nature biotechnology

VOLUME 29 NUMBER 1 JANUARY 2011
www.nature.com/naturebiotechnology

THE SCIENCE AND BUSINESS OF BIOTECHNOLOGY

Haplotyping whole genomes

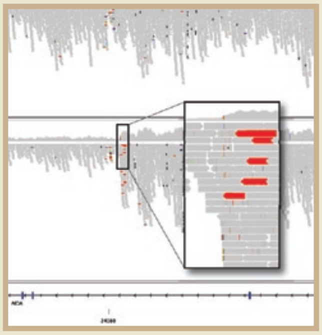
Genome-wide hydroxymethylation

Safe harbors in iPS cells

VOLUME 29 NUMBER 1 JANUARY 2011

nature biotechnology

The mixture of maternal and paternal sets of chromosomes in diploid organisms makes it difficult to determine haplotypes. Fan *et al*. and Kitzman *et al*. describe experimental approaches to genome-wide haplotyping using microfluidics and multiplexed large-insert cloning, respectively. Credit: Marina Corral & Erin Dewalt, based on "Colored LM of a normal male karyotype" by L. Willatt, East Anglian Regional Genetics Service/Photo Researchers, Inc.



Viewer for integrated genomics data, p. 24

EDITORIAL

NEWS

BIOENTREPRENEUR

BUILDING A BUSINESS

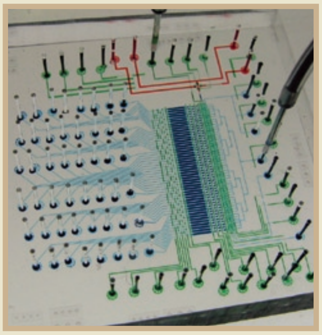
OPINION AND COMMENT

CORRESPONDENCE

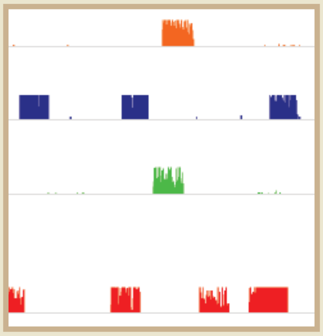
nature publishing group

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$4,048 (institution), US$4,658 (corporate institution). Canada add 5% GST #104911595RT001; Euro-zone: €202 (personal), €3,214 (institution), €4,011 (corporate institution); Rest of world (excluding China, Japan, Korea): £130 (personal), £2,077 (institution), £2,588 (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 © 2011 Nature America, Inc. All rights reserved. Printed in USA.

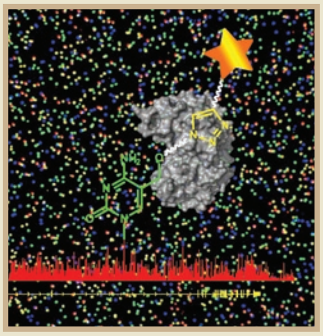
VOLUME 29 NUMBER 1 JANUARY 2011



Haplotyping with microfluidics, p 51



Haplotyping by multiplexed large-insert cloning, p 59



Global analysis of 5-hydroxymethylcytosine, p 68

COMMENTARY

FEATURE

PATENTS

NEWS AND VIEWS

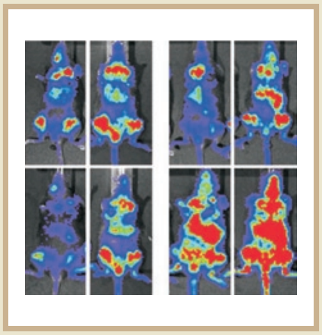
COMPUTATIONAL BIOLOGY

FEATURE

RESEARCH

ARTICLE

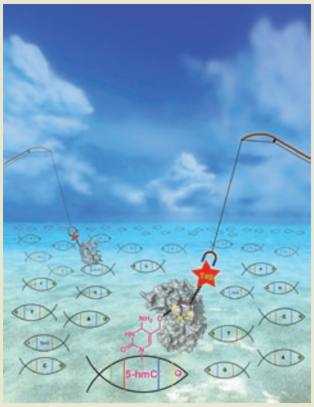
LETTERS



In vivo analysis of essential genes with shRNA, p 79

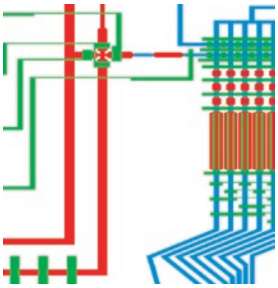
CAREERS AND RECRUITMENT

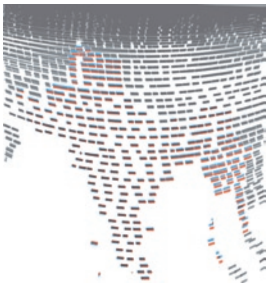
Genome-wide mapping of 5-hmC



Efforts to elucidate the functions of the recently discovered modification of mammalian DNA with 5-hydroxymethylcytosine (5-hmC) have been hampered by the lack of methods to map its distribution in the genome. Whereas high-throughput methods based on bisulfite sequencing or affinity purification have contributed to the understanding of the roles of other epigenetic modifications, such as DNA methylation and histone marks, no such method has been developed for 5-hmC. He and colleagues now use the T4 bacteriophage β-glucosyltransferase to chemically modify 5-hmC in a way that allows purification of 5-hmC DNA fragments. Normally, the enzyme transfers a glucose residue to the hydroxymethyl group of the modified cytosine but also accepts glucose molecules with minor modifications. Using 6-N$_3$-glucose and click chemistry, the authors biotinylate the 5-hmCs in genomic DNA and use the biotin tag to isolate hydroxymethylated DNA sequences. They identify 5-hmC in mammalian cell lines not previously known to contain the modification. By sequencing the isolated fragments, the authors create the first genome-wide maps of 5-hmC distribution in the cerebellum of adult mice and at postnatal day 7. They find an age-dependent and gene-specific increase in hydroxymethylation levels. **[Letters, p. 68]** *ME*

Haplotyping single cells and whole genomes

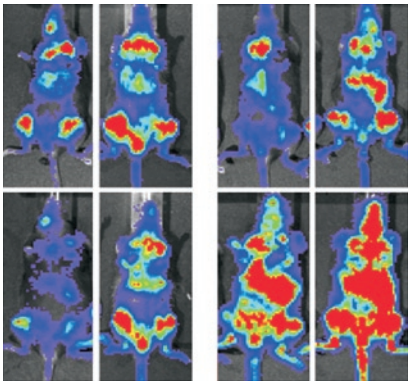




Two experimental approaches described in this issue make it possible to readily determine long haplotypes, in some cases of whole genomes from single cells. The process of haplotyping, called phasing, involves determining the combination of genetic variants that are present on each of the copies of homologous chromosomes in a diploid organism. Quake and colleagues separate and amplify the chromosomes of single cells using a microfluidic device. Phasing is readily achieved as chambers on the device contain either the maternal or paternal copy of a chromosome, which the authors analyze by genotyping microarray or by sequencing. Shendure and colleagues devise a strategy for multiplexed cloning and sequencing of large-insert genomic DNA clones. Clones are barcoded and pooled so that each pool is likely to contain DNA from only one copy of a chromosome, thereby enabling phasing by assembly of the clone sequences. The authors apply their method to sequence the first haplotype-resolved genome of a person with Gujarati Indian ancestry. Both methods allow phasing of rare variants, currently a major challenge, and both should be scalable, although Shendure's is more labor intensive than standard whole-genome sequencing and Quake's requires specialized devices. These approaches should be useful in genetics research, diagnostics and genomic analysis of single cells. [**Articles, p. 51; Letters, p. 59; News and Views, p. 38**] *CM*

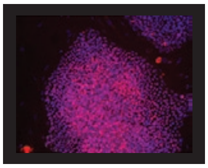
Written by Kathy Aschheim, Markus Elsner, Michael Francisco, Brady Huggett & Craig Mak

In vivo shRNA screening



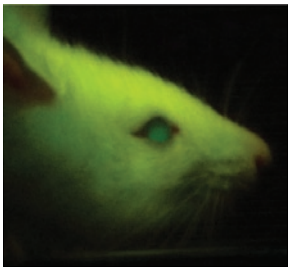
Acute gene knockdown in specific cell populations *in vivo* using short hairpin RNAs (shRNAs) is a powerful tool for elucidating gene function in a physiologically relevant context. The identification of genes that are important for survival and proliferation, and therefore present potential drug targets in cancer and other diseases, is complicated by the strong selective advantage of cells with inefficient shRNA expression. To facilitate the discovery of such genes, Lowe and colleagues have developed an inducible shRNA retroviral vector system that allows the identification of transduced and shRNA-producing cells and ensures robust transactivator expression. Based on the backbone of the microRNA miR30 and an shRNA under the control of the tetracycline-inducible promoter system, the construct encodes two fluorescent proteins. The expression of one fluorescent protein is coupled to the constitutively produced transactivator, thus marking transduced cells. The other is expressed upon tetracycline activation as part of the shRNA-containing transcript. Only cells with efficient shRNA expression will be positive for both fluorescent proteins. Robust transactivator expression can be ensured by either coupling it to an oncogene or by constructing a positive feedback loop. The authors show that their system can eradicate an aggressive cancer by inhibiting a single gene and substantially increase the sensitivity of detection for essential genes in a pooled shRNA experiment. [**Letters, p. 79**] *ME*

Safety in numbers



Inserting therapeutic transgenes in the human genome with randomly integrating viral vectors has been linked to several cases of leukemia and clonal cell amplification. A possible solution to this safety concern is genomic 'safe harbors'—loci for targeted integration that permit high transgene expression without perturbing the expression of endogenous genes. Thus far, a handful of particular loci have been studied. Sadelain and colleagues take a different approach, searching for safe harbor sites across the entire human genome. A genomic locus is deemed safe if it meets five criteria, which require that it is far from ultraconserved regions and genes, and especially far (300 kb) from cancer-related and microRNA genes. The authors generate induced pluripotent stem cells with cells from β-thalassemia patients and deliver the β-globin gene on a lentiviral vector. Analysis of ~5,840 integration sites shows that ~17% satisfy the five safe-harbor criteria. One clone carrying a single vector integration that meets the criteria expresses the β-globin transgene at 85% of the normal level after differentiation to erythroid cells. The integration does not appear to alter the expression of endogenous genes, as assessed by microarray analysis. Although this is a promising approach for defining genomic safe harbors, further validation is needed in long-term *in vivo* studies. [**Letters, p. 73; News and Views, p. 41**] *KA*

Shortcut to transgenic rodents



Cui and colleagues demonstrate the production of mice and rats with targeted mutations by injecting zinc finger nuclease (ZFN) mRNA and homologous donor DNA into zygotes. The method provides a faster alternative to embryonic stem cells for generating transgenic rodents by homologous recombination. Moreover, the approach can be applied to any rodent strain and to species for which embryonic stem cells have not been isolated. Transgenic animals have been generated for many years by injecting DNA into the pronuclei of zygotes, but the DNA integrates randomly rather than by homologous recombination. ZFNs can be readily targeted to a unique sequence in a mammalian genome. In the presence of donor DNA, sequence-specific ZFN-mediated cleavage stimulates homologous recombination by several orders of magnitude. [**Letters, p. 64; News and Views, p. 39**] *KA*

Patent roundup

Building a strong patent portfolio is the bedrock for any biotech business. But do you have a plan of action for when your intellectual property is infringed? [**Building a Business, p. 19**] *BH*

Two recent decisions by the Japanese Intellectual Property High Court bolster the market exclusivity period for brand biologic manufacturers. Tessensohn and Yamamoto guide innovators through the optimal process for receiving patent term extensions for their products. [**Patent Article, p. 34**] *MF*

Next month in nature biotechnology

- Genome editing in human cells with TALEs
- mRNA therapy
- Fc fusions for mucosal immunity
- Unnatural amino acids for controlling transcription

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
Production Editor: Amanda Crawford
Senior Illustrator: Katie Vicari
Illustrator: Marina Corral
Cover design: Erin DeWalt
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
Exectutive Editor: Veronique Kiermer
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

Similarity trials

A European guideline on biosimilar monoclonal antibodies suggests smaller trials with homogeneous, younger patient groups may suffice for marketing authorization.

The European regulatory authorities continue to make strides forward with biosimilars. Indeed, the latest advice from the European Medicines Agency (EMA) to those seeking to produce biosimilar monoclonal antibodies (mAbs) is really quite straightforward. In its draft *Guideline on Similar Biological Medicinal Products Containing Monoclonal Antibodies* released for consultation in November, the Committee for Medicinal Products for Human Use (CHMP) of the EMA, in essence, said two things. First, applicants seeking approval of biosimilar mAbs will have to undertake comparative clinical studies; they must show that the safety and efficacy performance of a biosimilar is indeed similar to that of a reference, originator compound. But, second, clinical work doesn't have to be too onerous, especially if applicants are smart about the way they choose the patient population. This latter point has received remarkably little attention but could be pivotal in shaping the biosimilars marketplace as the cost and complexity of running trials is the biggest barrier to market entry by generics companies.

The expensive part of the process of developing biosimilars was always going to be the clinical work. But the overarching guidelines that the CHMP published on biosimilars (not specifically mAbs) in 2005 were rather vague on the nature of those trials. Indeed, the guidelines barely deserved the name, so little guidance did they actually give, and so few lines did they draw. The 2005 CHMP documentation did little more than identify who companies needed to talk to at EMA when taking their product forward on, in essence, a case-by-case basis.

The 2010 antibody guidelines are much more specific. The key passages in the document still steer applicants firmly toward upfront discussion with the European regulator, which might be construed as a return to a case-by-case system. But the document also plants several highly significant signposts, all of which point in the same direction—that of simpler, smaller clinical studies.

The focus of the biosimilarity exercise, says the guideline, is to "demonstrate similar safety and efficacy compared to the reference product, not patient benefit *per se*." So the idea is not to reproduce the trial for the original approval, but to design a trial that shows the compounds are similar.

How would such a trial be done? Well the guideline has sage advice there, too.

First, choose a sensitive, experimental human model. In other words, select a clinical population where, from knowledge of the reference biologic, you would expect the drug to make a big impact. In that way, the similarity trial would be comparing one big impact with (one hopes) a second, similarly big impact. Comparing two large impacts would enable any differences to be more apparent.

Second, the guideline suggests using homogeneous patient populations. By taking variability out of the patient set, any variability between the brand and the biosimilar forms of the biologic would be more apparent, the CHMP argues. The guideline also points out that with a homogeneous population, the sample size needed to prove (or disprove) equivalence would be smaller and facilitate simpler interpretation. It warns against using patients who either have different disease severity or have been exposed to different prior treatments because these additional variables could complicate the interpretation of differences seen in the two drug arms of the study. The guideline even recommends using younger groups of patients where possible, because younger people would be less likely to be effected by "concomitant clinical conditions."

In short, a trial to establish similarity is quite unlike a trial designed to measure safety and efficacy.

Thus, with its new guideline on biosimilar mAbs, the EMA has not only provided specific and lucid advice to industry but, importantly, kept the debate about biosimilars and brands on a firm and rational footing. The US situation, in contrast, remains rather different.

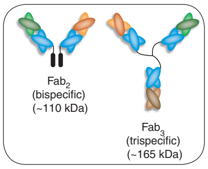
The 2010 Patient Protection and Affordable Care Act passed last March contained a section providing a legal framework for the approval of follow-on biologics. At the beginning of November, the US Food and Drug Administration held a two-day hearing in Silver Spring, Maryland, to discuss some of the issues with different stakeholders.

Understandably perhaps for a first meeting, much of the debate was polarized between generics firms, which called for low clinical hurdles to biosimilarity, and innovator companies, which called for high ones. There were polemics over whether a biosimilar approved in more than one indication would require comparative trials in each of those indications. There were distracting discussions on a phenomenon known as drift, wherein the manufacturing processes for a biologic and its biosimilar evolve away from each other. There were debates on whether trials should establish noninferiority or equivalence. And the possibility was discussed that a thorough analysis of post-translational modifications, three-dimensional structure and protein aggregation might obviate the need for clinical work altogether. That seems extraordinarily unlikely given the current state of technology and experience gained so far.

It is striking how far behind Europe the US regulatory pathway for follow-on biologics remains. Part of this is due to the weaker political impetus and greater lobbying strength of US-based innovator companies. But governmental momentum to push forward might change if it became evident that healthcare costs could be substantially reduced by means of follow-on products for expensive biologics and mAbs. If, as EMA suggests, abbreviated trials with homogeneous patient groups are sufficient to support a marketing authorization of a biosimilar, many more generics companies are likely to enter the biosimilars/follow-on market, increasing price competition and driving down healthcare costs.

Thus far, clinical trials for biosimilars in Europe haven't looked much smaller or less costly than those required for applying for a new marketing authorization. But the guideline on biosimilar mAbs provides the first tantalizing indication that smaller trials may indeed be possible.

IN this section



Boehringer buys into bispecific antibodies, p5

Biogen Idec's dramatic restructuring, p7

Spat over first GM mosquito trial, p9

Pfizer reaches out to academia—again

Pfizer is rolling out a grand plan to draw out drug-development-ready research from academia through a series of collaborations with leading medical centers worldwide. The first collaboration, announced in November, is with the University of California, San Francisco (UCSF), to which the pharma giant will commit $85 million. Coincidentally, London-based GlaxoSmithKline, is launching a similar outreach program, but with a very different approach. Like Pfizer, it wants to access leading academic researchers with targets ripe for translation into the clinic. Its scope, however, is more modest and targeted, focused on individual scientists.

For Pfizer, the overall aim in setting up these Global Centers for Therapeutic Innovation (CTIs) is to move novel biotherapeutics rapidly into human clinical trials—each project will aim to deliver a drug through phase 1 testing in five years. Pfizer expects five such initiatives to be up and running in 2011 in the United States, Europe and Asia. Assuming eight projects per CTI, this could bring dozens of differentiated biologics against new targets into the clinical pipeline.

The ambitious program departs from the traditional collaboration model. For starters, Pfizer will provide more than funding. The New York–based pharma will set up shop on each campus, contributing proprietary phage display libraries, peptide libraries and associated technologies for rapidly generating antibodies to be used as probes against the novel targets flagged by university researchers. Each CTI will be staffed with 20–25 Pfizer employees with expertise in cell-line generation, protein characterization and purification—the skill sets needed to rapidly identify and advance molecules into the clinic. All decision making, from the initial acceptance of proposals through the determination to start clinical testing, will be made by a joint steering committee. "The concept is to make a transition away from the vertically integrated R&D model into smaller, decentralized groups of a truly global nature," says Pfizer's Anthony Coyle, who is heading up the program out of the company's Cambridge, Massachusetts, facilities.

Pfizer's first Center for Therapeutic Innovation will sit on UCSF's Mission Bay campus (pictured), already home to several dozen biotech start-ups and an incubator network.

As important, the CTI model creates a 50-50 joint relationship where the goals of the investigators and the company are aligned and both sides are empowered to succeed. "There has got to be a change in the mindset from 'We own this, you do this for us,'" says Coyle. The CTIs will seek out investigators who have already developed a hypothesis around a novel disease mechanism and are keen to translate their discoveries into drugs. "These will be projects where we can articulate very clearly at the beginning what the first-in-man study will be," says Coyle. The strategy will be to define the mechanism of action and in parallel develop the appropriate drug to hit the target and also determine the right patient population to target with it.

The model "allows us to leverage all of the drug discovery capability in our organization—the ability to make clinical grade material, the finances to perform the right enabling toxicology studies, and the regulatory support to allow the investigator to realize the ambition and see the concept translated," says Coyle. He also hopes to bypass animal modeling. "What's becoming clear to me is that the time you spend on *in vivo* validation has zero impact, in most cases, on whether you will be successful going into the clinic. Here we propose to define the mechanism based on a human *in vitro* system, very quickly, which is again aided by having our phage library right there with the individuals doing the research."

Funding for CTI initiatives will follow a prenegotiated template Pfizer will put down at each institution. The company will pay for one to three post docs for each participating laboratory and the steering committee will have access to a flexible fund used either for additional biology or to allow the joint project team to move a compound into trials. There will be two clinical milestone payments, at proof of mechanism and successful proof of concept. All joint inventions will be jointly owned, with Pfizer holding an exclusive option to license a drug after proof of mechanism. In the event Pfizer exercises its option, any jointly developed enabling intellectual property (IP) would be licensed from the institution. If Pfizer declines, IP and other joint assets revert to the institution, which could then partner with someone else.

"There's going to be less of an establishment of value going into this, and more of it saved for the negotiation about the IP, which is downstream," says S. Claiborne Johnston, director of the UCSF Clinical and Translational Science Institute.

In the past, most collaborations, however, have failed to lead to new drugs. "I think they generally have failed because of the misalignment

IN brief

Amgen's bone-metastasis win

Xgeva for mets

The US Food and Drug Administration approved Amgen's monoclonal antibody (mAb) Xgeva (denosumab), which targets the receptor activator of NF-κB ligand (RANKL) to reduce skeletal-related events in individuals with bone metastasis from solid tumors. The agency's go-ahead, announced in November, was based on three pivotal phase 3 trials comparing the human anti-RANKL mAb with the standard of care bisphosphonate Zometa (zoledronic acid) from Novartis of Basel. Results from the three Xgeva trials—one in people with castration-resistant prostate cancer, one in breast cancer and one in individuals with solid tumors or multiple myeloma—show Xgeva's superiority in breast and prostate cancer trials and noninferiority to Zometa in solid tumors and multiple myeloma. Xgeva also reduced pain and improved quality of life compared with Zometa. "Xgeva could eat into the existing market share of the bisphosphonates," says Ranjith Gopinathan, industry analyst in life sciences at Frost & Sullivan. "Xgeva is expected to generate sales of about $2.4 billion in 2015," Gopinathan adds. The Thousand Oaks, California–based company's mAb was already approved back in June 2010 as Prolia to treat osteoporosis in postmenopausal women (*Nat. Biotechnol.* **28**, 640, 2010). The dose given to cancer patients is 12 times higher than to patients with bone loss indications, but so far the antibody has not shown the worrying side-effects associated with bisphosphonates, which include renal toxicity, atypical fractures of the thigh and osteonecrosis of the jaw. As questions mount over bisphosphonate use, clinicians may well favor treatment with the biologic. Robert Coleman, professor of medical oncology at Sheffield University, UK, believes Xgeva could potentially replace bisphosphonates as standard of care because of its efficacy, ease of administration (Xgeva is injected subcutaneously and Zometa is an intravenous infusion) and less severe side effects. "The only limitation could be the cost—many of the bisphosphonates are just about to come off patent, so doctors would need to balance cost and efficacy," says Coleman. Amgen is currently developing denosumab for rheumatoid arthritis, and rare giant cell tumors of the bone, which are very dependent on RANKL. *Suzanne Elvidge*

of the interests of the academic investigators and the industrial partners," says David Mack of the venture firm Alta Partners, in San Francisco, either because the academics were driven by other basic research questions or because of a lack of appreciation for the cost, risk and time that drug development takes. "They see that they've created an asset that is worth a lot, but actually it's not worth a lot because all of the risk is ahead of us—investment capital, development, technical risk."

But as grant funding proves ever harder to find, it's an opportune time for exploring new models. Plus, the venture capital industry is contracting significantly and is also shifting its focus, where possible, to more late-stage, downstream investments. The absence of an initial public offering market has made some of the investigators more realistic. "It's the right time for that kind of approach—getting them involved on a risk-sharing basis and setting some realistic near- to midterm milestones to achieve some value creation, even if it means then passing it on to Pfizer in exchange for a royalty," says Mack. The ability to hit the group running with a program and have immediate access to Pfizer's development resources may also be attractive to academics who are either uncomfortable or impatient with the venture capital process, where initial fund-raising could take years.

But more experienced academic entrepreneurs might not want to trade control or more potential upside in exchange for expediency. Paul Schimmel of the Scripps Research Institute in La Jolla, California, believes that "To preserve their freedom and work in an academic-like way, they'll probably want to turn to do that in the venture community and startups rather than the pharmaceutical industry, where it can get buried and disappear."

A tendency for people within companies to move is another ongoing issue. Regis Kelly, director of the California Institute for Quantitative Biosciences (QB3), a nonprofit institute spanning three University of California campuses in the San Francisco Bay Area, points to pharma's frequent management changes as a potential snag in making the partnerships thrive. For instance, in 2008, soon after Pfizer merged with Wyeth, it dissolved the Biotherapeutics and Bioinnovation Center (BBC)on UCSF's Mission Bay campus—set up in 2007 as a hybrid between academia and industry, to work on translational projects (*Nat. Biotechnol.* **27**, 308, 2009). For about a year, Kelly recalls, "there was a hiatus, where we couldn't start any new programs together."

Even as Pfizer focuses on decentralizing industry-academic partnerships, London-based GlaxoSmithKline (GSK) will soon adopt a virtual approach. GSK aims to create up to ten relationships with individual researchers throughout the world, forming a virtual project team with each of them in order to, like Pfizer, provide immediate access to GSK resources. "We're not talking about giving lots of money across to academia," says GSK's Patrick Vallance, who is leading the program. An experienced drug discoverer will work in tandem with the research group. "At the beginning it's very focused, with access to the whole of GSK's expertise," he says.

GSK is set to announce the first of its collaborations under the program, with Mark Pepys at University College, London (UCL), and Pepys' UCL spinout, Pentraxin Therapeutics, for the development of a small molecule to treat amyloidosis. GSK and Pentraxin are already working together to develop an antibody to treat the disease.

To some extent, Pfizer's CTI programs echo the spirit of Eli Lilly's Chorus initiative, started in 2007, in which a venture firm supplies the Indianapolis-based pharma with compounds for Lilly to rapidly advance through phase 1. But whereas both emphasize speed to the clinic from a similar preclinical starting point, the CTIs will also explore the biology around its targets in depth, at greater cost, but also presumably to its benefit. Indeed, although Pfizer is aware of the importance of targeted therapeutics and personalized medicine, "It's not an area we have invested a significant amount of time in," says Coyle. By focusing on translational medicine up front, "We're going to have a broader impact in the organization," he says.

Mark Ratner, *Cambridge, Massachusetts*

IN their words

"How many [new drugs] are approved each year—six, seven, eight, nine maybe? If the value of these few new drugs is worth 10, maybe 20 billion U.S. dollars, then where is the remainder of the $85 billion going?" Thomas Lonngren, outgoing chief of the European Medicines Agency, censures the industry for its profligate R&D spending. (*Wall Street Journal*, 15 December 2010)

"If I listened to you, I wouldn't be in this business. Without innovation, we are toast." Paul Hastings, CEO of Oncomed, responds to a question from the audience at 'Convergence' in San Francisco on 3 December on how his company can justify the high costs of its efforts. (*Xconomy*, 6 December 2010)

Boehringer splashes out on bispecific antibody platforms

A $2.2 billion deal in October between Macrogenics and Boehringer Ingelheim, followed the next month by another eye-catching $1.7-billion research collaboration between f-star and the German pharma are the latest in a string of blockbuster transactions in which big pharma has accessed exotic antibody platforms. The pharmaceutical industry, which has long since embraced traditional monoclonal antibody (mAb) therapies, is now looking to the next generation of therapeutics—conjugates, fragments or other derivatives (**Fig. 1**). Over the course of 2010, a raft of collaborations has put the wind into platform-companies' sails. But as the flow of investment into new technologies continues unabated, questions arise over how much value each antibody-modifying technology can add.

The promise of bispecific antibodies may have driven the latest deals (**Table 1**). Progress in bispecific antibodies thus far has been held back by their poor stability, short half-life and manufacturing uncertainties. Macrogenics CEO Scott Koenig says now it is possible to make them "in a way that satisfies commercialization objectives," and this has been important in attracting pharma. On the same day as the Rockville, Maryland–based biotech announced its collaboration with Boehringer, the biotech sealed a deal with Pfizer of New York.

In the case of Macrogenics, the biotech will receive R&D funding of $60 million for discovery work, using its dual-affinity retargeting (DART) platform to generate bispecific diabody scaffolds (pairs of Fv regions from different mAbs covalently joined by a short peptide linker) against Boehringer's targets. The bigger reward will come in milestones of up to $210 million for each of the programs, spanning a range of therapeutic areas, including the traditional mAb strongholds of cancer and immunology, but also branching out into respiratory and infectious diseases.

Similarly, f-star's deal with the Ingleheim-based pharma requires the startup to run seven targets through its modular antibody technology platform, which allows the introduction of additional antigen-binding sites into mAbs and mAb fragments by engineering the loops of Fc and Fv domains outside of complementarity-determining regions (CDRs). Non-CDR loops are randomized, libraries of antibody fragments are created by standard techniques, such as phage display, and molecules with desired binding properties are selected from these libraries. The aim is to generate fully functional fragments—coined 'Fcabs' by f-star—at one-third of an antibody's size. Again, f-star will receive an undisclosed initial technology licensing fee and be paid for doing the R&D. Beyond that, milestones of up to €180 ($241.3) million per program are in place for every product that makes it to market.

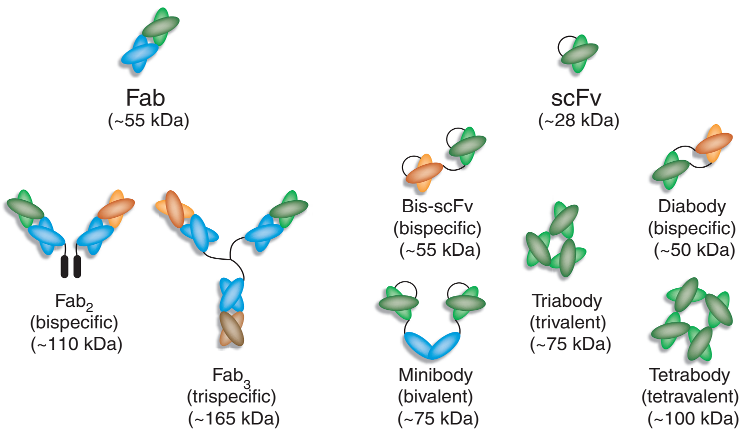


Figure 1 A cornucopia of different antibody fragments are currently in development. (Adapted from *Nat. Biotechnol.* **23**, 1126–1136, 2005.)

Although f-star CEO Kevin Fitzgerald acknowledges that attrition rates will likely limit the company's program from reaching every one of its milestones, the Boehringer deal provides validation for the Fcab platform. According to Fitzgerald, Fcabs' advantage over fragments fashioned from antibody variable regions is their long plasma half-life and that they can be combined into bispecific formats, while retaining all their effector sites. Antibodies with dual binding sites can be engineered to boost their therapeutic effect by engaging two targets at once (e.g., bringing T cells into close association with tumor cells or targeting two antigens on a single cell).

Currently, Munich-based Trion Pharma's Removab (catumaxomab) is the only bispecific antibody on the market. It was approved as a therapy for malignant ascites by the European Medicines Agency in 2009. Macrogenics' Koenig notes that 11 of the 15 largest pharma companies either have their own bispecific platforms or have licensing agreements around this format. This growing momentum will finally unlock the attraction of bispecific formats, he believes.

Further proof of the mounting interest in bispecific antibodies comes from another specialist, Micromet. The Bethesda, Maryland company closed a $70.5 million fundraising series in November, to further develop its bispecific T-cell engagement (BiTE) platform. BiTE antibodies, which comprise the minimal antigen-binding domains of two single-chain Fvs (scFvs) arranged in tandem on a polypeptide chain,

Table 1 Selected companies producing bispecific antibodies

Company	Technology platform	Mechanism
Macrogenics	DART (dual-affinity re-targeting) platform	Dual specificity 'antibody-like' therapeutic proteins capable of targeting many different epitopes with a single recombinant molecule
f-star	Modular antibody technology	Allows small antibody fragments with full antibody functionality (Fcab) or full-length antibodies with additional functionality (mAb) to be created
Micromet	BITE (bispecific T-cell engager) technology	Lead product Blinatumomab (MT-103) is a BiTE antibody designed to direct T cells against CD19 on B cell–derived acute lymphoblastic leukemias and non-Hodgkin's lymphomas
Bicycle Therapeutics	Bicycle technology	Mini-antibodies with two binding loops covalently attached to organochemical cores
Domantis/GSK	Dual targeting domain antibodies (dAbs)	The fully human dual targeting dAbs bind two targets simultaneously and can be manufactured in dimer, Fab-like or IgG-like formats. Domantis is developing Dual Targeting dAbs against cytokines for inflammatory diseases, tumor antigens on tumor cells and angiogenic factors for solid tumors

Source: Company websites

Box 1 Tweaking mAb lifetime and functionality

Pharma is eyeing platform engineering tools as a means to extract further returns from its investments in classic mAbs, believes Daniel Junius, CEO of ImmunoGen. This is demonstrated in the Waltham, Massachusetts–based company's agreement last October to apply its tumor-activated prodrug (TAP) technology to antibodies nominated by partner Novartis of Basel. The TAP platform enables the attachment of chemotherapeutics to existing antibodies through a disulfide bond created by introducing a dithiopyridyl group into the mAb of interest through treatment with either *N*-succinimidyl-4-[2-pyridyldithio]-pentanoate or *N*-succinimidyl-3-[2-pyridyldithiol]-propionate. Immunogen received $45 million upfront and is entitled to $200 million in milestones on any product that makes it to market. "There has been an increase in interest in the whole antibody-drug conjugate space" based on a small amount of "very compelling" clinical data, Junius says.

Pharma is also keen to boost mAbs' patent life. Adrian Tombling, patent attorney at Withers & Rogers in London, says all the broad patents on antibody production technologies that engendered so much royalty stacking and litigation in the 1990s have "pretty much all gone." At the same time, patents claiming to have discovered a particular target associated with a particular disease phenotype would not be granted in Europe.

Now, says Tombling, it is necessary to make a series of antibodies that differ by one or two amino acids to get broad protection for a single molecule. "You have to do more work before filing—there's a clear parallel with patenting classic chemical structures where you would generate a series of compounds," he says. *NM*

bind to T cells, directing them to antigens on the surface of tumor cells.

Pharma has other reasons to be interested in antibody platforms (**Box 1**). Lundbeck of Denmark, for instance, was looking to build on its central nervous system (CNS) franchise and expertise when it entered a €38 ($51) million deal with GenMab in October. "Lundbeck was attracted by increasing evidence that antibodies could be effective in CNS diseases," says GenMab CEO Jan van de Winkel. The biotech, headquartered in Copenhagen, will develop antibody formats known as Unibodies against three unnamed nervous system targets. Unibodies lack immune effector functions, such as complement activation, a trait which would prove advantageous for CNS targets.

Assimilation of independent, public mAb pioneer companies, including Idec, Medarex, Cambridge Antibody Technology, Celltech, Abgenix and Genentech, into the pharma fold has reduced overall access to established antibody discovery platforms. This has left a handy gap in the market for one of the most intriguing deals in the history of biotech, in which San Francisco–based startup Ablexis pulled together a consortium of five pharma companies to jointly support the development of its AlivaMab transgenic mouse platform for generating human antibodies. Of the five, only Pfizer was named, and Ablexis has also been cagey about the terms, saying each would pay a nonrefundable "seven figure" fee to join the consortium and make an "eight figure" payment upon delivery of the AlivaMab mouse strains. These AlivaMab strains express chimeric single-chain antibodies that contain a variable region of a human immunoglobulin light chain and a nonhuman heavy chain constant region (but are devoid of the first constant domain CH1).

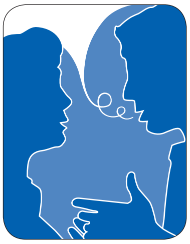
Another company looking to fill the gap for *in vivo* platforms is Crescendo Biologics, a Cambridge, UK–based company, developing a transgenic mouse platform for generating fully human fragments—in this case, VH (subset H) chains. Crescendo will combine this with a proprietary *in vitro* ribosome display system for fast optimization. Although the company claims this is the first transgenic platform for generating VH chains, these entities do occur naturally in sharks and camelids. Haptogen, a Scottish biotech located in Aberdeen developing shark-derived antibodies by means of phage display against haptens, was acquired by Wyeth (now Pfizer) in October 2007. But Ablynx, of Ghent, Belgium, which generates VH (subset H) chain antibodies—or nanobodies as it calls them—in llamas, has several partners, most recently extending an existing deal with Merck Serono of Geneva.

The move to structure antibody deals around targets illustrates the fact that pharma is increasingly choosing to rent, rather than buy, platforms. But for Melanie Lee, former R&D director of Celltech and UCB Pharma, targets have become an issue. "The platforms are productive and everyone has antibodies coming out of their ears; the issue increasingly is 'what are you going to target'?" Lee, who is now working in a different field, as CEO of Syntaxin, a specialist in botulinum toxins, feels emerging antibody platforms are merely producing variants on an existing theme. "I'm not optimistic you can take advantage of each new platform, unless they deliver high plasma concentration or a long half-life," Lee says.

Such concerns have done nothing to stem the flow of new platforms. In October 2009, Greg Winter, an antibody engineering pioneer and founding scientist of Cambridge Antibody Technology (subsequently acquired by MedImmune) and Domantis (now owned by GlaxoSmithKline of London) launched another antibody company, Bicycle Therapeutics, out of the Laboratory of Molecular Biology (LMB) in Cambridge. And the latest and startling discovery made at LMB, showing traditional wisdom is wrong and antibodies can in fact operate within cells (*Proc. Natl. Acad. Sci.* doi:10.1073/pnas.1014074107, 2010), seems certain to generate yet more interest.

Nuala Moran, *London*

IN their words



"Unfortunately, the campaign fell into sound bites and most people voted for it with the expectation that there were going to be stem cell cures in a year, that Superman would walk again." John Simpson of the Santa Monica–based Consumer Watchdog blasts the California Institute for Regenerative Medicine, which is seeking another $3 billion in bonds amid controversy over salaries of top administrators and the lack of medical breakthroughs. (*Los Angeles Times*, 22 November 2010)

"The NIH should focus on the barrier it controls, the NIH-funded patents." Jamie Love of advocacy group Knowledge Ecology International, after the agency denied three patients affected by Fabryzyme shortages the right to override Genzyme's patents. (*Pharmalot*, 7 December 2010)

"Why are you doing this [registry]?" Debra Leonard, director of clinical laboratories and pathology residency training at Weill Cornell Medical College, questions the need for the NIH's impending genetic-testing registry given the existence of two other databases. (*GenomeWeb*, 6 December 2010)

"We're entering a new world. Pharma companies are going to spend a lot of time developing first-in-class drugs." Frederick Frank, a biotech banker with Barclays Capital, on the shifting business focus of big pharma. (*Reuters*, 8 November 2010)

Biogen Idec restructures, sharpens neurology focus

It seems odd for a company in good financial health to undertake a dramatic restructuring, but Biogen Idec of Weston, Massachusetts, has just done exactly that. The surprise announcement in November encompasses staff cuts of 650 people—about 13% of the total headcount—and the termination or disposal of 11 development programs, including all of the company's oncology and cardiology pipelines.

The decision was taken by Biogen Idec's new chief executive George Scangos, only four months after his arrival from S. San Francisco, California–based Exelixis. The news emerged with very little warning and with Biogen's financial position looking strong. Annual revenues are about $4.5 billion, of which $1.2–1.3 billion is "free cash flow," according to Scangos. The cash comes largely from its three top products—especially Avonex (interferon (IFN) β-1a) an intramuscular injection for multiple sclerosis (MS), Tysabri (natalizumab, a humanized monoclonal antibody (mAb) against α-4/β-1 integrin), also for MS, and blockbuster antibody Rituxan (rituximab, a chimeric mAb against CD20) for non-Hodgkin's lymphoma, chronic lymphocytic leukemia and rheumatoid arthritis, held jointly with Genentech of S. San Francisco.

Scangos's single most surprising move is to exit the development of lixivaptan, Biogen's most advanced pipeline product. Lixivaptan is a selective vasopressin V2 receptor antagonist licensed from Cardiokine, of Philadelphia, in 2007, and already in phase 3 testing for hyponatremia, a metabolic condition in which body fluids are deficient in sodium. Biogen has had to pay Cardiokine $25 million compensation to end the collaboration.

But bolder and more far-reaching is Scangos's decision to get out of oncology development. He hopes to outlicense all the company's development compounds in that field. This includes GA-101 for non-Hodgkin's lymphoma; volociximab, an α-5/β-1 integrin inhibitor for solid tumors; and BIIB021, an Hsp90 inhibitor also for solid tumors. The licensing of these compounds is unlikely to garner large revenues in the near term, says Scangos. But he believes several have commercial merit and the company intends to keep some interest in them.

Biogen has also renegotiated its deal with Genentech over the co-marketing of Rituxan in the US. Genentech will take over all US sales and marketing of the product, allowing Biogen to scrap its Rituxan US sales force. All this has led to a significant headcount reduction, with 650 job losses out of a workforce of nearly 5,000.

Bloomberg via Getty Images

George Scangos, Biogen Idec's new chief executive.

Scangos refuses to break down the job losses, but admits half of them are due to discontinued programs. The restructuring—which also includes the closure of the company's site in San Diego—will save some $300 million a year.

So is Scangos a hatchet man brought in to trim the corporate fat so as to placate Biogen's notoriously aggressive private shareholders, led by Carl Icahn? Rituxan aside, he is keeping little of the legacy inherited from the Idec merger in 2003, and these moves do resemble proposals put forward by these shareholders to split the company into its component parts on the basis that the Idec merger was flawed from the start. Scangos denies it: "None of this was done out of financial desperation." He says the restructuring is a forward-looking move, with a vision to refocus the company on its neurology and hematology pipelines, in particular successors to Avonex and Tysabri for MS. In fact, on December 20, Biogen announced it would be acquiring a subsidiary of Neurimmune that is working on three preclinical antibody candidates targeting synuclein, tau and TDP-43 for neurodegenerative diseases.

"Oncology is maybe the most crowded space in all of drug development, with thousands of compounds in development," says Scangos, who as Exelixis CEO worked in the sector for more than a decade. "About 80% of the biotech industry, and all the pharmaceutical majors, must be working on it." To compete in that

IN brief

Roche cuts Genentech jobs

In November, Roche revealed sweeping cuts in its workforce as part of its so-called Operational Excellence Program to reduce costs. Over 4,800 people, 6% of its workforce, will lose their jobs and another 700 will be affected by outsourcing, all of which will save the Basel-based company an estimated CHF2.4 ($2.43) billion annually. But the news isn't all bad. Only 600 of the job cuts are in R&D, with most jobs losses in sales and manufacturing, and none at the US flagship Genentech, according to Robin Snyder spokesperson for the S. San Francisco–based biotech. The cuts in R&D include shuttering RNA interference programs in Kulmbach, Germany; Nutley, New Jersey; and Madison, Wisconsin. And there may be a silver lining to these moves, as projects cast off by Roche may create opportunities for startups that could revitalize the Swiss biotech industry. Indeed Basel-based biotechs Actelion and Basilea were set up to develop programs axed by Roche. In California, somewhat ironically, the reshuffling, which will cause the loss of over 800 local manufacturing jobs, was announced two weeks after the election in which a state-wide proposition rescinding corporate tax credits was defeated. Opponents of the measure, which included Genentech, the biggest contributor to the 'no' campaign, with over $1.6 million in donations, argued that rescinding the tax credits would result in job losses. Snyder claims that the timing of the announcement of cuts in the California workforce was unrelated to the election. *Laura DeFrancesco*

Company Bridge Awards

The National Cancer Institute has awarded $9.9 million to four companies developing diagnostics for cancer therapies. The new Small Business Innovation Research (SBIR) Phase II Bridge Awards are designed to support previously funded companies to develop and commercialize their products further (*Nat. Biotechnol* **27**, 678, 2009). Phase II recipients are Advanced Cell Diagnostics, 20/20 GeneSystems, AmberGen and Praevium Research; each company receives up to $3 million over three years. Advanced Cell Diagnostics of Hayward, California, is developing a CTCscope system to assess molecular profiles in circulating tumor cells (CTCs) that have the potential to metastasize in other parts of the body. 20/20 GeneSystems of Rockville, Maryland, is focusing on 'PredicTOR' as a companion diagnostic for therapies targeting the mTOR pathway. AmberGen, of Watertown, Massachusetts, will advance a gene expression–based test to predict colorectal cancer recurrence, and its response to treatment. Praevium Research of Santa Barbara, California, received funding for a miniaturized optoelectronic device, which could allow clinicians to image cancer tissue in real time, without the need for tissue excision and biopsies. *Nidhi Subbaraman*

space, he says, a company needs a very special competitive advantage.

"When I came to Biogen and looked at what we had, it was clear that if we were going to compete successfully in oncology we were going to have to invest substantially more," he says. Despite its cash resources, the company could not afford to pursue oncology programs at the same time as making the necessary investments in neurology—a field where Scangos believes Biogen is good enough to compete at the top level in every aspect, from early research to marketing. "I can't say the same in oncology, and something had to go, so the choice became rather clear."

Certainly, with several products in the MS market, consolidating in neurology makes sense for Biogen. The US medical profession still seems to regard Avonex as a preferred treatment for MS. Avonex revenues continue to increase at ~11% a year despite competition from Jerusalem-based Teva's Copaxone (glatiramer acetate), Rebif (IFN β-1a) from Merck, of Darmstadt, Germany, and Pfizer of New York, and Betaseron (IFN β-1b, which differs from human protein by a C17→S mutation) owned by Bayer, of Leverkusen, Germany.

Tysabri is also popular with clinicians but has problems of its own. A substantial number of patients using it have developed a potentially fatal brain infection called PML (progressive multifocal leukoencephalopathy). According to analyst Ian Somayia at Piper Jaffray in New York, only one in a thousand patients developed PML during phase 3 trials of the drug, but the rate has since shown signs of increasing with time. A serum assay is being developed to identify patients who are carrying the virus that causes PML, to mitigate the risk of treatment. But the test is not yet approved and not even known to be effective. So Biogen is making the most of its success while the going is good, and increased the US price by nearly 19% last June.

But a look at the emerging competition shows a serious threat to Biogen's core MS franchise. Novartis is just unleashing onto the US market its MS treatment Gilenya (fingolimod, a sphingosine-1-phosphate agonist), the first oral therapy to receive US Food and Drug Administration approval for the disease. Christopher Raymond at equity research firm Robert W. Baird says the Gilenya approval could have an impact on Avonex as well as Tysabri, producing a "more difficult road" for Biogen.

Other rival therapies are also in development. US patent protection for Avonex expires in 2013. There then arises the prospect of follow-on biologic competition if in the meantime the US authorities have developed an approval methodology for follow-ons. EU biosimilars competition is likely in the 2012 time frame, according to Shenouda at Stifel Nicolaus. Patents on other cash-generating products, in particular Rituxan, will expire between 2013 and 2018; indeed, revenues from Rituxan are already falling anyway.

One of the factors needing Scangos's urgent attention was the balancing of Biogen's pipeline. There are milestones coming up between 2011 and 2014 for three critical products in phase 3 testing for relapsing-remitting MS. BG-12 is an oral dimethyl fumarate that activates the NF-E2-related factor 2 pathway in MS. Zenapax (daclizumab) is a humanized antibody against interleukin-2, also for treating MS. The third candidate is PEGylated IFN β-1a. A fourth, prolonged-release Fampridine (4-aminopyridine, a potassium channel blocker) is expected to improve walking ability in adult patients with MS. It is already approved in the US and has been submitted for licensing to the European Medicines Agency in London. "We have a very rich late-stage pipeline of seven compounds either in phase 3 or about to go into it," says Scangos. "We also have a lot of interesting compounds at the early stage. What we don't have are many compounds around phase 2."

Thus, though the company probably has at least three or four new products coming to market in the next few years, it could be looking at a difficult interlude after that as the pipeline falters. This is why Biogen has to refocus on boosting its portfolio of development-stage agents, says Shenouda. Matt Roden of UBS agrees: "Acquisition of phase 2 pipeline assets and commercial growth drivers could be a positive."

There are also political changes coming to US healthcare, which will likely mean that drug price increases are no longer going to be an easy remedy for falling sales. "The next decade is going to be very competitive on health costs," Scangos predicts. "We have to be ready for that."

Two unknowns remain in the equation: Biogen's shareholders—and possible acquirers. "There are takeover rumors constantly surrounding Biogen," says Miami Beach–based Standpoint Research's Ronnie Moas. Biogen actually tried to find a buyer in 2007; several big pharma companies made enquiries but none took the bait. They may have been deterred from bidding because of considerable uncertainty about Tysabri at the time (it had only recently been allowed back on the US market after a temporary safety withdrawal).

Biogen is still a plausible takeout target, says UBS analyst Matt Roden. That may be one reason why restructuring has been undertaken so urgently; a company already in the throes of rapid rebuilding is less of a magnet for speculative offers than one whose value is perceived solely in the sales of its existing products.

Peter Mitchell, *London*

IN brief

Roche cuts Genentech jobs

In November, Roche revealed sweeping cuts in its workforce as part of its so-called Operational Excellence Program to reduce costs. Over 4,800 people, 6% of its workforce, will lose their jobs and another 700 will be affected by outsourcing, all of which will save the Basel-based company an estimated CHF2.4 ($2.43) billion annually. But the news isn't all bad. Only 600 of the job cuts are in R&D, with most jobs losses in sales and manufacturing, and none at the US flagship Genentech, according to Robin Snyder spokesperson for the S. San Francisco–based biotech. The cuts in R&D include shuttering RNA interference programs in Kulmbach, Germany; Nutley, New Jersey; and Madison, Wisconsin. And there may be a silver lining to these moves, as projects cast off by Roche may create opportunities for startups that could revitalize the Swiss biotech industry. Indeed Basel-based biotechs Actelion and Basilea were set up to develop programs axed by Roche. In California, somewhat ironically, the reshuffling, which will cause the loss of over 800 local manufacturing jobs, was announced two weeks after the election in which a state-wide proposition rescinding corporate tax credits was defeated. Opponents of the measure, which included Genentech, the biggest contributor to the 'no' campaign, with over $1.6 million in donations, argued that rescinding the tax credits would result in job losses. Snyder claims that the timing of the announcement of cuts in the California workforce was unrelated to the election. *Laura DeFrancesco*

Company Bridge Awards

The National Cancer Institute has awarded $9.9 million to four companies developing diagnostics for cancer therapies. The new Small Business Innovation Research (SBIR) Phase II Bridge Awards are designed to support previously funded companies to develop and commercialize their products further (*Nat. Biotechnol* **27**, 678, 2009). Phase II recipients are Advanced Cell Diagnostics, 20/20 GeneSystems, AmberGen and Praevium Research; each company receives up to $3 million over three years. Advanced Cell Diagnostics of Hayward, California, is developing a CTCscope system to assess molecular profiles in circulating tumor cells (CTCs) that have the potential to metastasize in other parts of the body. 20/20 GeneSystems of Rockville, Maryland, is focusing on 'PredicTOR' as a companion diagnostic for therapies targeting the mTOR pathway. AmberGen, of Watertown, Massachusetts, will advance a gene expression–based test to predict colorectal cancer recurrence, and its response to treatment. Praevium Research of Santa Barbara, California, received funding for a miniaturized optoelectronic device, which could allow clinicians to image cancer tissue in real time, without the need for tissue excision and biopsies. *Nidhi Subbaraman*

space, he says, a company needs a very special competitive advantage.

"When I came to Biogen and looked at what we had, it was clear that if we were going to compete successfully in oncology we were going to have to invest substantially more," he says. Despite its cash resources, the company could not afford to pursue oncology programs at the same time as making the necessary investments in neurology—a field where Scangos believes Biogen is good enough to compete at the top level in every aspect, from early research to marketing. "I can't say the same in oncology, and something had to go, so the choice became rather clear."

Certainly, with several products in the MS market, consolidating in neurology makes sense for Biogen. The US medical profession still seems to regard Avonex as a preferred treatment for MS. Avonex revenues continue to increase at ~11% a year despite competition from Jerusalem-based Teva's Copaxone (glatiramer acetate), Rebif (IFN β-1a) from Merck, of Darmstadt, Germany, and Pfizer of New York, and Betaseron (IFN β-1b, which differs from human protein by a C17→S mutation) owned by Bayer, of Leverkusen, Germany.

Tysabri is also popular with clinicians but has problems of its own. A substantial number of patients using it have developed a potentially fatal brain infection called PML (progressive multifocal leukoencephalopathy). According to analyst Ian Somayia at Piper Jaffray in New York, only one in a thousand patients developed PML during phase 3 trials of the drug, but the rate has since shown signs of increasing with time. A serum assay is being developed to identify patients who are carrying the virus that causes PML, to mitigate the risk of treatment. But the test is not yet approved and not even known to be effective. So Biogen is making the most of its success while the going is good, and increased the US price by nearly 19% last June.

But a look at the emerging competition shows a serious threat to Biogen's core MS franchise. Novartis is just unleashing onto the US market its MS treatment Gilenya (fingolimod, a sphingosine-1-phosphate agonist), the first oral therapy to receive US Food and Drug Administration approval for the disease. Christopher Raymond at equity research firm Robert W. Baird says the Gilenya approval could have an impact on Avonex as well as Tysabri, producing a "more difficult road" for Biogen.

Other rival therapies are also in development. US patent protection for Avonex expires in 2013. There then arises the prospect of follow-on biologic competition if in the meantime the US authorities have developed an approval methodology for follow-ons. EU biosimilars competition is likely in the 2012 time frame, according to Shenouda at Stifel Nicolaus. Patents on other cash-generating products, in particular Rituxan, will expire between 2013 and 2018; indeed, revenues from Rituxan are already falling anyway.

One of the factors needing Scangos's urgent attention was the balancing of Biogen's pipeline. There are milestones coming up between 2011 and 2014 for three critical products in phase 3 testing for relapsing-remitting MS. BG-12 is an oral dimethyl fumarate that activates the NF-E2-related factor 2 pathway in MS. Zenapax (daclizumab) is a humanized antibody against interleukin-2, also for treating MS. The third candidate is PEGylated IFN β-1a. A fourth, prolonged-release Fampridine (4-aminopyridine, a potassium channel blocker) is expected to improve walking ability in adult patients with MS. It is already approved in the US and has been submitted for licensing to the European Medicines Agency in London. "We have a very rich late-stage pipeline of seven compounds either in phase 3 or about to go into it," says Scangos. "We also have a lot of interesting compounds at the early stage. What we don't have are many compounds around phase 2."

Thus, though the company probably has at least three or four new products coming to market in the next few years, it could be looking at a difficult interlude after that as the pipeline falters. This is why Biogen has to refocus on boosting its portfolio of development-stage agents, says Shenouda. Matt Roden of UBS agrees: "Acquisition of phase 2 pipeline assets and commercial growth drivers could be a positive."

There are also political changes coming to US healthcare, which will likely mean that drug price increases are no longer going to be an easy remedy for falling sales. "The next decade is going to be very competitive on health costs," Scangos predicts. "We have to be ready for that."

Two unknowns remain in the equation: Biogen's shareholders—and possible acquirers. "There are takeover rumors constantly surrounding Biogen," says Miami Beach–based Standpoint Research's Ronnie Moas. Biogen actually tried to find a buyer in 2007; several big pharma companies made enquiries but none took the bait. They may have been deterred from bidding because of considerable uncertainty about Tysabri at the time (it had only recently been allowed back on the US market after a temporary safety withdrawal).

Biogen is still a plausible takeout target, says UBS analyst Matt Roden. That may be one reason why restructuring has been undertaken so urgently; a company already in the throes of rapid rebuilding is less of a magnet for speculative offers than one whose value is perceived solely in the sales of its existing products.

Peter Mitchell, *London*

Science snipes at Oxitec transgenic-mosquito trial

Oxitec released 3.3 milion sterile male transgenic *Aedes aegypti* mosquitoes in a field trial aimed at reducing wild mosquito populations to control dengue.

Early in November, at the annual meeting of the American Society of Tropical Medicine and Hygiene (ASTMH) in Atlanta, researchers from the British company Oxitec disclosed results from the world's first genetically modified (GM) mosquito field trials aimed at controlling the carrier for dengue fever. After the presentation at the meeting, *Science* (**330**, 1030–1031, 2010) published a news story claiming the trials had "strained ties" with Oxitec's collaborator, the Bill and Melinda Gates Foundation. Anthony James, the lead investigator on the Gates team, was also quoted as saying he would "never release GM mosquitoes the way Oxitec has now done in Grand Cayman." Although some concerns have been raised as to how information about the trial was disseminated, it seems that controversy over the environmental release of a GM organism has been overblown.

Oxitec's plans for transgenic mosquito trials have not been without controversy in the past. They have been criticized by environmental groups, such as Ottawa-based ETC Group and EcoNexus of Oxford, concerned about the risks of releasing an entirely new strain of organism into the environment. Activists warn that transgenic insect releases that reduce wild mosquito numbers might not only create an 'empty niche', which other potentially damaging insects might fill, but also affect organisms higher in the food chain that rely on mosquitoes as a dietary source.

The present spat, however, centers around disagreements over the rapid move to an open release of insects and in particular the way in which the existence of the trial was communicated to the community and public at large. Luke Alphey, CSO of Oxitec, concedes that researchers may have differing views on how to plan and execute such field tests; however, he says he hasn't received any complaints from the community nor has he been scolded by his Gates collaborator James, a professor at the University of California, Irvine. When contacted by *Nature Biotechnology*, James declined to comment, but a spokesperson for the Gates Foundation says of a different trial Oxitec is running in Mexico in collaboration with the Foundation that "we are happy with the way that is going."

For his part, Alphey says he was "surprised that *Science* chose to present the story the way they did." If there is a controversy around the way Oxitec prepared for the trials, he says, it has not officially been directed at his company.

Oxitec first commenced the Cayman trials in September 2009. Together with the islands' Mosquito Research and Control Unit (MRCU), the company liberated about 3.3 million sterile male transgenic *Aedes aegypti* mosquitoes into a region spanning about 16 hectares through 80 releases.

The OX513A mosquitoes used in the trial carry the LA513 transposon integrated into their genetic material via a *piggyBac* helper

Table 1 Progress in GM mosquito research

Species name/vector disease	Transposable element	Year transformed
Aedes aegypti/yellow fever	Mariner	1998
Aedes aegypti/yellow fever	Hermes	1998
Anopheles stephensi/Indo-Pakistani malaria	Minos	2000
Anopheles gambiae/African malaria	piggyBac	2001
Aedes aegypti/yellow fever	piggyBac	2001
Culex quinquefasciatus (Southern house mosquito)	Hermes	2001
Anopheles stephensi/Indo-Pakistani malaria	piggyBac	2002
Anopheles albimanus/New World malaria	piggyBac	2002
Aedes fluviatilis/Brazilian malaria	piggyBac	2006
Aedes albopictus (Asian tiger mosquito)	piggyBac	2010

Source: Morrison, N.I. *et al. Asia-Pac. J. Mol. Biol. Biotechnol.* (in the press).

IN brief

Temporary ban on clones

Food from cloned animals is under fire in Europe with the European Commission (EC) calling in October for a temporary commercial suspension. John Dalli, European commissioner for health and consumer policy, describes the proposal as "a realistic and feasible solution to respond to the present welfare concerns." A formal proposal for a five-year ban on the technology will be presented in the first half of 2011. Although sweeping, the proposed exclusion may not carry much weight in practice, because farmers mostly use cloning technology for their prized breeding stock, not to raise animals for food. EU breeders would be forbidden under the proposed ban to clone their best head of cattle in member states. Cloned embryos and semen of clones, however, could still be imported following a proposed traceability scheme. The offspring of clones, sired conventionally, would not be bound by these restrictions, EC spokesperson Frédéric Vincent points out, and consequently their meat and milk would not be banned. This decision avoids unleashing trade wars with the US but is likely to be opposed by the European Parliament. A public outcry followed a document release in August by the British Food Standards Agency that three bulls descending from embryos cloned in the US from an undisclosed company entered the food chain in the UK and Belgium. According to Vincent these occurrences are legal under current regulations, but probably uncommon. *Anna Meldolesi*

Filipinos back GM eggplant

Filipino farmers clamoring for the adoption of genetically modified (GM) eggplants in October passed a resolution to support multi-location field trials of the biotech crop. GM crop farmers and agriculture representatives from across the country endorsed a set of resolutions to support the advancement of biotech crops in the country including the pest-resistant eggplant. "When we consulted them, [farmers] asked, 'Are the seeds available already? Why is it taking so long?'" says Reynaldo Cabanao, president of the Asian Farmers Regional Network (ASFARNET). The GM eggplant was developed by the Agricultural Biotechnology Support Project II (ABSPII), a global public-private collaboration based at Cornell University in Ithaca, New York. It was engineered with the *Cry1Ac* gene from the bacterium *Bacillus thuringiensis* (*Bt*) to fend off the fruit and shoot borer, which can destroy up to 50% of the region's number-one food crop. Farmers who have witnessed the success of *Bt* corn are eager for *Bt* eggplant to be available, says Desiree Hautea, ABSPII coordinator for South East Asia, at the University of the Philippines, Los Baños. The GM eggplant is currently undergoing confined field tests adhering to biosafety regulations set by the Philippines Department of Agriculture, Bureau of Plant Industry. Multiple-site trials will follow, though commercialization plans remain undefined. *Nidhi Subbaraman*

IN brief

European R&D buoyant

The economic downturn had less effect than expected on biopharma companies in Europe, according to the newly released EU Industrial R&D investment Scorecard published by the European Commission. The report, which included data on industrial research spending for fiscal year 2009, ranked 400 EU-based companies and 1,000 firms based elsewhere. Many cash-strapped firms scaled back research in 2009, with R&D investments across all sectors worldwide down 1.9%. The biopharma group, however, consolidated its position as top R&D investor, with a 5.3% increase in 2008 in global R&D spending and a 2% hike in Europe alone. Swiss drug giant Roche of Basel ranked second among all sectors for R&D investment. John Shortmoor, pharma analyst at Datamonitor, is not surprised at the findings. "To sustain a presence—especially in a time when a significant number of marketed products are losing patent protection—requires constant product innovation and continued R&D investment." For biotech in particular, European firms increased their investment budget by 7.9% in 2009, outperforming their US counterparts, whose R&D spending dropped by 1.6%. "We cannot afford not to invest in R&D and risk losing our market position," says Nickie Inger Spile, vice president at Danish biotech Novozymes. The Bagsvaerd-based firm ranked number 10 for R&D spending among global biotechs. *Emma Dorey*

EU mAb biosimilars path

European regulators laid out the rules for copying biotech's blockbuster monoclonal antibody (mAb) therapies, paving the way for biosimilars developers to access the $36.4 billion market. The draft guidelines, published by the European Medicines Agency (EMA) in November, outline the process biosimilars developers must follow to gain approval for a mAb once a patent on the pioneer drug has expired. The studies and tests needed for approval are "less demanding than expected," comments Huub Schellekens at the departments of Pharmaceutical Sciences and Innovation Studies at Utrecht University, The Netherlands. The EMA will require *in vitro* pharmacokinetic and phamacodynamic studies to demonstrate that a biosimilar mAb is functionally equivalent to a reference mAb. In some cases, *in vivo* nonclinical studies may also be necessary. "The need for these studies should be decided on a case-by case basis," the guideline states. Factors that may warrant the need for such studies are, for instance, processing and formulation differences or insufficient evidence that a biosimilar is as safe and effective as the branded product. The EMA is willing to accept a drug's adverse event profile as proof of biosimilarity, and data from one clinical trial could be sufficient for approval in two different indications if the mechanism of action is the same. "This really opens the door [to biosimilars]," Schellekens points out. The guideline is available for public comment until May 31. *Gunjan Sinha*

plasmid (*BMC Biol.* **5**, 1–11, 2007). LA513 encodes the tetracycline-repressible transcription activator (tTA), a protein whose high-level expression is deleterious to cellular development, probably due to transcriptional 'squelching' and/or interference with ubiquitin-dependent proteolysis (*Nat. Biotechnol.* **23**, 453–456, 2005). When expressed, the tTA protein binds to the tetO operator sequence (upstream of tTA) and drives expression of tTA from a nearby minimal promoter, which in turn binds to tetO, creating a positive feedback system. Because tetracycline binds tTA, preventing the activator from interacting with tetO, batches of transgenic mosquitoes can be grown in the presence of the antibiotic (whereas in its absence, transgenic mosquito larvae die). The resultant transgenic *Aedes* eggs are collected for hatching at a trial site, and the smaller male pupae sorted from females and on maturity released into the field, where breeding with wild-type female mosquitoes results in sterile mating.

Field tests in Grand Cayman were conducted in two stages. The first set of small-scale releases assessed whether transgenic males could survive in the wild and mate with wild females. The presence of transgenic larvae showed that the transgenic males did survive and were capable of finding mates. These results formed the basis for a second trial, which began last year, to test the effect of the transgenic mosquitoes on suppressing the wild population. Adult mosquitoes as well as eggs were monitored using adult traps and ovitraps (black jars containing water and a paddle leading inside, on which mosquitoes lay eggs), respectively. Offspring from transgenic males also carried a fluorescent marker, allowing the transgenic larvae to be easily distinguished from wild counterparts.

According to Alphey's ASTMH presentation, results from the large release showed up to an 80% reduction in the numbers of wild mosquitoes ~11 weeks after the release. This reduction in the population was sustained for a further ~7 weeks until the end of the trial. It is possible that the approach could be even more effective in suppressing wild mosquitoes because in this case the study site was not isolated and surrounding areas contained high densities of wild mosquitoes.

William Black, a collaborator on the Gates project, was impressed by the results; the Cayman Islands trial "went very, very well," he says. David M. Brown, project manager at the department of microbiology and molecular genetics at the University of California, Irvine, agrees that the results enjoyed a very positive reception at the meeting. "There were [even] a few comments of gratitude," he says, as the Cayman Islands trial is an important step in pushing GM insect technology against dengue fever forward.

Alphey says preparatory work for the Grand Cayman trial was extensive and meticulous. Elected political representatives were informed and flyers were distributed. MRCU officials were educated and went on foot to answer questions the locals had about the trials. All vehicles and equipment carried phone numbers and clear labels, so any concerned observers could contact the authorities. There was good awareness, he says, "that the project was testing a new genetic method to control dengue using sterile males, that males don't bite, that not all species of mosquitoes would be controlled."

Even so, some commentators have questioned whether publicity about the trial could have been better handled. For example, many only became aware of the trial's existence after the Cayman Islands government posted a YouTube video announcing the trial (http://www.youtube.com/watch?v=tv6JsC2MQYI)—hardly the traditional forum for publicizing an environmental release of a transgenic organism.

Bart Knols, managing director at K&S Consulting in The Netherlands, says that because the material is now public but has not yet passed through peer review, the trial sponsors have potentially opened themselves up for criticism. According to Knols, public information connected with transgenic insect release trials, at a global and local level, needs to be managed carefully—if not for Oxitec's sake, he says, then for others, because if bad press did occur, it "may not affect Oxitec itself, it may affect other groups around the world who are working on [GM] insects. And then no one can take advantage of all these new tools that have been developed."

David Andow, McKnight University professor of insect ecology, at the University of Minnesota in St. Paul, also feels that Oxitec could have done a better job making the research community aware of its work. It is not clear whether the Cayman Islands evaluated the trials according to international standards such as the guidelines laid out in the Cartagena protocol, he says. "Communication would have gone a long way in making it clear to people like me whether or not [Cayman Islands officials] did that," he says.

Oxitec is continuing talks with the Malaysian government, which is considering releasing transgenic mosquitoes to address its local dengue problem. By comparison, Oxitec has "been good about publicizing the work they're doing in Malaysia," says Andow. "They essentially leapfrogged that [step] in the Cayman Islands." Now, Knols says, the Malaysian government may insist that Oxitec finish its trials in the Cayman Islands before beginning in Malaysia.

What seems to be clear is that the transgenic mosquito release in the Cayman Islands was viewed as a success. Indeed, William Petrie, director of the MRCU in the Cayman Islands, says the sterile transgenic mosquito release technique is head and shoulders above the population control methods currently in place there. Val Giddings, president of the Silver Spring, Maryland consultancy PrometheusAB and a former vice president at the Biotechnology Industry Organization (BIO), says that Oxitec's strategy, as a first attempt at using transgenic insects, is beyond reproach. Not only did the company pick a relatively isolated trial site and carry out the trials in collaboration with the government following the necessary protocol, it also used a species-specific technique in which the transgene would be extinguished in following generations.

Alphey says Oxitec is moving ahead with other projects using technological lessons learned from the study itself. Meanwhile, company researchers are preparing their data for peer review and publication. "I think what we've done is reasonable and appropriate," says Alphey, "Few people in the field would disagree with the proposition that new tools are required for dengue, and this is a significant step forward."

Nidhi Subbaraman, *New York*

Vatican panel backs GMOs

A panel of scientists convened by the Pontifical Academy of Sciences (PAS) has made a passionate endorsement of genetically modified organisms (GMOs) for global food security and development. The statement, published in 16 languages in the 30 November issue of the journal *New Biotechnology* (http://www.ask-force.org/web/Vatican-PAS-Statement-FPT-PDF/PAS-Statement-English-FPT.pdf) is the result of a workshop held in the Vatican in May 2009, involving 7 members of the PAS and 33 outside experts. It states that "there is a moral imperative" to make the benefits of genetic engineering technology "available on a larger scale to poor and vulnerable populations who want them," urging opponents to consider the harm that withholding this technology will inflict on those who need it most.

The Vatican's Pontifical Academy of Sciences, headquartered at Casina Pio IV shown here, holds a membership roster of the most respected names in 20th century science.

The panel's key recommendation is to free transgenic varieties "from excessive, unscientific regulation" that hampers agricultural progress by inflating the costs needed for crop R&D. Ingo Potrykus, a member of the panel and co-inventor of Golden Rice, who is at the Swiss Federal Institute of Technology in Zurich, is still waiting for its beta-carotene–enriched seeds to reach the fields and sees that decade-long delay as a bitter lesson for agricultural biotech. "There is lots of high-quality publicly funded research and lots of goodwill for public-private partnerships to use the technology for humanitarian ends, but nobody can invest a comparable amount of funds [to that spent by large agricultural firms]. It will be mandatory to change regulation if we have any interest in using the technology for public good and in the public sector and with nonindustrial crops."

Influential people in developing countries—African bishops included—distrust GMOs as the tools of a plot by multinational corporations to make poor farmers dependent on multinational corporations. A 2009 draft document of the African Synod states that a campaign favoring agbiotech "runs the risk of ruining small landowners, abolishing traditional methods of seeding, and making farmers dependent on companies producing GMOs." But Robert Paarlberg, an agricultural policy analyst at Wellesley College in Massachusetts, who attended the Vatican meeting, believes those wary of GMO varieties should have greater confidence in the capacity of their local political systems to keep intellectual property (IP) issues under control. "They need to understand that patent claims over transgenic seeds made in countries such as the US do not extend to Africa," as Paarlberg argues that "national patent laws in Africa are more restrictive towards claims of IP."

If more is not done to encourage public sector involvement in developing GMO products there is also a risk that transgenic product development might be restricted to those players able to cope with regulatory red tape and fees (that is, multinational companies). "The cause for the '*de facto* monopoly' is neither the technology itself, nor the IP involved, nor lack of interest in [it] from the public sector. The only cause is present regulation," says Potrykus.

The panel's statement calls specifically for a revision of the Cartagena Protocol on Biosafety, which deals with international trade in living GMOs. "Groups opposed to the technology used it as a vehicle to persuade governments in Africa to set in place European-style domestic regulatory systems regarding the approval of GMOs," says Paarlberg. But Calestous Juma, professor of the practice of international development at Harvard University, is pessimistic that the Cartagena agreement may be revised to incorporate the Vatican group's recommendations. Juma, who was not involved in the meeting, suggests that communities should create their own treaties to support the advancement of the field. "Little will be gained from seeking to operate under a treaty that is so overtly hostile to innovation," he says.

The group's conclusions do not represent the official Vatican position, the Holy See press office stressed. Yet Gonzalo Miranda, a bioethicist of the Pontifical Athenaeum Regina Apostolorum, believes its scientific authority should carry weight. "The Catholic Church encompasses different sensibilities on GMOs but the trend is toward a cautiously open attitude because evidence of benefits mounts as time goes by and harms don't materialize," he argues. The proceedings of the study week are "an important indication that the Vatican continues to keep the matter under review and to listen to expertise. This is more than many leaders around the world have done and the Vatican should be commended," says Juma.

Anna Meldolesi, *Rome*

NEWS maker

Biocentury Transgene

Biocentury Transgene is not only going head-to-head against Monsanto in China; it's also poised to conquer markets in developing countries.

Which company is the largest producer of genetically modified (GM) cotton seeds in China? Not Monsanto or Syngenta, but rather a 12-year old Shenzhen–based agbiotech company. Biocentury Transgene currently dominates China's *Bacillus thuringiensis* toxin (*Bt*) cotton seed market and is riding high on a successful formula—a combination of locally developed GM crop varieties, cut-rate seeds and low patent licensing fees, wrapped up in China's poor intellectual property (IP) protection. In an unparalleled feat, Biocentury has overtaken international players like Monsanto, and is now accelerating efforts to expand into Southeast and South Asia. The question is whether Biocentury can replicate its impressive growth outside China.

The first commercial GM crop available in China—Monsanto's *Bt* cotton—was introduced in 1997. The following year, the St. Louis-based agrochemical company had captured nearly 95% of the emerging Chinese *Bt* cotton seed market.

Around the same time, the Beijing-based Institute of Biotechnology, under the Chinese Academy of Agricultural Sciences (CAAS), received approval from the Chinese authorities for its locally developed strain of *Bt* cotton. In 1998, the CAAS scientists who developed the *Bt* cotton, with the Institute of Biotechnology, co-founded Biocentury in South China's special economic zone. Plant researcher Sandui Guo at the Institute of Biotechnology, China National Centre for Biotechnology Development, launched the new venture, raising 54 million yuan ($8.2 million) with support from the Shenzhen government and investment capital from a local investor, Wu Kaishong (now chairman of Biocentury Transgene).

Key to Biocentury's success is the affordability of its cotton seeds, which are sold, on average, at half the price of Monsanto's. And although agrochemical multinationals continue to argue that farmers can save money by reducing pesticide use, Chinese farmers mainly seek cheaper seeds, according to Biocentury Transgene's CEO Yasheng Yang. In 2009, Biocentury made 120 million yuan ($18.2 million) in cotton seeds sales, accounting for 95% of the company's revenue.

But price is not the only reason driving Biocentury's expansion. In China, grassroots endorsement often wins over glossy advertising campaigns. As a result, technical personnel from Biocentury work with farmers in cotton fields.

Another issue is that multinational agrochemical companies have lacked a product adapted for local conditions. As Dafang Huang, former director of CAAS's Institute of Biotechnology, observes, seeds cultivated in the United States do not fully fit the natural conditions in China. For example, in southern China's Yangtze River regions, locally developed varieties are better suited to the humid climate and soil conditions.

A final issue is that Monsanto and other multinational firms have held back from licensing technologies to local partners for seed development because of concerns about intellectual property (IP) protection. Although an industry insider notes that Biocentury itself has also suffered from lax IP protection because after licensing its technology some companies use it to produce their own seeds.

Biocentury has nontheless gained from linking collaborations with various local seed firms to cultivate *Bt* cotton varieties with high yield, charging very low licensing fees to its partners to boost its seed production. The Chinese government has also helped to broaden Biocentury's influence as it offers financial support to all state-owned seed firms, many of which license Biocentury's technologies.

As a result, Biocentury's growth has been nothing short of meteoric. By 2003, areas planted with cotton derived from Biocentury's *Bt* technologies surpassed those planted with Monsanto's competing crops—reaching 70% of the total Chinese cotton plantations in 2005. Last November, Biocentury claimed that over 90% of China's *Bt* cotton plants are derived from its technologies.

Expansion to international markets began in 2003 when Biocentury set up an Indian office. Its *Bt* cotton was approved in 2007 in India and the seeds promoted through a partnership with Nath Seeds of Aurangabad.

As a result of Biocentury's market entry, Monsanto was forced to drop the price of its seeds from the original 1,900 rupees ($41.20) to fewer than 1,000 rupees ($21.47), according to Yang. However, Monsanto seeds still account for 90% of the total cotton plantations.

Biocentury Transgene

Company staff member explains planting skills to a cotton farmer.

In Pakistan, Biocentury has formed a 50:50 joint venture with Guard of Lahore and is seeking other local partners to speed up commercialization on approval of its *Bt* cotton. The company has also penetrated into Vietnam and Bangladesh with newly established branches and local R&D partnerships. According to Yang, Biocentury sees its future as a leading agbiotech player, particularly in developing countries. But at the same time, Monsanto is also stepping up its marketing efforts in developing nations. And, in terms of scientific know-how in seed development and new varieties on offer, the US giant remains ahead.

Biocentury's total sales of $18 million pale by comparison to Monsanto's $10.5 billion in revenue for 2010. Even though licensed products and seeds cover 3 million hectares, Huang says the company obtains only $5.70 from each *Bt* cotton hectare because of China's very diverse and highly competitive seed market, which makes charging higher prices extremely difficult.

For this reason, company management is actively investing in R&D efforts to launch more competitive products. The Chinese company is well placed to carry out transgenic R&D because much of the technology is well established. Indeed, Biocentury has spent $6.6 million in R&D in the past three years, accounting for >10% of its total sales (in comparison corporate R&D spending in China is typically <2% of total sales). Biocentury has also strengthened its research collaborations with academic institutions like CAAS and licensed out technologies from some of the world's leading firms. Last January, it entered into a licensing agreement with Rehovot, Israel–based FuturaGene to develop the latter's salt-tolerance genes in cotton plants in China. FuturaGene and Biocentury will share revenues generated by sales of the newly developed cotton seeds.

Biocentury's supremacy in the *Bt* cotton market in China is uncontested. But repeating this feat outside China's borders looks altogether a different challenge.

Hepeng Jia *Beijing*

Can cancer clinical trials be fixed?

As oncology drug after oncology drug fails to achieve accelerated approval, sponsors are seeking other ways to speed trials. Malorye Allison investigates.

The usual tension between the US Food and Drug Administration (FDA), drug developers and desperate cancer patients heightened this past year as the agency seemed to further tighten the reins on its accelerated approval program. Notably, Basel-based Roche group member Genentech suffered two setbacks in its breast cancer portfolio. But even Plexxikon/Roche's PLX4032, which delivered outstanding early trial results in metastatic melanoma, an essentially untreatable malignancy, faces a longer road to market than might have been expected. The message seems to be that nothing short of good luck can speed even the most promising cancer drugs to market these days.

What's particularly frustrating to industry and patient advocates is that all three of these drugs are targeted to specific subpopulations of patients, exactly what the agency has been calling for.

Genentech's two strikes

The first blow to S. San Francisco–based Genentech was a 12-to-1 vote against Avastin's (bevacizumab, anti-VEGF monoclonal antibody) approval for breast cancer. The blockbuster angiogenesis inhibitor was granted accelerated approval for this indication in 2008. That decision came against the recommendation of the Oncology Drug Advisory Committee (ODAC), which voted 5 to 4 not to make the designation. But in July of this year an ODAC committee considering final approval for the drug found "no convincing evidence that Avastin was clinically beneficial"[1].

FDA oncologic drugs head Richard Pazdur won't discuss particular decisions but he says that progression-free survival (PFS), which was the endpoint in all the Avastin breast cancer trials, can be problematic. Often, he says, the first phase 2 data look very good, but in subsequent trials, PFS benefit diminishes to the point of being barely noticeable. That's exactly what the ODAC saw when it reviewed the Avastin data.

The accelerated approval was granted based on data that showed a median PFS of almost a year, twice as good as standard therapy. But the follow-up trials, which included nearly 2,000 people, showed PFS that ranged from just under one month to almost three months. The committee also noted that the drug can have substantial side effects and patients taking Avastin had more severe side effects than those getting Taxol (paclitaxel) alone. (Avastin was studied in combination with Taxol in all these trials, and compared to Taxol alone.) In December, the FDA rescinded approval of Avastin for metastatic breast cancer, Roche immediately asked for a hearing.

Then last August the FDA issued a rare "refuse to file" letter about Genentech's trastuzumab-DM1 (Herceptin conjugated with a toxin) accelerated approval application. The biologic license application (BLA) filing was considered aggressive because it was based on the results of just one study, but Genentech may have had some encouragement to go through with it based on earlier discussions with the agency[2].

The data for that BLA came from a single-arm, 110-patient phase 2 trial that showed the drug-conjugate shrank tumors in one-third of women with advanced, HER2-positive breast cancer[3]. The patients had received an average of seven prior treatments, including the only two HER2-targeting drugs currently available—Herceptin (trastuzumab) and GlaxoSmithKline's Tykerb (lapatinib, a kinase inhibitor). "We feel those data were meaningful because these women were so sick and this is an unmet need," says Genentech spokesperson Krysta Pellegrino.

But FDA reviewers complained that the women had not received all available therapies approved for metastatic breast cancer, particularly drugs not specifically for HER2-positive cancer. That position puzzled Genentech and some physicians, who consider HER2-positive disease a specific subset. "It is difficult for me to comprehend why patients should have received therapies validated in the HER2-negative population," says Edith Perez, director of the breast program at the Mayo Clinic in Jacksonville, Florida. "We desperately need new drugs for this population and T-DM1 looks promising."

"There are other, non-HER2–targeting chemotherapies and hormone therapies that benefit women with HER2-positive disease," according to Pazdur. He adds that it is crucial to test against all available therapies to gain accelerated approval. "You can't just show efficacy. It has to be better than all the standard treatments."

That study used objective response as an endpoint, which was likely another issue. "It's clear now that FDA wants survival data from randomized trials for new breast cancer treatments," Pellegrino says. To get that for T-DM1, Genentech plans to amend the ongoing EMILIA trial, which pits the drug conjugate against GlaxoSmithKline's Tykerb plus Xeloda (capecitabine, a 5-fluorouricil prodrug). The company plans to resubmit the T-DM1 BLA in mid-2012, which means the drug could be launched in 2013—about two years later than expected.

The fact that PFS is out and overall survival is in will make a big difference to any company going after a breast cancer indication. Genentech is "looking at the implication of these decisions across our portfolio," says Pellegrino.

The Avastin decision in particular puts the FDA in the difficult position of having to restrict access to a drug that is now being reimbursed and used in some very sick patients. Some of them are fighting to keep the drug covered by calling their congressmen, writing letters to leading newspapers and appearing on television news to say that Avastin is keeping them alive.

But the case also highlights another dilemma. When expensive, targeted therapies don't have companion diagnostic tests, payers have to decide whether to keep spending so much on drugs that work in so few. Some observers have expressed concerns that price is starting to have an impact on FDA approval decisions, but Pazdur says, "Cost is never a consideration."

Having to go head to head against so many therapies further complicates drug development in breast cancer. "All those legacy drugs muddy the waters," says George Sledge, president of the American Society of Clinical Oncology (ASCO). Researchers feel there should be a way around that. "Just because there are so many treatments, this shouldn't be a hurdle to new drugs," says Perez. "Some of the drugs used haven't even been approved but were grandfathered in," she adds.

Pazdur says they just have to do the right trials, but sponsors complain that trials are getting bigger and bigger already. In breast cancer, available drugs have so improved prognosis that the typical trial size has swelled from around 600–1,000, up to 3,000 or even 10,000. "The trial design is driven by events, so you need more patients to get the same number of events," explains Don Berry, of MD Anderson Hospital in Houston. The trials also have to be bigger because it's harder to show a big clinical effect in breast cancer. Showing a 25% improvement takes more patients than showing a 50% improvement. "Companies are not going to run a lot of these 5,000 patient trials," he says.

Although safety is much less of an issue in oncology, the recent withdrawal of

Table 1 Oncology drugs in clinical trials

Accelerated approval			
Year	1995–2000	2001–2003	2004–2008
Number of drugs	6	7	6
NMEs (%)	26	78	32
Drugs in phase 2	5	5	4
Drugs in phase 3	1	2	2
Median years from IND to approval	5.5[a]	9.3[a]	6.7[a]
Regular approval			
Number of drugs	17	2	13
% NMEs	74	21	68
Median time from IND to approval	8.5	6.2	7.0

NME, new medical entities; IND, investigational new drug.

[a] FDA disputes calculations of approval times owing to incomplete data[12]. (Source: Adapted from ref. 11.)

Pfizer/Wyeth's Mylotarg (gemtuzumab ozogamicin, monoclonal antibody against CD33 conjugated to a toxin) could be another factor in the agency's position. Mylotarg was made available for patients with acute myeloid leukemia in 2000 through the accelerated approval program. Subsequent post-marketing data revealed the drug was ineffective and had a higher rate of potentially lethal side effects than anticipated based on earlier data. Pfizer voluntarily withdrew the drug last summer.

Better safe than sorry?

If the agency was already getting a reputation for being “too strict on cancer drugs” the clincher was the launch of a phase 3 randomized trial of PLX4032. This BRAF-inhibitor's performance in a phase 1 study was good enough to prompt one expert to declare, “We are on the verge of a paradigm shift for metastatic melanoma”[4]. More than 80% of the 32 patients enrolled showed tumor shrinkage of at least 30%. Most patients do not respond to either of the two treatments currently available for this disease, so PLX4032 has gotten a great deal of attention. The company is planning to seek accelerated approval, but that may have little or no impact on how quickly the drug reaches the clinic.

PLX4032 targets the oncogenic BRAF mutation V600E, which occurs in about 60% of malignant melanomas as well as some solid tumors. During the phase 1 trial, researchers studied glucose uptake by PET and pathway inhibition in tumor samples. “We could see that the tissue was responding,” says Kathleen Glaub, president of Berkley, California–based Plexxikon. One study showed that when ERK phosphorylation was inhibited by 80% or more, there was a greater chance of clinical response[5]. A phase 2 study is also ongoing, but is closed to further enrollment.

The FDA has been increasingly resistant to granting accelerated approvals based on phase 1 or 2 data because confirmatory data have been taking so long to materialize, if they materialize at all. Once drugs became available, sponsors were finding it difficult to recruit enough patients for the additional randomized control trial data the agency wanted; too many patients wanted to get the drug and not risk being assigned to the control. The agency has been waiting for companies to at least start enrolling phase 3 trials before it will consider an accelerated approval data package.

So, in January 2010 the first patient was recruited in a controversial trial that will include almost 700 melanoma patients. Half those patients will receive the alkylating agent decarbazine, a drug that is widely considered completely ineffective and toxic. Because PLX4032 is a pill and decarbazine has to be infused, the study is not even blinded. What's most contentious about the trial, however, is not just that so many patients are being subjected to a dubious treatment, but they will not even be allowed to cross over to PLX4032 if their cancer progresses because the trial endpoint is overall survival.

“Our hands are tied,” says Keith Nolop, chief medical officer at Plexxikon. He says the study design was decided upon after extended discussions with the FDA about the type of data needed to clinch an approval. Malignant melanoma trials have never established a relationship between PFS and overall survival. The incredibly messy back and forths between the FDA and Genta of Berkeley Heights, New Jersey about its failed melanoma candidate (Genasense, antisense against Bcl-2) have made this particularly controversial territory. “I do not doubt for a moment that [Genasense] was in the back of the company's mind when they decided on the type of trial,” says George Sledge, president of ASCO. Indeed, “What we heard over and over again from the agency was that there is no association between PFS and survival in melanoma,” Nolop says.

The phase 3 study, he says, was designed to be certain they could meet FDA's requirements for survival data that follows patients all the way to their deaths. That's extremely frustrating to those experts who see many similarities between PLX4032 and Gleevec (imatinib), which was approved in 2001 without a randomized control trial.

These circumstances are also “leading to some tough conversations with patients and families,” Nolop admits. News of the PLX4032 trial even reached the front page of the *New York Times* in a story of two cousins, both struck by melanoma but randomized to opposing arms of the study[6]. The one who received PLX4032 has been leading a practically normal life since starting the drug, but he had to watch his cousin worsen and die, despite the family's pleas that he, too, should get the “superpills.” One physician quoted said he'd seen PLX4032 produce a “Lazarus effect,” where patients on their deathbeds rise up and feel well again.

In the article, physicians were divided on the necessity for the trial. The investigator who led the phase 1 trial, Keith Flaherty of Massachusetts General Hospital in Boston, reportedly said, “I know all that I need to know based on the results we already have. My use of this drug is not going to be informed by testing it against a drug we all hate and would rather never give a dose of again in our lives.” According to the *Times*, Donald Lawrence of Massachusetts General Hospital e-mailed colleagues about the trial using “moral outrage,” as the subject line. “Just had yet another conversation with a [patient] with a B-RAF mutation who will die in the next month or so because he can't get PLX4032,” he wrote. “Compromising the phase 3 trial is not justification for withholding an effective drug from dying patients.”

Those who support the trial are also uncomfortable about it. Cy Stein, a physician and researcher at Montefiore Medical Center in the Bronx, New York, and one of Genasense's discoverers, sides with the FDA. “I feel very sorry for those patients on decarbazine, but I don't see any other way to do it than a randomized trial,” he says. “Look at Coley Pharmaceuticals,” he says, referring to the discontinued ProMune, an immunomodulator in trials for non-small cell lung cancer. “It looked great in phase 2 but went nowhere in a randomized trial.”

When this article went to press, Plexxikon released new data from the phase 2 trial with 132 patients. The response rate at that point was 68%, with three complete responses and 66 patients showing tumor shrinkage of at least 30%. Median PFS was 6.8 months versus 2 months for historical controls. Roche/Plexxikon also announced that the companies were planning an 'expanded access' program for BRAF-positive melanoma patients who have not responded to at least one other therapy. It's not perfectly clear, and may never be, whether the companies or FDA decided the randomized trial was necessary. As Sledge points out, "We always think it is mean old FDA's fault." But sometimes companies are just not willing to take the risk of losing an approval opportunity.

When PFS counts

Meanwhile, like PLX4032, Pfizer's Alk/Met inhibitor crizotinib is in phase 3 trials for lung cancer and will likely be submitted for accelerated approval. If it sails through that process, this drug will have been developed with lightening speed, but only because of a twist of fate. A phase 1 trial of the drug in a variety of tumors was ongoing when news broke in August 2007 that Alk-translocations were key drivers of lung tumors[7]. The first candidates for that trial were identified four months later and enrolled a month after that.

"They got lucky with crizotinib," says Edward Kim, chief of Head and Neck Oncology at the MD Anderson Cancer Center. "You're not going to get that kind of luck too often."

The trial was remarkably successful. The investigators reported a dramatic and durable (some lasting 15 months) response in 70% of the patients with another 20–30% responding less well but still benefitting from the drug. Most importantly though, PFS doesn't carry the same baggage in lung cancer. Patients in the control arm of the phase 3 whose cancers progress will be allowed to sign up for the phase 2 trial and receive the new drug, which may be the biggest breakthrough to date in lung cancer treatment.

Biomarkers and adaptive trials

Given the increasing constraints around accelerated approval, researchers have been looking for other ways to get cancer treatments moving more quickly. Pazdur emphasizes that biomarkers can help, but only if they are used early in development. In addition, "The agency expects to see the same level of evidence for a marker as for a drug," he says.

But the agency has delayed release of its long-awaited guidance on how to qualify biomarkers for FDA approval. Not having the guidance is "a big hold up," says Kim. Companies and clinical researchers have been waiting since 2004 (ref. 8) for the rules to be outlined in detail. The *Qualification Process for Drug Development Tools* guidance, which addresses biomarkers and other tools, was posted in October of 2010 (ref. 9) and the review process, which was slated to be completed by the end of the year had not been completed by press time.

Finding and validating biomarkers, especially in clinical trials, have been uncertain and costly processes. MD Anderson has been pioneering adaptive trials to help improve these processes. Using a unique design and atypical statistical approaches, studies like I-SPY and BATTLE evaluate many more markers and compounds in a shorter than usual time frame[10]. (The trials match outcomes from a set of drugs to molecular signatures of patients and use that information to inform assignments to patients entering the trials later.) How much time these trials will save is debatable. As noted, Pazdur is expecting thorough clinical validation of any markers used in trials. He and the MD Anderson adaptive trial designers call such trials "exploratory" and believe additional trials will be needed to confirm their findings.

Small time savings are possible, but won't be the rule. "Adaptive designs can usually improve three things: time, cost and quality of data," says James A. Bolognese, senior director of clinical trial services at Cambridge, Massachusetts–based Cytel. It's possible to improve one or two of those factors without a negative impact on the third, but not likely that all three will be bettered. Most importantly, adaptive trials can help drugs fail faster. The MD Anderson team is hoping that will hold true with putative biomarkers as well.

Berry, the statistician behind the I-SPY and BATTLE trials, says that if it appears a drug is working in a subset, adaptive trials allow you to test it in the general population but quickly "prune out" the nonresponders. That way, you are not exposing individuals to a drug that's doing them no good, and you also have data on how the drug acts in patients without the biomarker.

MD Anderson's adaptive approach could gain momentum. Several major biotech and pharmaceutical companies are involved in these trials. That's strong evidence that industry has heard Pazdur's call to "Get biomarkers early" and are acting upon it.

Can it be fixed?

"We view accelerated approval as a success story," says Andrew Emmett, managing director for science and regulatory affairs at Washington, DC–based Biotechnology Industry Organization (BIO). He points to FDA data that show that drugs on the accelerated pathway are approved a year earlier on average. But he also sees signs that FDA has become "more conservative." In early 2000 about 80% of all new molecular entities in development for cancer were on the accelerated pathway; today that number is down to 32%[11] (**Table 1**). BIO and others are working to make sure the accelerated pathway stays open. But it's clear that how oncology trials are designed, launched and run is also a big part of the problem.

The group that is most highly invested in speeding up access to cancer drugs are the patient advocates, and some are increasingly focusing their frustration on Pazdur and his agency. "Science is moving forward, and he is pulling the agency backward," says Lorton, Virginia–based Abigail Alliance's Steve Walker. In the age of "biomarkers and personalized medicine, we have the opportunity for a completely new approach." In the *Times* article, Charles Sawyers, of the Memorial Sloan-Kettering Cancer Center in New York, was quoted as saying that because chemotherapy is so toxic, it's crucial to know exactly how well it works. "But with these drugs that have minimal side effects and dramatic response rates, where we understand the biology, I wonder, why do we have to be so rigorous? This could be one of those defining cases that says, 'Look, our system has to change.'"

Nat Goodman, a patient advocate and computational biologist at the Seattle-based Institute for Systems Biology wishes that more scientists would be open-minded about randomized control trials. "There is an assumption that the current paradigm works," he says. "But there has never been a test done that validates whether the current FDA process has an acceptable error rate."

Even if exceptions can be made for drugs like Gleevec, that will not solve the problem for the majority of cancer drugs, especially in indications like breast cancer, which is already crowded with potential therapies.

"If everything was Gleevec, well, then everything would be easy," says Sledge.

Malorye Allison, Acton, Massachusetts

1. Allison, M. *Nat. Biotechnol.* **28**, 879–880 (2010).
2. Merrill, J. *The Pink Sheet Daily*, Elsevier Business Intelligence, 27 August 2010.
3. Vogel, C.L. *et al. J. Clin. Oncol.* **27** Supplement, 1017 (2009).
4. Mulcahy, N. *Medscape Today*, <http://www.medscape.com/viewarticle/703912> 4 June 2009.
5. Bollag, G. *et al. Nature* **467**, 596–599 (2010).
6. Harmon, A. *The New York Times*, p.1, 18 September 2010.
7. Soda, M. *et al. Nature* **448**, 561–566 (2007).
8. Anonymous. *Innovation or Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products* (FDA, March 2004). <http://www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunitiesReports/ucm077262.htm>
9. Anonymous. *Qualification Process for Drug Development Tools. Draft Guidance* (US Department of Health and Human Services, FDA, CDER, October 2010). <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM230597.pdf>
10. Allison, M. *Nat. Biotechnol.* **28**, 383–384 (2010).
11. Richey, E.A. *et al. J. Clin. Oncol.* **27**, 4398–4405 (2009).
12. Lanthier, M. & Sridhara, R. *J. Clin. Oncol.* **28**, e226 (2010).

Breaking the mold

With the construction of two new manufacturing plants, the bioplastics market emerges. Daniel Grushkin reports.

Take it as a metaphor for the bioplastics industry as a whole, or as another example of a fledgling industry stuttering toward success. In October of last year, Cambridge, Massachusetts, biotech Metabolix produced a new grade of its renewable, bacterially produced plastic, one that could come in contact with food. The announcement promised to open a new market for its biodegradable polyhydroxyalkanoate (PHA), branded as Mirel. But the news was undercut a month later when the commercial phase of production at its plant in Clinton, Iowa, stalled for another three months and the company's stock price dropped by 20%.

Bioplastics—some made from starch, others processed enzymatically or created wholly within bacteria—have the same properties as plastic but either biodegrade or come from renewable sources. For decades, scientists and entrepreneurs have been searching to co-opt plastic produced from oil with more sustainable alternatives (**Box 1**), and some analysts predicted a meteoric rise. In 2008, the US Department of Agriculture estimated biotech's contribution to the plastics market would multiply by a factor of 200 to reach 20% market share by 2025. Yet what many saw as inevitable stalled over the last two years as economies and capital markets buckled in the economic crisis. "The benchmark assumptions have not materialized," the US Department of Agriculture wrote in an addendum to its 2008 Biobased Products Report[1].

Now, at end of the global recession, new developments signal an optimistic but more sober reality for bioplastics. The European Bioplastics Association predicts that the industry will grow by 36% annually for the next three years[2]. The epicenter—at least for now—is Brazil, more specifically, the southern town of Triunfo, where South America's largest petrochemical company, Braskem, has its polyethylene plants.

A bioplastic is born

Five years ago, Jens Riese, a partner in the Munich office of McKinsey, gave a presentation to a number of chemical industry analysts where he outlined an economic model by which a company could manufacture biobased polyethylene, the most commonly used plastic. "People were laughing. They said, 'Ethylene! That's the least likely product to go bio.'"

In September Riese was vindicated. An army of 2,000 technicians completed the labyrinth of pipes, compressors and cooling towers for a new bioplastics plant at Braskem's Truinfo complex. The São Paulo–based petrochemical company nestled the plant beside two decades-old, behemoth crackers—structures that separate oil and gas into their chemical components (much of which goes into plastic production).

Though dwarfed in size by its neighbors, the 'green' plant is now the largest bioplastic plant in the world. Instead of creating ethylene from gas and oil, it uses homegrown sugarcane ethanol as its feedstock. The ethylene is then converted conventionally in a plant two kilometers away using the Ziegler-Natta polymerization process (which won Karl Ziegler and Giulio Natta the 1963 Nobel Prize in chemistry). The only difference in the green plastic can be found chemically. The radioisotope famous for dating

Box 1 Resurrecting an old technology

When Henry Ford showcased his 'soy car', a vehicle whose body was made from a plastic derived from soy fibers, at the Michigan State Fair in 1941, he hoped to bolster American agriculture with a new application. The idea faded with the onset of World War II, but bioplastics never quite disappeared.

Already in 1913, the Elektrochemische Werke in Bitterfeld, Germany, was producing ethylene from ethanol as a precursor for refrigerants. Later plants in Great Britain created polyethylene from ethanol using a similar technique. The chemistry is a simple reaction—ethanol to ethylene and water ($C_2H_5OH \rightarrow C_2H_4 + H_2O$). The steps have been around since 1797—long before anyone heard of environmentalism.

And bioplastics were still being used until two decades ago when the movement was still nascent. Whereas most manufacturers migrated to petroleum feedstocks in the 1950s, six major chemical companies operated ethanol-to-ethylene plants in India, Pakistan, Peru, Australia and Brazil. They could no longer compete after oil prices collapsed in the 1990s. All but one producer of ethylene oxide in India disappeared.

Unsurprisingly, the driving force behind Braskem's green polyethylene plant, Antonio Morschbacker, manager of biopolymer technology, began his career at Brazil's Salgema plant, which had produced 100,000 tonnes of biobased PVC a year. Long after the plant closed, Morschbacker was still searching for a new competitive edge. For a relatively new company, formed by the merger of two smaller chemical companies in 2002, Braskem has great ambitions: to be among the five largest petrochemical companies in the world.

Morschbacker spent months speaking with customers about their experiences using bioplastics. Their complaints were the same. "The products didn't have the properties required for their applications," he says.

"I remembered the old technology we used to make PVC," Morschbacker says. "My inspiration was the old plant." Braskem's engineers' challenge was to refine the ethylene to 99.95% purity. At a lesser grade the polyethylene can't set. Development came speedily. "It took us between 6–8 months, no more than that," says Morschbacker.

The purification operation is Braskem's trade secret. The company has yet to receive a patent and won't reveal how it works. However, this much is known: it comes at the final step of converting the ethanol to ethylene. After the ethanol is vaporized, it's moved to a reactor studded with beds of catalyst. More than 95% of the ethanol in the chamber converts into ethylene and water. The ethylene is quickly quenched in a cooling tower where it is separated from the water. It's then compressed and passed to a scrubber that removes contaminants. At Braskem's final, proprietary step, the ethylene is distilled with propylene, compressed and cooled multiple times until it reaches a level of purity where it can be polymerized into polyethylene using the Ziegler-Natta polymerization process.

Morschbacker admits that if it took his R&D department such a short time to develop the final step of the process, anyone with the research facilities can follow with equal speed. The original science may go back over 200 years, but according to industry analyst and polymer consultant Jim Lunt, "This is part of the wave of the future."

Box 2 PHA double take

Tillman Gerngross, at Dartmouth's Thayer School of Engineering, Hanover, New Hampshire, may be PHA's greatest detractor. He headed Metabolix's product development group from 1993–1998. "I spent five years of my career working in this area, and was and am a strong environmentalist, but when I looked at the overall picture, it didn't look as rosy as they were trying to portray," he says.

Gerngross studied the energy that goes into making the plastic. And according to his numbers, oil beats it every time. To generate one kilogram of polyethylene requires 47 million joules of energy, whereas PHA (from growing the corn, converting it into sugar, fermenting it with microbes and finally purifying it) requires 81 million joules.

If the energy driving this process derives from fossil fuels, plastic from oil is simply cleaner. "You have a process that emits more CO_2 and methane, that uses more land, more water, that releases fertilizer into the water. Please tell me how that is greener than fossil fuels?" Gerngross says.

Metabolix's communications director, Kristi Guillemette, says that Gerngross's assessment is "outdated and inaccurate." Gerngross's findings were made a decade ago when the Telles plant was still just a schematic. A more recent life-cycle analysis of PHA was completed in 2008 by Bruce Dale at Michigan State University in Lansing. "The early study by Tillman was pretty much theoretical. He had to, in essence, guess. By the time we began the study there was a lot more process data. So we could have a much better idea of the energy cost—and it was less," says Dale.

With a decade of research and a pilot plant in place, Dale's findings present a far greener picture. The production of one kilogram of PHA—from growing the corn to plastic resin—removes 2.2 kilograms of CO_2 from the air, and only burns 2.5 million joules of nonrenewable energy, according to Dale. Dale has received support and data from both Metabolix and Archer Daniels Midland, and a closer look reveals some of the assumptions in Dale's study may have stacked the deck in PHA's favor. Dale assumes that the cornfields used for plastic were not tilled (saving energy), and that the energy consumed in converting the corn into PHA is powered by burning the residue from the fermentation process. Any additional energy is offset by the purchase of carbon credits.

Though Metabolix has trumpeted the study, to Dale's knowledge, none of his carbon-saving measures have been adopted. It's quite possible that the plant will do all these things. Metabolix promised a more complete life-cycle assessment of Mirel by the end of 2010, but it won't be fully accurate until the plant is made fully commercial, and that has been pushed back to 2011.

artifacts, C^{14}, lingers at 1.2 parts per trillion—marking the plastic's organic origin.

"We discovered not only a new route to produce plastics but a way to do it cleanly," says Sergio Gomes, Braskem's project manager. Polyethylene ordinarily has a pitiful carbon footprint. When processed from oil, 2.5 kilograms of CO_2 are emitted for every kilogram of plastic produced. Green polyethylene, on the other hand, reverses the number: 2.5 kilograms of CO_2 are removed from the atmosphere for every kilogram of green plastic produced. The environmental gains derive from the fact that sugarcane is not only a renewable resource, but absorbs CO_2 in order to grow. Gomes sees Braskem's bioplastic turning a plethora of everyday objects made of polyethylene—bags, bottles, car parts—into carbon sinks.

The concept has reeled in a number of major manufacturers. Johnson & Johnson of New Brunswick, New Jersey, announced it will use the plastic in its Sundown sunscreen; Proctor & Gamble of Cincinnati, will include it in Pantene, Max Factor and CoverGirl products; and Toyota of Aichi, Japan, has already signed up to buy 50,000 tonnes for car parts. The deal was so attractive for buyers that the plant sold out its annual capacity of 200,000 tonnes even before construction ended.

More than just a boon to Braskem, the initial windfall reveals a market hungry for environmentally friendly plastics. Product manufacturers may be skittish about using a new material, but if it's identical to the plastic they're already employing (as it is with green polyethylene) they're ready to buy.

"They're targeting existing known applications. That's different from what the bioplastics industry has been doing up to now," says Jim Lunt, a polymer development consultant with his own firm, Jim Lunt and Associates, based in Wayzata, Minnesota.

Race for market share

The avid demand has not been lost on Braskem's competition, which is scrambling to catch up. In 2008, Dow Chemical, the world's largest plastics manufacturer, headquartered in Midland, Michigan, announced it would build its own 350,000-tonne-capacity, sugarcane-based polyethylene plant in São Paulo—Brazil's sugarcane heartland. The plant was originally slated to go live last year, but the company was hit heavily by the economic crisis. Despite the setback, Dow says it has purchased the sugarcane fields to supply its production and is waiting for the right moment in the crop cycle to lay down the money to begin construction.

Nearby, at a complex in São Paulo, the Brussels-based chemical giant Solvay also ventured into bioplastics. It announced that next year it will begin constructing a sugarcane-based polyvinyl chloride (PVC) plant, which will add another 100,000 tonnes of bioplastics to the market.

For its own part, Braskem's desire to lead the pack has moved the company far afield from its petrochemical pedigree. Last year, it partnered with the enzymes producer Novozymes, based in Bagsvaerd, Denmark, to build a new plant that produces polypropylene from microbes. Sugar from cane will fuel the fermentation. The partners plan for the plastic to have a similar price point to oil-based polypropylene. "Our general starting point is that we want to be cost competitive with the market," says Tina Sejersgård Fanø, senior director of renewable feedstocks at Novozymes.

The Danish company is now engineering microorganisms to create polypropylene directly inside bacterial cells (it will not reveal the type of bacteria). Novozymes expects to complete the genetic engineering by 2014 when the process will be tested in a Braskem pilot plant that will presumably look more like a brewery than a chemical cracker.

As major chemical companies enter the bioplastics industry, a trend is evident. The chemical composition of plastics won't quickly change. What will change is the feedstock. Aside from environmental benefits, large petrochemical companies have been searching to stem their dependence on oil, whose price, despite its wild fluctuations, has been steadily rising.

At the root of Brazil's magnetic attraction is the fact that the country makes ethanol at a cheaper price than anywhere in the world. At $0.87 per gallon, few can compete. Ethanol sourced from US corn, for example, costs almost double. According to Braskem, its bioplastics can remain price competitive with oil-based plastic even at $40 a barrel.

Cradle to grave

Before the case closes on petroleum-based plastics, however, the environmental impact of tectonic shifts currently taking in the bioplastics marketplace deserves a deeper look. Though bioplastics address the carbon emissions posed by plastics production, they neglect the second half of society's trouble with plastics. Plastic generates 250 billion tonnes of material annually. A paltry 6.8% is recycled in the US. Where that plastic arrives at the end of its life cycle is the part of the equation that producers like Braskem don't address.

For the plastic to be truly green, it will have to be recyclable, and delving into the question of plastics recycling gets messy quickly. Some countries have more comprehensive waste disposal programs than others, and some plastics are more recyclable than others. What cannot be recycled ends up in incinerators, landfills or, worse, littering the landscape, where much of it runs into streams and rivers and eventually oceans.

"The trend now is to make conventional plastics from renewable resources, so the bioplastics industry is dividing into single-use disposable or compostable materials, and durable products, which are not designed to be compostable," says Lunt.

Genetic engineering to the rescue?

When considering bioplastics from the viewpoint of their end of life, there are no perfect plastic replacements. Of the 11 major varieties of bioplastics, each has drawbacks. Compressed starch plastic, which is molded from cornstarch and therefore biodegrades, has the largest market share of bioplastics. However, it doesn't have the material strength of polystyrene and, furthermore, absorbs water, making it a less-than-ideal replacement for carrying foodstuffs. To address the issue, bioplastics company Cereplast, located in Hawthorne, California, the US's largest starch plastic producer, blends in polylactic acid (PLA), a biodegradable plastic fermented from glucose, to make up for starch's weakness. For more durability, the company adds polypropylene and thereby the plastic loses its compostability.

Another example is NatureWorks, the largest bioplastics maker in the US. Its product, Ingeo, is wholly PLA. Though it's useful in fabrics and packaging, it must be heated above 140 °F to biodegrade, which means it will only properly end its life cycle in an industrial composter, which few municipal sanitation departments possess.

One molecule stands out—PHA. Six months before Braskem completed its plant in Brazil, Archer Daniels Midland of Decatur, Illinois, and Metabolix completed their own bioplastics plant in Clinton, Iowa—right beside an Archer Daniels Midland corn mill. The plant, a joint venture, called Telles, produces 50,000 tonnes of Mirel, its brand name for PHA. The plastic's feedstock can vary from corn, as in the Clinton plant, to sugarcane, or even to biomass, as it is produced in bacteria which can grow on any sugar feedstock.

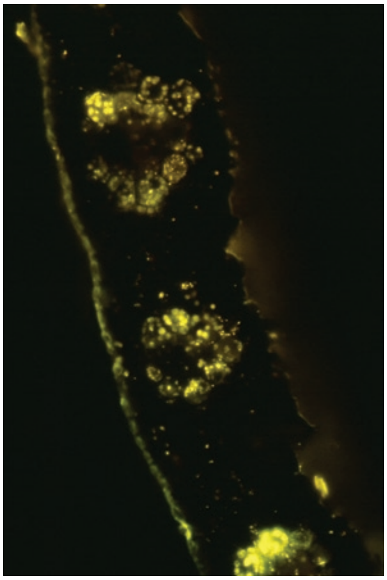


Figure 1 Plastic grass. Metabolix has engineered switchgrass to grow granules of PHA resin (yellow fluorescence). (Source: Metabolix)

PHA has potential because unlike bio-based polyethylene, it biodegrades in the environment (over a couple of days to a year depending on its thickness), which means a PHA bag washed offshore won't have a chance to clog the intestines of sea life because it decomposes both in water and in landfills (but see **Box 2**). The molecules degrade because they are broken down in bacteria and algae that have the enzymes to digest them. PHAs form as granular energy stores in the cytoplasm and are broken down when the organism needs fuel.

To produce the plastic, Metabolix has engineered a pathway in *Escherichia coli* to produce these PHA granules, but because *E. coli* doesn't have the enzymes to break it down, "they pretty much grow themselves to death filling up with the stuff," says Oliver Peoples, chief scientific officer of Metabolix.

Unlike starch, which is broken down hydrolytically, PHA will degrade only in the presence of microbes that have the enzymatic machinery. "If [PHA containers are] sitting in a perfectly normal dry place or filled with shampoo they'll sit pretty much forever. But if you take them out and throw them into your backyard compost, they'll go away," says Peoples.

The technology rose from the ashes of 60 years of failed PHA ventures. London-based Zeneca marketed the polymer as a product called Biopol in the 1990s at $30 a kilogram. The cost was too high for any mass applications, but the promise of the material caught other companies' interests. In 1996, Monsanto bought Biopol and tried to genetically engineer rapeseed into producing the polymer directly. The business was another failure for PHA. While other companies stumbled, the researchers at Metabolix were developing their own PHA technology. Monsanto sold them the intellectual property for Biopol in 2001.

Nine years later, Mirel finally arrived. The price has come down considerably from its original price—now $2.50 per kilogram—but it's still double the price of conventional plastic. Metobolix is trying to lower costs by engineering feedstock crops to produce PHA directly.

The company has been engineering sugarcane in Australia, switchgrass in Cambridge, Massachusetts (**Fig. 1**), and camelina, a fast growing oilseed from the mustard family, in Saskatoon, Canada. So far, camelina has seen the most success, and Metabolix expects to see its first commercially viable crop in two years. PHA will make up 10–20% of the seed weight. If it succeeds, the company will produce a plastic that "will come under the price of polypropylene and polyethylene," according to Metabolix CEO Richard Eno.

Tradeoffs and choices

As the industry leaps out of infancy, it faces a number of uncertainties. Bioplastic's environmental impact, though less than that of conventional plastics, opens a number of social questions about how we allocate natural resources—whether we are putting fuel and plastics production in direct competition with food and water consumption.

The existing variety of bioplastics depends on corn and sugar feedstocks, and according to Peoples, their markets are already tightly linked to oil because of ethanol's use as an alternative fuel. "The buffer that was once there has almost completely washed away," says Peoples. "That's going to have an impact on prices." How the rising bioeconomy will shift other economies is still unknown.

As with all environmental issues, solutions come piecemeal. The same can be said in the search for a suitable suite of bioplastics. The petroleum-based plastics industry itself took 50 years to mature. To achieve a comprehensive alternative, bioplastics may need just as much time.

Daniel Grushkin, Brooklyn, New York

1. Anonymous. *US Biobased Products Market Potential and Projections Through 2025* (US Department of Agriculture, February 2008). <http://www.usda.gov/oce/reports/energy/BiobasedReport2008.pdf>
2. Shen, L., Haufe, J. & Patel, M.K. *Product Overview and Market Projection of Emerging Bio-Based Plastics* (PRO-BIP 2009) (European Polysaccharide Network of Excellence, Utrecht, The Netherlands, November 2009).

Battling infringement

Steven A Bogen, James M Smith & John L DuPré

As more startups share competitive business information in their search for partners within big pharma or biotech, what legal protections are available if their intellectual property is infringed upon?

In the biopharmaceutical sector, in which large firms and small startups or universities often cross paths and frequently work together, a dispute over patent rights can be wildly unbalanced. In a David versus Goliath intellectual property (IP) battle, the large, well-financed Goliath can sometimes minimize or nullify David's hard-earned legal rights, thereby gaining freedom to operate. In fact, if a large company is willing to take risks, it's sometimes financially *beneficial* for the firm to just appropriate (rather than license) the technology.

In the following article, we expose some of the tactics used by large, well-established and well-financed companies to acquire a university's or small company's IP. We also offer suggestions that may be helpful in leveling the playing field, because if you're prepared, you may succeed in stopping the infringement and collecting damages.

Box 1 The consequence of eBay Inc. v. MercExchange, LLC

Before 2006, a university or a small company plaintiff forced out of business because of the delays in enforcing its patents could still have obtained a permanent injunction against the infringer. Permanent injunctions were usually granted to a patent holder after prevailing on liability (patent infringement and validity). The consequences of a permanent injunction were often so disastrous to an infringer that they served as a deterrent against egregious acts of intellectual property appropriation. Before 2006, driving the small company out of business did not eliminate this threat. That changed with the US Supreme Court decision of *eBay Inc. v. MercExchange, LLC*.

The eBay case was, in part, a reaction to the practice of patent enforcement by companies not competing in the marketplace. As a consequence of the eBay decision, injunctions prohibiting the manufacture, use and sale of infringing products are far less likely to be granted to companies not practicing the patented technology. If the company has been driven out of business because of the delays in enforcing its patents, then it clearly is not practicing the technology. In such circumstances, the courts may find that monetary remedies are sufficient to compensate for the injury caused by infringement. Namely, the small company can be totally compensated by money. Also, because the small company is no longer in business, the public interest may weigh against granting an injunction that shuts down an infringer. *eBay Inc. v. MercExchange, LLC* dramatically changed the risk profile for large, well-financed companies using technologies covered by patents from universities or small startups that have not yet commercialized the technology, and now the serious consequences to a large company of shutting down their manufacturing line and cutting off customers are less likely.

Exposing yourself

How does a large company find out about your technology anyway and decide to infringe it? The truth is, you can be exposed to the risk of IP appropriation through the normal process of seeking partners or financing your company. This is because at least some information must be disclosed in order for a prospective investor to perform due diligence—this can't be avoided. Companies seeking funding usually expect that a prospective investor will treat the information as confidential, sensitive business information, but the funding process presents inherent risks to confidentiality. And it's possible another company will act on the information it sees.

Many entrepreneurs think that if this does happen, a preliminary injunction can serve as a legal mechanism for protecting small companies in patent litigation. Although a court *can* grant a preliminary injunction, which would bar the defendant (in this case, the larger company) from making, using or selling the product before the full determination of the merits of a case, it is unlikely. In seeking a preliminary injunction, you need to demonstrate that there is a substantial likelihood of success on the merits of the case and that there is a threat of irreparable injury. Without a prior successful trial, it is difficult to demonstrate substantial likelihood of success.

It is also hard to establish the threat of irreparable injury, as the courts often conclude that any harm caused to you, the plaintiff, can be compensated with a monetary payment. Even if you were forced out of business, the harm can theoretically be compensated monetarily and thus is considered reparable.

There is a third thing working against you when seeking a preliminary injunction: the requirement for posting a bond. The bond is meant to compensate the defendant for financial losses if the defendant prevails. Perhaps you have the resources to post a multimillion-dollar bond, but few small startup companies or universities do.

If the court will not grant a *preliminary* injunction, then your goal is to secure a *permanent* injunction. This is the type of injunction

Steven A. Bogen is medical director of the clinical chemistry laboratory at Tufts Medical Center, Boston, Massachusetts, USA. James M. Smith and John L. DuPré are principals at Hamilton, Brook, Smith & Reynolds PC, Boston, Massachusetts, USA. e-mail: s.bogen@tuftsmedicalcenter.org, james.smith@hbsr.com or john.dupre@hbsr.com

that would be issued after a trial if you win. The problem for you as a small company is that, to get a permanent injunction, you must hit a proverbial home run. First, you need to be funded. Then, you need to work out the technical aspects of your technology, manufacture it, find a distribution channel and launch a product. (Although you can bring legal action, even without launching the product, you would have difficulty obtaining an injunction; see **Box 1**.) Next, you need to be awarded one or more patents with which to allege infringement. Finally, you need to actually win in court, at both the district and appellate levels. It's a lot; just your end of things that need to happen is a major uphill battle.

The climb steepens

Unfortunately, besides the hurdles you face on your own, there are many tactics a large company can use to stop a small company patent owner along the way. The first is financial. Until the trial, if the competing large company copies and markets essentially the same technology, then there will be two similar product offerings available to customers. The large company will likely enjoy brand-name recognition and lower distribution and service costs because, as an established company, it probably also handles other products. For these reasons, the large company will probably outsell you and might be able to bleed you dry (**Box 2**).

Once the litigation has started, the defendant might hire an aggressive law firm to generate large numbers of legal motions, document discovery requests and deposition notices. This will force your law firm to respond, driving up your legal bills and increasing your rate of cash burn. The entry fee—the absolute minimum a lawsuit is likely to cost to get to trial (not including appeals and a potential second trial for determining damages)—is about $1 million. More commonly, the figure is several million dollars. If the infringer hires an aggressive law firm, the number is likely to be even higher.

An additional tactic of a larger company is to simply find reasons to push the trial date out as far as possible, giving the large company more time to sell its product and drain your finances. Often, the judicial process moves at a glacial pace. A 2008 study of patent litigation measured the time to trial among 394 trials in 65 US federal districts. The national median time to a first trial, for assessing liability, is 2–3 years[1]. From start to finish, the process can take a decade. For a small company, the years required for enforcing IP can be lethal. Until an injunction is issued, the large, entrenched competitor may be dominating the market with the technology initially developed by you—and you may go out of business in the interim.

Another delaying tactic is to file a reexamination request with the US Patent and Trademark Office (USPTO), which undertakes reexamination of issued patents (35 U.S.C. 302). Patent reexaminations were intended to provide an inexpensive mechanism for challenging issued patents without requiring a court proceeding. Because it is inexpensive relative to litigation, the reexamination process is designed to provide an even footing for both small and large companies. There is, however, an alternative use for a patent reexamination that has little to do with its original purpose. It is possible to use it to delay court proceedings.

Within 3 months following the filing of a reexamination request, the USPTO will determine whether the requestor has raised a substantial new question of patentability. The threshold for this determination is not necessarily high. Often, that threshold can be met by the identification of new prior art, not considered by the patent examiner during the original prosecution, that creates a question as to novelty or obviousness of the patent claims. If these conditions are met, then the USPTO will reopen the patent examination.

Once the reexamination is started, the large company defendant can file a request with the court to stay, or place on hold, all legal proceedings pending the reexamination's outcome. There is no guarantee that the court will agree to a stay, but the argument can be compelling from the standpoint of judicial efficiency. Courts are not inclined to spend judicial time and resources if there is a chance the patent will not survive reexamination. Reexaminations can take at least 1 year—and often several—which for a stayed case can be a substantial delay.

A reexamination can also be filed even if the plaintiff already won at trial. If the reexamination request is filed after the determination of patent validity, then it may seem to the small company or lone inventor patentee like a sort of judicial double jeopardy. The defendant gets a second chance, with the same or similar arguments, to try to invalidate or narrow the patent claims. The fact that an invalidity argument was already adjudicated at trial has little weight in the reexamination. Juries are instructed to view issued patents with a presumption of validity unless proved otherwise. No such presumption of validity exists during a reexamination proceeding. The two different standards justify subjecting the patent holder to two routes for challenging patent validity.

Even if the defendant is ultimately found liable for infringement, the damages awarded may be relatively small compared with the potential upside in value appreciation the small company might have seen had the infringement never occurred. Patent law (35 U.S.C. 284) allows you to recover only damages adequate to compensate for the infringement but not less than a reasonable royalty. A lost-profits calculation is one possible basis for damages that usually yields a higher number than a reasonable royalty calculation. However, it only applies to companies that are in business and selling the product. If your company did not survive during the years of litigation, then you may have to settle for a reasonable royalty. Because universities are not in the business of manufacturing and selling the product, they cannot be awarded lost profits, only a reasonable royalty. This means that the defendant's worst-case outcome is that the damages will cost only what they would have paid anyway. In this case, there is no penalty to the large company for appropriating, rather than licensing, the IP other than the cost of litigation.

If the defendant is found to have willfully infringed, then that can result in greater damages. Many assume that the risk of double or treble damages will discourage large companies from aggressively appropriating

Box 2 Why litigation delays are lethal to small companies

Litigation battles that stretch 2 years or longer can cripple a small company. The enterprise is hemorrhaging cash and the court appears to be unconcerned with its plight by not moving the matter to a quick resolution. What should have been an innovative new product technology is now, because of the infringement, a me-too product. A small company may try to publicize the fact that they have the patent rights and are suing for patent infringement, but customers are often not concerned with the patents and IP ownership. On the other hand, the infringing, larger company might tell prospective customers that the small company may be out of business soon, further compromising potential reputation and sales. With no near-term solution in sight, the small company might exhaust its cash on the litigation battle and end up being acquired at a fire-sale price. This is an opportunity for the competitor or a third party to acquire the small company's IP rights at a steep discount—effectively driving the small company out of business.

a small company's (or university's) IP. Recent appellate court decisions, however, make it more difficult than ever to prove intent, rendering a finding of willful infringement an unlikely outcome.

Suggested steps

None of this looks particularly easy, and spending all this time enforcing your IP rights would detract from your real focus of running a business. Certainly, the concerns of the marketplace might seem a long way away if you're just launching your company with preclinical data. But there are steps you should take early on to mitigate risks, secure defensible patents and preserve confidentiality while seeking business partners or funding.

First, maintaining confidentiality early in the process is important. When providing a business plan to potential venture capital investors, obtaining a written confidential disclosure agreement is unlikely. Instead, document the recipient's existing confidentiality policy (it may be posted on its website). It may state that the information will be kept within the firm and its pre-established group of advisors. If there is no written policy, ask.

Second, boldly print or stamp the word "confidential" on each page of the business plan, not just on the front page.

Third, print the business plan on paper that does not allow clean copies (for example, Boise BEWARE security paper).

Fourth, do not provide a digital file of a business plan because it is easy to copy and retransmit.

Fifth, insert on each business plan page a discreet code linked to the business plan number. Retain a master list of the codes/business plan numbers so that even isolated pages from the document can be traced back to the recipient and you'll know where a leaked document page came from.

Sixth, minimize the number of business plans that you send out. Too many creates the appearance that the documents are in the public domain and increases the risk of disclosure by one recipient. Disclosure of the underlying information from a single recipient to an innocent third party can destroy the confidentiality of the information.

Seventh, avoid sensitive information in your business plan. That's easily done if the sensitive information is limited to a secret formula, but when it can be a new, previously unappreciated market opportunity, avoiding it becomes much harder.

Eighth, to the extent that confidential information is patentable, it is important to protect your IP rights as best you can before sending out business plans. So make sure you have complete and well-drafted patent applications filed before you disseminate business plans due to the possibility of a breach of confidentiality that could interfere with subsequent patent filings.

Ninth, enforcing IP also depends on the strength of your patents. This is an area in which the relationship to your patent counsel can be of paramount importance. Maintaining a continuation application pending throughout the enforcement period is a practice that has been criticized by some, but it can help balance the practices of infringers. If you have a continuation patent application pending, you can present newly cited prior art for consideration by the examiner without the limiting procedural requirements of reexamination. A continuation also enables the patent owner to tweak the claims as required in response to unexpected interpretations of claim language by the courts or redesigns by infringers attempting to avoid a lawsuit while copying the essence of the invention. A close and continuing relationship between an inventor and patent counsel can be the difference between success and failure in the patent enforcement process.

And finally, many of the constraints that the US legal system places on small companies and universities (in trying to enforce their IP) can be overcome by partnering with a large company against the infringer. In fact, this is an area in which your interests may align quite well. A large-company partner will have the financial resources to see the litigation through to the end. Because the partner is selling the product, it can obtain an injunction and argue for damages from lost profits rather than from a reasonable royalty. From the partner's standpoint, it has the opportunity for an exclusive license to your technology. The fact that a large competitor is infringing somewhat validates the market opportunity.

Conclusions

For universities and small companies, enforcing IP presents many unique challenges that large-company plaintiffs are unlikely to encounter. Part of the solution may be the creation of a new mechanism for prompt adjudication. Dragging the litigation out for up to a decade makes patent enforcement out of reach for many small businesses, and this can ultimately stifle competition and drive small companies out of business—both contrary to the general interests of the United States, if this happens to be where your company is located. Unfortunately, the only way to have a meaningful solution to this problem is with an act of US Congress. In the meantime, entrepreneurs should do all they can to keep infringers at bay.

1. Levko, A., Torres, V. & Teelucksingh, J. A closer look: 2008 patent litigation study: damages awards, success rates and time-to-trial. PwC <http://www.pwc.com/en_US/us/forensic-services/assets/2008_patent_litigation_study.pdf> (2008).

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

Essential information for synthetic DNA sequences

To the Editor:

Following a discussion by the workgroup for Data Standards in Synthetic Biology, which met in June 2010 during the Second Workshop on Biodesign Automation in Anaheim, California, we wish to highlight a problem relating to the reproducibility of the synthetic biology literature. In particular, we have noted the very small number of articles reporting synthetic gene networks that disclose the complete sequence of all the constructs they describe.

To our knowledge, there are only a few examples where full sequences have been released. In 2005, a patent application[1] disclosed the sequences of the toggle switches published four years earlier in a paper by Gardner *et al.*[2]. The same year, Basu *et al.*[3] deposited their construct sequences for programmed pattern formation into GenBank[3]. Examples of synthetic DNA sequences derived from standardized parts that have been made available in GenBank include the refactored genome of the bacteriophage phage T7 (ref. 4) and a BioBrick-based plasmid[5]. More recently, the full genome sequence of synthetic *Mycoplasma mycoides* JCVI-syn1.0 clone sMmYCp235-1 also has been made available in GenBank (accession no. CP002027)[6].

In contrast, most publications provide a variety of methods, information and/or partial sequences to explain the constructs used in a paper; for the research community, piecing together the full sequences of constructs is thus laborious, error-prone and sometimes impossible. A paper from your journal provides a recent example; although Kemmer *et al.*[7] provided admirably detailed Supplementary Information on the construction methods for their plasmids, they failed to provide access to the final sequences. Indeed, the gaps between key components are almost never reported, presumably because they are not considered crucial to the report. Yet, synthetic biology relies on the premise that synthetic DNA can be engineered with base-level precision.

Missing sequence information in papers hurts reproducibility, limits reuse of past work and incorrectly assumes that we know fully which sequence segments are important. For example, many synthetic biologists are currently realizing that translation initiation rates are dependent on more than the Shine-Dalgarno sequence[8]. Sequences upstream of the start codon are crucial for translation rates, yet are underreported. Similarly, it has been demonstrated that intron length can affect the dynamics of genetic oscillators[9]. Many more such examples are likely to emerge.

Because full sequence disclosure is critical, we wonder why the common requirement by many journals to provide GenBank entries for genomes and natural sequences has not been enforced for synthetic DNA and engineered genetic constructs. In an environment where word count is a constant battle, replacing plasmid construction method sections with references to annotated GenBank entries would be a welcome change. We therefore feel that including a completely annotated sequence of the construct would greatly contribute to the development of our discipline. We hope that in the future you will encourage the authors you publish to submit this information to GenBank or other appropriate databases. In the long term, we hope to establish a minimal information guideline around the Minimal Information about a Biomedical or Biological Investigation (MIBBI; http://mibbi.org/index.php/Main_Page) project and welcome contributions from the greater community.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Jean Peccoud[1], J Christopher Anderson[2], Deepak Chandran[3], Douglas Densmore[4], Michal Galdzicki[5], Matthew W Lux[1], Cesar A Rodriguez[6], Guy-Bart Stan[7] & Herbert M Sauro[3]

[1]Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, Virginia, USA. [2]Department of Bioengineering, QB3: California Institute for Quantitative Biological Research, University of California, Berkeley, California, USA. [3]Department of Bioengineering, University of Washington, Seattle, Washington, USA. [4]Department of Electrical and Computer Engineering, Boston University, Boston, Massachusetts, USA.[5]Biomedical and Health Informatics, University of Washington, Seattle, Washington, USA. [6]BIOFAB, Emeryville, California, USA. [7]Department of Bioengineering and Centre for Synthetic Biology and Innovation, Imperial College London, London, UK.
e-mail: peccoud@vt.edu

1. Gardner, T.S. & Collins, J.J. US patent 6,841,376 (2005).
2. Gardner, T.S., Cantor, C.R. & Collins, J.J. *Nature* **403**, 339–342 (2000).
3. Basu, S., Gerchman, Y., Collins, C.H., Arnold, F.H. & Weiss, R. *Nature* **434**, 1130–1134 (2005).
4. Chan, L.Y., Kosuri, S. & Endy, D. *Mol. Syst. Biol.* **1**, 2005.0018 (2005).
5. Shetty, R.P., Endy, D. & Knight, T.F. Jr. *J. Biol. Eng.* **2**, 5 (2008).
6. Gibson, D.G. *et al. Science* **329**, 52–56 (2010).
7. Kemmer, C. *et al. Nat. Biotechnol.* **28**, 355–360 (2010).
8. Salis, H.M., Mirsky, E.A. & Voigt, C.A. *Nat. Biotechnol.* **27**, 946–950 (2009).
9. Swinburne, I.A., Miguez, D.G., Landgraf, D. & Silver, P.A. *Genes Dev.* **22**, 2342–2346 (2008).

***Nature Biotechnology* replies:**

Kemmer *et al.*[1] have now lodged the sequences of the constructs used in their paper with GenBank HQ644133, HQ644134, HQ644135, HQ644136 and HQ644137. *Nature Biotechnology* and other Nature research journals currently require disclosure only of the sequences of genomes, deep sequencing and short-read data, short stretches of novel

sequence information (e.g., epitopes, functional domains, genetic markers or haplotypes) and their surrounding sequence information as well as any RNA interference, antisense or morpholino probes used in a paper (http://www.nature.com/authors/editorial_policies/availability.html); there is no consensus as yet that the sequence of every plasmid used in every paper should be lodged with GenBank or that such a policy would be beneficial to the wider community.

Even so, as Peccoud *et al.* point out, full sequence information is often essential to reproduce the findings reported in papers in the area of synthetic biology. As such, on a case-by-case basis, *Nature Biotechnology* will encourage authors of such papers to lodge the sequences of the constructs used in a paper in GenBank together with the corresponding accession numbers.

[1] Kemmer, C. *et al. Nat. Biotechnol.* **28**, 355–360 (2010).

Is transgenic maize what Mexico really needs?

To the Editor:

In the past three years, substantial progress has been made in updating knowledge on the present diversity of maize landraces and where these are still being grown within the Mexican territory. Here, we summarize some of these findings and briefly discuss their implications in relation to maize production and use in Mexico.

The term landrace was first described by Anderson and Cutler[1] as "a group of related individuals with enough characteristics in common to permit their recognition as a group." It refers to the varieties and populations of native maize in Mexico and has helped in the study of the genetic diversity of the crop.

As part of the implementation of the Biosafety Law—legislation that passed in March 2005 requiring the definition of both the areas of origin for crops native to Mexico and their genetic diversity—the Mexican government has been carrying out a survey of maize landraces since 2006. The program was financed with $1.5 million from the Ministry of Agriculture, Livestock, Rural Development, Fishery and Food (SAGARPA), the Ministry of Environment and Natural Resources (SEMARNAT) and the Inter-Ministerial Commission for Biosafety of Genetically Modified Organisms (CIBIOGEM).

Some of the key findings of this survey are as follows (for one of the results already published in Spanish, see http://www.biodiversidad.gob.mx/genes/origenDiv.html and ref. 2): first, a large number of maize landraces are currently being cultivated very widely in Mexico; second, diversity in maize landraces under cultivation is superior to what was originally believed to exist before the study started (in particular for the northern states of Mexico); and third, probable new maize landraces have been identified, diversity is higher than previously appreciated within landraces (such as Tuxpeño, which is the number-one provider of germplasm to most of the maize known in commercial breeding), and new teocintle (the most probable progenitor of maize) populations have been identified.

Maize genetic diversity exists as a result of the activities of small farm-holders (their plots currently represent 86% of the area where maize is cultivated in Mexico), who generally plant maize for subsistence[3] and depending on rainfall, permanently experiment and exchange seeds, and have designated many uses for the different variants cultivated[4,5]. It is because of these traditional agricultural practices that Mexico preserves and enhances the many different maize landraces we now know[6,7].

The new data acquired about the present number and distribution of maize landraces underline, on the one hand, the richness of genetic diversity of cultivars and, on the other, the reasons Mexico has for valuing and maintaining that diversity for future breeding needs. It is thus important that the very process by which those landraces are generated and maintained (that is, the practices of the small farmers) is preserved.

Currently there is no commercial production of transgenic maize in Mexico; only experimental trials have been approved. The question has been raised as to how Mexico will manage the commercialization of transgenic maize together with meeting its responsability of safeguarding the characteristics of the genetic diversity that has been revealed in the recent study. Much debate, some of it scientifically based, has taken place about the risks and benefits of allowing experimental trials of transgenic maize in a center of genetic diversity for the same crop. It is our opinion that some relevant questions about the potential impacts of transgenic maize on landraces have not been addressed either in these discussions or by experiments. For example, further experimental work is required to establish the potential for gene flow from transgenic maize to landraces, measures for managing this gene flow and the potential long-term impact of gene flow on landraces.

If gene flow from transgenic maize to landraces occurs, several other questions arise. How will intellectual property issues interact with the biological, social and economic reality of small-farmer agricultural practices that maintain and keep generating new variability in maize landraces in Mexico? What are the practical consequences for a small subsistence farmer cultivating native landraces of maize and finding his crops contain genes from transgenic plants? What is the legal position of such a farmer and is he/she likely to be infringing patents by cultivating or exchanging (knowingly or not) seeds that contain transgenes? What would be the stance of agbiotech companies in pursuing their intellectual property and licenses in such situations? Such questions need to be considered both at the small rural community level and nationally.

There is also the broader issue of the extent to which introduction of transgenic maize will provide solutions to existing problems for Mexican agriculture, such as the migration of male peasants (especially young people) to cities and abroad, an increasingly older rural population, the absence of effective mechanisms and incentives to cultivate maize landraces in a certified manner, weak market and grain distribution arrangements, and increasingly dominant patterns of food consumption based on foreign models of fast foods.

Mexico does not yet have in place a working and efficient mechanism for monitoring cross-pollination and gene flow under local agricultural conditions, despite claims that this is being instituted[8]. Information is lacking on the value that transgenic maize has for Mexican farming systems and its management requirements. Meanwhile, illegal transgenic maize introductions have been documented, and in some cases prosecuted, in Mexico. Moreover, there are concerns about the introduction of transgenic maize developed for pharmaceutical or other non-food purposes, and its impact on landraces[9,10].

Mexico needs to be able to define what kind of transgenic materials (for maize and any other relevant crop) it needs for its ecological, social and economic requirements. This responsibility must be carefully analyzed in order to provide farmers with adequate and necessary elements to help achieve a level of food security for the present and future of Mexican society, while conserving genetic diversity and helping develop adequately the social structures of the rural economy and society.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

DISCLAIMER
The manuscript reflects only the opinion of the authors and not the institution they represent.

Francisca Acevedo, Elleli Huerta, Caroline Burgeff, Patricia Koleff & José Sarukhán

Comisión Nacional para el Conocimiento y Uso de la Biodiversidad, México D.F., México.
e-mail: facevedo@conabio.gob.mx

1. Anderson, E. & Cutler, H.C. *Ann. Mo. Bot. Gard.* **29**, 69–86 (1942).
2. Kato, T.A., Mapes, L.M., Mera, L.M., Serratos, J.A. & Bye, R.A. *Origen y Diversificación del Maíz: una Revisión Analítica* (Universidad Autónoma de México, Comisión Nacional para el Conocimiento y Uso de la Biodiversidad, México, D.F. (2009).
3. Bellón, M.R. & Brush, S.B. *Econ. Bot.* **48**, 196–209 (1994).
4. Bellón, M.R. *et al. Diversidad y Conservación de Recursos Genéticos en Plantas Cultivadas, en Capital Natural de México, vol. II: Estado de la Conservación y Tendencias de Cambio.* (CONABIO, México, 2009).
5. Bourges, H. in *La Alimentación de los Mexicanos* (eds. Alarcón-Segovia, D. & Bourges, H.) 97–134 (El Colegio Nacional, México D.F., 2002).
6. Hernández-Xolocotzi, E. *Econ. Bot.* **39**, 416–430 (1985).
7. Pressoir, G. & Berthaud, J. *Heredity* **92**, 88–94 (2004).
8. Dalton, R. *Nature* **462**, 404 (2009).
9. Acevedo, F. *Nat. Biotechnol.* **22**, 803 (2004).
10. Acevedo G.F. *et al. La Bioseguridad en México y los Organismos Genéticamente Modificados: Como Enfrentar un Nuevo Desafío, en Capital Natural de México, vol. II: Estado de la Conservación y Tendencias de Cambio* (CONABIO, México, 2009).

Integrative genomics viewer

To the Editor:

Rapid improvements in sequencing and array-based platforms are resulting in a flood of diverse genome-wide data, including data from exome and whole-genome sequencing, epigenetic surveys, expression profiling of coding and noncoding RNAs, single nucleotide polymorphism (SNP) and copy number profiling, and functional assays. Analysis of these large, diverse data sets holds the promise of a more comprehensive understanding of the genome and its relation to human disease. Experienced and knowledgeable human review is an essential component of this process, complementing computational approaches. This calls for efficient and intuitive visualization tools able to scale to very large data sets and to flexibly integrate multiple data types, including clinical data. However, the sheer volume and scope of data pose a significant challenge to the development of such tools.

To address this challenge, we have developed the Integrative Genomics Viewer (IGV), a lightweight visualization tool that enables intuitive real-time exploration of diverse, large-scale genomic data sets on standard desktop computers. It supports flexible integration of a wide range of genomic data types including aligned sequence reads, mutations, copy number, RNA interference screens, gene expression, methylation and genomic annotations (**Supplementary Fig. 1**). The IGV makes use of efficient, multi-resolution file formats to enable real-time exploration of arbitrarily large data sets over all resolution scales, while consuming minimal resources on the client computer (**Supplementary Notes**). Navigation through a data set is similar to that of Google Maps, allowing the user to zoom and pan seamlessly across the genome at any level of detail from whole genome to base pair (**Supplementary Fig. 2**). Data sets can be loaded from local or remote sources, including cloud-based resources, enabling investigators to view their own genomic data sets alongside publicly available data from, for example, The Cancer Genome Atlas[1], 1000 Genomes[2] (http://www.1000genomes.org/) and ENCODE[3] (http://www.genome.gov/10005107) projects. In addition, IGV allows collaborators to load and share data locally or remotely over the internet.

IGV supports concurrent visualization of diverse data types across hundreds,

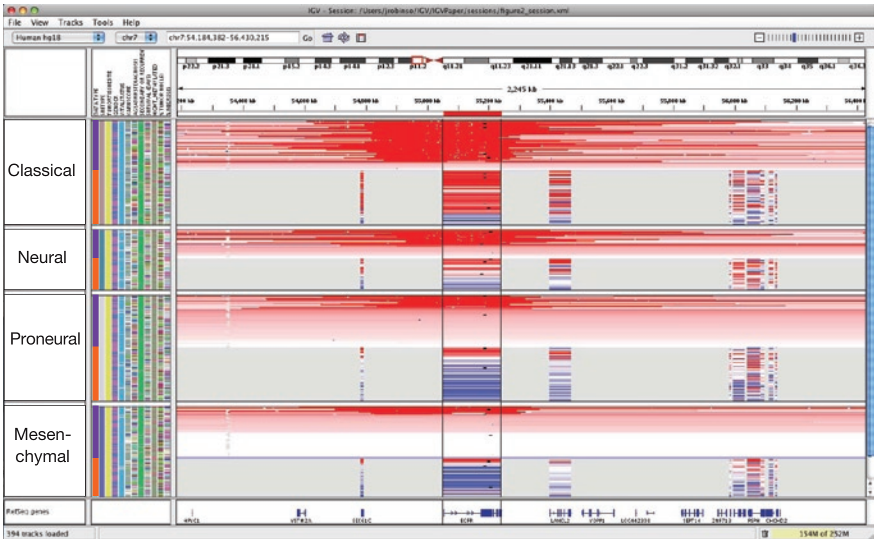


Figure 1 Copy number, expression and mutation data grouped by tumor subtype. This figure illustrates an integrated, multi-modal view of 202 glioblastoma multiforme samples from The Cancer Genome Atlas (TCGA). Copy number data are segmented values from Affymetrix (Santa Clara, CA, USA) SNP6.0 arrays. Expression data are limited to genes represented on all TCGA-employed platforms and displayed across the entire gene locus. Red shading indicates relative upregulation of a gene and the degree of copy gain of a region; blue shading indicates relative downregulation and copy loss. Small black squares indicate the position of point missense mutations. Samples are grouped by tumor subtype (2nd annotation column) and data type (1st sample annotation column) and sorted by copy number of the EGFR locus. Linking by sample attributes ensures that the order of sample tracks is consistent across data types within their respective tumor subtypes.

introduction of transgenic maize developed for pharmaceutical or other non-food purposes, and its impact on landraces[9,10].

Mexico needs to be able to define what kind of transgenic materials (for maize and any other relevant crop) it needs for its ecological, social and economic requirements. This responsibility must be carefully analyzed in order to provide farmers with adequate and necessary elements to help achieve a level of food security for the present and future of Mexican society, while conserving genetic diversity and helping develop adequately the social structures of the rural economy and society.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

DISCLAIMER
The manuscript reflects only the opinion of the authors and not the institution they represent.

Francisca Acevedo, Elleli Huerta, Caroline Burgeff, Patricia Koleff & José Sarukhán

Comisión Nacional para el Conocimiento y Uso de la Biodiversidad, México D.F., México.
e-mail: facevedo@conabio.gob.mx

1. Anderson, E. & Cutler, H.C. *Ann. Mo. Bot. Gard.* **29**, 69–86 (1942).
2. Kato, T.A., Mapes, L.M., Mera, L.M., Serratos, J.A. & Bye, R.A. *Origen y Diversificación del Maíz: una Revisión Analítica* (Universidad Autónoma de México, Comisión Nacional para el Conocimiento y Uso de la Biodiversidad, México, D.F. (2009).
3. Bellón, M.R. & Brush, S.B. *Econ. Bot.* **48**, 196–209 (1994).
4. Bellón, M.R. *et al. Diversidad y Conservación de Recursos Genéticos en Plantas Cultivadas, en Capital Natural de México, vol. II: Estado de la Conservación y Tendencias de Cambio.* (CONABIO, México, 2009).
5. Bourges, H. in *La Alimentación de los Mexicanos* (eds. Alarcón-Segovia, D. & Bourges, H.) 97–134 (El Colegio Nacional, México D.F., 2002).
6. Hernández-Xolocotzi, E. *Econ. Bot.* **39**, 416–430 (1985).
7. Pressoir, G. & Berthaud, J. *Heredity* **92**, 88–94 (2004).
8. Dalton, R. *Nature* **462**, 404 (2009).
9. Acevedo, F. *Nat. Biotechnol.* **22**, 803 (2004).
10. Acevedo G.F. *et al. La Bioseguridad en México y los Organismos Genéticamente Modificados: Como Enfrentar un Nuevo Desafío, en Capital Natural de México, vol. II: Estado de la Conservación y Tendencias de Cambio* (CONABIO, México, 2009).

Integrative genomics viewer

To the Editor:
Rapid improvements in sequencing and array-based platforms are resulting in a flood of diverse genome-wide data, including data from exome and whole-genome sequencing, epigenetic surveys, expression profiling of coding and noncoding RNAs, single nucleotide polymorphism (SNP) and copy number profiling, and functional assays. Analysis of these large, diverse data sets holds the promise of a more comprehensive understanding of the genome and its relation to human disease. Experienced and knowledgeable human review is an essential component of this process, complementing computational approaches. This calls for efficient and intuitive visualization tools able to scale to very large data sets and to flexibly integrate multiple data types, including clinical data. However, the sheer volume and scope of data pose a significant challenge to the development of such tools.

To address this challenge, we have developed the Integrative Genomics Viewer (IGV), a lightweight visualization tool that enables intuitive real-time exploration of diverse, large-scale genomic data sets on standard desktop computers. It supports flexible integration of a wide range of genomic data types including aligned sequence reads, mutations, copy number, RNA interference screens, gene expression, methylation and genomic annotations (**Supplementary Fig. 1**). The IGV makes use of efficient, multi-resolution file formats to enable real-time exploration of arbitrarily large data sets over all resolution scales, while consuming minimal resources on the client computer (**Supplementary Notes**). Navigation through a data set is similar to that of Google Maps, allowing the user to zoom and pan seamlessly across the genome at any level of detail from whole genome to base pair (**Supplementary Fig. 2**). Data sets can be loaded from local or remote sources, including cloud-based resources, enabling investigators to view their own genomic data sets alongside publicly available data from, for example, The Cancer Genome Atlas[1], 1000 Genomes[2] (http://www.1000genomes.org/) and ENCODE[3] (http://www.genome.gov/10005107) projects. In addition, IGV allows collaborators to load and share data locally or remotely over the internet.

IGV supports concurrent visualization of diverse data types across hundreds,

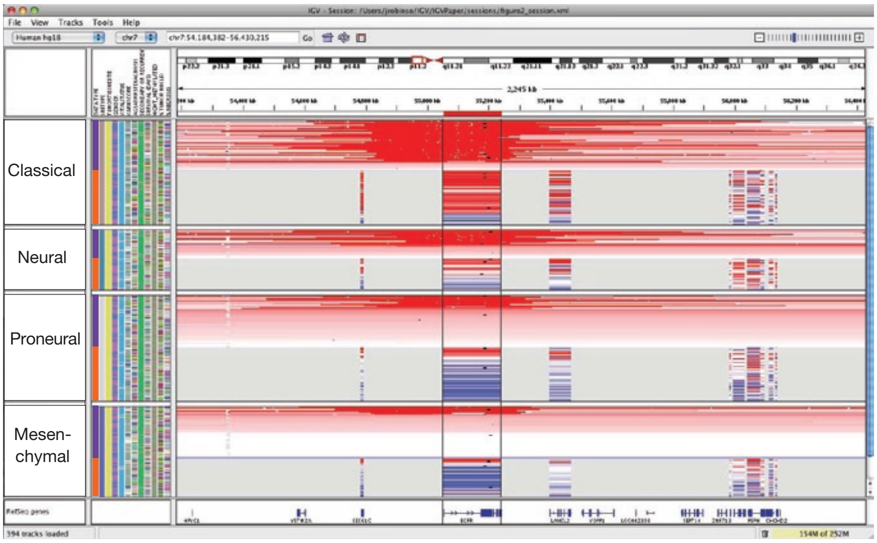

IGV - Session: /Users/jrobinso/IGV/IGVPaper/sessions/figure2_session.xml
File View Tracks Tools Help
Human hg18
chr7
chr7:54,184,382-56,430,215
Go
2,245 kb
Classical
Neural
Proneural
Mesen-
chymal
RefSeq genes
EGFR
394 tracks loaded

Figure 1 Copy number, expression and mutation data grouped by tumor subtype. This figure illustrates an integrated, multi-modal view of 202 glioblastoma multiforme samples from The Cancer Genome Atlas (TCGA). Copy number data are segmented values from Affymetrix (Santa Clara, CA, USA) SNP6.0 arrays. Expression data are limited to genes represented on all TCGA-employed platforms and displayed across the entire gene locus. Red shading indicates relative upregulation of a gene and the degree of copy gain of a region; blue shading indicates relative downregulation and copy loss. Small black squares indicate the position of point missense mutations. Samples are grouped by tumor subtype (2nd annotation column) and data type (1st sample annotation column) and sorted by copy number of the EGFR locus. Linking by sample attributes ensures that the order of sample tracks is consistent across data types within their respective tumor subtypes.

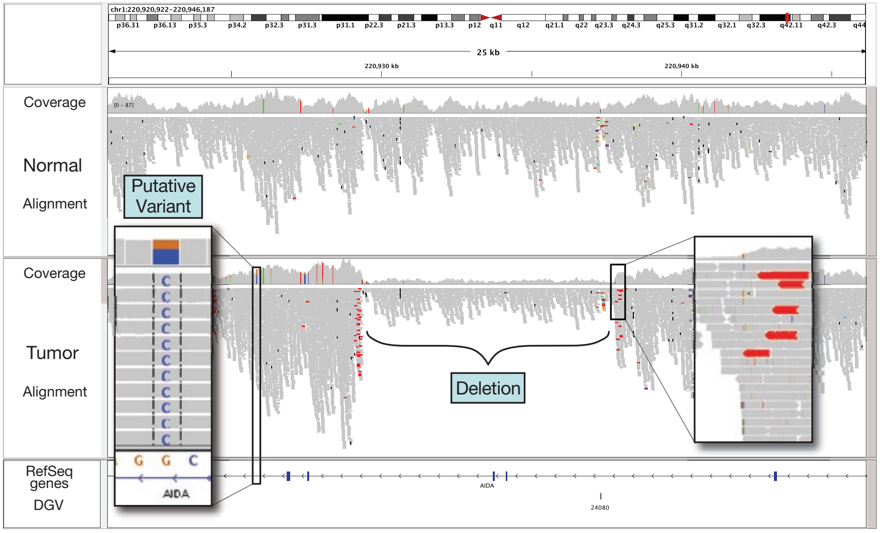


Figure 2 View of aligned reads at 20-kb resolution. Coverage plot and alignments from paired-end reads for a matched tumor/normal pair. Sequencing was performed on an Illumina (San Diego, CA) GA2 platform and aligned with Maq (http://maq.sourceforge.net/). Alignments are represented as gray polygons with reads mismatching the reference indicated by color. Loci with a large percentage of mismatches relative to the reference are flagged in the coverage plot as color-coded bars. Alignments with unexpected inferred insert sizes are indicated by color. There is evidence for a ~10-kb deletion (removing two exons of AIDA) in the tumor sample not present in the normal.

and up to thousands, of samples and correlation of these integrated data sets with clinical and phenotypic variables. A researcher can define arbitrary sample annotations and associate them with data tracks using a simple tab-delimited file format (**Supplementary Notes**). These might include, for example, sample identifier (used to link different types of data for the same patient or tissue sample), phenotype, outcome, cluster membership or any other clinical or experimental label. Annotations are displayed as a heatmap, but more importantly are used for grouping, sorting, filtering and overlaying diverse data types to yield a comprehensive picture of the integrated data set. This is illustrated in **Figure 1**, a view of copy number, expression, mutation and clinical data from 202 glioblastoma samples from The Cancer Genome Atlas project in a 3-kb region around the epidermal growth factor receptor (*EGFR*) locus[1,4]. The investigator first grouped samples by tumor subtype, then by data type (copy number and expression), and finally sorted them by median copy number over the *EGFR* locus. A shared sample identifier links the copy number and expression tracks, maintaining their relative sort order within the subtypes. Mutation data are overlaid on corresponding copy number and expression tracks, based on shared participant identifier annotations. Several trends in the data stand out, such as a strong correlation between copy number and expression and an overrepresentation of EGFR-amplified samples in the 'Classical' subtype.

IGV's scalable architecture makes it well suited for genome-wide exploration of next-generation sequencing (NGS) data sets, including both basic aligned read data as well as derived results, such as read coverage. NGS data sets can approach terabytes in size, so careful management of data is necessary to conserve computer resources and to prevent information overload. IGV varies the displayed level of detail according to resolution scale. At very wide views, such as the whole genome, IGV represents NGS data by a simple coverage plot. Coverage data are often useful for assessing overall quality and diagnosing technical issues in sequencing runs (**Supplementary Fig. 3**), as well as analysis of ChIP-Seq[5] and RNA-Seq[6] experiments (**Supplementary Figs. 4** and **5**).

As the user zooms below the ~50 kb range, individual aligned reads become visible (**Fig. 2**), and putative SNPs are highlighted as allele counts in the coverage plot. Alignment details for each read are available in popup windows (**Supplementary Figs. 6** and **7**). Zooming in further, individual base mismatches become visible, highlighted by color and intensity according to base call and quality. At this level, the investigator may sort reads by base, quality, strand, sample and other attributes to assess the evidence of a variant. This type of visual inspection can be an efficient and powerful tool for variant call validation, eliminating many false positives and aiding in confirmation of true findings (**Supplementary Figs. 6** and **7**).

Many sequencing protocols produce reads from both ends ('paired ends') of genomic fragments of known size distribution. IGV uses this information to color-code paired ends if their insert sizes are larger than expected, fall on different chromosomes or have unexpected pair orientations. Such pairs, when consistent across multiple reads, can be indicative of a genomic rearrangement. When coloring aberrant paired ends, each chromosome is assigned a unique color, so that intra- (same color) and inter- (different color) chromosomal events are readily distinguished (**Fig. 2** and **Supplementary Fig. 8**). We note that misalignments, particularly in repeat regions, can also yield unexpected insert sizes and can be diagnosed with the IGV (**Supplementary Fig. 9**).

There are a number of stand-alone, desktop genome browsers available today[7], including Artemis[8], EagleView[9], MapView[10], Tablet[11], Savant[12], Apollo[13] and the Integrated Genome Browser[14]. Many of them have features that overlap with IGV, particularly for NGS sequence alignment and genome annotation viewing. The Integrated Genome Browser also supports viewing array-based data (**Supplementary Table 1** and **Supplementary Notes**). IGV focuses on the emerging integrative nature of genomic studies, placing equal emphasis on array-based platforms, such as expression and copy-number arrays, NGS, as well as clinical and other sample metadata. Indeed, an important and unique feature of IGV is the ability to view all these different data types together and to use the sample metadata to dynamically group, sort and filter data sets (**Fig. 1**). Another important characteristic of IGV is fast data loading and real-time pan and zoom—at all scales of genome resolution and all data set sizes, including data sets comprising hundreds of samples. Finally, we have placed great emphasis on the ease of installation and use of IGV, with the goal of making both the viewing and sharing of their data accessible to end users who are not informatics specialists.

IGV is open source software and freely available (http://www.broadinstitute.org/igv/), including full documentation on use of the software.

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

ACKNOWLEDGMENTS
Thanks to M.-D. Nazaire, IGV software developer, J. Boehm, M. DePristo, M. Garber, M. Reich and B. Weir, who worked closely with the development team with suggestions for features and usability, and J. Bistline for expert assistance with figures. M. Garber provided source code for multiple alignment tracks. This work was supported by The National Institute of General Medical Sciences (R01GM074024), National Cancer Institute (R21CA135827) and National Human Genome Research Institute (U54HG003067).

AUTHOR CONTRIBUTIONS

J.T.R. and H.T. designed and developed the software; W.W., M.G., E.S.L., G.G. and J.P.M. contributed to the design of the interface and data views; J.P.M. and G.G. oversaw the project; and J.T.R., H.T., W.W., G.G. and J.P.M. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

James T Robinson[1], Helga Thorvaldsdóttir[1], Wendy Winckler[1], Mitchell Guttman[1,2], Eric S Lander[1–3], Gad Getz[1] & Jill P Mesirov[1]

[1]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts, USA. [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA.
e-mail: mesirov@broad.mit.edu; jrobinso@broadinstitute.org

1. Cancer Genome Atlas Research Network. *Nature* **455**, 1061–1068 (2008).
2. Durbin, R.M. *et al. Nature* **467**, 1061–1073 (2010).
3. The ENCODE Project Consortium. *Science* **306**, 636–640 (2004).
4. Verhaak, R.G. *et al. Cancer Cell* **17**, 98–110 (2010).
5. Guttman, M. *et al. Nature* **458**, 223–227 (2009).
6. Berger, M.F. *et al. Genome Res.* **20**, 413–427 (2010).
7. Nielsen, C., Cantor, M., Dubchak, I., Gordon, D. & Wang, T. *Nat. Methods* **7**, S5–S15 (2010).
8. Rutherford, K. *et al. Bioinformatics* **16**, 944–945 (2000).
9. Huang, W. & Marth, G. *Genome Res.* **18**, 1538–1543 (2008).
10. Bao, H. *et al. Bioinformatics* **25**, 1554–1555 (2009).
11. Milne, I. *et al. Bioinformatics* **26**, 401–402 (2010).
12. Fiume, M., Williams, V., Brook, A. & Brudno, M. *Bioinformatics* **26**, 1938–1944 (2010).
13. Lewis, S.E. *et al. Genome Biol.* **3**, RESEARCH0082.1–0082.14 (2002).
14. Nicol, J.W., Helt, G.A., Blanchard, S.G. Jr., Raja, A. & Loraine, A.E. *Bioinformatics* **25**, 2730–2731 (2009).

The problems with today's pharmaceutical business—an outsider's view

Mark Kessel

The pharmaceutical industry must devote greater resources, investment and effort to address its anemic drug pipeline in the long term, rather than focusing on its bottom line in the near term.

Is there any doubt that the leading drug companies are in desperate need of reinvention? Blockbuster drugs are coming off patent or being taken off the market for safety reasons and there are no replacement drugs on the horizon to make up the shortfall in profits. Furthermore, healthcare reform is likely to exacerbate the flaws in big pharma's traditional business model by imposing pay for performance, as is already the case in Europe. To state the obvious, over the past decade, the pharmaceutical industry has brought few drugs to market from its own development efforts. Commentators have stressed, and heads of big pharma have acknowledged, that the sector's R&D efforts need to be drastically changed. But alteration of the industry's culture and lumbering decision-making process will be slow and challenging and will require bold leadership.

Recognizing that the R&D engine cannot be repaired rapidly to fuel growth, big pharma has taken several steps in dealing with its diminished R&D productivity. First, it has looked to expand its markets geographically into developing countries; second, it has increased its emphasis on generic drugs and biosimilars; and finally, it has sought to diversify by migrating into new product categories.

Although the foregoing steps will lessen the projected shortfall in revenues associated with the expiration of patents in the coming years, the achievement of sustained growth in big pharma will of necessity depend to a large extent on another factor—its ability to increase the productivity of internal R&D efforts, while at the same time bolstering the pipeline with drugs acquired from the biotech sector. It is clear that from its internal productivity alone big pharma is unlikely to achieve the growth needed to fuel revenues. Successful implementation of this pipeline strategy will require the management at each company to optimize its current internal R&D efforts and its approach to acquiring drugs. In this article, I discuss some critical steps that could be implemented by pharmaceutical companies to better accomplish this strategy.

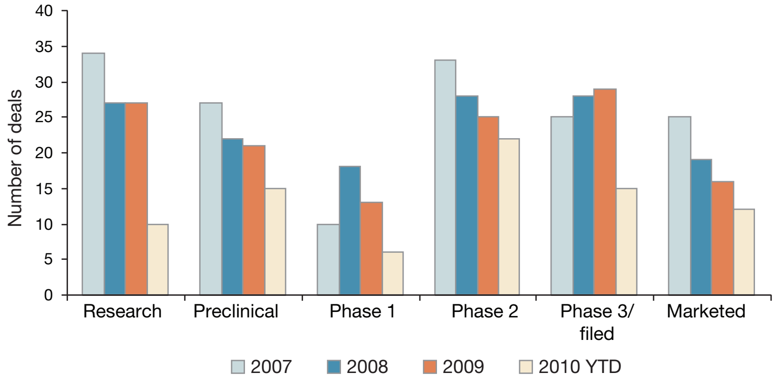


Figure 1 The number of big pharma deals with biotech have fallen in all stages. Source: Burrill & Co. (San Francisco); 2010 year to date (YTD) is through September 30.

A sector in crisis

The traditional business model at big pharma relies on (i) identifying promising new blockbuster drugs; (ii) conducting large, expensive clinical trials; and (iii) if successful, promoting the drugs with extensive marketing and sales presence in developed countries. Clearly, the traditional model cannot be sustained in today's environment. Internally developed pipeline productivity at big pharma has decreased significantly, averaging only about one new molecular entity a year per company. But the cost of bringing a new drug to market has continued to rise and is now estimated to exceed $1 billion. Even though the industry cites scientific breakthroughs, the timeline for developing a drug and getting it to market has not declined and can take as long as 15 years.

A major reason that big pharma must limit the number of compounds it introduces into its pipeline is that spending on R&D places great pressure on earnings. The public equity markets relentlessly focus on short-term performance and unduly punish companies that do not meet quarterly revenue and earnings expectations. It has been reported that analyst expectations for the industry are so diminished that they are now hoping that the pharmaceutical industry as a whole will reach a compounded annual growth rate of 1% of revenues over the next five years.

The loss of patents on blockbusters by big pharma is a major concern. In the next five years, of the top 10 best-selling drugs in the world, 9 will go off patent, and of the top 20,

Mark Kessel is partner at Symphony Capital, New York, New York, USA.
e-mail: mark@symphonycapital.com

18 will lose patent protection. As a result, ~$100 billion of sales will be lost during this period. This number may be understated, given the recent safety issues associated with some blockbuster drugs, such as GlaxoSmithKline's (GSK; Brentford, UK) diabetes drug Avandia (rosiglitazone). To compensate for these losses, big pharma has resorted to buying revenues by means of acquisitions to replace declining sales. At the same time, sales of existing drugs are less likely to benefit from direct-to-consumer advertising. Indeed, direct-to-consumer advertising will continue to garner greater scrutiny from regulators and have less of a favorable impact on sales of new and existing products.

The pressure asserted by generics is causing an ever-steeper decline in returns on marketing and sales on drugs coming off patent than was the case in the past. It has not gone unnoticed by big pharma that generics sales have outpaced sales of the pharmaceutical industry over the past ten years. This has been driven by an increase of demand, the expiration of patents and cost constraints imposed by governments and third-party payers. The expectation is that this trend will continue into the future. With the patent cliff looming, generics will have many small-molecule blockbusters to target. Although one of the anticipated benefits of healthcare reform for big pharma will be expanded coverage, pay for performance will be an increasing issue. This legislation will put added pressure on product pricing from government and third-party payers. GSK recently has reported a drop in profits, which it attributed to US healthcare reform and European government 'austerity' measures that have had an impact on the drug industry[1]. Can big pharma expect its drugs to garner premium pricing without showing a benefit over cheaper alternative therapies? Pay for performance is already the case in Europe. For example, the UK's National Institute for Clinical Excellence (NICE; London) rebuffed an expansion of Roche's (Basel) Tarceva (erlotinib) for an additional indication, having determined that it was not a cost-effective use of resources[2]. As a precursor of where the United States is going, consider that Medicare officials are already considering whether the government program should cover Dendreon's (Seattle) new prostate cancer dendritic cell vaccine costing $93,000 per patient. Another indication of the changing US landscape is the recent pronouncements from the US Food and Drug Administration (FDA) and the Centers for Medicare and Medicaid Services (CMS); these agencies have issued a proposal to allow drug companies to voluntarily request that the FDA and CMS conduct parallel reviews for marketing approval. Although the rules are supposedly not intended to change the approval standards, but rather to benefit sponsors, critics have asserted that this will shift the agencies' focus from efficacy and safety to comparative effectiveness.

Technology will make regulators and third-party payers better equipped to measure what benefits patients are deriving from the drugs. The net effect is that governments and payers will continue to bear down on prices, access, utilization and prescribing patterns.

In addition, the pharmaceutical sector is going to be faced with a more stringent regulatory pathway for approval of new drugs, as well as closer government scrutiny of the continued marketing of existing drugs. There is little doubt that the regulators are going to focus increasingly on patient safety and benefits when bringing new drugs to market[3]. The recent restrictions placed on GSK's Avandia because of data indicating an association with heart toxicity points in this direction.

The manner in which big pharma is perceived in political circles will also have an impact on its future prospects. The US Congress portrays the industry as insensitive to consumer safety. Indeed, the Obama Administration publicly vilified big pharma as part of its health reform initiative (while simultaneously courting its participation in providing funds to close the so-called donut hole, a coverage gap in the 2003 Medicare Part D health plan for prescription drugs).

Regulatory halting of sales of therapeutics for safety reasons, poorly handled product recalls and the imposition of unprecedented criminal and civil fines (reaching $2.3 billion in Pfizer's case), coupled with calls for CEOs to serve jail time for illegal drug promotion, settlements relating to bilking healthcare programs by inflating drug prices and investigations of paying bribes to boost sales and the development and marketing of drugs have also added to the public's wariness of the sector. The net effect of a plummeting reputation—down in some surveys as low as the tobacco and oil industries—has been to hurt the industry across numerous constituencies that have a bearing on the prospects of its products, including governments, regulators and consumers. For these reasons, the importance of disassociation and delineation from big pharma has not been lost on the biotech industry.

How is big pharma responding?

Big pharma is increasingly coming to recognize the shortcomings of its traditional business model. Pharma has not been good at identifying early molecules likely to succeed; its discovery and research productivity is wanting; it lacks an innovative culture; and it has not effectively captured external breakthroughs. On top of this, it needs to reduce its cost structure to maintain earnings (only a short-term panacea). So how is big pharma management dealing with these structural shortcomings? It is taking the more expedient approach of solving the near-term revenue shortfall while failing to aggressively address the longer-term problem—its anemic drug pipeline. This myopic focus is translating into several business strategies.

Buy revenues. One old habit that the pharmaceutical industry is finding hard to abandon is the quick fix to diminishing revenues—buy earnings. But this is only postponing the problem. In 2009, Pfizer (New York) acquired Wyeth (Madison, NJ, USA) for $68 billion and in October it announced the purchase of King Pharmaceuticals (Bristol, TN, USA) for $3.6 billion. The King acquisition will add a mere $0.02 to Pfizer's earnings per share over the next two years and not much more in the years to come. Following this path, Merck (Whitehouse Station, NJ, USA) acquired Schering-Plough (Madison, NJ, USA), and now Sanofi-aventis (Paris) is seeking to acquire Genzyme (Cambridge, MA, USA). But the market and security analysts see these transactions for what they are. The Wyeth acquisition was viewed as merely giving Pfizer a year's worth of breathing room as investors were expected to have difficulty in comparing financial results to Pfizer's year-ago figures. Another take on this acquisition was that although Wyeth presented Pfizer with an attractive biologics platform and some complementary products and businesses, it is not enough to sustain long-term revenue growth. Longer-term growth was viewed as dependent on the success of Pfizer's future drug development efforts. Lilly has stated that it is going to take a different path to dealing with its pipeline issues by limiting itself to small acquisitions rather than large-scale combinations.

Expand outside the United States. The global pharmaceutical market outside the United States is projected to grow more dramatically than in the United States owing to the growth of a substantial middle class in emerging economies. The key countries are likely to be China and Brazil, followed by India and Russia. China, for example, is expected to become the third largest market in 2011—up from eighth in 2006 and having increased by 27% in 2009 alone.

As these markets will be driven by generics, margins are likely to be lower than in the United States but higher than may be expected—more likely closer to those prevailing in Europe, as sales there, given Europe's economic environment, are likely to continue to experience volume and pricing pressures.

Given the projected growth for pharmaceuticals outside the United States, the pharmaceutical industry is acquiring a presence in these markets. Some companies have already started to pursue these markets aggressively. For example, at a recent conference[4], Hanspeter Spek, president of global operations at Sanofi-aventis, has indicated that his company ranks number one in emerging economies, which represent a key growth platform with fast-growing contributions coming from key markets, such as China and Brazil[4]. Sanofi has also acquired Zentiva in Hlohovec, Slovakia; Medley in Campinas, Brazil; and Kendrick in Ciudad de Mexico, Mexico. Similarly, Pfizer has participated in transactions with Indian (Aurobindo in Hyderabad, Claris Lifesciences in Ahmedabad, and Biocon in Bangalore) and Brazilian (Labóratorio Tueto Brasileiro in Anapolis) companies to gain access to these emerging markets, as well as to the local companies' branded and unbranded generics. Elsewhere, GSK is purchasing Nanjing MeiRui Pharma (Shanghai, China) and has partnered with Aspen Pharmacare (Durban, South Africa); and Abbott Pharmaceuticals (Deerfield, IL, USA) has purchased Solvay (Brussels) as both a geographic expansion into Europe and a foray into emerging markets. Merck recently announced it is looking to expand in India and other emerging markets through acquisitions and partnerships, seeking to become India's top or second largest pharmaceutical company by 2015.

Generics. Given the anticipated growth of generic sales, especially in the developing markets, pharmaceutical companies, such as Novartis (Basel) and Sanofi-aventis, have also placed an emphasis on building or purchasing generic operations. As mentioned above, Pfizer is partnering with Biocon of India, in part for its generics business. Similarly, Tokyo-based Daiichi Sankyo, the second-largest Japanese drug company, acquired the leading Indian company Ranbaxy (Gurgaon) to establish a global presence in the generics market.

Brand biologics. In the area of biologics, big pharma is recognizing the attractiveness of these franchises. Some of the allure relates to biologics having the potential for shorter development timelines than small molecules; what's more, brand biologics are likely to retain market dominance, even after expiration of their patents, due to high barriers to entry for competitors wishing to produce follow-on products or biosimilars. For Roche, biologics in 2009 are estimated to have accounted for 54% of its sales and are expected to rise to 59% by 2014. Indeed, the acquisition of biologic franchises and/or businesses has proven attractive across big pharma in recent years. Thus, for example, Sanofi-aventis is bidding over $18 billion to acquire Genzyme for its biological businesses, which even in niche diseases have been able to capture significant revenues and profits.

Orphan drugs. Indeed, big pharma is paying increased attention to the orphan drug market. These drugs, which target chronic, degenerative and other rare life-threatening diseases, have been shown by Genzyme and other biotech companies as having the potential to be highly profitable. The cost of treatments can run as much as $500,000 per patient per year. Compared with drugs for other indications, the development costs and regulatory hurdles for orphan drugs tend to be lower, whereas the margins are higher and the markets continue to expand. As a whole, markets for orphan diseases are expected to reach nearly $82 billion in 2011, up from about $59 billion in 2006. No wonder then that several pharmaceutical companies have entered this market, and some, like Pfizer, have even created specific business units focused exclusively on orphan diseases to help make up for the impending loss of revenues associated with their other drugs facing patent expiry[5].

Entering these markets will not be without major challenges, however. Most pharmaceutical companies are novices in the orphan drug field and will have to build drug development expertise for these indications and marketing infrastructure for niche populations. Given the plummeting reputation of big pharma as a whole, a major issue will be how to manage the image fallout from the high prices that will need to be charged to make such drugs commercially attractive.

Biosimilars. Another area that is witnessing intense investment from large pharmaceutical companies is biosimilars or follow-on biologics. Merck has announced it will invest over $1.5 billion in biosimilars by 2015; and Pfizer's partnership with Biocon gives Pfizer access to Biocon's biosimilar versions of insulin.

The pharmaceutical industry (and to an extent some biotech companies) recognizes an opportunity in follow-on biologics because the development risk and costs associated in manufacturing such products are likely to be significant barriers to entry. Such limited competition should make the market more attractive as it is not going to result in extensive pricing pressures as is the case in the generics market. Generics giant Teva Pharmaceuticals (Tel Aviv, Israel) has projected that ~$53 billion in branded biologics revenues will be exposed to biosimilars competition by 2015 based on patent expirations alone.

In Europe, where the market is still in its early stage, biosimilar products need to overcome concerns about comparative efficacy and safety. Another reason cited for the limited uptake relates to the national rules that prevent automatic substitution. In the United States, despite its description in the recent health reform legislation, there is still no regulatory pathway for the approval of biosimilars.

The challenge of demonstrating comparability, high development costs, coupled with a difficult market environment may cause companies, in the right circumstances, to seek to develop improved versions of the original drugs—so-called biobetters—rather than go down the path of developing biosimilars. And even though the net effect of an increasing number of biosimilars and biobetters will, like other drugs, create competition in terms of safety, efficacy and pricing, the pharma industry anticipates that governments and third-party payers will implement policies to foster the use of biosimilars in an effort to restrain costs associated with the reimbursement of expensive biologics[6].

Product diversification. Another strategy increasingly embraced by big pharma is to migrate into nonprescription health products for humans as well as animals and even reagent and supply service companies. Thus, Novartis recently moved into eye care products through its acquisition of Alcon (Hünenberg, Switzerland); GSK has augmented its business in consumer healthcare by acquiring Stiefel (Coral Gables, FL, USA) for its dermatology products; and Sanofi-aventis has acquired Chattem, an over-the-counter company based in Chattanooga, Tennessee. In animal health, Sanofi-aventis and Merck have expanded investment in Merial (Duluth, GA, USA), their animal health joint venture. And in the past year, Merck opted to buy Millipore (Billerica, MA, USA) for its US laboratory supply manufacturing operations.

The nutraceuticals sector—a market that the major food companies already target—is also garnering increasing attention from big pharma. These clinical nutritional products are developed to target specific diseases, ranging anywhere from diabetes to Alzheimer's disease, and their success is more dependent on physician recommendations than on shelf-space competition in supermarkets.

All the above businesses, compared with proprietary drugs, provide more stable long-term income flows. At the same time, however, such markets can be smaller, carry lower margins and face greater competition. Furthermore, it remains unclear whether big pharma has the

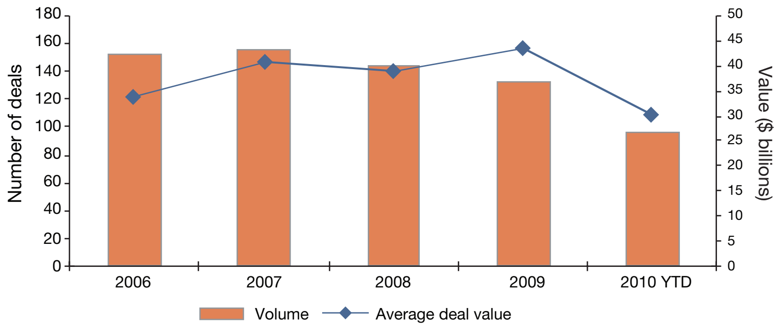


Figure 2 The number and value of biotech acquisitions by pharma have fallen. Source: Burrill; 2010 year to date (YTD) is through September 30.

requisite expertise to successfully exploit the benefits associated with these markets.

All these strategies attempt to address the projected shortfall in revenues through diversification across products and geography. Even so, they continue to ignore the critical problem—big pharma's drug pipeline, which remains anemic despite the commitment of vast resources to R&D.

The real problem

Decision making at large pharmaceutical companies is notorious for stymieing innovation, risk taking and long-term business development efforts. Although big pharma acknowledges the detrimental impact of its bureaucratic ways, it has not proven successful in becoming nimble like companies in the nascent biotech sector.

Examples of R&D inefficiency in the pharmaceutical industry are legion. As John Lechleiter, the CEO of Eli Lilly (Indianapolis), put it, companies are "taking too long, spending too much and producing far too little. Re-powering pharmaceutical innovation is an urgent need"[7]. Others have echoed his lament. Jeffrey Kindler, who recently announced he was stepping down as CEO of Pfizer, has gone on the record to acknowledge the detrimental impact of having 14 management layers between him and the bench scientists. His answer to this problem was to build the company's pipeline using a diversity of approaches that recognizes the increased need for flexibility in R&D spending and the dangers of sticking to the traditional industry research spending formula.

Another response to this earnings crisis has been to slaughter the R&D sacred cow. The approach here is to prune and streamline R&D so that the pipeline continues to advance while earnings are enhanced. In the past, marketing, sales and other costs were trimmed when big pharma encountered earnings pressure. This is no longer the case, with massive cuts in R&D personnel across the industry. According to recruitment specialists Challenger, Gray & Christmas (Chicago, IL, USA), the number of pharmaceutical industry jobs eliminated in the first ten months of 2010 was 45,263 (in 2009, it was 58,696 jobs). Since its merger with Wyeth, Pfizer has only four major R&D sites—down from 20 sites around the globe at the time of the acquisition in 2009—with most researchers in each laboratory focused on the same disease.

Another approach to R&D restructuring has been to 'spin out' scientists, products and patents. This can take several forms: creating entirely separate startup companies (e.g., GSK's Convergence Pharmaceuticals of Cambridge, UK) while retaining minority equity, with the potential for a return on investment if development of the spun-out programs progresses; out-licensing internally developed drugs to venture capital firms; divesting clinical services to contract research firms (e.g., Sanofi-aventis's deal with Covance of Princeton, New Jersey); or combining development and commercial efforts (e.g., GSK and Pfizer aggregating their HIV ventures to synergize their reach or Sanofi and Merck forming a joint venture for animal medicines). But rather than promoting a more efficient paradigm for furthering drug development, all of these measures are designed to reduce internal R&D costs, thereby increasing earnings in the short term.

The simple fact is that relying on acquiring drug candidates is also not likely to solve the problem. Given the absence of an initial public offering market and the decrease in capital available from other sources, one would assume that there would be a wealth of products in biotech companies desperate to collaborate with big pharma. However, big pharma deems only a limited number of products in the biotech sector as attractive opportunities (**Fig. 1**), suggesting that the number of transactions for coveted late-stage compounds is likely to be constrained. A recent study[8] by consultants Bain & Company (Boston) states that of 6,000 biotech projects available for late-stage licensing in 2009, only ~200 were likely to be attractive to large pharmaceutical companies and of these, fewer than 100 were potentially top-selling candidates. Taken together, they would account for only $30 billion in potential revenue—tens of billions below the shortfall in earnings due to patent expirations that pharma is facing. Why then are so many biotech development programs unattractive to pharma?

Clearly, big pharma still finds late-stage, phase 3, potential blockbuster compounds most alluring, and in the current environment it is also enamored with drugs that target smaller patient populations that fit its therapeutic areas and have high commercial potential. However, these drugs have to be able to deliver true advancements in meeting patient needs (to have a better chance in passing regulatory muster and to increase payer adoption).

Another factor limiting the pool is a mismatch of the drugs that the biotech industry is developing and the ones that big pharma finds attractive. Programs in the biotech sector can become highly misaligned with those in big pharma when the latter suddenly changes its product strategy, for example, when a particular type of therapeutic modality comes into vogue. For example, big pharma has been acquiring many of the first-generation biotech antibody pioneers while many other companies that have small-molecule programs have been left on the sidelines. Thus, those biotech companies that have been focusing on small molecules over the many years of the development cycle (which themselves were previously in vogue) have in recent years found their drugs less attractive to big pharma, whose managements have been scrambling to expand internal biologics expertise. Overall, in recent years, the number of biotech companies acquired by big pharma has also been dwindling (**Fig. 2**).

How quickly big pharma can disrupt an entire segment of the biotech industry is illustrated by Roche's recent decision to discontinue its research into RNA interference (RNAi) in a cost-cutting restructuring[9]. It had embarked into RNAi only three years ago after having invested $500 million in this area. This abrupt abandonment of the research and partnering efforts has the potential of sending the message to the industry and marketplace that RNAi technology is not worth pursuing, leaving the biotech companies with which they partnered and others pursuing RNAi to defend their continued development of the technology and scrambling to attract other partners for their programs.

Another issue is that big pharma lacks all the tools to separate the winners from the losers early on, and compensates by creating a high hurdle for licensing (as seen in the stagnant growth of licensing deals; **Fig. 3**). Add to this the need for truly innovative products to obtain reimbursement from third-party payers, and one can see why so many drugs now in the biotech pipeline fail to meet the grade. In sum, drugs that offer considerable advances in treating patients will be the ones that will receive market receptivity and, in turn, attract the attention of big pharma.

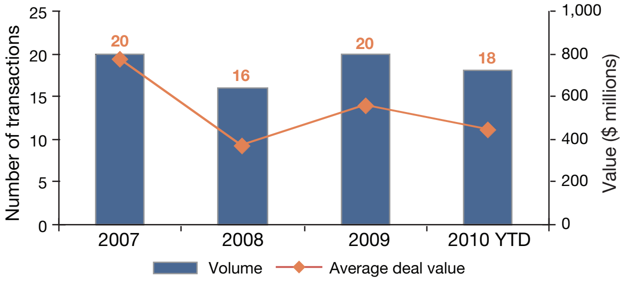


Figure 3 Licensing volumes and payments are declining as big pharma shifts priorities. Source: Burrill & Co.; 2010 year to date (YTD) is through September 30.

What big pharma should be doing

So what can big pharma do to enhance R&D success? The potential steps that seem sensible fall into two broad categories: (i) retooling internal R&D activities and (ii) exploiting business development.

Retooling internal R&D activities. There is an old adage that no one gets fired from big pharma for passing on a compound but might for bringing in one that fails. Therefore, there needs to be clear direction from top management that encourages and rewards risk taking. This will require a cultural change in the organization, which may take years to effect. The Bain Study[8], which interviewed leading global innovators who were responsible for some of the major breakthroughs in medicine, voiced the view that "broken innovation culture" lies at the core of big pharma's problems. This is manifested in a lack of dedication at the level of management to understand the science sufficiently, a reluctance to reach out to academia, high turnover in the R&D executive suites and a lack of willingness to undertake R&D in a different manner. Big pharma executives need to recognize that truly creative steps in product generation are not scalable like other manufacturing processes. Moving drugs from one phase to the next is not as important as getting the development program correct in the first place; and basing incentives on measurements related to other areas of the business, such as speed of throughput, applies pressures to mere numerical outcomes. Progress should be measured against the development plans thoughtfully devised[8].

Furthermore, there is no one-size-fits-all and each big pharma organization will need to determine which models can be implemented to enhance its own R&D productivity rather than merely replicating what others are doing.

It goes without saying that big pharma must determine the class of drugs and therapeutic areas it wants to exploit. This is not a static situation and pharma management needs to anticipate where the best prospects will reside. For example, in October 2010, Roche announced that it is planning to expand its business beyond its historic focus on oncology drugs if results from experimental treatments for a broader range of diseases prove favorable. Also, Sanofi-aventis, by its proposed acquisition of Genzyme, is making a major effort in biologics, albeit through an acquisition.

The first step that big pharma should consider in fostering a cultural change to enhance innovation is to adopt a more entrepreneurial environment for the way its R&D is conducted that aims to emulate the incentive-based culture of smaller biotech companies. For example, GSK has mimicked the biotech model and divided its research groups into smaller and supposedly more nimble segments. It is yet to be seen whether this reorganization proves as effective as biotech ventures in spurring innovation; indeed, success will likely require granting management of the different centers greater autonomy; aligning each center's research goals with incentives; finding the right talent; and minimizing the inherent bureaucracy and management silos imbedded within the large parent corporation.

Other companies have resorted to different structural reorganizations to accelerate the movement of products through the pipeline. For example, Eli Lilly (Indianapolis) created 'Chorus', an independent division designed to get compounds to proof of concept more quickly and cheaply than its regular development organization (http://www.choruspharma.com/about-us.html). Given the imminence of the patent cliff and the ever-present earnings pressure, it would behoove pharmaceutical management to look to these and other innovative mechanisms of getting compounds to human testing faster and accelerating 'go, no-go' decisions. What has been successful for one company, however, may not resonate with another. Trial and error will of necessity be the result. But clearly going down the traditional development path is not an option.

Yet, big pharma also needs to recognize that the improvement in how internal R&D operates will not, in and of itself, solve the immediate pipeline shortage. Big pharma will need to shift significant resources from internal development to a whole host of external resources.

Exploiting business development. In addition to addressing internal R&D shortcomings, big pharma needs to find ways to exploit external partners to enhance R&D success. Even Merck, famous for its go-it-alone attitude, has acknowledged the need to access expertise in academia and other companies. Several approaches can be adopted to further this end.

Historically, business development activities in the pharmaceutical industry have been constrained for several reasons. There has been reticence to partner outside core competencies. Also, there has been a bias to focus resources on the internally developed pipeline. Just as importantly, earnings considerations (rather than cash constraints) have diminished efforts to in-license promising drugs. The net effect has been to constrain external programs.

To exploit external resources more effectively, big pharma should consider combining the best of biotech with its own considerable attributes. Biotech companies bring intensity, entrepreneurialism and an agility to drug development. Conversely, large companies offer a global reach for commercial products, much needed cash and expertise across scientific, medical and regulatory disciplines, as well as the ability to run large clinical trials worldwide. How can management in pharmaceutical companies capitalize on these attributes to promote R&D efficiency?

Given the inherent risks and costs associated with drug development, the pharmaceutical industry has for some time been divesting itself from early-stage drug discovery activities and focusing instead on the later stages where the likelihood of success is much greater. This has created a gap in the process of translating basic research into potential drug candidates. Declining productivity combined with increasing costs within the pharmaceutical sector can serve as the basis of a new collaborative model whereby big pharma steps in and aligns itself with those aspects of drug discovery and development research at which academia excels, with those aspects big pharma does best, and also implementing those aspects that can best be handled collaboratively. In such a model, the collaborators would need to initially determine their respective goals and responsibilities. For example, an academic laboratory that has particular expertise around new biology that has resulted in a high-profile publication could set out to develop functional assays for a particular target in that pathway that would take the project to the stage where it becomes sufficiently validated for big pharma to step in; in turn, the pharma partner could provide much needed funding, development expertise and project management. Fundamental to this approach is a genuinely collaborative environment where academic contributions to translational research are rewarded with appropriately structured funding, intellectual property (IP) ownership and a share in future royalties. The collaboration between academic nonprofits engaging directly with big pharma to advance drugs through the clinic is not without major challenges, including differing cultures (blue-sky versus goal-oriented outlooks, openness and trade secrets), IP rights, conflicts of interest, as well as legal and public relations issues.

Some early steps in this direction are already being taken. In a partnership with academia that focuses on discovering new uses for existing compounds, Pfizer has agreed to give scientists at Washington University School of Medicine (St. Louis) unprecedented access to information regarding a vast number of small molecules and small-molecule candidates. Recently, Pfizer and the University of California San Francisco (UCSF) entered into a collaboration whereby the pharmaceutical company will pay up to $85 million to explore whether UCSF discoveries can be translated into biologic drugs. UCSF scientists will have access to Pfizer's drug development expertise. Ownership rights to the development results will be shared. In another example, Sanofi-aventis has teamed up with Harvard University (Cambridge, MA, USA) to collaborate in key therapeutic areas with a view to enhancing the pharma company's product portfolio.

Technology platforms can also provide novel ways to advance drug development. For instance, Pfizer recently entered into a research collaboration with Biovista (Charlottesville, VA, USA) to seek new indications using the latter's literature-searching platform to identify additional new indications for existing drugs. Another strategy related to platform technology, with inherent limitations, is to pool resources to access so-called precompetitive platforms. These can involve novel tests, imaging methods or safety-testing methodology. For example, Eli Lilly, Johnson & Johnson (New Brunswick, NJ, USA), Merck and Pfizer have invested in venture fund PureTech's Enlight Biosciences (Cambridge, MA, USA), which was established for the purpose of creating enabling technology platforms to meet needs shared across the industry.

In terms of in-licensing drugs, various approaches can be exploited to enhance the probability of success. For example, pharma can license the drug in a collaboration that leaves the biotech to advance the product unfettered during the early stages of development and then take over management of late-stage trials in which big pharma has the greater expertise; or, alternatively, pharma may option only the right to commercialize the compound and leave all the development efforts with the biotech company, thereby avoiding the development costs and risks.

A way to get access to compounds already under development is to use captive venture capital arms. For example, Novartis relies on its own venture capital fund while, at the same time, partnering with an external private equity group to share the costs and development risks. This model also incorporates a mechanism that enables Novartis to have an option to acquire the program if it is successful.

To avoid overpaying for compounds up front, big pharma has employed a risk-sharing arrangement whereby contingent payments are made upon the future commercialization of the licensed or acquired compounds.

In addition, when there is a constraint on R&D spending, one potential solution is to find outside resources to share the costs and risks. For example, TPG-Axon Capital (New York) and Quintiles Transnational's NovaQuest (Charlotte, NC, USA) partnering group has agreed to fund Eli Lilly's two Alzheimer's disease compounds in phase 3 testing. There are also firms, such as my own, Symphony Capital (New York), that are engaged in structured drug development financing to further the development of biotech compounds and, in the case of Symphony Capital, do so in a manner that also ameliorates an adverse impact on earnings.

Another major factor that can increase the chances of successfully accessing external drugs is to improve the manner in which collaborations are being pursued and conducted. Collaborations at big pharma require an internal champion, 'socializing' the opportunity across multiple departments and getting buy-in. Thus, potential partners must often negotiate strategy, marketing, portfolio managers and cross-functional committees, all of which weigh in on a decision for whether a collaboration should proceed. Is it any wonder that some transactions between big pharma and biotech have been reported to take two years; or that there is a high likelihood that this entire decision process will suffer from death by committee? What's more, as internal champions are moved to other areas, numerous biotech firms have encountered the frustration of stalled or terminated collaborations or at least having to start the collaboration process afresh with the new, designated pharma appointee (who may have interests different from the former incumbent). Pharmaceutical management needs to recognize that this painful incremental process of decision making needs to be streamlined so as not to turn off important external opportunities in the biotech sector.

Negotiating with different pharmaceutical companies can also be daunting for biotech companies. Several big pharmas follow a rigid template to conduct business development. Often, an investment in a program is determined on a numbers game relying solely on averages and probabilities, ignoring the more fundamental aspect of the innovation. Decision makers within pharma companies need to embrace a more flexible approach to structure the collaboration. Similarly, it is important for the teams engaged in the negotiations to have a clear understanding as to what is the key ingredient in the collaboration. They must determine at the outset if this is an IP transaction or whether there is a need to retain key people with the programs. Will big pharma just bring needed capital or will it provide additional expertise?

It is also important to conduct the negotiations to create a win-win outcome. Even though the current environment has tipped the negotiating balance in favor of big pharma, it would be detrimental if bargaining is driven so hard it ultimately creates serious friction between the participating parties to the detriment of the programs being conducted. Finalizing the agreement is only one aspect of a research collaboration on a drug candidate. The manner in which the collaboration is conducted afterwards is paramount, as a poorly executed collaboration can easily destroy a drug's prospects. It has been reported that over two-thirds of the collaborations in pharma fail or suffer significant downside events during the alliance[10]. Thus, effectively managing the collaboration is necessary to ensure that the value of the alliance is not jeopardized.

Often rigid policies impose constraints on business development teams, such as requiring them to rely on internal expertise in connection with the transaction. Pharmaceutical decision makers need to be allowed by upper management to engage outside expertise for domain and other resources, rather having to rely on suboptimum internal teams.

Finally, as the pharmaceutical industry continues to evolve, development tools, clinical outcome tracking and other activities may create new opportunities for business development. For example, it may make sense to outsource to a much greater extent various activities currently conducted within big pharma to academic institutions that contain the specific expertise, or to use health screening companies to get better outcomes data.

Conclusions

The current adverse environment confronting both the pharmaceutical and the biotech sectors is not likely to change dramatically in the near term. Big pharma's internal R&D activities, even if restructured and conducted more effectively as suggested, in and of themselves, are not going to yield the number of drugs necessary to deal with the coming patent cliff. At the same time, Wall Street in the current environment is not likely to fund the biotech industry to the extent needed to continue innovative drug development. Thus, big pharma's partnering activities are increasingly going to be needed by the biotech sector to further its drug development efforts.

The present environment provides big pharma with an opportune time to abandon its traditional R&D business model, which no longer produces the intended results, and expand its external collaborations in an effective manner. But this will require bold leadership to change the conservative culture that is embedded throughout big pharma. By creating the right culture and achieving the right balance between its internal and external R&D efforts, the pharmaceutical industry can reinvigorate its drug pipeline and drive growth and earnings in a sustainable manner. But the window of opportunity to attain that culture and balance may only remain open for a short time.

COMPETING FINANCIAL INTERESTS

The author declares no competing financial interests.

1. Anonymous. Results announcement for the third quarter 2010 (GlaxoSmithKline, London) <http://www.gsk.com/investors/reports/q32010/q32010.pdf> (21 October 2010).
2. Anonymous. NICE rebuffs Tarceva. *BioCentury Extra*, p.3 (16 June 2010).
3. PricewaterhouseCoopers. *Pharma 2020, Virtual R&D, Which Path Will You Take?* (PricewaterhouseCoopers, New York City, 2008).
4. Spek, H. CLSA Healthcare Forum, New York, December 1, 2010. <http://en.sanofi-aventis.com/binaries/EM_conference_Spek_Web_tcm28-29638.pdf>
5. Shaffer, C. *Nat. Biotechnol.* **28**, 881–882 (2010).
6. Ernst & Young. *Beyond Borders—Global Biotechnology Report* (E&Y, New York, 2009).
7. Anonymous. Big pharma aims for reinvention. *Financial Times* (12 May 2010).
8. Behnke, N. & Sueltenschmidt, N. *Changing Pharma's Innovation DNA* (Bain & Company, Boston, 2010).
9. Ledford, H. *Nature* **468**, 487 (2010).
10. Oliver Wyman. *Licensing to Win* (Oliver Wyman, New York, 2008). <http://www.oliverwyman.com/ow/pdf_files/OW_EN_HLS_2008_LicensingtoWin.pdf>

Patent term extensions for biologic innovators in Japan

John A Tessensohn & Shusaku Yamamoto

The Japanese Intellectual Property High Court recently issued two decisions that bolster the market exclusivity period for brand biologic manufacturers.

Last year, Japan established a regulatory pathway for generic versions of patented biologics, referred to as follow-on biologics or biosimilars. In addition, the Japanese patent term extension (PTE) and accelerated patent examination legislative regimes were established to offer innovators a means of optimizing the patent-protected life span of biopharmaceuticals in their portfolios. Now a landmark decision from the Intellectual Property High Court of Japan (IPHCJ; Tokyo; **Box 1** and **Table 1**) in Japan's first-ever biopharmaceutical PTE judicial appellate proceedings, which upheld the market exclusivity period for the brand biologic Enbrel (etanercept), has provided the biotech industry with greater certainty that competition from makers of biosimilars will be balanced by a patent monopoly period that allows a return on investment.

Box 1 IPHCJ

The IPHCJ is Japan's specialist intellectual property appellate court[13], modeled after the United States Court of Appeals for the Federal Circuit, which exercises supervisory jurisdiction over all JPO decisions, including PTE appeal decisions. The IPHCJ conducts *de novo* review of all JPO decisions. When adjudicating over complicated technical matters, like in the etanercept PTE cases, the IPHCJ's deliberations are briefed and aided by technical advisors (*saibansho chōsa-kan*)[14] who are chosen from a pool of university professors or credentialed experts in the field. Recent statistics[15,16] show that the IPHCJ has recently been reversing JPO decisions at a slightly increasing rate (**Table 1**).

Table 1 IPHCJ outcome trends of JPO appeal decisions—patents

Year	Total number of appeals	Number reversed	Percent reversed
2003	122	24	20
2004	157	16	10
2005	147	11	7
2006	162	23	14
2007	188	25	13
2008	171	30	18

Biosimilars in Japan

Japan quietly promulgated its biosimilar regulatory pathway[1] without the *Sturm und Drang* that polarized the US healthcare reform debate over its biosimilar regulatory pathway[2–4]. A study group under the auspices of the Japanese Ministry of Health Labor & Welfare (MHLW) began official review of biosimilars in 2007. After several public rounds of discussion and recommendations, the MHLW implemented its biosimilars guidelines on March 4, 2009 (**Box 2**).

On June 22, 2009, Japan approved its first biosimilar product—recombinant human growth hormone somatropin. Early doubts over the biosimilars business model[5] have apparently not curbed the enthusiasm of Japanese biosimilars business activity with foreign pharmaceutical giants like Merck (Whitehouse Station, NJ, USA)[6], Sanofi-aventis (Paris)[7] and GlaxoSmithKline (Brentford, UK)[8] to feverishly implement or ink development, licensing and marketing deals with Japanese partners. Japanese generics player Nippon Kayaku (Tokyo) has also partnered with Teva-Kowa Pharma (Tokyo) to develop for the Japanese market a biosimilar version of recombinant human granulocyte colony stimulating factor, a hormone used to accelerate the replenishment of white blood cells lost because of chemotherapy[9].

As national and multinational corporations increase their interest in developing biosimilar products, the question arises as to how brand manufacturers can best rally their intellectual property protection resources to defend against erosion of key markets. In this respect, the existing PTE and accelerated patent examination systems provide useful mechanisms by which biotech brand manufacturers can optimize their effective patent terms to mitigate biosimilars competition in the Japanese market (**Box 3** and **Fig. 1**).

The IPHCJ decisions

In a pioneering pair of biopharmaceutical appellate PTE proceedings, the IPHCJ (**Box 1**) overturned the Japan Patent Office's (JPO) denial of two PTE applications for a biotech therapeutic, etanercept[10].

The patents in suit were granted on November 1997 and July 1999, respectively. The regulatory clinical trials commenced on December 17, 1999, and the MHLW regulatory marketing approval was granted on January 19, 2005. The patentee applied for a full five-year

John A. Tessensohn & Shusaku Yamamoto are at Shusaku Yamamoto Patents, Chuo-Ku, Osaka, Japan.
e-mail: jtessensohn@shupat.gr.jp

extended term of nonworking as the health regulatory delay period of nonworking was 5 years, 1 month and 1 day for both patents. (A patentee is entitled to work one's invention, e.g., if the patent covers the product, the act of making of the product constitutes working. However, due to health regulatory issues, a patent directed to a pharmaceutical may not be able to be worked pending approval by the appropriate health regulatory authority. This period when the invention cannot be worked is the period of nonworking.)

The patentee filed their PTE applications with the patent office on April 18, 2005, but the JPO issued final rejections of the patentee's PTE applications on December 25, 2007. The patentee appealed the final rejections but on November 28, 2008, the board of appeals affirmed them on the ground that the patented claims that are the subject of the PTE applications did not explicitly cover the etanercept fusion protein product which was the subject of Japanese regulatory approval. The patentee filed suit with the IPHCJ to reverse the JPO's decisions.

Etanercept was approved in Japan to treat rheumatoid arthritis. It acts by binding tumor necrosis factor (TNF), one of the dominant inflammatory cytokines, or regulatory proteins, that play an important role in both normal immune function and the cascade of reactions causing the inflammatory process of rheumatoid arthritis. The binding of etanercept to TNF renders the bound TNF biologically inactive, resulting in considerable reduction in inflammatory activity. Additionally, etanercept binds to lymphotoxin-α, another cytokine involved in the inflammatory process of rheumatoid arthritis[11].

Etanercept is a recombinant, soluble, dimeric fusion protein, consisting of the extracellular ligand-binding region of recombinant human TNF receptor attached to the constant (Fc) region of human IgG. The receptor moiety of etanercept binds to circulating TNF (two molecules of TNF per receptor) and inhibits its attachment to endogenous TNF cell surface receptors, thereby rendering TNF inactive and inhibiting TNF-mediated mechanisms of inflammation.

The PTE claims in suit were directed to "TNF-R protein" but the JPO held that "the peptide of amino acids 258 to 489 corresponding to a polypeptide corresponding to the Fc region of human immunoglobulin G1 of etanercept was not expressly recited in the granted claims" and therefore did not encompass the etanercept fusion protein.

The specification of the patents in suit did stipulate that the TNF-R protein recited in the claims "may include, in addition to a polypeptide having TNF-R activity, other chemical moieties such as polypeptides…." The 'live' issue before the IPHCJ was whether or not "chemical moieties other than a protein polypeptide having TNF-R activity, such as a polypeptide" includes a polypeptide corresponding to the Fc region of human IgG1.

The JPO held that the polypeptides as "chemical moieties" that could be added to a protein having TNF-R activity were limited to "low molecular weight" polypeptides and the only permissible "chemical moieties" were also limited to "low molecular weight" chemical moieties. The JPO sought to justify this overly narrow claim interpretation of the TNF-R protein by cherry-picking, out of context, certain examples and sentences in the specification.

The court overrules overly narrow, pedantic claim interpretation

The IPHCJ was not sympathetic to the JPO's pedantic, selective and overly narrow interpretation of the examples and description in the specification. The IPHCJ found that there was ample and express support in various parts of the patent description, the patent specification as a whole, the common general knowledge as evidenced by numerous textbooks, that was entered into the record, and the relevant expert testimony adduced by the patentee that the JPO's low molecular weight limitation was an unreasonable and untenable interpretation.

Accordingly, on the strength of the patentee's submissions, Presiding Judge Tetsuhiro Nakano found that the JPO's restrictive claim interpretation of the TNF-R protein's scope was untenably incorrect and overturned the JPO's denials. Nakano found in favor of the patentee that the claims in their PTE applications did cover the etanercept fusion protein product that was the subject of Japanese regulatory approval, and were entitled to the PTE period of an additional five years, handing a complete unconditional victory to the biotech patentee.

On the basis of the above, biotech applicants should never shy from challenging questionably suspect JPO decisions to the IPHCJ, even though the IPHCJ maintains a large majority of the JPO decisions that it

Box 2 Japan's biosimilars framework

Japan's biosimilars guidelines incorporate a similarity standard that is akin to that of the European Union (EU; Brussels) approval pathway[17]. Under the MHLW's guidelines, biosimilars, by definition, refer to a biotechnological product that is produced by a subsequent-entry manufacturer and claimed to be comparable to a biopharmaceutical product already approved in Japan. In terms of development, the biosimilar is to be developed with current technologies and knowledge, and therefore need demonstrate only enough similarity to the brand product to guarantee safety and efficacy instead of absolute identity.

As with new biotechnological products, the MHLW requires not only the establishment of a well-defined manufacturing process, but also extensive characterization studies to demonstrate the molecular and quality attributes of the biosimilars under the US Food and Drug Administration's ICH Q5E guidance principles[18], with evaluation based on the data from nonclinical and clinical studies.

The MHLW's clinical efficacy guidelines require the conduct of clinical studies to verify the efficacy of the biosimilars and that the reference biopharmaceutical is bioequivalent and/or equivalent in quality where high similarity in terms of quality has been demonstrated. Studies to verify the bioequivalence/quality equivalence of the efficacy of the biosimilar with the originator biodrug should be appropriately designed and their validity explained. Specifically, the target number of cases should be set as necessary and valid, and the acceptable bioequivalence/quality-equivalence range pre-specified using clinically established endpoints. Where appropriate surrogate endpoints are available, the use of true endpoints will not always be required, but their validity should be fully explained on the basis of corroborative data or literature.

It is permissible to extrapolate to the biosimilars the other indications approved for a reference biodrug where it can be explained that efficacy is equivalent with respect to certain indications and that a similar pharmacological action can be expected in other indications through relevant pharmacokinetics and pharmacodynamics studies or analysis.

Lastly, post-marketing surveillance of all potential safety profiles, including immunogenicity, is required. The subsequent entrant must assure the traceability of adverse events during the respective surveillance period and notwithstanding any switch of the originator biodrug or drug with similar indications to the biosimilar, their substitution or combined application should in principle be avoided throughout the treatment period.

Box 3 The Japanese PTE

In Japan, PTE is available only if the granted patent cannot be worked, i.e., the granted patent covering the pharmaceutical cannot be sold or marketed to the public because of the requirement to obtain regulatory marketing approval from MHLW[19].

Users of the US patent system can use patent term adjustment to recoup lost patent term due to patent examination delays[20]. Unfortunately, there is no such patent term adjustment provision in Japan for patent examination delays, even though several pioneering biotech patents had severely shortened patent terms due to patent examination delays at the JPO[21].

Under Japan's patent law, the extended patent term corresponds to the "period which the patented invention could not be worked" owing to the requirement that pharmaceutical products must receive health regulatory approval before marketing (that is, the regulatory review period). The period of PTE commences on the day when the clinical trial is started or the day when the patent is registered, whichever is later, and ends on the day just before the day when the regulatory approval is mailed to the applicant of the regulatory approval[22]. It is not possible to file a PTE application for a pending patent application. The subject of PTE has a maximum duration of five years.

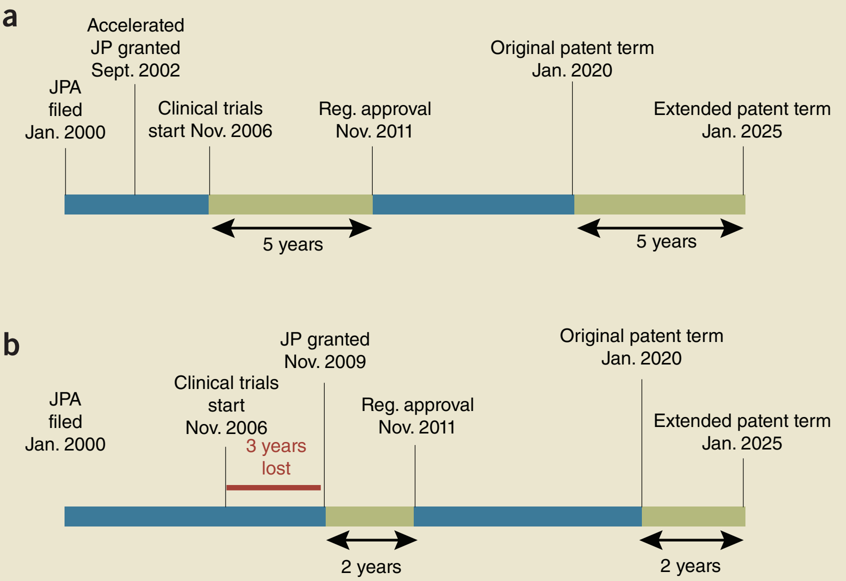

a
Accelerated JP granted Sept. 2002
JPA filed Jan. 2000
Clinical trials start Nov. 2006
Reg. approval Nov. 2011
Original patent term Jan. 2020
Extended patent term Jan. 2025
5 years
5 years
b
JP granted Nov. 2009
Original patent term Jan. 2020
Clinical trials start Nov. 2006
JPA filed Jan. 2000
Reg. approval Nov. 2011
Extended patent term Jan. 2025
3 years lost
2 years
2 years

Figure 1 Calculation of patent term extension periods. (**a**) Japanese patent term extension when a patent is granted before human testing is initiated. (**b**) Japanese patent term extension where trials commence before the patent is granted. JP, Japan Patent; JPA, Japan Patent Application.

To enjoy the optimal protection of the extended patent term, it is imperative to ensure the expeditious prosecution and grant of a valuable biotech patent application before commencing clinical trials. We illustrate this using two examples; one where a patent is granted before human testing is initiated (**Fig. 1a**) and another where trials commence before the patent is granted (**Fig. 1b**). In the examples, the Japanese patent application filing dates, the commencement and end dates of the clinical trial periods are all identical; only the patent grant dates are different.

In **Figure 1a**, the patentee is entitled to the maximum PTE benefit—the full five years—permissible under Japan's PTE period as the Japanese granted patent could not be worked in view of the health regulatory delay at MHLW. In contrast, the scenario in **Figure 1b** is that there was no accelerated patent grant and the clinical trials commenced before the patent grant date. The nonworking period caused by regulatory delay—three years—does not entitle the patent holder to any patent term extension whatsoever because during the relevant nonworking period, as the subject matter was not a granted patent, only a pending patent application.

In the **Figure 1b** scenario, the applicant was not deprived of the benefit of working its granted patent right as it only had a pending application right during the first three-year period of regulatory delay. Unfortunately, this lost three-year period of regulatory delay is irrecoverably wasted and cannot be the subject of any patent term extension because it occurred during the nonqualifying period.

Japan has several expedited examination procedures including "accelerated examination," the "patent prosecution highway" and the "super accelerated examination"[23]. As more biosimilars become available on the Japanese market, biotech innovators should ensure that their Japanese patent grant, accelerated or otherwise, is obtained before commencing clinical trials. This will maximize capture of the qualifying regulatory delay period for optimal patent term extension benefit and avoid the wasteful loss of potentially years of patent term.

adjudicates upon. It is worth expending the legal effort to do so in biopharmaceutical PTE matters because even a year of extended patent life for a blockbuster biopharmaceutical can translate into "additional millions or billions of dollars" of profits[12].

Conclusions

The pair of IPHCJ decisions discussed here in Japan's first biopharmaceutical PTE judicial appellate proceedings represent good news for biotech innovators. As such, companies with brand products should avail themselves of Japan's biotech-savvy courts, accelerated patent examination and PTE regimes so that valuable patent terms and profits can be optimized and protected. With increasing biosimilars competition, the biotech industry has little choice but to ensure smoother coordination with their research, patent and regulatory strategies to optimize all available patent and extended terms from regulatory delay; thereby securing dominance in the world's second-largest national healthcare market, Japan[12].

ACKNOWLEDGMENTS
The views and considerations of this article reflect only those of the authors and should not be attributed to Shusaku Yamamoto Co. or to any of its past, present or future clients.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Tessensohn, J.A. & Yamamoto, S. *World Intell. Prop. Law Rep.* **23**, 31–32 (2009).
2. Connolly, C. & Shear, M.D. Discord on health care dulls luster of new pacts. *Washington Post* 9 July 2009, A1.
3. Anonymous. Biotech bottleneck. *Washington Post* 29 July 2009, A16.
4. Pollack, A. Costly drugs known as biologics prompt exclusivity debate. *New York Times* 22 July 2009, B1.
5. Belsey, M.J., Harris, L.M., Das, R.R. & Chertkow, J. *Nat. Rev. Drug Discov.* **5**, 535–536 (2006).
6. Anonymous. Merck to storm Japan with generic biologics. *Nihon Keizai Shimbun* 3 March 2010.
7. Anonymous. Nichi-Iko eyes biodrug sales in 3 years with Sanofi-Aventis' help. *Nihon Keizai Shimbun* 1 June 2010.
8. GlaxoSmithKline acquired the ex-Japan development and marketing rights to JCR's biosimilar JR-013, a FOB recombinant human erythropoietin kappa that has been approved for renal anaemia in kidney dialysis patients and premature infants, JCR Pharmaceuticals Co. Ltd., *Press Release: Comprehensive Agreement on Biopharmaceuticals Business*, Dec. 18, 2009.
9. Anonymous. Teva-Kowa, Nippon Kayaku to develop generic for chemotherapy. *Nihon Keizai Shimbun* 20 April 2010.
10. *Immunex Corp. v. Commissioner of Japan Patent Office*, H-21 (gyo-ke) No. 10092 and H-21 (gyo-ke) No. 10093 dated Dec. 3, 2009, Intellectual Property High Court of Japan.
11. Wyeth, K.K. & Takeda Pharmaceutical Co. Ltd. Press release: announcing launch of ENBREL, treatment of rheumatoid arthritis. <http://www.takeda.com/press/article_1097.html> (29 March, 2005).
12. Jacobsen, T.M. & Wertheimer, A.I. *Modern Pharmaceutical Industry: A Primer* (Jones and Bartlett Publishers, Sudbury, MA, USA, 2010).
13. Tessensohn, J.A. & Yamamoto, S. *World. Intell. Prop. Law Rep.* **19**, 21–22 (2005).
14. Tessensohn, J.A. & Yamamoto, S. *World. Intell. Prop. Law Rep* 2010/1 18–19 (2010). <http://www.wipo.int/wipo_magazine/en/2010/01/article_0007.html>.
15. Japan Patent Office. *Annual Report 2006 (Japanese Version)* (JPO, Tokyo, 2006).
16. Japan Patent Office. *Annual Report 2009 (Japanese Version)* (JPO, Tokyo, 2009).
17. Schellekens, H. & Moors, E. *Nat. Biotechnol.* **28**, 28–31 (2010).
18. US Department of Health and Human Services *et al.* Guidance for industry Q5E comparability of biotechnological/biological products subject to changes in their manufacturing process. <http://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm128076.pdf> (June 2005).
19. Tessensohn, J.A. *Modified-Release Drug Delivery Technology*. (Rathbone, M.J. *et al.*, eds.) 45–51 (Informa Healthcare, New York, 2008).
20. *Wyeth v. Kappos*, 93 U.S.P.Q. 2d 1227 (Fed. Cir. 2010).
21. Tomomi, A. *et. al. Nat. Biotechnol.* **25**, 533–535 (2007).
22. Tessensohn, J.A. & Yamamoto, S. *Biotechnol. Law Rep.* **28** 483–496 (2009).
23. Tessensohn, J.A. & Yamamoto, S. *Nat. Biotechnol.* **27**, 815–818 (2009).

The next phase in human genetics

Vikas Bansal, Ryan Tewhey, Eric J Topol & Nicholas J Schork

Experimental haplotyping of whole genomes is now feasible, enabling new studies aimed at linking sequence variation to human phenotypes and disease susceptibility.

The maternal and paternal copies of each chromosome in the human genome have distinct combinations of nucleotides that are functionally important, but knowledge of this 'haplotype' information (**Fig. 1a**) has been absent from all but a handful of studies of genomes[1–3]. The reason for this is largely technical: determining haplotypes, or 'phasing,' is not trivial[4,5]. Two papers in this issue report experimental methods for phasing at the genome scale. Fan *et al.*[6] physically separate the chromosomes in single cells using a microfluidic device, essentially phasing before genome analysis. Kitzman *et al.*[7] prepare standard mixtures of maternal and paternal chromosomes for whole-genome sequencing using a new protocol that enables phasing through bioinformatics analysis. Haplotyping strategies such as these should transform human genome sequencing and the study of the phenotypic effects of combinations of genome sequence variants.

Haplotype information is essential for human genetic research because of the fundamental importance of diploidy in human biology, as seen in phenomena such as haploinsufficiency, recessive acting variants, dosage compensation and parent-of-origin imprinting effects. A simple example that shows the need for phasing is compound heterozygosity, which occurs when an individual is a heterozygote carrier of two different mutations at different loci in the same gene. In such cases, a phenotype may arise only when the two mutations are present on different chromosomes, disrupting both copies of the encoded protein[3]. In a more complex example, the two mutations may give rise to a mutated protein from one chromosome and aberrant allele-specific expression or methylation from the other. Assessing the combined effect of mutations implicated in compound heterozygosity requires phase information; simply knowing that an individual is heterozygous at the two loci is not enough. Resolving phase is also important for addressing a range of other problems in human genetics, including characterizing the genomes of under-studied populations[7], comparing chromosomal segments shared between *homo sapiens* and distant ancestors, and detecting complex genomic structural variation.

Existing methods for phasing an individual's whole genome have involved either analysis of related individuals[3], which is often not possible, or labor-intensive one-by-one cloning and sequencing of many large fragments of the genome[1], which is not scalable. Other approaches, such as pedigree- and population-based statistical phasing algorithms, have been used in traditional linkage and linkage disequilibrium mapping. But they are not comprehensive, typically resolving haplotypes only for specific genomic regions and particular variations co-segregating with a phenotype. Moreover, because these other strategies use probabilistic haplotype information, they do not directly observe all of the nucleotide content of a haplotype and are not appropriate for rare variants, which may not have been observed in enough people to be statistically informative. Imputation methods based on phasing algorithms, which infer missing genotype information from other available data, also suffer from the same limitations. Finally, previous haplotyping methods based solely on DNA sequencing reads can be incomplete without additional information to anchor those reads to a chromosome[4,5].

The two papers in this issue are among the first potentially scalable and accurate methods for experimentally phasing entire human genomes. Kitzman *et al.*[7] describe a cost-effective strategy for assembling long haplotypes by sequencing many haploid subsets of an individual genome using next-generation sequencing platforms (**Fig. 1b**). The first step in this approach is to generate a single whole-genome fosmid library with long inserts, in this case ~37 kb. This library is then randomly partitioned into pools such that each pool is essentially a haploid mixture of clones derived from either the maternal or paternal DNA at each genomic location. High-throughput sequencing of each pool provides haplotype information for each clone in that pool. Overlaps between haplotypes derived from different pools are then pieced together to assemble even longer haplotypes. Notably, the method of Kitzman *et al.*[7] is similar to the approach taken to sequence the genome of Craig Venter[1], which used fosmid libraries and standard Sanger sequencing to obtain long-range haplotype information as well as computational algorithms to assemble long haplotypes. The approach of Kitzman *et al.*[7], though, is more scalable and amenable to short-read sequencing.

Kitzman *et al.*[7] apply their method to study the genome of a female with ancestry from western India. They assemble haplotype contigs, or blocks, with a length of ≥386 kb for about half of the genome. Using the resulting phase information, they identify 10 genes (from a candidate set of 44 genes) that harbor two or more rare heterozygous functional mutations that are on different homologous copies of the gene and might therefore cause compound heterozygous phenotypic effects. The authors also use the phase information to identify haplotypic segments that are enriched for novel variants and differ substantially from previously sequenced HapMap populations, suggesting that haplotypes, in contrast to the genotype information captured in the HapMap initiative, contain much more information about ancestry. Lastly, the authors are able to use the fosmid data to detect complex structural variants—a difficult task when both homologous chromosomes are sequenced together.

Vikas Bansal, Eric J. Topol & Nicholas J. Schork are at Scripps Health, La Jolla, California, USA; Vikas Bansal, Ryan Tewhey, Eric J. Topol & Nicholas J. Schork are at The Scripps Translational Science Institute, La Jolla, California, USA; Eric J. Topol & Nicholas J. Schork are in The Department of Experimental Medicine, The Scripps Research Institute, La Jolla, California, USA.
e-mail: nschork@scripps.edu

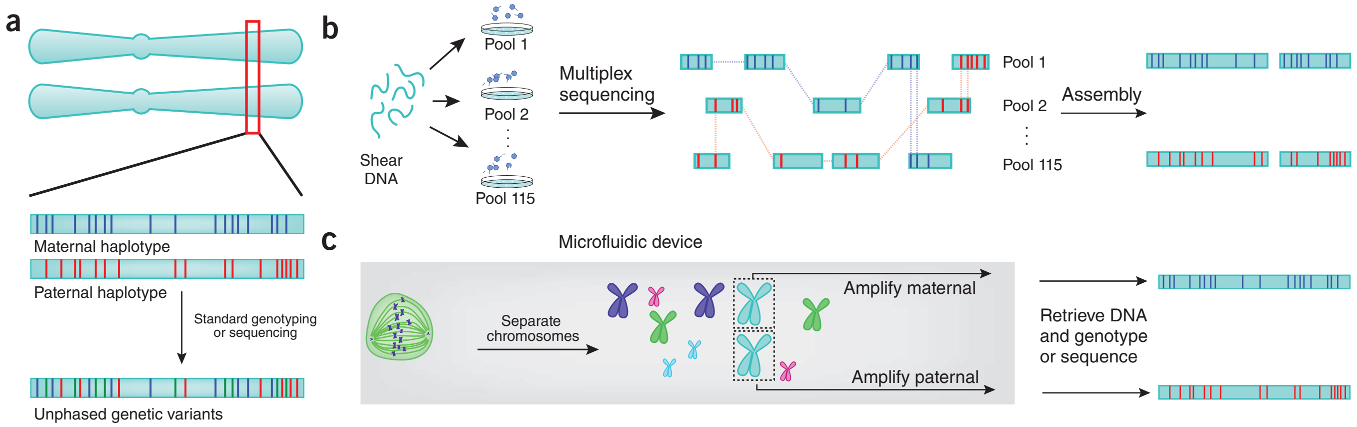


Figure 1 Experimental approaches to haplotyping genomes. (**a**) In standard analysis of a mixed pool of maternally and paternally inherited chromosomal DNA, haplotype information is lost. Blue vertical lines represent sequence variants on the maternal chromosome homolog; red vertical lines represent variants on the paternal homolog; green lines represent homozygous variants. (**b**) Kitzman *et al.*[7] exploit large-insert fosmid clone libraries that allow sequencing reads derived from a single chromosomal homolog to be associated with each other. Cloning and sequencing is multiplexed, enabling efficient construction of contigs that span large genomic regions. Although this approach may be easier to implement in most sequencing laboratories, it is less robust than the approach of Fan *et al.*[6] owing to the need for assembly. (**c**) Fan *et al.*[6] use a microfluidic device to separate and amplify homologous chromosomes during metaphase in single cells, enabling the individual chromosomes to be sequenced and nearly perfectly phased.

One drawback of the approach of Kitzman *et al.*[7] is that it requires stitching together phased contigs, albeit rather large ones. This may result in switching errors, where chromosomal segments are accurately haplotyped but misrepresent complete chromosomes. Such errors can occur once or many times over different chromosomes. These concerns are eliminated in the approach developed by Fan *et al.*[6], which resolves phase directly by isolating individual copies of each chromosome in a single cell with a microfluidic device. Thus, there is no need to assemble contigs. After isolation of single chromosomes, a haploid mixture of clones derived from either the maternal or paternal DNA can be genotyped using standard single-nucleotide polymorphism arrays or shotgun sequenced to generate haplotypes spanning entire chromosomes (**Fig. 1c**).

Fan *et al.*[6] validate their approach by haplotyping the genomes of a mother-father-child trio from the HapMap project. The experimentally determined haplotypes are highly concordant (99.8%) with previously calculated haplotypes inferred using family and population information, demonstrating the accuracy of the approach. The authors also demonstrate the potential of their method for clinical diagnostics by haplotyping a fourth individual, P0, whose genome has already been sequenced, at the highly polymorphic HLA locus. Phase information for this genomic region is very important for matching transplanted donor organs with a potential host. Although the approach taken by Fan *et al.*[6] is direct and the most optimal for phasing, the need for sophisticated microfluidic devices and sequencing technologies able to handle individual chromosomes may delay its routine use.

The work of Fan *et al.*[6] and Kitzman *et al.*[7] highlight the obvious, yet often overlooked, diploid nature of the human genome and expose the incompleteness of available individual genome sequences that do not phase genetic variants. Going forward, discussions of individual human genomes should refer either to a single maternally or paternally derived haplotypic complement of DNA or to the two genomes that each person possesses. This simple change in language should help emphasize the importance of studies that account for the phase information that is the hallmark of the human genome.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Levy, S. *et al. PLoS Biol.* **5**, e254 (2007).
2. Wang, J. *et al. Nature* **456**, 60–65 (2008).
3. Roach, J.C. *et al. Science* **328**, 636–639 (2010).
4. Bansal, V., Halpern, A.L., Axelrod, N. & Bafna, V. *Genome Res.* **18**, 1336–1346 (2008).
5. He, D., Choi, A., Pipatsrisawat, K., Darwiche, A. & Eskin, E. *Bioinformatics* **26**, i183–i190 (2010).
6. Fan, H.C., Wang, H., Potanina, A. & Quake, S.R. *Nat. Biotechnol.* **29**, 51–57 (2011).
7. Kitzman, J.O. *et al. Nat. Biotechnol.* **29**, 59–63 (2011).

Crafting rat genomes with zinc fingers

Meng Amy Li & Allan Bradley

Expressing zinc-finger nucleases in zygotes enables targeted transgene integration in the mouse and rat genomes.

Mammalian oocytes and zygotes are extraordinarily resilient receptacles that provide a conduit from designs etched in laboratory notebooks to living animals. The first transgenic mammals were generated three decades ago by injection of naked DNA into the pronuclei of mouse zygotes. Until now, however, pronuclear injection has allowed insertion of exogenous DNA only at random sites in the genome, and site-specific engineering has proved extremely difficult. In this issue, Cui and colleagues[1] have finally overcome this barrier, making use of zinc-finger nucleases (ZFNs) to stimulate targeted integration of transgenes by homologous recombination in mouse and rat zygotes. This technology will dramatically alter the speed

Meng Amy Li & Allan Bradley are at Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom
e-mail: abradley@sanger.ac.uk

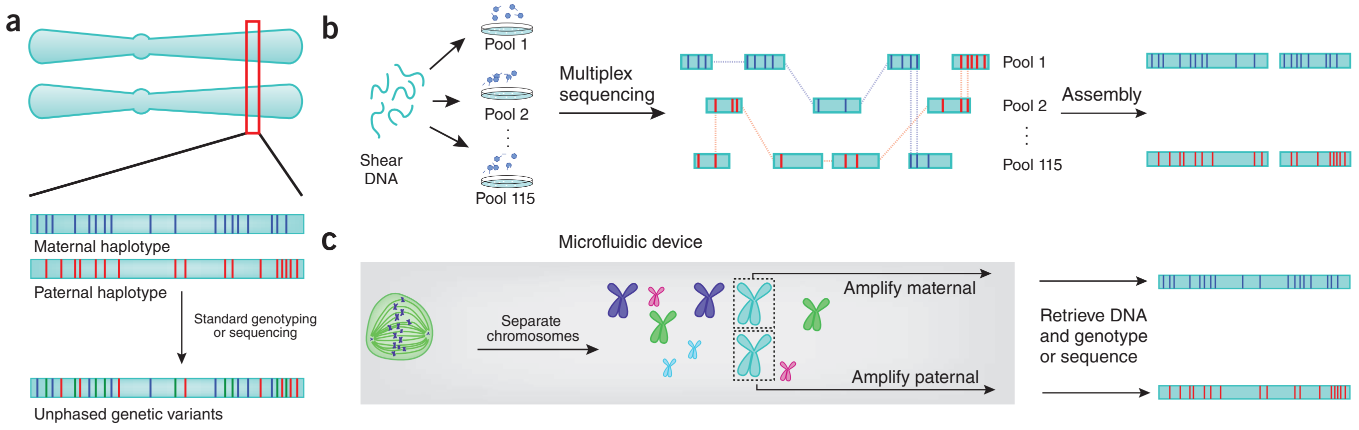


Figure 1 Experimental approaches to haplotyping genomes. (**a**) In standard analysis of a mixed pool of maternally and paternally inherited chromosomal DNA, haplotype information is lost. Blue vertical lines represent sequence variants on the maternal chromosome homolog; red vertical lines represent variants on the paternal homolog; green lines represent homozygous variants. (**b**) Kitzman *et al.*[7] exploit large-insert fosmid clone libraries that allow sequencing reads derived from a single chromosomal homolog to be associated with each other. Cloning and sequencing is multiplexed, enabling efficient construction of contigs that span large genomic regions. Although this approach may be easier to implement in most sequencing laboratories, it is less robust than the approach of Fan *et al.*[6] owing to the need for assembly. (**c**) Fan *et al.*[6] use a microfluidic device to separate and amplify homologous chromosomes during metaphase in single cells, enabling the individual chromosomes to be sequenced and nearly perfectly phased.

One drawback of the approach of Kitzman *et al.*[7] is that it requires stitching together phased contigs, albeit rather large ones. This may result in switching errors, where chromosomal segments are accurately haplotyped but misrepresent complete chromosomes. Such errors can occur once or many times over different chromosomes. These concerns are eliminated in the approach developed by Fan *et al.*[6], which resolves phase directly by isolating individual copies of each chromosome in a single cell with a microfluidic device. Thus, there is no need to assemble contigs. After isolation of single chromosomes, a haploid mixture of clones derived from either the maternal or paternal DNA can be genotyped using standard single-nucleotide polymorphism arrays or shotgun sequenced to generate haplotypes spanning entire chromosomes (**Fig. 1c**).

Fan *et al.*[6] validate their approach by haplotyping the genomes of a mother-father-child trio from the HapMap project. The experimentally determined haplotypes are highly concordant (99.8%) with previously calculated haplotypes inferred using family and population information, demonstrating the accuracy of the approach. The authors also demonstrate the potential of their method for clinical diagnostics by haplotyping a fourth individual, P0, whose genome has already been sequenced, at the highly polymorphic HLA locus. Phase information for this genomic region is very important for matching transplanted donor organs with a potential host. Although the approach taken by Fan *et al.*[6] is direct and the most optimal for phasing, the need for sophisticated microfluidic devices and sequencing technologies able to handle individual chromosomes may delay its routine use.

The work of Fan *et al.*[6] and Kitzman *et al.*[7] highlight the obvious, yet often overlooked, diploid nature of the human genome and expose the incompleteness of available individual genome sequences that do not phase genetic variants. Going forward, discussions of individual human genomes should refer either to a single maternally or paternally derived haplotypic complement of DNA or to the two genomes that each person possesses. This simple change in language should help emphasize the importance of studies that account for the phase information that is the hallmark of the human genome.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Levy, S. *et al. PLoS Biol.* **5**, e254 (2007).
2. Wang, J. *et al. Nature* **456**, 60–65 (2008).
3. Roach, J.C. *et al. Science* **328**, 636–639 (2010).
4. Bansal, V., Halpern, A.L., Axelrod, N. & Bafna, V. *Genome Res.* **18**, 1336–1346 (2008).
5. He, D., Choi, A., Pipatsrisawat, K., Darwiche, A. & Eskin, E. *Bioinformatics* **26**, i183–i190 (2010).
6. Fan, H.C., Wang, H., Potanina, A. & Quake, S.R. *Nat. Biotechnol.* **29**, 51–57 (2011).
7. Kitzman, J.O. *et al. Nat. Biotechnol.* **29**, 59–63 (2011).

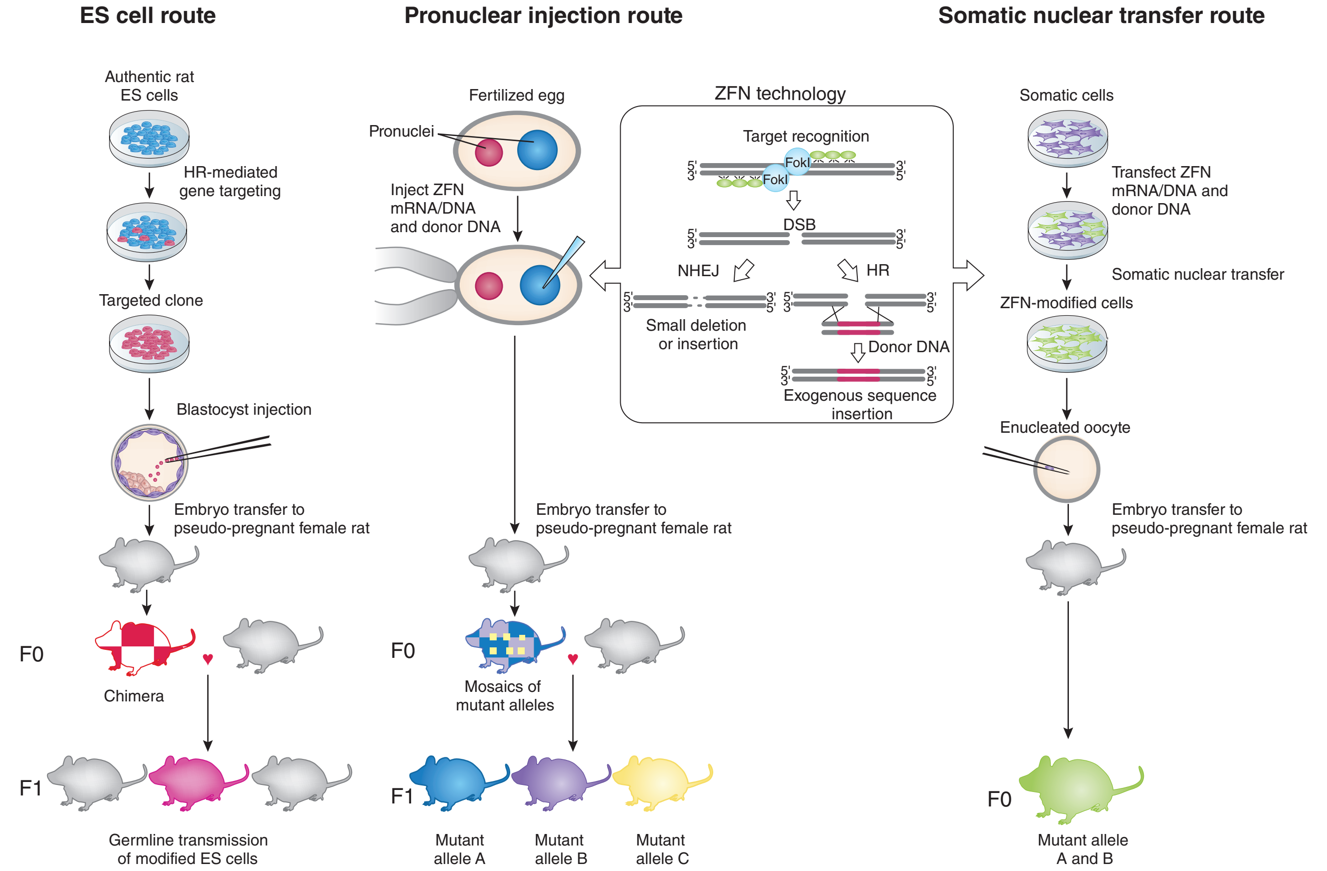
Crafting rat genomes with zinc fingers

Meng Amy Li & Allan Bradley

Expressing zinc-finger nucleases in zygotes enables targeted transgene integration in the mouse and rat genomes.

Mammalian oocytes and zygotes are extraordinarily resilient receptacles that provide a conduit from designs etched in laboratory notebooks to living animals. The first transgenic mammals were generated three decades ago by injection of naked DNA into the pronuclei of mouse zygotes. Until now, however, pronuclear injection has allowed insertion of exogenous DNA only at random sites in the genome, and site-specific engineering has proved extremely difficult. In this issue, Cui and colleagues[1] have finally overcome this barrier, making use of zinc-finger nucleases (ZFNs) to stimulate targeted integration of transgenes by homologous recombination in mouse and rat zygotes. This technology will dramatically alter the speed

Meng Amy Li & Allan Bradley are at Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom e-mail: abradley@sanger.ac.uk


ES cell route
Pronuclear injection route
Somatic nuclear transfer route
Authentic rat ES cells
HR-mediated gene targeting
Targeted clone
Blastocyst injection
Embryo transfer to pseudo-pregnant female rat
F0
Chimera
F1
Germline transmission of modified ES cells
Fertilized egg
Pronuclei
Inject ZFN mRNA/DNA and donor DNA
Embryo transfer to pseudo-pregnant female rat
F0
Mosaics of mutant alleles
F1
Mutant allele A
Mutant allele B
Mutant allele C
ZFN technology
Target recognition
FokI
FokI
5'
3'
3'
5'
DSB
NHEJ
HR
Small deletion or insertion
Donor DNA
Exogenous sequence insertion
Somatic cells
Transfect ZFN mRNA/DNA and donor DNA
Somatic nuclear transfer
ZFN-modified cells
Enucleated oocyte
Embryo transfer to pseudo-pregnant female rat
F0
Mutant allele A and B

Marina Corral

Figure 1 Three principal routes to achieve targeted genome modifications in the rat and other mammalian species. ES cell, embryonic stem cell; HR, homologous recombination; ZFN, zinc-finger nuclease; DSB, double-strand break; NHEJ, non-homologous end joining.

with which genetic alterations can be generated in a variety of mammalian species.

Evidence that mouse pro-nuclei are endowed with the machinery to support homologous recombination emerged more than 20 years ago, but the efficiencies were too low to be practically useful[2]. Instead, the technology to manipulate the mouse genome has relied on embryonic stem (ES) cells, which are extraordinarily receptive to homologous recombination. Mouse ES cells with targeted mutations are now available for >12,000 genes; these cells have thus become the genetic repository for the mouse. Unfortunately, the link between cultured ES cells and the production of mutant animals could not be readily established in other species. ES cell lines isolated from species other than the mouse rarely exhibit germline colonization, although recent success has been reported in the rat[3].

The lack of authentic ES cells held back targeted genome manipulation in most mammalian species for many years. This barrier was eventually overcome by exploiting the remarkable ability of a mammalian oocyte to reprogram a somatic cell nucleus, effectively converting it to a zygotic genome (**Fig. 1**). By performing gene targeting in cultured somatic cells and then using the engineered cells for nuclear transfer, it became possible to manipulate endogenous genes in several mammalian species[4]. Despite these successes, the technical difficulties have been substantial because of the low efficiencies of both gene targeting in somatic cells and subsequent reprogramming of their nuclei.

The research reported by Cui *et al.*[1] provides for the first time a route to directly manipulate the rat genome, an approach that bypasses the requirement for germline competent ES cells or somatic cell nuclear transfer (**Fig. 1**). The major difficulty in achieving gene targeting with naked DNA injected into pro-nuclei is the very low efficiency of targeted rather than random integration[2]. Gene targeting is stimulated by several orders of magnitude in somatic cells by provision of a double-strand break in the genome[5]. Cleaving the mammalian genome at a defined site was not possible before the development of ZFNs, modular proteins that couple zinc-finger DNA-binding domains to the nuclease domain of the restriction endonuclease FokI[6]. They function as homo- or heterodimers, cleaving DNA between the two binding sites (**Fig. 1**), and in principle can be designed to target any unique site in complex genomes.

After the initial demonstration of sequence-specific cleavage of the *Drosophila melanogaster* genome[7], ZFNs were shown to stimulate targeted integration of a template sequence by means of homologous recombination in fruitflies, plants and human cells. Cui *et al.*[1] and a recent report[8] both demonstrate that co-injection of a pair of ZFN mRNAs with a targeting vector into pro-nuclei stimulates the frequency of gene targeting in mice to workable levels of 2–20% (**Fig. 1**). Cui *et al.*[1] extend this approach to the rat. Interestingly, live-born founder animals obtained from these experiments[1,8] are mosaics that carry several different mutant alleles with deletions at the target locus, as well as correctly targeted alleles and unmodified wild-type alleles. Deletions are expected products following nonhomologous end-joining of cleaved DNA in the absence of targeting and have been described previously following expression of ZFNs in zygotes[9]. The transmission of multiple different mutant alleles from the same founder reflects germline mosaicism caused by expression and cleavage activity of ZFNs after DNA replication in maternal and/or paternal pro-nuclei. The ability to simultaneously generate a spectrum of mutations can be advantageous for genetic purposes.

Despite these advances, several questions about ZFN-stimulated pronuclear targeting remain to be addressed. Molecular biologists are familiar with unwanted off-target activity of restriction enzymes. To what extent do ZFNs cleave other sites in the genome? The comparatively small size of deletions generated at illegitimate sites suggests that off-target cleavages will be hard to trace. Does the physical damage of the host genome observed in many transgenic animals generated by pronuclear injection occur in this setting, too? The importance of the answers to these questions will undoubtedly depend on the frequency and type of unwanted events, their linkages to the desired genomic alterations and the context in which the technology is applied.

The mutant alleles generated using conventional gene targeting technology in mouse ES cells have become increasingly sophisticated over the past two decades. The repertoire of genetic alterations that can be achieved by ZFN-stimulated pronuclear targeting is fertile ground for exploration. Although mutant rats have recently been produced using rat ES cell technology[10] (**Fig. 1**), ZFN targeting applied directly to the zygote (**Fig. 1**) presents considerable advantages. Pronuclear injection of nucleic acids is well established, widely practiced and applicable to any strain. Moreover, transmission of the engineered allele from founder rats is readily achieved. Provided that ZFNs with the appropriate specificity can be generated, the community of rat researchers can look forward to rats with a myriad of defined genome modifications.

This technology will also find applications in a multitude of other species, which hitherto have required somatic cell reprogramming to achieve directed modifications of their genomes. Although vector-chromosome gene targeting has yet to be demonstrated in the pro-nuclei of farm animal zygotes, this is likely to be possible. ZFNs have been shown to stimulate gene targeting in a variety of species and can therefore be used in combination with somatic nuclear transfer, removing a bottleneck in achieving directed modification by this route. The promise of this technology will stimulate numerous applications, but the terms, conditions and costs associated with commercially provided ZFNs can be prohibitive and may limit their potential.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Cui, X. *et al. Nat. Biotechnol.* **29**, 64–67 (2011).
2. Brinster, R.L. *et al. Proc. Natl. Acad. Sci. USA* **86**, 7087–7091 (1989).
3. Buehr, M. *et al. Cell* **135**, 1287–1298 (2008).
4. McCreath, K.J. *et al. Nature* **405**, 1066–1069 (2000).
5. Jasin, M. *Trends Genet.* **12**, 224–228 (1996).
6. Kim, Y.G., Cha, J. & Chandrasegaran, S. *Proc. Natl. Acad. Sci. USA* **93**, 1156–1160 (1996).
7. Bibikova, M., Golic, M., Golic, K.G. & Carroll, D. *Genetics* **161**, 1169–1175 (2002).
8. Meyer, M., de Angelis, M.H., Wurst, W. & Kuhn, R. *Proc. Natl. Acad. Sci. USA* **107**, 15022–15026 (2010).
9. Geurts, A.M. *et al. Science* **325**, 433 (2009).
10. Tong, C., Li, P., Wu, N.L., Yan, Y. & Ying, Q.-L. *Nature* **467**, 211–213 (2010).

Out of harm's way

David A Williams & Adrian J Thrasher

Screening for safe harbor sites in the genome may improve the safety of gene therapy.

Successful treatment of devastating diseases of the immune system has been accomplished using high-efficiency gene transfer into hematopoietic stem and early progenitor cells with retroviral vectors. However, as emphasized by the occurrence of leukemia in several otherwise successful clinical trials, the integration of gamma-retroviral vectors containing powerful enhancers in the viral long terminal repeats carries a risk of insertional mutagenesis. In this issue, Papapetrou *et al.*[1] propose to reduce this risk by bioinformatic screening of randomly transduced cells to identify clones whose integration sites are located far from potentially dangerous regions of the genome. They demonstrate that screening insertion sites in hematopoietic cells derived from induced pluripotent stem (iPS) cells can readily identify apparently safe integrations that permit transgene expression at clinically relevant levels.

As seen in recent clinical trials for X-linked severe combined immunodeficiency, chronic granulomatous disease and Wiskott-Aldrich syndrome, activation of oncogenes by integrating viral vectors appears to provide a 'first hit' that in some cases is associated with secondary genetic events and ultimately full leukemic transformation. Although it is the potent and relatively indiscriminate properties of the gamma-retroviral long terminal repeats that ultimately do the damage, recent detailed mapping of integration sites also suggests that there may be a tethering between long terminal repeat sequences and active gene regulatory regions[2,3]. Retroviral vectors of all classes that lack long terminal repeats are likely to be safer because internal regulatory sequences can be designed to be less mutagenic, and their intrinsic integration mechanisms may be less prone to target sensitive chromosomal regions. Even so, there remains a finite risk of mutagenesis for all semi-randomly or randomly integrating vectors, as recently demonstrated in a clinical trial for β-thalassemia[4]. Strategies to mediate safe integration—and, ultimately, gene correction—are therefore of considerable interest.

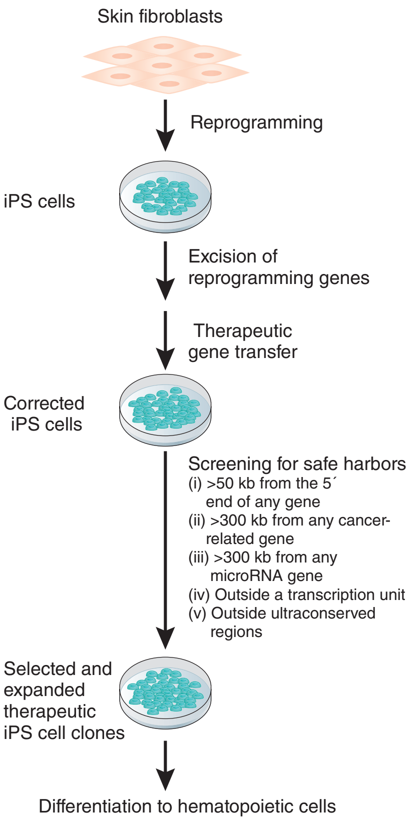

Skin fibroblasts
Reprogramming
iPS cells
Excision of reprogramming genes
Therapeutic gene transfer
Corrected iPS cells
Screening for safe harbors
(i) >50 kb from the 5´ end of any gene
(ii) >300 kb from any cancer-related gene
(iii) >300 kb from any microRNA gene
(iv) Outside a transcription unit
(v) Outside ultraconserved regions
Selected and expanded therapeutic iPS cell clones
Differentiation to hematopoietic cells

Figure 1 Schematic representation of the use of reprogramming and therapeutic gene transfer technology to select 'corrected' iPS cell clones harboring safe integrations of inserted vector sequences.

David A. Williams is at Children's Hospital Boston and Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts, USA. Adrian J. Thrasher is at the Centre for Immunodeficiency, Molecular Immunology Unit, Institute of Child Health, London, UK.
e-mail: DAWilliams@childrens.harvard.edu and a.thrasher@ich.ucl.ac.uk

Despite these advances, several questions about ZFN-stimulated pronuclear targeting remain to be addressed. Molecular biologists are familiar with unwanted off-target activity of restriction enzymes. To what extent do ZFNs cleave other sites in the genome? The comparatively small size of deletions generated at illegitimate sites suggests that off-target cleavages will be hard to trace. Does the physical damage of the host genome observed in many transgenic animals generated by pronuclear injection occur in this setting, too? The importance of the answers to these questions will undoubtedly depend on the frequency and type of unwanted events, their linkages to the desired genomic alterations and the context in which the technology is applied.

The mutant alleles generated using conventional gene targeting technology in mouse ES cells have become increasingly sophisticated over the past two decades. The repertoire of genetic alterations that can be achieved by ZFN-stimulated pronuclear targeting is fertile ground for exploration. Although mutant rats have recently been produced using rat ES cell technology[10] (**Fig. 1**), ZFN targeting applied directly to the zygote (**Fig. 1**) presents considerable advantages. Pronuclear injection of nucleic acids is well established, widely practiced and applicable to any strain. Moreover, transmission of the engineered allele from founder rats is readily achieved. Provided that ZFNs with the appropriate specificity can be generated, the community of rat researchers can look forward to rats with a myriad of defined genome modifications.

This technology will also find applications in a multitude of other species, which hitherto have required somatic cell reprogramming to achieve directed modifications of their genomes. Although vector-chromosome gene targeting has yet to be demonstrated in the pro-nuclei of farm animal zygotes, this is likely to be possible. ZFNs have been shown to stimulate gene targeting in a variety of species and can therefore be used in combination with somatic nuclear transfer, removing a bottleneck in achieving directed modification by this route. The promise of this technology will stimulate numerous applications, but the terms, conditions and costs associated with commercially provided ZFNs can be prohibitive and may limit their potential.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Cui, X. *et al. Nat. Biotechnol.* **29**, 64–67 (2011).
2. Brinster, R.L. *et al. Proc. Natl. Acad. Sci. USA* **86**, 7087–7091 (1989).
3. Buehr, M. *et al. Cell* **135**, 1287–1298 (2008).
4. McCreath, K.J. *et al. Nature* **405**, 1066–1069 (2000).
5. Jasin, M. *Trends Genet.* **12**, 224–228 (1996).
6. Kim, Y.G., Cha, J. & Chandrasegaran, S. *Proc. Natl. Acad. Sci. USA* **93**, 1156–1160 (1996).
7. Bibikova, M., Golic, M., Golic, K.G. & Carroll, D. *Genetics* **161**, 1169–1175 (2002).
8. Meyer, M., de Angelis, M.H., Wurst, W. & Kuhn, R. *Proc. Natl. Acad. Sci. USA* **107**, 15022–15026 (2010).
9. Geurts, A.M. *et al. Science* **325**, 433 (2009).
10. Tong, C., Li, P., Wu, N.L., Yan, Y. & Ying, Q.-L. *Nature* **467**, 211–213 (2010).

Out of harm's way

David A Williams & Adrian J Thrasher

Screening for safe harbor sites in the genome may improve the safety of gene therapy.

Successful treatment of devastating diseases of the immune system has been accomplished using high-efficiency gene transfer into hematopoietic stem and early progenitor cells with retroviral vectors. However, as emphasized by the occurrence of leukemia in several otherwise successful clinical trials, the integration of gamma-retroviral vectors containing powerful enhancers in the viral long terminal repeats carries a risk of insertional mutagenesis. In this issue, Papapetrou *et al.*[1] propose to reduce this risk by bioinformatic screening of randomly transduced cells to identify clones whose integration sites are located far from potentially dangerous regions of the genome. They demonstrate that screening insertion sites in hematopoietic cells derived from induced pluripotent stem (iPS) cells can readily identify apparently safe integrations that permit transgene expression at clinically relevant levels.

As seen in recent clinical trials for X-linked severe combined immunodeficiency, chronic granulomatous disease and Wiskott-Aldrich syndrome, activation of oncogenes by integrating viral vectors appears to provide a 'first hit' that in some cases is associated with secondary genetic events and ultimately full leukemic transformation. Although it is the potent and relatively indiscriminate properties of the gamma-retroviral long terminal repeats that ultimately do the damage, recent detailed mapping of integration sites also suggests that there may be a tethering between long terminal repeat sequences and active gene regulatory regions[2,3]. Retroviral vectors of all classes that lack long terminal repeats are likely to be safer because internal regulatory sequences can be designed to be less mutagenic, and their intrinsic integration mechanisms may be less prone to target sensitive chromosomal regions. Even so, there remains a finite risk of mutagenesis for all semi-randomly or randomly integrating vectors, as recently demonstrated in a clinical trial for β-thalassemia[4]. Strategies to mediate safe integration—and, ultimately, gene correction—are therefore of considerable interest.

David A. Williams is at Children's Hospital Boston and Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts, USA. Adrian J. Thrasher is at the Centre for Immunodeficiency, Molecular Immunology Unit, Institute of Child Health, London, UK. e-mail: DAWilliams@childrens.harvard.edu and a.thrasher@ich.ucl.ac.uk

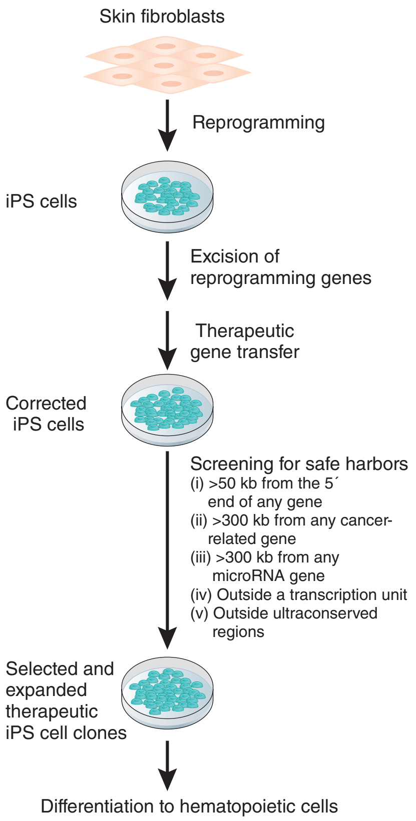

Skin fibroblasts
Reprogramming
iPS cells
Excision of reprogramming genes
Therapeutic gene transfer
Corrected iPS cells
Screening for safe harbors
(i) >50 kb from the 5´ end of any gene
(ii) >300 kb from any cancer-related gene
(iii) >300 kb from any microRNA gene
(iv) Outside a transcription unit
(v) Outside ultraconserved regions
Selected and expanded therapeutic iPS cell clones
Differentiation to hematopoietic cells

Figure 1 Schematic representation of the use of reprogramming and therapeutic gene transfer technology to select 'corrected' iPS cell clones harboring safe integrations of inserted vector sequences.

While the long-term goal of patient-specific mutation correction by homologous recombination is pursued, the use of genomic 'safe harbors' could certainly be helpful. Several laboratories have demonstrated that targeting specific loci for integration of vector sequences is associated with limited or no adverse effects on the expression of neighboring genes. This strategy is appealing because it is applicable to a wide range of diseases and applications. Most notably, targeting of the *AAVS1* locus located on chromosome 19 has been well-characterized using both zinc-finger nuclease technology to enhance homologous recombination[5–7] and Rep protein–mediated localization of AAV vectors[8]. This locus encodes the *PPP1R12C* gene, which is ubiquitously expressed, but targeting, at least with AAV vectors, is not necessarily functionally disruptive. Insertion at this site appears to provide a safe harbor with respect to genotoxicity and allows stable and long-term expression of transgenes in human and mouse embryonic stem cells[5,8]. These studies provide a 'proof of concept' for a safe harbor strategy, but application remains limited by the inefficiency of targeting in relevant primary cell types, such as hematopoietic stem cells, and dysregulated expression of transgenes when endogenous regulatory sequences are not present.

The development of induced pluripotent stem (iPS) cells[9]—a method of generating embryonic-like stem cells from somatic cells—offers a new approach that may address the limited capacity of hematopoietic stem cells to expand clonally for analysis *in vitro*. Evidence for the utility of iPS cells in correcting an inherited blood disorder was recently provided using a humanized mouse model of sickle cell anemia[10]. IPS cells generated from tail-tip fibroblasts were corrected by conventional gene targeting and successfully engrafted in the mice for disease rescue. In the context of genetic therapies, the major advantages of iPS cells include their capacity for nearly unlimited expansion *in vitro* and clonability. These characteristics allow molecular characterization of specific expandable clones. However, there are still major limitations to iPS cell technology that prevent immediate application to human disease. Further research is needed on generating transplantable cells and ensuring the integrity and safety of the derived cell products.

Papapetrou *et al.*[1] exploit iPS cell technology and propose a different approach to reducing the risks of viral vector insertional mutagenesis. The authors screen randomly generated clones according to five criteria that define the acceptable proximity of an integration site to potentially unsafe regions of the genome: transcription start sites, cancer-related genes, microRNA genes, transcription units and ultraconserved regions (**Fig. 1**). Thus, rather than target preselected genomic safe harbors, this method exploits bioinformatic analysis to screen all integration sites and identify those that are likely to be safe. In an iPS cell model for lentiviral correction of β-thalassemia, the authors succeed in generating a number of clones from patient fibroblasts of which a small—but practical—number are judged to carry safe insertions. One such clone expresses clinically relevant levels of the therapeutic transgene in differentiated cells *in vitro*.

Ultimately, the method proposed by Papapetrou *et al.*[1] requires validation in a true long-term engraftment setting in which the evolution of clonality can be monitored over time. However, the study is again a nice proof of concept that safe sites can likely be selected after gene transfer, and the technology can theoretically be applied to any disease-specific somatic cells.

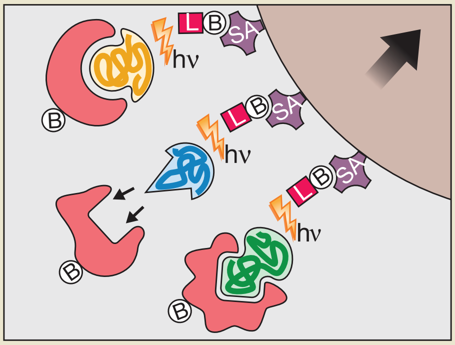
The use of gene transfer technology has now been shown to correct multiple genetic diseases and may soon become an accepted therapy for several immunodeficiency, metabolic and other diseases. Since the first reports of mutagenesis in patients, there has been remarkable progress in improving vector safety. New clinical trials are underway using modified vectors that are devoid of dangerous long terminal repeat sequences and that, in some cases, aim to replicate normal patterns of gene expression. Although it will take some time to determine the effectiveness and safety of these vectors, there is considerable cause for optimism. The development of iPS cell technology in parallel with progress in bioinformatics, as proposed by Papapetrou *et al.*[1], offers the possibility of additional novel approaches for the treatment of genetic diseases.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Papapetrou, E.P. *et al. Nat. Biotechnol.* **29**, 73–78 (2011).
2. Felice, B. *et al. PLoS ONE* **4**, e4571 (2009).
3. Cattoglio, C. *et al. Blood* published online, doi:10.1182/blood-2010-05–283523 (23 September 2010).
4. Cavazzana-Calvo, M. *et al. Nature* **467**, 318–322 (2010).
5. Hockemeyer, D. *et al. Nat. Biotechnol.* **27**, 851–857 (2009).
6. Perez, E.E. *et al. Nat. Biotechnol.* **26**, 808–816 (2008).
7. DeKelver, R.C. *et al. Genome Res.* **20**, 1133–1142 (2010).
8. Henckaerts, E. *et al. Proc. Natl. Acad. Sci. USA* **106**, 7571–7576 (2009).
9. Takahashi, K. & Yamanaka, S. *Cell* **126**, 663–676 (2006).
10. Hanna, J. *et al. Science* **318**, 1920–1923 (2007).

Aptamer-based proteomics arrays



Progress in the use of high-throughput proteomic analysis for biomarker discovery and diagnostics has been stymied by the challenge of quantifying tens of thousands of proteins whose abundance spans approximately twelve orders of magnitude. The use of mass spectrometry still poses technical difficulties and the inherent cross-reactivity of antibodies has limited the utility of antibody arrays. Gold *et al*. couple the use of slow off-rate modified aptamers (SOMAmers)—oligonucleotides containing functionalities that mimic amino acid side chains to enhance their specificity for targets—with the robustness of nucleotide arrays to provide an assay that can measure >800 human proteins in ~15 µl of human blood with low limits of detection (1 pM median) over a dynamic range from ~100 fM to 1 µM. First, proteins to be assayed (pink) bind tightly to their cognate SOMAmer, which is modified with biotin (B) and a fluorescent label (L), and bound protein-SOMAmer complexes are trapped on beads coated with streptavidin (SA). Then, as depicted, unbound proteins are washed away, and biotin-tagged bound proteins are released by exposure to UV light (hν). After a subsequent recovery step on SA-coated beads, the SOMAmers are eluted from their targets and quantified by hybridization to a customized DNA microarray. The fluorescent intensity of each probe spot is proportional to the amount of its target protein in the original sample, with the SOMAmers acting as both the binding agent and the quantifiable species. Gold *et al*. use this approach to identify 58 potential markers for chronic kidney disease. Ostroff *et al*. use the assay to analyze archived samples from >1,300 human subjects. From 44 candidate biomarkers, they identify a 12-protein panel with strong potential to diagnose non-small cell lung cancer. (*PLoS One* **5**, e15003, e15004, 2010) *PH*

Sequence-specific DNA-binding TALEs

Recent studies have mapped the relationships between the amino-acid sequences and DNA-binding specificities of transcription activator–like effector (TALE)-type transcription factors from the pathogenic plant genus *Xanthomonas*. Morbitzer *et al.* demonstrate that knowledge of this code allows the design of sequence-specific transcription factors that activate user-defined endogenous genes *in vivo* in plants. The researchers create custom TALEs that target a 19-bp sequence in the tomato promoter *Bs4S* or 19-bp sequences in the promoters of the *Arabidopsis thaliana* genes *EGL3* and *KNAT1*. Moreover, they show that TALEs targeting a 23-bp sequence have enhanced target specificity as compared to those targeting 19-bp sequences. Additional experiments enabled Morbitzer *et al.* to identify particular repeat units in TALE proteins that target G nucleotides specifically, an aspect of the binding code that had not been described previously. These results suggest that designer TALEs may represent an alternative to sequence-specific DNA targeting using zinc-finger domains. (*Proc. Natl. Acad. Sci. USA* **107**, 21617–21622, 2010) *CM*

Written by Laura DeFrancesco, Markus Elsner, Peter Hare & Craig Mak

Protein-sensing RNA control device

RNA-based molecules have been engineered to reprogram cells in response to externally applied small molecules or nucleic acids. To apply the same principles for modulating the effects of signaling events, RNA sensor-actuator devices need to be able to alter gene expression in response to changes in protein factors. Culler *et al.* have now constructed alternative splicing systems that either include or exclude exons based on the binding of a specific protein. The alternatively spliced exon contains a stop codon that prevents the translation of a downstream effector gene if included. The authors apply this strategy to the detection of the activity of the important cellular signaling molecules NF-κB and β-catenin using the expression of a fluorescent reporter as a readout. A potential application of this strategy is to selectively kill cells with overactive signaling pathways, for example, in cancer. To demonstrate the feasibility of such an approach, Culler *et al.* couple the NF-κB and β-catenin sensors to a gene encoding an enzyme that converts pro-drugs, causing 80% of the cells to undergo apoptosis in the presence of both pro-drug and pathway activation. (*Science* **330**, 1251–1255, 2010) *ME*

A peptide to get your GOAT

The appetite-suppressing gastric peptide hormone ghrelin is a promising therapeutic target for modulating weight gain and glucose control. As ghrelin needs to be acetylated with octanoate to be active, Barnett *et al.* set out to inhibit ghrelin *O*-acyltransferase (GOAT), the enzyme responsible for its activation. A fusion comprising ten ghrelin-derived amino acids, a stabilized octanoyl-CoA and the Tat motif (11 amino acids) inhibited GOAT *in vitro* and in cultured cells. Daily intraperitoneal doses of the peptide reduced weight gain, blood glucose and insulin-like growth factor 1 in normal, but not ghrelin-deficient, mice given a high-fat diet. These findings may open the way for new strategies to manage the growing incidence of obesity and type 2 diabetes in Western society. (*Science* **330**, 1689–1692, 2010) *PH*

Anti-inflammatory histone mimics

Exposure to pathogens results in complex responses that can both protect (immune response) and harm (inflammation cascade) the host. Nicodeme *et al.* present a new approach that could potentially prevent the deleterious effects of inflammation by inhibiting the formation of transcription complexes that upregulate inflammation-inducing genes. Targeting bromodomain and extra terminal domain (BET) proteins, which recruit proteins into transcription complexes, the researchers synthesized a histone mimic that binds BET proteins, keeping them from interacting with chromatin, which in turn prevents transcription complexes from forming. One synthetic inhibitor (I-BET), designed to bind peptides derived from acetylated histones, downregulated key inflammatory cytokines and chemokines when mouse macrophages were incubated with I-BET before treatment with lipopolysaccharide (LPS). This effect was specific for inflammation, as cytokines not induced by LPS were unaffected. I-BET treatment not only prevented complex formation but also prevented acetylation of histones on BET-sensitive promoters, though it is unclear whether I-BET inhibits acetylases directly or inhibits recruitment of acetylases to the transcription complex. Finally, the researchers showed that I-BET works *in vivo*. Injecting I-BET into mice prevented sepsis and death when given before LPS-induced shock, as well as when it was given after the signs of inflammation began to appear. In addition, I-BET prevented death in mice suffering from peritonitis and sepsis caused by cecal ligation. (*Nature*, published online, doi:10.1038/nature09589, 10 November 2010) *LD*

Trends in computational biology—2010

H Craig Mak

Interviews with leading scientists highlight several notable breakthroughs in computational biology from the past year and suggest areas where computation may drive biological discovery.

The field of computational biology encompasses a set of investigative tools as much as being a research endeavor in its own right. It is often difficult to gauge the utility and significance of a computational tool, at least until the research community has had sufficient time to explore, exploit and hone it in various applications. In an effort to identify recent notable breakthroughs in the field of computational biology, *Nature Biotechnology* surveyed leading researchers in the area, asking them to nominate papers of particular interest published in the previous year that have influenced the direction of their research. Some of the nominated papers had been featured in our pages and elsewhere; others were completely off our radar. Although we surveyed a small group of 15 scientists, the nominated papers (**Box 1**) provide a snapshot of some of the most exciting areas of current computational biology research.

All the papers featured in the following pages were nominated by at least two scientists. Our analysis not only highlights the richness of approaches and growth of the field, but also suggests that researchers of a particular type are driving much of cutting-edge computational biology (**Box 2**). Read on to find out what characterizes them and what they've been doing in the past year.

Next-generation sequence analysis

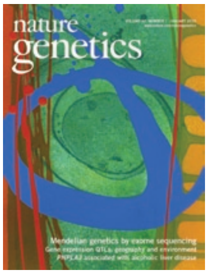
Imagine an experiment generating a billion data points every day, the equivalent of running millions of agarose gels—and taping (remember that!) the pictures into tens of thousands of laboratory notebooks, or hybridizing thousands of gene-expression microarrays. Computational biology has risen in prominence in recent years largely because of the increase in the data-generation capacity of high-throughput technology. More data create more opportunities and a more pressing need for systematic methods of analysis. And nowhere has that need been more evident than in the field of next-generation sequencing.

The latest sequencers take a week or two to generate about a billion short reads, stretches of about 50–400 bp of DNA sequenced from a longer molecule. Researchers face challenges on two levels when turning massive collections of reads into biologically meaningful information. The first set of challenges lies in processing the reads themselves: mapping them to their genomic locations, and then assembling them into longer contiguous stretches of DNA. The second set of challenges lies in interpreting large collections of reads, which may be assembled into whole genomes, to understand the functional effects of genetic variation. Thousands of genomes from humans, plants, animals and disease tissues have already been sequenced—and all are in need of better interpretation. Although algorithms, such as BLAST for searching and CLUSTALW for aligning, continue to be the workhorses of sequence analysis, several next-generation computational methods have emerged to cope with the DNA sequences captured in billions of short reads and thousands of genomes.

H. Craig Mak is Associate Editor, Nature Biotechnology

The advance. Two methods for *de novo* transcriptome assembly of short reads were published this year from Lior Pachter and colleagues[1] and from Aviv Regev and colleagues[2]. The transcriptome can be analyzed by sequencing cDNA reverse transcribed from RNA (RNA-Seq), but mapping and assembling the resulting reads are challenging owing to the complexities introduced by RNA splicing. The two methods are the first that robustly assemble full-length transcripts, including alternative splicing isoforms. In contrast to previous approaches, these two methods first map reads to the genome using software that takes possible splice junctions into account, thereby making assembly more manageable. Then, they apply graph-based algorithms to determine[1,2] and quantify[1] the most likely splice isoforms. The algorithms were applied to mammalian transcriptomes to follow global patterns of splicing during a developmental time course[1] and to identify novel, spliced, long, noncoding RNAs that had not been annotated by existing methods[2].

Progress toward the second challenge of genome interpretation was reported in papers that demonstrate the potential of genome sequencing for genetic analysis of human traits. The approach, pioneered by Jay Shendure at the University of Washington in Seattle, sequences the exonic regions of several genomes to identify protein-disrupting mutations linked to disease. "This study dramatically demonstrates how we can make new genetic discoveries by sequencing all the exons in a set of patients," says Steven Salzberg of the University of Maryland and a co-author with Pachter[1]. Since the publication of the first success of this strategy in January 2010 (ref. 3), several additional studies have taken a similar approach to study the genetics of human diseases.



What it means. Advances in transcript assembly from RNA-Seq data should allow alternative splicing to be studied genome-wide across

Steven Salzberg collaborated with Lior Pachter and Barbara Wold to develop a method for assembling short reads sequenced from cDNA into full-length spliced transcripts.

many biological conditions, such as in different tissues, over different time points and in response to genetic and chemical perturbations. In addition, RNA-Seq is now poised as a tool for discovering new RNAs that we may not have even known were transcribed from a genome. Armed with better knowledge of splicing patterns and comprehensive transcript catalogs, it should be possible to improve the annotation of genomes. "Thousands of people have accessed our software," says Salzberg, "and it is being integrated into easy-to-use graphical interfaces such as Galaxy"[4].

The flood of whole genomes and exomes should also drive sequence-analysis methods. Attending an International Cancer Genome Consortium meeting in Brisbane, Australia, in December, Debbie Marks of Harvard University noted, "We're still in the early days of whole-genome analysis…problems need to be articulated in ways that are computationally tractable." Many of the problems will require algorithmic advances, such as better *de novo* assembly of transcriptomes and genomes. However, much of the work that goes into interpreting a patient's whole genome to identify disease-causing mutations, for instance, involves filtering variants identified by sequencing against databases of known variants. As better databases should lead to improved genome analysis, it might be reasonable to expect the development and mining of biomedical databases to be a fertile source for computational advances.

1. Trapnell, C. *et al. Nat. Biotechnol* **28**, 511–515 (2010).
2. Guttman, M. *et al. Nat. Biotechnol.* **28**, 503–510 (2010).
3. Ng, S. *et al. Nat. Genet.* **42**, 30–35 (2010).
4. Goecks, J. Nekrutenko, A. & Taylor, J. *Genome Biol.* **11**, R86 (2010).

ORIGINAL PAPER

PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations

Discovery from data repositories

Electronic medical records are becoming a reality, promising lower health-care costs, improved patient treatments and, perhaps, scientific advances. As a result of incentives built into the Health Information Technology for Economic and Clinical Health (HITECH) Act passed as part of the Obama administration's health-care reform, in 10 years almost every US hospital could be using electronic medical records, up from 1.5–2% of hospitals today. "What's fun to think about," says Atul Butte of Stanford University, "is what kind of science can we derive from this?"

Electronic medical records can contain a wealth of history on physical exams and treatment regimes. Particularly amenable to automated analysis in the records are the standardized administrative billing codes used to charge for each procedure, test or clinical visit. These codes can track anything from diagnosis of coronary artery disease to the procedure for inserting a stent to keep blood vessels open. Several hospitals, with Vanderbilt University (Nashville, TN, USA) among the leaders, are pairing electronic medical records with the collection of tissue samples from every patient treated. These resources represent an unprecedented source of data on the genetic and physiological state of people linked to standardized, computable records of their phenome, or the set of all phenotypes including disease diagnosis and responses to treatment.

Atul Butte: "Ninety-nine percent of the work is not in software engineering or coding, it's in coming up with the right kind of question:…[one that] no one even realizes we can ask today."

The advance. Last year, Joshua Denny and colleagues at Vanderbilt University published the first study that demonstrates the feasibility of associating genetic modifications with data on phenotypic traits mined from electronic medical records[1]. The approach, which they called PheWAS (for phenome-wide association scans), is akin to the genome-wide association studies (GWAS) widely used today to find single-nucleotide polymorphisms (SNPs) that are genetically linked in a population to a particular disease trait—except that PheWAS is GWAS in reverse. GWAS associates genotypes with a given phenotype, such as height or a genetic disease. In contrast, PheWAS attempts to determine the range of clinical phenotypes associated with a given genotype.

The Vanderbilt group analyzed the medical records of ~6,000 patients who had been tested to see whether they carried a total of five SNPs previously associated with seven diseases (coronary artery disease, carotid artery stenosis, atrial fibrillation, multiple sclerosis, lupus, rheumatoid arthritis and Crohn's disease). To identify patient phenotypes in an automated fashion, they used billing codes in the electronic medical records to group patients into 'case' and 'control' populations for 776 phenotypes. Finally, a Chi-squared statistical test was used to evaluate whether patients harboring a specific SNP also tended to display a particular phenotype. The authors noted that, although there are many statistical challenges with this kind of analysis and there is much room for improvement of their method, four of the seven previously known disease-gene associations could be replicated, and several potential associations with other diseases were identified but not rigorously validated. These results highlight the possibility that novel biological discoveries might be made using this approach.

What it means. "Everyone wishes they could do this kind of study," remarks Butte, "but it represents a multimillion dollar investment. Vanderbilt is leading the way." Several other hospitals are making similar investments. The Mayo Clinic, for example, is coupling specimens collected from 20,000 patients with electronic medical records and other data gathered and standardized across the hospital system. "There has always been a question about whether electronic medical records would be of sufficient quality to allow genetic discovery," says Russ Altman, also at Stanford. "This paper sets the stage for widespread use of electronic records for genomic discovery."

More generally, the case of electronic medical records illustrates the potential value locked within unique biomedical databases and the challenges of realizing that value. For instance, a paper describing the PubChem BioAssay database[2] has caught the attention of several survey respondents. PubChem is the repository for small-molecule screening data generated by several NIH programs, and it receives similar data from many other organizations. "PubChem will become a key technology, in a manner similar to how freely available sequence databases in the 1990s enabled a generation

An overview of the PubChem BioAssay resource of computer-literate biologists to change the way biology is done," says Iain Wallace, a postdoctoral fellow in Gary Bader's group at the University of Toronto, which has been active in the development of databases of protein interactions. PubChem, which is funded by the NIH, brings to academics data that until now have been accessible primarily only to those in deep-pocketed pharmaceutical companies.

In the case of PubChem or Vanderbilt's electronic medical record database, careful statistical analyses will be required to robustly analyze these potential treasure troves of information. But rather than the algorithmic advances typically pursued in computational biology, according to Butte, "Ninety-nine percent of the work is not in software engineering or coding; it's in coming up with the right kind of question: given this data set, what question are we newly able to ask that everyone would love to know the answer to, but no one even realizes we can ask today?" Exposure is key, says Butte, "What I would love to see is a computational person going to surgical grand rounds at a hospital to figure out what the unsolved questions are, hearing about this tumor that spreads like crazy and saying, 'I can solve this problem computationally.' That would be the ideal." Unlike problems requiring clever new algorithms or massive clusters of computers, increasing exposure may be a particularly manageable challenge facing the field.

1. Denny J.C. *et al. Bioinformatics* **26**, 1205–1210 (2010).
2. Wang, Y. *et al. Nucleic Acids Res.* **38** database issue, D255–D266 (2010).

Learning to see

Determining the distribution of probes between different subcellular locations through automated unmixing of subcellular patterns

Why have computer scientists long endeavored to create software capable of accomplishing tasks humans can already do? In the case of biological research, one advantage of computational analysis is automation and fidelity. Whereas a trained person can look at one confocal microscope image and readily identify where a fluorescently labeled protein is localized in the cell, that person cannot hope to analyze the millions of images that can be gathered with automated technology. And even if several people were enlisted to the task, each may interpret the same image in different ways. This problem provides an apt introduction to machine learning, a technology that is finding success in biology.

Robert Murphy thinks that when computer science and biology come together "inside one person's head, that is a much more efficient process."

In machine learning, computer programs are trained to pick out patterns, which may be predefined by human supervisors or learned by the program directly from data. Such 'unsupervised' machine-learning tasks are often the hardest, in part because there are many possibilities for the computer to consider. Notably, many machine-learning tasks in disparate problem domains can be articulated using a common set of concepts. In this way, techniques developed for one problem, say mining data from text, can inspire solutions to other problems.

The advance. Robert Murphy and colleagues[1,2] at Carnegie Mellon University devised machine-learning algorithms that could accurately classify whether a pattern of fluorescent staining represents localization to one subcellular organelle or to a mixture of locations. Moreover, this 'pattern unmixing' can be done in an unsupervised way, without introducing bias from a human who predefines the categories. The need for this method is supported by studies in yeast in which up to a third of all fluorescently tagged proteins appeared to localize to several places in the cell.

The key to the approach is to segment an image into objects or shapes with quantifiable features. Then a pattern of objects can be defined as the probability that certain objects are found together. The best-performing algorithm identified patterns of objects using a technique called latent Dirichlet allocation, which has been successfully used to identify patterns of words representing conceptual topics from text documents. By analogy, visual objects representing the nucleus or Golgi apparatus are 'words' in an image, and patterns of protein localization that characterize the content of an image correspond to sets of words that co-occur in documents and define the topics in the text.

What it means. "This represents the first step toward a new way of thinking about interpreting images that is generative rather than descriptive," says Murphy. Whereas a descriptive approach may take an image of a cell expressing fluorescently tagged protein and tell you that the protein is in the nucleus, a generative approach builds a model that can produce images that look like other images, and in the process of building that model (that is, determining the parameters of the model), you learn about what characterizes a pattern in a way that is meaningful across a variety of situations. For instance, a drug in a screening assay may cause a protein to partially redistribute from one subcellular location to another, but given that organelles may look different in different cell types, without Murphy's approach, if the same screen is done on a different cell type, it is difficult to know that the same process is occurring. Machine learning has been previously applied to biology, but recent increases in the data-generation capacity of technology suggest that these kinds of approaches may play a growing role in biological discovery in the future.

Does this mean that more collaboration needs to occur between biologists and computer scientists classically trained in machine learning? Not necessarily, according the Murphy. "That's been going on for a long time already. In fact, there is a group of people who are knowledgeable in many of these different domains. There are people who in general may not push the frontier of computer science, but who use state-of-the-art techniques, and in some cases do end up pushing frontiers and identifying new problems that others in the field can then solve." The role of computational biologists is to be able to straddle domains. Murphy continues: "When the field started, it often grew by adventitious 'collisions' between computer scientists and biologists—over lunch, at a faculty meeting. That is a very inefficient way of moving forward. When those collisions can happen inside one person's head, that is a much more efficient process."

1. Coelho, L.P. *et al. Bioinformatics* **26**, i7–i12 (2010).
2. Peng, T. *et al. Proc. Natl. Acad. Sci. USA* **107**, 2944–2949 (2010).

CompBio 2.0

nature
CHEMISTRY IN MOTION
Attosecond snapshots reveal dynamics of valence electrons
150 as
HEART DISEASE
Scanning for serum lipid genes
CITIZEN SCIENCE
Crowd-based supercomputing
OCEAN MONITORING
Tap into submarine cables
NATUREJOBS
Marine biology

Businesses and broad segments of society have recently embraced decentralized mechanisms of information processing based on interactions among large groups of people. This advance in computing has not relied on new algorithms or clever data structures in the traditional sense

Box 1 Survey results

Thirty-three papers were nominated covering genomics, imaging, databases and data sets, protein-structure prediction, synthetic biology, genetics, antibody screening, systems biology and pharmacology. The 24 papers not discussed in the article are listed below.

Akavia, U.D. *et al.* An integrated approach to uncover drivers of cancer. *Cell* **143**, 1005–1017 (2010).

Ashley, E.A. *et al.* Clinical assessment incorporating a personal genome. *Lancet* **375**, 1525–1535 (2010).

Bandyopadhyay, S. *et al.* A human MAP kinase interactome. *Nat. Methods* **7**, 801–805 (2010).

Barash, Y. *et al.* Deciphering the splicing code. *Nature* **465**, 53–59 (2010).

Berger, S.I., Ma'ayan, A. & Iyengar, R. Systems pharmacology of arrhythmias. *Sci. Signal.* **3**, ra30 (2010).

Carro, M.S. *et al.* The transcriptional network for mesenchymal transformation of brain tumours. *Nature* **463**, 318–325 (2010).

Coulet, A., Shah, N.H., Garten, Y., Musen, M. & Altman, R.B. Using text to build semantic networks for pharmacogenomics. *J. Biomed. Inform.* **43**, 1009–1019 (2010).

Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D.B. Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8**, e1000294 (2010).

Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28**, 817–825 (2010).

Gibson, D.G. *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**, 52–56 (2010).

Lestas, I., Vinnicombe, G. & Paulsson, J. Fundamental limits on the suppression of molecular fluctuations. *Nature* **467**, 174–178 (2010).

Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

McGary, K.L. *et al.* Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc. Natl. Acad. Sci. USA* **107**, 6544–6549 (2010).

McLean, C.Y. *et al.* GREAT improves functional interpretation of *cis*-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).

Mungall, C.J. *et al.* Integrating phenotype ontologies across multiple species. *Genome Biol.* **11**, R2 (2010).

Pandey, G. *et al.* An integrative multi-network and multi-classifier approach to predict genetic interactions. *PLOS Comput. Biol.* **6**, e1000928 (2010).

Patel, C.J., Bhattacharya, J. & Butte, A.J. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS ONE* **5**, e10746 (2010).

Peng, H., Ruan, Z., Long, F., Simpson, J.H. & Myers, E.W. V3D enables real-time 3D visualization and quantitative analysis of large-scale biological image data sets. *Nat. Biotechnol.* **28**, 348–353 (2010).

Ravasi, T. *et al.* An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* **140**, 744–752 (2010).

Reddy, S.T. *et al.* Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat. Biotechnol.* **28**, 965–969 (2010).

Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).

Shaw, D.E. *et al.* Atomic-level characterization of the structural dynamics of proteins. *Science* **330**, 341–346 (2010).

Voelz, V.A., Bowman, G.R., Beauchamp, K. & Pande, V.S. Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1–39). *J. Am. Chem. Soc.* **132**, 1526–1528 (2010).

Zhang, C. *et al.* Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. *Science* **329**, 439–443 (2010).

of computational breakthrough but rather has been fueled by new communication media and tools, typically accessed online through websites and mobile devices. The social network Facebook (http://www.facebook.com/) has evolved from a platform solely devoted to keeping up with friends into a business that will generate an estimated $1.1 billion dollars in 2010. And notably, there is still room for specialized social networks, such as Jumo (http://www.jumo.com/), a fledgling effort geared toward philanthropy.

There have been some successes in adopting distributed computing tools and social networks into biological research. Social bookmarking tools such as del.icio.us (http://www.delicious.com/), CiteULike (http://www.citeulike.org) or Connotea (which was created by Nature Publishing Group; http://www.connotea.org/) allow users to tag and share papers and have existed for several years. Patients with the same medical conditions can connect with one another on social networks run by companies like PatientsLikeMe (http://www.patientslikeme.com/) or CureTogether (http://curetogether.com/). And for ~10 years, the Folding@home project (http://folding.stanford.edu/) has been leveraging participants' desktop computers or gaming consoles to study protein folding. But how else can computational paradigms that have permeated broader society be harnessed to drive scientific discovery?

The advance. Two papers[1,2] identified by our survey respondents highlight the potential impact of 'nontraditional' computing advances. The first is FoldIt, a multiplayer online game for predicting protein structures. David Baker and colleagues at the University of Washington in Seattle created a Web-based graphical interface that allowed players to manipulate a protein structure as if they were solving a visual puzzle[1]. This harnessed humans' spatial reasoning skills to improve computational predictions of the most likely protein conformation. Players competed against one another and were ranked on a scoreboard. When protein structures derived by FoldIt players were compared against structures predicted by a traditional computational approach, FoldIt predictions were as good or better in seven of ten test cases.

Web-Based, Participant-Driven Studies Yield Novel Genetic Associations for Common Traits

In another approach[2], researchers at Columbia University and Stanford collaborated with the consumer genetics testing company 23andMe (http://www.23andme.com/) to identify associations between genetic markers and human traits. What's notable in this study is that trait data were collected through Web-based surveys completed by consumers whose genetics had been analyzed by the company. Twenty-two traits, ranging from hair and eye color to the ability to smell the urinary metabolites of asparagus, were studied in nearly 10,000 people of northern European ancestry. The study identified single-nucleotide polymorphisms known to be associated with six of the traits, as well as novel associations with four traits. Pitfalls may be associated with approaches such as this, however, including recent concerns raised over the concordance between results of genetic tests conducted by different direct-to-consumer companies.

What it means. "Social networks need to appeal to people's selfish side," says Andrew Su, Associate Director of Bioinformatics at the Genomics Institute of the Novartis Research Foundation (San Diego, CA, USA), whose group has developed collaborative scientific tools for use publicly and within Novartis (Basel). "There needs to be some personal value derived from social networking; otherwise, where's the motivation to participate?" Arguably, the gaming aspect of FoldIt appealed to participants' competitive juices. In the 23andMe study, participants had already received their genetic data from the company, and the Web surveys served to increase the value of those data. The key, then, is to

Box 2 Cross-functional individuals

In the course of compiling this survey, several investigators remarked that it tends to be easier for computer scientists to learn biology than for biologists to learn computer science. Even so, it is hard to believe that learning the central dogma and the Krebs cycle will enable your typical programmer-turned-computational-biologist to stumble upon a project that yields important novel biological insights. So what characterizes successful computational biologists?

George Church, whose laboratory at Harvard Medical School (Cambridge, MA, USA) has a history of producing bleeding-edge research in many cross-disciplinary domains, including computational biology, says, "Individuals in my lab tend to be curious and somewhat dissatisfied with the way things are. They are comfortable in two domains simultaneously. This has allowed us to go after problems in the space between traditional research projects." A former Church lab member, Greg Porreca, articulates this idea further: "I've found that many advances in computational biology start with simple solutions written by cross-functional individuals to accomplish simple tasks. Bigger problems are hard to address with those rudimentary algorithms, so folks with classical training in computer science step in and devise highly optimized solutions that are faster and more flexible."

An overarching theme that also emerges from this survey suggests that tools for computational analyses permeate biological research according to three stages: first, a cross-functional individual sees a problem and devises a solution good enough to demonstrate the feasibility of a type of analysis; second, robust tools are created, often utilizing the specialized knowledge of formally trained computer scientists; and third, the tools reach biologists focused on understanding specific phenomena, who incorporate the tools into everyday use. These stages echo existing broader literature on disruptive innovations[1] and technology-adoption life cycles[2,3], which may suggest how breakthroughs in computational biology can be nurtured.

[1] Christiansen, C.M. & Bower, J.L. Disruptive technologies: catching the wave. *Harvard Business Review* (1995).

[2] Moore, G.A. *Crossing the Chasm: Marketing and Selling High-Tech Products to Mainstream Customers* (HarperBusiness, 1999).

[3] Rogers, E.M. *Diffusion of Innovations* (Free Press, 2003).

discover how to incentivize individuals in such a way that they support scientific discovery. One possibility is being tested by InnoCentive (partnering with Nature Publishing Group; http://www.innocentive.com/), which allows participants to pose scientific problems and offer cash prizes to other participants who provide a solution.

As in real life, different types of social interactions may justify different social networks, such as LinkedIn (http://www.linkedin.com/) for professional networking, which has thrived, even in the shadow of more general-purpose larger networks like Facebook. Several research-oriented efforts have been started, such as Sage Bionetworks (http://sagebase.org/), whose CEO, Stephen Friend, predicted earlier this year the coming obsolescence of "hunter-gatherer approaches, where large groups collect massive clinical and genomic information and expect that they as the data generator will be the data analyzer" (http://www.xconomy.com/national/2010/01/06/five-biotechnologies-that-will-fade-away-this-decade/). The two studies discussed above demonstrate successful applications of alternative paradigms for data analysis and data generation. When recruiting expertise to create these kinds of platforms, says Su, "it's hard to find people who have really traversed both computer science and biology. Discovery-oriented computational biologists with experience working on collaborative projects involving experimental scientists are particularly valuable."

[1] Cooper, S. *Nature* **466**, 756–760 (2010).

[2] Eriksson, N. *PLoS Genet.* **6**, e1000993 (2010).

nature biotechnology

Whole-genome molecular haplotyping of single cells

H Christina Fan[1], Jianbin Wang[1], Anastasia Potanina[2] & Stephen R Quake[1–3]

Conventional experimental methods of studying the human genome are limited by the inability to independently study the combination of alleles, or haplotype, on each of the homologous copies of the chromosomes. We developed a microfluidic device capable of separating and amplifying homologous copies of each chromosome from a single human metaphase cell. Single-nucleotide polymorphism (SNP) array analysis of amplified DNA enabled us to achieve completely deterministic, whole-genome, personal haplotypes of four individuals, including a HapMap trio with European ancestry (CEU) and an unrelated European individual. The phases of alleles were determined at ~99.8% accuracy for up to ~96% of all assayed SNPs. We demonstrate several practical applications, including direct observation of recombination events in a family trio, deterministic phasing of deletions in individuals and direct measurement of the human leukocyte antigen haplotypes of an individual. Our approach has potential applications in personal genomics, single-cell genomics and statistical genetics.

The sequencing of the human reference genome and the development of high-throughput short-read sequencing technologies have enabled partial decoding of an increasing number of individual human genomes[1–7]. However, all of these 'personal genomes' are incomplete, and should essentially be regarded as rough draft genomes. Although they all suffer from imperfections, such as gaps, miscalled bases and difficulties in determining large-scale structural variation, they are missing fundamental information of the unique haploid structure of homologous chromosomes. Haplotypes, the combinations of alleles at multiple loci along a single chromosome, are difficult to measure with current technologies but are an essential feature of the genome. A simple example of how the lack of this information limits the interpretation of existing genomes is to consider an individual having two mutations in a certain gene. If both mutations are on the same allele, then this individual would have one normal (that is, putatively functional) version of the gene and one mutated version. If the mutations are on different alleles, this individual would have two mutated versions and no normal version of the protein. In the absence of haplotype information, it is impossible to distinguish between these two cases.

Knowledge of complete haplotypes of individuals (personal haplotypes) would therefore be useful in personalized medicine. Notably, several studies have linked specific haplotypes to drug response and to resistance or susceptibility to diseases. A well-known example is the association of human leukocyte antigen (HLA) haplotypes with autoimmune diseases and clinical outcomes in transplantations[8–10]. Haplotypes within the apolipoprotein gene cluster may influence plasma triglyceride concentrations and the risk toward atherosclerosis[11]. Research suggests that a specific β-globin locus haplotype is associated with better prognosis of sickle cell disease[12], and other studies have linked haplotypes in the matrix metalloproteinase gene cluster with cancer development[13]. Haplotypes are also important in pharmacogenomics, an example being the association of β-2 adrenergic receptor to responses to drug treatment of asthma[14]. Deterministic haplotyping may greatly increase the power of genome-wide association studies in finding candidate genes associated with common but complex traits. It will also contribute to the understanding of population genetics and historical human migrations and the study of *cis*-acting regulation in gene expression.

Direct experimental determination of the haplotypes of an individual is challenging. The International HapMap Consortium has performed extensive SNP genotyping on different human populations, and by using family trios and statistical methods, has been able to catalog commonly occurring haplotype blocks in the human populations. However, in the best cases, when members of a family trio are analyzed, this approach leads to errors in resolving haplotype at approximately every ~3–8 megabases, and in the most general case, when an individual genome analyzed in the absence of family information, errors occur every 300 kilobases[15,16]. In the context of personalized genomics and medicine, the approaches used in the HapMap project have limited applicability, as materials from family members are not always available and computational approaches using statistical models have inherent statistical uncertainty and are limited to regions with strong linkage disequilibrium. Mate-pair shotgun genome sequencing has been demonstrated to achieve partial haplotype reconstruction of an individual but the haplotype blocks have limited sizes[5,17]. Other techniques have been demonstrated, including PCR in various forms[18–22], atomic force microscopy with carbon nanotubes[23], fosmid/cosmid cloning[24] and hybridization of probes to single DNA molecules[25]. Weaknesses of these methods include the inability to phase SNPs (that is, determine their relative arrangement on homologous chromosomes) more than tens of kilobases apart and/or a limitation in the number of markers that could be phased in a single assay. Whole-genome haplotyping can in principle be achieved by chromosome microdissection[26] or by the construction of somatic cell hybrids[27]. Yet the former is time-consuming and expensive, and the latter requires specialized and expensive equipment. So far, direct

[1]Department of Bioengineering, Stanford University, Stanford, California, USA. [2]Howard Hughes Medical Institute, Stanford University, Stanford, California, USA. [3]Department of Applied Physics, Stanford University, Stanford, California, USA. Correspondence should be addressed to S.R.Q. (quake@stanford.edu).

Received 5 October; accepted 24 November; published online 19 December 2010; doi:10.1038/nbt.1739

whole-genome haplotyping has not been accomplished for any individual. Here we address these issues using microfluidics.

RESULTS

Single-cell chromosome separation and amplification

We developed an approach termed direct deterministic phasing (DDP) in which the intact chromosomes from a single cell are dispersed and amplified on a microfluidic device (**Fig. 1**). The device consists of a cell-sorting region, where a single metaphase cell is identified microscopically and captured from a cell suspension; a chromosome release region, where metaphase chromosomes are released by protease digestion of the cytoplasm; a chromosome partitioning region, where the chromosome suspension is randomly separated into 48 partitions of a long narrow channel; an amplification region, where isolated chromosomes are individually amplified by multiple strand displacement amplification; and a product retrieval region, where amplified products are collected. The products are recovered independently, thus allowing direct genetic interrogation and genome-wide determination of haplotypes without the need for family information or statistical inference.

Whole-genome haplotyping of members in a HapMap CEU trio

We first verified DDP with three lymphoblastoid cell lines, GM12891, GM12892 and GM12878, representing a father-mother-daughter trio in the CEPH European (CEU) 1463 family. These cell lines have been extensively genotyped by the HapMap project.

For each single-cell experiment, the chromosomal origins of the contents of each microfluidic chamber were established by a 46-loci Taqman genotyping PCR. In this stage of metaphase, the chromosomes have duplicated but sister chromatids are still bound together at the centromere; therefore each metaphase cell has 46 separable chromosomes and no more than two chambers should contain templates for a given PCR genotyping assay. As expected, for assays that yielded PCR signals in two chambers, the alleles for both chambers matched that of the genomic DNA if the individual was homozygous for the tested locus, and the alleles of the two chambers were different if the individual was heterozygous for the tested locus (**Fig. 2**). There was no obvious bias in the distribution and no chromosome pairs were particularly difficult to separate. Because the chromosomes are randomly dispersed into chambers, it is possible that both homologous copies of a chromosome will co-locate in the same chamber. This probability can be made arbitrarily small by increasing the number of chambers, and in practice when co-location occurs we simply repeat the experiment with another cell.

Products from multiple chambers were pooled together into two mixtures such that each mixture contained one of the two homologous copies of most chromosomes. The two 'haploid' mixtures were separately genotyped on whole-genome genotyping arrays (Illumina's HumanOmni1-Quad BeadChip). For each individual, three to four single-cell experiments were performed, and each homologous chromosome had, on average, ~2 to 3 biological replicates. Phases were established for ~87.9%, ~89.9% and ~83.8% of ~970,000 refSNPs present on the array for GM12878, GM12891 and GM12892, respectively (**Fig. 2** and **Supplementary Data Sets 1–3**). By counting the number of inconsistent allele calls among biological replicates of each chromosome homolog, we estimated the error originating from amplification and genotyping for a single phase measurement to be 0.2–0.4%. The actual phasing error per SNP was much smaller because the final phases of most SNPs were determined by the consensus among replicates (**Supplementary Fig. 1**) and can be made as small as desired by increasing the number of replicates.

We compared our experimental phasing data of the child (GM12878) with haplotype data available from the HapMap project. In the HapMap

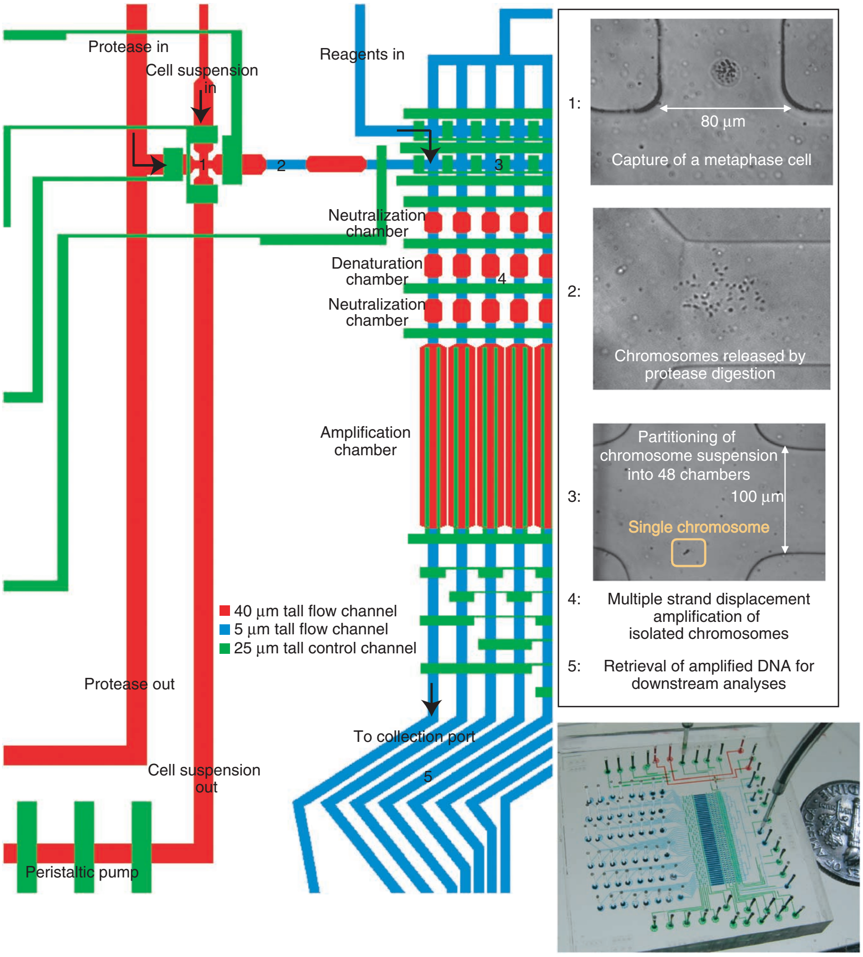

Protease in
Cell suspension in
1
2
Reagents in
3
Neutralization chamber
Denaturation chamber
4
Neutralization chamber
Amplification chamber
🟥 40 µm tall flow channel
🟦 5 µm tall flow channel
🟩 25 µm tall control channel
Protease out
To collection port
Cell suspension out
5
Peristaltic pump
1:
80 µm
Capture of a metaphase cell
2:
Chromosomes released by protease digestion
3:
Partitioning of chromosome suspension into 48 chambers
100 µm
Single chromosome
4:
Multiple strand displacement amplification of isolated chromosomes
5:
Retrieval of amplified DNA for downstream analyses

Figure 1 Microfluidic device designed for the amplification of metaphase chromosomes from a single cell. A single metaphase cell is recognized microscopically and captured in region 1. Protease (pepsin at low pH) is introduced to generate chromosome suspension in region 2. Chromosome suspension is partitioned into 48 units (region 3). Content in each partition is individually amplified (region 4). Specifically, chromosomes at low pH are first neutralized and treated with trypsin to digest chromosomal proteins. Chromosomes are denatured with alkali and subsequently neutralized for multiple strand displacement amplification to take place. As reagents are introduced sequentially into each air-filled chamber, enabled by the gas permeability of the device's material, chromosomes are pushed into one chamber after the next and finally arrive in the amplification chamber. Amplified materials are retrieved at the collection ports (region 5). In the overview image of the device, control channels are filled with green dye. Flow channels in the cell-sorting region and amplification region are filled with red and blue dyes, respectively.

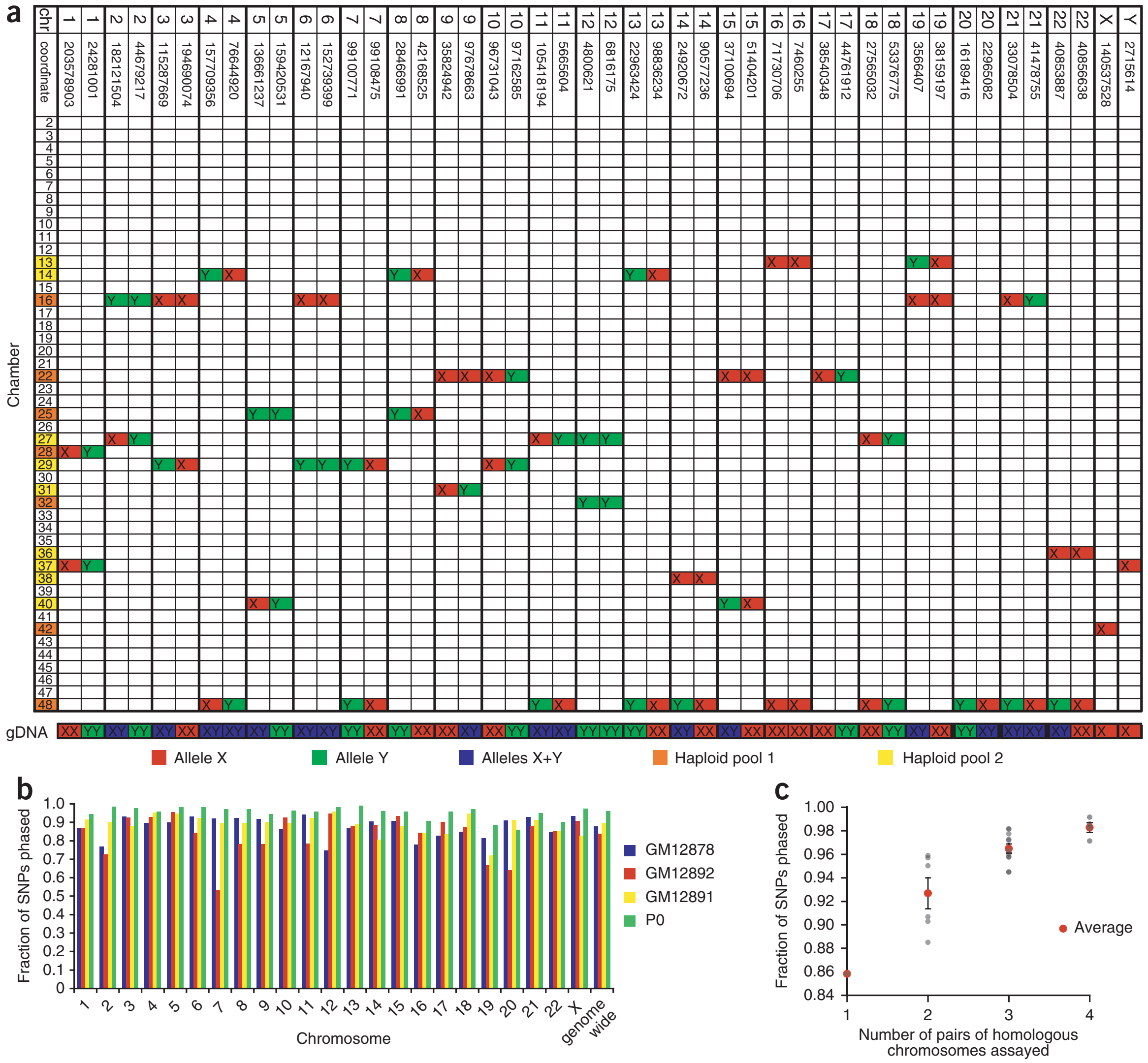


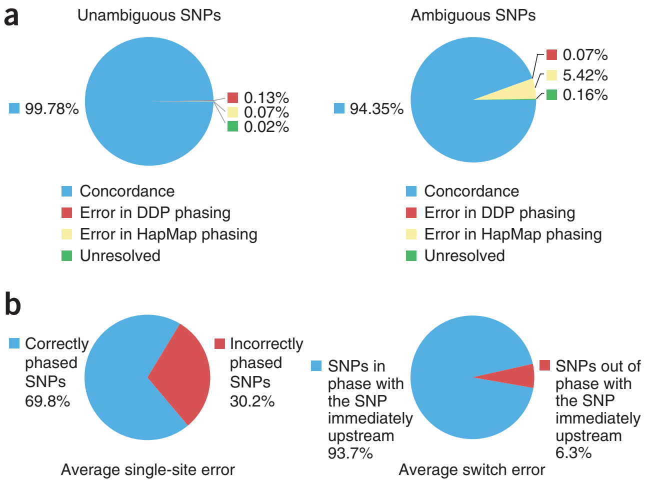
Figure 2 Whole-genome haplotyping. (**a**) Determining the chromosomal origin of amplification products in a microfluidic device using 46-loci PCR. This table represents results from an experiment using a single metaphase cell of P0's cultured whole blood. A row represents the content inside a chamber on the microfluidic device, and a column represents a locus, with specified chromosome and coordinate (NCBI Build 36.1). Each locus, except those on chromosomes 17 and 20, was found in two chambers. The two alleles of a SNP are highlighted in red and green. Heterozygous loci are labeled in blue. Chamber numbers labeled yellow were pooled together and genotyped on one HumanOmni1-Quad array, and chamber numbers labeled orange were pooled together and genotyped on another array. Genomic DNA extracted from cultured whole blood was also tested with the same 46-loci PCR. (**b**) Statistics of whole-genome haplotyping. The fraction of SNPs present on the array phased for each chromosome of each individual (GM12891, GM12892, GM12878 and a European individual 'P0') is shown as a colored bar. (**c**) Fraction of SNPs phased as a function of the number of pairs of homologous chromosomes assayed. This is based on the results from four single-cell experiments of P0. Each point represents the coverage of an autosome. The error bars represent s.e.m.

project, haplotypes in the CEU population were obtained by studying the genotypes of family trios. About 80% of the heterozygous SNPs of the child can be unambiguously phased given that one parent is homozygous for the SNP. The remaining ~20% of heterozygous SNPs in the child are ambiguous and require statistical phasing because both parents are heterozygous. Comparison of DDP and HapMap data on unambiguous SNPs provides an estimate of the accuracy of DDP. The concordance rate between the two data sets was 99.8%. The small number of inconsistencies arose from either error in DDP genotyping or error in genotyping in HapMap data (**Fig. 3a**). When considering ambiguous SNPs alone, the incongruence rate between the two data sets was 5.7%. The majority of these inconsistencies (96.0%) came from incorrect statistical phasing in the HapMap project, as we could confirm the phases of these ambiguous SNPs in the child from the experimentally determined phases of the two parents (**Fig. 3**). These data agree with previous evaluations of the accuracies of statistical phasing in CEU trios[15,28] and highlight the utility of direct experimental phasing even when family data are available.

Whole-genome haplotyping of a European individual

Having validated the DDP approach on well-characterized HapMap samples, we applied it to determine the haplotypes of an individual, labeled 'P0', whose genome has been sequenced[6] and clinically annotated[29]. As only a few cells are required for DDP, we collected a blood sample from a finger prick. Whereas some of the early microfluidic devices used for experiments with the

Figure 3 Comparison of statistically determined phases with experimentally determined phases. (**a**) Comparison of experimentally determined phases of ~160,000 heterozygous SNPs of GM12878 (child of the trio) and those determined by phase III of the HapMap project. Unambiguous SNPs refer to those that are homozygous for at least one parent and are deterministically phased using family data in HapMap. This comparison shows the accuracy of DDP. Ambiguous SNPs refer to those that are heterozygous for all members of the trio and statistical phasing is used in HapMap. This comparison provides an evaluation of statistical phasing. (**b**) Comparison of experimentally determined phases of P0 and those determined by PHASE. Seventy-six regions on the autosomal chromosomes were randomly selected and statistically phased three times. Each region carried 100 heterozygous SNPs and spanned an average of ~2 Mb. Switch error rate was calculated as the proportion of heterozygous SNPs with different phases relative to the SNP immediately upstream. Single-site error rate was calculated as the proportion of heterozygous SNPs with incorrect phase. A SNP was considered correctly phased if it had the dominant phase. For each region, the average values from the three runs were reported. Presented here are the average switch error and single-site error per region. The deterministic phases measured by DDP are taken as the ground truth.



family trio contained defects leading to the failure to retrieve products from some chambers, refinement in device fabrication yielded fully functional devices and thus improved the number of SNPs phased per single-cell experiment for P0. The average number of pairs of autosomal chromosomes separated per single cell of P0 was 17.5. We obtained ~96.1% coverage of the ~1.2 million SNPs present on the HumanOmni1S array using four single cells (**Fig. 2** and **Supplementary Data Set 4**). An additional ~861,000 SNPs were phased using materials from three single cells and the HumanOmni1-Quad array (**Supplementary Data Set 5**). For homologous chromosomes that were separated in all four single-cell replicates (that is, four biological replicates of each homologous copy), up to 99.2% of all SNPs assayed on a chromosome were phased (**Fig. 2**).

We noticed that the SNPs that were not phased tended to cluster together and closer inspection revealed that they were usually located in regions with higher GC content (**Supplementary Fig. 2**). Stronger molecular associations between DNA strands at regions with higher GC content might have led to more difficult amplification, and such phenomena associated with phi29 have been previously reported[30].

Phasing of SNPs was also achieved by direct sequencing. We lightly sequenced amplified material from three single copies of P0's chromosome 6, at an average read depth of 3.5× to 7.7× per copy (**Supplementary Table 1**). About 46,000 heterozygous SNPs on chromosome 6 determined by previous genome sequencing were phased, including several of the medically relevant rare variants that were identified in the clinical annotation of the genome[29]. For alleles called by threefold or greater coverage, the concordance rate of phasing by sequencing

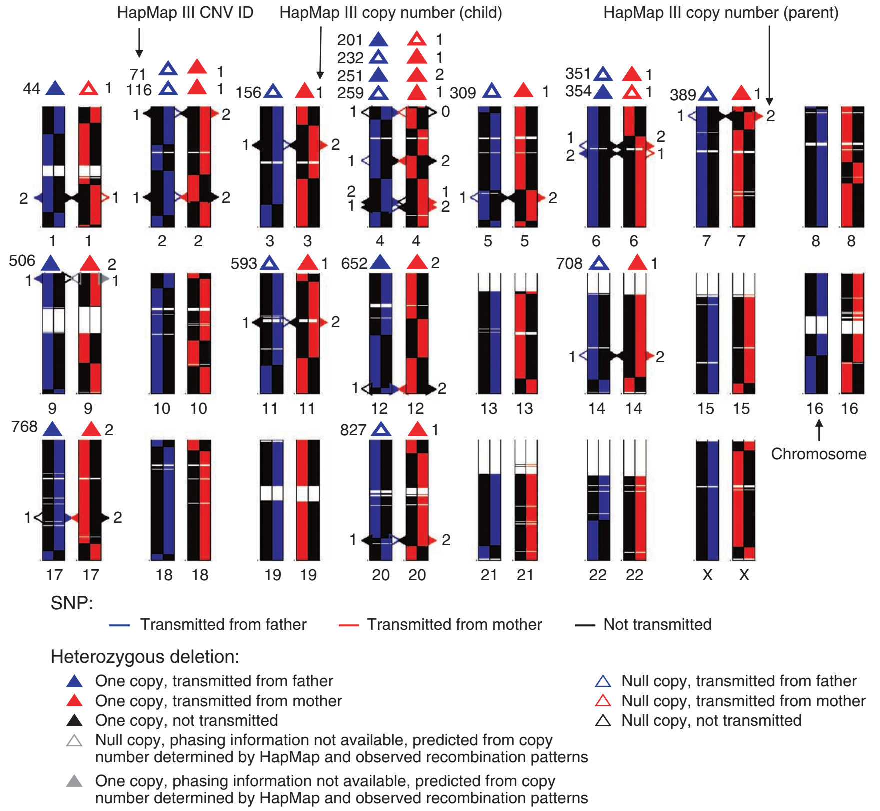
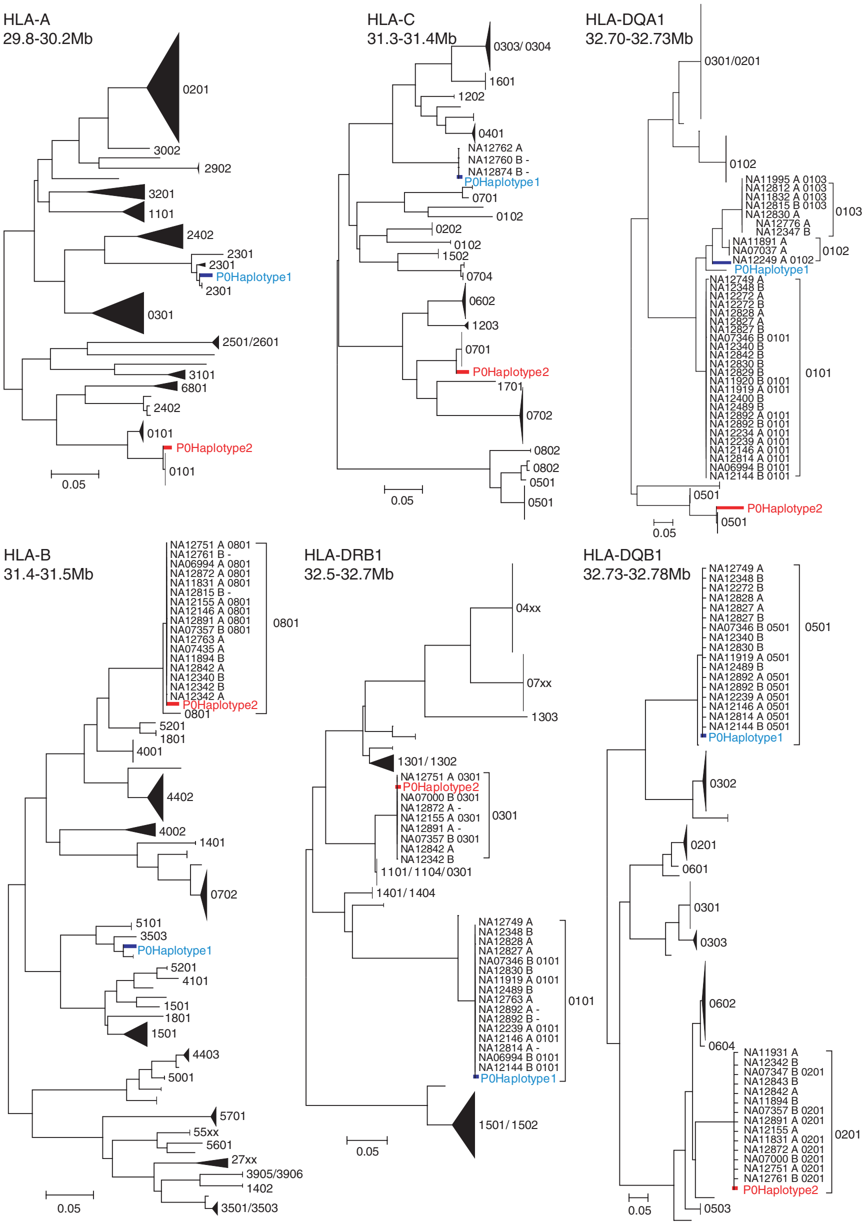

HapMap III CNV ID
HapMap III copy number (child)
HapMap III copy number (parent)
Chromosome
SNP:
Transmitted from father
Transmitted from mother
Not transmitted
Heterozygous deletion:
One copy, transmitted from father
One copy, transmitted from mother
One copy, not transmitted
Null copy, phasing information not available, predicted from copy number determined by HapMap and observed recombination patterns
One copy, phasing information not available, predicted from copy number determined by HapMap and observed recombination patterns
Null copy, transmitted from father
Null copy, transmitted from mother
Null copy, not transmitted

Figure 4 Direct observation of recombination events and deterministic phasing of heterozygous deletions in the family trio. Each allele with DDP data available for the child and the parent is represented by a colored line (blue, alleles transmitted to the child from the father; red, alleles transmitted to the child from the mother; black, untransmitted alleles). Centromeres and regions of heterochromatin are not assayed by genotyping arrays and are thus in white. Heterozygous deletions in the parents are represented as triangles along each homologous chromosome. A solid triangle represents one copy and a hollow triangle represents a null copy. The phases of deletions are determined for each parent independently. The triangles are color coded according to the state of transmittance as determined by the location of the deletion relative to spots of recombination. The phases of the deletions in the child are determined independent of the parents and are shown on top of the parental chromosomes. The integers on the left are the IDs of each region given by HapMap phase III. The numbers on the right are the copy number of a region in the child as determined by HapMap. Chromosomes are plotted with the same length.



		HLA-A	HLA-B	HLA-C	HLA-DRB1	HLA-DQA1	HLA-DQB1
HLA alleles by direct typing:	Genomic DNA	0101/2301	0801/5107	0701/1402	0301/0101	–	0201/0501
HLA haplotypes by phylogenetic analysis:	Haplotype 1	2301	?	?	0101	0101	0501
	Haplotype 2	0101	0801	0701	0301	0501	0201
HLA haplotypes by combining results from phylogenetic analysis and direct typing:	Haplotype 1	2301	5107	1402	0101	0101	0501
	Haplotype 2	0101	0801	0701	0301	0501	0201

Figure 5 HLA haplotypes of P0 determined using DDP. At each of the six classical HLA loci, the experimentally phased SNP haplotypes of P0 and 176 phased SNP haplotypes of CEU trios available from HapMap phase III were placed on a neighbor-joining tree. The two haplotypes of P0 are labeled in red and blue. For haplotypes in the CEU panel with HLA typing data, the four-digit HLA allele is presented next to the sample label. Most of each tree is compressed. Each compressed subtree is labeled with the HLA allele associated with members inside the subtree, if HLA allele information is available. The allelic identities of HLA-B and HLA-C on haplotype 1 were not determined with DDP because CEU individuals with similar SNP haplotypes as P0's SNP haplotypes did not have HLA typing data at these loci but could be inferred from the results of direct HLA typing of genomic DNA (first row of table). HLA-DQA1 was not directly typed.

and phasing by genotyping arrays was 99.8% (**Supplementary Fig. 3a**). This indicates that allele calling with haploid materials can be achieved accurately with relatively low coverage, an advantage over conventional genotyping by sequencing, which requires much higher fold coverage to guarantee accuracy of heterozygous SNPs. The amplification of minute amount of materials using the polymerase phi29 has been known to cause amplification bias and formation of nonspecific products that would undermine sequencing performance. Our group previously demonstrated improved performance of whole-genome amplification of single bacteria by reducing amplification volumes by ~1,000-fold using microfluidic devices similar to the one in this study[31,32]. The present sequencing experiments show that nonspecific products constituted a very small amount of the amplified materials and provide a characterization of the amplification bias for human chromosome–sized, single-molecule templates (**Supplementary Table 1** and **Supplementary Fig. 3b–d**). The coverage across the chromosome was nonuniform, yet distribution of reads over most of the chromosome in all sequenced copies was within two orders of magnitude (**Supplementary Fig. 3c**).

Comparison of experimental and statistical phasing

Statistical inference has been commonly used to estimate haplotypes in unrelated individuals, yet the lack of true haplotypes means that few studies have been conducted to evaluate the accuracy of these computational approaches. We used the statistical inference software PHASE[33–35] to infer haplotypes for P0 using CEU haplotypes, determined by family trios in the HapMap Project, as the background and compared the inferred haplotypes to the P0 haplotypes determined by DDP. Evaluation of a total of 76 ~2-Mb regions, each defined by 100 heterozygous SNPs, revealed an average of 6.3 block switches per region and an average block size of ~260 kb. An average of 30.2% of heterozygous SNPs were incorrectly phased using the statistical method (**Fig. 3b** and **Supplementary Fig. 4**). These results agree with two previous studies that compared statistical haplotype inference with real phases obtained from somatic cell hybrids and complete hydatidform moles[36,37], and illustrate the importance of direct experimental phasing especially when family data are not available.

Direct measurement of recombination events within a family trio

The availability of parental haplotypes allowed us to directly measure the products of recombination events that led to an individual's unique genome, which could previously only be inferred using three-generation families[38] or two-generation families with large sibships[39]. We aligned each homologous chromosome of the child to the pair of chromosomes of the parent from whom the chromosome was inherited. **Figure 4** illustrates the crossover events resulting from the paternal and maternal meioses. We detected 26 and 38 events in the male meiosis and female meiosis, respectively, with a median resolution of ~43 to 44 kb (**Supplementary Table 2**). The number of detected recombination events matched those in previous reports and supports the notion that the number of recombination events in females is generally higher than that in males[38,40]. At least 60% of these regions had recombination rates above the median sex average according to the deCODE genetic map[39]. In addition to the switchover of large blocks of homologous chromosomes as a result of recombination, we observed switchover at single sites, constituting ~0.4% of the total number of SNPs in each parent-child comparison; these are presumably products of gene conversion or cell culture–induced mutations, as well as DDP error.

Phasing of heterozygous deletions

Although copy number variants (CNVs) can be phased using statistical methods similar to those used to phase SNPs[41–43], direct experimental phasing of structural variation such as copy number polymorphisms has largely been unexplored. We experimentally phased the heterozygous deletions accessible with genotyping arrays, as determined by the HapMap Project, of the three individuals in the family trio. We phased 12 and 6 heterozygous deletions present within the family trio using genotyping array data (**Supplementary Table 3a**) and real-time PCR (**Supplementary Table 3b**), respectively. All of the phased heterozygous deletions within the trio agreed with the inheritance pattern (**Fig. 4**). We also phased all eight heterozygous deletions that had been detected by genome sequencing of P0 (ref. 6) using data from genotyping arrays and real-time PCR. Results from all platforms among all single-cell replicates were consistent (**Supplementary Table 4**). Phasing of other types of CNVs with the current approach of genotyping array and PCR is challenging, but we envision that deep sequencing of amplified materials would eventually allow each chromosome to be assembled and thus enable phasing of all CNVs.

Direct determination of the HLA haplotypes of an individual

An important application of DDP is the determination of the HLA haplotypes within an individual. The HLA loci are highly polymorphic and are distributed over ~4 Mb on chromosome 6. The ability to haplotype the HLA genes within the region is clinically important because this region is associated with autoimmune and infectious diseases[44], and the compatibility of HLA haplotypes between donor and recipient can influence the outcomes of transplantation[8]. Yet molecular techniques to measure HLA haplotypes in individuals are still limited[45].

To determine the HLA haplotypes, we first had to determine the HLA allele at each locus, which is usually achieved by direct sequencing. Here, we sought a simpler approach to determine the allele at each HLA locus by taking advantage of the experimentally determined SNP haplotypes of P0. Briefly, we used phylogenetic analyses to compare the SNP haplotypes of P0 within each HLA gene to those of CEU individuals whose HLA genes were typed previously (**Fig. 5**). The combination of the alleles at each HLA locus determined by phylogenetic analyses agreed with direct HLA typing of genomic DNA. Combining the results from all loci yielded the two HLA haplotypes of P0. One of the HLA haplotypes is the 8.1 ancestral haplotype, which is one of the most frequently observed haplotypes in Caucasians[46] and is associated with elevated risks of immunopathological diseases[47].

DISCUSSION

The DDP approach is scalable. Multiple cells can be processed simultaneously by modifying device design. Currently, the most labor-intensive procedure is the manual identification of metaphase cells. We anticipate that automation of this with a relatively simple engineering solution, such as the combination of computer vision and fluorescent labeling of mitotic cells, will dramatically increase throughput. The majority of the cost of the project went to the genotyping arrays, and as sequencing costs continue to drop, it may become more cost effective to sequence rather than to genotype.

DDP requires the presence of metaphase chromosomes because during metaphase chromosomes are most condensed and can be physically separated. DDP therefore requires sources of cells that can undergo mitosis, such as blood samples and cell lines. Yet DDP requires as little as a single cell, and thus may also have important applications in single-cell genomics, in fields such as preimplantation genetic diagnosis, prenatal diagnosis, aging, and cancer diagnosis and research.

To our knowledge, the work described here represents the first demonstration of a molecular-based, whole-genome haplotyping technique amenable for personal genomics. Whereas the bulk of the experiments described here focus on direct deterministic phasing of ~1 million variants accessible by genotyping arrays, DDP can be used to phase all variants in the genome. DDP of tagSNPs present on the genotyping arrays inherently provides phasing information for common variants that are in strong linkage disequilibrium with the tagSNPs. In addition, we showed that amplified materials from separated chromosome homologs could be directly sequenced, yielding phasing information for variants, including the rare and private ones, which are absent on standard genotyping arrays. Combining DDP SNP analysis with shotgun genome sequencing could allow the determination of the complete personal haplotype of an individual, even in the absence of family information.

METHODS

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

Accession code. Short-read sequence data have been deposited at the NCBI Sequence Read Archive (SRA) under accession no. SRA026722.

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

ACKNOWLEDGMENTS

We thank J. Melin and the Stanford Microfluidic Foundry for fabrication of microfluidic devices. We thank N. Neff and G. Mantalas for performing sequencing experiments. We thank D. Pushkarev for providing data and primers from the genome sequencing project of P0. We thank M. Anderson and D. Tyan at Stanford Histocompatibility Laboratory for providing HLA typing results. We thank Y. Marcy, P. Blainey, J. Jiang and A. Wu for helpful discussions. The project was supported by the US National Institutes of Health (NIH) Pioneer Award and an NIH U54 award. H.C.F. was supported by a scholarship from the Siebel Foundation. J.W. was supported by a scholarship from the China Scholarship Council.

AUTHOR CONTRIBUTIONS

H.C.F. and S.R.Q. conceived the experiments. H.C.F. designed the microfluidic device. A.P. developed protocols for device fabrication. H.C.F. and J.W. performed the experiments. H.C.F., J.W. and S.R.Q. analyzed the data and wrote 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. Wheeler, D.A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
2. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
3. Ahn, S.M. *et al.* The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.* **19**, 1622–1629 (2009).
4. Kim, J.I. *et al.* A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**, 1011–1015 (2009).
5. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
6. Pushkarev, D., Neff, N.F. & Quake, S.R. Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.* **27**, 847–850 (2009).
7. Schuster, S.C. *et al.* Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**, 943–947 (2010).
8. Petersdorf, E.W., Malkki, M., Gooley, T.A., Martin, P.J. & Guo, Z. MHC haplotype matching for unrelated hematopoietic cell transplantation. *PLoS Med.* **4**, e8 (2007).
9. de Bakker, P.I. *et al.* A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.* **38**, 1166–1172 (2006).
10. Stewart, C.A. *et al.* Complete MHC haplotype sequencing for common disease gene mapping. *Genome Res.* **14**, 1176–1187 (2004).
11. Groenendijk, M., Cantor, R.M., de Bruin, T.W. & Dallinga-Thie, G.M. The apoAI-CIII-AIV gene cluster. *Atherosclerosis* **157**, 1–11 (2001).
12. Nagel, R.L. *et al.* The Senegal DNA haplotype is associated with the amelioration of anemia in African-American sickle cell anemia patients. *Blood* **77**, 1371–1375 (1991).
13. Sun, T. *et al.* Haplotypes in matrix metalloproteinase gene cluster on chromosome 11q22 contribute to the risk of lung cancer development and progression. *Clin. Cancer Res.* **12**, 7009–7017 (2006).
14. Drysdale, C.M. *et al.* Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proc. Natl. Acad. Sci. USA* **97**, 10483–10488 (2000).
15. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
16. Frazer, K.A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
17. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
18. Zhang, K. *et al.* Long-range polony haplotyping of individual human chromosome molecules. *Nat. Genet.* **38**, 382–387 (2006).
19. Mitra, R.D. *et al.* Digital genotyping and haplotyping with polymerase colonies. *Proc. Natl. Acad. Sci. USA* **100**, 5926–5931 (2003).
20. Ding, C. & Cantor, C.R. Direct molecular haplotyping of long-range genomic DNA with M1-PCR. *Proc. Natl. Acad. Sci. USA* **100**, 7449–7453 (2003).
21. Michalatos-Beloin, S., Tishkoff, S.A., Bentley, K.L., Kidd, K.K. & Ruano, G. Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. *Nucleic Acids Res.* **24**, 4841–4843 (1996).
22. Ruano, G., Kidd, K.K. & Stephens, J.C. Haplotype of multiple polymorphisms resolved by enzymatic amplification of single DNA molecules. *Proc. Natl. Acad. Sci. USA* **87**, 6296–6300 (1990).
23. Woolley, A.T., Guillemette, C., Li Cheung, C., Housman, D.E. & Lieber, C.M. Direct haplotyping of kilobase-size DNA using carbon nanotube probes. *Nat. Biotechnol.* **18**, 760–763 (2000).
24. Burgtorf, C. *et al.* Clone-based systematic haplotyping (CSH): a procedure for physical haplotyping of whole genomes. *Genome Res.* **13**, 2717–2724 (2003).
25. Xiao, M. *et al.* Direct determination of haplotypes from single DNA molecules. *Nat. Methods* **6**, 199–201 (2009).
26. Ma, L. *et al.* Direct determination of molecular haplotypes by chromosome microdissection. *Nat. Methods* **7**, 299–301 (2010).
27. Douglas, J.A., Boehnke, M., Gillanders, E., Trent, J.M. & Gruber, S.B. Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. *Nat. Genet.* **28**, 361–364 (2001).
28. Marchini, J. *et al.* A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.* **78**, 437–450 (2006).
29. Ashley, E.A. *et al.* Clinical assessment incorporating a personal genome. *Lancet* **375**, 1525–1535 (2010).
30. Bredel, M. *et al.* Amplification of whole tumor genomes and gene-by-gene mapping of genomic aberrations from limited sources of fresh-frozen and paraffin-embedded DNA. *J. Mol. Diagn.* **7**, 171–182 (2005).
31. Marcy, Y. *et al.* Nanoliter reactors improve multiple displacement amplification of genomes from single cells. *PLoS Genet.* **3**, e155 (2007).
32. Marcy, Y. *et al.* Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc. Natl. Acad. Sci. USA* **104**, 11889–11894 (2007).
33. Stephens, M., Smith, N.J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**, 978–989 (2001).
34. Stephens, M. & Donnelly, P. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* **73**, 1162–1169 (2003).
35. Stephens, M. & Scheet, P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.* **76**, 449–462 (2005).
36. Kukita, Y. *et al.* Genome-wide definitive haplotypes determined using a collection of complete hydatidiform moles. *Genome Res.* **15**, 1511–1518 (2005).
37. Andres, A.M. *et al.* Understanding the accuracy of statistical haplotype inference with sequence data of known phase. *Genet. Epidemiol.* **31**, 659–671 (2007).
38. Broman, K.W., Murray, J.C., Sheffield, V.C., White, R.L. & Weber, J.L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63**, 861–869 (1998).
39. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nat. Genet.* **31**, 241–247 (2002).
40. Frazer, K.A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
41. Conrad, D.F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
42. Su, S.Y. *et al.* Inferring combined CNV/SNP haplotypes from genotype data. *Bioinformatics* **26**, 1437–1445 (2010).
43. McCarroll, S.A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40**, 1166–1174 (2008).
44. Shiina, T., Hosomichi, K., Inoko, H. & Kulski, J.K. The HLA genomic loci map: expression, interaction, diversity and disease. *J. Hum. Genet.* **54**, 15–39 (2009).
45. Guo, Z., Hood, L., Malkki, M. & Petersdorf, E.W. Long-range multilocus haplotype phasing of the MHC. *Proc. Natl. Acad. Sci. USA* **103**, 6964–6969 (2006).
46. Maiers, M., Gragert, L. & Klitz, W. High-resolution HLA alleles and haplotypes in the United States population. *Hum. Immunol.* **68**, 779–788 (2007).
47. Price, P. *et al.* The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunopathological diseases. *Immunol. Rev.* **167**, 257–274 (1999).
48. White, R.A. III, Blainey, P.C., Fan, H.C. & Quake, S.R. Digital PCR provides sensitive and absolute calibration for high throughput sequencing. *BMC Genomics* **10**, 116 (2009).
49. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599 (2007).

ONLINE METHODS

Microfluidic device design, fabrication and operation. The microfluidic device was made of polydimethylsiloxane (PDMS) and was fabricated using soft lithography by the Stanford Microfluidic Foundry. The two-layered device had rectangular 25-µm tall control channels at the bottom and rounded flow channels at the top. The device was bonded to a glass slide coated with a thin layer of PDMS. In the cell-sorting region of the device, flow channels were 40 µm high and 200 µm wide. In the amplification region of the device, flow channels were 5 µm and 100 µm wide and reaction chambers were 40 µm tall. A membrane valve was formed when a control channel crossed over with a flow channel and was actuated when the control channel was pressurized at 20–25 p.s.i. The area of each valve was 200 µm × 200 µm for the 40-µm flow channels, and 100 µm × 100 µm for the 5-µm flow channels. Membrane valves were controlled by external pneumatic solenoid valves that were driven by custom electronics connected to the USB port of a computer. A Matlab program was written to interface with the valves.

Cell culture. The Epstein-Barr virus–transformed lymphoblastoid cell lines GM12891, GM12892 and GM12878, belonging to the pedigree CEU 1463 (Coriell Cell Repositories), were cultured in RPMI 1640, supplemented with 15% FBS. Each culture was treated with 2 mM thymidine (Sigma) for 24 h at 37 °C to enrich the population of mitotic cells. Followed by multiple washings, cells were cultured in normal medium for 3 h and treated with 200 ng/ml nocodazole (Sigma) for 2 h at 37 °C to arrest cells at metaphase.

Whole blood (~250 µl) obtained from a finger-prick of Patient Zero ('P0') was treated with sodium heparin and cultured in PB-Max medium (Invitrogen) for 4 days. The culture was treated with 50 ng/ml colcemid (Invitrogen) for 6 h. The culture was layered on top of Accuspin System-Histopaque-1077 (Sigma) and centrifuged for 8 min at 590*g*. Nucleated cells at the interface was removed and washed once with HBSS.

Metaphase arrested cells incubated with 75 mM KCl at 25 °C for 10 to 15 min. Acetic acid was added to the cell suspension at a final concentration of 2% to fix the cells. After fixation on ice for 30 min, cells were washed multiple times and finally suspended in 75 mM KCl-1mM EDTA-1% Triton X-100. Cells were treated with 0.2 mg/ml RNaseA (Qiagen) before loading onto the microfluidic device.

Cell sorting, chromosome release and multiple strand displacement amplification. Before the loading of the cell suspension, the cell-sorting channel of the device was treated with Pluronic F127 (0.2% in PBS). Cell suspension was introduced into the device using an on-chip peristaltic pump and an off-chip pressure source. Metaphase cells could be distinguished from interphase cells microscopically by morphological differences. Once a single metaphase cell was recognized at the capture chamber, surrounding valves were actuated to isolate it from the remaining cell suspension. Pepsin solution (0.01% in 75 mM KCl, 1% Triton X-100, 2% acetic acid) was introduced to digest the cytoplasm and release the chromosomes. The chromosome suspension was pushed into a long narrow channel and partitioned into 48 180-pl compartments by actuating a series of valves along the channel. Trypsin (0.25%) in 150 mM Tris-HCl (pH 8.0) (1.2 nl) was introduced to neutralize the solution and to digest chromosomal proteins. Ten minutes later, denaturation buffer (Qiagen's Repli-G Midi kit's buffer DLB supplemented with 0.8% Tween-20; 1.4 nl) was introduced. The device was placed on a flat-topped thermal cycler set at 40 °C for 10 min. This was followed by the introduction of neutralization solution (Repli-G kit's stop solution; 1.4 nl) and incubation at 25 °C for 10 min. A mixture of reaction buffer (Qiagen's Repli-G Midi Kit), phi29 polymerase (Qiagen's Repli-G Midi Kit), 1× protease inhibitor cocktail (Roche) and 0.5% Tween-20 (16 nl) was fed in. The total volume per reaction was 20 nl and the device was placed on the flat-topped thermal cycler set at 32 °C for about 16 h. Amplification products from each chamber was retrieved from its corresponding outlet by flushing the chamber with TE buffer (pH 8.0) supplemented with 0.2% Tween-20. About 3–5 µl of products were collected from each chamber. Products were incubated at 65 °C for 3 min to inactivate the phi29 enzyme.

Initial genotyping with 46 loci. To determine the identity of chromosomes in each chamber, we performed Taqman PCR using a set of 46 genotyping assays (two assays per autosome and one assay per sex chromosome) on the products of each chamber on the 48.48 Dynamic Array (Fluidigm). The assays used are listed in **Supplementary Table 5**.

Whole-genome phasing using genotyping arrays. To generate sufficient materials for genotyping array experiments, we amplified DNA products from the microfluidic device a second time in 10 µl volume using the Repli-G Midi Kit's protocol for amplifying purified genomic DNA. Products from multiple chambers were pooled together into two mixtures such that each mixture contained one of the homologous copies of each chromosome. Each mixture, containing roughly one haploid genome of a cell, was genotyped on the HumanOmni1-Quad or HumanOmni1S BeadChips (Illumina). For GM12891, GM12878 and P0, four single cells were haplotyped. For GM12892, three single cells were haplotyped. Haplotyping data for cell lines were obtained from HumanOmni1-Quad array. Haplotyping data for P0 were obtained from both HumanOmni1-Quad and HumanOmni1S arrays. Genomic DNA extracted from each cell line was also genotyped on the HumanOmni1-Quad array. Genomic DNA of P0 was genotyped on the HumanOmni1S array.

For each chromosome homolog, the allelic identity of a SNP was determined from the consensus among the biological replicates. If equal numbers of both alleles were observed at the site, no consensus was drawn. We estimated the error of a single genotyping measurement by counting the number of inconsistent allele calls at sites typed more than once. For SNPs of which only one of the alleles was observed, the identity of the other allele was determined using the genotypes of genomic DNA. For the trio, the genotypes of genomic DNA were measured on the HumanOmni1-Quad BeadChip. The concordance rate of these genotypes with HapMap data was ~99.1% for each cell line. For P0, genotypes of genomic DNA were measured on the HumanOmni1S BeadChip. The combination of the consensus alleles from the two homologs at each SNP site should, in principle, agree with the genotype call of the genomic DNA control. SNPs that did not follow this rule (~0.3% for cell lines and ~0.4% for P0) were eliminated from downstream analyses.

Data files containing the phased haplotypes of the members of the trio and P0 are available as **Supplementary Data Set 1** (GM12891), **Supplementary Data Set 2** (GM12892), **Supplementary Data Set 3** (GM12878), **Supplementary Data Set 4** (P0 Omni1S) and **Supplementary Data Set 5** (P0 Omni1Quad). Each file contains whole-genome haplotypes of each individual of the CEU trio (GM12891, GM12892, GM12878) and P0. For the trio, refSNPs present on the Omni1-Quad array (Illumina) that were phased by DDP are included in these files. For P0, refSNPs present on the Omni1-Quad and SNPs present on the Omni1S arrays phased by DDP are included. For P0's data on Omni1-Quad arrays, only SNPs with both alleles directly observed were included. Alleles are presented relative to the forward strand, and SNPs with A/T and G/C alleles are not included. Column 1: SNP name; column 2: chromosome (chromosome X designated as '23'); column 3: position on chromosome (hg18); column 4: allele on homologous copy 1; column 5: allele on homologous copy 2.

Phasing of chromosome 6 using high-throughput sequencing. Three chambers containing amplified materials from a single copy of chromosome 6 were selected from the four single-cell experiments of P0 for paired-end sequencing on Illumina's Genome Analyzer II. Two chambers contained materials from chromosome 6 only, whereas the third chamber contained materials from a homolog of chromosomes 6, 16 and 18. Second-round amplified materials from these chambers were fragmented through a 30-min 37 °C incubation with 4 µl dsDNA Fragmentase (New England Biolabs) in a 20-µl reaction. Fragmented DNA was end-repaired, tailed with a single 'A' base, and ligated with adaptors. A 12-cycle PCR was carried out and PCR products with sizes of 300–500 bp were selected using gel extraction. Sequencing libraries were quantified with digital PCR[48]. Thirty-six base pairs were sequenced on each end.

Image analysis, base calling and alignment were performed using Illumina's GA Pipeline version 1.5.1. The first 32 bases on each read were aligned to the human genome (Build 36.1). SNP calling was carried out using Illumina's CASAVA version 1.6.0. Positions covered at least three times according to the 'sort.count' intermediate files were used in downstream analyses. A list of heterozygous SNPs was obtained from the sequenced genome of P0, using quality score >2.8 and heterozygous score of 20 (ref. 6). The phases of heterozygous SNPs were determined either from the

doi:10.1038/nbt.1739

direct observation of both alleles in the different homologs, or by inferring the identity of the unobserved allele if only one allele was detected.

Data sources. Genotypes, CNVs and phasing data of the three lymphoblastoid cell lines were downloaded from the website of the International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/). Genotypes of the merged phase I+II and III data were used. Phasing information from phase III was used. CNV data from phase III was used.

Comparison between experimentally determined phases of the family trio with HapMap data. We compared the experimentally determined phases of the heterozygous SNPs of the child (GM12878) to those determined by phase III of the HapMap Project. SNPs with A/T and G/C alleles were excluded from comparison. To determine the accuracy of experimental phasing and to locate spots of crossovers, the phases of heterozygous SNPs of each parent (GM12891, GM12892) were compared to those in the child (GM12878) inherited from that parent.

Phasing of heterozygous deletions. For the trio, a list of heterozygous deletions was obtained from phase III of the HapMap project. For P0, heterozygous deletions that were detected by previous genome sequencing and subsequently verified by digital PCR were studied[6]. For P0, the assays were the same as those used in a previous study[6]. For the trio, the sequences of primers and probes are listed in **Supplementary Table 6**. The assumption was that one of the chromosome homologs should give no calls for SNP markers or no PCR amplification within a region of heterozygous deletions. Digital PCR (Fluidigm's 48.770 digital array) was also performed using genomic DNA of each member of the trio to verify copy numbers.

Statistical phasing of P0 using PHASE. We evaluated the accuracy of statistical phasing by comparing statistically phased haplotypes and experimentally determined haplotypes of P0. We inferred the haplotypes of P0 using PHASE 2.1 (a Bayesian method–based program for haplotype reconstruction)[33–35]. Due to computational capacity, we randomly chose four regions on each autosomal chromosome (except chromosomes 4, 20, 21), each having 100 bi-allelic SNPs that were heterozygous in P0. We only selected SNPs with both alleles directly haplotyped and with perfect concordance with genotype determined by whole-genome sequencing. Each region covered a range of ~0.7 to ~3.3 Mb, with an average SNP to SNP distance of ~20 kb. We used the 176 phased CEU haplotypes in phase III of the HapMap project as known haplotypes for the inference. For each region, we ran the reconstruction three times with the same default settings but different random seeds and compared the results with the experimentally determined haplotypes. Switch error rate was calculated as the proportion of heterozygous SNPs with different phases relative to the SNP immediately upstream. Single-site error rate was calculated as the proportion of heterozygous SNPs with incorrect phase. A SNP was considered correctly phased if it had the dominant phase. For each region, the average values from the three runs were reported.

Determination of HLA haplotypes of P0. A total of 176 phased CEU haplotypes obtained from phase III of the HapMap project, together with experimentally phased haplotypes of P0, were used to construct neighbor-joining trees at the six classical HLA loci on chromosome 6. The coordinate boundaries of which haplotyped SNPs were used for each locus are presented in **Figure 5**. Only SNPs with both alleles directly observed were used. The number of SNPs used for HLA-A, HLA-B, HLA-C, HLA-DRB, HLA-DQA, and HLA-DQB were 420, 139, 89, 59, 14 and 34, respectively. Allele sharing distances were computed for each pair of haplotypes as $\frac{1}{n}\sum_{i=1}^{n} d_i$, where n is the number of loci and d_i equals 0 for matched alleles and 1 for unmatched alleles at the i^{th} SNP locus. Trees were constructed using MEGA 4.1 (ref. 49). A list of HLA alleles of individuals in the CEU panel typed in a previous study[9] was downloaded from http://www.inflammgen.org/. The allelic identity of each homologous chromosome of P0 at each HLA locus was determined by the allelic identities of its nearest neighbors in the tree.

doi:10.1038/nbt.1739

nature biotechnology

Haplotype-resolved genome sequencing of a Gujarati Indian individual

Jacob O Kitzman[1], Alexandra P MacKenzie[1], Andrew Adey[1], Joseph B Hiatt[1], Rupali P Patwardhan[1], Peter H Sudmant[1], Sarah B Ng[1], Can Alkan[1,2], Ruolan Qiu[1], Evan E Eichler[1,2] & Jay Shendure[1]

Haplotype information is essential to the complete description and interpretation of genomes[1], genetic diversity[2] and genetic ancestry[3]. Although individual human genome sequencing is increasingly routine[4], nearly all such genomes are unresolved with respect to haplotype. Here we combine the throughput of massively parallel sequencing[5] with the contiguity information provided by large-insert cloning[6] to experimentally determine the haplotype-resolved genome of a South Asian individual. A single fosmid library was split into a modest number of pools, each providing ~3% physical coverage of the diploid genome. Sequencing of each pool yielded reads overwhelmingly derived from only one homologous chromosome at any given location. These data were combined with whole-genome shotgun sequence to directly phase 94% of ascertained heterozygous single nucleotide polymorphisms (SNPs) into long haplotype blocks (N50 of 386 kilobases (kbp)). This method also facilitates the analysis of structural variation, for example, to anchor novel insertions[7,8] to specific locations and haplotypes.

The high quality of the human reference genome derives from the hierarchical sequencing of large-insert clones, such that the assembly corresponding to each clone represents a single haplotype[9]. One of the first 'personal genomes' exploited clone-based mate pairing and long, accurate Sanger reads to resolve variants into haplotype blocks (N50 of 350 kbp; that is, 50% of resolved sequence is within blocks of at least 350 kbp)[1]. Although new technologies[5] have subsequently enabled >1,000-fold reduction in genome sequencing costs, the short read-lengths and paucity of contiguity information are such that it remains challenging to determine haplotypes at a genome-wide scale. Genomic phase, the assignment of alleles to homologous chromosomes, was determined for SNPs using mate-paired reads on the SOLiD (sequencing by oligonucleotide ligation and detection) platform[10] for an individual genome, but only 43% of heterozygous variants were phased, and nearly all in blocks no greater than the insert size, that is, <3.5 kbp[10]. Experimental limitations on the size and complexity of mate-pair libraries based on *in vitro* circularization[11] make it difficult to improve upon this approach.

An alternative is to infer haplotypes from population-based linkage disequilibrium data or from pedigree analysis. For example, haplotypes were successfully inferred in the YH (YanHuang) genome for variants at which phased CHB/JPT HapMap data were available (CHB, Han Chinese from Beijing, China; JPT, Japanese from Tokyo, Japan)[12]. The genomes of a family of four have been sequenced and these relationships used to infer inheritance blocks[13]. Although they can be successful, inferential methods have limitations. Statistical phasing, whether based on genotyping[2] or sequencing[14], performs poorly when linkage disequilibrium is not high, and for rare variants. Phasing by pedigree analysis requires genome sequencing of many related individuals, increasing costs and limiting practical application.

We describe a cost-effective method for determining long-range haplotypes at a genome-wide scale by massively parallel sequencing of complex, haploid subsets of an individual genome (**Fig. 1**). We apply this method to the first reported whole-genome sequencing of a human of South Asian ancestry. The Indian subcontinent is home to myriad culturally and genetically diverse groups with distinct population histories[15]. We selected a female from the HapMap panel of 'Gujarati Indians in Houston' (GIH; NA20847) for sequencing. Notably, the imputation of genotypes for GIH was the least effective of all non-African populations in HapMap[2].

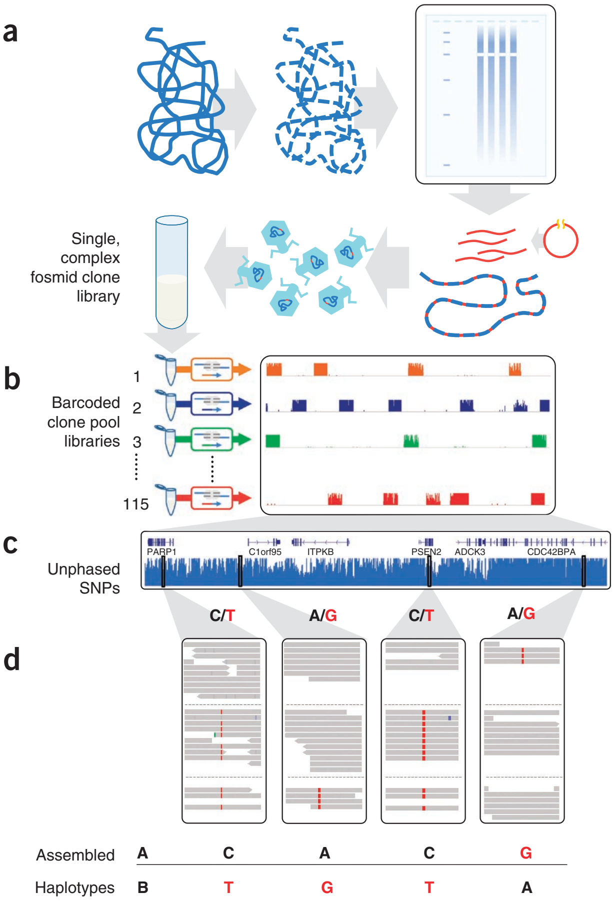
Genomic DNA from NA20847 was used to construct a single, complex fosmid library, containing clones packaged in phage for infecting *Escherichia coli* cells (>2 × 10^6 clones with ~37 kbp inserts) (**Fig. 1a** and **Supplementary Methods**). We then split a portion of this library to 115 pools, at a density such that each pool contained ~5,000 independent clones. Each pool was expanded by either scraping a single plate of infected cells and inoculating outgrowth culture, or by direct liquid outgrowth after infection. However, at no point does this method require the isolation of individual colonies. We next constructed 115 barcoded, shotgun sequencing libraries from fosmid DNA isolated from each of the 115 pools[16]. Libraries indexed with barcodes were combined and sequenced (Illumina GAIIx; PE76 or PE101 reads) to a mean 2.4× depth per haploid clone (**Fig. 1b**).

Because each pool captures an essentially random ~3% of the 6-gigabase (Gb) diploid genome (that is, ~5,000 fosmids × ~37 kbp inserts) sequence reads from each pool are overwhelmingly (99.1%) derived from only one homologous chromosome or the other at any single location. Upon mapping reads from each pool to the reference assembly, the approximate boundaries of 538,009 individual clones (37.2 ± 4.7 kbp) were identified by read depth (4,678 ± 1,229 clones per pool). Coverage was uniform across the genome (98.6% covered

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington, USA. [2]Howard Hughes Medical Institute, Seattle, Washington, USA. Correspondence should be addressed to J.S. (shendure@uw.edu) or J.O.K. (kitz@uw.edu).

Received 26 October; accepted 29 November; published online 19 December 2010; doi:10.1038/nbt.1740

Figure 1 Haplotype-resolved genome sequencing. (**a**,**b**) A single, highly complex fosmid library was constructed (**a**) and split into 115 pools (**b**), each representing ~3% physical coverage of the diploid human genome. Barcoded shotgun libraries from each pool were constructed, then combined and sequenced. As expected, reads from each library map to ~5,000 × ~37 kbp blocks, minimally redundant within each library. (**c**) Whole-genome shotgun sequencing of the same individual generated unphased variant calls. (**d**) Unphased variant calls were combined with haploid genotype calls to assemble haplotype blocks using a maximum parsimony approach[19] (reference allele in black, nonreference allele in red).



by one or more clones) and within each pool (82% of clones with mean read depth within a tenfold range) (**Supplementary Fig. 1**).

For unphased variation discovery, we performed conventional whole-genome resequencing to 15× depth (Illumina HiSeq; PE50) (**Supplementary Table 1** and **Supplementary Fig. 2**). After alignment to the reference, we called 3.3 × 10^6 SNPs and 3.4 × 10^5 short indels[17,18] (**Fig. 1c**). Nonreference sensitivity for SNPs was 91%, that is, HapMap variant genotypes at positions also called in our data, and genotype concordance to high-quality HapMap 3 genotypes[2] at called positions was 99.2% (n = 1,436,495). Other bulk statistics, including the heterozygous-to-homozygous call ratio, the fraction of called variants previously ascertained in the NCBI SNP database (dbSNP), the transition-to-transversion ratio, and the numbers and classes of coding variants, were consistent with expectations based on previously sequenced non-African genomes (**Supplementary Table 2**).

Several methods have been described for assembling haplotypes from sequence data[1,19–21]. We adopted a maximum parsimony approach[19] to combine the unphased variants from shotgun whole-genome sequencing with haploid genotype calls from sequencing of the 115 pools (**Fig. 1d**). The resulting assembly incorporated 94% of ascertained heterozygous SNPs into haplotype-resolved blocks, with an N90 of 89 kbp, an N50 of 386 kbp and an N10 of 1 megabase (Mbp) (**Fig. 2a**). Sixty-two percent of genes were fully encompassed by single blocks, and 73% were covered for over half their length.

To evaluate accuracy, we compared our haplotype assembly with HapMap phase predictions for NA20847 (**Fig. 2b**)[2]. For pairs of SNPs in exceptionally high-linkage disequilibrium ($D' > 0.90$ among GIH), we observed nearly perfect concordance (>99.7%). Because NA20847 was not part of a trio, HapMap predictions rely upon linkage disequilibrium between alleles to predict phase from genotypes. Correspondingly, concordance was reduced to ~71% when $D' < 0.10$, which is the case for most (66%) pairwise SNP combinations. Concordance is also reduced when one or both alleles in the pair is rare in GIH (**Fig. 2c**). Note that our haplotype assembly is experimental and specific to an individual, and therefore completely independent of population-based phenomena such as linkage disequilibrium and allele frequency. Consequently, these trends likely reflect errors in HapMap phasing[1].

South Asian history includes admixture between two ancestral groups, one genetically close to Europeans (ANI) and another more highly diverged from well-ascertained populations (ASI)[15]. Furthermore, principal components analysis revealed a distinct subgroup of Indian populations in general and GIH in particular, including NA20847, that may harbor substantial genetic ancestry from a third population distinct from ANI and ASI[15]. We compared haplotype blocks for this individual to HapMap allele frequencies in the GIH and CEPH European (CEU) populations to distinguish 'GIH-like' from 'CEU-like' haplotypes. Notably, novel SNPs are markedly enriched on the most GIH-like haplotypes (**Fig. 3**). We also scored haplotype blocks against allele frequencies from the 1000 Genomes Project[14] (**Supplementary Fig. 3**). Haplotypes that least resembled all three populations in that study (CEU, CHB/JPT and Yoruba) were also markedly enriched for novel SNPs. We propose that GIH-like blocks and other well-differentiated haplotypes may be derived from more poorly ascertained ancestral

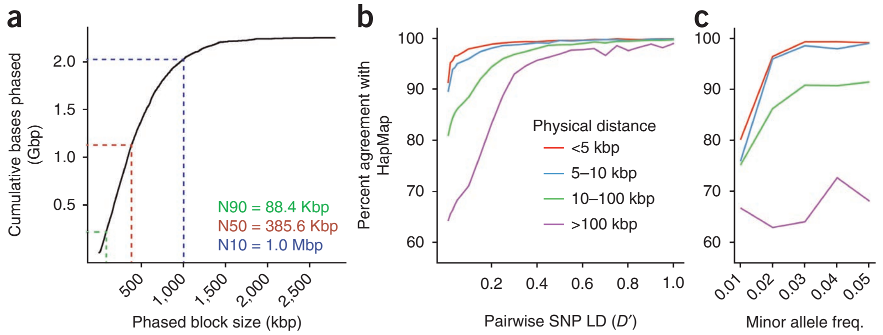


Figure 2 Haplotype assembly results. (**a**) Size distribution of blocks within the haplotype assembly up to a maximum block size of 2.79 Mbp. Half of the assembly comprised blocks longer than 386 kbp (N50). (**b**) Comparison of experimental phasing with HapMap population-based inference[2] for NA20847, with agreement of pairwise haplotype predictions as a function of physical distance and linkage disequilibrium. (**c**) Agreement of pairwise haplotype predictions as a function of physical distance and minor allele frequency (defined as the lower allele frequency of the pair in GIH). Key is the same as for **b**.

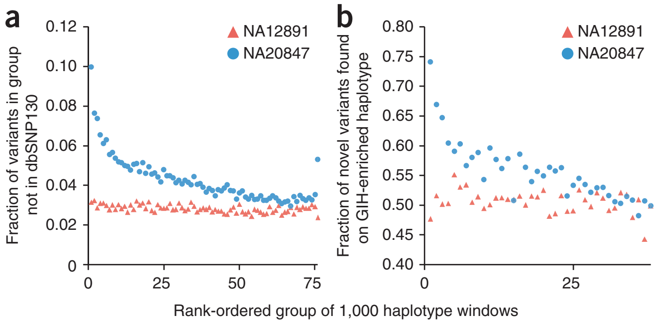


Figure 3 Enrichment of novel variants on ‘GIH-like’ haplotypes. (**a**) Haplotypes were scored and rank ordered within sliding windows of 20 HapMap variants[2] for greater similarity to GIH or CEU on the basis of population allele frequencies (left on *x* axis: more similar to GIH). Plotted is the fraction of novel variants (not in dbSNP v130) in rank-ordered groups of haplotype windows, demonstrating that the most ‘GIH-like’ haplotype windows are enriched for novel variants. Values from trio-phased[14] CEU individual NA12891 are shown for comparison (red). (**b**) Scores calculated in **a** for haplotype windows were compared between homologous chromosomes, and haplotypes were ranked based on the extent to which they scored as ‘GIH-like’ relative to their homolog. Plotted is the fraction of novel variants found on the more ‘GIH-like’ haplotype in rank-ordered groups of homologous haplotype windows. As above, the analysis was also performed for individual NA12891 using the rank ordering from individual NA20847. Haplotype blocks that are most differentiated relative to their homolog (higher ranked) with respect to GIH versus CEU similarity are enriched for novel variants relative to their homolog, consistent with the pattern observed in **a**.

populations, and therefore enriched for novel variants. Such haplotypes may represent a valuable source of information about human history on the South Asian subcontinent.

A substantial fraction of the human genome consists of gene-rich segmental duplications and otherwise structurally complex regions that continue to defy accurate diploid consensus assembly within individual genomes. We sought to evaluate whether haplotype-resolved sequencing is useful for the fine-mapping and haplotype-assignment of deletions, inversions and novel contigs.

We used shotgun read depth[22], discordant pairing in shotgun data[23] and array-based SNP calls[2] to estimate copy number and detect 58 deletions (>8 kbp), 15 of which were flanked by segmental duplications. Of these, 48 deletions (83%) were unambiguously confirmed by sequenced fosmid clones spanning the breakpoints, providing fine-scale resolution and confirming 30 as hemizygous (**Fig. 4a** and **Supplementary Table 3**). Heterozygous variants in flanking clones allowed for unambiguous incorporation of these deletions into haplotype-resolved blocks.

Inversions are challenging to detect because they are copy-number neutral and frequently mediated by repetitive sequences. As even fosmid end-sequencing tends to overcall inversions[6], the added information from interrogating full ~37-kbp inserts may be useful for discriminating true inversions from false positives (**Supplementary Fig. 4**). Indeed, we observed a number of unambiguous inversions by means of breakpoint-spanning clones (**Supplementary Fig. 5**). However, larger clones (>100 kbp) may be required to span the large duplication blocks where inversion breakpoints typically map[6]. NA20847 is heterozygous for the inversion-containing H2 haplotype at the *MAPT* locus (17q21) (**Supplementary Fig. 6**). Of note, we properly phased all 287 SNPs that tag the H2 haplotype across a 588-kbp span[24].

We also detected common human sequences unrepresented in the reference, that is, the ‘pan-genome’ (**Supplementary Table 4**)[7,8]. Of 16,904 contigs (total 12.8 Mbp) reported by two recent studies[7,8], we identified 8,993 in NA20847. We exploited the contiguity of fosmids to anchor ~30% of these (**Fig. 4b**), with 73% agreement (±50 kbp) with a previously anchored subset[8]. *De novo* assembly of remaining unmapped reads yielded 2,242 additional contigs after filtering, of which we anchored 396. To validate anchoring accuracy, we simulated novel insertions by deleting 600 intervals (250 bp–10 kbp) *in silico* from the reference and remapping reads to the modified reference. Unmapped reads were *de novo* assembled into 5,435 contigs that covered ~61% of simulated insertions. Of these, we predicted anchoring locations for 2,184 with an accuracy of 87%, with the remaining contigs unassigned because of limited clone coverage. The sensitivity and specificity with which novel contigs can be anchored by this approach is likely to improve with increased clone and shotgun coverage.

We recently demonstrated exome sequencing as a strategy for identifying causal variants in Mendelian disorders[25], for example, implicating compound heterozygote variants in *DHODH* in Miller syndrome[26]. In such studies, phasing reduces the number of candidate genes consistent with a recessive, compound heterozygous model[13]. For example, in this Gujarati Indian individual, unphased variant data included 44 genes consistent with compound heterozygosity (that is, two or more heterozygous, novel, nonsynonymous or splice-site variants that altered the same gene). But after phase was

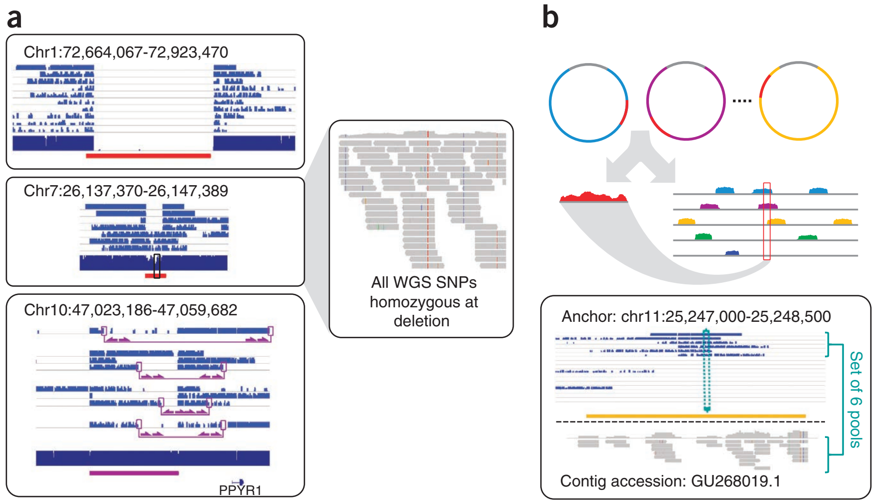

a
Chr1:72,664,067-72,923,470
Chr7:26,137,370-26,147,389
Chr10:47,023,186-47,059,682
PPYR1
All WGS SNPs homozygous at deletion
b
....
Anchor: chr11:25,247,000-25,248,500
Set of 6 pools
Contig accession: GU268019.1

Figure 4 Insertion anchoring and structural variation detection. (**a**) Homozygous deletion (top), hemizygous deletion (middle) and inversion (bottom) with fosmid clone support. Deletion calls were made using read depth and paired-read discordance. Inversions were called by paired-read discordance. SNPs within hemizygous deletions appear as stretches of hemizygosity by whole-genome shotgun sequencing. Purple connections indicate the additional support of strand discordance of read pairs spanning genomic DNA and the vector backbone. (**b**) Novel contigs not present in the reference assembly (red) but detected among clone pool–derived reads (light blue, purple, yellow) are anchored by searching for positions in the reference common to those pools but missing from most or all other pools. This approach anchors 1,733 recently reported insertion sequences[7,8] including contig GU268019.

taken into account, only ten were validated as *trans* heterozygous, with the remainder having both variants on the same haplotype.

This method requires significantly greater expertise and sample preparation than the haplotype-blind shotgun sequencing of an individual genome—specifically, the construction of a single fosmid library and >100 *in vitro* shotgun libraries, as compared with constructing one or a few *in vitro* shotgun libraries. A detailed consideration of the added effort and cost are provided in **Supplementary Table 5**. In summary, sample preparation can be completed in <2 weeks by a single technician at a cost (~$4,000) that is much greater than that of preparing a single shotgun library, but low relative to the overall cost of whole-genome sequencing. We use an unconventional method based on *in vitro* transposition[16] to significantly reduce the time and effort for producing >100 shotgun libraries. Current costs are primarily driven by commercial reagents for fosmid and shotgun library construction, and may therefore be amenable to optimization[16]. Furthermore, most steps are compatible with manual scaling and/or automation.

We also note that the total bases sequenced here (~87 Gb shotgun, ~110 Gb clone-based) is only modestly higher than for other individual human genomes sequenced to date. To estimate the minimal amount of clone sequencing required, we subsampled our data for either the number of independent clones or the depth of clone library sequencing (**Supplementary Fig. 7**). The primary effect was a reduction in the length of assembled haplotype blocks, rather than any decay in accuracy. For example, at 80% of clones and 60% of sequencing depth (which is 48% as much clone-based sequencing), the N50 dropped from 386 kbp to 238 kbp. However, most ascertained heterozygous variants remained phased (85.4%), and phasing remained highly concordant with HapMap (>99% at $D' > 0.9$). Other optimizations, for example, switching from plate-scraping to direct liquid outgrowth to improve clone uniformity (**Supplementary Fig. 1**), may further reduce sequencing requirements.

Haplotypes are essential to the information content that defines a diploid human genome, but have heretofore been intractable to genome-wide, experimental determination in the context of massively parallel sequencing. We anticipate that haplotype-resolved genome sequencing will be valuable in a broad range of scenarios, including the following. (i) Population genetics. Haplotype-resolved genome sequencing eliminates the need for population or pedigree-based haplotype inference. This will be most useful in populations that are poorly ascertained (e.g., South Asians) or have low linkage disequilibrium (e.g., Africans), and more generally for rare variants. (ii) Genetic anthropology. For example, the availability of the haplotype-resolved reference and Venter genomes was critical to the observation of a Neanderthal contribution to some modern humans[3]. (iii) Medical genetics of rare and common phenotypes. Haplotype information can facilitate the analysis of recessive Mendelian disorders[13], the determination of the parent of origin for *de novo* mutations, and the study of complex interactions among multiple SNPs[27]. (iv) Structural variation in both germline and cancer genomes. Our approach is more comprehensive than long-insert mate-pairing (whether by fosmids[6] or *in vitro* circularization[28]), as these methods determine the ends of large molecules but are blind to their internal contents. Also, the intermediate level of partitioning provided by fosmids may be more useful than whole chromosome amplification[29], as many germline and somatic structural events are intrachromosomal. (v) Allele-specific phenomena. Haplotype information may be essential for understanding the genetic basis of phenomena such as allele-specific expression and methylation[30]. (vi) *De novo* genome assembly. Massively parallel sequencing of highly complex pools of minimally redundant haploid clones may facilitate the high-quality *de novo* assembly of new genomes, an area that continues to be a major challenge for the genomics field despite the falling costs of DNA sequencing[11].

METHODS

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

Accession codes. Short read sequence data have been deposited at the NCBI Sequence Read Archive (SRA) under accession no. SRA026360. Assembled haplotype blocks and novel contigs are available from: http://krishna.gs.washington.edu/indianGenome/.

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

ACKNOWLEDGMENTS

We thank C. Lee and M. Malig for technical assistance, J. Akey, T. O'Connor and P. Green for helpful discussions, D. Reich for ancestry information on NA20847, the U.W. Genome Sciences Genomics Resource Center (GS-GRC) for sequencing and the 1000 Genomes Project for early data release. This work was supported by National Institutes of Health grants AG039173 (J.B.H.) and HG002385 (E.E.E.), a National Science Foundation Graduate Research Fellowship (J.O.K.), a Natural Sciences and Engineering Research Council of Canada Fellowship (P.H.S.) and a fellowship from the Achievement Rewards for College Scientists Foundation (J.B.H.). E.E.E. is an investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

The project was conceived and experiments planned by J.O.K., E.E.E. and J.S. J.O.K., A.P.M. and R.Q. carried out all experiments. J.O.K., A.A., J.B.H., R.P.P., P.H.S., S.B.N. and C.A. performed data analysis. J.O.K., A.P.M., A.A., J.B.H., R.P.P. and J.S. wrote the manuscript, and all authors reviewed it. All aspects of the study were supervised by J.S.

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. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
2. International HapMap Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
3. Green, R.E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
4. Anonymous. Human genome: Genomes by the thousand. *Nature* **467**, 1026–1027 (2010).
5. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
6. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
7. Li, R. *et al.* Building the sequence map of the human pan-genome. *Nat. Biotechnol.* **28**, 57–63 (2010).
8. Kidd, J.M. *et al.* Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nat. Methods* **7**, 365–371 (2010).
9. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
10. McKernan, K.J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19**, 1527–1541 (2009).
11. Schatz, M.C., Delcher, A.L. & Salzberg, S.L. Assembly of large genomes using second-generation sequencing. *Genome Res.* **20**, 1165–1173 (2010).
12. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
13. Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
14. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
15. Reich, D., Thangaraj, K., Patterson, N., Price, A.L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489–494 (2009).
16. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high density *in vitro* transposition. *Genome Biol.* **11**, R119 (2010).
17. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

18. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
19. Bansal, V. & Bafna, V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* **24**, i153–i159 (2008).
20. Kim, J.H., Waterman, M.S. & Li, L.M. Diploid genome reconstruction of Ciona intestinalis and comparative analysis with Ciona savignyi. *Genome Res.* **17**, 1101–1110 (2007).
21. Bansal, V., Halpern, A.L., Axelrod, N. & Bafna, V. An MCMC algorithm for haplotype assembly from whole-genome sequence data. *Genome Res.* **18**, 1336–1346 (2008).
22. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.* **41**, 1061–1067 (2009).
23. Hormozdiari, F. *et al.* Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* **26**, i350–i357 (2010).
24. Zody, M.C. *et al.* Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat. Genet.* **40**, 1076–1083 (2008).
25. Ng, S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
26. Ng, S.B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42**, 30–35 (2010).
27. Drysdale, C.M. *et al.* Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proc. Natl. Acad. Sci. USA* **97**, 10483–10488 (2000).
28. Korbel, J.O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
29. Ma, L. *et al.* Direct determination of molecular haplotypes by chromosome microdissection. *Nat. Methods* **7**, 299–301 (2010).
30. Tycko, B. Allele-specific DNA methylation: beyond imprinting. *Hum. Mol. Genet.* **19**, R210–R220 (2010).

ONLINE METHODS

Fosmid library pool construction. High molecular weight genomic DNA (HMW gDNA) was extracted from HapMap lymphoblastoid cell line GM20847 (Coriell) using the Gentra Puregene kit (Qiagen). A single, complex fosmid library (>2×10^6 clones) was created using the CopyControl pCC1Fos Fosmid Library Construction kit (Epicentre), as previously described[31]. After bulk infection, the library was split into 115 pools of ~5,000 clones each. Each pool was then individually expanded, either by scraping plates of infected cells and inoculating outgrowth culture, or by direct liquid outgrowth after infection. Clone DNA was extracted from each pool by alkaline lysis miniprep.

Massively parallel sequencing. Illumina-compatible shotgun sequencing libraries were prepared from each fosmid clone pool DNA and HMW gDNA using the Nextera DNA Sample Prep Kit (Epicentre), as described[16]. For each fosmid pool library, a 9-bp barcoded adaptor was added during PCR amplification[16]. Pool-derived libraries were combined before sequencing (PE76 or PE101 reads, plus index read, on an Illumina GA2x), and the index read was used to deconvolve the original clone pools from the combined reads. For unphased variant discovery, a single whole-genome shotgun library was sequenced across seven lanes (PE50 reads on an Illumina HiSeq).

Read mapping and variant discovery. Basecalling was performed with Illumina RTA v1.8 software. The resulting reads were aligned to the reference assembly (NCBI release GRCh37, UCSC release hg19) using BWA v0.5.8a[17]. The Genome Analysis Toolkit (GATK)[18] was used to recalibrate base quality scores, realign reads surrounding putative and known indels, and call single-nucleotide and indel variants from the whole-genome shotgun data. Quality filters were applied based on coverage, base and mapping quality score, and allelic and strand bias. Copy number genotypes were estimated genomewide by (G+C)-corrected read depth, as previously described[32]. Deletions >8 kbp were identified by intersecting regions of predicted copy less than 2 with split-read calls[23] and published SNP array-based calls[2] and requiring calls by two of the three methods.

Haplotype assembly. Clone coordinates were identified within each pool by searching for intervals of length 25–45 kbp with coverage significantly above background. Heterozygous SNP positions ascertained during whole-genome shotgun sequencing were regenotyped within each haploid clone pool. Clones with an excess of heterozygous positions, likely representing overlapping clones drawn from different haplotypes, were discarded. Haplotype blocks were created from overlapping clones using a custom reimplementation of HAPCUT[19], a parsimony maximization-based haplotype assembly algorithm. The effects of lower sequence coverage upon haplotype assembly accuracy and block length were simulated by leaving out a random subset of clones and/or reads.

Haplotype ancestry analysis. Phased blocks were divided into sliding windows of variants from HapMap[2] (20 SNPs/window) or the 1000 Genomes Project[14] (200 SNPs/window). For the HapMap-based comparison, similarity to GIH and CEU populations was scored for both haplotypes of NA20847 at every window based on the frequencies of alleles in NA20847 among GIH and CEU. Haplotype windows were then rank-ordered by the difference in similarity scores, such that haplotypes with high-frequency alleles among GIH but not CEU were more highly ranked. The fraction of all detected novel variants (not in dbSNP release 130) was then counted for each haplotype window for NA20847, and for comparison in the same rank-ordered windows, the trio-resolved CEU individual NA12871 (ref. 14). Pairs of homologous haplotype windows were rank ordered by differential similarity to GIH, and the fraction of novel variants on the GIH-enriched homolog was computed. For the 1000 Genomes-based comparison, haplotype windows were rank-ordered by divergence from CEU, YRI, and CHB+JPT populations and the fraction of novel variants per haplotype window computed for both NA20847 and NA12871 as before.

Pan-genome and novel contig mapping and anchoring. Whole-genome and clone pool-derived reads that did not align to the human genome reference (GRCh37/hg19) were mapped to novel contigs not present in the human reference genome assembly[7,8] to find contigs covered with ≥50 bp (phred-scaled mapping score ≥Q20). A subset of contigs were anchored by ≥2 reads with mates mapping to the reference. As further evidence of anchoring, intervals were identified in the reference assembly having read depth from clone pools also hitting a given contig but depleted among those pools not hitting that contig. Further novel sequences from NA20847 were assembled *de novo* from remaining unmapped reads using Velvet[33]. Contigs aligning to existing pan-genome sequences and contaminating sequences (*E. coli*, vector backbone, Epstein-Barr virus) were removed and remaining contigs were anchored as above. Sensitivity to detect and accurately anchor novel sequence was simulated by introducing *in silico* deletions into the reference, *de novo* assembling corresponding insertion contigs, anchoring as before, and measuring agreement between predicted anchoring location and the known site of simulated deletion.

31. Raymond, C.K. *et al.* Targeted, haplotype-resolved resequencing of long segments of the human genome. *Genomics* **86**, 759–766 (2005).
32. Sudmant, P.H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).
33. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).

doi:10.1038/nbt.1740

nature biotechnology

Targeted integration in rat and mouse embryos with zinc-finger nucleases

Xiaoxia Cui, Diana Ji, Daniel A Fisher, Yumei Wu, David M Briner & Edward J Weinstein

Gene targeting is indispensible for reverse genetics and the generation of animal models of disease. The mouse has become the most commonly used animal model system owing to the success of embryonic stem cell–based targeting technology[1], whereas other mammalian species lack convenient tools for genome modification. Recently, microinjection of engineered zinc-finger nucleases (ZFNs) in embryos was used to generate gene knockouts in the rat[2,3] and the mouse[4] by introducing nonhomologous end joining (NHEJ)-mediated deletions or insertions at the target site. Here we use ZFN technology in embryos to introduce sequence-specific modifications (knock-ins) by means of homologous recombination in Sprague Dawley and Long-Evans hooded rats and FVB mice. This approach enables precise genome engineering to generate modifications such as point mutations, accurate insertions and deletions, and conditional knockouts and knock-ins. The same strategy can potentially be applied to many other species for which genetic engineering tools are needed.

Conventional gene targeting in mouse embryonic stem (ES) cells is achieved by introduction of an antibiotic selection marker through homologous recombination. Targeted ES cells are then injected into wild-type blastocysts to generate chimeric animals, some of which contain targeted germ cells[5]. Time-consuming backcrossing is often necessary when ES cells are not available from the desired strain[5]. Moreover, in species without established ES cell lines, targeted gene modification is not feasible, largely limiting their use as model systems. For example, the rat is a preferred model over mice for studying many human diseases[6] but has lacked robust genetic modification tools until the application of ZFNs[2,3]. Although considerable progress has been made recently with rat ES cells[7,8], ZFN technology may overcome the limitations of ES cell technology.

ZFNs generate sequence-specific double-strand breaks[9,10] that are repaired mainly by either error-prone nonhomologous end joining (NHEJ) or high-fidelity homologous recombination (**Supplementary Fig. 1**). Embryonic injection of ZFNs has produced NHEJ-mediated knockout rats[2,3], mice[4] and zebrafish[11–13] with remarkable efficiency and germline transmission rates. However, mutations are unpredictable owing to the variable nature of DNA repair by NHEJ[14] and are limited to knockouts. On the other hand, successful homologous recombination in embryos using a homologous donor template provides accuracy and flexibility that the NHEJ process lacks, and enables gene addition.

Homologous recombination–mediated targeted integration was observed in mouse embryos after injection of donor DNA into eggs at a rate of <0.2%[15]. ZFN-mediated double-strand breaks have been shown in cultured human cells[16–18] and flies[19] to stimulate homologous recombination by several orders of magnitude. We set out to test and successfully achieved robust ZFN-assisted homologous recombination in both rat and mouse embryos.

Based on previous data in human cell lines[16,18], we first constructed donors with an eight base pair (bp) NotI restriction site inserted in between the ZFN binding sites, flanked by ~800 bp of immediate homology on each side (**Fig. 1a** and **Supplementary Tables 1** and **2**). Donor plasmid DNA and respective ZFN mRNA were co-injected into the pronucleus of one-cell embryos from Sprague Dawley rats and FVB mice followed by transfer of the injected eggs to pseudopregnant females.

Fetuses of NotI donors were harvested for analysis. Integration of the NotI site was detected using NotI digestion of PCR products of the target region amplified with primers outside of the homologous arms (F and R as shown in **Fig. 1a**). **Figure 1b** shows the expected digestion pattern of NotI integration in fetuses at the rat *Mdr1a* (1 of 15 13-day-old Sprague Dawley rat fetuses) and *PXR* loci (1 of 8 14-day-old Sprague Dawley rat fetuses) and the mouse *Mdr1a* locus (1 of 4 12.5-day-old FVB mouse fetuses). Sequencing of the PCR products confirmed the presence of a NotI site in all three loci, as well as deletions by NHEJ at both rat and mouse *Mdr1a* loci (**Supplementary Fig. 2**), indicating that these fetuses were mosaics. In addition, the *Mdr1a* locus was also successfully targeted in Long-Evans hooded rats. One of seven pups was identified as a founder (**Supplementary Fig. 3**), harboring two alleles: NotI insertion in one and an 11-bp deletion in the other. When the founder was bred to a wild-type Long-Evans hooded rat, 5 out of 12 F1 pups inherited the NotI allele, and 7 contained the 11-bp deletion allele.

Next, we constructed GFP donors, replacing the NotI site with a 1.5-kilobase (kb) human phosphoglycerate kinase (PGK) promoter-driven GFP cassette (**Fig. 2a** and **Supplementary Tables 1** and **2**). GFP is in the opposite orientation of transcription for the rat *Mdr1a* and *PXR* loci and the same orientation of transcription at the mouse *Mdr1a* locus. We analyzed live-born pups of GFP donors. DNA extracted from toe clips was amplified in four PCR reactions using primer sets (i) GF and GR, (ii) F and R, (iii) F and GF and (iv) R and GR (**Fig. 2a**). Set (i) amplified the GFP cassette, whereas set (ii) amplified the target region, favoring wild-type, deletion and small-insertion alleles over

Sigma Advanced Genetic Engineering (SAGE) Labs, Sigma-Aldrich Biotechnology, St. Louis, Missouri, USA. Correspondence should be addressed to X.C. (xiaoxia.cui@sial.com).

Received 9 April; accepted 10 November; published online 12 December 2010; doi:10.1038/nbt.1731

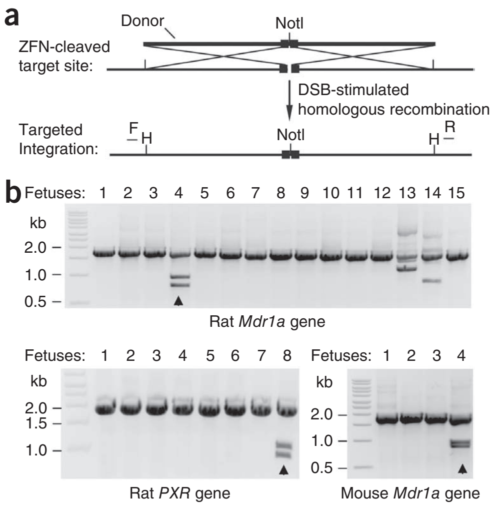


Figure 1 Targeted integration of NotI restriction site. (**a**) Schematic of donor and target site. Donors contain a NotI site inserted between the ZFN binding sequences (squares) with two flanking 800 bp homologous arms. F and R, forward and reverse primers (short bar) that sit outside of the homology. H, boundary of homology. DSB, double-strand break. (**b**) One pup (arrowhead) with NotI insertion was identified in each target using PCR with specific F and R primers followed by NotI digestion.

the targeted integration allele that is 1.5 kb larger than the wild type and rarely amplified by F and R primers when other alleles are present. Set (ii) also served as a positive control for genomic DNA quality. Sets (iii) and (iv) amplified the 5′ and 3′ respective junctions specific to targeted insertion and were the diagnostic reactions for targeted integration events. Expected product sizes are listed in **Supplementary Table 2**. **Figure 2b** shows that *Mdr1a* pup no. 3 and *PXR* pup no. 4 were positive for GFP and both junctions and contained the targeted insertion, whereas *Mdr1a* pup no. 19 was positive for GFP only, thus carrying a transgene. Sequencing revealed that *Mdr1a* pup no. 3 was a mosaic with three alleles: the targeted integrant and deletions of 513 bp and 6 bp, respectively (**Supplementary Table 3**). The complete panel of PCR reactions is shown in **Supplementary Figure 4**. NHEJ events in the pups were further analyzed (**Supplementary Fig. 5**). Using the same donor configuration, the mouse *Mdr1a* locus was targeted at 5% efficiency (**Supplementary Fig. 6**). The sequences of all junction PCR products were validated.

Southern blot analysis further confirmed integration of the GFP cassette to the target loci. Both flanking and internal probes were used. The flanking target probes (labeled as Mrd1a and PXR, respectively) were located outside of the homologous arms and hybridized to all alleles, whereas the internal probe (GFP) detected the GFP cassette (**Fig. 2a**). At the rat *Mdr1a* locus, the GFP probe recognized a single 4.5-kb band corresponding to targeted integration in pup no. 3 but not in the wild type (**Fig. 2c**), demonstrating that the GFP cassette was specifically inserted into the desired site, whereas the *Mdr1a* probe detected a single wild-type band of 3 kb in the wild-type sample, and three bands in pup no. 3, corresponding to the integration allele (4.5 kb), and the alleles with 6- and 513-bp deletions, respectively. PGK-GFP was expressed and visually detectable in the eyes of founder no. 3 under UV light (**Fig. 2d**). Founder no. 3 was then mated to a wild-type Sprague Dawley male. Approximately 50% of the F1 offspring inherited the targeted integration allele (**Fig. 3a**, **Supplementary Fig. 7** and **Supplementary Table 4**). Mating between integration-positive F1 animals generated homozygous F2 offspring that appeared normal. Southern blot analysis of the F2 animals is shown in **Figure 3b**.

For *PXR* founder no. 4, the flanking probe recognized two bands corresponding to the wild-type (4.7 kb) and integration (6.2 kb) alleles (**Fig. 2c**). However, there is an extra band around 4 kb hybridizing to the GFP probe that could have resulted from random integration (as in *Mdr1a* pup no. 19). When founder no. 4 was mated to wild-type Sprague Dawley females, ~50% of the F1 offspring was heterozygous for the GFP targeted integration allele, some of which also inherited the extra GFP locus. But the majority of F1 animals contained only the targeted integration allele (**Fig. 3c** and **Supplementary Table 4**).

Additional injection sessions produced another *Mdr1a* founder, no. 4-5, and two more *PXR* founders, no. 2-1 and no. 2-2 (**Supplementary Table 3** and **Supplementary Fig. 8**). Founder no. 4-5 contained a 147-bp deletion in addition to the targeted integration allele. Founder no. 2-1 contained a targeted integration allele and a wild-type allele. Normal breeding yielded ~50% F1 offspring heterozygous for the targeted integration allele (not shown). Founder no. 2-2 contained the targeted integration allele, an allele with a 236-bp deletion and an extra GFP locus.

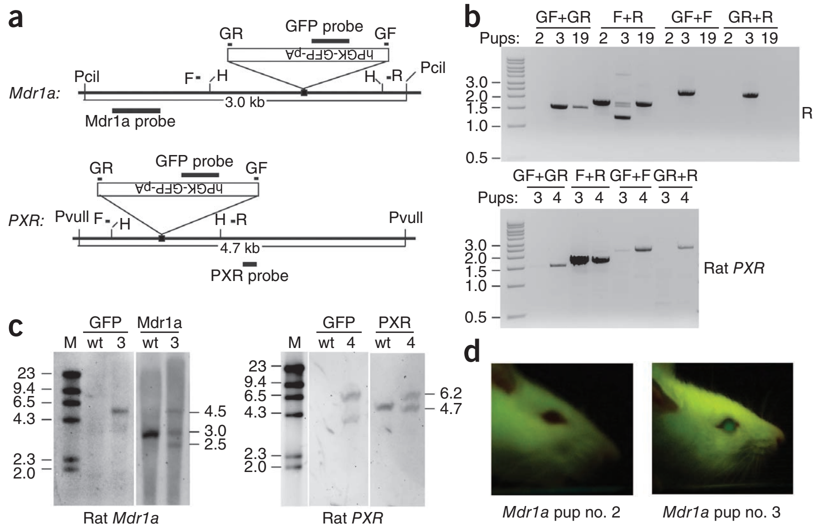


Figure 2 Targeted integration of a GFP cassette. (**a**) Schematic of target site (square) and GFP integration at *Mdr1a* and *PXR* loci. F, R and H, same as in **Figure 1a**. GF and GR, forward and reverse primers in GFP cassette. PvuII and PciI are restriction enzymes used in Southern blot analysis; neither cuts the 1.5-kb GFP insert, which inserts in the opposite orientation of transcription in both donors. Probes used in Southern blot analysis (thick bars) are marked at corresponding positions. (**b**) PCR analysis of GFP integration in selected *Mdr1a* and *PXR* rat pups. Pup IDs are labeled under the primers used. (**c**) Southern blot analysis of pups for GFP integration. GFP, GFP probe; Mdr1a and PXR, respective flanking probes. wt, wild-type Sprague Dawley genomic DNA. (**d**) GFP expression was visualized under UV light in the eyes of *Mdr1a* founder no. 3. Full-length blots are presented in **Supplementary Figure 9**.

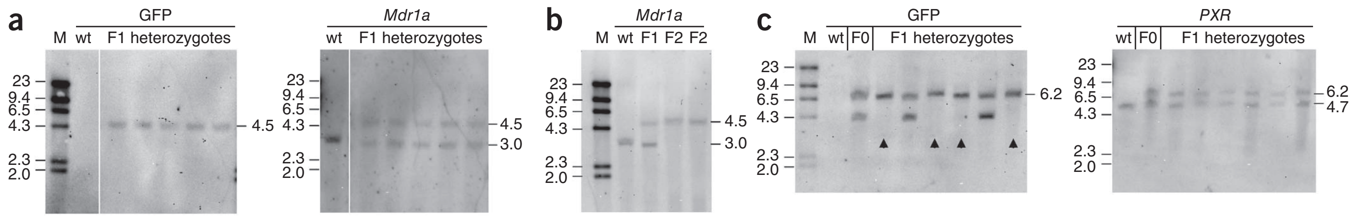


Figure 3 Germline transmission of site-specific GFP integration. Integration-positive offspring of *Mdr1a* founder no. 3 (F1 and F2 pups identified by PCR as in **Supplementary Fig. 7a** and not shown) and *PXR* founder no. 4 (F1 pups identified by PCR as in **Supplementary Fig. 7b**) were further confirmed with Southern blot analysis. GFP, GFP probe; Mdr1a and PXR, respective target probes. wt, wild-type Sprague Dawley genomic DNA. (**a**) *Mdr1a* heterozygotes. (**b**) *Mdr1a* F2 homozygotes. Wild-type and an F1 heterozygote are included as controls. (**c**) *PXR* F1 heterozygotes. Arrowheads indicate F1 animals in which targeted integration allele of *PXR* was segregated from the extra GFP locus. Full-length blots are presented in **Supplementary Figure 9**.

Table 1 summarizes the injection statistics that support the following conclusions: first, offspring of animals co-injected with ZFN mRNA and donor plasmid had similar overall survival and live-birth rates, as previously reported[2,20], indicating minimal toxicity. In addition, mutant animals appeared to be physically normal and bred well, except for *PXR* founder no. 2-2, which developed hydrocephalus of unknown cause and had to be euthanized. Second, NHEJ occurs at a higher rate than targeted integration. Thus, the presence of NHEJ-positive pups among live births should be used as a criterion for a successful injection session. For example, between *Mdr1a*/GFP and *PXR*/GFP injections, four sessions failed to produce founders with targeted integration, three of which did not generate any NHEJ-positive pups, implying possible variance in sample preparation and/or injection. **Table 1** combined data from all sessions, including those that probably failed.

Mosaicism is common among mutant animals produced with ZFNs and has been observed in knockout rats[2,3] and mice[4], where up to five different alleles were detected in individual founders. **Supplementary Table 3** summarizes the genotype of all targeted, integration-positive animals generated in this study, some of which contained up to three alleles. The degree of mosaicism in the founders likely correlates to the length of time ZFNs remain active in the embryos. Each cell division doubles the number of existing targetable (wild-type) alleles, allowing more independent NHEJ events to create different mutant alleles. In founders carrying more than two alleles, ZFNs must remain functional beyond the one-cell embryo stage.

Germline transmission of NHEJ-modified and targeted integration alleles is highly efficient. All alleles identified in founders *Mdr1a* no. 3 and *PXR* no. 4 were inherited in the F1 generation (**Supplementary Table 4**). The high germline transmission rate is consistent with the fact that mosaicism develops in embryos containing only a few cells, all of which have the potential to become germ cells, including those carrying alleles with low representation in the body. For example, sequencing of the PCR products in **Figure 2b** identified a 35-bp deletion only at the *PXR* locus with no wild-type allele in founder no. 4. However, 2 of the 57 F1 pups were wild type (**Supplementary Tables 3** and **4**). Overall, allele distribution among F1 offspring correlated roughly to the relative ratio among alleles detected in Southern blot analysis (**Fig. 2c** and **Supplementary Table 4**).

To our knowledge there has been no previous report of robust targeted integration in rat embryos of different backgrounds and complete germline transmission. While this manuscript was in revision, another group reported a similar targeting strategy in mice[21]. The same method may enable one to introduce precise modifications, such as point mutations, specific insertions and deletions, gene replacement, conditional knock-ins and knockouts, to an exact locus directly in embryos that then develop into mutant animals. Nevertheless, there are potential limitations to ZFN technology. Primarily, ZFNs have yet to be engineered to target any given sequence, which may limit the ability to introduce mutations at loci lacking ZFN target sites in the vicinity. Continuous improvement of ZFN design is necessary. Second, undesired modifications by ZFNs are also possible and have been detected at low rates in cultured human cells[22–24], although none have been observed in ZFN-engineered rodents so far[2–4]. In the meantime, advances in ZFN engineering continue to improve specificity[18,24–26]. In addition, unwanted mutations may segregate from the target loci and be eliminated from subsequent generations by breeding (**Fig. 3c** and **Supplementary Table 4**). Recently, the disruption of the p53 gene in a rat ES cell line and germline transmission were described[7]. In a separate report, germline transmission of a transgene was also demonstrated[8]. These are major improvements in ES cell technology that could help replicate in the rat the sophisticated targeting strategies that are already well developed in the mouse. However, ZFN technology possesses several advantages. First, ZFN-mediated homologous recombination in embryos does not require selectable markers. Second, the time frame needed to obtain mutant animals is shortened by bypassing ES cells and efficient germline transmission. More importantly, gene targeting using ZFN technology is not limited by the availability of ES cells, and time-consuming backcrossing is avoided. Finally, in theory, ZFN technology can be applied to any organism for which fertilized eggs can be collected, microinjected and transferred into pseudopregnant females.

Table 1 Injection statistics

Target	Donor	Strain	Embryos injected	Embryos transferred	Fetuses (f) or pups (p)	TI positive (mutation rate in %)	NHEJ positive (mutation rate in %)
Mdr1a	NotI	SD	97	81	15f	1 (6.7)	3 (2.0)
	MCS	LEH	125	80	7p	1 (14.3)	2 (28.6)
	GFP	SD	636	439	83p	2 (2.4)	21 (25.3)
	NotI	FVB	46	46	4f	1 (25.0)	3 (75.0)
	GFP	FVB	106	106	40f	2 (5.0)	3 (7.5)
PXR	NotI	SD	56	52	8f	1 (12.5)	3 (37.5)
	GFP	SD	670	472	36p	3 (8.3)	4 (11.1)

NotI: donor constructs with NotI site inserted in between the homologous arms (**Fig. 1a**). GFP: donor constructs with GFP cassette inserted in between the homologous arms (**Fig. 2a**). MCS: *Mdr1a* donor construct with multiple cloning site inserted in the NotI site in NotI donor. SD: Sprague Dawley rats; FVB, FVB/NTac mice; LEH: Long-Evans hooded rats. TI positive: the number of pups or fetuses harboring targeted integration; NHEJ positive: the number of pups or fetuses positive in mutation detection assay.

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 J. Books for technical assistance; R. DeKelver at Sangamo Biosciences for GFP plasmid; A. Harrington and L. Liaw for mouse microinjections; Z. Cui, G. Davis and T. Saunders for discussion; D. Carroll, K. Cunningham, T. Collingwood and F. Urnov for suggestions; and S. Fine for excellent assistance with figures.

AUTHOR CONTRIBUTIONS

X.C. designed the project, built donor constructs and carried out some of the diagnostic PCRs; D.J. built donor constructs and performed most of the PCR diagnosis; D.A.F. performed Southern blot analysis; Y.W. microinjected the rat samples; D.M.B. assembled ZFNs; X.C., D.J., D.A.F. and E.J.W. wrote 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. Capecchi, M.R. Gene targeting in mice: functional analysis of the mammalian genome for the twenty-first century. *Nat. Rev. Genet.* **6**, 507–512 (2005).
2. Geurts, A.M. *et al.* Knockout rats via embryo microinjection of zinc-finger nucleases. *Science* **325**, 433 (2009).
3. Mashimo, T. *et al.* Generation of knockout rats with X-linked severe combined immunodeficiency (X-SCID) using zinc-finger nucleases. *PLoS ONE* **5**, e8870 (2010).
4. Carbery, I.D. *et al.* Targeted genome modification in mice using zinc finger nucleases. *Genetics* **186**, 451–459 (2010).
5. Ledermann, B. Embryonic stem cells and gene targeting. *Exp. Physiol.* **85**, 603–613 (2000).
6. Aitman, T.J. *et al.* Progress and prospects in rat genetics: a community view. *Nat. Genet.* **40**, 516–522 (2008).
7. Tong, C., Li, P., Wu, N.L., Yan, Y. & Ying, Q.L. Production of p53 gene knockout rats by homologous recombination in embryonic stem cells. *Nature* **467**, 211–213 (2010).
8. Kawamata, M. & Ochiya, T. Generation of genetically modified rats from embryonic stem cells. *Proc. Natl. Acad. Sci. USA* **107**, 14223–14228 (2010).
9. Kim, Y.-G., Cha, J. & Chandrasegaran, S. Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proc. Natl. Acad. Sci. USA* **93**, 1156–1160 (1996).
10. Mani, M., Smith, J., Kandavelou, K., Berg, J.M. & Chandrasegaran, S. Binding of two zinc finger nuclease monomers to two specific sites is required for effective double-strand DNA cleavage. *Biochem. Biophys. Res. Commun.* **334**, 1191–1197 (2005).
11. Doyon, Y. *et al.* Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases. *Nat. Biotechnol.* **26**, 702–708 (2008).
12. Meng, X., Noyes, M.B., Zhu, L.J., Lawson, N.D. & Wolfe, S.A. Targeted gene inactivation in zebrafish using engineered zinc-finger nucleases. *Nat. Biotechnol.* **26**, 695–701 (2008).
13. Foley, J.E. *et al.* Rapid mutation of endogenous zebrafish genes using zinc finger nucleases made by oligomerized pool engineering (OPEN). *PLoS ONE* **4**, e4348 (2009).
14. Lieber, M.R. The biochemistry and biological significance of nonhomologous DNA end joining: an essential repair process in multicellular eukaryotes. *Genes Cells* **4**, 77–85 (1999).
15. Brinster, R.L. *et al.* Targeted correction of a major histocompatibility class II E alpha gene by DNA microinjected into mouse eggs. *Proc. Natl. Acad. Sci. USA* **86**, 7087–7091 (1989).
16. Moehle, E.A. *et al.* Targeted gene addition into a specified location in the human genome using designed zinc finger nucleases. *Proc. Natl. Acad. Sci. USA* **104**, 3055–3060 (2007).
17. Porteus, M.H. & Baltimore, D. Chimeric nucleases stimulate gene targeting in human cells. *Science* **300**, 763 (2003).
18. Urnov, F.D. *et al.* Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **435**, 646–651 (2005).
19. Bozas, A., Beumer, K.J., Trautman, J.K. & Carroll, D. Genetic analysis of zinc-finger nuclease-induced gene targeting in *Drosophila*. *Genetics* **182**, 641–651 (2009).
20. Filipiak, W.E. & Saunders, T.L. Advances in transgenic rat production. *Transgenic Res.* **15**, 673–686 (2006).
21. Meyer, M., de Angelis, M.H., Wurst, W. & Kühn, R. Gene targeting by homologous recombination in mouse zygotes mediated by zinc-finger nucleases. *Proc. Natl. Acad. Sci. USA* **107**, 15022–15026 (2010).
22. DeKelver, R.C. *et al.* Functional genomics, proteomics, and regulatory DNA analysis in isogenic settings using zinc finger nuclease-driven transgenesis into a safe harbor locus in the human genome. *Genome Res.* **20**, 1133–1142 (2010).
23. Hockemeyer, D. *et al.* Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases. *Nat. Biotechnol.* **27**, 851–857 (2009).
24. Perez, E.E. *et al.* Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nat. Biotechnol.* **26**, 808–816 (2008).
25. Miller, J.C. *et al.* An improved zinc-finger nuclease architecture for highly specific genome editing. *Nat. Biotechnol.* **25**, 778–785 (2007).
26. Szczepek, M. *et al.* Structure-based redesign of the dimerization interface reduces the toxicity of zinc-finger nucleases. *Nat. Biotechnol.* **25**, 786–793 (2007).

ONLINE METHODS

Rat work in this study was performed at SAGE Labs, which operated under approved animal protocols overseen by SAGE's Institutional Animal Care and Use Committee (IACUC). Mouse work in this study was a contracted service provided by Maine Medical Center Research Institute (MMCRI) transgenic and gene targeting core facility, which operated under approved animal protocols overseen by MMCRI's IACUC.

ZFN constructs. The design and assembly of ZFNs were described previously[11,18]. The obligate-heterodimer form of ZFNs was used throughout[25]. Full ZFN amino acid sequences and DNA sequences are provided in **Supplementary Methods**. ZFN binding sites below are underlined, and the spacers between the binding sites are in bold. The same pair of ZFNs was used to target both the mouse and rat *Mdr1a* loci due to sequence conservation.

Mdr1a: 5′-GCCATCAGCCCT**GTTCTT**GGACTGTCAGCTGGT
CGGTAGTCGGGA**CAAGAA**CCTGACAGTCGACCA-5′
PXR: 5′-CAAATCTGCCGTGTA**TGTGGG**GACAAGGCCAATGGCT
GTTTAGACGGCACAT**ACACCC**CTGTTCCGGTTACCGA-5′

Donor construction. Homologous arms were PCR amplified from FVB mouse or Sprague Dawley rat genomic DNA. The primers used are listed in **Supplementary Table 1**. All donors used pBluescript SK (+) backbone (Stratagene). In NotI donors, left arms were cloned into KpnI and NotI sites, and right arms, NotI and SacII sites. GFP donors were constructed by inserting the PCR-amplified PGK-GFP cassette into the NotI site in the respective NotI donors in either direction. Donor plasmid was purified using GenElute Endotoxin-Free plasmid maxiprep kit (Sigma) and quantified on a Nanodrop.

mRNA preparation for microinjection. ZFN constructs were linearized at the XbaI site. *In vitro* transcripts were generated using T7 MessageMax kit (Epicentre) and polyA tailing kit (Epicentre), following manufacturer's instructions, and precipitated with an equal volume of 5 M ammonium acetate by incubating on ice for 15 min followed by centrifugation at >15,000*g* at 4 °C for 15 min. Washed and dried RNA was dissolved in water, and concentration was determined using a Nanodrop. mRNA is then transfected to cultured cells to validate activity using mutation detection assay.

Mutation detection assay. Sequencing primers were used in pairs to amplify a 300–400 bp region surrounding the target site. Ten microliters of each PCR product was then incubated using the following program: 95 °C, 10 min, 95 °C to 85 °C, at −2 °C/sec, 85 °C to 25 °C at −0.1 °C/sec. One microliter each of nuclease S and enhancer (Transgenomic) was added to digest the above reaction at 42 °C for 20 min. The mixture is resolved on a 10% polyacrylamide TBE gel (Bio-Rad).

Microinjection of fertilized eggs. At SAGE Labs, Sprague Dawley rats purchased from Charles River Laboratories were housed in standard cages and maintained on a 12 h light/dark cycle with *ad libitum* access to food and water. Four- to five-week-old donors were injected with 20 units of pregnant mare serum gonadotropin (PMS) followed by 50 units of human chorionic gonadotropin (hCG) injection after 48 h and again before mating. Fertilized eggs were harvested a day later for injection. ZFN mRNA and donor DNA were mixed and injected into the pronucleus of fertilized eggs. The final concentration of each ZFN mRNA was 2.5 ng/µl, and that of donor DNA was 1 ng/µl. Recipients were injected with 40 µg of LH-Rh 72 h before mating. Microinjected eggs were transferred to pseudopregnant Sprague Dawley recipients.

At MMCRI, FVB/NTac mice were housed in static cages and maintained on a 14 h and 10 h light/dark cycle with *ad libitum* access to food and water. Three- to four-week-old females were injected with 5 units of PMS and 48 h later, with 5 units of hCG. Fertilized eggs were harvested 10–12 h after hCG injection for microinjection. ZFN mRNA and donor DNA were mixed and injected into the pronucleus of fertilized eggs. Each ZFN mRNA was at a final concentration of 1 ng/µl, and the donor DNA at 2 ng/µl. Microinjected eggs were transferred to pseudopregnant Swiss Webster (SW) recipients, which receive 40 µg of LH-Rh injection 72 h before mating.

PCR and NotI digestion conditions. QuickExtract (Epicentre) was used to extract DNA from tail or toe clips, following manufacturer's instructions. Accuprime High Fidelity Taq polymerase (Invitrogen) was used in all PCR reactions with cycling conditions recommended by the manufacturer. NotI digestion was done by adding 1 µl 10× BSA and 1 µl NotI to 8 µl of PCR reaction and incubating at 37 °C for 2 h.

Preparation of genomic DNA for Southern blot analysis. Tail clips or ear notches were used to prepare genomic DNA for Southern blot analysis. Tissues were first incubated in lysis buffer (100 mM Tris-HCl, pH 8.8; 50 mM EDTA, 0.5% SDS; 200 mM NaCl; 300 µg/ml proteinase K) at 55 °C for 2 to 5 h with occasional inversions. Supernatant was collected and precipitated. The washed and dried pellet was then dissolved in TE buffer (10 mM Tris-HCl, pH 8.0, 1 mM EDTA).

Probe labeling. The probes were labeled with PCR DIG Probe Synthesis Kit (Roche) with Sprague Dawley rat genomic DNA as template and primers listed in **Supplementary Methods.**

Southern blot analysis. Fifteen micrograms of genomic DNA was digested for 3 h to overnight at 37 °C with PciI (*Mdr1a*) or PvuII (*PXR*), concentrated by precipitation and resolved on a 0.8% agarose gel. Upon transfer to nylon membrane and UV cross-linking (120,000 µJ/cm^2), prehybridization and hybridization were carried out according to instructions for DIG Easy Hyb Granules (Roche) at 42 °C (GFP probe) or 37 °C (PXR or Mdr1a probes). The membrane was then developed using DIG Detection Kit (Roche), following manufacturer's instructions. The developed membrane was exposed to a ChemiDocXRS+ Imaging system (BioRad).

Visualizing GFP in founder rats. BlueStar high intensity LED flashlight and BlueBlock filter glasses (Nightsea) were used to visualize GFP expression in the *Mdr1a* and *PXR* GFP founder rats.

doi:10.1038/nbt.1731

nature
biotechnology

Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine

Chun-Xiao Song[1], Keith E Szulwach[2], Ye Fu[1], Qing Dai[3], Chengqi Yi[1], Xuekun Li[2], Yujing Li[2], Chih-Hsin Chen[4], Wen Zhang[1], Xing Jian[1], Jing Wang[1], Li Zhang[4], Timothy J Looney[4], Baichen Zhang[5], Lucy A Godley[6], Leslie M Hicks[5], Bruce T Lahn[4], Peng Jin[2] & Chuan He[1]

In contrast to 5-methylcytosine (5-mC), which has been studied extensively[1–3], little is known about 5-hydroxymethylcytosine (5-hmC), a recently identified epigenetic modification present in substantial amounts in certain mammalian cell types[4,5]. Here we present a method for determining the genome-wide distribution of 5-hmC. We use the T4 bacteriophage β-glucosyltransferase to transfer an engineered glucose moiety containing an azide group onto the hydroxyl group of 5-hmC. The azide group can be chemically modified with biotin for detection, affinity enrichment and sequencing of 5-hmC–containing DNA fragments in mammalian genomes. Using this method, we demonstrate that 5-hmC is present in human cell lines beyond those previously recognized[4]. We also find a gene expression level–dependent enrichment of intragenic 5-hmC in mouse cerebellum and an age-dependent acquisition of this modification in specific gene bodies linked to neurodegenerative disorders.

Parallel to the discovery of 5-hmC in mammalian genomes[4,5], Tet proteins were shown to use dioxygen to oxidize 5-mC to 5-hmC in mammalian DNA[5]. Tet proteins are a group of iron(II)/α-ketoglutarate–dependent dioxygenases similar to the AlkB family proteins and hypoxia-inducible factor (HIF) prolyl-hydroxylases[6,7]. As Tet1 and Tet2 appear to affect embryonic stem (ES) cell maintenance and normal myelopoiesis, respectively[8,9], these findings fostered speculation that this 5-hmC modification might also be an important epigenetic modification[10].

To elucidate the biology of 5-hmC, the first step is to identify the locations of 5-hmC within genomic DNA, but so far it has remained challenging to distinguish 5-hmC from 5-mC and to enrich 5-hmC-containing genomic DNA fragments.

Widely used methods to probe 5-mC, such as bisulfite sequencing and methylation-sensitive restriction digestion, cannot discriminate between 5-hmC and 5-mC[11,12]. Anti-5-hmC antibodies have only recently become commercially available. However, attempts to use the antibodies to immuno-enrich 5-hmC-containing genomic DNA from complex genomes for sequencing have yet to be successful[8]. A single-molecule, real-time sequencing technology has been applied to distinguish between cytosine, 5-mC and 5-hmC, but further improvements are necessary to affinity-enrich 5-hmC–containing DNA and to achieve base-resolution sequencing[13].

Here we present a chemical tagging technology to address both challenges. It has been shown that 5-hmC is present in the genome of the T-even bacteriophages. A viral enzyme, β-glucosyltransferase (β-GT), can catalyze the transfer of a glucose moiety from uridine diphosphoglucose (UDP-Glu) to the hydroxyl group of 5-hmC, yielding β-glucosyl-5-hydroxymethyl-cytosine (5-gmC) in duplex DNA[14,15] (**Fig. 1a**). We took advantage of this enzymatic process and used β-GT to transfer a chemically modified glucose, 6-N_3-glucose, onto 5-hmC for selective bio-orthogonal labeling of 5-hmC in genomic DNA (**Fig. 1b**). With an azide group present, a biotin tag or any other tag can be installed using Huisgen cycloaddition (click) chemistry for a variety of enrichment, detection and sequencing applications[16–18].

We used the biotin tag for high-affinity capture and/or enrichment of 5-hmC–containing DNA for sensitive detection and deep sequencing to reveal genomic locations of 5-hmC (**Fig. 1b**). The covalent chemical labeling coupled with biotin-based affinity purification provides considerable advantages over noncovalent, antibody-based immunoprecipitation as it ensures accurate and comprehensive capture of 5-hmC–containing DNA fragments, while still providing high selectivity.

We chemically synthesized UDP-6-N_3-Glu (**Supplementary Fig. 1** and **Supplementary Methods**) and attempted the glycosylation reaction of an 11-mer duplex DNA containing a 5-hmC modification as a model system (**Fig. 2**). Wild-type β-GT worked efficiently using UDP-6-N_3-Glu as the co-factor, showing only a sixfold decrease of the reaction rate compared to the native co-factor UDP-Glu (**Supplementary Fig. 2**). The 6-N_3-glucose transfer reaction finished within 5 min with as low as 1% enzyme concentration. The identity of the resulting β-6-azide-glucosyl-5-hydroxymethyl-cytosine (N_3-5-gmC) of the 11-mer DNA was confirmed by matrix-assisted laser desorption/ionization–time of flight (MALDI-TOF) analysis (**Fig. 2**). One can readily couple N_3-5-gmC with dibenzocyclooctyne-modified biotin (compound **1**) by copper-free

[1]Department of Chemistry and Institute for Biophysical Dynamics, the University of Chicago, Chicago, Illinois, USA. [2]Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia, USA. [3]Department of Biochemistry and Molecular Biology, the University of Chicago, Chicago, Illinois, USA. [4]Department of Human Genetics and Howard Hughes Medical Institute, the University of Chicago, Chicago, Illinois, USA. [5]Donald Danforth Plant Science Center, St. Louis, Missouri, USA. [6]Section of Hematology/Oncology, the University of Chicago, Chicago, Illinois, USA. Correspondence should be addressed to C.H. (chuanhe@uchicago.edu) or P.J. (peng.jin@emory.edu).

Received 25 August; accepted 15 November; published online 12 December 2010; doi:10.1038/nbt.1732

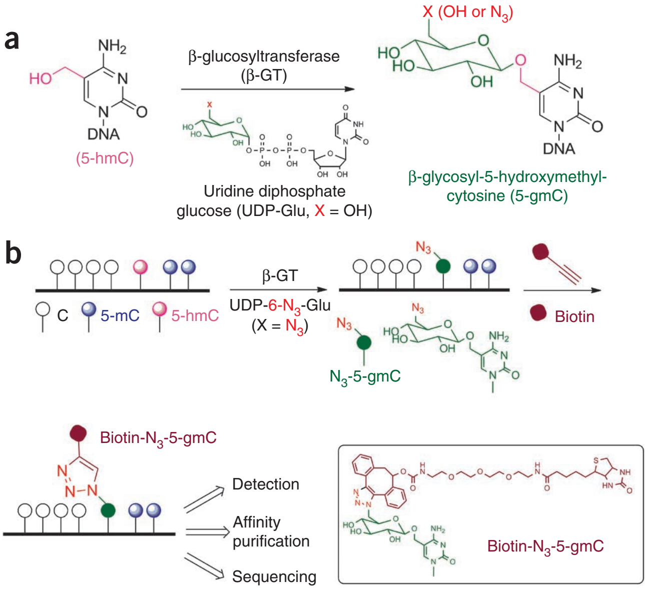


Figure 1 Selective labeling of 5-hmC in genomic DNA. (**a**) The hydroxyl group of 5-hmC in duplex DNA can be glucosylated by β-GT to form β-glucosyl-5-hydroxymethylcytosine (5-gmC) using UDP-Glu as a co-factor. (**b**) An azide group can be installed onto 5-hmC using chemically modified UDP-Glu (UDP-6-N_3-Glu), which in turn can be labeled with a biotin moiety using click chemistry for subsequent detection, affinity purification and sequencing.

click chemistry to introduce a biotin group (**Fig. 2**)[19,20]. Again, the identity of the 11-mer DNA with the biotin-N_3-5-gmC label was confirmed by MALDI-TOF analysis (**Fig. 2**). High-performance liquid chromatography (HPLC) analysis indicated that the click chemistry is high yielding (~90%) (**Supplementary Fig. 3**). High-resolution mass spectroscopy (HRMS) analysis of the corresponding HPLC hydrolysates further verified that biotin-N_3-5-gmC was formed (**Supplementary Fig. 4**).

The properties of 5-hmC in duplex DNA are quite similar to those of 5-mC in terms of its sensitivity toward enzymatic reactions such as restriction enzyme digestion and polymerization[13–15]. In an attempt to develop a method to differentiate these two bases in DNA, primer extension with a biotin-N_3-5-gmC–modified DNA template was tested. Addition of streptavidin tetramer (binds biotin tightly) completely stops replication by Taq polymerase specifically at the modified position as well as one base before the modified position (**Supplementary Fig. 5**). Therefore, this method has the potential to provide single-base resolution of the location of 5-hmC in DNA loci of interest.

Next, we performed selective labeling of 5-hmC in genomic DNA from various cell lines and animal tissues. Genomic DNA from various sources was sonicated into small fragments (~100–500 base pairs), treated with β-GT in the presence of UDP-6-N_3-Glu or regular UDP-Glu (control group) to yield N_3-5-gmC or 5-gmC modifications and finally labeled with cyclooctyne-biotin (**1**) to install biotin. Because each step is efficient and bio-orthogonal, this protocol ensures selective labeling of most 5-hmC in genomic DNA. The presence of biotin-N_3-5-gmC allows affinity enrichment of this modification and accurate quantification of the amount of 5-hmC in a genome using avidin–horseradish peroxidase (HRP).

We determined the total amount of 5-hmC in mouse cerebellum at different stages of development (**Fig. 3a**,**b**). The control group showed almost no signal, demonstrating the high selectivity of this method. The amount of 5-hmC depends on the developmental stage of the mouse cerebellum (**Fig. 3b**). A gradual increase from post-natal day 7 (P7, 0.1% of total nucleotides in the genome) to adult stage (0.4% of total nucleotides) was observed[21], which was further confirmed using antibody against 5-hmC through a dot-blot assay (**Supplementary Fig. 6a**). Our observation suggests that 5-hmC might play an important role in brain development. The 5-hmC level of mouse embryonic stem cells (mESC) was determined to be comparable to results reported previously (~0.05% of total nucleotides) (**Fig. 3c**,**d**)[5]. In addition, the amount of 5-hmC in mouse adult neural stem cells (aNSC) was tested, which proved comparable to that of mESC (~0.04% of total nucleotides) (**Fig. 3c**,**d**).

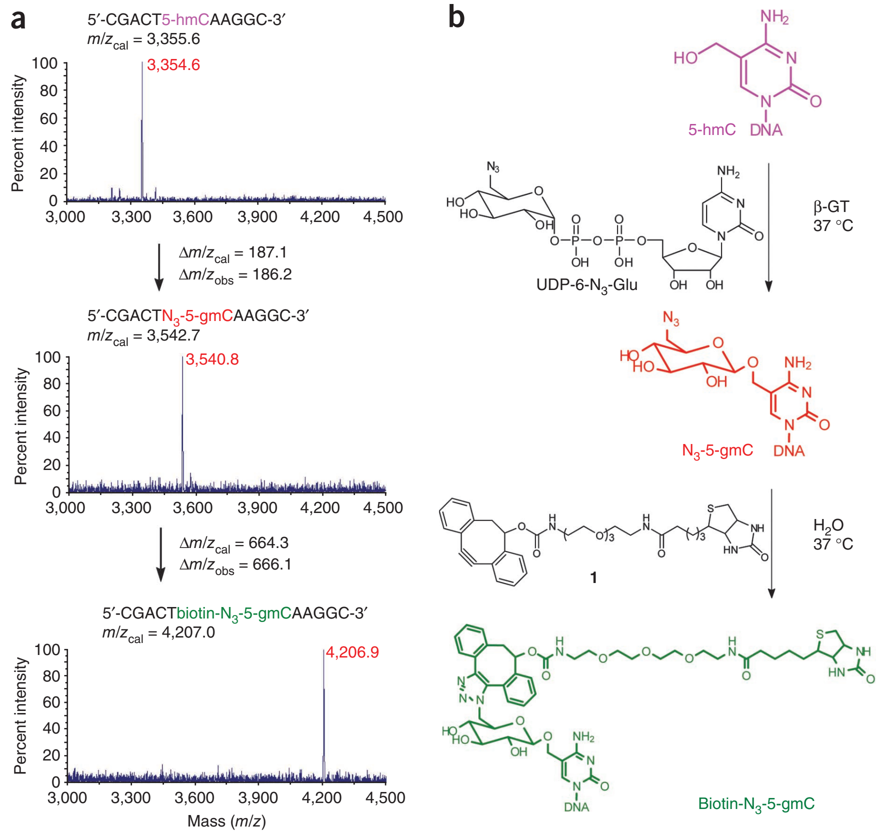


Figure 2 MS characterization of 5-hmC-, N_3-5-gmC- and biotin-N_3-5-gmC-containing 11-mer DNA in a model experiment. (**a**) MALDI-TOF of 5-hmC-, N_3-5-gmC- and biotin-N_3-5-gmC-containing 11-mer DNA, respectively, with the calculated molecular weight and observed molecular weight indicated. (**b**) Corresponding reactions of the β-GT–catalyzed formation of N_3-5-gmC and the subsequent copper-free click chemistry to yield biotin-N_3-5-gmC in duplex DNA. Reactions were performed in duplex DNA with the complementary strand; however, MS monitored the single-stranded DNA containing the modification.

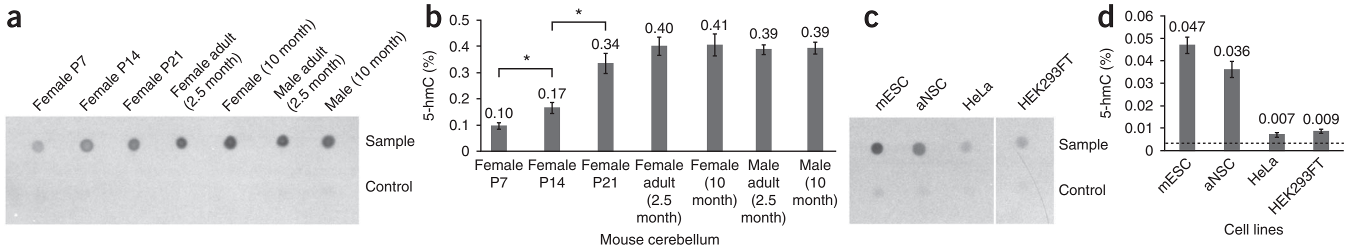


Figure 3 Quantification of 5-hmC in various cell lines and tissues. (**a**) Dot-blot assay of avidin-HRP detection and quantification of mouse cerebellum genomic DNA containing biotin-N_3-5-gmC. Top row: 40 ng of biotin-labeled samples using UDP-6-N_3-Glu. Bottom row: 40 ng of control samples using regular UDP-Glu without biotin label. The exact same procedures were followed for experiments in both rows. P7, P14 and P21 represent postnatal day 7, 14 and 21, respectively. (**b**) Amounts of 5-hmC are shown in percentage of total nucleotides of mouse genome. *, $P < 0.05$, Student's *t*-test; means ± s.e.m. for $n = 4$ experiments. (**c**) Dot-blot assay of avidin-HRP detection and quantification of genomic DNA samples from four cell lines (from same blot as in **a**), except that each dot contains 700 ng DNA. (**d**) Amounts of 5-hmC are shown in percentage of total nucleotides of the genome; means ± s.e.m. for $n = 4$ experiments. The dashed line indicates the limit of detection (~0.004%).

We also tested human cell lines (**Fig. 3c**,**d**). Notably, the presence of 5-hmC was detected in HeLa and HEK293FT cell lines, although in much lower abundance (~0.01% of total nucleotides) (**Fig. 3d**) than in other cells or tissues that have been previously reported to contain 5-hmC (previous studies did not show the presence of 5-hmC in HeLa cells due to the limited sensitivity of the methods employed[4]). These results suggest that this modification may be more widespread than previously anticipated. By contrast, no 5-hmC signal was detected in wild-type *Drosophila melanogaster*, consistent with a lack of DNA methylation in this organism[22].

To further validate the utility of the method for biological samples we confirmed the presence of 5-hmC in the genomic DNA from HeLa cells. A monomeric avidin column was used to pull down the biotin-N_3-5-gmC–containing DNA after genomic DNA labeling. These enriched DNA fragments were digested into single nucleotides, purified by HPLC and subjected to HRMS analysis. To our satisfaction, we obtained HRMS as well as MS/MS spectra of biotin-N_3-5-gmC identical to the standard from synthetic DNA (**Supplementary Fig. 4** and **Fig. 6b**,**c**). In addition, two 60-mer double-stranded (ds)DNAs, one with a single 5-hmC in its sequence and the other without the modification, were prepared. We spiked equal amounts of both samples into mouse genomic DNA and performed labeling and subsequent affinity purification of the biotinylated DNA. The pull-down sample was subjected to deep sequencing, and the result showed that the 5-hmC–containing DNA was >25-fold higher than the control sample (**Supplementary Fig. 7**).

Next, we performed chemical labeling of genomic DNA from mouse cerebellum, subjecting the enriched fragments to deep sequencing such that 5-hmC–containing genomic regions could be identified. Initially, we compared male and female adult mice (2.5 months old), sequencing multiple independent biological samples and multiple libraries prepared from the same genomic DNA. Genome-scale density profiles are nearly identical between male and female and are clearly distinguishable from both input genomic DNA and control DNA labeled with regular glucose (no biotin) (**Fig. 4a**). Peak identification revealed a total of 39,011 high-confidence regions enriched consistently with 5-hmC in both male and female (**Fig. 4a** and **Supplementary Table 1**). All of the 13 selected, enriched regions were subsequently successfully verified in both adult female and male cerebellum by quantitative PCR (qPCR), whereas multiple control regions did not display enrichment (**Supplementary Fig. 8**).

DNA methylation is widespread in mammalian genomes, with the exception of most transcription start sites (TSS)[23–25]. Previous studies have mostly assessed DNA methylation by bisulfite sequencing and methylation-sensitive restriction digests. It has since been appreciated that neither of these methods adequately distinguishes 5-mC from 5-hmC[11,12]. To determine the genome-wide distribution of 5-hmC, we generated metagene 5-hmC read density profiles for RefSeq transcripts. Normalized 5-hmC read densities differ by an average of 2.10 ± 0.04% (mean ± s.e.m.) in adult male and female cerebellum samples, indicating that the profiles are accurate and reproducible. We observed enrichment of 5-hmC in gene bodies as well as in proximal upstream and downstream regions relative to TSS, transcription termination sites (TTS) and distal regions (**Fig. 4b**). This is in contrast to previously generated methyl-binding domain–sequencing (MBD-Seq)[26], as well as our own methylated DNA immunoprecipitation sequencing (MeDIP-Seq) from mouse cerebellum genomic DNA, in which the majority (~80%) of 5-mC–enriched DNA sequences were derived from satellite and/or repeat regions (**Supplementary Fig. 9**). Further analyses also reveal that both intragenic and proximal enrichment of 5-hmC is associated with more highly expressed genes, consistent with a role for 5-hmC in maintaining and/or promoting gene expression (**Fig. 4b**). Proximal enrichment of 5-hmC ~875 bp upstream of TSSs and ~160–200 bp downstream of the annotated TTSs further suggests a role for these regions in the regulation of gene expression through 5-hmC.

Quantification of bulk 5-hmC in the cerebellum of P7 and adult mice indicates genomic acquisition of 5-hmC during cerebellum maturation (**Fig. 3a**). We further explored this phenomenon by sequencing 5-hmC–enriched DNA from P7 cerebellum and compared these sequences to those derived from adult mice. Metagene profiles at RefSeq transcripts confirmed an increase in proximal and intragenic 5-hmC in adult relative to P7 cerebellum, although there was little to no difference and minimal enrichment over input genomic DNA in distal regions (**Fig. 4c** and **Supplementary Table 2**). Peak identification using P7 as background identified a total of 20,092 enriched regions that showed significant differences between P7 and adult tissues. Of those, 15,388 (76.6%) occurred within 5,425 genes acquiring intragenic 5-hmC in adult females (**Supplementary Fig. 10** and **Supplementary Table 3**).

Gene ontology pathway analysis of the 5,425 genes acquiring 5-hmC during aging identified significant enrichment of pathways associated with age-related neurodegenerative disorders as well as angiogenesis and hypoxia response (**Fig. 4d** and **Supplementary Table 4**). This is of particular interest given that all these pathways have been linked to oxidation stress response and that the conversion of 5-mC to 5-hmC requires dioxygen[5]. Furthermore, an assessment of the gene

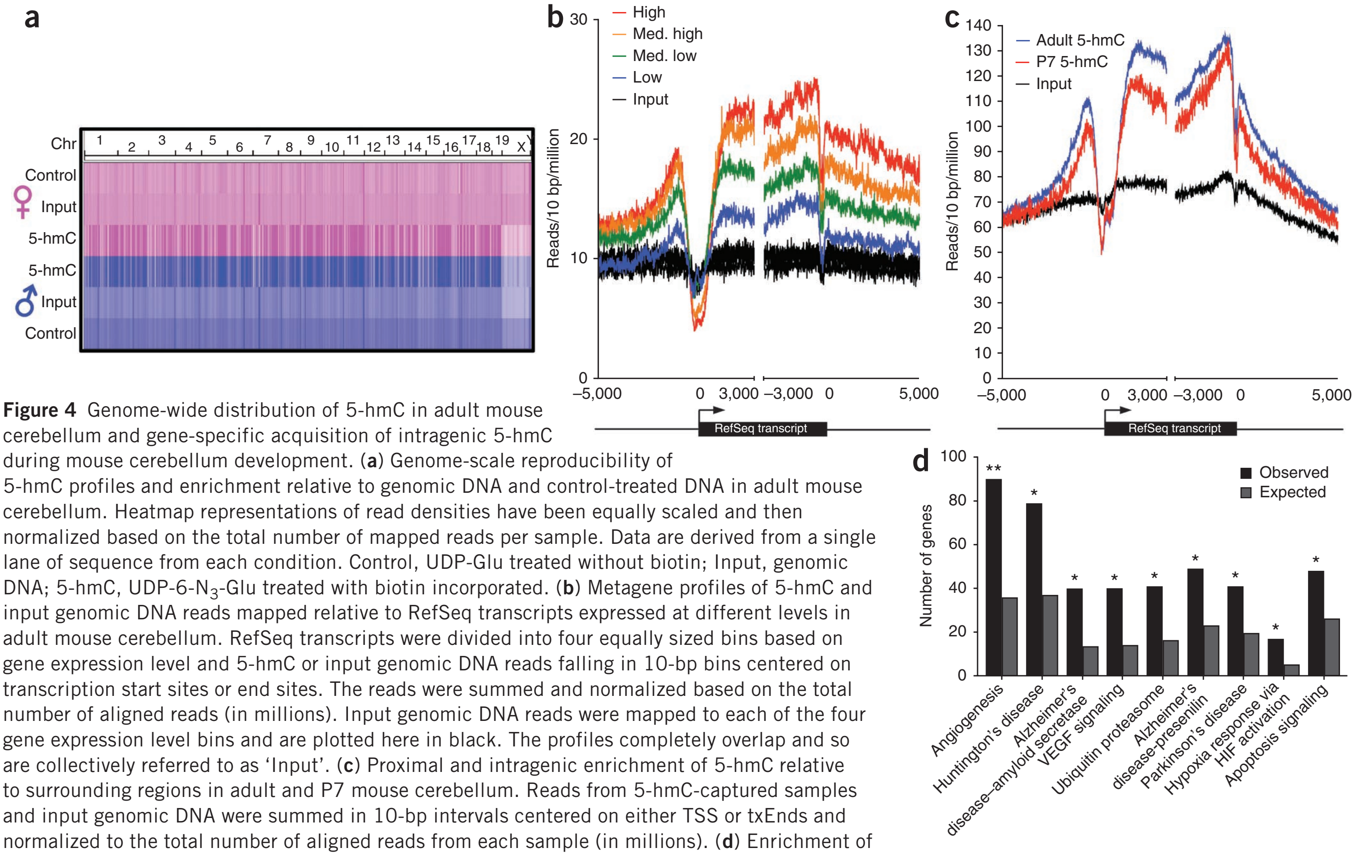


Figure 4 Genome-wide distribution of 5-hmC in adult mouse cerebellum and gene-specific acquisition of intragenic 5-hmC during mouse cerebellum development. (**a**) Genome-scale reproducibility of 5-hmC profiles and enrichment relative to genomic DNA and control-treated DNA in adult mouse cerebellum. Heatmap representations of read densities have been equally scaled and then normalized based on the total number of mapped reads per sample. Data are derived from a single lane of sequence from each condition. Control, UDP-Glu treated without biotin; Input, genomic DNA; 5-hmC, UDP-6-N_3-Glu treated with biotin incorporated. (**b**) Metagene profiles of 5-hmC and input genomic DNA reads mapped relative to RefSeq transcripts expressed at different levels in adult mouse cerebellum. RefSeq transcripts were divided into four equally sized bins based on gene expression level and 5-hmC or input genomic DNA reads falling in 10-bp bins centered on transcription start sites or end sites. The reads were summed and normalized based on the total number of aligned reads (in millions). Input genomic DNA reads were mapped to each of the four gene expression level bins and are plotted here in black. The profiles completely overlap and so are collectively referred to as 'Input'. (**c**) Proximal and intragenic enrichment of 5-hmC relative to surrounding regions in adult and P7 mouse cerebellum. Reads from 5-hmC-captured samples and input genomic DNA were summed in 10-bp intervals centered on either TSS or txEnds and normalized to the total number of aligned reads from each sample (in millions). (**d**) Enrichment of pathways associated with age-related neurodegenerative diseases in genes acquiring intragenic 5-hmC in adult mice relative to P7 mice. Shown are the number of genes that acquired 5-hmC in adult cerebellum and the number of genes expected based on the total number of genes associated with that pathway in mouse. **, $P < 10^{-10}$; *, $P < 10^{-5}$.

list revealed that 15/23 genes previously identified as causing ataxia and disorders of Purkinje cell degeneration in mouse and human acquired intragenic 5-hmC in adult mice (**Supplementary Fig. 11** and **Supplementary Table 5**)[27]. Together, these observations suggest that 5-hmC may play a role in age-related neurodegeneration.

Recently, β-GT was used to transfer a radiolabeled glucose for 5-hmC quantification[28]. (Our paper was under review when ref. 28 was published.) A major advantage of our technology is its ability to selectively label 5-hmC in genomic DNA with any tag. With a biotin tag attached to 5-hmC, DNA fragments containing 5-hmC can be affinity purified for deep sequencing to reveal distribution and/or location of 5-hmC in mammalian genomes. Because biotin is covalently linked to 5-hmC and biotin-avidin/streptavidin interaction is strong and highly specific, this technology promises high robustness as compared to potential anti-5-hmC, antibody-based, immune-purification methods[8]. Other fluorescent or affinity tags may be readily installed using the same approach for various other applications. For instance, imaging of 5-hmC in fixed cells or even live cells (if labeling can be performed in one step with a mutant enzyme) may be achieved with a fluorescent tag. In addition, the chemical labeling of 5-hmC with a bulky group could interfere with restriction enzyme digestion or ligation, which may be used to detect 5-hmC in specific genome regions. The attachment of biotin or other tags to 5-hmC also dramatically enhances the sensitivity and simplicity of the 5-hmC detection and/or quantification in various biological samples[28]. The detection limit of this method can reach ~0.004% (**Fig. 3d**) and the method can be readily applied to study a large number of biological samples.

With the technology presented here, we observed the developmental stage–dependent increase of 5-hmC in mouse cerebellum. Compared to postnatal day 7 at a time of massive cell proliferation in the mouse cerebellum, adult cerebellum has a significantly increased level of 5-hmC, suggesting that 5-hmC might be involved in neuronal development and maturation. Indeed, we also observed an increase of 5-hmC in aNSCs upon differentiation (unpublished data).

This technology enables us to selectively capture 5-hmC–enriched regions in the cerebellums from both P7 and adult mice, and determine the genome-wide distribution of 5-hmC by deep sequencing. Our analyses revealed general features of 5-hmC in mouse cerebellum. First, 5-hmC was enriched specifically in gene bodies as well as defined gene proximal regions relative to more distal regions. This differs from the distribution of 5-mC, where DNA methylation has been found both within gene bodies as well as in more distal regions[23–25,29]. Second, the enrichment of 5-hmC is higher in gene bodies that are more highly expressed, suggesting a potential role for 5-hmC in activating and/or maintaining gene expression. It is possible that conversion of 5-mC to 5-hmC is a pathway to offset the gene repression effect of 5-mC during this process without going through demethylation[30]. Third, we observed an enrichment of 5-hmC in genes linked to hypoxia and angiogenesis. The oxidation of 5-mC to 5-hmC by Tet proteins requires dioxygen[5,8]. A well-known oxygen sensor in mammalian systems that are involved in hypoxia and angiogenesis is the HIF protein, which belongs to the same mononuclear iron-containing dioxygenase superfamily as the active domain of the Tet proteins[7]. It is tempting to speculate that oxidation of 5-mC to 5-hmC by Tet proteins may constitute another oxygen-sensing and regulation pathway in mammalian cells. Lastly, the association of 5-hmC with genes that have been implicated in neurodegenerative disorders suggests that this base modification could potentially contribute to the pathogenesis of human neurological disorders. Should a connection between 5-hmC levels and human disease be established, the affinity purification approach shown in the current work could be used to purify and/or enrich 5-hmC–containing DNA fragments as a simple and sensitive method for disease prognosis and diagnosis.

In summary, we have developed an efficient and selective method to label and capture 5-hmC from genomic DNA. We have demonstrated the feasibility of using this approach to determine the genome-wide distribution of 5-hmC. Future application of this technology would enable us to understand the role(s) of the 5-hmC modification at molecular, cellular and physiological levels.

METHODS

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

Accession codes. The sequencing data have been deposited in NCBI's Gene Expression Omnibus with accession number GSE25398.

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

ACKNOWLEDGMENTS

We would like to thank S. Warren for the helpful discussion and critical reading of the manuscript. This study was supported partly by the US National Institutes of Health (GM071440 to C.H. and NS051630/MH076090/MH078972 to P.J.) and the University of Chicago.

AUTHOR CONTRIBUTIONS

C.H., C.-X.S. and P.J. designed the experiments with help from Y.F. and B.T.L. Experiments were performed by C.-X.S., K.E.S., Y.F., C.Y. and Q.D. with the help of W.Z. and X.J.; Q.D. and J.W. carried out the chemical synthesis; K.E.S., X.L., Y.L. and P.J. provided the mouse cerebellum, mouse aNSC and fly samples, and performed deep sequencing; C.-H.C., L.Z., T.J.L. and L.A.G. helped with the mouse ESC, human HeLa, human HEK and related samples; B.Z. and L.M.H. performed the mass spectrometry analysis from HeLa cells. C.H., C.-X.S. and P.J. wrote the paper. All authors discussed the results and commented on 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. Klose, R.J. & Bird, A.P. Genomic DNA methylation: the mark and its mediators. *Trends Biochem. Sci.* **31**, 89–97 (2006).
2. Reik, W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* **447**, 425–432 (2007).
3. Gal-Yam, E.N., Saito, Y., Egger, G. & Jones, P.A. Cancer epigenetics: modifications, screening, and therapy. *Annu. Rev. Med.* **59**, 267–280 (2008).
4. Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324**, 929–930 (2009).
5. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930–935 (2009).
6. Yi, C., Yang, C.G. & He, C. A non-heme iron-mediated chemical demethylation in DNA and RNA. *Acc. Chem. Res.* **42**, 519–529 (2009).
7. Hausinger, R.P. FeII/alpha-ketoglutarate-dependent hydroxylases and related enzymes. *Crit. Rev. Biochem. Mol. Biol.* **39**, 21–68 (2004).
8. Ito, S. *et al.* Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466**, 1129–1133 (2010).
9. Ko, M. *et al.* Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature* 10.1038/nature09586 (7 Nov 2010).
10. Loenarz, C. & Schofield, C.J. Oxygenase catalyzed 5-methylcytosine hydroxylation. *Chem. Biol.* **16**, 580–583 (2009).
11. Huang, Y. *et al.* The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS ONE* **5**, e8888 (2010).
12. Jin, S.G., Kadam, S. & Pfeifer, G.P. Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. *Nucleic Acids Res.* **38**, e125 (2010).
13. Flusberg, B.A. *et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* **7**, 461–465 (2010).
14. Josse, J. & Kornberg, A. Glucosylation of deoxyribonucleic acid. III. alpha and beta-glucosyl transferases from T4-infected *Escherichia coli*. *J. Biol. Chem.* **237**, 1968–1976 (1962).
15. Lariviere, L. & Morera, S. Structural evidence of a passive base-flipping mechanism for beta-glucosyltransferase. *J. Biol. Chem.* **279**, 34715–34720 (2004).
16. Kolb, H.C., Finn, M.G. & Sharpless, K.B. Click chemistry: diverse chemical function from a few good reactions. *Angew. Chem. Int. Ed.* **40**, 2004–2021 (2001).
17. Speers, A.E. & Cravatt, B.F. Profiling enzyme activities in vivo using click chemistry methods. *Chem. Biol.* **11**, 535–546 (2004).
18. Sletten, E.M. & Bertozzi, C.R. Bioorthogonal chemistry: fishing for selectivity in a sea of functionality. *Angew. Chem. Int. Ed.* **48**, 6974–6998 (2009).
19. Baskin, J.M. *et al.* Copper-free click chemistry for dynamic in vivo imaging. *Proc. Natl. Acad. Sci. USA* **104**, 16793–16797 (2007).
20. Ning, X., Guo, J., Wolfert, M.A. & Boons, G.J. Visualizing metabolically labeled glycoconjugates of living cells by copper-free and fast huisgen cycloadditions. *Angew. Chem. Int. Ed.* **47**, 2253–2255 (2008).
21. Munzel, M. *et al.* Quantification of the sixth DNA base hydroxymethylcytosine in the brain. *Angew. Chem. Int. Ed.* **49**, 5375–5377 (2010).
22. Lyko, F., Ramsahoye, B.H. & Jaenisch, R. DNA methylation in *Drosophila melanogaster*. *Nature* **408**, 538–540 (2000).
23. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766–770 (2008).
24. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).
25. Edwards, J.R. *et al.* Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. *Genome Res.* **20**, 972–980 (2010).
26. Skene, P.J. *et al.* Neuronal MeCP2 is expressed at near histone-octamer levels and globally alters the chromatin state. *Mol. Cell* **37**, 457–468 (2010).
27. Lim, J. *et al.* A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell* **125**, 801–814 (2006).
28. Szwagierczak, A., Bultmann, S., Schmidt, C.S., Spada, F. & Leonhardt, H. Sensitive enzymatic quantification of 5-hydroxymethylcytosine in genomic DNA. *Nucleic Acids Res.* **38**, e181 (2010).
29. Maunakea, A.K. *et al.* Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466**, 253–257 (2010).
30. Wu, S.C. & Zhang, Y. Active DNA demethylation: many roads lead to Rome. *Nat. Rev. Mol. Cell Biol.* **11**, 607–620 (2010).

ONLINE METHODS

Preparation of genomic DNA. All animal procedures were performed according to protocols approved by Emory University Institutional Animal Care and Use Committee. Genomic DNA from tissues and cell lines was purified using Wizard genomic DNA purification kit (Promega) with additional Proteinase K treatment and rehydrated in 10 mM Tris (pH 7.9). Genomic DNA samples were further sonicated in Eppendorf tubes into 100–500 bp by Misonix sonicator 3000 (using microtip, three pulses of 30 s each with 2 min of rest and a power output level of 2) or Bioruptor UCD-200 sonicator (Diagenode, Sparta). (The output selector switch was set on High (H), and sonication interval was 30 s with 30 cycles of sonication performed. In addition, samples were resuspended and centrifuged briefly every five cycles to keep the constancy of DNA shearing.) Cerebellums from P7 and 10-week-old C57BL/6 were used. Mouse feeder-free E14Tg2A ES cells (mESC) were cultured as reported[31]. Adult neural stem cells (aNSCs) were isolated and cultured as described previously[32].

Construction, expression and purification of wild-type β-GT. β-GT was cloned from the extract of T4 bacteriophage (American Type Culture Collection) into the target vector pMCSG19 by the ligation independent cloning method[33]. The resulting plasmid was transformed into BL21 star (DE3)-competent cells containing pRK1037 (Science Reagents) by heat shock. Positive colonies were selected with 150 g/ml ampicillin and 30 g/ml kanamycin. One liter of cells was grown at 37 °C from a 1:100 dilution of an overnight culture. The cells were induced with 1 mM of isopropyl-β-D-thiogalactoside when OD_{600} reached 0.6–0.8. After overnight growth at 16 °C with shaking, the cells were collected by centrifugation, suspended in 30 ml Ni-NTA buffer A (20 mM Tris-HCl, pH 7.5, 150 mM NaCl, 30 mM imidazole and 10 mM β-mercaptoethanol) with protease inhibitor phenylmethylsulfonyl fluoride. After loading to a Ni-NTA column, proteins were eluted with a 0–100% gradient of Ni-NTA buffer B (20 mM Tris-HCl, pH 7.5, 150 mM NaCl, 400 mM imidazole and 10 mM β-mercaptoethanol). β-GT-containing fractions were further purified by MonoS (GE Healthcare) (buffer A: 10 mM Tris-HCl, pH 7.5; buffer B: 10 mM Tris-HCl, pH 7.5 and 1 M NaCl). Finally, the collected protein fractions were loaded onto a Superdex 200 (GE Healthcare) gel-filtration column equilibrated with 50 mM Tris-HCl (pH 7.5), 20 mM $MgCl_2$ and 10 mM β-ME. The purity of the purified protein was determined by SDS-PAGE to be >95%. β-GT was concentrated to 45 μM and stored frozen at −80 °C with an addition of 30% glycerol.

Oligonucleotide synthesis. Oligonucleotides containing 5-hmC were prepared using Applied Biosystems 392 DNA synthesizer. 5-Hydroxymethyl-dC-CE phosphoramidite (Glen Research) was used to incorporate 5-hmC at the desired position during solid-phase synthesis, followed by postsynthetic deprotection by treatment with 30% ammonium hydroxide first and then 25–30% wt/wt solution of sodium methoxide in methanol (Alfa Aesar) overnight at 25 °C. The 11-mer DNA was purified by reversed-phase HPLC and confirmed by MALDI-TOF. Other DNA was purified by denaturing PAGE. Concentrations of the oligonucleotides were estimated by UV at 260 nm. Duplexes were prepared by combining equimolar portions of the each strand in annealing buffer (10 mM Tris, pH 7.5, 100 mM NaCl), heating for 10 min at 95 °C followed by slow cooling overnight.

5-hmC labeling reaction and click chemistry. The 5-hmC labeling reactions were performed in a 100-μl solution containing 50 mM HEPES buffer (pH 7.9), 25 mM $MgCl_2$, 300 ng/μl sonicated genomic DNA (100-500 bp), 250 μM UDP-6-N_3-Glu, and 2.25 μM wild-type βGT. The reactions were incubated for 1 h at 37 °C. After the reaction, the DNA substrates were purified by Qiagen DNA purification kit or by phenol-chloroform precipitation and reconstituted in H_2O. The click chemistry was performed with addition of 150 μM dibenzocyclooctyne modified biotin (compound **1**) into the DNA solution, and the reaction mixture was incubated for 2 h at 37 °C. The DNA samples were then purified by Qiagen DNA purification kit, which were ready for further applications.

Affinity enrichment of the biotinylated 5-hmC (biotin-N_3-5-gmC). Genomic DNAs used for deep sequencing were purified/enriched by Pierce Monomeric Avidin Kit (Thermo) twice following manufacturer's recommendations. After elution, the biotin-N_3-5-gmC containing DNA was concentrated by 10 K Amicon Ultra-0.5 ml Centrifugal Filters (Millipore) and purified by Qiagen DNA purification kit. Starting with 30 μg total genomic DNA, we can obtain 100-300 ng enriched DNA samples following the labeling and pull-down protocol described here. The deep sequencing experiment can be performed with as low as 10 ng DNA sample.

Primer extension assay. Reverse primer (14-mer, 5′-AAGCTTCTGGAGTG-3′, purchased from Eurofins MWG Operon and PAGE purified) was end-labeled with T4 polynucleotide kinase (T4 PNK) (New England Biolabs) and 15 μCi of [γ-^{32}P]-ATP (PerkinElmer) for 0.5 h at 37 °C, and then purified by Bio-Spin 6 column (Bio-Rad). For primer extension assay, REDTaq DNA polymerase (Sigma) was used. We first mixed 0.2 pmol template and 0.25 pmol γ-^{32}P-labeled primers with dNTP in the polymerase reaction buffer without adding polymerase. The mixture was heated at 65 °C for 2 min and allowed to cool slowly for 30 min. Streptavdin in PBS was then added if needed and allowed to mix at 25 °C for 5 min. REDTaq DNA polymerase was then added (final volumn 20 μl) and the extension reaction was run at 72 °C for 1 min. The reaction was quenched by 2× stop solution (98% formamide, 10 mM EDTA, 0.1% xylene cyanol, 0.1% bromophenol blue) and loaded on to a 20% denaturing polyacrylamide gel (7 M urea). Sanger sequencing was performed using Sequenase DNA Sequencing Kit (USB) with 1 pmol template and 0.5 pmol [γ-^{32}P]-labeled primer. The results were visualized by autoradiography.

Large-scale HeLa 5-hmC pull-down. Twenty dishes (15 cm) of HeLa cells were harvested and resuspended at 20 ml of 10 mM Tris (pH 8.0), 10 mM EDTA. Sodium dodecyl sulfate (SDS) and Proteinase K were added to final concentrations of 0.5% and 200 μg/ml, respectively, and the solution was allowed to incubate at 55 °C for 2 h. After adding NaCl to a final concentration of 0.2 M, the sample was extracted twice with equal volumes of phenol/chloroform/isoamyl alcohol (25:24:1) and once with chloroform. Chloroform was evaporated by placing the tube in 55 °C water bath for 1 h with cap open. RNase A was then added to a final concentration of 25 μg/ml and the solution incubated for 1 h at 37 °C. DNA was then extracted once with phenol/ chloroform/isoamyl alcohol (25:24:1) and once with chloroform and precipitated with 1.5 volumes of ethanol. Genomic DNA was washed twice with 20 ml of 70% ethanol, dried and resuspended in 10 mM Tris (pH 7.9) at 37 °C. Genomic DNA was then sonicated by Bioruptor UCD-200 sonicator into 100–1,000 bp as noted before. The 5-hmC labeling reaction was carried out in a 4 ml solution containing 50 mM HEPES buffer (pH 7.9), 25 mM $MgCl_2$, 550 ng/μl sonicated HeLa genomic DNA, 250 μM UDP-6-N_3-Glu and 2.25 μM wild-type β-GT. The reaction was incubated for 1 h at 37 °C, purified by phenol-chloroform precipitation and reconstituted in 4 ml H_2O. We added 20 μl of 30 mM dibenzocyclooctyne-modified biotin (compound **1**) and incubated the mixture for 2 h at 37 °C. The DNA sample was purified again by phenol-chloroform precipitation and then enriched for biotin-N_3-5-gmC by monomeric avidin column as noted before. The pull-down DNA was concentrated and digested by nuclease P1 (Sigma), venom phosphodiesterase I (Type VI) (Sigma) and alkaline phosphatase (Sigma) according to published protocols[34]. The sample was purified by HPLC C18 reversed-phase column as noted in **Supplementary Figure 3**. The peaks corresponding to the biotin-N_3-5-gmC from synthetic DNA were collected, lyophilized and subjected to HRMS analysis. For HRMS analysis, lyophilized fractions were dissolved in 100 μl of 50% methanol and 5–20 μl samples were injected for LC-MS/MS analysis. The LC-MS/MS system is composed of an Agilent 1200 HPLC system and an Agilent 6520 QTOF system controlled by MassHunter Workstation Acquisition software (B.02.01 Build 2116). A reversed-phase C18 column (Kinetex C18, 50 mm × 2.1 mm, 1.7 μm, with 0.2 μm guard cartridge) flowing at 0.4 ml min^{-1} was used for online separation to avoid potential ion suppression. The gradient was from 98% solvent A (0.05% (vol/vol) acetic acid in MilliQ water), held for 0.5 min, to 100% solvent B (90% acetonitrile (vol/vol) with 0.05% acetic acid (vol/vol) in 4 min. MS and MS/MS data were acquired in extended dynamic range (1,700 *m*/*z*) mode, with post-column addition of reference mass solution for real time mass calibration.

Dot-blot assays and quantification of genomic DNA containing 5-hmC. Labeled genomic DNA samples (biotin-N_3-5-gmC, 40 ng for mouse cerebellum samples, 700 ng for other samples) were spotted on an Amersham Hybond-N$^+$ membrane (GE Healthcare). DNA was fixed to the membrane by Stratagene UV Stratalinker 2400 (auto-crosslink). The membrane was then blocked with 5% BSA and incubated with avidin-HRP (1:20,000) (Bio-Rad), which was visualized by enhanced chemiluminescence. Quantification was calculated using a working curve generated by 1–8 ng of 32 bp synthetic biotin-N$_3$-5-gmC–containing DNA. Polyclonal antibody against 5-hmC (Active Motif) was also used for dot-blot assay (1:10,000 dilution).

5-hmC-enrichment test. Two solutions of 60-mer dsDNA (see **Supplementary Fig. 7**) were prepared as noted. Mouse DNA (30 μg) was spiked with 3 pg from each DNA solution. We did 5-hmC labeling and enrichment as noted. The pull-down DNA (10 ng) was end-repaired, adenylated, ligated to adapters (size selection 140–400 bp) and sequenced on an Illumina Genome Analyzer according to the manufacturer's recommendations for Illumina ChIP-Seq to identify spike enrichment.

Reads were mapped to the *Mus musculus* reference genome (NCBI37/mm9), excluding sequences that were not finished or that have not be placed with certainty (i.e., exclusion of sequences contained in the chrUn_random.fa and chrN_radom.fa files provided by the UCSC genome browser) and appended to contain fasta sequences corresponding to the positive and negative spiked controls. Sequence alignment was accomplished using bwa[35] and default alignment settings.

Deep sequencing of mouse cerebellum genomic DNA. DNA libraries were generated following the Illumina protocol for "Preparing Samples for ChIP Sequencing of DNA" (Part# 111257047 Rev. A). 25 ng genomic DNA, 5-hmC-captured DNA, or control captured DNA (in the absence of biotin) were used to initiate the protocol. In some instances < 25 ng DNA was eluted in the no-biotin control treatment. In these cases the entire amount of eluted DNA was used to initiate library preparation. DNA fragments ~150–300 bp were gel purified after the adaptor ligation step. PCR amplified DNA libraries were quantified on an Agilent 2100 Bioanalyzer and diluted to 6 pM for cluster generation and sequencing. 38-cycle single end sequencing was performed using Version 4 Cluster Generation and Sequencing Kits (Part #15002739 and #15005236 respectively) and Version 7.0 recipes. Image processing and sequence extraction were done using the standard Illumina Pipeline.

Sequence alignment and peak identification. FASTQ sequence files were aligned to *Mus musculus* reference genome (NCBI37/ mm9) using Bowtie[36] (**Supplementary Fig. 12**). The--best alignment and reporting option was used for all conditions, corresponding to no more than 2 bp mismatches across each 38 bp read. 5-hmC peak identification was performed using nonduplicate reads with MACS[37]. Parameters were as follows: effective genome size = 2.72e+09; tag size = 38; band width = 100; model fold = 10; *P* value cutoff = 1.00e-05; ranges for calculating regional lambda are: peak_region, 1,000, 5,000, 10,000.

For identification of high-confidence peaks consistently detected in adult female and male samples, data from all lanes were merged per condition (5-hmC enriched, nonenriched genomic DNA input) for each sex and used in the analysis described above. Using a combined input genomic DNA sequence set (male input plus female input) as background, we observed 78.7% overlap in identified peaks between male and female samples. As a more stringent analysis, we also used sex-matched input genomic DNA as background/control samples for peak identification. A total of 91,751 peaks were identified in adult female cerebellum and a total of 240,147 peaks were identified in adult male cerebellum using these parameters; 39,011 peaks overlapped ≥1 bp between sexes and are reported as the set of high-confidence peaks consistently detected adult cerebellum (**Supplementary Table 1**). Regions enriched for 5-hmC in adult cerebellum relative to P7 cerebellum were identified using a single lane of adult female 5-hmC reads as the treatment and the single lane of P7 reads as the background and/or control sample (**Supplementary Table 2**). A total of 20,092 regions were identified as enriched for 5-hmC in adult female cerebellum relative to P7 cerebellum. Of these, 15,388 (76.6%) were intragenic to 5,425 unique RefSeq transcripts. Genes acquiring 5-hmC during development (**Supplementary Table 3**) are those with peaks overlapping ≥1bp of a RefSeq gene.

Generation of metagene profiles and heatmaps. Metagene RefSeq transcript profiles were generated by first determining the distance between any given read and the closest TSS or TTS and then summing the number of 5′ends within 10 bp bins centered on either TSS or txEnds. Ten bp bins were then examined 5 kb upstream and 3 kb downstream to assess the level of 5-hmC in gene bodies relative to TSS and txEnds. The RefSeq reference file was obtained through the UCSC Genome Browser Tables (downloaded 05/20/2010).

Read densities (Reads/10bp) were calculated for each individual lane of sequence listed in **Supplementary Table 1** and then normalized per million reads of aligned sequence to generate a normalized read density. For samples sequenced on multiple lanes, normalized read densities were averaged. To generate the metagene profile for adult cerebellum we averaged normalized read densities from male and female. We observed excellent consistency in normalized read densities between both technical replicates (independent library preparation and sequencing the same library on multiple lanes) as well as between biological replicates (male and female adult samples). For genomic DNA input libraries from male and female samples normalized read densities differed by 3.41 ± 0.05% (mean ± s.e.m.). For 5-hmC libraries from male and female samples normalized read densities differed by 2.10 ± 0.04% (mean ± s.e.m.).

To assess 5-hmC in genes expressed at different levels, we obtained adult cerebellum gene expression data from the NCBI GEO sample GSM82974. Signal intensities were downloaded directly, divided into four bins of equal size, and converted into RefSeq mRNA IDs. We then mapped 5-hmC reads to the TSS and txEnds as described above. Heatmap representations of sequence densities were generated using Integrated Genomics Viewer tools and browser (IGV 1.4.2, http://www.broadinstitute.org/igv) with a window size (-w) of 25 and a read extend (-e) of 200.

MeDIP-Seq, MBD-Seq data and analysis. MBD-Seq data were downloaded from NCBI GEO number GSE19786, data sets SRR037089 and SRR037090 (ref. 26). Methyl cytosine containing DNA was immunoprecipitated as previously described[32] using 4 μg sonicated genomic DNA from adult female mouse cerebellum. We used 25 ng immunoprecipitated DNA to generate libraries for sequencing as described above.

MeDIP-Seq and MBD-Seq reads were aligned to the NCBI37, mm9 using identical parameters as that used for 5-hmC reads. Using these parameters SRR037089 provided 15,351,672 aligned reads, SRR037090 provided 15,586,459 aligned reads and MeDIP-Seq provided 14,104,172 aligned reads. Reads were identified as either RepeatMasker (Rmsk, NCBI37, mm9) or RefSeq (based on 05/20/10 UCSC download) if overlapping ≥1 bp of a particular annotation. The fraction of total reads corresponding to each was then determined. The expected fraction of reads based on the fraction of genomic sequence corresponding to either Rmsk or RefSeq was also plotted for comparison.

qPCR validation of 5-hmC–enriched regions. Input genomic DNA and 5-hmC enriched DNA were diluted to 1 ng/μl and 1 μl was used in triplicate 20 μl qPCR reactions each with 1× PowerSYBR Green PCR Master Mix (ABI), 0.5 μM forward and reverse primers, and water. Reactions were run on an SDS 7500 Fast Instrument using the standard cycling conditions. Primers were as follows, including the gene with which the identified peak associated the genomic location. Fold-enrichment was calculated as 2^-dCt, where dCt = Ct (5-hmC enriched) – Ct (Input). Chr3: 65106415-65106915_Kcnab1: Forward (AAGCTATGCCCGTGTCACTCA), Reverse (TGCATCAAGCGACACACAGA); Chr15: 27460605-27461105_Ank: Forward (ATCGGCAGAAGGTAGGAGGAA), Reverse (CCTCACTTGTCTCCCTGCTTATC); Chr8: 24136542-24137042_Ank1: Forward (GAGACCCTCTTGGGACAGTTACC), Reverse (TGGGTTACATTCCTCACTCGAA); Chr19: 16420423-16420923_Gnaq: Forward (ATGAGTGAACCATCCCATGCA), Reverse (TCAGCCAGTGCCTCGTGAT); Chr1: 36417273-36417773_4632411B12Rik: Forward (TGCAACAAGTGCCTGACATACA), Reverse (TTGTGTGTGCAATCATTGTTCATT); Chr11: 53835569-53836069_Slc22a4: Forward (CCTCCAGTCCAGGCAGTGAT), Reverse (CGTCAAAGGAGTCCTGGTCAA); Chr15: 99352255-99352755_Faim2: Forward (CCTCCTTAGGGCCATTCTCAA), Reverse (CGGACCTGATGGGCATAGTAG); Chr16: 7197547-7198047_A2bp1: Forward (TCTACTCCCGTTTCACCGTTTATAT), Reverse (GCCCATGCAGCCAGTTG); Chr17: 12879263-12879763_Igf2r:

doi:10.1038/nbt.1732

Forward (AGAGGGACATGGGCATCACA), Reverse (ACCGCTGACTGCCAGTACCT); Chr17: 32919340-32919840_Zfp871: Forward (GACCCAGGAGAGAAAGCATGAG), Reverse (TGACTCCGTGAACAGGAATGG); Chr2: 25147087-25147587_Grin1: Forward (AGAGAGATAGAGGTGGAAGTCAGGTT), Reverse (AGGAGCCTGGAGCAGAAATG); Chr5: 117916917-117917417_Ksr2: Forward (GAACAGTGTAAGGTCCACCCAAGT); Reverse (GGAAAAACGGGTTCGGAAAG); Chr7: 88013448-88013948_Zscan2: Forward (TGGCACACTTGAGCAAATCCTA); Reverse (TGCCAACTATTGGAATGGAAAATA); Control primers: Chr17: 31829767-31830267_Control1: Forward (GAACAGCCAGCAACCTTCTAAAA), Reverse (CAACAGCGTCATGGGATAACA); Chr12: 98299598-98300098_Control2: Forward (ACAACCCGCCCACCAAT), Reverse (TTTAGCTACCCCCAAGTTTAATGG).

GO pathway analysis. Peaks enriched for 5-hmC in adult female relative to P7 were overlapped with RefSeq annotations and those overlapping ≥1 bp were retained. A unique set of genes with ≥1 enriched 5-hmC region was then generated and used as input for the binomial gene list comparison tool provided by the Protein Analysis Through Evolutionary Relationships (PANTHER) classification system[38,39].

Chemical synthesis. Compound **1** was prepared according to previous literatures[20,40]. UDP-6-N_3-UDP was chemically synthesized as detailed in **Supplementary Methods**.

Statistical methods. We used unpaired two-tailed Student's *t*-tests (assuming equal variance) to determine significance and calculate *P*-values between mouse samples of different age. A minimum of three data points was used for each analysis.

31. Silva, J. *et al.* Promotion of reprogramming to ground state pluripotency by signal inhibition. *PLoS Biol.* **6**, e253 (2008).
32. Szulwach, K.E. *et al.* Cross talk between microRNA and epigenetic regulation in adult neurogenesis. *J. Cell Biol.* **189**, 127–141 (2010).
33. Donnelly, M.I. *et al.* An expression vector tailored for large-scale, high-throughput purification of recombinant proteins. *Protein Expr. Purif.* **47**, 446–454 (2006).
34. Crain, P.F. Preparation and enzymatic hydrolysis of DNA and RNA for mass spectrometry. *Methods Enzymol.* **193**, 782–790 (1990).
35. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
36. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
37. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
38. Thomas, P.D. *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* **13**, 2129–2141 (2003).
39. Thomas, P.D. *et al.* Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucleic Acids Res.* **34**, W645–650 (2006).
40. Jung, M.E. & Miller, S.J. Total synthesis of isopavine and intermediates for the preparation of substituted amitriptyline analogs—facile routes to substituted dibenzocyclooctatrienes and dibenzocycloheptatrienes. *J. Am. Chem. Soc.* **103**, 1984–1992 (1981).

doi:10.1038/nbt.1732

nature biotechnology

Genomic safe harbors permit high β-globin transgene expression in thalassemia induced pluripotent stem cells

Eirini P Papapetrou[1,2], Gabsang Lee[3], Nirav Malani[4], Manu Setty[5], Isabelle Riviere[1,2,6], Laxmi M S Tirunagari[1,2], Kyuichi Kadota[1,7], Shoshannah L Roth[4], Patricia Giardina[8], Agnes Viale[9], Christina Leslie[5], Frederic D Bushman[4], Lorenz Studer[1,3] & Michel Sadelain[1,2]

Realizing the therapeutic potential of human induced pluripotent stem (iPS) cells will require robust, precise and safe strategies for genetic modification, as cell therapies that rely on randomly integrated transgenes pose oncogenic risks. Here we describe a strategy to genetically modify human iPS cells at 'safe harbor' sites in the genome, which fulfill five criteria based on their position relative to contiguous coding genes, microRNAs and ultraconserved regions. We demonstrate that ~10% of integrations of a lentivirally encoded β-globin transgene in β-thalassemia-patient iPS cell clones meet our safe harbor criteria and permit high-level β-globin expression upon erythroid differentiation without perturbation of neighboring gene expression. This approach, combining bioinformatics and functional analyses, should be broadly applicable to introducing therapeutic or suicide genes into patient-specific iPS cells for use in cell therapy.

The advent of induced pluripotent stem (iPS) cells enables for the first time the derivation of unlimited numbers of patient-specific stem cells[1–3] and holds great promise for regenerative medicine[4,5]. Recent studies have explored the potential of iPS cell generation combined with gene and cell therapy for disease treatment in mice and humans[4,5]. However, for the promise of iPS cell technology in therapeutic applications to be fully realized, clinically translatable methodologies for the introduction of therapeutic, suicide, drug resistance or reporter genes into human iPS cells will be needed. The foreign genetic material should ideally be delivered into 'safe harbors', that is, regions of the genome where the integrated material is adequately expressed without perturbing endogenous gene structure or function, following a process that is amenable to precise mapping and minimizing occult genotoxicity. Retroviruses, such as HIV, efficiently integrate in the human genome with a strong bias toward actively transcribed genes[6]. This semi-random integration pattern favors expression of retrovirally encoded transgenes but entails a risk of perturbing the expression of neighboring genes, including cancer-related genes[7–10]. We hypothesized that screening iPS cell clones harboring a single vector copy would enable us to retrieve safe harbor sites that met the following five criteria: (i) distance of at least 50 kb from the 5′ end of any gene, (ii) distance of at least 300 kb from any cancer-related gene, (iii) distance of at least 300 kb from any microRNA (miRNA), (iv) location outside a transcription unit and (v) location outside ultraconserved regions (UCRs) of the human genome[11]. As the most common insertional oncogenesis event is transactivation of neighboring tumor-promoting genes[7,12], the first two criteria exclude the portion of the human genome located near promoters of genes, in particular, cancer-related genes (**Supplementary Table 1**). The latter were defined as genes functionally implicated in human cancers or the human homologs of genes implicated in cancer in model organisms (available at http://microb230.med.upenn.edu/protocols/cancergenes.html). Proximity to miRNA genes was adopted as an exclusion criterion because miRNAs are implicated in the regulation of many cellular processes, including cell proliferation and differentiation. As vector integration within a transcription unit can disrupt gene function through the loss of function of a tumor suppressor gene or the generation of an aberrantly spliced gene product[10], our fourth criterion excludes all sites located inside transcribed genes. Finally, we excluded UCRs—regions that are highly conserved over multiple vertebrates and known to be enriched for enhancers and exons[11].

We investigated this approach in an iPS cell model for the genetic correction of β-thalassemia major using a well-characterized globin lentiviral vector[13,14] (**Fig. 1a**). We generated a total of 20 iPS cell lines from skin fibroblasts or bone marrow mesenchymal stem cells (MSCs) (**Fig. 1b**) from four individuals with β-thalassemia major of various genotypes (**Supplementary Table 2**). All putative thalassemia iPS cell lines (referred to as thal-iPS) exhibited characteristic human embryonic stem (hES) cell morphology (**Fig. 1c** and **Supplementary Fig. 1**). Seven putative thal-iPS cell lines (**Supplementary Table 2**) were selected for further characterization. They expressed human pluripotent cell markers (Tra-1-81, Tra-1-60, SSEA-3, SSEA-4 and Nanog)

[1]Center for Cell Engineering, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [2]Molecular Pharmacology and Chemistry Program, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [3]Developmental Biology Program, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [4]Department of Microbiology, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [5]Computational Biology Program, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [6]Cell Therapy and Cell Engineering Facility, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [7]Thoracic Service, Department of Surgery, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [8]Thalassemia Program, Pediatric Hematology/Oncology Division, Weill Cornell Medical College, New York, New York, USA. [9]Genomics Core Facility, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. Correspondence should be addressed to M.S. (m-sadelain@ski.mskcc.org).

Received 24 September; accepted 26 October; published online 12 December 2010; doi:10.1038/nbt.1717

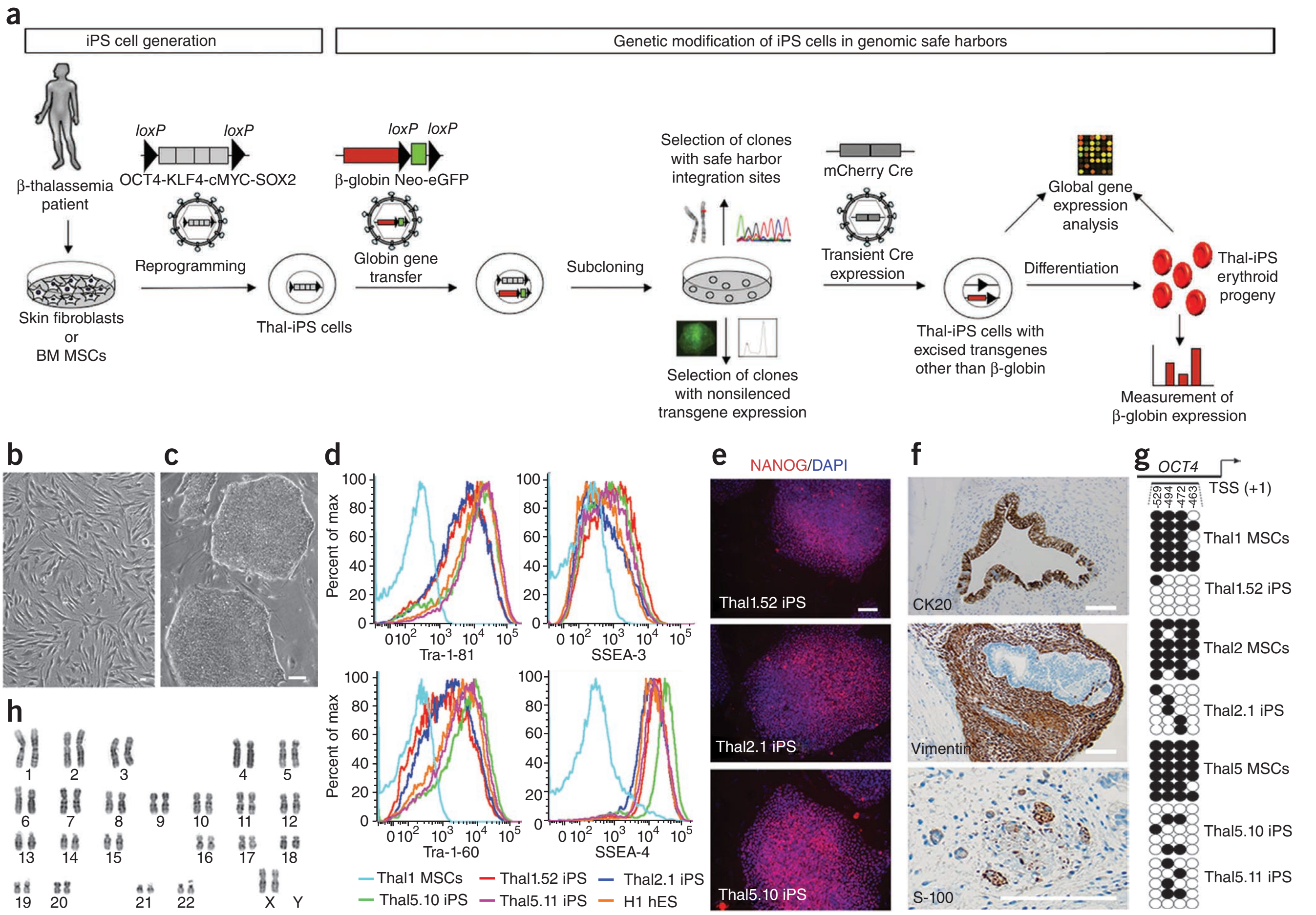


Figure 1 Safe harbor selection strategy and characterization of thal-iPS cell lines. (**a**) Following the establishment of patient-specific iPS cell lines, which in this study were generated from skin fibroblasts or bone marrow mesenchymal stem cells (BM MSCs) from β-thalassemia major patients, with an excisable single polycistronic vector co-expressing OCT4, KLF4, cMYC and SOX2 (illustrated in **Supplementary Fig. 10a**), the genetic rescue strategy is as follows: thal-iPS cells are transduced with a lentiviral vector expressing β-globin and an excisable Neo-eGFP selection cassette and subcloned into single cells, and single-vector-copy integrants are selected according to vector chromosomal position. The levels of β-globin expression afforded by the vector integrated at different genomic positions are analyzed in the erythroid progeny of each selected clone. Microarray analysis is used to examine perturbation of endogenous gene expression by the integrated provirus. The reprogramming vector can be efficiently excised before globin gene transfer or after—together with the Neo-eGFP selection cassette of the globin vector—through transient expression of Cre (**Supplementary Fig. 10**). (**b**) MSCs from β-thalassemia major patient 1. (**c**) Thal1.52 iPS cell line. (**d**) Expression of pluripotency markers in iPS cell lines thal1.52, thal2.1, thal5.10 and thal5.11. Thal1 MSCs: MSCs from β-thalassemia major patient 1; H1 hES: hES cell line H1. (**e**) NANOG expression in iPS cell lines. (**f**) Immunohistochemical analysis of a teratoma derived from line thal1.52. Upper panel: cytokeratin (CK) 20-positive intestine-like epithelium (endoderm); middle panel: vimentin-positive fibroblastic spindle cells (mesoderm); lower panel: S-100–positive peripheral nerve (ectoderm). (**g**) Bisulfite sequencing analysis of the *OCT4* promoter in the indicated thal-iPS cell lines and the MSCs from which they were derived. Each horizontal row of circles represents an individual sequencing reaction with white circles representing unmethylated CpG dinucleotides and black circles representing methylated CpG dinucleotides. The numbers indicate the CpG position relative to the transcriptional start site (TSS). (**h**) Karyotype analysis of thal2.1 iPS cell line. Scale bars, 50 μm.

and pluripotency-related genes at similar levels to hES cell lines (**Fig. 1d**–**e** and **Supplementary Figs. 1**–**3**). Their pluripotency was assessed by formation of teratomas comprising tissues derived from all three germ layers after grafting into immunodeficient mice (**Fig. 1f** and **Supplementary Figs. 4** and **5**). They could be efficiently differentiated *in vitro* into mesoderm derivatives, such as beating putative cardiomyocytes (**Supplementary Movie 1**) and hematopoietic progenitor cells (see below). Genotyping confirmed the β-thalassemia mutations (**Supplementary Table 2** and **Supplementary Fig. 6**). Silencing of all four transgenes was demonstrated by flow cytometry (in thal-iPS cell lines derived using vectors encoding the four reprogramming factors OCT4, SOX2, KLF4 and c-MYC together with distinct fluorescent proteins[15], **Supplementary Fig. 7**), as well as quantitative reverse-transcription (qRT)-PCR (**Supplementary Fig. 8**). Demethylation of the *OCT4* promoter was assessed and confirmed in the thal-iPS cell lines thal1.52, thal2.1, thal5.10 and thal5.11 (**Fig. 1g**). All seven thal-iPS cell lines tested exhibited normal male or female karyotypes (**Fig. 1h** and **Supplementary Fig. 9**). To generate transgene-free thal-iPS cells, we selected two thal-iPS cell lines, thal5.10 and thal5.11, found to contain six copies of the single polycistronic vector flanked by *loxP* sites (fSV2A) used for reprogramming (**Supplementary Fig. 10a**), after all six copies of the fSV2A vector they both contained were mapped to the genome (**Supplementary Table 3**). Several excised thal-iPS cell lines were derived from them after transient Cre expression by an integrase-deficient lentiviral vector (Cre-IDLV). Complete excision of all six copies of the fSV2A vector (**Supplementary Fig. 10a**–**d**,**f**) and absence of integration of the Cre-IDLV vector (**Supplementary Fig. 10c**,**e**,**f**) were thoroughly documented. Altered expression of endogenous genes in the vicinity of the six integrated vectors or of the residual promoterless (U3-deleted) lentiviral long terminal repeats (LTR) before and after vector excision, respectively, was excluded by microarray analysis (**Supplementary Fig. 11**). Characterization of two vector-excised lines, thal5.10-Cre8 and thal5.11-Cre23 (derived from lines thal5.10 and thal5.11, respectively), confirmed their preserved pluripotency (**Supplementary Figs. 1**–**3**, **5**). Comparative genomic hybridization (CGH) of the excised line thal5.10-Cre8 and the parental MSCs revealed no genetic abnormalities (**Supplementary Fig. 12**).

To establish thal-iPS cell clones harboring a therapeutic β-globin gene, we generated a lentiviral vector, TNS9.3/fNG, expressing the human β-globin gene *cis*-linked to its DNAse I hypersensitive site (HS) 2, HS3 and HS4 locus control region elements, derived from the previously described TNS9 vector[13] (**Fig. 2a**). To determine the probability of retrieving sites that meet the safe harbor criteria, we analyzed 5,840 integration sites of our TNS9.3/fNG vector in the thal5.11-Cre23 iPS cell line. This survey revealed that 17.3% of all integrations met all five criteria (**Supplementary Table 4**), supporting the feasibility of recovering iPS cell clones harboring vector integrations in safe harbors from a relatively small set of clones. We thus transduced the thal-iPS cell lines thal1.52, thal2.1, thal5.10 and thal5.11 at low multiplicity of infection to isolate thal-iPS cell clones harboring a single TNS9.3/fNG vector copy. Fifteen clones found to harbor a single TNS9.3/fNG copy by quantitative PCR (**Supplementary Table 5**) were randomly selected. Single-vector integration and clonality could be thoroughly established by Southern blot analysis after digestion using two different restriction enzymes and two different probes (**Fig. 2a**,**b** and **Supplementary Fig. 13**) in 13 of them, and the vector integration sites were mapped to the human genome (**Fig. 2c**–**f** and **Table 1**). One of the 13 clones, clone thal5.10-2, was found to harbor an integration that meets all five safe harbor criteria (**Table 1**). Two additional safe harbor sites were found among 23 other sites we mapped in multiple-copy thal-iPS cell clones (**Supplementary Table 6**).

To assess vector-encoded β-globin gene expression, we derived hematopoietic progenitors through embryoid body differentiation of the 13 single-copy thal-iPS cell clones and we further differentiated them along the erythroid lineage (**Fig. 3a** and **Supplementary Fig. 14**). By the end of this process, the majority of cells exhibited characteristic hematopoietic cell morphology, expression of the

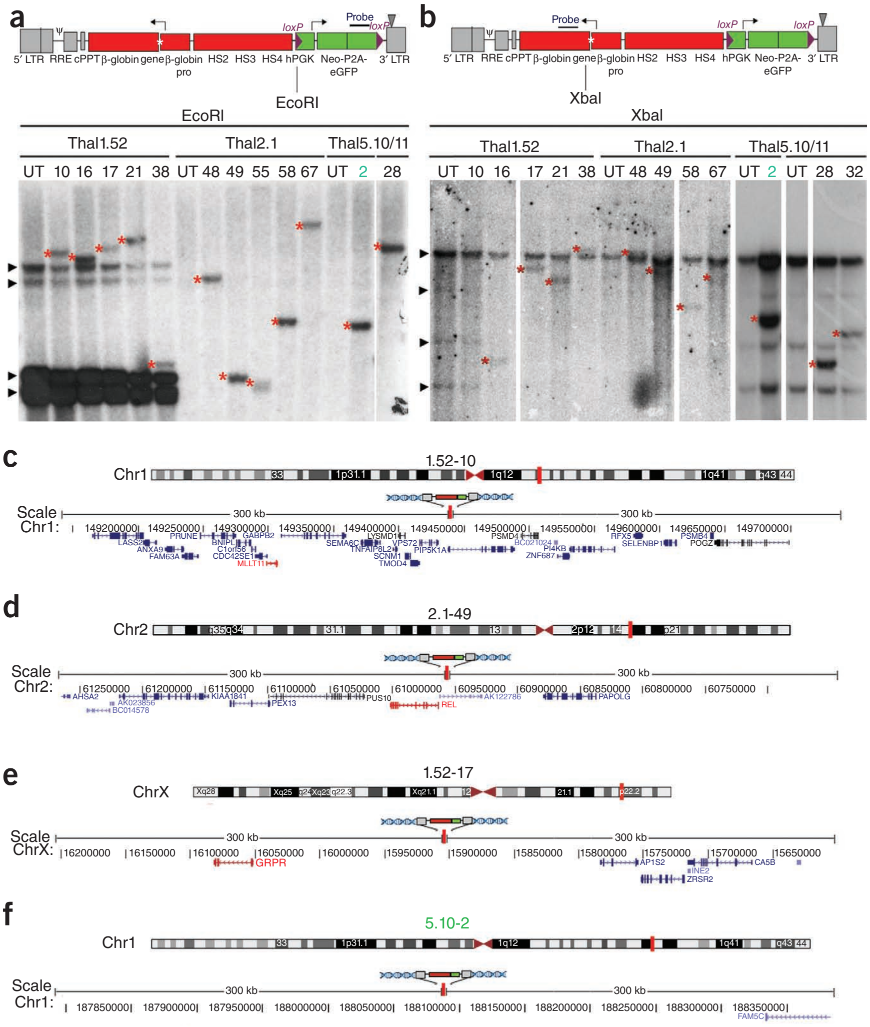


Figure 2 Single-vector copy, clonality and mapping of the integration site. (**a**,**b**) Upper panel: schematic representation of the TNS9.3/fNG lentiviral vector. An asterisk depicts a 4-bp insertion in the 5′ untranslated region (UTR) of the β-globin gene, which allows discrimination of the longer vector-encoded transcript from the endogenous β-globin transcript. TNS9.3/fNG also contains the human phosphoglycerate kinase (hPGK) promoter-driven neomycin phosphoryltransferase (Neo) and enhanced green fluorescent protein (eGFP) genes flanked by *loxP* sites. LTR: long terminal repeat; RRE: rev-responsive element; cPPT: central polypurine tract; HS: DNAse I hypersensitive site. Lower panels: Southern blot analysis to ascertain single integrations of the TNS9.3/fNG vector and clonality. Genomic DNA was digested with EcoRI (**a**) or XbaI (**b**). The probe used in **a** is eGFP (shown in the upper panel). The probe in **b** spans exons 1 and 2 of the β-globin gene (shown in the upper panel). The parental thal-iPS cell lines thal1.52, thal2.1, thal5.10 and thal5.11 and the clone number are depicted above the lanes. UT: untransduced. Arrowheads in **a** indicate bands corresponding to the reprogramming vectors pLM-GO, pLM-YS and pLM-CM (**Supplementary Fig. 8d**) present in the thal1.52 line and the clones derived from it. Arrowheads in **b** indicate endogenous bands (corresponding to the endogenous β-globin locus). Asterisks depict unique vector integration bands. (**c**–**f**) Examples of chromosome ideograms (upper panels) and graphics (lower panels) depicting 300 kb of human genome on both sides of the globin vector integration site in iPS clones thal1.52-10 (**c**), thal2.1-49 (**d**), thal1.52-17 (**e**) and the safe harbor clone thal5.10-2 (**f**). A vertical red line depicts the position of the vector insertion. Numbers depict positions in the corresponding human chromosome. All RefSeq genes present in the genomic region spanning 600 kb illustrated in the graphic are shown in blue. Genes implicated in cancer (**Supplementary Table 1**) are shown in red. Chromosome ideograms and graphics were generated with the UCSC Genome Graphs tool.

Table 1 Analysis of the globin vector integration site in 13 single-vector-copy thal-iPS cell clones with respect to the five safe harbor criteria

iPS line	Clone	Chromosome	Position	Orientation	Within 50 kb of 5′ end of any gene	Within 300 kb of cancer gene 5′ or 3′ end	Within 300 kb of miRNA 5′ or 3′ end	Inside a gene transcription unit	Within UCR
Thal1.52	1.52-10	1	149,440,726	+	+	+	−	+	−
Thal1.52	1.52-16	21	43,242,505	−	+	+	−	−	−
Thal1.52	1.52-17	X	15,909,610	−	−	+	−	−	−
Thal1.52	1.52-21	11	9,936,357	−	−	+	−	+	−
Thal1.52	1.52-38	2	232,144,880	+	+	+	+	−	−
Thal2.1	2.1-48	15	22,575,838	−	+	+	−	−	−
Thal2.1	2.1-49	2	60,964,901	−	+	+	−	+	−
Thal2.1	2.1-55	19	1,183,244	−	+	+	−	+	−
Thal2.1	2.1-58	X	33,033,291	−	−	−	−	+	−
Thal2.1	2.1-67	19	926,157	+	+	+	−	−	−
Thal5.10	**5.10-2**	**1**	**188,083,272**	**+**	**−**	**−**	**−**	**−**	**−**
Thal5.11	5.11-28	11	45,963,999	+	−	−	−	+	−
Thal5.11	5.11-32	16	1,381,251	−	+	−	−	−	−

Clone thal5.10-2 meets all five safe harbor criteria.

erythroid cell markers glycophorin A and transferrin receptor (CD71) and macroscopic hemoglobinization (**Supplementary Figs. 14** and **15**). The erythroid nature of these thal-iPS cell derivatives was further corroborated by the marked induction of well-characterized, erythroid-specific genes (**Supplementary Fig. 16**). Notably, the erythroid progeny of all wild-type and untransduced thal-iPS cell lines expressed α-globin, as well as embryonic and fetal ε- and γ- globins, albeit not the adult β-globin transcript, similarly to the erythroid progeny of the H1 hES cell line (**Fig. 3b–d** and **Supplementary Fig. 17**) and in accordance with previous reports[16–18]. Expression of vector-encoded β-globin was not detected in undifferentiated thal-iPS cell clones, as expected (**Supplementary Fig. 17**). Upon erythroid differentiation, 12 of the 13 single-copy thal-iPS cell clones expressed detectable vector-encoded β-globin. Expression levels, normalized to endogenous α-globin expression, ranged from 9% to 159% (mean, 53%) of a normal endogenous β-globin allele (**Fig. 3b,c**), similar to those we and others have obtained by lentiviral-mediated globin gene transfer in murine and human erythroid cells[14]. β-globin expression was confirmed and quantified at the protein level by high-performance liquid chromatography (HPLC) analysis in four clones (**Fig. 3d**, **Supplementary Table 7** and **Supplementary Fig. 18**). Notably, clone thal5.10-2, which expressed 85% of the level afforded by a normal endogenous β-globin allele (**Fig. 3b**), demonstrates that a globin vector, integrated in a site meeting all five of our safe harbor criteria (**Table 1**) and located >300 kb from the nearest gene 5′ end, is capable of expressing β-globin at a high level.

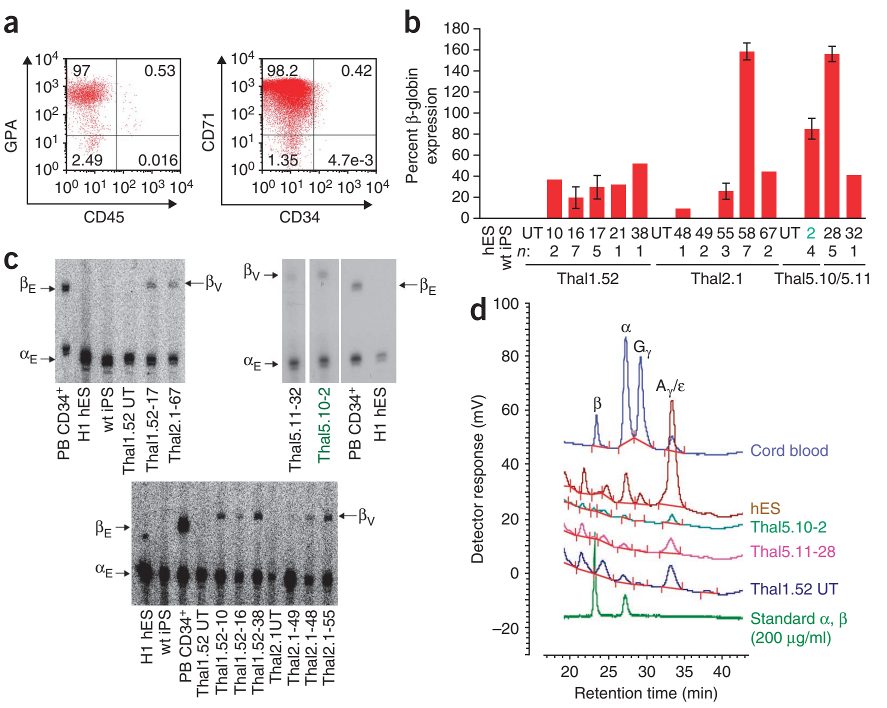


Figure 3 β-globin expression in the erythroid progeny of single-vector-copy thal-iPS cell clones. (**a**) Expression of erythroid cell markers CD71 and glycophorin A (GPA) in the erythroid progeny of thal-iPS cell line 1.52. (**b**) β-globin expression in the erythroid progeny of 13 single-vector-copy thal-iPS cell clones assessed by qRT-PCR. Expression levels are expressed per gene copy, relative to the average endogenous β-globin expressed in the *in vitro* differentiated erythroid progeny of peripheral blood CD34⁺ cells from four healthy individuals and normalized to endogenous α-globin expression. hES: erythroid progeny of hES cell line H1, wt iPS: erythroid progeny of iPS cell line FDCT, derived from fibroblasts of an 11-year-old healthy individual[30]. Numbers below graphs depict thal-iPS clone numbers derived from lines thal1.52, thal2.1, thal5.10 and thal5.11. *n*: number of independent differentiations for each clone. UT: untransduced. Error bars denote s.e.m. (**c**) β-globin expression in the erythroid progeny of a subset of single-vector-copy thal-iPS cell clones and controls assessed by quantitative primer extension. β_E: endogenous β-globin (80 bp); α_E: endogenous α-globin (60 bp); β_V: vector-encoded β-globin (84 bp). PB CD34⁺: erythroid cell derivatives of *in vitro* differentiated peripheral blood (PB) CD34⁺ cells from a normal donor; H1 hES: erythroid cell derivatives of the H1 hES line; wt iPS: erythroid cell derivatives of iPS cell line FDCT, derived from fibroblasts of a healthy individual; thal1.52 UT, thal2.1 UT: erythroid cell derivatives of untransduced lines thal1.52 and thal2.1, respectively; thal1.52-17, thal2.1-67, thal1.52-10, thal1.52-16, thal1.52-38, thal2.1-49, thal2.1-48, thal2.1-55, thal5.11-32, thal5.10-2: erythroid cell derivatives of the respective single-vector-copy thal-iPS clones. (**d**) Chromatograms of HPLC analysis of α- and β-globin expression in the erythroid progeny of clones thal5.10-2 and thal5.11-28. Cord blood, H1 hES and untransduced (UT) thal1.52 cells were used as controls. For quantification of these data, see **Supplementary Table 7**.

Expression of genes located within 300 kb of the vector insertion site was assessed in six single-copy thal-iPS cell clones in both the undifferentiated state, as well as in the erythroid progeny by microarrays. This analysis revealed that three out of five integrations eliminated by our safe harbor criteria did indeed result in perturbed expression of neighboring genes (**Supplementary Figs. 19** and **20**). Dysregulated expression was detected in a total of five genes present at a distance ranging from 9 to 275 kb from the vector insertion, whereas we did not detect any genes beyond 300 kb of the insertion to be significantly differentially expressed ($P < 0.05$). Of note, the safe harbor integration site in clone thal5.10-2 is in a genomic region with no genes within 300 kb on either side. The microarray analysis did not reveal any statistically significant differentially expressed genes elsewhere in the genome in this clone or any other.

Our data demonstrate that the generation and identification of transgene-expressing iPS cell clones, in which transgene expression is obtained at therapeutic levels in iPS cell progeny from selected chromosomal sites, are feasible by screening a limited number of single-copy clones and applying five safe harbor criteria for their selection. Approximately half (47.7%) of the clones we obtained under optimized transduction conditions harbored a single vector copy (**Supplementary Table 5**), and clonality could be confirmed in 13 out of 15 (86.7%) of them. As the frequency of integrations in sites that meet our five safe harbor criteria is 17.3% (**Supplementary Table 4**), the overall efficiency of our strategy is 7.1%. Three out of five clones eliminated by our safe harbor criteria showed perturbed expression of neighboring endogenous genes, which was not the case in clone thal5.10-2, demonstrating the usefulness of selecting genetically modified iPS cell clones based on this strategy and these criteria. Notably, applying our criteria to a series of gamma-retroviral and lentiviral integration sites associated with oncogenic events or perturbed endogenous gene expression would effectively eliminate all of these well-characterized deleterious integrations[7–10].

This approach has the prospect of broad application in genetic engineering of human iPS cells. Genetic correction through addition of a therapeutic gene into safe harbors in patient-specific iPS cells provides a realistic alternative strategy to targeted gene repair, especially for genetically heterogeneous disorders associated with multiple mutations. In contrast to genome editing strategies, our approach does not require customized targeting vectors with long isogenic ends[19] or complex genotoxicity screens that are needed when using endonucleases[20,21]. In the latter case, the risk of occult genotoxicity mediated by off-target effects of double-stranded DNA cleaving agents needs to be balanced against the long-term experience with risk assessment of retroviral vector integration, which can be thoroughly analyzed, as we demonstrate here. Apart from genetic correction, future clinical applications of iPS cells will likely require addition of drug resistance, reporter or suicide genes to permit *in vivo* selection, tracking or cell eradication, respectively. To this end, the identification of suitable genomic locations for transgene knock-in is of great importance. Recent studies suggest that genomic sites, such as the adeno-associated virus integration site 1 (AAVS1)[20,22] and the human ROSA26 locus[23], can support transgene expression, but data on the safety of these sites are lacking. The screening strategy we describe here should prove useful for the *de novo* discovery and characterization of putative universal genomic safe harbors. The requirements for a safe harbor are (i) avoidance of genotoxicity and (ii) support of the appropriate expression level and regulation of the integrated transgene. Notably, β-globin gene expression in the safe harbor clone thal5.10-2 was in the therapeutic range, which, based on clinical observations in individuals with homozygous β-thalassemia and hereditary persistence of fetal hemoglobin, is on the order of 30% of α-globin expression[24].

The potential genotoxicity of the reprogramming process used upstream of our safe harbor strategy also needs to be taken into account. In this study we used an excisable vector system and selected patient-specific iPS cell lines harboring a relatively low number of reprogramming vector copies and determined their position in the genome. Since Cre-mediated excision leaves behind a promoterless, U3-deleted LTR, we propose that lines can be selected—as we demonstrate here—on the basis of (i) exclusion of all integrations within exons, to avoid frame-shift, premature termination of translation or translation of abnormal proteins and (ii) ascertainment of lack of perturbation of gene expression by residual LTR fragments that reside within transcription units. Based on our large integration site data set in human iPS cells, 97% of all lentiviral vector integrations are outside exons. The need to screen thal-iPS cell lines for residual LTR insertions may be eliminated if efficient generation of human iPS cells using nonintegrating systems becomes a realistic option.

Ascertainment of lack of perturbation of gene expression in the host cell in both a local and genome-wide range, as shown here (**Supplementary Figs. 19** and **20**), provides an important initial safety test. This can be complemented by additional tests for features of neoplastic transformation[25] and, eventually, by serial transplantation studies of iPS cell–derived hematopoietic stem cells in immunodeficient mice, currently precluded by the inability to efficiently generate engraftable human hematopoietic stem cells derived from ES and iPS cells[5,26]. Further evaluation of safe harbors could also include long-term studies in transgenic mice bearing transgenes in syntenic regions, as well as bioinformatics-assisted searches in the cumulated databases of common retroviral integration sites found in patients treated with retroviral vectors and not associated with any side effect[27–29], although this information is of limited value in the absence of transgene expression data at these sites.

In conclusion, the present study provides a framework and a strategy combining bioinformatics and functional analyses for identifying safe harbors for transgene integration in the human genome. As our understanding of the function of the human genome and of genome-wide interactions advances, the definition of safe harbors will likely be refined over time, eventually building a registry of dependable genomic locations for the safe and effective genetic engineering of human cells.

METHODS

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

Accession code. GEO: GSE24901.

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

ACKNOWLEDGMENTS

We thank X. Wang and N. Wu for assistance with HPLC analysis; L. Ferro, E. Reed, J. Miller, M. Leversha and M. Tomishima for technical assistance; F. Boulad, Memorial Sloan-Kettering Cancer Center New York for bone marrow specimens; and A. Athanassiadou for advice on β-thalassemia genotyping. pCMVΔR8.91N/N was kindly provided by E. Poeschla, Mayo Clinic, Rochester, Minnesota. This work was supported by the Starr Foundation (Tri-Institutional Stem Cell Initiative, Tri-SCI-018), the New York State Stem Cell Science, NYSTEM (N08T-060) and National Heart, Blood, and Lung Institute grant HL053750 (M.S.). F.D.B., S.L.R. and N.M. were supported by National Institutes of Health grants AI052845 and AI082020 (F.D.B.). G.L. was supported by a New York Stem Cell Foundation Druckenmiller fellowship.

AUTHOR CONTRIBUTIONS

E.P.P. conceived and designed the study, designed and performed experiments, analyzed data and wrote the manuscript; G.L. performed iPS cell differentiation

experiments; N.M. performed bioinformatics analyses; M.S. and C.L. analyzed microarray data; L.M.S.T. provided technical assistance; K.K. performed histological analyses of teratomas; S.L.R. generated and analyzed integration site data; P.G. provided skin biopsy samples from β-thalassemia patients; A.V. generated microarray data; I.R., F.D.B. and L.S. analyzed data; M.S. conceived and designed the study, 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. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861–872 (2007).
2. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**, 1917–1920 (2007).
3. Park, I.H. *et al.* Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* **451**, 141–146 (2008).
4. Hanna, J. *et al.* Treatment of sickle cell anemia mouse model with iPS cells generated from autologous skin. *Science* **318**, 1920–1923 (2007).
5. Raya, A. *et al.* Disease-corrected haematopoietic progenitors from Fanconi anaemia induced pluripotent stem cells. *Nature* **460**, 53–59 (2009).
6. Schroder, A.R. *et al.* HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**, 521–529 (2002).
7. Hacein-Bey-Abina, S. *et al.* LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science* **302**, 415–419 (2003).
8. Ott, M.G. *et al.* Correction of X-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of MDS1–EVI1, PRDM16 or SETBP1. *Nat. Med.* **12**, 401–409 (2006).
9. Howe, S.J. *et al.* Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. *J. Clin. Invest.* **118**, 3143–3150 (2008).
10. Cavazzana-Calvo, M. *et al.* Transfusion independence and HMGA2 activation after gene therapy of human beta-thalassaemia. *Nature* **467**, 318–322 (2010).
11. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
12. Kustikova, O. *et al.* Clonal dominance of hematopoietic stem cells triggered by retroviral gene marking. *Science* **308**, 1171–1174 (2005).
13. May, C. *et al.* Therapeutic haemoglobin synthesis in beta-thalassaemic mice expressing lentivirus-encoded human beta-globin. *Nature* **406**, 82–86 (2000).
14. Sadelain, M., Boulad, F., Lisowki, L., Moi, P. & Riviere, I. Stem cell engineering for the treatment of severe hemoglobinopathies. *Curr. Mol. Med.* **8**, 690–697 (2008).
15. Papapetrou, E.P. *et al.* Stoichiometric and temporal requirements of Oct4, Sox2, Klf4, and c-Myc expression for efficient human iPSC induction and differentiation. *Proc. Natl. Acad. Sci. USA* **106**, 12759–12764 (2009).
16. Chang, K.H. *et al.* Definitive-like erythroid cells derived from human embryonic stem cells coexpress high levels of embryonic and fetal globins with little or no adult globin. *Blood* **108**, 1515–1523 (2006).
17. Qiu, C., Olivier, E.N., Velho, M. & Bouhassira, E.E. Globin switches in yolk sac-like primitive and fetal-like definitive red blood cells produced from human embryonic stem cells. *Blood* **111**, 2400–2408 (2008).
18. Chang, K.H. *et al.* Globin phenotype of erythroid cells derived from human induced pluripotent stem cells. *Blood* **115**, 2553–2554 (2010).
19. Giudice, A. & Trounson, A. Genetic modification of human embryonic stem cells for derivation of target cells. *Cell Stem Cell* **2**, 422–433 (2008).
20. Hockemeyer, D. *et al.* Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases. *Nat. Biotechnol.* **27**, 851–857 (2009).
21. Zou, J. *et al.* Gene targeting of a disease-related gene in human induced pluripotent stem and embryonic stem cells. *Cell Stem Cell* **5**, 97–110 (2009).
22. Smith, J.R. *et al.* Robust, persistent transgene expression in human embryonic stem cells is achieved with AAVS1-targeted integration. *Stem Cells* **26**, 496–504 (2008).
23. Irion, S. *et al.* Identification and targeting of the ROSA26 locus in human embryonic stem cells. *Nat. Biotechnol.* **25**, 1477–1482 (2007).
24. Safaya, S., Rieder, R.F., Dowling, C.E., Kazazian, H.H. Jr. & Adams, J.G. 3rd Homozygous beta-thalassemia without anemia. *Blood* **73**, 324–328 (1989).
25. Werbowetski-Ogilvie, T.E. *et al.* Characterization of human embryonic stem cells with features of neoplastic progression. *Nat. Biotechnol.* **27**, 91–97 (2009).
26. Ji, J. *et al.* OP9 stroma augments survival of hematopoietic precursors and progenitors during hematopoietic differentiation from human embryonic stem cells. *Stem Cells* **26**, 2485–2495 (2008).
27. Deichmann, A. *et al.* Vector integration is nonrandom and clustered and influences the fate of lymphopoiesis in SCID-X1 gene therapy. *J. Clin. Invest.* **117**, 2225–2232 (2007).
28. Aiuti, A. *et al.* Multilineage hematopoietic reconstitution without clonal selection in ADA-SCID patients treated with stem cell gene therapy. *J. Clin. Invest.* **117**, 2233–2240 (2007).
29. Schwarzwaelder, K. *et al.* Gammaretrovirus-mediated correction of SCID-X1 is associated with skewed vector integration site distribution in vivo. *J. Clin. Invest.* **117**, 2241–2249 (2007).

ONLINE METHODS

Lentiviral vector construction and production. The four bicistronic vectors pLM-GO, pLM-YS, pLM-RK and pLM-CM, encoding violet excited GFP (vexGFP)-P2A-OCT4, mCitrine-P2A-SOX2, mCherry-P2A-KLF4 and mCerulean-P2A-cMYC, respectively, used to generate iPS cells from subject 1, have been previously described[15]. The single polycistronic vector pLM-SV2A was constructed as follows: Klf4-P2A-cMYC and cMyc-E2A-SOX2 cassettes were generated by overlapping PCR using Pfu polymerase (Stratagene) with primers introducing the respective intervening 2A peptide preceded by a Gly-Ser-Gly linker and restriction enzyme sites in the ends of each cassette. The Klf4-P2A-cMYC cassette was inserted into NcoI and EcoRI sites of the polylinker of cloning plasmid pSL1180. The cMYC-E2A-SOX2 cassette was digested with ClaI (site within the cMYC cDNA) and EcoRI and ligated downstream of the previous cassette. The OCT4 cDNA and a linker encoding T2A preceded by a Gly-Ser-Gly linker were ligated between AgeI and NcoI sites 3′ to the previously ligated cassettes. The entire OCT4-T2A-KLF4-P2A-cMYC-E2A-SOX2 cassette was transferred as an AgeI/SalI fragment into the pLM lentiviral vector backbone[15] downstream of the human phosphoglycerate kinase (hPGK) promoter and upstream of the woodchuck hepatitis virus post-transcriptional regulatory element (WPRE) to generate the pLM-SV2A vector. The final vector was sequence-verified by DNA sequencing. Expression and correct processing of all four factors, OCT4, KLF4, SOX2 and c-MYC was confirmed by western blot analysis in transduced human MRC-5 fibroblasts (data not shown). The 'floxed' polycistronic vector pLM-fSV2A was derived from SV2A after insertion of annealed oligonucleotides containing a *loxP* site in a NheI site in the deleted U3 region of the 3′ LTR. The TNS9.3/fNG vector was derived from TNS9 (ref. 13) after insertion of a hPGK-Neo-P2A-eGFP cassette flanked by *loxP* sites and insertion of a 4-bp sequence in the β-globin 5′ UTR. The bicistronic cassette expressing the neomycin phosphoryltransferase (Neo) gene and eGFP linked by a P2A peptide preceded by a Gly-Ser-Gly linker was generated by overlapping PCR and inserted into AgeI and BglII sites of the pSL1180 cloning vector. The hPGK promoter was inserted between NotI/AgeI sites upstream of the Neo-P2A-eGFP cassette. To introduce *loxP* sites, we inserted annealed oligonucleotides in a MluI site, upstream of hPGK and in a SalI site downstream of eGFP, respectively. The entire floxed cassette was transferred as a MluI/SalI fragment into TNS9.3. For construction of an integrase-deficient lentiviral vector for Cre-mediated excision, an mCherry-P2A-Cre recombinase cassette was generated by overlapping PCR and inserted into the pLM lentiviral vector backbone under the transcriptional control of the cytomegalovirus immediate early promoter (pCMV). All oligonucleotide sequences are provided in **Supplementary Table 8**.

Vector production was performed by triple co-transfection of the plasmid DNA encoding the vector, pUCMD.G and pCMVΔR8.91 into 293T cells, as previously described[31]. For packaging of the integrase-deficient lentiviral vector encoding mCherry and Cre, pCMVΔR8.91 was replaced by pCMVΔR8.91N/N[32] (kindly provided by E. Poeschla).

Human iPS cell generation. Skin punch biopsy specimens were obtained after informed consent from patients with β-thalassemia major at the Thalassemia Center at Cornell University. To establish fibroblast cell cultures, we sliced the biopsy into <1 mm fragments, which were transferred into 60 mm plates containing Eagle's Modified Essential Medium with 10% FBS (FBS) (Hyclone). A cover slip was placed on top of each biopsy fragment and the plates were left undisturbed for 7–10 d to allow migration of cells.

Cryopreserved whole bone marrow specimens obtained from the Bone Marrow Transplantation Center at Memorial Sloan-Kettering Cancer Center (MSKCC) were thawed and, after density gradient separation over Ficoll, mononuclear cells were plated on tissue culture–treated dishes in Complete MesenCult Medium (Stem Cell Technologies). After ~2 weeks, adherent, fibroblast-like cells (**Fig. 1b**) were harvested by trypsinization and expanded.

iPS cell generation was performed as previously described[15]. Skin fibroblasts or MSCs at passages 2–7 were plated in gelatin-coated, 6-well plates at a density of 1 × 10^5 cells per well and transduced 24 h later with lentiviral vectors encoding OCT4, SOX2, KLF4 and c-MYC in the presence of 4 μg/ml polybrene. Media were changed 24 h later and replaced every day thereafter with hES cell media supplemented with 6 ng/ml FGF2 (R&D Systems) and 0.5 mM VPA (Sigma). Fifteen to 25 d after transduction, colonies with hES cell morphology were mechanically dissociated and transferred into plates pre-seeded with mitomycin C–treated mouse embryonic fibroblasts (MEFs) (GlobalStem). Cells were thereafter passaged with dispase and expanded to establish iPS cell lines.

The vector systems used for iPS cell generation were as follows: a combination of four bicistronic lentiviral vectors co-expressing OCT4, KLF4, c-MYC and SOX2 with a distinct fluorescent protein[15] (subject 1), a polycistronic vector co-expressing all four factors in a single transcript, SV2A (subject 2) and its derivative fSV2A, containing a *loxP* site in the 3′ LTR (subjects 4 and 5) (**Supplementary Fig. 9d**).

iPS cell characterization. Flow cytometry analysis, immunofluorescence, OCT4 promoter methylation analysis, karyotyping and teratoma formation assays were performed as described[15,30].

For assessment of expression of pluripotency genes, total RNA from thal-iPS cell lines was isolated with Trizol (Invitrogen). Reverse transcription was performed with Superscript III (Invitrogen) and qPCR was performed with primers shown in **Supplementary Table 9** using SYBR Green. Reactions were carried out in duplicate in an ABI PRISM 7500 Sequence Detection System (Applied Biosystems). Expression was calculated by relative quantification using the $\Delta\Delta C_t$ method with actin as endogenous control.

For teratoma formation assays, undifferentiated iPS cells were suspended in hES medium containing 10 μM of the Rho-associated kinase (Rock) inhibitor Y-27632 (Tocris)[33]. Approximately 2 × 10^6 cells were injected intramuscularly into NOD-SCID *IL2Rg*$^{-null}$ mice (Jackson Laboratory). Five to six weeks later, the tumors were surgically dissected and fixed in 4% formaldehyde. Cryosectioned samples were stained with hematoxylin and eosin for histological analysis and with antibodies against cytokeratin (CK) 20, vimentin and S-100 for immunohistochemical analysis. All animal experiments were conducted in accordance with protocols approved by MSKCC Institutional Animal Care and Use Committee (IACUC) and following National Institutes of Health guidelines for animal welfare.

Assessment of reprogramming vector silencing. qRT-PCR was performed with the primers and probes shown in **Supplementary Table 9**. Reactions were carried out in duplicate in an ABI PRISM 7500 Sequence Detection System (Applied Biosystems). Expression was calculated by relative quantification using the $\Delta\Delta C_t$ method with GAPDH as endogenous control.

Karyotyping. Standard G-banding analysis was performed at the MSKCC molecular cytogenetics core laboratory. Chromosome analysis was performed on a minimum of 10 4,6-diamidino-2-phenylindole (DAPI)-banded metaphases. All metaphases were fully karyotyped.

β-thalassemia genotyping. Genomic DNA was extracted from thal-iPS cell lines, dermal fibroblasts and MSCs using the DNeasy kit (Qiagen). 200 ng of DNA was used as template in a PCR reaction using the primer pair β-thal-1 (**Supplementary Table 9**). The 714-bp PCR product was gel-purified and sequenced.

Cre-mediated vector excision. iPS cell lines thal5.10 and thal5.11 were dissociated into single cells with accutase for 30 min at 37 °C. After incubation for 1 h on gelatin-coated plates to allow adherence and subsequent removal of MEFs, cells were plated at 1 × 10^5 cells per well of a 6-well plate with Matrigel (BD Biosciences) in MEF-conditioned media supplemented with 6 ng/ml FGF2 and 10 μM of Y-27632. The next day the cells were transduced with vector supernatants in the presence of 4 μg/ml polybrene for 16 h. The transduced cells were dissociated with accutase 48 h later, vigorously triturated into single cells and replated at titrated densities (from 100 to 500 cells per cm^2) on a layer of mitomycin C–treated MEFs with 10 μM of Y-27632. An aliquot of the cells was used for flow cytometry analysis of mCherry-Cre expression. After 10–15 d, single cell colonies were mechanically dissociated and replated into 6-well dishes on mitomycin C–treated MEFs. One week later, ~100–200 cells from each clone were manually picked under a stereoscope into 0.2 ml tubes and lysed in 25 μl lysis buffer with 100 μg/ml proteinase K, as previously described[34]. We used 3 μl of cell lysate to screen for excision of the fSV2A vector by multiplex PCR with three primer pairs: fSV2A- 1, fSV2A-2 and LTR (**Supplementary Fig. 10b**). The PCR products were analyzed by agarose gel electrophoresis. Six out of 47 screened clones from line thal5.10 and 7 out of 42 clones from line thal5.11 were found to no longer contain the integrated provirus. To assess complete vector excision and lack of integration of the Cre-expressing vector, we performed qPCR with primers and probes specific for the gag region of the lentiviral vectors and the human albumin gene (**Supplementary Table 9**). For Southern blot analysis 5 μg of genomic DNA was digested with XmaI or BglI and probed with a radiolabeled SalI-KpnI fragment spanning the WPRE.

Globin gene transfer and selection of single vector copy thal-iPS cell clones. iPS cell lines thal1.52 and thal2.1 were prepared and transduced with varying MOI of the TNS9.3/fNG vector, as described above. The transduced cells were dissociated 48 h later with accutase and vigorously triturated into single cells. An aliquot of the cells was used for flow cytometry analysis of eGFP expression. Transduced cells with gene transfer <30% (as estimated by the percentage of eGFP$^+$ cells) were replated at a density of 1,500 cells/cm^2 on a layer of Neo-resistant mitomycin C–treated MEFs (GlobalStem). G418 (Invitrogen) was added at a concentration of 12.5 μg/ml between days 5 and 9 after transduction. Approximately 20 d post-transduction, Neo-resistant colonies were manually picked and replated into 6-well dishes on mitomycin C-treated MEFs. One week later, ~100–200 cells from each clone were manually picked and lysed as described above. We used 5 μl of cell lysate for measurement of TNS9.3/fNG vector copy number (VCN) with multiplex quantitative PCR (qPCR) using sets of primers and probes specific for the globin vector (GV1) and for the human albumin gene (**Supplementary Table 9**). To determine absolute VCN, we generated a standard curve using serial dilutions of a plasmid containing both vector and albumin gene amplicons. Reactions were carried out in triplicate in an ABI 7500 detection system (Applied Biosystems). We digested 5–10 μg of genomic DNA extracted from single vector copy iPS clones with NcoI, XbaI or EcoRI and analyzed it by Southern blot analysis, as described[13,34], using a radiolabeled NcoI-BamHI fragment spanning exons 1 and 2 of the human β-globin gene or the eGFP cDNA as probe.

Integration site analysis. fSV2A vector integrations were mapped by linear amplification mediated (LAM)-PCR, using digestion with Tsp509I, as described[35]. PCR products were TOPO cloned and sequenced.

For analysis of TNS9.3/fNG vector integration sites in thal-iPS cells with relation to the 5 safe harbor criteria, line thal5.10-Cre8 was transduced with the vector at high MOI (~100) as described above and genomic DNA was extracted from the polyclonal population 5 d after transduction. Integration sites were isolated by ligation mediated (LM)-PCR, sequenced by 454/Roche pyrosequencing, processed and analyzed, as previously described[36].

TNS9.3/fNG vector integration sites in single-copy clones were mapped by inverse PCR (iPCR). We digested 1 μg of genomic DNA with HinP1I or HpyCH4IV, and diluted and incubated it with T4 DNA ligase. After phenol/chloroform extraction and ethanol precipitation, DNA was digested with XbaI or SalI and used as template in a PCR reaction with primers iPCR F and iPCR R (**Supplementary Table 9**). The PCR product was analyzed on a 3% agarose gel and all bands visualized with ethidium bromide were excised, purified and sequenced.

Integration sites were judged to be authentic if the sequences were adjacent to vector LTR ends and had a unique hit when aligned to the draft human genome (University of California Santa Cruz, UCSC hg18) using BLAT (http://genome.ucsc.edu/)[37]. Genomic annotations were also obtained from UCSC hg18 Genome Browser and mapped against the integration sites.

Integration sites were confirmed by PCR with LTR universal forward primer for fSV2A vector integrations or GV forward primer for TNS9.3/fNG integrations and reverse primers specific for the genomic sequence adjacent to the integration. All primer sequences are shown in **Supplementary Table 9**.

For computing the frequencies of integration sites with relation to our safe harbor criteria, gene data were obtained from UCSC RefSeq Gene and wgRna (miRNA) track version 1/31/10. A few of the gene symbols in RefSeq gene track did not match up to the RefSeq gene database from NCBI. Therefore, the mismatched gene symbols from UCSC database were converted to the proper gene symbols as found on NCBI. A few sources used a gene alias instead of gene symbol in which case the corresponding gene symbol was found using NCBI's gene info table: ftp://ftp.ncbi.nih.gov/gene/DATA/gene_info.gz. UCRs in the human genome were obtained from reference 11 and the data were downloaded from http://users.soe.ucsc.edu/~jill/ultra.html. As the genomic coordinates used in the publication were from an older assembly, we converted the coordinates to the hg18 freeze using UCSC lift genome annotations tool.

Erythroid differentiation. For hematopoietic differentiation of human iPS cells, embryoid bodies were generated and cultured, as previously described[30]. Briefly, intact human iPS cell colonies were collected with dispase and plated in low-attachment dishes (Corning) in DMEM with 20% FBS, 1% nonessential amino acids (NEAA), 1 mM L-glutamine and 0.1 mM β-mercaptoethanol (MTG), supplemented with 40 ng/ml bone morphogenetic protein 4 (BMP4) and 40 ng/ml vascular endothelial growth factor (VEGF). Two days later, the medium was switched to X-VIVO 15 (Lonza) with 1% NEAA, 1 mM L-glutamine and 0.1 mM MTG supplemented with 40 ng/ml BMP4, 40 ng/ml VEGF, 20 ng/ml FGF2, 40 ng/ml stem cell factor (SCF), 40 ng/ml Flt3 ligand (Flt3L) and 40 ng/ml thrombopoietin (TPO) and replaced every 3 d. At day 8 of embryoid body culture cells were dissociated with accutase and passed through a 22G needle 3–4 times. For further erythroid differentiation the cells were plated in low-attachment dishes in X-VIVO 15 supplemented with 20% BIT (Stem Cell Technologies), 1% NEAA, 1 mM L-glutamine, 0.1 mM MTG, 100 ng/ml SCF, 6 U/ml erythropoietin and 10^{-6} M dexamethasone. Media were replenished every 3 d for 15 d. Benzidine staining was performed as described[38]. Erythroid differentiation of CD34$^+$ cells from mobilized peripheral blood of two healthy individuals was performed as previously described[38].

Flow cytometry. Undifferentiated iPS cells were dissociated with accutase, stained with Alexa Fluor 647-conjugated anti-Tra-1–81 or anti-Tra-1-60 or anti-SSEA3 or anti-SSEA4 and PE-Cy5-conjugated anti-HLA-ABC antibodies (BD Biosciences). The erythroid progeny of iPS cells were incubated with allophycocyanin (APC)-conjugated anti-CD34 or APC-conjugated anti-glycophorin A (GPA), PerCP-conjugated anti-CD45 and PE-conjugated anti-CD71 (BD Biosciences). Data were acquired in a LSRII cytometer (BD Biosciences) and analyzed with the FlowJo software (version 8.8.4; Tree Star).

Analysis of β-globin expression. Total RNA was isolated with Trizol (Invitrogen). For quantitative RT-PCR, reverse transcription was performed with Superscript III (Invitrogen) and qPCR was performed with primers and probes specific for the human α- and β-globin transcripts (**Supplementary Table 9**). RNA from human CD34$^+$ cells isolated from mobilized peripheral blood from four healthy individuals and differentiated *in vitro* along the erythroid lineage, as described above, was used as reference. Reactions were carried out in triplicate in an ABI PRISM 7500 Sequence Detection System (Applied Biosystems). Vector-encoded β-globin expression per gene copy was calculated by relative quantification using the $\Delta\Delta C_t$ method with α-globin as endogenous control, relative to the average expression in four reference samples (accounting for two endogenous β-globin alleles).

The results of quantitative RT-PCR were corroborated in selected samples by quantitative primer extension assay with [^{32}P]dATP end-labeled primers PE-alpha and PE-beta (**Supplementary Table 9**) specific for the human α- and β- globin transcripts, respectively, as previously described[13]. Briefly, the radiolabeled primers were annealed to 0.25–1 μg of RNA and reactions were performed using the Primer Extension System-AMV Reverse Transcriptase kit (Promega). The predicted product length is 60 bp for the α-globin transcript, 80 bp for the endogenous β-globin transcript and 84 bp for the vector-encoded β-globin transcript. Radioactive bands were quantified by phosphorimager analysis (BioRad).

Tissue specificity of vector-encoded β-globin expression was assessed by qRT-PCR in undifferentiated thal-iPS cell clones transduced with the TNS9.3/fNG vector and their erythroid progeny using primers and probes specific for β-globin and GAPDH (Applied Biosystems).

HPLC analysis was performed at the MSKCC analytical pharmacology core laboratory. Frozen cord blood and cell pellets were thawed at 24 °C. The blood samples were proportionally diluted volumetrically to be in the calibration standard curve range (10–400 μg/ml). Cell pellets were incubated with 0.1%

doi:10.1038/nbt.1717

of sodium lauryl sulfate solution in an ice-water batch for 15 min. All samples were then centrifuged at 14,000*g* for 5 min and the supernatants were filtered with 0.45 μm polyethersulfone syringe filters before the assay. A gradient elution with a VYDAC Protein C4 column of 250-mm length (inner diameter, 4.6 mm; particle size, 5 μm) and a mobile phase containing acetonitrile and 0.1% trifluoroacetic acid (mobile phase A: 4/1, vol/vol and mobile phase B: 2/3, vol/vol) were used and the mobile phase composition was changed from 10% B to 46% B over 40 min. The separation of the sub-chains of hemoglobin from any potential interference was monitored at 220 nm and the flow rate was set at 1.0 ml/min. Calibration curves were determined for the α- and β-globin chains to permit conversion of peak areas to individual sub-chain amounts against the external reference standards.

Expression microarray analysis. Whole genome gene expression analysis was performed on Illumina BeadArrays at the MSKCC genomics core laboratory. The summarized data from the chips were normalized by variance stabilization normalization (vsn) using the vsn package in Bioconductor[39]. We used the vsn method to correct for possible batch effects in the data. The differential gene expression analysis was performed using limma Bioconductor package[40]. The limma package uses linear models to assess differential expression and uses empirical Bayesian methods to provide stable results even when the number of arrays is small. Multiple hypothesis correction was performed using the Benjamini-Hochberg method. Expression of each gene within 300 kb on either side of the globin vector integration site was compared to expression in all other clones with different vector insertions, as well as in untransduced lines.

CGH array analysis. Comparative genomic hybridization analysis of iPS line thal5.10-Cre8 (vector-excised) and the thal5 MSCs this line was derived from was performed on the Agilent 1M CGH platform at the MSKCC genomics core laboratory. The data were normalized using GC-RMA normalization. Circular Binary Segmentation[41] from the DNAcopy package of Bioconductor was used to determine any significant copy number alterations. A segment mean of < -0.3 or $> +0.3$ is generally considered to be an aberration and we did not find any aberrations in the analyzed sample using this threshold.

30. Lee, G. *et al.* Modelling pathogenesis and treatment of familial dysautonomia using patient-specific iPSCs. *Nature* **461**, 402–406 (2009).
31. Papapetrou, E.P., Kovalovsky, D., Beloeil, L., Sant'angelo, D. & Sadelain, M. Harnessing endogenous miR-181a to segregate transgenic antigen receptor expression in developing versus post-thymic T cells in murine hematopoietic chimeras. *J. Clin. Invest.* **119**, 157–168 (2009).
32. Saenz, D.T. *et al.* Unintegrated lentivirus DNA persistence and accessibility to expression in nondividing cells: analysis with class I integrase mutants. *J. Virol.* **78**, 2906–2920 (2004).
33. Watanabe, K. *et al.* A ROCK inhibitor permits survival of dissociated human embryonic stem cells. *Nat. Biotechnol.* **25**, 681–686 (2007).
34. Papapetrou, E.P., Ziros, P.G., Micheva, I.D., Zoumbos, N.C. & Athanassiadou, A. Gene transfer into human hematopoietic progenitor cells with an episomal vector carrying an S/MAR element. *Gene Ther.* **13**, 40–51 (2006).
35. Schmidt, M. *et al.* High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). *Nat. Methods* **4**, 1051–1057 (2007).
36. Wang, G.P., Ciuffi, A., Leipzig, J., Berry, C.C. & Bushman, F.D. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res.* **17**, 1186–1194 (2007).
37. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
38. Papapetrou, E.P., Korkola, J.E. & Sadelain, M. A genetic strategy for single and combinatorial analysis of miRNA function in mammalian hematopoietic stem cells. *Stem Cells* **28**, 287–296 (2009).
39. Huber, W., von Heydebreck, A., Sueltmann, H., Poustka, A. & Vingron, M. Parameter estimation for the calibration and variance stabilization of microarray data. *Stat. Appl. Genet. Mol. Biol.* **2**, Article 3 (2003).
40. Smyth, G.K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, Article 3 (2004).
41. Venkatraman, E.S. & Olshen, A.B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657–663 (2007).

nature
biotechnology

Toolkit for evaluating genes required for proliferation and survival using tetracycline-regulated RNAi

Johannes Zuber[1,5], Katherine McJunkin[1,2,5], Christof Fellmann[1,4], Lukas E Dow[1], Meredith J Taylor[1], Gregory J Hannon[1–3] & Scott W Lowe[1–3]

Short hairpin RNAs (shRNAs) are versatile tools for analyzing loss-of-function phenotypes *in vitro* and *in vivo*[1]. However, their use for studying genes involved in proliferation and survival, which are potential therapeutic targets in cancer and other diseases, is confounded by the strong selective advantage of cells in which shRNA expression is inefficient. We therefore developed a toolkit that combines Tet-regulated miR30-shRNA technology, robust transactivator expression and two fluorescent reporters to track and isolate cells with potent target knockdown. We demonstrated that this system improves the study of essential genes and was sufficiently robust to eradicate aggressive cancer in mice by suppressing a single gene. Further, we applied this system for *in vivo* negative-selection screening with pooled shRNAs and propose a streamlined, inexpensive workflow that will facilitate the use of RNA interference (RNAi) for the identification and evaluation of essential therapeutic targets.

RNAi technologies enable specific suppression of the expression of any gene through a conserved cellular machinery[2]. shRNAs can be expressed from DNA-based vectors integrated into the genome[3,4], thus enabling the study of stable loss-of-function phenotypes *in vitro* or in mouse models[5,6]. Genetic screens using focused or genome-wide shRNA libraries can identify putative therapeutic targets—for example, by identifying genes that are selectively required for the proliferation or survival of cancer cells[7–10]. A particularly useful approach to evaluate such genes is Tet-On RNAi, which uses a two-component conditional expression system that requires a reverse tetracycline transactivator (rtTA) and a tetracycline-responsive element (TRE) promoter driving shRNA expression[6,11,12]. In this system, shRNA expression is induced by doxycycline, enabling synchronous and reversible gene knockdown in established cell populations. Thus, in animal models, target inhibition can be triggered transiently after disease manifestation, thereby mimicking intervention with a targeted therapeutic.

Despite the power of RNAi technology, technical challenges remain. Unlike conventional or conditional gene deletion, RNAi-based loss-of-function studies rely on strong shRNA expression throughout the experiment. In viral vector systems, an unfavorable proviral integration site can prevent shRNA expression through promoter interference, epigenetic silencing or other inhibitory effects[13–16]. Additionally, genomic instability can result in random deletion of proviral transgenes. Although such events might be rare, their collective impact is most significant when an shRNA targets a gene essential for cell survival or proliferation. Here, even a few cells that fail to express the shRNA will outcompete those in which RNAi is effective and thereby mask the phenotype of gene knockdown.

To address these limitations, we designed an inducible shRNA expression system that enables precise tracking of retroviral transduction and shRNA induction through two fluorescent reporters. TRE-dsRed-miR30/shRNA-PGK-Venus-IRES-NeoR (TRMPV) produces two transcripts (**Fig. 1a**): when active, the TRE drives expression of a dsRed fluorescent protein and a microRNA (miR)-embedded shRNA[6], whereas the phosphoglycerate kinase (*PGK*) promoter drives constitutive expression of both the yellow-green fluorescent protein Venus and, using an internal ribosomal entry site (IRES), the neomycin resistance gene (NeoR). Available variants of TRMPV feature alternative drug selection markers, alternative fluorescent proteins and/or an improved TRE with reduced basal activity ($\text{TRE}^{\text{tight}}$)[17] (**Supplementary Fig. 1**).

To evaluate the utility of this vector for studying genes required for proliferation, we chose to target Replication Protein A, subunit 3 (Rpa3), an essential factor for DNA replication whose knockdown causes cell cycle arrest in dividing cells[18]. As a neutral control, we used an shRNA targeting *Renilla* luciferase (shRen, **Supplementary Fig. 2a**,**b**). In immortalized Rosa26-rtTA-M2 (ref. 19) mouse embryonic fibroblasts (MEFs), TRMPV vectors containing Rpa3 shRNAs effectively knocked down Rpa3 in the presence of doxycycline (**Fig. 1b** and **Supplementary Fig. 2c**,**d**). As expected, transduced cells constitutively expressed only Venus, whereas doxycycline treatment caused strong, reversible induction of TRE-driven dsRed in most cells (**Fig. 1c** and **Supplementary Fig. 2e**). Therefore, shRNA-expressing cells can be identified as a double positive ($\text{Venus}^+\text{dsRed}^+$) population.

To determine whether TRMPV could track the depletion of cells expressing antiproliferative shRNAs within a population, we performed a competitive proliferation assay by mixing untransduced and TRMPV-transduced cells and quantifying the percentage of each population over time. As expected, $\text{Venus}^+\text{dsRed}^+$ cells were maintained in control shRen (**Fig. 1c**) but depleted in shRpa3 cultures on doxycycline,

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA. [2]Watson School of Biological Sciences, Cold Spring Harbor, New York, USA. [3]Howard Hughes Medical Institute, Cold Spring Harbor, New York, USA. [4]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland. [5]These authors contributed equally to this work. Correspondence should be addressed to S.W.L. (lowe@cshl.edu).

Received 1 July; accepted 3 November; published online 5 December 2010; doi:10.1038/nbt.1720

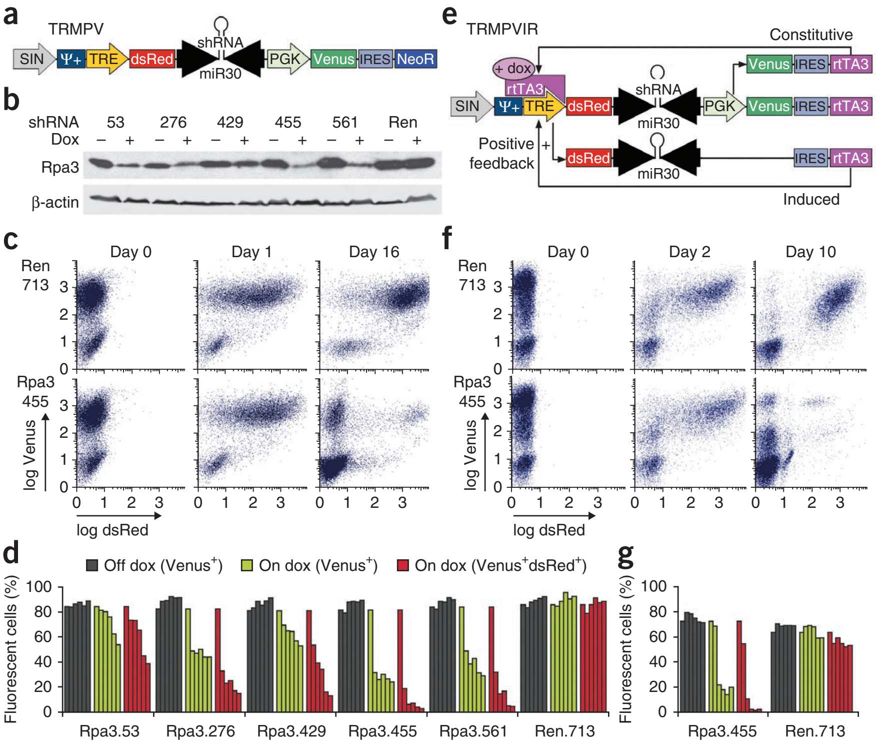


Figure 1 Dual-color TRMPV vectors enable Tet-regulated shRNA expression for suppression of genes involved in cell proliferation and survival. (**a**) Vector schematic of TRMPV, which was constructed in the pQCXIX self-inactivating (SIN) retroviral backbone. Ψ+, extended retroviral packaging signal. (**b**) Western blot of immortalized Rosa26-rtTA-M2 MEFs transduced with TRMPV harboring different Rpa3 shRNAs (numbered by start position) or a *Renilla* luciferase (Ren) shRNA. After selection, cells were cultured in the presence or absence of doxycycline for 4 d before collection. β-actin served as loading control. Uncropped blots are shown in **Supplementary Figure 2d**. (**c**) Representative flow cytometry plots of Rosa26-rtTA-M2 MEFs transduced with TRMPV.shRen.713 and TRMPV.shRpa3.455 in competitive proliferation assays. Cells selected for TRMPV were mixed with untransduced cells and passaged in doxycycline for 16 d. (**d**) Quantification of fluorescent cells in representative competitive proliferation assays. Each series of bars is a time course from left to right: day 0, 4, 8, 12, 16, 20. In all series, day 0 bar represents percentage Venus-positive cells before doxycycline treatment. In the presence of doxycycline, transduced cells were gated either on Venus or on both Venus and dsRed. (**e**) Vector schematic of TRMPVIR showing constitutive and inducible transcripts produced by the vector. IRES-dependent rtTA3 expression from the inducible transcript creates a positive feedback loop of TRE induction. (**f**) Representative flow cytometry plots of *Eμ-myc;Trp53*$^{-/-}$ lymphoma cells transduced with TRMPVIR.shRen.713 and TRMPVIR.shRpa3.455 in competitive proliferation assays over 10 d. Cells were incompletely transduced with TRMPVIR before day 0 (rather than selected and admixed with untransduced cells). (**g**) Quantification of fluorescent cells in representative competitive proliferation assays. Each series of bars is a time course from left to right: day 0, 2, 4, 6, 8, 10; see **d** for details.

as untransduced Venus$^-$dsRed$^-$ cells accumulated. A population of Venus$^+$dsRed$^-$ cells also expanded in shRpa3 cultures, reflecting the accumulation of clones that fail to induce TRE expression owing to preexisting positional effects or an active silencing process. Regardless of the cause, these clones lack shRNA induction, allowing them to evade cell cycle arrest in the presence of doxycycline.

This phenomenon, which we have observed in a wide variety of rtTA-expressing cell types, highlights the value of linking a fluorescent reporter tightly to shRNA expression. Accordingly, when we quantified shRpa3-transduced cells over time using only the constitutive Venus reporter, we observed a moderate depletion (**Fig. 1d**). In contrast, when quantifying Venus$^+$dsRed$^+$ cells, we observed a much more substantial depletion, which was greatest for the most potent shRNAs (shRpa3.455 and shRpa3.561) (**Fig. 1b**,**d** and **Supplementary Fig. 2c**). Thus, TRMPV provides a sensitive tool to assess inhibitory effects of shRNA expression on cell proliferation.

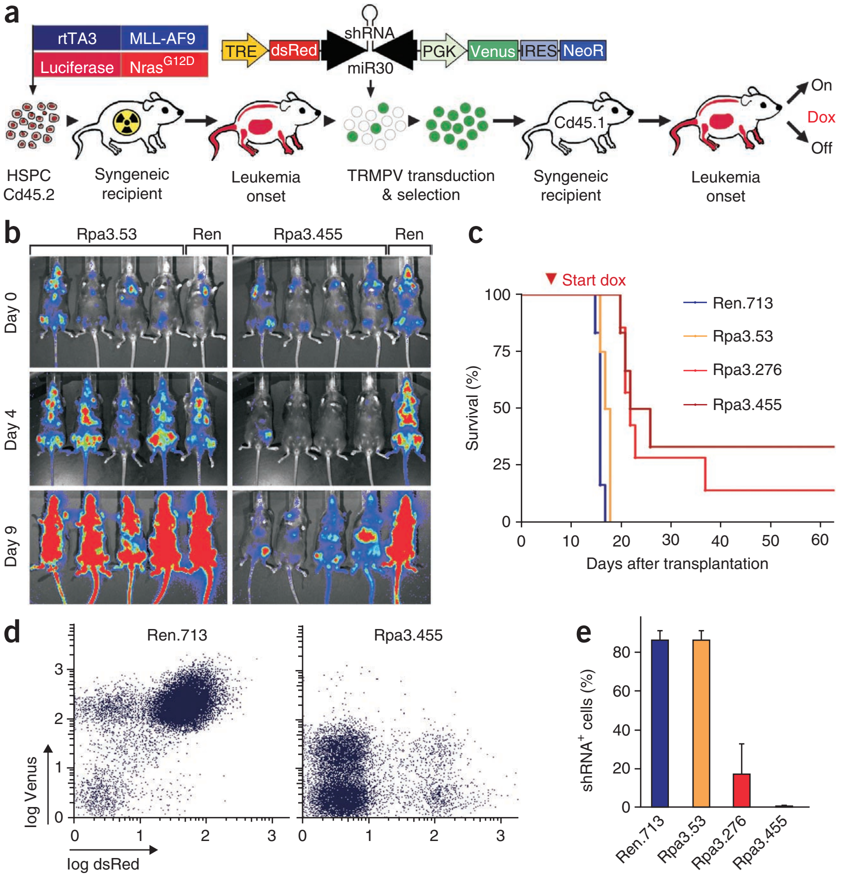
Next, we aimed to use TRMPV to assess antiproliferative phenotypes in cancer cells, a dynamic setting where the potential outgrowth of clones that escape shRNA induction presents a major concern. In the two-component Tet-On system, escape can occur either by an effect on the TRE locus itself or by insufficient rtTA expression. To minimize rtTA-related failure, we conceived two strategies to ensure strong, sustained rtTA expression in every cell. First, for mouse models driven by oncogenic transgenes, we hypothesized that linking the rtTA to the oncogene in a bicistronic transcript would select for strong rtTA expression based on the proliferative advantage conferred by the oncogene. Furthermore, 'oncogene addiction' (reviewed in ref. 20) would ensure maintained rtTA expression throughout the experiment.

We evaluated this strategy in a mouse model of acute myeloid leukemia (AML) induced by coexpression of a human mixed-lineage leukemia fusion gene (MLL-AF9; AF9 also known as MLLT3) and mouse Nras^{G12D}. To generate an oncogene-linked 'Tet-On-competent' model, we designed a retrovirus expressing MLL-AF9 in a bicistronic transcript downstream of an optimized rtTA (rtTA3)[21], whereas Nras^{G12D} was coexpressed with firefly luciferase (Luci) to enable disease monitoring by bioluminescence imaging. Hematopoietic stem and progenitor cells were cotransduced with rtTA3-IRES-MLL-AF9 and Luci-IRES-Nras^{G12D} and transplanted into recipient mice, which developed aggressive AML (**Supplementary Fig. 3**). Leukemia cells were collected from moribund mice, transduced with TRMPV vectors and analyzed in competitive proliferation assays. As in MEFs, shRpa3 induction led to efficient depletion of Venus$^+$dsRed$^+$ cells and an outgrowth of untransduced and Venus$^+$dsRed$^-$ cells (**Supplementary Fig. 4**). Therefore, oncogene-linked rtTA vectors can generate genetically defined mouse models suitable for characterizing genes required for proliferation or survival.

We also developed a strategy to enforce rtTA expression in systems where oncogene linkage is not an option—for example, in noncancer settings or established human cancer cell lines. We designed a dual-color construct that delivers both a TRE-driven shRNA and rtTA in a single retroviral vector (TRE-dsRed-miR30/shRNA-PGK-Venus-IRES-rtTA3, or TRMPVIR; **Fig. 1e**). In this construct, the *PGK* promoter drives constitutive expression of Venus and basal levels of rtTA. In the presence of doxycycline, the activated TRE promoter induces strong expression of a transcript containing the dsRed-shRNA cassette and, further downstream, rtTA3, whose translation is initiated through the IRES. Thus, in the induced state, this configuration is predicted to generate a positive feedback loop that produces more rtTA from the TRE transcript, ensuring robust activity of the Tet-On system.

To test this vector, we transduced an established mouse Eμ-myc;Trp53$^{-/-}$ lymphoma cell line with TRMPVIR-shRpa3 or shRen. Doxycycline

Figure 2 TRMPV enables RNAi-based evaluation of genes involved in tumor maintenance *in vivo*. (**a**) Schematic of the generation and application of a Tet-On-competent mouse model of leukemia. Hematopoietic stem and progenitor cells (HSPC, Cd45.2$^+$) are transduced with retroviruses that coexpress oncogenes, rtTA3 and firefly luciferase and subsequently transplanted into recipient mice. Resulting Tet-On leukemias are collected, transduced with TRMPV, selected and retransplanted into secondary Cd45.1$^+$ recipients. After leukemia onset, shRNAs are induced by doxycycline and effects analyzed using different readouts. (**b**) Representative bioluminescence imaging of recipient mice transplanted with TRMPV-transduced AML cells. Mice were treated with doxycycline at disease onset (day 0). (**c**) Kaplan-Meier survival curve of recipient mice of AML cells transduced with indicated TRMPV shRNAs. Mice were treated with doxycycline at disease onset as assayed by imaging (7 d after transplantation). (**d**) Representative flow cytometry plots of donor-derived (Cd45.2$^+$) cells in bone marrow of moribund doxycycline-treated mice in **c**. (**e**) Quantification of Venus$^+$dsRed$^+$ cells in Cd45.2$^+$ bone marrow cells collected from doxycycline-treated recipient mice ($n = 4$) at a terminal disease stage. Mean and s.e.m. are plotted.



treatment led to strong induction of dsRed expression, and Venus$^+$dsRed$^+$ shRpa3-expressing cells were outcompeted by untransduced and Venus$^+$dsRed$^-$ cells, whereas shRen-expressing cells were maintained over time (**Fig. 1f**). Again, quantification of Venus$^+$dsRed$^+$ cells provided a more sensitive assessment of deleterious shRNA effects than Venus alone (**Fig. 1g**). In summary, both the TRMPVIR vector and oncogene linkage of rtTA are effective strategies to provide robust rtTA expression and thereby facilitate the evaluation of Tet-regulated shRNAs.

To determine whether TRMPV could be used to characterize genes required for tumor maintenance *in vivo*, we transduced Tet-On-competent MLL-AF9;Nras^{G12D} AML cells with shRen or Rpa3 shRNAs of three different potencies. Drug-selected cells were transplanted into B6.SJL (Cd45.1$^+$) recipient mice (**Fig. 2a**), which allowed transplant tracing using Cd45.2-specific antibodies. Mice were monitored for leukemia by bioluminescence imaging (**Fig. 2b**) and treated with doxycycline upon leukemia onset. Mice succumbed to disease in ~17 d when shRen or the weak shRpa3.53 was expressed, whereas the intermediate shRpa3.276 or strong shRpa3.455 delayed disease progression and conferred a significant survival advantage (**Fig. 2b**,**c**). Nevertheless, even mice that initially responded eventually relapsed and succumbed to disease.

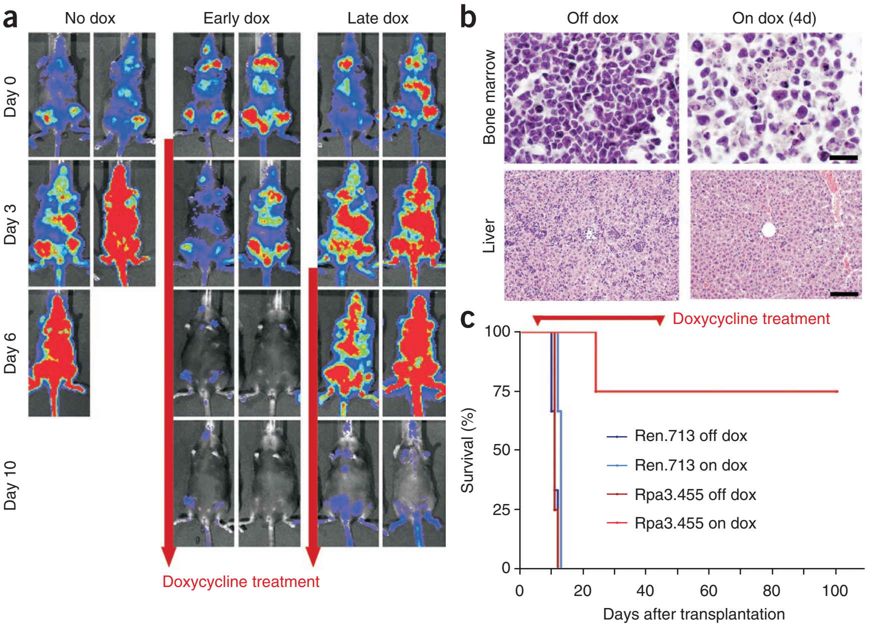
At terminal disease stage, we analyzed Cd45.2$^+$ leukemia infiltrates in bone marrow by flow cytometry. All TRMPV-shRen leukemias were strongly positive for both dsRed and Venus (**Fig. 2d**,**e**). Notably, all relapses of doxycycline-treated leukemias harboring shRpa3.276 or shRpa3.455 showed a strong depletion of dsRed-positive cells and an outgrowth of Venus$^+$dsRed$^-$ or double-negative clones that had evaded shRNA expression, presumably owing to silencing or loss of the TRE cassette or the entire provirus, respectively (**Fig. 2d**,**e**). Expression of the weak shRpa3.53, which caused depletion *in vitro* (**Fig. 1d**), had no significant effect on the frequency of Venus$^+$dsRed$^+$ cells *in vivo*, suggesting that this setting requires more potent shRNAs to detect the effects of Rpa3 suppression. Overall, these results illustrate how the ability to quantify shRNA-expressing cells at an advanced disease stage provides a rapid and sensitive strategy to detect inhibitory RNAi-mediated phenotypes *in vivo*.

We hypothesized that TRMPV would help to identify and isolate clones that homogeneously induce TRE expression. Clonal TRMPV-transduced MLL-AF9;Nras^{G12D} populations each showed homogeneous levels of constitutive Venus expression (**Supplementary Fig. 5a**). After doxycycline treatment, the levels of dsRed fluorescence induced varied substantially between clones, with some showing no induction (**Supplementary Fig. 5b**). Hence, TRMPV facilitates the selection of pure clonal populations capable of strong shRNA induction.

We transplanted selected clonal TRMPV-shRpa3.455 leukemias into recipient mice, which were either left untreated, treated with doxycycline upon disease onset or treated at an advanced disease stage (**Fig. 3a**). After early or late doxycycline treatment, shRpa3 induction caused rapid disease regression, even in mice showing wasting from advanced leukemia. Histology showed clearance of leukemic blasts from bone marrow, spleen and liver within 4 d (**Fig. 3b** and **Supplementary Fig. 5c**). All doxycycline-treated TRMPV-shRpa3 mice achieved complete disease remission, whereas untreated TRMPV-shRpa3 mice and recipients of clonal TRMPV-shRen control leukemias (treated or untreated) died rapidly (**Fig. 3c**). When relapse occurred in doxycycline-treated TRMPV-shRpa3 mice, imaging revealed that the disease typically expanded from a focal bone marrow infiltrate, suggesting the growth of a resistant cell clone. Indeed, all relapse leukemias were dsRed-negative (data not shown). Notably, 75% of doxycycline-treated TRMPV-shRpa3 mice remained healthy with no detectable luciferase signal during 22 weeks of follow-up—even when doxycycline was withdrawn after 40 d of treatment (**Fig. 3c**). Thus, we have established a robust system to identify genes that are essential for the maintenance and progression of cancers in mice.

We reasoned that the features of TRMPV might also facilitate multiplexed RNAi screening to identify genes required for disease

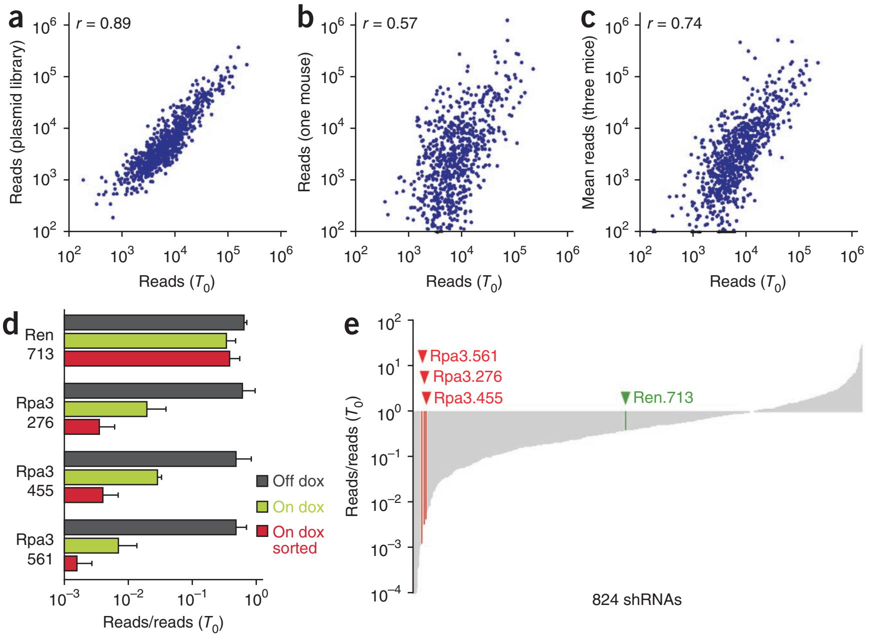
Figure 3 TRMPV-induced suppression of Rpa3 cures clonal MLL-AF9;Nras^{G12D} AML. (**a**) Bioluminescence imaging of recipient mice of clonal MLL-AF9;Nras^{G12D} AML harboring TRMPV.shRpa3.455. After leukemia onset as assayed by imaging (day 0) mice were either left untreated (no dox), treated with doxycycline (early dox) or treated at a more advanced disease stage (late dox). (**b**) Bone marrow and liver histology of untreated and doxycycline-treated mice 4 d after leukemia onset. Scale bars: 20 μm for bone marrow, 100 μm for liver. (**c**) Kaplan-Meier survival curve of recipient mice of clonal TRMPV.shRpa3.455 or TRMPV.shRen.713 leukemias. After disease onset as assayed by bioluminescent imaging (day 7 after transplantation), mice were either left untreated (off dox) or treated with doxycycline for 40 d (on dox).



maintenance. In these approaches, integrated proviruses serve as sequence tags to identify shRNAs that are specifically depleted (negatively selected) from a pooled library. We spiked the three most potent Rpa3 shRNAs and neutral shRen into a library of 820 TRMPV shRNAs (**Supplementary Table 1**). This pool was transduced into Tet-On-competent MLL-AF9;Nras^{G12D} AML cells under conditions that predominantly result in a single retroviral integration per cell (**Supplementary Fig. 6**), and transduced cells were subsequently drug selected. To read out shRNA representation, shRNA cassettes were amplified from genomic DNA using hybrid primers tagged with Illumina adapters and quantified by deep sequencing. shRNA representation in libraries generated directly after drug selection (T_0) strongly correlated with the original plasmid pool, suggesting that retroviral transduction and drug selection did not affect the library composition (**Fig. 4a**).

To test the utility of TRMPV for negative selection screens *in vivo*, 4 × 10^6 Venus$^+$ T_0 cells were transplanted into recipient mice, which were either left untreated or treated with doxycycline starting 4 d after transplantation. AML cells were collected from moribund leukemic mice 14 d after injection. In each individual untreated mouse, >95% of the shRNAs in the pool were detected by deep sequencing, and the distribution of normalized sequence reads per shRNA correlated with the preinjection (T_0) population (**Fig. 4b** and **Supplementary Fig. 7a**). We reasoned that deviations from the input representation may arise from leaky shRNA effects and/or stochastic changes in cell representation during disease engraftment and progression. Consistent with the latter possibility, the correlation between T_0 and untreated leukemia samples was substantially improved by averaging read numbers from three replicate mice (**Fig. 4c**, r = 0.74). Thus, the representation of a complex shRNA library can be maintained *in vivo,* and nonspecific changes in shRNA representation during disease progression can be deconvoluted by comparison of replicate mice.

Figure 4 Pooled negative selection RNAi screening *in vivo* detects shRpa3 depletion in MLL-AF9;Nras^{G12D} AML. (**a**) Scatter plot illustrating the correlation of normalized reads per shRNA between the plasmid library and transduced selected leukemia cells before transplantation (T_0); r, nonparametric (Spearman) correlation coefficient. (**b**) Scatter plot of normalized reads per shRNA in T_0 cells compared to an untreated leukemic recipient mouse. (**c**) Scatter plot of normalized reads per shRNA in T_0 cells compared to average reads in three untreated recipient mice. (**d**) Relative abundance of Rpa3 and *Renilla* luciferase shRNAs in leukemias isolated from untreated (off dox) and doxycycline-treated (on dox) recipient mice, each compared to the initial representation before transplantation (T_0). Leukemias from doxycycline-treated mice were analyzed both without and with purification of shRNA-expressing cells (Venus$^+$dsRed$^+$) before DNA isolation. (**e**) Relative abundance of all 824 shRNAs in Venus$^+$dsRed$^+$-sorted leukemia cells from doxycycline-treated mice compared to T_0 cells. The mean of normalized reads in doxycycline-treated mice (n = 3) was divided by normalized reads in T_0 cells; shRNAs are plotted according to the resulting ratios in ascending order. All three shRNAs targeting Rpa3 were among the 25 most depleted shRNAs, whereas neutral shRen.713 was not altered.



In doxycycline-treated mice, cells harboring antiproliferative shRNAs might evade shRNA expression and persist in the population, and shRNA cassettes amplified from these cells could contaminate sequencing libraries and dampen changes in representation. Indeed, we observed a substantial fraction of Venus$^+$dsRed$^-$ cells in doxycycline-treated mice (**Supplementary Fig. 8**). To determine whether eliminating such noninducers would increase the sensitivity of our readout, we isolated Venus$^+$dsRed$^+$ cells before assessing shRNA representation. All Rpa3 shRNAs showed moderate depletion compared to T_0 values in libraries amplified from unsorted leukemia cells (ratios < 0.1, **Fig. 4d**), but the detectable depletion was increased by four- to tenfold in libraries from sorted Venus$^+$dsRed$^+$ cells (ratios < 0.01). All three Rpa3 shRNAs were among the top 25 depleted when comparing treated mice to either untreated mice or to T_0 cells (**Fig. 4e** and **Supplementary Fig. 7b**,**c**). These results provide evidence that pooled *in vivo* RNAi screens for genes involved in tumor maintenance are feasible. In contrast to a recent report using stably expressed shRNAs[22], our inducible system enables acute target inhibition after disease onset and thereby effectively models therapeutic intervention.

Here we describe a toolkit that facilitates the use of RNAi for studying genes involved in cell proliferation and survival—an experimental setting that is complicated by clonal selection against efficient RNAi. By coupling shRNAs directly to a fluorescent reporter and using robust transactivator systems, we produced the TRMPV system, which enables the identification, tracking and isolation of only those cells that productively express an shRNA. We show that this feature facilitates the study of tumor maintenance genes and increases the sensitivity of negative selection RNAi screens. Indeed, by combining TRMPV-based cell sorting with deep sequencing as a readout, we noted >250-fold depletion for all three Rpa3 shRNAs—a substantial increase in sensitivity over ratios observed for negatively selected shRNAs using microarray-based platforms (two- to tenfold)[9,23,24]. A second source of false negatives in RNAi screens arises from the prevalence of nonfunctional shRNAs in libraries now available[25]. A recently developed method to identify potent shRNAs in a massively parallel assay (C.F., J.Z., G.J.H. & S.W.L. *et al.*, unpublished data) enables the production of prevalidated shRNA libraries that, when expressed from TRMPV, will represent a widely applicable resource.

Our methods provide a rapid pipeline for thorough validation and characterization of genes encoding putative drug targets. Candidate shRNAs identified by high-throughput screening can be quickly tested individually for deleterious effects in TRMPV bulk assays. The most promising targets can be further evaluated using our clonal assay for their role in tumor maintenance. Finally, inducible miR30-based shRNAs linked to fluorescent reporters have been implemented in germline transgenic mice (P. Premsrirut, L.E.D., C. Miething, K.M., S.W.L. *et al.*, unpublished data), enabling the study of target inhibition in both diseased and normal tissues. Overall, these technologies delineate a rational, inexpensive workflow to rigorously identify and characterize new therapeutic targets.

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 R. Dickins and C. Miething for discussions in setting up this system, as well as A. Lujambio, A. Rappaport, M. Saborowski and C. Vakoc for testing TRMPV in other models. We also thank Y. Dou (Univ. of Michigan) for providing human MLL-AF9. We gratefully acknowledge B. Ma and S. Muller for excellent technical assistance. We also thank E. Hodges, K. Chang, M. Rooks and the McCombie laboratory for help with Solexa sequencing as well as A. Gordon for bioinformatics support. P. Moody and T. Spencer were of great assistance in flow cytometry. Finally, we thank S. Kogan for histology. This work was supported by the Howard Hughes Medical Institute, the Starr Foundation and the Don Monti Memorial Research Foundation. J.Z. is the Andrew Seligson Memorial Fellow at Cold Spring Harbor Laboratory, and K.M. is the Robert and Teresa Lindsay Fellow of the Watson School of Biological Sciences. L.E.D. is supported by an Overseas Biomedical Research Fellowship of the National Health and Medical Research Council of Australia.

AUTHOR CONTRIBUTIONS

J.Z. and K.M. designed and performed experiments. C.F. and L.E.D. contributed new reagents and performed experiments. M.J.T. managed mouse monitoring and husbandry. G.J.H. and S.W.L. supervised this project. J.Z., K.M. and S.W.L. wrote 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. Martin, S.E. & Caplen, N.J. Applications of RNA interference in mammalian systems. *Annu. Rev. Genomics Hum. Genet.* **8**, 81–108 (2007).
2. Hannon, G.J. RNA interference. *Nature* **418**, 244–251 (2002).
3. Brummelkamp, T.R., Bernards, R. & Agami, R. A system for stable expression of short interfering RNAs in mammalian cells. *Science* **296**, 550–553 (2002).
4. Paddison, P.J., Caudy, A.A., Bernstein, E., Hannon, G.J. & Conklin, D.S. Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells. *Genes Dev.* **16**, 948–958 (2002).
5. Hemann, M.T. *et al.* An epi-allelic series of p53 hypomorphs created by stable RNAi produces distinct tumor phenotypes *in vivo. Nat. Genet.* **33**, 396–400 (2003).
6. Dickins, R.A. *et al.* Probing tumor phenotypes using stable and regulated synthetic microRNA precursors. *Nat. Genet.* **37**, 1289–1295 (2005).
7. Barbie, D.A. *et al.* Systematic RNA interference reveals that oncogenic *KRAS*-driven cancers require TBK1. *Nature* **462**, 108–112 (2009).
8. Luo, J. *et al.* A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. *Cell* **137**, 835–848 (2009).
9. Ngo, V.N. *et al.* A loss-of-function RNA interference screen for molecular targets in cancer. *Nature* **441**, 106–110 (2006).
10. Scholl, C. *et al.* Synthetic lethal interaction between oncogenic KRAS dependency and STK33 suppression in human cancer cells. *Cell* **137**, 821–834 (2009).
11. Stegmeier, F., Hu, G., Rickles, R.J., Hannon, G.J. & Elledge, S.J. A lentiviral microRNA-based system for single-copy polymerase II-regulated RNA interference in mammalian cells. *Proc. Natl. Acad. Sci. USA* **102**, 13212–13217 (2005).
12. Gossen, M. *et al.* Transcriptional activation by tetracyclines in mammalian cells. *Science* **268**, 1766–1769 (1995).
13. Lund, A.H., Duch, M. & Pedersen, F.S. Transcriptional silencing of retroviral vectors. *J. Biomed. Sci.* **3**, 365–378 (1996).
14. Ellis, J., Hotta, A. & Rastegar, M. Retrovirus silencing by an epigenetic TRIM. *Cell* **131**, 13–14 (2007).
15. Markstein, M., Pitsouli, C., Villalta, C., Celniker, S.E. & Perrimon, N. Exploiting position effects and the gypsy retrovirus insulator to engineer precisely expressed transgenes. *Nat. Genet.* **40**, 476–483 (2008).
16. Pikaart, M.J., Recillas-Targa, F. & Felsenfeld, G. Loss of transcriptional activity of a transgene is accompanied by DNA methylation and histone deacetylation and is prevented by insulators. *Genes Dev.* **12**, 2852–2862 (1998).
17. Agha-Mohammadi, S. *et al.* Second-generation tetracycline-regulatable promoter: repositioned tet operator elements optimize transactivator synergy while shorter minimal promoter offers tight basal leakiness. *J. Gene Med.* **6**, 817–828 (2004).
18. Wold, M.S. Replication protein A: a heterotrimeric, single-stranded DNA-binding protein required for eukaryotic DNA metabolism. *Annu. Rev. Biochem.* **66**, 61–92 (1997).
19. Hochedlinger, K., Yamada, Y., Beard, C. & Jaenisch, R. Ectopic expression of Oct-4 blocks progenitor-cell differentiation and causes dysplasia in epithelial tissues. *Cell* **121**, 465–477 (2005).
20. Weinstein, I.B. Cancer. Addiction to oncogenes–the Achilles heal of cancer. *Science* **297**, 63–64 (2002).
21. Das, A.T. *et al.* Viral evolution as a tool to improve the tetracycline-regulated gene expression system. *J. Biol. Chem.* **279**, 18776–18782 (2004).
22. Meacham, C.E., Ho, E.E., Dubrovsky, E., Gertler, F.B. & Hemann, M.T. *In vivo* RNAi screening identifies regulators of actin dynamics as key determinants of lymphoma progression. *Nat. Genet.* **41**, 1133–1137 (2009).
23. Silva, J.M. *et al.* Profiling essential genes in human mammary cells by multiplex RNAi screening. *Science* **319**, 617–620 (2008).
24. Schlabach, M.R. *et al.* Cancer proliferation gene discovery through functional genomics. *Science* **319**, 620–624 (2008).
25. Bassik, M.C. *et al.* Rapid creation and quantitative monitoring of high coverage shRNA libraries. *Nat. Methods* **6**, 443–445 (2009).

ONLINE METHODS

Retroviral vectors and shRNAs. TRMPV (TRE-dsRed-miR30/shRNA-PGK-Venus-IRES-NeoR) and its variants (**Supplementary Fig. 1**) were constructed based on pSIN-TRE-PIG[6] in the pQCXIX self-inactivating retroviral backbone (Clontech). Using standard cloning techniques, we inserted dsRed (from pDsRed2, Clontech) and the miR30 context downstream of the TRE promoter and replaced PGK-PuroR-IRES-GFP stepwise with a PGK-Venus-IRES-NeoR cassette using fragments from MSCVneo (Clontech) and pSLIK[26]. In variants, we changed TRE to TRE-tight (from pTRE-tight, Clontech), dsRed to Turbo-RFP (from TRIPZ, Open Biosystems) and NeoR to HygroR (from pMSCVhyg, Clontech, after destroying the EcoRI site in HygroR using QuickChange mutagenesis, Stratagene). To construct TRMPVIR, we replaced NeoR by rtTA3 (ref. 21) (amplified from pSLIK[26]). MSCV-rtTA3-IRES-MLL-AF9 was constructed by sequentially cloning a human *MLL-AF9* cDNA (provided by Yali Dou, University of Michigan) and an rtTA3-IRES cassette into an empty MSCV backbone (Clontech). Placing rtTA upstream of the IRES generally produced more robust rtTA expression that was less affected by the nature of the coexpressed oncogene than in the reverse orientation. Luci-IRES-Nras^{G12D} has been described previously[27]. Detailed cloning strategies and primer sequences are available on request. All vectors are available upon request; vector sequences can be obtained through GenBank (accession numbers in **Supplementary Fig. 1**).

miR30-shRNAs were designed by adapting BIOPRED*si*[28] small interfering RNA predictions, except for Rpa3.455 (clone V2MM_14728, Open Biosystems) and for Rpa3.561, which was identified using a sensor assay for high-throughput shRNA evaluation (C.F., J.Z., G.J.H., S.W.L. *et al.*, unpublished data). shRNAs were designated by the number of the first nucleotide of the 22-nt target site in the mRNA transcript at the time of design; for shRNA sequences, see **Supplementary Table 2**. shRNAs were cloned into the miR30 context as 116-nt XhoI–EcoRI fragments, which were generated by amplifying 97-mer oligonucleotides (Sigma-Aldrich) using 5′miR30-XhoI (TACAATACTCGA GAAGGTATATTGCTGTTGACAGTGAGCG) and 3′miR30-EcoRI (ACTT AGAAGAATTCCGAGGCAGTAGGCA) primers and the Platinum Pfx kit (Invitrogen) with the following conditions: 50 μl reaction containing 0.05 ng oligonucleotide template, 1× Pfx buffer, 1 mM MgSO$_4$, 0.3 mM of each dNTP, 0.8 μM of each primer, and 1.25 U Pfx polymerase; cycling: 94 °C for 2 min; 33 cycles of 94 °C for 15 s, 54 °C for 30 s and 68 °C for 25 s; 68 °C for 5 min. A customized shRNA library targeting selected mouse genes was designed based on BIOPRED*si* predictions and generated by cloning a complex pool of oligonucleotides synthesized on 55k customized arrays (Agilent Technologies) followed by large-scale capillary sequencing of individual clones. A pool of 824 shRNAs was assembled by combining plasmid DNA of 820 sequence-verified clones and adding *Renilla* luciferase and three potent Rpa3 control shRNAs at equimolar concentrations.

Cell culture, retroviral transduction and competitive proliferation assays. Rosa26-rtTA-M2 MEFs were isolated from Rosa26-rtTA-M2 transgenic mice[19] and immortalized by infection with a lentivirus expressing a CMV-driven SV40 large T antigen. To ensure homogeneous presence of the transgenic PuroR-containing rtTA-M2 allele, MEFs were selected with puromycin (2.5 μg ml^{-1}, Sigma-Aldrich) before experimental use. MLL-AF9;Nras^{G12D} leukemia cells were cultured in RPMI 1640 (Gibco-Invitrogen) supplemented with 10% FBS, 100 U ml^{-1} penicillin and 100 μg ml^{-1} streptomycin at 37 °C with 7.5% CO$_2$. *Eμ-myc;Trp53*$^{-/-}$ lymphoma cells were cultured in 45% DMEM, 45% IMDM (Gibco-Invitrogen), 10% FBS, 4 mM L-glutamine, 50 μM β-mercaptoethanol, 100 U ml^{-1} penicillin and 100 μg ml^{-1} streptomycin at 37 °C with 7.5% CO$_2$. Retroviral constructs were transfected into HEK293T Phoenix packaging cells as previously described[29]; chloroquine (25 μM, Sigma-Aldrich) was added to enhance plasmid stability. Transductions of leukemia and lymphoma cells were performed in six-well plates. The medium was removed except for ~0.5 ml containing 0.5–1 × 10^6 suspension cells, and 2.5 ml fresh virus-containing supernatant supplemented with Polybrene (4 μg ml^{-1}, Sigma-Aldrich) was added; plates were then centrifuged for 25 min at 515*g*. Efficiency of retroviral transduction was assessed 48 h after infection by flow cytometry (EasyCyte, Guava Technologies). Drug selection of TRMPV-transduced MEFs and AML cells was conducted using 1 mg ml^{-1} G418 (Geneticin, Gibco-Invitrogen). To induce shRNA expression, doxycycline (Sigma-Aldrich) was added at final concentrations of 1 μg ml^{-1} for leukemia and lymphoma cells and 2 μg ml^{-1} for MEFs.

To analyze shRNA effects on proliferation and survival in bulk populations, TRMPV-transduced cells were admixed with untransduced cells and cultured in the absence or presence of doxycycline. Venus and dsRed fluorescence were quantified on an EasyCyte (Guava Technologies) or LSR-II (BD Biosciences) flow cytometer. Single-cell-derived populations of TRMPV-transduced AML cells were isolated by means of limiting dilution by plating six dilutions containing ~16, 8, 4, 2, 1 and 0.5 cells. After expansion under G418 selection, populations were analyzed by flow cytometry. Clonal populations showing highly homogeneous Venus levels were also tested for shRNA (dsRed) induction by doxycycline treatment and flow cytometry after 24–48 h.

Assessment of shRNA efficacy. Rpa3 knockdown was assessed in immortal Rosa26-rtTA-M2 MEFs that were transduced with approximately a single copy of TRMPV (<20% overall transduction). After selection in G418 (1 mg ml^{-1}, resulting in a >95% Venus$^+$ population), cells were maintained in medium containing puromycin (2.5 μg ml^{-1}, Sigma-Aldrich), to ensure homogenous presence of the transgenic PuroR-containing rtTA-M2 allele, and G418 (1 mg ml^{-1}) and were treated with or without doxycycline (2 μg ml^{-1}) for 4 d before lysis. For immunoblotting, samples were lysed in Laemmli buffer, separated by SDS-PAGE and transferred to PVDF membranes (Immobilon-P, Millipore), which were incubated with antibodies to Rpa3 (M-18, 1:100 in TBS with 0.05% Tween-20 and 0.1% Triton X-100; Santa Cruz Biotechnology) and β-actin (AC-15, 1:5,000; Sigma-Aldrich). Densitometry was performed using ImageJ (US National Institutes of Health, available at http://rsbweb.nih.gov/ij/). To assess *Renilla* luciferase knockdown, NIH3T3 mouse fibroblast cells were transduced with a retrovirus that expresses the luciferase (MSCV-Renilla-PGK-HygroR), selected with hygromycin B (200 μg ml^{-1}, Roche) and reinfected with MSCV/LTRmiR30-PIG (LMP)[6] containing control or luciferase-targeted shRNAs. After puromycin selection (2 μg ml^{-1}), luciferase activity was quantified in whole cell protein extracts using a *Renilla* luciferase assay (Promega). Luciferase activity (absolute light units) was normalized to total protein levels. To determine knockdown efficacy at single copy, shRen.713 was cloned into the pCol-TGM targeting vector (P. Premsrirut, L.E.D., C. Miething, K.M., S.W.L *et al.*, unpublished data) and electroporated into KH2 ES cells. After selection, individual clones were expanded, infected with MSCV-Renilla and treated with or without doxycycline for 4 d. Luciferase activity was determined as above and normalized to the off-doxycycline condition.

Animal studies. The Cold Spring Harbor Animal Care and Use Committee approved all mouse experiments included in this work. To generate Tet-On MLL-AF9;Nras^{G12D} leukemia, hematopoietic stem and progenitor cells were isolated from C57BL/6 fetal livers (embryonic days 13.5–15) and retrovirally transduced as previously described[27,30]. For leukemia transplantation, 1 × 10^6 cells were injected in the tail vein of sublethally irradiated recipient mice (5.5 Gy, 24 h before transplantation). Whole-body bioluminescent imaging was performed using an IVIS100 system (Caliper LifeSciences) as described[27]. For shRNA induction, mice were treated with doxycycline in both drinking water (2 mg ml^{-1} with 2% sucrose; Sigma-Aldrich) and food (625 mg kg^{-1}, Harlan Laboratories). Leukemic mice were put to death at terminal disease stage (whole body signal in bioluminescent imaging, severe leukocytosis in peripheral blood smears, moribund appearance) by CO$_2$. Leukemia cells were collected from bone marrow (by flushing tibias and femurs with DMEM) and spleen (by gently mashing enlarged spleens in DMEM between two glass slides) and filtered through 100-μm cell strainers (BD Falcon). Resulting single cell suspensions were cultured, frozen (in 90% FBS, 10% DMSO) or directly used in flow cytometry. For the latter, erythrocytes were lysed using 150 mM NH$_4$Cl, 10 mM KHCO$_3$, 0.1 mM EDTA, and cells were resuspended in PBS containing 5% FBS and 0.1% NaN$_3$. For immunophenotyping, we used FITC-, PE-Cy5- or Pacific Blue–conjugated antibodies specific for CD45.2 (Ly5.2), Mac-1, Gr-1, CD19, B220, CD3, TER119, c-Kit and Sca-1 (all BioLegend); data were collected on Guava EasyCyte (Guava Technologies) or LSR-II (BD Biosciences) flow cytometers.

Pooled negative-selection RNAi screening *in vivo*. Tet-On MLL-AF9;Nras^{G12D} AML cells were transduced with a pool of 824 TRMPV constructs using

doi:10.1038/nbt.1720

conditions that predominantly lead to a single retroviral integration per cell (4.5% transduction in 2.5 × 10^7 cells total; see **Supplementary Fig. 6**). To preserve the library representation, a minimum of 1 × 10^6 cells (>1,000 cells per shRNA) were infected and maintained throughout the experiment. After G418 selection, 4 × 10^6 cells were transplanted into recipient mice, which were left untreated or treated with doxycycline 4 d after transplantation, the time point at which leukemia cells can typically be detected in bone marrow flow cytometry in this model. At a moribund disease stage (~14 d after transplantation), AML cells were collected from bone marrow and spleen and mixed in a 1:1 ratio. For samples from doxycycline-treated mice, ~1.5 × 10^7 Venus$^+$dsRed$^+$ cells were isolated using a FACSAriaII flow cytometer (BD Biosciences). Genomic DNA was isolated by two rounds of phenol extraction using PhaseLock tubes (5prime) followed by isopropanol precipitation. Deep sequencing template libraries were generated by PCR amplification of shRNA guide strands using primers that tag the product with standard Solexa/Illumina adapters (p7+loop, CAAGCAGAAGACGGCATACGATAGTGAAGCCACAGATGTA; p5+miR3′, AATGATACGGCGACCACCGACTAAAGTAGCCCCTTGAATTC). For each sample, DNA from at least 5 × 10^6 cells was used as template in multiple parallel 50-μl PCR reactions, each containing 1 μg template, 1× AmpliTaq Gold buffer, 0.2 mM of each dNTP, 0.3 μM of each primer and 2.5 U AmpliTaq Gold (Applied Biosystems), which were run using the following cycling parameters: 95 °C for 10 min; 35 cycles of 95 °C for 20 s, 52 °C for 30 s and 72 °C for 25 s; 72 °C for 7 min. PCR products (117 nt) were combined for each sample, precipitated and purified on a 2% agarose gel (QIAquick gel extraction kit, Qiagen). Libraries were analyzed on an Illumina Genome Analyzer at a final concentration of 8 pM**;** 18 nt were sequenced using a primer that reads in reverse into the guide strand (miR30EcoRISeq, TAGCCCCTTGAATTCCGAG GCAGTAGGCA). To provide a sufficient baseline for detecting shRNA depletion in experimental samples, we aimed to acquire >1,000 reads per shRNA in the T_0 sample. In practice, this depth of coverage required ~10 million reads per sample to compensate for disparities in shRNA representation inherent in the pooled plasmid preparation or introduced by PCR biases. With these conditions, we acquired T_0 baselines of >1,000 reads for 796 (96.6% of all) shRNAs. Sequence processing was performed using a customized Galaxy platform[31]. For each shRNA and condition, the number of matching reads was normalized to the total read number per lane and imported into a database for further analysis (Access 2003, Microsoft).

26. Shin, K.J. *et al.* A single lentiviral vector platform for microRNA-based conditional RNA interference and coordinated transgene expression. *Proc. Natl. Acad. Sci. USA* **103**, 13759–13764 (2006).
27. Zuber, J. *et al.* Mouse models of human AML accurately predict chemotherapy response. *Genes Dev.* **23**, 877–889 (2009).
28. Huesken, D. *et al.* Design of a genome-wide siRNA library using an artificial neural network. *Nat. Biotechnol.* **23**, 995–1001 (2005).
29. McCurrach, M.E. & Lowe, S.W. Methods for studying pro- and antiapoptotic genes in nonimmortal cells. *Methods Cell Biol.* **66**, 197–227 (2001).
30. Schmitt, C.A. *et al.* Dissecting p53 tumor suppressor functions in vivo. *Cancer Cell* **1**, 289–298 (2002).
31. Taylor, J., Schenck, I., Blankenberg, D. & Nekrutenko, A. Using Galaxy to perform large-scale interactive data analyses. *Curr. Protoc. Bioinformatics* **10**, 10.15 (2007).

doi:10.1038/nbt.1720

Pay for executives at private life sciences companies continues to steadily increase

Bruce Rychlik

In 2010, compensation targets for top life sciences executives once again moved upwards at a faster rate than those of their technology-sector peers.

The latest edition of CompStudy (http://www.compstudy.com), a compensation study released by J. Robert Scott and Ernst & Young, revealed that total target cash compensation in 2010 at private life sciences firms was up 4.2% over 2009 levels (**Fig. 1**). This year's report is the 9th edition of the annual compensation benchmark, and contains results gleaned from over 1,000 executives at nearly 200 companies. Produced by executive search firm J. Robert Scott and global services leader Ernst & Young, in collaboration with Noam Wasserman at Harvard Business School, the 2010 CompStudy provides position-by-position compensation data for top executives at private life sciences companies (**Box 1** and **Fig. 2**).

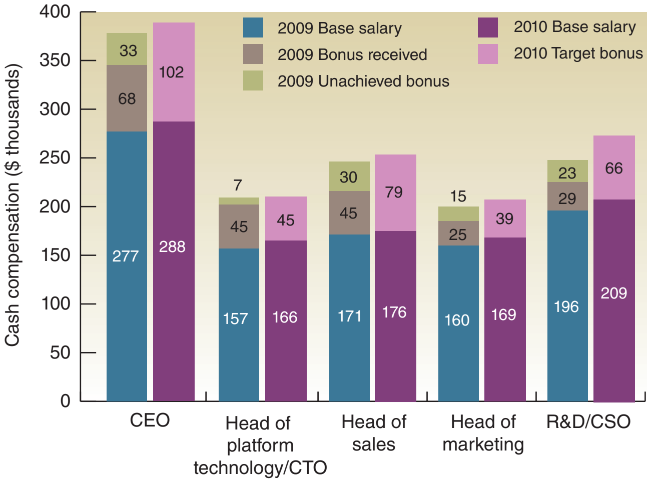


Figure 1 Cash compensation for life sciences nonfounders.

Historically, the average year-over-year change in nonfounder base salary has been a 5% increase. It is interesting to note that despite the continued turbulence in the broader economic climate, increases over the last three years have deviated from this number only slightly. The 2008 edition of the study reflected a 5.4% jump in base salaries, followed by increases of 3.2% in 2009 and 4.2% in 2010.

This contrasts with the data in the technology-sector edition of CompStudy. Although the average year-over-year increase for surveyed technology executives has also been right around 5%, there has been much more year-to-year variation. This has been particularly pronounced over the last few years; base salaries for nonfounder executives at private technology companies were flat from 2008 to 2009, and increased by 3.3% between 2009 and 2010.

This widening spread in rates of increase further widens the gap in pay between the two industries. As Wasserman mentioned during this year's webcast announcing the results, his analysis of historical CompStudy data (2002–2009) found that nonfounder CEOs at life sciences companies averaged 27% higher base salaries than their technology-sector peers. "This could be a simple function of supply and demand," commented Erik Lundh, managing director, life sciences, at J. Robert Scott's San Francisco office. "CEOs of life sciences companies have particular and highly sought-after knowledge and competencies that are fairly unique relative to their peers in other sectors."

The year-over-year change in pay levels, meanwhile, is probably more indicative of the way in which executive performance is measured. According to David Johnson, the service line leader for Ernst & Young's executive compensation advisory service, private life sciences businesses tend to be more research and development–oriented than companies in other industries. As a result, their compensation and performance metrics are based largely on product milestones, rather than revenue generation.

On the cash side of the equation, base salaries only tell part of the story; for senior executives, bonuses make up a large portion of their total annual compensation. In 2010, target bonuses as a percentage of base salary ranged from a low of 20.8% for general counsels to a high of 35.9% for heads of sales (**Fig. 1**). As a percentage of base salary, target bonuses have remained relatively steady throughout the history of the CompStudy. However, there has been significant variation in the amount actually paid out.

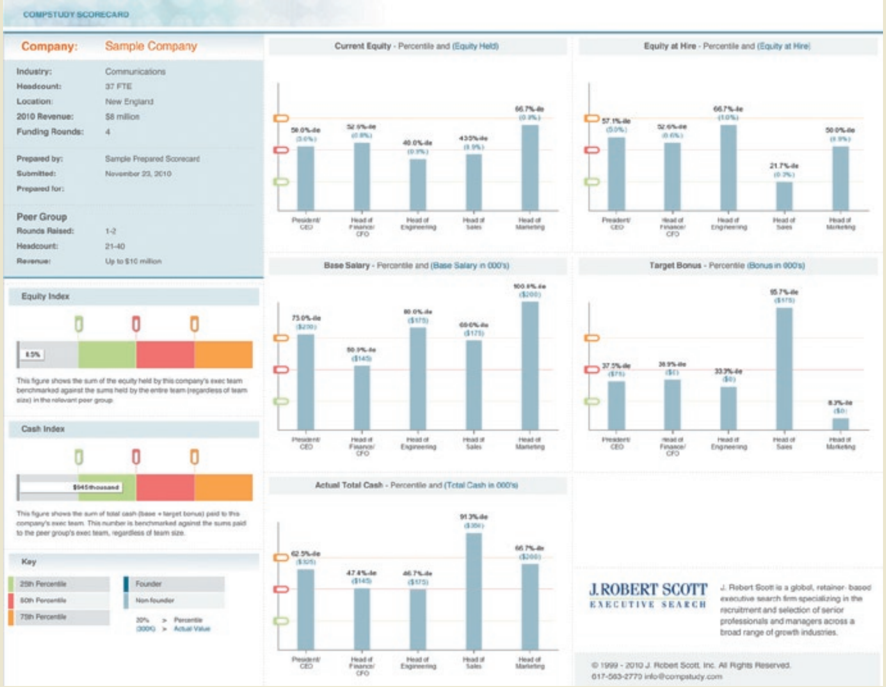
The bonus payout any one executive receives at year-end in life sciences is typically a function of both company and individual executive performance, but on an aggregate level bonus payouts are a good proxy for the health of an industry in a given year. The 2010 CompStudy

Bruce Rychlik is at J. Robert Scott, Boston, Massachusetts, USA.
e-mail: bruce.rychlik@fmr.com

Box 1 Methodology

For the 9th consecutive year executive search firm J. Robert Scott, in conjunction with global services professionals at Ernst & Young and Noam Wasserman of Harvard Business School, has released data on compensation trends for executives and board members at private, US life sciences companies. The 2010 CompStudy is a result of survey data collected from more than 1,000 executives at 185 private life sciences companies from across the United States in the following three industry segments: therapeutics, medical devices, and instrumentation and/or tools. The report presents the correlation between executive compensation and a number of variables including financing stage, company size (both in terms of product stage and headcount), founder/nonfounder status, industry segment and geography.

Last year, the CompStudy.com reporting platform was updated to include interactive, fully customizable charts. This year, a Company Scorecard feature has been added (**Fig. 2**). Available only to participants, each scorecard provides each company with a dashboard view of their executives' compensation, benchmarked against a peer group of their choice.



COMPSTUDY SCORECARD

Company: Sample Company

Industry: Communications
Headcount: 37 FTE
Location: New England
2010 Revenue: $8 million
Funding Rounds: 4

Prepared by: Sample Prepared Scorecard
Submitted: November 23, 2010
Prepared for:

Peer Group
Rounds Raised: 1-2
Headcount: 21-40
Revenue: Up to $10 million

Equity Index

This figure shows the sum of the equity held by this company's exec team benchmarked against the sums held by the entire team (regardless of team size) in the relevant peer group.

Cash Index

This figure shows the sum of total cash (base + target bonus) paid to this company's exec team. This number is benchmarked against the sums paid to the peer group's exec team, regardless of team size.

J. Robert Scott is a global, retainer- based executive search firm specializing in the recruitment and selection of senior professionals and managers across a broad range of growth industries.

© 1999 - 2010 J. Robert Scott, Inc. All Rights Reserved.
617-563-2770 info@compstudy.com

Figure 2 A sample CompStudy Company Scorecard.

Participants in the annual study receive free access to the reporting platform; for temporary access, please contact the author.

contains data on 2009 "achieved bonuses," and across all positions the average payout was 63% of target, up from 55% in last year's study. These numbers also compare favorably to the payouts received by executives at private technology companies, which averaged payouts of 57% in 2009 and 56% in 2010.

In addition to base and incentive pay, another important compensation component for CEOs and other corporate officers comes in the form of severance. This year's CompStudy data revealed that 57% of nonfounder CEOs had a prenegotiated severance package with a median length of 12 months. In every other executive role covered in the study, no more than one-third of executives had any guaranteed severance. As Lundh points out, because senior executives in the life sciences industry often have highly specialized skills, it can often take them longer to find their next ideal role. A prenegotiated severance package can therefore help to ensure that key executives are not constantly looking over their shoulders.

For life sciences companies, especially ones in formative stages, equity compensation holds great weight and is deemed important to the overall mix. Many of the executives surveyed by the CompStudy have taken a cash discount, relative to what their skill and experience might net them at a large pharmaceutical company. This early-stage company discount is offset, in part, by equity grants. Although founder equity stakes vary greatly and depend on several variables, nonfounder holdings have historically remained fixed. Median equity holdings for the nonfounder CEO at the earliest-stage companies are 4.6%; in those companies having raised four or more rounds of financing, median equity is 4.1% of the fully diluted shares.

As these data illustrate, a nonfounding CEO can expect her or his holdings to be marginally protected from dilution. Based on the data collected over the past decade, target values of fully diluted equity held by nonfounding CEOs hover around 5%. The method by which their equity stakes are grossed back up to the target level is also important. Over 40% of companies in this year's CompStudy reported using incentive stock options as equity instruments. Only 11.6% of companies reported granting shares of common stock outright. The mode of grant can become critical as a company nears a liquidity event, and as Ernst & Young's David Johnson points out, companies should "keep exit dynamics in mind when creating equity programs, so as to not get burned at exit."

For founding CEOs, there is little equity protection; median equity drops from 10.5% for CEOs at the earliest stages of financing to 5.0% with four or more rounds raised. These compensation differences for founding and nonfounding executives are of great interest to Noam Wasserman, professor at the Harvard Business School, who contributes to the design of the report each year. He addresses some of the current key issues surrounding private company founders and other "frustrations" that arise in creating and growing an enterprise on his site, http://founderresearch.blogspot.com.

In summary, although compensation at private life sciences companies is not completely uncorrelated with the state of the broader economy, it does appear less affected by fluctuations than pay at private technology companies. As a result, the compensation spread between the two growth industries has continued to widen over the last few years.

ACKNOWLEDGMENTS

The author would like to thank E. Brown and J. McDonough at J. Robert Scott for their contributions to the CompStudy report and this article.

COMPETING FINANCIAL INTERESTS

The author declares no competing financial interests.

PEOPLE

The Ontario Genomics Institute (Toronto) has announced the appointment of **Mark J. Poznansky** (left) as president and CEO, replacing **Christian Burks**. Poznansky has served on the OGI board of directors since 2004 and has been chairman since 2008. He was previously president, CEO and scientific director of the Robarts Research Institute in London, Ontario, and also served as president and CEO of Viron Therapeutics. For the past two years, he has managed his own consultancy for clients in government, hospitals, universities and the private sector.

"Dr. Poznansky brings a wealth of experience to OGI, in genomics, in business development and in government affairs," says Mark Lievonen, acting chair of the board of directors. "His knowledge of life sciences and in running institutes and businesses will serve OGI well as it continues its plans to aid the growth of the life sciences sector in Ontario and increase the province's reputation as a world leader in genomics research."

Advanced Cell Technology (Marlborough, MA, USA) has announced the unexpected passing of its chairman and CEO, **William M. Caldwell IV**, on December 13, 2010. Caldwell had been CEO since 2005 and chairman since 2006. **Gary Rabin**, a member of ACT's board since 2007 and managing partner of GR Advisors, has been named as interim chairman and CEO until a permanent replacement is named.

Former OSI Pharmaceuticals CEO **Colin Goddard** has been appointed to the board of directors of Human Genome Sciences (Rockville, MD, USA). Goddard joined OSI Pharmaceuticals as a scientist in 1989, advancing through its management ranks before being named CEO and a director in 1998 and chairman in 2000.

Receptos (San Diego) has named **Faheem Hasnain** (left) president, CEO and a member of the board of directors. Hasnain was most recently president and CEO of Facet Biotech. **William H. Rastetter** has stepped down as interim CEO of Receptos and will continue to serve as chairman of the board.

Pulmatrix (Cambridge, MA, USA) has named **Eva Jack** chief business officer, a newly created position. She most recently served as managing director at MedImmune Ventures, the corporate venture fund of MedImmune, which included serving as a board member or board observer for several portfolio companies. Previously, she was director of MedImmune's business development group.

John H. Johnson has been named chairman of the board of directors of Tranzyme Pharma (Research Triangle Park, NC, USA). He currently serves as president of Lilly Oncology and senior vice president at Eli Lilly. Previously, Johnson served as CEO of ImClone Systems and was also a member of ImClone's board until the company became a wholly owned subsidiary of Lilly in November 2008.

Jeffrey Kindler has resigned as chairman and CEO of Pfizer (New York), a little over a year after the company's $67 billion acquisition of Wyeth. He assumed the CEO post in July 2006. "The combination of meeting the requirements of our many stakeholders around the world and the 24/7 nature of my responsibilities has made this period extremely demanding on me personally," Kindler said in a statement, citing the need to "recharge [his] batteries." He is succeeded as chairman by independent director **George Lorch** and as CEO by Pfizer's global head of pharmaceuticals, **Ian Read**.

Originally slated to step down from his post on December 17, 2010, **Robert Klein** has agreed to remain as head of the California Institute for Regenerative Medicine (San Francisco) until a suitable replacement is found. State officials discovered that the leading candidate, Alan Bernstein, was ineligible because he was not a US citizen. Bernstein, a Canadian national, is executive director of the Global HIV Vaccine Enterprise. Klein says he anticipates leading the agency for no more than 12 months: "Within the year, and hopefully sooner than that, we'll be able to find someone with Dr. Bernstein's tremendous talent who is a US citizen who can fulfill this role."

Privately held contract research organization WIL Research (Fairfield, NJ, USA) has appointed **Howard Moody** as chief information officer and **Alan Findlater** as chief commercial officer. Moody brings more than 20 years of experience in the contract research and laboratory services industries, most recently serving as CIO for MDS Pharma Services. Findlater joins WIL Research from Covance, where he was global vice president of sales and client services for the nonclinical safety assessment and analytical services divisions.

Steve Self, development director at e-Therapeutics (Newcastle upon Tyne, UK), has agreed to become an executive director on the company's board. He previously served as group R&D director at Merck Generics and worked for a private equity bank on major US pharmaceutical acquisitions. He currently serves as a director at KMS Therapeutics.

Rep. **Fred Upton** (R-Mich.) has been named chairman of the House Energy and Commerce Committee, which has jurisdiction over FDA operations. Upton, who has served on the committee since 1991, will succeed Rep. **Henry Waxman** (D-Calif.). Upton plans to make legislation to repeal or restructure the Patient Protection and Affordable Care Act a committee priority. His stated positions include supporting federal funding of human embryonic stem cell research, opposing negotiating Medicare Part D drug prices and opposing drug importation.

William Vickery (left) has been named head of corporate and business development of Hybrigenics (Paris), a newly created position. Previously, he served as senior director of business development at ExonHit Therapeutics and as a business development director at Roche Pharmaceuticals.