nature biotechnology

THE SCIENCE AND BUSINESS OF BIOTECHNOLOGY

VOLUME 28 NUMBER 9 SEPTEMBER 2010
www.nature.com/naturebiotechnology

Castor bean genome

Synthetic biology meets antibody screening

Dissecting GPCR signaling

BIOPHARMACEUTICAL
Benchmarks

VOLUME 28 NUMBER 9 SEPTEMBER 2010

nature biotechnology

Seeds of the castor bean plant, *Ricinus communis*. Chan *et al*. report a draft genome of this oilseed crop, the first for a member of the Euphorbiaceae (p 951). Credit: Pablo Rabinowicz

FDA vote against Avastin for breast cancer, p 879

EDITORIAL

NEWS

BIOENTREPRENEUR

BUILDING A BUSINESS

OPINION AND COMMENT

CORRESPONDENCE

nature publishing group

Nature Biotechnology (ISSN 1087-0156) is published monthly by Nature Publishing Group, a trading name of Nature America Inc. located at 75 Varick Street, Fl 9, New York, NY 10013-1917. Periodicals postage paid at New York, NY and additional mailing post offices. **Editorial Office:** 75 Varick Street, Fl 9, New York, NY 10013-1917. Tel: (212) 726 9335, Fax: (212) 696 9753. **Annual subscription rates:** USA/Canada: US$250 (personal), US$3,520 (institution), US$4,050 (corporate institution). Canada add 5% GST #104911595RT001; Euro-zone: €202 (personal), €2,795 (institution), €3,488 (corporate institution); Rest of world (excluding China, Japan, Korea): £130 (personal), £1,806 (institution), £2,250 (corporate institution); Japan: Contact NPG Nature Asia-Pacific, Chiyoda Building, 2-37 Ichigayatamachi, Shinjuku-ku, Tokyo 162-0843. Tel: 81 (03) 3267 8751, Fax: 81 (03) 3267 8746. **POSTMASTER:** Send address changes to *Nature Biotechnology*, Subscriptions Department, 342 Broadway, PMB 301, New York, NY 10013-3910. **Authorization to photocopy** material for internal or personal use, or internal or personal use of specific clients, is granted by Nature Publishing Group to libraries and others registered with the Copyright Clearance Center (CCC) Transactional Reporting Service, provided the relevant copyright fee is paid direct to CCC, 222 Rosewood Drive, Danvers, MA 01923, USA. Identification code for *Nature Biotechnology*: 1087-0156/04. **Back issues:** US$45, Canada add 7% for GST. CPC PUB AGREEMENT #40032744. Printed by Publishers Press, Inc., Lebanon Junction, KY, USA. Copyright © 2010 Nature America, Inc. All rights reserved. Printed in USA.

VOLUME 28 NUMBER 9 SEPTEMBER 2010

COMMENTARY

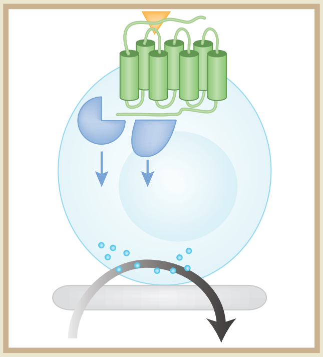
FEATURE

PATENTS

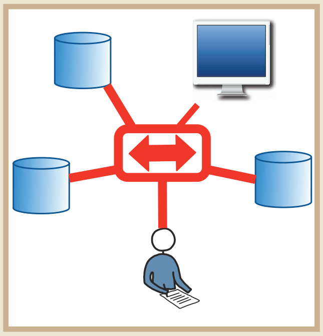
NEWS AND VIEWS

COMPUTATIONAL BIOLOGY

PERSPECTIVE



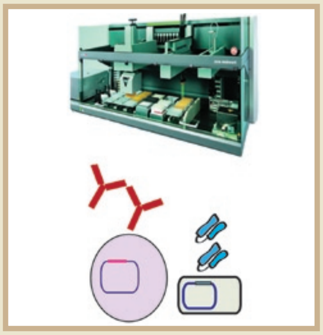
Deconvolving label-free GPCR signaling, p 928



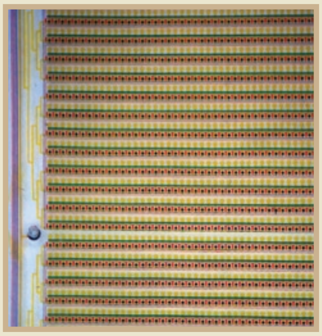
Language to describe biological pathways, p 935

Mushroom genomics, p 957

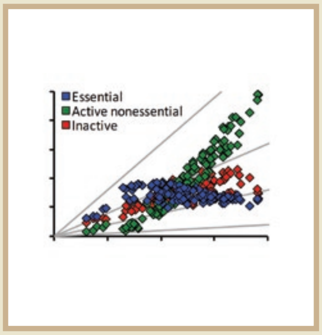
VOLUME 28 NUMBER 9 SEPTEMBER 2010



Sequencing substitutes for antibody screening, p 965



Microfluidics assay for protein-DNA binding, p 970



Reconstructing metabolic networks from genomes, p 977

RESEARCH

ANALYSIS

ARTICLES

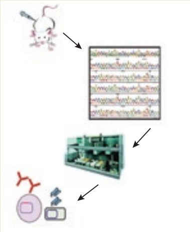
LETTERS

RESOURCE

CAREERS AND RECRUITMENT

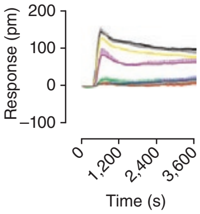
IN THIS ISSUE

Sequencing shortcuts antibody screening



Isolation of antigen-specific antibodies or antibody fragments, whether through B-cell immortalization or from recombinant libraries, generally requires laborious screening. Georgiou and colleagues circumvent this step by combining high-throughput sequencing of variable (V) genes expressed in bone marrow plasma cells with bioinformatic analysis of the V-gene repertoire. Consistent with the role of bone marrow plasma cells in making the majority of antibodies in circulation, the V genes encoding immunogen-specific antibodies typically represent 1–10% of all V-gene transcripts expressed in recently immunized mice. High-throughput sequencing enables the authors to identify the most abundant sequences encoding variable heavy and variable light chains. They then pair these based on their relative frequencies and use automated gene synthesis to express immunoglobulins or antibody fragments in mammalian cells or bacteria. When tested against three immunogens, ~80% of the antibodies are antigen specific. Much of the process is amenable to automation, and the costs involved should decline rapidly as sequencing and gene synthesis technologies become less expensive. **[Letters, p. 965]** *PH*

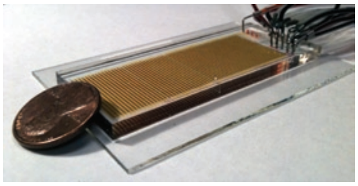
Label-free GPCR signaling analysis



Long mainstays in drug development, biochemical assays of labeled second-messenger signaling molecules are problematic because they can often be toxic to cells or interfere with intrinsic cellular processes. Kostenis and colleagues demonstrate approaches for studying G protein–coupled receptor (GPCR) signaling using label-free assays based on the cellular phenomenon of dynamic mass redistribution. Dynamic mass redistribution assays provide an optical readout of the integrated response of whole cells to drugs, but until now the whole-cell responses have not been mapped to individual G protein signaling pathways. Kostenis and colleagues describe methods for doing so using small molecules that inhibit or mask individual pathways. They apply these strategies to study endogenously expressed GCPRs in primary human keratinocytes, to dissect complex relationships between pathways and receptors, and to study G_{12}/G_{13} signaling, a pathway for which biochemical assays are not available. The ability to investigate GCPR signaling responses without labels is useful for characterizing the effects of drugs on therapeutically relevant primary cells. [**Analysis, p. 943; News and Views, p. 928**] *CM*

Written by Michael Francisco, Peter Hare, Craig Mak, & Lisa Melton

Protein–DNA binding on a chip

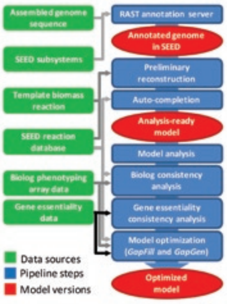


DeRisi, Quake and colleagues describe an updated microfluidic device for high-throughput measurements of the *in vitro* sequence preferences and binding affinities of DNA-binding proteins. In contrast to an earlier version of the device, the new chip requires no prior knowledge of the binding specificity of the protein of interest and can be used for motif discovery. It also has the unique ability to quantify binding affinities, which sets it apart from assays based on gel mobility shifts or standard protein microarrays. This advantage is conferred by microfluidic valves that trap binding reactions at equilibrium conditions, eliminating washing steps and allowing direct measurement of the concentration of soluble DNA available for binding. The authors determine the sequence specificity of 28 yeast transcription factors, of which two had not been characterized before and several others had been difficult to analyze using existing approaches. They also determine absolute binding affinities (K_d) by measuring concentration-dependent binding for four factors. These quantitative biophysical data on protein-DNA interactions should enable better understanding of transcriptional regulatory networks. [**Letters, p. 970**] *CM*

Castor bean genome

The castor bean plant (*Ricinus communis*) is an oilseed crop best known as a source of both castor oil and ricin, a highly toxic ribosome inactivating protein. Triacylglycerols extracted from its seeds contain 90% ricinoleate, a hydroxylated fatty acid that confers unique properties needed in high-temperature industrial lubricants and certain ingredients of medicinal and cosmetic products. Rabinowicz and colleagues use Sanger shotgun sequencing to assemble a draft (with 4.6-fold coverage) of the castor bean genome sequence. This is the first genome sequence for a member of the Euphorbiaceae, a large and commercially important plant family that includes cassava (*Manihot esculenta*), rubber plant (*Hevea brasiliensis*) and physic nut (*Jatropha curcas*). The authors find that many of the genes important to the metabolism of castor oil are single copy and thus potentially amenable to efforts to improve its yield and quality. In contrast, genome-wide sequencing reveals the ricin gene family to comprise 28 members—approximately fourfold more than previously suspected. The availability of the genome sequence may not only open the way for castor oil production without the associated risks posed by ricin but also enable biosecurity agencies to trace the origins of ricin, should it be used in bioterrorism. [**Articles, p. 951**] *PH*

Metabolic modeling made easier



Reconstructing a metabolic model from the genome sequence of an organism is a useful but arduous approach for predicting phenotypes from genotypes. Henry and colleagues describe a web-based resource that automates most of this process and apply it to create >100 new metabolic models of bacteria. The model reconstruction process was recently codified into 96 steps, 73 of which are now automated in the software described. The approach annotates the genes in a genome sequence, maps these genes to metabolic reactions, computes a 'biomass reaction' for simulating growth and then optimizes the model using several established techniques. After some final tweaking by hand, which is detailed in a supplementary tutorial, the model is ready for analysis. Henry and colleagues apply their tool across a diverse set of microbial genomes ranging from metabolically self-sufficient bacteria to parasites that rely on their hosts to provide many essential metabolic functions. The authors show how the models can be used to improve genome annotation and to assess global trends in microbial metabolism. [**Resource, p. 977**] *CM*

Blueprint of a model mushroom

Mushroom-forming fungi are an important source of food, industrially relevant enzymes and natural products with antitumor or immunostimulatory properties. Yet their biology remains poorly understood, in large part owing to both the difficulty in culturing most species on defined media and the dearth of molecular genetic tools to manipulate them. The genome sequence of *Schizophyllum commune*, reported by Wösten and colleagues, promises to enhance our understanding of the biology of basidiomycetes and thus potentially their utility in biotechnology. The 38.5-Mb genome of this wood-degrading fungus, the only mushroom-forming species amenable to targeted gene inactivation by homologous recombination, encodes 13,210 predicted genes. Of these, approximately a third encode proteins with no homolog in other fungi and more than half cannot be annotated with gene ontology terms. Genome-wide expression analysis reveals both differential gene expression in the four developmental stages studied and an extraordinary prevalence of antisense transcription: ~20% of all transcripts originate from an antisense transcript. It also leads the authors to identify two transcription factors that are shown, using targeted gene inactivation, to regulate mushroom formation. The genome encodes an impressive repertoire of lignin- and carbohydrate-degrading enzymes, which could facilitate more efficient production of lignocellulosic biofuel. [**Articles, p. 957**] *PH*

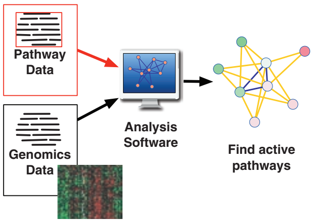
Patent Roundup

After 30 years of precedent, the issuance of gene patents has come under scrutiny in Europe and the US. Hoffenberg analyzes the courts' decisions and offers a possible solution in the new legal environment. [**Patent Article, p. 925**] *MF*

Recent patent applications in drug discovery [**New patents, p. 927**] *MF*

Two UK technology transfer offices have agreed to swap selected IP assets. The cooperation between Cancer Research Technology and the Medical Research Council Technology offices is intended to exploit IP packages in areas where the groups may be working but may not necessarily have the right models or clinical expertise. [**News, p. 883**] *LM*

Standards for pathway data



Incompatibility of data storage formats has hindered the sharing and analysis of digital representations of biological pathways. To address this problem, a large community of researchers present BioPAX, a standardized language for exchanging pathway data. The language has been under development for 8 years and is supported by >40 databases and software tools. BioPAX defines an ontology for describing biological entities, such as genes, proteins and metabolites, and the mechanisms by which they interact to accomplish biological processes. Signaling, metabolic, molecular interaction, genetic interaction and gene regulatory pathway information can be represented. It is hoped that using BioPAX to encode knowledge about pathways, an effort that is well underway, will make it easier to integrate pathway information from diverse sources and to develop computational tools that use this information to help interpret experimental data. [**Perspective, p. 935**] *CM*

Next month in nature biotechnology

- Focus on epigenetics
- Elasticity expands hematopoietic stem cells
- Chondrocyte differentiation of human ES cells

nature
biotechnology

www.nature.com/naturebiotechnology

EDITORIAL OFFICE biotech@us.nature.com
75 Varick Street, Fl 9, New York, NY 10013-1917
Tel: (212) 726 9200, Fax: (212) 696 9635
Chief Editor: Andrew Marshall
Senior Editors: Laura DeFrancesco (News & Features), Kathy Aschheim (Research), Peter Hare (Research), Michael Francisco (Resources and Special Projects)
Business Editor: Brady Huggett
Associate Business Editor: Victor Bethencourt
News Editor: Lisa Melton
Associate Editors: Markus Elsner (Research), Craig Mak (Research)
Editor-at-Large: John Hodgson
Contributing Editors: Mark Ratner, Chris Scott
Contributing Writer: Jeffrey L. Fox
Senior Copy Editor: Teresa Moogan
Managing Production Editor: Ingrid McNamara
Senior Production Editor: Brandy Cafarella
Production Editor: Amanda Crawford
Senior Illustrator: Katie Vicari
Illustrator/Cover Design: Kimberly Caesar
Senior Editorial Assistant: Ania Levinson

MANAGEMENT OFFICES
NPG New York
75 Varick Street, Fl 9, New York, NY 10013-1917
Tel: (212) 726 9200, Fax: (212) 696 9006
Publisher: Melanie Brazil
Executive Editor: Linda Miller
Chief Technology Officer: Howard Ratner
Head of Nature Research & Reviews Marketing: Sara Girard
Circulation Manager: Stacey Nelson
Production Coordinator: Diane Temprano
Head of Web Services: Anthony Barrera
Senior Web Production Editor: Laura Goggin

NPG London
The Macmillan Building, 4 Crinan Street, London N1 9XW
Tel: 44 207 833 4000, Fax: 44 207 843 4996
Managing Director: Steven Inchcoombe
Publishing Director: Peter Collins
Editor-in-Chief, Nature Publications: Philip Campbell
Marketing Director: Della Sar
Director of Web Publishing: Timo Hannay

NPG Nature Asia-Pacific
Chiyoda Building, 2-37 Ichigayatamachi, Shinjuku-ku, Tokyo 162-0843
Tel: 81 3 3267 8751, Fax: 81 3 3267 8746
Publishing Director — Asia-Pacific: David Swinbanks
Associate Director: Antoine E. Bocquet
Manager: Koichi Nakamura
Operations Director: Hiroshi Minemura
Marketing Manager: Masahiro Yamashita
Asia-Pacific Sales Director: Kate Yoneyama
Asia-Pacific Sales Manager: Ken Mikami

DISPLAY ADVERTISING display@us.nature.com (US/Canada)
display@nature.com (Europe)
nature@natureasia.com (Asia)
Global Head of Advertising and Sponsorship: Dean Sanderson, Tel: (212) 726 9350, Fax: (212) 696 9482
Global Head of Display Advertising and Sponsorship: Andrew Douglas, Tel: 44 207 843 4975, Fax: 44 207 843 4996
Asia-Pacific Sales Director: Kate Yoneyama, Tel: 81 3 3267 8765, Fax: 81 3 3267 8746
Display Account Managers:
New England: Sheila Reardon, Tel: (617) 399 4098, Fax: (617) 426 3717
New York/Mid-Atlantic/Southeast: Jim Breault, Tel: (212) 726 9334, Fax: (212) 696 9481
Midwest: Mike Rossi, Tel: (212) 726 9255, Fax: (212) 696 9481
West Coast: George Lui, Tel: (415) 781 3804, Fax: (415) 781 3805
Germany/Switzerland/Austria: Sabine Hugi-Fürst, Tel: 41 52761 3386, Fax: 41 52761 3419
UK/Ireland/Scandinavia/Spain/Portugal: Evelina Rubio-Hakansson, Tel: 44 207 014 4079, Fax: 44 207 843 4749
UK/Germany/Switzerland/Austria: Nancy Luksch, Tel: 44 207 843 4968, Fax: 44 207 843 4749
France/Belgium/The Netherlands/Luxembourg/Italy/Israel/Other Europe: Nicola Wright, Tel: 44 207 843 4959, Fax: 44 207 843 4749
Asia-Pacific Sales Manager: Ken Mikami, Tel: 81 3 3267 8765, Fax: 81 3 3267 8746
Greater China/Singapore: Gloria To, Tel: 852 2811 7191, Fax: 852 2811 0743

NATUREJOBS naturejobs@us.nature.com (US/Canada)
naturejobs@nature.com (Europe)
nature@natureasia.com (Asia)
US Sales Manager: Ken Finnegan, Tel: (212) 726 9248, Fax: (212) 696 9482
European Sales Manager: Dan Churchward, Tel: 44 207 843 4966, Fax: 44 207 843 4596
Asia-Pacific Sales & Business Development Manager: Yuki Fujiwara, Tel: 81 3 3267 8765, Fax: 81 3 3267 8752

SPONSORSHIP g.preston@nature.com
Global Head of Sponsorship: Gerard Preston, Tel: 44 207 843 4965, Fax: 44 207 843 4749
Business Development Executive: David Bagshaw, Tel: (212) 726 9215, Fax: (212) 696 9591
Business Development Executive: Graham Combe, Tel: 44 207 843 4914, Fax: 44 207 843 4749
Business Development Executive: Reya Silao, Tel: 44 207 843 4977, Fax: 44 207 843 4996

SITE LICENSE BUSINESS UNIT
Americas: Tel: (888) 331 6288 institutions@us.nature.com
Asia/Pacific: Tel: 81 3 3267 8751 institutions@natureasia.com
Australia/New Zealand: Tel: 61 3 9825 1160 nature@macmillan.com.au
India: Tel: 91 124 2881054/55 npgindia@nature.com
ROW: Tel: 44 207 843 4759 institutions@nature.com

CUSTOMER SERVICE www.nature.com/help
Senior Global Customer Service Manager: Gerald Coppin
For all print and online assistance, please visit **www.nature.com/help**
Purchase subscriptions:
Americas: *Nature Biotechnology,* Subscription Dept., 342 Broadway, PMB 301, New York, NY 10013-3910, USA. Tel: (866) 363 7860, Fax: (212) 334 0879
Europe/ROW: *Nature Biotechnology,* Subscription Dept., Macmillan Magazines Ltd., Brunel Road, Houndmills, Basingstoke RG21 6XS, United Kingdom. Tel: 44 1256 329 242, Fax: 44 1256 812 358
Asia-Pacific: *Nature Biotechnology,* NPG Nature Asia-Pacific, Chiyoda Building, 2-37 Ichigayatamachi, Shinjuku-ku, Tokyo 162-0843. Tel: 81 3 3267 8751, Fax: 81 3 3267 8746
India: *Nature Biotechnology,* NPG India, 3A, 4th Floor, DLF Corporate Park, Gurgaon 122002, India. Tel: 91 124 2881054/55, Tel/Fax: 91 124 2881052

REPRINTS reprints@us.nature.com
Nature Biotechnology, Reprint Department, Nature Publishing Group, 75 Varick Street, Fl 9, New York, NY 10013-1917, USA.
For commercial reprint orders of 600 or more, please contact:
UK Reprints: Tel: 44 1256 302 923, Fax: 44 1256 321 531
US Reprints: Tel: (617) 494 4900, Fax: (617) 494 4960

nature biotechnology

The identity problem

The US Food and Drug Administration (FDA) decision to approve a generic heparin derivative without clinical safety or efficacy data raises the possibility that clinical trials might not always be required for the approval of follow-on biologics.

At the end of July, the FDA granted marketing approval to an anticoagulant, the low-molecular-weight heparin (LMWH) enoxaparin sodium injection. The product, co-developed by Momenta Pharmaceuticals and Sandoz, was approved under generics regulations and was designated equivalent to and substitutable for Lovenox, manufactured by Sanofi-aventis. The FDA's decision to consider enoxaparin under the Abbreviated New Drug Application (ANDA) pathway has raised some eyebrows. It not only runs counter to the established European regulatory framework for biosimilars but also suggests that the FDA might, in some cases at least, not require extensive clinical trials for follow-on biologics.

Enoxaparin is not the first generic biologic approved in the United States. Six other generic biologics have been approved under ANDAs, a major advantage of which is that no clinical trial data are required and products can be designated therapeutically equivalent to a brand. Under this path, FDA can also designate a generic as automatically substitutable for the brand in the pharmacy. Of the biologics, only Lovenox and calcitonin generics have been approved under an ANDA and received brand substitutability status. Sandoz's Omnitrope (human growth hormone) was licensed using a New Drug Application (NDA) filed under 505(b)2, a regulatory pathway that allows submission of clinical data demonstrating safety and efficacy but does not offer substitutability.

When Momenta filed its ANDA for enoxaparin in August 2005, the challenge facing the FDA was how to demonstrate its therapeutic equivalence to Lovenox. The problem, ostensibly, is that enoxaparin is complex and difficult to define chemically. Its manufacture involves the alkaline depolymerization of heparin from pig intestinal mucosa. Heparin itself is a mixture of linear polysaccharide chains consisting of repeating disaccharide units composed of glucuronic or iduronic acid and N-sulfated or N-acetylated glucosamine; during its depolymerization to enoxaparin, additional distinctive chemical modifications may occur. The resulting final product has a mean molecular mass of 4.5 kDa but consists of polysaccharide chains varying in length, composition and distribution. The brand drug has never been fully characterized, nor has the contribution of its components to therapeutic efficacy been established.

In information that accompanied the approval, the FDA indicated how it compared the brand and generic versions, an assessment with possible implications for other biologics. The agency's analysis was based on five criteria.

First, the comparison involved gross physicochemical properties (molecular mass distribution and overall chemical composition): the FDA wanted to see that similar oligosaccharide chain lengths were present in the same relative abundance in generic and brand enoxaparin. Second, the drug sources were compared: the source heparin had to have a similar distribution of disaccharide building blocks, and beta-elimination of the heparin benzyl ester had to be shown during depolymerization. A third comparison looked at the products' molecular nature: besides confirming the presence of a pharmacologically important 1,6-anhydro ring at the ends of 15–25% of enoxaparin chains, the FDA looked at the spectrum of disaccharide building blocks and the oligosaccharide lengths and sequences—factors affected by the temperature, depolymerization time and other conditions of preparation.

Besides these physiochemical analyses, the FDA required not only biological and biochemical assay data to demonstrate equivalent anticoagulant activity *in vitro* and *in vivo* but also pharmacodynamic data from healthy human volunteers. Taking all the data together, FDA concluded "that generic enoxaparin will have the same active ingredient components as those of Lovenox's enoxaparin (within the context of its variability), even though the contribution of each component has not been fully elucidated."

As immune reactions, such as pruritus, urticaria and anaphylactic and anaphylactoid responses, have been observed with Lovenox, the agency also requested that manufacturers demonstrate equivalent immunogenicity for generic enoxaparin. Again, it stopped short of requiring clinical studies, accepting data from *in vitro* and *ex vivo* assays and from animal studies.

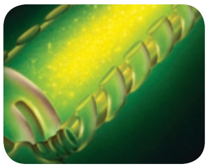
Overall, the generic enoxaparin approval indicates that current analytical technology and integrated, multivariate data analysis can convince the FDA of the equivalence of two complex, biologically derived preparations. The FDA has determined that the product is the *same* if it meets its criteria of identity (even if the product is not identical)—a substantial departure from European guidelines for LMWH biosimilars designed for products that contain a *similar* active ingredient. The good news, it seems, is that as long as the FDA is satisfied that the data package sufficiently establishes the sameness of the active ingredient, the need for clinical data diminishes.

The billion-dollar question—as yet unanswered—is whether the FDA will consider similar supporting data for complex biologics approved under the Biologic License Application pathway as sufficient to demonstrate therapeutic equivalence without large clinical trials. As the complexity of the biologic being reproduced becomes greater—from peptides, to hormones and growth factors, and all the way to monoclonal antibodies—the capacity of current technology is likely to approach its limits. In this respect, another generic product may soon provide an answer.

An ANDA for a generic of Copaxone, Teva's treatment for multiple sclerosis, has been before the FDA since July 2008. Copaxone is perhaps the quintessential complex peptide drug, a heterogeneous mixture containing a huge number of synthetic polypeptides. Its analysis will certainly push the envelope for current technology and illustrates how important sophisticated technical capability will be to sponsors wishing to work in this area.

Even the Copaxone case, though, may still not provide much guidance for recombinant biologics. Like enoxaparin, Copaxone's complexity largely stems from the active ingredient. In contrast, variation in most recombinant products—and thus the analytical challenge—arises not in the active ingredient but in post-translational modifications, proteolysis, oxidation and aggregation that occur during manufacture, formulation and storage. All of which adds up to a different challenge again.

IN this section



Avastin's commercial march suffers setback

Roche

Avastin may not become the world's best-selling cancer drug after all. Genentech had projected US Avastin sales reaching $10 billion by 2015 but the FDA has just rejected the drug for breast cancer.

Until July this year, when it slammed into a negative vote from a US Food and Drug Administration (FDA) advisory committee over its use in metastatic breast cancer, Avastin (bevacizumab) had been on an impressive roll. As the first drug specifically designed to block angiogenesis by inhibiting vascular endothelial growth factor (VEGF), it has accumulated several registered indications over only a few years. Its first approvals were based on significant overall survival improvements in colorectal and lung cancers. From there, S. San Francisco–based Genentech gained one indication after another for the drug in various solid tumors (**Table 1**). Two of these were through the FDA's accelerated approval process. Now, many are wondering whether the breast cancer vote is just a bump in the road or whether it signals the end of Avastin's winning streak.

Some of the drug's more recent approvals were based on objective response and progression-free survival (PFS), less rigorous criteria than overall survival. Avastin also comes with considerable side effects, and there is no effective response-biomarker yet. Those features seem to have helped bolt the door shut at the FDA's recent Oncology Drugs Advisory Committee (ODAC) meeting where new data from breast cancer studies were reviewed. The committee voted overwhelmingly against approval. "There was no convincing evidence that Avastin was clinically beneficial in the studies we looked at," says Wyndham Wilson, who chaired the committee. Wilson is head of the Lymphoma Therapeutics Section and senior investigator at the National Cancer Institute, Bethesda, Maryland.

Avastin was granted accelerated approval as a treatment for breast cancer in 2008 based on the E2100 trial (*N. Engl. J. Med.* **357**, 2666–2676, 2007). Genentech followed up with data from the AVADO and RIBBON1 studies (*Breast Cancer Res. Treat.* **122**, 181–188, 2010) to support full approval. Each study tested the drug with a different chemotherapy regimen. All three studies showed improvements in PFS.

"Anytime you are looking at a surrogate endpoint, like PFS, you have to be very cautious," Wilson says. The committee was looking for evidence that the patients were either living longer or that there were "meaningful" clinical improvements that offset the drug's side effects. They found neither. "I hope physicians will take heed of the committee's decision and do not continue prescribing the drug," Wilson adds. The FDA usually follows ODAC committee recommendations.

As part of Genentech's response to the decision, Sandra Horning, senior vice president, global head, clinical development hematology/oncology, noted in a press release, "we are dis-

Table 1 Avastin current cancer approvals (worth $5.7 billion in 2009)

Disease	Date	Context	Endpoint
Renal cell carcinoma	July 2009	Given in combination with interferon alfa for patients with metastatic disease	Progression-free survival. No statistically significant advantage in overall survival was documented.
Glioblastoma	May 2009	Accelerated approval for use as a single agent for patients with progressive disease after prior therapy	Durable objective response rates as demonstrated using WHO radiographic criteria and the presence of stable or decreasing corticosteroid use.
Breast cancer	February 2008	Accelerated approval for use with paclitaxel (Taxol) in individuals with metastatic, HER2-positive disease who have not failed prior chemotherapy	Progression-free survival. No statistically significant advantage in overall survival was documented.
Non–small cell lung cancer	October 2006	Labeling extension allowing the drug to be administered in combination with carboplatin and paclitaxel as a first-line treatment for unresectable, locally advanced, recurrent or metastatic nonsquamous disease	Overall survival was statistically significant.
Colorectal cancer	June 2006	Labeling extension for use with 5-fluorouracil–based chemotherapy, for second-line treatment of metastatic disease	Overall survival was statistically significant.
Colorectal cancer	February 2004	First-line treatment for metastatic disease	Overall survival was statistically significant.

IN brief

Sanofi-aventis snaps up microRNA maker

Regulus

Regulus headquarters.

Sanofi-aventis has ventured into microRNA (miRNA) territory by entering a partnership with Regulus Therapeutics worth $750 million. Most of the deals Sanofi-aventis has consummated of late, including—if it pulls it off— the acquisition of Cambridge, Massachusetts–based Genzyme, show its eagerness to catch up with its pharma peers in shoring up biologicals for its portfolio (*Nat. Biotechnol.* **27**, 581–582, 2009). The agreement with Regulus of San Diego, announced in June, is the first time the Paris-based pharma has dipped its toe into miRNA-targeted therapeutics. Sanofi-aventis will pay Regulus $25 million upfront to gain access to the biotech's miRNA platform, including their fibrosis program targeting miRNA-21. Sanofi is not the first big pharma to invest in miRNA, however. That honor goes to London-based GlaxoSmithKline, for an April 2008 agreement, extended in February, also with Regulus. In two years since GlaxoSmithKline signed its $600 million deal with Regulus, the field has "exploded," says Chris Hillier, professor at Glasgow Caledonian University, and the founder of miRNA specialist, Sistemic, located in Glasgow, Scotland. "The field has moved from embryonic, where it was clearly important but poorly understood [*Nat. Biotechnol.* **25**, 631–638, 2007], to seeing application in drug discovery, diagnostics, quality control and now therapeutics," Hillier says. Crucially, in May 2008, Santaris Pharma of Horsholm, Denmark, announced that the first miRNA-targeted therapeutic, SPC3649, entered phase 1 trials in hepatitis C infections. Further validation came from a discovery deal Santaris signed with Wyeth of New Jersey, in January 2009 that included miRNA targets, and an agreement signed in June with Boulder, Colorado–based MiRagen in cardiovascular diseases. New insights into how miRNAs influence mRNA to prevent translation have confirmed their key role in biological cascades, and thus, in complex disease. As drugs, miRNAs are deemed exceptionally attractive, because they can control multiple genes and biological pathways. Furthermore, worries that this wide-ranging influence would lead to off-target effects have been lifted by evidence that particular miRNAs affect pathways in an orchestrated and specific way. The finding that small molecules—in addition to oligonucleotides—can modulate miRNAs is another advantage. "Pharma companies can start to look at small molecules that change miRNAs in particular ways, switching phenotype A to phenotype B," says Hillier. "In other words, [pharma companies] can manipulate miRNA using technologies they are comfortable with." *Nuala Moran*

appointed by the committee's recommendation and believe Avastin should continue to be an option for women with this incurable disease," adding, "We will continue to discuss the data from the more than 2,400 women who participated in three phase 3 studies with the FDA."

Analysts don't seem to think the breast cancer decision will make much of an impact on Avastin sales. "In the worst-case scenario, we are hypothesizing that they will lose $80 million to $100 million per quarter in sales, but it is more likely to be in the $40 to $50 million range," says Ed Kissel, vice president, Quantitative Analysis at IntrinsiQ, Waltham, Massachusetts. He points out that doctors started using the drug in breast cancer, even before the accelerated approval was granted. "A lot of it comes down to reimbursement and any barriers that might be set by CMS [Centers for Medicare & Medicaid Services] or third-party payers to restrict access," Kissel says.

The larger issue is whether the drug has reached its sales peak. In the US, Avastin currently rakes in about $800 million a quarter for Genentech. More than 1,100 Avastin-related clinical trials are listed at http://www.clinicaltrials.gov/ (ClinicalTrials.gov), and ~25% of those are in phase 3. Roche and Genentech are sponsoring about 450 trials with the drug themselves. "They've had a great marketing strategy," Kissel says. "Avastin didn't create a new market," he explains. "They could say 'Use our product in conjunction with standard of care and you'll get better progression-free survival', and that was an easy sell for oncologists."

If regulators become more demanding, the company will have a much tougher time getting additional approvals. A patient advocate was the only ODAC member who voted for Avastin at the breast cancer committee meeting. But some patients are unenthusiastic about the drug. Fran Visco, president of the National Breast Cancer Coalition and a 22-year survivor of the disease says, "We are looking for something that has a major impact, and in no cancer has this drug had a major impact." Visco adds that she is concerned by the drug's side effects and "animal studies that suggest it may actually cause tumors to spread."

More data on breast cancer are sure to come, as more than 150 breast cancer trials are currently studying the drug, according to ClinicalTrials.gov. Among the most highly watched is BETH (treatment of HER2 positive breast cancer with chemotherapy plus trastuzumab versus chemotherapy plus trastuzumab plus bevacizumab). One of the investigators in this trial is Dennis Slamon, director of clinical translational research at UCLA's Jonsson Comprehensive Cancer Center in Los Angeles, and one of the pioneering clinicians involved in the development of Genentech's Herceptin (trastuzumab). That drug has had a major impact in breast cancer, and Slamon says that compelling laboratory and clinical evidence suggests that pairing it with Avastin could be synergistic. "HER2-positive patients have astronomical levels of VEGF," he says, pointing to a study conducted by his group measuring HER2neu and VEGF isoform levels in breast tumor samples (*Clinical Cancer Research* **10**, 1706–1716, 2004).

It thus seems possible that HER2 expression may be a surrogate marker for VEGF overexpression. A phase 2 trial of Herceptin/Avastin (no chemotherapy) in women with HER2-positive, recurrent or metastatic breast cancer showed a 48% objective response rate. "It's an active regimen and I was ready to try a biologics-only phase 3," he says. However, BETH includes chemotherapy because "that is just too big a step" for most investigators to consider now, Slamon adds.

Slamon continues, "The problem with Avastin is that it's being used in a nonselective way." Without a marker that tells which patients are most likely to benefit "the effect is masked," he says. The search for markers of response to Avastin may be a hot topic, but still only about 150 of the current trials with the drug listed on ClinicalTrials.gov appear to involve biomarkers. The ODAC committees increasingly seem to want biomarker data. "It is incumbent when studying agents with target specificity to try and understand if there are subtypes," says Wilson. The company is also aiming for an approval in ovarian cancer. Phase 3 trial results have so far showed a significant improvement in PFS. No survival benefit was seen, but the study is not finished yet.

The other massive uncertainty in this picture revolves around CATT (the Comparison of Age-Related Macular Degeneration Treatments Trials). Results from those studies, due in 2011, could bring new market share to Avastin, but at a substantial cost to Genentech. CATT is a head-to-head comparison of Avastin and Genentech's Lucentis (ranibizumab), which is very similar to Avastin in all but price. Lucentis was approved for wet-AMD in 2006, but some ophthalmologists have used Avastin instead, because it is so much cheaper. A release about CATT suggests Medicare alone would save $3 billion per year if patients on Lucentis switched to Avastin.

Finally, there are many other VEGF inhibitors in development. But Kissel thinks it will be even tougher for "me too's" to grab any of Avastin's market share because the standard of care has improved substantially. "They won't be competing against chemo anymore," he says, "The standard now is chemo plus a biologic, and PFS [is] longer."

Malorye Allison *Acton Massachusetts*

Pfizer explores rare disease path

The world's largest pharmaceutical company is thinking small by setting up a dedicated rare disease R&D unit in Cambridge, Massachusetts. Pfizer's new group, announced in June, will focus initially on treatments for muscular dystrophy and other serious diseases caused by genetic mutations, in addition to hemophilia, for which the company already markets a treatment. Pfizer's incursion into rare diseases is the latest signal that businesses built around niche indications are no longer the exclusive domain of biotech enterprises, such as Cambridge, Massachusetts–based Genzyme. The New York–based pharma company now joins Merck, of Whitehouse Station, New Jersey, GlaxoSmithKline (GSK) of London and Novartis of Basel, all of which have initiated rare disease programs in recent years. Whether such initiatives will remain small in scale—as part of numerous initiatives under way in big pharma to diversify their businesses—or expand to such an extent that they rival programs at biotech companies that have traditionally targeted rare disease is an open question.

"[Pharma] companies are realizing that for niche diseases, you can charge a significant premium," says Simos Simeonidis, managing director senior biotechnology analyst Rodman & Renshaw. Patient numbers are small by pharma standards, but because the drugs are lifesaving, even though they are expensive, insurers must pay up. Investors and corporate decision makers are slowly waking up to the potential value of these drugs that some have started calling 'minibusters'.

Shelly Harrison

Profit in niche indications. Pfizer's Jose Carlos Gutierrez Ramos, senior vice president of Biotherapeutics Research and Development group, will oversee the new research unit focused on rare diseases.

In February, GSK announced plans to form a new stand-alone rare diseases unit. The new unit, described in a company release as operating under a "lean structure," will work with the company's existing capabilities and seek strategic collaborations with other companies. Analysts, however, have been reserved in their assessments of these initiatives, noting that many of them are very small, resembling almost a token effort, rather than a full commitment to rare disease research. Indeed, Pfizer's new initiative, still in its very earliest stages, consists of two employees and a few laboratory benches.

But with the price incentive, rare disease research programs represent a good opportunity. Simeonidis gives Alexion's first product Soliris (eculizumab), approved for treating paroxysmal nocturnal hemoglobinuria, as an example. The average price for the lifesaving drug developed by the Cheshire, Connecticut–based company, is roughly $400,000 a year, and insurance companies are paying for it. At the same time, insurance companies in some countries are refusing to pay for drugs like Avastin (bevacizumab), which costs $50,000–$100,000 per year, that may extend life by a couple of months (this issue, page 879).

Tax incentives and seven years' protection against competition, spelled out in the Orphan Drug Act of 1983, encouraged biotech companies to pursue rare diseases or orphan indications. In previous decades, such niche markets were considered too small for multinational pharmaceutical companies that had large marketing arms to drive billion dollar sales of drugs for common ailments in the general population. But as a singular pursuit of the blockbuster model and me-too drugs becomes unsustainable, pharma companies are looking more closely at niche opportunities. What's more, an advantage of a more scientific nature is also beginning to attract large players. As rare diseases are typically caused by a known genetic variant, on paper at least, developing a cure should be more straightforward than for many common, multifactorial diseases with mass markets, such as type 2 diabetes. "The progress that's been made in genetically describing many of these diseases allows us to be [in a better position] to find drugs that can work for the diseases," says Ed Mascioli, vice president, biotherapeutics R&D, orphan and genetic diseases for Pfizer.

The scientific rationale is strengthening risk-benefit calculations and pharma are jumping on board. Damien Conover, a senior stock analyst

IN brief

Provenge twists again

Just when Dendreon thought it had reached the promised land, with the approval of its prostate cancer vaccine Provenge (sipuleucel-T), the Seattle-based company is back in the hot seat. In July, the Center for Medicare and Medicaid Services (CMS), which oversees Medicare, announced an investigation into whether it should pay for the cancer vaccine, approved in April (*Nat. Biotechnol.* **28**, 531–532, 2010). It sounds fair that a treatment regime costing $93,000 that offers 4 months' increase in median survival should be under such scrutiny, especially given that 75% of the potential patients, by Dendreon's reckoning, would be covered by Medicare. Yet, Dendreon consultant Jayson Slotnik of Foley Hoag in Washington, D.C. points out that this kind of analysis is rarely undertaken so soon after approval. "What new data will they be looking at?" he asks. CMS will not discuss an ongoing investigation, leaving to conjecture the reason for their decision to pursue this course. In a letter to CMS, Dendreon requests that the investigation be abandoned, or, at the least, brought to a speedy conclusion (the process takes a year) based on consistent results from four clinical trials, which recently appeared in the *New England Journal of Medicine*. Even that did not go smoothly, as an accompanying editorial questioned aspects of the trials. Meanwhile, on August 6, the FDA issued a warning to Dendreon about misleading promotions. *Laura DeFrancesco*

Lilly snaps up Alnara

Eli Lilly of Indianapolis, has acquired Alnara Pharmaceuticals, a two-year-old startup with a single drug—an enzyme supplement—currently under review by the US Food and Drug Administration. Alnara's lead product, Trizytek (liprotamase), is a nonporcine pancreatic enzyme therapy for patients with cystic fibrosis and other conditions in which the pancreas fails to produce enough enzymes needed to digest and absorb food. With the new deal, Lilly will gain a foothold in the enzyme replacement market, whereas the Cambridge, Massachusetts–based Alnara will benefit from the larger company's experience in the US, particularly in regulatory affairs, to help steer Trizytek into the clinic. "The deal sits with Lilly's new strategy of looking for niche markets where there are low levels of competition and less likelihood of pricing pressure," observes William Kridel, managing director of specialist investment banking group Ferghana Partners in New York. Kridel adds that Lilly may go on to do other such specialty deals. Trizytek contains protease, amylase and lipase enzymes made by microbial processes, and will be offered as an alternative to existing products made with pig enzymes. Alnara hopes to have the product on the market late this year. Trizytek once belonged to Altus Pharmaceutics, which folded following the recent economic downturn. Altus transferred rights to liprotamase to the Cystic Fibrosis Foundation Therapeutics, which were then repurchased by Alnara. The terms of the deal were not disclosed. *Susan Aldridge*

for Morningstar, points out that "what Pfizer is doing with their new rare disease unit is similar to what we're seeing at GlaxoSmithKline and what we've already seen at Novartis. It is a little bit emblematic of what we're seeing across the industry, which is a shift toward rare diseases, away from the primary care model that had served the big pharmaceutical firms pretty well over the last couple of decades."

Until now, success stories in rare diseases have been the province of small biotech. BioMarin, of Novato, California, for example, has three drugs on the market, all approved for rare disease indications: Naglazyme (galsulfase) for the treatment of mucopolysaccharidosis VI (MPS VI), Aldurazyme (laronidase) for MPS I and Kuvan (sapropterin dihydrochloride) for phenylketonuria. Elsewhere, Brussels-based chemical company Solvay succeeded with Creon, its pancreatic lipase therapy for cystic fibrosis, which led to its acquisition by Abbott Laboratories of Abbott Park, Illinois. And since 1994, Genzyme has enjoyed a monopoly on Gaucher disease treatment with its drug Cerezyme.

Although one of the advantages for biotech companies that have traditionally targeted orphan disease indications has been the lack of competition from big pharma, the entry of multinational drug companies into the area might not be all bad news. According to Simeonidis, many biotech companies have steered away from rare diseases because investors have preferred to emphasize larger markets, which they perceived as providing greater product returns and being more attractive for a potential pharma buyout. Simeonidis believes that having big pharma in the sandbox could be immensely helpful for biotech companies seeking investor support for rare disease indications. And of course, playing with big pharma companies translates to increased opportunities for partnerships, licensing agreements and acquisitions.

Competition from pharma will affect mostly big biotech, analyst Conover believes. Genzyme is currently the prime example of a biotech company that, in the long term, may experience competitive pressure, as the new Pfizer rare disease unit will be focusing on Gaucher disease. Since 1994 Genzyme has offered the only effective therapy for Gaucher disease.

Simeonidis thinks Pfizer will find it hard to compete with Genzyme at least in the near term. "If you look three to five years down the line, you could see a company like Pfizer having an advantage over Genzyme in this arena because of the difference in the amount of resources available... but right now, I would not think a company like Biomarin or Genzyme would be threatened by the presence of Pfizer."

Biotechs enjoy a competitive edge when it comes to orphan drug marketing, as it is very different from that of mass market, blockbuster medicines. "Case management is very important, as is keeping track of individual patients and helping them navigate the reimbursement maze," says Joseph Schwartz, a bioanalyst for Leerink Swann in Boston. Many analysts question whether Pfizer and other large pharma are making the investment required to develop and market a successful minibuster orphan drug therapy, or simply shot-gunning different biotech business strategies to see if any of them stick. Biotech-like ideas have been known to fizzle in pharma hands, including GSK's EpiNova and New York–based Pfizer's Biotherapeutics and Bioinnovation Center in San Francisco, which was shuttered in 2009.

Industry observers also point out that pharma seems to undermine their own rare disease initiatives by putting them under the direction of junior scientists who do not have the authority to leverage company resources, or failing to invest enough money and manpower in the unit. One company that has been a trendsetter in rare diseases is Novartis. The pharma stands out from the pack in having not an isolated rare diseases unit within the company, but an innovative, company-wide rare diseases program that investigates small indications first, using it as a launch pad for common diseases after proven success. For example, Novartis's Ilaris (canakinumab), recently approved to treat familial cryopyrin-associated periodic syndrome (CAPS), has been granted orphan drug status. Ilaris is now also being developed as a possible therapy for type 2 diabetes and chronic obstructive pulmonary disease, which are potential blockbuster applications.

Patient recruitment remains a formidable challenge, however, and that may be one reason why companies tend to shy away from rare diseases. Timothy Wright, global head of translational sciences for Novartis, says, "I wouldn't say we've overcome that, but we've taken on the challenge, working with support groups, patient advocacy groups and with key investigators in the field who have large groups of patients with certain diseases and are connected to networks."

Pharma companies hoping to emulate Novartis's success in rare diseases will have to either develop these capabilities and resources within the company, or acquire them. Biotechs are brimming with assets related to rare diseases, along with intellectual property, so this trend toward orphan diseases in pharma could bring new opportunities rather than a competitive threat.

Catherine Shaffer *Ann Arbor, Michigan*

India's Cipla sets sights on Avastin, Herceptin and Enbrel

Indian generic giant Cipla has begun its foray into biosimilars with an eye firmly on biotech's blockbusters. The Mumbai-based chemical generics manufacturer is taking aim at top-selling biologics—Roche's Avastin (bevacizumab) and Herceptin (trastuzumab) and Pfizer/Amgen's Enbrel (etanercept)—which last year brought in a combined $17 billion. With no expertise in biologics, Cipla has had to shop around to build its biologic capabilities. To this end, on June 15, the company made a $65 million investment in Shanghai-based BioMab and Indian firm MabPharm located in Goa. Although low-cost versions of biotech's most successful brand biologics represents a substantial opportunity, Cipla will be not only playing catch-up but also competing for market share with multinational pharmaceutical companies that have already ramped up their capacity and expertise in producing biologics (*Nat. Biotechnol.* **27**, 299–301, 2009). On the other hand, if major generics players from emerging economies meet the technical standards required for entry into the Western biosimilars market, this may force big pharma to price their follow-on products more competitively.

STR/AFP/Getty Images.

Cipla built its $1.17 billion generics business by offering cheap copies of anti-AIDS drugs. The Mumbai-based firm now aims to copy ten monoclonal drugs against rheumatoid arthritis, cancers and allergic asthma.

"This is a major decision," says Yusuf Hamied, Cipla's chairman, referring to the June announcement. The deal will be setting a precedent in that a player with very little presence in biotech extends its strategy to biologics by gearing up for antibody production. "A time will come when the world will be selling only biotech drugs. When that day arrives Cipla will be prepared," says Hamied.

The news was also welcomed by William Haddad, founder and long-time chair of the Generic Pharmaceutical Association in Arlington, Virginia, and currently chairman and CEO of New York–based Biogenerics. "The Cipla-China BioMab agreement should send shivers up the backs of the brand biotech companies as it undermines all the anti-generic biotech arguments," he said. "For me the great irony is that the third world will have access to lifesaving biotech medicines that are affordable, whereas patients in the so-called developed nations will not have access to them at prices they can afford or that insurance companies will cover."

At the outset, the Cipla-China partnership is targeting ten monoclonal antibody (mAb) drugs and fusion proteins against rheumatoid arthritis, cancers and allergic asthma for marketing in India and China, particularly drugs that are presently not protected by patent or whose patent term is due to expire.

"We are very happy to be partnering with Cipla," says Xu Shengping, CEO of Shangai-based BioMabs, which is setting up a new biosimilar facility in Shanghai under the collaboration with Cipla. Their technology will also be used by MabPharm's facility in Goa. "We expect to launch the first product at the end of 2011," Hamied says.

The Indian biosimilar space is already strewn with a handful of local firms developing and marketing a broad range of products (**Table 1**). The space has been bolstered by government incentives and the prospect of less stringent approval requirements than in the US and Europe. "We have a special scheme for biosimilar makers; it even goes as far as fully supporting clinical trials," says Department of Biotechnology secretary Maharaj Kishan Bhan. For instance, phase 1 and 2 trials for biogenerics have been waived by the drugs controller general of India under the Ministry of Health and Family Welfare (US Food and Drug Administration's Indian counterpart), and phase 3 trials with 100 patients are enough for establishing bioequivalence. This helps bring down development costs to $10–$20 million, enabling Indian companies to offer their biosimilars 25–40% cheaper than branded biologics, says Syamala Ariyanchira an independent pharma-biotech industry analyst in Bangalore.

Indian firms may be rushing into the biosimilar space now, but their interest is on the second wave of blockbuster products that will go off-patent between 2012 and 2016 in Europe and the US. Such products, which include mAbs and fusion proteins, present several challenges compared with simpler biologics, warns Jay

IN brief

TTO patent swap

Two medical research funders have agreed to exchange selected intellectual property (IP) assets in a bid to boost commercialization. Cancer Research Technology (CRT) of London and the UK's Medical Research Council Technology (MRCT) will offer each other the rights to discoveries funded by their respective parent organizations, the charity Cancer Research UK and the government-backed Medical Research Council (MRC). As part of the exchange, CRT will work on an MRC-derived project in cancer, whereas MRCT will reciprocate outside oncology with revenue sharing to be agreed on a case-by-case basis. MRCT and CRT are both 'super-TTOs', technology transfer offices, in that both run drug development facilities. CRT's Development Laboratory and MRCT's Centre for Therapeutics Discovery each produce preclinical data packages on small molecules and biologicals to add value to the original patented IP. Although the agreement between the two commercialization arms is broad in principle, the first swaps are likely to concern projects that would feed these internal development pipelines. According to Keith Blundy, CEO of Cancer Research Technology, "There are projects that both groups are already working on, but we are not necessarily 'kitted out' in the relevant clinical area. We may not have the biological models needed to progress the project." *John Hodgson*

Brazil bans Bayer

A judge has prohibited Bayer Cropscience from marketing Liberty Link corn, a genetically modified crop resistant to Ignite and Liberty herbicides, in Brazil. If the Leverkusen, Germany–based company fails to suspend marketing, planting, transportation and import immediately, it will be fined R$50,000 ($28,500) a day. This ruling issued in July by an environmental court in the southern state of Parana is only the second time a Brazilian court has overturned a commercial GM crop already approved by the country's technical commission on biosafety (CTNBio), says the commission's coordinator Jairon Nascimento. The first marketing suspension was in 1998 when a judge blocked Roundup Ready soybeans from Monsanto of St. Louis. It took a further six years to ascertain the commission's competence to make biosafety decisions related to GM crops, after which a flurry of commercial GM crop approvals followed. The court took action after a civil suit brought by several agriculture and consumer advocacy groups, who argued that CTNBio's May 2007 approval of Liberty Link maize relied on an inadequate review and neglected post-release safety monitoring. The judge in the Liberty Link case, Pepita Durski Tramontini Mazini, found that CTNBio failed to ensure adequate post-release monitoring of the crop or the potential effects on regional biomes. "The [post-release monitoring] plan is under analysis in CTNBio, but [the court] has not considered this fact," Nascimento says. *Lucas Laursen*

Table 1 Indian companies marketing biosimilars in India

Company (location)	Biosimilar	Product description
Dr Reddy's Lab (Hyderabad)	Grafeel	Filgrastim (recombinant granulocyte-macrophage colony-stimulating factor, G-CSF)
	Reditux	Biosimilar rituximab (mAb targeting CD20)
	Cresp	Darbepoetin alfa (recombinant erythropoietin)
Intas (Ahmedabad)	Neukine	Filgrastim (recombinant G-CSF)
	Neupeg	PEGylated G-CSF
	Intalfa	Recombinant human interferon alpha-2b
	Epofit	Recombinant erythropoietin
Shantha Biotech/ Merieux Alliance (Hyderabad)	Shanferon	Recombinant interferon alpha-2b
	Shankinase	Recombinant streptokinase
	Shanpoietin	Recombinant erythropoietin
Reliance Life Sciences (Mumbai)	ReliPoietin	Recombinant erythropoietin
	ReliGrast	Recombinant G-CSF
	ReliFeron	Recombinant interferon alpha-2b
	MIRel	Recombinant reteplase (tissue plasminogen activator)
Wockhardt (Mumbai)	Wepox	Recombinant erythropoietin
	Wosulin	Recombinant insulin
Biocon (Bangalore)	Eripro	Recombinant human erythropoietin
	Biomab	Bioximilar nimotuzumab (humanized mAb targeting epidermal growth factor receptor)
	Nufil	Filgrastim, recombinant G-CSF
	Myokinase	Recombinant streptokinase biosimilar
	Insugen	Recombinant human insulin

Desai, CEO of Universal Consulting in Mumbai. Biosimilars are never exact replicas of the originals and the ability of Indian companies to generate biosimilars that satisfy the US Food and Drug Administration and European Medical Agency will be the "acid test" for the Indian players, he says.

Huub Schellekens of the departments of Pharmaceutical Sciences and Innovation Studies at Utrecht University in The Netherlands is similarly blunt. "[US and European] markets will be dominated by big pharma," he believes. "It takes between 50 and 100 million euros [$64 and $129 million, respectively] to develop a biosimilar that meets the regulations in Europe, the US and Japan… that's in addition to post-marketing costs and pharmacovigilance demands," he adds. "I do not see how a small company, especially from India or China, even if they have the technical skills and money to develop a high-quality biosimilar could be able to compete with Teva, Sandoz or Hospira." In this context, Gayatri Saberwal, a scientist at the Institute of Bioinformatics and Applied Biotechnology in Bangalore, says, "the easiest way for Indian firms to get a toehold in Western markets is to become a contract manufacturer of biosimilars for large Western companies." Penetrating the Western biosimilars market may be tough, but it is definitely a battle worth fighting, says Kapil Khandelwal, CEO of Makven Capital, a healthcare advisory services firm in Bangalore. "If you don't do it, somebody else will," he says.

The main attributes needed for a biosimilar to succeed in the global marketplace will be safety, efficacy and, to a lesser extent, pricing. "Simply conducting a small clinical trial would not form the basis for approval of a biosimilar or a biological in Europe or the USA," says Robin Thorpe, head of Biotherapeutics at the UK's National Institute for Biological Standards and Control (NIBSC). Developing nations may not set the bar as high, says Thorpe. He points out that European requirements for biosimilars, which include detailed clinical studies comparing a biosimilar with an approved reference product, are often not adopted in developing nations.

Subpar quality of biosimilar products originating in emerging economies is already causing concern. In one study of a streptokinase biosimilar (*Nat. Biotechnol.* **23**, 413, 2005), NIBSC scientists found that one batch from India contained no detectable streptokinase protein or activity; what's more, two batches from the same manufacturer had only 10% and 20%, respectively, of the labeled potency. Schellekens voices similar worries. "We have tested many products from Asia and South America in our lab in Utrecht but most do not meet our quality standards. All the products that have failed to be approved in Europe came from Asia."

For Haddad, the notion that only a few Western companies have the expertise to successfully make copies of biologics that will meet the standards of the US, European and Japanese regulatory agencies is prejudiced. "Such arguments are political and not scientific," says Haddad. "You must move from chemistry to biology, but the learning curve has been crossed and scientists in the generic biotech industry match the competence of the multinationals." Geena Malhotra, Cipla's research manager, says the technology to characterize the innovator biotech products is so well established today "that working with the right partner, Cipla is confident that it can develop the biosimilars."

The first biosimilars approved have predominantly been simpler recombinant proteins, such as recombinant human growth hormone (**Table 1**). But Cipla is pursuing mAbs, complex, large molecules, copies of which will probably need extensive and extended clinical evaluation, before approval in any of the major markets.

According to Cipla's Malhotra, the company decided to pursue mAbs for immunology and oncology indications because it represents a good fit with their therapeutic and marketing experience, and their partner can provide the technological know-how. Cipla's partner BioMab cites mAbs as 'ideal drugs', given their strong specificity, proven efficacy and limited side effects, as reasons for pursuing generic versions—a fact also recognized by the Chinese government, which has specially established the National Engineering Research Center for Antibody Medicine to promote mAb therapies

Eric Langer, managing partner in biotech and life sciences marketing firm BioPlan of Rockville, Maryland, believes Cipla has identified an opportunity to cut costs of expensive innovator products, which enjoy high-volume sales. "The targets are likely to be for big-market products, produced at larger scales—like monoclonals," says Langer. Cipla may be making a strategic decision here, says Ariyanchira as "the number of competitors is very few in this area."

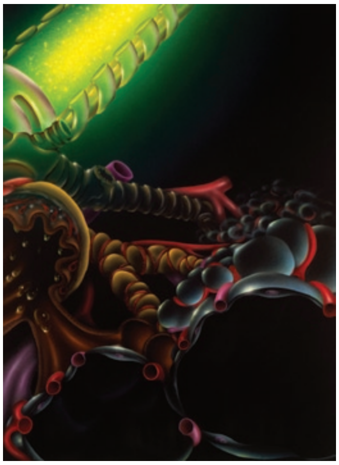
Schellekens cautions, however, that the regulatory demands will be substantial as there are more quality issues with mAbs than there are for recombinant proteins. "And with quality being the Achilles' heel of the Indian/Chinese biosimilars, it will take a lot of time and convincing before we will see doctors here [in Europe] using these [mAbs]."

At least one Indian company has decided to stay clear of biosimilars. "Biogenerics are not going to become a commercial success and I do not know why every company in India is rushing to start the copycat business all over again," says Krishna Ella, CEO of the Hyderabad-based Bharat Biotech. "The big pharma in the West would like us to keep perpetually busy copying their drugs while they innovate and bring out newer and better ones," he says. "The trouble is our companies look at short term for four years or so, and not long term."

Killugudi Jayaraman *Bangalore*

Cochrane meta-analysis on alpha-1 antitrypsin prompts furor

A July 2010 review from the Cochrane Collaboration has questioned the use of the purified versions of the biologic alpha-1 antitrypsin (AAT) to treat lung disease, a conclusion quickly challenged by advocacy groups and manufacturers. AAT is a protease inhibitor (also known as alpha-1 protease inhibitor) that is thought to protect pulmonary tissue against the destructive activity of a wide variety of proteases, in particular elastase. AAT deficiency is an inherited disorder that can cause chronic obstructive pulmonary disease with pulmonary emphysema, as elastase concentrations rise and start to break down elastin needed for lung elasticity. The condition, which can also lead to liver disease, is treated with augmentation therapy using AAT. The product is purified from blood plasma for intravenous delivery to patients and sold by US firm Talecris, in Research Triangle Park, North Carolina, and Kamada, in Ness Ziona, Israel, among others. Market leader Talecris, which is in the process of merging with Barcelona-based Grifols, earned $319 million in 2009 from sales of its AAT line, Prolastin.

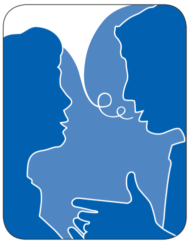

Pulmonary emphysema (diagram) caused by inherited alpha-1 antitripsin (AAT) deficiency is often treated with ATT augmentation therapy, but a recent review blasts this treatment as useless.

The Cochrane report did not go down well. The Cochrane Collaboration is a not-for-profit, independent organization headquartered in Oxford, UK, with contributors from more than 100 countries. Its sole purpose is to produce reviews of data from clinical trials to support evidence-based medicine, and these tend to be closely scrutinized by prescribing physicians in deciding treatment options. In this instance, Peter Gøtzsche, director of the Nordic Cochrane Centre in Copenhagen, concluded that AAT augmentation therapy "cannot be recommended, in view of the lack of evidence of clinical benefit and the cost of treatment," which can run to $150,000 annually for weekly infusions.

The review examined data from the only two randomized clinical trials conducted in AAT, both produced by the same group of investigators and encompassing 140 patients. The authors had planned to include head-to-head trials in which both groups received AAT in different doses or regimens but did not. "Such trials have little interest as long as it has not been shown that augmentation therapy… has any clinical value compared with placebo or no treatment," they wrote. Asger Dirksen of the University of Copenhagen, lead investigator in both trials, originally participated in the development of the Cochrane review protocol but asked that his name be removed from the final paper.

According to the review, mortality data were not reported in either trial nor did the researchers report an average number of lung infections or hospital admissions. The annual number of exacerbations and the quality of life were similar in the treated and untreated groups, the review noted. The report challenged the lack of detail on other outcome measures of lung function, notably

IN their words



"I am a fan of the work… that led to the decoding of the Neanderthal genome. But we don't need any more Neanderthals on the planet, right? We already have enough of them." Genome researcher Craig Venter, whose team created the first bacteria with a synthetic genome, on his lack of plans to produce a 'synthetic human' anytime soon. (*Der Spiegel*, 29 July 2010)

"Above all, what I'm looking for is businesses that are not dependent on patents. This is my fourth patent cliff in my career and I'm looking to avoid a fifth." Sanofi Aventis CEO Chris Viebacher, CEO of Sanofi Aventis, which has been recently linked with a buyout of Genzyme. (*Associated Press,* 30 July 2010)

"It's absolutely historic, and it's remarkable that we achieved this with the symbol of Spain." Veterinarian Julio César Diez of Palencia expounds on Got, Spain's first cloned fighting bull, which instead of facing a matador will spend its time siring other bulls. (*New York Times,* 30 July 2010)

IN brief

Stem cell clinic patrol

A web-based effort to report and investigate bogus stem cell clinics' claims has been launched. The International Society for Stem Cell Research (ISSCR), an independent, nonprofit organization has set up the first global policing site aimed at helping individuals and their doctors separate hypers and fraudsters from legitimate researchers and experiments. Starting on June 1, the portal on the website http://www.closerlookatstemcells.org/ allows people to submit the names of clinics whose cure claims they want the Deerfield, Illinois–based ISSCR to evaluate. The driving force behind the new effort is the rise of clinics located in more than two dozen countries, which promote cures for conditions ranging from multiple sclerosis and arthritis to diabetes and baldness (*Nat. Biotechnol.* **27**, 790–792, 2009). The evaluation will ask the clinics to present scientifically validated evidence for their treatment claims. They will also be asked to describe how their operations are scrutinized by appropriate national regulatory agencies. For the stem cell research community, self-interest is also at work. "The reason we stepped in is because [the websites] are using hype around stem cells to their own advantage, and it is going to invite a backlash against legitimate investigation," says George Daley, director of the stem cell transplantation program at Harvard Medical School and past president of ISSCR. By the first week in August, 280 submissions had been received. *Stephen Strauss*

EC woos SMEs

The European Commission (EC) is inviting biotech firms to apply for research grants, if partnered with academia. For the first time, a quarter of the biotech-specific grants will require the participation of small and medium enterprises (SMEs). The EC plans to hand out €240.3 million ($310.2 million) in direct research grants in 2011, up 26% from the €190 million ($245.3 million) this year. The bio-boost, part of a scheduled ramp-up to €6.4 billion ($8.2 billion) in research funding across all disciplines, is spread across three main areas: agriculture and fisheries, food, health and wellbeing and life sciences and biotech (€70.6 ($90.9) million). The 'cooperation' grants, which require researchers from three or more countries to collaborate, are part of the Seventh Framework Programme (FP7). The number of calls for proposals in industrial biotech, biorefineries and in emerging biotech areas has grown this year, according to the EC, which lists its calls for proposals online. Biotech researchers may also find relevant calls for proposals in neighboring research areas within the cooperation theme or through career grants from the European Research Council, which will provide €661 ($850.7) million across the life sciences. Researchers from academia and industry can learn more on September 13 and 14 at an EC-hosted information day and conference about the "knowledge based bio-economy" (http://www.kbbe2010.be/). *Lucas Laursen*

IN their words

"If you have a speed limit but no one enforcing it, you'll have people speeding. You need to proactively set up a radar system and surveil it." Harvard's George Church, on why he is in favor of federal regulation of synthetic biology to track how the technology is used and by whom. (*The Boston Globe*, 9 July 2010)

"This is the junkiest of junk science." Gilbert Ross from the American Council on Science and Health criticizes a proposal by cardiologists at Imperial College London to hand out a statin pill with each burger to help customers at fast-food restaurants counter the fatty content in their dinners. (*HealthFactsandFears.com*, 16 August 2010)

forced expiratory volume. The authors also point out that disclosure of financial conflicts of interest in the second trial, which was sponsored by Talecris, may be lacking.

The Alpha-1 Foundation immediately issued a forceful response to the Cochrane review's conclusions. In its release, Robert Stockley, director of R&D at Queen Elizabeth Hospital in Birmingham, UK, noted that the review's "conclusion was based on retrospective analysis of published data from only two small pilot placebo-controlled studies that were not powered to evaluate the effectiveness of augmentation therapy. This flies in the face of carefully crafted guidelines from the American Thoracic Society, the European Respiratory Society, the American College of Chest Physicians, and the American Association for Respiratory Care—all prestigious organizations that recommend augmentation therapy for the treatment of patients with lung disease due to Alpha-1." Foundation director Robert Sandhaus pointed out that several large observational studies have shown that augmentation therapy slows the progression of lung disease. The largest of these studies, in over 1,100 individuals with AAT deficiency, shows longer survival for individuals on augmentation therapy. Separately, Albert Farrugia, senior director, global patient access, for the Plasma Protein Therapeutics Association, called the Cochrane review "an arbitrary and dogmatic interpretation of clinical findings."

Gøtzsche's stance should come as no surprise. In response to an article on AAT clinical practice co-authored by Sandhaus in the *New England Journal of Medicine* last year, he questioned the use of quantitative CT [computed tomography] scans in the Dirksen trials instead of forced expiratory volume to show reductions in progression of emphysema, and said that supporting augmentation therapy "goes against the principles of evidence-based medicine."

"Our review is clear," says Gøtzsche, a statistician who in the past has challenged the power of placebo-controlled clinical trials, was one of the early critics of the overuse of mammography and has railed against ghostwriting practices and lack of other disclosures in medical publications.

Mark Ratner *Cambridge, Massachusetts*

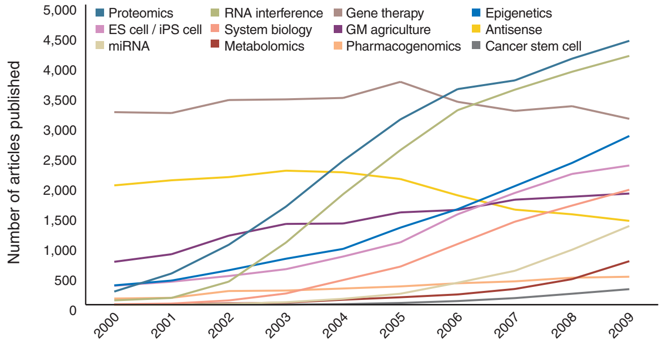
Trends in biotech literature 2009

Wayne Peng

Proteomics, small RNA-related and stem cell research continue their rapid growth in the literature, with epigenetics and systems biology showing recent expansion. The past decade has witnessed a boom in biotech publications from Asian countries, except for Japan, with Chinese authors now publishing more papers in the area than their US peers but accruing fewer citations.

Historic trends in biotech fields

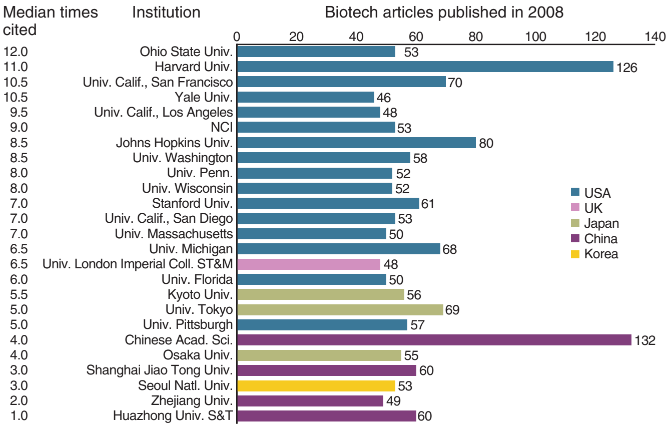
RNA interference, proteomics, microRNA and epigenetics are all expanding quickly.



GM, genetically modified; ES, embryonic stem; iPS, induced pluripotent stem; miRNA, microRNA. Source: National Center for Biotechnology Information, PubMed.
Data obtained by using fields (e.g., "gene therapy") as search term.
ES cell/iPS cell = ("ES cells" OR "iPS cells" OR "induced pluripotent stem cells" OR "embryonic stem cells")
GM agriculture = ("genetically modified" OR "genetically engineered") AND ("food" OR "crop" OR "plant" OR "meat")

Top 25 institutions publishing in biotech

Some Chinese institutions publish a high volume, but papers from US institutions are most cited.



Data obtained by searching 12 predefined 'biotech' fields for articles published in 2008.
Source: ISI-Thomson Reuters, Web of Science

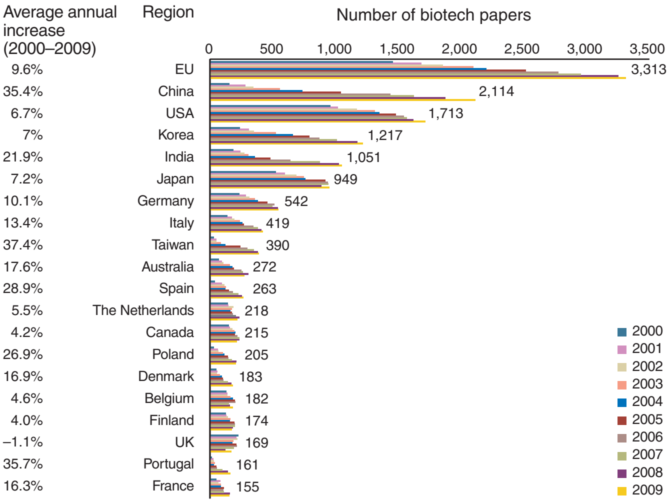
Biotech journal impact

Primary research journal	**2009 impact factor**
Nature Biotechnology	29.495
Cell Stem Cell	23.563
Nature Chemical Biology	16.058
Molecular Systems Biology	12.125
Genome Research	11.342
PNAS	9.432
Molecular and Cellular Proteomics	8.791
Biotechnology Advances	8.250
Review journal	**2009 impact factor**
Nature Reviews Drug Discovery	29.059
Annual Review of Pharmacology	22.468
Pharmacological Reviews	17.000
Annual Review of Biomedical Engineering	11.235
Current Opinion in Biotechnology	7.820
Trends in Biotechnology	6.909

Source: ISI-Thomson Reuters, Journal Citation Report

Number of biotech articles by region

China now publishes more 'biotech' papers than the US.



Source: National Center for Biotechnology Information, PubMed
EU represents the aggregated number of all EU member countries.

Top cited papers by fields

Field	Author	Title	Citation	Number of times cited
iPS cells/ES cells	Takahashi, K. *et al.*	Induction of pluripotent stem cells from adult human fibroblasts by defined factors.	*Cell* **131**, 861–872 (2008)	1,319
Genomic medicine	Zeggini, E. *et al.*	Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes.	*Nat. Genet.* **40**, 638–645 (2008)	376
microRNA	Vasudevan, S., Tong, Y. & Steitz, J.A.	Switching from repression to activation: MicroRNAs can up-regulate translation.	*Science* **318**, 1931–1934 (2008)	362
Next-generation sequencing	Parsons, D.W. *et al.*	An integrated genomic analysis of human glioblastoma Multiforme.	*Science* **321**, 1807–1812 (2008)	350
Kinase	Karaman, M.W. *et al.*	A quantitative analysis of kinase inhibitor selectivity.	*Nat. Biotechnol.* **26**, 127–132 (2008)	266
Nanobiotech	Poland, C.A. *et al.*	Carbon nanotubes introduced into the abdominal cavity of mice show asbestos-like pathogenicity in a pilot study.	*Nat. Nanotechnol.* **3**, 423–428 (2008)	217
Epigenetics	Meissner, A. *et al.*	Genome-scale DNA methylation maps of pluripotent and differentiated cells.	*Nature* **454**, 766–770 (2008)	211
Cancer stem cell	Quintana, E. *et al.*	Efficient tumor formation by single human melanoma cells.	*Nature* **456**, 593–598 (2008)	196
Diagnostics	Nagrath, A.M. *et al.*	Isolation of rare circulating tumor cells in cancer patients by microchip technology.	*Nature* **450**, 1235–1239 (2008)	196
Gene therapy	Maguire, A.M. *et al.*	Safety and efficacy of gene transfer for Leber's congenital amaurosis.	*N. Engl. J. Med.* **358**, 2240–2248 (2008)	187
Imaging	Qian, X. *et al.*	*In vivo* tumor targeting and spectroscopic detection with surface-enhanced Raman nanoparticle tags.	*Nat. Biotechnol.* **26**, 83–90 (2008)	179
Food biotechnology	Besselink, M.G.H. *et al.*	Probiotic prophylaxis in predicted severe acute pancreatitis: a randomized, double-blind, placebo-controlled trial.	*Lancet* **371**, 651–659 (2008)	151
Metabolic engineering	Atsumi, S., Hanai, T. & Liao, J.C.	Nonfermentative pathways for synthesis of branched-chain higher alcohols as biofuels.	*Nature* **451**, 86–89 (2008)	115
Agricultural biotechnology	Ming, R. *et al.*	The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus)	*Nature* **452**, 991–996 (2008)	86
Environmental biotechnology	Frias-Lopez, J. *et al.*	Microbial community gene expression in ocean surface waters.	*Proc. Natl. Acad. Sci. USA* **105**, 3805–3810 (2008)	77
Synthetic biology	Stricker, J. *et al.*	A fast, robust and tunable synthetic gene oscillator.	*Nature* **456**, 516–519 (2008)	64

Source: ISI-Thomson Reuters, Web of Science. Citation data as of 7/13/10.

Wayne Peng is Emerging Technology Analyst, Nature Publishing Group

Oncology's energetic pipeline

Surging interest in cancer bioenergetics has brought drug developers into the fray, but the field awaits a clinical success. Ken Garber explores the extent to which the concept is entering the mainstream.

In a plenary talk at this year's annual meeting of the American Association for Cancer Research in Washington, D.C., Bert Vogelstein predicted that future cancer cures will mostly come from targeting mutant pathways, not mutant genes. Despite successes like BRAF inhibitors in metastatic melanoma[1], only a minority of tumors is susceptible to such mutant gene targeting. As a result, "a lot of research is turning toward pathway-targeted therapies," said Vogelstein, a cancer researcher at Johns Hopkins University in Baltimore. Two such approaches, he said, are DNA damage response modulators and angiogenesis inhibitors. A third is metabolism.

Genes and metabolism are linked, said Vogelstein, whose group has worked on the genetics of colorectal cancer for over two decades. "Many of the [genetic] alterations that are found in tumors are simply selected for because they allow cells to grow in conditions in which certain metabolites are [growth] limiting," he added.

Industry is now convinced that the metabolic features of cancer cells play a vital role in the disease and aren't just downstream consequences of cellular transformation. "There's not a single big pharma company that's not moving into cancer metabolism," said Eyal Gottlieb, a researcher at the Beatson Institute for Cancer Research in Glasgow, UK. Several clinical trials are underway.

But whereas the explosion in drug development activity is welcome, uncertainties abound. "There's still a sense that this is a high-risk area," says Valeria Fantin, associate director of the molecular oncology group at Agios Pharmaceuticals, based in Cambridge, Massachusetts. No drug that directly targets tumor metabolism has yet to succeed in the clinic, and doubts persist about safety. New research is calling into question the assumption that blocking any single metabolic pathway will be effective. And, like the kinase inhibitors that now dominate targeted cancer therapy, metabolic cancer drugs will probably require individual tumor profiling to match patients to the right treatment. Such profiling techniques, like the drugs, are in their infancy.

From genetics to energetics

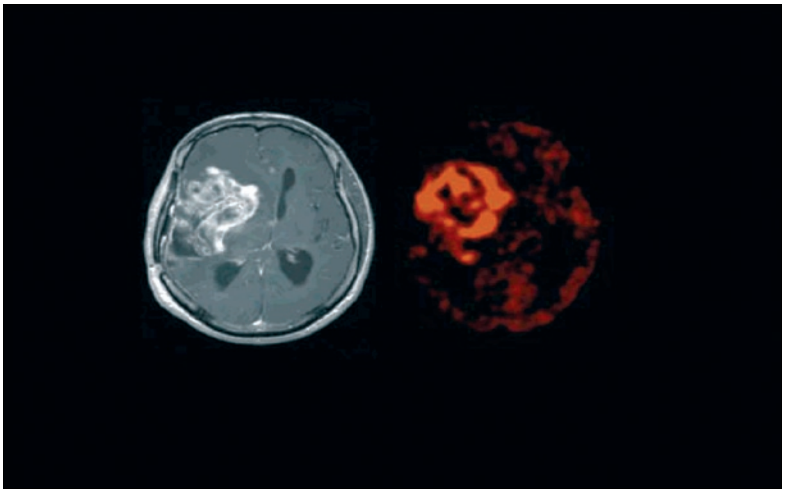
The field of cancer bioenergetics dates back to the 1920s, when German biochemist Otto Warburg compared slices of tumor and normal tissue, finding that tumors consume more glucose and produce more lactate, whereas oxygen consumption remains the same. Eukaryotes generate energy in the form of ATP through a combination of glycolysis (the anaerobic conversion of glucose to lactate) and oxidative phosphorylation (the combustion of glucose or other fuels in the presence of oxygen). Cells typically shift to glycolysis when oxygen is in short supply, enabling sprinters and weight lifters to continue generating ATP. Warburg's fundamental observation, that tumors rely heavily on glycolysis even in the presence of oxygen, has held up for more than 80 years. Warburg argued that this glycolytic shift, later dubbed the "Warburg effect," was the cause of cancer.

Warburg convinced few, and by the time he died in 1970, cancer bioenergetics was in decline, displaced first by the study of tumor viruses and then the discovery of proto-oncogenes and tumor suppressor genes. By the end of the 20th century, cancer was viewed as a disease of mutated genes, not as a consequence of deranged energy metabolism.

But the past decade has seen a resurrection in cancer bioenergetics, for two reasons. One is that the use of positron emission tomography (PET), in which a radioactive analog of glucose is injected into the bloodstream and taken up preferentially by tumors, has greatly expanded. The analog is only partly metabolized and then gets stuck in the glycolytic process, building up to high levels in glycolytic cells (**Fig.1**). Greater reliance on glycolysis by tumor cells, as compared with normal cells, makes PET imaging of tumors possible—visual validation for the Warburg effect. Second, researchers began linking tumor metabolism to the activation of oncogenes, especially *MYC* and *AKT*, and to the loss of tumor suppressor genes, mainly *TP53*, suggesting that genetic changes are driving the Warburg effect. This supplied a mechanism. The discoveries that mutations in metabolic enzymes caused several rare cancer syndromes also helped.

In the early 2000s, a few small companies began developing metabolic modulators for cancer, although much of biotech—and all of pharma—stayed on the sidelines. That's changing fast. "The last year has been quite critical in terms of companies trying to venture into this space," says Fantin. In February, for example, AstraZeneca and Cancer Research UK, both in London, announced a three-year alliance to develop small molecules that target cancer metabolism.

Genetic validation has been crucial. For example, last year Nick Papadopoulos' laboratory at Johns Hopkins showed that KRAS and BRAF mutations in colorectal cancer cell lines led to permanent upregulation of glucose transporters and enhanced glycolysis,



Brain MRI of a glioblastoma tumor (left) with corresponding FDG-glucose PET image from the same patient (right), showing higher glucose uptake within the tumor than outside it. (Reprinted with permission of Michelakis, E.D. *et al., Br. J. Cancer* **99**, 989–994, 2008.)

and that some nonmutant lines developed *de novo* KRAS or BRAF mutations in response to low glucose[2]. This implies that these signature cancer mutations arise to cope with the low glucose environment of tumors. (The researchers were able to block tumor growth with 3-bromopyruvate, a glycolysis inhibitor.)

The field's biggest boost was isocitrate dehydrogenase (IDH). In 2008, Duke and Johns Hopkins researchers, in the course of sequencing 22 human glioblastomas, found five that harbored mutations in the gene encoding *IDH1*. This was a major surprise, as *IDH1*, a metabolic enzyme, had never been implicated in cancer. It turns out that over 70% of low and medium grade gliomas carry mutations in *IDH1* and *IDH2*, whereas a smaller percentage of glioblastomas (12%) and acute leukemias (15%) have mutations. These discoveries set off a frantic effort to understand what a mutated metabolic enzyme was doing in such common and lethal cancers.

Oncometabolism makes its debut

The search has already yielded a surprise. At first blush, it appeared that IDH1 mutations caused a loss of function of the enzyme, and a group from the University of North Carolina (UNC) in Chapel Hill thus concluded that *IDH1* was a tumor suppressor gene. Normally, the IDH enzymes catalyze the conversion of isocitrate to α-ketoglutarate (α-KG), both metabolites being derived from glucose via the Krebs cycle. α-KG is also required for the activity of about 60 dioxygenases. The UNC group inserted mutant *IDH1* into cells and saw α-KG go down and levels of hypoxia-inducible factor (HIF) go up[3]. HIF, a master regulator of cell response to hypoxia, is normally held in check by one of the α-KG-dependent dioxygenases, so it made sense that *IDH1* mutations, by reducing levels of α-KG, would lead to HIF upregulation, often seen in tumors. It was a neat and coherent explanation.

But Agios researchers told a very different story last November. They performed metabolic profiling on IDH1 mutant cells and found high levels of a single metabolite, 2-hydroxyglutatarate (2HG)[4]. They determined that mutant IDH1, instead of performing its normal function (generation of α-KG), acted on existing α-KG, catalyzing its transformation to 2HG. Indeed, they found high levels of 2HG in IDH1 mutant tumors. Agios concluded that *IDH1* was not a tumor suppressor but an oncogene, and 2HG an "oncometabolite"—a small molecule breakdown product that promotes tumors.

The burning question then is how is 2HG promotes tumors. "We're actively working on it, it's just not ready for prime time yet," said Agios CEO David Schenkein. Many possibilities exist, including effects on the tumor microenvironment (2HG is secreted) as well as on mitochondria. In theory, by disabling mitochondria, 2HG could be triggering the Warburg effect. Researchers are looking especially closely at changes in histone demethylases, a member in the family of dioxygenases. If 2HG is affecting that group of enzymes, it might be epigenetically reprogramming cells for malignant transformation or growth.

The IDH enzymes, regardless of mechanism, are obvious drug targets, because tumor mutations are so common. "We've been in full discovery mode for quite a period of time," says Schenkein. Agios is also developing tests to identify IDH mutant tumors, and plans to prospectively select such individuals for its early clinical trials.

However, there is controversy. Yue Xiong, leader of the UNC group, contends that IDH is both an oncogene and a tumor suppressor gene, because he observes a drop in the α-KG pool in mutant cells, suggesting that the mutant enzyme both actively generates 2HG and, through inactivation of normal enzyme activity, depletes α-KG. (Agios has not observed α-KG depletion.) Paradoxically, either would promote tumor growth.

Glycolysis paradox solved?

The field of cancer metabolism is rife with paradoxes. A major one concerns glycolysis and ATP. The Warburg effect begs the question, why would tumors switch from oxidative phosphorylation to glycolysis when oxidative phosphorylation is much more efficient, producing roughly 36 ATPs per glucose molecule compared to only 2 from glycolysis? Many theories have been proposed: adaptation to hypoxia; resistance to apoptosis by means of mitochondrial dysfunction; more rapid energy generation via glycolysis; and acidification of the tumor microenvironment from lactate production, promoting tumor invasion and metastasis.

All of these are plausible, but none have carried the day. Recent studies point in another direction: perhaps tumors don't need ATP so much as they need to build proteins, lipids and nucleic acids for all the new cells being created. The lower ATP production in glycolytic tumors may not be a true paradox, some say. "That's making the assumption that ATP is what tumors care about," says Matthew Vander Heiden, a cancer researcher at the Massachusetts Institute of Technology in Cambridge. "I would argue that that is not the case… . There is no evidence out there whatsoever that tumors are ever limited for ATP production." Rather, he says, the greater need is for carbon skeletons required for DNA, RNA, protein and lipid synthesis. The carbons come off the glycolytic pathway in the form of sugars. Tumor cells enhance this process with the help of PK-M2, a tumor-specific isoform of the glycolytic enzyme pyruvate kinase.

PK-M2, as a result, is now a drug target. "It gives us the ability to go after a metabolic enzyme that's clearly dysregulated in cancers [and] may help explain the Warburg effect," says Agios's Schenkein. PK-M2, when altered by growth factor signaling, causes accumulation of glycolytic intermediates, "basically changing different choke points in the pipeline that will allow things to flow off in other side channels," explains Vander Heiden, co-author on two 2008 *Nature* papers on PK-M2 activity and its regulation[5,6], which built on pioneering work by the late German biochemist Erich Eigenbrodt at the University of Giessen in Germany.

The PK-M2 story itself contains two paradoxes. One is that tumor growth factor signaling makes the enzyme less active, not more. So glycolysis slows down. This goes against the dogma that tumors engage in unrestrained glycolysis. It appears that slowing glycolysis enables the accumulation of partially metabolized glucose intermediates and their diversion to macromolecule synthesis, instead of channeling them through glycolysis. The second paradox is more vexing: a less active pyruvate kinase means less lactate, and tumor cells have very high levels of lactate. "That one I don't think we understand," says Vander Heiden. Lactate may be coming from other sources via other enzymes, but the paradox remains unresolved for now.

There is also disagreement on whether to inhibit or activate PK-M2 pharmacologically. Pharma and biotech, in general, seek to inhibit the enzyme. But Vander Heiden, whose group has identified a number of pyruvate kinase activators, says that activation is probably a better way to go. "If what proliferating cells want to do is turn off pyruvate kinase, I think you want to turn it on," he says.

Addicted to glutamine

In any case, consensus is building that tumors alter their metabolism to build macromolecules, not to generate ATP. This theory hinges on glutamine. Since the 1950s it has been known that tumor cells use large amounts of this amino acid, but it was largely ignored until 2007. That year, researchers in Craig Thompson's laboratory at the University of Pennsylvania in Philadelphia labeled glucose and glutamine with radioactive carbon and traced their metabolic activities in brain cancer cells. To their surprise, almost all the carbon used to feed energy production came from glutamine, not glucose[7]. Equally surprising, glutamine provided most of the nicotinamide adenine dinucleotide phosphate (NADPH) needed to make the fatty acids required for new cell membranes and modifications of signaling proteins. (NADPH is required for certain enzymes to catalyze oxidation-reduction, particularly biosynthetic, reactions.) The Johns Hopkins laboratory of Chi Dang independently reported that glutaminase, an enzyme critical for glutamine metabolism, is induced to high levels by the *MYC* oncogene[8].

These studies suggest that tumors may not be as dependent on glycolysis as once thought. Glutamine, metabolized to glutamate and then to α-KG, can feed into the Krebs cycle. A recent paper showed that in the absence of glucose, cells can shift to a dependence on glutamine and survive[9]. So it may be necessary to block the metabolism of both glucose and glutamine to have a therapeutic impact on many tumors.

Compounds that inhibit glutamine metabolism are in development. Researchers at Johns Hopkins are testing an older inhibitor of glutaminase as well as synthesizing novel compounds. "We're actually moving this forward quite rapidly," said Dang. Initial development of these glutaminase inhibitors and other metabolic modulators (**Table 1**) is being funded with $18 million in grants from Stand Up to Cancer, a research initiative organized by the entertainment industry.

High on the target list is lactate dehydrogenase (LDH), the enzyme that catalyzes the final step of glycolysis. *LDHA*, normally expressed in muscle and liver, is a target gene of MYC and is upregulated in many tumors. Many *LDHA* inhibitors are now in development. Such inhibitors, in theory, should block glycolysis without serious side effects, because full *LDHA* activity should be necessary only in anaerobic conditions. RNA interference and small-molecule experiments in animals have shown potent and relatively selective antitumor effects[10].

"Whatever the winning compound is, I believe you will see a clinical effect—there's just no doubt about it," says Dang. But not all tumors will respond, he adds. "Some could actually revert to oxidative phosphorylation in response to that kind of therapy, and then [they] could escape by some other route after that." Combining LDHA inhibitors with glutaminase inhibitors, thus blocking both glycolysis and oxidative phosphorylation, is one obvious solution. But such complete metabolic shutdown could be dangerous. Only clinical trials will tell.

Deconstructing failure

Any serious effort to target cancer metabolism must rationalize the clinical experience with 2-deoxyglucose (2-DG), the prototypical glycolysis inhibitor. 2-DG works exactly like the radioactive glucose analog in PET scans; it's partly metabolized, then builds up to high levels in the cytoplasm of glycolytic cells. This blocks glycolysis and, in theory, halts tumor growth. Threshold Pharmaceuticals in Redwood City, California, launched a clinical trial in 2004, combining 2-DG with the chemotherapy drug Taxotere (docetaxel).

The trial ended in 2008. The drug combination was well tolerated, according to Threshold, which has not published the results, but only one partial response (out of 34 patients) was seen. As a single agent, 2-DG in prostate cancer showed little or no activity in an independent trial. Threshold has suspended development of the drug.

Did 2-DG fail because it was the wrong drug or the wrong strategy? Many experts blame the drug. Peng Huang, a researcher at the M.D. Anderson Cancer Center in Houston, says that blood glucose levels are so high that 2-DG has no chance to outcompete normal glucose for uptake into cancer cells. Researchers "really cannot [dose] 2-deoxyglucose to a concentration that can even minimally inhibit glycolysis *in vivo*," says Huang. An M.D. Anderson colleague, Waldemar Priebe, disagrees, saying that 2-DG does accumulate in cells because it's not metabolized like glucose. The reason for 2-DG's failure, Priebe says, is its extremely short half-life. "We found out that you cannot detect 2-DG in plasma… after one hour," he says. "It disappeared very rapidly."

But Threshold disagrees. "From our phase 1 study, 2-DG is relatively stable in plasma," writes Threshold spokesperson Denise Powell, citing a mean half-life of over five hours.

Priebe's group has now designed 2-DG prodrugs that last at least six hours in the plasma of animals. They've licensed the

Table 1 Selected drugs targeting tumor metabolism

Sponsor	Compound	Target and indication	Stage
University of Alberta (Edmonton)	DCA	PDK in glioblastoma multiforme	Phase 2
University of California Los Angeles		PDK in breast and lung cancer	Phase 2
University of Florida (Gainesville)			Phase 1
Thallion Pharmaceuticals (Montreal)	TLN-232 peptide	PK-M2	Phase 2[a]
TopoTarget (Copenhagen)	APO866	Nicotinamide phosphoribosyltransferase (Nampt) for cutaneous T-cell lymphoma	Phase 2
Cornerstone Pharmaceuticals (Cranbury, New Jersey)	CPI-613	PDH in pancreatic cancer and others	Phase 1
Gemin X Pharmaceuticals (Montreal)	GMX 1777	Nampt in metastatic melanoma	Phase 1
Myrexis Pharmaceuticals (Salt Lake City, Utah)	MPC-9528	Nampt	Preclinical
Stand Up to Cancer[b]	Aminooxyacetate	Aspartate amino transferase	Preclinical
	FX11	LDHA	Preclinical
	Metformin	AMP-activated protein kinase (AMPK) activator	Preclinical
	Phenformin	AMPK activator	Preclinical
	BPTES	Glutaminase	Preclinical
M.D. Anderson Cancer Center	3-bromopyruvate analogs	Hexokinase	Preclinical
Intertech Bio	2-DG ester prodrugs	Hexokinase	Preclinical
Agios Pharmaceuticals	Undisclosed	PK-M2	Undisclosed
Agios Pharmaceuticals	Undisclosed	IDH	Discovery
ScheBo Biotech (Giessen, Germany)	Undisclosed	PK-M2	Undisclosed

[a]Trial terminated due to legal dispute with licensor. [b]Stand up to Cancer "Cutting off the fuel supply" participating institutions include the University of Pennsylvania, Translational Genomics Research Institute, Johns Hopkins University, Salk Institute and Princeton University.

compounds to the startup Intertech Bio in Houston, which Priebe says is doing toxicology studies in preparation for an eventual investigational new drug submission and a clinical trial in brain cancer.

Dichloroacetate: for the win

The field of cancer bioenergetics still awaits a clinical success. The best immediate hope is dichloroacetate (DCA). DCA is a small molecule widely present in the environment at low concentrations. (It's a by-product of water chlorination). It inhibits pyruvate dehydrogenase kinase (PDK), a key enzyme in glucose metabolism. PDK phosphorylates and deactivates the pyruvate dehydrogenase (PDH) complex, a series of linked enzymes in mitochondria that collectively act as gatekeeper to the Krebs cycle, thus ontrolling the rate of glucose oxidation. DCA, by inhibiting PDK, activates PDH and directs pyruvate into the Krebs cycle, stimulating glucose oxidation and ATP generation. DCA has been used experimentally for decades to treat acquired and congenital mitochondrial diseases, because it shifts the balance of energy production from glycolysis to oxidative phosphorylation, thus improving energy metabolism and reducing lactate acidosis.

The inspiration to try DCA in cancer first occurred to Evangelos Michelakis, a cardiologist at the University of Alberta in Edmonton, about six years ago. Using cancer cell lines, Michelakis showed that DCA caused a multiplicity of anticancer effects, including the opening of potassium channels, reduced mitochondrial membrane potential and increased free radical production, all of which can lead to apoptosis. Meanwhile, Dang's group at Johns Hopkins reported that *PDK* was a HIF target gene, and that inhibiting the enzyme could slow tumor growth. In 2007 the University of Alberta launched a clinical trial of DCA in glioblastoma.

Michelakis published results for five patients in May 2010 (ref. 11). Brain MRI images showed evidence of tumor regression in three patients, and four were still alive 18 months after starting DCA. (One has since died, says Michelakis, and another remains essentially tumor-free.) In short, three did better than expected. But the clinical results, placed in context, say little. Two of the three responders also received the chemotherapy drug Temodar (temozolomide), so it's unclear how much DCA affected the outcome. The trial suspended enrollment early in 2009 after enrolling fewer than half of the planned 50 individuals. "We didn't go all the way, partly because [on dose] we felt we had the answer, but also because we sort of ran out of money," says Michelakis.

So is DCA working in cancer? "It's just too early to tell," says Peter Stacpoole, a University of Florida in Gainesville doctor and pharmacologist, who was the first to characterize DCA in 1969. "That would be a leap of faith. I'd like to believe it, but I have to keep my powder dry on that one."

One problem is that DCA inhibits PDK *in vivo* only at high micromolar concentrations. "The problem that this drug has is [low] potency," Michelakis says. "Combined with something else, you hope for a synergistic effect." DCA could, in theory, complement many other therapies. Despite lack of patent protection, the drug is moving forward in two new US trials.

DCA may have opened the door to more specific, nanomolar-potency PDK inhibitors. Many such compounds already exist. AstraZeneca, Novartis in Basel and other companies developed them in the last decade for diabetes, because they lower blood sugar in the absence of insulin. Peripheral neuropathy put an end to those efforts, but pharma is probably revisiting these compounds for cancer. "I'd be shocked if they weren't looking at this right now," says Stacpoole. (Novartis and Astra Zeneca declined to comment on their PDK inhibitor programs.)

Confronting complexity

Whether or not it's helping patients, DCA demonstrates for the first time that altering metabolism in humans is safe, says Vander Heiden. "Five people took the drug, got changes in how their mitochondria did metabolism in their tumors, and the patients tolerated it well," he notes. Pharma, he says, still is skeptical that potent metabolic modulators will be safe. After all, every cell metabolizes glucose. "The most exciting thing about this [DCA] paper," say Vander Heiden, "is it demonstrates that a therapeutic window is possible."

Efficacy for this drug class remains the biggest unknown. Clinical success may be a long time coming. Reasons for this include metabolic complexity, redundancy and variability. "The whole metabolic field is very different than the normal signaling cascades we are familiar with," says the Beatson Institute's Gottlieb. "We are only now beginning to appreciate the complexity of signaling [cascades]. The metabolic field is far more complex." One facet of that complexity is functional redundancy, developed over the roughly 2 billion years since eukaryotes first appeared. "It's very difficult to perturb the system by taking one component out," says Gottlieb. "It will rewire itself."

Gottlieb and his collaborators are taking a systems biology approach. "You cannot apply your knowledge based on a few papers," he says. "You have to… analyze fluxes, changes of fluxes of thousands of reactions in parallel, in order to try to understand how cells will react to metabolic perturbation."

And tumors differ in their metabolic activity. For example, most rely heavily on glycolysis, but many don't; some consume mostly glucose, others glutamine; some depend more on fatty acid oxidation for energy than others. "The challenge in the field is really to understand the metabolic profiles of tumors before we can really think about what combinations to use," says Dang. To this end, Dang's group is using a combination of assays involving gene expression, proteomics and metabolomics to profile individual tumors. "Measure, say, 250 metabolites in a tumor, and hopefully whatever profile you get is actually reflective and relatively static rather than fluctuating," he says.

Such metabolic profiling will take a lot of work—and time—to develop, validate and translate to the clinical setting. "That's going to be the one big thing that we need to work out over the next several years," says Dang, "while all the laboratories are developing small molecules that target specific weak points." The field of cancer metabolism is roughly where the signal transduction field was a decade or more ago: ripe with targets, starting to churn out potent and specific compounds to inhibit them, but only beginning to grasp the biological complexity that is the essence of the disease.

Ken Garber, Ann Arbor, Michigan

1. Garber, K. *Nat. Biotechnol.* **28**, 763–764 (2010).
2. Yun, J. *et al. Science* **325**, 1555–1559 (2009).
3. Zhao, S. *et al. Science* **324**, 261–265 (2009).
4. Dang, L. *et al. Nature* **462**, 739–744 (2009).
5. Christofk, H. *et al. Nature* **452**, 181–186 (2008).
6. Christofk, H. *et al. Nature* **452**, 230–233 (2008).
7. DeBerardinis, R.J. *et al. Proc. Natl. Acad. Sci. USA* **104**, 19345–19350 (2007).
8. Gao, P. *et al. Nature* **458**, 762–765 (2009).
9. Choo, A. *et al.* Mol. *Cell* **38**, 487–499 (2010).
10. Le, A. *et al. Proc. Natl. Acad. Sci. USA* **107**, 2037–2042 (2010).
11. Michelakis, E.D. *et al. Sci. Transl. Med.* **2**, 31ra34, 1–8 (2010)

Disclosing discoveries

Renee Kaswan

Toiling away at the university, you've just made your once-in-a-lifetime discovery. Here's how to survive what comes next.

Ah, your big scientific breakthrough! This should be your moment of triumph. A patent, recognition, tenure, wealth and scientific advancement for the good of mankind—indeed, all of this is possible. But there's an alternative world, too: aggressive intellectual property (IP) lawyers, lawsuits and the university claiming ownership with sole discretion and total authority over the destiny of your invention.

Commercializing an invention for any independent researcher is a journey fraught with challenges, whether working outside or inside a university system. And as your invention increases in value, the risks and difficulties in IP transfer increase. But if you aspire to be an inventor-entrepreneur, a good university technology transfer office (TTO) can and will support you as you bootstrap your start-up. This article explains how to navigate these waters.

But the first step is to take stock of your interests, skills and limitations. Do you really want to be an inventor-entrepreneur? Would you relinquish tenure and leave the university to develop your discovery? For your particular project, are the resources available to you within the university system more valuable than autonomy? Before answering, remember this: an idea is a far cry from a product or business, and most companies fail, so don't be too quick to give up your day job.

If you *do* choose independence, your next step is to determine how to manage the rights to the IP you have created. Most US schools have compulsory IP assignment policies. Before you or others invest time and money to make your discovery successful, secure a written waiver of assignment from the university. Otherwise, as the value of your IP rises so will the threat of litigation.

Renee Kaswan is founder of IP Advocate, Atlanta, Georgia, USA.
e-mail: rkaswan@ip-advocate.org

Dealing with Bayh-Dole

Perhaps the biggest challenge to researchers is that most universities misinterpret and/or misapply the federal law that governs how they commercialize inventions. The Bayh-Dole Act of 1980 was written by lawyers representing university technology transfer programs with the intent of promoting commercial investment into research and thus enabling the use of federally supported inventions. In this way, the benefits would become available to the public who funded the research. The act permitted universities to obtain titles to federally supported discoveries and serve as stewards of patentable inventions produced by faculty and other research personnel. This is very unlike conducting research at a company (**Box 1**).

Although Bayh-Dole requires that the university act as *coordinator* for inventions made with federal funds by its personnel, it does not require that the university own this IP or act as the sole means of commercialization. But most universities implement the act by compelling faculty and other research inventors—and sometimes students—to disclose their inventions to the institution's TTO and then *require* them to assign patent applications to the university's exclusive ownership. Most schools use this same approach for all inventions—whether federally funded or not. In general, as a researcher at the school you are compelled to comply, although each school's policies and practices differ.

The requirement for faculty to place *all* inventions with a single office on campus (a few universities do have a separate office for biomedical inventions) creates a bureaucratic bottleneck by making all faculty inventors subject to the same, often overworked or underfunded, staff. That same policy effectively squeezes inventions of all sorts, from biotech and nanotechnology to software through the same office.

This leaves you at their mercy. Woe to you if the university's TTO is low on funds for the year and does not want to pay a patent application fee or if it just lost its licensing officer for biotech to a better offer. Your discovery will languish. This is just one of many pitfalls you must watch out for (**Box 2**).

The fact is that technology transfer is a challenging job and one that has long odds no matter who is doing it—university TTO, inventor or otherwise. Still, the TTO can be a valuable partner and resource for you, and many technology transfer managers take a collaborative approach from the start, working diligently to bring innovations to market in partnership with the academic inventor. But the relationship is essentially a 30-year marriage to the institution—and the institution's legal counsel generally regards you as a mere corporate employee. Make certain your equity rights are well documented long before your active participation becomes unnecessary.

Prepare yourself

The most important actions you can take are to educate yourself and remain engaged in the process. Knowledge is key, and you will have the most control over the process *before* you disclose your discovery. So learn the ropes before disclosing or signing anything.

In this vein, get to know your TTO and its people before you need them. Start with an informal meeting with the TTO or invite someone from the office to give a talk to your group. Ask a lot of questions; get a feel for how they work. Assemble the documents you'll need to prepare for legal review and dissect them. You might want to insert addenda or strike out clauses in order to protect your laboratory's interests in the IP rights. You'll want your employment agreement, research contracts, the university's IP and conflict-of-interest policy, the state law on employer claims on inventions and any sponsored research conditions, including Bayh-Dole.

Many IP policies are contracts of adhesion, meaning they provide a unilateral right for the university to make changes without

Box 1 Faculty versus corporate research

Unlike corporate employees, university faculty are “hired to conduct research” not “hired to invent.” Corporations assign research projects specifically to their employees, whereas faculty are encouraged to initiate their own research ideas and innovations.

Corporations fund their employees’ research programs, and therefore the shareholders are the rightful beneficiaries of the intellectual property produced. Taxpayers fund research through federal research grants, so it is appropriate for the government to hold universities accountable for being proactive in managing inventions—if a university chooses to manage inventions—and have them benefit the public.

The American Association of University Professors’ charter describes the public benefits of these fundamental principles of academic freedom for research and free speech. The Association, as well as regulation of academic freedom, began in the 1940s in reaction to the widespread political corruption of the academic mission to seek and disseminate knowledge.

For better or worse, in 1980 when Congress passed the Bayh-Dole Act, universities added the function of technology transfer to their traditional role of cultivating knowledge. By including intellectual property trustee and clearinghouse functions, academia’s mission and responsibilities became more confounded.

requiring consent of the inventors. So if you are satisfied with the present IP policy, get the director of technology to sign an agreement stating that your rights in your disclosed invention cannot be altered without your written, voluntary consent.

Inventors who want to protect their ownership rights in their property can draft (or have a lawyer draft) a memorandum of agreement, or memorandum of understanding, and then make it an addendum to the standard invention disclosure agreement. Without your signature on this transfer of ownership, the university cannot sell or license your invention to anyone else.

The transfer of title for the patent application and issued patent must be registered at the US Patent and Trademark Office. Legal ownership change occurs in this assignment contract, and this is the point at which the property title is transferred. Think of this as transferring the title of your car: you sign the title assignment with the state but you have the bill of sale with the dealer on the price paid. The federal assignment form that transfers an inventor’s constitutional ownership rights to another party is, by federal law, made ‘for due consideration’, and that inventor can determine what that ‘consideration’—payment, in other words—is going to be; it would be foolish to give any employer carte blanche to define the payment.

Your consideration can be: participation in contract negotiations, veto power over license decisions, income for the laboratory, consultant salary for you, revocation of patent assignment for unmet diligence requirements, right to audit the university and licensee, administrative dispute resolution procedures, right to create a start-up company and license your own invention at nominal cost, definition of net income, right to publish, right to open source your invention or right to place your invention with an independent agent, such as GreenCentre (http://www.greencentrecanada.com/) or Science Commons (http://sciencecommons.org/).

You should determine what rights you hope to safeguard for your laboratory, your research, your students and yourself before you ever disclose a discovery to the TTO. Be mindful of the propensity to underestimate the unknown. Because of differing perspectives and experiences, your impression of the relative value of your invention versus the difficulties and expense to commercialize it will be quite different than the TTO’s. Neither of you is sure of what the other knows or doesn’t know, so use finesse, or else egos can collide.

For intercollegiate research, steer the invention to the TTO that best serves your commercialization goals. Some researchers invite a colleague from another university to participate in their research to get access to their office and avoid their resident one. If you are working with a researcher or group from another university, carefully look at both TTOs before deciding which to approach.

Understand your rights

So, the rights to your invention originally reside with you—the US Constitution makes that pretty clear. Technology transfer officers generally claim the university will own the patent rights to all federally funded research based upon the Bayh-Dole Act. Even so, a recent US Court of Appeals for the Federal Circuit decision ruled that the Bayh-Dole Act did not grant universities automatic ownership of federally funded research. In the case of *Board of Trustees of the Leland Stanford Junior University v. Roche Molecular Systems, Inc., et al.,* which involved patents for HIV test kits using PCR, the court rejected Stanford’s argument that one of the co-inventors’ assignment of rights to another entity, Cetus (Emeryville, Calif., no longer in business), was voided by the university’s rights to federally funded inventions under the Bayh-Dole Act. The court’s ruling states: “Bayh-Dole does not automatically void *ab initio* [from the beginning] the inventors’ rights in government-funded inventions.” This case may be appealed, so stay tuned.

Even so, the court’s decision means the university must have a contract with the inventor to get rights—the Bayh-Dole Act doesn’t change that. This does not mean you should automatically fight your university for rights to your invention. Truthfully, you need one another, and although withholding your signature on title assignment is your leverage, their leverage is firing you, harassing you or suing you. Many universities have sued students and faculty who refused to sign patent assignments (for more information, see http://www.ipadvocate.org/forum/dispute.cfm?Type=Disputes). It is not advisable to go to war with your employer, and if you won’t sign a transfer of title and they won’t waive their rights to title, you’ll likely fall into a stalemate situation.

Unfortunately for academic inventors, most courts give the university great latitude and presumption of rights because they are nonprofit public institutions. Faculty and student inventors usually lack the financial resources to outlast the legal gamesmanship.

If you decide you don’t want to work with your TTO, figure out what policy and contractual exceptions might apply to your situation before you make any disclosures. Your funding proposals may already have designated ownership of future IP; a seasoned principal investigator will coordinate contractual ownership from the earliest stages.

Early on, administrative dispute resolution procedures may be successful in getting a waiver of your IP rights. Beware, though, because once there is significant money at stake, many corporate-style board members and administrators will circumvent their institution’s IP policies and send all decisions and interactions to ruthless litigators. At which point, as a researcher without deep pockets behind you, you will be at a serious disadvantage.

Narrow your scope

When you apply for research funding, carefully define the scope of your work. This helps avoid confusion later about whether

Box 2 Eyes wide open

Be aware of these potholes in your path from university discovery to start-up success:

The full life cycle of commercialization can take up to 30 years, during which time technology transfer office (TTO) staff—and university administration—will come and go. You will need to be aware of any changes.

There are generally no performance requirements that would keep the TTO accountable to its inventors, so there is little recourse for faculty or students if the TTO is underperforming (for example, if the TTO has an unreasonable duration to review a disclosure and forward it to patent counsel or release a legal waiver to the inventors).

You are typically promised a share of royalties based on a verbal commitment that the TTO will be diligent in patenting and marketing your invention. This obviously leaves you vulnerable and at the mercy of the TTO.

TTOs often take a 'just in case' approach—they tend to claim control of all inventions that appear to have potential value rather than just the ones they have the resources and ability to commercialize. Conflicts arise whenever the TTO is uncertain or afraid to make an error in deciding to either commit limited financial support or release the invention back to the researchers.

Successful intellectual property draws a lot of attention, which can often be harmful—corporations sometimes pirate, stall or challenge rather than purchase intellectual property rights. Universities and small start-ups are easily outmaneuvered by so-called patent assassins, opportunist litigators (both their own and their opponents') and corporate business schemers.

A university partner with political clout can be invaluable when unexpected obstacles are thrown your way. However, if that political engine litigates against you, success becomes very elusive indeed. Be certain your alliance is legally secured before challenges arise.

or not a particular discovery was made with government funds and helps clarify the resulting questions about ownership.

For federal funding, read the implementing regulations to the Bayh-Dole Act at 37 CFR 401.1 (you can find that online) for information about how the scope of "planned and committed" activities dictate what is a "subject invention"—that is, an invention covered by Bayh-Dole. Drafting objectives toward science, not applications, leaves inventions of applications outside the scope of the federal funding arrangement.

Next, be sure you know your university's IP policy. It's usually incorporated by reference into your employment contract. Some policies specify that you must disclose all inventions, whereas others are more flexible. Some require all employees to agree to assign rights to future inventions to the university, and following the recent *Stanford v. Roche* case, many are incorporating new language: "I hereby assign." If you can't avoid agreeing to assign, work to narrow the scope of obligation or change the burden of proof on making determinations. Typically, those conditions will be qualified with whatever is agreed to in a research contract.

When determining scope of obligation, ask yourself these questions: Is the invention within the planned and committed activities of the research? If it's not, then is it within the academic responsibilities under your employment contract (thus, could it be part of consulting)? If it is not expressly part of your academic duties, then have you made significant use of university facilities when you didn't have to? Is the invention covered in exclusions under state labor law (which may limit employer claims on inventions, even those making use of facilities)? If you anticipate that you will need to circumvent claims of ownership by the resident university, make choices that will bolster your claim on your invention and discourage a lawsuit against you.

After you've decided to make the disclosure to your university, the TTO reviews it to see if it meets the requirements of Bayh-Dole, and then the TTO decides how it wants to proceed. A TTO can always decline to manage a particular innovation. For federally supported inventions, the option to manage patent rights may then go to the agency that sponsored your research, and from there, to the public domain (if no application is filed), to agency licensing programs (if the agency decides to obtain the title) or to the inventor (if the inventor requests the title and demonstrates the capacity to use the invention in the public interest). Currently, university requests to award the title to the inventor are approved expeditiously by most federal granting agencies.

If you want ownership and your research was federally funded, ask for help getting the waiver you need for personal ownership. If it was funded by a foundation, talk to the program officer and enlist his or her support to obtain ownership, perhaps with a conditional deal, such as, "If I get personal ownership, we will work together."

The TTO staff is unlikely to have the exact experience your discovery merits—especially if the invention is transformative new science. The goal here is not to find that elusive 'perfect match' of a TTO that understands your invention exactly—that's virtually impossible. You should be doing this as part of ongoing negotiations with your TTO. You'd be wise to reveal the minimum necessary to accomplish your task at hand, whether working with the university or a corporate sponsor. The balance between paranoia and naivety is precarious. Don't get fooled into giving away valuable information because of a few ego strokes.

The goal is to discover what your commercialization partners know and fill any gaps as needed—on all sides. A biotech researcher may not necessarily know the business world, just as your tech transfer specialist may not fully comprehend the significance of your work. But a good tech transfer manager will know how to assemble the right expertise for a deal ahead of the transaction. You should ask whether you can help identify a team to work on the invention in collaboration with the TTO.

Get it in writing

Write out your understanding of what the TTO has agreed to do as a memorandum of understanding and give it to the TTO. Then request that they make any corrections and return it to you. It's not necessarily legally binding but at least everyone's expectations will be in writing.

Another option is to approach the TTO—on behalf of the team you lead—and ask for help drafting a participation agreement. This assigns a portion of income and governance of research commercialization decisions back to the principal investigator and laboratory team under rules they have adopted up front. After all, well-designed participation agreements build strong laboratory programs by attracting further funding, top students and quality faculty. The participation agreement sets the ground rules for everyone involved. The TTO may support this approach, and it doesn't necessarily put anyone on the defensive, so work gets done faster and closer to the goals you set because the aim is to clarify the next steps for everyone.

If the TTO doesn't seem like a good fit, examine other options. Perhaps you can collaborate with a co-investigator at a university with a more favorable TTO and that office can take the lead. To set this up, consider subcontracting a bit of research to that school or creating a collaboration. Joint inventions can migrate to the other school, but make sure they want to deal with you and also that your colleague there is reliable. Consult your research contract to make sure there's no downside.

An inventor can file for a provisional patent for $150 for one year to buy time to sort out the details. The provisional patent application (**Supplementary Note**) will not be published, so the invention will not be exposed to competitors. If you happen to publish your work, the provisional patent protects the invention's priority for patenting, but only if it clearly teaches the invention so that one with ordinary skill in the art can practice it without undue experimentation—that is, only if the application enables the invention. The provisional patent is hardly a panacea, but it can be used discriminately. Filing a provisional patent may trigger a university's disclosure requirement, and it could constitute conflict of interest if you don't disclose—so be careful.

If the university really wants your signature assigning all rights, ask if they will provide legal representation to you to make sure your interests—and those of any others working with you—are protected. But even when all parties are collaborative and friendly, the fact is, it's *their* policy and *their* terms, written by *their* lawyers—all in regard to *your* IP.

And make sure your grad students have their own legal representation; they shouldn't rely on yours. A good TTO will respect this. But take care in how you present the request—if it's adversarial rather than transactional you'll only invite pain in the form of legal hassles.

Collaboration works best

You are an expert on your invention—it's your creation, after all. Without your enthusiastic cooperation, it is highly unlikely the licensing officer can or will proceed to commercialize your invention. But you're likely not an expert on the legal or business side of things. So be courteous, realistic and reasonable, but still work out terms you can be satisfied with no matter who takes over the reins of the technology licensing office.

Your own dedication to your cause and tenacity in pursing your goals are your best assets throughout commercialization.

Note: Supplementary information is available on the Nature Biotechnology website.

To discuss the contents of this article, join the Bioentrepreneur forum on Nature Network:
http://network.nature.com/groups/bioentrepreneur/forum/topics

The Cartagena Protocol and genetically modified mosquitoes

To the Editor:

The Cartagena Protocol on Biosafety[1] is the fundamental document of the United Nations on the responsible use of genetically modified (GM) organisms. Although the protocol applies to GM mosquitoes intended for disease control, its terms were negotiated primarily with concerns over the safety and trade of GM crops in mind. A sub-working group has been assigned by the Ad Hoc Technical Expert Group (AHTEG) on Risk Assessment and Risk Management to develop risk assessment guidelines for GM mosquitoes. Its first guidance document has recently been published following an April 2010 meeting in Ljubljana, Slovenia[2] and will be submitted to the Parties of the Protocol at their meeting next month. This is an important document outlining the potential risks of GM mosquitoes to biodiversity and human health; however, several overarching issues were considered to be beyond its scope. In this letter, I outline some of these issues and call for a broader discussion on GM mosquitoes to address their unresolved biosafety concerns.

As pointed out in the guidance document, several strategies are being developed to control vector-borne diseases using GM mosquitoes, each requiring its own risk assessment and management considerations. One strategy involves the release of genetically sterile males that, upon mating with wild females, produce unviable offspring, thus resulting in population suppression. The technology for this strategy has already been developed for *Aedes aegypti*[3]—the main vector of dengue fever—and its biosafety implications are relatively manageable because transgenes are only expected to persist in the wild for a few generations after release. Other self-limiting strategies are being developed that eliminate transgenes over subsequent generations.

Another strategy being developed involves the use of a 'gene drive system' to spread disease-refractory genes into mosquito populations, thus rendering entire populations incapable of transmitting diseases[4]. In support of this strategy, a transposable element has been observed to spread through the worldwide population of *Drosophila melanogaster* in a few decades. Progress is being made in the development of genes refractory to malaria and dengue fever, and synthetic gene drive systems are being developed for *A. aegypti* and other mosquito species. If successful, then just a few GM mosquitoes with these constructs would be capable of propagating transgenes over the entire geographical range of a species. Gene drive systems are being developed that are expected to be less capable of spreading between populations; however, this is yet to be shown in an environmental setting.

Perhaps the most important issue inadequately addressed by the guidance document is the ability of mosquitoes engineered with gene drive systems to propagate transgenes across national borders in the absence of an international agreement. Regarding gene flow, the document expresses the need to consider "methods to reduce the persistence of the transgene in the environment" in cases where GM mosquitoes have been shown to have adverse effects. As a form of risk management, it also encourages consideration of methods for "ensuring that they [GM mosquitoes] do not establish themselves beyond the intended receiving environment." However, the acceptability of an open release of GM mosquitoes with gene drive systems that are not shown to have adverse effects is left relatively ambiguous.

The *A. aegypti* mosquito, versions of which have been engineered to have a repressible female-specific flightless/sterile phenotype based on the use of the flight muscle promoter of Actin-4 gene.

A strict interpretation of the Cartagena Protocol, on the other hand, suggests that the requirements for a release of GM mosquitoes with invasive gene drive systems may be almost impossible to satisfy. The Advance Informed Agreement (AIA) procedure applies before the first environmental release of GM organisms in another country and grants the importing country the right to request the exporting country to perform a risk assessment at its own expense, part of which is to determine the likelihood of an "unintentional transboundary movement." If these movements are difficult to prevent, which is certainly the case for GM mosquitoes with invasive gene drive systems, then an environmental release is not allowed.

One way around this problem is a multilateral agreement, consistent with the protocol, which would acknowledge that any release of these mosquitoes is intentionally international and has been agreed to by the affected nations. The problem with a multilateral agreement, however, is its scale and feasibility. GM mosquitoes with invasive gene drive systems have the potential to spread transgenes over entire continents. In the context of Zambia's ban on GM food aid in 2002—during a famine that threatened hundreds of thousands of lives—a unanimous, almost worldwide agreement on GM mosquitoes seems challenging, if not impossible.

Despite this, invasive gene drive systems, such as homing endonuclease genes and

Medea elements, are being developed with the intent of driving genes refractory to malaria and dengue fever into mosquito populations. Gene drive was not an issue that was considered when the terms of the Cartagena Protocol were first negotiated and, as noted in the guidance document, the fact that mosquitoes are a vector of human disease poses "new considerations and challenges during the risk assessment process." Questions arise as to whether the risks of this technology should be weighed against the potential to control disease on a global scale. These issues must be addressed in a clear and open way, making further discussion essential.

A related issue is the exemption of GM mosquitoes in transit or destined for contained use from the AIA procedure. The AIA procedure was written, in part, with the intent of protecting developing countries against threats to biosafety due to a lack of resources to conduct their own risk assessment. Even so, during negotiations of the protocol, countries with strong biotech industries successfully argued that GM organisms in transit or containment pose negligible risks and thus the AIA procedure would restrict trade unnecessarily if applied to them. For GM mosquitoes with invasive gene drive systems, the risks are non-negligible because breaches of containment are impossible to rule out and, once released, just a few escapees could be capable of spreading transgenes on a global scale. The exemption must therefore be re-examined in these cases.

The scenario of containment is particularly relevant to GM mosquitoes because, before an open release, trials are being discussed that would take place in field cages exposed to the ambient environment in a location that the species naturally inhabits. This is an important step in a phased assessment of risks and efficiency; however, before these trials, developing countries are not entitled to request that the importing country pay for a preliminary risk assessment because the AIA procedure does not apply. This issue is not mentioned in the guidance document and was likely considered to be beyond its scope; however, the Cartagena Protocol clearly provides inadequate protection in this scenario, and further discussion is essential before field trials become a reality.

Another pressing issue hinted at in the guidance document is the inapplicability of the Cartagena Protocol to modified mosquitoes that do not fit the definition of "modern biotechnology." The protocol applies to living modified (LM) organisms developed using *in vitro* nucleic acid techniques; however, it does not apply to mosquitoes modified by other means having similar implications for biodiversity and human health.

The most noteworthy variety of non-LM mosquitoes is an *A. aegypti* line infected with the wMelPop strain of *Wolbachia*, an inherited bacterium capable of manipulating its host's reproductive biology in a manner that promotes its spread through a population. As it turns out, this *Wolbachia* infection is associated with several physiological changes beneficial for disease control, including reduced mosquito lifespan, reduced dengue viral load and reduced ability to obtain blood meals with age[5]. However, the existence of physiological changes in conjunction with invasiveness draws into question the wider implications these changes have on biosafety and highlights the fact that biosafety issues are not limited to genetic modification.

To address this issue, the guidance document states that "although the focus of this guidance is on LM mosquitoes, in principle, it may also be useful for the risk assessment of similar non-LM mosquito strategies." This is an important point; however, it is in no way legally binding. Non-LM mosquitoes are beyond the scope of the Cartagena Protocol. However, given that their biosafety implications are as serious as those for LM mosquitoes, further discussion is needed on how they should be regulated. Even-handed regulation will ensure that one strategy is not chosen over another purely for its immunity to onerous requirements.

In conclusion, the guidance document of the sub-working group represents an important first step towards incorporating the biosafety issues posed by GM mosquitoes into the Cartagena Protocol. It raises a number of important considerations regarding risk assessment that may be largely adequate for releases of sterile and self-limiting GM mosquitoes. However, for strategies involving mosquitoes capable of replacing entire populations with disease-incompetent varieties, several issues still need to be resolved. For these strategies, a balance must be sought between the precautionary principle, respect for the sovereignty of states and the ethical mandate to prevent disease on a global scale. Further discussion is needed to address the international regulatory challenges posed by GM mosquitoes in working towards the goal of global vector-borne disease control.

COMPETING FINANCIAL INTERESTS

The author declares no competing financial interests.

John M Marshall

Department of Infectious Disease Epidemiology, Imperial College London, London, UK.
e-mail: john.marshall@imperial.ac.uk

1. http://www.cbd.int/biosafety/protocol.shtml
2. http://www.cbd.int/doc/meetings/bs/bsrarm-02/official/bsrarm-02-05-en.pdf
3. Fu, G., *et al. Proc. Natl. Acad. Sci. USA* **107**, 4550–4554 (2010).
4. Marshall, J.M. & Taylor C.E. *PLoS Med.* **6**, e1000020 (2009).
5. Moreira, L.A. *et al. PLoS NTDs* **3**, e568 (2009).

The *FEBS Letters*/BioCreative II.5 experiment: making biological information accessible

To the Editor:

Current publications lack structured representations of the entities and relationships they report on. As a consequence, information retrieval is hampered and much of the scientific literature is poorly accessible unless it is organized in domain-specific databases by expert curation[1]. However, manual curation is a slow process and databases lag behind, failing to cover much of the published information. The combined effort of the IMEx group deals with only ~20% of the estimated 10,000 protein interaction articles published yearly (**Supplementary Methods**). To explore new publication strategies, the *FEBS Letters* experiment asked authors to supply structured annotations for their publications that were linked to databases with the intervention of professional bio-curators[2]. The BioCreative II.5 challenge then compared these annotations provided by authors and curators to automated systems[3]. Combining these two efforts has generated the first quantitative

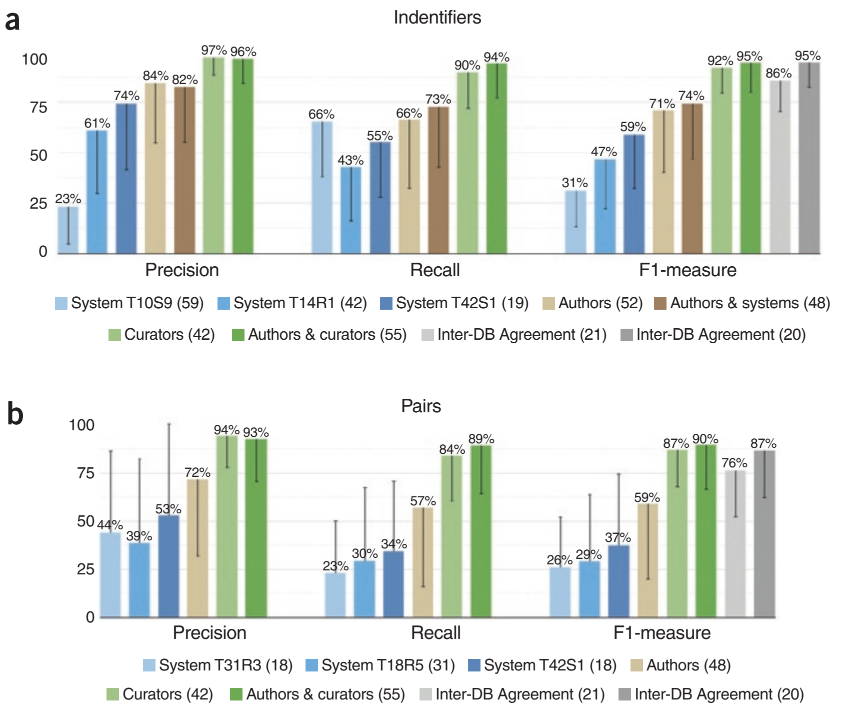


Figure 1 Evaluation results. The performance of annotation sources when (**a**) identifying an article's interacting proteins and (**b**) identifying binary protein interaction pairs (both by their (UniProt) database identifiers). Data obtained from five data sources: automatic extraction methodologies ('System T10 S9', 'System T14 R1' and 'System T42 S1'; where T10, T14 and T42 indicate team runs and S and R indicate online and offline submission number, respectively; curators ('Curators'), authors ('Authors'), authors and curators combined ('Authors & Curators'); or authors and extraction methodologies combined ('Authors & Systems'). Figures in parentheses next to each source indicate the number of articles that source annotated. Results are displayed as average precision, recall and the balanced F measure (F_1-measure) per article (F_1 weighs precision and recall equally; error bars: s.d.) for both protein identifier and interaction pair annotations from all three sources. For reference, inter-annotator agreement (IAA) is shown: 'Inter-DB Agreement' indicates IAA between curators of different databases (light gray bars); 'Intra-DB Agreement' DBs indicates IAA between curators of the same database. Light brown bars show the performance of the authors alone, whereas the dark brown bars are results from an ensemble classifier using both automatic extraction methodologies' and authors' annotations to evaluate the impact of joining both outputs. The test set articles used by the automatic extraction methodologies (Systems; blue bars) do not overlap with the articles annotated by the other sources. The Authors & Systems combination therefore is based on results from the automatic extraction methodologies' training set. The training set, Authors, Curators, Authors & Curators, and Authors & Systems have 33 articles in common. For more details on the distribution of articles, see **Supplementary Methods**.

data to support the debate on ways to supplement publications with structured information.

One way to extend the number of interactions captured by public repositories is to enlist authors in the annotation effort. As suggested by Orchard *et al.*[4], authors could be asked to submit the relevant information during the editorial process, as defined by the minimum information requirement for reporting protein interaction experiments (MIMIx). The pros and cons of possible approaches for adding structured information to scientific publications have been discussed. Gerstein *et al.*[5] and Hahn *et al.*[6] argued over the extent to which quality, consistency and stable support could be expected from authors, and to what extent automatic systems can help in this process. However, the debate is stalling because of the absence of comparable data about these approaches.

In February 2008, the FEBS editorial board designed an experiment asking authors to provide structured information on protein interactions along the lines of the MIMIx recommendations. Trained curators from the Molecular Interaction (MINT) protein-protein interaction (PPI) database[7] then reviewed these annotations and included them in their repository. A structured, human- and machine-readable summary was appended to the traditional abstract as a structured digital abstract (SDA). From March to December 2008, 76 authors were invited to take part in this experiment; 57 accepted, and the published articles contained these SDAs, reporting the identifiers of the interacting proteins, the interaction type (physical interaction, co-localization, enzymatic reaction and so forth) and the experimental method used to verify the interaction.

To complement this study, the BioCreative (http://www.biocreative.org/) organizers challenged text-mining researchers to reproduce a subset of the annotations provided by these SDAs. As data for this challenge ('BioCreative II.5'), Elsevier (Amsterdam) provided 122 PPI-describing articles with their SDAs, and over 1,000 negative articles from *FEBS Letters,* which did not describe PPIs with experimental evidence.

The evaluation presented here was designed to assess the performance of automatic extraction methodologies as compared with the results of curators and authors on (i) identification of the binary interaction pairs and (ii) the corresponding identification of the interacting proteins described in the article (by their (UniProt) database identifiers). The results were restricted to those proteins participating in an interaction supported by experimental evidence. Additionally, BioCreative II.5 assessed the capacity of automatic systems to retrieve articles reporting experimentally validated protein interaction information. This third task is of direct interest to database curators in the process of selecting relevant articles.

Curators also generated annotations based on author data (authors & curators). Finally, we created an ensemble system that produced annotations by combining author and system data (systems & authors). To evaluate the results from all five sources (**Fig. 1**), a 'consensus annotation' was created. Three independent curators consolidated their annotations until a consensus was reached. The performance of each source was evaluated in terms of precision, recall (coverage) and balanced F-measure (the harmonic mean between precision and recall) against this consensus. The results, including the evaluation of an inter-annotator agreement (IAA) study as reference, are shown in **Figure 1**. Detailed background information can be found in the **Supplementary Methods** and

Supplementary Results. From this analysis, we can draw four major conclusions.

First, the majority of authors publishing PPI articles in *FEBS Letters* during the period of the experiment (76 in total) were willing and able to provide structured information: 57 of the authors agreed to participate in the experiment, 56 provided curation-relevant PPI data and 17 out of 22 authors who responded to a questionnaire accompanying the experiment expressed their interest in SDAs.

Second, because of the relatively low accuracy of authors' submissions, the use of authors' annotations did not result in saving of curators' time; however, by combining the annotations of authors with other sources, we are able to show that the data quality increases relative to their individual results (**Fig. 1**), 'Curator' versus 'Authors & Curator', and 'Authors' versus 'Authors & Systems'). This claim is also explained by an investigation of the overlap of annotations between the three sources (**Supplementary Results**).

Third, adding SDAs to the editorial process was not only desirable for the *FEBS Letters* editorial processes, but also technically possible. These SDAs continue to form part of the two FEBS (Federation of European Biochemical Societies) periodicals *FEBS Letters* (published by Elsevier) publications and *FEBS Journal* (published by Blackwell). Furthermore, journals incorporating SDAs are likely to increase the impact and visibility of their articles: unambiguously labeled articles will be more relevant in database search results.

Finally, the (online) systems used to assist in the annotation process required on average 2.3 min (s.d. = 1.4 min) to provide annotations on the test set articles, whereas 22 of the authors reported spending 68 min (s.d. = 72 min) on average. Our measurements of MINT curators over 36 PPI articles show an average of 50 min (s.d. 26 min) per article. According to both authors and curators, retrieving the correct identifiers is the most time-consuming task. Therefore, having automated systems provide a list of relevant identifiers is likely to save authors and curators a significant amount of time. As we show in the **Supplementary Results**, systems are able to return identifiers that both authors and curators miss. And, combining the identifiers reported by authors with results from multiple automated systems ('Systems & Authors' in **Fig. 1**) achieved an *F*-measure of 0.75—better than the authors alone.

In addition, from the BioCreative II.5 article classification results, we established that text mining could help to select PPI-reporting articles for database curators (highest accuracy = 92%). However, given the results of the systems on interaction pairs, it is also clear that complex curation needs cannot be transferred to purely automated approaches. On the other hand, with the relatively good results of systems on identifiers, integrating text mining in the process is likely to generate a significant reduction of time. We need to acknowledge the problem of the relatively small sample of articles used in this first quantitative study, a problem that would be solved by sustained integration of author annotations in the publication and/or curation process, as in the ongoing *FEBS Letters* and *FEBS Journal* experiments. An investigation of integrating annotation systems into the human curation process will form part of the next BioCreative challenge (BC III). In the **Supplementary Discussion**, we propose a simple infrastructure for assembling text mining, authors, and curators in this process based on Leitner and Valencia[8].

Articles labeled with SDAs simplify retrieval and identification of relevant articles for readers. Furthermore, author-provided SDAs would directly benefit areas of research requiring large amounts of high-quality biological data, such as systems biology[9], as more structured data would be produced. In general, SDAs are likely to generate a fundamental impact on the relation between scientists and publications. Our results show that, by combining text mining, authors and curators, it is possible to increase the quality of the annotations and it is plausible that the curation process will be more efficient.

All participants in the BioCreative II.5 project are listed in the **Supplementary List of Participants**.

Note: Supplementary information is available on the Nature Biotechnology website.

AUTHOR CONTRIBUTIONS

F.L. and A.C.-a. contributed equally to this work. G.C. and A.V. contributed equally to this work. F.L. was in charge of the BioCreative challenge and analysis of the results, and the comparisons to the FEBS data; A.C.-a. was in charge of the *FEBS Letters* experiment, data curation, author studies and inter-annotator agreement studies; S.A.M. performed the ensemble experiments; A.C. and L.L. contributed to the inter-annotator agreement studies and data curation; M.K. contributed to the BioCreative II.5 challenge; L.H. was responsible for analysis of alternative metrics to assess the BioCreative results; G.C. was responsible for the contributions from the MINT database and the *FEBS Letters* experiment; A.V. was responsible for the BioCreative challenge.

ACKNOWLEDGMENTS

The authors acknowledge the BioCreative II.5 participants. G.C. thanks P. McCabe, F. Wieland, G. Superti-Furga, F. Goni, I. Mowbray and A. Klinkenberg for their support. L. Salwinski participated in the inter-annotator agreement study. The work on BioCreative was supported by the European Commission FP6 NoEs ENFIN LSHG-CT-2005-518254 and EMBRACE LSG-CT-2004-512092, and by grants BIO2007-66855 from the Spanish Ministerio de Ciencia e Innovación and by the Spanish National Bioinformatics Institute (http://www.inab.org/). The work on the MINT database is supported by FEBS, AIRC and Telethon grants. The participation of L.H. and S.A.M. in BioCreative II.5 was funded by US National Science Foundation grant IIS-0844419.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Florian Leitner[1], Andrew Chatr-aryamontri[2], Scott A Mardis[3], Arnaud Ceol[2], Martin Krallinger[1], Luana Licata[2], Lynette Hirschman[3], Gianni Cesareni[2,4] & Alfonso Valencia[1]

[1]Structural and Computational Biology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain. [2]Department of Biology, University of Rome Tor Vergata, Rome, Italy. [3]The MITRE Corporation, Bedford, Massachusetts, USA. [4]Research Institute 'Fondazione Santa Lucia,' Rome, Italy, and IRCSS Santa Lucia, Rome, Italy.
e-mail: cesareni@uniroma2.it; valencia@cnio.es

1. Seringhaus, M.R. & Gerstein, M.B. *BMC Bioinformatics* **8**, 17 (2007).
2. Ceol, A., Chatr-Aryamontri, A., Licata, L. & Cesareni, G. *FEBS Lett.* **582**, 1171–1177 (2008).
3. Leitner, F. *et al. IEEE/ACM Trans. Comput. Biol. Bioinformatics* **7**, 385–399 (2010).
4. Orchard, S. *et al. Nat. Biotechnol.* **25**, 894–898 (2007).
5. Gerstein, M., Seringhaus, M. & Fields, S. *Nature* **447**, 142 (2007).
6. Hahn, U., Wermter, J., Blasczyk, R. & Horn, P. *Nature* **448**, 130 (2007).
7. Chatr-aryamontri, A. *et al. Nucleic Acids Res.* **35**, D572–D574 (2007).
8. Leitner, F. & Valencia, A. *FEBS Lett.* **582**, 1178–1181 (2008).
9. Aloy, P. & Russell, R.B. *Nat. Rev. Mol. Cell Biol.* **7**, 188–197 (2006).

Biofortified sorghum in Africa: using problem formulation to inform risk assessment

To the Editor:

Most of the genetically modified (GM) crops approved to date (e.g., corn, cotton and soybean improved for insect resistance or herbicide tolerance) do not have compatible wild relatives near their intended area of cultivation, and those that do are not being cultivated in the center of diversity of the species. However, many GM crops being developed to solve agronomic or nutritional problems in developing countries may be grown near centers of origin and diversity of the crop, where these plants were first domesticated and remain major crops[1]. Furthermore, they are often being developed by publicly funded, nonprofit institutions[2]. Such developers, and the regulatory authorities that oversee them, often have relatively limited experience and resources for risk assessment and are faced with some of the first decisions regarding risks associated with gene flow in centers of diversity.

Although the potential for negative effects of gene flow from GM crops in centers of diversity must be considered, some would argue that another kind of risk will be increased if the benefit offered by these products is delayed[3,4]. It is essential, therefore, that data required for risk assessment, including those related to gene flow, are limited to information necessary to allow sound regulatory decisions. Numerous studies related to gene flow from GM crops have been conducted or proposed to address interesting research questions, including evaluations of distance and rates of gene flow, fitness of hybrids, ecosystem dynamics and other parameters[5]. Although some of these studies are useful for decision making, many lack a clear identification of the harm and how the study relates to a causal pathway from the GM crop to that harm. This accumulation of data under the name of 'risk assessment' can lead to considerable confusion about what is necessary for a regulatory decision[6].

The use of appropriate problem formulation to identify data needs has gained attention recently in discussions on risk assessment of GM crops[7–12]. Problem formulation begins with the identification of the protection goals of the law or other instrument that triggered the risk assessment (e.g., protection of biodiversity). A proper problem formulation then derives adverse effects (harm) as operational assessment endpoints (e.g., the abundance of a valued species) based on the protection goals. This is followed by the development of possible scenarios of harm (that is, how there may be adverse change to the assessment endpoints given what is known about the crop plant, the introduced traits and the environment; a risk scenario or conceptual model). Testable hypotheses can then be formulated and an experimental plan to test them can be determined.

The advantage of following the steps of problem formulation is that it focuses data acquisition on clear questions to help decision makers, rather than on attempts to exhaustively characterize all possible outcomes following cultivation of GM crops. It is important to recognize that for risk assessment to be effective, harm must be defined before data acquisition. Definitions of harm are necessarily subjective, and subjectivity in risk assessment cannot be eliminated by doing more scientific research. Thus, extensive collection of data cannot substitute for clear decision-making criteria[6,8,10].

In the following article, we present a case study that shows how these concepts can be applied to risk assessment for GM nutritionally enhanced sorghum intended for cultivation in the center of diversity of the crop and provides a model to help focus the criteria for risk assessments of other GM crops in their centers of diversity. The case is based on a discussion among a panel of individuals (including the authors of this correspondence) with expertise and experience in risk assessment, gene flow, sorghum biology and sorghum as a crop in Africa. This was assembled at the Donald Danforth Plant Science Center in St. Louis in October 2008 by the Program for Biosafety Systems, an organization involved in capacity building for regulation of biotech, to discuss the environmental risks associated with gene flow to wild relatives in the case of African biofortified sorghum (ABS). This panel was not convened to make a determination of the level of risk, but to discuss how it is possible to assess the risk. The steps of problem formulation were used to guide this discussion.

Sorghum is a major crop and staple food in sub-Saharan Africa. ABS is being developed with funding from the public sector in a humanitarian effort to bring better nutrition to the people of Africa (see http://www.grandchallenges.org/ImproveNutrition). Biotech is being used to introduce genes into sorghum for increased lysine and threonine, increased protein digestibility, reduced phytic acid to enhance the availability of iron and zinc, as well as increased levels of the vitamin A precursor beta carotene. The specific genes inserted into ABS and their modes of action were considered during our discussion. The genes are being combined in a single unit that will behave as a locus, to be expressed in the seed endosperm only. These sorghum lines will soon be ready for field trials and for breeding to introduce the genes into suitable local varieties.

The center of origin and diversity for sorghum is in the Ethiopia-Sudan region of Africa[13]. Existing data suggest that gene flow does occur readily between the crop and nearby or sympatric weedy populations, although very rarely to distant, more-or-less truly 'wild' populations[13–15]. According to theory, even neutral genes from cultivated sorghum, which are not expected to have a selective advantage or disadvantage by definition, may persist in the wild populations, even if gene flow should be rare[16,17]. The discussion panel agreed that when GM sorghum is grown in standard conditions for the cultivation of sorghum, transgenes are likely to be transferred to and persist in the wild populations, as with other genes from cultivated sorghum. For the purposes of a risk assessment, in this case, it should not be necessary to carry out any additional studies to test for the likelihood or frequency of gene flow to wild sorghum.

The important question the panel identified for environmental risk assessment of gene flow from ABS in Africa is whether there may be harmful consequences when the transgenes enter the wild populations through gene flow. To answer this question using problem formulation, the first steps are to determine the protection goals and identify assessment endpoints that fit those goals. In many countries, protection goals are defined by law. If no legal definition exists, it may be necessary to define the goals in the risk assessment, perhaps using precedent from similar assessments elsewhere.

Identification of the harm presents one of the greatest challenges for risk assessors. As noted before, 'harm' is subjective and cannot be deduced scientifically; science can help us predict whether there will be consequences of actions, but it cannot determine whether those consequences are acceptable[18]. In this case study, harm is defined as adverse changes to ecological assessment endpoints. We recognize that assessment endpoints could also be cultural, political or economic but did not consider those endpoints in our discussion.

In this case, we considered specific adverse changes to valued entities (that is, harms) and scenarios by which they could result from gene flow from ABS to wild sorghum (**Table 1**). The harms we identified include loss of valuable genetic diversity in the crop, loss in abundance or diversity of valued flora or fauna, and loss of crop yield. More than one scenario could lead to each of the identified harms, and each scenario is based on our knowledge of the biology of sorghum, the introduced traits, the environment where it will be grown and population genetics theory. Some of these scenarios are those typically associated with gene flow, such as loss of diversity due to a selective sweep or genetic swamping. Other scenarios are more specifically related to knowledge about the biology of the crop and the introduced traits. For example, the panel recognized that bird feeding is a serious problem already in sorghum but did not agree about whether this would have an impact in wild relatives of sorghum, or that there was a reason to expect the traits being introduced into ABS would make the seeds more attractive to birds; however, birds are known to prefer seeds with low-tannins[19], and the panel agreed that an unintended reduction in the level of tannins should be considered (**Table 1**; harm 1, scenario 4). During the problem formulation phase of risk assessment, it must be decided which scenarios are plausible, warranting further investigation, and which scenarios are so unlikely that they do not need to be considered[7–12]. We included most of the scenarios that we discussed, although there was some disagreement about which were plausible.

Having clearly outlined the harms and possible ways that ABS might lead to them, we developed a testable hypothesis of 'no harm' for each scenario identified, which can then be corroborated or refuted with existing or new observations. A testable hypothesis for each of these potential scenarios for harm is shown in **Table 1**. Various other hypotheses could have been formulated related to each scenario. It is not necessary to test every possible hypothesis, but ideally the hypothesis to test is one that will give most confidence that the scenario leading to harm is not likely[8]. In the case of ABS, the panel (including the authors of this correspondence) agreed that the scenarios by which the identified harms could occur are only likely if there are unintended changes associated with the transformation. All of these hypotheses of no difference can be tested by conducting a thorough comparison of the GM and non-GM sorghum for the specific characteristics in the hypotheses, to evaluate the likelihood that the identified harms will not occur from ABS.

Such a thorough characterization of a GM crop, which includes characteristics related to agronomic performance, survival and reproduction, disease and insect susceptibility, nutritional composition and known toxicants, is standard practice during GM crop development. Comparative assessment to detect differences between the GM crop and a comparator, usually its non-GM counterpart, forms the foundation of risk assessment for GM crops currently[12,16]. This is generally conducted in the laboratory and in field trials, which may be carried out over multiple seasons and in multiple locations. Field trials are conducted with appropriate measures for confinement of plant material, including the restriction of gene flow. If any potentially harmful unanticipated changes are detected during this characterization, further assessment or risk management options would be considered. It should be noted that certain unanticipated changes such as disease or pest susceptibility could have a significant effect on the comparative yield of the GM crop, in which case the product may not be deployed owing to poor agronomic performance not related to biosafety.

The panel also determined that an additional study to compare characteristics related to survival and reproduction in 'ABS × wild' hybrids and 'non-ABS × wild' hybrids could be conducted to test the hypothesis that transgene interaction with wild genetic backgrounds will not significantly increase the survival and reproduction of hybrids. Each of the harms identified is possible if there is an increase in survival and reproduction due to such an interaction (**Table 1**). Interactions between transgenes and 'wild' genes are not expected to increase hybrid survival

Table 1 A plan to assess the potential environmental risks of gene flow ABS to wild sorghums in Africa

Harm	Risk scenarios	Hypotheses	Experimental plan
Harm 1. Loss of valuable genetic diversity in the crop or compatible species	Scenario 1. Loss of allelic diversity in the wild sorghum due to a 'selective sweep'. A selective sweep following the movement of transgenes into the wild populations would likely leave the populations more genetically uniform in parts of the genome closely linked to the transgenes under strong selection[17,20]. This requires a substantial selective advantage for plants with the ABS transgene.	Hypothesis. ABS traits will not increase survival or reproduction of sorghum.	A thorough comparison of ABS and non-GM sorghum for characteristics related to survival and reproduction, disease and insect susceptibility, nutritional composition, and known toxicants.
	Scenario 2. Loss of allelic diversity due to 'genetic swamping'. 'Genetic swamping', whereby the wild species becomes genetically inextinguishable from the crop plant ('extinction by assimilation') is often cited as a risk from gene flow, but circumstances that would lead to such swamping are likely to be rare[17]. Genetic swamping from crop sorghum to wild sorghum does not occur currently; therefore, harm via this route would require a substantial increase in the hybridization frequency associated with ABS.	Hypothesis. ABS traits will not change the hybridization frequency of sorghum.	
	Scenario 3. Loss of abundance of wild sorghum due to 'outbreeding depression'. In certain circumstances, populations may decline if there is a reduction in the ability of hybrids to survive and reproduce following hybridization[17]. If the ABS transgenes reduce survival and reproduction, populations of wild sorghum could decline following hybridization with ABS.	Hypothesis. ABS traits will not reduce the survival or reproduction of sorghum.	
	Scenario 4. Loss in abundance of wild sorghum due to increased bird preference. Higher levels of tannins in sorghum seeds can make them less palatable to birds[18]. If the level of tannins decreases in ABS compared with those in other cultivated sorghums, birds may preferentially feed on the wild sorghum with the ABS traits over other nonsorghum seed sources. It is difficult to predict how this change in bird behavior could affect the dynamics of wild sorghum populations. It could potentially decrease the abundance of wild sorghums.	Hypothesis. ABS traits will not decrease tannin levels (increase bird preference) in sorghum.	
Harm 2. Loss in abundance or diversity of valued flora (native)	Scenario. Loss of native plants due to competition with wild sorghum. Loss of abundance or diversity of flora is possible if the wild sorghums that carry the transgene become invasive in unmanaged ecosystems (outside of agriculture) and outcompete other native plants.	Hypothesis. ABS traits will not increase the survival and reproduction of sorghum.	
Harm 3. Reduction in abundance or diversity of valued fauna (wildlife or domestic animals)	Scenario 1. Reduction of a critical food source for native fauna due to competition with wild sorghum. The loss of plant species abundance or diversity (flora, as in harm 2) could also have a detrimental impact on the abundance or diversity of native fauna that co-habit with wild sorghum.	Hypothesis. ABS traits will not increase the survival and reproduction of sorghum.	
	Scenario 2. Increased toxicity to native fauna. Native fauna, as well as domestic animals, that feed on wild sorghum could be affected if the GM traits introduced into sorghum should lead indirectly to an increase in the toxicants in sorghum. Endogenous toxins known to occur in sorghum include cyanide, tannins and nitrate.	Hypothesis. ABS traits will not increase the endogenous toxicity of sorghum.	
	Scenario 3. Decreased nutritional value for native fauna. Should the introduced ABS traits affect an unintended change that decreases the value of the existing nutritional composition of sorghum, there may be a detrimental effect on animals that feed regularly on wild sorghum.	Hypothesis**.** ABS traits will not decrease the existing nutritional value of sorghum.	
Harm 4. Significant decrease in yield of crops	Scenario 1. Increased abundance or persistence of wild sorghum in cultivated plantings. If the ABS traits lead to an increase in the weediness that renders wild sorghum, which can already be a problematic weed, more difficult to control in cultivated plantings or more competitive with crop plants, the result could be a loss of crop yields.	Hypothesis. ABS traits will not increase the survival and reproduction of sorghum.	
	Scenario 2. Increased reservoirs for crop pests. Crop yield could also be affected if the ABS traits are associated with an increase of disease or pest infestation in the wild sorghums, as these plants could serve as a reservoir for the pests and contribute to an increase in pest incidence in crop plantings. Changes in amino acid composition using mutation breeding in the past have been associated with decreased seed hardness and increased fungal and insect susceptibility[21].	Hypothesis. ABS traits will not increase disease or insect infestation in sorghum.	
Harms 1–4	Scenario. Increased selective advantage, invasiveness or weediness due to interactions between the transgene and wild genes. If there is an interaction between the transgene and the genes in the wild sorghum which results in an increase in survival and reproduction, the 'harms' that have been identified above might be possible, where an increase in survival and reproduction is part of the risk scenario/hypothesis.	Hypothesis. ABS transgene-interaction with wild genetic backgrounds will not increase the survival and reproduction of the hybrids.	A thorough comparison of characteristics related to survival and reproduction in 'ABS x wild' hybrids and 'non-ABS x wild' hybrids.

Towards patient-based cancer therapeutics

The Cancer Target Discovery and Development Network*

Orienting cancer drug discovery to the patient requires relating the genetic features of tumors to acquired gene and pathway dependencies and identifying small-molecule therapeutics that target them.

Small-molecule drug discovery was originally a compound-based activity. The process begins with the discovery of a biologically active compound, often a naturally occurring small molecule. The next step involves the identification of a disease that may benefit from treatment with the compound, followed by optimization and development of the eventual drug (or drugs through synthetic modifications). Penicillin is an early example of a drug that arose from this approach. Despite many advances in drug discovery in the intervening decades, compound-based drug discovery is still common today. Rapamycin (Rapamune, sirolimus), for instance, was discovered as a secondary metabolite of a *Streptomyces* strain and was explored without success as an antifungal agent before emerging as an effective immunosuppressive agent. Synthetic derivatives of rapamycin have now been approved or are being investigated as therapeutics in cancer (Torisel, temsirolimus; Afinitor, everolimus; ridaforolimus) and in other diseases.

The ability of recombinant DNA to provide nearly unlimited access to human proteins resulted in a second approach that is also common today—target-based drug discovery. Here, therapeutic targets are selected using insights gained most often from biochemistry, cell biology and model organisms. Small molecules are identified that modulate the targets (often by small-molecule screening) followed by optimization and clinical testing. Although this is a robust process, the common failure of candidate drugs in late-stage clinical testing, owing to unforeseen toxicity or lack of efficacy, reveals limits in our ability to select targets using surrogates of human physiology, such as *in vitro* assays and animal models.

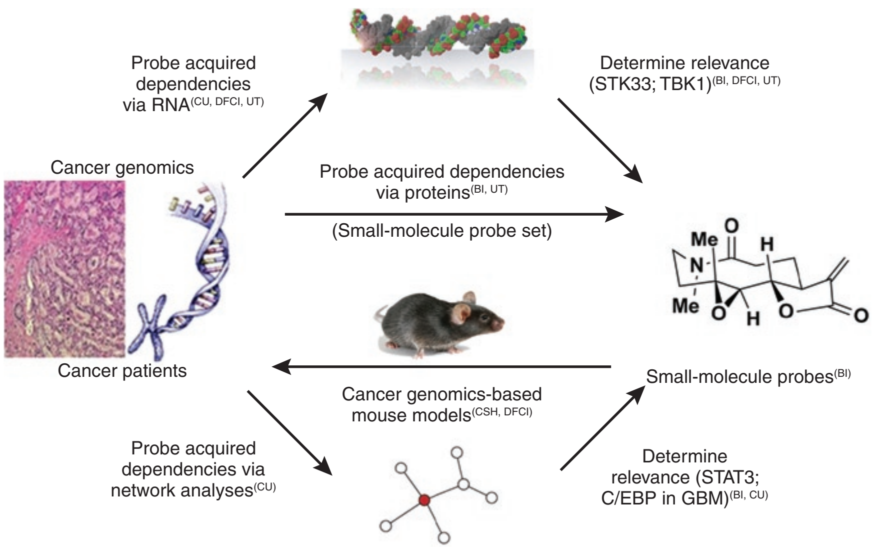

Probe acquired dependencies via RNA$^{\text{(CU, DFCI, UT)}}$
Determine relevance (STK33; TBK1)$^{\text{(BI, DFCI, UT)}}$
Cancer genomics
Probe acquired dependencies via proteins$^{\text{(BI, UT)}}$
(Small-molecule probe set)
Me
O
H
N
Me
O
H
O
O
Cancer patients
Small-molecule probes$^{\text{(BI)}}$
Cancer genomics-based mouse models$^{\text{(CSH, DFCI)}}$
Probe acquired dependencies via network analyses$^{\text{(CU)}}$
Determine relevance (STAT3; C/EBP in GBM)$^{\text{(BI, CU)}}$

Figure 1 The NCI's Cancer Target Discovery and Development (CTD^2) Network aims to relate the genetic features of cancers to acquired cancer dependencies and to identify small molecules that target the dependencies. The centers where the approach is being undertaken are abbreviated in parentheses: BI, Broad Institute of Harvard and MIT; CSH, Cold Spring Harbor Laboratory; CU, Columbia University; DFCI, Dana-Farber Cancer Institute; and UT, University of Texas, Southwestern Medical Center at Dallas.

Advances in human genetics suggest that a third approach—patient-based drug discovery—may offer an alternative with a lower rate of attrition when translated to human trials. Molecular characterization of patient tissues is providing remarkable insights into the root cause of many disorders. As these insights often point to targets and processes that are believed to be especially challenging for small-molecule therapeutics—targets such as transcription factors and regulatory RNAs and processes such as disrupting specific protein-protein interactions—scientists have been innovating in chemistry, cell-culture science and mechanism-of-action studies, among other fields. As a consequence, these hard-to-drug yet key targets and processes are being pursued with new optimism.

Although heritable disorders and infectious diseases are the subject of intensive patient-based drug-discovery efforts, recent insights into the genetics and biology of human cancers have made this family of diseases a prime target for this approach. High-throughput genetic, epigenetic and proteomic analyses of cancer tissues are providing unprecedented molecular insights into genes and pathways causally related to oncogenesis, tumor progression and drug sensitivity and resistance. This points to a path entailing the determination of genomic features of patients' tumors and the discovery and development of new types of therapeutics that target the dependencies (that is, addictions) arising from the specific patterns of genetic or epigenetic alterations within them[1]. This path has been validated in a growing number of extraordinary cases[2,3]. But its generalization is a tall order, one far from the reality of current routine clinical medicine and

**A full list of authors and affiliations appears at the end of the paper.*

not without additional challenges for payers, patients and healthcare providers[2,4].

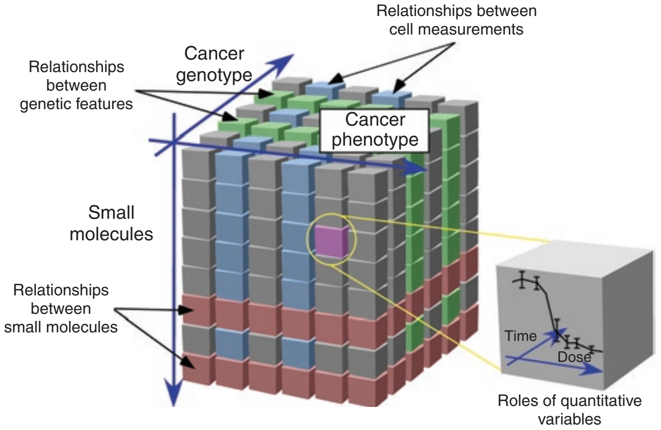

Relationships between cell measurements
Cancer genotype
Relationships between genetic features
Cancer phenotype
Small molecules
Relationships between small molecules
Time
Dose
Roles of quantitative variables

Figure 2 Conceptual image of a matrix of data relating cancer genotype, cancer phenotype and sensitivity to highly specific small-molecule modulators of cancer-relevant proteins. The CTD^2 Network is performing quantitative cellular measurements using small molecules (both with and without a knowledge of their targets) and genetically characterized cancer cell lines (copy number variation, mutational status and gene expression). Computational analyses are being performed that correlate the pattern of sensitivity with the genetic features of the cancer cell lines[9–11]. These analyses yield hypotheses for cancer genotype–drug efficacy relationships that can be tested *in vitro* and *in vivo* using systems developed within the CTD^2 Network.

The National Cancer Institute's approach

To pursue this path comprehensively and prospectively, the US National Cancer Institute (NCI) created the Cancer Target Discovery and Development (CTD^2) Network (http://ocg.cancer.gov/programs/tddn.asp). The Network currently comprises five interacting centers (**Fig. 1**). The mission of the CTD^2 Network is to decode cancer genotypes so as to read out acquired pathway and oncogene addictions of the specific tumor subtypes and to identify small molecules that target these dependencies. The Network builds on the data and insights gained from The Cancer Genome Atlas, Therapeutically Applicable Research to Generate Effective Treatments initiative and other cancer genomic efforts that are systematically cataloging the genetic and epigenetic alterations of specific cancers (e.g., mutational status and changes in gene expression, DNA methylation and chromosomal segment copy numbers). The CTD^2 Network is probing the consequences of these alterations on the dependencies or co-dependencies different cancers have on specific oncogenes or their interacting genes (that is, 'oncogene addiction' and 'nononcogene co-dependencies')[5]. Cataloging these Achilles' heels and linking them to the causal genetic alterations will be critically important for therapies that are personalized to individuals, including combination therapies aimed at targeting many such dependencies at once. It will also be important for anticipating resistance mechanisms and identifying clinical biomarkers.

The CTD^2 Network is currently taking five integrated approaches to determine the targets and processes upon which defined cancer genotypes become dependent. First, techniques that enable the systematic under- or overexpression of selected mRNA transcripts are being used to identify candidate genes. Second, computational network analyses are being performed on cancer genomic data sets to reveal critical master regulatory hubs in the circuitries of cancers, that act as integrators of the complex spectrum of genetic alterations that determine specific tumor subtypes. Third, a small-molecule probe set has been assembled, having members that modulate the activity of defined proteins and pathways that constitute candidate tumor dependencies. These compounds are being tested in many genomically characterized cancer cell lines, and small-molecule sensitivities are thus being correlated to the genetic features of the cancer cells. (In each of these three approaches, the CTD^2 Network measures the fitness of cancer cells having defined genetic features following targeted perturbations.) Fourth, simultaneously, probe-development projects are being undertaken to yield novel small molecules that modulate the functions of cancer therapeutic targets revealed by these approaches. Finally, the consequences of these and other agents that interfere with gene function are being, or will be, tested in, for example, mouse models of cancer having genetic alterations that closely mimic the patient-derived cancers (**Fig. 1**).

Probing acquired dependencies using RNA

The CTD^2 Network is exploiting the extraordinary advances in modulating gene function using RNA interference–based knockdown or RNA overexpression methods. Three examples illustrate the principles behind this approach to identifying acquired somatic genotype–specific dependencies.

The CTD^2 center at the University of Texas Southwestern Medical Center at Dallas is screening genomic small inhibitory (si) RNA libraries against a large panel of non-small cell lung cancer cell (NSCLC) lines derived from human tumors to identify, as a particular NSCLC subtype or clade, siRNAs that are lethal only to cancers that share a similar cancer genotype[6]. Clade-specific lethal siRNAs are being used to identify metabolic vulnerabilities that occur in a particular cancer subtype, vulnerabilities that might be exploited for developing genetically matched anti-cancer therapeutics.

The CTD^2 center at the Dana-Farber Cancer Institute in Boston is screening short-hairpin (sh)RNA libraries to identify different types of cancer vulnerabilities. For example, in a screen of 20 human cancer cell lines, Barbie *et al.*[7] have looked for kinases selectively required for cell survival that depend on oncogenic KRAS and found that, second only to KRAS itself, the noncanonical kinase TBK1 was a synthetic lethal partner.

At Columbia University in New York, the CTD^2 center is using pooled shRNA libraries to complement the computational analysis of master regulators of high-grade glioma subtypes and of glucocorticoid resistance in T-cell acute lymphoblastic leukemia.

Probing acquired dependencies by network analyses

Context-specific regulatory networks of the tumor cell are being assembled and interrogated computationally to reveal otherwise cryptic master regulator proteins whose gain or loss is necessary and sufficient for tumor initiation or progression. These proteins are emerging as master 'integrators' of a spectrum of genetic and epigenetic alterations contributing to the malignant phenotype and thus provide promising novel biomarkers as well as targets for therapeutic intervention (**Fig. 2**).

For instance, at the CTD^2 center at Columbia, C/EBP and STAT3 were recently identified as synergistic master regulators of the mesenchymal subtype of glioblastoma by computational analysis of a regulatory network dissected from a large collection of gene expression profiles of human high-grade gliomas[8]. Validation was achieved by two experimental approaches: shRNA-mediated silencing of these two genes reduced tumor aggressiveness in orthotopic xenografts and co-ectopic expression reprogrammed murine neural stem cells along an aberrant mesenchymal lineage.

Probing acquired dependencies by modulating proteins

The dramatic clinical consequences of linking genetic features of cancers to drug efficacies, including response rates of >80%, are well known, yet these advances today benefit <1% of those suffering from cancer[3]. The CTD^2 centers at the Broad Institute in Cambridge, Massachusetts, and the University of Texas, Southwestern Medical Center, are relating the genetic features of cancers to small-molecule probes or drug efficacies broadly. The CTD^2 Network is extending earlier efforts[9–11] in several ways: first, it is assembling and synthesizing highly specific small molecules (currently a collection of 225 probes and drugs) that target a wide range of proteins and that exploit advances in probe discovery[12,13]; second, it is creating small-molecule screening collections with novel chemical properties; third, it is making quantitative cellular measurements in a wide range of human cancer cell lines treated with the compounds; and fourth, it is identifying the genetic features in these cells that correlate with sensitivities of the small-molecule probes or drugs.

The CTD^2 Network is studying the novel compounds it identifies using cell lines whose genomic features (e.g., copy number, mutation or expression) have been richly characterized and parallel many of the changes found in human cancers[14,15] (although not without exception[16,17]). The intent of this effort is to identify (i) therapeutic targets of cancers linked to specific genetic features associated with cancers[10]; (ii) combinations of targets that, by using guided combination therapy, yield high rates of durable responses; and (iii) potential resistance mechanisms associated with such targets[18]. The resulting data and resources will be publicly available through the project's web site at the end of this year (http://ctd2.nci.nih.gov).

Probe-development projects for novel cancer targets

The CTD^2 Network also aims to accelerate the development of genetically matched cancer drugs by discovering novel small-molecule probes of candidate cancer targets not yet modulated by small molecules. The goal is to identify these gaps and to undertake collaborative probe-development projects involving high-throughput screening, follow-up and medicinal chemistry and biology, and mechanism-of-action studies. Advances in the science of probe discovery, especially in fundamental synthetic chemistry, the culturing and co-culturing of cells using conditions closer to natural physiological environments, and in small-molecule assay development, have enabled the discovery of compounds that modulate challenging cancer-relevant targets and processes[12,13].

CTD^2 investigators are especially interested in projects involving targets such as transcription factors and processes such as gene regulation and cellular differentiation. For example, small-molecule probe-development projects are underway involving both transcription factors (including STAT3, C/EBP (β and δ)[8] and MYC) and chromatin-modifying enzymes (including histone methyltransferases and histone demethylases) that have been identified from genomic studies of cancer.

Probing genetic alterations in mouse models of human cancers

Genomic characterization of human cancers has revealed many genes that are altered. Transgenic or knockout mice that contain germline alterations in the candidate cancer gene can be used to assess oncogenic function. However, their generation and analysis precludes high-throughput evaluation of mutated genes. Transplantable mouse models offer the advantage of speed because genetic lesions are introduced into stem or progenitor cells that are then transplanted into recipient animals. Such models exist for a number of cancer types, including lymphoma, glioblastoma and carcinomas of the liver[19–21]. These models can be used to screen large numbers of genes for oncogenicity and acquired dependencies[22] and to determine the efficacy of small-molecule probes that have been optimized for animal testing.

Conclusions

The CTD^2 Network was formed by the NCI to serve as a link in the overall effort to discover safe and effective patient-based cancer drugs and to facilitate their clinical development through the identification of the genetic features of human cancers that predict drug efficacy, resistance mechanisms and clinical biomarkers. The Network aims to relate these features to their unique dependencies and to identify small molecules that target them, even when this entails hard-to-drug targets and processes—an empirical path that begins and ends with cancer patients.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Weinstein, I.B. & Joe, A.K. *Nat. Clin. Pract. Oncol.* **3**, 448–457 (2006).
2. Aggarwal, S. *Nat. Rev. Drug Discov.* **9**, 427–428 (2010).
3. Thompson, C.B. *Cell* **138**, 1051–1054 (2009).
4. Bach, P.B. *N. Engl. J. Med.* **360**, 626–633 (2009).
5. Luo, J., Solimini, N.L. & Elledge, S.J. *Cell* **136**, 823–837 (2009).
6. Whitehurst, A.W. *et al. Nature* **446**, 815–819 (2007).
7. Barbie, D.A. *et al. Nature* **462**, 108–112 (2009).
8. Carro, M.S. *et al. Nature* **463**, 318–325 (2010).
9. Deininger, M.W., Goldman, J.M., Lydon, N. & Melo, J.V. *Blood* **90**, 3691–3698 (1997).
10. McDermott, U. *et al. Proc. Natl. Acad. Sci. USA* **104**, 19936–19941 (2007).
11. Ramanathan, A., Wang, C. & Schreiber, S.L. *Proc. Natl. Acad. Sci. USA* **102**, 5992–5997 (2005).
12. Frye, S.V. *Nat. Chem. Biol.* **6**, 159–161 (2010).
13. Workman, P. & Collins, I. *Chem. Biol.* **17**, 561–577 (2010).
14. Lin, W.M. *et al. Cancer Res.* **68**, 664–673 (2008).
15. Sos, M.L. *et al. J. Clin. Invest.* **119**, 1727–1740 (2009).
16. Lee, J. *et al. Cancer Cell* **9**, 391–403 (2006).
17. van Staveren, W.C. *et al. Biochim. Biophys. Acta* **1795**, 92–103 (2009).
18. Turke, A.B. *et al. Cancer Cell* **17**, 77–88 (2010).
19. Zender, L. *et al. Cell* **125**, 1253–1267 (2006).
20. Zheng, H. *et al. Nature* **455**, 1129–1133 (2008).
21. Zuber, J. *et al. Genes Dev.* **23**, 877–889 (2009).
22. Zender, L. *et al. Cell* **135**, 852–864 (2008).

The complete list of authors and affiliations is as follows:
Stuart L. Schreiber, Alykhan F. Shamji, Paul A. Clemons, Cindy Hon, Angela N. Koehler, Benito Munoz, Michelle Palmer, Andrew M. Stern & Bridget K. Wagner are at the Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA; Scott Powers, Scott W. Lowe, Xuecui Guo, Alex Krasnitz, Eric T. Sawey, Raffaella Sordella, Lincoln Stein & Lloyd C. Trotman are at Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA; Andrea Califano, Riccardo Dalla-Favera, Adolfo Ferrando, Antonio Iavarone, Laura Pasqualucci, José Silva & Brent R. Stockwell are at Columbia University, New York, New York, USA; William C. Hahn, Lynda Chin, Ronald A. DePinho, Jesse S. Boehm, Shuba Gopal, Alan Huang, David E. Root & Barbara A. Weir are at the Dana-Farber Cancer Institute, Cambridge, Massachusetts, USA; Daniela S. Gerhard & Jean Claude Zenklusen are at the US National Cancer Institute, Bethesda, Maryland, USA; and Michael G. Roth, Michael A. White, John D. Minna, John B. MacMillan & Bruce A. Posner are at the University of Texas, Southwestern Medical Center, Dallas, Texas, USA.
e-mail: stuart_schreiber@harvard.edu

Global health or global wealth?

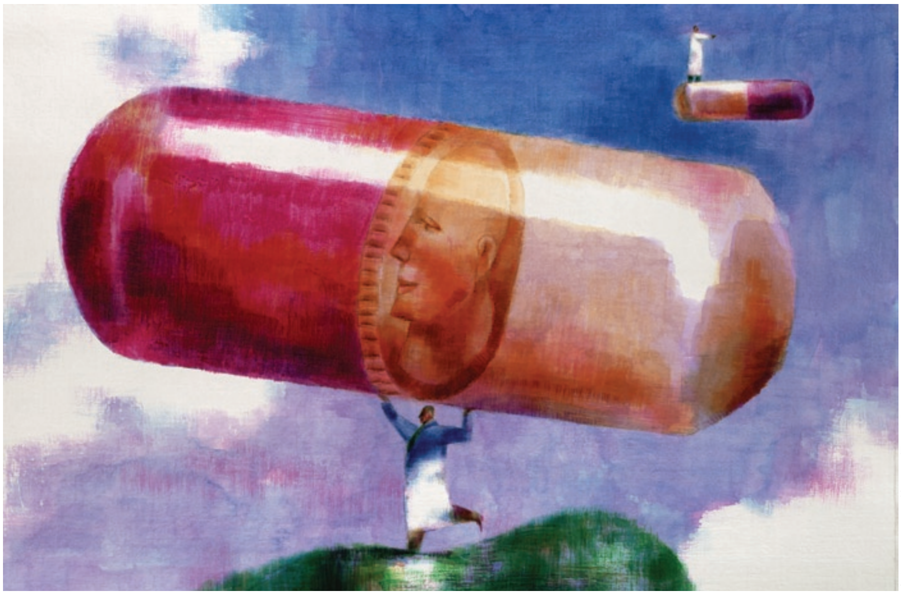
Rahim Rezaie & Peter A Singer

As health biotech enterprises in emerging economies move from imitation to innovation, will they become less relevant to local global health priorities?

Health enterprises in the emerging economies, particularly in China, India and Brazil, have made significant contributions to local and global health needs through low-cost manufacturing of health products. Moreover, a key policy objective for these countries is to foster a strong and innovative health biotech/pharmaceutical sector. The health biotech companies in this sector are innovating close to the 'coalface' of global health problems, making appropriateness, translation, uptake and affordability of the resulting solutions more likely. Even so, the shift from imitation to innovation in these countries—a trend largely stimulated by the adoption of the World Trade Organization's (WTO; Geneva) Trade-Related Aspects of Intellectual Property Systems (TRIPS)—is raising questions about the future market trajectory of the sector.

A basic question is whether this movement toward innovative products would mean movement away from poorer market segments, both at home and abroad. Stated differently, as enterprises in the emerging markets take on more costly innovative projects, would they be compelled to choose between global health and global wealth? Alternatively, is it possible for health entrepreneurs in the emerging economies, as their firms become more sophisticated technologically and financially, to address the needs of the poor while simultaneously taking advantage of more lucrative markets?

Here we argue that the objectives of global health and global wealth can be achieved simultaneously, provided targeted support mechanisms are in place to enable product development for the poorest market segments, for which a purely entrepreneurial model may not be suitable.



© Images.com/Corbis/Todd Davidson

As health biotech ventures in emerging economies take on more costly innovative projects, they must balance their need to recoup investment and the mission of addressing local health priorities

Rahim Rezaie and Peter A. Singer are at the McLaughlin-Rotman Centre for Global Health, University Health Network and University of Toronto, Toronto, Ontario, Canada.
e-mail: peter.singer@mrcglobal.org

Accessing global markets: implications for global health

Through service-provision arrangements with multinational pharmaceutical companies (MNCs), more and more enterprises in the developing world, such as China's Wuxi PharmaTec (Shanghai) and India's Advinus Therapeutics and Jubilant Organosys (both in Bangalore), are integrating themselves into the global product development value-chain. Furthermore, there is a growing trend toward collaborative development of innovative health products. For example, China's Hutchison Medipharma (Shanghai) and Shenzhen Sunway (Shenzhen) work with several global pharmaceutical MNCs to develop health technologies. The concern from a global health perspective is that these trends may, over time, shift the focus of domestic health biotech sectors in emerging economies toward the needs of more lucrative global markets and reflect the priorities of pharmaceutical MNCs.

This argument presumes, however, that the focus of the MNCs is static and will remain targeted on developed world markets. In fact, growth in pharmaceutical markets in emerging economies has begun to outstrip that of developed markets. DataMonitor has reported average annual growth rates of approximately 10% for Brazil and India and 21% for China between 2004 and 2008 with similar growth forecasts until 2013. Therefore, pharmaceutical MNCs and large biotech companies have significantly increased their focus on the markets of emerging economies.

The needs of developed and developing countries in terms of health products overlap in significant ways. Some health conditions, such as infectious diseases and, in particular, neglected tropical diseases, are disproportionately or almost exclusively affecting the developing world. Other health conditions, such as noncommunicable chronic diseases (NCDs), exist in both the developed and developing countries and have become the main source of disability and mortality worldwide. Indeed, in some instances, they are becoming growing epidemics in the emerging economies[1]. Therefore, the contributions of health enterprises based in the emerging economies to NCDs are by themselves highly relevant to global health.

Tapping global markets for products and services strengthens domestic firms by enhancing their technological, inancial and marketing resources. A closer look at the pipeline of innovative products in Indian firms, such as Wockhardt, Piramal Life Sciences (both in Mumbai), Dr. Reddy's Laboratories (Hyderabad) and Biocon (Bangalore) reveals their considerable relevance to domestic populations. As such, efforts that strengthen domestic industries can ultimately advance global health by enhancing the ability of domestic enterprises to address locally relevant diseases in a more innovative manner.

Therefore, the confluence of factors associated with global integration and targeting of global markets by firms in the emerging economies, together with the changing mindset and practice on the part of large pharmaceutical MNCs, hold the potential to temper any broad-based movement away from most local and global health needs by the former group of firms. These trends also enhance innovative capacity of domestic industries.

Entrepreneurial model: a critical vehicle for global health

Traditionally, the primary preoccupation of health entrepreneurs in emerging economies was copying and manufacturing existing products discovered and developed elsewhere. Over the past decade, many firms in these markets have extended this activity by adapting health technologies to developing world contexts and have leveraged process innovations and reduced labor costs to offer quality products at prices that are more affordable. For example, process innovations by India's Biocon (Bangalore) and Shantha Biotechnics (Hyderabad) helped reduce the cost of insulin and hepatitis B vaccine in the Indian market by >40% and >95%, respectively. However, the incremental nature of these innovations means that the firms involved incur substantially less cost than if they were to discover and develop the original products independently; something that will increasingly be required in the post-TRIPS period. Although there will continue to be considerable scope for incremental and business model innovations, these cannot, by themselves, address neglected disease areas for which no effective prevention, diagnosis or treatments are available. The latter challenges demand radical innovations where new solutions are explored with respect to unmet medical needs—a strategy that entails considerable financial risks and demands access to advanced technologies and know-how.

Notwithstanding cost advantages in the emerging economies, a purely entrepreneurial model, left on its own, is unlikely to address disease areas with relatively low monetary market potential. Thus far, health enterprises in developing countries have served to effectively shrink the proportion of populations without access to many health products, but they cannot eliminate the access gap on their own. We propose that limited but selective and well targeted interventions by domestic governments and the global health community can allow firms in emerging economies to expand their target market to include more of the poorest market segments, both at home and abroad.

Supporting the entrepreneurial model

Previous efforts to advance global health have included the formation of a variety of public-private-partnerships (PPPs)[2], advance market commitments (AMCs)[3,4], priority review vouchers[5] and patent pools to share intellectual property[6]. Several key organizations serve to inform and enable these and other initiatives. For example, BioVentures for Global Health (San Francisco, CA) engages biopharmaceutical companies in global health, in part by highlighting market opportunities with a global health impact. The Results for Development Institute's (R4D; Washington, DC) Center for Global Health R&D Policy Assessment (http://www.healthresearchpolicy.org) provides independent reviews of proposed solutions that aim to accelerate global health R&D. These initiatives have been supported from public and philanthropic sources as well as the private sector in pursuit of global health goals. However, these mechanisms have, to date, been primarily (though not exclusively) used by multinationals and biotech companies based in wealthy countries, although they can also be adapted to better meet the needs of emerging economy firms. For example, the Drugs for Neglected Diseases Initiative (DNDi; Geneva) produced a new anti-malarial ASMQ (a combination drug including artesunate and mefloquine) with the Oswaldo Cruz Foundation (Rio de Janeiro, Brazil), and the Program for Appropriate Technology in Health (PATH) developed a meningitis vaccine with the Serum Institute of India (Pune, India).

We believe that tapping into the fast-growing capabilities of health enterprises in the emerging economies is an effective way to complement these current initiatives. At a time when the innovative capacity of emerging-economy firms is growing rapidly, access to technologies relevant to global health is getting easier. Patent pools, such as that initiated by GlaxoSmithKline (GSK; London)[6], demonstrate an increased willingness by major patent holders to share intellectual property rights for the true diseases of poverty. Indeed, GSK has recently announced that it will also publish its research results related to a group of over 13,000 promising compounds against malaria[7]. Another key initiative, which serves to increase access to knowledge is the global access strategy used by the Bill & Melinda Gates Foundation (Seattle, WA) and Grand Challenges Canada (http://www.grandchallenges.ca). Combining the innovative capacities of firms in the emerging economies with increased access to knowledge and technologies required for more radical innovations will accelerate development of health products targeted at diseases of the poor.

Below, we briefly discuss three more mechanisms that could provide support to health biotech companies in emerging economies[8–11] in their pursuit of global health objectives.

Orphan drug–like legislation in emerging economies. In the US, orphan drug legislation[12] provides a host of incentives for products targeted at markets with low purchasing power because of their low prevalence, generally fewer than 200,000 people. This concept has been proposed as a model for domestic governments in emerging economies to apply to diseases of the poor[13], where low market potential stems not from low disease prevalence but from the reduced purchasing power of affected populations.

An approach based on the orphan drug model would focus resources toward diseases of the poor by offering support for upstream R&D activities and reduce investment risks for entrepreneurs through mechanisms such as extended periods of market exclusivity and expedited regulatory approval. Furthermore, to account for specific national contexts, it could include further measures that address issues such as eventual procurement and delivery of related products to target populations.

Global Health Accelerator. The Global Health Accelerator (GHA)[14] aims to tap the innovative capacity of emerging economy companies to advance development of products directed against diseases of the poor. For example, in the field of neglected tropical diseases, a total of 78 companies in the four emerging economies we studied have marketed 69 drugs, diagnostics and vaccines, with 54 more in the pipeline. Although these companies are often successful at addressing local diseases, they do not, however, have the financial or human resources to focus on distant markets.

The GHA platform is envisaged as providing to health entrepreneurs in the emerging economies a suite of services that include international market and regulatory assessment, identification of commercialization partners and distribution channels, and facilitation of access to financing. In addition, it would include a global health prize to recognize excellent examples of Southern innovation against diseases of the poor[10].

Global health funds targeted to emerging economies. The PATH-assisted Program for the Advancement of Commercial Technology–Child and Reproductive Health (PACT-CRH) has transferred a number of health technologies to Indian companies and provided over US$7 million in loans to 11 enterprises. In July, the Wellcome Trust (London, UK) also partnered with India's Department of Biotechnology (New Delhi) to launch the 'R&D for Affordable Healthcare' initiative, a £45 ($71.5) million program to support technology transfer and development by public and private sectors in India (http://www.wellcome.ac.uk/News/Media-office/Press-releases/2010/WTX060350.htm). In another initiative, Charles A. Gardner of the Global Forum for Health Research has proposed the provision of direct grants to innovating SMEs in developing countries, motivated by 28 years of his experience with the US Small Business Innovation Research (SBIR) program (http://healthresearchpolicy.org/sites/healthresearchpolicy.org/files/IDCR%20SBIR.pdf). Elsewhere, David Stevens and colleagues at the R4D Institute have put forward a 'Local Currency Guaranteed Development Bond SME Loan Program' (http://www.resultsfordevelopment.org).

In the United States and other developed countries, venture capital (VC) has had a critical role in the development of biotech industries. There is now also a fund based in New York—the Acumen Fund—that takes a VC-like approach to funding global health challenges. In addition, VC funds that invest in health technologies have sprung up in the emerging countries themselves. These include the Andhra Pradesh Industrial Development Corporation's Ventureast Biotechnology Venture Fund (Hyderabad, India), Bioveda China Fund (Shanghai, China) and Bioventures (Cape Town, South Africa). Although such funds sometimes struggle to balance social benefits while realizing financial return on investments, there is the potential to learn from these experiences to promote global health.

Conclusions

There is not an inevitable trade-off between global health and global wealth. Both goals can be pursued in parallel. But this will require concerted action on the part of the global health community, including governments in emerging economies and international donors, to optimize the potential to global health goals of innovative emerging economy firms. The window of opportunity for action will not remain open for long. The global health ommunity needs to better tailor existing measures (PPPs, AMCs, patent pools and priority review vouchers) to involve and support emerging-economy companies. New entrepreneurial support mechanisms, such as orphan drug-like legislation in emerging economies, the Global Health Accelerator and new funds, could add the growing capabilities of these firms to the repertoire of assets available for global health. Doing so will enable the global health community to seize this window of opportunity and ensure that innovative capacity is tapped not only in the industrialized countries but also in the emerging economies, so that the health needs of the poor can be more fully addressed.

ACKNOWLEDGMENTS

We gratefully acknowledge editorial comments from J. Clark. Also, we are thankful for S. Al-Bader's review of earlier drafts of this paper. C. Gardner provided helpful examples. The McLaughlin-Rotman Centre for Global Health is primarily funded by Genome Canada through the Ontario Genomics Institute, the Grand Challenges in Global Health Initiative (Bill & Melinda Gates Foundation), the Water Efficient Maize for Africa Project (Bill and Melinda Gates Foundation), the Ontario Research Fund and the Canadian Institutes of Health Research (CIHR). Some of our projects also receive funding from the Rockefeller Foundation and BioVentures for Global Health, and through in-kind contributions from Burrill & Company and Wulff Capital. Other matching partners are listed at http://www.mrcglobal.org/. Rahim Rezaie is supported by a doctoral scholarship from the CIHR.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturebiotechnology/.

1. Daar, A.S. *et al. Nature* **450**, 494–496 (2007).
2. Gustavsen, K. & Hanson, C. *Health Aff.* **28**, 1745–1749 (2009).
3. Towse, A. & Kettler, H. *Bull. World Health Organ.* **83**, 301–307 (2005).
4. Grabowski, H. *Health Aff.* **24**, 697–700 (2005).
5. Ridley, D.B., Grabowski, H.G. & Moe, J.L. *Health Aff.* **25**, 313–324 (2006).
6. Dentzer, S. *Health Aff.* **28**, w411–w416 (2009).
7. Lister, S. GlaxoSmithKline to share malaria research in hope of finding a cure. *Times Online* (2010). <http://www.timesonline.co.uk/tol/news/science/medicine/article6994620.ece>
8. Frew, S. E. *et al. Nat. Biotechnol.* **26**, 37–53 (2008).
9. Frew, S.E. *et al. Nat. Biotechnol.* **25**, 403–417 (2007).
10. Rezaie, R., Frew, S.E. & Sammut, S.M. *Nat. Biotechnol.* **26**, 627–644 (2008).
11. Al-Bader, S., Frew, S.E. & Essajee, I. *Nat. Biotechnol.* **27**, 427–445 (2009).
12. <http://www.fda.gov/RegulatoryInformation/Legislation/FederalFoodDrugandCosmeticActFDCAct/SignificantAmendmentstotheFDCAct/OrphanDrugAct/default.htm> (2009; accessed 10 January 2010).
13. Frew, S.E., Kettler, H.E. & Singer, P.A. *Health Aff.* **27**, 1029–1041 (2008).
14. Frew, S., Liu, V. & Singer, P. *Health Aff.* **28**, 1760–1773 (2009).

First-in-human clinical trials with vaccines—what regulators want

Karen B Goetz, Michael Pfleiderer & Christian K Schneider

Several factors should be taken into account when it comes to the first exposure of humans to a novel vaccine.

Vaccines have a long history of excellent safety and a highly positive benefit/risk profile. Even so, the lack of specific guidance from regulatory agencies specifically relating to the first application of a new experimental vaccine in humans has hampered product development. Most of the regulatory guidance documents for manufacturers are too broad and sometimes only vague where vaccines are concerned. As regulators deeply involved both in the development of the European Medicines Agency's (EMA; London) new regulatory framework on risk identification and mitigation, and in assessment and authorization of clinical trial applications for biotechnological and biological products (especially vaccines), we have been repeatedly approached by companies and vaccine developers regarding regulatory issues for first-in-human clinical trials. Here, we discuss these considerations as they relate to vaccines within the context of the current EMA guideline for risk identification and mitigation for first-in-human clinical trials based on the apparently considerable uncertainty among developers. We describe how regulators apply the guideline and where we see the limitations or the need to take alternative approaches. The discussion primarily focuses on prophylactic and therapeutic vaccines against infectious diseases as this classic field of products is associated with particular uncertainty.

© Bettmann/CORBIS

In the so-called Cutter incident in 1955, Cutter Laboratories of Berkeley, California, failed to fully inactivate a batch of polio vaccine (vials shown). This is one of the rare examples where documented adverse events were associated with the use of a vaccine in humans.

General considerations

The EMA's Committee for Medicinal Products for Human Use (CHMP) has assembled a Guideline on Strategies to Identify and Mitigate Risks for First-in-Human Clinical Trials with Investigational Medicinal Products as a joint effort of European regulators and scientists from various disciplines[1]. This guideline is applicable to any new molecular entity, both chemical and biotechnological and/or biological. Its main principle, which is now also widely applied by regulators assessing clinical trial applications in Europe, is an approach of risk identification and risk mitigation. This is done by assessing the mode of action, the nature of the target and the relevance of the animal species used for testing of nonclinical safety and toxicity. These issues are particularly pertinent to the design of first-in-human clinical trials of products that have a seemingly potentially higher risk in the first administration to humans than the nth iteration/reformulation of an established product. The most important consideration is to commence testing with a conservative calculation of a safe starting dose and sequential inclusion of subjects in the trial to limit exposure.

Unfortunately, little if any specific guidance is available for first-in-human trials specific to vaccines. The guidance for industry issued by the US Food and Drug Administration (FDA) concerning dose calculation for a first-in-human clinical study[2] describes in detail the initial dose finding but states explicitly that it is not pertinent for vaccines. Only general guidance concerning the principles for conduct of clinical studies is available from the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). EMA guidance specific to nonclinical and clinical evaluation of vaccines is available but also includes only limited guidance specific to first-in-human studies[3–5].

Because vaccines resemble pathogen antigens, which usually have antigenic features distinct from physiological structures found in human tissues, the risk accompanied with administration of these products is usually considered relatively low; frequently reported adverse events in clinical trials are in most cases manageable and transient

Karen B. Goetz, Michael Pfleiderer and Christian K. Schneider are at the Paul-Ehrlich-Institut, Federal Institute for Vaccines and Biomedicines, Langen, Germany; Michael Pfleiderer is the chairman of the European Medicines Agency (EMA) Committee for Medicinal Products for Human Use (CHMP) Vaccines Working Party (VWP) and the German member of the EMA CHMP Biologics Working Party (BWP); and Christian K. Schneider is also at the Twincore Centre for Experimental and Clinical Infection Research, Hannover, Germany and is a member of the EMA Committee for Medicinal Products for Human Use (CHMP), and the chairman of the EMA Committee for Advanced Therapies (CAT).
e-mail: goeka@pei.de

(e.g., fever and local reactogenicity). In addition, knowledge of immunological processes and the role of specific cells and mediators in this context continues to advance, facilitating our understanding of the mechanism of action of individual components of vaccines.

The overall safety of vaccines is corroborated by the fact that during decades of vaccine development and application, cases of severe damage caused by these products have been uncommon. However, rare examples of adverse events have been observed. In 1955, for example, insufficiently inactivated batches of polio vaccine caused an outbreak of polio due to the presence of wild-type poliovirus strains. This became known world wide as the Cutter incident, in which 40,000 children developed mild polio, 200 were permanently paralyzed and 10 died[6]. Another example is an aggravated or atypical disease following vaccination and exposure to wild-type viruses caused by a measles and respiratory syncitial virus[7] or experimental severe acute respiratory syndrome (SARS)[8] vaccine. Regarding live-attenuated vaccines, data suggesting elevated mortality observed in developing countries following vaccination with medium- and high-titer measles vaccines demonstrate again the need for cautious approaches when entering into early clinical trial phase[9]. But these examples also highlight that root causes for problems can often be identified, and principles for risk identification and mitigation can be developed.

Infections themselves can trigger immunological sequelae that can even be more harmful than the actual infection itself (e.g., rheumatic fever after infection with group A streptococci, such as strep throat or scarlet fever, or Guillain-Barré syndrome (GBS) following viral infections or infection with *Campylobacter jejuni* or certain other bacteria). Guillain-Barré syndrome is a rapidly progressing ascending paralysis, mediated by a cross-reactive attack of antibodies (e.g., cross-reacting against the GM1 ganglioside)[10]. Such knowledge is relevant to the risk assessment of a novel vaccine against an infection for the following reasons: first, vaccines often present an antigen in an artificial context (that is, as repetitive structures, such as in virus-like particles, as fragments of epitopes or as capsules); second, in many cases, vaccines are administered together with an adjuvant that enhances or modulates the immune response (see below); and third, vaccines provide an antigen dose that is both different from that seen in a natural infection and most times presented to the immune system by a different route. It may, thus, be (theoretically) possible that a vaccine could lead to clinical symptoms similar to infections that trigger downstream immunological sequelae and thus study protocols could benefit from implementing respective endpoints. This is borne out by the observation that especially in live vaccines, but also in some others, there have been sporadic reports of rheumatic fever and Guillain-Barré syndrome. At the same time, such complications are very rare and the causality is not always clear. For instance, sometimes concomitant minimal (respiratory) infections are present in a subject when an experimental vaccine is tested, but these are, of course, no reason to delay a vaccination. Thus, for the time being, there are doubts of a causal relation between vaccination and the onset of autoimmune diseases, apart from isolated cases.

Another aspect that must be taken into account is that vaccines are biological products. As such, even small changes to the established manufacturing processes may significantly alter product safety and/or efficacy. For example, simultaneous elimination of thiomersal and human serum albumin from a European tick–borne encephalitis vaccine drastically increased cases of moderate and severe fever after the first dose of primary immunization, which could only be corrected by reintroducing human serum albumin into the vaccine formulation[11]. These events demonstrate that the manufacturing process is an integral part of the concept and that changes in the formulation of a given vaccine may benefit from risk identification and mitigation considerations.

Finally, both novel adjuvants[12] that enhance the immune response and novel routes for antigen delivery (e.g., antigen delivery based on gene transfer) will affect the perception of risks and require specific regulatory strategies. With novel adjuvant or emerging new vaccine formats, including vaccines against pathogens for which no vaccine exists so far, safety considerations have to be put on a broader scale as has previously been done for rather straightforward cases like insufficient inactivation of a live virus.

On the other hand, vaccines have an excellent safety record and most new vaccines can a priori be considered low-risk medicinal products. It needs to be emphasized that we do not have to assume that a vaccine with a new mechanism of action or a novel structure is a high-risk vaccine._Likewise, not every new medicinal product should automatically be considered a high-risk medicinal product[13]. The first-in-human trial is a critical turning point between preclinical studies and first human exposure and subsequent larger clinical trials in hundreds or (for many vaccines) thousands of subjects. For sponsors, relevant risk assessment for first-in-human clinical studies means careful design and conduct of studies that reduce potential risk to humans. In comparison to therapeutic proteins or other medicinal products, however, the prophylactic character and mechanism of action of vaccines warrant particular attention. Indeed, some of the concepts introduced in the aforementioned EMA guideline[1] may not even be readily applicable.

First, pharmacokinetics should be considered relevant only if, for example, a vaccination approach involves either a novel or different means of delivery (the first pass effect for oral application versus the usual intramuscular route) or a novel live vaccine (where shedding rates can differ). Pharmacodynamics in vaccines is usually gauged by immunogenicity (appearance and increase of antibody titers).

Second, vaccines often include an adjuvant or are administered concomitantly with an immunomodulator that has its own impact on the overall risk assessment. As such agents can influence the behavior of a vaccine or the host's responses to a vaccine[14,15], it is often important to assess their effects (including, for instance, pharmacokinetics and pharmacodynamics) both separately from the vaccine antigen (as its own entity) and in combination with the antigen.

Third, the target population for vaccine trials is usually healthy and young—often infants from six weeks of age and up, children or adolescents. This requires special diligence concerning benefit/risk assessment.

Fourth, unlike other medicinal products, efficacy measurements are often indirect; thus, the elicited immune response is the active principle of a vaccine and needs to be part of the risk assessment.

And finally, the risk profile of a vaccine may be different over time and dependent on exposure to both vaccine and pathogen infection. For vaccines, acute risks have to be distinguished from sub-acute or chronic (long-term) risks after (repeated) product administration.

Although the general criteria and considerations mentioned in the EMA guideline[1] should always be taken into account, **Figure 1** displays criteria that are more specific to vaccines and may be helpful for developers. The relative importance of these criteria should be decided on a case-by-case basis for each product; if developers are in doubt, they should consult with regulators (either the national agencies of EU member states or the EMA) when designing a trial.

The safe starting dose: is the MABEL relevant?

The calculation of a safe starting dose is a central aspect for a first-in-human trial. The classic approach to calculate the starting dose for a first-in-human trial for a classic medicinal product (not a vaccine) is based on toxicity in

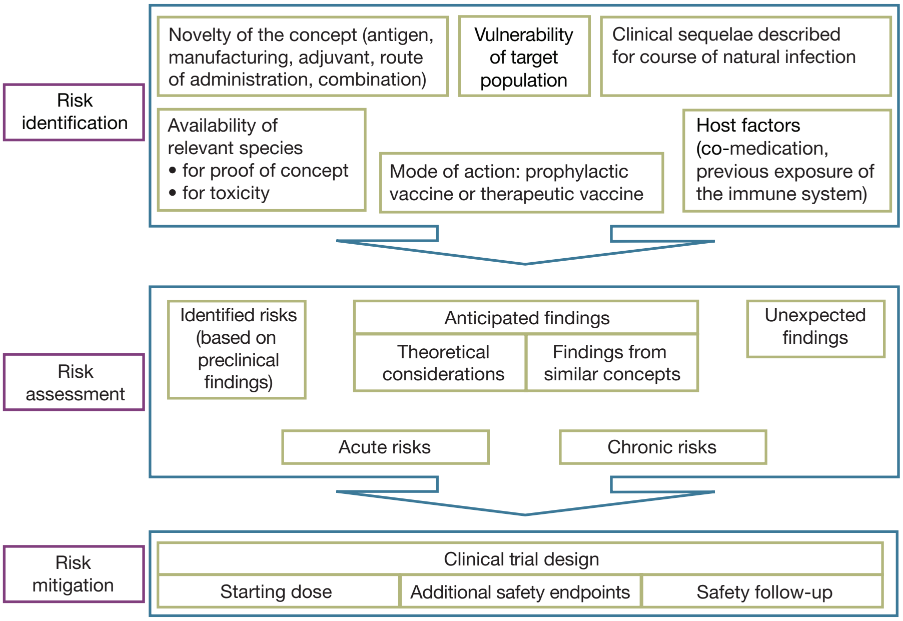

Risk identification
Novelty of the concept (antigen, manufacturing, adjuvant, route of administration, combination)
Vulnerability of target population
Clinical sequelae described for course of natural infection
Availability of relevant species
• for proof of concept
• for toxicity
Mode of action: prophylactic vaccine or therapeutic vaccine
Host factors (co-medication, previous exposure of the immune system)
Risk assessment
Identified risks (based on preclinical findings)
Anticipated findings
Theoretical considerations
Findings from similar concepts
Unexpected findings
Acute risks
Chronic risks
Risk mitigation
Clinical trial design
Starting dose
Additional safety endpoints
Safety follow-up

Figure 1 Risk assessment for a vaccine intended for first-in-human administration.

the relevant animal model specifically on the no-observed-adverse-effect level (NOAEL). This approach was also chosen for the agonistic anti-CD28 monoclonal antibody TGN1412, but was apparently insufficient to prevent the highly elevated pharmacodynamic effect, a massive cytokine-release syndrome[16]. Thus, the EMA guideline advocates an alternative approach, that is, a calculation based on the minimal-anticipated-biological-effect level (MABEL), being the dose level at which a minimal biological effect in humans is expected by *in vitro* or *in vivo* data. It is based on the occurrence of any biological effect, not only toxicity. Thus, the MABEL approach usually results in a much lower dose than that calculated with the NOAEL approach, which relates to toxicity findings. Here, the classic paradigm of a dose-dependent effect (including toxicity) is implicitly assumed—a principle already questionable for certain biotechnological medicinal products that can exhibit distinct pharmacodynamic effects at a low dose. For vaccines, additional—or even alternative—considerations need to be made because often thresholds for eliciting an immune response exist. Thus, the principle of little dose increases in cohorts might not be applicable here regarding the toxicity (if any) of the vaccine itself and consequences of the elicited immune response. In addition, if no correlate of clinical protection yet exists for the respective vaccination or if thresholds of antibodies are different between serotypes included in a vaccine (e.g., pneumococcal vaccines), the respective dose for a MABEL would be difficult to determine.

If a similar vaccine exists, for instance, a conventional Bacillus Calmette Guérin (BCG) vaccine in relation to new BCG-based tuberculosis-vaccine developments, information about the immunological pathways and clinical effects (efficacy and safety) may be extrapolated to indicate a safe starting dose and a possible test setting for the first-in-human clinical trial. If such a 'prototype' product is not available, a safe starting dose can be achieved by dissecting the different aspects that characterize a vaccine (**Fig. 2**), which are discussed below.

Vaccine antigen. Traditionally, the vaccine antigen consists of a live-attenuated pathogen, an inactivated pathogen or a recombinant or chemically synthesized antigen that resembles the natural antigen of the pathogen. An attenuated vaccine strain has impaired replication competence in humans. Here, a dose escalation starting from a low dose of the vaccine can indeed be feasible when it comes to test safety of the pathogen itself. But for inactivated pathogens and synthetic antigens, this may be less feasible because the protein or polysaccharide in itself might not exert any toxicity at all. These considerations apply both to dose escalation in nonclinical safety studies and for first-in-human trials. A MABEL approach may well be feasible only for certain types of antigens and may produce misleading results and even a false feeling of safety for other types.

Concerning direct toxicity, nonclinical studies might already be of help here. If there is a direct toxicity of the antigen, for instance, this normally will be apparent in nonclinical toxicity studies.

Adjuvants and immunomodulators. Adjuvants are an important component of a vaccine and a rapidly growing number of new adjuvant systems are used to enhance the immune response either through ideal presentation of the antigen or by immunomodulating effects. They are traditionally composed of mineral salts (e.g., alum), or more advanced developments derived from microorganisms like muramyl dipeptide, monophosphoryl lipid A or trehalose dimycolate. The mechanism of action of adjuvant emulsions includes the formation of a depot at the injection site enabling the slow release of antigen and the stimulation of antibody-producing plasma cells. Other adjuvants may be particulate antigen delivery systems (that is, liposomes, polymeric microspheres, nano-beads, immunostimulating complexes and virus-like particles), polysaccharides or nucleic acid–based adjuvants[17]. They may consist of combinations of two or more adjuvant systems (e.g., AS04) or not even be part of the formulation at all but concomitantly administered (that is, cytokines).

Some of these adjuvants are well known or at least 'established' through their long use. For newer adjuvants, however, less exposure data are available and the ideal dose of adjuvant with a certain antigen (content) has to be sought each time a novel vaccine is developed.

The dose of a novel adjuvant may feasibly be found through a MABEL approach. For example, a dose-dependent effect might exist for adjuvants targeting Toll-like receptors. A MABEL approach could also be used for immunomodulating adjuvants like cytokines. However, a threshold effect could exist here as for some antigens or other adjuvants.

Novel adjuvants can be species specific (e.g., cytokines), posing an additional challenge to find a relevant animal model (see discussion further below). Thus, even individual testing of the adjuvant or the immunomodulator in a separate first-in-human study might become necessary. Experience gained with a specific adjuvant in another vaccine could be considered supportive data, but it cannot be excluded that the same adjuvant causes serious side effects in combination with a different antigen. In any case, a thorough risk assessment is necessary also for the adjuvant.

The elicited immune response. The elicited immune response surely represents the main 'active' principle of vaccination. The vaccine antigen (such as a protein or polysaccharide) may in itself be harmless (and cause only unspecific local reactions), but the immune response against it could be harmful. Antibodies can cross-react with physiological structures and the concept of 'molecular mimicry'[18] is one of the hypotheses by which autoimmunity is explained. Antibodies, as well as $CD4^+$ T-helper cells that are part of activating and promoting a specific immune response against an antigen, might not only detect the target antigen they are intended for (that is, the pathogen) but also cross-react in an unwanted fashion with other structures that have a similar formation. T-cell recognition is 'degenerate'[19], meaning that T cells also react with structures that have less than 100% identity with the T-cell receptor's primary target.

Thus, for risk estimation of a novel vaccine candidate one needs to consider the immune response that is elicited by a vaccine as a potentially 'toxic' principle. Because activation of the immune system and the resulting immune response is not necessarily dose dependent and may be associated with particular threshold considerations, MABEL might also not be feasible here. One possible solution could be nonclinical studies. Unfortunately, for observation of a potential cross-reactivity animal toxicology data might not always be sufficiently helpful because the biological structures of animal and human organs regarding epitopes are different. Cross-reactivity of the sera of accinated animals with animal organs might not necessarily imply that the same would happen in humans and, likewise, absence of cross-reactivity or autoimmunity in animals would not imply safety in humans.

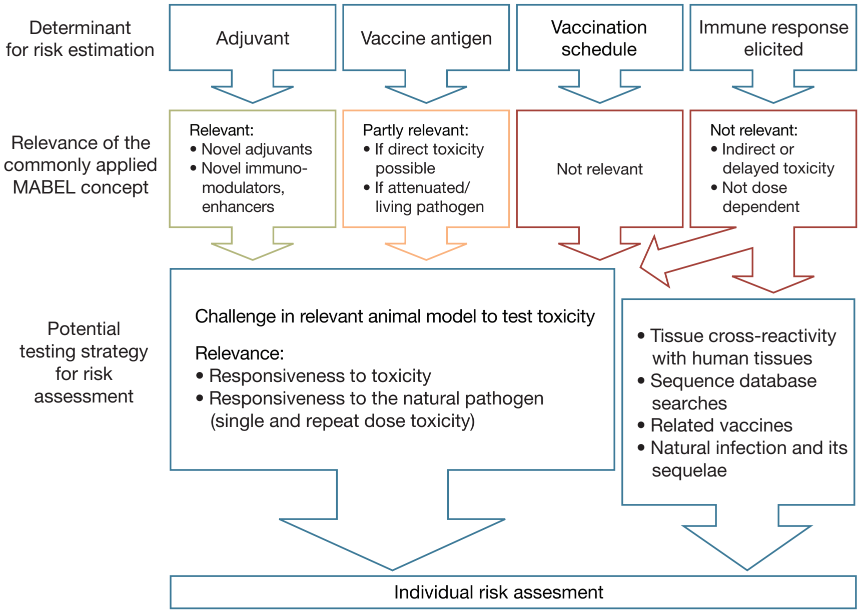

Determinant for risk estimation
Adjuvant
Vaccine antigen
Vaccination schedule
Immune response elicited
Relevance of the commonly applied MABEL concept
Relevant:
• Novel adjuvants
• Novel immuno-modulators, enhancers
Partly relevant:
• If direct toxicity possible
• If attenuated/ living pathogen
Not relevant
Not relevant:
• Indirect or delayed toxicity
• Not dose dependent
Potential testing strategy for risk assessment
Challenge in relevant animal model to test toxicity
Relevance:
• Responsiveness to toxicity
• Responsiveness to the natural pathogen (single and repeat dose toxicity)
• Tissue cross-reactivity with human tissues
• Sequence database searches
• Related vaccines
• Natural infection and its sequelae
Individual risk assesment

Figure 2 Factors to be considered for the starting dose of a vaccine for first-in-human administration.

On the other hand, *in vitro* tissue cross-reactivity studies performed with animal sera can be helpful. This approach, in which animals are vaccinated and their sera tested for cross-reactivity with human tissue sections, is routinely carried out in nonclinical toxicology testing of monoclonal antibodies. In the case of a vaccine product, the animal species might not necessarily have to be 'relevant' (see discussion below) because the animals are used rather to obtain the antibodies that can then be tested for toxicity. In addition, for the choice of the species, one may have to take into account that unrelated species may produce cross-reactive antibodies. These might be deleted in highly related species due to a tolerance for self that is shared by related but not by unrelated species. Nevertheless, such results can be useful to create a 'worst-case scenario' for cross-reactivity and may be helpful when considering risks. It is acknowledged that cross-reactivity studies have their inherent difficulties as data may be misleading if artifacts arise due to tissue preparation and fixation procedures. Even if true binding or cross-reactivity is observed, this might not necessarily point to a safety concern as the tissue structure may not be accessible under physiological conditions in humans. A shortcoming of this approach is that it tests only the antibody response; a T-cell response cannot be tested. This is problematic as T cells might be the main driver for toxicity in humans. Even so, such findings should be regarded as potential safety signals and will be helpful in assessing risks. For a first-in-human study (and subsequent clinical studies), such signals trigger the implementation of relevant clinical safety endpoints to detect potential clinical manifestations of such bindings or confirm that it does not occur in patients. These will be aimed mostly at subclinical changes, for instance, echocardiography in case binding to human cardiac tissue sections had been observed. It is fully acknowledged that such events can be rare and that the true risk of occurrence might not even become apparent before marketing authorization and use in large numbers of people. Nevertheless, for novel vaccination, principles such as precautionary measures are helpful in risk identification and mitigation strategies.

This discussion shows that for a first-in-human trial for a novel vaccine one also needs to consider the definition of risk. The EMA guideline[1] was written to detect and mitigate acute risks like cytokine release syndrome. For vaccines, such acute events derive either from an allergic reaction or are elicited by an adjuvant that triggers a skewed immune activation. The cross-binding of sera, however, would not be included in such acute risks because autoimmunity or other symptoms elicited by a real cross-binding of antibodies take a longer time. Antibodies have to be formed after vaccination and clinical manifestation normally requires a certain time following binding of those antibodies to the target organ structures. Therefore, these 'long-term' risks are hardly suitable to determine any inter-subject interval for administration in a sequential dosing concept but are rather meant to define suitable endpoints as discussed above. Another potential long-term risk is the possibility for a paradoxical enhancement of the disease (e.g., overstimulating immune cells by prolonging presentation of the pathogen antigen–antibody complexes).

The vaccination schedule. For most vaccines, single-dose administration is insufficient to establish immune protection as well as boosterable long-term persistence of the immune response. Adverse effects not triggered by a first dose might be triggered or will become detectable only during completion of the primary vaccination regimen or at the time of booster vaccination. These phenomena are known as positive re-exposure. The risk might then increase with the frequency of administration of a specific vaccine to achieve an acceptable immune response and might be particularly high for vaccines that must be administered at regular intervals. In view of these effects, the vaccination schedule also needs to be taken into consideration for the definition of the safe starting dose. As such safety issues cannot readily be predicted, a simulation of the vaccination schedule in animals should be made. In addition, appropriate safety evaluations in vaccinees and regular and extensive follow-up visits suitable to detect late effects (up to several months) must be implemented.

Quality/CMC considerations

At the time the step from animals to humans is made in drug development, the product should already be very well characterized. The potency of bacterial or viral antigens in the vaccine should be given special attention as this is a crucial factor to mediate toxicity and other adverse reactions. Therefore, specifications for potency should ideally be set sufficiently narrow. Specifications being too wide might project into false dose estimations, thus leaving room for uncertainty regarding the validity of dosing assumptions defining the starting dose. Assays measuring impurities, sterility and inactivation of biological agents have to be available at this early point in development. If possible, components (e.g., reagents, adjuvant and excipients) should be referenced (for trials in Europe) to the European Pharmacopoeia where monographs are available.

Of course, the manufacturing should be undertaken according to good manufacturing practice. Newly developed components have to be described in great detail, including chemical definition and biological structure, normally all the way down to amino acid sequence. For recombinant vaccines, as much data as possible on post-translational modifications, like addition of sugar structures (glycosylation), should be provided. Depending on the nature of a novel vaccine, similarity to human cell structures, receptors, nucleic acid or other possibly 'immunoactive' structures have to be described and evaluated with respect to (unwanted) interaction within the organism (see also discussion elsewhere in this article). If changes were made in the production process after the nonclinical studies, comparability would have to be shown between pre- and post-change product as per the relevant guideline[20] to demonstrate that the nonclinical data supporting a first-in-human use can still be applied.

Nonclinical considerations

Although animals present 'good models' for a variety of human physiological functions, they also have significant limitations when it comes to species-specific aspects; diseases induced by infectious agents relevant to humans may not exist in animals or may cause different symptoms. Likewise, certain adverse reactions can be seen only in humans and some adverse events of special interest cannot be predicted or reproduced in animals (e.g., a potential impact on functions of the central nervous system, especially learning difficulties or development of speech). Thus, a 'relevant' animal model is needed that maps the respective disease to be prevented as well as organ systems influenced by the new agent. In this respect, it is important to discuss the concept of the 'relevant species' specifically for vaccines. For a vaccine as well as for certain other biologicals, there needs to be a distinction of 'relevance' in respect of susceptibility and the clinical course of infection with the pathogen (proof of concept) and in respect of reliable prediction of safety and toxicology of the vaccine in humans.

Regarding safety evaluation, the ICH S6 guideline[21] defines a "relevant risk" species as one in which the test material is pharmacologically active due to the expression of the receptor or, in the case of monoclonal antibodies, an epitope. This is not feasible for a vaccine because here the medicinal product—the vaccine—itself is in most cases not the active principle (but the immune response against it is) and the target structure is the pathogen or infected cell containing the pathogen. This needs to be taken into account in the planning of the nonclinical development strategy.

When it comes to proof of concept, the relevant species might have to be defined differently. Here, the relevant species is one that is susceptible to infection with the pathogen and at best also resembles clinical features of humans suffering from infection and its subsequent resolution. Relevant animal models for most kinds of vaccine-targeted diseases exist (e.g., ferrets for influenza and chimpanzees for hepatitis A and B), but for specific scenarios investigators might have to combine different approaches to describe the human infection and the way the vaccine will prevent it. When selecting an animal model, which type of immune response is elicited in conjunction with the adjuvant, for example, the kind of T-cell response (cytotoxic T cells or T-helper cell responses), is also among the factors to be considered.

If a novel adjuvant is species specific (e.g., a cytokine), then the relevant animal model might have to be chosen based on the activity of the adjuvant in the respective animal species. On a case-by-case basis such an immunomodulator might have to be exchanged with the homolog active in the respective species. Also the immune response against a given vaccine antigen might be different in animals and in humans. Thus, extrapolation of data is difficult and often not feasible. Nevertheless, a nonclinical proof-of-concept study is usually mandatory before a first-in-human trial can be commenced because it adds valuable data to the overall concept of vaccine development and is needed to decide on the benefit/risk estimation to allow the first-in-human trial to be initiated (that is, to provide a rationale that the vaccine is likely to fulfil its purpose). A practical shortcoming can be the nonavailability of some animal models (e.g., the aforementioned chimpanzee model for animal protection reasons). A crucial point is to carefully consider physiological systems that might be or will be affected and how a vaccine could affect the response of different immunological cells that would be observed during natural infection; for example, overstimulated T cells can result in unexpected acute and chronic adverse events. Use of worst-case-scenario data from animals obtained with different doses of antigen and adjuvant and/or immunomodulator can be helpful in the estimation of the likelihood of such events to occur in humans, up to a full-blown systemic inflammatory response syndrome with its adverse impact on heart, liver, kidney and the central nervous system. For some organ systems and physiological scenarios, computer models are available that derive their accuracy from data that have been collected in all kinds of previous studies in humans of different age, gender and with co-morbidities; these can be helpful tools to estimate potential reactions only seen in human organisms[22]. Whether such models are useful for the development of vaccines has to be considered on a case-by-case basis and may best be discussed with regulators upfront.

Usually, only single and repeated dose toxicity studies in at least one animal species are required before first-in-human administration (repeat dose toxicity for most vaccines that are applied at least twice). For vaccines that target children and/or women of child-bearing potential, the influence on the reproductive system has to be explored. Here, different animal models might be defined as 'relevant' compared with the other nonclinical studies. For the emerging class of genetically modified biological systems, the risk of possible gene transfer into humans (or the human germ line) also needs to be quantified. Reproductive toxicity includes male and female reproductive capacity as well as the possible influence of transferred genes on the development of the embryo/fetus during pregnancy. This might indeed be an issue, given the complex changes to the maternal organism during pregnancy, including maternal-fetal exchange (hormones, antibodies and so forth). Therefore, the possible influence on fetal development (bone structure, central nervous system, organs and so forth) has to be closely surveyed as well.

Clinical trial design considerations

One central aspect of clinical trial design is the translation of potential findings from the nonclinical and *in vitro* studies (e.g., unexpected cross-reactivity of induced antibodies in animals with human tissue, to suitable clinical endpoints). As discussed, surveillance of subjects should be designed on a risk-based approach including acute and chronic risks (**Figs. 1** and **2**). Because they can affect immunological responses, several intrinsic and extrinsic factors influence the conduct and structure of a clinical trial design.

Intrinsic factors derive from subjects enrolled in the trial. They can cover, for example, concurrent diseases (e.g., HIV and malaria) and genetic polymorphisms, including major histocompatibility complex (MHC) haplotypes, receptor sensitivity or organ function. Ethnic factors, drug habits and nutritional status also directly affect the immune system.

Extrinsic factors derive from the socioeconomic background of the region where a trial takes place. Crucial factors for vaccines are the climate (that is, the ability to maintain the cold chain for the product), diagnostic and case definition practices. Drug compliance influences the trial subjects' view on multi-dose vaccinations as well as repeated visits for blood draws and adverse event checks. Some of these factors cannot be controlled or avoided (e.g., MHC haplotypes), but should nevertheless be considered in clinical trial protocol design as relevant. For example, developers elect to conduct a trial in a region where disease incidence or prevalence is high because only in this region would subjects be sufficiently motivated to comply with the trial protocol. Also crucial are local views on regulatory practice and good clinical practice as well as methodology and endpoints for the trial. This last instance, of course, influences all drug trials and is not unique for vaccines, but local ethical or religious views determine the acceptability of certain vaccines, as can be seen by the difficulty in eradicating polio, and might even be more an issue with vaccines preventing sexually transmittable diseases. To take into account these factors, the EMA has drafted a reflection paper containing examples of product groups and special extrinsic factors influencing studies[23] and the ICH has issued 'frequently asked question' paper E5 (ref. 24).

Statistical methods for limiting trial size. In first-in-human studies, only a very small number of participants are enrolled to minimize risk in light of the usually—at this point—nonexisting benefit for the enrolled study subjects (if, for example, the vaccine dose when deciding to follow the MABEL approach is too low and is maybe immunogenic but not yet protective). As most first-in-human studies have dose escalation in their procedure reliable measures for proceeding to the next dose cohort have to be implemented. Besides orientation from nonclinical animal challenge studies, several statistical methods limit the number of study subjects while at the same time allowing good estimates of nontoxic (minimal toxicity dose, MTD) and beginning effect (minimal effective dose, MED) levels. This includes, for example, the standard 3+3 cohort analysis and the continual reassessment method (CRM). The CRM is usually used to estimate the maximum tolerated dose but might be used as well to define the MED and MTD when starting from a dose level estimated to be between nontoxic and a beginning effect, as previously observed in nonclinical studies[25,26].

Surveillance of subjects. Safety is not restricted to 'tolerability' as this rather relates to local tolerance of the vaccine only. First, as usually only healthy participants are included in these studies, all possible control mechanisms must be applied. These include recording of routine laboratory parameters, including those specific to the expected interaction of the vaccine with the physiological environment, such as differential blood count and blood chemistry. Systematic evaluation should also include the recording of parameters in organs previously observed to be affected in animals (e.g., liver enzyme levels associated with hepatotoxicity) and those deduced from tissue cross-reactivity studies. Imaging techniques like (contrast) magnetic resonance imaging, computer tomography, ultrasound or X-ray of suspected vulnerable tissues as well as regular medical surveillance (electrocardiography or clinical examination) before, during and after the application of the new vaccine are also common. In addition, a first-in-human administration should not only be performed in a suitable hospital environment that provides the investigator with all necessary equipment, including an adjourning intensive care unit, but also cover a time span estimated to include all possible short-term adverse events and/or serious adverse events. After this period has elapsed, subjects are released from the trial center and examined as outpatients at regular intervals until the end of the expected interference induced by the vaccine (that is, long-term adverse events). Agencies often request long-term follow-up visits up to six months from the start of the trial, depending on the perceived potential risk for long-term events like autoimmunity.

If a genetically modified organism is used in a vaccine, there could be the risk of shedding (feces, urine) or direct transmission by means of a local inflammatory reaction at the injection site (e.g., smallpox or tuberculosis vaccination). Here, special environmental risk assessments are needed[27], and risk estimation thus implies not only vaccinees but also the persons coming into contact with them. Where possible, the existence of individuals of vulnerable immunological status (e.g., those with immunosuppression, premature newborns, the elderly, people with atopic diseases and those with severe co-morbidities in which an infection could be life-threatening) that are in contact with trial participants should be considered in the trial protocol and before the trial commences.

Pediatric studies. In contrast with most conventional pharmaceuticals, the target population for many vaccines is infants and children. First use in a pediatric population is, therefore, a particularly critical step that again needs careful consideration with respect to additional animal studies that might potentially be required (juvenile animals), further dose reduction and different dosing schemes. In addition, studies in children regardless of age are ethically difficult if no comparator yet exists and the disease to be prevented is at the same time not life threatening. Thus, justification of the trial design has to be very thorough, covering availability of a comparator (at least established medicinal use), impact and epidemiology of the disease as well as resulting age escalation/de-escalation planned.

For the different age groups, separate studies are usually required by European Union (EU; Brussels) regulators, especially in view of the new EU Paediatric Regulation[28], which entered into force in 2007. Here, the crucial point of decision is whether testing of the different subgroups should be done by age de-escalation or whether the disease to be prevented has its peak in the first few weeks and months of age and thus, the age group with the highest risk of infection as well as the maximum benefit by the vaccination should be vaccinated first. This approach should be agreed upon on a case-by-case basis involving (in Europe) the Paediatric Committee of EMA and, in general, the regulatory authorities concerned in the respective member states where the clinical trial is conducted. Vaccination of infants as the first age subgroup in the pediatric field has been agreed upon for the new live tuberculosis vaccines as infants are at the highest risk of tuberculosis in the first two years of life (thus, there is a dire need) and vaccination with the established BCG vaccine takes place shortly after birth (ideal comparator).

Guidance for this field is provided in various documents by EMA (http://www.ema.europa.eu/htms/human/paediatrics/sci_gui.htm). As in the EU, all different pediatric age groups (up to 18 years of age) will usually have to be evaluated separately in accordance with the European Paediatric Regulation; possibly only very small numbers for the individual trials will be available.

Conclusions

Most vaccines have an excellent safety record. As vaccination against infectious diseases contributes hugely to public and individual health all over the world, one needs to exercise caution when discussing risks associated with vaccines so as to avoid false and misleading signals for the public and politicians. Such a balanced view is particularly important with the emergence of new infectious agents (e.g., SARS and H1N1 influenza), the continued battle against neglected (tropical) diseases and the re-emergence of pathogens and vectors worldwide displaying increasing resistance to existing therapeutic agents. In this context, the establishment of new vaccines both against known and novel infectious diseases, as well as the improvement of established vaccines through novel techniques (e.g., genetic modification), is of the utmost importance. In addition, vaccination usually represents the cheapest and at the same time the most effective means of disease prevention worldwide.

In this context, a balanced and reasonable approach for first-in-human studies of a novel vaccine candidate is crucial to ensure safety of trial participants. The principles of the EMA guideline need to be applied in a reasonable and scientific way based on how prophylactic and therapeutic vaccines against infectious diseases function. Some principles, like the MABEL or NOAEL approaches, might require very careful adaptation to the specific needs and/or aspects of any given product, including a novel vaccine, as we discuss above. If a first-in-human trial for a vaccine does apply the MABEL strategy (e.g., for a novel adjuvant) and implements gradual dose increases, this merely represents the first step in defining the safety of administration. It must not be mistaken as 'dose finding' for immunogenicity, safety and tolerability. These are integral parts of further vaccine development to arrive at a dose that is maximally safe and immunogenic.

The discussion in this article demonstrates that the definition of a starting dose for a novel vaccine might not be straightforward; indeed, 'automatic' use of the MABEL approach might lead to misleading results. When uncertainty or doubt arises, we strongly advise manufacturers to seek discussion with regulatory agencies. This can be done either on a national level with the respective national competent authority in the EU member states (e.g., in Germany, the Paul-Ehrlich-Institut) or on a European level by using the CHMP Scientific Advice Procedure[29] (http://www.ema.europa.eu/htms/human/raguidelines/sa_pa.htm.). The former approach has the advantage of a direct discussion with the competent authority later responsible for evaluating and granting the clinical trial application in the respective EU member state. On the other hand, approaching European authorities has the advantage of receiving a European position on respective issues. Regulators are increasingly open for dialog, even at very early stages of development as well as for the development of future vaccines; such a dialog is to be considered an increasingly important factor for success.

The principles discussed in this article apply primarily to prophylactic and therapeutic vaccines against infectious agents. However, many of the principles discussed here might also readily be applied to other classes of vaccines, including therapeutic 'anti-tumor vaccines'. Because of their different immunological mode of action, these products should not be grouped with traditional vaccines and thus, according to their specific mode of action and nonprophylactic timing of use, have been classified as 'immunotherapy medicinal products'.

It is important to emphasize that not every novel vaccine or adjuvant system bears a high risk and 'higher risk' might likewise imply 'more effective' concepts (e.g., enhanced immunogenicity or protection against pathogens where no functional vaccine principle exists yet). The European guideline for first-in-human trials is intended to be a step forward to develop innovative compounds more safely. It is a certainty that even this guideline and all precautionary principles will never reduce the risk to zero. Transition from nonclinical studies will always be a risk but is also a necessity to develop more efficacious medicines against human diseases.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

DISCLAIMER

The views expressed in this article are our personal views and may not be understood or quoted as being made on behalf of the EMA committees or reflecting the position of the EMA committees or one of the CHMP Working Parties.

1. European Medicines Agency. *Guideline on Strategies to Identify and Mitigate Risks for First-in-Human Clinical Trials with Investigational Medicinal Products CHMP/SWP/28367/07* (EMA, London; 2007). <http://www.ema.europa.eu/pdfs/human/swp/2836707enfin.pdf>
2. US Food and Drug Administration (FDA). *Guidance for Industry and Reviewers, Estimating the Safe Starting Dose in Clinical Trials for Therapeutics in Adult Healthy Volunteers* (FDA, Washington, DC; 2002). <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm078932.pdf>
3. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). *Guideline E8: General Considerations for Clinical Trials.* (ICH, Geneva; 1997). <http://www.ich.org/LOB/media/MEDIA484.pdf>
4. European Medicines Agency (EMA). *Note for Pre-clinical Pharmacological and Toxicological Testing of Vaccines CPMP/SWP/465/95.* (EMA, London; 1997). <http://www.ema.europa.eu/pdfs/human/swp/046595en.pdf>
5. European Medicines Agency (EMA). *Guideline on Clinical Evaluation of New Vaccines CHMP/VWP/164653/05.* (EMA, London; 2006). <http://www.ema.europa.eu/pdfs/human/vwp/16465305enfin.pdf>
6. Nathanson, N. & Langmuir, A.D. *Am. J. Hyg.* **78**, 16–28 (1963).
7. Polack, F.P. *Pediatr. Res.* **62**, 111–115 (2007).
8. Yang, Z.Y. *et al. Proc. Natl. Acad. Sci. USA* **102**, 797–801 (2005).
9. Knudsen, K.M. *et al. Int. J. Epidemiol.* **25**, 665–673 (1996).
10. Hughes, R.A., Hadden, R.D., Gregson, N.A. & Smith, K.J. *J. Neuroimmunol.* **100**, 74–97 (1999).
11. Marth, E. & Kleinhappl, B. *Vaccine* **20**, 532–537 (2001).
12. Aguilar, J.C. & Rodriguez, E.G. *Vaccine* **25**, 3752–3762 (2007).
13. Schneider, C.K. *Expert Rev. Clin. Pharmacol* **1**, 327–331 (2008).
14. European Medicines Agency (EMA). *Guideline on Adjuvants in Vaccines for Human Use EMA/CHMP/VEG/134716/*2004 (EMA, London; 2005). <http://www.ema.europa.eu/pdfs/human/vwp/13471604en.pdf>
15. European Medicines Agency (EMA). *Explanatory Note on Immunomodulators for the Guideline on Adjuvants in Vaccines for Human Use EMA/CHMP/VWP/244894/*2006 (EMA, London; 2006). <http://www.ema.europa.eu/pdfs/human/vwp/24489406en.pdf>
16. Suntharalingam, G. *et al. N. Engl. J. Med.* **355**, 1018–1028 (2006).
17. Plotkin, S. & Offit, O. (eds). *Vaccines,* edn. 5 (Saunders Elsevier, New York; 2009).
18. Wucherpfennig, K.W. *J. Autoimmun.* **16**, 293–302 (2001).
19. Gran, B., Hemmer, B., Vergelli, M., McFarland, H.F. & Martin, R. *Ann. Neurol.* **45**, 559–567 (1999).
20. European Medicines Agency (EMA). *Guideline on Comparability of Biotechnology-Derived Medicinal Products after a Change in the Manufacturing Process—Non-Clinical and Clinical Issues EMA/CHMP/BMWP/101695/2006* (EMA, London;2007). <http://www.ema.europa.eu/pdfs/human/biosimilar/10169506enfin.pdf>
21. European Medicines Agency (EMA). *ICH Topic S 6. Note for Guidance on Preclinical Safety Evaluation of Biotechnology-Derived Pharmaceuticals (CPMP/ICH/302/95).* (EMA, London; 1998). <http://www.ema.europa.eu/pdfs/human/ich/030295en.pdf>
22. Gibson, G.G. & Rostami-Hodjegan, A. *Xenobiotica* **37**, 1013–1014 (2007).
23. European Medicines Agency (EMA). *Reflection Paper on the Extrapolation of Results from Clinical Studies Conducted Outside Europe to the EU-Population EMA/CHMP/EWP/692702/2008* (EMA, London; 2009). <http://www.ema.europa.eu/pdfs/human/ewp/69270208en.pdf>
24. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). *Topic E 5 (R1): Ethnic Factors in the Acceptability of Foreign Clinical Data* (ICH, Geneva; 2008). <http://www.ich.org/LOB/media/MEDIA481.pdf>
25. Neuenschwander, B., Branson, M. & Gsponer, T. *Stat. Med.* **27**, 2420–2439 (2008).
26. Thall, P.F. & Lee, S.J. *Int. J. Gynecol. Cancer* **13**, 251–261 (2003).
27. European Medicines Agency (EMA). *Guideline on Environmental Risk Assessments for Medicinal Products Consisting of, or Containing, Genetically Modified Organisms (GMOs) EMA/CHMP/BWP/473191/2006* (EMA, London; 2008). <http://www.ema.europa.eu/pdfs/human/bwp/47319106en.pdf>
28. European Parliament. *Official J. Eur. Union* **27**, L378/1–19 (2006).

Biopharmaceutical benchmarks 2010

Gary Walsh

Over the past four years, several new types of experimental biologic treatment have received commercial registration, but the emergence of biosimilars represents the biggest shift in the biologic approval landscape.

The rate of approval of new biopharmaceuticals has slowed over the past four years. Only 25 new biological entities (NBE) came onto the US or EU market since 2006, when we last updated the biopharmaceuticals marketplace. With a total of 58 approvals (including biosimilars and 'me-too' products), the number of biopharmaceuticals on the market now numbers just over 200 products.

The largest change from our previous survey[1] is the rise of biosimilars. Fourteen such drugs were approved in Europe, and biosimilar regulatory pathways were finalized in many other world regions, including the United States.

In terms of experimental therapies that have now been registered, the past four years have witnessed the approvals of the first (preventive and therapeutic) cancer vaccines and the first bispecific monoclonal antibody (mAb). However, we still await the approval of a gene therapy–based product, and the commercialization of small interfering RNAs (siRNAs) and therapies based on human embryonic stem (hESC) cells or induced pluripotent stem (iPS) cells remains some ways off.

Several precedents have been set in the types of manufacturing systems used to produce biologics: the first products have been produced in insect and yeast (*Pichia pastoris*) cell–based systems; and yeast and plant cell–based systems have been used to produce products with engineered glycosylation patterns. Increased focus upon innovation and streamlining within upstream and downstream processing is also evident, with disposable systems coming increasingly to the fore.

Gary Walsh is at the Industrial Biochemistry Program, Department of Chemical and Environmental Sciences, and the Materials and Surface Science Institute, University of Limerick, Limerick City, Ireland.
e-mail: *gary.walsh@ul.ie*

In the following article, I provide an update on biopharmaceuticals approved during the past four and a half years (from January 2006 until June 2010), examining which types of biopharmaceuticals have been launched and for what indications. As in previous articles, I have not included tissue-engineering products, which the US Food and Drug Administration (FDA) classifies as pure medical devices.

New arrivals

Among the 58 biopharmaceuticals (**Table 1** for definition) that gained approval over the past four and a half years within the European Union and/or the United States are 30 hormones, growth factors and other regulatory molecules, 13 mAb-based products, 4 blood-related proteins, 2 subunit vaccines and 9 additional products, including fusion proteins and therapeutic enzymes. Whereas an average approval rate of 13 products a year suggests a vibrant sector, further analysis reveals a more modest underlying performance, as just over 40% (25) were genuinely new biopharmaceutical entities. In contrast, nearly 50% (28) of the products approved were biosimilars, reformulated or me-too versions of previously approved substances (**Box 1** and **Table 2**). Additionally, five of the products approved for the first time in one region had previously been approved in a different region before 2006.

The underlining figure of 25 genuinely new biopharmaceutical approvals compares unfavorably to the approval rates reported in previous benchmark articles of 2006 and 2003 (27 and 30 products, approved over only three-and-a-half-year time spans, respectively). The year 2008 was particularly disappointing in this regard; only four genuinely new biopharmaceutical entities gained approval in the United States—Arcalyst (rilonacept; Regeneron, Tarrytown, NY, USA), Cimzia (certolizumab pegol; UCB, Brussels), Nplate (romiplostim; Amgen, Thousand Oaks, CA, USA) and Recothrom (recombinant human (rh)Thrombin; Zymogenetics, Seattle)—with no truly new product coming on stream in Europe.

Although general trends should not be projected from data describing such a short time frame, the overall number of new biological entities (NBEs) to gain approval over the past few years has been disappointing and is markedly lower than rates recorded over earlier periods[2] (**Fig. 1**). Moreover, in addition to modest approval numbers, few of those products approved are likely to reach blockbuster status, as many of the new biopharmaceuticals are approved for rare or orphan indications. Only four products—Arzerra (ofatumumab; GlaxoSmithKline (GSK), Brentford, UK); Removab (catumaxomab; Fresenius Biotech, Munich), Provenge (sipuleucel-T; Dendreon, Seattle) and Vectibix (panitumumab; Amgen, Thousand Oaks, CA, USA) are indicated to treat cancer, with two others aiming to prevent cancer (the cervical cancer vaccines Cervarix (GSK) and Gardasil (quadrivalent L1-encoded virus-like particle (VLP) vaccine incorporating human papillomavirus (HPV) genotypes 6, 11, 16 and 18; Merck, Whitehouse Station, NJ, USA). Biosimilars and reformulated products by and large enter a marketplace where significant product competition already exists. For example, the period witnessed the approval of eight (mainly biosimilar) erythropoietins (EPOs), which join a stable of five previously approved EPOs that have market advantage, albeit in the context of an overall market valued at almost $10 billion annually. Likewise, the six (biosimilar) filgrastim-based products approved for the treatment of neutropenia join three filgrastims previously approved by the traditional US Biological License Application (BLA) pathway.

Despite these reservations, the biopharmaceuticals sector still represents a significant and growing proportion of the overall

Box 1 Me too

Table 2 New biopharmaceuticals by category

Category	Products
Genuinely new biopharmaceuticals	Actemra/RoActemra, Arcalyst, Arzerra, Atryn, Cervarix, Cimzia, Elaprase, Elonva, Gardasil/Silgard, Ilaris, Kalbitor, Lucentis, Myozyme, Nplate, Preotach, Prolia, Provenge, Recothrom, Removab, Scintimun, Simponi, Soliris, Stelara, Vectibix and Victoza
Biosimilars	Abseamed, Binocrit, Biograstim, Epoetin-α hexal, Filgrastim hexal, Filgrastim ratiopharm, Nivestim, Omnitrope, Ratiograstim, Retacrit, Silapo, Tevagrastim, Valtropin and Zarzio
Reformulated me-too and related	Accretropin, Biopoin, Eporatio, Extavia, Exubera, Fertavid, Lumizyme, Mircera, Novolog mix, PEGintron/ribetol combo, Pergoveris, Opgenra, Vpriv and Xyntha
Previously approved elsewhere	Increlex, Macugen, Naglazyme, Orencia and Tysabri

The single biggest category of approvals (28 products) is variants of previously approved products. Beyond the 14 biosimilars approved within the European Union (**Table 2**, listed by trade name) the remaining products within this grouping are effectively reformulated versions of pre-existing products. Cangene's Accretropin, for example, is the eighth recombinant somatropin approved by the FDA. Extavia (interferon-β 1-b) has the same composition, pharmaceutical form and indications as Betaferon/Betaseron (approved since 1993) and is manufactured for Novartis by Bayer Schering. Pfizer's ill-fated inhalable insulin product, Exubera, contained recombinant human insulin as its active ingredient. Schering-Plough's Fertavid contains follitropin-β (rhFSH) as its active substance; the identical substance is found in Puregon, which gained approval initially in 1996. Lumizyme is effectively replacing the previously approved Myozyme in the United States. Roche's Mircera is a PEGylated form of the recombinant EPO found in Neorecormon, approved initially in 1997. The active element in Novo Nordisk's NovoLog is insulin aspart (a fast-acting engineered insulin). This particular product is a 50:50 formulation mix of soluble insulin aspart and insulin aspart-protamine crystals. Insulin aspart had been previously approved both formulated on its own (Novolog and Novorapid) and as a mix (Novomix 30 and Novolog mix 70/30). As its name suggests, Schering's PEGintron/Rebetol combo pack contains PEGintron and Rebetol capsules, PEGintron having gained approval as a standalone product in 2000. Serono's Pergoveris simply contains a fixed-dose combination of follitropin alfa (rhFSH) and lutropin alfa (rhLH), which have been individually marketed for years as Gonal F and Luveris, respectively. The active ingredient present in Howmedica's Opgenra (a recombinant bone morphogenetic protein) is identical to that present in its other product, Osigraft, approved within the European Union since 2001. Vpriv, like the previously approved product Cerezyme, is a recombinant glucocerebrosidase enzyme. Finally, the active ingredient in Wyeth's (Madison, NJ, USA) Xyntha is a recombinant B domain–deleted coagulation factor VIII, containing the same active ingredient as the company's previously approved product Refacto, which it is replacing. The same CHO cell line is used for its manufacture but details of both upstream and downstream processing have been revised, with the aim of limiting still further the risk of prion/viral contamination of the product. The primary manufacturing alterations introduced include the use of a chemically defined culture medium containing recombinant insulin, but which is free from albumin or other ingredients derived from human and/or animal sources, the replacement of the immunoaffinity purification step with an affinity step dependent upon a synthetic ligand and the introduction of a nanofiltration step.

Finally, five products (Increlex, Macugen, Naglazyme, Orencia and Tysabri), although approved for the first time within the European Union within the indicated time period, had actually gained approval before 2006 in the United States.

pharmaceutical market. Sales of all biologics totaled $94 billion by 2007 and represented the fastest growing segment of the $600 billion pharmaceutical industry. Recombinant therapeutic proteins (excluding antibodies) recorded aggregate global sales of $61 billion in 2009, whereas mAb-based products notched up an additional $38 billion[3], yielding an overall 2009 global biopharmaceuticals market value of $99 billion.

Among the most prominent blockbusters are mAb-based products indicated for treating cancer—Rituxan/MabThera (rituximab; Genentech/Roche, Basel/Biogen Idec, Cambridge, MA, USA), Herceptin (trastuzumab; Genentech), Avastin (bevacizumab; Genentech), Erbitux (cetuximab; ImClone, Branchburg, NJ, USA/Bristol-Meyers Squibb, New York) and Vectibix—as well as anti-tumor necrosis factor alpha (TNF-α) antibodies (Enbrel (etanercept; Amgen), Remicade (infliximab; Centocor, Horsham, PA, USA), Humira (adalimuma; Cambridge Antibody Company/Abbott, Abbott Park, IL, USA) and Cimzia (certolizumab; UCB, Brussels)—with each of these two product groups generating $18 billion in sales in 2009. The next most lucrative grouping are insulin and insulin analogs, collectively generating $13.3 billion in sales, followed by EPO-based products whose collective sales value stands at $9.5 billion. Five of the ten top-selling products (**Table 3**)—and four of the top five—are mAb based, confirming the preeminence of this product group within the biopharma sector.

Looking at each region independently during this period, a total of 49 biopharmaceuticals were approved in the European Union (**Fig. 2**). However, this includes 14 biosimilars, eight reformulated or me-too products and five products previously approved in the United States. Overall, therefore, only 22 genuinely new biopharmaceutical active ingredients debuted in Europe over those four and a half years. Although the European regulatory reporting structure make unambiguous identification of genuinely new molecular (either chemical or biopharmaceutical) entities challenging, it appears that a grand total of 120 such products came on the market from January 2006 to June 2010. Genuinely new biopharmaceuticals therefore represent only 18% of all new approvals (down from 22% in our last reporting period of 2003–2006).

In the United States, the same time period witnessed the approval of 21 genuinely new biopharmaceuticals—similar to the European Union. A grand total of 99 new molecular entities and original BLAs were approved within the United States in the same time frame, suggesting that 21% of all genuinely new drug approvals were biopharmaceuticals, down from the 24% reported for 2003–2006.

Eleven reformulated or me-too products were also approved in the United States, giving a total of 32 biopharmaceuticals for that region. In looking at absolute numbers of biopharmaceuticals approved, the difference between Europe and the United States (49 versus 32) is almost entirely due to the advent of biosimilar approvals in the European Union.

With the exception of Recothromb, all NBE's approved within this article's period of review are parenterally administered. Recothromb (recombinant human thrombin) is indicated for the control of minor bleeding during surgery and is applied directly to the site of bleeding. The approval in 2006 of Pfizer's (New York) inhaled insulin product, Exubera, represented a watershed in terms of biopharmaceutical delivery by means of inhalation. However, the product's subsequent withdrawal from the market due to poor patient demand represented a setback of equal magnitude, and eventually triggered the discontinuation of Novo Nordisk's and Lilly's inhalable insulin programs. However, these disappointments did not deter MannKind (Valencia, CA, USA), which recently received a letter of acceptance from the FDA regarding their new drug application (NDA) for their inhalable insulin product, Afrezza.

New antibody approvals

Half (13 of 25) of the genuinely new biopharmaceuticals to come on the market in the period under review are antibodies. 2009 was a particularly noteworthy year in this context, with seven mAb products coming on the market for the first time in the United States and/or the European Union. Four of those products—Arzerra, Ilaris (canakinumab; Novartis, Basel), Simponi (golimumab; Centocor) and Stelara (ustekinumab; Centocor) are fully human products. These join just two previously approved fully human antibodies (Vectibix, approved in 2006, and Humira, approved in 2000). Another technically noteworthy product is Removab, the first bispecific mAb to come on the market, approved in the European Union in 2009 (**Box 2** and **Fig. 3**).

Although 13 new mAb approvals over the survey period represent a respectable performance, several of these products target relatively modest markets or face the prospect of stiff competition from already approved products. Ilaris and Soliris (eculizumab, Alexion; Cheshire, CT, USA), for example, are directed to orphan indications (cryopyrin-associated periodic syndrome and paroxysmal nocturnal hemoglobinuria, respectively), whereas Cimzia and Simponi join three previously approved TNF-α inhibitors (Humira, Enbrel and Remicade). However, TNF-α inhibitors are used mainly in the treatment of rheumatoid arthritis and Crohn's disease, which represent large, lucrative markets. The current global market for rheumatoid arthritis therapies approaches $11 billion, whereas the market for biologics to treat inflammatory bowel disease (most notably Crohn's) could reach $5 billion by the end of next year[5,6].

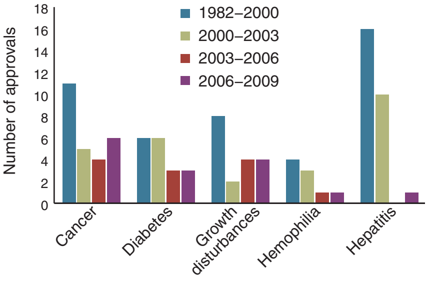


Figure 1 Number of approved biopharmaceuticals in five major markets.

Cancer vaccines

The approval of two preventive cancer vaccines represents another milestone within the current survey period. The vaccines, Gardasil and Cervarix, protect against the types of human papillomavirus (HPV) that cause most cervical cancer, which is second only to breast cancer in global incidences in women. HPV infections represent the most prevalent sexually transmitted disease worldwide, with 50% of young women being infected within five years of becoming sexually active. Two HPV strains (HPV 16 and HPV 18) are highly carcinogenic and are believed responsible for as many as 70% of invasive cervical cancers. Approximately half a million new cases of cervical cancer are diagnosed annually, culminating in an annual death rate approaching 300,000 (ref. 7). Cervarix is a divalent vaccine, comprising recombinantly produced VLPs of truncated major capsid L1 proteins from HPV types 16 and 18. Gardasil, on the other hand, is a quadrivalent vaccine containing recombinant VLP forms of the major capsid protein from HPV types 6, 11, 16 and 18. In addition to vaccinating against cervical cancer, the latter product also affords protection against genital warts, 90% of which are caused by HPV strains 6 and 11. Industry analysts forecast that these vaccines could each ultimately generate annual revenues exceeding $1 billion[8], although the market expansion into some states and/or world regions may be affected by resistance to the products on moral or religious grounds.

The first therapeutic cellular vaccine, Dendreon's Provenge, for metastatic prostate cancer, received approval in the United States in 2010, after a long and tortuous path. Along with a recombinant prostatic acid phosphatase-granulocyte-macrophage colony-stimulating factor (PAP-GM-CSF) fusion product, Provenge comprises patient-derived cells enriched using the marker CD52, activated *ex vivo* with the fusion protein and then returned to the patient. Because of the scale-up challenges facing this autologous cell therapy, the company expects to manufacture only 2,000 doses in the first year of production, which will serve a fraction of the patients with the disease.

Biosimilars

Throughout the early 2000s, the European Union developed legislative and regulatory provisions for the approval of biosimilars, and the European Medicine Agency (EMA) developed a suite of both overarching and product-specific associated regulatory guidelines. EU biosimilar regulations necessitate the generation of comparative data between the proposed new biosimilar product and the reference product, to which it claims (bio)similarity. The application dossier (relative to the one for the original reference product) must contain a full-quality module (details of manufacture and analysis, for instance), as well as abbreviated clinical and nonclinical data modules. The robustness of the European guidelines has been validated by the approval of 14 biosimilar products (based on seven distinct active biosimilar ingredients; **Box 1**). These include two recombinant human growth hormone (hGH) products (somatropins) and seven recombinant granulocyte colony-stimulating factors (G-CSFs; filgrastims). More significant technically has been the approval of five recombinant EPO-based biosimilars, illustrating the feasibility of developing biosimilar products displaying complex glycocomponents. EPO displays one *O*-linked and three *N*-linked glycosylation sites, and its carbohydrate components constitute almost 40% of its molecular mass. Despite this level of complexity, the biosimilar products displayed glycoprofiles sufficiently similar to the reference medicines to satisfy European regulators. During the same period, however, EU applications for five different additional biosimilars (based upon three distinct active ingredients, two interferons and one insulin) were either rejected or withdrawn within the period under review in this article.

Currently, European regulators through the Committee for Human Medicinal Products' (CHMP) Biosimilar Medicinal Products Working Party are updating guidelines specific for EPO products, as well as developing guidelines for biosimilar follicle-stimulating hormone (FSH; follitropin alfa), interferon-β and, perhaps most notably, mAb-based products. The size, and structural and functional complexities of mAb-based products and their modes of action render the development of biosimilar versions particularly challenging[9]. Even so, many first-generation products have reached or are reaching the end of patent protection—including Herceptin, Rituxan, Remicade and Humira—and their market value renders them attractive biosimilar targets (**Table 3**). Although European guidelines will not be final for several months, the CHMP has already provided scientific advice relating to the development of several biosimilar mAb-based products.

Development of a biosimilar (sometimes referred to as follow-on biologic) pathway in the United States has been more tortuous, but such a framework was finally ratified in March when President Barack Obama signed his healthcare reform bill into law. This should facilitate the approval of a plethora of follow-on products in that jurisdiction over the coming years, with companies such as Merck, Cangene (Winnipeg, MB, Canada), Sandoz (Holtzkirchen, Germany), Teva (Petach Tikva, Israel), Dr Reddy's (Hyderabad, India) and Biocon (Bangalore, India) positioning themselves to take advantage of the market opportunity[10]. Other regions, too, have developed biosimilar-type regulatory pathways. Health Canada, for example, issued guidelines in March of this year for subsequent-entry biologics, whereas EMA guidelines have been directly adopted in Australia. Similar regulations have been adopted in Japan, Switzerland, Turkey and several other parts of the world.

It also remains to be seen if actual revenues generated by biosimilar products will ultimately equal the hype and controversy associated with their initial development. Cost savings achieved relative to originator product are likely to be modest (10–30%), with some forecasting that originator products will retain the bulk of the market[11]. The EU biosimilars market was estimated at $60 million in 2008, and estimates for the US market by 2013 are a modest $30 million[10]. Even so, biosimilar-type products have and will derive significant market share in regions outside these markets. Global biosimilar sales surpassed the $1.3 (€1) billion milestone in 2007. Moreover, an estimated $33.2 (€25) billion worth of biologics will have lost patent protection by 2016 (ref. 12). The eventual and almost inevitable approval of biosimilar versions of current mAb and other blockbusters in European countries and the United States will increase biosimilar market value substantially within these regions. Some analysts predict that several biosimilars will approach or surpass blockbuster status, $1.3 (€1) billion annual sales by 2017 (ref. 13).

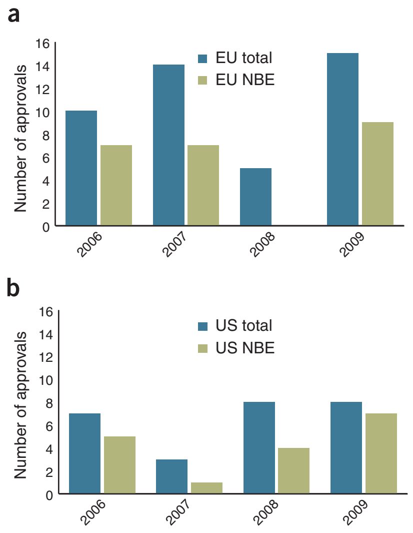


Figure 2 Biopharmaceutical approval numbers, by region, from 2006 to 2009. 'Total' is the total number of biopharmaceuticals approved in (**a**) the EU and in (**b**) the US each year. NBE is the number of biopharmaceutical entities genuinely new to the indicated region, approved in that region each year. Note that of the 23 NBEs recorded for Europe, 5 of those products had gained approval in the USA before 2006.

Production systems

Analysis of products approved from 2006–June 2010 confirm that systems based on mammalian cells and *Escherichia coli* remain the workhorses of biopharmaceutical production. Of the 58 products approved, 32 are produced in mammalian (mainly Chinese hamster ovary; CHO) cell lines, whereas 17 are produced using *E. coli*. Four are produced using *Saccharomyces cerevisiae* (Victoza, liraglutide; Novo Nordisk, Bagsværd, Denmark), Gardasil, Valtropin (somatropin; BioPartners, Barr, Switzerland) and the active ingredient in Novolog mix insulins (Novo Nordisk), whereas Atryn (antithrombin; Genzyme) and Macugen (pegaptanib; Eyetech, New York) remain the sole examples of biopharmaceuticals produced in transgenic animals and by means of direct synthesis, respectively. Notable milestones under this heading, however, include the approval of the first NBEs produced in a baculovirus-insect cell–based system (Cervarix and Provenge's fusion protein component) and in the yeast *P. pastoris* (Kalbitor, ecallantide; Dyax, Cambridge, MA, USA).

At least two plant-produced recombinant proteins (CaroRx and human intrinsic factor) are now approved for healthcare (though strictly not pharmaceutical) application in Europe. CaroRx is a mAb that binds to *Streptococcus mutans*, a primary causative agent of bacterial tooth decay. Product application to the tooth surface can prevent bacterial adherence, hence reducing the incidence of dental caries. Human intrinsic factor on the other hand is approved as a dietary supplement for the treatment of Vitamin B-12 deficiency.

A more significant milestone in plant-based production systems would be the approval of a plant-produced, parenterally administered product. Protalix Biotherapeutics' (Karmiel, Israel) Taligurase alfa (recombinant glucocerebrosidase produced in cultured carrot cells), currently in phase 3 testing, is a lead contender in this regard. Glucocerebrosidase replacement therapy is used to treat Gaucher disease—a rare lysosomal storage disorder—and current products are either extracted directly from placental tissue or produced by recombinant means in CHO cells. These products are treated with an exoglucosidase enzyme as part of downstream processing to remove sialic acid caps on the product's glycocomponent. This unmasks mannose residues, facilitating direct product uptake by macrophages (the target cell type) via cell surface mannose receptors. The plant-produced taligurase alfa is targeted to plant cell storage vacuoles during its biosynthesis, using a plant-specific, C-terminal sorting signal. The resulting product naturally displays terminal mannose residues on its glycocomponent, apparently as a result of an endogenous vacuolar carbohydrase. This eliminates the need for a subsequent exoglucosidase-mediated downstream processing step.

Protein engineering

Seventeen of the 25 genuine NBEs approved from 2006–2009 have been engineered in some way. Of the 11 antibodies approved, six are fully human, one is bispecific (**Box 2**) and the remaining ones are humanized. Two mAb fragments (Cimzia and Scintimun (besilesomab); Behringwerke, Marburg, Germany) gained approval within the current timeframe. Scintimun, used for diagnostic purposes, is derived from a traditional murine mAb. In contrast, Cimzia is both humanized and PEGylated (covalently attached to the polyethylene glycol (PEG) polyether compound). This anti–TNF-α Fab fragment was initially approved in 2008 for the treatment of Crohn's disease but, as of May 2009, it is also indicated for the treatment of rheumatoid arthritis. The single PEG moiety is a 40 kDa branched structure, attached through thiol functional chemistry to the molecule's sole cysteine residue (Cys227), present at the C-terminal end of the mAb fragment. PEGylation extends the plasma half-life of the product, enabling its once-monthly subcutaneous administration.

Box 2 A first: bispecific antibodies

Removab, approved within the European Union in 2009, is the first bispecific mAb to come on the market (**Fig. 3**). The antibody comprises a mouse κ-light chain, a rat λ-light chain, a mouse IgG2a-heavy chain and a rat IgG2b-heavy chain, and it is indicated for the treatment of malignant ascites in patients displaying epithelial cell adhesion molecule (EpCAM)-positive carcinomas.

The mAb, which is administered intraperitoneally, displays two different antigen-binding sites, a mouse-derived EpCAM-binding Fab region and a rat-derived CD3-binding Fab region. EpCAM is overexpressed on the majority of epithelial tumors, and the bispecific nature of the antibody effectively brings CD3-expressing T lymphocytes into close proximity with tumor cells. Moreover, the Fc region of the antibody facilitates docking of various immune effector cells (for example, phagocytes and natural killer cells), which, in combination and in synergy with the T lymphocytes, can induce tumor cell destruction through multiple tumoricidal mechanisms.

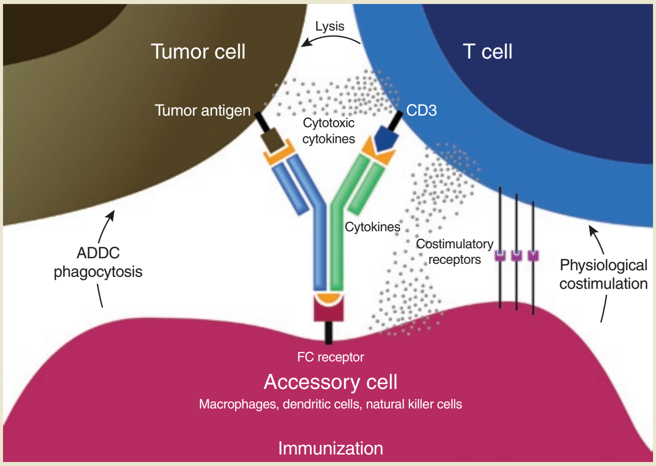


Figure 3 Structure of Removab, the first bispecific antibody to achieve approval. (Source: Fresenius Biotech, Munich)

Two of the remaining engineered products (Arcalyst and Nplate) are dimeric fusion proteins, whereas Novo Nordisk's Victoza (which was approved initially in Europe in 2009 and gained US approval this January) is a glucagon-like peptide 1 (GLP-1) analog with an attached fatty acid. GLP-1 is a member of the incretin hormone family, a group of gastrointestinal hormones that stimulate insulin biosynthesis and release. Victoza differs from the native 30 amino acid molecule in that one lysine residue is substituted by an arginine and a C16 fatty acid is acylated to the remaining lysine. These modifications increase the hormone's plasma half-life from about 2 minutes to 13 hours, facilitating once-daily product administration for the treatment of type 2 diabetes. The product has blockbuster potential, which is perhaps unsurprising given that the total antidiabetic drug market approached $23 billion in 2009 (ref. 14).

Glycoengineering

The majority of therapeutic proteins display one or more post-translational modifications (PTMs), and these PTMs invariably influence the biochemical and therapeutic properties of such proteins[15]. Glycosylation represents the most complex and the most widespread PTM, being associated with 40% of all approved products. The use of mammalian cell lines in the production of glycosylated biopharmaceuticals—despite some well-recognized limitations—is largely dictated by their ability to generate products with therapeutically acceptable glycoprofiles.

A notable trend relates to engineering the glycocomponent of glycosylated biopharmaceuticals to modify or enhance some therapeutic attribute. Earlier approaches involved downstream processing (e.g., CHO-produced glucocerebrosidase; discussed above) or the incorporation of additional glycosylation sites into the protein backbone—exemplified by Amgen's EPO analog Aranesp. Within the past 2–3 years, this latter approach has been extended to additional biopharmaceuticals, at least on an experimental level. Hyperglycosylated variants of interferon-α, for example, display a 25- to 50-fold increase in plasma half-life[16], whereas hyperglycosylated variants of FSH-enhanced ovulation and embryo maturation achieved upon administration to female mice[17].

An alternative engineering approach entails the chemical conjugation of presynthesized oligosaccharides to the protein's backbone. For example, it has recently been reported that conjugation of an oligosaccharide bearing terminal mannose 6-phosphate (MP-6) enhances cellular uptake of a lysosomal α-glucosidase (used to treat Pompe disease), likely through enhanced binding to cell surface MP-6 receptors on muscle cells[18]. Further studies have illustrated that this modification correlates with substantial therapeutic improvements in Pompe disease mouse models[19].

The glycoengineering approach most intensely pursued in recent years, however, entails the direct engineering of the actual glycosylation capacity of various producer cell types. Thus, for example, a knockout CHO cell line (so-called Potelligent technology; BioWa, Princeton, NJ, USA) has been developed, which is capable of producing completely defucosylated antibodies displaying improved cancer-killing ability[20] (**Box 3** and **Fig. 4**).

Recent advances have also been recorded in engineering the glycosylation capacity of both yeast and plant cells. Despite potential technical and economic advantages over that of mammalian systems, neither of these systems has proven suitable for the production of glycosylated products. Glycoprotein expression

Table 3 The ten top-selling biopharmaceutical products in 2009

Product	Sales value ($ billions)	Company
Enbrel (etanercept)	6.58	Amgen, Wyeth, Takeda Pharmaceuticals
Remicade (infliximab)	5.93	Centocor (Johnson & Johnson), Schering-Plough, Mitsubishi Tanabe Pharma
Avastin (bevacizumab)	5.77	Genentech, Roche, Chugai
Rituxan/MabThera (rituximab)	5.65	Genentech, Biogen-IDEC, Roche
Humira (adalimumab)	5.48	Abbott, Eisai
Epogen/Procrit/Eprex/ESPO (epoetin alfa)	5.03	Amgen, Ortho, Janssen-Cilag, Kyowa Hakko Kirin
Herceptin (trastuzumab)	4.89	Genentech, Chugai, Roche
Lantus (insulin glargine)	4.18	Sanofi-aventis
Neulasta (pegfilgrastim)	3.35	Amgen
Aranesp/Nespo (darbepoetin alfa)	2.65	Amgen, Kyowa Hakko Kirin

Source: LaMerie Business Intelligence, Barcelona

Box 3 Fucose-knockout technology

Antibodies continue to represent the most prominent category of biopharmaceuticals and thus far all approved products are of the IgG class. IgGs are glycosylated at an asparagine residue (Asn297) found within the antibody's Fc region, which plays a somewhat indirect role in triggering antibody-dependent cell-mediated cytotoxicity (ADCC), the principal mechanism by which mAbs trigger the destruction of cancer cells (**Fig. 4**).

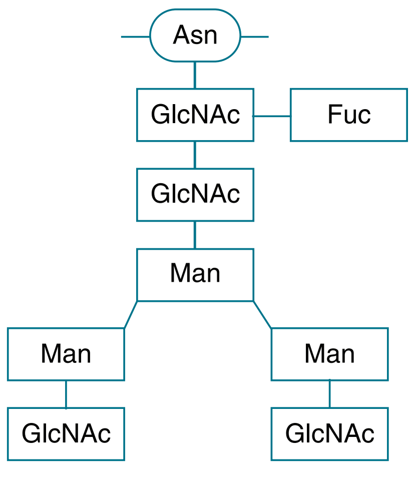


Figure 4 Representative oligosaccharide structure found in association with the Fc moiety of human IgG molecules.

Removal of the fucose residue normally resident in the glycocomponent enhances ADCC activity by up to 100-fold, a finding with obvious potential in terms of developing next-generation antibody-based oncology products. A CHO-knockout cell line has been generated that is devoid of the FUT8 gene, which encodes the fucosyltransferase enzyme that normally attaches this fucose residue to the sugar backbone. These so-called Potelligent cells therefore are capable of generating completely defucosylated antibody with consequent improved potential cancer-killing ability. BioWa (Princeton, NJ, USA; a wholly owned subsidiary of Japan's Kyowa Hakko Kirin group) has licensed the Potelligent technology to Novartis for the development of enhanced ADCC antibodies.

in yeast invariably results in the attachment of mannose-enriched sugar side chains, largely devoid of sialic acid caps, which reduces serum half-life. Glycoprotein expression in plant-based systems typically results in hyperglycosylated products containing xylose and fucose moieties that are immunogenic in man. Moreover, the sugar side chains present are usually devoid of sialic acid caps, a feature that can negatively influence their serum half-life.

Advances have been made in engineering yeast glycosylation capabilities, rendering probable their ability to produce glycosylated biopharmaceuticals displaying therapeutically acceptable glycoprofiles. From a commercial standpoint, much of this engineering has culminated in the development by Merck's wholly owned subsidiary Glycofi of engineered *P. pastoris* strains capable of producing uniformly glycosylated, sialic acid–capped products[21]. This entailed knocking out four genes (to prevent yeast-specific glycosylation) and introducing 14 additional glycosylation genes.

In plant systems, one of the most notable recent advances has been the development of systems lacking core xylose and fucose transferase activity. Greenovation Biotech (Heilbronn, Germany) has developed a glycoengineered knockout moss (*Physcomitrella patens*) lacking these activities. The moss is grown in a confined fermentor under photoautothropic conditions. The medium consists of little more than water and minerals, with light and CO_2 serving as energy and carbon sources, respectively. Elsewhere, Biolex Therapeutics (Pittsboro, NC, USA) has developed an alternative system based on engineered duckweed (*Lemna minor*) in which the endogenous fucosyl and xylosyl transferase activities are inhibited by means of an RNA interference (RNAi)-based mechanism. Interim results from a phase 2b trial of Biolex's lead product (Locteron, an interferon-α 2b) were announced in April of this year.

Upstream and downstream processing

Issues, such as healthcare reform, increased demands upon healthcare budgets and increased competition due to the advent of biosimilars continue to place downward pressure upon manufacturing costs. The past few years have seen the development of new approaches for upstream and downstream processing, including the increasing prominence of disposable systems, productivity gains and attempts to streamline downstream processing procedures.

The adoption of single-use disposable bioreactors continues to gain momentum. Prominent examples of such systems include GE Healthcare's (Bucks, UK) Wave Bioreactor and Xcellerex's (Marlborough, MA, USA) XRD Bioreactor, to name a few. GE's disposable 'cellbags' are available with up to 500-liter capacity and are made from multilayered plastic, with the inner layer being a biocompatible ethylene vinyl acetate—polyethylene copolymer. The cellbag is presterilized using γ-radiation and culture mixing is achieved by a rocking mechanism. The Xcellerex XRD bioreactor, in contrast, employs disposable bags of up to 2,000-liter capacity. The bags contain a magnetically coupled internally mounted agitator for mixing and are housed in a stainless steel shell during upstream processing.

Proponents of single-use bioreactors cite advantages, such as reduced capital equipment requirements, faster set-up times, minimal validation, the elimination of cleaning-in-place requirements and associated short production turnaround times. Even so, the recurrent cost of single-use bags and the expenses associated with their disposal as hazardous waste could be a disadvantage. Moreover, many established manufacturing sites will have already invested heavily in installing and validating traditional stainless steel–based systems, and such fixed systems remain the only real option available if high volume (>2,000 liter) production batch sizes are required[22]. Single-use systems, therefore, will likely be attractive only in certain manufacturing situations.

Expression levels achieved by mammalian cell culture systems also continue to improve. Yields on the order of 5 g/liter are now common and further gains will be underpinned by the ongoing development of selection methods for high-producing mammalian cell lines[23] as well as further media optimization and approaches to prolong the life span of cells in culture[24].

As culture yields continue to increase, the production bottleneck for some high-volume products, at least, is shifting toward downstream processing[25]. Moreover, downstream processing costs can constitute up to 80% of total manufacturing costs. Viral filters, for example, can cost $25,000 per production run, whereas a process-scale protein A column used for antibody purification could cost up to $1.5 million[26]. Such costs further fuel the desire for innovation in this area, with the main emphasis thus far falling upon process streamlining and simplification. Charged depth filters have been developed, for example, which aim to not only clarify product streams by removing cell debris but concurrently remove selected contaminants, such as DNA and selected host cell proteins. Less-used purification modalities, such as aqueous two-phase systems, crystallization and precipitation, are coming under renewed evaluation. Another ongoing line of innovation entails developing nonconventional chromatographic supports, or chromatographic application in expanded bed or other nontraditional modes.

Nucleic acid–based products

Since 2006, no nucleic acid–based products have gained approval for human use in either the European Union or United States. The first human gene therapy trial was initiated in 1989. In the intervening two decades, a total of 1,443 nucleic acid–based therapies have entered clinical trials (http://www.wiley.com/legacy/wileychi/genmed/clinical/), the majority in the United States. Cancer represents by far the most popular indication (64% of all trials), with cardiovascular disease, monogenetic disorders and infectious diseases each accounting for roughly 8% of trials. Over 1,000 trials in over 20 years and yet no products have been approved for human use in either Europe or the United States.

Recent years have witnessed some advances, however, most notably the approval of several gene-based products (DNA vaccines) for veterinary application. Fort Dodge's (part of Pfizer's Animal Health division, since its merger with Wyeth) West Nile Innovator was first approved by the US Department of Agriculture in 2005. Indicated for vaccination of horses against West Nile virus, this plasmid DNA–based vaccine incorporates gene sequences for West Nile virus surface antigens. Upon intramuscular administration, antigen expression follows cellular uptake, thereby triggering protective immunity. Additional DNA veterinary vaccines approved include the following: Apex-IHN (a DNA vaccine encoding the viral glycoprotein of infectious hematopoietic necrosis virus; Novartis Animal Health, Basel) approved in 2005 in Canada for use in salmon; LifeTide-SW5 (a DNA vaccine against growth hormone releasing hormone; VGX Animal Health, Woodland, TX, USA) approved in Australia in 2007 for prevention of fetal loss in swine; and Canine Melanoma Vaccine (a DNA vaccine encoding human tyrosinase; Merial, Duluth, GA, USA) approved in the United States in 2007 for treatment of canine malignant melanoma.

Although several dozen gene products indicated for human application have reached late-stage clinical trials—or have completed trials—no application for marketing licenses have met with regulatory approval in Europe or the United States. Although it initially granted fast track designation, the FDA rejected a product application in 2008 for Introgen Therapeutics' (Austin, TX, USA) product Advexin for the treatment of squamous cell carcinoma of the head and neck. Advexin consists of an engineered adenoviral vector harboring a functional copy of the human p53 gene. In the same year, a European marketing application for Advexin was withdrawn when the EMA issued a provisional opinion indicating that the application could not be approved based upon outstanding regulatory concerns relating to product efficacy, safety and quality.

More recently, the prospect of a gene therapy–based medicine entering the marketplace received another setback when the EMA issued a negative opinion relating to Cerepro in December 2009, ultimately prompting the company to withdraw the marketing application in March of this year. Cerepro (Ark Therapeutics, London, UK and Kuopio, Finland) is based on the application of an adenoviral vector housing the herpes simplex virus–derived thymidine kinase gene to a site of tumor resection in high-grade malignant glioma; by converting a subsequently administered prodrug, ganciclovir, to its toxic form (deoxyguanosine triphosphate), the encoded enzyme was designed to remove residual tumor cells left after surgery and therefore enhance outcomes. The EMA's negative opinion, however, was based upon failure to show sufficient efficacy, effectively triggering rejection based upon a risk-to-benefit analysis.

Currently, European regulators are considering another gene therapy–based marketing application from Amsterdam Molecular Therapeutics (Amsterdam). The company submitted an application for its lead product, Glybera, to the EMA in January of this year, so a decision is likely some way off. Glybera consists of an engineered adenoviral vector housing a human lipoprotein lipase (LPL) gene and aims to treat LPL deficiency.

Antisense and RNAi-based products also continue their development. Several antisense oligonucleotides are currently in phase 3 testing, such as Mipomersen (antisense to apolipoprotein B; Isis, Carlsbad, CA, USA). In February, Isis announced that this cholesterol-lowering product had met its endpoints in a phase 3 trial, although the ensuing optimism was somewhat dampened by concerns over high liver enzyme levels associated with some trial participants.

Genta's (Berkeley Heights, NJ, USA) lead antisense product, Genasense, continues its long sojourn in phase 3 clinical trials. This drug aims to inhibit production of BCL-2, a protein believed to prevent apoptosis of cancer cells. Its development for the treatment of a variety of cancers when used in conjunction with standard therapies continues, although an NDA for the treatment of advanced melanoma was filed as far back as 2003. Two years ago, the FDA requested additional clinical data to support Genta's application for treatment of chronic lymphocytic leukemia.

The first RNAi-based experimental therapies only entered clinical trials in 2004; the failure last year in phase 3 clinical trials of the most advanced experimental product (Opko Health's bevasiranib, siRNA-targeting vascular endothelial growth factor for treating wet age-related macular degeneration) represented a setback. As for antisense and gene therapy, technical difficulties associated with product delivery still beset this field; in addition, new chemistries with improved product stability are required for siRNA therapeutics. Even so, the recent demonstration that systemically delivered siRNA inhibited the expression of an anticancer target in human patients ($n = 3$) with solid tumors provides some encouragement[27].

Future prospects

The total global market for protein-based therapies is projected to grow at between 7% and 15% annually over the next several years[28] and protein-based products are likely to represent four of the five top-selling drugs globally by 2013 (ref. 29).

Although the coming years will likely witness the approval of a nucleic acid–based product, the majority of likely approvals will be protein based with mAb-based approvals continuing to dominate. Currently, 240 mAb products are in clinical trials, along with an additional 120 recombinant proteins[30]. Biosimilars, too, are likely to come to the fore over the next several years. In addition to the establishment of regulatory approval routes in Western markets, accelerating biosimilar sales will also be driven by rapidly growing markets in expanding economies, such as China's and India's. The market value of China's biopharma sector reached the $10 billion mark in 2008 (ref. 31), whereas the Indian market has reached almost $2 billion[32].

The proportion of engineered products coming on line will also continue to increase. Whereas engineering has traditionally focused upon the protein backbone, PTM engineering is now coming to the fore. Bench-level advances in glycoengineering will likely translate into PTM-engineered approvals in the intermediate term, with mAb-dependent, cell-mediated, cytotoxicity-optimized, glycoengineered antibodies likely to lead the way.

Recent and ongoing advances in stem cell biology also bode well for the development of stem cell therapies in the intermediate to longer-term future, and some 85 clinical trials based upon adult stem cell therapy are now underway[33]. The ability to reprogram somatic cells to form iPS cells was groundbreaking in that it provided a new source of autologous tissue for cell therapy, independent of embryonic stem cells, and potentially facilitated replacement cell therapy without the risk of immunological rejection. More recently, researchers have achieved direct cellular transdifferentiation (direct conversion of one differentiated cell type into another). Particularly noteworthy is the recent finding that fibroblasts can be converted directly into neurons through expression of just three cellular transcription factors[34].

Overall, therefore, future prospects for the biopharma sector remain bright, with ongoing research and innovation providing very deep roots, indeed, for securing and nurturing future product development.

1. Walsh, G. Biopharmaceutical benchmarks. *Nat. Biotechnol.* **24**, 769–776 (2006).
2. Rader, R. Paucity of biopharma approvals raises alarm. *GEN* **28**, 3–15 (2008).
3. Evers, P. *The Future of the Biological Market* (Business Insights, March 2010).
4. R&D Pipeline News. Special edition, March 2010. La Merie Business Intelligence, available at www.pipelinereview.com
5. Scheinecker, C. *et al.* Tocilizumab. *Nat. Rev. Drug Discov.* **8**, 273–274 (2009).
6. Melmed, G. *et al.* Certolizumab pegol. *Nat. Rev. Drug Discov.* **7**, 641–642 (2008).
7. Keam, S. Harper, D.M. Human papillomavirus types 16 and 18 vaccine (recombinant, AS04 adjuvanted, adsorbed). *BioDrugs* **22**, 205–208 (2008).
8. Crum, C., Jones, C., Kirkpatrick, P. Quadrivalent human papillomavirus recombinant vaccine. *Nat. Rev. Drug Discov.* **5**, 629–630 (2006).
9. Schneider, C.K. & Kalinke, U. Towards biosimilar monoclonal antibodies. *Nat. Biotechnol.* **26**, 985–990 (2008).
10. Carlson, B. Biosimilar market fails to meet projections. *GEN* **29**, 43–45 (2009).
11. Mackler, B.F. Biosimilars and follow on branded biologics. *GEN* **29**, 89–92 (2009).
12. The top 10 biosimilars players; positioning performance and SWOT analysis (*Business Insights*, April 2009) Available at www.globalbusinessinsights.com
13. Westphal, N.J. & Malecki, M.J. *Biosimilars 2007–2017: Shifting Payer and Physician Opinion Increases the Hurdles to Uptake* (Decision Resources, October, 2008).
14. Drunker, D.J. *et al.* Liraglutide. *Nat. Rev. Drug Discov.* **9**, 267–268 (2010).
15. Walsh, G. & Jefferis, R. Post-translational modifications in the context of therapeutic proteins. *Nat. Biotechnol.* **24**, 1241–1252 (2006).
16. Ceaglio, N. *et al.* Novel long-lasting interferon alpha derivatives designed by glycoengineering. *Biochimie* **90**, 437–449 (2008).
17. Trousdale, R.K. *et al.* Efficacy of native and hyperglycosylated follicle-stimulating hormone analogs for promoting fertility in female mice. *Fertil. Steril.* **91**, 265–270 (2009).
18. Zhu, Y. *et al.* Carbohydrate-remodelled acid α-glucosidase with higher affinity for the cation-independent mannose 6 phosphate receptor demonstrates improved delivery to muscles of Pompe mice. *Biochem. J.* **389**, 619–628 (2005).
19. Zhu, Y. *et al.* Glycoengineered acid α-glucosidase with improved efficacy at correcting the metabolic aberrations & motor function deficits in a mouse model of Pompe disease. *Mol. Ther.* **17**, 954–963 (2009).
20. Natsume, A. Niwa, R., Satoh, M. Improving effector functions of antibodies for cancer treatment: enhancing ADCC and CDC. *Drug Des. Devel. Ther.* **3**, 7–16 (2009).
21. Hamilton, S.R. *et al.* Humanization of yeast to produce complex terminally sialylated glycoproteins. *Science* **313**, 1441–1443 (2006).
22. DePalma, A. Single use systems make headway with sceptics. *GEN* **29**, 27–31 (2009).
23. Browne, S.M. & Al-Rubeai, M. Selection methods for high producing mammalian cell lines. *Trends Biotechnol.* **25**, 425–432 (2007).
24. Durocher, Y. & Butler, M. Expression systems for therapeutic glycoprotein production. *Curr. Opin. Biotechnol.* **20**, 700–707 (20 09).
25. Liu, C. & Downey, W. Contract manufacturing demands remain strong. *GEN* **29**, 53–59 (2009).
26. DePalma, A. Removing impediments in downstream processing. *GEN* **29**, 135–139 (2009).
27. Davis, M.E. *et al.* Evidence of RNAi in humans from systemically administered siRNA via targeted nanoparticles. *Nature* **464**, 1067–1071 (2010).
28. Hiller, A. Fast growth foreseen for protein therapeutics. *GEN* **29**, 153–155 (2009).
29. Goodman, M. Sales of biologics to show robust growth through to 2013. *Nat. Rev. Drug Discov.* **8**, 837 (2009).
30. Sheridan, C. Fresh from the biologic pipeline—2009. *Nat. Biotechnol.* **28**, 307–310 (2010).
31. Chinese Market for Biopharmaceuticals, Asia Market Information and Development Company, March 2009. Available via www.reportlinker.com
32. Chakraborty, C. & Agoramoorthy, G. A special report on India's biotech scenario: advancement in biopharmaceutical and healthcare sectors. *Biotechnol. Adv.* **28**, 1–6 (2010).
33. Netterwald, J. Stem cell technologies regenerative medicine. **29**, 39–42 (2009).
34. Vierbuchen, T. *et al.* Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* **463**, 1035–1041 (2010).

Will the patentability of genes survive?

Howard Leslie Hoffenberg

Recent court decisions in the United States and Europe have brought the patentability of genes under attack.

After an approximately 30-year reign during which patents were issued for genes, the issuance of such patents has come under scrutiny in both the United States and Europe. On March 29, 2010, in *Association for Molecular Pathology v. United States Patent and Trademark Office,* a US district court invalidated 15 patent claims to genes used to diagnosis susceptibility to breast cancer and likely responsiveness to certain therapeutics[1–3]. In Europe, an administrative review panel within the European Patent Office invalidated claims to these same genes in a counterpart European patent[4]. In the United Kingdom, the court invalidated a patent for an identified gene on the grounds of lack of industrial applicability[5].

The question becomes whether gene patents will survive when the issue reaches higher courts. This article discusses the reasoning by the US and European courts for invalidating the patent claims to genes and concludes with a look at the economics and politics at play, as well as one possible solution—compulsory licensing.

Reversing conventional wisdom

For approximately three decades, the reasoning that a purified gene isolated from the remainder of the contents of a living cell from which it came, in a quantity or concentration greater than that in the living cell, was a human intervention that substantially altered that which was naturally occurring so as to have new character and use, was considered sound. For example, a gene as it was found in a cell could not be used in an assay. To be used in an assay, the gene had to be isolated and increased in quantity and concentration. Accordingly, the isolated and concentrated gene had a different character and use than the naturally occurring gene. The genes that were the subject of *Association for Molecular Pathology*, held by Myriad Genetics, were located using correlation studies between cancer and DNA markers, which were in turn used to map the location of the gene within the genome. The process of identification and sequence analysis took over two years and cost over $100 million. In his decision, Judge Robert Sweet revisited this conventional wisdom and concluded that “DNA's existence in an ‘isolated’ form alters neither this fundamental quality of DNA as it exists in the body nor the information it encodes. Therefore, the patents at issue directed to ‘isolated DNA’, containing sequences found in nature are unsustainable as a matter of law and are deemed unpatentable subject matter under 35 U.S.C. §101”[6].

Judge Sweet reasoned that US Supreme Court precedent mandated that for an article of manufacture and/or composition of matter to be a patentable subject, it had to be “markedly different” from a product of nature. He also concluded that there is no patentable subject matter absent a change that results in the creation of a “fundamentally new product.” Judge Sweet picked up the “markedly different” standard from the Supreme Court case of *Diamond v. Chakrabarty*[7]. In *Chakrabarty*, the invention in question was bacteria that “ate up” oil in an oil spill. The Court wrote that: “the patentee has produced a new bacterium with markedly different characteristics from any found in nature and one having the potential for significant utility. His discovery is not nature's handiwork, but his own...”[8].

The question becomes whether gene patents will survive when the issue reaches higher courts.

However, in what sense did the Supreme Court use the term “markedly differently characteristics”? Did the Court use this phrase in the sense of establishing a standard, or in the sense of judicial hyperbole to praise the invention and bolster its decision of patentability? In its *Manual of Patent Examining Procedure*, the US Patent and Trademark Office analyzed the *Chakrabarty* decision and did not extract that the Court imposed a “markedly different characteristics” standard[9]. Further, there is also a question of whether the phrase “fundamentally new product” appears in Supreme Court precedent. Even more so, by applying a “markedly different” standard, Judge Sweet relieved himself of fully reasoning out the patentability of gene fragments, which in fragmented form do not appear naturally in a cell, involve a measure of ingenuity to deduce the operative region of the gene and have properties that increase the efficacy of molecular diagnostics.

Judge Sweet redressed the *a priori* reasoning that isolation of a gene was a human intervention that substantially altered the naturally occurring gene to impart new character and use by honing in on a passage from *The American Wood-Paper Co. v. The Fibre Disintegrating Co.*, where the high court stated, “There are many things well known and valuable in medicine or in the arts which may be extracted from divers[e] substances. But the extract is the same, no matter from what it has been taken. A process to obtain it from a subject from which it has never been taken may be the creature of invention, but the thing itself when obtained cannot be called a new manufacture”[10].

Judge Sweet rejected as nonanalogous the Fourth Circuit case of *Merck & Co., Inc. v. Olin Mathieson Chem. Corp.* In *Mathieson*, the Court found a claim for vitamin B_{12} produced by artificial fermentation in a concentration greater than 450 LLD units per milligram to be patentable over naturally occurring vitamin B_{12}, which is found in cow liver and rumen in “minute quantities.” The court distinguished the highly concentrated vitamin B_{12} from a purified substance as being different in kind from that found in nature. In particular, the court wrote that, “From the natural fermentates, which, for this purpose,

Howard Leslie Hoffenberg is at The IP Law Offices of Howard L. Hoffenberg, Esq., Los Angeles, California, USA.
http://www.ipcounselor.com/
e-mail: howard@ipcounselor.com

were wholly useless and were not known to contain the desired activity in even the slightest degree, products of great therapeutic and commercial worth have been developed. The new products are not the same as the old, but new and useful compositions entitled to the protection of the patent"[11].

The patent owner in *Association for Molecular Pathology* argued that *Mathieson* was on point because native DNA was unsuitable to be a primer or probe in molecular diagnostic tests. Judge Sweet rejected this on the grounds that the isolated DNA possessed the identical nucleotide sequence as the natural DNA sequence and that the isolated DNA functioned as a primer or probe primarily due to the nucleotide sequence identity between native and isolated DNA[12].

A question arises whether Judge Sweet properly concluded that *Mathieson* was inapposite. A gene in a living cell is present in such low abundance that it cannot be used as found in the cell or purified out of cells in any quantity to be useful for an assay. Only through Myriad's technology of isolating the gene (or fragments) did a meaningful assay arise for breast cancer susceptibility and responsiveness to certain therapeutics. Further, once the human intervention in the isolation imparts the quality of being useful in an assay, is it or is it not superfluous that there is no additional human intervention to change the chemical form and structure? In *Mathieson*, the concentrated vitamin B_{12} retained the same chemical form as natural vitamin B_{12} so as to be physiologically active.

Judge Sweet found it unnecessary to address an argument that because DNA represents the physical embodiment of biological information, on this basis it is a phenomenon of nature and exempted from being patentable subject matter. Heretofore, patentability of chemical compositions has been premised on their physical structure. This argument to exempt a chemical composition on the grounds that it conveys information ventures into uncharted legal waters.

The patents that were the subject of the lawsuit contained method claims for a molecular diagnostic. Judge Sweet held these claims to be unpatentable by applying a "machine-or-transformation" test newly articulated by the US Court of Appeals for the Federal Circuit in *Bilski v. Kappos*[13]. This test was articulated by the Federal Circuit in the context of business method patents. Applying this test, Judge Sweet found the diagnostic methods to be unpatentable mental steps. At the time Judge Sweet made his ruling, *Bilski* was under review by the US Supreme Court. The Court has since issued its decision (*Nat. Biotechnol.* **28**, 767 (2010).) In brief, the Court recognized the machine-or-transformation test as being only one calculus to assess patentable subject matter and held that there could be other tests. This at least provides a basis for making creative argument in support of the patentability of the diagnostic claims found to be unpatentable abstract mental steps.

Challenges in Europe

In Europe, Article 5 of the European Patent Convention currently provides that:

1 The human body…and the simple discovery of one of its elements, including the sequence or partial sequence of a gene, cannot constitute patentable inventions.

2. An element isolated from the human body or otherwise produced by means of a technical process, including the sequence or partial sequence of a gene, may constitute a patentable invention, even if the structure of that element is identical to that of a natural element.

In a counterpart European patent originally issued covering both the genes and diagnostic methods, several oppositions were filed against the Myriad patent[15,16]. Ultimately, a second instance panel of review of the European Patent Office (EPO) sustained the validity of a narrower version of the patent claiming diagnostics but not claiming genes or gene fragments.

In the United Kingdom, the England and Wales High Court of Justice (EWHC) put the brakes on investigators running to the patent office as soon as a gene is identified and/or postulated without sufficient experimental data as to its function and implication regarding a disease state. In more detail, investigators using bioinformatics, and not wet chemistry, identified and/or postulated a particular human protein called neutrokine-α and deduced the nucleotide sequence of a gene that coded for this protein. A European patent was successfully obtained from and defended in the EPO claiming this gene[17]. The patent did not contain a description of a real and practical way to exploit the gene. In a revocation action, the EWHC declined to follow the EPO and invalidated the patent for want of industrial applicability in that its only known use was in research to learn how the gene itself might be implicated in a disease state[18].

Conclusions

Economics seems to have been a contributing factor in the district court's decision. With Myriad Genetics charging $2,000 to $3,000 for its molecular diagnostic test resulting in a financial barrier for women receiving potentially lifesaving medical care, emotions ran deep and the political pressures were great for a result-oriented decision to make the diagnostic available at a more affordable price. Other diagnostic laboratories have claimed to be able to provide a similar test at a much lower cost, but were precluded from doing so by Myriad's patents. Hard cases make bad law[19].

Rather than developing bad law, perhaps one solution lies in compulsory licensing of some patents. With copyrighted material, Congress has mandated compulsory licensing under certain circumstances[20]. The Supreme Court has opened the door to compulsory licensing in its decision in *eBay Inc. v. MercExchange, L.L.C.* that a permanent injunction in a case of patent infringement is not automatic[21]. Germany and other countries have compulsory licensing. Compulsory licensing seems to be the vehicle for fairness and for everyone to get a "slice of the pie." Innovative companies will be able receive a return on their investment in research and development and be encouraged to do so. Consumers will have access to the technology at reasonable prices and lives will be saved and good health achieved. It is up to disinterested parties to add the weight to make compulsory licensing a reality.

COMPETING FINANCIAL INTERESTS
The author declares no competing financial interests.

1. Revised and reissued on April 5, 2010.
2. *Association for Molecular Pathology v. United States* Patent and Trademark Office, __ F.Supp. __, 2010 WL 1233416 (S.D.N.Y., 2010).
3. Claims 1, 2, 5, 6, 7 and 20 of US5,747,282; claims 1, 6 and 7 of US5,837,492; claim 1 of US5,693,473; claim 1 of US5,709,999; claim 1 of US5,710,001; claim 1 of US5,753,441 and claims 1 and 2 of US6,033,857.
4. Case number T 0666/05 - 3.3.04.
5. *Eli Lilly & Co. v. Human Genome Sciences Inc.*, [2008] EWHC 1903 (Pat) and [2010] EWCA Civ 33.
6. 2010 WL 1233416 at p.2.
7. *Diamond v. Chakrabarty*, 447 U.S. 303, 308, 100 S.Ct. 2204, 65 L.Ed.2d 144 (1980).
8. 447 U.S. 303, 310.
9. USPTO. Manual of Patent Examining Procedure, Section 2105.
10. *The American Wood-Paper Co. v. The Fibre Disintegrating Co.*, 90 U.S. (23 Wall.) 566, 593-94, 23 L.Ed. 31 (1874).
11. *Merck & Co., Inc. v. Olin Mathieson Chem. Corp.*, 253 F.2d 156, 165 (4th Cir.)
12. 2010 WL 1233416 at p.45.
13. *In re Bilski*, 545 F.3d 943 (Federal Circuit 2008).
14. *Bilski v. Doll*, 129 S.Ct. 2735, 174 L.Ed.2d 246, 77 USLW 3442, 77 USLW 3653, 77 USLW 3656 (U.S. Jun 01, 2009) (NO. 08-964).
15. EP 705903.
16. Case number T 0666/05 - 3.3.04.
17. EP 1577391.
18. *Eli Lilly & Co. v. Human Genome Sciences Inc.*, [2008] EWHC 1903 (Pat) and [2010] EWCA Civ 33.
19. *Winterbottom v. Wright* (1842) 10 M&W 109 and Karl Nickerson Llewellyn (1893–1962).
20. 17 USC § 115 entitled "Scope of exclusive rights in nondramatic musical works: Compulsory license for making and distributing phonorecords."
21. *eBay Inc. v. MercExchange*, LLC, 547 U.S. 388, 126 S.Ct. 1837.

Recent patent applications in drug discovery

Patent number	Description	Assignee	Inventor	Priority application date	Publication date
WO 2010083617	New pyrazolopyrimidine compounds that are protein kinase inhibitors; useful for treating (hyper) proliferative diseases and angiogenesis-related diseases or as a research tool in drug discovery.	Oncalis (Schlieren, Switzerland)	Capraro H	1/21/2009	7/29/2010
JP 2010164448	A dispensing apparatus used in genome-based drug discovery, with an overflow tank to temporarily accommodate cleaning liquid overflowing through an opening from the liquid storage tank storing cleaning liquid for the nozzle cleaning portion.	Matsushita Electrical Industrial (Kadoma, Japan)	Shimokawa K, Yamashita S	1/16/2009	7/29/2010
US 20100167418	A method of identifying a candidate modulator of integrin activity, comprising contacting integrin polypeptide with candidate agent and detecting binding of candidate agent to the integrin polypeptide.	Immune Disease Institute (Boston)	Luo BH, Springer TA	12/29/2008	7/1/2010
WO 2010073519	A new Alzheimer's disease model animal having a continuous increase in the concentration of amyloid-β protein in the brain; useful for monitoring *in vivo* production of amyloid-β protein or symptoms of Alzheimer's disease and in drug discovery.	Japan Health Science Foundation (Tokyo)	Kagawa S, Takikawa O	12/26/2008	7/1/2010
US 20100152280	A new oligomeric compound that is a modulator of systemic RNA interference defective-1 expression; useful for diagnosing and treating cancer and viral infection, and in drug discovery.	Isis Pharmaceuticals (Carlsbad, CA, USA)	Bennett CF, Dobie KW	5/24/2004	6/17/2010
WO 2010060019, US 20100130725	A method of characterizing a molecule, comprising collecting the biosensor response of a molecule producing a primary profile, collecting the biosensor response of a marker panel in a cell panel, extracting a specific set of biosensor parameters from each biosensor signal, normalizing each biosensor parameter against a positive control and comparing the molecule biosensor index to a library of modulator biosensor indexes.	Corning (Corning, NY, USA)	Fang Y, Ferrie AM, Lahiri J, Tran E	11/24/2008	5/27/2010, 5/27/2010
US 20100122906	A biochemical concentrator for drug discovery comprising a solution containing a target species, and an electric field generator for effecting movement of the target species within the solution and changing species flux within regions of solution.	Holm-Kennedy JW	Holm-Kennedy JW	3/16/2005	5/20/2010
US 20100119413	A system for analyzing biological samples in a microtiter plate for drug discovery research; comprises a processor that controls a gripper assembly to grip the microtiter plate along opposing sides in portrait or landscape orientation.	Beckman Coulter (Fullerton, CA, USA)	Avgerinos PN, Rizzotte SH, Turner DA	5/18/2008	5/13/2010
US 20100099190	A cultured cell construct comprising spheroids of mesenchymal stem cells; useful for developing biodevices used for diagnosis, reconstructive medicine and drug discovery.	Transparent (Chiba, Japan)	Chung U, Itaka K, Kataoka K, Nishiyama N, Ohba S, Wang W, Yamasaki Y	10/21/2008	4/22/2010
JP 2010071851	A flowthrough cell-type biosensor apparatus for use in, e.g., drug discovery. The biosensor has a switching valve with a suction element that maintains the pump in an airtight state when sucking the sample solution from a container using the pump.	Ulvac (Chigasaki, Japan)	Ito A, Mizutani T, Take R, Tanaka S	9/19/2008	4/2/2010
JP 2010051243	A new *Actinomyces* strain deposited as NITE P-621 and obtained by introducing a mycinose biosynthetic gene into the bacteria; useful for manufacturing a rosamicin derivative having antimicrobial activity and for drug discovery.	Sansho (Osaka, Japan)	Anzai Y, Fujiwara T, Iisaka Y, Kato F, Kinoshita K, Moroboshi T	8/28/2008	3/11/2010

Source: Thomson Scientific Search Service. The status of each application is slightly different from country to country. For further details, contact Thomson Scientific, 1800 Diagonal Road, Suite 250, Alexandria, Virginia 22314, USA. Tel: 1 (800) 337-9368 (http://www.thomson.com/scientific).

A holistic view of GPCR signaling

Terry Kenakin

Dynamic mass redistribution assays measure the complexity of G protein–coupled receptor signaling.

Despite the central importance of G protein–coupled receptors (GPCRs) in drug discovery, the biochemical assays that are widely used in the pharmaceutical industry do not capture the integrated response of a cell to GPCR activation. In this issue, Schröder *et al.*[1] show that a method for measuring dynamic mass redistribution offers a far more comprehensive picture of GPCR signaling compared with traditional assays, making it possible to visualize temporal receptor-activation profiles, to link whole-cell responses to individual signaling pathways and to assay GPCR-mediated drug effects, all in primary human cells. The study suggests that dynamic mass redistribution assays could provide a powerful tool for testing drug candidates in pharmacology and drug discovery.

GPCRs are seven-transmembrane-helix proteins that are coupled to intracellular signaling proteins such as G proteins (G_i/G_o, G_s, G_q and G_{12}/G_{13} proteins) and the more recently discovered β-arrestin. Traditional assays of GPCR activation use fluorescent labels to measure changes in the levels of second-messenger molecules, such as inositol phosphate or cyclic AMP. These measurements do not capture the complex ways in which biochemical signaling pathways are integrated in whole cells. Indeed, there are now many examples in which assays of integrated responses yield drug-response profiles that differ from measurements of single pathways.

Dynamic mass redistribution is a cellular process that occurs when molecules within a cell change their intracellular locations. This phenomenon can be detected by passing polarized light through the bottom portion of cells and measuring changes in the wavelength of the light using resonant waveguide technology (**Fig. 1**). The resulting optical trace can be recorded for several minutes after stimulating cells with ligands, providing a real-time readout of pharmacologically mediated changes in cellular mass. Because it is noninvasive, the assay is applicable to virtually any cell type, including primary cells relevant to a disease.

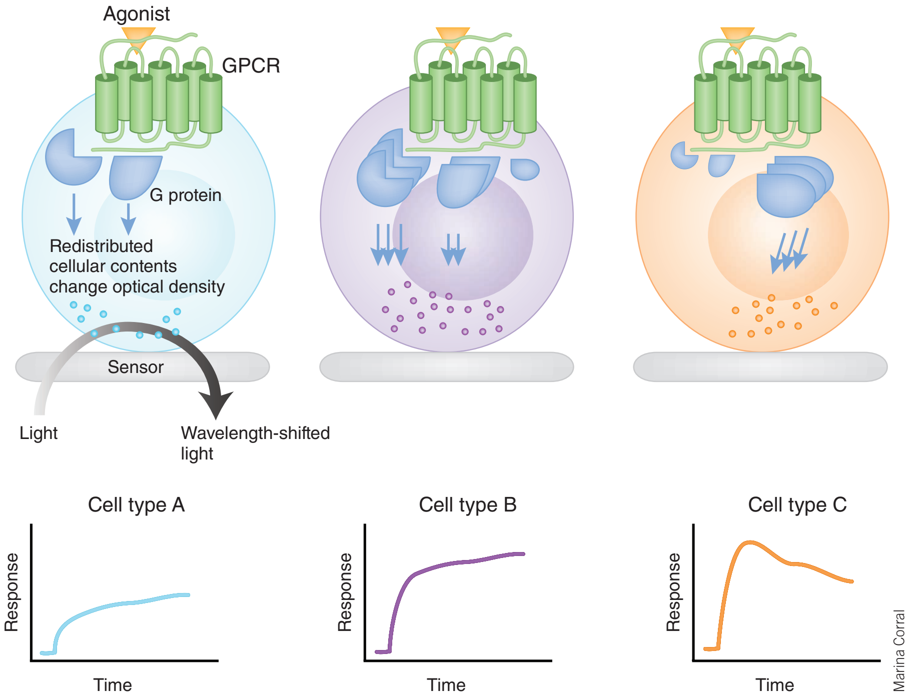


Figure 1 Measuring cell type–dependent drug efficacy using dynamic mass redistribution. A dynamic mass redistribution assay generates optical traces that represent the summation of all signaling pathways activated by a GPCR (left). The optical trace depends on the cell type, as different cells may express signaling pathway components in different relative stoichiometries (middle). If an agonist produces the same receptor active state in different cells, then different cells may produce responses of different intensity (potency). Quantitative differences can be used to derive a cell-independent measure of efficacy. If agonist binding activates different signaling pathways, then the relative stoichiometries of cell signaling components can change the shape and the intensity of the response (right).

It has been shown previously that ligands specific to GPCRs produce detectable dynamic mass redistribution signals. Schröder *et al.*[1] advance the field by demonstrating that dynamic mass redistribution responses from whole cells can be mapped to individual G-protein pathways, including the G_{11}/G_{12} pathway, for which biochemical assays are not available. They accomplish this with small molecules that inhibit or mask specific pathways. They also show how dynamic mass redistribution assays can be used to discover novel signaling complexity. For example, they find that the free fatty acid receptor FFA1, previously believed to signal only through the G_q/G_{11} pathway, also activates the Gi pathway. In another example, they uncover interactions between two signaling pathways: pretreatment with a compound that increases intracellular levels of the second

Terry Kenakin is at GlaxoSmithKline Research and Development, Research Triangle Park, North Carolina, USA.
e-mail: terry.p.kenakin@gsk.com

messenger cyclic AMP boosts the response to an agonist that signals through a different second messenger, inositol phosphate.

These advances are especially noteworthy in light of the authors' use of dynamic mass redistribution assays to measure the effects of an agonist in primary human keratinocytes, illustrating how this technology can measure signals from cells containing native levels of cell surface receptors. The ability to deconvolve signaling responses in primary human cells is particularly exciting for pharmacologists, who have often had to rely on animal systems, which can lead to incorrect predictions of clinical utility.

Another notable result of Schröder *et al.*[1] is that dynamic mass redistribution assays can measure cell type–specific responses (**Fig. 1**). The authors demonstrate differences in Gs signaling, but not G_i/G_o signaling, between HEK and CHO cells. This capability is important because recent evidence suggests that the activity of some agonists is cell-type dependent, contrary to historical assumptions[2-5]. These agonists are referred to as 'functionally selective' or 'biased' agonists. Biased agonists produce ligand-specific active-state conformations of receptors, which can differentially activate cellular pathways[6–9]. In these cases, the efficacy of a drug is affected by the cell type treated because cells vary in the relative stoichiometries of intracellular signaling components. As any new investigative drug could exhibit biased agonism, methods to assess the relevance of such activity in human systems are important.

The observations of Schröder *et al.*[1] offer new opportunities and pose new challenges to drug discovery. The opportunities stem from the potential to link receptor coupling mechanisms to the cell's phenotypic response to agonists, which could identify important drug phenotypes. The challenges will be to apply the wealth of cell-specific agonist data generated by dynamic mass redistribution assays to accurately predict agonism in therapeutic contexts. When the amount of detail provided by an assay is too great, it becomes difficult to classify drug effects, a useful exercise for enhancing drug properties through medicinal chemistry.

So how can pharmacologists incorporate dynamic mass redistribution technology into the fabric of discovery, exploiting the obvious advantages without being overwhelmed by the complexity of the data? One possibility would be to use the method to detect differences in the activity of a drug in different cell types relative to a reference agonist, thereby identifying valuable biased agonists[10]. Label-free assays are perfectly suited for this as almost any cell type can be studied. If the cell types contain sufficiently different dominant signaling pathways (that is, G-protein or β-arrestin–dominant signaling), then the relative activity of test agonists could be used to guide medicinal chemists in optimizing biased agonism. Although much remains to be done to mine the potential of dynamic mass redistribution for understanding GPCR function, Schröder *et al.*[1] have provided excellent examples of the types of pharmacological experiments that are needed to link whole-cell optical signals to cellular stimulus-response effects.

COMPETING FINANCIAL INTERESTS
The author declares no competing financial interests.

1. Schröder, R. *et al. Nat. Biotechnol.* **28**, 943–949 (2010).
2. Furchgott, R.F. in *Advances in Drug Research*, vol. **3**, (eds. Harper, N.J. & Simmonds, A.B.) 21–55. Academic Press, London (1966).
3. Colquhoun, D. *Trends Pharmacol. Sci.* **6**, 197 (1985).
4. Black, J.W. & Leff, P. *Proc. R. Soc. Lond. B* **220**, 141–162 (1983).
5. Stephenson, R.P. *Br. J. Pharmacol.* **11**, 379–393 (1956).
6. Kenakin, T. *Trends Pharmacol. Sci.* **16**, 232–238 (1995).
7. Luttrell, L.M. & Gesty-Palmer, D. *Pharmacol. Rev.* **62**, 305–330 (2010).
8. Kenakin, T. & Miller, L.J. *Pharmacol. Rev.* **62**, 265–304 (2010).
9. Violin, J.D. & Lefkowitz, R.J. *Trends Pharmacol. Sci.* **28**, 416–422 (2007).
10. Peters, M.F. & Scott, C.W. *J. Biomol. Screen.* **14**, 246–255 (2009).

Pushing the envelope on HIV-1 neutralization

Joseph G Joyce & Jan ter Meulen

The identification of broadly neutralizing monoclonal antibodies against HIV-1 may aid efforts to design a vaccine.

Despite a massive research effort stretching back more than 20 years, development of a prophylactic HIV-1 vaccine has been hindered by the ability of the virus to evade host immune defenses through rapid antigenic variation and epitope masking. Neutralizing antibodies generated in infected individuals often provide a useful starting point for vaccine design, but antibody responses against HIV-1 are in most cases highly specific for the original viral strain and do not keep pace with newly evolving quasi-species. Two new studies in *Science* by Wu *et al.*[1] and Zhou *et al.*[2] present an approach that may lead to an HIV-1 vaccine that can induce a broad and potent neutralizing antibody response. They describe the identification[1] and structural characterization[2] of human broadly neutralizing monoclonal antibodies directed toward the conserved CD4 receptor binding site on the viral envelope glycoprotein gp120. Importantly, the antibodies were selected directly from infected donor sera using a recently developed technique of direct clonal B-cell sorting[3], and the selection probes were rationally designed to identify antibodies specific for the targeted receptor binding site. The hope is that these antibodies will facilitate the engineering of vaccine candidates capable of focusing the immune response on highly conserved protective epitopes and inducing a broadly neutralizing antibody repertoire in immunized individuals.

Joseph G. Joyce and Jan ter Meulen are in the Department of Vaccine Research, Merck Research Laboratories, West Point, Pennsylvania, USA.
e-mail: joseph_joyce@merck.com

The selection strategy employed by Wu *et al.*[1] is a critical aspect of the work as it offers a way of identifying broadly neutralizing antibodies in the small percentage of infected individuals able to produce a protective antibody repertoire (called 'nonprogressers' or 'elite controllers'). Over the last decade, several such antibodies have been found[4]. Their immunologic targets include structurally conserved or functionally important epitopes, such as the CD4 binding site, chemokine co-receptor binding sites, the high-mannan glycan shield, the membrane proximal region of the viral envelope protein gp41 and the gp41 pre-hairpin intermediate. However, the coverage breadth of the antibodies to these targets is generally limited to ~40–50% of viral strains across all clades, and their potency varies widely.

Vaccine researchers have used several experimental techniques for generating neutralizing monoclonal antibodies, including human hybridoma generation, immortalization of B cells with Epstein Barr virus and combinatorial display[5]. However, the quality of the antibodies discovered with such methods is directly related to the quality of

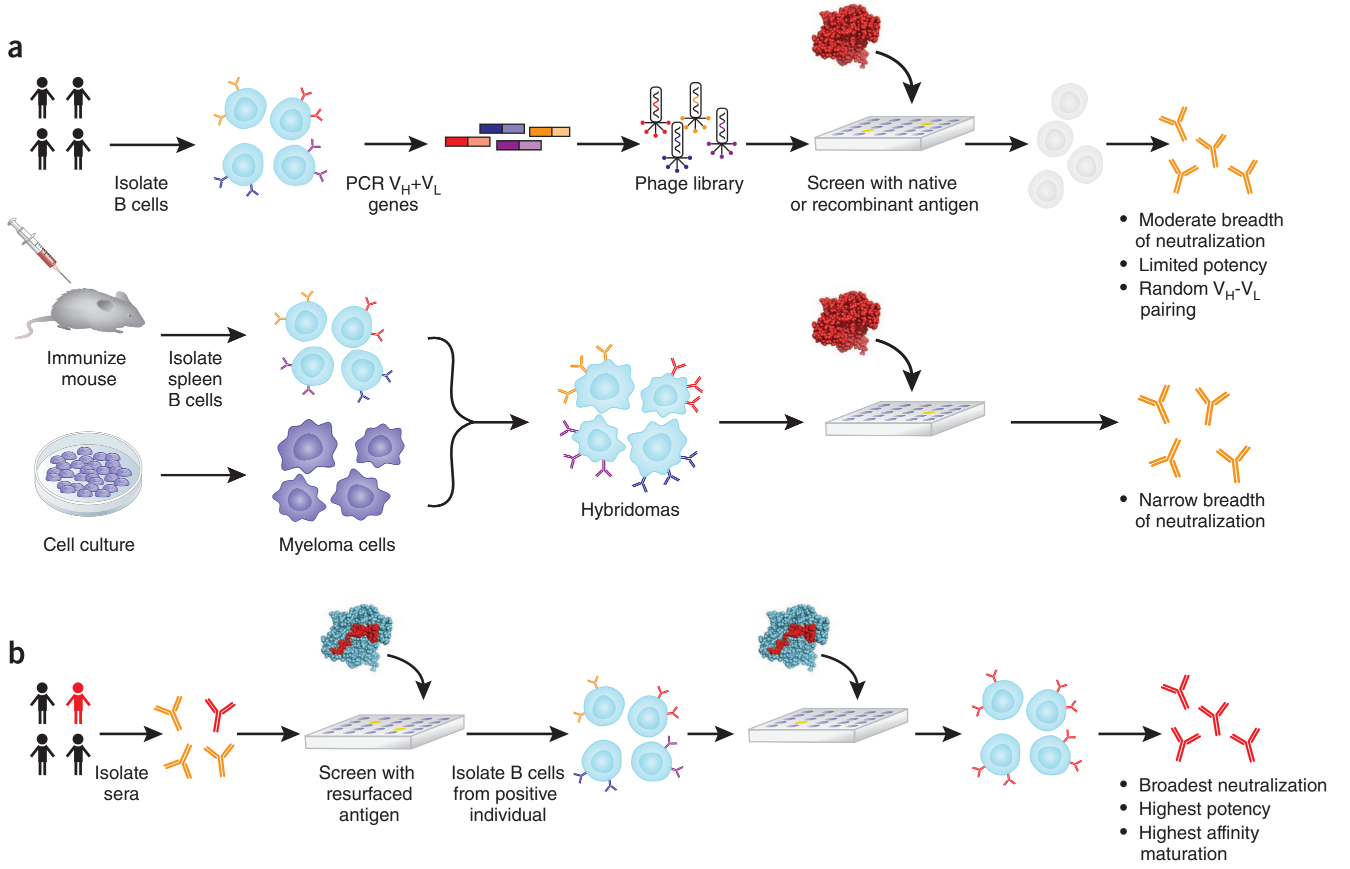

a
Isolate B cells
PCR V_H+V_L genes
Phage library
Screen with native or recombinant antigen
• Moderate breadth of neutralization
• Limited potency
• Random V_H-V_L pairing
Immunize mouse
Isolate spleen B cells
Cell culture
Myeloma cells
Hybridomas
• Narrow breadth of neutralization
b
Isolate sera
Screen with resurfaced antigen
Isolate B cells from positive individual
• Broadest neutralization
• Highest potency
• Highest affinity maturation

Figure 1 Neutralization breadth of anti-HIV monoclonal antibodies depends on antibody-generation technology and antigen configuration. (**a**) Established techniques for monoclonal antibody generation and selection include hybridomas and phage display antibody (Fab) or single-chain Fv libraries. For hybridoma generation, an animal is immunized with the desired antigen (HIV-1 gp120 is denoted in red). After spleen cell harvest and myeloma fusion, clonal supernates are screened and monoclonal antibodies of the desired specificity are identified with an appropriate functional assay. For library selection, human V_H and V_L antibody genes are isolated from naive or infected individuals and randomly cloned into filamentous bacteriophage for surface expression. Phage specific for the desired antigenic target are identified by multiple rounds of panning and the antibody genes cloned and expressed. The characteristics of most HIV-1–neutralizing monoclonal antibodies isolated in this fashion are indicated. Mouse hybridomas have not yielded broadly neutralizing antibodies against HIV-1 to date[5], although the technology has been successful for other infectious agents. (**b**) In the approach discussed here[1,2], an engineered resurfaced antigen is constructed by displaying the HIV-1 CD4 binding site (red) on an SIV gp120 framework (gray). Infected donor sera are screened for binding to antigen, and the memory-B-cell repertoire from a positive individual is propagated and screened with resurfaced antigen. Competition analysis with known CD4 binding site–directed monoclonal antibodies and affinity determination by surface plasmon resonance are used to select clones with improved breadth and potency compared to monoclonal antibodies identified by the strategies in **a**.

the antigens used for panning and selection. When antigen selection is uninformed by structural knowledge of immunologically relevant conformations—as has often been the case in HIV-1 vaccine research—the resulting antibodies are likely to be suboptimal.

The new work[1,2] uses two strategies that offer a significant advantage over previous efforts (**Fig. 1**). First, the antigen probes were designed using a technique called 'resurfacing' in which a relevant neutralizing epitope—in this case the HIV-1 CD4 binding site—is presented in the context of an immunologically irrelevant scaffold—here, a simian immunodeficiency virus (SIV) gp120 framework. The investigators used knowledge of immunologically relevant CD4 binding site conformations garnered from previous studies with neutralizing antibodies along with computational modeling to precisely define the desired epitope.

Second, monoclonal antibodies were directly isolated from individual B cells of HIV-1–infected individuals by antigen-specific, memory-B-cell sorting[3]. In this approach, donors with a strong positive serum reactivity to the probes were selected and their memory-B-cell repertoires were screened to identify clones that bind the antigen with high affinity. Nonimmortalized B-cell culture has been used previously to identify the HIV-1 broadly neutralizing monoclonal antibodies PG9 and PG16, with the primary screening assay being virus neutralization[6]. Interestingly, both antibodies recognize a cryptic epitope on the native HIV-1 envelope trimer but do not bind gp120 or gp41 in enzyme-linked immunosorbent assays.

The source of antibodies is an important distinguishing feature of the current work[1,2] because phage display and other cloned antibody libraries may show selection biases, often yielding monoclonal antibodies of relatively low affinity and moderate specificity. The most potent HIV-1 broadly neutralizing monoclonal antibodies have all been isolated directly from human B cells and share the common characteristic of extensive maturation relative to germline sequence, reflecting the immune response of the host to the evolving HIV-1 infection. One of the newly identified broadly neutralizing antibodies, VRC01, exhibited ~30% and ~20% divergence for V_H and V_L chains, respectively[1], and the antibody contained an additional disulfide bond as well as residue deletions within its L_κ-chain[2]. Similarly, PG9 and PG16 vary by

~25% from their germline parent, with no single mutation accounting for their broad cross-reactivity[7].

The discovery of broadly neutralizing monoclonal antibodies such as VRC01 raises expectations that an effective prophylactic vaccine can be generated by reverse engineering of appropriate immunogens. In principle, the antibody can be used to design candidate immunogens that focus the immune response on the desired protective epitope. For example, the crystal structure of VRC01 bound to the resurfaced gp120 probe used in its identification[2] could inform modifications of the epitope surface to increase binding affinity or mask nonproductive irrelevant antibody responses. Such modifications might include single-residue substitutions, addition of glycan shielding sites or introduction of conformational constraints. Designed immunogens could then be tested in animal models for functional responses such as competitive binding or neutralization.

Thus far, attempts to reverse engineer immunogens from neutralizing monoclonal antibodies have not met with success, although promising results were recently reported for both HIV-1 (ref. 8) and influenza[9]. Challenges include correct structural presentation of complex, discontinuous epitopes and focusing of the immune response on desired regions of the molecule. The resurfaced gp120 probe used to identify VRC01 presents the HIV-1 CD4 binding site in the context of an SIV framework[1], and immunization with this protein would be expected to produce antibodies against both target and framework. It is an open question whether the framework-specific response would be immunodominant and whether the proportion of antibodies directed to the HIV-1 CD4 binding site would be high enough to effect protection. In addition, the antibody response directed to sterically restricted or transient conformational intermediates, such as those presented on gp41 and CD4-inducible epitopes on gp120, may be thermodynamically or kinetically limited in potency.

Finally, many HIV-1–specific broadly neutralizing monoclonal antibodies have a distinctive architecture that may itself pose a challenge to vaccine design. Such antibodies are characterized by extended complementarity-determining regions, lipid-binding capability and use of domain swapping. If these structural features are genetically restricted in the general population and are critical to neutralization potency, there is at present no way to bias the immune response towards production of such antibodies. Furthermore, the extended complementarity-determining regions and hydrophobic combining sites of HIV-1 broadly neutralizing monoclonal antibodies that bind near the membrane surface may mediate polyreactivity with human proteins, such as cardiolipin and various nuclear antigens. One model postulates that polyreactive antibodies are actively eliminated from the repertoire during B-cell maturation owing to their anti-self activity, and that this accounts for the poor ability to elicit them with designed immunogens.

Despite the remaining obstacles, there is considerable cause for optimism as greater understanding of the immune system opens the door to rational manipulation. Challenges, such as the need to induce highly matured antibodies to specific epitopes, may be addressable through optimized immunization regimens and novel adjuvants. For example, a heterologous prime-boost regimen that involved priming with a canarypox vector followed by boosting with recombinant gp120 antigens has demonstrated a modicum of efficacy in a large phase 2 clinical trial[10]. Taken together, recent developments in HIV-1 research raise the prospect of an effective vaccine in the not-too-distant future.

COMPETING FINANCIAL INTERESTS
The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturebiotechnology/.

1. Wu, X. *et al. Science* **329**, 856–861 (2010).
2. Zhou, T. *et al. Science* **329**, 811–817 (2010).
3. Scheid, J.F. *et al. J. Immunol. Methods* **343**, 65–67 (2009).
4. Kwong, P.D. & Wilson, I.A. *Nat. Immunol.* **10**, 573–578 (2009).
5. Hammond, P.W. *mAbs* **2**, 157–164 (2010).
6. Walker, L.M. *et al. Science* **326**, 285–289 (2009).
7. Pancera, M. *et al. J. Virol.* **84**, 8098–8110 (2010).
8. Bianchi, E. *et al. Proc. Natl. Acad. Sci. USA* **107**, 10655–10660 (2010).
9. Bommakanti, G. *et al. Proc. Natl. Acad. Sci. USA.* **107**, 13701–13706 (2010).
10. Rerks-Ngarm, S. *et al. N. Engl. J. Med.* **361**, 2209–2220 (2009).

LINCing chromatin remodeling to metastasis

Carlo M Croce

A long intergenic noncoding RNA may promote metastatic progression by coordinating the activity of histone-modifying enzymes.

In recent years, noncoding RNAs, such as microRNAs and long intergenic noncoding RNAs (lincRNAs), have been implicated as important regulators of oncogenesis and metastatic progression[1,2]. Now, two studies in *Nature*[3] and *Science*[4], both from the laboratory of Howard Chang, have revealed how a lincRNA drives tumor metastasis and remodels chromatin. Beyond their importance for cancer biology, these papers show that a noncoding RNA can synchronize the activities of different histone-modifying enzymes to regulate gene expression.

As developmental genes can have roles in cancer, Chang and colleagues[3] first set out to investigate whether the *HOXC* locus, which contains many developmental genes, is involved in metastasis. They began by hybridizing total RNA derived from normal human breast epithelia, primary breast carcinomas and distant breast cancer metastases to ultra-dense HOX tiling arrays. Several noncoding RNAs and protein-coding exons were differentially expressed in the breast cancer samples. Among the noncoding RNAs, the lincRNA *HOTAIR* was found to be particularly strongly associated with unfavorable prognosis and metastatic disease[3]. *HOTAIR* is one of >3,000 lincRNAs that are actively transcribed and highly conserved in the human genome[5]. Functional studies have suggested that lincRNAs are involved in chromatin remodeling and in processes such as dosage compensation, imprinting, homeotic gene expression and cancer[5–8].

HOTAIR is known to recruit the polycomb repressive complex 2 (PRC2), with its H3 lysine 27 (H3K27) histone methylation activity, to specific genomic loci, especially the *HOXD* locus[6]. Chang and colleagues[3] showed that forced expression of *HOTAIR* in carcinoma cells causes a genome-wide retargeting of PRC2 to chromatin sites and altered histone H3K27 methylation, thereby altering the expression of genes known to inhibit breast cancer progression (such as cell adhesion molecules of

Carlo M. Croce is in the Human Cancer Genetics Program, Ohio State University, Columbus, Ohio, USA.
e-mail: carlo.croce@osumc.edu

the protocadherin family and the angiogenesis-related EPH receptor A1)[3]. Inhibition of *HOTAIR* expression reduced invasiveness and metastatic potential *in vivo*, suggesting that dysregulation of *HOTAIR* may have an important role in tumor invasion and metastasis.

The metastatic role of *HOTAIR* is consistent with its known physiological functions. *HOTAIR* is involved in specifying the chromatin states of fibroblasts during development, and the PRC2 occupancy observed by Chang and colleagues[3] upon *HOTAIR* overexpression resembles that observed in embryonic fibroblasts. Whereas the enhanced cellular motility and matrix invasion associated with fibroblasts early in development are necessary for embryogenesis, they are deadly when reactivated in the context of malignant disease[3,6].

The importance of PRC2 in human tumors is well established. EZH2, the subunit of PRC2 that carries the H3K27 methylase activity, is overexpressed in a wide variety of cancers, including prostate and breast cancer, and is implicated in the silencing of tumor suppressors. But until now it has remained unclear how the silencing complexes are recruited to specific genes. The demonstration of PRC2 recruitment by *HOTAIR* provides the long-sought mechanistic link (**Fig. 1**), showing for the first time that a lincRNA promotes metastatic chromatin remodeling.

In a related study[4], the Chang laboratory investigated the function of *HOTAIR* once it is bound to a target sequence. Working with HeLa cells and human primary foreskin fibroblasts, they found that *HOTAIR* serves as a scaffold for at least two histone-modification complexes: the 5′ region of *HOTAIR* binds PRC2, and the 3′ region binds the LSD1/CoREST/REST complex (**Fig. 1**). This important observation suggests that *HOTAIR* synchronizes the assembly of two different complexes to specific targets for coupled histone H3 lysine 27 methylation and lysine 4 demethylation.

Individual histone modifications are rarely encountered in isolation, but how the placement of different marks is coordinated is largely mysterious. Although it remains to be seen whether other lincRNAs possess the scaffolding function of *HOTAIR*, lincRNAs may provide a widespread mechanism for coordinating the activity of several histone-modifying enzymes. LincRNAs are not unique in their ability to do this. For example, the H3K4me3 histone methyltransferase MLL2 assembles a protein complex containing UTX, which demethylates H3K27me3 (ref. 9). However, this example involves protein-protein interactions. RNA-mediated association of two distinct chromatin remodeling complexes to coordinate methylation has not been shown previously.

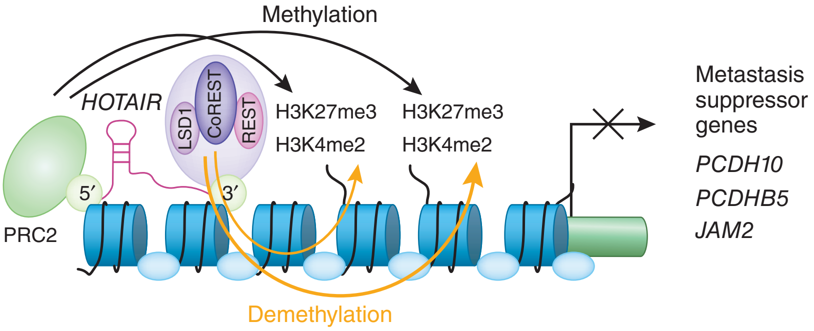


Figure 1 Schematic presentation of *HOTAIR* function in breast cancer progression. Upregulated *HOTAIR* in breast cancer cells provides a scaffold for PRC2 and LSD1-CoREST. These two protein complexes bind to the 5′ and 3′ portions of *HOTAIR*, respectively. The resulting molecular complex is bound to the promoter of genes encoding metastasis suppressors (such as *PCDH10, PCDHB5* and *JAM2*) to coordinately regulate the histone modifications H3K27me3 trimethylation and H3K4me2 demethylation, which in turn, silence expression of the target genes.

The work of Chang and colleagues[3,4] raises many questions for future research. Which transcription factor(s) or genetic mutations cause the upregulation of *HOTAIR*? Are epigenetic changes at the *HOTAIR* locus involved? Further experiments are necessary to assess the importance of lincRNAs such as *HOTAIR* in cancer, particularly the progression of the primary tumor to metastatic disease. For example, targeted overexpression of *HOTAIR* in breast epithelial cells and other normal or malignant epithelial cells could establish whether dysregulation of this gene alone causes metastasis or whether additional genetic or epigenetic changes are required[10].

From a therapeutic perspective, it may be possible to target the interaction between *HOTAIR* and the two protein complexes with short interfering RNAs, microRNAs or small molecules. Before testing this, however, it is necessary to further investigate the role of *HOTAIR* overexpression in cancer invasion and metastasis in animal models.

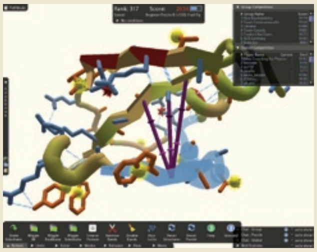
The studies of Chang and colleagues[3,4] show that lincRNAs might be as important in oncogenesis as microRNAs and classical protein-coding genes. Many lincRNAs in addition to *HOTAIR* are dysregulated in different cancers, and it will be fascinating to investigate the mechanisms of their involvement in tumorigenesis and to assess their suitability as therapeutic targets.

COMPETING FINANCIAL INTERESTS

The author declares no competing financial interests.

1. Calin, G.A. & Croce, C.M. *Nat. Rev. Cancer* **6**, 857–866 (2006).
2. Calin, G.A. *et al. Cancer Cell* **12**, 215–229 (2007).
3. Gupta, R.A. *et al. Nature* **464**, 1071–1076 (2010).
4. Tsai, M.-C. *et al. Science* **329**, 689–693 (2010).
5. Khalil, A.M. *et al. Proc. Natl. Acad. Sci. USA* **106**, 11667–11672 (2009).
6. Rinn, J.L. *et al. Cell* **129**, 1311–1323 (2007).
7. Ponting, C.P. *et al. Cell* **136**, 629–641 (2009).
8. Huarte, M. *Cell* **142**, 409–419 (2010).
9. Agger, K. *et al. Natur e* **449**, 731–734 (2007).
10. Fabbri, M. *et al. Proc. Natl. Acad. Sci. USA* **104**, 15805–15810 (2007).

The structure-solving crowd



Take a very hard biophysics problem, turn it into a computer game, allow anyone on the internet to play—scientists and nonscientists alike—and a large network of people will reach better solutions than those produced by sophisticated supercomputers. This paradigm-rattling insight is described by Cooper *et al.* in a recent paper about 'Foldit', an online game in which players compete to fold proteins into their lowest-energy conformations. Foldit presents improperly folded proteins to be optimized and directs players to the incorrect parts of the structures by highlighting high-energy areas (such as exposed hydrophobic residues, steric clashes and cavities) in color. Players, working alone or in teams, can manually tug and tweak the proteins and run simplified energy-minimization programs. Out of ten blind puzzles, humans outscored the protein-folding software Rosetta in five and played to a draw in three. (Neither side did well on the remaining two.) Human players were especially good at finding solutions that involved substantial backbone rearrangements, whereas Rosetta remained trapped in local energy minima. The authors note the "complexity, variation and creativity" of the human protein-folding strategies and the intricate social strategies that emerged, via chat and a wiki, to support the players. (*Nature* **466**, 647–651, 2010) *KA*

Tau and Fyn conspire in Alzheimer's

The molecular mechanisms underlying the toxicity of amyloid-β oligomers in Alzheimer's disease are not well understood. A major player seems to be the microtubule-associated protein tau. Tau is normally localized to the axons of neurons, whereas toxic effects are mainly observed in the dendrites, making it difficult to explain tau's involvement. During the course of the disease, tau becomes abnormally phosphorylated, detaches from the microtubules and relocates to other neuronal compartments. Ittner *et al.* elucidate how this might be detrimental for neurons. Besides microtubules, tau also interacts with a number of nonreceptor tyrosine kinases. A prominent example is Fyn, which is involved in organizing the postsynaptic machinery by phosphorylating a subunit of the NMDA receptor (NMDAR), thereby increasing its affinity to the scaffolding protein PSD95. Increased binding of NMDAR to PSD95 has been shown to cause neurotoxicity in other diseases. Ittner *et al.* now find that access of Fyn to the dendrites is regulated by tau in mice. In *Tau*$^{-/-}$ cells, Fyn is excluded from the dendrites, and expression of a truncated tau mutant that is excluded from the dendrites leads to sequestration of Fyn in the soma. When tau detaches from the microtubules upon phosphorylation, it can access the dendritic compartment, thereby increasing the Fyn concentration and consequently NMDAR phosphorylation. The authors show that therapeutically targeting the PSD95-NMDAR interaction with a cell-permeable peptide improves survival and memory in a mouse model of Alzheimer's disease. (*Cell* **142**, 387–397, 2010) *ME*

Written by Kathy Aschheim, Laura DeFrancesco, Markus Elsner & Peter Hare

Chemical inducers of HSC expansion

Hematopoietic stem cell (HSC) transplantation has been used for more than four decades to treat patients with life-threatening diseases of the blood and bone marrow. However, the challenge of identifying defined culture conditions to expand human HSCs *ex vivo* has limited full realization of the clinical potential of the approach. Boitano *et al.* address this issue by assaying CD34 and CD133 expression in cultured human HSCs screened with a library comprising 100,000 heterocyclic compounds. One of these, a purine derivative named SR1, expands CD34^{+} cells from humans, monkeys and dogs but not mice. Culturing human HSCs with SR1 increases by 17-fold the number of cells capable of hematopoietic reconstitution in immunodeficient mice. Mechanistic studies suggest that SR1 binds directly to and inhibits the aryl hydrocarbon receptor, which was not previously known to have a role in human HSC biology. (*Science*, published online 5 August 2010; doi:10.1126/science.1191536) *PH*

Higher resolution optical imaging

Conventional microscopes can resolve the position of individual fluorophores only to about half the imaging light's wavelength. Additional knowledge about the specimen can be used to increase the resolution. If, for example, well-separated individual fluorophores are imaged, the accuracy of the position measurement is theoretically limited only by the number of photons that can be collected for each light-emitting molecule. This principle lies at the heart of so-called super-resolution microscopy techniques, such as PALM or STORM. In practice, the achievable resolution has been limited to 5–10 times worse than the theoretical estimates. Pertsinidis *et al.* now show that using closed-loop feedback control to lock the signal of individual fluorescent molecules can correct for positional noise caused by thermal fluctuations. Moreover, addressing the CCD array at any desired subpixel location can minimize systematic localization errors caused by defects and dirt on the optics and especially irregularities in the charge-coupled device (CCD) arrays used for light detection. With their feedback control system, they can achieve accuracies of up to 0.5 nm, close to the theoretical limit for the number of photons collected. They use their new technology to investigate the intersubunit distances in E-cadherin dimers. Although the technology has only been applied to pairs of fluorophores, imaging applications with many molecules seem possible. (*Nature* **466**, 647–651, 2010). *ME*

Saving cone cells

In certain forms of retinitis pigmentosa (RP), cone cells persist for awhile after rod cells have died. This provides a window of opportunity to rescue cones, the loss of which results in total blindness. Busskamp *et al.* were able to do just that in mouse models of RP by delivering the gene for a well-studied bacterial halorhodopsin (light-activated chloride pump from *Natronomonas pharaonis*, cNpHR) via an adeno-associated vector (AAV). The opsin (or green fluorescent protein (GFP) in control animals), under the control of cell-specific promoters, was delivered into the subretinal space of 21-day-old mice, and isolated retinas were later tested for gene expression, light responses and the ability to relay information to ganglion cells. They found that opsin expression persisted until the mice were 110 days old and that the retinas responded to both light-on and light-off signals as well as directional signals. The transduced mice performed better than control mice on behavioral tests. Finally, in isolated human retinas transduced with a lentivirus vector, which expresses more rapidly than AAV (human retinas persist in culture for only 2–3 weeks), the researchers detected gene expression after 2–3 days, as well as light responses not seen in control retinas. (*Science* **329**, 413–417, 2010) *LD*

The BioPAX community standard for pathway data sharing

Emek Demir[1,2,*], Michael P Cary[1], Suzanne Paley[3], Ken Fukuda[4], Christian Lemer[5], Imre Vastrik[6], Guanming Wu[7], Peter D'Eustachio[8], Carl Schaefer[9], Joanne Luciano[10], Frank Schacherer[11], Irma Martinez-Flores[12], Zhenjun Hu[13], Veronica Jimenez-Jacinto[12], Geeta Joshi-Tope[14], Kumaran Kandasamy[15], Alejandra C Lopez-Fuentes[16], Huaiyu Mi[17], Elgar Pichler[18], Igor Rodchenkov[19], Andrea Splendiani[20,21], Sasha Tkachev[22], Jeremy Zucker[23], Gopal Gopinath[24], Harsha Rajasimha[25,26], Ranjani Ramakrishnan[27], Imran Shah[28], Mustafa Syed[29], Nadia Anwar[1], Özgün Babur[1,2], Michael Blinov[30], Erik Brauner[31], Dan Corwin[32], Sylva Donaldson[19], Frank Gibbons[31], Robert Goldberg[33], Peter Hornbeck[22], Augustin Luna[34], Peter Murray-Rust[35], Eric Neumann[36], Oliver Reubenacker[37], Matthias Samwald[38,39], Martijn van Iersel[40], Sarala Wimalaratne[41], Keith Allen[42], Burk Braun[11], Michelle Whirl-Carrillo[43], Kei-Hoi Cheung[44], Kam Dahlquist[45], Andrew Finney[46], Marc Gillespie[47], Elizabeth Glass[29], Li Gong[43], Robin Haw[7], Michael Honig[48], Olivier Hubaut[5], David Kane[49], Shiva Krupa[50], Martina Kutmon[51], Julie Leonard[42], Debbie Marks[52], David Merberg[53], Victoria Petri[54], Alex Pico[55], Dean Ravenscroft[56], Liya Ren[14], Nigam Shah[57], Margot Sunshine[34], Rebecca Tang[43], Ryan Whaley[43], Stan Letovksy[58], Kenneth H Buetow[59], Andrey Rzhetsky[60], Vincent Schachter[61], Bruno S Sobral[25], Ugur Dogrusoz[2], Shannon McWeeney[27], Mirit Aladjem[34], Ewan Birney[6], Julio Collado-Vides[12], Susumu Goto[62], Michael Hucka[63], Nicolas Le Novère[6], Natalia Maltsev[29], Akhilesh Pandey[15], Paul Thomas[17], Edgar Wingender[64], Peter D Karp[3], Chris Sander[1] & Gary D Bader[19]

Biological Pathway Exchange (BioPAX) is a standard language to represent biological pathways at the molecular and cellular level and to facilitate the exchange of pathway data. The rapid growth of the volume of pathway data has spurred the development of databases and computational tools to aid interpretation; however, use of these data is hampered by the current fragmentation of pathway information across many databases with incompatible formats. BioPAX, which was created through a community process, solves this problem by making pathway data substantially easier to collect, index, interpret and share. BioPAX can represent metabolic and signaling pathways, molecular and genetic interactions and gene regulation networks. Using BioPAX, millions of interactions, organized into thousands of pathways, from many organisms are available from a growing number of databases. This large amount of pathway data in a computable form will support visualization, analysis and biological discovery.

Increasingly powerful technologies, including genome-wide molecular measurements, have accelerated progress toward a complete map of molecular interaction networks in cells and between cells of many organisms. The growing scale of these maps requires their representation in a form suitable for computer processing, storage and dissemination by means of software systems. The BioPAX project aims to facilitate knowledge representation, systematic collection, integration and wide distribution of pathway data from heterogeneous information sources. This will enable these data to be incorporated into distributed biological information systems that support visualization and analysis.

BioPAX supports efforts working toward a complete representation of basic cellular processes. Biology has come a long way since the Boehringer-Mannheim wall chart of metabolic pathways[1] and the Nicholson Metabolic Map[2]. Since then, several groups have developed methods and databases for organizing pathway information[3–16], but only recently have groups collaborated as part of the BioPAX project to develop a generally accepted standard way of representing these pathway maps. Complete molecular process maps must include all interactions, reactions, dependencies, influence and information flow between pools of molecules in cells and between cells. For ease of use and simplicity of presentation, such network maps are often organized in terms of subnetworks or pathways. Pathways are models delineated within the entire cellular biochemical network that help us describe and understand specific biological processes. Thus, a useful definition of a pathway is a set of interactions between physical or genetic cell components, often describing a cause-and-effect or time-dependent process, that explains observable biological phenomena. How do we represent these pathways in a generally accepted and computable form?

Challenges posed by the many fragmented pathway databases

The total volume of pathway data mapped by biologists and stored in databases has entered a rapid growth phase, with the number of

[*] A full list of author affiliations appear at the end of this paper.

Published online 9 September 2010; doi:10.1038/nbt.1666

online resources for pathways and molecular interactions increasing 70%, from 190 in 2006 to 325 in 2010 (ref. 17). In addition, molecular profiling methods, such as RNA profiling using microarrays, or protein quantification using mass spectrometry, provide large amounts of information about the dynamics of cellular pathway components and increase the power of pathway analysis techniques[18,19]. However, this growth poses a formidable challenge for pathway data collection and curation as well as for database, visualization and analysis software, as these data are often fragmented.

The principal motivation for building pathway databases and software tools is to facilitate qualitative and quantitative analysis and modeling of large biological systems using a computational approach. Over 300 pathway or molecular interaction–related data resources[17] and many visualization and analysis software tools[3,20–22] have been developed. Unfortunately, most of these databases and tools were originally developed to use their own pathway representation language, resulting in a heterogeneous set of resources that are extremely difficult to combine and use. This has occurred because many different research groups, each with their own system for representing biomolecules and their interactions in a pathway, work independently to collect pathway data recorded in the literature (estimated from text-mining projects[23] to be present in at least 10% of the >20 million articles currently indexed by PubMed). As a result, researchers waste time collecting information from different sources and converting it from one form of representation to another. Fragmented pathway data results in substantial lost opportunity cost. For instance, visualization and analysis tools developed for one pathway database cannot be reused for others, making software development efforts more expensive. Therefore, it is imperative to develop computational methods to cope with both the magnitude and fragmented nature of this expanding, valuable pathway information. Whereas independent research efforts are needed to find the best ways to represent pathways, community coordination and agreement on standard semantics is necessary to be able to efficiently integrate pathway data from multiple sources on a large scale.

BioPAX requirements and implementation

A common, inclusive and computable pathway data language is necessary to share knowledge about pathway maps and to facilitate integration and use for hypothesis testing in biology[24]. A shared language facilitates communication by reducing the number of translations required to exchange data between multiple sources (**Fig. 1**). Developing such a representation is challenging owing to the variety of pathways in biology and the diverse uses of pathway information. Pathway representations frequently use abstractions for metabolic, signaling, gene regulation, protein interaction and genetic interaction, and these serve as a starting point toward a shared language[25]. Also, several variants of this common language may be required to answer relevant research questions in distinct fields of biology, each covering unique levels of detail addressing different uses, but these should be rooted in common principles and must remain compatible.

BioPAX addresses these challenges. We developed BioPAX as a shared language to facilitate communication between diverse software systems and to establish standard knowledge representation of pathway information. BioPAX supports representation of metabolic and signaling pathways, molecular and genetic interactions and gene regulation. Relationships between genes, small molecules, complexes and their states (e.g., post-translational protein modifications, mRNA splice variants, cellular location) are described, including the results of events. Details about the BioPAX language are available in online documentation at http://www.biopax.org/. The BioPAX language provides terms and descriptions, to represent many aspects of biological pathways and their annotation. It is implemented as an ontology, a formal system of describing knowledge (**Box 1**) that helps structure pathway data so that they are more easily processed by computer software (**Fig. 2**). It provides a standard syntax used for data exchange that is based on OWL (Web Ontology Language) (**Box 1**). Finally, it provides a validator that uses a set of rules to verify whether a BioPAX document is complete, consistent and free of common errors. BioPAX is the only community standard for biological pathway exchange to and from databases, but it is related to other standards (discussed below in the "What is not covered?" section).

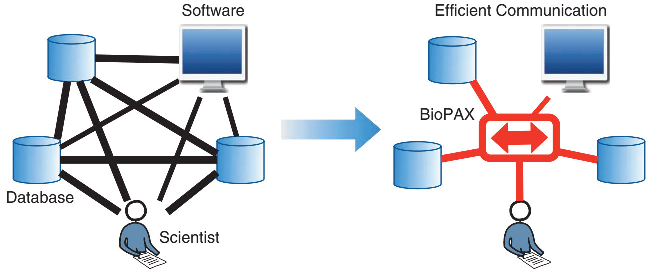


Figure 1 BioPAX is a shared language for biological pathways. BioPAX reduces the effort required to efficiently communicate between pathway users, databases and software tools. Without a shared language, each system must speak the language of all other systems in the worst case (black lines). With a shared language, each system only needs to speak that language (central red box).

Example of a pathway in BioPAX

Pathway models are generally described with text and with network diagrams. Here we use the AKT signaling pathway[26,27] as an example to show how a typical pathway diagram that can only be interpreted by people (**Fig. 3**, top left) would be represented using BioPAX (**Fig. 3**, right). The AKT pathway is a cell surface receptor–activated signaling cascade that transduces external signals to intracellular events through a series of steps including protein-protein interactions and protein kinase–mediated phosphorylation. The pathway eventually activates transcription factors, which turn on genes to promote cell survival. By representing the pathway using the BioPAX language (**Fig. 3** and **Supplementary Tables 1** and **2**), it can be analyzed by computational approaches, such as pathway analysis of gene expression data.

Representing a pathway using the BioPAX language sometimes necessitates being more explicit to avoid capturing inconsistent data. For instance, the typical notion of an 'active protein' is dependent on context, as the same molecule could be active in one cellular context, such as a cellular compartment with a set of potentially interacting molecules, and inactive in another context. Thus, capturing the specific mechanism of activation, such as phosphorylation modification, is usually required, and the presence of downstream events that include the modified form signifies that the molecule is active. Interactions where the mechanism of action is unknown can also be specified.

What does BioPAX include?

BioPAX covers all major concepts familiar to biologists studying pathways, including metabolic and signaling pathways, gene regulatory networks and genetic and molecular interactions (**Supplementary Table 3**). The BioPAX language is distributed as an ontology definition (**Fig. 4**) with associated documentation, a validator for checking a BioPAX document for errors and other software tools (**Table 1**).

Box 1 What is an ontology?

An ontology is a formal system for representing knowledge[64]. Such representation is required for computer software to make use of information. Example ontologies include organism taxonomies[65] and the Gene Ontology[40]. A formal representation allows consistent communication of knowledge among individuals or computer systems and helps manage complexity in information processing as knowledge is broken down into clear concepts that can be considered independently. Ontologies also enable integration of knowledge between independent resources linked on the World Wide Web. Such linked, structured data form the basis of the semantic web, an extension of the web that promises improved information management and search capability[61]. Representing and sharing knowledge using ontologies is simplified by availability of the standard web ontology language (OWL; http://www.w3.org/TR/owl-features/). Tools to edit OWL, such as Protégé[63], have been developed by the semantic web community and adopted in the life sciences. Implementing BioPAX using OWL enables both the ontology and the individuals and values to be stored in the same XML-based format, which makes data transmission easier. Using OWL also enables BioPAX users to take advantage of existing software tools for editing, transmitting, querying, reasoning about and visualizing OWL data.

An ontology is composed of classes, properties (representing relations) and restrictions and is used to define individuals (instances of classes, also known as objects) and values for their properties. Classes (also known as concepts or types) are often arranged into a hierarchy (or taxonomy) where child classes are more specific than, and inherit the properties of, parent classes. For example, in BioPAX, the BiochemicalReaction class is a subclass of the Conversion class. Classes may have properties (also known as fields, attributes or slots), which express possible relations to other classes (that is, they may have values of specific types). For example, a SmallMolecule is related to the ChemicalStructure class by the property structure. Restrictions (also known as constraints) define allowable values and connections within an ontology. For example, molecularWeight must be a positive number. Individuals are instances of classes where values occupy the properties of those instances. BioPAX defines the classes, properties and restrictions required to represent biological pathways and leaves creation of the individuals to users (data providers and consumers).

Pathway abstractions frequently used in several pathway databases and software programs are supported as follows:

- Metabolic pathways are described using the 'enzyme, substrate, product' abstraction[28] where substrates and products of a biochemical reaction are often small molecules. An enzyme, often a protein, catalyzes the reaction, and inhibitors and activators can modulate the catalysis event. Metabolic pathways use BioPAX classes: PhysicalEntity, Conversion, Catalysis, Modulation, Pathway.
- Signaling pathways involve molecules and complexes participating in biochemical reactions, binding, transportation and catalysis events (**Fig. 3**)[5,9,29–31]. These pathways may also include descriptions of molecular states (such as cellular location, covalent and noncovalent modifications, as well as fragments of sequence cleaved from a precursor) and generic molecules (such as the family of homologous Wnt proteins). Signaling pathways use BioPAX classes: PhysicalEntity, Conversion, Control, Catalysis, Modulation, MolecularInteraction, Pathway.
- Gene regulatory networks involve transcription and translation events and their control[12,14]. Transcription, translation and other template-directed reactions involving DNA or RNA are captured in a 'template reaction' in BioPAX, which maps a template to its encoded products (e.g., DNA to mRNA). Multiple sequence regions on a single strand of the template, such as promoters, terminators, open reading frames, operons and various reaction machinery binding sites, are active in a template reaction. Transcription factors (generally proteins and complexes), microRNAs and other molecules, participate in a 'template reaction regulation' event. Gene regulatory networks use BioPAX classes: PhysicalEntity, TemplateReaction, TemplateReactionRegulation.
- Molecular interactions, notably protein-protein[32–36] and protein-DNA interactions[37], involve two or more 'physical entities'. BioPAX follows the standard representation scheme of the Proteomics Standards Initiative Molecular Interaction (PSI-MI) format[38]. Molecular interactions use BioPAX classes: PhysicalEntity, MolecularInteraction.
- Genetic interactions occur between two genes when the phenotypic consequence of perturbing both genes is different than expected given the phenotypes of each single gene perturbation[39]. BioPAX represents this as a pair of genes that participate in a 'genetic interaction' measured using an observed 'phenotype'. Genetic interactions use BioPAX classes: Gene, GeneticInteraction.

Metabolic-, signaling- and gene regulatory–pathway abstractions are process oriented. They imply a temporal order and can be thought of as extensions of the standard chemical reaction pathway notation to accommodate biological information. Molecular and genetic interactions, however, imply a static network of connections among system components, instead of the temporally ordered process of reactions that defines a metabolic or signaling pathway. BioPAX supports combining these different types of data into a single model that is useful to gain a more complete view of a cellular process.

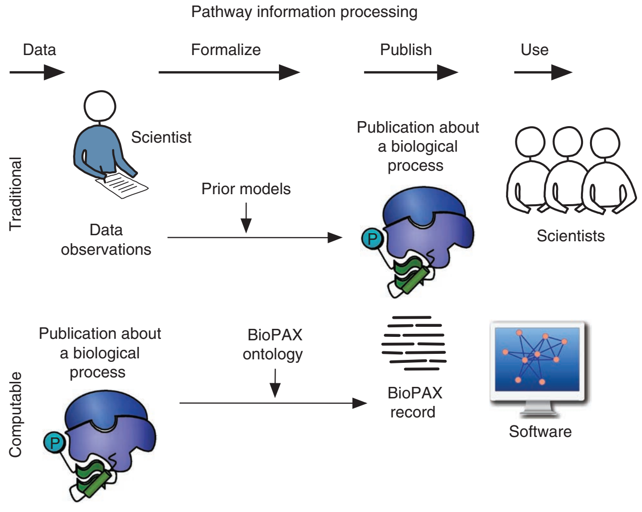


Figure 2 BioPAX enables computational data gathering, publication and use of information about biological processes. Traditional pathway information processing: observations considering prior models published as text and figures. Computable pathway information processing: scientist's description represented using formal, computable framework (ontology) published in a format readable by computer software for analysis by scientists.

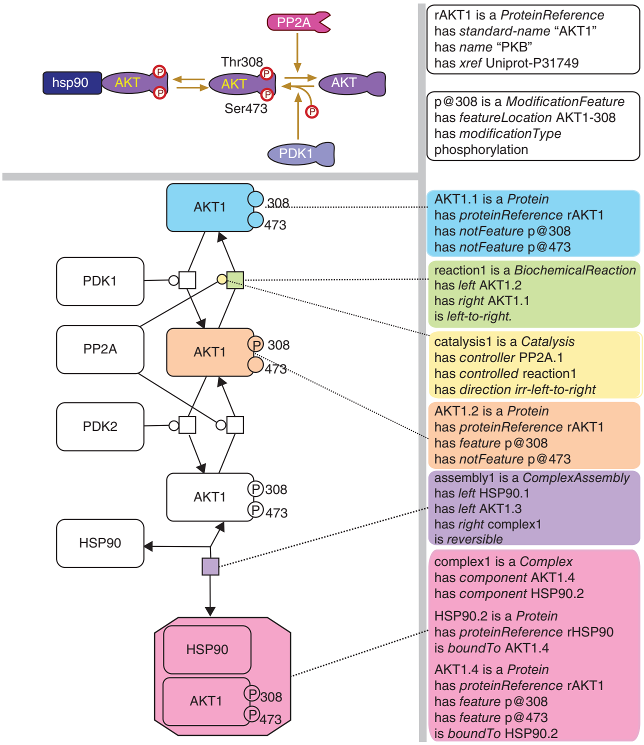


Figure 3 The AKT pathway as represented by a traditional method (top left; from http://www.biocarta.com/), a formalized SBGN diagram (left; from http://www.sbgn.org/[62]) and using the BioPAX language (right). An important advantage of the BioPAX representation is that it can be interpreted by computer software and used in multiple ways, including automatic diagram creation, information retrieval and analysis. Online documentation at http://www.biopax.org/ contains more details about how to represent diverse types of biological pathways. Actual samples of pathway data in BioPAX OWL XML format are available in **Supplementary Tables 1** and **2**.

BioPAX provides many additional constructs, not shown in **Figure 4**, that are used to store extra details, such as database cross-references, chemical structure, experimental forms of molecules, sequence feature locations and links to controlled vocabulary terms in other ontologies (**Supplementary Fig. 1**). BioPAX reuses a number of standard controlled vocabularies defined by other groups. For example, Gene Ontology[40] is used to describe cellular location, PSI-MI vocabularies[38] are used to define evidence codes, experimental forms, interaction types, relationship types and sequence modifications, and Sequence Ontology[41] is used to define types of sequence regions, such as a promoter region on DNA involved in transcription of a gene. Other useful controlled vocabularies can be referenced, such as the molecule role ontology[42].

BioPAX defines additional semantics that are currently only captured in documentation. For instance, physical entities represent pools of molecules and not individual molecules, corresponding to typical semantics used when describing pathways in textbooks or databases. A molecular pool is a set of molecules in a bounded area of the cell, thus it has a concentration. Pools can be heterogeneous and can overlap, as in the case of a protein existing in multiple phosphorylation states.

BioPAX also defines a range of constructs that are represented as ontology classes. Some of these represent biological entities, such as proteins, and are organized into classes that conceptualize the pathway knowledge domain. Others are used to represent annotations and properties of the database representation of biological entities. For instance, BioPAX provides 'xref' classes to represent different kinds of references to databases that can be useful for data integration. These are represented as subclasses of UtilityClass for convenience. A future version of BioPAX would ideally capture these semantics and structure these concepts more formally.

Uses of pathway data encoded in BioPAX

Once pathway data are translated into a standard computable language, such as BioPAX, it is easier for software to access them and thereby support browsing, retrieval, visualization and analysis (**Fig. 5**). This enables efficient reuse of data in different ways, avoiding the time-consuming and often frustrating task of translating them between formats (**Fig. 1**). Additionally, it enables uses that would be impractical without a standard format, such as those dependent on combining all available pathway data.

BioPAX can be used to help aggregate large pathway data sets by reducing the required collection and translation effort, for instance using software such as cPath[43]. Typical biological queries, such as 'What reactions involve my protein of interest?' generate more complete answers when querying these larger pathway data sets. Another frequent use is to find pathways that are active in a particular biological context, such as a cell state determined by a genome-scale molecular profile measurement. For instance, pathways with multiple differentially expressed genes may be transcriptionally active in one biological condition and not in another. Functional genomics and pathway data can be imported into software and combined for visualization and analysis to find interesting network regions. A typical workflow involves overlaying molecular profiling data, such as mRNA transcript profiles, on a network of interacting proteins to identify transcriptionally active network regions, which may represent active pathways[44]. A number of recent papers have used this pathway analysis workflow to highlight genes and pathways that are active in specific model organisms or diseased tissues, such as breast cancer, using gene and protein expression, copy number variants and single-nucleotide polymorphisms[19,44–49]. BioPAX has also been used in a number of these studies to collect and integrate large amounts of pathway information from multiple databases for analysis. For instance, protein expression data were combined with pathway information to highlight the importance of apoptosis in a mouse model of heart disease[50]. Multiple groups have found that tumor-associated mutations are significantly related by pathway

Table 1 What is included in BioPAX

Content	Description
Ontology specification	Web Ontology Language (OWL) XML file, developed using free Protégé ontology editor software[63].
Language documentation	Explanation of BioPAX entities, example documentation, best practice recommendations, use cases and instructions for carrying out frequently used technical tasks.
Example files	Example files for biochemical pathway, protein and genetic interaction, protein phosphorylation, insulin maturation, gene regulation and generic molecules in OWL XML.
Graphical representation	Recommendations for graphical representation using Systems Biology Graphical Notation (SBGN) as a guide.
Paxtools software	Java programming library supporting import/export, conversion and validation. Can be used to add BioPAX support to software.
List of data sources and supporting software	Databases making data available in BioPAX format, software systems for storing, visualizing and analyzing BioPAX pathways.

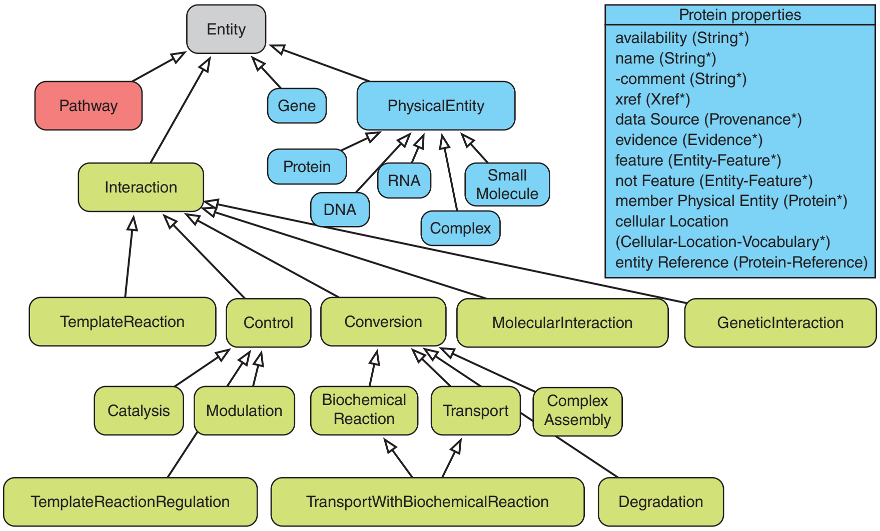


Figure 4 High-level view of the BioPAX ontology. Classes, shown as boxes and arrows, represent inheritance relationships. The three main types of classes in BioPAX are Pathway (red), Interaction (green) and PhysicalEntity and Gene (blue). For brevity, the properties of the Protein class only are shown as an example at the top right. Asterisks indicate that multiple values for the property are allowed. Refer to BioPAX documentation at http://www.biopax.org/ for full details of all classes and properties.

information[47,48]. And recently, in a study of rare copy number variants in 996 individuals with autism spectrum disorder, a core set of neuronal development–related pathways were found to link dozens of rare mutations to autism that were not significantly linked to the disorder on their own by traditional single-gene association statistics[49]. These studies highlight the importance of pathway information in explaining the functional consequence of mutations in human disease. BioPAX pathway data can also be converted into simulation models, for instance using differential equations[51] or rule-based modeling languages[52], to predict how a biological system may function after a gene is knocked out.

BioPAX is useful for exchanging information among and between data providers and analysis software. Pathway database groups can share the effort of pathway curation by making their pathways available in BioPAX format and exchanging them with others. For example, pathways in BioPAX format from the Reactome[8] database are imported by the US National Cancer Institute/Nature Pathway Information Database[9]. Data providers can use existing BioPAX-enabled software to add useful new features to their systems. For example, the Cytoscape network visualization software[20] can read and display BioPAX-formatted data as a network. The Reactome group used this feature to create a pathway visualization tool for their website. Because Reactome data were available in BioPAX format, and Cytoscape could already read BioPAX format, this new feature was easy to implement.

The Paxtools Java programming library for BioPAX has been developed to help software developers readily support the import, export and validation of BioPAX-formatted data for various uses in their software (http://www.biopax.org/paxtools/). Using Paxtools and other tools, a range of BioPAX-compatible software has been developed, including browsers, visualizers, querying engines, editors and converters (**Supplementary Table 4**). For instance, the ChiBE and VisANT pathway-visualization tools read BioPAX format[22], and the WikiPathways website[53], a community wiki for pathways, is working on using BioPAX to help import pathways from several sources, including manually edited pathways from biologists. The Pathway Tools software[21] and CellDesigner pathway editor[54] are developing support for BioPAX-based data exchange. In addition, tools for the storage and querying of Resource Description Framework (http://www.w3.org/RDF/) data sets, generated within the Semantic Web community, can be used to effectively process BioPAX data.

What is not covered?

The BioPAX language uses a discrete representation of biological pathways. Dynamic and quantitative aspects of biological processes, including temporal aspects of feedback loops and calcium waves, are not supported. However, BioPAX addresses this need by coordinating work (as described below) with the SBML and CellML mathematical modeling language communities[55,56] and a growing software tool set supporting biological process simulation[57]. Detailed information about experimental evidence supporting elements of a pathway map is useful for evaluating the quality of pathway data. This information is only included in BioPAX for molecular interactions, because that was already defined by the PSI-MI language[58] and it was reused The BioPAX work group makes use of PSI-MI–controlled vocabularies and other concepts and works with the PSI-MI work group to build these vocabularies in areas of shared interest, such as genetic interactions. Although BioPAX does not aim to standardize how pathways are visualized, work is coordinated with the

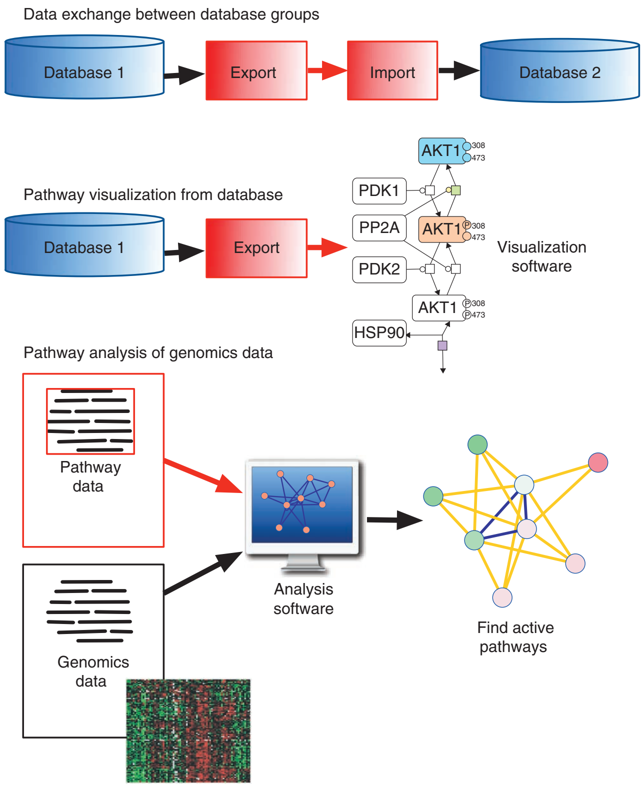


Figure 5 Example uses of pathway information in BioPAX format. Red-colored boxes or lines indicate the use of BioPAX.

Systems Biology Graphical Notation (SBGN; http://sbgn.org/) community, via members of both communities who attend BioPAX and SBGN meetings, to ensure that SBGN can be used to visualize BioPAX pathways. Currently, most BioPAX concepts can be visualized using SBGN process description and SBGN activity flow diagrams and a mapping of BioPAX to SBGN entity relationship diagrams is under development. BioPAX development is coordinated with the above standardization efforts through regular communication between workgroups to ensure complementarity and compatibility. For instance, controlled vocabularies developed by PSI-MI and BioPAX can be used to annotate SBML and CellML models (**Fig. 6**). BioPAX aims to be compatible with these and other efforts, so that pathway data can be transformed between alternative representations when needed. PSI-MI to BioPAX and SBML to BioPAX converters are available (**Supplementary Table 4**).

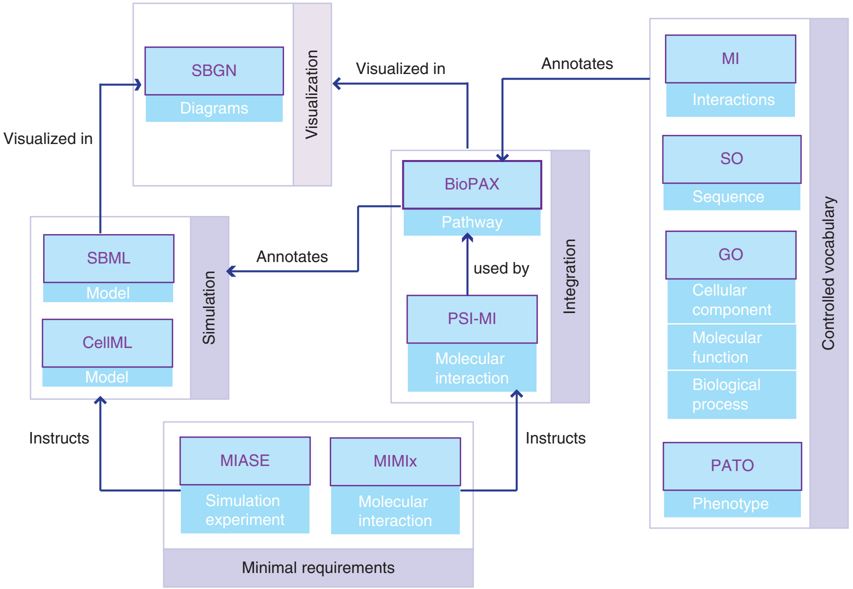


Figure 6 The relationship among popular standard formats for pathway information. BioPAX and PSI-MI are designed for data exchange to and from databases and pathway and network data integration. SBML and CellML are designed to support mathematical simulations of biological systems and SBGN represents pathway diagrams.

How does the BioPAX community work?

Whereas BioPAX facilitates communication of current knowledge, it is challenging for all knowledge-representation efforts to anticipate new forms of information. As new types of pathway data and new knowledge representation languages and tools become available, the BioPAX language must evolve through the efforts of a community of scientists that includes biologists and computer scientists.

BioPAX is developed through community consensus among data providers, tool developers and pathway data users. More than 15 BioPAX workshops have been held since November 2002, attended by a diverse set of participants. Incremental versions, also called levels, of the BioPAX language were progressively developed at these workshops to focus the group's efforts on attainable intermediate goals. Broader input came from mailing lists and a community wiki. Community members participated in developing functionality they were interested in, which was integrated into specific levels (**Supplementary Table 5**). Level 1 supports metabolic pathways. Level 2 adds support for molecular interactions and post-translational protein modifications by integrating data structures from the PSI-MI format. Level 3 adds support for signaling pathways, molecular state, gene regulation and genetic interactions (**Supplementary Table 3**). It is anticipated that newer BioPAX levels replace older ones, so use of the most recent BioPAX level 3 is currently recommended. To ease the burden on users and developers, BioPAX aims to be backwards compatible where practical. Level 2 is backwards compatible with level 1; however, level 3 involved a major redesign that necessitated breaking backwards compatibility. This said, many core classes have remained the same in levels 1, 2 and 3, and software is provided for updating older BioPAX pathways to level 3 (via Paxtools). All BioPAX material (**Table 1**) is made freely available under open source licenses through a central website (http://www.biopax.org/) to encourage broad adoption. The database and tool support (**Supplementary Table 4**) of a common language aids the creation, analysis, visualization and interpretation of integrated pathway maps.

In addition to the creation of a shared language for data and software, the process of achieving community consensus spurs innovation in the field of pathway informatics. Community discussion helps resolve technical knowledge representation issues faced by many data providers and users and facilitates the convergence to common terminology and representation. Solutions are discovered in independent research groups and incorporated in new data models and community best practices, which then enable identification of new issues. Thus, community workshops support a positive feedback cycle of knowledge sharing that has led to an accepted BioPAX language and development of better software and databases. We expect this to continue and to support new scientific uses of pathway information, motivated by end-user access to valuable integrated pathway information and efficiency gain for database and software development groups. This will especially benefit new pathway databases and software tools that adopt standard representation and software components from the start.

Future community goals

The BioPAX shared language is a starting point on the path to developing complete maps of cellular processes. Additional near and long-term goals remain to be realized to enable effective integration and use of biological pathway information, as described below.

Data collection. Data must be collected and translated to a standard format for them to be integrated. This process is underway, as the descriptions of millions of interactions in thousands of pathways across many organisms from multiple databases are now available in BioPAX format. However, vast amounts of pathway data remain difficult to access in the literature and in databases that don't yet support standard formats. Increasing use of standards requires promoting and supporting data curation teams and automating more of the data collection process using software. Easy-to-use tools for tasks like pathway editing must also be developed so that biologists can share their data in BioPAX format without substantial resource investment. Ideally, appropriate software would allow authors to enter data directly in standard formats during the publication process, to facilitate annotation and normalization by curators before incorporation into databases for use by researchers[53].

Validation and best practice development. To aid data collection, major data providers and others must develop community best practice guidelines and rules to help diverse groups use BioPAX consistently when multiple ways of encoding the same information exist. This will enable data providers to benefit from automatic syntactic and semantic validation of their data so they can ensure they are sharing data using standard representation and best practices[59,60]. Data collection and automatic validation will facilitate convergence to generally accepted biological process models.

Semantic integration. Several models of the same biological process may usefully co-exist. Ideally, different models could be compared for analysis and hypothesis formulation. Even so, comparison is difficult because the same concept can be represented in several ways owing to use of multiple levels of abstraction (such as the hRas protein versus the Ras protein family), use of different controlled vocabularies, data incompleteness or errors. Future research needs to develop semantic integration solutions that recognize and aid resolution of conflicts.

Visualization. Pathway diagrams are highly useful for communicating pathway information, but it is challenging to automatically construct these diagrams in a biologically intuitive way from pathway data stored in BioPAX. The SBGN pathway diagram standardization effort provides a starting point toward achieving this goal (**Fig. 3**). Intuitive and automatically drawn biological network visualizations may one day replace printed biology textbooks as the primary resource for knowledge about cellular processes.

Language evolution. As uses of pathway information and technology evolve, so must the BioPAX language. For instance, future BioPAX levels should capture cell-cell interactions, be better at describing pathways where sub-processes are not known or need not be represented, more closely integrate third-party controlled vocabularies and ontologies to ease their use and better encode semantics for easier data validation and reasoning.

Many groups within the BioPAX community, including most pathway data providers and tool developers, are working to achieve the above goals. For instance, Pathway Commons (http://www.pathwaycommons.org/) aims to be a convenient single point of access for all publicly accessible pathway information and the WikiPathways project (http://www.wikipathways.org/) seeks to enable pathway curation by individuals[53]. Also, the semantic web community is developing a set of technologies that promise to ease the integration of information dispersed on the World Wide Web[61]. These technologies will aid pathway data integration because BioPAX is compatible with them through use of the W3C standard Web Ontology Language, OWL. All of the above research and development activities support the vision of data providers sharing computable maps of biological processes in a standard format for convenient use by a community of pathway researchers.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

Funded by the US Department of Energy workshop grant DE-FG02-04ER63931, the caBIG program, the US National Institute of General Medical Sciences workshop grant 1R13GM076939, grant P41HG004118 from the US National Human Genome Research Institute and Genome Canada through the Ontario Genomics Institute (2007-OGI-TD-05) and US National Institutes of Health grant R01GM071962-07. Thanks to many people who contributed to discussions on BioPAX mailing lists, at conferences and at BioPAX workshops, especially A. Ruttenberg and J. Rees.

AUTHOR CONTRIBUTIONS

All authors helped develop the BioPAX language, ontology, documentation and examples by participating in workshops or on mailing lists and/or provided data in BioPAX format and/or wrote software that supports BioPAX. See **Supplementary Table 5** for a full list of author contributions.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Gasteiger, E. *et al.* ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* **31**, 3784–3788 (2003).
2. Nicholson, D.E. The evolution of the IUBMB-Nicholson maps. *IUBMB Life* **50**, 341–344 (2000).
3. Demir, E. *et al.* PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. *Bioinformatics* **18**, 996–1003 (2002).
4. Krull, M. *et al.* TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Res.* **34**, D546–D551 (2006).
5. Fukuda, K. & Takagi, T. Knowledge representation of signal transduction pathways. *Bioinformatics* **17**, 829–837 (2001).
6. Davidson, E.H. *et al.* A genomic regulatory network for development. *Science* **295**, 1669–1678 (2002).
7. Kohn, K.W. Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol. Biol. Cell* **10**, 2703–2734 (1999).
8. Matthews, L. *et al.* Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* **37**, D619–D622 (2009).
9. Schaefer, C.F. *et al.* PID: the Pathway Interaction Database. *Nucleic Acids Res.* **37**, D674–D679 (2009).
10. Bader, G.D. & Hogue, C.W. BIND—a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics* **16**, 465–477 (2000).
11. Kitano, H. A graphical notation for biochemical networks. *BIOSILICO* **1**, 169–176 (2003).
12. Gama-Castro, S. *et al.* RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.* **36**, D120–D124 (2008).
13. Mi, H. *et al.* The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.* **33**, D284–D288 (2005).
14. Keseler, I.M. *et al.* EcoCyc: a comprehensive view of *Escherichia coli* biology. *Nucleic Acids Res.* **37**, D464–D470 (2009).
15. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **38**, D473–D479 (2010).
16. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32** Database issue, D277–280 (2004).
17. Bader, G.D., Cary, M.P. & Sander, C. Pathguide: a pathway resource list. *Nucleic Acids Res.* **34**, D504–D506 (2006).
18. Huang, W., Sherman, B.T. & Lempicki, R.A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
19. Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D. & Ideker, T. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* **3**, 140 (2007).
20. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
21. Karp, P.D. *et al.* Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief. Bioinform.* **11**, 40–79 (2010).
22. Hu, Z. *et al.* VisANT 3.0: new modules for pathway visualization, editing, prediction and construction. *Nucleic Acids Res.* **35**, W625–W632 (2007).
23. Hoffmann, R. *et al.* Text mining for metabolic pathways, signaling cascades, and protein networks. *Sci. STKE* **2005**, pe21 (2005).
24. Racunas, S.A., Shah, N.H., Albert, I. & Fedoroff, N.V. HyBrow: a prototype system for computer-aided hypothesis evaluation. *Bioinformatics* **20** Suppl 1, i257–i264 (2004).
25. Cary, M.P., Bader, G.D. & Sander, C. Pathway information for systems biology. *FEBS Lett.* **579**, 1815–1820 (2005).
26. Vivanco, I. & Sawyers, C.L. The phosphatidylinositol 3-Kinase AKT pathway in human cancer. *Nat. Rev. Cancer* **2**, 489–501 (2002).
27. Koh, G., Teong, H.F., Clement, M.V., Hsu, D. & Thiagarajan, P.S. A decompositional approach to parameter estimation in pathway modeling: a case study of the Akt and MAPK pathways and their crosstalk. *Bioinformatics* **22**, e271–e280 (2006).
28. Karp, P.D. An ontology for biological function based on molecular interactions. *Bioinformatics* **16**, 269–285 (2000).
29. Joshi-Tope, G. *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* **33** Database issue, D428–D432 (2005).
30. Mi, H., Guo, N., Kejariwal, A. & Thomas, P.D. PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Res.* **35**, D247–D252 (2007).
31. Demir, E. *et al.* An ontology for collaborative construction and analysis of cellular pathways. *Bioinformatics* **20**, 349–356 (2004).

32. Bader, G.D., Betel, D. & Hogue, C.W. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* **31**, 248–250 (2003).
33. Salwinski, L. *et al.* The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* **32**, D449–D451 (2004).
34. Chatr-aryamontri, A. *et al.* MINT: the Molecular INTeraction database. *Nucleic Acids Res.* **35**, D572–D574 (2007).
35. Kerrien, S. *et al.* IntAct—open source resource for molecular interaction data. *Nucleic Acids Res.* **35**, D561–D565 (2007).
36. Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34**, D535–D539 (2006).
37. Matys, V. *et al.* TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110 (2006).
38. Kerrien, S. *et al.* Broadening the horizon—level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.* **5**, 44 (2007).
39. Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327**, 425–431 (2010).
40. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
41. Eilbeck, K. *et al.* The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.* **6**, R44 (2005).
42. Yamamoto, S., Asanuma, T., Takagi, T. & Fukuda, K.I. The molecule role ontology: an ontology for annotation of signal transduction pathway molecules in the scientific literature. *Comp. Funct. Genomics* **5**, 528–536 (2004).
43. Cerami, E.G., Bader, G.D., Gross, B.E. & Sander, C. cPath: open source software for collecting, storing, and querying biological pathways. *BMC Bioinformatics* **7**, 497 (2006).
44. Cline, M.S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* **2**, 2366–2382 (2007).
45. Efroni, S., Carmel, L., Schaefer, C.G. & Buetow, K.H. Superposition of transcriptional behaviors determines gene state. *PLoS ONE* **3**, e2901 (2008).
46. Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A.F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18** Suppl 1, S233–S240 (2002).
47. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
48. Wu, G., Feng, X. & Stein, L. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol.* **11**, R53 (2010).
49. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368–372 (2010).
50. Isserlin, R. *et al.* Pathway analysis of dilated cardiomyopathy using global proteomic profiling and enrichment maps. *Proteomics* **10**, 1316–1327 (2010).
51. Moraru, I.I. *et al.* Virtual Cell modelling and simulation software environment. *IET Syst. Biol.* **2**, 352–362 (2008).
52. Hlavacek, W.S. *et al.* Rules for modeling signal-transduction systems. *Sci. STKE* **2006**, re6 (2006).
53. Pico, A.R. *et al.* WikiPathways: pathway editing for the people. *PLoS Biol.* **6**, e184 (2008).
54. Kitano, H., Funahashi, A., Matsuoka, Y. & Oda, K. Using process diagrams for the graphical representation of biological networks. *Nat. Biotechnol.* **23**, 961–966 (2005).
55. Lloyd, C.M., Halstead, M.D. & Nielsen, P.F. CellML: its future, present and past. *Prog. Biophys. Mol. Biol.* **85**, 433–450 (2004).
56. Hucka, M. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524–531 (2003).
57. Sauro, H.M. *et al.* Next generation simulation tools: the Systems Biology Workbench and BioSPICE integration. *OMICS* **7**, 355–372 (2003).
58. Hermjakob, H. *et al.* The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat. Biotechnol.* **22**, 177–183 (2004).
59. Racunas, S.A., Shah, N.H. & Fedoroff, N.V. A case study in pathway knowledgebase verification. *BMC Bioinformatics* **7**, 196 (2006).
60. Laibe, C. & Le Novere, N. MIRIAM Resources: tools to generate and resolve robust cross-references in Systems Biology. *BMC Syst. Biol.* **1**, 58 (2007).
61. Berners-Lee, T. & Hendler, J. Publishing on the semantic web. *Nature* **410**, 1023–1024 (2001).
62. Le Novere, N. *et al.* The Systems Biology Graphical Notation. *Nat. Biotechnol.* **27**, 735–741 (2009).
63. Knublauch, H., Fergerson, R.W., Noy, N.F. & Musen, M.A. The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. in *The Semantic Web–ISWC 2004: Third International Semantic Web Conference, Hiroshima, Japan, November 7-11, 2004: Proceedings* (eds. McIlraith, S.A., Dimitris Plexousakis, D. & van Harmelen, F.) 229—243 (Springer, 2004).
64. Sowa, J.F. *Knowledge Representation: Logical, Philosophical, and Computational Foundations* (Brooks/Cole, 2000).
65. Wheeler, D.L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **35**, D5–D12 (2007).

[1]Computational Biology, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [2]Center for Bioinformatics and Computer Engineering Department, Bilkent University, Ankara, Turkey. [3]SRI International, Menlo Park, California, USA. [4]Institute for Bioinformatics Research and Development, Japan Science and Technology Agency, Tokyo, Japan. [5]Université libre de Bruxelles, Bruxelles, Belgium. [6]European Bioinformatics Institute, Hinxton, Cambridge, UK. [7]Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [8]NYU School of Medicine, New York, New York, USA. [9]National Cancer Institute, Center for Biomedical Informatics and Information Technology, Rockville, Maryland, USA. [10]Predictive Medicine, Belmont, Massachusetts, USA. [11]BIOBASE Corporation, Beverly, Massachusetts, USA. [12]Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico. [13]Biomolecular Systems Laboratory, Boston University, Boston, Massachusetts, USA. [14]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA. [15]McKusick-Nathans Institute of Genetic Medicine and the Departments of Biological Chemistry, Pathology and Oncology, Johns Hopkins University, Baltimore, Maryland, USA. [16]Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico. [17]Artificial Intelligence Center, SRI International, Menlo Park California, USA. [18]No affiliation declared. [19]Donnelly Center for Cellular and Biomolecular Research, Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada. [20]Faculté de Médecine, Université Rennes 1, Rennes, France. [21]Rothamsted Research, Harpenden, UK. [22]Cell Signaling Technology, Inc., Danvers, Massachusetts, USA. [23]Broad Institute, Cambridge, Massachusetts, USA. [24]Center for Food Safety and Applied Nutrition, US Food and Drug Administration, Laurel, Maryland, USA. [25]Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA. [26]Neurobiology, Neurodegeneration and Repair Laboratory, National Eye Institute, National Institutes of Health, Bethesda, Maryland, USA. [27]Department of Behavioral Neuroscience. Oregon Health & Science University, Portland, Oregon, USA. [28]US Environmental Protection Agency Durham, North Carolina, USA. [29]Mathematics & Computer Science Division, Argonne National Laboratory, Argonne, Illinois, USA. [30]University of Connecticut Health Center, Farmington, Connecticut, USA. [31]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts, USA. [32]Lexikos Corporation, Boston, Massachusetts, USA. [33]Biotechnology Division, National Institute of Standards and Technology, Gaithersburg, Maryland, USA. [34]Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda Maryland, USA. [35]Unilever Centre for Molecular Sciences Informatics, Department of Chemistry, University of Cambridge, Cambridge, UK. [36]Clinical Semantics Group, Lexington, Massachusetts, USA. [37]Center for Cell Analysis and Modeling, University of Connecticut Health Center, Storrs, Connecticut, USA. [38]Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland. [39]Konrad Lorenz Institute for Evolution and Cognition Research, Altenberg, Austria. [40]Department of Bioinformatics, Maastricht University, Maastricht, The Netherlands. [41]University of Auckland, Auckland, New Zealand. [42]Syngenta Biotech Inc., Research Triangle Park, North Carolina, USA. [43]Department of Genetics, Stanford University, Stanford, California, USA. [44]Yale Center for Medical Informatics, Yale University, New Haven, Connecticut, USA. [45]Loyola Marymount University, Los Angeles, California, USA. [46]Physiomics PLC, Magdalen Centre, Oxford Science Park, Oxford, UK. [47]St. John's University, Jamaica, New York, USA. [48]Columbia University, New York, New York, USA. [49]SRA International, Fairfax, Virginia, USA. [50]Novartis Knowledge Center, Cambridge, Massachusetts, USA. [51]University of Ottawa, Ottawa, Ontario, Canada. [52]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA. [53]Vertex Pharmaceuticals, Cambridge, Massachusetts, USA. [54]Human and Molecular Genetics Center, Medical College of Wisconsin, Milwaukee, Wisconsin, USA. [55]Gladstone Institute of Cardiovascular Disease, San Francisco, California, USA. [56]Department of Plant Breeding and Genetics, Cornell University, Ithaca, New York, USA. [57]Centre for Biomedical Informatics, School of Medicine, Stanford University, Stanford, California, USA. [58]Computational Sciences, Informatics, Millennium Pharmaceuticals Inc., Cambridge, Massachusetts, USA. [59]Center for Biomedical Informatics and Information Technology, National Cancer Institute, Bethesda, Maryland, USA. [60]Institute for Genomics and Systems Biology, The University of Chicago and Argonne National Laboratory, Chicago, Illinois, USA. [61]Total Gas & Power, Paris, France. [62]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto, Japan. [63]Biological Network Modeling Center, California Institute of Technology, Pasadena, California, USA. [64]Department of Bioinformatics, Göttingen, Germany. Correspondence should be addressed to G.D.B. (biopax-paper@biopax.org).

nature biotechnology

Deconvolution of complex G protein–coupled receptor signaling in live cells using dynamic mass redistribution measurements

Ralf Schröder[1,5], Nicole Janssen[2,5], Johannes Schmidt[1], Anna Kebig[2], Nicole Merten[1], Stephanie Hennen[1], Anke Müller[1], Stefanie Blättermann[1], Marion Mohr-Andrä[2], Sabine Zahn[3], Jörg Wenzel[3], Nicola J Smith[4], Jesús Gomeza[1], Christel Drewke[1], Graeme Milligan[4], Klaus Mohr[2] & Evi Kostenis[1]

Label-free biosensor technology based on dynamic mass redistribution (DMR) of cellular constituents promises to translate GPCR signaling into complex optical 'fingerprints' in real time in living cells. Here we present a strategy to map cellular mechanisms that define label-free responses, and we compare DMR technology with traditional second-messenger assays that are currently the state of the art in GPCR drug discovery. The holistic nature of DMR measurements enabled us to (i) probe GPCR functionality along all four G-protein signaling pathways, something presently beyond reach of most other assay platforms; (ii) dissect complex GPCR signaling patterns even in primary human cells with unprecedented accuracy; (iii) define heterotrimeric G proteins as triggers for the complex optical fingerprints; and (iv) disclose previously undetected features of GPCR behavior. Our results suggest that DMR technology will have a substantial impact on systems biology and systems pharmacology as well as for the discovery of drugs with novel mechanisms.

G protein–coupled receptors (GPCRs) are among the most important drug target classes[1]. For many members of this receptor family, it is now well established that they oscillate among multiple conformations that can be differentially stabilized by ligands, thus permitting access to only a subset of the complete repertoire of receptor behaviors[2–8]. This phenomenon, also referred to as biased agonism or functional selectivity, has important implications for GPCR-related drug discovery because it raises the possibility to design signaling pathway–specific therapeutics.

Activation of downstream signaling events of GPCRs has been traditionally recorded with assays based on quantification of distinct intracellular second messengers[5,9–11] and/or translocation of β-arrestin proteins[5,12–16]. It is becoming increasingly apparent, however, that integrated cellular responses—rather than individual components of signaling pathways—need to be analyzed, because different classes of GPCRs typically produce one or more specific second messengers and because GPCRs may engage additional non–G protein effectors[17–21]. An optical biosensor technology, based on measurement of the cellular process of dynamic mass redistribution, was recently developed to monitor such integrated signaling responses. In DMR technology, polarized light is passed through the bottom of a biosensor microtiter plate containing the cell samples, and a shift in wavelength of reflected light is indicative of redistribution of cellular constituents triggered upon receptor activation (**Fig. 1a**)[5,22,23]. The wavelength shift may vary in magnitude, direction (positive or negative) and over time depending on how different activated signaling pathways cause various intracellular molecules to relocate.

DMR technology enables GPCR function to be analyzed without labeling. In the assay, receptor activity is measured as an optical trace that represents the generic response of living cells, reminiscent of the holistic responses obtained in tissue or organ bath experiments. Label-free technologies could therefore be decisively advantageous to monitor even complex signaling processes, particularly in primary cells, which are difficult to analyze with traditional biochemical methods and which are difficult to transfect with labeled components of the GPCR signaling cascade for optical studies[23–25]. It is likely that these obstacles have so far precluded assessment of drug candidates in their native environment.

Although label-free recording of DMR is already being applied in pharmaceutical companies on a more empirical basis to assess feasibility for high-throughput screening or for pharmacological ligand profiling[25–29], no in-depth analytical study had been conducted that compared this technology platform with the methods traditionally used in early drug discovery.

Therefore, we applied DMR to monitor signaling of a number of GPCRs from all four coupling classes (G_i/G_o, G_s, G_q and G_{12}/G_{13}) and compared the receptor's functionality for inducing a whole-cell response with the more classical biochemical approaches to define GPCR signal transduction. An experimental strategy is presented to identify the post-receptor trigger underlying the optical response profiles, thereby allowing optical traces to be precisely assigned to distinct GPCR-mediated signaling pathways. We then take advantage

[1]Molecular, Cellular, and Pharmacobiology Section, Institute of Pharmaceutical Biology, University of Bonn, Bonn, Germany. [2]Pharmacology and Toxicology Section, Institute of Pharmacy, University of Bonn, Bonn, Germany. [3]Department of Dermatology, University of Bonn, Bonn, Germany. [4]Molecular Pharmacology Group, Neuroscience and Molecular Pharmacology, Faculty of Biomedical and Life Sciences, University of Glasgow, Glasgow, UK. [5]These authors contributed equally to this work. Correspondence should be addressed to E.K. (kostenis@uni-bonn.de) or K.M. (k.mohr@uni-bonn.de).

Published online 15 August 2010; doi:10.1038/nbt.1671

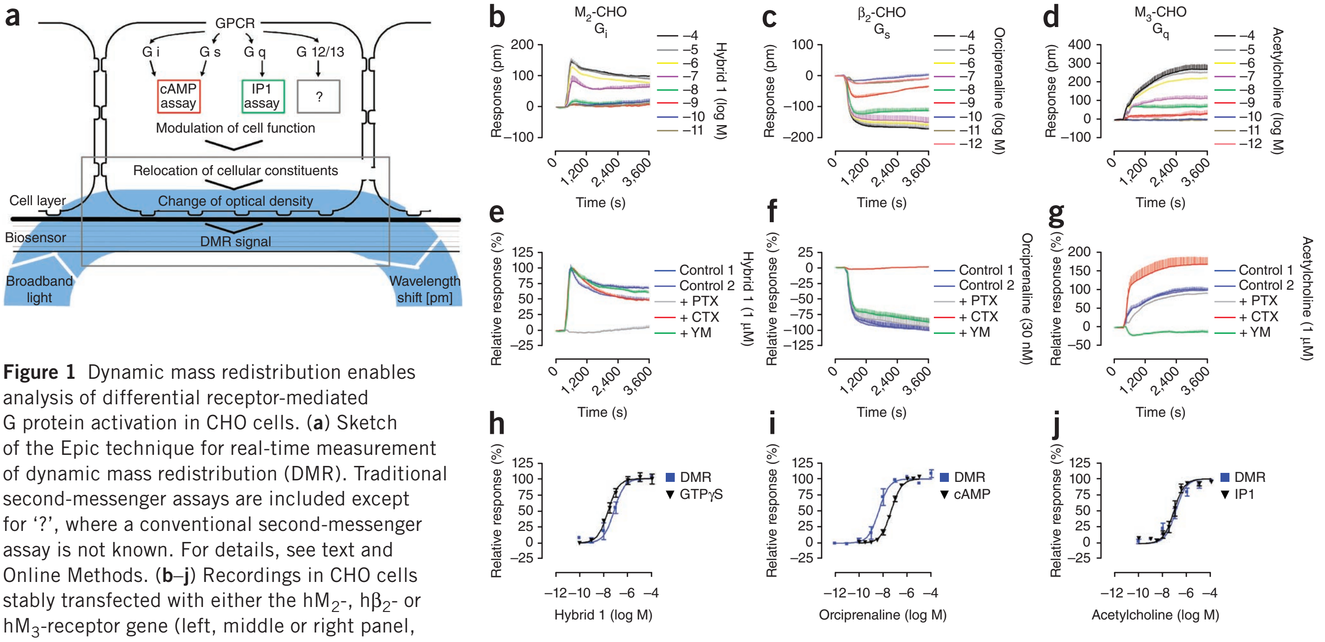


Figure 1 Dynamic mass redistribution enables analysis of differential receptor-mediated G protein activation in CHO cells. (**a**) Sketch of the Epic technique for real-time measurement of dynamic mass redistribution (DMR). Traditional second-messenger assays are included except for '?', where a conventional second-messenger assay is not known. For details, see text and Online Methods. (**b**–**j**) Recordings in CHO cells stably transfected with either the hM$_2$-, hβ$_2$- or hM$_3$-receptor gene (left, middle or right panel, respectively). (**b**–**d**) The cell lines were challenged with the indicated concentrations of the corresponding agonist (M$_2$, hybrid 1; β$_2$, orciprenaline; M$_3$, acetylcholine), and wavelength shift was monitored as a measure of receptor activation. Representative data (mean + s.e.m.) of at least four independent experiments. (**e**–**g**) Validation by toxin pretreatment of agonist-induced G$_{i,}$ G$_s$ and G$_q$ activation: pertussis toxin (100 ng/ml; PTX), cholera toxin (100 ng/ml; CTX), YM 254890 (300 nM; YM). Representative data (mean + s.e.m.) from at least three independent experiments. Control 1 corresponds to YM signature and control 2 to PTX and CTX signatures. (**h**–**j**) Comparison of concentration-effect curves resulting from DMR assays and traditional assays. Calculated log EC$_{50}$ values are: (**h**) M$_2$ DMR: −7.21 ± 0.27, GTPγS: −7.63 ± 0.07; (**i**) β$_2$ DMR: −8.35 ± 0.24, cAMP: −7.34 ± 0.03; (**j**) M$_3$ DMR: −7.02 ± 0.08, IP1: −7.02 ± 0.07. All data are means (±s.e.m.) of at least three independent experiments.

of this strategy to explore complex GPCR signaling patterns in both recombinant and primary human cells. Notably, our results provide evidence that simultaneous visualization of signaling pathways by DMR, but not by recording of defined downstream signaling events in single component functional assays, enables unexpected signaling phenomena to be identified, thus implying the need to shift from single-component to system analysis. We suggest that optical recording of DMR represents an enabling technology with substantial impact on both the dissection of complex biological signaling patterns of GPCRs in basic research and the understanding of mechanisms of drug action in GPCR drug discovery.

RESULTS

DMR reports signaling of G$_i$/G$_o$-, G$_s$- and G$_q$-linked receptors

To establish that DMR measures signaling downstream of different G protein classes, CHO cells stably transfected to express the G$_i$/G$_o$-sensitive muscarinic receptor M$_2$, the G$_q$-linked muscarinic receptor M$_3$, or the G$_s$-sensitive adrenergic β$_2$ receptor were challenged with increasing concentrations of their respective agonists, and DMR was recorded as a function of receptor activity. For all three receptors, real-time optical recordings were concentration dependent and varied depending on the primary signaling pathway of each receptor (**Fig. 1b–d**). Ligand activity was undetectable in native CHO cells, demonstrating that the DMR traces required the presence of the respective GPCRs (**Supplementary Fig. 1**).

To further establish whether heterotrimeric G proteins are responsible for orchestrating the observed temporal response patterns, we chose to pharmacologically silence the G protein signaling pathways using pertussis toxin (PTX) to block G$_i$/G$_o$ signaling, YM-254890 (hereafter referred to as YM) to suppress G$_q$ signaling (ref. 30), and cholera toxin (CTX) to mask G$_s$ signaling. Indeed, M$_2$ receptor traces were completely abrogated by PTX but unaffected by YM and CTX (**Fig. 1e**), identifying G$_i$/G$_o$ proteins as upstream triggers for this optical fingerprint. On their own, PTX, YM or CTX did not induce a DMR response (**Supplementary Fig. 2**). Furthermore, G protein activation, as reflected by GTPγS binding assays, was in good agreement with the DMR data (**Fig. 1**), thus supporting the notion that the optical traces resulted from a G$_i$-mediated signaling event. Corresponding observations were made for G$_s$- and G$_q$-DMR assays: the G$_s$ signatures of the β$_2$ receptor were exclusively masked by CTX but not by PTX or YM (**Fig. 1f**), and the G$_q$ signatures of the M$_3$ receptor were masked by YM but not by PTX or CTX (**Fig. 1g** and **Supplementary Fig. 3**). Again, traditional second-messenger assays suggested that optical traces are a consequence of engaging the respective signaling pathways assigned to both receptors (**Fig. 1i**,**j**).

DMR can identify signaling along the G$_{12}$/G$_{13}$ pathway

Whereas second-messenger assays are well suited to detect activation of G$_i$-, G$_s$- and G$_q$-sensitive receptors, such assays are not yet available to detect G$_{12}$/G$_{13}$ signaling, apart from high-content screening or approaches that assume contribution of these G proteins by recording mostly far removed downstream events[31]. To demonstrate that DMR provides information about signaling through this fourth Gα protein family, we took advantage of the atypical cannabinoid receptor GPR55, which is the only GPCR known to date with exclusive bias toward the G$_{12}$/G$_{13}$ pathway[8,32,33]. Human embryonic kidney (HEK293) cells were chosen to establish a stable GPR55-expressing clone, because these cells were virtually unresponsive in DMR assays to lysophosphatidylinositol (LPI), currently the most suitable GPR55 agonist[32–34] (**Supplementary Fig. 4a**). Of note, wild-type CHO cells responded with robust

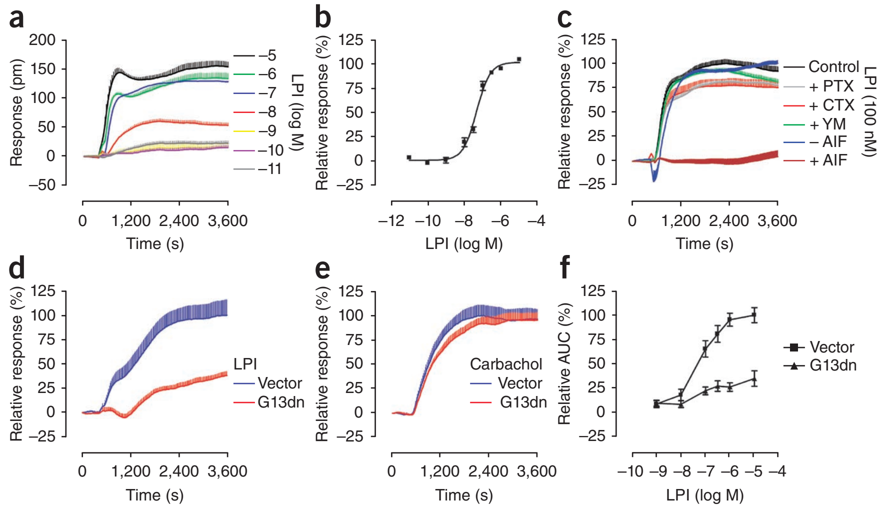


Figure 2 Dynamic mass redistribution visualizes signaling along the G_{12}/G_{13} pathway. (**a**) GPR55-HEK cells were challenged with the indicated concentrations of the GPR55 agonist lysophosphatidylinositol (LPI), and wavelength shift was monitored over time as a measure of receptor activation. Data shown are representative (mean + s.e.m.) of at least three independent experiments. (**b**) Concentration-effect curve for LPI in GPR55-HEK cells resulting from DMR traces in three independent experiments. The calculated log EC_{50} value is −7.34 ± 0.05. (**c**) LPI-mediated alteration of cell activity in GPR55-HEK cells is not blunted by pretreatment with toxin (5 ng/ml PTX, 100 ng/ml CTX) or pathway inhibitor (300 nM YM-254890), but is sensitive to preincubation of cells with 300 μM of the pan–G protein agonist AlF_4^- (AIF). Shown are representative data (mean + s.e.m.) from at least three independent experiments. (**d**–**f**) LPI- but not carbachol-mediated DMR is substantially diminished in GPR55-HEK cells cotransfected to express a dominant negative form of G13 (G13dn, G13Q226L,D294N). GPR55 cells cotransfected to express G13dn or empty pcDNA3.1 vector DNA were treated with 1 μM LPI (**d**) or 100 μM carbachol (**e**) and DMR was monitored over time. Depicted are representative optical traces (**d**,**e**) and concentration effect relationships of five such experiments (**f**).

DMR signals to LPI challenge, whereas the same cells did not show any significant LPI effect in traditional second-messenger assays covering the G_i, G_s and G_q pathways and were therefore judged unsuitable for transfection with and functional exploration of GPR55 (**Supplementary Fig. 4b–d**). GPR55-HEK cells displayed concentration-dependent optical traces upon exposure to LPI (**Fig. 2a**,**b**). Notably, GPR55-mediated DMR was insensitive to inhibition of G_i, G_s or G_q signaling by PTX, CTX or YM, but was silenced when cells were pretreated with the pan–G protein activator aluminum fluoride (AlF_4^-) (**Fig. 2c**). This effect could not be explained by a general blunting of cell responsiveness in DMR assays (**Supplementary Fig. 5**). These data suggest a G-protein origin of the GPR55 trace that is independent of coupling to G_i/G_o, G_s and G_q proteins, a conclusion that is further corroborated by the lack of second-messenger production in GPR55-HEK cells (**Supplementary Fig. 6**). Notably, GPR55-HEK cells transfected to coexpress dominant negative G13 (G13dn, G13Q226L,D294N) did not display altered GPR55 cell surface expression (**Supplementary Fig. 7**) but did show substantially diminished LPI-induced DMR responses, whereas optical traces elicited by carbachol to stimulate endogenously expressed G_q-sensitive muscarinic receptors, as a control, were virtually unaffected (**Fig. 2d–f**). Taken together, these data suggest that DMR is competent to visualize signaling along the G_{12}/G_{13} pathway and therefore represents a methodology applicable to probe functionality of GPCRs from all four coupling classes, which at present is beyond reach of most GPCR assay platforms.

DMR response profiles are cell type dependent

As signaling-dependent relocation of cellular constituents is likely to depend on cellular background, we also examined G_i/G_o-, G_q- and G_s-mediated DMR responses in HEK293 cells. The G_i-coupled prostaglandin D_2 receptor CRTH2 revealed a signature profile comparable to that observed for the G_i-coupled M_2 receptor in CHO cells (**Fig. 3a**; compare **Fig. 1b**). Similar DMR traces were also obtained when a panel of additional G_i/G_o-coupled receptors were analyzed (**Supplementary Fig. 8**), supporting the notion that optical traces may indeed be suggestive of engagement of particular signaling pathways. However, a positive DMR signal was observed when the G_s signaling cascade was activated in HEK cells by the lipid mediator prostaglandin E_1 (PGE_1), which acts via the two endogenously expressed G_s-linked EP2-EP4 receptors, or by orciprenaline, which stimulates endogenous β_2 receptors (**Fig. 3** and **Supplementary Fig. 9**); this was in contrast to the downward-deflected G_s signature in CHO cells (**Fig. 1c**). Similar cellular context dependency was also observed when forskolin, a direct adenylyl cyclase activator that bypasses the receptor, was applied. DMR responses of forskolin are essentially superimposable on those induced by stimulation of G_s GPCR agonists in both CHO and HEK293 cells (**Supplementary Fig. 10**). Differentiation of signatures with pathway modulators (**Fig. 3b–d**,**g–i**), specific receptor antagonists (data not shown) and second-messenger assays (**Fig. 3e**,**j**) confirmed and validated that optical traces for the tested receptors faithfully reflect stimulation of signaling pathways previously assigned to them. Taken together, these disparate DMR measurements suggest that unique differences exist in the spatiotemporal organization of the G_s downstream signaling network in these two cell lines.

Holistic DMR recordings uncover signaling promiscuity

The free fatty acid receptor FFA1 has previously been classified as a G_q/G_{11} sensitive receptor[35–37]. Stimulating FFA1-HEK cells with the small-molecule agonist TUG424 (ref. 38) induced robust DMR responses, distinct in shape from those obtained for G_i- and G_s-coupled receptors in this cellular background (compare **Fig. 4a** with **Fig. 3a**,**f**). However, unlike what would be expected for a G_q-sensitive receptor, FFA1-mediated DMR was only partly sensitive to inhibition by the G_q inhibitor YM (**Fig. 4b**; compare black and blue trace), although the same concentration of YM was sufficient to completely silence FFA1 activity in assays quantifying generation of inositol phosphates (IP1 assays), the classical approach to measure functional activity of G_q/G_{11}-sensitive receptors (**Fig. 4c**). Apparently, FFA1 engages signaling pathways in addition to the G_q/G_{11} pathway in this particular cellular background. Indeed, FFA1 also signals through the G_i pathway, as inferred from both the partial PTX sensitivity of the DMR signal (**Fig. 4b**; compare black and gray trace) and the partial inhibition by PTX of ERK1/2 MAP kinase phosphorylation (**Supplementary Fig. 11**). In agreement with a dual G_q/G_i coupling profile, only the combination of PTX and YM was required and

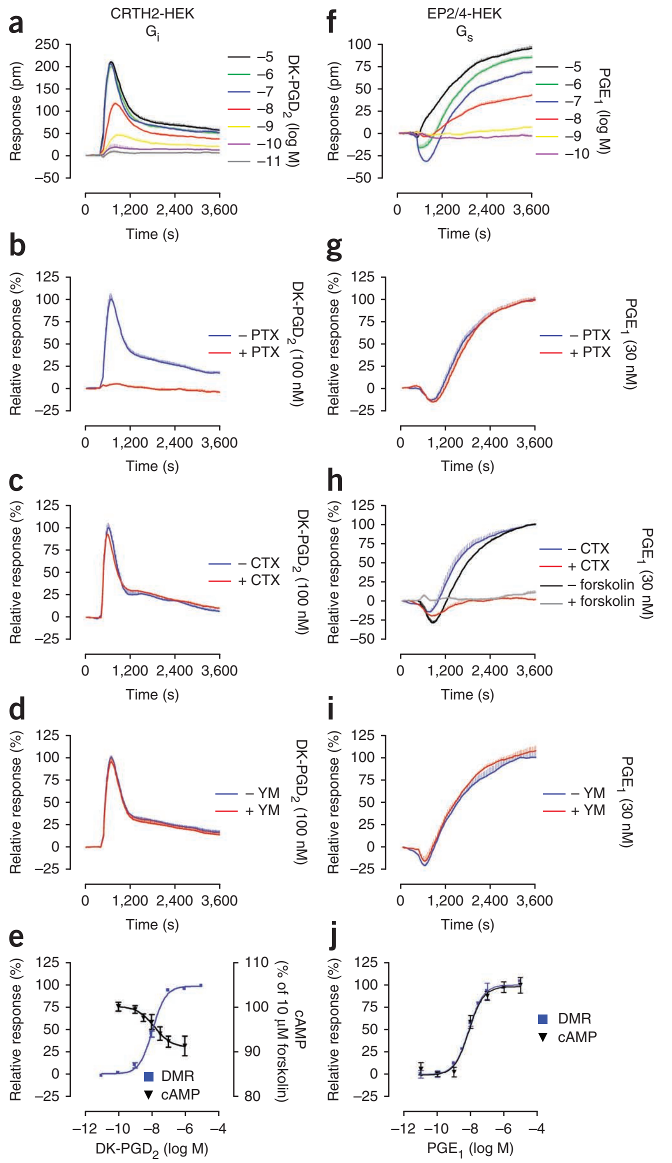


Figure 3 Dynamic mass redistribution enables measurement of differential receptor-mediated G protein activation in HEK293 cells. Shown are DMR and second-messenger assays performed with the following cell lines and receptors. (**a**–**e**) HEK293 cells stably expressing CRTH2. (**f**–**j**) HEK293 cells endogenously expressing EP2-EP4 receptors. In **a**,**f**, cells were challenged with the indicated concentrations of agonists and wavelength shift was monitored as a measure of receptor activation. Shown are representative data (mean + s.e.m.) of at least three independent experiments. In **b**,**g**, pretreatment of cells with 5 ng/ml PTX inhibits signaling of the G_i-sensitive CRTH2 receptor but not signaling of the G_s-sensitive EP2-EP4 receptors. In **c**,**h**, pretreatment of cells with 100 ng/ml CTX (or 10 μM forskolin) masks signaling of the G_s-sensitive EP2-EP4 receptors but does not affect G_i traces of CRTH2. In **d**,**i**, pretreatment of cells with 300 nM YM does not alter CRTH2 or EP2-EP4 traces. All data are normalized to the maximum response obtained in the absence of pharmacological inhibitors. In **e**,**j**, comparison of DMR assays with traditional cAMP second-messenger assays. In **e**, CRTH2-mediated decrease of intracellular cAMP is calculated as percent inhibition of adenylyl cyclase stimulated with 10 μM forskolin. Calculated log EC$_{50}$ values are: (**e**) DMR: −7.96 ± 0.04, cAMP: −7.85 ± 0.34; (**j**) EP2-EP4–mediated responses in DMR and cAMP assays are normalized to the maximum responses obtained by 10 μM PGE$_1$ in each assay. Calculated log EC$_{50}$ values are: DMR: −8.11 ± 0.07, cAMP: −8.11 ± 0.12. All data are means (±s.e.m.) of at least three independent experiments.

sufficient to completely erase the FFA1 response (**Fig. 4b**; compare black and red trace). Notably, however, G_i activity of FFA1 was barely detectable in traditional second-messenger cAMP inhibition assays (**Fig. 4d**), which is in good agreement with previous observations[35,36]. These findings highlight a strength of DMR technology: not only does DMR offer access to high content integrated cellular information, it also provides mechanistic insight if used in conjunction with inhibitors to deconvolute complex signaling pathways.

DMR can analyze GPCR function in human primary cells

Analyzing GPCR-mediated signal transduction in primary human cells—the cell type in which medicines are intended to mediate their therapeutic effect—is highly desirable for GPCR drug candidates. To test whether DMR is sufficiently sensitive to detect GPCR signaling in a native environment, we chose the cAMP-elevating agent PGE$_1$ as a stimulus, known to affect cell growth and cytokine production of human keratinocytes[39], and monitored DMR in both immortalized (HaCaT cells) and primary human keratinocytes obtained from six patients who underwent skin surgery. HaCaT and primary human keratinocytes responded with concentration-dependent optical traces reminiscent of those already observed in PGE$_1$-treated HEK293 cells (**Fig. 5a**,**b**; compare **Fig. 3f**). Indeed, PGE$_1$ traces in both HaCaT and primary human keratinocytes reflect activation of the two G_s-coupled EP2-EP4 receptors because the responses were sensitive to inhibition by a combination of EP2-EP4 antagonists (**Fig. 5c**,**d**), and were invisible following CTX- but not PTX- or YM-treatment (**Fig. 5e**,**f** and data not shown). Notably, although cAMP and DMR assays were both sufficiently sensitive to quantify PGE$_1$ activity in primary human cells, DMR was superior with respect to the quality of the signal window under conditions of low receptor expression (**Fig. 5g**,**h**).

DMR uncovers unknown signaling paradigms

It has been shown that persistent activation of the G_s signaling pathway can augment muscarinic M_3 receptor–mediated inositol phosphate production[40]. At first glance, the results reported here might appear to be in good accordance with this earlier report as we detected enhanced muscarinic M_3 receptor signaling in the presence of cAMP-elevating agents such as CTX (**Figs. 1g** and **6a**) and forskolin (**Fig. 6b**; red versus black trace). However, these enhanced M_3 signaling responses in DMR assays were sensitive to pretreatment of the cells with PTX, implying a G_i/G_o-mediated event (**Fig. 6a**,**b**; red versus blue trace). In contrast, M_3 DMR was completely insensitive to PTX pretreatment when intracellular cAMP was not elevated before application of the muscarinic agonist (**Fig. 6a**,**b**; gray versus black trace). These observations indicate that elevated intracellular cAMP—when present before the muscarinic agonist—serves as a stimulus to enable the M_3 receptor to engage an additional signaling pathway. Notably, detection of G_i activity under conditions of elevated cAMP can also be accomplished by immunocapture GTPγS binding assays (**Fig. 6c**) but not by traditional cAMP inhibition assays, in which receptor agonist and forskolin need to be co-applied simultaneously, not sequentially, to obtain measurable cAMP level changes as exemplified for the bona fide G_i-linked muscarinic M_2 receptor (**Supplementary Fig. 12**).

DISCUSSION

GPCRs constitute the single largest family of cell surface receptors, attracting great interest as therapeutic targets in all major disease

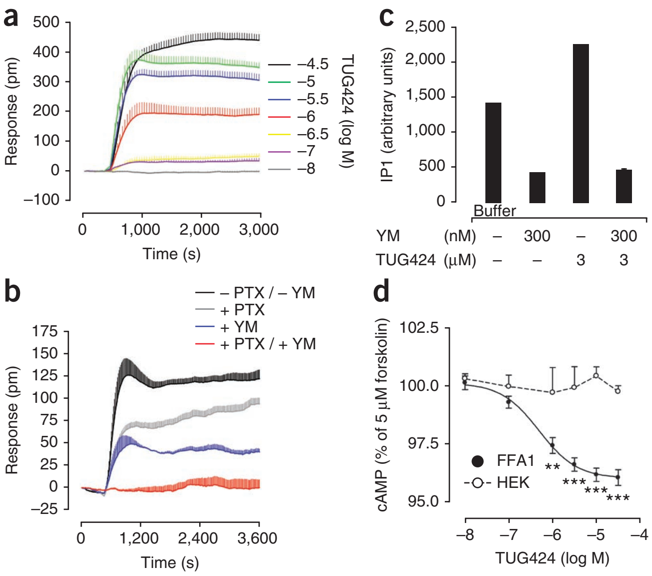


Figure 4 Parallel visualization of all signaling pathways unveils an additional signaling route of the free fatty acid receptor FFA1. (**a**) DMR recordings of FFA1-HEK cells treated with the indicated concentrations of the small-molecule agonist TUG424 (ref. 38). (**b**) The DMR signature obtained with 3 μM of the small-molecule FFA1 agonist TUG424 is partly sensitive to pretreatment of FFA1-HEK cells with PTX (5 ng/ml) or YM (300 nM) but completely abrogated in the presence of a combination of PTX and YM. (**c**) FFA1-mediated production of the second-messenger IP1 is completely blunted in the presence of 300 nM YM. FFA1-HEK cells were stimulated with 3 μM TUG424, and the resulting accumulation of inositol phosphates (IP1) was detected with an HTRF-IP1 assay kit as described in the Online Methods. (**d**) FFA1 activation of the G_i signaling pathway is statistically significant in cAMP inhibition assays. FFA1-HEK—or HEK293 cells for control—were stimulated with 5 μM forskolin, and inhibition of cAMP formation was quantified with an HTRF-cAMP assay kit as outlined in the Methods section. The cAMP level induced by stimulation with 5 μM forskolin (Fsk) was set to 100%. Shown are mean values and s.e.m. of three to six independent experiments. For statistical analysis, individual concentrations were compared by two-way ANOVA with Bonferroni's correction for multiple comparisons. $**P < 0.01$, $***P < 0.001$.

areas[1]. Accordingly, assay technologies enabling discovery of novel GPCR ligands are likely to substantially influence the drug discovery process. Recently, label-free technology platforms based on dynamic mass redistribution of intracellular proteins such as the Corning Epic Biosensor (for operating principle see **Fig. 1a**) or alteration of electric impedance have emerged for the study of GPCRs[5,22–25,41]. However, no in-depth analytical study to date has thoroughly 'validated' and/or compared the novel DMR technology with the more traditional biochemical and second-messenger assays that have been the mainstay of GPCR drug development.

Our results show that DMR technology can capture receptor activation of all four GPCR coupling classes (G_i/G_o, G_s, G_q and G_{12}/G_{13}), which at present is unachievable by most other technology platforms. DMR technology therefore represents a universal, pathway-unbiased yet pathway-sensitive approach toward investigation of G protein–mediated effects. The ability of the technology to detect signaling

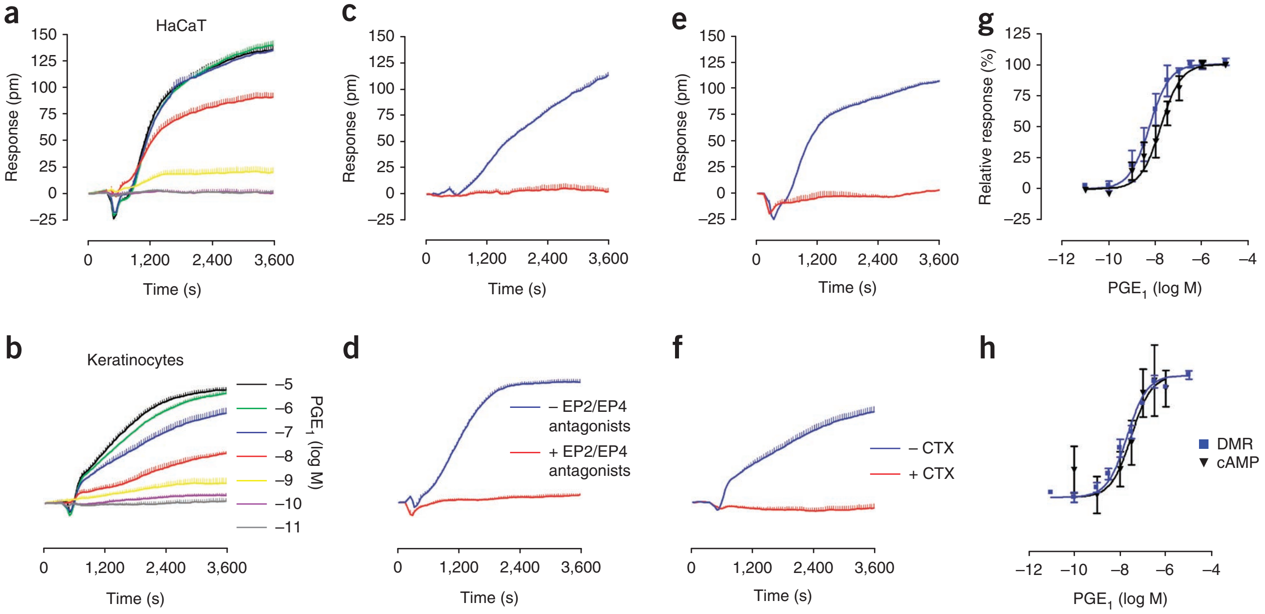


Figure 5 Dynamic mass redistribution enables analysis of GPCR functionality in immortalized and primary human keratinocytes. Top panels, immortalized human keratinocytes (HaCaTs); bottom panels, primary human keratinocytes. (**a**,**b**) The cell lines were challenged with the indicated concentrations of PGE_1, and wavelength shift over time was monitored as a measure of receptor activation. (**a**) Representative data (+ s.e.m.) of at least four independent experiments. (**b**) Representative data (+ s.e.m.) of cells from one human donor. Cells of five additional human donors yielded comparable optical traces (not shown). (**c**,**d**) PGE_1-mediated DMR is inhibited by pretreatment with a combination of the EP2-EP4 receptor antagonists AH6809 and L161,982, respectively. Optical traces of 30 nM PGE_1 (**c**) or 100 nM PGE_1 (**d**) in the absence and presence of a combination of 10 μM AH6809 and 3 μM L161,982. All data are representative data (+ s.e.m.) of at least four independent experiments. (**e**,**f**) DMR signatures of 100 nM PGE_1 are masked when cells are pretreated with 250 ng/ml cholera toxin (CTX). (**e**) Representative data (+ s.e.m.) of at least four independent experiments. (**f**) One representative dataset (+ s.e.m.) from one out of five human subjects. (**g**,**h**) Comparison of DMR assays with traditional endpoint cAMP second-messenger assays. Calculated log EC_{50} values are: (**g**) HaCaT: DMR: –8.27 ± 0.09, cAMP: –7.78 ± 0.09; (**h**) keratinocytes: DMR: –7.70 ± 0.06, cAMP: –7.60 ± 0.12. All data are means (±s.e.m.) of at least four independent experiments.

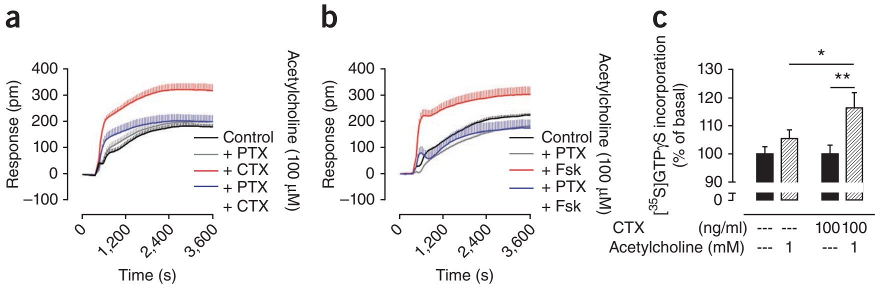


Figure 6 The muscarinic M_3 receptor adapts to adenylyl cyclase activation with a changed signaling repertoire. (**a**,**b**) Set of DMR experiments addressing (**a**) indirect and (**b**) direct activation of adenylyl cyclase by cholera toxin (CTX) and forskolin (Fsk), respectively. Acetylcholine (100 µM)-induced DMR traces were measured under control conditions and after pretreatment with pertussis toxin (100 ng/ml; PTX), CTX (100 ng/ml) or Fsk (10 µM) as indicated. Representative data (mean + s.e.m.) from at least three independent experiments. Note that PTX sensitivity emerges only after pretreatment with either CTX or forskolin. (**c**) GTPγS binding assay on membranes prepared from M_3-CHO cells. M_3-CHO cells were grown to confluence and were left untreated or were pretreated with CTX (20 h, 100 ng/ml) before membrane preparation. GTPγS incorporation was determined in the absence and presence of 1 mM acetylcholine followed by immunoenrichment of G_i proteins with an antiserum to the C-terminal region common to G_i proteins, as described in the Online Methods (mean + s.e.m., $n = 3$). P values <0.05 were considered statistically significant according to one-way analysis of variance (ANOVA) with Bonferroni's correction for multiple comparisons, as appropriate.

along the G_{12}/G_{13} pathway may be of great relevance to future 'GPCR deorphanization' strategies, particularly as receptors previously considered to be non-signaling might exclusively signal through G_{12}/G_{13}. Although lack of pathway bias is highly advantageous for deorphanization studies, the possibility must be considered that DMR traces, if opposing in direction and possessing identical kinetics, may yield zero signatures and therefore mask activity of biologically relevant molecules. Nevertheless, the use of pathway blockers should uncover such hidden pathway activation.

DMR technology and traditional second-messenger assays also diverge greatly in another aspect: DMR displays an overall cellular response, most likely encompassing a variety of cellular events downstream of the GPCR[5,22–24], which is in stark contrast to quantification of defined second messengers that only partially determine the overall response. This likely explains why agonist potencies determined with both methods may, but do not necessarily have to, converge. Indeed, this study revealed that the sensitivity of DMR is at least equal or even superior (**Fig. 1i**) to that of second-messenger recording for the detection of receptor-dependent, G protein–mediated signaling.

Complexity of optical traces obviously raises the possibility that an unimaginable wealth of intracellular players may be involved in defining the fine details of signature amplitude and duration. It will be exciting to unravel the individual components shaping complex optical response patterns, perhaps using libraries of signaling pathway inhibitors or genome-wide genetic screens with siRNA libraries. Our study does not solve the signature riddle completely, but provides a major mechanistic advance toward understanding the complex optical traces. Namely, heterotrimeric G proteins represent the postreceptor trigger responsible for orchestrating the complex response profiles for the various receptors and cellular backgrounds examined here, which was demonstrated using a combination of toxins and pharmacological pathway inhibitors.

The experimental power of these tools in label-free detection has been shown in this study for many different receptors and various cellular backgrounds, including primary human keratinocytes. Given the emerging successes in directing differentiation of embryonic or pluripotent stem cells to mature cells such as neurons or endothelial cells[42,43], label-free DMR detection raises the exciting possibility of expanding studies of drug action mechanisms and even drug screening processes to physiologically relevant cells. Native signaling has already been addressed in publications using label-free DMR detection[25]. All of these reports, however, have involved the analysis of immortalized cell lines[25], which are much less close to tissue biology than the primary human cells used here.

Our study also demonstrates how the collation of signaling routes within one dynamic, all-encompassing response, and its mechanistic deconvolution with appropriate pharmacological tools, can visualize unexpected signaling phenomena. Identification of an additional signaling pathway for the free fatty acid receptor FFA1 is one such example. In fact, the application of DMR technology to reveal ligand efficacy along the G_i pathway is particularly noteworthy because this aspect of FFA1 behavior is hardly detectable in the traditional cAMP inhibition assay (**Fig. 4d**). Identification of cAMP as an intracellular stimulus that confers signaling multiplicity onto the muscarinic M_3 receptor is another example. Although M_3-G_i interaction has been inferred indirectly many years ago on the basis of partial PTX sensitivity of M_3-mediated responses[44,45], a defined stimulus for this event has remained elusive so far. It is therefore important to stress that this particular mode of cellular cross-talk has been uncovered because DMR visualizes the summation of individual GPCR signaling routes during a single experiment, and because DMR, in contrast to traditional biochemical assays, does not require pharmacological manipulation of the second-messenger adenylyl cyclase–cAMP pathway to probe G protein (G_i) activity.

In summary, comparative analysis of traditional biochemical methods with the more recently developed DMR technology platform uncovers the experimental power of whole-cell label-free detection. Not only does DMR provide a temporally resolved readout for the summation of receptor-triggered signaling events in recombinant and primary living cells with unprecedented sensitivity and accuracy, but it is this cumulative readout of cellular activity that may disclose further levels of biological complexity in the regulation of signal transduction processes. We therefore anticipate that DMR, as a holistic readout of cell function, will advance systems biology and systems pharmacology and thereby promote the discovery of therapeutics with novel mechanisms.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank U. Rick, M. Vasmer-Ehses and T. Kögler for expert technical assistance and Corning Inc. for providing us with the Epic system. This work was supported by the DFG (Deutsche Forschungsgemeinschaft) grants KO 1582/3-1 to E.K., MO 821/2-1 to K.M. and WE 4428/1-1 to J.W. and by the Dr. Hilmer Foundation

(PhD fellowship to S.H.). A.K. is a member of the DFG-funded Research Training School GRK 677. N.J.S is an NHMRC/NHF of Australia Overseas Research Fellow. We thank M. De Amici and U. Holzgrabe (University of Milan and University of Würzburg) for kindly providing Hybrid 1 (ref. 47), T. Ulven (University of Southern Denmark) for TUG424 (ref. 38) and Astellas Pharma Inc. (Osaka, Japan) for providing us with YM-254890 (ref. 30).

AUTHOR CONTRIBUTIONS

R.S., N.J., J.S., A.K., N.M., S.H., A.M., S.B., M.M.-A. and N.J.S. designed and performed the experiments; E.K. and K.M. designed the research and wrote the manuscript; S.Z., J.W. and G.M. provided important biological samples or research tools; C.D., J.G., N.J.S., G.M. and R.S. provided important ideas and edited the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturebiotechnology/.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Overington, J.P., Al-Lazikani, B. & Hopkins, A.L. How many drug targets are there? *Nat. Rev. Drug Discov.* **5**, 993–996 (2006).
2. Baker, J.G. & Hill, S.J. Multiple GPCR conformations and signalling pathways: implications for antagonist affinity estimates. *Trends Pharmacol. Sci.* **28**, 374–381 (2007).
3. Bosier, B. & Hermans, E. Versatility of GPCR recognition by drugs: from biological implications to therapeutic relevance. *Trends Pharmacol. Sci.* **28**, 438–446 (2007).
4. Galandrin, S., Oligny-Longpre, G. & Bouvier, M. The evasive nature of drug efficacy: implications for drug discovery. *Trends Pharmacol. Sci.* **28**, 423–430 (2007).
5. Kenakin, T.P. Cellular assays as portals to seven-transmembrane receptor-based drug discovery. *Nat. Rev. Drug Discov.* **8**, 617–626 (2009).
6. Urban, J.D. *et al.* Functional selectivity and classical concepts of quantitative pharmacology. *J. Pharmacol. Exp. Ther.* **320**, 1–13 (2007).
7. Kenakin, T. Agonist-receptor efficacy. II. Agonist trafficking of receptor signals. *Trends Pharmacol. Sci.* **16**, 232–238 (1995).
8. Ryberg, E. *et al.* The orphan receptor GPR55 is a novel cannabinoid receptor. *Br. J. Pharmacol.* **152**, 1092–1101 (2007).
9. Eglen, R.M. Functional G protein-coupled receptor assays for primary and secondary screening. *Comb. Chem. High Throughput Screen.* **8**, 311–318 (2005).
10. Williams, C. cAMP detection methods in HTS: selecting the best from the rest. *Nat. Rev. Drug Discov.* **3**, 125–135 (2004).
11. Willoughby, D. & Cooper, D.M. Live-cell imaging of cAMP dynamics. *Nat. Methods* **5**, 29–36 (2008).
12. Barnea, G. *et al.* The genetic design of signaling cascades to record receptor activation. *Proc. Natl. Acad. Sci. USA* **105**, 64–69 (2008).
13. Hamdan, F.F., Audet, M., Garneau, P., Pelletier, J. & Bouvier, M. High-throughput screening of G protein-coupled receptor antagonists using a bioluminescence resonance energy transfer 1-based β-arrestin2 recruitment assay. *J. Biomol. Screen.* **10**, 463–475 (2005).
14. Lefkowitz, R.J. & Whalen, E.J. β-arrestins: traffic cops of cell signaling. *Curr. Opin. Cell Biol.* **16**, 162–168 (2004).
15. Olson, K.R. & Eglen, R.M. β-galactosidase complementation: a cell-based luminescent assay platform for drug discovery. *Assay Drug Dev. Technol.* **5**, 137–144 (2007).
16. Verkaar, F. *et al.* G protein-independent cell-based assays for drug discovery on seven-transmembrane receptors. *Biotechnol. Annu. Rev.* **14**, 253–274 (2008).
17. Azzi, M. *et al.* β-arrestin-mediated activation of MAPK by inverse agonists reveals distinct active conformations for G protein-coupled receptors. *Proc. Natl. Acad. Sci. USA* **100**, 11406–11411 (2003).
18. Galandrin, S. & Bouvier, M. Distinct signaling profiles of β1 and β2 adrenergic receptor ligands toward adenylyl cyclase and mitogen-activated protein kinase reveals the pluridimensionality of efficacy. *Mol. Pharmacol.* **70**, 1575–1584 (2006).
19. Hansen, J.L., Theilade, J., Haunso, S. & Sheikh, S.P. Oligomerization of wild type and nonfunctional mutant angiotensin II type I receptors inhibits $G\alpha_q$ protein signaling but not ERK activation. *J. Biol. Chem.* **279**, 24108–24115 (2004).
20. Hoffmann, C., Ziegler, N., Reiner, S., Krasel, C. & Lohse, M.J. Agonist-selective, receptor-specific interaction of human P2Y receptors with β-arrestin-1 and -2. *J. Biol. Chem.* **283**, 30933–30941 (2008).
21. Violin, J.D. & Lefkowitz, R.J. β-arrestin-biased ligands at seven-transmembrane receptors. *Trends Pharmacol. Sci.* **28**, 416–422 (2007).
22. Fang, Y., Ferrie, A.M., Fontaine, N.H., Mauro, J. & Balakrishnan, J. Resonant waveguide grating biosensor for living cell sensing. *Biophys. J.* **91**, 1925–1940 (2006).
23. Fang, Y., Li, G. & Ferrie, A.M. Non-invasive optical biosensor for assaying endogenous G protein-coupled receptors in adherent cells. *J. Pharmacol. Toxicol. Methods* **55**, 314–322 (2007).
24. Fang, Y. & Ferrie, A.M. Optical biosensor differentiates signaling of endogenous PAR1 and PAR2 in A431 cells. *BMC Cell Biol.* **8**, 24 (2007).
25. Rocheville, M. & Jerman, J.C. 7TM pharmacology measured by label-free: a holistic approach to cell signalling. *Curr. Opin. Pharmacol.* **9**, 643–649 (2009).
26. Dodgson, K., Gedge, L., Murray, D.C. & Coldwell, M. A 100K well screen for a muscarinic receptor using the Epic label-free system—a reflection on the benefits of the label-free approach to screening seven-transmembrane receptors. *J. Recept. Signal Transduct. Res.* **29**, 163–172 (2009).
27. Lee, P.H. *et al.* Evaluation of dynamic mass redistribution technology for pharmacological studies of recombinant and endogenously expressed G protein-coupled receptors. *Assay Drug Dev. Technol.* **6**, 83–94 (2008).
28. McGuinness, R.P. *et al.* Enhanced selectivity screening of GPCR ligands using a label-free cell based assay technology. *Comb. Chem. High Throughput Screen.* **12**, 812–823 (2009).
29. Peters, M.F., Vaillancourt, F., Heroux, M., Valiquette, M. & Scott, C.W. Comparing label-free biosensors for pharmacological screening with cell-based functional assays. *Assay Drug Dev. Technol.* **8**, 219–227 (2010).
30. Takasaki, J. *et al.* A novel $G\alpha_q$/11-selective inhibitor. *J. Biol. Chem.* **279**, 47438–47445 (2004).
31. Riobo, N.A. & Manning, D.R. Receptors coupled to heterotrimeric G proteins of the G12 family. *Trends Pharmacol. Sci.* **26**, 146–154 (2005).
32. Henstridge, C.M. *et al.* The GPR55 ligand l-α-lysophosphatidylinositol promotes RhoA-dependent Ca^{2+} signaling and NFAT activation. *FASEB J.* **23**, 183–193 (2009).
33. Ross, R.A. The enigmatic pharmacology of GPR55. *Trends Pharmacol. Sci.* **30**, 156–163 (2009).
34. Henstridge, C.M. *et al.* GPR55 ligands promote receptor coupling to multiple signalling pathways. *Br. J. Pharmacol.* **160**, 604–614 (2010).
35. Briscoe, C.P. *et al.* The orphan G protein-coupled receptor GPR40 is activated by medium and long chain fatty acids. *J. Biol. Chem.* **278**, 11303–11311 (2003).
36. Itoh, Y. *et al.* Free fatty acids regulate insulin secretion from pancreatic beta cells through GPR40. *Nature* **422**, 173–176 (2003).
37. Stoddart, L.A., Brown, A.J. & Milligan, G. Uncovering the pharmacology of the G protein-coupled receptor GPR40: high apparent constitutive activity in guanosine 5′-*O*-(3-[^{35}S]thio)triphosphate binding studies reflects binding of an endogenous agonist. *Mol. Pharmacol.* **71**, 994–1005 (2007).
38. Christiansen, E. *et al.* Discovery of potent and selective agonists for the free fatty acid receptor 1 (FFA(1)/GPR40), a potential target for the treatment of type II diabetes. *J. Med. Chem.* **51**, 7061–7064 (2008).
39. Zhang, J.Z., Maruyama, K., Iwatsuki, K., Ono, I. & Kaneko, F. Effects of prostaglandin E1 on human keratinocytes and dermal fibroblasts: a possible mechanism for the healing of skin ulcers. *Exp. Dermatol.* **3**, 164–170 (1994).
40. McGraw, D.W., Almoosa, K.F., Paul, R.J., Kobilka, B.K. & Liggett, S.B. Antithetic regulation by β-adrenergic receptors of G_q receptor sinaling via phospholipase C underlies the airway beta-agonist paradox. *J. Clin. Invest.* **112**, 619–626 (2003).
41. McGuinness, R. Impedance-based cellular assay technologies: recent advances, future promise. *Curr. Opin. Pharmacol.* **7**, 535–540 (2007).
42. Gaspard, N. *et al.* Generation of cortical neurons from mouse embryonic stem cells. *Nat. Protoc.* **4**, 1454–1463 (2009).
43. James, D. *et al.* Expansion and maintenance of human embryonic stem cell-derived endothelial cells by TGFβ inhibition is Id1 dependent. *Nat. Biotechnol.* **28**, 161–166 (2010).
44. Burford, N.T., Tobin, A.B. & Nahorski, S.R. Differential coupling of m1, m2 and m3 muscarinic receptor subtypes to inositol 1,4,5-trisphosphate and adenosine 3′,5′-cyclic monophosphate accumulation in Chinese hamster ovary cells. *J. Pharmacol. Exp. Ther.* **274**, 134–142 (1995).
45. Schmidt, M., Nehls, C., Rumenapp, U. & Jakobs, K.H. m3 Muscarinic receptor-induced and Gi-mediated heterologous potentiation of phospholipase C stimulation: role of phosphoinositide synthesis. *Mol. Pharmacol.* **50**, 1038–1046 (1996).

ONLINE METHODS

Materials and reagents. Tissue culture media reagents were from Invitrogen or Sigma-Aldrich, prostaglandin E_1 (PGE_1), 13,14-dihydro-15-keto-prostaglandin D_2 (DK-PGD_2) and AH6809 from Cayman, L161,982 from Tocris, forskolin (Fsk) from Applichem, $[^{35}S]GTP\gamma S$ from Perkin Elmer and Hank's balanced salt solution (HBSS) from Invitrogen. All other chemicals were obtained from Sigma-Aldrich unless explicitly indicated.

Cell culture and cell lines stably expressing individual GPCRs. The following receptors (human sequences) and cell lines were used: Flp-In-Chinese hamster ovary cells (Flp-In-CHO) stably expressing the M_2 or the M_3 receptor referred herein as M_2-CHO and M_3-CHO, CHO cells stably expressing the GPR55 or the β_2 receptor (GPR55-CHO and β_2-CHO) and untransfected CHO cells. CHO cells were cultured in Ham's nutrient mixture F-12 (HAM-F12) supplemented with 10% (v/v) FCS (FCS), 100 U/ml penicillin, 100 µg/ml streptomycin. The medium was complemented with 2 mM L-glutamine for M_2-CHO and M_3-CHO, 1 mM L-glutamine and 200 µg/ml G418 for β_2-CHO and 400 µg/ml G418 for GPR55-CHO.

HEK293 cells and HEK293 cells stably expressing CRTH2 (CRTH2-HEK cells), HEK293-Flp-In T-REx cells stably transfected with FFA1 (FFA1-HEK) and AD-HEK293 stably transfected with 3xHA-GPR55 (GPR55-HEK, kindly provided by Andrew Irving, University of Dundee, UK) were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% (v/v) FCS, 100 U/ml penicillin and 100 µg/ml streptomycin. For CRTH2-HEK and GPR55-HEK the medium was supplemented with 500 µg/ml G418, and for FFA1-HEK with 100 µg/ml hygromycin B and 15 µg/ml blasticidin (both from InvivoGen). For receptor expression FFA1-HEK cells were treated with 1 µg/ml doxycycline for 16 h.

Immortalized keratinocytes (human adult low calcium temperature, or HaCaT, cells) were grown in RPMI-1640 supplemented with 10% (v/v) FCS, 100 U/ml penicillin and 100 µg/ml streptomycin.

Primary human keratinocytes were obtained from skin samples of healthy patients and were cultured in KGM2 (Promocell) supplemented with 100 U/ml penicillin and 100 µg/ml streptomycin. All patients had provided written informed consent before excision. The study was approved by the ethics committee of the University of Bonn (concession-no. 090/04).

All cells were cultivated with 5% CO_2 at 37 °C.

Transient transfections of GPR55-HEK cells. To effectively deliver cDNA coding for dominant-negative G13 (G13Q226L,D294N) into GPR55-HEK cells an electroporation method was used as described previously[46]. DMR measurements were performed 48 h after transfection.

Dynamic mass redistribution (DMR) assays (Corning Epic Biosensor measurements). A beta version of the Corning Epic System was used consisting of a temperature-control unit, an optical detection unit, and an on-board robotic liquid handling device. Functional principle: a confluent cell layer adheres to the bottom of a well equipped with an optical biosensor. Ingoing broadband light is directed to travel along the bottom. The electromagnetic field extends into the cell layer for a depth of about 150 nm and loses energy depending on the optical density of the adjacent cell area, and the outgoing wavelength is measured. GPCR-mediated signaling affects optical density and thereby shifts the outgoing wavelength (measured in picometers) relative to pre-stimulus condition and is recorded over time. The magnitude of this wavelength shift is proportional to the amount of relocated intracellular matter: an increase in mass contributes positively and a decrease negatively to the overall response[22,23].

Cells were seeded onto 384-well Epic sensor microplates and cultured for 20–24 h to obtain confluent monolayers. GPR55 cells were treated as described previously[34].

Before the assay, cells were washed with assay buffer (HBSS with 20 mM HEPES) and transferred to the Epic reader for 2 h at 28 °C. DMR was monitored before and after addition of compound solutions. The incubation time for pretreatment with PTX or CTX was 16–20 h, for YM-254890 2.5 h, for aluminum fluoride 1.5 h and for forskolin 1–2.5 h.

$[^{35}S]GTP\gamma S$ assay. Membranes were prepared from M_2-CHO cells and $[^{35}S]GTP\gamma S$ incorporation measured as described previously[47,48]. For muscarinic M_3 receptors $[^{35}S]GTP\gamma S$ binding assays included an immunocapture step with an antiserum to the C terminus of G_i and were performed according to a previously published procedure[49].

Second-messenger accumulation assays (over expressed receptors). cAMP and IP1 accumulation were quantified with the HTRF-cAMP dynamic kit or the HTRF-IP1 kit, respectively (both from Cisbio) as per manufacturer's instructions and as described previously[50] on a Mithras LB 940 reader (Berthold Technologies).

cAMP accumulation assay (endogenously expressed receptors). cAMP accumulation was quantified with the competitive immunoassay HitHunter cAMP-HS+-kit (DiscoveRx Corp.) as per manufacturer's instructions using the Mithras LB 940 reader.

Calculations and data analysis. Quantification of DMR signals for concentration effect curves was performed either by calculation of the area under the curve (AUC) between 0 and 3,600 s (**Figs. 1i**,**j**, **2b**,**f**, **3j** and **5g**,**h**) or by the maximum value between 300 and 1,200 s (**Figs. 1h** and **3e**) for those traces that displayed fast kinetics and clear peak maxima. All optical DMR recordings were buffer and solvent corrected. For data normalization, indicated as relative response (%), top levels of concentration effect curves were set 100% and bottom levels 0%. Data calculation and EC_{50} value determination by nonlinear regression was performed using GraphPad Prism 4.02 (GraphPad Software).

Statistical analysis. Where appropriate, differences in means were examined by one- or two-way analysis of variance (ANOVA) with Bonferroni's multiple comparison post-hoc test using GraphPad Prism 5.01 (GraphPad Software). A *P* value <0.05 was considered statistically significant.

46. Pantaloni, C. *et al.* Alternative splicing in the N-terminal extracellular domain of the pituitary adenylate cyclase-activating polypeptide (PACAP) receptor modulates receptor selectivity and relative potencies of PACAP-27 and PACAP-38 in phospholipase C activation. *J. Biol. Chem.* **271**, 22146–22151 (1996).
47. Antony, J. *et al.* Dualsteric GPCR targeting: a novel route to binding and signaling pathway selectivity. *FASEB J.* **23**, 442–450 (2009).
48. Jäger, D. *et al.* Allosteric small molecules unveil a role of an extracellular E2/transmembrane helix 7 junction for G protein-coupled receptor activation. *J. Biol. Chem.* **282**, 34968–34976 (2007).
49. Smith, N.J., Stoddart, L.A., Devine, N.M., Jenkins, L. & Milligan, G. The action and mode of Binding of thiazolidinedione ligands at free fatty acid receptor 1. *J. Biol. Chem.* **284**, 17527–17539 (2009).
50. Schröder, R. *et al.* The C-terminal tail of CRTH2 is a key molecular determinant that constrains $G\alpha_i$ and downstream signaling cascade activation. *J. Biol. Chem.* **284**, 1324–1336 (2009).

doi:10.1038/nbt.1671

nature biotechnology

Genome sequence of the model mushroom *Schizophyllum commune*

Robin A Ohm[1], Jan F de Jong[1], Luis G Lugones[1], Andrea Aerts[2], Erika Kothe[3], Jason E Stajich[4], Ronald P de Vries[1,5], Eric Record[6,7], Anthony Levasseur[6,7], Scott E Baker[2,8], Kirk A Bartholomew[9], Pedro M Coutinho[10], Susann Erdmann[3], Thomas J Fowler[11], Allen C Gathman[12], Vincent Lombard[10], Bernard Henrissat[10], Nicole Knabe[3,18], Ursula Kües[13], Walt W Lilly[12], Erika Lindquist[2], Susan Lucas[2], Jon K Magnuson[8], François Piumi[6,7], Marjatta Raudaskoski[14], Asaf Salamov[2], Jeremy Schmutz[2], Francis W M R Schwarze[15], Patricia A vanKuyk[16], J Stephen Horton[17], Igor V Grigoriev[2] & Han A B Wösten[1]

Much remains to be learned about the biology of mushroom-forming fungi, which are an important source of food, secondary metabolites and industrial enzymes. The wood-degrading fungus *Schizophyllum commune* is both a genetically tractable model for studying mushroom development and a likely source of enzymes capable of efficient degradation of lignocellulosic biomass. Comparative analyses of its 38.5-megabase genome, which encodes 13,210 predicted genes, reveal the species's unique wood-degrading machinery. One-third of the 471 genes predicted to encode transcription factors are differentially expressed during sexual development of *S. commune*. Whereas inactivation of one of these, *fst4*, prevented mushroom formation, inactivation of another, *fst3*, resulted in more, albeit smaller, mushrooms than in the wild-type fungus. Antisense transcripts may also have a role in the formation of fruiting bodies. Better insight into the mechanisms underlying mushroom formation should affect commercial production of mushrooms and their industrial use for producing enzymes and pharmaceuticals.

The importance of mushroom-forming fungi in agriculture, human health and ecology underscores their biotechnological potential for a wide range of applications. The most conspicuous forms of these species, most of which are basidiomycetes, are their fleshy, spore-bearing fruiting bodies. Although these are primarily of economic value because of their use as food[1,2] (worldwide production of edible mushrooms amounts to ~2.5 million tons annually), mushrooms also produce antitumor and immunostimulatory molecules[1,2], as well as enzymes used for bioconversions[3]. Moreover, they have been identified as promising cell factories for the production of pharmaceutical proteins[4].

Despite their economic importance, relatively little is known about how mushroom-forming fungi obtain nutrients and how their fruiting bodies are formed. The vast majority of mushroom-forming fungi cannot be genetically modified, or even cultured under laboratory conditions. The basidiomycete *Schizophyllum commune*, which completes its life cycle in ~10 d, is a notable exception insofar as it can be cultured on defined media and there are a wealth of molecular tools to study its growth and development. It is the only mushroom-forming fungus for which genes have been inactivated by homologous recombination. The importance of *S. commune* as a model system is also exemplified by the fact that its recombinant DNA constructs will express in other mushroom-forming fungi[5]. In contrast, constructs that have been developed for ascomycetes are often not functional in mushroom-forming basidiomycetes.

S. commune is one of the most commonly found fungi and can be isolated from all continents, except for Antarctica. *S. commune* has been reported to be a pathogen of humans and trees, but it mainly adopts a saprobic lifestyle by causing white rot[6]. It is predominantly found on fallen branches and timber of deciduous trees. At least 150 genera of woody plants are substrates for *S. commune*, but it also colonizes softwood and grass silage[7]. The mushrooms of *S. commune* that form on these substrates are used as a food source in Africa and Asia.

In the life cycle of *S. commune*[8], meiospores germinate to form a sterile monokaryotic mycelium, in which each hyphal compartment

[1]Department of Microbiology and Kluyver Centre for Genomics of Industrial Fermentation, Utrecht University, Utrecht, The Netherlands. [2]Department of Energy Joint Genome Institute, Walnut Creek, California, USA. [3]Department of Microbiology, Friedrich Schiller University, Jena, Germany. [4]Department of Plant Pathology and Microbiology, University of California, Riverside, California, USA. [5]CBS-KNAW Fungal Biodiversity Centre, Utrecht, The Netherlands. [6]INRA, Biotechnologie des Champignons Filamenteux, Marseille, France. [7]Universités Aix-Marseille I & II, Marseille, France. [8]Chemical and Biological Process Development Group, Pacific Northwest National Laboratory, Richland, Washington, USA. [9]Biology Department, Sacred Heart University, Fairfield, Connecticut, USA. [10]Architecture et Fonction des Macromolecules Biologiques, Université Aix-Marseille I & II, Marseille, France. [11]Department of Biological Sciences, Southern Illinois University, Edwardsville, Illinois, USA. [12]Department of Biology, Southeast Missouri State University, Cape Girardeau, Missouri, USA. [13]Division of Molecular Wood Biotechnology and Technical Mycology, Büsgen-Institute, University of Göttingen, Göttingen, Germany. [14]Department of Biochemistry and Food Chemistry, University of Turku, Biocity A, Turku, Finland. [15]Wood Protection & Biotechnology, Empa, Swiss Federal Laboratories for Materials Testing and Research, St. Gallen, Switzerland. [16]Molecular Microbiology, Institute of Biology, Leiden University, Leiden, The Netherlands. [17]Department of Biological Sciences, Union College, Schenectady, New York, USA. [18]Present address: Dartmouth Medical School, Hanover, New Hampshire, USA. Correspondence should be addressed to J.S.H. (hortons@union.edu), I.V.G. (IVGrigoriev@lbl.gov) or H.A.B.W. (h.a.b.wosten@uu.nl).

Received 14 April; accepted 12 May; published online 11 July 2010; doi:10.1038/nbt.1643

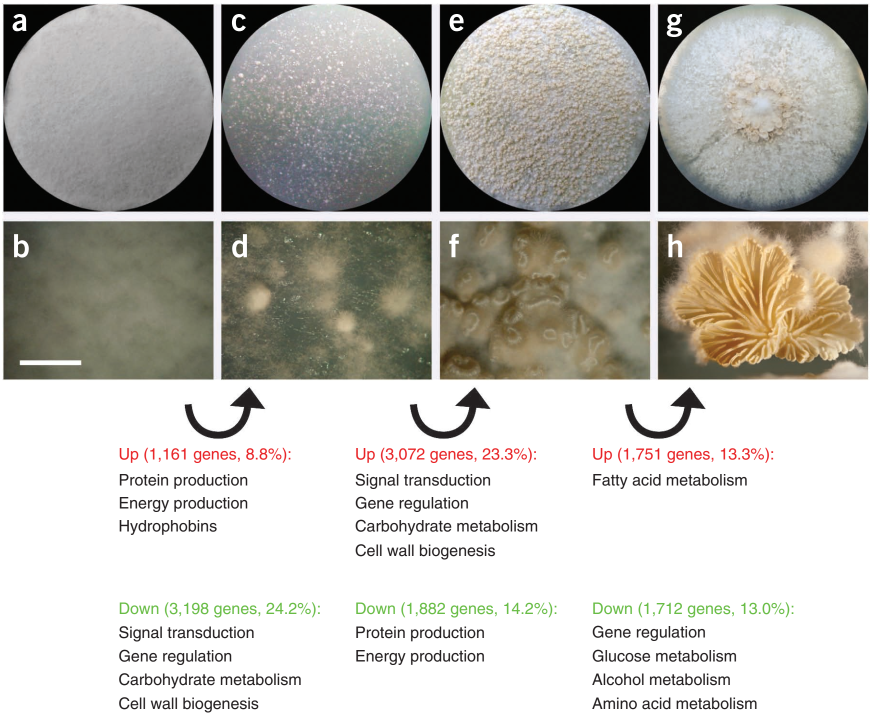


Figure 1 Development of *S. commune*. (**a**–**h**) Four-day-old (**a**–**f**) and 8-day-old (**g**,**h**) colonies grown from homogenates illustrate typical developmental stages in the life cycle of *S. commune*. A monokaryon generates sterile aerial hyphae that form a fluffy white layer on top of the vegetative mycelium (**a**,**b**). Aerial hyphae of a dikaryon interact with each other to form stage I aggregates (**c**,**d**), which, after a light stimulus, develop into stage II primordia (**e**,**f**). These primordia further differentiate into sporulating mushrooms (**g**,**h**). Enrichment analysis shows that particular functional terms are over-represented in genes that are up- or downregulated during a developmental transition. These terms are indicated below the panels. **a**,**c**,**e**,**g** represent cultures grown in 9-cm Petri dishes, whereas **b**,**d**,**f**,**h** represent magnifications thereof. Scale bar, 1 cm (**h**), 2.5 mm (**b**,**d**) and 5 mm (**f**).

contains one nucleus. Initial growth of this mycelium occurs beneath the surface of the substrate, with formation of aerial hyphae a few days after germination (**Fig. 1a,b**). Monokaryons that encounter each other fuse, and a fertile dikaryon forms when the alleles of the mating-type loci *matA* and *matB* of the partners differ. A short exposure to light is essential for fruiting, whereas a high concentration of carbon dioxide and high temperatures (30–37 °C) are inhibitory. Mushroom formation is initiated with the aggregation of aerial dikaryotic hyphae. These aggregates (**Fig. 1c**,**d**) form fruiting-body primordia (**Fig. 1e**,**f**), which further develop into mature fruiting bodies (**Fig. 1g**,**h**). Karyogamy and meiosis occur in the basidia within the mature fruiting body, and the resulting basidiospores can give rise to new monokaryotic mycelia.

Here we report the genomic sequence of the monokaryotic *S. commune* strain H4-8 and illustrate the potential of this basidiomycete as a model system to study mushroom formation. Besides the importance of understanding the sexual reproduction of *S. commune* for the commercial production of mushrooms, insight into the basis of this species' capacity to degrade lignocellulose may inspire more effective strategies to degrade lignocellulosic feedstocks for biofuel production.

RESULTS

The genome of *S. commune*

Sequencing of the genomic DNA of *S. commune* strain H4-8 with 8.29× coverage (**Supplementary Table 1**) revealed a 38.5-megabase genome assembly with 11.2% repeat content (**Supplementary Results 1**). The assembly is contained on 36 scaffolds (**Supplementary Table 2**), which represent 14 chromosomes[9]. We predict 13,210 gene models, with 42% supported by expressed sequenced tags (ESTs) and 69% similar to proteins from other organisms (**Supplementary Tables 3** and **4**). Clustering of the proteins of *S. commune* with those of other sequenced fungi (a phylogenetic tree of the organisms used in the analysis is shown in **Supplementary Fig. 1**) identifies 7,055 groups containing at least one *S. commune* protein (**Supplementary Table 5**). Analysis of these clusters suggested that 39% of the *S. commune* proteins have orthologs in the Dikarya and are thus conserved in the Basidiomycota and Ascomycota (**Supplementary Table 6**). Notably, a similar percentage of proteins (36%) are unique to *S. commune*, as based on OrthoMCL analysis. Of these proteins, 46% have at least one inparalog (a gene resulting from a duplication within the genome) in *S. commune*. The uniqueness of the *S. commune* proteome is also illustrated by the over- and under-representation of protein family (PFAM) domains compared to other fungi (**Supplementary Results 2**) and the fact that only 43% of the predicted genes (5,703 out of the 13,210) could be annotated with a gene ontology (GO) term.

Global gene expression analysis

We used massively parallel signature sequencing (MPSS) to compare whole-genome expression at the four developmental stages, defined by monokaryons, stage I aggregates, stage II primordia and mature fruiting bodies (**Fig. 1**). The majority of genes are either expressed in all four stages (4,859 genes) or not expressed in any of them (5,308 genes) (**Fig. 2** and **Supplementary Table 7**). Of the 13,210 predicted genes, 59.8% are expressed in at least one developmental stage (**Supplementary Table 7**). Fewer of the unique *S. commune* genes meet this criterion, whereas a higher percentage was observed for genes that share orthologs with Agaricomycetes or more distant fungi (**Supplementary Table 6**). This suggests that *S. commune* genes lacking homology to any reported sequences are more stringently regulated than orthologs of genes reported for other species. This is consistent with the observation that genes that are apparently unique to *S. commune* are over-represented in the pool of genes that are differentially expressed during the four developmental stages studied (**Supplementary Tables 8** and **9**).

Antisense transcription is a widespread phenomenon in *S. commune* (**Fig. 2b**,**c**). Of the tags that could be related to a gene model, 18.7% originate from an antisense transcript; and 42.3% of the predicted genes have antisense expression during one or more of the four developmental stages studied (**Supplementary Tables 7** and **10**). Northern hybridization with strand-specific probes confirmed the existence of antisense transcripts of *sc4* (DOE JGI Protein ID 73533; data not shown). Whereas a relatively large number of genes expressed in the antisense direction are uniquely expressed in stage II (2,888 genes), relatively few genes are expressed in the antisense direction in all stages (1,195 genes) (**Fig. 2b**). Our data suggest that 4,302 genes are expressed

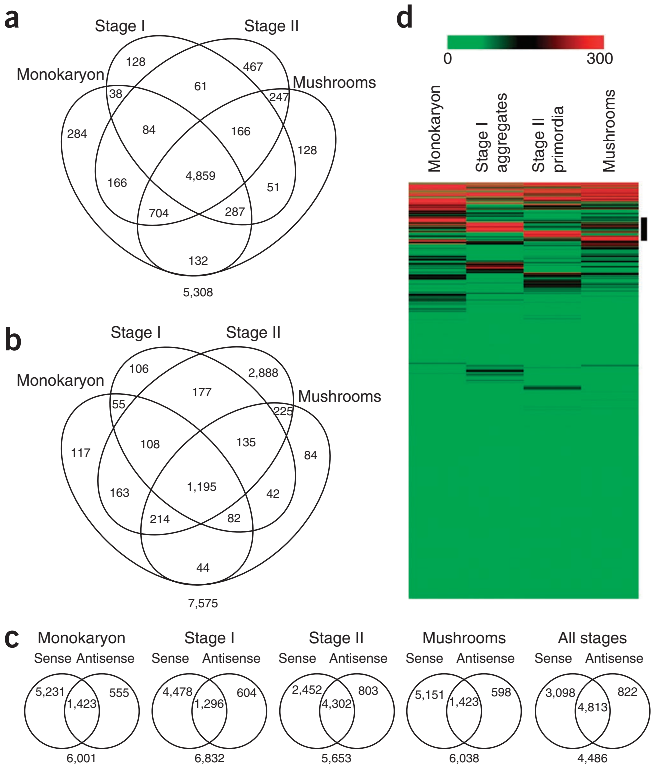


Figure 2 Gene expression in four developmental stages of *S. commune*. (**a**,**b**) The cutoff for expression is 4 tags per million (TPM). Venn diagrams show the overlap of genes expressed in the sense (**a**) and antisense (**b**) directions in the four developmental stages. For example, **a** shows that 61 genes are expressed in the sense direction in stage I and stage II, 4,859 genes are expressed in the sense direction in all stages, 132 genes are expressed in the sense direction in the monokaryon and mature fruiting bodies, and 5,308 genes are not expressed in the sense direction in any of the stages. (**c**) Venn diagrams of the overlap in genes that show sense and antisense expression in each developmental stage, and in all stages combined. (**d**) Heat map of expression of the *S. commune* genes in the four developmental stages. The bar at the top of the panel represents expression values between 0 and 300 TPM. Genes with expression values >300 TPM are also indicated in red. The bar on the right indicates a cluster of 366 highly expressed and differentially regulated genes. Annotation information for the genes in this cluster is given in **Supplementary Table 18**.

in both the sense and antisense directions during stage II (**Fig. 2c**). This overlap is larger for genes expressed during this phase of the life cycle than for the other developmental stages studied.

Fruiting-body development

We performed an enrichment analysis of functional annotation for the expression profiles of the developmental stages defined by monokaryons, stage I aggregates, stage II primordia and mature fruiting bodies. Functional terms involved in protein or energy production, or associated with hydrophobins, are over-represented in genes upregulated during formation of stage I aggregates (**Fig. 1** and **Supplementary Table 9**). Genes involved in signal transduction, regulation of gene expression, cell wall biogenesis and carbohydrate metabolism are enriched in the group of genes downregulated during the formation of stage I aggregates. These functional terms are enriched in the upregulated genes during formation of stage II primordia, whereas terms involved in protein and energy production are enriched in the downregulated genes (**Fig. 1** and **Supplementary Table 9**). Genes encoding transcription factors and genes involved in amino acid, glucose and alcohol metabolism are enriched in the group of genes downregulated during the formation of mature fruiting bodies.

As whole-genome expression was previously analyzed during mushroom formation in *Laccaria bicolor*[10], we next investigated whether the regulation of orthologous gene pairs of *L. bicolor* and *S. commune* might be correlated during fruiting. When we compared microarray expression profiles of free-living mycelium and mature fruiting bodies of *L. bicolor* to the MPSS expression profiles of monokaryotic mycelium and mature fruiting bodies of *S. commune*, we found that 6,751 expressed genes from *S. commune* had at least one expressed ortholog in *L. bicolor.* We determined the correlation of changes in expression of the functional annotation terms to which these orthologous pairs belong. There were 15 gene ontology terms, 2 KEGG terms, 4 KOG terms and 4 PFAM terms that showed a positive correlation in expression ($P < 0.01$; **Supplementary Table 11**). These terms include metabolic pathways (such as valine, leucine and isoleucine biosynthesis) and regulatory mechanisms (such as transcriptional regulation by transcription factors and signal transduction by G-protein α subunit). This indicates that regulation of these processes during mushroom formation is conserved in *S. commune* and *L. bicolor*.

Analysis of the matA and matB gene loci

Formation of a fertile dikaryon is regulated by the *matA* and *matB* mating-type loci. Proteins encoded in these loci activate signaling cascades (**Supplementary Results 3**) upstream of target genes. The target genes include those encoding enzymes and proteins that fulfill structural functions, such as hydrophobins (**Supplementary Results 4**), needed for the formation of fruiting bodies.

The *matA* locus of *S. commune* strain H4-8 appears to have more homeodomain genes than any fungal mating-type locus described thus far. This locus consists of two subloci, *Aα* and *Aβ*, which are separated by 550 kilobases (kb) on chromosome I of strain H4-8.

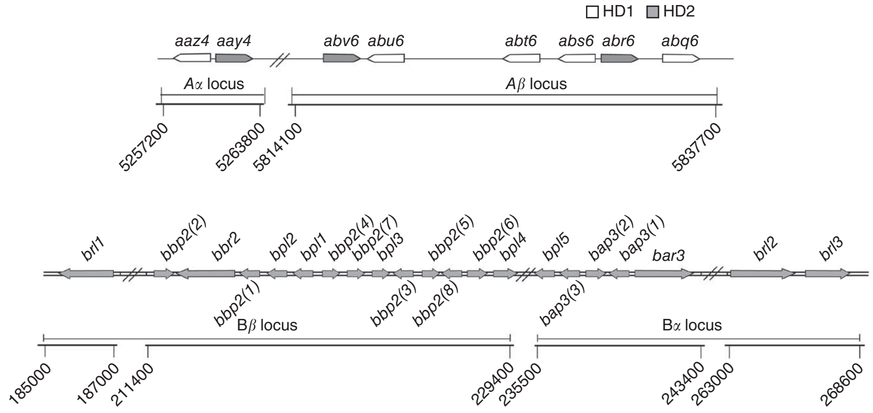


Figure 3 Distribution of genes encoding HD1 and HD2 homeodomain proteins in the *matA* locus and genes encoding pheromone receptors and pheromones in the *matB* locus of *S. commune* strain H4-8. The *matA* and *matB* loci are positioned on scaffolds 1 and 10, respectively. We identified an additional pheromone receptor gene, *brl4*, on scaffold 8.

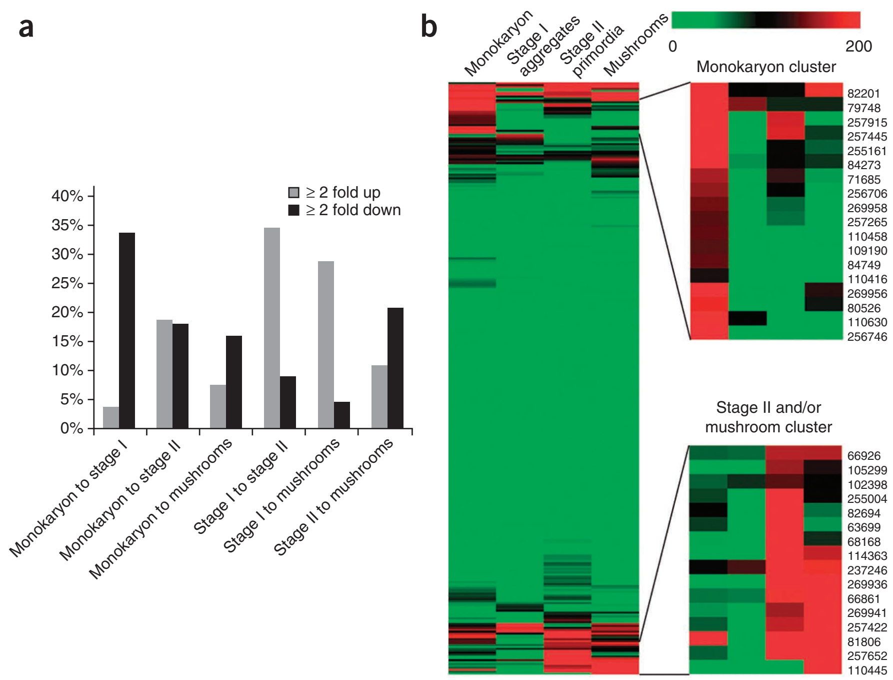


Figure 4 Expression of the 471 transcription factors in the genome of *S. commune*. (**a**) The histogram shows the percentage of transcription factor genes that are differentially expressed between stages of development. (**b**) The heat map shows a cluster containing predominantly monokaryon-specific transcription factors and a cluster containing predominantly stage II- and/or mushroom-specific transcription factors. These clusters are enlarged to the right of the heat map. The latter group contains two fungus-specific transcription factor genes, *fst3* and *fst4*.

Annotation revealed that the $A\alpha$ locus of H4-8 contains two divergently transcribed genes, which encode the Y and Z homeodomain proteins of the HD2 and HD1 classes, respectively (**Fig. 3** and **Supplementary Table 12**). These two genes, *aay4* and *aaz4*, have been described previously[1]. A homeodomain gene has also been identified previously in the $A\beta$ locus of H4-8 (ref. 11). Our genomic sequence revealed that this locus actually contains six predicted homeodomain genes: *abq6* (HD1), *abr6* (HD2), *abs6* (HD1), *abt6* (HD1, but lacking the nuclear localization signal), *abu6* (HD1) and *abv6* (HD2) (**Fig. 3** and **Supplementary Table 12**).

Annotation of the genomic sequence of *S. commune* reveals that the *matB* system contains more genes than previously envisioned. The *matB* locus comprises two linked loci, $B\alpha$ and $B\beta$, which both encode pheromones and pheromone receptors[1] (**Fig. 3**). Previously, one pheromone receptor gene was identified in both $B\alpha3$ and $B\beta2$ of strain H4-8 (called *bar3* and *bbr2*, respectively)[12]. The genome sequence of *S. commune* reveals four additional genes with high sequence similarity to these pheromone receptor genes, which we call *B* receptor–like genes 1 to 4 (*brl1* to *brl4*; **Fig. 3**). Three of these genes are located near *bar3* and *bbr2* on scaffold 10, whereas one (*brl4*) is located on scaffold 8. MPSS analysis shows that the *brl* genes are expressed (**Supplementary Table 13**). In fact, of all receptor and receptor-like genes, *brl3* shows the highest expression under the conditions tested.

Three and eight pheromone genes have previously been identified at the $B\alpha3$ and $B\beta2$ loci, respectively[13]. We identified one additional pheromone gene, named *B* pheromone–like-5 (*bpl5*), at the $B\alpha3$ locus. Moreover, four additional pheromone-like genes were detected at the $B\beta2$ locus, called *bpl1* to *bpl4* (**Fig. 3**). Of these, only *bpl2* showed no expression in MPSS analysis (**Supplementary Table 13**). The $B\alpha$ gene *bpl5* and three of the new $B\beta$ pheromone-like genes show deviations from the consensus farnesylation signal, CAAX (where C is cysteine, A is aliphatic and X is any residue), with the variant motifs CASR, CTIA, CRLT and CQLT for Bpl5, Bpl1, Bpl2 and Bpl3, respectively. Previously, one of the pheromone genes (*bbp2(6)*) was shown to function with the deviant farnesylation signal CEVM[12]. This suggests that in *S. commune* only one amino acid residue in the consensus sequence of the farnesylation signal needs to be aliphatic.

Transcription factors

The genome of *S. commune* reveals genes encoding 471 putative transcription factors, of which 311 are expressed during at least one developmental stage (**Supplementary Table 14**). Of these genes, 56% are expressed in all developmental stages; 268 were expressed in the monokaryon, 200 during formation of stage I aggregates, 283 during formation of stage II aggregates and 253 during formation of mushrooms. We identified a cluster of monokaryon-specific transcription factors and a group of transcription factors upregulated in stage II primordia or in mature mushrooms, or both (**Fig. 4**). The latter group includes *fst3* (NCBI Protein ID: 257422) and *fst4* (NCBI Protein ID: 66861),

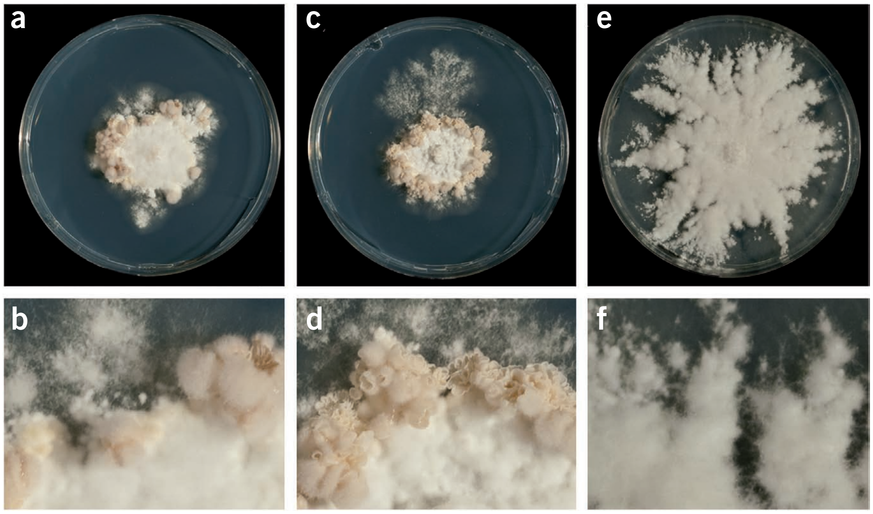


Figure 5 Transcription factors affecting fruiting body formation. (**a**,**b**) Wild-type dikaryon fruiting-body formation. (**c**–**f**) Fruiting-body formation in dikaryons in which *fst3* (**c**,**d**) or *fst4* (**e**,**f**) has been inactivated. Lower panels (**b**,**d**,**f**) show a magnification of part of the colonies shown in upper panels (**a**,**c**,**e**). Scale bar, 5 mm (**b**,**d**,**f**).

Table 1 Comparison of the number of FOLymes and CAZymes of *S. commune* with those of other fungi

	FOLymes											CAZymes			
Species	LO1	LO2	LO3	LDA1	LDA2	LDA3	LDA4	LDA5	LDA6	LDA7	LDA8	GH	GT	PL	CE
S. commune	2	0	1	1	0	2	1	0	4	4	1	240	75	16	30
C. cinerea	17	1	1	18	0	0	0	0	1	2	0	211	71	13	54
L. bicolor	9	1	0	4	0	0	0	0	3	2	0	163	88	7	20
P. placenta	2	0	0	3	0	0	0	0	0	1	2	124	51	4	13
P. chrysosporium	0	16	1	3	0	1	1	0	1	4	0	181	66	4	20
C. neoformans	0	0	0	0	0	0	0	0	0	0	0	75	64	3	8
U. maydis	0	0	0	0	0	1	0	1	1	1	0	101	64	1	19
S. cerevisiae	0	0	0	0	0	0	0	0	0	3	0	46	68	0	3
A. nidulans	1	0	1	0	1	0	0	0	1	0	0	250	91	21	32
N. crassa	5	0	2	1	0	0	0	1	1	1	1	173	76	4	23

LO1, laccases; LO2, peroxidases; LO3, cellobiose dehydrogenases; LDA1, aryl alcohol oxidases; LDA2, vanillyl-alcohol oxidases; LDA3, glyoxal oxidases; LDA4, pyranose oxidases; LDA5, galactose oxidases; LDA6, glucose oxidases; LDA7, benzoquinone reductases; LDA8, alcohol oxidases; GH, glycoside hydrolases; GT, glycosyl transferases; PL, polysaccharide lyases; CE, carbohydrate esterases.

which encode transcription factors that contain a fungus-specific Zn(II)2Cys6 zinc-finger DNA binding domain.

We inactivated the *fst3* and *fst4* genes via targeted gene deletions. The Δ*fst3* and Δ*fst4* monokaryons showed no phenotypic differences from the wild-type monokaryons. In contrast, the Δ*fst4* Δ*fst4* dikaryon did not fruit, but produced more aerial hyphae when compared to the wild type (**Fig. 5**). This suggests that Fst4 is crucial in the switch between the vegetative and reproductive phases of the *S. commune* life cycle. In contrast, the Δ*fst3* Δ*fst3* dikaryon formed more, albeit smaller, reproductive structures than those of the wild type (**Fig. 5**). As spatial and temporal regulation of fruiting-body formation and sporulation were not altered in the Δ*fst3* Δ*fst3* strain, we conclude that Fst3 inhibits the formation of clusters of mushrooms.

Wood degradation by *Schizophyllum commune*

As a white-rot fungus[6], *S. commune* degrades all woody cell wall components; in contrast, brown-rotters efficiently degrade cellulose but only modify lignin, leaving a polymeric residue. Lignin-degrading enzymes, which are commonly classified as FOLymes[14], comprise lignin oxidases (LO families) and lignin-degrading auxiliary enzymes that generate H_2O_2 for peroxidases (LDA families). The LO family consists of laccases (LO1), lignin peroxidases, manganese peroxidases, versatile peroxidases (LO2) and cellobiose dehydrogenases (CDHs; LO3). *S. commune* contains 16 FOLyme genes and 11 genes that encode enzymes distantly related to FOLyme enzymes (**Table 1** and **Supplementary Table 15**). The genome lacks genes encoding peroxidases of the LO2 family. However, it contains a CDH gene (LO3), two laccase genes (LO1) and 13 LDA genes, including four genes encoding glucose oxidases (LDA6) and benzoquinone reductases (LDA7) (**Table 1**).

S. commune appears to possess a more diverse assortment of FOLymes than the brown-rot fungus *Postia placenta* and the fungi that are known not to have ligninolytic activity (that is, *Ustilago maydis*, *Cryptococcus neoformans*, *Aspergillus nidulans*, *Neurospora crassa* and *Saccharomyces cerevisiae*; **Table 1**). In contrast, it has fewer FOLymes than either the coprophilic fungus *Coprinopsis cinerea* and the white-rot fungus *Phanerochaete chrysosporium*, which are predicted to possess 40 and 27 members, respectively[14].

Regarding polysaccharide degradation, *S. commune* has the most extensive machinery for degrading cellulose and hemicellulose of all of the basidiomycetes we examined. The Carbohydrate-Active Enzyme database (CAZy) identified 240 candidate glycoside hydrolases, 75 candidate glycosyl transferases, 16 candidate polysaccharide lyases and 30 candidate carbohydrate esterases encoded in the genome of *S. commune* (**Table 1** and **Supplementary Table 16**). Compared to the genomes of other basidiomycetes, *S. commune* has the highest number of glycoside hydrolases and polysaccharide lyases. *S. commune* is rich in genes encoding enzymes that degrade pectin, hemicellulose and cellulose (**Supplementary Table 17**). In fact, *S. commune* has genes in each family involved in the degradation of these plant cell wall polysaccharides. The *S. commune* genome is particularly rich in members of the glycosyl hydrolase families GH93 (hemicellulose degradation) and GH43 (hemicellulose and pectin degradation), and the lyase families PL1, PL3 and PL4 (pectin degradation) (**Supplementary Table 17**). The pectinolytic capacity of *S. commune* is further complemented by the presence of pectin hydrolases from families GH28, GH88 and GH105.

DISCUSSION

The phylum Basidiomycota contains roughly 30,000 described species, accounting for 37% of the true fungi[15]. The Basidiomycota comprises two class-level taxa (Wallemiomycetes and Entorrhizomycetes) and the subphyla Pucciniomycotina (rust), Ustilaginomycotina (smuts) and Agaricomycotina[16]. The Agaricomyotina include the mushroom- and puffball-forming fungi, crust fungi and jelly fungi. Genomic sequences are currently available for five members of the Agaricomycotina: *P. chrysosporium*[17], *L. bicolor*[10], *P. placenta*[18], *C. neoformans*[19] and *C. cinerea*[20]. Our 38.5-megabase assembly of the *S. commune* genome represents the first genomic sequence for a member of the family Schizophyllaceae. Thirty-six percent of the encoded proteins have no ortholog in other fungi. Only 43% of the predicted genes could be annotated with a gene ontology term, underscoring that much about the proteome of *S. commune* remains unknown. This percentage resembles that seen in other basidiomycetes: 30% in *L. bicolor*[10], 48% in *P. placenta*[18] and 49% in *P. chrysosporium*[17].

S. commune invades wood primarily by growing through the lumen of vessels, tracheids, fibers and xylem rays. Adjacent parenchymatic cells in the xylem tissue are invaded via simple and bordered pits. As a consequence of this approach to invasion, cellulose, hemicellulose or pectin can serve as the primary carbon source for *S. commune*. Indeed, the genome of *S. commune* probably encodes at least one gene in each family involved in the degradation of cellulose, hemicellulose and pectin. The large number of predicted pectinase genes is consistent with earlier studies describing *S. commune* as one of the best pectinase producers among the basidiomycetes[21]. *S. commune* also encodes carbohydrate-active enzymes that degrade other polymeric sugars, such as those acting on starch, mannans and inulins. Consistent with the wide variety of substrates that support its growth, *S. commune* has the most complete polysaccharide breakdown machinery of all basidiomycetes examined.

We know much less about how fungi degrade lignin than how they digest plant polysaccharides. Fungi are assumed to use FOLymes to degrade lignin[14]. Although members of the LO2 family of lignin oxidases are known to degrade lignin, it remains controversial whether laccases (LO1) and cellobiose dehydrogenases (CDHs; LO3) share this capacity. *S. commune* contains 16 genes encoding FOLymes. There are no members of the LO2 family, but the genome contains one CDH gene and two laccase genes. CDHs may participate in the degradation of cellulose, xylan and, possibly, lignin by generating hydroxyl radicals in a Fenton-type reaction. Laccases catalyze the one-electron oxidation of phenolic, aromatic amines and other electron-rich substrates with the concomitant reduction of O_2 to H_2O. They are classified as having either low or high redox potential[22], but it is not clear whether the two *S. commune* gene products belong to the high– or low–redox potential enzyme categories.

When the genomes of the white-rot fungi *S. commune* and *P. chrysosporium*[17] and the brown-rot fungus *P. placenta*[18] are compared, it is clear that *S. commune* has evolved its own set of FOLymes. *P. chrysosporium* lacks genes encoding laccases (LO1). It is thought to degrade lignin with the enzymes encoded by 16 isogenes of peroxidases (LO2), one CDH gene (LO3) and four genes of the multicopper oxidase superfamily. In contrast, *P. placenta* contains two laccase-encoding genes (LO1) but lacks members of the LO2 and LO3 families. As *S. commune* and *P. placenta* lack true LO2 FOLymes, one would expect a low number of LDAs that are responsible for H_2O_2 production for the peroxidases. This is not the case. *S. commune* contains more LDAs than *P. chrysosporium*. For instance, *S. commune* contains four glucose oxidase (LDA6) genes, whereas fungi seldom express more than one of these. In the absence of peroxidases of the LO2 family, it is expected that the glucose oxidases of *S. commune* serve another function. Glucose oxidases convert glucose into gluconic acid. This acid solubilizes inorganic phosphate and thus aids in the uptake of the nutrient[23].

The *matA* and *matB* mating-type loci of *S. commune* regulate the formation of a fertile dikaryon after the fusion of monokaryons that encounter one other. The genome sequence of this species now reveals that the mating type loci of *S. commune* contain the highest number of reported genes within such loci in the fungal kingdom. The *matB* locus comprises two linked loci, *B*α and *B*β, which both encode pheromones and pheromone receptors[1]. Nine allelic specificities have been identified for both loci, resulting in 81 different mating types for *matB*. It was previously reported that the *B*α*3* and *B*β*2* loci of H4-8 contain three and eight pheromone genes, respectively, and each contain one pheromone receptor gene[12,13]. We identified five additional pheromone genes and four additional pheromone receptor–like genes in the genome of H4-8. These newly identified receptor-like genes are present in a *matB* deletion strain, which has no pheromone response with any mate (T.J.F., unpublished data). This raises the question of whether the four receptor genes function in *matB*-regulated development. Expression of these genes, as discerned using MPSS, suggests that they do not represent pseudogenes.

The *matA* locus consists of two subloci, Aα and *A*β, of which 9 and 32 allelic specificities, respectively, are expected to occur in nature[1]. These loci are separated by 550 kb on chromosome I of strain H4-8. Such a large distance has not been found in other fungi that have a tetrapolar mating system. The functionally well-characterized *A*α locus showed no substantial differences from the published descriptions[1]. It is composed of two genes encoding Y and Z homeodomain proteins of the HD2 and HD1 classes, respectively. The Y and Z proteins, as in other basidiomycetes, interact in non-self combinations to activate the A-pathway of sexual development[1,24]. Notably, a nuclear localization signal is present in Y but not in Z. This is consistent with non-self interaction of the two proteins taking place in the cytosol, followed by the translocation of the active protein complex into the nucleus[1].

The *A*β locus of *S. commune* has been studied much less than the *A*α locus. Notably, *A*β reflects the highest degree of homeodomain-gene complexity for any fungal mating-type locus described to date. It contains four homeodomain genes of the HD1 class and two of the HD2 class. The *A*β locus of *S. commune* thus resembles that of *C. cinerea*, which consists of two pairs of functional HD1 and HD2 homeodomain genes (b and d)[25]. The large number of genes in *matA*β would explain why recombination analyses predict as many as 32 mating specificities for this locus[26]. Overall, *S. commune* seems ideal for identifying the evolutionary pathways that have created high numbers of allelic specificities for enhancing outbreeding versus inbreeding rates.

As little is known about molecular processes that control formation of fruiting bodies in basidiomycetes, other than the role of the mating-type loci[8], we compared genome-wide expression profiles at four developmental stages. MPSS showed that relatively few genes were specifically expressed in the monokaryon (284 genes) and in stage I aggregates and the mature mushrooms (128 genes in both cases). Notably, 467 genes were specifically expressed in stage II primordia. This suggests that this stage represents a major developmental switch, an idea supported by the fact that genes involved in signal transduction and regulation of gene expression are enriched in the group of upregulated genes during formation of stage II primordia. A positive correlation of expression of these gene groups during mushroom formation in both *S. commune* and *L. bicolor* suggests that regulation of mushroom formation is a conserved process in the Agaricales.

Our analysis of gene expression in *S. commune* reveals a high frequency of antisense expression. About 20% of all sequenced mRNA tags originated from an antisense transcript, and >5,600 of the predicted genes showed antisense expression in one or more developmental stages. Antisense transcription was most pronounced in stage II primordia. At this stage, >4,300 genes were expressed in both the sense and antisense directions, and >800 genes were expressed in the antisense direction only. Previously, MPSS has revealed antisense transcripts in *Magnaporthe grisea*[27]. Little is known about the function of these transcripts in fungi. The circadian clock of *N. crassa* is entrained in part by the action of an antisense transcript derived from a locus encoding a component of the circadian clock[28], possibly through RNA interference. It is tempting to speculate that antisense transcripts also regulate mRNA levels in *S. commune*. Natural antisense transcripts in eukaryotes have also been implicated in other processes, such as translational regulation, alternative splicing and RNA editing[29]. The antisense transcripts of *S. commune* may likewise have such functions. In all these cases, the antisense transcripts could function in a developmental switch that occurs when stage II primordia are formed.

The apparently high conservation of gene regulation in the Agaricales led us to study the 471 genes predicted to encode transcriptional regulators. Of these, 268 were expressed in the monokaryon, whereas 200, 283 and 253 were expressed during formation of stage I aggregates, stage II primordia and mushrooms, respectively. The relatively high number of transcription factors expressed during formation of stage II primordia again points to a major switch that probably occurs during this developmental stage.

We identified a group of monokaryon-specific transcription factors and a group of transcription factors that are upregulated in stage II primordia or mature mushrooms, or in both. The *fst3* and *fst4* genes encode transcriptional regulators belonging to the latter group. Growth and development were not affected in monokaryotic strains in which *fst3* or *fst4* were inactivated. Phenotypic differences were, however, observed in the dikaryon. The Δ*fst4* Δ*fst4* dikaryon did not fruit but produced more aerial hyphae than the wild type. In contrast, the Δ*fst3* Δ*fst3* dikaryon formed more, albeit smaller, fruiting bodies than the wild type. This suggests that Fst4 is involved in the switch between the vegetative and the reproductive phase, and that Fst3 inhibits formation of clusters of mushrooms. Inhibition of such clusters could be important in a natural environment to ensure that sufficient energy is available for full development of fruiting bodies. As *fst3* and *fst4* have homologs in other mushroom-forming fungi, it is tempting to speculate that they have similar functions in these organisms. This is supported by the observation that the homologs of *fst3* and *fst4* are upregulated in young fruiting bodies of *L. bicolor* compared to free-living mycelium[10]. In mature fruiting bodies of *L. bicolor*, the expression level of the homolog of *fst3* remains constant compared to young fruiting bodies, whereas the *fst4* homolog returns to the level expressed in the free-living mycelium.

In conclusion, the genomic sequence of *S. commune* will be an essential tool to unravel mechanisms by which mushroom-forming fungi degrade their natural substrates and form fruiting bodies. The large variety of genes that encode extracellular enzymes that act on polysaccharides probably explains why *S. commune* is so common in nature. Moreover, the genome sequence suggests that *S. commune* may have a unique mechanism to degrade lignin. Our MPSS data has provided leads on how mushroom formation is regulated, highlighting both the roles of certain transcription factors and the possible involvement of antisense transcription. Better understanding of the physiology and sexual reproduction of *S. commune* will probably have an impact on the commercial production of edible mushrooms and the use of mushrooms as cell factories.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Data availability and accession codes. *S. commune* assemblies, annotations and analyses are available through the interactive JGI Genome Portal at http://jgi.doe.gov/Scommune. Genome assemblies, together with predicted gene models and annotations, were also deposited at DDBJ/EMBL/GenBank under the project accession number ADMJ00000000. MPSS data have been deposited in NCBI's Gene Expression Omnibus with accession number GSE21265.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and the University of California, Lawrence Berkeley National Laboratory under contract no. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under contract no. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract no. DE-AC02-06NA25396. The work was also supported by the Dutch Technology Foundation STW, the Applied Science division of the Netherlands Organization for Scientific Research and the Technology Program of the Dutch Ministry of Economic Affairs.

AUTHOR CONTRIBUTIONS

J.S.H., H.A.B.W., T.J.F., W.W.L., A.C.G., E.K. and S.E.B. wrote the proposal; L.G.L., J.F.d.J., E.L. and R.A.O. isolated RNA and DNA and made libraries; E.L. and S.L. coordinated sequencing of the genome; J.S. assembled the genome; I.V.G., R.A.O., K.A.B., J.E.S., S.E.B., E.K. and H.A.B.W. coordinated the annotation process; I.V.G., A.A., A.S., J.E.S., T.J.F., E.R., A.L., F.P., U.K., J.K.M., J.F.d.J., R.P.d.V., P.M.C., V.L., B.H., W.W.L., A.C.G., P.A.v.K., K.A.B., J.S.H., E.K., S.E., N.K., R.A.O. and M.R. annotated genes; R.A.O., J.F.d.J., F.W.M.R.S. and L.G.L. performed experiments; R.A.O., J.F.d.J., F.W.M.R.S., L.G.L. and H.A.B.W. interpreted and designed experiments; R.A.O., A.A., J.E.S., E.R., J.F.d.J., R.P.d.V., F.W.M.R.S., K.A.B., E.K., J.S.H., I.V.G. and H.A.B.W. wrote the paper; R.A.O. and H.A.B.W. coordinated writing of the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Kothe, E. Mating-type genes for basidiomycete strain improvement in mushroom farming. *Appl. Microbiol. Biotechnol.* **56**, 602–612 (2001).
2. Kües, U. & Liu, Y. Fruiting body production in basidiomycetes. *Appl. Microbiol. Biotechnol.* **54**, 141–152 (2000).
3. Lomascolo, A., Stentelaire, C., Asther, M. & Lesage-Meessen, L. Basidiomycetes as new biotechnological tools to generate natural aromatic flavours for the food industry. *Trends Biotechnol.* **17**, 282–289 (1999).
4. Berends, E., Scholtmeijer, K., Wösten, H.A.B., Bosch, D. & Lugones, L.G. The use of mushroom-forming fungi for the production of N-glycosylated therapeutic proteins. *Trends Microbiol.* **17**, 439–443 (2009).
5. Alves, A.M. *et al.* Highly efficient production of laccase by the basidiomycete *Pycnoporus cinnabarinus*. *Appl. Environ. Microbiol.* **70**, 6379–6384 (2004).
6. Schmidt, O. & Liese, W. Variability of wood degrading enzymes of *Schizophyllum commune*. *Holzforschung* **34**, 67–72 (1980).
7. de Jong, J.F. *Aerial Hyphae of Schizophyllum commune: Their Function and Formation.* PhD thesis, Univ. Utrecht (2006).
8. Wösten, H.A.B. & Wessels, J.G.H. The emergence of fruiting bodies in basidiomycetes. in *The Mycota. Part I: Growth, Differentiation and Sexuality* (eds. Kües, U. & Fisher, R.) 393–414 (Springer, Berlin, 2006).
9. Asgeirsdottir, S.A., Schuren, F.H.J. & Wessels, J.G.H. Assignment of genes to pulse-field separated chromosomes of *Schizophyllum commune*. *Mycol. Res.* **98**, 689–693 (1994).
10. Martin, F. *et al.* The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. *Nature* **452**, 88–92 (2008).
11. Shen, G.P. *et al.* The Aalpha6 locus: its relation to mating-type regulation of sexual development in *Schizophyllum commune*. *Curr. Genet.* **39**, 340–345 (2001).
12. Fowler, T.J., Mitton, M.F., Vaillancourt, L.J. & Raper, C.A. Changes in mate recognition through alterations of pheromones and receptors in the multisexual mushroom fungus *Schizophyllum commune*. *Genetics* **158**, 1491–1503 (2001).
13. Fowler, T.J., Mitton, M.F., Rees, E.I. & Raper, C.A. Crossing the boundary between the *Bα* and *Bβ* mating-type loci in *Schizophyllum commune*. *Fungal Genet. Biol.* **41**, 89–101 (2004).
14. Levasseur, A. *et al.* FOLy: an integrated database for the classification and functional annotation of fungal oxidoreductases potentially involved in the degradation of lignin and related aromatic compounds. *Fungal Genet. Biol.* **45**, 638–645 (2008).
15. Kirk, P.M., Cannon, P.F., David, J.C. & Stalpers, J.A.. *Ainsworth and Bisby's Dictionary of the Fungi* (CAB International, Wallingford, UK, 2001).
16. Hibbett, D.S. *et al.* A higher-level phylogenetic classification of the Fungi. *Mycol. Res.* **111**, 509–547 (2007).
17. Martinez, D. *et al.* Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. *Nat. Biotechnol.* **22**, 695–700 (2004).
18. Martinez, D. *et al.* Genome, transcriptome, and secretome analysis of wood decay fungus *Postia placenta* supports unique mechanisms of lignocellulose conversion. *Proc. Natl. Acad. Sci. USA* **106**, 1954–1959 (2009).
19. Loftus, B.J. *et al.* The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*. *Science* **307**, 1321–1324 (2005).
20. Stajich, J.S. *et al.* Insights into evolution of multicellular fungi from the assembled chromosomes of the mushroom *Coprinopsis cinerea* (*Coprinus cinereus*). *Proc. Natl. Acad. Sci. USA* **107**, 11889–11894 (2010).
21. Xavier-Santos, S. *et al.* Screening for pectinolytic activity of wood-rotting basidiomycetes and characterization of the enzymes. *Folia Microbiol. (Praha)* **49**, 46–52 (2004).
22. Xu, F. *et al.* A study of a series of recombinant fungal laccases and bilirubin oxidase that exhibit significant differences in redox potential, substrate specificity, and stability. *Biochim. Biophys. Acta* **1292**, 303–311 (1996).
23. Xiao, C. *et al.* Isolation of phosphate-solubilizing fungi from phosphate mines and their effect on wheat seedling growth. *Appl. Biochem. Biotechnol.* **159**, 330–342 (2009).
24. Spit, A., Hyland, R.H., Mellor, E.J. & Casselton, L.A. A role for heterodimerization in nuclear localization of a homeodomain protein. *Proc. Natl. Acad. Sci. USA* **95**, 6228–6233 (1998).
25. Casselton, L.A. & Olesnicky, N.S. Molecular genetics of mating recognition in basidiomycete fungi. *Microbiol. Mol. Biol. Rev.* **62**, 55–70 (1998).
26. Raper, J. *Genetics of Sexuality of Higher Fungi* (The Roland Press, New York, 1966).
27. Gowda, M. *et al.* Deep and comparative analysis of the mycelium and appressorium transcriptomes of *Magnaporthe grisea* using MPSS, RL-SAGE, and oligoarray methods. *BMC Genomics* **7**, 310 (2006).
28. Kramer, C., Loros, J.J., Dunlap, J.C. & Crosthwaite, S.K. Role for antisense RNA in regulating circadian clock function in *Neurospora crassa*. *Nature* **421**, 948–952 (2003).
29. Lavorgna, G. *et al.* In search of antisense. *Trends Biochem. Sci.* **29**, 88–94 (2004).

ONLINE METHODS

Strains and culture conditions. *S. commune* was routinely grown at 25 °C on minimal medium (MM) with 1% (wt/vol) glucose and with or without 1.5% (wt/vol) agar[30]. Liquid cultures were shaken at 225 r.p.m. Glucose was replaced with 4% (wt/vol) glycerol for cultures used in the isolation of genomic DNA. All *S. commune* strains used were isogenic to strain 1-40 (ref. 31). Strain H4-8 (*matA43 matB41*; FGSC no. 9210) was used for sequencing. EST libraries were generated from H4-8 and from a dikaryon that resulted from a cross between H4-8 and strain H4-8b (*matA4 matB43*)[32]. Strains 4-39 (*matA41 matB41*; CBS 341.81) and 4-40 (*matA43 matB43*; CBS 340.81) were used for MPSS. These strains show a more synchronized fruiting compared to a cross between H4-8 and H4-8b. Partial sequencing of the haploid genome revealed that strains 4-40 and 4-39 have minor sequence differences (<0.2%) with strain H4-8 (data not shown).

Isolation of genomic DNA, genome sequencing and assembly. Genomic DNA of *S. commune* was isolated as described[30] and sequenced using a whole-genome shotgun strategy. All data were generated by paired-end sequencing of cloned inserts with six different insert sizes using Sanger technology on ABI3730xl sequencers. The data were assembled using the whole-genome shotgun assembler Arachne (http://www.broad.mit.edu/wga/).

EST library construction and sequencing. Cultures were inoculated on MM plates with 1% (wt/vol) glucose using mycelial plugs as an inoculum. Strain H4-8 was grown for 4 d in the light, whereas the dikaryon H4-8 × H4-8.3 was grown for 4 d in the dark and 8 d in the light. Mycelia of the dikaryotic stages were combined and RNA was isolated as described[30]. The poly(A)$^+$ RNA fraction was obtained using the Absolutely mRNA Purification kit and manufacturer's instructions (Stratagene). cDNA synthesis and cloning followed the SuperScript plasmid system procedure with Gateway technology for cDNA synthesis and cloning (Invitrogen). For the monokaryon, two size ranges of cDNA were cut out of the gel to generate two cDNA libraries (JGI library codes CBXY for the range 0.6 kb–2 kb and CBXX for the range >2 kb). For the dikaryon, cDNA was used in the range >2 kb, resulting in library CBXZ. The cDNA inserts were directionally ligated into vector pCMVsport6 (Invitrogen) and introduced into ElectroMAX T1 DH10B cells (Invitrogen). Plasmid DNA for sequencing was produced by rolling-circle amplification (Templiphi, GE Healthcare). Subclone inserts were sequenced from both ends using Big Dye terminator chemistry and ABI 3730 instruments (Applied Biosystems).

Annotation methods. Gene models in the genome of *S. commune* were predicted using Fgenesh[33], Fgenesh+[33], Genewise[34] and Augustus[35]. Fgenesh was trained for *S. commune* with a sensitivity of 72% and a specificity of 74%. Augustus *ab initio* gene predictions were generated with parameters based on *C. cinerea* gene models[20]. In addition, about 31,000 *S. commune* ESTs were clustered into nearly 9,000 groups. These groups were either directly mapped to the genomic sequence with a threshold of 80% coverage and 95% identity, included as putative full-length genes, or used to extend predicted gene models into full-length genes by adding 5′ and/or 3′ UTRs. Because multiple gene models were generated for each locus, a single representative model at each locus was computationally selected on the basis of EST support and similarity to protein sequences in the NCBI nonredundant database. This resulted in a final set of 13,210 predicted genes, of which 1,314 genes have been manually curated. In 66 cases, models were created or coordinates were changed.

All predicted gene models were functionally annotated by homology to annotated genes from the NCBI nonredundant set and classified according to Gene Ontology[36], eukaryotic orthologous groups (KOGs)[37], KEGG metabolic pathways[38] and Protein Family (PFAM) domains[39].

Repeat content. RepeatModeler 1.0.3 (http://www.repeatmasker.org/RepeatModeler.html) was used to generate *de novo* repeat sequence predictions for *S. commune*. Repeats were classified by comparison to the RepBase database (http://www.girinst.org/repbase/index.html). RepeatModeler produced 76 families of repeats used as a search library in RepeatMasker (http://www.repeatmasker.org/).

Orthologs of *S. commune* proteins in the fungal kingdom. Proteins of *S. commune* were assigned to orthologous groups with OrthoMCL version 2.0 (ref. 40) with an inflation value of 1.5. Members of such groups were assigned as orthologs (in the case of proteins from another species) or inparalogs (in the case of proteins from *S. commune*). Orthologs were determined in *C. cinerea*[20], *L. bicolor*[10], *P. placenta*[18], *P. chrysosporium*[17], *C. neoformans*[19], *U. maydis*[41], *S. cerevisiae*[42], *A. nidulans*[43] and *N. crassa*[44]. All-versus-all BLASTP analysis was performed using NCBI standalone BLAST version 2.2.20, with an *E* value of 10^{-5} as a cutoff. Custom scripts were used to further analyze the orthologous groups resulting from the OrthoMCL analysis. The evolutionary conservation for each orthologous group was expressed as the taxon this orthologous group was most specifically confined to (see **Supplementary Fig. 1**).

Representation analysis. FuncAssociate 2.0 (ref. 45) was used to study over- and under-representation of taxon-specific genes and of functional-annotation terms in sets of differentially regulated genes. Default settings were used, with a *P* value of 0.05 or 0.01 as the cutoff.

Protein families. The PFAM database version 24.0 (ref. 39) was used to identify PFAM protein families. Custom scripts in Python were written to group genes on basis of their PFAM domains. Differences in the number of predicted proteins belonging to a PFAM family across the fungal domains were determined using Student's *t*-test. When Agaricales were compared to the rest of the Dikarya, or when *S. commune* was compared to the Agaricales, only groups with a minimum of five members in at least one of the fungi were analyzed. When *S. commune* was compared to the rest of the Dikarya, only groups with a minimum of five members in at least four of the fungi were analyzed. In all cases, a *P* value of 0.05 was used as a cutoff. Similar results were obtained using the nonparametric Mann-Whitney *U*-test.

CAZy annotation. Annotation of carbohydrate-related enzymes was performed using the CAZy annotation pipeline[46]. Ambiguous family attributions were processed manually along with all identified models that presented defects (such as deletions, insertions or splicing problems). Each protein was also compared to a library of experimentally characterized proteins found in CAZy to provide a functional description.

FOLy annotation. Lignin oxidative enzymes (FOLymes)[14] were identified by BLASTP analysis of the *S. commune* gene models against a library of FOLy modules using an *e* value <0.1. The resulting 68 protein models were analyzed manually using the BLASTP results as well as multiple-sequence alignments and functional inference based on phylogeny[47]. Basically, a protein was identified as a FOLyme when it showed a similarity score above 50% with sequences of biochemically characterized enzymes. When the similarity score was <50% the proteins were scored as a FOLyme-related protein.

MPSS expression analysis. Total RNA was isolated from the monokaryotic strain 4-40 and from the dikaryon resulting from a cross between 4-40 and 4-39. A 7-day-old colony grown on solid MM at 30 °C in the dark was homogenized in 200 ml MM using a Waring blender for 1 min at low speed. Two milliliters of the homogenized mycelium was spread out over a polycarbonate membrane placed on top of solidified MM. Vegetative monokaryotic mycelium was grown for 4 d in the light. The dikaryon was grown for 2 and 4 d in the light to isolate mycelium with stage I aggregates and stage II primordia, respectively. Mature mushrooms 3 d old were picked from dikaryotic cultures that had grown for 8 d in the light. RNA was isolated as described[30]. MPSS was performed essentially as described[48] except that after DpnII digestion MmeI was used to generate 20-bp tags. Tags were sequenced using the Clonal Single Molecule Array technique (Illumina). Between 4.2 and 7.6 million tags of 20 bp were obtained for each of the stages. Programs were developed in the programming language Python to analyze the data. Tag counts were normalized to tags per million (TPM). Those with a maximum of <4 TPM in all developmental stages were removed from the data set. This data set consisted of a total of 40,791 unique tags. Of these tags, 61.7% and 58.6% could be mapped to the genome sequence and the predicted transcripts, respectively, using a perfect match as the criterion. The mapped tags accounted for 71.4% and 70.8%

doi:10.1038/nbt.1643

of the total number of tags, respectively. For comparison, 97.4% of the ESTs from *S. commune* strain H4-8 could be mapped to the assembly. Unmapped tags can be explained by sequencing errors in either tag or genomic DNA. Moreover, RNA editing may have altered the transcript sequencing to produce tags that do not match the genome perfectly. It may also be that the assigned untranslated region is incomplete or that the DpnII restriction site that defines the 5′ end of the tag is too close to the poly(A) tail of the mRNA. TPM values of tags originating from the same transcript were summed to assess their expression levels. A transcript is defined as the predicted coding sequence extended with 400-bp flanking regions at both sides.

Comparison of gene expression in *L. bicolor* and *S. commune*. Whole-genome expression analysis of *L. bicolor*[10] and *S. commune* was done essentially as described[49]. For *L. bicolor*, the microarray values from replicates were averaged. Expression values of genes were increased by 1, and the ratio between monokaryon and mushrooms (for *S. commune*), and between free-living mycelium and mature fruiting bodies (for *L. bicolor*), was log-transformed. All expressed genes from *S. commune* that had at least one expressed ortholog in *L. bicolor* were taken into account, resulting in a total of 6,751 orthologous pairs. These pairs were classified on the basis of functional-annotation terms. Correlation of changes in expression of these gene classes was expressed as the Pearson correlation coefficient. Only gene ontology terms with 10–200 pairs were used in the analysis. In the case of PFAM domains, a minimum of ten ortholog pairs were used.

Deletion of transcription factors *fst3* and *fst4*. The transcription factor genes *fst3* (NCBI Protein ID: 257422) and *fst4* (NCBI Protein ID: 66861) were deleted using the vector pDelcas[32]. Transformation of *S. commune* strain H4-8 was done as described[30]. Regeneration medium contained no antibiotic, whereas selection plates contained 20 μg ml^{-1} nourseothricin. Deletion of the target gene was confirmed by PCR. Compatible monokaryons with a gene deletion were selected from spores originating from a cross of the mutant strains with wild-type strain H4-8.3.

30. van Peer, A.F., de Bekker, C., Vinck, A., Wösten, H.A.B. & Lugones, L.G. Phleomycin increases transformation efficiency and promotes single integrations in *Schizophyllum commune*. *Appl. Environ. Microbiol.* **75**, 1243–1247 (2009).
31. Raper, J.R., Krongelb, G.S. & Baxter, M.G. The number and distribution of incompatibility factors in *Schizophyllum*. *Am. Nat.* **92**, 221–232 (1958).
32. Ohm, R.A. *et al.* An efficient gene deletion procedure for the mushroom-forming basidiomycete *Schizophyllum commune*. *World J. Microbiol. Biotechnol.* 10.1007/s11274–010–0356–0.
33. Salamov, A.A. & Solovyev, V.V. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516–522 (2000).
34. Birney, E. & Durbin, R. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* **10**, 547–548 (2000).
35. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, 215–225 (2003).
36. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
37. Koonin, E.V. *et al.* A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* **5**, R7 (2004).
38. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, D277–D280 (2004).
39. Finn, R.D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **38**, D211–D222 (2010).
40. Li, L., Stoeckert, C.J. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
41. Kämper, J. *et al.* Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. *Nature* **444**, 97–101 (2006).
42. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546–567 (1996).
43. Galagan, J.E. *et al.* Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature* **438**, 1105–1115 (2005).
44. Galagan, J.E. *et al.* The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* **422**, 859–868 (2003).
45. Berriz, G.F., Beaver, J.E., Cenik, C., Tasan, M. & Roth, F.P. Next generation software for functional trend analysis. *Bioinformatics* **25**, 3043–3044 (2009).
46. Cantarel, B.L. *et al.* The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* **37**, D233–D238 (2009).
47. Gouret, P. *et al.* FIGENIX: intelligent automation of genomic annotation: expertise integration in a new software platform. *BMC Bioinformatics* **6**, 198 (2005).
48. Brenner, S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630–634 (2000).
49. McCarroll, S.A. *et al.* Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat. Genet.* **36**, 197–204 (2004).

doi:10.1038/nbt.1643

nature biotechnology

Draft genome sequence of the oilseed species *Ricinus communis*

Agnes P Chan[1,10], Jonathan Crabtree[2,10], Qi Zhao[1], Hernan Lorenzi[1], Joshua Orvis[2], Daniela Puiu[3], Admasu Melake-Berhan[1], Kristine M Jones[2], Julia Redman[2], Grace Chen[4], Edgar B Cahoon[5], Melaku Gedil[6], Mario Stanke[7], Brian J Haas[8], Jennifer R Wortman[2], Claire M Fraser-Liggett[2], Jacques Ravel[2] & Pablo D Rabinowicz[1,2,9]

Castor bean (*Ricinus communis*) is an oilseed crop that belongs to the spurge (Euphorbiaceae) family, which comprises ~6,300 species that include cassava (*Manihot esculenta*), rubber tree (*Hevea brasiliensis*) and physic nut (*Jatropha curcas*). It is primarily of economic interest as a source of castor oil, used for the production of high-quality lubricants because of its high proportion of the unusual fatty acid ricinoleic acid. However, castor bean genomics is also relevant to biosecurity as the seeds contain high levels of ricin, a highly toxic, ribosome-inactivating protein. Here we report the draft genome sequence of castor bean (4.6-fold coverage), the first for a member of the Euphorbiaceae. Whereas most of the key genes involved in oil synthesis and turnover are single copy, the number of members of the ricin gene family is larger than previously thought. Comparative genomics analysis suggests the presence of an ancient hexaploidization event that is conserved across the dicotyledonous lineage.

The castor bean plant is a tropical perennial shrub that originated in Africa, but is now cultivated in many tropical and subtropical regions around the world. It can be self- and cross-pollinated and worldwide studies reveal low genetic diversity among castor bean germplasm[1,2]. Approximately 90% of the oil from castor bean seeds is composed of the unusual hydroxylated fatty acid ricinoleic acid[3]. Because of the nearly uniform ricinoleic acid content of castor oil, and the unique chemical properties that this fatty acid confers to the oil, castor bean is a highly valued oilseed crop for lubricant, cosmetic, medical and specialty chemical applications. Castor bean has also been proposed as a potential source of biodiesel; the high oil content of its seeds[4] and the ease with which it can be cultivated in unfavorable environments contribute to its appeal as a crop in tropical developing countries. It is believed that castor oil was first used as an ointment 4,000 years ago in Egypt, from where it spread to other parts of the world, including Greece and Rome, where it was used as a laxative 2,500 years ago[5].

An important obstacle to widespread cultivation of castor bean is the high content of ricin, an extremely toxic protein[6], in its seeds. Ricin is considered one of the deadliest natural poisons when administered intravenously or inhaled as fine particles. Ricin was first isolated more than a century ago[7]. It has been reportedly used as a weapon[6] and attempts to use ricin as a specific immunotoxin for therapeutic purposes in different cancers have been reported[8,9]. Its biochemical activity has been characterized as a type 2 ribosome-inactivating protein (RIP), composed of two subunits linked by a disulfide bond: a 32 kDa ricin toxin A (RTA) chain that harbors the ribosome-inactivating activity, and a 34 kDa ricin toxin B (RTB) chain, with a galactose-binding lectin domain. RTA is an N-glycosidase that depurinates adenine in a specific residue of the 28S ribosomal RNA[10,11]. The RTB chain allows ricin to enter eukaryotic cells by binding to cell surface galactosides and subsequent endocytosis. Other RIPs are common in plants, although they are not toxic because they are usually monomeric and lack a lectin domain. These proteins constitute the type 1 RIPs[12].

Ricin is synthesized as a precursor encoding both subunits in the endoplasmic reticulum of endosperm cells and is translocated and accumulated in protein bodies[13]. The precursor is proteolytically processed in the endoplasmic reticulum and in the protein bodies, where it is stored as the mature heterodimer. Ricin is very similar to the *R. communis* agglutinin (RCA)[14]. However, whereas ricin is a weak hemagglutinin, RCA has low toxicity and strong hemagglutinin activity. In addition, RCA is a tetrameric protein composed of two RTA- and two RTB-like subunits.

The relative ease with which ricin can be purified has raised concerns about its possible use in bioterrorism. For this reason, the United States produces only limited amounts of castor oil and is among the world's largest importers of castor oil and its derivatives. Moreover, much of the West's supply relies on importing castor oil

[1]J. Craig Venter Institute (JCVI), Rockville, Maryland, USA. [2]Institute for Genome Sciences (IGS), University of Maryland School of Medicine, Baltimore, Maryland, USA. [3]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA. [4]United States Department of Agriculture, Agricultural Research Service, Western Regional Research Center, Crop Improvement and Utilization, Albany, California, USA. [5]Center for Plant Science Innovation and Department of Biochemistry, University of Nebraska-Lincoln, Lincoln, Nebraska, USA. [6]International Institute of Tropical Agriculture, Oyo State, Ibadan, Nigeria. [7]Institut für Mikrobiologie und Genetik, Abteilung Bioinformatik, Universität Göttingen, Göttingen, Germany. [8]Broad Institute of the Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts, USA. [9]Department of Biochemistry and Molecular Biology, University of Maryland School of Medicine, Baltimore, Maryland, USA. [10]These authors contributed equally to this work. Correspondence should be addressed to P.D.R. (prabinowicz@som.umaryland.edu).

Received 30 June; accepted 2 August; published online 22 August 2010; doi:10.1038/nbt.1674

Table 1 Genome assembly and annotation statistics for the draft sequence of the castor bean genome

	All scaffolds	Scaffolds longer than 2 kb
Fold genome coverage	4.59	4.59
Number of scaffolds	25,828	3,500
Total span	350.6 Mb	325.5 Mb
N50 (scaffolds)	496.5 kb	561.4 kb
Largest scaffold	4.7 Mb	4.7 Mb
Average scaffold length	14 kb	93 kb
Number of contigs	54,000	24,500
Largest contig	190 kb	190 kb
Average contig length	6 kb	13 kb
N50 (contigs)	21.1 kb	
GC content	32.5%	
Gene models	31,237	
Gene density	11,220 bp/gene	
Mean gene length	2,258.6 bp	
Mean coding sequence length	1,004.2 bp	
Longest gene	15,849 bp	
Mean number of exons per gene	4.2	
Mean exon length	251 bp	
Longest exon	6,590 bp	
GC content in exons	44.5%	
Mean intron length	381 bp	
Longest intron	33,291 bp	
GC content in introns	31.8%	
Mean intergenic region length	6,846 bp	
Longest intergenic region	691,597 bp	
GC content in intergenic regions	30.7%	

from developing countries periodically threatened by political and economic instability. Therefore, knowledge of the genetics and enzymology of fatty acid metabolism in castor bean seeds is important in efforts to ensure a sustained supply of hydroxy fatty acids without the complications posed by the toxicity of ricin. A better understanding of the biology of ricin accumulation may permit the development of less toxic varieties, and more developed genomic information about the species may improve public safety by tracing the origins of samples used in potential bioterror attacks.

RESULTS

Genome sequencing and annotation

The castor bean genome, which is distributed across ten chromosomes, is estimated by flow cytometry to be ~320 Mb[15]. Especially as there is, to our knowledge, no available genetic map and limited genomic information for the species, we set out to generate a draft sequence of the castor bean genome by producing ~2.1 million high-quality sequence reads from plasmid and fosmid libraries (Online Methods), and then using the Celera assembler to build consensus sequences or contigs that were linked to form 25,800 scaffolds using the two end-sequences from individual clones (mate-paired reads). The assembly covered the genome ~4.6×, spanning 350 Mb, which is consistent with previous genome size estimations. If only the 3,500 scaffolds larger than 2 kb are considered, the assembly spans 325 Mb with an N50 of 0.56 Mb (**Table 1**).

We searched the genome sequence assembly for repetitive DNA using a combination of sequence alignment to databases of repetitive sequences and RepeatScout to identify repeats *de novo*. Overall, >50% of the genome was identified as repetitive DNA (excluding low-complexity sequences), most of which could not be associated with known element families. One-third of the repetitive elements were retrotransposons, and <2% were DNA transposons (**Table 2**). The most abundant known repeats are long terminal repeat elements (22.7% Gypsy-type and 9.5% Copia-type).

Protein-coding genes were annotated using multiple gene-prediction programs, homology searches against sequence databases and the cDNA spliced-alignment tool PASA (program to assemble spliced alignments). To aid the genome annotation, we also generated 52,165 expressed sequence tags (ESTs) from five cDNA non-normalized libraries. Using PASA, these and other castor bean cDNA sequences from GenBank could be aligned to 5,491 predicted genes and to 688 genomic regions where no gene had been predicted, allowing the creation of additional gene models. Once all gene-prediction programs and homology searches had been run, these data were consolidated into consensus gene predictions using the program Evidence Modeler (EVM; Online Methods). EVM showed better sensitivity and specificity than any of the individual gene finders used (**Supplementary Table 1**). In this way, we identified 31,237 gene models (**Table 1**). Using TIGR's paralogous families pipeline, 58.5% of the castor bean gene models were grouped in 3,020 predicted protein families, each comprising at least two members (**Supplementary Fig. 1** and **Supplementary Table 2**).

Polyploidization analysis

Although the castor bean genome assembly is fairly fragmented, it contains several megabase-sized scaffolds. We took advantage of these to investigate the extent of genome duplications in castor bean and contribute to the elucidation of the evolutionary history of the dicotyledonous lineage. Different models have been proposed to explain the origin of genome duplications in dicots. Whereas one supports the occurrence of an ancestral hexaploidization event common to all dicots[16], the other model suggests that all dicot genomes share one duplication event[17]. As analysis of genomic duplications in the castor bean genome provides an opportunity to contribute to resolving this controversy, we searched for putative paralogous genes using reciprocal best BLAST matches between all castor bean genes. We then selected the 30 pairs of scaffolds that contained the largest numbers of paralogous gene pairs, and displayed the 22 unique scaffolds containing those 30 pairs of scaffolds in a dot plot. This approach identified six triplicated regions (regions for which two additional paralogous regions exist in the genome). We also identified nine duplicated regions (unmarked strings of dots) for which we cannot determine whether or not a third paralogous region exists (**Fig. 1**). We then carried out a more precise and comprehensive search for evidence of genomic triplications by first building Jaccard clusters of paralogous genes using an all-versus-all BLASTP search. We identified and displayed blocks of syntenic genes using Sybil[18] and manually inspected the results to identify triplicated regions. Using this method, we identified 17 triplicated regions (**Supplementary Fig. 2**) that included those found using the reciprocal best BLAST matches method. The fact that the triplications were found in multiple groups of scaffolds suggests that the castor bean genome underwent a hexaploidization event.

Table 2 Classification of repetitive sequences in the draft sequence of the castor bean genome

	Length occupied (bp)	Total repeats (%)	Genome (%)
Retrotransposons	61,199,930	36.07	18.16
Gypsy	38,595,566	22.75	11.45
Copia	16,078,721	9.48	4.77
Line	465,220	0.27	0.14
Sine	1,867	0.00	0.00
Other	6,058,556	3.57	1.80
Unclassified elements	105,387,872	62.12	31.26
DNA transposons	3,065,391	1.81	0.91
Total transposable elements	169,653,193	25.33	50.33
Low complexity sequences	6,348,051	0.95	1.88

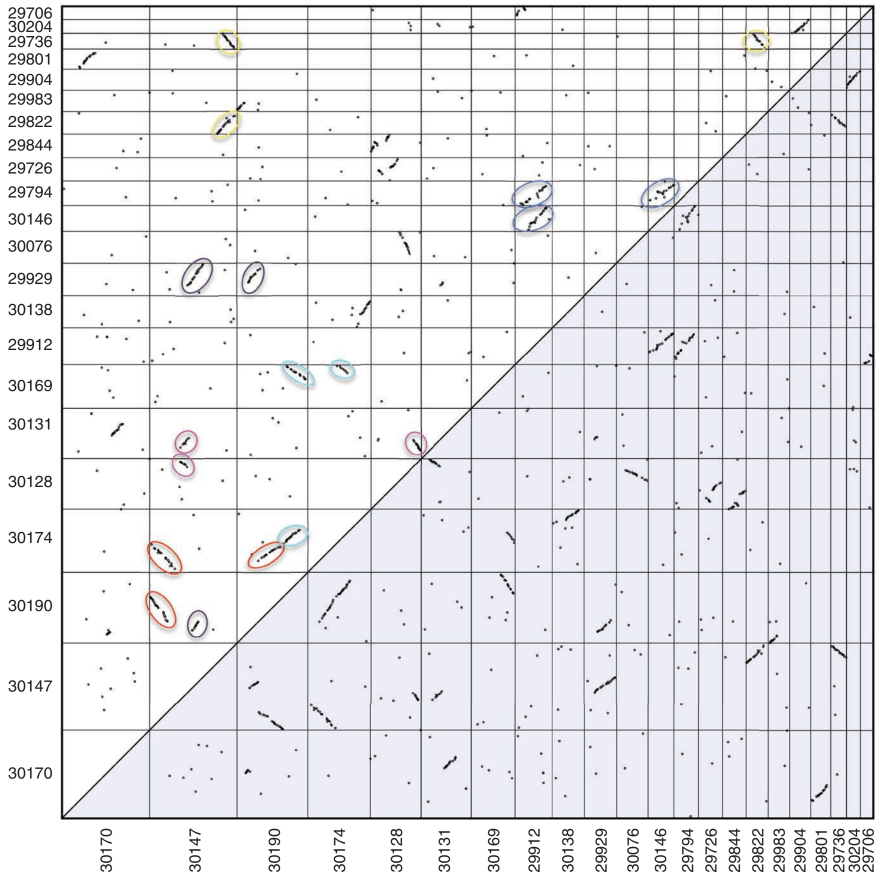


Figure 1 Reciprocal best BLAST matches between castor bean genes. Strings of paralogous genes that correspond to triplicated regions are highlighted in the same color. The 30 pairs of scaffolds that contained the highest numbers of paralogous gene pairs are shown.

To determine whether the triplication of the castor bean genome corresponds to ancestral polyploidization events previously described in the dicot lineage, we compared triplicated regions in the castor bean genome with the *Arabidopsis thaliana*[19], poplar[20], grapevine[16], and papaya[21] genomes by generating Jaccard clusters in a pairwise manner between castor bean and each of the other genomes. Of the 17 triplications, 8 (including 5 of the 6 triplications identified by reciprocal best BLAST matches) contained blocks of five or more syntenic gene pairs between each of the three castor bean regions and all of the other dicot genomes. Castor bean paralogous gene blocks generally showed a one-to-one, one-to-two and one-to-four relationship with their grapevine, poplar and *A. thaliana* orthologs, respectively (**Fig. 2** and **Supplementary Fig. 3**). Some exceptions were observed in the comparison with *A. thaliana* that were expected due to the further re-arrangements that exist in its genome[19]. Comparison between the castor bean and papaya genomes is less clear due to the fragmentation of both genome assemblies. Our results support the presence of a hexaploidization event common to all dicots, as well as one additional genome duplication in poplar, and two further duplications in the *A. thaliana* genome.

The ricin gene family

As the presence of ricin makes castor bean an important subject for biosecurity research, we analyzed the lectin gene family that includes the genes for ricin and RCA. The ricin gene encodes three domains: an N-terminal RIP domain and two C-terminal lectin domains. It has been reported that this gene family comprises 6–8 members, detected by Southern-blot hybridization using a ricin cDNA probe[22,23]. However, our draft of the castor bean genome reveals 28 putative genes in the family, including potential pseudogenes or gene fragments. To increase the reliability of our analysis of this gene family by improving the sequence and assembly quality, we manually finished sequence gaps or ambiguities inside the ricin-like gene models. In this way, the sequence and assembly of eight scaffolds was improved and the 28 gene family members (**Fig. 3**) were contained in a total of 17 scaffolds, each containing 1–5 ricin-agglutinin gene family members (**Supplementary Table 3**). These results suggest that the members of this lectin gene family tend to be clustered in the castor bean genome. The largest cluster spans 70 kb and includes a group of five family members interrupted by one gene that does not belong to the gene family. The other clusters contain two or three gene family members in regions ranging between 0.7 and 17 kb. Ten scaffolds contained only a single gene-family member, and four of them were longer than 250 kb, suggesting that these four genes were not part of clusters. However, it is uncertain if the other six scaffolds that contain only one member of the family are part of clusters because they are shorter than 12 kb. Probably some of these tandem duplications were not discriminated in previous studies using Southern-blot analysis, resulting in an underestimation of the gene family size. Furthermore, although we did not manually curate structural annotation, we found two cases in which adjacent ricin-like gene fragments could belong to pseudogenes that accumulated frame shifts and stop codons (**Fig. 3**).

The length of the different members of the family identified by automatic annotation was variable, ranging from 66 to 584 amino acids. Although some of the shorter genes could be nonfunctional or pseudogenes, start and stop codons could be predicted, making it difficult to determine whether they are functional or not. Moreover, four of them were truncated as a consequence of their location at the end of a contig or scaffold. Sequence comparison to ricin and RCA coding sequences in GenBank uncovered one full-length gene model (60629.m00002) identical to the ricin-coding sequence and another full-length gene model (60637.m00004) showing 99% identity to the sequence encoding RCA. These gene models likely correspond to the reported ricin and RCA sequences, respectively. An additional predicted gene (60628.m00003) shows complete identity to the ricin-coding sequence, although presumably, the sequence coding for about 150 of the 576 amino acids is missing from this gene model because it is located at the end of a scaffold. Three other gene models are truncated in a similar way (60626.m00001; 60639.m00003; 60627.m00002) and show 100% identity to the ricin-coding sequence, although the available sequences are much shorter (149 to 188 amino acids). Thus, it is uncertain whether these genes represent complete identical copies of the gene encoding ricin. The rest of the gene family members showed different degrees of similarity to the ricin- or RCA-coding sequences. Overall, 7 of the 28 genes of the lectin family encode proteins that contain the RIP and the two lectin domains, 9 encode proteins with only the RIP domain and 9 encode proteins with one or two lectin domains only (**Fig. 3**). cDNA alignments showed evidence of expression of the genes encoding ricin and RCA as well as one of the homologs (60638.m00018) for which a putatively complete gene was modeled (data not shown). Furthermore, evidence of RIP activity has been recently reported for the proteins encoded by the seven full-length ricin-like genes[24].

Oil metabolism genes

In light of the importance of castor bean as an oilseed crop, we examined the annotation of 71 gene models that showed similarity to known genes involved in the biosynthesis of fatty acids and triacylglycerols, which in castor bean correspond mainly to ricinoleic acid and triricinolein[25]. Of these 71 gene models, the annotation of 67 was manually improved (**Supplementary Table 4**). Castor bean has not only evolved an oleic acid hydroxylase to synthesize ricinoleic acid, but has also developed the capacity to efficiently accumulate high levels of ricinoleic acid in its seed oil. Therefore, we focused on a few key genes in the ricinoleic acid biosynthetic and metabolic pathways. The oleic acid hydroxylase gene (*FAH*), which produces ricinoleic acid from oleoyl-phosphatidycholine, likely evolved from the widely occurring *FAD2* gene for the Δ12-oleic acid desaturase[26]. BLAST searches of these genes against the entire castor genome confirmed that there is only one copy of each of these genes (28035.m000362 and 29613.m000358, respectively). Among the key enzymes involved in the incorporation of ricinoleic acid into oils are diacylglycerol acyltransferases (DGATs), which catalyze the final step in triacylglycerol assembly. Two classes of endoplasmic reticulum–associated DGATs (DGAT1 and DGAT2) occur in castor bean, as well as a homolog of a soluble DGAT[27–29]. The gene models coding for these enzymes are also single copy (29912.m005373, 29682.m000581 and 29889.m003411, respectively). In addition to DGAT-coding genes, it is likely that other genes have evolved to maintain high and specific flux of ricinoleic acid from its synthesis on phosphatidylcholine to its storage in triacylglycerols in castor bean seeds.

Remarkably, even though ricinoleic acid accounts for nearly 90% of the fatty acids in castor bean seeds, it represents <5% of the fatty acids in phosphatidylcholine[30]. Although the mechanism for ricinoleic acid flux among lipid classes is not clear, a number of specialized acyltransferase and phosphatidylcholine metabolic enzymes likely participate in these reactions, including phospholipid:diacylglycerol acyltransferase 1 (PDAT1; 29912.m005286)[31] and the recently identified phosphatidylcholine:diacylglycerol cholinephosphotransferase[32] (PDCT; 29841.m002865). Information on copy number, genomic context and regulatory regions of these and other metabolic genes will be important for the biotechnological transfer of ricinoleic acid production to established oilseed crops that lack ricin and its associated health risks. In addition, it is likely that the correct combination of specialized metabolic genes identified from the castor bean genome sequence will enable the engineering of triricinolein accumulation to amounts substantially higher than the modest levels achieved to date in model oilseeds[33,34].

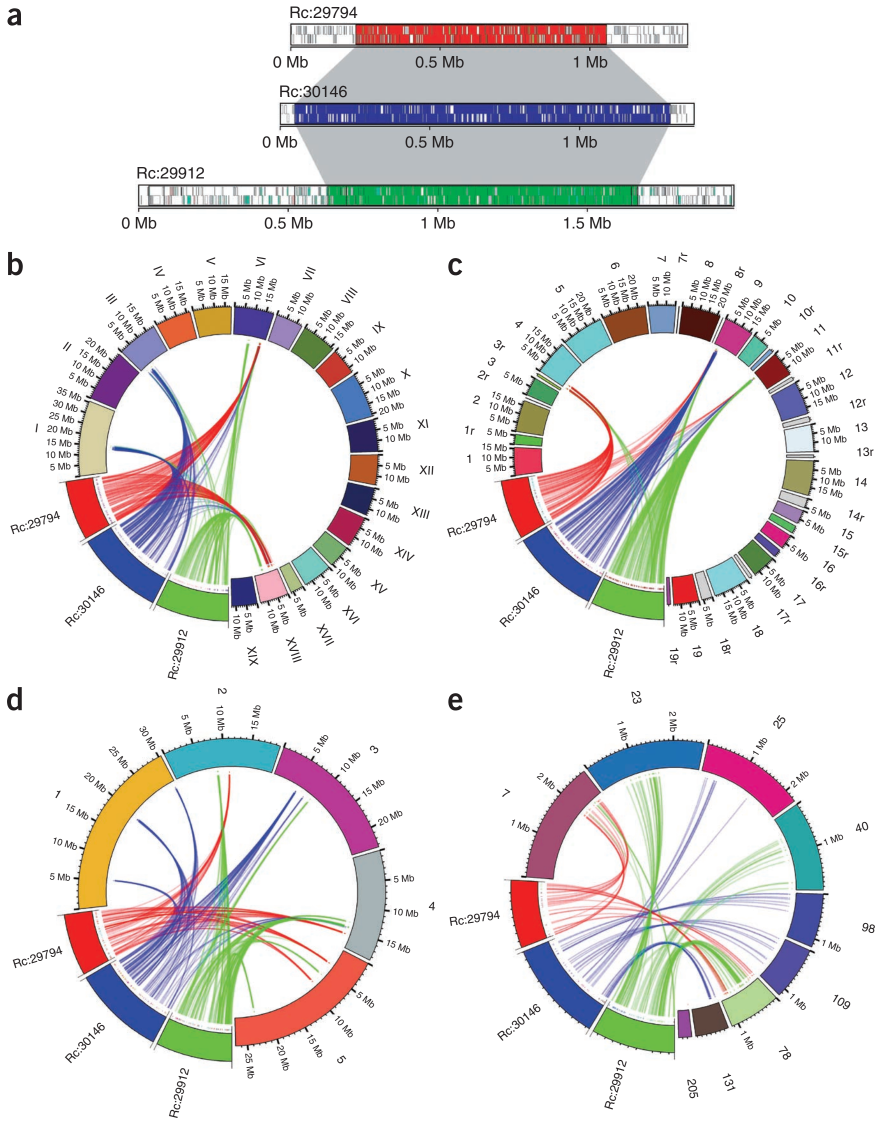


Figure 2 Collinearity between three paralogous castor bean genomic regions and their putative orthologs in other dicot genomes. (**a**) An example of a conserved paralogous triplication in the castor bean genome. (**b**–**e**) Putative orthologous gene pairs are shown as colored lines connecting the castor bean scaffolds (noted as Rc:scaffold number) to chromosomes or scaffolds in the other dicot genome. In most cases, one copy of the paralogous castor bean genes corresponds to two genes in poplar (**b**), one gene in grapevine (**c**) and four genes in *A. thaliana* (**d**). The castor bean–papaya relationship (**e**) is inconclusive. Numbers around the circles correspond to linkage group numbers (**b**), chromosome numbers (**c** and **d**) or scaffold numbers (**e**). Grapevine scaffolds that were mapped to chromosomes but their exact location is unknown are noted with an 'r' (random). The size of the castor bean genomic regions is proportional in all circles. Additional castor bean paralogous regions and their corresponding orthologs from other dicots are shown in **Supplementary Figure 3**.

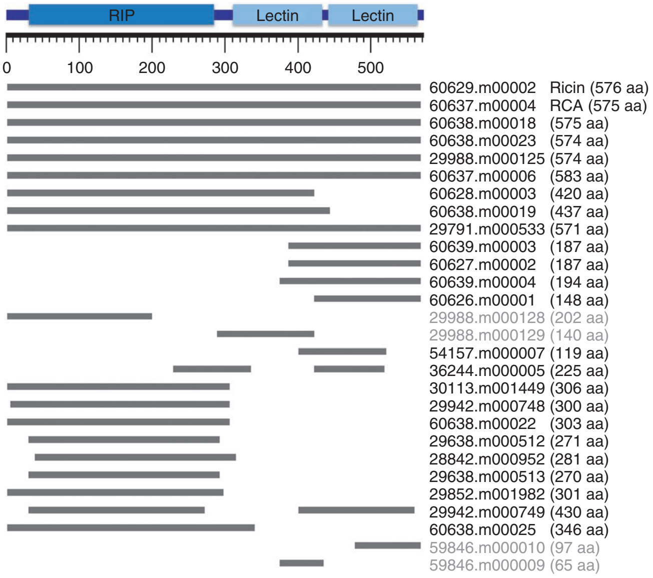


Figure 3 Schematic representation of the members of the ricin/RCA lectin gene family in castor bean. Ricin protein domains are represented at the top by blue boxes, and gray boxes represent protein sequences from this gene family aligned to the ricin precursor protein sequence used as reference. The ruler indicates the amino acid coordinates. The ricin and RCA genes are indicated and the amino acid sequence length for each gene model is shown in parenthesis. Pairs of adjacent gene models that could belong to a single pseudogene are shown in gray.

Disease resistance genes

To contribute to research aimed at understanding and improving biotic stress resistance in members of the Euphorbiaceae, especially for cassava[35], we compiled a list of predicted castor bean proteins with a functional annotation related to disease resistance. One hundred and twenty-one predicted disease-resistance proteins were identified (**Supplementary Table 5**) using our automated annotation pipeline. The majority of these predicted proteins belong to the nucleotide binding–leucine-rich repeat class, followed by the less common extracellular leucine-rich repeat–containing proteins[36], and dirigent-like proteins that have been associated with disease resistance[37]. The castor bean gene models coding for these resistance genes were found distributed in 69 scaffolds and were often found in clusters of genes from the same class. However, in some cases (for example, scaffold 30190), different resistance gene classes are found in the same cluster (**Supplementary Table 5**). These data will be useful for comparative studies on resistance genes in cassava, as well as other crop members of the Euphorbiaceae.

DISCUSSION

The sequence of the castor bean genome constitutes an important resource to study genome evolution, not only in the Euphorbiaceae family but also in plants in general. Besides its value for comparative genomics, and the insights it has yielded regarding synthesis of the highly toxic protein ricin[38] and the accumulation of castor oil, the castor bean genome promises to be invaluable in developing improved diagnostic and forensic methods for ricin detection and cultivar identification for tracing sample origins. Molecular diagnostic methods[39] and worldwide analyses of castor bean populations[1,2] have been reported and the availability of the castor bean genome sequence will accelerate efforts to advance such studies and technologies.

In addition to its relevance for biosecurity, availability of the castor bean genome could have implications for the production of biofuels and thus contribute to reducing greenhouse gas production. The industry of castor oil as a biodiesel component is being developed in Brazil[4], where the use of biofuels is highly advanced. Furthermore, castor oil can also be used as lubricity additive to replace sulfur-based lubricant components in petroleum diesel, helping to reduce sulfur emissions[40].

Unfortunately, the presence of ricin poses a problem for castor bean as a widely cultivated oilseed crop. Therefore, considerable effort has been directed to engineering ricinoleic acid production in seeds of the model plant *A. thaliana* as a prelude to transferring the required genes to an established ricin-free oilseed crop such as soybean. The initial strategy has involved the seed-specific expression of the castor bean *FAH* gene for the FAD2-related Δ12 oleic hydroxylase[26], the key enzyme in ricinoleic acid synthesis[41,42]. However, transgenic expression of *FAH* resulted in the accumulation of ricinoleic acid and other hydroxy fatty acids to only 15–20% of the total fatty acids in *A. thaliana* seeds[41,42]. Even co-expression of *FAH* with one additional ricinoleic acid metabolic gene, including the castor bean gene for DGAT2, yielded only small increases in ricinoleic acid accumulation in seeds of transgenic *A. thaliana* that were far less than the levels typically found in castor bean seeds[33,43]. These results also reflect the modest production of other unusual fatty acids that has been achieved by expression of FAD2 variants such as the Δ12 epoxygenase and fatty acid conjugases in seeds of transgenic plants[44,45]. These results suggest that expression of a single biosynthetic gene, such as *FAH* alone or together with a gene involved in the metabolism of a given unusual fatty acid, is insufficient to reproduce the oil composition observed in castor bean seeds. Thus, additional information on regulatory and metabolic genes is needed to fully transfer high levels of unusual fatty acid production and accumulation to engineered oilseed crops[43,46,47]. We believe that the castor bean genome sequence and its annotation constitute the foundation for identifying the regulatory and metabolic networks controlling castor-oil biosynthesis. When combined with metabolomics studies, these castor bean genome resources will enable metabolic engineering for improving castor oil production in crop plants lacking ricin.

Our analysis of the castor bean genome contributes to the debate on the polyploidization events that occurred in dicotyledonous genomes, supporting the presence of an ancestral hexaploidization event. Extending this type of analysis to cassava will benefit the cassava research community as it will synergize with the recently released genome sequence of cassava (http://www.phytozome.net/cassava), which is an important food and, more recently, industrial crop in poor, tropical countries. It has been proposed that cassava is an allopolyploid[48], and preliminary comparative genomics analyses between cassava and castor bean showed evidence of genomic duplications in cassava relative to castor bean (S. Rounsley, University of Arizona, Tucson, personal communication). These analyses suggest that the allopolyploidization event may have occurred in the cassava genome relatively recently, after the split between the two lineages. Further genome-wide comparative studies will provide insights on the genome evolution of cassava and the Euphorbiaceae family. Such information will help advance cassava breeding, which is a key means for developing countries to generate improved cassava lines with increased levels of stress resistance and nutritional content.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Accession codes. GenBank nuccore: AASG02000000 and GenBank gene: XP_002509419.1–XP_002540639.1. (The annotation data can also be freely accessed through the project's website (http://castorbean.jcvi.org/), which includes a genome browser and a BLAST server.)

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

This work was supported by the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health, Department of Health and Human Services, under NIAID Contract N01-AI-30071 (*R. communis* subproject) awarded to C.M.F.-L., J.R., J.R.W. and P.D.R.; Federal Bureau of Investigation grant J-FBI-04-186 to C.M.F.-L., J.R. and P.D.R.; and National Science Foundation grant DBI 0701919 to E.B.C. We thank the Joint Technology Center at J. Craig Venter Institute for carrying out all sequencing work, and K. Wurdack for his assistance with phylogenetics.

AUTHOR CONTRIBUTIONS

A.P.C., J.C., H.L., B.J.H. and J.R.W. performed genomic analyses. Q.Z., J.O. and M.S. conducted genome annotation. D.P. worked on the genome assembly. A.M.-B., K.M.J. and J.R. made DNA preparations, library constructions, and closure work. G.C., E.B.C. and M.G. performed manual annotations. C.M.F.-L. and J.R. conceived the project. P.D.R. conceived and directed the project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Allan, G. *et al.* Worldwide genotyping of castor bean germplasm (*Ricinus communis L.*) using AFLPs and SSRs. *Genet. Resour. Crop Evol.* **55**, 365–378 (2008).
2. Foster, J.T. *et al.* Single nucleotide polymorphisms for assessing genetic diversity in castor bean (*Ricinus communis*). *BMC Plant Biol.* **10**, 13 (2010).
3. da Silva Ramos, L.C., Shogiro Tango, J., Savi, A. & Leal, N.R. Variability for oil and fatty acid composition in castorbean varieties. *J. Am. Oil Chem. Soc.* **61**, 1841–1843 (1984).
4. da Silva Nde, L., Maciel, M.R., Batistella, C.B. & Maciel Filho, R. Optimization of biodiesel production from castor oil. *Appl. Biochem. Biotechnol.* **130**, 405–414 (2006).
5. Scarpa, A. & Guerci, A. Various uses of the castor oil plant (*Ricinus communis* L.). A review. *J. Ethnopharmacol.* **5**, 117–137 (1982).
6. Knight, B. Ricin–a potent homicidal poison. *BMJ* **1**, 350–351 (1979).
7. Lord, J.M., Roberts, L.M. & Robertus, J.D. Ricin: structure, mode of action, and some current applications. *FASEB J.* **8**, 201–208 (1994).
8. Schnell, R. *et al.* A Phase I study with an anti-CD30 ricin A-chain immunotoxin (Ki-4.dgA) in patients with refractory CD30+ Hodgkin's and non-Hodgkin's lymphoma. *Clin. Cancer Res.* **8**, 1779–1786 (2002).
9. Fidias, P., Grossbard, M. & Lynch, T.J. Jr. A phase II study of the immunotoxin N901-blocked ricin in small-cell lung cancer. *Clin. Lung Cancer* **3**, 219–222 (2002).
10. Endo, Y., Mitsui, K., Motizuki, M. & Tsurugi, K. The mechanism of action of ricin and related toxic lectins on eukaryotic ribosomes. The site and the characteristics of the modification in 28 S ribosomal RNA caused by the toxins. *J. Biol. Chem.* **262**, 5908–5912 (1987).
11. Macbeth, M.R. & Wool, I.G. Characterization of *in vitro* and *in vivo* mutations in non-conserved nucleotides in the ribosomal RNA recognition domain for the ribotoxins ricin and sarcin and the translation elongation factors. *J. Mol. Biol.* **285**, 567–580 (1999).
12. Lord, J.M., Hartley, M.R. & Roberts, L.M. Ribosome inactivating proteins of plants. *Semin. Cell Biol.* **2**, 15–22 (1991).
13. Lord, J.M. Synthesis and intracellular transport of lectin and storage protein precursors in endosperm from castor bean. *Eur. J. Biochem.* **146**, 403–409 (1985).
14. Roberts, L.M., Lamb, F.I., Pappin, D.J. & Lord, J.M. The primary sequence of Ricinus communis agglutinin. Comparison with ricin. *J. Biol. Chem.* **260**, 15682–15686 (1985).
15. Arumuganathan, K. & Earle, E.D. Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* **9**, 208–218 (1991).
16. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
17. Velasco, R. *et al.* A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* **2**, e1326 (2007).
18. Crabtree, J., Angiuoli, S.V., Wortman, J.R. & White, O.R. Sybil: methods and software for multiple genome comparison and visualization. *Methods Mol. Biol.* **408**, 93–108 (2007).
19. The *Arabidopsis* Genome Initiative Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
20. Tuskan, G.A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596–1604 (2006).
21. Ming, R. *et al.* The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **452**, 991–996 (2008).
22. Halling, K.C. *et al.* Genomic cloning and characterization of a ricin gene from *Ricinus communis*. *Nucleic Acids Res.* **13**, 8019–8033 (1985).
23. Tregear, J.W. & Roberts, L.M. The lectin gene family of *Ricinus communis*: cloning of a functional ricin gene and three lectin pseudogenes. *Plant Mol. Biol.* **18**, 515–525 (1992).
24. Leshin, J. *et al.* Characterization of ricin toxin family members from *Ricinus communis*. *Toxicon* **55**, 658–661 (2010).
25. McKeon, T.A., Chen, G.Q. & Lin, J.T. Biochemical aspects of castor oil biosynthesis. *Biochem. Soc. Trans.* **28**, 972–974 (2000).
26. van de Loo, F.J., Broun, P., Turner, S. & Somerville, C. An oleate 12-hydroxylase from *Ricinus communis* L. is a fatty acyl desaturase homolog. *Proc. Natl. Acad. Sci. USA* **92**, 6743–6747 (1995).
27. He, X., Turner, C., Chen, G.Q., Lin, J.T. & McKeon, T.A. Cloning and characterization of a cDNA encoding diacylglycerol acyltransferase from castor bean. *Lipids* **39**, 311–318 (2004).
28. Kroon, J.T., Wei, W., Simon, W.J. & Slabas, A.R. Identification and functional expression of a type 2 acyl-CoA:diacylglycerol acyltransferase (DGAT2) in developing castor bean seeds which has high homology to the major triglyceride biosynthetic enzyme of fungi and animals. *Phytochemistry* **67**, 2541–2549 (2006).
29. Saha, S., Enugutti, B., Rajakumari, S. & Rajasekharan, R. Cytosolic triacylglycerol biosynthetic pathway in oilseeds. Molecular cloning and expression of peanut cytosolic diacylglycerol acyltransferase. *Plant Physiol.* **141**, 1533–1543 (2006).
30. Thomaeus, S., Carlsson, A.S. & Stymne, S. Distribution of fatty acids in polar and neutral lipids during seed development in *Arabidopsis thaliana* genetically engineered to produce acetylenic, epoxy and hydroxy fatty acids. *Plant Sci.* **161**, 997–1003 (2001).
31. Dahlqvist, A. *et al.* Phospholipid:diacylglycerol acyltransferase: an enzyme that catalyzes the acyl-CoA-independent formation of triacylglycerol in yeast and plants. *Proc. Natl. Acad. Sci. USA* **97**, 6487–6492 (2000).
32. Lu, C., Xin, Z., Ren, Z., Miquel, M. & Browse, J. An enzyme regulating triacylglycerol composition is encoded by the ROD1 gene of *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* **106**, 18837–18842 (2009).
33. Burgal, J. *et al.* Metabolic engineering of hydroxy fatty acid production in plants: RcDGAT2 drives dramatic increases in ricinoleate levels in seed oil. *Plant Biotechnol. J.* **6**, 819–831 (2008).
34. Cahoon, E.B. *et al.* Engineering oilseeds for sustainable production of industrial and nutritional feedstocks: solving bottlenecks in fatty acid flux. *Curr. Opin. Plant Biol.* **10**, 236–244 (2007).
35. Hillocks, R.J. & Jennings, D.L. Cassava brown streak disease: a review of present knowledge and research needs. *Int. J. Pest Manage.* **49**, 225–234 (2003).
36. van Ooijen, G., van den Burg, H.A., Cornelissen, B.J. & Takken, F.L. Structure and function of resistance proteins in solanaceous plants. *Annu. Rev. Phytopathol.* **45**, 43–72 (2007).
37. Fristensky, B., Horovitz, D. & Hadwiger, L.A. cDNA sequences for pea disease resistance response genes. *Plant Mol. Biol.* **11**, 713–715 (1988).
38. Musshoff, F. & Madea, B. Ricin poisoning and forensic toxicology. *Drug Test Anal* **1**, 184–191 (2009).
39. Audi, J., Belson, M., Patel, M., Schier, J. & Osterloh, J. Ricin poisoning: a comprehensive review. *J. Am. Med. Assoc.* **294**, 2342–2351 (2005).
40. Goodrum, J.W. & Geller, D.P. Influence of fatty acid methyl esters from hydroxylated vegetable oils on diesel fuel lubricity. *Bioresour. Technol.* **96**, 851–855 (2005).
41. Broun, P. & Somerville, C. Accumulation of ricinoleic, lesquerolic, and densipolic acids in seeds of transgenic *Arabidopsis* plants that express a fatty acyl hydroxylase cDNA from castor bean. *Plant Physiol.* **113**, 933–942 (1997).
42. Smith, M.A., Moon, H., Chowrira, G. & Kunst, L. Heterologous expression of a fatty acid hydroxylase gene in developing seeds of *Arabidopsis thaliana*. *Planta* **217**, 507–516 (2003).
43. Lu, C., Fulda, M., Wallis, J.G. & Browse, J. A high-throughput screen for genes from castor that boost hydroxy fatty acid accumulation in seed oils of transgenic *Arabidopsis*. *Plant J.* **45**, 847–856 (2006).
44. Li, R., Yu, K., Hatanaka, T. & Hildebrand, D.F. Vernonia DGATs increase accumulation of epoxy fatty acids in oil. *Plant Biotechnol. J.* **8**, 184–195 (2010).
45. Cahoon, E.B. *et al.* Conjugated fatty acids accumulate to high levels in phospholipids of metabolically engineered soybean and *Arabidopsis* seeds. *Phytochemistry* **67**, 1166–1176 (2006).
46. Cernac, A. & Benning, C. WRINKLED1 encodes an AP2/EREB domain protein involved in the control of storage compound biosynthesis in *Arabidopsis*. *Plant J.* **40**, 575–585 (2004).
47. Thelen, J. & Ohlrogge, J. Metabolic engineering of fatty acid biosynthesis in plants. *Metab. Eng.* **4**, 12–21 (2002).
48. Umanah, E.E. & Hartmann, R.W. Chromosome numbers and karyotypes of some Manihot species. *Am. Soc. Hortic. Sci.* **98**, 272–274 (1973).

ONLINE METHODS

Whole genome shotgun sequencing. Castor bean inbred cultivar Hale[49] (NSL 4773) seeds were obtained from the National Center for Genetic Resources Preservation (NCGRP) at Ft. Collins, Colorado (Germplasm Resources Information Network). Nuclear DNA from etiolated castor bean seedlings grown in a growth chamber was purified as described[50] and was randomly sheared by nebulization, end-repaired with consecutive BAL31 nuclease and T4 DNA polymerase treatments and size-selected using gel electrophoresis on 1% low-melting-point agarose. After ligation to BstXI adapters, DNA was purified by three rounds of gel electrophoresis to remove excess adapters, and the fragments were ligated into the vector pHOS2 (a modified pBR322 vector) linearized with BstXI. The pHOS2 plasmid contains two BstXI cloning sites immediately flanked by sequencing-primer binding sites. Six libraries with small average insert size (3.5–9 kb) were constructed by electroporation of the ligation reaction into *E. coli.* strain GC10. In addition, two fosmid libraries were constructed using 30 μg of DNA that was sheared by bead beating and end-repaired (as described above). Fragments between 39 and 40 kb were isolated with a pulse field electrophoresis system and ligated to the blunt-end CopyControl pCC1FOS vector (Epicentre). Lambda phage packaging and infection were performed following the manufacturer's instructions. All clones were plated onto large format (16 × 16 cm) diffusion plates prepared by layering 150 ml of antibiotic-free Luria Bertani (LB)-agar onto a previously set 50-ml layer of LB-agar containing ampicillin or chloramphenicol as required by the vector. Colonies were picked for template preparation using Qbot or QPix colony-picking robots (Genetix), inoculated into 384-well blocks containing liquid medium and incubated overnight with shaking. High-purity plasmid DNA was prepared using the DNA purification robotic workstation custom-built by Thermo CRS and based on the alkaline lysis miniprep[51] and isopropanol precipitation. The DNA precipitate was washed with 70% ethanol, dried and resuspended in 10 mM Tris HCl buffer containing a trace of blue dextran. The typical yield of plasmid DNA from this method is ~600–800 ng per clone, providing sufficient DNA for at least four sequencing reactions per template. Sequencing was carried out using the di-deoxy sequencing method[52]. Two 384-well cycle-sequencing reaction plates were prepared from each plate of plasmid template DNA for opposite-end, paired-sequence reads. Sequencing reactions were completed using the Big Dye Terminator chemistry (Applied Biosystems) and standard M13 forward and reverse primers. Reaction mixtures, thermal cycling profiles and electrophoresis conditions were optimized to reduce the volume of the Big Dye Terminator mix and to extend read lengths on the AB3730xl sequencers (Applied Biosystems). Sequencing reactions were set up using a Biomek FX (Beckman Coulter) pipetting workstation. Robotics was used to aliquot and combine templates with reaction mixes consisting of deoxy- and fluorescently labeled di-deoxy-nucleotides, DNA polymerase, sequencing primers and reaction buffer in a 5 μl volume. Bar-coding and tracking systems promoted error-free template and reaction mix handling. After 30–40 consecutive cycles of amplification, reaction products were precipitated with isopropanol, dried at 25 °C, resuspended in water and transferred to an AB3730xl DNA analyzer.

A total of 2,276,000 paired-end sequence reads were attempted yielding 2,079,000 high-quality sequences, of which 12% correspond to fosmid clones (40 kb insert size), 60% to 9 kb insert size clones, 10% to 5 kb insert size clones and 18% to 3.5 kb insert clones. The average read-length was 839 bp. All reads were assembled into contigs using the Celera assembler[53] version 3.20 that utilizes an 'overlay-layout-consensus' approach to produce consensus sequences or contigs. Celera also uses mate-pair read information to build scaffolds where contigs are ordered and oriented relative to each other. The Celera assembler was run using the default parameters for large genomes. In addition to the normal contigs, the assembler creates so-called 'degenerate contigs' which have some kind of problem, such as excessive deviation from the expected level of coverage. We manually inspected the degenerate contigs and recovered ~12.4 Mb of sequences that contained plant gene-like sequences as determined by BLAST analysis. The consensus sequences were entered in an in-house genome annotation relational database called RCA1.

As the genomic DNA used for sequencing was purified from non-axenic seedlings, plant-associated bacteria were likely to be present in our sequence. Therefore, contigs smaller than 2 kb that did not show a high level of identity to plant organelle sequences (BLASTN E value cutoff $< 10^{-50}$), and showed sequence similarity to bacterial proteins from available bacterial genome sequences with BLASTX E values $< 10^{-20}$ were removed.

Closure of sequence gaps. To increase the quality of the ricin gene family annotation, we performed finishing work on eight scaffolds that contained members of this gene family to close sequence gaps or ambiguities within the corresponding gene models. Closure was conducted by editing the ends of sequence traces, primer walking on plasmid templates, sequencing genomic PCR products that spanned the gaps or by transposon insertion and sequencing of selected fosmids clones[54].

Gene prediction and genome annotation. All *R. communis* scaffolds were processed through the TIGR eukaryotic annotation pipeline. Before running the gene prediction software, we used RepeatMasker to mask the genomic sequence using a library of known plant repeats from an in-house plant repeat database and novel castor bean repeats identified by running RepeatScout, an algorithm that identifies sequences that are overrepresented in the assembly[55]. To prevent incorrect annotation of repeats as genes, we took a conservative approach and any sequence repeated at least ten times in the genome was considered repetitive. Manual inspection of the list of repeats generated by RepeatScout was carried out to remove members of known gene families that were wrongly reported as repeats. Further screening by manual review was carried out to remove putative gene families that were mistakenly identified as repeats, resulting in a final set of 1,517 consensus repeat sequences. With the so-constructed repeat library, 50.33% of the castor bean genome was masked as repetitive sequences. Low complexity sequences and tandem repeats were identified but not masked because they are often part of protein coding sequences. The RepeatScout library masked 49.88% of the genome whereas the known plant repeat library masked 8.24% of the genome. Repeats were classified using 2,994 Viridiplantae repeats from RepBase[56] and the consensus repetitive sequences identified by RepeatScout (**Table 2**).

Four gene finders were run on the masked genome: FgenesH gene prediction algorithm trained with a dicotyledonous matrix[57]; Augustus trained with *Arabidopsis*[58]; GlimmerHMM trained with *Arabidopsis*[59]; and SNAP trained with *Arabidopsis*[60].

We used PASA[61] to align 53,516 castor bean cDNA sequences to the castor bean genome. We used all available castor bean cDNA sequences from GenBank at the time, and 52,165 ESTs from five cDNA non-normalized libraries constructed from mRNAs from leaves, flowers, roots and two different seed developmental stages. cDNA clones were sequenced from the 5′ end, except for the root cDNA clones, which were sequenced from both ends to increase the chances of obtaining full-length cDNA sequences. PASA also assembles the aligned cDNA sequences into so-called 'PASA assemblies'. Using the unmasked castor genome sequence, PASA aligned and assembled ~73% of the castor bean cDNA sequences. For a cDNA sequence to be aligned to the genome, it should have at least 95% identity along 90% of its length, and consensus splice sites should be present at all inferred exon/intron boundaries. After alignment, PASA generated 8,132 nonredundant cDNA assemblies, of which 5,491 overlapped predicted gene models and 688 identified nonannotated regions. These PASA assemblies were used for identification of new gene models as well as to validate or update existing ones. Other PASA assemblies were not incorporated into gene models owing to intron/exon structure conflicts or because the fragmentary nature of the genome assembly precluded the alignments to meet the stringency criteria.

Sequence homology to nucleotide and protein datasets was computed using the Analysis and Annotation Tool (AAT) package[62] on the unmasked castor bean genome. AAT utilizes a two-step approach consisting of a fast database homology search followed by a rigorous, splice-aware local alignment. The datasets used for AAT analyses included: (i) *Oryza sativa* peptides (October 2006 release); (ii) *Arabidopsis* proteins (TAIR 6, September 2006 release); (iii) an in-house nonredundant amino acid database; (iv) a database of transcript assemblies that contains clustered and assembled ESTs and other cDNA sequences from plant species[63] for which over 1,000 sequences are available in GB (http://plantta.jcvi.org/).

Proteins having the highest scoring amino acid alignment to our gene models were incorporated into the gene models using GeneWise[64] to increase protein prediction reliability.

doi:10.1038/nbt.1674

All gene structures predicted by the methods described above as well as the alignments to protein and nucleotide databases were combined into consensus gene models using EVM[65], a software package developed at The Institute for Genomic Research (TIGR, now the J.C. Venter Institute or JCVI) that integrates data from multiple gene prediction programs as well as protein and cDNA similarity searches, to achieve the most accurate annotation possible with automated tools. It uses a nonstochastic, weighted-evidence combining technique that accounts for both the type and abundance of evidence to compute weighted consensus gene structures. All potential gene structure components were scored based on manually set weights so that exon and intron structures supported by PASA alignments and high-quality protein alignments had the highest relevance in determining a gene model's final structure, and the structure predicted by *ab initio* gene-finding software were given lower weights according to their accuracy for castor bean. Evidence from transcript assemblies alignments, protein alignments and gene prediction software were given a weight of 1, whereas GeneWise protein alignments received a weight of 5, and the weight of PASA assemblies was set at 20. Dynamic programming then was applied by EVM to find the highest scoring consensus gene structure, supported by all available evidence.

Gene models produced by EVM were then updated by new PASA assembly alignments. PASA extended untranslated regions and added small missed exons. This resulted in a total of 31,237 gene models of which 19,768 have either EST or protein support (5,316 gene models have castor bean EST support determined by PASA, and 16,848 have protein evidence support determined by AAT searches). 3,150 models were labeled as 'partial' because they missed either start and/or stop codons. 354 gene models contained an internal gap, which is represented by 'Ns' in the nucleotide sequence and 'Xs' in protein sequence, indicating the location and predicted size of the gap.

A dataset of 60 castor bean genes manually modeled based on highly conserved cDNA and protein alignments across multiple plant species were used as reference to evaluate the gene prediction algorithms' performance in comparison with EVM consensus predictions (**Supplementary Table 1**). Although this is a small set of genes, we used the exons to estimate the specificity and sensitivity of exon prediction by the different gene-finder programs as described[65]. Future iterations of the annotation can be improved by using a larger set of genes for training and evaluation of the gene prediction software, as more castor bean cDNA sequences become available.

Gene models were automatically named and their function was assigned by computationally extracting this information from BLASTP searches against the TAIR6 *Arabidopsis* peptides, Uniprot-Swissprot and experimentally verified Panda (http://www.ebi.ac.uk/panda), Panther (http://www.pantherdb.org/) and Interpro (http://www.ebi.ac.uk/interpro) databases. Gene models whose hits in those databases were defined as "unknown function" were labeled "conserved hypothetical protein" in our genome annotation. Gene models with no match in these databases above the selected threshold were labeled "hypothetical protein."

Automated Gene Ontology GO term assignments were done by extrapolating GO terms from matching *Arabidopsis* proteins using BLASTP with an E value threshold of 10^{-40}. Castor bean gene models with no match to *Arabidopsis* were screened against Pfam domains and assigned the Pfam associated GO term, if matches were above the selected cutoff. Altogether, this resulted in the assignment of 43,657 GO terms to 14,991 *R. communis* proteins.

Putative signal peptide sequences were identified using SignalP[66] and TargetP (http://www.cbs.dtu.dk/services/TargetP), and transmembrane regions were predicted by tmHMM[67]. Castor bean protein domains were also compared against the Pfam database of conserved families[68]. Proteins were organized into putative paralogous families based on conserved domain composition, taking into account both previously identified domains from public databases and potential novel domains identified using independent methods[69,70].

Noncoding RNAs were identified by searching against various RNA libraries. tRNAscan-SE[71] was run on the assembled genomic sequence to identify tRNAs. All 20 tRNAs were found in the genome with a total of 717 copies. rRNA sequences were annotated based on homology to previously published rRNA sequences in plants. snRNA were searched by blasting against the NONCODE database[72].

We assigned Enzyme Commission (EC) classification developed by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology, to provide metabolic pathway annotation. Castor bean proteins were searched against PRIAM profiles[73] using PSI-BLAST, and EC numbers were assigned for hits with an E value $< 10^{-10}$.

Annotation data are displayed in the project website (http://castorbean.jcvi.org/), which includes a generic genome browser (http://gmod.org/wiki/GBrowse), where gene models can be viewed in their sequence and genomic context. We used a gene model nomenclature that is composed by the scaffold ID number, followed by a period and the gene model number that consists of the letter 'm' followed by the gene model number. This number can be used to locate genes in the castor bean genome browser. Gene models in the genome browser are linked to Manatee pages, which include additional annotation information (http://manatee.sourceforge.net).

The castor bean predicted proteome could be matched to over 3,000 protein domains from Pfam[68], several of which are not present in *Arabidopsis* or poplar, including secondary metabolism genes (**Supplementary Fig. 1**). However, these results may have a substantial error due to inaccuracies of the automatic annotation both in poplar and castor bean.

We also searched for tandem gene duplications and found a total of 2,610 (8% of the total) genes forming part of tandem arrays.

Identification of genome duplications. A total of 167,984 predicted polypeptides from *R. communis*, *Vitis vinifera*, *Populus trichocarpa*, *Arabidopsis thaliana* and *Carica papaya* were subjected to an all-versus-all BLASTP analysis using WU-BLASTP 2.0MP, with the default BLOSUM62 substitution matrix, no low-complexity sequence filter, and an E-value cutoff of 10^{-5}. The castor bean subset of the BLAST results was analyzed to extract 5,536 pairs of castor genes that are reciprocal best hits and reside on distinct sequence contigs.

Each of the 721 (of 25,828) castor scaffolds with at least five annotated protein-coding genes was examined for runs of five or more genes that are collinear and are reciprocal best hits of collinear genes in another castor bean scaffold. Images were generated from these results and inspected manually for the presence of regions that appear to be triplicated in the castor bean genome, on the basis of overlapping runs of collinear matching gene pairs. The regions thus determined were also cross-checked against dot plots showing the relative positions of the paralogous gene pairs.

To further analyze these putative triplications four sets of Jaccard[74] orthologous (protein) clusters[18] were computed between castor and each of the four other genomes: Jaccard clusters were first defined within each genome by taking all BLASTP matches with E value $\leq 10^{-10}$, ≥80% identity and ≥70% sequence coverage and then forming clusters by transitively merging all pairs of proteins with Jaccard coefficient ≥0.6. In the second step, pairs of Jaccard clusters in distinct genomes were merged if each contained a protein with a best hit in the other cluster, taking into consideration only BLASTP matches with E value $\leq 10^{-10}$ and ≥70% sequence coverage (but imposing no other restriction on percent identity). For each triplication, Circos[75] was used to display the three castor regions and any collinear cluster matches between genes in those regions and those in the respective target genomes.

49. Brigham, R. Registration of castor variety Hale (Reg. No. 3). *Crop Sci.* **10**, 457 (1970).
50. Rabinowicz, P.D. *et al.* Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat. Genet.* **23**, 305–308 (1999).
51. Sambrook, J. & Russell, D.W. *Molecular Cloning. A Laboratory Manual* 3rd edn., (Cold Spring Harbor Laboratory Press, 2001).
52. Sanger, F., Nicklen, S. & Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467 (1977).
53. Myers, E.W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287**, 2196–2204 (2000).
54. Birren, B., Green, E.D., Klapholz, S., Myers, R.M. & Roskams, J. *Genome Analysis. A Laboratory Manual. Analyzing DNA* Vol. 1 (Cold Spring Harbor Laboratory Press, 1997).
55. Price, A.L., Jones, N.C. & Pevzner, P.A. De novo identification of repeat families in large genomes. *Bioinformatics* **21** Suppl 1, i351–i358 (2005).
56. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).

doi:10.1038/nbt.1674

57. Salamov, A.A. & Solovyev, V.V. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516–522 (2000).
58. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19** Suppl 2, ii215–ii225 (2003).
59. Majoros, W.H., Pertea, M. & Salzberg, S.L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
60. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
61. Haas, B.J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
62. Huang, X., Adams, M.D., Zhou, H. & Kerlavage, A.R. A tool for analyzing and annotating genomic sequences. *Genomics* **46**, 37–45 (1997).
63. Childs, K.L. *et al.* The TIGR Plant Transcript Assemblies database. *Nucleic Acids Res.* **35**, D846–D851 (2007).
64. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
65. Haas, B.J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
66. Bendtsen, J.D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783–795 (2004).
67. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).
68. Finn, R.D. *et al.* Pfam: clans, web tools and services. *Nucleic Acids Res.* **34**, D247–D251 (2006).
69. Haas, B.J. *et al.* Complete reannotation of the *Arabidopsis* genome: methods, tools, protocols and the final release. *BMC Biol.* **3**, 7 (2005).
70. Wortman, J.R. *et al.* Annotation of the *Arabidopsis* genome. *Plant Physiol.* **132**, 461–468 (2003).
71. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
72. He, S. *et al.* NONCODE v2.0: decoding the non-coding. *Nucleic Acids Res.* **36**, Database issue, D170–D172 (2008).
73. Claudel-Renard, C., Chevalet, C., Faraut, T. & Kahn, D. Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res.* **31**, 6633–6639 (2003).
74. Jaccard, P. The distribution of the flora in the alpine zone. *New Phytol.* **11**, 37–50 (1912).
75. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).

nature biotechnology

Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells

Sai T Reddy[1,2], Xin Ge[1], Aleksandr E Miklos[3,4], Randall A Hughes[3,4], Seung Hyun Kang[1], Kam Hon Hoi[2], Constantine Chrysostomou[1], Scott P Hunicke-Smith[3], Brent L Iverson[3,5], Philip W Tucker[3,6], Andrew D Ellington[3–5] & George Georgiou[1–3,6]

Isolation of antigen-specific monoclonal antibodies (mAbs) and antibody fragments relies on high-throughput screening of immortalized B cells[1,2] or recombinant antibody libraries[3–6]. We bypassed the screening step by using high-throughput DNA sequencing and bioinformatic analysis to mine antibody variable region (V)-gene repertoires from bone marrow plasma cells (BMPC) of immunized mice. BMPCs, which cannot be immortalized, produce the vast majority of circulating antibodies. We found that the V-gene repertoire of BMPCs becomes highly polarized after immunization, with the most abundant sequences represented at frequencies between ~1% and >10% of the total repertoire. We paired the most abundant variable heavy (V_H) and variable light (V_L) genes based on their relative frequencies, reconstructed them using automated gene synthesis, and expressed recombinant antibodies in bacteria or mammalian cells. Antibodies generated in this manner from six mice, each immunized with one of three antigens were overwhelmingly antigen specific (21/27 or 78%). Those generated from a mouse with high serum titers had nanomolar binding affinities.

The ability of the mammalian humoral immune response to generate a vastly diverse antibody repertoire in response to an antigen has been exploited for a range of biotechnology applications in diagnostics, therapy and basic research[7]. Since the development of the hybridoma technology by Kohler and Milstein 35 years ago[1], several methods for the generation of mAbs have been developed. Such methods include B-cell immortalization through genetic reprogramming by means of Epstein-Barr virus[8] or retrovirus-mediated gene transfer[2], cloning of V genes by single-cell PCR[9,10] and approaches for *in vitro* discovery that involve the display and screening of recombinant antibody libraries[3–6,11]. Both *in vitro* and *in vivo* methods for antibody discovery are critically dependent on high-throughput screening to determine antigen specificity. Recently, B-cell analysis has been expedited by soft lithography and microengraving techniques that allow for high-throughput identification of antigen-specific B cells[12,13]. However, this is at the cost of considerable technical complexity due to the need to amplify V genes and expand B cells. Similarly, the success of *in vitro* antibody discovery techniques depends on a range of screening parameters, which include the nature of the display platform, antigen concentration, binding avidity during enrichment, the number of rounds of screening (by panning or sorting), and the design and diversity of synthetic antibody libraries[7,14,15].

We have developed a simple and rapid method for antibody isolation without the need for screening. We exploited high-throughput DNA sequencing to analyze the V_L and V_H gene repertoires derived from the mRNA transcripts of fully differentiated mature B cells, antibody-secreting BMPCs, from immunized mice. After bioinformatic analysis, several abundant V_L and V_H gene sequences could be identified within the repertoire of each immunized mouse. V_L and V_H genes were paired according to their relative frequencies within the repertoire. Antibody genes were rapidly synthesized by oligonucleotide and PCR assembly by automated liquid-handling robots. Recombinant antibodies were expressed in bacterial and mammalian systems as single-chain variable fragments (scFv) and full-length IgG, respectively (**Fig. 1**). Finally, we confirmed that the resulting antibodies were overwhelmingly antigen specific (21/27 or 78%), thus confirming that our approach enables rapid and direct isolation of mAbs without screening.

B-cell maturation terminates with the formation of plasma cells, which represent <1% of all lymphoid cells but are responsible for the overwhelming majority of antibodies in circulation[16,17]. The bone marrow constitutes the major compartment where plasma cells reside and produce antibodies for prolonged periods of time, whereas plasma cells present in secondary lymphoid organs are often short lived. In mice, a stable and highly enriched antigen-specific BMPC population of ~10^5 cells (10–20% of all BMPCs) appears 6 d after secondary immunization and persists for prolonged periods[18]. In contrast, the increase in size of the splenic plasma cell population is highly transient, peaking at day 6 and rapidly declining to <10^4 cells by day 11. Notably, BMPCs are long lived and thus responsible for making the stable circulating population of antibodies in serum, which in turn is likely to play a dominant role in pathogen neutralization and other protective humoral immune responses[16].

To examine the dynamics of the V-gene repertoires in BMPCs, especially early after challenge with antigens, we immunized pairs

[1]Department of Chemical Engineering, University of Texas at Austin, Austin, Texas, USA. [2]Department of Biomedical Engineering, University of Texas at Austin, Austin, Texas, USA. [3]Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, Texas, USA. [4]Applied Research Laboratories, University of Texas at Austin, Austin, Texas, USA. [5]Department of Chemistry and Biochemistry, University of Texas at Austin, Austin, Texas, USA. [6]Section of Molecular Genetics and Microbiology, University of Texas at Austin, Austin, Texas, USA. Correspondence should be addressed to G.G. (gg@che.utexas.edu).

Received 11 June; accepted 2 August; published online 29 August 2010; doi:10.1038/nbt.1673

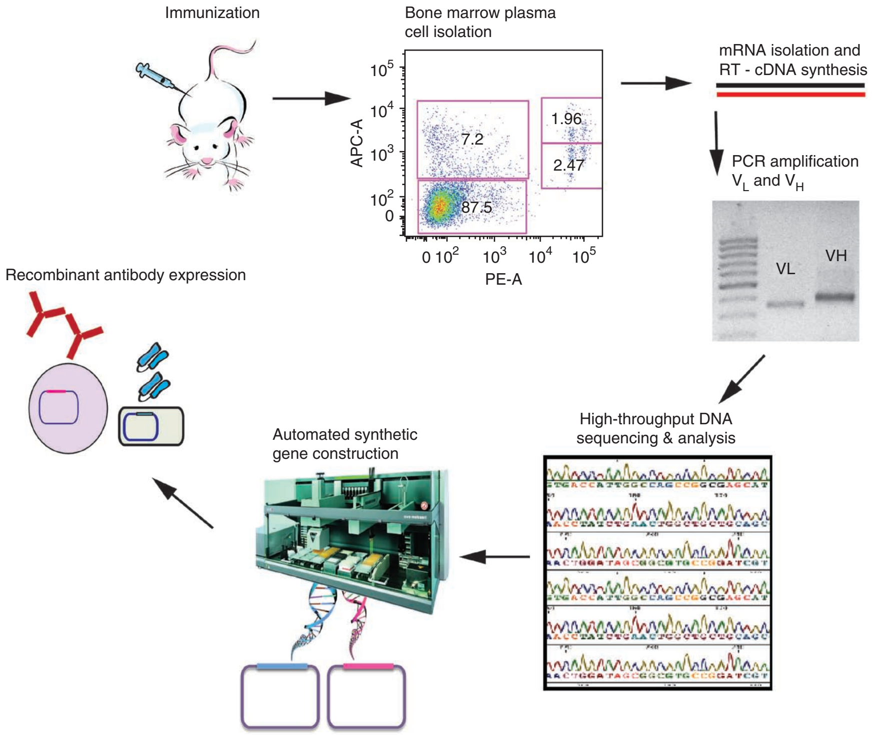

Immunization
Bone marrow plasma cell isolation
APC-A
10^5
10^4
10^3
10^2
0
7.2
1.96
2.47
87.5
0 10^2 10^3 10^4 10^5
PE-A
mRNA isolation and RT - cDNA synthesis
PCR amplification V_L and V_H
VL
VH
High-throughput DNA sequencing & analysis
Automated synthetic gene construction
Recombinant antibody expression

Figure 1 Isolation of monoclonal antibodies by mining the antibody variable (V)-gene repertoires of bone marrow plasma cells. Immunized mice are euthanized and $CD45R^-$ $CD138^+$ plasma cells are isolated. After mRNA isolation and first-strand cDNA synthesis, variable light (V_L) and variable heavy (V_H) gene DNA is generated. High-throughput 454 DNA sequencing and bioinformatic analysis is performed to determine the V_L and V_H repertoire. The most abundant V_L and V_H genes are identified and the sequences paired by using a simple relative-frequency rule. The respective antibody genes are synthesized using automated, robotically assisted gene synthesis. Finally, antigen-specific antibody single chain variable fragments or full-length IgGs are expressed in bacteria or mammalian cells, respectively. APC-A, CD45R-allophycocyanin-area; PE-A, CD138-R-phycoerythrin-area.

of mice with chicken egg ovalbumin, human complement serine protease (C1s), human B-cell regulator of IgH transcription (Bright) or adjuvant only. Antigen was co-injected with complete Freund's adjuvant followed by a secondary booster immunization in incomplete Freund's adjuvant. Mice were euthanized 6 d after secondary immunization and BMPCs ($CD45R^-$ $CD138^+$) were isolated to high purity (**Supplementary Fig. 1**). Total RNA was extracted and reverse transcribed for synthesis of first-strand cDNA. Well-characterized[11], degenerate V-gene primer mixes were used for second-strand amplifications, resulting in V_L and V_H PCR products of high purity (**Supplementary Fig. 2**), which were then submitted for high-throughput DNA sequencing of long reads using the Roche 454-GS FLX technology.

Unlike recent high-throughput sequencing analyses that explored V-gene repertoire diversity in zebrafish[19], humans[20,21] or synthetic libraries[22], our goals were to (i) identify highly expressed V genes whose products were likely to be antigen specific and (ii) determine the relative V-gene transcript abundance in the BMPC repertoires of immunized mice. These two tasks do not require exhaustive coverage of the V-gene repertoire; we have found that obtaining ~5,000 V-gene sequences per BMPC sample is sufficient to provide the information needed for antibody discovery, thus minimizing DNA sequencing costs. 454 reads were first processed by multiple sequence and signal filters, and then subjected to a simple and rapid bioinformatic analysis that relied on homologies to conserved framework regions within V genes to identify the most common complementarity determining region 3 (CDR3) sequences (**Supplementary Fig. 3**). This approach correctly identified ~94% of V_H and ~92% of V_L sequences in the Kabat database (**Supplementary Table 1**). Of a total of 415,018 reads, 23.2% contained CDR3 of V_H (CDRH3) and 26.6% contained CDR3 of V_L (CDRL3) sequences (**Supplementary Table 2**), representing 6,681–16,743 and 7,112–21,241 CDRH3 and CDRL3 sequences reads per mouse, respectively. For each mouse, frequency distributions of the CDR3s were calculated. Sequencing of the same samples, from separate cDNA library preparations by different facilities, gave quantitatively similar rankings for the abundances of CDR3 sequences. The same rank order frequencies are observed for all of the highly expressed CDR3s (**Supplementary Table 3**). This is important, because as discussed below, our approach for antibody discovery exploits the rank-order frequency of V genes to identify the most highly expressed clones. V-gene sequences containing a particular CDR3 were accepted as full length if they covered all three CDRs. Pairwise identities and frequencies were calculated by multiple sequence alignments, followed by germline analysis (**Supplementary Fig. 3** and Online Methods). A graphical user interface application was developed to enhance data analysis and visualization of the results (**Supplementary Fig. 4** and **Supplementary Data**).

Analysis of the BMPC repertoires from the six mice each immunized with one of the three antigens led to several interesting observations. First, in all immunized mice, including those receiving the same antigen, >92% of the CDRH3 sequences were unique to an individual mouse. The CDRL3 repertoires were less diverse, and in some instances, BMPCs from mice immunized with different antigens expressed high levels of the same CDRL3 (data not shown). A lower degree of V_L diversity, especially in early responses (as was the case here), is consistent with CDRL3 being derived from a single-gene recombination event (V-J), as opposed to two recombination events (V-D-J) for CDRH3. Second, and most importantly, ~10–20% of the total repertoire of all immunized mice were on average composed of only four CDRH3 sequences (**Supplementary Table 4**). For example, in the two mice immunized with C1s, the frequencies of the most abundant CDRH3s were 7.93% and 10.99% of the total repertoire. Third, as expected for early responses, the most highly abundant CDR3s were assembled from a diverse array of germline V-gene segments, with an average somatic mutation rate of only two and five amino acid substitutions for V_L and V_H, respectively (**Supplementary Tables 5** and **6**). Not surprisingly, certain germline V-gene families were represented preferentially in mice responding to particular antigens. For example, in mice immunized with C1s, 17.2% and 36.4% of the entire V_H-gene repertoire was composed of members of the *IGHV1* family, whereas the V_H-gene repertoire in mice injected only with adjuvant were dominated by sequences from the *IGHV5* or *IGHV6* families (**Supplementary Fig. 5**).

In most instances, the V genes encoding a highly abundant CDR3 were dominated by one sequence; the second most abundant V-gene sequence (somatic variant) was present at 10% the level and differed

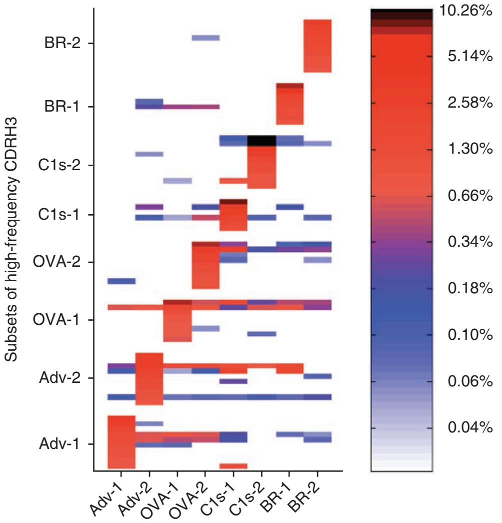


Figure 2 Comparison of high-frequency CDRH3s reveals unique V_H genes in each of eight mice immunized with one of three antigens or an adjuvant control. Heat map showing the distribution of highly represented CDRH3s in mice injected with Adjuvant (Adv), ovalbumin (OVA), C1s and Bright (BR). The *y* axis represents the ten highest frequency CDRH3 sequences identified in each mouse. The *x* axis compares the frequency of these prevalent CDRH3 sequences across all other mice. White, sequences found at frequencies that are not statistically significant (0.00–0.03%). Black, sequences found at a frequency of >10%.

from the dominant sequence by one or two amino acids. However, there were some instances where abundant CDRH3s were encoded by several V genes that were represented at comparable frequencies (**Supplementary Fig. 6** and **Supplementary Table 7**). Notably, the V_H repertoires were quite distinct even among genetically identical littermates immunized with the same antigen on the same day. For mice immunized with C1s or Bright, each mouse developed a distinct and diverse set of abundant CDRH3 sequences (**Fig. 2** and **Supplementary Table 8**). This suggests that each mouse generates its own unique and highly expressed V_H-gene repertoire, which may allow for the discovery of a panel of diverse antibodies. One exception, however, was that a few CDRH3 sequences abundant in both ovalbumin-immunized mice were also present at high frequency in other mice, suggesting that the corresponding antibodies may be polyspecific. Not surprisingly, some CDRH3 sequences from animals that received adjuvant only, found at moderate levels, were also present in immunized mice (**Fig. 2**). Antibodies encoding these sequences were probably specific to adjuvant or to common natural antigens. CDRL3 diversity was lower, with several promiscuous sequences represented at high frequency in several mice (**Supplementary Table 9**). Fourth, even though the BMPC V_H repertoires were largely composed of sequences unique to each mouse, principal component analysis of CDRH3s shared between mice revealed distinct clustering of the data for each cohort (that is, same cage and litter) immunized at the same time but with different antigens (**Supplementary Fig. 7**). This signature likely reflects environmental factors, such as the antigenic history of the animal groups, and suggests that V-gene repertoire analysis may provide valuable diagnostic information.

It should be noted that a few copies (typically <5) of the most abundant CDRH3 sequences raised to a given antigen were observed at very low levels (typically <0.1%) in the CDRH3 repertoires of mice receiving other antigens. As several of the respective V genes were shown to encode antigen-specific antibodies (see below), we believe that the presence of these sequences in mice immunized with other antigens might originate from low levels of cross-sample contamination, a conclusion supported by the biased distributions of common CDRH3 sequences within the same cohort (**Supplementary Fig. 8**). Because of the high sensitivity of 454 DNA sequencing, even with the utmost care it is not possible to completely rule out low-level contamination (sequence noise) during library preparation/multiplex sequencing. Although an important consideration for studies aiming to compare unbiased repertoires[19,20], sequence noise does not affect the methodology described herein, as the most abundant V genes in the BMPC repertoire are 20- to >100-fold more abundant than the sequence noise level.

Manual screening of small combinatorial libraries of scFvs in *Escherichia coli* using BMPC V genes led to a low yield of antigen-specific clones (less than four positive clones per 96-well plate; data not shown). Upon further analysis, most of these scFvs displayed low apparent affinity by enzyme-linked immunosorbent assay (ELISA) and/or poor expression and aggregation. We reasoned that this was a consequence of combinatorial pairing; even if a V_L and a V_H gene represented 5% of the cDNA pool, assuming no PCR biases in scFv assembly, the probability of correct pairing is only 0.25%. Discovery of positive clones would thus require an extensive amount of screening. To overcome these problems, and to avoid screening altogether, we hypothesized that V_L and V_H genes represented at approximately the same frequency likely arise from the same plasma cell and, hence, are naturally paired. To test this hypothesis, we synthesized the four or five most abundant full-length V_L and V_H genes from each mouse (excluding V_H sequences that were cross-represented in adjuvant-only mice), which accounted for a minimum of 0.5% of the repertoire, expressed the recombinant antibodies, and tested for antigen binding. Synthetic genes were constructed by robotically assisted, high-throughput DNA synthesis (Online Methods). Briefly, gene fragments (200–500 nucleotides long) were generated using inside-out nucleation PCR reactions. The design of these fragments and relevant overlaps was automated using customized software to facilitate robotic synthesis and assembly (**Supplementary Data**). Alignment and 'padding' of the sequences at either end yielded genes of identical length and permitted the use of a generic overlapping assembly strategy that ensured the greatest oligonucleotide reuse (**Supplementary Fig. 9**). In this manner, up to 48 V_L and 48 V_H genes could be synthesized and validated for the correct open reading frame by one researcher within 1 week, at a reagent cost of <$2,000. In most cases, V_L and V_H pairing was determined by rank ordering of CDR3 frequency within the repertoire. In cases where two V_L or V_H genes were found at very similar frequencies, we constructed multiple V_L-V_H combinations. Paired V genes were then expressed as scFv fragments in *E. coli*. ELISA analysis of bacterial lysates indicated that the resulting antibodies were overwhelmingly antigen specific (~78%): we obtained 21/27 antigen-specific antibodies from six mice immunized with three different protein antigens (**Table 1**). To further evaluate the utility of this simple pairing strategy, we constructed a combinatorial library of scFvs comprising the four most abundant V_L and V_H genes from each of the two mice immunized with C1s. scFv antibodies were expressed in *E. coli*. Binding analysis by ELISA revealed that all of the highest antigen-binding clones possessed the same V_L-V_H gene combinations predicted by our pairing strategy (**Supplementary Table 10**).

As mouse 2 immunized with C1s (C1s-2) displayed the highest serum titers (**Supplementary Table 11**), we selected antibodies from this animal for biophysical characterization of antigen binding affinity by surface plasmon resonance. Antibodies were expressed from synthetic genes and purified as monomeric scFv fragments in

Table 1 Antigen binding of antibody single-chain variable fragments (scFvs) from high frequency V_L and V_H genes

V_L-V_H pair	% V_L	CDRL3	% V_H	CDRH3	scFv binding
α–OVA					
1.1L-1.1H	11.70	WQGTHFPLT	7.11	GSSYYAMDY	+
1.2L-1.2H	4.40	QQYNSYPLT	1.10	LLWLYAMDY	+
1.3L-1.3H	3.38	QQSNSWYT	0.57	DVYDGYAMDY	+
1.4L-1.4H	2.20	QHHYGTPPWT	0.54	NPYAMDY	–
2.1L-2.1H	5.32	WQGTHFPLT	7.61	RTTVSRDWYFDV	+
2.2L-2.2H	4.05	QQYNSYPLT	3.23	YYYGSSAMDY	+
2.3L-2.3H	3.46	QQYSSYPLT	2.22	DGWYYFDY	+
2.4L-2.4H	2.01	QQHYSTPWT	2.10	EDDYDLFAY	+
α–C1s					
1.1L-1.1H	12.95	WQGTHFPQT	7.93	GNYYYAMDY	+
1.2L-1.1H	6.94	QQWSSYPQLT	7.93	GNYYYAMDY	+
1.3L-1.2H	3.81	QNDHSYPLT	2.64	DMISYWYFDV	+
1.4L-1.3H	3.16	QQGQSYPFT	1.67	EDYGNYWYFDV	+
1.4L-1.4H	3.16	QQGQSYPFT	1.67	EGYYYGSSYFDY	–
2.1L-2.1HA	17.10	FQGSHVPLT	10.99	SDRYDGYFDY	+
2.1L-2.1HB	17.10	FQGSHVPLT	9.93	SDRFDGYFDY	+
2.2L-2.2H	2.62	QQSNEDPWT	3.30	WLLLAY	+
2.3L-2.2H	2.20	WQGTHFPH	3.30	WLLLAY	+
2.3L-2.3H	2.20	WQGTHFPH	1.65	SDGYYYFDY	+
2.4L-2.4H	1.64	QQHYSTPFT	1.15	YYDYDKAYYFDY	–
α–Br					
1.1L-1.1H	6.64	LQYASSPFT	7.20	HDYGNYVDY	+
1.2L-1.2H	4.73	WQGTHFPRT	5.62	DGNYQEDYFDY	–
1.3L-1.3H	4.51	QQNNEDPRT	1.91	EGYAYDVDY	+
1.4L-1.4H	3.59	QQRSSYPLT	1.20	YDYGKDFDY	+
2.1L-2.1H	7.24	WQGTHFPQT	2.57	RGDGNYFFDY	+
2.2L-2.2H	4.50	QQGQSYPWT	2.27	GDEAWFAY	–
2.3L-2.3H	3.12	LQYASSPYT	2.03	EGDFDY	–
2.4L-2.4H	2.58	FQGSHVPWT	1.63	GGNYDYAMDY	+

E. coli whole-cell lysates expressing antibody scFvs that were constructed by pairing the most abundant V genes (as shown above). V_L and V_H gene pairing was determined by relative frequency (%) of the respective V genes in the BMPC repertoires. ELISA analysis was performed to determine antigen binding (Online Methods). +, more than threefold stronger ELISA signal on antigen-coated wells relative to wells coated with unrelated antigen (BSA and/or gelatin). OVA, ovalbumin; BR, bright.

E. coli and as full-length IgG antibodies in HEK 293F cells. Pairing of the most abundant light (2.1L) and heavy (2.1H-B) V genes (frequencies, 17.10% CDRL3 and 9.93% CDRH3) from mouse C1s-2 yielded an antibody with a K_D of 20 nM as a scFv ($k_{on} = 2.3 \times 10^4$ M^{-1} sec^{-1}; $k_{off} = 5.0 \times 10^{-4}$ sec^{-1}) and unexpectedly, a slightly lower monovalent K_D of 50 nM ($k_{on} = 2.4 \times 10^4$ M^{-1} sec^{-1}; $k_{off} = 1.2 \times 10^{-3}$ sec^{-1}) as an IgG. From the same mouse, pairing of C1s genes 2.2L with 2.2H (frequencies, 2.62% CDRL3 and 3.30% CDRH3) resulted in an IgG that displayed low binding affinity (K_D of ~500 nM, data not shown). However, the pairing of C1s genes 2.3L with 2.2H (frequencies, 2.20% CDRL3 and 3.30% CDRH3) generated an IgG with subnanomolar binding affinity ($K_D = 0.43$ nM; $k_{on} = 4.5 \times 10^5$ M^{-1} sec^{-1}; $k_{off} = 1.9 \times 10^{-4}$ sec^{-1}; **Supplementary Fig. 10** and **Supplementary Table 12**), indicating that the natural pairing is likely 2.3L–2.2H. Furthermore, the antibodies were suitable for functional assays, such as sandwich ELISA and immunoprecipitation of C1s from human serum (**Supplementary Figs. 11** and **12**).

Our approach capitalizes on mining the antibody repertoire of BMPCs, a population of B cells that is responsible for the synthesis of the large majority of circulating immunoglobulins in animals[16]. Although we have validated this methodology in mice, there is no reason to believe that the same approach cannot be readily extended to primates, including humans. Furthermore, it is possible that this technology could be extended for antibody discovery with more complex antigens such as viral and bacterial pathogens, as in these situations BMPCs may still develop polarity. We note, however, that the polarization of the BMPC repertoire in instances when the antigen may contain multiple highly immunogenic epitopes requires further evaluation. The mechanisms that dictate the selection of B-cell differentiation into plasma cells and homing into the bone marrow are complex and appear to partially relate to high antigen affinity[23,24]. As the highly abundant BMPCs correspond to abundant circulating antibodies, it seems plausible to hypothesize that these antibodies have been selected by the immune system (at least partly) because they display more potent pathogen neutralization. Therefore, antibodies generated by the mining of the BMPC repertoire may prove particularly useful for therapeutic purposes. The hybridoma technology and other B-cell immortalization methods interrogate the antibody-producing cells in pre–plasma cell B-cell populations, specifically in memory B cells, or in circulating short-lived plasmablasts[9]. Fully differentiated plasma cells are not amenable to most of these analyses, as they do not survive outside their biological niches. Very recently, microwell arrays and single-cell cloning were used to isolate antibodies from spleen plasma cells[12]. Nonetheless, despite the use of a sophisticated screening technology, only small numbers of antigen-specific clones could be isolated. Consequently, information on the repertoire and relative abundance of V genes could not be obtained by this method.

Our use of high-throughput DNA sequencing, bioinformatic analysis and automated gene synthesis can lead to the isolation and expression of mAbs with minimal effort. In our hands, it takes ~10 person hours for sample preparation for DNA sequencing. With automated bioinformatic processing of the 454 sequencing data, no extra effort is required to identify highly abundant V_L and V_H genes for DNA synthesis. Synthetic genes can be constructed either by an automated facility (as described herein) or through commercial gene-synthesis vendors. Furthermore, antibody genes can be codon optimized as desired for either bacterial or mammalian expression and subsequent characterization studies. Thus, in terms of effort by dedicated personnel (not including DNA sequencing and synthesis, which are carried out by multi-user services) and time line required for antibody discovery, our method compares very favorably to methods involving hybridomas, B-cell immortalization, and B-cell screening and/or single-cell cloning. Currently, the most expensive part of our antibody discovery process is DNA sequencing followed by gene synthesis. However, the cost for these technologies is declining at a rapid and exponential pace, resembling Moore's law for microelectronics[25,26]. Taken within this context, we envisage that the expense for our approach to antibody discovery will eventually not be a limitation.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Accession codes. GenBank: HQ15046,47,49,51-91.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank C. Das for her invaluable technical assistance on mammalian cell culture and transfection and M. Byrom for assistance with mouse experiments. This work was funded by grants from the Clayton Foundation, the Cockrell Chair in Engineering (to G.G.) and by a fellowship from Natural Sciences and Engineering Research Council of Canada (to X.G.).

AUTHOR CONTRIBUTIONS

S.T.R. and G.G. developed the methodology, designed experiments, analyzed the data and wrote the manuscript; X.G., C.C., and K.H.H. carried out the bioinformatic analysis; S.T.R., A.E.M., R.A.H., S.H.K. and K.H.H. performed the experiments; S.P.H.-S. performed 454 DNA sequencing; B.L.I., P.W.T. and A.D.E. helped analyze the data.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Köhler, G. & Milstein, C. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* **256**, 495–497 (1975).
2. Kwakkenbos, M.J. *et al.* Generation of stable monoclonal antibody-producing B cell receptor-positive human memory B cells by genetic programming. *Nat. Med.* **16**, 123–128 (2010).
3. Clackson, T. *et al.* Making antibody fragments using phage display libraries. *Nature* **352**, 624–628 (1991).
4. Feldhaus, M.J. *et al.* Flow-cytometric isolation of human antibodies from a nonimmune *Saccharomyces cerevisiae* surface display library. *Nat. Biotechnol.* **21**, 163–170 (2003).
5. Harvey, B.R. *et al.* Anchored periplasmic expression, a versatile technology for the isolation of high-affinity antibodies from *Escherichia coli*-expressed libraries. *Proc. Natl. Acad. Sci. USA* **101**, 9193–9198 (2004).
6. Schaffitzel, C., Hanes, J., Jermutus, L. & Plückthun, A. Ribosome display: an in vitro method for selection and evolution of antibodies from libraries. *J. Immunol. Methods* **231**, 119–135 (1999).
7. Hoogenboom, H.R. Selecting and screening recombinant antibody libraries. *Nat. Biotechnol.* **23**, 1105–1116 (2005).
8. Traggiai, E. *et al.* An efficient method to make human monoclonal antibodies from memory B cells: potent neutralization of SARS coronavirus. *Nat. Med.* **10**, 871–875 (2004).
9. Wrammert, J. *et al.* Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nature* **453**, 667–671 (2008).
10. Meijer, P.-J. *et al.* Isolation of human antibody repertoires with preservation of the natural heavy and light chain pairing. *J. Mol. Biol.* **358**, 764–772 (2006).
11. Mazor, Y., Blarcom, T.V., Mabry, R., Iverson, B.L. & Georgiou, G. Isolation of engineered, full-length antibodies from libraries expressed in *Escherichia coli. Nat. Biotechnol.* **25**, 563–565 (2007).
12. Jin, A. *et al.* A rapid and efficient single-cell manipulation method for screening antigen-specific antibody-secreting cells from human peripheral blood. *Nat. Med.* **15**, 1088–1092 (2009).
13. Love, J.C., Ronan, J.L., Grotenbreg, G.M., van der Veen, A.G. & Ploegh, H.L. A microengraving method for rapid selection of single cells producing antigen-specific antibodies. *Nat. Biotechnol.* **24**, 703–707 (2006).
14. Cobaugh, C.W., Almagro, J.C., Pogson, M., Iverson, B. & Georgiou, G. Synthetic antibody libraries focused towards peptide ligands. *J. Mol. Biol.* **378**, 622–633 (2008).
15. Persson, H., Lantto, J. & Ohlin, M. A focused antibody library for improved hapten recognition. *J. Mol. Biol.* **357**, 607–620 (2006).
16. Manz, R.A., Hauser, A.E., Hiepe, F. & Radbruch, A. Maintenance of serum antibody levels. *Annu. Rev. Immunol.* **23**, 367–386 (2005).
17. Shapiro-Shelef, M. & Calame, K. Regulation of plasma-cell development. *Nat. Rev. Immunol.* **5**, 230–242 (2005).
18. Manz, R.A., Thiel, A. & Radbruch, A. Lifetime of plasma cells in the bone marrow. *Nature* **388**, 133–134 (1997).
19. Weinstein, J.A., Jiang, N., White, R.A., Fisher, D.S. & Quake, S.R. High-throughput sequencing of the zebrafish antibody repertoire. *Science* **324**, 807–810 (2009).
20. Boyd, S.D. *et al.* Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci. Transl. Med.* **1**, 12ra23 (2009).
21. Glanville, J. *et al.* Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci. USA* **106**, 20216–20221 (2009).
22. Ge, X., Mazor, Y., Hunicke-Smith, S.P., Ellington, A.D. & Georgiou, G. Rapid construction and characterization of synthetic antibody libraries without DNA amplification. *Biotechnol. Bioeng.* **106**, 347–357 (2010).
23. Radbruch, A. *et al.* Competence and competition: the challenge of becoming a long-lived plasma cell. *Nat. Rev. Immunol.* **6**, 741–750 (2006).
24. Phan, T.G. *et al.* High affinity germinal center B cells are actively selected into the plasma cell compartment. *J. Exp. Med.* **203**, 2419–2424 (2006).
25. Carlson, R. The changing economics of DNA synthesis. *Nat. Biotechnol.* **27**, 1091–1094 (2009).
26. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).

ONLINE METHODS

Immunization. Purified C1s (CalBiochem), purified chicken egg ovalbumin (Sigma), or recombinant bacterially expressed human B-cell regulator of IgH transcription (Bright) were resuspended in sterile-filtered PBS at 1.0 mg/ml. On the day of primary immunization, 25 μl of antigen solution was thoroughly mixed with 25 μl of complete Freund's adjuvant (CFA; Pierce Biotechnology) and 50 μl of sterile PBS and stored on ice. Female BALB/c mice (Charles Rivers Laboratories) 6–8 weeks old were housed in conventional barrier space and were maintained on a normal chow diet. Before injections, mice were bled from the tail vein and ~25 μl of blood was collected and stored at −20 °C for later analysis. Day 1 was designated as the day primary immunizations were performed. 100 μl of the antigen-CFA mixture per mouse was injected with a 26-gauge needle subcutaneously into the backpad. Mice were monitored daily by animal housing staff and cages were changed twice per week.

For secondary immunization, 25 μl of antigen solution was thoroughly mixed with 25 μl of incomplete Freund's adjuvant (IFA; Pierce Biotechnology) and 50 μl of sterile PBS and stored on ice. On day 21, mice were given the secondary immunization intraperitoneally at 100 μl of antigen-IFA mixture per mouse. On day 26, mice were euthanized by CO_2 asphyxiation and blood, femurs and tibia were collected.

Isolation of BMPCs. Muscle and fat tissue were removed from the harvested tibias and femurs. The ends of both tibia and femurs were clipped with surgical scissors and bone marrow was flushed out with a 26-gauge insulin syringe (Becton Dickinson, BD). Bone marrow tissue was collected in sterile-filtered buffer no. 1 (PBS with 0.1% BSA/2 mM EDTA). Bone marrow cells were collected by filtration through a 70-μm cell strainer (BD) with mechanical disruption and washed with 20 ml of PBS and collected in a 50 ml tube (Falcon, BD). Bone marrow cells were then centrifuged at 335*g* for 10 min at 4 °C. Supernatant was decanted and the cell pellet was resuspended with 3.0 ml of red blood cell lysis buffer (eBioscience) and shaken gently at 25 °C for 5 min. Cell suspension was then diluted with 20 ml of PBS and centrifuged at 335*g* for 10 min at 4 °C. Supernatant was decanted and cell pellet was resuspended in 1.0 ml of buffer no. 1.

Each isolated bone marrow cell suspension was incubated with 2.5 μg and 1.5 μg of biotinylated rat anti-mouse CD45R(B220) and biotinylated rat anti-mouse CD49b (eBioscience), respectively. Cell suspension was rotated at 4 °C for 20 min. Cell suspensions were then centrifuged at 930*g* for 6 min at 4 °C, supernatant was removed and the cell pellet was resuspended in 1.5 ml of buffer no. 1. Streptavidin conjugated M280 magnetic beads (Invitrogen) were washed and resuspended according to manufacturer's protocol. 50 μl of magnetic beads were added to each cell suspension and the mixture was rotated at 4 °C for 20 min. Cell suspensions were then placed on Dynabead magnet (Invitrogen) and supernatants (negative fraction, cells unconjugated to beads) were collected and cells bound to beads were discarded.

Prewashed streptavidin M280 magnetic beads were incubated for 30 min at 4 °C with biotinylated rat anti-mouse CD138 (BD Pharmingen) with 0.75 μg antibody per 25 μl of magnetic beads. Beads were then washed according to manufacturer's protocol and resuspended in buffer no. 1. The negative cell fraction (depleted of $CD45R^+$ and $CD49b^+$ cells) collected as above was incubated with 50 μl of CD138-conjugated magnetic beads and the suspension rotated at 4 °C for 30 min. Beads with $CD138^+$ bound cells were isolated by the magnet, washed 3 times with buffer no. 1, the negative ($CD138^−$) cells unbound to beads were discarded (or saved only for analysis). The positive $CD138^+$ bead-bound cells were collected and stored at 4 °C until further processed.

Preparation of V_L and V_H genes. $CD45R^−$ $CD138^+$ BMPCs isolated as described herein were centrifuged at 930*g* at 4 °C for 5 min. Cells were then lysed with TRI reagent and total RNA was isolated according to the manufacturer's protocol in the Ribopure RNA isolation kit (Ambion). mRNA was isolated from total RNA with oligodT resin and the Poly(A) purist kit (Ambion) according to the manufacturer's protocol. mRNA concentration was measured with an ND-1000 spectrophotometer (Nanodrop).

The isolated mRNA was used for first-strand cDNA synthesis by reverse transcription with the Maloney murine leukemia virus reverse transcriptase (MMLV-RT, Ambion). cDNA synthesis was performed by RT-PCR using 50 ng of mRNA template and oligo(dT) primers according to manufacturer's protocol of Retroscript kit (Ambion). After cDNA construction, PCR amplification was performed to amplify the V_L and V_H genes with a standard mix of degenerate primers[27]. A complete list of primers can be found in **Supplementary Table 13**. A 50 μl PCR reaction consisted of 0.2 mM of forward and reverse primer mixes, 5 μl of Thermopol buffer (NEB), 2 μl of unpurified cDNA, 1 μl of Taq DNA polymerase (NEB) and 39 μl of double-distilled H_2O. The PCR thermocycle program was: 92 °C for 3 min; 4 cycles (92 °C for 1 min, 50 °C for 1 min, 72 °C for 1 min); 4 cycles (92 °C for 1 min, 55 °C for 1 min, 72 °C for 1 min); 20 cycles (92 °C for 1 min, 63 °C for 1 min, 72 °C for 1 min); 72 °C for 7 min; 4 °C storage. PCR gene products were gel purified and submitted to SeqWright and Genomic Sequencing and Analysis Center at the University of Texas Austin for Roche GS-FLX 454 DNA sequencing.

High-throughput sequencing of V_L and V_H repertoires. V-gene repertoires isolated from BMPC of eight mice were sequenced using high-throughput 454 GS-FLX sequencing (University of Texas, Austin, TX; SeqWright). In total, 415,018 sequences were generated, and 454 data quality-control filtered and grouped >97% of the sequences into data sets for each mouse according to their Multiplex Identifiers usages.

Bioinformatic analysis. (1) CDR3 identification. A search method was developed based on conserved flanking sequence motifs found upstream and downstream of CDR3. Searching motifs for CDRH3 and CDRL3 were determined based on amino acids that occur with an average frequency of 99% at specific positions in V genes from the Kabat database. (**Supplementary Table 1**). V_H sequences were searched for the motif DXXX(Y/F)(Y/F)C (Kabat # 86-92) and WGXG(T/S) (Kabat # 103-107) at N- and C- termini of CDRH3, respectively. Analogously, V_L genes were found by searching for the motifs DXXXY[F/Y]C (Kabat # 82-88) and FGXGT (Kabat # 98-102). This approach correctly identifies >94% of V_H and 92% of V_L full-length sequences in the Kabat database. Any sequences or reverse complements containing these motifs were extracted as either V_H or V_L genes, respectively. Only the sequences with in-frame CDR3 and without stop codons were further analyzed. For each sample, the most highly represented CDR3 sequences (typically represented at frequencies >1%) were discovered, and their relative abundances in all the other seven samples were calculated. To find a consensus full-length V_H/V_L gene sequence, sequences containing high-frequency CDR3s of interest were analyzed for pairwise homology by BLAST, and the sequence with the highest score was chosen. **Supplementary Figure 3** summarizes the bioinformatics analysis of the V-gene sequences. Analysis was performed using Perl scripts in a Unix environment, which were converted into a graphical user interface using the Matlab 7.1 GUI builder for enhanced visualization of results (**Supplementary Data**).

(2) Analysis of CDR3 expression across samples from different mice. CDR3 sequences found in multiple samples were extracted and analyzed for their prevalence in all mice. First, principle component analysis was performed using Matlab to analyze the variance of CDR3 expression in different mice (**Supplementary Fig. 7**). The majority of the variance between mouse samples was categorized into seven principle components. Second, the percent of CDRH3 sequences found in multiple samples was calculated. Because it could not be determined whether replicate sequences were due to contamination or a true biological effect, a permutation test was performed to determine whether the percentage of sequences shared across four samples was biased by samples analyzed on a specific day. The percent of shared sequences was calculated for all 70 possible combinations of the eight samples selected four times, subsequently ranked by percentage overlap. The top three ranked combinations were considered significant and not attributed to random combinations.

(3) Frequency distribution of abundant CDRH3. A heat map was generated to illustrate the prevalence of highly abundant CDRH3s from each sample in mice receiving different antigens. Only CDRH3 sequences with statistically significant frequencies in the top 5% of the distribution (frequency cutoff ~0.03%) were represented (**Fig. 2**).

(4) Homology analysis of full-length V genes. Full-length V genes were found for sequences containing identical CDR3s. First, sequences were placed

doi:10.1038/nbt.1673

in frame by docking CDR3 motifs. Second, full-length V-gene sequences were accepted if they did not contain stop codons and covered all three CDR regions. Nonidentical, full-length V genes (containing at least one amino acid difference) were aligned to determine pairwise homology using the multiple sequence alignment tool in Geneious Software (Biomatters Ltd.; **Supplementary Fig. 6** and **Supplementary Table 7**).

(5) Germline analysis. The top four full-length consensus V_L and V_H genes were analyzed by the IMGT/V-Quest Tool[28]. Additionally, the top 30 ranked CDRH3 sequences of four mice (adjuvant-1, adjuvant-2, C1s-1 and C1s-2) were further analyzed for V(D)J recombination using the IMGT/V-QUEST tool. The V segment germline usage and V_H gene somatic mutations were identified after the IMGT/V-QUEST analysis. These data are reported in **Supplementary Figure 5** and **Supplementary Table 6**.

Construction of synthetic antibody genes. The coding sequences for the selected V_L and V_H genes were designed using the GeneFab software component of our in-house protein fabrication automation (PFA) platform[29]. After reverse translation of the primary amino acid sequences for each V_L and V_H using an *E. coli* class II codon table, the coding sequences for each V_L and V_H were paired based upon their relative frequency from the sequencing data (most abundant V_L with the most abundant V_H, and so forth). The antibody V_L and V_H sequences were built into scFvs with a polyglycine-serine linker $(GGGGS)_4$ between the V_L and V_H sequences. The scFv genes were aligned using the sequence encoding the common $(GGGGS)_4$ linker sequence and a universal randomly generated stuffer sequence was applied to the ends of the scFv sequences to ensure that all of the constructs were of the same length (808 bp). This design format reduced the number of oligonucleotides needed for gene synthesis as oligonucleotides with identical sequences between the different scFv constructs could be reused. SfiI restriction endonuclease sites were added, flanking each gene sequence to facilitate cloning of the synthetic gene constructs into compatible pMoPac16 vectors[30].

The scFv genes were synthesized from overlapping oligonucleotides using a modified thermodynamically balanced inside-out nucleation PCR[31]. The 80 mer oligonucleotides necessary for the construction of the various scFv genes were designed using the GeneFab software with a minimal overlap of 30 nucleotides between oligonucleotide fragments. The oligonucleotides were synthesized using standard phosphoramidite chemistry at a 50 nmol scale using a Mermade 192 oligonucleotide synthesizer (Bioautomation) using synthesis reagents from EMD Chemical and phosphoramidites from Glen Research. All of the oligonucleotide liquid-handling operations necessary for assembling the various genes were done on a Tecan Evo 200 workstation (Tecan) with reagent management and instrument control done through the FabMgr software component of the PFA platform[29]. The gene assembly PCRs were performed using KOD-Hotstart polymerase using buffers and reagents supplied with the enzyme (Novagen). To facilitate cloning of the V_L and V_H genes separately into vectors for IgG expression, the genes for the various V_L and V_H pairs were either built as gene fusions similar to the scFvs except without the $(GGGGS)_4$ linker or as separate genes. These constructs contained sites for the restriction enzymes BssHII and BsiWI flanking the V_L gene and the BssHII and NheI sites flanking the V_H gene.

Antibody expression and antigen binding analysis. Antibody fragments were expressed as scFv fusions to the human light chain constant region Cκ (scAbs), followed by a C-terminal polyhistidine (polyHis) tag. Cloning was accomplished by SfiI digestion of antibody genes and ligation into the expression vector pMoPac16 followed by electroporation transformation into *E. coli Jude 1* cells, which were then plated on Luria Broth (LB, Miller) agar plates supplemented with 100 μg/ml ampicillin. Single colonies were used to inoculate cultures in microtiter 96-well plates with 200 μl/well of Terrific Broth (TB, Miller) supplemented with 2% glucose and 100 μg/ml ampicillin; plates were shaken for 16 h at 30 °C. 10 μl of each well was used to inoculate 200 μl/well of fresh 96-well plates containing TB media supplemented with 100 μg/ml ampicillin and 1 mM of isopropyl-β-D-thiogalactopyranoside (IPTG, Calbiochem).

After a 4 h IPTG induction at 25 °C with shaking, plates were centrifuged at 3,600*g* for 10 min at 4 °C, the supernatant was decanted and cell pellets were resuspended in 20% BugBuster HT (Novagen) in PBS at 150 μl/well. Plates were then shaken at 25 °C for 30 min, and then centrifuged at 3,600*g* for 15 min at 4 °C. 50 μl/well of cell lysates were then added to an ELISA 96-well plate that was precoated with antigen (e.g., ovalbumin, C1s, Bright) at 2 μg/ml in PBS and preblocked with 0.5% BSA or 1% gelatin. A standard indirect ELISA protocol was followed with the detection anti-polyHis antibody (Sigma) conjugated to horseradish peroxidase (HRP) and developed with TMB substrate (Dako) for 15–45 min and stopped with 2N H_2SO_4. The absorbance was measured at 450 nm with a 96-well spectrophotometer (BioTek). Positive wells were identified when the absorbance value was at least threefold above background binding to BSA.

For IgG expression, synthetic V_L and V_H genes were digested with BssHII/BsiWI and BssHII/NheI, respectively, and then ligated into the vectors pMAZ-IgL and pMAZ-IgH, respectively[32]. pMAZ-IgL carries the constant human kappa light chain antibody region and pMAZ-IgH carries the constant human heavy chain antibody region of IgG1. Vectors were transformed into *E. coli Jude 1* cells and plated on LB agar plates supplemented with 100 μg/ml ampicillin. Single colonies were selected and verified for correct V gene sequence. *E. coli* cells carrying pMAZ-IgL and pMAZ-IgH vectors were then grown in 2 ml TB supplemented with 100 μg/ml ampicillin; after overnight growth, plasmid DNA was isolated and purified. 20 μg each of purified pMAZ-IgL and pMAZ-IgH were used for cotransfection and transient expression from HEK 293F cells following the Freestyle MAX expression system (Invitrogen). HEK 293F cells were grown for 96 h after transfection and medium was harvested and IgG was purified by a protein-A agarose chromatography column.

Surface plasmon resonance. C1s was covalently immobilized on a CM5 chip (GE Healthcare) at a level of ~200 response units via standard amine coupling chemistry as described in the manufacturer's protocol. BSA was similarly coupled for baseline correction. All kinetic analyses were performed at 25 °C in HBS-EP (10 mM HEPES, 150 mM NaCl, 50 μM EDTA, 0.005% P-20, pH 7.4) on a BIAcore 3000 (GE Healthcare). Antibodies were injected over immobilized antigen at a flow rate of 50 μl/min or 100 μl/min and the chip was regenerated with a single 10s injection of 20 mM NaOH. Each sensogram was run in duplicate. Kinetic and equilibrium constants were determined by global fitting to a bivalent model using BIAevaluation software (GE Healthcare).

Software. Software is available upon request and on our website: (http://www.che.utexas.edu/georgiou/home.htm).

27. Krebber, A. *et al.* Reliable cloning of functional antibody variable domains from hybridomas and spleen cell repertoires employing a reengineered phage display system. *J. Immunol. Methods* **201**, 35–55 (1997).
28. Brochet, X., Lefranc, M.P. & Giudicelli, V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res.* **36**, W503–W508 (2008).
29. Cox, J.C., Lape, J., Sayed, M.A. & Hellinga, H.W. Protein fabrication automation. *Protein Sci.* **16**, 379–390 (2007).
30. Hayhurst, A. *et al.* Isolation and expression of recombinant antibody fragments to the biological warfare pathogen Brucella melitensis. *J. Immunol. Methods* **276**, 185–196 (2003).
31. Gao, X., Yo, P., Keith, A., Ragan, T.J. & Harris, T.K. Thermodynamically balanced inside-out (TBIO) PCR-based gene synthesis: a novel method of primer design for high-fidelity assembly of longer gene sequences. *Nucleic Acids Res.* **31**, e143 (2003).
32. Mazor, Y., Barnea, I., Keydar, I. & Benhar, I. Antibody internalization studied using a novel IgG binding toxin fusion. *J. Immunol. Methods* **321**, 41–59 (2007).

nature biotechnology

De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis

Polly M Fordyce[1,2,6], Doron Gerber[3,6], Danh Tran[4], Jiashun Zheng[1], Hao Li[1,5], Joseph L DeRisi[1,2,6] & Stephen R Quake[2,4,6]

Gene expression is regulated in part by protein transcription factors that bind target regulatory DNA sequences. Predicting DNA binding sites and affinities from transcription factor sequence or structure is difficult; therefore, experimental data are required to link transcription factors to target sequences. We present a microfluidics-based approach for *de novo* discovery and quantitative biophysical characterization of DNA target sequences. We validated our technique by measuring sequence preferences for 28 *Saccharomyces cerevisiae* transcription factors with a variety of DNA-binding domains, including several that have proven difficult to study by other techniques. For each transcription factor, we measured relative binding affinities to oligonucleotides covering all possible 8-bp DNA sequences to create a comprehensive map of sequence preferences; for four transcription factors, we also determined absolute affinities. We expect that these data and future use of this technique will provide information essential for understanding transcription factor specificity, improving identification of regulatory sites and reconstructing regulatory interactions.

Recent evidence suggests that knowledge of both strongly and weakly bound sequences and their interaction affinities is required for an accurate understanding of transcriptional regulation. Weak-affinity sites are evolutionarily conserved, make significant contributions to overall transcription[1,2] and may allow closely related transcription factors to mediate different transcriptional responses[3]. In addition, quantitative models require both strongly and weakly bound sequences and their binding affinities to recapitulate transcriptional responses[4–7].

Unfortunately, quantitative data detailing transcription factor binding are often lacking, even for model organisms. *In vivo* immunoprecipitation-based methods, such as ChIP-chip[8] and ChIP-SEQ[9], provide genome-wide information about promoter occupancy. However, these techniques require knowledge of physiological states under which transcription factors are bound to promoters, cannot distinguish whether a transcription factor contacts DNA directly or is tethered by means of another DNA-binding protein, and do not measure affinities.

In vitro methods complement *in vivo* data by measuring binding affinities, distinguishing whether transcription factors directly bind DNA, and allowing manipulation of post-translational modifications and buffer conditions. Furthermore, *in vitro* methods can be used without knowledge of the conditions under which transcription factors are active. However, current *in vitro* methods cannot simultaneously discover both high- and low-affinity target sequences and measure their affinities. Electromobility shift assays[10], DNAse footprinting[11] and surface plasmon resonance[12] require prior knowledge of potential binding sites, precluding motif discovery. Conversely, selection techniques (e.g., SELEX) and one-hybrid systems[13] discover motifs from a large sequence space, but recover only the most strongly bound sequences, without affinity information. Protein binding microarrays (PBMs)[3,14–18] can discover both strongly and weakly bound sequences but cannot measure reactions at equilibrium, preventing affinity measurements. PBMs also suffer from reduced sensitivity: a recent study using PBMs to probe transcription factor binding in *S. cerevisiae* failed to recover consensus motifs for 49 of 101 transcription factors with previous evidence of direct DNA binding[15]. Embedding immobilized DNA in hydrogels[19] extends the PBM technique to allow affinity and kinetic measurements, but this approach can analyze binding to only ~100 DNA sequences at a time.

An alternative approach is mechanically induced trapping of molecular interactions (MITOMI), a technique that uses a microfluidic device to measure binding interactions at equilibrium, allowing construction of detailed maps of binding energy landscapes. The first-generation MITOMI device measured 640 parallel interactions and required DNA libraries that were specific to a particular transcription factor[20].

Here we report a second-generation MITOMI device (MITOMI 2.0) capable of measuring 4,160 parallel interactions. Devices were fabricated in polydimethylsiloxane (PDMS) using multilayer soft lithography; each device had 4,160 unit cells and ~12,555 valves

[1]Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, California, USA. [2]Howard Hughes Medical Institute, Chevy Chase, Maryland, USA. [3]Department of Life Sciences, Bar Ilan University, Ramat Gan, Israel. [4]Departments of Bioengineering and Applied Physics, Stanford University, Palo Alto, California, USA. [5]Center for Theoretical Biology, Peking University, Beijing, China. [6]These authors contributed equally to this work. Correspondence should be addressed to S.R.Q. (quake@stanford.edu) or J.L.D. (joe@derisilab.ucsf.edu).

Received 17 March; accepted 5 August; published online 29 August 2010; doi:10.1038/nbt.1675

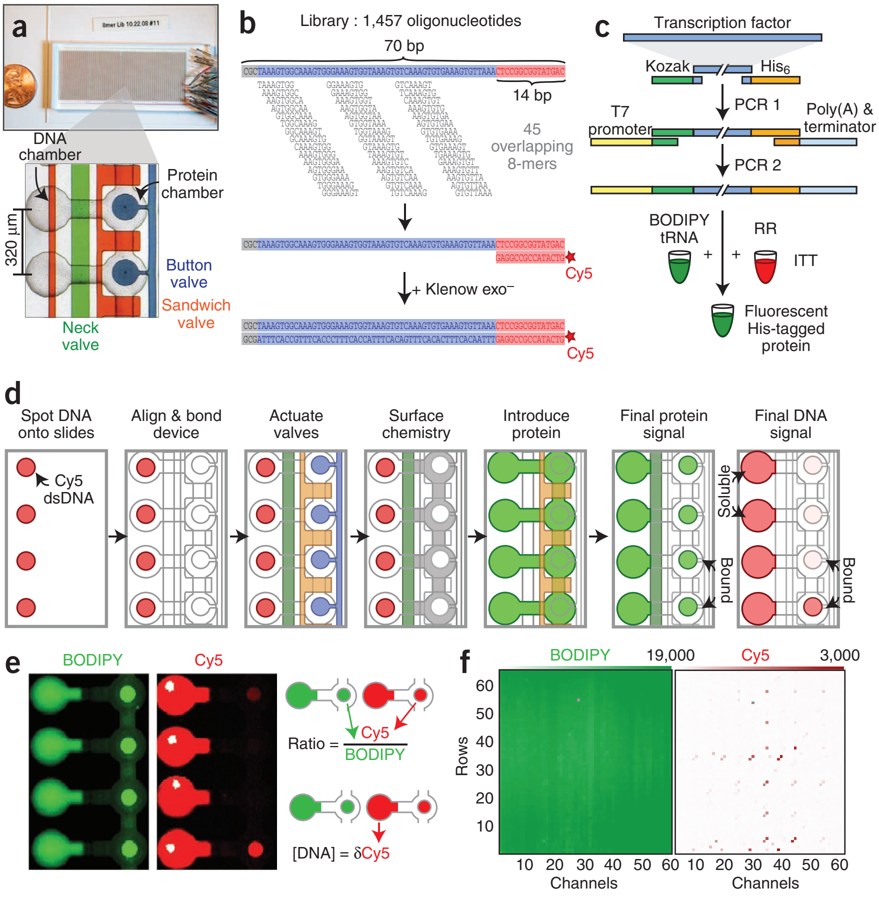


Figure 1 Overall experimental design and procedure. (**a**) Microfluidic device hybridized to glass slide. Unit cells contain two chambers (a 'DNA chamber' and a 'protein' chamber) controlled by three valves: a 'neck' valve (green) separates the two chambers; a 'sandwich' valve (orange) isolates unit cells; and a 'button' valve (blue) protects molecular interactions. (**b**) DNA 8-mer library design. Each 70-bp oligonucleotide contains 45 overlapping 8-mers, a 3-bp GC-clamp at the 5′ end and an identical 14-bp sequence at the 3′ end for Cy5 labeling and primer extension. (**c**) PCR generation of linear templates for protein expression. In PCR1, template-specific primers attach a Kozak sequence, 6× His tag and universal overhangs. In PCR2, universal primers add a T7 promoter, poly-A tail and T7 terminator. *In vitro* transcription and translation (ITT) of this template in rabbit reticulocyte lysate (RR) with BODIPY-labeled, lysine-charged tRNA produces labeled, His-tagged protein. (**d**) Overview of experimental procedure. Devices are manually aligned to a spotted microarray. Neck valves are closed to protect DNA within chambers, and slide surfaces are derivatized with anti-pentaHis antibodies below the button (white) and passivated elsewhere (gray). Lysate containing fluorescently labeled His-tagged transcription factors is introduced and neck valves are opened to allow interaction between transcription factors and DNA; sandwich valves are closed to isolate each unit cell. After an incubation, button valves are pressurized to protect protein–DNA interactions, unbound DNA and proteins are washed out, and the device is scanned. δ is a proportionality constant. (**e**) Scanned picture showing final protein (BODIPY, left) and DNA (Cy5, right) intensities in the chamber and under the button. (**f**) Arrays showing example protein intensities (left) and DNA intensities (right) under the button for each unit cell within a device.

to control fluid flow (**Fig. 1a** and **Supplementary Fig. 1**). Each unit cell contained a DNA chamber and a protein chamber, controlled by micromechanical valves—a 'neck' valve, 'sandwich' valves and a 'button' valve (**Fig. 1a**). Unit cells were programmed with particular DNA sequences by aligning and bonding the device with a noncovalently spotted DNA microarray containing a library of 1,457 double-stranded Cy5-labeled oligonucleotides. To accommodate all 65,536 DNA 8-mers, we designed each 70-bp oligonucleotide to contain 45 overlapping, related 8-mer de Bruijn sequences[21] (**Fig. 1b**). Each oligonucleotide sequence appeared in at least two unit cells.

To evaluate the performance of this technique, we measured DNA binding for 28 *S. cerevisiae* transcription factors from ten different families (**Supplementary Table 1**). Of these, there was prior evidence for 26 transcription factors, of direct, sequence-specific DNA binding, and 2 transcription factors had no previously annotated literature motifs, despite multiple previous attempts[14,15,22].

All transcription factor protein was produced by *in vitro* transcription and translation. PCR-generated linear expression templates were added directly to rabbit reticulocyte lysate off-chip in the presence of a small fraction of BODIPY-labeled, lysine-charged tRNA to produce BODIPY-labeled, His-tagged transcription factors (**Fig. 1c** and **Supplementary Fig. 2**). In each experiment, ~50 μl of extract (~100 ng of protein) was loaded into the device.

After alignment to DNA microarrays, slide surfaces within the protein chamber were derivatized with anti-pentaHis antibodies beneath the button valve and passivated elsewhere (**Fig. 1d**). Introduction of His-tagged transcription factors into both chambers solubilized spotted DNA, allowing transcription factors and DNA to interact. Transcription factor–DNA complexes were captured on the surface beneath the button valve during a ~1 h incubation; rapid closure of the button valve trapped interactions at equilibrium concentrations before a final wash to remove unbound material before imaging[20].

BODIPY intensities under the button valve reflect the number of surface-bound protein molecules; Cy5 intensities under the button valve reflect the number of DNA molecules bound by surface-immobilized protein (**Fig. 1d–f**). Therefore, the ratio of Cy5 to BODIPY fluorescence is linearly proportional to the number of protein molecules with bound DNA, or protein fractional occupancy. Cy5 intensities within the DNA chamber reflect the amount of soluble DNA available for binding.

All 28 transcription factors showed oligonucleotide-specific variations in bound Cy5 intensities, demonstrating marked preferences for individual oligonucleotides (**Fig. 2a** and **Supplementary Fig. 3**). By contrast, the distribution of intensities for rabbit reticulocyte extract alone was well fit by a Gaussian distribution (reduced $\chi^2 = 1.0$, $P = 0.47$), establishing that binding is due to expressed transcription factors and not components of the *in vitro* transcription and translation system (**Fig. 2a**).

Variations in fluid flow between channels can lead to differences in the number of protein molecules beneath each button valve. To account for these differences and generate a quantity proportional

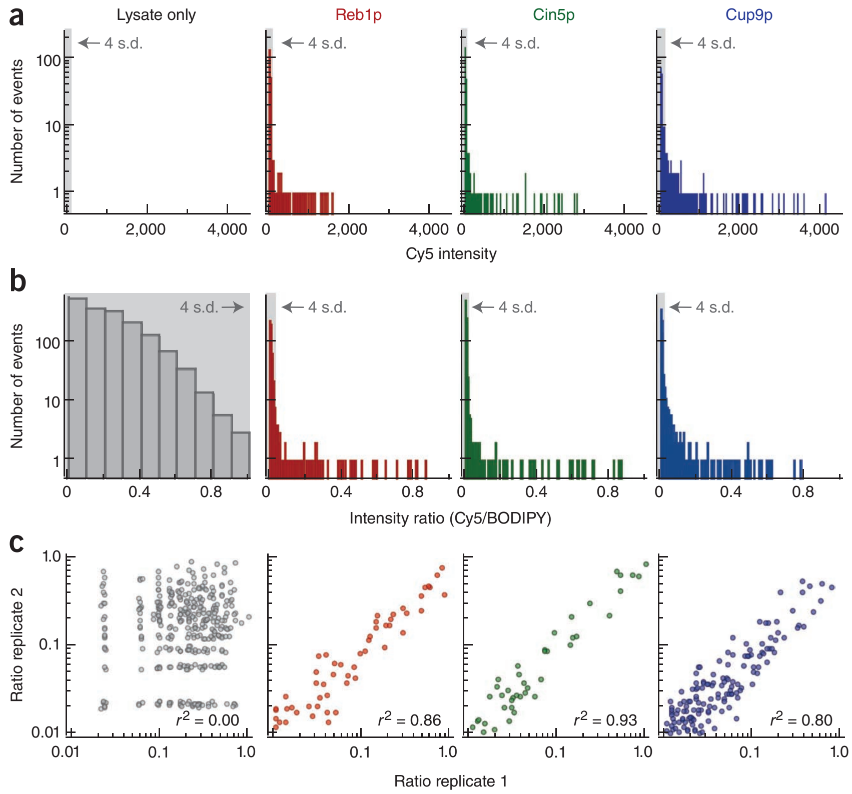


Figure 2 Detailed analysis of measured Cy5 intensities and fluorescence intensity ratios (Cy5/BODIPY-FL) for rabbit reticulocyte lysate alone, Reb1p, Cin5p and Cup9p. (**a**) Distribution of measured Cy5 intensities for all oligonucleotides. Light gray box indicates measurements within 4 s.d. of the mean (as determined by a Gaussian fit). Measured Cy5 intensities for rabbit reticulocyte lysate alone are well fit by a Gaussian distribution (reduced $\chi^2 = 1.0$, $P = 0.47$). For all transcription factors, measured Cy5 intensities deviate significantly from a Gaussian distribution, with measured events many s.d. above the mean. (**b**) Distribution of measured intensity ratios for all oligonucleotides. Light gray box indicates measurements within 4 s.d. of the mean (as determined by a Gaussian fit). Measured intensity ratios in the presence of transcription factors deviate significantly from a normal distribution (**Supplementary Table 2**). (**c**) Correlation between ratios measured for the same oligonucleotide at two separate locations within the device.

to fractional occupancy, Cy5 intensities were normalized by BODIPY intensities to yield a dimensionless intensity ratio (Cy5 intensity/BODIPY intensity) (**Fig. 1e**). Intensity ratios also showed strong preferences for individual oligonucleotide sequences, with no clear preference detected for rabbit reticulocyte lysate alone (**Fig. 2b**, **Supplementary Fig. 4** and **Supplementary Table 2**). Intensity ratios were well correlated both between measurements of the same 70-mer oligonucleotide at different locations within a given device (**Fig. 2c** and **Supplementary Table 3**) and between experiments (**Supplementary Fig. 5**).

Binding affinity can be described by a single-site binding model relating intensity ratio (r) to DNA concentration ($[D]$); K_d, the DNA concentration at which measured intensities reach half their maximum value ($r_{\max}$) provides a quantitative measure of binding affinity.

$$r = \frac{r_{\max} \cdot [D]}{[D] + K_d} \qquad (1)$$

At low DNA concentrations, measured intensity ratios are approximately inversely proportional to K_d. Calibrated measurements of DNA chamber intensities in our experiments establish that soluble DNA concentrations are indeed low (150 ± 25 nM, mean ± s.e.m.) (**Supplementary Fig. 6**), suggesting it might be possible to accurately estimate interaction affinities from intensity ratios measured at a single, low DNA concentration.

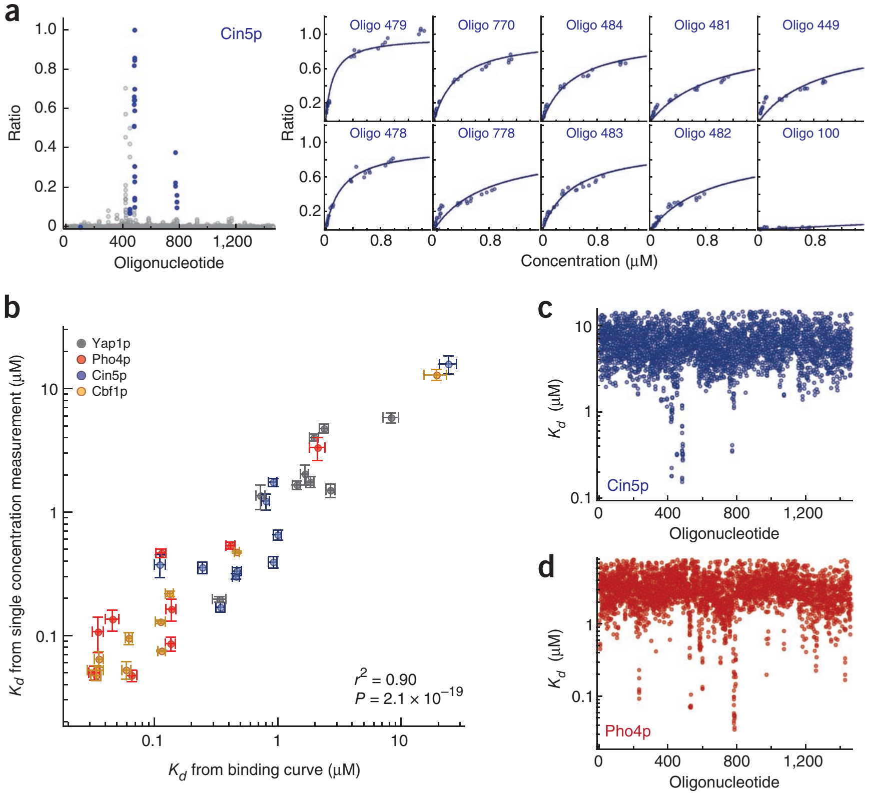
To test this hypothesis, we first measured concentration-dependent binding for four transcription factors (Cbf1p, Cin5p, Pho4p and Yap1p) from two different families, each interacting with ten oligonucleotides from the 8-mer DNA library. We then globally fit equation (1) over all oligonucleotides at all concentrations to get accurate K_d measurements (**Fig. 3a** and **Supplementary Figs. 7–9**).

Next, we calculated K_d values for the exact same oligonucleotides from single-concentration measurements. The low DNA concentration used for these measurements prevented direct determination of $r_{\max}$, a parameter that depends on quantities that vary between experiments (e.g., amount and intensity of BODIPY and Cy5 dyes incorporated during protein and DNA library production, respectively), and must be empirically determined. K_d values from concentration-dependent binding can be used to 'calibrate' the appropriate $r_{\max}$ value (**Supplementary Methods** and **Supplementary Tables 4** and **5**). Single-concentration K_d values calculated using calibrated $r_{\max}$ values were in excellent agreement with those derived from concentration-dependent binding ($r^2 = 0.90$, $P = 2.1 \times 10^{-19}$) (**Fig. 3b**). Furthermore, once calibrated, $r_{\max}$ values can be used to calculate K_d values for all oligonucleotides with signals above background, providing absolute affinities for all 1,457 oligonucleotides with only a few additional measurements (**Fig. 3c,d** and **Supplementary Fig. 10**). The range of K_d values calculated here for Pho4p and Cbf1p agree with those measured in previous studies (~10 nM–10 μM)[20], validating our approach. Relative differences in binding affinities between oligonucleotides (the Gibbs free energy upon binding, ΔΔG) can also be calculated using these calibrated $r_{\max}$ values (**Supplementary Fig. 11**).

Even in the absence of additional information to calibrate $r_{\max}$ values, however, measured intensity ratios provide accurate information about binding affinity. To demonstrate this, we assumed an $r_{\max}$ value of 1 for all transcription factors and again compared measured and calculated K_d values. K_d measurements were well correlated ($r^2 = 0.67$, $P = 1.8 \times 10^{-10}$), although individual curves were systematically offset (**Supplementary Fig. 12**a). ΔΔG describes relative affinity differences between oligonucleotides and is therefore less sensitive to these offsets, with stronger correlations ($r^2 = 0.76$, $P = 8.0 \times 10^{-13}$) (**Supplementary Fig. 12b**).

Measured intensity ratios reflect interaction affinities between a given transcription factor and a 70-bp oligonucleotide. Identifying transcription factor target sites requires determination of the precise subsequences responsible for transcription factor binding within each oligonucleotide. Traditionally, analysis of transcription factor binding requires designation of sequences into bound and unbound populations, followed by a search for sequences overrepresented in the bound population, which ignores relative strengths of binding interactions,

Figure 3 Comparison between K_d values derived from direct measurements of concentration-dependent binding and K_d values calculated from ratio measurements at a single concentration. (**a**) Cin5p measurements. Measured ratio signals for all oligonucleotides (gray) and selected oligonucleotides (blue) (left); concentration-dependent binding for selected oligonucleotides fit to a single-site binding model (right). (**b**) K_d calculated from single-concentration measurements compared with K_d derived from fits concentration-dependent binding for Cin5p (blue), Pho4p (red), Yap1p (gray) and Cbf1p (gold). (**c**) Calculated K_d values for all oligonucleotides for Cin5p. (**d**) Calculated K_d values for all oligonucleotides for Pho4p.



and can be sensitive to the precise threshold used to delineate populations. Here we used a pipeline that incorporates all intensity information for all oligonucleotides to generate a position-specific affinity matrix (PSAM)[23] describing the change in binding affinity upon mutation of a specific position within a consensus sequence (**Supplementary Fig. 13**). Notably, PSAMs describe actual binding affinities for any combination of nucleotides and can be used to calculate predicted affinities to arbitrary sequences.

First, we analyzed all measured intensity ratios using fREDUCE, an enumerative algorithm that searches for sequences whose occurrence within oligonucleotides correlates strongly with their measured signal[24]. For all 28 proteins, fREDUCE returned sequences with strong correlations (**Supplementary Table 6** and **Supplementary Fig. 14**).

Next, the highest-correlated 7- and 8-bp fREDUCE sequences were converted to PSAMs using MatrixREDUCE[23], an algorithm that fits all measured intensity ratios with a statistical mechanical model assessing the effects of individual base-pair substitutions on binding affinity. Because investigations of MatrixREDUCE performance have recommended the use of initial seed sequences derived from enumerative analysis to ensure optimization of global minima[24], the fREDUCE sequences were used as seeds. MatrixREDUCE assumes that the free energy contributions of each position in the binding site are independent; although this is known to be false in some instances, we use linear motifs here to compare our results with the largest possible set of previous literature.

To choose the single PSAM that best explains measured binding, we compared occupancies predicted by each PSAM for all oligonucleotides in the DNA library with measured intensity ratios (**Supplementary Fig. 15**). Predicted and measured values were well-correlated for almost all transcription factors (**Supplementary Table 7**). For all 26 transcription factors with described motifs, the final recovered motif was in agreement with those previously reported in the literature (**Fig. 4**)[14,15,22]. We also derived PSAMs for two transcription factors that were previously resistant to characterization, Msn1p and Nrg2p, establishing considerably enhanced sensitivity over both ChIP-based and PBM techniques.

Two well-characterized basic helix-loop-helix proteins (Pho4p and Cbf1p) provide a test of the ability to detect both high- and low-affinity target sequences. Pho4p binds both high-affinity (5′-CACGTG-3′) and low-affinity (5′-CACGTT-3′) sites[25]; Cbf1p binds to a degenerate 5′-RTCACRTG-3′ motif[20,26]. For both proteins, we recovered the expected motif variants (**Fig. 4** and **Supplementary Fig. 15**).

Detailed analysis of differences between measured and calculated binding profiles can provide additional information about binding preferences. For example, oligonucleotides with high measured intensity ratios but low predicted occupancies could indicate binding to additional motifs. In addition, this comparison allows investigation of whether free energy contributions at each position within the sequence are truly independent.

For most transcription factors, optimized PSAMs successfully described gross binding properties (e.g., Pho4p, Cin5p, Msn2p and Sko1p; **Supplementary Fig. 16**), albeit with outliers at weak binding energies that may represent cooperative interactions between base-pair substitutions. For a few transcription factors (Rpn4p, Cup9p, Cad1p, Matα2p and Pdr3p), correlations between measured and predicted binding were much weaker ($r^2 < 0.25$). To determine if low correlations resulted from binding to additional target sequences, we used BioPROSPECTOR[27], MDScan[27], MEME[28] and WEEDER[29] to scan for overrepresented sequences within oligonucleotides with high measured intensity ratios (*Z*-score > 25 for Rpn4p or 75 for Cup9p) but low predicted occupancies (*Z*-score < 3).

For Rpn4p, although both PBM studies and our initial analysis identified binding to a 5′-GCCACC-3′ motif, ChIP and expression data suggest a T-rich 5′ extension of this motif upstream of Rpn4p target genes. Notably, analysis of the 13 oligonucleotides with discordant measured and predicted binding returned this precise extension, establishing that unexpected binding data can yield biologically relevant results (**Supplementary Fig. 17**).

The Cup9p-optimized PSAM also agreed with previous PBM[15] results (**Fig. 4**); however, 14 sequences showed stronger-than-predicted binding (**Supplementary Fig. 18**). Analysis of these sequences yielded

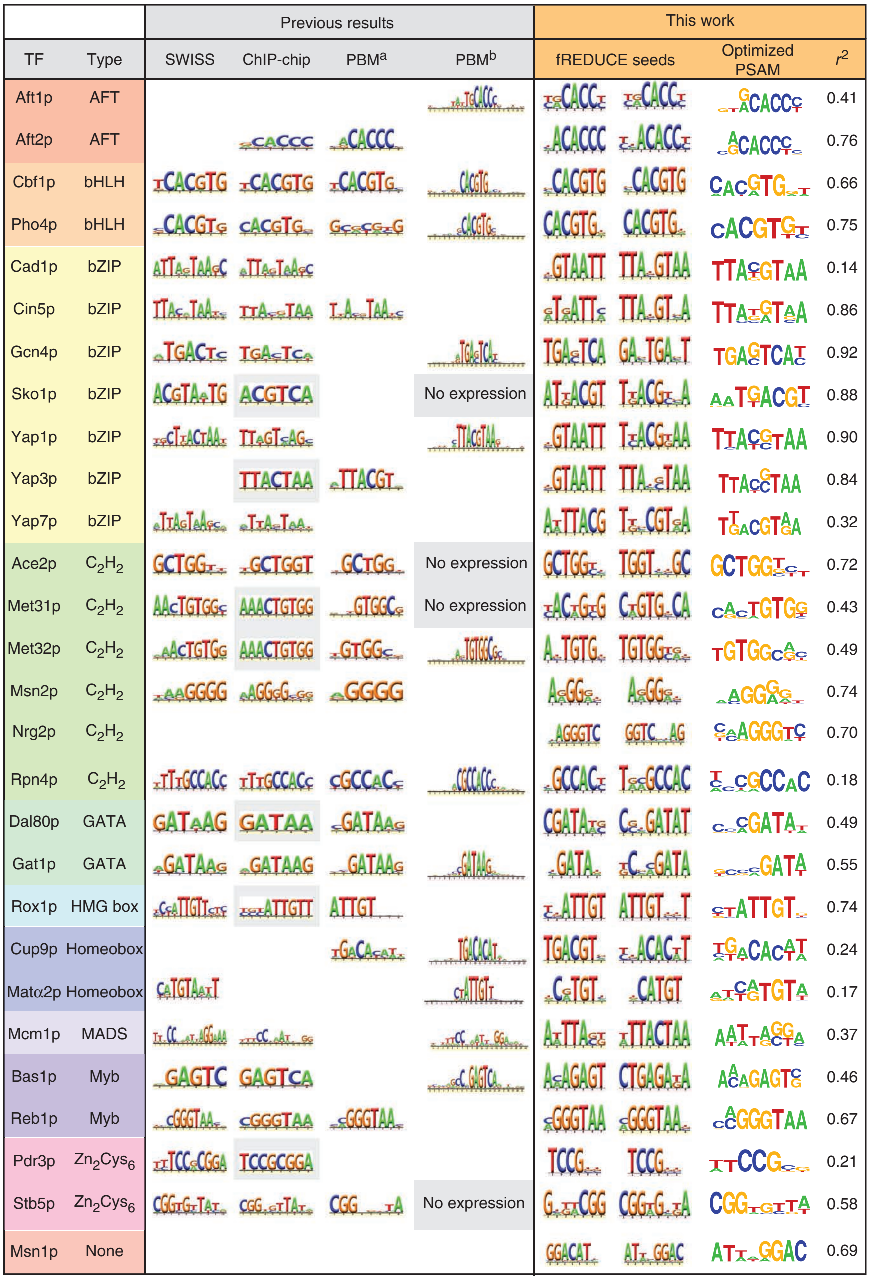

Previous results
This work
TF
Type
SWISS
ChIP-chip
PBM^a
PBM^b
fREDUCE seeds
Optimized PSAM
r^2
Aft1p AFT 0.41
Aft2p AFT 0.76
Cbf1p bHLH 0.66
Pho4p bHLH 0.75
Cad1p bZIP 0.14
Cin5p bZIP 0.86
Gcn4p bZIP 0.92
Sko1p bZIP No expression 0.88
Yap1p bZIP 0.90
Yap3p bZIP 0.84
Yap7p bZIP 0.32
Ace2p C_2H_2 No expression 0.72
Met31p C_2H_2 No expression 0.43
Met32p C_2H_2 0.49
Msn2p C_2H_2 0.74
Nrg2p C_2H_2 0.70
Rpn4p C_2H_2 0.18
Dal80p GATA 0.49
Gat1p GATA 0.55
Rox1p HMG box 0.74
Cup9p Homeobox 0.24
Matα2p Homeobox 0.17
Mcm1p MADS 0.37
Bas1p Myb 0.46
Reb1p Myb 0.67
Pdr3p Zn_2Cys_6 0.21
Stb5p Zn_2Cys_6 No expression 0.58
Msn1p None 0.69

Figure 4 Comparison between motifs found for all 28 *S. cerevisiae* transcription factors and previous literature results (SWISS, SwissRegulon[30]; ChIP-chip, Harbison library[22]; PBM[1], protein binding microarray[14]; PBM[2], protein binding microarray[15]). For ChIP-chip data, boxes shaded in gray represent literature-derived motifs. For PBM[2] results, white boxes represent proteins applied to arrays that did not yield motifs; boxes shaded in gray represent proteins that were not expressed sufficiently to be applied to arrays. fREDUCE Seeds: 7- and 8-bp fREDUCE motifs that correlate most strongly with measured intensities; Optimized PSAM: MatrixREDUCE PSAM represented as an AffinityLogo; r^2: Pearson correlation coefficient between all measured ratio values and protein occupancies predicted by the optimized PSAM.

motifs similar to the optimized PSAM, but with an 'ACGT' core (**Supplementary Fig. 18**, gray box). To assess the affinity of Cup9p for this candidate alternate motif, we measured concentration-dependent binding of Cup9p to the primary motif, candidate secondary motif and several related motifs (**Supplementary Fig. 19a**). A random 2-bp substitution abolished binding, but mutating these bases or the entire second half of the motif to the candidate secondary motif reduced affinity only ~20-fold (**Supplementary Fig. 19b**), confirming weak-affinity binding. Interestingly, this motif is found only 29 times in the genome outside of coding regions, primarily at the boundary of subtelomeric repeats and upstream of genes regulated by iron depletion, metal toxicity or oxidative stress (**Supplementary Table 8**). Although the physiological role of these putative binding sites is unknown, these results demonstrate the ability of MITOMI 2.0 to detect weak but potentially biologically relevant transcription factor binding sites.

For the remaining three transcription factors (Cad1p, Matα2p and Pdr3p), low correlations between predicted and measured binding likely resulted from experimental variability and not binding to additional motifs. Correlations between technical replicates across the device were relatively low (**Supplementary Table 3**), owing to either binding to a limited number of oligonucleotides (Cad1p, **Supplementary Fig. 3**) or large variations in protein coverage (for Matα2p and Pdr3p). Consistent with this, these transcription factors do not bind any oligonucleotides with stronger-than-expected affinity.

The data presented here demonstrate increased sensitivity over current state-of-the-art techniques, detecting sequence-specific binding for several proteins that have failed to yield results in multiple experiments (Cad1p, Msn1p, Nrg2p, Sko1p, Yap7p and Pdr3p). Moreover, these data represent the most comprehensive investigation of biophysical binding affinities to date, including ΔΔG values for 28 transcription factors and K_d values for four transcription factors from two different families (Cbf1p, Cin5p, Pho4p and Yap1p) binding to 1,457 individual sequences. These data can be used to test basic assumptions underlying current models of transcription factor–DNA specificity and to more accurately model cooperativity between nucleotide-binding sites ('nonadditivity').

The DNA library used here is not organism-specific, making this technique useful for a wide range of organisms, including higher eukaryotes and pathogens. In addition, the programmable nature of MITOMI 2.0 allows subsequent detailed examination of unexpected binding phenomena or systematic mutational analysis of candidate motifs through direct observations of concentration-dependent binding. Although these experiments probed transcription factor binding to double-stranded DNA, MITOMI 2.0 can be used, with only minimal changes, to investigate single-stranded DNA binding and RNA binding. When paired with advances in rapid whole-genome sequencing, we anticipate that MITOMI 2.0 characterization of all recognizable transcription factors in a given proteome will allow transcriptional networks and regulons to be quickly identified and ultimately modeled.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Accession codes. Gene Expression Omnibus: GPL10817.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

P.M.F. was supported by a Howard Hughes Medical Institute/Helen Hay Whitney Foundation Postdoctoral Fellowship. J.L.D., S.R.Q. and this work were supported by the Howard Hughes Medical Institute. We thank A. Potanina for assistance with fabrication of microfluidic devices, O. Homann for implementation of PSAM functionality with MochiView and D. Breslow, F. Caro, S. Churchman, M. Dimon, T. Kiers, A. Kistler, C. Nelson, K. Sorber, E. Yeh and I. Zuleta for careful reading of the manuscript.

AUTHOR CONTRIBUTIONS

P.M.F. designed experiments, designed, created and printed the DNA library, made linear expression templates, fabricated microfluidic devices, performed microfluidic experiments assessing concentration-dependent binding and binding to the 8-mer library, analyzed data and wrote the manuscript. D.G. designed experiments, designed and fabricated microfluidic devices and performed microfluidic experiments assessing binding to the 8-mer library. D.T. fabricated microfluidic devices and performed microfluidic experiments assessing binding to the 8-mer library. J.Z. and H.L. analyzed data. S.R.Q. designed experiments, analyzed data and wrote the manuscript. J.L.D. designed experiments, assisted with printing the DNA library, analyzed data and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Tanay, A. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res.* **16**, 962–972 (2006).
2. Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U. & Gaul, U. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature* **451**, 535–540 (2008).
3. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723 (2009).
4. Kim, H.D. & O'Shea, E.K. A quantitative model of transcription factor-activated gene expression. *Nat. Struct. Mol. Biol.* **15**, 1192–1198 (2008).
5. Segal, E. & Widom, J. From DNA sequence to transcriptional behavior: a quantitative approach. *Nat. Rev. Genet.* **10**, 443–456 (2009).
6. Gertz, J., Siggia, E.D. & Cohen, B.A. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature* **457**, 215–218 (2009).
7. Yuh, C.H., Bolouri, H. & Davidson, E.H. Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. *Development* **128**, 617–629 (2001).
8. Iyer, V.R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533–538 (2001).
9. Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
10. Garner, M.M. & Revzin, A. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the *Escherichia coli* lactose operon regulatory system. *Nucleic Acids Res.* **9**, 3047–3060 (1981).
11. Galas, D.J. & Schmitz, A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.* **5**, 3157–3170 (1978).
12. Jost, J.P., Munch, O. & Andersson, T. Study of protein-DNA interactions by surface plasmon resonance (real time kinetics). *Nucleic Acids Res.* **19**, 2788 (1991).
13. Meng, X., Brodsky, M.H. & Wolfe, S.A. A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nat. Biotechnol.* **23**, 988–994 (2005).
14. Badis, G. *et al.* A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol. Cell* **32**, 878–887 (2008).
15. Zhu, C. *et al.* High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res.* **19**, 556–566 (2009).
16. Berger, M.F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.* **24**, 1429–1435 (2006).
17. Berger, M. *et al.* Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133**, 1266–1276 (2008).
18. De Silva, E.K. *et al.* Specific DNA-binding by apicomplexan AP2 transcription factors. *Proc. Natl. Acad. Sci. USA* **105**, 8393–8398 (2008).
19. Bonham, A.J., Neumann, T., Tirrell, M. & Reich, N.O. Tracking transcription factor complexes on DNA using total internal reflectance fluorescence protein binding microarrays. *Nucleic Acids Res.* **37**, 94 (2009).
20. Maerkl, S.J. & Quake, S.R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**, 233–237 (2007).
21. Ralston, A. De Bruijn sequences-a model example of the interaction of discrete mathematics and computer science. *Math. Mag.* **55**, 131–143 (1982).
22. Harbison, C.T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).
23. Foat, B.C., Morozov, A.V. & Bussemaker, H.J. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* **22**, e141–e149 (2006).
24. Wu, R., Chaivorapol, C., Zheng, J., Li, H. & Liang, S. fREDUCE: detection of degenerate regulatory elements using correlation with expression. *BMC Bioinformatics* **8**, 399 (2007).
25. Vogel, K., Horz, W. & Hinnen, A. The two positively acting regulatory proteins PHO2 and PHO4 physically interact with PHO5 upstream activation regions. *Mol. Cell. Biol.* **9**, 2050–2057 (1989).
26. Wieland, G. *et al.* Determination of the binding constants of the centromere protein Cbf1 to all 16 centromere DNAs of *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **29**, 1054–1060 (2001).
27. Liu, Y. *et al.* A suite of web-based programs to search for transcriptional regulatory motifs. *Nucleic Acids Res.* **32**, W204–W207 (2004).
28. Bailey, T.L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
29. Pavesi, G. *et al.* MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. *Nucleic Acids Res.* **34**, W566–W570 (2006).
30. Pachkov, M., Erb, I., Molina, N. & Van Nimwegen, E. SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic Acids Res.* **35**, D127–D131 (2007).

ONLINE METHODS

Oligonucleotide sequence files and data for all transcription factors are available for download at http://derisilab.ucsf.edu.

DNA library and transcription-factor production. All possible 65,536 8-bp DNA sequences were assembled into a maximally compact de Bruijn sequence that was subsequently divided over 1,457 oligonucleotides. Sequences were hybridized to a Cy5-labeled oligonucleotide and extended using Klenow fragment (exo-) (New England Biolabs) to produce Cy5-labeled dsDNA. Cy5-labeled dsDNA was diluted to a final concentration of 1.25 μM in 3× SSC with polyethylene glycol (PEG) (Fluka) and D-(+)-trehalose dihydrate (Fluka) (for enhanced subsequent solubility) and printed onto custom 2″ × 3″ ThermoFisher Scientific SuperChip Epoxysilane slides (ThermoFisher Scientific) using a DeRisi lab custom microarrayer.

A two-step PCR reaction was used to amplify transcription factor coding sequences and add appropriate upstream and downstream sequences for efficient transcription and translation in rabbit reticulocyte lysate (Promega) (**Supplementary Fig. 2**).

Microfluidic device fabrication and experimental procedure. Flow and control molds were fabricated on 4″ silicon wafers using positive (SPR 220-7.0) and negative (SU-8) photoresists, respectively. PDMS devices were produced and the MITOMI experimental procedure was performed as described previously[20].

Initial data analysis and normalization. Median Cy5 and BODIPY fluorescence intensities varied somewhat between experiments. To facilitate comparisons between transcription factors, Cy5 intensity distributions were fit to a Gaussian distribution and this Gaussian mean was subtracted from all measurements to center the background distribution around zero. Fluorescence intensity ratios were calculated by dividing Cy5 fluorescence intensities by BODIPY fluorescence intensities; ratios were similarly normalized such that the background was centered around zero, and further normalized such that the maximum measured intensity was 1.

Motif finding pipeline. We searched for 7- and 8-bp sequences that correlated most strongly with measured intensity ratios using fREDUCE. Both doubly- (R, Y, S, W, K, M) and triply- (B, D, H, V) degenerate IUPAC bases were included, and both the forward sequence and its reverse complement were analyzed. The most strongly correlated 7-bp and 8-bp sequences were then used as seeds for MatrixREDUCE analysis, with additional unspecified base pairs added to either side of the 7-bp seed to standardize length.

Occupancy profile calculations. We calculated predicted occupancy profiles from PSAMs using a slight modification of the MatrixREDUCE formalism to reflect the fact that, in our assay, transcription factors are surface-immobilized and DNA sequences are in solution (**Supplementary Methods**).

doi:10.1038/nbt.1675

nature biotechnology

High-throughput generation, optimization and analysis of genome-scale metabolic models

Christopher S Henry[1], Matthew DeJongh[2], Aaron A Best[3], Paul M Frybarger[2,3], Ben Linsay[4] & Rick L Stevens[4,5]

Genome-scale metabolic models have proven to be valuable for predicting organism phenotypes from genotypes. Yet efforts to develop new models are failing to keep pace with genome sequencing. To address this problem, we introduce the Model SEED, a web-based resource for high-throughput generation, optimization and analysis of genome-scale metabolic models. The Model SEED integrates existing methods and introduces techniques to automate nearly every step of this process, taking ~48 h to reconstruct a metabolic model from an assembled genome sequence. We apply this resource to generate 130 genome-scale metabolic models representing a taxonomically diverse set of bacteria. Twenty-two of the models were validated against available gene essentiality and Biolog data, with the average model accuracy determined to be 66% before optimization and 87% after optimization.

Current sequencing technology is producing thousands of sequenced genomes each year, transforming the fields of genomics and bioinformatics and increasing demand for new tools that enable high-throughput generation of functioning genome-scale metabolic models. With a functioning genome-scale metabolic model, culture conditions can be predicted[1], phenotypes can be predicted and reconciled with experimental data[2], and poorly annotated regions of the metabolic network can be identified[3]. In short, genome-scale metabolic models are central to the use of sequence data to produce detailed and quantitative predictions of organism behavior. The process of reconstructing genome-scale metabolic models has been broken down into 96 steps[4], clearly outlining its complexity and explaining in part the slow pace of creation of new models. Here we introduce the Model SEED, a web-based resource (available at http://www.theseed.org/models/) designed to speed the creation of new metabolic models by automating most of these steps. Several steps, however, are not currently amenable to automation and must still be performed manually, which is why we designate the models we create as 'draft models'. We call this resource the Model SEED because it is built upon the foundation of accurate genome annotations provided by the SEED framework for annotation and analysis[5,6]. At the core of the Model SEED is a model reconstruction pipeline (**Fig. 1** and **Supplementary Fig. 1**), which integrates and augments technologies for genome annotation[5,6], construction of gene-protein-reaction (GPR) associations, generation of biomass reactions, reaction network assembly[7], thermodynamic analysis of reaction reversibility[8,9] and model optimization[2,9,10] to generate draft genome-scale metabolic models. Whereas existing automated reconstruction methodologies only address portions of the reconstruction process[7,10–13], the Model SEED is capable of generating functioning draft metabolic models of an organism starting from an assembled genome sequence. The integration of the Model SEED pipeline with the SEED framework also enables a tight coupling between genome annotation and metabolic reconstruction that is essential for the high-throughput generation of metabolic models.

Preliminary model reconstruction

We applied the Model SEED pipeline to generate draft models for a taxonomically diverse set of 130 bacterial organisms (**Fig. 2**). In the first step of the pipeline, the genome sequences for these 130 organisms were imported into the SEED using the RAST server (http://rast.nmpdr.org/)[6], which performs gene calling and annotation of genome sequences in ~24 h. Once a genome sequence has been annotated by RAST, users can utilize powerful tools for manual curation of annotations before proceeding with the subsequent steps in the pipeline. The pipeline continues with the 'preliminary reconstruction' step, which uses the RAST annotations to generate a preliminary model for each organism (Online Methods). These preliminary models consist of a reaction network complete with GPR associations, predicted Gibbs free energy of reaction values and an organism-specific biomass reaction including nonuniversal cofactors, lipids and cell wall components. Each preliminary model network includes all reactions associated with one or more enzymes encoded in the organism's genome as well as a set of spontaneous reactions that do not require enzymatic catalysis (**Supplementary Table 1**).

The GPR associations for each reaction in the network are generated based on the genome annotations and a mapping between biochemical reactions and the standardized functional roles assigned to genes during RAST annotation[7]. This mapping is used to differentiate between cases where protein products from multiple genes form a complex to catalyze a reaction, and cases where protein products from multiple genes can independently catalyze the same reaction (Online Methods). Although these GPR

[1]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois, USA. [2]Department of Computer Science, Hope College, Holland, Michigan, USA. [3]Department of Biology, Hope College, Holland, Michigan, USA. [4]Computer Science Department and Computation Institute, University of Chicago, Chicago, Illinois, USA. [5]Computing, Environment, and Life Sciences Directorate, Argonne National Laboratory, Argonne, Illinois, USA. Correspondence should be addressed to C.S.H. (chenry@mcs.anl.gov).

Received 22 February; accepted 30 July; published online 29 August 2010; doi:10.1038/nbt.1672

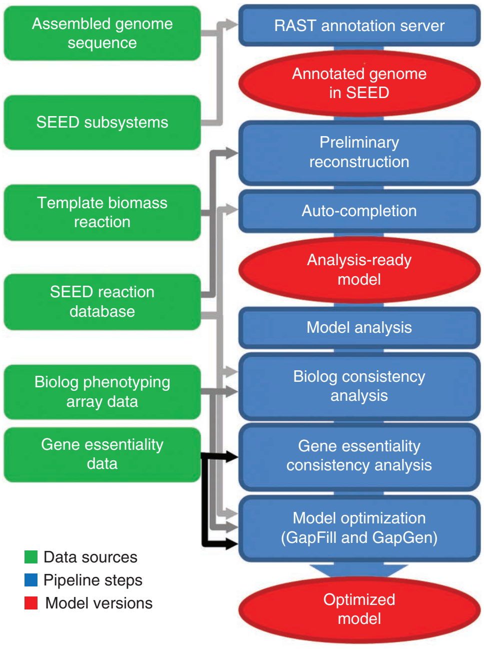


Figure 1 Model SEED genome-scale metabolic reconstruction pipeline. In the first step of the Model SEED pipeline, the assembled genome sequence is annotated by the RAST server and imported into the SEED analysis system. Next, a preliminary model is generated consisting of intracellular and transport reactions associated with genes on the basis of RAST annotations, spontaneous reactions and an organism-specific biomass reaction. In the auto-completion step of the pipeline, additional intracellular and transport reactions are added to create an analysis-ready model capable of simulating biomass production using only transportable nutrients. FBA is then used to generate phenotype predictions in the model analysis step. The final three steps of the pipeline involve the removal and addition of reactions from the model to fit Biolog phenotyping array data (when available) and gene essentiality data (when available) to produce an optimized model.

associations are generated based on well-curated annotations, they should be visually inspected to ensure accuracy.

The cofactor specificity of enzymes is also determined in the draft models based on genome annotations. For example, if an enzyme is known to use NADP$^+$ as an electron acceptor, then the functional role assigned to the associated gene will contain this information (e.g., "Non-phosphorylating glyceraldehyde-3-phosphate dehydrogenase (NADP) (EC 1.2.1.9)"); this functional role will subsequently be associated with a biochemical reaction that specifies NADP$^+$ as the cofactor. If the cofactor is unknown, then a standard cofactor is used (such as NAD$^+$). As annotation of cofactor specificity is often imprecise, cofactors should be visually inspected to ensure that the correct cofactors are used for the organism being modeled.

In metabolic models, biomass reactions are included to enable the simulation of cell growth and division via the simultaneous production of all small-molecule building blocks of biomass (e.g., amino acids, lipids, nucleotides and cofactors); the product of the biomass reaction is one gram of biomass, whereas the reactants are the constituent metabolites that combine to form one gram of biomass. During the preliminary reconstruction phase, the pipeline generates an organism-specific draft biomass reaction based on a reaction template (**Supplementary Table 2**). When a component of biomass is nonuniversal (e.g., cofactors, cell wall components), this template includes criteria specifying the metabolic subsystems and functional roles a genome must contain for the component to be added to the organism-specific biomass reaction. This template was tailored to produce nearly complete biomass reactions for the 130 demonstration organisms, but the nonuniversal portions of this template will require expansion to produce complete biomass reactions for some families of bacteria not included in the demonstration set (e.g., cyanobacteria). Biomass reactions must also include stoichiometric coefficients that indicate the relative abundance of each small molecule in the total biomass of an organism. Because the experimental data required to calculate these coefficients are not typically available, the Model SEED employs a set of rules to produce approximate coefficients for each biomass reaction (see Online Methods). These coefficients must be adjusted and fit to available experimental data before draft models may be used to produce quantitative predictions of organism growth rates. The approximate coefficients generated by the Model SEED are only sufficient for qualitative predictions of the conditions in which an organism will grow[14].

Automatic completion of model gaps

The models generated during the preliminary reconstruction usually contain gaps that prevent the production of one or more components of the biomass reaction. In the 'auto-completion' process, an optimization algorithm identifies the minimal set of reactions that must be added to each model to fill these gaps[10,15]. The reactions added during this process are different for each draft model, as metabolic requirements vary among organisms and different genome annotations contain different gaps. Reactions are selected from a comprehensive database of mass- and charge-balanced reactions standardized to aqueous conditions at neutral pH. This database combines all

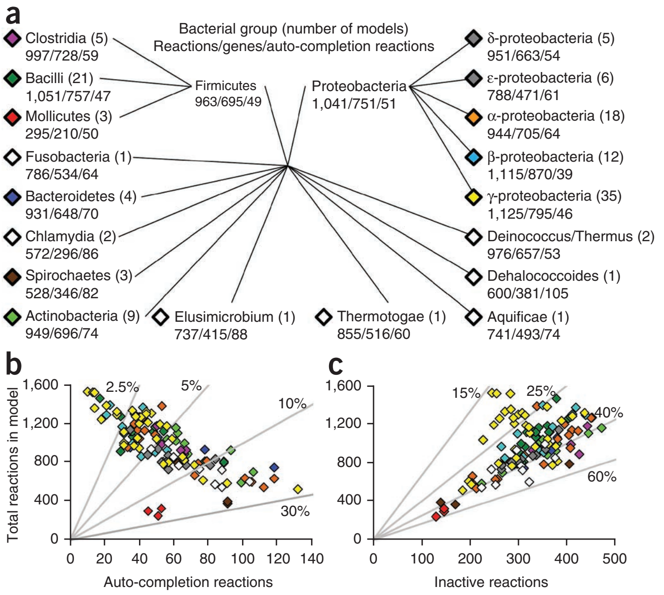


Figure 2 Properties of SEED models organized by taxonomy. (**a**) The 18 taxonomic groups containing the SEED models are displayed along with the number of models contained within each group and the average number of reactions, genes and auto-completion reactions included within the group models. The tree is arranged such that closely related taxonomic groups are co-localized. (**b**,**c**) Total number of reactions in each SEED model plotted against the number of reactions added during the auto-completion process (**b**) and the number of reactions that are inactive in FBA (**c**). Each point corresponds to a single SEED model, and the points are color coded by the taxonomic groups listed in **a**.

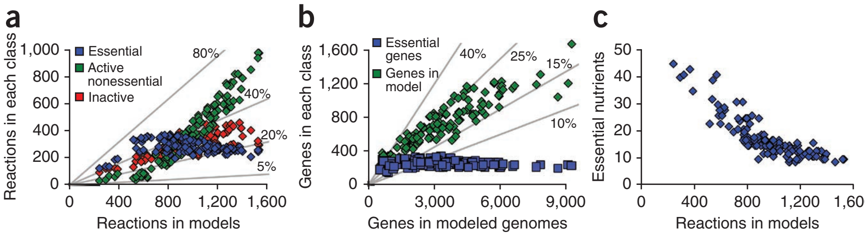


Figure 3 Properties of SEED models predicted using FBA. (**a**) FBA was used to classify each reaction in each model as essential (blue), nonessential but capable of carrying flux (green) or incapable of carrying flux (red), and the number of reactions in each class was plotted against the total number of reactions in each SEED model. (**b**) Gene essentiality in the SEED models, as predicted using FBA, compared with the total number of genes included in the genome of each modeled organism. Number of essential genes, blue; number of genes in model, green. Lines indicate the percentage of total model reactions (**a**) or organism genes (**b**) that was captured in each region of the plots. (**c**) Number of essential nutrients that are required for growth of each SEED model, as predicted by FBA, compared with the total number of reactions for each model.

the biochemistry contained in the KEGG[16,17] and 13 published genome-scale metabolic models[10,18–29] into a single, nonredundant set. Because this database is standardized at neutral pH, model reactions may require adjustment when intracellular conditions deviate significantly from the standard. The auto-completion process ensures that every SEED model is capable of simulating cell growth, and it produces a list of metabolic functions predicted to be missing from the genome annotations (**Supplementary Table 3**).

When applied to our set of 130 demonstration organisms, the auto-completion process added an average of 56 reactions to each model (**Supplementary Table 3**). In general, the number of reactions added during auto-completion increased as the total number of reactions in the model decreased (**Fig. 2b**). One explanation for this trend is that many of the smaller models are associated with endosymbiotic or pathogenic organisms, which depend upon host cells to perform many metabolic functions. As a result, these organisms import many essential metabolites rather than synthesizing them *de novo*, and poorly annotated transporters are often missing from preliminary reconstructions. For example, our model of the endosymbiont *Buchnera aphidicola*, which consisted of only 517 reactions before auto-completion, required the largest number of auto-completion reactions (132 reactions). Many of the reactions added involve the transport of essential metabolites for which biosynthesis pathways appear to be lacking. Some of the intracellular reactions added represent metabolic functions that are predicted to be missing from the *B. aphidicola* annotations. However, most represent metabolic functions that are provided to *B. aphidicola* by its host (e.g., lipopolysaccharide biosynthesis pathways)[30]. This result demonstrates how functions added during the auto-completion process suggest hypotheses about metabolic interactions between obligate intracellular organisms and their hosts.

The auto-completion results also enable the identification of regions of the metabolic network where gaps in the genome annotations for the 130 organisms appear to be most prevalent. Over 50% of the reactions added to the SEED models during the auto-completion process are associated with metabolic processes involved in either cofactor biosynthesis (ubiquinone biosynthesis, menaquinone and phylloquinone biosynthesis and thiamin biosynthesis) or cell wall biosynthesis (LOS core oligosaccharide biosynthesis, teichoic and lipoteichoic acids biosynthesis and KDO2-lipid A biosynthesis). This explains the notable exception, involving the three mollicute models (red points in **Fig. 2b**), to the inverse relationship between the number of auto-completion reactions and the model size. Because these mollicutes lack a cell wall, none of the cell wall biosynthesis reactions were added during the auto-completion process.

As a case study for how auto-completion results can drive the improvement of genome annotations, we performed a directed search to identify genes responsible for a reaction added to the *Mycobacterium tuberculosis* H37Rv model during the auto-completion process (namely, *2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate synthase (EC 4.2.99.20)*). This reaction performs an essential step of the menaquinone biosynthesis pathway. In our directed search, all genes annotated with this reaction in other genomes were identified, and BLASTP was used to search for homologs for these genes in the *M. tuberculosis* genome. One such gene, *Rcas_1310* in *R. castenholzi*, was found to be homologous with *Rv0554* in *M. tuberculosis* with an e-value of 2.2×10^{-8}. The *Rv0554* gene clusters on the *M. tuberculosis* genome with six other genes involved in the menaquinone biosynthesis pathway, lending additional confidence to this functional assignment. To our knowledge, this is the first time the *4.2.99.20* activity has been associated with a gene in *M. tuberculosis*, and this association fills an important gap in a required metabolic pathway.

Model analysis

We call the draft models generated by the preliminary reconstruction and auto-completion processes 'analysis-ready' because they can simulate the production of biomass from transportable nutrients. On average, our 130 demonstration models include 965 reactions (**Fig. 3a**), 688 genes (**Fig. 3b**) and 876 metabolites. In the 'model analysis' step of the Model SEED pipeline, flux variability analysis (FVA)[31] is used to classify the reactions in the SEED models as essential, active or inactive (**Figs. 2c** and **3a**). Reactions classified as inactive cannot carry flux during simulated growth and are indicative of gaps in the metabolic network where additional manual curation is required. In the 130 SEED models, the average fraction of inactive reactions is 31.7% (**Fig. 2c**); not surprisingly, this is larger than the fraction of inactive reactions typically found in manually refined published models (16%). The remaining reactions that are not inactive in the SEED models are classified as either essential (if they must carry flux for growth to occur) or active (if they can carry flux but aren't essential for growth). The smaller SEED models tend to have fewer essential reactions (**Fig. 3a**), which is likely a result of metabolites being imported rather than synthesized and of biomass reactions involving fewer cofactors.

Flux balance analysis (FBA) is also used in the model analysis process to predict the essential genes in the SEED models. Despite wide variations in the genome sizes of our demonstration organisms, the number of essential metabolic genes remained relatively constant around an average value of 237 (**Fig. 3b**). This result implies that bacteria with larger genomes do not maintain redundant copies of essential genes to improve robustness. This conclusion is supported by previous studies[32], which reveal that larger genomes include a greater fraction of genes encoding secondary metabolic functions, transcriptional control and signaling mechanisms to improve versatility. Although the number of predicted essential genes remained relatively constant across all models, the specific reactions associated with these genes varied substantially. Only 47 reactions were associated with essential genes in nearly every model analyzed,

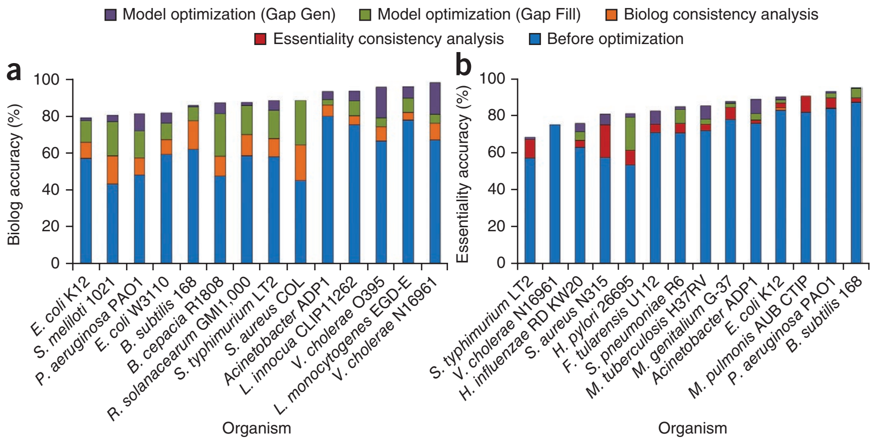


Figure 4 Accuracy of models generated by the Model SEED pipeline. (**a**,**b**) The accuracy of the SEED models in predicting Biolog phenotyping array data (**a**) and gene essentiality data (**b**) steadily improved during the model-refining steps of the pipeline. Before optimization (blue bars), the SEED models had an average overall accuracy of 66%; this increased to 71% after the Biolog consistency analysis (orange bars), to 75% after the gene essentiality consistency analysis (red bars), to 83% after the GapFill stage of the model optimization (green bars) and to 88% after the GapGen stage of the model optimization (purple bars). The gene essentiality consistency analysis affected only the GPR associations in the models, so it did not affect the accuracy of the Biolog phenotyping array predictions.

whereas 740 reactions were associated with essential genes in fewer than ten models analyzed (**Supplementary Table 4**).

Flux balance analysis is also used in the model analysis step to identify the nutrients that are essential for growth in each SEED model. In general, the number of essential nutrients decreases as the number of reactions in the models increases (**Fig. 3c**). Although defined growth conditions are unknown for many of the modeled organisms, this analysis reveals a wide range of predicted nutrient requirements. These predictions are invaluable to efforts to culture these organisms in defined media conditions[1]. All predictions generated from the SEED models are available on the Model SEED website.

Comparison with existing models and phenotype data

Biolog phenotyping arrays[18,20,21,33–35] and gene essentiality data sets[36,37] are available for 22 of the 130 demonstration organisms, and these data sets were used to validate and optimize the models for these organisms (**Fig. 4**). After the auto-completion process, the models had an average predictive accuracy of 60% for Biolog data, 72% for essentiality data and 66% overall (blue bars in **Fig. 4**). A modified version of the Growmatch algorithm[2] was included in the Model SEED pipeline to identify and correct the possible errors in the models that cause the incorrect predictions. This model optimization process consists of four steps: (i) Biolog consistency analysis to identify missing transport reactions; (ii) gene essentiality consistency analysis to identify conflicts between GPR relationships and essentiality data; (iii) gap filling to identify overconstrained or missing reactions; and (iv) gap generation to address underconstrained or extra reactions. These four optimization steps improved the average accuracy of the SEED models to 89% for Biolog data, 85% for essentiality data and 87% overall (**Fig. 4**, **Supplementary Methods** and **Supplementary Tables 5–7**). No genome-scale metabolic models have been published for eight of the organisms with available Biolog data and four of the organisms with available gene essentiality data. Nonetheless, the draft models of these organisms are as accurate as the draft models of organisms for which published models do exist (**Table 1**).

Table 1 Prediction accuracy of SEED models

Organism	Published model exists	Biolog accuracy (%)		Essentiality accuracy (%)	
		Original	Optimized	Original	Optimized
B. cepacia R1808	No	47.5	87.3	–	–
E. coli W3110	No	59.3	81.8	–	–
F. tularensis U112	No	–	–	70.9	82.5
L. innocua CLIP11262	No	75.5	93.8	–	–
L. monocytogenes EGD	No	77.8	96.0	–	–
M. pulmonis AUB CTIP	No	–	–	81.8	90.5
R. solanacearum GMI	No	58.6	87.7	–	–
S. meliloti 1021	No	43.2	80.6	–	–
S. pneumoniae R6	No	–	–	70.6	84.8
V. cholerae N16961	No	67.1	98.2	75.0	75.0
V. cholerae O395	No	66.5	95.7	–	–
New model average	No	61.9	90.1	74.6	83.2
Acinetobacter ADP1	Yes	80.0	93.3	75.7	88.8
B. subtilis 168	Yes	62.0	86.0	87.2	95.0
E. coli K12	Yes	57.1	79.3	82.7	89.9
H. influenzae RD KW20	Yes	–	–	62.9	75.7
H. pylori 26695	Yes	–	–	53.2	80.9
M. genitalium G-37	Yes	–	–	77.7	87.5
M. tuberculosis H37RV	Yes	–	–	71.9	85.1
P. aeruginosa PAO1	Yes	48.1	81.5	83.9	92.9
S. aureus COL	Yes	45.2	88.7	–	–
S. aureus N315	Yes	–	–	57.3	80.6
S. typhimurium LT2	Yes	58.0	88.6	57.0	68.2
Models with published counterpart average	Yes	58.4	86.2	71.0	84.5

Empty elements in the table indicate a lack of Biolog or essentiality data for the corresponding organism.

Genome-scale models have already been published for 19 of the organisms selected for metabolic reconstruction by the Model SEED pipeline[10,18–29]. Comparison of SEED models with their published counterparts shows that, on average, 86% of the genes in the published models are also included in the SEED models (**Supplementary Methods** and **Supplementary Table 8**). Most genes found exclusively in the published models were not included in the SEED models because either the functions assigned to these genes in the SEED are inconsistent with the reactions mapped to them in the published models or the functions are not specific enough to allow for mapping to explicit reactions. One example of additional content included in the SEED models that was not included in the published models is the sedoheptulose bisphosphate bypass in *Escherichia coli*. This bypass, exclusively in the SEED *E. coli* model, converts D-erythrose

Table 2 Example uses of the Model SEED resource

Research question	Unique capability of Model SEED	Insights or results generated
What are the essential genes in my newly sequenced organism?	Functioning draft models enable essential genes to be predicted.	357 correctly predicted essential genes in four microbes not previously modeled, and 30,316 essential genes predicted in all models
What defined culture conditions will my organism grow in?	Functioning metabolic models enable culture conditions to be predicted.	1,391 Biolog growth conditions correctly predicted in eight microbes not previously modeled, and essential nutrients predicted for all models
What are some global trends in microbial metabolic behavior?	Functioning draft models for many diverse microbes enable the exploration of such trends.	**Figures 2** and **3** show global trends in gene essentiality, reaction activity, essential nutrients and annotation gaps.
How accurate are the annotations for my organism of interest?	Functioning models convert annotations into predictions of experimentally observable phenotypes.	**Figure 4** shows the accuracy of models generated from annotations for 22 organisms based on comparison with experimentally observed phenotypes.
What are the knowledge gaps in genome annotation in general?	Recurring annotation gaps can be identified by comparing gaps found in every model.	Cofactor biosynthesis and cell wall biosynthesis account for 50% of annotation gaps found (**Supplementary Tables 2** and **6**).
What alternative pathways are present in an organism's metabolic reaction network?	Comprehensive reaction database, functional role mappings and updated annotations enable identification of alternative pathways.	Sedoheptulose bisphosphate bypass identified in the pentose phosphate pathway of *E. coli*, which is unique to the SEED *E. coli* model. Bypass in *E. coli* experimentally confirmed in ref. 38.
How can I identify and fill the gaps in my genome annotations?	Directed searches may be performed for functions added during model auto-completion and optimization.	Auto-completion process identified EC 4.2.99.20 as missing in *M. tuberculosis*, and a directed search identified peg.554 (Rv0554) as a candidate for this function.

4-phosphate and dihydroxyacetone phosphate to D-sedoheptulose 7-phosphate in the pentose phosphate pathway. It has been experimentally demonstrated to exist in transaldolase-deficient *E. coli* mutants[38] and is associated with the secondary activities of two glycolytic enzymes (6-phosphofructokinase and fructose-bisphosphate aldolase).

Manual curation

When comparing the steps in the Model SEED pipeline (**Fig. 1**) with the steps outlined in the published metabolic reconstruction protocol[4], we found that the pipeline replicates 73 of the first 82 steps in the protocol. The preliminary reconstruction step of the pipeline automates most of the first 42 steps of the protocol. The only steps missing are experimental data collection, assigning gene and reaction localization (mostly for eukaryotic models), addition of intracellular transport reactions (SEED models only include cytosol and extracellular compartments), determination of biomass reaction coefficients and loading models into the COBRA toolbox. The auto-completion and model analysis portions of the Model SEED pipeline automate all of the protocol steps 43–66 and 67–80, respectively, with the only exception being reconnection of inactive reactions. The Model SEED does not attempt to reconnect inactive reactions because this requires manual curation to differentiate the inactive reactions that are a result of misannotation from those that should be reconnected. The model optimization process implemented in the Model SEED corresponds with steps 81–82 of the published protocol.

The models produced by the Model SEED still require some manual curation before they can match most published models in quality and accuracy. We have included a tutorial on this curation process within the Model SEED website and in the **Supplementary Methods**. The infrastructure provided in the Model SEED facilitates this curation process by providing a functioning draft model with testable predictions, enabling validation of models with experimental data and supporting comparison of models in the Model SEED database (including many published models). We are also developing tools to directly support the iterative refinement of draft models within the Model SEED website.

DISCUSSION

Here we demonstrate the Model SEED as a resource for the generation, optimization and analysis of draft genome-scale metabolic models for 130 taxonomically diverse bacteria. Unlike existing resources such as KEGG[16,17] or MetaCyc[39] that focus on cataloging gene functions, metabolic reactions and pathways, the Model SEED produces functioning metabolic models that not only describe what pathways are present but also predict how those pathways are used by each organism. These unique capabilities make the Model SEED a valuable resource for numerous applications in biology (**Table 2**). The model validation (**Fig. 4b** and **Supplementary Table 9**) and large-scale gene essentiality predictions (**Fig. 3a**) demonstrate that SEED models can correctly identify many essential metabolic genes. The Biolog prediction validation (**Fig. 4a**) and essential nutrient predictions (**Fig. 3c**) demonstrate that culture conditions can be predicted. The model optimization (**Fig. 4**) and auto-completion results (**Fig. 2**) show how the Model SEED is useful as a means of assessing annotation quality. And the global trends in model predictions and statistics (**Figs. 2** and **3**) demonstrate an ability to study universal trends in microbial behavior. By providing biologists with a means of rapidly producing a functioning draft metabolic model for an organism with the click of a button, the Model SEED makes genome-scale metabolic models more accessible to the wider scientific community. The Model SEED also enables the rapid rebuilding of models to integrate improved annotations and new experimental data. Rapid update of genome-scale metabolic models is essential for keeping up with the emergence of new high-throughput experimental data sets and for enabling researchers worldwide to benefit from new discoveries in organism metabolism.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

This work was supported by the US Department of Energy under contract DE-ACO2-06CH11357, by the National Institute of Allergy and Infectious Diseases under contract HHSN266200400042C and by the National Science Foundation under grants MCB-0745100, CCF-0829929 and DBI-0850546. M.D. was also supported by the Argonne National Laboratory Guest Faculty Program and the United States Fulbright Scholarship Program. We acknowledge the SEED annotators and development team for producing the annotations and computational infrastructure that make this work possible. We thank R. Overbeek, V. Vonstein, R. Olson, T. Disz, S. Devoid, F. Xia and T. Paczian for assistance with the use of the SEED genome annotation and analysis tools.

AUTHOR CONTRIBUTIONS

C.S.H. developed and operated the Model SEED resource and the related algorithms with assistance from all other authors. All authors participated in curation and testing of the SEED models. M.D. contributed significantly to reaction network annotation, model testing and model curation. A.A.B. contributed significantly to template biomass reaction design, model curation and selection of organisms for reconstruction. P.M.F. contributed to reaction network annotation. R.L.S. contributed to model curation and the design of pipeline and algorithms. B.L. contributed to Model SEED web interface development. All authors contributed to the writing and revision of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Yus, E. *et al*. Impact of genome reduction on bacterial metabolism and its regulation. *Science* **326**, 1263–1268 (2009).
2. Kumar, V.S. & Maranas, C.D. GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. *PLoS Comput. Biol.* **5**, e1000308 (2009).
3. Feist, A.M. & Palsson, B.O. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat. Biotechnol.* **26**, 659–667 (2008).
4. Thiele, I. & Palsson, B. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* **5**, 93–121 (2010).
5. Overbeek, R., Disz, T. & Stevens, R. The SEED: A peer-to-peer environment for genome annotation. *Commun. ACM* **47**, 46–51 (2004).
6. Aziz, R.K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).
7. DeJongh, M. *et al.* Toward the automated generation of genome-scale metabolic networks in the SEED. *BMC Bioinformatics* **8**, 139 (2007).
8. Jankowski, M.D., Henry, C.S., Broadbelt, L.J. & Hatzimanikatis, V. Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys. J.* **95**, 1487–1499 (2008).
9. Henry, C.S., Zinner, J., Cohoon, M. & Stevens, R. *i*Bsu1103: a new genome scale metabolic model of *B. subtilis* based on SEED annotations. *Genome Biol.* **10**, R69 (2009).
10. Suthers, P.F. *et al.* A genome-scale metabolic reconstruction of *Mycoplasma genitalium, i*PS189. *PLOS Comput. Biol.* **5**, e1000285 (2009).
11. Notebaart, R.A., van Enckevort, F.H., Francke, C., Siezen, R.J. & Teusink, B. Accelerating the reconstruction of genome-scale metabolic networks. *BMC Bioinformatics* **7**, 296 (2006).
12. Tsoka, S., Simon, D. & Ouzounis, C.A. Automated metabolic reconstruction for *Methanococcus jannaschii*. *Archaea* **1**, 223–229 (2004).
13. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, W182–W185 (2007).
14. Pramanik, J. & Keasling, J.D. Effect of *Escherichia coli* biomass composition on central metabolic fluxes predicted by a stoichiometric model. *Biotechnol. Bioeng.* **60**, 230–238 (1998).
15. Satish Kumar, V., Dasika, M.S. & Maranas, C.D. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics* **8**, 212 (2007).
16. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
17. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30**, 42–46 (2002).
18. Feist, A.M. *et al.* A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* **3**, 121 (2007).
19. Reed, J.L., Vo, T.D., Schilling, C.H. & Palsson, B.O. An expanded genome-scale model of *Escherichia coli* K-12 (*i*JR904 GSM/GPR). *Genome Biol.* **4**, R54 (2003).
20. Durot, M. *et al.* Iterative reconstruction of a global metabolic model of *Acinetobacter baylyi* ADP1 using high-throughput growth phenotype and gene essentiality data. *BMC Syst. Biol.* **2**, 85 (2008).
21. Oh, Y.K., Palsson, B.O., Park, S.M., Schilling, C.H. & Mahadevan, R. Genome-scale reconstruction of metabolic network in *Bacillus subtilis* based on high-throughput phenotyping and gene essentiality data. *J. Biol. Chem.* **282**, 28791–28799 (2007).
22. Goelzer, A. *et al.* Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of *Bacillus subtilis*. *BMC Syst. Biol.* **2**, 20 (2008).
23. Schilling, C.H. *et al.* Genome-scale metabolic model of *Helicobacter pylori* 26695. *J. Bacteriol.* **184**, 4582–4593 (2002).
24. Oliveira, A.P., Nielsen, J. & Forster, J. Modeling *Lactococcus lactis* using a genome-scale flux model. *BMC Microbiol.* **5**, 39 (2005).
25. Feist, A.M., Scholten, J.C., Palsson, B.O., Brockman, F.J. & Ideker, T. Modeling methanogenesis with a genome-scale metabolic reconstruction of. *Methanosarcina barkeri*. *Mol. Syst. Biol.* **2**, 2006 0004 (2006).
26. Jamshidi, N. & Palsson, B.O. Investigating the metabolic capabilities of *Mycobacterium tuberculosis* H37Rv using the in silico strain *i*NJ661 and proposing alternative drug targets. *BMC Syst. Biol.* **1**, 26 (2007).
27. Nogales, J., Palsson, B.O. & Thiele, I. A genome-scale metabolic reconstruction of *Pseudomonas putida* KT2440: *i*JN746 as a cell factory. *BMC Syst. Biol.* **2**, 79 (2008).
28. Duarte, N.C., Herrgard, M.J. & Palsson, B.O. Reconstruction and validation of *Saccharomyces cerevisiae i*ND750, a fully compartmentalized genome-scale metabolic model. *Genome Res.* **14**, 1298–1309 (2004).
29. Becker, S.A. & Palsson, B.O. Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: an initial draft to the two-dimensional annotation. *BMC Microbiol.* **5**, 8 (2005).
30. Douglas, A.E. Nutritional interactions in insect-microbial symbioses: aphids and their symbiotic bacteria *Buchnera*. *Annu. Rev. Entomol.* **43**, 17–37 (1998).
31. Mahadevan, R. & Schilling, C.H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* **5**, 264–276 (2003).
32. Konstantinidis, K.T. & Tiedje, J.M. Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc. Natl. Acad. Sci. USA* **101**, 3160–3165 (2004).
33. von Eiff, C. *et al.* Phenotype microarray profiling of *Staphylococcus aureus* menD and hemB mutants with the small-colony-variant phenotype. *J. Bacteriol.* **188**, 687–693 (2006).
34. Bochner, B.R. Global phenotypic characterization of bacteria. *FEMS Microbiol. Rev.* **33**, 191–205 (2009).
35. Keymer, D.P., Miller, M.C., Schoolnik, G.K. & Boehm, A.B. Genomic and phenotypic diversity of coastal *Vibrio cholerae* strains is linked to environmental factors. *Appl. Environ. Microbiol.* **73**, 3705–3714 (2007).
36. Gerdes, S. *et al.* Essential genes on metabolic maps. *Curr. Opin. Biotechnol.* **17**, 448–456 (2006).
37. Zhang, R., Ou, H.Y. & Zhang, C.T. DEG: a database of essential genes. *Nucleic Acids Res.* **32**, D271–D272 (2004).
38. Nakahigashi, K. *et al.* Systematic phenome analysis of *Escherichia coli* multiple-knockout mutants reveals hidden reactions in central carbon metabolism. *Mol. Syst. Biol.* **5**, 306 (2009).
39. Karp, P.D., Riley, M., Paley, S.M. & Pellegrini-Toole, A. The MetaCyc Database. *Nucleic Acids Res.* **30**, 59–61 (2002).

ONLINE METHODS

The Model SEED pipeline consists of seven consecutively applied steps (Fig. 1): (i) annotation; (ii) preliminary reconstruction; (iii) auto-completion; (iv) FBA analysis; (v) Biolog consistency analysis; (vi) gene essentiality consistency analysis; and (vii) reaction network optimization. Each of these steps is described in detail below.

Model SEED reconstruction pipeline: preliminary reconstruction. The Model SEED reconstruction pipeline produces analysis-ready genome-scale metabolic models starting with high-quality genome annotation in the context of the SEED framework, and optimizes them when phenotype and gene essentiality data are available (**Fig. 1**). In the first step of this pipeline, the assembled genome sequence is annotated by the RAST server and imported into the SEED. In the second step, a preliminary metabolic model is constructed consisting of (i) the spontaneous reactions, enzymatic reactions and transport reactions that make up an organism's metabolism; (ii) the set of GPR relationships that describe how reaction activity depends upon an organism's genes; and (iii) a biomass reaction that describes the essential small-molecule building blocks of the organism. Enzymatic intracellular and transmembrane transport reactions are included in the preliminary model if one or more of the functional roles associated with these reactions in the SEED (http://www.theseed.org/models/) have been assigned to one or more of the genes in the annotated genome. The functional role-to-reaction mappings in the SEED are used to construct the GPR relationships that encode how genes work together to form the protein complexes that catalyze enzymatic reactions. Additionally, if neighboring nonhomologous genes are associated with the same reaction in a model, the protein products for these genes are also assumed to function together in a single enzyme complex. These GPR relationships are essential for correctly predicting the impact of gene knockout on organism viability and behavior by using a genome-scale model. The biomass reaction in the preliminary model is assembled based on the template biomass reaction in the SEED (**Supplementary Table 2**), which was constructed from a curation of the biomass reactions included in 19 existing genome-scale metabolic models[10,18–29]. The template biomass reaction includes 83 small-molecule reactants, 39 of which are universal building blocks included in the biomass reaction of every organism (e.g., nucleotides for RNA and amino acids for protein). The remaining 44 reactants are included in a subset of the biomass reactions based on specific criteria that must be satisfied by evidence available in the annotated genome. These criteria include cell wall type (Gram positive, Gram negative, other) and subsystem variant codes that indicate specifically how an organism implements certain metabolic functions.

The stoichiometric coefficients in biomass reactions typically indicate the relative abundance of each small-molecule building block in an organism's biomass. Model SEED uses the following rules to generate stoichiometric coefficients that very roughly approximate the relative abundance of biomass components in each modeled organism: (i) relative abundances for amino acids, nucleotides, protein, DNA, RNA and cofactors are based on measured values in *E. coli*[18] for gram-negative organisms and *Bacillus subtilis*[21] for gram-positive organisms; (ii) a growth-associated ATP maintenance of 60 mmol per gram biomass per hour is assumed, which is approximately the value used in genome-scale models published to date; (iii) all cofactors are assumed to be present in equal mass; and (iv) the net mass of all biomass components sums to one gram.

Model SEED reconstruction pipeline: auto-completion. The preliminary metabolic models assembled during the second step of the Model SEED pipeline typically contain gaps in their reaction networks that prevent the production of one or more essential building blocks in the biomass reaction. As a result of these gaps, preliminary models are incapable of simulating cell growth under any conditions. In the third step of the Model SEED pipeline, these gaps are identified and eliminated through a process called auto-completion. In the auto-completion process, an optimization is performed to identify the minimal set of new reactions that must be added to the preliminary model to enable the production of biomass in the minimal confirmed growth medium for the modeled organism (**Supplementary Table 3**). If the minimal confirmed growth medium for an organism is unknown, any transportable metabolite is allowed to be consumed from the medium during the auto-completion process. The reactions added during the auto-completion process are selected from a comprehensive database of spontaneous reactions, enzymatic reactions and trans-membrane transport reactions maintained as a part of the SEED. This database consists of ~12,000 reactions and 15,044 compounds, and it combines all the biochemistry contained in the KEGG[16,17] and 13 published genome-scale metabolic models[10,18–29] into a single, nonredundant set. Often the gaps in the reaction network of a preliminary model may be filled by many different distinct sets of reactions. Equation (1) shows the novel objective function used in the auto-completion optimization to select for the set of reactions that represents the best possible hypothesis of what is actually missing from the genome annotations.

$$\text{Minimize}\sum_{i=0}^{R}\left(1+P_{T,i}+P_{K,i}+P_{SS,i}+P_{F,i}-f_{SS,i}-f_{p,i}\right)z_i \qquad (1)$$

In this objective function, z_i is a binary variable created for any reaction not currently included in the model. Separate z_i variables are created for the forward and reverse directions of each reaction, and if a reaction included in the model is irreversible, a z_i variable is introduced for the direction of the reaction not included in the model. Thus auto-completion solutions also involve making some existing reactions in the model reversible.

$P_{T,i}$ is a penalty on the addition of transport reactions during the auto-completion process. This penalty equals 4 for transport reactions involving compounds in the biomass reaction, 2 for all other transport reactions and 0 for intracellular reactions. This penalty ensures that completion of intracellular biosynthesis pathways is favored over the addition of transport reactions. $P_{K,i}$ is a penalty favoring addition of KEGG reactions. This penalty equals 0 for KEGG reactions and 2 for non-KEGG reactions. Addition of KEGG reactions is favored to avoid the addition of simplified lumped reactions included in many existing models. $P_{SS,i}$ is a penalty favoring the addition of reactions mapped to SEED functional roles and subsystems. This penalty equals 0 if the reaction is mapped to at least one functional role in a SEED subsystem, 1 if the reaction is mapped to at least one functional role not found in a subsystem and 3 if the reaction is not mapped to any functional roles. Reactions mapped to SEED functional roles and subsystems are favored because these reactions take part in the core pathways of metabolism and represent the most well-curated portion of the known biochemistry. $P_{f,i}$ is a penalty on the addition of reactions proceeding in a thermodynamically unfavorable direction. This penalty equals 0 if a reaction is proceeding in a favorable direction[9], and $5 + 0.1(\Delta_r G'^{\circ} - 10$ kcal/mol) if the reaction is proceeding in an unfavorable direction, where $\Delta_r G'^{\circ}$ is the estimated Gibbs free energy change of reaction[8]. If $\Delta_r G'^{\circ}$ cannot be calculated, $P_{f,i}$ equals 6 for unfavorable reactions. $f_{ss,i}$ is a bonus applied to reactions involved in subsystems already well represented in the preliminary model. $f_{ss,i}$ is equal to the number of reactions in the preliminary model associated with the subsystem over the total number of reactions in the database associated with the subsystem. Similarly, $f_{p,i}$ is a bonus applied to reactions involved in short linear pathways (called scenarios[7]) already well represented in the preliminary model. $f_{p,i}$ is equal to the number of reactions in the preliminary model associated with the scenario over the total number of reactions in the database associated with the scenario.

The auto-completion objective is combined with the following set of constraints to form a complete mixed integer linear optimization problem (MILP), which may then be solved directly using the CPLEX 11.1 optimization package typically in a few hours and nearly always in <24 h:

$$N_{Super}\bullet v = 0 \qquad (2)$$

$$0 \le v_i \le 1{,}000 z_i \;\; i = 1,\ldots,r \qquad (3)$$

$$v_{bio} > 10^{-3}\ \text{g/g CDW h} \qquad (4)$$

In the auto-completion optimization, equation (2) represents the mass balance constraints that enforce the quasi-steady-state assumption of FBA, N_{Super} is the stoichiometric matrix for the superset of KEGG/model reactions with reversible reactions decomposed into separate forward and backward components and v is the vector of fluxes through the superset reactions.

Equation (3) enforces the bounds on the reaction fluxes (v_i) and the values of the reaction use variables (z_i). Equation (4) forces the flux through the biomass reaction, v_{bio}, to a nonzero value, ensuring that the z_i variables associated with the reactions needed to enable model growth are set to 1 during the optimization. Once the auto-completion optimization has produced a set of z_i and v_i values that optimally satisfy all constraints, the reactions with z_i values equal to 1 are added to the preliminary model to produce an analysis-ready model. The abbreviation CDW in the units of equation (4) stands for cell dry weight.

Model SEED reconstruction pipeline: analysis-ready model optimization. The remaining steps of the Model SEED pipeline involve the optimization of the analysis-ready model to better fit any experimental growth phenotype data that are available. Because these steps of the pipeline require data for fitting, they can be applied only to those organisms for which experimental data exist. The first optimization step of the pipeline, called Biolog consistency analysis, is performed only for organisms with available Biolog phenotyping array data[34]. In this step, the list of nutrients for which transport reactions exist in the model is compared against the list of nutrients the organism is known to metabolize based on available Biolog phenotyping array data. If no transport reaction exists in the model for a nutrient that is known to be metabolized, the transport reaction associated with the nutrient is added to the model. Because the transport reactions added in this process are not associated with a gene, it is impossible to discern the specific mechanism used to drive the movement of the nutrient across the cell membrane. Therefore, every transport reaction added to the SEED models during the Biolog consistency analysis follows a mechanism of proton symport for negative ions and proton antiport for positive ions.

The second optimization step of the pipeline, called gene essentiality consistency analysis, is performed only for organisms with available gene essentiality data. In this step, the data are used to identify and correct errors in annotations and GPR relationships included in the analysis-ready model. An algorithm is used to automatically search for instances of inconsistency between model annotations and available gene essentiality data. Three types of inconsistency are examined during the consistency analysis: (i) identical functional roles are assigned to an essential gene and one or more nonessential genes, (ii) identical functional roles are assigned to multiple essential genes without indicating that the protein products of these genes form a complex and (iii) one or more essential genes and one or more nonessential genes are all annotated to encode portions of the same protein complex. Once inconsistent annotations are identified, they are grouped by associated metabolic function, and a variety of annotation corrections are automatically proposed. Proposed corrections are then manually reviewed for implementation in the model.

The third optimization step in the pipeline, called model optimization, involves using the GrowMatch algorithm[2] with additional global optimization steps as described[9]. The model optimization proceeds in two stages: (i) GapFill to correct errors in the model that prevent growth *in silico* when growth is observed *in vivo* (false-negative predictions) and (ii) GapGen to correct errors in the model that allow growth *in silico* when growth is not observed *in vivo* (false-positive predictions). In the GapFill stage, a series of mixed integer linear optimization problems (MILPs) is solved to produce a set of possible solutions. Each solution represents a minimal set of modifications to the model reaction network that results in a maximal reduction in false-positive predictions. The modifications proposed by the GapFill algorithm include the addition of new reactions to the model reaction network or switching an existing reaction from being irreversible to being reversible. The most physiologically reasonable solution is then manually identified for implementation in the refined model.

The GapGen stage of the model optimization is similar to the GapFill stage in that a series of MILPs is solved to produce a small number of solutions, one of which is manually selected for implementation to maximally reduce prediction errors. In the GapGen stage, however, false-positive predictions are eliminated, and reactions are made irreversible or removed entirely rather than being added. The GapGen stage of the model optimization provides a valuable means of identifying reactions in the models that were underconstrained by the reversibility prediction method used.

Model validation using FBA. FBA is first used in the Model SEED pipeline to verify that every model produced by the pipeline is ready for analysis, by confirming that the model is capable of simulating biomass production in the minimal defined growth medium for the modeled organism. If no minimal defined growth medium is known for the organism, FBA is used to ensure that the model is capable of simulating biomass production using only nutrients for which transmembrane transport reactions exist in the model.

In the assessment and optimization of the SEED models, FBA is used to calculate the maximum possible growth *in silico* for every experimental condition with available data. Model accuracy is assessed by determining that fraction of experimental conditions where the growth predicted *in silico* and growth observed *in vivo* are either both zero or both nonzero.

doi:10.1038/nbt.1672

Cleantech: brave new world or coming home?

Mari Paul

Cleantech can offer biotech professionals the same mission, means of support and use of talents and skills they can have in drug discovery and development.

For almost 20 years I have watched and encouraged biotech start-ups in medical research as they dream, strive and struggle against overwhelming odds to bring lifesaving new products to the market. With the economic downturn and the steep decrease in early-stage funding from the exodus of venture capitalists, it has been discouraging to see the industry drop new research start-ups and cut the ranks of the surviving biotechs, leaving much talent looking for employment. But every door closing makes us look for open windows.

Endeavoring to do some pathway engineering of my own, I asked Michael Arbige, vice president of technology at Genencor International and a member of BayBio's board of directors, his advice for seeking new career options, specifically in the now-hot field of renewable or green biotechnologies, otherwise known as "cleantech."

Mari Paul: Mike, you've seen biotech and cleantech grow from their beginnings at Genentech and Genencor. Can you give us a quick backgrounder?

Michael Arbige: Industrial biotechnology, or cleantech, is not a new revolution. It started in the late 1970s with the beginnings of the biotech industry. I did my first biomass experiment back then, and there were others who were already doing them. Around that time, many early biotech companies such as Genentech, Amgen and Biogen had industrial applications, and many large industrial companies like Kodak and Corning Glass also had biotechnology projects. Interestingly enough, creating the molecules and vast infrastructure needed to successfully commercialize products for industrial biotech was often a bigger challenge than for drugs.

One of the initial targets we considered the Holy Grail for industrial biotech was called calf chymosin, or rennin, which is an enzyme that turns milk into cheese. The market size for the molecule was similar to that for many of the early drug targets, but in the end it turned out to be very difficult to express and produce economically. Some of those drugs, such as insulin, were in fact much easier to produce, but we used many of the same skills and development theories with reninn that were used in making drugs.

Then came protein engineering. Genentech decided that the early drugs they were trying to commercialize might be difficult to get through FDA hurdles and that this technology might best be applied to an industrial target, so a group of us that included Genentech's best protein chemists, molecular biologists and process scientists started working with Procter & Gamble in modifying an enzyme called protease to create a cost-effective, novel product that could be used to create a better laundry detergent. These sciences continued to evolve, and we eventually moved on to creating protein production systems in fungi, again accessing cutting-edge scientists needed to make this successful. In the end we were producing tanker cars full of proteins every day just for this detergent product—more protein than Genentech produces in a year.

According to Michael Arbige, creating the molecules and vast infrastructure needed to successfully commercialize industrial biotech products was often a bigger challenge than for drugs.

These technologies were eventually applied to the production and manufacturing of peptides and small molecules like 1,3-propanediol, which we developed in conjunction with DuPont and is now at full-scale manufacturing. And today the buzz in biotech is around systems and synthetic biology using pathway engineering to optimize the delivery and effect of molecules—the same technologies and people skills being used in the cleantech space to make fuels. The only difference is the target and the speed with which you can get things to market. Some of our probiotics and products used in foods have a longer regulatory cycle, but some are simply regulated by the Environmental Protection Agency and can be faster to the market. We love to work in partnerships; our collaboration with P&G started with protease in the early 1980s and is the longest-running biotech-industrial partnership. Our initial efforts with P&G allowed them to take phosphates out of detergents, which had a huge environmental impact.

People working in cleantech want to solve some of the most pressing environmental challenges of our time, says Lisa Zanetto.

Our science relationships are pretty much the same as those of a medical biotech company. We're currently working with over 50 universities doing basic scientific research as well as across many scientific functions. Our major challenge is scalability, as regulatory issues are less of a challenge for us than for the drug companies. We also learned early on about cost containment as well as the value of saving the environment.

MP: I often cite three things a job has to do for you: provide a means of support; use your

Mari Paul is at Life Science Leaders, San Francisco, California, USA.
e-mail: mari@lifescienceleaders.com.

talents and skills; and give you a mission. I asked Lisa Zanetto, Genencor's director of human resources, to compare cleantech and biotech and how cleantech can use a biotech employee's talents and skills.

Lisa Zanetto: Our largest R&D site is in Palo Alto, California, in the heart of Silicon Valley, so we're lucky to have access to some of the best talent in the world. We often interview candidates in the areas of molecular biology, biochemistry, protein engineering, recovery, formulations and fermentation from the biotech and pharmaceutical industry.

One difference I found between careers at Genencor and at my former employer Galileo is the breadth of products you can be exposed to: at Galileo we were heavily involved with one or two products, whereas at Genencor you can work on many. But science is only one difference—inside the same science, one still has to adapt to each company's different business model, culture and leadership style.

MP: Can you comment on why someone might "come home" to cleantech?

LZ: Many of our employees join the company for the opportunity to change the world: our employees are deeply committed to finding answers to some of the most pressing environmental challenges of our time. They want to make a difference, and we provide them that opportunity. And they can see the impact of their contribution in a relatively short period of time. The most important thing for any of us is to find a value match in something one has a passion for and can thrive and be the best at.

MA: You may have heard of our recent project with Huntsman—a project that, if fully incorporated for pretreating all of the cotton produced worldwide, would save 10 trillion liters of water annually, which is more than 1,000 liters for each person on earth. This technology space includes all the big drivers of human survival: water, food, energy and the environment.

Also, the folks that come to this technology get to see the fruits of their labors. I personally have over 200 patents on products that people use every day to my and my teams' credit. I go home at Thanksgiving and look around the table and see a number of items that I have had an impact on creating, and I have a really good reason to be thankful.

MP: So the mission for both is clear. Some lives we can save with medicine, and some we can save with water or other solutions that preserve the environment. Cleantech can offer biotech professionals the same means of support, use of talents and skills, and mission that they can have working on drug products. Cleantech is less of a new world and more of a home to which we can return.

COMPETING FINANCIAL INTERESTS

The author declares no competing financial interests.

PEOPLE

Selecta Biosciences (Watertown, MA, USA) has named **Werner Cautreels** president, CEO and a member of the board of directors. Cautreels brings more than 25 years of experience as a pharmaceutical executive, most recently as CEO and global head of R&D of Solvay Pharmaceuticals until its acquisition by Abbott Laboratories in February.

"We are extremely pleased that Werner has become Selecta's CEO and was attracted to the vast potential of Selecta's nanotechnology to develop new classes of synthetic nanoparticle vaccines and immunotherapies," says George Siber, a member of Selecta's board and former CSO of Wyeth Vaccines. "Werner's leadership and success in business strategy and drug development will be a great asset as Selecta navigates the broad set of opportunities with its proprietary technology."

Christopher B. Begley has announced his intention to retire as CEO of Hospira (Lake Forest, IL, USA). Begley will serve as CEO until his successor is named, and will then remain as executive chairman of the board of directors, ensuring continuity of leadership and an orderly transition of his CEO responsibilities. In addition, chief operating officer **Terrence C. Kearney** has announced his intention to retire by the end of 2010. He will be succeeded by current vice president of supply chain **James H. Hardy**, who has been appointed senior vice president of operations effective at the end of December 2010.

Torbjorn Bjerke has been appointed CEO of Karolinska Development (Stockholm). He is currently CEO for Orexo, a position he has held since 2007, and will assume his new position at Karolinska Development when a successor is found. He succeeds **Conny Bogentoft**, who will assume the position of chief scientific officer at Karolinska.

Genocea Biosciences (Cambridge, MA, USA) has named **Chip Clark** to the position of chief business officer. He joins Genocea with 20 years of industry experience, most recently as co-founder and chief business officer of Vanda Pharmaceuticals.

Michael Giuffre has been appointed to the board of directors of DiaMedica (Winnipeg, Manitoba, Canada). Giuffre is a clinical professor of cardiac sciences and pediatrics at the University of Calgary and a board member of the Alberta Medical Association. As a biotechnology consultant, he has been involved with RedSky, MDMI and MedMira.

Charles A. Johnson has been appointed executive vice president of research and development and chief medical officer at Inspire Pharmaceuticals (Durham, NC, USA). Since 2007 he has served as chief medical officer at APT Pharmaceuticals. He previously spent 13 years at Genentech, where he held several senior leadership positions including vice president and head of the immunology and tissue repair clinical group.

Gregory L. Miller has joined Concert Pharmaceuticals (Lexington, MA, USA) as head of business and corporate development. He most recently served as senior director, business development and corporate strategy at AMAG Pharmaceuticals.

Avid Radiopharmaceuticals (Philadelphia, PA, USA) has announced the appointment of **Mark A. Mintun** to the newly created position of chief medical officer. Mintun brings extensive research, clinical and management experience to Avid, having most recently served as vice chair for research and director of the Center for Clinical Imaging Research at Mallinckrodt Institute of Radiology at Washington University School of Medicine.

SkyePharma (London) has announced that it has appointed **Axel Müller** as CEO, succeeding **Ken Cunningham**, who announced his intention to step down in May to focus on a portfolio of nonexecutive appointments. Müller has more than 25 years of experience in the pharmaceutical industry, most recently serving as CEO of Acino Holding from June 2008 to March 2010. Previously he was president of Siegfried Generics and managing director and vice president, international of Aceto Holding.

Human Genome Sciences (Rockville, MD, USA) has announced the appointment of **Tuomo Pätsi** as vice president, HGS Europe. He formerly served as regional vice president–Northern Europe at Celgene International.

OncoGenex Pharmaceuticals (Bothell, WA, USA) has appointed **David V. Smith** to its board of directors. He brings to the board substantial financial expertise in financial controls, analytics and process. He currently serves as executive vice president and chief financial officer of Thoratec.

NeoGenomics (Ft. Myers, FL, USA) has announced that **Mark W. Smits** has been named vice president of sales and marketing. Most recently, Smits was with Thermo Fisher as vice president for the Fisher Healthcare Division. In addition, NeoGenomics announced the promotion of **Grant Carlson** as vice president of business development.

Amarin Corporation (Dublin) has named **Colin W. Stewart** as president, CEO and a member of the company's board of directors. He has more than 30 years of experience in executive management and commercial positions for pharmaceutical companies, including five years as president and CEO of CollaGenex Pharmaceuticals. **Declan Doogan**, who had been serving as the Amarin's interim CEO, will continue as chief medical officer.

Invida Group (Singapore) has announced that it has added three new members to its senior management team: **Sumeet Sud** has been named to the newly created role of chief marketing officer, and **Girdhar Balwani** and **Thomas Birsinger** have been named country managers for India and Vietnam, respectively. Sud has over 18 years of experience at companies including Talecris Biotherapeutics, Pfizer and Merck. Balwani joins Invida with over 25 years of experience in the pharma industry, most recently at UCB as regional general manager for nine countries in Asia Pacific. Birsinger brings experience working in the US, Thailand, Vietnam and Korea on both the manufacturing and distribution sides of the business, most recently with Zuellig Pharma in Vietnam and Korea.