VOLUME 28 NUMBER 1 JANUARY 2010
www.nature.com/naturebiotechnology

nature biotechnology

THE SCIENCE AND BUSINESS OF BIOTECHNOLOGY

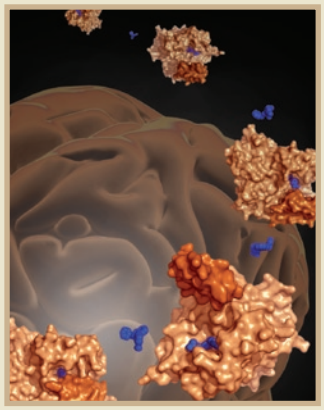
Allosteric modulators of phosphodiesterase 4

Towards the human pan-genome

Stem cells improve cancer models

VOLUME 28 NUMBER 1 JANUARY 2010

nature biotechnology



Structures of phosphodiesterase 4 in open and closed conformations and several allosteric modulators. Burgin, Gurney and colleagues solve crystal structures of phosphodiesterase 4 bound to ligands and design allosteric modulators with pro-cognitive effects. (p 63)

© 2010 Nature America, Inc. All rights reserved.

Chinese tech exchange launches, p 9

EDITORIAL

NEWS

BIOENTREPRENEUR

BUILDING A BUSINESS

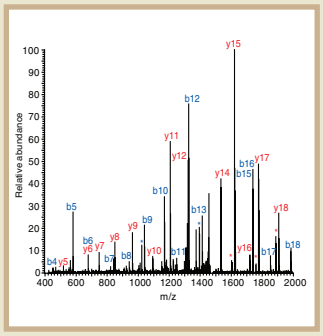
OPINION AND COMMENT

CORRESPONDENCE

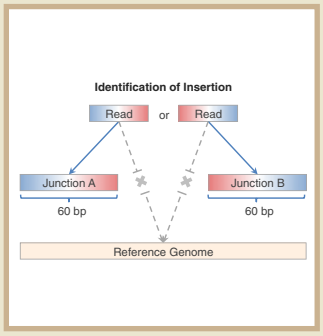
Nature Biotechnology (ISSN 1087-0156) is published monthly by Nature Publishing Group, a trading name of Nature America Inc. located at 75 Varick Street, FI 9, New York, NY 10013-1917. Periodicals postage paid at New York, NY and additional mailing post offices. **Editorial Office:** 75 Varick Street,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

BPA WORLDWIDE

npg
nature publishing group



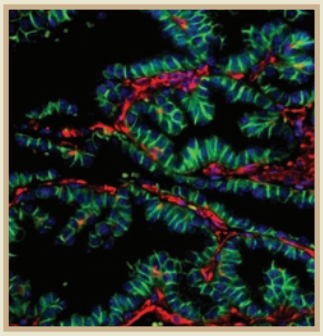
Normalizing proteomics data, p 40



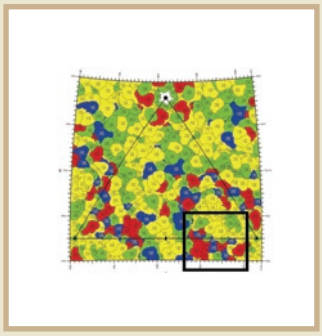
A library of genome breakpoints, p 47

Panning for novel human sequences, p 57

© 2010 Nature America, Inc. All rights reserved.



Better cancer models, p 71



Muscle delivery with chimeric AAV, p 79

© 2010 Nature America, Inc. All rights reserved.

ARTICLES

LETTERS

RESOURCE

CAREERS AND RECRUITMENT

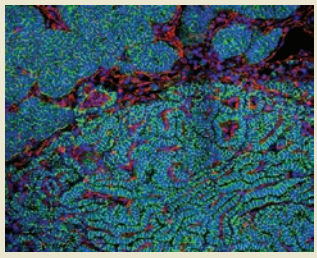
ADVERTISEMENT

BioPharma Dealmakers

What are the latest business and partnering strategies adopted by companies working to develop new cancer treatments and diagnostics? Freelance journalist Crispin Littlehales looks at how different organizations are synergizing strengths to build new franchises in cancer. Also in this installment of BioPharma Dealmakers, freelance journalist Barbara Nasto takes a look at the latest trends in outsourcing to contract research organisations (CROs). The BioPharma Dealmakers section follows Letters after page 89 and is produced with the commercial support from the organizations featured in the Advertorial Partnering Profiles.

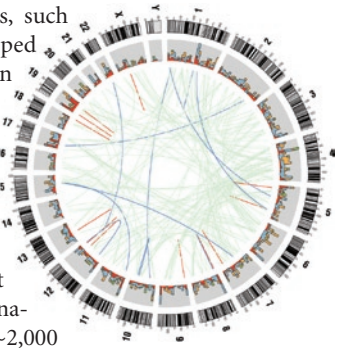
© 2010 Nature America, Inc. All rights reserved.

Models that better mimic human cancer



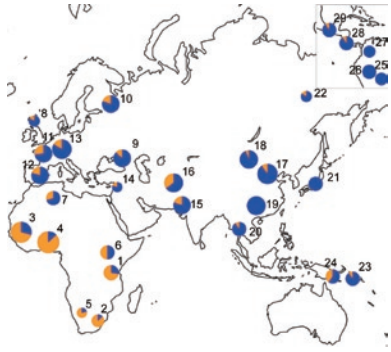
The low predictive value of mouse cancer models for human disease is a major challenge for cancer research. Whereas human tumors develop from individual cells in the context of normal tissue, cancer research mostly relies on models employing xenografts or carrying oncogenic mutations throughout the whole animal or tissue. Using engineered embryonic stem cells, Heyer and colleagues now model the tissue context of cancer development more faithfully in chimeric mice. By injecting embryonic stem (ES) cells with a knockout of a tumor suppressor gene and an inducible oncogene under the control of a lung-specific promoter into normal blastocysts, they generate mice that develop tumors within a year of oncogene induction. These tumors have a histology resembling human lung tumors and show sensitivity to targeted therapy that depends on the causing oncogene, similar to the corresponding human cancers. The authors analyze the activation of the signaling pathway in tumors caused by *HER2*, *KRAS* or *EGFR* mutations and find specific differences that could be exploited therapeutically. After the initial ES cell line with a tumor suppressor and a tissue-specific promoter is established, testing of additional oncogenes is relatively fast, with the first cancers being observed <8 months from the time of project initiation. [**Articles, p. 71**] *ME*

Breaking structural variation



Genomes contain rearrangements, such as insertions, deletions or swapped regions of varying sizes. Gerstein and colleagues demonstrate the utility of compiling a collection of known regions where these rearrangements border normal DNA sequence, known as the 'breakpoints' of the rearrangement. This collection is called a breakpoint library, and in its current incarnation contains the breakpoints of ~2,000 rearrangements that have previously been experimentally mapped to nucleotide resolution. The breakpoint library can be used to analyze the molecular and evolutionary mechanisms by which genomes are broken and repaired—a common source of rearrangements. For instance, Gerstein and colleagues find that breakpoints generated by nonallelic homologous recombination are associated with genomic regions that form relatively rigid DNA helices. Having a breakpoint library also enables them to create a software-based approach, BreakSeq, for identifying the breakpoints to nucleotide resolution from the short reads of a next generation–sequenced genome. Such resolution is necessary to ascertain the effects of rearranging protein-coding regions of the genome. BreakSeq and the breakpoint library should be valuable for ongoing human genome sequencing efforts. [**Resource, p. 47**] *CM*

Building the human pan-genome



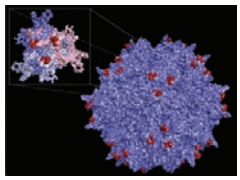
One person's DNA is not the same as another's. Wang and colleagues explore just how different two human genomes may be by comparing the genomes of a person of Asian ancestry and one of African ancestry to the NCBI reference sequence. In contrast to previous studies that resequenced the same Asian and African individuals by mapping short sequencing reads to the reference genome, the authors use a new algorithm to assemble the short reads into longer sequences without the aid of the reference. This *de novo* assembly generates contiguous stretches of base pairs that are long enough to reveal novel sequences not present in the reference. Intriguingly, the novel sequences contain protein-coding regions, and some are found predominantly in specific human populations. In total, about 5 Mb of sequence is unique to each genome, and there is an estimated 19–40 Mb of sequence in the worldwide human population not present in the current reference genome. Taken together, these findings indicate that each of our genomes is not simply a slightly rearranged version of the reference sequence with some single-nucleotide changes. Instead, the genetic makeup of each person may be augmented by millions of DNA bases pairs—an insight that requires that we rethink the way we use, collect and analyze individual genome sequences. [**Analysis, p. 57**] *CM*

ProcesSI$_N$g quantitative proteomics data

The intrinsic variation between mass spectrometry data collected from replicate samples is a major hurdle to realizing the full potential of shotgun proteomics. Schnitzer and colleagues address this challenge by integrating fragment-ion intensities with mass spectrometry features used previously to quantify proteins from label-free proteomics data. Their normalized spectral index (SI$_N$) not only outperforms alternatives, such as spectral counting and the 'area under the curve' method, but is also relatively simple to use. The features required to calculate the index are readily extracted from mass spectrometry data and no software is required to interpret the data. [**Letters, p. 83; News and Views, p. 40**] *PH*

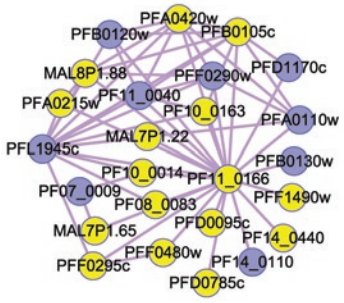
Written by Kathy Aschheim, Markus Elsner, Michael Francisco, Peter Hare & Craig Mak

Muscle-bound AAV



The ability to control the tropism of viral vectors would be very beneficial in gene therapy. Various routes to this goal have been explored, many involving modification of the capsid surface to alter interactions of the virus with cellular receptors. Asokan and colleagues have pursued such an approach for adeno-associated virus (AAV), a vector that has high clinical potential because it infects non-dividing cells and has little pathogenicity or immunogenicity. Using existing structural data on the heparan sulfate receptor 'footprint' (the residues on the viral surface that contact the receptor), the authors swap a six-residue stretch with the corresponding sequence from other AAV strains. One of the resulting chimeric viruses shows promise for application in muscle diseases as it transduces a wide range of muscle groups with high efficiency. Moreover, unlike all natural AAV isolates studied to date, it does not accumulate in the liver. [**Letters, p. 79**] *KA*

Plasmodium functional genomics



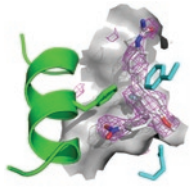
Full characterization of all of the genes of the malarial parasite *Plasmodium falciparum* is likely to be critical to developing effective strategies to control this devastating infectious disease. But largely owing to the inadequacy of current techniques for *P. falciparum* reverse genetics, the functions of more than half of its genes have yet to be identified. Probabilistic interaction networks based on data collected after chemically induced growth perturbations have been used to suggest functions for hypothetical proteins of many model organisms. But the relatively monotonous regulation of *P. falciparum* transcription, most frequently attributed to the stable environment provided by the host's red blood cells and the observation that it has only a third as many transcription factors as most eukaryotes, has discouraged researchers from using this approach to annotate its genes. Bozdech and colleagues contradict the perception that *P. falciparum* transcript abundance is hardwired by demonstrating relatively robust changes in transcript levels in response to 20 chemicals that perturb growth or development. They integrate these data with phylogenetic information, domain-domain interaction data and results from yeast two-hybrid screens to create an interaction network that predicts the functions of >50% (>2,500) hypothetical proteins from *P. falciparum*. By potentially revealing new components that contribute to invasion and pathogenesis, this approach offers a promising strategy to streamline selection of drug targets and vaccine candidates to combat malaria. As proof of this concept, Bozdech and colleagues predict proteins that are associated with merozoite invasion, one of the most promising targets in malaria intervention programs. Subcellular imaging confirms that 31 proteins predicted to contribute to this process indeed localize to intracellular compartments associated with the invasion apparatus. [**Resource, p. 91**] *PH*

Patent roundup

The recognition of RNA-mediated gene suppression as an important experimental tool as well as its potential commercial application is reflected in the patent landscape, with an increasing number of patent applications seeking exclusive rights to RNA interference (RNAi)–based discoveries. Chi-Ham *et al.* summarize the patent thicket and point out legal uncertainties on who will own key RNAi intellectual property.
[**Patent Article, p. 32**] *MF*

Recent patent applications in induced pluripotent stem cells.
[**New patents, p. 37**] *MF*

Cognitive enhancement without emesis



Hydrolysis of the ubiquitous second messenger cyclic AMP is accomplished largely by a single enzyme family called phosphodiesterase 4 (PDE4). PDE4 has long been a target of high interest in the pharmaceutical industry as PDE4 inhibitors are effective against diverse pathologies, including inflammation and cognitive dysfunction. Yet none of these inhibitors is approved for use in patients, and clinical development has been slowed by emetic side effects. Burgin, Gurney and colleagues now present the first PDE4 crystal structures that show both the catalytic domain and either of two regulatory domains. The structures reveal that the regulatory domains function by closing over the catalytic pocket and blocking access to it. Using the new structural information, the authors undertake a large medicinal chemistry effort to find allosteric modulators that might have greater tolerability than earlier inhibitors. Of the >800 compounds synthesized, 140 are partial inhibitors with allosteric inhibition kinetics. The most promising of these show high pro-cognitive efficacy in mice and much lower emetic side effects than earlier inhibitors in several animal models. [**Articles, p. 63; News and Views, p. 38**] *KA*

Next month in nature biotechnology

- Rational lipid design improves siRNA delivery
- Endothelial cells from hES cells
- Antibody half-life linked to efficacy
- Predicting plant gene function
- Hepatitis C virus imaged in real time

© 2010 Nature America, Inc. All rights reserved.

nature biotechnology

Gathering clouds and a sequencing storm

Why cloud computing could broaden community access to next-generation sequencing.

© 2010 Nature America, Inc. All rights reserved.

Those with any doubts as to whether we have entered the decade of the sequencer need only pay a visit to the Broad Institute in Cambridge, Massachusetts. There in the lobby, a wall of flat screen TVs displays an endless stream of A's, T's, C's and G's, with a mind-boggling multiple-digit readout counting up the number of DNA base pairs sequenced. The Broad is one of a hundred or so research centers around the world currently generating thousands of gigabases of DNA sequence every week. But whereas researchers at these centers have a wide array of core computing resources and expertise at their disposal for analyzing the reams of data they generate, smaller laboratories that intend to purchase next-generation sequencers are not so fortunate. For the latter, more funding and effort should be devoted not only to the development of on- and off-site data management solutions but also to disseminating software from core facilities to the broader community.

In the coming year, it is not unreasonable to expect that the amount of sequence data generated around the world will outstrip that generated in the past decade. The National Institutes of Health's Cancer Genome Atlas (CGA; http://cancergenome.nih.gov/) is already ramping up its effort to sequence hundreds of genomes in 20 different types of cancer. And almost everywhere one looks, other sequencing projects are springing up or gathering pace; examples include the 1000 Genomes Project (a high-resolution map of human genomic variation from 1,000 individuals; http://www.1000genomes.org/), the Personal Genome Project (the exomes of tens of healthy volunteers; http://www.personalgenomes.org/), the 1001 Genomes Project (sequence variation in 1,001 strains of *Arabidopsis thaliana;* http://1001genomes.org/); and the Mouse Genomes Project (the genomes of 17 mouse strains; http://www.sanger.ac.uk/modelorgs/mousegenomes).

Next-generation sequencing platforms are playing an increasingly prominent role in resequencing efforts. But their role in *de novo* sequencing and assembly is also broadening from simple microbes to filling gaps and providing finer sequence resolution and coverage in higher organisms—the characterization of the human pan-genome on p. 57 being one example.

These efforts, together with a burgeoning number of additional applications for next-generation platforms in small RNA discovery, transcriptomics, chromatin immunoprecipitation and copy number variation studies, mean that deep sequencing instruments are likely to become indispensable and ubiquitous tools in the biology laboratory. But for this technology to be truly widely adopted, challenges related to data handling and analysis remain to be addressed.

Next-generation sequencers produce a prodigious stream of data. A single Illumina instrument, for example, can generate up to 90 billion bases per run. This represents terabytes of raw image data that require at a minimum 4 GB of RAM and 750 GB of local storage capacity to carry out the data handling and analysis.

Whereas genome centers are set up to deal with such gargantuan files, most academic laboratories are in a completely different situation. They have no large central computing pool and data storage capacity. They are more likely to generate data in an *ad hoc* manner, rather than in a steady stream amenable to an automated data management pipeline. And they often lack sequencing specialists and support staff working under the same roof who can create software tailored to their needs and solve computational problems.

Some algorithms, such as those for mapping short DNA reads to a reference genome, have progressed to a high level of sophistication and are widely accessed. This is because these programs—written by cross-disciplinary individuals—have now been optimized by computer scientists to enhance user friendliness and remove bugs. In other areas, such as short RNA read mapping or analysis of genome structural variation, progress has been slower, in part because the problem is more complex and in part because the data have not been available.

One potential solution to the data handling/storage problem for smaller research groups is the use of cloud computing (see p. 13). In this approach, a user rents processing time on a computer cluster (e.g., from Amazon) through a virtual operating system (or 'cloud'), which can load software and provide an access point for running highly parallelized tasks. Sequencing data can be sent to the cluster either by disk or the internet (although the size of data sets presents its own problems for the latter).

The first software (CrossBow) capable of performing alignment and single nucleotide polymorphism analysis on multiple whole-human data sets on a computing cloud was published just 6 weeks ago (*Genome Biol.* **10**, R134, 2009). Essentially, the package makes it possible to analyze an entire human genome in a single day while sitting with a laptop at your local Starbucks.

It remains unclear, however, whether the cost of routinely renting time on the cloud would be cost effective in the long term, particularly if a user intends to analyze billions of base pairs of genome sequence on a regular basis. What's more, if the wide uptake of sequence analysis on clouds depends on the availability of user-friendly, debugged software, bioinformaticians might not be willing to spend the time to familiarize themselves with hadoop, the open source program needed to process large data sets on a cloud—especially when their jobs focus on developing algorithms for their own local computer clusters.

Thus, for next-generation sequencing to move out of genome centers, more effort must focus on creating software compatible for use in a cloud or better still, infrastructure software (similar to Apache for web servers) that would allow community-generated software for all types of sequence analysis to be plugged into it. This approach is likely to be particularly valuable for smaller laboratories lacking software development resources. And although it will not solve all the data management and analysis problems associated with next-generation platforms, it could give many the opportunity to adopt a powerful and rapidly advancing technology that would otherwise remain out of reach.

IN this section

Public biotechs on the rebound p5

Academics loosen ties with industry p7

China launches high-tech stockmarket p9

As Genzyme flounders, competitors and activist investors swoop in

Genzyme's contamination woes have fueled headlines for the past 12 months. Just as the company seemed to have resolved the viral contamination at its Allston, Massachusetts plant, a new problem arose in the fall: steel, rubber and fiber fragments were found in vials from the same facility. As the company was forced to withdraw or scale back affected products for people with rare, life-threatening conditions, other companies have moved quickly to fill the vacuum. The US Food and Drug Administration (FDA) took steps to accelerate approval of Gaucher's and Fabry's disease treatments, made by Genzyme's main competitors. In November, as the gloom deepened, Genzyme's CEO Henri Termeer sold a noticeable chunk of his stock through an automatic sale.

The series of black eyes for Genzyme has not only opened some of the company's key products up to competition earlier than anticipated, but also thrashed the stock and dented the image of the Cambridge, Massachusetts–based company, once known mainly for its innovation and remarkably profitable approach in developing drugs for rare diseases. Genzyme's serial manufacturing deficiencies have sparked questions about whether the management is competent or the company is just a victim of bad luck. Similar contamination problems have occurred at several other companies, but the double whammy of viral and particulate contamination have raised some eyebrows.

The FDA has been tightening the reins on biotech manufacturers, and Genzyme's situation may just be a reflection of that. Alan Burns, of Sartorius Stedim Biotech in Concord, California, points out one of the challenges with biotech drugs is that the manufacturing processes are so varied. "Some of the processes under which biotech drugs are made are not as well-controlled as what you see in a typical pharmaceutical company, where most of the products are similar with respect to their manufacturing process," says Burns, whose company sells equipment and services for biomanufacturing. During the early days of the industry, the FDA had to "allow more variability within biotech production processes," he notes, or biotech drugs would not have been able to be made at all.

But as the agency has become stricter about manufacturing standards, companies have been standardizing procedures and putting in more rigorous quality control measures. The goal is that even if contamination does occur, the affected product never reaches the public. "The FDA has shown a new interest in contamination, and this was a little bit initiated by other cases," says Huub Schellekens of the Department of Pharmaceutical Sciences at Utrecht University in The Netherlands. "Contamination is more or less a general problem, and I think it was just Genzyme's turn."

For Genzyme, the trouble started last February when the FDA sent a warning letter detailing "significant objectionable conditions" at the Allston plant, which has six bioreactors. These issues were expected to take several months to address. Then, four months later, Genzyme announced vesivirus 2117 had been detected in a bioreactor producing Cerezyme (imiglucerase; recombinant human (rh) β-glucocerebrosidase) for treating Gaucher's disease. Production for Cerezyme and another drug Fabrazyme (agalsidase β; rh α-galactosidase A), for Fabry's disease, had recently been cut so that Myozyme (alglucosidase alfa; rh alpha-glucosidase), for treating Pompe's disease, could be added to the plant's schedule. As a result, manufacturing of all three products ended up being severely compromised when the plant was shut down to be disinfected.

This is not Genzyme's first encounter with vesivirus: in 2008, the virus had contaminated the Allston plant and the company's brand new manufacturing plant in Geel, Belgium. There is no evidence that the virus harms people; rather, it diminishes the productivity of the cell lines used to pump out recombinant proteins.

Because these are drugs for extremely rare and life-threatening diseases, patients and their doctors quickly became frantic as supplies plummeted. Europe's Committee for Medicinal Products for Human Use reduced the amount of Cerezyme given to each member nation by 80%. The medical advisory board of the National Gaucher Foundation in Tucker, Georgia, recommended that patients with the most advanced disease should get priority access to the drug and that others should receive lower doses to stretch out the small supply.

It is not clear yet how this shortage has affected any patient's wellbeing. But the ramifications for Genzyme are not good either way: if patients were harmed by the temporary shortage, it's a tragedy; if the very low doses worked, demand for Cerezyme might go down, even when the company restores its manufacturing facility to full capacity.

In addition, over the past few months, the FDA has given rivals Shire, located in Basingstoke, England, and Protalix Biotherapeutics of

Jonathan Wiggs/Globe Staff

Sandra E. Poole took control of Genzyme's Allston plant last summer, charged with overseeing the company's vital clean-up efforts.

© 2010 Nature America, Inc. All rights reserved.

IN brief

Sarkozy's great biotech loan


AP Photo/Christophe Ena

Sarkozy reveals the spending plans.

President Nicolas Sarkozy has approved the national 'Grand Emprunt', French for 'big loan', a €35 ($73.7 billion) economic stimulus package to fund French industry and infrastructure. The borrowing scheme unveiled December 16 is heavily focused on education, research and innovation, with at least €5.5 billion ($7.9 billion) flowing into the life sciences, biotech, clean-tech and academic research. The Grand Emprunt is by far the biggest, though not the first, government-driven plan to benefit the biotech sector. In November, the Kurma Biofund was launched, as a joint partnership between the public financing body Caisse des Depots et Consignations (CDC) Entreprises, Paris, and venture capital group Natexis Private Equity, Paris. The €50 million ($71.5 million) fund, which will increase to €100 million ($143 million) next year, is open to newly-created biotech companies spinning off European academic centers. InnoBio, announced in October, is a €139 million ($199 million) fund aimed at boosting the development of small-to-medium enterprises working in drug discovery and related technology platforms such as imaging, diagnostics and bioproduction. This dedicated biotech fund was created by the government's Strategic Fund for Innovation (FSI), which will contribute €52 million ($74.4 million), with the rest provided by nine corporate pharma partners, including €25 million ($35.8 million) pledged by Paris-based Sanofi Aventis and London-based GlaxoSmithKline. That the national CDC deposit fund is part of these financial investment instruments shows the French government's resolve to push biotech onto its agenda. But André Choulika, France Biotech President, laments the small sums invested. "InnoBio represents the cost of ten days of R&D in big pharma," he says, "nevertheless it could act as a catalyst to attract additional private funding." Existing schemes coupled with the 'Grand Emprunt' could signal a turning point for the biotech industry. According to presidential advisors Arnold Munnich, head of pediatrics, Necker Hospital, Paris, and economist Bernard Belloc, University of Toulouse, the French government understands that long-term growth depends on bridging the gap between academia and industry and is putting its muscle behind public-private partnerships to ensure university tech transfer and research evaluation are upgraded and professionalized. Ramin Chaybani from Novoptim, a Paris-based business development consultancy for European biotechs, points out, "The diagnosis is correct. Now we need to wait and see whether these measures have a catalytic effect." *Golbahar Pahlavan*

Carmiel, Israel, a major boost by accelerating review of, and facilitating early access for patients to, competing products that are not yet formally approved. Shire's velaglucerase alfa for Gaucher's and Replagal for Fabry's disease, and Protalix's Uplyso (taliglucerase alfa) for Gaucher's are all moving much more quickly thanks to the Allston debacle. Reports suggest that relatively few patients were switched to these still-experimental drugs, but, at the very least, Genzyme's problems have helped to increase patient and physician awareness about these rival products. In December, Pfizer even jumped into the fray, inking a major deal with Protalix for the plant-generated Uplyso, including a $60 million upfront payment.

One of the key steps Genzyme has taken to upgrade its manufacturing processes is to assign Sandra E. Poole head of the Allston facility. Poole is a senior vice president and for the past few years oversaw the construction and European approval process for the company's new manufacturing facility in Geel, Belgium. Genzyme also developed a new test to detect vesivirus, which the company says was introduced by culture materials used in the manufacturing process.

The Allston plant was back in operation by last August, but November heralded reports of more contamination. This time, "foreign particles," steel and other materials, were uncovered in less than 1% of product vials packaged at the Allston plant. No patients appear to have been harmed by these foreign particles, and because some of the affected drugs were the only treatments available, the FDA did not demand a recall. Instead, the agency issued a warning cautioning physicians to carefully inspect any vials of drugs produced at the plant before using them. Shortly after, Genzyme also announced that the agency had completed its review of Allston and had "provided Genzyme with a Form-483 outlining remaining deficiencies, which were mainly related to the fill/finish capabilities at the facility."

It's not uncommon to see visible particles in some types of biotech products, mainly due to protein aggregation. But via e-mail, Patricia Hughes, from the Division of Manufacturing and Product Quality at the FDA's Center for Drug Evaluation and Research noted that, "According to the USP [United States Pharmacopeia], all parenterally administered articles [e.g., Genzyme's recombinant products] must be prepared in [a] manner designed to exclude particulate matter." Certain types of contaminants, such as fibers or metal shavings, can also indicate "poorly controlled operations or inadequate maintenance," she added. If a company is getting that kind of debris in the product, "Some part of your downstream filling system may not be set up right," says Bill Bees, senior vice president of operations at Cangene in Toronto, Ontario, which has a fill finish facility in Baltimore. For example, if a piece of equipment has been improperly assembled, metal shavings can arise when the ill-fitting pieces rub together.

Biomanufacturing experts acknowledge that contamination is a widespread concern in the industry, but many were disconcerted by the comments of Genzyme's senior vice president Geoff McDonough in the media. In an interview with the *Boston Globe* newspaper, McDonough stated that Genzyme's particle contamination rate was "within industry standards." The response of one manufacturing expert, speaking off the record, was "With that comment, they threw us all under the bus."

Genzyme CEO Termeer responded to the crisis by announcing changes to operating procedures and a new focus on quality in manufacturing. But the biggest changes yet may be foisted on the company by activist investors. On December 10, *The Wall Street Journal* reported Genzyme had named an independent director to its board after pressure from stockholder Relational Investors of San Diego. This new director is Robert Bertolini, who was chief financial officer at Kenilworth, New Jersey-based Schering-Plough when it pulled off a remarkable turnaround in 2008. According to the *Wall Street Journal*, Relational Investors head, Ralph Whitworth, who is renowned for activism, wants to see even more changes at Genzyme.

The good news finally seems to be trickling through. In December, the company announced the first new shipments of Cerezyme since the Allston plant was closed for cleaning. Shipping of Fabrazyme was also imminent as *Nature Biotechnology* went to press. Also in December the company announced an agreement with the FDA on a regulatory path for the long-delayed Lumizyme (a form of Myozyme for adults). The firm predicts its manufacturing capacity will increase fourfold from 2006 to 2012, both through new facilities and expansion of established plants. But it's going to take a massive effort for the company to undo all the damage. "They are taking the right steps by minimizing Allston's role while it's remediated," says Josh Schimmer, managing director and biotechnology analyst at Leerink Swann of Boston, "but they are having to scramble because they have done irreparable harm to the franchise." Schimmer approves of Bertolini's appointment and the fact that Genzyme is finally responding to shareholders' wishes. "The organization can be structured in such a way that these problems don't recur," he says.

Malorye Allison *Acton, Massachusetts*

© 2010 Nature America, Inc. All rights reserved.

Optimism in public biotech rises as credit crunch recedes

deCODE genetics, the public biotech company based in Iceland that served as the poster child for genetics-driven drug discovery, filed for Chapter 11 bankruptcy in November 2009. Riding the hype around the Human Genome Project, deCODE's stock once traded above $25 per share. The demise of deCODE says as much about the inadequacy of biotech investment models to sustain the development of new drugs from novel biology as it does about the current financial climate. But observed from a distance, the bankruptcy looks like just another biotech casualty of the credit crunch.

In this respect, all is not gloom and doom. A survey of the cash status of 294 public biotechs by *Nature Biotechnology* reveals the sector has been more resilient to the credit shortfall than initial predictions might have suggested (*Nat. Biotechnol.* **27**, 493, 2009).

As *Nature Biotechnology* went to press, 22 biotech firms had filed for bankruptcy since January 2008, including 10 in 2009, according to figures from the Biotechnology Industry Organization in Washington, DC. Nine companies, though not officially bankrupt, have also closed business this year. Delistings of public companies from the exchanges have also remained common: 21 healthcare companies were delisted from the NASDAQ in 2009 through November, whereas 22 were delisted in 2008.

But our analysis of the 294 public biotech filings to the end of September reveals that financing conditions have also started to ease over the past year. Roughly 28% of all biotechs were operating with less than one year's cash reserves on September 30, an improvement from an earlier analysis showing 39% based on a mix of 4Q 2008 and 1Q 2009 filings.

Kári Steffánson, deCODE's president and CEO.

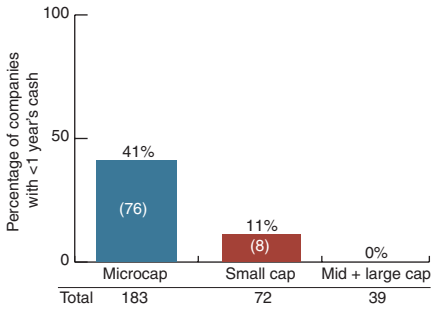


Figure 1 Percentage of biotech firms operating with less than one year's cash, segmented by market cap. Microcap, <$250 million; small cap, $250 million to <$1 billion; midcap, $1 billion to <$5 billion; large cap, ≥ $5 billion.

One thing has not changed, however: microcap and small cap firms have been the hardest hit (**Fig. 1**). Even here, though, the outlook is improving, with 41% of microcap firms operating with less than a year's cushion compared with 50% last year (see *Nat. Biotechnol.* **27**, 493, 2009). Strong merger and acquisition (M&A) activity has provided a lifeline for several companies, and the resurgent stock market may mean that access to public financing could improve. At the time of writing, the NASDAQ Biotechnology Index had climbed some 20 points since early March. This has observers predicting better days. "There is a lot more hope on the horizon now than there was a year ago, when there was none at all," says Chris Wasden, managing director at PricewaterhouseCoopers in New York.

Financing figures back up that hope. Through November, the biotech industry had raised in 2009 $6.6 billion through public offerings, initial or otherwise—a tremendous improvement over the $2.25 billion raised through public offerings in all of 2008. A month-by-month breakdown of those figures (**Table 1**) shows that the surge came in the second half. That reverses the trend from 2008, when the public markets withered as the year progressed.

Table 1 Public offerings (initial and follow-on) raised by month, 2008–2009 ($M)

	Jan.	Feb.	March	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
2009	55.8	423.4	0.0	48.8	1,063.2	97.4	691.6	867.0	1,126.3	1,022.1	371.1	—
2008	413.6	689.9	89.9	0.0	99.9	177.3	12.8	546.0	219.9	0.0	0.0	0.0

Source: Bioworld.

© 2010 Nature America, Inc. All rights reserved.

IN brief

New EU states ranked

The biotech industry in most of the EU's new member states lags behind that of their Western European neighbors, despite many declaring biotech a national priority. The 14 allbio report produced by EuropaBio and Zurich-based Venture Valuation is the first to gather data on the state of the biotech industry in the 12 new member states and 2 candidate countries. The report identifies 260 biotech companies—70% are in the service sector—operating in these countries and at least 18 therapeutic products in development. Hungary, Poland, the Czech Republic and Estonia lead the group, although the report found no correlation between a country's gross domestic product and the strength of its biotech sector. According to Patrik Frei of Venture Valuation who authored the report, despite a highly educated workforce, these countries lack support structures for small and medium-sized enterprises and funding for intellectual property and technology transfer. "It is essential to get these fundamentals right because private equity and venture capital will not invest without them," Frei notes. Erno Duda, who heads the Hungarian Biotechnology Association, notes that the Hungarian biotech sector has grown from 10 to 110 companies in only a few years. "Hungary has a strong science base, especially in medicinal chemistry," he says. "The biggest obstacle now is lack of management knowledge. But we are beginning to have the critical mass of companies to make a cluster." *Susan Aldridge*

15 states sue Amgen

New York and 14 other states have filed a lawsuit alleging that Amgen offered doctors illegal kickbacks to increase sales of its blockbuster anemia drug. The multi-state case charges Thousand Oaks, California–based Amgen of overfilling vials of erythropoietin-stimulating agent Aranesp (darbepoetin alfa) by 16–19% to provide medical practices with free product for which they could then bill insurers. An ex-employee turned whistleblower, who first filed a complaint in 2006, alleges that Amgen's sales force promoted the overfill and the revenue it would bring from third-party payers such as Medicaid. Drug wholesaler ASD Healthcare and drug-purchasing International Nephrology Network, both based in Frisco, Texas, are also named in the suit. The plaintiffs are requesting treble damages, which could amount to several billions of dollars; the defendants deny the charges. Wells Wilkinson, director of the Boston-based consumer watchdog Prescription Access Litigation, calls the scheme a creative variation on a widespread marketing tactic of inducing doctors to inflate drug prices on reimbursement. "Recent lawsuits have been seeking increasingly larger fines from drug companies," he notes. "These are encouraging signs that the federal government is going to take a stronger stand against prescription drug fraud." Until the penalties are sufficiently severe, companies will continue to use these types of "egregious" marketing ploys, he predicts. *Asher Mullard*

There is also brightening news regarding employment. After a staggering 52 companies reduced their workforce in the first quarter, restructuring and downsizing activity tapered off significantly (**Table 2**); indeed, 2009 could end up resembling 2008 in totality—a victory of sorts. In this regard, M&A activity has hampered job growth, Wasden says, as merging companies create redundancies, and thus pink slips, and it usually takes months before those people find their way to new jobs or start-ups.

Even if the crisis is passing, opinions vary on what sort of rebound biotech will see. When the economy retracts as strongly as it did in 2008, history suggests a recovery of similar slant—a 'V-shaped' recovery, as opposed to a 'W' shape, wherein the economy would surge and fall again. Wasden thinks the recovery will fall somewhere in between.

Table 2 Company restructurings

2006	35
2007	57
2008	114
2009 (first three quarters)	98

Source: BioCentury.

"The general view is that the worst is behind us," he says. "And I'm actually more optimistic than a lot of people. I see it as more of a 'U' recovery."

This means slower, gradual growth, which might not give much comfort to those firms treading water or in danger of slipping beneath the waves. But given the way 2009 began, any signs of recovery are fueling optimism.

Brady Huggett, Business Editor

The biotech Stradivarius

Thanks to a 'biotech' intervention, the modern fiddle in the picture fooled more than 100 listeners in a blind test. Professor Francis Schwarze of the Swiss Federal Laboratory for Materials Testing and Research treated Norwegian spruce and sycamore with two fungi to recreate the effects of cold climate thought to cause the superior quality of the wood used by Antonio Stradivari in the 17th century. Schwarze commissioned violin craftsman Michael Rhoneimer of Baden, Switzerland to build an instrument which was tested alongside untreated fiddles and a $2 million Stradivarius. Listeners were asked to identify the Strad, and while 113 picked the biotech fiddle, only 39 correctly identified the Strad.

Michael Rhoneimer

IN their words

"It's not credible for him not to be considered a lobbyist," Craig Holman, of consumer advocacy group Public Citizen, argues that BIO's Jim Greenwood should officially declare himself a lobbyist. (*Pharmalot*, December 7, 2009)

"I do believe the sections relating to the creation of a market for biosimilar products is one area of the bill that strikes the appropriate balance in providing lower cost options." Rep. Lynn Jenkins (R. Kansas), one of dozens of lawmakers who, during the health-care debate in the House, parroted language verbatim from text provided by Genentech (S. San Francisco, California). (*New York Times*, November 14, 2009)

"The Iceland-based subsidiary [Islensk Erfdagreining] that performs all of deCODE's human genetics work—manages its population resources, conducts its research and services, offers and processes its tests and genome scans, and whose scientists and laboratories are licensed to undertake this work—is not in bankruptcy." DeCODE genetics CEO Kari Stefansson clarifies that his company's bankruptcy will not affect the Icelandic database that holds the medical and genetic records of the island's population. (*Nature News*, November 23, 2009)

"We're excited about the potential of this. It can be a fundamental game-changer." John Maraganore, of Cambridge, Massachusetts-based Alnylam, on the company's internal small-interfering RNA program against transthyretin amyloidosis. (*Xconomy*, December 12, 2009)

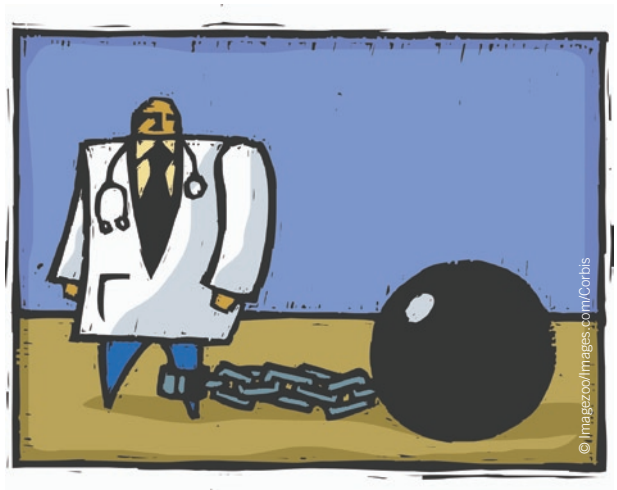
New product approvals

Kalbitor (ecallantide)	Dyax (Cambridge, Massachusetts)	The US Food and Drug Administration approved Kalbitor for acute attacks of hereditary angioedema (HAE) in patients 16 years and older. The first subcutaneous HAE treatment, the drug is a reversible plasma kallikrein inhibitor.

© 2010 Nature America, Inc. All rights reserved.

Report concludes industry–academia partnerships on the wane

A survey published in the November issue of the journal *Health Affairs* reports that academic researchers' links with industry are on the ebb (*Health Affairs* **28**, 1814–1825, 2009). The survey's results contrast with a growing perception in the media and political circles that industry's influence on academic research is on the increase—claims that have been fueled by high-profile cases in which researchers have failed to fully disclose corporate ties. An ongoing investigation by US Senate Finance Committee member Chuck Grassley into federally funded-researchers who flout the rules on ties with drug makers adds to such views. The data in the study showed that of the 3,080 scientists interviewed, 52.8% reported some kind of relationship with industry within the past three years. Overall, 20% of research faculty received industry funding, a drop from the 28% faculty who took part in a similar survey in 1995.



Universities' conflict of interest policies may explain, at least in part, the chill in academic-industry relationships.

"Universities and academia are eager to form relationships with industry, and politicians expect it. And yet at the same time, we keep finding complaints about conflicts." Art Caplan, professor of bioethics at University of Pennsylvania and director of its Center for Bioethics, describes the inherent contradictions in the situation. "It's almost as if the people we elect as public officials say 'get married to industry', but then our ethics say 'be careful about consummating the marriage'."

The report's respondents were drawn randomly from life science departments at the top 50 US National Institutes of Health (NIH)-funded universities. Industry ties included consulting, paid speaking, research funding through a grant or contract as the principal investigator, and sitting on scientific advisory boards. A subset focusing on scientists working in biotech found industry collaborations to be 17%, compared with 23% in 1985 and 21% in 1995.

Over the three year period covered in the report, industry supplied an average of $33,477 in research funds, excluding overheads, per respondent, a figure which constitutes 8.7% of all research funds received by faculty. Clinical faculty members received a greater proportion of their funds from industry than did nonclinical faculty members (10.5% versus 2.5%). Industry funding was significantly higher within clinical departments than in nonclinical departments (47.3% versus 26.3%). Of the faculty with industry support, the median amount of industry funding in 2006 was $99,000—similar to the (Consumer Price Index–Medical adjusted) value of $91,500 in 1996.

One explanation for the drop may be the fallout from a series of scandals that have rocked the medical community and adversely affected public opinion. Grassley's probe into irregularities in disclosures of industry ties has damaged the reputations of several high-profile academics, including Stanford University's Alan Schatzberg, Brown University's Martin Keller and Emory University's Charles Nemeroff (*Nat. Biotechnol.* **27**, 411–414, 2009). Such scandals are nothing new: almost a decade ago, University of Pennsylvania researcher Jim Wilson was enveloped by accusations of malfeasance when he failed to disclose a stake in Sharon Hill, Pennsylvania-based Genovo, the company that developed the ornithine decarboxylase gene therapy used in Wilson's trial that led to the death of teenager Jesse Gelsinger. In that case, even though Genovo provided no direct sponsorship or support for Wilson's trial, the perceived bias still severely damaged his reputation. Such cases thus not only have led to a tightening in ethical oversight of industry–faculty interactions, but also may have increased reticence on the part of academics to partner with companies.

The customary solution for such conflicts of interest (COI) is disclosure and transparency. But a report published in November by the US Department of Health and Human Services states that >90% of universities rely on professors to disclose their own conflicts, rather than file reports with the NIH as required, a factor that may tacitly encourage under-reporting. Universities are reluctant to force NIH-funded researchers to disclose financial relationships with drug makers for fear that they may lose those researchers—along with the researchers' star power and associated revenue for the university—to other institutions with fewer restrictions.

If many agree that oversight of COI in academia is inadequate, only a few have offered clear proposals for improving the situation. A report by the Washington, DC-based Institute of Medicine published last April (http://www.iom.edu/en/Reports/2009/Conflict-of-Interest-in-Medical-Research-Education-and-Practice.aspx) calls on health research centers, journals, professional societies and others to strengthen their COI policies through greater transparency, by rejecting gifts and by insisting that advisory boards do not themselves have members with industry ties. "You want experts making decisions," says Bernard Lo, Director of the Program in Medical Ethics at the University of California, San Francisco, who chaired the Institute of Medicine committee. "But the burden should be on those who say we can't find a qualified expert who doesn't have significant conflicts."

Some universities and a few states have attempted to address complaints about ethics in industry–faculty relationships by implementing new, tighter laws and regulations. An acknowledged problem with these tougher policies is that they tend to restrict productive interactions. "Some relationships are desirable," says Lo. "They benefit the public and industry as well as academia and should be fostered."

Minnesota is among a handful of states that require drug and device makers to disclose payments to clinicians. Minnesota's 1993 laws restricting interactions between industry and academic scientists have become a model for other states and may be emulated at the federal level. Last October, however, Michael Gonzalez-Campoy, an endocrinologist from Sunfish Lake, Minnesota, gave testimony about the effect of the 16-year-old legislation on the practice of medicine in the state. According to Gonzalez-Campoy, the legislation has led to a decline in medical education, confusion among patients and damage to the reputations of academics, who are forced to report relationships with industry in a manner that suggests it is always unethical.

"Entire practices have closed their doors to the marketing side of industry. Some of these include prominent institutions, such as Allina,

© 2010 Nature America, Inc. All rights reserved.

IN brief

China's GM rice first

Chinese officials have approved a strain of genetically engineered rice, placing the country in position to be first in the world to produce biotech rice on a commercial scale. China's Ministry of Agriculture in December said it had issued safety certificates for the rice but that additional production trials are required before full commercialization can begin. The trials may take two to three years. The rice variety, engineered to fend off pests with toxins from the bacterium *Bacillus thuringiensis* (Bt) was developed by scientists at Huazhong Agricultural University in Wuhan, China. News of the approval came in November, only a week after Chinese officials had announced approval of the country's first transgenic maize. The feed crop is engineered to produce phytase, an enzyme that helps animals better utilize phosphorus in maize. China's most widely grown transgenic crop, Bt cotton, was approved in 1997. China isn't the first nation to approve biotech rice, but it may be the first to commercialize it. US regulatory officials in 1999 approved, or deregulated, Bayer CropScience's transgenic herbicide-tolerant rice, but the North Carolina-based company never commercialized it. "Farmers are concerned that it will hurt their export markets," in countries that don't allow transgenic rice, says Doug Gurian-Sherman, a senior scientist at Cambridge, Massachusetts–based Union of Concerned Scientists. Farmers' fears were realized in 2006 when an unapproved transgenic rice variety contaminated US commercial rice, resulting in lost exports. *Emily Waltz*

Pea trials flee to US

Field trials of transgenic peas developed by a European university may relocate overseas to ensure a biotech-friendly environment. The University of Hannover in Germany is eyeing North Dakota as a safe place to evaluate several genetically modified (GM) pea lines intended as animal feed, under field conditions, marking the first time that EU-funded plant research has been forced to emigrate. "Vandals are seen as heroes by some media. [Field trial] locations have to be disclosed precisely so that the eco-terrorists can program their GPS," says Hans-Jörg Jacobsen, whose laboratory engineered the GM peas to express one or more antifungal genes. The relocation will be part of a scientific collaboration still under negotiation with the North Dakota State University (NDSU). Pollen flow is not a problem because peas are self-fertilizing plants, but in Germany, field testing could get into trouble anyway, and Jacobsen predicts there is an 80% chance the fields would be destroyed. "We face a militant resistance, which is extremely difficult to handle by a scientist which usually has only a small budget and limited personnel," sympathizes Jens Katzek from BIO Mitteldeutschland, a cluster promoting biotech. US trials are not expected to begin before 2011 for logistical reasons and will be performed ensuring "the highest level of containment and separation from commercial pea production channels," says Kevin McPhee a plant geneticist at NDSU. *Anna Meldolesi*

Mayo Clinic, Fairview and Park Nicollet. In turn, pharmaceutical companies have trimmed their local sales forces and several would rather not do business in Minnesota—they are gone from the state," Gonzalez-Campoy said.

Gonzalez-Campoy testified, he did not, however, offer any information about patient outcomes. In fact, there are no studies that address how patients are affected by COI issues. The activist group PharmedOut has submitted an open letter to Francis Collins, director of the NIH, signed by a large group of scientists, physicians and ethicists, asking that the NIH fund studies on medical ethics, COI in medicine and research and prescribing behavior (http://www.pharmedout.org/NIHLetter.pdf)

The issue has ruffled the feathers of many academics who work in translational research. Some feel under attack for their interactions with industry and characterize regulators and ethicists as a group of pencil pushers out of touch with the realities of science and patient care. "Conflict of interest is a meaningless term. It implies malignancy…If I have an interest in a company, I want that company to succeed, and that company is interested in me because of my objectivity and reputation and scientific integrity. If I compromise that, I'm of no use to anybody," says Thomas Stossel, professor of medicine at Harvard Medical School. According to Stossel, attempts made by universities to eliminate or reduce COI tend to stifle beneficial relationships. For example, Harvard's current policy prohibits a researcher from owning equity interest in a company and at the same time receiving money from the company. Stossel, who has equity interests in ZymeQuest of Beverly, Massachusetts. and Critical Biologics Corporation (CBC) of Cambridge, Massachusetts, believes that equity is a positive incentive for scientists. "So here we have a rule that's predicated on the assumption that […] the inventor intends to cheat, lie or steal. That's really disrespectful."

Harvard University is undergoing a comprehensive review of its COI policy, and it declined to comment for this article. At other universities, similar policies seek to protect the integrity of education of graduate and postdoctoral students—essentially, to prevent the university from becoming a branch of the research and development department of grant-funding companies. However, for an early-stage biotech company, offering stocks in lieu of cash for consulting is an easy way to stretch a startup budget. If such liaison makes later sponsorship of the same scientist's research difficult or impossible, it could become an obstacle to product commercialization.

Taking the industry's perspective, Paul Pomerantz, of the Drug Information Association, based in Horsham, Pennsylvania, notes, "No doubt the current scrutiny of the drug industry–academic relations has had a chilling effect on the dollars that are available to academia." He adds that strong ties are desirable to ensure clinical studies benefit from academic rigor.

There could be a completely different explanation for the findings reported in *Health Affairs*: a drop in industry funding associated with the economy. Lila Feisee, who is the resident expert on these issues at BIO, believes this is unlikely because academic–industry licensing activity is holding steady. "We do hear that maybe sometimes, when companies work with universities, it could be a smoother process. Some tech transfer offices are more savvy than others."

A consensus on the subject is unlikely, at least until more data are made available. Until then, Eric G. Campbell, the lead author of the *Health Affairs* paper and director of research at the Institute for Health Policy, Harvard Medical School, thinks that relationships with drug makers will continue to shrink. "Industry is less viable as a long-term funding source," he says. Its funding "tends to be small in amount and short in duration—and it's getting smaller."

Catherine Shaffer *Ann Arbor, Michigan*

SELECTED research collaborations

Partner 1	Partner 2	$ (millions)
Incyte (Wilmington, Delaware)	Novartis (Basel, Switzerland)	1,310
Nabi Biopharmaceuticals (Rockville, Maryland)	GlaxoSmithKline (GSK; London)	540
PanGenetics (Utrecht, The Netherlands)	Abbott Laboratories (Abbott Park, Illinois)	360
Trellis Bioscience (South San Francisco, California)	MedImmune (Gaithersburg, Maryland)	338

© 2010 Nature America, Inc. All rights reserved.

Investor volatility plagues drug companies on new Chinese exchange

Small and medium-sized companies listing on ChiNext, a new Chinese stock market launched six weeks ago in Shenzhen, have been buffeted by investor volatility and insider trading. Despite the roller coaster ride, proponents say the exchange will offer a much-needed exit for Chinese domestic companies that generate profits. Indeed, the four Chinese drug companies that have so far floated on the exchange all achieved high initial valuations (**Table 1**). Even so, it will likely be some time before the new exchange will offer a *bona fide* alternative to more traditional markets, such as the New York–based NASDAQ exchange, let alone attract companies and investors from beyond China's borders.

The ChiNext market, also called the China Growth Enterprise Market (GEM), has been in the making for more than eight years, says Hui-Hsing Ma of German venture capital (VC) firm TVM Capital, of Munich. Chinese domestic companies have been clamoring for such an exchange, says Ma, because government 'offshoring' restrictions have prohibited them from listing on foreign bourses. The long time from inception to launch is primarily a result of caution on the part of the Chinese authorities, who are anxious to avoid any risk of failure for domestic investors and are also ploughing money into startup funds for biotech (see **Box 1**).

A wild debut for China's newly launched stock market for small, high-tech enterprises in Shenzhen.

This cautious approach initially appeared to have paid off: during ChiNext's first day of trading on October 30, all 28 listed companies soared in price. Intense broker speculation even forced regulators to step in and suspend trading at one point. Over the next two weeks, shares rose by 10%, reaching a trading limit of 88.1 yuan ($12.9) for Sichuan Jifeng Agricultural Machinery Chain, located in the southwest Sichuan Province.

On December 4, however, state regulators ordered that same company to close its trading account over alleged share price manipulation. Investors panicked, share prices plunged and more than two-thirds of the listed equities ended far below their starting prices. Twenty of them dropped by the maximum daily limit of 10%, triggering another suspension of trading, with Tianjin-based Chase Sun Pharmaceutical among the ten worst performers.

ChiNext is centered on a range of high-tech businesses, from electronics and clean

Table 1 Four drug firms make their debut on ChiNext stock exchange on 30 October

Company name	Headquarters	Focus area	Flotation price	Gain or loss, 30 Oct. to 4 Dec.
Chongqing Lummy Pharmaceutical	Chongqing municipality	Drugs related to infectious diseases, cancer and gastrointestinal disorders	37.48 yuan	–5.5%
Anhui Anke Biotech	Anhui province	Developing, manufacturing and marketing of pharmaceuticals and healthcare products	47 yuan	–8.6%
Beijing Beilu Pharmaceutical	Beijing	Traditional Chinese medicine–based drugs for diabetes and other indications	39 yuan	–8%
Chase Sun	Tianjin municipality	Developing, manufacturing and marketing of pharmaceuticals and healthcare products	101 yuan	–9%

Source: Shenzhen Stock Exchange (http://www.szse.cn)

Details
Incyte has entered a collaboration and license agreement with Novartis for INCB18424, an oral JAK1/JAK2 (Janus kinase) inhibitor now in phase 3 for myelofibrosis, and INCB28060, an oral cMET (mesenchymal-epithelial transition factor kinase) inhibitor, poised to start phase 1 trials as a therapy for multiple cancers. Incyte will retain exclusive rights for development and commercialization in the US and Novartis outside the US. Incyte will receive an up-front payment of $150 million and an immediate $60 million milestone payment for the initiation of the European phase 3 trial that began July 2009, plus potential payments up to $1.1 billion for future milestones.
Nabi has entered a licensing deal with GSK for its experimental vaccine NicVAX, now in late-stage trials, to treat nicotine addiction. Nabi receives $40 million in up-front payment in a deal worth potentially more than $540 million in option fees and milestones.
Abbott will pay PanGenetics $170 million up front plus additional milestone payments up to $190 million for global rights to PG110, its fully humanized antibody to nerve growth factor. PG110 is now in phase 1 testing in subjects with osteoarthritis. If successful, Abbott plans to test the compound on chronic lower back pain, cancer pain and diabetic neuropathic pain.
MedImmune will pay Trellis for the global, exclusive license to its preclinical-stage antibodies directed against respiratory syncytial virus (RSV) in a licensing deal worth up to $338 million. Trellis's RSV antibodies were identified using its CellSpot discovery platform, a high-throughput screening of human B cells that allows isolation of rare human antibodies.

© 2010 Nature America, Inc. All rights reserved.

IN brief

GM crop biosafety lab folds

A fully equipped laboratory for studying pathogen-resistant transgenic plants will close its doors by the year's end. The International Centre for Genetic Engineering and Biotechnology (ICGEB) Biosafety Outstation in Ca'Tron di Roncade, Treviso, Italy, was set up to study potential risks concerning genetically modified crops and plant pathogens of importance to the developing world. The outstation's facilities, part of the ICGEB, were refurbished with financing from Treviso-based Cassamarca Foundation, supported by banking group Unicredit. But the bank's financial woes have prevented the foundation from renewing the €4-million ($5.7 million), 5-year contract, says Mark Tepfer, leader of the outstation's Plant Virology group. Tepfer will transfer some his projects to his permanent appointment at the French National Institute for Agricultural Research in Paris. "I'm fairly optimistic that we'll find a way to continue," he adds. The ICGEB operates under a treaty signed by 59 countries within the United Nations system to conduct research and education in biomedicine, crop improvement, environmental remediation and biopharmaceutical and biopesticide production throughout the developing world. ICGEB administrator Decio Ripandelli hopes to shift some of the outstation's research and education programs to the Trieste and New Delhi groups. Ripandelli says he lobbied the Cassamarca Foundation to put the facilities, including a high-containment greenhouse, into a "pharmacological coma" to avoid restarting from scratch but the foundation is noncommittal. Ripandelli says, "It's really a pity and a scandal if the facilities are not used." *Lucas Laursen*

Plant genomics' ascent

Grants supporting plant genome research in the US have reached an all-time high. Over 2009, the National Science Foundation (NSF) doled out nearly $102 million, the largest sum since the annual grant program began in 1998. The funding aims to increase understanding of plant gene function and the interaction of plant genomes and the environment. "This funding lets you tackle bigger problems," says David Salt, a former grant recipient and plant biologist at Purdue University. "It lets you devise more integrated and collaborative projects." The NSF chose 32 projects focused on "economically important crop plants" ranging from West African cultivated rice to poplar trees, according to the foundation. The largest award, worth more than $10.4 million over four years, went to a proposal to help complete the international effort to sequence the tomato genome. James Giovannoni at Cornell University's Boyce Thompson Institute for Plant Research in Ithaca, New York, leads the project. The NSF also chose for the first time a switchgrass research project. With a grant worth more than $4.5 million, Thomas Juenger and his team at the University of Texas at Austin will explore over the next four years how switchgrass responds to drought and other stresses caused by climate change, to expand the knowledge needed to develop switchgrass as a biofuel crop. *Emily Waltz*

© 2010 Nature America, Inc. All rights reserved.

Box 1 China sets up raft of state-backed VC funds

In an unrelated but perhaps even more significant move, the Chinese government has, in one stroke, set up 20 new venture funds to invest in technology start-ups, including biotech, as a prime focus. The venture funds, worth up to ¥8 billion, are sponsored by the National Development and Reform Commission, a leading economic development body.

The funds will be administered through seven provincial-level governments, starting with Beijing, Jilin, Shanghai, Anhui, Hunan, Chongqing and Shenzhen. More provinces are expected to follow suit. "This is a new signal that the government is starting to look for new economic engines," says Zheng Yufeng, senior manager of Investment Banking Division of Beijing-based Zero2ipo Group. Zheng points out that nearly 20% of this fund has been allocated to biotech and pharma firms.

But Hui-Hsing Ma of TVM Capital believes the impact of these funds on China's biotech industry is likely to be slower and less transparent than private stock exchange investment. Few Chinese provincial governments are sophisticated biotech investors and they naturally tend to be biased towards their own regional agenda, she says, although venture capitalists looking for Chinese deals would welcome them as co-investors.

energy to software, robot design and biotech. Four biotech/pharma companies are listed on the exchange (**Table 1**)—Chongqing Lummy Pharmaceutical in the Chongqing municipality, Anhui Anke Biotechnology in Anhui province, Beijing Beilu Pharmaceutical and Chase Sun in Tianjin municipality—as developers of innovative drugs rather than the traditional generics or device firms. To qualify for ChiNext, companies must generate at least ¥50 million ($7.4 million) in the previous fiscal year, and made profits of at least ¥5 million in that year, or ¥10 million in two years. This excludes most high-risk, development-stage biotech companies, which have no products on the market.

Despite the restrictions on company eligibility, the immaturity of the Chinese market and signs that local investors are looking to turn a quick profit, several analysts are upbeat. The new exchange will provide not only investment opportunities but also a new financing vehicle for fledgling Chinese companies.

"Despite irrational near-term market performance, long-term prospects are rosy," says Zhang Yuanda, deputy secretary-general of China Association for SMEs (small and medium-sized enterprises). Indeed, although Ma says it is very early days, the high initial valuation achieved "is a definite incentive for this exit/financing option," indicating that ChiNext could be an important source of capital for the Chinese biotech industry at least. That said, when Chinese biopharmaceutical firm Nuokang announced in November it wanted to raise up to $69 million from an initial public offering (IPO) (it raised $32.9M in the end), it chose to file it on the NASDAQ rather than in China, even though Nuokang is profitable and thus would easily qualify for a ChiNext listing.

The strict eligibility requirements for ChiNext also offer advantages, according to Ma. "This makes it less risky than other pre-revenue exchanges in the region, like the Tokyo's MOTHERS and Korea's KOSDAQ," says Ma.

However, the need for profitability means that the exchange will be unlikely to provide another source of funds for cash-hungry Western biotech firms or for their backers looking for a better-value exit than selling up to big pharma or a rival biotech firm. The regulatory, financial and language hurdles of the Chinese markets continue to be too daunting: "China is a difficult place to figure out", says Drew Senyei, managing director of VC firm Enterprise Partners in La Jolla, California. In any case, many venture capitalists think that a return of the US or European IPO market is the only real hope for cash-hungry biotechs. "We are cautiously optimistic about the US IPO market for biotech, but are much less so about the other markets," says Jamie Topper, general partner at Frazier Healthcare Ventures, of Menlo Park, California.

Markus Hosang, general partner at VC firm BioMedPartners in Basel, is also unenthusiastic about the idea of a Chinese float. "We do look for M&A exits for our companies on a global basis, but an exit through IPO would probably have to take place on one of the European stock markets," he says.

David Seemungal of Cubase Consulting, London, reckons the more successful Western biotech companies will probably shy away from the Chinese market until they perceive that more robust intellectual property enforcement is in place. And far from benefiting from China's accelerating move into biotech, the West could even lose out, says Seemungal. "There could be a brain drain of biotech expertise from ailing Western biotechs to Chinese-based biotech companies newly established on the back of Chinese state funding," he says.

Peter Mitchell *London, and* **Fu Jing** *Beijing*

Roger Beachy

Plant scientist Roger Beachy has joined the Obama administration to lead the National Institute of Food and Agriculture (NIFA), the new research funding arm of the US Department of Agriculture (USDA). Beachy, whose research led to the first transgenic crop, was previously the long-time head of the not-for-profit Donald Danforth Plant Science Center in St. Louis. Emily Waltz talks to Beachy about his plans for the new agency.

Did you have any idea you were on President Obama's short list for this job?
I had no idea. Rajiv Shah, who had just been appointed USDA's chief scientist, attended a meeting in St. Louis and during his visit he quizzed me about what I thought NIFA should be like. A month later he called and asked if I would consider taking the job as director.

Do you think you were hired because the current administration wants to push the agbiotech agenda?
Not at all. They wanted a scientist who has a reputation for having accomplished both fundamental and applied science and has a grasp on the importance of international programs. The best science I have done in my career has not been biotech; it has been what I have taught about virus structure and pathology and how viruses move between cells.

What are society's most urgent agricultural challenges, as NIFA sees them?
Sustainable food production and nutrition, readiness for climate aberrations that will impact productivity and developing renewable options like biofuels and industrial and pharmaceutical materials. To address these challenges, we will create sub-institutional structures within NIFA. One of the institutes would address biofuels, climate and environment; another would address food safety and nutrition; a third would address food production and sustainability; and a fourth institute would focus on youth, families and communities.

What opportunities does your new position offer that your previous job didn't?
I'm a scientist and I'd like to see agricultural science benefit humankind. This job gives me a far greater opportunity to do that than my previous roles. We need a new generation of scientists who understand the importance of the environment, of sustainability, food production, biofuels and climate, and I'm not sure that has been as much of a focus at the USDA as it should have been.

How will NIFA differ from its predecessor, the Cooperative State Research, Education and Extension Service (CSREES)?
With NIFA there will be a greater focus [than under CSREES] on competitive grants, and greater linkage between fundamental and applied research with extension and education. We want to ensure that the knowledge we gain from research reaches farmers and consumers; from the lab to the field to the fork. Our agency will be unique in that regard. The NIH [National Institutes of Health] doesn't have the same capability of going from the lab to the bedside.

How will you accomplish these linkages?
We will request that a significantly greater percentage of the research grants that NIFA awards include a component for extension or education. In the past, about 25–30% of our grants included these components. We'd like to double that.

Do you think that the financial support from Monsanto at the Donald Danforth Plant ScienceCenter will affect how you form relationships with industry at NIFA?
No. As president of the Danforth Center I encouraged relationships with private companies, including Monsanto. But it should be understood that those relationships did not result in significant influences over the mission of the Center. It's unfortunate that some people think that that relationship has tainted me in some way, although I guess it's not unexpected.

Will more money find its way into research grants or will we simply see a reshuffling of funding priorities?
We had a budget increase for a competitive grants program, called Agricultural and Food Research Initiative (AFRI), this year from $201 million to $262.5 million, which suggests that the Obama administration is keen to invest more heavily in agriculture research. And the farm bill states that we are eligible for up to $700 million in AFRI funds [2008–2012 period]. That's a good start, but we need more than a billion dollars per year to meet the major societal challenges that involve agriculture.

"Without additional support, there will likely be few genetically enhanced crops developed by public sector researchers in the marketplace in the near future."

Why do we see such an emphasis on transgenic strains of major crops rather than other crops that would benefit small-scale farmers and consumers?
There is relatively little profit in minor crops like blueberries and sweet potatoes compared with the large commodity crops. So the major seed companies aren't very interested in developing them; that is left to the public sector and small seed companies. And while public sector science is putting a lot of effort into researching these smaller crops, the cost of navigating the regulatory process is so high that it essentially eliminates public sector participation in commercialization. Noncommercial researchers also lack the expertise and infrastructure to provide regulatory authorities with the necessary documentation for regulatory approval. Without additional support, there will likely be few genetically enhanced crops developed by public sector researchers in the marketplace in the near future.

Are you going to attempt to change the regulatory process so that these minor crops can make it to the market?
In the early days of agbiotech, regulations were

© 2010 Nature America, Inc. All rights reserved.

fairly minimal, which kept development costs low. The safety of a product was judged on the product itself and not the method used to develop it. Regulatory agencies have lost some of that focus in the past ten years. Now crops made with genetic intervention are viewed through a different lens than those made by classical breeding. I am very interested in having a regulatory structure that is science based and gets back to what we originally had.

"I think it's important that we stop talking only about risks and talk more about risk-benefit analyses."

How will you go about making these regulatory changes?

I've been on the job for four weeks, so I don't have an answer yet. But it is an interest of mine. NIFA is not a regulatory agency and is not part of the regulatory process, and to put a lot of immediate effort into changing the regulatory structure before we have a sense of how much need there is for change would not be prudent.

Can NIFA hope to achieve an impact beyond the US?

In the past, the USDA supported a larger number of foreign students and other trainees who would attend our agricultural universities and then return to their home countries to implement their knowledge. The resources for such programs have shrunk in recent years. In the next few months, we will create a Center for International Programs, reporting to me, that will seek to rebuild international partnerships based on local agriculture, rather than imported goods.

Some scientists have criticized the USDA for becoming conservative in the kinds of crop research it supports. Do you think this is true?

I agree there has been a narrow focus and it's partly because Congress has gotten involved in telling the USDA what to fund. That hasn't made it easy to be more exploratory in research. During my tenure we expect to award larger grants that are longer in term. We hope this will engage a broader range of scientists and engineers who haven't traditionally come to USDA for funding. For example, we would bring together biomedical researchers, plant biologists and extension agents to work on increasing the nutritional value of food.

What will it take to get grant managers at the USDA to think differently and direct funding toward a broad range of scientists?

When I arrived, I was impressed by the willingness of the management team to consider doing things in a different way. They are ready for change. But we may also bring in a few people as advisors or staff to help stimulate the change.

What are your ideas on how to provide accurate, science-based information that the public will actually read?

Communication of any type of science in lay language is terribly important. The USDA has not always kept good track of the impact of its research. Instead, we leave it to the universities to publicize discoveries. We need to find more proactive ways to let people know that we are part of those discoveries.

Will NIFA fund research that examines the potential risks of biotech crops?

We've had more than 15 years of successful deployment of biotech crops. That history alone tells us a lot about how safe transgenes are under current regulatory guidelines. I think it's important that we stop talking only about risks and talk more about risk-benefit analyses.

So if a scientist applies for a grant to study only the risks of a crop, is that person out of luck?

If there is a legitimate concern about the safety of a product, absolutely there is an opportunity for support from NIFA.

Emily Waltz, Nashville, Tennessee

© 2010 Nature America, Inc. All rights reserved.

Up in a cloud?

Cloud computing offers solutions for companies wrestling with large-scale data sets, but security issues will likely continue to restrict its use to precompetitive or nonconfidential data. Clare Sansom reports.

In October, servers operated by the perhaps aptly named Danger group, a subsidiary of Seattle-based Microsoft, suffered a complete 4-day outage that caused many of Danger's customers, all heavy users of T-mobile's Sidekick phone, to irrevocably lose terabytes of valuable data. Daniel Eran Dilger, writing in the online news magazine *Apple Insider* (http://www.appleinsider.com), claimed that this exemplified "the dark side of cloud computing." Other commentators, however, still maintain that cloud computing is poised to become the dominant model for data-intensive computing. As a data-intensive industry, biotech stands to gain or lose by how, and to what extent, it takes advantage of this much-hyped concept.

The cloud computing model, which provides users access to data analysis services over the internet at remote locations, offers cash-strapped biotechs and less agile big pharma companies clear advantages in terms of cost, efficiency and flexibility. The problem is, however, that despite attempts to develop software houses into 'middle-layer' companies helping specialist industries access the cloud and buffer their data, most corporations view this model as too risky to host data integral to their business. For the foreseeable future, cloud computing is likely to find most use for early stage gene and sequence data increasingly viewed as precompetitive research.

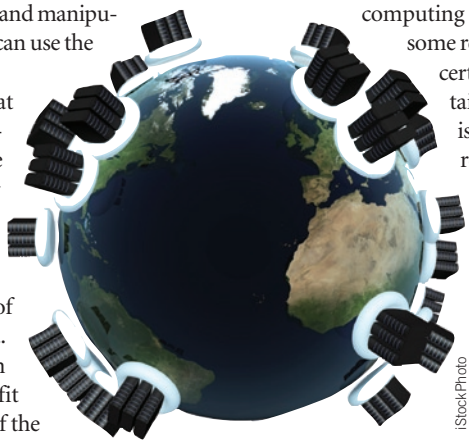
Cloud basics

'The cloud' refers to the delivery of computer resources as a web-based, scalable service, as opposed to the use of local workstations or clusters. Some say it is just an extension of the internet: the term itself was derived from the use of cloud symbols to depict the internet in schematic network diagrams. However, the basic service concept dates from the 1960s, before the internet was conceived, when artificial intelligence pioneer John McCarthy, inventor of the Lisp programming language, predicted that "computation may someday be organized as a public utility." And, like public services, cloud computing is provided by 'household names': Google, Yahoo, Amazon, Microsoft and AT&T.

The concept of cloud computing resembles that of grid computing, which predates it by some years. Grids divide a program into many pieces and distribute these among clusters of machines. It is, therefore, only appropriate for high performance, central processing unit-intensive tasks, and setting up on a grid system requires considerable expertise. Although academic groups and public consortia make extensive use of the grid, it has been relatively little taken up by the biotech sector.

In contrast, the virtues of cloud computing include flexibility and ease of use. Both the startup company setting up a single machine for a short-term project and the multibillion-dollar company storing and manipulating terabytes of data can use the cloud.

Those biotechs that are currently exploring cloud solutions cite several advantages, key among them cost savings and improvements in processing speed and handling of large volumes of data. Small companies in particular can benefit from transferring out of the company some expertise in, for example, systems and database administration.



iStockPhoto

Software as a service, aka cloud computing, offers advantages to both large and small companies.

"Why would you want to install a server when you can do the same job by accessing a cloud provider via a browser with a few add-ons, and have no worries about installing and maintaining databases, and keeping up with software licenses?" asks Keith Chessell, CEO of Solcom, a consultancy based in the Isle of Wight, UK, that provides cloud computing solutions to biotech companies.

The availability of virtual servers effectively 'on demand' and only when required can help both startup companies with little or no prior investment in computational infrastructure, and companies whose demand for high-performance computing is 'spiky'. "If you need a 10,000-node cluster for 2 days, and then not again for months, it will be far more efficient to 'rent' that CPU power from a cloud provider than to set up an in-house cluster," says Charles Wuischpard, CEO of Linux specialists Penguin Computing, based in San Francisco. Penguin provides access to its cloud-based, massively parallel supercomputing cluster, Penguin On Demand (POD), for biotech clients.

This 'software on demand' or 'software as a service' paradigm is as applicable to mundane computational tasks as it is to scientific computing. Google docs and Googlemail are already replacing desktop office systems and in-house e-mail packages. Biotechs can benefit from handing over these tasks, leaving their computing teams free to work on scientific and technical problems. "We let Google do what they do best, so we can concentrate on what we do best," says Todd Pierce, vice-president for information technology at biotech giant Genentech of S. San Francisco, California, now part of Basel-based Roche.

Trade-offs

Despite these perceived advantages, most biotech companies are not yet embracing cloud computing fully; indeed, there are some reasons why they almost certainly never will for certain types of data. These issues relate to security, reliability and intellectual property rights. Many companies are reluctant to allow their sensitive data—or even any data—to leave their own proprietary firewall-protected networks. Rick Franckowiak, director for systems engineering at Johnson & Johnson Pharmaceutical Research & Development in New York, explains the trade-offs. "We are not willing to compromise on data security for efficiency…but the risks are very much tied to how we implement our solutions."

And even cloud advocates, like Paul Miller, UK-based consultant and founder of The Cloud of Data, are clear that the use of third-party data centers can be problematic. "In many enterprise contexts [including biotech], it is unrealistic to expect a wholesale embrace of the public cloud," he says. Several cloud providers are now seeking to work with companies with security concerns to develop cloud-type solutions that remain inside company firewalls, in so-called 'private clouds'.

This is particularly pertinent as governments and pharmaceutical regulatory bodies tighten standards for data security, particularly for clinical data. Indeed, cloud providers that can agree to work within such regulations have the best chance of attracting business within what they

© 2010 Nature America, Inc. All rights reserved.

see as a lucrative sector. But take-up has still been patchy. Some companies are on track to migrate much of their computational resources to public and private clouds within a few years; others are embracing mixed solutions, setting up protected private clouds or moving only less sensitive data; and others see no value in it at all.

Early adopter

One of the first life sciences companies to investigate cloud technology—as long ago as 2003—was pharmaceutical giant Merck of Whitehouse Station, New Jersey. At that time, Merck was outperforming most of its competitors in generating genotype, gene expression and clinical trait data. Eric Schadt, who worked for Seattle-based Rosetta Inpharmatics, then a Merck subsidiary, remembers, "Merck needed to integrate terabytes of data, and they built one of the largest computer networks in either academia or pharma to do so. This cost millions of dollars and eventually reached 10,000 nodes, but with the advent of second-generation sequencing it was clear this still would not be enough to meet their data analysis needs." Merck could have decided to continue, growing even larger in-house clusters of even faster processors, but the cost in infrastructure—space, air conditioning, water cooling, systems staff—was becoming unsustainable. They chose to explore migrating their computing into the cloud service that had just been launched by Amazon of Seattle. "When Amazon first offered this service, it was very expensive, but it has recently become more cost-effective than further extending the large in-house cluster," says Schadt.

When, early in 2009, Merck closed Rosetta and with it much of its genomics operations, the data it had generated were inherited by a new not-for-profit, open-access medical research organization, Sage Bionetworks, in Seattle. Sage inherited the Merck/Rosetta Amazon cloud presence as well as its 10,000-node network and is now also exploring collaborations with Microsoft and others to investigate further cloud services. Schadt, who now works with Sage alongside his role as CSO at sequencing company Pacific Biosciences in Menlo Park, California, explains "Sage will ultimately need access to the cloud because genetic data is pouring off sequencers at a rate that almost no company has the in-house facilities to handle. Security is not really an issue here, because these data are early stage: companies will ultimately be able to select genomic data and pull it out of the Sage cloud for further analysis behind their firewalls."

Pacific Biosciences is also exploring cloud solutions, although at an earlier stage. The company's first commercial sequencing machines, due to be released in 2010, will produce sequence data at a roughly similar rate to the Illumina Genome Analyzers and Basel-based Roche's 454 FLX machines, but Schadt expects this to grow until, in a few years, they are producing "hundreds of gigabases [of sequence] every day." He and his colleagues are developing methods to provide sequence data storage and analysis services to Pacific's academic and biotech customers. He seems confident that security problems can be solved. "Amazon and Google have built their reputations on the need to keep data secure. The fact that major pharma companies are confident enough to put even some of their data in these companies' clouds implies that the cloud providers have the issue in hand," he says.

Hybrids

Yet, as Microsoft's recent problems prove, no provider, not even the largest, may be considered completely reliable. Savvy biotechs may prefer to hedge their bets by working with several providers, and 'middle-layer' specialist software companies can help them do so (**Table 1**).

A good example is rPath, based in Raleigh, North Carolina. "We package applications with operating systems, middleware, libraries and everything else they need to run so they can be easily deployed to and moved between a variety of cloud and grid-based platforms," says Jake Sorofman, vice president of marketing at rPath. "Our solutions allow applications to be ported on demand from Amazon to Rackspace or Globus in a matter of minutes," adds Sorofman. "Companies working with us can simply redeploy to another provider if there is an outage, provided their data is adequately backed up." Interestingly, however, few bioinformatics companies are moving into this area. Almost all middle-layer companies are computer specialists with clients in a variety of data-intensive disciplines.

One bioinformatics company that is bucking this trend and offering a middle-layer cloud solution is Seattle-based Geospiza. Companies that use Applied Biosystems' next-generation SOLiD sequencing method but do not have their own sequencing laboratories can now benefit from a pioneering cloud-based partnership between Geospiza, ABI, Life Technologies and Amazon. Geospiza president, Rob Arnold, explains how the collaboration works. "A researcher with no direct access to sequencers sends DNA samples to a service lab [that uses ABI sequencers], the lab does the sequencing and the sequence data is directly transferred to a cloud system for storage and analysis. Our software system includes an e-commerce style front end that allows researchers to choose and control their analyses directly, and samples and data are tracked through the process life cycle, from order placement to data visualization." All data storage and analysis can be performed using cloud systems, both Geospiza's own managed clouds and Amazon's public cloud.

Table 1 Cloud service providers

Company	Services offered	Sectors/clients
Amazon	Application and data hosting Scalable public and private clouds	Big pharma (Eli Lilly) Large biotech (ABI) Small/startup (IXICO) Public sector (Sanger Centre/EBI)
Globus (Argonne, Illinois)	Grid services	Public sector (US Department of Energy (DoE))
Google	Software as a service Document hosting, e-mail	Large biotech (Genentech) Small/startup (De Novo)
IBM (New York)	Application and data hosting	Big pharma (Johnson & Johnson)
Microsoft	Application and data hosting	Big pharma (Johnson & Johnson) Large biotech (Genentech) Small/startup (Spirogen, London) Public sector (Sage)
Rackspace (London)	Application and data hosting, specializing in small companies	Small/startup (Spirogen)
Star Internet	Data hosting, networking	Small/startup (Science Warehouse)
Cycle Computing (Buffalo, New York)	Implementing Condor Grid workload management	Big pharma (Johnson & Johnson)
Geospiza	Sequence analysis and hosting	Large biotech (ABI)
rPath	Automating application development and maintenance	Public sector (US DoE)
Solcom	Cloud project management and software	Small/startup (Spirogen)
Univa UD (Chicago and Austin, Texas)	Managing public and private clouds	Small/startup (Pacific Biosciences)

© 2010 Nature America, Inc. All rights reserved.

© 2010 Nature America, Inc. All rights reserved.

Leader of the pack

Amazon currently seems to be the first choice mainstream cloud provider for much of the industry. Companies using or exploring their services include, besides the sequencing companies, *in silico* drug discovery company De Novo Pharmaceuticals of Cambridgeshire, UK, image analysis company IXICO of London, large pharmaceutical companies, including Indianapolis-based Eli Lilly and Johnson & Johnson of New Brunswick, New Jersey, and public sector genomics institutes. Amazon's cloud services offer two principal advantages over its competitors: extreme flexibility, and a level of security that meets the highest standards now available. Amazon's aptly named Elastic Compute Cloud (EC2) offers a fully and transparently scalable range of services. At one end of the scale, a small company, or an academic group, can pay a few cents an hour to rent a single "virtual Linux box" and associated software in Amazon's cloud, even paying with a credit card. At the other end, multinational companies can use a similar environment to set up virtual supercomputing services without the overheads associated with a large on-site cluster. Eli Lilly's system uses the Amazon Virtual Private Network, which enables the pharma giant to reap the benefits of cloud computing while keeping its existing firewall and other security systems intact.

Amazon's commitment to data security, and to meeting the specific security needs of life sciences companies, is exemplified by its compliance with the Health Insurance Portability and Accountability Act (HIPAA) in the US. Besides regulating some health insurance policies, this law establishes standards for the privacy and security of health-related data. Its complex directives have not been universally welcomed by US medical researchers, but compliance by companies with no health focus can indicate that they take the security concerns of the biotech industry seriously. "Amazon has published white papers on security and stated that they believe the company will meet the HIPAA requirements; this was one of the principal reasons why we selected their Elastic Compute Cloud," says Norman Taylor, an IT consultant with IXICO.

IXICO's core technology provides image analysis and data management for clinical trials; its customers include large pharmaceutical companies, biotechs and academic groups worldwide. "Our developers are currently testing a pilot version of TrialTracker, our online data management system, on the Amazon cloud," says Taylor. "We believe that this solution will suit some of our clients but know that cloud is not the answer for all client needs."

Resistance is futile

Whereas these examples show drug companies of different sizes exploring cloud computing for their computational needs, in many instances, companies simply will not require the kind of data handling capabilities that cloud computing offers. As David Brown, informatics manager at the Cambridge-based antibacterial startup Prolysis, puts it, "Our data storage requirements are modest." Others, such as Evotec, a drug discovery company based in Hamburg, Germany, reject the concept as inherently insecure. Shane Pereira, external marketing communications manager at Evotec, explains, "Biotech and pharmaceutical companies have millions of dollars' worth of revenue resting on their scientific and clinical data…. Such data has to be totally ring-fenced and secured, and any breach in protocol could cause catastrophic loss of revenue."

But even companies that reject the cloud for their own research may not be completely isolated from its reach. Few researchers can say that they will never access public-domain bioinformatics resources, and fewer still will never need to order consumables. Both these services are now being provided by institutes and companies that employ cloud technology. Data from the Ensembl collection of eukaryotic genome databases, a joint project of the Wellcome Trust Sanger Institute of Hinxton, UK, and the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) of Heidelberg, Germany, have recently been ported to the Amazon public cloud. "We have always made Ensembl's data freely available to all, including companies," says Ewan Birney, who leads the EMBL-EBI nucleotide databases group. Other EMBL-EBI and Sanger databases will be ported to both public and in-house cloud providers.

Public genomics institutes such as these have data transfer and processing needs that dwarf those of all but the largest pharmaceutical companies. And these are still growing: the 1000 Genomes Project, a "deep catalogue of human variation," co-chaired by Richard Durbin of the Sanger Institute and David Altshuler of Harvard of Cambridge, Massachusetts, will ratchet these requirements up yet another notch. Cloud computing offers these institutes significant savings in the cost of and time taken by data processing. "We can put a complete Illumina data set in the cloud and analyze it cheaper than and just as quickly as, we can in-house," says Phil Butcher, an informatics expert at the Sanger Institute. As all these data are in the public domain, there are fewer security issues with them than with data generated within the biotech industry.

However, the Sanger Institute does still face a problem with the public cloud. "Our biggest hurdle is data transfer into and out of the cloud. The internet is not designed for moving terabytes of data around…sometimes it is most efficient and cost-effective to put it on a hard drive and send it to the cloud provider by Fedex," says Butcher. Simon Twigger's group at the Medical College of Wisconsin, Milwaukee, offers proteomics data analysis using Amazon's cloud through the ViPDAC (Virtual Proteomics Data Analysis Cluster) facility. Again, ViPDAC is freely available to all, including commercial clients, and Twigger keeps no records of user affiliation.

The procurement companies that the biotech industry relies on to supply consumables are also increasingly working with the cloud. Science Warehouse, based in Leeds, UK, a web-based procurement company, works principally in the science sector, mostly with academia but increasingly with biotech. The company has developed a proprietary cloud system for its databases, renting servers from Star Internet, based in Gloucester, UK, and developing all its software in-house. "Purchasing information can be sensitive, although not as much so as data used for drug discovery…but we have invested a lot of resources in making our proprietary software as secure as possible," says Science Warehouse' sales and marketing director, Jonathan Betts.

Making the right choice

If an overall mood can be discerned it is probably one of cautious optimism that many companies' business models can be adapted to include effective use of the cloud. It seems that there is no one business model for cloud computing that will suit all biotech companies. Few executives and scientists are completely dismissive, but none is as visionary as John Wilbanks, vice president for science at Creative Commons. Wilbanks sees cloud computing as becoming ubiquitous as "a service that people sell to make networking easier," with data increasingly in the public domain, shared through common standards.

In reality, biotech is heading for a 'mixed economy', incorporating public and private clouds as well as traditional, in-house systems. The security issue is still very much a live one, although opinions differ as to exactly how important it is. Johnson & Johnson's Franckowiak perhaps spoke for the industry when, after careful study, he concluded that if chosen wisely, cloud offers a partial solution to his company's computing needs. "Cloud computing can solve the problem of overtaxed internal resources...but not, at least not yet, for the highest-risk applications involving sensitive data," he says.

Clare Sansom, London & Cambridge

bioentrepreneur
from bench to boardroom

Backing your brand

Sue Charles & Chris Fisher

Branding is one of the most important and powerful aspects of building a business, but it requires delicate maintenance to keep it working for you instead of against you.

© 2010 Nature America, Inc. All rights reserved.

Truly successful businesses—biotech or otherwise—not only need a great product offering but also must establish their position as leading brands that are immediately recognizable in their sector. 'Branding' can be confusing, however, with companies and customers having their own concept of what a 'brand' really is.

In the biotech sector, companies often neglect brand identity because, unlike other industries, they lack tangible products. As most energy is expended on product development—and convincing investors to give their money to finance further product development—it's easy to let branding drop down the list of priorities.

But branding is not just limited to products; it can be essential for effectively communicating the value of your company's equity to stakeholders. This is especially critical in an industry in which perceptions and confidence are major drivers of value. And it remains true whether the economy is expanding or contracting (**Box 1**).

Box 1 Branding in a recession

Does the economic downturn have an effect on how much businesses spend on their brand? It can, but customer loyalty is critical during a recession; you still need to communicate with all your stakeholders during difficult times. In this climate, it is important for companies to balance their communications budgets—to spend nothing sends the wrong messages, yet to overspend can be more damaging. Take annual reports, for example, which are a key branding tool; we are now seeing a trend of companies moving away from very expensive and glossy publications to something simpler and more understated. This can be a highly effective approach if used in a smart way.

It is vital to ensure that your brand is in the best position when the economic climate changes for the better. At the same time, downturns can offer many possibilities to reposition a company so it gains major competitive advantages when the market swings back. For example, Biota (Melbourne, Australia), a biotech that develops anti-infectives, has been able to make a good recovery because of the H1N1 flu pandemic. The company has emerged from the industry downturn and has positioned itself nicely for its next strategic phase. Within our own company, College Hill Life Sciences (London), we chose to invest in expansion into the United States to internationalize our offering in 2009, at the height of the downturn. Now that we are coming out of the gloom, this positioning is paying dividends.

Your first move

So, how do you create a brand identity? You should start by coming up with the right name and logo. Do you really want to call your company Birmingham Proteins if you're planning to broaden your horizons to antibodies and RNAi in five years' time? Similarly, you should be wary of cultural mores. Names in one language might not translate well in others; thus, naming your new startup in Naples *Gen Italia* might not be optimal if you are seeking to raise serious money from English-speaking investors.

Sue Charles is managing partner and Chris Fisher is associate partner, creative director at College Hill Life Sciences, London, UK.
e-mail: sue.charles@collegehill.com or chris.fisher@collegehill.com

Choose a name and logo wisely and make it link to what you do. If you have a strong brand identity, people should be able to instantly recognize what your business does, what it offers and what they can expect. For example, Genentech (South San Francisco, California) is a global brand with a strong personality built up through delivering on its promises. In addition, Genentech has one of the most memorable and sought-after website addresses in the industry—http://www.gene.com. If people do not know your company name, cannot relate to it or if it means different things to different people, then you have failed in creating your brand identity.

Choosing a logo is influenced by many factors, including your competitor's logos and your general business environment; the technical practicalities; the complexity of the overall visual identity and the role the logo plays; the personality of the company; and how the logo will be influenced by long-term business objectives.

But this is just the beginning. Successful branding is not achieved by just a logo or name. You must capture a company's or product's purpose and successfully communicate it to stakeholders through images, words and dialog. It's your promise to your audience—you must tell them what they can expect from your company and what differentiates you from your competitors.

Brand elements

You can only build a brand once you have a clear picture of what your business is. What are your goals and aspirations? How do you want audiences to perceive your company? A startup biotech should always begin by developing its identity and looking at its positioning. Do your market research—what are you offering, where do you sit in the market and where do you want to go? To answer these questions, a brand agency will use appropriate processes, workshops and market research to establish the key information and messages

that will feed into building your brand.

Successful companies have a clear positioning and identity, know where they want to be in the future and know how to develop and use their brand to meet business objectives. Don't ever try to be everything to everyone—you won't be. And be true to both your vision and reality (**Box 2**).

Anecdotal evidence suggests that a CEO's reputation accounts for up to 40% of a company's reputation. For successful branding, it is crucial that CEOs lead by example and capture the essence of the brand through their words and actions—the person at the top should always be seen as the embodiment of their company's brand. Genentech CEO Arthur Levinson is a notable example: he made the scientist-as-chief-executive persona work for him and the company, which was always seen as innovative and slightly away from the pharma mold—a positioning that might not have been possible with an ex-pharma man at the top. (In April, following Genentech's $47 billion merger with Roche (Basel, Switzerland), Levinson left day-to-day operations after decades at the helm, though he still leads the board of directors.)

The CEO of a company needs to understand what the brand is, all the influences on it and what it means to different stakeholders. Additionally, the CEO's behavior can have a dramatic effect on how the brand is perceived. Indeed, from the top down it's vital for all company employees to believe in and understand the brand.

Box 2 Essential elements

Although many scientific entrepreneurs might shudder at the 'touchy-feely' world of marketing, it is a vital part of any business—whether the company deals in lollipops or locked nucleic acids. Below are the essential elements to consider when trying to build a successful brand.

Vision, mission and values. These will make clear your long-term strategic goals, your tactics to achieve these goals and what you stand for and believe in.

Personality and behavior. Are you serious, corporate and financially led? Or are you more visionary, approachable and diverse?

Brand proposition. What is your brand promising to your audiences, and what will they experience?

Name and logo. Any company needs a recognizable written name and graphical logo (preferably not a hackneyed DNA helix!).

Visual identity and framework. This visual platform will serve as both the underlying structure and the styled content of your communications.

Key messages. These make clear who you are, what you do, what you offer and what people can expect from you.

Benefits and labeling (of technology, products and services). You should use persuasive and clear terminology that is understood by your audiences and that conveys the purpose and benefits of your product.

Market positioning. Be sure to create ownership of a defined market for your company that allows it to prosper alongside competitors. There is an audience that will choose you.

Target audience. Identify the type of people you want to engage with, who will also benefit from knowing about you.

Brand consistency. From corporate stationery and presentations to e-mail signatures and phone greetings, everything about your brand should be consistent and managed.

Brand guidelines. How should the brand be managed, what design templates should staff be using, where and when should a logo be used and what is the company's style and tone? All these questions must be outlined in an easy-to-use reference manual.

Follow-through. You must deliver on your promises. All audiences will have a memory of you through their experiences; make sure it is a great one.

Brand strategy

A strong and successful brand drives a business forward, represents the goals of the stakeholders and creates loyalty and confidence. This is accomplished through an effective brand strategy (**Box 3**).

The key is to have clear and consistent communications that deliver what you promise. For example, which business is more likely to have a brand that resonates with its audience: one that is poorly organized with bad communications and technology no one understands or one that always delivers and has clear and effective communications? In biotech we are predominantly selling the 'hopes and dreams' of tomorrow, and the better you can convey that, the more successful your brand will be.

Your brand must always be a true reflection of your company—you cannot call yourself a global company if your business consists of just two people working in a small office 100 miles north of London. Unless your goal is to rapidly grow your business across the world, there will be a disconnect between the brand and the audience's experience of it, and the brand will become irrelevant. You should find your place in the market and ensure your branding reflects both your current status and your near-term vision.

Successful companies quickly embrace the importance and value of brand communications. They don't have to be the loudest—it's not about volume; it's about projecting the right image with measured communications to help attract investors, customers and employees. It is critical for all employees to be on board, from the CEO to the frontline staff. Everyone in your company needs to play a part in creating an effective brand.

If you have the resources, working with the right brand communications agency can also help because the agency has experts with the right skills and know-how to create, build and drive your brand. If you're a smaller biotech that has no in-house communication, branding or design experts, it's important that you find an outside agency that has the experience to develop the branding strategy. In larger companies, external expertise can also provide a more balanced view of reality than in-house teams, which may be too wrapped up in corporate messages and details to fully understand what will resonate with external audiences.

Whatever the choice of agency, you need to be sure you like their work and how they do it before hiring. The right one will understand the challenges of your market, the concerns of your leadership team and the expectations of your audiences. The agency should be able to develop a brand that will create interest in your company and build emotional connections with your customers.

An agency should also question your views and decisions; for example, some leadership teams believe that they regularly need to

© 2010 Nature America, Inc. All rights reserved.

refresh their brand as they hit certain targets and milestones. But if you're delivering what you're promising, then there probably is no need to change your brand. Your agency should be working with you to look at how your brand will be affected before you hit your milestones. As you mature, you need to update your brand and messages—but these changes should, in general, be evolutionary. The only time you should be looking to review your brand is after a business-changing situation, such as a merger or acquisition, or if your company is going to move in a new direction.

A recent report on the cancer sector highlighted that Genentech was the leading company as measured by sales, products, R&D and corporate equity. Roche was ranked 15th. Genentech was also the number-one company when ranked by overall image performance[1]. It will be interesting to follow these metrics in the aftermath of the Roche takeover of Genentech.

On the merger and acquisition theme, the industry will be watching with interest how the Invitrogen (Carlsbad, California) and Applied Biosystems (Foster City, California) brands will fare under the new umbrella name of Life Technologies (Carlsbad) following their merger at the end of last year. The branding strategy appears set up to allow well-known brands to be acquired but retain their individual brand identity with the added endorsement from Life Technologies.

Media channels

Never have brands and communications been more under the spotlight than with the explosion of online blogging, social networking sites and 24/7 news feeds. It's important that you and your company maximize the potential of the Internet to promote the brand and connect with audiences across the world.

There are many social networking sites that can be used to create exposure for a brand and gain some excellent and cost-effective coverage, but you must always understand how these sites can be used to both your advantage and your disadvantage and why they are useful before you jump in. Again, the right agency will be able to provide the best advice for your individual needs. A good way to see how social networking sites are being used in the industry is to monitor how your competitors and the industry leaders are using these communications channels, especially what is being communicated and how frequently. Social media sites can be equally destructive to a brand, and any online strategies must be carefully managed. Two good questions to ask are: will the site add credibility and valuable interest to my business, and will the investment in time bring a valuable return? These tools should be considered when planning any communications strategy.

Maximizing the potential of the Internet begins with your own website, which is the first contact most people will have with your company. It is vital to make the right first impression and attract repeat visitors—your company website is your shop front.

Our experience has shown that companies aren't always focused on their websites. There have been many occasions when a company has revised its business strategy and yet the brand and website have remained the same. The outside world will always see your website as a reflection of your company—make it relevant and professional. Ensure that it's up to date and that messages are congruent with all other channels of communication. It doesn't have to cost a fortune. An effective way of evaluating the relevance and standard of your website is to involve key stakeholders who represent all areas of your business, making sure everyone has a voice. This can also be an excellent way of promoting initiative and strengthening relationships both throughout the company and with suppliers and key contacts. The challenge is facilitating and monitoring the input; this is where an agency can provide a valuable and independent link to manage and extract the contribution of stakeholders and ultimately provide the advice needed to evolve the website.

Box 3 Planning your voice

Determining and conveying your company's brand takes a great deal of managing and forethought. Here are some questions to keep in mind.

What messages is your company communicating? Are you the market leader and the most responsible? Are you the most advanced? Do you have the biggest potential or the best-in-class products?

How will you disseminate these messages? Will you use traditional print and online media or webinars and podcasts? Will you embrace the new tools of social networking and reach out to new audiences?

When will you communicate your messages? What are the best times for maximum impact and coverage? When and what are your competitors communicating? Are there any legal requirements that affect timing? Will different time zones affect the impact of your communications?

Where will you communicate these messages? What do your audiences read and how do they get their information? How can you maximize coverage and exposure, and which events and conferences are worth attending?

Who are your target audiences and how will your brand engage with them? Audiences might be customers, investors, partners, suppliers, current and future employees, media and the local community—how will you tailor your messages so they have significance to all your audiences?

Don't hurt yourself

For all the striving to establish a brand, a few wrong moves can wreck all that hard work. There are several factors that can have a negative effect on your brand.

Inaccuracy. If you do not deliver on your promises, customers and stakeholders will not have any trust in your business. A good example is advertising: it is too easy to project a certain image, set an expectation, invite customers to call your company and then not properly brief and train the employee who is answering the phone—or worse, leave a caller trapped in an automatic answering system. A consumer's experience of your brand can damage or improve how your brand is perceived.

Poor communication. You need to invest in encouraging regular feedback from audiences. Ensure they fully understand your business and allow them the opportunity to provide constructive feedback on your products and/or services. An effective mechanism could be on a website, where you would be inviting feedback with a clear incentive and benefit.

Complacency. It is very easy to sit back on your reputation when everything is going well and you are an industry leader, but that's the time when your brand is at most risk from adverse publicity. Make sure you stick to your brand strategy and that you have planned for

© 2010 Nature America, Inc. All rights reserved.

potential issues. Just as importantly, remain competitive and push the brand further. GlaxoSmithKline (Brentford, UK) has made some recent changes that are a good example. The appointment of Andrew Witty as CEO has been accompanied by a visible change in the communications output, as the company both defends and pushes its brand as an open and trusted leader in the pharmaceutical sector despite accusations that the industry is profiteering from the threat of pandemic flu.

Inconsistency. Make sure your brand communication strategy is consistent so stakeholders will get the right information at the right time in the right format. To ensure this is achieved every time, it is critical to establish clear procedures so that every element meets the company's standards and expectations.

Parting thoughts

A brand helps present your company to the rest of the world. You can create a successful brand only through a long-term relationship with your stakeholders, whether they are customers, employees, partners, investors or suppliers. Your business must assess how the brand delivers on both internal and external expectations. Ultimately, creating a successful brand will help drive value. Successful branding can essentially be accomplished in six ways: understand who you are (and who you are not); ensure your investment in branding adds value; communicate messages clearly; have an engaging visual identity; make sure your brand proposition is what your audiences actually experience; and take advantage of the speed and globalization of communication to spread the word.

1. Carlin, P. & Gordon, E. Why cancer KOLs love Genentech. *Scrip World Pharmaceutical News* **3446**, 22 May 2009, pp.28–30.

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

© 2010 Nature America, Inc. All rights reserved.

Harmonizing biosecurity oversight for gene synthesis

To the Editor:

As highlighted in your December issue, commercial gene synthesis companies routinely sell long strands of made-to-order DNA to researchers around the world. Observers have long speculated that individuals, terrorist organizations and governments could potentially use this DNA to create pathogens and other biosecurity threats. Last November, two separate industry groups—the International Association Synthetic Biology (IASB) and the International Gene Synthesis Consortium (IGSC)—issued competing standards[1,2] specifying the precautions that companies should take before they provide artificial DNA to customers. So far, roughly a dozen gene synthesis companies in the US, Europe and China have promised to follow one or the other of these standards. More recently, the US Department of Health and Human Services has similarly announced its own draft 'Framework Guidance' recommending steps that commercial gene synthesis providers should take to screen incoming orders[3]. Readers who follow private sector 'standards wars' will immediately recognize that the current situation is unstable and only one of these three standards is likely survive in the long term. Furthermore, the prevailing standard will almost certainly set security policy in this area for many years. As we explain below, the choices are profound. On the one hand, the US government's draft Guidance embraces an automated 'Best Match' approach that defines threats based on how closely they resemble the so-called Select Agent list. We argue that this procedure is clearly less capable (but also less expensive) than either private standard, both of which require human experts to investigate gene function any time a customer sequence resembles a pathogen or toxin found in government's enormously larger Genbank database. On the other hand, the two private standards—despite their strong substantive similarities—also present an important choice. This is because the IASB standard was developed openly in public meetings, whereas the IGSC standard was written behind closed doors by five large companies. For better or worse, the winning standard will almost certainly have a profound impact on future transparency both within the gene synthesis industry and in its dealings with the wider public.

Since 1999, when suppliers started making DNA to order, industry observers understood that artificial DNA could potentially be used to 'resurrect' pathogens such as smallpox that are no longer found in nature, build genetically engineered weapons similar to those pursued by the Soviet Union in the 1980s or exploit the new science of synthetic biology to create artificial pathogens. Within a few years, most gene synthesis companies had developed and implemented programs to screen incoming orders for security threats. Most companies did this by comparing customer-submitted sequences against GenBank (http://www.ncbi.nlm.nih.gov/Genbank/). They then paid human experts to determine whether the closest homologs encoded functions that could be used to make weapons. Where the answer was problematic, companies would conduct a further investigation to make sure that the customer existed, had a legitimate use for the DNA and had thought through any safety or biosecurity issues. Today, most current and proposed screening protocols follow this same basic structure.

The problem is that each DNA synthesis company imiplements this system differently. This means that across the industry, practices are highly nonuniform, with a few firms paying relatively little attention to biosecurity. Things would be better, industry observers agreed, if companies could be persuaded to converge on a uniform (and hopefully high) standard.

In principle, there are two ways to accomplish this. The first and most familiar method is government regulation. In late 2006, the US government's National Science Advisory Board for Biosecurity (Washington, DC) asked the federal government to prepare formal guidance on how gene synthesis companies should screen incoming orders[4]. Shortly afterward, the US government convened a formal interagency task force to develop this vision into formal guidelines. However, there is also a second path: private agreement. In April 2008, Europe's leading gene synthesis industry trade association, the IASB, hosted a meeting in Munich where leading companies from the United States and Europe discussed what they could do to improve biosecurity at the grassroots level[5]. Participants quickly agreed to develop a new code of conduct. By mid-2008, both industry and government were hard at work developing screening standards.

As it turned out, the private track was slightly faster. IASB produced a first draft "Code of Conduct for Best Practices in Gene Synthesis" in late 2008, which members continued to comment on into the first half of 2009. In keeping with the practice at most companies, this Code of Conduct required commercial gene synthesis providers to compare incoming orders against GenBank and use human experts to determine what the closest GenBank matches encoded. If the GenBank match was problematic—most notably, if it encoded a protein associated with virulence—members would have to conduct additional customer investigations before filling the order. On November 3, IASB held a meeting in Cambridge, Massachusetts, USA, to produce the final text of IASB's Code[2]. In the course of these discussions, participants agreed to narrow the draft Code so that members would only have to investigate sequences that corresponded to known pathogens. (The compromise was based on a judgment that current technology is incapable of making threats from the DNA found in other organisms.) Seven companies, including

© 2010 Nature America, Inc. All rights reserved.

four European and one Chinese gene synthesis companies, had signed this final Code by late November. IASB's US counterpart trade association, SynBIA (http://www.synbia.org/), also announced its support.

Alas, industry standard-setting is a messy business. Shortly after IASB announced plans to finalize its draft, two gene synthesis companies, DNA2.0 (Menlo Park, CA, USA) and Geneart (Regensburg, Germany), put together their own, competing proposal[6]. Unlike IASB's code of conduct, the DNA2.0/Geneart method did not require companies to screen against GenBank or pay human experts to determine what the closest matches encoded. Instead, the companies argued that commercial DNA providers should compare customer sequences against a predefined list of threats. The main advantage of this method, its authors pointed out, is that it can be done by computer and so promises to be both "fast" and "cheap." At the same time, a really exhaustive threat list would have to be coextensive with GenBank—and such a thing will not be possible for many years. This makes the Geneart/DNA2.0 approach far less capable than systems that rely on human experts.

By October, it was obvious that DNA2.0 and Geneart were quietly reaching out to companies around a new and possibly revised standard. Nothing happened, however, until several weeks after IASB finalized its code of conduct. Then, on November 19, five companies—DNA2.0, Geneart, Blue Heron (Bothell, WA, USA), IDT (Coralville, IA, USA) and Genscript (Piscataway, NJ, USA)—announced that they had formed the IGSC and developed a "Harmonized Screening Protocol" to guide the industry[3,7]. Encouragingly, this new document tracked the IASB's code in almost every detail. Indeed, it even dropped DNA2.0/Geneart's previous suggestion that screening should be based on predefined lists. Instead, like IASB, the IGSC code states that companies should "screen the complete DNA sequence of every synthetic gene order … against all entries found in one or more of the internationally coordinated sequence reference databanks (i.e., NCBI/GenBank, EBI/EMBL or DDBJ)." It goes on to state that when this procedure identifies "a potential pathogen or toxin sequence," orders should receive further review "by a human expert."

Many industry observers find this family resemblance between the IASB and IGSC standards puzzling. After all, IASB's Cambridge meeting had been open, and two IGSC companies (Blue Heron and Geneart) had actually attended. Why write an entirely new Protocol when IASB's Code of Conduct already existed? The reason, we think, is that IGSC membership is limited to large companies or, as one IDT representative recently put it, companies that operate a "significant business in gene synthesis"[7]. As they dominate the DNA-to-order market, the five IGSC members preferred to write their own Protocol behind closed doors. Joining IASB—which is open to all gene synthesis companies, regardless of size—would dilute their control. IGSC has also indicated in its Protocol that members may change the text in the future.

As *Nature Biotechnology* goes to press, the world's gene synthesis companies thus have two voluntary standards from which to choose. As in most private sector 'standards wars'—think of the battle between VHS and Betamax video formats—the winner is likely to be decided quickly. And that will set up the endgame. Everyone knows that big pharma and other large customers are bound to embrace the winning standard. Once that happens, gene synthesis companies from Iowa to Shanghai will have to adopt it if they want to stay in business.

By late last year, the US government regulation process was also in its final stages. One might have thought that the government would have been happy to endorse the high standards adopted by IASB and IGSC. Instead, on November 27, it published draft guidelines[1] that call for a very different approach. Under this 'Best Match' method, companies would only perform a follow-up investigation if a customer's sequence were "more closely related to a Select Agent or Toxin sequence than to a non-Select Agent or Toxin Sequence." Like earlier suggestions based on predefined lists, Best Match is easy to automate and therefore fast and cheap. But it is also less capable. Indeed, even the guidelines admit that the "non-Select Agents or Toxins" that Best Match ignores "may pose a biosecurity threat." This is very different from the IASB and IGSC approaches, which require human experts to examine all pathogen and toxin matches. (The federal draft guidelines do point out that companies "*may* [emphasis supplied] choose to investigate such sequences as part of their best practices," but this is only a suggestion.)

Clearly, commercial gene synthesis has come to a crossroads. For many years, the idea of a common, industry-wide screening standard was either theoretical or left to the indefinite future. Now, suddenly, policymakers have three to choose from. How should society approach this embarrassment of riches?

For now, the US government's draft guidelines pose the most pressing issue. The problem is obvious: if the US government thinks that Best Match is sufficient, why should any company pay human experts to carry out the IASB or IGSC standards? And indeed, several IGSC members have already said that they think Best Match is sufficient[8]. In truth, this is a cost-benefit judgment and no one can be sure that having human experts investigate GenBank matches is worth the effort. Nevertheless, it seems strange for government to tell companies that current screening programs are, in effect, too ambitious. Given that most companies have already volunteered for a high standard, government should do no less. Probably the simplest way to do this will be to revise the draft guidelines so that companies practice both Best Match and human investigation each time GenBank searches turn up pathogen and toxin genes. Readers interested in the issue should e-mail comments on the guidelines to asprfrcorrespondence@hhs.gov on or before January 26, 2010.

In the long run, society will also need to choose between the IASB and IGSC standards. We have already argued that the only key difference concerns openness. This matters for several reasons. First, we see no good reason why small gene synthesis companies should not have a voice in setting their own standards. To the extent that standards affect competitiveness, they should be fair to all. Moreover, synthetic biology as an industrial field is rapidly developing, so that today's small gene synthesis company could easily be tomorrow's large biofuels player or contract research organization. In this context, it seems arbitrary to limit IGSC's membership based on current (and possibly transient) market share.

The second reason for openness has to do with public trust. Openness—and inviting criticism—were key in showing the public that the IASB's Code of Conduct had been carefully designed. This openness will be needed again. For example, the IASB, IGSC and draft US federal guidelines say relatively little about how companies should investigate their customers. What should companies do for the handful of orders—estimated to be one or two per thousand[9]—where a customer (i) seeks a potentially dangerous sequence and (ii) is an individual or else is affiliated with an unknown start-up company? Suggestions based on computerized checks against, say, Dun and Bradstreet listings—which are notoriously subject to fraud[10]—strike us as clearly insufficient. But when human intervention is needed, companies must have detailed guidance as to which methods are and are not acceptable. These are hard questions that must not only be solved correctly but also shown to the public to have been solved correctly.

© 2010 Nature America, Inc. All rights reserved.

Finally, openness is efficient. IASB and one of us (S.M.M.) at the University of California, Berkeley's Goldman School of Public Policy are now collaborating on pilot software for a proposed online forum (VIREP) where experts who investigate genes in the course of screening will be able to deposit their research. Sharing these data will save companies the trouble of investigating the same genes over and over again. Nothing will happen, however, unless companies are open with one another. It is difficult to see how this can happen under standards that are set by a few self-described 'significant' companies.

The high standards set forth in the IASB and IGSC standards could still collapse. And even if they don't, no one can be sure whether future standards will be set in an open and transparent way. Still, gene synthesis companies have come a long way since IASB members first agreed to pursue an industry-wide Code of Conduct. The goal is in reach.

Markus Fischer[1] & Stephen M Maurer[2]

[1]Entelechon GmbH, Regensburg, Germany.
[2]Goldman School of Public Policy, University of California, Berkeley, California, USA.
e-mail: smaurer@berkeley.edu

1. International Association Synthetic Biology. Code of conduct for best practices in gene synthesis. <http://tinyurl.com/iasbcode/> (2009).
2. International Gene Synthesis Consortium. Harmonized screening protocol: gene sequence & customer screening to promote biosecurity. <http://www.genesynthesisconsortium.org/Harmonized_Screening_Protocol.html> (2009).
3. Department of Health and Human Services. Screening framework guidance for synthetic double-stranded DNA providers. *Fed. Regist.* **74**, 62319–62327 <http://edocket.access.gpo.gov/2009/E9-28328.htm> (2009).
4. National Science Advisory Board for Biosecurity. Addressing biosecurity concerns related to the synthesis of select agents. <http://oba.od.nih.gov/biosecurity/pdf/Final_NSABB_Report_on_Synthetic_Genomics.pdf> (2006).
5. International Association Synthetic Biology. Workshop report: technical solutions for biosecurity in synthetic biology. <http://tinyurl.com/iasbreport/> (2008).
6. Check Hayden, E. *Nature* **461**, 22 (2009).
7. Anonymous. World's top gene synthesis companies establish tough biosecurity screening protocol. <http://www.einpresswire.com/article/56890-world-s-top-gene-synthesis-companies-establish-tough-biosecurity-screening-protocol> (2009).
8. Wadman, M. US drafts guidelines to screen genes. *Nature News* <http://www.nature.com/news/2009/091204/full/news.2009.1117.html> (4 December 2009).
9. Maurer, S., Fischer, M., Schwer, H., Stähler, C. & Stähler, P. Working paper: making commercial biology safer: what the gene synthesis industry has learned about screening customers and orders. <http://gspp.berkeley.edu/iths/Maurer_IASB_Screening.pdf> (2009).
10. Tozzi, J. Watchdogs: shell schemes are on the rise. *BusinessWeek.com* <http://www.businessweek.com/smallbiz/content/nov2009/sb2009115_791003.htm> (5 November 2009).

Correcting the record

To the Editor:

As executive director, and on behalf of the International Life Sciences Institute (ILSI; Washington, DC), I write to correct certain statements regarding ILSI made by David Schubert in a letter published in the September issue[1]. In suggesting that University of Georgia Professor Wayne Parrott was incorrectly characterized in a news article as a "public sector" scientist, Schubert accurately noted that Parrott is a scientific advisor to ILSI, but failed to acknowledge he was not speaking in any capacity relating to ILSI. Schubert then proceeded inaccurately to characterize ILSI as a "lobby group" and repeated a claim—made in an anonymous article on a Wikipedia-style website—that ILSI has a "hidden agenda to protect the interests of the food, chemical, and drug industries." Schubert further claimed that ILSI had been "banned from participating in World Health Organisation [WHO; Geneva] activities." These statements are both derogatory and untrue.

First, ILSI has no "hidden agenda." ILSI is a respected world leader in scientific inquiry relating to nutrition, food safety, toxicology, risk assessment and the environment. ILSI has achieved this status through its tripartite operating model of bringing together scientists from academia, government and industry, which, in addition to producing broadly informed scientific output, also helps to ensure that ILSI's work is balanced and directed toward the public good rather than the commercial interests of its members. ILSI also publishes the results of its work, irrespective of whether it could be considered beneficial or harmful to its members' interests.

Second, ILSI is not a "lobby group"; it is a public charity exempt from taxation under US Internal Revenue Code § 501(c)(3) and, as such, is prohibited by law from attempting to influence legislation as a substantial part of its activities. Since its inception, ILSI has gone beyond this legal prohibition and has refrained from lobbying. ILSI's board of trustees has always endorsed this stance. The most recent (March 2009) articulation of this tenet is found in the International Life Sciences Institute Code of Ethics and Organizational Standards of Conduct (available on the ILSI website; http://www.ilsi.org/documents/code%20of%20ethics%20(2009).pdf). This document states that, "Advocacy of any kind is strictly limited to promotion of the use of evidence-based science as an aid in decision-making. ILSI does not conduct lobbying activities." Thus, although ILSI does, from time to time, provide scientific information to public decision-making bodies, we do not endorse specific public policy outcomes other than the application of evidence-based science

Finally, WHO has not "banned" ILSI from participating in its activities. Schubert stated inaccurately that WHO has "banned" ILSI from taking part in its activities. It is true that in 2005, certain interest groups and labor unions wrote to the WHO to complain about the fact that ILSI receives industry funding, and expressed concern, among other things, that an ILSI-WHO scientific workshop scheduled for the spring of 2006 could "influence recommendations for the WHO on its Guidelines for Drinking Water Quality due in 2008." A subsequent report of the WHO Board Standing Committee on Nongovernmental Organizations (NGOs) clarified that WHO's collaboration with ILSI was intended to harness ILSI's technical expertise but did not include participation in "normative activities," such as setting microbiological or chemical standards for food and water. Schubert incorrectly implies that the WHO barred ILSI from participation in any of its activities because of doubts about its scientific integrity. However, the truth is that ILSI had never participated in the standard-setting matters in question because the WHO had consistently reserved the right to set such standards without direct collaboration with outside organizations. Moreover, the WHO continues to recognize ILSI as an accredited NGO and continues to work with ILSI, as it has since 1991.

On the basis of the above evidence, the derogatory statements made about ILSI in Schubert's letter go well beyond the mere expression of opinion, and instead constitute assertions of fact that are simply untrue.

© 2010 Nature America, Inc. All rights reserved.

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

Suzanne Harris

International Life Sciences Institute, Washington, DC, 20005, USA.
e-mail: suzie@dc.ilsi.org

1. Schubert, D. *Nat. Biotechnol.* **27**, 802–803 (2009).

David Schubert replies:

Suzanne Harris has four concerns about my letter in the September 2009 issue of *Nature Biotechnology*[1]. She claims that I implied Parrott's comments were made as a representative of the International life Sciences Institute (ILSI), that I misstated the ILSI agenda and lobbying status, and that I incorrectly stated that ILSI was banned from World Health Organization activities.

The first is simply not true. Nowhere was this implied, for my only goal was to demonstrate that Parrott is not an unbiased academic observer in the transgenic food debate because he has associations with industry-sponsored institutions, such as ILSI. He has, in fact, in the past co-authored letters with industry-backed scientists to *Nature Biotechnology* similar to that of Harris[2].

The remaining concerns relate to the legal definition of lobbying or are subjective in nature. With respect to the latter, I will only furnish a few additional references, from which interested readers can make up their own minds.

With respect to the agenda of ILSI, although I may be wrong, it seems logical to me that an organization that is heavily funded by the world's largest food, tobacco and transgenic seed companies is going to promote the interests of their support group. Although I did quote a referenced website regarding another group's assessment of the ILSI 'agenda', I recommend an examination of additional documents that some may say reach a similar conclusion. These include citations relating to ILSI activities in "Integrity in Science"[3] published by the Center for Science in the Public Interest (CSPI; Washington, DC) and an article by Michael Jacobson[4] that outlines the various ways industry is able to manipulate science and public health policy.

With respect to lobbying, I was not aware of the legal definition of a lobbying group, and in this context both my cited source for this claim and I misused the word. I apologize for this mistake. It should be pointed out, however, that there are many ways to influence policy independently of formal lobbying, including those outlined by Jacobson[4], as well as the 'sound science' approach promoted by Newt Gingrich and the Bush administration[5].

Finally, with respect to the ban of ILSI from WHO activities, I did not claim that they were banned from all WHO activities. Because of space limitations, I cited a text that was heavily referenced regarding the details of the WHO incident. Additional references include the Associated Press[6] and CSPI[3].

My conclusion that Wayne Parrott is not simply a public sector plant biologist and should not have been introduced as such remains the same and was in fact confirmed by *Nature Biotechnology*[7]. However, it should be the responsibility of *Nature Biotechnology* to document these conflicts of interest, not a concerned reader, such as myself. A similar conflict with industry-funded plant biologists representing themselves as neutral commentators in the transgenic food debate was documented in these pages many years ago[8].

1. Schubert, D. *Nat. Biotechnol.* **27**, 802–803; author reply 803 (2009).
2. Beachy, R. *et al. Nat. Biotechnol.* **20,** 1195–1196; author reply 1197 (2002).
3. http://www.cspinet.org/integrity/.
4. Jacobson, M.F. Lifting the veil of secrecy: corporate support for health and environmental professional associations, charities, and industry front groups. *CSPI and its Integrity in Science Project*, <http://cspinet.org/new/pdf/lift_the_veil_intro.pdf> (8 September 2003).
5. Schubert, D. Bush's "sound science": turning a deaf ear to reality. *The San Diego Union Tribune* <http://legacy.signonsandiego.com/uniontrib/20040709/news_lzle-9schubert.html> (9 July 2004).
6. Heilprin, J. WHO to rely less on US research. *Associated Press Online*. <http://www.sfgate.com/cgi-bin/article/article?f=/n/a/2006/01/27/national/w150409S47.DTL> (27 January 2006).
7. Anonymous. *Nature Biotechnology* replies. *Nat. Biotechnol.* **27**, 803 (2009).
8. Sharpe, V.A. & Gurian-Sherman, D. Competing interests. *Nat. Biotech.* **21**, 1131 (2003).

International trade and the global pipeline of new GM crops

To the Editor:

In a previous issue, Paul Christou and colleagues[1] highlighted the patchwork of laws and regulations governing tolerance levels for *approved* genetically modified (GM) material in non-GM food and in the labeling and traceability of GM products. A related but different problem is that of 'asynchronous approval' of *new* GM crops across international jurisdictions, which is of growing concern due to its potential impact on global trade. Different countries have different authorization procedures and, even if regulatory dossiers are submitted at the same time, approval is not given simultaneously (in some cases, delays can even amount to years). For instance, by mid-2009 over 40 transgenic events were approved or close to approval elsewhere but not yet approved—or not even submitted—in the European Union (EU; Brussels) (for more details, see **Supplementary Data**). Yet, like some other jurisdictions, the EU also operates a 'zero-tolerance' policy to even the smallest traces of nationally unapproved GM crops (so-called low-level presence). The resultant rejection of agricultural imports has already caused high economic losses and threatens to disrupt global agri-food supply chains[2–8].

To assess the likelihood of future incidents of low-level presence of unapproved GM material in crop shipments and to understand related impacts on global trade and the EU's agri-food sector, we compiled a global pipeline of new GM crops. Our motivation was to obtain a realistic estimate of how many new GM crops will be commercialized in the next years, by whom and in which countries—and when these new crops will be authorized by the different trading partners of the EU. In this context, we invited a select panel of national regulators, industry representatives, experts from national and international research institutes and actors from the global food and feed supply chain to a workshop organized at the Institute for Prospective Technological Studies of the European Commission's Joint Research Centre in November 2008 to discuss for the first time the issue of low-level presence in view of a growing global pipeline of new GM crops. (For more details, see **Supplementary Notes**.)

© 2010 Nature America, Inc. All rights reserved.

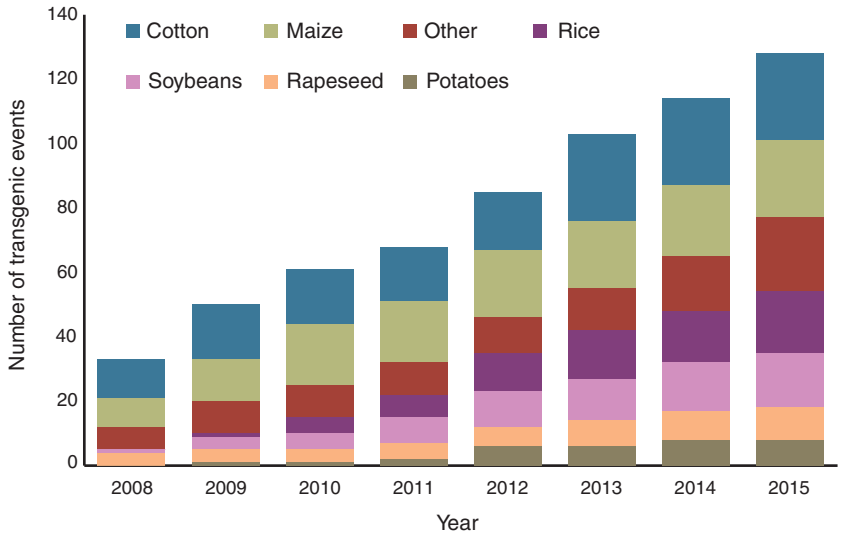


Figure 1 Current numbers and estimations of future numbers of GM crops worldwide. Although the commercialization of the crops shown may be technically possible by 2015, the practical—or rather regulatory—feasibility may be more questionable (e.g., for rice in particular), given that in some of the developer countries no GM (food) crops have been authorized so far. See **Supplementary Data**.

On the basis of this workshop and subsequent desk research, we predict that by 2015 there could be over 120 different transgenic events in commercialized GM crops worldwide—compared with around 30 GM events in commercially cultivated GM crops in 2008 (**Fig. 1**)[9]. Although the currently common traits in GM crops (insect resistance and herbicide tolerance) will continue to be the most common traits in 2015, optimized crop composition is expected to gain increasing importance (**Table 1**). Moreover, we expect that about half the new transgenic events that could be brought to market until 2015 will have been developed by players in Asia (33 in India, 20 in China, 5 in the rest of Asia) and Latin America, with the other half coming from companies in the United States and the EU. (For more details, see **Supplementary Data**.)

Apart from the implication that a quadrupling of the number of transgenic events in commercialized GM crops between 2008 and 2015 is likely to increase the negative impact of low-level presence on international trade (if no fundamental change takes place in the way new GM crops are currently approved in different countries), our study also indicates that new transgenic events are likely to be introduced and that new crops will be targeted. Even so, the current crops dominating the GM landscape (soybeans, maize, cotton and canola) will continue to dominate the picture in 2015 (**Fig. 1**). Likewise, the long-anticipated product quality traits are likely to come forth only slowly: of a total of 91 new GM crops that are expected to be commercialized between 2009 and 2015, only 18 represent quality innovations (**Table 1**). This conclusion is supported by a survey reported in the August issue, where Graff *et al.*[10] sought to answer the question of why quality-improving innovations from agbiotech have not been more readily forthcoming, and through a survey carried out in 2004 they identified 49 quality innovations, which they expected to be commercialized by 2015 (21 of those between 2010 and 2015).

More importantly, though, what our present study indicates is that an increasing number of GM crops are being developed by new players outside the United States or Europe (in particular, by actors in Asia). These new players develop the crops for their population-rich home markets and may therefore be less affected by the marketability of the crops abroad. Just in November, for example, China took a major step towards endorsing the use of a major staple crop, GM rice[11]. However, as has been seen in the recent cases where traces of GM maize in soybeans led to the rejection of the soybean shipments[7], under certain regulatory settings (in particular zero tolerance towards low-level presence) the cultivation of one type of crop may even affect the marketability of other types of crops. This means that if third countries want to authorize GM varieties of crops that are welfare-enhancing for their societies, in future they may also consider the potential impact of 'cross low-level presence' in different, but export-relevant, crops. The extent to which this situation shapes the approval and development of future agbiotech innovations remains to be seen. Unfortunately, past experience with the use of GM crops shows that irrational fear of export losses represents a significant impediment to biosafety policymaking[12].

Table 1 Numbers of current and possible numbers of expected GM traits worldwide

Trait category[a]	Commercial in 2008	Commercial pipeline	Regulatory pipeline	Advanced development	Total by 2015[b]
Insect resistance	21	2	11	25	59
Herbicide tolerance	11	5	4	13	33
Product quality[c]	2	1	5	12	20
Virus resistance	5	0	2	3	10
Abiotic stress tolerance	0	0	1	6	7
Other	0	0	2	11	13

[a]Crops in the commercial pipeline are already authorized in at least one country but not yet marketed by the developer, crops in the regulatory pipeline are submitted for authorization in at least one country but are not yet authorized anywhere and crops in advanced development are not yet submitted for authorization but it is expected that they will pass the regulatory process by 2015. [b]Numbers do not add up to total numbers given in **Figure 1** because of stacking of traits in some new GM crops. [c]Product quality comprises crop composition traits as well as improved shelf life; crop composition is optimized for maize, canola, soybeans, potatoes and rice, and the targeted compounds cover fatty acids, amino acids, starch, beta-carotene and enzymes (these crops are optimized for use as food, feed, biofuel or industrial inputs). Source: **Supplementary Data**.

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

Alexander J Stein & Emilio Rodríguez-Cerezo

European Commission, Joint Research Centre (JRC), Institute for Prospective Technological Studies, Seville, Spain.
e-mail: contact@ajstein.de; emilio.rodriguez-cerezo@ec.europa.eu

Disclaimer: The views expressed are purely those of the authors and may not in any circumstances be regarded as stating an official position of the European Commission.

1. Ramessar, K., Capell, T., Twyman, R.M., Quemada, H. & Christou, P. *Nat. Biotechnol.* **26**, 975–978 (2008).
2. Vermij, P. *Nat. Biotechnol.* **24**, 1301–1302 (2006).
3. Mitchell, P. *Nat. Biotechnol.* **25**, 1065–1066 (2007).
4. DG AGRI. *Economic Impact of Unapproved GMOs on EU Feed Imports and Livestock Production* (European Commission, Brussels, 2007). <http://ec.europa.eu/agriculture/publi/reports/>
5. Brookes, G. *Economic Impacts of Low Level Presence of Not Yet Approved GMOs on the EU Food Sector* (Graham

© 2010 Nature America, Inc. All rights reserved.

Brookes Consulting, Gloucester, 2008). <http://www.agindustries.org.uk/document.aspx?fn=load&media_id=3118&publicationId=396>
6. Backus, G.B.C. *et al. EU Policy on GMOs: a Quick Scan of the Economic Consequences* (LEI Report 2008-070, Wageningen University and Research Centre, The Hague, 2008). <http://www.lei.wur.nl/UK/publications+en+products/LEI+publications/default.htm?id=932>
7. Hornby, C. *et al.* EU Rejects More US Soy with GM corn traces. *Reuters*, September 18, 2009 <http://www.reuters.com/article/companyNews/idUKTR58H4GG20090918>
8. Fischer Boel, M. GMOs: letting the voice of science speak. (European Commissioner for Agriculture, SPEECH/09/474, European Commission, Brussels, 2009). <http://europa.eu/rapid/pressReleasesAction.do?reference=SPEECH/09/474>
9. Stein, A.J. & Rodríguez-Cerezo, E. *The Global Pipeline of New GM Crops: Implications of Asynchronous Approval for International Trade* (JRC Technical Report EUR 23486 EN, European Communities, Luxemburg, 2009). <http://ipts.jrc.ec.europa.eu/publications/pub.cfm?id=2420>
10. Graff, G.D., Zilberman, D. & Bennett, A.B. *Nat. Biotechnol.* **27**, 702–704 (2009).
11. Batson, A. & Areddy, J.T. Beijing gives nod to modified rice. *Wall Street Journal*, p.A13 <http://online.wsj.com/article/SB125959909959569901.html> (1 December 2009).
12. Gruère, G & Sengupta, D. *Food Pol.* **34**, 399–406 (2009).

High-density resequencing DNA microarrays in public health emergencies

To the Editor:

Despite epidemiological intelligence and microbiological surveillance systems or programs, virus emergence mostly occurs by surprise. Influenza A virus offers a paradigm of such a situation: the world was focusing on a threat due to highly pathogenic (H5N1) avian influenza A virus in Asia[1]. Instead, it faced a novel reassortant swine-origin H1N1 influenza A virus (H1N1pdm) with pandemic potential on the American continent[2]. Seasonal influenza surveillance and pandemic preparedness are potent incentives to develop rapid, specific, sensitive and robust diagnostic tools for laboratories, particularly the World Health Organization's (WHO; Geneva) National Influenza Centers and other laboratories in different countries specialized in epidemic responses. Here, we describe the use of random non-PCR–based nucleic acid amplification and high-density resequencing DNA microarrays to rapidly identify reassorted influenza A virus strains of swine origin. Such an approach could prove useful for public health authorities attempting to detect and identify reassortment events in future outbreaks.

Worldwide, laboratories have developed or implemented identification assays targeting A(H5N1), A(H7N7), A(H9N2) and sometimes A(H2N2) viruses besides seasonal human influenza viruses A(H1N1) and A(H3N2). Very few anticipated that, one day, there would be a need for the detection of A(H1N1) with swine virus–derived components. Therefore, when the global alert spread worldwide about a new influenza A(H1N1) virus with a novel genetic make-up on April 24, 2009 (ref. 2), all frontline laboratories were abruptly faced with the task of rapidly identifying samples—in the case of our laboratory, as early as two days after the initial global alert. For national health authorities across the world, it was critical that local laboratories were able not only to identify H1N1 viruses but also to differentiate seasonal human H1N1 viruses from swine-like H1N1 strains.

Despite prompt dissemination of a specific real-time reverse transcriptase (RT)-PCR set of protocols by the US Centers for Disease Control (CDC; Atlanta) and the very rapid availability of viral sequences on the Global Initiative on Sharing Avian Influenza Databank website (GISAID; http://platform.gisaid.org/dante-cms/live/struktur.jdante?aid=1131/), delays imposed by primer and probe manufacturers prevented immediate implementation of specific H1N1 assays in many centers across the world. Several national influenza centers resorted to performing RT-PCR of the influenza M (matrix) segment followed by sequencing of amplicons. Our laboratory, which focuses on the early identification of microbiological threats but is not specialized in influenza in particular, found another solution. We first used random non-PCR–based nucleic acid amplification of the samples[3] followed by high-density resequencing on DNA microarrays (PathogenID v2.0) to identify and characterize four genes of the novel reassortant influenza virus from swine origin.

We designed the PathogenID v2.0 microarray with a total of 126 viral sequences from a whole range of viruses. It was engineered to detect and describe four genes in order to type and subtype influenza viruses from a large diversity of natural and permanent hosts (humans, birds, horses and pigs) with a minimum of sequences tiled on the array (Leclercq *et al.*, unpublished data; see **Supplementary Data** for PathogenID

Table 1 Identification and characterization of the new H1N1pdm RNA alone or in mixture with other influenza A RNAs

Strain(s) tested	Gene tiled	Call rate (%)[a]	Identification by BLASTN analysis of the sequence reconstructed by the DNA microarray		
			Type	Subtype	Sequence origin (for H1N1pdm)
A/Paris/2590/2009 (H1N1)v tested alone	PB2	69	A		Avian
	H	82.1		H1	North American swine lineage
	N	75.2		N1	European swine lineage
A/Paris/2590/2009 (H1N1)v tested with mixture A (A/Bayern/7/95 (H1N1) and A/Wyoming/03/2003 (H3N2))	PB2	72.2	A		Avian
	H	80.4		H1	North American swine lineage
	N	58.2[b]		N1	European swine lineage
A/Paris/2590/2009 (H1N1)v tested with mixture B (A/Duck/Cambodia/D4(KC)/2006 (H5N1))	PB2	64.8	A		Avian
	H	82.7		H1	North American swine lineage
	N	70.5		N1	European swine lineage

[a]The call rate is the percentage of bases called by the resequencing algorithm[11]. Identification of each segment was performed with the same corresponding sequence tiled on the array.
[b]In this case, another sequence of N1 was used to identify N1 segment from H1N1pdm.

© 2010 Nature America, Inc. All rights reserved.

Brookes Consulting, Gloucester, 2008). <http://www.agindustries.org.uk/document.aspx?fn=load&media_id=3118&publicationId=396>
6. Backus, G.B.C. *et al. EU Policy on GMOs: a Quick Scan of the Economic Consequences* (LEI Report 2008-070, Wageningen University and Research Centre, The Hague, 2008). <http://www.lei.wur.nl/UK/publications+en+products/LEI+publications/default.htm?id=932>
7. Hornby, C. *et al.* EU Rejects More US Soy with GM corn traces. *Reuters*, September 18, 2009 <http://www.reuters.com/article/companyNews/idUKTR58H4GG20090918>
8. Fischer Boel, M. GMOs: letting the voice of science speak. (European Commissioner for Agriculture, SPEECH/09/474, European Commission, Brussels, 2009). <http://europa.eu/rapid/pressReleasesAction.do?reference=SPEECH/09/474>
9. Stein, A.J. & Rodríguez-Cerezo, E. *The Global Pipeline of New GM Crops: Implications of Asynchronous Approval for International Trade* (JRC Technical Report EUR 23486 EN, European Communities, Luxemburg, 2009). <http://ipts.jrc.ec.europa.eu/publications/pub.cfm?id=2420>
10. Graff, G.D., Zilberman, D. & Bennett, A.B. *Nat. Biotechnol.* **27**, 702–704 (2009).
11. Batson, A. & Areddy, J.T. Beijing gives nod to modified rice. *Wall Street Journal*, p.A13 <http://online.wsj.com/article/SB125959909959569901.html> (1 December 2009).
12. Gruère, G & Sengupta, D. *Food Pol.* **34**, 399–406 (2009).

High-density resequencing DNA microarrays in public health emergencies

To the Editor:

Despite epidemiological intelligence and microbiological surveillance systems or programs, virus emergence mostly occurs by surprise. Influenza A virus offers a paradigm of such a situation: the world was focusing on a threat due to highly pathogenic (H5N1) avian influenza A virus in Asia[1]. Instead, it faced a novel reassortant swine-origin H1N1 influenza A virus (H1N1pdm) with pandemic potential on the American continent[2]. Seasonal influenza surveillance and pandemic preparedness are potent incentives to develop rapid, specific, sensitive and robust diagnostic tools for laboratories, particularly the World Health Organization's (WHO; Geneva) National Influenza Centers and other laboratories in different countries specialized in epidemic responses. Here, we describe the use of random non-PCR–based nucleic acid amplification and high-density resequencing DNA microarrays to rapidly identify reassorted influenza A virus strains of swine origin. Such an approach could prove useful for public health authorities attempting to detect and identify reassortment events in future outbreaks.

Worldwide, laboratories have developed or implemented identification assays targeting A(H5N1), A(H7N7), A(H9N2) and sometimes A(H2N2) viruses besides seasonal human influenza viruses A(H1N1) and A(H3N2). Very few anticipated that, one day, there would be a need for the detection of A(H1N1) with swine virus–derived components. Therefore, when the global alert spread worldwide about a new influenza A(H1N1) virus with a novel genetic make-up on April 24, 2009 (ref. 2), all frontline laboratories were abruptly faced with the task of rapidly identifying samples—in the case of our laboratory, as early as two days after the initial global alert. For national health authorities across the world, it was critical that local laboratories were able not only to identify H1N1 viruses but also to differentiate seasonal human H1N1 viruses from swine-like H1N1 strains.

Despite prompt dissemination of a specific real-time reverse transcriptase (RT)-PCR set of protocols by the US Centers for Disease Control (CDC; Atlanta) and the very rapid availability of viral sequences on the Global Initiative on Sharing Avian Influenza Databank website (GISAID; http://platform.gisaid.org/dante-cms/live/struktur.jdante?aid=1131/), delays imposed by primer and probe manufacturers prevented immediate implementation of specific H1N1 assays in many centers across the world. Several national influenza centers resorted to performing RT-PCR of the influenza M (matrix) segment followed by sequencing of amplicons. Our laboratory, which focuses on the early identification of microbiological threats but is not specialized in influenza in particular, found another solution. We first used random non-PCR–based nucleic acid amplification of the samples[3] followed by high-density resequencing on DNA microarrays (PathogenID v2.0) to identify and characterize four genes of the novel reassortant influenza virus from swine origin.

We designed the PathogenID v2.0 microarray with a total of 126 viral sequences from a whole range of viruses. It was engineered to detect and describe four genes in order to type and subtype influenza viruses from a large diversity of natural and permanent hosts (humans, birds, horses and pigs) with a minimum of sequences tiled on the array (Leclercq *et al.*, unpublished data; see **Supplementary Data** for PathogenID

Table 1 Identification and characterization of the new H1N1pdm RNA alone or in mixture with other influenza A RNAs

Strain(s) tested	Gene tiled	Call rate (%)[a]	Identification by BLASTN analysis of the sequence reconstructed by the DNA microarray		
			Type	Subtype	Sequence origin (for H1N1pdm)
A/Paris/2590/2009 (H1N1)v tested alone	PB2	69	A		Avian
	H	82.1		H1	North American swine lineage
	N	75.2		N1	European swine lineage
A/Paris/2590/2009 (H1N1)v tested with mixture A (A/Bayern/7/95 (H1N1) and A/Wyoming/03/2003 (H3N2))	PB2	72.2	A		Avian
	H	80.4		H1	North American swine lineage
	N	58.2[b]		N1	European swine lineage
A/Paris/2590/2009 (H1N1)v tested with mixture B (A/Duck/Cambodia/D4(KC)/2006 (H5N1))	PB2	64.8	A		Avian
	H	82.7		H1	North American swine lineage
	N	70.5		N1	European swine lineage

[a]The call rate is the percentage of bases called by the resequencing algorithm[11]. Identification of each segment was performed with the same corresponding sequence tiled on the array.
[b]In this case, another sequence of N1 was used to identify N1 segment from H1N1pdm.

© 2010 Nature America, Inc. All rights reserved.

v2.0 microarray design). The possibility of using consensus sequences was one of the hypotheses tested in the design of the new version of the microarray. For each of the four genes, from 12 to 148 different sequences were first aligned. From the alignments, a small number of clusters gathering sequences with <15% variation were determined. For each cluster, one to three consensus sequences were computed and used for tiling (see **Supplementary Table 1**).

The H1N1pdm virus that emerged in Mexico in 2009 resulted from a reassortment between the triple reassortant swine influenza virus that had caused considerable problems for pig farmers for several years and a Eurasian swine lineage. The eight segments of the resulting virus were thus from different host origins (that is, (i) avian (for PB2 and PA genes), (ii) human (PB1) and (iii) classic (H1, NP and NS) or avian-like (M, N1) swine viruses)[4]. From the start of the outbreak, our laboratory faced the urgent need to investigate several suspected cases of human infection by the novel virus.

Among the samples processed during the outbreak, some of the isolates (for example, sample 70; designated as A/Paris/2618/2009 (H1N1)) were identified as seasonal human influenza viruses, whereas others (for example, sample 49, designated as A/Paris/2590/2009 (H1N1)) were clearly harboring genes different from those of viruses previously known to circulate in humans. Six original clinical specimens were tested with our PathogenID v2.0 resequencing microarray (**Supplementary Methods**). Four of the six tested samples were correctly identified as H1N1pdm virus (3/5) or seasonal H1N1 influenza virus (1/1) by BLASTN analysis of all viral sequences obtained (**Supplementary Table 2**). A low viral load in the original clinical specimens was thought to explain why only two viral segments (M and H1) were detected and identified in the case of sample 49, and no sequence at all was detected in the remaining two samples (data not shown). In these two cases, a quantification by real-time RT-PCR specific for the M gene showed that they were under the detection limit by the DNA resequencing microarray, which has been estimated to be around 2.5×10^8 viral genomes copies after amplification[3].

As positive controls, we have successfully tested one prototype human strain A/Bayern/7/95 (H1N1)[5] and two strains isolated from pigs, one representing the classical lineage A/Swine/England/117316/86 (H1N1)[6] and one representing the Eurasian lineage (also called avian-like) A/Swine/England/195852/92 (H1N1)[6,7]. Amplification products of viral RNA of A/Bayern/7/95 (H1N1) cell culture supernatant and of A/Paris/2618/2009 (H1N1) isolate both hybridized with one of the H1 consensus sequences tiled on the array. The BLASTN analysis of the sequences reconstructed by the PathogenID v2.0 microarray showed these strains were seasonal human influenza viruses (data not shown). For the isolate of the novel reassortant A/Paris/2590/2009 (H1N1), all sequences reconstructed from the microarray are presented in **Supplementary Figure 1**. The hemagglutinin (HA) sequence was reconstructed by hybridization to a H1 tiled consensus sequence that is different from the previous H1 consensus to which the seasonal strains hybridized. This shows that the method is useful as a discriminating diagnostic. BLASTN analysis allowed the determination of the segment origin for PB2 (which encodes subunits of viral RNA polymerase), H1 and N1 genes, which were derived from avian, classic swine virus and avian-like swine virus respectively (**Table 1**). Data on the M gene are not shown because this gene (only represented by one sequence on the chip) has been chosen to provide detection of influenza viruses and identification of their types and not further characterization. Geographical and host origins of PB2 and H1 segments were visualized by a phylogenetic tree generated with a wide range of prototype influenza A viruses (**Fig. 1**).

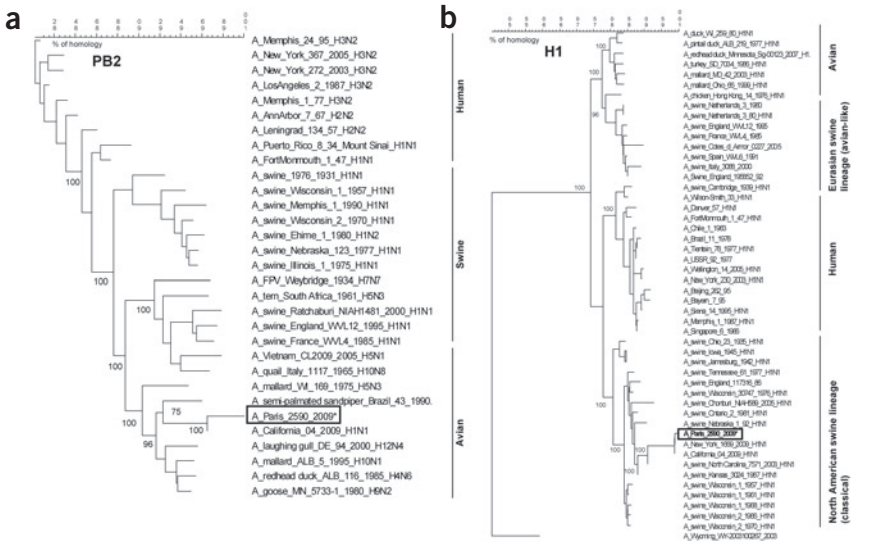


Figure 1 Phylogenetic trees of PB2 and H1 genes belonging to influenza A viruses isolated over a long period of time across a wide geographical area and from a variety of host species. (**a**,**b**) Phylogenetic analyses were based on available sequences of PB2 (**a**) and H1 (**b**) genes from either GISAID or NCBI databases. Trees were generated by the neighbor-joining method with BioNumerics software for windows (version 5.1, Applied Maths) by bootstrapping approach (with 100 replicates). *, sequences generated by DNA resequencing microarray.

A reassortment between this novel H1N1pdm and the highly pathogenic avian (H5N1) virus would be a pivotal event as it could generate a new virus with increased pathogenicity and transmissibility. Another event of reassortment leading to the transfer of the neuraminidase (NA) of current seasonal (H1N1) influenza strains to the pandemic prone H1N1pdm would very probably generate a new virus with partial surface antigen shift (on the HA only) but with a high level of resistance to oseltamivir (Tamiflu), a specific and effective drug against influenza[8]. This would have major consequences for public health and it would be paramount to detect such a reassortment as soon as possible, should it occur. As in many detection algorithms, the first set of protocols made available worldwide by the CDC, NA detection and characterization is generally not performed by first-line laboratories. In-depth characterization is therefore delayed until diagnostic confirmation and is undertaken by world-level reference centers.

The avian strain A/Duck/Cambodia/D4(KC)/2006 (H5N1) on its own was successfully tested for PB2, HA and NA by the DNA microarray (data not shown). Various mixtures of equivalent amounts of RNA of influenza viruses belonging to the same subtype (H1N1) or in combination with different subtypes (H3N2 or H5N1) were then processed and hybridized onto the PathogenID v2.0 microarray. In all

© 2010 Nature America, Inc. All rights reserved.

tested cases, all segments of H1N1pdm virus were detected and identified after a BLASTN analysis. The generated sequences were not affected by the presence of other viral sequences, even within the same molecular subtype. This strategy allowed a large viral genetic diversity to be covered and proved to be discriminating, even in the presence of a mixture of viral RNAs (see **Table 1** and **Supplementary Table 3**).

Non-PCR amplification systems, such as loop-mediated isothermal PCR, nucleic acid sequence-based amplification and rolling circle amplification, are used for viral diagnosis, but all these approaches need specific primers. Random PCR-based amplification processes are available but although random priming can amplify an unknown target, it often yields lower amounts of DNA than specific primers, which can reduce the overall sensitivity of the process. In previous studies, we showed that Phi29-based multiple displacement amplification is much more sensitive than random PCR[3] because amplification is random and thus independent of any specific or orientated process (unlike quantitative RT-PCRs[9] or specific hybridization DNA microarrays[10] that can be developed only after a new viral strain is identified).

Taken together with the advantages of our amplification protocol, we feel the PathogenID v2.0 DNA microarray offers several advantages over existing diagnostics for detecting novel pathogens. First, the microarray is able to generate sequences that are not already tiled. Second, PathogenID v2.0 is ready for use (indeed, its deployment in analyzing field strains might have enabled earlier identification of the novel H1N1 virus when it first emerged, which is thought to have been several months before the outbreak was officially recognized)[4]. Using PathogenID v2.0, any laboratory around the world could potentially carry out rapid and accurate viral identification. In addition, there would be no lag time between public health authorities issuing an alert and laboratories around the world implementing tests. Although high-throughput pyrosequencing technologies, such as the 454 FLX system marketed by Roche Diagnostics (Basel), can also be deployed for in-depth characterization of novel viruses, the cost of sequencing instrumentation and reagents, as well as the delay in interpreting the data, makes these systems less useful to public health authorities seeking the first occurrence of a novel virus in a given country. Although the sequence information provided by resequencing microarrays are more limited than those produced by high-throughput pyrosequencing technologies, their ability to generate results as early as 24 h after the beginning of the experiment is a substantial advantage. The versatility, the rapidity and the high discriminating power of the PathogenID v2.0 microarray could prove critical to detect and identify reassortment events and therefore prompt health authorities to take efficient decisions for patient treatment and for outbreak management.

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

ACKNOWLEDGMENTS

This study was supported by Programme Transversal de Recherche (PTR DEVA n°246) from Institut Pasteur (France). We thank the sponsorship of Total-Institut Pasteur for their financial support of this research program. We thank the EU-funded RIVERS program (Sixth Framework Program SSP-5-B-INFLUENZA 04 405: Resistance of Influenza Viruses in Environmental Reservoirs and Systems). We thank all members of Laboratory for Urgent Response to Biological Threats (CIBU) and Système d'Astreinte Microbiologique for their contribution and G.K. Kennedy, P. Dickinson and Affymetrix's staff, who were involved in the design and manufacture of this second generation of microarrays (UC1 AI062613 (Kennedy) from the US National Institute of Allergy and Infectious Diseases, National Institutes of Health).

Nicolas Berthet[1], India Leclercq[2,3], Amélie Dublineau[2], Sayuri Shigematsu[2], Ana Maria Burguière[2], Claudia Filippone[1], Antoine Gessain[1] & Jean-Claude Manuguerra[2]

[1]Institut Pasteur, Unit of Epidemiology and Pathophysiology Oncogenic Virus and CNRS URA3015, Paris, France. [2]Institut Pasteur, Laboratory for Urgent Response to Biological Threats (CIBU), Paris, France. [3]Université Paris Diderot, Paris, France.
e-mail: jean-claude.manuguerra@pasteur.fr

1. World Health Organization. Avian Influenza A(H5N1) in humans and poultry in Viet Nam Update 1 (WHO, Geneva; 2004).
2. World Health Organization. Influenza A(H1N1) - Update 1 (WHO, Geneva; 2009).
3. Berthet, N. *et al. BMC Mol. Biol.* **9**, 77 (2008).
4. Smith, G.J. *et al. Nature* **459**, 1122–1125 (2009).
5. World Health Organization. Influenza vaccine for 1997–1998 season (WHO, Geneva; 1997).
6. Brown, I.H. *et al. J. Gen. Virol.* **78**, 553–562 (1997).
7. Reid, A.H., Fanning, T.G., Janczewski, T.A. & Taubenberger, J.K. *Proc. Natl. Acad. Sci. USA* **97**, 6785–6790 (2000).
8. Meijer, A. *et al. Emerg. Infect. Dis.* **15**, 552–560 (2009).
9. Wang, R., Sheng, Z.M. & Taubenberger, J.K. Detection of novel (swine origin) H1N1 influenza A virus by quantitative real-time RT-PCR. *J. Clin. Microbiol.* **47**, 2675–2677 (2009).
10. Lu, Q. *et al.* Detection of the 2009 swine-origin influenza A (H1N1) virus by a subtyping microarray. *J. Clin. Microbiol.* **47**, 3060–3061 (2009).
11. Cutler, D.J. *et al. Genome Res.* **11**, 1913–1925 (2001).

© 2010 Nature America, Inc. All rights reserved.

Clinical comparability and European biosimilar regulations

Huub Schellekens & Ellen Moors

Clinical trials required by European regulators to compare biosimilar products with corresponding biologic brands are surplus to requirements and may even be a barrier for the development of biosimilars of more complicated biologics.

© 2010 Nature America, Inc. All rights reserved.

In 2004, the European Union (EU; Brussels) adopted legislation to establish a comprehensive regulatory pathway for bringing biosimilars to market. Currently, the European Commission has approved six of these products. On the basis of the European Public Assessment Reports (EPARs), which summarize the regulatory data used as a basis for granting marketing authorization, the quality of the biosimilars seems to be equal to or better than the originals. The mandatory clinical trials outlined in the regulations have shown these products to be effective and safe. Even so, the guidelines also require a comparability exercise intended to show the quality, safety and efficacy of the biosimilar to be comparable to the original product. We contend that this comparability exercise is of debatable value; indeed, it may even be a barrier for the development of biosimilars of more complicated biologics. For this reason, we suggest that the requirement for comparability studies for biosimilars be dropped. At the same time, the revised regulatory pathway used in the EU should be expanded to include complex pharmaceuticals other than biologics.

Biosimilars emerge

The first recombinant DNA–derived therapeutic proteins were introduced in the 1980s. These were mainly copies of endogenous human proteins, such as erythropoietin (EPO), insulin, growth hormones and cytokines. Such recombinant proteins were followed by the first monoclonal antibodies (mAbs) produced by hybridoma technology—products that have become important treatment options in clinical practice for many diseases, including anemia, diabetes, cancer, hepatitis and multiple sclerosis[1].

The patents for many of these first wave of biopharmaceuticals have expired or are about to expire, opening the possibility for marketing noninnovator versions of these products. When the patent of a classic small-molecule drug expires, generics may be marketed if their therapeutic equivalence to the original drug has been established. Conventional generics are considered to be therapeutically equivalent to a reference once pharmaceutical equivalence (that is, identical active substances) and bioequivalence (that is, comparable pharmacokinetics) have been established and do not require formal clinical efficacy and safety studies. This relatively modest requirement is one of the major reasons generics can be marketed far below the price of innovator drugs.

The generic approach cannot, however, be applied to copies of therapeutic proteins because of their complexity. After a fierce debate between regulatory bodies, the brand biotech industry and companies planning to introduce noninnovator versions of protein drugs, a consensus was reached on the need for clinical data to substantiate the clinical equivalence of these products.

Because it is impossible to show two protein products to be identical, the term 'biosimilars' was introduced in the EU and 'follow-on protein products' or 'biogenerics' in the United States. Pioneering law in this area, the EU adopted legislation in 2004 to establish a comprehensive regulatory pathway for bringing biosimilars to market. Subsequently, the European Medicine Agency (EMEA, London) and its scientific Committee for Medicinal Products for Human Use (CHMP) developed guidance documents to provide more detail on the requirements[2–9]; the United States is expected to establish similar pathways in the coming months[10] that allow separate marketing approval after patent expiration and adopt other provisions to protect intellectual property, such as data exclusivity of the reference products[3,11–13]. To be allowed on the market, the biosimilar product should be shown to be similar to the reference product in terms of quality, safety and efficacy.

The EMEA (headquarters in Canary Wharf, London, pictured here) has pioneered the regulatory oversight of biosimilars.

Huub Schellekens is in the Departments of Pharmaceutical Sciences and Innovation Studies and Ellen Moors is in the Department of Innovation Studies, Utrecht University, Utrecht, The Netherlands.
e-mail: h.schellekens@uu.nl

Table 1 Different biosimilars in the EU in March 2009

International generic name	Brand name reference product	Trade name biosimilar
Somatotropin	Genotropin	Omnitrope
Somatotropin	Humatrope	Valtropin
Epoetin alfa	Eprex	Abseamed, Binocrit and Epoetin alfa Hexal
Epoetin zeta	Eprex	Retacrit and Silap
Filgrastim	Neupogen	Biograstim, Filgrastim, Ratiopharm, Ratiograstim and Tevagrastim
Filgrastim	Neupogen	Filgrastim Hexal and Zarzio

Currently, based on advice from the EMEA CHMP, the European Commission (Brussels) has approved biosimilar versions of recombinant somatropin[14,15], recombinant human EPO[16–20] and recombinant filgrastim[21,22] (**Table 1**; the EPARs summarizing product characteristics and a scientific discussion of the data supporting each EMEA approval can be found in refs. 22–30).

Not all biosimilar applications have been successful thus far. The European regulator rejected Alpheon, a biosimilar version of interferon (IFN)-α-2a (ref. 31); and another biosimilar application concerning three different human insulin formulations was withdrawn in 2008 by Marvel Lifesciences (Mumbai, India)[32].

On the basis of our analysis of the criteria that the EMEA CHMP has applied to the evaluation of biosimilars and the strengths and weaknesses of European regulations, we consider in the following sections the question of whether the biosimilar regulatory pathway should be expanded to other complex pharmaceuticals than biologics.

What is a biosimilar?

Neither the EU legislation nor the EMEA CHMP guidelines provides a definition of a biosimilar other than it is a product comparable in quality, safety and efficacy to a reference product. The acceptable differences between biosimilar and reference products in these three major attributes are not stated. Thus, only the evaluation of what the EMEA CHMP accepts and rejects will define what a biosimilar is.

Table 2 lists the types of differences between a biosimilar and innovator product that have been allowed thus far by the EMEA CHMP. The list includes completely different host cells and formulations, differences in the level of impurities and in the types and levels of glycosylation. These variations are known to have the potential to have a major effect on a product's clinical efficacy and safety. The clinical studies of biosimilars tested thus far, however, have shown that for the products under review, these differences have not compromised efficacy or influenced the level of adverse drug reactions in humans compared with the brand product.

Thus, the clinical data, which are mandatory for a marketing authorization request for biosimilars, enable evaluation of the biological consequences of both the differences found in the aspects of the biologics that can be characterized and the aspects that are missed by current analytical tools.

What's more, when the CHMP/EMEA's evaluations to date are examined, any difference in host cell expression system, purity and formulation appears acceptable if the clinical data show no negative effect. This raises the question of whether a comparison of the quality attributes of a biosimilar with the reference product is relevant.

The preclinical and clinical comparability exercise

The foundation of the EMEA CHMP regulatory framework for biosimilars is the EU legislation in the Human Code of 2004 (ref. 2). This stipulates that biologics recalcitrant to full characterization not only fall outside of traditional generic regulations but also require supplementary preclinical testing or human trials. An important part of the documentation for classic generics is comparative pharmacokinetic data. These data are also expected in a biosimilar approach. With classic generics, the comparative pharmacokinetic data are a surrogate for clinical trials. In contrast, for a biosimilar marketing application, clinical data are mandatory. This raises the question of why the comparative pharmacokinetic data are needed.

For classic small-molecule drugs, an 80–125% acceptance range for comparative pharmacokinetic data is used by regulators. According to the EMEA CHMP guidelines, this range does not apply to biosimilars and a range specific to every product should be predefined and justified. However, for many if not all biotech-derived therapeutic proteins, this either is impossible or can be established only in extensive clinical trials.

In practice no equivalence margin has been predefined in any of the studies of the biosimilars; mostly, the classic acceptance range of 80–125% was used *post hoc* (**Table 3**). In the majority of cases, either comparative pharmacokinetic data were not provided at all or one or more parameters were not within this *post hoc* defined acceptance range. In all cases, however, the EMEA CHMP has accepted these results based on the argument that clinical trials are required to demonstrate comparable efficacy and safety. The agency has also not provided a scientific and regulatory rationale for comparative pharmacokinetics.

The same holds true for clinical comparability. The usefulness of these studies is debatable (**Table 3**). In fact, only in the case of filgrastim were the direct clinical comparisons between biosimilar and innovative product done according to the regulations. In all other cases, this comparison was either lacking or incomplete or data showed that the biosimilar actually lacked clinical comparability!

Biosimilars rejected or withdrawn

Alpheon, a biosimilar version of Roferon-A (IFN-α-2a), was rejected by the EMEA in June 2006. The reasons included quality and clinical differences between Alpheon and the reference product, inadequate data on the stability of the active substance, inadequate validation of the process for the finished product and insufficient validation of immunogenicity testing[31].

Another biosimilar application concerning three different human insulin formulations with Humulin as reference product was withdrawn by Marvel[32]. The main concerns of the CHMP were that the comparability of the Marvel insulins and the Eli Lilly (Indianapolis) Humulin insulins had not been shown and the Indian company had not supplied enough information on how the active substance or the finished products were made and that the processes used to make them had not been validated.

Table 2 Quality differences between biosimilars and reference drug products

Different host cells	Different levels of impurities	Different formulation	Different glycosylation
Valtropin	Abseamed, Binocrit and Epoetin alfa Hexal	Retacrit and Silap	Abseamed, Binocrit and Epoetin alfa Hexal
	Zarzio and Filgrastim Hexal	Biograstim, Filgrastim, Ratiopharm, Ratiograstim and Tevagrastim	Retacrit and Silap
		Zarzio and Filgrastim Hexal	

© 2010 Nature America, Inc. All rights reserved.

That said, in the cases of the biosimilar IFN-α2a and Marvel insulins, there were not only major comparability issues but also other problems such as validation of analytical tools. So it remains a question whether the comparability issues alone would have resulted in a negative opinion of the CHMP.

The scope of the EU biosimilar regulations

With the exception of a few small peptides like somatostatin and calcitonin, it is impossible with current technology to fully characterize biologics, including highly purified biotechnologically derived therapeutic proteins, such as somatotropin, epoetins and filgrastim. Copies of these latter molecules were the first biosimilars to be approved by the European Commission under a new regulatory pathway. Even so, some inconsistencies remain.

The current European biosimilar regulatory pathway is restricted to biologics but apparently not all copies of biologic molecules qualify. The EMEA CHMP guidelines state that comparability exercises to demonstrate similarity are more likely to be applied to highly purified products, which can be thoroughly characterized (e.g., biotech-derived medicinal products), than to other types of biologics. The implicit assumption is, therefore, that the pathway does not apply to more poorly purified biologics that are even more complex and difficult to characterize than a highly purified recombinant biologic. This seems to contradict the rationale that biologic complexity, and our inability to adequately characterize complex proteins, necessitates the comparability exercise.

Biologics are defined in the EMEA CHMP guidelines as products of living cells. There are, however, other compounds that are as complex as biologics, whose characteristics are also highly dependent on the production process and which are impossible to characterize fully.

A good example is glatiramer acetate (Copaxone), which is used for the treatment of multiple sclerosis. Glatiramer acetate is a member of the glatiramoid class of products. It is the acetate salt of synthetic polypeptides containing four naturally occurring amino acids: L-glutamic acid, L-alanine, L-lysine and L-tyrosine. The glatiramer acetate in Copaxone is not a single molecular entity but rather a heterogeneous polypeptide mixture that contains a huge, perhaps incalculable number of polypeptides, which has not been fully characterized.

The precise mechanisms by which the Copaxone product exerts its pharmacological effects in individuals with multiple sclerosis are not fully elucidated but the drug is presumed to act as a modulator of the immune system[33]. Upon subcutaneous injection, Copaxone degrades into smaller peptides and free amino acids locally, resulting in low or undetectable serum concentrations of the drug or its metabolites. Moreover, glatiramer acetate need not be in the systemic circulation to exert its anti-inflammatory effects.

There have been attempts to develop a generic version of Copaxone (**Fig. 1**), which clearly failed. Analysis of different batches of this attempted generic also showed big differences between batches illustrating the difficulties in consistently producing this type of product.

Another example is the iron-sucrose complex (Venofer) that is used for the intravenous treatment of iron deficiencies. Recently, several copies of Venofer have been introduced that differ slightly in physical chemical characteristics and show considerable differences in efficacy and safety[34].

Considering the complexity of Copaxone and Venofer, it seems reasonable that they should be excluded from the EMEA's generic pathway. Taking this further, it seems reasonable that the biosimilar pathway should not be restricted just to biologics but should be applicable to all medicinal products that are complex and difficult to characterize.

Table 3 Preclinical and clinical equivalence discrepancies

Pharmacokinetics not according to guidelines and/or outside acceptance range	Clinical trials not according to guidelines and/or showing differences
Omnitrope (no comparison with reference product)	Omnitrope (no direct comparison with reference product)
Abseamed, Binocrit and Epoetin alfa Hexal (acceptance range not predefined, AUC after iv treatment outside range)	Valtropin (only partial comparison with reference product)
Retacrit and Silap (acceptance range not predefined. Correction needed to meet range)	Abseamed, Binocrit and Epoetin alfa Hexal (no comparison of subcutaneous administration)
Filgrastim Hexal and Zarzio (at the lower doses and after a multiple subcutaneous dose of 5 μg/kg outside acceptance range)	Retacrit and Silap (no comparison of subcutaneous administration; no comparability for mean dosage)

Conclusions

Europe was the first region in the world with a comprehensive legislative and regulatory pathway for the introduction of biosimilars. The two cornerstones of the EMEA CHMP guidelines are the need for clinical data and the comparability exercise to show biosimilarity in quality, efficacy and safety. Six biosimilars have been approved under this pathway. Their evaluation as described in EPARs confirms the need for clinical data to confirm the efficacy and safety of these products.

Even so, in our opinion, the merits and/or added value of the comparability exercise are questionable. The comparison of quality characteristics between the biosimilar and the reference product will always show differences. And with the improvement of the analytical tools, our ability to find differences will only increase. In most cases, the consequences of these differences are unknown; for example, how does one assess the effects of a reduced level of *O*-glycosylated isoform in epoetin zeta compared with Eprex? In any case, the quality differences become irrelevant if the clinical data show the biosimilar to be clinically equivalent to the reference product.

According to the EPARs, the biosimilar epoetin alfa and one of the biosimilar filgrastims

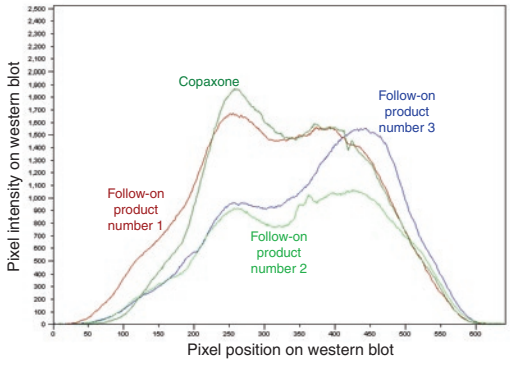


Figure 1 Western blot analysis of Copaxone and three follow-on products. Reprinted by permission of Teva Pharmaceuticals.

© 2010 Nature America, Inc. All rights reserved.

have fewer impurities and less modified product than their reference products. We have recently analyzed the physical chemical characteristics of both biosimilar epoetins and have found the quality of these biosimilars to exceed the original product.

An additional aspect for regulators to consider is technology obsolescence. Since the introduction of the first recombinant DNA–derived therapeutic proteins, the technology to produce and purify these products has greatly improved. Biosimilar manufacturers are consequently using state-of-the-art technology; in contrast, brand manufacturers of the original products are often locked into old technologies because changing methods has major financial and regulatory consequences. With this in mind, it seems much more logical for regulators to expect biosimilars to be produced by the best technology on offer rather than to mandate that they are of comparable quality to the brands.

Furthermore, there are many reasons to question the usefulness of comparative pharmacokinetic trials. The assays to determine product levels are often too imprecise; the relation between pharmacokinetic parameters and clinical effect of biologics is unclear; the dose-response curve of therapeutic proteins is often bell shaped (meaning that widely differing protein levels have the same clinical effect); and the acceptance range for pharmacokinetics parameters between biosimilar and reference product are difficult or impossible to predefine and justify.

The regulatory demand for clinical comparison between biosimilar and reference product is also questionable considering the practical consequences. The majority of biosimilars approved did not meet the conditions of the different guidelines mainly because of reasons beyond the control of both the regulators and the manufacturers of the biosimilars; indeed, in one case, a direct comparison between the biosimilar and the reference product was completely absent in the dossier. Apparently, contrary to the regulations, a direct clinical comparison is not essential for evaluating the clinical efficacy and safety of a biosimilar.

Removing the mandatory comparability exercise from the guidelines does not mean that comparisons between biosimilar and original product are not important during their development. Manufacturers do comparisons between their biosimilars and the original products to set specifications for their production and purification and to validate their production methods and analytical tools. Comparative data may also be helpful for a biosimilar manufacturer to claim extrapolation of indication. And it may be important for marketing reasons.

Dropping the obligation to do the comparability exercise will also make it easier to develop more complex biosimilars, such as mAbs and vaccines, and will avoid the ethical and practical problems concerning the comparability studies of products with survival as the primary clinical end point.

European legislators and regulators have had the courage to be the first to introduce a pathway for the introduction of biosimilars. This has enabled the introduction of six high-quality products and also the possibility to evaluate the strengths and weaknesses of the regulations. We urge regulators to consider the experience with the first biosimilars to streamline current regulations and apply an even playing field so that the same regulatory principles are applied to complex pharmaceuticals other than recombinant proteins.

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

1. Avidor, Y., Mabjeesh, N.J. & Matzkin, H. *South Med. J.* **96**, 1174–1186 (2003).
2. European Parliament and Council. *Off. J. Eur. Union* **47**, 34–57 (2004).
3. Anonymous. Guideline on similar biological medicinal products. (EMEA, London, 2005; accessed 4 March 2008). <http://www.emea.europa.eu/pdfs/human/biosimilar/043704en.pdf>
4. Anonymous. Guideline on similar biological medicinal products containing biotechnology-derived proteins as active substance: quality issues. (EMEA, London, 2006; accessed 4 March 2008). <http://www.emea.europa.eu/pdfs/human/biosimilar/4934805en.pdf>
5. Anonymous. ANNEX to guideline on similar biological medicinal products containing biotechnology-derived proteins as active substance: non-clinical and clinical issues guidance on similar medicinal products containing recombinant erythropoietins. (EMEA, London, 2006; accessed 4 March 2008). <http://www.emea.europa.eu/pdfs/human/biosimilar/9452605en.pdf>
6. Anonymous. Annex to guideline on similar biological medicinal products containing biotechnology-derived proteins as active substance: non-clinical and clinical issues guidance on similar medicinal products containing somatropin. (EMEA, London, 2006; accessed 4 March 2008). <http://www.emea.europa.eu/pdfs/human/biosimilar/9452805en.pdf>
7. Anonymous. Annex to guideline on similar biological medicinal products containing biotechnology-derived proteins as active substance: non-clinical and clinical issues guidance on similar medicinal products containing recombinant human soluble insulin. (EMEA, London, 2006; accessed 4 March 2008). <http://www.emea.europa.eu/pdfs/human/biosimilar/3277505en.pdf>
8. Anonymous. Annex to guideline on similar biological medicinal products containing biotechnology-derived proteins as active substance: non-clinical and clinical issues guidance on similar medicinal products containing recombinant granulocyte-colony stimulating factor. (EMEA, London, 2006; accessed 4 March 2008). <http://www.emea.europa.eu/pdfs/human/biosimilar/3132905en.pdf>
9. Anonymous. Similar biological medicinal products containing biotechnology-derived proteins as active substance: Non-clinical and clinical issues. CHMP/42832/05 <http://www.emea.europa.eu/pdfs/human/biosimilar/4283205en.pdf> (Accessed 21 December 2009)
10. <http://www.patentdocs.org/2009/12/followon-biologics-news-briefs-no-10.html>
11. Crommelin, D.J. *et al. Eur. J. Hosp. Pharm. Sci.* **1**, 11–17 (2005).
12. Roger, S.D. *Nephrology* **11**, 341–346 (2006).
13. Schellekens, H. *Nephrol. Dial. Transplant.* **20**, 31–36 (2005).
14. <http://www.emea.europa.eu/humandocs/PDFs/EPAR/Omnitrope/H-607-PI-en.pdf> (Accessed 4 March 2008)
15. <http://www.emea.europa.eu/humandocs/PDFs/EPAR/valtropin/H-602-PI-en.pdf> (Accessed 4 March 2008)
16. <http://www.emea.europa.eu/humandocs/PDFs/EPAR/abseamed/H-727-en6.pdf> (Accessed 4 March 2008)
17. http://www.emea.europa.eu/humandocs/PDFs/EPAR/binocrit/H-725-en6.pdf
18. <http://www.emea.europa.eu/humandocs/PDFs/EPAR/epoetinalfahexal/H-726-en6.pdf> (Accessed 4 March 2008)
19. <http://www.emea.europa.eu/humandocs/PDFs/EPAR/silapo/H-760-PI-en.pdf> (Accessed 4 March 2008)
20. <http://www.emea.europa.eu/humandocs/PDFs/EPAR/retacrit/H-872-en6.pdf> (Accessed 4 March 2008)
21. <http://www.emea.europa.eu/humandocs/PDFs/EPAR/Omnitrope/060706en6.pdf> (Accessed 4 March 2008)
22. <http://www.emea.europa.eu/humandocs/PDFs/EPAR/valtropin/H-602-en6.pdf> (Accessed 4 March 2008)
23. <http://www.emea.europa.eu/humandocs/PDFs/EPAR/Omnitrope/060706en6.pdf> (Accessed 4 March 2008)
24. <http://www.emea.europa.eu/humandocs/PDFs/EPAR/valtropin/H-602-en6.pdf> (Accessed 4 March 2008)
25. <http://www.emea.europa.eu/humandocs/PDFs/EPAR/epoetinalfahexal/H-726-en6.pdf> (Accessed 4 March 2008)
26. <http://www.emea.europa.eu/humandocs/PDFs/EPAR/binocrit/H-725-en6.pdf>
27. <http://www.emea.europa.eu/humandocs/PDFs/EPAR/abseamed/H-727-en6.pdf> (Accessed 4 March 2008)
28. <http://www.emea.europa.eu/humandocs/PDFs/EPAR/silapo/H-760-en6.pdf> (Accessed 4 March 2008)
29. <http://www.emea.europa.eu/humandocs/PDFs/EPAR/retacrit/H-872-en6.pdf> (Accessed 4 March 2008)
30. Anonymous. CHMP assessment report for Filgrastim Hexal (EMEA, London; 2008) <http://www.emea.europa.eu/humandocs/PDFs/EPAR/FilgrastimHexal/H-918-en6.pdf> (Accessed 28 February 2009)
31. Anonymous. Questions and answers on recommendation for refusal of marketing application for alpheon. (EMEA, London, 2006; accessed 3 March 2008). <http://www.emea.europa.eu/pdfs/human/opinion/19089606en.pdf>
32. Anonymous. Questions and answers on the withdrawal of the marketing authorisation application for Insulin Human Rapid Marvel; Insulin Human Long Marvel; Insulin Human 30/70 Mix Marvel. (EMEA, London; 2008; accessed 28 February 2008). <http://www.emea.europa.eu/humandocs/PDFs/EPAR/insulinhumanrapidmarvel/419308en.pdf>
33. Blanchette, F. & Neuhaus, O. *J. Neurol.* **255 Suppl 1**, 26–36 (2008).
34. Toblli, J.E., Cao, G., Oliveri, L. & Angerosa., M. *Port. J. Nephrol. Hypert.* **23**, 53–63 (2009).

© 2010 Nature America, Inc. All rights reserved.

The intellectual property landscape for gene suppression technologies in plants

Cecilia L Chi-Ham, Kerri L Clark & Alan B Bennett

Reviewing the major features in the patent landscape of RNA-mediated gene suppression may aid the development of patent strategies that will support the next generation of genetically modified crops.

© 2010 Nature America, Inc. All rights reserved.

RNA-mediated gene suppression is a powerful technology to suppress the expression of targeted genes within plants, as well as most other organisms. In 2002, RNA interference (RNAi) was proclaimed by *Science* as the "breakthrough technology of the year" and by *Fortune* as a "billion dollar breakthrough." The recognition of RNAi-mediated gene suppression as an important experimental tool and its potential commercial application is further reflected in the patent landscape related to RNAi-mediated gene suppression, with an increasing number of patent applications seeking exclusive rights to RNAi-based discoveries. Recent publications summarizing the RNAi-based patent thicket in applications to human medicine point out legal uncertainties over who will own key RNAi intellectual property (IP) and the apprehension that this has created among investors[1,2].

Although a commercial human RNAi-based therapeutic is yet to be released, RNA-mediated gene suppression was used to produce the very first commercial genetically modified (GM) crop, the FLAVR SAVR tomato, in 1994. Here, we examine the scientific evolution of RNA-mediated gene suppression technologies used in agricultural biotech and the associated patent landscape. There is current and emerging IP in the United States with broad claims that are likely to influence the freedom to operate (FTO) for RNA-mediated gene suppression technologies used in the development of GM plants. However, early patented methods of RNA-mediated gene suppression, including antisense and co-suppression, are nearing the end of their patent life. As this IP approaches expiration it opens gaps in the patent landscape that may offer greater FTO. This survey of the major landmarks in the patent landscape of RNA-mediated gene suppression is one step in informing IP strategies that can support the next generation of genetically modified crops.

Cecilia L. Chi-Ham, Kerri L. Clark and Alan B. Bennett are at the Public Intellectual Property Resource for Agriculture, University of California Davis, Davis, California, USA.
e-mail: abbennett@ucdavis.edu

Discovery and application of RNA-mediated gene suppression

Antisense RNA-mediated gene suppression. Gene suppression triggered by naturally occurring antisense RNA was first identified in bacteria in 1983, suggesting the possibility for applying this or similar strategies to suppress gene expression in other organisms[3]. Ecker and Davis were the first to use antisense technology to induce transient inhibition of exogenous gene expression in plant cells[4] (**Fig. 1**). Subsequent studies used antisense transcripts to suppress constitutive expression of exogenous genes in whole tobacco plants[5,6]. Rothstein and colleagues also showed that gene suppression was heritable, suggesting this technology had practical applications in generating stable transgenic crops with improved agronomic traits. Subsequent scientific landmarks included the demonstration that antisense RNA could also be used to modulate expression of endogenous plant genes including those encoding chalcone synthase to modify flower pigmentation in petunia[7], polygalacturonase to modify fruit-ripening characteristics in tomato[8,9] and the photosynthetic RuBisCo in tobacco[10]. Although the mechanism of antisense RNA-mediated gene suppression was not clearly understood in these early experiments, the hypothesis suggested that the formation of an RNA-RNA hybrid destabilized the mRNA, resulting in its rapid degradation.

Antisense RNA-mediated gene suppression was quickly adopted by plant scientists as a broadly applicable method to downregulate expression of target genes. Antisense RNA was embraced in plant research because reverse genetic approaches used in other biological systems, such as homologous recombination and gene-tagging mutagenesis, were either not applicable or not yet well developed in plants. Antisense RNA-mediated gene suppression was quickly shown to be a powerful tool for both basic transgenic research and for the development of commercial products, which highlighted the value of this technology (**Box 1, Table 1**).

Co-suppression of gene expression. In 1990, an RNA-based gene suppression phenomenon, termed co-suppression, was serendipitously discovered in plants when researchers attempted to generate new purple varietals of petunia flowers by overexpressing pigment genes[11–13] (**Fig. 1**). The unexpected results of white variegated flowers were proposed to be the result of not only post-transcriptional suppression of the endogenous pigment gene but also the suppression of the exogenous overexpressed homologous gene, hence, the term co-suppression. This was the first reported demonstration that sense transgenes, like antisense transgenes, could induce gene silencing in plants.

Co-suppression and antisense RNA-mediated gene suppression technologies, however, faced similar technical limitations. The technologies did not silence gene expression completely and as a result, large numbers of transgenic plants had to be screened to identify those with desirable levels of gene suppression. The elucidation of the mechanisms underlying co-suppression and the induction of what is now termed RNA

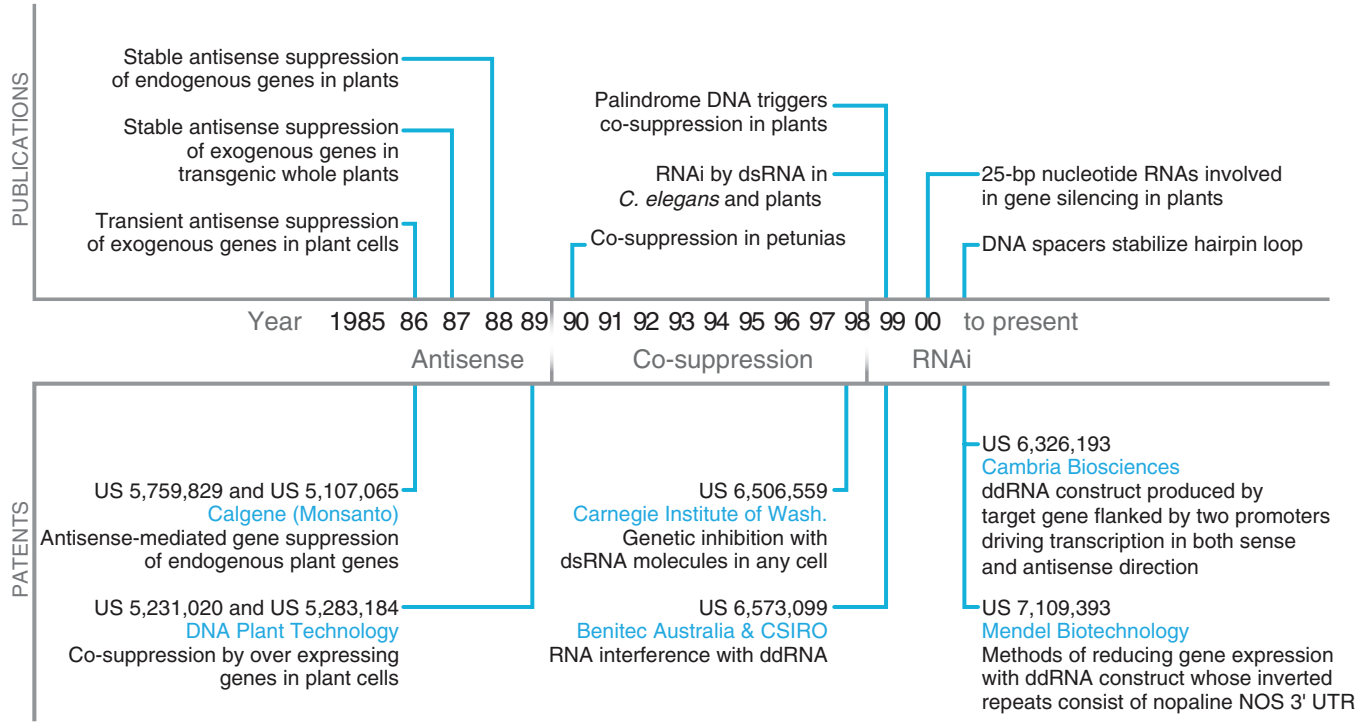


Figure 1 Scientific and US intellectual property milestones in RNA-mediated gene suppression in plant biotech. NOS, nopaline synthase.

© 2010 Nature America, Inc. All rights reserved.

interference led to the improvements required to efficiently and reproducibly suppress gene expression.

RNA interference. The term 'RNA interference' was coined in 1998 by Nobel laureates Fire, Mello and colleagues to describe a gene-silencing phenomenon induced by double-stranded RNA (dsRNA)[14]. The mechanism of RNAi-induced gene silencing was defined after several seminal discoveries that unveiled a common gene suppression pathway in plants, animals and nematodes. Fundamental to deciphering the RNAi mechanism was identifying RNA as the trigger for gene silencing. Central to this discovery was the unconventional observation that plant viral resistance, in some cases, could also be instigated by the coat-protein mRNA and not the translated viral coat protein[15,16]. The role of mRNA as a gene-silencing trigger was additionally established independently while the co-suppression phenomenon in petunia was further investigated[17,18]. Both Metzlaff and Stam proposed that the aberrant mRNAs paired with complementary endogenous mRNA and triggered RNA cleavage resulting in post-transcriptional gene silencing. Together, the data demonstrated that co-suppression was a consequence of the rapid degradation of mRNA that shared a high degree of sequence homology. Although it was proposed that the complementary RNA bound with targeted RNA and triggered cleavage, the role of dsRNA in triggering cleavage was unproven.

In 1998, Fire, Mello and colleagues and Waterhouse and colleagues independently demonstrated definitively that dsRNA was required to induce gene silencing in nematodes and co-suppression in plants, respectively (**Fig. 1**). Fire and colleagues injected purified sense and antisense oligonucleotides separately or together into *Caenorhabditis elegans*. Co-introduction of sense and antisense strands, which formed dsRNA, resulted in gene suppression two orders of magnitude higher than when the oligonucleotide strands were introduced separately[14,19]. Waterhouse and colleagues hypothesized that co-suppression was triggered by dsRNA that was formed by the hybridization of complementary transgene mRNAs or complementary regions of the same transcript[20]. They also theorized that the presence of complementary regions in the same transcript could be mimicked by the insertion of multiple transgene copies in a palindromic orientation resulting in a read-through transcript similar to that observed by Stam[18]. To test this hypothesis, they designed a plasmid such that a single transcribed RNA would form a double-stranded hairpin structure. Transformation of rice by this dsRNA plasmid into suppressed target gene expression significantly more than plasmids encoding a single gene in the sense or antisense orientation[20]. This powerful technology is now known as DNA-directed RNA (ddRNA). Hamilton and Baulcombe further defined the mechanism of RNAi when they discovered small RNA species, of ~25 nucleotides, in plants undergoing co-suppression, that were absent in nonsilenced plants[21] (**Fig. 1**). They also noted these species were complementary to the silenced gene. These short interfering RNAs (siRNAs) are now known to be the functional form of RNAi. Collectively, these foundational findings became the basis to develop tools to efficiently trigger gene silencing in plants and animals.

Gene-silencing DNA constructs and plasmids. The deduction that dsRNA triggers RNAi led to the development of methods to produce dsRNA and efficient gene silencing. Using synthetic oligonucleotides, as Fire and Mello did, induces transient gene suppression but is dependent on efficient transformation of the plant cell by the oligonucleotides, a traditionally inefficient process. However, *in vivo* transcription of the dsRNA can potentially yield stable gene suppression. Waterhouse and colleagues demonstrated that ddRNA

Box 1 RNA gene suppression in the agbiotech pipeline

The discovery and application of gene-suppression strategies, such as antisense and RNAi, to modify phenotypes led companies to use these strategies to develop transgenic crops. In 1994, the use of antisense RNA–mediated gene suppression technology gave birth to the first commercial transgenic crop approved by the US and other international regulatory agencies: Calgene's FLAVR SAVR tomato[32]. Engineered to express the endogenous polygalacturonase gene in the antisense orientation, FLAVR SAVR showed reduced expression of this cell wall–degrading enzyme, consequently delaying the softening of tomatoes. FLAVR SAVR's international commercial release was a historic milestone marking antisense technology's usefulness in applied research. Other transgenic plants in the US regulatory pipeline that use co-suppression or antisense to downregulate gene expression are listed in **Table 1**. Currently, there are no commercial GM crops that use RNAi. With the recent advent of optimized methods to efficiently trigger RNAi gene silencing and interest in quality crop traits that require precise regulation of steps in metabolic pathways, we anticipate new RNAi-based, genetically modified cultivars will enter the commercialization pipeline. However, the extent to which this will be observed depends on the FTO within the patent landscape.

Table 1 A summary of commercial developments in agbiotech developed with RNA-mediated gene suppression.

Company	Crop	Trait gene	RNA-based gene suppression approach	Regulatory approval (animal feed, human food and/or environmental)	Phenotypic description
Calgene (now Monsanto)	Tomato (FLAVR SAVR)	Polygalacturonase	Antisense	US, Canada, Mexico, Japan	Delayed fruit ripening
Zeneca London, UK	Tomato	Polygalacturonase	Antisense and co-suppression	US, Canada, Mexico	Delayed fruit ripening
DNA Plant Technology	Tomato	Aminocyclopropane cyclase	Co-suppression	US, Canada, Mexico	Delayed fruit ripening
Vector Tobacco Durham, NC	Tobacco	Quinolinic acid phosphoribosyltransferase	Antisense	US	Reduced nicotine levels
DuPont Canada Agricultural Products Ontario, Canada	Soybean	Fatty acid desaturase	Co-suppression	US, Canada, Japan, Australia	High oleic acid soybean
Florigene Pty. Ltd.	Carnation	1-aminocyclopropane-1-carboxylic acid	Co-suppression	Australia, European Union	Longer vase life
US Department of Agriculture	Plum	Plum pox virus coat protein	Co-suppression	US	Viral resistance

Data source: Agbios (http://www.agbios.com)

© 2010 Nature America, Inc. All rights reserved.

that encodes an RNA hairpin could suppress target gene expression by up to 90% in transgenic plants[20]. The level of gene suppression was further increased to almost 100% by inserting a DNA spacer (intron) between the complementary inverted sequences to stabilize the hairpin loop structure[22] (**Fig. 2**). Additionally, Brummell and colleagues, using complementary inverted nopaline synthase 3′ untranslated transcription region (UTR), demonstrated that the inverted complementary regions used to form a hairpin could be composed of DNA sequences other than those of the target gene[23]. This ddRNA construct is suitable for high-throughput cloning as only a single copy of the target gene needs to be present (**Fig. 2**).

DsRNA can also be produced by transcribing individual strands of sense and antisense transcripts. Individual strands of RNA can be produced by transcribing the RNA from separate plasmids or from a single plasmid using either two promoters to drive expression of a sense or antisense DNA, or two opposing promoters to drive the transcription of the sense and antisense RNA (**Fig. 2**). A number of plasmids based on these various structural arrangements are now available to facilitate cloning of RNAi plasmids for gene silencing in plants. The ongoing improvement and development of new biotech methodologies continues to fuel the evolution of plasmids for RNA-based gene suppression.

Landscape of RNA-mediated gene suppression patents in the United States

Antisense RNA-mediated gene suppression. The antisense RNA-mediated gene suppression technology used to create the FLAVR SAVR tomato was also the basis for one of the first patents with broad claims in the field of gene suppression. Calgene (now St. Louis-based Monsanto) was awarded two US patents, US 5,107,065 and US 5,759,829 (refs. 24,25) (**Fig. 1**). These are dominant patents in the RNA-mediated gene suppression landscape because of their broad claims, which describe antisense-mediated suppression of any gene indigenous to the plant cell. At the time, the mechanism of gene suppression was thought to occur through complementary hybridization of the antisense and endogenous, sense mRNA transcripts. However, in 2005, Monsanto, knowing the mechanism of RNAi-mediated gene silencing, performed a retrospective examination of the transgene insertions that comprised the FLAVR SAVR tomato[26] and the example cited in US 5,107,065 and US 5,759,829. The structural analysis of the inserted transgenes revealed the presence of tandem T-DNA insertions resulting in the possible formation of an inverted double-stranded loop (**Fig. 2**), which is now known as a structure that triggers RNAi-induced gene silencing. Monsanto's antisense patents were awarded in the early 1990s and expired in 2009.

Co-suppression of gene expression. Prior to the identification of dsRNA as a trigger for co-suppression in plants, the method to induce co-suppression, although unreliable, was to overexpress a sense gene in plants. DNA Plant Technology (Oakland, CA, USA) was awarded two US patents, US 5,231,020 and US 5,283,184 (refs. 27,28), both with broad claims to achieve endogenous gene silencing by overexpressing a sense gene in any plant

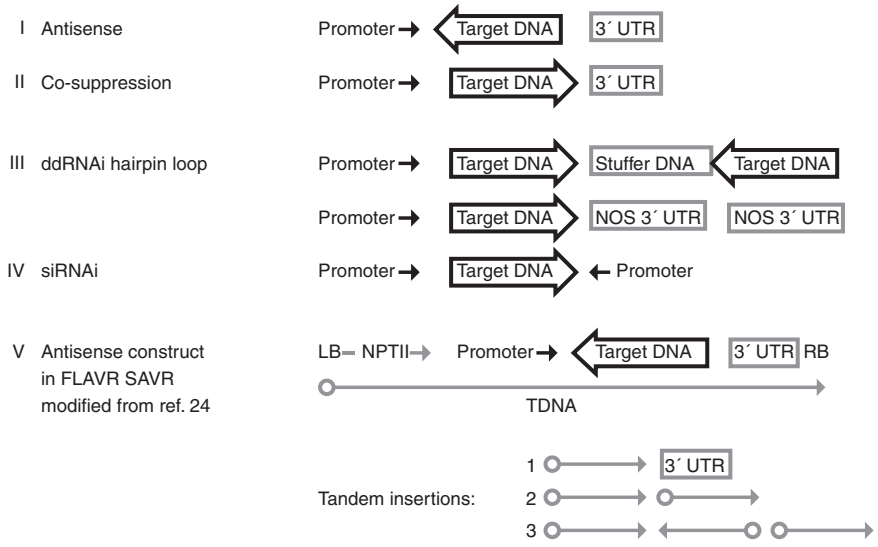


Figure 2 RNA Gene-silencing constructs commonly used in plant genetic modification. NOS, nopaline synthase.

(**Fig. 1**). Both patents specify altering the phenotype of the plant by suppressing endogenous genes in a plant cell. Like Monsanto's antisense IP, these broad co-suppression patents were filed in the early 1990s and are either expired or will expire in 2010.

RNAi-induced gene suppression. RNAi can be induced by the delivery of siRNAs or DNA constructs encoding complementary RNA into cells. Fire, Mello and colleagues at the Carnegie Institute of Washington (Washington, DC, USA) were awarded a broad patent, US 6,506,559 (ref. 29), also known as the Carnegie patent, which is now widely considered the most fundamental patent in the RNAi field. The patent encompasses the general method of using dsRNA formed with complementary copies of the target gene to downregulate gene expression in a cell (plant or animal) *in vitro*. Based on the construction of the claims, it is arguable that the exclusivity of the claims is limited to gene suppression *in vitro*. It is unclear if the patent rights include applications of dsDNA to suppress gene expression *in vivo*, such as would be the case in living transgenic plants.

Fire and Mello's fundamental patent also spawned related-child patent applications including one US application serial no. 10/283,267 with pending claims specific for use of dsRNA in plants. The examination of this application was suspended. However, the subject matter has the potential to be dominant in the plant biotech patent landscape. The United States Patent and Trademark Office (USPTO) recently published three additional patent applications submitted by Carnegie in 2007 (serial nos. 11/826,385; 11/905,449; 11/905,368). These applications encompass the use of dsRNA *in vivo* in plant and animal cells, which, if awarded, may substantially affect the RNAi IP landscape.

RNAi DNA constructs. The development of the ddRNA technology by Waterhouse and colleagues, unlike synthetic siRNA, enabled the stable production of dsRNA in a cell and stable gene suppression in transgenic plants. This research formed the basis of one of the most pertinent US patents in animal biotech fields, US 6,573,099 (ref. 30). This patent was co-awarded to Benitec Australia (Melbourne) and the Commonwealth Scientific and Industrial Research Organization (CSIRO, Canberra) and described the use of ddRNA constructs in animal cells. Benitec's US 6,573,099 patent family encompasses 11 patent applications. Three of the current applications, including serial nos. 09/646,807, 11/218,999 and 11/180,928 contain pending claims that describe the process of using ddRNA plasmids to suppress gene expression in eukaryotic cells and plants.

Indicative of the potential dominance of Benitec's US 6,573,099 within the US RNAi landscape, this key patent has been challenged in the USPTO. In 2004, after Benitec sued Nucleonics (Horsham, PA, USA) and two other companies for alleged infringement of their patent, Nucleonics aggressively challenged Benitec's IP by requesting two successive reexaminations of US 6,573,099 by the USPTO. In April 2008, Benitec cancelled several claims and the remaining claims were rejected by the USPTO as being obvious in view of prior art, including the Carnegie patent US 6,506,559, which Nucleonics licensed (**Fig. 2**). In response to the USPTO, Benitec/CSIRO claims that they were the first to invent. Though most international patent systems follow the first-to-file priority, the USPTO follows the first-to invent priority. Until the IP issues are resolved by the USPTO, uncertainty remains on one of the most critical patent estates for deploying RNAi in animals and plants.

To expedite DNA cloning manipulations in generating RNAi constructs, Brummell and colleagues devised a strategy amenable to high-throughput gene-silencing experiments. This method is the subject of Mendel Biotechnology's (Hayward, CA, USA) US patent 7,109,393. The patent claims a method to suppress gene expression in a plant cell by expressing a cassette that encodes two inverted nopaline synthase 3′ UTRs interrupted by spacer DNA; this structure forms a loop at the 3′ termini of the target gene to be silenced (**Fig. 1**). The patent's subject matter is not specific to a particular target gene and thus broadly encompasses any target gene. However, the invention's claims are limited to the use of inverted repeats from the gene encoding nopaline synthase, in particular the 5′ and 3′ UTRs, leaving open the possibility of using other inverted repeat sequences to achieve the same effect in suppressing expression of a target gene. Mendel's patent issued in 2006 and is expected to expire in 2020.

A novel plasmid designed to produce dsRNA using dual promoters (US 6,326,193) was patented by Cambria Biosciences (Woburn, MA, USA). The specified inventions include a plasmid that contains a single DNA segment that can be transcribed in both directions by separate promoters placed in opposite orientations (**Fig. 2**). Other narrower patent claims describe the use of the expression plasmids as a biological pest control agent. The term of this patent is expected to expire in 2019.

Discussion

Scientific research advancements are often nurtured by a highly synergistic environment, as was evidently the case for the deduction of the mechanism of RNAi-mediated gene suppression (**Fig. 2**). In contrast, when claiming patent rights, patent law favors clear and well-defined invention boundaries. Failure to invent around or license technologies claimed in patents may trigger legal repercussions and even prevent product commercialization. Thus, IP due diligence to evaluate potential legal risks is an important step in the commercialization of products and requires both

© 2010 Nature America, Inc. All rights reserved.

technical and legal expertise[31]. A key aspect of this process is assessing the FTO of a product, that is, does the developed product infringe on third-party proprietary IP? Because IP laws vary between countries and patents have national boundaries, it is essential to perform an FTO IP analysis for each country in which products will be developed and deployed. Here, we presented an overview of the scientific development and IP landscape of RNA-mediated gene suppression technologies used in agbiotech in the United States. The application of RNA-mediated gene suppression to produce GM organisms evolved from strategies based on expression of target genes in antisense orientation, to co-suppression by overexpressing sense transcripts and then to producing dsRNA. There now exist both emerging as well as expiring patents in the United States for the general use of RNAi in plants, and DNA constructs that mediate dsRNA production.

Currently, there is substantial FTO for the use of RNA-mediated gene suppression in plants in the United States. The broadest US patent in RNAi is awarded to the Carnegie Institute of Washington, US 6,506,559. Patent rights were awarded for the *in vitro* use of dsRNA in controlling gene expression in plants and animals. It is unclear if the breadth of the claims would also encompass the *in vivo* use of RNAi. However, a conservative IP strategy would be to consider licensing this patent, which Carnegie offers nonexclusively. Both the patent families of Benitec/CSIRO and Carnegie Institute contain related pending patent applications seeking to gain patent rights toward very similar subject matter related to the production and use of dsRNA molecules for gene suppression in plants. Owing to the potential for overlapping subject matter and the commercial potential of this technology, the prospect of ongoing opposition to these pending applications is likely. Until the patent application prosecution processes are completed, it is difficult to assess the repercussions of these developing claims in the RNAi patent landscape.

Gene-suppression technologies expected to have greater FTO in the near future are the RNAi predecessors technologies: antisense and co-suppression. Though there is broad IP issued to these technologies, their patent lives are expired or nearing expiration. The broadest patent claims on RNA-based gene suppression based on antisense-suppression was awarded to Calgene (now Monsanto) and expired in 2009. Co-suppression by overexpression of sense genes in plant cells was another novel invention aimed at suppressing expression in plant cells. Broad patent claims for this technology were awarded to DNA Plant Technology in the 1980s and are scheduled to expire soon. Although antisense-based and co-suppression induced gene suppression methods were considered technically less optimal than dsRNA, modern high-throughput genetic screening methods and the expiration of these patents may make these alternatives more attractive.

This RNAi patent landscape highlights the legal complexities in any given technology space. It illustrates some of the FTO IP considerations developers must consider in generating new agricultural or pharmaceutical products. Evolving patent landscapes create a great deal of uncertainty in making product development and investment decisions that rely on a realistic FTO assessment. Currently the average processing time for US patent applications is 40 months; however, in emergent technology areas this time frame can be prolonged. This is the case for the patent application by Carnegie Institution and Benitec/CSIRO, which remain under prosecution 10 years after the initial filing date. RNAi is continuing to develop as a fundamental tool in both plant and animal biotech and an ongoing assessment of the patent landscape will be important to equip scientists and investors with knowledge for evaluating FTO in this technology sector.

ACKNOWLEDGMENTS

We thank J. Harvey and R. Riley-Vargas for critical discussion and contributions to this paper. We are grateful to N. Fong for graphic assistance.

1. Schmidt, C. *Nat. Biotechnol.* **25**, 273–275 (2007).
2. Howard, K. *Nat. Biotechnol.* **21**, 1441–1446 (2003).
3. Simons, R.W. *Gene* **72**, 35–44 (1988).
4. Ecker, J.R. & Davis, R.W. *Proc. Natl. Acad. Sci. USA* **83**, 5372–5376 (1986).
5. Rothstein, S.J., Dimaio, J., Strand, M. & Rice, D. *Proc. Natl. Acad. Sci. USA* **84**, 8439–8443 (1987).
6. Delauney, A.J., Tabaeizadeh, Z. & Verma, D.P. *Proc. Natl. Acad. Sci. USA* **85**, 4300–4304 (1988).
7. van der Krol, A.R. *et al. Nature* **333**, 866–869 (1988).
8. Sheehy, R.E., Kramer, M. & Hiatt, W.R. *Proc. Natl. Acad. Sci. USA* **85**, 8805–8809 (1988).
9. Smith, C.J.S. *et al. Nature* **334**, 724–726 (1988).
10. Rodermel, S.R., Abbott, M.S. & Bogorad, L. *Cell* **55**, 673–681 (1988).
11. van der Krol, A.R., Mur, L.A., Beld, M., Mol, J.N. & Stuitje, A.R. *Plant Cell* **2**, 291–299 (1990).
12. van der Krol, A.R., Mur, L.A., de Lange, P., Mol, J.N. & Stuitje, A.R. *Plant Mol. Biol.* **14**, 457–466 (1990).
13. Napoli, C., Lemieux, C. & Jorgensen, R. *Plant Cell* **2**, 279–289 (1990).
14. Fire, A. *et al. Nature* **391**, 806–811 (1998).
15. Pang, S.Z., Slightom, J.L. & Gonsalves, D. *BioTechnology (NY)* **11**, 819–824 (1993).
16. Lindbo, J.A. & Dougherty, W.G. *Virology* **189**, 725–733 (1992).
17. Metzlaff, M., O'Dell, M., Cluster, P.D. & Flavell, R.B. *Cell* **88**, 845–854 (1997).
18. Stam, M., Viterbo, A., Mol, J.N. & Kooter, J.M. *Mol. Cell. Biol.* **18**, 6165–6177 (1998).
19. Montgomery, M.K., Xu, S. & Fire, A. *Proc. Natl. Acad. Sci. USA* **95**, 15502–15507 (1998).
20. Waterhouse, P.M., Graham, M.W. & Wang, M.B. *Proc. Natl. Acad. Sci. USA* **95**, 13959–13964 (1998).
21. Hamilton, A.J. & Baulcombe, D.C. *Science* **286**, 950–952 (1999).
22. Smith, N.A. *et al. Nature* **407**, 319–320 (2000).
23. Brummell, D.A. *et al. Plant J.* **33**, 793–800 (2003).
24. Shewmaker, C.K., Kridl, J.C., Hiatt, W.R., Knauf, V. US Patent no. 5,107,065 (1992).
25. Shewmaker, C.K., Kridl, J.C.; Hiatt, W.R., Knauf V. US Patent no. 5,759,829 (1998).
26. Sanders, R.A. & Hiatt, W. *Nat. Biotechnol.* **23**, 287–289 (2005).
27. Jorgensen, R.A. & Napoli, C.A. US Patent no. 5,231,020 (1993).
28. Jorgensen, R.A. & Napoli, C.A. US Patent no. 5,283,184 (1994).
29. Fire, A., *et al.* US Patent no. 6,506,559 (2003).
30. Graham, M.W. US Patent no. 6,573,099 (2003).
31. Fenton GM, C Chi-Ham and S Boettiger. Freedom to operate: The law firms approach and role. in *Intellectual Property Management in Health and Agricultural Innovation: A Handbook of Best Practices* (eds. A Krattiger, RT Mahoney, L Nelsen, *et al.* MIHR, Oxford, UK & PIPRA, Davis, US) (2007).
32. Kramer, M. & Redenbaugh, K. *Euphytica* **79**, 293–297 (1994).

© 2010 Nature America, Inc. All rights reserved.

Recent patent applications in induced pluripotent stem (iPS) cells

Patent number	Description	Assignee	Inventor	Priority application date	Publication date
WO 2009137624	A method of generating and expanding human hemangio-colony forming cells *in vitro*, "comprising" culturing cell culture "comprising" pluripotent stem cells and adding growth factors (bone morphogenic protein) to a culture comprising embryoid bodies.	Advanced Cell Technology (Worcester, MA, USA)	Lanza R, Lu S	5/6/2008	11/12/2009
WO 2009136867	Effecting dedifferentiation of a partially differentiated cell or of maintaining pluripotency or self-renewing characteristics of an undifferentiated cell by increasing the amount or activity of an estrogen-related receptor protein in the cell.	Agency for Science, Technology & Research (Singapore)	Feng B, Jiang J, Kraus P, Lufkin T, Ng HH	5/6/2008	11/12/2008
WO 2009137629	Producing a pluripotent stem cell–derived enucleated erythroid cell, involving providing a pluripotent stem cell and differentiating the pluripotent stem cell into an enucleated erythroid cell by culturing the pluripotent stem cell with OP9 mouse stromal cells or human mesenchymal stem cells.	Advanced Cell Technology (Worcester, MA, USA)	Lanza R, Lu S	5/6/2008	11/12/2009
US 20090280096	A pluripotent stem cell modified to overexpress Pdx1 and Ngn3; useful for manufacturing a medicament for treating an individual in need of pancreatic cell therapy.	Bonham K, Kubo A, Snodgrass HR, Stull R, Vistagen Therapeutics (S. San Francisco, CA, USA)	Bonham K, Kubo A, Snodgrass HR, Stull R	5/9/2008	11/12/2009
JP 2009254340	A cell culture apparatus and conveyance container for embryonic stem cells, iPS cells and self-skeletal myoblasts, comprising stainless steel with certain surface uneveness and surface chromium concentration.	Nagai K	Nagai K	4/19/2008	11/5/2009
WO 2009133971	Producing an iPS cell comprising introducing at least one kind of non-viral expression vector incorporating at least one gene that encodes a reprogramming factor into a somatic cell.	Kyoto University (Kyoto, Japan)	Okita K, Yamanaka S	5/2/2008	11/5/2009
WO 2009131262	Manufacturing stem cells by mixing and reacting floated somatic cells and the virus solution to prepare a somatic cell-virus mixture and cultivating the somatic cells in which the genes are induced in a culture dish.	Mirae Biotech (Seoul)	Jeon K, Kim EY, Park SP	4/25/2008	10/29/2009
WO 2009128533	A method for producing mesenchymal stem cells capable of differentiating into myoblasts by culturing a pluripotent stem cell derived from human or animal, involving subculturing the pluripotent stem cell by keeping it in an undifferentiated state, culturing the subcultured pluripotent stem cell under conditions that enable induction of differentiation of the cell into a fat cell *in vitro*, and separating and collecting the CD105-positive cell.	National University Corp., Nagoya University (Aichi, Japan)	Ninagawa N, Torihashi S	4/18/2008	10/22/2009
WO 2009122747	A net-like structure having a hemopoietic progenitor cell, obtained by seeding an iPS cell derived from human on a feeder cell, and culturing the iPS cell under the conditions suitable for the induction of differentiation of the iPS cell into the hemopoietic progenitor cell.	Kyoto University (Kyoto, Japan), Tokyo University (Tokyo)	Eto K, Nakauchi H, Nishiki-i H, Takahashi K, Takayama N, Yamanaka S	4/1/2008	10/8/2009
WO 2009140655	A viral vector for transduction of a somatic cell to induce a pluripotent stem cell, comprising in serial array a stem cell responsive promoter element, a pluripotent stem cell transcription factor element, a stem cell non-responsive promoter element and a reporter element.	Primegen Biotech (Irvine, CA, USA)	Javier C, Kannemeier C, Pham J, Sundsmo J	5/15/2008	5/15/2009

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

© 2010 Nature America, Inc. All rights reserved.

Putting the lid on phosphodiesterase 4

Miles D Houslay and David R Adams

Structural insights into the regulation of phosphodiesterase 4 lead to the discovery of allosteric modulators with reduced side effects.

© 2010 Nature America, Inc. All rights reserved.

Phosphodiesterase 4 (PDE4), the major enzyme for degrading cAMP in cells, has long presented a tantalizing target for therapeutic intervention[1], not least because of its genetic associations with schizophrenia, stroke, asthma, osteoporosis and prostate cancer[2]. Although numerous PDE4 inhibitors have been developed, their deployment has been plagued by side effects such as nausea and emesis[3]. In this issue, an exciting contribution from Burgin *et al.*[4] provides a quantum leap forward in our appreciation of PDE4 structure, its regulation by phosphorylation and protein-protein interactions, and the structural basis of inhibitor action. The authors use these insights to develop a series of >800 compounds, many of which exert an allosteric downregulatory action on PDE4 activity that results in greatly reduced side effects in several animal models. The study thus provides a new paradigm for development of PDE4 inhibitors that might at last allow the rich therapeutic potential of these compounds to be realized.

PDE4 is a member of a large superfamily of phosphodiesterases that provide the sole means for degrading cyclic nucleotides. There are several approved drugs that target phosphodiesterases, including a selective inhibitor of PDE3 (cilostazol) for treatment of intermittent claudication and selective inhibitors of cGMP-specific PDE5 (e.g., Viagra) for erectile dysfunction, proving that this superfamily is druggable. The critical cAMP signalling pathway has also been targeted by many therapeutics that interact with G-protein-coupled receptors, thereby controlling the generation of cAMP. However, the development of therapeutics that inhibit cAMP degradation by PDE forms found at specific intracellular locales and with cell type–specific patterns of expression would allow a new range of critical disease processes to be addressed.

Four PDE4 genes (A/B/C/D) encode ~25 isoforms distinguished by unique N-terminal regions for targeting to specific intracellular signaling complexes involved in compartmentalized cAMP signalling[2]. These isoforms are further classified based upon the presence of regulatory upstream conserved region (UCR) domains located between the isoform-specific extreme N-terminal region and the core catalytic domain, which shows high homology across all isoforms. Thus, long PDE4 forms have both UCR1 and UCR2, short forms have UCR2 and super-short forms have a truncated UCR2 (ref. 2). These regions orchestrate activity changes upon regulatory phosphorylation by kinases such as PKA and ERK, and also affect targeting and sensitivity to inhibition[2]. Biochemical studies have indicated that UCR1 and UCR2 interact[5] and that UCR2 may bind the catalytic unit[6].

A central question in understanding PDE4 function relates to how UCR2 channels its regulatory action into the catalytic unit. This issue has fundamental relevance to important cellular processes such as desensitization to cAMP, in which PKA activation of PDE4 plays a pivotal role[1]. It also has implications for drug discovery, as regulatory events in the region N-terminal to the catalytic unit—phosphorylation of UCR1 by PKA or protein binding to UCR2 and to the sequence connecting UCR2 to the catalytic unit—can selectively alter the potency of certain (e.g. rolipram and RS25344), but by no means all, inhibitors[3]. Significantly, it has been proposed that these changes in sensitivity to inhibitors such as rolipram are linked to the adoption of distinct PDE4 conformational states and that these states may relate either to the side effects of inhibitors or to 'bonus' effects, in which functional responses to certain inhibitors are greater than those expected on the basis of simple competitive inhibition[3].

A molecular understanding of these issues has remained elusive, even after the appearance of crystal structures of the unliganded PDE4 core catalytic domain nearly a decade ago[7] and, subsequently, of numerous ligand-bound complexes. Now, Burgin *et al.*[4] have acquired the first PDE4 structures that show the core catalytic domain bound to UCR2, in complex with various inhibitors. Their work sheds light on some of the most important and longstanding questions relating to the mechanisms of regulatory control in the PDE4 family and has major implications for the design of novel PDE4 inhibitors.

UCR2 has long been known to exert an autoinhibitory action on the core catalytic domain[6], although the structural basis for this effect was unclear. The new structures identify an α-helical NQVSE[F/Y]ISXTFLD sequence within UCR2 that can fold across the catalytic pocket, thereby gating access to it (**Fig. 1a**). A dimeric catalytic domain assembly, seen in nearly all previous PDE4 crystal structures, is also observed here. Unfortunately, residues linking UCR2 to the core catalytic domain are disordered, and so it is not immediately clear whether UCR2 folding over the catalytic site is intramolecular or intermolecular within a dimeric assembly. Indeed, it is also possible that the gating sequence in these structures is adventitiously provided by adjacent protein molecules in the crystal lattice rather than from within discrete dimers. Nevertheless, accompanying mutagenesis studies carried out by Burgin *et al.*[4] strongly corroborate the location and mode of UCR2 docking on the catalytic unit as seen in their crystal structures. Their data are also consistent with earlier biochemical and two-hybrid studies showing not only that UCR2 interacts with the catalytic unit but that UCR1 interacts with UCR2 to form a regulatory module[5,6].

Past efforts to make inhibitors that are selective for each of the four PDE4 sub-families have been confounded by the structural similarity of their catalytic sites[8], although a degree of success has been achieved with PDE4D-selective

Miles D. Houslay is in the Department of Neuroscience and Molecular Pharmacology, University of Glasgow, Glasgow, G12 8QQ, Scotland, UK and David R. Adams is in the Department of Chemistry, Heriot-Watt University, Edinburgh EH14 4AS, Scotland, UK. e-mail: M.Houslay@bio.gla.ac.uk

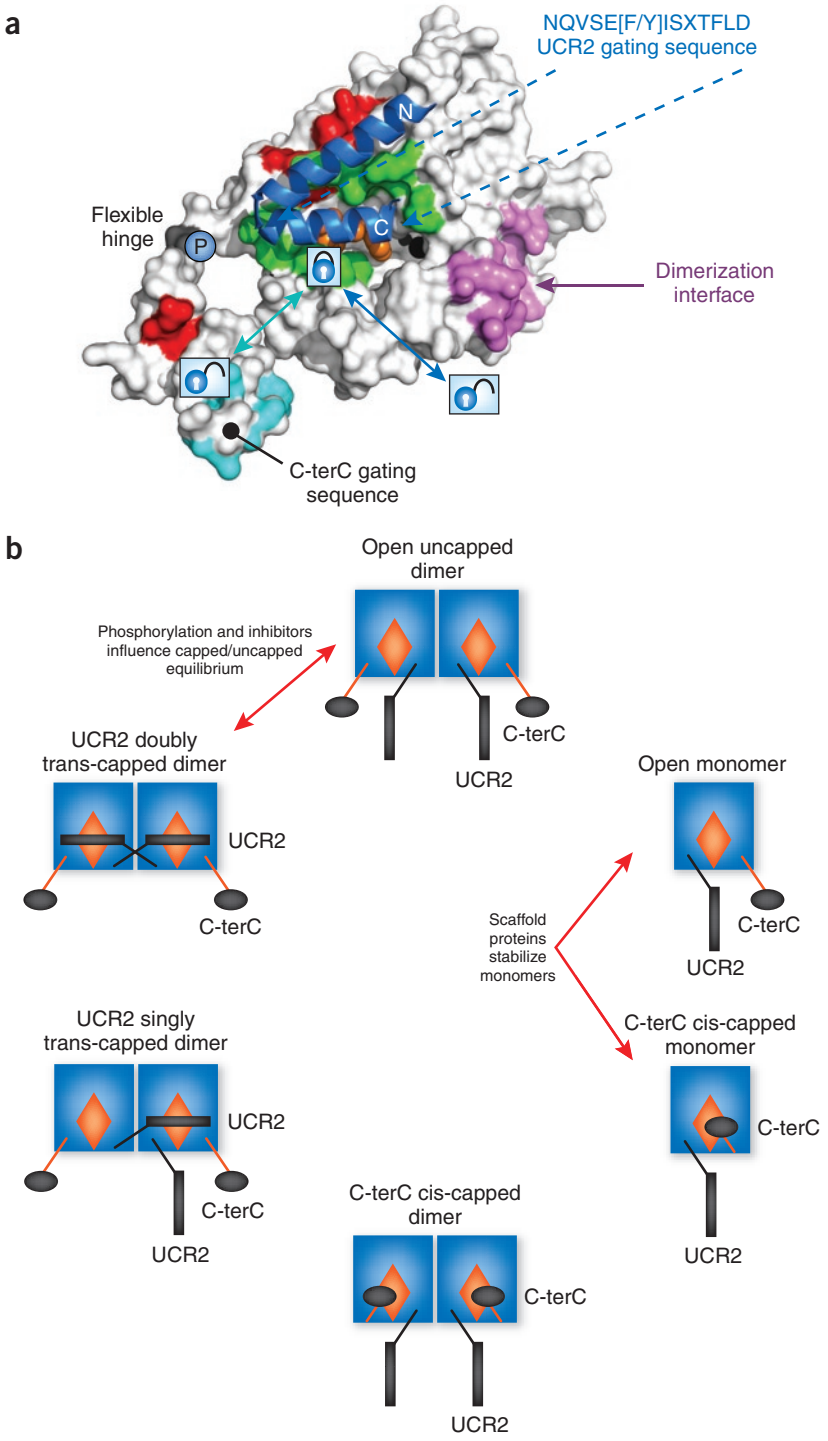


Figure 1 Structural basis of PDE4 regulation. (**a**) Model for gated access to PDE4 catalytic pocket. The PDE4 core catalytic domain forms a dimeric assembly (not shown) in which intermolecular docking of UCR2 (blue ribbon) to the rim of the catalytic pocket (green surface) gates access (blue arrow). Phosphorylation by ERK (dark grey site, P) stabilizes the UCR2-capped state. An additional gating mechanism may be provided by intramolecular capping of the active site by the C-terminal regulatory region (cyan arrow). Gating regulates entry and binding of inhibitors/substrate (orange) within the catalytic pocket and is fine-tuned by phosphorylation of and protein interactions with the extended N-terminal regulatory domain (including UCR1) as well as by the binding of scaffolding proteins to core catalytic unit loci, as exemplified here with binding sites shown for RACK1 (red) and β-arrestin (cyan), which themselves may alter gating by UCR2 and the C-terminal regulatory sequence. The model is based on new PDB entries 3G4G, 3G45 and 3G58 from Burgin *et al.*[4] augmented by entries 3KKT and 1XM6, in which the C-terminal sequence is also defined. The dimer interface is based upon various reported PDE4 catalytic unit structures, and the binding sites for β-arrestin and RACK1 are derived from peptide mapping coupled with mutagenesis studies (see, e.g., refs. 2,3,7 and 8). (**b**) Capping by either UCR2 or the C-terminal regulatory helix sequence generates multiple conformational states of PDE4. This schematic shows a series of key conformational states that can be deduced from the structural information in Burgin *et al.*[4]. Capping by the C-terminal regulatory is likely to be intramolecular given the short C-terminal flexible hinge sequence and the standard catalytic domain dimer configuration. The hinge sequence is disordered in the structures of Burgin *et al.*[4], but ordered in a recently released (unpublished) C-terminal-capped structure (PDB: 3KKT). For simplicity, capping by UCR2 is depicted as intermolecular because the linking region between UCR2 and the core catalytic unit appears too short for self-capping, although it should be noted that this region is highly variable among the four sub-families. These states will be subject to further elaboration when PDE4 is phosphorylated by various kinases and through association with diverse scaffold proteins. Capped forms are expected to show lower activity and increased affinity for inhibitors that interact with residues in the UCR2 or C-terminal regulatory gating sequences. Structural detail for the assembled UCR2-UCR1 regulatory module and linked PDE4 N-terminal sequences remains to be determined.

© 2010 Nature America, Inc. All rights reserved.

inhibitors[3]. The new structures suggest an explanation for these findings as they reveal a sequence difference within the UCR2 gating helix, with PDE4D containing a Phe at position 6 and PDE4A/B/C having Tyr at this position. This residue is shown to have a critical role in inhibition as its side chain is orientated into the catalytic pocket and can contact inhibitors when the enzyme adopts the UCR2-capped configuration. Indeed, a simple Tyr-Phe mutation at this one position converts the inhibitor profile of PDE4B into that of PDE4D, and vice versa with the reverse mutation in PDE4D.

The study by Burgin *et al.*[4] opens new opportunities for the rational design of inhibitors distinguished by their preference for interaction with residues within the UCR2 gating sequence. Among current inhibitors, RS25344 interacts favorably with the gating sequence, stabilizing the UCR2-capped state, whereas roflumilast interacts less favorably, preferentially occupying the uncapped catalytic pocket. These discoveries demystify the complex inhibition kinetics of compounds, such as rolipram, that have affinity for the catalytic pocket in both UCR2-uncapped and -capped states, resulting in the aforementioned sensitivity to PDE4 phosphorylation and protein-protein associations[3].

Burgin *et al.*[4] used this new understanding of UCR2 gating to design inhibitors with enhanced dependence on interactions with the gating UCR2 helix over interactions with catalytic pocket residues, exploiting the unique

Phe-containing gating sequence of PDE4D to develop inhibitors selective for PDE4D isoforms. This allowed them to generate compounds showing >10,000-fold preference for the UCR2-capped PDE4D dimeric state coupled with novel allosteric inhibition kinetics that imply negative cooperativity between the two catalytic units. The authors' model envisages this new type of inhibitor binding to one subunit in the UCR2-capped state and inducing a conformational change that leads to substantially lower inhibitor affinity together with reduced catalytic activity in the second, uncapped subunit (**Fig. 1b**). This yields partial competitive inhibition kinetics, as previously noted for PDE4A when 'converted' into a high-affinity rolipram-binding conformation through interaction with SRC family protein kinases, a phenomenon that was dependent upon interaction of their SH3 domain with LR2, the linking region between UCR2 and the PDE4 catalytic unit[9].

Previous studies have shown that binding of proteins and antisera to UCR2 can markedly decrease or increase PDE4 activity. The results of Burgin *et al.*[4] now put this into perspective as such protein-protein interactions can be expected to either enhance or disrupt the ability of UCR2 to dock onto the catalytic unit and thereby influence activity. Importantly for drug design, their study shows, for the first time, that active-site directed inhibitors can influence the conformational state of PDE4. Thus, inhibitors such as rolipram that have an intrinsic capacity to stabilize UCR2 docking will tip the balance in favor of the UCR2-capped state (**Fig. 1b**). In effect, protein association and inhibitor binding cooperate to promote UCR2 docking, altering sensitivity to inhibitors in the process. Remarkably, a regulatory sequence at the C-terminal end of the catalytic unit (**Fig. 1**), which provides one of the two binding sites involved in the association of β-arrestin and RACK1 scaffold proteins with the PDE4D5 isoform[10], was also found to be able to occupy the gate position in place of UCR2. This highlights a further means to regulate PDE4 by protein associations. The linkage between inhibitor binding and protein association suggests that inhibitors that perturb the equilibria between capped and uncapped states may exert functional effects beyond simply elevating cAMP levels through PDE4 inhibition. Indeed, they may influence cAMP signalling more subtly by affecting PDE4 targeting within cells, as has been shown for the action of rolipram on PDE4A4 distribution[11].

PDE4 phosphorylation by ERK confers cross-talk between two pivotal signalling pathways[1,2]. The new structures provide striking insights into the structural basis for this control and, in particular, for its dependence on UCR2 (ref. 2). This can now be explained as arising from stabilization of the UCR2-capped configuration by an interaction between a conserved Arg close to the UCR2 gating sequence and the Ser that ERK phosphorylates in the catalytic unit.

Burgin *et al.*[4] have revealed an exquisite dual-gating regulatory mechanism for PDE4 in which access to the catalytic pocket can be controlled by either the UCR2 helix or by a C-terminal helix that forms part of the core catalytic unit. This model provides key insights into how members of this critically important multi-enzyme family are regulated and provides a firm structural basis for future functional studies. Indeed, we can deduce from this model that the functioning of PDE4, the outcome of phosphorylation and its sensitivity to inhibition are likely to be fine-tuned by the binding of diverse protein partners to the N- and C-terminal regions.

Despite huge investment from the pharmaceutical industry over the last decade and extremely promising pre-clinical data, clinical deployment of current generations of PDE4 inhibitors has been severely compromised by nausea and emesis side-effects, which limit the effective therapeutic window[3]. On the basis of the proposal that the emesis side effect is related to brain PDE4D inhibition[12], the recent focus of endeavour has been to generate compounds that show reduced inhibition towards PDE4D compared with other PDE4 sub-families in the hope of reducing emesis. Burgin *et al.*[4] debunk this notion. Indeed, exploiting their structural insights, they actually set out to make highly PDE4D-selective inhibitors. Intriguingly, their lead compounds not only were brain-penetrant, and so potentially can access the emesis centre in the area postrema, but actually exhibited a greatly reduced emesis propensity coupled with excellent cognition-enhancing properties. With these results the authors chart a course for developing safer PDE4-selective inhibitors.

1. Conti, M. & Beavo, *J. Annu. Rev. Biochem.* **76**, 481–511 (2007).
2. Houslay, M.D., Baillie, G.S. & Maurice, D.H. *Circ. Res.* **100**, 950–966 (2007).
3. Houslay, M.D., Schafer, P. & Zhang, K.Y. *Drug Discov Today* **10**, 1503–1519 (2005).
4. Burgin, A. *et al. Nat. Biotechnol.* **27**, 63–70 (2010).
5. Beard, M.B. *et al. J. Biol. Chem.* **275**, 10349–10358 (2000).
6. Lim, J., Pahlke, G. & Conti, M. *J. Biol. Chem.* **274**, 19677–19685 (1999).
7. Xu, R.X. *et al. Science* **288**, 1822–1825 (2000).
8. Wang, H. *et al. Biochem. J.* **408**, 193–201, (2007).
9. McPhee, I. *et al. J. Biol. Chem.* **274**, 11796–11810 (1999).
10. Bolger, G.B. *et al. Biochem. J.* **398**, 23–36, (2006).
11. Terry, R. *et al., Cell. Signal.* **15**, 955–971 (2003)
12. Robichaud, A. *et al. J. Clin. Invest.* **110**, 1045–1052 (2002).

Enriching quantitative proteomics with SI_N

Mihaela E Sardiu & Michael P Washburn

A new metric called the normalized spectral index (SI_N) provides a simple way to quantify and compare label-free proteomics data.

Quantitative proteomics using mass spectrometry (MS) is increasingly finding application in areas ranging from systems biology to the identification of clinical biomarkers. But accurate quantification of large numbers of proteins in label-free shotgun experiments remains challenging. In this issue, Griffin *et al.*[1] cleverly incorporate three types of information commonly generated in MS experiments—unique peptide number, spectral count and fragment-ion intensity—to create a scoring function that facilitates quantitative analysis. Robust analysis of data using this approach should enable more reliable quantitative comparisons of label-free MS data both within and across laboratories.

The traditional approach for obtaining quantitative proteomics data involves comparing samples differentially labeled with light and heavy isotopes in a single MS run[2]. More recently, the relative abundances of proteins in mixtures have been determined without labeling[2]. Label-free methods permit comparison of multiple data sets without the cost and inconvenience of isotopic labeling. In both isotope-based and label-free experiments, quantification is achieved by analyzing either of two sources of information. In one approach, protein abundance is determined from the shape of peaks of eluting peptides using the area under the curve, or the summed intensity of each peak corresponding to a peptide[2]. A second approach, called spectrum

Mihaela E. Sardiu and Michael P. Washburn are at Stowers Institute for Medical Research, Kansas City, Missouri, USA.
e-mail: mpw@stowers.org

© 2010 Nature America, Inc. All rights reserved.

Phe-containing gating sequence of PDE4D to develop inhibitors selective for PDE4D isoforms. This allowed them to generate compounds showing >10,000-fold preference for the UCR2-capped PDE4D dimeric state coupled with novel allosteric inhibition kinetics that imply negative cooperativity between the two catalytic units. The authors' model envisages this new type of inhibitor binding to one subunit in the UCR2-capped state and inducing a conformational change that leads to substantially lower inhibitor affinity together with reduced catalytic activity in the second, uncapped subunit (**Fig. 1b**). This yields partial competitive inhibition kinetics, as previously noted for PDE4A when 'converted' into a high-affinity rolipram-binding conformation through interaction with SRC family protein kinases, a phenomenon that was dependent upon interaction of their SH3 domain with LR2, the linking region between UCR2 and the PDE4 catalytic unit[9].

Previous studies have shown that binding of proteins and antisera to UCR2 can markedly decrease or increase PDE4 activity. The results of Burgin *et al.*[4] now put this into perspective as such protein-protein interactions can be expected to either enhance or disrupt the ability of UCR2 to dock onto the catalytic unit and thereby influence activity. Importantly for drug design, their study shows, for the first time, that active-site directed inhibitors can influence the conformational state of PDE4. Thus, inhibitors such as rolipram that have an intrinsic capacity to stabilize UCR2 docking will tip the balance in favor of the UCR2-capped state (**Fig. 1b**). In effect, protein association and inhibitor binding cooperate to promote UCR2 docking, altering sensitivity to inhibitors in the process. Remarkably, a regulatory sequence at the C-terminal end of the catalytic unit (**Fig. 1**), which provides one of the two binding sites involved in the association of β-arrestin and RACK1 scaffold proteins with the PDE4D5 isoform[10], was also found to be able to occupy the gate position in place of UCR2. This highlights a further means to regulate PDE4 by protein associations. The linkage between inhibitor binding and protein association suggests that inhibitors that perturb the equilibria between capped and uncapped states may exert functional effects beyond simply elevating cAMP levels through PDE4 inhibition. Indeed, they may influence cAMP signalling more subtly by affecting PDE4 targeting within cells, as has been shown for the action of rolipram on PDE4A4 distribution[11].

PDE4 phosphorylation by ERK confers cross-talk between two pivotal signalling pathways[1,2]. The new structures provide striking insights into the structural basis for this control and, in particular, for its dependence on UCR2 (ref. 2). This can now be explained as arising from stabilization of the UCR2-capped configuration by an interaction between a conserved Arg close to the UCR2 gating sequence and the Ser that ERK phosphorylates in the catalytic unit.

Burgin *et al.*[4] have revealed an exquisite dual-gating regulatory mechanism for PDE4 in which access to the catalytic pocket can be controlled by either the UCR2 helix or by a C-terminal helix that forms part of the core catalytic unit. This model provides key insights into how members of this critically important multi-enzyme family are regulated and provides a firm structural basis for future functional studies. Indeed, we can deduce from this model that the functioning of PDE4, the outcome of phosphorylation and its sensitivity to inhibition are likely to be fine-tuned by the binding of diverse protein partners to the N- and C-terminal regions.

Despite huge investment from the pharmaceutical industry over the last decade and extremely promising pre-clinical data, clinical deployment of current generations of PDE4 inhibitors has been severely compromised by nausea and emesis side-effects, which limit the effective therapeutic window[3]. On the basis of the proposal that the emesis side effect is related to brain PDE4D inhibition[12], the recent focus of endeavour has been to generate compounds that show reduced inhibition towards PDE4D compared with other PDE4 sub-families in the hope of reducing emesis. Burgin *et al.*[4] debunk this notion. Indeed, exploiting their structural insights, they actually set out to make highly PDE4D-selective inhibitors. Intriguingly, their lead compounds not only were brain-penetrant, and so potentially can access the emesis centre in the area postrema, but actually exhibited a greatly reduced emesis propensity coupled with excellent cognition-enhancing properties. With these results the authors chart a course for developing safer PDE4-selective inhibitors.

1. Conti, M. & Beavo, *J. Annu. Rev. Biochem.* **76**, 481–511 (2007).
2. Houslay, M.D., Baillie, G.S. & Maurice, D.H. *Circ. Res.* **100**, 950–966 (2007).
3. Houslay, M.D., Schafer, P. & Zhang, K.Y. *Drug Discov Today* **10**, 1503–1519 (2005).
4. Burgin, A. *et al. Nat. Biotechnol.* **27**, 63–70 (2010).
5. Beard, M.B. *et al. J. Biol. Chem.* **275**, 10349–10358 (2000).
6. Lim, J., Pahlke, G. & Conti, M. *J. Biol. Chem.* **274**, 19677–19685 (1999).
7. Xu, R.X. *et al. Science* **288**, 1822–1825 (2000).
8. Wang, H. *et al. Biochem. J.* **408**, 193–201, (2007).
9. McPhee, I. *et al. J. Biol. Chem.* **274**, 11796–11810 (1999).
10. Bolger, G.B. *et al. Biochem. J.* **398**, 23–36, (2006).
11. Terry, R. *et al., Cell. Signal.* **15**, 955–971 (2003)
12. Robichaud, A. *et al. J. Clin. Invest.* **110**, 1045–1052 (2002).

Enriching quantitative proteomics with SI_N

Mihaela E Sardiu & Michael P Washburn

A new metric called the normalized spectral index (SI_N) provides a simple way to quantify and compare label-free proteomics data.

Quantitative proteomics using mass spectrometry (MS) is increasingly finding application in areas ranging from systems biology to the identification of clinical biomarkers. But accurate quantification of large numbers of proteins in label-free shotgun experiments remains challenging. In this issue, Griffin *et al.*[1] cleverly incorporate three types of information commonly generated in MS experiments—unique peptide number, spectral count and fragment-ion intensity—to create a scoring function that facilitates quantitative analysis. Robust analysis of data using this approach should enable more reliable quantitative comparisons of label-free MS data both within and across laboratories.

The traditional approach for obtaining quantitative proteomics data involves comparing samples differentially labeled with light and heavy isotopes in a single MS run[2]. More recently, the relative abundances of proteins in mixtures have been determined without labeling[2]. Label-free methods permit comparison of multiple data sets without the cost and inconvenience of isotopic labeling. In both isotope-based and label-free experiments, quantification is achieved by analyzing either of two sources of information. In one approach, protein abundance is determined from the shape of peaks of eluting peptides using the area under the curve, or the summed intensity of each peak corresponding to a peptide[2]. A second approach, called spectrum

Mihaela E. Sardiu and Michael P. Washburn are at Stowers Institute for Medical Research, Kansas City, Missouri, USA.
e-mail: mpw@stowers.org

© 2010 Nature America, Inc. All rights reserved.

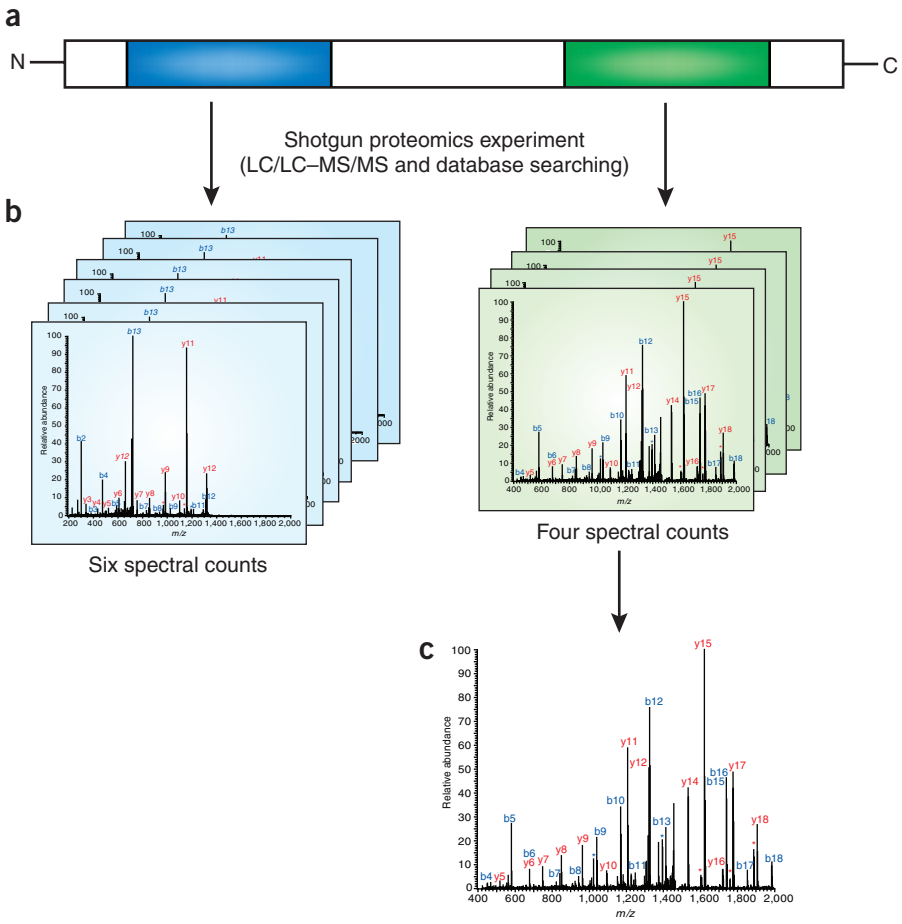


Figure 1 The normalized spectral index (SI_N) leverages unique peptide numbers, spectral counts and fragment ion-intensities to determine protein abundances. For a hypothetical protein with two unique peptides, the SI_N takes into account the fragment ion intensities for all spectra counted for each identified peptide derived from that protein. By contrast, conventional spectrum counting involves summing the number of tandem mass spectra detected for a given protein. (**a**) A protein of length *L* yields two unique peptides (blue and green) and these two peptides are detected and identified using tandem mass spectrometry and database searching out of thousands of peptides from a complex protein mixture. (**b**) The blue peptide yields six spectral counts and the green peptide yields four spectral counts. (**c**) After each of the spectra in **b** are interpreted with database searching and all ions (blue and red) are matched to peptides, the SI_N equation sums the intensities of the matched ion in any given spectrum and quantifies the protein of interest by normalizing this against the sum of ion intensities for all detected proteins and the length of the protein of interest.

counting, involves summing the total number of tandem mass spectra that are detected and identified for a given protein[2].

In the new method of Griffin *et al.*[1], the abundance of proteins in a mixture is determined by combining spectrum counting with another type of information: the fragment ion intensities from tandem mass spectra for each spectral count of a protein (**Fig. 1**). Spectrum counts and fragment ion intensities are integrated into an equation termed the normalized spectral index, or SI_N, that takes into account both protein length and the number of unique peptides per protein. For a given protein in a sample, the SI_N is defined as the sum of fragment ion intensities for all spectra counted for a protein (*SI*) normalized by the sum of *SI* over all proteins (*n*) and by the length of the protein (*L*):

$$SI_N = \left[\sum_{k=1}^{pn}\left(\sum_{j=1}^{sc} i_j\right)_k\right] \Big/ \left[\sum_{j=1}^{n} SI_j\right] \Big/ L$$

where *sc* is the spectrum count for a peptide *k*, *i* is the fragment ion intensity of peptide *k*, and *pn* is the total number of unique peptides of the protein. Although several methods rely solely on spectral counts, which represent the number of times the spectra are assigned to the same peptide, the SI_N approach first evaluates each individual spectrum and then sums the fragment ion intensity in an individual spectrum to form the spectral index. As the

© 2010 Nature America, Inc. All rights reserved.

spectra generated for any peptide differ in the same mass spectrometer run, the weighted approach of the spectral index improves quantitative analysis.

The authors tested the ability of the SI_N and methods that use spectrum counts or the area under the curve to quantify each of several proteins spiked into a solution of bovine serum albumin. The SI_N approach proved more robust for calculating protein abundance and also minimized the variation of technical replicates and the effects of different sample loads. Finally, using protein samples from lung tissue, the authors found a strong correlation between quantification estimated using the SI_N and that using western blot analysis.

The novelty of the SI_N approach is the use of fragment-ion intensities for every spectral count (**Fig. 1**). The intensity features of tandem mass spectra have been previously shown to be valuable for improved database matching[3] and quantitative proteomic analysis using a specialized MS setup[4]. In contrast, the information required to determine the SI_N is readily available and can be implemented with standard proteomics pipelines

It is worth emphasizing that all of the data used by Griffin *et al.*[1] were generated by ion-traps with lower resolution than state-of-the-art mass spectrometers. With these instruments it is difficult to integrate the intensity of the area under the curve for an eluting peptide because it is challenging to determine where a peak begins and ends. More expensive instrumentation systems with higher resolution and high mass accuracy should permit more accurate determination of the intensity of the area under the curve[5]. It remains to be seen how effectively the SI_N strategy can be adapted to mass spectrometers other than ion-trap systems.

The work by Griffin *et al.*[1] reinforces the importance of normalizing quantitative proteomics data sets[1,6,7]. Taking into account protein length and the total intensity of a data set improves the use of spectrum counts alone (as in the normalized spectral abundance factor[7]). Consideration of protein length is important as longer proteins typically yield more spectral counts than shorter proteins in proteomics analyses[7]. Normalizing to the total ion intensity accounts for the variation in signal intensity across the proteomics experiment.

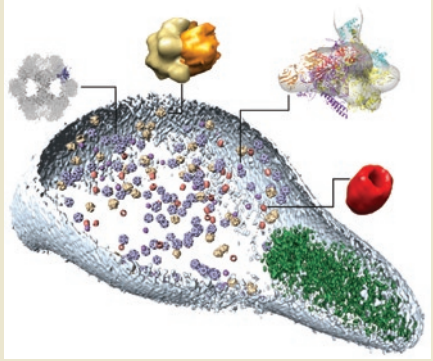
When these concepts are applied to other methods for quantitative proteomics, additional improvements are likely to be achieved. In particular, the impact of normalization approaches on labeled quantitative proteomics analysis should be investigated. Griffin *et al.*[1] provide some evidence that the SI_N can be used to estimate the abundance of proteins in a complex cellular lysate[1]. This is similar to prior work in which spectral counting was proposed to estimate the absolute protein expression of *Saccharomyces cerevisiae* and *Escherichia coli* proteins[6]. Further research linking label-free quantitative proteomic data sets and absolute protein expression levels is needed to explore the tantalizing possibility that the SI_N approach might find broad application in quantifying proteins in complex mixtures.

An area where the work of Griffin *et al.*[1] could have an immediate impact involves the quantitative analysis of protein interaction networks. Several large-scale projects currently underway aim to assemble thousands of human protein complexes into an interaction network[8,9]. For reasons of cost and simplified experimental pipelines, these studies have so far been conducted using label-free approaches[8,9]. The development of improved label-free quantitative proteomic analysis tools, like the SI_N equation, should advance both network assembly and retroactive analysis of existing protein interaction network data sets.

1. Griffin, N.M. *et al. Nat. Biotechnol.* **28**, 83–89 (2010).
2. Kline, K.G., Finney, G.L. & Wu, C.C. *Brief. Funct. Genomic. Proteomic.* **8**, 114–125 (2009).
3. Elias, J.E. *et al. Nat. Biotechnol.* **22**, 214–219 (2004).
4. Venable, J.D. *et al. Nat. Methods* **1**, 39–45 (2004).
5. Mann, M. & Kelleher, N.L. *Proc. Natl. Acad. Sci. USA* **105**, 18132–18138 (2008).
6. Lu, P. *et al. Nat. Biotechnol.* **25**, 117–124 (2007).
7. Zybailov, B. *et al. J. Proteome Res.* **5**, 2339–2347 (2006).
8. Sardiu, M.E. *et al. Proc. Natl. Acad. Sci. USA.* **105**, 1454–1459 (2008).
9. Sowa, M.E., Bennett, E.J., Gygi, S.P. & Harper, J.W. *Cell* **138**, 389–403 (2009).

Small but not simple

Even for the simplest organisms that can be grown in laboratory media, such as bacteria of the mycoplasma family, we are far from understanding all of the design principles and essential functions needed to sustain life. For instance, more than a quarter of the 370 essential protein-coding genes of *Mycobacterium genitalium* have no known function[1]. Three recent papers in *Science*[2–4], by a consortium of research groups led by Peer Bork, Luis Serrano and Anne-Claude Gavin, illustrate the complexity of *Mycobacterium pneumoniae* through comprehensive analyses of its transcriptome[2], proteome[3] and metabolome[4]. *M. pneumoniae* has one of the smallest known genomes of a self-replicating bacterium, comprising 816 kb and encoding just 689 proteins, only 8 of which are predicted to be transcription factors. The observation that 89 of the 117 new transcripts identified are antisense to annotated genes reveals a hitherto unappreciated level of gene regulation[2]. Moreover, many genes were found to produce more than one transcript[2]. Proteome analysis showed that at least 90% of the proteins studied are part of at least one of the 178 protein complexes identified[3]. More than half of these complexes had not been described previously. Reconstruction of the metabolic network revealed that many redundancies and branched pathways common to other organisms are not present in *M. pneumoniae*[4]. However, despite its low number of metabolic enzymes and transcriptional regulators, the bacterium is able to perform a large variety of metabolic reactions and to adapt quickly to changes in the environment. The former can be explained by the large fraction of multifunctional enzymes, whereas the latter suggests a level of regulation different from that of bacteria with larger genomes. Taken together, the papers highlight the complexity of even the simplest bacteria and underscore the challenges in understanding and reconstructing minimal life forms. More information can be found in commentaries by Venter and colleagues[1] and Ochman and Raghavan[5].



Markus Elsner

1. Glass, J.I., Hutchison, III, C.A., Smith, H.O. & Venter, J.C. *Mol. Sys. Bio.* **5**, 330 (2009).
2. Güell, M. *et al. Science* **326**, 1268–1271 (2009).
3. Kühner, S. *et al. Science* **326**, 1235–1240 (2009).
4. Yus, E. *et al. Science* **326**, 1263–1268 (2009).
5. Ochman, H. & Raghaven, R. *Science* **326**, 1200–1201 (2009).

© 2010 Nature America, Inc. All rights reserved.

Genome sequencing on nanoballs

Gregory J Porreca

Advances in technology deliver cheaper human genome sequencing.

© 2010 Nature America, Inc. All rights reserved.

The rapid pace of innovation in the field of genome sequencing continues with a recent publication in *Science* by Drmanac *et al.*[1]. The authors resequenced three full human genomes using a next-generation technology that combines highly efficient imaging on ordered arrays with an inexpensive ligation-based chemistry. These technological improvements further reduce the cost of human genome sequencing.

Next-generation sequencing technologies generate up to billions of short reads in a run. All of these approaches use either polymerase or ligase to identify each base with a fluorescent signal that is read by a microscope and a digital camera. Development of these systems has focused on manipulating and arraying DNA such that it can be seen by the camera and sequenced, and on devising a sequencing chemistry with sufficient accuracy and read-length. For effective human genome sequencing, the individual DNA spots should be small (≤1 μm) and present at high density (approaching 1 million spots per mm^2). Furthermore, the reads must be long enough (>30 bp) to allow unambiguous alignment to the reference sequence, which is the first step in identifying variants.

Drmanac *et al.*[1] have met these goals with a platform that integrates several technologies: (i) a library-generation protocol that transforms fragments of genomic DNA into highly engineered molecules; (ii) a method for generating spots, called DNA nanoballs, and arraying them in highly dense grids for efficient imaging; and (iii) a nonprogressive chemistry (that is, errors do not accumulate because each base is read from a fresh sequencing primer) that uses ligation with partially degenerate sequencing primers[1–3] to yield accurate ~70-bp reads split across eight priming sites (**Fig. 1**).

What is most intriguing is how the platform approaches several technical optima that in concert drive down cost. First, the amount of reagent used is dictated by the area and height of the instrument's flow-cell chamber. Drmanac *et al.*[1] recognized that the chamber's height does not affect sequencing performance because the submicron-sized DNA features are attached to its surface. So they devised a process to manufacture thin chambers and perfected a way to flow liquid through them, potentially enabling significant cost savings over current systems with thicker flow-cells.

Second, throughput is driven both by the speed of the camera and by how many spots can be packed into a single image. The authors used the electron-multiplied charge-coupled device (CCD) present in several other sequencing systems[2] (http://www.polonator.org/), which is faster and more sensitive than what is found in the most popular next-generation platforms on the market. They combined this camera with one of their key innovations, a patterned array of DNA nanoballs. These compact chains of amplified DNA assemble into a densely packed grid of spots on the flow-cell surface, maximizing the yield of useful sequenced bases from camera pixels. Nanoballs offer a higher array packing density than bridge amplification[4], because they physically exclude other DNA molecules from their spot on the grid, and a much easier and cheaper workflow than emulsion PCR[5], because they are prepared in a simple reaction that does not waste most of the amplification reagents on empty emulsion bubbles. The result is an instrument capable of maximal throughput, given today's camera technology, and therefore minimal capital cost per base pair.

It is difficult to directly compare sequencing cost between different platforms. This is because, for genomic resequencing, cost is driven by the coverage required to achieve the desired accuracy. Different platforms may require different levels of coverage to achieve the same accuracy, so comparisons have to be made by fixing either coverage, to measure differences in cost and accuracy, or accuracy, to measure differences in coverage and cost[6]. There are other considerations as well. Different mutations (e.g., homozygous versus heterozygous substitutions, insertions or deletions) are generally sequenced with different accuracy. Moreover, the reference standard used to verify mutations must be more accurate than the sequence in question, and this is difficult to achieve with the genome-wide single-nucleotide polymorphism chips that are often used. All of these factors combine to confound a simple bases-per-dollar comparison.

So what can be said about the relative cost of this approach? Drmanac *et al.*[1] sequenced three genomes at a coverage of 45–87× and at an average reagent cost of $4,400 per genome. By their estimates, one false-positive sequencing error occurred every 100,000 bases—an accuracy on par with, or better than, that of other popular sequencing platforms. Complete Genomics, the company associated with the study by Drmanac *et al.*[1], has positioned itself as a service provider of full human genome sequences rather than as a vendor of sequencing instruments and reagents. Of course, the cost of reagents does not include equipment, labor and data-handling, and is not the same as the price charged to customers for a genome sequence. One appropriate comparison is therefore with Illumina's Personal Genome Sequencing Service, which delivers a full human genome sequence (on an iMac computer) for $48,000. Complete Genomics currently charges $20,000 for a sequence of similar accuracy[1,4,7].

With time, it is certain that prices across all vendors will fall. Complete Genomics has a target price of $5,000 per genome for 'bulk' orders[7], a substantial drop that is certainly possible in the near term. Reducing prices significantly beyond that will likely require further innovation. For instance, substantial increases in instrument throughput could be achieved by switching from CCD cameras to the much faster and cheaper complementary metal oxide semiconductor (CMOS) technology. But CMOS is less sensitive, so the DNA nanoballs would have to be made brighter, which may require considerable research and development.

As Complete Genomics makes progress in process automation and robustness, they may be able to address applications beyond human genome sequencing, including gene expression analysis, chromatin immunoprecipitation and metagenomics. For these, part of the difficulty will be in the process scaling and multiplexing required to accommodate the ultra-high throughput of their machines. In addition, for quantitative applications, it will be important to ensure they can calibrate for biases introduced by the library- and nanoball-generation protocols.

Gregory J. Porreca is at Good Start Genetics, Boston, MA.
e-mail: gporreca@gsgenetics.com

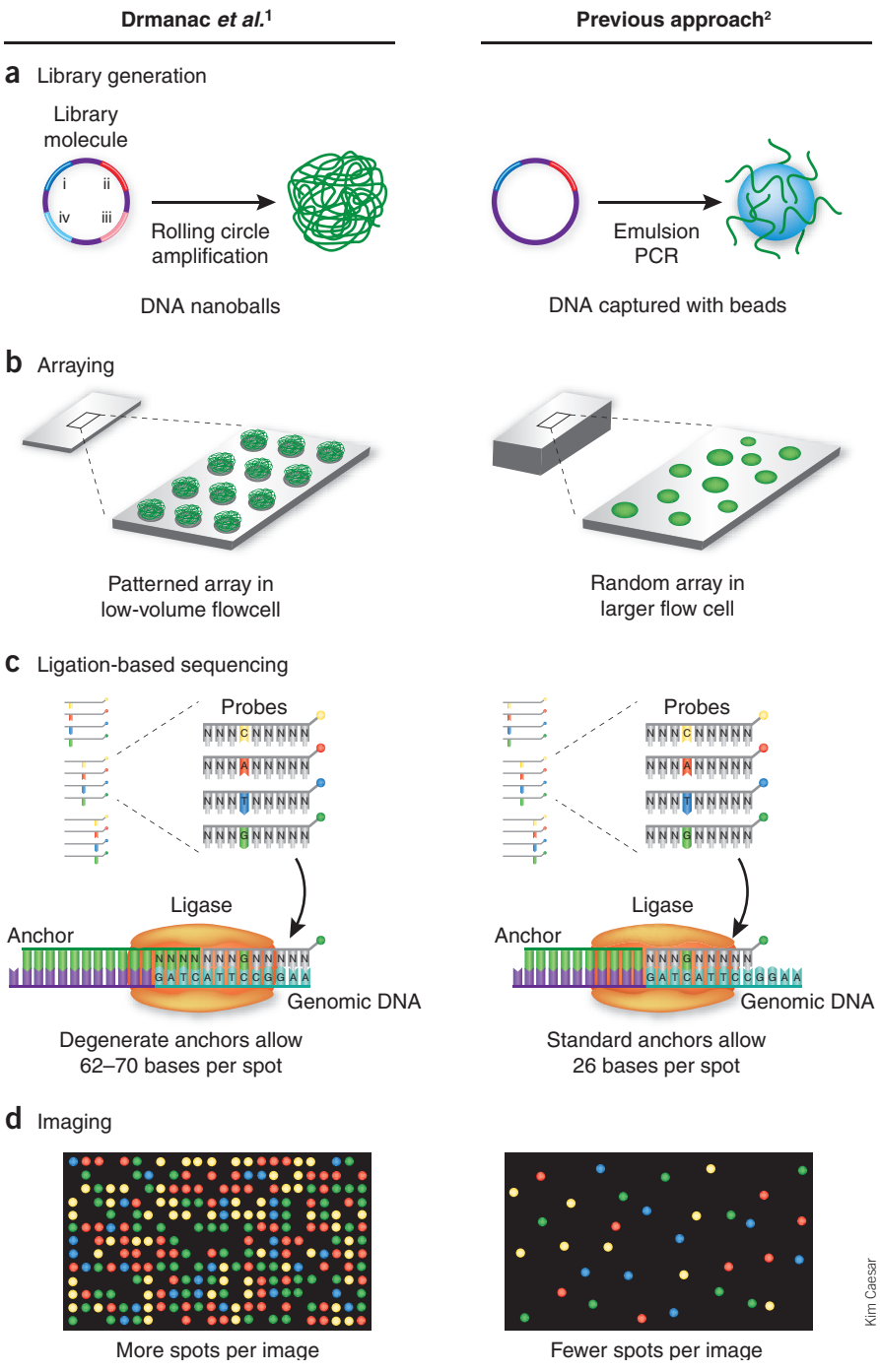


Figure 1 Comparison of the sequencing process in Drmanac *et al.*[1] (left) and in a previous sequencing-by-ligation method[2] (right). (**a**) Genomic DNA is converted into library molecules. Each molecule contains four segments of genomic DNA (i–iv), flanked by priming sites (shown in purple). Each library molecule is converted into a linear concatemer of itself to become a DNA nanoball. (**b**) Billions of DNA nanoballs are added to a silicon slide that contains a grid-like pattern of binding sites, which causes the nanoballs to self-assemble into a dense grid of spots for sequencing, maximizing the number of useful sequenced bases in each image (see **d**). (**c**) Ligation-based sequencing chemistry is used to interrogate bases of genomic DNA in the library molecules. Each cycle of sequencing tags the DNA nanoballs with a fluorophore whose color identifies the base (A, C, T or G) present at a specific position. The chemistry allows 5–10 contiguous bases to be read from each of the eight priming sites in the library molecule. (**d**) Digital images of the patterned arrays are taken after each sequencing reaction. The images are computationally analyzed to generate billions of raw sequence reads. These reads are then processed with assembly and analysis software to accurately identify mutations.

As costs continue to drop, the use of sequencing in diagnostics is expected to increase dramatically. This large market imposes significant requirements on any technology it adopts. Costs must be very low to displace existing technologies, and accuracy must be extremely high. False-positive mutation calls drive up assay cost by requiring expensive and time-intensive verification, and false-negative calls (which generally cannot be verified) are a source of diagnostic error. Thus, accuracy must be high, quantifiable and thoroughly measured in advance of releasing an assay into the clinic. What's more, continual monitoring of assay performance and compliance with clinical laboratory best-practices are imperative if a sequence is to be considered actionable medical advice.

In the span of a few short years, the mature technology of capillary sequencing has been supplanted by new sequencing approaches that offer tremendous increases in how much we can afford to sequence and how quickly we can do it. As the technology advances, focus will shift from the initial feats of sequencing single genomes to the ongoing challenge of producing lots of sequence accurately and efficiently. In this endeavour, the platform of Drmanac *et al.*[1] is sure to remain in the mix.

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

1. Drmanac, R. *et al. Science,* published online doi:10.0026/science.1181498 (5 November 2009). 2009 Nov 5 [Epub ahead of print]
2. Shendure, J.A. *et al. Science* **309**, 1728–1732 (2005)
3. Church, G.M. *et al.* US appl. no. 2007/0207482 (2007).
4. Bentley, D.R. *et al. Nature* **456**, 53–59 (2008).
5. Mckernan, K.J. *et al. Genome Res.* **19**, 1527–1541 (2009).
6. Fuller, C.W. *et al. Nat. Biotechnol.* **27**, 1013–1023 (2009).
7. Karow, J. Complete genomics details low-cost sequencing tech in paper; collaborators encouraged by results. *InSequence* <http://www.genomeweb.com/sequencing/complete-genomics-details-low-cost-sequencing-tech-paper-collaborators-encourage> (10 November 2009).

© 2010 Nature America, Inc. All rights reserved.

Short-read genome assembler

A novel algorithm lies behind recent insights into the genomes of the giant panda (*Nature*, published online 13 December 2009, doi:10.1038/nature08696), the cucumber (*Nat. Genet.*, **41**, 1275–1281, 2009) and new versions of two human genomes (p. 57–62). Wang and colleagues developed an approach, called SOAPdenovo, to assemble short sequencing reads (~35–75 bp) of large genomes into multikilobase sized chunks without needing a reference genome, a process termed *de novo* assembly. The approach combines efficient data structures for representing short reads with methods for correcting sequencing errors before genome assembly. Together, these advances enabled two human genomes to be assembled so that half of all bases are contained in stretches of sequence at least 5.9–7.4 kb long compared to 1.5 kb for the previous best *de novo* assembly methods. Although this size was smaller than the 20–100 kb achieved using Sanger sequencing (which generates longer reads), it was long enough to discover fragments containing novel coding regions of the human genome. And with the help of long-insert paired-end libraries, the sequences could be arranged into linear genome scaffolds hundreds of kilobases long, which is closer in size to assemblies derived from Sanger sequencing. As these studies suggest, *de novo* assembly of large mammalian and plant genomes from next-generation sequencing technology is now feasible. (*Genome Res.*, published online December 17, 2009, doi:10.1101/gr.097261.109) *CM*

Discrete logic models signaling

With the advent of ever more detailed maps of cellular signaling pathways, it has become apparent that sophisticated analysis methods are needed to understand the behavior of the whole system. Approaches based on differential equations require measuring or estimating a multitude of parameters. A simpler method, discrete logic modeling, which represents the signaling networks as a series of interconnected 'on' and 'off' switches, only requires knowledge of protein-protein interactions and whether they activate or inhibit each other. Although this modeling strategy has been successfully applied in some cases, so far no general approach to optimizing the model using experimental data had been developed. Now, Sorger and colleagues present an algorithm that can modify an input signaling network to optimize the fit to experimental results. By introducing a tunable parameter that balances goodness of fit with model complexity, the authors avoid overfitting and optimize the predictive power of the model. This is validated by constructing a model of HepG2 cell signaling and optimizing it with experimental data of phosphorylation cascades after different stimulations. Several predictions of the model have already been validated in the literature. (*Mol. Syst. Biol.* **5**, 331, 2009) *ME*

Taking down hepatitis C

Chronic infection with hepatitis C virus (HCV) remains a public health problem, as current therapies work in only half of the 170 million people infected worldwide, many of whom will develop serious complications. Lanford and colleagues show that in chronically infected primates, targeting an endogenous, highly expressed liver microRNA results in an enduring reduction in viral load. The target, microRNA (miR)-122, upregulates HCV replication by binding to the viral 5′-untranslated end. Four chimpanzees were given 12 weekly intravenous injections of a locked nucleic acid–modified oligonucleotide directed against that region. In animals given the highest dose, circulating virus as well as liver HCV RNA was reduced by almost three orders of magnitude. Two lines of experiments showed that resistance to treatment did not develop: there was no viral rebound during treatment and deep sequencing did not reveal any mutations in samples taken throughout and after treatment. As viral loads took several months to rebound to pretreatment levels after therapy was stopped, the approach shows therapeutic promise. (*Science*, published online December 3, 2009; doi:10.1126/science1178178) *LD*

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

Receptor-selective aglycosylated Abs

The ability of antibodies to bind all six members of the Fc gamma receptor (FcγR) family is essential for many aspects of adaptive immunity, including the antibody-dependent cell-mediated cytotoxicity (ADCC) response that underlies the effects of IgG-based therapeutics, such as the anticancer drug trastuzumab (Herceptin). Although the critical dependence of FcγR engagement on IgG glycosylation has restricted the manufacture of therapeutic antibodies to mammalian expression systems, Sazinsky *et al.* (*Proc. Natl. Acad. Sci. USA* **105**, 20167–20171, 2008) demonstrated the ability to uncouple FcγR binding from antibody glycosylation by mutation of the Fc region. Jung *et al.* now further demonstrate the utility of microbial systems for antibody engineering by identifying mutants of aglycosylated trastuzumab that bind the high-affinity FcγR1 receptor with affinity similar to that of their glycosylated counterparts. Importantly, however, the mutant antibodies bind none of the five other FcγRs, including the inhibitory FcγRIIb receptor that is well documented to prevent dendritic cell activation. Accordingly, mutant aglycosylated trastuzumab—but neither clinical-grade trastuzumab nor glycosylated mutant trastuzumab—potentiate the killing of HER2-overexpressing cancer cells *in vitro* by potent activation of dendritic cells. Although these effects have yet to be tested *in vivo*, they suggest that the efficacy of therapeutic antibodies might be enhanced by engineering aglycosylated variants to mediate more potent ADCC. (*Proc. Natl. Acad. Sci. USA*, published online December 18, 2009, doi:10.1073/pnas.0908590107) *PH*

Haploid genetic screens for human cells

The haploid genetic screens routinely used by yeast researchers to recognize recessive mutations have long been the envy of those working with multicellular eukaryotes. RNA interference–based strategies never silence gene expression completely and are often fraught with undesired off-target effects. Moreover, knockout strategies with diploid cells are complicated by the fact that most mammalian genes require disruption of both alleles to confer a phenotype different from the wild-type condition. To address this problem, Carette *et al.* use gene-trap retroviruses for large-scale gene disruption and tagging in a previously described derivative of the KBM7 chronic myeloid leukemia cell line that is haploid for all chromosomes except chromosome 8. Their efforts to screen gene trap–mutagenized KBM7 cells for resistance to influenza virus, cytolethal distending toxin and ADP-ribosylating toxins identify previously uncharacterized genes required for the actions of several intensively studied pathogens. This strategy should be amenable to screens that assay modulation of any reporter gene of interest. (*Science* **326**, 1231–1235, 2009) *PH*

© 2010 Nature America, Inc. All rights reserved.

RESOURCE

Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library

Hugo Y K Lam[1,13], Xinmeng Jasmine Mu[1,2,13], Adrian M Stütz[3], Andrea Tanzer[4], Philip D Cayting[5], Michael Snyder[2,12], Philip M Kim[6–9], Jan O Korbel[3,10,13] & Mark B Gerstein[1,5,11]

Structural variants (SVs) are a major source of human genomic variation; however, characterizing them at nucleotide resolution remains challenging. Here we assemble a library of breakpoints at nucleotide resolution from collating and standardizing ~2,000 published SVs. For each breakpoint, we infer its ancestral state (through comparison to primate genomes) and its mechanism of formation (e.g., nonallelic homologous recombination, NAHR). We characterize breakpoint sequences with respect to genomic landmarks, chromosomal location, sequence motifs and physical properties, finding that the occurrence of insertions and deletions is more balanced than previously reported and that NAHR-formed breakpoints are associated with relatively rigid, stable DNA helices. Finally, we demonstrate an approach, BreakSeq, for scanning the reads from short-read sequenced genomes against our breakpoint library to accurately identify previously overlooked SVs, which we then validate by PCR. As new data become available, we expect our BreakSeq approach will become more sensitive and facilitate rapid SV genotyping of personal genomes.

© 2010 Nature America, Inc. All rights reserved.

Structural variation of large segments (>1 kb), including copy-number variation and unbalanced inversion events, is widespread in human genomes[1–6] with ~20,000 SVs presently reported in the Database of Genomic Variants (DGV)[2]. These SVs have considerabe impact on genomic variation by causing more nucleotide differences between individuals than single-nucleotide polymorphisms[4–6] (SNPs). In several genomic loci, rates of SV formation could even be orders of magnitude higher than rates of single nucleotide substitution[7,8]. To measure the influence on human phenotypes of common SVs (that is, those present at substantial allele frequencies in populations) and *de novo* formed SVs, several studies have mapped SVs across individuals. They reported associations of SVs with normal traits and with a range of diseases, including cancer, HIV, developmental disorders and autoimmune diseases[9–14]. Although most SVs listed in DGV are presumably common, *de novo* SV formation is believed to occur constantly in the germline and several mutational mechanisms have been proposed[15].

Nevertheless, so far our understanding of SVs and the way we analyze SV maps is limited by the limited resolution of most recent surveys, such as those solely based on microarrays, which have not revealed the precise start and end coordinates (that is, breakpoints) of the SVs. This has hampered our understanding of the extent and effects of SVs in humans, as mapping at breakpoint resolution can reveal SVs that intersect with exons of genes or that lead to gene fusion events[5,16].

The lack of nucleotide-resolution maps has further prevented systematic deduction of the processes involved in SV formation, such as whether common SVs emerged initially as insertions or deletions at ancestral genomic loci. Instead, operational definitions have been applied for classifying common SVs into gains, losses, insertions and deletions based on either allele frequency measurements, or the 'human reference genome' (hereafter also referred as the reference genome) that was originally derived from a mixed pool of individuals[17]. Thus, inference of the ancestral state of an SV locus is crucial for relating SV surveys to primate genome evolution and population genetics.

The lack of data at nucleotide resolution has also limited the number of SVs for which the likely mutational mechanisms of origin have been inferred. These mechanisms are thought to include (i) NAHR involving homology-mediated recombination between paralogous sequence blocks; (ii) nonhomologous recombination (NHR) associated with the repair of DNA double-strand breaks (that is, nonhomologous end-joining) or with the rescue of DNA replication-fork stalling events (that is, fork stalling and template switching[18]); (iii) variable number of tandem repeats (VNTRs) resulting from expansion or contraction of simple tandem repeat units; and (iv) transposable element insertions (TEIs) involving mostly long and short interspersed elements (LINEs and SINEs) and combinations thereof, along with other types of TEI-associated events (e.g., processed pseudogenes).

Finally, owing to the lack of resolution of most SV maps, junction sequences (the flanking sequences of breakpoints) have thus far not

[1]Program in Computational Biology and Bioinformatics, [2]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut, USA. [3]Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [4]Institute for Theoretical Chemistry, University of Vienna, Vienna, Austria. [5]Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, USA. [6]Terrence Donnelly Centre for Cellular and Biomolecular Research, [7]Banting and Best Department of Medical Research, [8]Department of Molecular Genetics, [9]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [10]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [11]Department of Computer Science, Yale University, New Haven, Connecticut, USA. [12]Present address: Department of Genetics, Stanford University School of Medicine, Stanford, California, USA. [13]These authors contributed equally to this work. Correspondence should be addressed to J.O.K. (jan.korbel@embl.de) or M.B.G. (mark.gerstein@yale.edu).

Received 31 July; accepted 8 December; published online 27 December 2009; doi:10.1038/nbt.1600

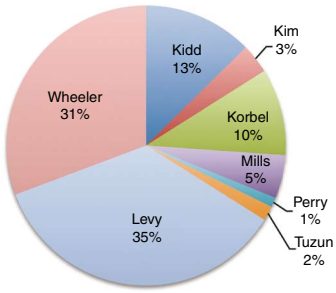


Figure 1 Composition of the SV breakpoint library. SVs in the library were based on different SV-mapping and breakpoint-sequencing strategies. A large fraction (44%) of the breakpoints were based on data generated using 454/Roche sequencing, including resequencing of an individual human genome (Wheeler[21], 602 SVs) and sequencing of breakpoints in two individuals after high-resolution and massive paired-end mapping (Korbel[5] and Kim[16], 264 SVs). The remaining 56% of the breakpoints were identified using other approaches, including Sanger capillary sequencing of breakpoints identified by whole-genome shotgun sequencing and assembly of an individual human genome (Levy[44], 694 SVs), fosmid-paired-end sequencing carried out in multiple individuals (Tuzun[3] and Kidd[6], 281 SVs), breakpoints mined from SNP discovery DNA resequencing traces (Mills[45], 98 SVs), and tiling-array-based comparative genomic hybridization followed by breakpoint sequencing (Perry[25], 22 SVs). Fewer than five breakpoints were reported in two genomes sequenced using short 36-bp reads (Illumina/Solexa)[22,23], presumably owing to the complex DNA sequence patterns frequently associated with breakpoints[5,6,25].

been exploited for testing the presence of SVs in an individual in a similar fashion to the way SNPs can be directly detected by oligonucleotide chips with probes designed for each polymorphism.

Recent advances in microarray technology and large-scale DNA sequencing have paved the way for high-resolution SV maps. To date, nearly 2,000 SVs have been fine-mapped at nucleotide level and efforts such as the 1000 Genomes Project (http://1000genomes.org/), which will soon sequence >1,000 human genomes, might in the near future report many more SVs at such resolution (**Supplementary Fig. 1**). Thus far, however, no study has leveraged the potential of collectively analyzing breakpoint-level SV data.

Here we present a comprehensive analysis of a library of nearly 2,000 SVs assembled from eight recent surveys that involve individuals from three distinct populations. We demonstrate four uses of the breakpoint library—mapping structural variation at high resolution, revealing ancestral states of variants, inferring mechanisms of variant formation and correlating the inferred mechanisms with DNA sequence features. We found several lines of evidence consistent with a nonuniform distribution of SV formation mechanisms and with locus-specific sequence properties, such as DNA helix stability, chromatin accessibility and the propensity for a DNA sequence to recombine, which may predispose genomic regions to SV-mutational processes.

RESULTS

Generation of a standardized SV breakpoint library

We compiled a set of breakpoints from eight published sources (**Fig. 1**). In accordance with a previously proposed operational definition[19], we defined SVs to be deletions, insertions and inversions reported relative to the reference genome with a size of 1 kb or larger. As our initial library encompassed SVs mapped using different types of evidence, sequencing technologies and genome assembly versions, an essential first step was library standardization. We therefore implemented a computational pipeline for generating a unified, nonredundant breakpoint library (Online Methods).

The pipeline yielded a nonredundant set of 1,889 SVs that were initially annotated as deletions (1,409), insertions (419) or inversions (61) relative to the reference genome (**Supplementary Fig. 2**). This set, which represents the most exhaustive compilation to date of SV breakpoints in phenotypically normal individuals, is available as **Supplementary Table 1** and at http://sv.gersteinlab.org/breakseq. It also has been deposited into the BreakDB database[20] (http://sv.gersteinlab.org/breakdb).

High-resolution mapping of SVs from short-read sequencing data

Personal genomics endeavors based on next-generation sequencing technology[21–23] typically detect genomic variation by mapping relatively short sequencing reads directly onto the reference genome. Although many short indels (<1 kb) can be accurately identified with such an approach, SVs >1 kb are commonly missed, or not identified at nucleotide (that is, breakpoint-level) resolution. This is probably because of the difficulty in constructing accurate sequence alignments from short reads (e.g., 36 mers), especially if they involve long sequence gaps or span breakpoints.

We thus devised an approach, BreakSeq, for detecting SVs by aligning raw reads directly onto SV breakpoint junctions of the alternative, nonreference, alleles contained in our library (**Fig. 2a**, Online Methods). Briefly, the genomic coordinates of each breakpoint in the standardized library are used to extract 30 bp of flanking sequence from the reference genome. These 30-bp flanking sequences are concatenated into 60-bp junction sequences. Thus, a deletion event is represented with a single junction sequence in the library (containing the sequence flanking its single breakpoint), whereas an insertion has both left and right junction sequences (containing the sequence flanking each of its two breakpoints). DNA reads from personal genomes are aligned against the junction sequences. Successful alignment requires a read to overlap a junction sequence by at least 10 bp on each side of the breakpoint. This approach is conceptually similar to using a library of exon splice junctions in transcriptome analyses, which leads to considerably better coverage of alternatively spliced transcripts than restricting the analysis to reference genome sequences lacking splice junctions[24].

To demonstrate the utility of our approach for mapping personal SVs at high resolution, we mapped short reads from three personal genomes sequenced with Illumina/Solexa technology. These included two previously published genomes[22,23] from individuals of Nigerian (Yoruba from Ibadan, YRI) and Han Chinese (HCH) origins. The third genome was from a HapMap individual of European ancestry (CEPH) that was sequenced recently in the pilot phase of the 1000 Genomes Project. To prioritize the SV calls generated by BreakSeq, we developed a scoring system based on supportive read-matches (the number of reads that map to a breakpoint; Online Methods) and distinguished low-support SV calls (with 1 to 4 supportive read-matches) from high-support SV calls. For the HCH, CEPH (NA12891) and YRI (NA18507) genomes, we identified 158, 219 and 179 SVs, respectively (**Supplementary Table 2**). Several SVs were shared among the three, suggesting that they may represent common alleles. For example, among the high-support calls, we found that 57 SVs were shared between the YRI and HCH genomes, 62 between the YRI and NA12891 genomes, 52 between the HCH and NA12891 genomes, and 42 were common to all three genomes.

To validate these results, we used PCR to test 24 insertion and 33 deletion calls predicted in NA12891 relative to the

© 2010 Nature America, Inc. All rights reserved.

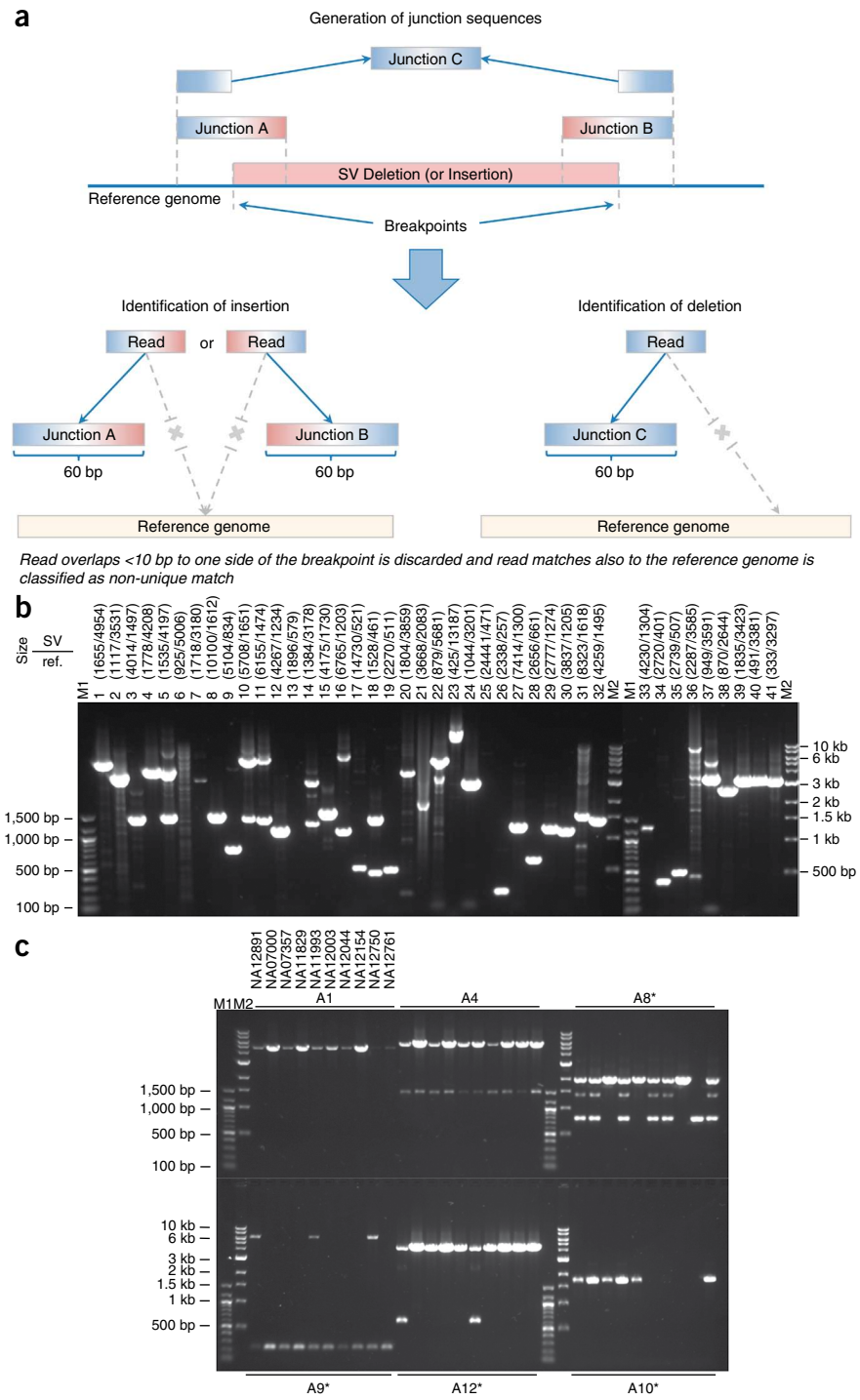


Figure 2 Mapping breakpoints using the library. (**a**) Overview of the BreakSeq approach. Breakpoints are used to generate junction sequences spanning breakpoints (upper)—the 30 bp of sequence flanking each side of the breakpoint (60 bp total). Then, DNA reads are aligned to the junction sequences (lower). Alignment results are interpreted as follows. In the case of insertions relative to the reference genome (left), sequences A and B represent the left and right breakpoint junction sequences of the nonreference SV allele, respectively. In the case of deletions (right), sequence C represents the junction sequence of the nonreference SV allele. Solid lines with arrows, successful alignments. Dashed lines with crosses, no proper alignment. (**b**) Representative PCR validation of detected SVs in NA12891. Primers flanking each SV were used to amplify 41 different genomic regions (see **Supplementary Table 3** for genomic coordinates and primer sequences). Expected band sizes for the reference and nonreference SV alleles are given at the top of each lane. The difference in size of the products for the reference and nonreference alleles confirmed the presence of the SVs for all loci except 6, 13 (confirmed by LongAmp Taq in a separate experiment), 21, 25 and 36. M1 is a 100-bp marker and M2 is a 1-kb marker. (**c**) A subset of SVs, which were confirmed by sequencing, was analyzed in nine additional genomic DNA samples (HapMap individuals with ancestry in Europe) to test for SV frequency within the CEPH population. An asterisk indicates that the SV is present polymorphically.

reference genome (**Supplementary Table 3**). Specifically, PCR amplification of predicted nonreference SV alleles[5] was used as a means for validation. In 48 cases the predicted SVs were validated, and in one case the reaction was inconclusive (**Fig. 2b** and **Supplementary Fig. 3**). Furthermore, seven reactions neither revealed the reference allele nor the predicted SV allele. (This primer failure rate can be explained by repetitive and GC-rich sequences that occur in association with SVs.) Finally, in a single case only the reference allele was found, suggesting either a false-positive prediction or the inability to amplify the event band of a predicted size of 7.5 kb.

We then sequenced 12 of the PCR-validated amplicons with Sanger capillary sequencing and confirmed the predicted breakpoint in all—that is, the Sanger-sequenced junction was identical to that in the library, with few single base-pair differences (presumable SNPs). We also analyzed a panel including nine unrelated CEPH individuals for the presence of six of the sequenced SVs and found that most SVs (four) were present polymorphically, whereas the remaining SVs likely represent rare alleles (**Fig. 2c** and **Supplementary Table 3**). All together, 48 out of 57 predicted SVs (84.2%) were confirmed successfully, and the validation rate was estimated at 98% (48 out of 49) based on the PCR reactions that could be scored, demonstrating high specificity. Notably, as about half of our validated SVs were low-support SV calls, our validations demonstrate that accurate calls are generated both at high- and low-support levels. This suggests that BreakSeq may perform reasonably even in conjunction with low-coverage sequencing projects.

Inferring ancestral states of SV loci by comparing breakpoint junctions to primate genomes

Global SV surveys have so far reported SV events such as insertions and deletions using operational definitions—that is, comparisons

© 2010 Nature America, Inc. All rights reserved.

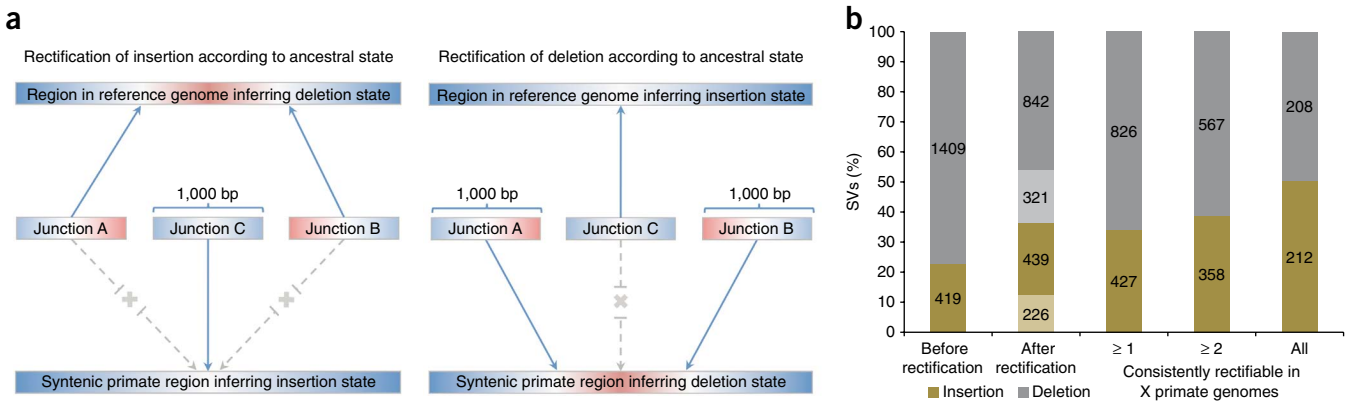


Figure 3 Ancestral state classification. (**a**) Junction sequences are aligned onto syntenic regions of a nonhuman primate genome to infer SV ancestral states. For rectifying an SV insertion event (from deletion) according to ancestral state (left), sequences A and B represent the junction sequences of the reference SV allele, whereas sequence C represents the junction sequence of the nonreference SV allele. For rectifying an SV deletion event (from insertion) according to ancestral state (right), sequence C represents the junction sequence of the reference SV allele and sequences A and B represent the junction sequences of the nonreference SV allele. Solid lines with arrows indicate successful alignments and dashed lines with crosses indicate no proper alignment. (**b**) Results of classifying SVs as insertions or deletions according to ancestral state. An SV event is defined as 'rectifiable' (indicated by darker color) if unambiguous high-quality alignments to putative ancestral regions could be constructed for the loci in any primate genomes (regardless of whether the classification is changed), and as 'unrectifiable' (represented by lighter color) if not.

© 2010 Nature America, Inc. All rights reserved.

with the human reference genome or allele frequency measurements. However, we reasoned that a systematic assessment of SV formation requires an unambiguous discrimination of SV event types—that is, one minimally affected by ascertainment biases. As the human reference genome presumably contains a mixture of common and rare SV alleles, it can serve only as a provisional reference for classifying SVs as insertions or deletions. Likewise, allele frequency measurements are of limited use in the context of classifying SVs into 'gains' and 'losses', as they may be affected by population-specific allele frequencies. In fact, ancestral state assignments facilitate systematic surveys of SVs in the context of studies focusing on human genome evolution, SV formational processes as well as minor and/or major allele assignment (as the ancestral allele often corresponds to the major one).

We therefore devised a framework that automatically assigns ancestral states of SV genomic loci based on a comparison of SV breakpoint junction sequence with the corresponding syntenic segments from the chimpanzee, orangutan and macaque genomes. Our approach (**Fig. 3a** and Online Methods) involves extracting ± 500-bp flanking sequences around each breakpoint junction, combining them into putative ancestral regions (stretches resembling the allele present in the reference genome and stretches resembling the alternative allele), and then comparing the regions with syntenic primate genome sequences to deduce the most likely ancestral state. We defined SV loci as 'rectifiable' if unambiguous high-quality alignments to putative ancestral regions could be constructed for the loci in any primate genomes.

Overall, ancestral states of 1,281 (70%) out of 1,828 SV indel events could be assigned. For the vast majority of these (1,142), the chimpanzee genome contributed to the ancestral state assignment. For an additional 139 cases located in hard-to-align regions in the chimpanzee genome (e.g., sequence assembly gaps), the ancestral state was inferred based on aligning junctions to the orangutan and macaque genomes. After ancestral state assignment, 665 SVs (36%) were classified as insertions and 1,163 (64%) as deletions. Furthermore 925 out of the 1,281 events were consistently rectifiable in at least two genomes. Of those, 420 were consistently rectifiable in all three genomes, with an approximate balance between insertions (212) and deletions (208) (**Fig. 3b**). We note that this balance differs substantially from earlier provisional SV classifications, which were strongly biased toward deletions, probably owing to the difficulty of many SV detection approaches in identifying insertions relative to the reference genome.

Inferring mechanisms of SV formation

Breakpoint junction sequences can also be used to deduce the molecular mechanisms of origin for SVs. To systematically classify SVs in our library, we evaluated previously reported signatures of particular formation mechanisms (such as VNTR, TEI, NAHR and NHR) with a computational pipeline (**Fig. 4a** and Online Methods). TEIs can be identified by the underlying genomic signatures of transposable elements; VNTRs, by underlying tandem repeats and low-complexity DNA; NAHRs, by the extended stretches of high sequence identity at the breakpoint junctions; and NHRs, by events lacking the former patterns. Parameters of the pipeline were chosen so as to yield results comparable to those achieved manually; in this regard, we confirmed the applicability of the chosen parameters by performing a sensitivity analysis (Online Methods and **Supplementary Fig. 4**).

We found, consistent with earlier findings based on considerably smaller data sets[5,25], that NHR events constitute the most abundant mechanism of SV formation in the genome (**Fig. 4b**). Our analyses inferred NHR as the formation mechanism for nearly half of all SVs in our set (45%), whereas 28% involved NAHRs, 21% involved TEIs, 5% involved VNTRs and 2% were ambiguous (the full list of events is available in **Supplementary Table 1**). Although VNTRs have the ability to contract and expand more than a kilobase, most of the 92 VNTRs identified in this study involved simple repeat units <1 kb in size. We thus reasoned that they do not fall strictly into the stringent SV definition given above and excluded VNTRs from most of the remaining analyses below. Additionally, for NAHR and TEI mechanisms, we focused on the high-confidence sets in the analyses unless indicated otherwise (Online Methods).

We then analyzed SV formation mechanisms of 1,281 rectifiable SV-indel events. As discussed above, SVs were provisionally mostly reported as deletions owing to ascertainment biases[5,16,21], regardless of the respective formation mechanisms. For example, despite the fact that retrotransposons are thought to move within the genome by a 'copy-and-paste process' involving reverse transcription of RNA intermediates and insertion of full-length or fragmented mobile

© 2010 Nature America, Inc. All rights reserved.

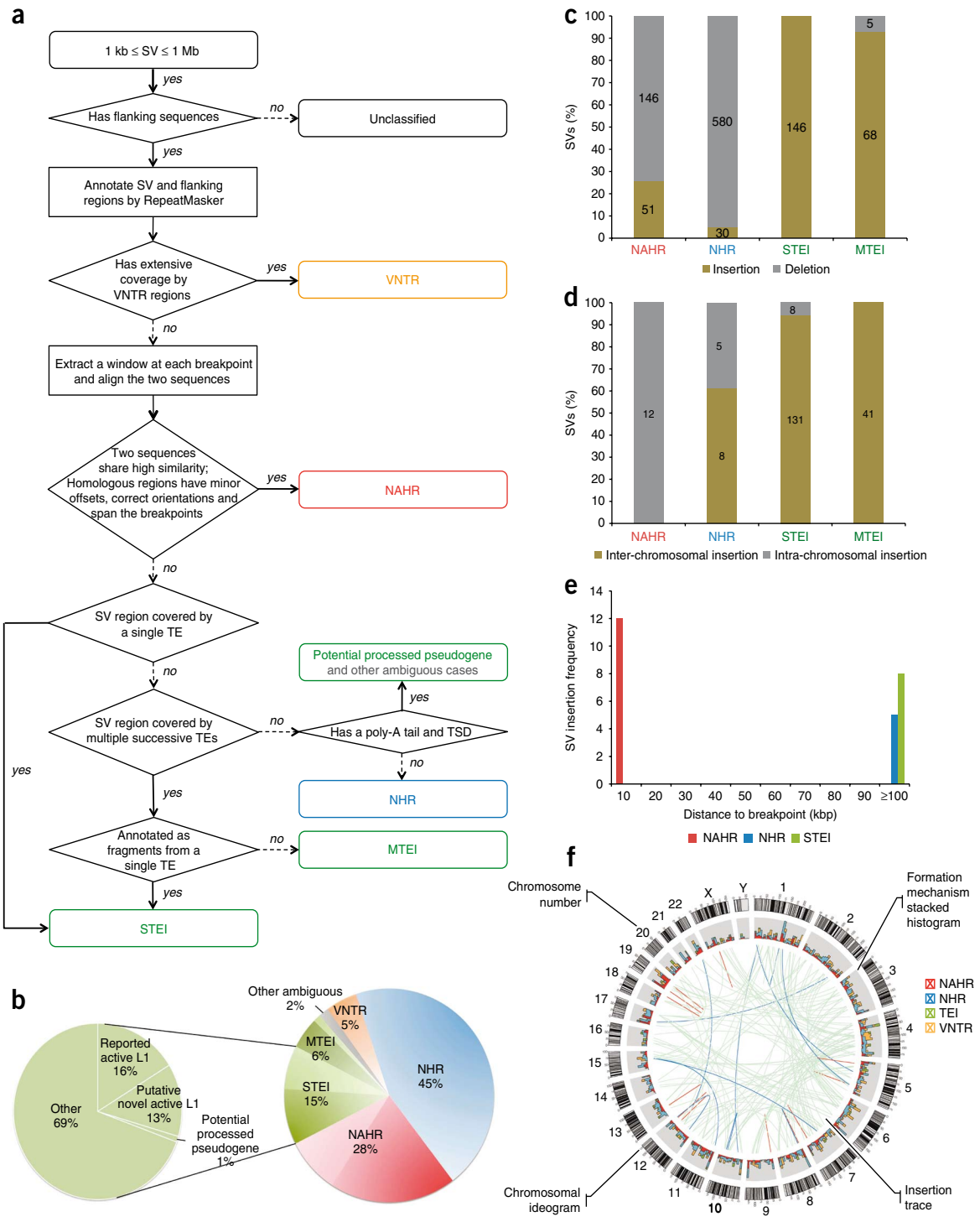


Figure 4 Inferring mechanisms of SV formation. (**a**) Pipeline for classifying SV-formation mechanisms. TE, transposable element. TSD, target site duplication. (**b**) Mechanisms of formation inferred for SVs in the library (larger circle on right). In NAHR (red) and MTEI/STEI (green), darker wedges represent high-confidence classification subsets, and lighter wedges are extended subsets. STEI is further subdivided in the left circle according to the fraction of previously reported L1 insertions[26], novel L1 insertions and processed pseudogene insertions in our data set. STEI, single transposable element insertion; MTEI, multiple transposable element insertion. (**c**) SV-indel distribution for all rectifiable events, broken down by formation mechanism. (**d**) Distribution of inter- versus intra-chromosomal events for all consistently rectifiable insertions, broken down by formation mechanism. (**e**) Distances of putative ancestral loci to insertion sites for all consistently rectifiable intra-chromosomal insertions, showing that intra-chromosomal NAHR insertions usually involve nearby sequences, whereas TEIs and NHR-associated insertions usually involve distant sequences. (**f**) Genome-wide view of insertion trace. The outermost circle represents chromosomal ideograms; the second circle represents SV formational mechanisms of 1,554 events in a stacked histogram. The lines in the innermost circle indicate the origin of the insertion sequences in the human genome for all 321 consistently rectifiable insertions.

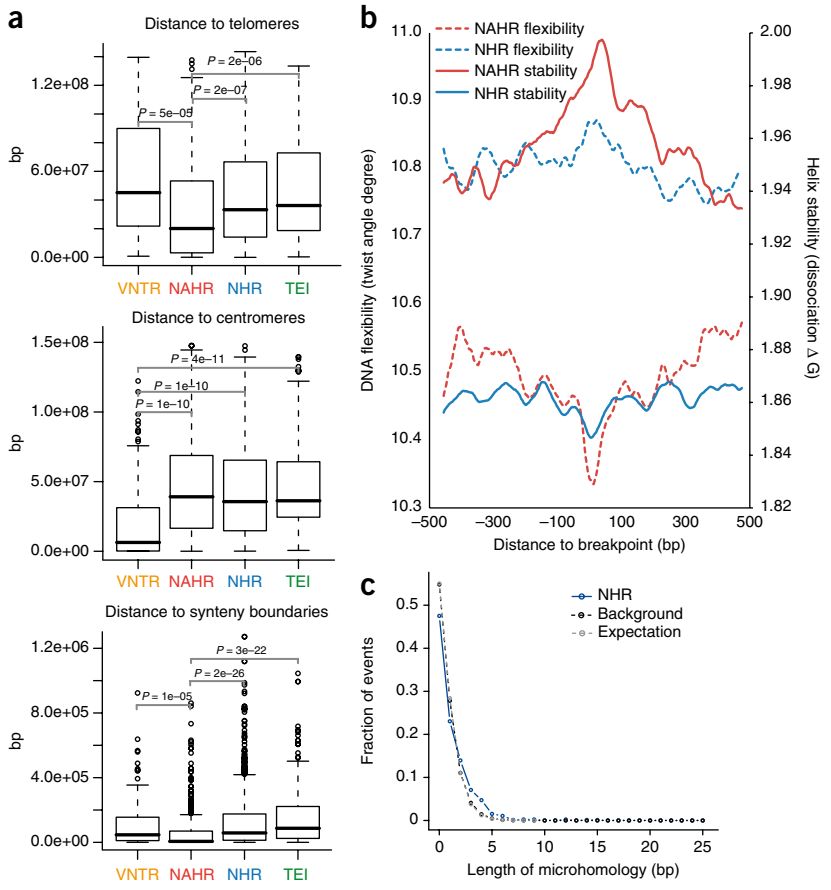


Figure 5 Analysis of breakpoint features. (**a**) Distance to chromosomal landmarks. Brackets indicate significantly different classes ($P < 0.05$ in Wilcoxon rank sum test after multiple hypothesis test correction by the Holm method). NAHR events are found to be significantly closer to telomeres and human-chimpanzee synteny block boundaries than the other mechanistic classes; VNTRs are significantly enriched in centromeric and pericentromeric regions. (**b**) DNA flexibility (dashed lines and left *y*-axis) and helix stability (solid lines and right *y*-axis) around NAHR and NHR breakpoints. (**c**) Distribution of NHR events with different lengths of microhomologies at the breakpoints. Microhomologies are significantly enriched in NHR breakpoints compared to a random background (KS test, $P = 2.43E\text{-}11$).

elements[26], most TEIs were previously annotated as deletions. Nevertheless, our ancestral state analysis revised the actual locus origin for a considerable number of SVs, and helped to resolve this apparent contradiction. Our results show that nearly all SVs associated with transposable elements for which ancestral states could be assigned were categorized as insertions (98%).

Through manual inspection, we found that the remaining transposable element–associated deletions can be reasonably explained as NHR-mediated SV deletions in regions of concentrated transposon annotations, which are difficult to distinguish from retrotranspositions. This shows that using the class name TEI was justified in retrospect, and that our ancestral analysis pipeline is able to produce results consistent with prior knowledge on the formation mechanism of TEI. On the other hand, even after classification by ancestral states, NAHR and NHR events were mostly annotated as deletions (**Fig. 4c**), which may be due to biases of these formation mechanisms toward deletions (as previously reported for NAHR[7]) or due to biases in SV detection methods toward ascertaining deletions in ancestral loci.

Further analysis of TEI events showed that they involved LINEs, SINEs, LTR-elements, composite retrotransposons and processed pseudogenes. Our results show that LINE-1s (L1s) represent the most abundant class at the given size range (>1 kb) as expected[27], with 71% of the TEIs mediated by LINE/L1 transposable elements. Although many transposable elements in the human genome have lost their ability to retrotranspose autonomously, several full-length elements, including 147 L1s, are still implicated in recent or ongoing retrotransposition activity[26]. Interestingly, our results suggest the possible recent activity in the human population of at least 84 L1 elements, which were reported by our pipeline as 'full-length' with poly-A tracts and target-site duplications. To the best of our knowledge, 38 of these putative active mobile elements have not yet been implicated with recent L1 activity (**Fig. 4b**, **Supplementary Table 1** and **Supplementary Fig. 5**). The remaining TEIs include three potential processed pseudogenes that were identified on the basis of their spliced primary transcripts, poly-A tracts and target site duplications (**Fig. 4b** and **Supplementary Table 4**).

We then focused on SVs associated with NAHR and NHR. Because these SVs mostly involve deletions relative to ancestral sequence, we reasoned that they might represent a particularly interesting class of SVs with potential impact on conserved DNA sequence. In fact, we found that 41% and 33% of the NAHR and NHR-based deletions, respectively, intersect with annotated exons from RefSeq genes (Online Methods) and thus may have a functional impact. On the other hand, insertions generated by NAHR or NHR have thus far received little attention, presumably due to difficulties in tracing these. Therefore, we extended our analysis to infer the most likely loci of origin of the inserted DNA sequences for 427 consistently rectifiable insertions (Online Methods). We found that NAHR insertions usually involve nearby sequence stretches stemming from the same chromosome as would be expected from the NAHR duplication mechanism. On the contrary, TEIs were found to originate randomly from inter-chromosomal locations in the genome, probably owing to the nature of retrotransposition of RNA intermediates. Furthermore, NHR-based insertions commonly involve both intra- and inter-chromosomal rearrangements (**Fig. 4d–f**).

Insights into SV formational biases

Finally, we analyzed the relationship between mechanisms of SV formation and sequence features located near to the breakpoints (including chromosomal landmarks, recombination hotspots, repeat sequences, GC content, short DNA motifs and microhomology regions). Briefly, we first extracted the DNA sequences flanking both sides of each breakpoint junction. In the case of insertions, junction sequences included flanking DNA reconstructed from the inserted sequence. We also generated two random background sets, one by randomly picking sequences from the reference genome (global background), and the

© 2010 Nature America, Inc. All rights reserved.

other by randomly picking DNA sequences from the local sequence context specific to each mechanistic class (local background). We then identified sequence features in the flanking regions of each breakpoint and calculated their enrichment with *P*-values based on randomization tests (Online Methods). We also tested for significant differences between SV formation mechanisms with respect to each feature using a Wilcoxon rank sum test (**Fig. 5a** and **Supplementary Fig. 6**).

We correlated SVs with chromosomal landmarks and found that NAHR events are significantly ($P \leq$ 1E-05) more proximal to telomeres and human-chimp synteny block boundaries than the other mechanistic classes. Moreover, we observed that VNTRs are significantly ($P \leq$ 1E-10) enriched in centromeric and pericentromeric regions, as expected (**Fig. 5a**). These results demonstrate a nonuniform distribution of SV formation mechanisms in the human genome (**Fig. 4f**).

We correlated SVs with recombination hotspots[28] and observed that they are significantly enriched for NAHR events (1.5-fold enrichment; P = 2.96E-03). Recombination hotspots are typically enriched for segmental duplications[29], which may act as mediators for NAHR during meiotic recombination. We further observed biases toward recombination hotspots for TEIs (**Supplementary Table 5**), but not for NHR-mediated events. Whereas the accumulation of TEIs might in part be due to the formation of such elements by NAHR-mediated recombination involving interspersed repeat sequence, the lack of an enrichment for NHR indicates that DNA double-strand breaks occurring during recombination might be insufficient for initiating double-strand repair mediated by nonhomologous end-joining.

We assessed associations between SV formation mechanisms and common repeat elements in the genome. For example, NAHR events have previously been reported to be associated with various types of genomic DNA repeats, in particular segmental duplications[5,6,16]. After classification of NAHR events by our pipeline, we confirmed that significant ($P \sim 0$) associations with segmental duplications are present both for NAHR-insertions (3.9-fold) and NAHR-deletions (7.4-fold). Furthermore, we found NAHR significantly ($P \sim 0$) associated with the SINE/Alu class of mobile elements. On the other hand, LINE elements (both the L1 and L2 classes) were significantly ($P \leq$ 1E–03) depleted among the NAHR events in our set whereas NHR events did not show significant enrichment (or depletion, except marginally for L2) with genomic repeat-structure (**Supplementary Table 5**).

We analyzed various features related to the physical properties of DNA at SV breakpoint junctions. In contrast to NHR, NAHR events were found to be biased toward GC-rich regions (**Supplementary Table 5**). A possible explanation for this bias is the known GC-richness of recombination hotspots[30], which we found to be significantly (P = 2.96E–03) enriched for NAHR events. Further, our results may indicate SV formation biases owing to DNA duplex stability. We thus extended our analyses by two additional features: DNA helix stability predicted by calculating the average of the dissociation free energy of each overlapping dinucleotide[31], and DNA flexibility based on the calculation of the average of the twist angle among each overlapping dinucleotide[32]. Our results indicate that in contrast to NAHR, NHR events are associated with high DNA flexibility and low helix stability, both of which are believed to be markers of fragility[33]. This is possibly due to sequence-specific biases for SV formation (**Supplementary Table 5**). We went on to characterize the change of these fragility marker signatures in a region of ± 500 bp around the breakpoint by smoothing the signal with a 50-bp sliding window. Interestingly, we observed that the strength of the marker signatures was most extreme at or very close to the SV breakpoints (**Fig. 5b**).

We reasoned that our comprehensive breakpoint junction library may enable us to identify simple DNA sequence motifs associated with SV breakpoints. Thus, we used the MEME tool[34] to carry out a comprehensive search for DNA motifs (6–12 nt, Online Methods) and found a significant enrichment (2.1-fold; $P \sim 0$) of the dinucleotide repeat $(TG)_6$ near breakpoints of NHR events, a sequence motif that fits with their relatively neutral GC content as shown above. We further analyzed all the NHR breakpoint sequences and found that the maximum consecutive occurrence of the TG-dinucleotide was 26. The MEME search did not reveal significantly enriched sequence motifs near NAHR or TEI events. Nevertheless, we used the MAST tool[34] to search for the DNA sequence motif 'CCNCCNTNNCCNC' that recently was reported to be associated with chromosomal recombination hotspots[35], and found a significant enrichment (1.5-fold; $P \sim 0$) of the motif near NAHR-associated SVs, but not near NHR- or TEI-associated SVs.

Previous studies have observed the occurrence of stretches of short repeating sequences of 2 to ~10 bp (that is, microhomologies) at the breakpoints of NHR events[18,36]. We used our breakpoint junction library to scan NHR breakpoints for microhomology stretches of different lengths, and observed statistical enrichment relative to a random background (1.4-fold on average; KS test, P = 2.43E-11; **Fig. 5c** and **Supplementary Table 6**) as expected. This suggests a strong association of microhomology stretches with SV formation by nonhomologous end-joining[36] or fork stalling and template switching[18].

DISCUSSION

In this study we presented a comprehensive library of 1,889 nonredundant SVs identified by breakpoint-resolution mapping in eight studies. Our approach, BreakSeq, leverages a breakpoint junction library for SV detection. Whereas other computational approaches for SV detection (such as paired-end mapping[5,37], DNA read-depth analysis[38–40] and split-read alignment analysis[41]) remain essential for identifying previously unknown SVs (a process that typically involves targeted PCR and sequencing), our approach serves as a tool for rapidly identifying specific SV alleles in personal genomics data. Specifically, by mining personal genomes for sequences present in the breakpoint junction library, BreakSeq leverages alternative, nonreference genomic sequence data to rapidly detect previously described SVs that short-read based personal genomics surveys commonly fail to ascertain. As such, BreakSeq enables a step towards overcoming reference bias, which is the favoring in ascertainment of SV alleles present in the human reference genome sequence.

We foresee that the utility of BreakSeq will increase as data sets grow (e.g., when SV calls from the 1000 Genomes Project are published). As our approach has a linear time complexity (Online Methods), it is easily extendable to larger data sets. In this regard, the size of our junction library currently comprises 0.004% of the reference genome in terms of nucleotide bases, and even a 100-fold increase of its size (>0.2 million SVs; ~10 times of DGV) will result in a data set considerably smaller than the reference genome. Thus, applying BreakSeq in personal genomics studies adds negligible computing efforts (compared to SNP genotyping) and at the same time dramatically improves SV calling. The library will be updated regularly to serve the personal genomics community in enabling precise SV detection with various next-generation sequencing platforms.

In addition to enabling accurate SV mapping, our junction library allows characterizing SV ancestral states. Whereas the ancestral states of SNPs and small indels have been inferred according to ancestral alignments in earlier studies[42,43], we here report systematic ancestral state inference for SVs. When applying our new classification approach to 1,281 SVs, we found that overall there is a balance of insertions and deletions, unlike most currently published SV sets that

© 2010 Nature America, Inc. All rights reserved.

display a considerable bias toward deletions. It should be noted that the nonhuman primate genomes used in our ancestral state inference correspond to single animals, which certainly do not represent idealized ancestral genomes. Nonetheless, we reasonably assume that SV loci can be classified at high confidence when ancestral states can be consistently inferred across three distinct primates.

Furthermore, we have developed a computational pipeline for classifying SVs according to their formation mechanisms and for analyzing various DNA sequence characteristics of the affected genomic loci. Together with the ancestral state analysis, this allowed us to analyze SV formation processes with respect to likely ancestral loci, an analysis that revealed some insights into SV formation. For example, our analyses suggest that the physical properties of the underlying DNA sequence influence locus-specific propensities for different SV formation mechanisms. We observed that NAHR-based SVs are associated with a relatively high GC content and with recombination hotspots, indicating that double-strand breaks occurring specifically during meiotic recombination contribute to NAHR-associated SV formation. On the other hand, NHR breakpoint regions appear to have lower DNA stability and higher flexibility, features that may increase the chance of double-strand breaks in general. Overall, our analysis reveals formational biases underlying SV formation and conforms to the fact that NAHR is driven by recombination between repeat sequences, whereas NHR is likely driven by DNA repair and replication errors.

By applying BreakSeq on a large scale, we envisage that it could be used for genotyping and determining SV allele frequencies. In fact, it should be possible to put each of the breakpoint sequences in our library directly onto a commercially available SNP chip, which could be used to precisely assess SV genotypes simultaneously with all of the SNPs in an individual. (This should add only a small number of probes to the ~1 M probes already on commercial chips.)

Lastly, we note that as our approach depends on current SV lists, it is inevitably affected by their existing biases owing to presently applied technologies. Likely biases include the difficulty in mapping insertions relative to the reference genome and in ascertaining SVs in repetitive regions, for example, segmentally duplicated sequences. We anticipate that in the near future, as technologies advance in terms of read-lengths, inherent biases against repeat-rich sequences will be further reduced and the mapping of SVs onto our junction library will further improve, making it essentially comparable to SNP genotyping. In this regard, as thousands of human genomes will be sequenced in the coming years, there will be a huge demand for reliable and accurate SV mapping and SV genotyping.

METHODS

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

Data accession. BreakSeq website (http://sv.gersteinlab.org/breakseq).

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

ACKNOWLEDGMENTS

We acknowledge support from the National Institutes of Health, the A.L. Williams Professorship funds and the European Molecular Biology Laboratory. We thank R. Alexander and E. Khurana for proofreading the manuscript, and A. Abyzov, Z. Zhang, T. Rausch and J. Du for helpful discussions. Finally, we thank the 1000 Genomes Project for early data access.

AUTHOR CONTRIBUTIONS

H.Y.K.L., X.J.M. and J.O.K. contributed equally to this work; M.B.G. and J.O.K. co-directed this work; M.B.G., J.O.K., H.Y.K.L. and X.J.M. designed the research; H.Y.K.L., X.J.M., A.M.S., A.T., P.D.C., M.S., P.M.K. and J.O.K. performed or provided direction for the analyses and/or experiments; M.B.G., J.O.K., H.Y.K.L., X.J.M., P.M.K. and A.T. wrote the manuscript.

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

1. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
2. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
3. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
4. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
5. Korbel, J.O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
6. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
7. Turner, D.J. *et al.* Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nat. Genet.* **40**, 90–95 (2008).
8. van Ommen, G.J. Frequency of new copy number variation in humans. *Nat. Genet.* **37**, 333–334 (2005).
9. Korbel, J.O. *et al.* The genetic architecture of Down syndrome phenotypes revealed by high-resolution analysis of human segmental trisomies. *Proc. Natl. Acad. Sci. USA* **106**, 12031–12036 (2009).
10. Sharp, A.J. *et al.* Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat. Genet.* **38**, 1038–1042 (2006).
11. McCarroll, S.A. *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat. Genet.* **40**, 1107–1112 (2008).
12. de Cid, R. *et al.* Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nat. Genet.* **41**, 211–215 (2009).
13. Gonzalez, E. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**, 1434–1440 (2005).
14. Aitman, T.J. *et al.* Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature* **439**, 851–855 (2006).
15. Hastings, P.J., Lupski, J.R., Rosenberg, S.M. & Ira, G. Mechanisms of change in gene copy number. *Nat. Rev. Genet.* **10**, 551–564 (2009).
16. Kim, P.M. *et al.* Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome Res.* **18**, 1865–1874 (2008).
17. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
18. Lee, J.A., Carvalho, C.M. & Lupski, J.R.A. DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131**, 1235–1247 (2007).
19. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
20. Korbel, J.O. *et al.* PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol.* **10**, R23 (2009).
21. Wheeler, D.A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
22. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
23. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
24. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
25. Perry, G.H. *et al.* The fine-scale and complex architecture of human copy-number variation. *Am. J. Hum. Genet.* **82**, 685–695 (2008).
26. Mills, R.E., Bennett, E.A., Iskow, R.C. & Devine, S.E. Which transposable elements are active in the human genome? *Trends Genet.* **23**, 183–191 (2007).
27. Xing, J. *et al.* Mobile elements create structural variation: analysis of a complete human genome. *Genome Res.* **19**, 1516–1526 (2009).
28. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321–324 (2005).
29. Sharp, A.J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
30. Meunier, J. & Duret, L. Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.* **21**, 984–990 (2004).
31. Breslauer, K.J., Frank, R., Blocker, H. & Marky, L.A. Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. USA* **83**, 3746–3750 (1986).
32. Sarai, A., Mazur, J., Nussinov, R. & Jernigan, R.L. Sequence dependence of DNA conformational flexibility. *Biochemistry* **28**, 7842–7849 (1989).
33. Bailey, J.A. & Eichler, E.E. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat. Rev. Genet.* **7**, 552–564 (2006).
34. Bailey, T.L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202 (2009).

© 2010 Nature America, Inc. All rights reserved.

35. Myers, S., Freeman, C., Auton, A., Donnelly, P. & McVean, G. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat. Genet.* **40**, 1124–1129 (2008).
36. Linardopoulou, E.V. *et al*. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* **437**, 94–100 (2005).
37. Lee, S., Cheran, E. & Brudno, M. A robust framework for detecting structural variations in a genome. *Bioinformatics* **24**, i59–i67 (2008).
38. Campbell, P.J. *et al*. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40**, 722–729 (2008).
39. Chiang, D.Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods* **6**, 99–103 (2009).
40. Wang, L.Y., Abyzov, A., Korbel, J.O., Snyder, M. & Gerstein, M. MSB: a mean-shift-based approach for the analysis of structural variation in the genome. *Genome Res.* **19**, 106–117 (2009).
41. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
42. Paten, B. *et al*. Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.* **18**, 1829–1843 (2008).
43. Spencer, C.C. *et al*. The influence of recombination on human genetic diversity. *PLoS Genet.* **2**, e148 (2006).
44. Levy, S. *et al*. The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
45. Mills, R.E. *et al*. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* **16**, 1182–1190 (2006).

© 2010 Nature America, Inc. All rights reserved.

ONLINE METHODS

Data preparation. Our initial breakpoint library altogether represented 1,961 SVs identified at high precision based on the Nationl Center for Biotechnology Information (NCBI) build 36 of the human genome. It was compiled from eight different published sources based on paired-end mapping[5,16], fosmid-paired-end sequencing[3,6], Sanger capillary sequencing[44], resequencing of an individual human genome using second-generation sequencing[21], DNA resequencing traces for SNP discovery projects (support by at least two reads was required for an SV to be included in our data set)[45], and high-resolution array-based comparative genomic hybridization[25]. For the 253 SVs identified through fosmid-paired-end sequencing[3,6], 387 published sequenced clones originally used to identify SVs in NCBI build 35 were realigned to the NCBI build 36 human genome before inclusion in the library. A split-read analysis was then carried out using BLAT to infer the breakpoints of the events. For the 98 SVs from resequencing traces[45], the liftover tool available at the UCSC genome browser (http://genome.ucsc.edu/) was used to convert the breakpoint coordinates from human NCBI build 35 to build 36. All SVs in our analysis were between 1 kb and 1 Mb in length (that is, we removed events >1 Mb, reasoning that they may be lower in confidence). After accounting for redundancy, our standardized breakpoint library consisted of 1,889 SVs that were used in all subsequent calculations and analyses.

SV mechanism classification pipeline. Four major steps were involved in our procedure to classify SV formation mechanisms. First, SVs were examined for extensive coverage by tandem repeats and regions of low complexity (here, low-complexity DNA refers to micro-satellite DNA, polypurine/polypyrimidine stretches, and regions of extremely high AT or GC content, as defined by the RepeatMasker program; http://www.repeatmasker.org/) to identify instances of expansion or contraction of VNTRs. Second, ±100-bp flanking sequences derived from both breakpoint junctions were aligned against each other to scan for blocks of extensive homology. SVs were classified as 'high-confidence NAHR' if the homologous blocks had a minimum sequence identity of 85%, a minimum length of 50 bp for the identical sequences, a maximum offset of 20 bp between the homologous blocks, correct orientations and covered the breakpoints. SVs displaying at least three but not all of the above criteria were classified as 'extended NAHR'. Third, SVs aligning to known interspersed mobile elements carrying the common diagnostic features of corresponding transposable elements, that is, target site duplications and poly-A tracts[26], were classified as 'high-confidence TEIs'. Events missing one or more of the diagnostic features were classified as 'extended TEIs'. TEIs were furthered categorized as single transposable element insertions (STEIs) if a single element was involved and multiple transposable element insertions (MTEIs) if multiple elements appeared to be involved. Furthermore, full-length TEIs were discriminated from transposable element fragments and transposable element subfamilies were also recorded. Through identification of spliced protein-coding gene sequences and TEI-diagnostic features, processed pseudogenes likely inserted via a TEI-associated mechanism were also identified. Finally, SVs lacking signatures of any of the above diagnostic sequence features were classified as NHR events.

Sensitivity analysis for the SV mechanism classification. Sensitivity analysis was performed on five key parameters used in the mechanism classification pipeline (**Supplementary Fig. 4**). Classification results were examined as each parameter was varied over a large range while fixing the other parameters at default values. First, the cutoff for the length of homologous blocks in the flanking sequences alignment for classifying NAHR events (NAHRhomolen) was varied from 10 to 150 bp with a step size of 10 bp. Second, the cutoff for the percentage identity of homologous blocks in the flanking sequences alignment for classifying NAHR events (NAHRpct) was varied from 70 to 100% with a step size of 1%. Third, the cutoff for the coverage of VNTR regions in the SV was varied from 0 to 100% with a step size of 5%. Fourth, the window size used to examine the consistency of the transposable element boundary with a breakpoint for classifying STEI and MTEI events (TEIwin) was varied from 10 to 400 bp with a step size of 10 bp. Finally, the gap size used to examine whether adjacent transposable elements can be joined for classifying MTEI events (TEIgap) was varied from 0 to 300 bp with a step size of 10 bp. Default values for NAHRhomolen, NAHRpct, VNTRcutoff, TEIwin and TEIgap used in the pipeline were 50, 85, 50, 200 and 150, respectively.

Analysis of ancestral state. For a 'deletion' relative to the reference genome, a ± 500-bp flanking sequence at each breakpoint was extracted to obtain two sequences of 1,000 bp representing both the left (A) and right (B) breakpoint junction sequences. Then a 1,000-bp junction sequence at the breakpoint of the alternative allele, representing 500 bp upstream and downstream of the left and right breakpoints, respectively (C), was also extracted. If C aligned onto a nonhuman primate genome (that is, a potential ancestral genomic locus) at high-quality and with better length and sequence identity (represented by the BLAT score) than A and B, then the event was rectified as an insertion relative to the ancestral genome. Conversely, for an 'insertion' relative to the reference genome, the A, B (alternative allele) and C (reference allele) junction sequences of the event were extracted. If A and B both displayed an alignment better than C onto a nonhuman primate genome, the event was rectified as a deletion relative to the ancestral genome.

All the alignments were performed using BLAT on the chimpanzee (panTro2), macaque (rheMac2), and orangutan (ponAbe2) genomes, the sequences of which were downloaded from the UCSC genome browser (http://genome.ucsc.edu/). The Net alignments[46,47] from UCSC were also downloaded and the top level was chosen to verify that the alignment of the junction sequences were in the syntenic regions of the corresponding SVs. Because all the primate ancestral genomes are highly similar, the alignment identity and coverage were required to be >90%. Furthermore, the length ratio of target versus query was required not to exceed a deviation of 10%.

SVs were classified as 'rectifiable' if unambiguous high-quality alignments to putative ancestral regions could be constructed in any nonhuman primate genome. Particularly, an SV was classified as 'rectified' if its state was changed from its original state to another after the analysis (from deletion to insertion, or vice versa). The state of each SV was then assigned based on the closest nonhuman primate genome (e.g., from chimpanzee to orangutan and to macaque) in which a corresponding syntenic region existed. SVs were considered as 'consistently rectifiable' if they were rectified to the same state with no inconsistent ancestral assignment inferred.

Insertion trace. After rectification based on the ancestral state analysis, all insertions that were consistently rectifiable were aligned onto the human reference genome to scan for the presumable origin of the inserted sequences. Because the inserted sequence of an event rectified from a deletion is already present in the reference genome, any alignments overlapping with >50% of the SV region were discarded and the next best match was chosen. BLAT alignments tracing inserted sequences were required to have a sequence identity >90%.

Enrichment calculation. To calculate the enrichment and *P*-value for each feature and repeat association with breakpoints, a nonparametric randomization test based on sampling was employed. For the observed samples, the exact coordinates of the breakpoints were taken for location-dependent computation and sequences flanking the breakpoints were extracted for sequence-dependent computation. A random global background was generated by randomly sampling a set of coordinates, or sequences with the same length, of the same amount from the reference genome (build 36). Similarly, a local background was generated by randomly sampling in a 10-kb window at the breakpoints. The sampling was repeated 1,000 times with replacement and the observed statistic of the breakpoints was tested against the sampling distribution based on the whole genome. The enrichment value was calculated by comparing the observed statistic over the mean of the statistics of the samplings. Then, the *P*-value of the enrichment was calculated by counting the number of samplings that yielded a statistic as extreme as, or more extreme than, the observed one. The enrichment was reported as significant for any $P < 0.05$.

Correlation of chromosomal landmarks. Distance to telomeres was calculated from the midpoint of an SV to the end of the chromosome in the same arm. Distances to centromeres and pericentromeric gaps were calculated from the midpoint of an SV to the closest centromeric or pericentromeric gap boundary on the same chromosome. Distance to the closest synteny block boundary was calculated by computing the distance from each breakpoint to

© 2010 Nature America, Inc. All rights reserved.

doi:10.1038/nbt.1600

the closest synteny block boundary and then taking the average for the two breakpoints. Synteny block boundaries were taken from the human-chimpanzee Net alignment file[46,47] available at the UCSC genome browser and the 'gap' type was excluded from the analysis. A Wilcoxon rank sum test was then done to compare the distance measurements of different formation mechanisms in a pair-wise fashion, followed by a correction for multiple hypothesis testing using the Holm method.

Feature computation. We considered the following features at SV breakpoints in our analysis: GC content, helix stability and DNA flexibility. All features were computed for sequences within 50 bp of the breakpoints or randomly extracted from the genome. GC content was calculated by computing the percentage of guanine and cytosine nucleotides over the given length of the sequence. Helix stability of the DNA duplex was predicted by calculating the average of the dissociation free energy of each overlapping dinucleotide[31]. Similarly, DNA flexibility was estimated by calculating the average of the twist angle among all overlapping dinucleotides[32]. To observe the change of the DNA flexibility and helix stability around a breakpoint, values at each nucleotide were smoothed using a sliding window of 50 bp, which was slid across an interval of 1 kb centered on the breakpoint.

Repeat association. The association of repeat elements and pseudogenes was calculated by intersecting the relevant data sets. Each element was overlapped with a breakpoint and the average number of overlapping elements for all the input breakpoints was calculated. Repeat elements in the human genome build 36 were downloaded from the RepeatMasker track of the UCSC genome browser (March 2006 assembly). Only the elements annotated with repeat classes SINE and LINE were included in this analysis. In total, there were 1,783,897 SINE elements and 1,407,547 LINE elements of which 1,193,509 were Alu elements and 927,909 were L1 elements, respectively. For the pseudogene analysis, we used PseudoPipe[48] to identify pseudogenes in the genome based on the protein annotations in the Ensembl database (release 48). This analysis involved 2,454 duplicated pseudogenes and 10,999 processed pseudogenes.

Motif discovery. MEME was used to discover sequence motifs near SV breakpoints and to generate position weight matrices (PWMs) for significantly enriched motifs. The input data to MEME were sequences of 200 bp centered on the breakpoints. Motif width was allowed to range from 6 bp to 12 bp. For SVs classified as NAHR-mediated we also looked for an overrepresentation of a previously described sequence motif specific to recombination hotspots[35]. The recombination-hotspot motif was converted into a PWM by considering the average genomic frequencies of the four bases ACGT (0.295, 0.205, 0.205, 0.295) and by adding pseudocounts of 1. After identifying the motifs, MAST was applied to search for a motif match in the original set and the global background set. The *P*-value cutoff for each motif match was $P < 0.0001$ and a randomization test was performed as described above to calculate the enrichment *P*-values for each motif.

Microhomology enrichment analysis. The lengths of the microhomology sequences at the breakpoints of NHR-mediated events were compared with the local background and a theoretical distribution. The theoretical expectation was calculated by assuming independence between genomic positions and a uniform distribution of the four nucleotides (ATCG) in the genome. The formula $P \times (1-P)^2 \times (i+1)$ was used to calculate the probability of observing homology of a specific length, where i is the length of homology and P is the probability of observing the same pair of nucleotides at the given genomic positions (that is, $P = p(\mathrm{A})^2 + p(\mathrm{T})^2 + p(\mathrm{C})^2 + p(\mathrm{G})^2$ and $p(\mathrm{A,C,G,T}) = (0.295, 0.205, 0.205, 0.295)$ were estimated from the local background). A one-sided Kolmogorov-Smirnov test (KS-test) was performed to test the enrichment of microhomologies in NHR compared to the local background. The size of the effect was calculated as the fold enrichment of microhomology stretches between NHR and the background.

Mapping SVs with a junction library. The breakpoint junction mapping approach that we developed works as follows. The junction library for SV mapping is created by joining 30 bp flanking sequences on each side of a breakpoint. A deletion event is represented with a single junction sequence in the library, while an insertion has both a left and right junction sequence corresponding to each of its breakpoints. DNA reads from personal genomes are aligned against the junction library. Reads are required to overlap a breakpoint by at least 10 bp on each side. All successfully mapped reads are then aligned against the reference genome. Only those reads that do not map onto the reference genome are labeled as 'unique' in the personal genome; the other reads are labeled as 'nonunique'. A short-read aligner, Bowtie[49], is used to perform all the alignments (allowing for two mismatches). To score the SV candidates on the basis of supportive hits, the following formula is used:

$$S_i = \max(0, \log_2 T_i - \log_2 R_i)$$

where S_i is the score representing the effective number of hits (supportive hits) in $\log_2$ scale for SV i, with unique and nonunique hits denoted as T_i and R_i respectively. If T_i or R_i is 0, the log term is replaced by 0. A score of 1 thus indicates 2 supportive hits, whereas scores >2 (high-support) indicate the presence of >4 supportive hits.

The mapping process showed a linear time complexity in practice. On average, it required 8 h to run our junction-mapping program (open-sourced and available for download at http://sv.gersteinlab.org/breakseq) against a sequenced genome at 40× physical coverage on a 3GHz quad-core computer node with 16GB physical memory. All identified SVs for the YRI and HCH genomes are listed in **Supplementary Table 2**; for NA12891, in accordance with prepublication agreements for 1000 Genomes Project data, we only provide the coordinates of SVs identified on a single chromosome (that is, chromosome 6).

Intersection of the breakpoint junction library with RefSeq genes. RefSeq gene annotations were downloaded from the UCSC Genome Browser. Intersection of the SVs in our breakpoint junction library and RefSeq genes were found by comparing the start- and end- coordinates of the two datasets. For insertion events whose inserted sequences could be traced, the positions from which the insertions were derived were compared to the RefSeq gene annotations. In particular, 60 out of 146 NAHR deletions and 193 out of 580 NHR deletions intersected with annotated exons from RefSeq genes. Insertions were also found to have an impact on coding regions, with 19 out of 51 NAHR insertions and 11 out of 30 NHR insertions intersecting with the exons. These included cases where exons at the insertion site were altered by the insertion event (19 NAHRs and 7 NHRs) and where the inserted sequence was itself derived from exonic DNA (3 NAHRs and 6 NHRs).

PCR validation. We tested by PCR validation 24 insertion and 33 deletion calls predicted in NA12891 relative to the reference genome (**Supplementary Table 3**). Specifically, we designed PCR primers as previously described[5] and amplified the predicted nonreference SV alleles. For the PCR, 10ng of genomic DNA (Coriell Institute were used with the SequalPrep Long PCR Kit (Invitrogen) in 20 µl volumes using the following PCR conditions in a C1000 thermocycler (BioRad): 94 °C for 3 min, followed by 10 cycles of 94 °C for 10 s, 60 °C for 30 s and 68 °C for 10 min and 25 cycles of 94 °C for 10 s, 56 °C for 30 s and 68 °C for 10 min (+10 s/cycle), followed by a final cycle of 72 °C for 10 min. Some of the reactions that failed with the SequalPrep enzyme were amplified with the LongAmp Taq DNA Polymerase (NEB) or the iProof High Fidelity DNA Polymerase (Biorad). PCR products were analyzed on a 1% agarose gel stained with Sybr Safe Dye (Invitrogen). Marker M1 was a 100-bp ladder whereas M2 corresponded to a 1-kb ladder (500, 1,000, 1,500, 2,000, 3,000, etc) (NEB). Primers and polymerases are listed in **Supplementary Table 3**.

46. Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA* **100**, 11484–11489 (2003).
47. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
48. Zhang, Z. *et al.* PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics* **22**, 1437–1439 (2006).
49. 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).

© 2010 Nature America, Inc. All rights reserved.

nature biotechnology

Building the sequence map of the human pan-genome

Ruiqiang Li[1,2,7], Yingrui Li[1,7], Hancheng Zheng[1,3,7], Ruibang Luo[1,3,7], Hongmei Zhu[1], Qibin Li[1], Wubin Qian[1], Yuanyuan Ren[1], Geng Tian[1], Jinxiang Li[1], Guangyu Zhou[1], Xuan Zhu[1], Honglong Wu[1,6], Junjie Qin[1], Xin Jin[1,3], Dongfang Li[1,6], Hongzhi Cao[1,6], Xueda Hu[1], Hélène Blanche[4], Howard Cann[4], Xiuqing Zhang[1], Songgang Li[1], Lars Bolund[1,5], Karsten Kristiansen[1,2], Huanming Yang[1], Jun Wang[1,2] & Jian Wang[1]

© 2010 Nature America, Inc. All rights reserved.

Here we integrate the *de novo* assembly of an Asian and an African genome with the NCBI reference human genome, as a step toward constructing the human pan-genome. We identified ~5 Mb of novel sequences not present in the reference genome in each of these assemblies. Most novel sequences are individual or population specific, as revealed by their comparison to all available human DNA sequence and by PCR validation using the human genome diversity cell line panel. We found novel sequences present in patterns consistent with known human migration paths. Cross-species conservation analysis of predicted genes indicated that the novel sequences contain potentially functional coding regions. We estimate that a complete human pan-genome would contain ~19–40 Mb of novel sequence not present in the extant reference genome. The extensive amount of novel sequence contributing to the genetic variation of the pan-genome indicates the importance of using complete genome sequencing and *de novo* assembly.

The Human Genome Project[1] established the foundation for human genomics studies. Subsequent analyses unveiled genetic variations and identified their effects on phenotypic diversity and differences in disease susceptibility[2]. Guided by the National Center for Biotechnology Information (NCBI) reference genome, initial studies of human genetic variation focused largely on identifying[3] and cataloging[4,5] single-nucleotide polymorphisms (SNPs) and studying their association to human diseases[6]. Structural variation (which is thought to contribute more variant sequences than SNPs) has also been extensively identified and analyzed in the human genome[7–10].

The availability of a number of individual human genomes[11–15] has provided an unprecedented opportunity to investigate detailed genetic differences at the individual level. Preliminary analyses have revealed that these genomes contain sequences that could not be mapped onto the human reference genome (novel sequences), resulting in the proposal that the majority of these sequences likely belong to the gap regions in the current version of the human genome assembly[12]. When fosmid clones from HapMap samples were sequenced, 525 sequences were identified that mapped instead to highly polymorphic structural variant regions, among which 172 sequences appeared to be specific to the individual rather than to be sequences missing as a result of gaps in the reference genome[8]. Thus, the variable part of the sequence composition in the human genome (individual-specific sequences) may contribute considerable sequence divergence in addition to substitutions, base-pair level indels, rearrangements or copy number changes present in the commonly shared part of human genomes. Substantial effort has been taken to refine genomic reference sequences (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/), but thorough genome-wide characterization is still necessary to gain a comprehensive understanding of individual-specific sequences.

These intriguing findings and the limited amount of information about the extent and range of diversity that these individual-specific sequences contribute to human genetic variation prompted us to begin to build the human pan-genome—that is, the nonredundant collection of all human DNA sequence present in the entire human population. We assembled *de novo* the Asian and African complete individual genome sequences and compared them to the NCBI reference human genome. Our findings showed that human genomes contain a large amount of novel sequence that is both population and individual specific. Additional analyses allowed us to investigate the amount of sequence variation that is expected to exist between any two individuals as well as obtain information about the presence of potentially functional genetic elements within these novel sequences.

Our study also shows that combining individual-specific sequences with shared core human genome sequences will enable the creation of a human pan-genome that will be important for better understanding personal genomes and their use in medical genomics studies. Based on our findings here, it is also clear that establishing a complete human pan-genome will require using extensive sequencing data rather than relying primarily on array-based technologies that are dependent on the current reference genome.

[1]BGI-Shenzhen, Shenzhen 518083, China. [2]Department of Biology, University of Copenhagen, Copenhagen, Denmark. [3]School of Bioscience and Biotechnology, South China University of Technology, Guangzhou, China. [4]Fondation Jean Dausset, Centre d'Étude du Polymorphisme Humain (CEPH), Paris, France. [5]Institute of Human Genetics, University of Aarhus, Aarhus, Denmark. [6]Genome Research Institute, Shenzhen University Medical School, Shenzhen, China. [7]These authors contributed equally to this work. Correspondence should be addressed to Jun Wang (wangj@genomics.org.cn) or Jian Wang. (wangjian@genomics.org.cn).

Received 21 October; accepted 30 November; published online 7 December 2009; doi:10.1038/nbt.1596

Table 1 Summary of YH and NA18507 novel sequence identification

	YH		NA18507	
Step	**Number**	**Total length (bp)**	**Number**	**Total length (bp)**
1. Genome assembly	185,086	2,874,204,399	314,877	2,682,734,144
2. Sequences nonexistent in the reference genome (NCBI build 36)	7,211	5,125,070	7,330	4,798,833
3. Individual-specific novel sequence	6,144	4,957,235	7,305	4,790,170
3.1. Aligned to the sequence of at least one of the following human genomes	5,193	4,072,922	6,556	4,280,560
a. On NA18507/YH genome	2,626	2,655,416	4,514	3,412,855
b. On Watson's genome (raw reads)	2,221	348,668	2,229	343,668
c. On Venter's genome	3,665	1,011,133	4,427	1,357,377
d. Aligned to the 363 kb gap regions	105	87,199	147	77,255
e. Aligned to the GenBank human clones	2,871	2,697,530	3,709	2,719,344
3.2. Aligned to other mammal genomes	507	311,467	362	176,535
3.3. Unknown	444	272,424	387	184,142

We compared the assembly of YH (Asian) and NA18507 (African) genome sequences against the human reference genome (NCBI build 36) and identified sequences in YH and NA18507 not present in the reference genome. Potential plant or microbial contaminations were filtered, and the remaining sequences that could not be aligned to the reference genome were defined as individual-specific sequences. The fraction of sequences that aligned to other available human sequences (>90% identity) or showed homology with other mammalian genomes (Blast, 1e-20) provided evidence that these were human sequences. The 26 gaps (363 kb) in the human reference genome sequence were previously closed as described[17].

© 2010 Nature America, Inc. All rights reserved.

RESULTS

Short-read assembly and novel sequence detection

For our analysis, we used the raw data of the Asian (YH) genome that we previously sequenced and the raw data of the African (NA18507) genome that we downloaded from NCBI. These data were both produced using the Illumina Genome Analyzer (GA) and consisted of 117.7 Gb and 135 Gb of sequencing reads, respectively, with read-lengths of ~35 bp[13,14]. We also used an additional 82.5 Gb paired-end reads that we recently generated from YH with library insert sizes ranging from 200 bp to 9.6 kb (**Supplementary Table 1**), raising the total amount of YH sequence data used to 200.2 Gb.

We carried out *de novo* short-read assembly (Online Methods) and obtained a total assembled sequence size of 2.87 Gb for YH and 2.68 Gb for NA18507 (**Table 1**). The N50 scaffold size of the two genomes was, respectively, 446.3 kb and 61.9 kb, and the N50 contig size was 7.4 kb and 6.0 kb.

We aligned the YH and NA18507 assembly scaffolds against the NCBI human reference genome[16] (Online Methods) and found that 5.1 and 4.8 Mb of the sequence (see Discussion), respectively, was absent in the reference genome, where absent sequences were defined to be those >100 bp long and with <90% identity. We filtered the novel sequence to eliminate possible contamination by comparing these sequences to all known plant and microbe genomes, which resulted in a final total of 5.0 Mb of novel sequence in the YH genome and 4.8 Mb in NA18507 (**Table 1** and **Supplementary Data Sets 1** and **2**). We also assessed whether using a lower (80%) identity cutoff would substantially alter the results and found little difference, indicating that the identified novel sequences have no close homologs in the reference genome (**Supplementary Fig. 1**). In this study, we define the term 'novel sequences' to denote sequences that are present in at least one human individual but not in the NCBI reference genome.

We then bootstrapped subsets of read data for assembly and checked the coverage of these novel sequences. At a sequencing depth of 40-fold or above, >95% of all novel sequences could be assembled with the subset data (**Supplementary Fig. 2**), which indicates that the two assemblies covered essentially the complete set of nonrepeat novel sequences of the donors' genomes, including the unique sequences and consensus sequences of repeats. The sum of novel sequences and the NCBI reference genome provided a human pan-genome for further analyses.

Initial characterization of novel sequences

We first validated the identified novel sequences by comparing them to all previously published human genome assemblies and human genome clone sequences to search for any matches or homologs (>90% identity). We found that ~2.7 Mb of YH (54%) and NA18507 (56%) novel sequences overlapped with each other; 348.7 kb of the YH and 343.7 kb of the NA18507 sequences could be aligned to Watson's raw sequencing reads; 1.0 Mb of the YH and 1.4 Mb of the NA18507 sequences could be aligned to Venter's (so-called HuRef)[11] genome; 2.7 Mb of each could be aligned to known human clones deposited in GenBank; and 87.2 kb of the YH and 77.3 kb of the NA18507 sequences could be aligned to the 26 recently closed gaps on the NCBI reference genome[17]. In all, 4.1 Mb (82.1%) of the YH and 4.3 Mb (89.4%) of the NA18507 novel sequences could be aligned to other human sequences, indicating that these are valid DNA sequences (**Table 1**).

We further compared the remaining unaligned novel sequences to all available sequenced mammalian genomes, and found that 311.5 kb of the YH and 176.5 kb of the NA18507 sequences had homologs in these genomes (E-value = 1e-20). In all, only 272.4 kb (5.5%) of the YH and 184.1 kb (3.8%) of the NA18507 novel sequences could not be aligned to any known human or mammalian genome sequence. Additionally, for each individual genome (Venter's, YH, NA18507), only part of the identified novel sequences could also be found in the others, which indicated there is also an individual-specific distribution of the novel sequences.

We then investigated the length distribution of novel sequences, which revealed that 1,171 (16.25%) of the YH and 1,201 (16.38%) of the NA18507 novel sequences had lengths >1 kb, and that 33 and 9 sequences had lengths >10 kb (**Supplementary Fig. 3**). (See **Supplementary Discussion** for more information about the length of novel sequences.)

To assess how many of the YH and NA18507 novel sequences were likely to be insertions or deletions, we aligned flanking sequences at both ends of the novel sequences onto the NCBI human reference genome. We anchored 3.1 Mb (62%) of the YH and 2.9 Mb (61%) of the NA18507 novel sequence to the reference chromosomes (**Supplementary Table 2**). Among these, we found that about half (46% in YH and 43% in NA18507) were insertions in the individual genomes or deletions in the NCBI human reference. Only a small fraction (1.7% in YH and 1.4% in NA18507) appeared to be sequences that were highly divergent from sequences in the same location on the reference genome. About 878.9 kb of the YH and 833.3 kb of the NA18507 novel sequences mapped to gap regions in the

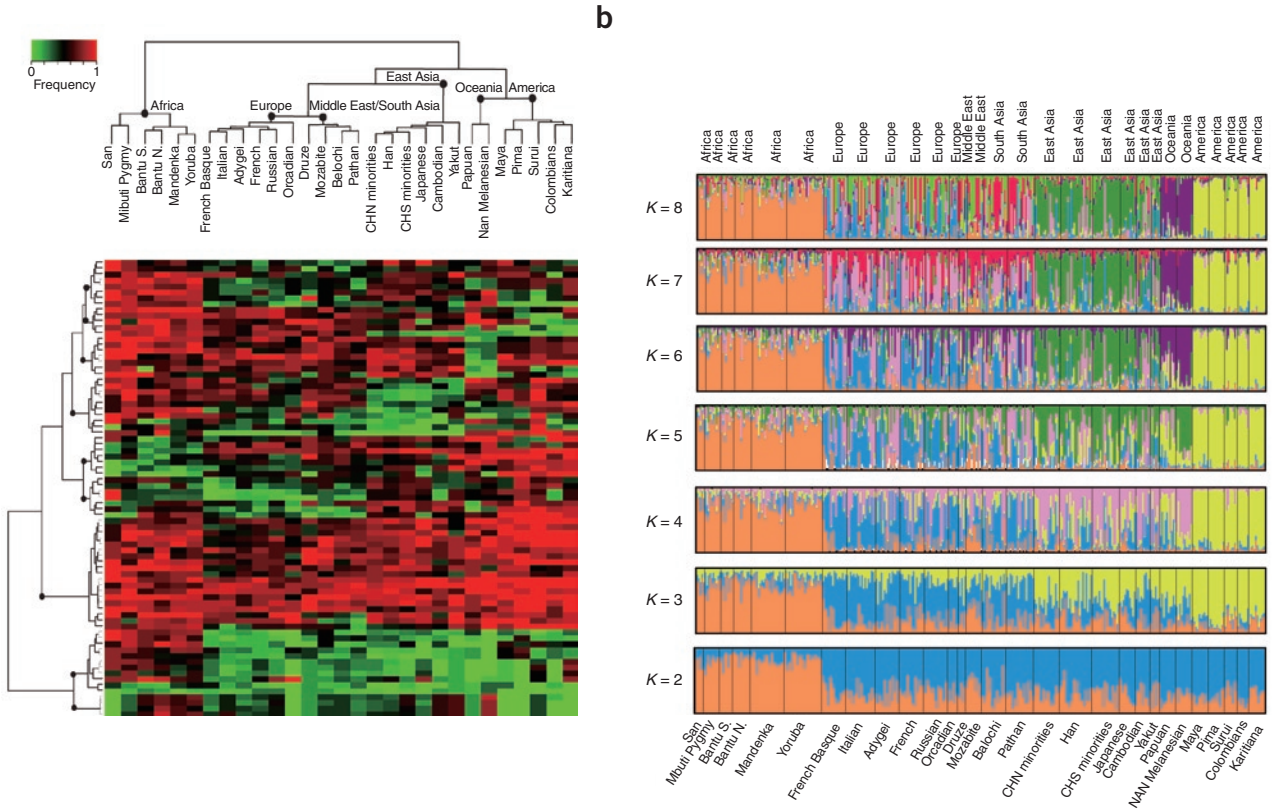


Figure 1 Population-specific patterns in novel sequences. (**a**) Frequency of individual-specific sequences (rows) in each population (columns) and neighbor-joining tree of the populations. PCR amplification was used to detect the presence or absence of each sequence in each individual. The novel sequence frequency in each population was calculated over multiple individuals belonging to the same population (on average ~8 individuals per population). For each sequence, the relative frequency in each population is represented by color intensity (red, higher frequency; green, lower frequency). Displayed here are the 83 novel sequences with <90% frequency variation over all samples. Groups of novel sequences that displayed different population-specific patterns are described in detail in **Supplementary Discussion**. (**b**) Population structure inferred by Bayesian clustering using novel-sequence frequency information. Each individual is shown as a thin vertical line, which is partitioned into *K* colored components representing estimated membership fractions in *K* genetic clusters. Population groups are separated by black lines. The population names are at the bottom of the figure and geographic locations are at the top.

© 2010 Nature America, Inc. All rights reserved.

reference chromosomes. The remaining anchored sequences were located in complex structural variant regions.

Population pattern of novel sequences

If these novel sequences constitute true variation in the sequence composition of the human genome, the expectation is that, as with SNPs and other types of sequence variation, these novel sequences will have population-specific characteristics. To determine the frequency variation of the novel sequences in humans in different worldwide populations, we randomly selected 164 novel sequences that did not overlap between YH and NA18507 (91 from YH and 73 from NA18507) (**Supplementary Table 3**). We used PCR to amplify these sequences in 351 individual samples from 41 worldwide populations of the HGDP-CEPH Human Genome Diversity cell line panel[18,19] (**Supplementary Tables 4** and **5**).

We then carried out phylogenic and genetic structural analyses using the novel sequences. For the profiled populations, we built a neighbor-joining tree with the novel sequence frequencies between populations as distance without prior information of individual origins (**Fig. 1a**). The tree topology generally agreed with previously defined population relationships[20–24]. Six groups that clustered by genetic analysis fit well with the main geographic boundaries of Africa, Europe, Middle East/South Asia, East Asia, Oceania and America, and structural analysis[25] also displayed consistent results (**Fig. 1b**). The novel sequences also showed distinct frequency clustering in the African, European, Middle Eastern/South Asian, East Asian, Oceanian and Native American geographic populations (**Fig. 1a**).

Phylogenic analysis of the maternally inherited mitochondria genome and paternally inherited Y chromosome have previously demonstrated the out-of-Africa migration of modern human populations[26]. Interestingly, we found that several novel sequences showed a variety of different patterns of frequency change along these defined human migration paths[27] (**Fig. 2**). For example, **Figure 2a** shows a novel sequence that has a very high frequency in the South African San population, which becomes lower as one moves to the North African population, and becomes still lower in more distant populations, until it is completely absent in the most geographically distant populations in the European (Russian), Oceanian (NAN Melanesian), and Native American (Surui) populations. Another novel sequence (**Fig. 2b**) has the opposite frequency changes, showing increasing frequency along the migration path. In addition, **Figure 2c** shows a novel sequence where the frequency becomes lower as populations follow the out-of-Africa path and has a substantial frequency reduction through the Middle Eastern, South Asian, European and Oceanian populations, but still maintains a high frequency in the East Asian and Native American populations. **Figure 2d** shows a fourth example where the novel sequence has a rapid frequency reduction in East Asian and Oceanian populations as compared to that for the European populations.

(See **Supplementary Discussion** for more information of population studies carried out on the novel sequences.)

Estimating individual sequence differences and human pan-genome size

Our analysis of novel sequence variation between populations was concordant with SNP frequency differences between populations. Similarly, we would expect to see a concordance between SNP frequencies and novel sequence frequencies between individuals. We therefore compared the SNP differences identified in previous studies[20–24] against a PCR-validated sampling of novel sequences (**Supplementary Fig. 4**) and indeed observed a positive correlation.

Given these findings, we set out to assess the DNA sequence composition differences between two individuals. We used a test based on standard SNP variation rates between two individuals, and defined sequence composition differences as 'yes or no' calls that a novel sequence or SNP was present or absent between the two individuals; rearrangement of homologous sequences or copy number changes of repeat sequences were not

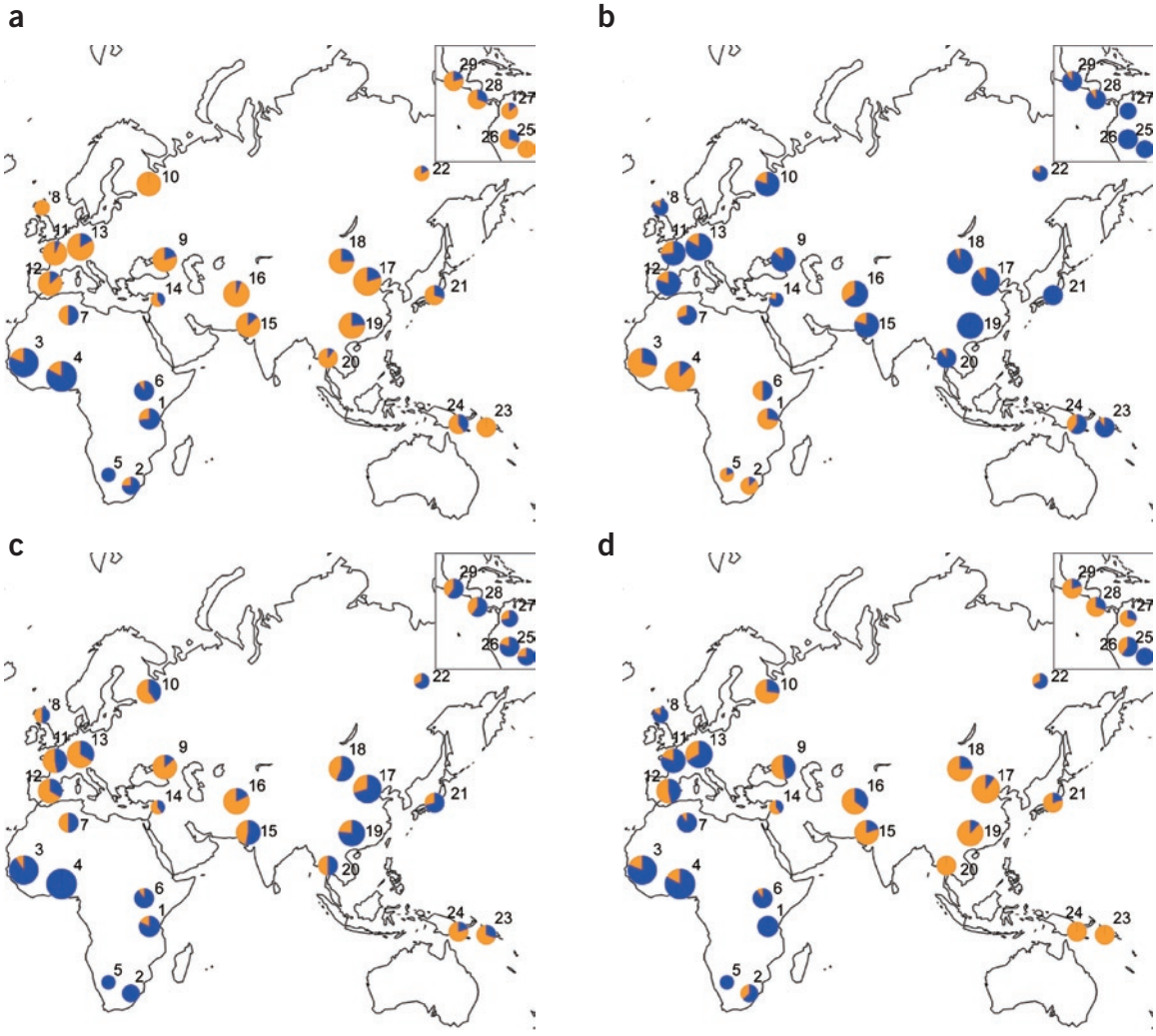


Figure 2 Examples of novel sequences with variant frequencies across populations. (**a**) NA18507, Scaffold_13185, shows a very high frequency in African populations and declines as populations grow more geographically distant. (**b**) YH, Scaffold_1781, shows a very low frequency in African populations and increases as populations grow more geographically distant. (**c**) YH, Scaffold_14717, shows a very low frequency in European populations. (**d**) NA18507, Scaffold_80603, shows a very low frequency in Asian populations. Each pie represents a single population; pie position on the map denotes the approximate geographical location of the population; pie size represents the number of DNA samples analyzed. Blue in each pie indicates the frequency of the novel sequence in the population. Key shows name of population corresponding to each number; number of samples for each population is given in parentheses.

© 2010 Nature America, Inc. All rights reserved.

included. Alignment of the YH and NA18507 genomes and comparison of the SNP data sets from these two genomes showed that the individual-specific sequence differences were at least 8 Mb in total and that the SNPs differed by 0.155% (**Supplementary Fig. 5**). The 8-Mb difference included 4 Mb SNPs and 4 Mb individual-specific sequences.

To assess variation between individuals that are more related based on the closeness of their populations, we compared the YH assembly to a preliminary assembly of the Korean (SJK[28]) genome with 28-fold coverage (data not shown). Our analysis revealed a 1.6-Mb individual-specific sequence difference and a 0.092% SNP difference between YH and SJK. The current sequence for the SJK genome was not sufficient for complete assembly, thus we used a calculation based on **Supplementary Figure 2** from which we could infer that the individual-specific sequence difference would cover ~1.8 Mb. These two analyses provide an individual-specific sequence increase from 1.8 Mb to 4 Mb and a SNP difference increase from 0.092% to 0.155%, indicating SNP rate and individual-specific sequence differences are positively correlated. We therefore estimated that the length of individual-specific sequences between a random pair of human individuals would range between 1.8 Mb and 4 Mb, and with the inclusion of the composition differences from SNPs, it would be in the range of 4.2 Mb to 8 Mb.

To estimate the size of the pan-genome, we used the above range for individual sequence differences and (given the correlation between the individual-specific sequences and SNP differences) used a transformation of Watterson's θ_w[27,29] (Online Methods), to evaluate the sequence differences in the population. From this we calculated a population mutation parameter for individual-specific sequences of 3.5e-4 to 7.4e-4 per base. Given a world population of over six billion people, we estimated that a complete human pan-genome would include an additional 19–40 Mb of novel sequences over the reference genome.

To assess the accuracy of this estimate, we carried out a preliminary assessment of the pan-genome size by sequentially adding Venter's HuRef[11], YH and NA18507 to the NCBI human reference genome. This preliminary pan-genome had a cumulative length that fell within our expected range (**Fig. 3**). We also estimated that common polymorphic, individual-specific sequences (those having >1% frequency in the human population) would be about 5–10 Mb in total length, and that these should be able to be defined after complete sequencing of about 100–150 individuals randomly selected from the world population.

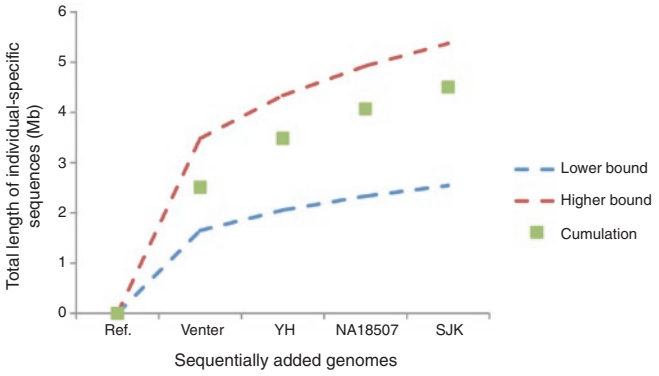


Figure 3 Cumulative length of individual-specific sequences resulting from sequentially adding genomes to the pan-genome. When adding a new genome, the novel sequences >100 bp and with <90% identity to previously added sequences were considered new individual-specific sequences and added to the data set. The real length of all novel sequences after adding each genome (green point) is in the range from the lower-bound (blue dashed line) and higher-bound (red dashed line) of the proposed individual-specific sequence population model.

Genes contained in novel sequences

To gain insight into the presence or absence of genes between the novel sequences and reference genomes, we aligned the 162 human NCBI RefSeq genes that could not be mapped to the NCBI reference genome onto the assembled YH and NA18507 novel sequences. We found that 72 and 69 of these genes could be fully or partially (>100 bp) found in the YH and NA18507 genomes, respectively, and that 55 of those individual genes overlapped between the two genomes (**Supplementary Tables 6** and **7**). Functional analysis showed that about a third of the RefSeq genes present in YH and NA18507 are members of highly variable gene families (such as mucin 2, major histocompatibility complex HLA-DQA1 and non-coding RNA SNORA66), whereas the majority of the remaining genes (57 (79%) of YH and 53 (77%) of NA18507) are currently considered hypothetical genes and have unknown functions.

Analysis of these novel sequences at the protein level by aligning the novel sequences to all human RefSeq proteins using tBlastN (E value = 1e-5) gave 1,151 and 1,087 hits in the YH and NA18507 novel sequences, respectively (**Supplementary Tables 8** and **9**). As with the RefSeq gene mapping results, the majority (915 (79%) in YH and 809 (74%) in NA18507) of the hits were to hypothetical proteins of unknown function, indicating that these genes have not been well studied. Among the hits that matched functionally classified proteins, the most abundant were members of the double homeobox protein (*DUX*) family (113 hits in YH and 58 hits in NA18507), which are known to be associated with heterochromatin[30] and also to include a number of pseudogenes[31]. Additional abundant protein categories were made up of gene families that are known to be quickly evolving and have many variant copies or may have copy number differences between genomes, which reflects the findings in the above gene analysis. These protein categories included mucin[32], zinc finger[33] and olfactory receptor proteins[34] (**Supplementary Fig. 6**).

To check whether the genes predicted by homology are likely to be functional, we investigated the conservation level of these genes across species. In total, 200 YH novel sequences (35 of which are predicted genes) and 155 NA18507 novel sequences (14 of which are predicted genes) have identified homologous regions present in all three of the chimpanzee, macaque and mouse genomes. Using 'intergenic' sequences of ~2–5 kb in length that are at least 5 kb distant from genes annotated by Ensembl as a neutral control and the well-annotated (with "NM-" prefix) RefSeq genes coding sequence present in the NCBI human reference genome as a positive control, we saw a bimodal distribution in the sequence identity of the homologous novel sequences (**Fig. 4**): the left peak of the distribution conformed to the neutral control and the right peak conformed to the positive control. The predicted coding sequences were clearly enriched at the high cross-species identity level (>90%), which is consistent with the sequence identity distribution of known annotated coding sequences. This strongly indicates that at least a portion of the homology-predicted genes might be functional and biologically important.

DISCUSSION

This study provided a genome-wide quantitative exploration of novel sequences in different individual genomes, and initiated an effort to construct the human pan-genome. We identified an extensive amount of novel sequences, which were found to be common variant sequences with different frequencies across populations. We also estimated the extent of sequence variation between two human individuals.

Cross-species conservation analysis revealed that some genes contained in these novel sequences are conserved among mammalian genomes, suggesting that these genes might be biologically functional and thus may be related to differences in gene networks between human individuals. Our

© 2010 Nature America, Inc. All rights reserved.

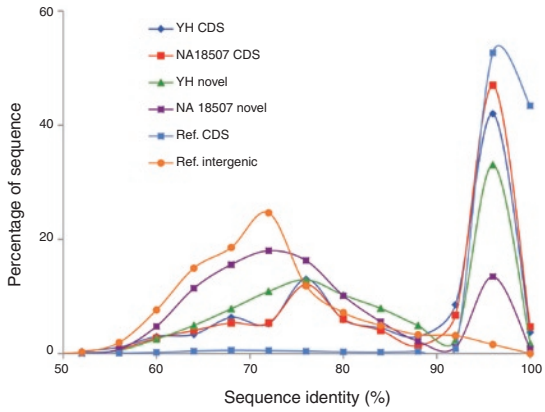


Figure 4 Distribution of sequence identity (in percentage) calculated from multiple alignments between human, chimpanzee, macaque and mouse genomes. YH novel sequences (green triangles) and NA18507 novel sequences (purple squares) had a bimodal distribution of sequence identity in the multiple alignments, whereas the distribution peak with between 90% and 100% identity is enriched in both YH novel coding sequence (dark blue diamonds) and NA18507 novel coding sequence (red squares). Coding sequences of the reference human genome (light blue squares) and the intergenic region (orange circles) were used as positive controls and neutral controls of conservation level, respectively. CDS, coding sequence.

finding that individual genomes contain a considerable amount of novel sequence indicates that similar analyses may be useful for medical genomics studies to augment array-based technologies that rely on the reference genome. Complete sequencing and assembly of personal genomes may allow larger numbers of various types of genetic variation to be identified that lead to more complete information about the genetic determinants of phenotype and disease.

Hence, it is important and practical to sequence and carry out *de novo* assembly on more human genomes to discover the common polymorphic sequences in the human population and to obtain a complete human pan-genome. Theoretically, our current pan-genome built from four individual genomes has already covered >90% of novel sequences that had a frequency >0.5 in the human population and about a half of those at a 0.1 frequency. With continuous innovation in sequencing technology, sequencing is becoming a practical and affordable method for analyzing a large number of complete human genomes, making it feasible to establish a more comprehensive understanding of the human genome, to make discoveries in medical genomics and to develop new applications for personalized medicine.

METHODS

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

Data access. DDBJ/EMBL/GenBank: ADDF000000000 (YH) and DAAB000000000 (NA18507). The versions described in this paper are the first versions, ADDF010000000 (YH) and DAAB010000000 (NA18507). NCBI: sequencing reads of YH genome, NCBI Short Read Archive SRA009271. The assembled genomes and all of the associated analyses are freely available at http://yh.genomics.org.cn.

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

ACKNOWLEDGMENTS

This project is supported by the Chinese Academy of Science (GJHZ0701-6), the National Natural Science Foundation of China (30725008; 30890032), Shenzhen local government, the Danish Platform for Integrative Biology, the Ole Rømer grant from the Danish Natural Science Research Council. L. Goodman edited the manuscript. J. Sun, M. Zhao, Y. Liu, Y. Zheng and H. Wang helped on designing the primers. W. Jin helped on experimental validation. San A, J. Wang, Y. Huang, M. Jian, M. Chen, Y. Huang, Xiaoli Ren, H. Liang, H. Zheng, S. Lin helped on the data production.

AUTHOR CONTRIBUTIONS

Ruiq. L., Y.L., Ha. Z. and Ruib. L. contributed equally to this work. H.Y., Ju. W. and Ji. W. managed the project. Ju. W., Ruiq. L., L.B. and Y.L. designed the analyses. Ju. W., Ruiq. L., Y.L., Ha. Z., Ruib. L., Ho. Z., Q.L., W.Q., G.Z., H.W., J.Q., X.J., D.L., Hon. C., S.L. and K.K. performed the data analyses. H.B. and How. C. contributed the DNA samples. Y.R., X.H. and Xu. Z. performed PCR validation. G.T., J. L., Xi. Z. performed sequencing. Ju. W., Ruiq. L., Y.L. and Ruib. L. wrote the paper.

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

1. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
2. Frazer, K.A., Murray, S.S., Schork, N.J. & Topol, E.J. Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10**, 241–251 (2009).
3. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
4. Hinds, D.A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).
5. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
6. Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**, 95–108 (2005).
7. Khaja, R. *et al.* Genome assembly comparison identifies structural variants in the human genome. *Nat. Genet.* **38**, 1413–1418 (2006).
8. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
9. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
10. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
11. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
12. Wheeler, D.A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
13. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
14. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
15. Li, Y. & Wang, J. Faster human genome sequencing. *Nat. Biotechnol.* **27**, 820–821 (2009).
16. Li, R. *et al.* De novo assembly of the human genomes with massively parallel short read sequencing. *Genome Res.* (in the press).
17. Bovee, D. *et al.* Closing gaps in the human genome with fosmid resources generated from multiple individuals. *Nat. Genet.* **40**, 96–101 (2008).
18. Cann, H.M. *et al.* A human genome diversity cell line panel. *Science* **296**, 261–262 (2002).
19. Cavalli-Sforza, L.L. The Human Genome Diversity Project: past, present and future. *Nat. Rev. Genet.* **6**, 333–340 (2005).
20. Tishkoff, S.A. *et al.* The genetic structure and history of Africans and African Americans. *Science* **324**, 1035–1044 (2009).
21. Wang, S. *et al.* Genetic variation and population structure in native Americans. *PLoS Genet.* **3**, e185 (2007).
22. Rosenberg, N.A. *et al.* Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
23. Li, J.Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).
24. Jakobsson, M. *et al.* Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**, 998–1003 (2008).
25. Falush, D., Stephens, M. & Pritchard, J.K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
26. Underhill, P.A. & Kivisild, T. Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu. Rev. Genet.* **41**, 539–564 (2007).
27. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22**, 231–238 (1999).
28. 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).
29. Wong, G.K. *et al.* A population threshold for functional polymorphisms. *Genome Res.* **13**, 1873–1879 (2003).
30. Beckers, M. *et al.* Active genes in junk DNA? Characterization of DUX genes embedded within 3.3 kb repeated elements. *Gene* **264**, 51–57 (2001).
31. Holland, P.W., Booth, H.A. & Bruford, E.A. Classification and nomenclature of all human homeobox genes. *BMC Biol.* **5**, 47 (2007).
32. Dekker, J., Rossen, J.W., Buller, H.A. & Einerhand, A.W. The MUC family: an obituary. *Trends Biochem. Sci.* **27**, 126–131 (2002).
33. Krishna, S.S., Majumdar, I. & Grishin, N.V. Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res.* **31**, 532–550 (2003).
34. Young, J.M. *et al.* Extensive copy-number variation of the human olfactory receptor gene family. *Am. J. Hum. Genet.* **83**, 228–242 (2008).

© 2010 Nature America, Inc. All rights reserved.

ONLINE METHODS

Data availability. The data described in this study are freely available in the YH genome database (http://yh.genomics.org.cn/download.jsp). The full data set includes: (i) previous and newly generated Illumina Genome Analyzer (GA) sequencing reads in FASTQ format; (ii) *de novo* genome assembly of both YH and NA18507 genome sequences (contigs, scaffolds); (iii) identified novel sequences in the two genomes with corresponding alignment information; (iv) PCR gel figures and validation results in HGDP-CEPH panels. Sequencing reads of YH genome are also available at the NCBI Short Read Archive (SRA009271).

Public data used. The human genome reference assembly (NCBI Build 36.3), HuRef assembly, RefSeq mRNA and protein sequences, EST sequences and core nucleotide database were downloaded from the NCBI database (http://www.ncbi.nlm.nih.gov). Protein sequences and annotations were downloaded from the UniProt database (http://www.uniprot.org/downloads). Read sequences of Watson's genome were provided by the Baylor College of Medicine. Read sequences of sample NA18507 were provided by Illumina, Inc., which is also publicly available in NCBI Short Read Archive (accession number SRA000271).

Data production. Library construction and read sequencing by Illumina Genome Analyzer II platform followed the manufacturer's instructions. Fluorescent images and base-calling were performed by Illumina data process pipeline (IlluminaPipeline-1.3.2).

Genome assembly. Whole genome short read *de novo* assemblies of the YH and NA18507 genomes were performed by SOAPdenovo[16] software (**Supplementary Data**, http://soap.genomics.org.cn), which is based on the De Brujin graph algorithm. The algorithm details and step-by-step YH and NA18507 assembly results have been described in the assembler manuscript[16]. Here is a brief summary of the algorithm.

First, sequencing errors that were primarily accumulated at the 3′-end of reads were corrected according to 17-mer frequency. In this step, all the raw read sequences from short insert-sized libraries (<500 bp paired-end insert size or single-ended) were broken down into 17 mers, and the frequency of each 17 mer was counted. Low frequency (<5 in this study) 17 mers that were likely to be sequencing errors were edited to the closely similar high-frequency 17 mers.

The sequencing-error corrected reads were then loaded into memory, and De Brujin graph data format was used to build the overlap graph, with 25-mers as vertex and read paths across the 25-mers as edges. Thus, two reads will be joined if they have at least a 25 bp overlap. The repeat sequences and sequencing errors would make the graph very complex. To reduce the complexity of the graph and filter noise connections, we removed the 'tips' which are short (<50 bp) and low-coverage dead ends in the graph, removed the low-coverage connections that nodes were linked by only one or a few reads in the graph, and merged the bubbles where redundant paths having the same input and output nodes while with minor differences (polymorphisms or difference between homologous sequence copies). After error correction, we broke the graph at repeat boundaries and outputted the unambiguously continuous sequence fragments as contigs.

Next, we realigned the short reads onto the contigs, and transferred the read paired-end information into contig linkage information. The unreliable linkages between two contigs that have equal or less than three read-pairs were filtered. The contig linkage graph was linearized by masking repeat contigs, which have multiple conflicted connections to the other contigs. And the remaining contigs with compatible connections to each other were constructed into scaffolds. The paired-end information was used step by step that started from short paired-ends to longer paired-ends.

The final step of *de novo* assembly is to close the gaps inside constructed scaffolds. We collected read pairs with one end located at the edge of contigs and another end located in the gaps, and performed local assembly to extend the contig sequence into the gaps. The final gap closed scaffolds were used for all analysis in the project.

Identification of novel sequences. We aligned all assembled contigs to the human reference genome (NCBI Build 36.3) using BLAT[35] with -fastmap option enabled. Alignment position of each contig indicates a candidate location of the scaffold to which the contig belongs. For contigs with multiple hits, the top ten hits with highest sequence identity and >90% coverage of the contig remained as candidate locations. Then we checked candidate locations of contigs within a scaffold to build scaffold-reference alignment to maintain orientations in as linear a fashion as possible between scaffolds and the NCBI reference genome. The alignment with the longest length in linear orientation between a scaffold and the reference was picked as 'best-hit' of the scaffold. We then aligned the scaffolds against the located regions on the NCBI reference genome by LASTZ.

The unmapped sequences derived from LASTZ alignment were treated as candidate novel sequences. We then aligned the scaffolds with unmapped sequences to the whole NCBI reference genome again using BLASTn[36]. The scaffold fragments with <90% identity to any region of the NCBI reference genome was defined as novel sequences. Novel sequences with <100 bp were filtered.

The identified novel sequences were first aligned to gap-closure fosmids[17] and the genomes of HuRef, Watson, YH (if novel sequences are from NA18507), and NA18507 (if novel sequences are from YH). Novel sequences that had alignments with >90% identity to any these genomes were kept apart, and the remaining novel sequences were then aligned to the other mammalian genomes. Hits with an E-value < 1e-20 were retained as valid alignment in classification. Novel sequences that aligned to human and mammalian genomes were retained, and those aligned to non-mammalian genomes were treated as potential contaminations and were filtered in this analysis. The remaining novel sequences that had no alignment to any sequences in GenBank database were classified as unknown.

Population profiling of novel sequences by PCR. To validate novel sequences identified in our assemblies and survey their frequency in human populations, we extracted novel sequences with lengths >500 bp and randomly selected 233 novel sequences for PCR validation in 347 human DNA samples from a worldwide population that were provided by HGDP-CEPH (**Supplementary Fig. 7**) and four HapMap CHB samples. The appropriate temperature of these PCR experiments was 58 ± 2 °C. The high-quality amplification of 164 novel sequences was used in this analysis.

Phylogeny tree construction. The frequency of novel sequences in each ethnic group was calculated. We then used the frequency information to cluster the ethnic groups and the novel sequences by hierarchical clustering in heatmap function implemented in R scripts. The distance between objects was measured by standard complete linkage clustering (farthest neighbor method) by comparing the frequency of novel sequences between two clusters.

Genetic structure. The genetic structures of world ethnic groups were calculated using STRUCTURE[37] (version 2.2) by *K*-means partitional clustering. The monoploid model was used to adapt for novel sequence present/absent information. The PCR validation results in 288 individuals in this study were transformed to a 0/1 matrix as the data input of STRUCTURE, where 0 denoted for absence of the novel sequence and 1 for presence. STRUCTURE calculated membership coefficients to place all the individuals to *K* clusters, where *K* value was set from 2 to 8 in our study.

Evaluate the sequence differences in the population. To estimate the pan-genome size, we used the range of individual sequence difference and transformed Watterson's θ, which was primarily designed for SNP divergence, to suit this estimation. First, we estimated the average composition difference in DNA sequence between two individuals was 1.8Mb to 4Mb, where the two boundaries were defined by SJK-YH (both are Asians) and YH-NA18507 differences. Second, by definition of Watterson's θ, the total amount of identified individual-specific sequences would approximately conform the following formula:

$$K = \theta \times L \times a, a = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n}$$

Where n is the sample size (1 for a haploid, 2 for a diploid), L is the size of a single human genome and θ is the averaged individual specific sequence rate among all samples. We therefore estimate the range of $(\theta \times L)$ to be ~0.9–1.9 Mb. Third, by extrapolating the above results to the whole human population with a size of ~6.5 billion (a is calculated to be about 23.9), we estimated the total amount of human individual-specific sequences to be ~19–40 Mb.

35. Kent, W.J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
36. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
37. Falush, D., Stephens, M. & Pritchard, J.K. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol. Ecol. Notes* **7**, 574–578 (2007).

© 2010 Nature America, Inc. All rights reserved.

nature biotechnology

ARTICLES

Design of phosphodiesterase 4D (PDE4D) allosteric modulators for enhancing cognition with improved safety

Alex B Burgin[1,4], Olafur T Magnusson[2,4], Jasbir Singh[3], Pam Witte[1], Bart L Staker[1], Jon M Bjornsson[2], Margret Thorsteinsdottir[2], Sigrun Hrafnsdottir[2], Timothy Hagen[3], Alex S Kiselyov[3], Lance J Stewart[1] & Mark E Gurney[2]

© 2010 Nature America, Inc. All rights reserved.

Phosphodiesterase 4 (PDE4), the primary cAMP-hydrolyzing enzyme in cells, is a promising drug target for a wide range of conditions. Here we present seven co-crystal structures of PDE4 and bound inhibitors that show the regulatory domain closed across the active site, thereby revealing the structural basis of PDE4 regulation. This structural insight, together with supporting mutagenesis and kinetic studies, allowed us to design small-molecule allosteric modulators of PDE4D that do not completely inhibit enzymatic activity (I_{max} ~ 80–90%). These allosteric modulators have reduced potential to cause emesis, a dose-limiting side effect of existing active site–directed PDE4 inhibitors, while maintaining biological activity in cellular and *in vivo* models. Our results may facilitate the design of CNS therapeutics modulating cAMP signaling for the treatment of Alzheimer's disease, Huntington's disease, schizophrenia and depression, where brain distribution is desired for therapeutic benefit.

Spatial and temporal signaling by the second messenger cAMP is highly regulated by signal transduction proteins that manage local synthesis and degradation of this cyclic nucleotide. A single superfamily of cyclic nucleotide phosphodiesterases (PDEs) is responsible for the hydrolysis of cAMP and cGMP. The importance of these enzymes is emphasized by the high degree of sequence conservation among the catalytic domains of all eleven superfamily members (PDE1–11)[1], with a single active-site amino acid determining nucleotide selectivity[2]. PDE families are distinguished by N-terminal motifs encoding unique regulatory domains, although the structural basis for PDE regulation is not understood for any PDE family member.

PDE4 is the primary cAMP-specific hydrolase and is represented by four genes (*PDE4A*, *B*, *C* and *D*). All four proteins contain signature regulatory domains called upstream conserved region 1 (UCR1; ~55 amino acids) and upstream conserved region 2 (UCR2; ~78 amino acids)[3]. UCR2 negatively regulates cAMP hydrolysis and also is needed for high-affinity binding of the PDE4 inhibitor rolipram[4,5]. UCR1 and UCR2 form a regulatory module that is disrupted by protein kinase A (PKA) phosphorylation of UCR1[6]. Phosphorylation increases enzymatic activity and sensitivity to rolipram inhibition[7,8]. PDE4 isoforms also differ in N-terminal sequences that determine cellular sublocalization[9] or assembly into multi-component protein complexes[10–12].

PDE4 is a therapeutic target of high interest for central nervous system (CNS), inflammatory and respiratory diseases[13–16]; however, no PDE4 inhibitors have yet been brought to market because of issues related to tolerability, such as emesis and diarrhea[17,18]. The PDE4 isoforms whose inhibition causes emesis have not yet been determined as existing active site–directed PDE4 inhibitors do not show any isoform selectivity. Gene deletion studies in mice have implicated the PDE4D isoform[19], but because rodents are unable to vomit, confirmation is needed in non-rodent species. The emetic response to PDE4 inhibitors is mediated in part by a brainstem noradrenergic pathway[20] and can be reduced by limiting distribution to the brain[21]. However, when distribution to brain is desired for therapeutic benefit, as in CNS indications such as impaired cognition, schizophrenia and depression, a different strategy is required.

The PDE4 inhibitors that have been explored in human clinical trials bind the active site competitively with cAMP and therefore completely inhibit enzyme activity at high concentrations. Although this traditional approach to PDE4 inhibitor design has demonstrated therapeutic benefit[18,22], competitive inhibitors are likely to alter cAMP concentrations beyond normal physiological levels, perturbing the tight temporal and spatial control of cAMP signaling within cells and leading to side effects. We reasoned that an understanding of the structural basis of PDE4 regulation may allow the development of an alternative strategy for targeting PDE4. Here we present crystal structures of the PDE4 UCR2 regulatory domain in contact with small molecules and the catalytic domains of both PDE4D and PDE4B. The structures show that phosphodiesterase activity is regulated by controlling access to the active site. The insight provided by the structures and by supporting mutational data allowed us to design PDE4 allosteric modulators that only partially inhibit cAMP hydrolysis. Such compounds are more likely to lower the magnitude of PDE4 inhibition and maintain cAMP signaling, thereby reducing target-based toxicity. A similar approach has been described previously for allosteric modulators of G protein–coupled receptors[23] and for atypical retinoid ligands of nuclear hormone receptors[24]. Analysis

[1]deCODE biostructures, Bainbridge Island, Washington, USA. [2]deCODE genetics, Reykjavik, Iceland. [3]deCODE chemistry, Woodridge, Illinois, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to A.B.B. (aburgin@embios.com) or M.E.G. (mgurney@decode.com).

Received 7 July; accepted 4 December; published online 27 December 2009; doi:10.1038/nbt.1598

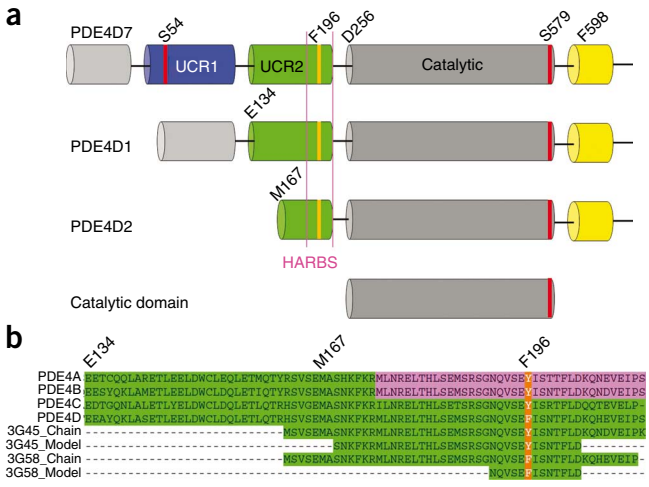


Figure 1 PDE4D Constructs. (**a**) Each of the four PDE4 genes generates multiple distinct isoforms that have unique N-terminal regions involved in targeting (gray). Representative long (PDE4D7), short (PDE4D1) and supershort (PDE4D2) isoforms are illustrated. Sites of PKA phosphorylation (S54) and ERK phosphorylation (S579) are indicated with red lines. All residues in PDE4D are based on the numbering of the reference PDE4D3 isoform, GenBank accession number AAA97892. The UCR1 and UCR2 domains are highlighted in blue and green, respectively. The C-terminal domain is highlighted in yellow. (**b**) Alignment of the UCR2 domain from PDE4A, PDE4B, PDE4C and PDE4D is shown. Regions within UCR2 shown to be important for high-affinity rolipram binding (HARBS) to PDE4A[4] and PDE4B[5] are highlighted in pink and delineated by pink line in **a**. The F196 (PDE4D)-Y274 (PDE4B) polymorphism is highlighted in orange. The amino acids present in the protein constructs used for crystallization ("Chain") and those amino acids modeled from visible electron density ("Model") are also aligned.

of our crystal structures suggests that PDE4 allosteric modulators may interact selectively with particular conformations of the enzyme[25] or with PDE4 in complex with accessory proteins[14]. We show that PDE4D allosteric modulators are potent in cellular and *in vivo* assays and have greatly reduced potential for emesis.

RESULTS

PDE4–UCR2 regulatory domain crystal structures

Each PDE4 gene has multiple transcription units resulting in multiple splice isoforms, which can be categorized into three basic isoforms[26] (**Fig. 1a**). Long isoforms have UCR1 and UCR2; short isoforms lack UCR1; and super-short isoforms lack UCR1 and have a truncated UCR2 domain. Previous studies have suggested that some PDE4 inhibitors preferentially bind the long forms of the enzyme, suggesting they may interact with UCR1 and/or UCR2. For example, it has been suggested that the PDE inhibitor RS25344 is more potent in cell extracts containing long forms of PDE4D, such as PDE4D7, than in those containing only the PDE4D catalytic domain[27]. In addition, various studies have described high-affinity and low-affinity 3H-rolipram binding sites[28], and mutational studies of PDE4A and PDE4B have mapped the high-affinity binding site to UCR2[4,5] (**Fig. 1b**). We confirmed these results for PDE4D by comparing the activity of rolipram, RS25344 and the closely related analog PMNPQ[29] to the well-characterized competitive inhibitor roflumilast[30] using a real-time, coupled enzyme assay that measures initial rates of cAMP hydrolysis. RS25344 and PMNPQ are >10,000× more potent, and *(R)*-rolipram is >50× more potent against PDE4D7 than against the PDE4D catalytic domain (**Supplementary Fig. 1**). There is no significant difference in affinity for RS25344 and PMNPQ between a long isoform and a supershort isoform (**Supplementary Fig. 2**).

The simplest explanation for these results is that the same region of UCR2 contacts both the inhibitor and the active site. With this assumption, we used Gene Composer software[31,32] to design a series of N-terminal PDE4D truncations (**Fig. 1b**) and used these proteins in crystallization trials in the presence of RS25344, PMNPQ and rolipram (**Supplementary Notes**). We screened for ligand-dependent crystallization conditions, obtained crystals and solved a structure of an N-terminally truncated PDE4D construct (N terminus at residue 163) in complex with RS25344 (**Fig. 2a**,**b**). Although the first 27 residues are disordered and not visible in electron density maps, a well-defined helix spanning residues Asn191 to Asp203 was observed over the active site. Close examination shows multiple interactions that would stabilize this structure and would explain why RS25344 selectively inhibits the biological isoforms. Notably, a phenylalanine residue (Phe196) extends from the UCR2 helix into the active site,

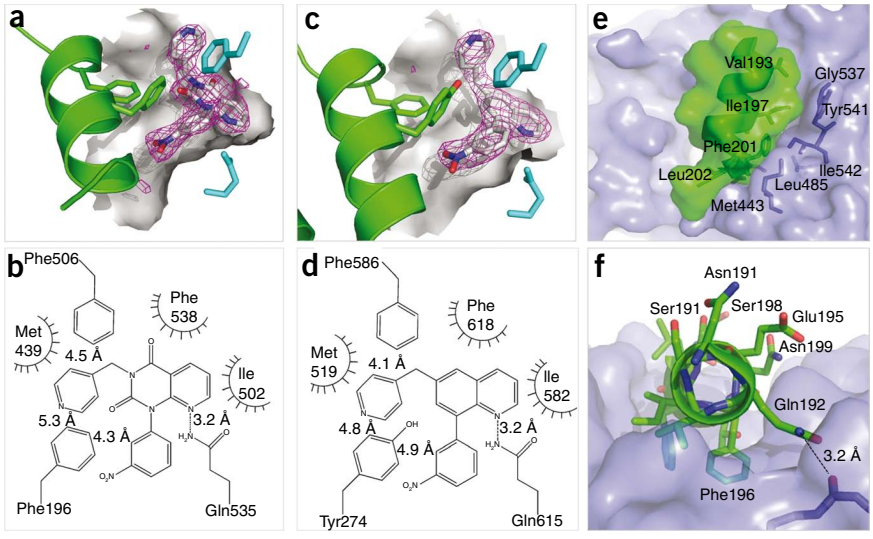


Figure 2 Capping the active site by UCR2. (**a**) A surface representation of the PDE4D catalytic domain (gray) bound with RS25344 (F_o-F_c omit maps in magenta) interacting with UCR2 (green) (PDB ID: 3G4G). (**b**) Schematic representation of RS25344 interacting with catalytic domain residues (Met439, Phe506, Phe538, Ile502, Gln535) and UCR2 (Phe196). (**c**) A surface representation of the PDE4B catalytic domain (gray) bound with PMNPQ (F_o-F_c omit maps in magenta) interacting with UCR2 (green) (PDB ID: 3G45). (**d**) Schematic representation of PMNPQ interacting with catalytic domain residues (Met519, Phe586, Phe618, Ile582, Gln615) and UCR2 (Tyr274). (**e**) Key hydrophobic interactions between the UCR2 domain (PDE4D(PDB ID: 3G4G)) and the catalytic domain. The UCR2 helix is surface rendered, and key residues are highlighted in green. The catalytic domain is surface rendered, and key residues are highlighted in blue. (**f**) The UCR2 helix is shown without surface rendering. Hydrophilic residues oriented toward solvent are labeled. RS25344 is not illustrated in order to simplify visualization of the interactions between UCR2 and the catalytic domain.

© 2010 Nature America, Inc. All rights reserved.

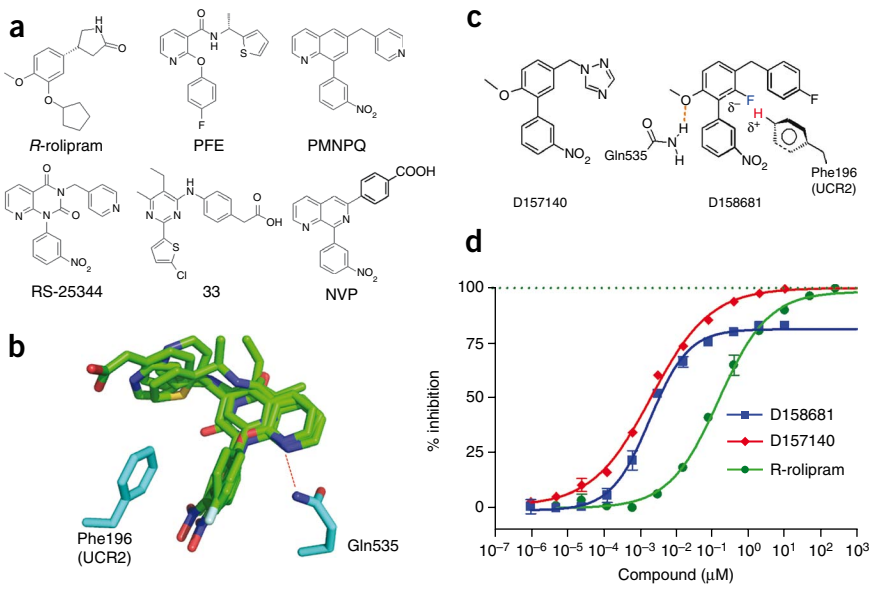


Figure 3 A common pharmacophore for PDE4 inhibitors accessing the UCR2 binding pose. (**a**) Literature compounds shown in our studies to bind UCR2 or the C-terminal helix (PFE[39]; PMNPQ and RS25344 (ref. 27); 33 (ref. 53); NVP[54]. (**b**) Superposition of literature compounds (green) based on our co-crystal structures of rolipram (PDB ID: 3G4K), PMNPQ (PDB ID: 3G58) or RS25344 (PDB ID: 3G4I) bound to the catalytic domain of PDE4D illustrate a common pharmacophore with two aromatic arms clamping Phe196 of UCR2 (cyan) and the compound forming a hydrogen bond (dashed red line) with Gln535 (cyan). (**c**) Chemical structures of a PDE4 full inhibitor (D157140) and a PDE4 allosteric modulator (D158681); also illustrated is the schematic interaction of D158681with PDE4D showing a hydrogen bond acceptor interaction with Gln535, and Ar_1 (4-F phenyl) working in cooperation with Ar_2 (3-NO2 phenyl) and the fluoro phenyl core to clamp onto Phe196 (UCR2). (**d**) *In vitro* inhibition of PDE4D7 by D15740 and *(R)*-rolipram demonstrating full inhibition of cAMP hydrolysis. D158681 displays partial inhibition kinetic behavior. Error bars are shown but typically are smaller than the size of the symbol (mean ± s.d., $n = 3$).

forming a network of interactions with the small-molecule inhibitor (**Fig. 2a**,**b**). We also observed a C-terminal helix in PDE4D in co-crystallization trials with PMNPQ, which overlays the active site similarly to UCR2 (**Supplementary Fig. 3** online). This helix was clearly defined in only one of the four monomers in the asymmetric unit, suggesting only a weak interaction with PMNPQ. Presumably more potent compounds that exploit this binding pose could be designed.

Close examination of the UCR2 helix also shows that there is good shape complementarity with the catalytic domain groove and that multiple interactions could stabilize the interaction of UCR2 and the catalytic domain in the absence of RS25344 (**Fig. 2e**). For example, Gln192 is positioned to make a hydrogen bond to the main chain carbonyl of Asn528 (**Fig. 2f**). More importantly, Phe201 from UCR2 fits into a hydrophobic cleft made by Ile542, Met439 and Leu485; and residues Val193, Ile197 and Leu202 form one face of the UCR2 helix that can interact with residues Gly537, Tyr541, Ile542 and Met443 across a hydrophobic surface of the catalytic domain (**Fig. 2e**). As expected, hydrophilic residues across UCR2 are oriented toward solvent, with Glu195 and Asn199 forming one face and Asn191, Ser194 and Ser198 forming a second solvent-accessible face (**Fig. 2f**). According to this model, Phe196 and Phe201 both play central roles in stabilizing the structure by allowing the UCR2 helix to contact the ligand (via Phe196) or to contact the catalytic domain directly (via Phe201); as expected, mutations at these positions substantially reduce affinity for RS25344 and PMNPQ (**Supplementary Table 1**). It is also important to note that antibodies targeting the observed UCR2 helix (K116 developed against peptide Val193 to Thr213) activate wild-type PDE4D3 (a short form of PDE4D) *in vitro*[33], indicating that the observed UCR2 helix also inhibits PDE4D activity in the absence of inhibitors.

If the observed UCR2 structure is important for the regulation of PDE4D, a similar structure is expected in PDE4B. We therefore engineered similar truncated constructs in PDE4B and obtained a PMNPQ-dependent PDE4B crystal structure (**Fig. 2c**,**d**). The resulting structure showed nearly identical interactions of PMNPQ with the catalytic domain, except that a homologous UCR2 tyrosine residue (Tyr274 in PDE4B homologous to Phe196 in PDE4D) was observed to stack on the nitro-substituted aromatic group of PMNPQ (4.9 Å). The tyrosine-phenylalanine difference results in a slightly weaker interaction between Tyr274 and Phe586, and would eliminate the edge-to-face π-π interaction between Tyr274 and the aromatic pyridine group of PMNPQ. These results predict that both RS25344 and PMNPQ would preferentially inhibit PDE4D over PDE4B. We confirmed this hypothesis by demonstrating that RS25344 and PMNPQ are ~10 times more potent against PDE4D7 than PDE4B1, equivalent long forms of PDE4 containing UCR1 and UCR2 (**Supplementary Fig. 4**), and that the Phe196/Tyr274 sequence variance is responsible for this selectivity difference (**Supplementary Fig. 5**). In addition, previous reports have shown that PDE4B and PDE4D have different kinetic properties despite the fact that their catalytic domains are highly conserved[34]. We observed very similar differences, with the apparent K_M for the PDE4B enzyme being significantly higher than for the PDE4D enzyme (7.70 ± 0.57 μM versus 1.50 ± 0.14 μM, $P < 0.0001$) (**Supplementary Fig. 6**). Notably, modeling of the PDE4D-UCR2 and PDE4B-UCR2 structures with catalytic domain structures containing 5′AMP[35] suggests that Tyr274 in PDE4B could hydrogen bond to the 2′OH of AMP or cAMP (**Supplementary Fig. 7**). To test whether this interaction was the source of the kinetic difference, we introduced the Phe196Tyr mutation into PDE4D. The apparent K_M value for this point mutant (7.6 μM) was equivalent to PDE4B (7.7 μM), demonstrating that, even in the absence of inhibitors, the UCR2 domain can affect the catalytic properties of PDE4D (**Supplementary Fig. 6**).

Design of PDE4 allosteric modulators

The RS25344 and PMNPQ co-crystal structures revealed a previously unknown binding mode for PDE4 inhibitors involving UCR2. Even though they are not traditional active site–directed inhibitors, both compounds are highly emetic in ferrets and other pre-clinical models[20,36]. Nonetheless, we explored the structure-activity relationship (SAR) of compounds interacting with UCR2 to understand the potential pharmacology of this binding mode. Early in our SAR studies, we discovered compounds that did not fully inhibit PDE activity (**Fig. 3** and **Supplementary Fig. 8**), offering the possibility that such partial inhibitors might have a natural 'safety valve' that could translate into improved tolerability, particularly with regard to emesis. Thus, we focused our SAR studies on understanding the chemical requirements for the design of partial versus full inhibitors of PDE4.

© 2010 Nature America, Inc. All rights reserved.

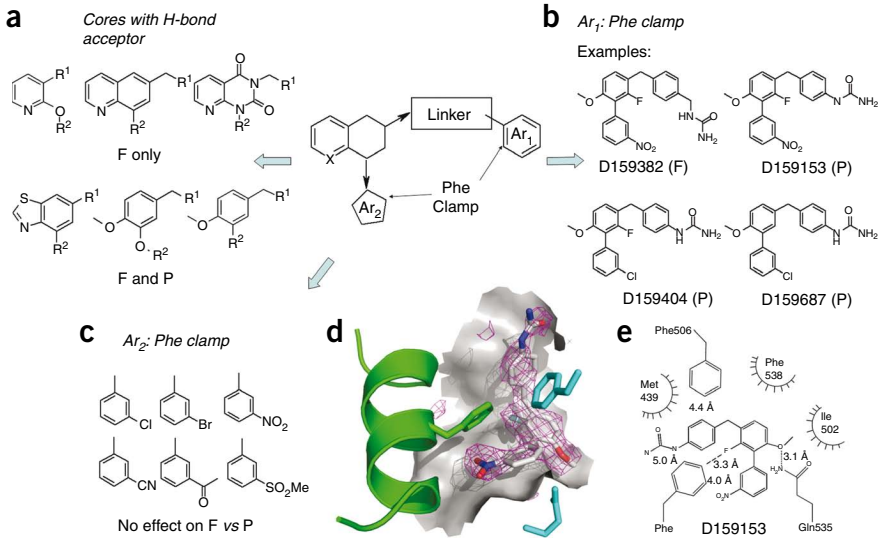


Figure 4 Critical pharmacophore elements that determine partial kinetic behavior of PDE4 allosteric modulators. Key elements of the pharmacophore include a planar scaffold providing a hydrogen bond acceptor, a linker and two aromatic substituents that create a clamp to hold UCR2 in the closed conformation across the active site. F and P signify compounds with "full" or "partial" inhibition kinetic behavior. (**a**) Scaffolds providing a hydrogen-bond acceptor to Gln535. (**b**) Aromatic Ar_1 substituents providing a part of the UCR2 clamp. (**c**) Aromatic Ar_2 substituents providing a part of the UCR2 clamp. (**d**) Co-crystal structure of a representative methoxyphenyl allosteric modulator (D159153) bound to PDE4D showing the UCR2 helix in the closed conformation. The regulatory helix is shown as a ribbon (green), the Fo-Fc omit map for the ligand is highlighted in magenta, and the active site surface is rendered in gray with key residues colored cyan. (**e**) Binding mode of D159153 to PDE4D indicating critical interactions.

Compounds able to bind UCR2 share a common pharmacophore and binding pose (**Fig. 3a**,**b**). The pharmacophore consists of four elements: a planar scaffold providing a hydrogen bond to Gln535, a linker and two aromatic substituents which create a clamp that holds UCR2 in the closed conformation (**Fig. 4**). Based on this common pharmacophore, we explored various heteroaromatic and aromatic cores featuring H-bond acceptors such as the pyridine- or quinoline-based scaffolds reported in the literature (**Fig. 4a**). Potent compounds were obtained, but all were full inhibitors of the enzyme. However, synthetic chemistry studies centered on benzothiazole, catechol and biaryl chemotypes (**Fig. 4a**) afforded compounds with both full and partial kinetic behavior (**Fig. 3c**,**d**). We hypothesized that the partial kinetic behavior might be due to weakening the interaction with the Q switch and P clamp elements of the binding pocket[37]. For example, a central methoxyphenyl core features an sp3 oxygen instead of a heterocyclic sp2 nitrogen as the hydrogen bond acceptor to Gln535 and therefore has both reduced directionality and greater flexibility. In addition, the monocyclic methoxyphenyl core decreases hydrophobicity and π-π stacking with Phe538 and Ile502 compared with RS25344 or PMNPQ. Allosteric modulators with partial inhibition behavior fully displaced 3H-rolipram from the high-affinity binding site on UCR2 and were >10,000-fold less potent against the truncated PDE4D catalytic domain lacking UCR2 (**Supplementary Fig. 9**).

In the next series of experiments, we investigated the effect of Ar_1 and Ar_2 substituents on full versus partial inhibition. We focused our SAR studies on elaboration of the biaryl derivatives and discovered that subtle changes to the Ar_1 functionality affected full versus partial kinetic behavior. For example, D159382, featuring a benzylurea substituent for Ar_1, was found to be a full inhibitor (**Fig. 4b**), whereas in striking contrast, the respective phenylurea derivative D159153 lacking the CH_2 link displayed partial inhibition behavior. A variety of substituents were tolerated at Ar_2 (**Fig. 4c**). The most active compounds contained relatively small *meta*-substituents, which improved potency but did not affect full versus partial kinetic behavior. A *meta*-chloro substituent was preferred to limit potential for genotoxicity. Over the course of the medicinal chemistry effort, we synthesized 805 compounds, of which 140 were allosteric modulators with partial inhibition kinetics. The SAR studies were supported by an extensive structural biology effort. We obtained 21 co-crystal structures of compounds bound to forms of PDE4D or PDE4B containing UCR2 (**Fig. 4d**,**e**). Full and partial inhibitors close UCR2 across the PDE4 active site in the same spatial orientation (r.m.s. deviation = 0.34 Å), so differences in the positioning of UCR2 do not explain the differences in kinetic behavior (data not shown).

To optimize PDE4D selectivity, we explored fluoro-substituted derivatives that could provide favorable electrostatic interactions with the partially positively charged edge of Phe196 in PDE4D. We found that this substitution pattern enhances selectivity for PDE4D because the same fluoro group introduces electrostatic and steric repulsion with Tyr274 of PDE4B (**Fig. 3c**). This allowed us to design compounds that were 60–100 times more selective for PDE4D than for PDE4B (**Supplementary Table 2**). PDE4A and C also contain a tyrosine at the key position in UCR2, with the consequence that they behave similarly to PDE4B in terms of selectivity (**Supplementary Table 3**). As UCR2 is unique to PDE4, optimized PDE4D allosteric modulators such as D159687 were >1,000 time more selective against PDE4 compared with other PDEs (**Supplementary Table 3**).

Our kinetic and biophysical data suggest that PDE4 behaves as a dimer with negative cooperativity between two binding sites (**Supplementary Fig. 10**). In the absence of modulator, both active sites are equivalent and the enzyme obeys simple Michaelis-Menten kinetics with respect to cAMP substrate. To explain partial inhibition of PDE4 in the presence of modulator, we propose that only one UCR2 domain can be placed in the closed conformation, resulting in the formation of an asymmetric PDE4 dimer. It was previously found that the stoichiometry of 3H-(*R*)-rolipram equilibrium binding to PDE4B2 is ~0.5 (ref. 5), which is consistent with our model of an asymmetric PDE4 dimer in the closed conformation. Modulator binding at one site decreases the turnover rate at the second active site, and this explains why the maximum inhibition of the enzyme is >50%.

Activity of PDE4 allosteric modulators in cellular assays

The effect of PDE4 allosteric modulators on cAMP hydrolysis was examined in human HEK293 cells[38,39]. PDE4D allosteric modulators are about 15× less active in the HEK293 cAMP assay than expected based on their biochemical median inhibitory concentration (IC_{50}) (**Supplementary Table 2**). We therefore wondered whether allosteric modulators and full inhibitors would behave differently in a biological context. Multiple compounds were profiled in a Sephadex-stimulated human whole blood assay of leukotriene E4 (LTE4) production by eosinophils[39,40] (**Supplementary Fig. 11**). PDE4D allosteric modulators were more potent in the human whole blood LTE4 assay than

© 2010 Nature America, Inc. All rights reserved.

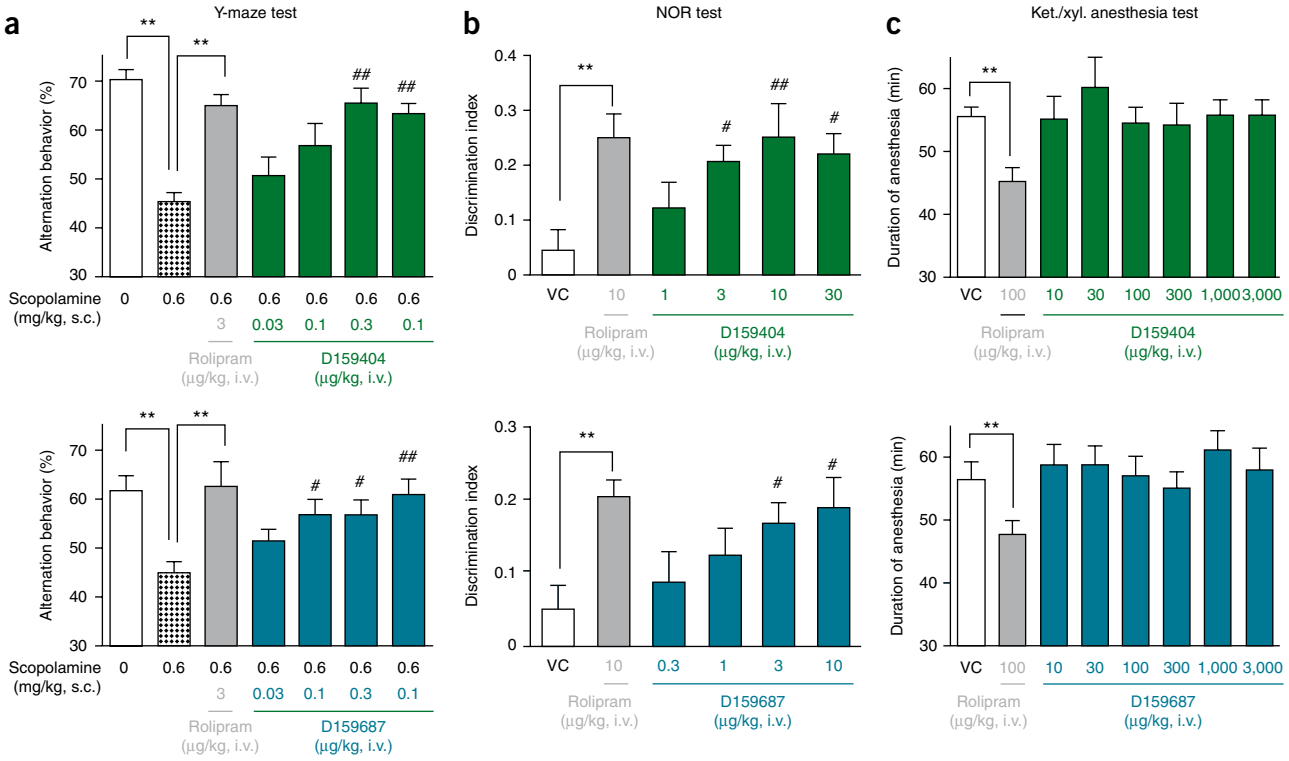


Figure 5 Effects of compounds on mouse models of cognition and a behavioral correlate of emesis. Effects on ddY strain mice of selective PDE4D allosteric modulators in the scopolamine-impaired Y-maze test (Y-maze, **a**), the novel object recognition test (NOR, **b**) with dosing 3 h after T1, and the ketamine/xylazine anesthesia duration test (Ket/xyl, **c**). (**a**) PDE4D allosteric modulators (D159404 and D159687) reversed the cholinergic deficit in the scopolamine-impaired Y-maze test to the same extent as rolipram. Each column represents mean ± s.e.m. (n = 9–10); $^{**}P < 0.01$ (Wilcoxon rank sum test); $\#P < 0.05$, $\#\#P < 0.01$ versus scopolamine-treated group (Dunnett's multiple comparison test). (**b**) PDE4D allosteric modulators improve object discrimination in the NOR test to a similar extent as rolipram. Columns represent mean ± s.e.m. (n = 10–13) for the discrimination index (DI) at the 24 h retention test (T2); $^{**}P < 0.01$ (Student's t-test); $\#P < 0.05$; $\#\#P < 0.01$ versus vehicle-treated group (Dunnett's multiple comparison test). (**c**) PDE4D allosteric modulators do not reduce the duration of ketamine/xylazine-induced anesthesia at doses 1,000× greater than their MED for pro-cognitive benefit in the NOR test (the least-sensitive cognitive test). Each column represents mean ± s.e.m. (n = 12–20); $^{**}P < 0.01$ (Student's t-test).

© 2010 Nature America, Inc. All rights reserved.

in the HEK293 cAMP assay and equally effective as roflumilast (**Supplementary Table 2**). This result demonstrates that allosteric modulators of PDE4 can provide complete inhibition of a biological response, even though they have partial enzyme inhibition kinetics *in vitro*. We hypothesize that the concentration of cAMP within cells is likely maintained within a narrow window such that a maximal effect on signaling does not require complete inhibition of PDE4.

Efficacy of PDE4 allosteric modulators in tests of cognition

We next profiled two PDE4D full inhibitors (D157140 & D159382) and four PDE4D allosteric modulators (D158681, D159153, D159404 and D159687) against rolipram as a reference compound in rodent cognition assays (**Supplementary Table 2**). Rolipram has been shown to have benefit in numerous rodent models of cognition including models of cholinergic deficit and memory consolidation[13]. The selected compounds all distribute into mouse brain after intravenous dosing (**Supplementary Table 2** and **Supplementary Fig. 12**).

Cholinergic deficit was modeled in the mouse using the scopolamine-impaired Y-maze test (**Fig. 5a**). Scopolamine induces a cholinergic deficit resulting in an 'amnesic' state owing to reduced cholinergic signaling through muscarinic receptors that are positively coupled to adenylate cyclase. The 8-min test measures the performance of immediate spatial working memory by monitoring the spontaneous preference of rodents to novelty. Alternating exploration of the three arms of a Y-maze by animals is assessed (continuous spontaneous alteration)[41]. An unimpaired score for alternation behavior is usually ~70%, and a minimum dose of scopolamine was used to reduce the impaired score to chance (~50%) levels. All compounds showed a dose response by intravenous administration and a maximum cognitive benefit similar to rolipram. In fact, the allosteric modulators were reproducibly slightly more potent (**Supplementary Table 4**). D159404 and D159687 were also evaluated by oral dosing (**Supplementary Fig. 13**); the minimum effective dose (MED) was 10 µg/kg for both compounds, consistent with their bioavailability in mouse (F = 28% for D159404 and 32% for D159687).

The effect of PDE4 allosteric modulators on long-term memory formation was assessed using the novel object recognition (NOR) test (**Fig. 5b**). This model uses the spontaneous preference of rodents to novelty in an environment and measures their ability to recognize an object previously seen (episodic memory). The level of discrimination between the novel and the familiar object progressively decreases with increasing time. By delaying administration of the PDE4 modulator, we assessed the effect of the compounds on memory processes dependent upon CREB phosphorylation by cAMP-dependent PKA[42]. Under this paradigm of delayed administration, rolipram provides cognitive benefit as shown by an improvement in the discrimination index measured at an inter-trial interval of 24 h, but there was no cognitive benefit with delayed administration of an anti-cholinesterase such as donepezil (data not shown). As with the Y-maze test, all compounds showed a dose response with intravenous administration and the

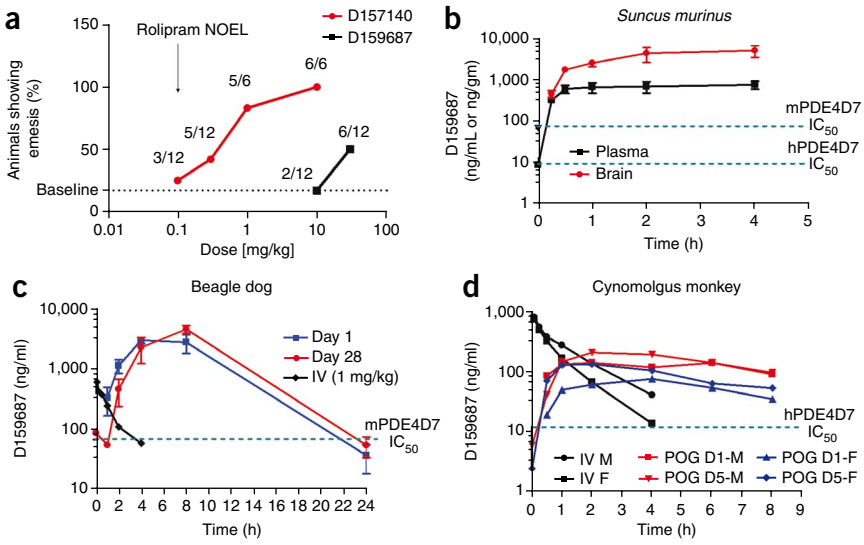


Figure 6 Effects of compounds on emesis. Comparison of emetic threshold in three species by oral dosing. (**a**) *S. murinus* shrews were dosed with D159687, a PDE4D allosteric modulator, or with D157140, a PDE4D full inhibitor. Emetic activity was expressed as the number of animals showing emesis per number of animals tested at the respective dose ($n = 6$ or 12). Baseline emesis of 17% is consistent with prior experience with the *S. murinus* model (data not shown). The NOEL (no observable effect level) for emesis for D159687 was 10 mg/kg. D157140 was emetic at 0.1 mg/kg. (**b**) For toxicokinetic analysis, the test compounds were administered to groups of 18 male *S. murinus* each after a wash-out period of about 1 week after the emesis test. At $t = 0$, 0.25, 1, 2 and 4 h post-dose, three animals each were terminally bled, and plasma and brain samples were collected for toxicokinetic analyses. The *in vitro* IC_{50} for inhibition of human (hPDE4D7) and nonprimate (mPDE4D7) enzymes are indicated as blue dotted lines. (**c**) Toxicokinetic profile of D159687 in the female beagle dog. Data are shown for day 1 and day 28 of a 28-d repeat dose study at 100 mg/kg ($N = 4$ per dose) by powder in capsule (PIC). Emesis was not observed in any animal. D159687 was emetic at 600 mg/kg in a single-dose range finding study. Like mouse, dog has the nonprimate Tyr196 amino acid polymorphism in PDE4D. (**d**) Toxicokinetic profile of D159687 in the cynomolgus monkey. One male and one female monkey were dosed intravenously (IV) with D159687 at 1 mg/kg or orally by gavage (POG) for 5 d at 10 mg/kg in 0.5% HPMC (hydroxypropyl methylcellulose), 0.5% poloxamer 188, 0.4% Tween 80. Toxicokinetic data were collected on day 1 and day 5 of the multi-dose study. Emesis was not observed in either monkey. D159687 was emetic at 30 mg/kg in a single-dose range finding study. The *in vitro* IC_{50} of D159687 against human PDE4D7 is indicated as a blue dotted line.

maximum cognitive benefit was similar to rolipram (**Supplementary Table 4**). The allosteric modulators D159404 and D159687 consistently showed slightly greater potency, and provided cognitive benefit when administered orally with MED consistent with bioavailability (**Supplementary Fig. 13**). D159404 and D159687 also provided cognitive benefit in rat in the NOR test with dosing before the first training session with MED ≤ 10 μg/kg (**Supplementary Fig. 14**).

Emetic potential of PDE4 allosteric modulators

We next profiled our compounds in the ketamine/xylazine test, which has been proposed as a behavioral correlate of emesis in the mouse (**Fig. 5c**)[43]. As rodents are unable to vomit, reduction of the duration of ketamine/xylazine-induced anesthesia has been introduced as a behavioral correlate that is sensitive to PDE4D gene deletion[19,43]. Consistent with the 30–40× selectivity for mouse PDE4D over PDE4B (**Supplementary Table 2**), PDE4D full inhibitors potently reduce anesthesia in this model at doses similar to their MED for cognitive benefit (**Supplementary Table 4**). In contrast, and even at 1,000× the MED for cognitive benefit, PDE4D allosteric modulators had little or no effect on anesthesia duration (**Supplementary Table 4**).

To further investigate the topic of emesis, we profiled D159687 for emetic activity in *Suncus murinus* (Asian house shrew), the beagle dog and the cynomolgus monkey (**Fig. 6**). *S. murinus* have an emetic response to motion, ethanol overdose and many classes of drug that

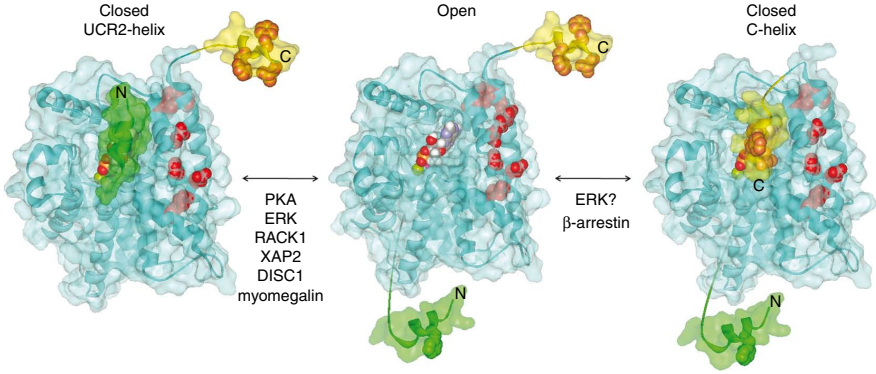


Figure 7 A model to explain how PDE4 regulatory domains control PDE4 activity. The PDE4 enzyme can exist in multiple conformations with UCR2 (green) or C-terminal (yellow) regulatory helices allowing access ("open") or blocking access ("closed") to the active site. In the model, movement of regulatory helices is affected by phosphorylation states of the enzyme and by the presence of partner proteins. Residues shown to be important for binding RACK1 and β-arresting are highlighted in red and orange, respectively. The model explains how a very diverse set of biological inputs could simultaneously modulate PDE4 activity *in vivo*. For example, a combination of PKA phosphorylation and β-arrestin binding would drive the enzyme into a fully activated/open state by limiting the ability of both UCR2 and C-terminal helices to bind across the active site, whereas a combination of ERK phosphorylation and XAP2 binding would lock the enzyme into a fully inactive and/or closed conformation. If these phosphorylation states and partner proteins alter the equilibrium between open and closed conformations, the catalytic activity could be finely controlled across a very wide dynamic range in response to subtle changes in phosphorylation levels or partner protein concentrations.

© 2010 Nature America, Inc. All rights reserved.

are emetic in human[44,45]. The PDE4D-selective allosteric modulator was 100× less emetic than rolipram in *S. murinus*, 3,000× less emetic than rolipram in the beagle dog and 500× less emetic in monkey. The emetic potential of PDE4 inhibitors can be reduced by reducing their distribution to brain[21]. To our knowledge, there has been no previous demonstration of reduced emetic potential for a PDE4D selective compound that preferentially distributes to brain. Thus, the mechanism of action of PDE4D allosteric modulators reduces potential for emesis while maintaining pro-cognitive efficacy.

DISCUSSION

The structural, kinetic and mutational results presented above can be used to create a general model for the regulation of PDE4 activity by controlling access to the active site (**Fig. 7**). This mechanism is probably general to other PDE families, as many of these contain unique upstream regulatory domains that likely function analogously to UCR2 in PDE4 (e.g., GAF domains in PDE2, 5, 6, 10 and 11; Ca^{2+}/calmodulin domains in PDE1; and the PAS domain in PDE8). Our model for regulation of PDE4 activity is supported by previous studies examining phosphorylation states of PDE4. For example, it has been shown that UCR2 negatively regulates PDE4 activity and that UCR1 phosphorylation by PKA releases this inhibitory activity[7,8,33]. According to the model, in the absence of PKA phosphorylation at Ser54, UCR2 adopts a 'closed' conformation. PKA phosphorylation alters interactions between UCR1 and UCR2 (interactions that are not visualized in the current structures), causing the UCR2 helix to adopt the 'open' active conformation (**Supplementary Fig. 15**). PKA phosphorylation activates the enzyme by increasing V_{max} without affecting K_M (ref. 8). This kinetic behavior is explained by the open versus closed UCR2 conformations, since the UCR2 helix is reducing the concentration of active enzyme and thereby suppressing V_{max}. It also has been shown that ERK phosphorylation inhibits the activity of some PDE4 isoforms and that this inhibition requires UCR2 (ref. 46). This suggests that phosphorylated Ser579 interacts directly with the UCR2 helix to stabilize the closed conformation. It is interesting to note that modeling suggests that the phosphoserine (phospho-Ser579 in PDE4D3; phosphor-Ser659 in PDE4B3) would be in close proximity to a conserved arginine residue on UCR2 and a conserved lysine on the catalytic domain (**Supplementary Fig. 16**). This could provide a network of interactions that would directly stabilize the UCR2 closed conformation and result in inhibition after ERK phosphorylation of PDE4D or PDE4B.

The model (**Fig. 7**) also explains recent peptide-array, yeast-two-hybrid and immunoprecipitation studies demonstrating that partner proteins can decrease PDE4 activity[47,48]. For example, the immunophilin XAP2 has been shown to interact with the N-terminal region of UCR2 not visible in our crystal structures and decrease enzymatic activity. This binding also increases the sensitivity to inhibition by rolipram[49], consistent with XAP2 stabilizing the UCR2 closed conformation and interactions between UCR2 and rolipram. DISC1, encoded by a gene that is disrupted in a familial form of schizophrenia[50], interacts with UCR2 and with residues on the catalytic domain of PDE4D[51]. PKA phosphorylation disrupts DISC1 binding to PDE4B and activates the enzyme[47]. These results can be explained by DISC1 stabilizing the closed conformation while phosphorylation of UCR1 at Ser54 disrupts this interaction, thereby activating the enzyme. Binding of partner proteins may also activate PDE4 by shifting the equilibrium to the open conformation. For example, myomegalin has been shown to bind the N-terminal region of UCR2 and as a result interact with PDE4D5 and PDE4D3 (long and short forms of PDE4D), but not with the supershort isoform PDE4D2[48]. Myomegalin may hold UCR2 in an open conformation, as immunoprecipitates contain active PDE4D3. This could limit the toxicity of PDE4D allosteric modulators in the heart[12].

Our medicinal chemistry efforts have focused on developing PDE4D selective modulators that distribute into brain for the enhancement of cognition. Lead compounds display substantially wider therapeutic windows with respect to emesis than do earlier, active site–directed PDE4 inhibitors. We confirm that PDE4D inhibition is highly emetic (e.g., D157140 in *S. murinus*). We have shown that PDE4D allosteric modulators have much less effect on cAMP levels in a cellular model than do full inhibitors of PDE4, yet are highly potent and efficacious in cellular and *in vivo* models of cAMP signaling. We speculate that PDE4 allosteric modulators may better maintain spatial and temporal aspects of cAMP signaling, thereby improving tolerability because of the limit placed on the magnitude of PDE4 inhibition.

PDE4 inhibitors as a class have been observed to cause vasculopathy in rodents and other species, although this toxicity has not been reported in humans. The vascular lesions consist of medial necrosis of small arteries accompanied by perivascular fibrosis and the presence of inflammatory cells, particularly in the mesentery, but also in multiple other organs[52]. Mesenteric vasculopathy is difficult to monitor in human clinical trials. We so far have not observed PDE4D allosteric modulators to cause mesenteric vasculopathy in rats, mice or dogs, nor do we see induction of TIMP-1 or interleukin-6 in rodents (unpublished data), although further studies in additional species and of longer duration will be needed to confirm these initial data.

The data presented here indicate that the activity of PDE4D can be modulated by bifunctional compounds bridging Phe196 in UCR2 and Gln535 in the active site. We have altered binding in the active site by reducing hydrophobic interactions with the P clamp residues Phe538 and Ile502 and potential interactions with residues near the di-metal ion center (M site). By so doing, we created allosteric modulators with partial enzyme inhibition kinetics. PDE4 allosteric modulators inhibit PDE4 >50%, as closing UCR2 across one active site in the PDE4 dimer also decreases the catalytic activity of the second active site. Thus, PDE4 allosteric modulators that do not completely inhibit enzyme activity reduce emetic potential while maintaining efficacy in cellular and *in vivo* models.

METHODS

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

Accession codes. Protein Data Bank: Coordinates have been deposited with accession codes 3G4G, showing RS25344 bound to the PDE4D catalytic domain and UCR2 (**Fig. 2**); 3G45, showing PMNPQ bound to the PDE4B catalytic domain and UCR2 (**Fig. 2**); 3IAD, showing D159153 bound to the PDE4D catalytic domain and UCR2 (**Fig. 4**); 3G58, showing PMNPQ bound to the PDE4D catalytic domain and the C-terminal regulatory helix (**Fig. 2b** and **Supplementary Fig. 3**); 3G4I, showing RS25344 bound the PDE4D catalytic domain (**Fig. 2b**); 3G4K, showing rolipram bound to the PDE4D catalytic domain (**Fig. 2b**); 3G4L, showing roflumilast bound to the PDE4D catalytic domain (**Supplementary Notes**).

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

ACKNOWLEDGMENTS

The development of Gene Composer software used to design protein constructs was supported in part by the National Institute of General Medical Sciences–National Center for Research Resources, co-sponsored PSI-2 Specialized Center Grant U54 GM074961 for the Accelerated Technologies Center for Gene to 3D Structure. The authors would like to thank M. Smith, M.H. Haraldsson, G.V. Halldorsdottir, B.B. Sigurdsson, G. Bragason, I. Saemundsdottir, B. Gudmundsdottir, T.J. Dagbjartsdottir, K. Astradsdottir, S. Gunnarsdottir, B. Eiriksdottir, N. Zhou, D. Sullins, P. Rauen, A. Motta, W. Zeller, J. Christensen and M. O'Connell for contributions to the research. We also thank Dr. Akira Ito

© 2010 Nature America, Inc. All rights reserved.

and colleagues at Dainippon Sumitomo and Dr. Klaus Mendla and colleagues at Boerhinger Ingelheim for contributions to the animal studies.

AUTHOR CONTRIBUTIONS

A.B.B., P.W., B.L.S., L.J.S. and M.E.G. contributed to the structural and molecular biology experiments. O.T.M., J.M.B, M.T., S.H. and M.E.G. contributed to kinetic, safety and efficacy studies. J.S., T.H., A.S.K. and M.E.G. contributed to medicinal chemistry experiments. A.B.B. and M.E.G. wrote the manuscript.

COMPETING INTERESTS STATEMENT

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/.

© 2010 Nature America, Inc. All rights reserved.

1. Omori, K. & Kotera, J. Overview of PDEs and their regulation. *Circ. Res.* **100**, 309–327 (2007).
2. Wang, H., Robinson, H. & Ke, H. The molecular basis for different recognition of substrates by phosphodiesterase families 4 and 10. *J. Mol. Biol.* **371**, 302–307 (2007).
3. Bolger, G. *et al.* A family of human phosphodiesterases homologous to the dunce learning and memory gene product of *Drosophila melanogaster* are potential targets for antidepressant drugs. *Mol. Cell. Biol.* **13**, 6558–6571 (1993).
4. Jacobitz, S., McLaughlin, M.M., Livi, G.P., Burman, M. & Torphy, T.J. Mapping the functional domains of human recombinant phosphodiesterase 4A: structural requirements for catalytic activity and rolipram binding. *Mol. Pharmacol.* **50**, 891–899 (1996).
5. Rocque, W.J. *et al.* Human recombinant phosphodiesterase 4B2B binds (R)-rolipram at a single site with two affinities. *Biochemistry* **36**, 14250–14261 (1997).
6. Beard, M.B. *et al.* UCR1 and UCR2 domains unique to the cAMP-specific phosphodiesterase family form a discrete module via electrostatic interactions. *J. Biol. Chem.* **275**, 10349–10358 (2000).
7. MacKenzie, S.J. *et al.* Long PDE4 cAMP specific phosphodiesterases are activated by protein kinase A-mediated phosphorylation of a single serine residue in Upstream Conserved Region 1 (UCR1). *Br. J. Pharmacol.* **136**, 421–433 (2002).
8. Sette, C. & Conti, M. Phosphorylation and activation of a cAMP-specific phosphodiesterase by the cAMP-dependent protein kinase. Involvement of serine 54 in the enzyme activation. *J. Biol. Chem.* **271**, 16526–16534 (1996).
9. Shakur, Y., Pryde, J.G. & Houslay, M.D. Engineered deletion of the unique N-terminal domain of the cyclic AMP-specific phosphodiesterase RD1 prevents plasma membrane association and the attainment of enhanced thermostability without altering its sensitivity to inhibition by rolipram. *Biochem. J.* **292**, 677–686 (1993).
10. Bolger, G.B. *et al.* The unique amino-terminal region of the PDE4D5 cAMP phosphodiesterase isoform confers preferential interaction with beta-arrestins. *J. Biol. Chem.* **278**, 49230–49238 (2003).
11. Bolger, G.B. *et al.* Scanning peptide array analyses identify overlapping binding sites for the signalling scaffold proteins, beta-arrestin and RACK1, in cAMP-specific phosphodiesterase PDE4D5. *Biochem. J.* **398**, 23–36 (2006).
12. Lehnart, S.E. *et al.* Phosphodiesterase 4D deficiency in the ryanodine-receptor complex promotes heart failure and arrhythmias. *Cell* **123**, 25–35 (2005).
13. Blokland, A., Schreiber, R. & Prickaerts, J. Improving memory: a role for phosphodiesterases. *Curr. Pharm. Des.* **12**, 2511–2523 (2006).
14. Houslay, M.D., Schafer, P. & Zhang, K.Y. Keynote review: phosphodiesterase-4 as a therapeutic target. *Drug Discov. Today* **10**, 1503–1519 (2005).
15. DeMarch, Z., Giampa, C., Patassini, S., Bernardi, G. & Fusco, F.R. Beneficial effects of rolipram in the R6/2 mouse model of Huntington's disease. *Neurobiol. Dis.* **30**, 375–387 (2008).
16. Zhang, H.T. Cyclic AMP-specific phosphodiesterase-4 as a target for the development of antidepressant drugs. *Curr. Pharm. Des.* **15**, 1688–1698 (2009).
17. Giembycz, M.A. Life after PDE4: overcoming adverse events with dual-specificity phosphodiesterase inhibitors. *Curr. Opin. Pharmacol.* **5**, 238–244 (2005).
18. Spina, D. PDE4 inhibitors: current status. *Br. J. Pharmacol.* **155**, 308–315 (2008).
19. Robichaud, A. *et al.* Deletion of phosphodiesterase 4D in mice shortens alpha(2)-adrenoceptor-mediated anesthesia, a behavioral correlate of emesis. *J. Clin. Invest.* **110**, 1045–1052 (2002).
20. Robichaud, A., Savoie, C., Stamatiou, P.B., Tattersall, F.D. & Chan, C.C. PDE4 inhibitors induce emesis in ferrets via a noradrenergic pathway. *Neuropharmacology* **40**, 262–269 (2001).
21. Aoki, M. *et al.* Studies on mechanisms of low emetogenicity of YM976, a novel phosphodiesterase type 4 inhibitor. *J. Pharmacol. Exp. Ther.* **298**, 1142–1149 (2001).
22. Giembycz, M.A. Development status of second generation PDE4 inhibitors for asthma and COPD: the story so far. *Monaldi Arch. Chest Dis.* **57**, 48–64 (2002).
23. Conn, P.J., Christopoulos, A. & Lindsley, C.W. Allosteric modulators of GPCRs: a novel approach for the treatment of CNS disorders. *Nat. Rev. Drug Discov.* **8**, 41–54 (2009).
24. Altucci, L., Leibowitz, M.D., Ogilvie, K.M., de Lera, A.R. & Gronemeyer, H. RAR and RXR modulation in cancer and metabolic disease. *Nat. Rev. Drug Discov.* **6**, 793–810 (2007).
25. Hoffmann, R., Wilkinson, I.R., McCallum, J.F., Engels, P. & Houslay, M.D. cAMP-specific phosphodiesterase HSPDE4D3 mutants which mimic activation and changes in rolipram inhibition triggered by protein kinase A phosphorylation of Ser-54: generation of a molecular model. *Biochem. J.* **333**, 139–149 (1998).
26. Houslay, M.D. & Adams, D.R. PDE4 cAMP phosphodiesterases: modular enzymes that orchestrate signalling cross-talk, desensitization and compartmentalization. *Biochem. J.* **370**, 1–18 (2003).
27. Saldou, N. *et al.* Comparison of recombinant human PDE4 isoforms: interaction with substrate and inhibitors. *Cell. Signal.* **10**, 427–440 (1998).
28. Souness, J.E. & Rao, S. Proposal for pharmacologically distinct conformers of PDE4 cyclic AMP phosphodiesterases. *Cell. Signal.* **9**, 227–236 (1997).
29. Brideau, C., Van Staden, C., Styhler, A., Rodger, I.W. & Chan, C.C. The effects of phosphodiesterase type 4 inhibitors on tumour necrosis factor-alpha and leukotriene B4 in a novel human whole blood assay. *Br. J. Pharmacol.* **126**, 979–988 (1999).
30. Reid, P. Roflumilast Altana Pharma. *Curr. Opin. Investig. Drugs* **3**, 1165–1170 (2002).
31. Lorimer, D. *et al.* Gene Composer: database software for protein construct design, codon engineering, and gene synthesis. *BMC Biotechnol.* **9**, 36 (2009).
32. Raymond, A. *et al.* Combined protein construct and synthetic gene engineering for heterologous protein expression and crystallization using Gene Composer. *BMC Biotechnol.* **9**, 37 (2009).
33. Lim, J., Pahlke, G. & Conti, M. Activation of the cAMP-specific phosphodiesterase PDE4D3 by phosphorylation. Identification and function of an inhibitory domain. *J. Biol. Chem.* **274**, 19677–19685 (1999).
34. Wang, P. *et al.* Expression, purification, and characterization of human cAMP-specific phosphodiesterase (PDE4) subtypes A, B, C, and D. *Biochem. Biophys. Res. Commun.* **234**, 320–324 (1997).
35. Xu, R.X. *et al.* Crystal structures of the catalytic domain of phosphodiesterase 4B complexed with AMP, 8-Br-AMP, and rolipram. *J. Mol. Biol.* **337**, 355–365 (2004).
36. Robichaud, A., Tattersall, F.D., Choudhury, I. & Rodger, I.W. Emesis induced by inhibitors of type IV cyclic nucleotide phosphodiesterase (PDE IV) in the ferret. *Neuropharmacology* **38**, 289–297 (1999).
37. Card, G.L. *et al.* Structural basis for the activity of drugs that inhibit phosphodiesterases. *Structure* **12**, 2233–2247 (2004).
38. McCahill, A. *et al.* In resting COS1 cells a dominant negative approach shows that specific, anchored PDE4 cAMP phosphodiesterase isoforms gate the activation, by basal cyclic AMP production, of AKAP-tethered protein kinase A type II located in the centrosomal region. *Cell. Signal.* **17**, 1158–1173 (2005).
39. Chambers, R.J. *et al.* A new chemical tool for exploring the role of the PDE4D isozyme in leukocyte function. *Bioorg. Med. Chem. Lett.* **16**, 718–721 (2006).
40. Souness, J.E. *et al.* Suppression of eosinophil function by RP 73401, a potent and selective inhibitor of cyclic AMP-specific phosphodiesterase: comparison with rolipram. *Br. J. Pharmacol.* **115**, 39–46 (1995).
41. Mihara, T. *et al.* Pharmacological characterization of a novel, potent adenosine A1 and A2A receptor dual antagonist, 5-[5-amino-3-(4-fluorophenyl)pyrazin-2-yl]-1-isopropylpyridine-2(1H)-one (ASP5854), in models of Parkinson's disease and cognition. *J. Pharmacol. Exp. Ther.* **323**, 708–719 (2007).
42. Bailey, C.H., Bartsch, D. & Kandel, E.R. Toward a molecular definition of long-term memory storage. *Proc. Natl. Acad. Sci. USA* **93**, 13445–13452 (1996).
43. Robichaud, A. *et al.* Assessing the emetic potential of PDE4 inhibitors in rats. *Br. J. Pharmacol.* **135**, 113–118 (2002).
44. Hirose, R. *et al.* Correlation between emetic effect of phosphodiesterase 4 inhibitors and their occupation of the high-affinity rolipram binding site in Suncus murinus brain. *Eur. J. Pharmacol.* **573**, 93–99 (2007).
45. Ueno, S., Matsuki, N. & Saito, H. Suncus murinus: a new experimental model in emesis research. *Life Sci.* **41**, 513–518 (1987).
46. MacKenzie, S.J., Baillie, G.S., McPhee, I., Bolger, G.B. & Houslay, M.D. ERK2 mitogen-activated protein kinase binding, phosphorylation, and regulation of the PDE4D cAMP-specific phosphodiesterases. The involvement of COOH-terminal docking sites and NH2-terminal UCR regions. *J. Biol. Chem.* **275**, 16609–16617 (2000).
47. Millar, J.K. *et al.* DISC1 and PDE4B are interacting genetic factors in schizophrenia that regulate cAMP signaling. *Science* **310**, 1187–1191 (2005).
48. Verde, I. *et al.* Myomegalin is a novel protein of the golgi/centrosome that interacts with a cyclic nucleotide phosphodiesterase. *J. Biol. Chem.* **276**, 11189–11198 (2001).
49. Bolger, G.B. *et al.* Attenuation of the activity of the cAMP-specific phosphodiesterase PDE4A5 by interaction with the immunophilin XAP2. *J. Biol. Chem.* **278**, 33351–33363 (2003).
50. Millar, J.K. *et al.* Genomic structure and localisation within a linkage hotspot of Disrupted In Schizophrenia 1, a gene disrupted by a translocation segregating with schizophrenia. *Mol. Psychiatry* **6**, 173–178 (2001).
51. Murdoch, H. *et al.* Isoform-selective susceptibility of DISC1/phosphodiesterase-4 complexes to dissociation by elevated intracellular cAMP levels. *J. Neurosci.* **27**, 9513–9524 (2007).
52. Zhang, J. *et al.* Histopathology of vascular injury in Sprague-Dawley rats treated with phosphodiesterase IV inhibitor SCH 351591 or SCH 534385. *Toxicol. Pathol.* **36**, 827–839 (2008).
53. Naganuma, K. *et al.* Discovery of selective PDE4B inhibitors. *Bioorg. Med. Chem. Lett.* **19**, 3174–3176 (2009).
54. Hersperger, R., Bray-French, K., Mazzoni, L. & Muller, T. Palladium-catalyzed cross-coupling reactions for the synthesis of 6, 8-disubstituted 1,7-naphthyridines: a novel class of potent and selective phosphodiesterase type 4D inhibitors. *J. Med. Chem.* **43**, 675–682 (2000).

ONLINE METHODS

Animal experiments were reviewed and approved by the Laboratory Animal Ethical Committee in Iceland. Policies and procedures at other performance sites were in accordance with applicable national regulations.

Kinetic assay of PDE4 activity. A real-time, kinetic assay was developed for accurately determining initial rates of cAMP hydrolysis by purified PDE4. The assay is based on coupling the formation of the PDE4 reaction product, 5′-AMP to the oxidation of NADH, by the use of three coupling enzymes (adenylate kinase, pyruvate kinase and lactate dehydrogenase), which allows for a convenient spectrophotometric (or fluorescent) readout of reaction rates. Conditions were established wherein the PDE4 reaction was fully rate determining and optimal amounts of coupling enzymes and other reagents were determined.

Assays were performed in 96-well plates in a total volume of 200 µl/well. Compounds were dissolved in DMSO or DMSO for controls was added to plates in a volume of 10 µl followed by addition of 180 µl of assay mix. Plates were pre-incubated at 25 °C for 15 min and the reactions were initiated by the addition of 10 µl of cAMP followed by thorough mixing. Reaction rates were measured by monitoring the decrease in absorbance at 340 nM for a period of 20 min in a SpectraMax 190 (Molecular Devices) plate reader. Initial rates (slopes) were determined from linear portions of the progress curves. Final concentrations of assay components were as follows: 50 mM Tris, pH 8, 10 mM $MgCl_2$, 50 mM KCl, 2.5% DMSO, 0.5 mM TCEP, 1 mM PEP, 0.4 mM NADH, 40 µM ATP, 40 µM cAMP, 4.5 units adenylate kinase, 0.86 units pyruvate kinase, 1.13 units lactate dehydrogenase, ~1 nM PDE4.

A 12-point (one row) dose-response dilution scheme was used for determining IC_{50} values (**Fig. 1b–d**, S1, S2, S8 and S9). Compounds were measured in three replicates per point. Two rows (12 wells) were designated for positive and negative controls, respectively. Positive controls contain assay mix, DMSO and cAMP whereas negative controls contain only assay mix and DMSO. Data are expressed as percent inhibition, which is calculated as shown in equation (1).

$$\%\,\text{Inhibition} = (\text{AvgPos} - \text{sample})/(\text{AvgPos} - \text{AvgNeg})*100 \quad (1)$$

Assay variability within plates was assessed by calculating a Z′-factor as shown in equation (2).

$$Z' = 1 - [3(\text{StdPos} + \text{StdNeg})/(\text{AvgPos} - \text{AvgNeg})] \quad (2)$$

A Z′-value of 0.6 and higher was considered adequate and data with Z′ values <0.6 were discarded. Data were analyzed using Graphpad Prism. Data were fit to a standard four-parameter sigmoidal dose-response equation (equation (3)).

$$Y = \min + (\max - \min)/(1 + (X/IC_{50})^{\text{Hill}}) \quad (3)$$

The K_M for substrate with PDE4D and PDE4B was determined using varying cAMP (1–400 µM) and the data were fit to the standard Michaelis-Menten equation (equation (4)).

$$V = V_{max}/[\text{cAMP}] + K_M \quad (4)$$

cAMP assay in forskolin-stimulated human kidney HEK293 cells. HEK293 cells (10^4/well) were grown in DMEM media supplemented with 10% FBS (FBS) and 1% penicillin-streptomycin in 96-well plates coated with poly-D-lysine. Cells were grown at 37 °C in an atmosphere of 5% CO_2. Cells were seeded, grown overnight and used at 70–85% confluency. After overnight growth the medium was removed and cells carefully washed 1× with warm PBS. The cells were then incubated with serum-free media containing 5 µM forskolin (FSK) and the test article for 20 min at 37 °C to stimulate cAMP production. Positive controls (12 wells) contained only media plus FSK and negative controls (12 wells) contained media alone. Each test article was measured at 12 different concentrations using a serial dilution scheme and in triplicates at each concentration. After stimulation, cells were washed, lysed and the cAMP concentration in the lysates was measured using a commercial enzyme-immunoassay (Biotrak, Amersham Biosciences) according to the manufacturer's instructions. Dose-response data were analyzed using Prism (GraphPad) and fit to a standard four-parameter, sigmoidal dose-response equation (see equation (3), above).

Sephadex-stimulated LTE4 assay in human whole blood. Blood was collected into heparinized Vacutainer tubes from healthy male or female volunteers with informed consent. Fresh blood (352 µl) and test article (8 µl) dissolved in DMSO:H_2O (1:1) were pre-incubated at 37 °C for 15 min in a 96-well plate, followed by addition of 40 µl of a slurry of Sephadex G-15 suspended in PBS (0.16 g/ml). Each sample plate contained two test articles measured in triplicate at each concentration, eight positive controls (no compound) and eight negative controls (no Sephadex). Samples were incubated at 37 °C for 4 h, followed by addition of 8 µl of 15% EDTA solution and centrifugation at 115*g*. Quantification of LTE4 in the resulting plasma samples was determined using a commercial enzyme-linked immunosorbent assay (ELISA) from Cayman Chemicals following the manufacturers instructions. Dose-response data were analyzed using Prism (GraphPad) and fit to a standard three-parameter, sigmoidal dose-response equation with a fixed hillslope of unity.

Behavioral testing. For each of the behavioral tests, generally three or more doses of test compound were evaluated over a close log interval, (e.g., 10, 30, 100 µg/kg). Rolipram was used as a reference compound each day to validate results obtained across multiple days of testing. Test compounds were dissolved in PEG300/DMF/saline (4/1/5) for intravenous (i.v.) dosing. For the Y-maze test, a dose of scopolamine (0.6 mg/kg) was used that reduced alternation by ddY mice to chance (50%) levels. In the Y-maze, spontaneous alternation behavior is defined as entry into all three arms on consecutive occasions. For example, mice move consecutively from arm A → arm B → arm C. Scopolamine was injected at $t = -30$ min, the test compound was administered intravenously at $t = -5$ min, and Y-maze alternation was tested from $t = 0$ to $t = +8$ min. Percent alternation behavior is scored as [(no. of alternation behaviors recorded)/(no. of total arm entries − 2)] × 100. Male ddY mice were used for these tests as they show greater alternation behavior than other common strains of mice. Generally ddY mice were 30–35 g; 8–10 mice were used per group.

For the novel object recognition or NOR test, two identical objects were presented to mice in the testing arena (trial 1 or T1) for 5 min. One of two identical objects presented to the mice during T1 was replaced by a novel object during trial 2 (T2), which was performed 24 h after T1 (that is, an inter-trial interval of 24 h). The test compound was administered after T1 at $t = +3$ h for i.v. dosing. Effects on memory consolidation were assessed at $t = +24$ h during T2 by presenting one familiar object and one novel object for 5 min. The time spent by the mouse exploring each object was recorded. The discrimination index was calculated as time spent exploring (Novel − Familiar)/(Novel + Familiar). Adult male ddY mice (30–35 g) were used for the NOR test.

For the ketamine/xylazine test, male ddY mice of 24–32 g were anesthetized with combined subcutaneous injection of xylazine (10 mg/kg) and ketamine (80 mg/kg). Test compounds were administered by i.v. injection at $t = +15$ min after induction of anesthesia. The duration of anesthesia was defined as the time between the ketamine/xyalazine injection and recovery of the righting reflex; generally this was between 50–60 min. Recovery of the righting reflex was scored when the mice returned spontaneously to the prone position when placed in dorsal recumbency. The NOEL for rolipram in the ketamine/xylazine test is 30 µg/kg.

Statistical analysis. Statistical comparison of behavioral data was by Wilcoxon rank sum test or Dunnet's multiple comparison test. The MED was identified as the minimum dose at which cognitive benefit or reduction in anesthesia duration became statistically significant ($P \leq 0.05$). WinNonlin (Pharsight) was used for the analysis of pharmacokinetic data.

Assessment of emesis. Male *S. murinus* >5 weeks old (body weight 40–60 g) were dosed by oral gavage with the compound dissolved and/or suspended in vehicle consisting of 0.5% HPMC (hydroxypropyl methylcellulose), 0.9% benzyl alcohol, 0.4% Tween 80 in a constant volume of 10 ml/kg body weight. Thereafter, the *S. murinus* were placed individually in a transparent observation box. Control animals were treated with the same volume of vehicle. After gavage, the animals were monitored continuously for 180 min and the incidence of vomiting episodes was recorded. Studies in the beagle

© 2010 Nature America, Inc. All rights reserved.

dog and cynomolgus monkey were performed similarly with a 180-min observation period after oral dosing.

Chemical synthetic methods. Synthetic methods are reported in ref. 55. PMNPQ, was prepared as described[56]. RS25344 was prepared as described[57].

55. Singh, J. *et al.* Biaryl inhibitors for treating pulmonary and cardiovascular disorders. PCT/US2008/084193 (2008).
56. Deschenes, D. *et al.* Substituted 8-arylquinoline phosphodiesterase-4 inhibitors. WO 94/22852 (2000).
57. Wilhelm, R. *et al.* Optionally substituted pyrido[2,3-D]pyridine-2,4(1H,3H)-diones and pyrido[2,3-D]pyrimidine-2(1H,3H)-ones. US 5,264,437 (1993).

© 2010 Nature America, Inc. All rights reserved.

doi:10.1038/nbt.1598

nature biotechnology

Chimeric mouse tumor models reveal differences in pathway activation between ERBB family– and KRAS-dependent lung adenocarcinomas

Yinghui Zhou[1,6], William M Rideout III[1,6], Tong Zi[1], Angela Bressel[1], Shailaja Reddypalli[1], Rebecca Rancourt[1], Jin-Kyeung Woo[1], James W Horner[2], Lynda Chin[3], M Isabel Chiu[1], Marcus Bosenberg[4], Tyler Jacks[5], Steven C Clark[1], Ronald A DePinho[2], Murray O Robinson[1] & Joerg Heyer[1]

© 2010 Nature America, Inc. All rights reserved.

To recapitulate the stochastic nature of human cancer development, we have devised a strategy for generating mouse tumor models that involves stepwise genetic manipulation of embryonic stem (ES) cells and chimera generation. Tumors in the chimeric animals develop from engineered cells in the context of normal tissue. Adenocarcinomas arising in an allelic series of lung cancer models containing *HER2* (also known as *ERBB2*), *KRAS* or *EGFR* oncogenes exhibit features of advanced malignancies. Treatment of $EGFR^{L858R}$ and $KRAS^{G12V}$ chimeric models with an EGFR inhibitor resulted in near complete tumor regression and no response to the treatment, respectively, accurately reflecting previous clinical observations. Transcriptome and immunohistochemical analyses reveal that PI3K pathway activation is unique to ERBB family tumors whereas *KRAS*-driven tumors show activation of the JNK/SAP pathway, suggesting points of therapeutic intervention for this difficult-to-treat tumor category.

Cancer is a complex disease driven by mutation and epigenetic alterations in oncogenic and tumor-suppressive pathways linked to the control of cell proliferation, survival and apoptosis[1]. The mutational alteration of genes in these cancer pathways can strongly influence how tumors respond to targeted agents. In non-small cell lung cancer, previous work has identified a mutually exclusive spectrum of activating mutations in *KRAS*, *EGFR* or *HER2*, coupled with disruption of the *p16INK4a* (also known as *CDKN2A*) and *p53* (also known as *TP53*) tumor suppressor genes[2]. *EGFR* mutations and/or amplification have been correlated with dramatic tumor regression upon treatment with targeted EGFR inhibitors (e.g., erlotinib (Tarceva))[3], whereas *KRAS* activating mutations (associated with smoking) correlate with resistance to EGFR inhibitors and other targeted agents[4]. The inability to inhibit activated KRAS has fueled efforts to identify tractable drug targets downstream of mutant KRAS signaling as alternative therapeutic points of intervention[5]. Furthermore, EGFR-driven tumors, initially sensitive to EGFR inhibitors, tend to develop resistance to targeted therapy by acquiring additional mutations in EGFR or activation of alternative receptor tyrosine kinases (RTKs)[6–12].

One possible approach to overcoming resistance to RTK-targeted therapy is to identify and validate essential downstream signaling pathways commonly activated by different RTKs, such that combined inhibition of these pathways leads to more durable therapeutic responses. The mutual exclusivity of mutations in *HER2*, *KRAS* and *EGFR* in human lung tumors suggests that they activate common signaling pathways to support tumorigenesis[13,14]. We sought to test this hypothesis by creating animal models for and analyzing tumors harboring mutations in KRAS or ERBB family members (such as *EGFR* and *HER2*). Current cancer modeling strategies include conventional germline transgenics, somatic mutation induction using virally delivered recombinase and tissue reconstitution. Each of these methods has its own caveats: germline cancer models containing more than one modified allele require lengthy timelines, and breeding costs to intercross the desired genetic modifications are high; models using viral transduction or tissue reconstitution are limited by the efficacy of the technique in the desired target tissue. Here we report a new technology that rapidly and consistently generates mouse cancer models that not only retain the context between tumor and adjacent stroma but can also be applied to the field in a timely and cost-efficient manner. We show that tumors from KRAS or ERBB family chimera models display not only common, but more importantly distinct signaling features, a finding that may have implications for treatment options and patient outcomes.

RESULTS

HER2-induced respiratory failure in transgenic mice

To model the oncogenic involvement of *HER2* in the lung, we generated 12 independent transgenic mouse lines carrying either the

[1]AVEO Pharmaceuticals, Cambridge, Massachusetts, USA. [2]Belfer Institute for Applied Cancer Science, Departments of Medical Onocology, Medicine and Genetics, Dana-Farber Cancer Institute and Department of Medicine and Genetics, Harvard Medical School, Boston, Massachusetts, USA. [3]Belfer Institute for Applied Cancer Science, Department of Medical Oncology, Dana-Farber Cancer Institute, Department of Dermatology, Harvard Medical School, Boston, Massachusetts, USA. [4]Department of Dermatology, Yale University School of Medicine, New Haven, Connecticut, USA. [5]Koch Institute and Department of Biology and Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [6]These authors contributed equally to this work. Correspondence should be addressed to J.H. (jheyer@aveopharma.com).

Received 29 May; accepted 16 October; published online 20 December 2009; doi:10.1038/nbt.1595

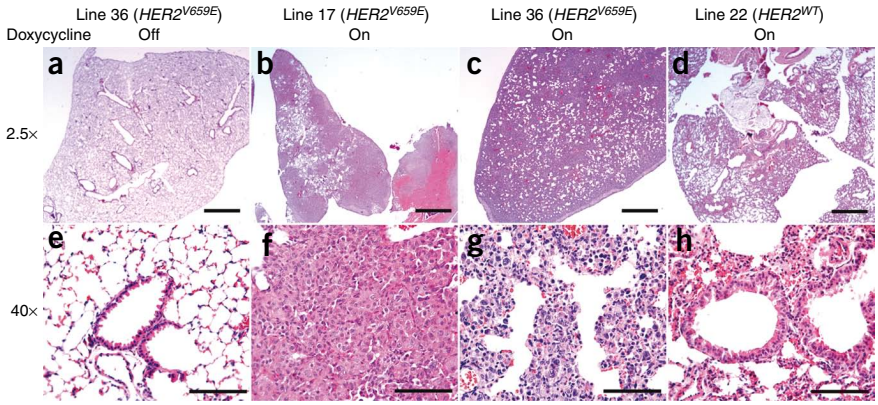


Figure 1 Phenotype of inducible transgenic *HER2* expression in the lung. (**a**–**h**) Line 17 has a low level of $HER2^{V659E}$ expression. Line 36 has high level of $HER2^{V659E}$ expression. Line 22 has a high level of $HER2^{wt}$ expression. Without doxycycline, the lungs of line 36 have a normal appearance (**a**,**e**). After 4 weeks of doxycycline induction, line 17 showed marked epithelial hyperplasia in the lung resulting in near-confluent growth and obliteration of much of the alveolar architecture. Tumor necrosis is present in the central portion of the smaller lung section (lower right) (**b**,**f**). After 2 weeks of doxycycline induction, line 36 showed marked and uniform epithelial hyperplasia in the lung (**c**,**g**). After 5 weeks of doxycycline induction, line 22 showed mild to moderate epithelial hyperplasia in the lung (**d**,**h**). Scale bars: **a**–**d**, 1 mm; **e**–**h**, 0.1 mm.

wild-type or neu mutation (*V659E*) of the human *HER2* gene under the control of the tetracycline-regulatable promoter (*tetO*)[15] and crossed them with *CCSP-rtTA* transgenic mice in which the reverse tetracycline-controlled transcriptional activator (rtTA) is expressed in the bronchiolar epithelium cells of the lung[16]. Three transgenic lines with lung-specific, inducible HER2 expression (line 36 ($HER2^{V659E}$ high), line 17 ($HER2^{V659E}$ low) and line 22 ($HER2^{WT}$) (data not shown)) were selected to test for *HER2*-dependent tumorigenesis in the lung. Unexpectedly, after receiving doxycycline for as little as 2 weeks, double transgenic mice exhibited respiratory distress, requiring euthanasia. Lungs from these mice were typically two to three times heavier than controls and, on histological analysis exhibited epithelial hyperplasia throughout the alveolar space and bronchi, with only rare focal early adenomas (**Fig. 1**). We surmised that acute onset diffuse hyperplasia caused rapidly progressive respiratory distress, thus precluding emergence of more advanced malignancies. To circumvent such physiological compromise with standard transgenic models, we exploited engineered embryonic stem (ES) cells and chimera formation to produce mice that are mosaic for engineered inducible *HER2* and other pro-oncogenic alleles.

ES cell–based mouse models of lung cancer

To develop a more clinically relevant stochastic model wherein aspiring cancer cells arise in the context of surrounding normal tissue, we used standard mouse chimera formation with genetically modified ES cells to generate mice possessing cells engineered with lung cancer–relevant alleles. All genetic alterations (disruption of tumor suppressor genes and introduction of inducible oncogenes) were performed in ES cells (**Fig. 2**). Thus, the tissues of chimeras were composed of cells from both the genetically modified ES cells and the genetically wild-type host blastocyst as evidenced by the patched coat colors of these mice.

Two major challenges surfaced when developing the chimeric model approach: functional validation of individual genetic elements and maintenance of pluripotency of the resultant ES clones. To overcome these challenges, we modified the ES cells stepwise and, at each step, functionally tested the introduced genetic elements as well as the ability of such targeted ES cell clones to contribute well to a host embryo. Only ES clones that passed both criteria were enlisted into the next round of modifications.

For the lung $HER2^{V659E}$ chimera model (termed LH chimera model), a 129/SvEv ES cell line was modified (i) by two rounds of targeted recombination to inactivate the *Ink4a/Arf* (also known as *Cdkn2a*) tumor suppressor gene resulting in ES-cell line 10E3, and

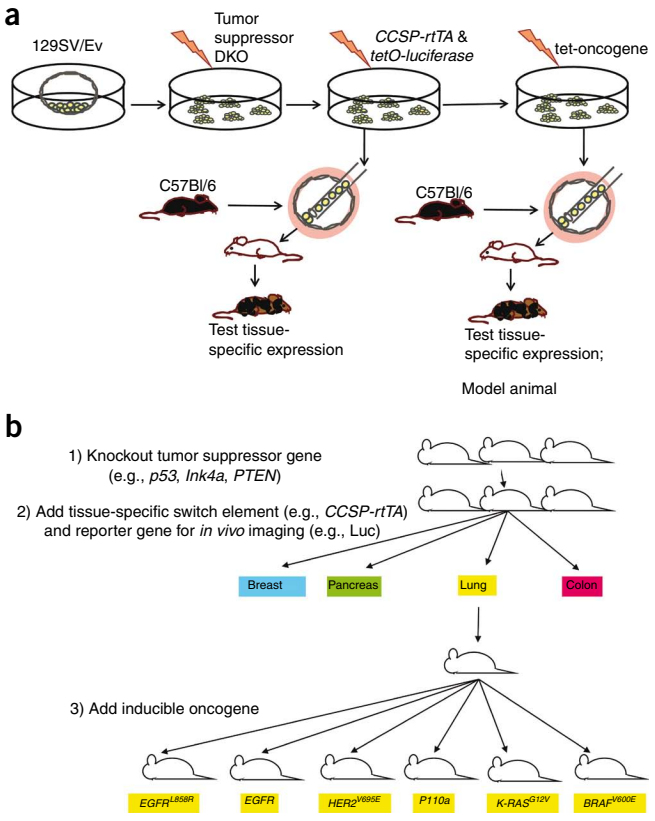


Figure 2 Illustration of the chimeric mouse tumor model approach. (**a**) Tumor model ES cell lines were constructed through sequential genetic modifications with *in vitro* and *in vivo* validation after each step. The first step involved biallelic modification of a tumor suppressor gene by homologous recombination (DKO, double knockout) and *in vivo* testing of chimera formation. The second step involved co-transfection of *rtTA* under the control of a tissue-specific promoter and *luciferase* under the control of a tetracycline-responsive promoter (*tetO-luciferase*) followed by testing of ES cell clones *in vivo* for tissue-specific induction of *luciferase* in chimeras. The third step involved transfection of validated ES cell clones from step two with a tetracycline-inducible oncogene construct (*tetO-oncogene*) and subsequent *in vivo* validation of tissue-specific oncogene induction in chimeras. (**b**) The flexibility of the chimeric mouse tumor model approach. By varying the tumor suppressor gene modification, the tissue-specific promoter driving the switch element, and the inducible oncogene in steps 1, 2 and 3, respectively, a library of models can be generated on the same genetic strain background and easily archived in liquid nitrogen.

© 2010 Nature America, Inc. All rights reserved.

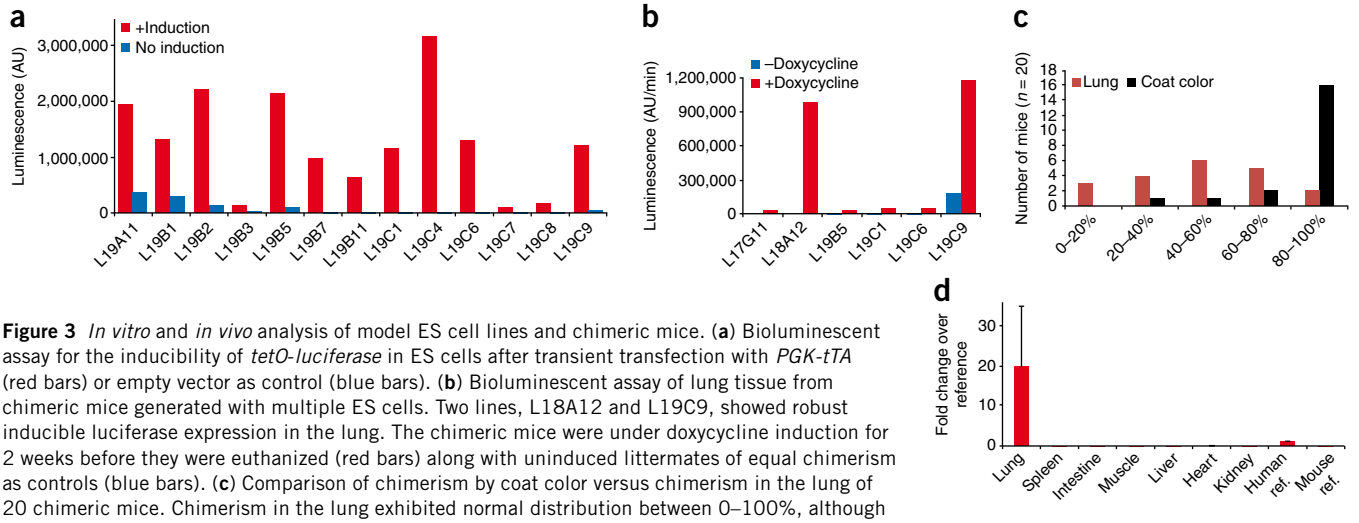


Figure 3 *In vitro* and *in vivo* analysis of model ES cell lines and chimeric mice. (**a**) Bioluminescent assay for the inducibility of *tetO-luciferase* in ES cells after transient transfection with *PGK-tTA* (red bars) or empty vector as control (blue bars). (**b**) Bioluminescent assay of lung tissue from chimeric mice generated with multiple ES cells. Two lines, L18A12 and L19C9, showed robust inducible luciferase expression in the lung. The chimeric mice were under doxycycline induction for 2 weeks before they were euthanized (red bars) along with uninduced littermates of equal chimerism as controls (blue bars). (**c**) Comparison of chimerism by coat color versus chimerism in the lung of 20 chimeric mice. Chimerism in the lung exhibited normal distribution between 0–100%, although 16 of 20 mice had >80% coat color chimerism. In contrast the chimerism in the lung was determined by quantification of the wild type and null Ink4a/Arf alleles by Southern blot analysis (Online Methods). (**d**) Tissue-specific expression of *HER2* in the lung of $HER2^{V659E}$-dependent chimeric mice. The levels of *HER2* were determined by qRT-PCR and normalized against human reference RNA.

© 2010 Nature America, Inc. All rights reserved.

(ii) by cotransfection of *CCSP-rtTA*[17], a *tetO-luciferase* transgene and a selection cassette (*PGK-puromycin*). Puromycin-resistant clones containing both transgenes were tested *in vitro* for *tetO* promoter–driven expression of *luciferase* (**Fig. 3a**).

Sixteen ES cell clones that expressed luciferase only in the presence of transiently transfected *tTA in vitro* were enlisted into chimera formation studies and assessed for ES cell contribution; 10/16 ES clones gave rise to chimeras with varying degrees (5–95%) of ES cell contribution evident by coat color. The chimeras were tested for doxycycline-inducible, lung-specific expression of luciferase (**Fig. 3b** and **Supplementary Tables 1** and **2**). Two ES clones, L18A12 and L19C9, which gave robust chimera formation, (>50% ES contribution) combined with lung-specific, doxycycline-inducible expression of luciferase, were selected as the platform for testing the chimera approach to cancer models. An inducible oncogene was added to the ES lines by co-transfection with *tetO-HER2*^{V659E} and *PGK-hygromycin*. Hygromycin-resistant and *tetO-HER2*^{V659E}–positive ES cell clones, that showed *tetO* promoter–driven *HER2*^{V659E} expression *in vitro*, were used to generate chimeric mice (**Supplementary Fig. 1a**). Chimerism in the lung ranged from 20–80% (**Fig. 3c**) and lung tissue from these chimeras was analyzed for *HER2* transgene expression by RT-PCR and immunohistochemistry. Chimeric mice derived from ES cell clones, LH33B3 and LH33B6 (*Ink4a/Arf* $^{-/-}$; *CCSP-rtTA*; *tetO-luciferase*; and *tetO-HER2*^{V659E}), showed a low copy number, and robust, doxycycline-inducible and lung-specific expression of the *HER2*^{V659E} transgene (**Supplementary Figs. 1b,c** and **Fig. 3d**).

HER2^{V659E} ES cell chimeras develop adenocarcinomas

ES cell clones, LH33B3 and LH33B6, were used to generate chimeric mice which, at 4 weeks of age, were administered doxycycline in water or food. In sharp contrast to the rapid-onset, diffuse hyperplasia phenotype of the conventional *HER2*^{V659E} germline transgenic model, 125/174 of the doxycyline-treated *HER2*^{V659E} chimeras developed visible lung tumor nodules within 4–7 months coinciding with respiratory compromise or a failure to thrive (**Fig. 4a**). Clinical symptom onset ranged from 2 months to 1 year and correlated with degree of coat color chimerism. Lung tumors from symptomatic mice histologically resembled invasive adenocarcinomas, which were not observed in the conventional *CCSP-HER2* wild-type and *V659E* transgenic models (**Fig. 1** versus **Fig. 4b**). A serial necropsy study of *HER2*^{V659E} chimeric mice revealed that adenocarcinoma formation required at least 6–8 weeks of doxycycline induction (**Supplementary Fig. 2**), which could explain why only hyperplasia and, rarely, adenomas were found in the conventional *HER2* transgenic mouse, which succumbed within 2–4 weeks (**Fig. 1**).

HER2^{V659E} chimeric mice on doxycycline for 4–7 months had an average lung weight sixfold greater than normal lungs due to extensive tumor burden (**Fig. 4c**). Most lung specimens possessed multiple tumor nodules (2–5 mm in diameter) separated by normal lung tissue. In <20% of the cases, typically those mice with the highest degree coat-color chimerism, the lungs were nearly effaced by highly malignant tumor cells. In low degree chimeras that survived >8 months on doxycycline, we typically observed one to two large, highly malignant nodules within otherwise normal lungs. These solitary nodules exhibited the capacity to reach 1 cm in diameter with similar histopathology to human lung adenocarcinomas and thus closely recapitulated advanced stages of human lung cancer progression (**Supplementary Fig. 3a**).

Immunohistochemical analysis of the *HER2*^{V659E} lung adenocarcinomas confirmed expression of (i) the human *HER2*^{V659E} transgene, (ii) cytokeratin, and (iii) surfactant protein C (SPC) in all tumor cells. Expression of the Clara Cell 10 kDa protein (CCSP) was not observed in the tumor cells (**Fig. 4b**, v–viii). We conclude that these tumors originated from lung epithelial cells descended from the genetically modified ES cells. *HER2*^{V659E} chimeric mice receiving doxycycline for <8 weeks showed small tumor nodules adjacent to terminal bronchioles, suggesting that these tumors originated from the bronchioalveolar epithelium (**Fig. 4b**, i).

The inducible nature of this model system afforded the opportunity to study whether expression of the transgenic oncogene was required for both tumorigenesis and maintenance[18]. First, *HER2*^{V659E} chimeric mice developed lung tumors only under continuous doxycycline induction. In the absence of doxycycline, no chimeric mice developed lung tumors in over a year ($n = 30$). Instead, lymphomas and sarcomas arose by 1 year of age consistent with the reported Ink4a/Arf null tumor spectrum[19]. Second, a time-course study showed that, within

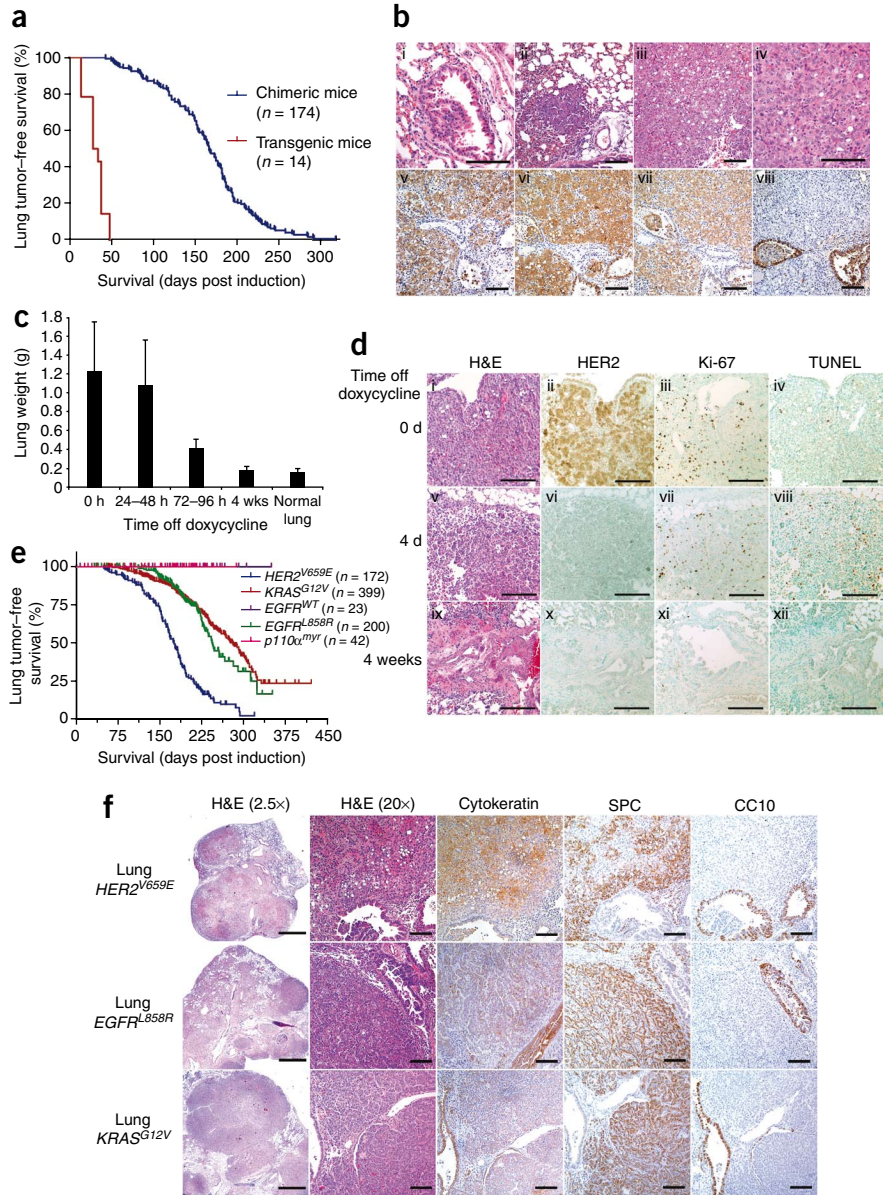


Figure 4 Characterization of lung tumors in lung $HER2^{V659E}$ chimeric mice. (**a**) Kaplan-Meier plot of survival of the germline transgenic and the chimeric lung $HER2^{V659E}$ models. (**b**) Characterization of adenocarcinomas in $HER2^{V659E}$-dependent chimeras. i. Association of an emerging tumor with terminal bronchioles after 5 weeks of doxycycline induction. ii. An adenoma developed within the context of normal lung tissue. iii, iv. Low (20×) and high (40×) power view of an adenocarcinoma showing high nuclear/cytoplasm ratios with clear nuclear pleomorphism, aberrant mitosis and invasion into the surrounding normal tissue. v, vi, vii, viii. Immunohistochemical staining for HER2, cytokeratin, SPC and CC10 (CCSP). (**c**) Histogram of lung weight from chimeras that are either on doxycycline, or off doxycycline for 24–48 h, 72–96 h or 4 weeks. The tumors completely regressed after 4 weeks of doxycycline withdrawal. (**d**) Analysis of tumor regression after doxycycline withdrawal. i, ii, iii, iv. Under doxycycline induction, the tumor cells have strong HER2 expression and are actively proliferating (Ki67 staining). No apoptosis was observed in these tumors (TUNEL staining). v, vi, vii, viii. After 4 d of doxycycline withdrawal, tumors are still clearly visible but HER2 expression can no longer be detected in tumor cells. The number of proliferating (Ki67 positive) cells is slightly reduced. The number of apoptotic (TUNEL positive) cells sharply increased. ix, x, xi, xii. After 4 weeks of doxycycline withdrawal, all tumors completely disappeared and were replaced by fibrotic scar tissue. As a consequence, no HER2 expression is detected and few proliferating (Ki67 positive) or apoptotic (TUNEL positive) cells are present. (**e**) Kaplan-Meier plot of survival of five chimeric lung models expressing $HER2^{V659E}$, $KRAS^{G12V}$, $EGFR^{WT}$, $EGFR^{L858R}$ or $p110\alpha^{myr}$, respectively. (**f**) Histology of $HER2^{V659E}$-, $KRAS^{G12V}$-, $EGFR^{L858R}$-dependent tumors. H&E staining showed that most tumors had a sheet-like growth pattern with clear nuclear pleomorphism, features of adenocarcinomas but not adenomas. All tumors are cytokeratin positive (indicating that the tumors originated from epithelial cells), and SPC positive (a type II pneumocyte marker) and CC10 negative (a Clara Cell marker). Scale bars: 0.1 mm (1 mm for 2.5× H&E staining images).

© 2010 Nature America, Inc. All rights reserved.

96 h after doxycycline withdrawal, tumors had lost expression of human $HER2^{V659E}$, decreased proliferation and increased apoptosis, coinciding with a 50% drop in average lung weight (**Fig. 4c**,**d**). After 4 weeks of doxycycline withdrawal, the average lung weight returned to baseline (**Fig. 4c**), and histological inspection showed regional scar tissue in place of tumors (**Fig. 4d**); thus, both tumor initiation and maintenance required continuous expression of $HER2^{V659E}$.

Rapid generation of multiple lung tumor models

The proof of concept $HER2^{V659E}$-lung adenocarcinoma model encouraged us to test in chimeras the various signature oncogenes prevalent in human lung adenocarcinoma. The ES cell line, L19C9, was used as a platform for the introduction of various inducible oncogenes including (i) $PIK3CA^{myr}$ (myristylated-p110α)[20], (ii) $EGFR^{WT}$ and (iii) $EGFR^{L858R}$. An additional model driven by $KRAS^{G12V}$ was made by co-transfection of *CCSP-rtTA*, *tetO-*$KRAS^{G12V}$, *tetO-luciferase* and *PGK-puromycin* into an *Ink4a/Arf*$^{-/-}$ ES line (10E3). Two to three ES cell clones for each model were identified that met the selection criteria outlined above for the $HER2^{V659E}$ chimeric model (**Fig. 2**).

These models have differing oncogenic potential. Notably, the $PIK3CA^{myr}$-driven ($n = 42$) and the $EGFR^{WT}$-driven ($n = 23$) models developed hyperplasia with no progression to malignancy within 9–12 months of doxycycline treatment (data not shown). In contrast, the $KRAS^{G12V}$-driven ($n = 399$) and $EGFR^{L858R}$-driven ($n = 200$) models developed invasive adenocarcinomas within 4–12 months of doxycycline treatment (**Figs. 4e**,**f** and data not shown).

Although the median tumor latencies of $KRAS^{G12V}$ (9 months) and $EGFR^{L858R}$ models (8 months) were 2–3 months longer than that of the $HER2^{V659E}$ model (6 months, **Fig. 4e**), tumors from all three models showed similar histopathological features (**Fig. 4f**). Strong proliferation (numerous Ki67-positive cells) and few apoptotic cells (a lack of TUNEL-positive cells) were seen in all tumors (data not shown).

In summary, the chimeric model approach enabled us to generate a library of lung tumor models in rapid, stepwise fashion; for

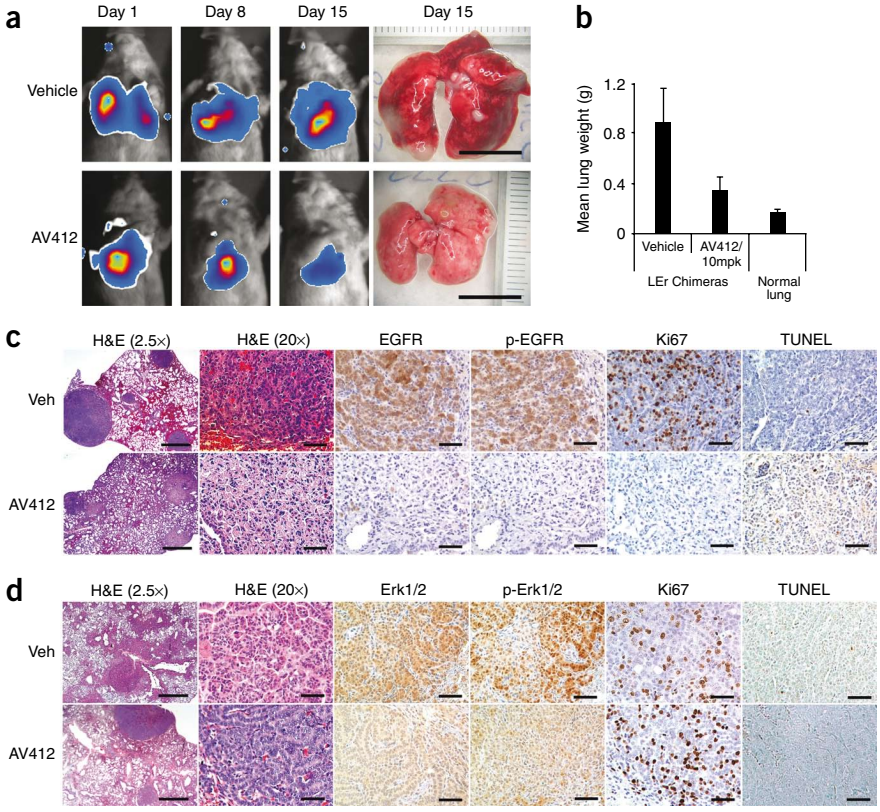


Figure 5 Differential response of $EGFR^{L858R}$- and $KRAS^{G12V}$-dependent tumors to the irreversible EGFR inhibitor AV412. (**a**) Bioluminescent imaging of $EGFR^{L858R}$ chimeric mice before (day 1) and after 1 or 2 weeks (day 8 and day 15) of treatment with AV412 or vehicle control. Treatment with AV412 (10 mpk daily for 2 weeks) resulted in a greater than tenfold drop in bioluminescence whereas the vehicle had no effect. Upon autopsy, lungs in the control group were enlarged with multiple white bulging nodules whereas the AV412-treated lungs were much smaller with flat yellow nodules, consistent with tumor regression. Scale bars, 1 cm. (**b**) Comparison of lung weight of vehicle-treated versus AV412-treated $EGFR^{L858R}$ model chimeric mice. The average lung weight of the AV412-treated mice was almost one-third of that of the control group, much closer to the lung weight of normal mice. (**c**) Histology and immunohistochemical comparison of vehicle treated versus AV412-treated $EGFR^{L858R}$-dependent chimeric mice. After 2 weeks of AV412 treatment (10 mpk daily), almost all tumor cells were eliminated. As a result, very few EGFR or pEGFR-positive cells can be detected in the lung. There were no proliferating cells (Ki67) and some residual apoptosis (TUNEL). (**d**) Histology and immunohistochemical comparison of vehicle-treated versus AV412-treated $KRAS^{G12V}$-dependent chimeric mice. After 2 weeks of AV412 treatment (30 mpk daily), all tumor cells were still viable with active Erk signaling. There was no difference in the number of proliferating cells between vehicle versus AV412-treated tumors. No sign of apoptosis can be seen. Scale bars: 0.1 mm (1 mm for 2.5× H&E staining images).

© 2010 Nature America, Inc. All rights reserved.

example, starting from a wild-type ES cell line, it took 13 months (including a latency of 3–4 months of doxycycline induction) to develop the initial tumors in the first complete doxycycline-inducible $HER2^{V659E}$ model, lines LH33B3 and LH33B6. The subsequent lung models, starting from an ES cell line already containing *Ink4a/Arf*$^{-/-}$; *CCSP-rtTA*; and *tetO-luciferase*, required on average 7–8 months from transfection of the oncogene to the development of observable tumors. Additionally, established model ES cell lines were routinely frozen, stored in liquid nitrogen and recovered without loss of pluripotency or the ability to generate primary tumors in chimeras (**Supplementary Fig. 1d**,**e**). Comparable timelines are not achievable with traditional breeding technologies when starting from establishment of a transgenic line or when combining multiple transgenic lines through breeding (especially if homozygous tumor-suppressor gene modification is desired).

Oncogene-dependent response to EGFR inhibition

To further validate these chimeric mouse models, which featured pronounced similarity to human lung adenocarcinomas (assessed by histopathology and gene expression analysis (**Supplementary Fig. 3**)), and to evaluate their utility in cancer drug development, we treated $EGFR^{L858R}$ model animals with AV412 (an irreversible EGFR kinase inhibitor with sub-nanomolar median effective concentration, or EC_{50}, against the L858R mutation *in vitro*[21]). Chimeras with advanced tumor development were identified by bioluminescent imaging from a cohort of 13 that were on doxycycline for 5 months and showed signs of decreased respiratory function. These animals were divided into vehicle (n = 3) and treatment groups (n = 4) based on equivalent bioluminescent signal intensity and treated orally daily with vehicle (0.5% tragacanth) or AV412 (10 mg/kg body weight), respectively. During the course of drug treatment, bioluminescent imaging was regularly performed to monitor tumor burden *in vivo*, followed by end-of-study necropsy, and histopathological analysis of the lungs. As expected, $EGFR^{L858R}$ mice showed a strong response to AV412 treatment in line with previous reports on various EGFR inhibitors in $EGFR^{L858R}$-driven models[22,23]. The bioluminescent signal in the AV412-treated mice greatly diminished after 1 week of dosing and declined further during the second week; in contrast, bioluminescent signals in the control group remained strong throughout (**Fig. 5a**). Consistent with a decrease in tumor burden, the average lung weight of the AV412-treated group was only slightly higher than normal lungs and one-third the weight of the vehicle-treated group (**Fig. 5b**). Tumor nodules in the vehicle control group were positive for EGFR and Ki67, whereas in the AV412 group, most tumors had regressed, lost their normal architecture (leaving only fibrous scar tissue), and were negative for EGFR and Ki67 (**Fig. 5c**); results were reminiscent of the doxycycline-withdrawal experiment (**Fig. 4d**). Neither vehicle- nor AV412-treated lungs showed signs of apoptosis by TUNEL staining probably due to the absence of *EGFR*-driven, inhibitor-sensitive tumor cells after successful treatment. In contrast, $KRAS^{G12V}$ chimeras (n = 4) with strong bioluminescent signals did not respond to AV412 treatment (**Fig. 5d**) despite EGFR expression detected in at least 50% of the tumors observed immunohistochemically (data not shown). The chimeric models recapitulated published clinical findings that human lung adenocarcinomas bearing mutated *KRAS* or mutated *EGFR* alleles are resistant and sensitive, respectively, to EGFR inhibitors[4,24]. These results provide a proof

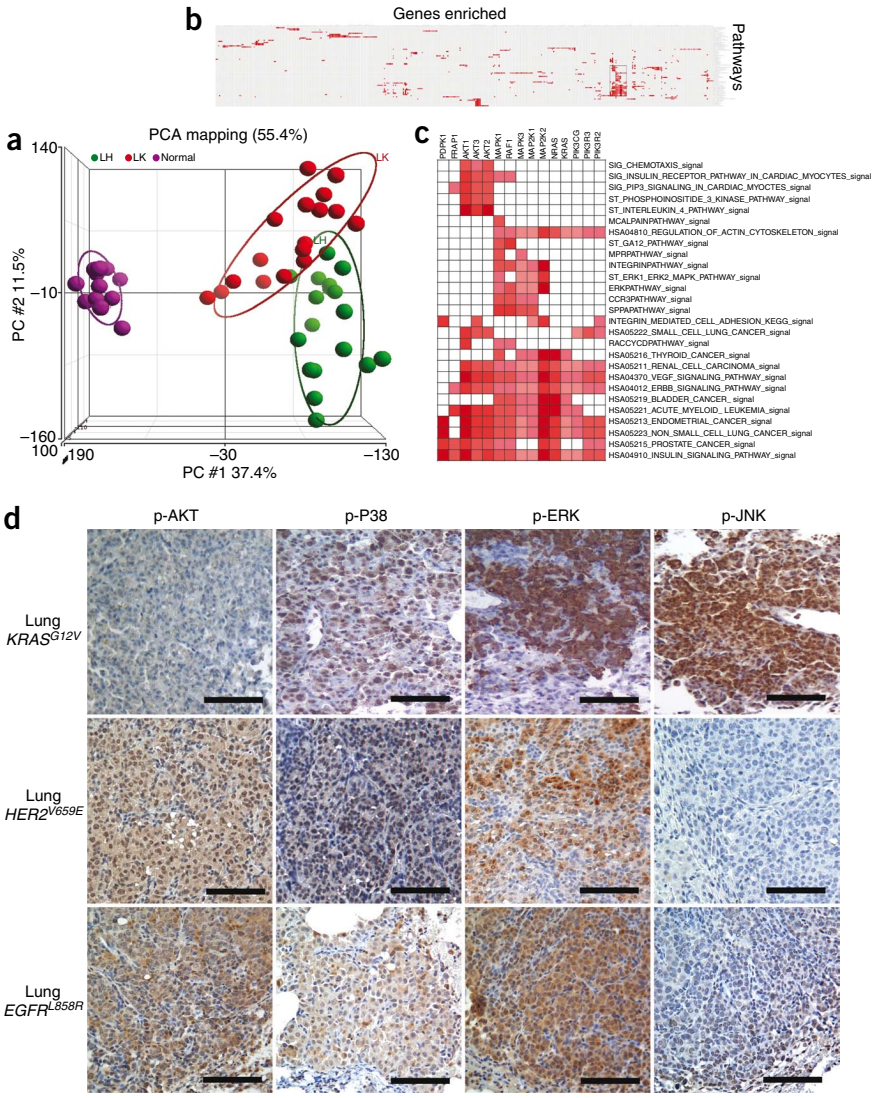


Figure 6 Signaling pathway comparison in $KRAS^{G12V}$-, $HER2^{V659E}$- and $EGFR^{L858R}$-dependent lung tumors. (**a**) Principle component analysis of microarray data on 13 normal lung samples, 18 $KRAS^{G12V}$ and 18 $HER2^{V659E}$ tumor samples. The first three principle components are shown, which account for >50% of the variations. The $KRAS^{G12V}$ and $HER2^{V659E}$ tumors form distinct groups, but they are closer to each other than to normal lung samples. LH, lung $HER2^{V659E}$ chimera model; LK, lung $KRAS^{G12V}$ chimera model. (**b**) Using Gene Set Enrichment Analysis, 72 of 344 canonical pathways in the MySignatureDB were identified as significantly enriched in $HER2^{V659E}$-dependent tumors ($P < 5\%$ and FDR < 0.25). Leading edge analysis, which highlights pivotal genes shared among gene sets, revealed that multiple genes involved in PI3K/AKT and MAPK/ERK signaling were the key components in one-third of these pathways. An independent analysis using Ingenuity Pathway Analysis also confirmed that both the PI3K/AKT pathway and the ERK/MAPK pathway are significantly enriched ($P < 0.00025$ and $P < 0.002$). No canonical pathway was identified as significantly enriched in $KRAS^{G12V}$-dependent tumors by microarray analysis. The black box depicts the region shown in **c**. (**c**) Close-up view of the result of the leading edge analysis, focusing on genes involved in PI3K/AKT and MAPK/ERK signaling underlying one-third of the enriched pathways. (**d**) Comparison of pathway activation in $KRAS^{G12V}$-, $HER2^{V659E}$- and $EGFR^{L858R}$-dependent tumors by immunohistochemistry of phospho-AKT (S473), phospho-P38, phospho-ERK and phospho-JNK. All three tumor models have active ERK and p38 signaling. The $KRAS^{G12V}$-dependent tumors show inactive PI3K/AKT signaling and activated JNK signaling, whereas the $HER2^{V659E}$- and $EGFR^{L858R}$-dependent tumors do not show activated JNK signaling but show activated AKT signaling. Scale bars, 0.1 mm.

© 2010 Nature America, Inc. All rights reserved.

of concept that chimeric mouse lung models can provide effective preclinical tools reflecting clinically observed responses to therapy and that further characterization of these models might assist in the development of additional targets for intervention.

Differential pathway activation in chimera models

To assess common as well as divergent spectrums of activated signaling pathways in the primary tumors, we compared the transcriptomes of $KRAS^{G12V}$-dependent ($n = 18$) and $HER2^{V659E}$-dependent ($n = 18$) lung tumors against normal lung controls ($n = 13$). Principal component analysis demonstrated that although $HER2^{V659E}$- and $KRAS^{G12V}$-dependent tumors were both distinct from normal lung, the identity of the initiating oncogene in these tumors had a clear impact on the transcriptome (**Fig. 6a**). Comparison of microarray data from the chimeric lung KRAS (LK) model and the lox-STOP-lox (LSL)-*KRAS* lung model[25] revealed high concordance between the two models (**Supplementary Fig. 4**). Further analysis by *t*-test identified 965 differentially expressed probes between $KRAS^{G12V}$- and $HER2^{V659E}$-dependent tumors (cutoff, $P < 0.0001$ and >1.4-fold difference). To pinpoint pathways specifically affected by the initiating oncogene, we tested canonical signaling pathways using two different tools: Gene Set Enrichment Analysis and Ingenuity Pathway Analysis. Both analyses revealed that multiple genes involved in PI3K/AKT and MAPK/ERK signaling were specifically enriched in $HER2^{V659E}$-dependent tumors and they constitute key components of canonical signaling pathways differentially regulated by KRAS and HER2 (**Fig. 6b**,**c**).

Next, as a direct measurement of pathway activation, the phosphorylation status of AKT, p44/42 MAPK, p38 MAPK and JNK/SAPK was examined by immunohistochemistry in ten $KRAS^{G12V}$-, $HER2^{V659E}$- and $EGFR^{L858R}$-dependent tumors each (**Fig. 6d**). Consistent with the microarray findings, immunohistochemical staining for phosphorylated AKT (pAKT) at positions S473 and T308 was very pronounced in $HER2^{V659E}$- and $EGFR^{L858R}$-dependent tumors but almost completely absent in $KRAS^{G12V}$-dependent tumors, which was also in agreement with the observation that RTK-dependent tumors are more sensitive to PI3K/mTOR inhibitors than *KRAS*-dependent tumors[5]. Similarly, staining for p38 MAPK phosphorylation was slightly weaker in $KRAS^{G12V}$-dependent tumors than in $HER2^{V659E}$- and $EGFR^{L858R}$-dependent tumors. On the other hand, immunohistochemical staining of phospho-p44/42 MAPK was equally strong in both $KRAS^{G12V}$- and $HER2^{V659E}/EGFR^{L858R}$-dependent tumors, despite the RNA gene expression level difference. Even more interesting, strong staining for

SAPK/JNK phosphorylation was detected in $KRAS^{G12V}$-dependent, but absent in $HER2^{V659E}$- and $EGFR^{L858R}$-dependent tumors. These findings indicate the utility of our preclinical chimeric lung tumor models in elucidating potential nodes for therapeutic intervention *in vivo*.

DISCUSSION

Cancer drug development is characterized by a high rate of failure; ~95% of cancer investigational new drug applications fail during expensive clinical trials owing to toxicity or lack of efficacy[26] and even some of the successes have been modest when compared to the efficacy observed in the preclinical models. This is exemplified by near universal efficacy of antiangiogenic drugs as monotherapies in xenograft models compared to the modest clinical results in most tumor types[27]. Considering the inaccuracy of clinical response prediction and the broadening spectrum of oncology drugs that target activated receptors or pathways in tumors, angiogenesis or interactions between tumor and stroma, preclinical models need to encompass the aforementioned facets of tumor biology and development.

Primary mouse tumor models capture the complex heterotypic interactions between tumor cells and their microenvironments with the added benefit that interactions between the tumor and the host are fully functional (that is, no species-specific incompatibilities between receptors and ligands, for example, mouse HGF and human cMET[28,29]). The availability of an allelic series of key driver mutations would enable cancer drug validation in preclinical settings that are more clinically relevant. State-of-the-art mouse models of cancer that develop tumor types relevant to human disease with short latencies include adeno-*Cre*-activated $Kras^{G12V}$-driven lung adenocarcinomas[25], doxycycline-[18] and 4-hydroxytamoxifen-inducible[30] melanomas, and RCAS-tva retroviral infection *in vivo*[31,32]. The virally triggered models share the advantage of spatio-temporal regulation of the oncogenic insult, but require additional manipulations of the mice to induce oncogenesis. Doxycycline-inducibility provides simplicity of oncogene activation, but combination with germline models usually results in activation throughout the target tissue. Another established route to making complex, orthotopic primary tumor models has been retroviral transduction of oncogenes into lineage-restricted progenitors followed by transplantation into host animals (e.g., hematopoietic lineages[33,34], liver[35] and mammary gland[36,37]). This approach is powerful but limited by several factors, most notably the ability to isolate progenitor cells from the desired target tissue and to reintroduce the manipulated cells effectively into the host animal.

In this study, we implemented an approach based on ES cells and mouse chimeras to engineer multiple cancer alleles and generate genetically sophisticated mouse models of lung adenocarcinoma. These models provide (i) more advanced tumor development compared to a traditional transgenic *HER2*-driven model; (ii) the ability to create a model library employing oncogenes relevant to humans; (iii) models for *in vivo* compound testing on primary tumors; and (iv) a tool for discovering variation in downstream signaling pathways, which may lead to identification of drug targets for tumor types that are resistant to current targeted therapies (e.g., the resistance of $KRAS^{G12V}$-, $EGFR^{L585R,T790M}$-driven (ref. 7) and *EGFR* exon 20 insertion[38]–driven tumors to EGFR inhibitors). The ES cell-chimeric method is capable of targeting all tissue and cell types through varied combinations of tissue-specific promoters, drug-inducible oncogenes and virally introduced recombinase genes to activate or inactivate genetic elements in the chimeric tissue. Other advantages relative to germline transgenic models are the time and cost of maintaining breeding colonies to intercross the desired genetic modifications; in contrast, by creating an array of genetically engineered ES cell lines, we are capable of rapidly producing a large, complex tumor model library containing multiple genetic modifications (**Fig. 2**).

As a proof of concept that chimera-based models could serve as accurate predictors of clinical response, two of the chimeric lung adenocarcinoma models ($EGFR^{L858R}$ and $KRAS^{G12V}$) exhibited a differential response to an irreversible EGFR inhibitor similar to observed clinical responses to reversible EGFR-inhibitors for non-small cell lung cancer patients that harbored mutations in *EGFR* and *KRAS*[4,24,39,40]. The $EGFR^{L858R}$-driven tumors in chimeras were sensitive to EGFR inhibition, but treatment was not maintained long enough for resistance to appear an important clinical complication not yet validated in our models.

In addition, the variety of primary tumors and uniformity of strain background (129SvEv) from our ES cell models library allows us to compare and contrast biological traits (e.g., the transcriptome, signaling pathway activation and drug sensitivity). For example, the microarray and immunohistochemistry data from the *KRAS* and *HER2*/*EGFR* lung adenocarcinoma models showed preferences for downstream pathway activation unique to each initiating oncogene. The observation that *KRAS* tumors signal through ERK and JNK pathways suggests that combination therapy against these two pathways may provide survival benefits for individuals with *KRAS* mutations, whereas the reliance of tumors harboring activated *EGFR* or *HER2* on the PI3K pathway suggests that inhibition of both the ERBB kinase as well as the PI3K pathway could result in tumor growth inhibition[41]. Notably, the differential engagement of signaling pathways between the models are consistent with the response observed of mutant *KRAS* and *EGFR* tumors to PI3K inhibitors in other preclinical studies[5].

In summary, despite the challenge of working with primary tumor models, our initial studies with chimera models have demonstrated their utility in modeling cancer, emulating tumor drug response and identifying biomarkers of tumor biology driven by differing oncogenes.

METHODS

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

Accession code. Gene Expression Omnibus[42] microarray data have been deposited under GEO Series accession number GSE18784.

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

ACKNOWLEDGMENTS

We thank S. Kollipara, I. Agekeum, Q. Xiao, D. Potz, K. Jesmer, Q. Shen, J. Brodeur and A. Cooper for their expert technical help. We also thank K. Garland, P. Bains-Vallee and S. Perry for excellent animal research support. The *CCSP-rtTA* construct was provided by J. Whitsett at the University of Cincinnati. Finally, we are grateful to R. O'Hagan for helpful discussions and critical reading of the manuscript.

AUTHOR CONTRIBUTIONS

Y.Z. established ES cell lines, analyzed the expression data and participated in data interpretation. W.M.R. established ES cell lines, chimeras and participated in data interpretation. T.Z. performed and analyzed immunohistochemistry. A.B. phenotyped and analyzed all mice. S.R. cloned all vectors and targeting constructs. R.R. established chimeric mice and phenotyped mice. J.-K.W. performed and analyzed RT-PCR and luciferase assays. J.W.H. established HER2 transgenic mice. M.B. performed all pathology analysis. L.C., M.I.C., S.C.C., R.A.D. and M.O.R. participated in the planning and data interpretation. T.J. and J.H. conceived the chimera model and participated in planning and data analysis. W.M.R., Y.Z., R.A.D. and J.H. wrote the manuscript and all authors edited it.

COMPETING INTERESTS STATEMENT

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

© 2010 Nature America, Inc. All rights reserved.
npg

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

1. Weinberg, R.A.. *The Biology of Cancer* (Garland Science, New York, 2007).
2. Forbes, S.A. *et al.* The catalogue of somatic mutations in cancer (COSMIC). in *Current Protocols in Human Genetics* (Wiley, Hoboken, New Jersery, USA, 2008).
3. Kumar, A., Petri, E.T., Halmos, B. & Boggon, T.J. Structure and clinical relevance of the epidermal growth factor receptor in human cancer. *J. Clin. Oncol.* **26**, 1742–1751 (2008).
4. Pao, W. *et al.* KRAS mutations and primary resistance of lung adenocarcinomas to gefitinib or erlotinib. *PLoS Med.* **2**, e17 (2005).
5. Engelman, J.A. *et al.* Effective use of PI3K and MEK inhibitors to treat mutant Kras G12D and PIK3CA H1047R murine lung cancers. *Nat. Med.* **14**, 1351–1356 (2008).
6. Pao, W. *et al.* Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. *PLoS Med.* **2**, e73 (2005).
7. Kobayashi, S. *et al.* EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.* **352**, 786–792 (2005).
8. Bell, D.W. *et al.* Inherited susceptibility to lung cancer may be associated with the T790M drug resistance mutation in EGFR. *Nat. Genet.* **37**, 1315–1316 (2005).
9. Gorre, M.E. *et al.* Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. *Science* **293**, 876–880 (2001).
10. Engelman, J.A. *et al.* MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. *Science* **316**, 1039–1043 (2007).
11. Sergina, N.V. *et al.* Escape from HER-family tyrosine kinase inhibitor therapy by the kinase-inactive HER3. *Nature* **445**, 437–441 (2007).
12. Stommel, J.M. *et al.* Coactivation of receptor tyrosine kinases affects the response of tumor cells to targeted therapies. *Science* **318**, 287–290 (2007).
13. Kosaka, T. *et al.* Mutations of the epidermal growth factor receptor gene in lung cancer: biological and clinical implications. *Cancer Res.* **64**, 8919–8923 (2004).
14. Suzuki, M. *et al.* Exclusive mutation in epidermal growth factor receptor gene, HER-2, and KRAS, and synchronous methylation of nonsmall cell lung cancer. *Cancer* **106**, 2200–2207 (2006).
15. Moody, S.E. *et al.* Conditional activation of Neu in the mammary epithelium of transgenic mice results in reversible pulmonary metastasis. *Cancer Cell* **2**, 451–461 (2002).
16. Fisher, G.H. *et al.* Induction and apoptotic regression of lung adenocarcinomas by regulation of a K-Ras transgene in the presence and absence of tumor suppressor genes. *Genes Dev.* **15**, 3249–3262 (2001).
17. Perl, A.K., Tichelaar, J.W. & Whitsett, J.A. Conditional gene expression in the respiratory epithelium of the mouse. *Transgenic Res.* **11**, 21–29 (2002).
18. Chin, L. *et al.* Essential role for oncogenic Ras in tumour maintenance. *Nature* **400**, 468–472 (1999).
19. Serrano, M. *et al.* Role of the INK4a locus in tumor suppression and cell mortality. *Cell* **85**, 27–37 (1996).
20. Singh, B. *et al.* p53 regulates cell survival by inhibiting PIK3CA in squamous cell carcinomas. *Genes Dev.* **16**, 984–993 (2002).
21. Suzuki, T. *et al.* Pharmacological characterization of MP-412 (AV-412), a dual epidermal growth factor receptor and ErbB2 tyrosine kinase inhibitor. *Cancer Sci.* **98**, 1977–1984 (2007).
22. Kobayashi, S. *et al.* An alternative inhibitor overcomes resistance caused by a mutation of the epidermal growth factor receptor. *Cancer Res.* **65**, 7096–7101 (2005).
23. Politi, K. *et al.* Lung adenocarcinomas induced in mice by mutant EGF receptors found in human lung cancers respond to a tyrosine kinase inhibitor or to down-regulation of the receptors. *Genes Dev.* **20**, 1496–1510 (2006).
24. Eberhard, D.A. *et al.* Mutations in the epidermal growth factor receptor and in KRAS are predictive and prognostic indicators in patients with non-small-cell lung cancer treated with chemotherapy alone and in combination with erlotinib. *J. Clin. Oncol.* **23**, 5900–5909 (2005).
25. Sweet-Cordero, A. *et al.* An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. *Nat. Genet.* **37**, 48–55 (2005).
26. Sharpless, N.E. & Depinho, R.A. The mighty mouse: genetically engineered mouse models in cancer drug development. *Nat. Rev. Drug Discov.* **5**, 741–754 (2006).
27. Quesada, A.R., Medina, M.A. & Alba, E. Playing only one instrument may be not enough: limitations and future of the antiangiogenic treatment of cancer. *Bioessays* **29**, 1159–1168 (2007).
28. Bhargava, M. *et al.* Scatter factor and hepatocyte growth factor: activities, properties, and mechanism. *Cell Growth Differ.* **3**, 11–20 (1992).
29. Rong, S. *et al.* Tumorigenicity of the met proto-oncogene and the gene for hepatocyte growth factor. *Mol. Cell. Biol.* **12**, 5152–5158 (1992).
30. Dankort, D. *et al.* Braf(V600E) cooperates with Pten loss to induce metastatic melanoma. *Nat. Genet.* **41**, 544–552 (2009).
31. Lewis, B.C., Klimstra, D.S. & Varmus, H.E. The c-myc and PyMT oncogenes induce different tumor types in a somatic mouse model for pancreatic cancer. *Genes Dev.* **17**, 3127–3138 (2003).
32. Pao, W., Klimstra, D.S., Fisher, G.H. & Varmus, H.E. Use of avian retroviral vectors to introduce transcriptional regulators into mammalian cells for analyses of tumor maintenance. *Proc. Natl. Acad. Sci. USA* **100**, 8764–8769 (2003).
33. Daley, G.Q., Van Etten, R.A. & Baltimore, D. Induction of chronic myelogenous leukemia in mice by the P210bcr/abl gene of the Philadelphia chromosome. *Science* **247**, 824–830 (1990).
34. Heard, J.M., Roussel, M.F., Rettenmier, C.W. & Sherr, C.J. Multilineage hematopoietic disorders induced by transplantation of bone marrow cells expressing the v-fms oncogene. *Cell* **51**, 663–673 (1987).
35. Zender, L. *et al.* Identification and validation of oncogenes in liver cancer using an integrative oncogenomic approach. *Cell* **125**, 1253–1267 (2006).
36. Edwards, P.A., Hiby, S.E., Papkoff, J. & Bradbury, J.M. Hyperplasia of mouse mammary epithelium induced by expression of the Wnt-1 (int-1) oncogene in reconstituted mammary gland. *Oncogene* **7**, 2041–2051 (1992).
37. Wu, M. *et al.* Dissecting genetic requirements of human breast tumorigenesis in a tissue transgenic model of human breast cancer in mice. *Proc. Natl. Acad. Sci. USA* **106**, 7022–7027 (2009).
38. Greulich, H. *et al.* Oncogenic transformation by inhibitor-sensitive and -resistant EGFR mutants. *PLoS Med.* **2**, e313 (2005).
39. Lynch, T.J. *et al.* Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.* **350**, 2129–2139 (2004).
40. Jackman, D.M. *et al.* Impact of epidermal growth factor receptor and KRAS mutations on clinical outcomes in previously untreated non-small cell lung cancer patients: results of an online tumor registry of clinical trials. *Clin. Cancer Res.* **15**, 5267–5273 (2009).
41. Junttila, T.T. *et al.* Ligand-independent HER2/HER3/PI3K complex is disrupted by trastuzumab and is effectively inhibited by the PI3K inhibitor GDC-0941. *Cancer Cell* **15**, 429–440 (2009).
42. Edgar, R., Domrachev, M. & Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).

© 2010 Nature America, Inc. All rights reserved.

ONLINE METHODS

All animal work was carried out according to AVEO's Institutional Animal Care and Use Committee guidelines and in accordance with AAALAC (Association for Assessment and Accreditation of Laboratory Animal Care) policies and certification.

DNA constructs. The *CCSP-rtTA* construct was provided by J. Whitsett at the University of Cincinnatti[17]. The *TetO-KRAS^{G12V}* construct was provided by L.C. *TetO-HER2*, *TetO-EGFR* and *TetO-PIK3CA* were generated by subcloning resequenced full-length human cDNAs into p*TetO-pA*. All mutations were made by mutagenesis using the QuickChange II XL Kit from Stratagene and verified by sequencing. *TetO-luciferase* was generated by subcloning firefly *luciferase* into p*TetO-pA*. The *Ink4a* conditional targeting vector was provided by R.A.D[19].

Ink4a double knockout. ES clone H12, with one *Ink4a* allele targeted by homologous recombination[19], was obtained from the Dana-Farber Cancer Institute. H12 was first modified by Cre-mediated recombination to generate C23. The second allele of C23 was targeted again with the same targeting vector, resulting in clone 5F8, which gave rise to clone 10E3 (*Ink4a$^{-/-}$*) after Cre-mediated excision of the second targeted allele. The ES clones were injected into blastocysts to confirm pluripotency by forming chimeras with chimerism by coat color >50%.

Chimeric lung *HER2^{V659E}*, *EGFR^{WT}*, *EGFR^{L858R}* and *PIK3CA^{myr}* models. Clone 10E3 was transfected with *CCSP-rtTA*, *tetO-luciferase* and *PGK-puro*. Puromycin-resistant clones were genotyped and subjected to Bright-Glo *in vitro* bioluminescent assay. Chimeras made from selected clones were analyzed for rtTA expression and inducible luciferase activity in the lung. One of the clones passing the *in vivo* test, L19C9, was co-transfected with the inducible oncogene *tetO-HER2^{V659E}* and *PGK-hygro* to generate lung *HER2* model clones, or with an inducible oncogene (*tetO-EGFR^{WT}*, *tetO-EGFR^{L858R}* or *tetO-PIK3CA^{myr}* (myristylated-p110α)) and *PGK-neo* to generate the respective model ES clones. Chimeras from these ES cell clones were analyzed for the expression of the human oncogenes in the lung by RT-PCR.

Chimeric lung *KRAS^{G12V}* model. The lung *KRAS^{G12V}* model ES clones were generated by co-transfecting 10E3 with *CCSP-rtTA*, *tetO-KRAS^{G12V}*, *tetO-luciferase* and *PGK-puromycin*. Puromycin-resistant clones were subcloned and PCR genotyped. The inducibility of the *KRAS^{G12V}* and *luciferase* in positive ES clones was tested by RT-PCR and bioluminescent assay, respectively, after transiently transfecting the cells with *PGK-tTA*. Chimeras from these clones were analyzed for the expression of *KRAS* in the lung by RT-PCR.

Chimeric mice production. Confluent ES cell cultures were washed twice with PBS and trypsinized. Equal volume of embryonic stem cell media (ESCM) without leukemia inhibitory factor (LIF) was added and cells were triturated several times to make single-cell suspension. We added 5 ml of ESCM-LIF to the cells, spun them down and then resuspended them in 4 ml of ESCM-LIF, plated them on a 60-mm tissue culture plate and incubated them for 30 min to remove feeder cells. Cells in suspension were spun down again and resuspended in minimal volume for injection. Typically 10–15 ES cells were injected into one C57BL/6 blastocyst (3.5 d post coitum (d.p.c.)) and 12–16 blastocysts were transferred into the uterine horns of a pseudo pregnant Swiss Webster female (2.5 d.p.c.). Pups are born 17 d after the transfer and chimeras can be identified by their agouti coat color. Chimeras are housed in a specific-pathogen-free animal facility. Chimeras are weaned 3–4 weeks after birth and ear tagged. Doxycycline was delivered in either water containing 2 mg/ml doxycycline in 10 mg/ml sucrose in dark bottles changed twice weekly or in food pellets containing 2,500 p.p.m. of doxycycline.

Chimerism determination. The chimerism of ES cell models was evaluated by coat color chimerism and DNA chimerism in the lung. Coat color chimerism was judged by the percentage of agouti-colored fur for each chimera after weaning. This subjective evaluation was carried out by one person for all the chimeras used in this publication. Chimerism in the lung was determined by hybridizing DNA extracted from the lung to a single radiolabeled *Ink4a* probe. When digested with BamHI, this probe detects a 6-kb band from WT tissue coming from the host blastocyst and a 4-kb band from *Ink4a/Arf$^{-/-}$* null ES cell–derived tissue. The chimerism was calculated based on quantification of these two bands. *Ink4a/Arf$^{+/-}$* ES cell DNA was used as a control and gave a reading of 48–50%.

Real-time RT-PCR. Real-time quantitative (q)RT-PCR was carried out in 384-well format on ABI 7900HT systems using QuantiTech SYBR green RT-PCR kit (Qiagen). Each sample was analyzed in quadruplicate reactions with 50 ng of RNA used in each 20 μl reaction. *HPRT* was used as loading control. The relative expression of each human oncogene was expressed as fold change over Human Universal Reference RNA (Stratagene), which was calculated using the delta Ct method in the SDS 2.1 software.

Immunohistochemistry. Formalin-fixed and paraffin-embedded slides were first washed in antigen retrieval buffer, and then incubated with primary antibody at 4 °C overnight. The next day, after washing, the samples were incubated with horseradish peroxidase–conjugated secondary antibody for 2 h. They were then washed and incubated with 3,3′-diaminobenzidine tetrahydrochloride and H_2O_2 under close monitoring for color development. The CC10 and SPC antibodies were purchased from Chemicon. The Ki67 antibody was purchased from NeoMarkers. The HER2, EGFR, phospho-EGFR, phospho-AKT, phospho-ERK, phospho-p38, and phospho-JNK antibodies were purchased from Cell Signaling. The TUNEL assay was carried out with a kit from Sigma.

Luciferase assay. ES cells grown on 96-well plate were lysed in 1× Bright-Glo (Stratagene) and imaged using a Fuji Luminometer and signals quantified using ImageQuant 4.0. Lung tissue was first homogenized in PBS using a Sonicator. 100 μl of the homogenate was transferred to a 96-well plate and 100 μl of 2× Bright-Glo was added and mixed and the plate was imaged the same way.

Bioluminescent Imaging. Mice were injected with 250 mg of luciferin (Invitrogen) and anesthetized with Isoflurane. The mice were then transferred into the chamber of the NightOwl imaging system under anesthesia. Mice were typically imaged 8–10 min after injection. The lung chimeras were typically imaged for 5–10 min under the maximum sensitivity setting.

Expression profiling by microarray. RNAs extracted from tumors were labeled, mixed with labeled mouse universal reference RNA (Stratagene), and hybridized on Agilent Mouse Genome two-color microarray. Each sample was hybridized on 2 chips with dye swap. The chips were scanned with an Agilent scanner. Feature extraction and data preprocessing were carried out using Agilent software. The principal component analysis and *t*-test were carried out using Partek Genomic Suite 6.1. Gene set enrichment analysis (GSEA) and leading edge analysis were done using GSEA 2.0 and MySignatureDB 2.5. (Broad Institute).

AV412 treatment. AV412 was formulated fresh daily in 0.5% tragacanth. Vehicle or the drug was given to tumor-bearing chimeras using oral gavage. Treatment was well tolerated as both the vehicle and treatment groups showed <5% weight loss (data not shown). Mice were treated for 2 weeks at the relevant dose and at the end of treatment, mice were euthanized and the lungs taken for analysis.

© 2010 Nature America, Inc. All rights reserved.

nature biotechnology

Reengineering a receptor footprint of adeno-associated virus enables selective and systemic gene transfer to muscle

Aravind Asokan[1,2], Julia C Conway[1], Jana L Phillips[1], Chengwen Li[1], Julia Hegge[4], Rebecca Sinnott[1], Swati Yadav[1], Nina DiPrimio[1], Hyun-Joo Nam[3], Mavis Agbandje-McKenna[3], Scott McPhee[5], Jon Wolff[4] & R Jude Samulski[1]

© 2010 Nature America, Inc. All rights reserved.

Reengineering the receptor footprints of adeno-associated virus (AAV) isolates may yield variants with improved properties for clinical applications. We generated a panel of synthetic AAV2 vectors by replacing a hexapeptide sequence in a previously identified heparan sulfate receptor footprint with corresponding residues from other AAV strains. This approach yielded several chimeric capsids displaying systemic tropism after intravenous administration in mice. Of particular interest, an AAV2/AAV8 chimera designated AAV2i8 displayed an altered antigenic profile, readily traversed the blood vasculature, and selectively transduced cardiac and whole-body skeletal muscle tissues with high efficiency. Unlike other AAV serotypes, which are preferentially sequestered in the liver, AAV2i8 showed markedly reduced hepatic tropism. These features of AAV2i8 suggest that it is well suited to translational studies in gene therapy of musculoskeletal disorders.

New viral strains constantly evolve in nature through iterative mutagenesis[1,2]. The breadth of tissue tropism displayed by various AAV isolates, such as AAV8 and AAV9, is beneficial for gene transfer by systemic delivery[3–6]. In some cases, however, it would be desirable to direct homing of AAV vectors to specific organs. All naturally occurring AAV serotypes and variants tested to date have a propensity to accumulate within the liver, albeit with varying efficiency[4]. Consequently, strategies to redirect AAV capsids from the liver to target organs would be very useful from a clinical standpoint. Tissue-specific promoters and, more recently, microRNA-based gene regulation strategies, have been used to sharply segregate gene expression patterns among different tissue types[7,8]. However, such regulatory strategies do not preclude sequestration of AAV vector genomes in off-target organs such as the liver after systemic administration.

To develop AAV vectors with improved tropism for clinical applications, we reengineered the heparan sulfate receptor footprint on the AAV2 capsid surface using information available from structural studies, including crystallographic data and cryo-electron microscope analysis of AAV capsids and their cognate receptors[9–12]. The heparan sulfate footprint on the AAV2 capsid consists of the basic amino acid residues R484, R487, K527, K532, R585 and R588, which form a continuous basic patch[10–12]. R585 and R588, located within the so-called GH loop, form the inner walls of the spikes located on the icosahedral threefold axis (**Fig. 1a**), and other residues occupy the floor surrounding these regions[12]. Mutation of either R585 or R588 disrupts the basic cluster and abolishes heparan sulfate binding[10,11].

Using site-directed mutagenesis, we substituted the hexapeptide motif 585-RGNRQA-590, which contains R585 and R588, with corresponding amino acids from different AAV serotypes and nonhuman primate isolates (**Fig. 1b**,**c**) to generate a series of AAV2 inner loop (AAV2i) mutants. Earlier studies established that mutating R585 and/or R588 on the AAV2 capsid to C, M, A or E is sufficient to attenuate heparan sulfate binding[10,11]. In the current study, AAV2i mutants containing Q, A, S or N in position 585 and T, N, A or G in position 588 were also unable to bind heparan sulfate under physiological conditions, as demonstrated by affinity column binding assays (**Supplementary Fig. 1**). In general, titers of all AAV2i mutants were similar to that of the parental AAV2, and their efficiency at transducing various cell types *in vitro* was reduced by several orders of magnitude (**Supplementary Fig. 2**).

Using live animal bioluminescence imaging, we studied vector biodistribution in normal BALB/C mice after intravenous administration at low dosage. One week after administration, most AAV2i mutants were deficient in transduction as evidenced by low bioluminescent signal (**Fig. 1b**). A notable exception was AAV2i8, which displayed a systemic transduction profile (**Fig. 1b**) regardless of the duration of gene expression or the intravenous route of administration (tail or portal vein; **Supplementary Fig. 3**).

Based on the above observations with AAV2i8 containing a 585-QQNTAP-590 motif, we tested several AAV2i mutants with 585-QXXTXP-590 or 585-NXXTXP-590 motifs derived from other strains of AAV. AAV2i mutants with residues Q/N585, T588 and P590 showed systemic transduction profiles similar to that of the AAV8 control (**Fig. 1c**). In contrast, the AAV2 control showed a greater tropism for liver, as established earlier[4]. The higher transduction efficiency of AAV2i8 compared with AAV2i10, AAV2i11, AAV2irh.2 and AAV2irh.38 highlights the subtle

[1]Gene Therapy Center and [2]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [3]Macromolecular Structure Group, University of Florida, Gainesville, Florida, USA. [4]Mirus BioCorporation, Madison, Wisconsin, USA. [5]Asklepios Biopharmaceutical, Inc., Chapel Hill, North Carolina, USA. Correspondence should be addressed to A.A. (aravind_asokan@med.unc.edu).

Received 6 August; accepted 4 December; published online 27 December 2009; doi:10.1038/nbt.1599

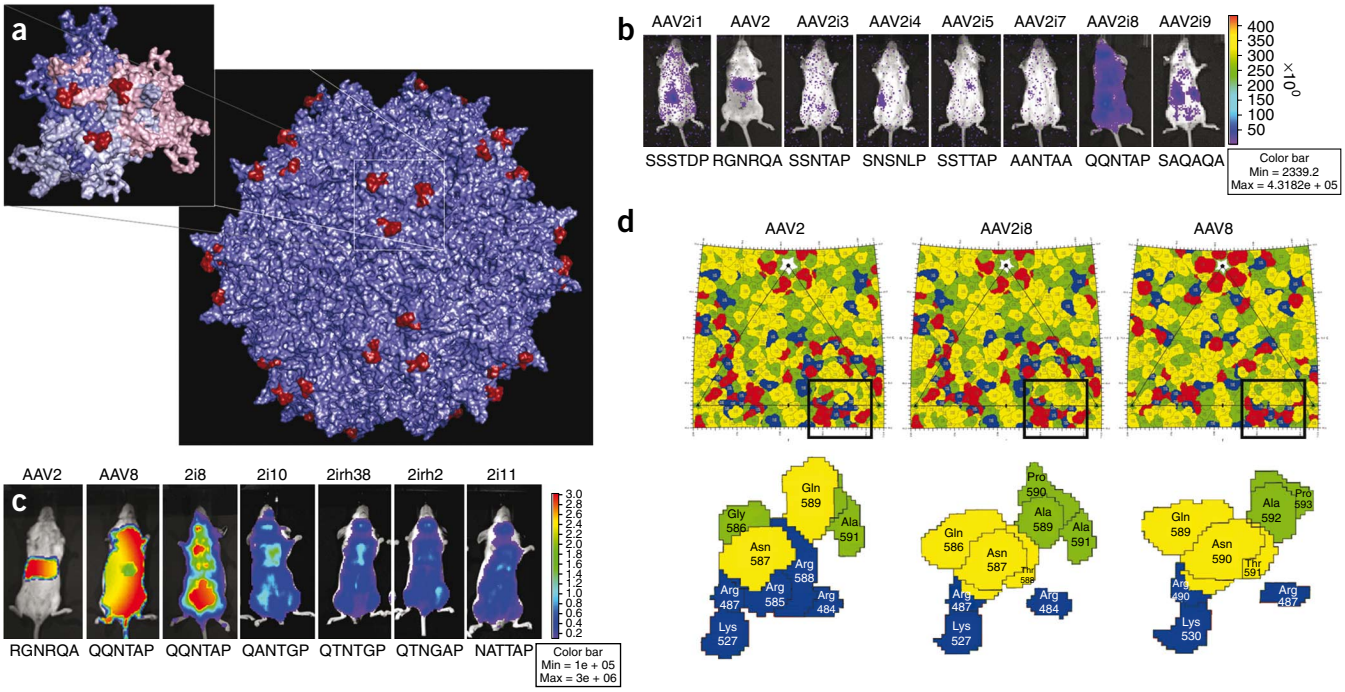


Figure 1 Structure-function correlates of AAV2i vectors with reengineered receptor footprints. (**a**) Three-dimensional structural model of the AAV2 capsid highlighting the 585–590 region containing basic residues implicated in heparan sulfate binding. Inset shows VP3 trimer, with residues 585-RGNRQA-590 located on the innermost surface loop highlighted in red. VP3 monomers are colored salmon, blue, and gray. Images were rendered using Pymol. (**b**) Representative live animal bioluminescent images of luciferase transgene expression profiles in BALB/c mice ($n = 3$) injected intravenously (tail vein) with AAV2i CMV-Luc vectors (dose 1×10^{10} vg in 200 µl PBS). Photographs and bioluminescent images were obtained at 1 week after injection. The overlay demonstrates decreased transduction efficiency for most AAV2i mutants with the exception of AAV2i8. Bioluminescence scale ranges from 0–4×10^5 relative light units (photons/sec/cm^2). Residues within the 585–590 region in each AAV2i mutant are indicated below corresponding mouse image data. (**c**) Representative live animal bioluminescent images of luciferase transgene expression profiles in BALB/c mice ($n = 3$) injected intravenously (tail vein) with AAV2, AAV8, AAV2i8 and structurally related AAV2i mutants (dose 1×10^{11} vg in 200 µl PBS) packaging the CBA (chicken beta actin)-Luc cassette. All AAV2i mutants show a systemic transduction profile similar to that of AAV8, with AAV2i8 showing enhanced transduction efficiency. Bioluminescence scale ranges from 0–3×10^6 relative light units (photons/sec/cm^2). Residues within 585–590 region in each AAV2i mutant is indicated below corresponding mouse image data. (**d**) Comparison of AAV2, AAV2i8 and AAV8 capsid surface residues based on schematic "Roadmap" projections. A section of the asymmetric unit surface residues on the capsid crystal structures of AAV2 and AAV8, as well as a model of AAV2i8, are shown. Close-up views of the heparan sulfate binding region and residues 585–590 reveal a chimeric footprint on the AAV2i8 capsid surface. Red, acidic residues; blue, basic residues; yellow, polar residues; green, hydrophobic residues. Each residue is shown with a black boundary and labeled with VP1 numbering based on the AAV2 capsid protein sequence.

© 2010 Nature America, Inc. All rights reserved.

synergy between residues within the hexapeptide motif in conferring systemic tissue tropism. Notably, the 585-QQNTAP-590 motif did not result in systemic tropism when incorporated into the corresponding region on AAV1 or AAV3 capsids (**Supplementary Fig. 4**). Taken together, these results highlight the complexity of the structural coordinates required to attain an atypical systemic transduction profile.

To examine the surface footprint of AAV2i8, we generated model surface maps of this mutant and of parental AAV2 and AAV8 capsids using stereographic roadmap projections, which allow simultaneous projection of amino acids, charge distribution and capsid surface topology onto a two-dimensional surface[13]. Substitution of 585-RGNRQA-590 with QQNTAP results in disruption of the continuous basic patch (blue residues) formed by the cluster of arginine and lysine residues (**Fig. 1d**). In addition, our model of the AAV2i8 footprint shows an overall chimeric distribution of amino acid residues with respect to AAV2 and AAV8 (refs. 9,14). The chimeric nature of AAV2i8 is corroborated in the observation that these capsids were only modestly neutralized when exposed to anti-AAV2 serum or human serum (**Supplementary Tables 1** and **2**). Thus, reengineering receptor footprints on AAV capsids can simultaneously alter antigenicity.

Based on its promising transduction profile, the lab-derived AAV2i8 strain was further characterized. We quantified luciferase transgene expression and genome copy numbers in cardiac, skeletal muscle and liver tissue lysates at 2 weeks after vector administration in BALB/C mice. As shown in **Figure 2a**, AAV8 ubiquitously transduced muscle and liver tissue with high efficiency, consistent with the systemic transduction profile in **Figure 1c**. AAV2 also transduced liver preferentially, although less efficiently than AAV8, and showed only modest transduction in muscle tissue. The chimeric AAV2i8 transduced cardiac and skeletal muscle tissue with a high efficiency similar to that of AAV8 and was detargeted from the liver. These findings were supported by data on biodistribution of vector genome copies in muscle and liver tissues determined by Q-PCR (**Fig. 2b**). For AAV2 and AAV8, high amounts of vector genome copies were recovered from liver compared with cardiac or skeletal muscle. For AAV2i8, sequestration in liver was ~40-fold lower compared with AAV2 or AAV8.

Further biodistribution studies confirmed the redirection of AAV2i8 from liver to muscle. AAV2i8 transduced a wide range of muscle groups in the murine forearms and hind legs as well as intercostal, facial and abdominal muscles (**Fig. 2c**). Cardiac and diaphragm muscle were transduced with high efficiency, whereas low levels of vector genome copies were recovered from other major organs, such as brain, lung and spleen. These results distinguish the tissue tropism of the chimeric

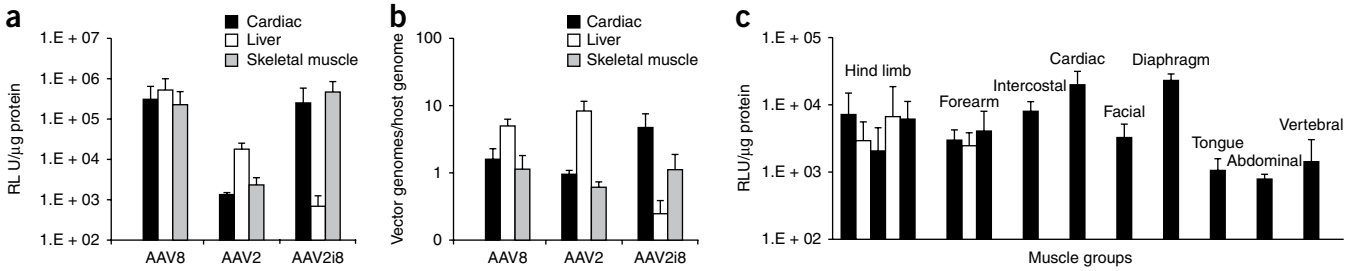


Figure 2 Selective muscle tropism of AAV2i8. (**a**) Quantification of luciferase transgene expression in three major tissues: cardiac (black bars), skeletal muscle (pooled hind limb and abdominal; gray bars) and liver (white bars). Tissue lysates were obtained from BALB/c mice ($n = 3$) at 2 weeks after administration of AAV2, AAV2i8 and AAV8 (dose 1×10^{11} vg, tail vein) and subjected to luminometric analysis. AAV2i8 shows high transduction in cardiac and skeletal muscle and low transduction in liver. Luciferase levels are shown as relative light units normalized to protein levels determined using a Bradford assay. Error bars indicate s.d. (**b**) Vector genome copy numbers (luciferase transgene) in three major tissues: cardiac (black bars), skeletal muscle (pooled hind limb and abdominal; gray bars) and liver (white bars). Host genomic as well vector DNA was extracted from tissue lysates obtained from BALB/c mice ($n = 3$) at 2 weeks after administration of AAV2, AAV2i8 and AAV8 (dose 1×10^{11} vg, tail vein). Host and vector genome copy number were determined by Q-PCR with specific primer sets against the lamin gene and luciferase transgene, respectively. AAV2i8 shows enhanced muscle sequestration and decreased accumulation in liver tissue compared with AAV2 and AAV8. (**c**) Luciferase transgene expression in major muscle sub-groups obtained from BALB/c mice ($n = 3$) at 2 weeks after intravenous administration of AAV2i8 (dose 1×10^{11} vg, tail vein) packaging the CBA (chicken beta actin)-Luc cassette. Tissue lysates from five different muscle groups from the hind limb skeletal muscle (alternating black and white bars), three groups from the forelimb (alternating black and white bars), intercostals, cardiac, facial, diaphragm, tongue, abdominal and vertebral muscle types (black bars) were subjected to luminometric analysis. Luciferase levels are shown as relative light units normalized to protein levels determined by a Bradford assay. Error bars indicate s.d.

AAV2i8 capsid from that of any naturally occurring AAV serotype or isolate characterized thus far (**Supplementary Fig. 5**).

Our results confirm previous findings that attenuation of heparin binding in general can result in liver detargeting and systemic dissemination of AAV2-derived vectors. Earlier studies demonstrated a strong correlation between heparin binding and liver tropism in the case of AAV2 and AAV6 (refs. 10,15). Disruption of the basic receptor footprint through mutagenesis of R585 and/or R588 residues (in AAV2) or K531 (in AAV6) attenuated heparan sulfate binding, which correlated with decreased liver tropism[10,15]. In addition, it was demonstrated that an R484E;R585E AAV2 mutant is detargeted from the liver and retains the ability to transduce muscle tissue with modest efficiency, similar to the parental AAV2 (ref. 16). Although the R484E;R585E AAV2 vector has not shown high transduction efficiency in larger animal models[17] or been compared directly with AAV8 or AAV9, these early studies clearly show the potential to control tissue tropism by manipulating receptor-binding domains. Recently, a novel AAV mutant with cardiac-specific transduction was generated through directed evolution[18]. The laboratory-derived M41 clone displayed a tenfold higher transduction efficiency in cardiac tissue compared with the liver. Whereas similar trends were observed for several of our mutants, an important advantage of AAV2i8 and mutants with a 585-QXXTXP-590 motif is their ability to efficiently transduce not only cardiac muscle but the entire range of muscle groups with a transduction efficiency 2–3 orders of magnitude higher than that observed in the liver (**Fig. 2c** and **Supplementary Fig. 5**). To our knowledge, such an efficient switch in tropism from liver to muscle has not been demonstrated previously.

Next, we used an isolated hind limb perfusion technique[19] to examine the efficiency with which AAV2i8 traverses the blood vessel barrier. AAV2i8 transduced hind limb skeletal muscle as efficiently as AAV8 at low volume of injection, at moderate and high vector dosage (**Fig. 3a–c**). At low vector dose, AAV8 displayed three- to tenfold increases in transduction efficiency at higher volumes of injection. However, AAV2i8 traversed blood vessels and transduced underlying skeletal muscle with high efficiency regardless of the volume of injection.

The atypical tropism of AAV2i8 distinguishes it from natural AAV serotypes 8 and 9 and suggests that engineered AAV vectors can be tailored for specific clinical applications. The mechanism underlying the switch in AAV2i8 tropism from liver to muscle is currently unknown. Our results support the notion that the chimeric vector possesses a unique surface footprint that facilitates specific interactions with receptors distinct from those used by AAV2 and AAV8. Another possible explanation of our findings is that the increased circulation half-life of AAV2i8 allows sequestration in tissues other than the liver through heparan sulfate–independent uptake mechanisms.

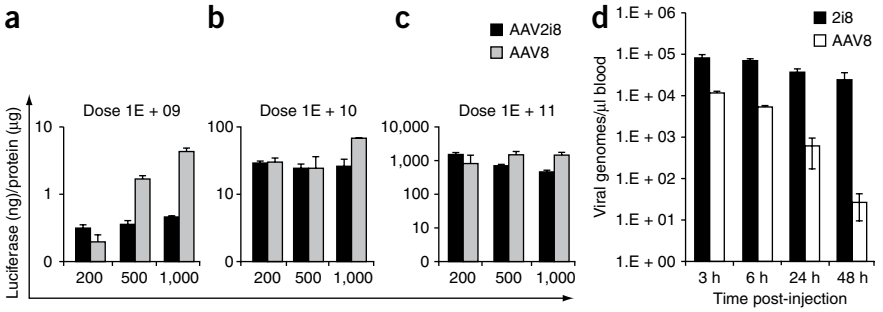


Figure 3 Blood transport profile of AAV2i8. (**a**–**c**) Luciferase transgene expression in pooled skeletal muscle subgroups from right and left hind limb of BALB/c mice ($n = 4$) after isolated perfusion of AAV2i8 (black bars) or AAV8 (gray bars) into each saphenous vein. Tissue lysates prepared after administration of three different doses (1×10^9 (**a**), 1×10^{10} (**b**), 1×10^{11} (**c**) vg) in low (200 µl), medium (500 µl) or high (1 ml) volume of injection were subjected to luminometric analysis. Luciferase levels are shown as relative light units ormalized to protein levels determined using a Bradford assay. (**d**) Vector genome copy numbers recovered from blood at different time intervals after administration through the tail vein ($n = 3$). Whole blood DNA was extracted and analyzed by Q-PCR with primers against the luciferase transgene. AAV2i8 shows prolonged circulation compared with AAV8. Error bars indicate s.d.

© 2010 Nature America, Inc. All rights reserved.

AAV2i8 showed markedly reduced blood clearance and appears to persist well over 48 h in blood (**Fig. 3d**). Moreover, muscle-specific luciferase transgene expression levels increased gradually over the course of several weeks (**Supplementary Fig. 6**). In contrast, AAV8 vector genome copy number rapidly decreased, approaching background levels within the same time period. These results and previous observations that other AAV serotypes with systemic tissue tropism have long circulation half-lives[4] suggest that strategies to manipulate blood circulation time of AAV capsids might afford control over vector tropism.

From the standpoint of vector development and clinical safety, AAV2i8 is an attractive candidate for gene therapy of muscular dystrophies, which requires transduction of a wide range of muscle types after systemic administration[6,20]. The selective muscle tropism of AAV2i8 and its ability to evade sequestration by liver, when exploited in conjunction with transcriptional regulatory elements such as muscle-specific promoters, should allow exquisite control over vector biodistribution as well as cardiac or skeletal muscle–specific transgene expression. Preliminary isolated limb perfusion studies in nonhuman primates comparing AAV2i8 with AAV8 and AAV9 have shown promising results (data not shown) in this regard.

In summary, we have developed a strategy to engineer synthetic AAV strains with atypical transduction profiles. Extrapolation of this approach to receptor binding domains other than the heparan sulfate binding domain and to other natural AAV isolates might yield new chimeric vectors with unique tissue tropisms and antigenicity suitable for translational disease-specific applications.

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 would like to thank the American Heart Association (A.A.; no. 0735637N); National Institute of Arthritis and Musculoskeletal and Skin Diseases (A.A. & R.J.S.; R21AR055712); National Institute of Allergy and Infectious Diseases (R.J.S. and A.A.; R01AI072176); National Heart, Lung, and Blood Institute (A.A.; R01HL089221); Senator Paul Wellstone Center for Muscular Dystrophy (R.J.S.; U54AR056953); National Institute of General Medical Sciences (M.A.-M.; R01GM082946); and Asklepios Biopharmaceutical for research support. We would also like to MirusBio Corp. for assistance with isolated limb perfusion studies.

AUTHOR CONTRIBUTIONS

A.A. conceived the strategy, designed the project and analyzed data. A.A., S.M. and R.J.S. supervised the project and prepared the manuscript. J.C.C. and R.S. built capsid mutants, J.L.P. and C.L. carried out animal experiments, and J.H., S.M. and J.W. designed and carried out isolated limb perfusion studies. S.Y. carried out Q-PCR studies and N.D., H.-J.N. and M.A.-M. carried out molecular modeling studies.

COMPETING INTERESTS STATEMENT

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. Hensley, S.E. *et al.* Hemagglutinin receptor binding avidity drives Influenza A virus antigenic drift. *Science* **326**, 734–736 (2009).
2. Palmenberg, A.C. *et al.* Sequencing and analyses of all known human rhinovirus genomes reveal structure and evolution. *Science* **324**, 55–59 (2009).
3. Gao, G., Vandenberghe, L.H. & Wilson, J.M. New recombinant serotypes of AAV vectors. *Curr. Gene Ther.* **5**, 285–297 (2005).
4. Zincarelli, C., Soltys, S., Rengo, G. & Rabinowitz, J.E. Analysis of AAV serotypes 1–9 mediated gene expression and tropism in mice after systemic injection. *Mol. Ther.* **16**, 1073–1080 (2008).
5. Inagaki, K. *et al.* Robust systemic transduction with AAV9 vectors in mice: efficient global cardiac gene transfer superior to that of AAV8. *Mol. Ther.* **14**, 45–53 (2006).
6. Wang, Z. *et al.* Adeno-associated virus serotype 8 efficiently delivers genes to muscle and heart. *Nat. Biotechnol.* **23**, 321–328 (2005).
7. Wang, B. *et al.* Construction and analysis of compact muscle-specific promoters for AAV vectors. *Gene Ther.* **15**, 1489–1499 (2008).
8. Brown, B.D. *et al.* Endogenous microRNA can be broadly exploited to regulate transgene expression according to tissue, lineage and differentiation state. *Nat. Biotechnol.* **25**, 1457–1467 (2007).
9. Xie, Q. *et al.* The atomic structure of adeno-associated virus (AAV-2), a vector for human gene therapy. *Proc. Natl. Acad. Sci. USA* **99**, 10405–10410 (2002).
10. Kern, A. *et al.* Identification of a heparin-binding motif on adeno-associated virus type 2 capsids. *J. Virol.* **77**, 11072–11081 (2003).
11. Opie, S.R., Warrington, K.H. Jr., Agbandje-McKenna, M., Zolotukhin, S. & Muzyczka, N. Identification of amino acid residues in the capsid proteins of adeno-associated virus type 2 that contribute to heparan sulfate proteoglycan binding. *J. Virol.* **77**, 6995–7006 (2003).
12. Levy, H.C. *et al.* Heparin binding induces conformational changes in Adeno-associated virus serotype 2. *J. Struct. Biol.* **165**, 146–156 (2009).
13. Xiao, C. & Rossmann, M.G. Interpretation of electron density with stereographic roadmap projections. *J. Struct. Biol.* **158**, 182–187 (2007).
14. Nam, H.J. *et al.* Structure of adeno-associated virus serotype 8, a gene therapy vector. *J. Virol.* **81**, 12260–12271 (2007).
15. Wu, Z. *et al.* Single amino acid changes can influence titer, heparin binding, and tissue tropism in different adeno-associated virus serotypes. *J. Virol.* **80**, 11393–11397 (2006).
16. Muller, O.J. *et al.* Improved cardiac gene transfer by transcriptional and transductional targeting of adeno-associated viral vectors. *Cardiovasc. Res.* **70**, 70–78 (2006).
17. Raake, P.W. *et al.* Cardio-specific long-term gene expression in a porcine model after selective pressure-regulated retroinfusion of adeno-associated viral (AAV) vectors. *Gene Ther.* **15**, 12–17 (2008).
18. Yang, L. *et al.* A myocardium-tropic AAV evolved by DNA shuffling and *in vivo* selection. *Proc. Natl. Acad. Sci. USA* **106**, 3946–3951 (2009).
19. Hagstrom, J.E. *et al.* A facile nonviral method for delivering genes and siRNAs to skeletal muscle of mammalian limbs. *Mol. Ther.* **10**, 386–398 (2004).
20. Gregorevic, P. *et al.* Systemic delivery of genes to striated muscles using adeno-associated viral vectors. *Nat. Med.* **10**, 828–834 (2004).

© 2010 Nature America, Inc. All rights reserved.

ONLINE METHODS

Mutagenesis and vector production. Domain swapping studies with the AAV2 capsid protein subunit were carried out using the Quik-change site-directed mutagenesis kit using primers (IDT) designed as per manufacturer's instructions (**Supplementary Table 3**). All AAV vectors packaging the cytomegalovirus or CBA (chicken beta actin)-driven luciferase transgene cassette were produced and purified using previously published procedures[21]. Vector genome titers were determined through Q-PCR using primers specific for the firefly luciferase transgene (**Supplementary Table 3**).

Vector characterization. Heparin binding studies with AAV2i mutants were carried out as described earlier with some modifications[10,11]. Briefly, heparin-agarose beads (Sigma) packed in disposable microspin columns (Bio-Rad) were loaded with different AAV vectors and subjected to four wash cycles using phosphate buffer containing 15 mM NaCl followed by elution with varying NaCl concentrations in phosphate buffer. Vector genomic DNA was extracted from fractions collected from each step using a DNeasy kit (Qiagen) and subjected to dot blot analysis using a ^{32}P-labeled probe specific for the luciferase transgene.

The potential role of cell surface heparan sulfate or sialylated glycans in AAV infection was characterized in CHOpgsD and HEK293 cell lines, respectively. Briefly, 2×10^5 HEK293 cells were left untreated or treated for 2 h at 37 °C with 50 mU/ml neuraminidase type III (sialidase) from *Vibrio cholerae* in culture media without serum. At 24 h after infection with AAV2i mutants (multiplicity of infection: 1,000), cell lysates were obtained using passive lysis buffer and subjected to a luminometric assay (Promega). Studies with CHOpgsD cells were carried out using similar conditions.

Vector administration and animal studies. Housing and handling of BALB/c mice used in the current study were carried out in compliance with National Institutes of Health guidelines and approved by IACUC at University of North Carolina-Chapel Hill (protocols #06–300 and #09–117). AAV2i mutants or parental AAV serotype vectors were administered through the intramuscular (right hind limb; gastrocnemius muscle) in a volume of 50 µl PBS or intravenous route (tail vein) in a volume of 200 µl PBS. Luciferase expression in animals was imaged using a Xenogen IVIS100 imaging system (Caliper Lifesciences) after intraperitoneal injection of luciferin substrate (120 mg/kg; Nanolight). Image analysis was carried out using the Living Image software. Luciferase transgene expression in various tissue types was determined as described earlier using a luminometric assay (Promega). Vector genome copy numbers were determined after extraction of genomic DNA at different time intervals from whole blood (10 µl collected from tail vein in heparinized capillary tubes) and various tissue types using a DNeasy kit (Qiagen).

Isolated limb perfusion studies were carried out in BALB/c mice as described earlier[19]. Mice were anesthetized with 1–2% isoflurane throughout the procedure. Prior to injection of AAV vectors at three different doses (1×10^9, 1×10^{10}, or 1×10^{11} vg), a tourniquet was placed on the upper hind limb to restrict blood flow into and out of the hind limb. AAV2i8 or AAV8 vectors were injected into the saphenous vein at a rate of 8 ml/min using a needle catheter connected to a programmable Harvard PHD 2000 syringe pump (Harvard Instruments). Mice received acetaminophen (100 mg/kg/day, in drinking water) for the first 48 h after each surgical procedure. Mice were euthanized at 2 weeks post-injection and limb muscles harvested and separated into six groups (quadriceps, biceps, hamstring, gastrocnemius, shin and foot). Luciferase activity from each muscle group was determined using a luminometric assay and total level of luciferase expression per gram of muscle tissue was determined.

Molecular modeling. Structural models of the Vp3 protein subunit of AAV2i mutants were generated using the SWISS-MODEL online three-dimensional (3D) model building server with the crystal structure of AAV2 supplied as template (PDB ID: 1lp3a)[9]. **Figure 1a** showing 3D capsid surface topology was generated using the programs Pymol (http://www.pymol.org/) and Roadmap[13].

21. Grieger, J.C. *et al.* Production and characterization of AAV vectors. *Nat. Protoc.* **1**, 1412–1428 (2006).

© 2010 Nature America, Inc. All rights reserved.

nature biotechnology

Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis

Noelle M Griffin[1,2], Jingyi Yu[1,2], Fred Long[1,2], Phil Oh[1,2], Sabrina Shore[1,2], Yan Li[1,2], Jim A Koziol[3] & Jan E Schnitzer[1,2]

© 2010 Nature America, Inc. All rights reserved.

Replicate mass spectrometry (MS) measurements and the use of multiple analytical methods can greatly expand the comprehensiveness of shotgun proteomic profiling of biological samples[1–5]. However, the inherent biases and variations in such data create computational and statistical challenges for quantitative comparative analysis[6]. We developed and tested a normalized, label-free quantitative method termed the normalized spectral index (SI_N), which combines three MS abundance features: peptide count, spectral count and fragment-ion (tandem MS or MS/MS) intensity. SI_N largely eliminated variances between replicate MS measurements, permitting quantitative reproducibility and highly significant quantification of thousands of proteins detected in replicate MS measurements of the same and distinct samples. It accurately predicts protein abundance more often than the five other methods we tested. Comparative immunoblotting and densitometry further validate our method. Comparative quantification of complex data sets from multiple shotgun proteomics measurements is relevant for systems biology and biomarker discovery.

Quantitative proteomics is widely used for examining differences in global protein expression between cellular states and in disease biomarker and target discovery[3,7–10]. Current methods for use with labeled or label-free approaches are based on a single MS feature of abundance, such as spectral or peptide count or chromatographic peak area or height (**Supplementary Notes**). Comparing and quantifying differential expression remains an important challenge for this field[11,12]. To date, MS fragment ion intensities appear only to be used for candidate-based quantification, such as the quantification of small molecules relative to a labeled version of the analyte of interest[13]. A similar approach is single or multiple reaction monitoring where transitions from selected precursor to specific fragment ions are monitored and compared to a standard[14,15]. Fragment ion intensities are also used in iTRAQ quantification, where the intensity of the reporter fragment ion is directly related to the abundance of the precursor from which it's derived[16]. To date, fragment-ion approaches have not been applied in a label-free manner or used in large-scale shotgun proteomics analysis. Here, we explore their utility as an abundance feature.

We previously discovered that multiple MS measurements of a sample are required for large-scale shotgun proteomics platforms to achieve statistically significant comprehensiveness in protein identifications[1] (**Supplementary Notes**). This is critical for biomarker discovery where proteins differentially expressed between normal and disease samples can be identified only if samples are analyzed systematically and equivalently to completeness. This requires four to eight MS measurements of each distinct sample[1,2,17]. Unfortunately, because replicate data contain inherent biases and variations, MS signals are frequently corrupted by systematic or even apparently random changes (**Supplementary Notes**).

We set out to develop and test various methods to quantify, normalize and compare complex label-free proteomic data. We concurrently developed and tested various methods to normalize these features to control for measurement biases and variations. We sought MS features of abundance recorded in all data sets that can be easily extracted, and thus can be universally mined. These include spectral count (SC, number of MS/MS spectra per peptide) and unique peptide number (PN). We also include fragment ion (MS/MS) intensities as a new feature easily extracted from typical MS data and, to our knowledge, not incorporated previously into unlabeled, normalized quantification.

The spectral index (SI) is the cumulative fragment ion intensity for each significantly identified peptide (including all its spectra) giving rise to a protein and is defined as

$$\mathrm{SI} = \sum_{k=1}^{pn}\left(\sum_{j=1}^{sc} i_j\right)_k \qquad (1)$$

where *sc* is the spectral count for the peptide *k*, *i* is the fragment ion intensity of peptide *k*, *j* is the *j*th spectral count of *sc* total spectral counts for peptide *k* and *pn* is the number of peptides identified for that protein. Therefore, this equation inherently incorporates fragment ion intensity values with SC and PN for each protein.

To test the reproducibility of the raw MS abundance features, we graphed the mean diamonds and confidence circles (see Online Methods) of multiple MS measurements of the same liver endothelial plasma membrane sample, with the null hypothesis that all replicates are equal. The mean PN, SC and SI across data

[1]Proteogenomics Research Institute for Systems Medicine, San Diego, California, USA. [2]Sidney Kimmel Cancer Center, San Diego, California, USA. [3]The Scripps Research Institute, La Jolla, California, USA. Correspondence should be addressed to J.E.S. (jschnitzer@prism-sd.org).

Received 21 September; accepted 16 November; published online 13 December 2009; doi:10.1038/nbt.1592

sets were not sufficiently reproducible and showed significant differences ($P < 0.05$; **Fig. 1a–c**), easily visualized by the non-overlapping mean diamonds and confirmed by analysis of variance (ANOVA; Online Methods). Thus, normalization is required to enable meaningful quantitative comparison within and between samples.

We began with a simplistic approach to normalize the MS data sets by using the 'housekeeping' protein actin. The SI of each protein was divided by the SI of actin in each MS measurement to yield SI_{act}.

$$SI_{act} = \left[\sum_{k=1}^{pn}\left(\sum_{j=1}^{sc} i_j\right)_k\right] \Big/ \left[\sum_{k=1}^{pn}\left(\sum_{j=1}^{sc} i_j\right)_k\right]_{act} \quad (2)$$

This normalization approach was applied to the 5,923 proteins identified in common across all the liver replicate MS measurements.

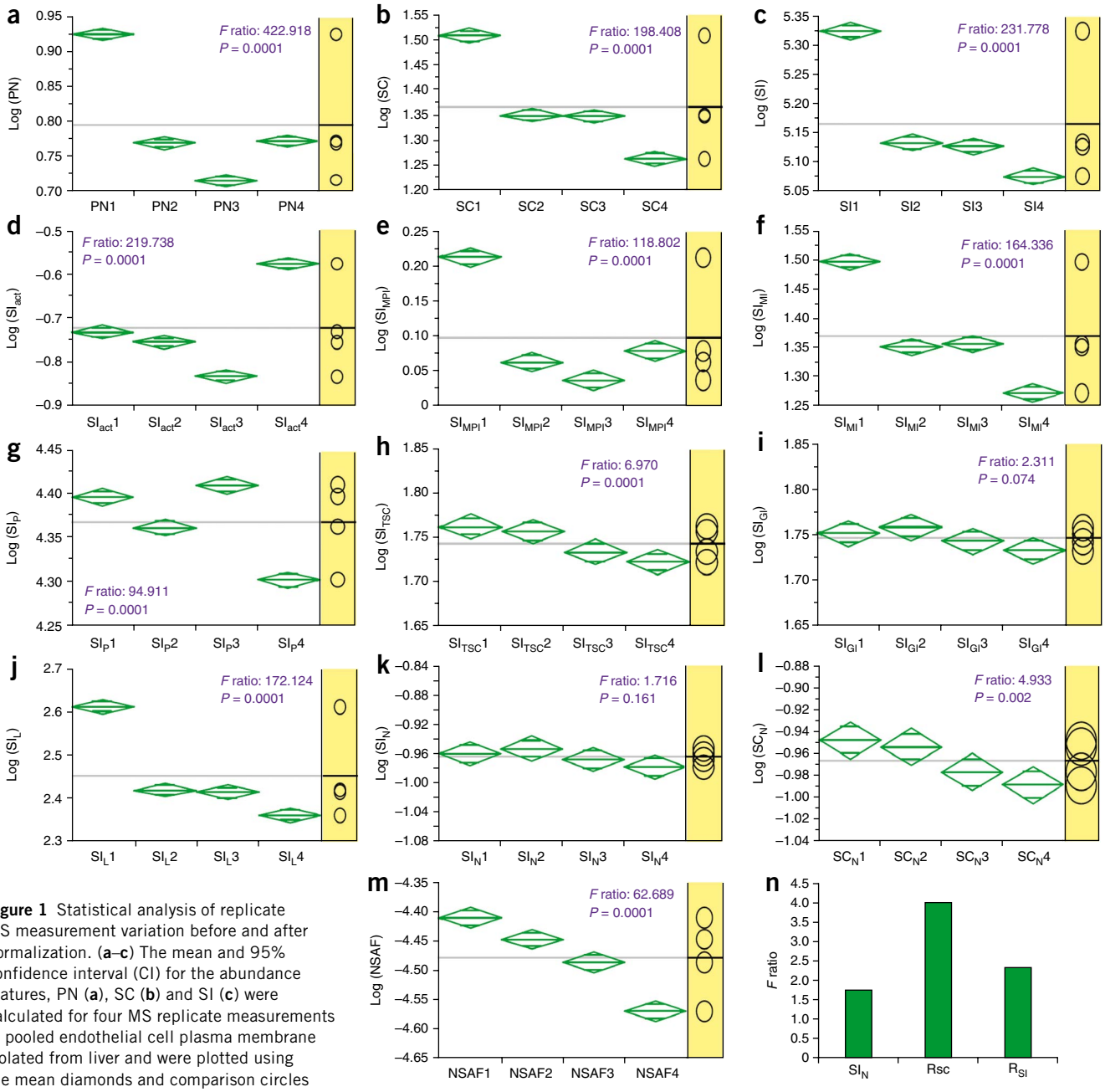


Figure 1 Statistical analysis of replicate MS measurement variation before and after normalization. (**a**–**c**) The mean and 95% confidence interval (CI) for the abundance features, PN (**a**), SC (**b**) and SI (**c**) were calculated for four MS replicate measurements of pooled endothelial cell plasma membrane isolated from liver and were plotted using the mean diamonds and comparison circles methods. If the CIs, as indicated by the diamonds, do not overlap, the groups are significantly different. For statistical analysis of difference in mean intensities or other features, between multiple replicate samples, one-way ANOVA was performed. Our null hypothesis was that all replicate samples were equal. If our null hypothesis is true then we expect the *F*-ratio to be ~1 (d.f. = 5,919). Our significance level was $P < 0.05$. The *x* axis represents each of the four replicate data sets, and the *y* axis represents the log of abundance feature being examined ($n = 5{,}923$). (**d**–**l**) The indicated normalization methods were applied separately to the SI (**d**–**k**) or SC (**l**) data sets and tested for differences as described above. (**m**,**n**) We applied NSAF[20] and Rsc[21] methods to the replicate data sets (see Online Methods for equations) and tested for differences. (**n**) Graph of the comparison of *F*-ratios obtained from statistical testing of SI_N, Rsc and R_{SI}, where R_{SI} is the Rsc equation with SI substituted for SC.

© 2010 Nature America, Inc. All rights reserved.

A significant difference was still detected between the replicates ($P < 0.05$; **Fig. 1d**). The results were similar when we replaced SI with SC or PN, or when we tested different standards or tissue samples (data not shown).

Next we used mean SI values to normalize the data. The mean protein intensity (MPI) was calculated by dividing the total SI for all identified proteins $\left(\sum_{j=1}^{n} SI_j\right)$ by the total number of proteins identified, n. The SI of each protein was subsequently normalized by MPI:

$$SI_{MPI} = SI/\left[\left(\sum_{j=1}^{n} SI_j\right)/n\right] \quad (3)$$

Similarly, we incorporated the total SC to generate another mean intensity normalization (MI) method (equation (4)). MI was calculated by dividing the total SI for the data set $\left(\sum_{j=1}^{n} SI_j\right)$ by the total SC of the data set $\left(\sum_{j=1}^{n} SC_j\right)$. The SI of each protein was subsequently normalized by MI to yield SI_{MI} for each protein:

$$SI_{MI} = SI/\left[\left(\sum_{j=1}^{n} SI_j\right)\Big/\left(\sum_{j=1}^{n} SC_j\right)\right] \quad (4)$$

However, the replicates were still significantly different ($P < 0.05$; **Fig. 1e,f**). Equivalent normalization of the SC and PN data sets was even less effective (data not shown).

Next, we incorporated PN, SC and MS/MS intensities of each peptide into subsequent normalizations, because the SI values are dependent on these features. Each individual SI was normalized by either the total PN (PNt) for the protein, p (equation (5)), total SC (TSC) (equation (6)), or global/total intensity (GI) (equation (7)):

$$SI_p = SI/\left(\sum_{j=1}^{PNt} p_j\right) \quad (5)$$

$$SI_{TSC} = SI/\sum_{j=1}^{n} SC_j \quad (6)$$

$$SI_{GI} = SI/\sum_{j=1}^{n} SI_j \quad (7)$$

The replicate SI_p and SI_{TSC} data sets were still significantly different ($P < 0.05$; **Fig. 1g,h**). SI_{GI} showed no significant difference between the data sets (**Fig. 1i**), indicating that it succeeded in normalizing the replicate data sets.

Although the SI_{GI} method provided a dramatic improvement, we aimed for further enhancement. As large proteins can contribute more peptides than smaller ones, their abundance may be overestimated[18,19]. To correct for protein length (number of amino acids), we first normalized SI by protein length (SI_L) (**Supplementary Table 1**, equation (8)). Although this improved on SI alone (**Fig. 1j**), significant differences ($P < 0.05$) were still evident between the samples. As SI_{GI} successfully normalized different samples, we incorporated protein length into this method, resulting in SI_N:

$$SI_N = SI_{GI}/L \quad (9)$$

No significant difference could be detected between the SI_N normalized data sets (**Fig. 1k**) and SI_N was superior to SI_{GI}.

We applied the SI_N normalization similarly to the SC data sets, by replacing SI with SC to yield SC_N.

$$SI_N = SI_{GI}/L$$

$$SI_N = \left[\sum_{k=1}^{pn}\left(\sum_{j=1}^{sc} i_j\right)_k\right]\Big/\left[\sum_{j=1}^{n} SI_j\right]\Big/L \quad (10)$$

SC_N failed to adequately reduce the variation between the data sets (**Fig. 1l**), showing the superiority of SI over SC. As expected, PN was even worse (data not shown).

Next, we compared SI_N to two published SC methods, normalized spectral abundance factor (NSAF)[20] and $\log_2$(protein ratio) from spectral counts (Rsc)[21]. NSAF failed to adequately normalize our replicate liver data sets (**Fig. 1m**). Substituting SI for SC in the NSAF approach proved much better but there was still a significant difference between the replicates (data not shown). Rsc proved better than NSAF

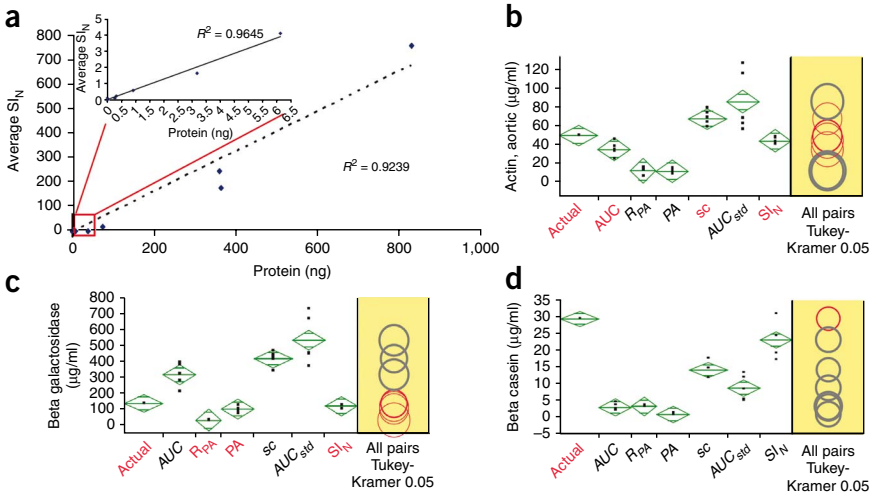


Figure 2 Correlation of SI_N with protein abundance. (**a**) BSA was spiked with a protein mix containing 19 standard proteins spanning a wide dynamic range (0.5–50,000 fmol) (Online Methods), which was separated by SDS-PAGE, trypsin digested and analyzed by two-dimensional LC. SI_N values for each spiked protein were calculated, averaged and plotted against the amount of the protein standard added. Due to the large range in protein abundance, many of the data points cluster close to the origin, thus this region was magnified for ease of visualization. The R^2 correlation was 0.9239. (**b**–**d**) Statistical analysis comparing the quantification of proteins across replicate measurements using six quantification methods (relative to known value). The mean and 95% CI for protein abundance, as determined by various relative quantitative methods, were plotted for three representative proteins from a standard protein mixture[23] and compared to the actual loaded amount using ANOVA. Individual means were compared using the Tukey-Kramer honestly significant difference (HSD method[33,34]). Quantitative methods that were not significantly different from the actual protein abundance (ANOVA, α–significance level = 0.05) are highlighted in red.

© 2010 Nature America, Inc. All rights reserved.

Table 1 Comparison of quantitative methods in the determination of the correct amount of spiked protein

Method	No. of correct abundance determinations (α = 0.05)	Percent of correct abundance determinations
SI_N	13 (2)[a]	81.25 (93.75)[b]
SC	6	37.50
AUC	8 (1)[a]	50 (56.25)[b]
AUC_{std}[22]	3	18.75
PA	7	43.75
R_{PA}[21]	7	43.75

Summary of the statistical analysis used to compare the ability of five quantitative methods to accurately determine the correct amount of a protein in a standard mixture across replicate data sets. The number of correct abundance determinations in which the predicted protein amount, as determined by each method, did not deviate significantly from the mean of the actual protein amount was determined using ANOVA (α–significance level = 0.05). SI_N, normalized spectral index; SC, spectral count; AUC, area under the curve; AUC_{std}, area under the curve for a protein relative to the spiked standard; PA, peak area—peak area or AUC for a protein expressed as a percentage of the total PA for all identified proteins (see Online Methods for equations).

[a]No. of times selected for best method at predicting abundance when all methods were outside the α-significance level for a particular protein standard (**Fig. 2d**). [b]Includes the best method for predicting abundance when all methods were outside the α-significance level.

in reducing the variation but was inferior to the SI_N method, (**Fig. 1n**). When SC was replaced by SI in the Rsc equation (R_{SI}), SI yet again outperformed SC, demonstrating the substantial improvement that can be gained by using SI over SC, regardless of the normalization approach (**Fig. 1n**).

To validate SI_N as an abundance feature, we performed experiments in which either BSA (2.65 μg) or a complex plasma membrane fraction (40 μg) was spiked with a mixture of 19 protein standards across a wide dynamic range (0.5–50,000 fmol) (**Supplementary Methods**). The SI_N for each of the standard proteins was calculated and plotted as a function of protein load. R^2 = 0.9239 (**Fig. 2a** and **Supplementary Fig. 1a**). The slope of the regression line is 1.223 (95% CI of 1.101–1.345), meaning the magnitude of SI_N for any given change in protein abundance can be calculated (**Supplementary Data**).

To determine how SI_N compares to the most commonly used abundance features, namely SC and area under the curve (AUC), we analyzed the ability of SI_N, SC and various AUC methods[21,22], to accurately predict the amount of each protein in a published data set of a standard protein mixture[23] (Online Methods and **Supplementary Notes**). For 13 of the 16 proteins in the standard mixture, no significant difference could be determined between the SI_N predicted amount and the actual amount (**Fig. 2b**,**c** and **Table 1**). For the remaining three proteins, SI_N came closest to predicting the actual protein amount two out of three times (**Fig. 2d** and **Table 1**). Thus, SI_N accurately predicted protein amount 81.25% of the time and was the best method at predicting protein amount 93.75% of the time.

The next best method was total AUC, which was accurate 56.25% of the time, including the one time AUC was closest at predicting protein abundance when all methods were outside the α-significance level of 0.05. The other AUC methods[21,22] faired just as poorly as the 'raw' AUC method in accurately determining the protein amount (18.75% and 43.75%, respectively). SC predicted correct protein abundance for the proteins only 37.5% of the time (**Table 1**).

To determine whether SI_N could control for variation in sample load, we compared two different MS data sets taken from the same sample, but analyzed different protein amounts (**Fig. 3**). Proper normalization should scale the individual 40- and 150-μg samples, facilitating a direct comparison based on relative abundance. Before normalization, the mean SI values between the 40- and 150-μg samples were clearly and significantly different ($P < 0.05$; **Fig. 3a**). First, we corrected the SI values for the 40-μg sample by the dilution factor, SI40*150/40 (**Fig. 3b**). However, the samples were still significantly different (*t*-ratio 24.459). Although the Rsc method was more effective than the NSAF method, significant variation was still apparent (*t*-ratio 11.916; **Fig. 3d**,**e**). SI_N was the only method tested that

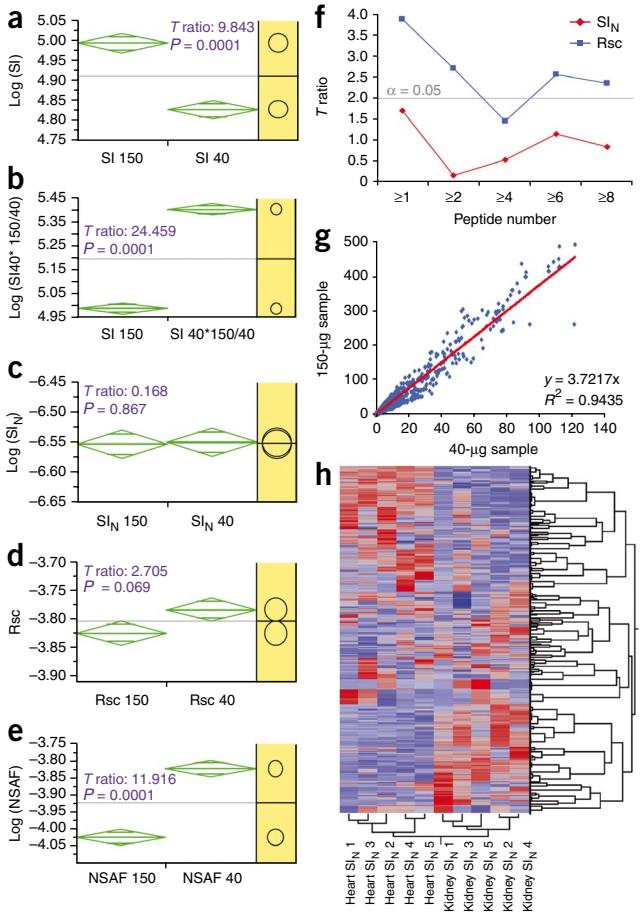


Figure 3 Statistical analysis of normalization methods applied to variable protein load and distinct sample data sets. The indicated normalization methods were each applied to the 40-μg and 150-μg MS data sets from normal lung endothelial cell plasma membranes. (**a**–**e**) Mean and 95% CI for a raw SI data set (**a**) and data sets normalized by the dilution factor (**b**), SI_N (**c**), Rsc (**d**) and NSAF (**e**) were plotted using the mean diamonds and comparison circles. The *x* axis represents the two different protein loads, and the *y* axis represents the log of the normalized abundance feature (number of common proteins, n = 2,660). (**f**) *T*-ratios for statistical testing of SI_N and Rsc are plotted as a function of peptide cut-off numbers (number of peptides and/or proteins commonly identified between the samples). The α = 0.05 significance line is plotted in gray. *T*-ratios above this line indicate that samples are different. (**g**) The SI_N values converted to estimated nanogram amounts (based on initial sample load) for 2,660 proteins common between the 40 μg and 150 μg data sets were plotted against each other. The slope of the line is 3.72, R^2 = 0.94. (**h**) Two-way clustering of ~3,000 proteins identified in heart and kidney endothelial cell plasma membrane samples. Each column in the matrix represents a single two-dimensional LC-MS/MS run for either heart or kidney, based on the SI_N normalized MS data. Proteins (rows) and tissues (columns) are clustered based on their similarities in protein intensity profile. Colors within the heatmap range from light blue (least prevalent) to dark red (most prevalent), illustrating the relative abundance of each protein within a particular sample.

© 2010 Nature America, Inc. All rights reserved.

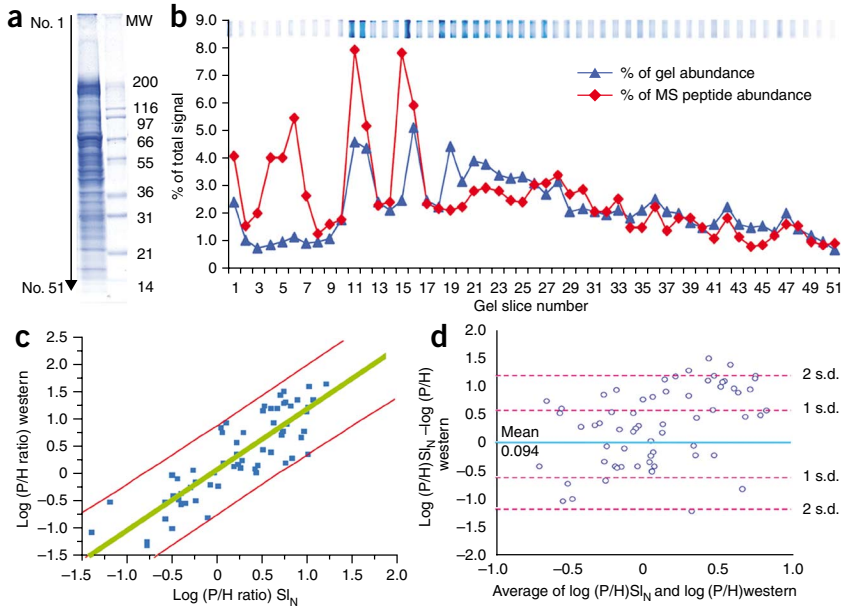


Figure 4 Comparative analysis of proteins quantified by SDS-PAGE and MS analysis. (**a**) Proteins in endothelial cell plasma membranes from rat lung were separated by SDS-PAGE, stained with Coomassie blue and cut into 51 slices. Each gel slice was subjected to densitometry and MS analysis. (**b**) The densitometry intensities for each slice were compared to the SI on the same axis, with the *x* axis being the gel slice number. (**c**) Sixty-four proteins found in both lung endothelial cell plasma membranes (P), and the entire lung homogenate (H) were analyzed by western blot analysis to quantify protein signal by densitometry. The P/H ratio for each protein from the western analysis is plotted against its P/H ratio from the SI_N values (multiple measurements). Spearman's Rho correlation between western and SI_N ratio is $\rho = 0.86$ and all the points fall within 95% CI (red line). (**d**) The Bland-Altman plot for the two methods with 1 and 2 s.d. of the mean.

eliminated the variation introduced by dissimilar or even unknown protein loads and thereby facilitated comparisons of proteins between these samples (**Fig. 3c**). When the *T*-ratios obtained from testing SI_N and Rsc are plotted at incremental peptide cut-off levels (confidently identified peptides, common between samples see Online Methods), SI_N consistently outperforms Rsc as each *T*-ratio falls below the significance level of 2, meaning that no significant difference can be found between the SI_N normalized data sets (**Fig. 3f**).

Next, we used SI_N values to estimate protein amounts (Online Methods). Linear regression analysis of the estimated nanogram amounts for each of the 2,660 common proteins from the 40- and 150-µg samples fit a straight line ($R^2 > 0.94$) with a slope of 3.72, in excellent agreement with the 150:40 ratio of 3.75 (**Fig. 3g**). This test provides additional strong statistical validation for the applicability of the SI_N method to thousands of proteins for large-scale quantification of protein expression.

As a more stringent test, we determined whether the SI_N method could facilitate unsupervised hierarchical clustering of biologically distinct data sets to identify correlated expression patterns. Using endothelial cell plasma membranes isolated from kidney and heart, each with five replicate MS measurements, we performed two-way unsupervised clustering (not imparting any prior knowledge onto the data set) on the complete data sets using SI_N values for all commonly identified proteins (across all data sets; **Fig. 3h**). The clustering algorithm successfully clustered replicates according to tissue type, whereas distinct samples (heart versus kidney) could be visually separated. This confirmed that SI_N is quite successful in recognizing sameness of replicates while exposing the differences of distinct biological samples.

Next, we compared intensities obtained from a Coomassie-stained SDS-PAGE gel (**Fig. 4a**) to the SI abundance values generated by liquid chromatography (LC)-MS/MS analysis. Each gel slice was treated as a distinct sample with an MS and densitometric measurement to determine relative abundance. The two profiles overlapped substantially for most of the gel (**Fig. 4b**). This strong correlation verifies further the utility of SI as a quantitative tool. Notably, MS also detected proteins in gel regions containing high-molecular-weight proteins with little Coomassie staining, consistent with this stain's well-known inability to stain some high-molecular-weight proteins. The sensitivity and utility of the SI method is readily apparent.

To determine whether we could detect quantitative differences for individual proteins expressed in two distinct samples, we applied the SI_N method to data sets from total lung homogenates (H) versus lung endothelial plasma membrane (P) subfractions. Because the P fraction is physically derived from H, the protein composition of P is a subset of the proteins present in H. To determine those proteins that are enriched in P compared to H, we used SI_N and western blot analysis to generate P/H expression ratios for individual proteins. Statistical comparison of the P/H ratios for 64 proteins, ranging from low to high abundance, produced a Spearman's Rho correlation coefficient of 0.86, indicating an excellent positive correlation between the two methods (**Fig. 4c**). As the Bland Altman plot[24] showed a mean difference of 0.09, and the data points fall within 2 s.d. of the mean (95% CI for the difference between the two methods), there is very little evidence to indicate that the quantitative methods are significantly different (**Fig. 4d**).

In this study, we have developed and tested a number of methods to quantify and normalize complex proteomics data obtained from a variety of MS methodologies. Extensive statistical testing and validation of our methods systematically demonstrated their utility through a wide variety of tests applied to diverse MS data sets. With the SI_N method, we have successfully compensated for experimental and random bias and noise, thus showing that protein abundances, as reflected by mean SI_N from replicate samples, are essentially the same. Conversely, the analysis of biologically distinct samples, with noise and bias controlled by proper normalization, enables meaningful direct quantitative comparisons reflecting their true biological diversity.

We aimed to generate an abundance index with the convenience of SC, but greater confidence at low peptide numbers without the added complexity of peak area or AUC measurements. We tested the benefit of combining PN, SC and MS/MS ion intensities into one metric, SI, as opposed to using these features in isolation. This approach proved to be more statistically robust than the SC or AUC methods (**Figs. 1** and **2** and **Supplementary Figs. 2** and **3**) and obviates the need for samples spiked with protein standards.

Using the fragment ion intensity, specifically only those intensities that match the peptide of interest (these are inherently more reflective of the precursor), may facilitate more accurate measurements as there is less chance of including the signal of co-eluting precursors or

© 2010 Nature America, Inc. All rights reserved.

background noise (**Supplementary Discussion**). Also, the peak height of the MS/MS fragments are summed for SI_N, whereas for the AUC, the precursor ions are integrated. Overlapping peptides in the precursor/MS scan increases the chance for error with the integration process. This is particularly important as most mass spectrometers operate in conjunction with LC systems, thereby producing MS scans permeated by chemical noise. In addition, the chromatogram becomes noisier as the sample complexity increases. For example, even our typical 36-h LC-LC-MS/MS runs for complex samples still produce overlapping chromatograms due to co-elution peaks, making AUC quantification troublesome. Most groups use LC or LC-LC setups with much shorter elution times. This will continue to be an issue, even in the advent of high-resolution instruments, as improvements in MS resolution can only really be appreciated when the chromatographic resolution improves in tandem. This has yet to be fully realized. It is not, however, an issue for the SI_N calculation as the MS/MS spectra is inherently less complex and no integration is performed.

SI_N could accurately determine the correct amount of each protein standard in a mixture better than all other methods tested (**Fig. 2b–d** and **Table 1**). Despite giving the AUC methods the best possible advantage (**Supplementary Notes**), SI_N consistently outperformed them in determining protein abundance (**Supplementary Figs. 2** and **3**). SC performed as modestly as the AUC methods. SI_N could accurately determine the relative abundance for thousands of proteins in complex samples, without the need for spiking with protein standards (**Figs. 3g** and **4**). SI_N also facilitated the identification of a subset of proteins enriched in P relative to H. We showed outstanding correlation between the SI_N and western blot analysis ratios (**Fig. 4c**,**d**), validating this enrichment.

As abundance features, SI, SC, PN and AUC are not reproducible across replicate data sets (**Fig. 1a–c** and **Supplementary Fig. 4a**). Methods of normalizing complex LC-MS/MS data are only just emerging[21,25–29], but no comparison is generally shown between the pre- and post-normalized data. Stringent validation has been rather sporadic and the results and efficacy have been variable[26,30–32]. Therefore, we undertook a systematic and logical approach to normalize our data sets. Of the methods tested, only SI_{GI} and SI_N removed the variability between replicates (**Fig. 1i**). For SI_N, we added a protein size parameter to the SI_{GI} calculation, resulting in a clear benefit over SI_{GI} alone (**Fig. 1i**,**k**). Even when we substituted SI for SC, PN or AUC in the normalization methods, SI consistently outperformed all other features, regardless of the sample, measurement or normalization approached applied (**Fig. 1k**,**l**,**o** and **Supplementary Fig. 4c**). Even the raw SI values show less variation between the replicates than AUC values normalized using the SI_N approach (**Supplementary Fig. 4**).

SI_N does not appear to overnormalize, but rather can reduce replicate variability to maintain sameness in data sets from a single sample while maintaining quantitative differences between distinct samples (**Fig. 3h**). As SI_N can also successfully normalize data sets with different loading amounts (**Fig. 3c**), the dilution factor between samples can be accurately derived from the data simply by calculating the slope of the regression line (**Fig. 3g**). SI_N can also control for the variation introduced by different MS methodological analysis of the same sample (**Supplementary Fig. 5**) to facilitate comparison and quantification across all data sets (**Supplementary Fig. 6**). This may have important implications for the comparison of data sets acquired in different laboratories.

In summary, combining and normalizing several MS abundance features—including for the first time, fragment ion intensities—should find broad utility in MS quantification. When we compared our new methods to each other and to previously reported methods[21,22], the best method was SI_N. This scoring function was developed through logical systematic application of enabling parameters which, when combined, produced improved normalization and quantification. Our method allows the quantitative comparison of biologically distinct data sets with high confidence and relative ease, and should therefore greatly facilitate the use of label-free quantitative proteomic approaches for differential protein expression analysis. This method is all the more valuable in the era of systems biology and biomarker discovery where distinct samples must be analyzed both quantitatively and comprehensively through the replicate MS measurements necessary to gain confidence in determining expression differences and novel biomarkers.

METHODS

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

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

ACKNOWLEDGMENTS

This work was supported by National Institute of Health grants (to J.E.S): RO1HL074063, R33CA118602 and P01CA104898.

AUTHOR CONTRIBUTIONS

N.M.G. designed, developed and analyzed the methods, provided some of the mass spectrometry data, performed the spiking experiments and analysis and wrote the manuscript; J.Y. initiated the project, designed, tested and implemented the methods; F.L. developed the scripts for data extraction; P.O. performed western blot analysis and densitometry; S.S. performed western blot analysis; Y.L. provided key mass spectrometry data; J.A.K. provided direction for statistical analysis; J.E.S supervised the project, designed specific tests and helped to write the manuscript. All authors have read and agreed to all the content in this manuscript.

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

1. Durr, E. *et al.* Direct proteomic mapping of the lung microvascular endothelial cell surface in vivo and in cell culture. *Nat. Biotechnol.* **22**, 985–992 (2004).
2. Li, Y. *et al.* Enhancing identifications of lipid-embedded proteins by mass spectrometry for improved mapping of endothelial plasma membranes in vivo. *Mol. Cell. Proteomics* **8**, 1219–1235 (2009).
3. Oh, P. *et al.* Subtractive proteomic mapping of the endothelial surface in lung and solid tumours for tissue-specific therapy. *Nature* **429**, 629–635 (2004).
4. Slebos, R.J. *et al.* Evaluation of strong cation exchange versus isoelectric focusing of peptides for multidimensional liquid chromatography-tandem mass spectrometry. *J. Proteome Res.* **7**, 5286–5294 (2008).
5. Kislinger, T., Gramolini, A.O., MacLennan, D.H. & Emili, A. Multidimensional protein identification technology (MudPIT): technical overview of a profiling method optimized for the comprehensive proteomic investigation of normal and diseased heart tissue. *J. Am. Soc. Mass Spectrom.* **16**, 1207–1220 (2005).
6. Wong, J.W., Sullivan, M.J. & Cagney, G. Computational methods for the comparative quantification of proteins in label-free LCn-MS experiments. *Brief. Bioinform.* **9**, 156–165 (2008).
7. Oh, P. *et al.* Live dynamic imaging of caveolae pumping targeted antibody rapidly and specifically across endothelium in the lung. *Nat. Biotechnol.* **25**, 327–337 (2007).
8. Shiio, Y. *et al.* Quantitative proteomic analysis of Myc oncoprotein function. *EMBO J.* **21**, 5088–5096 (2002).
9. Shiio, Y. *et al.* Quantitative proteomic analysis of myc-induced apoptosis: a direct role for Myc induction of the mitochondrial chloride ion channel, mtCLIC/CLIC4. *J. Biol. Chem.* **281**, 2750–2756 (2006).
10. Chiang, M.C. *et al.* Systematic uncovering of multiple pathways underlying the pathology of Huntington disease by an acid-cleavable isotope-coded affinity tag approach. *Mol. Cell. Proteomics* **6**, 781–797 (2007).
11. Service, R.F. Proteomics. Proteomics ponders prime time. *Science* **321**, 1758–1761 (2008).
12. Service, R.F. Proteomics. Will biomarkers take off at last? *Science* **321**, 1760 (2008).
13. Kolodziej, E.P., Gray, J.L. & Sedlak, D.L. Quantification of steroid hormones with pheromonal properties in municipal wastewater effluent. *Environ. Toxicol. Chem.* **22**, 2622–2629 (2003).
14. Wolf-Yadlin, A., Hautaniemi, S., Lauffenburger, D.A. & White, F.M. Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. *Proc. Natl. Acad. Sci. USA* **104**, 5860–5865 (2007).

© 2010 Nature America, Inc. All rights reserved.

15. Kuhn, E. *et al.* Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and 13C-labeled peptide standards. *Proteomics* **4**, 1175–1186 (2004).
16. Ross, P.L. *et al.* Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **3**, 1154–1169 (2004).
17. Koziol, J.A., Feng, A.C. & Schnitzer, J.E. Application of capture-recapture models to estimation of protein count in MudPIT experiments. *Anal. Chem.* **78**, 3203–3207 (2006).
18. Ishihama, Y. *et al.* Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell. Proteomics* **4**, 1265–1272 (2005).
19. Rappsilber, J., Ryder, U., Lamond, A.I. & Mann, M. Large-scale proteomic analysis of the human spliceosome. *Genome Res.* **12**, 1231–1245 (2002).
20. Zybailov, B. *et al.* Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *J. Proteome Res.* **5**, 2339–2347 (2006).
21. Old, W.M. *et al.* Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteomics* **4**, 1487–1502 (2005).
22. Silva, J.C. *et al.* Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol. Cell. Proteomics* **5**, 144–156 (2006).
23. Klimek, J. *et al.* The standard protein mix database: a diverse data set to assist in the production of improved Peptide and protein identification software tools. *J. Proteome Res.* **7**, 96–103 (2008).
24. Bland, J.M. & Altman, D.G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1**, 307–310 (1986).
25. Callister, S.J. *et al.* Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J. Proteome Res.* **5**, 277–286 (2006).
26. Wang, W. *et al.* Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal. Chem.* **75**, 4818–4826 (2003).
27. Lukas, T.J. *et al.* Informatics-assisted protein profiling in a transgenic mouse model of amyotrophic lateral sclerosis. *Mol. Cell. Proteomics* **5**, 1233–1244 (2006).
28. Forner, F. *et al.* Quantitative proteomic comparison of rat mitochondria from muscle, heart, and liver. *Mol. Cell. Proteomics* **5**, 608–619 (2006).
29. Choi, H., Fermin, D. & Nesvizhskii, A.I. Significance analysis of spectral count data in label-free shotgun proteomics. *Mol. Cell. Proteomics* **7**, 2373–2385 (2008).
30. Baggerly, K.A. *et al.* A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. *Proteomics* **3**, 1667–1672 (2003).
31. Wagner, M., Naik, D. & Pothen, A. Protocols for disease classification from mass spectrometry data. *Proteomics* **3**, 1692–1698 (2003).
32. Anderle, M. *et al.* Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum. *Bioinformatics* **20**, 3575–3582 (2004).
33. Kramer, C.Y. Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics* **12**, 309–310 (1956).
34. Tukey, J.W. Some selected quick and easy methods of statistical analysis. *Trans. N.Y. Acad. Sci.* **16**, 88–97 (1953).

© 2010 Nature America, Inc. All rights reserved.

ONLINE METHODS

Sample preparation. Sprague-Dawley female rats (150–250 g; Charles River Laboratories) were used unless otherwise indicated, and all animal procedures were carried out in accordance with the Sidney Kimmel Cancer Center committee on Animal Usage and Care (IACUC) standards. As described previously[35,36], luminal vascular endothelial cell plasma membranes were directly isolated from rat lung and liver tissues with quality control showing ≥20-fold enrichment for known endothelial makers and ≥20-fold depletion of markers of other cell types and subcellular organelles. Sample purity was assessed with multiple antibodies against protein markers for endothelial membrane and other cellular compartments.

Mass spectrometry analysis. Proteins were prefractionated on SDS-PAGE gels before two-dimensional (2D) LC-MS/MS and reversed phase (RP)-MS/MS. For RP-MS/MS using either an LCQ or LTQ mass spectrometer, digested peptides were extracted from each gel slice and lyophilized[2]. For 2D LC-MS/MS using either an LCQ or LTQ, peptides extracted from each gel slice were first pooled into seven groups, then lyophilized[2]. Data acquisition from both the LCQ and LTQ was carried out in data-dependent mode. Full MS scans were recorded on the eluting peptides over the 400–1400 m/z range with one MS scan followed by three MS/MS scans of the most abundant ions. A dynamic exclusion was applied for repeat count of 2, a repeat duration of 0.5 min and an exclusion duration of 3 min. A dynamic exclusion window was applied for a duration of 10 min for 2D LC-MS/MS.

Database search. The acquired MS/MS spectra were converted into mass lists using the Extract_msn program from Xcalibur and searched against a protein database containing human, rat and mouse sequences (total entries, 262,200) using the Sequest program in Bioworks 3.1 for Linux (Thermo Fisher Scientific). The searches were performed allowing for tryptic peptides only with peptide mass tolerance of 1.5 Da for LTQ data, 2.0 Da for LCQ data and a minimum of 21 fragmented ions in one MS/MS scan. Accepted peptide identification was based on a minimum ΔCn score of 0.1; minimum cross-correlation score of 1.8($z = 1$), 2.5($z = 2$), 3.5($z = 3$). False-positive identification rate was determined by the ratio of number of peptides found only in the reversed database to the total number of peptides found in both forward and reverse databases. The false-positive identification rates were ≤1%[2]. The positive protein identification results were extracted from Sequest.out files, filtered and grouped with DTASelect software using the above criteria. Proteins were identified based on two unique, significantly identified peptides.

Fragment ion intensity (intensity of the ions in the MS/MS spectrum that are assigned to a given peptide), peptide number (number of unique peptides identifying a protein) and spectral counts (number of MS/MS spectra assigned to a particular peptide) were extracted from the DTAselect output files using a script written in-house (**Supplementary Data**). AUC was calculated as described below. For the purpose of this manuscript, fragment ion intensity is defined as the total intensity of all detected b and y fragment ions (MS/MS spectra) for a specific peptide. The fragment ion intensity of each peptide that passes the threshold for identification that gives rise to a significantly identified protein (see above) is summed. The combination of these summed fragment ion intensities from all MS/MS spectra and peptides relating to a given protein is combined and is referred to as the spectral index (SI) for that protein. For faster data acquisition, we used centroid algorithms for all of the MS analysis. In general, the centroid algorithms will sum the intensities if the ions have very close values, that is, isotope clusters. Therefore, the fragment ion intensities obtained are those that are recorded in Bioworks at the time of data acquisition.

NSAF and Rsc. NSAF is described[20] as

$$(\mathrm{NSAF})_k = (\mathrm{Spc/L})_k \Big/ \sum_{i=1} (Spc/L)_i$$

where Spc is the spectral count for protein k and L is the length of protein k.

The Rsc is described[21] as:

$$\mathrm{Rsc} = \log_2\left[(n_2+f)/(n_1+f)+\log_2\left[(t_1-n_1+f)/(t_2-n_2+f)\right]\right]$$

where, for each protein, R_{SC} is the $\log_2$ ratio of abundance between Samples 1 and 2; $n1$ and $n2$ are spectral counts for the protein in Samples 1 and 2, respectively; $t1$ and $t2$ are total numbers of spectra over all proteins in the two samples; and f is a correction factor[37] set to 1.25 as used in the original Rsc study[21].

Protein abundance calculation. SI_N were converted to estimated nanogram amounts, by including the initial sample load in the final calculation using the following equation:

$$RPQ_p = \frac{SI_N}{\sum_{i=1}^{j} (SI_N)_i} * Q * 1{,}000$$

where j = number of all proteins identified with ≥2 unique peptides, and the subscript i refers to the ith protein of j total proteins, and Q is the amount of sample (μg) used in a given measurement.

Statistics. JMP IN 5.1 (SAS Institute) was used for all statistical analysis.

Data sets distribution: skewness and kurtosis. t-tests and ANOVAs are common statistical tests used for determining differences between sample means but require data to be normally distributed to achieve analytical rigor. Our raw SC, PN and SI data sets were not normally distributed (**Supplementary Fig. 7**) as measured by the skewness and kurtosis of the frequency distribution. The skewness is a measure of distributions symmetry. For symmetrical distributions the skewness = 0, for right- and left-tailed distributions the skewness is >0 and <0, respectively. Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. Data sets with high kurtosis (>0) tend to have a distinct peak near the mean, decline rather rapidly and have 'heavy' tails. Data sets with low kurtosis (<0) tend to have a flat top near the mean rather than a sharp peak. Kurtosis = 0 for a normal distribution.

To maintain statistical rigor and to avoid inflated variance, we performed a $\log_{10}$ transformation of our data sets, which produces a reasonable normality as determined from the histogram and Q-Q plots (**Supplementary Fig. 7**). Thus, for comparative statistical analysis, we similarly transformed all the data sets after performing the normalizations described below. It should be noted that equivalent results to those described below were obtained with nonparametric analyses (data not shown).

To visualize normalized data sets, we graphed the mean (center line) and 95% confidence intervals (CI), indicated as diamonds on the graphs, of normalized spectral indexes. If the CIs shown by the mean intervals do not overlap, the groups are significantly different. The reverse is not necessarily true and significance is determined from the summary statistics associated with the analysis (see below). The confidence circles are another way of visualizing the diamonds and aids in determining CI overlap.

To determine whether there was any evidence that the replicate values were significantly different before and after application of the normalization methods, we applied a t-statistic (2 replicates) or ANOVA, one-way (>2 replicates) to look for differences in normalized mean abundance features. For the statistical analysis, we used only the proteins that were identified in common across all replicate data sets for a particular comparison. Our null hypothesis was that both (2 replicates) or all (>2 replicates) samples were equal. For the t-statistic (2 replicates), the normalized values were deemed significantly different if a large t-ratio (as determined from the t-tables) and a small P-value ($P < 0.05$) were produced from the t-statistic. We use a t-ratio <2 in absolute value for significance as it approximates the 0.05 significance level. For analysis of difference in mean intensities between multiple replicate (>2) samples, analysis of variance (ANOVA, one-way) was performed. Our null hypothesis was that all replicate samples were equal. If our null hypothesis was true then we expect the F-ratio to be ~1. (Informally, the smaller the F statistic (equivalently, the larger the P-value), the closer the agreement across the replicates.) Our significance level was $P < 0.05$. If there is no statistically significant difference between the replicates (as indicated by F-ratio ~1) we conclude that the normalization method succeeded in controlling for the variation between the replicate data sets.

© 2010 Nature America, Inc. All rights reserved.

doi:10.1038/nbt.1592

Unsupervised hierarchical clustering. Cluster analysis was performed on a data set from five replicate MS measurements of endothelial cell plasma membranes isolated from kidney and heart samples using JMP 5.1, and using Wards hierarchical method[38].Ward's method is a hierarchical method designed to optimize the minimum variance within clusters (minimizes within-group dispersions). The clustering was unsupervised meaning without labeled classes, optimization criterion, feedback signal or any other information beyond the raw data. Simply, we did not differentiate in any way the heart samples from the kidney samples. A two-way clustering was performed, which is a data mining technique that allows simultaneous clustering of the rows and columns of a matrix[39–41]. The SI_N values for each protein was normalized across each row (all ten samples) using the following standard approach: (SI_N − (mean SI_N) row/(s.d.)row).

Western blot analysis. All antibodies were purchased commercially or obtained as gifts from other researchers. Custom polyclonal antibodies were provided by BioSource and 21st Century Biochemicals. Western blot analysis was carried out as described previously[35,36]. Densitometry analysis was carried out using Scion Image software for PCs.

Quantification of proteins in the standard protein mixture database using multiple quantitative methods. Raw data files from ten replicate analysis of the standard protein mixture[23] carried out by an LTQ mass spectrometer, were downloaded from the ISB website (http://regis-web.systemsbiology.net/PublicDatasets/). We chose these data sets because it is the same mass spectrometer as we use in our own laboratory and thus have all the software necessary for searching the data and extracting the required information. We searched the data against the same databases highlighted in the original paper using Sequest with Bioworks 3.2. The resulting data was sorted and grouped as described above using DTAselect for the calculation of spectral count and SI_N values. We used the peak area calculation function in Bioworks 3.2 (which incorporates the ICIS algorithm) to calculate the AUC for each significantly identified peptide that was matched to a standard protein in the sample. We used the default parameters for the AUC as follows: mass tolerance 1.5 amu, 5 point smoothing, minimum threshold for peak integration is 50,000. The AUC for each protein was presented multiple ways, including methods corresponding to normalization approaches for AUC published in the literature. As SI_N is a normalized index, we thought it only fair to present the AUC data before and after they have been normalized by various published methods. These include:

Total AUC. The area under the curve (AUC) for each protein is presented as the sum of AUCs for all significantly identified peptides identifying each protein in the run (un-normalized data).

PA. This corresponds to the percentage peak area (PA), which is the default 'normalization' in the Bioworks program, where the total AUC for a protein is expressed as a percentage of the total AUC for all identified proteins.

AUC_{std}. The average AUC for the three most intense peptides per protein is calculated and then normalized by the AUC of a protein standard[22]. This approach is very similar to that described in ref. 28 and also very similar to other popular AUC methods that normalize to the AUC of a spiked internal standard.

R_{PA}. Sum the AUC for each peptide, then each peptide is corrected by dividing the peptide by the sum of all peptide intensities. Similar peptides are compared across replicates and average peptide ratios are generated to reflect protein abundance, R_{PA}[21].

The AUC, as calculated by each method outlined above, was determined for each protein in the standard protein mixture and compared to the spectral count and SI_N values generated for the same proteins across the replicates. Only 16 of the 18 proteins were consistently detected in each of the 10 replicates, so these 16 common proteins were compared across the replicates. The amount determined for each protein by each quantitative method was averaged across all replicates and compared to the actual loaded amount using ANOVA. The mean value of each protein was compared to the actual loaded amount using the Tukey-Kramer HSD method[33,34] (which follows the same principle as the *t*-test, but corrects for multiple testing).

35. Oh, P. & Schnitzer, J.E. Isolation and subfractionation of plasma membranes to purify caveolae separately from glycosyl-phosphatidylinositol-anchored protein microdomain. in *Cell Biology: A Laboratory Handbook* (ed. C.J.) 34–36 (Academic Press, Orlando, FL, USA, 1998).
36. Schnitzer, J.E. *et al.* Separation of caveolae from associated microdomains of GPI-anchored proteins. *Science* **269**, 1435–1439 (1995).
37. Beissbarth, T. *et al.* Statistical modeling of sequencing errors in SAGE libraries. *Bioinformatics* **20** Suppl 1, i31–i39 (2004).
38. Kendall, M. *Multivariate Analysis*, edn. 2 (Macmillan, New York, 1980).
39. Mirkin, B. *Mathematical Classification and Clustering* (Kluwer Academic Publishers, Dordrecht, The Netherlands; 1996).
40. Cheng, Y. & Church, G.M. in *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology 93-1032000*, August 19–23, 2000 (AAAI Press, Menlo Park, CA, 2000).
41. Hartigan, J. Direct clustering of a data matrix. *J. Amer. Stat. Assoc.* **67**, 123–129 (1972).

© 2010 Nature America, Inc. All rights reserved.

nature biotechnology

RESOURCE

Transcriptional profiling of growth perturbations of the human malaria parasite *Plasmodium falciparum*

Guangan Hu[1,3], Ana Cabrera[2,3], Maya Kono[2], Sachel Mok[1], Balbir K Chaal[1], Silvia Haase[2], Klemens Engelberg[2], Sabna Cheemadan[1], Tobias Spielmann[2], Peter R Preiser[1], Tim-W Gilberger[2] & Zbynek Bozdech[1]

Functions have yet to be defined for the majority of genes of *Plasmodium falciparum*, the agent responsible for the most serious form of human malaria. Here we report changes in *P. falciparum* gene expression induced by 20 compounds that inhibit growth of the schizont stage of the intraerythrocytic development cycle. In contrast with previous studies, which reported only minimal changes in response to chemically induced perturbations of *P. falciparum* growth, we find that ~59% of its coding genes display over three-fold changes in expression in response to at least one of the chemicals we tested. We use this compendium for guilt-by-association prediction of protein function using an interaction network constructed from gene co-expression, sequence homology, domain-domain and yeast two-hybrid data. The subcellular localizations of 31 of 42 proteins linked with merozoite invasion is consistent with their role in this process, a key target for malaria control. Our network may facilitate identification of novel antimalarial drugs and vaccines.

© 2010 Nature America, Inc. All rights reserved.

Together with AIDS/HIV and tuberculosis, human malaria represents one of the three most dangerous infectious diseases of humankind[1]. In 2007, 1.38 billion people were estimated to be at risk of infection with *P. falciparum*, the protozoan endoparasite responsible for up to 2 million annual human deaths from malaria[2,3]. The lack of an effective vaccine and the rapid spread of resistance to most antimalarial drugs are major concerns for the control of this unicellular eukaryote. In particular, the complexity of the *P. falciparum* life cycle, which is associated with many unique morphological and metabolic states, has challenged efforts to identify parasite-specific molecular mechanisms that can be targeted by new malaria intervention strategies[4].

The genome of *P. falciparum* encodes ~5,300 genes. This obligate endoparasite has lost many basic metabolic abilities, such as a majority of the enzymes of amino acid synthesis, but expanded its repertoire of proteins involved in many parasite-specific functions, such as interaction with its host, antigenic variation and host-cell invasion[5]. This is consistent with the difficulty in predicting functions for the majority of *P. falciparum* proteins. Genome-wide approaches offer an attractive method to accelerate functional annotation of the *P. falciparum* genome.

The haploid state of the genome throughout the majority of the *P. falciparum* life cycle and lack of inducible knockout or RNAi-mediated knockdown systems for this parasite limits the application of forward and reverse genetic approaches to assess gene function in this species[6,7]. Moreover, the low efficiency of the available transfection technologies makes genetic modification of *P. falciparum* too costly and time consuming for genome-wide analyses. Although the potential of systems biology approaches to derive functional gene predictions is widely appreciated[8], previous efforts to predict the functions of uncharacterized *P. falciparum* gene products were based on gene interaction networks derived mainly from probabilistic integration of transcriptome data collected at different stages of the *P. falciparum* life cycle[9–11]. Largely because many genes with unrelated functions exhibit similar transcriptional profiles across the *P. falciparum* life cycle[12,13], these approaches provided relatively low-confidence predictions of gene function.

Although studies with model organisms such as yeast and *Caenorhabditis elegans* suggest that microarray analyses of global transcriptional responses to growth perturbations can substantially improve the accuracy and coverage of probabilistic interaction networks[14,15], the utility of monitoring changes in gene expression in response to growth perturbations for predicting *P. falciparum* gene function has been controversial. Some perturbations, including those associated with several antimalarial drugs, such as chloroquine and several antifolates, induced only low-amplitude mRNA changes with no particular link to their presumed mode of action[16,17]. On the other hand, exposure of *P. falciparum* parasites to febrile temperatures[18], artesunate[19] and an inhibitor of sphingomyeline synthase[20] induced biologically relevant transcriptional changes that led to the identification of proteins associated with these processes.

Here we demonstrate that DNA microarray-based profiling of growth perturbations in *P. falciparum* can generate a high-resolution transcriptional data set that reflects functional relationships between *P. falciparum* genes. We use this data set to construct a gene interaction network that predicts the functions of 2,545 *P. falciparum* hypothetical proteins with confidence levels comparable to those of similar

[1]Division of Genetics and Genomics, School of Biological Sciences, Nanyang Technological University, Singapore. [2]Bernhard Nocht Institute for Tropical Medicine, Department of Molecular Parasitology, Hamburg, Germany. [3]These authors contributed equally to this work. Correspondence should be addressed to T.-W.G. (gilberger@bni.uni-hamburg.de) or Z.B. (zbozdech@ntu.edu.sg).

Received 8 September; accepted 6 December; published online 27 December 2009; doi:10.1038/nbt.1597

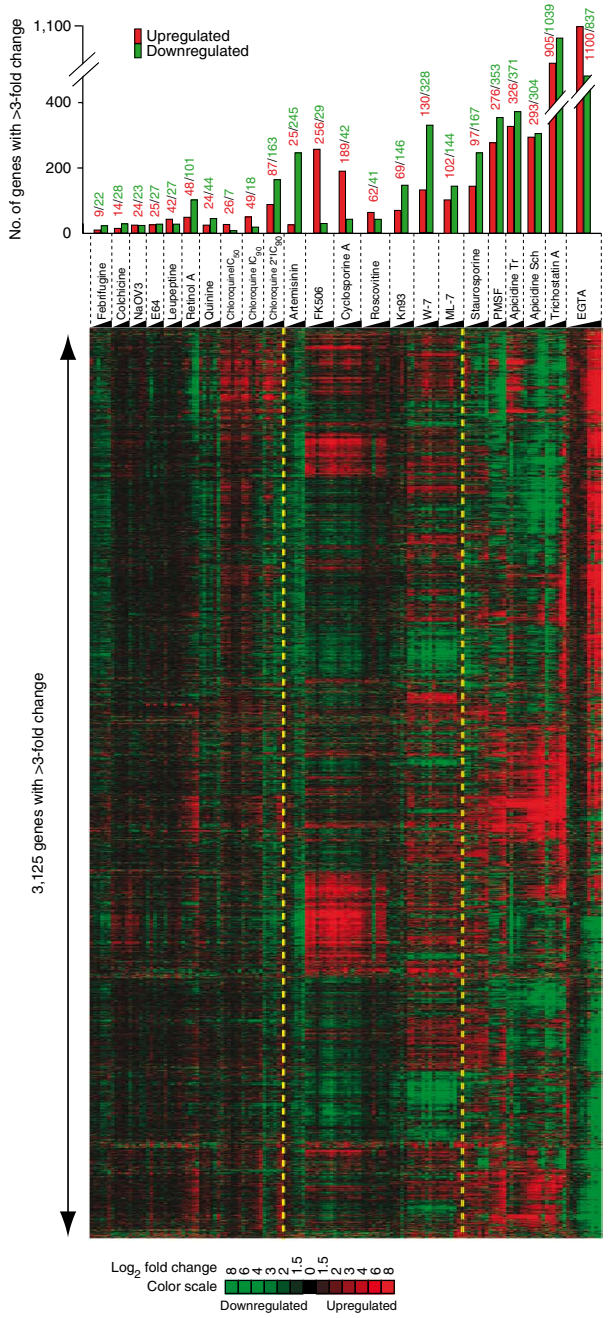


Figure 1 Overview of the gene expression responses of *P. falciparum* to growth perturbation induced by drug or inhibitor treatments. The heatmap summarizes global transcriptional responses to 20 compounds conducted in 23 time-course experiments with a total of 144 microarrays. A total of 3,125 genes that show at least a threefold change in mRNA abundance in at least one experiment are included in the overview data set. The color scale indicates upregulation or downregulation of each individual mRNA transcript compared to the corresponding time point in control untreated cells (**Supplementary Table 1**). The bar diagram (top) indicates the total number of genes that show more than threefold upregulation (red bar) or downregulation (green bar) in each treatment experiment. The number of up- and downregulated genes is also indicated. The treatment experiments were ordered according to the total number of genes with altered expression (more than threefold) and grouped (yellow dashed lines) according to the number of genes with altered levels of their mRNA levels (see text). The treatment experiments were conducted in the time courses indicated along the horizontal axis and genes were arranged using hierarchical clustering.

approaches applied for well-studied model organisms[21,22]. We focused mainly on the late stage (schizont) of the *P. falciparum* intraerythrocytic developmental cycle (IDC) to target the key process of parasite invasion and identified a subnetwork that encompasses 416 genes likely to participate in this process. Using a green fluorescent protein (GFP)-tagging approach, we demonstrate that 31 of 42 genes selected from the subnetwork localize within cellular compartments directly associated with host-cell invasion.

RESULTS

Transcriptional profiling of growth perturbations

We carried out microarray measurements of *P. falciparum* global transcriptional responses to 20 growth-inhibiting compounds (**Fig. 1** and **Supplementary Table 1**). For each compound, synchronized *P. falciparum* cells were exposed to inhibitory concentrations (IC) of 50 (IC_{50}) or 90 (IC_{90}) determined individually for each drug and RNA samples were collected from multiple time points (**Supplementary Table 1**).

A total of 3,125 genes exhibited at least a threefold increase or decrease in transcript level after exposure to at least one chemical stimulus for at least one of the time points after initiating growth perturbation (**Fig. 1** and **Supplementary Table 2**). Using a threefold change in transcript abundance as a cutoff for transcriptional modulation, we loosely classify the transcriptional responses into three compound classes.

The first class induced <50 genes (~1% of the genome) and had an overall transcriptional effect on <250 genes (~5% of the genome). This includes compounds like colchicine, Na_3VO_4, E64, leupeptin and two of the three tested antimalarial drugs, chloroquine and quinine (**Fig. 1**). These results are reminiscent of those in reports that revealed unusually low levels of transcriptional responses to highly toxic antimalarial drugs[16,17]. Despite their low amplitudes, these responses were, however, highly reproducible and specific to each compound[16,17]. In agreement with this, we observed highly reproducible responses of *P. falciparum* to chloroquine (data not shown) that were also dose dependent (26, 49 and 87 genes were induced more than threefold and 194, 257 and 330 genes, more than twofold with IC_{50}, IC_{90} and 2*IC_{90} concentrations, respectively).

We found only moderate overlap between our results and previously published data[17,19]. Compared with these studies, only 12.5% and 10% of the genes whose expression was altered by chloroquine and artemisinin, respectively, were also found to be differentially expressed. Differences in experimental design that might account for these dissimilarities may relate to the considerably higher drug concentrations used previously, different representations of the developmental stages in starting cultures (e.g., asynchronized parasites for chloroquine studies[17]) and different approaches to data analyses (e.g., filtering of genes with stage-specific expression in the artesunate study[19]). Despite these discrepancies, our experiments and the previous published work showed genes with highly reproducible and dose-dependent responses to these malaria drugs. This suggests that, despite their low amplitudes and broad gene representations, transcriptional changes in response to chemical stimuli may reflect physiologically relevant processes involving functionally related genes.

The second class of compounds induced transcription of >50 genes (~1%) and overall involved 250–500 genes (~5–10%). This includes inhibitors of calcium/calmodulin-dependent protein kinases (CDPK; ML-7 and W-7) and the calcineurin pathway (FK506 and cyclosporine A), all of which inhibited the development of the schizont stage (**Supplementary Fig. 1**). We observed striking similarities

© 2010 Nature America, Inc. All rights reserved.

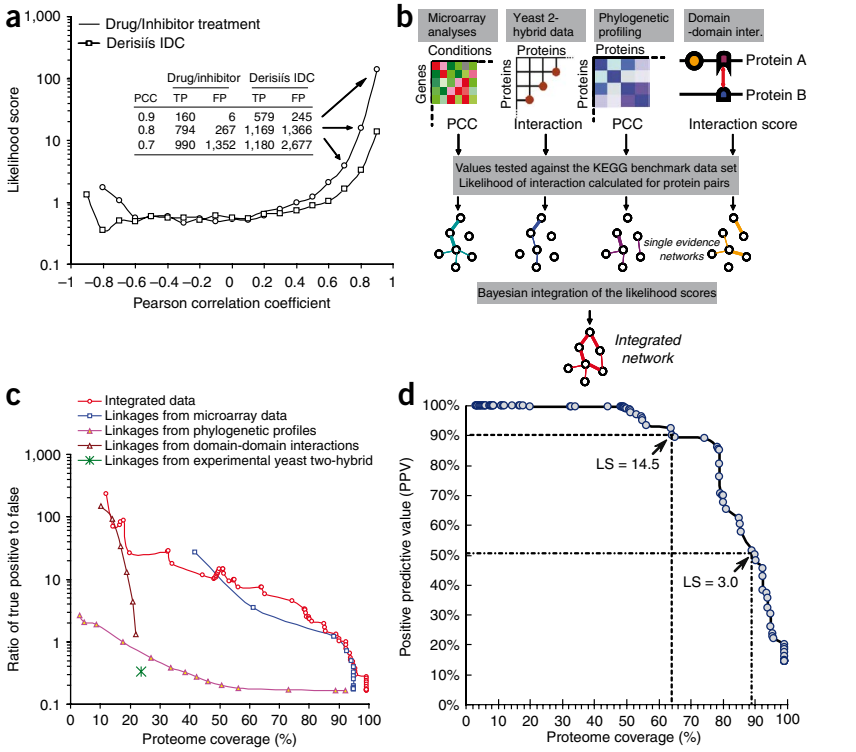


Figure 2 Reconstruction of the PlasmoINT interaction network. (**a**) The plot depicts the likelihood of functional relationships along the correlation of mRNA abundance profiles for all gene pairs in the microarray data. Pearson Correlation Coefficients (PCC) were calculated for every pair of the 492 *P. falciparum* genes with KEGG functional assignments in both perturbation data sets (Drug/inhibitor) and the IDC transcriptome[12]. The numbers of false-positive (FP) and true-positive (TP) gene pairs in the high PCC bins are indicated in the inset table. (**b**) Flow chart describing assembly of the interaction network. The four input data sets were evaluated for protein interaction using a relevant scoring system and score values were tested against the KEGG benchmark to derive the interaction likelihood scores that were used as an input evidence for Bayesian integration. For more details on KEGG benchmark scoring and network building, see **Supplementary Table 3.** (**c**) The relationship between proteome coverage of the individual input data sets (microarray data, phylogenetic profiles, domain-domain interaction and yeast two-hybrid system) and TP/FP ratio thresholds illustrates the contribution of each individual input to the integrated network data set. (**d**) The predictive precision rates (positive predictive value, PPV) at different likelihood score cutoffs were evaluated by tenfold cross-validation and plotted against the proteome coverage. Each dot of the ratio represents an average of ten cross-validations at a particular likelihood score cutoff. The vertical dashed line shows the likelihood score cutoffs and proteome coverage corresponding to the PPV (PPV = TP/(TP + FP)) 50% and 90% (likelihood score thresholds (LS) of 3 and 14.5). At these ratios, TP/FP was equal to 1 (~50% confidence) and 9 (~90% confidence), respectively.

© 2010 Nature America, Inc. All rights reserved.

in transcriptional responses induced by inhibitors within each class, which suggests that their inhibitory effect in *P. falciparum* may be very specific (**Fig. 1**). Moreover, there is only a limited overlap between the transcriptional responses induced by the CDPK and calcineurin inhibitors. This suggests that these two types of intracellular signaling pathways play specific, nonoverlapping roles in *P. falciparum* parasites that are both connected to transcriptional regulation.

The third class of compounds was able to induce transcription of >250 genes (~5%) and overall involved >500 genes (~10%). These include EGTA, phenylmethylsulfonyl fluoride, staurosporine, trichostatin A and apicidin (**Fig. 1**). With the exception of apicidin, these responses were compatible with an arrest in IDC development, indicating that the inhibitory effects of these compounds are associated with mechanisms that regulate the *P. falciparum* life cycle (**Supplementary Fig. 2**). In contrast, apicidin and to some degree trichostatin A (both histone deacetylase inhibitors) caused a general deregulation of the IDC transcriptional cascade by derepression of genes that are normally suppressed at both the trophozoite and schizont stages.

Reconstruction of a probabilistic gene functional network

To evaluate co-transcriptional properties of functionally related genes, we calculated the Pearson correlation coefficient (PCC) between transcription profiles of a subset of 492 genes that can be assigned to at least one pathway defined by the Kyoto Encyclopedia of Genes and Genomes (KEGG)[23]. Overall, we observed a disproportionately high number of functionally related genes being transcriptionally co-regulated (PCC > 0.6) (**Fig. 2a** and Online Methods). In comparison with the *P. falciparum* IDC transcriptome[12], the enrichment of functionally related genes was improved by 1.6-, 3.5- and 11-fold for the 0.7, 0.8 and 0.9 PCC thresholds, respectively (**Fig. 2a**). This high occurrence of transcriptional co-regulation among functionally related genes suggests a good potential of the perturbation data set for functional gene predictions. Hence, we used it as a core data set for the assembly of a probabilistic network in which we integrated this data set with additional inputs: (i) phylogenetic profiles with sequence homology values (E-values) of all 5,363 *P. falciparum* protein sequences to their orthologs in 210 sequenced genomes; (ii) domain-domain interactions [24]; and, (iii) yeast two-hybrid interactions[25] (**Fig. 2b** and Online Methods). In addition, the perturbation microarray data were combined with the IDC transcriptomes from three *P. falciparum* laboratory strains[26] and four field isolates[27].

To reconstruct the probabilistic network, we used the KEGG gold standard data set to calculate the likelihood score of protein interaction evidence from all four input data sets (**Supplementary Table 3**) and subsequently integrated these scores into the final score using a Bayesian integration approach (**Fig. 2b** and Online Methods). Overall, we established integrated likelihood scores for 14,168,597 functional linkages between 5,374 *P. falciparum* proteins (99.2% of the proteome). In general, the integrated likelihood scores provided higher proteome coverage than each of the individual input data sets at all probability thresholds (**Fig. 2c**). In contrast to the domain-domain interaction data set, which provides high-accuracy predictions for a small proportion of the proteome (~10%), the transcriptome data and phylogenetic profiles can provide high proteome coverage. However, their predictive values are consistently lower. In our calculations, we observed low accuracy for the protein-protein interaction data set based on the two-hybrid system[25]. This data set therefore provides a low contribution to the final likelihood scores (**Fig. 2c**).

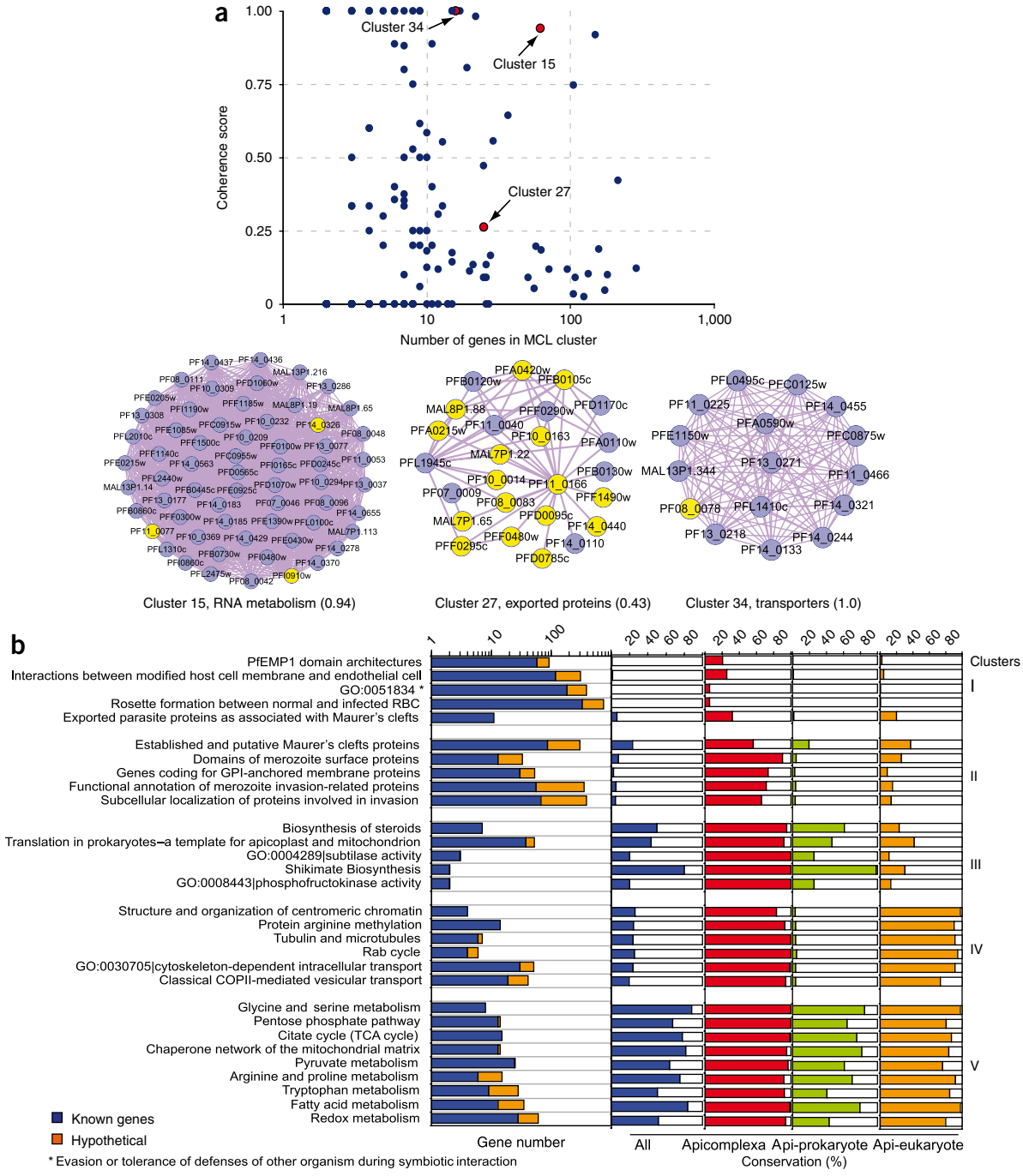


Figure 3 MCL- and WNC-based functional predictions and their functional categorizations. (**a**) Summary of the 208 MCL clusters depicted as a scatter plot with the number of genes plotted against their coherence score. Coherence score 0 corresponds to MCL cluster without any functionally characterized proteins. Examples of three clusters with high- and medium-coherence scores are indicated in the scatter plot and also drawn (below) with the functionally characterized (purple) and hypothetical proteins (yellow) linked by edges that correspond to functional links with >90% confidence. (**b**) The conservation of different functional pathways across 210 genomes including 155 prokaryotes, 6 apicomplexa and 49 other eukaryotes is summarized and indicated for selected functional gene groups (for the full list, see **Supplementary Table 5**). The conservation of each pathway is calculated independently as the fraction of the number of species containing potential homologs (reciprocal BLASTP hit, E-value ≤ 10^{-10}) according to four categories: total 210 genomes (the second panel, blue bar), apicomplexa (third panel, red bar), prokaryotes plus apicomplexa (fourth panel, green bar) and eukaryote plus apicomplexa (right panel, orange bar). Pathways were classified into five categories: genes specific to *P. falciparum* (cluster I), genes conserved in apicomplexa (II), genes conserved in apicomplexa and prokaryotes (III), genes conserved in apicomplexa and other eukaryotes (IV) and genes conserved in all 210 genomes (V). The total number of functionally characterized and hypothetical genes in each category are displayed similarly to **a**. Api-eukaryote, genes conserved in apicomplexans and eukaryotes; api-prokaryote, genes conserved in apicomplexans and prokaryotes.

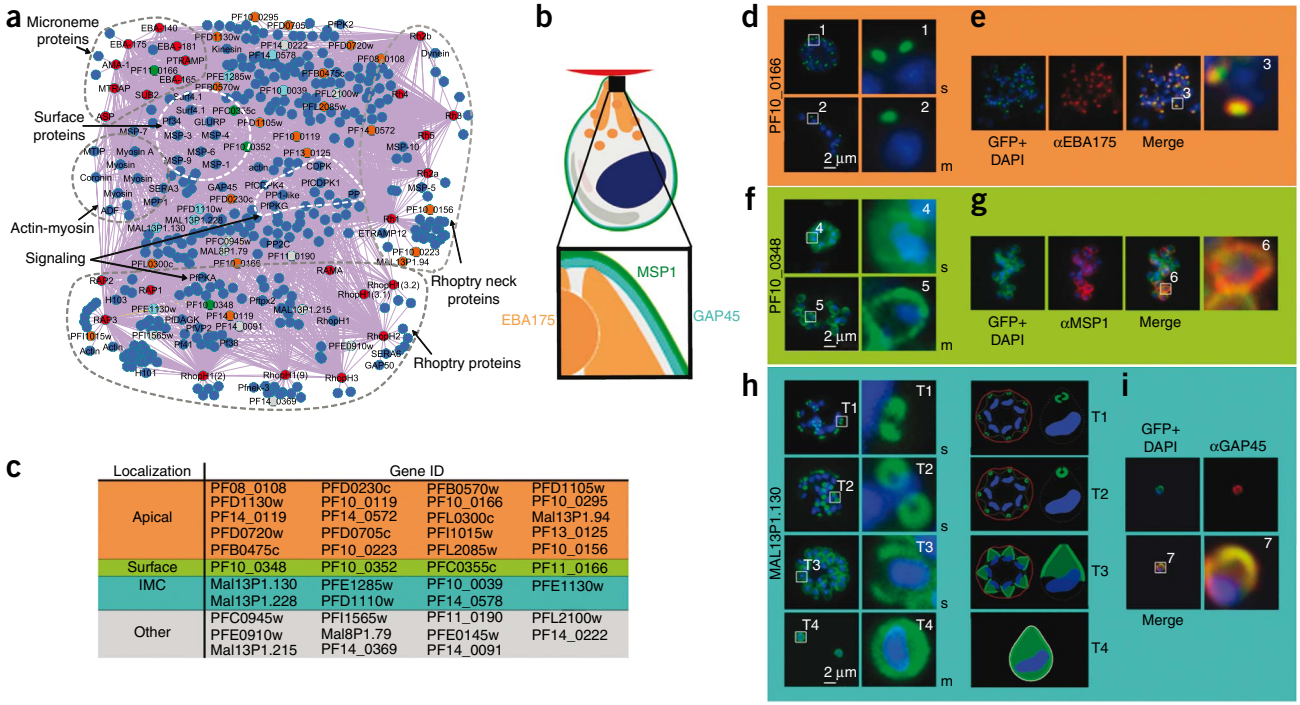
© 2010 Nature America, Inc. All rights reserved.

Using the calculated functional linkages, we assembled two interaction networks based on likelihood score thresholds that correspond to 50% (339,721 linkages for 89% of proteome) and 90% confidence precision rates (72,748 linkages for 68% of proteome) (**Fig. 2d** and Online Methods).The connectivity of both the 50% and 90% confidence networks fits a power-law distribution with power (λ) values of 0.93 and 1.14, respectively (**Supplementary Fig. 3**). This distribution represents a typical scale-free network, well-known for protein-protein interaction networks in eukaryotic cells[28]: a small number of highly connected nodes (hubs) are linked to a larger number of less connected nodes and so on.

Modular analysis and network-based functional predictions

In the next step, we used two parallel approaches to explore the assembled network for the prediction of *P. falciparum* hypothetical protein function. First, we used the Markov cluster (MCL) algorithm[29] to define significant clusters of highly interconnected genes in the network. We used a coherence score to test enrichment of every single cluster for genes involved in a particular pathway. This analysis not only tests the quality of the network but also generates functional predictions for hypothetical genes that fall into these clusters (**Fig. 3a**). For this work, we used the 90% confidence network to provide the most conservative assessment of the network quality. Second, we used the weighted neighbor-counting (WNC) method to derive functional prediction for the hypothetical proteins. For this, we explored the 50% confidence network to maximize the number of functional predictions for hypothetical proteins. The confidence of these predictions was assessed by a 'leave-one-out' analysis[30] that is based on the efficiency of recalling functional predictions of previously characterized genes (**Supplementary Fig. 4**).

MCL identified 208 modules in the 90% confidence network, resulting in 3,029 genes being assigned to at least one of the 106 modules with functional assignments (**Fig. 3a** and **Supplementary Table 4**). The MCL modules represent many pathways conserved across the eukaryotic species (e.g., RNA metabolism) or specific to *P. falciparum* (e.g., proteins exported to the host cell cytoplasm, "exported proteins"), as well as coherent functional groups (e.g., transporters) (**Fig. 3a**). The functions of 1,376 hypothetical genes can be predicted



Localization	Gene ID			
Apical	PF08_0108	PFD0230c	PFB0570w	PFD1105w
	PFD1130w	PF10_0119	PF10_0166	PF10_0295
	PF14_0119	PF14_0572	PFL0300c	Mal13P1.94
	PFD0720w	PFD0705c	PFI1015w	PF13_0125
	PFB0475c	PF10_0223	PFL2085w	PF10_0156
Surface	PF10_0348	PF10_0352	PFC0355c	PF11_0166
IMC	Mal13P1.130	PFE1285w	PF10_0039	PFE1130w
	Mal13P1.228	PFD1110w	PF14_0578	
Other	PFC0945w	PFI1565w	PF11_0190	PFL2100w
	PFE0910w	Mal8P1.79	PFE0145w	PF14_0222
	Mal13P1.215	PF14_0369	PF14_0091	

Figure 4 Blueprint of the protein network implicated in merozoite invasion. (**a**) Subnetwork associated with merozoite invasion process. This subnetwork has a total of 2,417 links (purple lines) that are derived from the 90% confidence network and link the 25 reference genes to 25 core apical proteins (marked with red circles) with 418 proteins that include the experimentally validated (colored circles) and other proteins (blue circles). The forty-two proteins whose intracellular localization were studied are represented by a corresponding color; apical proteins (orange), merozoite surface proteins (green), IMC (turquoise) and other localization (gray). The core proteins and other previously characterized proteins were grouped manually based on their functional assignments. The dotted lines outline areas with functionally related proteins previously linked with invasion, such as microneme proteins, actin and myosin. (**b**) Schematic representation of an invasive merozoite. The apical organelles are depicted in orange, the IMC in turquoise and the surface in green. Examples for compartment-specific marker proteins are given. (**c**) Synopsis of subcellular localization of 42 proteins predicted to be involved in invasion. Proteins are grouped into either apical (orange), surface (green), IMC (turquoise) or other (gray; cytosolic, apicoplast or mitochondrial), according to their predominant localization. (**d**–**i**) Representative localization for one member of each group in late schizonts and free merozoites. Boxed regions are numbered and depicted in higher magnification to the right. The nucleus is stained with DAPI (blue). PF10_0166-GFP (green) localized to the apical region of schizonts (s) and free merozoites (m) in unfixed parasites (**d**). PF10_0166-GFP co-localized with the microneme protein EBA175 (red) in fixed parasites (**e**). PF10_0348-GFP (green) localized to the surface of schizonts and free merozoites in unfixed parasites (**f**). PF10_0348-GFP co-localized with the surface protein MSP-1 (red) in fixed parasites (**g**). Dynamics of MAL13P1.130-GFP (green) during schizogony in unfixed parasites. In early schizogony (T1), MAL13P1.130-GFP emerged as a cramp-like-structure at the apical tip of forming merozoites (**h**). This structure develops to be ring-like (T2) before becoming evenly distributed at the periphery of the nascent merozoite (T3-4). The third row shows a schematic representation. For confocal three-dimensional reconstitution, see **Supplementary Movies 1–3**. MAL13P1.130-GFP co-localized with the IMC protein GAP45 (red) in fixed parasites (**i**).

© 2010 Nature America, Inc. All rights reserved.

by their association to these modules, whose confidence is represented by the coherence scores. The MCL analysis suggests that the assembled network detects functionally related genes with sufficient precision. The WNC approach allows (functional) explorations of unknown genes even outside of the identified modules and generated predictions for 2,545 hypothetical proteins (95% in the genome) that can be assigned to 216 functional terms (**Supplementary Fig. 5** and **Supplementary Table 5**).

Taking advantage of the phylogenetic profiles (see above), we investigated the overall evolutionary conservation of the derived functional groups with the newly assigned genes (**Fig. 3b**). Only a small number of functional gene groups are restricted to *P. falciparum* and exhibit either no or low sequence homology to genes in other organisms, including closely related apicomplexan species. The majority of these represent the subtelomeric gene families encoding several classes of surface antigens, such as *var*, *rifin* and *stevor*, and proteins associated with Maurer's clefts (**Fig. 3b**, cluster I). Parasite invasion dominates the functional cluster that is highly conserved among apicomplexans but diverges from all other eukaryotic and prokaryotic species (**Fig. 3b**, cluster II). Cluster III depicts several *P. falciparum* functions that have a prokaryotic origin such as steroid biosynthesis (a term assigned by KEGG, corresponding to *P. falciparum* isoprenoid synthesis), translation in genes of the mitochondria and apicoplasts (non-photosynthetic plastids found in most Apicomplexa) and three homologs of proteins involved in subtilisin protease activity. Moreover, the WNC analysis assigned many new proteins to the majority of the highly conserved functional groups that are either of eukaryotic (cluster IV) or prokaryotic origin (cluster V). It is possible that many of the newly annotated genes represent evolutionarily diverse factors of these otherwise well-conserved, and thus potentially essential, pathways. The precision rates for these functional terms provide a measure of confidence for these functional predictions and help to identify candidates for previously unrecognized molecular factors that are essential for the growth, development and virulence of *P. falciparum*.

Proteins implicated in *P. falciparum* merozoite invasion

Invasion of the host's red blood cells by a specialized invasive form called the merozoite is a key step in the *P. falciparum* life cycle. To validate the predictive potential of our approach, we explored the utility of our network to identify genes associated with merozoite invasion. Merozoite invasion involves multiple molecular mechanisms ranging from specific ligand-receptor interactions, actin-myosin motility, protease activities, protein translocation and signaling[31–33]. It is mediated by an unknown number of proteins and is of high interest for drug and vaccine development because interference with this crucial biological process holds the potential to disrupt the parasite's life cycle. Although >50 proteins have been previously linked with this process, gaps remain in our understanding of the molecular mechanisms that mediate the entire invasion process. To provide a comprehensive picture of the invasion process, we generated a subnetwork of proteins that are directly connected to 25 previously established invasion-associated proteins in the 90% confidence interaction network (**Fig. 4a**). Overall, this subnetwork contains 418 proteins, including 155 with a predicted function and 263 hypothetical proteins (**Supplementary Table 6**). The subnetwork compiles the majority of proteins previously linked with invasion-like apical organelle proteins, glycosylphosphatidylinisotol-anchored surface proteins, actin-myosin motor components and signal transduction proteins. It also includes 43 out of 56 proteins recently predicted to be associated with cellular compartments of the merozoite invasion machinery[33]. Finally, 230 out of all 263 hypothetical proteins represented in the invasion subnetwork were also predicted by WNC as merozoite invasion factors.

For the functional validations, we initially selected 70 proteins from this invasion process protein subnetwork. For this selection, we prioritized proteins with a high WNC score (**Supplementary Table 5**) and gene length ≤2 kb (to facilitate cloning and expression of these proteins in *P. falciparum* transfection experiments). Open reading frames were fused with GFP and expressed ectopically in *P. falciparum* under the control of an appropriate promoter mimicking the expression profile of the endogenous allele[34]. Of these, 63 proteins could be expressed as GFP-fusion proteins in transgenic parasites, of which 42 resulted in a defined intracellular localization (**Fig. 4** and **Supplementary Fig. 6b**). From the remaining 21 GFP fusions, 11 were not expressed at sufficient levels and 10 were discarded because of retention in the endoplasmic reticulum that might be caused by the bulky GFP moiety, as described previously[34] (data not shown).

The remaining 42 proteins can be grouped according to their localization (**Fig. 4b**,**c**). The largest group consists of 20 proteins that showed a predominantly apical distribution in maturing schizonts and in free merozoites after rupture (**Fig. 4d**,**e** and **Supplementary Fig. 6a**). The second group is represented by four proteins with GFP distributed in the periphery of the parasite (**Fig. 4f**,**g** and **Supplementary Fig. 6a**). The third group (7 proteins) localizes to the inner membrane complex (IMC)[35], a membranous system underlying the plasma membrane and involved in the structural integrity and motility of invasive parasites[35–37]. These proteins display a unique spatial dynamic during schizogony reflecting the biogenesis of this compartment (**Fig. 4h**,**i**, **Supplementary Fig. 6a** and **Supplementary Movies 1–3**). The remaining 11 proteins revealed localizations that are not obviously associated with invasion, although this does not exclude them from playing a role in this process (**Supplementary Fig. 6a**,**b**). Examples are proteins that localize to the cytosol including the putative kinase PFC0945w and the profilin homolog PFI1565w. In summary, 31 out of 42 selected proteins are associated with structures known to be directly involved in invasion. This demonstrates that the functional predictions based on our approach can lead to the identification of new putative targets for malaria intervention strategies.

DISCUSSION

Until now, the potential of using transcriptional profiling of growth perturbation for functional analyses of malaria parasites has been underappreciated. We demonstrate that functionally related genes share similar transcriptional profiles to a diverse panel of chemical perturbations, which suggests that many of these genes share regulatory mechanisms responsive to external stimuli (**Fig. 2a**). This suggests that transcriptional profiling may be a viable approach for functional genomics of human malaria parasites and can provide insights into parasite biology. Although mRNA decay was proposed to make a major contribution to the regulation of gene expression in *P. falciparum*[38], our data suggest that the responses to chemically induced growth perturbations are associated with transcription[39], rather than mRNA stability. We find essentially no relationship between our mRNA profiles and the previously established pattern of mRNA decay (data not shown).

The sensitivity of *P. falciparum* transcription to chemical stimuli has enabled us to make gene-function predictions not included in previous network-based approaches[10,11,40]. Our 90%-confidence network (termed PlasmoINT) contains close to 6 times more linkages and 2.5 times more proteins than PlasmoMAP[10], hitherto the most reliable published *P. falciparum* interaction network. In addition, there are five times as many linkages, which are supported by two or more types of evidence (**Supplementary Table 7**). These additions can be attributed mainly to the extensive transcriptional data and

© 2010 Nature America, Inc. All rights reserved.

inclusion of the annotations from the functional genomic database, the Malaria Parasite Metabolic Pathways[41]. This enables us to provide more accurate reconstructions of the majority of metabolic and cellular pathways (**Supplementary Fig. 7**) and thus more confident functional gene predictions. We also compared the Gene Ontology (GO) terms assigned to the *P. falciparum* genes by PlasmoINT with those assigned by the ontology-based pattern identification (OPI) method[40]. There is, however, only a limited congruity between these two studies with only 13%, 22% and 37% of the genes matching the predictions between the OPI and PlasmoINT-assigned GO terms at 4th, 3rd, and 2nd level, respectively. Although the relatively low level of consistency between these two methods is surprising, it is worth noting that the 47% recall precision of PlasmoINT contrasts (Online Methods and **Supplementary Fig. 4**), with only 18% precision for OPI. Similarly, the increased precision of the PlasmoINT prediction may result from the inclusion of the perturbation data set, which captures the finer pattern of transcriptional regulation in response to growth perturbation compared to the development stage–specific expression used by OPI. In addition to the supplementary material, the data presented in this manuscript have been compiled to a searchable database available online (http://zblab.sbs.ntu.edu.sg/), which we plan to update periodically.

As invasion of the host cell is essential for survival of *P. falciparum* and is a key target for new malaria intervention strategies, we used the functional annotations obtained from our interactome to experimentally validate proteins predicted to be associated with the invasion process. Of the 42 proteins that could be localized in the parasite, 31 were predominantly targeted either to the apical organelles, the parasite periphery or the IMC (**Fig. 4** and **Supplementary Fig. 6**): all key compartments for host cell invasion. Interestingly, 11 out of the 31 proteins contain neither a predicted signal peptide nor a transmembrane domain. Both of these are characteristic for proteins previously associated with the invasion machinery, highlighting the power of this approach. For instance, network prediction enabled us to identify novel proteins associated with the IMC such as MAL13P1.228, PF14_0578 or PFE1130w. This notion is further supported by the identification and localization of PFB0570w and PFD1105w, two proteins previously associated with the rhoptries, (exocytotic organelles containing many proteins with adhesive functions[42,43]), PF10_0348 and PF10_0352, two proteins of the merozoite surface protein super-family[44,45], and MAL13.P1.130 and PFD1110w, two newly localized IMC proteins[46,47]. Further confirmation of the utility of this study came from the identification and localization of PFD0230c. This unique serine protease was recently identified in a forward chemical genetic screen as one of the key regulators for merozoite egress[48]. Although it will be crucial to further validate these novel proteins and to extend their characterization, this subnetwork of proteins predicted to be involved in invasion offers a comprehensive blueprint of this process at the molecular level. These results may be useful for functional studies of each identified protein and rational drug and vaccine development.

METHODS

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

Accession codes. Gene Expression Omnibus: GSE19468.

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

ACKNOWLEDGMENTS

This project was funded by the Academic Research Council of the Singapore Ministry of Education (grant no. ARC 11/05 M45080011), Singaporean National Medical Research Council (grant # IRG07Nov030) and the Deutsche Forschungsgemeinschaft (GI312 and GRK1459). The authors also thank B.D. Wastuwidyaningtyas and S. Tan for excellent technical assistance with the microarray experiments, K. Jurries for graphical assistance and A. Law, R. Stanway, T. Voss and H. Hoppe for critical reading of the manuscript. We are grateful to M. Blackman (National Institute for Medical Research, London) for providing the MSP-1 antibody, to P. Sharma (National Institute of Immunology, New Delhi) for providing the GAP45 antibody and to Jacobus Pharmaceuticals for providing WR99210.

AUTHOR CONTRIBUTIONS

G.H. and Z.B. performed the computation and data analysis. G.H., A.C., P.R.P., T.S., T.-W.G. and Z.B. drafted the paper. S.M., G.H., S.C. and B.K.C. performed the microarray experiments. A.C., M.K., S.H., K.E. and T.S. cloned the genes, generated the transgenic parasites and carried out the microscopy. All authors read and approved the final manuscript.

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

1. Vitoria, M. *et al.* The global fight against HIV/AIDS, tuberculosis, and malaria: current status and future perspectives. *Am. J. Clin. Pathol.* **131**, 844–848 (2009).
2. Hay, S.I. *et al.* A world malaria map: Plasmodium falciparum endemicity in 2007. *PLoS Med.* **6**, e1000048 (2009).
3. Molyneux, D.H. Control of human parasitic diseases: Context and overview. *Adv. Parasitol.* **61**, 1–45 (2006).
4. Nwaka, S. & Hudson, A. Innovative lead discovery strategies for tropical diseases. *Nat. Rev. Drug Discov.* **5**, 941–955 (2006).
5. Gardner, M.J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498–511 (2002).
6. Balu, B., Shoue, D.A., Fraser, M.J., Jr. & Adams, J.H. High-efficiency transformation of Plasmodium falciparum by the lepidopteran transposable element piggyBac. *Proc. Natl. Acad. Sci. USA* **102**, 16391–16396 (2005).
7. Maier, A.G. *et al.* Exported proteins required for virulence and rigidity of *Plasmodium falciparum*-infected human erythrocytes. *Cell* **134**, 48–61 (2008).
8. Winzeler, E.A. Applied systems biology and malaria. *Nat. Rev. Microbiol.* **4**, 145–151 (2006).
9. Kato, N. *et al.* Gene expression signatures and small-molecule compounds link a protein kinase to Plasmodium falciparum motility. *Nat. Chem. Biol.* **4**, 347–356 (2008).
10. Date, S.V. & Stoeckert, C.J. Jr. Computational modeling of the Plasmodium falciparum interactome reveals protein function on a genome-wide scale. *Genome Res.* **16**, 542–549 (2006).
11. Wuchty, S. & Ipsaro, J.J. A draft of protein interactions in the malaria parasite *P. falciparum*. *J. Proteome Res.* **6**, 1461–1470 (2007).
12. Bozdech, Z. *et al.* The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. *PLoS Biol.* **1**, E5 (2003).
13. Le Roch, K.G. *et al.* Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* **301**, 1503–1508 (2003).
14. Hughes, T.R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
15. MacCarthy, T., Pomiankowski, A. & Seymour, R. Using large-scale perturbations in gene network reconstruction. *BMC Bioinformatics* **6**, 11 (2005).
16. Ganesan, K. *et al.* A genetically hard-wired metabolic transcriptome in *Plasmodium falciparum* fails to mount protective responses to lethal antifolates. *PLoS Pathog.* **4**, e1000214 (2008).
17. Gunasekera, A.M., Myrick, A., Le Roch, K., Winzeler, E. & Wirth, D.F. *Plasmodium falciparum*: genome wide perturbations in transcript profiles among mixed stage cultures after chloroquine treatment. *Exp. Parasitol.* **117**, 87–92 (2007).
18. Oakley, M.S. *et al.* Molecular factors and biochemical pathways induced by febrile temperature in intraerythrocytic Plasmodium falciparum parasites. *Infect. Immun.* **75**, 2012–2025 (2007).
19. Natalang, O. *et al.* Dynamic RNA profiling in *Plasmodium falciparum* synchronized blood stages exposed to lethal doses of artesunate. *BMC Genomics* **9**, 388 (2008).
20. Tamez, P.A. *et al.* An erythrocyte vesicle protein exported by the malaria parasite promotes tubovesicular lipid import from the host cell surface. *PLoS Pathog.* **4**, e1000118 (2008).
21. Groth, P., Weiss, B., Pohlenz, H.D. & Leser, U. Mining phenotypes for gene function prediction. *BMC Bioinformatics* **9**, 136 (2008).
22. Kim, W.K., Krumpelman, C. & Marcotte, E.M. Inferring mouse gene functions from genomic-scale data using a combined functional network/classification strategy. *Genome Biol.* **9** Suppl 1, S5 (2008).
23. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, D277–D280 (2004).
24. Lee, H., Deng, M., Sun, F. & Chen, T. An integrated approach to the prediction of domain-domain interactions. *BMC Bioinformatics* **7**, 269 (2006).
25. LaCount, D.J. *et al.* A protein interaction network of the malaria parasite Plasmodium falciparum. *Nature* **438**, 103–107 (2005).

© 2010 Nature America, Inc. All rights reserved.

26. Llinas, M., Bozdech, Z., Wong, E.D., Adai, A.T. & DeRisi, J.L. Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains. *Nucleic Acids Res.* **34**, 1166–1173 (2006).
27. Mackinnon, M.J. *et al.* Comparative transcriptional and genomic analysis of Plasmodium falciparum field isolates. *PLoS Pathog.* **5**, e1000644 (2009).
28. Barabasi, A.L. & Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).
29. Krogan, N.J. *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637–643 (2006).
30. Deng, M., Zhang, K., Mehta, S., Chen, T. & Sun, F. Prediction of protein function using protein-protein interaction data. *J. Comput. Biol.* **10**, 947–960 (2003).
31. Cowman, A.F. & Crabb, B.S. Invasion of red blood cells by malaria parasites. *Cell* **124**, 755–766 (2006).
32. Soldati, D., Foth, B.J. & Cowman, A.F. Molecular and functional aspects of parasite invasion. *Trends Parasitol.* **20**, 567–574 (2004).
33. Haase, S. *et al.* Characterization of a conserved rhoptry-associated leucine zipper-like protein in the malaria parasite *Plasmodium falciparum*. *Infect. Immun.* **76**, 879–887 (2008).
34. Treeck, M. *et al.* A conserved region in the EBL proteins is implicated in microneme targeting of the malaria parasite *Plasmodium falciparum*. *J. Biol. Chem.* **281**, 31995–32003 (2006).
35. Baum, J. *et al.* A conserved molecular motor drives cell invasion and gliding motility across malaria life cycle stages and other apicomplexan parasites. *J. Biol. Chem.* **281**, 5197–5208 (2006).
36. Baum, J. *et al.* A malaria parasite formin regulates actin polymerization and localizes to the parasite-erythrocyte moving junction during invasion. *Cell Host Microbe* **3**, 188–198 (2008).
37. Morrissette, N.S. & Sibley, L.D. Disruption of microtubules uncouples budding and nuclear division in Toxoplasma gondii. *J. Cell Sci.* **115**, 1017–1025 (2002).
38. Shock, J.L., Fischer, K.F. & DeRisi, J.L. Whole-genome analysis of mRNA decay in *Plasmodium falciparum* reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle. *Genome Biol.* **8**, R134 (2007).
39. De Silva, E.K. *et al.* Specific DNA-binding by apicomplexan AP2 transcription factors. *Proc. Natl. Acad. Sci. USA* **105**, 8393–8398 (2008).
40. Zhou, Y. *et al.* Evidence-based annotation of the malaria parasite's genome using comparative expression profiling. *PLoS One* **3**, e1570 (2008).
41. Ginsburg, H. Caveat emptor: limitations of the automated reconstruction of metabolic pathways in *Plasmodium*. *Trends Parasitol.* **25**, 37–43 (2009).
42. Chattopadhyay, R. *et al.* PfSPATR, a *Plasmodium falciparum* protein containing an altered thrombospondin type I repeat domain is expressed at several stages of the parasite life cycle and is the target of inhibitory antibodies. *J. Biol. Chem.* **278**, 25977–25981 (2003).
43. Wickramarachchi, T., Devi, Y.S., Mohmmed, A. & Chauhan, V.S. Identification and characterization of a novel *Plasmodium falciparum* merozoite apical protein involved in erythrocyte binding and invasion. *PLoS ONE* **3**, e1732 (2008).
44. Pearce, J.A., Mills, K., Triglia, T., Cowman, A.F. & Anders, R.F. Characterisation of two novel proteins from the asexual stage of *Plasmodium falciparum*, H101 and H103. *Mol. Biochem. Parasitol.* **139**, 141–151 (2005).
45. Wickramarachchi, T. *et al.* A novel *Plasmodium falciparum* erythrocyte binding protein associated with the merozoite surface, PfDBLMSP. *Int. J. Parasitol.* **39**, 763–773 (2009).
46. Bullen, H.E. *et al.* A novel family Of apicomplexan glideosome associated proteins With an inner-membrane anchoring role. *J. Biol. Chem.* **284**, 25353–25363 (2009).
47. Rayavara, K. *et al.* A complex of three related membrane proteins is conserved on malarial merozoites. *Mol. Biochem. Parasitol.* **167**, 135–143 (2009).
48. Arastu-Kapur, S. *et al.* Identification of proteases that regulate erythrocyte rupture by the malaria parasite *Plasmodium falciparum*. *Nat. Chem. Biol.* **4**, 203–213 (2008).

© 2010 Nature America, Inc. All rights reserved.

ONLINE METHODS

Parasite culture, treatment and microarray. The perturbation time courses were performed with 2% hematocrit and 5% parasitemia cultures. Parasites were treated with appropriate drug or compound concentrations and collected at 5–8 time points taken at regular time intervals (30–120 min). A total of 247 microarray experiments were carried out, including 29 drug treatment time courses with 20 compounds and corresponding untreated controls from different drug or inhibitor treatment (**Supplementary Table 1**). Genome-wide gene expression profiling was conducted using long oligonucleotides representing all 5,363 *P. falciparum* genes as previously described[49]. The expression data were normalized using linear normalization and background filtering as implemented by the NOMAD database (http://derisilab.ucsf.edu) and described[12]. Subsequently each gene profile was represented by an average expression value calculated as an average of all oligonucleotides representing a particular gene. For the final data set we considered only the genes for which at least 80% of time points in each time course yielded a positive expression signal.

For the final microarray input data sets for the reconstruction of the gene functional network, we incorporated the perturbation data set with the IDC transcriptome of laboratory strains (3D7, Dd2 and HB3, 148 microarray experiments)[12] and four lab isolates[27]. To indicate the strength of functional association of each gene pair by gene expression profiles, PCCs were calculated independently across each data set first and intergraded by a new technique that we term the "optional average" method. Briefly, Fisher's z-transform[50] was used to average two PCCs from two independent IDC transcriptomes and compared to the PCC from perturbation data. If the latter is smaller, the final PCC is the PPC from perturbation data. Otherwise, the final PCC is equal to the average PCC from two tested data sets defined by the Fisher's z-transform.

The input data sets for the network construction. For the network assembly we incorporate the microarray data set (above) with three additional inputs. (i) The phylogenetic profiles were calculated for all *P. falciparum* genes obtained from the PlasmoDB version 5.4 (http://www.plasmodb.org/download/). Using BLASTP, the protein sequences of *P. falciparum* were compared with 210 reference organisms, including 155 prokaryotes and 55 eukaryotes available from the NCBI and the ENSEMBL. For each protein a vector was generated with elements p_{ij} where $p_{ij} = -1/\log E_{ij}$ where Eij represents the E-value of the gene (i) ortholog in the genome (j). As a metric of phylogenetic profile similarity, the mutual information was calculated with the histograms of p_{ij} values, binned in 0.01 intervals, as previously described[51]. The mutual information scores were divided into 15 bins for the KEGG benchmark test (**Supplementary Table 3**). (ii) For the domain-domain interaction evidence, we carried out Hidden Markov Model–based predictions of all functional domains defined by the PFAM database in all 5,363 *P. falciparum* proteins. For this we use the set of domain-domain interactions as defined previously[24]. Based on the confidence scores provided by the Lee database[24], the gene pairs were subsequently divided into six bins and tested against the KEGG benchmark. (iii) From the yeast two-hybrid system protein-protein interactions were obtained from the previous publication[25] and all 2,811 interactions among 1,308 *P. falciparum* proteins were tested against the KEGG benchmark as one bin (**Fig. 2b**).

Calculation of the likelihood scores using the KEGG gold standard benchmark data set. The KEGG 'gold standard' benchmark data set includes 492 annotated *P. falciparum* genes that can be assigned to 71 metabolic or cellular pathways defined by the KEGG database[23]. This defines 11,046 positive pairs of genes that belong to pathways with >3 genes. The negative set includes 61,721 gene pairs that do not fall into a common pathway. **Supplementary Table 3** online shows the parameters of naive Bayesian network of all data sets based on this reference data set. The ratio of true to false positive in **Figure 2c** is calculated using the KEGG benchmark data set and it reflects measure of agreement of the functional relationship of each gene pair as a function of the individual scoring systems (e.g., PCC for microarray data and phylogenetic profiling). The calculated likelihood scores reflect the functional relationships between *P. falciparum* genes and are applicable as input values for assembling a probabilistic interactome network.

Building the interaction network Integration of the data sets by the Bayesian probabilistic model was carried out as previously described[10]. In principle, the final likelihood score is determined as:

$$\text{Likelihood Score (LS)} = LS_{PPC} \times LS_{PHY} \times LS_{PPI} \times LS_{Domain}$$

PPC, microarray input; *PHY*, phylogenetic profile input; *PPI*, yeast two-hybrid input; *Domain*, domain-domain interaction input.

We performed a tenfold cross-validation to evaluate the overall performance of the prediction. Briefly, first the positive and negative benchmarks were randomly divided into ten separate equal sets, and nine of them were used as the training set to calculate the likelihood scores and the remaining one set as the test to identify the positives and negatives. We ran this process ten times so that each of the ten sets was a test set and the remaining nine constituted the training set. Finally, all true positives (TP) and false positives (FP) were summed up under different likelihood score cutoffs to evaluate the ratio of true positives to false positives. The positive predictive values (PPV=TP/(TP+FP)) were calculated as the fraction of true positives to the total number of true positive and false positive (**Fig. 2d**).

The modular analysis and the weighted neighbor counting for network-based gene function prediction. We searched the local modules in the network using the Markov Cluster (MCL) algorithm, which is a fast and scalable unsupervised graph clustering algorithm[52]. To define the parameter of granularity, we followed a previously published method[53] by optimizing the functional coherence and size of the clusters[54]. The networks and subnetworks were designed and visualized using Cytoscape 2.5 (ref. 55).

The neighbor-counting method weighted by the likelihood score was used for the functional gene predictions in which the likelihood score of each linkage could represent the functional similarity between two proteins:

$$f(i,j) = \sum LS(m)\delta(j) / \sum LS(m)$$

where the f(i,j) is the probability of gene i having function j. The LS(m) is the likelihood score of the m^{th} neighbor of gene i. $\delta(j) = 1$ if the gene has function j, else $\delta(j) = 0$. Without threshold, we assigned an unannotated protein with *k* functions having the top *k* statistic scores. The performance of the predictions were evaluated by plotting precision against recall over various thresholds as described[56]. For a given threshold, precision and recall are defined as:

$$\text{Precision} = \sum_i^V k_{i,\beta} / \sum m_{i,\beta} \qquad \text{Recall} = \sum_i^V k_{i,\beta} / \sum n_i$$

where n_i is the number of known functions of protein i; $m_{i,\beta}$ is the number of functions predicted for protein i at threshold β and $k_{i,\beta}$ is the number of functions predicted correctly for protein i . V is the set of all functionally known genes.

DNA constructs, transfection and intracellular localizations. PCR amplification for the GFP constructs was carried out using cDNA with the gene-specific primers summarized in **Supplementary Table 8**. PCR products were digested with KpnI and AvrII and ligated into the transfection vector $pARL_{ama\text{-}1}$-GFP[34]. To avoid cytotoxic effects due to overexpression of the putative proteases, only 1 kb N-terminal fragments of PF08_0108 and PFD0230c were cloned. To ensure late expression, the promoter of the *ama-1* gene was used to drive transcription. *P. falciparum* asexual stages (3D7) were transfected as described previously[57]. Positive selection for transfectants was achieved using 10 nM WR99210.

The western blot analyses were carried out as previously described[58] using the mouse anti-GFP (1:1000, Roche) and sheep anti-mouse IgG horseradish peroxidase (1:3000, Roche). Images of unfixed GFP-expressing parasites were captured using a Zeiss Axioskop 2plus microscope with a Hamamatsu Digital camera (ORCA C4742-95) using Zeiss axiovision software. Immunofluorescence microscopy was performed on 4% formaldehyde/0.0075% glutaraldehyde-fixed parasites incubated for 1 h with primary antibodies in the following dilutions: rabbit anti-MSP-1 (1:2,000), rabbit anti-GAP45 (1:2,000) and rabbit anti-EBA-175 (1:2,000). Subsequently, cells were incubated with Alexa-Fluor 594 goat anti-rabbit IgG or Alexa-Fluor 488 goat anti-mouse IgG antibodies (1:2,000, Molecular Probes) and with DAPI at 1 μg/ml (Roche).

© 2010 Nature America, Inc. All rights reserved.

49. Hu, G., Llinas, M., Li, J., Preiser, P.R. & Bozdech, Z. Selection of long oligonucleotides for gene expression microarrays using weighted rank-sum strategy. *BMC Bioinformatics* **8**, 350 (2007).
50. Huttenhower, C., Hibbs, M., Myers, C. & Troyanskaya, O.G. A scalable method for integration and functional analysis of multiple microarray data sets. *Bioinformatics* **22**, 2890–2897 (2006).
51. Date, S.V. & Marcotte, E.M. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat. Biotechnol.* **21**, 1055–1062 (2003).
52. Krogan, N.J. *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637–643 (2006).
53. Wuchty, S. & Ipsaro, J.J. A draft of protein interactions in the malaria parasite *P. falciparum*. *J. Proteome Res.* **6**, 1461–1470 (2007).
54. Lee, I., Date, S.V., Adai, A.T. & Marcotte, E.M. A probabilistic functional network of yeast genes. *Science* **306**, 1555–1558 (2004).
55. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
56. Deng, M., Zhang, K., Mehta, S., Chen, T. & Sun, F. Prediction of protein function using protein-protein interaction data. *J. Comput. Biol.* **10**, 947–960 (2003).
57. Fidock, D.A. & Wellems, T.E. Transformation with human dihydrofolate reductase renders malaria parasites insensitive to WR99210 but does not affect the intrinsic activity of proguanil. *Proc. Natl. Acad. Sci. USA* **94**, 10931–10936 (1997).
58. Struck, N.S. *et al.* Re-defining the Golgi complex in *Plasmodium falciparum* using the novel Golgi marker PfGRASP. *J. Cell Sci.* **118**, 5603–5613 (2005).

© 2010 Nature America, Inc. All rights reserved.

doi:10.1038/nbt.1597

Executive pay goes up at private life sciences companies despite tumultuous economic climate

Bruce Rychlik & Evan Brown

Compensation to top executives at private life sciences companies continued along its upward trajectory, standing in stark contrast to flat pay levels at technology firms.

© 2010 Nature America, Inc. All rights reserved.

The 4th quarter of 2008 marked one of the worst three-month stretches for US equity markets since the Great Depression. Concurrent to the markets' year-end tumble, compensation committees across the life sciences industry were convening to determine 2009 pay levels for their executives. Surprisingly, the latest compensation study released by executive search firm J. Robert Scott and Ernst & Young, in collaboration with Noam Wasserman at Harvard Business School, rebranded this year as CompStudy (http://www.compstudy.com/), revealed that total target cash compensation at private life sciences companies for 2009 was up by nearly 4.5%. Contrast that to the largely flat pay at technology companies, including downward movement in a few key positions as well as bonuses as a percentage of base salary, and the result is even more surprising.

This year's report is the 8th edition of the annual compensation benchmark and contains results gleaned from the largest sample size to date—over 1,000 executives at more than 200 life sciences companies. The 2009 CompStudy provides position-by-position compensation data for top executives at privately held companies, information that is perennially difficult to come by for executives and investors.

Of great interest to participants and readers of the survey is that, for the first time, results are published in an online version (**Fig. 1**) in addition to an abridged print version. We did this as part of our continuing effort to make the data more accessible and more useful to our participants. This also allows J. Robert Scott to publish real-time updates and analyses as they become available.

Bruce Rychlik and Evan Brown are at J. Robert Scott, Boston, Massachusetts, USA.
e-mail: bruce.rychlik@fmr.com

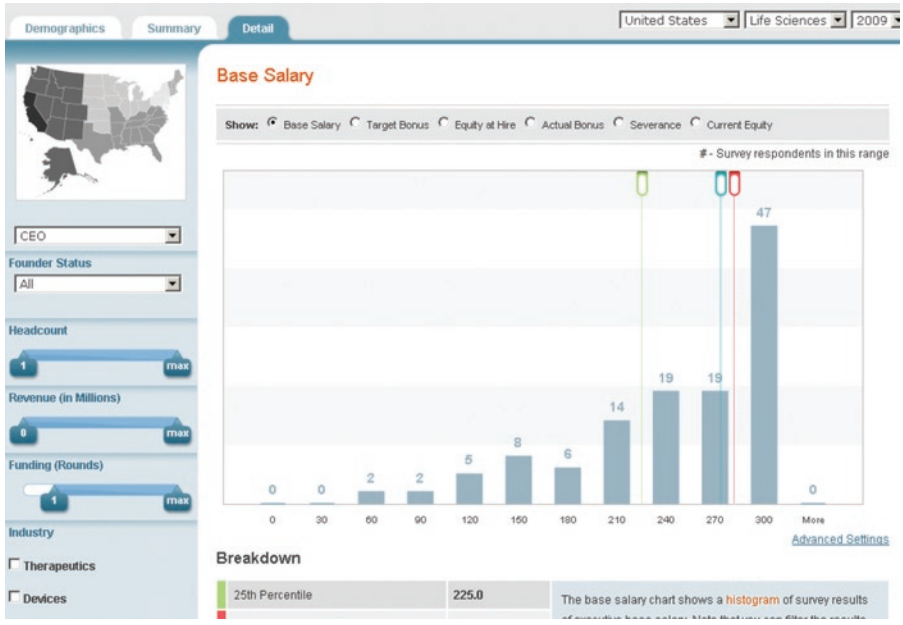


Figure 1 In the online version of the 2009 CompStudy, J. Robert Scott publishes real-time data updates and analyses as they become available.

The survey data were officially released December 3, 2009, through a webcast produced by Ernst & Young, with nearly 500 life sciences executives in attendance. When asked for their prediction during the 2008 webcast, 53% of the audience said they expected base salaries to remain flat or decrease in 2009. An additional 43% believed that compensation would increase less than 5%. The latter group of respondents proved correct, as average base salaries were up nearly 4% (**Fig. 2**). It is interesting to note, however, that over half the 2009 participants thought salaries would be flat or down, and not a single position surveyed in the CompStudy this year saw a year-over-year decrease in average base pay for life sciences executives.

In the 2008 Life Sciences Report, the average base salary rose 5.2% across all positions surveyed between 2007 and 2008. Clearly, while the 2009 figure was down slightly from 2008, the life sciences industry was not nearly as affected by the stagnating economy as was the technology sector. In the companion report in technology, one finds that base salaries for tech executives were up only 0.8%, a marked drop from the 5% annualized growth seen over the last decade.

Jonathan Fortescue, managing director at J. Robert Scott and a panelist on the life sciences webcast, stated, "The life sciences clearly march to the beat of a different drummer," and that this expressed confidence shown in

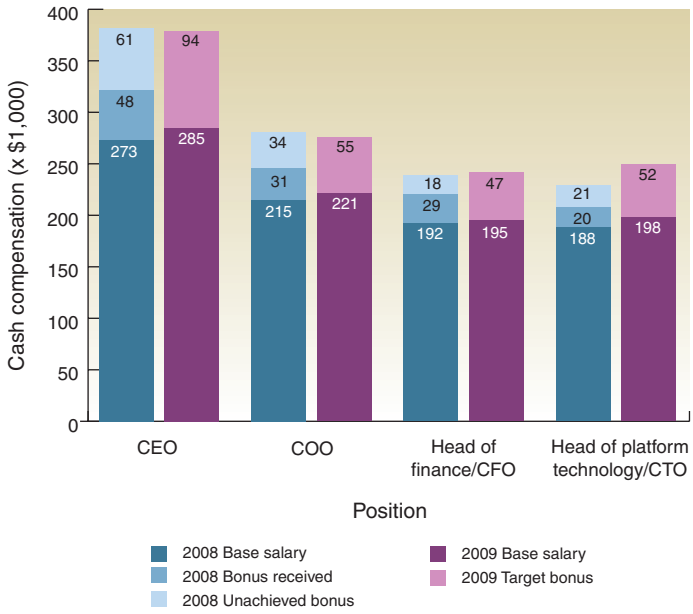


Figure 2 Cash compensation for selected positions at private life sciences companies.

the industry reaffirms it as a growth-driver for the US economy.

David Johnson, the service line leader for Ernst & Young's executive compensation advisory practice, echoed this sentiment, saying that private companies are more optimistic about their outlook than are public companies. This is likely helped even more by the longer business cycle seen in the life sciences, where success is not measured just in quarterly profit statements.

Bonuses provide another interesting take on the compensation outlook. Generally, target bonuses are set according to position and base salary, and the percentage of the target that an executive receives at year end is based on personal and company-wide performance.

Despite a slight decrease in the average target bonus for CEOs in 2009, bonus targets (as a percentage of base salary) generally held steady in 2009 (**Fig. 2**). Aaron Lapat, managing director at J. Robert Scott, commented that steady bonus targets and increasing base salaries are likely due to the nature of the sector. According to Lapat, "In terms of private life sciences companies, these executives may have experienced an increase in pay because their compensation tends to be based more on clinical milestones [than solely on financial performance]."

Of additional significant interest each year is determining what nonfounding CEOs receive in terms of equity. The survey indicates that as a company raises additional rounds of institutional funding, a nonfounding CEO can expect to be somewhat protected from dilution in terms of their holdings. Five percent of the fully diluted equity of the company seems to be the target, a number that has remained remarkably consistent over the last several years. Median equity holdings for the nonfounder CEO at the earliest stage companies is 4.9%; in those companies having raised four or more rounds, median equity is 4% of the fully diluted shares.

For founding CEOs, there is little equity protection; median equity drops from 10.5% for CEOs at the earliest stages of financing to 5% with four or more rounds raised. Compensation differences such as these for founding and nonfounding executives are of great interest to Professor Wasserman, who contributes to the design of the report each year. At his site (http://founderresearch.blogspot.com/), he addresses some of the current key issues surrounding private company founders and other "frustrations" that arise in creating and growing an enterprise.

The report also looks at organizational changes across the spectrum of financing rounds. One can see which positions are most frequently added as a company progresses with financing and builds out its management teams, and which positions company founders are most likely to hold. Interestingly, 'head of finance' is the position most commonly added post-founding; 'head of operations' and 'head of business development' are commonly added only after a company has raised multiple rounds of financing. Most of these additions represent the formalization of roles that the founders formerly held; the CEO is handing out some of his hats as company growth necessitates.

Not surprisingly, the most likely role for a founder to hold is CEO, although founder CEOs also get replaced at the fastest clip. In companies that have raised one or fewer rounds of institutional financing, 70% of CEOs are founders, whereas only 42% of CEOs at companies that have raised four or more rounds are also company founders.

Conclusions

In summary, in spite of a gloomy economic climate and decreased venture funding, executives at private life sciences companies saw modest pay raises and target bonuses consistent with past years' levels. In an industry where human capital with specific domain and functional experience is so highly prized, these compensation figures represent a vote of confidence in the sector, and an effort to ensure the very top talent remains motivated and engaged.

© 2010 Nature America, Inc. All rights reserved.

PEOPLE

Michael Moore (left) has been named nonexecutive chairman of Oxford BioTherapeutics (Oxford, UK). Moore brings several decades of senior management experience in the biotech sector and joins the board after five years as CEO of Piramed Ltd. Previously, he was CSO and research director of Xenova Group. In addition, Oxford BioTherapeutics recently announced the appointment of **Jim Cornett** as COO, **Mike Gresser** as CSO and **Jon Terrett** as vice president, oncology, to lead its US R&D operations.

Moore comments, "I am delighted to join Oxford BioTherapeutics at this exciting time in the company's development."

Xencor (Monrovia, CA, USA) has named **Bruce Carter** (left) chairman of the board. Carter joined ZymoGenetics in 1986 as vice president of R&D. After Novo Nordisk acquired the company, he was promoted to corporate executive vice president and CSO for Novo Nordisk. He led the negotiations that spun out ZymoGenetics as an independent company in 2000 and most recently served as ZymoGenetics' CEO. He continues to serve as chairman of the board of ZymoGenetics.

Mirna Therapeutics (Austin, TX, USA) has named **Chris Earl**, **Corey S. Goodman** and **Evan Melrose** as outside directors to its board. Earl was the first president and CEO of BIO Ventures for Global Health and also served as managing director of Perseus Capital and the Perseus-Soros BioPharmaceutical Fund. Goodman is currently an adjunct professor at the University of California, San Francisco, and most recently served as president of Pfizer's Biotherapeutics & Bioinnovation Center. Melrose is the founder of PTV Sciences and was formerly a director with Burrill & Company.

Hana Biosciences (S. San Francisco, CA, USA) has announced the appointment of **Howard Furst** to its board. Furst has over 20 years of experience in the healthcare industry and is currently a partner at Deerfield Management. Hana also announced the departure of **Arie Belldegrun** from the board after six years of service.

The BioIndustry Association (London) has announced the appointment of **Nigel Gaymond** as its new CEO. Gaymond began in sales and marketing at IBM in the UK and moved to the British Consulate-General in Boston, where he ran the commercial department in assisting the UK's biotech, healthcare and agriculture exports. Upon his departure, he established the consultancy Gaymond International.

Millipore (Billerica, MA, USA) has named **Robert S. Langer** to the company's board of directors, replacing **Daniel Bellus**, a Millipore director since 2000, who will retire from the board on March 8, 2010. Langer is the David H. Koch Institute Professor at the Massachusetts Institute of Technology. He serves on the board of MIT's McGovern Institute and the Whitehead Institute.

Affymetrix (Santa Clara, CA, USA) has named **Andrew J. Last** as chief commercial officer. He was previously vice president, global marketing and strategic planning of BD Biosciences' cell analysis unit and general manager of Pharmingen.

Pepscan Holding (Lelystad, The Netherlands) has appointed **Wim E.M. Mol** as CEO. He brings more than 20 years of experience in the pharma industry, most recently at Schering Plough where he was vice president responsible for the global scientific development and commercial strategy of a phase 3 project in an alliance with a subsidiary of Merck-Serono.

Gary Palmer, who most recently served as vice president of medical affairs at Genomic Health, has been appointed chief medical officer at On-Q-ity (Waltham, MA, USA), a diagnostics company developing products to improve cancer therapy effectiveness.

Infinity Pharmaceuticals (Cambridge, MA, USA) has announced that **Adelene Q. Perkins** will become the company's president and CEO and join the board of directors in accordance with an existing management succession plan. Founder, CEO and current chairman **Steven H. Holtzman** will continue full-time involvement with the company as executive chair of the board. Perkins joined Infinity in 2002 and currently serves as president and chief business officer.

Privately held NormOxys (Wellesley, MA, USA) has appointed **Martin Tolar** president and CEO. Tolar has held senior positions in pharma and biotech, most recently at CoMentis, where he served as CSO and later executive vice president and chief business officer. In addition, NormOxys announced the appointment of **David Clark** as its first chief medical officer. Clark previously served in Pfizer's clinical development division.

Diagnostics company Vermillion (Fremont, CA, USA) has named **William C. Wallen** to its board of directors. Wallen is CSO and senior vice president, R&D for IDEXX Laboratories.

Dennis Winger has been appointed to serve on the board of directors of Nektar Therapeutics (San Carlos, CA, USA). He is currently a director of Vertex Pharmaceuticals, Cephalon and Accuray, having retired in 2008 from Applera where he served as senior vice president and CFO.

Vertex Pharmaceuticals (Cambridge, MA, USA) has announced the appointment of **Nancy J. Wysenski** as executive vice president and chief commercial officer. Wysenski previously held the position of COO at Endo Pharmaceuticals and she was a co-founder, president and CEO of EMD Pharmaceuticals.

William D. Young, formerly chairman and CEO of Monogram Biosciences, has been named chairman of the board of Biogen Idec (Cambridge, MA, USA), succeeding **Bruce R. Ross**, who has retired from the board. Young previously served as Genentech's COO. Biogen Idec also announced that **Marijn E. Dekkers**, a director since May 2007, has stepped down from the board. Dekkers is the incoming CEO of Bayer HealthCare.

© 2010 Nature America, Inc. All rights reserved.
npg