nature biotechnology

THE SCIENCE AND BUSINESS OF BIOTECHNOLOGY

VOLUME 27 NUMBER 2 FEBRUARY 2009

www.nature.com/naturebiotechnology

Target capture comes a long way

High-response signature peptides

Getting to the hub of cancer prognosis

nature
biotechnology

www.nature.com/naturebiotechnology

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

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

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

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

DISPLAY ADVERTISING display@natureny.com (US/Canada)
display@nature.com (Europe)
nature@natureasia.com (Asia)
Global Head of Display Advertising Sales: John Michael, Tel: 44 207 843 4960, Fax: 44 207 843 4996
Global Account Development Manager: Graham Combe, Tel: 44 207 843 4914, Fax: 44 207 843 4749
Asia-Pacific Sales Manager: Ken Mikami, Tel: 81 3 3267 8765, Fax: 81 3 3267 8746
Display Account Managers:
New England: Sheila Reardon, Tel: (617) 399 4098, Fax: (617) 426 3717
New York/Mid-Atlantic/Southeast: Jim Breault, Tel: (212) 726 9334, Fax: (212) 696 9481
Midwest: Mike Rossi, Tel: (212) 726 9255, Fax: (212) 696 9481
West Coast South: George Lui, Tel: (415) 781 3804, Fax: (415) 781 3805
West Coast North: Bruce Shaver, Tel: (415) 781 6422, Fax: (415) 781 3805
Germany/Switzerland/Austria: Sabine Hugi-Fürst, Tel: 41 52761 3386, Fax: 41 52761 3419
United Kingdom/Ireland/France/Belgium/Eastern Europe: Jeremy Betts, Tel: 44 207 843 4968, Fax: 44 207 843 4749
Scandinavia/The Netherlands/Italy/Spain/Portugal/Israel/Iceland: Graham Combe, Tel: 44 207 843 4914, Fax: 44 207 843 4749
Greater China/Singapore: Gloria To, Tel: 852 2811 7191, Fax: 852 2811 0743

NATUREJOBS naturejobs@natureny.com (US/Canada)
naturejobs@nature.com (Europe)
nature@natureasia.com (Asia)
US Sales Manager: Peter Bless, Tel: (212) 726 9248, Fax: (212) 696 9482
European Sales Manager: Andrew Douglas, Tel: 44 207 843 4975, Fax: 44 207 843 4996
Asia-Pacific Sales Manager: Ayako Watanabe, Tel: 81 3 3267 8765, Fax: 81 3 3267 8746

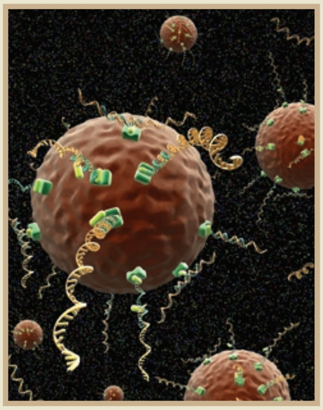
SITE LICENSE BUSINESS UNIT
Americas: Tel: (888) 331 6288 institutions@natureny.com
Asia/Pacific: Tel: 81 3 3267 8751 institutions@natureasia.com
Australia/New Zealand: Tel: 61 3 9825 1160 nature@macmillan.com.au
India: Tel: 91 124 2881054/55 npgindia@nature.com
ROW: Tel: 44 207 843 4759 institutions@nature.com

CUSTOMER SERVICE www.nature.com/help
Senior Global Customer Service Manager: Gerald Coppin
For all print and online assistance, please visit **www.nature.com/help**
Purchase subscriptions:
Americas: *Nature Biotechnology,* Subscription Dept., 342 Broadway, PMB 301, New York, NY 10013-3910. Tel: (866) 363 7860, Fax: (212) 689 9108
Europe/ROW: *Nature Biotechnology,* Subscription Dept., Macmillan Magazines Ltd., Brunel Road, Houndmills, Basingstoke RG21 6XS, United Kingdom. Tel: 44 1256 329 242, Fax: 44 1256 812 358
Japan: *Nature Biotechnology,* NPG Nature Asia-Pacific, Chiyoda Building, 2-37 Ichigayatamachi, Shinjuku-ku, Tokyo 162-0843. Tel: 81 3 3267 8751, Fax: 81 3 3267 8746
India: *Nature Biotechnology,* NPG India, 3A, 4th Floor, DLF Corporate Park, Gurgaon 122002, India. Tel: 91 124 2881054/55, Tel/Fax: 91 124 2881052

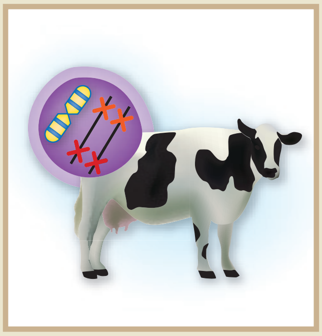
REPRINTS reprint@natureny.com
Nature Biotechnology, Reprint Department, Nature Publishing Group, 75 Varick Street, Fl 9, New York, NY 10013-1917, USA.
For commercial reprint orders of 600 or more, please contact:
UK Reprints: Tel: 44 1256 302 923, Fax: 44 1256 321 531
US Reprints: Tel: (212) 726 9278, Fax: (212) 679 0843

VOLUME 27 NUMBER 2 FEBRUARY 2009

nature biotechnology



Artist's rendering of streptavidin-coated magnetic beads used to pull down ultra-long biotinylated RNA 'baits' designed to capture specific genomic DNA fragments. Gnirke *et al*. use the approach for targeted Illumina sequencing, represented by the processed image of a massively parallel sequencing experiment (p 182). Credit: Ken Eward ©BioGrafx.



Human polyclonals from cattle, p 146

EDITORIAL

NEWS

BIOENTREPRENEUR

BUILDING A BUSINESS

OPINION AND COMMENT

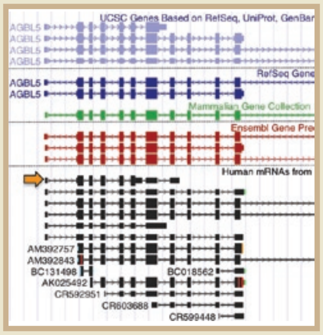
CORRESPONDENCE

BPA WORLDWIDE

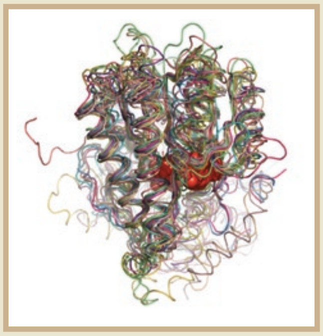
npg
nature publishing group

Nature Biotechnology (ISSN 1087-0156) is published monthly by Nature Publishing Group, a trading name of Nature America Inc. located at 75 Varick Street, Fl 9, New York, NY 10013-1917. Periodicals postage paid at New York, NY and additional mailing post offices. **Editorial Office:** 75 Varick Street, Fl 9, New York, NY 10013-1917. Tel: (212) 726 9335, Fax: (212) 696 9753. **Annual subscription rates:** USA/Canada: US$250 (personal), US$3,060 (institution), US$4,050 (corporate institution). Canada add 7% GST #104911595RT001; Euro-zone: €202 (personal), €2,430 (institution), €3,488 (corporate institution); Rest of world (excluding China, Japan, Korea): £130 (personal), £1,570 (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 © 2009 Nature Publishing Group. Printed in USA.

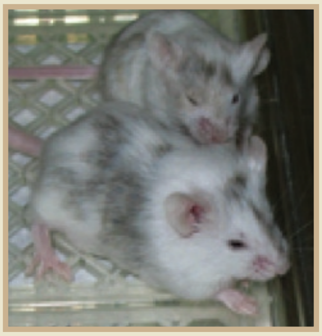
VOLUME 27 NUMBER 2 FEBRUARY 2009



Getting more from genome browsers, p 153



Understanding protein promiscuity, p 157



Cells for reprogramming screen, p 169

COMMENTARY

BOOK REVIEW

FEATURE

PATENTS

NEWS AND VIEWS

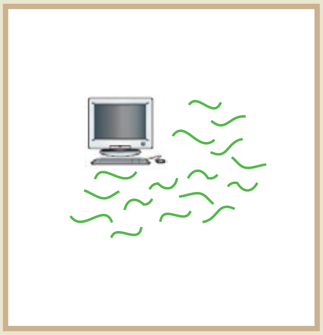
COMPUTATIONAL BIOLOGY

PRIMER

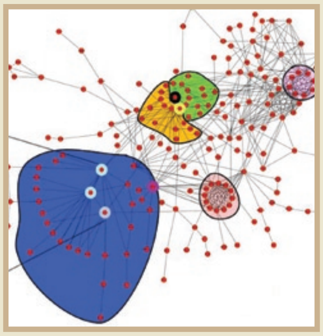
RESEARCH

REVIEW

VOLUME 27 NUMBER 2 FEBRUARY 2009



High-response signature peptides, p 190



Breast cancer hubs, p 199

BRIEF COMMUNICATIONS

ARTICLES

LETTERS

NATUREJOBS BIOTECHNOLOGY

CAREERS AND RECRUITMENT

NATURE BIOTECHNOLOGY CLASSIFIED

See back pages.

André Stevens

Farming out human-antibody production

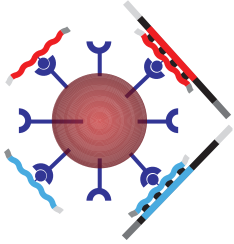
The dependence on human donors for supplying antigen-specific human polyclonal antibodies is a major constraint on their widespread clinical use. But prospects for using large domesticated animals that express human immunoglobulin genes have been stymied by the dominant expression of the endogenous immunoglobulin genes and limited understanding of immunoglobulin gene function and organization in ruminants. Taking a key step in a complex genetic engineering program to produce high yields of polyclonal antibodies in cattle, Kuroiwa *et al.* test whether human immunoglobulin genes can support bovine B-cell development and robust humoral immunity in the absence of normal bovine immunoglobulin-gene expression. They first show that, unlike mice and humans, cattle possess two independent pathways for B-cell development, each comprising a different functional IgM locus. Only the homozygous inactivation of both IgM loci confers B-cell deficiency. The authors then introduce a human artificial chromosome carrying the entire unrearranged human immunoglobulin heavy and κ-light-chain loci into a double IgM-knockout background. A calf produced after multiple rounds of cloning generates a high proportion of antigen-specific immunoglobulin after hyperimmunization with anthrax protective antigen. Approximately 80% of plasma immunoglobulins from the animal are chimeric (carrying a bovine κ- or λ-light chain). Although knocking out bovine light-chain loci would likely boost yields of fully human product further, the feasibility of producing relatively high levels of polyclonal antibodies (>500 μg/ml) in cattle further clears the way for optimizing this system. Antisera from hyperimmunized cattle would not only be enriched for antibodies of specific therapeutic value but could also pose less risk of viral contamination than antisera pooled from thousands of human donors. **[Articles, p. 173]** *PH*

Predicting strong-signal peptides

The potential of targeted mass spectrometry to monitor changes in biomarker levels in response to disease progression and therapy is appreciated increasingly. Ideally, levels of a protein of interest should be monitored using a peptide that uniquely identifies it. But, as only a fraction of all peptides present in a complex biofluid are detected by the most advanced instruments available for discovery experiments, the best candidates for signature peptides are often not even observed during empirical investigations. Nontargeted mass spectrometry–based technologies are also unable to easily translate insights from nonproteomic technologies into sensitive, focused assays of the presence or abundance of a signature, or proteotypic, peptide. Fusaro *et al.* build on previous algorithms for *de novo* prediction of signature peptides from *in silico* digests by using the Random Forest algorithm to predict peptides that not only are proteotypic but also produce the highest ion-current response for that protein. They demonstrate the accuracy of their predictor with ten validation sets—including three with data obtained from plasma—that involve different experimental conditions, database-search algorithms, quantification methods and sample complexities. The model correctly predicts 12 of 18 validated peptides that can reliably assay the abundance of six proteins for which no data are available in a comprehensive database. [**Articles, p. 190**] *PH*

Written by Kathy Aschheim, Laura DeFrancesco, Michael Francisco, Peter Hare, Brady Huggett, Craig Mak, Andrew Marshall & Lisa Melton

Target capture for the long haul



Several methods to more efficiently use sequencing resources capture large targeted regions of the genome in cases where simple PCR is insufficient. Gnirke *et al.* describe a new method based on hybridization-capture to a heterogeneous pool of ultra-long, biotinylated RNA 'baits' in solution. The RNA probes are *in vitro* transcribed from PCR-amplified 200-base-pair oligos synthesized on a microarray, hybridized in vast excess to genomic DNA in solution and then pulled down using streptavidin-coated magnetic beads. This method addresses limits of the efficiency and reproducibility of previously described methods that exploit array-capture or multiplex amplification. The authors demonstrate targeted capture and sequencing of 2.5 Mb of protein-coding exons from 1,900 human genes as well as the targeted sequencing of four gene-containing regions, each spanning 0.22 to 0.75 Mb of the genome. [**Articles, p. 182**] *CM*

iPS cell screening tool

Reprogramming of somatic cells into induced pluripotent stem (iPS) cells relies on a small number of transgenes, such as *Oct4*, *Sox2*, *Klf4* and *c-Myc*. Some of these are known oncogenes, however, and random integration of any transgene into the genome carries a risk of insertional mutagenesis. These concerns would be allayed if the reprogramming genes could be replaced by small molecules that have no detrimental effects of their own. Jaenisch and colleagues have generated mouse cells that will facilitate screening for such small molecules. Starting with an iPS-cell chimeric mouse carrying multiple copies of doxycycline-inducible *Oct4*, *Sox2*, *Klf4* and *c-Myc*, they segregate the transgenes through two rounds of breeding to produce mice bearing single copies in all possible combinations. Cell lines containing all four factors yield iPS cells after treatment with doxycycline, but three-factor lines do so only if the fourth factor is added, providing a system to screen for small molecules that can substitute for the missing factor. [**Brief Communications, p. 169**] *KA*

The hub of breast cancer prediction

Tissue-specific gene signatures have long been suggested as predictive of disease outcome, but now Taylor *et al.* move up a level, linking disruptions in the dynamic modular organization of the human protein-protein interaction network with breast cancer outcome. Previous work in yeast identified two classes of highly connected proteins, called 'date' and 'party' hubs; party hubs were proteins expressed with their interacting partners in at least one of five environmental conditions; date hubs were selectively expressed with only a few partners in any given condition. Taylor *et al.* find that similar patterns of co-expression hold true for proteins in the human interactome when their expression across 79 tissues is examined. They identify 'intramodular' hubs with party-like cross-tissue patterns of co-expression. These contrast with 'intermodular' hubs with date-like co-expression in few tissues. Moreover, the co-expression of some hubs with their partners appears to be disrupted in breast cancer cells—party-like hubs become date-like and vice versa. Taylor *et al.* go on to exploit this as a means of improving predictions of breast cancer survival based on gene expression profiles taken from tumor cells. [**Letters, p. 199**] *CM*

Patent Roundup

Two insiders give advice on dealing with an 'obviousness' rejection from the patent office. [**Building a Business, p. 117**] *BH*

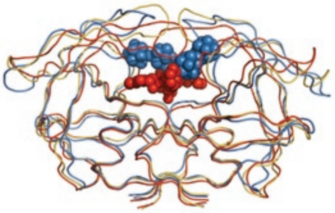
The European Patent Office has invoked a 'morality' clause to refuse to issue patents for stem cells involving the destruction of human embryos. [**News, p. 103**] *LM*

A recent decision by the US Board of Patent Appeals and Interferences on Kubin and Goodwin's patent application is unlikely to threaten all DNA sequence patents. [**Correspondence, p. 120**] *AM*

Rai and colleagues systematically create a patent landscape covering the engineering and use of zinc-finger proteins and ask whether one company's near monopoly of intellectual property rights ultimately helps or hinders development of the platform. [**Patent Article, p. 140**] *MF*

Recent patent applications in gene expression. [**New patents, p. 145**] *MF*

Protein promiscuity



Protein promiscuity—the ability of proteins to perform different functions or to interact with different partners—has been recognized for a long time. However, only recently have researchers come to appreciate how ubiquitous and wide-ranging it is. Whereas promiscuity can be problematic, when, for example, a drug interacts with proteins that are not its intended target, it can also be harnessed for good, as to encourage a range of interactions or develop new or stronger ones for pharmaceutical and industrial applications. Nobeli and colleagues look at the levels of promiscuous behaviors proteins engage in, the conditions that drive promiscuity and the mechanisms underlying it. Finally, they discuss how protein engineers and drug developers alike can exploit this feature of proteins to design novel proteins with applications in research and industry. [**Review, p. 157**] *LD*

Next month in nature biotechnology

- Silencing via splicing
- Liver-specific mutagenesis for cancer-gene discovery
- Ligand-dependent exponential RNA amplification
- Investigating human embryonic stem cell teratomas

nature biotechnology

The worst of times, the best of times

Big pharma should be more proactively investing in cash-hungry public biotech companies.

These days, cash is king in the biotech sector. There are the 'withs': most of the top 20 pharmaceutical companies and a few larger biotechs currently sitting on capital reserves of several billion dollars. And there are the 'withouts': an alarmingly large number of public biotech firms with less than a year's cash that are currently gasping for the financial oxygen locked away in deep-frozen tundra of the equity markets. As a result, many biotechs with promising products now face financial oblivion. If the pharmaceutical industry is really serious about fostering a diverse universe of external product opportunities, then it should rethink how it uses some of its cash reserves to invest directly in public biotech firms that currently languish at bargain valuations. In the long term, the result could prove a win-win situation for pharma and biotech alike.

The publicly quoted biotech company sector now finds itself in an extremely precarious situation. According to the Biotechnology Industry Organization, around 38% of 370 small biotech companies have less than one year's worth of cash. Nearly 100 publicly listed biotechs are operating on less than six months' cash.

Refinancing options are dwindling. The public markets remain shut and convertible debt is increasingly hard to come by. 2008 was the worst fundraising year in the past nine years. Biotech companies raised only $5.7 billion from public equity, a 58% decline from the previous year. And there is little expectation that 2009 will be any better. Compared with the average raised each year by the biotech sector during the previous five years ($9.8 billion), public companies are looking at a shortfall in funding for 2008 of billions of dollars. There is virtually no chance that the deficiency can be made up in increases in other financing sources. The valuations attributed to research collaborations or licensing deals will also be driven down by the general financial slump as witnessed by the fall in the amount of money raised by the biotech sector through partnering in 2008 from $22.4 to $20.0 billion.

One could argue that devalued biotech companies ought to make attractive acquisition targets. Indeed, in 2008 there was a record number of biotech-pharma mergers and acquisitions (M&As): 31 compared with 19, 24 and 23 M&As in 2007, 2006 and 2005, respectively. However, no one is expecting a dramatic upswing in the capacity of big pharma and biotech to complete more deals or acquisitions in the next 12 months. Although M&As will mop up a few assets, the bureaucracy associated with such transactions will mitigate against this being a wholesale solution. It's not a matter of cash reserves—Pfizer has about $17 billion in cash reserves and Bristol-Myers Squibb, $7.2 billion, for example. The problem is there are only a few biotech companies that big pharma can 'plug and play' into its existing internal strategic R&D priorities. And Pfizer at least appears set on bulking up with Wyeth rather than investing in biotech small fry.

This brings up another option—private investments in public equity (PIPEs), an off-market transaction in which a company issues new stock to an investment group, usually at a substantial discount to prevailing market rates. Since the end of 2008, many venture capital (VC) funds have been rewriting their investment criteria to allow them to take advantage of distressed public biotech markets. The problem is that the venture capitalists have already been doing this for some time. The well of VC funds for PIPEs and public equity trades is not limitless and investors do not have sufficient management resources to handle the transactions.

Could the pharmaceutical industry or big biotech step in to fill the gap? Perhaps. In December, Novo A/S, which already makes VC investments and holds a controlling interest in Danish biotech giant Novo Nordisk, announced that it had established a $500 million 'growth equity' fund to do just this type of investment.

The question is will other pharma or big biotech companies follow suit? And if they do get involved, the innate risk-aversion and conservatism of pharma management makes it likely that investments will be selective—for companies with late-stage clinical products. All the other public biotechs—no matter how promising and innovative their phase 1 or phase 2 programs—will be left out in the cold.

In the near term then, the plights of pharma and biotech in the present financial environment could not be more different. Cash-rich pharma can simply hunker down and look forward to occasionally cherry-picking from a smorgasbord of devalued biotech assets for projects to suit its needs. In contrast, public biotech companies face a bleak year ahead. At best, many companies will have to radically restructure, shelve all but one or two key programs or consolidate with another of biotech's walking wounded. At worse, a substantial number of firms will just cease to exist in 2009.

Of course, the pharmaceutical industry is not a charity aimed at saving small biology-based companies. If pharma acts in any sphere, it has to be based on self-interest. But this is not a time for business as usual. With the productivity of internal R&D programs at pharmaceutical companies continuing to plummet, drug pipelines increasingly empty and a raft of patent expirations on the horizon, a whole swathe of innovative biotech companies simply disappearing will have significant long-term repercussions. The loss of these companies and their products might not be a problem now. But if the current crop of small-cap biotechs with early-phase products does turn into the lost generation, pharma may be looking at a massive new lacuna in its pipelines five years from now.

Pharma needs to pay attention to the plight of current biotech firms and do more to support its drug discovery engine. The time has come for big pharma to ask not what biotech can do for it, but what it can do for biotech.

Venture capital shifts strategies, startups suffer

Even before Lehman Brothers' collapse triggered the global banking crisis at the end of last September, venture capital (VC) firms were finding it harder to raise new funds. In the nine months from January to September 2008, US VC firms raised only $19.7 billion from their limited partners compared with $32 billion for the entire previous year and $30 billion in 2006 (Dow Jones Venture figures). About 25–30% of this money typically goes into the healthcare sector.

The shortfall did not, however, discourage VCs in the US from deploying their funds (**Table 1**). Overall, investment this side of the Atlantic in the first three quarters of 2008 was on par with the same period of 2007, standing at $22.3 billion from just under 2,000 deals. And for US healthcare in the third quarter alone, VC investment held steady at $2.2 billion, with over half of that going to biopharmaceuticals (**Fig. 1**).

However, the crisis has changed the structure of VC investing, as firms shift more cash into their existing portfolios at the expense of startups. "Not only is it taking longer for these firms to get an exit," says Jonathan McQuitty of VC firm Abingworth, based in London and Boston, "but also there are fewer potential co-investors than [there were] previously... That means we have to have some pretty hefty reserves."

Moreover, the crisis has hit some VC firms' resources far worse than others—not through bad management, but by unfortunate timing, says McQuitty. Several VC firms had not yet begun raising new funds when disaster struck in mid-2008, and now they cannot do so, he says. "Smaller or less experienced VC firms are going from a situation where fund-raising was merely a bumpy ride to where it simply doesn't happen," he says. "Several funds have missed the wave and are now heading for a wipeout, possibly taking them out of the seed financing game for months or even years." Even top-notch healthcare VCs like New Enterprise Associates (NEA) in Chevy Chase, Maryland, and Abingworth itself have had trouble meeting their fund-raising targets, he admits.

There were four major medical technology financing events in the third quarter of 2008.

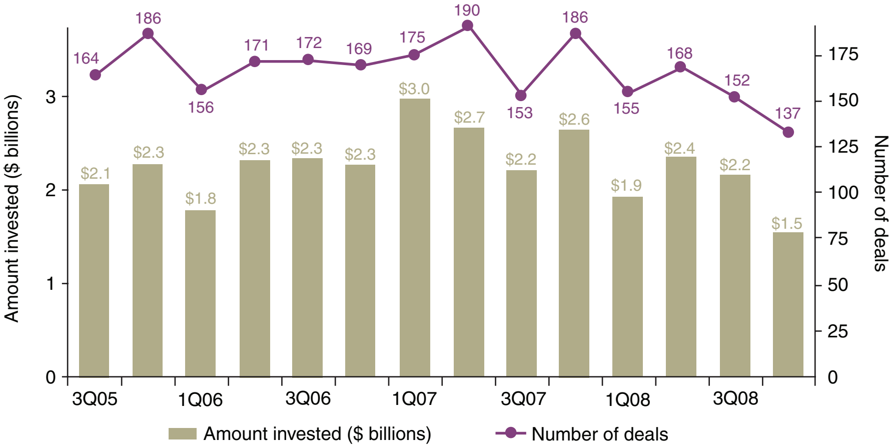


Figure 1 Venture capital investment into US healthcare companies by quarter over the past three years (adapted from Dow Jones Venture presentation).

Pacific Biosciences of Menlo Park, California, raised $100 million to fund development of its single-molecule real-time DNA sequencing platform. Proteolix in South San Francisco obtained $79 million to underwrite phase 2 trials of a drug aimed at protein degradation pathways in cancer and autoimmune diseases. Portola Pharmaceuticals, also in South San Francisco, raised $60 million to pay for clinical trials of therapeutics for vascular disease. And the device firm CVRx, based in Minneapolis, restocked its cash reserves with $83 million for trials of its hypertension therapy—not strictly biotech, but the event attracted many of the VCs most active in biotech, such as NEA and Frazier Healthcare Ventures in Menlo Park, which also co-invested in the Portola financing. Another leader, Advanced Technology Ventures, participated in both the Proteolix and Portola financings. Nomura Phase4 Ventures in London also co-invested in the Proteolix deal but otherwise kept a low profile in the second half of 2008—as did NEA.

Other leading biotech VCs active in the quarter were Intel Capital, which co-led the Pacific financing, and Kleiner Perkins Caufield & Byers, which also backed the Pacific deal. Virtually all the Pacific VC investors were already deeply committed to the company—a feature of recent VC financings, where investors are being very careful to see their established portfolio through clinical trials at the expense of neglecting seed financing. On the other hand, the top three third-quarter deals were later-stage fundings (*Nat. Biotechnol.* **26**, 1212, 2008).

Only a few VC firms are still regularly investing in seed or early-stage companies. Post-Lehman, in fact, VCs have begun to reexamine the very ground they are standing on. "In the past six months the pace of VC investing in biotech has decreased as people stand back and reassess fundamental business models," says Jamie Topper, general partner at Frazier. The firm is one of the top five US biotech VCs, with half of its biotech investing previously going into early-stage (seed or series A) financing. But now the firm is rethinking that strategy—at least in the short term—until it sees just how deep and dark the crisis is going to be. "Every dedicated healthcare VC is questioning the desirability of funding early-stage biotechs," says Topper.

While total US healthcare VC investment dropped 25–50% in the final quarter of 2008, Topper estimates, early-stage biotech funding plummeted about 50–75%. He expects that trend to continue for the first half of 2009. "Our current fund will probably place 5–10% at most in early-stage biotech, compared to the usual 25%," he says. Biotech deals in 2009 are likely to be later stage, possibly even private investments

IN brief

Merck joins the biotech game

AP Photo/Keystone, Laurent Gillieron

Merck's CEO Richard Clark introduces a new strategy for biogenerics.

Merck's CEO Richard Clark has unveiled plans to enter the biotech drug market by creating Merck BioVentures (MBV), a global division focused on developing biotech drugs, in particular copycat versions of existing biologics. The initiative represents Merck's shot at replenishing a dwindling pipeline and an attempt to position itself as a major competitor in the biotech field. The unit is expected to burn $1.5 billion over the next seven years, with a manufacturing capacity fully operational by 2012. The news comes at a time when the Whitehouse Station, New Jersey–based pharma faces dwindling sales of cholesterol-lowering blockbusters Zetia (ezetimibe) and Vytorin (ezetimibe and simvastatin) and the expiration of some key patents. Merck's new biotech division will take advantage of its GlycoFi technology, purchased in 2006. This glyco-engineering platform—a faster, less expensive production method than mammalian-based culture—will enable the company to circumvent generic manufacturing restrictions and be competitive in its pricing approach. MBV already has a candidate drug in phase 1, MK2578 (pegylated erythropoietin), designed to compete with Thousand Oaks, California–based Amgen's Aranesp (darbepoetin alfa), and at least five other products projected to be in late-stage development by 2012. "It was important to make a decision around manufacturing and leverage our internal capabilities," says Frank Clyburn, MBV general manager. Biogenerics represent an important market opportunity as $10 billion worth of biologic drugs are expected to come off patent by 2010, with an additional $10 billion by 2015. Given that the new Democratic administration is expected to push biogenerics legislation through Congress, the timing is propitious, although a generic-drug-style abbreviated pathway looks increasingly unlikely. As clinical trial costs will, mostly likely, be added to the cost of developing a follow-on biologics environment, the investment and expertise needed for success could be considerable. But considering the large number of leading biologics, such as Epogen (epoetin alfa), Enbrel (etanercept) and Avastin (bevacizumab), facing patent expirations through 2017, and the diminished late-stage risk involved in producing follow-on biologics, Merck's strategy is timely. Basel-based Novartis and Petach Tikva, Israel–based Teva already market follow-on biologics in Europe and India, and several other companies also have the cash and the technology to enter the race. "Over the longer term, we will also apply our unique humanized GlycoFi yeast technology platform to the development of novel biologics," says Caroline Lappetito, Merck's director of global communications. *–Victor Bethencourt*

Table 1 Top ten biggest rounds for private biotech firms in 2008.

Company	Amount invested ($ millions)	Round number	Date closed
OncoMed Pharmaceuticals	169	2	12 December
Portola	130	3	9 July
Pacific Biosciences	100	5	14 July
Radius Health	82.5	3	20 November
Ganymed Pharmaceuticals	82.2	4	18 November
Proteolix	79	3	8 September
ESBATech	62.5	2	7 August
Merrimack Pharmaceuticals	60	6	10 June
Biolex Therapeutics	60	4	6 October
Intrexon	55	3	7 May

Source: BCIQ: BioCentury Online Intelligence

in public equities (PIPEs). "Some of the later-stage public market biotech and medical device companies have taken big hits on valuations, making them very attractive investments for us," he says. Moreover, Frazier is planning to shift its investment away from the biotech sector to growth equity, funding companies that are already profitable but need more capital to expand, often in the pharma or healthcare services sector. This naturally delivers lower multiples on exit, but it mitigates development risk, which is the top priority right now.

Abingworth's McQuitty agrees. "Several larger funds have even gone to a PIPE-only strategy; they feel they want to take time out from investing in private companies at all." Abingworth is, he assures, still willing to do early-stage investing, though possibly less so than before.

Figures from Dow Jones Venture confirm that funding has indeed shifted away from seed financing toward the later-stage companies, falling from 23% in the first quarter of 2008 to 18% in the third quarter (in all VC sectors, not just healthcare). This trend, however, was already apparent before 2008.

VC funds associated with pharmaceutical companies (e.g., GSK Ventures or MedImmune Ventures) are bucking this trend and are still actively interested in funding early technology ventures. "They have not just investment focus but also strategic focus; they want to access an innovation," says Topper. "That's good for the industry, because for other VCs [like Frazier] it is very hard to make a rational argument for taking a chance on an early startup when you are not sure just how they are going to finance themselves later." An example of this 'strategic financing' is the backing of CVRx by J&J Development in New Brunswick, New Jersey, owned by Johnson & Johnson (J&J), also in New Brunswick. J&J led the $83.9 million financing, in cooperation with NEA and several other top-ranking healthcare VC investors, including Frazier and InterWest Partners in Menlo Park and Houston.

Another trend inspired by the credit drought is stronger interest in co-investing and pre-syndication. "We are aware that later funding rounds may not be easy, so there is a move to pre-syndicating either the whole funding process or a big chunk, say 80%, right at the start," says Abingworth's McQuitty. "The company directors don't then have to be out pounding the pavement looking for their next rounds."

Jens Eckstein, partner at TVM Capital in Boston, points to more subtle shifts in the structure of seed financing. "The trend is for incubating rather than 'official' seed funding—a stealth mode with tight control on spending," he says. This has coincided with larger first rounds combining series A and B, funded by syndicates strong enough to advance the biotech companies enough money to keep going longer, without having to look for further financing. The result has been some series A fundings worth as much as $30 million.

Eckstein believes European biotech has been hit significantly harder than the US, with very few big VCs still active. Besides Sofinnova in Paris, Abingworth and TVM itself, several firms have dropped out of the sector entirely. Dow Jones Venture numbers back this up: VC investing into European healthcare companies flopped from €468 million in 61 deals in the first quarter to €164 million in 32 deals in the second quarter. Significantly, Germany has now taken the lead in European VC investing, as UK activity fell off a cliff in 2008.

But all is not lost, insists Eckstein. "A number of VC firms have kept their powder dry, with money in hand still to invest, having only recently closed their latest funds." He also notes the merger and acquisition environment remains strong, with cash-rich pharma in buying mode by necessity as they face pipeline issues. "There has been no real slowdown in deal flow," he says. "But the pressure is on startups to think through their plans more critically."

Peter Mitchell *London*

Biologic approvals in 2008

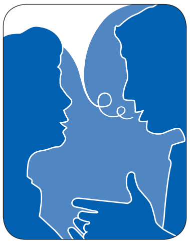
Amgen (Thousand Oaks, California), Regeneron (Tarrytown, New York), ViroPharma (Exton, Pennsylvania) and ZymoGenetics (Seattle, Washington) all received US Food and Drug Administration (FDA) approvals last year for novel types of biologics (**Table 1**). For a full list of FDA approvals of drugs from public biotech companies, see **Supplementary Table 1** online.

Table 1 Selected biologic approvals from public biotech companies in 2008[a]

Company/partner	Product (generic)	Indication
Amgen	Nplate (romiplostim, a 60-kDa peptide with a thrombopoietin receptor (Mpl)-binding domain	Thrombocytopenia in adults with idiopathic thrombocytopenic purpura
Biogen Idec (Boston)/ Elan (Dublin)	Tysabri (natalizumab)	Moderately to severely active Crohn's disease
Cangene (Winnipeg, Manitoba, Canada)	Accretropin (somatropin)	Growth failure in children with growth hormone deficiency and short stature associated with Turner's syndrome
Genentech (S. San Francisco, California)/Roche (Basel)	Avastin (bevacizumab)	Metastatic breast cancer
Regeneron	Arcalyst (rilonacept, single-chain fusion of the extracellular binding domains of interleukin (IL)-1 receptor I and IL-1 receptor accessory protein coupled to Fc portion of a human IgG)	CIAS1-associated periodic syndrome
ViroPharma	Cinryze (serum-derived complement factor C1-esterase inhibitor)	Prevent angioedema attacks in individuals with hereditary angioedema
ZymoGenetics (Seattle)/ Bayer (Leverkusen, Germany)	Recothrom (recombinant thrombin)	General aid to achieving hemostasis during surgery

[a] As defined by *Nat. Biotechnol.* **26**, 753–762 (2008).

IN their words



"If people don't want to buy bonds in General Electric, what's going to make them want to invest in an early-stage biotech company?"

Randy Scott, chairman of Genomic Health, laments the bleak situation that biotechs face in raising capital in the current financial climate. (*The Wall Street Journal*, January 11, 2009)

"Whatever it takes to make friends and influence people—whether it's building a school or handing out Viagra."

A CIA operative on how the agency occasionally wins over Afghanistan warlords by offering Pfizer's (New York) impotence drug. (*Washington Post,* December 26, 2008)

"Practically anything you can put a name onis branded in a doctor's office, short of branding, like a Nascar driver, on the doctor's white coat."

Physician Robert Goodman decries pharma's marketing efforts to brand almost everything in the doctor's office. (*The New York Times,* December 30, 2008)

"It's practically a paradise for conducting clinical trials."

A spokesperson for Quintiles, the world's largest contract research organization, on the company's success in enrolling 204 infants for a vaccine study in just three days in India. (*Pharmalot*, December 18, 2008)

"The senator's worried that something's ghostwritten. I mean, give me a break."

Lila Nachtigall, a New York University professor and director of its Women's Wellness Center, on Senator Charles Grassley's enquiry into the role of Wyeth in the writing of a journal article she authored extolling hormone treatment. (*The New York Times*, December 12, 2008)

"2009 will be a year of anticipation for the venture capital industry as the economic turmoil will engender a fair amount of Darwinian change."

National Venture Capital Association President Mark Heesen celebrates the bicentenary of Darwin's birth by suggesting it will be survival of the fittest in the biotech sector. (NVCA press release, December 18, 2008)

Osiris seals billion-dollar deal with Genzyme for cell therapy

Adult stem cell products could be lining up to penetrate the inflammation and orthopedics markets, judging by the recent $1.38 billion paid by Genzyme for Osiris's mesenchymal stem cell technology. Genzyme's bid to develop and commercialize two of Osiris Therapeutics' mesenchymal stem cell (MSC) products, announced in November, was hardly surprising. With more than 10 years of experience in developing stem cells in the clinic, Columbia, Maryland–based Osiris Therapeutics has become a frontrunner among companies commercializing adult stem cells. The company's two stem cell therapies, Prochymal and Chondrogen, are well advanced into clinical studies for graft-versus-host disease (GvHD), Crohn's disease and knee cartilage regeneration. But one issue continues to puzzle researchers: no one is quite sure how these cell products work.

The deal with Genzyme places Osiris on a firm financial footing. As well as milestone payments, the $130 million up-front payment comes in addition to cash from a US Department of Defense contract for Prochymal use in acute radiation sickness. Although Genzyme has no equity stake in Osiris, which will keep all US and Canadian rights to its MSC products and indications, the Cambridge, Massachusetts–based company does get commercial rights to the MSC treatments in all other countries. Potential markets include osteoarthritis, and may ultimately expand to include cardiovascular disease, diabetes and chronic obstructive pulmonary disease.

Osiris was one of the first companies to work with adult MSCs and is "the leader in the space," according to Chris Mason, Professor of Regenerative Medicine Bioprocessing at University College London with no financial ties to the firm. "It is a pragmatic company, they have done a good job of manufacturing high-quality cells leading to robust clinical data," he adds.

"We've been talking to Osiris in one way or another for five or six years now because we've been in cell therapy for a long time ourselves," says Stephen Potter, senior vice president for corporate development at Genzyme, which received approval for the first cell therapy (Carticel) ever approved by the US Food and Drug Administration (FDA), in 1997. But there was no real movement toward a deal "until we started to see the phase 2 data in their first GvHD trial," he notes. "That really started to elevate our discussion, and I think they [Osiris] started to see the value in having a partner who had a more global reach."

Osiris Therapeutics

The head office of Osiris in Columbia, Maryland. The company is pursuing adult stem cell formulations for a raft of indications including joint injury, cardiovascular disease and diabetes.

Osiris has advanced its lead product Prochymal well into phase 3 clinical trials to treat severe refractory GvHD, a disease that causes life-threatening immune system reactions in the skin, gastrointestinal tract and kidney after a bone marrow transplant. Data from all three phase 3 Prochymal trials are expected mid-2009, after which the company expects to submit a biologic license application to the FDA. There is currently no approved drug to treat GvHD, and the FDA has granted Prochymal fast track designation and orphan disease status, which will expedite regulatory reviews and ensure marketing exclusivity for seven years. Prochymal has also been granted expanded access by FDA and Health Canada, so it is now available to any child aged 2 months to 17 years with end-stage GvHD without the need for compassionate-use paperwork.

In addition to GvHD indications, Prochymal is also in phase 3 trials for Crohn's disease, a gastrointestinal tract disorder. Studies suggest that MSCs can help regenerate intestinal tissue damaged by Crohn's that is necessary for absorbing nutrients. Prochymal is also in late-stage development for acute radiation syndrome, a condition that shares similar clinical manifestations to GvHD and Crohn's disease. In July 1997, Osiris and Genzyme entered an agreement to study radiation sickness. Then in January 2008, the partners were awarded a US Department of Defense contract worth $224.7 million to develop and stockpile Prochymal to counter nuclear terrorism and other radiological incidents. The biodefense deal was an important harbinger for Genzyme, says Potter. "It's always nice to get a test drive," he says.

Genzyme has a proven track record of marketing nonblockbuster products to produce good margins and returns for investors. More specifically, Genzyme's extensive product portfolio includes Epicel (cultured epidermal autografts) for treating patients with severe burns and Carticel (autologous cultured chondrocytes) marketed to the orthopedic surgery community for regeneration and repair of cartilage defects. This franchise complements Osiris' other MSC formulation: Chondrogen, a meniscus regeneration product, which is virtually the same substance as intravenously-administered Prochymal, except that it is formulated for direct intra-capsular injection into the knee.

The substance of this deal is MSCs made from adult volunteer bone marrow donors. Osiris selects donors from a pool of 18- to

30-year-old individuals. The cells are first isolated by density gradient separation and further purified by a selective adherence technique to eliminate non-MSC contaminants to 99.5% purity. From here, Osiris expands the culture to yield 10,000 doses of Prochymal from a single donor. According to University College's Mason, the company's novel method for expanding MSC cells, their experience in processing and packaging marrow-derived MSCs into an off-the-shelf product, was instrumental in clinching success. "The cells are high quality, the material is reproducible and the results excellent," Mason adds. From the investor point of view, it makes for a highly scalable and efficient business model.

> "Do I think the MSC therapy will be effective at counteracting a broad range of autoimmune diseases? Absolutely."

When infused into the patient, the MSCs are drawn to the sites of damage and inflammation, whether it is ischemic tissue in the myocardium immediately after a heart attack or the mucosal crypts in the colon of patients with Crohn's disease or the tissues affected by GvHD, which occurs in ~50% of bone marrow transplant patients. In fact, the colitis of Crohn's disease can resemble the histopathology of the GvHD-affected colon, which is what originally prompted investigators at Osiris to explore and develop the product in the former indication.

Because Prochymal and Chondrogen are allogeneic, some degree of immune response might be expected, but to date it appears that MSCs are either immune privileged or they turn off or disable immune cells. Of ~850 patients treated with Osiris's MSCs over the past decade, with some GvHD patients having received as many as 12 consecutive doses, there have been no infusion reactions either on initial or subsequent administration, even with doses given months later. "It's an innate property, not something we do to them," says Osiris CEO Randal Mills. "It's the stem cell equivalent of O-negative blood. They just lack the cell surface antigens."

MSCs seem to achieve their therapeutic effects by working as anti-inflammatory agents. Although the complete mechanism remains uncertain, MSCs down-modulate the immune response by suppressing tumor necrosis factor α and interferon γ production while boosting interleukin (IL)-10 and IL-4 secretion by T-helper 2 cells. Because of their pluripotency, MSCs might be expected to engraft the tissues they are intended to regenerate, but in the case of infused MSCs, engraftment appears to be transient. "In Crohn's, we do see the MSCs go to the intestine," says Mills. "They will engraft at the site of inflammation, but they don't persist there." The cells act more as a drug than as a reconstructing agent. "This is not strictly 'regenerative medicine', but it has that effect and the results are excellent," Mason points out.

Whatever the reason, Prochymal appears to work. Phase 2 results in 32 adult patients with acute GvHD showed an overall response rate of 94%, with a complete remission or response rate of 77% representing 24 patients. And results for end-stage refractory GvHD in 12 children, ranging from infant to 15 years, showed a 100% response rate with a complete remission or response in 58% or 7 children.

Osiris is also pursuing MSC formulations for type 1 diabetes, chronic obstructive pulmonary disease and the prevention of heart failure following acute myocardial infarction, all of which are in phase 2 clinical trials. And there's probably more on the way. "There's a lot of animal data suggesting multiple sclerosis as a target," says Mills. Rheumatoid arthritis, acute organ rejection and scleroderma are also targets. "Do I think the MSC therapy will be effective at counteracting a broad range of autoimmune diseases? Absolutely."

George S. Mack *Columbia, South Carolina*

New product approvals

RoActemra (tocilizumab)/F. Hoffmann-La Roche (Basel) and Chugai (Tokyo)

The European Commission approved RoActemra for adult patients with moderate to severe rheumatoid arthritis (RA) who have either responded inadequately to, or who were intolerant to, previous therapy with one or more disease-modifying anti-rheumatic drugs or tumor necrosis factor antagonists. RoActemra is the first humanized interleukin-6 (IL-6) receptor-inhibiting monoclonal antibody developed for RA.

Stelara (ustekinumab)/Janssen-Cilag (Beerse, Belgium) and Centocor (Horsham, Pennsylvania)

The European Commission approved Stelara for moderate to severe plaque psoriasis in adults who failed to respond to, or who have a contraindication to, or are intolerant to other systemic therapies including ciclosporin, methotrexate and PUVA (psoralen plus ultraviolet A light). Stelara is a human monoclonal antibody that targets the p40 sub-unit of cytokines interleukin-12 (IL-12) and interleukin-23 (IL-23). Centocor (a Johnson & Johnson subsidiary) discovered the drug and has exclusive marketing rights in the U.S.

IN brief

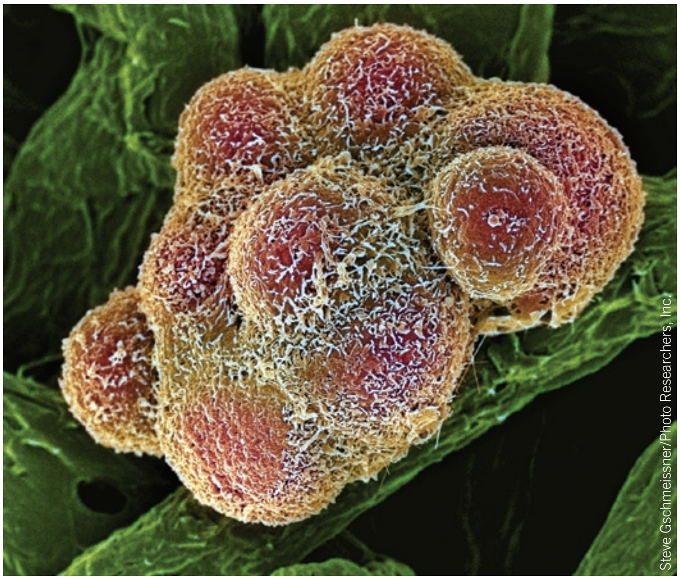
GM poplars to grow next door

Researchers at the Ghent, Belgium–based Flanders Institute of Biotechnology (VIB) have gained ground in a long-running battle over the planting of genetically modified (GM) poplar trees by applying for permits to plant the trees across the border. The Belgian government initially refused VIB's application to run field trials on home turf, but now the Dutch government, which has already issued a 'positive opinion', may grant them permission. The transgenic poplars are deficient in the enzyme cinnamoyl-CoA reductase, which reduces the lignin content making them more suitable for bioethanol production, although so far their benefits have only been demonstrated in the lab. The VIB had hoped for a green light from the Belgian Biosafety Council to run the trials closer to its research facilities and pilot-scale biorefinery. Instead, researchers will be forced to make regular trips to neighboring Holland to monitor and harvest the trees. Willy De Greef, secretary general of EuropaBio, the Brussels-based association for European bioindustry, says, "VIB is a public institute, which doesn't have the resources of a multinational. I don't even dare to think about what it does to their annual research budget." He says if European laws governing the planting of GM field trials were more consistently adhered to across member states, such situations wouldn't arise. A final decision from the Dutch government is due in spring 2009. *–Hayley Birch*

FDA goes public-friendly

In an effort to reach out to the American public, the US Food and Drug Administration (FDA) in December announced a collaboration with online health information provider WebMD to disseminate health and drug safety information. The deal gives the FDA pages within WebMD's website and print magazine. WebMD.com reaches a far larger audience than the FDA's website with nearly 50 million unique visitors each month compared with the FDA's 6 million. "It's important to put the information where the people are going, and not expect them to come to us," says FDA's Jason Brodsky, director of consumer health information. The WebMD-FDA site http://www.webmd.com/fda/ links to the agency's guides to reporting adverse events and understanding product recalls, and offers safety tips on drugs, medical devices, food and cosmetics. The agency plans to add multi-media content and features on the safe use of products. For example, the agency will offer a guide to parents on vaccines, warn consumers about unlawful distribution of unapproved drugs and answer questions such as, What are biologics? European agencies are also attempting to improve online access to health information. The UK's National Health Service (NHS) plans to launch in April NHS Evidence, a web-based service that consolidates clinical data and experience, prescribing and safety information, and technology appraisals. And the European Commission in December adopted a legislative proposal aimed to improve patient access to information about drugs. *–Emily Waltz*

Avastin-Tarceva combination fails in lung cancer



These breast cancer cells are not all identical. Companies are combining cancer agents that target different tumorigenic pathways to boost therapeutic success.

Steve Gschmeissner/Photo Researchers, Inc.

At the Chicago Multidisciplinary Symposium in Thoracic Oncology last November, Genentech of S. San Francisco, California, and OSI Pharmaceuticals of Melville, New York, presented data showing that a combination of Genentech's anti-vascular endothelial growth factor (VEGF) monoclonal antibody (mAb) Avastin (bevacizumab) and OSI's small-molecule inhibitor of epidermal growth factor receptor (EGFR) Tarceva (erlotinib) had failed to improve overall survival in patients with advanced non-small-cell lung cancer (NSCLC). The inability of the drug cocktail to meet the primary endpoints of the phase 3 trial is the latest of several setbacks for targeted combination therapies—regimens in which two or more separate agents targeting different pathways are simultaneously administered to patients to synergize their effect—once hailed as the future of oncology. Now that results from a clutch of similar trials of VEGF/EGFR inhibitor cocktails are in hand, observers are beginning to question whether the initial excitement over combination therapies was warranted.

Scientific rationale dictates that the use of two or more agents to target different pathways perturbed in cancer cells has a better chance of blocking tumor survival and metastasis than just one drug. Early combination therapies paired molecularly targeted mAbs or small molecules with traditional chemotherapies. But as more pathway-specific therapies gain approval, companies are exploring whether newer agents targeting particular molecular pathways involved in tumor signal transduction can achieve an additive or even synergistic effect, while avoiding overlapping toxicities.

Although the results of Genentech and OSI's BeTa Lung trial combining Avastin and Tarceva were disappointing, they were not entirely negative. Median progression-free survival (PFS), a secondary endpoint in this study, increased to 3.4 months, compared with 1.7 months for Tarceva alone. The combined therapy also had a positive impact on another secondary endpoint—objective response rate, a measure of tumor shrinkage—which doubled from 6.2% in the Tarceva monotherapy arm to 12.6% in the combination arm. But the primary endpoint—overall survival—failed to pan out as hoped. So what went wrong?

First, NSCLC is a highly heterogeneous cancer type, Edward Kim, an oncologist at MD Anderson Cancer Center in Houston points out. Targeted therapies for NSCLC provide benefit only to a subset of patients who have the appropriate genetic changes in their cells. So the signal of benefit among these individuals may have been swamped by 'noise' from 'genetically unsuitable' trial participants. NSCLC might be a particularly difficult cancer to treat with pathway-specific drug combinations, and success might be more likely in other cancers, suggests Kim.

Second, the availability of numerous therapies for lung cancer may be obscuring the readout. "Improvement in OS [overall survival] is becoming an increasingly challenging clinical endpoint in lung cancer clinical trials, as patients are receiving more additional therapies after each progression than they did in the past," writes Roche, based in Basel, on behalf of its subsidiary Genentech, in an e-mail. This can confound attempts to tease apart the effects of each component therapy in overall survival outcomes. Although the results from the BeTa Lung study are a clear setback for Genentech, the company still was able to give it a positive spin. "The commercial impact of these results is negligible as they do not affect the approved indications for bevacizumab and erlotinib," Roche adds.

Three other phase 3 trials have been testing EGFR- and VEGF-blocking combos in colorectal cancer (**Table 1**). So far, results have been discouraging. The Panitumumab Advanced Colorectal Cancer Evaluation (PACCE) study was the first to report results, in 2007. This trial compared regimes of a chemotherapeutic agent and Avastin with or without Thousand Oaks, California–based Amgen's human mAb Vectibix (panitumumab), which targets EGFR. An interim analysis revealed that the Vectibix cohorts experienced greater toxicity, and lower efficacy, prompting Amgen to pull Vectibix out of the trial.

Similarly, at the American Society of Clinical Oncology (ASCO) conference in Chicago last June, initial results were also disappointing from the CAIRO2 study—in which colorectal cancer patients were administered a combination of chemotherapy (Xeloda (capecitabine) and Eloxatin (oxaliplatin)) and Avastin, with or without New York–based ImClone Systems' anti-EGFR chimeric mAb Erbitux (cetuximab). Patients in the Erbitux-containing arm of this trial experienced more side effects and reduced benefit compared with controls. A third trial, CALGB 80405, is ongoing, and is looking at chemotherapy plus a combination of anti-EGFR/VEGF agents in colorectal cancer.

Despite these seeming failures, it would be premature to write an obituary for the strategy. "The development of targeted combination therapies is in its infancy," says David Chang, head of oncology at Amgen, who remains sanguine. "I don't think the disappointing results are a showstopper by any means," he says. Many other drugs, for instance, perform poorly until they are tested in the right setting or the right patients. Chang cites the case of single-agent Avastin, which produced negative results in early phase 3 trials as a second-line breast cancer therapy, but subsequently fared better in colorectal cancer and as first-line therapy in breast cancer.

Amgen is currently running, or has completed, 14 clinical trials of targeted combinations. "We're investing heavily into the concept of achieving a more comprehensive anti-angiogenic effect, for example, by inhibiting two major pathways in angiogenesis," says Chang. In this case, the idea is to hit not only VEGF using Avastin, but also the angiopoietin type 2 (Ang-2) pathway targeted by the investigational drug AMG-386, a peptibody (Fc fragment linked to 20-residue peptide that binds Ang-2) currently in phase 1 trials for ovarian cancer.

Roche is adopting a similarly robust attitude. "The majority of the relevant studies are in the early phases of development, so it is too early to comment on the outcome of this approach," says a company spokesperson, "and Roche will continue to investigate combining such therapies." They are currently running the phase 3 ATLAS trial using an anti-VEGF/anti-EGFR combo for the first-line treatment of patients with advanced NSCLC.

From the studies to date, one lesson is becoming apparent: combining pathway-targeted cancer therapeutics is not as side effect–free as might have been hoped. "All these targeted agents that are currently available affect major cellular pathways," says Amgen's Chang. "When you inhibit one, that might be tolerable; when you inhibit two or more, that may have effects that are not acceptable for clinical use."

As more targeted agents are approved, the choice of combinations will become much more complicated. Thus far, with only a handful of pathway-directed agents on the market, the approach has been largely empirical, guided by trial and error. This could soon change. "We're just not sure which drugs should be paired with which," says MD Anderson's Kim. "We need to figure this out, and find the best markers to indicate which patients should receive certain combinations."

Steps are already being taken in this direction. Just as Genentech's Herceptin (trastuzumab) shows efficacy only in HER2-positive breast cancers, so other targeted drugs work against a certain genetic background. The efficacy of Amgen's Vectibix in colorectal cancer, for example, is restricted to individuals without mutations in the *KRAS* signaling gene. If other targeted therapies show similar selectivity, then combinations of these therapies will have to take this fact into account.

On this basis, Paul Workman of the UK's Institute of Cancer Research in London says that it is unfair to judge the whole field of combination targeted therapy on the basis of the VEGF/EGFR inhibitor studies to date. "The full benefits of the approach will only come to fruition when we can really apply genetic stratification and pathway-activation profiling," he argues.

Perhaps the major challenge facing combination targeted therapies is to move away from a pragmatic empiricism to a more rational, scientifically based strategy. This requires integrating new insights into the pathway perturbations that drive various cancers with knowledge of how specific targeted agents act. "An understanding of the underlying fundamental biology should allow the right targeted therapeutics to be matched to generate a synergistic effect rather than the additivity that we've been used to," says Amgen's Chang. Capitalizing on combination therapies, in Workman's view, mandates a comprehensive systems biology perspective that will serve as its scientific foundation.

For all the false starts and dashed hopes, there is still an upbeat feeling about combination targeted therapies, both from the perspective of helping patients and commercial success. "If a drug combination works or improves efficacy compared with what a single agent would do, that in itself will increase market penetration or expand indications beyond the drugs' original use," says Chang. "But it is the scientific rationale that really drives interest in pursuing combination therapies."

The painful lessons that have been and continue to be learned in this pursuit should, however, eventually strengthen the field. "As we get more drugs, we'll have to ask the tougher questions, and that is what will force us to be more scientifically rigorous," says Kim.

Dan Jones *Brighton, UK*

Table 1 Selected efficacy trials of VEGF/EGFR combination therapies

Company	Trial description	Results
Amgen	Phase 3 (PACCE) trial of chemotherapy (folinic acid, 5-fluorouracil plus Eloxatin) and Avastin with or without Vectibix in 231 patients with advanced colorectal cancer.	Trial halted after preliminary review of data indicated increased toxicity and no increase in benefit for the treatment arm[a].
ImClone	Phase 3 (CAIRO) trial of Xeloda, Eloxatin and Avastin, with or without Erbitux in individuals with previously untreated metastatic colorectal cancer.	Median PFS 10.7 months versus 9.8 months in Erbitux arm; response rates 40.6% versus 43.9% and median overall survival 20.4 months versus 20.3 months. No significant difference in PFS or overall survival between patients with a *KRAS* mutation or those without.
OSI	Phase 3 (BeTa Lung)[b] trial of Tarceva and Avastin or Tarceva and placebo in 636 individuals with advanced NSCLC.	Did not meet primary endpoint of increasing overall survival. But median PFS on combination increased to 3.4 months versus 1.7 months for Tarceva alone and objective response rate rose to 12.6% versus 6.2% on Tarceva alone.

[a]Amgen is now studying Vectibix in combination with chemotherapy in colorectal cancer patient stratified according to *KRAS* status. [b]Results of a second study (ATLAS) in which OSI is evaluating Avastin and Tarceva for NSCLC patients whose disease has not progressed on Avastin or other chemotherapy are also expected in the first half of the year. PFS, progression-free survival.
Source: IDDB

IN brief

Stem cells caught in morality clause

The European Patent Office (EPO) will not be issuing patents for stem cells that have been obtained through the destruction of human embryos. The ruling announced last November invokes so-called 'morality clauses', invalidating the University of Wisconsin-Madison's key patent for a method of obtaining embryonic stem cell cultures from primates, including humans (the Wisconsin Alumni Research Foundation/Thomson patent will still be upheld in the US). Although the European ruling expressly rejects destruction of the human embryo, there is still some confusion. Aurora Plomer, professor of law and bioethics, University of Sheffield, says the ruling has "left open the question of whether specific moral exclusion extends to downstream derivative products, that is, products based on stem cell lines whose original derivation would have involved destruction of a human embryo." In Europe the situation remains quite fluid, with researchers bypassing the EPO by filing applications directly to their national patent office. But the stem cell ruling may have further implications, such as "increased costs for the industry, as investors revert to discrete selective national filings to secure patent protection on [human embryonic stem cell] inventions in favorable environments," says Plomer. This ruling comes as experts warn that the UK may lose its place as leader in the field, as Obama's administration has pledged to inject more money into federal funding of stem cell work. *–Nayanah Siva*

Land use stirs biofuels ruckus

The Biotech Industry Organization (BIO) has been asking the US Environmental Protection Agency (EPA) to publicly release its new methodology for calculating biofuels' life cycle greenhouse gas emissions, which will include emissions from indirect land use changes. The biotech industry needs "an actual measurement" of the effects biofuels have on the agricultural market, and how those effects are "translated into the actual land use around the world," says Paul Winters, BIO communications director. The calculations are required by the Energy Independence and Security Act (EISA) of 2007 and help determine which biofuels qualify for inclusion in the annual US quota for renewable fuel blended into gasoline, thus allowing the petroleum industry to purchase the biofuels to meet this quota. (For 2009, this quota is set at about 11 billion gallons.) Though some argue that indirect land use effects cannot be reliably measured, Tim Searchinger, visiting scholar at Princeton University in Princeton, New Jersey, counters that his analysis suggests even "the most heroic of assumptions" won't show that greenhouse gas emissions are reduced over "a reasonable period" by the use of biofuels in the gas supply. Regardless, BIO's biofuel members have a meaningful stake in the EPA's calculations. EPA has not set a date for the release of its Notice of Proposed Rule Making for EISA 2007. *–Susan Kim*

IN brief

Cuba's first GM corn

Cuba will be planting its first genetically modified (GM) corn to help reduce its dependence on costly food imports. The Cuban Center for Genetic Engineering and Biotechnology (CIGB) of Havana will begin the experimental plantation of 125 acres with the GM corn, provisionally called FR-Bt1. This corn is currently undergoing regulatory approval for its environmental release. "Cuban rules are very strict… but in Cuba there is a political will for employing the technology," explains Carlos Borroto, deputy director of the state-run center, and head of the Cuban National Program of Agricultural Biotechnology. The FR-Bt1, whose technical details cannot be revealed due to confidentiality clauses in the registration process, is aimed at animal feed and will be used exclusively in Cuba. The GM crop is engineered to resist the country's main pest: the lepidopteron *Spodoptera frugiperda*. The FR-Bt1 corn was developed by a large CIGB team, led by Camilo Ayra, in collaboration with other research bodies. The entire project was financed with public funds from the Cuban Council of State. "Because the corn has shown an elevated level of multiplication, some 2.5 acres could produce enough seeds to plant 300 acres," says Borroto. Although the use of GM organisms is debated in Cuba, public perception is mostly positive because these developments do not seek commercial gain but the nation's food sufficiency. The outcome of these field trials is expected for April 2009. *–Veronica Guerrero*

EU pushes advanced therapies

This month, the EU Committee for Advanced Therapies (CAT) will be holding its first workshop to discuss the implementation of a new legislation designed to harmonize gene therapies, cell therapies and tissue-engineered products within Europe. The lack of EU-wide regulatory frameworks for such novel therapies has, in the past, hampered the biotech industry's growth and hindered patient access. The recently passed EU Advanced Therapies Regulation lays down rules on the authorization, supervision and pharmacovigilance of newly emerging therapies. The committee, which is responsible for preparing draft opinions on quality, safety and efficacy of advanced therapies for final approval by the Committee for Medicinal Products for Human Use (CHMP), is part of the European Medicines Agency (EMEA). It includes representatives from CHMP, member states, clinicians and patient organizations. The regulation outlines a centralized marketing authorization procedure and special incentives for small and medium-sized enterprises (SMEs). Christiane Abouzeid, of the BioIndustry Association, believes that the CAT will help small companies by providing expert advice on complex products. An industry spokesperson notes that incentives for companies and investors within the new Advanced Therapies Regulation will more than offset any short term "pain" while procedures are set up. *–Susan Aldridge*

FDA holds court on *post hoc* data linking *KRAS* status to drug response

In mid-December, Amgen of Thousand Oaks, California, and its competitor ImClone Systems of New York jointly went in front of the US Food and Drug Administration (FDA) Oncologic Drug Advisory Committee (ODAC) to request permission to shrink the market for their products on the basis of genetic stratification of their target patient populations. Both argued, on the basis of retrospective analyses correlating mutation status with therapeutic response, that their respective anti–epidermal growth factor receptor (EGFR) monoclonal antibodies Vectibix (panitumumab) and Erbitux (cetuximab) for advanced colorectal cancer should be relabeled for use in only the 60% of individuals whose tumors harbor the wild-type *KRAS* gene. While the FDA continues to gather opinions and debate internally its criteria for biomarker validation, thus far the agency continues to be reluctant to consider retrospective data, even if such data indicate that a group of patients could be spared futile therapy.

Post hoc re-evaluation of clinical data runs counter to conventional statistical practice at the FDA. According to the agency's standard line of thinking, biomarker and therapeutic should be developed in parallel and endpoints designed prospectively in order for the validity of a hypothesis (and a related null hypothesis) to be tested. For its part, the FDA acknowledges that the science of drug development tied to prognostic indicators is moving at breakneck speed and that new developments may provide reasons for re-evaluating its stance—for example, in situations where patients could be spared futile treatment on the basis

ImClone Systems

ImClone and Amgen were hoping to include label warnings about *KRAS* mutations on their products to assist physicians in making treatment decisions for their patients.

SELECTED research collaborations

Partner 1	Partner 2	$ (millions)
Archemix (Cambridge, Massachusetts)	GlaxoSmithKline (GSK, London)	1,420
Dynavax (Berkeley, California)	GlaxoSmithKline (London)	810
Apitope (Bristol, UK)	Merck Serono (Geneva)	€154
BRAIN (Zwingenberg, Germany)	Genencor/Danisco (Palo Alto, California)	*

*Terms not disclosed.

of biomarkers identified after a clinical trial has begun or even after a product has been approved for marketing. As a result, the FDA is gathering information on how to assess retrospective biomarker usefulness, and the members of the ODAC panel seemed to be using the confab with Amgen and ImClone scientists as a sounding board to air their questions.

The FDA panel was upfront with its concerns and cited problems intrinsic to retrospective studies. The greatest angst was over "re-analysis of failed clinical trials" in search of alleged efficacy in subsets of biomarker-patient groups that may be undertaken without consideration for missing data and questionable assay techniques. This was a clear warning shot intended to discourage drug developers from rummaging through their discarded and well-worn products with the idea of manipulating data to make a drug fit some selectively back-tested and redistilled class of patient, biomarker and disease. "We are in agreement with that," says Hagop Youssoufian, senior vice president of clinical research and development at ImClone Systems, now a wholly owned subsidiary of Eli Lilly of Indianapolis, Indiana. "But that was never the purpose of the *KRAS* analysis anyway," he says. Both Erbitux and Vectibix won FDA approval as second-line therapies in the US without regard to biomarker status.

The ODAC panel didn't spend much time contesting the conclusions about *KRAS* status and drug response put forward by Amgen and ImClone speakers; instead, it acted more as devil's advocate. By nearly everyone's account, the all-day meeting offered a vigorous debate that was "constructive" and even friendly. "We viewed it as a collaborative effort to really try to move the field forward," says David Reese, who has been Amgen's global development leader for the Vectibix program. "We shared our data as an example of how these things actually transpire in the real world to help inform their thinking." Amgen's Vectibix won US approval in September 2006, and ImClone's Erbitux got its US approval in February 2004. But during the period between 2005 and 2007, new data were emerging (**Table 1**) that demonstrated *KRAS* status to be an important indicator with regard to the use of either Vectibix or Erbitux in colorectal cancers.

Meanwhile, European Union (EU) approval for Vectibix came at the end of 2007 complete with a label restriction to the *KRAS*-wild-type tumor subgroup. And Erbitux received a similar label restriction based on *KRAS* status more recently. In the EU, where public health systems prevail, product approval following European Medicines Agency (EMEA) recommendations does not guarantee that a patient will receive a drug. An insurer can veto a product's use if there is no compelling evidence of efficacy in a given situation, which makes off-label use close to impossible for most patients. Therefore, prognostic biomarkers are particularly valued because they often provide a more compelling risk/benefit ratio for individual patients.

In theory, for Amgen and ImClone's case, performing a new clinical trial with a *KRAS* status hypothesis and new endpoints would solve the FDA's quandary, but the reality is that such an idea just won't fly now. "With the

Box 1 FDA considerations for retrospective drug-biomarker analysis

The FDA has circulated six items that would likely be a minimum starting point for them to assess a retrospective analysis from a clinical trial:

- The trial must be adequate, well-conducted and well-controlled;
- The sample size must be sufficiently large to be likely to ensure random allocation to each of the study arms for factors (such as KRAS status) that were not used as stratification variables for randomization;
- Tumor tissue must be obtained in ≥95% of the registered and randomized study subjects and an evaluable result (presence of wild-type or mutant KRAS) must be available for ≥90% of the registered and randomized study subjects;
- Before analysis, the FDA must have reviewed the assay methodology and determined that it has acceptable analytical performance characteristics (for example, sensitivity, specificity, accuracy, precision) under the proposed conditions for clinical use;
- Genetic analysis must be performed according to the qualified assay method by individuals who are masked to treatment assignment and clinical outcome results;
- Before analysis of clinical outcomes based on the genetic testing, agreement with the FDA must be reached on the analytical plan for hypothesis testing for proposed labeling and promotional claims.

Source: Adapted from FDA's Oncologic Drugs Advisory Committee meeting briefing document, 16 December 2008

Details
Archemix and GSK have partnered to develop new aptamer therapeutics against interleukin-23 and six undisclosed targets with relevance to inflammatory disease. Archemix will receive $27.5 million upfront and is eligible to receive up to $200 million in development, regulatory and sales milestones for each of the seven aptamer products. The biotech would also receive tiered royalties up to lower double-digits.
The two companies plan to develop and market inhibitors of endosomal toll-like receptors (TLRs) to treat autoimmune and inflammatory diseases. GSK will pay $10 million upfront for an exclusive option to license four programs. The deal includes Dynavax's DV1079, a bifunctional TLR7 and TLR9 inhibitor about to start phase 1 testing. The biotech is entitled to about $200 million in milestone fees for each program.
The companies have signed a deal to collaborate on the development and commercialization of Apitope's peptide therapeutic for multiple sclerosis, which has just completed a preliminary clinical study. Apitope will receive up to €154 million in upfront, development and sales milestone payments. Apitope's peptide ATX-MS-1467 is designed to induce immunological tolerance to autoantigens involved in multiple sclerosis.
BRAIN has joined forces with Genencor to use metagenomics to develop enzymes for the production of products to replace petrochemicals in biofuels, plastics, rubber, adhesives and cosmetics. Genencor will use its capabilities in metabolic pathway engineering and biomanufacturing of industrial bioproducts. BRAIN will provide Genencor access to its technologies, in particular its metagenome resources of some 150 million genes of yet uncultured microorganisms. Enzymes and biosynthetic pathways of interest will be genetically engineered in microbial production strains for the production of biochemicals.

Table 1 *KRAS* status and response to EGFR antibodies in colorectal cancer

Publication	Treatment (panitumumab or cetuximab)	Number of subjects (wild type; mutant)	Objective response, *n* (%)	
			Mutant	Wild type
Lancet Oncol. **6**, 279–286 (2005)	Panitumumab or cetuximab or cetuximab + chemotherapy	31 (21; 10)	2 (20)	8 (38)
Cancer Res. **67**, 2643–2648 (2007)	Panitumumab or cetuximab or cetuximab + chemotherapy	48 (32; 16)	1 (6)	10 (31)
J. Clin. Oncol. **25**, 4132 (2007)	Cetuximab ± chemotherapy	37 (20; 17)	0 (0)	17 (46)
J. Clin. Oncol. **2**, 4021 (2007)	Cetuximab ± chemotherapy	81 (49; 32)	2 (6.3)	13 (26.5)
J. Clin. Oncol. **25**, 3230–3237 (2007)	Cetuximab	80 (50; 30)	0 (0)	5 (10)
AACR Meeting Abstracts **2007**, 5671 (2007)	Cetuximab ± chemotherapy	78 (49; 27)	0 (0)	24 (49)

overwhelming consistency of data we have for Erbitux and Vectibix, no one has any inclination to administer these drugs to patients with *KRAS*-mutant tumors," says ImClone's Youssoufian. "You would be randomizing and subjecting patients to a drug that could, at minimum, be neutral, if not associated with adverse effects."

The data are out for oncologists to see, and nearly every clinician is ostensibly aware that *KRAS* mutant–harboring tumors are no longer candidates for anti-EGFR products. In addition, the National Comprehensive Cancer Network's guidelines make a strong recommendation to reserve Erbitux and Vectibix for tumors with the wild-type *KRAS*. So the message is sinking in. "The problem for us as a company," says Youssoufian, "is that we cannot proactively disseminate that message because it's simply not in our label." And Amgen's Reese says, "We think it's the correct science, and we have an obligation to communicate the appropriate information to patients and physicians, and the mechanism for us to do that is the label. That formed the basis of our proposal."

If Amgen and ImClone were lingering under any illusion that the discussion might lead ODAC to recommend that the agency include *KRAS* status in their drug labels, they were proven wrong. The panel clearly didn't feel inclined to comment—at least for now. "Really, everyone has already implemented *KRAS* testing, so what is the big deal if the agency makes a decision right away or not?" says senior biotech analyst Aaron Reames of Wachovia Securities. "The issue is not with the anti-EGFR monoclonal antibodies in this case but rather the precedent that this decision will set."

Biostatistician Richard Simon of the National Cancer Institute of Bethesda, Maryland, favors the Amgen-ImClone proposal to tie *KRAS* status to anti-EGFR therapy. But he understands the suspicion engendered by retrospective studies. "When I've given talks, I have gotten questions from people at the FDA who express concern about companies reanalyzing the same clinical trial with regard to multiple biomarkers," he says. "The kind of retrospective studies that we're used to seeing are not very reliable because they don't have a single biomarker hypothesis. There is skepticism because we've seen so much garbage come from *post hoc* data dredging."

The FDA presented Amgen and ImClone with six considerations for a valid retrospective analysis (see **Box 1**). One of the more important issues was the type of *KRAS* assay used (*Nat. Biotechnol.* **26**, 839–840, 2008). Both companies are largely in compliance with the points in question, but if or exactly when their labels will be revised is anybody's guess. An Amgen spokesperson will say only that the company remains in "productive discussions" with the FDA.

The potential financial consequences of a label change for Amgen and ImClone are also unclear. Senior biotech analyst Eric Schmidt of Cowen does not think limiting Vectibix and Erbitux will have a substantial impact on the companies' profits. Of the estimated 40% colorectal cancer patients with *KRAS*-mutant tumors, he says: "They were likely receiving shorter courses of therapy due to their unresponsiveness." He expects no more than a 10–30% decline in product revenues; however, he envisions a potential offset resulting from more frequent testing for *KRAS* status. "Awareness that a patient is *KRAS* wild type could drive adoption of Erbitux and Vectibix in more second-line patients," he says. From Schmidt's point of view, there's no significant dollar loss or gain resulting from limiting use of the drugs. "I think these companies simply believe this is the right thing to do, and that they are facilitating better medicine, better utilization of healthcare dollars and better citizenship," he says.

George S. Mack *Columbia, South Carolina*

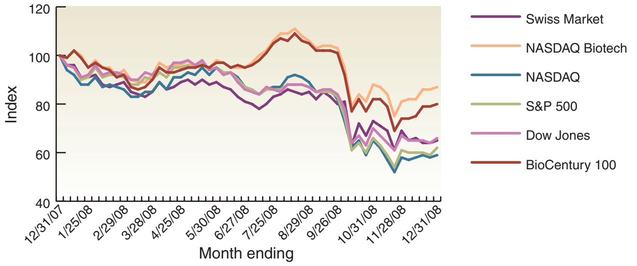
2008—down, but not out

Walter Yang

The biotech sector received $20 billion less in funding in 2008 than the previous year. Public equity markets were particularly hard hit, with IPOs (initial public offerings), follow-ons and PIPEs (private investment in public equity) all down 35% or more compared to 2007. Venture financings were also off one-fifth from the $6.8 billion posted in 2007. Funding from partnerships came in at $20 billion, compared with $22 billion for 2007. In general, biotech indexes performed better than the Dow Jones Industrial Average and Standard & Poor's 500.

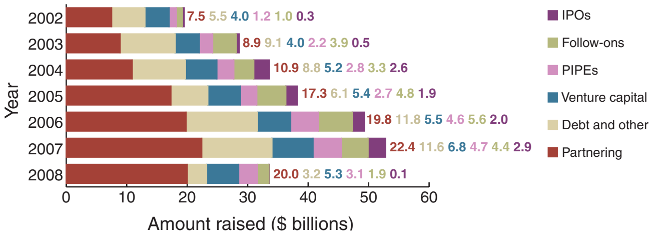
Stock market performance

The overall markets were off more than 30% on the year, whereas the BioCentury 100 index was down 20%.



Global biotech industry financing

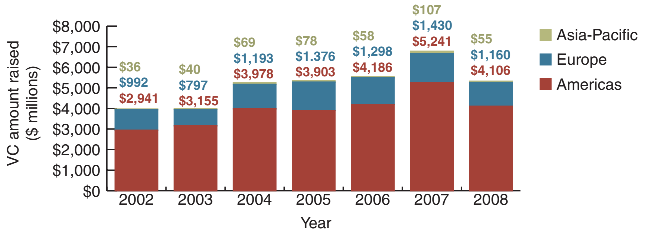
Including partnership promises to US companies, biotech financing dropped to $34 billion from $53 billion in 2007.



PIPEs, private investment in public equity; IPOs, initial public offerings. Source: BCIQ: BioCentury Online Intelligence, Burrill & Co.

Global biotech venture capital (VC) investment

Private companies raised about $5 billion in 2008, similar to levels seen in 2004–2006.

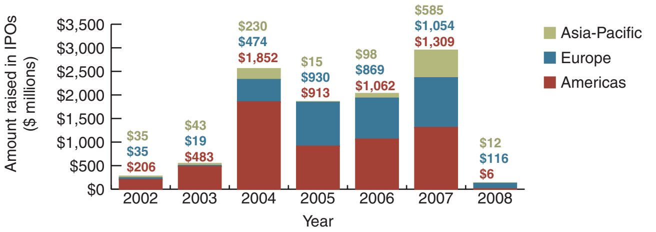


	2002	2003	2004	2005	2006	2007	2008
Americas	200	184	186	182	195	217	207
Europe	91	77	79	92	81	101	93
Asia-Pacific	6	5	6	9	8	9	6

Table indicates number of VC investments. Source: BCIQ: BioCentury Online Intelligence

Global biotech initial public offerings (IPOs)

Only six IPOs were completed in 2008, and only one of those listed in the US.



	2002	2003	2004	2005	2006	2007	2008
Americas	5	8	34	18	25	23	1
Europe	3	1	12	24	21	21	3
Asia-Pacific	2	5	6	3	3	7	2

Table indicates number of IPOs. Source: BCIQ: BioCentury Online Intelligence

Notable 2008 deals

IPOs	Company (lead underwriters)	Amount raised ($ millions)	Percent change in stock price since offer	Date completed
	MolMed (Banca IMI, Societe Generale)	$85.4	–50%	29-Feb
	Ipsogen (Bryan, Garnier & Co.)	$18.6	–19%	10-Jun
	PCI Biotech (Fondsfinans)	$12.0	–45%	10-Jun
	Fluorotechnics	$7.0	–1%	24-Oct
	Bioheart (Dawson James)	$5.8	–81%	19-Feb
	Genera Biosystems (Domain Capital)	$4.8	–52%	21-May

Venture capital	Company (lead investors)	Amount raised ($ millions)	Round number	Date closed
	OncoMed (Adams Street Partners)	$169.0	2	12-Dec
	Portola (No lead)	$130.0	3	9-Jul
	Pacific Biosciences (Deerfield, Intel Capital)	$100.0	5	14-Jul
	Radius Health (MPM Capital, Wellcome Trust, MPM Bio IV NVS)	$82.5	3	20-Nov
	Ganymed (ATS Beteiligungsverwaltung)	$82.2	4	18-Nov
	Proteolix (Nomura Phase4 Ventures)	$79.0	3	8-Sep
	Gemin X (Caxton Advantage Life Sciences Fund, Caxton Global)	$76.0	3	30-Jun

Mergers and acquisitions	Target	Acquirer	Value ($ millions)	Date announced
	Genentech	Roche	$43,700	21-Jul
	Millennium	Takeda	$8,200	10-Apr
	Applied Biosystems	Invitrogen	$6,700	12-Jun
	ImClone	Eli Lilly	$6,500	6-Oct
	LifeCell	Kinetic Concepts	$1,700	7-Apr
	Speedel	Novartis	$900	10-Jul

Licensing /collaboration

Researcher	Investor	Value ($ millions)	Deal description
Actelion	GlaxoSmithKline	$3,246.8	Co-develop and co-commercialize phase 3 insomnia compound almorexant outside of Japan
Isis	Genzyme	$1,900.0	Develop mipomersen and follow-on compounds to treat high cholesterol
Acceleron	Celgene	$1,878.0	Jointly develop and commercialize ACE-011 to treat cancer; option for three other programs
Archemix	GlaxoSmithKline	$1,427.5	Develop and commercialize aptamer therapeutics to treat inflammatory diseases
Amgen	Takeda	$1,177.0	Received Japanese rights to 12 compounds and partnered to develop motesanib
PDL (now Facet)	Bristol-Myers	$1,155.0	Develop and commercialize multiple myeloma monoclonal antibody (mAb) elotuzumab (HuLuc63); option to include PDL241
Exelixis	Bristol-Myers	$1,000.0	Develop and commercialize phase 3 XL184 and phase 1 compound XL28
Alnylam	Takeda	>$1,000.0	Five-year collaboration to develop RNA interference therapeutics

Walter Yang is research director at BioCentury

Glial cells on the radar

Long thought of as passive bystanders, glial cells are coming under increasing scrutiny as mediators of inflammatory disease in the nervous system. Now, some drug makers are hoping they can be targeted pharmacologically. Cormac Sheridan reports.

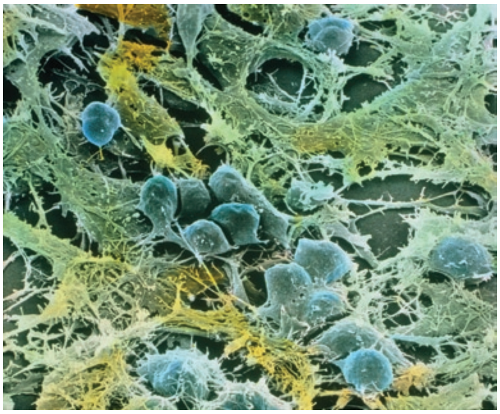
After years of neglect, glial cells are finally registering on drug developers' radar. This reflects increasing knowledge not only about their immunological functions, particularly those related to the production of inflammatory cytokines, but also about their myriad interactions with neurons. The insights are opening up new possibilities for treating a wide range of central nervous system (CNS) conditions, including Alzheimer's disease, Parkinson's disease, neuropathic pain, epilepsy, spinal cord injury, multiple sclerosis and traumatic brain injury. Indeed, the US National Institute on Drug Abuse's (NIDA) phase 2 trial of Alameda, California–based Avigen's small-molecule ibudilast (AV-411) for opioid addiction, initiated two months ago, is just one of several early-stage clinical trials of glial cell modulators currently underway (**Table 1**). And the results of these programs are generating more than idle interest; just a few weeks ago, San Diego-based biopharmaceutical company MediciNova announced a takeover bid for Avigen and its portfolio.

From Dangerfield to drugs

Northwestern University's Linda Van Eldik says glial cells are often jokingly referred to as the 'Rodney Dangerfields' of the brain—a reference to the hapless comedian who wove an entire career from his famous catchphrase: "I don't get no respect." Traditionally considered passive 'bystander' cells, which merely offered physical support to neurons, glia—the word is actually derived from the Greek for glue—are, in fact, central actors in the development and regulation of the CNS, even if they do not participate directly in nerve signal transmission

Although glial cells were originally identified in the nineteenth century, Joyce DeLeo of Dartmouth Hitchcock Medical Center in Lebanon, New Hampshire, says that, even a decade ago, the field hadn't progressed much further than basic characterization of the cells. That changed with the development of improved tools, such as antibodies, agents that modulate glial cell function, more sophisticated electrophysiology instruments and transgenic mice, which now allow scientists to probe function with more precision. Increasing multidisciplinarity has also boosted information flow, as researchers who had previously focused exclusively on neurons have begun to consider glial cells in their experimental designs.

Glial cells come in many forms, but three major types are currently recognized: astrocytes, oligodendrocytes and microglia. The first are primarily involved in creating the blood-brain barrier and regulating nerve synapses, oligodendrocytes in myelinating axons, and microglia in removing unwanted cells. Acting together, these cells play numerous roles in the neural system: maintaining neural homeostasis, storing energy, buffering pH, balancing ion concentration and recycling neurotransmitters.



P.M. Motta & D. Palermo/Photo Researchers

Moving off target from neurons to glia (shown) may provide relief for chronic pain sufferers.

In particular, astrocytes and microglia are the immunocompetent cells of the CNS and are activated following tissue injury, infection or inflammation. What's more, evidence is growing that dysregulated glia might be mediators of disease pathogenesis in themselves. Thus, the expression of mutant-superoxide dismutase 1, mutant ataxin, mutant huntingtin and monamine oxidase B in astrocytes has been associated with disease processes in amyotrophic lateral sclerosis (ALS), spinocerebellar ataxia (SCA), Huntington's disease and Parkinson's disease. And glutamate-mediated excitotoxicity associated with glial cells has been proposed as a significant component in ALS, SCA and Huntington's disease.

Given that they divide actively, are ten times more numerous than neurons in the brain, constitute around 70% of the brain's total cell population and express many of the mutant proteins involved in neurodegenerative disease, it is no surprise that glial cells are beginning to pique the curiosity of drug developers around the world. And the interest centers not only on targeting them in neurodegenerative disease but also on modulating glial-neuron interactions for such indications as neuropathic pain.

Glia and neuropathic pain

According to Dartmouth's DeLeo—who has been working on glial cell biology since the 1980s and has played a central role in identifying the links between pain and glial cell cytokine production—the entire field is still largely at the target discovery phase (**Box 1**). "We haven't found any significant intracellular pathways that are specific to one type of glial cell," she says. "That's the challenge." There is, however, "receptor heterogeneity," she says. Different glial cell types can be targeted by specific extracellular proteins, such as toll-like receptors (TLRs), integrins and ion channels.

DeLeo has also published data indicating the involvement of TLR-4 on microglia in the activation of the CNS innate immune response and the subsequent development of pain in rodent models of neuropathy[1]. However, modulation, rather than outright inhibition of glial cells and their receptors, is the pharmacological goal. "We can live with chronic pain, but if you totally alter how the brain recognizes foreign pathogens, that's deleterious," she says.

As a cofounder of Boston-based Solace Pharmaceuticals, an early-stage firm focused exclusively on pain, DeLeo is now attempting to commercialize her insights. On the basis of her work and that of several other pain biologists, Solace has in-licensed its lead compound, SLC022 (the methylxanthine-derivative propentofylline), which modulates glial cell activity and has been shown to modulate glutamic acid decarboxylase, a synthase of the inhibitory neurotransmitter γ-aminobutyric acid (GABA). Aventis Pharma (Frankfurt) had completed phase 3 trials of SLC022 in Alzheimer's but discarded the compound for lack of efficacy. According to DeLeo, the compound is due to enter a phase 2 trial involving approximately 200 sufferers of chronic neuralgia, brought on by herpes zoster infection, later this year.

Elsewhere in Boston, Ru-Rong Ji, associate professor at Brigham & Women's Hospital and

Table 1 Selected glial cell modulators in development

Company (location)	Compound	Mechanism	Indications	Stage
Allon Therapeutics (Vancouver)	AI-108AI-208	Promote microtubule assembly by binding tubulin	Alzheimer's disease	Phase 2
			Stroke damage	Phase 2
Avigen	AV-411 (ibudilast)	Phosphodiesterase inhibitor[a]	Neuropathic pain	Phase 2a[b]
			Opioid dependence	Phase 2
Evotec	EVT 302	Selective MAO-B inhibitor	Smoking cessation	Phase 2a
			Alzheimer's disease	Phase 1
Key Neurotech Pharmaceuticals	KMN38-7271	Cannabinoid receptor agonist	Traumatic brain injury	Phase 2
Neurim Pharmaceuticals (Tel Aviv)	NEU-120	MAO-B inhibitor	Parkinson's disease	Phase 2[a]
Solace Pharmaceuticals	SLC022 (pro-pentofylline)	Modulates glutamic acid decarboxylase, a GABA synthase	Neuropathic pain	Phase 2

[a]Alternative mechanism of action may be associated with glial cell activity. [b]Approved in Japan and other countries in Asia for treating bronchial asthma, and dizziness following cerebral stroke.

Harvard Medical School, has also been looking at glial cells and neuropathic pain, specifically the molecular mechanism underlying a feedback loop between glial cells and neurons involved in propagating the pain response. His group's recent work indicates that matrix metalloproteinase 9 (MMP-9) and MMP-2, respectively, are responsible for the development of early- and late-phase neuropathic pain in rodent models[2]. "Current treatments just block neurotransmission," he says. In other words, they are effective only while the drug is present. But Li is looking to block, or at least damp down, the feedback between glia and neurons so the pain can be controlled more effectively. Furthermore, his approach suggests different treatments could be tailored for short-term neuropathic pain associated with surgery, for example, or for chronic pain, such as that associated with diabetic neuropathy.

Ji's group has already demonstrated the feasibility of his approach in rodents, using L5 spinal nerve ligation, a widely used animal model of neuropathic pain. Both MMP-9 and MMP-2, which are released from neurons after injury, are responsible for cleaving the pro-inflammatory cytokine interleukin-1β (IL-1β) into its active form. MMP-9–mediated cleavage of the cytokine appears to activate microglia—the brain's macrophages—by means of the p38 mitogen-activated protein kinase (MAPK) pathway, whereas MMP-2–mediated cleavage of the same molecule activates astrocytes through extracellular signal-regulated kinase. Activation of these cells stimulates further release of IL-1β, which exacerbates the propagation of the pain signal. "IL-1β in itself is a trigger," he says.

His group has additional, as yet unpublished, data indicating that IL-1β can both enhance excitatory synaptic transmission and reduce inhibitory synaptic transmission in extracts of spinal cord tissue. Intriguingly, the administration of tissue inhibitors of metalloproteases-1 (TIMP-1) and TIMP-2, endogenous MMP inhibitors, has a powerful effect on reducing pain. Ji is now considering how to develop therapeutic approaches based on this work, an effort that is at an early stage, he says.

One of the key activities of Avigen's glial cell modulator ibudilast is to suppress IL-1β as well as tumor necrosis factor α and IL-6. Avigen has been developing ibudilast for the treatment of neuropathic pain and opiate dependence. In November 2007, Avigen reported data from an Australian phase 2b trial in neuropathic pain, showing that ibudilast was well tolerated at all doses up to 80 mg/day, with plasma levels associated with a reduction in reported pain scores. Plans for a phase 2b trial in the US were put on hold in December while Avigen looked for a partner. In the meantime, the company is testing ibudilast in a phase 2 trial, funded by NIDA and initiated in collaboration with the New York State Psychiatric Institute and Columbia University, designed to wean 30 heroin abusers off morphine. The primary endpoint is reducing symptoms of opioid withdrawal, with results expected later in the year.

Cannabinoid receptors

Given the key role of glial cells in maintaining homeostasis in the brain and spine, intense interest also surrounds the development of compounds that target astrocytes, oligodendrocytes and microglia, either directly or indirectly, for a wide range of CNS indications. Magdeburg, Germany–based KeyNeurotek Pharmaceuticals is targeting one of the most challenging and poorly understood conditions, traumatic brain injury. It has a phase 2a clinical trial underway of KMN38-7271 (formerly Bay 38-7271), a cannabinoid receptor agonist, in severely brain-injured patients, a condition with no approved therapy. Its drug candidate, which it in-licensed from Leverkusen, Germany–based Bayer HealthCare in 2005, binds to both cannabinoid subtypes, cannabinoid 1 (CB1) receptor and cannabinoid 2 (CB2) receptor, and the company posits a dual mode of action for the compound.

CB1 receptor is normally expressed on neuronal cells and plays a role in regulating neuronal excitation and neurotransmitter release. When brain injury occurs, neurons in the affected region die by necrosis or apoptosis, events that trigger inflammatory responses involving microglia and astrocytes. "In parallel you see upregulation of the CB2 receptor. Before this activity there are no CB2 receptors present," says KeyNeurotek CEO Frank Striggow. Activating these CB2 receptors, which are expressed on glial cells, is thought to dampen excessive brain inflammation. Binding of KMN38-7271 to CB1 receptors, whose expression profile remains unchanged during injury, says Striggow, is considered to have a neuroprotective role by restoring ion homeostasis to surviving neurons. Whether these effects will translate into any clinical benefit should become apparent later this year, when KeyNeurotek hopes to report data from the placebo-controlled study, which commenced in October 2006. "We hope to see some indication of a therapeutic effect," he says. The study will follow a range of clinical parameters and outcomes, including death, length of coma and level of residual disability. The company is also planning to test the compound in stroke.

KeyNeurotek and Bayer have, between them, completed four phase 1 trials in healthy volunteers and have found the compound to be safe and well tolerated across a range of doses. "I strongly believe that the agonists are much safer than the antagonists. The endogenous cannabinoids are very important for the brain," says Striggow. Nevertheless, there is, at the very least, a theoretical possibility that excessive activation of cannabinoid receptors could result in psychogenic effects. "You have to work with the right dose."

Monoamine oxidase B inhibitors

Another German firm, Hamburg-based Evotec, is exploring the potential of EVT 302, a highly selective inhibitor of monoamine oxidase B (MAO-B), in Alzheimer's disease and smoking cessation. Like KeyNeurotek, it too in-licensed the compound from a large pharmaceutical firm, in this case, Basel-based Roche. It is a successor to an earlier MAO-B inhibitor lazabemide, which demonstrated an effect on disease progression in a phase 3 trial in Alzheimer's disease before being abandoned because of safety issues that arose during the extension period of the study. "The Roche data have never been published but it was very encouraging," says John Kemp, chief R&D officer at Evotec (and Roche's former head of CNS research).

EVT 302 belongs to a different chemical class, however, although it acts through the same mechanism. It is based on the observation that activated astrocytes have elevated MAO-B activity, a phenomenon that is evident in several neurodegenerative conditions, including Alzheimer's disease. "You get this dramatic upregulation of MAO-B, particularly surrounding the amyloid plaques," says Kemp. MAO-B cleavage of its normal substrate, dopamine, leads to the formation of hydrogen peroxide, which, in turn, can generate tissue-damaging hydroxyl free-radicals. Evotec has completed a phase 1 study in Alzheimer's disease, but it is not actively pursuing that indication at present, however, because of budgetary reasons.

Instead, Evotec is running a phase 2 trial of the compound in smoking cessation, where it may have a dual mode of action. As well as interfering with the dopamine reward system, which is stimulated by nicotine, EVT 302 may also alleviate withdrawal symptoms, as dopamine levels are depressed in long-term smokers. It aims to secure a partner in both indications after completion of the phase 2 program, which will involve a second trial later this year.

Its potential could extend to other conditions as well. "Parkinson's disease is obviously one," says Kemp, noting that a couple of MAO-B inhibitors are already approved in that indication, namely Azilect (rasagiline), which is jointly marketed by Teva, of Petach Tikva, Israel, and Copenhagen-based Lundbeck; and Zelapar (selegiline), which Valeant Pharmaceuticals, of Aliso Viejo, California, markets. Additional support for the therapeutic rationale was recently published by a group at the Buck Institute for Age Research, in Novato, California, and collaborators. These investigators found that transgenic mice, engineered to express elevated levels of MAO-B within their astrocytes, demonstrated several phenotypes associated with Parkinson's disease[3].

Evotec's interest in glial cell modulators could also lead to other opportunities. A group based at Inserm, in Bordeaux, France, and collaborating institutions recently reported that D-serine released from astrocytes acts as a co-agonist, along with the neurotransmitter glutamate, of the *N*-methyl-D-aspartic acid (NMDA) receptor in a specific sub-region of the rat hypothalamus[4]. Through this linkage, the astrocytes can exert a direct influence on NMDA functioning, which is linked to learning and memory. "We have a strong interest in NMDA function," says Evotec's Kemp. "You can affect the level of D-serine in various ways."

Moving off target in epilepsy

The same basic concept also applies to epilepsy, another indication in which glial cell modulation could offer an alternative and, possibly, more benign approach to therapy. "If you hit activated glia you hit pathways that are less likely to be important in normal physiological conditions," says Annamaria Vezzani, at the Mario Negri Institute for Pharmacological Research, in Milan. Up to now, antiepileptic drugs have mainly targeted neurons, and they can impose a heavy burden of side effects on those who take them. Vezzani's group has been among the leaders in identifying a link between IL-1β–driven brain inflammation, mediated by microglia and astrocytes, and the occurrence of, or predisposition toward, seizures[5]. In animal models, elevated levels of the cytokine in specific areas can lower the threshold at which animals become susceptible to seizures. "It's not model specific. It's reproducible in various models," she says. Her group has also found glial cell activation of the IL-1β signaling system in sections of human brain tissue.

Targeting these inflammatory processes is therefore a distinct possibility, which she is exploring with an undisclosed US company. "It's quite interesting. There are drugs already in clinical use that can be exploited in epilepsy," she says. The arthritis drug Kineret (anakinra), a recombinant form of the IL-1 receptor antagonist (IL-1Ra), marketed by Thousand Oaks, California–based Amgen, is one possibility, she says. "The brain does not produce this antagonist in great amounts." Caspase-1 inhibitors, which prevent the conversion of pro-IL-1β to its active form, represent another possibility.

Even so, this effort, like so many others involving glial cell modulation, is still at a preliminary stage. The steady accumulation of knowledge about the multifarious roles of glial cells, in both healthy and diseased states, has greatly added to the complexity of our picture of the CNS. However, it has also greatly increased the range of potential targets for therapeutic intervention. Pinpointing those that could lead to real treatment advances is going to be one of the key tasks in the further development of this field.

Cormac Sheridan, Dublin

Box 1 Screening for glial activators

Linda Van Eldik and Marty Watterson, codirectors of the Center for Drug Discovery and Chemical Biology at Northwestern University in Chicago have taken a functional approach to drug discovery in an effort to find molecules that influence glial cell activity. Their team has developed cell-based screens for identifying compounds that selectively inhibit the production of cytokines from activated glial cells. One drug candidate that emerged from this program, minozac, is now in development at Transition Therapeutics, of Toronto, following its acquisition of the compound's original licensee, Neuromedix, also of Toronto. The precise mechanism of action is not fully understood as yet. "We're affecting signaling pathways inside the cell," says Watterson. Even so, the investigators have already demonstrated striking efficacy in animal models of Alzheimer's disease, in terms of ability both to suppress cytokine production in the presence of amyloid-β peptide and to improve deficits in behavioral tests.

More recently, based on minozac's oral bioavailability, ability to penetrate the blood-brain barrier and lack of toxicity, Van Eldik and Watterson have used the same scaffold to develop a follow-on compound, 069A. This time around, however, they added a chemical group—a pyridinyl pharmacophore—that is found in p38 MAPK inhibitors. Early results indicate that this mode of action also demonstrates efficacy in a mouse model of Alzheimer's, in which disease is induced by administration of human amyloid-β peptide[6]. "That's a lead compound but it's not a candidate for clinical development yet," says Watterson. Further medicinal chemistry refinement will be necessary before it can be taken toward the clinic.

Like most others who are active in this field, Van Eldik and Watterson are seeking to modulate excessive glial cell activity rather than suppress it globally. "You're bringing it back to normal homeostasis," Van Eldik says. "Usually the cytokines are in very low levels."

1. Tango, F.Y. *et al. Proc. Natl. Acad. Sci. USA* **102**, 5856–5861 (2005).
2. Kawasaki, Y. *et al. Nat. Med.* **14**, 331–336 (2008).
3. Mallajosyula, J.K. *et al. PLoS ONE* **3**, e1616 (2008).
4. Panatier, A. *et al. Cell* **125**, 775–784 (2006).
5. Vezzani, A. *et al. Epilepsia* **49** Suppl. 2, 24–32 (2008).
6. Munoz, L. *et al. J. Neuroinflammation* **4**, 21 (2007).

Avoiding the obvious

Sherry L Murphy & Kenneth D Sibley

Obviousness is one of the most common reasons for examiners rejecting patent applications. What can you do to limit the chances of such a setback?

After years of hard work, you finally have an invention and file for a patent application. Some time later, the application is reviewed by an examiner at the US Patent and Trademark Office (USPTO), and an Office Action is issued that states various reasons why the application is being rejected. One of the reasons is that your invention is 'obvious'.

Obvious? How is that possible?

The patentability requirement of nonobviousness is a hurdle often faced by inventors negotiating meaningful patent protection from the USPTO. Recent decisions from the courts have made an obviousness rejection more difficult to overcome (**Box 1**). But, as we outline below, a better understanding of the obviousness hurdle and how to overcome it may mean the difference between success and failure in the prosecution process and ultimately obtaining protection for your innovations.

I could have thought of that

The legal analysis underlying a conclusion of obviousness is complex, and the USPTO, courts, attorneys and commentators have grappled with the concept for just about as long as patent protection has been available in the US[1]. Therefore, we start with a gross oversimplification: obviousness is exactly what one would think. In essence, it is the examiner concluding, 'I could have thought of that'.

To be more technically correct, obviousness is the legal conclusion that (back when the invention was made) a hypothetical person with adequate background and skill in the relevant technical field would have followed an existing motivation to solve a known problem in order to combine the teachings from the references (unearthed by the examiner) that were in existence at the time of filing, and thus create the invention (note that the 'known problem' does *not* have to be the same problem that the inventor is solving).

In practice, an examiner may peruse the application, look at the listing of claims being made and pick out certain aspects of those claims. He or she may then search the literature that was available prior to the filing date of the application and find those aspects, piece by piece, in one or more references. In drafting the obviousness rejection, the examiner will articulate some rational basis as to why one might combine these references to produce the claimed invention.

As is apparent from such a scenario, the conclusion of obviousness is necessarily made looking back, and therefore, hindsight is an important problem. Examiners do what they can with their limited time to review the application and the plethora of papers thrown at them by attorneys. However, to overcome such a rejection, the inventor may need to explain to the examiner (and to the attorney) why the conclusion of obviousness is not a reasonable one and why the picture painted by the examiner is not an accurate portrayal of the state of the technical field at the time of invention.

Prepare to attack

As an inventor, you may think, "Okay, so the examiner takes two or three patents (or other references) that do not even deal with the same problem as my invention and then calls the invention 'obvious'? That must be easy to overcome." Not necessarily. That is why your input is key.

After the initial obviousness rejection from the USPTO, the 'burden of persuasion' is on you, the applicant. There are different ways to go about responding to the rejection, which can be generally divided into the two categories of 'attack' and 'rebut'.

To illustrate the difference in practical terms, which would you prefer: being asked by your attorney to talk about the field and state of the art at the time of filing (gathering evidence to attack) or being told to conduct a complicated and expensive set of experiments in a short, fixed time period (gathering evidence to rebut)? Naturally, attacking is preferable. Therefore, the discussion that follows is focused on this. Do remember, however, that you are attacking the rejection, not the examiner. It is the examiner's job to play the devil's advocate, and by doing this job properly, the examiner can actually make your patent (once issued) stronger by building a strong foundational administrative history[2].

The presentation of evidence against obviousness—that is, a presentation of relevant facts from which a legal conclusion of nonobviousness should follow—is an area of frequent deficiency among applicants. For many years, a *de minimis* approach has generally been used to respond to obviousness rejections. However, recent developments in the law instruct otherwise.

Therefore, it is important for you—the inventor—and your attorney to dig deep and present the best evidence in support of your invention. Furthermore, the best evidence and arguments should be presented sooner rather than later during prosecution.

Consideration of what evidence to present against a finding of obviousness should be done in concert with the attorney's arguments and overall strategy. Evidence is normally presented into the prosecution record by submitting published articles, affidavits or declarations, which are testimonial evidence and/or testimonial presentation of other evidence (e.g., unpublished results; **Box 2**), and/or by visiting the USPTO and meeting with the examiner (**Box 3**).

Upon presenting evidence and arguments against the obviousness rejection, the examiner must consider all evidence anew and determine

Sherry L. Murphy is an associate and Kenneth D. Sibley is a shareholder at Myers Bigel Sibley & Sajovec, 4140 Parklake Ave, Suite 600, Raleigh, NC 27612, USA.
e-mail: smurphy@myersbigel.com

Box 1 The ugly 11: recent decisions on obviousness of biotech inventions

Recently in *KSR v. Teleflex*[8], the US Supreme Court struck down a patent on adjustable gas and brake pedals for vehicles. In doing so, the court declared a more flexible test of obviousness, which now makes a finding of obviousness easier to accomplish across technological fields.

As a testament to this trend, in February 2008, Bruce Kisliuk, a director of US Patent and Trademark Office Technology Center 1600 (Biotechnology and Biochemistry), listed 11 recent decisions, 10 of which were rendered in 2007, that examiners were to refer to for their determination of obviousness. These decisions tally as follows:

Won (not obvious)

Takeda v. Alphapharm: for ACTOS thiazolidinedione, used to control blood sugar in patients with type 2 diabetes, no motivation was found to select this particular compound as a lead compound because related literature mentioned unwanted side effects and because of unexpected results of nontoxicity.

Forest Labs, Inc. v. Ivax: for Lexapro selective serotonin reuptake inhibitor, used to treat depression, no motivation or reasonable expectation of success was found to resolve a racemate of citalopram.

Lost (obvious)

Pharmastem v. Viacell: treatments using stem cells from umbilical cord blood for hematopoietic reconstitution were found to be merely confirming what was already expected in the literature.

Pfizer v. Apotex: for Norvasc, a high blood pressure treatment, motivation was found, given the problems faced, to select the anion from a limited list of FDA-approved anions to form the pharmaceutically acceptable salt.

McNeil-PPC v. Perrigo: for Pepcid Complete antacid, motivation was found in the art to use impermeable coating on the antacid to make it more palatable.

In re Omeprazole: for Prilosec OTC, a heartburn treatment, the court found it obvious to substitute one active alkaline-reactive compound for another.

Ex parte Kubin: a sequence of polynucleotides encoding NAIL polypeptides was found obvious in view of the known amino acid sequence and given the state of the art at the time of invention.

Daiichi Sankyo v. Apotex: treating bacterial ear infection with topical administration of the antibiotic ofloxacin was found obvious in view of a similar antibiotic used to treat middle ear infection.

Aventis v. Lupin: for ALTACE, a high blood pressure treatment, the purified stereoisomer was found obvious, predictable and separable by conventional methods. (Compare to *Forest Labs.*)

Syngenta Seeds v. Monsanto: for a transgenic corn plant that produces an insecticidal protein, it was found obvious to substitute codons having higher guanine-cytosine content in order to create plant-preferred codons.

On remand (to be determined)

In re Sullivan: for antivenom used to treat rattlesnake bites, it was remanded to the US Patent and Trademark Office because rebuttal evidence submitted by the applicant must be considered on the record.

obviousness based on the entire record. The examiner may thereafter maintain the rejection, withdraw the rejection or issue a new rejection. However, the examiner should clearly state his or her findings of fact, both to allow an opportunity to challenge those findings and to build a clear record[3].

Inventor insight

How can you help as an inventor? Prepare for the rejection early. Thoroughly search the literature and discuss possible arguments with the attorney. Tell the attorney any information that may be useful in attacking an obviousness rejection, such as your reasoning or other evidence that one trained in the field would come to a different conclusion after consideration of the references used by the examiner.

For example, in the case of *Forest Laboratories, Inc. v. Ivax*[4], the US Court of Appeals for the Federal Circuit, in considering whether it would have been obvious to resolve the positive enantiomer compound found in citalopram (named escitalopram), examined the state of the science at the time of invention and found that a person with ordinary skill in the art would generally have been motivated to develop new compounds rather than undertake the difficult and unpredictable task of resolving a known racemate. The court also found that a person of ordinary skill would have had no reasonable expectation of success in resolving the racemate, given the relatively new and unpredictable technique of high performance liquid chromatography at the time of the invention and evidence of failed attempts to purify the citalopram racemate at that time (in the mid-1980s).

The *Forest Labs* case highlights the importance of painting a picture of the state of the art when the invention was made. One commentator has likened the inventive moment to finding just the right needles in the haystack[5]. If the roadmap to those needles was available after the fact (hindsight), it may be easy to forget the massive haystack that had enveloped those needles before the roadmap was known.

Along these lines, good record keeping is important. You should save the praise the invention received after it was unveiled. Publications in top journals and overall praise from peers are good forms of evidence against obviousness. Also save the comments of the peers who doubted your predictions. One of the best weapons against a conclusion of obviousness could be a rejected grant proposal with a comment from a peer stating the invention won't work.

You, the inventor, can play devil's advocate too: if you received your invention as a grant proposal, what criticisms might you make? How would you back up those criticisms? This thought process can be extremely effective in generating the type of evidence your attorney needs to attack the rejection.

Consideration of commercial impact, to the extent that the commercial impact can be attributed to the invention (and not merely to aggressive marketing), may also lead to evidence in the form of unexpected results or technological advantages not previously appreciated by peers.

Whenever possible, evidence should be presented with arguments that are clear, succinct and easily understandable. Technical jargon should be avoided. Though examiners are

Box 2 Interviewing with the examiner

At some point during prosecution, you might take a trip to the US Patent and Trademark Office (USPTO) with your attorney and talk to the examiner face-to-face. This allows the inventor to tell the examiner the story behind the discovery in person. Alternatively, a video conference or telephone interview may be conducted.

The key to success in interviewing with the examiner is to do the interview when the time is right. The time is *not* right, for example, when there are many rejections or if they are better dealt with by correspondence.

Keep the following points in mind when interviewing at the USPTO:

- Be prepared; you have a fixed period of time.
- Bring only those items that are essential, including people. Usually that means you, your attorney and sometimes one or two others.
- Treat the examiner as a partner, even when things are not going your way. If you become too adversarial with the examiner, your attorney may need to ‘switch sides’ and defend the overall process to get things back on track.
- Stay in tune with the overall progress of the interview and whose turn it is to speak. At some point, the attorney needs to be quiet and let you tell the story of your invention. The attorney steps in on legal issues.
- Do not expect immediate gratification. The attorney and examiner may focus on reducing the substantive portion of the meeting to writing for the prosecution record. In addition, the examiner may want to consider the arguments further, do more research and discuss the case with supervisors. This is okay; your goal is to educate the examiner, not achieve a hasty win.

technically trained, they have only a limited period of time to become acquainted with the technical details of your particular invention. Even more importantly, in litigating the issued patent, the validity decision falls on judges, who, for the most part, are not technically trained. According to Judge Arthur M. Smith, “This is a challenge which can be met only by very clear writing addressed to this ‘non-technical’ audience”[6].

A stitch in time saves nine

If this all sounds hard, there are things that can be done to make it easier. You can build a plan of response to the obviousness rejection into the application at the time of initial filing—a poorly or hastily written application cannot be fixed later!

Have in-depth discussions with your attorney to prepare a comprehensive and accurate patent application that will explain the invention in the proper context and aid in attacking an obviousness rejection during prosecution. For example, the application may detail the general state of related technology at the time of filing and how the invention is unique. The application should give the reader an accurate picture of the haystack of ideas you faced at the time your invention was developed. This picture not only will set the stage for the presentation of the invention to the examiner but also will serve to refresh your and your attorney’s memory when faced with an obviousness rejection some years later.

Keep in mind that the patent application “constitute[s] one of the most difficult legal instruments to draw with accuracy”[7]. Many other (less meritorious) patent applications are written so that even an experienced patent attorney is left to wonder, ‘What is the invention, anyway?’[6]. Careful research and preparation will make your well-drafted application stand out on the examiner’s desk at the outset of prosecution.

Conclusion—inventors taking action

Inventor input is crucial to overcome the obviousness hurdle during patent prosecution. When faced with a rejection based on obviousness, it is important for an inventor to consider and discuss with an attorney the available evidence that may be used to attack the rejection. In view of the available evidence, the inventor and attorney should review the examiner’s stated reasoning behind the rejection and point out flaws in that reasoning. Inventors should also participate in drafting and editing any prepared declaration and be prepared to speak directly with the examiner in an interview.

Active participation by the inventor in attacking an obviousness rejection not only will aid in procuring the patent but also will build a strong prosecution administrative history for a patent that may later be litigated.

Box 3 Rule 132 declarations—laying a proper evidentiary foundation

A declaration is testimonial evidence, typically from an inventor but often from other qualified experts or witnesses. As such, a proper ‘evidentiary foundation’ should be laid to qualify the evidence as reliable. For instance, who is speaking? Why is the speaker qualified to be making these statements? Are there any possible biases of this speaker that should be taken into consideration?

Not to say that a patent examiner is going to apply the intricate Federal Rules of Evidence in his consideration of a declaration. However, the same sort of evidential foundations are needed to explain to the examiner why he or she should consider and trust the information stated within.

Therefore, the declaration should begin by identifying in detail *who* is speaking and possible biases of which the examiner should be aware. Next, *what* is being discussed (journal articles, experimental results or other documentary evidence) is offered and properly identified in terms of *when*, *where*, *how* and *why*[9]. According to Edward Imwinkelried, a noted writer on legal advocacy, “the testimony therefore covers five topics: the witness’s qualification as an expert, the general theory, the facts of the case, the opinion and the explanation of the opinion”[9].

Careful review of the declaration for content and accuracy cannot be overstated. Be truthful and honest, and avoid selective presentation of data. Remember, if the patent is litigated, you will be cross-examined on your statements by an experienced attorney who is being paid a lot to make you look bad.

1. Federico, P.J. in *Nonobviousness—The Ultimate Condition of Patentability* (ed. John F. Witherspoon) Part 1 101–111 (Bureau of National Affairs, Inc., Washington, DC, 1980).
2. Lupo, R.V. in *Nonobviousness—The Ultimate Condition of Patentability* (ed. John F. Witherspoon) Part 4 201–218 (Bureau of National Affairs, Inc., Washington, DC, 1980).
3. Rollins, A.D. *J. Pat. & Trademark Off. Soc.* **70**, 403–407 (1988).
4. 501 F.3d 1263 (Fed. Cir. 2007).
5. Arnold, T. & Nation, F.R. in *Nonobviousness—The Ultimate Condition of Patentability* (ed. John F. Witherspoon) Part 4 1–44 (Bureau of National Affairs, Inc., Washington, DC, 1980).
6. Smith, A.M. & J. Pat. Off. Soc. **41**, 24–25 (1959).
7. *Sperry v. State of Florida*, 137 USPQ 578, 580 (S. Ct. 1963) (quoting *Topliff v. Topliff*, 145 U.S. 156, 171).
8. 550 U.S. 398 (2007).
9. Imwinkelried, E.J. in *Evidentiary Foundations*, 5th edn. 352 (LexisNexus, New York, 2002).

The next generation

To the Editor:

nature
biotechnology
Finger prick to protein detection on a chip
One-stop MS data analysis
Genomic changes in human ES cells

Your editorial in the December issue[1] argues that the education of the next generation of biotechnologists should include active development and cultivation of entrepreneurial skills. It is suggested that while the success of early biotech breakthroughs has seen "many academic institutions set up teaching programs to capture the rapid advances being made in recombinant technology," the majority of these programs have "largely ignored the mysteries of commercialization". I believe this is true.

Most educational programs in science, particularly those within academia, tend not focus on how to relate work done in the laboratory to the 'real' (commercial) world outside those walls. What I would argue, however, is that any shift in education to include the cultivation of entrepreneurial skills should be accompanied by an equal emphasis on the development of programs, courses and exercises in how to communicate with the public and reflect on potential social and ethical aspects of the work in question.

These skills represent a vital element of what it takes to achieve commercial success in today's post-genetic-modification (GM)-controversy world. This is already being recognized in the emerging field of nanotechnology, in which new educational programs (be they at a high school, bachelor or postgraduate level) are including information and activities relating to social dimensions and ethical questions around the science. In nanotechnology, this emphasis on the importance of scientists being aware of and engaging with these types of issues is said to be based on 'learning the lessons' of what happened with biotech, specifically the controversy surrounding GM crops. The key idea here is that commercial success is not only about what you can do, but also about what society thinks you should do. For the next generation of biotechnologists to be educated as successful entrepreneurs, I would argue that they, too, need to learn the lessons from controversies in their field and find ways to incorporate the cultivation of skills in social and ethical reflection into their education. Without this, they run the serious risk of their products lacking one of the most crucial elements for success, that of social robustness.

Fern Wickson

Centre for the Study of the Sciences and the Humanities, University of Bergen, PO Box 7805, 5020 Bergen, Norway.
e-mail: Fern.Wickson@svt.uib.no

1. Anonymous. *Nat. Biotechnol.* **26**, 1313 (2008).

To the Editor:

Your December editorial calls on the old guard of biotech to devote more energy and time to developing the upcoming generation of young entrepreneurs[1]. You note that this demographic group is characterized by a revolutionary streak—not only showing "openness to new ideas," but also a persistence in pursuing them. To shed some more light on the views of the next generation, I present below the results of an informal survey of entrepreneurially orientated students that asked three questions: What motivates them to pursue a career in biotech? What do they identify as the major opportunities and challenges in biotech? And how willing would they be to take on a job in a biotech startup in the current financial environment?

A standardized open-ended questionnaire was e-mailed directly to students from 16 major biotech clubs across North America during December (**Box 1**). The universities were selected arbitrarily based on accessibility to student e-mails from preexisting contact via extracurricular involvement. A total of 703 individuals were e-mailed with 161 (23%) from all 16 institutions (**Supplementary Table 1** online) responding in full to the e-mail. They ranged from 18 to 27 years of age across undergraduate (22%), doctoral (51%), medical (16%) and MBA (11%) student populations. Incomplete or incoherent responses were discarded from the results. There was little variability in the responses according to geography, although the sample size was not sufficiently large to enable differentiation (**Fig. 1**).

When we asked students what motivated them to pursue a career in biotech, the top reason they gave was the opportunity to help others. "I want to be able to cure a million patients at the same time" was a common refrain among the respondents. This provides direct evidence for the "Yes we can" philosophy that you attributed to this generation in your editorial. The second and third most cited reasons, as expected, were intellectual stimulation and monetary incentives (**Supplementary Data** online).

The second question regarding major opportunities and challenges for biotech yielded answers across the spectrum. In response, students cited such goals as the potential of personalized medicine to transform healthcare or for environmental

Box 1 Questionnaire to upcoming biotech entrepreneurs

1. Age?
2. University affiliation?
3. Program of Study (MBA, Medical, Doctoral, Undergraduate)?
4. What motivated you to choose biotechnology as a career pathway?
5. What are the major opportunities in biotechnology today?
6. What are the major challenges or obstacles to pursuing biotechnology today?
7. Would you join a biotech startup in today's difficult financial environment?

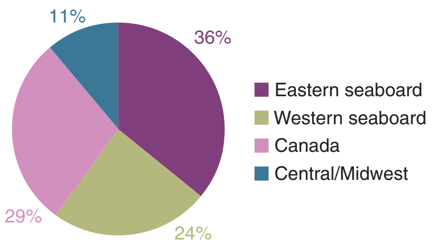


Figure 1 Geographical distribution of survey respondents. A significant proportion of the respondents were from Canada because the magazine from which the e-mails were drawn was based in Canada. Nevertheless, the number of respondents from each region corresponded roughly to the strength of the region's biotech industry, as would be expected. Students from biotech-intensive regions, such as Massachusetts or California, were less likely to be concerned about job uncertainty than students from other regions.

biotechnology, particularly synthetic biology, to alleviate the impact of climate change—all rote answers typically touted by the biotech industry. At the same time, however, students also provided some more surprising answers.

On one side of the spectrum, students mentioned global health. When asked to elaborate, they mentioned their desire to get involved at some point in nonprofit organizations, such as the Institute for OneWorldHealth, which invests in research targeting neglected diseases ignored by multinational pharmaceuticals. Others, especially among MBA students, cited the example of Endeavor, a nonprofit recently covered in the *The Economist*, which aims to cultivate entrepreneurship in emerging economies[2]. The program offers the opportunity for MBA students to intern at biotech companies in Latin America or South Africa and gain experience in navigating an immature regulatory system and market. This interest by MBA students from top-flight schools is promising, given the potential for such biotech companies in emerging economies not only to create jobs, but to innovate novel, affordable solutions to local health problems and create self-sustaining economic cycles[3].

On the other end of the spectrum, undergraduate and medical students specifically, brought up the topic of cognitive enhancers. This response may have been primed by a recent widely read article in *Nature* that argued in favor of cognitive enhancers[4]. When asked to elaborate in informal follow-up e-mails, students cited the shifting demographics toward a rapidly aging population that is likely to suffer from psychiatric illnesses, which creates a tremendous opportunity to simultaneously target a large market and improve the quality of life of a significant proportion of the population.

One of the interesting demographics that emerged was the relatively high number of medical students involved in student biotech clubs. Many older doctors hold significant reservations about collaborations with industry and commercialization[5]; perhaps the "openness to new ideas" cited by your editorial is evident in this younger generation of doctors.

The final part of the questionnaire attempted to assess the level of risk taking of this current generation of upcoming biotech entrepreneurs. First, students were asked to identify the factors in the biotech industry that most concerned them. Lack of early-stage financing or venture capital support was the number one reason given followed by the uncertainty in the job market due to the financial crisis. Interestingly respondents that were younger and based in larger biotech hubs, such as San Francisco or Boston, cited the second concern regarding job uncertainty much less frequently than older students based in smaller hubs. This may reflect a combination of a more risk-taking, entrepreneurial culture in the large hubs or the less risky nature of joining a startup in an area that already has hundreds of other startups right next door.

Yet despite these reservations, when we directly asked students if they would consider joining a biotech startup in this financial environment if the opportunity presented itself, an overwhelming majority (67%) said yes. Clearly, those in the sampled group are willing to take on the risks associated with startups necessary to see their mission through to completion. Accordingly, as it seems to attract the best talent, industry should devote resources and partnerships with nonprofits tackling the big issues like neglected diseases. For example, offering sabbaticals to scientists to work at such institutions or companies in emerging economies could benefit for-profit companies by creating knowledge exchange, while renewing staff morale by exposing them to new ideas and work with big impact.

All of the above suggests that companies should prioritize efforts to maintain a spirit of idealism and entrepreneurship important to the next generation, allowing new recruits to engage in high-risk endeavors similar to Google's 20% free-time rule for its engineers. Finally, the most important thing industry can do is to advocate on behalf of the interests of the coming generation by tackling the controversial regulation necessary for drugs like cognitive enhancers head on and advocating legislation supportive of early-stage financing that funds high-risk ideas.

ACKNOWLEDGEMENTS
We would like to thank students who participated from Yale University, Harvard University, Stanford University, University of California at San Diego, Tufts University, Northwestern University, University of Washington, Caltech, University of British Columbia, Columbia University, Darmouth College, University of North Carolina, Chapel-Hill, University of Michigan, Duke University, University of Ottawa and the University of Toronto.

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

Justin Chakma

BioSynergy Magazine, 761 Bay Street, Suite 2406, Toronto, Ontario, MRG 2R2, Canada and the McLaughlin-Rotman Centre for Global Health, University Health Network and University of Toronto, MaRS Centre, South Tower Suite 406, 101 College Street, Toronto, Ontario, M5G 1L7, Canada.
e-mail: justin.chakma@utoronto.ca

1. Anonymous. *Nat. Biotechnol.* **26**, 1313 (2008).
2. Anonymous. Spreading the Gospel. *The Economist* (July 31 2008).
3. Singer, P.A. *et al. Nature* **449**, 163 (2007).
4. Greely, H. *et al. Nature* **456**, 702–705 (2008).
5. Blumenthal, D. *N. Engl. J. Med.* **349**, 2452–2459 (2003).

Gunvalson and PTC Therapeutics' community outreach

To the Editor:
We thought your news story in the November issue "Gunvalson decision sends shockwaves though industry"[1] was a balanced summary of the case and the broader implications for the clinical trial process.

However, one important point that is omitted from the article is the support PTC has received from patient advocacy groups and parents in the case. The Parent Project Muscular Dystrophy (PPMD), a US-based nonprofit organization founded and run by parents of children with muscular dystrophy, joined by its international counterpart, United Parent Projects Muscular Dystrophy, filed an *amicus curiae* brief with the appellate court in support of PTC's position on the appeal. In addition, an *amicus curiae* brief supporting PTC was filed by the family of a boy with muscular dystrophy. This support from the patient community results from PTC's long-standing efforts to engage clinicians, regulators and patient advocates in the development of PTC124, in the belief that an open and direct approach to patient communications is in the best interests of all stakeholders. It is this openness that is threatened by the lower court's decision.

Note added in proof: On December 16, 2008, in a 3–0 decision, the Third Circuit Court of Appeals vacated the lower court's order, finding that there was no clear promise made to the Gunvalsons and thus no basis for their reliance. This is an important ruling for all companies conducting clinical trials in the area of rare or orphan diseases.

COMPETING INTERESTS STATEMENT

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

Stuart W Peltz

PTC Therapeutics, 100 Corporate Court, South Plainfield, New Jersey 07080, USA.
e-mail: speltz@ptcbio.com

1. Allison, M. *Nat. Biotechnol.* **26**, 1201–1202 (2009).

DNA sequence patents are not in the grave yet

To the Editor:

Some DNA sequence patent holders may be feeling like Mark Twain when he read his premature obituary. We believe the patent article by Miles Yamanaka[1] in the October issue entitled "A nail in the coffin of DNA sequence patents?" is unduly alarmist. The headline and final sentence both imply that the decision by the Board of Patent Appeals and Interferences (BPAI) on the patent application of Kubin and Goodwin (application no. 09/667,859) threatens all DNA sequence patents. This is misleading because it is overly broad. In *Kubin*, the BPAI does suggest a higher standard for nonobviousness[2], a criterion that the US Court of Appeals for the Federal Circuit unwisely rendered largely inoperable for DNA sequence patents in its 1995 *Deuel* decision[3]. In the eyes of most analysts, *Kubin* is a sensible corrective. But, even assuming that the Federal Circuit goes along with the BPAI's reasoning, precisely how *Kubin* will affect DNA patents as a whole is hardly clear.

nature
biotechnology
New paths to iPS cells
Nanotube tags for protein arrays
Anthocyanin-enriched tomatoes

Read narrowly, the BPAI decision precludes only claims on DNA sequence based on prior characterization of a protein's amino acid sequence. On that reading, *Kubin* merely captures a judgment that deriving a nucleic acid sequence from a corresponding amino acid sequence is straightforward to those with ordinary skill in the art, despite some degeneracy of the genetic code. (Yamanaka acknowledges the possibility of this narrow reading when he states "the *Kubin* decision will make it harder to obtain claims to a polynucleotide encoding a protein *when that encoded protein is already known*" [emphasis added].) But claims to DNA sequence derived from amino acid sequence are mainly confined to some 'first generation' gene patents based on cloning genes for known proteins. Most DNA sequence patents that we study in our work, for example, are not based on prior characterization of a protein, but start from a genetic discovery or DNA sequence variation.

Even if *Kubin* is read more broadly, to render invalid all composition of matter claims to DNA sequence patents where the procedure for finding the sequence is obvious to the ordinary genomic scientist, the case should not affect claims to inventions identified by procedures that are not obvious at the time of patent application. *Kubin* does not call into question patents on DNA sequences that arise from genuine invention; rather it corrects the anomalously low threshold for nonobviousness established by *Deuel*. *Kubin* is not a "nail in the coffin of DNA sequence patents," but rather a mechanism for culling marginal patents based on an accurate reading of the state of the science.

Robert Cook-Deegan[1] & Arti K Rai[2]

[1]Center for Genome Ethics, Law & Policy, Institute for Genome Sciences & Policy, Duke University, Box 90141, Durham, North Carolina 27708, USA. [2]Elvin R. Latty Professor of Law, Duke Law School, Science Drive and Towerview Road, Durham, North Carolina 27708, USA.
e-mail: bob.cd@duke.edu

1. Yamanaka, M. *Nat. Biotechnol.* **26**, 1085–1086 (2008).
2. *Ex parte Kubin*, 83 USPQ2d 1410 (Bd. Pat. App. & Int. 2007).
3. *In re Deuel*, 51 F.3d 1552 (Fed. Cir. 1995).

A lifeline for the biotech sector

Mark Kessel

With many small biotech companies teetering at the edge of a financial precipice, the US government should act swiftly to enact tax benefits allowing a refund of net operating losses.

As the biotech industry heads into 2009, it is facing strong negative headwinds. Layoffs, the shelving of promising drug development programs and bankruptcies are continuing at an alarming pace. The consensus is that those investors still able to invest in biotech will be focused mainly on biotech companies that are profitable, nearing profitability or have at least 18 to 24 months of cash on hand. Furthermore, to preserve cash, biotech companies will be reluctant to spend cash on anything but their most advanced program. Thus, those companies with the greatest need for capital are likely to face a liquidity crisis without realistic access to capital or credit, and many companies will be forced to put earlier-stage programs on hold. Several financial stimulus initiatives have already been proposed to the US Congress, such as a reduction of capital gains tax on invested funds or tax credits set against research. But there is another form of financial assistance, refunds on net operating losses (NOLs), that would not only be relatively easy to implement but also target those companies in most need of financial life support.

© Roy Morsch/CORBIS

Could a NOL refund provide a lifeline to the small-cap biotechs running out of cash?

How big is the problem?

The financial meltdown is threatening the lives of many small biotech companies—the very same companies that are the future to developing life-enhancing drugs. The signs of impending financial catastrophe are already apparent:

About 30% of biotech companies trading in the capital markets have less than six months of cash to fund operations—more if one looks at smaller-cap firms (**Fig. 1**).

Total capital raised by both public and private companies fell by 55% in 2008 compared with 2007 (see p. 114)

Only one biotech initial public offering (IPO) took place in the United States in 2008 and it raised less than $6 million compared to 41 IPOs in 2007 that raised nearly $2 billion.

If the new US administration is serious about the 'innovation economy' that it has been espousing, then it should care about maintaining the United States as the undisputed global leader in biotech. Allowing small, emerging biotech companies to fail will result in the loss of essential expertise needed to continue product development programs. Even big pharma recognizes that waiting to pick up biotech products in a fire sale would be counterproductive because it needs not only the products but also the external pool of R&D expertise to augment its anemic drug pipelines. This is essential for pharmaceutical companies to replace products coming off patent and also in the longer term to augment their own underproductive drug development efforts.

A solution

If there is one financial characteristic common to all small biotech companies, it is enormous NOLs. Given the huge expenditures and length of time that it takes before a biotech company can derive meaningful revenues from its pipeline, it should come as no surprise that most biotech companies fail to turn a profit for extended periods of time, if at all. So why not devise a means for small biotech companies to avail themselves of the one asset that they all have?

In general terms, under US tax regulations, accumulated and unused losses from any particular year, known as NOLs, are allowed to offset income in future years, thereby resulting in a reduction of taxes during those years that the NOLs are still utilizable. But under the present system, biotech companies must wait to become profitable before they can avail themselves of this benefit.

By simply allowing biotech companies to elect to receive a one-time refund of their accumulated NOLs at some discounted rate

Mark Kessel is at Symphony Capital LLC, 875 Third Avenue, 18th floor, New York, New York 10022, USA.
e-mail: mark@symphonycapital.com

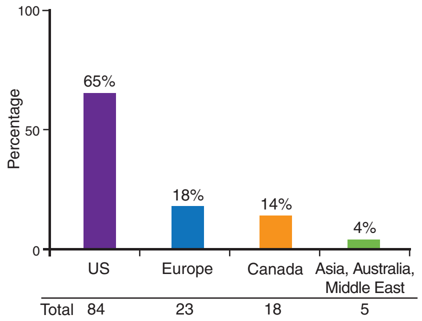


Figure 1 A greater proportion of small biotech companies in the United States are in dire financial straits than in other parts of the world. Figure shows number of small biotech companies (market caps less than $250 million) in a particular region that have less than a year's cash as a percentage of total small biotech companies in that region. Source: *Nature Biotechnology*

(rather than allowing them to utilize NOLs when they become profitable), the US government would provide tax relief when it is most needed. The proposal need not be complicated, and to ensure it benefits the right companies, it should contain the following provisions.

Qualifying companies would elect to receive a one-time refund of accrued NOLs at a significant discount in lieu of claiming qualified research expenses for the applicable tax year;

- these refunds would have to be reinvested in US-based R&D development;
- any NOLs used in the determination of the amount of the refund would be permanently extinguished;
- and only companies up to a certain size and with losses would be eligible.

The proposal—foregoing a larger tax benefit in the future to receive a smaller tax benefit today—should be attractive for small-cap biotechs struggling for survival in these dismal capital markets.

Why government needs to act

What is at stake if nothing is done by the new administration and US Congress to assist the biotech industry? Already, US biotech companies are forced to preserve cash to ride out this perfect storm. There is a vicious cycle facing these companies today—to get funding from currently available sources, a biotech company needs to have sufficient cash to fund its operations for an extended period of time. Thus, many of these companies are shelving life-changing R&D programs for new treatments to extend their corporate longevity. Is this in the interest of the American populace?

The Obama administration also should not overlook the jobs that are at stake. Direct employment in the bioscience industry exceeds 1.3 million. But this is not the only job impact as there are an additional 6.2 million jobs related to this industry. This aggregate of approximately 7.5 million employees is more than double that of the comparable figures for the automotive industry that, unlike the bioscience industry, is no longer a global leader and is likely to continue to decline further.

The drugs being created by the biotech industry account for the vast majority of the new important treatments approved by the US Food and Drug Administration (FDA). In 2008 alone, of the 24 drugs that were approved by the FDA, the majority came out of the pipeline of the biotech industry. Societal value associated with gains in life expectancy aside, life-extending drugs creates enormous leverage in national wealth. It has been reported that from 1970 to 2000, gains in life expectancy added about $3.2 trillion per year to national wealth, with half of these gains due to advances against heart disease alone. Looking into the future, a permanent 1% reduction in mortality just from cancer has a present value to current and future generations of Americans of nearly $500 billion and a cure would be worth approximately $50 trillion.

A win-win

Unlike the other bailouts handed out by the US government, the proposal from the point of view of the US taxpayers has minimal treasury revenue impact as companies that avail themselves of this tax benefit would only be claiming NOLs at a substantial discount in return for foregoing the ability to claim the full NOLs for future tax years. Also, the loss of tax revenues will pale by comparison to the funds granted to other industries that do not have the biotech industry's societal benefits and job growth potential.

If the new administration and the US Congress recognize the unprecedented threat to the health of the biotech sector and the benefits to be derived from a refund of NOLs, the legislation should not be difficult to enact. At the same time, the cost of inaction is equally clear: unless legislation is forthcoming quickly, many smaller, emerging biotech companies will lay off employees, postpone or abandon cutting-edge R&D programs or, worse, face extinction. Can the US government simply sit by and watch the innovative and entrepreneurial heart of one of this country's most successful industrial sectors face financial oblivion?

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/

Charting a course through a perfect storm

Arthur Klausner

A venture capitalist gives his perspective on the outlook for life sciences ventures amid the perfect storm of the current economic downturn.

Make no mistake. What was originally thought to be just a 'Wall Street problem', which then spilled over onto Main Street, is now sending aftershocks to the epicenter of venture capital's world—Silicon Valley's Sand Hill Road. Sequoia Capital's (Menlo Park, CA, USA) now infamous Slide Presentation of Doom[1] proclaiming "R.I.P. Good Times" underlines the seriousness of the situation. There is no longer any question that life sciences venture capitalists (VCs) are in for a tough ride along with everybody else. So, given the current pressures on the venture capital community, what is the outlook for the present and next generation of life sciences startups?

©Jordi Elias/Illustration Works/Corbis

Private biotechs need to be particularly strong swimmers to avoid being sucked under during the current turbulent times.

Descent into the maelstrom

I often get the feeling that entrepreneurs view VCs as being at the top of some sort of alternative-universe food chain. These somewhat mysterious (but usually nattily dressed) investors fly in to board meetings—perhaps somewhat less often via private jets than before—opine on key issues facing the company while simultaneously keeping opposable thumbs glued to their Blackberries and then zip off to their next oh-so-important assignations.

What some operating executives fail to keep in mind, however, is that virtually all VCs in fact invest other people's money. As such, we answer to the stewards of those sources of capital—the people who run pension funds, university endowments, fund-of-funds and the like. VCs are thus under pressure today because our investors are under pressure. In the parlance of money managers, venture capital resides in the 'alternative asset' sector, along with other private equity, leveraged buyouts, hedge funds and real estate. Not surprisingly, the current economic environment has decreased the overall risk tolerance of these broad-based investors, and as such they are not exactly clamoring to increase their exposure to alternative assets in general or to potentially risky venture capital in particular.

But these investors have another problem beyond woeful 12-month rates of return: the 'denominator problem'. Say there was a pension targeting a 10% target allocation for alternative assets. Further assume that this group was appropriately allocated at the beginning of 2008, and that their public stock portfolio tanked by a typical 30–40% during the just-ended brutal year. Unless the valuation of their venture capital holdings declined accordingly (which is quite unlikely because these valuation changes tend to lag somewhat), they now find themselves significantly overallocated to venture capital on a percentage-of-assets basis. Although some such investors have been frantically rewriting their rules to allow increased variance around their allocation targets, this imbalance has both short- and medium-term effects on the VC funds in which these groups normally invest.

Taking the medium-term outlook first, limited-partner investors will have to think especially long and hard about backing future venture funds, even those sponsored by VCs they already know and love. Thus, it's pretty straightforward to project a significant medium-term contraction of the venture industry as a whole.

Of more near-term impact (and what many entrepreneurs don't realize) is that when a VC announces the closing of a new fund, we do not receive all of that precious capital up front. Rather, we draw down the committed funds on an as-needed basis to make new and follow-on investments in portfolio companies and to cover the management fee that pays everybody's salary and keeps the proverbial lights on in the office. This arrangement makes good sense, as VCs are trying to generate return on investment in the healthy double-digits, whereas the risk-free interest rate associated with any capital left sitting in the bank could seriously drag down overall performance.

Now, however, even previously reliable investors into venture capital funds may have trouble meeting their future funding requirements. Although reports of actual defaults on capital calls have thus far been rather scarce (with the collapsed Washington Mutual bank being a particularly high-profile exception), limited partners are for the first time requesting detailed projections of distributions from venture capital funds as well as schedules of anticipated capital calls. Some limited partners

Arthur Klausner is at Pappas Ventures,
2520 Meridian Parkway, Suite 400, Durham,
North Carolina 27713, USA.
e-mail: aklausner@pappasventures.com.

have also asked for additional time beyond the traditional two-week notice period on capital calls. Others (with Harvard's endowment and CalPERS being mentioned the most) have sold or considered selling major portions of their venture capital fund holdings (along with their associated future funding requirements).

The last time this type of situation took place—but on a much smaller scale—was directly after the internet bubble burst in the early 2000s. Numerous entrepreneurs who had become paper millionaires many times over via the imputed value of their publicly traded shareholdings had chosen to become limited-partner investors in tech-focused venture capital funds as a way of further leveraging their gains. A $5-million commitment into a particular venture capital fund might have made perfect sense when the entrepreneur's net worth was over $100 million. But if that individual's stock subsequently declined from something over $100 per share to a number that rounded to zero, the future capital calls on that VC commitment quickly morphed into something between onerous and impossible.

Fellow limited-partners reacted in two diametrically opposed ways to the distress being felt by defaulting investors. One group, clearly moved by the 'there-but-for-the-grace-of-God-go-I' argument, urged leniency. By contrast, other limited partners took the 'I-came-up-with-my-money-so-everybody-else-had-better-do-the-same' position and demanded that harsh penalties be imposed upon non-performing investors. The VCs themselves were caught in the middle, of course, wanting to keep all their investors (reasonably) happy while at the same time bringing in as much of the originally committed capital as possible. Turning back to the current situation, it will be uncomfortably interesting to see how things play out when it is major institutional investors that are having trouble meeting capital calls.

Finally, to top off today's perfect storm, there is the possibility of the US capital gains tax increasing from 15% to 20%. Although the venture capital industry may indeed be woefully short on new capital gains these days, this change would further serve to render this asset class slightly less attractive than before.

Keeping portfolio companies afloat

The Biotechnology Industry Organization estimates that almost 50% of small-cap biotech companies hold less than a year's worth of cash, so it's abundantly clear that many public firms are now or will soon be seeking access to capital. In the past when things started to look really bleak for the finances of the life sciences sector, the public markets miraculously rebounded just in time to save the industry—but there is little indication that such a resurgence is forthcoming this time around. Unfortunately, the picture isn't a whole lot prettier on the private side of the industry either.

Sources of capital. It's no secret that initial public offerings (IPOs) have been dormant for well over a year now, and those optimistic companies that had filed to go public are one by one giving up the ghost. Estimates regarding a potential return of the IPO market seem to center around late-2009/early-2010, but the scary fact is that nobody really knows.

Venture debt (which from the outset has always seemed like a bit of an oxymoron) has largely gone up in smoke as well. Only a few such providers are still in the game, including the likes of Silicon Valley Bank, Oxford Finance and Square 1 Bank. Most of the others have either exited the business entirely or are demanding terms and covenants that are onerous enough to effectively discourage any potential business. In addition, the amount of venture debt available to any one particular company is down significantly from even a year ago. As one VC remarked recently, "When it comes to venture debt, $5 million is the new $15 million…."

Corporate collaborations remain attractive, but they are somewhat unpredictable and 'lumpy'. Despite the impressive 'biobuck' totals that are trumpeted in press releases, it's very difficult to ink a mega-deal on command. More insidiously, any roster of announced deals by definition fails to mention all the companies that tried to find a lucrative partnership but ultimately failed to do so.

VCs, as the initial sources of capital, also become the 'funders of last resort' for private (and sometimes public) companies looking to survive choppy financial waters. But venture capital funding is becoming scarcer, pricing tougher and syndicates weaker owing to the pressures described above.

In most life sciences VCs' portfolios, I wager that there are one or more pairs of private companies that fit the following scenario. Both have impressive clinical-stage projects, large potential markets, experienced management teams and enthusiastic investors—and each company needs to raise additional capital within the next 12 months. The backers of the first company have significant capital set aside for future investment. The second company, however, has been around for several years and has come close to exhausting the reserves of its financial backers. If you didn't know this last detail, you might think each company would be worth about the same amount at their next financing. But based on the distressed nature of one of these deals, there could ultimately be as much as a fourfold or more difference in price.

Lest we forget, on a more positive note, the Obama administration in the United States should be good for biomedical research—in particular stem cell research (which has its own difficulties from a commercialization standpoint) and personalized medicine (where there is clearly money to be made, but where and how to do so remains less clear).

Operating issues. The future pricing and political environment does not exactly look like clear sailing either, with several key issues dominating the life sciences horizon. With an incoming US administration that placed healthcare reform front and center during its campaign, look for lower drug prices going forward, but also increased drug volumes as more individuals are insured. Biotech's innovative products may not be hit as hard as the pharmaceutical industry overall, but healthcare reform will certainly be a net negative for drug developers.

There is also the looming specter of biogeneric legislation. Any new rules that allow the equivalent of generic versions of biologic therapeutics are naturally going to decrease the value to innovators of developing these types of products in the first place. In addition, experts have raised real concerns regarding the true therapeutic and safety equivalence of biosimilars. Nevertheless, there is a (debatably) large amount of money that could be saved via biogenerics, so in today's tough economic times it certainly feels like some sort of legislation is on the way to create an alternative approval process for these copycat drugs that doesn't involve full-scale clinical trials.

In terms of regulatory issues, the current attitude at the US Food and Drug Administration (FDA) is as cautious and safety-conscious as professional observers have ever seen it. Despite the uptick in new drug approvals seen in 2008, this ultraconservative posture opens up the possibility that some potentially efficacious drugs could ultimately fail to win approval. What's more, in the past year the agency has taken to missing PDUFA (Prescription Drug User Fee Access) deadlines with alarming consistency. Although, as of this writing, no one yet knows who the permanent FDA commissioner will be, it seems that a history of taking a hard line with industry will be a virtual prerequisite for the new appointee. That said, some degree of (rational) leadership should be significantly better for all partisans than no real leadership at all.

The forecast

Life science ventures are more financially constrained than venture-backed tech companies for one basic reason. If you actually have revenues, then you are likely to have a choice regarding how much you decide to invest to increase your company's growth. But for product-lacking biotech companies, it's virtually impossible to 'bootstrap' when you don't have any sales. Luckily, most of our life sciences portfolio companies' belts have already been pretty tight for a while owing to the lack of an IPO market. So if you're already up and running, all you may really be able to do is try to run even leaner and meaner than before while looking for creative financial solutions.

For life science VCs, the keys to success haven't really changed that much:

- Seek undervalued assets, be they in the form of products, projects or companies (and there are certainly more bargains out there than ever).
- Develop a potential 'fast-to-liquidity' strategy.
- Align with deep-pocketed, like-minded investment syndicate members.

Neither are they so different for startups:

- Understand the needs and desires of your potential acquirers as you design your development programs and potentially even define your exit.
- Stay away from large infrastructure by employing virtual and semi-virtual models to maximize cash efficiency.
- Keep your focus laser sharp (although admittedly this lack of product diversification will tend to increase volatility).

As before, the ultimate key to company success is figuring out a plausible answer to the following troublesome drug development algebra: 90% attrition for compounds entering human clinical trials; drug development costs averaging ~$800 million; and total development timelines approximating 12–15 years. So, at their very heart, new companies need to have individually tailored strategies designed to beat these odds. Specifically, each and every startup looking for funding should be able to describe why their drug development program has a higher than average chance of success (perhaps a proven mechanism of action or a therapeutic area where preclinical models are particularly predictive); will require less money than average (maybe focusing on indications with short, inexpensive clinical trials); or should take less time than average (perhaps some sort of 'repurposing' that involves new potential uses of either previously approved drugs or at least compounds that have already demonstrated safety in clinical trials targeting other therapeutic indications).

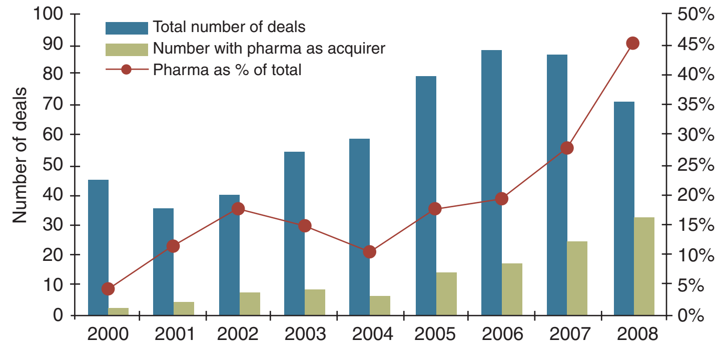


Figure 1 Ebb and flow of biotech mergers & acquisitions. The total number of M&A deals has been increasing over the years (blue bars). Although 2008 saw a bit of a decline overall, big pharma (green bars) has actually continued to pick up their deal pace. Source: *BioCentury* and Piper Jaffray.

Although it's one thing for a life sciences startup to create value, it's quite another for shareholders to actually be paid off for that value. If and when the IPO markets return, going public is likely going to continue to represent just another in a series of financing events for developing companies, rather than a champagne-popping liquidity event for investors. Thus, acquisition by a larger entity—either by pharma or big biotech—is becoming the only real way for investors in the sector to truly succeed.

There is no question that startup companies provide the innovative research that big pharma needs in the face of their questionable R&D efficiency and the specter of its key revenue-driving products continuing to go generic over the next several years. The much-publicized layoffs within the industry have served to further weaken the R&D capability of the major players, and the concept of outsourcing now pervades all aspects of their businesses. And, most importantly, big pharma has huge stockpiles of cash. As shown in **Figure 1**, although we may have seen some leveling off of total biotech merger activity over the past couple of years, big pharma's piece of this critical pie is increasing.

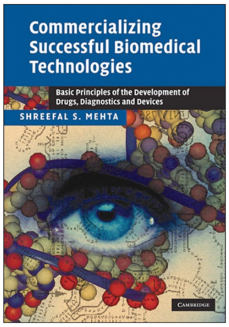
So the final question becomes whether the acquirers of life sciences startups will continue to pay fair value (plus an acquisition premium) to access their next generation of innovative products—or will they as a group show newfound restraint and take advantage of the industry's financing woes to turn the situation into a true buyer's market? Although industry pundits are closely monitoring incoming deal data to determine if prices are in fact eroding, I simply do not believe that predatory pricing will occur on a large-scale basis. The reason is that the marketplace for the best ideas still exists. Pharma companies simply cannot afford to let the most attractive drug development projects end up in the hands of their fiercest competitors, and as such I believe they have no choice but to continue to pry open their corporate wallets for the most promising of these assets.

Conclusions

With less venture money chasing new life sciences companies as we move into 2009, private biotech valuations are naturally going to decline. Only the most attractive startups will be able to attract financing, and long-term investment returns will naturally rise. The mergers and acquisitions market will continue to provide lucrative exits for valuable therapeutic programs. Put simply, 2009/2010 will be terrific times to invest, should you be lucky or skillful enough to have the venture capital funds available to put to work. That said, investment and company-building strategies going forward are going to have to continue to adapt to the difficult overall environment. Strict focus combined with capital efficiency are clearly the watchwords of the next generation of life sciences companies. In the meantime, the financial storm that currently envelops biotech is likely to sink a significant number of poorly resourced companies. But when the streamlined biotech fleet emerges on the other side, it will certainly be stronger for it.

[1] http://www.techcrunch.com/2008/10/10/sequoia-capitals-56-slide-powerpoint-presentation-of-doom/)

The bioentrepreneur's road map


Commercializing Successful Biomedical Technologies
Basic Principles of the Development of Drugs, Diagnostics and Devices
SHREEFAL S. MEHTA
CAMBRIDGE

Commercializing Successful Biomedical Technologies: Basic Principles of the Development of Drugs, Diagnostics and Devices

by Shreefal S Mehta

Cambridge University Press, 2008
360 pages, hardcover, $80.00
ISBN-13: 9780521870986

Reviewed by Michael R Bielski

Bringing regulated biomedical technologies (small molecules, biologics, medical devices, diagnostics and combination products) from the labor atory to the bedside is one of the more complex and resource-intensive journeys an aspiring entrepreneur can embark upon. Unlike other high technologies, which may go from concept to market to obsolescence in a matter of months (computer processors, anyone?), a drug may take 14 years and hundreds of millions of dollars to receive approval by the US Food and Drug Administration (FDA) and ultimately generate revenue. Along the way, numerous issues requiring a variety of skills and talents must be addressed, many of which are not immediately apparent to an inventor/researcher-turned-CEO. In *Commercializing Successful Biomedical Technologies*, biomedical academic and entrepreneur Shreefal Mehta highlights the key issues that must be understood to improve the chance of bringing biomedical technology innovation to market.

Mehta's objective is not simply to highlight the issues, but rather to point out the considerable amount of attention that should go into preparing to address them. Emphasizing this point, he states his hope that "better thinking and planning in the development of regulated products will help improve the efficiency, success and quality of biomedical technology commercialization, increasing the number of innovative products that can be delivered to help people." In other words, taking the time to identify relevant issues and create a comprehensive biomedical technology commercialization plan to mitigate their associated risks before they arise greatly increases the chance of a successful biomedical technology startup.

In practice, creating a sound biomedical technology commercialization plan is a multi-disciplinary endeavor that requires business, legal, technical, regulatory and marketing expertise throughout. In an effort to reduce the complexity associated with synthesizing these disciplines and facilitate the creation of a viable commercialization plan, Mehta develops a seven-stage framework that breaks down the requirements for developing a commercialization plan into its key components: (i) plan (industry context), (ii) position (market research), (iii) patent (intellectual property rights), (iv) product (new product development), (v) pass! (regulatory plan), (vi) production (manufacture) and (vii) profits (reimbursement). This framework allows the aspiring entrepreneur to systematically assess the components of a sound commercialization plan in a way that doesn't allow him or her to lose sight of the big picture, yet still manage the details required for successful commercialization.

Mehta wisely addresses the practical limitations of using a linear road map to organize the iterative and path-dependant process of biomedical product development. He points out, "The linear road map shows the components that must be assessed to build a sound commercialization plan, but the processes are all carried out in parallel, with shifting emphasis on each component as one proceeds down the plan." The text prepares the reader to organize the commercialization process while at the same time pointing out the inevitable fact that feedback from one component will ultimately influence or change the understanding of another previously researched component. For example, limited access to intellectual property rights may change market strategy, which in turn may alter the regulatory pathway required to develop an FDA-approved product.

Throughout, Mehta highlights the significant differences that are specific to biomedical technology commercialization as compared to the rest of the free-market industries in the United States. Most importantly, biomedical technologies are heavily regulated, requiring a significant commercialization planning to account for this issue. Another complexity not often faced by other technology sectors is that the user—that is, the patient—does not make the purchasing decision; it is made instead by a provider such as a physician or pharmacy benefit manager. Adding more complexity to the situation, the government or insurance company is the payer, requiring the entrepreneur to consider reimbursement strategies early on—or otherwise bear the burden of having a tremendous product with nobody willing to prescribe or pay for it. Factor in other issues ranging from what truly is a patentable invention in the life sciences to how to commercially manufacture products in accordance with current good manufacturing practices and you begin to see how much planning must go into bringing biomedical technologies to market.

As in many cases, the best way to allow readers to fully grasp a technology commercialization concept is to apply the concepts covered in the text to real-world technology examples. The text does include excellent case studies and excerpts throughout; however, the reader will crave more—an issue that will hopefully be addressed in subsequent editions and/or by supplementary materials.

Although there are many general technology commercialization materials available to students, researchers and entrepreneurs on the market, the advice given often lacks the context of the biomedical technology space—specifically the regulatory and reimbursement issues that significantly affect the planning required to bring biomedical technologies to market. Often, readers and teachers interested in biomedical technology commercialization are forced to supplement general technology commercialization materials with specific biomedical commercialization materials. Although no single text could possibly provide all the information necessary for an individual to translate biomedical technology research into a biomedical technology product, Mehta does an excellent job of identifying and organizing the major issues associated with biomedical technology commercialization in a framework that students, researchers and entrepreneurs can understand.

Michael R. Bielski is at the Center for Biotechnology, Stony Brook University, Stony Brook, New York 11794, USA.
e-mail: michael.r.bielski@gmail.com

The cancer vaccine roller coaster

Bruce Goldman & Laura DeFrancesco

The cancer vaccine field is littered with promising products that failed to show clinical efficacy. Could it finally be on the verge of a first US approval?

If any field epitomizes the boom and bust cycles of biotech, it would be cancer vaccines. Over the years, numerous tumor immunotherapies have gone through rounds of early-stage successes, only to fail in phase 3 clinical trials. Experts point to many reasons for the failures, from "jumping the gun" before enough was known about the biology or the therapies to letting business considerations—going for low cost and short time lines—trump science; what Peter Bross, chief of clinical evaluations at the US Food and Drug Administration's (FDA's) Center for Biologicals Evaluation and Research calls companies simply not doing their homework. Put these problems together with poorly designed clinical trials of heterogeneous cancer patient populations with late-stage disease, add a lack of familiarity of the regulatory authorities in assessing tumor vaccine products, mix in manufacturing scale-up headaches and the resulting recipe is all but toxic to investors. As Bruce Booth of Atlas Ventures (Waltham, MA, USA) puts it, realizing the potential of cancer vaccines is "full of complexity."

But some researchers and analysts are keeping the faith, hoping that a more comprehensive understanding of tumor immunology will lead the way to more fruitful approaches (**Table 1**). Several promising phase 3 programs are nearing completion, so 2009 may well be the year of the cancer vaccine. "There have been other technologies that failed in their first iteration.... As long as modifications are made and something new comes out of it, I think you'll generate interest," says Reni Benjamin, senior biotech analyst at Rodman and Renshaw (New York).

In the meantime, the question is whether there is enough money to support the approach in the coffers of biotechs or coming from the pharmaceutical industry, which has been burned repeatedly (**Table 2**). And what lessons from the ever-growing list of failures—and some possible successes—will inform future practitioners in the field?

Beginnings

Cancer vaccinology is predicated on the notion of awakening the immune system to the presence of cancer by presenting it with antigens associated with tumor cells. Once the immune system is roused, the concept is that it would be capable not only of mounting a sustained bodywide search for similarly suspicious cells, but also of retaining a memory of the abnormal antigens, permitting a renewed, rapid assault should the tumor recur.

The notion that the immune system could be enlisted to launch an attack on an existing tumor has been around at least since the late 1800s, when the New York City–based physician William Coley noticed that metastases at several sites regressed in a sarcoma patient after she developed a bacterial incision-wound infection. Coley's attempts to exploit this discovery were handicapped by the then-crude state of knowledge. But to this day, remnants of this approach can be seen in the use of general immune stimulants, like attenuated bacteria (e.g., mycobacterial components in Bacille Calmette Guerin, BCG) and interleukins, in treating bladder cancer and melanoma, respectively, as well as their inclusion in combination therapies in literally hundreds of clinical trials.

The discovery and identification of tumor-associated antigens, which now number in the hundreds (see **Table 3** for some examples), stimulated a second approach to cancer vaccines, an approach still highly visible among the therapies being tested today. Roughly half of ongoing clinical trials enlist a tumor-associated antigen or collection of antigens (**Fig. 1** and **Table 4**). Many such trials have ended in failure, which we now know is because these antigens muster only weak immune responses because they are normal human proteins merely overexpressed on tumor cells (to which the patient would be tolerant) or they too closely resemble such proteins or they elicit only a weak response from the patient's compromised immune system. It is now known that multiple co-stimulatory signals are needed to generate a robust T-cell response against a tumor-associated antigen; if these signals are not supplied, T-cell anergy and peripheral tolerance follows. Such tepid immune responses are not nearly what would be needed to eradicate advanced cancers, which early on accounted for most patients treated in clinical trials. Contemporary trials using tumor-associated and more promising tumor-specific antigens now use various immune stimulatory molecules, such as granulocyte macrophage colony stimulating factor (GM-CSF), and generalized adjuvants, such as keyhole limpet hemocyanin (KLH), to boost the response.

Many approaches have explicitly tried to engage cell-mediated immunity either using isolated antigen-presenting cells (APCs) or attempting to stimulate them *in situ* (**Fig. 2**). Techniques were developed for extracting dendritic cells, a major APC, loading them up with tumor antigens in various ways and reintroducing them into patients. Early attempts here failed, and in some cases, actually led to poorer outcomes than if the individual had been untreated, as immature dendritic cells, it was later learned, were as likely to suppress the immune system as to stimulate it. Methods for characterizing the right types of dendritic cell and other APCs are now being worked out, and it's become clearer how to activate these cells through cytokines, such as GM-CSF, to optimize antigen presentation (one such immunotherapeutic candidate in late-stage

Bruce Goldman is a freelance writer living in San Francisco, and Laura DeFrancesco is Senior Editor, Nature Biotechnology.

Table 1 Selected early stage cancer vaccine programs

Company (location)	Product	Composition	Indication	Trial phase
Antigen Express (Worcester, MA, USA; a subsidiary of Generex Biotechnology, Toronto)	Her-2/neu breast cancer vaccine	Her-2/neu epitope peptide conjugated at N terminus to the C terminus of the key moiety of the MHC class II–associated invariant chain (Ii protein) containing a four–amino-acid (LRMK) modification	Breast cancer	Phase 2
Apthera (Scottsdale, AZ, USA)	NeuVax	Immunopeptide (E25) from Her-2/neu administered together with GM-CSF	Early-stage breast cancer	Phase 1/ 2
Argos Therapeutics (Durham, NC, USA)	AGS-003	Autologous dendritic cells loaded with total RNA from resected tumors	Renal cancer	Phase 2
Immunocellular Therapeutics (Los Angeles, CA, USA)	ICT-107	Autologous dendritic cells treated with tumor-specific peptides from 6 antigens expressed on glioblastomas	Brain cancer	Phase 1
Immunotope (Doylestown, PA, USA)	IMT-1012	Peptide vaccine containing 12 tumor-associated peptides discovered through proteomics, including A-kinase anchor protein 9, midasin (MIDAS-containing protein RAD50), talin 1, vinculin vimentin and centrosome-associated protein 350	Advanced ovarian and breast cancer	Phase 1
Pevion Biotech (Bern, Switzerland)	Pevi-Pro	Influenza virosomes expressing three Her2/neu epitopes	Breast cancer	Phase 1
Vaxon Biotech (Paris)	Vx-001	A peptide vaccine comprising the cryptic peptide human telomerase reverse transcriptase (TERT_{572}) and its HLA-A*0201-restricted modified variant ($\text{TERT}_{572\text{Y}}$)	NSCLC	Phase 1

Table 2 Selected deals in the cancer vaccine sector

Date	Company (location)	Partner (location)	Product	Deal terms	Status
12/08	Oncothyreon	Merck KgaA	Stimuvax for NSCLC	$13 million for manufacturing rights	Phase 3
11/08	Argos Therapeutics	Private investment	RNA-loaded autologous dendritic cells (and other immunotherapies)	$35.2 million for series C funding	AGS-033 for renal cell carcinoma in phase 1/2
4/08	Cell Genesys	Takeda (Osaka)	GVAX for prostate cancer	$50 million up front, plus milestones worth up to $270 million	Collaboration terminated after GVAX trial stopped (12/08)
3/07	Oxford Biomedica	sanofi aventis (Bridgewater, NJ, USA)	TroVax (allogeneic modified vaccinia strain Ankara expressing 5T4 (OBA1) antigen) for renal cancer	$690 million in royalties and milestones	Phase 2 vaccinations stopped due to excess deaths (7/08), analysis continues of vaccinated patients
12/04	CancerVax	Merck/Serono	Canvaxin for melanoma	$25 million cash up front, $12 million equity purchase; equally share development costs, up to $253 million in milestones	Partnership terminated (11/05); CancerVax merged with Micromet in 2006
4/03	Biovest	Accentia	BiovaxID (autologous idiotypic determinant from B-cell lymphoma conjugated to KLH and combined with GM-CSF) for non-Hodgkin's lymphoma	$20 million; Accentia owns 81% of Biovest	Phase 3
7/01	IDM Pharma (Irvine, CA, USA)	sanofi aventis	Uvidem (autologous dendritic cell vaccine loaded *ex vivo* with tumor antigens derived from resected tumor) for melanoma	$33 million	Partnership terminated (1/08)

clinical trials, Dendreon's Provenge (Seattle; sipuleucel-T) for prostate cancer, may prove to be among the first therapeutic cancer vaccines to receive FDA approval; **Box 1**).

Just in the past five years, information has surfaced, pointing to a whole new problem with cancer immunotherapy—active immunosuppression in the tumor microenvironment. Tumors have been long suspected to evade immune detection by, for example, Darwinian evolution of cells whose defining surface antigens are suppressed or creating positive pressure gradients that make it harder for circulating immune cells to penetrate them (**Fig. 3**). But now, it has emerged that in addition to evasion, tumors actually can induce local immunosuppression through the stimulation of regulatory T cells or the recruitment of myeloid-derived suppressor (MDS) cells. The former, primarily through their production of transforming growth factor (TGF)-β, inhibit CD8$^+$ cytotoxic T cells (CTLs), T helper 1 (T$_H$1) cells and natural killer (NK) cells, which are the main mediators of immune surveillance against tumors. MDS cells, a mixed population of relatively immature myeloid cells, also suppress cellular immune responses primarily by producing arginase 1 and nitric oxide synthase 2A.

One means of potentiating the power of cancer vaccines and unleashing the immune system, according to leading academics, would be to counteract tumor-mediated immune suppression. This could be accomplished by targeting the regulators of the regulators, so to speak. For example, several molecules have been identified (e.g., CTL antigen 4, CTLA-4) that engage with regulatory T cells. Animal studies have shown that blocking such interactions, either with monoclonal antibodies (mAbs) or gene knockouts abrogates immune suppression. Indeed, several dozen clinical trials, according to the US National Institutes of Health (http://www.clinicaltrials.gov), are currently underway using mAbs against CTLA-4 in combination with chemotherapy or vaccines.

Table 3 Examples of tumor-specific antigens

Tumor-specific, shared antigens
Cancer only
MAGE-3
NY-ESO-1
TRAG-3
Expressed in some normal tissues
WT-1
PRAME
SURVIVIN-2B
Overexpressed in cancer
Her-2
MUC-1
Survivin
Mutated, unique
p53
α-actinin-4
Malic enzymes

Source: GSK

Immunotherapy's many faces

Cancer immunotherapy means different things to different people. In the case of cancers that are known to express viral antigens (e.g., cervical cancer and some melanomas that express human papilloma virus), immunotherapy takes the form of a classic immunoprotective, prophylatic vaccine like smallpox or polio where a viral antigen is presented to the immune system. In those cases where cancers overexpress a particular endogenous surface antigen (e.g., Her-2 in some breast cancers or CD-20 in some lymphoma cells), mAbs directed against those surface markers (Genentech's Herceptin (trastuzumab) and Genentech's and Biogen-Idec's Rituxan (rituximab), respectively) provide passive immunity, which can keep a tumor in check for a while. There are many such mAbs for various cancers under development. As currently applied, these mAbs are not preventive but rather therapeutic, though Herceptin has been approved for ever earlier stages in breast cancer, where it might, at least in theory, protect against recurrences by preventing metastases from taking hold.

Active immunotherapies, on the other hand, are designed to incite the individual's own immune system to mount a response to an antigen or group of antigens exclusive to or predominantly associated with the patient's tumor. They can take the form of peptide/protein vaccines or cellular vaccines.

The former type of vaccine generally falls into two categories. The first is based on shared peptide or protein antigens that occur commonly in a particular cancer or group of cancers (epidermal growth factor receptor (EGFR) vIII, for example, which is found in 30–40% of glioblastomas, or MAGE-3, which is expressed on many lung tumors). The proteins can be injected directly or expressed on attenuated virus particles, or nonproliferative bacterial or yeast cells (**Box 2**). An alternative approach is to isolate antigens from an individual patient and present these back to the person in a form designed to elicit immune surveillance, such

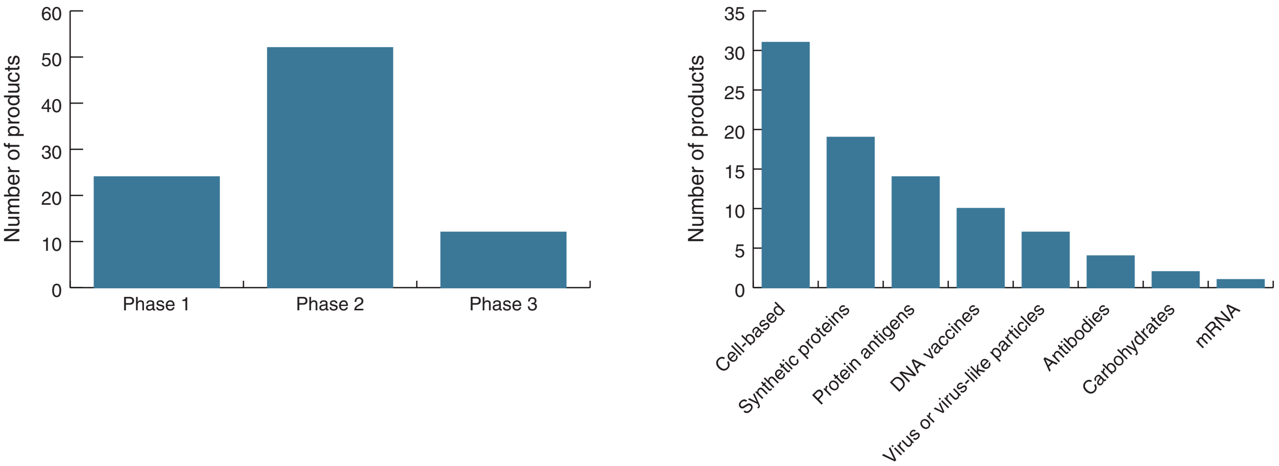


Figure 1 Cancer vaccine types. (Source: Tufts Center for the Study of Drug Development)

as vaccines designed to stimulate responses against antibody idiotypes found on lymphomas or the use of heat shock proteins to present unique tumor peptides (**Box 3**).

Cellular cancer vaccines can also be divided into two broad groups: allogeneic or autologous. The former, so-called 'off-the-shelf' vaccines, are usually collections of tumor cell lines, administered as aggregates to present several potential tumor antigens to the patient's immune system. Autologous whole cells, on the other hand, are isolated from, and returned to, individuals after some *ex vivo* manipulation to activate or induce maturation of APCs. An example of this type of vaccine would be a product based on isolation of APCs from a patient that is engineered to express some soluble factor (or factors) that generates an immune response to a common antigen (e.g., prostate-specific antigen in the case of prostate cancer, or p53/telomerase more generally (**Box 1**)).

Compared with cellular vaccines, peptide vaccines have the advantage of being similar to existing vaccine approaches used for decades in immunization programs against infectious agents. Such vaccines are less tricky to manufacture on a large scale than cellular vaccines. In 2002, for example, the FDA placed a hold on CancerVax's (Carlsbad, CA, USA) phase 3 trial of cellular vaccine Canvaxin because of manufacturing concerns. What's more, the longer clinical history and widespread use of peptide/protein vaccines means that regulators are more familiar with their oversight and less likely to raise issues unanticipated by product sponsors.

The perilous path

Cancer vaccines represent a relatively small portion of the oncology drugs in commercial development. The Tufts Center for the Study of Drug Development (Boston) reports that only one-fifth of oncology biologic therapeutics in company pipelines are vaccines (**Fig. 4**).

Table 4 Selected cancer vaccines in late clinical trials

Company (location)	Product	Description	Indication	Trial phase
Whole-cell-based autologous cells (personalized)				
Avax Technologies (Philadelphia)	M-Vax	Autologous cell vaccine in which patient tumor cells are treated with the hapten dinitrophenyl	Metastatic melanoma with at least one tumor to create vaccine	Phase 3
Dendreon	Provenge	Autologous dendritic cells exposed *ex vivo* to fusion protein combining prostate alkaline phosphatase and GM-CSF	Asymptomatic, metastatic hormone-refractory prostate cancer	Phase 3
Geron (Menlo Park, CA, USA)	GRNVAC1	Autologous dendritic cells transfected with mRNA for human telomerase and a portion of lysosome- associated membrane protein (enhances antigen presentation)	AML in remission	Phase 2
IDM Pharma	Bexidem	Autologous interferon-γ-activated macrophages (monocyte-derived activated NK cells).	Superficial bladder cancer	Phase 2/3
	Uvidem	Autologous dendritic cell vaccine loaded *ex vivo* with tumor antigens derived from resected tumor	Melanoma with M1a or M1b stage disease and/or in-transit lesions; stage III and IV melanoma	Phase 2
	Collidem		Colorectal cancer	Phase 1/2
Introgen Therapeutics (Austin, TX, USA)	INGN 225	Dendritic cells treated with an adenovector carrying the human p53 gene	Advanced metastatic SCLC Breast	Phase 2
MolMed (Milan)	M3TK	T cells bioengineered to express MAGE 3 tumor antigen	Metastatic melanoma	Phase 2 (enrollment halted)
Northwest Biotherapeutics (Bethesda, MD, USA)	DC-Vax Prostate	Dendritic cells loaded with recombinant prostate-specific membrane antigen (PSMA)	Hormone-dependent, nonmetastatic prostate cancer	Phase 3
	DC-Vax Brain	Dendritic cells loaded with tumor extract	Newly diagnosed glioblastoma multiforma requiring surgery, radiation and chemotherapy	Phase 2
Prima Biomed (Sydney, Australia)	CVac	Dendritic cells primed with a mucin-1 and a mannan-fusion protein adjuvant	Late-stage ovarian cancer	Phase 2
Whole-cell-based allogeneic tumor cells (off-the-shelf)				
Cell Genesys	GVAX pancreatic	Two allogeneic cultured cancer lines, irradiated and bioengineered to secrete GM-CSF.	Metastatic pancreatic cancer	Phase 2
	GVAX leukemia	One allogeneic leukemia cell line irradiated and bioengineered to secrete GM-CSF	Newly diagnosed AML, chronic CML and myelodysplastic syndrome	Phase 2
NovaRx (San Diego)	Lucanix	Four non-small cell lung cancer cell lines carrying antisense oligonucleotides against transforming growth factor β-2	Advanced NSCLC	Phase 3

Table 4 Selected cancer vaccines in late clinical trials (continued)

Company (location)	Product	Description	Indication	Trial phase
Onyvax (London)	Onyvax-P	Three human cell lines representing different stages of prostate cancer	Hormone-resistant prostate cancer	Phase 2
Unique-antigen-based (personalized): purified peptide or protein				
Antigenics	HSPPC-96 Oncophage	Heat shock protein vaccine purified from autologous tumor cells	Recurrent glioma	Phase 2 (investigator-initiated trial)
			Resected renal-cell carcinoma (RCC)	Phase 3 (completed)
Biovest International	BiovaxID	Tumor-specific idiotype conjugated to keyhole limpet hemocyanin, plus GM-CSF	Mantle cell lymphoma Indolent follicular B-cell non-Hodgkin's lymphoma	Phase 2 Phase 3
Shared antigen (off-the-shelf): purified protein or peptide				
Apthera (Scottsdale, AZ, USA)	NeuVax	Immunogenic peptide derived from the Her-2/neu protein plus GM-CSF	Early-stage Her-2-positive breast cancer	Phase 2/3
CellDex	CDX-110	A 14-amino-acid segment of a mutated EGFR	Glioblastoma multiforme	Phase 2/3
Cytos Biotechnology (Schlieren, Switzerland)	CYT004-MelQbG10	Modified fragment of the Melan-A/MART-1 protein coupled to the carrier QbG10	Advanced-stage melanoma	Phase 2
Generex Biotechnology	Ii-Key/HER2/neu cancer vaccine	Peptide vaccine containing Ii-Key modified Her-2/neu protein fragment	Node-negative breast cancer	Phase 2
GlaxoSmithKline Biologicals (Brussels, Belgium)	MAGE-A3 antigen-specific cancer immunotherapeutic	Liposomally packaged cancer vaccine against MAGE-3 antigen	Metastatic MAGE-A3-positive melanoma NSCLC following surgery	Phase 3 Phase 3
IDM Pharma	IDM-2101	Nine CTL epitopes from four tumor-associated antigens, including two proprietary native epitopes and seven modified epitopes and one universal epitope (a source of T-cell help)	NSCLC	Phase 2
Immatics Biotechnologies (Tuebingen, Germany)	IMA901 IMA910	Peptide vaccine comprising multiple fully synthetic tumor-associated peptides	Renal cancer Colorectal cancer	Phase 2 Phase 1/2
Norwood Immunology (Chelsea Heights, Australia)	Melanoma cancer vaccine	Melanoma-specific peptides gp100 and MAGE-3	Melanoma	Phase 2
Oncothyreon	Stimuvax	Liposomal vaccine containing a synthetic 25–amino-acid-peptide sequence from MUC-1	Stage III NSCLC	Phase 3
Pharmexa (Hoersholm, Denmark)	GV1001	Recombinant protein vaccine targeting human telomerase reverse transcriptase, plus GM-CSF	Pancreatic Liver Lung	Phase 3 Phase 2 Phase 2

AML, amyotrophic lateral sclerosis; CML, chronic myelogenous leukemia.

Although modern cancer vaccine development dates back to the 1980s, none has been approved in the United States (though there are five products on the market elsewhere; **Table 5**). Thus, the rate of approval of cancer vaccines lags far behind other biologics—as of 2006, seven of twelve vaccines in phase 3 clinical trials had entered clinical study a decade earlier.

To date, an estimated 7,000 people have participated in late-stage clinical trials of active cancer immunotherapies. These have largely been an exercise in frustration, as candidates—including a few that looked quite good in early trials—have fallen by the wayside in pivotal phase 3 trials. Some recent losers that have gone quietly into the night:

- PANVAC (Therion Biologics, Cambridge, MA, USA), an off-the-shelf vaccine consisting of attenuated poxvirus carrying genes encoding two tumor-associated antigens (carcinoembryonic antigen and mucin 1, MUC-1) and three immunostimulatory molecules (intracellular adhesion molecule 1, B7.1 and lymphocyte function–associated molecule 3) for use in advanced pancreatic cancer, failed to meet clinical endpoints after promising early trials, leading the company to close its doors and file for bankruptcy protection in December 2006.

- Theratope (Biomira, Edmunton, AB, Canada; now Oncothyreon, Seattle), an off-the-shelf vaccine, consisting of a synthetic mimic (STn-crotyl) of the tumor-associated, *O*-linked epitope of MUC-1 (STn-serine), tethered to an immunostimulatory protein (KLH) and delivered along with an adjuvant from Seattle-based Corixa (Detox-B, an oil droplet emulsion containing monophosphoryl lipid A and cell wall skeleton from *Mycobacterium phlei*) for use in metastatic breast cancer, showed no improvement in either time to progression or overall survival. The company hasn't completely abandoned the target; in partnership with Merck KGaA (Darmstadt, Germany), it has developed a "more sophisticated" approach for eliciting a T-cell response, according to Marita Hobman, director of intellectual property

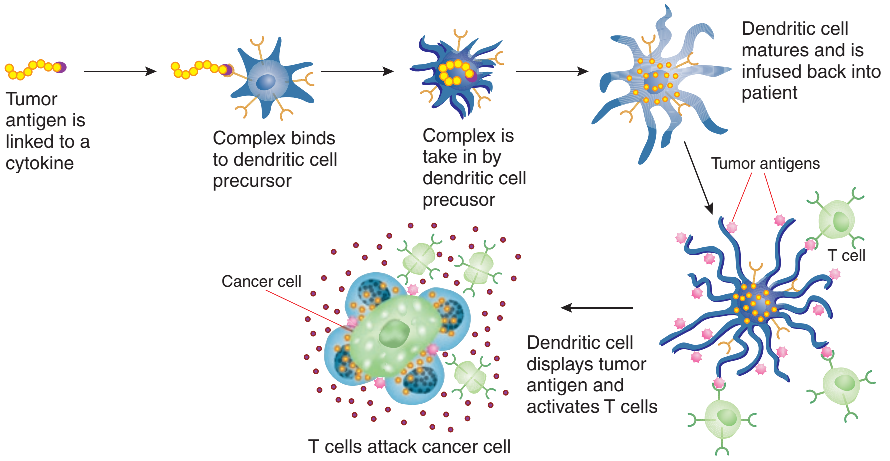

Tumor antigen is linked to a cytokine
Complex binds to dendritic cell precursor
Complex is take in by dendritic cell precusor
Dendritic cell matures and is infused back into patient
Tumor antigens
T cell
Dendritic cell displays tumor antigen and activates T cells
Cancer cell
T cells attack cancer cell

Figure 2 Dendritic cells that attack cancer. (Source: National Cancer Institute)

management and business development at Oncothyreon.

- Canvaxin (CancerVax, now MicroMet, Munich), an off-the-shelf mix of three irradiated melanoma cell lines bearing over a dozen defined tumor-associated antigens, plus an adjuvant (BCG) for use in stage III melanoma, yielded worse outcomes in treated patients than in controls, unlike earlier trials in which patients had been more carefully selected for human leukocyte antigen (HLA) alleles correlating with better outcomes. After Canvaxin failed, CancerVax merged with Micromet, which is developing passive immunotherapies using mAbs against various tumor antigens.

- GVAX (Cell Genesys, S. San Francisco, CA, USA), an off-the-shelf, whole-cell vaccine, consisting of infusions of cells from existing prostate cancer lines engineered to express GM-CSF for use in hormone-refractory prostate cancer, yielded excess deaths in treated patients versus controls, leading to abandonment of the trial.

Although there is a clear preponderance of off-the-shelf vaccines in this group of failures, the fate of individualized vaccines has not necessarily been much better. Two companies with vaccines targeting antibody idiotypes associated with tumors—Favrille (San Diego) and Genitope (Fremont, CA, USA)—both shut down their trials when their products failed to reach statistical significance, essentially ending their programs in late 2008.

Getting it right

A cancer vaccine has to jump through several hoops, says Johns Hopkins University

Box 1 A whole-cell vaccine nears approval?

Dendreon is developing a whole-cell-based candidate, Provenge, for metastatic, hormone-resistant, prostate cancer (HRPC). The vaccine is a patient-specific, vaccine produced by incubating an individual's own blood, enriched for dendritic cells and other APCs with a recombinant fusion protein composed of prostatic acid phosphatase (PAP) and GM-CSF.

Although not tumor specific, PAP is highly tissue specific. Although expressed in the majority of prostate tumors, PAP is only minimally expressed in tissues other than the prostate gland, says Dendreon's Frohlich, who is chief medical officer. It is immunologically distinct from acid phosphatases found in other tissues. Because HRPC patients' prostates have already been surgically removed or irradiated, autoimmunity doesn't pose much of a practical problem.

The question of whether tolerance to this antigen can be broken was addressed in a preclinical study performed by Dendreon investigators and published in 2001 (ref. 2), in which their product induced autoimmune prostatitis in rats (a clear sign of immune mobilization against the antigen). This convinced Dendreon that it could raise an immune response in a clinical setting as well. A phase 1/2 trial published in 2000 (ref. 3), demonstrating strong antigen-specific T- and B-cell responses to the approach, was consistent with this finding.

That year, Dendreon launched two trials of Provenge, each with about 120 asymptomatic patients. As Frohlich explains, it was then believed that asymptomatic patients would progress more slowly than symptomatic patients, buying time for the initially subtle effects of immunotherapy to kick in before the disease reached a stage that was intractable to immunotherapy.

In the interest of getting a fast readout, Dendreon had picked as its primary endpoint time to progression (TTP)—assumed to be a reasonable surrogate for survival. But Dendreon was to find out otherwise. During the course of the trial, medical opinion leaders decided that overall survival is a better endpoint than TTP, which carries a subjective component. However, the trial continued with the previously agreed upon endpoint of TTP.

This proved ironic. When the first trial was unblinded, TTP had missed statistical significance ($P = 0.05$) by the barest of margins, ($P = 0.052$) whereas overall survival analysis demonstrated a 41% reduction in the risk of death, with a high level of statistical significance ($P = 0.01$).

But survival was not a prespecified primary endpoint. And whereas an outside advisory committee voted 13–4 for approval, in April 2007, the FDA instead insisted on another trial to confirm the survival results.

In the aftermath of the FDA's failure to approve Provenge despite clear signs of efficacy, angry patient advocates peppered the agency with letters of protest. But proponents of strict adherence to trial protocols liken the argument that a therapy ought to be approved on the basis of an unplanned analysis to moving the goal posts[4].

Dendreon is soldiering on with a new 512-patient trial with the primary endpoint of survival. Preliminary results, announced in October 2008, demonstrated a 20% reduction in the risk of death in the treatment arm, only slightly less than the 22% reduction, which the company believes is necessary to achieve statistical significance. Final results are due later this year, and if Provenge makes the grade, it may yet turn out to be the first whole-cell-based active cancer immunotherapy approved by the FDA. But many years, many tens of millions of dollars, and perhaps more than a few lives might have been saved had the Dendreon's phase 3 trial not been marred by an unfortunate choice of an endpoint.

oncologist Hyam Levitsky, co-inventor of GVAX and member of the board of the cancer vaccine company Antigenics (New York). "In an existing tumor, the body has already been exposed to those antigens, so there may already have been an initial immune response. But very often, the immune system is defeated and rendered tolerant to the antigens that the vaccine is targeting. A successful vaccine has to overcome this tolerance, and that's not trivial." Moreover, Levitsky says, the vaccine frequently has to work in what can be a hostile environment. "The tumor has essentially taken over and altered the landscape, stealing various attributes of the normal immune system to turn down immune response."

The antigens to use in a vaccine to circumvent the challenge of breaking immune tolerance without generating autoimmunity should be tumor specific. But such antigens are rarely found, says Jeffrey Weber, head of the Comprehensive Melanoma Research Center at the H. Lee Moffitt Cancer Center (Tampa, FL, USA). "These are few and far between. You can discover any number of mutated, tumor-specific antigens, but you seldom find any that turn up on more than 5% of tumors of any given type." And even when you find one, he says, that doesn't mean it will be highly immunogenic.

In practice, cancer antigens targeted by active immunotherapies have more often been tumor associated: overexpressed on tumors, but nonetheless present at lower frequencies in normal tissues. In trials of vaccines based on these antigens, the necessity of breaking tolerance—for example, by pairing the selected antigen with a powerful adjuvant—has clashed with the need to avoid an excessive immune assault on healthy tissues where the antigen also resides. "You can vaccinate the hell out of somebody against melanoma self-antigens that are overexpressed on cancer, and you won't induce severe side effects—or any immune response to speak of," says Weber. "But if you administer the same vaccine along with one dose of anti-CTLA-4 antibodies, you can induce life-threatening autoimmune colitis or skin rash or hepatitis."

Another problem plaguing trials of cancer immunotherapies has been the intractability of the cancers targeted. In theory, any cancer should be amenable to immunotherapy, but in practice, only a few cancers have received most of the attention, at least historically. Melanoma, which early on was found to have tumor-specific antigens, has been targeted frequently using the protein or peptide approach—mostly without success, as no really tumor-specific melanoma antigens have yet been exploited, only tumor-associated antigens. But those cell-based approaches, in which autologous proteins or extracts are used for priming, require access to a sufficient tumor mass. This more or less excludes melanoma or even breast cancer, where the tissue tends to be fibrotic and where tumors tend to be diagnosed increasingly early, while they are still relatively small.

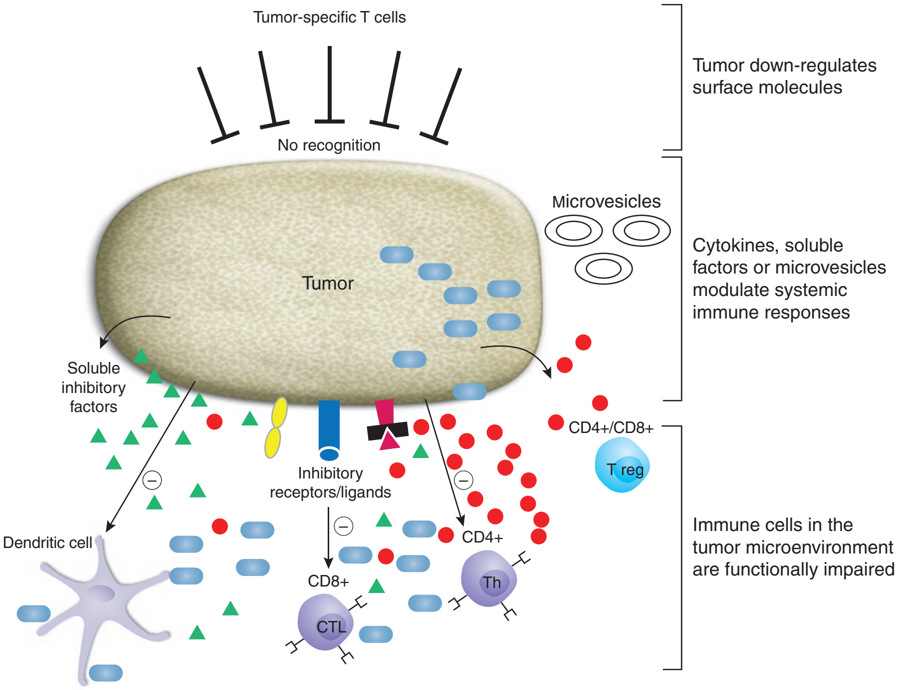

Tumor-specific T cells
No recognition
Tumor down-regulates surface molecules
Microvesicles
Tumor
Cytokines, soluble factors or microvesicles modulate systemic immune responses
Soluble inhibitory factors
CD4+/CD8+
T reg
Inhibitory receptors/ligands
Dendritic cell
CD4+
Th
CD8+
CTL
Immune cells in the tumor microenvironment are functionally impaired

Figure 3 Tumor cell's interactions with the immune system. (Reprinted from Whiteside, T.L. Immune suppression in cancer: effects on immune cells, mechanisms and future therapeutic intervention. *Semin. Cancer Biol.* **16**, 3–15, 2006, with permission from Elsevier.)

Recognizing that the immune response takes time to develop, some vaccine developers have turned to slow-growing prostate cancer or kidney tumors, where the time to progression is longer. And then, of course, greater prevalence of certain tumor types, such as lung, create a large patient pool with which to populate clinical trials, whereas the dearth of decent treatments for these indications speaks most loudly to the need for ramped-up clinical experimentation.

Certainly, the tendency to use individuals who are in advanced cancer stages has made proof of clinical efficacy more difficult to achieve. Of course, individuals with late-stage disease, who have often been treated with other therapeutic agents that have failed, tend to be more available. And sponsoring companies prefer this population because they expect that positive treatment effects will be observed more quickly in advanced-stage patients than in early-stage or fully resected ones. But decades' worth of clinical trials of cancer vaccines conducted across multiple tumor types not surprisingly suggest that immunotherapies are more likely to work best in patients with earlier-stage, less-aggressive tumors[1] or in individuals whose tumor burden has been reduced to the microscopic level by surgery or chemotherapy.

"It's at this level of microscopic disease where I think cancer vaccines are most likely to succeed," Levitsky says. "Well over 50% of the common cancers can be treated into a state of minimal residual disease. What we lose patients to is typically not the inability to get the disease into that minimal state, but rather the inability to completely eradicate the residual component." All too often, a seemingly excised tumor returns. "From a public-health point of view," he says, "the impact of an effective immunotherapy—delivered at the point of minimal residual disease—that could wipe out the last traces of a tumor, would be truly staggering. Ironically, that's probably the most difficult time to demonstrate efficacy in a clinical trial."

Standard measures of a cancer therapy's efficacy—tumor shrinkage or growth arrest—are worthless for patients with minimal residual disease. How can you score tumor shrinkage if the patient no longer appears to have a tumor?

Box 2 Pharma perseveres with off-the-shelf vaccines

Off-the-shelf vaccines have the advantage that they can be produced in bulk, making them more attractive to big pharma than individualized vaccines that are tailored to each patient. One such vaccine, CDX-110, developed by John Sampson at Duke University (Durham, NC, USA) and Amy Heimberger at MD Anderson Hospital (Houston), has been in-licensed by Celldex Therapeutics (Phillipsburg, NJ, USA, which merged with Avant Immunotherapics, Needham, MA,USA, in late 2007) and has attracted the attention of Pfizer (New York).

The vaccine targets EGFRvIII (a 14-amino-acid segment of a mutated EGFR) that not only appears solely on glioblastoma cells, but also has never been expressed in any other kind of cell at any time in development. Its biological activity is clearly germane to the tumor's aggressiveness, so knocking it out should directly impair the tumor's viability. It is located on cell surfaces, making it accessible to attack. And, because it's a mere peptide rather than a full-sized protein, it's simple to manufacture.

EGFRvIII is found on 30–40% of glioblastomas, an aggressive form of brain cancer, which even when surgically excised, irradiated and exposed to chemotherapy, typically recurs within six months. The mutant receptor is characterized by a 267–amino acid deletion within its extracellular domain, which changes the molecule's configuration, locking it into a perpetual signaling mode that drives relentless cell replication. Thus its 100% tumor specificity: no cell with elevated expression of this mutant receptor could possibly be normal. In addition, the deletion creates a novel splice junction, which CDX-110 spans.

The dearth of effective therapies for glioblastoma makes it possible to test the new vaccine as a front-line therapy in patients who have just had their tumors thoroughly resected. In a phase 2 trial (ACTIVATE), 22 patients with EGFRvIII-positive glioblastomas were given standard treatment, followed by serial injections of the vaccine. Time to progression (TTP) more than doubled to more than 14 months compared with 6.4 months for a set of EGFRvIII-positive historical controls. Overall survival improved commensurately, proving surprisingly enduring, with about two-thirds of the injected patients surviving for two years, more than one-third for at least three years and a fifth still alive after four years. In a second phase 2 study (ACT II), in which 23 subjects received the vaccine simultaneously with chemotherapy, patients' median TTP reached 16.6 months, and median survival time 33.1 months—a point at which, historically, all EGFRvIII-positive patients would long since have died.

This jump in long-term survival is puzzling, as EGFRvIII-positive glioblastomas typically contain large numbers of EGFRvIII-negative cells—a status that ought to shield them from the vaccine. One possible explanation says Sampson, is that EGFRvIII-positive cells are stem cells for the tumor, another possibility is that EGFRvIII-positive cells either make other tumor cells more proliferative or harder to kill. Chuck Baum, senior vice-president and head of the oncology development at Pfizer (New York), which licensed the rights to the vaccine in April, 2008, suggests that part of the apparently powerful effect CDX-110 has on even EGFRvIII-negative tissue may be due to a phenomenon called 'epitope spreading': as tumor cells lyse, they release their internal contents into the surrounding medium, giving local APCs access to previously occult tumor antigens (e.g., mutant intracellular proteins). "You can get a broadening response with time, and eventually end up with a more effective immune response than the one you started with," Baum says.

Avant and Pfizer are working on a large phase 2/3 trial, with preliminary results expected by mid-2009. With the aggressive lethality that characterizes glioblastoma, those results shouldn't be long in coming. "It's not going take 20 years to find out if the vaccine worked," says Heimberger.

Another vaccine in the pipeline that targets an antigen that is both tumor specific and present in enough patients' tumors to allow an off-the-shelf, mass-produced vaccine is GlaxoSmithKline (GSK) Biologicals (Brussels) MAGE-3 (melanoma antigenic epitope 3). Since October 2007, the vaccine-producing division of the pharmaceutical giant has been recruiting non-small-cell lung cancer (NSCLC) patients for a large phase 3 clinical trial. The 400-center study, spanning 33 countries, will be the largest-ever in lung cancer and, for that matter, the largest-ever study of an active cancer immunotherapy for any indication.

Vincent Brichard, senior vice-president for cancer immunotherapeutics at GSK, says that about 40% of all NSCLC tumors express MAGE-3, which is expressed only transiently during fetal development. In adults, MAGE-3 expression is confined to the testes, opaque to immune surveillance so that tolerance doesn't develop.

This makes it possible to conceive of an off-the-shelf immunotherapy targeting the widely shared antigen. The GSK approach uses the entire 360 amino acid–long MAGE-3 protein to maximize the number of epitopes. GSK's vaccine bolsters the T-cell immunogenicity of its recombinant MAGE-3 protein by packaging it in liposomes, which Brichard says enhances delivery to APCs and by administering the vaccine with an adjuvant mix that has been optimized to produce a potent T-cell response to MAGE-3. This immunostimulatory potion combines GSK's own adjuvant, monophosphoryl lipid A (MPA), with QS-21, a complex lipid mix licensed from Antigenics, and CpG oligonucleotide (a Toll-like receptor 9 agonist developed by Coley Pharmaceutical Group, a Canadian biotech purchased in late 2007 by Pfizer).

QS-21 is also known to induce a strong antibody response. Antibodies could conceivably play a role against even an intracellular molecule such as MAGE-3 to the extent that lysed tumor cells release entire, undegraded protein molecules that could be targeted by antibodies. The resulting antigen-antibody complexes would, in turn, be highly available for uptake by APCs.

At the annual meeting of the American Society of Clinical Oncology in Chicago in June 2008, GSK presented the results of a randomized, 182-patient phase 2 trial of MAGE-3. Early-stage NSCLC patients who had had their tumors completely resected and then received several injections of the MAGE-3 vaccine had roughly one-third the recurrence rates of those given placebo injections, results mirroring those from several other MAGE-3 tests.

GSK's huge phase 3 trial—whose primary endpoint, like that of the recent phase 2, is disease-free survival—departs from its phase 2 counterpart in two respects: first, about half of the patients in the trial will receive chemotherapy before vaccination, a regimen never tested in phase 2 (Brichard notes, though, that the trial's large size leaves plenty of statistical power for an independent analysis of those eschewing chemotherapy). Second, there has been a change in the adjuvant mix's composition—the addition of CpG —since the phase 2 trial. Brichard says the new mix proved more immunostimulatory compared with the earlier formulation in another head-to-head trial, but it is into such small cracks that surprises can flow.

An alternative is to monitor recurrences or, more accurately, deaths among treated versus untreated patients. But that can take a long time. In renal-cell carcinoma, for example, the median time to recurrence for patients who have had their tumors fully resected and show no signs of residual tumor is 6.8 years.

Further complicating cancer immunology trials is the fact that each approach tends to be novel, creating trial-design and regulatory issues. A designated antigen can be either tumor associated or tumor specific, and tumor-specific antigens can be shared by many patients or unique to each patient. Shared antigens offer the prospect of off-the-shelf vaccines, with attendant economies of scale. But they also often lack tumor specificity, thus incurring the drawbacks of immune tolerance.

Whereas the failed Biomira vaccine, Theratope, is an example of a highly purified, well-defined antigen with a single epitope, CancerVax's Canvaxin candidate was a whole-cell mixture with multiple antigens, some of them undoubtedly not even identified, let alone characterized. The former approach runs the risk of eliciting too narrow an immune response. The latter may trigger unwanted cross-reactions, says Weber, and the difficulty of assessing its potency in any given person poses regulatory issues.

The roughly 30 different active cancer immunotherapies now in late-stage clinical trials (**Table 3**) also differ in their methods of manufacture and the indications for which they're being tested. Trial designs differ greatly, too. Among the variables: early- versus late-stage disease trade-offs, different endpoints and widely divergent timelines due to differential prognoses in different indications.

"Any therapy that's totally novel and first in class is a double-edged sword," says Mark Frohlich, senior vice president of clinical affairs and chief medical officer of Dendreon. "On the one hand, there's a lot of excitement from patients and regulators who want to approve something that's new and different. On the other hand, if it's a new product, those same regulators also need to make sure they're doing everything to ensure public safety and establish a solid precedent." Frohlich speaks from experience, having undergone an epic regulatory ordeal with Dendreon's cell-based prostate-cancer vaccine Provenge (**Box 1**).

For all these reasons, clinical trials of active cancer immunotherapies are high-stakes

Box 3 Personalized vaccines—a viable option?

Oncophage, a personalized vaccine developed by Antigenics, consists of an extract containing heat-shock protein-peptide complexes prepared from an individual patient's excised tumor. This approach is based on work by company co-founder Pramod Srivastava, now director of the University of Connecticut's Center for Immmunotherapy of Cancer and Infectious Diseases (Farmington, CT, USA), showing that APCs have receptors for heat shock proteins. This provides a pathway whereby unique tumor-specific antigens (the products of random mutations in rapidly dividing cancer cells) could become immunogenic. In principle, this preparation can target any tumor type, but in practice, its application is limited to tumors of sufficient size to obtain enough material.

In several early-stage trials involving individuals with advanced disease over several indications, there were striking cases of complete tumor regression as well as instances of partial shrinkage or stable disease, although the overall results were unspectacular. Given the virtual absence of side effects, Antigenics forged ahead, launching phase 3 trials in advanced metastatic melanoma[5] and renal-cell carcinoma (RCC)[6].

The melanoma trial proved difficult, as excised tumors were often too small to produce enough vaccine and subsequent clinical development was halted. But subjects with early-stage disease who got ten or more injections saw a big survival improvement over controls receiving currently approved treatments.

In the RCC trial, initiated in 2000, more than 700 patients whose tumors had been resected were randomized to either Oncophage or the current standard of care, which consists of 'watchful waiting' as there are no approved, effective treatments for such patients. As reported in *The Lancet* last July[6], an analysis triggered in November 2005 by the accumulation of a specified number of inidivduals whose disease had progressed found no statistically significant difference between treated and untreated patients for either relapse-free survival or overall survival. Disturbingly, though, the independent review committee also determined that 40% of the patients originally logged by principal investigators in the multicenter trial as having had recurrences, in reality had residual disease before treatment began. Excluding these patients from the analysis diluted the power of the study, which would have run much longer had these classification errors not been made. "One of the reasons the trial did not meet its endpoints relates to the fact that the data were evaluated prematurely," says Christopher Wood of MD Anderson, who was the principal investigator and author of the *Lancet* paper.

In March 2007, study investigators conducted a second analysis of more mature data, using an RCC classification system that hadn't existed when the Oncophage trial had begun. Vaccinated patients classified as "intermediate stage" in this new system (a subset consisting of stage 1–3b, which overlapped but was not equivalent to the older system's "early stage" group) enjoyed a statistically robust ($P = 0.026$) relapse-free survival benefit, suffering recurrences at just over half the rate of untreated patients—as well as a trend toward statistically significant improvement in overall survival ($P = 0.126$).

"What needs to be done, and hopefully will be done," Wood says, "is another trial that focuses on that earlier-stage group. But the amount of money to do that would be enormous, because you're narrowing down the population even further, so it would be hard to accrue. And because they're earlier-stage disease, recurrences are less frequent, so it could take ten years before you reach statistical significance."

The direct costs of the seven-year phase 3 trial exceeded $60 million, says Garo Armen, Antigenics' CEO. "That, plus the necessary maintenance of our manufacturing, quality control, quality assurance and manufacturing-related research functions throughout the trial, came to $250 million."

Armen says about 500 trial patients will continue to be followed up for relapse-free survival and overall survival for another three years, at a cost to the company of $1.5 million. Meanwhile, in April, 2008, Oncophage was approved in Russia, where a large number of the phase 3 patients had been recruited. Antigenics has also filed with the European Medicines Agency (EMEA) for approval in Europe. The EMEA policy of granting conditional approval would allow Oncophage to be launched commercially in Europe provided Antigenics commits to conducting

(continued)

Box 3 Personalized vaccines—a viable option? (continued)

a full-sized confirmatory phase 3 trial. Such an option is not available in the United States.

Like Antigenics's Oncophage, BiovaxID, developed by Biovest International, (Worcester, MA, USA) a majority-owned subsidiary of Accentia Biopharma (Tampa, FL, USA), is personalized and targets antigens that are both tumor specific and unique to each patient. But its construct is entirely different. The vaccine's lead indication is indolent follicular non-Hodgkin's lymphoma, in which a particular cancerous clone of antibody-producing B-lymphocytes proliferates.

The BiovaxID concept began in the Stanford University laboratory of Ron Levy and was pushed forward by Larry Kwak, who took the idea with him to the National Cancer Institute (NCI), where a phase 2 trial was initiated in 1995. Kwak now heads the lymphoma division at MD Anderson and consults for Biovest.

All the constituent cells of a B-cell lymphoma produce antibodies with identical idiotypes characteristic of the cancerous cells' hyperproliferative common ancestor. Those malignant B-cells also carry the idiotype on their surfaces. BiovaxID is an anti-idiotype vaccine consisting of hybridoma-produced identical copies of those overabundant (and, therefore, easily characterized) antibodies conjugated to the immune stimulant KLH. The vaccine is administered along with GM-CSF.

In 1999, favorable phase 2 results led the NCI to initiate a double-blind phase 3 trial, which Biovest took over under a Cooperative Research and Development Agreement. In the trial, 76 patients in remission following standard chemotherapeutic regimens, received serial BiovaxID injections. Importantly, only subjects who had sustained a complete remission after chemotherapy were included, because previous studies showed that patients in complete remission mount a better immune response (humoral and/or cellular) compared to those who do not, which also correlates with clinical outcome, according to Angelos Stergiou, chief medical officer and head of clinical research at Biovest.

This summer, Biovest reported that BiovaxID treatment prolonged disease-free survival by over one year, from 30.6 months for control subjects to 44.2 months for BiovaxID-treated subjects, ($P = 0.047$). The announced results will be formally presented at the next American Society of Clinical Oncology Conference in Orlando and will be submitted for peer review. In addition to seeking approval in the United States, the company is approaching regulatory agencies in other countries and plans on launching a compassionate-use program, referred to as Name-Patient Program, for BiovaxID, in parts of Europe early in 2009, according to Stergiou..

The apparent BiovaxID success follows failures of two other nearly identical candidates, advanced by Genitope (Fremont, CA, USA) and Favrille (San Diego), to reach statistically significant results in phase 3 trials over the past year. Like BiovaxID, these were anti-idiotype vaccines conjugated to KLH and delivered with GM-CSF, but Stergiou speculates that the method of preparation of the antibody may be the difference. Both Genitope and Favrille produced their antibodies as recombinant proteins rather than using the hybridoma methodology employed by Biovest.

Another potentially big difference is that in the BiovaxID trial, all vaccinated patients were in a state of complete remission. The other candidates' protocols, in contrast, allowed patients in partial remission—with visible tumor masses—to remain under study. Although this sped enrollment, it may have hindered efficacy.

BiovaxID's approach could, in theory, be applicable to other non-Hodgkin's lymphomas as well as multiple myeloma. But, as a personalized vaccine treatment that must be produced batch by batch for each patient, even the most efficiently manufactured product is likely to be expensive for patients—although Stergiou refuses to put a price tag on the vaccine at the moment.

propositions. Putting a novel approach through its paces among those most likely to benefit—patients who are least ill and for whom obtaining statistically significant results will thus presumably take the most patience—is a costly venture. Even the most sponsor-friendly phase 3 cancer immunotherapy trials—a modest 300-patient study of an easily manufactured, mass-produced off-the-shelf vaccine, in an indication with fast clinical readouts—and associated surgeries, imaging assays and so forth are going to cost about $20 million, according to one knowledgeable company official. With more and larger trials of personalized, more technology-intensive vaccines, the cost soars to hundreds of millions.

The way forward

All segments of the sector—immunologists, entrepreneurs, even regulators—appear to agree that the entrée into cancer vaccines was premature. Thomas Okarma, CEO of Geron (Menlo Park, CA, USA), which has a product in trials, calls early attempts at vaccines "immunological kindergarten." In addition, certain assumptions about the immune system may not be correct. Early failures with cancer vaccines led to the belief that tumors could not generate an immune response, according to Eli Gilboa, at the University of Miami, Florida. In fact, he says, "It appears [tumors] can [generate an immune response] for a time, but they have elaborated mechanisms for avoiding the immune system. Focusing more on how to mitigate tumor-induced immune suppression will be key going forward."

A second assumption that is proving false is that chemotherapy and immunotherapy are incompatible. According to Gilboa, evidence is emerging that some forms of chemotherapy are not incompatible, but in fact can synergize immunotherapy.

Another area of agreement is that we have reached the end of the single-agent era. Key to confronting the two main issues facing vaccine developers—the ability of tumors to induce tolerance and the hostile microenvironment in tumors—are combination therapies, but this creates certain problems, according to Robert Schreiber, cancer researcher at Washington University School of Medicine (St. Louis). "Most companies are locked into using their own products and therefore do not like to use combination therapies. And the FDA is particularly leery of trying too many combinations at once," he says. Schreiber sees a role for university-industry partnerships in getting around this potential logjam. "Once a successful regimen has been identified, there will be many companies that come knocking at the door. Since large-size clinical trials are very expensive, I see a great opportunity for industry and academia/foundations to pair up at this time," he says.

As for cell-based therapies, the jury is still out on whether autologous vaccines will be the ticket or whether there is a place for allogeneic, off-the-shelf ones. Logistics (read 'cost') seems to dictate that only allogeneic vaccines will be

Table 5 Approved and marketed cancer vaccines

Company (location)	Product	Description	Indication	Status
Antigenics	OncoPhage	Heat shock protein vaccine purified from autologous tumor cells	Renal cell carcinoma	Approved in Russia Granted fast track status by US FDA
Biovest International	BiovaxID	Tumor-specific idiotype conjugated to keyhole limpet hemocyanin, plus GM-CSF	Various B-cell–related cancers	Compassionate use in France, Germany, Italy, Greece, Spain and the UK. Granted fast track status by US FDA
Corixa (acquired by GSK in 2005)	Melacrine	Lysate from two melanoma cell lines, Detox adjuvant (proprietary) with monophosphoryl lipid A and mycobacterial cell wall skeleton	Melanoma	Approved in Canada
CreaGene (Seoul)	CreaVaxRCC	Autologous monocytes treated with GM-CSF and IL-4 to create immature dendritic cells activated with tumor extracts plus cytokines	Metastatic renal cell carcinoma	Approved in Korea
Genoa Biotechnologia (Brazil)	Hybricell	Autologous monocytes treated with cytokines and converted to dendritic cells that are fused with patient-derived tumor cells	Various cancers	Approved in Brazil
Vaccinogen (Frederick, MD, USA)	OncoVax	Metabolically active, irradiated, autologous tumor cells with BCG	Colon cancer	Approved in Europe, available in Switzerland Granted Fast Track status by FDA
Mologen (Berlin)	dSlim/Midge	Allogeneic tumor cells modified *ex vivo* to express interleukin 7 (IL-7), GM-CSF and double stem-loop immunomodulating oligodeoxyribonucleotides.	Kidney cancer	Compassionate use in India
Center of Molecular Immunology (Cuba)	CimaVax EGF	EGF conjugated to rP64k	Lung cancer	Cuba, Peru

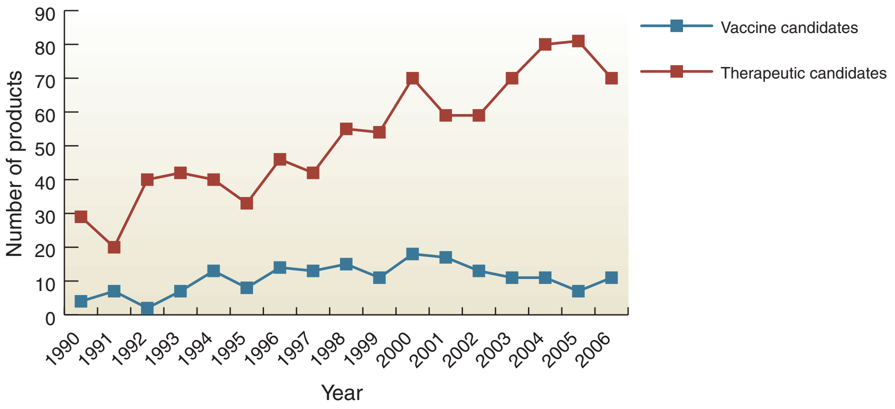


Figure 4 New cancer therapeutics and vaccines entering clinical study per year from 1990 to 2006. (Source: Tufts Center for the Study of Drug Development)

commercially viable, although the evidence to date suggests that it may not be a clinically viable approach.

Even non-cell-based personalized vaccines raise a host of regulatory issues. Is each vaccine a different product? Do vaccines produced for different patients have different immunogenicities? Different cross-reactivities? Different potencies?

The answer might rest in finding more shared, but tumor-specific, antigens, which are in the minority among products in trials today. And down the road, advances in related fields might provide cancer vaccinologists with the tools they need to create off-the-shelf vaccines. For example, Geron, which has an autologous vaccine in trials now, is using this as a proof of concept according to Okarma. The company also has in place the technology for making dendritic cells from stem cells, which would enable the company to prepare an off-the-shelf, activated dendritic cell, something not available at present.

Keith Wonnacott, chief of FDA's cell therapies, joins the chorus of immunologists and academics optimistic that some cancer vaccine will succeed. "We anticipate success, and that lessons will be learned. Much of what has gone on has been helpful. We would love to see success," he says. Whether that optimism is justified, only time will tell.

1. Choudhury, A. *et al. Adv. Cancer Res.* **95**, 147–202 (2006).
2. Valone, FH, *et al. Cancer J.*, **7**, S53–61 (2001).
3. Small, E.J. *et al. J. Clin. Oncol.* **18**, 3894–3903 (2000).
4. Allison, M. *Nat. Biotechnol.* **26**, 967–969 (2008).
5. Testori, A. *et al. J. Clin. Oncol.* **26**, 955–962 (2008).
6. Wood, C. *et al. Lancet* **372**, 145–154 (2008).

Proprietary science, open science and the role of patent disclosure: the case of zinc-finger proteins

Subhashini Chandrasekharan, Sapna Kumar, Cory M Valley & Arti Rai

A closer look at the large patent estate now covering both the engineering and use of zinc-finger proteins.

Recent advances in the ability to engineer customized zinc-finger proteins (ZFPs), which can bind virtually any DNA sequence of interest, have generated excitement among both academic and industrial researchers. Engineered ZFPs can be used to alter chromatin structure, regulate endogenous gene expression levels, and introduce targeted modifications in genes. In one salient case, a chimeric zinc finger–nuclease (ZFN) successfully stimulated homologous recombination and thus repaired a mutant IL2Rγ (*IL2RG*) gene associated with X-linked severe combined immune deficiency (SCID)[1]. ZFP-based therapeutics developed by Sangamo Biosciences for diabetic neuropathy and peripheral arterial disease are undergoing phase 1 and 2 clinical trials[2], and a ZFN-mediated approach for disrupting the CCR5 receptor in patient T cells as a strategy to increase resistance to HIV is in preclinical development[3]. These advances have given researchers hope that ZFP- and ZFN-based approaches may help improve both the efficiency and the precision of gene therapy. Other potential commercial applications for ZFPs include plant genetic engineering, the production of biopharmaceutical molecules such as growth factors and antibodies, and the nascent field of synthetic biology. ZFN technology has also been used successfully to make targeted gene modifications in several model organisms such as *Drosophila*[4,5], *C. elegans*[6], plants[7,8] and most recently zebrafish[9,10], illustrating the range of uses for ZFNs in basic research as powerful molecular biology tools.

As might be expected with any research platform that has many potential commercial uses, a large patent estate now covers both the engineering and the use of ZFPs. Notably, the patent estate was initially owned by several different companies and academic institutions, thereby creating the possibility that subsequent users and developers would face prohibitive costs in negotiating multiple licenses—the classic scenario of a patent "anticommons"[11]. However, one company, Sangamo, has now consolidated the majority of this patent estate. The dominant patent position held by Sangamo has raised the recurrent question of whether a company's monopoly control over an important and versatile research platform will ultimately help or hinder optimal development of that platform. Because such development can occur within both the private and public sectors, there is also the subsidiary issue of whether patents will be enforced against academic researchers in the same manner as they might be enforced against private-sector competitors.

Previous studies[12] suggest that academic researchers do not seem concerned about being sued by private-sector patentees. For example, a survey of academic biomedical researchers found that only 5% report checking for patents related to their research[13]. These studies further indicate that private-sector patent owners practice "rational forbearance" and do not sue academic researchers because of the difficulties and disadvantages of asserting patent rights in such circumstances[14,15]. Currently, the conventional view is that academic biomedical research is more likely to be impeded by lack of access to privately held research inputs such as materials, data and know-how than by patents[12,13,16,17].

To explore the impact of ZFP patents, and specifically Sangamo's dominant patent position, on academic and commercial research and development, we systematically created a map of existing patents in the ZFP arena, presented here for the first time. We also conducted interviews with academic researchers in the field to develop a nuanced understanding of the complex interactions between private and public ZFP research endeavors. Our findings are consistent with the view that, for academics, lack of access to information and materials is a greater problem than the threat of patent lawsuits. However, because some of the access problems would have been alleviated if statutory obligations regarding patent disclosure had been met, our research also suggests the heretofore unrecognized possibility of an overlap between patents and access to information and research materials. More complete patent disclosure might also have obviated the need to generate various open-science alternatives to the Sangamo platform.

The ZFP/ZFN intellectual property landscape

Using a keyword-based search query (**Fig. 1**), we determined that the number of ZFP-related patents granted in the United States increased steadily from 1997 to 2001, with four patents granted in 1997 and 26 granted in 2001. Since 2001, the numbers of patents issued each year has remained fairly constant, and to date the largest number of patents (28) was granted in 2006 (**Fig. 1**). The search query similarly identified 189 pending US applications for the same time period (data not shown). Sangamo Biosciences is the single largest owner of issued

Subhashini Chandrasekharan and Arti Rai are at the Center for Public Genomics, Center for Genome Ethics, Law & Policy, Institute for Genome Sciences & Policy, Duke University, Durham, North Carolina 27708, USA; Arti Rai is also at Duke University School of Law, Durham, North Carolina 27708, USA; Sapna Kumar is at the Chambers of the Hon. Kenneth Ripple, US Court of Appeals for the Seventh Circuit, Chicago, Illinois 60604, USA; and Cory M. Valley is at the University of Maryland School of Law, Baltimore, Maryland 21201, USA.
e-mail: rai@law.duke.edu

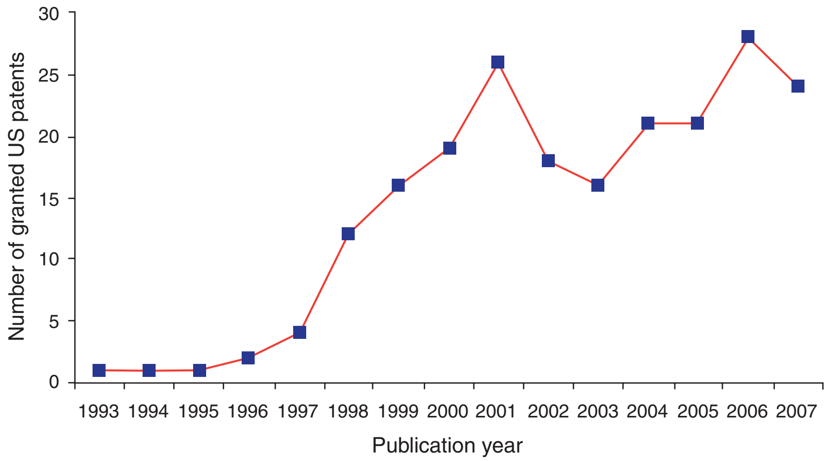


Figure 1 US ZFP patents granted in 1993–2007. Using Delphion analysis tools, we queried the USPTO database with the following search algorithm: (((zinc finger protein) <in> (TITLE,ABSTRACT,CLAIMS)) *OR* ((ZFP) <in> (TITLE,ABSTRACT,CLAIMS)) *OR* ((Zinc finger) <in> (TITLE,ABSTRACT,CLAIMS)) *OR* ((zinc finger binding protein) <in> (TITLE,ABSTRACT,CLAIMS))). The query was designed to capture any patent containing one or more of the search terms in the "Title" or the "Abstract" or the "Claims" fields. Search terms were selected from keywords specific to ZFPs that frequently appear in a subset of relevant patents (for example, patents owned by Sangamo) and in published articles. Claims-based searches are important to reduce noise, as they avoid terms found only in the description (specification) section of the patent. The claims define the "metes and bounds" of the invention, whereas the description often uses particular terms in the context of providing general background information. All US patents issued on or before 31 December 2007 were included.

US patents on ZFPs (42 patents). But a number of other institutions are also well represented; for example, the Massachusetts Institute of Technology (MIT) owns 13 patents and the Scripps Institute owns 9 patents (**Fig. 2**).

From the pool of patents generated by our search query, patents that directly pertain to the engineering and use of engineered zinc-finger proteins were identified through analysis of the claims (**Supplementary Table 1** online). The 42 patents owned by Sangamo include 8 patents on rules and libraries for constructing sets of 'two-zinc-finger' domains, each of which can bind to a specific sequence of six nucleotides. These were previously owned by UK-based Gendaq Ltd., which was acquired by Sangamo in July 2001 (ref. 18).

Sangamo has also actively licensed intellectual property (IP) from a number of academic institutions. This IP includes five patents from MIT, three from the Scripps Institute, two from Harvard University and six from Johns Hopkins University (JHU)[18]. The patents licensed from MIT, Scripps and Harvard and two of the six patents licensed from JHU are a subset of the patents listed in **Supplementary Table 1**. An additional four patents licensed from JHU relate to ZFN technology (**Supplementary Table 2** online). Sangamo acquired these technologies from MIT, Harvard, the California Institute of Technology (Caltech) and JHU under worldwide exclusive licenses for all fields of use, including the rights to sublicense[18]. The only exception to this pattern is for patents licensed from the Scripps Institute, where the licenses exclude Sangamo from specific fields of use, including diagnostics, therapeutics and genetic engineering in plants[18]. Thus, although initial ownership of ZFP-related patents was dispersed, creating the potential for high transaction costs and anticommons effects, Sangamo's energetic acquisition and licensing activity has consolidated many of the requisite patent rights.

Four patents on the engineering and design of the *Fok*I endonuclease, which is used to generate designer ZFNs, have been licensed from JHU (**Supplementary Table 2**). Sangamo Biosciences recently obtained exclusive rights to related technologies for genetic engineering and gene modification using ZFNs from the University of Utah, US Application no. US20050208489A1: *Targeted chromosomal mutagenesis using zinc finger nucleases*, and from Caltech, US Application no. US20050026157A1: *Use of chimeric nucleases to stimulate gene targeting*[18]. Sangamo also purchased ZFN-related IP from STELL Inc. in 2004, US20030232410A1: *Methods and compositions for using zinc finger endonucleases to enhance homologous recombination*[19]. Assuming that these applications are granted, Sangamo will have consolidated key IP surrounding the use of ZFNs for gene correction and gene repair in the United States.

An analysis of the different categories of patents (**Supplementary Tables 1** and **2**) reveals that at least 24 of the 55 patents owned by or licensed exclusively to Sangamo cover technologies for the design, selection and optimization of engineered ZFPs. Our analysis also indicates that several patents owned by Sangamo are foundational for the ZFP field, with limited possibilities for a 'workaround'. Perhaps most salient is a trio of patents (US Patent nos. 71777766, 6785613 and 6453242) that broadly claim the dominant 'modular' strategy for ZFP design (at least with respect to three-finger ZFPs that bind to sequences containing nine nucleotides). This modular strategy relies on assembling a multifinger protein from individual zinc-finger modules where each module has been determined to bind specifically to a particular three-nucleotide subunit and, ideally, to the subunit as further specified by its location within the sequence of nine nucleotides. Also significant is US Patent no. 6794136, which covers "iterative optimization in the design of binding proteins": this patent broadly covers methods for further improving binding specificity once a ZFP candidate for a particular nucleotide sequence has been identified.

More than three-quarters of the patents owned by or licensed to Sangamo (44 of 55) concern inventions that could be categorized as research methods and tools, with 24 patents covering methods for the design and selection of ZFPs and another 20 patents covering methods to regulate or modify endogenous gene expression using engineered ZFPs and/or ZFP transcription factors. The earliest issued patent in this set will not expire until 2018, making it unlikely that academic or commercial researchers will be able to wait for the technologies to pass into the public domain.

Impact on commercial R&D

The ZFP patent landscape that we have created confirms Sangamo's dominant position in ownership of patents covering relevant research tools and methods, including foundational patents on enabling technologies. This position could have at least two benefits. First, a dominant patent position facilitates Sangamo's ability to attract private capital[20]. Given Sangamo's considerable R&D expenses[21] and lack of marketable products, this private capital is necessary even though Sangamo has also received some federal funding, including two grants totaling nearly $4 million from National Institute of Standards and Technology. Not surprisingly, Sangamo executives have repeatedly stated that a strong patent portfolio has been vital to the company's success[22,23]. Second, as mentioned earlier, Sangamo's consolidation of relevant IP rights may ease negotiation cost burdens for commercial entities that want to work in this area, as they will have to negotiate licenses with only one institution instead of several. Such licensing negotiations may be an option that Sangamo actively seeks. Economic theory would suggest that a rational, profit-maximizing monopolist that cannot develop a platform by itself in certain areas of application will often be inclined to license, so as to promote development in those areas by others[24,25]. Collaboration and licensing might be particularly desirable for a small company such as Sangamo that has limited capacity to pursue in-house development for all possible applications of its technology.

However, economic theory has also identified a variety of situations in which increased negotiation costs in concluding licensing deals, as well as other distortions, could impede a monopolist's optimal deployment of a research platform[24]. As an empirical matter, the historical record shows that patents that conferred monopoly control over foundational technologies in the aircraft and automobile industries impeded development[26].

Sangamo's out-licensing strategies provide support for both the optimistic and the pessimistic views of monopoly control. For application areas outside Sangamo's main focus on ZFP-based medical therapeutics, the company has granted several companies access to its IP. For example, through its "Enabling Technology Program," Sangamo has longstanding collaborations with Pfizer, Amgen and NovoNordisk for more efficient pharmaceutical production of proteins[18]. More recently, Sangamo granted Dow AgroSciences exclusive rights (including sublicensing rights) to ZFP and ZFN technologies for modifying plant genomes and altering plant gene expression[2].

In contrast, several reports indicate that the inability to conclude a licensing arrangement with Sangamo played a crucial role in the failure of the plant biotechnology start-up Phytodyne, founded by researchers at Iowa State University. Phytodyne received significant venture capital investment and financial support from the state of Iowa and was developing plant genetic engineering applications viewed as highly promising by the industry. It is difficult to ascertain the long-term impact of this failure on innovation in plant genetic engineering, particularly because Dow is now actively engaged in similar R&D. However, to the extent that small enterprises such as Phytodyne may be better positioned to pursue breakthrough innovation than larger firms like Dow[27,28], this example illustrates the potential negative effects of patent monopolies.

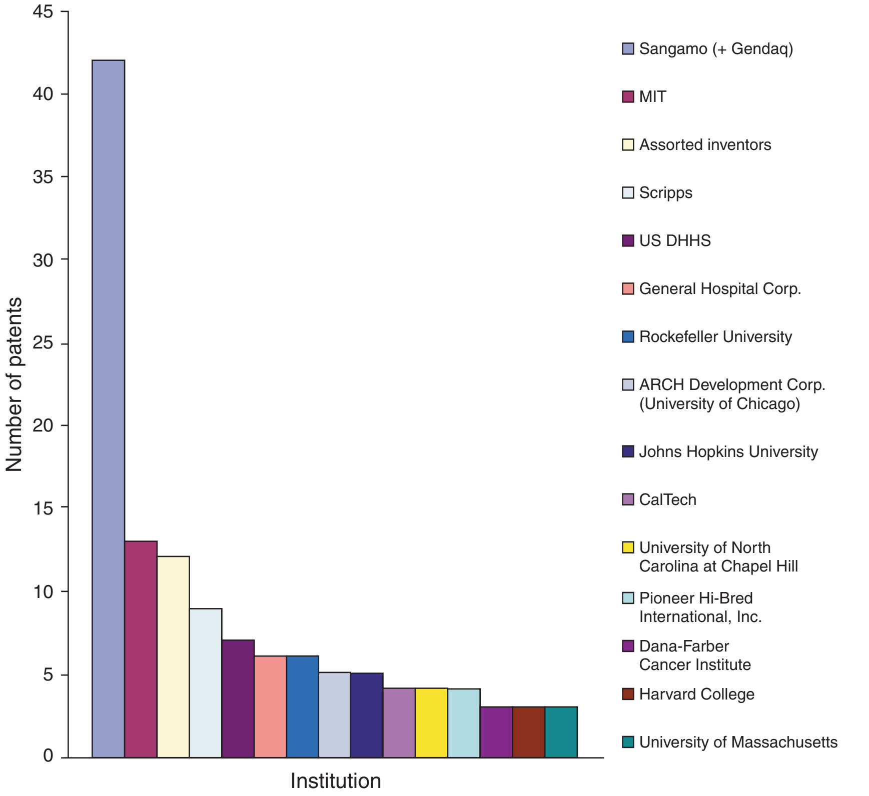


Figure 2 Ownership (assignees) of US ZFP patents by institution, 1993–2007. Institutions with three or more US ZFP patents are shown. Data are complete as of 31 December 2007.

Impact on academic research

Academia provides an important venue for improvement of research platforms, in addition to the commercial sector. As noted earlier, survey research indicates that, with respect to such platforms, academic scientists routinely ignore patents, and private-sector patentees correspondingly refrain from enforcing their patents[12,16,17]. To determine whether Sangamo patents were impeding academic research and, if so, to what extent, we interviewed a number of prominent ZFP researchers, including researchers who have licensed patents to and collaborate with Sangamo. Academic scientists indicated that they routinely used patented technologies owned by Sangamo without securing a license. Thus, consistent with prior work, we found that ZFP researchers engage in infringement under the expectation that Sangamo will refrain from suing academics.

Several scientists did, however, express concern about lack of access to Sangamo's ZFPs and ZFNs. Researchers would like to collaborate with Sangamo because it possesses a platform capable of engineering ZFPs for many triplet nucleotide sequences as well as the information necessary for performing further optimization that is sometimes required to obtain high-specificity ZFPs and ZFNs. Sangamo does not disclose detailed information about this proprietary platform. Additionally, although Sangamo has signed material transfer agreements with several academic research groups to provide ZFPs and/or ZFNs, it appears to be highly selective in its choice of collaborators[22].

Sangamo recently entered an agreement with Sigma-Aldrich under which Sigma will use Sangamo's technology platform to provide ZFP and ZFN reagents that bind any DNA sequence in which a researcher is interested[2,21]. Although this agreement is likely to improve academic researchers' access to Sangamo's highly specific ZFPs (at least to the extent that researchers can afford to pay Sigma's $25,000 fee), researchers will still be unable to access Sangamo's platform directly.

The role of patent disclosure

Sangamo's unwillingness to disclose proprietary know-how about its platform is not unusual—secrecy is a routine competitive strategy in the commercial sector. More problematic is the strong possibility that at least part of this proprietary information should, under standard doctrines of patent disclosure, be disclosed in the Sangamo patents themselves. Patent law requires that a patent teach a "person having ordinary skill in the art" how to practice the claimed invention. According to several ZFP scientists with whom we spoke, actually practicing the trio of foundational patents that cover the design of "specific" three-finger proteins would require access to Sangamo's proprietary database or 'rule set' on matching ZFP modules with particular three-base DNA subunits. These Sangamo patents do not, however, disclose any such database or rule set. Thus, in this case, even though these patents are not being asserted against scientists, they confer 'practical excludability' because they do not meet the statutory obligation of enabling scientists to practice the inventions that the patents cover[13].

The Sangamo case study also highlights the fact that patents and access to tangible materials and know-how, which are thought of as two distinct problems, might actually overlap in interesting ways. If the 'patent bargain' of exclusivity in exchange for disclosure were being satisfied, problems encountered by academics over access to physical materials and data might be alleviated. The patent disclosure would provide at least some of the information not disclosed by scientific publication that is necessary to make such materials independently. This is especially salient because academic researchers report that a major reason for not making research materials independently is "inability" to do so, due to lack of equipment, information or expertise[12,13,17]. Improving patent disclosure would not resolve the problem that, absent a formal research exemption from infringement liability in patent law, using the statutorily required patent disclosure to make or practice the invention for academic research would technically constitute willful infringement. However, given the reluctance of companies to sue academic researchers, concerns about infringement may be more hypothetical than real.

Unfortunately, problems associated with inadequate patent disclosure in biotechnology are likely to get worse rather than better. Even if it is enforced incompletely[29], the high standard of disclosure for DNA sequence patents has historically made disclosure in biotechnology better than in other areas. However, as biotechnology begins to look more like information technology, with the ZFP databases and design rule sets providing one illustration of this trend, the notoriously poor disclosure standards associated with information technology may be poised to infiltrate biotechnology[30]. Notably, as many commentators have pointed out, the case law that governs information technology patents often allows broad, vague claims that are unsupported by adequate disclosure[31,32].

Policing the patent bargain of exclusivity in exchange for appropriate disclosure should be the function of the US Patent & Trademark Office (USPTO). But given the high volume of pending patent applications and the rapidly changing state of the art, especially in biotechnology, developing mechanisms by which experts outside the USPTO could help flag problems of underdisclosure (either during the examination process or post-grant) would be a welcome improvement. Whether academic researchers would be inclined to participate in such mechanisms is not clear. Because academic scientists largely rely on peer-reviewed publications rather than patent disclosures for know-how, and rarely experience patents as threatening or impeding their research activities, there may be little incentive for the academic community to engage in such an outside review process.

Open-science alternatives

The Zinc Finger Consortium, a prominent academic program founded by ZFP researchers J. Keith Joung and Dan Voytas, was created in part to address concerns about access to materials and Sangamo's proprietary databases[33]. Two web-based tools for identifying potential ZFP target sites in DNA sequences are also freely available, Zinc Finger Tools[34], developed by Carlos Barbas's team at the Scripps Research Institute, and a second program, Zinc Finger Targeter (ZiFiT), designed by members of the Consortium[35]. The Consortium has also generated an archive of plasmids encoding over 140 zinc-finger modules (derived from publicly available archives of zinc fingers) that bind specific nucleotide triplets. The plasmids are made available to all interested academic researchers via the nonprofit distribution service AddGene. These various finger modules have been reported to bind to many ANN and GNN triplets and to CNN and TNN triplets to a lesser degree. These zinc-finger modules appear to infringe various Sangamo patents, but nevertheless Sangamo has not blocked their distribution for research purposes. Reagent availability through the Consortium is subject to a vaguely worded licensing agreement stating that certain uses of the zinc-finger modules requires a license from Sangamo[36]. But the extent to which Sangamo attempts to enforce this clause is unclear. Recent work from Consortium labs[37] has furthermore demonstrated that the efficacy rate for engineering ZFPs using these modules is significantly lower than the more robust rates originally reported in the literature by other groups[38,39]. Thus, it may be that actual enforcement against academic or commercial users of Consortium modules is unnecessary because most commercial applications would be likely to require the higher-efficiency ZFPs produced by Sangamo.

In July 2008, Keith Joung and his colleagues improved on prior Consortium technology by reporting a novel and robust method for generating custom ZFNs with activities superior to those produced by the previously standard modular design approach and with activities and toxicities comparable to those of an optimized ZFN produced by the proprietary Sangamo method[33,40]. The presence of roughly comparable proprietary and open-science alternatives may produce a productive tension resembling the competition between the public and private human genome sequencing endeavors[41]. Alternately, it may result in peaceful coexistence of the two platforms, as illustrated by the diffusion of microarray technologies. Open approaches for disseminating 'spotted glass' microarray technology pioneered by Pat Brown and colleagues in the early 1990s aimed to offer academic researchers a lower-price alternative to Affymetrix's costly microarrays[42,43]. Although Affymetrix sued commercial developers of spotted microarray technology, it never asserted its IP rights against academic users[43,44]. A decade later, both platforms continue to be widely used in academic research. An early response from Sangamo suggests that it does not perceive OPEN as a major challenge. Indeed, Sangamo has indicated that the coexistence of the open-science alternative may even be favorable to its position, as having more academic scientists performing ZFP-based research may enhance the value of the company[21]. With Sangamo's patents broadly covering uses such as regulation of gene expression in different organisms, commercial development of downstream applications would almost always require rights to use IP controlled by Sangamo.

The reagents associated with the OPEN platform will be made publicly available to academic researchers at a price of approximately $5,000 a set[21,45]. Not only will this be more affordable to academic researchers than the $25,000 charged by Sangamo/Sigma, but the availability of OPEN reagents may eventually provide sufficient competition to cause a reduction in the price of the Sangamo/Sigma reagents.

Conclusions

Sangamo's strategic acquisition of patents has given the company a powerful monopoly over an important platform technology. As economic theory would predict, Sangamo has often (but not always) licensed its platform technology in a manner that is both profit maximizing and likely to enhance social benefit. To date, Sangamo has also tolerated an open-science alternative to its proprietary platform. The coexistence of open and proprietary alternatives may be productive or, at a minimum, peaceful.

Two features of the ZFP/ZFN case are particularly noteworthy. First, because of problems with patent disclosure, patents may effectively be posing a barrier to academic research in this field. Second, resolving deficiencies in patent disclosure could mitigate the problem of academic access to physical materials and know-how, perhaps even obviating the need to develop open-science alternatives. Thus our study raises the possibility that even when academics are not defendants in patent suits, and enjoy a *de facto* (if not *de jure*) exemption from patent infringement liability, the patent system may nonetheless be failing to fulfill the constitutional mandate that patents "promote the progress of…the useful Arts."

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

ACKNOWLEDGMENTS
The authors would like thank R. Cook-Deegan and K. Joung for helpful comments and criticisms. The authors gratefully acknowledge the support of the National Human Genome Research Institute and the Department of Energy (CEER Grant P50 HG003391, Duke University Center of Excellence for ELSI Research.

1. Urnov, F.D. *et al. Nature* **435**, 646–651 (2005).
2. Sangamo Biosciences Annual SEC filing Form 10K 2008. http://www.secinfo.com/d14D5a.t1bzr.d.htm (accessed 23 January 2009).
3. Perez, E.E. *et al. Nat. Biotechnol.* **26**, 808–816 (2008).
4. Beumer, K., Bhattacharyya, G., Bibikova, M., Trautman, J.K. & Carroll, D. *Genetics* **172**, 2391–2403 (2006).
5. Bibikova, M., Beumer, K., Trautman, J.K. & Carroll, D. *Science* **300**, 764 (2003).
6. Morton, J., Davis, M.W., Jorgensen, E.M. & Carroll, D. *Proc. Natl. Acad. Sci. USA* **103**, 16370–16375 (2006).
7. Wright, D.A. *et al. Plant J.* **44**, 693–705 (2005).
8. Lloyd, A., Plaisier, C.L., Carroll, D. & Drews, G.N. *Proc. Natl. Acad. Sci. USA* **102**, 2232–2237 (2005).
9. Meng, X., Noyes, M.B., Zhu, L.J., Lawson, N.D. & Wolfe, S.A. *Nat. Biotechnol.* **26**, 695–701 (2008).
10. Doyon, Y. *et al. Nat. Biotechnol.* **26**, 702–708 (2008).
11. Heller, M.A. & Eisenberg, R.S. *Science* **280**, 698–701 (1998).
12. Walsh, J.P., Cho, C. & Cohen, W.M. *Res. Policy* **36**, 1184–1203 (2007).
13. Cohen, W.M. & Walsh, J.P. in *Innovation Policy and the Economy* Vol. 8 (Jaffe, A.B., Lerner, J. & Stern, S., eds.) 1–30 (University of Chicago Press, Chicago, 2007).
14. Fore, J., Jr., Wiechers, I.R. & Cook-Deegan, R. *J. Biomed. Discov. Collab.* **1**, 7 (2006).
15. Pressman, L. *et al. Nat. Biotechnol.* **24**, 31–39 (2006).
16. Campbell, E.G. *et al. J. Am. Med. Assoc.* **287**, 473–480 (2002).
17. Walsh, J.P., Cho, C. & Cohen, W.M. *Science* **309**, 2002–2003 (2005).
18. Sangamo Biosciences Annual SEC filing Form 10K 2007. http://www.secinfo.com/dsvrp.u4fp.htm (accessed 23 January 2009).
19. Sangamo Biosciences Quarterly SEC filing Form10Q 2004. http://google.brand.edgar-online.com/displayfilinginfo.aspx?FilingID=3125032-801-132173&type=sect&TabIndex=2&companyid=72223&ppu=%252fdefault.aspx%253fsym%253dSGMO (accessed 23 January 2009).
20. Mann, R. & Sager, T. *Res. Policy* **36**, 193–208 (2007).
21. Pearson, H. *Nature* **455**, 160 (2008).
22. Scott, C.T. *Nat. Biotechnol.* **23**, 915–918 (2005).
23. Kaiser, J. *Science* **310**, 1894–1896 (2005).
24. Farrell, J. & Weiser, P. *Harv. J. Law Technol.* **17**, 85–134 (2003).
25. Kitch, E. *J. Law Econ.* **20**, 265–290 (1977).
26. Merges, R. & Nelson, R. *Columbia Law Rev.* **90**, 839–916 (1990).
27. Acs, Z. & Audretsch, D. *Innovation and Small Firms* (MIT Press, Cambridge, Massachusetts, USA, 1990).
28. Audrestsch, D. *Innovation and Industry Evolution* (MIT Press, Cambridge, Massachusetts, USA, 1995).
29. Holman, C. *Albany Law J. Sci. Technol.* **17**, 1–85 (2007).
30. Rai, A. & Boyle, J. *PLoS Biol.* **5**, e58 (2007).
31. Bessen, J. & Meurer, M. *Patent Failure* (Princeton University Press, Princeton, New Jersey, USA, 2008).
32. Burk, D. & Lemley, M. *Berkeley Technol. Law J.* 17, 1155–1206 (2002).
33. Wright, D.A. *et al. Nat. Protoc.* **1**, 1637–1652 (2006).
34. Mandell, J.G. & Barbas, C.F. III. *Nucleic Acids Res.* **34** (Web Server issue), W516–W523 (2006).
35. Sander, J.D., Zaback, P., Joung, J.K., Voytas, D.F. & Dobbs, D. *Nucleic Acids Res.* **35** (Web Server issue) W599–605 (2007).
36. Zinc Finger Consortium Sets from AddGene (AddGene Licenses for Plasmids with Sangamo Zinc-Finger Technology) http://www.addgene.org/pgvec1?f=a&cmd=showfile&file=sangamo (accessed 23 January 2009).
37. Ramirez, C.L. *et al. Nat. Methods* **5**, 374–375 (2008).
38. Segal, D.J. *et al. Biochemistry* **42**, 2137–2148 (2003).
39. Bae, K.H. *et al. Nat. Biotechnol.* **2**, 275–280 (2003).
40. Maeder, M.L. *et al. Mol. Cell* **31**, 294–301 (2008).
41. Eisenberg, R.S. & Nelson, R.R. *Acad. Med.* **77**, 1392–1399 (2002).
42. Hager, J. *Methods Enzymol.* **410**, 135–168 (2006).
43. Fox, J.L. *Nat. Biotechnol.* **17**, 325–326 (1999).
44. Rouse, R. & Hardiman, G. *Pharmacogenomics* **4**, 623–632 (2003).
45. Kaiser, J. *ScienceNOW Daily News* (24 July 2008).

Corrected after print 9 February 2009.

Erratum: Proprietary science, open science and the role of patent disclosure: the case of zinc-finger proteins

Subhashini Chandrasekharan, Sapna Kumar, Cory M Valley & Arti Rai

***Nat. Biotechnol.* 27, 140–144 (2009); published online 9 February 2009; corrected after print 9 February 2009**

In the version of this article published in print, the second affiliation for Arti Rai was inadvertently inserted into the middle of the affiliation for Sapna Kumar. The two affiliations should have read as follows: "Arti Rai is also at Duke University School of Law, Durham, North Carolina 27708, USA; Sapna Kumar is at the Chambers of the Hon. Kenneth Ripple, US Court of Appeals for the Seventh Circuit, Chicago, Illinois 60604, USA;". The error has been corrected in the HTML and PDF versions of the article.

Recent patent applications in gene expression

Patent number	Description	Assignee	Inventor	Priority application date	Publication date
WO 2008148858	A method of diagnosing or prognosing HIV-related diseases comprises collection of a blood sample from a subject, isolation of the monocytes from this blood sample and determination of gene expression in the monocytes.	Institute of Tropical Medicine (Antwerp, Belgium), Free University of Brussels (Brussels), VIB (Ghent, Belgium)	de Baetselier P, Raes G, van den Bergh R, Vanham G	6/8/2007	12/11/2008
WO 2008150884	A new regulatable gene expression construct for affecting the processing of RNA comprises a nucleic acid molecule encoding an RNA comprising a riboswitch operably linked to a coding region.	Yale University (New Haven, CT, USA)	Breaker RR, Wachter A	5/29/2007	12/11/2008
WO 2008148115	Evaluating multiple sclerosis (MS) in a patient comprises determining a gene expression profile for a blood sample of a patient, comparing the gene expression profile and classifying gene expression profile as MS profile or non-MS profile.	Ore Pharmaceuticals (Gaithersburg, MD, USA)	Bigwood D, Eastman E, Kaldjian E	5/25/2007	12/4/2008
WO 2008141682	A method of preparing oligonucleotides as probes useful in gene expression analysis; involves providing a hydroxyl-containing compound, preparing the phosphitylated compound in the presence of a first activator and reacting it in a second activator without isolation.	Girindus (Bensberg, Germany)	Groessel O, Hohfeld A, Kirchhoff C, Lange M, Schoenberger A	5/22/2007	11/27/2008
US 20080295202	A new isolated polynucleotide comprising a promoter sequence or coding sequence for soybean SC194 protein; useful for gene expression and for altering marketable flower traits, such as color, morphology and fragrance in flowering plants.	Li Z	Li Z	5/17/2007	11/27/2008
US 20080295201	A new polynucleotide comprising a promoter or coding sequence for lipid transfer protein 2 (LTP2); useful for regulating gene expression and altering marketable flower traits such as color, morphology and fragrance in flowering plants.	Li Z	Li Z	5/17/2007	11/27/2008
US 20080293164	An assay method comprising providing a sample containing a target biomolecule, providing a sensor protein conjugated to a signaling chromophore, providing a conjugated polymer and applying a light source; useful, e.g., for detecting gene expression.	Sirigen (San Diego)	Baldocchi R, Fu T, Gaylord BS, Hong JW, Sun C	10/6/2006	11/27/2008
WO 2008140334	A new isolated promoter polynucleotide comprising at least two specific sequence motifs; useful for controlling transcription of operably linked polynucleotides in plants for expressing pharmaceutical products and desired phenotypes.	Allan AC, Chagne D, Espley R, Hellens RP	Allan AC, Chagne D, Espley R, Hellens RP	5/11/2007	11/20/2008
WO 2008137090, US 20080286273	A method of predicting patient response to cancer treatment, comprising measuring in a biological sample from a patient the levels of gene expression and correlating the signature score with a predicted response to cancer treatment.	Siemens Medical Solutions USA (Malvern, PA, USA)	Buffa FM, Harris AL, Krishnan S, Krishnapuram B, Lambin P, Nuyten D, Nuyten DSA, Rao RB, Seigneuric RG, Starmans M, Starmans MHW, Steck H, Wouters BG	5/2/2007	11/13/2008, 11/20/2008
WO 2008136971	A method of diagnosing whether a human subject has, or is at risk for, developing pancreatic cancer, by detecting the level of expression of miR gene products from a tissue sample and comparing the gene expression detected to a database comprising part of the data.	Ohio State University Research Foundation (Columbus, OH, USA)	Croce CM	4/30/2007	11/13/2008
WO 2008136902	A new isolated double-stranded nucleic acid for reducing expression of a target gene in a mammalian cell.	City of Hope (Duarte, CA, USA), Integrated DNA Technologies (Coralville, IA, USA)	Behlke MA, Kim D, Rossi JJ	5/1/2007	11/13/2008

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).

Year of the ox

Yann Echelard

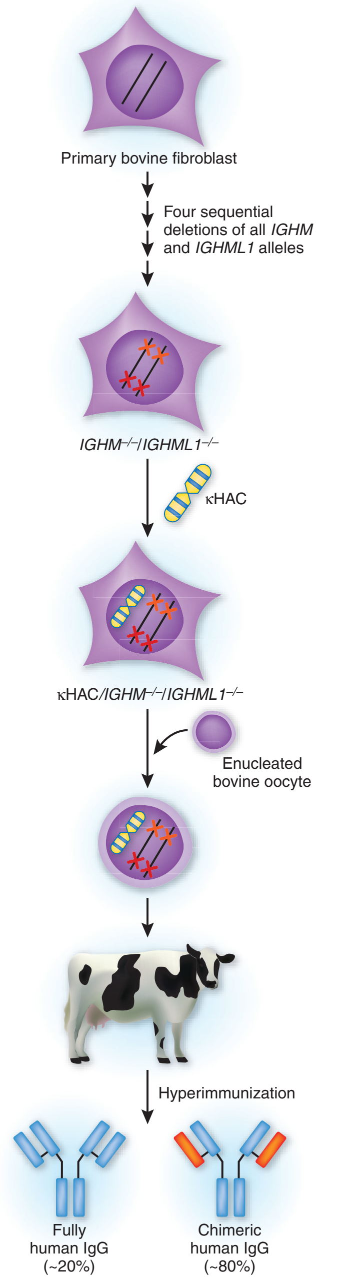
High levels of human polyclonal antibodies have been produced in a transgenic large animal.

With a market now worth well over $2 billion in the United States[1], human polyclonal antibodies purified from thousands of plasma pools have become standard therapy for many viral infections and immune disorders[2] and for neutralization of toxins[3]. Despite their clinical potential, however, the use of polyclonal antibodies remains limited by issues related to their supply, cost and safety[4]. In this issue, Kuroiwa *et al.*[5] bring us a step closer to large-scale production of relatively homogenous recombinant polyclonal antibodies, which could alleviate these problems and expand application of this therapeutic modality to new indications[6].

Unlike monoclonal antibodies, which recognize a single epitope, polyclonal antibody preparations bind multiple epitopes on the disease-causing agent and can thereby neutralize distinct variants of toxins or infectious particles, making them the agents of choice for treating certain medical emergencies and acute illnesses. Hyperimmune globulins—sourced from human or animal donors with high titers of antibodies against specific antigens—are in high demand to curb immunosuppression associated with transplants; prevent Rh hemolytic disease; treat and prevent infections such as hepatitis B, hepatitis A, rabies, respiratory syncytial virus, cytomegalovirus and varicella-zoster; and neutralize toxins, including diphtheria, botulism, digoxin and snake and spider toxins[2,3].

The ability to produce human antibodies in mice expressing human immunoglobulin genes has long been appreciated[7], and mice now provide a convenient source of hybridomas for generating candidate therapeutic human monoclonal antibodies. However, the small body size of mice makes them unsuitable for synthesizing large amounts of hyperimmune globulins. Aiming to translate this approach to a large animal, Kuroiwa and colleagues previously expressed the human immunoglobulin heavy chain and λ-light chain from a human artificial chromosome in cloned cows[8]. Although these animals did produce human antibodies, the levels were too low to be of practical utility as the active endogenous immunoglobulin loci suppressed expression of the human genes[6,8].

The feasibility of introducing human immunoglobulin genes and knocking out endogenous immunoglobulin genes in cattle has been far from certain. First, it is not straightforward to perform multiple genetic modifications in large animals: embryonic stem cell lines are not available and generation intervals are long (around three years in cattle). Successive rounds of transfection and selection in primary cells, which have a limited life span, each followed by somatic cell nuclear transfer (to regenerate the cell line) are necessary to introduce the targeting constructs and the human immunoglobulin loci. Second, the accumulation of epigenetic errors caused by successive rounds of nuclear transfer has been reported to compromise the viability of offspring. Finally,


Primary bovine fibroblast
Four sequential deletions of all *IGHM* and *IGHML1* alleles
$IGHM^{-/-}/IGHML1^{-/-}$
κHAC
κHAC/$IGHM^{-/-}/IGHML1^{-/-}$
Enucleated bovine oocyte
Hyperimmunization
Fully human IgG (~20%)
Chimeric human IgG (~80%)

Kim Caesar

Figure 1 Cattle capable of producing human polyclonal antibodies are produced by multiple cycles of transfection, selection and nuclear transfer[5]. Each of the four bovine IgM heavy chain alleles is knocked out by homologous recombination followed by somatic cell nuclear transfer to extend the life span of the selected cell lines. An artificial chromosome carrying the human immunoglobulin heavy and κ-light chain loci (κHAC) is transferred to the multitargeted bovine cell lines by microcell-mediated chromosome transfer, and three additional nuclear transfer steps are performed to obtain a healthy transgenic calf with the κHAC/$IGHM^{-/-}/IGHML1^{-/-}$ genotype. Vaccination of this animal with an antigen of interest produces ~20% fully human antibodies and ~80% chimeric antibodies (bearing human heavy chains and bovine light chains).

Yann Echelard is at GTC Biotherapeutics, Inc., 5 Mountain Road, Framingham, Massachusetts 02130, USA.
e-mail: yann.echelard@GTC-bio.com

it has been unclear whether human immunoglobulins alone could support bovine humoral immunity in the absence of endogenous bovine immunoglobulins.

These formidable challenges make the achievement of a transgenic calf expressing high levels of human antibodies all the more remarkable. Kuroiwa *et al.*[5] focused first on successively inactivating all the bovine IgM heavy chain genes (**Fig. 1**), whose expression is essential for B-cell development. Because ruminants, unlike mice and humans, have two functional IgM loci (*IGHM* and *IGHML1*), four alleles had to be targeted to knock out IgM heavy-chain production. In the next step, the $IGHM^{-/-}$ $IGHML1^{-/-}$ fibroblasts were transfected with an artificial human chromosome (κHAC) bearing the unrearranged human heavy and κ-light-chain loci (**Fig. 1**). In the end, after seven rounds of cloning, a healthy transgenic calf, with all bovine IgM heavy chain alleles inactivated and bearing the human artificial chromosome, was obtained. This animal expressed 60-fold more human immunoglobulins than animals described previously[8]—a yield that is potentially competitive from a cost perspective with producing nonhuman hyperimmune globulins.

Vaccination of the calf with anthrax-protective antigen yielded high titers of anthrax-specific immunoglobulins. Although ~80% of serum IgGs were functional chimeric antibodies comprising human heavy chains and bovine light chains, the remainder were fully human (**Fig. 1**). Once purified, the hyperimmune globulins fully protected mice challenged with anthrax spores and in an *in vitro* toxin neutralization assay outperformed a control anthrax hyperimmune globulin preparation derived from human donors.

Importantly, the calf's immunization response was similar to that of wild-type cattle, confirming that the human immunoglobulin loci can support the humoral response in the absence of bovine IgM. Other transgenic calves produced in this study appeared to produce similar levels of human IgGs, suggesting that a herd of cattle with this genotype could provide an abundant source of human hyperimmune globulins. Nevertheless, extensive purification will be necessary to obtain preparations containing only fully human IgGs, which will increase production costs. Knocking out the bovine Igλ locus, which contributes ~90% of light chains in cattle, in this line could further increase the proportion of fully human immunoglobulins and improve process yields.

Although the seven years that have elapsed since this group reported transchromosomic calves expressing human immunoglobulin loci[8] might seem like a long development time, it is worth remembering that the line generated in the present study[5] or subsequent lines could support production of multiple products. Each new hyperimmune globulin product would be dependent on the antigen used in the immunization protocol, rather than the bovine line or the purification process. Scale-up should be relatively straightforward, although this might require cloning rather than natural breeding.

It is still too early to confidently predict the commercial success of human hyperimmune globulins from transgenic cattle. Uncertainties remain concerning, for example, the impact of purification on production costs and the feasibility of using somatic cell nuclear transfer to generate large numbers of animals. Clinical studies—the costliest and riskiest aspect of drug development—must also be completed. But given the flexibility and scalability of using transgenic large animals, this approach may be well placed to compete with traditional human- and animal-derived intravenous immunoglobulins, hyperimmune globulins, and monoclonal and polyclonal antibodies produced in cell culture[9] in applications spanning infectious diseases, oncology, neurological conditions and immune modulation. As we enter the Chinese year of the ox, it seems fitting to look forward to clinical trials of polyclonal antibodies obtained from transgenic cattle.

1. Robert, P. *Int. Blood/Plasma News* **25**, 169 (2008).
2. Lemieux, R., Bazin, R. & Néron, S. *Mol. Immunol.* **42**, 839–848 (2005).
3. Newcombe, C. & Newcombe, A.R. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **848**, 2–7 (2007).
4. Farrugia, A. & Poulis, P. *Transfus. Med.* **11**, 63–74 (2001).
5. Kuroiwa, Y. *et al. Nat. Biotechnol.* **27**, 173–181 (2009).
6. Robl, J.M. *Cloning Stem Cells* **9**, 12–16 (2007).
7. Lonberg, N. & Huszar, D. *Int. Rev. Immunol.* **13**, 65–93 (1994).
8. Kuroiwa, Y. *et al. Nat. Biotechnol.* **20**, 889–894 (2002).
9. Rasmussen, S.K., Rasmussen, L.K., Weilguny, D. & Tolstrup, A.B. *Biotechnol. Lett.* **29**, 845–852 (2007).

One photon up, one photon down

Enrico Gratton & Michelle Digman

A microscopy technique based on stimulated Raman scattering achieves label-free imaging with very high sensitivity.

Identifying different molecular species in microscopic images is still a considerable challenge in many areas of biology. Most commonly, especially in experiments on live cells and tissues, what is detected is not the native molecule but a fluorescently labeled analog, which is assumed to mimic the behavior of the unlabeled molecule. In a recent *Science* paper, Xie and colleagues[1], describe a new techniqúe, stimulated Raman scattering (SRS) microscopy, that is capable of imaging unlabeled molecules in live cells and tissues with diffraction-limited resolution and high sensitivity.

Since the 1920s, when Chandrasekhara Raman[2] first explained the loss of energy to vibrations from a beam of monochromatic light traversing a liquid sample, investigation of the quasi-elastic interactions of light with matter has become a major technique for analyzing the vibrational spectrum of molecules in the condensed state and the composition of biological samples[3]. The advent of the laser in the 1960s and the introduction of resonance Raman scattering, which increase the sensitivity to specific vibrations near a chromophore, made this a relatively simple and accessible method. The appeal of this approach in microscopy is that the molecular vibrations excited by the Raman effect are exquisitely dependent on local molecular arrangements and therefore serve as a fingerprint of individual molecules or classes of molecules.

Until now, the small intensity of Raman scattering has meant that only techniques based on coherent anti-Stokes Raman scattering (CARS) have had the sensitivity necessary for diffraction-limited microscopy[4]. In the CARS effect, two laser beams impinge on the sample. The pump laser excites a vibration and the probe laser produces the anti-Stokes transition (that is, addition to the probe laser of energy contained in excited molecular vibrations), resulting in a new emission wavelength different from those of the pump and probe lasers. This new wavelength characterizes the energy of the

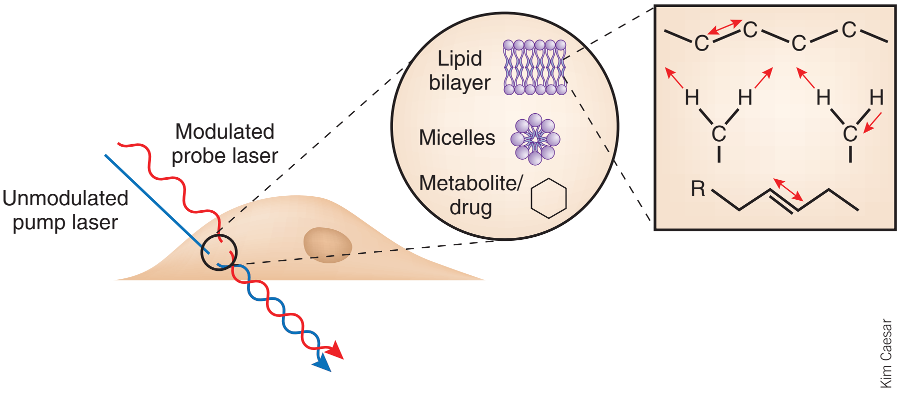
Enrico Gratton and Michelle Digman are at the Laboratory for Fluorescence Dynamics, University of California, Irvine, California 92697, USA.
e-mail: egratton@uci.edu

vibration that was added to the probe beam, and it can be detected with a very high signal-to-noise ratio. During the 1990s, Xie's group combined CARS with laser scanning microcopy to achieve diffraction-limited resolution and high sensitivity by exposing a sample to collinear laser beams of different energy[4]. However, in CARS microscopy, attribution of the signals to specific vibrations and therefore to specific molecules remains difficult owing to signal distortion and the relatively large nonresonant background.

In their new work, Xie and colleagues[1] show that these problems can be overcome using SRS microscopy. In an SRS experiment, two laser beams that differ by exactly the energy needed to excite a specific molecular vibration impinge simultaneously and collinearly on the sample (**Fig. 1**). One laser is used to pump the vibrations and the other to produce the stimulated emission process that forces these vibrations to return to their ground state. The authors demonstrate that the spectral response of SRS is identical to the spontaneous Raman signal and that the background signal is eliminated, making it possible to use the extensive Raman literature to assign vibrational spectra to specific molecules.

Xie and colleagues[1] achieve very high sensitivity, which is crucial for imaging biological samples, by a clever detection set-up based on modulation of the pump and stimulated laser beams. In their design, the laser used for stimulated emission is modulated at high frequency, and this modulation transfers to the pump beam if molecules with the particular Raman excitation band are present. This transfer of modulation from one beam to another is detected using a lock-in amplifier, delivering very high sensitivity. Because two laser beams are interacting in the sample, all the advantages of multiphoton microscopy, such as optical sectioning and diffraction-limited resolution, can be obtained without using pinholes[5]. The pump and probe beam are then scanned through the sample using conventional galvanometer scanners, and the changes in the modulation of the pump beam are recorded at the different positions in the sample. The amplitude of the modulation at each pixel is proportional to the concentration of the molecules with the particular Raman absorption band. Tuning to a specific vibration is done by changing the difference in energy between the pump and probe laser beams.

The idea of applying different laser beams of different energy to produce specific interactions in the sample is not new and has been used with direct two-photon excitation of electronic states[5], the CARS effect[5] and stimulated emission[6]. All these schemes allow optical sectioning without a pinhole and very high selectivity and sensitivity. One important difference between two-photon excitation and the stimulated emission methods (either from vibrational or from electronic states) is that the stimulated emission methods depend separately and linearly on the intensity of the pump and probe beams. Therefore, only one of the two laser beams has to be of relatively high power, reducing instrument cost and sample damage. From the technical point of view, we anticipate that other techniques for modulating the laser beams, such as those used by Dong *et al.*[6] for their stimulated emission microscope, will further improve the signal-to-noise ratio. One technical limitation of SRS microscopy as proposed by Xie and colleagues[1] is that the pump and probe beams are measured in the same direction as that of the laser propagation. This can be problematic in the case of thick tissues, although improved detection systems could alleviate this limitation.



Kim Caesar

Figure 1 Before passing through the sample, the pump laser (blue) is unmodulated and the probe laser (red) is modulated, that is, its amplitude is varied with a high frequency. The lasers are focused on different areas of the sample with different molecular compositions. When an area contains molecules with vibrations of the energy equivalent to the energy difference between the two lasers, the modulation is transferred from the probe to the pump laser. The amplitude of the modulation, which is proportional to the concentration of the molecule of interest, can be measured with high sensitivity.

Given its high selectivity for specific vibrations, SRS microscopy lends itself in principle to the study of a large variety of molecules, including metabolites and small-molecule drugs. A potential limitation of the method is that the low intensity of Raman scattering requires relatively high concentrations of the molecule of interest. Xie and colleagues[1] estimate that the detection limit for their model substrate retinol is 50 µM, although this value will vary widely according to the scattering cross-section of the particular molecule of interest.

The concentration of vibrations that can be measured in a reasonable amount of time by SRS is adequate to detect vibrations arising from lipids, as these molecules are very abundant in biological samples. The sensitivity of SRS to specific lipids is of great interest as the detection and characterization of lipids in biological tissues using fluorescent molecules is still challenging. Fluorescence-based techniques rely on the assumption that fluorescent lipid analogs behave exactly as their nonfluorescent original molecules, which is difficult to demonstrate. By circumventing this issue, the SRS technique may contribute substantially to our knowledge of the microscopic organization and local composition of biological membranes. In addition, SRS will be useful for studying lipid metabolism and transport pathways, either by monitoring the appearance and disappearance of characteristic vibrations at different subcellular locations or by characterizing the composition of transport intermediates.

A promising application of SRS is the study of drug delivery to complex tissues. In a proof-of-principle experiment, Xie and colleagues[1] compare the transport characteristics of retinoic acid and dimethyl sulfoxide in mouse skin. This suggests that Raman microscopy has the unique ability to visualize the penetration and sites of accumulation of unlabeled compounds *in situ*.

1. Freudiger, C.W.M. *et al. Science* **322**, 1857–1861 (2008).
2. Raman, C.V. & Krishnan, K.S. *Nature* **121**, 711 (1928).
3. Peticolas, W.L. *Biochimie* **57**, 417–428 (1975).
4. Zumbusch, A.H., Holtom, G.R. & Xie, X.S. *Phys. Rev. Lett.* **82**, 4142–4145 (1999).
5. Denk, W., Strickler, J.H. & Webb, W.W. *Science* **248**, 73–76 (1990).
6. Dong, C.Y., So, P.T., French, T. & Gratton, E. *Biophys. J.* **69**, 2234–2242 (1995).

Knocking sense into regulatory pathways

Guri Giaever & Corey Nislow

Simultaneous targeted perturbations illuminate the structure and function of regulatory networks.

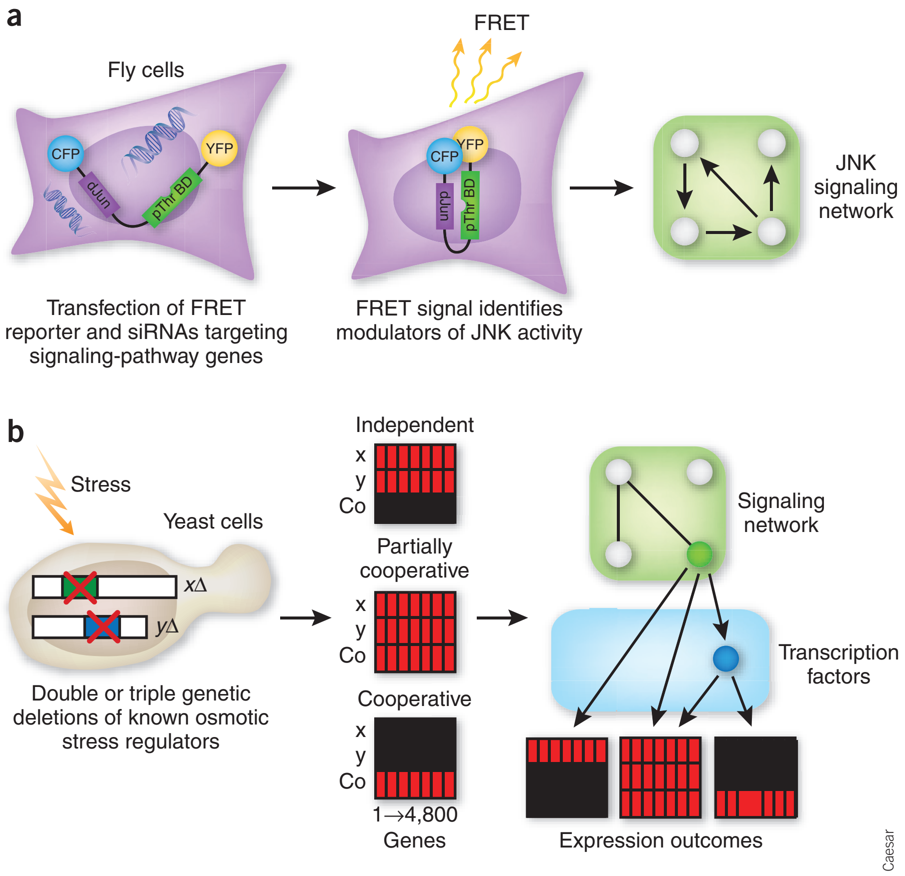
The complex architectures of cellular signaling pathways are beginning to yield their secrets thanks to sophisticated combinations of experimental and computational tools. Two recent papers describe complementary new approaches to identifying the relationships between signaling components (kinases and phosphatases), the transcription factors they regulate and subsets of target genes. Both studies exploit the possibility of introducing multiple, simultaneous perturbations into model systems: Bakal *et al.*[1], in *Science*, study RNA interference of two genes simultaneously in *Drosophila* cells, and Capaldi *et al.*[2], in *Nature Genetics*, carry out double and triple genetic knockouts in yeast. Both groups use their perturbation data to order components in a signaling pathway by epistasis analysis. Yet each dissects a different level of the regulatory hierarchy—kinase cascades in Bakal *et al.*[1] and transcriptional circuits in Capaldi *et al.*[2]—illustrating the power and versatility of multiperturbation strategies.

A grand challenge of genetics and systems biology in the next decade will be to integrate vast amounts of new data on gene and protein functional interactions into frameworks that define cellular pathways. Specifically, we will need to understand how cellular parts assemble into pathways, how multiple pathways are coordinated and how pathways are insulated and integrated to form a functioning cell[3]. Pathways do not generally act linearly, nor do they act in isolation. To dissect pathway architecture genetically, it will be necessary to perform epistatic analysis on data derived from experiments that employ multiple mutations or knockouts. This will be a challenging task, regardless of the phenotypic readout. Furthermore, the derived pathways might not reflect physical interactions between pathway components but rather a 'logical' architecture—that is, there may be more than one pathway that can predict the observed phenotypes.

Bakal *et al.*[1] used a Förster resonance energy transfer (FRET)-based reporter of phosphorylation of the JUN NH_2-terminal kinase (JNK) to screen the effects of 1,565 double-stranded (ds) RNAs that target all known and predicted *Drosophila* kinases, phosphatases, regulatory subunits and adapters (**Fig. 1a**). Although 24 known and novel regulators of JNK phos phorylation were identified, several well-known regulators were not. To address this problem, the researchers screened the 1,565 dsRNAs against 12 cell cultures sensitized by simultaneous transfection with a second dsRNA that targeted a canonical component of the JNK pathway. The resulting 17,724 (~1565 × 12) double perturbations identified 55 additional regulators of JNK. Next, they applied an integrative network algorithm that incorporates genetic and phosphoproteomics data to construct a JNK phosphorylation network, allowing them to propose a model of the architecture of JNK signaling.

This study is impressive in the number of double dsRNA combinations tested and in the

Guri Giaever and Corey Nislow are at the Terrence Donnelly Centre for Cellular and Biomedical Research, 160 College St., University of Toronto, Toronto M5S3E1, Ontario, Canada.
e-mail: guri.giaever@utoronto.ca, corey.nislow@utoronto.ca



Kim Caesar

Figure 1 Multiple-perturbation experiments enable reconstruction of signaling pathways. (**a**) Strategy used by Bakal *et al.*[1] to identify the signaling network regulating JNK activity. *Drosophila* cells containing a dJUN-FRET sensor-reporter of JNK phosphorylation activity are transfected with one or two dsRNAs (left panel). FRET signals identify dsRNAs that modulate JNK activity (middle panel). A probabilistic computational framework is then applied to reconstruct the phosphorylation signaling network (right panel). (**b**) Strategy used by Capaldi *et al.*[2] to predict the transcriptional activation network of the Hog1 MAPK pathway. Yeast mutants with combinatorial deletions of genes known to be involved in the Hog1 pathway are generated (left panel). Gene expression profiles of the mutant strains are analyzed to derive the effects of individual deletions (labeled x and y) as well as the 'cooperative effect' (labeled 'Co') of two deletions together on all genes in the genome (middle panel). Genes are then clustered by whether x and y regulate them independently, partially cooperatively or cooperatively. The resulting modes of regulator interaction allow the order of the genes in the pathway to be inferred (right panel).

blending of experimental and computational approaches, but its scale also hints at the magnitude of the work that remains. Surely, additional double dsRNA screens beyond the 12 performed will reveal more JNK regulators. Moreover, there is room to optimize experimental protocols for dsRNA knockdown (e.g., in sequence selection and delivery) and for measuring the degree of silencing of each gene and off-target effects. Finally, the studies of Bakal *et al.*[1] were all carried out with unstimulated *Drosophila* cells. Given JNK's known roles in maintaining cell, tissue and organism fidelity in the face of cellular stress, additional experiments will be needed to determine if and how the architecture of the JNK network is affected by stress.

JNK is known to exert its influence through numerous transcription factors, which in turn regulate a diverse set of target genes. How might the influences of the JNK regulators extend into the downstream transcriptional program? Capaldi *et al.*[2], working with budding yeast, suggest a way of tackling this question. The authors focused on building a quantitative model of the Hog1 MAPK-dependent pathway, which regulates the osmotic stress response. They performed gene expression profiling experiments on single-, double- and triple-knockout mutants of *hog1*, *msn2/4*, *sko1*, *sok2* and *hot1*—all known components of the Hog1 pathway.

Their key insight was to computationally tease apart the effects of knocking out single genes from what they call a "cooperative component," which quantifies whether two genes function independently, cooperatively (epistasis) or partially cooperatively (**Fig. 1b**). Perhaps most importantly, their method computes the cooperative component for each gene in the genome, providing a fine-grained view of how pairs of regulators interact functionally over the entire genome. The resulting regulatory map shows, for the first time, how the Hog1 MAPK signal propagates through different combinations of transcription factors to regulate distinct subsets of genes. By applying their analytical method to expression profiles of salt- versus glucose-induced osmotic stress, Capaldi *et al.*[2] also suggest how different branches of the regulatory hierarchy are used in a context-dependent manner to respond to different types of osmotic stress.

In contrast to the Bakal study, Capaldi *et al.*[2] use complete knockouts of components of a well-understood pathway. These precise deletions have the advantage of being invariant from cell to cell and assay to assay, but have the disadvantage that they cannot be considered essential genes. An important feature of both studies is that they rely on phenotypic readouts other than fitness to define pathway architecture. Comparing the networks derived from multiple phenotypic measures should greatly expand the 'dynamic range' of network biology.

As these two studies show, the analysis of multiple mutants and multiply perturbed cells provides crucial information for reconstructing the 'wiring diagram' of the cell, but the impact goes further. Drugs, for example, are simply cellular perturbations that can be conditionally applied. This concept was recently explored by chemically perturbing multiple mutants to enrich for genetic interactions and to order the components in a DNA repair pathway[4,5]. In a clinical context, a patient taking a therapeutic agent represents a unique genetic background combined with a chemical perturbation. Thus, it is conceivable that the approaches discussed here might eventually help guide the analysis of complex perturbation experiments in therapeutic settings.

1. Bakal, C. *et al. Science* **322**, 453–456 (2008).
2. Capaldi, A.P. *et al. Nat. Genet.* **40**, 1300–1306 (2008).
3. Hartwell, L. *Nature* **387**, 855–857 (1997).
4. St. Onge, R.P. *et al. Nat. Genet.* **39**, 199–206 (2007).
5. Lehar, J., Stockwell, B.R., Giaever, G. & Nislow, C. *Nat. Chem. Biol.* **4**, 674–681 (2008).

Sequencing in real time

Michael L Metzker

DNA synthesis by single polymerase molecules has been visualized at the speed of catalysis, heralding a new sequencing technology of unparalleled throughput.

DNA sequencing methods generally work by halting the process of copying the template strand in one way or another, using dideoxynucleotides (in Sanger sequencing), reversible terminators or natural nucleotides[1]. Now, a report by Eid *et al.*[2] in *Science* shows that sequence information can be obtained by continuous monitoring of DNA synthesis itself. This strikingly different approach, which records the incorporation of fluorescently labeled nucleotides into single primer strands in real time, promises to increase the speed and read-length of DNA sequencing and to open new avenues in basic research on DNA polymerases and nucleotide analogs.

Most of the next-generation sequencing systems, such as those from Roche/454 (ref. 3), Illumina/Solexa[4] and Life Technologies/Agencourt Personal Genomics[5], are not single-molecule methods as they rely on DNA amplification. A single-molecule technique was recently reported by Helicos Biosciences[6], but its dependence on reversible terminators limits it to the analysis of short DNA fragments. DNA polymerases perform optimally with nucleotide concentrations in the low micromolar range, a requirement that presents a challenge to single-molecule detection methods, which typically use fluorophores at pico- to nanomolar concentrations.

Michael L. Metzker is at the Human Genome Sequencing Center and the Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza N1409, Houston, Texas 77030, USA.
e-mail: mmetzker@bcm.edu

Eid *et al.*[2], of Pacific Biosciences, solved this problem with the company's zero-mode waveguide array[7], a nanostructured device that reduces the observation volume to the zeptoliter range—an improvement of more than three orders of magnitude over confocal fluorescence microscopy. At this superresolution volume, an estimated 0.01–1 molecule enters the detection layer by diffusion, providing a very low background signal and a signal-to-noise ratio of ~25:1.

To enable parallel sequencing, the authors used a chip with thousands of nanoscale wells containing an immobilized DNA polymerase bound to a primed DNA template to be sequenced (**Fig. 1**). To allow uninterrupted monitoring of nucleotide incorporation, they labeled nucleotides with four distinguishable fluorescent dyes on the terminal phosphate group rather than on the base, creating nucleotide analogs that apparently do not interfere with DNA synthesis by φ29 DNA polymerase, a highly processive, strand-displacing polymerase.

The residence time of the phospholinked nucleotides in the polymerase active site is governed by the rate of catalysis and is on the millisecond time scale. The bound nucleotide generates a recorded fluorescent pulse as no other fluorescent molecules are present in the detection volume of the zero-mode waveguide. Formation of a phosphodiester bond releases the fluorophore, which quickly diffuses away, reducing fluorescence to background levels and generating a natural, unmodified DNA product (**Fig. 1**). Translocation of the template marks an

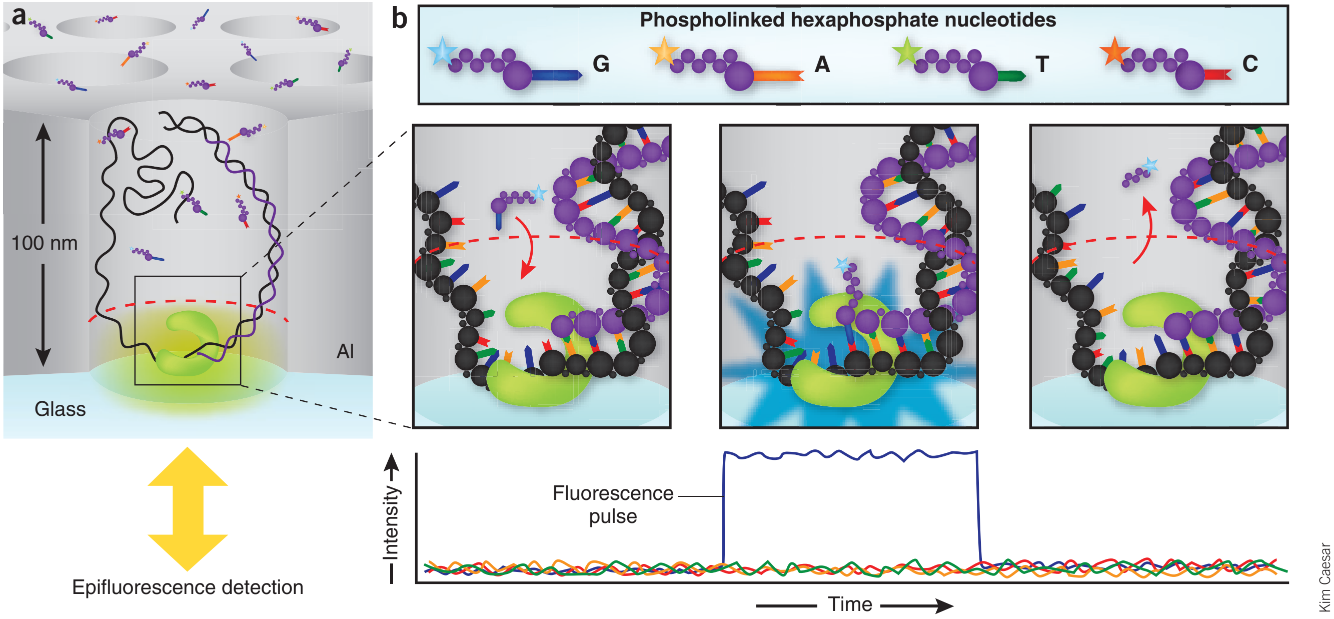

a
100 nm
Al
Glass
Epifluorescence detection
b
Phospholinked hexaphosphate nucleotides
G
A
T
C
Fluorescence pulse
Intensity
Time
Kim Caesar

Figure 1 Real-time single-molecule sequencing. (**a**) A single φ29 DNA polymerase (green), immobilized at the bottom of a zero-mode wavelength nanostructure in an aluminium (Al) casing, adds fluorescently tagged nucleotides to a primed DNA template (black). (**b**) As light penetrates only the bottom 20–30 nm of the well, a fluorescent pulse is recorded only when the correct nucleotide is bound in the active site. Upon phosphodiester-bond formation and diffusion of the fluorescent polyphosphate byproduct out of the detection layer, a dark interphase period is observed. Translocation of the template prepares the DNA polymerase for the next incoming nucleotide. Fluorescent dyes are attached to the terminal phosphates of hexaphosphate nucleotides. Multiple lasers excite fluorophores only when they enter the region below the broken red line (zero-mode waveguide layer) and excitation and emission wavelengths pass through the same objective (epifluorescence detection).

interphase period before binding and incorporation of the next nucleotide. If processivity and fidelity are indeed unaffected by the phospholinked nucleotides, read-lengths of many thousands of bases should be possible.

Using a synthetic template with two dye-labeled phospholinked nucleotides, Eid *et al.*[2] demonstrated an impressive sequencing rate of approximately five bases per second, averaged from 740 single-molecule reads. By contrast, rates of at least 38 bases per second have been reported for ϕ29 DNA polymerase–mediated synthesis from primed M13 template[8]. The potential for long reads was also shown using a closed circular 72-bp template, from which polymerization was maintained for thousands of seconds, yielding several kilobases of synthesized DNA. The strand-displacing capability of the φ29 DNA polymerase enables resequencing of closed circular templates; multiple passes are likely to enhance sequence accuracy.

The team assessed accuracy by analyzing a known 150-bp linear template with a threshold detection method based on dye-weighted summation. When the read was aligned with a known reference, they identified 27 errors, including deletions, insertions and mismatches. This corresponds to a read accuracy of ~83% (131/158). Expanding the number of reads to 449 in a separate experiment revealed a lower accuracy of <80%. The authors attributed sequencing errors to very short interphase intervals (deletions), dissociation of the complementary nucleotide before phosphodiester-bond formation (insertions) and spectral misassignment of fluorescent dyes exhibiting significant emission overlap (mismatches). Because most errors are stochastic events, accuracy was improved to >99% by sequencing the same template molecule 15 times or more.

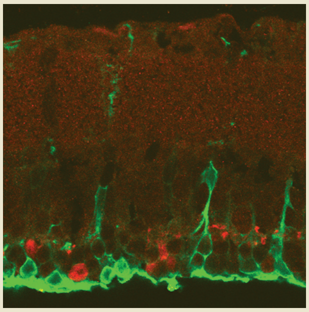
The Pacific Biosciences system can also measure enzyme kinetics of single polymerase molecules. The authors presented data revealing that φ29 DNA polymerase in this system occasionally pauses or exhibits distinct polymerization rates of approximately two bases per second and approximately four bases per second—modes that presumably interconvert. This provides fundamental insights into the kinetics of base incorporation (albeit in a surface-bound system) that would likely be overlooked in bulk solution experiments. It should also be possible to use this technology to study kinetic parameters affected by modified template bases and nucleotides, native and mutant polymerases, and polymerase activators and inhibitors.

Attaining the goal of the $1,000 genome will require substantial improvements in sequencing throughput, accuracey and cost. The array developed by Eid *et al.*[2] contains 93 × 33 wells, and the current method of populating them with DNA polymerases results in about a third of the wells being occupied with one polymerase molecule. Based on a daily throughput of roughly 400 kb, with 15-fold resampling of templates, the estimated throughput of the system is 27 kb of accurate consensus read data per well, or 30 mb per array. Although the authors speculate that increasing the density to 14,000 functioning zero-mode waveguides could produce a daily equivalent of 1× coverage of a human diploid genome, these calculations do not factor in the need for circular consensus sequencing.

In addition to increasing array density (to upwards of 100,000 wells per chip), other potential improvements include engineering more effective polymerases, creating brighter fluorescent dyes that reduce sequencing errors and developing methods that enable self-assembly of single polymerase molecules into each well and efficient circularization of large DNA fragments for consensus sequencing. Many of these challenges will likely be overcome in the near future, marking the entry of a formidable competitor in the next-generation sequencing market.

1. Metzker, M.L. *Genome Res.* **15**, 1767–1776 (2005).
2. Eid, J. *et al. Science* **323**, 133–138 (2009).
3. Margulies, M. *et al. Nature* **437**, 376–380 (2005).
4. Bentley, D.R. *et al. Nature* **456**, 53–59 (2008).
5. Valouev, A. *et al. Genome Res.* **18**, 1051–1063 (2008).
6. Harris, T.D. *et al. Science* **320**, 106–109 (2008).
7. Levene, M.J. *et al. Science* **299**, 682–686 (2003).
8. Soengas, M.S., Gutiérrez, C. & Salas, M. *J. Mol. Biol.* **253**, 517-529 (1995).

Retinal rejuvenation



Blindness caused by the loss of cone or rod photoreceptor cells in the retina may be amenable to cell-replacement therapy. Building on a previous paper showing a functional benefit from grafting of postmitotic rod photoreceptors, Reh and colleagues studied transplantation of retinal cells differentiated from human embryonic stem cells, a cell type that can be expanded without limit. Two to three weeks after injection of 50,000–80,000 retinal cells into the subretinal space of wild-type adult mice, some of the cells had migrated to the outer nuclear layer (the normal location of photoreceptors), adopted photoreceptor morphology and expressed the photoreceptor markers rhodopsin and recoverin. When the same protocol was applied to adult *Crx*$^{-/-}$ mice, which are incapable of any photoreceptor electroretinographic response, 15 of 23 animals acquired the ability to respond to flashes of light. These results establish that human embryonic stem cells differentiated to retinal cells *in vitro* can restore some degree of visual function *in vivo*. (*Cell Stem Cell* **4**, 73–79, 2009) *KA*

Vaccines and antibody maturation

Vaccine development is often hampered by a weak and low-avidity antibody response, leading to, at best, incomplete protection and, at worst, greater severity of the infection. Working on respiratory syncytial virus (RSV), a leading cause of infant hospitalization, Delgado *et al.* identify maturation of antibody affinity as a key factor for safe and efficient immunization. They compare the immune response to nonreplicating vaccines, including a failed vaccine from the 1960s that caused an enhanced respiratory disease, with the response to wild-type virus in mice. Whereas mice challenged with wild-type RSV produce a repertoire of antibodies with increasingly high affinity over time, no affinity maturation is observed with nonreplicating vaccines. The authors identify the activation of Toll-like-receptors (TLRs) as the main determinant for antibody maturation. When combined with a cocktail of specific activators of different TLRs, nonreplicating vaccines show affinity maturation similar to wild type virus and protect mice from enhanced respiratory disease. This underscores the importance of TLR activation in developing better adjuvants and enhancing immunization strategies. (*Nat. Med.* **15**, 34–41, 2009) *ME*

Keeping sepsis at bay

Sepsis kills as many people annually as heart attacks, yet current therapies often fail to stop fatal progression to organ failure. Now Németh and colleagues report that, in a mouse model of sepsis, the infusion of bone marrow stromal cells (BMSCs) does just that. Injecting a million BMSCs around the time that sepsis was induced (by ligating and then puncturing the cecum) led to a 50% reduction in death and spared liver, kidney and spleen. In mice receiving transplants, serum concentrations of inflammatory cytokines (tumor necrosis factor-α and interleukin (IL)-6) rose only slightly, whereas concentrations of the anti-inflammatory IL-10 were elevated. A fluorescent dye allowed the path of transplanted cells to be followed from the blood through the lungs (and spleen and liver). The transplanted cells were surrounded by macrophages in the lung, which led the researchers to ask what role such cells might play in the outcome. In a combination of *in vivo* and *in vitro* experiments using mice deficient in various cytokines, the researchers show that the BMSCs, after binding bacterial lipopolysaccharide, reprogram macrophages into releasing IL-10, which in turn prevents the infiltration of neutrophils into organs, a source of organ damage and pathogenesis. Autologous and allogeneic BMSCs worked equally well, which bodes well for a potential human therapeutic. (*Nat. Med.* **15**, 42–49, 2009) *LD*

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

Bacteria shorten mosquito life

Unlike fine wine, pathogen-carrying mosquitoes do not improve with age. Pathogens such as those that cause dengue fever and malaria must mature in their mosquito hosts for about 2 weeks before they are able to cause disease. McMeniman *et al.* devise an ingenious way of controlling pathogen transmission, taking advantage of this incubation period by developing a strain of bacteria that shortens the lifespan of *Aedes aegypti*, the dengue vector. By serially passaging a strain of the bacterial symbiont *Wolbachia* in mosquito cell culture for 3 years, the researchers weaned the bacterium from its natural host so that it can target *A. aegypti.* In laboratory trials, the lifespan of *Wolbachia*-infected mosquitoes is halved—a reduction predicted to be sufficient to reduce pathogen transmission and the incidence of human disease. The microbial control agent should spread rapidly through natural populations because infected female mosquitoes pass the bacteria to their offspring and cytoplasmic incompatibility prevents uninfected females from reproducing with infected males. And because the bacteria kill mosquitoes long after they have reached sexual maturity, the approach should not compromise reproductive fitness. It may thus be less prone to the emergence of insect resistance, which is problematic with alternatives such as insecticide application. (*Science* **323**, 141–144, 2009) *CM*

FcγRIIa inhibitors get in the groove

Encouraged by evidence that inhibition of the primate-specific Fc-γ receptor IIa (FcγRIIa) may provide new therapies for autoimmune diseases such as rheumatoid arthritis and lupus erythematosus, Pietersz *et al.* exploit knowledge of the three-dimensional structure of the FcγRIIa ligand-binding site to design >100 small-molecule inhibitors predicted to target the groove formed by receptor dimerization. They assess these *in vitro* by screens for inhibition of platelet activation and aggregation, as well as capacity to inhibit tumor necrosis factor-α secretion from macrophages, and test *in vivo* the five most promising candidates using a collagen-induced arthritis (CIA) model involving transgenic mice expressing human FcγRIIa. The strongest inhibitor not only shows better long-term suppression of CIA than methotrexate, immunosuppressive anti-CD3 antibody or FcγRIIa-specific antibody fragments, but also does not inhibit CIA in a CIA-susceptible mouse not expressing FcγRIIa. Although none of the compounds are able to control established disease, their ability to antagonize FcγRIIa activity downstream of immune-complex formation could make them promising leads in the pursuit of less immunosuppressive anti-inflammatories for certain autoimmune diseases. (*Immunol. Cell Biol.* **87**, 3–12, 2009) *PH*

Understanding genome browsing

Melissa S Cline & W James Kent

How can genome browsers help researchers to infer biological knowledge from data that might be misleading?

As genomic knowledge expands, new forms of data become available to help interpret genomic sequences. However, biological data can be noisy: living systems are complex and measurement technologies are rarely perfect. Two excellent approaches for reducing noise are data aggregation and visualization. When combined, multiple forms of evidence tend to be more accurate than a single source, as each distinct form reduces overall uncertainty[1]. The human mind is an outstanding data analysis tool. Although it absorbs textual data poorly, it can assimilate visual data in great detail[2], and can process it efficiently to identify common themes[3].

Genome browsers facilitate genomic analysis by presenting alignment, experimental and annotation data in the context of genomic DNA sequences. These include the University of California Santa Cruz (UCSC) Genome Browser (http://genome.ucsc.edu/), Ensembl (http://www.ensembl.org/), and National Center for Biotechnology Information (NCBI) Map Viewer (http://www.ncbi.nlm.nih.gov/mapview/). They differ in their user interfaces, but address similar tasks, as described in **Supplementary Notes** online and reviewed elsewhere[4]. We focus here on the UCSC Genome Browser.

Figure 1 shows the display for a representative gene queried using the UCSC Genome Browser. The browser displays several tracks, or collections of data, some of which are hidden by default. The user controls which tracks are displayed by means of pull-down menus below the image. The track names are hyperlinked to pages that detail how the data were computed, outline any specific display conventions and may offer additional display options. Each track item within the browser is hyperlinked to a details page providing further information on that item, such as publications in PubMed and sequences in GenBank. The importance of studying these details cannot be overstated. Although genome browsers can simplify the task of generating hypotheses, the user must still evaluate the facts carefully to ensure that the hypotheses are likely to be valid.

Melissa S. Cline is in the Department of Molecular, Cellular and Developmental Biology and W. James Kent is at the Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, California 95064, USA.
e-mail: kent@soe.ucsc.edu

Gene structure and transcripts

Arguably the most important tracks are those that indicate the genes. No data indicate 'the genes' unambiguously. Genes are detected through experimental evidence (namely, observed transcription), and rare transcripts are often difficult to distinguish from measurement errors. To address this uncertainty, there are many gene and gene prediction tracks, each with its own evidence standards.

The high-confidence, low-coverage end of the spectrum contains tracks that derive gene structures from specific full-length transcripts (**Fig. 1a**, line 2). The track indicating genes[5] from the Mammalian Gene Collection (MGC) shows transcripts sequenced from selected high-quality clones. RefSeq Genes[6] shows expert-curated transcripts, along with some provisional transcripts awaiting curation.

For increased coverage, UCSC Genes[7] (**Fig. 1a**, line 1) and Ensembl Genes[8] (**Fig. 1a**, line 3) show predicted transcripts that are derived from mRNA, expressed sequence tag (EST) and protein-sequence alignments. Unlike the RefSeq and MGC transcripts, these transcripts do not always correspond to any single mRNA sequence, but represent composites of sets of similar aligned sequences with good overall evidence.

Aligned sequences offer the broadest but noisiest transcript data. The human mRNA (**Fig. 1a**, line 4) and spliced EST (**Fig. 1b**) tracks show GenBank sequences that align well to the genome. ESTs are short fragments obtained from a single sequencing pass, whereas mRNAs are obtained by high-quality sequencing of entire cDNAs. In general, ESTs describe more transcript isoforms, whereas mRNAs describe fewer isoforms but do so with greater accuracy. However, any aligned sequence is only as good as its underlying clone. If a clone is of poor quality, even the best sequencing protocols will yield misleading sequences. Thankfully, such sequences can often be identified—and disregarded—by following commonsense rules, such as those described below.

Interpreting aligned sequences

First, sequences that align with many errors should be trusted less, because they might not be bona fide products of the locus. Colored vertical lines indicate mismatches and insertions, and double horizontal lines indicate gaps. Sometimes, mismatches arise through normal genetic variation. Such cases can be identified by comparison against data from dbSNP[9] (**Fig. 1a**, line 10).

Second, one should not trust any variation evidenced from only one aligned sequence. For example, BE891408 (**Fig. 1b**, arrow vi) seems to suggest two novel exons, although no other alignment contains these exons. Furthermore, the details page of this EST indicates an older publication date. Together, these facts indicate that this EST should be disregarded.

Third, two or more questionable alignments support each other only if they were derived independently. Aligned sequences are often redundant, with multiple sequences derived from the same clone or from related clones in the same laboratory. Such cases are not independent observations, but one observation recorded multiple times. The browser display is also redundant, as all MGC genes transcripts also appear under human mRNAs (**Fig. 1a**, arrows i and iv). This detail would be easy to miss, and could lead to misinterpretation of sequence-variation frequencies.

Fourth, one should be careful with alignments that suggest partial or erroneous cellular processing. This includes mRNAs that are not spliced or have retained introns (such as BC062326, **Fig. 1a**, arrow iii); mRNAs with premature stop codons, that fall well before the last splice site (such as AK023398, **Fig. 1a**, arrow ii); and run-on alignments that extend past the bounds of the loci (such as DA949381, **Fig. 1b**, arrow v). When such alignments are not supported by others, they probably indicate biological noise.

Finally, a short transcript does not imply a short transcribed region. Aligned sequences are often incomplete, especially in the untranslated regions (UTRs). Sequences are frequently cloned with incomplete UTRs for technical reasons, and sequencers often stop reading prematurely. Thus, variation in alignment lengths might not represent transcript variation; absence of evidence is not evidence of absence. Some tracks can indicate genuine variation: transcription factor binding site data can suggest alternative promoter usage, and the Poly(A) track[10,11] (**Fig. 1a**, line 5) can suggest alternative polyadenylation. For example, the polyA sites near the center of **Figure 1a** suggest that some of the shorter transcripts are actually complete isoforms.

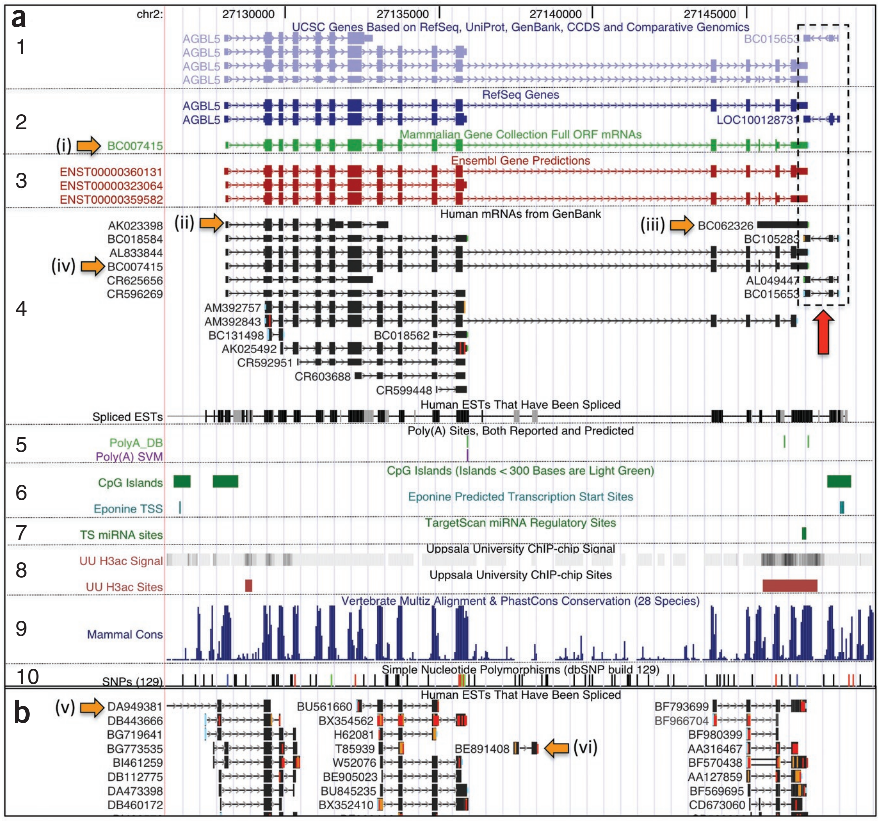


Figure 1 Illustrative screen shots from the current UCSC Genome Browser. (**a**) Selected tracks for the human *AGBL5* locus. 1. UCSC Genes[7]; 2. RefSeq Genes[6] and MGC[5] Genes; 3. Ensembl Genes[8]; 4. Human mRNAs and Spliced ESTs; 5. Poly(A)[10,11]; 6. CpG Islands[14] and Eponine TSS[15]; 7. TS miRNA sites; 8. Uppsala ChIP[16], 9. Conservation[12]; 10. SNPs (129)[9]. Most tracks are shown in pack display mode, with each item displayed separately. The CpG Islands, spliced ESTs, SNPs (129) and TS microRNA sites tracks are shown in dense mode, with all items condensed to a single display line. Darker portions of the EST track indicate regions of stronger evidence, which suggests greater likelihood that the regions are transcribed. In lines 1–4, each track item represents a transcript. Exons are shown as rectangles: taller rectangles indicate coding (CDS) segments, whereas shorter rectangles represent untranslated regions. Introns are shown as lines connecting exons, with arrowheads indicating the direction of transcription. Most transcripts shown are transcribed left to right, in the 5′ to 3′ direction on the sense strand. The dashed box, marked with the red arrow, indicates transcripts of the BC015653 locus on the antisense strand. The human mRNAs track is colored to show mRNA codons that are nonsynonymous to the genome. Orange arrows indicate (i,iv) an mRNA found in both the MGC genes and human mRNAs tracks (BC007415), (ii) an mRNA with a premature stop codon (AK023398) and (iii) an unspliced mRNA (BC062326). (**b**) Excerpt of the spliced ESTs track shown in pack mode, colored to indicate bases that differ from the genomic sequence. Orange arrows indicate (v) a run-on EST and (vi) an alignment consisting of two blocks that are not contained in any other aligned sequence.

Conservation and regulatory data

Figure 1a (line 9) shows genomic conservation, as inferred by MultiZ phylogenetic alignments of genomic sequences[12]. Overall, conservation is strongest in coding exons, weaker in UTRs and weakest in introns and intergenic regions. Strong conservation suggests functional importance, and highly conserved noncoding regions often contain regulatory signals.

TargetScan[13] (**Fig. 1a**, line 7) predicts microRNA binding sites in the highly conserved 3′ UTR. One might assume that this region is highly conserved to preserve these sites. Although this might be true, caution is warranted. TargetScan's track description page indicates that predictions are derived from MultiZ alignments: the predictions depend on conservation. This exemplifies the importance of investigating all of the details before drawing conclusions.

Figure 1a (lines 6 and 8) shows transcriptional start sites suggested by three separate lines of evidence: CpG islands[14], predicted transcription start sites[15] and experimentally determined acetylated histone H3 sites[16]. Each of these signals can be misleading: some genes have no CpG islands, transcription factor binding predictors often overpredict and histone measurement is noisy. However, in aggregate, such data can yield a strong, synergistic prediction.

Moving beyond visualization

After examining a locus, it is often valuable to save data in a text-based format for subsequent analysis. This can be done using the Table Browser[17], accessible through the 'Tables' link. It allows users to select a track, and extract the data from that track for a specific region (defaulting to the last region visualized), or genome-wide. For example, selecting the SNPs (build 129) track and position button allows users to extract a list of SNPs for the region last visualized.

Although genome browsers allow one to scan visually for loci with certain attributes, it can be easier to identify loci with those attributes and then evaluate them visually. This can be done with the Table Browser's filter and intersection functionality. Filtering allows one to limit the track items according to data within the track. For example, one can obtain a list of predicted p53 binding sites by selecting the TFBS Conserved track and filtering for items with names matching "*p53*" (the 'describe table schema' button outlines the available fields). By intersecting the filtered track with the GIS ChIP-PET track[18], one can identify predicted p53 binding sites that are supported experimentally. For output, one can select a set of hyperlinks to the Genome Browser. Or, one can save the output as a custom track and further refine this track through additional filter and intersection actions. This allows users to build sophisticated queries to identify genomic regions sharing a combination of traits.

Conclusion

This primer describes a small subset of the analyses possible with genome browsers, but illustrates some basic principles. Virtually any genomic data can be erroneous, and one should be wary of data suggested by only a single observation. Nonetheless, the combination of multiple observations can suggest reliability, especially when the observations come from varying forms of evidence. Genome browsers facilitate such combination by presenting data visually, in a genomic context. Additional analysis scenarios are described under the recommended resources in **Supplementary Box 1** online.

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

ACKNOWLEDGMENTS

Many people have made significant contributions to this manuscript. In particular, we wish to thank Mark Diekhans, Rachel Karchin, Donna Karolchik, Rachel Harte, Trey Lathe, Mary Mangan and Brooke Rhead for their insights. Bert Overduin and Deanna Church generously offered insights on how to perform similar analyses with the Ensembl and NCBI Map Viewer browsers, so that this paper might assist a broader community of users. This work was funded by the National Human Genome Research Institute (2 P41 HG002371-06 to UCSC Center for Genomic Science, 3 P41 HG002371-06S1 ENCODE supplement to UCSC Center for Genomic Science) and National Cancer Institute (contract no. N01-CO-12400 for Mammalian Gene Collection). M.S.C. is supported under National Institutes of Health GM-040478, and thanks Manuel Ares, Jr. for his mentorship and encouragement. We are grateful to the many researchers who have contributed tracks to the UCSC Genome Browser, as the value of any browser is determined by the underlying data.

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. Shafer, G.A. *A Mathematical Theory of Evidence* (Princeton University Press, Princeton, 1976).
2. Miller, G.A. *Psychol. Rev.* **63**, 81–97 (1956).
3. Bauer, M. & Johnson-Laird, P. *Psychol. Sci.* **4**, 372–378 (1993).
4. Furey, T.S. *Hum. Genomics* **2**, 266–270 (2006).
5. Gerhard, D.S. *et al. Genome Res.* **14**, 2121–2127 (2004).
6. Pruitt, K.D., Tatusova, T. & Maglott, D.R. *Nucleic Acids Res.* **35**, D61–D65 (2007).
7. Hsu, F. *et al. Bioinformatics* **22**, 1036–1046 (2006).
8. Flicek, P. *et al. Nucleic Acids Res.* **36**, D707–D714 (2008).
9. Sherry, S.T. *et al. Nucleic Acids Res.* **29**, 308–311 (2001).
10. Tian, B., Pan, Z. & Lee, J.Y. *Genome Res.* **17**, 156–165 (2007).
11. Tian, B., Hu, J., Zhang, H. & Lutz, C.S. *Nucleic Acids Res.* **33**, 201–212 (2005).
12. Siepel, A. *et al. Genome Res.* **15**, 1034–1050 (2005).
13. Lewis, B.P. *et al. Cell* **115**, 787–798 (2003).
14. Gardiner-Garden, M. & Frommer, M. *J. Mol. Biol.* **196**, 261–282 (1987).
15. Down, T.A. & Hubbard, T.J. *Genome Res.* **12**, 458–461 (2002).
16. Rada-Iglesias, A. *et al. Genome Res.* **18**, 380–392 (2008).
17. Karolchik, D. *et al. Nucleic Acids Res.* **32**, D493–D496 (2004).
18. Ng, P. *et al. Nat. Methods* **2**, 105–111 (2005).

nature biotechnology

Protein promiscuity and its implications for biotechnology

Irene Nobeli[1], Angelo D Favia[2] & Janet M Thornton[2]

Molecular recognition between proteins and their interacting partners underlies the biochemistry of living organisms. Specificity in this recognition is thought to be essential, whereas promiscuity is often associated with unwanted side effects, poor catalytic properties and errors in biological function. Recent experimental evidence suggests that promiscuity, not only in interactions but also in the actual function of proteins, is not as rare as was previously thought. This has implications not only for our fundamental understanding of molecular recognition and how protein function has evolved over time but also in the realm of biotechnology. Understanding protein promiscuity is becoming increasingly important not only to optimize protein engineering applications in areas as diverse as synthetic biology and metagenomics but also to lower attrition rates in drug discovery programs, identify drug interaction surfaces less susceptible to escape mutations and potentiate the power of polypharmacology.

As our understanding of biology increases, so does the evidence that many of our assumptions about what goes on at a molecular level are too naive to capture the complexity of life. Following the realization that the central dogma—'DNA makes RNA makes protein'—is only partly true, another simplistic assumption, that of one protein having one function, is now also being challenged[1–7]. The idea that multiple functions can be associated with single molecular entities, or closely related homologs (referred to as functional promiscuity, cross- or poly-reactivity, poly- or multi-specificity), although widely accepted in fields like immunology and detoxification metabolism, is only recently being discussed in a wider context. The specificity of enzymes, for example, has been thought of as the cornerstone of catalysis, and this has affected the procedures by which biochemical characterization of proteins has been carried out (often discovery of one function has ended the search for others). This is surprising given that functional promiscuity is ultimately a result of binding/interaction promiscuity, which is so common that it is not generally disputed.

Perhaps one explanation is that functional promiscuity may often be invisible, resulting in an observable phenotype only under certain conditions. In various 'underground metabolism' examples[8], wild-type enzymes catalyze reactions acting on substrate analogs that are themselves endogenous metabolites. This reveals a network of reactions carried out at very low, usually undetectable levels, but which can become important if the substrate or enzyme concentration changes owing to other factors. This 'underground' network is one reason why predicting phenotype from genotype remains challenging[9] and why organisms can often be much more robust than expected after deletion of genes involved in major metabolic pathways[10,11]. Another reason why binding and functional promiscuity have not been equally acknowledged is that binding promiscuity resulting from chance encounters of macromolecules and ligands occurs commonly and may not affect an organism. Finally, many functions discovered *in vitro* may not be relevant *in vivo*. However, actively searching for promiscuous activities usually reveals one.

Why is understanding promiscuity important? From a theoretical perspective, the notion of promiscuity is intertwined with that of molecular recognition, and the latter underlies the biochemistry of living systems. Moreover, the evolution of function has often been associated with the initial presence of promiscuity[12], and so a better understanding of promiscuity would improve our knowledge of the processes of evolution (**Box 1**). From a more practical standpoint, enhanced understanding of promiscuity can facilitate progress in protein engineering and drug design for both biomedical or industrial applications. For example, the action of drugs relies on the promiscuous character of their protein targets, and their side effects relate to the promiscuity of unintended targets. Thus, to design drugs less likely to have deleterious side effects and more likely to withstand resistance mutations, we need to understand how to exploit and manage a protein's tendency for binding promiscuity. In biotechnology, the promiscuous nature of proteins could also be exploited to evolve enzymes with different reaction/substrate specificities[13] or to use proteins in innovative ways (e.g., in the context of synthetic biology projects)[14].

This review summarizes our current knowledge of the extent of protein interaction and functional promiscuity, and the different molecular mechanisms that give rise to promiscuity. This basic understanding is fundamental for the success of many areas of applied biology and is not simply an academic exercise. Given that promiscuity is ubiquitous, we do not attempt a comprehensive review of all known cases but limit ourselves to a few illustrative cases that question our one-protein-one-function view. We go on to present a classification that is meant to form the basis of our understanding of how and when promiscuity may occur. In the second part of this review, we then present the implications of

[1]Institute of Structural and Molecular Biology, School of Crystallography, Birkbeck, University of London, Malet Street, London, WC1E 7HX, UK. [2]EMBL Outstation–Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. Correspondence should be addressed to I.N. (i.nobeli@bbk.ac.uk).

Published online 9 February 2009; doi:10.1038/nbt1519

Box 1 Promiscuity and evolution

The implications of functional promiscuity for the evolution and survival of organisms are vast. The consensus appears to be that the presence of proteins with alternative activities, even if initially evolutionarily neutral, can become important for the survival of an organism, if the conditions in the environment change. This can give a selective advantage to members of the population where the secondary activity is higher, or better regulated, thus resulting in the incorporation of such activities under selective pressure. Recent data show that the human genome comprises a large pool of neutral elements that are biochemically active but whose evolution is not constrained in mammals[16]. Presumably, many of these have arisen from gene duplication and are thus a source of functional versatility that could be mined, should the need arise. For a more extensive discussion of the role of promiscuity in evolution, the reader is directed to a number of papers on the subject and the references therein[1,4,8,12,17,90].

promiscuity, and the challenges and opportunities that a new view on function would bring with it. In particular, we present issues and practical applications of functional promiscuity that are relevant to the pharmaceutical and biotechnology sectors.

Classifications of promiscuity

Previous reviews on promiscuity have suggested different classification schemes. Copley[2], whose review centers on enzymes, makes a distinction between ‘moonlighting’ proteins (enzymes that occasionally play structural or regulatory roles[15]) and those with multiple catalytic capabilities (enzymes catalyzing secondary adventitious reactions). Copley also defines four types of catalytic promiscuity: first, using similar substrates to perform one chemical reaction; second, producing multiple products due to imperfect control of the reactants and use of different sites in catalysis; third, using the same residues in the active site but performing different overall reactions; and finally, catalyzing distinct transformations involving different mechanisms. Hult and Berglund[5] recognize three major types of enzymatic promiscuity that can be combined: condition promiscuity (catalysis in a variety of temperatures, pH, etc.), substrate promiscuity and catalytic promiscuity (different chemical transformations performed), which they subdivide into accidental or natural, and induced. Finally, Bornscheuer and Kazlauskas[6] classify catalytic promiscuity according to whether the reactions involve different functional groups, different mechanisms or, more commonly, both.

We suggest that the various manifestations of promiscuity (e.g., catalysis of multiple reactions versus catalysis of a single reaction with different substrates) are inevitably linked to the conditions that trigger them (e.g., differential expression) and the molecular mechanisms underlying them (e.g., flexibility). In addition, our focus differs from that of previous reviews that have concentrated on the functional promiscuity of enzymes with an emphasis on the chemistry; in contrast, we attempt to encompass the wider spectrum of functions in the protein world, and thus our emphasis is on biology. Overall, our classification recognizes that promiscuity manifests itself at different levels, is triggered by different conditions, employs different mechanisms and results in different effects, but almost any mechanism is available to any level, can be triggered by any condition and can have any effect (**Fig. 1**).

Levels of promiscuity

We start with the observation that different molecular entities may lead to promiscuity at different molecular ‘levels’. In other words, alternative functions may be associated with the following: a single gene (a unique DNA sequence); a single transcript (a uniquely transcribed RNA sequence); a single protein (unique amino acid sequence) and either close homologs (high sequence identity, members of the same family) or remote homologs (low sequence identity, members of the same superfamily) in the same species or in different organisms. Here we concentrate mostly on functional and interaction promiscuity at the protein level. We refer to multiple functions for a single protein molecule as ‘individual’ or ‘pure’ promiscuity to distinguish it from the ‘family’ promiscuity, which includes the protein molecule and its close and remote homologs, often observed in duplicated and evolutionarily related proteins.

Promiscuity at the individual gene or transcript level. The pilot phase of the ENCODE project[16], which aimed to identify all functional elements of 1% of the human genome, showed that a single DNA or RNA sequence is unlikely to map to a single function in eukaryotic organisms, confirming that functional promiscuity at the gene or transcript level is normal in higher organisms. Alternative splicing of pre-messenger RNA is common and can result in protein variants with altered binding properties, enzymatic activities, localization and/or stability.

Promiscuity at the individual protein level. This pure promiscuity is more common than was once thought. Various mechanisms can give rise to identical, or near-identical, protein molecules having multiple functions, which is discussed below in more detail.

Promiscuity in the context of related proteins (family promiscuity). Functional promiscuity within families and superfamilies is very common. Gene duplication events leading to paralogous proteins that then diverge in function give rise to a large number of related proteins with different specializations within the same organism. Similarly, speciation events lead to orthologous proteins whose sequences and functions drift through evolutionary time under different selection pressures. Clearly, pure promiscuity is related to the promiscuity observed within a family or superfamily of proteins. The hypothesis that multiple functions within a family evolve (via duplication and subsequent specialization) from promiscuous ‘generalist’ ancestors[17] is both plausible and appealing.

Functional promiscuity in the context of families and superfamilies of proteins has been discussed in the literature by us[18,19] and others[20,21]. In the case of enzymes, the general consensus is that proteins with sequence identity >40%, when exhibiting functional variation, often share the same chemistry (reaction mechanism) but may act on diverse substrates. Related enzymes with low sequence identities (<30%), on the other hand, are more likely to diverge in chemistry as well as substrate. Many related enzymes are known to have conserved only part of a chemical reaction (e.g., the formation of an enolate anion intermediate stabilized by a metal ion in the enolase superfamily[22]) and these are known as mechanistically diverse superfamilies[21,23]. The pattern of function conservation in evolution appears to be superfamily dependent, and variation exists in how much of the substrate substructure conserved within a superfamily is actually part of the reactive substructure in the same superfamily[24]. In general, the diversity of function within protein superfamilies appears to be distributed in a power-law fashion: a few superfamilies exhibit a lot of functional diversity, whereas the majority do not. The biological significance of this observation is debatable. We can only underestimate the functional variation within protein superfamilies because of the incompleteness of the data and the inaccuracy of methods used for function assignment[18]. As a consequence, we cannot

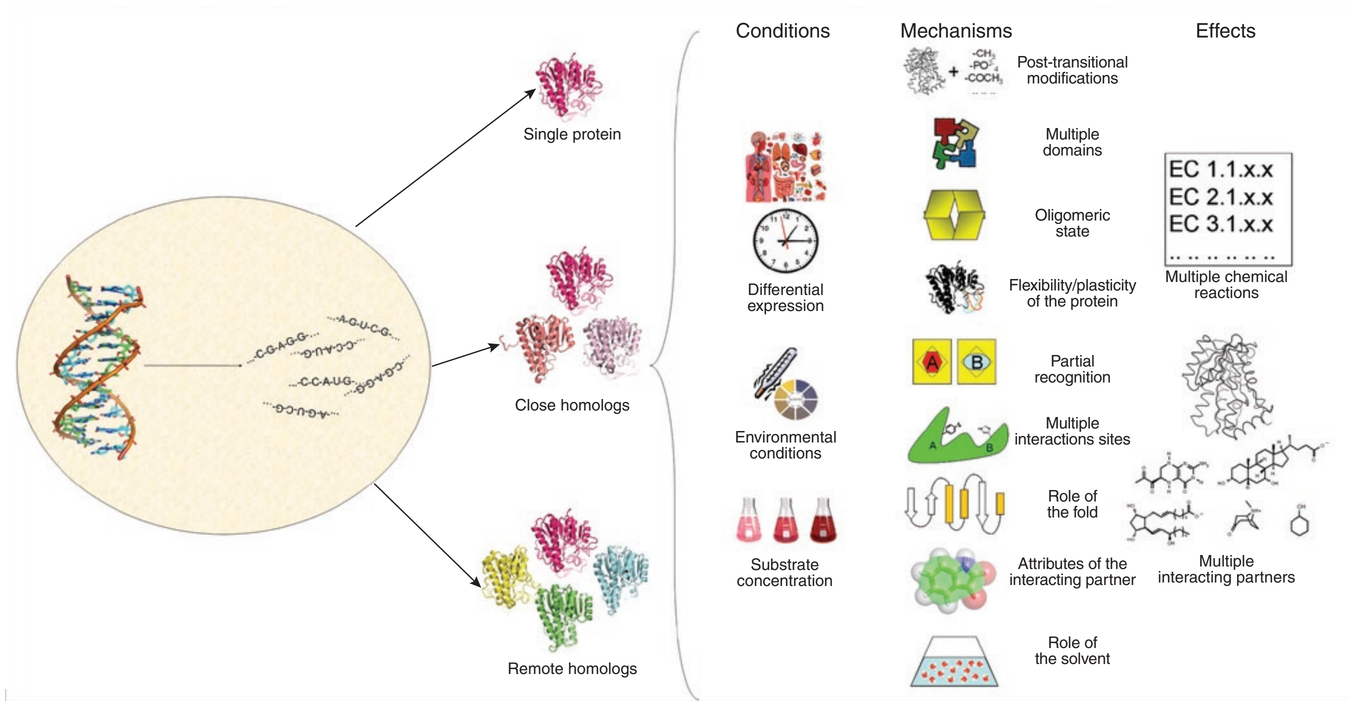


Figure 1 Schematic representation of protein promiscuity: levels, triggering conditions, molecular mechanisms and overall effects.

know whether some families are inherently more promiscuous than others because of physical or biochemical constraints, or whether this is an evolutionary accident.

Manifestations of promiscuity

In general, promiscuity results in the existence of proteins with multiple functions, as in the case of a protein exhibiting both catalytic and structural roles. As this topic has been reviewed previously[1,2,5,7], we mention briefly two major effects: proteins interacting with multiple partners and enzymes catalyzing multiple reactions.

Multiple substrates or partners. An obvious manifestation of multiple functions is a protein interacting with multiple partners. Examples of this abound, but some particularly striking examples are the following: the recognition of multiple antigens by the same germline antibody[25], the recognition of foreign molecules by nuclear receptors such as the pregnane X receptor[26] and the efflux of structurally dissimilar xenobiotics by transporter pumps[27]. Interaction promiscuity between G protein–coupled receptors (GPCRs) and the alpha subunits of G proteins is well documented, despite the fact that selectivity in recognition in these cases is important for the correct signaling pathways to be activated[28]. Numerous examples exist of enzymes that bind related or unrelated substrates leading to alternative products, including members of the short-chain dehydrogenase/reductase family[29], transketolase/transaminases[30] and the ubiquitous kinases[31].

Included here is the interaction of a protein chain with other chains (as homo- or hetero-multimers), as this is often a sign of multiple functions (e.g., the nuclear factor kappa B family of transcription factors comprising five proteins whose combinations of multimers give rise to recognition of diverse target sequences[32]). Finally, the binding of metals or ions is also known to alter the function of proteins; for example, cation binding drastically affects the efficiency and substrate specificity of nucleic acid polymerases[33], and the number of ions can also be crucial for function of the multisite calcium-binding protein calmodulin[34].

Multiple chemical reactions. Currently, many known examples of multiple chemical reactions can be found among enzymes (for a comprehensive list of references, see previous reviews on the subject[1,35]), and evidence for this type of promiscuity goes back a long way. Pocker & Stone[36] showed in the 1960s that erythrocyte carbonic anhydrase, which evolved to catalyze the reversible hydration of CO_2, also has weak esterase activity on a variety of phenyl and naphthyl acetates.

Conditions that drive promiscuity

Several conditions are often associated with protein promiscuity. These conditions may relate to timing or localization of expression of the protein, the environment surrounding the protein or the concentration of a ligand or substrate or cofactor.

Differential expression. This can happen in both space and time. The classic 'space' example is that of crystallins, which make up most of the total soluble protein of the lens in the vertebrate eye but are also expressed in a variety of other tissues. The mammalian alpha B–crystallin, for example, has been implicated in the modulation of intermediate filament organization under physiological stress[37] and in the autoimmune response in multiple sclerosis[38]. A typical 'time' (and space) example is a mechanism commonly used by viruses to make up for their relatively small number of genes by reusing proteins in different contexts. BGLF4, the only serine/threonine protein kinase identified in Epstein-Barr virus (EBV), is known to phosphorylate multiple proteins, but its expression patterns and subcellular localization within EBV-replicating cells suggests that it also plays various roles during the different stages of the virus replication[39].

Environmental conditions. Protein function depends clearly on environmental conditions, such as the pH and temperature. A recent example comes from the study of the enzymatic activity of thymidine kinase from an extremophile eubacterium. This enzyme discriminates against unnatural substrates at the high temperatures encountered by this organism but exhibits substrate promiscuity at much lower temperatures[40].

Concentration of ligand. An early example of this mechanism was that of the mammalian cytosolic aconitase, which was shown to act either as an enzyme or as an iron-responsive-element binding protein, depending on the levels of iron present in the cell[41]. Many nuclear hormone receptors also display this type of promiscuity. Retinoid X receptors (RXR) that are potently activated by 9-cis-retinoic acid may also be activated by docosahexaneoic acid (DHA) in the retina, where DHA is present in high concentrations, but not in other tissues, as the affinity of RXR for DHA is only in the micromolar range[42]. Where the role of a protein is supported by the presence of another molecule, for example, NAD(H) in dehydrogenase/reductase reactions, changing the state of this molecule (here, a cofactor) is a mechanism to change the direction of the reaction being catalyzed.

Mechanisms of promiscuity

Several molecular mechanisms make promiscuity possible, some that are part of the protein mechanism and others part of the interacting partner, which might also be a macromolecule.

Post-translational modifications. Post-translational modifications (phosphorylation, glycosylation, acetylation, and alkylation, among others) are usually associated with activation of a specific function, but evidence suggests that they also constitute an important mechanism for controlling alternative functions. For example, the product of the single enolase gene in the murine malaria parasite has recently been shown to exist in various phosphorylation states, which are distributed not only in the cytoplasm but also in the nucleus, cell membranes and cytoskeletal elements, suggesting a link between phosphorylation and alternative nonglycolytic functions of the different isoforms[43]. Modifications by SUMO proteins (small ubiquitin-like modifiers) are also very important for moonlighting proteins. SUMOylation is the most likely mechanism by which LeCp, a plant cysteine protease normally localized in the cytoplasm, can be transported to the nucleus, where it acts as a transcription factor[44].

Multiple domains. Fused domains are used in nature as a way for enabling single proteins to achieve multiple functions, often the catalysis of consecutive reactions in a metabolic pathway. Twenty years ago, the arom pentafunctional enzyme in yeast was shown to be a mosaic of domains homologous to the monofunctional domains catalyzing individual reactions in the shikimate pathway of *Escherichia coli*[45]. This mechanism can also be used to enrich the functional repertoire of protein superfamily members, which, when fused with different domains, give rise to polypeptide chains that are only partly related and accommodate a diverse range of interacting partners. An extensive range of mechanisms and examples has previously been reviewed[46,47].

Oligomeric state. A protein's oligomeric state can determine its function. For example, pyruvate kinase is a metabolic enzyme as a tetramer and a thyroid hormone binder as a monomer[48]. Moreover, isoenzymes of pyruvate kinase differ in their oligomeric state between normal proliferating and tumor cells (e.g., M2-pyruvate kinase is mostly tetrameric in lung tissue but mostly dimeric in tumors[49]).

Flexibility of the protein. Perhaps the single most important mechanism by which promiscuity can be achieved is structural flexibility. The induced fit theory of recognition[50] enjoyed several years of popularity, following the realization that Fischer's lock-and-key view of enzyme-substrate binding failed to explain the binding of apparently noncomplementary partners and the binding promiscuity involving ligands of very different shapes. In recent years, a more elegant explanation has gained ground. In this view of proteins[12], the protein energy landscape is rugged with many local minima. In the presence of a ligand (whose conformation may also fluctuate), one of the preexisting protein conformations becomes energetically more favorable, the equilibrium is shifted toward that conformation and a complex is formed[51]. Many complexes of varying degrees of complementarity and functional activity may thus form from single amino acid sequences. Evidence for the preexistence of conformational isomers in equilibrium comes both from crystal structures and kinetic data. For example, James *et al.*[52] show that the antibody SPE7 exhibits four different binding-site conformations in six crystal structures. The same authors describe kinetic experiments that are not consistent with a simple bimolecular association of the antibody with a ligand but can instead be explained by the presence of multiple conformers in equilibrium before the binding event. More recently, NMR experiments of Lange *et al.*[53] probing the microsecond dynamics of ubiquitin have revealed an ensemble that covers all conformations observed in 46 crystal structures of this protein (most of which are complexes), reinforcing the evidence for the conformational selection model of recognition. It is worth noting that the theory of selection of preexisting multiple conformations does not necessarily contradict induced fit; the two can be reconciled, as hypothesized by Grunberg *et al.*[54].

Whatever the route to the establishment of different complexes, abundant structural evidence exists of a single protein adopting different conformational states in different complexes. Ekroos and Sjogren[55] have shown that large conformational changes are associated with the binding of two different drugs to the cytochrome P450 3A4 CYP enzyme. The HIV protease can be seen in a variety of conformations bound to different drug molecules (**Fig. 2**). Many more examples of protein flexibility, its origins and especially its implications in drug design can be found in a comprehensive review by Teague[56]. Proteins bound to other proteins may also exhibit large conformational differences, as demonstrated by the very different conformations of glycoprotein Ibα (GpIbα) bound to thrombin in two separate crystal structures[57,58].

The quintessential example of the relationship between flexibility and binding promiscuity comes from the interactions of antibodies and antigens. Antibodies use what is basically the same scaffold to recognize any possible nonself molecule that may be presented to them, and they achieve this by remarkable flexibility of the germline antibody. As antibodies mature to become more specific, their flexibility is reduced. Zimmermann *et al.* demonstrated that the decrease in the flexibility of antibodies during maturation is achieved with mutations that rigidify the combining site by restricting the motion of the complementarity-determining region loops[59].

How is flexibility itself achieved? The presence of loops appears to be a common mechanism for recognizing multiple partners, as exemplified by the short-chain dehydrogenase/reductase family where the C-terminal domain's loop conformations can drastically change the shape and size of the binding site (**Fig. 3**), allowing substrates as diverse as small alcohols and large coenzyme A derivatives to enter and bind[60]. Similarly, the conformations of the eight loops present in the amidohydrolase superfamily are responsible for the wide diversity of substrates that these enzymes are known to hydrolyze[61]. Alternatively, multidomain proteins can take advantage of the flexibility of inter-domain hinges to adapt a binding site to the partner, as in the case of the HIV protease binding site 'flaps' that open and close over the binding site. Finally, external signals may be responsible for the conformational changes.

For example, the binding of calcium atoms promotes a conformational change that exposes hydrophobic patches on the surface of the regulatory protein calmodulin, which are then recognized by peptide sequences in the target enzymes. Promiscuity in the interactions of calmodulin is achieved by exposing different recognition motifs, using different calcium stoichiometries and adopting different conformations[62]. Clearly, flexibility does not need to be localized near the active site (allosteric sites may be involved instead). In fact, Hou *et al.*[63] suggest that extreme functional promiscuity may be more easily achieved by distributing flexibility throughout the protein scaffold, as in the case of the promiscuous detoxifying enzyme GSTA1-1.

Perhaps surprisingly, the abundance of conformational changes associated with molecular recognition may not always favor promiscuity. The model of Savir and Tlusty[64], which is based on statistical mechanics principles, suggests that optimal discrimination between competing targets requires a finite mismatch between the ligand and the target, resulting in a conformational proofreading step. In other words, conformational changes may be selected in evolution for their ability to enhance specificity in recognition in the presence of competing noisy interactions.

Partial recognition. Conformational changes to the binding site may not always be necessary for promiscuity. Molecular recognition can be partial, achieved through imperfect complementarity between a ligand and a target, and thus rigid binding sites may also accept a variety of partners, as long as their shape and chemical complementarity is tolerated. Partial recognition is the most likely mechanism behind the fact that many enzymes can catalyze the reactions of a whole family of ligands, albeit with different catalytic efficiencies. Binding through molecular mimicry of structurally similar ligands is likely to be the most common mechanism for the formation of transient complexes of noncognate partners. In addition, evolution may have encouraged the use of molecular mimicry to help regulate the function of a protein, as in the case of barnase, a RNase that must be inhibited intracellularly: its natural inhibitor barstar partially mimics the RNA substrate by binding to the barnase phosphate-binding site[65].

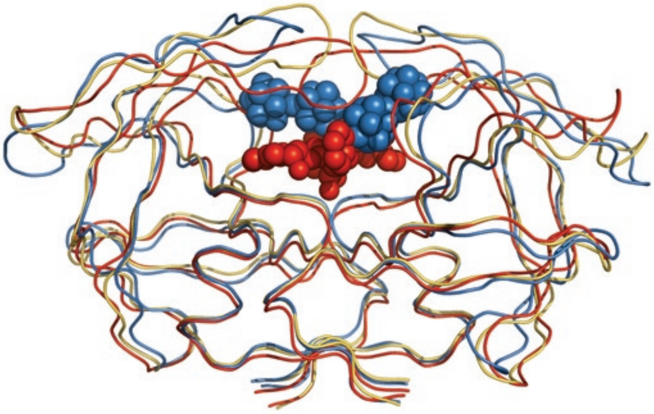


Figure 2 Structural flexibility and promiscuous binding. Structural superposition (based on Cα atoms) of three crystal structures of human immunodeficiency virus (HIV) protease available from the Worldwide Protein Data Bank http://www.wwpdb.org/. The proteins representations are colored in yellow for the apo form, blue and red for the two inhibitor complexes (PDB codes: 1hsi, 1hsh and 1ztz, respectively). The two inhibitors are shown as spheres colored according to the corresponding protein color.

Multiple interaction sites or single site with diverse interacting residues. The availability of multiple binding sites and the possibility of accommodating multiple ligands in a variety of ways in a single site constitute major mechanisms of achieving promiscuity. This is referred to as 'differential ligand positioning' in Mariuzza's classification of paths to multispecificity[66]. This mechanism may have an advantage over conformational flexibility when many different structures need to be recognized, as in the case of antibodies. Using a single conformation in promiscuous interactions avoids having to establish many different conformations from a single sequence that might present some protein-folding challenges[67].

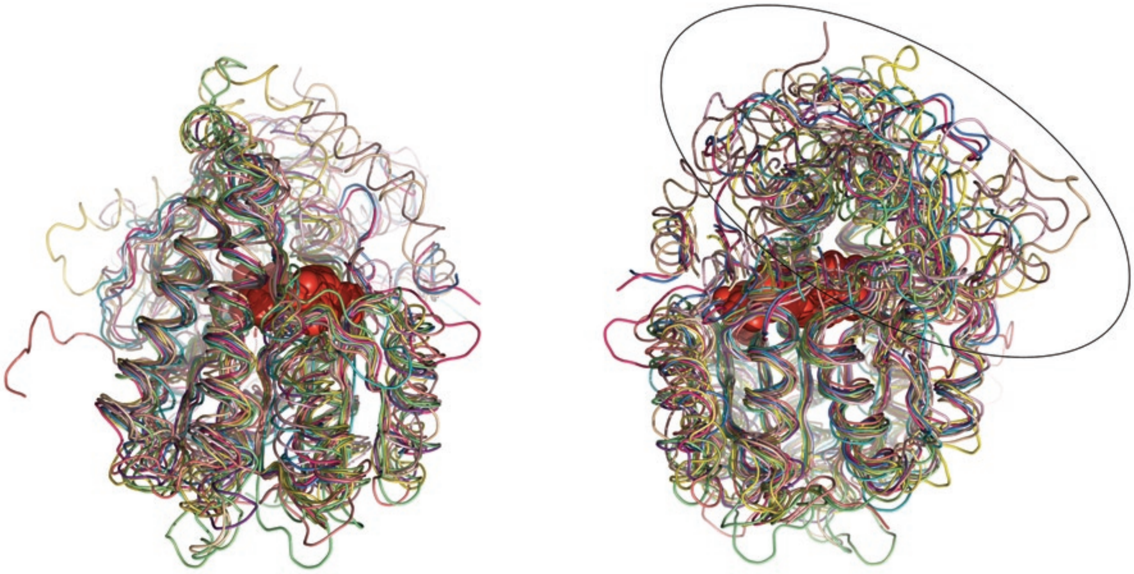


Figure 3 The short-chain dehydrogenase/reductase proteins. Superposition of 15 short-chain dehydrogenase/reductase protein structures based on their Cα atoms; two different views, obtained by a 180° rotation on the main axis, are shown. The C-terminal part of the proteins, where most of the variation occurs, is circled in the right-hand view. Cofactors are depicted as spheres (in red).

Allosteric interactions are also important in facilitating promiscuity. Spiller *et al.*[68] show that mutations that affect the promiscuity of the active site can be found away from it, and they may do so by "transmitting subtle changes into more significant active site perturbations." Directed evolution and the immune system point to the existence of "mutational hotspots" that could influence function across long distances. In a comprehensive structural analysis of affinity-matured antibodies, Orencia *et al.*[69] found that most mutations were located at positions away from the active site. Although there are obviously many more nonbinding site residues than there are residues that are involved in interactions, one would expect mutations that affect the specificity of the antibody to be concentrated inside the binding site, but this does not seem to be the case.

The number of residues that guarantee specificity may indeed be very low. Relaxing the substrate specificity of an enzyme has been achieved with single substitutions in the case of L-Ala-D/L-Glu epimerase from *E. coli* and the muconate lactonizing enzyme II from *Pseudomonas* sp., both members of the enolase superfamily[70]. Similarly, a single amino acid substitution was sufficient to broaden the substrate spectrum of a sialic acid aldolase[71]. The availability of such simple pathways to evolve promiscuous intermediate enzymes may not be surprising in the light of the fact that, if natural evolution is to be accomplished, only a small number of mutations should lead to a selective advantage; otherwise, the probability of accumulating these mutations would be too low[70].

Role of the fold. Some evidence suggests a relationship between fold and promiscuity. Certainly, some folds have been recognized as being particularly plastic, a common example being that of the (beta/alpha)$_8$ barrels, the plasticity of which may be due to their modular architecture and the ability to present catalytic residues from each of the eight strands surrounding the active site[72]. Raillard *et al.*[73] suggest that close sequence similarity coupled with functional diversity may be a good criterion for identifying "functional plasticity islands" in sequence space, and thus proteins that would be useful in protein engineering. Presumably such proteins would also be good candidates for displaying some level of interaction promiscuity. Taylor Ringia *et al.*[74] suggest that the unusual divergence of sequence among homologs performing the same function (as is observed for *O*-succinylbenzoate synthases (OSBS) from eubacteria and archaea) may reflect not only the fact that the reaction requires only modest rate accelerations, but it may also suggest functional promiscuity within the members of the family. Indeed, OSBS from *Amycolatopsis* was initially thought to be acting only as an *N*-acylamino acid racemase. However, it is worth remembering that the difference in plasticity of folds is itself still controversial. Panchenko *et al.*[75] found no significant difference between the degree of structural variation per unit of sequence of different folds or structural classes of proteins, although several exceptions were found.

It has also been suggested that increased stability of a protein fold may render a protein more evolvable because functionally beneficial but destabilizing mutations can be tolerated better[76,77]. Evolvability is not necessarily linked to the initial presence of promiscuity, but the stability of a fold could correlate with a broad energy well (that is, the presence of many low-energy conformers, each of which might interact optimally with a different ligand). One would also expect the evolvability of a fold to be reflected in the distribution of functions within that fold. Indeed, the observation that some protein folds and superfamilies are associated with a large number of functions and structurally diverse ligands, whereas others show remarkable function and substrate conservation[18,19] provides some support for the link between fold, evolvability and functional promiscuity.

Flexibility of the interacting partner. Structural flexibility of the partner plays an equally important role when multiple partners are recognized by the same active site, especially if the protein is relatively rigid. Thus, one would expect to find many ligands bound in conformations that do not correspond to their global energy minimum. Indeed, Stockwell and J.M.T.[78] have shown that some of the most ubiquitously recognized small molecules (ATP, NAD and FAD) adopt very diverse conformations when binding to proteins, and moreover, these conformations often comprise torsion angles outside the preferred low-energy ranges[78]. More generally, over 60% of ligands in a large set of pharmaceutically relevant complexes have been shown to bind in conformations not corresponding to any local minima[79]. Interestingly, the same study[79] showed a correlation between the strain energy that can be tolerated in a complex and the flexibility of the ligand (number of rotatable bonds), indicating that more flexible ligands can tolerate higher strains and thus could be expected to participate in more binding events (and thus, presumably, in more promiscuous ones) than less flexible ones (although it would be difficult to test this hypothesis, as it is not clear how one could isolate this effect from others).

Chemical scaffolds in the interacting partner. A different ligand attribute that may enhance or promote promiscuity is the fact that many ligands share chemical groups. In addition, enzymes usually act on a very small part of the molecule and may have evolved to recognize only a small part of the structure (e.g., a given functional group). One could then imagine that groups with many neighbors in the small-molecule chemical space might be more prone to promiscuous binding, as they may share their targets with those of their neighbors. Classic examples are those of the phosphate, adenine/adenylate moieties and five- or six-member sugar rings that are among the most commonly occurring scaffolds in biology, and that are consequently recognized by a large number of unrelated protein families. Motifs recognizing these groups may be shared even among nonhomologous proteins[80]. The fact that some scaffolds are more ubiquitous in nature may be an evolutionary accident, but we note that two of the three groups mentioned above contain multiple hydrogen bonding opportunities combined with a hydrophobic surface, and this may make them particularly suitable for recognition by multiple arrangements of protein residues.

Size and complexity of interacting partners. The promiscuity of proteins is almost inevitably linked to the promiscuity of ligands, and binding promiscuity of small molecules is widespread. Azzaoui *et al.*[81] have examined the *in vitro* pharmacology profiles of >3,000 ligands against 79 targets and found that >20% were promiscuous according to their definition (not so surprising, considering that their target data set comprised primarily classes known to be promiscuous, such as GPCRs, transporters and nuclear receptors). Examining basic physicochemical properties of their ligands led to the conclusion that larger, hydrophobic ligands containing nitrogen atoms are more likely to be promiscuous and small polar ligands are more likely to be selective. Interestingly, if molecular weight is correlated with complexity (which it is, to a large extent, and for various measures of complexity[82]), then, according to this study, one would expect larger, more complex molecules to be more promiscuous. However, Hopkins *et al.*[83] observed an inverse correlation between promiscuity and molecular weight; in Radhakrishnan and Tidor's study[84], smaller ligands were shown to be more promiscuous than large ones. In addition, Hann *et al.*[85] suggested that larger compounds, being more complex, are statistically less likely to form useful interactions with protein partners. We would add that small ligands are potentially allowed inside more binding sites, whereas large ones are inevitably excluded by steric interactions from the smaller binding sites.

There is no agreement then on whether large size and complexity of small molecules are conducive to promiscuity. This may be partly explained by different definitions of promiscuity, and the fact that if complexity reduces the chances of a very good interaction, it does not necessarily exclude the presence of multiple weak interactions (too weak to be considered as 'activities'). Overall, we believe the promiscuity potential of a small molecule might be a balance between the number of ways that the molecule can be recognized and the number of geometric constraints that are imposed on the active site. An extreme example would be water, which is small, forms limited interactions and is found in virtually all binding sites (of course, water is unique, as most proteins evolved a surface that is compatible with being solvated in aqueous solutions). More realistically, a metal, like zinc, might have a maximum number of only four coordinating residues, so evolving binding sites to recognize zinc might be relatively easy, a theory that is supported by the large number of unrelated proteins that are known to bind zinc[86]. A medium-sized molecule like a nucleotide might be able to contact up to about ten residues, although not all contacts will need to be formed all of the time. In this case, there are many more ways to achieve a minimum binding energy, but the geometric constraints involved will be stricter and more numerous. A fair number of unrelated proteins have evolved to bind nucleotides. On the other hand, a very large molecule, like vitamin B12 might require a very precise geometric arrangement in the binding site but could also use a large number of atoms to contact its protein target. In this case, we expect fewer proteins to have evolved the ability to bind this molecule. We suggest that having more ways to recognize a molecule increases its promiscuity, up to the point where constraining the geometry within the binding site becomes an issue, and then an increase in complexity results in higher specificity instead (**Fig. 4**). However, we do not expect a good correlation between complexity and promiscuity in general. Additional issues such as hydrophobicity, flexibility and the strength of competitive interactions with the solvent are likely to be at least as important as the issue of size and complexity, so that any trends observed are likely to be different for different ligand classes.

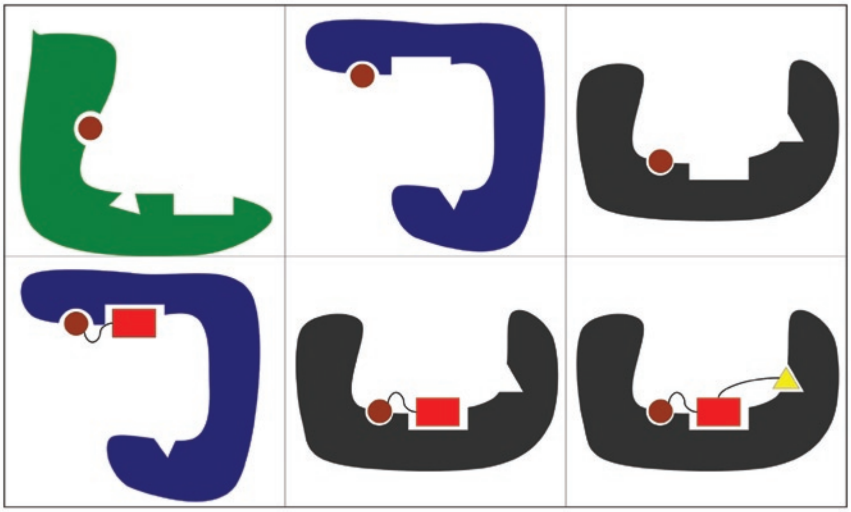


Figure 4 Toy schemata illustrating how size and complexity of a protein's interacting partners can affect the chances of an effective binding. Three hypothetical proteins are shown (in green, blue and black), each comprising three binding sites of different shape. A very small molecule (brown sphere) can be easily accommodated in any of the three proteins. A larger molecule (brown sphere + red rectangle) can only be accommodated by two of the proteins (blue and black). Finally, the most complex molecule (brown sphere + red rectangle + yellow triangle) can only find a complementary binding site in the third protein (black).

Polymers as interacting partners. Promiscuity in recognition is often related to the recognition of biological polymers. The obvious explanation for this is that polymers comprise groups of atoms that repeat periodically and groups that are shared by many members of a class. The peptide backbone in polypeptides or glycosidic bonds in polysaccharides are examples. If recognition relates to these groups, then cross-reactivity can be expected from polymers of the same class, size and conformation. The recognition of aromatic side chains by chymotrypsin is an example of selectivity that extends to a large number of substrates. Similarly, enzymes in the fatty acid elongation pathway can accept substrates with different carbon chain lengths.

Role of the solvent. A binding site or an active site is determined by the presence or absence of solvent molecules. One effect of the solvent is that it creates a buffer zone, allowing a much larger variety of polar interactions to be established than would have been possible using the protein residues alone. The solvent not only directly affects the size and shape of the available binding site but importantly also the dielectric constant and consequently the pK_a values of the host side chains and electrostatic potential of the site. As an added complication, solvent molecules diffuse in and out of the site depending on the conformation of the residues that enclose it. Thus, the role of the solvent in promiscuity is inseparable from the role of conformational flexibility of the host protein.

Energetics of promiscuous behavior

Most promiscuous proteins are thought to use a combination of hydrophobic interactions, hydrogen bonding and flexibility to bind to multiple ligands[87]. Much of our knowledge of energetic origins of promiscuity comes from studies on antibody-antigen recognition. The current consensus is that during affinity maturation the antibodies become more rigid, increasing the specificity and affinity in the recognition of epitopes. Recently, Dimitrov *et al.*[88] showed that the transition from a monospecific to a polyspecific antibody is accompanied by a change in the thermodynamic pathway of binding: highly unfavorable changes in entropy were observed on antigen binding and this together with slow association kinetics point to an increased flexibility of the antigen binding site. However, the overall change in the free energy of binding is not substantially different between the native and urea-exposed antibody, indicating that the specificity transition of the antibody exhibits the phenomenon of enthalpy-entropy compensation[89].

What is perhaps more surprising and largely unexplained at the atomic level is that mutations in a sequence in directed evolution experiments seem to affect the promiscuous activities of a protein more than its native function[90]. Evolution must have played a role in building some robustness for a native function, but how this robustness is achieved in practice is not clear. Differences in the energetic contributions of binding (that is, the idea that efficient native functions are driven by the enthalpy of specific interactions such as hydrogen bonding, as opposed to the entropy of promiscuous binding[91]) may be responsible but it is debatable whether this principle is generally true[92]. Large favorable nonpolar contributions in combination with negligible electrostatic terms, as calculated by molecular dynamics simulations, have been suggested as the basis of promiscuity in the case of the PDZ domains in complex with various proteins[93]. However, as specific interactions could be more prone to individual chance mutations than hydrophobic 'sticky patches', it is not obvious how the idea of enthalpy-driven specificity could be reconciled with the robustness of native functions. It has been suggested that the balance of enthalpic and entropic contributions to binding is important in predicting the response of an inhibitor to mutations associated with drug resistance[94]. It is likely that the balance of forces that drive binding can be used as indicators of whether an interaction is 'native' or 'promiscuous'.

Evidence from the literature overwhelmingly supports the hypothesis that every protein is capable of recognizing multiple partners, as exemplified *in vitro* by the number of ligands usually found to bind with at least micromolar affinities in high-throughput screening. Moreover, recognition seems to exploit all possible routes, as shown in the case of different structural classes of ligands bound to streptavidin[95], where thermodynamics results for the complexes were very diverse, and the structural similarities of the complexes did not translate to energetic similarities (neither in terms of total energy, nor in terms of the balance of entropy versus enthalpy).

Exploiting promiscuity for molecular design

Protein engineering has exploited the phenomenon of interaction promiscuity to design proteins with novel functions. The literature on protein engineering feats is vast and there are numerous comprehensive reviews for the interested reader[96–99]. Here we restrict our discussion to how an understanding of promiscuity could help practical applications in the biotechnology industry.

Protein engineering. How can we evolve new proteins exploiting promiscuity? The most obvious application of promiscuity in protein engineering is that of producing a new enzyme by reinforcing an existing moonlighting or promiscuous reaction. Changing the substrate of one promiscuous reaction to mimic that of a more efficiently catalyzed reaction (by the same enzyme) can achieve a significant increase in its rate, as was shown in the manipulation of the *N*-acylamino acid racemase substrates to mimic the substrate of the *O*-succinylbenzoate synthase (OSBS) reaction catalyzed by the OSBS member of the enolase superfamily[74].

However, if the starting enzyme cannot catalyze the desired reaction, even with low efficiency, but one of its relatives does, then a guided single amino acid substitution may be enough to produce a functionally promiscuous intermediate. Schmidt *et al.*[70] proved this point by rationally mutating single residues in the active sites of *E. coli* L-Ala-D/L-Glu epimerase and *Pseudomonas* sp. P51 muconate lactonizing enzyme II that enabled both proteins to catalyze the OSBS reaction while retaining their original function (at a reduced capacity). Mutations were guided by comparison with the OSBS protein and affected residues that were obviously interfering with the desired substrate for the OSBS reaction.

One important lesson for protein engineers comes from the work of Aharoni *et al.*, who applied directed evolution to serum paraoxonase, a bacterial phosphotriesterase, and carbonic anhydrase II. They have shown that evolution of new functions is driven by mutations that have little effect on the native function but large effects on the promiscuous functions[90]. Tampering with promiscuous functions will prove to be harder, as the relevant residues may sit on a much more rugged energy surface, where small perturbations of the sequence can have detrimental effects on the protein stability or the actual function itself. Protein engineering can further exploit the idea that evolution of a new function for enzymes can take place through a nonspecific (promiscuous) intermediate, possibly one that mimics an ancestral state of low enzymatic specificity (consistent with Jensen's patchwork hypothesis of evolution[17]). This is supported both by the aforementioned work of Aharoni *et al.*[90] and by that of Matsumura and Ellington[100], who evolved a β-galactosidase from a β-glucuronidase via an intermediate with broad substrate specificity. However, shortcuts may be available to protein engineers. It is encouraging that, at least in some cases, it is possible to go directly from one function to another by means of a single amino acid mutation, as was the case of a single substitution on HisA (an isomerase from the histidine biosynthesis pathway) that resulted in the enzyme gaining the TrpF activity of a similar isomerase from the tryptophan biosynthesis pathway[101]. Predicting when such a swift change is possible is not easy, and this may be a relatively rare event.

Protein engineers may also be interested in the lessons learned from promiscuity in secondary metabolism molecular biosynthesis. Unlike primary metabolic pathways that commonly produce a single product, secondary metabolic pathways often produce a variety of natural products, many of which have no known target or use[102]. Gibberellin-producing pathways that achieve the synthesis of over 100 different products across species and several products within a single species are reminiscent of combinatorial diversity-oriented synthetic chemistry. Clearly, we need to improve our understanding of how enzymes in these pathways achieve the production of so many derivatives to exploit their principles in protein engineering; doing so may provide useful hints at engineering pathways that can efficiently explore part of chemical space.

Finally, the recent advent of the fields of metagenomics and synthetic biology may provide new ways of exploiting promiscuity through protein engineering. Metagenomics, which explores unculturable microbes in soil, water and the human body is relevant for several reasons. First, at least some of the novel enzymes discovered are likely to belong to known protein families, thus expanding their functional repertoire (promiscuity at the family level). Importantly, the moonlighting activities of related proteins from different species often evolve independently (as is the case with many proteins in yeasts[103]), suggesting that many new promiscuous activities are likely to be discovered in metagenomic samples. Second, the discovery of catalysts with unusual stability at extreme conditions and their comparison to homologous sequences active at normal conditions could highlight the types of mutations required to achieve stability without compromising function. Finally, discovering alternative substrates and reactions for a family of catalysts could provide more starting points for evolving related sequences.

Synthetic biology[104], which constructs proteins with well-defined properties to form parts of new synthetic biology-based systems, demands a much improved, more intelligent and, most likely, more modular approach to *de novo* protein design. The inherent interaction promiscuity of macromolecules will obviously be both an advantage and a hurdle in this process, as it will allow us to understand and exploit the sequence/structure–function relationship of proteins, but at the same time, it could jeopardize the performance of synthetic systems, which may be affected by the side effects of promiscuous behavior.

Drug design. Drug binding is the ultimate paradigm of a promiscuous activity: without binding promiscuity, we would not be able to introduce man-made compounds that alter or inhibit the function of their targets. At the same time, promiscuity is an obstacle for the successful design of drugs. Toxicity represents the biggest hurdle in drug development, with many promising lead compounds being dropped at a late stage in development after clinical trials raise safety concerns. Ligands predicted to be more promiscuous and less selective were more commonly found among those leads whose development had to be terminated[81]. Not only is the high attrition rate at a late stage of development costly, but also modern attitudes to animal welfare mean that companies will be increasingly under social pressure to limit, if not eliminate, tests on mammals. Thus, *in silico* prediction of unwanted side effects caused by the promiscuous behavior of drugs and their targets is a modern holy grail for the pharmaceutical industry. Considerable effort is being put now into computational[105,106] and experimental[107,108] screening of a series of common or suspected off-targets in the hope that any problems will be identified at the early stages of drug development, before the costs associated with candidates rise steeply. Much of the interest in binding promiscuity comes from studies of kinase inhibitor binding, as serine/threonine and tyrosine kinases were predicted to correspond to more than a fifth of the druggable genome[109]. The exploitation of pockets other than the conserved ATP-binding site (using allosteric inhibitors as in the case of human mitogen-activated protein (MAP) kinases[110]) is now a widespread mechanism to increase selectivity in kinase inhibitor design. In addition, there is considerable interest in several nontarget proteins (the hERG (human ether-a-go-go-related gene) potassium channel, pregnane X receptor, cytochrome P450s, P-glycoprotein and phase 2 metabolizing enzymes), as they bind promiscuously small hydrophobic molecules, presenting a great challenge to drug development[111].

The binding promiscuity of both proteins and their partner ligands is not always unwelcome and can in fact be exploited for drug development. An example is the use of old drugs for new targets, which has been hailed as a promising solution to reducing both the cost and time of drug development[112]. New targets do not need to share an obvious sequence or structure to the old targets, although careful examination of the physicochemical properties of the binding sites can reveal similarities that explain the promiscuous behavior of the drug. For example, celecoxib (Celebrex), a selective cyclooxygenase (COX)-2 inhibitor that has no effect on the constitutive COX-1, binds carbonic anhydrase isoenzymes with nanomolar affinity, and thus shows promise as a therapy for glaucoma and cancer[113]. One advantage of this approach is that it offers an opportunity for developing countries to explore the benefits of drugs no longer covered by patents for the treatment of neglected diseases, while at the same time alleviating the problem of successfully expanding the chemical space that is safe for drug development.

Finally, there is another side to the implications of promiscuity in drug development, which has come from the realization that many drugs (notably those targeting the central nervous system or cancer) do not act on a single target but instead target multiple proteins simultaneously. It appears that in such cases, promiscuity may be not only harmless but actually necessary[114,115], rendering the rational design of promiscuous polypharmacology drugs as a promising new tool in the drug industry[83,116]. The design of the promiscuous drugs will clearly need to take into account not only the inherent promiscuity of ligand scaffolds or specific binding site environments but also the potential genomic variations that exist between individuals, as well as the lifestyle choices likely to influence the expression patterns and concentrations of chemicals in their bodies.

The development of resistance in microbial or viral targets presents a big hurdle in drug development. Enzymes can rapidly acquire mutations that lead to drug resistance but leave their cognate catalytic function unaffected because drug binding is a promiscuous event, and thus not as robust and resistant to mutations as a native function. Thus, targeting directly the residues responsible for the native function (e.g., catalytic rather than substrate-binding residues) should make it harder for an organism to develop resistance[3].

Although we have only referred above to drugs, the discussion applies of course to other types of synthetic products, such as cosmetics,

Box 2 A role for computational biology?

Predicting protein promiscuity is a problem of daunting complexity for bioinformaticians. Indeed, earlier work has shown that bioinformatics methods need improving to reliably uncover promiscuous reactions[117], and our own *in silico* work in protein function prediction[118,119] has curbed our optimism. Even so, we do not doubt that certain areas of bioinformatics research will be important for progress in this field. Some of these avenues have already been pursued by computational biologists, but referencing individual studies is outside the scope of this review.

Data analysis. There is a great deal of data on protein promiscuity to be found in function-related databases. Collating information on promiscuous proteins would be a necessary first step, and existing enzyme, pathway or ontology databases can provide a lot of information on proteins with multiple EC numbers, reactions or substrates, or function categories for genes. Clues as to moonlighting might also be found from expression data. Unexpected expression patterns that do not correlate with our knowledge of existing networks are often indicative of a moonlighting function for a protein. Such data would be complemented by data on alternative splicing of a single gene, which will give hints to any additional roles in the cell.

Sequence-based methods. Large protein families with many relatives may indicate a trend toward promiscuity. Is there a correlation between number of orthologs and number of paralogs and how could it be explained?

Structure-based methods. Analyzing binding site characteristics could reveal those that make proteins more amenable to promiscuity.

Docking profiles. Probing the binding site with panels of selected ligands or other proteins can assess how restrictive the site is toward different types of molecules.

Flexibility. *In silico* studies of the flexibility of proteins can reveal how this may contribute to recognizing multiple partners.

Redundancy in pathways. The evidence of redundancy in metabolic and regulatory networks should be examined carefully, as it may also contain evidence for protein functional promiscuity.

Calculation of promiscuity indices. This could be based on *in silico* or experimental data and could help rank proteins and their partners according to their interaction promiscuity.

Mapping of small-molecule space to protein space. This would reveal any preferences of protein families for sets of chemical groups and possibly allow the engineering of mutatants capable of binding small molecules from neighboring parts of the chemical space.

These are only some possible directions that could be explored to improve our chances of successfully exploiting promiscuity. Experimental verification of any rules learned and predictions made will be indispensable.

nutraceuticals or agrochemicals. Thus, the potential of exploiting promiscuity in industry is vast.

Future directions

We conclude that promiscuity is not a rare phenomenon in biology, that the molecular mechanisms involved are numerous and we have a limited understanding of them, and that a greater understanding of selectivity versus promiscuity will be of enormous industrial and academic value. One goal is, given a protein sequence and/or structure, to computationally predict both the potential partners and actual, or potential, *in vivo* function(s) of this protein (**Box 2**). This is clearly a big challenge, not only because it has proved so hard to predict interaction partners *in silico* but also because function itself is not simply a product of the laws of physics and chemistry but primarily of those of evolution. In our view, the great number of challenges associated with understanding and exploiting functional promiscuity should be seen as a great source of opportunities for the future, for experimental and computational scientists alike.

ACKNOWLEDGMENTS
The authors thank David Moss for his suggestions for improving the manuscript.

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

1. O'Brien, P.J. & Herschlag, D. Catalytic promiscuity and the evolution of new enzymatic activities. *Chem. Biol.* **6**, R91–R105 (1999).
2. Copley, S.D. Enzymes with extra talents: moonlighting functions and catalytic promiscuity. *Curr. Opin. Chem. Biol.* **7**, 265–272 (2003).
3. Colman, P.M. & Smith, B.J. Specificity and promiscuity in protein-ligand and protein-protein interactions. *Aust. J. Chem.* **56**, 763–767 (2003).
4. Khersonsky, O., Roodveldt, C. & Tawfik, D.S. Enzyme promiscuity: evolutionary and mechanistic aspects. *Curr. Opin. Chem. Biol.* **10**, 498–508 (2006).
5. Hult, K. & Berglund, P. Enzyme promiscuity: mechanism and applications. *Trends Biotechnol.* **25**, 231–238 (2007).
6. Bornscheuer, U.T. & Kazlauskas, R.J. Catalytic promiscuity in biocatalysis: Using old enzymes to form new bonds and follow new pathways. *Angew. Chem. Int. Edn Engl.* **43**, 6032–6040 (2004).
7. Jeffery, C.J. Moonlighting proteins. *Trends Biochem. Sci.* **24**, 8–11 (1999).
8. D'Ari, R. & Casadesus, J. Underground metabolism. *Bioessays* **20**, 181–186 (1998).
9. Sriram, G., Martinez, J.A., McCabe, E.R., Liao, J.C. & Dipple, K.M. Single-gene disorders: what role could moonlighting enzymes play? *Am. J. Hum. Genet.* **76**, 911–924 (2005).
10. Kim, J. & Copley, S.D. Why metabolic enzymes are essential or nonessential for growth of *Escherichia coli* K12 on glucose. *Biochemistry* **46**, 12501–12511 (2007).
11. Miller, B.G. & Raines, R.T. Identifying latent enzyme activities: substrate ambiguity within modern bacterial sugar kinases. *Biochemistry* **43**, 6387–6392 (2004).
12. James, L.C. & Tawfik, D.S. Conformational diversity and protein evolution–a 60-year-old hypothesis revisited. *Trends Biochem. Sci.* **28**, 361–368 (2003).
13. Farinas, E.T., Bulter, T. & Arnold, F.H. Directed enzyme evolution. *Curr. Opin. Biotechnol.* **12**, 545–551 (2001).
14. Andrianantoandro, E., Basu, S., Karig, D.K. & Weiss, R. Synthetic biology: new engineering rules for an emerging discipline. *Mol. Syst. Biol.* **2**, 2006 0028 (2006).
15. Watanabe, H., Takehana, K., Date, M., Shinozaki, T. & Raz, A. Tumor cell autocrine motility factor is the neuroleukin/phosphohexose isomerase polypeptide. *Cancer Res.* **56**, 2960–2963 (1996).
16. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
17. Jensen, R.A. Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30**, 409–425 (1976).
18. Todd, A.E., Orengo, C.A. & Thornton, J.M. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143 (2001).
19. Nobeli, I., Spriggs, R.V., George, R.A. & Thornton, J.M. A ligand-centric analysis of the diversity and evolution of protein-ligand relationships in E.coli. *J. Mol. Biol.* **347**, 415–436 (2005).
20. Devos, D. & Valencia, A. Practical limits of function prediction. *Proteins* **41**, 98–107 (2000).
21. Gerlt, J.A. & Babbitt, P.C. Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu. Rev. Biochem.* **70**, 209–246 (2001).
22. Gerlt, J.A., Babbitt, P.C. & Rayment, I. Divergent evolution in the enolase superfamily: the interplay of mechanism and specificity. *Arch. Biochem. Biophys.* **433**, 59–70 (2005).
23. Glasner, M.E., Gerlt, J.A. & Babbitt, P.C. Evolution of enzyme superfamilies. *Curr. Opin. Chem. Biol.* **10**, 492–497 (2006).
24. Chiang, R.A., Sali, A. & Babbitt, P.C. Evolutionarily conserved substrate substructures for automated annotation of enzyme superfamilies. *PLOS Comput. Biol.* **4**, e1000142 (2008).
25. Arevalo, J.H., Taussig, M.J. & Wilson, I.A. Molecular basis of crossreactivity and the limits of antibody-antigen complementarity. *Nature* **365**, 859–863 (1993).
26. Kliewer, S.A., Goodwin, B. & Willson, T.M. The nuclear pregnane X receptor: a key regulator of xenobiotic metabolism. *Endocr. Rev.* **23**, 687–702 (2002).
27. Paulsen, I.T. Multidrug efflux pumps and resistance: regulation and evolution. *Curr. Opin. Microbiol.* **6**, 446–451 (2003).
28. Wong, S.K. G protein selectivity is regulated by multiple intracellular regions of GPCRs. *Neurosignals* **12**, 1–12 (2003).
29. Kallberg, Y., Oppermann, U., Jornvall, H. & Persson, B. Short-chain dehydrogenase/reductase (SDR) relationships: a large family with eight clusters common to human, animal, and plant genomes. *Protein Sci.* **11**, 636–641 (2002).
30. Han, Q., Fang, J. & Li, J. Kynurenine aminotransferase and glutamine transaminase K of *Escherichia coli*: identity with aspartate aminotransferase. *Biochem. J.* **360**, 617–623 (2001).
31. Allende, C.C. & Allende, J.E. Promiscuous subunit interactions: a possible mechanism for the regulation of protein kinase CK2. *J. Cell. Biochem. Suppl.* **30–31**, 129–136 (1998).
32. Wietek, C. & O'Neill, L.A. Diversity and regulation in the NF-kappaB system. *Trends Biochem. Sci.* **32**, 311–319 (2007).
33. Lazcano, A., Diaz-Villagomez, E., Mills, T. & Oro, J. On the levels of enzymatic substrate specificity: implications for the early evolution of metabolic pathways. *Adv. Space Res.* **15**, 345–356 (1995).
34. Valeyev, N.V., Bates, D.G., Heslop-Harrison, P., Postlethwaite, I. & Kotov, N.V. Elucidating the mechanisms of cooperative calcium-calmodulin interactions: a structural systems biology approach. *BMC Syst. Biol.* **2**, 48 (2008).
35. Kirschner, K. & Bisswanger, H. Multifunctional proteins. *Annu. Rev. Biochem.* **45**, 143–166 (1976).
36. Pocker, Y. & Stone, J.T. The catalytic versatility of erythrocyte carbonic anhydrase. VII. Kinetic studies of esterase activity and competitive inhibition by substrate analogs. *Biochemistry* **7**, 3021–3031 (1968).
37. Head, M.W. & Goldman, J.E. Small heat shock proteins, the cytoskeleton, and inclusion body formation. *Neuropathol. Appl. Neurobiol.* **26**, 304–312 (2000).
38. van Noort, J.M. *et al.* The small heat-shock protein alpha B-crystallin as candidate autoantigen in multiple sclerosis. *Nature* **375**, 798–801 (1995).
39. Wang, J.T. *et al.* Detection of Epstein-Barr virus BGLF4 protein kinase in virus replication compartments and virus particles. *J. Gen. Virol.* **86**, 3215–3225 (2005).
40. Lutz, S., Lichter, J. & Liu, L. Exploiting temperature-dependent substrate promiscuity for nucleoside analogue activation by thymidine kinase from Thermotoga maritima. *J. Am. Chem. Soc.* **129**, 8714–8715 (2007).
41. Kennedy, M.C., Mende-Mueller, L., Blondin, G.A. & Beinert, H. Purification and characterization of cytosolic aconitase from beef liver and its relationship to the iron-responsive element binding protein. *Proc. Natl. Acad. Sci. USA* **89**, 11730–11734 (1992).
42. Noy, N. Ligand specificity of nuclear hormone receptors: sifting through promiscuity. *Biochemistry* **46**, 13461–13467 (2007).
43. Pal-Bhowmick, I., Vora, H.K. & Jarori, G.K. Sub-cellular localization and post-translational modifications of the Plasmodium yoelii enolase suggest moonlighting functions. *Malar. J.* **6**, 45 (2007).
44. Matarasso, N., Schuster, S. & Avni, A. A novel plant cysteine protease has a dual function as a regulator of 1-aminocyclopropane-1-carboxylic Acid synthase gene expression. *Plant Cell* **17**, 1205–1216 (2005).
45. Duncan, K., Edwards, R.M. & Coggins, J.R. The pentafunctional arom enzyme of *Saccharomyces cerevisiae* is a mosaic of monofunctional domains. *Biochem. J.* **246**, 375–386 (1987).
46. Vogel, C., Bashton, M., Kerrison, N.D., Chothia, C. & Teichmann, S.A. Structure, function and evolution of multidomain proteins. *Curr. Opin. Struct. Biol.* **14**, 208–216 (2004).
47. Bashton, M. & Chothia, C. The generation of new protein functions by the combination of domains. *Structure* **15**, 85–99 (2007).
48. Parkison, C., Ashizawa, K., McPhie, P., Lin, K.H. & Cheng, S.Y. The monomer of pyruvate kinase, subtype M1, is both a kinase and a cytosolic thyroid hormone binding protein. *Biochem. Biophys. Res. Commun.* **179**, 668–674 (1991).
49. Mazurek, S., Boschek, C.B., Hugo, F. & Eigenbrodt, E. Pyruvate kinase type M2 and its role in tumor growth and spreading. *Semin. Cancer Biol.* **15**, 300–308 (2005).
50. Koshland, D.E., Jr. Correlation of Structure and Function in Enzyme Action. *Science* **142**, 1533–1541 (1963).
51. Ma, B., Shatsky, M., Wolfson, H.J. & Nussinov, R. Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. *Protein Sci.* **11**, 184–197 (2002).
52. James, L.C., Roversi, P. & Tawfik, D.S. Antibody multispecificity mediated by conformational diversity. *Science* **299**, 1362–1367 (2003).
53. Lange, O.F. *et al.* Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science* **320**, 1471–1475 (2008).
54. Grunberg, R., Leckner, J. & Nilges, M. Complementarity of structure ensembles in protein-protein binding. *Structure* **12**, 2125–2136 (2004).
55. Ekroos, M. & Sjogren, T. Structural basis for ligand promiscuity in cytochrome P450 3A4. *Proc. Natl. Acad. Sci. USA* **103**, 13682–13687 (2006).

56. Teague, S.J. Implications of protein flexibility for drug discovery. *Nat. Rev. Drug Discov.* **2**, 527–541 (2003).
57. Celikel, R. *et al.* Modulation of alpha-thrombin function by distinct interactions with platelet glycoprotein Ibalpha. *Science* **301**, 218–221 (2003).
58. Dumas, J.J., Kumar, R., Seehra, J., Somers, W.S. & Mosyak, L. Crystal structure of the GpIbalpha-thrombin complex essential for platelet aggregation. *Science* **301**, 222–226 (2003).
59. Zimmermann, J. *et al.* Antibody evolution constrains conformational heterogeneity by tailoring protein dynamics. *Proc. Natl. Acad. Sci. USA* **103**, 13722–13727 (2006).
60. Oppermann, U. *et al.* Short-chain dehydrogenases/reductases (SDR): the 2002 update. *Chem. Biol. Interact.* **143–144**, 247–253 (2003).
61. Seibert, C.M. & Raushel, F.M. Structural and catalytic diversity within the amidohydrolase superfamily. *Biochemistry* **44**, 6383–6391 (2005).
62. Yamniuk, A.P. & Vogel, H.J. Calmodulin's flexibility allows for promiscuity in its interactions with target proteins and peptides. *Mol. Biotechnol.* **27**, 33–58 (2004).
63. Hou, L. et al. Functional promiscuity correlates with conformational heterogeneity in A-class glutathione S-transferases. *J. Biol. Chem.* **282**, 23264–23274 (2007).
64. Savir, Y. & Tlusty, T. Conformational proofreading: the impact of conformational changes on the specificity of molecular recognition. *PLoS ONE* **2**, e468 (2007).
65. Guillet, V., Lapthorn, A., Hartley, R.W. & Mauguen, Y. Recognition between a bacterial ribonuclease, barnase, and its natural inhibitor, barstar. *Structure* **1**, 165–176 (1993).
66. Mariuzza, R.A. Multiple paths to multispecificity. *Immunity* **24**, 359–361 (2006).
67. Sethi, D.K., Agarwal, A., Manivel, V., Rao, K.V. & Salunke, D.M. Differential epitope positioning within the germline antibody paratope enhances promiscuity in the primary immune response. *Immunity* **24**, 429–438 (2006).
68. Spiller, B., Gershenson, A., Arnold, F.H. & Stevens, R.C. A structural view of evolutionary divergence. *Proc. Natl. Acad. Sci. USA* **96**, 12305–12310 (1999).
69. Orencia, M.C., Hanson, M.A. & Stevens, R.C. Structural analysis of affinity matured antibodies and laboratory-evolved enzymes. *Adv. Protein Chem.* **55**, 227–259 (2000).
70. Schmidt, D.M. *et al.* Evolutionary potential of (beta/alpha)8-barrels: functional promiscuity produced by single substitutions in the enolase superfamily. *Biochemistry* **42**, 8387–8393 (2003).
71. Woodhall, T., Williams, G., Berry, A. & Nelson, A. Creation of a tailored aldolase for the parallel synthesis of sialic acid mimetics. *Angew. Chem. Int. Edn. Engl.* **44**, 2109–2112 (2005).
72. Babbitt, P.C. & Gerlt, J.A. Understanding enzyme superfamilies. Chemistry As the fundamental determinant in the evolution of new catalytic activities. *J. Biol. Chem.* **272**, 30591–30594 (1997).
73. Raillard, S. *et al.* Novel enzyme activities and functional plasticity revealed by recombining highly homologous enzymes. *Chem. Biol.* **8**, 891–898 (2001).
74. Taylor Ringia, E.A. *et al.* Evolution of enzymatic activity in the enolase superfamily: functional studies of the promiscuous o-succinylbenzoate synthase from Amycolatopsis. *Biochemistry* **43**, 224–229 (2004).
75. Panchenko, A.R., Wolf, Y.I., Panchenko, L.A. & Madej, T. Evolutionary plasticity of protein families: coupling between sequence and structure variation. *Proteins* **61**, 535–544 (2005).
76. Bloom, J.D., Labthavikul, S.T., Otey, C.R. & Arnold, F.H. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. USA* **103**, 5869–5874 (2006).
77. Wagner, A. Robustness, evolvability, and neutrality. *FEBS Lett.* **579**, 1772–1778 (2005).
78. Stockwell, G.R. & Thornton, J.M. Conformational diversity of ligands bound to proteins. *J. Mol. Biol.* **356**, 928–944 (2006).
79. Perola, E. & Charifson, P.S. Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J. Med. Chem.* **47**, 2499–2510 (2004).
80. Denessiouk, K.A. & Johnson, M.S. When fold is not important: a common structural framework for adenine and AMP binding in 12 unrelated protein families. *Proteins* **38**, 310–326 (2000).
81. Azzaoui, K. *et al.* Modeling promiscuity based on *in vitro* safety pharmacology profiling data. *ChemMedChem* **2**, 874–880 (2007).
82. Allu, T.K. & Oprea, T.I. Rapid evaluation of synthetic and molecular complexity for in silico chemistry. *J. Chem. Inf. Model.* **45**, 1237–1243 (2005).
83. Hopkins, A.L., Mason, J.S. & Overington, J.P. Can we rationally design promiscuous drugs? *Curr. Opin. Struct. Biol.* **16**, 127–136 (2006).
84. Radhakrishnan, M.L. & Tidor, B. Specificity in molecular design: a physical framework for probing the determinants of binding specificity and promiscuity in a biological environment. *J. Phys. Chem. B* **111**, 13419–13435 (2007).
85. Hann, M.M., Leach, A.R. & Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* **41**, 856–864 (2001).
86. Andreini, C., Banci, L., Bertini, I. & Rosato, A. Counting the zinc-proteins encoded in the human genome. *J. Proteome Res.* **5**, 196–201 (2006).
87. Redinbo, M.R. Promiscuity: what protects us, perplexes us. *Drug Discov. Today* **9**, 431–432 (2004).
88. Dimitrov, J.D., Lacroix-Desmazes, S., Kaveri, S.V. & Vassilev, T.L. Transition towards antigen-binding promiscuity of a monospecific antibody. *Mol. Immunol.* **44**, 1854–1863 (2007).
89. Kang, J. & Warren, A.S. Enthalpy-entropy compensation in the transition of a monospecific antibody towards antigen-binding promiscuity. *Mol. Immunol.* **44**, 3623–3624 (2007).
90. Aharoni, A. *et al.* The 'evolvability' of promiscuous protein functions. *Nat. Genet.* **37**, 73–76 (2005).
91. Padlan, E.A. Anatomy of the antibody molecule. *Mol. Immunol.* **31**, 169–217 (1994).
92. James, L.C. & Tawfik, D.S. The specificity of cross-reactivity: promiscuous antibody binding involves specific hydrogen bonds rather than nonspecific hydrophobic stickiness. *Protein Sci.* **12**, 2183–2193 (2003).
93. Basdevant, N., Weinstein, H. & Ceruso, M. Thermodynamic basis for promiscuity and selectivity in protein-protein interactions: PDZ domains, a case study. *J. Am. Chem. Soc.* **128**, 12766–12777 (2006).
94. Ohtaka, H. *et al.* Thermodynamic rules for the design of high affinity HIV-1 protease inhibitors with adaptability to mutations and high selectivity towards unwanted targets. *Int. J. Biochem. Cell Biol.* **36**, 1787–1799 (2004).
95. Weber, P.C., Pantoliano, M.W. & Salemme, F.R. Crystallographic and thermodynamic comparison of structurally diverse molecules binding to streptavidin. *Acta Crystallogr. D Biol. Crystallogr.* **51**, 590–596 (1995).
96. Brannigan, J.A. & Wilkinson, A.J. Protein engineering 20 years on. *Nat. Rev. Mol. Cell Biol.* **3**, 964–970 (2002).
97. Leisola, M. & Turunen, O. Protein engineering: opportunities and challenges. *Appl. Microbiol. Biotechnol.* **75**, 1225–1232 (2007).
98. Bornscheuer, U.T. & Pohl, M. Improved biocatalysts by directed evolution and rational protein design. *Curr. Opin. Chem. Biol.* **5**, 137–143 (2001).
99. Arnold, F.H. Combinatorial and computational challenges for biocatalyst design. *Nature* **409**, 253–257 (2001).
100. Matsumura, I. & Ellington, A.D. In vitro evolution of beta-glucuronidase into a beta-galactosidase proceeds through non-specific intermediates. *J. Mol. Biol.* **305**, 331–339 (2001).
101. Jurgens, C. *et al.* Directed evolution of a (beta alpha)8-barrel enzyme to catalyze related reactions in two different metabolic pathways. *Proc. Natl. Acad. Sci. USA* **97**, 9925–9930 (2000).
102. Fischbach, M.A. & Clardy, J. One pathway, many products. *Nat. Chem. Biol.* **3**, 353–355 (2007).
103. Gancedo, C. & Flores, C.L. Moonlighting proteins in yeasts. *Microbiol. Mol. Biol. Rev.* **72**, 197–210 (2008).
104. Pleiss, J. The promise of synthetic biology. *Appl. Microbiol. Biotechnol.* **73**, 735–739 (2006).
105. Merlot, C. In silico methods for early toxicity assessment. *Curr. Opin. Drug Discov. Devel.* **11**, 80–85 (2008).
106. Campillos, M., Kuhn, M., Gavin, A.C., Jensen, L.J. & Bork, P. Drug target identification using side-effect similarity. *Science* **321**, 263–266 (2008).
107. Fedorov, O. *et al.* A systematic interaction map of validated kinase inhibitors with Ser/Thr kinases. *Proc. Natl. Acad. Sci. USA* **104**, 20523–20528 (2007).
108. Trubetskoy, O.V. *et al.* High throughput screening assay for UDP-glucuronosyltransferase 1A1 glucuronidation profiling. *Assay Drug Dev. Technol.* **5**, 343–354 (2007).
109. Hopkins, A.L. & Groom, C.R. The druggable genome. *Nat. Rev. Drug Discov.* **1**, 727–730 (2002).
110. Ohren, J.F. *et al.* Structures of human MAP kinase kinase 1 (MEK1) and MEK2 describe novel noncompetitive kinase inhibition. *Nat. Struct. Mol. Biol.* **11**, 1192–1197 (2004).
111. Ekins, S. Predicting undesirable drug interactions with promiscuous proteins in silico. *Drug Discov. Today* **9**, 276–285 (2004).
112. Chong, C.R. & Sullivan, D.J., Jr. New uses for old drugs. *Nature* **448**, 645–646 (2007).
113. Weber, A. *et al.* Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. *J. Med. Chem.* **47**, 550–557 (2004).
114. Mencher, S.K. & Wang, L.G. Promiscuous drugs compared to selective drugs (promiscuity can be a virtue). *BMC Clin. Pharmacol.* **5**, 3 (2005).
115. Roth, B.L., Sheffler, D.J. & Kroeze, W.K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discov.* **3**, 353–359 (2004).
116. Morphy, R., Kay, C. & Rankovic, Z. From magic bullets to designed multiple ligands. *Drug Discov. Today* **9**, 641–651 (2004).
117. Gómez, A., Domedel; N., Cedaño' J., Pinol, J. & Querol, E. Do current sequence analysis algorithms disclose multifunctional (moonlighting) proteins? *Bioinformatics* **19**, 895–896 (2003).
118. Macchiarulo, A., Nobeli; I. & Thornton, J.M. Ligand selectivity and competition between enzymes in silico. *Nat. Biotechnol.* **22**, 1039–1045 (2004).
119. Favia, A.D., Nobeli, I., Glaser, F. & Thornton, J.M. Molecular docking for substrate identification: the short-chain dehydrogenases/reductases. *J. Mol. Biol.* **375**, 855–874 (2008).

nature biotechnology

Transgenic mice with defined combinations of drug-inducible reprogramming factors

Styliani Markoulaki[1,3], Jacob Hanna[1,3], Caroline Beard[1], Bryce W Carey[1,2], Albert W Cheng[1,2], Christopher J Lengner[1], Jessica A Dausman[1], Dongdong Fu[1], Qing Gao[1], Su Wu[1], John P Cassady[1,2] & Rudolf Jaenisch[1,2]

Proviruses carrying drug-inducible *Oct4*, *Sox2*, *Klf4* and *c-Myc* used to derive 'primary' induced pluripotent stem (iPS) cells were segregated through germline transmission, generating mice and cells carrying subsets of the reprogramming factors. Drug treatment produced 'secondary' iPS cells only when the missing factor was introduced. This approach creates a defined system for studying reprogramming mechanisms and allows screening of genetically homogeneous cells for compounds that can replace any transcription factor required for iPS cell derivation.

The generation of iPS cells from mouse and human somatic cells through the forced expression of defined transcription factors[1–4] constitutes a major breakthrough in regenerative biology[5]. However, current reprogramming strategies require viral transduction and/or potentially oncogenic transcription factors. Understanding the molecular changes underlying iPS cell derivation will facilitate the development of safer reprogramming strategies, for example, by replacing the virally transduced factors with small molecules[6–8].

Screening approaches using infected cells are hampered by the genetic variability caused by random integration of multiple proviral copies[9,10]. Recently, we generated a 'secondary' transgenic system that eliminates such heterogeneity[9,10]. In this approach, mouse embryonic fibroblasts (MEFs) heterozygous for the ROSA26-M2 reverse tetracycline transactivator (*ROSA26-M2rtTA*) were infected with doxycycline (dox)-inducible lentiviruses carrying the four reprogramming factors (*Oct4* (*Pou5f1*), *Sox2*, *Klf4* and *c-Myc* (*Myc*)) and induced to generate primary iPS cells by addition of dox. These cells were used to obtain chimeric mice with genetically identical somatic cells that can be isolated and reprogrammed *in vitro* by addition of dox. However, such secondary somatic cells require isolation from chimeric mice and contain copies of all four reprogramming factors, thus impeding their use in drug screens aimed at identifying components that can substitute for a given transcription factor. Here we describe the generation of genetically homogeneous mice and MEF lines containing different combinations of a defined set of dox-inducible proviral genomes. This was achieved through random segregation of the integrated lentiviruses after germline transmission from primary iPS cell–derived chimeras (**Fig. 1a**). We used the previously described Pro B cell–derived iB-iPS#9 cell line[10], which carries a single copy each of *c-Myc* and *Sox2* and two copies each of *Klf4* and *Oct4* ($O_2S_1K_2M_1$) (**Fig. 1b** and **Supplementary Fig. 1** online). To produce transgenic offspring, we crossed an iB-iPS#9 chimera that transmitted the transgenes through the germline in 100% of the offspring to wild-type females (**Fig. 1a**), and 91 individual offspring were genotyped. This analysis identified mice carrying all possible combinations of one, two, three or all four reprogramming factors (**Supplementary Fig. 2** and **Supplementary Methods** online).

We determined whether germline transmission of the inducible transgenes would interfere with their ability to reprogram secondary somatic cells upon exposure to dox. Peripheral blood samples were collected from 90 adult progeny obtained from the iB-iPS#9 chimera and cultured in the presence of dox (**Supplementary Methods**). Initial colonies (**Fig. 1c**) appeared after 7–16 d of dox induction in all seven samples derived from mice positive for *ROSA26-M2rtTA* and all four factors (**Supplementary Tables 1** and **2** online). All lines were expanded without dox, had an embryonic stem (ES) cell–like morphology and expressed SSEA-1 and Nanog (**Fig. 1c**). Four lines (iPS 9.27B, 9.48B, 9.67B and 9.74B) carried a single copy each of *Oct4*, *Sox2*, *Klf4* and *c-Myc* ($O_1S_1K_1M_1$) (**Fig. 1d**). Several iPS cell lines were injected into blastocysts (**Supplementary Table 3** online) and produced chimeras with germline contribution (**Fig. 1e**). To determine whether the copy number of *Oct4* and/or *Klf4* affected the reprogramming process, we analyzed the reprogramming efficiency and kinetics of CD11b$^+$ cells. No major differences were observed between cells carrying multiple or single copies of the reprogramming factors from F_1 and F_2 donor mice (**Fig. 1f** and **Supplementary Fig. 3** online). These results, together with the derivation of iPS cell lines from all *ROSA26-M2rtTA*$^{+/-}$ mice that carried at least one copy of each factor (**Supplementary Tables 1** and **2** online), demonstrate that the lentiviral transgenes are not silenced after transmission through the germline. Also, multiple somatic cell types (tail tip–derived fibroblasts, keratinocytes, liver cells and lymphocytes) from mice carrying single copies of each of the reprogramming factors were efficiently reprogrammed (**Supplementary Figs. 4** and **5** online).

We generated somatic cell lines with different combinations of factors by crossing transgenic male 9.27 ($O_1S_1K_1M_1$; **Fig. 1d**) with wild-type females. MEF cultures were established from individual embryos and genotyped for the segregated transgenes (**Fig. 2a**).

[1]The Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA. [2]Massachusetts Institute of Technology, Department of Biology, 31 Ames Street, Cambridge, Massachusetts 02142, USA. [3]These authors contributed equally to this work. Correspondence should be addressed to R.J. (jaenisch@wi.mit.edu).

Received 11 September 2008; accepted 19 December 2008; published online 18 January 2009; doi:10.1038/nbt.1520

‘Single-copy four-factor’ ($O_1S_1K_1M_1$) MEF lines reproducibly generated iPS cells with ~1% efficiency (**Fig. 2b** and **Supplementary Fig. 6** online). In contrast, no iPS cell colony formation was observed with ‘three-factor’ lines, that is, $O_1S_1K_1$ ($n = 3$), $O_1S_1M_1$ ($n = 2$), $S_1K_1M_1$ ($n = 1$) and $O_1K_1M_1$ ($n = 3$) (**Fig. 2b**). However, when these MEF lines were transduced with the missing factor and grown in the presence of dox, iPS cell colonies appeared within 14–21 d (**Fig. 2b** and **Supplementary Fig. 7** online) at efficiencies similar to the highest reported efficiencies for fibroblasts[9]. All lines grew independently of dox, expressed pluripotency markers and induced teratomas *in vivo* (**Fig. 2b** and **Supplementary Fig. 8** online).

In contrast to previous reports[11,12], reprogramming of tail-derived or embryonic fibroblasts (similar to peripheral blood cells) was not possible from three-factor lines lacking *c-Myc* (**Fig. 2b**), possibly because of suboptimal stoichiometry of the three factors. Indeed, infection of $O_1S_1K_1$ (no *c-Myc*) fibroblasts with a lentivirus expressing *Klf4*, but not with lentiviruses expressing *Oct4*, *Sox2* or green fluorescent protein (*GFP*: control), allowed derivation of iPS cell lines (**Fig. 2c**), suggesting that higher levels of *Klf4* can substitute for the action of *c-Myc*. When M_1K_1 MEFs were treated with dox before transduction with *Sox2* and *Oct4*, we observed enhanced reprogramming efficiency and obtained Nanog-GFP$^+$ iPS cells already after 12–14 d instead of 22–24 d (**Fig. 2d**). In contrast, dox pretreatment of O_1S_1 MEFs before re-infection with *c-Myc* and *Klf4* lentiviruses did not alter reprogramming kinetics or efficiency (**Fig. 2d**). This indicates that early induction of *c-Myc* and *Klf4* sensitizes fibroblasts for the ectopic expression of *Oct4* and *Sox2* and enhances their reprogramming speed and efficiency. These results are consistent with the hypothesis that c-Myc and/or Klf4 might induce epigenetic changes that facilitate the interaction of Oct4 and Sox2 with their targets, resulting in more rapid reprogramming[1].

About 12% of the mice developed skin epithelial tumors, even though they were not treated with dox, suggesting leaky transgene expression in our system. Tumors were only observed in mice carrying

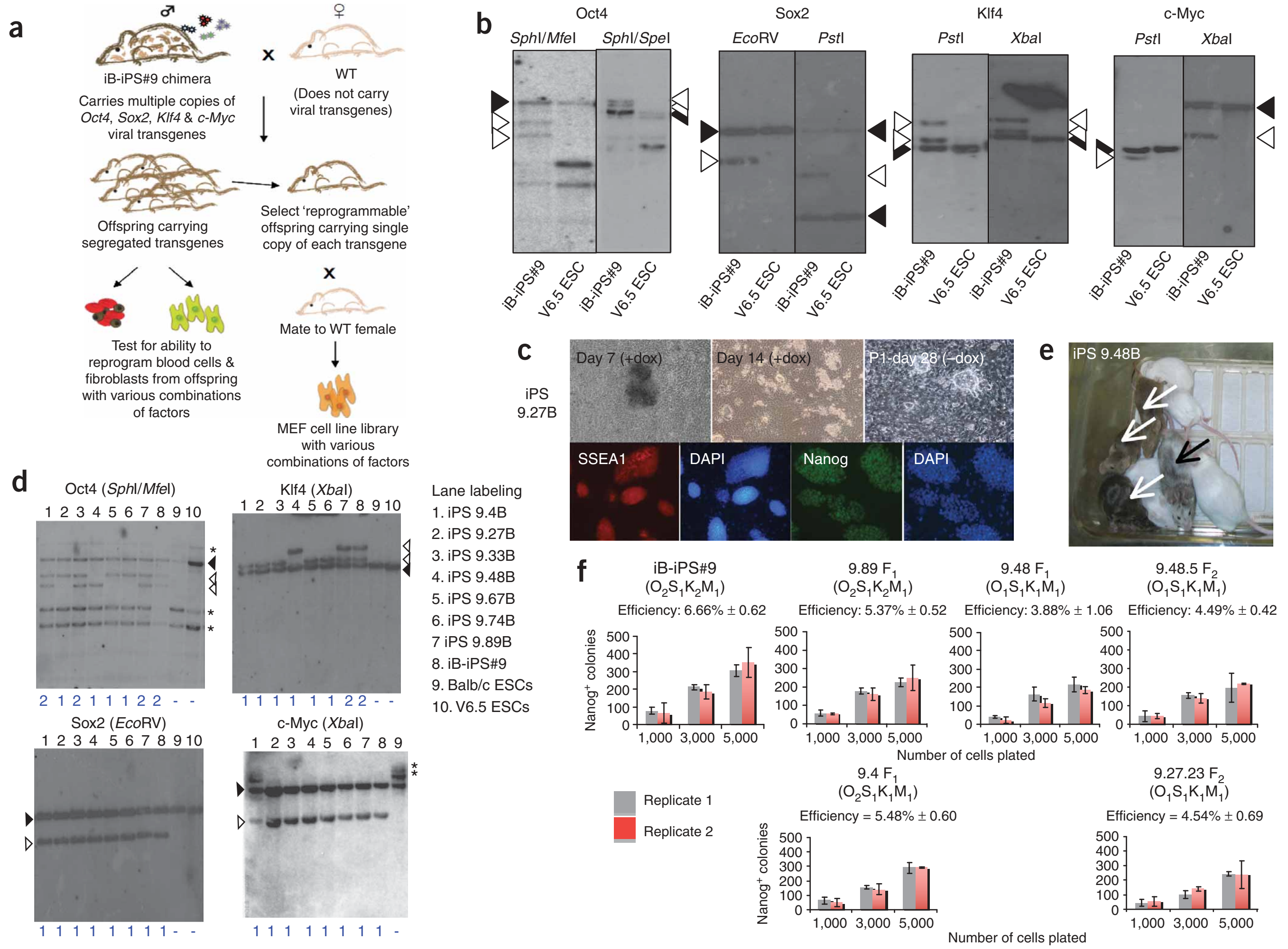
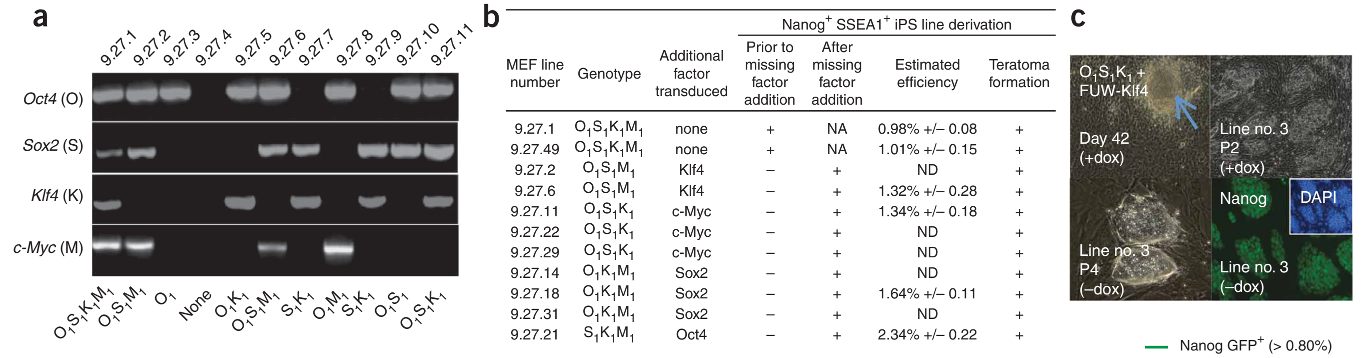


Figure 1 ‘Reprogrammable’ mice carrying single copies of reprogramming factors. (**a**) Experimental outline. iB-iPS#9 chimera[10] is mated to generate offspring with different transgene copy numbers. Blood and tail fibroblasts were collected from adult offspring, and MEF cultures were established from day E13.5 embryos. (**b**) Southern blot analysis of iB-iPS#9 cell line and V6.5 ES cells (ESC) as controls (**Supplementary Methods**). Filled arrowheads, endogenous bands; open arrowheads, proviral integrations. (**c**) Top panels: iPS cell colony formation from F_1 offspring 9.27 ($O_1S_1K_1M_1$). Immunofluorescence analysis of the same iPS cell line that grew independently of dox is shown in the lower panel. (**d**) Southern blot analysis of iPS cell lines derived from blood of F_1 progeny. *, nonspecific background bands. (**e**) iPS cells contribute to chimeras (black arrow) that exhibit germline transmission (transgenic offspring: white arrows). (**f**) Reprogramming efficiency of CD11b$^+$ cells 28 d after dox induction. Efficiencies were calculated as the fraction of Nanog$^+$ colonies to cells seeded. Error bars, s.d. in duplicate wells. The generation (F_1 or F_2) and transgene copy number (subscript) are shown. “B” indicates iPS cell line derived from peripheral blood. WT, wild type.



			Nanog$^+$ SSEA1$^+$ iPS line derivation			
MEF line number	Genotype	Additional factor transduced	Prior to missing factor addition	After missing factor addition	Estimated efficiency	Teratoma formation
9.27.1	$O_1S_1K_1M_1$	none	+	NA	0.98% +/– 0.08	+
9.27.49	$O_1S_1K_1M_1$	none	+	NA	1.01% +/– 0.15	+
9.27.2	$O_1S_1M_1$	Klf4	–	+	ND	+
9.27.6	$O_1S_1M_1$	Klf4	–	+	1.32% +/– 0.28	+
9.27.11	$O_1S_1K_1$	c-Myc	–	+	1.34% +/– 0.18	+
9.27.22	$O_1S_1K_1$	c-Myc	–	+	ND	+
9.27.29	$O_1S_1K_1$	c-Myc	–	+	ND	+
9.27.14	$O_1K_1M_1$	Sox2	–	+	ND	+
9.27.18	$O_1K_1M_1$	Sox2	–	+	1.64% +/– 0.11	+
9.27.31	$O_1K_1M_1$	Sox2	–	+	ND	+
9.27.21	$S_1K_1M_1$	Oct4	–	+	2.34% +/– 0.22	+

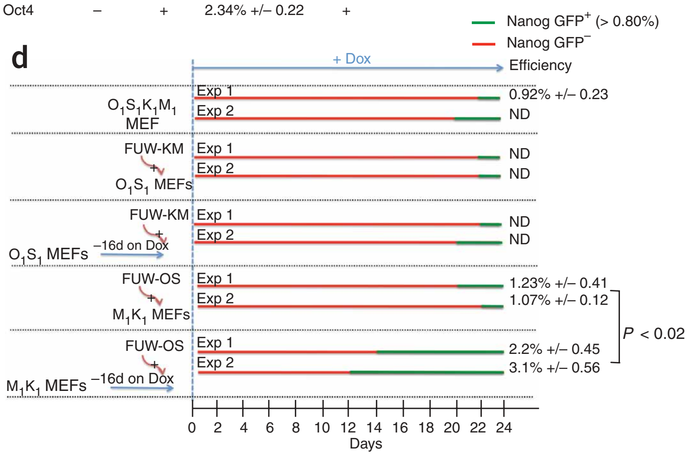


Figure 2 Library of MEF lines carrying different combinations of reprogramming factors. (**a**) PCR genotyping of select independent *ROSA26-M2rtTA*$^+$ MEF lines from mating offspring 9.27 ($O_1S_1K_1M_1$) to wild-type females. Genotype is indicated at the bottom. (**b**) iPS cell derivation from MEF lines carrying combinations of three or more factors. Missing factor was introduced by infection with TetO-FUW lentivirus (FUW) carrying the missing transcription factor. NA, not applicable; ND, not determined. The efficiencies reported are based on Nanog$^+$ colonies fixed 30 d after plating 10,000 cells and addition of dox. (**c**) iPS cells from three-factor MEF lines lacking *c-Myc* after transduction with *Klf4*. 200,000 $O_1S_1K_1$ MEFs were infected with the indicated control virus and cultured in the presence of dox without passaging. Image of primary colony on day 42 of dox induction after infection with FUW-*Klf4*. Primary colonies were picked and passaged without dox and expressed Nanog. Nine independent lines derived from two experiments. (**d**) Kinetics of Nanog-GFP knock-in allele expression in two-factor lines, pre-treated or not with dox, after transduction of the missing factors. 20,000 infected cells were seeded per well. Two wells were harvested every 48 h for detection of Nanog-GFP by FACS. Nanog-GFP was defined by achieving >0.8% GFP$^+$ cells. Blue dashed line, day of infection (d0). Pretreatment with dox was done for 16 d. Two independent experimental sets are shown. Efficiency was determined after 28 d of dox treatment as number of Nanog-GFP$^+$ colonies per 10,000 cells initially seeded.

all three *ROSA26-M2rtTA*, *c-Myc* and *Oct4* alleles (**Supplementary Fig. 9** online), indicating that *Oct4* reactivation can also act in concert with *c-Myc* in tumor formation.

In addition to their potential use in high-throughput drug screens, somatic cell lines and mouse strains that are genetically identical and possess different combinations of drug-inducible reprogramming factors at minimal copy numbers will be useful for the study of reprogramming mechanisms and for unraveling the mechanism of action of certain chemical compounds that modulate iPS cell generation, which remain largely unknown[1]. Such studies will enhance our understanding of how each of the reprogramming factors contributes to the rewiring of the transcriptional network and epigenetic state in differentiated somatic cells during the reprogramming process[1].

Mice carrying the inducible reprogramming factors will be deposited at the Jackson Laboratory for distribution as soon as animals that are homozygous for a given transgene combination have been obtained.

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

ACKNOWLEDGMENTS

We thank G. Bell, R. Flannery and members of the Jaenisch lab for assistance and comments. R.J. is supported by grants from the National Institutes of Health: 5-RO1-HDO45022, 5-R37-CA084198 and 5-RO1-CA087869. J.H. is a Novartis Fellow by the Helen Hay Whitney Foundation.

AUTHOR CONTRIBUTIONS

S.M., J.H. and R.J. conceived the project. S.M. and J.H. designed and performed all experiments and data analyses. C.B., S.W. and J.P.C. assisted in Southern blot analysis, C.J.L., B.W.C. and A.W.C. assisted in iPS cell derivation, efficiency experiments and expression analysis. J.A.D., D.F. and Q.G. assisted in genotyping, teratoma formation and pathology analysis. J.H., S.M. and R.J. 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/

1. Jaenisch, R. & Young, R. *Cell* **132**, 567–582 (2008).
2. Takahashi, K. & Yamanaka, S. *Cell* **126**, 663–676 (2006).
3. Wernig, M. *et al. Nature* **448**, 318–324 (2007).
4. Takahashi, K. *et al. Cell* **131**, 861–872 (2007).
5. Hanna, J. *et al. Science* **318**, 1920–1923 (2007).
6. Shi, Y. *et al. Cell Stem Cell* **3**, 568–574 (2008).
7. Marson, A. *et al. Cell Stem Cell* **3**, 132–135 (2008).
8. Huangfu, D. *et al. Nat. Biotechnol.* **26**, 1269–1275 (2008).
9. Wernig, M. *et al. Nat. Biotechnol.* **26**, 916–924 (2008).
10. Hanna, J. *et al. Cell* **133**, 250–264 (2008).
11. Wernig, M. *et al. Cell Stem Cell* **2**, 10–12 (2008).
12. Nakagawa, M. *et al. Nat. Biotechnol.* **26**, 101–106 (2008).

nature biotechnology

Antigen-specific human polyclonal antibodies from hyperimmunized cattle

Yoshimi Kuroiwa[1,2], Poothappillai Kasinathan[1], Thillainayagen Sathiyaseelan[1], Jin-an Jiao[1], Hiroaki Matsushita[1], Janaki Sathiyaseelan[1], Hua Wu[1], Jenny Mellquist[1], Melissa Hammitt[1], Julie Koster[1], Satoru Kamoda[2], Katsumi Tachibana[2], Isao Ishida[2] & James M Robl[1]

Antigen-specific human polyclonal antibodies (hpAbs), produced by hyperimmunization, could be useful for treating many human diseases. However, yields from available transgenic mice and transchromosomic (Tc) cattle carrying human immunoglobulin loci are too low for therapeutic applications. We report a Tc bovine system that produces large yields of hpAbs. Tc cattle were generated by transferring a human artificial chromosome vector carrying the entire unrearranged, human immunoglobulin heavy (h*IGH*) and κ-light (h*IGK*) chain loci to bovine fibroblasts in which two endogenous bovine *IgH* chain loci were inactivated. Plasma from the oldest animal contained >2 g/l of hIgG, paired with either human κ-light chain (up to ∼650 μg/ml, fully human) or with bovine κ- or λ-light chain (chimeric), with a normal hIgG subclass distribution. Hyperimmunization with anthrax protective antigen triggered a hIgG-mediated humoral immune response comprising a high proportion of antigen-specific hIgG. Purified, fully human and chimeric hIgGs were highly active in an *in vitro* toxin neutralization assay and protective in an *in vivo* mouse challenge assay.

hpAbs, produced from donated human plasma, have been used therapeutically for many years[1,2]. In an effort to improve effectiveness for specific disease applications, some products have been made in immunized humans[3], despite substantial challenges and restrictions. These include limitations on the types of vaccines used, number of immunizations permitted, types of adjuvant, amount of plasma that can be collected and dependence on voluntary donations. Alternatively, human plasma donors have been screened to select those with naturally high reactivity to specific antigens. Because hpAbs could be useful for treating many life-threatening human diseases, such as bacterial and viral infections, cancer and various autoimmune syndromes, an alternative hpAb production system is greatly needed[4,5].

Transgenic mice carrying the human immunoglobulin loci produce antigen-specific hpAbs in response to hyperimmunization[4], demonstrating that the mouse immune system can support human immunoglobulin gene rearrangement, affinity maturation and human antibody production following hyperimmunzation. Although human antibody–producing mice are ideal for generating human monoclonal antibodies, their small body size makes them unsuitable for producing practical amounts of therapeutic hpAbs. Large farm animals, such as cattle, could be a desirable source for therapeutic hpAbs because their size would enable them to produce a large quantity of antibodies after hyperimmunization with desired antigens.

Previously, we reported the generation of transchromosomic (Tc) cattle carrying a human artificial chromosome (HAC) vector comprising the entire, germline-configured, h*IGH* and h*IGL* chain loci[6]. Although human immunoglobulin gene rearrangement appeared normal in Tc cattle, the level of hIgG produced in their plasma was very low (∼10 μg/ml). We suspect that dominant expression of endogenous bovine IgG (bIgG) suppressed expression of hIgG. In mice transgenic for human immunoglobulin, disruption of endogenous murine immunoglobulin genes by gene targeting resulted in a significant increase in production of hIgG[4]. Therefore, inactivation of the endogenous bovine immunoglobulin gene(s) could enhance production of hpAbs in Tc cattle.

In comparison with those of mouse and human, little is known about immunoglobulin gene function and organization in cattle. Among IgH chain classes, the IgM heavy chain of mouse and human is encoded by a single gene, *IGHM*, which is the first to be expressed during early B cell development and is essential for B-cell development[7–9]. In contrast, large farm animals, such as sheep, goat and cattle, appear to possess two IgM loci[10]: the classical *IGHM* as well as an IgM-like (*IGHML1*) locus. In cattle, two distinct IgM sequences have been registered: U63637 (or AY149283) encodes *IGHML1* (located on chromosome 11; refs. 11–14), whereas AY230207 (or AY158087) encodes *IGHM* (mapped to chromosome 21; refs. 14,15). Although it is unknown whether the additional *IGHML1* locus is functional, if it supports B-cell development and IgG production in the absence of *IGHM*, then two heavy-chain gene knockouts (four targeting events) would be required to inactivate bovine immunoglobulin production.

Another potentially challenging problem associated with the use of a Tc bovine hpAb-production system is whether or not the human

[1]Hematech, Inc, 4401 S. Technology Dr., Sioux Falls, South Dakota 57106, USA. [2]Kyowa Hakko Kirin Co., Ltd., 1-6-1 Ohtemachi, Chiyoda-ku, Tokyo 100-8185, Japan. Correspondence should be addressed to J.M.R. (jrobl@hematech.com) or Y.K. (ykuroiwa@hematech.com).

Received 22 November 2008; accepted 23 December 2008; published online 18 January 2009; doi:10.1038/nbt.1521

immunoglobulin genes could support bovine B-cell development and humoral immunity in the absence of functional bovine immunoglobulin gene expression. Because the immune system in large farm animals is distinctly different from that of the mouse and human[16–22], successful production of hpAbs in the mouse is not necessarily indicative of success in cattle.

In this study, we first addressed the question of *IGHML1* function by generating and evaluating a series of IgM knockout cattle. We found that, surprisingly, each of the two IgM loci is fully functional and inactivation of both IgM loci is required for complete B-cell deficiency in cattle. Second, we investigated the function of a HAC vector (κHAC) comprising both h*IGH* and h*IGK* loci, in IgM double-knockout ($IGHM^{-/-}IGHML1^{-/-}$) cattle. We report here a detailed characterization of our first, κHAC/$IGHM^{-/-}IGHML1^{-/-}$ calf. Production of this calf (468) required five sequential genetic modifications and seven consecutive cloning events. Calf 468 continuously produced >2 g/l of hIgG in plasma, 10–20% of which (up to 649.1 μg/ml) was fully human hIgG (hIgG/hκ-chain). After hyperimmunization with anthrax protective antigen (PA), both fully human hIgG/hκ-chain and chimeric hIgG antibodies were found to be highly effective in an *in vitro* toxin-neutralization assay (TNA) and in an *in vivo* mouse protection assay. These results demonstrate the feasibility of using a bovine system to produce a large volume of highly active hpAbs for human therapy.

RESULTS

Generation and analysis of $IGHM^{-/-}$, $IGHML1^{-/-}$ and $IGHM^{-/-}IGHML1^{-/-}$ knockout cattle

We previously generated IgM knockout cattle by using a sequential gene targeting system, based on the U63637 sequence, which was the only one registered at that time[23]. By constructing and screening a genomic library made from the IgM knockout bovine fibroblast cell line, we found that our previous IgM knockout was indeed $IGHML1^{-/-}$; both alleles of *IGHML1*, designated as alleles *U* and *u*, were disrupted by the knockout cassettes, whereas the *IGHM* alleles, designated as *AY* and *ay*, were still intact (**Supplementary Fig. 1a** online).

To elucidate the involvement of both IgM loci, *IGHM* and *IGHML1*, in B-cell development in cattle, we generated $IGHML1^{-/-}$, $IGHM^{-/-}$ and $IGHM^{-/-}IGHML1^{-/-}$ knockout animals. To specifically knock out both alleles of the *IGHM* gene, we constructed the allele-specific knockout vectors pbCμayKOhyg and pbCμAYKObsr (from the alleles *ay* and *AY*, respectively), which were identified from the genomic library used previously[23] (**Supplementary Fig. 1b**). The wild-type bovine fibroblast line 6939 was transfected with pbCμayKOhyg to target allele *ay*, and $IGHM^{-/+}$ colonies were identified by PCR. Seventeen $IGHM^{-/+}$ colonies were identified from 210 (8.1%) hygromycin B-resistant colonies. To rejuvenate cells, we produced cloned embryos, collected four 40-d cloned fetuses and established fibroblast cell lines. All four cell lines were confirmed to be $IGHM^{-/+}$ by genomic PCR (**Supplementary Fig. 1c**). Evaluation of a polymorphic sequence within the PCR products demonstrated that the vector was exclusively integrated into allele *ay* of the *IGHM* gene in all four fetuses. One $IGHM^{-/+}$ cell line was then subjected to a second round of gene targeting to disrupt the second allele, *AY*, of *IGHM*, using a second knockout vector (pbCμAYKObsr). Fourteen $IGHM^{-/-}$ colonies were identified from 146 (9.6%) blasticidine-resistant colonies. After embryonic cloning of colonies, six rejuvenated cell lines were produced from fetuses recovered at 40 d. All proved to be $IGHM^{-/-}$ by genomic PCR (**Supplementary Fig. 1d**). Sequence analysis of the PCR products (AYKObsrF2 × AYKObsrR2) demonstrated that the second knockout vector was exclusively integrated into allele *AY* of the *IGHM* gene in all six fetuses. To generate the double-knockout $IGHM^{-/-}IGHML1^{-/-}$ cell lines, we further transfected the $IGHML1^{-/-}$ cell line established previously[23] with the knockout vectors (pbCμayKOhyg and pbCμAYKObsr) to sequentially disrupt the two alleles, *ay* and *AY*, of the *IGHM* gene. After two additional rounds of gene targeting (29 $IGHM^{-/+}IGHML1^{-/-}$ colonies were identified from 453 (6.4%) hygromycin B-resistant colonies; 26 $IGHM^{-/-}IGHML1^{-/-}$ colonies were identified from 215 (12.1%) blasticidine-resistant colonies), four fetuses were collected at 40 d and shown to be $IGHM^{-/-}IGHML1^{-/-}$ by genomic PCR (**Supplementary Fig. 1e**). Targeting frequencies at *IGHM* were substantially higher than those at the *IGHML1* locus (0.17–0.45%)[23], presumably due to use of allele-specific targeting vectors.

To verify specific disruption of each of the genes, we evaluated expression by RT-PCR analysis (primers; BL17 × mBCμR2) on spleen cells from $IGHM^{-/-}$, $IGHML1^{-/-}$, $IGHM^{-/-}IGHML1^{-/-}$ and wild-type control fetuses after 180 d of gestation (**Supplementary Fig. 1f**). All fetuses originated from the same primary bovine fibroblast line 6939, as described above. After sequence analysis of the amplified transcripts, we confirmed specific disruption of *IGHM* or *IGHML1* gene expression and expression of *IGHML1* or *IGHM*, in the $IGHM^{-/-}$ or $IGHML1^{-/-}$ fetuses, respectively. Gene expression was not detected from either of the two IgM genes in $IGHM^{-/-}IGHML1^{-/-}$ fetuses (**Supplementary Fig. 1g**). Although both *IGHM* and *IGHML1* transcripts were detected in wild-type fetuses, the level of expression of *IGHML1* appeared to be much lower than that of *IGHM*, indicating that *IGHML1* is a minor IgM class in the presence of *IGHM* in wild-type cattle.

$IGHM^{-/-}$, $IGHML1^{-/-}$ and $IGHM^{-/-}IGHML1^{-/-}$ cell lines were used to generate calves (**Table 1**) for comparison of B-cell development, immunoglobulin protein secretion and antigen-specific humoral immune response. Flow-cytometry analysis of peripheral blood mononuclear cells showed clear B-cell populations (CD21^{+},

Table 1 Production of cloned calves from genetically modified fibroblast cell lines

			Pregnant at (%)[a]				Calves survived more than 2 months (%)[a]
Cell line ID	Genotype	Recipients	40 d	90 d	150 d	270 d	
F056-2	$IGHM^{-/-}$	62	34 (55)	21 (40)	20 (38)	19 (37)	15 (29)
1638	$IGHM^{-/-}IGHML1^{-/-}$	49	23 (47)	9 (18)	7 (14)	2 (4)	2 (4)
261R	κHAC/$IGHM^{-/-}$	454	261 (57)	113 (25)	95 (21)	68 (15)	71 (16)
A254-2	κHAC/$IGHM^{-/-}IGHML1^{-/-}$	37	23 (62)	11 (30)	7 (19)	2 (5)	0 (0)
443	κHAC/$IGHM^{-/-}IGHML1^{-/-}$	213	84 (39)	6 (3)	5 (2)	2 (1)	1 (0.5)
Total		815	425 (52)	160 (20)	134 (17)	93 (12)	90 (11)

[a]Percentages were calculated by dividing the number of fetuses or calves by that of recipients implanted.

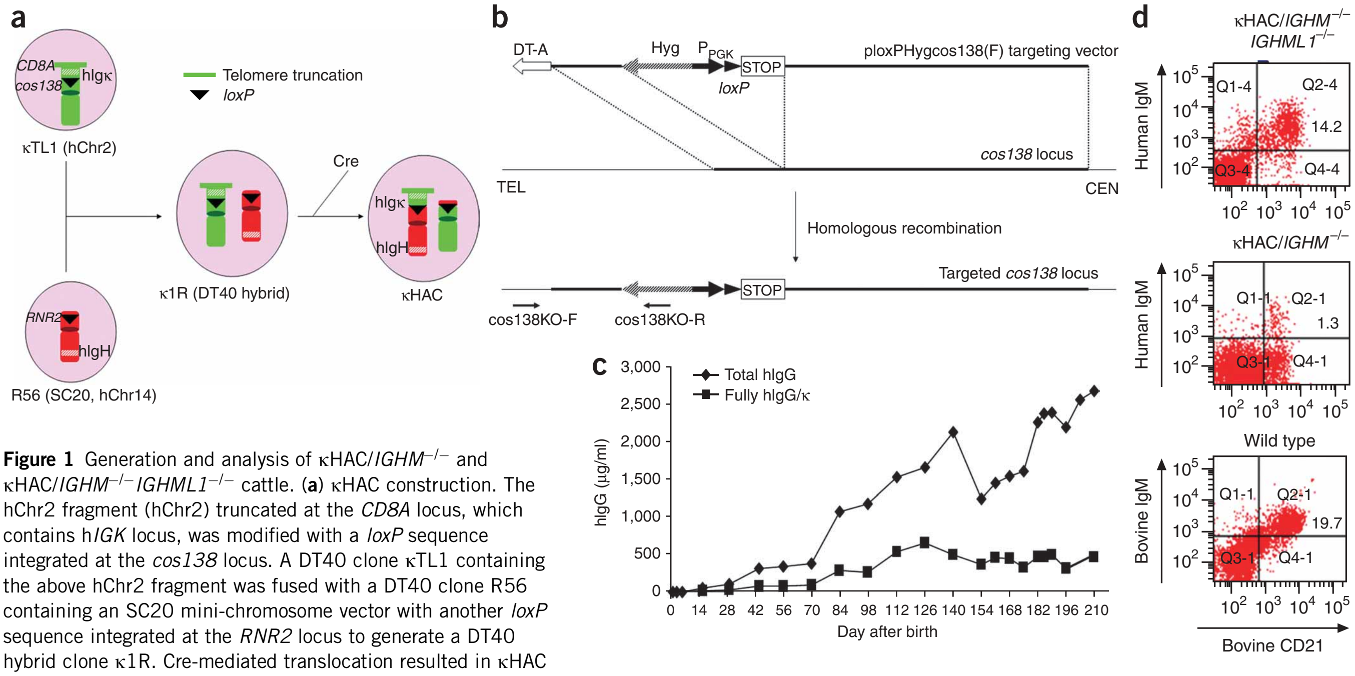


Figure 1 Generation and analysis of κHAC/*IGHM*$^{-/-}$ and κHAC/*IGHM*$^{-/-}$*IGHML1*$^{-/-}$ cattle. (**a**) κHAC construction. The hChr2 fragment (hChr2) truncated at the *CD8A* locus, which contains h*IGK* locus, was modified with a *loxP* sequence integrated at the *cos138* locus. A DT40 clone κTL1 containing the above hChr2 fragment was fused with a DT40 clone R56 containing an SC20 mini-chromosome vector with another *loxP* sequence integrated at the *RNR2* locus to generate a DT40 hybrid clone κ1R. Cre-mediated translocation resulted in κHAC comprising both entire *hIGH* and h*IGK* loci. (**b**) Targeting vector ploxPHygcos138(F). The vector comprises a 5′ homologous arm, a 3′ homologous arm, STOP cassette containing transcriptional and translational stop sequence, a *loxP* sequence, PGK promoter, DT-A (diphtheria toxin A gene) and *hyg* gene. A primer pair, cos138KO-F x cos138KO-R, was used to identify the event of homologous recombination. (**c**) Total hIgG and fully human hIgG/hκ-chain (hIgG/κ) levels in calf 468 serum for 7 months after birth. (**d**) B-cell population in Ileal Peyer's patch of κHAC/*IGHM*$^{-/-}$ *IGHML1*$^{-/-}$, κHAC/*IGHM*$^{-/-}$ and control bovines, stained with anti-CD21 and hIgM (or bIgM) antibodies.

IgM^{+}B220^{+}) in both *IGHM*$^{-/-}$ and *IGHML1*$^{-/-}$ calves, whereas no B cells were detected in *IGHM*$^{-/-}$*IGHML1*$^{-/-}$ calves (**Supplementary Fig. 1h**). B220^{+}IgM^{-} cells were detected in *IGHM*$^{-/-}$*IGHML1*$^{-/-}$ calves and could be pro-B cells (the stage before IgM cell surface expression), because an IgM knockout can not ablate pro-B-cell generation. We have also performed RT-PCR analysis for $V_HD_HJ_H$-rearranged bovine IgD and IgG transcripts and were not able to detect the transcripts in *IGHM*$^{-/-}$*IGHML1*$^{-/-}$ bovines, suggesting that the B220^{+}IgM^{-} cells are neither IgD^{+} nor IgG^{+} B cells. Furthermore, the cells were not CD21^{+}, which should be the case for either IgD^{+} or IgG^{+} B cells. Within the first day after birth, before colostrum administration, we detected secreted IgM protein in sera of the *IGHM*$^{-/-}$ (4–11 μg/ml) and the *IGHML1*$^{-/-}$ (4–7 μg/ml) calves, at levels comparable to controls (8–21 μg/ml). No secreted IgM protein was detected in the *IGHM*$^{-/-}$*IGHML1*$^{-/-}$ calves (the detection limit of this enzyme-linked immunosorbent assay (ELISA) is 0.4 μg/ml). IgG protein was detected in the sera of the *IGHM*$^{-/-}$ (8–11 μg/ml), *IGHML1*$^{-/-}$ (6–11 μg/ml) and, surprisingly, in *IGHM*$^{-/-}$*IGHML1*$^{-/-}$ (4–11 μg/ml) calves. The IgG protein detected in sera of *IGHM*$^{-/-}$ *IGHML1*$^{-/-}$ calves is likely to have come from the mother, possibly through the placenta, because *IGHG* transcripts were not detected in peripheral blood mononuclear cells (**Supplementary Fig. 1i**).

When calves were 3–4 months of age, we detected high levels of IgG protein in the *IGHM*$^{-/-}$ and *IGHML1*$^{-/-}$ calves (30–42 mg/ml and 34–41 mg/ml, respectively). Furthermore, both types of calves responded to immunization with titers comparable to wild-type controls (**Supplementary Fig. 1j,k**). These data demonstrate that, in contrast to the mouse and human, cattle possess two fully functional IgM loci, *IGHM* and *IGHML1*, each capable of supporting B-cell development and antigen-specific humoral immune response. For a complete inactivation of immunoglobulin gene function in cattle, both loci need to be disrupted.

Generation and analysis of κHAC/*IGHM*$^{-/-}$ and κHAC/*IGHM*$^{-/-}$ *IGHML1*$^{-/-}$ cattle

Previously[6], we introduced a HAC, carrying both h*IGH* and h*IGK* chain loci (ΔΔHAC), into cattle to produce hIgG. To improve the level of expression of hIgG, we considered constructing a different HAC for this study. As rearrangement and expression of the *IGK* locus precedes that of the λ-light chain locus (*IGL*)[24] in human and mouse, the h*IGK* locus might compete with bovine immunoglobulin light chain loci (b*Igl*) better than the h*IGL* locus because the immunoglobulin λ-light chain is the predominant light chain expressed in cattle[21]. Furthermore, human κ-chain normally represents more than half of the total human immunoglobulin light chain (κ/λ ratio = 60/40)[25] expressed in human. Based on this rationale, we attempted to construct a HAC vector comprising the entire loci for both h*IGH* and h*IGK* chain genes (κHAC) using a chromosome-cloning system[26] (**Fig. 1a,b**). The κHAC was introduced into either *IGHM*$^{-/-}$ or *IGHM*$^{-/-}$*IGHML1*$^{-/-}$ bovine fibroblasts by microcell-mediated chromosome transfer and calves were generated by embryonic cloning (**Table 1**).

The κHAC/*IGHM*$^{-/-}$*IGHML1*$^{-/-}$ cell line (A254-2) generated calves at lower efficiency than the κHAC/*IGHM*$^{-/-}$ cell line (**Table 1**), possibly because of the additional two rounds of embryonic cloning (total of six) required to knock out both IgM loci. Of the two male calves produced, calf 445 died shortly after birth, whereas calf 443 survived to 40 d and produced 541 μg/ml of total hIgG (fully human hIgG/hκ-chain + chimeric) in the serum. This level was substantially higher than that in our previous ΔΔHAC calves (∼10 μg/ml). We established a fibroblast cell line from calf 443 and conducted an additional (seventh) round of embryonic cloning, which gave rise to one healthy calf, 468. The scheme for generation of calf 468 is summarized in **Supplementary Figure 2** online.

From birth, calf 468 showed a substantial increase in hIgG; reaching > 1 g/l in serum at 84 d of age (**Fig. 1c**). Human IgM was also detected

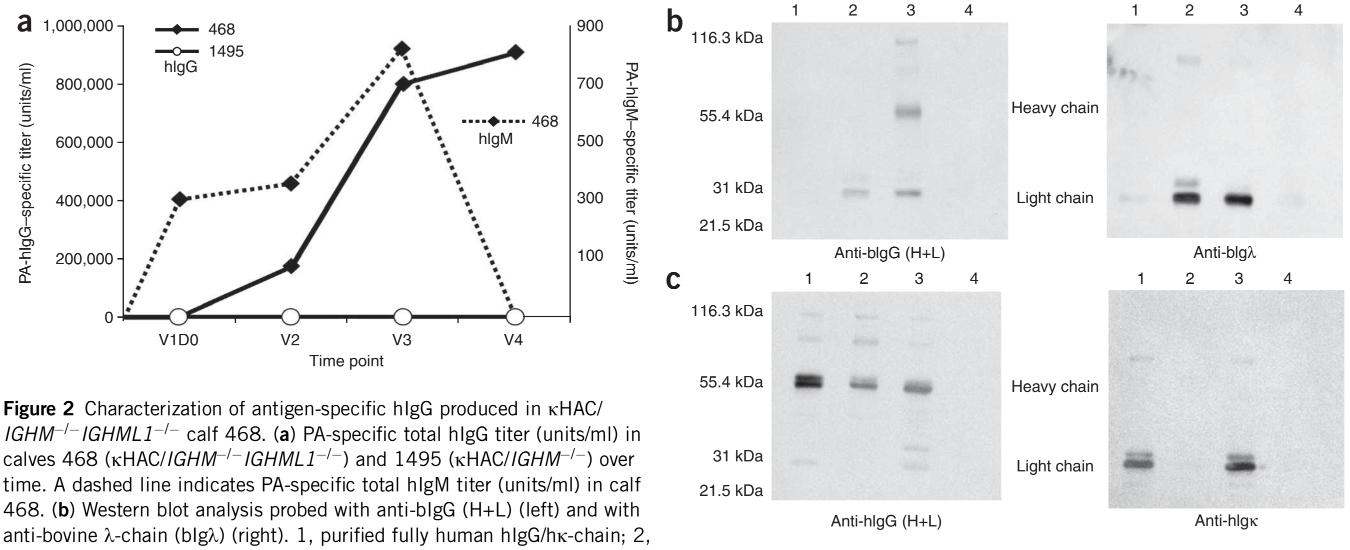


Figure 2 Characterization of antigen-specific hIgG produced in κHAC/$IGHM^{-/-}IGHML1^{-/-}$ calf 468. (**a**) PA-specific total hIgG titer (units/ml) in calves 468 (κHAC/$IGHM^{-/-}IGHML1^{-/-}$) and 1495 (κHAC/$IGHM^{-/-}$) over time. A dashed line indicates PA-specific total hIgM titer (units/ml) in calf 468. (**b**) Western blot analysis probed with anti-bIgG (H+L) (left) and with anti-bovine λ-chain (bIgλ) (right). 1, purified fully human hIgG/hκ-chain; 2, purified chimeric hIgG; 3, purified commercial bIgG; 4, purified commercial hIgG. (**c**) Western blot probed with anti-hIgG (H+L) (left) and with anti-human κ-chain (right). 1, purified fully human hIgG/hκ-chain; 2, purified chimeric hIgG; 3, purified commercial hIgG; 4, purified commercial bIgG.

and the hIgM/hIgG ratio was 1.2% on average (**Supplementary Table 1** online). In contrast, hIgG level in κHAC/$IGHM^{-/-}$ calves never exceeded ~10 μg/ml. Because the bovine immunoglobulin light chain genes (b*Igl* and b*Igk*) were not inactivated, we also measured the level of fully human hIgG (hIgG/hκ-chain) by a sandwich ELISA. Fully human hIgG/hκ-chain consisted of ~10–20% of total hIgG detected in the serum, which reached levels as high as 649 μg/ml (**Fig. 1c**). The hIgG subclass distribution in calf 468 was similar to that observed in human (**Supplementary Table 2** online).

To evaluate B-cell development in κHAC/$IGHM^{-/-}$ and κHAC/$IGHM^{-/-}IGHML1^{-/-}$cattle, flow cytometry analysis was performed on cells from Ileal Peyer's patch, the major lymphoid tissue where B-cell development, proliferation and immunoglobulin diversification extensively occur in cattle and other gut-associated lymphoid tissue animals[16,20,22] (**Fig. 1d**). κHAC/$IGHM^{-/-}IGHML1^{-/-}$ bovines showed improved B-cell development (hIgM$^+$/CD21$^+$ mature B cells) compared to κHAC/$IGHM^{-/-}$ animals and were comparable to controls. The data suggest that bovine B-cell development can be supported by hIgM in the complete absence of bIgM.

Characterization of antigen-specific hIgG produced in κHAC/$IGHM^{-/-}$ $IGHML1^{-/-}$ cattle

At the age of 112 d, we immunized calf 468 with anthrax PA[27] to examine the hIgG-mediated humoral immune response (**Fig. 2a**). At vaccination booster 2 (V2), calf 468 began to show a PA–specific hIgG response and reached a high titer at V4. The titer at V4 was higher than the bIgG titer in a control calf 1495 (κHAC/$IGHM^{-/-}$) and comparable to PA-specific bIgG titers obtained in wild-type cattle after V16 (**Supplementary Table 3** online). Furthermore, the anti-PA titer obtained from calf 468 was substantially higher than the hIgG anti-PA titer in a human reference serum (AVR801; **Supplementary Table 3**) obtained from donors after four vaccinations with Anthrax Vaccine Adsorbed. On the other hand, as expected, there was no detectable PA-specific hIgG titer in the κHAC/$IGHM^{-/-}$ control calf 1495 (**Fig. 2a**). This suggests that the κHAC/$IGHM^{-/-}IGHML1^{-/-}$ genotype is useful in generating high titer, antigen-specific hIgG after antigen immunization.

For characterization of the PA-specific hIgG produced in cattle, human IgG was purified from the plasma of calf 468 collected after V4 by plasmapheresis. To check the purity of the fully human hIgG/hκ-chain fraction, we performed SDS-PAGE and western blot analysis using anti-bovine IgG (heavy and light chains; H+L) and anti-bovine immunoglobulin λ-chain polyclonal antibodies. There were neither bIgG heavy nor light chain bands detected in the fully human hIgG/hκ-chain fraction (**Fig. 2b**). On the contrary, detection of hIgG heavy and human κ-light chains by anti-hIgG (H+L) polyclonal antibodies and anti-human κ-chain polyclonal antibodies, respectively (**Fig. 2c**), indicates that the fully human hIgG/hκ-chain fraction indeed contains both hIgG heavy and human κ-light chains. We also analyzed human heavy chain and bovine light chain chimeric hIgG obtained from the flow-through fraction of the anti-human κ-chain Sepharose column. The chimeric hIgG fraction was positive for hIgG heavy and bovine light chains, but negative for human κ-light and bIgG heavy chains (**Fig. 2b**,**c**).

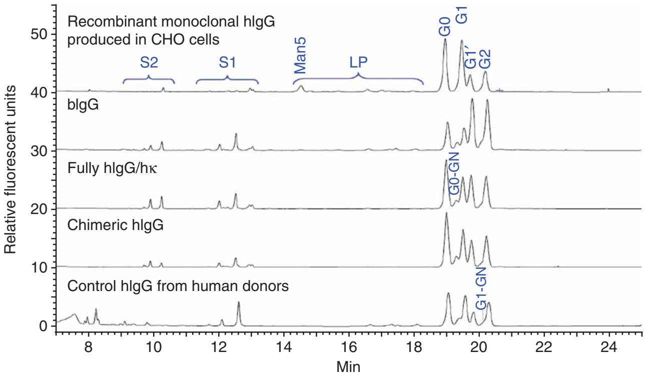


Figure 3 Glycosylation analysis of antigen-specific hIgG produced in κHAC/$IGHM^{-/-}IGHML1^{-/-}$ calf 468. Capillary gel electrophoresis with helium-cadmium laser–induced fluorescent detection (CE-LIF) on recombinant monoclonal hIgG produced in CHO cells, bIgG from wild-type cattle, fully human hIgG/hκ-chain (hκ) from calf 468, chimeric hIgG from calf 468 and hIgG from human donors (polyglobin-N). S1-S2, monosialyl and bisialyl acids (sialic acid content); LP, mannose and/or afucosylation (fucosylation content); G0, G1, G1′, G2; gal structure (gal content), G0-GN, G1-GN; GlcNAc (GlcNAc content).

Table 2 Toxin neutralization activities

	No. of vaccinations	IgG conc (g/l)	TNA (ED_{50})[a]	TNA (EC_{50})[b] (µg)
Wild-type bovine pooled hyperimmune purified bIgG	16	10.4	10,090	1.0
Calf 468–derived purified total hIgG	4	17.7	12,377	1.4
Calf 468–derived purified chimeric hIgG	4	18.4	13,143	1.4
Calf 468–derived purified fully human hIgG/hκ-chain	4	21.1	11,890	1.8
Human pooled immune serum (AVR 801)	4	5.3	111	57.0

[a] TNA ED_{50} is the dilution of the antibody solution or serum that neutralizes 50% of total cell cytotoxicity by the anthrax toxin. [b] TNA EC_{50} is the amount (µg) of antibody required to neutralize 50% of total cell cytotoxicity by the anthrax toxin.

Furthermore, the percentage of PA-specific IgG fraction was estimated by using a PA-immobilized Sepharose affinity column. Purified bIgG from the control calf 1495 (κHAC/$IGHM^{-/-}$), as well as fully human hIgG/hκ-chain and chimeric hIgG from calf 468, were loaded onto the PA affinity column and the PA-specific IgG fraction was eluted at pH 2.5. Compared to control calf 1495, an unusually high proportion of PA-specific antibody, both fully human hIgG/hκ-chain (13%) and chimeric hIgG (35%), was produced by hyperimmunization of calf 468 (**Supplementary Table 4** online).

Glycosylation analysis of antigen-specific hIgG produced in κHAC/$IGHM^{-/-}$ $IGHML1^{-/-}$ cattle

As the IgG heavy chain is glycosylated at its Fab and Fc regions in a species-specific manner[28], we investigated *N*-linked oligosaccharides both in the fully human hIgG/hκ-chain and chimeric hIgG fractions by capillary gel electrophoresis with helium-cadmium laser–induced fluorescent detection (CE-LIF; **Fig. 3** and **Supplementary Table 5** online). When compared with monoclonal hIgG produced in CHO (Chinese hamster ovary) cells and polyclonal hIgG control from human donors, the glycosylation profile of hIgG (both hIgG/hκ-chain and chimeric hIgG) produced in calf 468 appears to be more similar to that of the polyclonal hIgG control. One minor difference between the bovine-derived hIgG and the control human-derived polyclonal hIgG is in the LP peak, which is thought to contain fucose-less sugar chains. However, the LP peak is similarly minor even in the human control. S1 and S2 peaks contain a sugar chain to which sialic acid is added. The sialic acids, *N*-acetylneuraminic acid (NANA) and *N*-glycolylneuraminic acid (NGNA), were analyzed by reversed-phase high-performance liquid chromatography (HPLC) and fluorescence detection (**Supplementary Table 6** online). Total content of sialic acid is similar between calf 468–derived hIgG and the control human-derived hIgG. However, the ratio of NANA/NGNA is different as expected: calf 468–derived hIgG has predominantly NGNA (similar to the control bIgG[28]), whereas the control polyclonal hIgG exclusively has NANA. With respect to branched sugar chains (G0-G2), the contents of galactose (galactose residue per *N*-glycan) and *N*-acetylglucoseamine (GlcNAc) (G0-GN and G1-GN) are similar between the calf 468–derived hIgG and the control human-derived hIgG.

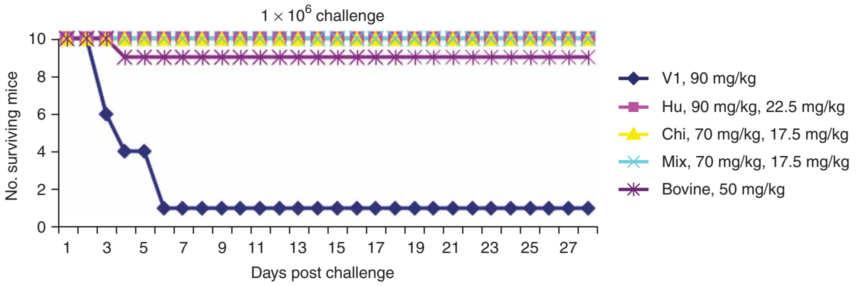


Figure 4 *In vivo* mouse protection assay of PA-specific hIgG produced in κHAC/$IGHM^{-/-}$ $IGHML1^{-/-}$ calf 468. V1, purified total hIgG from calf 468 at V1 of PA-immunization; Hu, purified fully human hIgG/hκ-chain from calf 468 at V4 of PA-immunization; Chi, purified chimeric hIgG from calf 468 at V4 of PA-immunization; Mix, purified total hIgG from calf 468 at V4 of PA-immunization; Bovine, hyperimmunized pooled wild-type bIgG at V16 of PA-immunization.

TNA and mouse protection assay of PA-specific hIgG produced in κHAC/$IGHM^{-/-}$ $IGHML1^{-/-}$ cattle

The purified fully human hIgG/hκ-chain and chimeric hIgG fractions containing the binding activity against PA antigen at V4 were evaluated by the TNA[29,30] (**Table 2**). The TNA of hIgG produced in calf 468 (both purified fully human hIgG/hκ-chain and chimeric hIgG fractions) is comparable to that of hyperimmunized wild-type bIgG and much higher than that of the human reference. Our PA-challenge mouse protection assay (**Fig. 4**) involved challenging mice with 1×10^6 anthrax (Sterne strain) spores. Mice were given either 90 mg/kg of total hIgG produced in calf 468 at V1 (contained little activity); 90 mg/kg of fully human hIgG/hκ-chain or 70 mg/kg of chimeric hIgG or 70 mg/kg of total hIgG (hIgG/hκ-chain + chimeric hIgG) from calf 468 at V4; or 50 mg/kg of hyperimmunized pooled wild-type bIgG at V16. IgG doses were standardized to contain equivalent TNA activity in the purified fraction. With the negative control (bovine-derived hIgG at V1), nine out of ten mice died, whereas both fully human hIgG/hκ-chain and chimeric hIgG collected from calf 468 at V4 completely protected all ten mice. The hyperimmune pooled bIgG resulted in the death of one of the ten mice tested. This complete protection activity was also observed with 22.5 mg/kg and 17.5 mg/kg of fully human hIgG/hκ-chain and chimeric hIgG, respectively, from calf 468 at V4 (**Fig. 4**). These data suggest that hIgG produced in calf 468 (both fully human hIgG/hκ-chain and chimeric hIgG) was fully functional and effective in neutralizing the toxin activity *in vitro* and *in vivo*.

DISCUSSION

This study demonstrates the feasibility of producing a large quantity of highly active, antigen-specific hpAbs in a large farm animal species. Calf 468 produced over 2 g/l of total serum hIgG (fully human and chimeric). Moreover, we showed that hyper-immunization with a PA antigen resulted in high *in vitro* and *in vivo* neutralization potency. The high activity may be attributed to an unusually high percentage of PA-specific, fully human and chimeric hIgG. In the human reference serum AVR801, the percentage of PA-specific hIgG is estimated to be 2.1%[31]. The high antigen specificity should be beneficial for therapeutic applications.

To generate a Tc calf capable of producing a large volume of functional hIgG, several difficult challenges were addressed. We have shown that, unlike mouse and human, cattle have two independent pathways for B-cell development regulated by the two distinct fully functional IgM heavy chain loci, *IGHM* and *IGHML1*, and that the inactivation of both these loci is critical for producing large quantities of hIgG. This is the first demonstration of a mammalian species that has multiple fully functional *IgH* loci. As other ungulates, such as sheep and goat, may also possess a similar *IGHML1* locus in addition to the classical *IGHM* gene[10], the double-knockout approach of the two IgM loci may be equally useful for production of hpAbs in other large farm animals.

Another challenge was to produce a viable calf following four gene-targeting events and insertion of a HAC: a total of five sequential genetic modifications and six cloning procedures. Calf 468 was actually produced from a seventh cloning procedure. As additional κHAC/$IGHM^{-/-}IGHML1^{-/-}$ calves have comparably high levels of serum hIgG (**Supplementary Table 7** online), the κHAC/$IGHM^{-/-}$ $IGHML1^{-/-}$ genotype appears to be useful for producing a large quantity of hIgG. However, the low rate of development to term and relatively high incidence of mortality after birth are impediments for commercial production. Results presented in several studies show dramatic declines in the efficiency of cloning with successive cloning procedures[32–36]. One possible reason for the decrease in efficiency is the accumulation of epigenetic errors, including imprinting errors, induced by embryonic cloning. To solve this potential problem, we have incorporated a plan to produce IgM double-knockout ($IGHM^{-/-}IGHML1^{-/-}$) cell lines by mating. Our preliminary results of breeding between highly recloned male ($IGHM^{-/+}IGHML1^{-/-}$) and female ($IGHM^{-/+}IGHML1^{-/+}$) parents indicate that the rate of development to term of calves and of survival after birth is improved to a level similar to that of calves derived from *in vitro* fertilization.

Stability of κHAC was examined from both mitotic and structural perspectives. The former was done by fluorescent *in situ* hybridization (FISH) analysis using human COT1 DNA as a probe. As >90% of cells observed generally retain κHAC as a single copy–independent chromosome both in peripheral blood lymphocytes and fibroblasts for at least several years, κHAC appears to be mitotically stable during development. We tested structural stability by genomic PCR mapping with 16 markers dispersed over the entire HAC structure. Most of the animals tested were positive for all the markers (13 out of 15 animals), with the exception of two calves which were missing some markers. Overall, κHAC is retained at a high rate and with high fidelity during development.

In the current Tc bovine system, ~80% of total serum hIgG produced is chimeric; consisting of human IgG heavy and bovine immunoglobulin light chains. As the chimeric hIgG is fully functional and is likely not highly immunogenic, a mixture of chimeric and fully human hIgG could be safer and more useful than fully animal-derived polyclonal antibodies, for single, or minimal repetitive, dose treatments. However, fully human hIgG would be preferred for applications that require long-term, repetitive treatments. Fully human hIgG could be derived by purification, as demonstrated in this study. Notably, the serum level of fully human hIgG/hκ-chain in calf 468 was ~500 μg/ml in spite of the presence of the bovine immunoglobulin light chain genes. This level is comparable to that of hIgG-producing transgenic mice in which both murine *Igh* and *Igk* genes are knocked out[8]. The human IgG heavy chain may preferentially pair with human κ-light chain, rather than with bovine immunoglobulin light chain. However, knocking out the bovine immunoglobulin light chain genes would be preferable for higher yields of fully human hIgG.

In the present type of genetic modification (κHAC/$IGHM^{-/-}$ $IGHML1^{-/-}$), other classes of chimeric IgG heavy chain—for example, trans-class switched or trans-spliced IgG heavy chain—could be generated[37,38]. Because the bovine Cγ region is still intact in the $IGHM^{-/-}IGHML1^{-/-}$ double knockout, a heavy chain comprising human $V_HD_HJ_H$ and bovine Cγ sequences could be produced. To investigate this possibility, we performed RT-PCR with one primer located in human V_H and the other in bovine Cγ sequence from two κHAC/$IGHM^{-/-}IGHML1^{-/-}$ newborn calves. We detected human $V_HD_HJ_H$ and bovine Cγ-comprising transcript in the sample from one animal (**Supplementary Fig. 3** online) and the result was confirmed by sequencing. Another issue concerning the $IGHM^{-/-}IGHML1^{-/-}$ double knockout is the possibility that bovine $V_HD_HJ_H$ and bovine Cγ-comprising transcripts could be generated from an in *cis* class switch mechanism on the bovine *IgH* locus once hIgM$^+$ B cells are activated. To investigate this possibility, we conducted RT-PCR to amplify $V_HD_HJ_H$-rearranged bovine *IGHG* transcripts from two animals and detected bovine *IGHG* transcripts at low levels, with confirmation by sequence analysis (**Supplementary Fig. 3** online). Both chimeric heavy chain and fully bovine bIgG heavy chain are removed by our purification process and are not detected after purification (**Fig. 2b**).

It has been suggested that cattle can use gene conversion for immunoglobulin gene diversification[19–22]. Gene conversion might cause small segments of bovine V (or pseudo V) sequence to be placed into the human V sequence. We investigated this possibility using RT-PCR to amplify human $V_HD_HJ_H$-rearranged human Cγ transcripts from four κHAC/$IGHM^{-/-}IGHML1^{-/-}$ animals (primers used in this RT-PCR also amplify bovine sequence). The RT-PCR products were subcloned for sequence analysis (31 subclones were analyzed). Excluding the CDR3 region (D_H segment), sequence analysis showed >90% homology with human sequence (V_H and J_H) and no obvious trace of bovine sequence was detected.

Both polyclonal antibodies collected from human plasma donors and monoclonal antibodies produced by fermentation have been extraordinarily beneficial for treating a wide variety of human diseases. Our Tc bovine system for production of hpAbs may help to expand the repertoire of diseases that can be successfully treated using antibody-based therapeutics.

METHODS

All animal procedures were performed in compliance with Hematech's guidelines, and protocols were approved by the Institutional Animal Care and Use Committee.

Construction of genomic library and library screening. Genomic DNA was extracted from the $IGHML1^{-/-}$ fibroblast cell line 4658, originally derived from a primary bovine fibroblast line 6939 and a λ-phage-based genomic library was constructed using λFIX II vector through a custom library construction service (Lofstrand). A PCR product amplified with a primer pair (bCμf2 × bCμr2) was ^{32}P-labeled using Rediprime II DNA Labeling System kit (Amersham Biosciences) according to the manufacturer's manual, to use as a probe. This probe was able to hybridize to exon 2-3 of both *IGHM* and *IGHML1* genes. Plaque hybridization was carried out under a standard protocol. Positive phage plaques hybridized with the probe were propagated and DNA was extracted and purified using Wizard Lambda Preps DNA Purification System kit (Promega) according to the manufacturer's manual. The phage clones were classified into four alleles based on sequence identity. Alleles *U* and *u* contained the *puro* and *neo* STOP knockout cassettes[23], and essentially matched the sequence of the *IGHML1* locus (U63637 and AY149283) as expected. Alleles *AY* and *ay* were intact and matched the sequence of the *IGHM* locus (AY230207 and AY158087).

Construction of targeting vectors. The 7.5 kb of *Sal*I-*Bgl*II genomic fragment (5′ homologous arm) and 2.0 kb of *Bgl*II-*Bam*HI fragment (3′ homologous arm) around the exon 2 of alleles *ay* and *AY* of *IGHM* gene were subcloned into pBluescript II SK(–) (Stratagene), and then *hyg* or *bsr*, STOP cassette (Stratagene) and DT-A (diphtheria toxin A) genes were inserted (pbCμayKOhyg vector and pbCμAYKObsr vectors, respectively), as previously described[23]. For ploxPHygcos138 (F), genomic sequence of *cos138* was amplified with a primer; cos138-F6B × cos138-R6B, and cloned to the *Bam*HI site in pBluescript II SK(–). Hyg-PGK-*loxP* cassette[26] was cloned to the *Spe*I site in the *cos138* genomic sequence, followed by DT-A subcloning. Primer sequences: cos138-F6B (5′-TCGAGGATCCCACATAGACATTCAACCGCAAAGCAG-3′), cos138-R6B (5′-TCGAGGATCCAGGCCCTACACATCAAAAAGTGAAGCAG-3).

Construction of κHAC vector. κHAC vector was constructed using a previously described chromosome-cloning system[6,26]. Briefly, a DT40 clone, containing a hChr2 fragment truncated at the *CD8A* locus, was electroporated (550 V, 25 μF) with ploxPHygcos138 (F) targeting vector (25 μg) to integrate a *loxP* sequence at the *cos138* locus. Colonies were selected by hygromycin B (1.5 mg/ml) for 2 weeks and their DNA was subjected to PCR screening with cos138KO-F × cos138KO-R primers under the following conditions: 98 °C for 10 s, and 65 °C for 8 min in 40 cycles. A clone κTL1 was identified and fused to a DT40 clone (R56) containing the stable and germline-transmittable human microchromosome vector, SC20. The SC20 vector contained a *loxP* sequence integrated at the *RNR2* locus[26]. The resulting DT40 hybrids contained the two human chromosome fragments. The DT40 hybrid clone (κ1R) was then transfected with a Cre recombinase-expression vector to induce Cre/*loxP*-mediated chromosomal translocation between the hChr2 fragment and the SC20 vector. The stable transfectants were analyzed using nested PCR[26] to confirm the occurrence of chromosomal translocation. FISH analysis and fluorescent-activated cell sorting (FACS) of green fluorescent protein–expressing cells[26] were also used to confirm the presence of κHAC. Primer sequences: cos138KO-F (5′-TCTTTCTCTCACCTAATTGTCCTGGC-3′), cos138KO-R (5′-AGGACTGGCACTCTTGTCGATACC-3′).

Genetic modification of bovine fibroblasts. Bovine fetal fibroblasts were cultured and transfected as previously described[23]. Briefly, fibroblasts were electroporated with 30 μg of pbCμayKOhyg or pbCμAYKObsr vector at 550 V and 50 μF. After 48 h, the cells were selected under 200 μg/ml of hygromycin B or 10 μg/ml of blasticidine-HCl for 2 weeks and resistant colonies were picked up and transferred to replica plates; one was for genomic DNA extraction and the other was for embryonic cloning. Microcell-mediated chromosome transfer was done with the κHAC vector as described previously[6].

Genomic PCR analyses. Genomic DNA was extracted from the replica 24-well plates, fetuses or ear biopsies from calves, using a Puregene DNA extraction kit (GentraSystem). For genotyping *IGHML1*$^{-/-}$, primer pairs PuroF2 × PuroR2 and NeoF3 × NeoR3 were used as described previously[23]. To identify heterozygous *IGHM*$^{-/+}$ genotype, primer pair ayKOhygF2 × ayKOhygR2 was used. Forty cycles of PCR were performed by incubating the reaction mixtures in the following conditions: 98 °C for 10 s, and 68 °C for 8 min. To identify homozygous *IGHM*$^{-/-}$ genotype, primer pair AYKObsrF2 x AYKObsrR2 was used, together with ayKOhygF2 × ayKOhygR2 primers, as above. For genotyping *IGHM*$^{-/-}$*IGHML1*$^{-/-}$, all the four primer pairs ayKOhygF2 × ayKOhygR2, AYKObsrF2 × AYKObsrR2, PuroF2 × PuroR2 and NeoF3 × NeoR3 were used. All the PCR products were run on 0.8% agarose gels. Primer sequences: ayKOhygF2 (5′-TGGTTGGCTTGTATGGAGCAGCAGAC-3′), ayKOhygR2 (5′-TAGGATATGCAGCACACAGGAGTGTGG-3′), AYKObsrF2 (5′-GGTAGTGCAGTTTCGAATGGACAAAAGG-3′), AYKObsrR2 (5′-TCAGGATTTGCAGCACACAGGAGTG-3′), PuroF2 (5′-GAGCTGCAAGAACTCTTCCTCACGC-3′), PuroR2 (5′-ATGTACCTCCCAGCTGAGACAGAGGG-3′), NeoF3 (5′-TTTGGTCCTGTAGTTTGCTAACACACCC-3′), NeoR3 (5′-GGATCAGTGCCTATCACTCCAGGTTG-3′). In addition, Southern hybridization using each of the drug-resistant genes as a probe was performed to confirm a single-site integration of each of the knockout cassettes.

RT-PCR analysis. RNA was extracted from spleens of fetuses or peripheral blood mononuclear cells from calves using an RNeasy mini kit (Qiagen) and first-strand cDNA synthesis was done using the superscript first strand synthesis system for RT-PCR (Invitrogen). PCR was done using primer pairs; mBCμF2 × mBCμR2 (**Supplementary Fig. 1c**), BL17 (located in the leader exon of bovine immunoglobulin heavy chain) × mBCμR2 and BL17 × bCγ1R2 in 40 cycles composed of 98 °C for 10 s, 62 °C for 30 s, 72 °C for 1 min. For detection of bovine *β-actin* mRNA expression, bBAF and bBAR primers were used in the same PCR condition. To exclude the possibility of genomic DNA contamination, another RT-PCR was performed without reverse transcriptase. The PCR products were run on 0.8% agarose gel. Primer sequences: mBCμF2 (5′-GCATGCTGACCATCACAGAG-3′), mBCμR2 (5′-GTTCAGGCCATCATAGGAGG-3′), BL17 (5′-CCCTCCTCTTTGTGCTGTCA-3′), bCγ1R2 (5′-GGGAGCTCAGGGGGTGGGCAACAGTCA-3′), bBAF (5′-ACATCCGCAAGGACCTCTAC-3′), bBAR (5′-AACCGACTGCTGTCACCTTC-3′).

Flow cytometry analysis. Peripheral blood was collected from 180-d-old fetuses or calves by jugular venipuncture into heparinized tubes. Ileum and cecum were also collected in AIMV cell culture medium (Invitrogen-GIBCO). Whole white blood cells (leukocytes) were isolated from heparinized blood using RBC-lysis buffer (Sigma). Lymphocytes from Ileal Peyer's patch were isolated by mechanical disruption and filtered using a 40 μm nylon cell strainer (BD Biosciences) before density-centrifugation using Ficoll-Paque PLUS (GE Healthcare Biosciences). Sheep anti-bovine IgM-biotin (Bethyl) and F(ab′)$_2$ goat anti-human IgM-biotin (Serotec) followed by streptavadin-PE-Cy5 (Caltag) were used to label surface IgM on the B cells. To label surface B220 marker on developing bovine B cells, we used mouse anti-bovine B220 (CD45R) antibody clone GS5A (VMRD) followed by anti-mouse IgG1-PE secondary antibody (Caltag). Mouse anti-bovine CD21 Clone MCA1424 (Serotec) directly labeled with PE was used to detect surface CD21 marker on bovine B cells. Staining was done by a standard protocol and then analyzed by FACScan or FACSAria flow cytometer (BD Biosciences).

Western blot. Immunoglobulin heavy and light chains were separated by SDS PAGE using 4–12% precast Bis-Tris gels (Invitrogen) and transferred to polyvinylidene difluoride membranes that were directly probed with specific horseradish peroxidase (HRP)-conjugated antibodies following blocking. The HRP-conjugated antibodies were: goat anti-bIgG (heavy and light; H+L) HRP (KPL) for bIgG heavy chain, goat anti-bIgG (Fab′)$_2$ HRP (Jackson ImmunoResearch) for bovine light chain, donkey anti-hIgG (H+L) HRP (Jackson ImmunoResearch) for hIgG heavy chain, and goat anti-hIgκ light chain HRP (Bethyl) for human κ-light chain. All HRP-conjugated antibodies for bovine and human IgGs were confirmed to have no species cross-reactivity.

ELISA. ELISA assays were sandwich type using an affinity-purified capture antibody and an appropriate HRP-enzyme–labeled detection antibody. For bIgM detection, sheep anti-bIgM affinity-purified (Bethyl) as a capture and sheep anti-bIgM-HRP as a detection antibody were used. For bIgG detection, sheep anti-bIgG affinity-purified as a capture and sheep anti-bIgG-HRP as a detection antibody were used. Detection was performed by a standard protocol.

hIgG was analyzed by using a commercial ELISA test (Bethyl). All assay steps were carried out as per manufacturer. Briefly, human reference serum (Standard) supplied in the kit was diluted to 500 ng/ml and then to 7.8 ng/ml in 1:2 serial dilutions (total of seven dilutions) in PBS/0.1% Tween 20 (PBS/Tween). Nunc Maxisopr Immuno plates were coated with affinity-purified goat anti-hIgG capture antibody at 10 μg/ml concentration, 100 μl/well at 25 °C for 1.5 h. Plates were washed three times with 200 μl of PBS/Tween buffer using a plate washer. Standards were loaded (500 ng/ml to 7.8 ng/ml) at 100 μl/well in duplicate wells. Four 1:2 serial dilutions of each serum samples were loaded in duplicates at 100 μl/well. Plates were covered and incubated at 25 °C for 1 h. Plates were then washed three times with PBS/Tween as described earlier. Sheep anti-hIgG HRP-conjugate antibody was diluted 1:100,000 in PBS/Tween and loaded at 100 μl/well for all wells. Plates were then incubated for 1 h at 25 °C and washed again three times with PBS/Tween. 1:1 mix of TMB/H_2O_2 substrate system (KPL) was added at 100 μl/well and color development was allowed for 20–25 min. Color reaction was stopped by adding 100 μl/well of 10% phosphoric acid Stop reagent and plates were read in a microplate reader at 450 nm. A standard curve (log to linear) was drawn with OD$_{450}$ reading on the *y* axis and log$_{10}$ concentrations (ng/ml) on *x* axis and average sample readings were interpolated in the graph to obtain ng/ml concentrations of each sample, using an automated Excel worksheet module. Final μg/ml concentration of hIgG was calculated by taking the mean of all dilutions of each sample in the linear portion of the curve.

OVA-immunization. *IGHM*$^{-/-}$, *IGHML1*$^{-/-}$ and *IGHM*$^{-/-}$*IGHML1*$^{-/-}$ calves and control wild-type calves were immunized with Ovalbumin (OVA) antigen (Sigma) at 1 mg/dose formulated with Montanide ISA 25 adjuvant (Seppic) as water-in-oil emulsion. The calves were immunized three times at 3-week intervals (primary immunization followed by first booster after 3 weeks and second booster after 6 weeks). Vaccine was administered by intramuscular injection (2 ml dose containing 1 mg/ml OVA plus 1 ml of ISA-25 adjuvant) in the neck region. Serum samples were collected before each immunization (V1, V2 and V3) and 7 d and 14 d after each immunization for antibody titer analysis. Blood was drawn into serum separator tubes, allowed to clot and serum was separated by centrifugation. Serum was then aliquoted in 0.5–1 ml volumes and stored frozen until assays were performed. Anti-OVA antibody titers were determined by OVA-specific IgG ELISA.

IBR-immunization. *IGHM*$^{-/-}$ calves and wild-type control calves were immunized with Triangle 4, which contained IBR antigen (Fort Dodge Animal Health). Vaccine was administered by subcutaneous injection in the neck region at 2 ml per dose. The animals were boosted four more times, with an interval of 3 weeks for each booster for the first to fourth vaccinations, and an interval of 6 weeks between the fourth and fifth vaccinations. Serum samples were taken right before each immunization (V1 to V4) and 7 d and 14 d after each immunization for antibody titer analysis. Blood was drawn into serum separator tubes (tiger-top), allowed to clot and serum was separated by centrifugation. Serum was then aliquoted in 0.5–1 ml volumes and stored frozen until assays were performed. Anti-IBR antibody titers were determined by IBR-specific IgG ELISA with a commercial bovine rhinotrachetis virus antibody test kit (IDEXX).

PA-immunization. κHAC/*IGHM*$^{-/-}$*IGHML1*$^{-/-}$ calves and κHAC/*IGHM*$^{-/-}$ control calves were immunized with anthrax recombinant protective antigen (rPA) antigen (List Biological) at 2 mg/dose formulated with Montanide ISA 206 adjuvant (Seppic) as a water-in-oil-in-water emulsion. The calves were immunized four times with 4-week intervals. Vaccine was administered by intramuscular injection (2 ml per dose containing 2 mg/ml PA plus 1 ml of ISA-206 adjuvant) in the neck region. Serum samples were collected before each immunization (V1 to V4) and 7 d, 10 d and 14 d after each immunization for antibody titer analysis. Blood was drawn into serum separator tubes, allowed to clot and serum was separated by centrifugation. Serum was then aliquoted in 0.5–1 ml volumes and stored frozen until assays were performed. Anti-PA antibody titers were determined by PA-specific IgG ELISA as follows.

To determine PA-specific hIgG titers, 96-well Immuno 2-HB ELISA plates were coated by adding 100 µl per well of 2 µg/ml of rPA (List Biological) in PBS at pH 7.4 and incubating overnight (12–16 h) at 4 °C. rPA-coated plates were then washed three times with 200 µl of PBS/0.05% Tween 20. Serum samples were diluted in PBS/0.05% Tween 20 buffer with 5% membrane blocking agent (non-fat dry milk) in four serial dilutions. High-titer purified hIgG from calf 468 with a predetermined end-point titer was used as the standard and seven 1:3 serial dilutions from 1:9,000 to 1:6,561,000 were prepared in PBS/0.05% Tween 20 buffer for the standard curve. Reciprocal of the end-point dilution was used as titer units, and for the standard, the end-point titer was determined and assigned as 7,400,000 units. A positive-control serum with predetermined titer (900,000 units) and a negative-control serum with no titer were also diluted serially in PBS/0.05% Tween 20 buffer with non-fat dry milk and were used as internal controls to monitor consistency of the assays. The calibrator standard serum dilutions, positive-control serum, negative-control serum and test serum samples were added in duplicate wells at 100 µl/well in rPA-coated plates and incubated for 1 h at 37 °C. Plates were washed three times with PBS/0.05% Tween 20 buffer to remove unbound proteins and 100 µl of donkey anti-hIgG-HRP–labeled antibody (Jackson Immuno Research) diluted at 1:50,000 in PBS/T buffer with non-fat dry milk added to each well. Plates were incubated for 1 h at 37 °C and washed three times with in PBS/0.05% Tween 20. Finally, the bound anti-PA antibodies were detected by adding 100 µl/well TMB +H_2O_2 substrate mix (KPL) and incubated for 10 min at 25 °C. The reaction was stopped by adding 100 µl 10% phosphoric acid and read in Microplate Reader (Biotek Instruments) at 450 nm. A four-parameter standard curve was generated using seven serial dilution values and serum sample values were calculated by interpolation on the curve by KC-4 software. Average titer values from three or four test dilutions were calculated for each test serum sample. Similarly, PA-specific bIgG titers were determined.

TNA assay. The TNA assay was performed as described previously[37] with some modifications. In brief, cells were plated in a 96-well assay plate and allowed to adhere overnight in a 37 °C, 5% CO_2 incubator so that they would reach a density of 40–60% confluency the following morning. Sera from calves that had been vaccinated with rPA as described above was prepared in a twofold sequential dilution and distributed into a separate 96-well plate. A fixed dose of lethal toxin (a mixture of rPA and rLF) was added to each of the serum dilutions and the mixtures were incubated for 1 h in a 37 °C, 5% CO_2 incubator. The lethal toxin/serum mixtures were then added to the cells in the individual wells of the 96-well plate and incubated for 4 h. This 4-h incubation provides the time for any remaining active lethal toxin to lyse the cells. Cells were washed, stained with thiazol blue (MTT; Sigma) and incubated for 1 h at 37 °C. To determine the cell viability, we plotted OD_{570} readings (with background subtracted out) against the dilutions of the serum samples. This analysis allows for the calculation of either an end-point titer or an effective-dose 50% (ED_{50}), which is the dilution of sera in which one-half of the lethal toxin is neutralized.

Mouse protection assay. Groups of ten female A/J mice (Jackson Laboratories) at ~7 weeks of age were challenged with the Sterne strain of anthrax spores (Colorado Serum). Spores were administered at a dose of 1×10^6 spores by intraperitoneal (IP) injection. Spores had been prepared by washing three times in sterile water to remove the saponin that is present in the commercial preparation. Washed spores were stored in sterile water and the titer of spores was determined on nutrient agar plates. Spores were diluted with sterile water so that the appropriate dose per mouse was in a 200 µl volume. Mice were treated with purified IgG preparations at 4 h after challenge. Total purified hIgG from calf 468 contained both fully human and chimeric hIgG molecules. Mice received total hIgG, fully human hIgG, chimeric hIgG, pre-immune total hIgG (negative control) or a pooled bIgG positive control. All antibody treatments were administered by IP injection in a 200 µl volume. Mice were observed twice daily for 28 d and moribund animals were euthanized.

Purification of fully human hIgG/hκ-chain and chimeric hIgG fractions. Plasma bags were thawed at 25 °C overnight and total protein concentration was determined. One volume of purified water was added to the plasma, followed by adjusting to pH 4.8 with 20% acetic acid. Caprylic acid was slowly added to the sample (with continuous mixing) to a final concentration of 6.0%. The sample was mixed for 30 min and filtered using a depth filter device. The filtrate was adjusted to neutral pH and loaded onto an anti-hIgG Fc affinity column (6CP Sepharose) equilibrated with PBS. The column was eluted with pH 3 solution to recover IgG. The 6CP column elution peak was neutralized and then passed through an anti-bIgG Fc column (HC15 Sepharose) to remove residual bIgG. To separate fully human hIgG from chimeric hIgG, the IgG sample was applied onto an anti-human $F(ab')_2$ κ Sepharose column. The flow-through fraction contained chimeric hIgG, whereas the pH 3.0 eluted peak was fully human hIgG. Samples were then dialyzed into PBS and stored at 2–8 °C.

Glycosylation analysis. *N*-linked oligosaccharide profiling was done as follows. A sample of antibody (0.5 mg) was diluted with water (total 49 µl) in a sample tube (1.5 ml). 2-mercaptoethanol (1 µl) and PNGase F (10 units, 10 µl) were added to the mixture and incubated at 37 °C for 20–24 h. After addition of ethanol (150 µl), the mixture was centrifuged at 15,000*g* for 15 min. The supernatant containing the released oligosaccharides was transferred to a new sample tube and evaporated to dryness. *N*-linked oligosaccharides in the mixture were labeled with 2-aminobenzoic acid (2-AA) according to the method reported previously[39]. Briefly, water (20 µl) was added to the dried oligosaccharide sample. A derivatization reagent was freshly prepared by dissolution of 2-AA and sodium cyanoborohydride (30 mg and 20 mg, respectively) in methanol (1 ml) containing 4% sodium acetate and 2% boric acid. This reagent (100 µl) was then added to the oligosaccharide solution. The mixture was kept at 80 °C for 1 h. After cooling followed by addition of water (30 µl), the oligosaccharide mixture was purified using a solid-phase extraction column (Oasis HLB cartridges, 1 ml, Waters). The reaction solution was diluted with 1.0 ml of acetonitrile-water (95:5), mixed vigorously and applied to a cartridge previously equilibrated with the same solvent (1 ml × 2). After washing the cartridge with acetonitrile-water (95:5, 1 ml × 2), the fluorescence-labeled oligosaccharides were eluted with acetonitrile-water (20:80, 1 ml) and the eluate was evaporated to dryness by a centrifugal evaporator. The residue was dissolved in water (100 µl), and a portion (typically 5 µl) was used for the analysis by CE-LIF. Capillary electrophoresis was performed on a ProteomeLab PA800 system (Beckman Coulter) equipped with a helium-cadmium laser–induced fluorescence detector (excitation 325 nm, emission 405 nm) using a DB-1 capillary (100 µm internal diameter, 30 cm effective length, 40 cm total length, Agilent/J&D Scientific) in 100 mM Tris-borate buffer (pH 8.3) containing 10% PEG35000 as the running buffer. PEG was added to diminish electroendoosmotic flow and improve the resolution. For pressure injection, sample solutions were introduced into the capillary at 1 p.s.i. for 10 s. Separation was performed by applying 25 kV at 25 °C at reverse polarity.

Sialic acid content analysis was carried out as follows. A sample of antibody (0.4 mg) was diluted with water (total 100 µl) in a sample tube (1.5 ml). Hydrolysis solution (water/acetic acid; 27:8; 100 µl) was added to the sample, and incubated at 80 °C for 2.5 h. Then 1,2-diamino-4,5-methylenedioxybenzene solution (200 µl) was added and the mixture was kept at 60 °C for 2 h in the dark. After cooling, 1 M NaOH (200 µl) was added to stop the reaction. The derivatized sialic acids were separated by reversed phase HPLC using a C18 column (9 × 150 mm, Symmetry, Waters) and mobile phase (water/acetonitrile/methanol; 84:9:7) at 0.6 ml/min. Detection was performed using fluorescence detector (excitation 373 nm, emission 448 nm). Sialic acid content was calculated from a standard curve generated from known concentrations of NANA and NGNA derivatized in a same manner as the sample.

Embryonic cloning. Cloned fetuses and calves were produced using chromatin transfer procedure as described previously[6,23]. Both $IGHM^{-/-}IGHML1^{-/-}$ and κHAC/$IGHM^{-/-}IGHML1^{-/-}$ calves were maintained with ∼7 mg/ml of exogenous bIgG supplied as bovine intravenous immunoglobulin from wild-type cattle donors.

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

ACKNOWLEDGMENTS
We thank the teams at Transova Genetics for their efforts in embryo transfer and at the Research Development Center for their assistance in fetal and sample collection, calf delivery and care. We thank the teams at Hematech Inc. for their assistance in gene targeting, cell culture, embryonic cloning and any other technical procedures. We also thank Ralph Kubo, Sandra Rickert, Mareto Hosono and Tomoyuki Tahara for their useful comments on the immunological study.

AUTHOR CONTRIBUTIONS

Y.K. and J.M.R. led the work and wrote the manuscript. P.K. and J.K. led animal cloning. T.S. and H.W. led immunological analyses and immunization. J.J. led purification and protein chemistry. H.M. and J.S. carried out gene targeting experiments. J.M. conducted the mouse challenge assay. M.H. performed flow cytometry analysis. S.K. and K.T. implemented sugar chain analyses. I.I. initiated the work.

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. Lemieux, R., Bazin, R. & Neron, S. Therapeutic intravenous immunoglobulins. *Mol. Immunol.* **42**, 839–848 (2005).
2. Jolles, S., Sewell, W.A.C. & Misbah, S.A. Clinical uses of intravenous immunoglobulin. *Clin. Exp. Immunol.* **142**, 1–11 (2005).
3. Newcombe, C. & Newcombe, A.R. Antibody production: polyclonal-derived biotherapeutics. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **848**, 2–7 (2007).
4. Lonberg, N. Human antibodies from transgenic animals. *Nat. Biotechnol.* **23**, 1117–1125 (2005).
5. Waltz, E. Polyclonal antibodies step out of the shadows. *Nat. Biotechnol.* **24**, 1181 (2006).
6. Kuroiwa, Y. *et al.* Cloned transchromosomic calves producing human immunoglobulin. *Nat. Biotechnol.* **20**, 889–894 (2002).
7. Kitamura, D., Roes, J., Kuhn, R. & Rajewsky, K.A. B cell-deficient mouse by targeted disruption of the membrane exon of the immunoglobulin µ chain gene. *Nature* **350**, 423–426 (1991).
8. Tomizuka, K. *et al.* Double trans-chromosomic mice: Maintenance of two individual human chromosome fragments containing Ig heavy and κ loci and expression of fully human antibodies. *Proc. Natl. Acad. Sci. USA* **97**, 722–727 (2000).
9. Yel, L. *et al.* Mutations in the mu heavy chain gene in patients with agammaglobulinemia. *N. Engl. J. Med.* **335**, 1486–1493 (1996).
10. Hayes, H.C. & Petit, J.E. Mapping of the β-lactoglobulin gene and of immunoglobulin M heavy chain-like sequence to homologous cattle, sheep and goat chromosomes. *Mamm. Genome* **4**, 207–210 (1993).
11. Hosseini, A., Campbell, G., Prorocic, M. & Aitken, R. Duplicated copies of the bovine J_H locus contribute to the Ig repertoire. *Int. Immunol.* **16**, 843–852 (2004).
12. Chowdhary, B.P., Fronicke, L., Gustavsson, I. & Scherthan, H. Comparative analysis of the cattle and human genomes: detection of ZOO-FISH and gene mapping-based chromosomal homologies. *Mamm. Genome* **7**, 297–302 (1996).
13. Tobin-Janzen, T.C. & Womack, J.E. Comparative mapping of *IGHG1, IGHM, FES* and *FOS* in domestic cattle. *Immunogenetics* **36**, 157–165 (1992).
14. Gu, F., Chowdhary, B.P., Andersson, L., Harbitz, I. & Gustavsson, I. Assignment of the bovine immunoglobulin gamma heavy chain (IGHG) gene to chromosome 21q24 by in situ hybridization. *Hereditas* **117**, 237–240 (1992).
15. Zhao, Y., Kacskovics, I., Rabbani, H. & Hammarstrom, L. Physical mapping of the bovine immunoglobulin heavy chain constant region gene locus. *J. Biol. Chem.* **278**, 35024–35032 (2003).
16. Reynaud, C.A., Mackay, C.R., Muller, R.G. & Weill, J.-C. Somaticgeneration of diversity in a mammalian primary lymphoid organ: the sheep ileal Peyer's patches. *Cell* **64**, 995–1005 (1991).
17. Jenne, C.N., Kennedy, L.J., McCullagh, P. & Reynolds, J.D. A new model of sheep Ig diversification: shifting the emphasis toward combinatorial mechanisms and away from hypermutation. *J. Immunol.* **170**, 3739–3750 (2003).
18. Butler, J.E. Immunoglobulin diversity, B-cell and antibody repertoire development in large farm animals. *Rev. Sci. Tech.* **17**, 43–70 (1998).
19. Butler, J.E. Immunological diversity, B-cell and antibody repertoire development in large farm animals. *Rev. Sci. Tech.* **17**, 43–70 (1998).
20. Meyer, A. *et al.* Immunoglobulin gene diversification in cattle. *Int. Rev. Immunol.* **15**, 165–183 (1997).
21. Aitken, R. *et al.* Structure and diversification of the bovine immunoglobulin repertoire. *Vet. Immunol. Immunopathol.* **72**, 21–29 (1999).
22. Lucier, M.R. *et al.* Multiple sites of Vλ diversification in cattle. *J. Immunol.* **161**, 5438–5444 (1998).
23. Kuroiwa, Y. *et al.* Sequential targeting of the genes encoding immunoglobulin-µ and prion protein in cattle. *Nat. Genet.* **36**, 775–780 (2004).
24. Alt, F.W., Blackwell, T.K. & Yancopoulos, G.D. Development of the primary antibody repertoire. *Science* **238**, 1079–1087 (1987).
25. Hood, L., Gray, W.R., Sanders, B.G. & Dreyer, W.Y. Light chain evolution: antibodies. *Cold Spring Harbor Symp. Quan. Bio.* **32**, 133–146 (1967).
26. Kuroiwa, Y. *et al.* Manipulation of human minichromosomes to carry greater than megabase-sized chromosome inserts. *Nat. Biotechnol.* **18**, 1086–1090 (2000).
27. Brey, R.N. Molecular basis for improved anthrax vaccines. *Adv. Drug Deliv. Rev.* **57**, 1266–1292 (2005).
28. Raju, T.S., Briggs, B.J., Borge, M.S. & Jones, J.S.A. Species-specific variation in glycosylation of IgG: evidence for species-specific sialylation and branch-specific galactosylation and importance for engineering recombinant glycoprotein therapeutics. *Glycobiology* **10**, 477–486 (2000).
29. Hering, D., Thompson, W., Hewetson, J., Little, S., Norris, S. & Pace-Templeton, J. Validation of the anthrax lethal toxin neutralization assay. *Biologicals* **32**, 17–27 (2004).
30. Pittman, P.R., Parker, & Friedlander, A.M. Anthrax vaccine: immunogenicity and safety of a dose-reduction, route-change comparison study in humans. *Vaccine* **20**, 1412–1420 (2002).
31. Semenova, V.A. *et al.* Mass value assignment of total and subclass immunoglobulin G in a human standard anthrax reference serum. *Clin. Diagn. Lab. Immunol.* **11**, 919–923 (2004).
32. Wakayama, T. *et al.* Cloning of mice to six generations. *Nature* **407**, 318–319 (2000).
33. Kubota, C., Tian, X.C. & Yang, X. Serial bull cloning by somatic cell nuclear transfer. *Nat. Biotechnol.* **22**, 693–694 (2004).
34. Mann, M.R. *et al.* Disruption of imprinted gene methylation and expression in cloned preimplantation stage mouse embryos. *Biol. Reprod.* **69**, 902–914 (2003).
35. Rideout, W.M., Eggan, K. & Jaenisch, R. Nuclear cloning and epigenetic reprogramming of the genome. *Science* **293**, 1093–1098 (2001).
36. Tamashiro, K.L. *et al.* Cloned mice have an obese phenotype not transmitted to their offspring. *Nat. Med.* **8**, 262–267 (2002).
37. Knight, K.L., Kingzette, M., Crane, M.A. & Zhai, S.K. Transchromosomally derived Ig heavy chains. *J. Immunol.* **155**, 684–691 (1995).
38. Nolan-Willard, M., Berton, M.T. & Tucker, P. Coexpression of mu and gamma 1 heavy chains can occur by a discontinuous transcription mechanism from the same unrearranged chromosome. *Proc. Natl. Acad. Sci. USA* **89**, 1234–1238 (1992).
39. Kamoda, S., Ishikawa, R. & Kakehi, K. Capillary electrophoresis with laser-induced fluorescence detection for detailed studies on *N*-linked oligosaccharide profile of therapeutic recombinant monoclonal antibodies. *J. Chromatogr. A.* **1133**, 332–339 (2006).

nature biotechnology

Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing

Andreas Gnirke[1], Alexandre Melnikov[1], Jared Maguire[1], Peter Rogov[1], Emily M LeProust[2], William Brockman[1,5], Timothy Fennell[1], Georgia Giannoukos[1], Sheila Fisher[1], Carsten Russ[1], Stacey Gabriel[1], David B Jaffe[1], Eric S Lander[1,3,4] & Chad Nusbaum[1]

Targeting genomic loci by massively parallel sequencing requires new methods to enrich templates to be sequenced. We developed a capture method that uses biotinylated RNA 'baits' to fish targets out of a 'pond' of DNA fragments. The RNA is transcribed from PCR-amplified oligodeoxynucleotides originally synthesized on a microarray, generating sufficient bait for multiple captures at concentrations high enough to drive the hybridization. We tested this method with 170-mer baits that target $>$15,000 coding exons (2.5 Mb) and four regions (1.7 Mb total) using Illumina sequencing as read-out. About 90% of uniquely aligning bases fell on or near bait sequence; up to 50% lay on exons proper. The uniformity was such that $\sim$60% of target bases in the exonic 'catch', and $\sim$80% in the regional catch, had at least half the mean coverage. One lane of Illumina sequence was sufficient to call high-confidence genotypes for 89% of the targeted exon space.

The development and commercialization of a new generation of increasingly powerful sequencing methodologies and instruments[1–4] have lowered the cost per nucleotide of sequencing data by several orders of magnitude. Within a short time, several individual human genomes have been sequenced on next-generation instruments[3,5–7], with plans and funding in place to sequence more (http://www.1000genomes.org/).

Sequencing entire human genomes will be an important application of next-generation sequencing. However, many research and diagnostic goals may be achieved by sequencing a specific subset of the genome in large numbers of individual samples. For example, there may be substantial economy in targeting the protein-coding fraction, the 'exome', which represents only $\sim$1% of the human genome. The economy is even greater for many key resequencing targets, such as genomic regions implicated by whole-genome association scans and the exons of sets of protein-coding genes implicated in specific diseases. Efficient and cost-effective targeting of a specific fraction of the genome could substantially lower the sequencing costs of a project, independent of the sequencing technology used.

Sequencing targeted regions on massively parallel sequencing instruments requires developing methods for massively parallel enrichment of the templates to be sequenced. Recognizing the inadequacy of traditional singleplex or multiplex PCR for this purpose, several groups have developed 'genome-partitioning' methods for preparing complex mixtures of sequencing templates that are highly enriched for targets of interest[8–15]. Only two of these methods have been tested on target sets complex enough to match the scale of current next-generation sequencing instruments.

The first method, microarray capture[9,12,13], uses hybridization to arrays containing synthetic oligonucleotides that match the target sequence to capture templates from randomly sheared, adaptor-ligated genomic DNA; it has been applied to $>$200,000 coding exons[12]. Array capture works best for genomic DNA fragments that are $\sim$500 bases long[12], thereby limiting the enrichment and sequencing efficiency for very short dispersed targets, such as human protein-coding exons that have a median size of 120 bp[16].

The second method, multiplex amplification[14], uses oligonucleotides that are synthesized on a microarray, subsequently cleaved off and amplified by PCR, to perform a padlock and molecular-inversion reaction[17,18] in solution where the probes are extended and circularized to copy, rather than directly capture, the targets. Uncoupling the synthesis and reaction formats in this manner is advantageous because it allows reusing and quality testing of a single lot of oligonucleotide probes. However, the padlock reaction is not nearly as well understood as a simple hybridization and has not been properly optimized for this purpose. As published[14], multiplex amplification missed $>$80% of the targeted exons in any single reaction and showed highly uneven representation of sequencing targets, poor reproducibility between technical replicates, and uneven recovery of alleles. A more recent nonsequencing-based study using a similar approach suggests that the uniformity, reproducibility and efficiency of multiplex amplification can be improved[15].

[1]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [2]Agilent Technologies Inc., 5301 Stevens Creek Blvd., Santa Clara, California 95051, USA. [3]Department of Biology, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, Massachusetts 02139, USA. [4]Department of Systems Biology, Harvard Medical School, 200 Longwood Ave., Boston, Massachusetts 02115, USA. [5]Present address: Google, Inc., 5 Cambridge Center, Cambridge, Massachusetts 02142, USA. Correspondence should be addressed to A.G. (gnirke@broad.mit.edu).

Received 10 November 2008; accepted 5 January 2009; published online 1 February 2009; doi:10.1038/nbt.1523

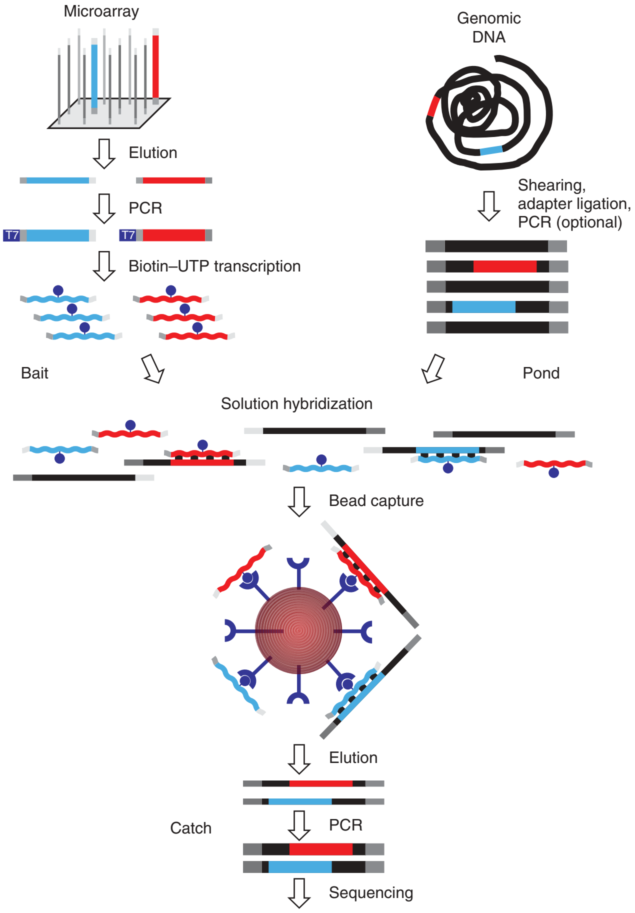


Figure 1 Overview of hybrid selection method. Illustrated are steps involved in the preparation of a complex pool of biotinylated RNA capture probes (bait; top left), whole-genome fragment input library (pond; top right) and hybrid-selected enriched output library (catch; bottom). Two sequencing targets and their respective baits are shown in red and blue. Universal adaptor sequences are gray. The excess of single-stranded nonself-complementary RNA (wavy lines) drives the hybridization.

Here we describe a method that overcomes some of the weaknesses of previous methods. It combines the simplicity and robust performance of oligonucleotide hybridization with the advantages of amplifying array-synthesized oligonucleotides and performing the selection reaction in solution.

RESULTS

Hybrid selection method

We developed a method for capturing sequencing targets that combines the flexibility and economy of oligonucleotide synthesis on a microarray with the favorable kinetics of hybridization in solution (**Fig. 1**). A complex pool of ultra-long 200-mer oligonucleotides is synthesized in parallel on an Agilent microarray and then cleaved from the array. Each oligonucleotide consists of a target-specific 170-mer sequence flanked by 15 bases of a universal primer sequence on each side to allow PCR amplification. After the initial PCR, a T7 promoter is added in a second round of PCR. We then use *in vitro* transcription in the presence of biotin-UTP to generate a single-stranded RNA hybridization bait for fishing targets of interest out of a 'pond' of randomly sheared, adaptor-ligated and PCR-amplified total human DNA. The hybridization is driven by the vast excess of RNA baits that cannot self-anneal. The 'catch' is pulled down with streptavidin-coated magnetic beads, PCR amplified with universal primers and analyzed on a next-generation sequencing instrument. The method allows preparation of large amounts of bait from a single oligonucleotide array synthesis that can be tested for quality, stored in aliquots and used repeatedly over the course of a large-scale targeted sequencing project.

Capturing and sequencing exon targets

For a pilot study, we used a set of 1,900 human genes randomly chosen to ensure unbiased sampling regardless of length, repeat content or base composition. We designed 22,000 bait sequences of 170 bases in length, targeting all 15,565 protein-coding exons of these genes. The baits were tiled without overlap or gaps such that the entire coding sequence was covered. This simple design minimizes the number of synthetic oligonucleotides required; for 75% of all coding exons in the human genome, a single oligonucleotide would be sufficient. As the median size of protein-coding exons is only 120 bp[16], many baits extend beyond their target exon. Our test baits for catching exons constituted 3.7 Mb, and the targeted exons comprised 2.5 Mb (67%).

Our pond consisted of genomic DNA, derived from a human cell line (Coriell NA15510), that had been randomly sheared, ligated to standard Illumina sequencing adapters, selected to include lengths of 200–350 bp (mean insert size ~250 bp) and PCR amplified for 12 cycles. We hybridized 500 ng of this whole-genome fragment library with 500 ng biotinylated RNA bait, PCR amplified the hybrid-selected DNA and generated 36-base sequencing reads off the Illumina adaptor sequence at the ends of each fragment. We obtained 85 Mb of sequence that aligned uniquely to the human genome; 76 Mb was on or within 500 bp of a bait.

Of the specifically captured 76 Mb of sequence, 49 Mb (65%) lay directly on a bait. The proportion of this sequence directly within the exons (36 Mb total) closely matched the proportion of exonic sequence within the bait. Overall, 58% and 42% of the 85 Mb uniquely aligning human sequence mapped to baits and exons, respectively.

The high stringency of hybridization selects for fragments that contain a substantial portion of the bait sequence. As a result, fragments for which both ends map near to or outside of the ends of the bait sequence are overrepresented relative to fragments that overlap less (that is, fragments that end near the middle of a bait). Merely end-sequencing the fragments with short 36-base reads therefore leads to elevated coverage near the end of the baits, with many reads falling outside the target, and a pronounced dip in coverage in the center. This effect is evident in the cumulative coverage profile representing 7,052 free-standing single-bait targets (**Fig. 2a**).

To improve coverage in the middle, we replaced end sequencing of the catch with shotgun sequencing of the catch. Specifically, we changed the Illumina adaptor on the whole-genome fragment library to a generic adaptor, independent of a sequencing method, and amplified the catch with PCR primers carrying a *Not*I site at their 5′ ends. *Not*I-digestion of the PCR product generates sticky ends and facilitates concatenation by co-ligation for subsequent reshearing and shotgun sequencing of the hybrid-selected DNA. This modification to the protocol shifted the coverage to the middle (**Fig. 2b**). About 90 of 102 Mb of unique human sequence (88%) aligned within 500 bases of a bait. The proportion of bait sequence in the specific catch (90 Mb) rose from 65% to 77% (69 Mb; 51 Mb thereof on exon). The fraction of bait and exon sequence in the uniquely aligning human Illumina sequence was 67% and 50%, respectively.

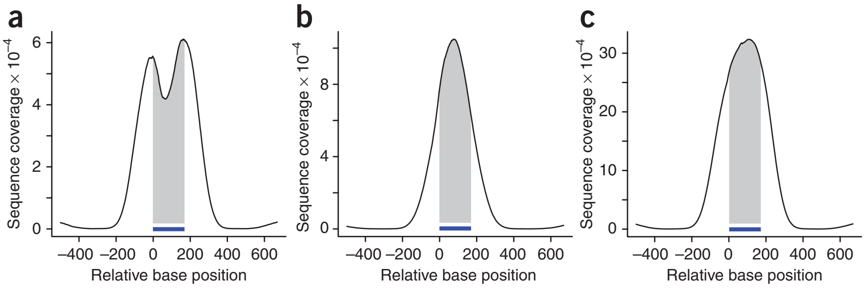


Figure 2 Coverage profiles of exon targets by end sequencing and shotgun sequencing. Shown are cumulative coverage profiles that sum the per-base sequencing coverage along 7,052 single-bait target exons. Only free-standing baits that were not within 500 bases of another one were included in this analysis. (**a**) End sequencing with 36-base reads produced a bimodal profile with high sequence coverage near and slightly beyond the ends of the 170-base baits (indicated by the horizontal bar). (**b**) Shotgun sequencing of a capture from a different pond library (containing fragments with generic rather than Illumina-specific adapters) with 36-base reads after concatenating and reshearing gave more coverage on bait (shaded area) than near bait. (**c**) Resequencing of the first capture with 76-base end reads had a similar effect, although the peak was slightly wider and the on-bait fraction of the peak area slightly smaller. Note that the scale on the *y*-axis and hence the absolute peak height is different in each case. The different scales reflect the different numbers of sequenced bases, which are much lower for GA-I lanes (**a**,**b**) than for a GA-II lane (**c**).

Although shearing the catch improved the proportion of bait sequence, the process adds an additional round of library construction with associated costs, amplification steps and potential biases. It also generates reads containing uninformative adaptor sequence as a by-product. During the course of these experiments, it became possible to increase the sequence read-length on the Illumina platform. We reasoned that simply increasing the read-length would also increase coverage in the middle and thus obviate the need for shotgun-library construction. Indeed, we performed end sequencing of the very same catch that had produced the bimodal coverage profile shown in **Figure 2a**, this time running 76-base instead of 36-base reads on one lane of an Illumina GA-II instrument. The longer reads resulted in a unimodal, center-weighted cumulative coverage profile (**Fig. 2c**). This lane generated 492 Mb of sequence that aligned uniquely to the genome, of which 445 Mb were on or near a bait. Of the specifically captured sequence, 321 Mb (72%) was directly on the bait itself and 235 Mb (53%) was contained within the exons. About 65% of the unique human sequence was on bait; 48% was on exons proper. The average coverage of bases was 86-fold within baits and 94-fold within coding exons.

Specificity

The percentage of the uniquely aligning human sequence that falls on or near a bait (e.g., 445/492 = 90% for the 76-base end reads) provides an upper bound for estimating the specificity of hybrid selection. In this experiment, 358 Mb (42%) of the 851 Mb of raw sequence did not align uniquely to the human genome (**Table 1**) and was not considered. By comparison, typically ∼55% of raw bases in whole-genome-sequencing lanes do not align uniquely. The raw bases likely contain hybrid-selected human sequence that is not unique. The lower bound, assuming that all discarded sequence represented repetitive human background sequence rather than low-quality reads, was 445/851 = 52%. To obtain a more precise number, we aligned the raw reads again to the human genome, this time allowing multiple placements, and determined the fraction of all human alignable sequence that lay on or within 500 bp of a bait. Based on this calculation, our best estimate for the specificity of this catch was 82%.

Of note, the specifically captured sequence included near-target hits that were not on exons proper. The percentage of uniquely aligning Illumina sequence that actually lay on coding sequence, that is, the upper bound of the overall specificity of targeted exon sequencing, was 48% in this experiment. **Table 1** shows a detailed breakdown of raw and uniquely aligned Illumina sequences and measures of specificity for the three targeted exon-sequencing experiments.

Table 1 Detailed breakdown of Illumina sequences generated from exon catches

Length and kind of Illumina sequencing reads	36-base GA-I end sequences	36-base GA-I shotgun sequences	76-base GA-II end sequences
Aggregate length of target[a]	2.5 Mb	2.5 Mb	2.5 Mb
Aggregate length of baits	3.7 Mb	3.7 Mb	3.7 Mb
Total raw unfiltered sequence	152 Mb	219 Mb[b]	851 Mb
Raw sequence not aligned uniquely to genome[c]	67 Mb	116 Mb	358 Mb
Uniquely aligned human sequence	85 Mb	102 Mb	492 Mb
Uniquely aligned sequence on target	36 Mb	51 Mb	235 Mb
Uniquely aligned sequence near target[d]	40 Mb	38 Mb	210 Mb
Uniquely aligned sequence on or near target	76 Mb	90 Mb	445 Mb
Fraction of uniquely aligned sequence on or near target[e]	89%	88%	90%
Fraction of raw bases uniquely aligned on or near target[f]	50%	41%[g]	52%
Fraction of uniquely aligned bases on target[h]	42%	50%	48%

[a] Protein-coding exon sequence only. [b] Each unit of concatenated catch contains 44-46 bases (∼18%) of generic adaptor sequence. Therefore, ∼18% (39 Mb) of the 219 Mb is not of human origin. [c] All raw sequence that fails to align uniquely to the human reference genome including low-quality sequence. [d] Outside but within 500 bp of a target exon. [e] Upper bound for estimating the specificity of hybrid selection. [f] Lower bound for estimating the specificity of hybrid selection. [g] The denominator (219 Mb) includes ∼39 Mb of sequence from the generic adapters. Excluding these 39 Mb, the lower bound for the estimated specificity of this catch is 90/180 = 50%. [h] Upper bound for the overall specificity of targeted exon sequencing.

Regional capture and sequencing

Next, we designed and tested a pool of 170-mer baits for targeted sequencing of four genomic regions ranging from 0.22 to 0.75 Mb in size (**Supplementary Table 1** online). The combined span of the regions was 1.68 Mb. The target regions included a large portion of ENCODE region ENr113 as well as the genes *IGF2BP2*, *CDKN2A*, *CDKN2B* and *CDKAL1*. For a pilot experiment, we designed nonoverlapping 170-mers that largely excluded repeated sequences (allowing no more than 40 bases of repetitive sequence in each). The baits totaled 0.75 Mb in length, whereas the remaining 0.93 Mb was not covered owing to repetitive sequence content. We fished in a pond containing 350- to 500-bp fragments of human genomic DNA (Coriell NA15510). The catch was analyzed with the

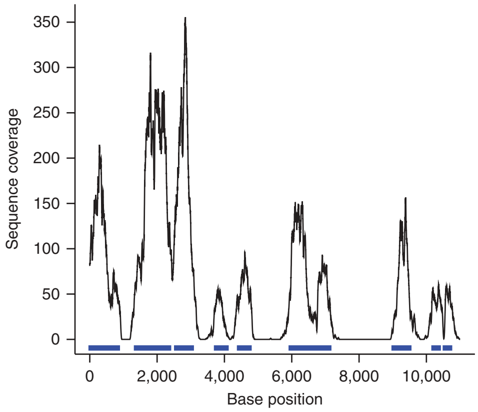


Figure 3 Sequence coverage along a contiguous target. Shown is base-by-base sequence coverage along a typical 11-kb segment (chr4:118635000–118646000) out of 1.7 Mb. Sequence corresponding to bait is marked in blue. Segments that had more than 40 repeat-masked bases per 170-base window were not targeted by baits and received little or no coverage with sequencing reads aligning uniquely to the genome.

shotgun sequencing approach above, with 36-base reads. The experiment preceded the development of the 76-base reads.

We generated one lane of Illumina GA-I sequence, yielding 191 Mb that aligned uniquely to the human reference sequence. Of this sequence, 179 Mb (94%) fell within the four targeted genome segments. About 164 Mb was on bait whereas 15 Mb aligned uniquely within the 0.95 Mb that was not covered by baits. Essentially all unique sequence within the bait-free zones was within 500 bp of a bait sequence, suggesting that it had been caught by specific hybridization to a bait. A typical coverage profile along 11 kb is shown in **Figure 3**. As expected, the coverage was not uniform and had peaks at unique segments that were represented in the bait pool and deep valleys or holes at mostly repetitive regions outside the baits. The average depth of coverage for the 0.75 million genome bases covered by bait in the four target regions was 221.

Evenness of coverage

Uniformity of capture, along with specificity, is the main determinant for the efficiency and practical utility of any bulk enrichment method for targeted sequencing. The larger the differences in relative abundance, the deeper one has to sequence to cover the underrepresented targets. We sought to display the data in a form that is independent of the absolute quantity of sequence (**Fig. 4**). Specifically, we normalized the coverage of each base to the mean coverage observed across the entire set of targets. This allows comparison of results from experiments with widely differing sequence yields, different template preparation methods or different sequencing instruments.

The two graphs in **Figure 4** show the fraction of bases contained within a bait at or above a given normalized coverage level; the normalized coverage was obtained by dividing the observed coverage by the mean coverage, which was 18 for the shotgun-sequenced exon capture (**Fig. 4a**) and 221 for the regional capture (**Fig. 4b**).

In the exon-capture experiment, >60% of the bases within baits received at least half the mean coverage, and almost 80% at least one-fifth. Twelve percent had no coverage in this particular sequencing lane. The normalized coverage-distribution plot for targeted regional sequencing is considerably flatter, indicating even better capture uniformity: 80% of the bases within baits received at least half the mean coverage; 86% received at least one-fifth; 5% was not covered in this experiment.

We attribute the differences in performance mainly to the fact that exon targets are generally short and isolated and often targeted by a single capture oligonucleotide (with few additional ones to choose from without widening the segment covered by bait). In contrast, the regional capture benefits from synergistic effects between adjacent baits, that is, an overhanging genome fragment caught by one bait contributing to the coverage underneath neighboring ones. The slightly longer DNA fragments used in this experiment (350–500 bases compared to 200–350 bases for exon capture) may have contributed to this effect. Additional coverage-distribution data, including graphs that were truncated at a normalized coverage of 5 instead of 1 to show the tail of the distribution, are available in **Supplementary Figures 1 and 2** online.

Effects of base composition

Separating the exon-capture baits into five categories based on their GC content revealed a systematic difference in coverage—with targets having GC content in the range of 50–60% receiving the highest coverage and those with very high (70–80%) or very low (30–40%) GC content getting the least coverage (**Supplementary Fig. 3** online). The effects of base composition most likely reflect genuine systematic differences in hybridization behavior. However, it is also conceivable that GC bias at other steps in the process contribute to this effect. For example, we know from microarray assays that PCR can deplete oligonucleotide sequences with extreme base compositions up to about fivefold (data not shown). In addition, bias at the oligonucleotide-synthesis step may play a role. PCR amplification of the catch and sequencing itself is also known to introduce bias[19,20].

Reproducibility

To assess the reproducibility of targeted exon sequencing, we compared the results from independent technical replicates. Specifically, we performed two separate hybrid selections with ~250-bp fragments prepared from the same source DNA (Coriell NA15510) and generated one lane of Illumina shotgun sequence each. The ratio of the mean normalized sequence coverage for individual exons in the two experiments was distributed closely around 1, indicating much less experiment-to-experiment than target-to-target variability (**Fig. 5a**). Base-by-base coverage profiles for individual exons were remarkably

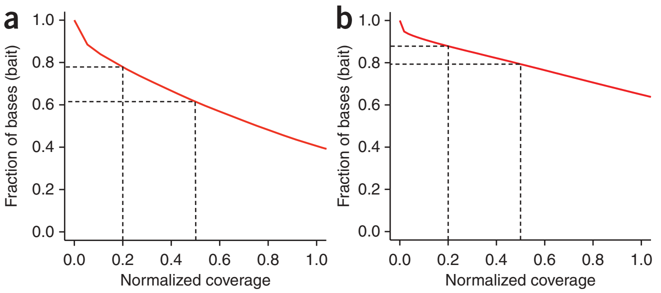


Figure 4 Normalized coverage-distribution plots. Shown is the fraction of bait-covered bases in the genome achieving coverage with uniquely aligned sequence equal or greater than the normalized coverage indicated on the *x*-axis. (**a**,**b**) The absolute per base coverage was divided by the mean coverage of all bait positions (18 in **a**; 221 in **b**). The curve for the shotgun-sequenced exon capture (**a**) is steeper than the curve for the regional capture (**b**), indicating a less uniform representation of sequencing targets in the exon catch. Dashed lines point to the fraction of bases achieving at least half or one-fifth the mean coverage.

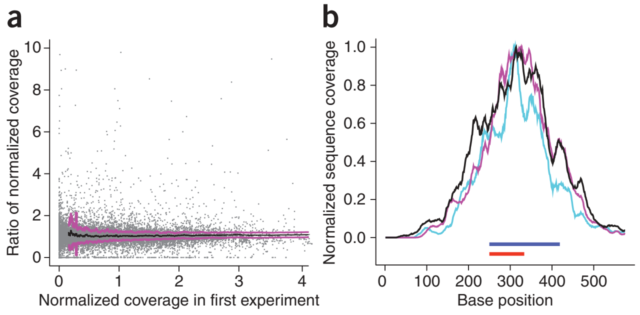


Figure 5 Reproducibility of hybrid selection. (**a**) For each exon (*n* = 15,565), the ratio of the mean coverage in two independent hybrid-selection experiments performed on the same source DNA (NA15510) was plotted over its mean coverage in one experiment. Coverage was normalized to adjust for the different number of sequencing reads. The average ratio (black line) is close to 1. S.d. is indicated by purple lines. (**b**) Base-by-base sequence coverage along one target in three independent hybrid selections, two of them performed on NA15510 (purple and teal lines) and one on NA11994 source DNA (black). Note the similiarities at this fine resolution of the three profiles, which were normalized to the same height. The position of target exon (ENSE00000968562) and bait is indicated by red and blue bars, respectively.

similar between the two technical replicates (purple and teal lines in **Fig. 5b**), consistent with the notion that variability in coverage is by and large systematic rather than stochastic. The coverage profile along the same exon in a different source DNA (Coriell NA11994) followed a similar pattern (black line in **Fig. 5b**). Additional data that demonstrate the sample-to-sample consistency of targeted sequencing of whole-genome amplified DNA samples can be found in **Supplementary Figure 4** online.

The number of exon positions where we called a high-confidence genotype in the two technical replicates was 1,586,379 and 1,578,975, respectively, that is, ~64% of the 2.5 Mb of targeted exon sequence. A total of 1,459,172 nucleotide positions were called in both. Of these, only 14 disagreed, indicating an overall discordance rate of ~10^{-5}, which is consistent with our threshold for genotype calls, that is, a logarithm of odds ratio (LOD) ≥5.

The excellent reproducibility permits sequencing of essentially the same subset of the genome in different experiments. It also allows accurate predictions of target coverage at a given number of total sequencing reads. According to a normalized coverage distribution plot for exon as opposed to bait sequence (**Supplementary Fig. 1a**), quadrupling the number of sequenced bases would increase the fraction of exon sequence called at high confidence to >80%. This can be easily achieved by longer reads and higher cluster densities on a newer Illumina GA-II instrument. Indeed, a single lane of 76-base end-sequencing reads provided high-confidence genotypes for 89% (~2.2 Mb) of the targeted exon space.

Accuracy of single-nucleotide polymorphism (SNP) detection

To assess the accuracy of SNP detection, we fished for exons in three different human samples (Coriell NA11830, NA11992 and NA11994) that had been previously genotyped for the International HapMap project. With one lane of Illumina GA-I sequence for each sample, we were able to call 7,712 sequencing-based genotypes in coding exons for direct comparison with previously obtained genotypes. Each cell line had ~3,850 genotypes in HapMap within our target exons, of which ~22% were heterozygous. As expected, the detection sensitivity of 67% (7,712 high-confidence genotype calls for 11,544 HapMap genotypes) closely matched the percentage of exon bases scanned with high confidence (64%) in these particular GA-I sequencing lanes.

The discordance rate at high-confidence sites was low (0.6%) and close to the estimated error rate of HapMap genotypes[21]. Of note, the HapMap discordancy for the very same loci in whole-genome Illumina sequencing experiments was essentially the same (0.6%). Hence, there is no evidence that the hybrid-selection process *per se* compromises the accuracy.

To resolve a representative subset of the discrepancies, we genotyped two DNA samples (Coriell NA11830 and NA11992) by mass-spectrometric primer-extension assays (Sequenom). A list of all 44 discordant genotypes plus 22 Sequenom genotypes is shown in **Supplementary Table 2** online. In 19 of 22 informative cases (86%), the Sequenom assay confirmed the sequencing-based result. Three cases were bona fide hybrid-selection sequencing errors that missed the nonreference allele at heterozygous positions. Bias against the nonreference allele may be due to preferential capture of the reference allele present in the capture probes, to preferential alignment against the reference genome or both.

Overall, the two alleles at heterozygous loci were represented almost equally on average. Based on 1,722 heterozygous SNP calls, the fraction of reads supporting the reference allele had a mean of 0.53 and a s.d. of 0.12. The nearly balanced recovery of both alleles increases the power to detect heterozygotes. Consequently, the sensitivity to detect SNPs is mainly limited by sequence coverage rather than by systematic or stochastic allelic bias or drop-out effects.

DISCUSSION

We have developed a hybrid-selection method for enriching specific subsets of a genome that is flexible, scalable and efficient. It combines the economy of oligodeoxynucleotide synthesis on an array with the favorable kinetics of RNA-driven hybridization in solution and works well for short dispersed segments and long contiguous regions alike. With further optimization, routine implementation of hybrid selection would enable deep, targeted next-generation sequencing of thousands of exons as well as of megabase-sized candidate regions implicated by genetic screens. Targeting based on hybrid selection may be potentially useful for a variety of other applications as well, where traditional singleplex PCR is either too costly or too specific in that specific primers may fail to produce a PCR product that represents all genetic variation in the sample. Examples are enrichment of precious ancient DNA that is heavily contaminated with unwanted DNA, deep sequencing of viral populations in clinical samples, or metagenomic analyses of environmental or medical specimens.

Previous methods for hybrid selection have used cloned DNA, such as bacterial artificial chromosomes or cosmids, to create capture probes for cDNA[22,23] or genomic DNA fragments[24]. Clone-based probes are suboptimal for several reasons. Readily available clones often contain extraneous sequences and are not easily configured into custom pools. Moreover, cDNAs are inefficient for capturing very short exons (data not shown). Instead of using cloned DNA, we use pools of ultra-long custom-made oligonucleotides that are synthesized in parallel on a microarray and offer much greater flexibility. In principle, one can target any arbitrary sequence. As with all hybridization-based methods, repeat elements have to be either circumvented at the bait design stage or physically blocked during the hybridization. We currently do both. There are also fundamental limits to the power of hybridization to discriminate between close paralogs, members of gene families, pseudogenes or segmental duplications.

We perform a simple pull-down with streptavidin-coated magnetic beads, a generic laboratory technique that does not require customized equipment. It can be performed in almost any tube or multi-well plate format, and there are numerous precedents for processing many samples in parallel. Our method is also largely independent of the sequencing platform. As shown here, it works well in combination with the Illumina platform whereby the hybrid-selected material can be either end sequenced or shotgun sequenced. Direct end-sequencing with longer reads is clearly preferred as it is far less complex and requires fewer amplification steps. Our protocol can also be easily adapted for the Roche 454 Sequencer (data not shown), which produces fewer but even longer reads, and, presumably, for other sequencing platforms as well.

The length of the baits allows thorough washes at high stringency to minimize contamination with nontargeted sequences that would cross-hybridize to the bait or hybridize to legitimate target fragments via the common adaptor sequence. A related source of background, indirect pull-down of repetitive passenger DNA fragments, is suppressed by addition of C_0t-1 DNA to block repeats during the hybridization.

To prepare the bait, we amplify the complex pool of synthetic oligonucleotides twice by PCR. The risk of introducing bias during the amplification is more than compensated by its advantages: first, PCR selects for full-length synthesis products; second, it helps amortizing the fixed cost of chemical oligonucleotide synthesis over a large number of DNA samples; third and most importantly perhaps, it allows storage and testing at various stages of aliquots and obviates the need for frequent chemical re-synthesis and quality control of a given set of DNA oligonucleotides.

The sensitivity is in part due to the use of single-stranded RNA as capture agent. While a 5′-biotinylated double-stranded PCR product is equally specific (data not shown), it is not as good a hybridization driver. In a hybrid selection with single-stranded RNA, each bait is present in vast (several hundred-fold) excess over its cognate target. The excess RNA drives the hybridization reaction toward completion and reduces the amount of input fragment library needed. Further, saturating the available target molecules with an excess of bait prevents all-or-none single-molecule capture events that give rise to the stochastic and skewed representation of targets and alleles in multiplex amplification[14]. It also helps normalizing differences in abundance and hybridization rates of individual baits to some extent.

An important parameter for capturing short and dispersed targets such as exons is fragment size. Longer fragments extend beyond their baits and thus contain more sequence that is slightly off-target. On the other hand, shearing genomic DNA to a shorter size range generates fewer fragments that are long enough to hybridize to a given bait at high stringency. By virtue of the high excess of bait, our protocol works well for fishing in whole-genome libraries with a mean insert size of ~250 bp, i.e., only slightly longer than the average protein-coding exon and minimum target size (164 and 170 bp, respectively). In contrast, microarray capture has a lower effective concentration of full-length probes, requires more input fragment library to drive the hybridization and becomes less efficient with input fragment libraries that have insert sizes much smaller than 500 bp[12]. Array capture is therefore better suited for longer targets, for which edge effects and target dilution by over-reaching baits or overhanging fragment ends are negligible. In fact, capturing fragments larger than the oligonucleotides is beneficial for this application as it helps extend coverage into segments next to repeats that must be excluded from the baits. Because of synergistic effects between neighboring baits, contiguous regions are less demanding targets than short isolated exons.

One advantage of hybrid selection is that long capture probes are more tolerant to polymorphisms than the shorter sequences typically used as primers for PCR or multiplex amplification. We have seen very little allelic bias and few cases of allelic drop out at SNP loci. The concordance of sequencing-based genotype calls and known HapMap genotypes was excellent (99.4%). For the majority of discrepancies that we looked at, the sequencing genotype was validated by a specific SNP-genotyping assay. We have not examined other genetic variation such as indels, translocations and inversions; the capture efficiency may be lower for such sequence variants because they differ more from the reference sequence used to design the baits.

In conclusion, the technology described here should allow extensive sequencing of targeted loci in genomes. Still, it remains imperfect with some unevenness in selection and some gaps in coverage. Fortunately, these imperfections appear to be largely systematic and reproducible. We anticipate that additional optimization, more sophisticated bait design based on physicochemical as well as empirical rules, and comprehensive libraries of pre-designed and pre-tested oligonucleotides will enable efficient, cost-effective, and routine deep resequencing of important targets and help identify biologically and medically relevant mutations.

METHODS

Capture probes (bait). Libraries of synthetic 200-mer oligodeoxynucleotides were obtained from Agilent Technologies. The pool for exon capture consisted of 22,000 oligonucleotides of the sequence 5′-ATCGCACCAGCGTGTN_{170} CACTGCGGCTCCTCA-3′ with N_{170} indicating the target-specific bait sequences. Baits were tiled along exons without gaps or overlaps starting at the left-most coding base in the strand of the reference genome sequence shown in the UCSC genome browser (that is, 5′ to 3′ or 3′ to 5′ along the coding sequence, depending on the orientation of the gene) and adding additional 170-mers until all coding bases were covered. The synthetic oligonucleotides for regional capture consisted of 10,000 200-mers that targeted 4,409 distinct 170-mer sequences, of which 3,227 were represented twice (that is, the sequence above plus its reverse complement) and 1,182 were represented thrice. For baits designed to capture a predefined set of targets, we chose the minimal set of unique olignonucleotides and added additional copies (alternating between reverse complements and the original plus strands) until the maximum capacity of the synthetic oligonucleotide array (currently up to 55,000) was reached. Note that the PCR product and the biotinylated RNA bait is the same for forward- and reverse-complemented oligonucleotides. Synthesizing plus and minus oligonucleotides for a given target may provide better redundancy at the synthesis step than synthesizing the very same sequence twice, although we have no hard evidence that reverse complementing the oligonucleotides has any measurable benefit. Complete lists of sequencing targets and oligonucleotide sequences are available as **Supplementary Table 1** and **Supplementary Data 1–3** online. Oligonucleotide libraries were resuspended in 100 µl TE0.1 buffer (10 mM Tris-HCl, 0.1 mM EDTA, pH 8.0). A 4-µl aliquot was PCR amplified in 100 µl containing 40 nmol of each dNTP, 60 pmol each of 21-mer PCR primers A (5′-CTGGGAATCGCACCAGCGTGT-3′) and B (5′-CGTGGATGA GGAGCCGCAGTG-3′), and 5 units PfuTurboCx Hotstart DNA polymerase (Stratagene). The temperature profile was 5 min. at 94 °C followed by 10 to 18 cycles of 20 s at 94 °C, 30 s at 55 °C, 30 s at 72 °C. The 212-bp PCR product was cleaned up by ultrafiltration (Millipore Montage), preparative electrophoresis on a 4% NuSieve 3:1 agarose gel (Lonza) and QIAquick gel extraction (Qiagen). The gel-purified PCR product (100 µl) was stored at –70 °C. To add a T7 promoter, a 1-µl aliquot was reamplified in 200 µl as before, except that the forward primer was T7-A (5′-GGATTCTAATACGACTCACTATAGG GATCGCACCAGCGTGT-3′) and 12 to 15 PCR cycles were sufficient. Qiagen-purified 232-bp PCR product (1 µg) was used as template in a 100-µl MAXIscript T7 transcription (Ambion) containing 0.5 mM ATP, CTP and GTP, 0.4 mM UTP and 0.1 mM Biotin-16-UTP (Roche). After 90 min. at 37 °C, the unincorporated nucleotides and the DNA template were removed by gel filtration and TURBO DNase (Ambion). The yield was typically 10–20 µg of biotinylated RNA as determined by a Quant-iT assay (Invitrogen), that is, enough for 20–40 hybrid selections. Biotinylated RNA was stored in the presence of 1 U/µl SUPERase-In RNase inhibitor (Ambion) at –70 °C.

Whole-genome fragment libraries (pond). Whole-genome fragment libraries were prepared using a modification of Illumina's genomic DNA sample preparation kit. Briefly, 3 μg of human genomic DNA (Coriell) was sheared for 4 min. on a Covaris E210 instrument set to duty cycle 5, intensity 5 and 200 cycles per burst. The mode of the resulting fragment-size distribution was ~250 bp. End repair, nontemplated addition of a 3′-A, adaptor ligation and reaction clean-up followed the kit protocol except that we used a generic adaptor for libraries destined for shotgun sequencing after hybrid selection. This adaptor consisted of oligonucleotides C (5′-TGTAACATCACAGCATCAC CGCCATCAGTCxT-3′ with 'x' denoting a phosphorothioate bond resistant to excision by 3′–5′ exonucleases) and D (5′-[PHOS]GACTGATGGCGCACTAC GACACTACAATGT-3′). The ligation products were cleaned up and size-selected on a 4% NuSieve 3:1 agarose gel followed by QIAquick gel extraction. A standard prep starting with 3 μg of genomic DNA yielded ~500 ng of size-selected material with genomic inserts ranging from ~200 to ~350 bp, that is, enough for one hybrid selection. To increase the yield we typically amplified an aliquot by 12 cycles of PCR in Phusion High-Fidelity PCR master mix with HF buffer (NEB) using Illumina PCR primers 1.1 and 2.1, or, for libraries with generic adapters, oligonucleotides C and E (5′-ACATTGTAGTGTCGTAG TGCGCCATCAGTCxT-3′) as primers. After QIAquick clean-up, if necessary, fragment libraries were concentrated in a vacuum microfuge to 250 ng per μl before hybrid selection.

Hybrid selection. A 7-μl mix containing 2.5 μg human Cot-1 DNA (Invitrogen), 2.5 μg salmon sperm DNA (Stratagene) and 500 ng whole genome fragment library was heated for 5 min. at 95 °C, held for 5 min. at 65 °C in a PCR machine and mixed with 13 μl prewarmed (65 °C) 2× hybridization buffer (10× SSPE, 10× Denhardt's, 10 mM EDTA and 0.2% SDS) and a 6-μl freshly prepared, prewarmed (2 min. at 65 °C) mix of 500 ng biotinylated RNA and 20 U SUPERase-In. After 66 h at 65 °C, the hybridization mix was added to 500 ng (50 μl) M-280 streptavidin Dynabeads (Invitrogen), that had been washed 3 times and were resuspended in 200 μl 1M NaCl, 10 mM Tris-HCl, pH 7.5, and 1 mM EDTA. After 30 min. at 20 °C, the beads were pulled down and washed once at 20 °C for 15 min. with 0.5 ml 1× SSC/0.1% SDS, followed by three 10-min. washes at 65 °C with 0.5 ml prewarmed 0.1× SSC/0.1% SDS, resuspending the beads once at each washing step. Hybrid-selected DNA was eluted with 50 μl 0.1 M NaOH. After 10 min. at 20 °C, the beads were pulled down, the supernatant transferred to a tube containing 70 μl 1 M Tris-HCl, pH 7.5, and the neutralized DNA desalted and concentrated on a QIAquick MinElute column and eluted in 20 μl. We routinely use 500 ng of pond and bait per reaction but have seen essentially identical results in proportionally scaled-down 5-μl reactions with 100 ng each.

Catch processing and sequencing. For fragment libraries carrying standard Illumina adaptor sequences, 4 μl of hybrid-selected material was amplified for 14 to 18 cycles in 200 μl Phusion polymerase master mix and PCR primers 1.1 and 2.1 and the PCR product cluster amplified and end sequenced for 36 or 76 cycles. Hybrid-selected material with generic adaptor sequences (8 μl) was amplified in 400 μl Phusion High-Fidelity PCR master mix for 14 to 18 cycles using PCR primers F (5′-CGCTCAGCGGCCGCAGCATCACCGCCATCAGT-3′) and G (5′-CGCTCAGCGGCCGCGTCGTAGTGCGCCATCAGT-3′). Initial denaturation was 30 s at 98 °C. Each cycle was 10 s at 98 °C, 30 s at 55 °C and 30 s at 72 °C. Qiagen-purified PCR product (~1 μg) was digested with *Not*I (NEB), cleaned-up (Qiagen MinElute) and concatenated in a 20-μl ligation reaction with 400 U T4 DNA ligase (NEB). After 16 h at 16 °C, reactions were cleaned up and sonicated. Sample preparation for Illumina sequencing followed the standard protocol except that the PCR amplification was limited to ten cycles.

Genotyping. Specific custom SNP genotyping was performed in 24-plex PCR and primer-extension reaction format using MassARRAY iPLEX chemistry and mass-spectrometric detection (Sequenom).

Computational methods. All coverage and SNP statistics are for single lanes (1/8 of a flow cell) of sequencing data. Illumina reads were collected from the instrument and aligned to the human genome using the ImperfectLookupTable (ILT) of the ARACHNE genome assembly suite[25] which is available with documentation at http://www.broad.mit.edu/wga. Briefly, a lookup table of the locations of every 12-mer in the genome was computed. For a single read, each 12-mer in the read was looked up, and all occurrences of each 12-mer were considered putative placements. Each putative placement of the read in the genome was interrogated for number of mismatches. No insertions or deletions were considered. To ensure high quality and unique placements, only reads with four or fewer errors and a next-best placement at least three errors worse were considered. Coverage at each reference position was accumulated from the unique alignments. All aligned bases were included in the basic coverage calculations. High-confidence base calls (and coverage calculations based thereon) excluded bases that failed a signal clarity filter. The filter was that the ratio of brightest dye color to next-brightest dye color had to be 2 or greater. Typically, ~80% of aligned bases passed this filter. Genotypes at each position were inferred with a straightforward Bayesian model. The likelihood of the observed data P(data|genotype) assuming each genotype at each position was computed with the assumptions that each allele is equally likely to be observed and miscalls occur with a rate of 1/1,000. These genotypes were combined with a prior probability over the genotypes defined by the reference. The prior probability used was: P(homozygous reference) = 0.999, P(heterozygous ref/nonref) = 0.001, P(nonref) = 0.00001. This yields the posterior probability P(genotype|data). The most likely genotype was selected. The confidence in our call of the specific genotype was the ratio of the best to next-best theory. We used a best-to-next-best ratio of 10^5 (LOD score 5) as threshold for calling a high-confidence genotype. The confidence in our belief that there was a SNP (independent of the specific genotype) was the ratio of the best theory to the reference. We used a best-to-reference ratio of 10^5 as our minimum confidence cutoff for reporting a SNP. Genome coordinates are zero-offset and for NCBI Build 35 (hg17). Raw unaligned Illumina sequences in SRF (sequence read format) from the hybrid-selection experiments described here are available at DNS (http://www.broad.mit.edu/annotation/hybrid_selection/).

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

ACKNOWLEDGMENTS

We thank the staff of the Broad Institute Genome Sequencing Platform and Genetic Analysis Platform for generating sequencing and genotyping data. This work was supported by National Human Genome Research Institute grant HG03067-05 (to E.S.L.) and funds of the Broad Institute.

AUTHOR CONTRIBUTIONS

A.M. and P.R. developed the wet lab protocol. J.M., W.B., T.F., C.R., S.G. and D.B.J. developed computational tools and analyzed data. E.M.L. synthesized the 200mer oligodeoxynucleotide pools. G.G. and S.F. prepared and sequenced fragment libraries. A.G., E.S.L and C.N. designed and directed the project and wrote the paper.

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. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
2. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
3. Bentley, D.R. *et al.* Accurate whole genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
4. Smith, D.R. *et al.* Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res.* **18**, 1638–1642 (2008).
5. Ley, T.J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).
6. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–66 (2008).
7. Wheeler, D.A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
8. Dahl, F., Gullberg, M., Stenberg, J., Landegren, U. & Nilsson, M. Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic Acids Res.* **33**, e71 (2005).
9. Albert, T.J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* **4**, 903–905 (2007).

10. Dahl, F. *et al.* Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proc. Natl. Acad. Sci. USA* **104**, 9387–9392 (2007).
11. Fredriksson, S. *et al.* Multiplex amplification of all coding sequences within 10 cancer genes by Gene-Collector. *Nucleic Acids Res.* **35**, e47 (2007).
12. Hodges, E. *et al.* Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* **39**, 1522–1527 (2007).
13. Okou, D.T. *et al.* Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods* **4**, 907–909 (2007).
14. Porreca, G.J. *et al.* Multiplex amplification of large sets of human exons. *Nat. Methods* **4**, 931–936 (2007).
15. Krishnakumar, S. *et al.* A comprehensive assay for targeted multiplex amplification of human DNA sequences. *Proc. Natl. Acad. Sci. USA* **105**, 9296–9301 (2008).
16. Clamp, M. *et al.* Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci. USA* **104**, 19428–19433 (2007).
17. Nilsson, M. *et al.* Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science* **265**, 2085–2088 (1994).
18. Hardenbol, P. *et al.* Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat. Biotechnol.* **21**, 673–678 (2003).
19. Dohm, J.C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**, e105 (2008).
20. Quail, M.A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nat. Methods* **5**, 1005–1010 (2008).
21. Frazer, K.A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
22. Lovett, M., Kere, J. & Hinton, L.M. Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. *Proc. Natl. Acad. Sci. USA* **88**, 9628–9632 (1991).
23. Parimoo, S., Patanjali, S.R., Shukla, H., Chaplin, D.D. & Weissman, S.M. cDNA selection: efficient PCR approach for the selection of cDNAs encoded in large chromosomal DNA fragments. *Proc. Natl. Acad. Sci. USA* **88**, 9623–9627 (1991).
24. Bashiardes, S. *et al.* Direct genomic selection. *Nat. Methods* **2**, 63–69 (2005).
25. Jaffe, D.B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91–96 (2003).

nature biotechnology

Prediction of high-responding peptides for targeted protein assays by mass spectrometry

Vincent A Fusaro[1,2], D R Mani[1], Jill P Mesirov[1] & Steven A Carr[1]

Protein biomarker discovery produces lengthy lists of candidates that must subsequently be verified in blood or other accessible biofluids. Use of targeted mass spectrometry (MS) to verify disease- or therapy-related changes in protein levels requires the selection of peptides that are quantifiable surrogates for proteins of interest. Peptides that produce the highest ion-current response (high-responding peptides) are likely to provide the best detection sensitivity. Identification of the most effective signature peptides, particularly in the absence of experimental data, remains a major resource constraint in developing targeted MS–based assays. Here we describe a computational method that uses protein physicochemical properties to select high-responding peptides and demonstrate its utility in identifying signature peptides in plasma, a complex proteome with a wide range of protein concentrations. Our method, which employs a Random Forest classifier, facilitates the development of targeted MS–based assays for biomarker verification or any application where protein levels need to be measured.

Proteomic discovery experiments in case-and-control comparisons of tissue or proximal fluids frequently generate lists comprising many tens to hundreds of candidate biomarkers[1]. Integrative genomic approaches incorporating microarray data and literature mining are also increasingly being used to guide identification of candidate protein biomarkers. To further credential biomarker candidates and move them toward possible clinical implementation, it is necessary to determine which of the proteins from lists of candidates differentially abundant in diseased versus healthy patients can be detected in body fluids, such as blood, that can be assayed with minimal invasiveness[1].

This process, termed verification, has historically been approached using antibodies. High-quality, well-characterized collections of antibodies suitable for protein detection in tissue are now being developed[2]. But unfortunately, the required immunoassay-grade antibody pairs necessary for sensitive and specific detection in blood exist for only a tiny percentage of the proteome. Thus, for the majority of proteins, suitable reagents for their detection and quantification in blood (or other biofluids) do not yet exist and alternative technologies are needed to bridge the gap between discovery and clinical-assay development. This problem is an important aspect of the larger need in biology and medicine for quantitative methods to measure the presence and abundance of any protein of interest.

Targeted MS is emerging as an assay technology capable of selective and sensitive detection and quantification of potentially any protein of interest (or modification thereof) in the proteome[3–6]. In stable isotope dilution–multiple reaction monitoring (MRM)-MS, peptides (precursors) from candidate proteins of interest are selectively detected and caused to fragment (products) in the mass spectrometer. The resulting product ions are used to quantify the peptide, and therefore, the protein from which it was derived, by calculating the ratio of the signal response of the endogenous peptide to a stable isotope–labeled version of the peptide added as an internal standard[3–6].

The first step in developing an MRM-MS–based assay involves selecting a subset of peptides to use as quantitative surrogates for each candidate protein. ‘Signature peptides’[1] correspond to the subset of ‘proteotypic peptides’[7] that, in addition to being sequence unique and detectable, are also the highest responding peptides for each protein. Current methods rely on selecting signature peptides based on detection in the initial MS discovery data[3,5], identification in databases of MS experimental data[8,9] or computational approaches to predict proteotypic peptides[10–13]. When multiple peptides are detected for a candidate protein for which experimental data are available, selection is primarily based on high peptide-response. Other considerations such as high-performance liquid chromatography (HPLC) retention time, amino acid composition, uniqueness in the genome and charge state also play a role. After selecting signature peptides, the targeted MRM-MS assay must be optimized for each peptide to select appropriate precursor-to-product ion transitions[5,14]. Because some peptides fail the optimization process due to poor chromatography, solubility problems, interference with matrix or failure to recover the peptide after digestion in plasma, it is common for laboratories to evaluate approximately five peptides per protein. This usually insures that at least one peptide per protein is suitable for developing a quantitative assay[3,5].

Two key problems usually arise with the selection of signature peptides for assay development. First, only a fraction of peptides present in a complex sample are detected in discovery proteomic experiments. This undersampling problem is well known and leads

[1]Broad Institute of Massachusetts Institute of Technology and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [2]Bioinformatics Program, Boston University, 24 Cummington Street, Boston, Massachusetts 02215, USA. Correspondence should be addressed to S.A.C. (scarr@broad.mit.edu) or J.P.M. (mesirov@broad.mit.edu).

Received 17 October 2008; accepted 3 January 2009; published online 25 January 2009; doi:10.1038/nbt.1524

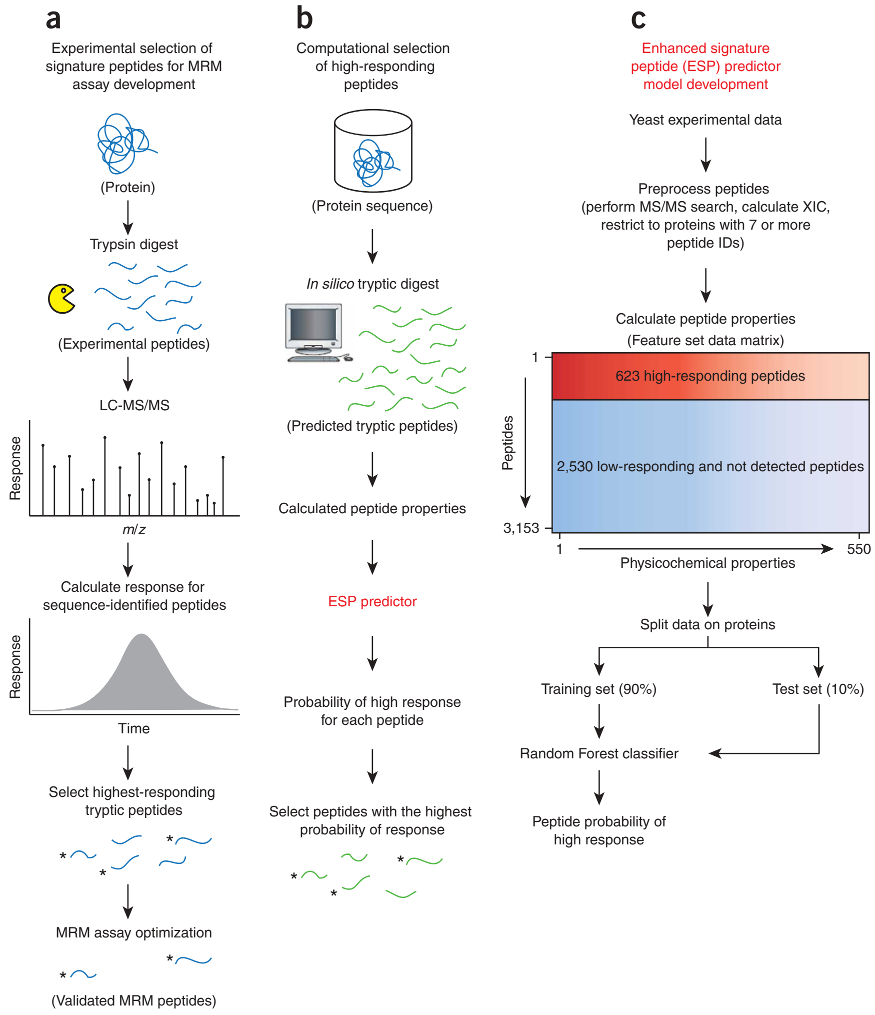

a
Experimental selection of signature peptides for MRM assay development
(Protein)
Trypsin digest
(Experimental peptides)
LC-MS/MS
Response
m/z
Calculate response for sequence-identified peptides
Response
Time
Select highest-responding tryptic peptides
MRM assay optimization
(Validated MRM peptides)
b
Computational selection of high-responding peptides
(Protein sequence)
In silico tryptic digest
(Predicted tryptic peptides)
Calculated peptide properties
ESP predictor
Probability of high response for each peptide
Select peptides with the highest probability of response
c
Enhanced signature peptide (ESP) predictor model development
Yeast experimental data
Preprocess peptides (perform MS/MS search, calculate XIC, restrict to proteins with 7 or more peptide IDs)
Calculate peptide properties (Feature set data matrix)
1
623 high-responding peptides
Peptides
2,530 low-responding and not detected peptides
3,153
1
550
Physicochemical properties
Split data on proteins
Training set (90%)
Test set (10%)
Random Forest classifier
Peptide probability of high response

Figure 1 ESP application and model development overview. (**a**) A typical proteomic workflow to select signature peptides for targeted protein analysis using MRM. Candidate proteins are experimentally analyzed, and five signature peptides per protein are selected based primarily on high peptide-response and sequence composition, among other factors. After optimization, the remaining peptides are referred to as validated MRM peptides. (**b**) We computationally digest each candidate protein, *in silico* (no missed cleavages, 600–2,800 Da), to produce a set of predicted tryptic peptides. Peptide sequences are input into the ESP predictor and we select the five peptides with the highest probability of response for each protein. To validate the ESP predictions, we compare the top five predicted peptides to the experimentally determined five highest-responding peptides from **a**, denoted by asterisks (3 out of 5, in this example). (**c**) We developed the ESP predictor using peptides from a yeast lysate experimental analysis. We trained the ESP predictor using Random Forest on 90% of the peptides and held out 10% to test the model, referred to as Yeast test. We split the data at the protein level to avoid any bias in training and testing the model on peptides from the same protein and to keep the training and test data completely separated.

to poor reproducibility of peptide and protein detection, even in replicate samples[15]. As a result, the best signature peptides for any given candidate may not be the ones observed in the discovery experiment. Second, it is of interest to quantify candidate proteins identified by methods other than proteomics, such as genomic experiments or literature mining. These candidate proteins may represent biomarkers or key components in signaling or metabolic pathways. In these situations, *de novo* prediction of signature peptides is required.

Here we describe the enhanced signature peptide (ESP) predictor, a computational method to predict high-responding peptides from a given protein. We (i) validate the method on ten diverse experimental data sets not used in training the ESP predictor, (ii) show that ESP predictions are significantly better at selecting high-responding peptides than existing computational methods[10,12,13], (iii) demonstrate that the ESP predictor can be used to define the best peptides for targeted MRM-MS–based assay development in the absence of experimental proteomic data for the protein and (iv) identify the most relevant physicochemical properties used to predict high-responding peptides in the context of electrospray ionization (ESI)-MS.

RESULTS

Method overview

We developed a model to predict the probability that a peptide from a given protein will generate a high response in an ESI-MS experiment. We define peptide response as the sum of the extracted ion chromatogram (XIC) based on the monoisotopic peak for all charge states and modifications detected from sequence-identified peptides. This measure is more consistent with the intended application of the ESP predictor, which is to predict signature peptides from an *in silico* digest of a candidate protein (**Fig. 1a,b**).

We used liquid chromatography (LC)-ESI-MS analyses of a yeast lysate sample, from three proteomic laboratories, to derive a training set to model peptide response (**Fig. 1c**). For each protein, we standardized the peptide response, using the *z*-score (*z*), and selected a threshold to define 'high' ($z \geq 0$) and 'low' ($z \leq -1$) responding peptides. We also derived a set of 'not detected' peptides from an *in silico* tryptic digest (no missed cleavages, mass 600–2,800 Da), but we considered only peptides not sequence identified in any form, including missed cleavages. Because we are only interested in detecting high-responding peptides, we combined the 'low' and 'not detected' peptides together to create the final training set of 'high' versus 'low/not detected'.

To develop a predictive model, one must encode the peptides as an *n*-dimensional property vector. These properties represent specific characteristics of the peptides such as mass, hydrophobicity and gas-phase basicity. We considered 550 physicochemical properties (**Supplementary Table 1** online) to model peptide response[16,17]. For each physicochemical property, we computed the property value by averaging over all amino acids in each peptide. Thus, the training set comprised a matrix of 'peptides by properties' along with the class labels, 'high' or 'low/not detected'.

We modeled peptide response using the Random Forest[18] algorithm. Random Forest is a nonlinear ensemble classifier composed of many individual decision trees. We chose Random Forest because the algorithm, and its R implementation[19], conveniently includes many features especially suited to this type of analysis. Specifically, Random Forest effectively handles data sets with large numbers of correlated

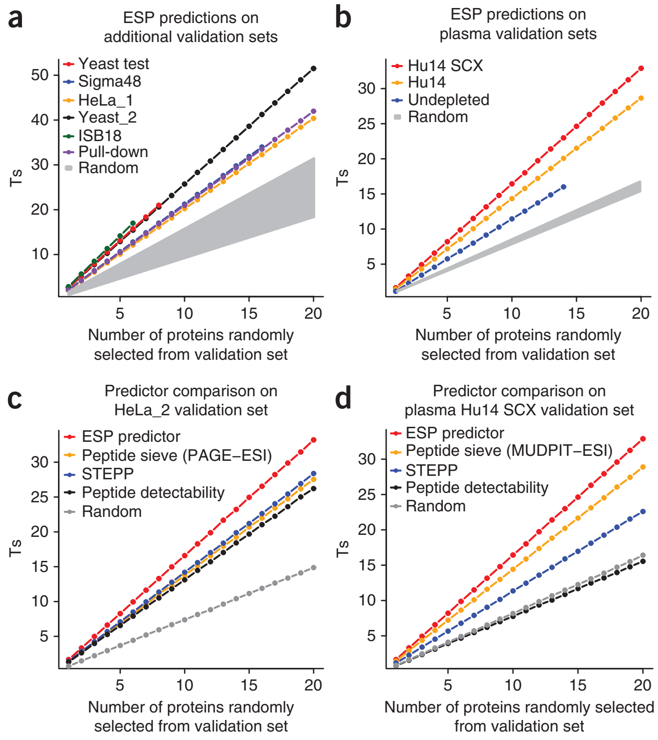
Table 1 Description of validation sets

Validation set[a]	Experiment type (ESI)	Proteins[b]	Theoretical peptides[c]	PS ≥1[d]	PS ≥2[e]	Ts[f]	Mixture complexity[g]	Database search	Quantification
ISB-18	LC-MS	6	153	100%	100%	17[h]	Low	Spectrum Mill	XIC
Yeast test	LC-MS	8	226	100%	88%	21[h]	Medium	Spectrum Mill	XIC
Plasma	LC-MS	14	633	71%	36%	16[i]	Very High	Spectrum Mill	XIC
Sigma48	LC-MS	16	438	88%	69%	34[h]	Low	Spectrum Mill	XIC
Plasma Hu14	LC-MS	30	1,403	87%	43%	43[h]	Very high	Spectrum Mill	XIC
Yeast_2	LC-MS	94	1,930	97%	82%	242[h]	Medium	Spectrum Mill	XIC
HeLa_1	LC-MS	149	4,944	90%	65%	301[h]	High	Mascot	MSQuant
HeLa_2	GeLC-MS	300	15,172	86%	54%	498[h]	High	Mascot	MSQuant
Pull-down	GeLC-MS	172	8,062	92%	68%	358[h]	Medium	Mascot	MSQuant
Plasma Hu14 SCX	SCX-LC-MS	45	1,935	93%	49%	74[h]	High	Spectrum Mill	XIC

[a]All validation sets were analyzed using an LTQ-Orbitrap except the plasma and ISB-18 data, which were analyzed using an LTQ-FT. [b]Only proteins with six or more theoretical peptides (*in silico* digest) and at least five sequence-identified peptides were considered for validation. [c]*In silico* tryptic digest with no missed cleavages and a mass range of 600–2,800 Da. [d]Protein sensitivity (PS). The percent of proteins with one or more peptides predicted by the ESP predictor to be among the five highest responding. The weighted mean of all validation sets based on number of proteins is 89%. [e]The percent of proteins with two or more peptides predicted by the ESP predictor to be among the five highest responding. The weighted mean of all validation sets based on number of proteins is 60%. [f] Test statistic (Ts). The sum of correctly predicted peptides among the five highest-responding peptides for all proteins in the validation set. [g]Simple comparison of the number of proteins present in each sample mixture. For example, plasma has more than 10,000 proteins (very high) compared to sigma48, which has 48 proteins (low). [h]$P < 0.0001$, [i]$P = 0.0363$ based on null distribution for the entire validation set, by permutation test.

features, provides insight into the model by determining the most relevant properties during training[18–21] and exhibits better performance for this data set than using a Support Vector Machine[22] does. Notably, the structure of the decision trees that make up the final model are learned using only the training set, and the model is fixed for subsequent testing and validation.

We also attempted to reduce the dimensionality of the training set by considering two feature-selection techniques, Fisher Criterion Score[23] and the area under the receiver operator curve (ROC)[24]. We used the best features ranked by each of the feature selection methods to build Support Vector Machine models using three different kernels and Random Forest. Random Forest exhibited the best performance using all 550 properties, implying that feature selection is not helpful in this context (**Supplementary Fig. 1** online).



Metrics to evaluate the ESP predictor

We created the ESP predictor to select high-responding peptides from candidate proteins, in the absence of MS experimental data, with the intention of developing an MRM-MS assay. Therefore, we developed metrics to assess the success of such predictions. When developing an MRM-MS assay, it is necessary to evaluate the assay performance of about five peptides per protein in the biological matrix of interest (typically plasma) to reliably obtain at least one peptide with suitable limits of detection and quantification. The expense and time associated with generating synthetic peptides and evaluating the assay performance of each for MRM quantification (typically involving generation of a ten-point concentration response curve for each peptide) make evaluation of more than five peptides per protein impractical.

We evaluated the ESP predictor on ten validation sets not used in training to assess its performance (**Table 1**). We experimentally analyzed each validation set using ESI-MS and selected the five highest-responding, fully tryptic (no missed cleavages) peptides from each protein (**Fig. 1a**). Then, using the ESP predictor, we ranked the predicted probability of high response for all tryptic peptides

Figure 2 ESP predictor validation and method comparison. ESP predictions outperform existing computation models and are statistically significant for all validation data sets based on a random permutation test. We plotted the mean number of cumulative correctly predicted peptides (Ts) for random combinations of 1–20 proteins. We calculated the 95% confidence interval of the mean, but the error bars were too small to display. The null distribution for *P*-value calculation is derived using a predictor that randomly selects the top five high-responding peptides for a protein (**Supplementary Fig. 2**). (**a**) ESP predictor performance on multiple validation sets, with the performance of a random predictor shown in gray. Each validation set produces its own set of random distributions, depending on the number of peptides per protein. We grouped all random distributions into a single shaded area. (**b**) ESP predictions on plasma validation sets. The samples represent undepleted plasma, top 14 most-abundant proteins depleted, and depleted and then fractionated using SCX (also referred to as MUDPIT). Random selection of the top five peptides resulted in the gray area. (**c**) Comparison between the ESP predictor, proteotypic predictors and random predictions on a HeLa GeLC-MS cell lysate. (**d**) Comparison between the ESP predictor, proteotypic predictors and random predictions on a depleted and fractionated plasma sample. This is the sample type most commonly used for MRM biomarker verification. See **Tables 1** and **2** for more details. STEPP, SVM technique for evaluating proteotypic peptides.

Table 2 Comparison of computational methods

Method	Validation set	PS ≥1[a]	PS ≥2[b]	Ts[c]
ESP Predictor	HeLa_2 (GeLC-MS)	86%	54%	498[d]
STEPP[13]	HeLa_2 (GeLC-MS)	80%	44%	425[d]
Peptide sieve (PAGE-ESI)[10]	HeLa_2 (GeLC-MS)	77%	43%	413[d]
Peptide detectability[12]	HeLa_2 (GeLC-MS)	77%	41%	394[d]
ESP predictor	Plasma Hu14 SCX	93%	49%	74[d]
Peptide sieve (MUDPIT-ESI)	Plasma Hu14 SCX	82%	46%	65[d]
STEPP	Plasma Hu14 SCX	69%	36%	51[e]
Peptide detectability	Plasma Hu14 SCX	62%	13%	35[f]

The ESP predictor demonstrates the best performance compared to existing computational methods. Refer to **Table 1** for additional validation set information. STEPP, SVM technique for evaluating proteotypic peptides.

[a]Protein sensitivity (PS): The percent of proteins with one or more peptides predicted by the ESP predictor to be among the five highest responding. [b]The percent of proteins with two or more peptides predicted by the ESP predictor to be among the five highest responding. [c]Test statistic (Ts). The sum of correct peptides among the five highest-responding peptides for all proteins in the validation set. [d]$P < 0.0001$, [e]$P = 0.0029$, [f]$P = 0.6685$ based on null distribution for the entire validation set, by permutation test.

generated from an *in silico* digest of the same proteins and selected the top five peptides for each protein (**Fig. 1b**). We calculated two metrics designed to assess how well the ESP predictor selected the five highest-responding peptides for all proteins in each validation set. First, we calculated the protein sensitivity, which is the percent of proteins with one or more peptides predicted by the ESP predictor to be among the five highest responding. Second, we calculated a P-value to test the hypothesis that the ESP predictions are significantly better than random predictions, using a permutation test. In gauging the performance of the ESP predictor, a combination of high protein sensitivity and low *P*-value is desirable. A high protein sensitivity indicates that more proteins in the data set have at least one correctly predicted high-responding peptide, whereas statistical significance requires $P < 0.05$. We also compared the ESP predictor to three publicly available computational methods for predicting proteotypic peptides[10,12,13].

Validation of the ESP predictor

We wanted to demonstrate the advantage of applying a single model to predict high-responding peptides in varied data spanning a wide range of different ESI experimental types, mixture complexities, database search algorithms and XIC quantification methods. For a fair assessment of how well the ESP predictor selects the five highest-responding peptides, we restricted the validation sets to proteins with six or more theoretical peptides and five or more sequence-identified peptides. The results indicate the ESP predictor performance is consistent across all ten validation sets despite very different types of proteomic data (**Table 1**). On average, the ESP predictor achieves a success rate of 89% at selecting one or more high-responding peptides per protein. Across all validation sets, the ESP predictor correctly selects approximately two out of five high-responding peptides from an average of 42 theoretical peptides per protein.

Next, we used a permutation test to confirm that the ESP predictions are statistically more significant, across multiple proteins, than random predictions and current computational methods (**Fig. 2** and **Supplementary Fig. 2** online). The predictions on nine of the ten validation sets tested were significantly better than random ($P < 0.0001$). Only the predictions on the most complex mixture, undepleted plasma, were less significant ($P = 0.036$). The predictions for the undepleted plasma are better understood in the context of predictions for the Plasma Hu14 (with the 14 most abundant proteins depleted) and Plasma Hu14 SCX (depleted and fractionated) validation sets (**Fig. 2b**). The number of correct peptides selected significantly increases (**Table 1**) as the mixture complexity decreases, suggesting less ion suppression and better quantification due to less interference.

We also compared the performance of the ESP predictor on the HeLa_2 and Plasma Hu14 SCX validation sets to three computational methods designed to predict proteotypic peptides (**Table 2**). We demonstrate, using the HeLa_2 and Plasma Hu14 SCX validation sets, that our method for selecting high-responding peptides performs significantly better (based on Ts, **Table 2**) than methods designed to predict proteotypic peptides (**Fig. 2c,d**). Compared to the HeLa_2 validation set, these other methods exhibit more variability with the Plasma Hu14 SCX validation set, whereas ESP still performed well. Performance on fractionated plasma is especially important because it represents a sample type frequently used in MRM biomarker verification. It is relevant to note that these studies constitute the first evaluation of the performance of peptide response predictors in the context of plasma, the most difficult proteome of all with respect to complexity and dynamic range of protein abundance.

To further demonstrate the robustness of the ESP predictor, we examined three quantification methods to calculate peptide response using the HeLa_1 validation set. In addition to MSQuant (**Table 1** and **Fig. 2a**), we also searched the raw data using Spectrum Mill, which reports peptide intensity. We also calculated the XIC based on the monoisotopic peak from the raw data. All three methods exhibited similar performance (**Supplementary Fig. 3** online). This suggests the ESP predictor is agnostic to the method of calculating peptide response, as long as it is done consistently.

The ESP predictor selects optimal signature peptides for MRM-MS assays

Having validated that the ESP predictor is successful at predicting high-responding peptides, we sought to determine if the predictions can be used to select signature peptides to configure MRM-MS assays in plasma. We tested the ability of the ESP predictor to select the correct signature peptides for a set of 14 proteins (9 cardiovascular biomarkers, 4 nonhuman proteins and prostate-specific antigen). For each of these proteins, we had previously experimentally defined the validated MRM peptides and then configured successful MRM-MS assays using these peptides. We used the ESP predictor to select five candidate signature-peptides and compared the results to the validated MRM peptides for each protein (**Fig. 1a,b**). The ESP predictor correctly selected two validated MRM peptides per protein, on average, yielding a protein sensitivity of 93% (**Fig. 3** and **Supplementary Data** online for all plots).

We then evaluated the usefulness of a proteomic database in defining signature peptides for MRM-MS assay configuration for these 14 proteins. Using the MRM feature of the Global Proteome Machine (GPM) respository[8], a well-known and comprehensive database of proteomic experimental data, we obtained an average of only 0.8 validated MRM peptides per protein. Most importantly, for six of these proteins, no prior MS experimental data existed in GPM (CD40, BNP, HRP, IL-33, leptin, and MBP). For these six proteins, ESP correctly predicted 12 out of 18 validated MRM peptides. Across all 14 proteins, only 11 of the 39 validated MRM peptides were found in GPM, whereas ESP correctly predicted 29 of the 39 validated MRM peptides (**Supplementary Table 2** online). For the eight proteins for which data were available in GPM, there was good agreement between the ESP predictor and GPM in predicting validated MRM peptides (**Fig. 3c**). These results point to potential issues in using proteomic data in databases for MRM-MS assay configuration, as recently noted by others[25], and underscore the need for a computational approach to select signature peptides in the absence of MS experimental data.

Important physicochemical properties

One major benefit of using Random Forest is that it facilitates model interpretation by determining an importance score for each physicochemical property. We followed a procedure similar to that described previously[26] to determine the number of important properties. Briefly, we randomly split the yeast training data into train (80%) and test (20%) sets. We then trained a model using all 550 properties and recorded the test error using the variable importance measure to rank the properties. Note that the variable importance measure was calculated once using all properties to avoid overfitting. Next, we repeatedly removed the least important half of the properties and recorded the test-set error at each step. We repeated this entire process 100 times to produce an error distribution (**Fig. 4a**). Because the test error was distributed normally, we used a two-tailed *t*-test to determine the minimum number of properties at which the test error distributions were no longer significantly different ($P < 0.05$). We selected 35 properties as the most important and grouped them into five major categories (**Fig. 4c** and **Supplementary Methods** online for more information about the 35 properties). Even though we used all properties in the final Random Forest model, we selected these 35 properties to gain some insight into an interpretation of the model.

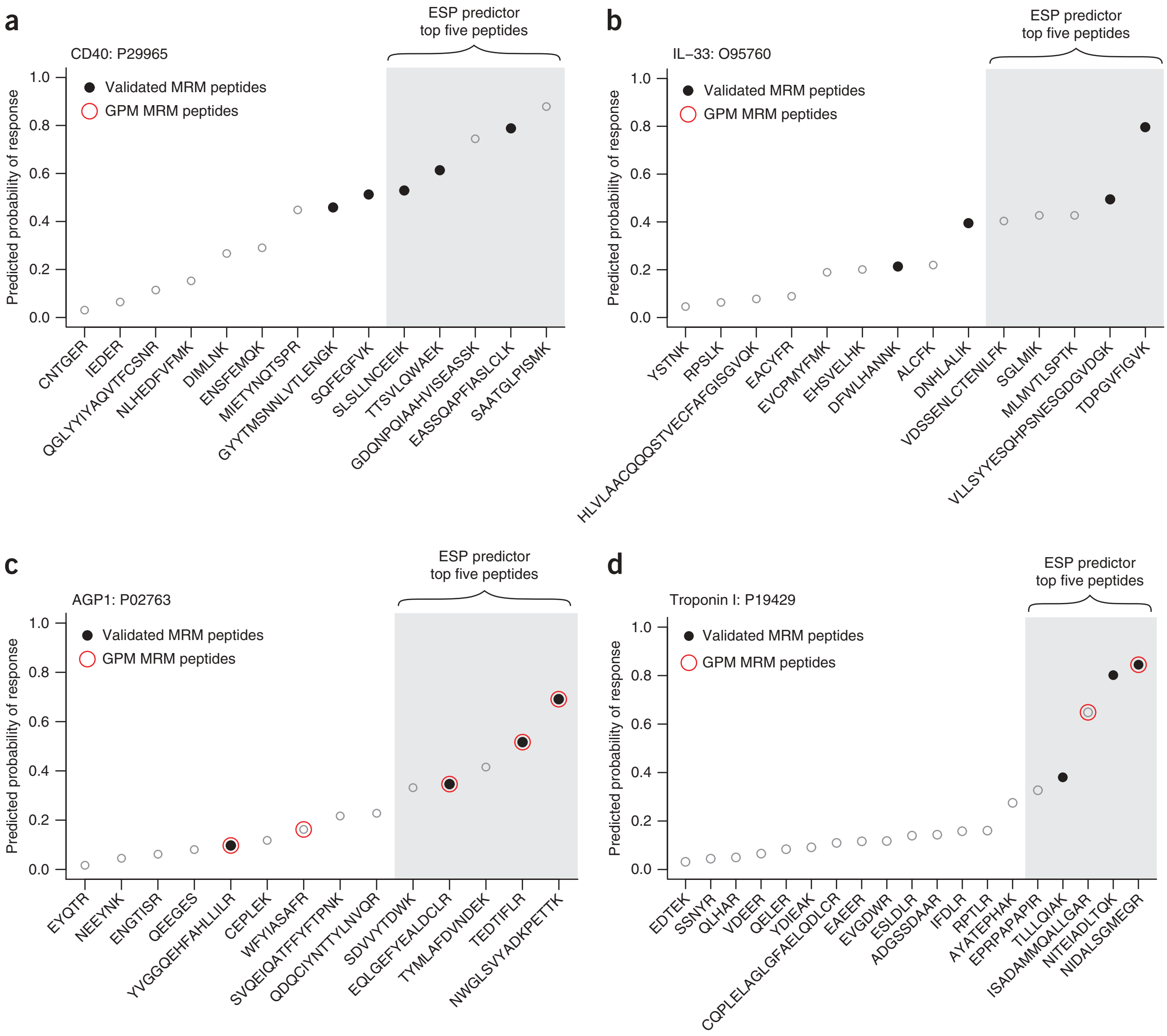
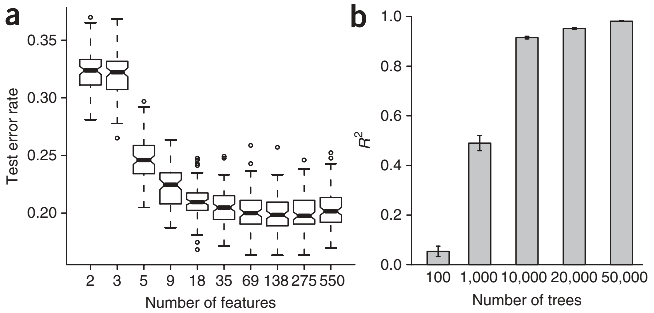


Figure 3 ESP predictions translate into experimentally validated MRM peptides. For each protein, we performed an *in silico* digest (600–2,800 Da) and ensured that the top five peptides predicted by the ESP predictor were unique in the Swiss-Prot human database. Although additional filtering criteria could easily be applied after analysis with the ESP predictor, we opted for no filtering (except top five uniqueness) to demonstrate the simplicity of using the ESP predictor to select candidate signature-peptides to configure an MRM-MS assay. For all plots, peptides are sorted by the ESP predicted probability of response (*y*-axis). The actual rank order of measured peptide response is shown in **Supplementary Table 2.** (**a**) The ESP predictor correctly selected all three validated MRM peptides (filled black circles) out of the five predicted candidate signature-peptides for troponin I. (**b**) The ESP predictor correctly selected two validated MRM peptides out of the five predicted candidate signature-peptides for IL-33. In **a** and **b**, two representative proteins not found in the GPM database are shown. (**c**) GPM correctly selected all four of the validated MRM peptides among the top five. Three peptides are common between the ESP predictor and GPM. (**d**) Only two peptides were suggested by GPM of which only one was a validated MRM peptide. In **c** and **d**, two representative proteins are shown where we overlaid the MRM peptides suggested by GPM (open red circles). Example **d** highlights the limitations of relying solely on database predictions because two validated MRM peptides would have been missed.



c

Rank	Peptide properties
1	Mass
2	Length
3	Positive charge
4	Gas phase basicity
5	Hydrophobicity index 3.0 pH
6	Fraction of site occupied by water
7	Hydrophobicity coefficient in RP-HPLC C18
8	Transfer energy organic solvent water
9	Partition coefficient
10	Energy transfer from out to in (95% buried)
11	Partition energy
12	Retention coefficient at pH 2
13	RF value in high salt chromatography
14	The stability scale from knowledge based atom atom potential
15	Apparent partition energies calculated from Chothia index
16	Average relative fractional occurence in E0i
17	Isoelectric point (Matlab bioinformatics toolbox)
18	Transfer free energy from vap to chx
19	Propensity of amino acids with pi helices
20	Propensity to be buried inside
21	Hydrophobicity parameters to predict bitter taste
22	Slopes proteins FDPB VFF neutral
23	Absolute entropy
24	Retention coefficient in NaH2PO4
25	Linker index
26	Relative preference value at C'
27	Normalized composition of mt proteins
28	Linker propensity index
29	Consensus normalized hydrophobicity scale
30	Transfer free energy from chx to oct
31	Flexibility parameter for two rigid neighbors
32	Refractivity
33	Normalized composition from animal
34	Transfer free energy to surface
35	Isoelectric point

🟥 Hydrophobicity
🟪 Energy
🟦 Structural
🟧 Charge
⬛ Other

Figure 4 Analysis of important physicochemical properties in predicting high-responding peptides. (**a**) The yeast training set was randomly split into training- (80%) and test- (20%) sets to produce 100 different Random Forest models (1,000 trees) at each step of halving the number of important properties. The box plot shows the test set error distribution. (**b**) The stability of property importance improves with increased number of trees in the Random Forest model. For a given number of trees, five models were built and the pairwise Spearman rank correlation coefficient of determination (R^2) was calculated for the ranked list of important features (error bars ± 1 s.d). (**c**) The top 35 features from the ESP predictor using 50,000 trees are listed.

Stability of the ranking of important physicochemical properties is highly dependent on the number of trees used in Random Forest. We built five Random Forest models using 100, 1,000, 10,000, 20,000 and 50,000 trees and analyzed the pair-wise Spearman rank correlation (ten correlations with five models) of the property ranking for each Random Forest model. Not surprisingly, the pair-wise correlation for a Random Forest with 100 trees indicates almost no correlation of the property rank between models ($R^2 = 0.06 \pm 0.02$, mean ± s.d.). However, the correlation continues to improve as we increase the number of trees (for 50,000 trees, $R^2 = 0.98 \pm 0.001$, mean ± s.d.). With 50,000 trees, the list of important physicochemical properties becomes more stable and reproducible (**Fig. 4b**). We observed no indication of overfitting with 50,000 trees, which is consistent with the behavior of Random Forest (**Supplementary Fig. 4** online).

DISCUSSION

The ESP predictor is more robust and performs significantly better than existing computational methods or random predictions across ten experimentally diverse validation sets. Based on our analyses, it provides a robust method to select candidate signature-peptides for MRM-MS protein quantification, especially in the absence of MS-based experimental data. When applied directly to MRM-MS–assay development for 14 proteins, our method achieved a success rate of 93%, and on average correctly selected two signature peptides per protein. In particular, we used the ESP predictor to successfully configure MRM-MS assays for six proteins in which MS discovery data were not found in a comprehensive proteomic database.

We showed, using two validation sets, that the ESP predictor performs significantly better than three previous methods designed to predict proteotypic peptides (**Fig. 2c,d**). We attribute the success of our method to the following two factors. First, a unique aspect of our study, relative to these prior studies, is the method used to determine the training set. Prior studies defined their training set based on peptides 'detected' or 'not detected' in an MS experiment. Our method focuses on predicting high-responding peptides. High-responding peptides are not only proteotypic (that is, detectable and unique) but constitute that subset of proteotypic peptides producing the highest MS response. Second, Random Forest is a committee of decision trees that vote on deciding a final classification, and each of these trees is based on random resampling in both feature and sample space. These characteristics of the Random Forest may be responsible for the ability of the model to generalize well beyond the training set (**Supplementary Figs. 1** and **5** online).

A major advantage of the ESP predictor is that a single model performs well across all common ESI experimental types. Unlike existing methods, which developed separate models for different ESI platforms[10] or even data set–specific models[11], we observe very consistent performance with a single model, indicating the model does not need to be retrained. We show this by testing the ESP predictor against validation sets from multiple database-search algorithms, quantification methods, mass spectrometers and experimental conditions. The ESP predictor would probably need to be retrained to be used on data produced using matrix-assisted laser desorption ionization (MALDI) MS, if a protease other than trypsin was used, or if other sample preparation procedures differ significantly from those used for the training set (e.g., not reducing and alkylating cysteines before digestion or using different LC solvent buffers).

Use of the Random Forest classifier provides insight into the model by calculating the most important physicochemical properties used to predict high-responding peptides. Because Random Forest is a non-linear model, it is not possible to determine the direction of response from each property (**Supplementary Fig. 6** online). It is difficult to compare relevant physicochemical properties (which are heavily influenced by the underlying experimental data) across previous studies because each study used different training sets, properties and computational models. For example, previously[13] cysteine was shown to be important in classifying a peptide as 'not proteotypic' because the sample was not alkylated, making cysteine-containing peptides unlikely to be detected by MS. To further illustrate this point, in our study, we applied two different feature-selection techniques and found minimal overlap with the top 35 properties reported by the final Random Forest model (**Supplementary Fig. 7** online). However, there is broad agreement that hydrophobicity, positive charge and energy terms are critical for predicting high-responding peptides and proteotypic peptides[10–13]. We grouped the top 35 properties into five categories: hydrophobicity, energy, structural, charge and other (**Fig. 4c**). In previous studies examining ESI response, it was observed that Gibbs free-energy transfer between amino acids has led to an increased response in peptides with nonpolar regions[27]. This supports our findings that hydrophobicity and energy properties influence peptide response. The structural properties may indicate likely cleavage sites during protein digestion, and we know peptides must carry a charge to be detected in a mass spectrometer[28]. It is worth mentioning that, although many of the properties appear similar in name (that is, hydrophobicity), often the amino acid values were determined under different experimental conditions. For example, a mathematical model has been developed[29] to calculate amino acid hydrophobicity based on HPLC performance of synthetic amino acids (rank 5 in **Fig. 4c**). On the other hand, a model of retention time (that is, hydrophobicity; rank 12 in **Fig. 4c**)[30] was developed based on HPLC performance using a synthetic 5-mer peptide in which individual amino acids were sequentially added in the middle. This suggests Random Forest is able to leverage subtle differences in amino acid property values to appropriately calculate peptide response.

In summary, we have shown that the ESP predictor is a robust method to predict high-responding peptides from a given protein based entirely on the peptide sequence. The ESP predictor greatly facilitates selection of optimal candidate signature-peptides for developing targeted assays to detect and quantify any protein of interest in the proteome. The ESP predictor fills a critical gap, enabling selection of candidate signature-peptides for proteins of interest in the absence of high-quality MS-based experimental evidence. Its use should improve the efficiency of biomarker verification, currently one of the most significant resource constraints in the development of biomarkers for early detection of disease, and the development of pharmacodynamic markers of therapeutic efficacy[1,31,32].

METHODS

Defining empiric peptide classification training set. The National Cancer Institute Clinical Proteomic Technology Assessment in Cancer Program (NCI-CPTAC) prepared a tryptic digest of a yeast lysate sample and sent it to three proteomic laboratories: Vanderbilt University, New York University (NYU) and the Broad Institute. All laboratories were expected to follow the same MS protocol on an LTQ-Orbitrap mass spectrometer. Vanderbilt analyzed the sample in duplicate on two instruments, NYU analyzed the sample in duplicate, and the Broad Institute performed six replicates. Thus, the yeast lysate was analyzed 12 times across four LTQ-Orbitraps. The raw files were searched using Spectrum Mill v3.4 beta with a precursor mass tolerance of 0.05 Da and fragment mass tolerance of 0.7 Da, specifying up to two missed cleavages and the following modifications: cysteine carbamidomethylation, carbamylation of N termini and lysine, oxidized methionine and pyroglutamic acid. The tandem MS (MS/MS) data were autovalidated at the protein level with a protein score of 25 and at the peptide level using a score of 13, percent similarity of 70%, forward-reverse score of 2, and rank 1-2 score difference of 2, for all charge states. In total, 4,230 peptides (570 proteins) were identified. The peptide identities, *m/z*, and retention time were exported to calculate the XIC for the monoisotopic peak.

The XIC for each peptide (in a given charge state) was calculated by determining the location (*m/z* and retention time) of the peptide peak. If a peptide was sequenced multiple times (that is, has many MS/MS spectra), the peptide with the best Spectrum Mill score on a per charge basis was used for this purpose. Peptides with the highest score indicate the highest confidence in matching the fragment spectra compared to spectra with lower scores for the same peptide.

In each LC-MS/MS run, different sets of peptides were sequence identified owing to the stochastic behavior of the mass spectrometer. Therefore, retention times were propagated across different LC-MS/MS runs using a quadratic regression model ($R^2 = 0.99$ for all LC-MS/MS runs). This yielded an approximate elution time, and allowed us to 'hunt' for peptides not sequence identified in a particular LC-MS/MS run. The XIC was calculated using a combination of retention time and *m/z* for each peptide.

An in-house program was developed to automatically calculate the XIC using the Thermo Software Development Kit. The XIC was calculated using a retention time tolerance of ± 2.5 min and *m/z* tolerance of ± 15 p.p.m. A summary table was created where the response for each peptide was obtained by summing the XIC values for all peptide variations (that is, peptides with multiple charge states and common modifications). This reduced the list to 3,637 peptides.

The yeast LC-MS/MS runs from each institute (Vanderbilt, NYU, Broad Institute) were then median normalized to account for any instrument or processing differences (which were expected to be minor because all samples were processed following the same protocol). The median normalization divides each LC-MS/MS run by its median XIC value and then multiplies it by the common median XIC (the median of the median of all 12 LC-MS/MS runs). A table of identified peptides was created, with their corresponding XIC (if present) in all 12 LC-MS/MS runs. The median of all 12 LC-MS/MS runs was selected as the 'official' XIC value for each peptide. Peptides with a coefficient of variance (s.d./mean×100%) >100% were rejected. In addition, any peptide with a median XIC of zero was rejected, indicating that it was not reliably detected in all LC-MS/MS runs.

Next, a set of peptides 'not detected' in the mass spectrometer was created. An *in silico* tryptic digest was performed for all sequence-identified proteins. A substring search was used to remove any *in silico* peptide where we had evidence of a sequence-identified peptide. For example, if the *in silico* peptide was LQTISALPK and the sequence-identified peptide was LQTISALPKGDELR, the *in silico* peptide was rejected because it is a substring of the sequence-identified peptide. Thus, the 'not detected' set of peptides was not seen in any form of the sequence-identified peptides. In addition, any peptide sequence that was not unique and any N- or C-terminal peptides (~4% of the peptides) were removed. The final peptide set contained a list of sequence-identified peptides (with their corresponding XIC) and peptides that were not sequence identified in any form.

To classify peptides as high- or low-responding, we considered only proteins with seven or more sequence-identified peptides. The peptide response within each protein was log transformed (excluding peptides 'not detected') to create a normal distribution and is justified by the Box Cox transformation[33]. The log-transformed data were then standardized, using the *z*-score (*z*), within each protein. High-responding peptides were selected with a $z \geq 0$ whereas low-responding peptides were selected with a $z \leq -1$. This procedure was used only to create the training set and does not apply to the validation sets, where we examined only the five highest-responding peptides. The 'not detected' peptides were then appended to the low-responding peptides to create a binary high ($n = 623$) versus low/not detected ($n = 2{,}530$) classifier.

Calculation of physicochemical properties for peptides. A diverse set of 550 physicochemical properties was used to calculate the peptide feature set. Properties such as length, number of acidic (glutamic acid, asparagine) and basic (arginine, lysine, histidine) residues were calculated by counting the number of amino acids in each peptide. The Bioinformatics package in Matlab was used to calculate the peptide mass and pI. The gas phase basicity was calculated from Zhang's model[17]. The remaining 544 physicochemical properties contained individual values for each amino acid. For each peptide and a given property, the constituent amino acid numerical values were averaged to produce a single value. Missing values were ignored. The average (rather than median or sum) was chosen because it is sensitive to outliers and normalizes for peptide length. It was assumed that the average physicochemical property across each peptide was sufficient to capture relevant information about peptide response. The model does not incorporate protein context such as flanking amino acids or protein information (e.g., protein molecular weight or protein pI). We view this as a separate problem from predicting high-responding peptides[34,35]. Calculations of the peptide feature set were performed in Matlab R2006b (MathWorks).

Random Forest classifier for predicting high-responding peptides. Random Forest is a nonlinear ensemble algorithm composed of many individual decision trees. Each tree is grown using a randomized tree-building algorithm. For each tree (*num_tree*), a bootstrap sample (that is, random data subset sampled with replacement) is selected from the training set. At each decision branch in the tree, the best spilt is chosen from a randomly selected subset of properties (rather than all properties), *num_feature*. With these two random steps each tree is different. Predictions result from the ensemble of all trees by taking the majority vote. Instead of relying on this binary classification, a probabilistic output (the fraction of trees that vote high) was used and referred to as probability of response.

The peptide training data were imbalanced. High-responding peptides, the class of interest, comprised only ~20% of the data. Most classifiers focus on optimizing overall accuracy at the expense of misclassifying the minority class (high-responding peptides). Down sampling is a common technique to handle imbalanced data sets[36]. In Random Forest, the number of training samples for each class was set to the size of the minority class ($n = 623$), and samples were selected via bootstrapping with replacement from both the minority and majority classes. This process was repeated for each tree and exhibits a significant improvement in performance and generalization[36].

Balanced class sizes were used to optimize *num_tree* and *num_feature* parameters in Random Forest. The *num_feature* parameter was optimized by setting *num_tree* to 1,000 and varying *num_feature* between 2 and 550 features. The optimal value for *num_feature* was determined to be 90 (**Supplementary Fig. 8** online). The *num_tree* parameter was optimized by increasing the number of trees until the variable importance measure was consistent and reproducible (**Fig. 4b**). The *num_tree* parameter was set to 50,000 trees.

The training data were used to calculate a no call region in order to judge the model performance on peptides confidently classified as either high or low/not detected. Peptides with a predicted probability of response between 0.38–0.65 were labeled as no call and the model was not penalized. Peptides with a predicted probability greater than or equal to 0.65 were classified as high and peptides with a predicted probability less than or equal to 0.38 were classified as low/not detected. The reject region was selected based on a false positive rate (1 – specificity) of 10%. This choice of reject region yielded calls on 74% of the training data.

The weighted accuracy was used to account for the imbalanced class size. The weighted accuracy is calculated as: $A_w = 0.5 * (\text{sensitivity} + \text{specificity})$ where sensitivity is the percent of true positives and specificity is the percent of true negatives. The yeast training data were split into training (90%) and test (10%) sets. The training and test set weighted accuracies were 81% and 76%, respectively. We also examined the area under the curve (AUC) for a receiver operating characteristic (ROC) plot[24] on the test data. The AUC is a standard measure of performance where a perfect classification would have an AUC of 1 and random classification would have an AUC of 0.5. The AUC for the test set was 83% ($P = 9.4\text{e-}9$) indicating the predictions are significantly better than random (**Supplementary Fig. 9** online). Random Forest and ROC calculations were performed in R (http://CRAN.R-project.org/) using the Random Forest package v. 4.5-18 (ref. 19) and ROCR library v. 1.0-2 (ref. 37), respectively.

Random Forest variable importance score. A measure of how each property contributes to the overall model performance is determined during Random Forest training. When the values for an important property are permuted there should be a noticeable decrease in model accuracy. Likewise, when the values for an irrelevant property are permuted there should be little change in model accuracy. The difference in the two accuracies are then averaged for all trees and normalized by the standard error to produce an importance measure, referred to as the variable importance score.

Permutation test to evaluate the significance of the ESP predictions. All proteins were required to contain at least six or more predicted tryptic peptides (from an *in silico* digest) and at least five or more sequence-identified peptides. For each protein, the five highest-responding peptides were selected (based on experimental data, **Fig. 1a** down to 'MRM-MS assay optimization'). Then, using the same protein, five peptides with the highest probability of response were selected using the ESP predictor (**Fig. 1b**). For each validation set, the actual test statistic (Ts) was calculated as the sum of the number of peptides in common between the top five peptides from the experimental and computational methods for each protein. Next, a random test statistic (Trs) was calculated by randomly sampling five peptides and taking the sum of the number of peptides in common with the top five experimentally derived peptides for each protein in the validation set. This process was repeated 10,000 times to produce a null distribution for each validation set. The resulting distribution was used to estimate a one-tailed *P*-value. Using this procedure, the statistical significance of the predictions made by the ESP predictor was calculated as the number of proteins (also selected at random from the respective validation set) increased. The permutation test implicitly accounts for differences in the number of peptides from each protein. The permutation test calculations were performed in R.

Analysis and MS summary for all validation sets. All protein mixtures were digested using trypsin and analyzed using reversed-phased nano LC-ESI-MS/MS on multiple LTQ Oribtrap and LTQ-FT mass spectrometers (Thermo). Specific conditions concerning chromatography, buffers, injection volume and MS analysis settings varied according to each validation set (full details for all validation sets are provided in the **Supplementary Methods**). Validation sets were subsequently processed using either Spectrum Mill 3.4 beta (Agilent Technologies) or Mascot v. 2.1.0.3 (Matrix Science) to determine sequence-identified peptides from the collected MS/MS spectra. Peptide response was calculated using either an in-house developed program to calculate the XIC, MSQuant v. 1.4.2 b5 (http://msquant.sourceforge.net/), or Spectrum Mill. The total peptide response was calculated by summing all forms of a given peptide (that is, multiple charge states and the following modifications: carbamidomethylation, carbamylated lysine, oxidized methionine and pyroglutamic acid). The following is a brief summary of each validation set:

ISB-18 is a publicly available standard protein mix consisting of 18 proteins provided by the Institute for Systems Biology (ISB)[38]. Only the LTQ-FT data were considered.

Yeast test refers to the 10% of proteins held-out from the training set in order to evaluate the model performance.

Plasma refers to neat plasma (that is, undepleted plasma).

Sigma48 refers to a set of 48 equimolar proteins (Universal Proteomics Standard Set, Sigma). The samples were digested using a trifluoroethanol-assisted digestion protocol[39].

Plasma Hu14 SCX refers to a plasma sample with the 14 most abundant proteins removed using a MARS Hu-14 column (Agilent Technologies) and then fractionated using strong cation exchange (SCX). Eleven fractions were collected and analyzed.

Yeast_2 refers to a separate independent analysis of a yeast mixture. Importantly, proteins in common with the yeast training set were removed.

HeLa_1 refers to HeLaS3 cell lysate digested in-solution.

HeLa_2 refers to HeLaS3 cell lysate analyzed by GeLC-MS (**Supplementary Methods**).

Pull-Down refers to a GeLC-MS affinity pull-down experiment from a HeLaS3 cell lysate.

Plasma Hu14 refers to a plasma sample with the 14 most abundant proteins removed using a MARS Hu-14 column.

MRM-MS assay development. The validated MRM peptides were defined from single protein digests for each of the 14 proteins. Peptide selection for the 14 target proteins was based upon experimental observation using commercially available protein standards. Briefly, the proteins were individually digested with trypsin and analyzed by nano LC-MS/MS in positive-ion electrospray on an LTQ linear ion trap mass spectrometer (Thermo) with data-dependent acquisition. Peptide-sequence identity was determined using Spectrum Mill on the collected MS/MS spectra. Approximately five candidate peptide standards per protein were chosen based primarily on high relative response. Exclusion criteria included large hydrophobic or small hydrophilic peptides, flanking tryptic ends with dibasic amino acids (KK, RR, KR, RK) at the N or C terminus and peptide identity corresponding to multiple endogenous plasma proteins. Peptide standards containing methionine and cysteine were avoided if possible. Stable isotope–labeled versions of each candidate peptide were synthesized for quantification and MRM response curves were optimized in plasma for each protein over a wide concentration range. All peptides that performed satisfactorily over the response curves are referred to as "validated MRM peptides."

All MRM experiments were performed on a 4000 Q Trap Hybrid triple quadrupole/linear ion trap mass spectrometer coupled to a Tempo LC system (Applied Biosystems). Data analysis was done using MultiQuant software (Applied Biosystems).

The GPM database was searched (December 19, 2008) by entering the protein Ensembl accession number and then selecting the 'MRM' link. For some proteins, a large number of peptides were listed. Only the top five peptides were considered based on the number of times observed in the GPM database.

Data and software availability. The yeast MS data used to develop the model are publicly available from Tranche (http://tranche.proteomecommons.org/). The ESP predictor is freely available as a module in the GenePattern integrative genomics software package (http://www.genepattern.org/) under the category 'proteomics'. The automated script to calculate the XIC using the Thermo Software Development Kit is available upon request. Source code and examples are available as **Supplementary Source Code** online. The data associated with this manuscript may be downloaded from the ProteomeCommons.org Tranche system 〈http://www.proteomecommons.org/data-downloader.jsp?fileName=90MaGKV4KHKHOyOvNGSXxtDhAEQbJA3KbZap6ruHxvUFDk%2BvOFyhawX%2BhSQa%2Bxa/KvG6oQCYON4nsZ/uDw55FfNDAU0AAAAAAAAMLw==〉.

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

ACKNOWLEDGMENTS

We thank the National Cancer Institute (NCI) Clinical Proteomic Technology Assessment in Cancer Program (NCI-CPTAC, http://proteomics.cancer.gov/programs/CPTAC/) for providing samples of yeast lysate and raw MS data generated by the CPTAC centers. We thank Rushdy Ahmad, Kathy Do, Amy Ham, Emily Rudomin, and Shao-En Ong for MS data generation, and Hasmik Keshishian and Terri Addona for generating the lists of validated MRM peptides. We also thank Shao-En Ong, Jacob Jaffe, Karl Clauser, Eric Kuhn, Pablo Tamayo, and Nick Patterson for helpful discussions. We would like to thank the reviewers for their insightful comments. This work was supported in part by grants to S.A.C. from the National Institutes of Health Grants 1U24 CA126476 as part of the NCI's Clinical Proteomic Technologies Assessment in Cancer Program, the National Heart, Lung, and Blood Institute, U01-HL081341 and The Women's Cancer Research Fund; to J.P.M. from the National Science Foundation and NIGMS the National Institutes of Health (NIGMS and NCI); to D.R.M. from the National Institutes of Health grant R01 CA126219, as part of NCI's Clinical Proteomic Technologies for Cancer Program.

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

1. Rifai, N., Gillette, M.A. & Carr, S.A. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat. Biotechnol.* **24**, 971–983 (2006).
2. Uhlen, M. & Hober, S. Generation and validation of affinity reagents on a proteome-wide level. *J. Mol. Recognit.* (2008).
3. Anderson, L. & Hunter, C.L. Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Mol. Cell. Proteomics* **5**, 573–588 (2006).
4. Gerber, S.A., Rush, J., Stemman, O., Kirschner, M.W. & Gygi, S.P. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci. USA* **100**, 6940–6945 (2003).
5. Keshishian, H., Addona, T., Burgess, M., Kuhn, E. & Carr, S.A. Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution. *Mol. Cell. Proteomics* **6**, 2212–2229 (2007).
6. Stahl-Zeng, J. *et al.* High sensitivity detection of plasma proteins by multiple reaction monitoring of N-glycosites. *Mol. Cell. Proteomics* **6**, 1809–1817 (2007).
7. Kuster, B., Schirle, M., Mallick, P. & Aebersold, R. Scoring proteomes with proteotypic peptide probes. *Nat. Rev. Mol. Cell Biol.* **6**, 577–583 (2005).
8. Craig, R., Cortens, J.P. & Beavis, R.C. Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* **3**, 1234–1242 (2004).
9. Deutsch, E.W., Lam, H. & Aebersold, R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EBMO reports* **9**, 429–434 (2008).
10. Mallick, P. *et al.* Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* **25**, 125–131 (2007).
11. Sanders, W.S., Bridges, S.M., McCarthy, F.M., Nanduri, B. & Burgess, S.C. Prediction of peptides observable by mass spectrometry applied at the experimental set level. *BMC Bioinformatics* **8** Suppl 7, S23 (2007).
12. Tang, H. *et al.* A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics* **22**, e481–e488 (2006).
13. Webb-Robertson, B.J. *et al.* A support vector machine model for the prediction of proteotypic peptides for accurate mass and time proteomics. *Bioinformatics* **24**, 1503–1509 (2008).
14. Jaffe, J.D. *et al.* Accurate inclusion mass screening: a bridge from unbiased discovery to targeted assay development for biomarker verification. *Mol. Cell. Proteomics* **7**, 1952–1962 (2008).
15. Malmstrom, J., Lee, H. & Aebersold, R. Advances in proteomic workflows for systems biology. *Curr. Opin. Biotechnol.* **18**, 378–384 (2007).
16. Kawashima, S. & Kanehisa, M. AAindex: amino acid index database. *Nucleic Acids Res.* **28**, 374 (2000).
17. Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **76**, 3908–3922 (2004).
18. Breiman, L. Random forest. *Mach. Learn.* **45**, 5–32 (2001).
19. Liaw, A. & Wiener, M. ClassificatIon and Regression by randomForest. *R News* **2**, 18–22 (2002).
20. Diaz-Uriarte, R. & Alvarez de Andres, S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**, 3 (2006).
21. Enot, D.P., Beckmann, M., Overy, D. & Draper, J. Predicting interpretability of metabolome models based on behavior, putative identity, and biological relevance of explanatory signals. *Proc. Natl. Acad. Sci. USA* **103**, 14865–14870 (2006).
22. Vapnik, V.. *The Nature of Statistical Learning Theory* (Springer, New York, 1995).
23. Bishop, C. *Neural Networks for Pattern Recognition* (Oxford University Press, Oxford, 1995).
24. Fawcett, T. *ROC Graphs: Notes and Practical Considerations for Researchers* (Technical report, HP Laboratories, Palo Alto, CA, USA, 2004).
25. Lange, V., Picotti, P., Domon, B. & Aebersold, R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* **4**, 222 (2008).
26. Svetnik, V. *et al.* Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **43**, 1947–1958 (2003).
27. Cech, N.B. & Enke, C.G. Relating electrospray ionization response to nonpolar character of small peptides. *Anal. Chem.* **72**, 2717–2723 (2000).
28. Cech, N.B. & Enke, C.G. Practical implications of some recent studies in electrospray ionization fundamentals. *Mass Spectrom. Rev.* **20**, 362–387 (2001).
29. Cowan, R. & Whittaker, R.G. Hydrophobicity indices for amino acid residues as determined by high-performance liquid chromatography. *Pept. Res.* **3**, 75–80 (1990).
30. Parker, J.M., Guo, D. & Hodges, R.S. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. *Biochemistry* **25**, 5425–5432 (1986).
31. Whiteaker, J.R. *et al.* Integrated pipeline for mass spectrometry-based discovery and confirmation of biomarkers demonstrated in a mouse model of breast cancer. *J. Proteome Res.* **6**, 3962–3975 (2007).
32. Zolg, J.W. & Langen, H. How industry is approaching the search for new diagnostic markers and biomarkers. *Mol. Cell. Proteomics* **3**, 345–354 (2004).
33. Sokal, R.R. & Rohlf, F.J. *Biometry the Principles and Practice of Statistics in Biological Research*, edn. 3 (W.H. Freeman and Company, 1995).
34. Thomson, R., Hodgman, T.C., Yang, Z.R. & Doyle, A.K. Characterizing proteolytic cleavage site activity using bio-basis function neural networks. *Bioinformatics* **19**, 1741–1747 (2003).
35. Yen, C.Y. *et al.* Improving sensitivity in shotgun proteomics using a peptide-centric database with reduced complexity: protease cleavage and SCX elution rules from data mining of MS/MS spectra. *Anal. Chem.* **78**, 1071–1084 (2006).
36. Chen, C., Liaw, A & Breiman, L. *Using Random Forest to Learn Imbalanced Data* (Technical Report 666. Statistics Department of University of California at Berkeley, Berkeley, 2004).
37. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940–3941 (2005).
38. 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).
39. Wang, H. *et al.* Development and evaluation of a micro- and nanoscale proteomic sample preparation method. *J. Proteome Res.* **4**, 2397–2403 (2005).

nature biotechnology

Dynamic modularity in protein interaction networks predicts breast cancer outcome

Ian W Taylor[1,2], Rune Linding[1,3], David Warde-Farley[4,5], Yongmei Liu[1], Catia Pesquita[6], Daniel Faria[6], Shelley Bull[1,7], Tony Pawson[1,2], Quaid Morris[4,5] & Jeffrey L Wrana[1,2]

Changes in the biochemical wiring of oncogenic cells drives phenotypic transformations that directly affect disease outcome. Here we examine the dynamic structure of the human protein interaction network (interactome) to determine whether changes in the organization of the interactome can be used to predict patient outcome. An analysis of hub proteins identified intermodular hub proteins that are co-expressed with their interacting partners in a tissue-restricted manner and intramodular hub proteins that are co-expressed with their interacting partners in all or most tissues. Substantial differences in biochemical structure were observed between the two types of hubs. Signaling domains were found more often in intermodular hub proteins, which were also more frequently associated with oncogenesis. Analysis of two breast cancer patient cohorts revealed that altered modularity of the human interactome may be useful as an indicator of breast cancer prognosis.

Transcriptome analyses have been extensively applied as molecular diagnostic and prognostic tools in breast cancer. Recently, the prognostic predictive performance of gene expression signatures has been improved by incorporating interactome data[1], suggesting that altered gene expression in breast cancer might disturb the higher-level organization of the interactome and affect disease outcome.

To investigate this possibility, we first identified proteins that have many interacting partners (so called 'hubs') in a network of protein-protein interactions curated from the literature and high-throughput sources[2] (**Supplementary Fig. 1a** online). Next, we obtained genome-wide expression data measured in 79 human tissues[3], and quantified the extent to which a hub and its interacting partners were co-expressed in the same tissues (**Supplementary Methods** online). We used the average Pearson correlation coefficient (PCC) of co-expression of a hub protein and its partners to identify whether interactions are context specific (that is, interacting proteins are not always co-expressed) or constitutive (that is, interacting proteins are always co-expressed). This revealed a multi-modal distribution that appeared to be the superposition of distinct populations of hubs centered over increasing average PCC values (**Fig. 1a**, red asterisks). Randomly reassigning the expression data to different gene products in the same network resulted in an approximately normal distribution of PCC values (**Fig. 1a**, black dashed line). The shoulder (marked with a black asterisk) is largely due to strongly correlated gene products that have a high probability of reforming interactions with their true interactors when randomized (data not shown). We observed a similar multi-modal distribution using a literature-curated source alone[4] (**Supplementary Fig. 1b**) or a different high-confidence human PPI database[5] (**Supplementary Fig. 1c**).

The human interactome thus has two classes of hubs. One class displays low correlation of co-expression with its partners. We call these hubs intermodular hubs, as first proposed for the yeast interactome[6,7]. A second class, termed intramodular hubs, displays more highly correlated patterns of co-expression (**Fig. 1a**). These features reflect a modular architecture. Restricting the analysis to interactions conserved between yeast and humans revealed a single peak at high average PCC, suggestive of largely intramodular hubs (**Fig. 1b**). Previous analyses showed that the assembly of intramodular hubs into macromolecular complexes constrains intramodular hub evolution[6]. This is visualized as a cluster of highly correlated interactions interconnecting intramodular hubs in the human interactome (**Supplementary Fig. 1a**; green edges between blue nodes).

Modular structure can confer higher-order function to interactomes, such that intermodular hubs provide temporally and spatially restricted linkages to intramodular hubs that in turn fulfill specific functions, often as multi-subunit macromolecular machines[8,9]. For example, most components of the 26S proteasome show highly correlated expression and function together to mediate protein degradation (**Supplementary Fig. 2a** online). However, three hub components (PSMB1, PSMB2 and PSMD9) are intermodular, reflecting tissue-specific modulation of the proteasome[10,11]. Using the Gene Ontology (GO) molecular function database[12], we found that intramodular hubs shared more functional similarity with their partners than did intermodular hubs (Student's *t*-test, $P < 0.02$, **Supplementary Fig. 2b**).

[1]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, 600 University Ave., Toronto, Ontario M5G 1X5, Canada. [2]Department of Molecular Genetics, University of Toronto, 1 King's College Circle, Room 4396, Toronto, Ontario M5S 1A8, Canada. [3]Cellular & Molecular Logic Team, Institute of Cancer Research (ICR), Section of Cell, Molecular & Systems Section, 237 Fulham Road, London, SW3 6JB, UK. [4]The Terrence Donnelly Centre for Cellular and Biomolecular Research, 160 College St., Toronto, Ontario M5S 3E1, Canada. [5]Department of Computer Science, University of Toronto, 10 King's College Road, Room 3303, Toronto, Ontario M5S 3G4, Canada. [6]Faculty of Sciences, University of Lisbon, Campo Grande, 1749-016, Lisbon, Portugal. [7]Dalla Lana, School of Public Health, University of Toronto, 155 College St., Toronto, ON M5T 3M7, Canada. Correspondence should be addressed to J.W. (wrana@lunenfeld.ca).

Received 25 September 2008; accepted 18 December 2008; published online 1 February 2009; doi:10.1038/nbt.1522

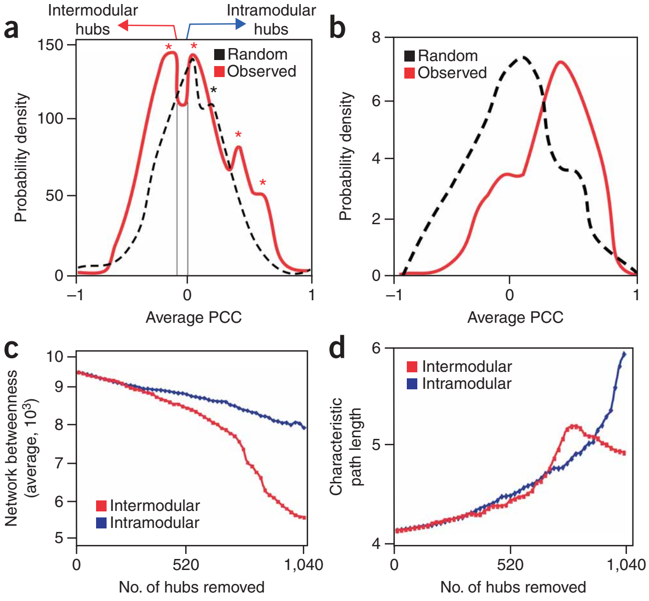


Figure 1 Evidence of dynamic network modularity in the human interactome. (**a**) The probability density of the average PCC of co-expression for human hub proteins with their interactors across 79 human tissues (red line) is compared to randomized data (dashed black line). (**b**) Same as (**a**) but only using human hub proteins conserved in yeast (red line) compared to randomized data (dashed black line). (**c**) Network betweenness as a function of removing equivalent numbers of intermodular or intramodular hubs. (**d**) Characteristic path length of the network as a function of removing equivalent numbers of intermodular or intramodular hubs.

Intermodular hubs have been proposed to be critical for global network connectivity[7]. We tested this by systematically removing either intermodular or intramodular hubs from the interaction network and analyzing the number of paths between nodes using a topological measure known as ‘betweenness’[13]. Betweenness measures information flow through networks, with high betweenness reflecting multiple paths between nodes and low betweenness few paths. In a biological context, betweenness measures the ways in which signals can pass through the interaction network. Betweenness was more strongly affected by removing inter- rather than intramodular hubs (**Fig. 1c**). Another topological measure of global network connectivity is the characteristic path length (CPL), which is the average of the shortest path between all nodes in a network[14]. Systematic removal of intermodular hubs increased CPL to a threshold beyond which CPL rapidly collapsed due to splintering of the large network into small subnetworks (**Fig. 1d**). In contrast, intramodular hub removal only increased CPL. The greater sensitivity of both betweenness and CPL to removal of intermodular hubs is consistent with the notion that the human interactome is modular with intermodular hubs connecting functional modules that are comprised of intramodular hubs.

Next, we asked whether hub types display characteristic biochemical features. We found that intermodular hubs were larger than intramodular hub proteins (Mann-Whitney U-test, $P < 0.005$, **Supplementary Fig. 3a** online). Analysis of domain numbers (modularity) and size (globularity) revealed intermodular hubs have more domains compared to a randomized distribution, whereas intramodular hubs have fewer domains than expected by chance ($P < 0.05$ and $P < 0.01$ respectively, **Fig. 2a**). Conversely, intramodular hubs have greater globularity (domain size) and intermodular hubs less ($P < 0.05$ and $P < 0.01$, respectively, **Fig. 2b**). Linear motifs (that is, post-translational modifications and short binding motifs[15]) are over- and underrepresented in intermodular and intramodular hubs, respectively ($P < 0.005$, **Fig. 2c**; **Supplementary Fig. 3b**).

We then explored domain types in the different hub classes. Cell signaling domains (as defined by the SMART database[16]) were enriched in intermodular hubs (sign test, $P < 0.001$), whereas

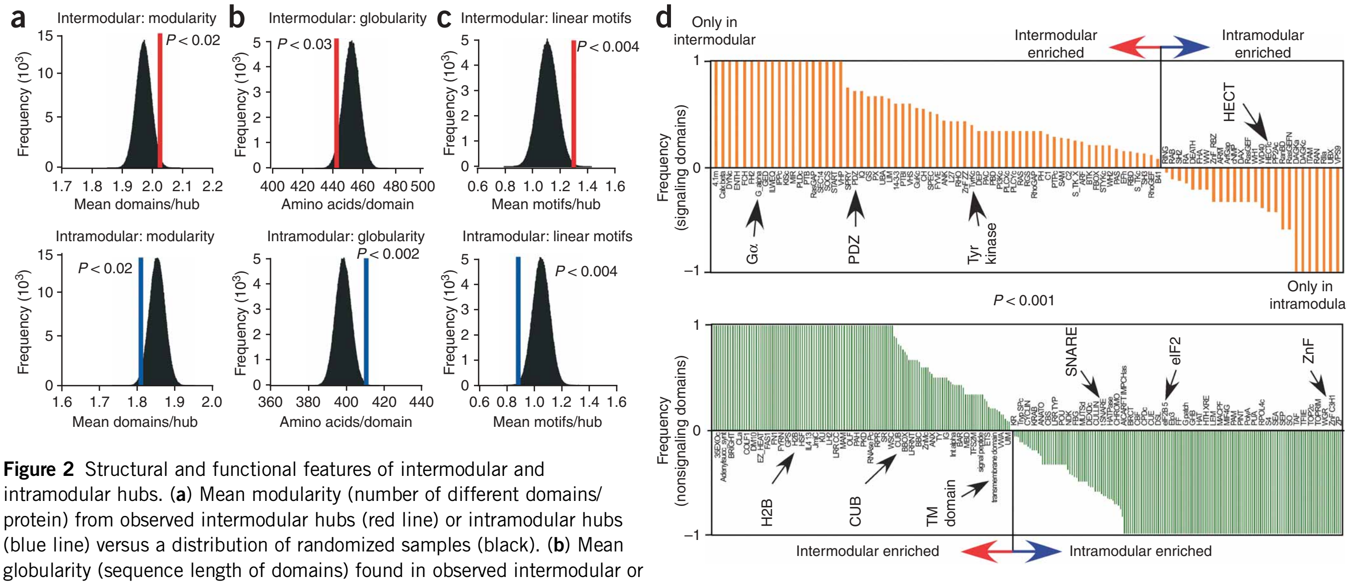


Figure 2 Structural and functional features of intermodular and intramodular hubs. (**a**) Mean modularity (number of different domains/ protein) from observed intermodular hubs (red line) or intramodular hubs (blue line) versus a distribution of randomized samples (black). (**b**) Mean globularity (sequence length of domains) found in observed intermodular or intramodular hubs compared to randomized distributions. (**c**) Mean number of experimentally validated linear motifs and phosphosites from the ELM and Phospho-ELM database in intermodular or intramodular hubs compared to randomized distributions. (**d**) Domain distribution between intermodular hubs and intramodular hubs. The frequency of individual domains in intermodular hubs minus their frequency in intramodular hubs was plotted for each of the signaling domains (top panel, orange bars) or non-signaling domains (bottom panel, green bars), as indicated**.** A frequency of 1 indicates domains are found exclusively in intermodular hubs, whereas a frequency of –1 indicates exclusively intramodular hubs. Note that to retain legibility only a fraction of nonsignaling domains are labeled.

nonsignaling domains were evenly distributed between the hub types (**Fig. 2d**). For example, tyrosine kinase, PDZ and Gα domains were found predominantly or exclusively in intermodular hubs (**Fig. 2d**). The two hub types have similar degree distributions (that is, number of interactions per hub; **Supplementary Fig. 4** online), indicating that the biochemical attributes of hub proteins are an inherent property of the hub type and are not a function of the number of interacting partners. Taken together, these results indicate that intra- and intermodular hubs display distinctive structural characteristics consistent with their roles in organizing communication and function of dynamic protein networks.

To explore this in detail we examined the well-characterized RAS subnetwork. RAS behaves as an intramodular hub, with many highly correlated regulatory partners, such as RALGDS and SOS (**Supplementary Fig 5a** online). In contrast, partners that employ RAS as an effector (that is, Insulin receptor adaptor protein, IRS1 (ref. 17)) or a regulator (that is, BRAF[17]) tended to be intermodular. The latter is connected to a large cluster of intramodular transcription factors, such as NFκB and p53. Also notable is that connections between the RAS module and the downstream intramodular cluster occur almost exclusively via intermodular hubs. This suggests a modular assembly of signaling networks with intermodular hubs organizing the interconnectivity of functional modules such as RAS and the downstream RAS transcriptional effectors.

During tumor progression, rewiring of signaling networks drives phenotypic alterations while maintaining the robustness of the network[8], suggesting that there may be differences in hub-type association with cancer. We queried Online Mendelian Inheritance in Man (OMIM)[18], the census of cancer genes[19], and oncogenic translocations and found that mutations of intermodular hubs were associated with cancer phenotypes more frequently than those of intramodular hubs (Fisher's exact test, $P < 0.05$, **Supplementary Figs. 5b**,**c** and **6** online). As intermodular hubs regulate the global functions of modular networks, these results suggest that alterations in network modularity may occur in cancer.

To investigate this we analyzed a well-described cohort of sporadic, nonfamilial breast cancer patients[20]. We first looked for significant differences in the average PCC of hub proteins and their interacting partners in patients who were disease free after extended follow-up (hereafter referred to as 'good outcome') and those who died of disease ('poor outcome') (**Supplementary Fig. 7** online). This revealed 256 hubs that displayed altered PCC as a function of disease outcome. One such hub was BRCA1, a protein that is mutated in a subset of familial breast cancers. The expression of BRCA1 was strongly correlated with the expression of its partners in tumors from surviving patients, but not well correlated with their expression in tumors from poor-outcome patients (**Fig. 3a**). In contrast, the transcription factor Sp1, which shares some interacting partners with BRCA1, was not significantly changed. Of the BRCA1 partners highly correlated in good outcome tumors, both MRE11 and BRCA2 were notable as they are members of the BRCA1-associated genome surveillance complex (BASC) and are misregulated in poor prognosis breast cancer[21,22]. Our results suggest that disorganization of the BASC by loss of coordinated co-expression of components is associated with poor outcome.

Analysis of interactions between the 256 hub proteins revealed that they form an interconnected network (**Fig. 3b**). Notably, we did not identify hubs that were themselves significantly up or downregulated in the good versus poor outcome groups, but rather we identified hubs that had altered PCC of expression between outcome groups (**Supplementary Fig. 7**). Of the 256 hubs identified in our study, only 23% (59 hubs) showed significantly altered expression in our cohort when analyzed using 'significance analysis of microarrays'[23].

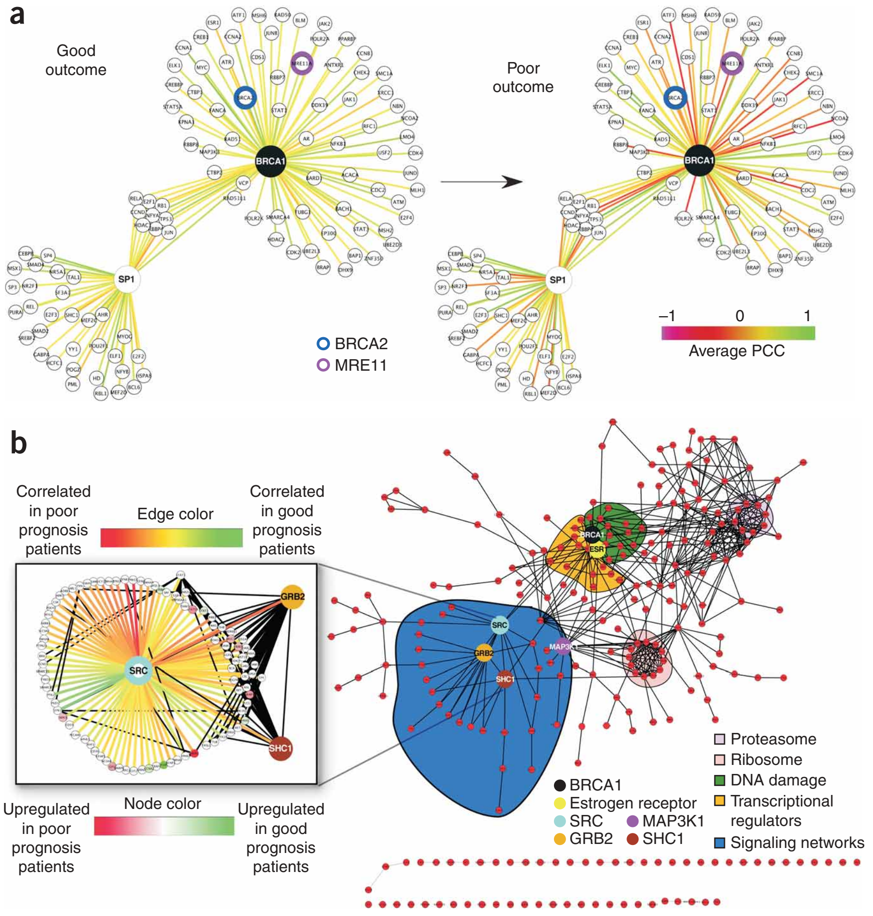


Figure 3 Differences in dynamic network properties in breast cancer tumors. (**a**) Network of the interacting partners of BRACA1 and SP1. BRCA1 and its interactors (e.g., BRCA2 and MRE11, as indicated) are highly ordered (green edges indicate correlated expression between protein pairs) in the surviving patients, whereas that organization is lost in patients who die of disease. Interactions involving Sp1 are not significantly altered. (**b**) Shown are all hubs (red nodes) that have, as a function of patient outcome, significantly different correlation of co-expression with their partners. Black edges connect hubs that have direct protein-protein interactions. Note that most hubs are components of a an interconnected network. The network includes many functional groups known to be misregulated in breast cancer pathogenesis (highlighted in legend). Inset shows a subnetwork focused on SRC and its interactors together with GRB2 and SHC1. Edge colors represent the correlation between SRC and each of its partners, while node colors represent changes in gene expression between outcome groups. Black edges indicate interactions not involving SRC. Note that while SRC is not significantly differently expressed between patient groups, it is a significant predictor hub because of differences in the coordinated co-expression of SRC and many of its partners.

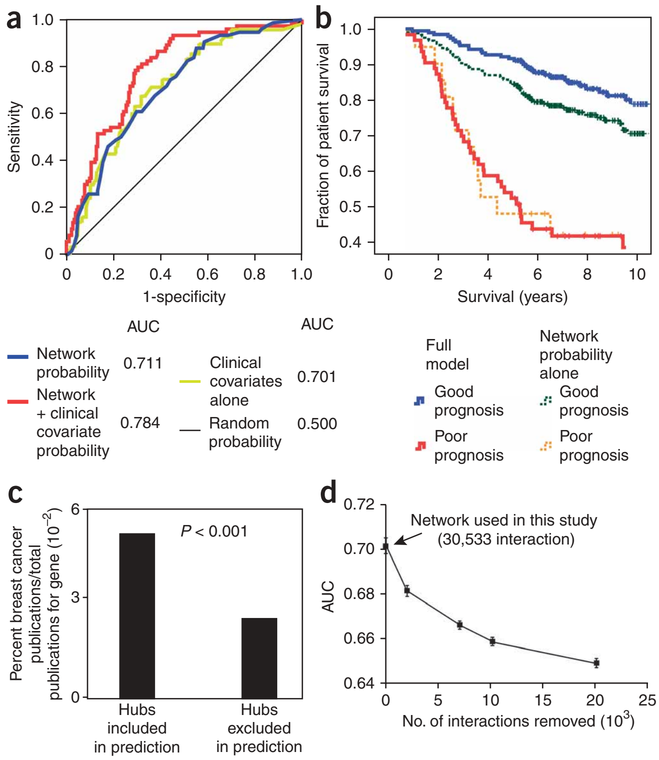


Figure 4 Dynamic network properties predict breast cancer outcome. (**a**) ROC curve of the probabilities for prognostic group membership from the affinity propagation clustering of patient dynamic network properties using fivefold cross-validation runs. Outcome prediction performances are shown for network probabilities alone (blue line), TNM tumor classifications alone (yellow line) and combining network properties of each tumor and TNM tumor classifications (red line). Random division of patients is shown with the black diagonal). (**b**) Kaplan-Meier disease-free survival curves. Patients were grouped into good and poor prognostic groups based on a fivefold cross-validation analysis of patient data. Patient survival is plotted for network probability alone (green and orange lines, as indicated) or network probability controlling for clinical covariates (red and blue lines). (**c**) Genes encoding hub proteins that are included in the prediction algorithm are cited significantly more frequently in the breast cancer literature than excluded hubs. (**d**) Algorithm performance declines as a function of decreasing interactome size. Interactions were randomly removed from the current interactome as indicated and performance of the dynamic network modularity algorithm assessed. Average AUC (+s.d.) at each of the reduced interactome sizes is plotted (black squares) and was calculated from 5-fold cross-validation runs performed in triplicate.

For example, no significant difference in the expression level of the oncogene product SRC was observed between groups (**Fig. 3b**, inset); however, the coordinated co-expression of SRC and its regulators or effectors (see inset **Fig. 3b**) was clearly affected. Unbiased analysis of the 256 hubs in this aberrant network demonstrated over-representation in literature (**Fig. 4**, Fisher's exact test, $P<0.001$) and microarray studies[20,24–26] of breast cancer (**Supplementary Fig. 8** online, Fisher's exact test, $P<0.02$) when compared to a similar network that did not change significantly between groups. These hubs include signaling proteins (MAP3K1, GRB2, SHC and SRC), an estrogen receptor (ESR1) and DNA damage response proteins (BRCA1, RAD51, MRE11). Single-nucleotide polymorphisms in MAP3K1 are associated with breast cancer susceptibility[27]. Thus, there are changes in dynamic network modularity that are associated with poor outcome in breast cancer, and these may provide a prognostic signature in breast cancer.

To develop a prognostic signature that could be used to classify gene expression profiles from individual patients, we computed the relative expression of hubs with each of their interacting partners, determined for which hubs the relative expression differed significantly between patients who survived versus those who died from disease, and then employed affinity propagation clustering[28]. Affinity propagation is a clustering algorithm that takes similarity measures between data points and iteratively refines them until there are high quality exemplars. Clustering of test patients using affinity propagation allowed us to assign a probability of poor prognosis for each patient (**Supplementary Methods** and **Supplementary Fig. 9** online). We used a fivefold cross-validation strategy in which the hub selection process was incorporated on the training set within the cross-validation loop to avoid overfitting and assessed performance using receiver operator characteristic (ROC) curves. This revealed a typical area under the curve (AUC) of 0.711 (**Fig. 4a**) and accuracy, sensitivity and specificity of 76%, 86% and 81%, respectively. This compared favorably with the retrospective[29] or prospective[30] performance of commercially available genomic breast cancer diagnostics (53%, 41% and 68% in ref. 30 and 70%, 71% and 67% in ref. 29 for predicting 10-year survival[31]).

We also assessed performance using interactomes in which hubs were randomly removed. We observed that the performance of the classifier was reduced as hubs were removed (**Fig. 4d**), indicating that our accuracy may be limited by the interactome density. As current interactomes are likely incomplete and contain biases[32], further interactome mapping by systematic approaches may lead to improved prognostic performance.

To test the ability of the classifier to predict survival, we grouped patients using the poor outcome probabilities. The threshold for probability of prognosis was set to 0.4 as this consistently yielded the highest accuracy of prediction. Analysis of these two groups revealed significantly different 5-year survival (Mantel-Cox Log Rank test, nominal $P < 0.001$). Only 48% of patients possessing the poor-prognosis modularity signature survived for >5 years (**Fig. 4b**). Conversely, 85% of those with a good prognostic signature survived for 5 years. The average overall error rate of prognosis using the test-set data at this prognostic cutoff was 29.1%.

We next asked whether prognostic accuracy could be improved by incorporating clinical data (patient age, tumor stage and tumor grade). A logistic regression model that incorporated these variables along with network probabilities resulted in better performance (AUC = 0.784) (**Fig. 4a**) and enhanced prognostic classification (error rate, 25%) (**Fig. 4b**). Clinical covariates alone showed similar performance as the network probability score (AUC = 0.701, **Fig. 4a**). We also repeated these analyses using expression data from the TransBIG[30] cohort of breast cancer patients and observed similar, if not better, performance (AUC = 0.718–0.827; **Supplementary Fig. 9a** online) and Kaplan-Meier survival curves. Thus, >80% of predicted good-prognosis patients survived >10 years compared with <35% of those in the poor-prognosis group (**Supplementary Fig. 9b**). These results demonstrate that the molecular changes of the tumor that are captured by measuring changes in the network modularity of tumor interactomes are significant and independent predictors of patient disease outcome and suggest that measuring these changes may improve the predictive value of prognostic indicators already used in the clinic.

Previous approaches have employed network information to improve classification performance of gene signatures by extracting co-expressed pathways (that is, functional modules) and then using these pathways to assess cancer outcome[1]. In contrast, we have searched for changes in the global modularity of the human interactome that indicate altered organization and information flow. Other mechanisms that affect network connectivity, such as alterations in protein stability and post-translational modification, may also influence network modularity on a global scale during cancer progression. Our studies motivate the search for multi-modal therapies that target hubs in networks that display altered modularity in disease. Furthermore, the favorable performance of our classification algorithms suggests that changes in network modularity may be a defining feature of tumor phenotype that, in turn, determines patient prognosis.

METHODS

Data integration to determine PCC of co-expression in interaction networks. We used a method analogous to that previously described[7]. The complete interactome from OPHID[2] as well as subsets of interactions mapped from yeast to man[33] or literature-curated interactions[4] were downloaded as well as expression data from 79 human tissues[3]. Hubs were defined to be nodes with more than five interactions, as these proteins are in the top 15% of the degree distribution of the network. For each hub the average PCC of co-expression for each interaction and the hub was assessed using a similar algorithm as previously described[7]. Random reassignment of the expression values to nodes in the network was used to ascertain if the observed network was nonrandom. The network was visualized using Cytoscape 2.5.1 (ref. 34).

Topological network analysis. Betweenness and CPL of networks were calculated using algorithms implemented by the tYNA web interface[13]. When assessing network robustness to hub removal, an equivalent number of intermodular and intramodular hubs were removed from the network in order of descending clustering coefficient.

To validate that the two hub classes are distinct, we investigated length, phosphorylation, linear motifs, globularity, domain architecture (**Supplementary Methods**). These were either computed directly from the hub sequence or by mapping to the appropriate database[35]. Significance levels were computed by sampling (**Supplementary Methods**).

Distribution of hub types by human disease phenotypes. For each hub, gene entries in OMIM[18] were extracted and manually curated for hubs (i) associated with cancer, malignancy or metastasis and (ii) found to be involved in oncogenic translocation fusions.

Network analysis between breast tumor samples. To assess differences in network organization between patients who were alive after extended follow-up versus those that died from disease, we used a nonparametric algorithm, within a cross-validation loop, to determine the difference in correlation of co-expression of hubs with their interactors. First, we calculated the PCC of hubs and their interactors for each patient group. We then calculated the absolute value of the difference of these PCCs. The magnitude is the difference in PCC of a hub between patient groups. To identify hubs that are significantly different between patient groups, we randomly assigned patients to one of two groups and repeated the analysis. This was done 1,000 times to calculate the random distribution. Real PCC differences for hubs between patient groups were compared to the random distribution to generate *P*-values. This defines a network signature of hubs whose modularity is different as a function of disease outcome. *P*-value cutoff and degree cutoff for hubs were optimized as a function of accuracy during cross-validation runs.

To measure prognostic accuracy of this network, we trained an affinity propagation algorithm[28] using the network signature to predict the patient outcome using fivefold cross-validation. Specifically, we partitioned the patient cohort into five approximately equally-sized portions, defined a network signature and trained our algorithm using four of these portions as described in detail in **Supplementary Methods**. To test the algorithm, we provided it with only the gene expression data for patients in this latter hold-out training set and compared its predictions of clinical outcome with the actual outcomes for these patients. We repeated this procedure for each hold-out set, amassing outcome predictions for every patient. To measure the variability in our predictions, we repeated the fivefold cross-validation procedure three times with different random partitions of the data.

Breast cancer patient prognostic predictive value is related to the total size of the protein interaction network. Interactions were randomly removed to obtain interactomes of reduced size, as indicated. The accuracy of prediction of outcome using dynamic network modularity at each indicated interactome size was then assessed by ROC curve analysis and is plotted as the average AUC (±s.d.) of three runs of fivefold cross-validation.

Kaplan-Meier survival curves were drawn for groups defined by the algorithm using patient survival data and drawn using SPSS for Mac, Rel.14.0.1.

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

ACKNOWLEDGMENTS

We acknowledge Lars Juhl Jensen, EMBL-Heidelberg, for critical assessment of this manuscript. We would like to thank L. Attisano for critical review and helpful discussions. Work in J.L.W.'s lab was supported by funds from Genome Canada through the Ontario Genomics Institute and the Ontario chapter of the Canadian Breast Cancer Fund. D.W.-F. was supported by an Natural Sciences and Engineering Research Council of Canada operating grant assigned to Q.M. J.L.W. is a CRC chair and an International Scholar of the Howard Hughes Medical Institute. I.W.T. is supported by a fellowship from the Canadian Breast Cancer Foundation.

AUTHOR CONTRIBUTIONS

I.W.T. contributed to the development of the project, the execution of all experiments and the writing of the manuscript. R.L. contributed to experiments in **Figure 2** and writing of the manuscript. D.W.-F. and Q.M. contributed to the development and implementation of the prognosis classification algorithm and the cross-validation strategies, as well as to the writing the manuscript. Y.L. contributed programming support for data in **Figures 1–4**. C.P. & D.F. contributed to the semantic similarity experiment in Supplementary **Figure 2b**. S.B. contributed to the development of the statistical frameworks throughout the manuscript and writing the manuscript. T.P. contributed to writing of the manuscript. J.L.W. contributed to the development of the project and the writing of the manuscript.

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

1. Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D. & Ideker, T. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* **3**, 140 (2007).
2. Brown, K.R. & Jurisica, I. Online predicted human interaction database. *Bioinformatics* **21**, 2076–2082 (2005).
3. Su, A.I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* **101**, 6062–6067 (2004).
4. Chatr-aryamontri, A. *et al.* MINT: the Molecular INTeraction database. *Nucleic Acids Res.* **35**, D572–D574 (2007).
5. von Mering, C. *et al.* STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.* **35**, D358–D362 (2007).
6. Fraser, H.B. Modularity and evolutionary constraint on proteins. *Nat. Genet.* **37**, 351–352 (2005).
7. Han, J.D. *et al.* Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**, 88–93 (2004).
8. Barabasi, A.L. & Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).
9. de Lichtenberg, U., Jensen, L.J., Brunak, S. & Bork, P. Dynamic complex formation during the yeast cell cycle. *Science* **307**, 724–727 (2005).
10. Tengowski, M.W., Feng, D., Sutovsky, M. & Sutovsky, P. Differential expression of genes encoding constitutive and inducible 20S proteasomal core subunits in the testis and epididymis of theophylline- or 1,3-dinitrobenzene-exposed rats. *Biol. Reprod.* **76**, 149–163 (2007).
11. Thomas, M.K., Yao, K.M., Tenser, M.S., Wong, G.G. & Habener, J.F. Bridge-1, a novel PDZ-domain coactivator of E2A-mediated regulation of insulin gene transcription. *Mol. Cell. Biol.* **19**, 8492–8504 (1999).
12. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
13. Yip, K.Y., Yu, H., Kim, P.M., Schultz, M. & Gerstein, M. The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks. *Bioinformatics* **22**, 2968–2970 (2006).
14. Yu, H., Greenbaum, D., Xin Lu, H., Zhu, X. & Gerstein, M. Genomic analysis of essentiality within protein networks. *Trends Genet.* **20**, 227–231 (2004).
15. Puntervoll, P. *et al.* ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.* **31**, 3625–3630 (2003).

16. Letunic, I. *et al.* SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.* **34**, D257–D260 (2006).
17. Karnoub, A.E. & Weinberg, R.A. Ras oncogenes: split personalities. *Nat. Rev. Mol. Cell Biol.* **9**, 517–531 (2008).
18. McKusick, V.A. Mendelian inheritance in man and its online version, OMIM. *Am. J. Hum. Genet.* **80**, 588–604 (2007).
19. Futreal, P.A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
20. van de Vijver, M.J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009 (2002).
21. Roukos, D.H. Prognosis of breast cancer in carriers of BRCA1 and BRCA2 mutations. *N. Engl. J. Med.* **357**, 1555–1556, author reply 1556.
22. Soderlund, K. *et al.* Intact Mre11/Rad50/Nbs1 complex predicts good response to radiotherapy in early breast cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **68**, 50–58 (2007).
23. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116–5121 (2001).
24. Chang, H.Y. *et al.* Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS Biol.* **2**, E7 (2004).
25. Liu, R. *et al.* The prognostic role of a gene signature from tumorigenic breast-cancer cells. *N. Engl. J. Med.* **356**, 217–226 (2007).
26. Sorlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. USA* **100**, 8418–8423 (2003).
27. Easton, D.F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093 (2007).
28. Frey, B.J. & Dueck, D. Clustering by passing messages between data points. *Science* **315**, 972–976 (2007).
29. Paik, S. *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **351**, 2817–2826 (2004).
30. Buyse, M. *et al.* Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J. Natl. Cancer Inst.* **98**, 1183–1192 (2006).
31. Haibe-Kains, B. *et al.* Comparison of prognostic gene expression signatures for breast cancer. *BMC Genomics* **9**, 394 (2008).
32. Bertin, N. *et al.* Confirmation of organized modularity in the yeast interactome. *PLoS Biol.* **e153** (2007).
33. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403 (2002).
34. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
35. Linding, R. *et al.* Systematic discovery of *in vivo* phosphorylation networks. *Cell* **129**, 1415–1426 (2007).

Corrigendum: What's fueling the biotech engine—2007

Saurabh Aggarwal

***Nat. Biotechnol.* 26, 1227–1233 (2008); published online 7 November 2008; corrected after print 9 February 2009**

In the version of this article initially published, 2007 US sales of therapeutic enzymes in Figure 3 were incorrectly listed as $0.7 billion. They should be $0.9 billion. In the same figure, the 2007 growth (%) was incorrectly listed as –14%; it should be 18%. On page 1230, human growth hormone sales growth rate was incorrectly reported as 45%; it should be 5%. The errors have been corrected in the PDF and HTML versions of this article.

Erratum: What's fueling the biotech engine—2007

Saurabh Aggarwal

***Nat. Biotechnol.* 26, 1227–1233 (2008); published online 7 November 2008; corrected after print 9 February 2009**

In the version of this article initially published, in Figure 5, the $2.40 billion pie section was mislabeled as Rituxan. It should be Procrit. The error has been corrected in the PDF and HTML versions of this article.

Erratum: Asymmetric RNA duplexes mediate RNA interference in mammalian cells

Xiangao Sun, Harry A Rogoff & Chiang J Li

***Nat. Biotechnol.* 26, 1379–1382 (2008); published online 23 November 2008; corrected after print 9 February 2009**

In the version of this article initially published, on page 1379, column 2, line 7, a parenthesis was misplaced. "…the passenger (often the sense strand)" should have read, "…the passenger (often the sense) strand". The error has been corrected in the HTML and PDF versions of the article.

Erratum: Profile: Alan Alda

George S Mack

***Nat. Biotechnol.* 26, 1325 (2008); published online 5 December 2008**

In the print version of this article, the photo of Alan Alda was erroneously credited. The photo was taken by Alan Alda.

Erratum: Biotech sector ponders potential 'bloodbath'

Peter Mitchell & Brady Huggett

***Nat. Biotechnol.* 27, 3–5 (2009); published online 8 January 2009; corrected after print 9 February 2009**

In the version of this article initially published, we neglected to define our categories for public biotech firms. In the following definitions, the dollar amounts refer to a firm's market capitalization or 'cap': microcap, <$250 million; small cap, $250 million to <$1 billion; mid-cap, $1 billion to <$5 billion; large cap, ≥$5 billion. The definitions have been added to Box 2, Figure 1 legend, in the HTML and PDF versions of the article.

Erratum: Doubts surround link between *Bt* cotton failure and farmer suicide

Cormac Sheridan

***Nat. Biotechnol.* 27, 9–10 (2009); published online 8 January 2009; corrected after print 9 February 2009**

In the version of this article initially published, in column 3, line 3 in the last paragraph, Debdatta Sengupta was referred to as "he." The text should have read "she" as Sengupta is a woman. The error has been corrected in the HTML and PDF versions of the article.

Fear factor

Genevive Bjorn

Job prospects are looking gloomy as the economic downturn runs its course, but there are bright spots for some.

As if years of shrinking budgets hadn't created a competitive enough postdoctoral job market in many fields and sectors, the worldwide economic downturn is now making some difficult-to-navigate career paths downright treacherous. Although hard data are yet to emerge, anecdotal evidence suggests that the slowdown has stymied the efforts of many postdoctoral jobseekers.

"The situation has changed so dramatically, so quickly, that it is difficult to know how to react," says Marc Kastner, dean of the school of science at the Massachusetts Institute of Technology (MIT) in Cambridge. He says he is "very concerned" about the prospects for young scientists.

Many prospective employers had placed advertisements and received applications before the crunch. But as hiring freezes and suspended candidate searches become more common worldwide, fewer jobs will be filled than initially planned, predicts Roger Davies, chairman of the physics department at the University of Oxford, UK. "We won't know the numbers until the jobs are filled in the first four months of the year," he says.

The finances of US universities have taken a big hit (*Nature* **457**, 11–12, 2009). MIT is making 5% spending cuts, but some US universities are taking more drastic steps. In December, Harvard University announced a hiring freeze, and others have followed suit.

Kevin Covey, a third-year Spitzer fellow at Harvard's Center for Astrophysics in Cambridge, Massachusetts, recently applied for a junior faculty position at Johns Hopkins University in Baltimore, Maryland. "About an hour after I sent in the application, the job was cancelled," he says. He is now applying for about a dozen other faculty and postdoc positions.

Public funding not guaranteed

Universities in Europe have so far avoided a freeze on recruitment, largely because most are publicly funded and don't rely on interest payments from large endowments to cover operating costs. Still, cuts may be coming as government revenues decline. And even before the economic slowdown, the UK Engineering and Physical Sciences Research Council had planned to reduce its grants portfolio for physics from £137 million ($206 million) for 2006–2007 to £97 million for 2008–2009, according to council reports.

"I'm using a blanket strategy of applying for every suitable post in the United Kingdom, which amounts to only half a dozen openings," says Daniel Mortlock, an astrophysics postdoc at Imperial College London.

Publicly funded research institutions have reason to be wary. The Gemini Observatory in Hilo, Hawaii, recently suspended searches for two postdoc positions, even though its 2009 budget—which funded those positions—was approved in November. Gemini is supported by a consortium whose members are in Argentina, Australia, Brazil, Britain, Canada, Chile and the United States. Its administrators are worried that the promised funding may not come through. "This is absolutely related to the current economic turmoil," says deputy director Jean-René Roy. "Government revenues are falling in various countries, and shortfalls roll down to funding for scientific research." He adds that Gemini has resumed one search and still hopes to be able to fill both positions.

There are striking exceptions to the belt-tightening, however. The European Space Agency (ESA) received strong support from European ministers during a meeting on November 25 and 26 in The Hague. They agreed to spend £9.9 billion ($13.5 billion) on space-science projects for 2009–2013, thus stabilizing ESA's workforce (*Nature* **456**, 552, 2008). "ESA does not see a reason to deviate from the human-resources policy that it applied before the global economic crisis," says Andreas Diekman, head of ESA's office in Washington, DC.

New careers, new competition

Mortlock says he may soon face the reality of having to leave astrophysics research for another field of science, or even of leaving science altogether. However, Clare Jones, a careers adviser for postgraduates in the career development center at the University of Nottingham, UK, warns that this is not the best time to attempt a complete switch into another field.

Although nontraditional career paths have promise, science PhDs and postdocs may find themselves competing with more and more jobseekers who have MBAs and master's degrees. According to the US Bureau of Labor Statistics, the unemployment rate for advanced-degree holders jumped to 3.1% in November from 2.5% in September; overall unemployment in the United States stands at 6.7% of the workforce.

Professional, scientific and technical services (the category that includes science PhDs employed as consultants, managers and researchers) took a hard hit in October and November, losing 15,727 jobs. Along with the collapse of financial services and layoffs in the drug and biotech industries, this means postdocs are now competing with seasoned insiders for fewer jobs.

One postdoc in virology (who asked to remain anonymous) applied for some 100 consulting jobs early last autumn. "Where I did get interviews, I was interviewing alongside MBAs from Chicago who had just been laid off from Bear Stearns," he says. "It's hard enough for science PhDs to break into consulting, let alone compete with seasoned investment bankers." But his persistence eventually paid off, as he recently received a job offer.

Genevive Bjorn is a freelance writer in Maui, Hawaii.

Box 1 Personal postdoc coaching

Anxious postdoc jobseekers in the United States have a new resource to help jumpstart their careers. Starting this month, the National Postdoc Association (NPA) will offer its members personal coaching services in conjunction with YouPlus, a coaching company.

Interested postdocs will be able to get advice primarily by telephone with follow-up and additional support by e-mail. The coaching process often involves homework assignments in combination with telephone counseling services. These phone conversations encourage introspection and reflection while discussing the postdoc's passions and strengths, and the available career options.

NPA members will be eligible to buy personal coaching services from YouPlus at discounted rates. The usual hourly rate is $90 an hour; NPA executive director Cathee Johnson-Phillips declined to reveal the discounted price.

Although it had been under discussion for some time, the agreement came after the financial downturn pushed negotiators to expedite the process. Johnson-Phillips and others in the NPA are trying to meet a long-felt need: members have, for some time, been asking for more help in developing their skills, in the hope of feeling more comfortable pursuing all the available career options.

Postgraduate researchers often benefit from one-to-one confidential career discussions, says Clare Jones, of the career development center at the University of Nottingham, UK.

"They feel free to express their concerns," she says. "They don't worry that it's going to have any impact on how others view them or their work." *G.B.*

Maja Zavaljevski, a postdoc in hematology at the University of Washington in Seattle, has been dismayed by her biotech job search. "I'm competing with experienced industry PhDs who've been recently laid off," she says. US chemical manufacturing, a category that includes the drug industry, employed 851,000 in November 2008, down from 860,500 people a year earlier.

Unexpected opportunities

Jones says many of her clients are nervous and discouraged. She advocates "creative job searching." For example, seek tips from former colleagues who have moved into industry and might spot an opening in the company before it is posted. Jones also recommends seeking possible moves within one's own institution, for example, looking into scientific management, business development or career advising.

Academics have other options, according to Matthias Haury, coordinating manager of the European Molecular Biology Laboratory's International Centre for Advanced Training in Heidelberg, Germany. Many companies in Europe, he says, still need qualified scientists to work as consultants, because so many academics disdain positions other than "pure research."

One result of the economic crunch could simply be longer job-search times, notes Ryan Wheeler, manager of the postdoctoral services office at the Scripps Research Institute in La Jolla, California. He suspects that most postdocs will weather the storm relatively unharmed by staying in their current jobs or negotiating postdoc extensions. Indeed, unemployment rates among scientists remain low compared with many professions (*Nature* **452**, 777, 2008). Wheeler suggests using any extra downtime to scrutinize one's skills, interests and values, perhaps through coaching or mentoring (**Box 1**). "The whole process," says Wheeler, "is less daunting if you have a better idea of what's important to you in a job."

This story was reprinted with some modification from Naturejobs, 15 January 2009,
doi: 10.1038/nj7227-342a

PEOPLE

Sutro Biopharma (S. San Francisco, CA, USA), formerly known as Fundamental Applied Biology, has announced the appointment of **William J. Newell** (left) as CEO. Newell most recently served as the president of Aerovance. He was also the chief business officer of QLT and senior vice president, corporate and business development at Celera Genomics. **Daniel S. Gold**, Sutro's previous CEO, becomes president and COO.

"I am very pleased to join Sutro Biopharma," says Newell. "Open Cell-Free Synthesis (OCFS) technology has enormous advantages for the development of therapeutic proteins including those with novel scaffolds, such as antibody derivatives like Fab's and scFv's, that are challenging to make using conventional mammalian, yeast and bacterial systems. This novel system takes cell-free protein synthesis to an entirely new and advanced level, and I look forward to helping the company realize the commercial potential of this technology."

Illumina (San Diego) has appointed **Bill Bonnar** as senior vice president of operations, a newly created role. He joins the company from KLA-Tencor, where he served in the same capacity.

Amir Elstein has joined Teva Pharmaceutical Industries' (Jerusalem) board of directors. He had been a member of Teva's senior management since 2005, most recently as executive vice president, global pharmaceutical resources. Elstein has resigned his executive position with the company to rejoin its board of directors. He previously served on the company's board from 1995 to 2004.

Auspex Pharmaceuticals (Vista, CA, USA) has announced the appointment of **Michael Grey** as president, CEO and a member of the board of directors, and **R. Gary Gilmore** as CFO. Grey was previously president and CEO of SGX Pharmaceuticals, which was recently acquired by Eli Lilly. Gilmore has over 20 years experience in corporate finance, most recently as CFO of Althea Technologies.

NextBio (Cupertino, CA, USA) has named **Andrew Grygiel** as vice president of marketing. He comes to the company with more than 20 years of high-technology marketing experience, most recently as senior vice president of global marketing at Orchestria. Previously he was vice president of global industries for EMC's information and content management software group. In addition, **Sergio Gurrieri** has been appointed senior director of business development, with responsibility for identifying new growth opportunities and strategic business development initiatives.

Robert M. Kennedy has joined Venomix (Kalamazoo, MI, USA) as vice president of research. He was most recently a director of International Discovery Sourcing Consultants and a founder of Metabalog. Previously he directed the cardiovascular chemistry, PET chemistry and combinatorial chemistry groups at Pfizer Global Research and Development in Ann Arbor, Michigan.

Ablynx (Ghent, Belgium) has named **Debbie Law** as its chief scientific officer. Law has over 13 years experience in the biotech industry, previously holding the position of vice president of research at PDL BioPharma as well as senior research positions with EOS Biotechnology and COR Therapeutics.

Valeant Pharmaceuticals (Aliso Viejo, CA, USA) has named **Anders Lonner** to its board of directors. Lonner has been group president and CEO and a board member of Meda AB since 1999. Before joining Meda, he was vice president, Nordic region for Astra and then CEO of KaroBio.

Kiwa Bio-Tech Products Group (Beijing) has announced the resignation of **Lianjun Luo** as chief financial officer, a position he held since March 2004. Luo remains as a director of the company.

Nicole Onetto, senior vice president and chief medical officer of ZymoGenetics and **Eric K. Rowinsky**, executive vice president and chief medical officer of ImClone Systems, have joined the board of directors of Symphony ViDA (Rockville, MD, USA), a drug development company established in October 2008 by Symphony Capital Partners in collaboration with OXiGENE (Waltham, MA, USA).

Bruce A. Peacock and **Steven J. Burakoff** have joined the board of directors of Ligand Pharmaceuticals (San Diego), In addition, **Jeff Perry**, who has served as a Ligand director since December 2005, has resigned from the board. Peacock is president and CEO of Alba Therapeutics, and has also been a venture partner with SV Life Sciences since May 2006. Burakoff is a director of the Tisch Cancer Institute at the Mount Sinai Medical Center, where he also serves as professor of medicine and professor of oncological sciences.

Antibiotics developer Nabriva Therapeutics (Vienna) has appointed **William Prince** chief medical officer. Prince has had 15 years of experience developing antibiotics, antifungals and antivirals at GlaxoSmithKline. He joins Nabriva Therapeutics from Surface Logix, where he served as chief development officer.

Champions Biotechnology (Arlington, VA, USA) has appointed three key executives: **Mark Schonau**, most recently chief financial officer (CFO) for Insys Therapeutics, as CFO; **Sara Parkerson**, formerly vice president of oncology services for Matria Healthcare, as director of personalized oncology services US; and **Elizabeth Bruckheimer**, a cancer biologist and pharmacologist, as director of preclinical development. **Durwood C. Settles**, who had served as acting CFO, has been named the company's controller.

Abraxis BioScience (Los Angeles) has announced that in connection with its strategic plan to spin off Abraxis Health, it has named **Lex H.T. Van der Ploeg** as senior vice president of integrative medicine and translational science. Van der Ploeg has over 25 years of experience in enzymology, biochemistry and genetics. For the past four years he has served as vice president, basic research, and site head of the Merck Research Laboratory in Boston.