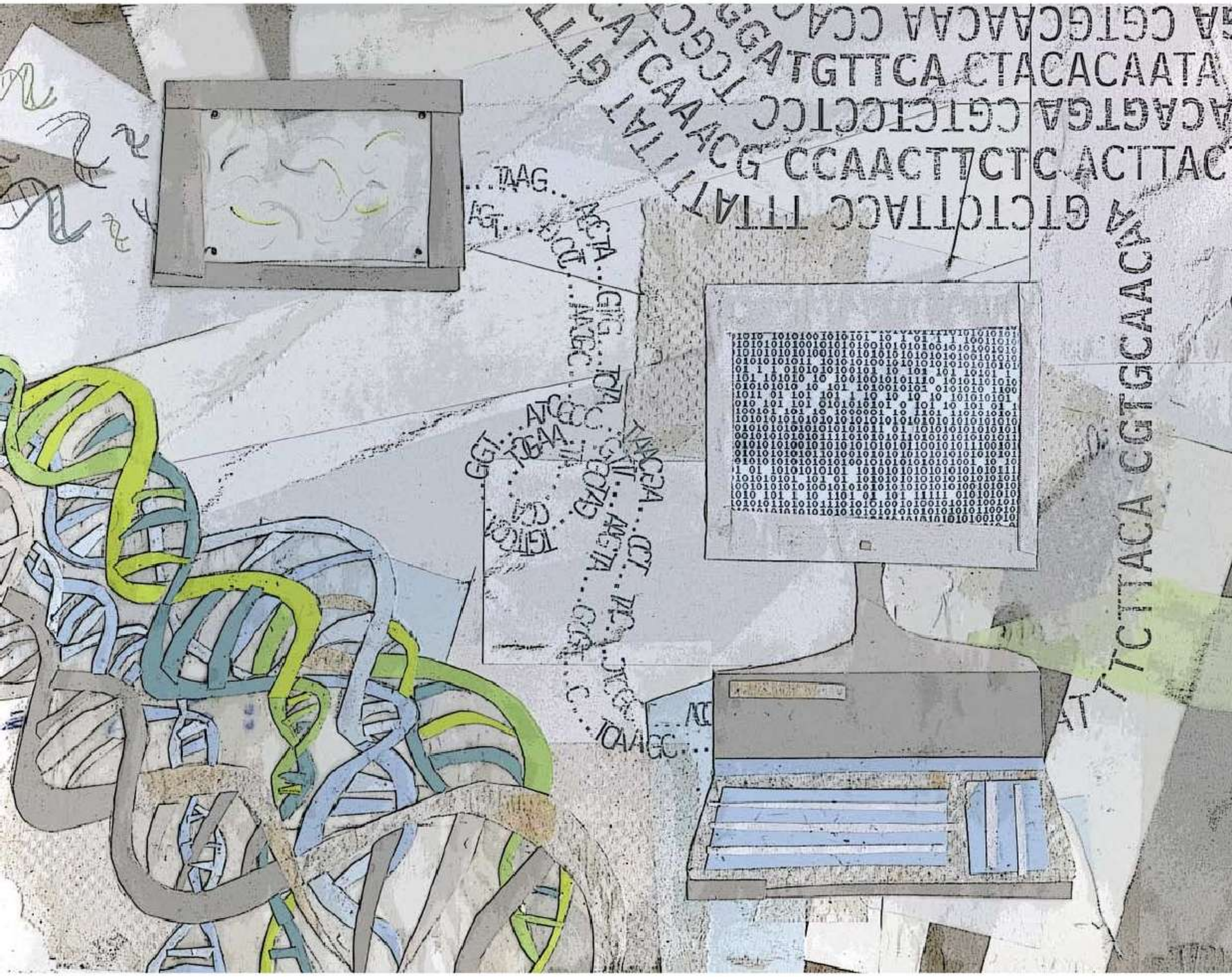
November 2009 | volume 6 | number 11

nature methods

www.nature.com/naturemethods

Techniques for life scientists and chemists



- 4C-ing rearrangements
- Improved reprogramming of human cells
- Strategies for ‘edgetics’
- Visual proteomics of *Leptospira interrogans*
- FOCUS: NEXT-GEN SEQUENCING DATA ANALYSIS

nature methods

Nature Methods (ISSN 1548-7091) 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 9200, Fax: (212) 689 9702. **Annual subscription rates:** USA/Canada: US$150 (personal), US$2,513 (institution), Canada add 5% GST #104911595RT001; Euro-zone: €153 (personal), €1,997. (institution); UK and Europe £99 (personal), £1,288 (institution); Rest of world (excluding China, Japan, Korea): £99 (personal), £1,288 (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. **Back issues:** US$20, Canada add 7% for GST.

POSTMASTER: Send address changes to *Nature Methods*, Subscriptions Dept., PO Box 5054, Brentwood, TN 37024-5054.
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 Methods*: 1548-7091/04. CPC PUB AGREEMENT #40032744. Printed by Publishers Press, Inc., Lebanon Junction, KY, USA. Copyright © 2009 Nature Publishing Group. Printed in USA.

CUSTOMER SERVICE
Senior Global Customer Service Manager: Gerald Coppin
For all print and online assistance, please visit **www.nature.com/help**
Purchase subscriptions:

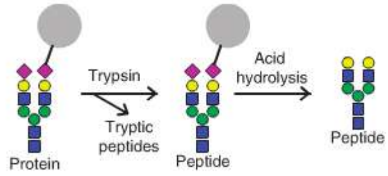
Americas: *Nature Methods*, Subscriptions Dept., PO Box 5054, Brentwood, TN 37024-5054. Tel: (800) 524 0384, Fax: (615) 377 0525

Europe/ROW: *Nature Methods*, Subscriptions Dept., Macmillan Magazines Ltd., Brunel Road, Houndmills, Basingstoke RG21 6XS, United Kingdom. Tel: 44 1256 329 242, Fax: 44 1256 812 358

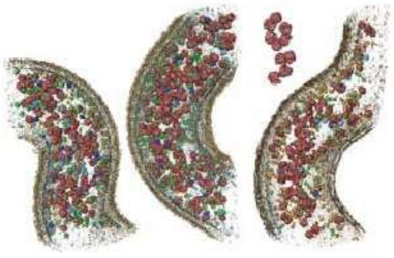
Japan: *Nature Methods*, Nature Japan K.K., MG Ichigaya Bldg. 5F, 19-1 Haraikatamachi, Shinjuku-ku, Tokyo 162-0841. Tel: 81 3 3267 8751, Fax: 81 3 3267 8746

India: Harpal Singh Gill, Macmillan Magazines Ltd, 5A/12 Ansari Road, Darya Ganj, New Delhi 110 002 India. Tel: 00 91 11 324 4186, Fax: 00 91 11 327 2010

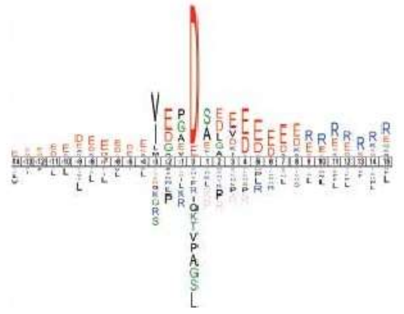
nature publishing group



809 | Glycopeptide profiling



817 | Visual proteomics



786 | Introducing iceLogo

An artistic interpretation of the sequencing process from the DNA molecule to the decoded bases. Cover by Erin Dewalt.

ONLINE FOCUS ON NEXT-GENERATION SEQUENCING DATA ANALYSIS

COMMENTARY

REVIEWS

BRIEF COMMUNICATIONS

ARTICLES

Online Focus on NGS data analysis

Next-generation sequencing (NGS) is causing much excitement because it allows an unprecedented in-depth look into many biological questions. It can also be the source of confusion or even frustration because data processing and analysis critically depend on software tools, of which there are legions. In a freely available online Focus, leading experts describe the analysis steps in key applications of NGS with a discussion of how different algorithms handle these steps. The goal of this Focus is to provide end users with a better understanding of the strengths and weaknesses of currently used software such that readers can make more informed choices when trying to make sense of the millions of reads that come out of an NGS platform.
Focus forward p801

Systematic isolation of interaction mutants

Proteins do not function in isolation but interact with networks of other proteins in the cell. A systematic approach to identify mutations that specifically affect one, or a subset, of these interactions would be useful for the understanding of gene and protein function. In this issue of *Nature Methods*, two groups use different approaches to achieve this goal. Michnick and colleagues describe a protein complementation assay based on the *Saccharomyces cerevisiae* cytosine deaminase, which can be used to select both for and against specific protein interactions. Vidal and colleagues use the forward and reverse yeast two-hybrid assay, applying this approach to isolate interaction mutants of *Caenorhabditis elegans* BCL2. Because the approach perturbs the protein interaction network one edge at a time, Vidal and colleagues define these mutations as 'edgetic'.
Brief Communication p813, Article p843, News and Views p797

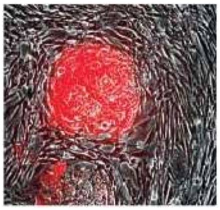
Orchestrating alternative splicing

Transcript diversity all lies in the splicing; *cis* elements on the mRNA recruit *trans*-acting splicing factors to carry out the splicing reaction that brings exons together. Wang and colleagues have now adapted nature's method for the laboratory. By adding sequence specific RNA-binding domains of the Pumilio protein to the mRNA, they created a motif that will recruit an engineered splicing factor. This factor comprises a Pumilio domain and a functional module that affects splicing. This approach allowed Wang and colleagues not only to modulate the splicing of endogenous genes but also to study the activity of splicing factors.
Article p825

Visualizing the proteome of *Leptospira interrogans*

An important component to understanding biology on a systems level is knowing the composition and spatial and temporal locations of the protein complexes responsible for carrying out most important biological functions. Aebersold and colleagues now combine the techniques of quantitative mass spectrometry and cryo-electron tomography (cryoET) to advance the emerging field of 'visual proteomics'. They use mass spectrometry to select and quantify suitable protein complexes for cryoET analysis via a process called template matching, in which cryoET signals from cellular protein complexes are matched to those of a reference structure. Aebersold and colleagues adapted statistical concepts from the mass spectrometry–based proteomics field to develop a rigorous scoring method for template matching. This allowed them to reduce false positive matches and accurately map the location of several specific protein complexes in the human pathogen *Leptospira interrogans*.
Article p817, News and Views p798

Better reprogramming through chemistry



Methods for the efficient generation of induced pluripotent stem cells are still needed. Ding and colleagues present small molecules that improve the efficiency of reprogramming human fibroblasts to pluripotency. Inhibition of the MEK-ERK pathway and of TGF-β signaling during reprogramming with Oct4, Sox2, Kef4 and c-Myc resulted in substantially improved reprogramming efficiency; also, embryonic stem cell–like colonies were visible as early as 14 days after transduction of retroviruses expressing the factors. In addition, the small molecule thiazovivin both increased the survival of induced pluripotent stem cells upon trypsinization and further improved the efficiency of the reprogramming process.
Brief Communication p805

nature methods
www.nature.com/naturemethods

EDITORIAL OFFICE methods@us.nature.com
75 Varick Street, Fl 9, New York, NY 10013-1917
Tel: (212) 726 9200, Fax: (212) 689 9702
Editor: Daniel Evanko
Senior Editor: Nicole Rusk
Associate Editors: Natalie de Souza, Allison Doerr
Senior Copy Editor: Irene Kaganman
Managing Production Editor: Ingrid McNamara
Senior Production Editor: Brandy Cafarella
Production Editor: Amanda Crawford
Senior Illustrator: Katie Vicari
Illustrator: Kimberly Caesar
Cover Design: Erin Dewalt
Senior Editorial Assistant: Hugh Ash

MANAGEMENT OFFICES
NPG New York
75 Varick Street, Fl 9, New York, NY 10013-1917
Tel: (212) 726 9200, Fax: (212) 696 9006
Publisher: Veronique Kiermer
Executive Editor: Linda Miller
Chief Technology Officer: Howard Ratner
Head of Nature Research & Reviews Marketing: Sara Girard
Circulation Manager: Stacey Nelson
Production Manager: Richard Bradley
Head of Web Services: Anthony Barrera
Web Production Editor: Colleen Diez

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

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

DISPLAY ADVERTISING display@us.nature.com (US/Canada)
display@nature.com (Europe)
nature@natureasia.com (Asia)
Global Head of Advertising and Sponsorship: Dean Sanderson, Tel: (212) 726 9350, Fax: (212) 696 9482
Global Head of Display Advertising: Andrew Douglas, Tel: 44 207 843 4975, Fax: 44 207 843 4996
Asia-Pacific Sales Director: Kate Yoneyama, Tel: 81 3 3267 8765, Fax: 81 3 3267 8746
Display Account Managers:
Global Account Development Manager: Graham Combe, Tel: 44 207 843 4914, Fax: 44 207 843 4749
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: Jeremy Betts, Tel: 44 207 843 4968, Fax: 44 207 843 4749
Scandinavia/Iceland/Spain/Portugal: Evelina Rubio-Hakansson, Tel: 44 207 843 4079, Fax: 44 207 843 4749
France/Belgium/The Netherlands/Italy/Israel/Eastern Europe: Nicola Wright, Tel: 44 207 843 4959, Fax: 44 207 843 4749
Asia-Pacific Sales Manager: Ken Mikami, Tel: 81 3 3267 8765, Fax: 81 3 3267 8746
Greater China/Singapore: Gloria To, Tel: 852 2811 7191, Fax: 852 2811 0743

NATUREJOBS naturejobs@us.nature.com (US/Canada)
naturejobs@nature.com (Europe)
nature@natureasia.com (Asia)
US Sales Manager: Ken Finnegan, Tel: (212) 726 9248, Fax: (212) 696 9482
European Sales Manager: Dan Churchward, Tel: 44 207 843 4966, Fax: 44 207 843 4596
Asia-Pacific Sales Manager: Ayako Watanabe, Tel: 81 3 3267 8765, Fax: 81 3 3267 8746

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

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

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

What's in a test?

Customers of genetic and genomic services need better education even more than tighter regulation.

The last three years have seen an expansion in direct to consumer (DTC) genetic services, from those that offer single-gene tests to companies that screen a customer's DNA at various loci for polymorphisms associated with certain diseases or even, for a hefty fee, offer whole-genome sequencing. For the first time the lay public has direct access to methods and data previously limited to professionals. This has raised concerns among public health and consumer advocates as well as governmental institutions and has led to calls for tighter regulation.

Critics are concerned about the analytical validity: that the tests perform accurately; the clinical validity: that the genetic variants tested for are associated with increased disease risk; and the clinical utility: that the information is helpful for the consumers.

Different countries have approached these concerns differently. In the US, with little federal oversight, regulation is largely up to the states and varies widely. In Germany, a recent law (Gesetz über genetische Untersuchung bei Menschen) in essence bans DTC genetic testing, and mandates that only physicians can order genetic tests and that the interpretation of the results must be bundled with counseling. In contrast, the House of Lords' Science and Technology committee in the UK issued a report in July on genomic medicine proposing self-policing by the industry.

Which is the right approach? There may not be a single answer for all DTC genetic tests. A DTC test that screens for a specific mutation implicated in a high-risk disease such as cancer or a disease that could affect one's children requires different considerations than DTC genomic services that test a wider range of loci with more uncertain links to diseases.

But in either case, restrictive regulations have drawbacks. For one, they are hard to enforce. It is difficult to envisage how they can be upheld with companies that sell their services over the internet. And what is to prevent the companies from following the letter rather than the spirit of the law? By partnering with healthcare providers that order the tests for a consumer, DTC companies could circumvent the mandate for physician involvement. Requiring a physician means that the cost of the tests will always be high, even if the technology becomes very cheap.

Are special regulations for DTC genetic services even necessary, or are existing 'truth in advertising' laws sufficient? Current US laws regulating the promotion of services can be enforced through statutory bodies like the Federal Trade Commission. If, for example, a company made untrue claims about clinical validity of its tests, the Federal Trade Commission could step in and either prohibit these claims or issue specific consumer alerts.

Do people really need to be protected from learning their genetic makeup firsthand? Advocates for DTC testing argue that the information is not as 'toxic' as some fear, often citing a recent study by the Reveal study group, which showed that learning about an increased risk for Alzheimer's disease did not increase depression and anxiety in test subjects. Although it may be true that people can handle bad news very well, it is important to note that the Reveal study has its biases. As Robert Cook-Degan, a member of the Reveal study group, pointed out, all test subjects received pretest counseling, and people with a history of anxiety or depression were not included. So there is as of yet no indication of what the societal impact of DTC genetic testing will be.

Companies could do their share to alleviate concerns. Sponsorship of a public database with research-based evidence supporting associations between genes and diseases with tools to view and interpret the DTC data would go a long way. Importantly, the privacy of the information needs to be safeguarded.

But the onus is also on consumers to educate themselves about what DTC genetic services do and do not offer. In the case of tests for a single mutation known to be associated with a disease, it is important to look at the details of the science involved. For example, mutations in *BRCA1* and *BRCA2* mainly indicate elevated risk of breast cancer in women with a family history of the disease. In the case of DTC genomic tests, the increased disease risk owing to certain alleles is often very small. There needs to be an understanding that genotyping or sequencing data should not be a node in the decision tree to medical intervention. Rather this information should form the basis for a more detailed talk with a physician or genetic counselor.

With sequencing costs dropping, it is likely that DTC genetic services will soon include affordable whole-genome sequencing. Consumers who have familiarized themselves with the limitations of these data will be better equipped for the 3 gigabases of information that may soon come their way.

Spin filter–based sample preparation for shotgun proteomics

To the Editor: Wiśniewski *et al.* recently reported a sample preparation method for proteome analysis using spin filter microcentrifugation devices[1]. The procedure described is almost identical to a method we reported in 2005 (ref. 2). In our paper, we described the use of spin filters to remove sodium dodecyl sulfate (SDS) and other contaminants, followed by the reduction, alkylation and tryptic digestion of proteins on the filter and finally the isolation of peptides by centrifugation. We described the application of the spin filter preparation method to purified proteins, protein mixtures, cell lysates and subcellular fractions, which are the major elements of the method described by Wiśniewski *et al.*[1]. Our spin filter method already has seen considerable use: we are aware of at least 18 publications in which it was applied (**Supplementary Note**).

These publications show that this approach is useful in some applications, but is not necessarily "universal" as Wiśniewski *et al.*[1] suggest. We and others have found that the use of spin filters has considerable limitations because of poor peptide recoveries when relatively small (<50 μg) protein samples are analyzed. Even at higher sample loads, digestion efficiencies and peptide recoveries are variable[3]. In our previous work with detergent-solubilized membrane vesicles, the spin filter preparation did not yield protein identifications, apparently owing to the difficulty of removing detergent (1% CHAPS) that interfered with protein digestion. In that work, we used a 'short' SDS–polyacrylamide gel electrophoresis (SDS-PAGE) separation ~1–2 cm into the gel, followed by in-gel tryptic digestion and multidimensional liquid chromatography–tandem mass spectrometry (LC-MS/MS) to identify several dozen vesicle-associated proteins[4]. To analyze cell and tissue proteomes, we also have used the in-solution digestion method of Wang *et al.*[5], which uses trifluoroethanol (TFE) instead of detergent to solubilize hydrophobic and membrane proteins.

We compared the performance of the spin filter method (performed as described by Wiśniewski *et al.*[1]) with that of the short SDS-PAGE and TFE methods for analysis of proteins from RKO colon carcinoma cells (**Table 1** and **Supplementary Data**). We analyzed samples corresponding to a high protein load (50 μg) and a low protein load (150 ng). Then we analyzed all digests under identical conditions by reverse phase LC-MS/MS (**Supplementary Methods**). At the high protein load, the spin filter preparation yielded 8% more protein identifications than the gel and TFE methods. However, at the low protein load, the spin filter method yielded just 44% of the protein identifications found with the gel method and only 31% of the identifications found with the TFE method.

Table 1 | Comparison of spin filter, short SDS-PAGE and TFE methods

Method	Protein load	Peptide identifications	Protein identifications
Spin filter		5,369	642
Short SDS-PAGE	50 μg	4,176	593
TFE		4,663	593
Spin filter		86	46
Short SDS-PAGE	150 ng	298	106
TFE		626	150

Samples of human RKO colon carcinoma cells containing the indicated amounts of protein were prepared in triplicate by the indicated methods and analyzed by reverse phase LC-MS/MS. Peptide identifications are total MS/MS spectrum-to-sequence database matches at 5% false discovery rate; protein identifications are nonredundant identifications with at least two identified peptides and parsimonious protein assembly. Reported values are the means of three technical replicate analyses.

Thus we conclude that spin filter-based approaches are subject to substantial losses of identifications at low sample loads, probably owing to binding of proteins and peptides to the spin filters. We note that Wiśniewski *et al.*[1] only analyzed complex cell proteomes with their spin filter method. However, nonspecific binding and protein or peptide losses would make this method a poor choice for the analysis of less complex samples (for example, multiprotein complex pull-downs), which often represent very small protein loads. The short SDS-PAGE approach is much better suited to such samples.

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

ACKNOWLEDGMENTS

This work was supported in part by the National Cancer Institute Clinical Proteomic Technologies Assessment for Cancer program grant 1U24CA126479. We thank K. Cheek for technical assistance and R. Slebos for valuable comments.

Daniel C Liebler & Amy-Joan L Ham

Department of Biochemistry, Vanderbilt University School of Medicine,
Nashville, Tennessee, USA.
e-mail: daniel.liebler@vanderbilt.edu

1. Wiśniewski, J.R., Zougman, A., Nagaraj, N. & Mann, M. *Nat. Methods* **6**, 359–362 (2009).
2. Manza, L.L., Stamer, S.L., Ham, A.J., Codreanu, S.G. & Liebler, D.C. *Proteomics* **5**, 1742–1745 (2005).
3. Eggler, A.L., Luo, Y., van Breemen, R.B. & Mesecar, A.D. *Chem. Res. Toxicol.* **20**, 1878–1884 (2007).
4. Lapierre, L.A. et al. *Am. J. Physiol. Gastrointest. Liver Physiol.* **292**, 1249–1262 (2007).
5. Wang, H. *et al. J. Proteome Res.* **4**, 2397–2403 (2005).

Wiśniewski & Mann reply: We welcome the correspondence by Liebler and Ham[1] because it gives us the opportunity to correct an embarrassing oversight in our Brief Communication describing the filter-aided sample preparation (FASP) method published earlier this year[2]. The method by Manza *et al.*[3] indeed has similarities to our protocol and had we been aware of it, we would certainly have cited it. Unfortunately, neither we nor the reviewers, nor the many people that have already used our protocol for a year were aware of the paper. More importantly, there are fundamental differences between the methods. Both perform digestion in a 'chemical reactor' (in this case a spin column) as do many other protocols in proteomics (for example, ref. 4). However, we completely eliminated sodium dodecyl sulfate (SDS) and other detergents by urea exchange, which we had previously introduced for complete sample solubilization[5]. This was the main advance of our protocol, which

enabled us to combine the advantages of in-gel and in-solution digestion workflows. It has been commonly held that SDS, once introduced into the sample, will make subsequent mass spectrometric analysis impossible. Manza *et al.*[3] explicitly state in their paper that they could not completely remove SDS and that its presence reduced the number of identified BSA peptides. Indeed, after multidimensional separation Manza *et al.*[3] identified 75 soluble cytosolic and 142 nuclear proteins. In contrast, our FASP approach allowed us to identify more than 7,000 proteins, about one-third of which were membrane or membrane-associated proteins[2]. In FASP, SDS is dissociated from proteins using urea. This presumably sequesters them into small micelles, which can pass through the filter pores, thus separating protein and detergent. The method of Manza *et al.*[3] does not use such a step and is therefore not effective at removing SDS or other detergents.

FASP achieves essentially complete protein unfolding during the whole process of detergent removal, which allows use of large-molecular-weight cut-off filters without a loss of small proteins. In contrast, Manza *et al.*[3] reported that it was necessary to limit filter size to the 3–5 kDa range. The ability of FASP to work with larger pore filters substantially reduces sample preparation time.

Liebler and Ham[1] also state that the method is not "universal" because it disproportionately loses protein at low sample loads. We did not specifically develop the FASP protocol for high-sensitivity work. However, we demonstrated identification of 1,700 proteins from HeLa cell material corresponding to only 1,250 cells (750 ng total protein)[2]. We have now tested FASP with tenfold lower amounts and did not observe a disproportionate reduction in peptide ion current, peptide or protein identifications (Supplementary Note). Current commercial spin filters are not optimized for FASP, and they are not optimal for working with very small protein amounts (<100 ng). Miniaturization of the filter units should reduce proteins losses proportionally. In any case, in describing FASP method as "universal," we were specifically referring to its ability to represent the proteome in an unbiased way, which we demonstrated by comparison to the transcriptome.

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

Jacek R Wiśniewski & Matthias Mann

Department of Proteomics and Signal Transduction, Max-Planck Institute for Biochemistry, Martinsried, Germany.
e-mail: mmann@biochem.mpg.de

1. Liebler, D.C. & Ham, A.-J.L. *Nat. Methods* **6**, 785 (2009).
2. Wiśniewski, J.R., Zougman, A., Nagaraj, N. & Mann, M. *Nat. Methods* **6**, 359–362 (2009).
3. Manza, L.L., Stamer, S.L., Ham, A.J., Codreanu, S.G. & Liebler, D.C. *Proteomics* **5**, 1742–1745, (2005).
4. Ethier, M., Hou, W., Duewel, H.S. & Figeys, D. *J. Proteome Res.* **5**, 2754–2759 (2006).
5. Nagaraj, N., Lu, A., Mann, M. & Wiśniewski, J.R. *J. Proteome Res.* **7**, 5028–5032 (2008).

Improved visualization of protein consensus sequences by iceLogo

To the Editor: Large sequence-based datasets are often scanned for conserved sequence patterns to extract useful biological information[1]. Sequence logos[2] have been developed to visualize conserved patterns in oligonucleotide and protein sequences and rely on Shannon's information theory to calculate conservation among all positions in a multiple-sequence alignment. A sequence logo, such as that created by the popular WebLogo tool[3], is a histogram-like presentation in which bars are vertical stacks of symbols; the stack height reflects the extent of conservation, and the height of individual symbols reflects their relative frequency at a given position. However, to our knowledge, no existing tool can compare, in a statistically sound manner, an experimental peptide or protein sequence set to (i) the background of species-specific natural occurrences of amino acids, (ii) a position-specific background set or (iii) a background set that is influenced by the experimental protocol. In addition, underrepresented elements—nontolerated amino acids or nucleotides—are generally not or not statistically well presented (**Supplementary Note 1**).

Here we introduce iceLogo, a free, open-source Java application for the analysis and visualization of consensus patterns in aligned peptide sequences (http://icelogo.googlecode.com/; a description of methods used by iceLogo and a user manual are available in **Supplementary Note 2**). Instead of relying on information theory, iceLogo builds on probability theory. The user first defines an appropriate reference set, tailoring it to ideally approximate the expected background distribution. These reference set distributions and associated standard deviations are then used to test the experimental set, which results in a probability value (Z score) that indicates whether or not the reference set and the experimental set are equal (null hypothesis) or are different (alternative hypothesis). This probability value implicitly takes into account sample size, avoiding misinterpretation of sequence logos from small experimental sequence sets. The reference set can be derived from a multiple-sequence alignment, from the natural amino acid composition or from Monte Carlo sampling a FASTA format database. The experimental sequence set is generally a multiple sequence alignment of peptides that are expected to share sequence features. Finally, the result of the probability analysis can be displayed in complementary illustrations such as position-specific bar charts, heatmaps and so-called iceLogos, which we developed to aid analysis, visualization and understanding of consensus sequences intuitively.

We illustrate the use of iceLogo and its visualization methods with two recent analyses[4,5] done in our laboratory. In an analysis of the substrate specificity of human granzyme B (ref. 4), we generated both a WebLogo (**Supplementary Fig. 1**) and an iceLogo (**Fig. 1a**) for 452 identified human granzyme B cleavage sites. The extended specificity of granzyme B for acidic residues surrounding the cleavage site[4] was very clear in the iceLogo, which compared the experimental set to the human proteome as reference set. However, in the WebLogo, this acidic stretch was hidden in the noise. iceLogo also generated amino acid parameter graphs for over 500 physicochemical and biochemical amino acid properties[6]; using the net charge amino acid parameter, this preferred acidic region was again clearly visualized by iceLogo (**Supplementary Fig. 2**). In a second study, we determined the substrate profile of the yeast N-terminal acetyltransferase A (NatA) complex[5], responsible for the majority of co-translational acetylation of nascent yeast protein N termini. We observed a strong preference for α-acetylation of proteins starting with serine and an elevated average frequency of serine (23%) at this position in the theoretical yeast proteome. Thus, any random set of yeast protein N termini would yield a significant ($P < 0.05$) sequence logo with serine at position 2 (when counting the initiating methionine), independent of NatA specificity consid-

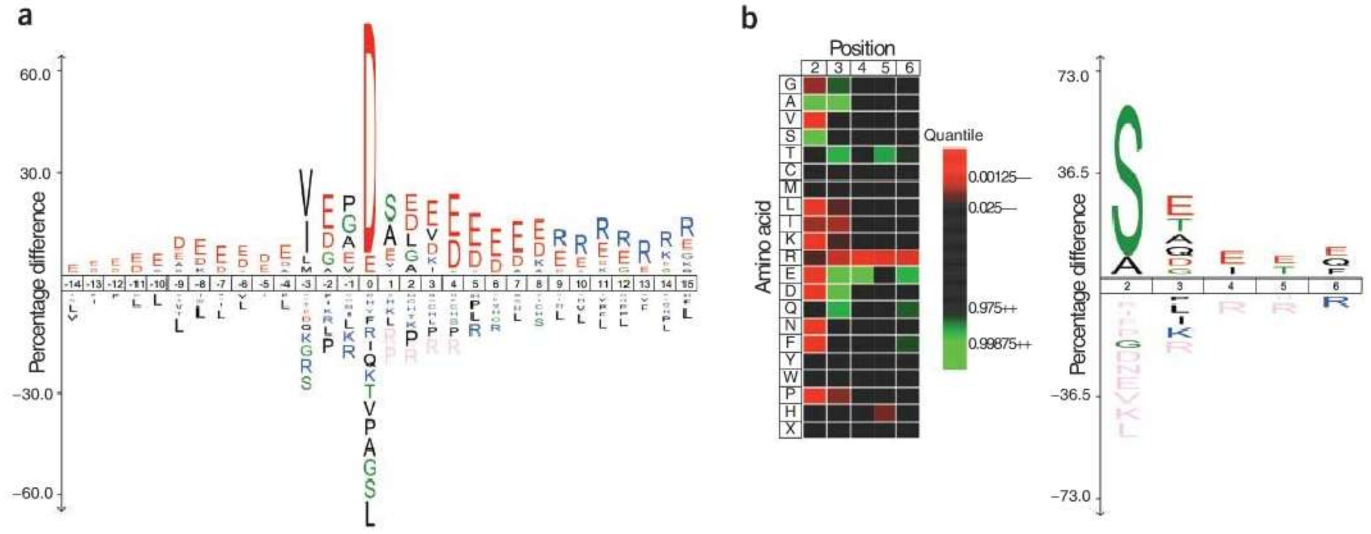


Figure 1 | Visualization of protein consensus sequences by iceLogo. (**a**) An iceLogo generated from 452 human granzyme B cleavage sites with the human Swiss-Prot proteome as reference set. In this logo, granzyme B cleavage occurred at the peptide bond between residues 0 and 1, with positive-numbered residues corresponding to amino acids C-terminal to this scissile bond, and negative-numbered residues are amino acids N-terminal to it. (**b**) Heatmap (left) and iceLogo (right) generated using 163 previously reported NatA targets[6], identifying serine-starting proteins as major targets for α-N-acetylation. The positional reference set was created by iterative sampling of 163 random yeast proteins from position 2 to 6 as indicated. In both iceLogos, only significant amino acids ($P < 0.05$) are shown or colored in the heatmap. The size or color intensity of an amino acid reflects the difference in the frequency of an amino acid in the experimental and its frequency in the reference set. The P value of each amino acid at every position was calculated by testing the experimental frequency against the frequency of each amino acid in the reference set.

erations. To take into account this naturally elevated frequency of serine, iceLogo randomly sampled amino acids anchored to all yeast protein N termini, which verified that NatA indeed strongly prefers acetylating serine-starting yeast proteins (**Supplementary Fig. 3**). The heatmap and iceLogo present general views of consensus features (**Fig. 1b**).

In summary, iceLogo provides a robust, probability-based analysis and improved visualization of consensus sequences in multiple aligned peptide sequences by allowing the user to define a custom background (reference set) of sequences fully tailored to the sample's origin, the studied protein features and the experimental technology.

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

ACKNOWLEDGMENTS

We thank P. Van Damme, F. Impens and K. Plasman for their data, useful suggestions and comments. K.H. is supported by a PhD grant from the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). The laboratory in Ghent acknowledges support by research grants from the Fund for Scientific Research Flanders (Belgium) (project numbers G.0077.06 and G.0042.07), the Concerted Research Actions (project BOF07/GOA/012) from the Ghent University, the Inter University Attraction Poles (IUAP06).

Niklaas Colaert[1,2,4], Kenny Helsens[1,2,4], Lennart Martens[3], Joël Vandekerckhove[1,2] & Kris Gevaert[1,2]

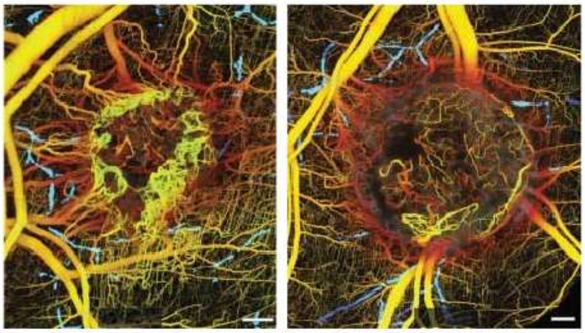
[1]Department of Medical Protein Research, Vlaams Instituut voor Biotechnologie and [2]Department of Biochemistry, Ghent University, Ghent, Belgium. [3]European Molecular Biology Laboratory Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [4]These authors contributed equally to this work.
e-mail: kris.gevaert@ugent.be

1. Hulo, N. *et al. Nucleic Acids Res.* **36**, D245–D249 (2008).
2. Schneider, T.D. & Stephens, R.M. *Nucleic Acids Res.* **18**, 6097–6100 (1990).
3. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. *Genome Res.* **14**, 1188–1190 (2004).
4. Van Damme, P. *et al. Mol. Cell Proteomics* **8**, 258–272 (2009).
5. Arnesen, T. *et al. Proc. Natl. Acad. Sci. USA* **106**, 8157–8162 (2009).
6. Kawashima, S. *et al. Nucleic Acids Res.* **36**, D202–D205 (2008).

MICROSCOPY

Faster than a speeding blood cell

A new *in vivo* imaging strategy produces detailed maps of tumor microvasculature and lymphatic vessels without injected labels.



OFDI visualization of blood and lymphatic vessels. OFDI reveals differences in the density and structure of vasculature in an untreated tumor (left) and in one treated with an antibody that blocks angiogenic activity of VEGFR-2 (right). Lymphatic vessels, in white, are also revealed in both images. Scale bars, 500 μm. Reprinted from *Nature Medicine*.

After more than a quarter-century of studying tumor biology, Massachusetts General Hospital (MGH) researcher Rakesh K. Jain had grown exasperated with the limitations of available tools for performing *in vivo* tumor imaging. "We were very frustrated about how deep we could go, and trying to compare what was going on deeper in the tumor with what was going on superficially," he says.

Fortunately, things recently changed for the better, thanks to a productive collaboration with MGH imaging specialist Brett Bouma's team, who had developed an imaging technique called optical frequency–domain imaging (OFDI) (Yun *et al.*, 2003). In OFDI, a monochromatic light source transmitted through a tissue-penetrating fiberoptic probe is 'chirped', so that it rapidly steps through different wavelengths as it scans the sample; a detector then receives the 'echoes' of this light as it bounces off of tissue surfaces, enabling the computational reconstruction of the positions of those surfaces based on the wavelength of each echo.

To maximize OFDI's usefulness for the applications Jain had in mind, Bouma's team boosted its detection capabilities, making it possible to accurately measure the extremely subtle signal perturbations resulting from Doppler shifts as blood cells flow toward and away from the probe (Vakoc *et al.*, 2009). "It allows very fine sensitivity to very weak back-reflections," says Bouma. "So if we have just one or two photons scattering from a blood cell, we can detect them and characterize the velocity of the flowing blood cell as it tumbles through a capillary."

With this ultrahigh sensitivity comes great imaging speed, and Bouma and Jain could rapidly acquire tremendous amounts of data that could subsequently be translated into detailed three-dimensional vascular maps. Although Doppler OFDI does not achieve the lateral resolution of multiphoton imaging, it can penetrate much further and routinely delivers images from depths of one millimeter or greater. OFDI also enabled the researchers to quickly image considerable volumes, making it possible to reconstruct complete vascular networks from different tumors and thereby perform comparative analyses. "Breast cancer cells in the breast will respond to various therapies, but once these cells go to the brain, they may not respond," says Jain. "This technology allows us to look at what the blood vessels look like at these two sites, and they look entirely different."

OFDI also offered surprising additional benefits, including the unanticipated capability to image lymphatic vessels, which appear as regions of low light-scattering intensity. As with the blood vessel imaging, these data could be acquired purely via intrinsic signals, with no need for injection of contrast agents or fluorescent tracers. "It was a huge bonus," says Jain; "by superimposing blood and lymphatic vessels, you can see how these two networks interact with each other. We could not get this kind of insight with any other technology before this."

These advantages make OFDI a potent tool for assessing how tumors develop or respond to treatment. In one set of experiments, the researchers monitored reduction in the density of vasculature and in the length and diameter of individual blood vessels in mammary tumors from mice treated with a monoclonal antibody that targets vascular endothelial growth factor receptor 2 (VEGFR-2) over both long (imaging every 2 days) and short (imaging every 2 hours) time scales.

They could even directly distinguish healthy cells from necrotic or apoptotic ones based on changes in tissue scattering; for example, OFDI revealed how tumors treated with diphtheria toxin undergo extensive cell death within 48 hours of exposure in parallel with accompanying reduction of blood vessel length.

Both investigators describe their success as the result of highly effective collaboration between their teams and intend to continue working closely together as they prepare to move this system into the clinic. "The technology … is stable and robust," says Bouma. "Practical deployment in actual tumors in humans just requires overcoming the barriers of access and [developing] probes and catheters and endoscopes." Certain tumors may already prove amenable to OFDI imaging. "Breast cancer may be accessible with this device," says Jain, "and it's probably where the earliest impact on human cancer will come."

Michael Eisenstein

RESEARCH PAPERS

Vakoc, B.J. *et al.* Three-dimensional microscopy of the tumor microenvironment *in vivo* using optical frequency domain imaging. *Nat. Med.* **15**, 1219–1223 (2009).

Yun, S.H. *et al.* High-speed optical frequency-domain imaging. *Opt. Express* **11**, 2953–2963 (2003).

CELL BIOLOGY

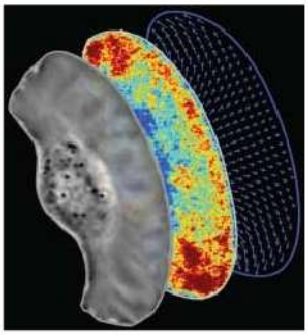
A fluid situation

By monitoring the size-dependence of particle distribution in the lamellipodium, fluid flow in moving cells can be measured.

Cell migration is a critical part of several biological processes. In recent years, studies of motile cells have focused largely on the role of the cytoskeleton in this process. Cells also contain fluid, however, which can respond to hydrodynamic and osmotic pressure in the moving cell. "But this intracellular fluid," says Kinneret Keren at the Technion Institute in Israel, "has been largely invisible."

Keren and her colleagues set out to study fluid flow in the lamellipodia of moving cells, with the idea that such a measurement may help illuminate the role of fluid flux in actin-based cell motility. They found, however, that this was no easy task.

Fluid movement in cells and embryos is typically studied by tracking the movement of single particles. But because of the dense actin meshwork in the lamellipodium, only very small particles (30 nanometers in diameter) could be used in this case. In contrast to the larger particles that are more typical in such studies (100 nanometer–or even micrometer-sized), the small particles diffuse so rapidly that tracking them for the purpose of detecting biased movement is very difficult in practice.



The intracellular fluid flow in the lamellipodium of rapidly moving fish keratocytes was measured by quantifying the steady-state size-dependent distribution of inert probes. Shown are a phase-contrast image, a ratio image (large probe, 655 QDs; small probe, AlexaFluor 488 dye) showing enhancement of large probes toward the leading edge and sides of the cell, and the fluid flow field. Image courtesy of K. Keren.

So the researchers decided to make diffusion work for rather than against them. They reasoned that a net fluid flow in the lamellipodium would have different effects on particles of different sizes. Larger particles with slower diffusion (but still of a size that could enter the lamellipodium) would be affected by flow, whereas smaller, rapidly diffusing particles would have a distribution that is less sensitive to flow. By measuring the ratio between different-sized fluorescent probes introduced into the lamellipodium, Keren and colleagues could determine whether there was indeed such a differential effect. In fish keratocytes, a cell type with rapid, consistent motion, they saw an enhanced localization of larger particles toward the leading edge. From this observed distribution, they inferred a forward-directed fluid flow in the lamellipodium.

GENOMICS

THE TRUE RNA-SEQ

With a modified polymerase and optimized oligonucleotide chemistry, Helicos' single-molecule sequencer takes on RNA.

RNA sequencing (RNA-seq) is actually a misnomer for the increasingly popular technique to determine the sequence of transcripts. It is not the RNA that is being sequenced but its reverse-transcribed cDNA derivative.

Presently available second-generation sequencing platforms all require many copies of the molecule that is to be sequenced and thus include an amplification step in their protocols. As RNA cannot be amplified, the detour via cDNA is necessary, and though the protocols for cDNA generation are well worked out, they are not immune to errors and bias, which can make data interpretation difficult.

Helicos BioSciences has recently introduced a third-generation, single-molecule DNA sequencer, the HeliScope, and now a team led by Fatih Ozsolak and Patrice Milos at the company have adapted the protocol to allow direct RNA sequencing, thus avoiding the cDNA detour and allowing a straight look at the transcriptome.

It is not a given that what works with robust DNA will also work with more fickle RNA. Whereas the scientists did not have to change the principle of Helicos' sequencing by synthesis, Milos says that the main challenge was in modifying all the components of the system—buffer, polymerase and nucleotide chemistry—so that they would work in the context of RNA; the exact nature of these modifications is not being disclosed.

The team started with synthetic 40-mer RNA oligoribonucleotides that they poly(A)-tagged to capture them on the poly(T) surface of the sequencer's flowcell. Their prototype flowcell was small, allowing only thousands of reads, as opposed to the 600–800 million reads in the HeliScope, but the average read number per area on the flowcell was very similar, indicating that the prototype can be scaled up. The average read length was around 20 nucleotides, with an error rate of approximately 4%.

Moving to a biological sample the researchers then sequenced poly(A)-containing RNA from yeast starting with 2 nanograms of material, about 100-fold less than other next-generation sequencing platforms require for RNA-seq. A three-day run yielded just over 41,000 reads of which 48% aligned to the yeast genome. Milos says that the team is now working on scaling the prototype methods up for the HeliScope.

Higher sequencing depth will be beneficial for error correction and quantitative transcript analysis, but another challenge, especially for the discovery of new transcripts and isoforms, Based on a biophysical model built to describe the flows, the researchers postulated that myosin II might generate the pressure at the rear of the lamellipodium that drives forward flow. Indeed, treatment with blebbistatin, an inhibitor of myosin II activity, resulted in a depletion of large particles at the leading edge of the cell. "The fact that the particle distribution was reversed in blebbistatin, which is what we expected, gave us a lot of confidence that we are actually measuring fluid flow with this approach," says Keren. Notably, although blebbistatin treatment slows down the movement of fish keratocytes, it does not halt them completely. At least in the *in vitro* context, therefore, and at least for this cell type, fluid flow is not essential for movement.

Fish keratocytes are among the fastest moving cells. They are involved in wound healing and migrate in a sheet at the surface of the animal. It is possible, as Keren speculates, that the fluid flow seen *in vitro* may also exist at the leading edge of the sheet *in vivo*. It will be of interest to extend this approach to other motile cell types, but this is likely not to be trivial. The measurement requires approximately one minute for particle equilibration within the lamellipodium, so it requires steady movement of the cells over this time frame.

"We study a very simple and probably idealized *in vitro* system for cell movement," says Keren, "but it allows us to show that you can measure and model fluid flow in the moving cell. So for the first time, you can see what the fluid is doing."

Natalie de Souza

RESEARCH PAPERS

Keren, K. *et al.* Intracellular fluid flow in rapidly moving cells. *Nat. Cell Biol.* **11**, 1219–1224 (2009).

is the short read length. Other next-generation sequencing platforms have used paired-end reads, short reads from either end of a longer molecule, to improve isoform discovery. Scientists at Helicos are currently developing a similar but distinct approach for the single-molecule sequencer.

Their strategy involves the capture of long molecules in the flowcell followed by sequencing of the initial 30 nucleotides. Then they turn off the laser, which captures the signal of the incorporated fluorophore-labeled nucleotides, and add unlabeled nucleotides to extend the strand for a defined length, after which they turn the laser back on and sequence the next 30 bases. The end result is intermittent sequence information on a long molecule that will make its characterization much easier than if it has to be assembled from short reads. Milos predicts that this strategy will, for example, be invaluable for finding long intergenic noncoding RNAs, transcripts that span the interval between exonic regions, and she adds: "I think people are very interested to learn if these are indeed true cellular RNAs."

This is just one application for single RNA molecule sequencing; it is likely that in 2010, when Helicos will make this technology available to customers, many more will become apparent.

Nicole Rusk

RESEARCH PAPERS

Ozsolak, F. *et al.* Direct RNA sequencing. *Nature* **461**, 814–818 (2009).

NEWS IN BRIEF

SYSTEMS BIOLOGY

Metabolic network in 3D

Zhang *et al.* used structural genomics to solve three-dimensional structures of the 478 proteins (120 by experiment and 358 by modeling) involved in the central metabolic network of the bacterium *Thermotoga maritima*. This comprehensive structural analysis allowed them to assign metabolic functions to the proteins and identify essential genes as well as investigate questions about the mechanism of metabolic network expansion and the evolution of protein folds.

Zhang, Y. *et al. Science* **325**, 1544–1549 (2009).

STEM CELLS

Reprogramming with OCT4

Four transcription factors, OCT4, SOX2, c-Myc and KLF4, are necessary to reprogram somatic cells into induced pluripotent stem (iPS) cells. But for clinical applications involving iPS cells, it is desirable to avoid overexpressing the oncogenes encoding c-Myc and KLF4. Now Kim *et al.* report that *OCT4* (*POU5F1*) alone is all you need to reprogram human neural stem cells into neural iPS cells that look like and behave like human embryonic stem cells.

Kim, J.B. *et al. Nature* **461**, 649–653 (2009).

CHEMICAL BIOLOGY

Inducing cell signaling

Hashiro *et al.* describe an alternative concept for inducing specific cell-signaling pathways, without requiring protein engineering: they take advantage of the fact that many proteins activate downstream signaling pathways upon localizing to the plasma membrane. By placing a synthetic ligand at the plasma membrane, they can induce translocation of the endogenous protein to the plasma membrane and subsequent activation of the downstream signaling cascade.

Hashiro, S. *et al. J. Am. Chem. Soc.* **131**, 13568–13569 (2009).

PROTEOMICS

Natural product discovery using proteomics

Nonribosomal peptide synthetases and polyketide synthetases are very large enzyme machines in microorganisms that synthesize interesting and potentially pharmacologically valuable metabolites. Bumpus *et al.* describe an assay to enrich for such high-molecular-weight proteins, identify them by mass spectrometry and use the sequence information to link gene expression to the metabolic product.

Bumpus, S.B. *et al. Nat. Biotechnol.* **27**, 951–956 (2009).

GENOMICS

Metagenomics of bug splatter

With a unique approach to sampling the diversity of species in a local environment, Kosakovsky Pond *et al.* collected biological matter (from insects, bacteria and other species) from the front bumper of a moving vehicle and subjected it to phylogenetic profiling. They built a complete pipeline for metagenomic analysis, which involved DNA sequencing of the bumper samples, quality control, sequence alignment via database matching and taxonomic assignment; all tools are available in the Galaxy platform.

Kosakovsky Pond, S. *et al. Genome Res.* advance online publication (9 October 2009).

CHEMISTRY

Keep your eye on the atom

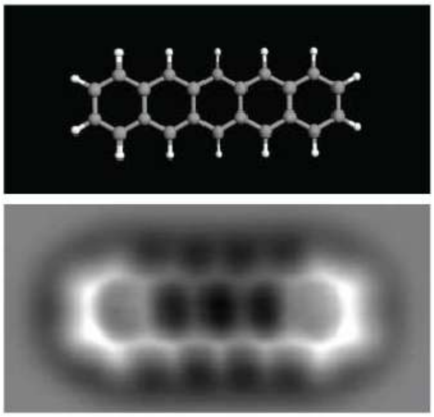
Researchers use atomic force microscopy to image the chemical structure of the small molecule pentacene, with atomic resolution.

How many times have chemists wished for a microscope so powerful that they could see right down to the atomic level of an individual molecule? Scientists at IBM Research in Zurich, Switzerland have now achieved this elusive goal, using atomic force microscopy (AFM).

An atomic force microscope uses a cantilever with a sharp tip to scan the topology of a surface. In noncontact mode, the cantilever is oscillated at a specific frequency; when the tip comes incredibly close to the sample surface, the frequency shifts as a result of forces between the tip and the sample.

To obtain atomic resolution imaging of single molecules, the IBM team had to first tackle some experimental challenges. To ensure a stable operation, they used a low temperature (5 K), noncontact AFM instrument in constant height mode, with a high stiffness cantilever. They also needed to know the exact atomic composition and geometry of the AFM tip to be able to precisely interpret the force measurements; they settled on a tip terminated with a single CO molecule. Unlike a typical metal AFM tip, "The CO tip is quite inert, and it prevents the molecule [under study] from being bonded to the tip," explains Leo Gross, a scientist on the IBM Research team. This, as it turned out, was the key to obtaining atomic resolution.

Gross and his colleagues demonstrated their technique by imaging the small molecule pentacene, which consists of five fused benzene rings. The atomic resolution AFM image speaks for itself. "We didn't expect to have such high resolution," says Gross, noting that the team was surprised that they could clearly image the five rings, the bonds between atoms and even the carbon-hydrogen bonds. But they also performed density functional theory calculations, which confirmed that the pentacene images were what they should have seen.



The atomic structure of pentacene. Ball-and-stick model of pentacene (top). AFM image of pentacene using a CO-modified tip (bottom). Image courtesy of IBM Research, Zurich; reprinted with permission from the American Association for the Advancement of Science.

There are of course other methods by which atomic-resolution molecular structures can be obtained, such as by nuclear magnetic resonance (NMR) spectroscopy or crystallography. But the ability to image individual molecules is unique to this AFM-based technique. With such an approach, for example, "we can say whether something happens to this one molecule; if we have a charge transfer or if we exchange one atom or dissociate one atom," explains Gross.

Beside resolving the structures of unknown molecules, the AFM technique could be used to address a number of interesting chemical questions. Gross and his colleagues are hoping to push the approach to identify atomic species, to distinguish a carbon atom from an oxygen or a nitrogen or a sulfur atom, for example. They also believe the technique could be used to answer fundamental questions about chemical bonding, such as bond order (how many electrons are participating in a bond) and bond length.

By using a reactive molecule at the tip, researchers could also potentially use the approach to directly probe the molecule under study, perhaps leading to new atomic-level insights about chemical reactivity and catalysis. The possibility for these kinds of manipulation experiments, says Gross, "are what distinguishes AFM from other microscopy techniques."

Allison Doerr

RESEARCH PAPERS

Gross, L. *et al.* The chemical structure of a molecule resolved by atomic force microscopy. *Science* **325**, 1110–1114 (2009).

RNA INTERFERENCE

Silence restored

Certain yeast previously assumed to lack RNA interference machinery instead have alternative enzyme variants, which can in turn be transplanted to truly deficient species.

Although pathways for RNA interference (RNAi) exist in a vast array of species, they are notably absent from one of biology's favorite model organisms. "For a long time, people have known that there's no RNAi in *Saccharomyces cerevisiae*," says David Bartel of the Whitehead Institute, "and many people wanting to use the tools of budding yeast to study RNAi have lamented that this is the case."

All budding yeast apparently lack Dicer, an enzyme that processes double-stranded RNAs into small interfering RNAs (siRNAs), although some retain homologs of Argonaute, a key component of the RNA-induced silencing complex (RISC), and Bartel and colleagues were keen to explore whether 'Argonaute-only' yeast can perform RNAi.

They searched for candidate siRNAs in several species, including close *S. cerevisiae* relative *S. castellii*, and identified many molecules exhibiting hallmarks of Dicer-mediated cleavage from endogenous double-stranded RNAs. Although they could not identify canonical Dicer in these species, a more open-ended search for Dicer-like RNase III cleavage domains revealed a novel protein of apparently analogous function, which they named DCR1.

DCR1 lacks standard functional domains found in Dicer homologs from other species but is fully capable of partnering with Argonaute to facilitate RNAi. Even more striking, however, was the finding that transplanting the genes for DCR1 and Argonaute from *S. castellii* to *S. cerevisiae* was enough to render the latter strain fully RNAi-competent. "It's worked to knock down the expression of every gene we've tried to target," says Bartel.

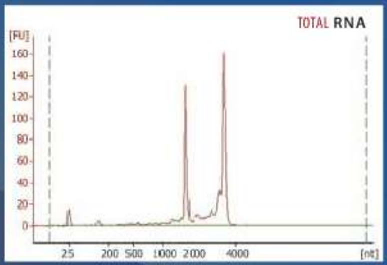
In *S. castellii*, one job of the RNAi machinery appears to entail silencing of retrotransposons, and transplantation of these enzymes has the same effect in formerly RNAi-deficient yeast. "This result shows that the RNAi machinery can recognize transposons it hasn't seen before and specifically silence them but not the other genes of the cell," says Bartel.

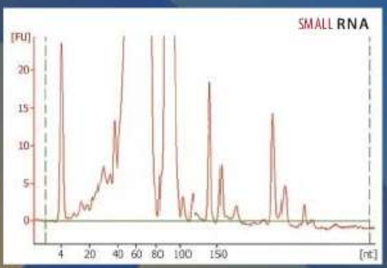
These unexpected findings should expand the *S. cerevisiae* genetic toolbox, and Bartel's Whitehead colleague and collaborator Gerald Fink is also keen to apply RNAi to tackle pathogenic yeast *Candida albicans*, another DCR1-expressing strain that has proven challenging to study.

Michael Eisenstein

RESEARCH PAPERS

Drinnenberg, I.A. *et al.* RNAi in budding yeast. *Science* advance online publication (10 September 2009).





which involves the reconstitution of a split enzyme reporter through a protein-protein interaction. The key to this particular version of the PCA method was the development of *FCY1* as a PCA reporter. *FCY1* encodes cytosine deaminase (yCD), which converts cytosine to uracil and permits both positive and negative selection. N- and C-terminal complementary yCD fragments fused to two different interacting proteins will reconstitute a functional yCD only upon a fruitful physical interaction. Protein interacting partners can be selected positively by requiring that *fcy1* deletion mutants grow in the absence of uracil. Conversely, loss of protein interactions are selected by growth in the presence of 5-fluorocytosine (5-FC), a compound that is converted to the toxic metabolite 5-fluorouridine triphosphate in an *FCY1*-dependent manner.

In this streamlined system[2], if a protein X interacts with both protein Y and a third partner, protein Z, then mutant forms of X can be selected such that they must interact with Y but fail to interact with Z, thereby preserving the X-Y edge but eliminating the X-Z edge in a protein-protein interaction network. After demonstrating that their system works in control experiments, Ear and Michnick[2] applied it to the analysis of Swi6, a regulatory protein that physically interacts with the DNA binding proteins, Swi4 and Mbp1, to form two different transcription complexes.

These two methods offer the potential for a more refined approach to dissecting the meaning of edges on a protein-protein interaction network. Each method builds on the ideas underlying experiments by Amberg, Botstein and colleagues, which combined systematic mutagenesis and physical interaction mapping[5]. They examined alanine mutants of surface-exposed residues on yeast actin[6] for defects in two-hybrid interactions with actin-binding proteins, to characterize a molecular 'footprint' for specific interactions in the context of the actin crystal structure. This approach has been extremely powerful and the two highlighted studies[1,2] expand the scope of proteins amenable to this scrutiny. Phenotypic analysis, including large-scale genetic interaction mapping[7], with edgetic mutations that are mapped to protein structure may reveal the cellular consequences of the loss of a specific edge on a network and trace those consequences back to the protein structure. In this regard, these studies provide a roadmap for integrating global approaches, including protein-protein interaction maps, structural genomics and genetic interaction networks.

The potential for genetic interaction mapping with edgetic alleles of query genes is interesting because the complete removal of genes or nodes is likely to have drastic effects on a network. The phenotypic consequences of removing one edge at a time may be more akin to mutations associated with natural variation. Thus, the knowledge gained from detailed analyses of these mutations in the context of an organism may provide more granular insights into the relationship between genotype and phenotype. Finally, edgetic alleles are relevant to chemical-genetic approaches designed to identify drugs that inhibit a specific protein-protein interaction because they provide an excellent model for understanding phenotypic consequences of drug treatment.

1. Dreze, M. *et al. Nat. Methods* **6**, 843–849 (2009).
2. Ear, P.H. & Michnick, S.W. *Nat. Methods* **6**, 813–816 (2009).
3. Dixon, S.J., Costanzo, M., Baryshnikova, A., Andrews, B. & Boone, C. *Annu. Rev. Genet.* advance online publication, doi:10.1146/annurev.genet.39.073003.114751 (27 August 2009).
4. Varadarajan, R., Nagarajaram, H.A. & Ramakrishnan, C. *Proc. Natl. Acad. Sci. USA* **93**, 13908–13913 (1996).
5. Amberg, D.C., Basart, E. & Botstein, D. *Nat. Struct. Biol.* **2**, 28–35 (1995).
6. Wertman, K.F., Drubin, D.G. & Botstein, D. *Genetics* **132**, 337–350 (1992).
7. Tong, A. *et al. Science* **303**, 808–813 (2004).

Adding a spatial dimension to the proteome

Kay Grünewald

A 'visual proteomics' approach combining quantitative mass spectrometry and cryo-electron tomography offers a window into the spatial arrangement of protein complexes in the human pathogen *Leptospira interrogans*.

Mass spectrometry–based proteomics is currently applied routinely in many biological studies, both for addressing specific questions and for generating 'parts lists' of proteins present under a given condition. More recently, studies quantitatively cataloging proteins in subcellular compartments and in whole cells have been reported. One of the next challenges is to merge these data with visual approaches to reveal the spatial arrangement of the cellular proteome. This is crucial for understanding the entirety and complexity of the functional interactions of the proteome inventory and goes beyond a sole inventory description. In this issue of *Nature Methods*, Aebersold and colleagues[1] present the so far most comprehensive example of a 'visual proteomics' approach using a combination of cryo-electron tomography and quantitative mass spectrometry.

Mass spectrometry–based proteomics is moving from being a qualitative technique to a quantitative one, made possible by methods applying stable isotope protein labels[2]. This is typically done by relative quantification between two experimental conditions. The introduction of stable isotope–labeled peptides as standards, however, allows the absolute quantification of the abundance of a given protein in a specific sample[3]. By using defined specimen preparation protocols, this information can then be used to calculate the respective copy numbers of a protein, for example, in a cell.

Cryo-electron microscopy refers to electron microscopic imaging applied to vitrified biological objects, that is, specimens embedded in amorphous ice. Preserving the water while avoiding traditional methods of specimen preparation that involve chemical fixation, dehydration and heavy

Kay Grünewald is at the Division of Structural Biology, The Wellcome Trust Centre for Human Genetics, Oxford, UK.
e-mail: kay@strubi.ox.ac.uk

metal staining enables the direct structural investigation of biological macromolecules. Various electron microscopic imaging modalities exist, of which cryo-electron tomography[4] has the unique capability to provide three-dimensional reconstructions of unique, highly complex pleomorphic objects. Tomograms of intact cells represent images of their entire unperturbed proteome and provide snapshots of the positions of the multitudes of functional interactions of the proteins in the cellular volume at the time of vitrification.

To unveil the complexity of the information embedded in these images, sophisticated pattern recognition techniques are needed[5]. One approach to this problem is template matching, that is, searching for a macromolecule of known structure in a reconstructed volume. Confident template matching of structures inside a crowded cell[6] is a highly complex and by far not a trivial task, particularly given the currently attainable resolution of 4–6 nanometers with cryo-electron tomography and the peculiarity of the missing wedge in tomographic data caused by the limited angular coverage of the tilt series during data acquisition[4]. Nevertheless, such confidence is critically required for interpreting the biological relevance and implication of the results.

The work by Beck *et al.*[1] breaks new ground in the efforts toward visual proteomics of whole cells, demonstrated on the human pathogen *Leptospira interrogans*. The cross-section diameter of this specimen (100–180 nm) makes it an ideal object for higher-resolution cryo-electron tomography. The authors[1] recently used quantitative mass spectrometry to detect the absolute amounts of proteins in *L. interrogans*[7]. Here they link these average measurements from combined cell lysates to the precise subcellular localization of selected protein complexes in single cells. They identified and localized these protein complexes in the tomograms via cross-correlation–based template matching using a scoring function adapted from a similar statistical concept used in proteomics (**Fig. 1**). They evaluated the weight of the different subscores *in silico* with test data. They also tested and validated the performance and robustness of the improved true positive discrimination rate using different stress conditions—heat shock, antibiotic stress and starvation.

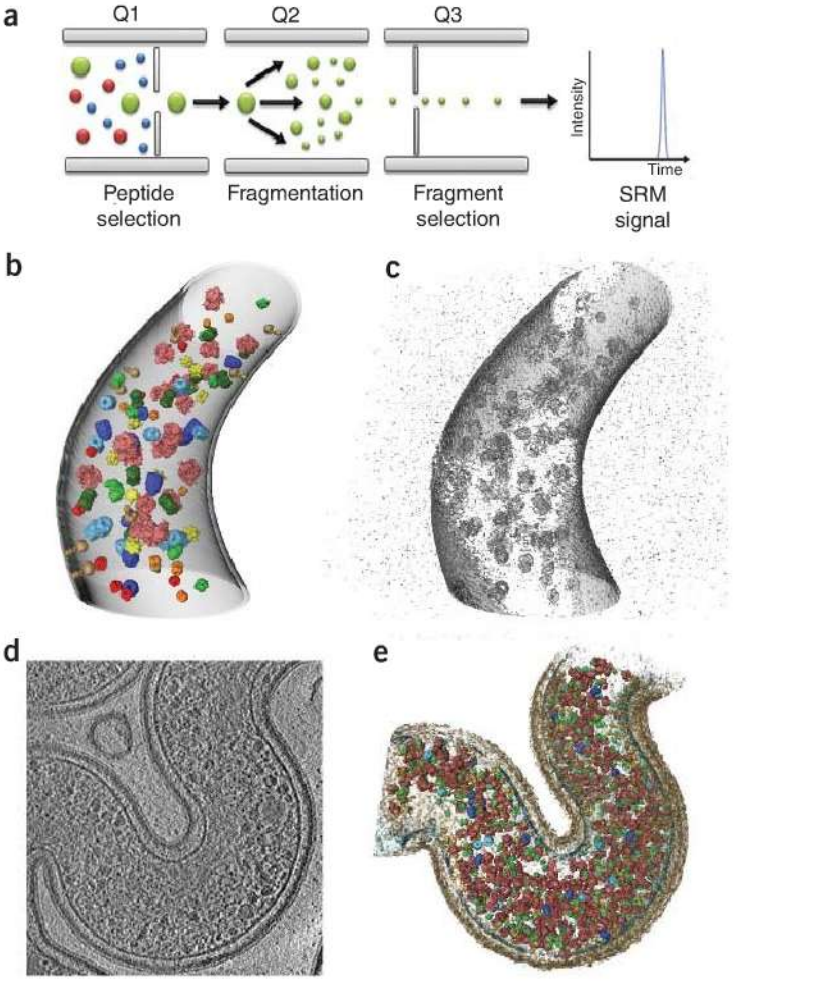


Figure 1 | Integration of quantitative proteomics and cryo-electron tomography data. Mass spectrometry–based proteomics data (**a**) reveal relative abundances for a number of suitable-sized search template complexes. SRM, selected reaction monitoring. On the basis of the absolute copy number of the proteins per cell, Beck *et al.*[1] created realistic *in silico* training sets (**b**), added a noise model (**c**) resembling the situation for cryo-electron tomography data, and performed template matching for the complexes. The results were used to improve the performance of the scoring function. Finally, an optimized scoring function was applied to real cryo-electron tomography data (**d**), resulting in a confident visualization of the spatial arrangement of protein complexes in the human pathogen *L. interrogans* (**e**). Figure panels courtesy of M. Beck.

The synergy between the two technologies used is evident: quantitative proteomics data are used as a validation criterion for the improved template matching scoring function to mine the tomograms and as such increase the confidence of detection. In turn the cryo-electron tomography data provide information on the spatial distribution of the proteome inside the cell. Such spatial information is highly important, as it provides the basis for understanding the structural arrangements of proteins in functional supercomplexes and machines.

At the same time, these tight protein interactions are likely to be still the major limiting factor in the performance of the detection in the cellular context at the current resolution of the tomograms; that is, identifying one component in a supercomplex is harder than for the isolated component and might favor a false positive detection of a similar component rather than detecting the component in the supercomplex. Beck *et al.*[1] did not validate this concept as the chosen macromolecules (of sufficient molecular mass and abundance for which templates where available) did not include examples interacting with each other to form a tight supercomplex. Future studies and application to other organisms or complex (sub)cellular compartments, for example, organelles, will show how robust and generally applicable the scoring function–based approach introduced by Beck *et al.*[1] is.

The particular shape and size of *L. interrogans* also clearly favored the analysis. For now, only a very limited number of similarly suited small and thin organisms are available. With the current technical constraints, one cannot provide a similar analysis and verification for, for example, a region of or an entire adherently grown mammalian cell. The elegant foundation laid out in this work will undoubtedly lead to various applications in 'visual proteomics' in the future, once the electron microscopy technology has advanced a step further, perhaps with the introduction of novel, more sensitive detectors, to allow sufficient resolution and larger field of view. Advances in the reproducibility of mass spectrometry–based proteomic quantification are surely anticipated[8]. Ultimately, the hybrid approach presented here by Beck *et al.*[1] will reveal crucial insights into the structure and function of living cells at the level of the individual protein complexes and their interactions.

1. Beck, M. *et al. Nat. Methods* **6,** 817-823 (2009).
2. Aebersold, R. & Mann, M. *Nature* **422**, 198–207 (2003).
3. Ong, S.E. & Mann, M. *Nat. Chem. Biol.* **1**, 252–262 (2005).
4. Lucić, V., Förster, F. & Baumeister, W. *Annu. Rev. Biochem.* **74**, 833–865 (2005).
5. Nickell, S. *et al. Nat. Rev. Mol. Cell. Biol.* **7**, 225–230 (2006).
6. Grünewald, K. *et al. Biophys. Chem.* **100**, 577–591 (2003).
7. Malmström J. *et al. Nature* **460**, 762–765 (2009).
8. Bell, A.W. *et al. Nat. Methods* **6,** 423–430 (2009).

nature methods

Focus on next-generation sequencing data analysis

A user's guide

An artistic interpretation of the sequencing process from the DNA molecule to the decoded bases by Erin Dewalt.

What used to take years and extensive collaborations—generating the raw sequence of the three gigabases in the human genome—can now be done in a few days by a single investigator using a single run on some of the latest next-generation sequencing machines. The drawback is that this massive amount of data comes in the form of short reads, and one needs to invest heavily in computational analysis and choose from a plethora of tools to make sense of it all.

The recurring theme when it comes to the choice of software is that a 'one-size-fits-all' program does not exist, but users have to mix and match, which requires knowledge about the analysis steps in a given application and how different software operates at each step. This Focus aims to provide some of this information.

Before choosing software, newcomers to next-generation sequencing will be faced with the choice of a platform, each with unique data characteristics. In this Focus the different sequencing technologies are discussed only peripherally; the main goal is to guide readers in their choice of software so they can extract a maximum of information from the data.

On the next two pages we introduce the authors who contributed to the Focus issue and provide a brief summary of the Commentary and the Reviews. Please visit our website for the full text.

We realize that for some of these applications new algorithms are still emerging at a rapid rate. The goal of our authors was not to give a comprehensive list of programs or to pit them against each other and declare a winner. Instead, in the Reviews the authors aim to explain the principles behind the programs and to take the readers through the different analysis steps of an application so that they can make informed choices about software suitable to their needs.

We are pleased to acknowledge the financial support of Applied Biosystems as principal sponsor. As always, *Nature Methods* carries sole responsibility for all editorial content and peer review.

Nicole Rusk

ONLINE CONTENTS

COMMENTARY

REVIEWS

The focus is freely available online thanks to the support of Applied Biosystems.
http://www.nature.com/nmeth/journal/v6/n11s/index.html

Editor, *Nature Methods* Daniel Evanko
Focus Editor Nicole Rusk
Publisher Veronique Kiermer
Senior Copy Editors Anita Gould, Irene Kaganman
Managing Production Editor Ingrid McNamara
Senior Production Editor Brandy Cafarella
Production Editor Amanda Crawford
Design Erin Dewalt
Sponsorship Graham Combe
Marketing Joanna Budukiewicz

npg
nature publishing group

Summary of the Online Focus on next-generation sequencing data analysis

Read the Reviews and Commentary at http://www.nature.com/nmeth/jurnal/v6/n11s/index.html

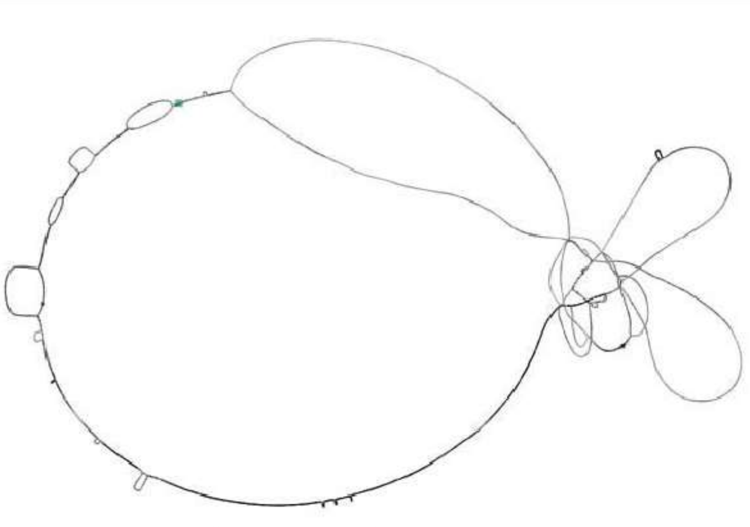
The first steps

The principles behind current alignment and assembly software.

Corresponding author Paul Flicek heads the Vertebrate Genomics Team at the European Bioinformatics Institute (EBI), which is responsible for the creation and maintenance of human variation databases derived from high-throughput sequencing and genotyping data. His main focus is the functional annotation of the genome. Flicek is a leading force behind Ensembl, the genome information system seeking to analyze, visualize and distribute genomic data. Co-author Ewan Birney, cofounder of Ensembl, is at the European Molecular Biology Laboratory and has a keen interest in functional genomics, evidenced by his participation in ENCODE (Encyclopedia of DNA elements). Birney's team wrote the popular assembler Velvet.

The alignment of sequence reads to a reference is the most fundamental analysis step once the read sequence is determined. Not surprisingly, many algorithms are devoted to this task. Rather than discuss each one in turn, an effort that would become outdated almost as soon as it is written, Flicek and Birney describe the working principle, strengths and weaknesses behind hash-based approaches and Burrows-Wheeler transform techniques, the two methods that underlie alignment algorithms.

When no reference genome exists, reads need to be assembled *de novo*, a computationally challenging endeavor, especially in large genomes with long repeat regions. The Review provides insight into how current assemblers work, what they can and cannot achieve and what is still needed. (*Nat. Methods* **6**, S6–S12, 2009) **http://www.nature.com/nmeth/journal/v6/n11s/full/nmeth.1376.html**



Bacterial genome assembled with a de Bruijn graph.

Bridging the gap

How to handle the tension between ever-increasing data volume and the need to process it.

Author John McPherson leads the platform for cancer genomics and high-throughput screening at the Ontario Institute for Cancer Research in Toronto, Canada. His main focus is the sequencing of cancer cells' genomes and epigenomes. Involved in the human genome project during his time as co-director of the Genome Sequencing Center at Washington University, McPherson gained familiarity with next-generation sequencing technology during an appointment as associate professor at the Baylor College of Medicine.

In this Commentary, McPherson gives a brief history of sequencing from its birth in 1977 to the advent of next-generation sequencers. He discusses the state of the art in sequencing technology and presents his views on how the gap between data output on the one hand and analysis on the other might be closed. (*Nat. Methods* **6**, S2–S5, 2009) **http://www.nature.com/nmeth/journal/v6/n11s/full/nmeth.f.268.html**

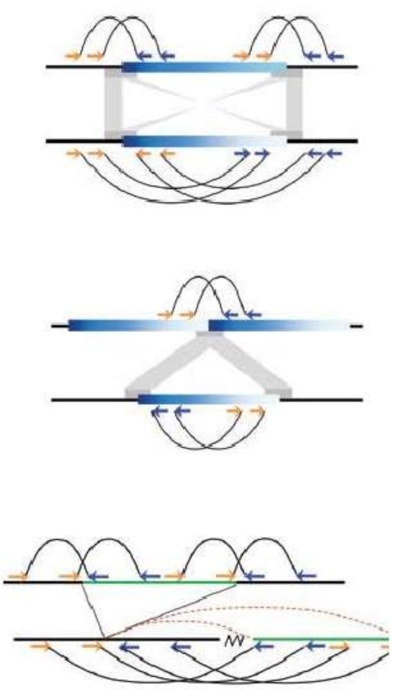
Discovering structural variants

How algorithms detect signatures characteristic of structural variants.

Corresponding author Michael Brudno is an assistant professor and Canada research chair in computational biology at the University of Toronto. Throughout his career Brudno has developed analysis tools for genomics data derived from microarray analyses and next-generation sequencing, and at present he focuses on the development of algorithms for mapping and assembling short-read data and for detecting genomic variation within a species.

As structurally complex as genomic rearrangements can be, they all leave characteristic signatures that can be found in short-read data, if one knows what to look for. In this Review, Brudno and colleagues describe the different classes of signatures, from basic insertions and deletions to inversions and duplications, that can be found either by mapping of paired end reads or by depth-of-coverage measurements. The authors introduce the reader to methods for detecting these signatures (clustering paired-end reads or partitioning the reference into windows) and to current software tools that implement these methods. Tables provide guidance as to the kind of signature each tool detects and show metrics for their performance. (*Nat. Methods* **6**, S13–S20, 2009)

http://www.nature.com/nmeth/journal/v6/n11s/full/nmeth.1374.html

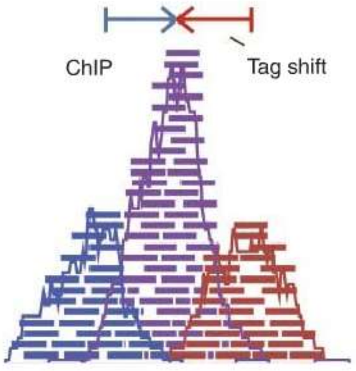


Insertion, deletions and inversions leave characteristic signatures in reads derived from them.

Computing ChIPs and RNA

Algorithms for handling the multistep analysis of data derived from ChIPs or transcriptomes.

Corresponding author Ali Mortazavi is a Beckman Institute Fellow at the California Institute of Technology, working on the applications of next-generation sequencing to genomics and regulatory biology. As part of his PhD thesis on the structure and evolution of mammalian gene networks in Barbara Wold's laboratory at the California Institute of Technology, Mortazavi co-authored one of the first chromatin immunoprecipitation–sequencing (ChIP-seq) experiments in the human (Johnson, D.S. *et al. Science* **316**, 1497–1502, 2007) and RNA-seq experiments in the mouse (Mortazavi, A. *et al., Nat. Methods* **5**, 621–628, 2009). His interests continue to lie in the integration of ChIP-seq and RNA-seq data to reconstruct gene regulatory networks.



A signal profile of ChIP data.

ChIP data—in contrast to data from the genome or transcriptome—are the result of enrichment rather than purification experiments. The challenge is to identify regions of enrichment, so-called peaks, over background. In this Review the authors present the different steps in a ChIP-seq experiment, from peak calling to background filtering and significance ranking, and explain how computational tools handle each of these steps. A table provides an overview of publicly available ChIP-seq software.

Analysis of a transcriptome entails transcript and isoform discovery as well as quantification of gene expression. Usually one of these two aspects is emphasized in a given study and this will determine which software is most appropriate. The authors describe how current software tools handle the different analysis steps in an RNA-seq experiment, from mapping of reads across splice junctions to assigning reads to gene models and quantifying expression. This allows the readers to choose a software package that best meets the needs of their experiment. A table provides information about publicly available RNA-seq software packages and their performance. (*Nat. Methods* **6**, S22–S32, 2009)

http://www.nature.com/nmeth/journal/v6/n11s/full/nmeth.1371.html

A chemical platform for improved induction of human iPSCs

Tongxiang Lin[1], Rajesh Ambasudhan[1], Xu Yuan[1], Wenlin Li[1], Simon Hilcove[1], Ramzey Abujarour[1], Xiangyi Lin[1], Heung Sik Hahm[1], Ergeng Hao[2], Alberto Hayek[2] & Sheng Ding[1]

The slow kinetics and low efficiency of reprogramming methods to generate human induced pluripotent stem cells (iPSCs) impose major limitations on their utility in biomedical applications. Here we describe a chemical approach that dramatically improves (>200-fold) the efficiency of iPSC generation from human fibroblasts, within seven days of treatment. This will provide a basis for developing safer, more efficient, nonviral methods for reprogramming human somatic cells.

Recent advances in generating human induced pluripotent stem cells (iPSCs)[1–3] have raised hopes for the utility of these cells in biomedical research and clinical applications[4]. However, iPSC generation is still a very slow (~4 weeks) and inefficient ($\leq$0.01%[1,2]) process that results in a heterogeneous population of cells. Identifying fully reprogrammed iPSCs from such a mixture is tedious and requires expertise in human pluripotent cell culture.

Although recently developed methods (for example, protein transduction)[3] for iPSC generation mitigate the risk of genomic insertions of exogenous reprogramming factors, the low efficiency and slow kinetics of reprogramming continue to present a problem for ultimate applications of human iPSCs. For example, genetic or epigenetic abnormalities could be enriched during the reprogramming process, where tumor suppressors may be inhibited and oncogenic pathways may be activated. Though recent studies have reported an improved efficiency of reprogramming by genetic manipulations[4] in addition to using the original four factors, they typically make the process even more complex and increase the risk of genetic alterations and tumorigenicity. Thus there is still a tremendous need for a safer, easier and more efficient procedure for human iPSC generation, which would also facilitate identifying and characterizing fundamental mechanisms of reprogramming.

During reprogramming mediated by four factors (4TF) encoded by *OCT4* (*POU5F1*), *SOX2*, *KLF4* and *c-MYC* (*MYC*), mesenchymal-type fibroblasts undergo dramatic morphological changes that result in iPSCs with distinct cell polarity, boundaries and cell-cell interactions. The cells start expressing E-cadherin, a marker for epithelial cells[5], which is also highly expressed in human embryonic stem cells (hESCs). We reasoned that factors that promote the mesenchymal to epithelial transition, such as TGFβ pathway antagonists, would have a direct impact on the reprogramming process. In addition, MEK-ERK pathway inhibition previously had been shown to be important in various reprogramming steps[6,7]. Furthermore, factors promoting cell survival could also be beneficial in improving reprogramming efficiency. Consequently, we focused on small molecules that can regulate these three processes and pathways, as the use of small molecules has many advantages[4,7,8] in studying biological processes and they are a safer choice than genetic manipulation. Here we describe a simple chemical platform that substantially enhances generation of fully reprogrammed human iPSCs from fibroblasts through a much faster and more efficient process.

We tested known inhibitors of TGFβ receptor and MEK on 1×10^4 human primary fibroblasts (CRL2097 or BJ) that we retrovirally transduced with cDNAs encoding the 4TFs, for their effect on reprogramming (**Fig. 1a**). On day 7 after infection, we added the compounds, individually or in combinations, and examined the cultures for iPSCs over the next 1–3 weeks.

On day 7 after treatment (day 14) we observed the strongest effect in the cultures treated with a combination of ALK5 inhibitor SB431542 (2 μM) and MEK inhibitor PD0325901 (0.5 μM), which resulted in ~45 large alkaline phosphatase (ALP)-positive colonies (**Fig. 1b**) with characteristic hESC-like morphology, of which over 24 colonies were TRA-1-81$^+$ (**Fig. 1c**), and six to ten colonies stained positive for SSEA4 in addition to NANOG, a mature pluripotency factor that was not ectopically introduced (**Fig. 1d**,**e**). Moreover, the treated cultures had high expression of endogenous mRNA for the pluripotency genes (**Fig. 1f**). In contrast, we did not observe any NANOG$^+$ colonies in the untreated control cultures or in cultures that we treated with PD0325901 alone (**Fig. 1e** and **Supplementary Fig. 1a**). However, in cultures treated with only SB431542, we observed 1–2 ALP$^+$ hESC-like colonies (**Supplementary Fig. 1a**). Notably, the combined effect of both inhibitors (**Supplementary Fig. 1b**,**c**), as well as the individual effect of SB431542 (data not shown) was dose-dependent.

When we maintained the SB431542 plus PD0325901–treated cultures for 30 d without splitting, we obtained about 135 iPSC colonies per well (**Fig. 1g**), a >100-fold improvement in efficiency over the conventional method. Consistent with previous reports[1], in untreated controls transduced with cDNAs encoding 4TFs, we observed one to two iPSC colonies in addition to several granulate

[1]Department of Chemistry, The Scripps Research Institute, La Jolla, California, USA. [2]Department of Pediatrics, The Whittier Institute for Diabetes, University of California San Diego, La Jolla, California, USA. Correspondence should be addressed to S.D. (sding@scripps.edu).

RECEIVED 10 JULY; ACCEPTED 23 SEPTEMBER; PUBLISHED ONLINE 18 OCTOBER 2009; DOI:10.1038/NMETH.1393

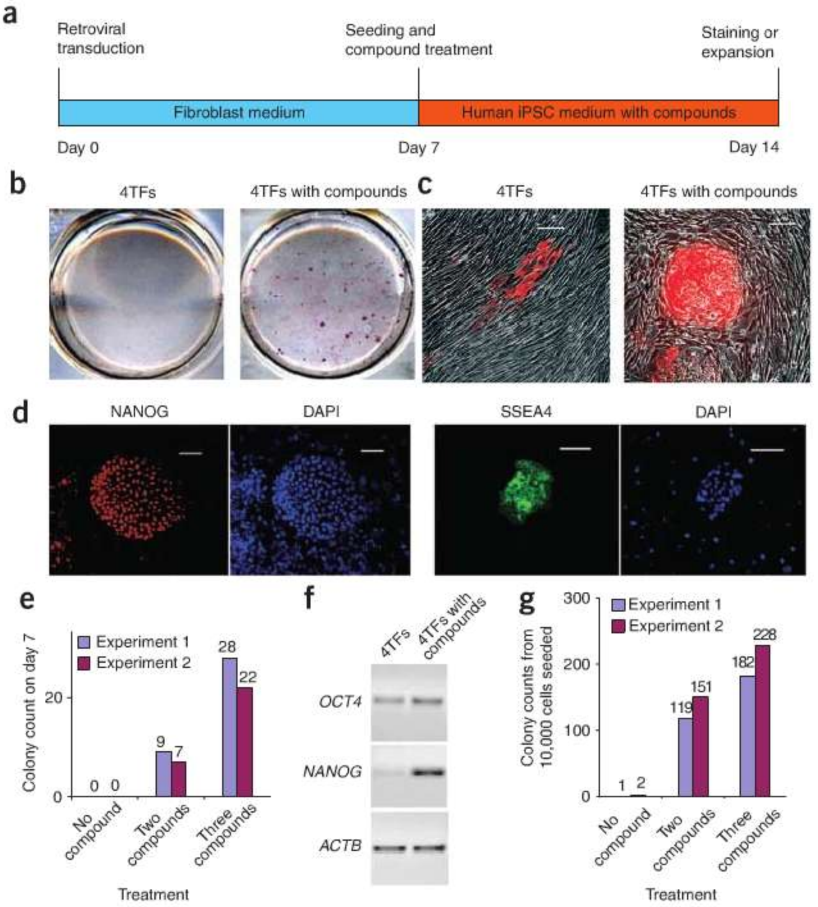


Figure 1 | Compound treatment for 7 d is sufficient to induce pluripotent stem cells from human fibroblasts transduced with the four reprogramming factors. (**a**) Timeline for human iPSC induction using combined SB431542 and PD0325901 treatment along with retroviral transduction of genes encoding 4TF. Treatment with compounds began with cell reseeding at day 7 after transduction and was maintained for 7 d. (**b**) Staining for ALP$^+$ colonies of untreated (left) or two-compound–treated (right) cultures on day 7 of treatment (experiment day 14). (**c**) A representative colony from either the two-compound treated (right) or control cultures (left), on day 14, showing Tra-1-81 immune reactivity. The images shown here is an overlay of phase contrast and fluorescence micrographs from the same field. Note the hESC-like colony morphology only in the treated cultures. (**d**) Fluorescence images of day 14 iPSCs showing hESC-specific protein markers NANOG and SSEA4 and DAPI-stained nuclei from the same field. Scale bars, 50 μm. (**e**) The numbers of NANOG$^+$ colonies on day 14 under the indicated treatment conditions. (**f**) Reverse transcriptase (RT)-PCR analysis of endogenous *OCT4* and *NANOG* mRNA expression in 4TF-expressing samples treated with two compounds or untreated controls. (**g**) Numbers of NANOG$^+$ colonies on day 30 under indicated treatment conditions, without splitting.

colonies after 30 d of culture. These granulate structures have been suggested to be partially reprogrammed colonies[1]. We also observed granulate colonies in SB431542-treated cultures, which outnumbered by several fold the few hESC-like colonies. Notably, the number of granulate colonies was dramatically lower after the combined SB431542 and PD0325901 treatment, which resulted in a concomitant increase in the number of hESC-like colonies. This suggested that a combined inhibition of ALK5 and MEK may guide partially reprogrammed colonies to a fully reprogrammed state, thereby improving the overall reprogramming process. Moreover, that we observed improved induction of iPSCs as early as 7 d after treatment, suggests that treatment with these small molecules not only improves the efficiency of the reprogramming process but may also accelerate its kinetics (**Fig. 1a**), although additional experiments are required to determine whether the reprogrammed cells at this stage indeed become fully independent of exogenous reprogramming factors earlier than in untreated cultures.

Although we could pick and expand the iPSC colonies, as in hESC cultures, splitting the cultures by trypsinization resulted in poor survival. From a recent screen performed in our laboratory using an in-house collection of compounds, we identified a small molecule, thiazovivin (**Supplementary Fig. 2**), which dramatically improves the survival of hESCs upon trypsinization (X.Y. and S.D.; unpublished data). Addition of thiazovivin to our cocktail of SB431542 and PD0325901 also improved the survival of iPSCs after splitting by trypsinization (**Fig. 2a**) and meant that we could obtain many reprogrammed colonies. From 10,000 cells that we originally seeded, a single 1:4 splitting on day 14 resulted in ~1,000 hESC-like colonies on day 30 (**Fig. 2b**), and two rounds of splitting (on day 14 and on day 21 (1:10)) resulted in ~11,000 hESC-like colonies (**Fig. 2b,c**) on day 30. These colonies had high levels of endogenous mRNA (**Fig. 2d**) and protein expression (**Fig. 2c,e**) of pluripotency markers, whereas the expression of the four transgenes could hardly be detected (**Fig. 2d**). In contrast, we did not obtain any iPSC colonies from untreated or two compound–treated samples that we trypsinized (**Supplementary Table 1**).

To examine whether the enhancement in reprogramming observed in thiazovivin-treated cultures was solely due to survival of colonies after splitting or whether it also augmented the reprogramming effect of combined SB431542 and PD0325901 treatment, we tested the three-compound cocktail, without cell splitting, on cells transduced with cDNAs encoding the 4TFs. In these cultures, by day 14, we observed ~25 large colonies that all expressed NANOG (**Fig. 1e**). By day 30 we observed ~205 very large NANOG$^+$ colonies (**Fig. 1g**), that were also TRA-1-81$^+$ and SSEA4$^+$ (data not shown), which translates to a more than 200-fold improvement in reprogramming efficiency over the no-compound treatment and a twofold increase over the two-compound treatment.

Two-compound treatment also resulted in more ALP$^+$ colonies compared to untreated controls when the reprogramming factors were introduced using a lentiviral, rather than a retroviral system (**Supplementary Fig. 3a**). Furthermore, the three-compound cocktail did not appear to influence reprogramming factor expression from retroviral vectors (**Supplementary Fig. 3b–f**).

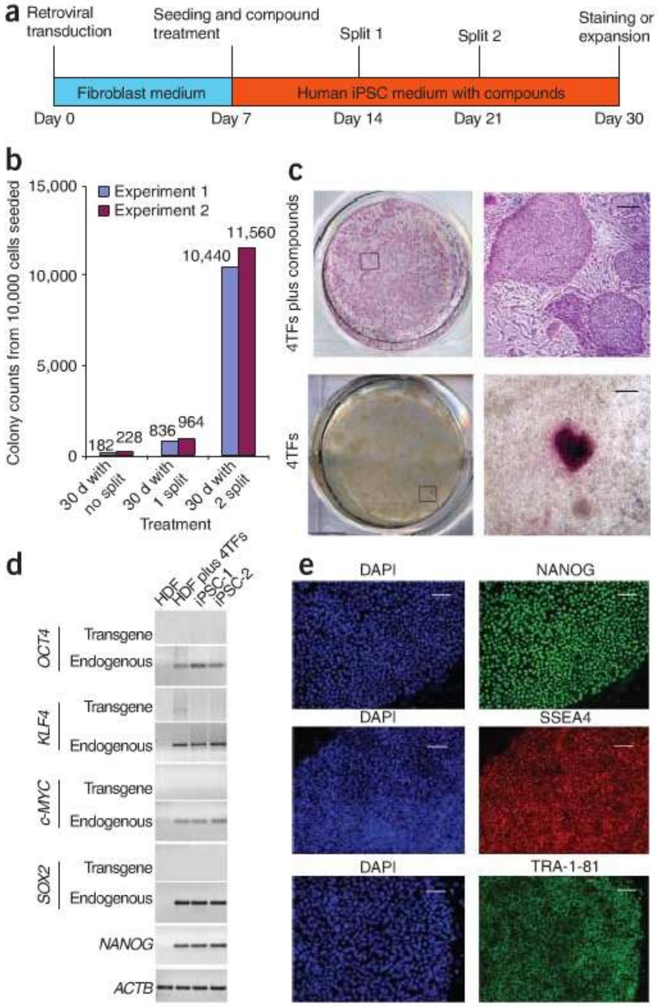


Figure 2 | Prolonged compound treatment and cell passaging dramatically increased the number of reprogrammed colonies. (**a**) Timeline of human iPSC induction using SB431542, PD0325901 and thiazovivin. (**b**) Numbers of NANOG$^+$ colonies on day 30 from three compound–treated cultures trypsinized as indicated. (**c**) ALP staining of day 30 cultures with (top) or without (bottom) three-compound treatment. Boxed areas in images on the left are shown on the right. Scale bars, 200 μm. (**d**) RT-PCR analysis of cDNA from iPSC colonies obtained with three-compound treatment showing reactivated mRNA expression of endogenous pluripotency markers. HDF, human dermal fibroblast. iPSC-1 and 2 refers to independent iPSC colonies. (**e**) Immunofluorescence images of day 30 iPSCs treated with three compounds showing protein expression of pluripotency markers NANOG, SSEA4 and TRA-1-81. Scale bars, 50 μm.

The iPSC colonies generated using the three-compound cocktail could be readily and stably expanded for long term under conventional hESC culture conditions (over 20 passages), and they closely resembled hESCs in terms of morphology, typical pluripotency marker expression and differentiation potentials. They exhibited a normal karyotype (**Supplementary Fig. 4**) and could be differentiated into derivatives of all three germ layers, both *in vitro* (**Fig. 3a,b**) and *in vivo* (**Fig. 3c**). These results also suggest that there was no short-term adverse effect associated with the use of the much more convenient trypsinization procedure, although we cannot entirely rule out long-term effects.

The finding that TGFβ and MEK-ERK pathway inhibition improved fibroblast reprogramming suggests critical roles for these two signaling pathways and mesenchymal to epithelial transition mechanisms in the process. Consistently, the addition of TGFβ has an inhibitory effect on four-factor–mediated reprogramming of fibroblasts (data not shown). TGFβ and its family members have important roles in self-renewal and differentiation of embryonic stem cells[9]. Moreover, TGFβ is a prototypical cytokine for induction of epithelial-mesenchymal transition and maintenance of the mesenchymal state[10]. A major end point of this signaling, in this context, is downregulation of E-cadherin[11]. E-cadherin has been shown to be important for the maintenance of pluripotency of embryonic stem cells and has been recently suggested to be a regulator of *NANOG* expression[12]. Therefore inhibition of TGFβ signaling, which results in de-repression of epithelial fate, could benefit the reprogramming process in multiple ways. ERK signaling also promotes epithelial-mesenchymal transition[11] and is downstream of TGFβ in the process[12]. We had previously shown that the effect of reversine, a small molecule that can reprogram myoblasts to a multipotent state, is mediated in part through inhibition of MEK-ERK[6]. This may explain the effect observed in reprogramming when it was combined with TGFβ inhibition.

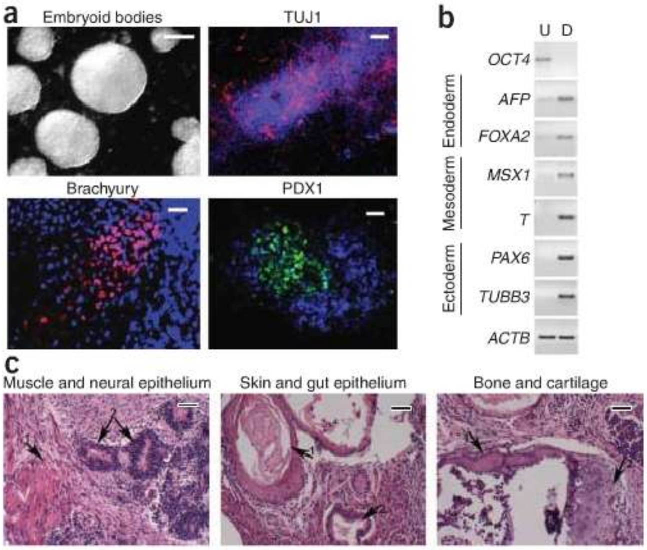


Figure 3 | *In vitro* and *in vivo* differentiation of iPSCs generated with three-compound treatment. (**a**) Micrographs show embryoid bodies generated from iPSCs and their *in vitro* differentiation into ectodermal, mesodermal and endodermal cell types, revealed by immunoreactivity to typical markers βIII Tubulin (TUJ1), Brachyury and PDX1, respectively. Scale bars, 100 μm (embryoid bodies) and 10 μm (all others). (**b**) RT-PCR analysis of cDNA from iPSC colonies showing mRNA expression of representative lineage markers and the absence of *OCT4* expression in differentiating cells. U, undifferentiated; D, differentiated. (**c**) Histological sections from teratomas generated in nude mice from iPSCs (3 independent colonies tested) consist of tissues from all three germ layers. Left: 1, muscle; 2, neural epithelium. Middle: 1, skin; 2, gut epithelium. Right: 1, bone; 2, cartilage. Scale bars, 20 μm.

Concurrent with this study, two other studies reported an improvement in reprogramming efficiency in human cells, either by epigenetic manipulation[13] or by reprogramming keratinocytes[14]. However, the chemical platform described here is unique, in that it modulates upstream signaling pathways and could radically improve reprogramming on a general cell type, like fibroblasts. The chemical conditions described here could provide a basic platform on which a non-viral and non-DNA-based[15], more efficient and safer reprogramming process could be developed, which could yield an unlimited supply of safer human iPSCs for various applications.

METHODS

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

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

ACKNOWLEDGMENTS

This work was supported by funding from US National Institutes of Health and Fate Therapeutics to S.D. We thank C.V. Wright (Vanderbilt University medical center) for the gift of Pdx1 antibody, R. Coleman for the assistance in the preparation of the manuscript and all the members of Ding lab for helpful discussions.

AUTHOR CONTRIBUTIONS

T.L. and S.D. initiated the study. S.D., R. Am. and T.L. made the hypothesis, designed the experiments and wrote the manuscript. T.L., R. An, X.Y., H.S.H., and W.L. conducted the experiments. S.H., R. Ab. and X.L. provided assistance in some of the experiments. E.H. generated teratomas, and A.H. supervised E.H. S.D. supervised the study.

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

1. Takahashi, K. *et al. Cell* **131**, 861–872 (2007).
2. Yu, J. *et al. Science* **318**, 1917–1920 (2007).
3. Muller, L.U.W., Daley, G.Q. & Williams, D.A. *Mol. Ther.* **17**, 947–953 (2009).
4. Feng, B., Ng, J.H., Heng, J.C.D. & Ng, H.H. *Cell Stem Cell* **4**, 301–312 (2009).
5. Hay, E.D. *Acta Anat.* **154**, 8–20 (1995).
6. Chen, S. *et al. Proc. Natl. Acad. Sci. USA* **104**, 10482–10487 (2007).
7. Shi, Y. *et al. Cell Stem Cell* **2**, 525–528 (2008).
8. Xu, Y., Shi, Y. & Ding, S. *Nature* **453**, 338–344 (2008).
9. Watabe, T. & Miyazono, K. *Cell Res.* **19**, 103–115 (2009).
10. Willis, B.C. & Borok, Z. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **293**, L525–L534 (2007).
11. Thiery, J.P. & Sleeman, J.P. *Nat. Rev. Mol. Cell Biol.* **7**, 131–142 (2006).
12. Chou, Y.F. *et al. Cell* **135**, 449–461 (2008).
13. Huangfu, D. *et al. Nat. Biotechnol.* **26**, 1269–1275 (2008).
14. Aasen, T. *et al. Nat. Biotechnol.* **26**, 1276–1284 (2008).
15. Zhou, H. *et al. Cell Stem Cell* **4**, 381–384 (2009).

Enrichment of glycopeptides for glycan structure and attachment site identification

Jonas Nilsson[1], Ulla Rüetschi[1], Adnan Halim[1], Camilla Hesse[1], Elisabet Carlsohn[2], Gunnar Brinkmalm[3] & Göran Larson[1]

We present a method to enrich for glycoproteins from proteomic samples. Sialylated glycoproteins were selectively periodate-oxidized, captured on hydrazide beads, trypsinized and released by acid hydrolysis of sialic acid glycosidic bonds. Mass spectrometric fragment analysis allowed identification of glycan structures, and additional fragmentation of deglycosylated ions yielded peptide sequence information, which allowed glycan attachment site and protein identification. We identified 36 N-linked and 44 O-linked glycosylation sites on glycoproteins from human cerebrospinal fluid.

Glycosylation is the most frequent and most complex post-translational modification of proteins[1]. Glycans are predominantly glycosidically N-linked to asparagine or O-linked to serine and threonine; these N- and O-linked glycans are often terminated with sialic acids[2]. Correct protein glycosylation is important in a wealth of biological processes including protein folding, intracellular sorting, secretion, uptake, and cell and host-microbial recognition[3,4]. Aberrant glycosylation of N- and O-linked glycoproteins in cancer cells has an important role in tumor growth and metastasis, and glycan analysis may be used as a diagnostic tool[5]. Little is known, however, about the site-specific glycosylation of proteins on a proteomic scale, and for this purpose new glycoproteomic techniques are needed to address the relationship between site-specific glycosylations and their biological function.

N- and O-linked glycans released by PNGase F treatment or by beta-elimination, respectively, can be profiled by mass spectrometry or by fluorescence detection after chromatographic separation[6]. Mild periodate oxidation and oxime coupling to probes has recently been introduced for the visualization of sialic acid on living cells[7]. However, these techniques do not provide any information regarding glycan attachment sites or glycan diversity within or between different glycoproteins in a mixture. Capture of periodate-oxidized glycoproteins on hydrazide beads has previously been used for high-throughput mapping of protein N-linked glycan sites[8,9]. With this approach, formerly N-glycosylated peptides are released by PNGase F treatment, and then the glycosylation sites can be assayed. However, no information regarding glycan structures is made available, and PNGase F treatment works only for N-linked glycoproteins.

It is therefore important to develop methods to analyze both the glycan structure and its protein attachment site. For this purpose we developed a glycopeptide capture-and-release method for the selective enrichment of sialic acid–containing glycopeptides (**Fig. 1a**). The strategy is based on the fact that sialic acids can be selectively oxidized by mild periodate oxidation and then captured onto hydrazide beads[8]. The captured glycoproteins are then subjected to trypsin digestion, such that only the glycopeptides remain attached to the beads. The glycosidic bond between the terminal sialic acid and the penultimate monosaccharide is sensitive to mild acid hydrolysis; thus captured glycopeptides can be selectively released by acid hydrolysis. The released glycopeptides can then be analyzed by mass spectrometry; the putative glycan structures of both N- and O-linked glycopeptides can be determined in parallel with peptide identification and in many cases identification of their glycan attachment sites.

As an initial test with a model glycoprotein, we oxidized human transferrin with 2 mM periodate at 0 °C and captured it on hydrazide beads. After trypsin treatment we removed the supernatant, containing the trypsin-released peptides, from the beads, and then cleaved the immobilized glycopeptides from the resin with 0.1 M formic acid at 80 °C for 1 h. (**Fig. 1a**). We tested the efficiency of the capture-and-release method (**Supplementary Fig. 1**). For mass spectrometry analysis, we used reversed-phase liquid chromatography coupled to an electrospray ionization (ESI) linear ion trap/Fourier transform ion cyclotron resonance (FTICR) instrument. The total-ion current chromatogram (**Fig. 1b**) demonstrated the presence of distinct chromatographic peaks composed of ions that matched the masses of Asn432- and Asn630-containing tryptic peptides plus the mass of five hexoses and four *N*-acetylhexosamines ($Hex_5HexNAc_4$).

The low-energy collision-induced dissociation fragment spectrum (MS^2) of the mass to charge ratio (*m/z*) 1,374.9 peak representing the Asn630-containing peptide (**Fig. 1c**) and the subsequent spectrum, in which the second-generation fragments (MS^3) are acquired (**Fig. 1d**), show that the glycans were composed of the complex biantennary motif in accordance with published findings[10] (**Supplementary Fig. 2**). The peptide ion

[1]Department of Clinical Chemistry and Transfusion Medicine, Institute of Biomedicine, University of Gothenburg, Sahlgrenska University Hospital, Gothenburg, Sweden. [2]The Proteomics Core Facility, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden. [3]Department of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology, University of Gothenburg, Sahlgrenska University Hospital, Mölndal, Sweden. Correspondence should be addressed to J.N. (jonas.nilsson@clinchem.gu.se).

RECEIVED 15 JULY; ACCEPTED 4 SEPTEMBER; PUBLISHED ONLINE 18 OCTOBER 2009; DOI:10.1038/NMETH.1392

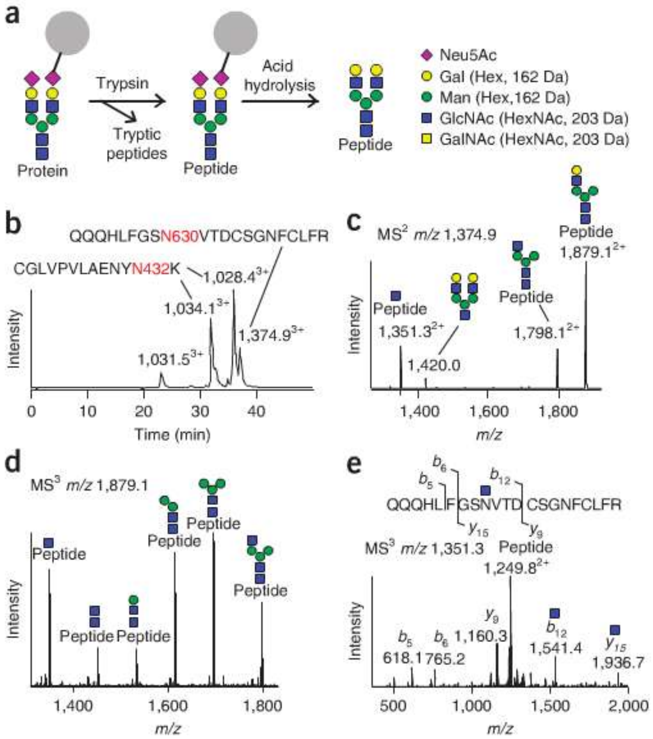


Figure 1 | Capture-and-release and mass spectrometry analyses of enriched glycopeptides. (**a**) Sialylated glycoproteins were periodate-oxidized, captured on hydrazide beads (gray) and trypsin-digested. Tryptic peptides from the captured glycoproteins were released into the supernatant. Glycopeptides were released from the beads by mild formic acid hydrolysis, which selectively cleaves off sialic acids. Neu5Ac, sialic acid; Gal, galactose; Man, mannose; GlcNAc, *N*-acetylglucosamine; GalNAc; *N*-acetylgalactosamine; Hex, hexose; HexNAc, *N*-acetylhexosamine. (**b**) Reversed phase liquid chromatography–ESI-FTICR ion-current chromatogram of enriched transferrin glycopeptides. All ions shown are in the $[M + zH]^{z+}$ form and charges are shown in superscript when $z > 1$. The N-terminal asparagine of the Asn630-containing glycopeptide (*m/z*, 1,374.9) was deaminated. The N-terminal of the iodoacetamide cysteine of the Asn432-containing glycopeptide was deaminated (*m/z* 1,028.4) or in native form (*m/z* 1,034.1). (**c**) MS^2 fragmentation of the Asn630-containing glycopeptide (*m/z* 1,374.9) showed glycosidic fragmentation. (**d**) Fragmentation (MS^3) at *m/z* 1,879.1 (in **c**) was compatible with a complex biantennary structure. (**e**) MS^3 at *m/z* 1,351.3 (in **c**, peptide ion with glycosidically linked HexNAc, Pep+HexNAc) gave peptide fragmentation. Apart from the biantennary glycan, triantennary and fucosylated biantennary glycans were also observed for the Asn432- and Asn630-containing glycopeptides. The elution peaks for the individual parent ions were integrated, and on assumption of equivalent enrichment and ionization efficiencies for different glycoforms, the relative intensities were found to be 98% biantennary, 0.5% fucosylated biantennary and 1.5% triantennary for the Asn432-containing glycopeptide. The Asn630-containing glycopeptide was found to be 97% biantennary, 2.6% fucosylated biantennary and 0.4% triantennary. The absolute identity of monosaccharides in glycan and glycopeptide fragment ions was putatively assigned because saccharide isomerism cannot be addressed in the collision-induced dissociation fragmentation of glycopeptides.

with glycosidically linked HexNAc (Pep+HexNAc) at *m/z* 1,351.3 was present in the MS^2 spectrum, and MS^3 analysis of this ion resulted in peptide fragmentation (**Fig. 1e**), which we used in database searching for peptide identification. The *m/z* 1,031.5 peak (**Fig. 1b**) originated from the Asn162-containing tryptic peptide from beta-2-glycoprotein 1 with a complex biantennary glycan, a contaminant in the transferrin sample. Complex triantennary, fucosylated bi- and triantennery glycans were major components, eluting simultaneously with *m/z* 1,031.5. We analyzed the fragmentation pattern to identify whether fucose was attached to the core or the antenna of the N-linked glycan (**Supplementary Fig. 3**). We also captured and characterized O-linked glycopeptides from bovine fetuin, another model glycoprotein (**Supplementary Fig. 4**).

We then applied the sialic acid capture-and-release strategy to analyze glycoproteins in three human cerebrospinal fluid (CSF) samples. We analyzed the supernatant fraction, which mainly contains nonglycosylated trypsin-released peptides from captured glycoproteins, by liquid chromatography–MS^2 shotgun analysis and identified 84 proteins by database searching (**Supplementary Table 1**). For the acid-released glycopeptide fractions (**Supplementary Fig. 5**), we estimated ~80% of the MS^2 fragmented precursor ions to be glycopeptides owing to the presence of the diagnostic ion at *m/z* 366, $[HexHexNAc+H]^+$, and because of the characteristic fragmentation patterns at glycosidic linkages.

We characterized N-linked glycopeptides with various glycoforms and tentatively quantified the relative glycoform abundance for each N-linked glycopeptide. The glycan structure has been reported previously only for a few of these CSF proteins[11] (**Supplementary Table 2**). The Pep+HexNAc ion was frequently present in the MS^2 spectra and was selected for MS^3 fragmentation. Mascot searches of the Pep+HexNAc fragment spectra yielded a list of 36 unique N-linked glycosylation sites from 23 glycoproteins, of which all have been reported previously[9].

The only O-linked glycan we detected from CSF was HexHexNAc-*O*-Ser/Thr, which is compatible with the core 1 structure Galβ3GalNAcα1-*O*-Ser/Thr (**Fig. 2**), but could, in principle, also be core 8 structure (Galα3GalNAcα1-*O*-Ser/Thr) for some of these O-linked glycosylation sites. Glycopeptides that contained $Hex_2HexNAc_2$ could be composed of a HexHexNAc elongated core 1 (for example, Galβ4GlcNAcβ3Galβ3GalNAcα1-*O*-Ser/Thr) or core 2 (Galβ4GlcNAcβ6(Galβ3)GalNAcα1-*O*-Ser/Thr) structures, but we found that they were composed of two separate HexHexNAc glycans on different serine or threonine residues because of the sequential loss of two terminal hexose units in the fragmentation spectra (**Supplementary Fig. 6**) and a lack of glycan fragments exceeding *m/z* 366. MS^2 of tryptic O-linked glycopeptides resulted in characteristic fragmentation patterns in which the Pep+HexNAc and peptide ions were the main peaks (**Fig. 2**).

MS^3 of peptide ions yielded peptide fragmentation patterns; a Mascot database search of those spectra yielded 44 different O-linked glycosylation sites from 22 different proteins (**Supplementary Table 3**). Only some of these glycosylation sites have been reported previously. In some cases we deduced the serine or threonine glycosylation site based on peptide fragmentation in the presence of an intact glycan in the sequence (**Fig. 2**). We noted that a Ser/Thr-X-X-Pro motif, in which X is any amino acid, was present in all but seven of the O-linked glycopeptides; this 'consensus' sequence was consistent with that reported in literature[12]. The brain-specific polypeptide GalNAc transferase GalNAc-T13 has been reported to glycosylate serine or threonine, followed by proline positioned three amino acids away in the sequence of model peptides[13] and may be responsible for the Ser/Thr-X-X-Pro glycosylation patterns we identified for CSF

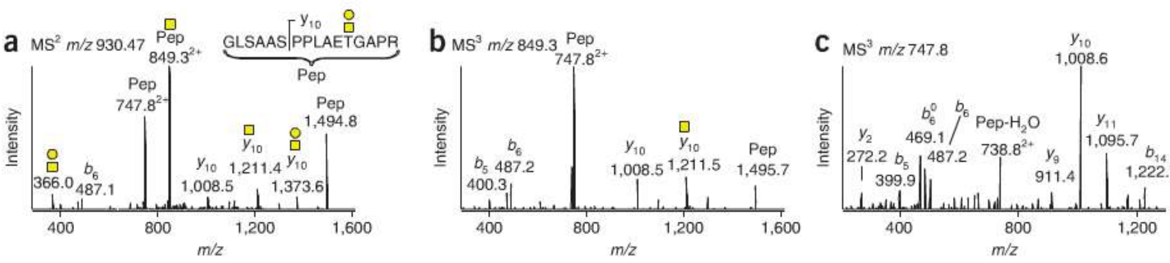


Figure 2 | Fragment mass spectra of the tryptic O-glycosylated SAAS peptide from ProSAAS. (**a**) MS^2 results in loss of hexose (*m/z* 849.3, Pep+HexNAc) and HexHexNAc (*m/z* 747.8, peptide (pep)). (**b**) MS^3 of Pep+HexNAc gave peptide fragmentation, with an intact HexNAc, which was used to pinpoint the glycan attachment site. (**c**) MS^3 of the peptide ion gave peptide fragmentation which was used to identify the protein. Monosaccharide symbols are explained in **Figure 1**. All ions are $[M + zH]^{z+}$ and their charges are shown as superscripts when $z > 1$.

O-linked glycoproteins. Notable examples of new O-linked glycosylations that we found in CSF included a glycan site on the C-terminal tryptic peptide of apolipoprotein E (**Supplementary Fig. 7**) and two sites on ProSAAS, which resided on the SAAS (**Fig. 2**) and LEN peptides (UniProtKB/Swiss-Prot database: Q9UHG2). We also identified two O-linked glycans on a tryptic peptide from alpha-dystroglycan, which depends on glycan interactions with laminin for its correct function[1]. Human Cystatin C is generally believed to be nonglycosylated, but we found a HexHexNAc glycan on a B/B haplotype–specific N-terminal peptide. The B/B haplotype has an A25T mutation, which normally is part of the signal peptide, but was still present in the identified glycopeptide. The effects on protein processing and other biological implications of these O-linked glycosylations remain to be investigated. Electron capture dissociation (ECD) and electron transfer dissociation fragmentation techniques for glycopeptide analysis with intact glycans[14] (**Supplementary Fig. 8**) are promising alternatives to precisely map O-linked glycan attachment sites.

The sialic acid capture-and-release strategy presented here is a selective and simple way to purify both N- and O-linked glycopeptides from complex biological mixtures. Although CSF has a low protein concentration (approximately 0.5 mg ml^{-1}), and two-thirds of the protein mass is nonglycosylated albumin, our enrichment method allowed us to attain a high degree of glycopeptide purity, facilitating mass spectrometry analyses. The only inadvertent modifications of glycopeptides we detected were the occasional presence of periodate-oxidized sialic acids (**Supplementary Fig. 7**) and hydrolysis of the Asp-Pro peptide bond, which was due to the formic acid treatment (**Supplementary Table 3**).

Our method has some important limitations. Sialic acids are necessary for capture onto beads; thus, non–sialic acid–terminated glycopeptides cannot be enriched. Additionally, because sialic acids are removed in the release step of the protocol, our technique will not reveal the amount of sialylation of individual glycans but selectively only their core saccharide structures. However, for analytical purposes, the removal of sialic acid is beneficial because the mass spectrometry fragmentation patterns are simplified and the peptide ions are more likely to be automatically chosen for secondary fragmentation. Lectin-based strategies can be used to enrich sialylated and mannosylated glycopeptides from model glycoproteins[15]. However, at least for complex biological samples, covalent capture techniques are preferable for efficient glycopeptide enrichment because harsh washing procedures of the solid phase are required to remove the vast majority of nonglycosylated peptides, which otherwise would dominate the ion chromatograms. Our method yields qualitative information of N- and O-linked glycosylation core structures and their attachment sites in sialylated glycoproteins in a proteomic mixture.

METHODS

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

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

ACKNOWLEDGMENTS

The work was supported by grants from the Swedish Research Council (8266), governmental grants to the Sahlgrenska University Hospital, Torsten and Ragnar Söderbergs Foundation, the Inga-Britt and Arne Lundberg Research Foundation, Knut and Alice Wallenberg Foundation and cNeupro.

AUTHOR CONTRIBUTIONS

J.N. conceived and designed the method, performed experiments, analyzed data and wrote the manuscript; U.R. designed the mass spectrometry setup, performed experiments, analyzed data and wrote the manuscript; A.H. performed experiments, analyzed data and wrote the manuscript; C.H. performed experiments and analyzed data; E.C. and G.B. designed the mass spectrometry setup and performed experiments; and G.L. designed the method and wrote the manuscript.

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

1. Varki, A. *et al. Essentials of Glycobiology* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA, 2009).
2. Varki, A. *Nature* **446**, 1023–1029 (2007).
3. Grewal, P.K. *et al. Nat. Med.* **14**, 648–655 (2008).
4. Ohtsubo, K. & Marth, J.D. *Cell* **126**, 855–867 (2006).
5. Fujimura, T. *et al. Int. J. Cancer* **122**, 39–49 (2008).
6. Wada, Y. *et al. Glycobiology* **17**, 411–422 (2007).
7. Zeng, Y., Ramya, T.N.C., Dirksen, A., Dawson, P.E. & Paulson, J.C. *Nat. Methods* **6**, 207–209 (2009).
8. Zhang, H., Li, X.J., Martin, D.B. & Aebersold, R. *Nat. Biotechnol.* **21**, 660–666 (2003).
9. Hwang, H. *et al. Mass Spectrom. Rev.* advance online publication, doi:10.1002/mas.20221 (8 April 2009).
10. Satomi, Y., Shimonishi, Y., Hase, T. & Takao, T. *Rapid Commun. Mass Spectrom.* **18**, 2983–2988 (2004).
11. Sihlbom, C., Davidsson, P., Sjögren, M., Wahlund, L.O. & Nilsson, C.L. *Neurochem. Res.* **7**, 1332–1340 (2008).
12. Wilson, I.B., Gavel, Y. & von Heijne, G. *Biochem. J.* **275**, 529–534 (1991).
13. Zhang, Y. *et al. J. Biol. Chem.* **278**, 573–584 (2003).
14. Alley, W.R., Mechref, Y. & Novotny, M.V. *Rapid Commun. Mass Spectrom.* **23**, 161–170 (2009).
15. Kubota, K. *et al. Anal. Chem.* **80**, 3693–3698 (2008).

A general life-death selection strategy for dissecting protein functions

Po Hien Ear & Stephen W Michnick

Clonal selection strategies are central tools in molecular biology. We developed a general strategy to dissect protein functions through positive and negative clonal selection for protein-protein interactions, based on a protein-fragment complementation assay using *Saccharomyces cerevisiae* cytosine deaminase as a reporter. We applied this method to mutational or chemical disruption of protein-protein interactions in yeast and to dissection of the functions of an allosterically activated transcription factor, Swi6.

Proteins can perform different functions, in part, through sets of distinct interactions with other proteins. Thus, a protein's functions can be dissected by selectively disrupting individual interactions[1]. To achieve such fine dissection of protein-protein interactions, we designed a simple positive and negative clonal selection strategy based on a protein-fragment complementation assay (PCA)[2,3] with a prodrug-converting enzyme as reporter (**Fig. 1a**). We chose the yeast *Saccharomyces cerevisiae* cytosine deaminase (yCD) as the reporter protein because life and death selection assays have been established for this enzyme in a broad spectrum of cells including those of bacteria[4], yeast[5] and mammals[6].

Conceptually, a yCD PCA could be used to dissect protein-protein interactions as follows. In yeast, deletion of the *FCY1* gene encoding yCD renders the strain defective for the pyrimidine salvage pathway because its capacity to convert cytosine to uracil is lost; thus cells that cannot synthesize uracil by the *de novo* pyrimidine pathway cannot grow in the absence of uracil (**Fig. 1b**). In the life selection assay, the interaction of proteins *X* and *Y* brings complementary fragments of yCD into proximity, allowing them to fold and reconstitute its catalytic activity. Complementing an *FCY1* knockout strain with yCD PCA would restore cell growth. Death selection is achieved when the same yCD-complemented strain is treated with 5-fluorocytosine (5-FC), a nontoxic compound that is converted to toxic 5-fluorouridine triphosphate (5-FUTP) in a pathway that depends on yCD activity[7].

We could combine these selection assays to dissect binary interactions between proteins. For instance, if protein X interacts with both protein Y and a third partner Z, and we wish to disrupt the X-Z, but retain the X-Y interaction, we could first perform the death selection yCD PCA with a library of mutants X* screened against Z and select for growth (disruption of the X-Z interaction) and then perform the life selection yCD PCA with mutants of X* against Y, selecting clones that grow (positive for the X*-Y interaction). The key strength of this approach is that, as both life and death selection result in a growth phenotype, identifying clones in both steps would be trivial compared to just using a life selection assay that would require an additional step for replicating clones on a control plate to recover mutants of interest.

To identify appropriate PCA fragments, we tested seven different combinations of yCD fragments (**Fig. 1c**) each fused downstream of the coding sequence of homodimerizing residues (250–281) of the GCN4 parallel coiled-coil leucine zipper (zip) via a sequence coding for a 15-amino-acid flexible linker peptide. We identified complementary N- and C-terminal fragments (referred to as F[1] and F[2]) with the highest yCD PCA activity consisting of residues 1–77 and 57–158 (yCD-F[1]1–77 and yCD-F[2]57–158) (**Fig. 1d**). To improve the activity of the yCD PCA, we first introduced into these yCD fragment fusions three mutations that had been shown to increase thermostability of full-length yCD[8] and found that this improved the activity of yCD PCA (**Supplementary Fig. 1**).

However, the observed improvement for the leucine zipper interaction was not sufficient to detect other interactions such as that between the human small GTPase H-Ras (Ras) and the Ras binding domain (RBD) of the serine/threonine kinase c-Raf (**Fig. 1e**). Thus, using the Ras-RBD interaction as a test system, we attempted to generate fragments that result in improved yCD PCA activity. We generated libraries of randomly mutated yCD-F[1]1–77 and yCD-F[2]57–158 by error-prone PCR (with 1–2 mutations per fragment) and fused them to the interacting partners, Ras and RBD (**Supplementary Fig. 2**). We screened the mutant fragment–Ras and –RBD fusions against each other and identified mutants that additionally increased yCD PCA activity (**Supplementary Table 1**). Among the clones collected, the mutations T95S and K117E in yCD-F[2]57–158, in combination with the three mutations that increased thermostability, showed the greatest sensitivity to 5-FC (highest yCD activity) (**Fig. 1e** and **Supplementary Table 2**). These optimized fragments are called henceforth, OyCD-F[1] and OyCD-F[2]. Mutations T95S and K117E (**Fig. 1f**) may improve activity by creating a salt bridge between Glu117 and wild-type Arg125 of the adjacent subunit (**Supplementary Fig. 3**). OyCD PCA activity did not result from spontaneous complementation of the fragments, as expression of proteins of interest fused to OyCD-F[1] with OyCD-F[2] alone resulted in no observable OyCD activity (**Supplementary Fig. 4**).

We next tested OyCD PCA sensitivity for detecting changes in dissociation constants and quantities of protein complexes. First,

Département de Biochimie, Université de Montréal, Montréal, Québec, Canada. Correspondence should be addressed to S.W.M. (stephen.michnick@umontreal.ca).

RECEIVED 5 MAY; ACCEPTED 18 AUGUST; PUBLISHED ONLINE 11 OCTOBER 2009; DOI:10.1038/NMETH.1389

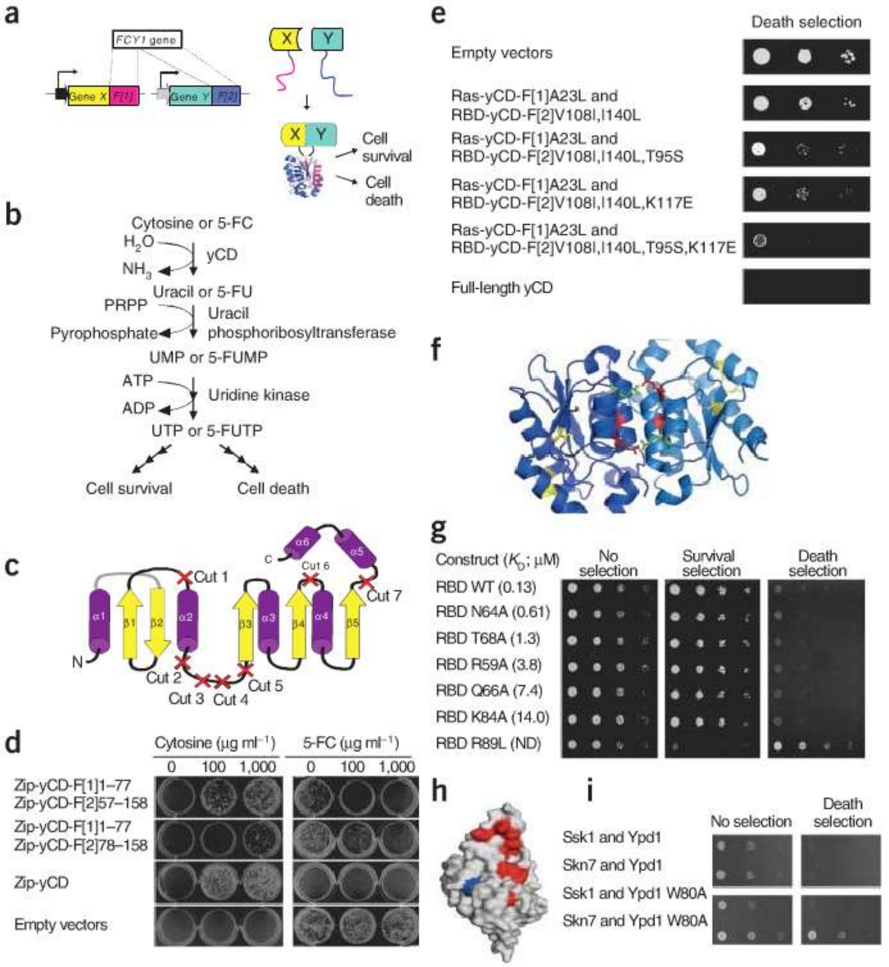


Figure 1 | Development and characterization of the OyCD PCA. (**a**) Two *FCY1* fragments, *F[1]* and *F[2]* were each fused to genes encoding one of two interacting proteins. This allows cell survival or cell death selection. (**b**) Enzymes of the pyrimidine salvage pathway. Both cytosine and 5-FC are substrates for yCD. (**c**) yCD schematic with the 7 cut sites indicated. (**d**) Life and death yCD PCA, on growth medium containing the indicated amounts of cytosine or 5-FC, for different fragment combinations (1–77 and 57–158 in comparison to 1–77 and 78–158), fused to zip peptides. In all PCAs, yeast colonies shown are from tenfold serial dilutions of starting material beginning with 10,000 cells. (**e**) Optimized yCD (OyCD) PCA tested by fusion of yCD fragments to RBD of c-Raf. Introduced mutations in yCD are indicated. (**f**) A model of the OyCD structure with optimizing T95S and K117E mutations (red), Arg125 (green) and increased thermostability-conferring mutations (yellow) based on yCD structure (Protein Data Bank: 1YSB). (**g**) Life (50 μg ml^{-1} cytosine) and death (50 μg ml^{-1} 5-FC) selection OyCD PCA for monitoring Ras interactions with mutant RBDs. Cells expressed OyCD-F[1]-Ras and RBD-OyCD-F[2] with wild-type (WT) RBD or the indicated RBD construct. ND, not determined. (**h**) A model of Ypd1 showing residues that mediate interactions with Skn7, Ssk1 and Sln1 (red) and showing Trp80 (ref. 10) (blue), which mediates Ypd1-Skn7 but not Ypd1-Ssk1 interactions. (**i**) OyCD PCA for interactions of Ypd1 (wild-type and W80A mutant) with Skn7 and Ssk1, using death selection (100 μg ml^{-1} 5-FC).

we tested the Ras-RBD interaction using binding mutants of the RBD with known dissociation constants[9] and found that we could detect Ras interactions with the wild-type RBD and with mutants of RBD with a dissociation constant (K_D) up to 14 μM (**Fig. 1g**). We then measured competitive and stoichiometric disruption of the homodimerizing mutant of the peptidyl-prolyl isomerase FKBP12 (FM1) by the high-affinity binding macrolide FK506 (**Supplementary Fig. 5**). The OyCD PCA death assay was sensitive enough to distinguish a threefold decrease in the number of FM1 homomers, caused by a twofold increase in FK506 concentration. This difference corresponds to approximately the change in the number of complexes that would be caused by a single disruptive mutation of a protein-protein interface, suggesting that conditions can be found for detecting changes in binding affinity of mutants.

To demonstrate that OyCD PCA could be used to dissect protein-protein interactions of a protein with different partners, we studied the interaction between Ypd1, a histidine-containing phosphotransfer protein required for signaling osmotic stress in yeast, and its response regulator proteins Skn7 and Ssk1. Ypd1 has a common binding domain for its regulatory proteins, but the interactions have been shown to be mediated in part through distinct residues[10] (**Fig. 1h**). Trp80 is known to mediate the specific interaction between Ypd1 and Skn7 but not between Ypd1 and Ssk1. We could clearly observe specific disruption of the interaction between Ypd1 W80A mutant and Skn7 while the Ypd1 W80A-SSk1 interaction was retained (**Fig. 1i**).

As proof of principle of a functional dissection of interacting proteins, we devised a strategy to dissect the transcriptional regulation of the *S. cerevisiae* transcription factor Swi6. Swi6 interacts with Mbp1 or Swi4 to form MBF or SBF transcription factor complexes, respectively, which regulate the expression of genes that control the G1- to S-phase transition of the yeast cell cycle[11] (**Fig. 2a**). Swi6 is a modular protein that contains two transcriptional activation regions (N- and C-TAR), an ankyrin repeat domain (AnkRD) and a C-terminal heterodimerizing domain (BD) that can interact with unique C-terminal Swi6 binding domains of Mbp1 (mBD) or Swi4 (sBD)[11] (**Fig. 2b**). To determine whether Swi6 binds to Mbp1 and Swi4 in distinct ways and therefore could differentially regulate MBF or SBF activities, we created a three-step screening strategy to identify Swi6 mutants for which MBF activity was lost but SBF activity was retained.

We first determined that MBF and SBF complexes could be detected by OyCD PCA (**Supplementary Fig. 6**). We then generated a library of full-length Swi6 mutants in which the sequences encoding the BD and the C-TAR domains were randomly mutated by error-prone PCR (Swi6*). We screened the resulting library of 10,000 clones against Mbp1 in the death assay and collected 8,000 'positive' clones (non-reconstitution of OyCD activity) (**Fig. 2c**). Second, we screened these Swi6 mutants against Swi4 in the OyCD life assay (reconstitution of OyCD activity). After the two steps of selection, we retested 90 clones carrying potential Swi6

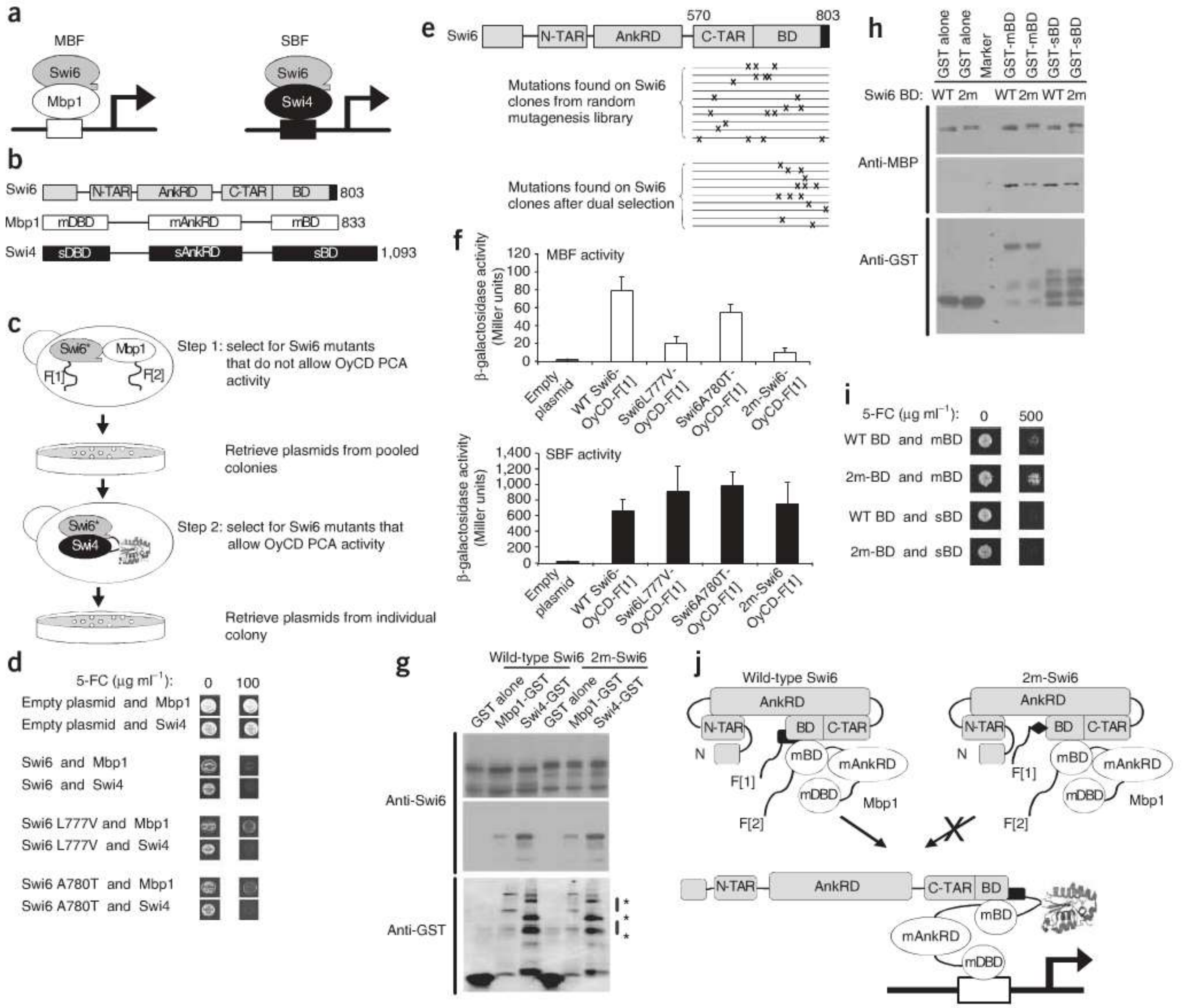


Figure 2 | Dissecting transcriptional activity of Swi6. (**a**) Schematic of MBF and SBF transcription factor complexes. (**b**) Domain structures of Swi6, Mbp1 and Swi4. BD, Swi6 C-terminal domain that binds Mbp1 and Swi4; mBD or sBD, C-terminal Swi6-binding domains of Mbp1 and Swi4, respectively. (**c**) Strategy for engineering a Swi6 mutant. Step 1: death selection screen of a mutant Swi6 OyCD fusion library (Swi6*) expressed with Mbp1 fusion. Selection is for clones lacking OyCD PCA activity (growth on 5-FC). Step 2: life selection of Swi6* clones from step 1, expressed with Swi4 fusion. Selection is for clones with OyCD PCA activity (growth on cytosine). (**d**) Examples of Swi6* clones that grew on 5-FC when expressed with Mbp1 but not with Swi4. (**e**) Distribution of Swi6 mutations in the initial library and after the two-step OyCD PCA screen. (**f**) MBF and SBF transcriptional activities of the indicated Swi6 fusion proteins in *swi6* deletion cells. Data are mean ± s.d. ($n = 4$). (**g**,**h**) GST pulldown assays with full-length proteins expressed in yeast (**g**) or with purified C-terminal binding domains expressed in bacteria (**h**), in the indicated combinations were analyzed by western blot. Wild-type or mutant (2m) Swi6 BD was fused to maltose binding protein (MBP). Unbound (top) and GST-bound (middle) fractions were analyzed with an antibody to Swi6 (anti-Swi6; **g**) or antibody to MBP (anti-MBP; **h**) as well as for expression of GST fusions (bottom) with an antibody to GST (anti-GST). Asterisks indicate degradation products of Mbp1-GST or Swi4-GST. (**i**) OyCD PCA using death selection with Mbp1, Swi4 and Swi6 C-terminal binding domains. (**j**) Model for allosteric regulation of Swi6. Swi6 undergoes a conformational change on binding Mbp1 and activates MBF activity. The 2m-Swi6 does not undergo this change. Black diamond indicates L777V and A780T mutations. N- and C-TAR, N- and C- transcriptional activation domains; AnkRD, ankyrin repeat domain; mDBD and sDBD, Mbp1 and Swi4 DNA binding domains, respectively.

mutants for interactions with Mbp1 or Swi4 using OyCD PCA (**Supplementary Fig. 7**). Nine clones had decreased OyCD PCA activity with Mbp1, with OyCD PCA activity with Swi4 remaining unaffected (**Fig. 2d**). Comparison of a set of sequences from the original Swi6* mutant library to those of the clones found after the life and death selection screen showed that mutants in the initial library were randomly distributed throughout C-TAR and BD, whereas the mutants selected after the second Swi4 screen had mutations located only in the BD, suggesting that mutants were selected for specific binding (**Fig. 2e** and **Supplementary Fig. 8**).

Finally, we screened the nine Swi6 mutants to identify those that disrupt MBF but not SBF activity using MBF and SBF transcription reporter assays[12]. Two single Swi6 mutants (Swi6L777V and Swi6A780T) had decreased MBF activity and unchanged SBF activity, and when we combined the two mutations, the double mutant (2m-Swi6) had an additional reduction in MBF activity and unchanged SBF activity (**Fig. 2f**). The remaining seven clones had no change in MBF or SBF activity (**Supplementary Fig. 7b**).

We next performed glutathione *S*-transferase (GST) pulldown experiments to analyze the interaction between the 2m-Swi6 and Mbp1. Notably, both full-length 2m-Swi6 and C-terminal fragments of 2m-Swi6 retained the ability to bind to both Mbp1 and Swi4 (**Fig. 2g,h**) although OyCD PCA activity was decreased (**Fig. 2d,i**), at similar expression levels of wild-type Swi6 and 2m-Swi6 (**Supplementary Fig. 9**).

A potential explanation for these results is provided by consideration of how binding of Mbp1 and Swi4 to Swi6 activate their transcriptional activity and how the OyCD PCA detects protein-protein interactions (**Fig. 2j**). In its inactive state, the AnkRD of Swi6 antagonizes Swi6 transactivation by direct binding to both N- and C-terminal TARs[13]. Residues 773–784 of Swi6 have been shown to be important in activation of Swi6 (ref. 13). Binding of Swi6 to Mbp1 or Swi4 causes the TARs to dissociate from the AnkRD and Swi6 to open up, allowing the TARs to engage the transcriptional machinery, a transition that requires participation of residues 773–784. As the two mutations in 2m-Swi6 are found in this region, it is possible that the 2m-Swi6 mutations decouple binding of Swi6 to Mbp1 from a change in conformation that is necessary for transactivation. The 2m-Swi6 could be locked in the inactive state, whether or not bound to Mbp1 (**Fig. 2j**).

PCAs are exquisitely sensitive to the topology of protein complexes because the reporter fragments must be free and close enough in space to fold[14,15]. We suggest that the OyCD PCA result for the Mbp1–2m-Swi6 interaction is thus not due to disruption of the interaction, but is caused by sequestering of the PCA fragment that is fused to the C terminus of Swi6 downstream from residues 773–784. In this model, Swi4 must engage the conformation change in Swi6 in a different way, thus allowing for formation of an active SBF complex.

We demonstrated an approach to dissect the functions of a protein by disrupting unique protein-protein interactions or by decoupling binding from conformation changes required for a specific protein function. As many functions of a protein are mediated by multiple protein-protein interactions, the strategy can allow for systematic dissection of protein function and provide mechanistic insights into how binding is coupled to specific functions. The OyCD PCA is general and applicable to study interactions of any full-length protein. Unlike in the yeast two-hybrid or split ubiquitin assays, the proteins (including nuclear chromatin–associated proteins) may be expressed in their appropriate cellular compartments and with posttranslational modifications that reflect their natural state under any specific conditions and in any cell type. Finally, the OyCD PCA should also be invaluable to efforts devoted to creating new chemical or protein probes for manipulation of cellular regulatory networks and to developing therapeutics.

METHODS

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

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

ACKNOWLEDGMENTS

We thank H. Bussey (McGill University) for providing yeast strains and p41XGal1 plasmids, B. Andrews (Donelly Centre, University of Toronto) for plasmids pBA487 and pBA251 and Swi6 antibodies, T. Clackson and V. Rivera (Ariad) for pC$_4$EN-F$_M$3, M.J. Booth for constructing p415Gal-Linker-OyCD-F[2], and M. Malleshaiah and C. Chan for comments. This research was supported by grants from the Canadian Institutes of Health Research (MOP-152556) and the Canada Research Chairs Program to S.W.M. P.H.E. acknowledges the Faculté des etudes supérieurs de l'Université de Montréal for scholarships.

AUTHOR CONTRIBUTIONS

P.H.E. and S.W.M. designed the experiments, analyzed the results and wrote the manuscript. P.H.E. performed the experiments.

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

1. Reguly, T. *et al. J. Biol.* **5**, 11 (2006).
2. Pelletier, J.N., Campbell-Valois, F.X. & Michnick, S.W. *Proc. Natl. Acad. Sci. USA* **95**, 12141–12146 (1998).
3. Michnick, S.W., Ear, P.H., Manderson, E.N., Remy, I. & Stefan, E. *Nat. Rev. Drug Discov.* **6**, 569–582 (2007).
4. Mahan, S.D., Ireton, G.C., Knoeber, C., Stoddard, B.L. & Black, M.E. *Protein Eng. Des. Sel.* **17**, 625–633 (2004).
5. Hartzog, P.E., Nicholson, B.P. & McCusker, J.H. *Yeast* **22**, 789–798 (2005).
6. Wei, K. & Huber, B.E. *J. Biol. Chem.* **271**, 3812–3816 (1996).
7. Fang, F., Hoskins, J. & Butler, J.S. *Mol. Cell. Biol.* **24**, 10766–10776 (2004).
8. Korkegian, A., Black, M.E., Baker, D. & Stoddard, B.L. *Science* **308**, 857–860 (2005).
9. Block, C., Janknecht, R., Herrmann, C., Nassar, N. & Wittinghofer, A. *Nat. Struct. Biol.* **3**, 244–251 (1996).
10. Porter, S.W. & West, A.H. *Biochim. Biophys. Acta* **1748**, 138–145 (2005).
11. Bahler, J. *Annu. Rev. Genet.* **39**, 69–94 (2005).
12. Andrews, B.J. & Moore, L.A. *Proc. Natl. Acad. Sci. USA* **89**, 11852–11856 (1992).
13. Sedgwick, S.G. *et al. J. Mol. Biol.* **281**, 763–775 (1998).
14. Remy, I., Wilson, I.A. & Michnick, S.W. *Science* **283**, 990–993 (1999).
15. Tarassov, K. *et al. Science* **320**, 1465–1470 (2008).

Visual proteomics of the human pathogen *Leptospira interrogans*

Martin Beck[1,5], Johan A Malmström[1,5], Vinzenz Lange[1], Alexander Schmidt[1], Eric W Deutsch[2] & Ruedi Aebersold[1–4]

Systems biology conceptualizes biological systems as dynamic networks of interacting elements, whereby functionally important properties are thought to emerge from the structure of such networks. Owing to the ubiquitous role of complexes of interacting proteins in biological systems, their subunit composition and temporal and spatial arrangement within the cell are of particular interest. 'Visual proteomics' attempts to localize individual macromolecular complexes inside of intact cells by template matching reference structures into cryo-electron tomograms. Here we combined quantitative mass spectrometry and cryo-electron tomography to detect, count and localize specific protein complexes in the cytoplasm of the human pathogen *Leptospira interrogans*. We describe a scoring function for visual proteomics and assess its performance and accuracy under realistic conditions. We discuss current and general limitations of the approach, as well as expected improvements in the future.

The biochemical processes of the living cell are catalyzed in large part by a multitude of functional modules, each of which is characterized by a specific cellular distribution in space and time. Frequently, such modules are complexes of interacting proteins. Quantitative mass spectrometry is commonly used to determine the composition of protein complexes and the proteome in general[1]. However, as such mass spectrometric measurements are carried out on the combined lysates of multiple cells, spatial information is lost, and properties unique to specific cells are averaged over the population of the lysed cells.

Cryo-electron tomography (cryoET) is an imaging technique for the three-dimensional (3D) observation of cells in a close-to-life state[2]. At the currently achievable resolution it should, in principle, be possible to identify and localize large protein complexes in frozen-hydrated specimens. This is accomplished by template matching[3] such that the signals representing a specific protein complex are correlated with the signals acquired on the cell by cryoET. Thereby, an extensive search is performed that scans the entire tomogram for structural templates contained in a database. The combined structural signatures of multiple protein complexes, detected in the cell by cryoET, have the potential to describe the spatial proteome organization of a specific cell, a procedure referred to as 'visual proteomics'[4]. The feasibility of such an approach has been discussed[3] and experimentally demonstrated for cell-free systems[5]; its application to intact cells has so far been limited to the unambiguous detection of ribosomes[6]. The more general application of the technology has been hampered by the challenge of discriminating true from false positive template matches, a task that is complicated by the relatively low signal-to-noise ratio (SNR) of cryo-electron tomograms. Structural diversity and the fact that different protein complexes may exhibit a similar shape and size but differ in abundance by over four orders of magnitude[7] further complicates detection.

To assess and alleviate the current limitations in template matching and to thus turn it into a generic method for suitable classes of templates, we (i) optimized a new scoring function *in silico* under realistic conditions, namely with a large dynamic range of protein concentrations and various noise model scenarios, (ii) extracted a relevant number of cross-correlation features from the tomograms and built reliable statistical models to distinguish true from false positive matches and (iii) applied thorough statistical validation of template matching for different protein complexes localized in numerous tomograms of a large number of individual cells. Collectively, these steps allowed us to confidently detect and localize various complexes in single cells.

The human pathogen *L. interrogans* has a strongly elongated and helically coiled cell shape. The diameter of a cross-section of a typical cell is no more than 100–180 nm and its length is 6–20 μm. These properties make *L. interrogans* an ideal specimen for cryoET, as the cytoplasm of these bacteria can be observed with extraordinarily high contrast without sacrificing resolution. The narrow cross-section allows excellent electron beam penetration and the elongated shape reduces the effects of molecular crowding[2]. We therefore chose *L. interrogans* as a model system to apply the template matching method to detect, count and localize an array of different protein complexes in electron tomograms of frozen-hydrated, individual cells at different states. The robustness and the accuracy of our visual proteomics approach critically depends on prior knowledge of the absolute quantity of the

[1]Institute of Molecular Systems Biology, The Swiss Federal Institute of Technology (ETH Zurich), Zurich, Switzerland. [2]Institute for Systems Biology, Seattle, Washington, USA. [3]Faculty of Science, University of Zurich, Zurich, Switzerland. [4]Center for Systems Physiology and Metabolic Diseases, Zurich, Switzerland. [5]These authors contributed equally to this work. Correspondence should be addressed to R.A. (aebersold@imsb.biol.ethz.ch).

RECEIVED 23 JULY; ACCEPTED 15 SEPTEMBER; PUBLISHED ONLINE 18 OCTOBER 2009; DOI:10.1038/NMETH.1390

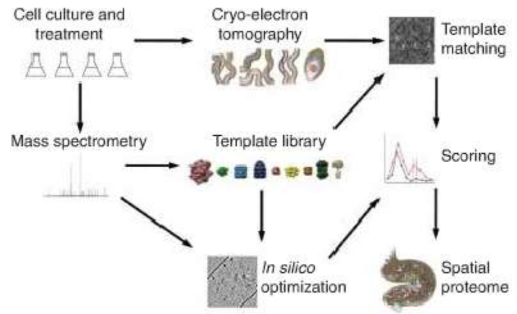


Figure 1 | An integrated workflow for visual proteomics. Differently stimulated cells were subjected to shotgun mass spectrometry and cryoET analyses. A template library was built that included the protein complexes identified in the proteome for which structures of satisfactory homology were available. Targeted, quantitative mass spectrometry was used to determine cellular concentration of the selected targets and to detect inducible changes in their abundance in different cellular states. Test datasets were generated based on the quantitative *Leptospira* proteome to estimate the accuracy of template detection and to train a new scoring function. The templates were cross-correlated with the electron optical density in the tomograms by template matching as described earlier[4] and assigned into the spatial context of the cell using the statistically evaluated, optimized scoring function.

targeted complexes in the cell, thus requiring the convergence of quantitative mass spectrometry and cryoET.

RESULTS

Workflow and selection of target protein complexes

The general experimental workflow of this study consists of the synergistic use of quantitative mass spectrometry to select and quantify protein complexes suitable for visual proteomics and cryoET to detect and localize them in close-to-life, frozen-hydrated cells (**Fig. 1**).

We used liquid chromatography–tandem mass spectrometry (LC-MS/MS) to generate an extensive proteome list for *L. interrogans* containing 2,221 proteins, representing 61% of the proteome predicted from the 3,658 open reading frames annotated in the *L. interrogans* genome (**Supplementary Fig. 1**). The data are available in PeptideAtlas. We performed a Psi-Blast analysis against protein sequences from all species, and identified 26 *L. interrogans* protein complexes that we initially considered suitable for template matching (**Supplementary Table 1**). The complexes in the set fulfilled the following criteria: (i) primary structures of the complex subunits are well conserved in bacterial species, (ii) 3D structure of bacterial homologs have been solved, and (iii) oligomeric assembly has the minimal mass and/or spatial elongation to make it detectable by cryoET.

We then used label-free quantitative proteomics based on inclusion list–guided LC-MS/MS[8] to identify components of the protein complexes on the target list. We analyzed extracts from *L. interrogans* cells in four states, (i) exponentially growing, unperturbed cells, (ii) heat-shocked cells, simulating fever, (iii) antibiotic (ciprofloxacin)-treated cells and (iv) starved cells. The data indicated the quantitative behavior of the *L. interrogans* proteome in general and specifically of the proteins associated with the pathways relevant for the selected protein complexes (**Fig. 2** and **Supplementary Results**). Apart from the starvation state, which was associated with obvious morphological changes of *L. interrogans* cells (**Fig. 2a**), the antibiotic treatment caused the most pronounced change in abundance of the protein complexes targeted for template matching. After 24 h of treatment with 5 μg ml^{-1} ciprofloxacin, we observed an increase in the amounts of ATP synthase by 35%, of ClpB by 50% and of Hsp15 proteins by 30- to 40-fold and a decrease in the amounts of ribosomes by 25% (**Fig. 2b**). Therefore we chose to investigate cells primarily in the unperturbed and antibiotic-treated condition.

Out of the initial 26 protein complexes, five were abundant in the cytoplasm and occurred at more than 1,000 copies per cell, another seven were of medium abundance and occurred at least at 100 copies per cell, and the remaining complexes were present at less than 100 copies per cell (**Supplementary Table 1**). We used the quantitative proteomic data to select nine protein complexes from the initial set that cover a matrix of different molecular weights and cellular abundances (**Table 1** and **Supplementary Figs. 1,2**). Three complexes, ATP synthase, RNA polymerase II and the ribosome are part of central cellular pathways and their abundance was modulated under different conditions (**Supplementary Results**). The remaining six selected complexes are involved in protein folding and degradation. Specifically, the GroEL and GroEL-GroES complex are chaperones that assist protein folding; ClpB, ClpP and HslU-HslV are unfoldases and proteases; and Hsp15 is a heat shock protein. In contrast to other organisms in which Hsp15 is involved in the recycling of 50S ribosomal units, Hsp15 and Hsp15-like in *L. interrogans* are homologous to Hsp20 (ref. 9), a family of small stress-induced proteins that form large, oligomeric assemblies[10,11].

Generation of a scoring function and *in silico* test data

The accurate detection, quantification and localization of the selected templates in the *L. interrogans* cells by cryoET depends on a statistically reliable determination of true positive discovery rates. The ensuing problem of correctly discriminating true from false cross-correlations is similar to the problem of discriminating true from false assignments of fragment ion spectra to peptide sequences in mass spectrometry–based proteomics. We therefore adapted the statistical concepts established in proteomics and that are implemented in the PeptideProphet[12] software tool to generate a discriminate scoring function for visual proteomics. It consists of a linear combination of the following three knowledge-based, empirical subscores: (i) distribution of cross-correlation scores from matching templates to background; (ii) distribution of cross-correlation scores from matching templates to competing templates (local score compared to the score of all other templates in the database); and (iii) distribution of cross-correlation scores from matching templates to three different decoy templates (arbitrarily chosen geometrical shapes: cube, ellipsoid and cylinder). Reliable detection by cryoET is also complicated by the fact that the features to which the templates are assigned are of vastly different abundance. Consequently, false matches of a low-abundance complex to a highly abundant competing structure might dramatically inflate the false positive error rate for the low-abundance complex. We therefore measured the absolute average cellular concentration of the targeted protein complexes under all the investigated conditions by selected reaction monitoring (SRM)–mass spectrometry[13] to generate test datasets containing the target protein complexes at their actual cytoplasmic concentration and to validate

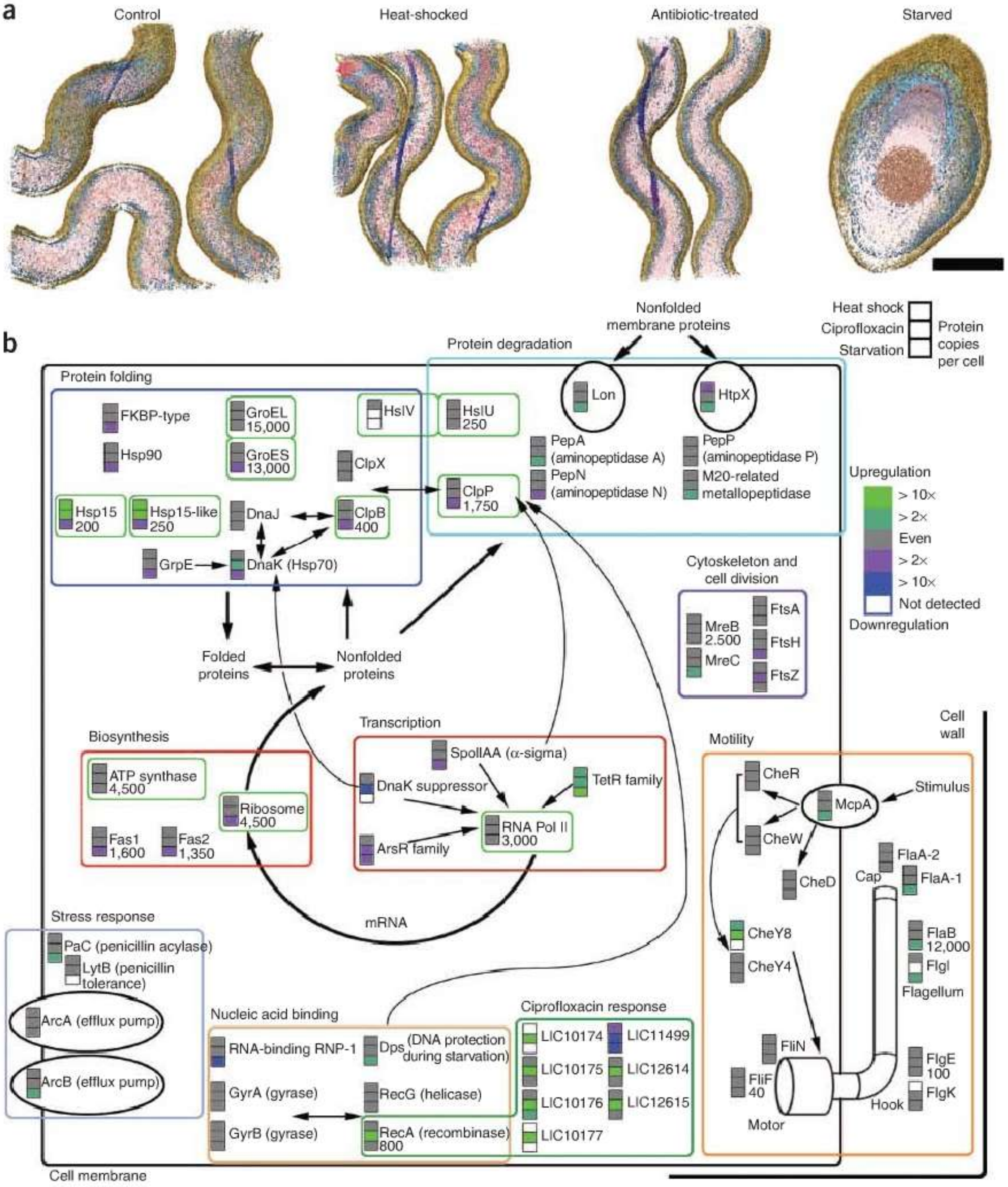


Figure 2 | Stress response of *L. interrogans* cells in the context of the protein complexes selected as templates for template matching. (**a**) Three-dimensional surface-rendered volumes of exponentially growing cells incubated for 1 h at 42 °C (heat shock), for 24 h in the presence of 5 μg μl^{-1} ciprofloxacin (antibiotic-treated) or for 7 d in the absence of nutrients (starved) and of nontreated cultures (control). The cytoplasm is colored in red, membranes in bright blue, periplasmic flagella in dark blue and the cell wall in brown. Scale bar, 200 nm. (**b**) Up- and downregulation for each protein is shown for the heat-shock, antibiotic (ciprofloxacin) treatment and starvation condition versus the control. Proteins are grouped based on their function. The relative abundance of many proteins was reduced under starvation, and heat-shock proteins were strongly upregulated under both stress conditions (heat shock and antibiotic treatment). Recombinase A and a cluster of (so far) hypothetical proteins had a very strong response upon treatment with ciprofloxacin (dark green). In contrast, the abundance of the ciprofloxacin target DNA gyrase was unchanged. Copy numbers per cell, as determined by SRM, are given for the control condition (if applicable).

the signals derived from real data with an independent method (**Supplementary Table 2**). The number of copies per cell of the targeted complexes ranged from 4,500 copies per cell for the ribosome in nonstimulated cells to 15 copies per cell HslU-HslV complex in antibiotic-treated cells (**Table 1**).

To optimize the relative weight of each subscore in the scoring function and to test whether the selected templates can, in principle, be detected at the electron optical properties of the acquired tomograms, we generated test datasets (phantom cells) *in silico* that contained two independent noise components[14] at various levels: contrast transfer function–convoluted (quantum) noise to simulate the interaction of the electron beam with the specimen and modulation transfer function–convoluted (detector) noise to simulate the

Table 1 | Cellular abundance of template-protein complexes

	Copy number per cell (s.d.; n = 4)		
Template	**Nonstimulated**	**Heat-shocked**	**Antibiotic-stimulated**
Ribosome	4,500 (500)	3,500 (700)	3,400 (300)
RNA polymerase II	3,000 (200)	3,000 (200)	3,000 (400)
ATP synthase	1,500 (500)	1,500 (400)	2,300 (600)
GroEL	1,100 (100)	1,300 (150)	1,300 (150)
GroES	1,900 (200)	1,900 (300)	1,700 (350)
ClpB	70 (10)	70 (10)	100 (40)
ClpP	140 (30)	110 (60)	140 (70)
Hsp15	40 (5)	310 (60)	1,300 (150)
HslU–HslV	20 (5)	15 (5)	15 (10)

In contrast to values given in **Figure 2**, stochiometric relations have been taken into account. For Hsp15, the sum of both closely related gene products is given.

imperfection of the charge-coupled device (CCD) camera (**Fig. 3** and **Supplementary Results**).

We then adjusted the total SNR by comparing the cross-correlation coefficients obtained by template matching ribosomes in real tomograms to test datasets. To account for the structural noise component during this process[15], we used two different conformations of the ribosome[16] (Protein Data Bank (PDB) identifiers: 1P85, 1P86, 1P87 and 1P6G) to generate test datasets and used a third structure for template matching (PDB identifiers: 2AW7 and 2AWB). The two scenarios developed above showed cross-correlation values (0.3–0.45 for top cross-correction values of the ribosomal template) that roughly matched the real data (0.30–0.35): (i) a total SNR of 0.5 with even quantum and detector noise contribution and (ii) a total SNR of 0.05 with a predominant quantum noise contribution. In both cases the test datasets looked similar to real tomograms (**Fig. 3**). We therefore concluded that these two noise scenarios matched real data the best and investigated the performance of template matching further for both cases.

Assessment of the visual proteomics performance *in silico*

We used the *in silico* test datasets to optimize the subscores by a linear discriminate analysis and to validate the performance of the true positive from false positive discrimination (**Fig. 3c–e**). The two simulation conditions that best matched the real data comprised a conservative and a more optimistic scenario, as the detector noise component substantially limited the attainable resolution by damping higher frequencies. In case of the conservative scenario (even contribution of quantum and detector noise at a total SNR of 0.5), at a sensitivity of 75%, we obtained the following specificities for the targeted complexes: ribosome, >90%; RNA polymerase II, ~50%; GroEL, ~60%; GroEL-GroES, ~70%; undistinguished GroEL or GroEL-GroES, >90%; HslU-HslV, >90%; and ATP synthase, ~50% (**Supplementary Fig. 3**). The smaller templates ClpB, ClpP and Hsp barely or never reached this sensitivity level, but when we set the sensitivity to 50%, we discovered these smaller templates with the following specificities: ClpB, ~40%; ClpP, ~45%; and Hsp, ~45%. In the case of the optimistic scenario with predominant quantum noise, the specificity for most of the templates was at least equal to the more conservative scenario at the 75% sensitivity but at a tenfold lower total SNR. Exceptions were RNA polymerase II and Hsp, which reached ~40% and ~25% specificity at 50% sensitivity, respectively. This is quite remarkable given the lower SNR of 0.05 and is partially caused by the low-pass filter applied during reconstruction that reduces quantum noise more effectively than detector noise. We also found that the performance of visual proteomics, particularly in case of the smaller templates is strongly dependent on template abundance. For example, when we investigated phantom cells that mimic the untreated cell state in which Hsp is present at very low abundance, the specificity dropped to less than 10%. Generally, we can conclude from the analysis of the *in silico* test data that template matching can theoretically detect protein complexes of distinct shape and sufficient size (**Fig. 3f**) over a cellular abundance dynamic range of a maximum of two orders of magnitude.

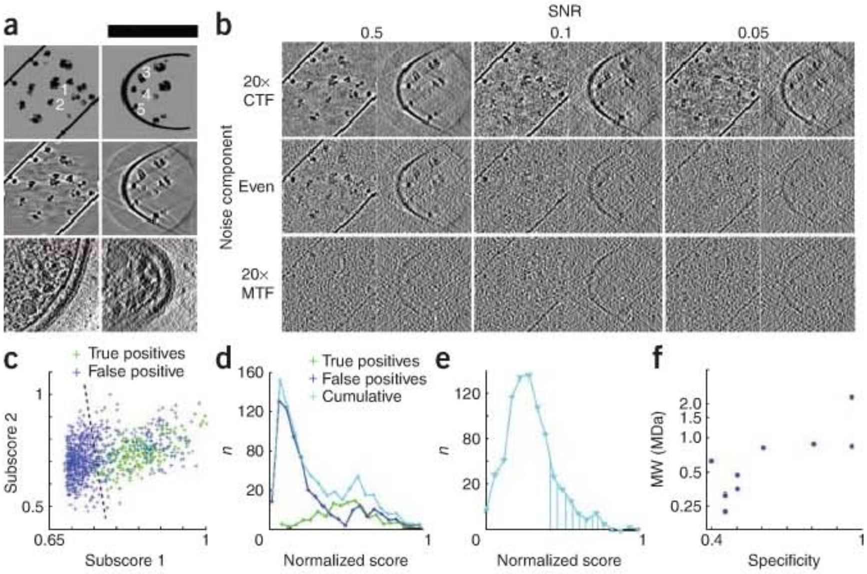


Figure 3 | Generation of *in silico* test data and development of a scoring function for template matching in cryo-electron tomograms. (**a**) Slices through reconstructions of 5 nm in thickness displayed along the *z* axis (left) and *x* axis (right) showing unprocessed test dataset (top), tomographic reconstruction without noise (middle) and the collected dataset (bottom). The positions of a ribosome (1), RNA polymerase II (2), GroEL (3), Hsp (4) and ATP synthase (5) are indicated. Scale bar, 150 nm. (**b**) Reconstructions with an SNR of 0.5, 0.1 and 0.05 with predominant quantum (top), even (middle) and predominant detector noise component (bottom). CTF, contrast transfer function; MTF, modulation transfer function. The noise models in the middle left (conservative) and top right (optimistic) are discussed in text. For each SNR value, slices through reconstructions of 5 nm in thickness displayed along the *z* axis (left) and *x* axis (right) are shown. (**c**) Linear discrimination of the subscores 1 and 2 in the *in silico* RNA polymerase II test datasets. Dashed line, optimal discrimination threshold. (**d**) Score distribution in the *in silico* RNA polymerase II test datasets. (**e**) Score distribution from real data. The marked area under the curve corresponds to the absolute abundance expectation value for the given cellular volume as determined by SRM. The curve shape is very similar to the theoretical distribution shown in **d**. (**f**) Molecular weight (MW) plotted on a logarithmic scale against the specificity achieved *in silico* at 50% sensitivity (conservative noise model).

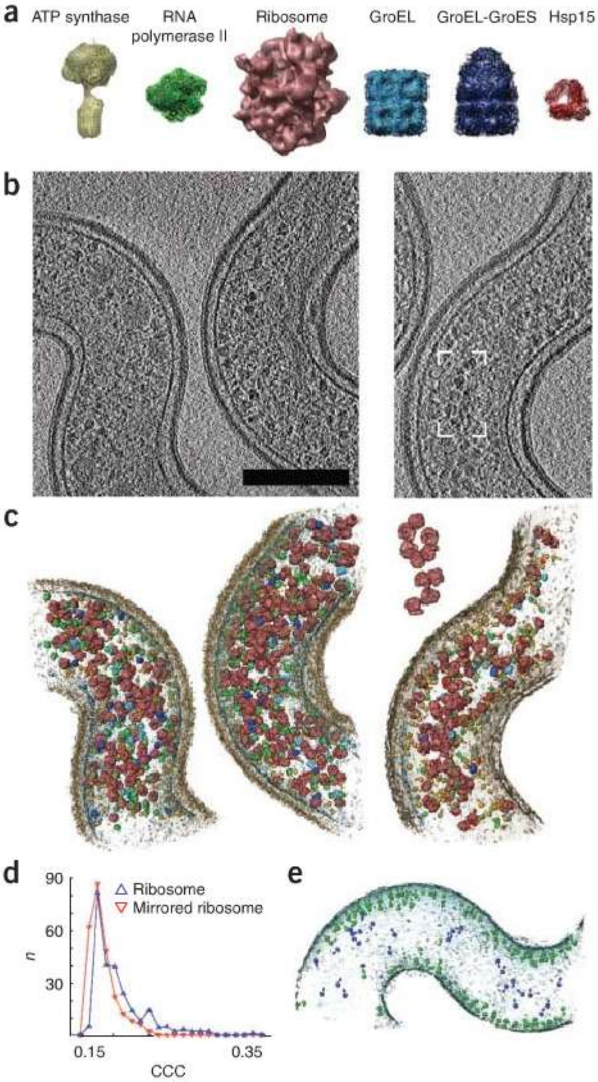


Figure 4 | Template matching in subvolumes of *L. interrogans* cells. (**a**) Template library of protein complexes superimposed with amino acid chain traces (in black, if applicable). The surface rendering has been done at the relevant resolution of the references applied to template matching. (**b**–**c**) Localization of the targeted protein complexes by template matching in a representative subvolume of nonstimulated (left) and antibiotic-treated (right) cells. Scale bar, 200 nm. Slices through the reconstructed volumes of 7 nm in thickness without post-processing (**b**) and surface-rendered models of the assigned templates (colored as in **a**) in context with the cell wall (transparent brown) and membrane (transparent blue) (**c**). The bracketed region in **b** and inset in **c** show a group of ribosomes resembling the pseudo-planar relative orientation of polyribosomes reported recently for bacterial lysates[17]. (**d**) Distribution of the top 250 cross-correlation coefficients (CCCs) extracted from a tomogram with the ribosome as template and mirrored ribosome as decoy template. The cross-correlation intensity is lower in case of the decoy template and the curve shape changes. (**e**) Surface-rendered view of a tomographic reconstruction of an *L. interrogans* cell showing ATP synthase localization in context with the cell membrane. Single particles with a plausible positioning and orientation (membrane-embedded and pointing into the cytoplasm) are colored in green, non-plausible false positives are in blue.

Assessment of the performance of visual proteomics in cells

The quantitative mass spectrometry data reflect the protein abundance as an average over a large number of cells. To detect cell to cell variations and variations in the local concentration of the targeted complexes, we acquired six tomograms using an identical experimental setup for each, the non-stimulated, heat-shocked and antibiotics-treated condition that collectively contained subvolumes of 37 individual *L. interrogans* cells, each covering ~10% of the average cell volume. To detect, localize and quantify the targeted protein complexes (**Table 1**) we applied the optimized template matching method. To compensate for potential variations in the local protein concentration, we selected tomograms with similar SNR from the larger data pool and applied the optimized scoring function to the selected volumes of each cellular state as a whole. We set the anticipated discovery rate to 80% of the cytoplasmic concentration of the template protein complexes. When the conservative noise model was used as a base for estimating the confidence for visual proteomics in real datasets, the following specificities are achieved in average: ribosome >90%, RNA polymerase II ~40%, GroEL ~80%, GroELS ~70%, undistinguished GroEL or GroELS >90%, ATP synthase ~50% (including angular bias correction; **Supplementary Results**) and Hsp 45% (in the ciprofloxacin-treated state). The detection of the low abundant target protein complexes turned out to be very challenging in real datasets: When the number of single particles that can be expected per tomogram ranged from 0 to 5, the resulting statistical models were noisy and not straight-forward to interpret. We therefore omitted HslU-HslV, ClpB and ClpP from our analysis.

The resulting score distributions of several real datasets were in good agreement with the abundance expectation value for the given template in the particular fraction of the cytoplasmic volume (as determined by SRM), demonstrating the power of SRM to function as an independent method for validating template matching (**Fig. 3d**,**e** and **Supplementary Fig. 4**). In addition, some templates provided auxiliary information for the orthogonal validation of the data as follows. (i) Ribosome spatial arrangement: a group of ribosomes that resembles the pseudo-planar relative orientation of polyribosomes reported recently for bacterial lysates[17] could be found (**Fig. 4a–c**). (ii) Ribosome handedness: when ribosomes with inverted handedness were used as decoy templates, they could be clearly discriminated from their native counterparts (**Fig. 4d**). (iii) ATP synthase cellular localization: an ATP synthase is membrane-embedded and points inwards from the cytoplasmic membrane. We therefore asked what fraction of the high-quality template matches conformed to these highly restrictive features. The ATP-synthase complexes matched in a tomogram of an *L. interrogans* cell using the optimized scoring function are shown in **Figure 4e**. We manually discriminated plausible positioning and orientation from nonplausible matches. These data indicate a false positive discovery rate in real tomograms (~10%) that is more optimistic than in test datasets (50%), demonstrating that topological accuracy is not necessarily in agreement with positional correctness (**Supplementary Results**).

The combined quantitative proteomics-cryoET method described here allowed us to link abundance measurements obtained from the combined cell lysate of many cells with the subcellular distribution of selected protein complexes in single cells and to detect variations from cell to cell and even in subcellular volumes. Cells in the nonstimulated state had an average ribosome concentration of ~20 μM (~40 mg ml^{-1}) in the cytoplasm, but the local concentration ranged from 5 to 30 μM (~10–65 mg ml^{-1}) independent of the different conditions investigated. The local fluctuations in case of total GroEL together with GroEL-GroES were larger and ranged from ~8 to 100 μM (~0.5–6.5 mg ml^{-1}). In some cases these local fluctuations can be explained by the following phenomena: some regions of *L. interrogans* cells are occupied by large spherical structures (~50–100 nm) of relatively homogenous electron optical density (**Supplementary Fig. 5**). The nature of these structures is unknown, but if present, they cause a local decrease of protein complexes by displacement. The most obvious response to stress in the proteome was the upregulation of Hsp: the cytoplasmic concentration of Hsp increased from 0.06 μM in nonstimulated cells to 30 μM (~45 μg ml^{-1}) during stress. However, even slight variations in the total SNR between different tomograms as well as in the local SNR can severely hamper the template detection, particularly in case of the 'small' protein complexes.

DISCUSSION

We found that the noise generated by the detectors widely used for cryoET (CCD cameras), but not by the interaction of the electron beam with the specimen itself, is currently a critical technical limitation of implementing the visual proteomics concept. This noise reduces the attainable resolution so that only 'very large' protein complexes such as the ribosome or GroEL can be discovered with high confidence. The development of alternative detection concepts is an active field of research and might enable a superior image digitization for cryoET in the near future, thereby pushing the size threshold of detectable complexes toward smaller molecular weights.

Our study showed that the very large molecular machines are, in most cases, of low abundance in the cytoplasm and therefore will be difficult to detect in general. Notable exceptions are protein complexes involved in transcription, translation and protein folding, which are the most abundant cytoplasmic proteins[7]. This problem will be even more severe for higher organisms, which have a larger dynamic range of protein concentrations than bacteria[7,18]. Much higher throughput in cryoET data acquisition, but also in the computational post-processing, will be essential to detect low-abundance protein complexes with higher confidence.

We demonstrate that molecular crowding can hamper the detection of protein complexes by template matching (**Supplementary Fig. 2**). This problem might be addressed by more sophisticated image classification techniques in the future. Also, development of preparation techniques used for cryoET, particularly specimen thinning, might enable the investigation of less crowded biological systems.

Another important limitation is structural diversity and protein complex oligomeric states: most templates used in this study form very stable assemblies. For example, in the case of the ribosome it is reasonable to assume that the predominant species in exponentially growing *L. interrogans* cells is engaged in translation and is therefore fully assembled, as has been shown *in vivo* for yeast[19]. The expansion of template libraries to cover as many structural species as possible is, however, an important focus for future development.

The biological system investigated here is exceptional in terms of specimen thickness as well as the targeted protein complexes, and therefore represents the currently feasible state of our visual proteomics approach that combines quantitative proteomics and cryo-ET. Nevertheless, a quantitative assessment of different structural species of short-lived protein complexes could enable the integration of other targets undergoing dynamic structural changes, and technical improvements increasing the resolution and signal to noise of cryoET might enable the application of the technology to more biological systems and might facilitate structure-based modeling in systems biology. The method described here can be applied generally to estimate confidence values of cross-correlation–based feature extraction in cryoET and therefore will also be useful for structure determination by subtomogram averaging[20]. *In silico* optimization of the data acquisition parameters upfront can be envisioned as well as *in silico* testing of new classifiers.

METHODS

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

Accession codes. PeptideAtlas: PAe000383, PAe000349–PAe000352 and PAe000401.

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

ACKNOWLEDGMENTS

This project has been funded in part by ETH Zurich, the Swiss National Science Foundation (grant 31000-10767), federal funds from the US National Heart, Lung, and Blood Institute, the US National Institutes of Health (contract N01-HV-28179), by SystemsX.ch the Swiss initiative for systems biology, in part by the proteomics in time and space (PROSPECTS) European network of excellence, and with funds from the European Research Council project 'Proteomics V3.0'. M.B. was supported by a long-term fellowship of the European Molecular Biology Organization and a Marie Curie fellowship of the European Commission; J.A.M. was supported by a fellowship from the Swedish society for medical research (SSMF); and A.S. and V.L. were supported by the Competence Center for Systems Physiology and Metabolic Diseases. We thank O. Medalia, W. Baumeister, members of the electron microscopy facility of ETH Zurich (EMEZ) for continued support and F. Förster for critical reading of the manuscript.

AUTHOR CONTRIBUTIONS

J.A.M. and M.B. planned the experiments, performed the experimental work and data analysis and wrote the manuscript. A.S. and V.L. participated in the experimental work and the data analysis. E.W.D. assembled the PeptideAtlas build. R.A. was the project leader and wrote the manuscript.

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

1. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
2. Lucic, V., Forster, F. & Baumeister, W. Structural studies by electron tomography: from cells to molecules. *Annu. Rev. Biochem.* **74**, 833–865 (2005).
3. Best, C., Nickell, S. & Baumeister, W. Localization of protein complexes by pattern recognition. *Methods Cell Biol.* **79**, 615–638 (2007).
4. Nickell, S., Kofler, C., Leis, A.P. & Baumeister, W. A visual approach to proteomics. *Nat. Rev. Mol. Cell Biol.* **7**, 225–230 (2006).
5. Frangakis, A.S. *et al.* Identification of macromolecular complexes in cryoelectron tomograms of phantom cells. *Proc. Natl. Acad. Sci. USA* **99**, 14153–14158 (2002).
6. Ortiz, J.O., Forster, F., Kurner, J., Linaroudis, A.A. & Baumeister, W. Mapping 70S ribosomes in intact cells by cryoelectron tomography and pattern recognition. *J. Struct. Biol.* **156**, 334–341 (2006).
7. Malmström, J. *et al.* Proteome-wide cellular protein concentrations of the human pathogen *Leptospira interrogans*. *Nature* **460**, 762–765 (2009).

8. Schmidt, A. *et al.* An integrated, directed mass spectrometric approach for in-depth characterization of complex peptide mixtures. *Mol. Cell. Proteomics* **7**, 2138–2150 (2008).

9. Nally, J.E., Artiushin, S. & Timoney, J.F. Molecular characterization of thermoinduced immunogenic proteins Q1p42 and Hsp15 of *Leptospira interrogans*. *Infect. Immun.* **69**, 7616–7624 (2001).

10. Kennaway, C.K. *et al.* Dodecameric structure of the small heat shock protein Acr1 from *Mycobacterium tuberculosis*. *J. Biol. Chem.* **280**, 33419–33425 (2005).

11. Kim, R., Kim, K.K., Yokota, H. & Kim, S.H. Small heat shock protein of *Methanococcus jannaschii*, a hyperthermophile. *Proc. Natl. Acad. Sci. USA* **95**, 9129–9133 (1998).

12. Keller, A., Nesvizhskii, A.I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–5392 (2002).

13. 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).

14. Forster, F., Pruggnaller, S., Seybert, A. & Frangakis, A.S. Classification of cryo-electron sub-tomograms using constrained correlation. *J. Struct. Biol.* **161**, 276–286 (2008).

15. Baxter, W.T., Grassucci, R.A., Gao, H. & Frank, J. Determination of signal-to-noise ratios and spectral SNRs in cryo-EM low-dose imaging of molecules. *J. Struct. Biol.* **166**, 126–132 (2009).

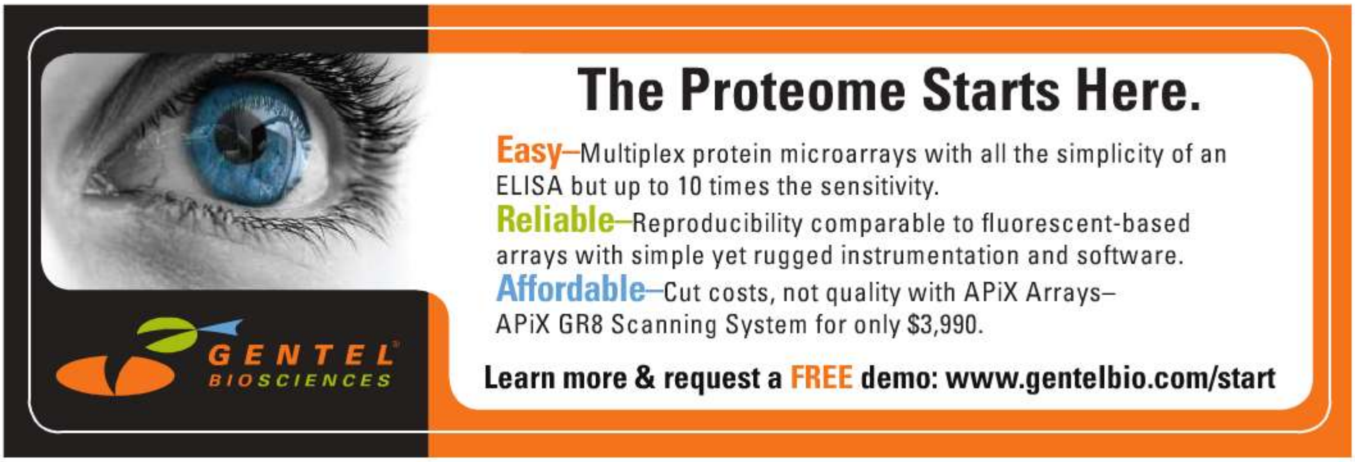
16. Gao, H. *et al.* Study of the structural dynamics of the *E. coli* 70S ribosome using real-space refinement. *Cell* **113**, 789–801 (2003).

17. Brandt, F. *et al.* The native 3D organization of bacterial polysomes. *Cell* **136**, 261–271 (2009).

18. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).

19. Dresios, J., Derkatch, I.L., Liebman, S.W. & Synetos, D. Yeast ribosomal protein L24 affects the kinetics of protein synthesis and ribosomal protein L39 improves translational accuracy, while mutants lacking both remain viable. *Biochemistry* **39**, 7236–7244 (2000).

20. Förster, F., Medalia, O., Zauberman, N., Baumeister, W. & Fass, D. Retrovirus envelope protein complex structure in situ studied by cryo-electron tomography. *Proc. Natl. Acad. Sci. USA* **102**, 4729–4734 (2005).



Engineering splicing factors with designed specificities

Yang Wang[1], Cheom-Gil Cheong[2], Traci M Tanaka Hall[2] & Zefeng Wang[1]

Alternative splicing is generally regulated by *trans*-acting factors that specifically bind pre-mRNA to activate or inhibit the splicing reaction. This regulation is critical for normal gene expression, and dysregulation of splicing is closely associated with human diseases. Here we engineered artificial splicing factors by combining sequence-specific RNA-binding domains of human Pumilio1 with functional domains that regulate splicing. We applied these factors to modulate different types of alternative splicing in selected targets, to examine the activity of effector domains from natural splicing factors and to modulate splicing of an endogenous human gene, *Bcl-X*, an anticancer target. The designer factor targeted to *Bcl-X* increased the amount of pro-apoptotic Bcl-xS splice isoform, thus promoting apoptosis and increasing chemosensitivity of cancer cells to common antitumor drugs. Our approach permitted the creation of artificial factors to target virtually any pre-mRNA, providing a strategy to study splicing regulation and to manipulate disease-associated splicing events.

As a key regulatory step in gene expression, alternative splicing is widespread in humans with most genes producing multiple splicing isoforms with distinct and sometimes opposing functions[1]. The choice of splicing isoforms is tightly regulated in different tissues and developmental processes, and disruption of such regulation is a common cause of human disease[2]. Therefore new approaches to specifically modulate alternative splicing will improve our understanding of splicing regulation and will have therapeutic potential.

Generally, the splicing process is regulated through *cis*-acting elements that recruit *trans*-acting splicing factors to affect use of nearby splice sites. Many splicing factors have modular organization, with separate sequence-specific RNA binding modules and splicing effector domains. For example, serine/arginine-rich proteins contain N-terminal RNA recognition motifs (RRMs) that bind to exonic splicing enhancers in pre-mRNAs and C-terminal arginine/serine–rich domains that promote exon inclusion[3]. Analogously, the heterogeneous nuclear ribonucleoprotein (hnRNP) A1 binds to exonic splicing silencers through its RRM and inhibits exon inclusion through a C-terminal glycine-rich domain[4]. Following this configuration, we envisioned engineering unique factors by combining an RNA recognition module to recognize targets with a functional module to affect splicing. Ideally, such an engineered splicing factor (ESF) should recognize any target and modulate splicing in desired ways.

Here we report what, to our knowledge, is the first attempt to engineer splicing factors with designed sequence specificity and activities. These ESFs can promote or suppress splicing and can specifically recognize target genes undergoing different types of alternative splicing. We used such ESFs to study the functional domains of natural splicing factors and designed a new ESF to specifically modulate splicing of an endogenous gene. The 'designer' ESF can shift splicing of *Bcl-X* (*BCL2L1*) to increase the pro-apoptotic isoform in multiple cultured cancer cells, therefore sensitizing these cells to common antitumor drugs.

RESULTS

Design principles of ESFs

To generate an RNA-binding module with predictable specificity, we used the unique RNA recognition mode of PUF proteins (named for *Drosophila melanogaster* Pumilio and *Caenorhabditis elegans fem-3* binding factor) that are involved in mediating mRNA stability and translation[5]. Most splicing factors recognize their targets through RRMs or K homology domains that bind to short RNA elements with moderate affinities. However, it is impractical to engineer an RNA recognition module using these domains because of their weak binding affinity and the absence of a predictive RNA recognition code. The PUF domain of human Pumilio1 binds tightly to cognate RNA sequences, and its specificity can be modified. It contains eight PUF repeats that recognize eight consecutive RNA bases with each repeat recognizing a single base[6] (**Fig. 1a**). Two amino acid side chains in each repeat recognize the Watson-Crick edge of the corresponding base and determine the specificity of that repeat, thus a PUF domain can be designed to specifically bind most 8-nucleotide RNAs[6,7]. To generate a sequence-specific ESF, we fused a modified PUF domain with a splicing regulatory domain: an arginine/serine–rich domain for a splicing activator or a glycine-rich domain for a splicing repressor. We also included a nuclear localization sequence to direct the ESF to the nucleus where splicing occurs and a Flag tag to facilitate detection (**Fig. 1b**).

[1]Department of Pharmacology, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [2]Laboratory of Structural Biology, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, North Carolina, USA. Correspondence should be addressed to Z.W. (zefeng@med.unc.edu).

RECEIVED 2 JULY; ACCEPTED 4 SEPTEMBER; PUBLISHED ONLINE 4 OCTOBER 2009; DOI:10.1038/NMETH.1379

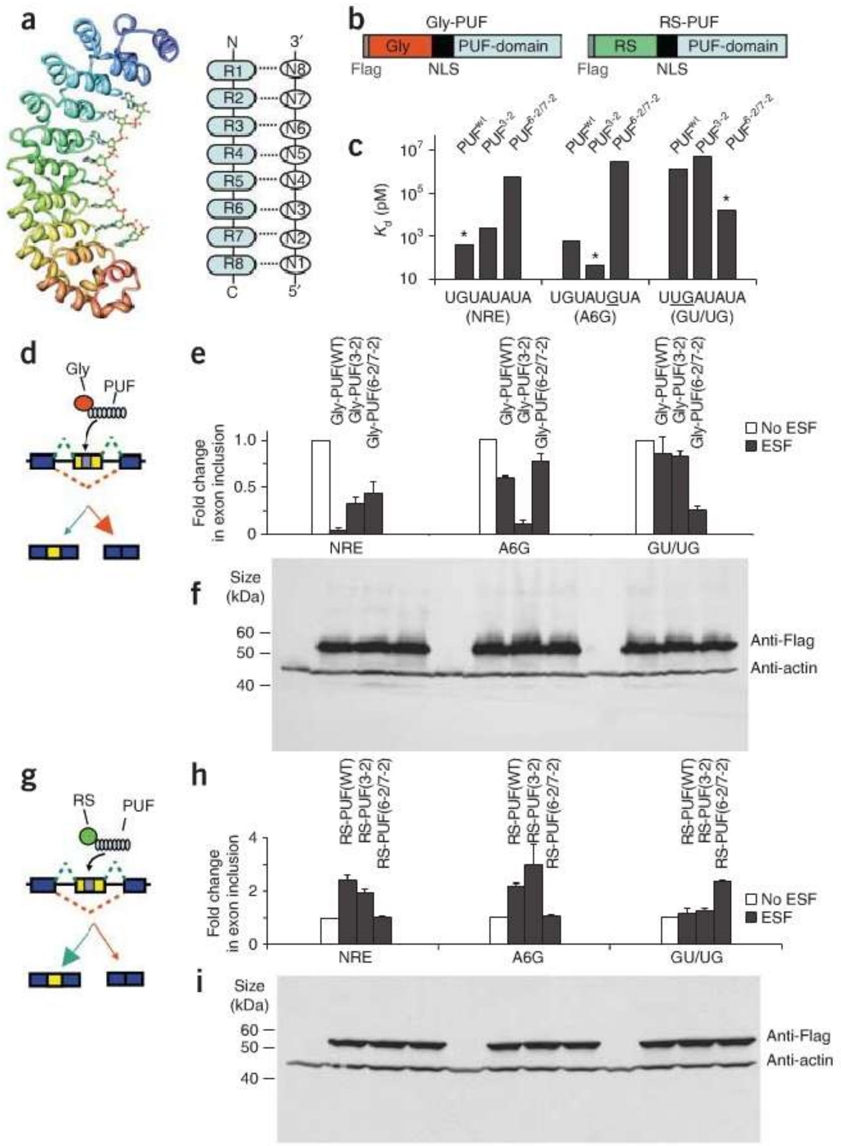


Figure 1 | Design of ESFs to modulate exon skipping. (**a**) The interaction between PUF domain and RNA is illustrated by crystal structure[6] (left) and diagram (right). The PUF domain repeats R1–R8 recognize nucleotides N8–N1, respectively. (**b**) Modular domain organization of ESFs. NLS, nuclear localization signal. (**c**) The dissociation constants (K_d) of PUF domain–8-mer target binding. Recognition sequences for PUF(WT) (Nanos response element (NRE)), for PUF(3-2) (A6G) and for PUF(6-2/7-2) (GU/UG) are indicated. Asterisks mark RNA–PUF domain cognate pairs. (**d**–**i**) Schematics of inhibition of cassette exon (yellow) inclusion by Gly-PUF–type ESFs (**d**) and promotion of exon inclusion by RS-PUF–type ESFs (**g**) in exon-skipping reporters containing different 8-nucleotide target sequences (light blue). The combinations of vectors encoding splicing reporters and ESF expression vectors listed in **c** were transfected into 293T cells, and inclusion of cassette exon was analyzed by RT-PCR 24 h after transfection. Fold change in cassette exon inclusion was plotted relative to that for the reporters alone without ESF (**e**,**h**). Error bars, s.d. (n = 3). Western blot analyses of ESF expression using antibodies to the Flag tag (anti-Flag), with detection of actin with actin antibody as loading controls (**f**,**i**); samples are in the same order as in **e**,**h**, respectively.

Specifically modulating exon inclusion with ESFs

To test the design concept, we created six ESFs by fusing either the glycine-rich (Gly) domain of hnRNP A1 or the arginine/serine–rich (RS) domain of ASF, also known as SF2 (ASF/SF2) with three different PUF domains (wild-type human Pumilio1 and two modified PUF domains) to obtain Gly-PUF–type or RS-PUF–type ESFs. Wild-type PUF specifically binds to Nanos response element RNA sequence, whereas the PUF(3-2) mutant (C935S,Q939E; PUF(3-2) indicates two mutated amino acid in PUF repeat 3) recognizes A6G sequence and the PUF(6-2/7-2) mutant (N1043S,Q1047E mutations in PUF repeat 6 and S1079N,E1083Q mutations in repeat 7) recognizes GU/UG sequence[6,7] (**Fig. 1c**). We generated exon-skipping reporters containing the corresponding 8-nucleotide target sequences of the ESFs in the alternatively spliced cassette exon. We transfected these reporters into 293T cells with the ESF expression constructs and analyzed changes in splicing using reverse transcriptase (RT)-PCR.

Consistent with our prediction, the Gly-PUF–type ESFs repressed inclusion of the cassette exon containing a cognate target sequence (**Fig. 1d–f**). Such inhibition was sequence-specific, with the maximal inhibition of exon inclusion occurring between cognate ESFs and reporters (**Fig. 1e**). The splicing repressor activities of Gly-PUF–type ESFs correlated roughly with the binding affinities of PUF motifs to their targets (**Supplementary Fig. 1**), suggesting that the binding affinity to its target contributes to splicing factor strength.

In line with the above observations, RS-PUF–type ESFs had opposite activity on splicing and promoted inclusion of cassette exons containing cognate targets (**Fig. 1g–i**). The activities of these ESFs correlated roughly with the binding affinities of PUF motifs to their targets (**Supplementary Fig. 1**), supporting the modular configuration of splicing factors[8,9]. In addition, both types of ESFs were effective in the different cell types tested, suggesting that their activities were not cell line–specific (**Supplementary Fig. 2**).

Despite a clear trend, the correlation between ESF activities and their binding affinities to targets was not completely linear (**Supplementary Fig. 1**), as we observed considerable exon skipping or, less prominently, inclusion, between some noncognate pairs of ESFs and targets (**Fig. 1d**: Nanos response element and A6G sequences). This nonlinearity is likely due to a combination of sequence-specific and nonspecific effects of splicing factors. We used a 1:5 ratio of Gly-PUF expression plasmid to splicing reporter plasmid, which represents the lower end of the range that gave robust activity and is substantially lower than the ratios typically used in splicing factor/splicing reporter transfection experiments[10]. We optimized this ratio by titrating the amount

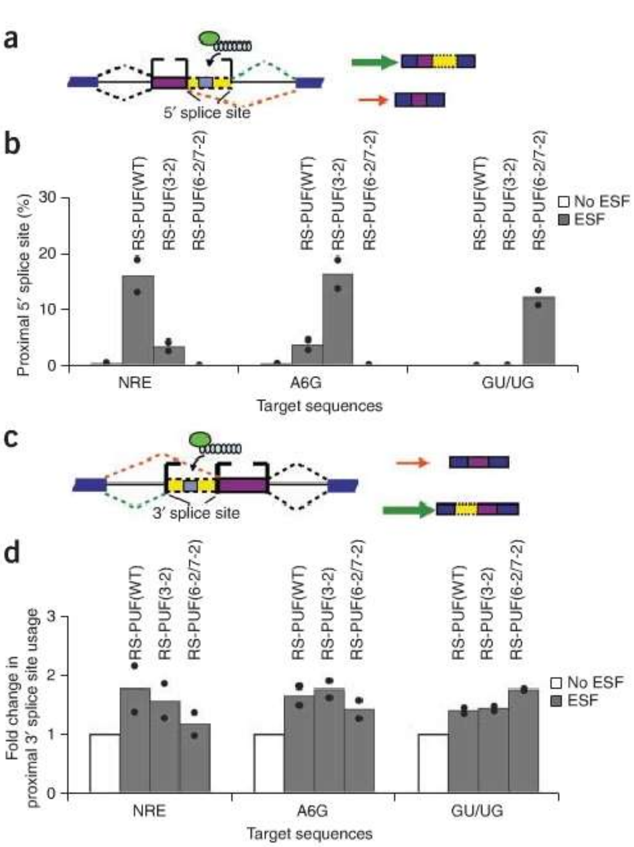


Figure 2 | Effect of ESFs on alternative splice site usage. (**a**) Splicing reporters containing tandem 5′ splice sites and different PUF target sequences (light blue) were transfected into 293T cells along with vectors encoding RS-PUF, which we expected to increase the use of the downstream 5′ splice site (large arrow). (**b**) Percentage of proximal 5′ splice site usage in cells transfected with splicing reporters and ESF expression vectors. Bars show means of two experiments, with dots indicating the data points. Instead of relative changes compared to 'no ESF' controls, we plotted percentage of proximal 5′ splice site usage to avoid exaggerating the relative change because there is little or undetectable proximal 5′ splice site usage in the absence of an ESF. (**c**) Splicing reporters containing tandem 3′ splice sites and different PUF target sequences (light blue) were transfected along with vectors encoding RS-PUF, which we expected to increase the use of upstream 3′ splice site (large arrow). (**d**) Fold changes of proximal 3′ splice site usage relative to 'no ESF' controls. Means of duplicate experiments are plotted, with dots indicating the data points.

of the ESF expression plasmid to a fixed amount (0.2 μg in 1 ml cell culture) of splicing reporter plasmid (**Supplementary Fig. 3**). Higher amounts of splicing factor expression vector caused exon skipping for all reporters, including those with pseudospecific target sequences (**Supplementary Fig. 3**). We had previously observed similar effects of hnRNP A1 and ASF/SF2 in tethering experiments with MS2 coat protein[11]. Therefore, as for other processes involving protein-RNA binding, the sequence specificity of splicing factors can only be considered in the context of certain concentration ranges.

ESFs specifically modulate use of alternative splice sites

In addition to regulating exon skipping or inclusion, some splicing factors regulate alternative use of splice sites by recognizing regulatory sequences between two alternative sites. ASF/SF2 can recognize exonic splicing enhancers and promote the use of intron-proximal sites, whereas hnRNP A1 can bind exonic splicing silencers and shift splicing toward the use of intron-distal sites[12–14]. Therefore we next examined whether ESFs can predictably regulate alternative 5′ or 3′ splice site usage with ESFs containing the same PUF domains (**Fig. 1c**). We designed reporters containing cognate target sequences between tandem 5′ splice sites[11] and transfected them with the RS-PUF–type ESFs. As expected, RS-PUFs increased the use of the downstream intron-proximal 5′ splice sites, with the strongest effect on reporters bearing their cognate sequence (**Fig. 2a,b**). Similarly, the RS-PUFs modulated alternative 3′ splice site usage (by promoting the use of upstream 3′ splice sites) in a sequence-specific fashion (**Fig. 2c,d**), indicating that these ESFs can modulate different alternative splicing events. Owing to the combination of sequence-specific and nonspecific effects, we also found that ESFs can affect splicing of pseudospecific targets in an alternative 3′ splice site reporter.

Examining various functional domains in ESFs

Compared to conventional tethering experiments using the MS2 coat protein or lambda N-B box systems[15], ESFs, which specifically recruit different proteins or domains to certain pre-mRNA regions, can recognize the pre-mRNA in a natural context without introducing foreign RNA and thus are more advantageous for *in vivo* applications. Previous studies using *in vitro* splicing systems had shown that arginine/serine–rich domains from various proteins (or even a short peptide of arginine/serine–rich repeats) can function as splicing activators[3,16]. To test whether other arginine-serine–rich domains can function as the effector module of ESFs, we generated new ESFs by fusing a PUF domain (PUF(3-2)) with arginine/serine–rich domains from other serine/arginine-rich proteins (9G8, SC35 and SRp40) or a short peptide of arginine/serine–rich repeats (six repeats of RS dipeptides $(RS)_6$), and examined whether these new ESFs promote exon inclusion when transfected with their target-containing splicing reporters (**Fig. 3a**). Compared to the non-ESF control, all RS-PUFs promoted inclusion of the cassette exon containing their cognate sequences (**Fig. 3b**). The short arginine/serine–rich repeat had similar activity to the other arginine/serine–rich domains, with the exception of the arginine/serine–rich domain from 9G8, which has slightly higher activity than the rest of the arginine/serine–rich domains (**Fig. 3c**).

Glycine-rich domains are also found in many known splicing factors, especially members of hnRNP A1 family. However, it is unclear whether these glycine-rich domains can function as general splicing inhibitors, as only the domain from hnRNP A1 represses splicing when tethered to the exonic region[4]. To test whether glycine-rich domains have intrinsic splicing inhibitory activity, we generated new ESFs by fusing PUF(3-2) with the glycine-rich domains from additional hnRNP A1 family members (hnRNP A2/B1 and hnRNP A3) or with a glycine-rich short peptide (19 amino acids) and examined whether they can inhibit exon inclusion when transfected along with a splicing reporter (**Fig. 3a**). We found that all Gly-PUF constructs suppressed the inclusion of the cassette exon containing their cognate sequence (**Fig. 3b**), indicating that the glycine-rich domains in members of hnRNP A1 family are likely responsible for their splicing suppression activity. Notably, a 19-amino-acid glycine-rich sequence was sufficient

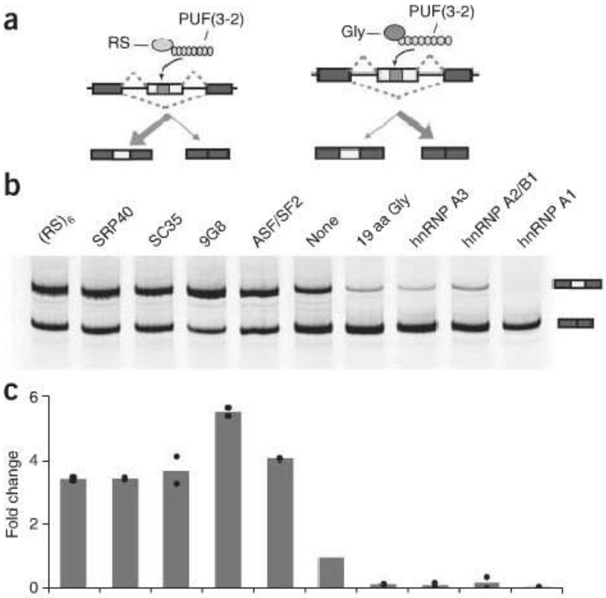


Figure 3 | Various arginine/serine–rich domains or glycine-rich domains can function as the effector module of ESFs. (**a**) Exon-skipping reporters containing the 8-nucleotide cognate sequence of PUF(3-2) ESFs in the cassette exon were used to examine the splicing activator activities of RS domains or fragment (left) or the splicing suppressor activities of glycine-rich domains or fragment (right). (**b**) Promotion or inhibition of exon inclusion by ESFs with indicated splicing effector domains was assayed by RT-PCR. (**c**) Fold change in cassette exon inclusion relative to the reporter alone without an ESF. Samples are in the same order as in **b**. The means of duplicate experiments are plotted, with dots indicating the data points.

for splicing-inhibitory activity (**Fig. 3b–c**), suggesting that other splicing factors with a glycine-rich domain may function as splicing suppressors when binding to exons. PUF domains alone did not affect splicing of the same reporters (**Supplementary Fig. 4**), suggesting that glycine-rich domains are responsible for splicing inhibition. We used the well-studied glycine-rich domain from hnRNP A1 for the subsequent experiments.

Designing an ESF to modulate endogenous gene splicing

An important application for ESFs is to modulate alternative splicing of endogenous genes, particularly disease-associated splicing events. To demonstrate this, we chose to target the *Bcl-X* pre-mRNA that produces two splicing isoforms of opposite function using alternative 5′ splice sites[17] (**Fig. 4a**). The long splicing isoform *bcl-xL* encodes a potent apoptosis inhibitor in long-lived postmitotic cells and is upregulated in many cancer cells to protect them against apoptotic signals. The short splice isoform Bcl-xS is pro-apoptotic and is expressed predominantly in cells with a high turnover rate (for example, developing lymphocytes)[18]. Others have used antisense oligonucleotide–based methods to inhibit the use of the intron proximal 5′ splice site and increase pro-apoptotic Bcl-xS[19–22].

The ratio of the two Bcl-x splicing isoforms is regulated by multiple *cis*-acting elements that are located near the two alternative 5′ splice sites[23]. To shift the splicing of *Bcl-X*, we designed a Gly-PUF–type ESF to recognize the 8-nucleotide sequence in the exon extension region, a sequence that otherwise does not affect splicing (data not shown). This designer ESF should be able to 'reprogram' the splicing regulation of *Bcl-X* to increase Bcl-xS isoform, which in turn can promote apoptosis (**Fig. 4a**). We mutated the first, third and fifth repeats of the wild-type PUF domain and verified that this modified domain (Q867E,Q939E,C935S,Q1011E, C1007S, named PUF(531)) can recognize its new target with very high affinity ($K_d \approx 4$ pM; **Supplementary Fig. 5**) compared to the

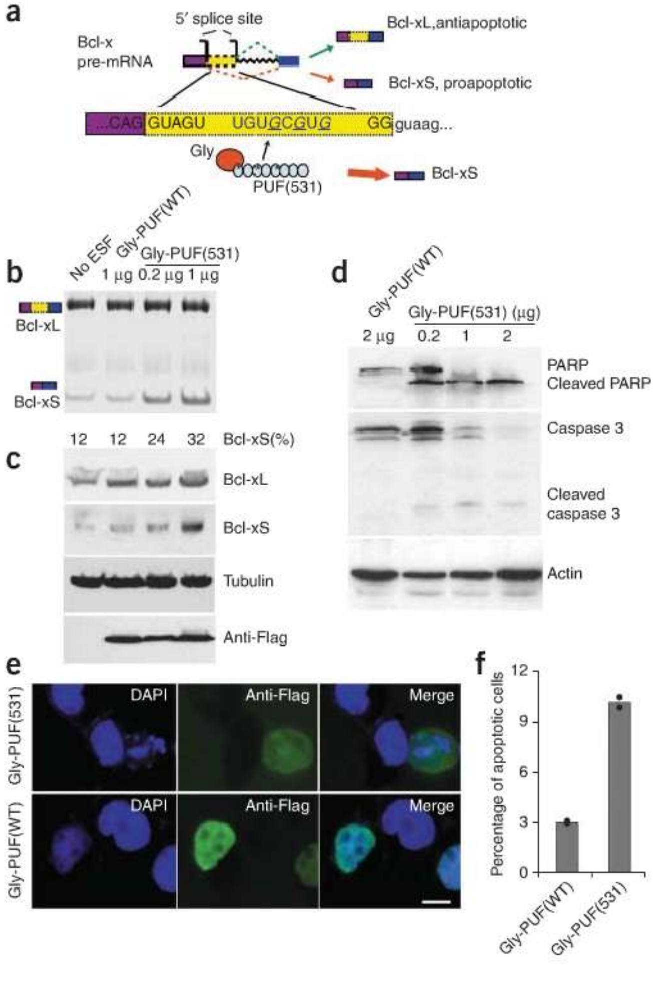


Figure 4 | Design of an ESF to modulate splicing of endogenous *Bcl-X* pre-mRNA. (**a**) PUF(531) was fused with the glycine-rich domain of hnRNP A1 to inhibit downstream 5′ splice sites (large red arrow). (**b**) HeLa cells were transfected with different amounts of expression constructs encoding Gly-PUF(531) and Gly-PUF(WT). Two *Bcl-X* isoforms were detected by RT-PCR and the percentage of *bcl-xS* mRNA was quantified by polyacrylamide gel electrophoresis. (**c**) Western blots of Bcl-xL and Bcl-xS in the presence of ESFs. Samples are in the same order as in **b**. The expression of ESFs is detected with antibody to Flag (anti-Flag), and the tubulin level was detected with an antibody as a control. The blot was exposed for a longer time for Bcl-xS because the available Bcl-x antibody detects Bcl-xL with much higher sensitivity. (**d**) Cleavage of PARP and caspase 3 in HeLa cells transfected with vectors encoding ESF expression constructs was detected by western blot at 24 h after transfection. Actin level was detected with an antibody as a control. (**e**) Fluorescence images showing localization of ESFs in transfected HeLa cells detected with anti-Flag. The cells were co-stained with DAPI to show nuclei. Some nuclei, especially in cells transfected with Gly-PUF(531), are fragmented owing to apoptosis. Scale bar, 5 µm. (**f**) Percentage of apoptotic cells (cells with fragmented nuclear DNA) measured in two independent experiments; >200 cells in total were counted from pictures of randomly chosen fields. Bars show the means, with dots indicating the data from the two experiments.

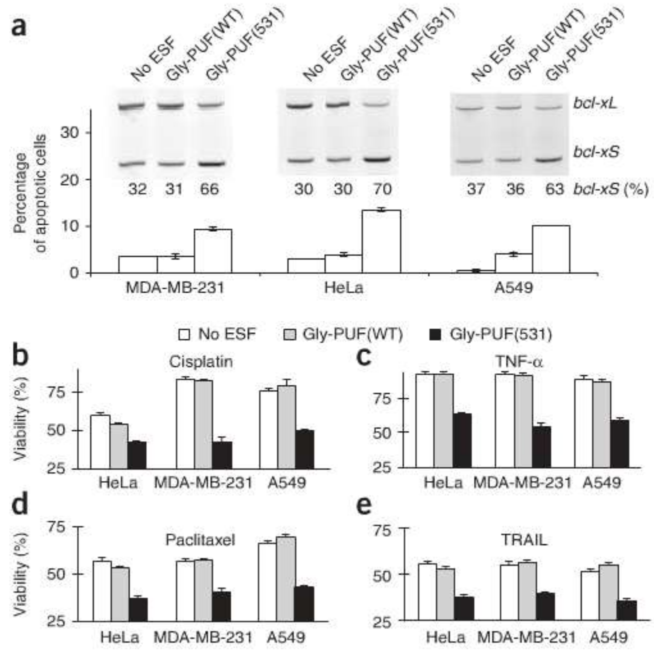


Figure 5 | ESFs affect *Bcl-X* splicing in multiple cancer cells. (**a**) Splicing modulation of *Bcl-X* in three cancer cell lines infected with lentivirus expressing Gly-PUF(531), control ESF or GFP (as 'no ESF' mock infection). *Bcl-X* splicing isoforms were detected by RT-PCR, with percentage of *bcl-xS* quantified. Propidium iodide–stained cells were analyzed by flow cytometry to determine the percentage of dead cells. (**b**) Effect of ESFs on cisplatin sensitivity of different cancer cells. Cells were infected with lentivirus expressing ESFs and controls, and cisplatin was added at 72 h after infection. Cell viability was measured with the WST-1 assay 24 h after drug treatment. Error bars, s.d. (n = 3). (**c**–**e**) Effect of ESFs on the sensitivities to TNF-α (**c**), paclitaxel (**d**) and TRAIL (**e**) in the indicated cell lines. Experimental conditions are the same as described for **b**. The significant differences (P < 0.05, judged by paired t-test) of cell viabilities were observed for all drug treatments between the Gly-PUF(531) and Gly-PUF(WT) infected cells. Error bars, s.d.; n = 3.

wild-type PUF affinity for this sequence ($K_d = 660 \pm 17$ nM). When we transfected the expression plasmid into HeLa cells in which Bcl-xL was the predominant form, Gly-PUF(531) increased splicing of the Bcl-xS isoform in a dose-dependent manner, whereas the control ESF (Gly-PUF(WT)) or PUF(531) alone did not affect the *bcl-xS* mRNA level (**Fig. 4b** and **Supplementary Fig. 4b**). We also observed the increase in the amount of Bcl-xS using western blots (**Fig. 4c**), suggesting that the change of steady-state Bcl-xS/Bcl-xL ratio is probably due to a splicing shift rather than destabilization of *bcl-xL* mRNA by PUF(531) binding. The amounts of Bcl-xL were not notably different under our experimental conditions, probably owing to a higher sensitivity of the Bcl-x antibody in detecting Bcl-xL than Bcl-xS[19,20] (**Supplementary Fig. 6**).

Designer ESF induced apoptosis

Induction of the pro-apoptotic Bcl-xS isoform by Gly-PUF(531) led to cleavage of caspase 3 and poly(ADP-ribose) polymerase (PARP), two known molecular markers in the apoptosis pathway (**Fig. 4d**). Using immunofluorescence microscopy, we observed that many cells expressing Gly-PUF(531) had fragmented nuclear DNA, indicating that they were undergoing apoptosis (**Fig. 4e**). Examination of >200 cells from randomly chosen fields in two independent experiments indicated that ~10% of cells transfected with Gly-PUF(531) expression plasmid had fragmented nuclear DNA versus only ~3% in control cells (**Fig. 4f**). We also noted that, as designed, the ESFs were localized predominantly in the nuclei of transfected cells (**Fig. 4e**), suggesting ESFs dissociate from their targets when the fully spliced mRNAs are transported into the cytoplasm.

In addition to HeLa cells, we tested Gly-PUF(531) in other cells including a breast cancer cell line (MDA-MB-231) and a lung cancer cell line (A549). We infected the cells with lentivirus expressing Gly-PUF(531) or control ESF and found that the designer ESF caused a considerable shift of splicing to produce more Bcl-xS isoform in all cell types tested (**Fig. 5a**). For reasons that are not completely clear, the lentivirus-infected Hela cells had an elevated basal amount of Bcl-xS. Such a splicing shift increased apoptosis as determined by flow cytometry of propidium iodide–stained cells (**Fig. 5a**). In the absence of an exogenous stimulus that induces apoptosis, the increases of apoptotic cells were modest (about threefold) but significant in all cell types tested ($P < 0.05$, paired t-test), consistent with Bcl-xL being an important apoptosis inhibitor for most cancers[18].

Designer ESF sensitized cancer cells to antitumor drugs

As the increase of Bcl-xS released inhibition of apoptosis, we expected a more apparent effect on cell survival during induction of apoptosis by chemotherapeutic drugs. We treated the different cancer cells with either Gly-PUF(531) or control Gly-PUF(WT) in combination with antitumor drugs (cisplatin and paclitaxel) or cytokines (TNF-α and TNF-related apoptosis inducing ligand (TRAIL)), which are commonly used in cancer treatments. To achieve robust expression during the entire period of drug treatment, we infected the MDA-MB-231, A549 and HeLa cells with lentivirus expressing Gly-PUF(531) or control ESF and then treated the infected cells with low doses of drugs for 24 h, conditions under which most mock-infected cancer cells were viable (**Fig. 5b–e**).

In all cell types tested, expression of Gly-PUF(531) sensitized cells to the antitumor drugs tested, leading to significant decreases of cell viability compared to controls ($P < 0.05$) as judged by WST-1 cell proliferation assay. Different cell lines responded to the combinations of antitumor drugs and ESFs with varying sensitivities (**Figs. 5b–e**), consistent with the diverse mechanisms by which these drugs kill cancer cells. The enhancements of drug sensitivity in ESF-treated cancer cells were in the same range as cells treated with small-molecule inhibitors of Bcl-xL[24,25].

DISCUSSION

A design similar to the one presented here could be adopted to study other RNA processing steps (for example, polyadenylation or RNA degradation) that are regulated through interactions between RNA *cis*-elements and protein *trans*-acting factors. The development of ESFs also provides a strategy to manipulate disease-associated splicing events, potentially leading to new treatment for diseases associated with aberrant splicing. Previous methods to modulate disease-related splicing have used antisense oligonucleotides to mask the splicing signals[26] or to recruit additional factors through an extended unpaired tail[27,28], which require relatively high doses of oligonucleotides to effectively change splicing. The design of ESFs uses a fundamentally different strategy: to directly recognize targets through protein-RNA interaction and thus to reprogram the splicing regulation code.

Compared to antisense methods, ESFs are more flexible because they can positively or negatively affect multiple types of alternative splicing. By optimizing various combinations of ESF modules, this approach will allow fine-tuned adjustment of alternative splicing. In addition, ESFs could be stably expressed *in vivo* using the current arsenal of gene therapy and expression tools, whereas the delivery and *in vivo* stability of antisense oligonucleotides are still difficult to control. Finally, in addition to manipulating disease-associated splicing, this new approach has immediate impact as a new system to study factors that regulate splicing or other RNA processing pathways.

The RNA-binding module of our ESFs, the PUF domain, allows target-discriminatory power similar to that of microRNAs, which recognize targets mainly by a 7-nucleotide seed match. Improving ESF specificity will minimize off-target effects, which may be achieved by recognizing longer target sequences using two tandem PUFs or creating a PUF with more repeats. A possible concern with the PUF domain is that repeats can be designed to recognize adenine, guanine or uracil, but amino acid residues that specify recognition of cytosine have not been determined. This might present a limitation if a target sequence must be chosen in a small region. However, a cytosine can be tolerated at position 5 in the target sequence, providing us with the ability to target a sequence with one cytosine. Future work to determine the recognition code for a cytosine by a PUF repeat could address this issue.

In addition to manipulating disease-associated splicing events, ESFs may be designed to create animal models of splicing diseases or to examine the physiological effects of particular alternative splicing choices. Another important application of ESFs is to study the self-regulation of splicing factors. Many splicing factors regulate their own splicing through a feedback loop[29,30], but it is difficult to dissect such a regulatory network because these factors usually also regulate the splicing of essential genes for cell survival. Designing ESFs to direct the splicing of their own pre-mRNAs provides a unique opportunity to construct new splicing regulatory feedback loops or even new regulatory networks, thus could be used to systematically model how splicing regulation is achieved.

METHODS

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

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

ACKNOWLEDGMENTS

We thank C. Burge and our colleagues for critical reading of our manuscript. This work was supported by a grant from the Beckman Foundation (Z.W.) and the Intramural Research Program of the National Institutes of Health, National Institute of Environmental Health Sciences (T.M.T.H).

AUTHOR CONTRIBUTIONS

Z.W. and T.M.T.H. conceived the ideas. Z.W. and Y.W. designed and conducted the splicing experiments. C.-G.C. modified the PUF domains. Z.W. and T.M.T.H. 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/naturemethods/.

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

1. Wang, E.T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
2. Cooper, T.A., Wan, L. & Dreyfuss, G. RNA and disease. *Cell* **136**, 777–793 (2009).
3. Graveley, B.R. & Maniatis, T. Arginine/serine-rich domains of SR proteins can function as activators of pre-mRNA splicing. *Mol. Cell* **1**, 765–771 (1998).
4. Del Gatto-Konczak, F., Olive, M., Gesnel, M.C. & Breathnach, R. hnRNP A1 recruited to an exon in vivo can function as an exon splicing silencer. *Mol. Cell. Biol.* **19**, 251–260 (1999).
5. Wickens, M., Bernstein, D.S., Kimble, J. & Parker, R.A. PUF family portrait: 3′UTR regulation as a way of life. *Trends Genet.* **18**, 150–157 (2002).
6. Wang, X., McLachlan, J., Zamore, P.D. & Hall, T.M. Modular recognition of RNA by a human pumilio-homology domain. *Cell* **110**, 501–512 (2002).
7. Cheong, C.G. & Hall, T.M. Engineering RNA sequence specificity of Pumilio repeats. *Proc. Natl. Acad. Sci. USA* **103**, 13635–13639 (2006).
8. Black, D.L. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* **72**, 291–336 (2003).
9. Wang, Z. & Burge, C.B. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14**, 802–813 (2008).
10. Bai, Y., Lee, D., Yu, T. & Chasin, L.A. Control of 3′ splice site choice in vivo by ASF/SF2 and hnRNP A1. *Nucleic Acids Res.* **27**, 1126–1134 (1999).
11. Wang, Z., Xiao, X., Van Nostrand, E. & Burge, C.B. General and specific functions of exonic splicing silencers in splicing control. *Mol. Cell* **23**, 61–70 (2006).
12. Eperon, I.C. *et al.* Selection of alternative 5′ splice sites: role of U1 snRNP and models for the antagonistic effects of SF2/ASF and hnRNP A1. *Mol. Cell. Biol.* **20**, 8303–8318 (2000).
13. Long, J.C. & Caceres, J.F. The SR protein family of splicing factors: master regulators of gene expression. *Biochem. J.* **417**, 15–27 (2009).
14. Martinez-Contreras, R. *et al.* hnRNP proteins and splicing control. *Adv. Exp. Med. Biol.* **623**, 123–147 (2007).
15. Keryer-Bibens, C., Barreau, C. & Osborne, H.B. Tethering of proteins to RNAs by bacteriophage proteins. *Biol. Cell* **100**, 125–138 (2008).
16. Philipps, D., Celotto, A.M., Wang, Q.Q., Tarng, R.S. & Graveley, B.R. Arginine/serine repeats are sufficient to constitute a splicing activation domain. *Nucleic Acids Res.* **31**, 6502–6508 (2003).
17. Boise, L.H. *et al.* bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. *Cell* **74**, 597–608 (1993).
18. Adams, J.M. & Cory, S. The Bcl-2 apoptotic switch in cancer development and therapy. *Oncogene* **26**, 1324–1337 (2007).
19. Taylor, J.K., Zhang, Q.Q., Wyatt, J.R. & Dean, N.M. Induction of endogenous Bcl-xS through the control of Bcl-x pre-mRNA splicing by antisense oligonucleotides. *Nat. Biotechnol.* **17**, 1097–1100 (1999).
20. Mercatante, D.R., Bortner, C.D., Cidlowski, J.A. & Kole, R. Modification of alternative splicing of Bcl-x pre-mRNA in prostate and breast cancer cells. analysis of apoptosis and cell death. *J. Biol. Chem.* **276**, 16411–16417 (2001).
21. Wilusz, J.E., Devanney, S.C. & Caputi, M. Chimeric peptide nucleic acid compounds modulate splicing of the *bcl-x* gene *in vitro* and *in vivo*. *Nucleic Acids Res.* **33**, 6547–6554 (2005).
22. Gendron, D. *et al.* Modulation of 5′ splice site selection using tailed oligonucleotides carrying splicing signals. *BMC Biotechnol.* **6**, 5 (2006).
23. Zhou, A., Ou, A.C., Cho, A., Benz, E.J. Jr. & Huang, S.C. Novel splicing factor RBM25 modulates Bcl-x pre-mRNA 5′ splice site selection. *Mol. Cell. Biol.* **28**, 5924–5936 (2008).
24. Oltersdorf, T. *et al.* An inhibitor of Bcl-2 family proteins induces regression of solid tumours. *Nature* **435**, 677–681 (2005).
25. Shoemaker, A.R. *et al.* A small-molecule inhibitor of Bcl-XL potentiates the activity of cytotoxic drugs *in vitro* and *in vivo*. *Cancer Res.* **66**, 8731–8739 (2006).
26. Hua, Y., Vickers, T.A., Baker, B.F., Bennett, C.F. & Krainer, A.R. Enhancement of SMN2 exon 7 Inclusion by antisense oligonucleotides targeting the exon. *PLoS Biol.* **5**, e73 (2007).
27. Cartegni, L. & Krainer, A.R. Correction of disease-associated exon skipping by synthetic exon-specific activators. *Nat. Struct. Biol.* **10**, 120–125 (2003).
28. Skordis, L.A., Dunckley, M.G., Yue, B., Eperon, I.C. & Muntoni, F. Bifunctional antisense oligonucleotides provide a trans-acting splicing enhancer that stimulates SMN2 gene expression in patient fibroblasts. *Proc. Natl. Acad. Sci. USA* **100**, 4114–4119 (2003).
29. Jumaa, H. & Nielsen, P.J. The splicing factor SRp20 modifies splicing of its own mRNA and ASF/SF2 antagonizes this regulation. *EMBO J.* **16**, 5077–5085 (1997).
30. Hutchison, S., LeBel, C., Blanchette, M. & Chabot, B. Distinct sets of adjacent heterogeneous nuclear ribonucleoprotein (hnRNP) A1/A2 binding sites control 5′ splice site selection in the hnRNP A1 mRNA precursor. *J. Biol. Chem.* **277**, 29745–29752 (2002).

High-resolution, long-term characterization of bacterial motility using optical tweezers

Taejin L Min[1,2], Patrick J Mears[1,2], Lon M Chubiz[3], Christopher V Rao[3], Ido Golding[1,2,4] & Yann R Chemla[1,2,4]

We present a single-cell motility assay, which allows the quantification of bacterial swimming in a well-controlled environment, for durations of up to an hour and with a temporal resolution greater than the flagellar rotation rates of ~100 Hz. The assay is based on an instrument combining optical tweezers, light and fluorescence microscopy, and a microfluidic chamber. Using this device we characterized the long-term statistics of the run-tumble time series in individual *Escherichia coli* cells. We also quantified higher-order features of bacterial swimming, such as changes in velocity and reversals of swimming direction.

Many microorganisms move around by swimming in liquid medium and can modulate their swimming behavior to move up gradients of chemicals, temperature or light. In liquid environments, *Escherichia coli* swims in a random pattern composed of 'runs', in which the cell maintains an approximately constant direction, and 'tumbles', in which the cell stops and randomly changes direction[1]. Runs and tumbles are generated by different states of the motors that rotate the bacterial flagella. Each cell has several flagellar motors that can rotate either clockwise or counterclockwise. When the motors turn counterclockwise, the flagella rotate together in a bundle and push the cell forward. When one or more of the motors turn clockwise, some flagella may break from the bundle and cause the cell to tumble and randomize its orientation. During chemotaxis, *E. coli* biases its 'random walk' based on temporal changes in chemical concentration. When the bacterium moves up a gradient of attractant, it detects an increase in attractant concentration and reduces its probability of tumbling. The result is that the cell tends to continue going up the gradient.

The modulation of bacterial swimming serves as a model system for the way a living cell processes signals from its environment and changes its behavior based on those signals[1,2]. Standard methods for assaying bacterial swimming and chemotaxis typically fall into two categories. The first consists of observing freely swimming cells, typically in a flow-cell setup. Chemoeffector variation is created in space or time[3–5], and the change in swimming behavior is then examined[6,7]. The second type of assay uses cells that are tethered to a surface, usually a microscope slide, so that the rotation of an individual flagellar motor can be followed[8,9].

These approaches have enabled the acquisition of large amounts of data that have yielded important insights into bacterial swimming and its modulation. However, both assays are limited in their ability to quantify whole-cell swimming (**Supplementary Note 1**). Here we describe an optical trap–based assay to investigate cell motility. This assay allowed us to quantify bacterial swimming in a well-controlled environment for durations up to 1 h and at data acquisition rates that are faster than the ~100 Hz flagellar rotation rates. We thus characterized the long-term statistics of the run-tumble time series in individual cells. Moreover, we characterized higher-order features of bacterial swimming, such as changes in velocity and reversals of swimming direction.

RESULTS

Experimental setup

Our single-cell motility assay involves a custom-made instrument combining optical tweezers, light and fluorescence microscopy and a microfluidic chamber (**Fig. 1a**). The optical tweezers consist of two traps generated by two orthogonally polarized beams from a single 1,064-nm diode-pumped solid-state laser[10]. The separation between the two traps is controlled by a piezo-actuated mirror stage. A custom flow-cell (**Supplementary Fig. 1** and Online Methods) serves as the experimental trap chamber and can be displaced relative to the two traps in all directions by a three-axis translational stage. For measurements of bacterial motility, we filled the chambers with a tryptone broth–based 'trapping medium', though other buffers are also appropriate (Online Methods). We injected bacteria into a top 'antechamber' and flowed them through a narrow inlet into the bottom channel, where they were captured by the traps. Trapping a rod-shaped bacterium by each end with two optical traps[11] allowed us to orient the cell at will in the plane of the chamber (**Fig. 1b**). We visualized trapped bacteria either by brightfield or epifluorescence microscopy (**Fig. 1c** and Online Methods).

Despite immobilization by the optical traps, cells displayed motile behavior, evinced by flagellar bundle rotation and counter-rotation ('rolling') of the cell body[12]. This behavior was detected directly and sensitively by the optical traps themselves, by imaging light from both orthogonally polarized trapping beams onto two separate position-sensitive photodetectors. Consistent with

[1]Department of Physics, [2]Center for the Physics of Living Cells, [3]Department of Chemical and Biomolecular Engineering and [4]Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. Correspondence should be addressed to I.G. (igolding@illinois.edu).

RECEIVED 15 MAY; ACCEPTED 8 SEPTEMBER; PUBLISHED ONLINE 4 OCTOBER 2009; DOI:10.1038/NMETH.1380

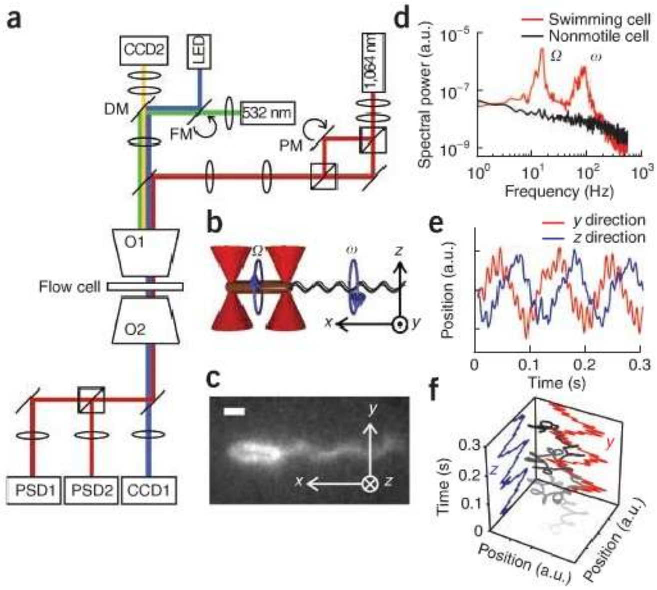


Figure 1 | Combined optical trap and fluorescence microscope setup. (**a**) Instrument layout showing the trapping beam (red), light-emitting diode (LED) illumination path for brightfield imaging (blue), fluorescence excitation beam (green), fluorescence emission (yellow), piezo-actuated mirror (PM), flip-mount mirror (FM), dichroic mirror (DM), microscope objectives (O1 and O2), position-sensitive detectors (PSD1 and PSD2) and charge-coupled device cameras (CCD1 and CCD2). (**b**) Schematic showing optical traps (red cones) and a trapped cell. Circular arrows indicate the rotational direction of the cell body (brown cylinder) and the flagellar bundle (black wavy lines). Also shown is the coordinate axis notation for the optical trap signal. (**c**) Fluorescence image of a Cy3-labeled, optically trapped cell. Scale bar, 1 μm. (**d**) Power spectrum of the optical trap signal from a swimming cell and a nonmotile cell. Swimming cell signal shows oscillatory peaks at 10 Hz and 100 Hz corresponding to body roll (Ω) and flagellar bundle rotation (ω) frequencies, respectively. (**e**) Typical optical trap signal of a swimming cell along y and z directions. (**f**) A three-dimensional plot (grayscale line) of the swimming cell signal. Color darkens with time. Rotational motion of the cell body (large radius rotations) and the flagellar bundle (small radius rotations) are easily recognizable.

previous reports on optically trapped cells[12,13], power spectra from the position-sensitive photodetector outputs upon trapping of a swimming cell revealed two peaks with frequencies ~100 and ~10 Hz (**Fig. 1d**). These oscillatory signals correspond to flagellar bundle rotation and cell body counterrotation or 'roll'[12,13], respectively (**Fig. 1b**). Our measured flagellar rotation (ω) and body-roll rates (Ω) were consistent with those observed in experiments with freely swimming cells[14], demonstrating that the optical traps did not inhibit motility other than in fixing the cell's position. Although we did not observe cell swimming directly, these oscillation frequencies provide information on the motile behavior of the cell (**Supplementary Note 2**).

In a typical experiment, we trapped an *E. coli* cell (strain RP437, wild-type for chemotaxis[15]) horizontally (defined as x in **Fig. 1b**). The motion of each trapped end in the orthogonal plane, along the vertical direction (y) and along the optical axis (z), was detected by one position-sensitive photodetector and revealed both frequencies of oscillation (**Fig. 1e**). The y and z components of the low-frequency signal were 90° out of phase, indicating that the cell end moved in a circular trajectory perpendicular to its body axis (**Fig. 1f**). The rotation was clockwise, as measured looking at the tail of the cell in the direction of swimming, consistent with the expected direction of body roll[1]. The higher-frequency oscillatory signal corresponding to flagellar bundle rotation also revealed a circular motion, in the counterclockwise direction, as expected (**Fig. 1f**).

Of primary importance to our work was characterizing the health of the optically trapped cells. The high photon-flux at near-infrared wavelengths generated by the optical traps has been shown to induce photodamage in cells[16,17]. As demonstrated previously[16], this damage can be largely mitigated by trapping cells under reduced oxygen conditions, for instance, by use of an oxygen-scavenging system. We optimized conditions to enhance cell viability in our trap (Online Methods). Under our tryptone broth–based 'trapping medium' (with oxygen scavenger), we found that trapped *E. coli* cells displayed behavior consistent with healthy cells, growing and dividing at a rate comparable to standard values from the literature (~2 h generation time at room temperature[18]) (**Supplementary Fig. 2**). Furthermore, we could follow swimming in individual cells for extended periods of time in the trap (up to ~1 h; data not shown). Our trapping protocol constitutes a substantial improvement over a previously reported trap-based study of bacterial swimming under oxygenated conditions[13], in which cells could be monitored only for very short times (<10 s).

Observation of single-cell run-tumble behavior

Closer examination of swimming traces revealed regions of alternating oscillatory and nonoscillatory ('erratic') signals (**Fig. 2**). By imaging the motion of a Cy3-labeled cell using epifluorescence microscopy and simultaneously monitoring the trap signals generated by this motion, we established that these oscillatory and erratic signals corresponded to runs and tumbles of the cell, respectively. Cell images taken during oscillatory periods (1.2, 2.2, 2.7 and 3.2 s) displayed a well-formed flagellar bundle extending from the tail of the cell as expected for a run, whereas those taken during erratic periods (1.7 s) exhibit a disrupted bundle, indicative of a tumbling conformation[19] (**Fig. 2a**).

To ascertain that the observed run-tumble behavior in trapped cells is physiologically relevant and rule out the possibility of an artifact induced by the optical traps, we performed two control experiments. In the first, we examined the motility of two mutant strains: a *cheY* deletion (strain CR20; see **Supplementary Table 1** for list of strains used in this study), which does not tumble, and a *cheZ* deletion (strain CR33), which mostly tumbles and does not run. Data traces obtained for these mutants displayed the expected phenotypes (**Fig. 3a–c**): 'runners' generated prolonged oscillatory signals, whereas 'tumblers' underwent continuous erratic motion. In the second control experiment, we quantified the run-tumble behavior of strain PS2001-pMS164 (ref. 20), in which a permanently active CheYD13K mutant protein is expressed from an inducible promoter, under the control of isopropyl β-D-1-thiogalactopyranoside (IPTG). This strain allowed us to modulate run-tumble statistics and to compare them to those obtained with our wild-type strain.

To quantify the swimming behavior of optically trapped cells, we developed an automated run-tumble detection routine using the continuous wavelet transform[21] to discriminate regions of oscillatory and nonoscillatory behavior (Online Methods and

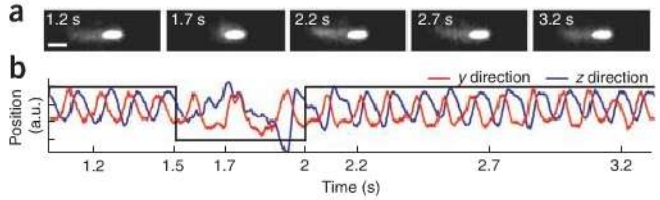


Figure 2 | Direct observation of tumbles in an optically trapped cell. (**a**) Fluorescence images of a trapped cell. Shown in the first frame is the trapped cell body (bright oval shape) and the flagellar bundle (faint cloud) formed to the left of the cell body. The second frame shows the cell tumbling, with the appearance of a disrupted flagellar bundle. Subsequent frames show the reformed flagellar bundle and the running cell. Each frame was obtained by averaging three successive images collected at a rate of 10 Hz, with the marked time point in the middle. Scale bar, 2 μm. (**b**) Optical trap signals in the y and z directions, recorded simultaneously with the fluorescence images. Black lines delineate the run (high) and tumble (low) periods. Only the low-frequency component corresponding to body roll is shown for clarity.

Supplementary Fig. 3). For a dataset of 43 wild-type cells constituting a total of 5,473 detected run events, our algorithm yielded an average run duration of 3.90 ± 0.30 s (mean ± s.e.m., $n = 43$), within the range of previously reported values (0.8–4 s)[7,20,22]. Analysis of 53 PS2001-pMS164 mutant cells at various induction levels revealed that, as expected, run durations were longer than in wild-type cells at low (1 μM) IPTG concentrations and shorter at high (100 μM) IPTG concentrations. The tumble bias, B—defined as the fraction of time the cell spends tumbling, $B = t_{\text{tum}}/(t_{\text{tum}} + t_{\text{run}})$ —exhibited a sigmoidal response to IPTG (**Fig. 3d**). The midpoint of the response was at ~20 μM and the enhancement in bias relative to wild-type cells had a factor of ~4. This behavior is in good agreement with published values[20], confirming our view that tumbles exhibited by trapped *E. coli* represent physiologically relevant events. We note, however, that trapped cells exhibited longer tumble durations than observed in free swimming cells (**Supplementary Note 3**).

Single-cell statistics of motility parameters

The ability to track an individual bacterium for an extended time period (**Fig. 4a**) allowed us to extract single-cell distributions of motility parameters. We determined the cumulative run duration distributions for 43 individual wild-type and 44 individual PS2001-pMS164 cells at a range of induction levels (**Fig. 4b,c** and **Supplementary Fig. 4a–d**). Single-cell run-duration distributions were predominantly exponential but also display considerable cell-to-cell variability. To determine more accurately the shape of the distributions, we normalized each curve by the individual-cell mean run duration (as determined by an exponential fit) along the time axis, maximizing the overlap of the individual distributions[7] (**Fig. 4d,e**).

By pooling all the normalized data, we characterized the 'average' single-cell run duration distribution (**Fig. 4d,e**). Both wild-type and PS2001-pMS164 strain data had exponential distributions at short times, but the former additionally exhibited a pronounced 'heavy tail' corresponding to very long runs, which was much smaller in the mutant strain. The curves for individual wild-type cells also indicate that very long runs are taken by the majority of cells, rather than in a few outliers. Notably, this behavior matched that previously reported in single-motor tethered cell studies[23] and may similarly represent the inherent stochasticity in the chemotactic signaling pathway in wild-type cells. Such a degree of stochasticity was not present in the PS2001-pMS164 strain, in which the concentration of signaling protein CheYD13K is externally controlled. The ability to collect sufficient statistics from individual trapped bacteria provides information not available in population distributions. Note that taking the population averages of the single-cell distributions shown in **Figure 4b** before normalization does not give an accurate representation of the average distribution (**Fig. 4d**), emphasizing the importance of collecting single-cell statistics.

Higher-order features in cell motility

Our preceding analysis of trapped cells characterized their motility in terms of the standard two-state, 'run-tumble' picture. Yet this abstraction of cell swimming is only a first approximation. Researchers in the field have already pointed to aspects of movement beyond this approximation, including changes in cell velocity after a tumble[7], reversal of swimming direction when the flagellar bundle changes its orientation[22,24] and changes in motor and swimming velocity as a function of multiple physiological and mechanical factors[6,13]. Most of these observations, however, were sporadic, limited by the noise or short time duration of available techniques.

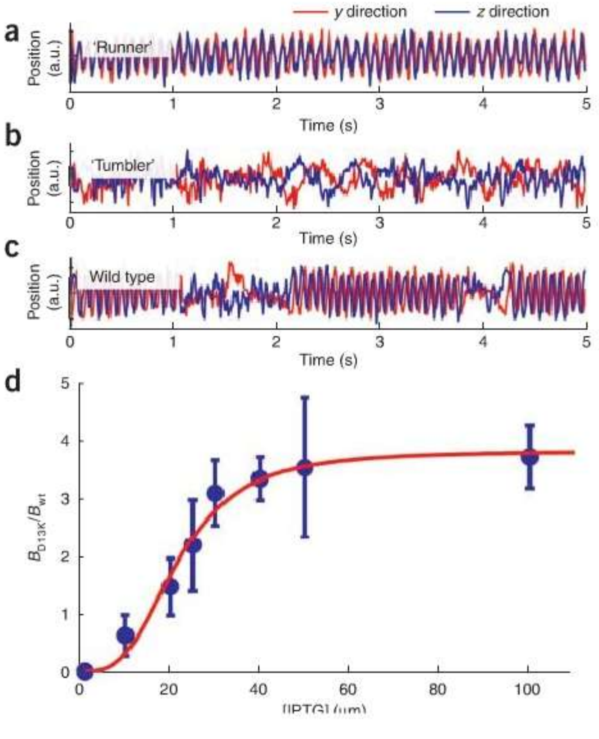


Figure 3 | Run-tumble phenotyping using the optical trapping assay. (**a**) A 'runner' mutant generates predominantly oscillatory signals. (**b**) A 'tumbler' mutant generates predominantly erratic signals. (**c**) A wild-type cell generates oscillatory signals interrupted by intermittent erratic signals. (**d**) Induction response (average tumble bias B_{D13K} of individually trapped cells at various levels of induction, normalized by the average tumble bias of wild-type cells B_{wt}) of the PS2001-pMS164 strain (data points). Higher CheYD13K levels increase the probability of tumbling. Error bars, s.e.m. (n = 6, 8, 8, 5, 13, 4, 3, 6, from lowest to highest IPTG concentration). Fitting to Hill function gives a Hill coefficient of ~3 (red line).

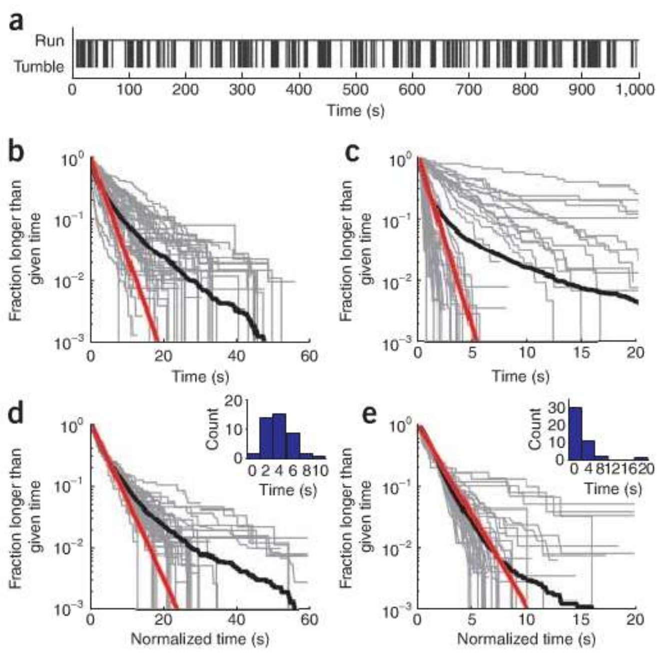


Figure 4 | Run duration statistics in individual bacteria. (**a**) A typical binary time series generated from the swimming signal of a single trapped cell. (**b**,**c**) Cumulative distribution of run durations, comprising 5,473 runs observed in 43 wild-type cells (**b**) and 7,317 runs observed in 44 inducible-bias mutants that showed 20 or more runs (**c**) (black lines). Each gray line shows the run-time statistics from a single cell. The thick black line is the population ensemble. The red line is an exponential fit to 90% of the data, encompassing the shorter events. Each gray line shows the fraction of runs observed from a single cell that were longer than a given time. The red line is an exponential fit to the first decade of the ensemble distribution. (**d**,**e**) Cumulative distribution for wild-type cells (**d**) and inducible bias mutants (**e**) in which individual run duration distributions were scaled so that the mean run duration equals the ensemble mean. This scaling procedure collapses data by effectively removing individual variability, thus revealing the underlying universal behavior in the population ensemble. Inset, histogram of mean tumble durations used in scaling.

Swimming traces collected by our technique also exhibited 'higher-order' swimming dynamics, in particular reversals in phase difference between y and z signals (**Fig. 5a**), indicating reversals in swimming direction (**Supplementary Fig. 5**) and changes in oscillation frequency (**Fig. 5b**), corresponding to changes in swimming speed[13,14]. To fully analyze such higher-order behavioral patterns, we used the continuous wavelet transform to determine not only the Ω values but also the phase difference, $\Delta\phi$, between y and z signals at every point in time (**Supplementary Fig. 6**).

We plotted two-dimensional histograms in Ω and $\Delta\phi$ for two representative cells (**Fig. 5c,d**). The majority of trapped wild-type cells (42 of 43 cells) exhibited reversals, illustrated by the two peaks along the horizontal $\Delta\phi$ axis in the histograms. Reversals occurred frequently (we detected an average of 21 reversal events per time trace) and exclusively after the cell tumbles, on average one out of every six tumbles, or every 21.2 ± 1.1 s (mean ± s.e.m., $n = 859$) (**Fig. 5e**). In certain cases (29 of 42 reversing cells), reversals were also accompanied by an observable change in Ω (**Fig. 5c**), and thus presumably swimming speed. Notably, a similar analysis of the flagellar bundle rotation signal indicated no corresponding changes in ω in the majority of cells (data not shown). These observations suggest that reversals may be important in the motility of cells (**Supplementary Note 4** and **Supplementary Fig. 7**).

Occasionally (in 6 of 42 cells), cells exhibited noticeable, discrete changes in Ω with no corresponding change in swimming direction (illustrated in the two peaks along the vertical Ω axis; **Fig. 5d**). Changes in speed occurred both spontaneously, without tumbling (69.5%) (**Fig. 5f**) or after a tumble (30.5%). Furthermore, the flagellar bundle exhibited no corresponding changes in ω (data not shown). These observations suggest that speed changes may represent different conformational states of the flagellar bundle (**Supplementary Note 5**).

In addition to these higher-order features, many cells exhibited asymmetric $\Delta\phi$ distributions (**Fig. 5c,d**), indicating a bias in swimming direction. Although we found no preferred swimming direction in the cell population, reflecting the fact that our traps do not impose directionality, many individual bacteria did display a distinct bias (**Supplementary Note 6**).

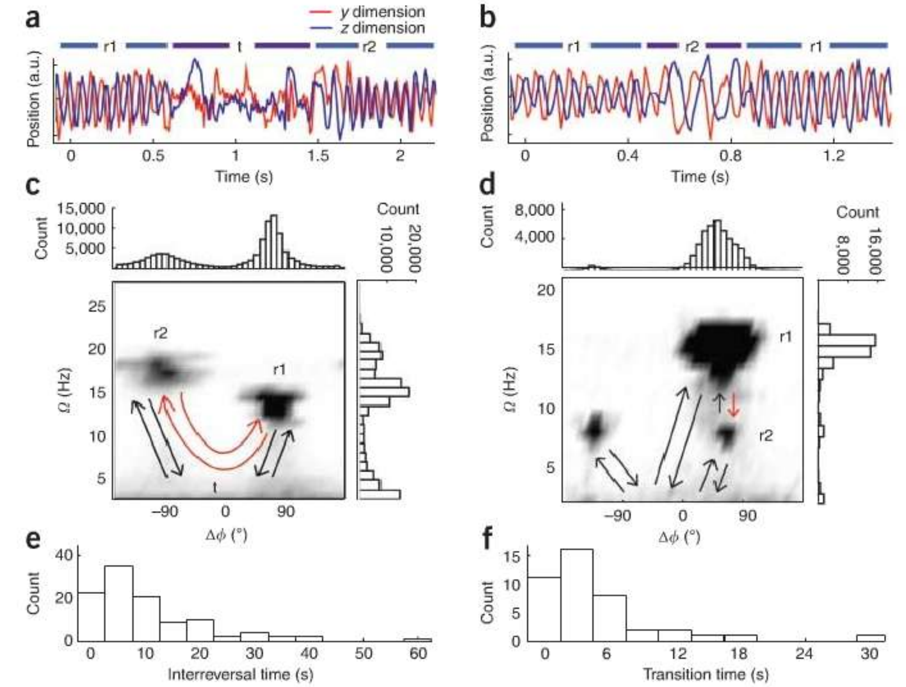


Figure 5 | Higher-order features in cell motility. (**a**) y-dimension and z-dimension signals showing a reversal in run direction (periods designated r1 and r2) after a tumble (designated t). (**b**) y-dimension and z-dimension signals showing change in body roll frequency (Ω) (r2) in the middle of a run (r1). (**c**,**d**) Images are two-dimensional histograms of Ω and phase difference between swimming signals in y and z dimensions ($\Delta\phi$). All possible transitions between different swimming modes are marked by arrows. (**e**,**f**) Waiting time distributions for the transitions highlighted by red arrows in **c** (**e**) and **d** (**f**). Data for **a**, **c** and **e** are from the same cell, and data for **b**, **d** and **f** from another cell.

DISCUSSION

By limiting the physical translocation of the bacterial cell while at the same time allowing high-accuracy measurement of its rotational motion, we followed bacterial swimming for long periods of time with high temporal resolution. The extensive run-tumble statistics thus collected from individual cells expand the range of measured distributions by over an order of magnitude over previous studies[7]. As an example of the consequences of this advance, our wild-type cell run distributions now reveal, to our knowledge for the first time, a pronounced tail similar to that observed in individual flagellar motors[23]. These findings suggest that stochastic variation in the amounts of chemotactic proteins is manifested in the long-term run-tumble behavior of swimming cells.

With our technique we investigated cell swimming beyond the classical, binary run-tumble picture and quantified the statistics of cell reversals, changes in swimming speed and direction bias. Although these features are unique to the swimming behavior of the whole cell (not revealed at the single flagellar motor level), they may also provide insight into the relationship of individual flagella and whole-cell swimming phenotypes. For example, cell reversals may reflect tumbling states in which multiple flagella rotate clockwise and disrupt the flagellar bundle. Tumbles involving a single flagellum are unlikely to induce reversals, as a partial bundle is likely to persist and bias the cell's direction during such tumbles[14]. Changes in swimming speed and bias in swimming direction may similarly reflect different states or spatial arrangements of the flagella.

This technique will be well-suited to investigate chemotaxis in individual cells. A critical requirement for a quantitative characterization of cell chemotactic response is the ability to create an arbitrary stimulus, in the form of spatiotemporally varying chemoeffector concentrations, and to follow the response of the cell in terms of its swimming behavior as well as changes in intracellular parameters such as gene expression. We describe possible approaches to achieving this goal in **Supplementary Note 7**. These advances will enable the development of an integrated device to quantify whole-cell swimming and chemotactic response.

METHODS

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

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

ACKNOWLEDGMENTS

We thank P. Cluzel (Harvard University) for the gift of the PS2001-pMS164 strain and the following people for their generous advice: C. Guet, H. Park, M. McLachlan, K. Neuman, S. Chattopadhyay, W. Ryu, T. Shimizu, R. Segev, G. Ordal, I. Nemenman, T. Emonet and all members of the Golding, Chemla, Rao, Selvin and Ha laboratories. The work was supported by the US National Science Foundation (grant 082265, Center for the Physics of Living Cells). Y.R.C. is supported by Burroughs-Wellcome Fund Career Awards at the Scientific Interface. T.L.M. was supported by National Institutes of Health Institutional National Research Service Award in Molecular Biophysics (PHS 5 T32 GM08276). C.V.R. is supported by National Institutes of Health grant GM054365.

AUTHOR CONTRIBUTIONS

Y.R.C. and I.G. conceived the cell-trapping project. T.L.M. developed the cell-trapping assay. T.L.M. and P.J.M. developed the measurement protocols, performed the experiments and analyzed the data. L.M.C. and C.V.R. constructed and tested bacterial strains used in this study. C.V.R. provided expertise on bacterial physiology and chemotaxis. P.J.M., T.L.M., I.G. and Y.R.C. wrote the paper.

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

1. Berg, H.C. *E. coli in motion*. (Springer, New York, 2004).
2. Alon, U. *An introduction to systems biology: design principles of biological circuits*. (Chapman & Hall/CRC, Boca Raton, Florida, USA, 2007).
3. Brown, D.A. & Berg, H.C. Temporal stimulation of chemotaxis in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **71**, 1388–1392 (1974).
4. Khan, S. *et al.* Excitatory signaling in bacteria probed by caged chemoeffectors. *Biophys. J.* **65**, 2368–2382 (1993).
5. Block, S.M., Segall, J.E. & Berg, H.C. Impulse responses in bacterial chemotaxis. *Cell* **31**, 215–226 (1982).
6. Staropoli, J.F. & Alon, U. Computerized analysis of chemotaxis at different stages of bacterial growth. *Biophys. J.* **78**, 513–519 (2000).
7. Berg, H.C. & Brown, D.A. Chemotaxis in *Escherichia coli* analyzed by 3-dimensional tracking. *Nature* **239**, 500–504 (1972).
8. Silverman, M. & Simon, M. Flagellar rotation and the mechanism of bacterial motility. *Nature* **249**, 73–74 (1974).
9. Sowa, Y. *et al.* Direct observation of steps in rotation of the bacterial flagellar motor. *Nature* **437**, 916–919 (2005).
10. Bustamante, C., Chemla, Y.R. & Moffitt, J.R. High resolution dual trap optical tweezers with differential detection. in *Single-Molecule Techniques: A Laboratory Manual* (eds., P. Selvin & T. Ha) 297–324 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA, 2009).
11. Ashkin, A., Dziedzic, J.M. & Yamane, T. Optical trapping and manipulation of single cells using infrared-laser beams. *Nature* **330**, 769–771 (1987).
12. Rowe, A.D., Leake, M.C., Morgan, H. & Berry, R.M. Rapid rotation of micron and submicron dielectric particles measured using optical tweezers. *J. Mod. Opt.* **50**, 1539–1554 (2003).
13. Chattopadhyay, S., Moldovan, R., Yeung, C. & Wu, X.L. Swimming efficiency of bacterium *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **103**, 13712–13717 (2006).
14. Darnton, N.C., Turner, L., Rojevsky, S. & Berg, H.C. On torque and tumbling in swimming *Escherichia coli*. *J. Bacteriol.* **189**, 1756–1764 (2007).
15. Parkinson, J.S. & Houts, S.E. Isolation and behavior of *Escherichia coli* deletion mutants lacking chemotaxis functions. *J. Bacteriol.* **151**, 106–113 (1982).
16. Neuman, K.C., Chadd, E.H., Liou, G.F., Bergman, K. & Block, S.M. Characterization of photodamage to *Escherichia coli* in optical traps. *Biophys. J.* **77**, 2856–2863 (1999).
17. Rasmussen, M.B., Oddershede, L.B. & Siegumfeldt, H. Optical tweezers cause physiological damage to *Escherichia coli* and *Listeria* bacteria. *Appl. Environ. Microbiol.* **74**, 2441–2446 (2008).
18. Neidhardt, F.C., Ingraham, J.L. & Schaechter, M. *Physiology of the bacterial cell: a molecular approach* (Sinauer Associates, Sunderland, Massachusetts, USA, 1990).
19. Turner, L., Ryu, W.S. & Berg, H.C. Real-time imaging of fluorescent flagellar filaments. *J. Bacteriol.* **182**, 2793–2801 (2000).
20. Alon, U. *et al.* Response regulator output in bacterial chemotaxis. *EMBO J.* **17**, 4238–4248 (1998).
21. Teolis, A. *Computational signal processing with wavelets.* (Birkhäuser, Boston, 1998).
22. Berg, H.C. & Turner, L. Cells of *Escherichia coli* swim either end forward. *Proc. Natl. Acad. Sci. USA* **92**, 477–479 (1995).
23. Korobkova, E., Emonet, T., Vilar, J.M., Shimizu, T.S. & Cluzel, P. From molecular noise to behavioural variability in a single bacterium. *Nature* **428**, 574–578 (2004).
24. Cisneros, L., Dombrowski, C., Goldstein, R.E. & Kessler, J.O. Reversal of bacterial locomotion at an obstacle. *Phys. Rev. E* **73**, 030901 (2006).

High-resolution identification of balanced and complex chromosomal rearrangements by 4C technology

Marieke Simonis[1,5], Petra Klous[1,5], Irene Homminga[2], Robert-Jan Galjaard[3], Erik-Jan Rijkers[4], Frank Grosveld[5], Jules P P Meijerink[2] & Wouter de Laat[1,5]

Balanced chromosomal rearrangements can cause disease, but techniques for their rapid and accurate identification are missing. Here we demonstrate that chromatin conformation capture on chip (4C) technology can be used to screen large genomic regions for balanced and complex inversions and translocations at high resolution. The 4C technique can be used to detect breakpoints also in repetitive DNA sequences as it uniquely relies on capturing genomic fragments across the breakpoint. Using 4C, we uncovered *LMO3* as a potentially leukemogenic translocation partner of *TRB@*. We developed multiplex 4C to simultaneously screen for translocation partners of multiple selected loci. We identified unsuspected translocations and complex rearrangements. Furthermore, using 4C we detected translocations even in small subpopulations of cells. This strategy opens avenues for the rapid fine-mapping of cytogenetically identified translocations and inversions, and the efficient screening for balanced rearrangements near candidate loci, even when rearrangements exist only in subpopulations of cells.

Chromosomal rearrangements (deletions, amplifications, inversions and translocations) occur naturally in the genome[1–6] and often cause disease, particularly when they affect gene expression. This can happen when the rearrangement leads to changes in gene copy number, creates fusion genes or results in a repositioning of regulatory elements such as enhancers. Accurate mapping of chromosomal rearrangements is required to find disease-associated genes and is important both for understanding the mechanism of disease and for optimal diagnosis and treatment decisions. Deletions and amplifications causing copy-number variation can be detected at high resolution using microarray-based comparative genomic hybridization (array-CGH). High-resolution mapping of translocations and inversions not accompanied by loss or gain of DNA can be done at a genome-wide scale by massive parallel paired-end sequencing[7,8]. Given the repetitive nature of the human genome, this is not trivial and massive genome-wide sequencing currently cannot be done routinely for every patient. Detection of balanced rearrangements still mostly relies on molecular-cytogenetic techniques such as chromosomal karyotyping. However, its limited resolution (5–50 megabases (Mb)) necessitates additional labor-intensive experiments to validate genomic breakpoints.

Here we demonstrate that the recently developed chromatin conformation capture on chip (4C) technology[9] can be used to quickly fine-map chromosomal breakpoints in large selected genomic regions that span at least 3 Mb on each side of a selected chromosomal location. Unlike other genomics strategies for mapping chromosomal rearrangements, 4C does not rely on finding the one fragment that carries the breakpoint; rather, it identifies rearrangements based on the capture and identification of many fragments across the breakpoints. As a consequence, 4C can even be used to detect chromosomal breakpoints when they are located within repetitive DNA sequences. Using a single microarray, 4C can be used to analyze multiple selected chromosomal regions simultaneously for breakpoints and rearranged partners throughout the genome, even when the rearrangements are balanced or complex. We used the 4C technology to uncover new rearrangements in cell lines and samples from individuals with T cell–derived acute lymphocytic leukemia (T-ALL) and identified *LMO3* as a potential leukemogenic translocation partner of T cell receptor beta locus (*TRB@*). Finally, we found that 4C, can be used to identify balanced rearrangements even when they are present in small subpopulations of cells. The fact that large megabase-sized regions around target sites are captured efficiently by 4C, no matter the three-dimensional (3D) structure of the DNA, has consequences also for the interpretation of results obtained by other chromatin conformation capture (3C)-based methods, which we discuss here.

RESULTS

4C identifies balanced chromosomal rearrangements

The 4C technology[10] (**Supplementary Fig. 1a**) is based on 3C technology[11]. It involves treatment of cells with formaldehyde to cross-link parts of the genome that are physically close in the nucleus. The DNA is then digested with a restriction enzyme (such as HindIII used here) and cross-linked DNA fragments are ligated. Inverse PCR with primers specific to a selected locus (the 'viewpoint') subsequently allows amplification of its

[1]Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences (KNAW) and University Medical Center Utrecht, Utrecht, The Netherlands. [2]Department of Pediatric Oncology, [3]Department of Clinical Genetics, [4]Department of Biochemistry and [5]Department of Cell Biology, Erasmus Medical Center, Rotterdam, The Netherlands. Correspondence should be addressed to W.d.L. (w.delaat@hubrecht.eu).

RECEIVED 21 JULY; ACCEPTED 16 SEPTEMBER; PUBLISHED ONLINE 11 OCTOBER 2009; DOI:10.1038/NMETH.1391

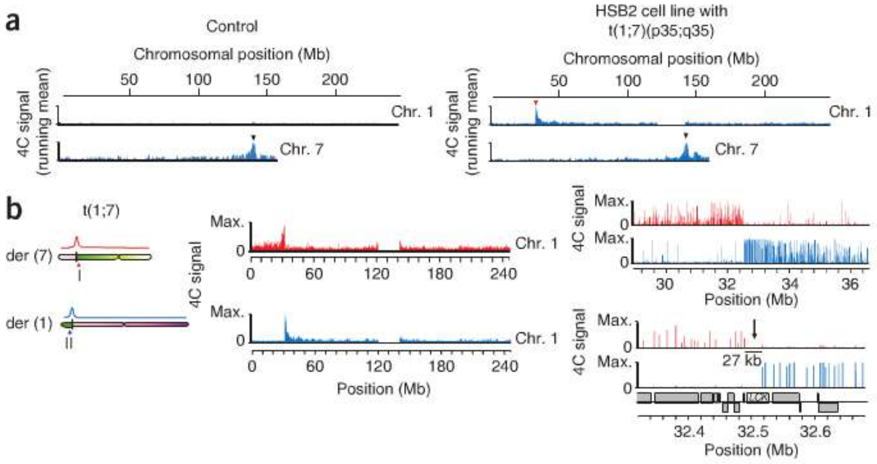


Figure 1 | 4C accurately detects a balanced translocation. (**a**) The 4C signals across chromosomes (chr.) 1 and 7 in cells from a healthy individual and the HSB-2 cell line carrying t(1;7)(p35;q35). The black and red arrowheads indicate positions of viewpoint sequences and translocation site, respectively. Running mean data were plotted, using a window size of ~60 kb. Scale on *y* axes (arbitrary units) is identical for all chromosomes, with the highest mean value set to maximum. (**b**) The 4C signals on chromosome 1, captured by viewpoint fragments I (red) and II (blue) located at opposite sides of the *TRB@* locus on chromosome 7. The regions on chromosome 1 captured by viewpoint fragments on chromosome 7 directly neighbor each other and flank the previously cloned breakpoint (arrow) located in the *LCK* gene. The highest signal in each sample was set to maximum (max.).

interacting partners. When analyzed on a 4C-tailored microarray (385,000 probes) that analyzes the entire human genome at an average resolution of 7 kb (ref. 9), the highest hybridization signals always map to a region of several megabases surrounding the viewpoint sequence (**Supplementary Fig. 1b**). Thus, 4C technology predominantly identifies flanking sequences of the viewpoint, and we reasoned it should therefore detect genomic rearrangements present in this chromosomal area.

To test this hypothesis, we applied the 4C technology to the T-ALL cell line HSB-2. This cell line contains a reciprocal translocation between the *TRB@* locus on band 3, sub-band 5 on the q arm of chromosome 7 (7q35) and the lymphocyte cell–specific protein-tyrosine kinase (*LCK*) locus on 1p35, t(1;7)(p35;q35) (ref. 12). We performed two independent 4C experiments, analyzing DNA interactions with viewpoint sequences located 462 kb centromeric and 239 kb telomeric of the breakpoint in *TRB@*. With both viewpoint sequences, we observed strong hybridization signals not only around the *TRB@* locus on chromosome 7 but also across a megabase-sized region on 1p35, specifically in HSB-2 cells (**Fig. 1**). These signals represented restriction fragments captured by the viewpoint sequences on chromosome 7 as an effect of their close physical proximity. Notably, the first restriction fragments captured on chromosome 1 in both experiments directly flanked the previously identified chromosomal breakpoint. Thus, 4C identifies translocation partners based on the appearance of unusually large clusters of signals on unrelated chromosomes and maps the translocation breakpoint to a position just upstream of the first captured restriction fragment on this chromosome.

Next, we tested whether 4C can be used to identify inversions by applying it to a sample from an individual with T-ALL that, based on a fluorescence *in situ* hybridization (FISH) analysis, carries an inversion on chromosome 7, inv(7)(p15q35). This abnormality leads to the rearrangement of the *TRB@* locus into the *HOXA* gene cluster and activation of the *HOXA* genes[13,14]. The same set of *TRB@* viewpoint sequences described above was used. The telomeric *TRB@* viewpoint sequence captured centromeric *HOXA* fragments and the centromeric *TRB@* viewpoint fragment captured telomeric *HOXA* fragments, thus revealing an inversion between the loci in the sample from an individual with T-ALL (**Fig. 2a**). The two captured regions directly neighbor each other and locate the breakpoint to a 6-kb region. Restriction-fragment paired-end sequencing of the breakpoint (**Supplementary Fig. 2**) confirmed the location of this breakpoint (**Fig. 2a**). Thus 4C technology can detect balanced translocations and inversions at high resolution.

4C identifies unbalanced chromosomal rearrangements

We explored the potential of 4C technology by applying it to an Epstein-Barr virus (EBV) transformed cell line derived from an individual with postaxial polydactyly (PAP). PAP is an autosomal-dominant heritable disorder characterized by extra ulnar or fibular digits. The cells had been previously characterized by karyotyping and FISH analyses and found to contain an unbalanced translocation between chromosomes 4 and 7, t(4;7)(p15.2;q35), with a micro-deletion of unknown size[15]. We performed two 4C experiments, analyzing DNA interactions with two viewpoint fragments located on either side of the rearranged part of chromosome 7. In contrast to what had been found for the balanced translocation, the chromosome 4 fragments captured by the two viewpoint sequences on chromosome 7 did not directly flank each other but were 2.8 Mb apart (**Fig. 2b** and **Supplementary Fig. 3**). We cloned one of the breakpoints and sequenced it, confirming the breakpoint location at 20.08 Mb (**Fig. 2b**). The sequence revealed that the breakpoint on chromosome 7 was more than 3 Mb away from the viewpoint sequence. Thus, 4C viewpoint sequences can be used to capture DNA fragments and characterize rearrangements even when the breakpoints are several megabases away. The data also showed that 4C technology is very suitable to fine-map poorly characterized rearrangements identified by chromosomal karyotyping. When directed to both sides of a genomic breakpoint, 4C can be used to immediately identify whether a translocation or inversion is balanced or accompanied by additional rearrangements such as a deletion (that is, unbalanced).

We next investigated whether 4C technology can be used to identify a deletion not associated with a translocation. For this, we applied 4C to a sample from an individual with T-ALL that was previously characterized by array-CGH to contain a homozygous deletion of the *p15-p16* loci on chromosome 9p21 (J.P.P.M.; unpublished data). Using a viewpoint fragment located ~2 Mb away from the predicted, but not fine-mapped, deletion, we observed a region lacking probe signals, demarcating the deleted area (**Supplementary Fig. 4a**). Notably, we observed increased hybridization signals for the region beyond the deletion. We

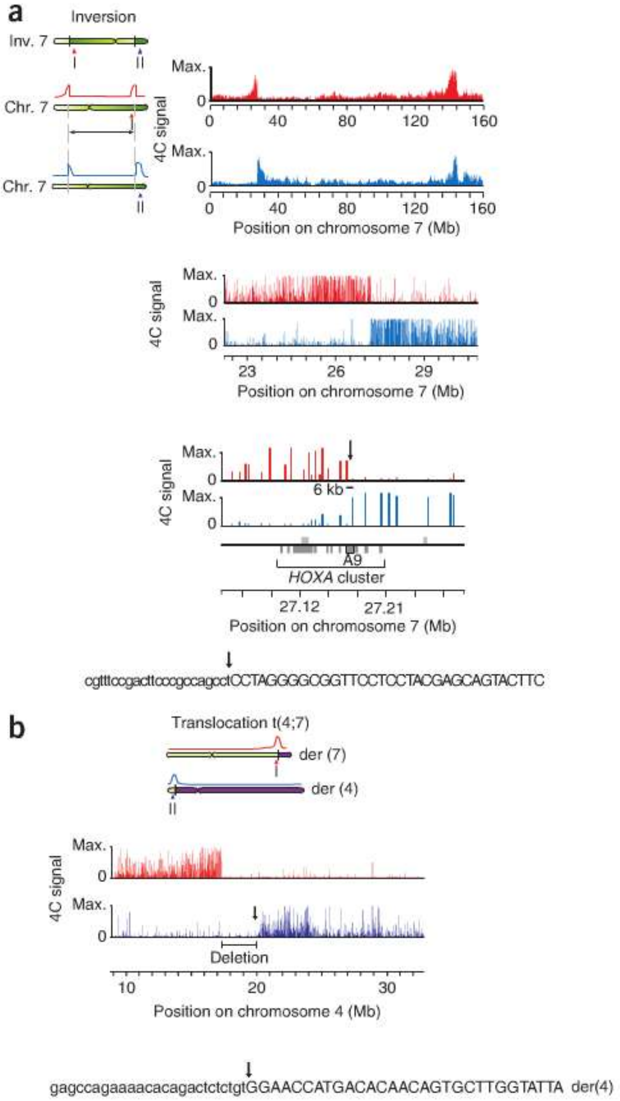


Figure 2 | 4C accurately detects a balanced inversion and an unbalanced translocation. (**a**) Balanced inversion detected in a sample from an individual with T-ALL. The 4C signals on chromosome 7 are fragments captured by viewpoint fragments I (red) and II (blue) located at opposite sides of the *TRB@* locus on chromosome 7. The breakpoint (arrow) was sequenced and confirmed to locate immediately adjacent to the captured sequences. (**b**) The 4C technology accurately detects an unbalanced translocation t(4;7). The 4C data show fragments captured on chromosome 4 by viewpoint fragments I (red) and II (blue) located on opposite sides of the breakpoints on chromosome 7. The two captured regions do not directly neighbor each other, demonstrating a deletion on chromosome 4. The breakpoint of derivate chromosome 4 (arrow) was sequenced. *Y* axes represent raw intensities of the microarray signals; maximum (max.) = 65,535 counts. For whole-chromosome views a running mean with a window size of 29 probes was applied.

expected this because the deletion brings the region in closer physical proximity to the viewpoint fragment. We used the 4C data to predict the breakpoints and design primers for PCR amplification across the ~2-Mb deleted region, which allowed us to map the two breakpoints at the base-pair resolution (**Supplementary Fig. 4b**). This indicates that 4C technology can be used to identify homozygous deletions as regions containing reduced hybridization signals across the deleted area in combination with increased hybridization signals on the other side of the deletion.

The 4C technology identifies translocations in cell mixtures

Tumors and tumor samples are often mosaic, and current high-throughput techniques cannot detect rearrangements present in small subpopulations of cells. We applied 4C to cell mixtures containing control cells (K562) and various amounts of HSB-2 cells carrying the t(1;7) described above. We found that even if only 5% of the analyzed pool of cells carried the translocation, this rearrangement could still be detected efficiently (**Fig. 3** and **Supplementary Fig. 5**). Thus, 4C can be used to identify balanced rearrangements in nonhomogeneous samples, enabling early detection of rearrangements in small tumors.

The 4C identifies new chromosomal rearrangements

We next asked whether 4C can be used to easily identify new rearrangements. *TCR* loci are frequently involved in chromosomal rearrangements in T-ALL cells because translocations can arise during the process of variable-diversity-joining (VDJ) recombination. We screened samples from five individuals with T-ALL for new genetic rearrangements associated with the *TRB@* locus. One sample had a translocation between *TRB@* and the p arm of chromosome 12, plus an additional deletion ~3 Mb away from the translocation breakpoint on chromosome 12 (**Fig. 4** and see **Supplementary Fig. 6** for other chromosomes). The translocation t(7;12)(q35;p12.3) is new in T-ALL. We mapped the breakpoint on chromosome 12 at 6-kb resolution, and cloned and sequenced it (**Fig. 4b**). The translocation positioned the enhancer of *TRB@* 70 kb downstream of the still intact LIM domain only gene *LMO3* (**Fig. 4c**). Microarray expression data showed that *LMO3*, normally silenced in T cells, was highly active in the corresponding T-ALL sample (**Supplementary Fig. 7**). The protein family members *LMO1* and *LMO2*, but not *LMO3*, have previously been found as oncogenic translocation partners of the *TCR* loci in T-ALL. Notably, *LMO3* was recently found to act as an oncogene in neuroblastoma[16]. Thus, 4C technology can be used to discover new oncogenes rearranged with frequently modified loci.

Multiplex 4C technology

Finally, we aimed to develop a 4C strategy that would simultaneously identify multiple recurrent rearrangements associated with a given disease using a single microarray (**Fig. 5** and **Supplementary Figs. 8,9**). In T-ALL, a set of loci, in particular *TRA@-TRD@*, *TRB@*, *BCL11B* and *MLL*, frequently recombine with various other genes[17,18]. We included these four loci, together with nine other loci that have been described either as their translocation partner or to be rearranged otherwise in T-ALL[17,18], in a 4C-multiplex strategy. The genomic sites interacting with each of the 13 viewpoints were PCR amplified separately and pooled in two mixes representing the chromosomal neighborhoods of 6 and 7 viewpoints, respectively. These mixes were differentially labeled and hybridized to the same microarray. The data show that multi-view 4C accurately identifies rearrangements in each of 10 T-ALL samples analyzed. We identified translocations between *TRD@-HOX11* and *TRD@-LMO2*, translocations between *BCL11B-Nkx-2.5* and *BCL11B-TLX3*, the translocation between *CALM-AF10*, the common SET-Nup214 deletion, a *TRA@–c-myc* translocation and an *MLL-AF-6* translocation

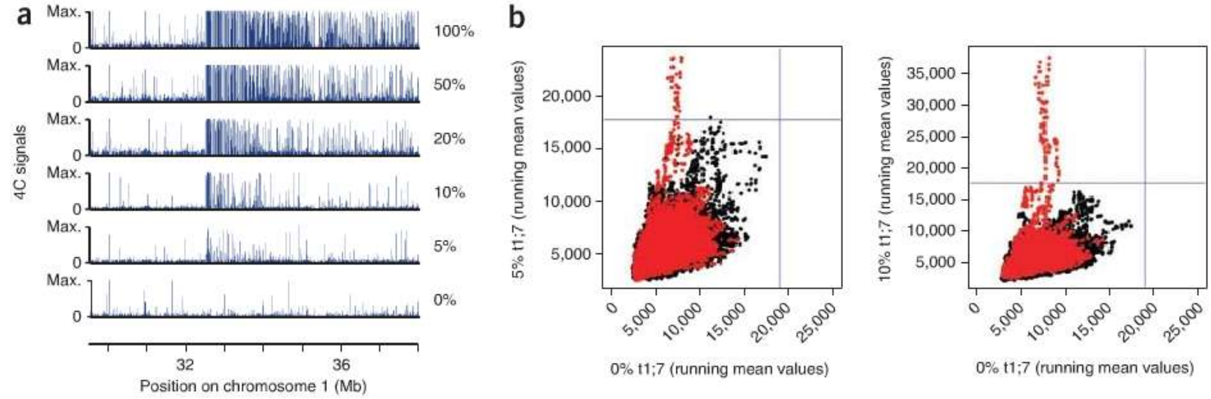


Figure 3 | 4C detects translocations in small subpopulations of cells. (**a**) The 4C technology was applied to mixtures of K562 cells and various amounts of HSB-2 cells, the latter containing translocation t(1;7). Percentages of cells carrying the translocation are indicated on the right. Using a viewpoint neighboring the breakpoint on chromosome 7, sequences on chromosome 1 were captured efficiently in all mixtures, even when only 5% of the cells carried the translocation. Scale on *y* axes are hybridization signal intensities (arbitrary units), with the highest value set to maximum (max.). (**b**) The 4C values from the translocation site clearly separate from other 4C data on *trans* chromosomes. The amount of captured fragments per chromosomal region was determined by calculating running mean values (window size 49 probes, ~300 kb). In the samples containing 5% (left) or 10% (right) HSB-2 cells, higher running mean values were found than in the sample that did not contain cells carrying the t(1;7) translocation (distance to median >3× median, blue line). The high 4C values are located on chromosome 1 (red), next to the breakpoint (**Supplementary Fig. 5**).

(**Fig. 5** and **Supplementary Figs. 8,9**). In one sample we found a novel *LMO1-TRB@* translocation (**Supplementary Fig. 9**). Analogous to *HOX11* (ref. 19,20) and *LMO2* (ref. 21), *LMO1* has now been found to translocate to both *TRB@* and *TRA@-TRD@*[17,18]. Retrospectively, we could confirm all these rearrangements by chromosomal karyotyping and/or FISH analyses (J.P.P.M.; data not shown). In another sample (10110) from an individual with T-ALL, we initially failed to identify the deletion previously mapped by array-CGH to locate between the *LMO2* gene and the locus containing the *RAG1*and *RAG2* genes on chromosome 11 (ref. 22). Instead, using a target sequence ~500 kb telomeric of *LMO2*, we found the region to be fused to an area on chromosome 1 that contained the *STIL* and *TAL1* genes, both also implicated in T-ALL (**Supplementary Fig. 10a**). To further map this rearrangement, we applied 4C to sequences on either side of the breakpoints of the two chromosomes. Together, the data revealed a complex chromosomal rearrangement involving a translocation, t(1;11) (**Supplementary Fig. 10a**). The breakpoint on derivate chromosome 11 was flanked by two small (100–200 kb) regions, one of which carried the T-ALL gene *LMO2*, and the other carrying the T-ALL genes *STIL* and *TAL1*. On either side this was followed by a large deleted region spanning 1–3 Mb (**Supplementary Fig. 10b**). Thus, the deletion identified by array-CGH[22] was shown by 4C to be part of a more complex chromosomal rearrangement involving a translocation. Notably, oncogenes like *LMO2* and *TAL1* are known to translocate to the *TCR* loci in T-ALL, but have not been documented previously to rearrange with each other. The fact that these oncogenes also recombine may support the idea that nuclear co-localization of all these loci at some stage of T-cell development is an important mechanism behind translocation partner selection in T-ALL. Collectively, the data showed that multiview 4C identifies many known and new translocations as well as complex chromosomal rearrangements associated with T-ALL with a single microarray. Clearly, this strategy can also be adapted to detect such rearrangements in samples from patients with other diseases.

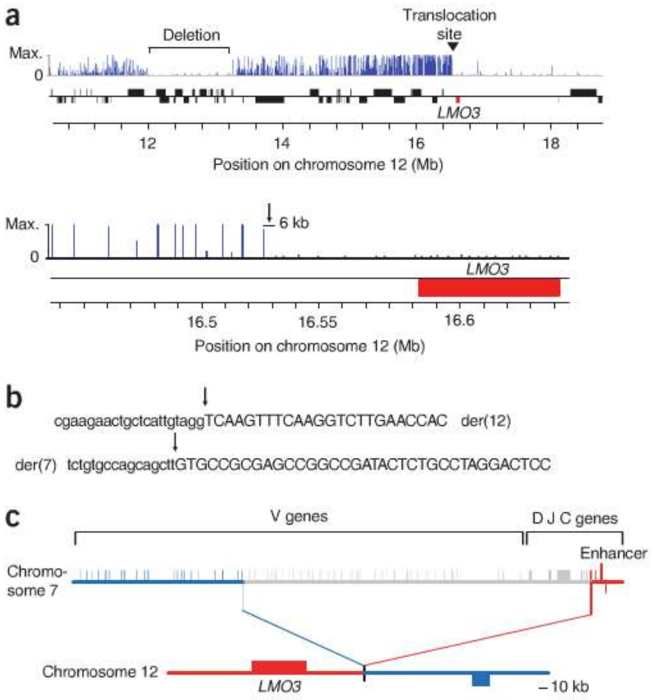


Figure 4 | 4C identifies novel translocation partners. (**a**) Uncharacterized samples from five individuals with T-ALL were screened with 4C, using a viewpoint fragment near the *TRB@* locus on chromosome 7. Shown 4C signals represent fragments on chromosome 12 captured by the viewpoint. Scale on *y* axes are hybridization signal intensities (arbitrary units), with the highest value set to maximum (max.). In one sample, many high 4C signals appeared specifically on chromosome 12, revealing a translocation, t(7;12)(q35;p12.3). A deletion is present several megabases from the translocation site (arrow) on chromosome 12 (top). The translocation site is present in a 6 kb region close to *LMO3* (bottom). (**b**) Sequences of both breakpoints of t(7;12)(q35;p12.3); nucleotides in upper case are from chromosome 12, in lower case from chromosome 7 and in italics are from unknown origin. (**c**) Schematic representation of the translocation site of t(7;12)(q35;p12.3). The enhancer of *TRB@* is positioned 70 kb downstream of the *LMO3* gene.

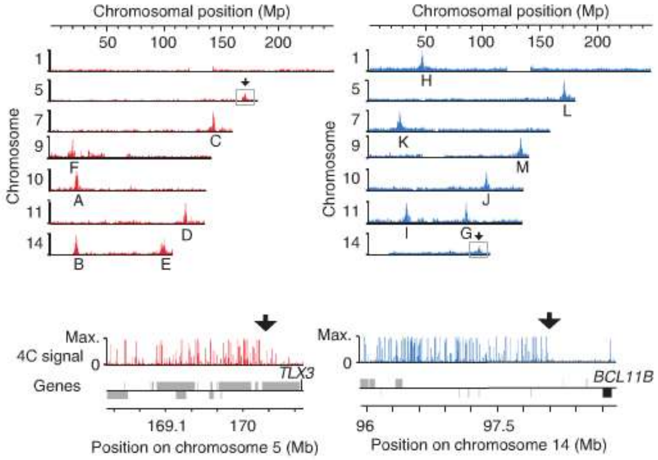


Figure 5 | 4C can be used to analyze multiple sites frequently involved in rearrangements in T-ALL cells on one microarray. The 4C analysis was performed on 13 viewpoints separately: A, *AF10*; B, TRA@-TRD@; C, TRB@; D, *MLL*; E, *BCL11B*; F, *CDKN2A-CDKN2B*; G, *CALM*; H, *TAL1*; I, *LMO2*; J, *HOX11*; K, *HOXA*; L, *TLX3*; and M, *NUP214*. PCR products are combined in two pools. A–F were labeled with Cy-5 (left) and G–M were labeled with Cy-3 (right), and both pools were hybridized to a microarray. The 4C data show high signals on all the viewpoints included in the pools. Additional signals in the Cy-5 pool (boxed) on chromosome 5 demonstrate *TLX3* is a translocation partner of one of the Cy-5 viewpoints. In the Cy-3 pool *BCL11B* was found as a translocation partner (square). Together the data demonstrate the tested sample carries a *BCL11B-TLX3* translocation.

DISCUSSION

The idea to use 4C technology to map genomic rearrangements is based on our observation that 4C technology predominantly identifies large genomic regions flanking the viewpoint[9,10]. Other groups have published similar strategies to uncover long-range DNA interactions in the nucleus[23–26], but these studies did not clearly reveal this important feature of 4C, probably because not enough captured sequences had been analyzed. To prevent the description of an anecdotal collection of interacting DNA elements, we recommend that for each viewpoint sufficient numbers of captured sequences are analyzed by 4C.

Folding of chromosomes is indeed not random, and some regions of the genome are captured and identified by 4C owing to the 3D structure of the chromosomes[9]. When interpreting 4C data, chromosomal rearrangements can easily be discerned from looped chromatin structures. First, signal intensities seen for genomic regions that are physically close on the linear chromosome template are much higher than those observed for distant regions that loop in 3D to the genomic viewpoint. Second, signal profiles from translocations, inversions and deletions each have a very distinct shape that is different from the more Gaussian curves seen for looped regions. Most distinctly, chromosomal rearrangements will yield a sharp transition in signal intensities at the probes that surround the breakpoint. Third and most importantly, the number of probes with positive signals near genomic breakpoints is usually much higher than observed for a region which loops toward the viewpoint. In one example, we found this to be true even in a cell mixture with only 5% of the cells carrying the translocation. Thus, it is not difficult to discriminate structural variation in the genome from 3D configurations detected by 4C. It is important that sufficient numbers of ligation products are analyzed simultaneously. Each diploid cell donates a maximum of two ligation products per viewpoint. For singleplex 4C, we therefore routinely pool at least 15 PCRs, each performed on 200 ng DNA template, for hybridization on a microarray, meaning that we simultaneously analyze ~0.5 million cells, or ~1 million interactions with the viewpoint. Under such conditions, translocations and other rearrangements can be readily identified, even when the breakpoint is 3 Mb away from the viewpoint or present in small subpopulations of cells. In the 4C-multiplex strategy, aimed to simultaneously screen at several sites for structural rearrangements, we performed fewer PCRs per viewpoint and PCR products from different viewpoints were mixed, reducing the net amount of PCR material per viewpoint hybridized to the array. As a result, signal intensities and the number of positive probes identified at the breakpoint dropped, but in each case we could still identify the underlying rearrangement, as confirmed by FISH analysis and by additional 4C experiments. In addition to the complexity of the sample, the distance between the breakpoint and the viewpoint is important. We recommend defining a viewpoint every 3 Mb when screening a large genomic region for breakpoints. The resolution provided by 4C technology is limited by the density of restriction enzyme digestion sites. Here average resolution was 7 kb, but with the same restriction enzyme (HindIII) this can be improved to 4–5 kb when using a higher-resolution array. Cross-hybridization to probes on a microarray always causes undesired background. However, this usually occurs at probes randomly distributed over the genome. It therefore has little impact on the detection of rearrangements by 4C, which depends on chromosomal clustering of probes with strong hybridization signals.

As 4C identifies rearrangements based on the capture of many genomic fragments across the breakpoint, it can be used to uncover balanced rearrangements when breakpoints are present in, or are surrounded by, repetitive sequences. Deletions are currently better detected by, for example, array-CGH, but if deletions are associated with (unbalanced) translocations, they are readily identified by 4C as well and accurately mapped to one of the two derivative chromosomes (**Figs. 2b**, **4** and **Supplementary Fig. 10**). The same is true for large (balanced) inversions, but smaller inversions will be more difficult to identify. Future systematic analyses of samples with inversions varying in size should reveal the detection limit, but we currently estimate that inversions smaller than ~1 Mb cannot be detected by 4C. Possibly the future use of next-generation sequencing instead of microarrays will provide a more quantitative analysis of captured sequences that may improve the detection limit of small rearrangements.

Paired-end sequencing is extremely powerful for genome-wide analysis of both unbalanced and balanced chromosomal rearrangements[7,8]. However, the repetitive nature of the human genome makes the detection of balanced rearrangements by paired-end mapping not trivial, particularly when they are present only in a subpopulation of cells. The 4C technology enables a more focused approach, screening for rearrangements in large genomic regions around candidate loci. Such candidate loci can be regions suspected to carry a rearrangement based on low-resolution techniques such as FISH and chromosomal karyotyping. They may be loci recurrently involved in rearrangements such as the T-cell receptor loci in leukemia samples and the immunoglobulin loci in lymphoma samples. These are large loci that can carry a break anywhere in a region of up to several megabases in size, making it very difficult to design ligation-mediated PCR strategies for the identification of rearrangement partners. Finally, candidate loci may represent genes that are aberrantly expressed without apparent variation in their DNA copy number[27]. We expect that a systematic analysis by 4C focusing on such genes in many samples will lead to the identification of new balanced rearrangements. In this respect, it is relevant to mention that 4C can also be applied to analyze solid tumor material (data not shown).

Array painting is another recently developed technique for fine-mapping translocation breakpoints. It involves isolating chromosomes based on size using flow-sorting and characterization of a selected (derivative) chromosome by hybridization to a microarray or by large-scale sequencing[28,29]. However, not all chromosomes and chromosome derivatives can be isolated based on size, and inversions cannot be detected using this technique. Finally, the fact that 4C can be used to detect translocations present in small subpopulations of cells makes it a potent technique to study rearrangements in mixtures of cells and mosaic tumors and creates the prospect of identifying tumor cells in early stages of metastasis.

METHODS

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

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

ACKNOWLEDGMENTS

We thank A. de Klein, W. van Ijcken, J. Gladdines, J. Veltman, A. Hoischen and E. Splinter for assistance. This work was supported by grants from the Dutch Cancer Society (KWF-EMCR 2006-3500) to J.P.P.M., grants from the European Union and the Cancer Genomics Centre to F.G., grants from the Dutch Scientific Organization (NWO) (912-04-082), Netherlands Genomics Initiative (050-71-324) and the Erasmus MC to W.d.L.

AUTHOR CONTRIBUTIONS

M.S. designed, performed and analyzed experiments and wrote the manuscript; P.K. performed and analyzed experiments; I.H. helped perform experiments; R.-J.G.: provided PAP cell line and helped design the PAP experiment; E.-J.R. designed the microarray. F.G. helped design experiments; J.P.P.M. provided leukemia samples, helped design T-ALL experiments and helped write the manuscript; and W.d.L. designed experiments, supervised the project and 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/naturemethods/.

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

1. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
2. Mehan, M.R., Freimer, N.B. & Ophoff, R.A. A genome-wide survey of segmental duplications that mediate common human genetic variation of chromosomal architecture. *Hum. Genomics* **1**, 335–344 (2004).
3. Sharp, A.J., Cheng, Z. & Eichler, E.E. Structural variation of the human genome. *Annu. Rev. Genomics Hum. Genet.* **7**, 407–442 (2006).
4. Easton, D.F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093 (2007).
5. Eichler, E.E. *et al.* Completing the map of human genetic variation. *Nature* **447**, 161–165 (2007).
6. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
7. Campbell, P.J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40**, 722–729 (2008).
8. Korbel, J.O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
9. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* **38**, 1348–1354 (2006).
10. Simonis, M., Kooren, J. & de Laat, W. An evaluation of 3C-based methods to capture DNA interactions. *Nat. Methods* **4**, 895–901 (2007).
11. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
12. Burnett, R.C., Thirman, M.J., Rowley, J.D. & Diaz, M.O. Molecular analysis of the T-cell acute lymphoblastic leukemia-associated t(1;7)(p34;q34) that fuses LCK and TCRB. *Blood* **84**, 1232–1236 (1994).
13. Soulier, J. *et al.* HOXA genes are included in genetic and biologic networks defining human acute T-cell leukemia (T-ALL). *Blood* **106**, 274–286 (2005).
14. Speleman, F. *et al.* A new recurrent inversion, inv(7)(p15q34), leads to transcriptional activation of HOXA10 and HOXA11 in a subset of T-cell acute lymphoblastic leukemias. *Leukemia* **19**, 358–366 (2005).
15. Galjaard, R.J. *et al.* Isolated postaxial polydactyly type B with mosaicism of a submicroscopic unbalanced translocation leading to an extended phenotype in offspring. *Am. J. Med. Genet. A.* **121**, 168–173 (2003).
16. Aoyama, M. *et al.* LMO3 interacts with neuronal transcription factor, HEN2, and acts as an oncogene in neuroblastoma. *Cancer Res.* **65**, 4587–4597 (2005).
17. Armstrong, S.A. & Look, A.T. Molecular genetics of acute lymphoblastic leukemia. *J. Clin. Oncol.* **23**, 6306–6315 (2005).
18. Graux, C., Cools, J., Michaux, L., Vandenberghe, P. & Hagemeijer, A. Cytogenetics and molecular genetics of T-cell acute lymphoblastic leukemia: from thymocyte to lymphoblast. *Leukemia* **20**, 1496–1510 (2006).
19. Hatano, M., Roberts, C.W. & Minden, M. Crist. W.M. and Korsmeyer, S.J. Deregulation of a homeobox gene, HOX11, by the t(10;14) in T cell leukemia. *Science* **253**, 79–82 (1991).
20. Kennedy, M.A. *et al.* HOX11, a homeobox-containing T-cell oncogene on human chromosome 10q24. *Proc. Natl. Acad. Sci. USA* **88**, 8900–8904 (1991).
21. Boehm, T. Foroni, L., Kaneko, Y., Perutz, M.F. & Rabbitts, T.H. The rhombotin family of cysteine-rich LIM-domain oncogenes: distinct members are involved in T-cell translocations to human chromosomes 11p15 and 11p13. *Proc. Natl. Acad. Sci. USA* **88**, 4367–4371 (1991).
22. Van Vlierberghe, P. *et al.* The cryptic chromosomal deletion del(11)(p12p13) as a new activation mechanism of LMO2 in pediatric T-cell acute lymphoblastic leukemia. *Blood* **108**, 3520–3529 (2006).
23. Ling, J.Q. *et al.* CTCF mediates interchromosomal colocalization between Igf2/H19 and Wsb1/Nf1. *Science* **312**, 269–272 (2006).
24. Lomvardas, S. *et al.* Interchromosomal interactions and olfactory receptor choice. *Cell* **126**, 403–413 (2006).
25. Wurtele, H. & Chartrand, P. Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended Chromosome Conformation Capture methodology. *Chromosome Res.* **14**, 477–495 (2006).
26. Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* **38**, 1341–1347 (2006).
27. Tomlins, S.A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–648 (2005).
28. Chen, W. *et al.* Mapping translocation breakpoints by next-generation sequencing. *Genome Res.* **18**, 1143–1149 (2008).
29. Fiegler, H. *et al.* Array painting: a method for the rapid analysis of aberrant chromosomes using DNA microarrays. *J. Med. Genet.* **40**, 664–670 (2003).

'Edgetic' perturbation of a *C. elegans* BCL2 ortholog

Matija Dreze[1,2,5], Benoit Charloteaux[1,3,5], Stuart Milstein[1,4,5], Pierre-Olivier Vidalain[1,4,5], Muhammed A Yildirim[1], Quan Zhong[1], Nenad Svrzikapa[1,4], Viviana Romero[1], Géraldine Laloux[1,2], Robert Brasseur[3], Jean Vandenhaute[2], Mike Boxem[1,4], Michael E Cusick[1], David E Hill[1] & Marc Vidal[1]

Genes and gene products do not function in isolation but within highly interconnected 'interactome' networks, modeled as graphs of nodes and edges representing macromolecules and interactions between them, respectively. We propose to investigate genotype-phenotype associations by methodical use of alleles that lack single interactions, while retaining all others, in contrast to genetic approaches designed to eliminate gene products completely. We describe an integrated strategy based on the reverse yeast two-hybrid system to isolate and characterize such edge-specific, or 'edgetic', alleles. We established a proof of concept with CED-9, a *Caenorhabditis elegans* BCL2 ortholog. Using *ced-9* edgetic alleles, we uncovered a new potential functional link between apoptosis and a centrosomal protein. This approach is amenable to higher throughput and is particularly applicable to interactome network analysis in organisms for which transgenesis is straightforward.

Classical 'forward' genetics and functional genomics or 'reverse' genetics have together assigned potential function(s) to tens of thousands of genes across dozens of organisms. With the availability of genome sequences and the development of automated phenotypic analyses, reverse genetics strategies based on null or nearly null alleles are rapidly becoming a major source of gene function information. However, functional interpretation of (nearly) null alleles is often complicated because gene products do not operate in isolation but act on each other within complex and dynamic interaction, or 'interactome', networks[1,2].

In interactome graphs, in which macromolecules and interactions between them are represented by 'nodes' and 'edges', respectively, knockouts or knockdowns should be modeled as eliminating a node and all its edges (**Fig. 1a**). More precise determination of molecular function(s) should occur with the development of new systematic strategies to generate alleles that perturb a single interaction, or edge at a time, while maintaining all others unperturbed. Systematic use of such 'edgetic' alleles should be useful to evaluate *in vivo* roles of individual interactions (**Fig. 1b**).

The reverse yeast two-hybrid (R-Y2H) and one-hybrid (R-Y1H) systems rely on genetic selections to identify mutations that disrupt protein-protein and DNA-protein interactions[3–6]. The efficiency of early R-Y2H versions was somewhat limited because most R-Y2H interaction-defective alleles correspond to truncation mutations unless a strategy is used to enrich for missense mutations[5,7–9]. Although dual-reporter systems had been developed to eliminate missense alleles, these systems only allow assaying two partners of a protein simultaneously and are limited if the two partners bind to the same region[10,11].

Here we describe an integrated strategy to systematically isolate edgetic alleles for subsequent *in vivo* characterization. We applied this strategy to *C. elegans* CED-9 (ref. 12), an ortholog of the human antiapoptotic oncoprotein BCL2. We efficiently identified edgetic alleles with various interaction defects caused by specific perturbations of CED-9 binding sites. A subset of *ced-9* edgetic alleles reintroduced *in vivo* caused phenotypes clearly distinct from the *ced-9* null phenotype, and suggestive of a physical and functional link between apoptosis and the centrosome. Our integrated pipeline interrogates interaction networks by perturbing edges instead of nodes and therefore complements the technological arsenal provided by gene knockouts and gene knockdowns to systematically investigate gene function.

RESULTS

Currently available *ced-9* alleles

CED-9 prevents apoptosis by sequestering the Apaf-1 ortholog CED-4 (ref. 13). Apoptosis is triggered when EGL-1 (a BCL2 homology domain 3 (BH3) protein) expression is turned on[14]. By physically interacting with CED-9, EGL-1 triggers conformational changes in CED-9, releasing CED-4 and allowing CED-4–mediated activation of the CED-3 caspase[15,16].

The *ced-9* gene was initially identified through the isolation of a dominant allele, *ced-9(n1950)*, which suppresses apoptosis[12,17]. In this allele, a single amino acid change (G169E) in the encoded protein prevents EGL-1–induced dissociation of otherwise wild-type CED-9/CED-4 complex formation, and thus CED-9(G169E)

[1]Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, and Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [2]Unité de Recherche en Biologie Moléculaire, Facultés Universitaires Notre-Dame de la Paix, Namur, Wallonia, Belgium. [3]Centre de Biophysique Moléculaire Numérique, Faculté Universitaire des Sciences Agronomiques de Gembloux, Gembloux, Wallonia, Belgium. [4]Present addresses: Alnylam Pharmaceuticals, Cambridge, Massachusetts, USA (S.M. and N.S.), Laboratoire de Génomique Virale et Vaccination, Centre National de la Recherche Scientifique, Unité de Recherche Associée 3015, Institut Pasteur, Paris, France (P.-O.V.) and Utrecht University, Utrecht, The Netherlands (M.B.). [5]These authors contributed equally to this work. Correspondence should be addressed to M.V. (marc_vidal@dfci.harvard.edu).

RECEIVED 6 APRIL; ACCEPTED 28 SEPTEMBER; PUBLISHED ONLINE 25 OCTOBER 2009; DOI:10.1038/NMETH.1394

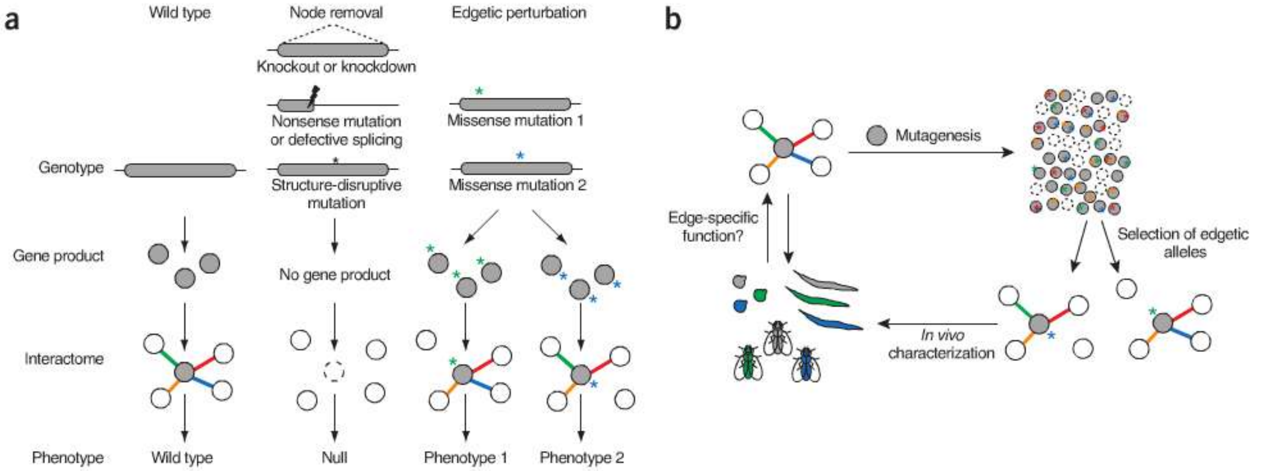


Figure 1 | Schematic representations of genotype-phenotype associations. (**a**) Possible phenotypes resulting from distinct network perturbations caused by different experimental strategies or mutation types. (**b**) Edgetic strategy applied to a protein of interest (gray node). Colors represent specific edges, their specific perturbation and the specific corresponding phenotypes. Dashed circles represent absent, truncated or unstable gene products.

can be considered edgetic by our definition. All four additional alleles currently available (*ced-9*(*n2812*), *ced-9*(*n2077*), *ced-9*(*n2161*) and *ced-9*(*n1653*ts)) result in complete or near complete CED-9 depletion, that is, CED-9 node removal (**Supplementary Fig. 1**).

Isolation of *ced-9* edgetic alleles

We tested whether our strategy could be used to isolate alleles encoding CED-9(G169E)-like proteins. We first determined that Y2H is suitable to (i) detect the CED-9/CED-4 interaction, (ii) reconstitute the EGL-1–induced dissociation of this interaction and (iii) recapitulate the CED-9(G169E) edgetic profile (**Fig. 2a**,**b** and **Supplementary Data 1**). We then developed a Y2H-based scheme starting from a library of random *ced-9* mutations produced by PCR amplification of a truncated *ced-9* gene, encoding CED-9 lacking its C-terminal transmembrane domain (CED-9ΔTM). The library was enriched for full-length open reading frames (ORFs) (**Fig. 2c** and **Supplementary Table 1**). From this library, we selected for alleles encoding CED-9(G169E)–like proteins based

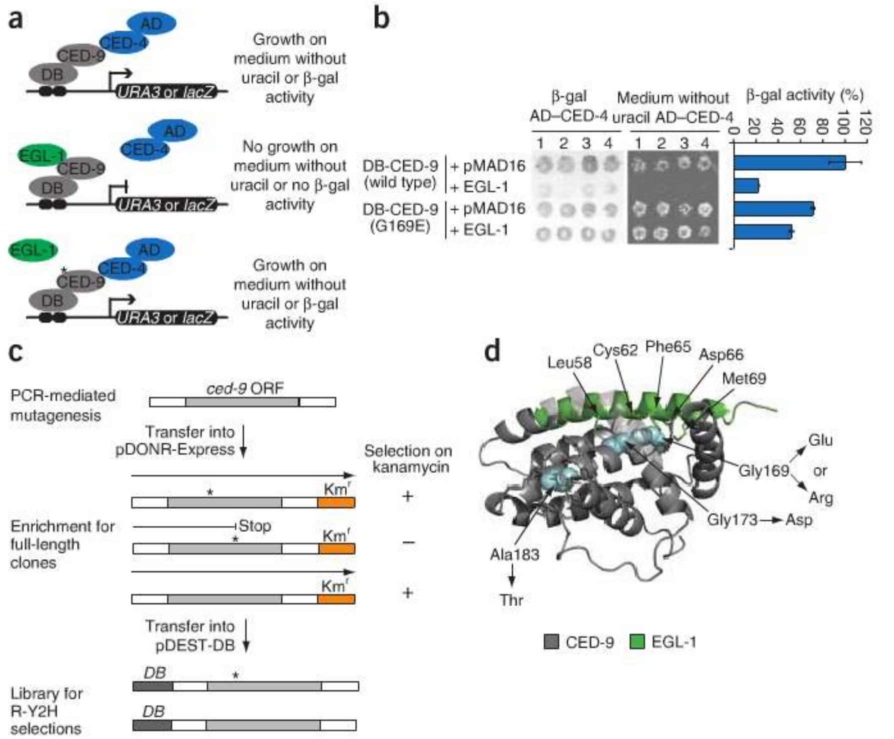


Figure 2 | Isolation of *ced-9* alleles insensitive to EGL-1. (**a**) Schematic of a modified Y2H assay used to identify edgetic alleles that maintain CED-9 interaction with CED-4 in presence of EGL-1. DB, Gal4 DNA-binding domain. AD, Gal4 activation domain. β-gal, β-galactosidase. (**b**) Y2H phenotypes of the interaction between AD–CED-4 and DB–CED-9 (wild type or G169E) in the absence or presence of EGL-1. Each assay was performed in quadruplicate (1–4): filter β-galactosidase assay (β-gal; left), growth assay on medium without uracil (middle) and a quantitative β-galactosidase assay (β-gal activity; right). Error bars, s.e.m. ($n = 4$). (**c**) To generate the CED-9ΔTM mutant library, ORFs mutagenized by PCR were cloned by Gateway reaction into pDONR-Express, a bacterial expression vector containing a kanamycin resistance–encoding gene (Km^r) placed in-frame with the ORF cloning site[8]. The selection of *Escherichia coli* transformants on kanamycin-containing plates was designed to eliminate nonsense mutations and out-of-frame changes, enriching the library with full-length ORFs that can then be transferred into the pDEST-DB Y2H vector by Gateway reaction for R-Y2H selections. White boxes surrounding *ced-9* ORF represent Gateway recombination sites. (**d**) Crystal structure of a CED-9/EGL-1 complex (Protein Data Bank (PDB) identifier: 1TY4)[18] with residues mutated in proteins encoded by *ced-9* edgetic alleles indicated in blue. Substitutions are indicated. EGL-1 residues less than 4 Å away from CED-9 mutated residues and CED-9 residues less than 4 Å away from A183 are shown as sticks. For clarity hydrogen atoms have been omitted.

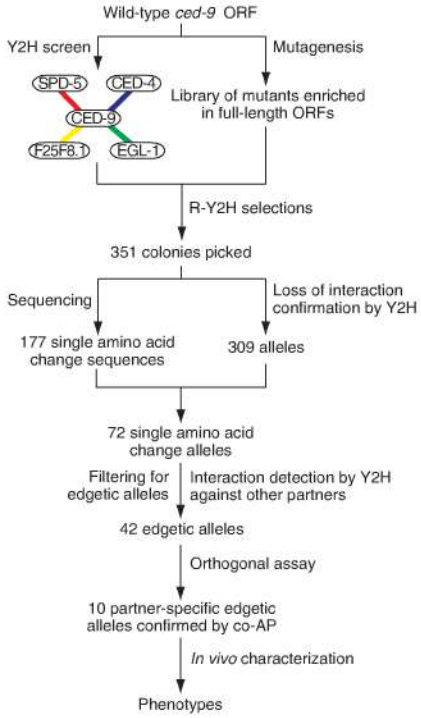

Wild-type *ced-9* ORF
Y2H screen
Mutagenesis
SPD-5
CED-4
CED-9
F25F8.1
EGL-1
Library of mutants enriched in full-length ORFs
R-Y2H selections
351 colonies picked
Sequencing
Loss of interaction confirmation by Y2H
177 single amino acid change sequences
309 alleles
72 single amino acid change alleles
Filtering for edgetic alleles
Interaction detection by Y2H against other partners
42 edgetic alleles
Orthogonal assay
10 partner-specific edgetic alleles confirmed by co-AP
In vivo characterization
Phenotypes

Figure 3 | Outline of the edgetic strategy. The interaction network of CED-9 was mapped by Y2H and confirmed by co-AP. In parallel, a CED-9ΔTM mutant library enriched for full-length ORFs was generated. Interaction-defective alleles were isolated by R-Y2H from the CED-9ΔTM mutant library. *ced-9* alleles PCR-amplified directly from yeast colonies were sequenced to identify potential mutations, and then reintroduced by gap repair into fresh yeast cells to confirm loss-of-interaction phenotypes by Y2H. Interaction-defective alleles were subsequently tested by Y2H against other CED-9 partners to distinguish between edgetic and non-edgetic alleles. The interaction profiles of a subset of partner-specific edgetic alleles were validated by co-AP. Validated alleles were expressed *in vivo*, and phenotypic consequences were examined.

on their ability to maintain the interaction with CED-4 in the presence of EGL-1, thus conferring growth on selective medium lacking uracil (**Fig. 2a**).

We identified four such alleles. Two of these alleles encoded proteins with substitutions of Gly169, the amino acid mutated in the protein encoded by *ced-9(1950)*, the dominant allele originally isolated in a forward genetic screen (**Fig. 2d**). One change corresponded exactly to the previously described *ced-9(1950)* G169E substitution, demonstrating the power of our Y2H genetic selection. The other change, G169R, was new but similar to G169E (substitution of a glycine by a bulky charged residue). The third allele encoded a protein with a G173D mutation, a substitution of a glycine close to Gly169 in the CED-9 sequence. In the CED-9/EGL-1 co-crystal[18], Gly169 and Gly173 are adjacent and are both in contact with the EGL-1 BH3 peptide (**Fig. 2d**). The A183T substitution in the protein encoded by the fourth allele affects a residue outside of the EGL-1 BH3 binding groove but in the same α-helix as Gly169 and Gly173. An A183Y substitution decreases the melting temperature of the CED-9/EGL-1 complex by 5 °C (ref. 19), consistent with our observation that the A183T mutation affects the CED-9/EGL-1 interaction. Altogether we found that a genetic selection in the appropriate yeast strain background can efficiently isolate *ced-9* edgetic alleles.

Integrated strategy to isolate edgetic alleles

We generalized the approach for new interactions (**Fig. 3**). We screened CED-9ΔTM against *C. elegans* cDNA[20] and ORFeome[21] libraries by Y2H. We recovered both EGL-1 and CED-4 as CED-9ΔTM interactors, validating our Y2H screen[14,22]. We also identified two new Y2H interactors: residues 829–1,198 of SPD-5 and full-length F25F8.1. SPD-5 is a centriole protein essential for centrosome maturation and mitotic spindle assembly[23]. F25F8.1 is an uncharacterized protein with no known orthologs outside of the *Caenorhabditis* genus. We validated all Y2H interactions by co-affinity purification (co-AP) in HEK293T cells using glutathione S-transferase (GST) pulldown[21] with CED-9ΔTM and full-length EGL-1, CED-4, SPD-5 and F25F8.1 (**Supplementary Fig. 2**). Having validated SPD-5 and F25F8.1 as likely bona fide biophysical interactors, we used our edgetic strategy to interrogate the biological relevance of these interactions.

From the CED-9ΔTM mutant library (**Figs. 2c** and **3**), we used R-Y2H to select mutants unable to interact with either CED-4 or SPD-5, using *SPAL10::URA3* (ref. 6), a counterselectable marker that causes toxicity in the presence of 5-fluoroorotic acid (5-FOA)[24]; loss of interaction results in the ability to grow on plates containing 5-FOA. As the CED-9/F25F8.1 interaction does not confer 5-FOA sensitivity, we screened for *ced-9* mutants unable to interact with F25F8.1 by looking for decreased *GAL1::lacZ*–induced β-galactosidase activity (**Supplementary Data 2**). We obtained a total of 351 potential *ced-9* alleles, with 192, 144 and 15 of their gene products unable to interact with CED-4, SPD-5 and F25F8.1, respectively.

After PCR amplification of these potential alleles, we sequenced them and analyzed interactions by Y2H against all CED-9 partners to confirm the interaction defects and determine their specificity (**Fig. 3**, **Supplementary Figs. 3–5**, **Supplementary Tables 2–5** and **Supplementary Data 2**). We found 42 alleles with an edgetic profile (that is, disrupting one or a subset of interactions), each affecting one of 33 different amino acids along the CED-9 sequence (~13% of the sequence; **Supplementary Table 5**). In contrast, 30 alleles impaired all CED-9 binding capacities, and we therefore considered them non-edgetic.

We used co-AP pulldowns in HEK293T cells as an orthogonal protein-interaction assay[21,25]. We tested 16 partner-specific edgetic alleles encoding proteins defective in their ability to interact with CED-4 (five alleles), SPD-5 (nine alleles) or F25F8.1 (two alleles). We validated a substantial proportion of R-Y2H edgetic alleles (10/16) by co-AP (**Supplementary Fig. 6** and **Supplementary Data 2**).

Structural analysis of edgetic and non-edgetic residues

Although our edgetic strategy does not require a priori knowledge of tertiary structure, we could use such information to investigate the properties of residues mutated in edgetic and non-edgetic alleles ('edgetic' and 'non-edgetic' residues, respectively). To assess whether affected residues were preferentially located in protein-binding sites, we quantified their surface exposure in the CED-9 tertiary structure (**Fig. 4a**). We defined as solvent-accessible the residues with 10% or more of solvent-accessible surface area in at least one of the three available CED-9 crystal structures[16,18,19],

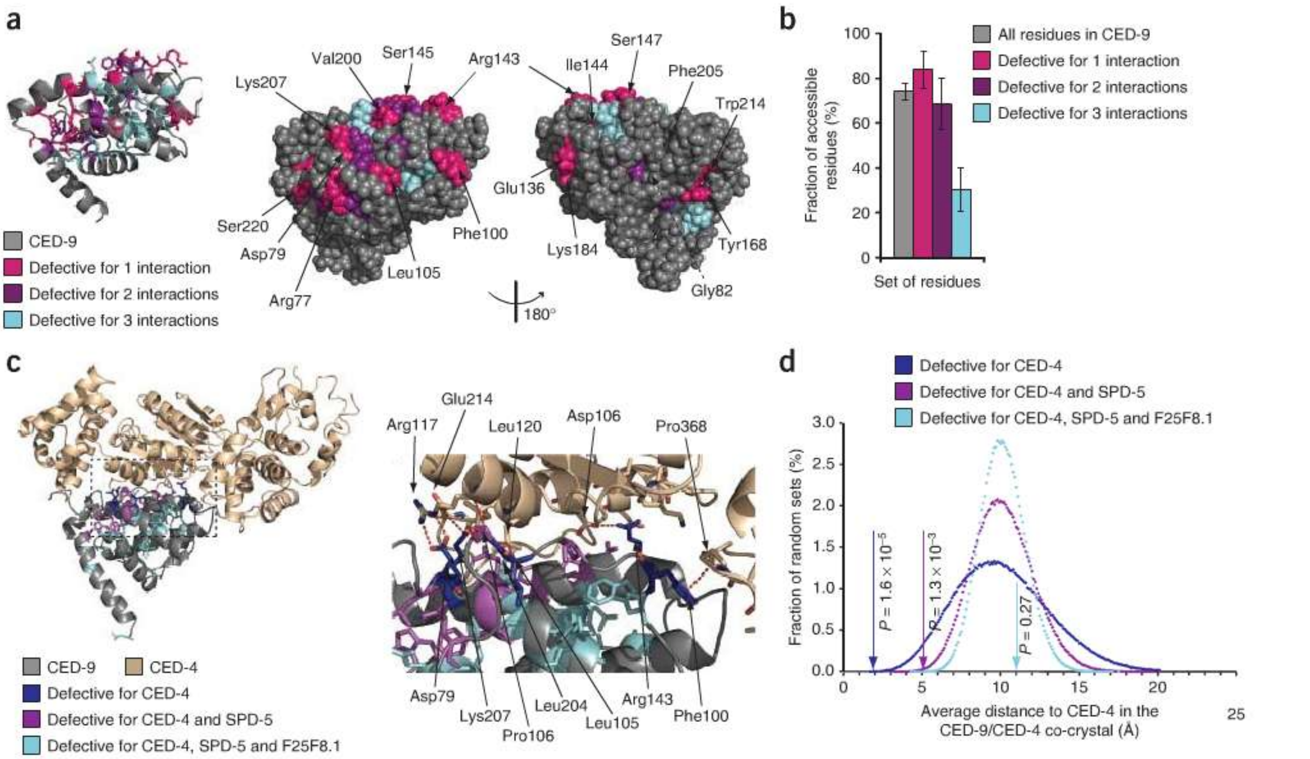


Figure 4 | Edgetic and non-edgetic residues in CED-9 and CED-9/CED-4 structures. (**a**) Ribbon diagram of CED-9 (PDB: 1OHU)[19] (left); residues mutated in R-Y2H alleles are shown as sticks (hydrogen atoms omitted). Space-filling representation of the same structure in the identical (middle) and opposite (right) orientations. Residues mutated in alleles defective for only one interaction are labeled. Gly82 (dashed line) is buried. (**b**) Fraction of residues (all residues versus the residues mutated in the indicated sets of *ced-9* alleles) accessible in at least one of the three CED-9 structures[16,18,19] using a 10% solvent-accessible surface area cutoff. Error bars, standard error for a binomial distribution. (**c**) Ribbon diagram of CED-9 complexed with one CED-4 monomer (PDB: 2A5Y)[16] (left); residues mutated in CED-4 interaction–defective alleles are shown as sticks. In the close-up of the same view (right), CED-4 residues that interact with CED-9 residues mutated in CED-4-specific edgetic alleles are also shown as sticks. Red dashed lines, interactions. Oxygen atoms are red and nitrogen atoms blue. Hydrogen atoms are omitted. (**d**) Distribution of the average distance to CED-4 in the CED-9/CED-4 co-crystal obtained for 1,000,000 random sets of 6, 14 or 24 residues as compared to the average distance (arrows) of the residues mutated in the indicated sets of CED-4 interaction–defective alleles.

taking into account variations between these structures. Edgetic residues, especially those mutated in proteins defective for one interaction, were on average more accessible than non-edgetic residues (**Fig. 4a**,**b** and **Supplementary Data 3**). The average surface exposure we observed for non-edgetic residues was significantly lower than expected by chance ($P < 10^{-6}$; empirical *P*-value; **Supplementary Fig. 7**).

These observations suggest that edgetic alleles of *ced-9* targeted relatively more accessible residues that are likely part of interaction regions. In contrast, non-edgetic alleles were defective for all three CED-9 interactions because of disruptive substitutions in the CED-9 core. The non-conservative nature of these substitutions corroborated this explanation (**Supplementary Table 5**). For instance, ten non-edgetic alleles (~1/3) encoded a protein with an α-helix residue mutated to proline. Notably, two non-edgetic alleles that we isolated contained a substitution of Y149, which is mutated in the protein encoded by the *ced-9*(*n1653*ts) allele (**Supplementary Fig. 1**) and is crucial for CED-9 structure[18]. This finding supports the proposal that non-edgetic alleles are defective for all interactions because of a disrupted CED-9 tertiary structure.

If non-edgetic mutations disrupt CED-9 tertiary structure and edgetic mutations affect specific interaction regions, non-edgetic alleles should tend to encode relatively unstable proteins. We expressed wild-type CED-9 as well as proteins encoded by 14 edgetic alleles and 14 non-edgetic alleles, all as GST fusion proteins in human cells (**Supplementary Fig. 8**). Proteins encoded by edgetic alleles were expressed at levels comparable to that of wild-type CED-9 fusion protein. In contrast, the non-edgetic mutant proteins could not be detected or were expressed at much lower levels than wild-type CED-9. As expected, two CED-9 truncated proteins also had reduced expression (Stop1 and Stop2; **Supplementary Fig. 8**). Non-edgetic mutations in Tyr149 (Y149C and Y149H) resulted in decreased stability and poor expression of CED-9 as previously reported for the Y149N mutant[18]. These data strongly suggest that edgetic and non-edgetic alleles result from distinct molecular defects: destabilization and degradation for the non-edgetic alleles and more subtle changes that do not affect overall stability for edgetic alleles.

Edgetic and non-edgetic residues in binding sites

To test our model for the structural basis of edgetic versus non-edgetic alleles, we took advantage of the CED-9/CED4 co-crystal[16], locating in this quaternary structure the residues mutated in all alleles defective for CED-4 interaction (**Fig. 4c**). All six distinct residues mutated in the alleles defective only for the CED-4 interaction were located at the CED-9/CED-4 interface,

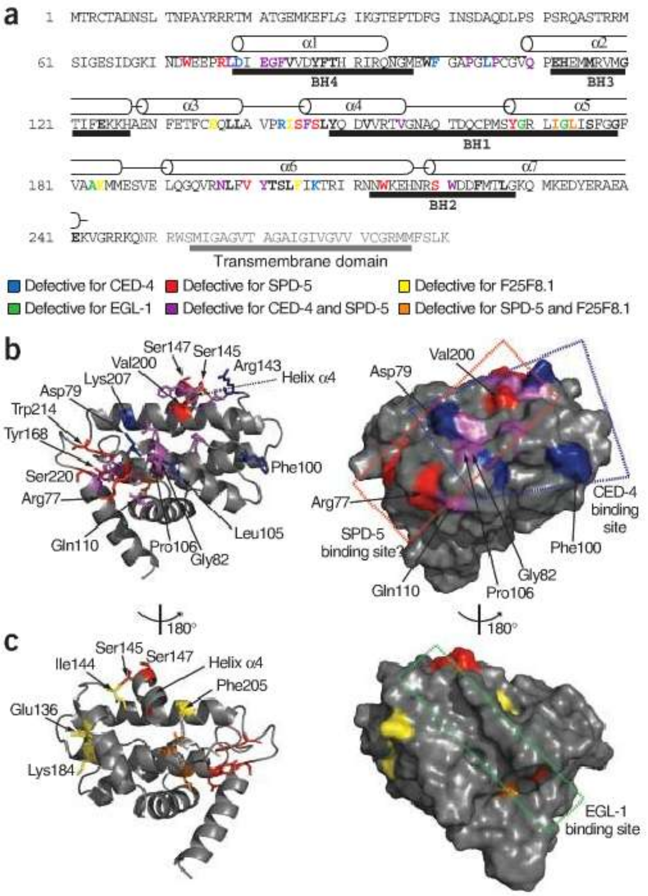

a
1 MTRCTADNSL TNPAYRRRTM ATGEMKEFLG IKGTEPTDFG INSDAQDLPS PSRQASTRRM
α1
α2
61 SIGESIDGKI NDWEEPRLDI EGFVVDYFTH RIRQNGMEWF GAPGLPCGVQ PEHEMMRVMG
BH4
BH3
α3
α4
α5
121 TIFEKKHAEN FETFCKQLLA VPRISFSLYQ DVVRTVGNAQ TDQCPMSYGR LIGLISFGGF
BH1
α6
α7
181 VAAFMMESVE LQGQVRNLFV YTSLFIKTRI RNNWKEHNRS WDDFMTLGKQ MKEDYERAEA
BH2
241 EKVGRRKQNR RWSMIGAGVT AGAIGIVGVV VCGRMMFSLK
Transmembrane domain
🟦 Defective for CED-4
🟥 Defective for SPD-5
🟨 Defective for F25F8.1
🟩 Defective for EGL-1
🟪 Defective for CED-4 and SPD-5
🟧 Defective for SPD-5 and F25F8.1
b
Ser147
Ser145
Val200
Arg143
Helix α4
Lys207
Asp79
Trp214
Tyr168
Ser220
Arg77
Phe100
Leu105
Gln110
Pro106
Gly82
Val200
Asp79
Arg77
SPD-5 binding site
CED-4 binding site
Phe100
Gly82
Gln110
Pro106
180°
180°
c
Ser145
Ser147
Ile144
Helix α4
Phe205
Glu136
Lys184
EGL-1 binding site

Figure 5 | Positioning edgetic residues in CED-9 structures. (**a**) Positions of edgetic residues in the CED-9 sequence. The portion of CED-9 present in the crystal (PDB: 1OHU)[19] and the α-helices observed in the corresponding structure are indicated above the sequence; BCL2 homology (BH) domains[19] are indicated under the sequence. Edgetic and non-edgetic residues are in bold font, edgetic residues are colored as indicated. (**b**) Ribbon diagram of the CED-9 structure (PDB: 2A5Y)[16] (left); residues mutated in edgetic alleles defective for CED-4 and/or SPD-5 interaction are shown as sticks. Helix α4 (the region undergoing EGL-1–induced conformational changes) is indicated[18]. Van der Waals surface of the same structure in identical orientation (right). The CED-4–binding site and the hypothetical SPD-5–binding site are shown. (**c**) Ribbon diagram of the CED-9 structure (PDB: 2A5Y)[16] (left) at opposite orientation with respect to **b**. Residues mutated in edgetic alleles defective for SPD-5 and/or F25F8.1 interactions are shown as sticks. Van der Waals surface of the same structure in identical orientation (right). The EGL-1 binding site is shown.

significantly more than expected by chance ($P = 1.2 \times 10^{-4}$; hypergeometric test). Half of the 14 residues mutated in the proteins encoded by edgetic alleles defective for CED-4 and one additional partner were at the CED-4 binding site, also more than expected by chance ($P = 0.022$; hypergeometric test). In contrast, only one of the 24 non-edgetic residues was in contact with CED-4, a significantly unlikely occurrence ($P = 9.5 \times 10^{-3}$; hypergeometric test).

We also compared the average distance to CED-4 of the residues in each set to the average distance of random sets of residues (**Fig. 4d**). Whereas 24 CED-9 residues picked at random had one chance in four to be further away from CED-4 than the non-edgetic residues ($P = 0.27$; empirical P-value), the edgetic residues were significantly closer to CED-4 than expected by chance ($P = 1.3 \times 10^{-3}$ and 1.6×10^{-5} for alleles defective for two and one interaction, respectively; empirical P-values). These results argue that mutations of edgetic residues likely result in the alteration of the CED-9/CED-4 interface, whereas mutations of non-edgetic residues likely disrupt the CED-9/CED-4 interaction by altering CED-9 structure.

As there is no obvious clustering of edgetic residues for any CED-9 interactor on the CED-9 primary sequence (**Fig. 5a**), suggesting that the binding sites for SPD-5, F25F8.1 and CED-4 are conformational, we used sets of edgetic residues to map the putative binding sites for SPD-5 and F25F8.1 (**Fig. 5b**,**c**, **Supplementary Figs. 9**,**10** and **Supplementary Data 4**). Our edgetic strategy enabled the isolation of partner-specific edgetic alleles for each CED-9 partner even though the CED-9 interaction surfaces seem intricate, with partly overlapping sites.

Node removal and edgetic perturbation *in vivo*

RNAi of *ced-9* (*ced-9*(RNAi)) in worms results in apoptosis-triggered embryonic lethality[26] because of increased germ-cell death (**Fig. 6a**). In addition to previously described defects in mitotic spindle assembly resulting in embryonic lethality[23], RNAi of *spd-5* also increased apoptosis in the germline, approximately half as much as *ced-9*(RNAi) (**Fig. 6a**). This is consistent with our identification of SPD-5 as a biophysical interactor of CED-9 and suggests that SPD-5, in addition to its role in mitosis, is also involved in apoptosis regulation. As with *ced-9*(RNAi), *spd-5*(RNAi)–induced germ-cell apoptosis was suppressed in a *ced-3* null background.

To evaluate whether the CED-9/SPD-5 interaction directly contributes to the embryonic lethality and germ-cell death observed upon *spd-5*(RNAi), we characterized worms carrying CED-9(W214R), a SPD-5–specific edgetic allele. In parallel, we analyzed worms carrying CED-9(K207E), a CED-4–specific edgetic allele, to evaluate the phenotypic consequences of perturbing the CED-9/CED-4 interaction.

We generated transgenic lines carrying genes encoding CED-9(K207E) and CED-9(W214R) by microparticle bombardment and crossed them into worms carrying the *ced-9*(*n2161*) null allele. Compared to worms rescued with a wild-type *ced-9* transgene, worms expressing CED-9(K207E) laid fewer embryos ($P = 0.04$; Student t-test), similar to *ced-9* null mutants (**Fig. 6b**). However wild-type *ced-9* transgene and both edgetic alleles rescued the embryonic lethality conferred by the *ced-9* null allele (**Fig. 6c**). Even though we cannot exclude that CED-9(K207E) could retain some residual capacity to bind CED-4 *in vivo*, the rescue observed with this transgene suggests that the antiapoptotic action of CED-9 during embryonic development is not exclusively correlated to CED-4 sequestration. These data also show that the embryonic lethality that occurs in worms subjected to *spd-5*(RNAi) is not necessarily due to loss of the CED-9/SPD-5 interaction as worms expressing CED-9(W214R) are viable.

The viability of transgenic worms expressing CED-9(K207E) or CED-9(W214R) allowed investigation of the role of CED-9/CED-4 and CED-9/SPD-5 interactions in germline apoptosis. We subjected worms to apoptotic challenges induced by *ced-4* (RNAi) and *cpb-3*(RNAi), which suppress and mildly increase germ-cell apoptosis, respectively[27,28]. Without an apoptotic challenge (*gfp*(RNAi)), worms expressing CED-9(K207E) and

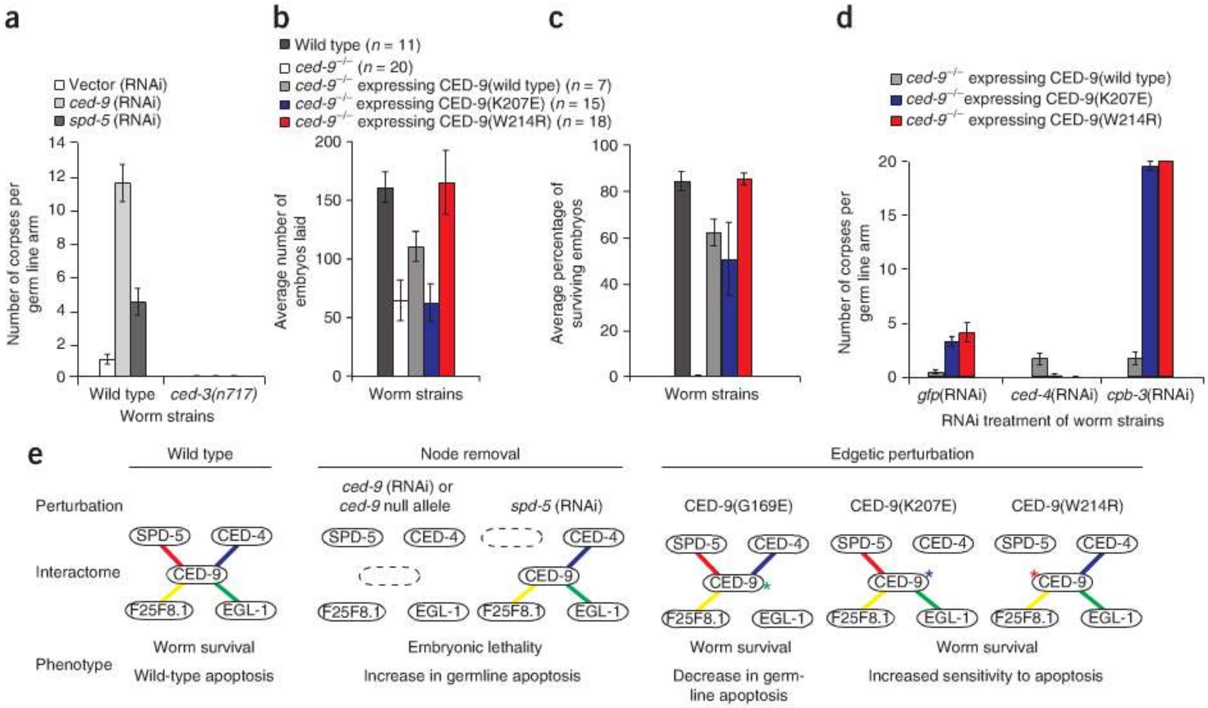


Figure 6 | Node removal and edgetic perturbation *in vivo*. (**a**) Corpse count per germline arm in wild-type (N2) or apoptosis-defective *ced-3(n717)* worms subjected to RNAi. Vector(RNAi) is the negative control. Errors bars, s.e.m. (n = 15). (**b**,**c**) Average number of embryos laid (**b**; error bars, s.e.m.) and fraction of worm broods reaching adulthood (**c**; error bars, s.e. for a binomial distribution) in the indicated strains. (**d**) Corpse count per germline arm in the indicated strains treated with RNAi targeting *gfp* (negative control), *ced-4* or *cpb-3*. Error bars, s.e.m. (n = 12, except for *gfp*(RNAi) and *cpb-3*(RNAi) of *ced-9*$^{-/-}$ expressing CED-9(wild-type) where n = 9 and 7, respectively). (**e**) Schematic of phenotypic consequences of selected network perturbations. Node removal is induced either by *ced-9* or *spd-5*(RNAi), while edgetic perturbation is caused by the CED-9(G169E)[12,17], CED-9(W214R) or CED-9(K207E) mutations. Wild-type, (*unc-69(e587)*) worm strain. *ced-9*$^{-/-}$: worm strain carrying a *ced-9* null allele (*ced-9(n1950n2161)*). CED-9 wild-type, CED-9(K207E) and CED-9(W214R) are expressed from constructs integrated into the *ced-9* null allele worm strain genetic background.

CED-9(W214R) exhibited a small increase in germline apoptosis compared to worms rescued with a wild-type *ced-9* allele (3.25, 4.08 and 0.44 dead cells, respectively, $P = 8.2 \times 10^{-5}$ and 0.05; Student *t*-test) (**Fig. 6d**), but germ-cell apoptosis in these worms was much less pronounced relative to *ced-9*(RNAi)–induced apoptosis (**Fig. 6a**). As anticipated, *ced-4*(RNAi) of CED-9(K207E) and CED-9(W214R) mutants suppressed apoptosis in the germline. RNAi treatment of *cpb-3* strongly increased germ-cell apoptosis in worms expressing CED-9(K207E) and CED-9(W214R), compared to the small increase observed in worms rescued with a wild-type *ced-9* allele (19.5, 20 and 1.72 dead cells, respectively, $P = 5.2 \times 10^{-12}$ and 5.8×10^{-8}; Student *t*-test). This finding suggests that, similarly to CED-9/CED-4, the CED-9/SPD-5 interaction also protects germ cells from apoptosis.

Node and edge removal can result in diverse phenotypic profiles, uncovering different aspects of the apoptosis module (**Fig. 6e**). The RNAi experiments we presented implicate *ced-9* and *spd-5* in *ced-3*–mediated germline apoptosis and reveal genetic links between these actors. As both *ced-9* and *spd-5* are essential genes, knockdowns lead to embryonic lethality, precluding characterization. Partner-specific edgetic alleles, which restore viability of mutant worms, underscore that in contrast to the CED-9/EGL-1 interaction, both CED-9/CED-4 and CED-9/SPD-5 protein-protein interactions contribute to negative control of germline apoptosis, especially in response to particular apoptotic triggers.

DISCUSSION

We present a strategy to select edgetic alleles defective for one or a few protein interactions, as a way to better understand their role in complex interaction networks. Applying this strategy to CED-9, we identified edgetic alleles whose gene products (i) lacked only a subset of interactions, (ii) had interaction defects that were likely due to specific changes in or close to protein interaction sites and (iii) had *in vivo* phenotypes different from those caused by null or near-null perturbations. An edgetic mutation that only affected the interaction between CED-9 and SPD-5 resulted in increased sensitivity to apoptotic stimuli. In contrast, the null phenotype for these genes was embryonic lethality. Hence, our platform is one alternative to define functions for essential genes beyond their null phenotype.

Though the biological function of the CED-9/SPD-5 interaction was not fully defined, there are two likely possibilities. The CED-9/SPD-5 interaction may be required to suppress apoptosis during spindle assembly, during centrosome assembly or at other times during cell division when SPD-5 is present. Alternatively SPD-5 may 'moonlight'[29] in the apoptotic pathway, given that loss of the CED-9/SPD-5 interaction sensitizes cells to apoptosis caused by loss of CPB-3, an RNA-binding protein with no known function in spindle assembly or cell division.

Biological systems consist of interaction networks in which many types of macromolecules associate with and act on each other, and biological properties of living organisms reflect the local and global properties of these networks. Several well-characterized inherited human disease alleles associated with particular disease phenotypes have been shown to correspond to edgetic perturbations[30]. Among the many biophysical interactions identified so far, a critical step is to identify the biologically relevant ones[31] and understand how they contribute to cellular systems. To address such questions, tools that probe interactions (edges) are needed. Considering that tens of thousands of interactions have been mapped for an increasing number of organisms, such edgetic perturbation strategies must be compatible with high-throughput settings. Our platform is a reverse genetics strategy to interrogate protein-protein interactions in the context of interactome networks. We propose the systematic use of 'edgetic perturbation' reagents, whether as alleles or small compounds, to analyze the properties of interaction networks.

METHODS

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

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

ACKNOWLEDGMENTS

This paper is dedicated to the memory of Stan Korsmeyer. We thank the members of the Vidal Lab and of the Dana-Farber Cancer Institute CCSB and particularly A.-R. Carvunis for helpful discussions. This work was supported by US National Institutes of Health (NIH) grants R01 HG001715 from the National Human Genomics Research Institute (NHGRI) and National Institute of General Medical Sciences (NIGMS) and grants R33 CA105405, R33 CA132073 and R21/R33 CA081658 from the US National Cancer Institute (NCI) (M.V.), U01 CA105423 from the NCI (principal investigator, S. Orkin; project leader, M.V.) and by Institute Sponsored Research funds from the Dana-Farber Cancer Institute Strategic Initiative awarded to CCSB. M.D. and G.L. were supported by a 'Research Fellow' fellowship from the Fonds de la Recherche Scientifique (FRS-FNRS, French Community of Belgium). B.C. was supported by the Belgian Program on Interuniversity Attraction Poles initiated by the Federal Office for Scientific, Technical and Cultural Affairs (IAP P6/19 PROFUSA). S.M. was supported by an NIH National Research Service Award training grant fellowship (T32CA09361). P.-O.V. was supported by a European Molecular Biology Organization long-term fellowship 61-2002. Support was provided by the Leukemia Research Foundation to M.B. M.V. and R.B. thank the 'Fonds de la Recherche Scientifique (FRS-FNRS, French Community of Belgium)'. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NCI, NHGRI, NIGMS or the NIH.

AUTHOR CONTRIBUTIONS

M.D., B.C., S.M., P.-O.V., M.A.Y., Q.Z., R.B., J.V., M.B. and M.V. conceived the experiments and analyses. S.M. and P.-O.V. generated the *ced-9* mutant library and performed Y2H screens and R-Y2H selections. M.D. and P.-O.V. cloned the alleles. P.-O.V. and M.D. performed the co-APs. M.D. and G.L. developed and implemented the modified Y2H assay. B.C. performed the structural analyses. B.C. and M.A.Y. performed the statistical analyses. N.S. generated the transgenic worms under the supervision of S.M. and M.B. S.M. performed survival and apoptosis challenge experiments. M.D. and V.R. performed the mutant alleles stability experiment. M.D., B.C., S.M., P.-O.V., M.E.C. and M.V. wrote the manuscript. All authors discussed the results. D.E.H. and M.V. conceived and co-directed the project.

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

1. Barabási, A.L. & Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).
2. Vidal, M. Interactome modeling. *FEBS Lett.* **579**, 1834–1838 (2005).
3. Shih, H.M. *et al.* A positive genetic selection for disrupting protein-protein interactions: identification of CREB mutations that prevent association with the coactivator CBP. *Proc. Natl. Acad. Sci. USA* **93**, 13896–13901 (1996).
4. Leanna, C.A. & Hannink, M. The reverse two-hybrid system: a genetic scheme for selection against specific protein/protein interactions. *Nucleic Acids Res.* **24**, 3341–3347 (1996).
5. Vidal, M., Braun, P., Chen, E., Boeke, J.D. & Harlow, E. Genetic characterization of a mammalian protein-protein interaction domain by using a yeast reverse two-hybrid system. *Proc. Natl. Acad. Sci. USA* **93**, 10321–10326 (1996).
6. Vidal, M., Brachmann, R.K., Fattaey, A., Harlow, E. & Boeke, J.D. Reverse two-hybrid and one-hybrid systems to detect dissociation of protein-protein and DNA-protein interactions. *Proc. Natl. Acad. Sci. USA* **93**, 10315–10320 (1996).
7. Endoh, H. *et al.* Integrated version of reverse two-hybrid system for the postproteomic era. *Methods Enzymol.* **350**, 525–545 (2002).
8. Gray, P.N., Busser, K.J. & Chappell, T.G. A novel approach for generating full-length, high coverage allele libraries for the analysis of protein interactions. *Mol. Cell. Proteomics* **6**, 514–526 (2007).
9. Kritikou, E.A. *et al. C. elegans* GLA-3 is a novel component of the MAP kinase MPK-1 signaling pathway required for germ cell survival. *Genes Dev.* **20**, 2279–2292 (2006).
10. Inouye, C., Dhillon, N., Durfee, T., Zambryski, P.C. & Thorner, J. Mutational analysis of STE5 in the yeast *Saccharomyces cerevisiae*: application of a differential interaction trap assay for examining protein-protein interactions. *Genetics* **147**, 479–492 (1997).
11. Serebriiskii, I., Khazak, V. & Golemis, E.A. A two-hybrid dual bait system to discriminate specificity of protein interactions. *J. Biol. Chem.* **274**, 17080–17087 (1999).
12. Hengartner, M.O., Ellis, R.E. & Horvitz, H.R. *Caenorhabditis elegans* gene *ced-9* protects cells from programmed cell death. *Nature* **356**, 494–499 (1992).
13. Yang, X., Chang, H.Y. & Baltimore, D. Essential role of CED-4 oligomerization in CED-3 activation and apoptosis. *Science* **281**, 1355–1357 (1998).
14. Conradt, B. & Horvitz, H.R. The *C. elegans* protein EGL-1 is required for programmed cell death and interacts with the Bcl-2-like protein CED-9. *Cell* **93**, 519–529 (1998).
15. del Peso, L., Gonzalez, V.M. & Nunez, G. *Caenorhabditis elegans* EGL-1 disrupts the interaction of CED-9 with CED-4 and promotes CED-3 activation. *J. Biol. Chem.* **273**, 33495–33500 (1998).
16. Yan, N. *et al.* Structure of the CED-4-CED-9 complex provides insights into programmed cell death in *Caenorhabditis elegans*. *Nature* **437**, 831–837 (2005).
17. Hengartner, M.O. & Horvitz, H.R. Activation of *C. elegans* cell death protein CED-9 by an amino-acid substitution in a domain conserved in Bcl-2. *Nature* **369**, 318–320 (1994).
18. Yan, N. *et al.* Structural, biochemical, and functional analyses of CED-9 recognition by the proapoptotic proteins EGL-1 and CED-4. *Mol. Cell* **15**, 999–1006 (2004).
19. Woo, J.S. *et al.* Unique structural features of a BCL-2 family protein CED-9 and biophysical characterization of CED-9/EGL-1 interactions. *Cell Death Differ.* **10**, 1310–1319 (2003).
20. Walhout, A.J. *et al.* Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287**, 116–122 (2000).
21. Li, S. *et al.* A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540–543 (2004).
22. Spector, M.S., Desnoyers, S., Hoeppner, D.J. & Hengartner, M.O. Interaction between the *C. elegans* cell-death regulators CED-9 and CED-4. *Nature* **385**, 653–656 (1997).
23. Hamill, D.R., Severson, A.F., Carter, J.C. & Bowerman, B. Centrosome maturation and mitotic spindle assembly in *C. elegans* require SPD-5, a protein with multiple coiled-coil domains. *Dev. Cell* **3**, 673–684 (2002).
24. Boeke, J.D., LaCroute, F. & Fink, G.R. A positive selection for mutants lacking orotidine-5'-phosphate decarboxylase activity in yeast: 5-fluoro-orotic acid resistance. *Mol. Gen. Genet.* **197**, 345–346 (1984).
25. Rual, J.F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
26. Lettre, G. *et al.* Genome-wide RNAi identifies p53-dependent and -independent regulators of germ cell apoptosis in *C. elegans*. *Cell Death Differ.* **11**, 1198–1203 (2004).
27. Ellis, H.M. & Horvitz, H.R. Genetic control of programmed cell death in the nematode *C. elegans*. *Cell* **44**, 817–829 (1986).
28. Boag, P.R., Nakamura, A. & Blackwell, T.K. A conserved RNA-protein complex component involved in physiological germline apoptosis regulation in *C. elegans*. *Development* **132**, 4975–4986 (2005).
29. Jeffery, C.J. Moonlighting proteins–an update. *Mol. Biosyst.* **5**, 345–350 (2009).
30. Zhong, Q. *et al.* Edgetic perturbation models of human genetic disorders. *Mol. Syst. Biol.* (in the press).
31. Venkatesan, K. *et al.* An empirical framework for binary interactome mapping. *Nat. Methods* **6**, 83–90 (2009).

Antibody production branches out

Alan Dove

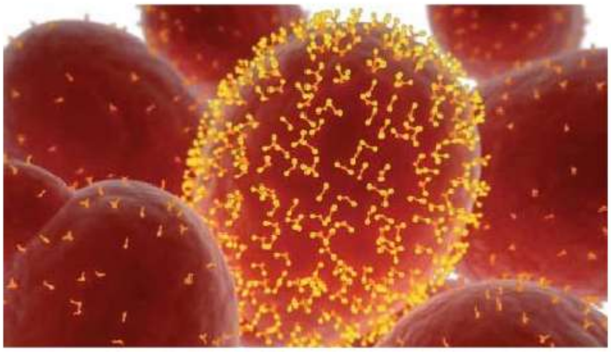
Antibodies, the molecular workhorses of protein research, have traditionally been one of the most difficult reagents to procure. Using innovative new technologies, though, a burgeoning antibody production industry is turning these molecules into commodities.

Kids today do not know how good they have it. Not long ago, generating a research-grade antibody to a protein was a major ordeal. From expressing and purifying the antigen to immunizing a suitable animal, generating hybridoma cells and purifying the final product, the process could occupy months of a graduate student's or postdoc's time.

Not anymore. "There are so many shops [that] will synthesize the peptides for you, immunize a mouse or a rabbit or a rat, whatever you want, and they'll send you the serum," says Tillman Gerngross, cofounder and CEO of Adimab. He adds that "getting polyclonal antibodies [to] something, or even monoclonals—that's a commoditized business at this point."

Commodity antibodies cover most research needs quite well, but for some targets and applications, they simply will not do. Because they come from standard laboratory mammals, for example, commercial antibodies can only target a limited set of epitopes; the injected animal recognizes conserved mammalian antigens as 'self', blunting or blocking its immune response to them. Scientists who have an eye toward clinical applications may also need better control over the antibody's structure and production than a commercial source can provide.

Fortunately, several groups have been developing novel antibody production platforms for these special cases, ranging from simple do-it-yourself expression systems to sophisticated bioreactors capable of making clinical-grade material.



An artist's rendering shows a yeast cell presenting antibodies on its surface, which bind antigen for detection and cell sorting by flow cytometry. Image courtesy of T. Gerngross.

Brewing B cells

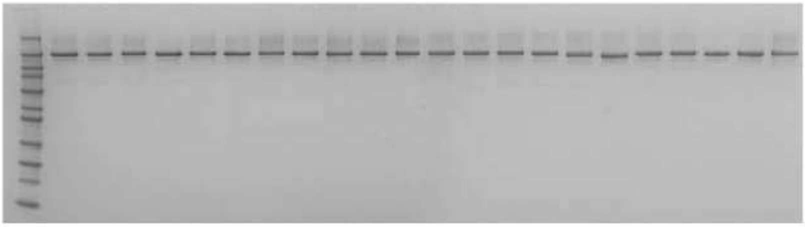
At Gerngross's company, the focus is on a very specific—and very valuable—question in clinical research. "Here's an antigen, how quickly can I get to a panel of fully humanized genes [to] that antigen? It is that metric that we sort of tried to make progress against," says Gerngross.

To do that, Adimab took their work completely outside the animal kingdom, to the yeast *Saccharomyces cerevisiae*. Gerngross and his colleagues generated a complete synthetic library of the human preimmune repertoire and engineered the yeast to make human immunoglobulin gamma (IgG) molecules representing that repertoire. Each yeast cell produces a specific IgG and presents it on its surface, like a B cell. The result is a fungal version of the human spleen, containing antibodies to every possible antigen.

With the system built, the team can now present antigens to it. "Once you have [IgG] on the surface, then you can use [flow cytometry] or magnetic bead separation to find those yeasts that make an antibody that is presented on the surface that is [to] your antigen," says Gerngross. Unlike a spleen, however, the yeast cells have not been preselected to eliminate self-reactive antibodies. As a result, the system contains a wide range of antibodies to the conserved mammalian epitopes that traditional animal-based schemes miss.

The strategy also gets around a limitation of yeast as an antibody production platform. "Different sequences will in some cases express very well in yeast and in other cases not express at all," Gerngross explains. Previous efforts that selected antibodies in mammals, then cloned the locus encoding IgG into yeast for expression, yielded mixed results. Because Adimab has built the entire preimmune repertoire into yeast, the company automatically preselects antibodies that will express well in that system.

With Gerngross's approach, generating a complete panel of antibodies to a given antigen takes about 3–4 weeks, after which the yeast can be transferred to a different medium that causes the cells to secrete their IgG molecules instead of presenting them on the surface. "You can purify [the antibody] out of a 24-well plate and that will typically yield somewhere between 50 and a few hundred micrograms," says Gerngross. Those quantities are sufficient for small-scale experiments, and though the yeast culture could be scaled up for larger batches, users can also move the selected antibody genes into other



A silver-stained gel shows the purity of antibodies produced by a yeast expression system. Image courtesy of T. Gerngross.

systems. The Adimab yeast strain itself is proprietary, but Gerngross says the antibody genes the company provides could be moved into another strain of yeast for production, or into mammalian cells.

Hunt and peck

Whereas Adimab has focused on duplicating the functions of B cells in simple fungi, other companies prefer to rely on a ready-made immune system for antibody selection. At Crystal Bioscience, for example, researchers are using chickens to generate antibodies.

According to Robert Etches, the company's president and CEO, the birds provide many of the same advantages as a yeast system: "because of the solid genetic distance between chickens and mice and humans and so on, the chicken can see antigens that other animals don't, because there is too great a similarity between a mouse protein and a human protein the mouse doesn't recognize it as nonself, whereas a chicken will."

Etches admits that it is not an entirely new idea. Indeed, chickens were one of the first model organisms used for immunological research, and investigators have long used them to generate antibodies to targets that are highly conserved in mammals. "The problem has been that there hasn't been a good way of making a chicken monoclonal antibody. That is the part of the technology that we've developed here at Crystal Bioscience. Our in-house technology is aimed at the production of monoclonal antibodies from immunized chickens," says Etches.

To do that, the team collects the spleen cells from immunized birds, then puts the cells inside small agarose capsules, isolating each B cell in its own sphere. Each sphere also contains antigen-coated beads. The beads trap antigen-reactive antibodies inside the spheres containing the B cells that produced them. Fluorescently tagged IgY, the chicken equivalent of mammalian IgG, then highlights all of the spheres whose B cells produce antibodies to that antigen.

Once they have sorted out the right cells, Etches and his colleagues isolate the genes encoding those cells' antibody variable regions. "You can then reconstruct that about any way you'd like: you can put it onto a human constant region; you can put it onto a mouse constant region; you can put it back onto a chicken constant region; you can basically make any kind of antibody that you want using those chicken [variable] sequences," he says.

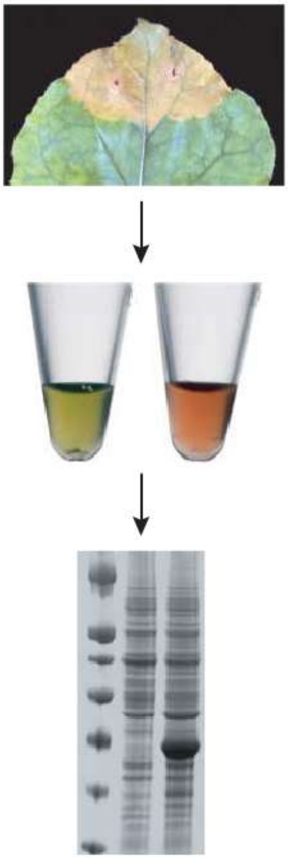
Although working with whole animals is somewhat slower than using libraries of yeast cells, Etches argues that his system has some compensatory advantages: "if you use an animal such as a chicken, you gain access to all of the affinity maturation processes that are in the humeral immune system to give you a really good antibody,... whereas in nonvertebrate systems, you don't have that."

Both the Crystal Bioscience and the Adimab platforms can be used to produce research-scale quantities of antibodies, but neither company is particularly interested in larger-scale production. Instead, both companies provide the antibody genes, which customers can then express in their system of choice. "We're not in the production business, we're in the information business. What we ultimately give you is sequence [encoding] an antibody that when you express it in [Chinese hamster ovary (CHO) cells] or wherever you wish to manufacture the antibody, [the sequence] encodes that binding event that elicits a desired therapeutic effect. How you make it is totally your problem," says Gerngross.

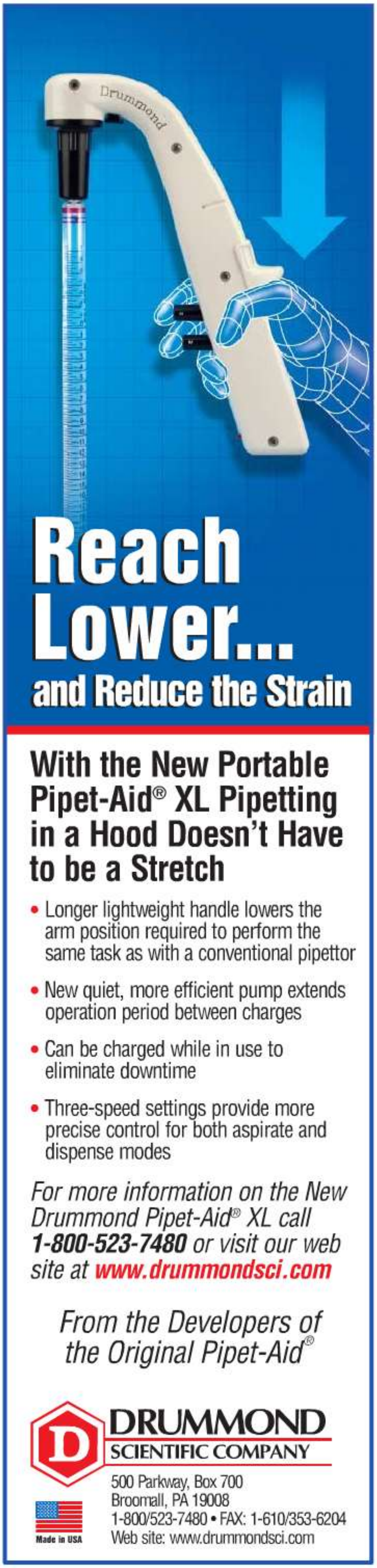
Turning over a new leaf

Researchers confronting the antibody production problem often turn to cultured mammalian cells, but they may soon find themselves shopping for expression systems in the greenhouse instead. "Plants have all the machinery necessary for not only the production of the proteins but also for the assembly of them correctly into the classic IgG [and] also the more complex [antibody classes] like secretory IgA [molecules], which also have the joining chain and the secretory component," says George Lomonossoff, professor of biological chemistry at the John Innes Centre in Norwich, UK.

Because plants express the proteins so well, researchers have long sought to use plants as an antibody expression platform. Trials in the late 1980s established that transgenic plants can do the job at least as well as mammalian cells in culture. "The problem with the transgenic approach is that it is quite slow, and you have to decide what you want to make, and then it's a whole process of transformation, regeneration, crossing and eventually getting true breeding lines [that]



Infiltrating a leaf with a plasmid containing a gene encoding a red fluorescent protein causes the leaf to express the protein at levels high enough to be seen visually in extract from the leaf and on a Coomassie-stained gel. Image courtesy of G. Lomonossoff.



have high-enough expression levels," he says.

Instead of making an entire transgenic plant, Lomonossoff and his colleagues introduce the antibody genes into the leaves of an unmodified plant. "You can actually insert the genes encoding the heavy and the light chain...on the same plasmid, and simply put the plasmids into *Agrobacterium* sp. and flood the cells, the leaf tissue, with those constructs," Lomonossoff explains, adding that "you can actually get very high levels of expression within a few days in leaf tissue."

Because the process is so quick, scientists can make multiple versions of an antibody and express them in different leaves, then select the ones with the highest affinity. For small-scale laboratory use, a few inoculated leaves can produce milligram quantities of protein, so even researchers without large plant facilities should find the technique straightforward.

The system can also be used to produce other types of complex proteins, including viral subunits that assemble into complete virion cores. One company, Medicago, is using that approach to produce an experimental H5N1 influenza vaccine, and Lomonossoff says he has received many reagent requests from academic researchers as well: "I think I sent four lots of plasmids off just today. We send a little sort of expression kit;... the idea is to get it as widely used as possible."

Relatively small bioreactors, such as these in the protein production core facility at EPFL, can produce large quantities of antibodies from traditional mammalian cell cultures. Image courtesy of F. Wurm.

Investigators who hope to commercialize an antibody may also find plants appealing, as plant expression is relatively easy to scale up, but Lomonossoff cautions against visions of field-grown pharmaceutical antibodies: "You can imagine if you grew stuff in a field, there'd be all sorts of things like bird droppings and earthworms and insects coming into your product potentially, and that concerns regulatory agencies."

Tried and true

Even with careful containment, new systems such as plants may never be practical choices for researchers who hope to see their antibodies used clinically. "The regulatory agencies are not very innovative and provocative," says Florian Wurm, professor of biotechnology and head of the laboratory of cellular biotechnology at the Ecole Polytechnique Federale de Lausanne (EPFL) in Lausanne, Switzerland. Wurm adds that "if you want today to get your product into the clinic in the fastest way, you don't experiment around with a system [that] has not been used before because you lose two years of time, minimum."

Pharmaceutical production problems may not interest academic scientists in the early phases of a project, but Wurm, who runs the core protein production facility at EPFL, advocates getting useful antibodies into industrial-grade systems as early as possible. "More frequently than I sometimes expect, I find people doing weird things, coming with very strange cell lines to us and saying 'can I make out of this a manufacturing process?'... Well, yes you can, in principle,... but I would advise you rapidly to switch to CHO" cells.

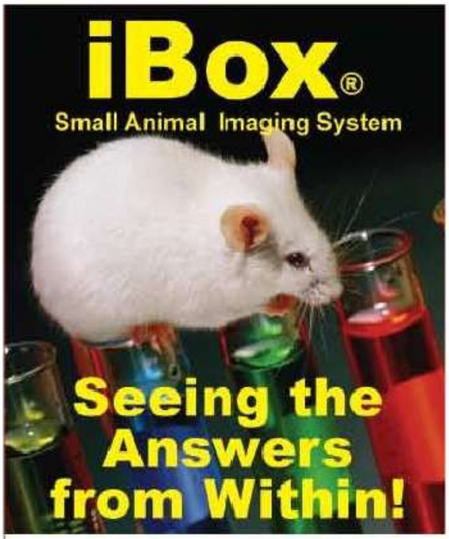
CHO cells have long been unfashionable in basic research laboratories, where the decades-old system is regarded as a creaky industrial workhorse that desperately needs to be replaced. Wurm begs to differ: "there is truly no communication between the two worlds. There's so much innovation...in this technology, the old one, but it's not sexy enough to publish in a leading journal."

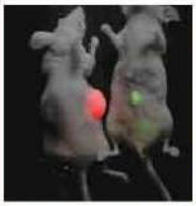
Through incremental advances, industrial researchers have progressively increased the density and longevity of a typical CHO cell culture. In the 1980s, getting a suspension culture with 2 million cells per milliliter to survive for a week was a triumph. Today, antibody production facilities routinely culture 15 million cells per milliliter for three weeks in chemically defined medium.

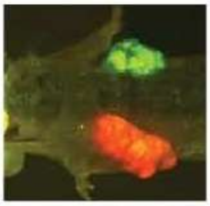
Wurm explains that a modern bioreactor filled with CHO cells and serum-free medium can yield 2 grams per liter of antibody, with very few contaminants. He adds that "in spite of the fact that [*Escherichia*] *coli* grows faster, in spite of yeast being a wonderful organism, you cannot get this purity and you cannot get the yield out of these microbial systems as we have achieved now out of mammalian cells."

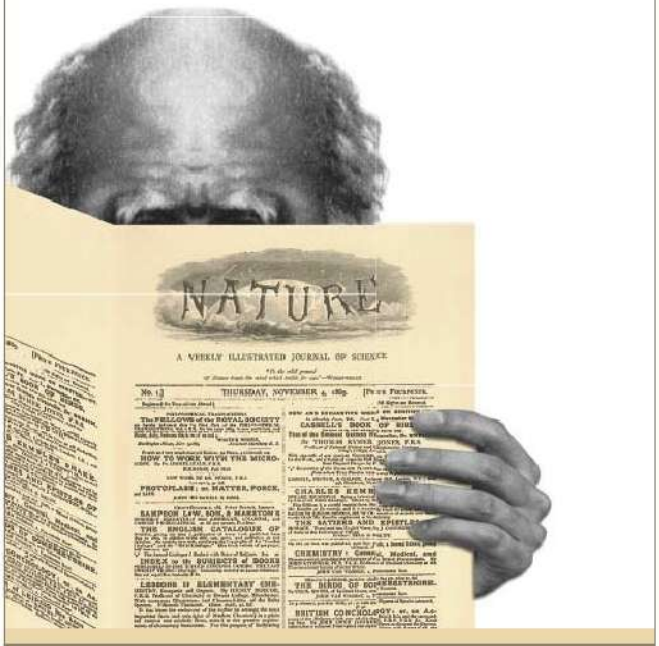
Although each system has its adherents and detractors, most agree that it is a good time to be working on antibodies. "Antibody products and similar products for therapy are growing every year between 10 and 15 percent in spite of the economic downturn, which no other industry in the world does, so I think it's a very exciting period to be in and trying to contribute to this field," says Wurm.

Alan Dove is a science writer based in Springfield, Massachusetts, USA (alan.dove@gmail.com).









SUPPLIERS GUIDE: COMPANIES OFFERING ANTIBODIES AND ANTIBODY SERVICES

Company	Web address
AbD Serotec	http://www.ab-direct.com/
Abgent	http://www.abgent.com/
Active Motif	http://www.activemotif.com/
Adimab	http://www.adimab.com/
AnaSpec	http://www.anaspec.com/
Assay Designs	http://www.assaydesigns.com/
BioGenes	http://www.biogenes.de/
Charles River	http://www.criver.com/
Chemicon (Millipore)	http://www.chemicon.com/
DIAsource Immuno Assays S.A.	http://www.diasource-diagnostics.com/
Epitomics	http://www.epitomics.com/
Fitzgerald	http://www.fitzgerald-fii.com/
GeneTel	http://www.genetel-lab.com/
GENOVAC	http://www.genovac.com/
GenWay Biotech	http://www.genwaybio.com/
GTC Biotherapeutics	http://www.gtc-bio.com/
Harlan Sera-Lab	http://www.harlanseralab.co.uk/
Innovative Research	http://www.innov-research.com/
Lonza Biologics	http://www.lonzabiologics.com/
Mabtech	http://www.mabtech.com/
Maine Biotechnology Services	http://www.mainebiotechnology.com/
MorphoSys	http://www.morphosys.com/
New England Peptide	http://www.newenglandpeptide.com/
Novus Biologicals	http://www.novusbio.com/
OriGene	http://www.origene.com/
ProSci Inc	http://www.prosci-inc.com/
Research & Diagnostic Antibodies	http://www.rdabs.com/
SDI	http://antibodies.sdix.com/
SouthernBiotech	http://www.southernbiotech.com/
Zymed (Invitrogen)	http://www.zymed.com/

EPICENTRE®
Biotechnologies

Next-generation sequencing library preparation: simultaneous fragmentation and tagging using *in vitro* transposition

The advent of next-generation sequencing has made possible genome analysis at previously unattainable depth. Roche, Illumina and Life Technologies, among others, have developed well-established platforms for deep sequencing. Regardless of the instrument, one of the bottlenecks for next-generation sequencing is the amount of time and resources required for template and library preparation. Here we describe Epicentre's Nextera™ technology (covered by issued and/or pending patents), which counters this bottleneck and simplifies the sample preparation procedure.

At present, next-generation sequencing platforms use slightly different technologies for sequencing, such as pyrosequencing, sequencing by synthesis or sequencing by ligation. However, most platforms adhere to a common library preparation procedure, with minor modifications, before a 'run' on the instrument. This procedure includes fragmenting the DNA (sonication, nebulization or shearing), followed by DNA repair and end polishing (blunt end or A overhang) and, finally, platform-specific adaptor ligation. This process typically results in considerable sample loss with limited throughput. To streamline the workflow, increase throughput and reduce sample loss, Epicentre has developed Nextera™ technology, a transposon-based method for preparing fragmented and tagged DNA libraries in as little as 4 hours. This flexible, scalable and efficient technique can be used to generate libraries for multiple sequencing platforms.

Method overview

Nextera technology uses *in vitro* transposition to prepare sequencer-ready libraries (**Fig. 1**). In a classic transposition reaction, transposases catalyze the random insertion of excised transposons into DNA targets with high efficiency. During cut-and-paste transposition, a transposase makes random, staggered double-stranded breaks in the target DNA and covalently attaches the 3′ end of the transferred transposon strand to the 5′ end of the target DNA. The transposase and transposon complex, also referred to as a Transposome™ complex, inserts an arbitrary DNA sequence at the point of insertion. We have discovered that the entire complex is not necessary for insertion, and free transposon ends are sufficient for integration.

Fraz Syed, Haiying Grunenwald & Nicholas Caruccio

Epicentre Biotechnologies, Madison, Wisconsin, USA. Correspondence should be addressed to F.S. (fraz.syed@epibio.com).

When free transposon ends are used in the reaction, the target DNA is fragmented and the transferred strand of the transposon end oligonucleotide is covalently attached to the 5′ end of the target fragment (**Fig. 1a**). The size distribution of the fragments can be controlled

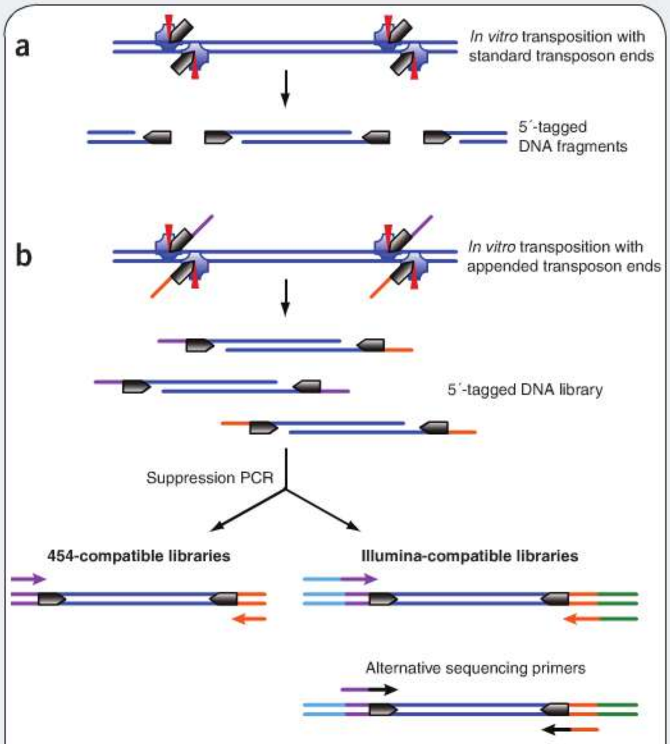

a
In vitro transposition with standard transposon ends
5′-tagged DNA fragments
b
In vitro transposition with appended transposon ends
5′-tagged DNA library
Suppression PCR
454-compatible libraries
Illumina-compatible libraries
Alternative sequencing primers

Figure 1 | Overview of Nextera fragmentation and tagging technology. (**a**,**b**) Using *in vitro* transposition, the starting DNA template is randomly fragmented and tagged at sites indicated by red arrows, using standard transposon ends (**a**) or transposon ends appended to unique adaptor sequences (colored bars) (**b**). After suppression PCR, the library can be amplified and sequenced using the appropriate platform-specific primers.

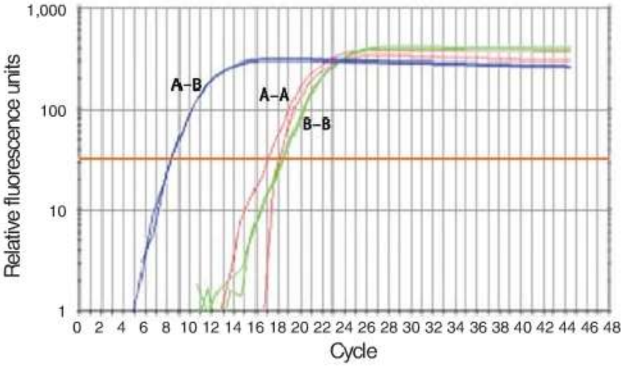


Figure 2 | Enrichment of A-B–tagged DNA fragments. Transposon ends were modified to contain two unique Roche/454–compatible tags (A and B), and a library was prepared following the scheme in **Figure 1b**. qPCR was performed using A-A, B-B and A-B primer pairs for 45 cycles. The horizontal orange line is the fluorescence threshold.

by changing the amounts of transposase and transposon ends (data not shown). Exploiting transposon ends with appended sequences results in DNA libraries that can be used in high-throughput sequencing (**Fig. 1b**).

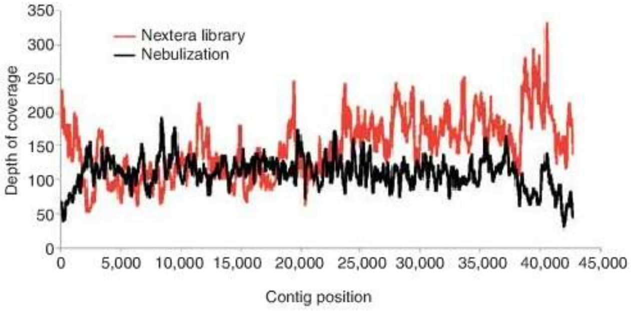
Creating tagged libraries

Nextera technology can be used to create single- and dual-tagged libraries. To generate platform-specific libraries, complementary tags can be added to the 5′ and 3′ ends of the fragmented DNA. Once the DNA is labeled at the 5′ end, a complementary tag can be added to the 3′ end by extension with a strand-displacing polymerase. PCR amplification with a single primer confirms tagging at both ends of the DNA fragments and faithful reproduction of input DNA sizes (data not shown). Thus, single-primer PCR amplifies the genomic library and produces double-stranded DNA fragments with complementary adaptors.

In addition to the complementary tags, two independent tags can be added to the fragmented DNA by appending to the transposon end sequence an engineered adaptor sequence. After extension, the sequencing adaptors enable amplification by emulsion PCR, bridge PCR and other methods. The amplified library can be subsequently sequenced.

Enrichment of dual-tagged fragment libraries

We used T7 bacteriophage genomic DNA to demonstrate dual tagging and enrichment of fragments containing both tags (**Fig. 2**). We modified transposon ends to contain Roche/454-compatible tags (A and B). After the tagging and fragmenting reaction, we heat-inactivated the transposase and performed limited-cycle PCR (suppression PCR). We analyzed a 1:100 diluted sample by quantitative PCR (qPCR). A mixed population of tagged DNA fragments was obtained, containing the desired A-B–tagged DNA, as well as DNA fragments with either A or B tags at both ends. However, qPCR showed enrichment of the A-B–tagged fragments in the amplified sample.



	Nebulization	Nextera library
Assembled contigs:	1 (42,679 bp)	1 (42,680 bp)
Percentage error:	0.48%	0.41%
Proportion Q40-plus bases:	0.999	0.998
Mapped reads:	99.48%	96.19%
Average coverage:	110	125

Figure 3 | Comparison of nebulization and Nextera-generated libraries. The Nextera library, enriched in DNA fragments tagged with Roche/454 A and B adaptors, was sequenced on a GS FLX instrument (Roche). A control library prepared from the same starting DNA, using nebulization, was also sequenced. Depth of coverage across the contig (top) and a summary of data (bottom) are shown for each library. Q40 is a quality score denoting the probability of a wrong base call at 1 in 10,000.

Nebulization versus *in vitro* transposition

We performed deep sequencing of the Nextera-enriched dual-tagged library and of a control library prepared by nebulization and the manufacturer's recommended protocol. Contig assembly, coverage and accuracy of the Nextera library data were comparable to those for the control library produced using nebulization (**Fig. 3**).

Conclusions

The current library preparation methods for next-generation sequencing are time-consuming and prone to considerable sample loss. Even before library preparation, the recovered DNA must be purified and end-polished. Epicentre's Nextera technology offers many advantages over current library preparation methods, such as a streamlined workflow that can result in substantial savings in time and cost. The method is scalable and requires less starting DNA than current procedures. As described here, Nextera technology adapts *in vitro* transposition, a powerful technique that can simultaneously fragment and tag genomic DNA, by using optimized transposases and incorporating engineered free transposon ends. By additional manipulation, libraries containing complementary or independent adaptor sequences can also be efficiently constructed and amplified before sequencing on most next-generation sequencing platforms.

ACKNOWLEDGMENTS

Sequencing data were generated in collaboration with Hilary Morrison, Jessica Welch, Ekaterina Andreischeva and William Reznikoff, Marine Biological Laboratory, Woods Hole, Massachusetts, USA.

This article was submitted to *Nature Methods* by a commercial organization and has not been peer reviewed. *Nature Methods* takes no responsibility for the accuracy or otherwise of the information provided.

sartorius stedim biotech

SENSOLUX® stand-alone version: noninvasive determination of pH and DO in shake flasks

The determination of growth kinetic parameters during early process development is of crucial importance. The more information output is obtained, the more optimization strategies can be applied to generate highly productive cell lines. The new SENSOLUX® technology enables an optical and noninvasive measurement of pH and dissolved oxygen saturation in shake flasks. This technology provides a simple detection method suitable for effective final clone screening and medium optimization.

Sartorius Stedim Biotech now introduces the SENSOLUX® technology, which enables an optical and noninvasive measurement of the pH value and the dissolved oxygen saturation (DO) during the cultivation of animal and human cells. The first member of this product line is the SENSOLUX® stand-alone, an intelligent shaker tray with an integrated sensor system. Used in combination with the new single-use SENSOLUX® Erlenmeyer Flasks (EF), it facilitates the easy, safe and highly informative online measurement of these crucial process parameters in incubation shakers. SENSOLUX® EF are equipped with two precalibrated sensor patches that are sensitive to pH and DO, respectively. The sensor system integrated into the tray monitors both parameters optically and noninvasively from outside the SENSOLUX® EF.

The SENSOLUX® technology is based on the principle of fluorescence. The sensor patches contain fluorescent dyes that can be excited with light of a given wavelength. The SENSOLUX® shaker tray contains nine independent optical sensors. Optical fibers integrated into the shaker tray transmit light of a particular wavelength to the sensor patches; at the same time, they also transmit the luminescence response from the patches to a measuring amplifier. The characteristics and intensity of the light emitted is influenced by changes in the concentration of the parameters pH or DO, respectively (**Fig. 1**).

We evaluated the SENSOLUX® technology under different conditions. Here we compare the accuracy of the pH and DO patches to that of standard pH and DO electrodes. We also analyze the cultivation of Chinese hamster ovary (CHO) cells by using the SENSOLUX® EF.

SENSOLUX® versus standard technology

We evaluated the SENSOLUX® pH patch technology in comparison to the standard electrochemical pH electrode by implementing the patch into a BIOSTAT® A plus 1-liter glass vessel. A pH profile (6.05–8.60) was induced by a controlled titration (0.4 M NaOH, 0.4 M HCl) in a phosphate-based buffer system (0.05 M per liter, pH 8.6, 0.15 M ion strength with NaCl). The temperature was controlled via the control unit at 37 °C, and data were detected with an interval of 6 s.

We achieved a maximum deviation smaller than ±0.1 pH units with a

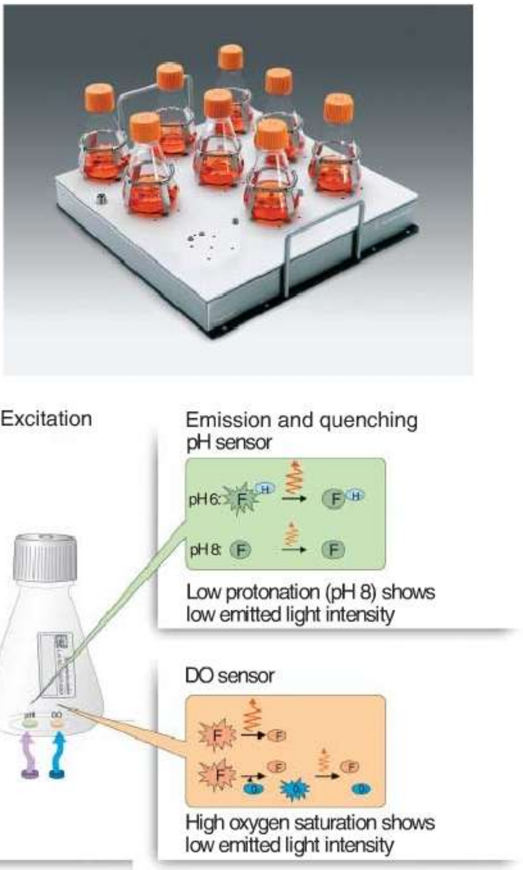

Excitation
Emission and quenching
pH sensor
pH 6:
F
H
pH 8:
F
Low protonation (pH 8) shows low emitted light intensity
pH
DO
DO sensor
F
High oxygen saturation shows low emitted light intensity

Figure 1 | SENSOLUX® stand-alone version: product picture and schematic.

Kathrin Schmale

Sartorius Stedim Biotech GmbH, Goettingen, Germany. Correspondence should be addressed to K.S. (kathrin.schmale@sartorius-stedim.com).

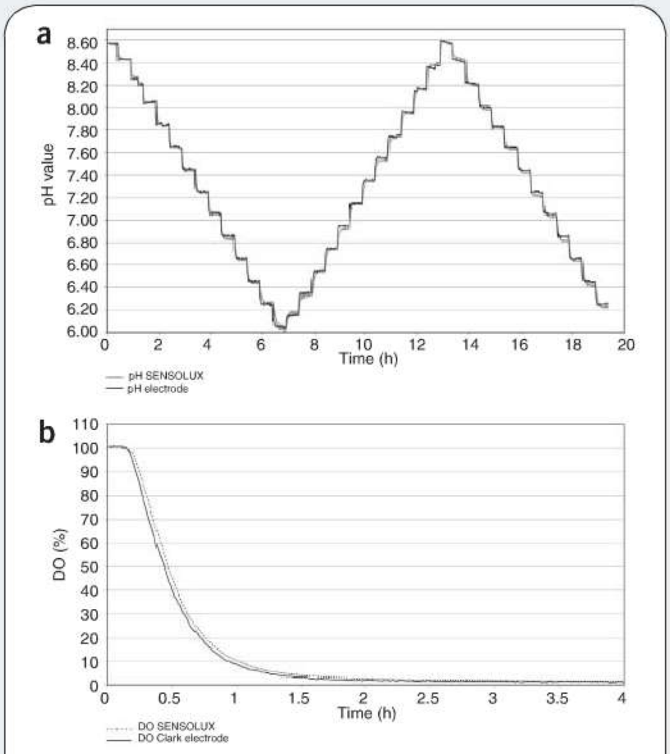


Figure 2 | SENSOLUX® patch technology versus standard electrochemical pH and DO electrodes. (**a**) A pH profile (6.05–8.6) was induced by titration (detection interval, 6 s). (**b**) Under a controlled gas atmosphere (50% DO air saturation), the DO was detected every 15 s.

detection interval of 6 s during a 19-h experiment (**Fig. 2a**). This signifies that 11,600 measurements can be carried out with this low deviation range. For real-time cultivation, this means, for example, a cultivation time of 16 d with a detection interval of 2 min. During our studies, the pH showed a drift of 0.01 units per day.

We also evaluated the SENSOLUX® DO patch technology in comparison to the standard Clark electrode by using a 250-ml Erlenmeyer flask equipped with both technologies. We used the same phosphate buffer system as described above and ensured a controlled environment by carrying out the experiment in the CERTOMAT® IS incubation shaker (37 °C, 40 r.p.m., orbit 50 mm). Using a sparger implemented in the E-Flask, nitrogen was added with a gassing rate of 0.5 liter per min. The DO drift of both technologies was monitored over 35 h at an interval of 15 s. The initial DO was 100% air saturation. To achieve a constant gas environment in the flasks, nitrogen and air were added (both 0.5 liter per min) during the complete test. After 10 h, a constant gas condition could be achieved in the system, and the drift study commenced.

The comparison between the DO patch technology and a standard Clark electrode showed similar curve progressions over the course of the experiment (**Fig. 2b**). We detected a deviation of 0.2% at a lower (1%) air saturation and 1.0% at a higher (100%) air saturation. Up to 8,600 light spots, the SENSOLUX® DO patches had an average drift of 0.08% per hour (data not shown; detection interval, 15 s). For a real-time cultivation, this implies a drift of 0.02% per day at 50% air saturation with a detection interval of 5 min.

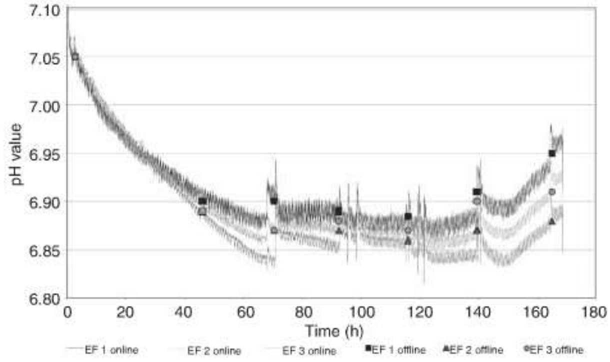


Figure 3 | Growth of CHO cells in single-use SENSOLUX® EF. CHO cells were batch-cultivated in three single-use SENSOLUX® EF 250 ml (EF1, EF2 and EF3) at 37 °C, 200 r.p.m. (25 mm orbit), 5% CO_2 and 85% humidity. The graph shows the pH value during the cultivation time as determined by SENSOLUX measurements (online) and standard measurement (offline).

Cultivation CHO cells in single-use SENSOLUX® EF

CHO cells were grown in SENSOLUX® EF 250 ml (proCHO5 medium, 4 mM L-glutamine, 1× HT) at 37 °C in an incubation shaker (200 r.p.m., 25 mm orbit) with a CO_2 saturation of 5% and a humidity of 85%. The working volume was 50 ml; starting cell density was 5×10^5 cells per ml. Samples for subsequent analyses were collected once a day. Beside the online, fluorescence-based measurement of pH and DO, samples were also collected for external pH measurement by using a standard pH electrode and amplifier. The sample vials were prewarmed (37 °C) and were closed immediately after sampling to avoid temperature and CO_2 shifts.

In the course of the cultivation in the SENSOLUX® EF, CHO cells achieved a maximum viable cell density of 5×10^6 cells per ml (data not shown). The cells were not oxygen limited during the total cultivation time (DO = 90–100%). A maximum deviation 0.04 units was detected when comparing the external measured pH with the online fluorescence-based detection method, which was in the described deviation range of 0.1 pH units (**Fig. 3**).

Conclusion

The SENSOLUX® patch technology produced similar results to those of the standard pH and DO electrodes, indicating that this technology is a useful alternative to the standard electrochemical pH and DO measurement. Both accuracy and drift of the new technology are comparable to the commonly used electrodes. Thus, the SENSOLUX® stand-alone version is a suitable tool for providing conclusive results in the early process-development phase—for example, advanced clone screening and medium optimization.

ACKNOWLEDGMENTS
I thank D. Riechers, J. Lueders, M. Hantke and Y. Stark of Sartorius Stedim Biotech GmbH, Goettingen, Germany, for their support and for providing their data.

This article was submitted to *Nature Methods* by a commercial organization and has not been peer reviewed. *Nature Methods* takes no responsibility for the accuracy or otherwise of the information provided.

hMSC differentiation marker detection using Thermo Scientific Solaris™ qPCR Gene Expression Assays

Human mesenchymal stem cells have become an important resource in developing strategies for regenerative therapies, owing to their ease of use and differentiation potential. Analytical tools, such as whole-genome expression array and validation with Solaris™ qPCR Assays, are essential to fully understand the key molecular events, such as microRNA-mediated gene modulation, that mark stem cell differentiation.

Although microarrays are useful for rapid whole-genome profiling, a complementary method with improved sample throughput, sensitivity and dynamic range is needed for follow-up studies. Quantitative real-time PCR (qPCR) is often the method of choice to validate gene expression results from whole-genome microarrays. Solaris™ qPCR Gene Expression Assays are predesigned on a genome-wide scale using a novel, tier-based algorithm to detect all variants of a given gene and distinguish among closely related family members. Solaris assays incorporate minor groove binder (MGB™)[1] and Superbase™ technologies (Epoch Biosciences, Inc) for increased sequence design space and enhanced specificity. Combining these two chemical strategies with a fluorescent (FAM) reporter dye and corresponding Dark Quencher™ fluorochrome (Epoch Biosciences, Inc) results in highly specific and sensitive assays that consistently function under universal thermocycling conditions. Here we describe an application of Solaris technology to validate the microarray expression data from early-stage osteogenic human mesenchymal stem cells (hMSCs).

microRNAs (miRNAs) are involved in many aspects of cellular processes; however, little is known about their role in the regulation of adult stem cell differentiation. In a recently published screen using a Thermo Scientific Dharmacon miRIDIAN microRNA Inhibitor and Mimic library, miR-148b was shown to increase alkaline phosphatase (ALPL) activity, an early marker of osteoblast differentiation[2]. Here we show how the novel Solaris platform was used to further characterize gene expression in hMSCs treated with differentiation medium or with miRNA mimics.

We assessed osteogenic differentiation in hMSCs treated with medium or with miRNA mimics (**Fig. 1**). For the medium treatments, hMSCs were grown in osteoblast differentiation medium or propagation medium. For the mimic treatment, we transfected miRIDIAN miR-148b mimic or miRNA mimic negative control 1 into hMSCs. Six days after induction of osteogenic differentiation, we collected the cells and assessed the culture for ALPL-positive cells[2]. Using the Thermo Scientific Cellomics VTi ArrayScan high-content imaging system, we observed an approximately eightfold increase in ALPL-positive cells treated with either differentiation medium or miRNA mimics relative to controls (data not shown).

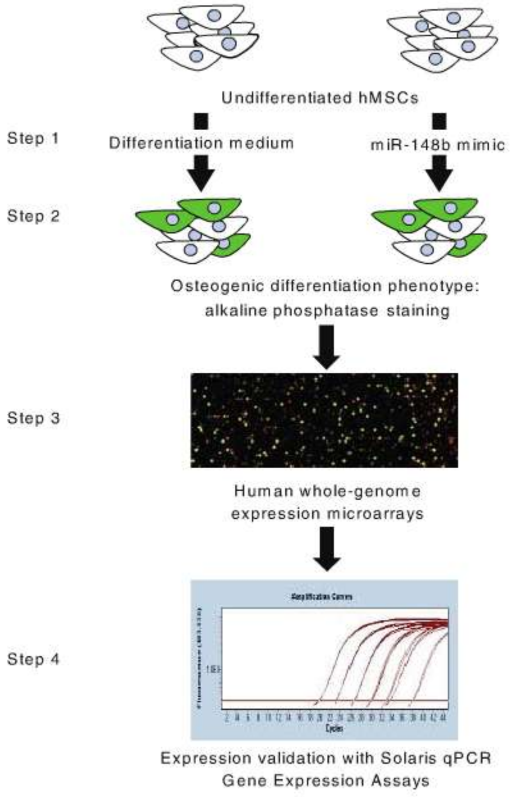


Figure 1 | Experimental workflow for characterization of gene expression changes in human mesenchymal stem cell (hMSC) osteogenic differentiation. Step 1: hMSCs (Lonza) are treated with differentiation medium or miRIDIAN miR-148b mimic for 6 d[2]. Step 2: cells are assessed for the ALPL early osteogenic marker. Step 3: RNA is isolated and microarray expression analysis performed to identify genes that are differentially regulated with each treatment. Step 4: Solaris qPCR Gene Expression Assays are used to validate differentially expressed gene targets identified from the microarray data.

Zaklina Strezoska, Yuriy Fedorov & Melissa L Kelley

Thermo Fisher Scientific, Lafayette, Colorado, USA. Correspondence should be addressed to M.L.K. (melissa.kelley@thermofisher.com).

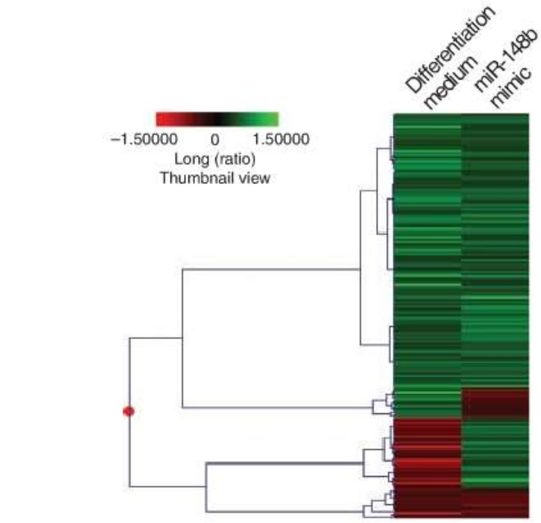


Figure 2 | Microarray analysis identifies genes that are differentially regulated in hMSC osteogenic differentiation by differentiation medium treatment and miRNA treatment. Isolated total RNA from treated hMSCs was hybridized against RNA from undifferentiated cells or mimic control transfected cells on Human Whole Genome (4x44k) Expression Microarrays (Agilent) per manufacturer's instructions. Three technical replicates were combined for each treatment, and a twofold cutoff (log ratio of greater than 0.3 or less than −0.3) and *P* values <0.001 were applied to identify genes that were differentially regulated. Agglomerative hierarchical clustering was performed using cosine correlation distance metrics. Each row of the heat map represents a gene.

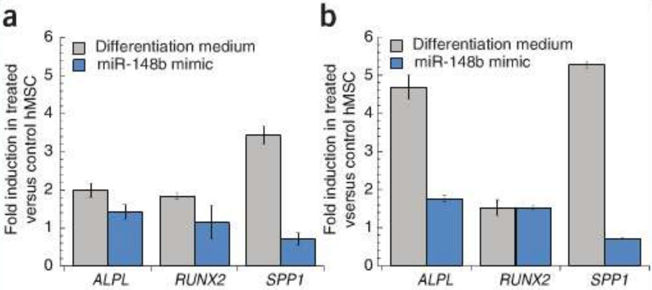


Figure 3 | Expression of three characterized osteoblast differentiation markers for differentiation medium–treated or miR-148b mimic–treated hMSCs. (**a**) Expression of the three gene markers as measured by microarray analysis (Agilent). (**b**) Expression of the three gene markers based on qPCR data obtained using Solaris qPCR Gene Expression Assays. qPCRs were carried out in a 12.5-μl reaction volume (final oligonucleotide concentrations of 800 nM of each primer and 200 nM of MGB probe) in Solaris Gene Expression master mix using a standard thermal cycling protocol on a Roche LightCycler480. Expression data were normalized to a *GAPDH* reference gene using the *ΔΔCq* method[6]. All data are presented as fold change over the reference samples (differentiation medium/propagation medium– or miR-148b/mimic control–treated cells) for *ALPL*, *SPP1* and *RUNX2*.

Microarray expression analysis identified 891 genes as differentially regulated as a result of treatment with differentiation medium, and 686 as differentially regulated by the miR-148b mimic treatment (analyzed using Rosetta Resolver software). Among these, 190 genes were regulated by both treatments (**Fig. 2**). The majority of these genes (143) were regulated in the same direction (up or down) by both treatments.

We examined differential expression of three characterized early osteoblast marker genes[3] in more detail: *ALPL* (alkaline phosphatase), *SPP1* (secreted phosphoprotein 1) and *RUNX2* (runt DNA-binding domain transcription factor). Based on the microarray analysis, *ALPL* and *RUNX2* were modestly induced approximately two fold under medium treatment and only modestly under miRNA treatment (**Fig. 3a**). *SPP1* expression was induced only by the medium treatment, by approximately 3.5-fold, and was slightly reduced by the miRNA treatment.

We then validated the expression levels of the same early osteogenic markers using Solaris qPCR Gene Expression Assays. We observed upregulation of all three osteogenic markers in differentiation medium–treated hMSCs: *ALPL* and *SPP1* were induced ~4.5-fold and >5-fold, respectively, whereas *RUNX2* was only mildly induced (**Fig. 3b**). The relatively low induction of *RUNX2* expression on day 6 is not surprising as this transcription factor is typically upregulated at the onset of osteogenic differentiation[4]. *ALPL* and *RUNX2* were mildly induced in miR-148b mimic–treated hMSCs. *SPP1* gene expression, however, was slightly decreased in miR-148b mimic–treated cells, in contrast to the marked induction observed with the differentiation medium treatment. This supports previously published data demonstrating a decrease in the *SPP1* expression caused by the miR-148b mimic[2].

The microarray and qPCR detection methods revealed relatively similar expression levels for both treatments, with the exception of *ALPL* (for which higher expression was indicated with qPCR detection). Although these two gene expression detection methods are commonly used for identification and validation, discrepancies between them are sometimes observed owing to the differences in sensitivity and dynamic range[5]. The similarities in gene expression for *ALPL* and *RUNX2* osteogenic markers and the 143 genes identified in the expression profiling that are commonly regulated between differentiation medium– and miRNA mimic–treated cells further support a role for miR-148b in the stimulation of osteogenic differentiation of hMSCs.

Follow-up qPCR studies using Solaris qPCR Gene Expression Assays will provide a more robust and quantitative assessment of these gene expression changes and a foundation for further study of the osteoblast differentiation mechanism and miRNA involvement in this process.

1. Lukhtanov, E.A., Lokhov, S.G., Gorn, V.V., Podyminogin, M.A. & Mahoney, W. Novel DNA probes with low background and high hybridization-triggered fluorescence. *Nucleic Acids Res.* **35**, e30 (2007).
2. Schoolmeesters, A. *et al.* Functional profiling reveals critical role for miRNA in differentiation of human mesenchymal stem cells. *PLoS One* **4**, e5605 (2009).
3. Kulterer, B. *et al.* Gene expression profiling of human mesenchymal stem cells derived from bone marrow during expansion and osteoblast differentiation. *BMC Genomics* **8**, 70 (2007).
4. Karsenty, G. Minireview: transcriptional control of osteoblast differentiation. *Endocrinology* **142**, 2731–2733 (2001).
5 Morey, J.S., Ryan, J.C. & Van Dolah, F.M. Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR. *Biol. Proced. Online* **8**, 175–193 (2006).
6. Pfaffl, M.W. Quantification strategies in real-time PCR in *A-Z of Quantitative PCR* (Bustin, S.A., ed.) 87–112 (International University, La Jolla, California, USA, 2004).

This article was submitted to *Nature Methods* by a commercial organization and has not been peer reviewed. *Nature Methods* takes no responsibility for the accuracy or otherwise of the information provided.