15. K. Shuai, C. Schindler, V. R. Prezioso, J. E. Darnell Jr., *Science* **258**, 1808 (1992).
16. R. W. Blakesley *et al.*, *Genome Res.* **14**, 2235 (2004).
17. K. D. Pruitt, T. Tatusova, D. R. Maglott, *Nucleic Acids Res.* **31**, 34 (2003).
18. International HapMap Consortium, *Nat. Rev. Genet.* **5**, 467 (2004).
19. W. J. Kent *et al.*, *Genome Res.* **12**, 996 (2002).
20. P. T. Spellman *et al.*, *Genome Biol* **3**, RESEARCH0046 (2002).
21. N. D. Trinklein *et al.*, *Genome Res.* **14**, 62 (2004).
22. B. Ren, B. D. Dynlacht, *Methods Enzymol.* **376**, 304 (2004).
23. ENCODE Consortium, unpublished data.
24. Mammalian Gene Collection (MGC) Project Team, *Genome Res.***14**, 2121 (2004).
25. International HapMap Consortium, *Nature* **426**, 789 (2003).
26. A. Felsenfeld, J. Peterson, J. Schloss, M. Guyer, *Genome Res.* **9**, 1 (1999).
27. A. Siepel, D. Haussler, in *Statistical Methods in Molecular Evolution*, R. Nielsen, Ed. (Springer, New York, in press).
28. M. Blanchette *et al.*, *Genome Res.* **14**, 708 (2004).
29. The Consortium thanks the ENCODE Scientific Advisory Panel for their helpful advice on the project: G. Weinstock, G. Churchill, M. Eisen, S. Elgin, S. Elledge, J. Rine, and M. Vidal. We thank D. Leja, and M. Cichanowski for their work in creating figures for this paper. Supported by the National Human Genome Research Institute, the National Library of Medicine, the Wellcome Trust, and the Howard Hughes Medical Institute.

Supporting Online Material
www.sciencemag.org/cgi/content/full/306/5696/636/DC2
Tables S1 to S3

VIEWPOINT

Systems Biology and New Technologies Enable Predictive and Preventative Medicine

Leroy Hood,[1]* James R. Heath,[2,3] Michael E. Phelps,[3] Biaoyang Lin[1]

Systems approaches to disease are grounded in the idea that disease-perturbed protein and gene regulatory networks differ from their normal counterparts; we have been pursuing the possibility that these differences may be reflected by multiparameter measurements of the blood. Such concepts are transforming current diagnostic and therapeutic approaches to medicine and, together with new technologies, will enable a predictive and preventive medicine that will lead to personalized medicine.

Biological information is divided into the digital information of the genome and the environmental cues that arise outside the genome. Integration of these types of information leads to the dynamic execution of instructions associated with the development of organisms and their physiological responses to their environments. The digital information of the genome is ultimately completely knowable, implying that biology is unique among the sciences, in that biologists start their quest for understanding systems with a knowable core of information. Systems biology is a scientific discipline that endeavors to quantify all of the molecular elements of a biological system to assess their interactions and to integrate that information into graphical network models (*1*–*4*) that serve as predictive hypotheses to explain emergent behaviors.

The genome encodes two major types of information: (i) genes whose proteins execute the functions of life and (ii) cis control elements. Proteins may function alone, in complexes, or in networks that arise from protein interactions or from proteins that are interconnected functionally through small molecules (such as signal transduction or metabolic networks). The cis control elements, together with transcription factors, regulate the levels of expression of individual genes. They also form the linkages and architectures of the gene regulatory networks that integrate dynamically changing inputs from signal transduction pathways and provide dynamically changing outputs to the batteries of genes mediating physiological and developmental responses (*5*, *6*). The hypothesis that is beginning to revolutionize medicine is that disease may perturb the normal network structures of a system through genetic perturbations and/or by pathological environmental cues, such as infectious agents or chemical carcinogens.

Systems Approaches to Model Systems and Implications for Disease

A model of a metabolic process (galactose utilization) in yeast was developed from existing literature data to formulate a network hypothesis that was tested and refined through a series of genetic knockouts and environmental perturbations (*7*). Messenger RNA (mRNA) concentrations were monitored for all 6000 genes in the genome, and these data were integrated with protein/protein and protein/DNA interaction data from the literature by a graphical network program (Fig. 1).

The model provided new insights into the control of a metabolic process and its interactions with other cellular processes. It also suggested several concepts for systems approaches to human disease. Each genetic knockout strain had a distinct pattern of perturbed gene expression, with hundreds of mRNAs changing per knockout. About 15% of the perturbed mRNAs potentially encoded secreted proteins (*8*). If gene expression in diseased tissues also reveals patterns characteristic of pathologic, genetic, or environmental changes that are, in turn, reflected in the pattern of secreted proteins in the blood, then perhaps blood could serve as a diagnostic window for disease analysis. Furthermore, protein and gene regulatory networks dynamically changed upon exposure of yeast to an environmental perturbation (*9*). The dynamic progression of disease should similarly be reflected in temporal change(s) from the normal state to the various stages of disease-perturbed networks.

Systems Approaches to Prostate Cancer

Cancer arises from multiple spontaneous and/or inherited mutations functioning in networks that control central cellular events (*10*–*12*). It is becoming clear from our research that the evolving states of prostate cancer are reflected in dynamically changing expression patterns of the genes and proteins within the diseased cells.

A first step toward constructing a systems biology network model is to build a comprehensive expressed-mRNA database on the cell type of interest. We have used a technology called multiple parallel signature sequencing (MPSS) (*13*) to sequence a complementary DNA (cDNA) library at a rate of a million sequences in a single run and to detect mRNA transcripts down to one or a few copies per cell. A database containing more than 20 million mRNA signatures was constructed for normal prostate tissues and an androgen-sensitive prostate cancer cell line, LNCaP, in four states: androgen-starved,

[1]Institute for Systems Biology, Seattle, WA, USA. [2]Department of Chemistry, California Institute of Technology, Pasadena, CA, USA. [3]Department of Molecular and Medical Pharmacology, The David Geffen School of Medicine at the University of California Los Angeles, Los Angeles, CA, USA.

*To whom correspondence should be addressed. E-mail: lhood@systemsbiology.org

androgen-stimulated, normal conditions, and an androgen-insensitive variant. In comparing the androgen-sensitive (typical of early-stage cancer) and androgen-insensitive (typical of late-stage cancer) stages (*14*, *15*), thousands of changes in mRNA expression were identified but, out of 554 expressed transcription factors, 112 changed between the early- and late-stage cell lines (80% of which were missed when cDNA arrays were used), and a similar number changed between the cancerous cells and normal tissue. By comparing the prostate database with a tissue-wide database of 58 million MPSS signatures from 29 normal tissues from Lynx Therapeutics, about 300 prostate-specific genes (Fig. 2) were identified, approximately 60 of which possessed signal peptides, suggesting that they may be secreted (*8*). Antibodies to one of these proteins recognized, by blood analyses, 5 out of 10 early and 5 out of 10 late prostate cancers (*16*). In contrast, the standard prostate cancer blood marker, PSA, recognized no early cancers but many of the late prostate cancers, including all of those missed by our marker. Thus, two markers are better than one, and by extension a panel of multiple markers might recognize most early and late prostate cancers.

Several groups have documented the fact that (unidentified) molecules in blood serum, detected by mass spectrometry, reflect various stages of cancer (*17*–*20*). Aebersold's group has succeeded in identifying many of these biomarkers through the use of a glycoprotein capture method, coupled with isotopic labeling and analyses by mass spectrometry (*21*, *22*). Molecular diagnostics will increasingly play a key role in providing direct measures of disease biology for selecting and following therapeutic responses.

Given enough measurements, one can presumably identify distinct patterns for each of the distinct types of a particular cancer, the various stages in the progression of each disease type, the partition of the disease into categories defined by critical therapeutic targets, and the measurement of how drugs alter the disease patterns. The fascinating question is how many parameters need to be measured in order to stratify and follow the progression of various prostate cancers, or to stratify and follow the progression of the most frequent 20 or 30 cancers, or eventually the most common diseases. Finally, changes in the tissue-specific markers might identify critical points within the network. It is the key nodal points within these perturbed networks that may be affected by drugs, either to convert the diseased network back toward normalcy or to permit the specific killing of the diseased cells. Thus, multiparameter blood measurements will not only be invaluable for diagnostics but also for rationalizing the discovery of appropriate drug targets. In this scenario, molecular diagnostics will become an invaluable tool for molecular therapeutics.

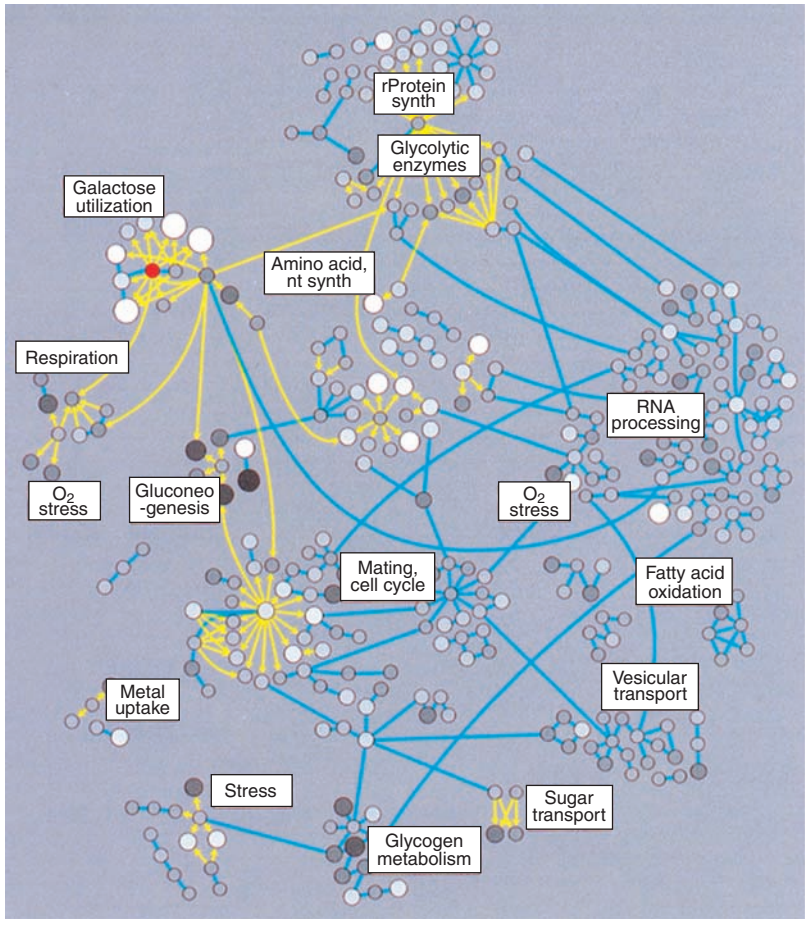

Fig. 1. A network perturbation model of galactose utilization in yeast. This model reflects the integration of mRNA levels for the 6000 yeast genes in each of 20 different genetic and environmental perturbations, as well as thousands of protein/protein and protein/DNA interactions from the literature. The software program Cytoscape (*54*) integrated these data into a network where the nodes represent proteins (encoded by genes) and the lines represent interactions (blue straight lines, protein/protein interactions; yellow lines with arrows, protein/DNA interactions). A gray scale represents the levels of mRNA, with black being abundant levels and white very low levels. The red node indicates that this network model reflects the knockout of the corresponding gene (and protein) *gal 4*—a key transcription factor. rProtein, ribosomal protein; nt, nucleotide; synth, synthesis.

Toward Analyses of Single Cells and Single Molecules

The systems biology approach toward constructing a predictive network model of a metabolic process in yeast required ~10^5 measurements. For the prostate cancer example, roughly 10^8 measurements were sufficient to begin constructing a large set of cancer markers that could be correlated back to the digital code of the genome. However, for constructing a predictive model of human disease, methods that can address the heterogeneity that characterizes biology—from the differences in how individual cells respond to environmental perturbations, to the diversity of cell types and environments within real tissues—will be critical.

In the prostate, there are neuroepithelial cells, various stromal cells, endothelial cells, and epithelial cells (from which 95% of cancers arise), each of which exhibits a continuous developmental cycle. One cannot reliably generate information for networks from mixed populations of cells. Various investigators have used cell sorting (*23*), manual dissection (*24*), or laser capture microdissection (LCM) (*25*) to obtain relatively homogeneous populations of cells. However, cell sorting and LCM themselves may cause processing-induced changes in gene expression (*26*, *27*), and manual microdissection rarely provides completely homogeneous cell types. Furthermore, even cells of one type typically represent different stages of a developmental or physiological process. Biologists would like to analyze individual cells for the key measurements of systems biology, so that network hypotheses could be generated from individual cells. The mRNAs from single cells have been analyzed after polymerase

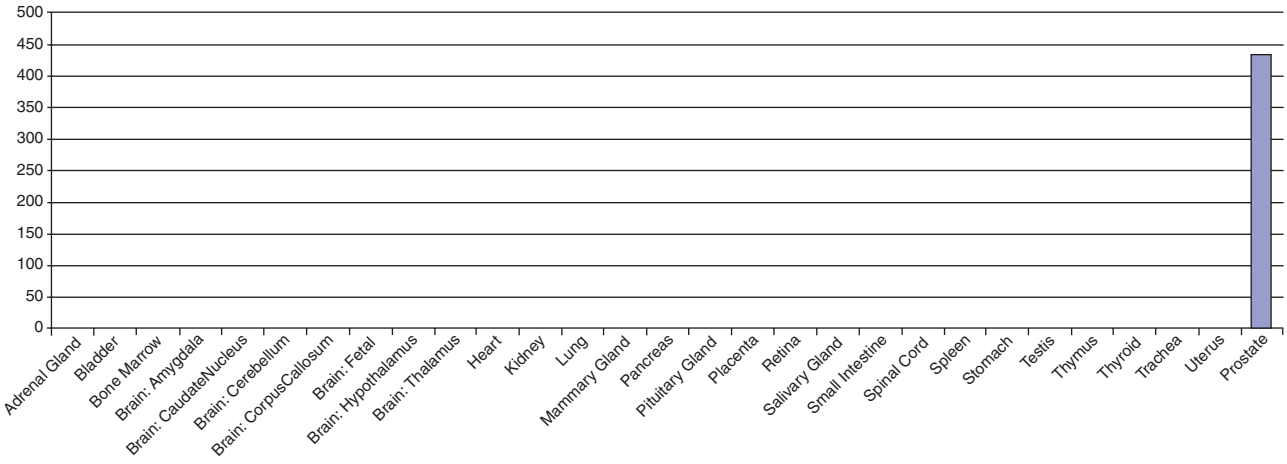

Fig. 2. A prostate-specific marker identified through quantitative profiling of all mRNAs across all 29 major organs in the human body. The gene *HOXB13* is expressed at 432 transcripts per million in the prostate tissue but is not expressed in the other 28 normal tissues. This method has been used to identify approximately 50 potential serum-based protein biomarkers for prostate cancer.

chain reaction (PCR) amplification, but there is no similar amplification technique for proteins. Thus, techniques are needed that are highly parallel, allow for multiple types of measurements (genes and proteins) and operations (such as cell sorting) to be integrated, are miniaturized (to analyze single cells and single molecules), and are automated. Here we highlight just a few of the technologies that are being driven by the needs of systems biology.

Microfluidics has existed as a useful biotechnology for some time (*28*–*30*). However, multilayer elastomer microfluidics (Fig. 3) is a powerful new technology that allows for the integration of many pumps, valves, and channels within an easily fabricated microchip. This means that multiple operations, such as cell sorting (*31*, *32*), DNA purification, and single-cell gene expression profiling (*33*), can be executed in parallel. This technology provides a bridge between biological materials and systems biology through large-scale multiparameter analysis, with applications ranging from molecular dissections of single cells (for example, from needle biopsies) and very small cell populations to multiparameter disease diagnostics from cells and blood.

Nanomechanical (*34*) and nanoelectronic (*35*, *36*) devices are emerging as highly sensitive, label-free, and real-time detectors of genes, mRNAs, and proteins. To date, demonstrations of these nanotechnologies have been at the single- or few-device level, but the reported detection sensitivities and dynamic ranges (*37*, *38*) have been spectacular. Nanofabrication methods for constructing large libraries of these devices (*39*–*43*) and integrating nanotechnologies with elastomer microfluidics (*44*) are moving forward. It is likely that within the next couple of years, miniaturized and automated microfluidics/nanotech platforms that integrate operations such as cell sorting and serum purification with measurements of 5 to 10 biomarkers from single cells or very small fluid volumes will emerge. New measurement types, such as quantifying the forces associated with protein/protein, protein/DNA, and protein/drug interactions, are possible. Other emerging nanotechnologies include tools for the

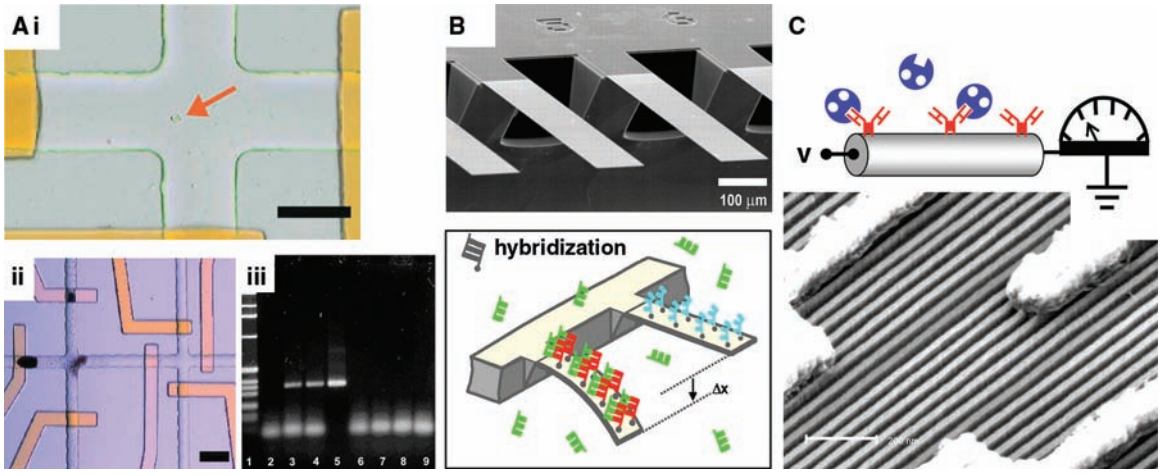

Fig. 3. Microfluidic and nanotechnology platforms. (**A**) An integrated microfluidics environment for single-cell gene expression studies. A single cell is introduced (**i**) into a 100-μm-wide channel. Before the cell is introduced, an affinity column (beads covered with oligo dT) is loaded [dark regions in (**ii**)]. The orange-colored regions in (ii) are valves that separate, for example, the empty chamber at the right from the region in which the column is being constructed. Three such valves constitute a peristaltic pump (not shown). Data from a real-time PCR analysis of the isolated mRNA (**iii**) illustrate the power of this integrated microfluidics approach. Lanes 3 and 4 correspond to one and nine cells, respectively, whereas the other lanes correspond to various controls [adapted from (*33*)]. (**B**) Array of nanomechanical biomolecular sensors. The cantilevers are fabricated to be only a few nanometers thick, with a molecular probe (such as single-stranded DNA) bonded to their top surface. DNA hybridization leads to steric crowding that forces the cantilever to bend. The bending can be detected optically or electronically [adapted from (*34*)]. (**C**) An electron micrograph showing a library of 16-nm-wide silicon nanowire biomolecular sensors. The scale bar is 200 nm, and the structures on top of the nanowires are electrical contacts. Nanowire sensors operate by binding molecular probes (such as antibodies) to the surface of a semiconducting nanowire. When the target protein binds to the probe, the conductivity properties of the nanowire are altered, and so the binding event is electronically detected. Both nanocantilevers and nanowires are capable of real-time biomolecular detection [adapted from (*55*)].

rapid sequence analysis of individual DNA molecules (*45*) and even nanoparticle-based in vivo cancer imaging probes (*46*).

These various technologies will be harnessed to generate preliminary network hypotheses for analyzing human diseases within the next few years. Those hypotheses must ultimately be tested in vivo. Such testing typically means molecular imaging, which encompasses methods ranging from bioluminescence and fluorescence (*47–50*) to positron emission tomography (PET) (*49–52*) and magnetic resonance imaging (MRI) (*48*). The challenge is to reduce the large numbers of elements delineated in the network analyses to one of a few targets of molecular imaging biomarkers that can provide critical tests of the network. For example, specific metabolic enzymes that are selectively expressed in prostate cancer cells would constitute such a target. We searched the genes that were differentially expressed between early- and late-stage prostate cancer cell lines (*15*) and determined that L-lactate dehydrogenase A, which catalyzes the formation of pyruvate from (S)-lactase, was only expressed, and at a high level, in the late-stage cancer cells. A specific PET tracer based on this reaction would serve to validate this finding and might also allow the identification of prostate cancer metastases. Molecular imaging is already being aligned with molecular therapeutics in the use of labeled drug candidates to provide direct measurements in patients by imaging pharmacokinetics of the drug throughout the body, titration of drugs to their disease targets, and measuring therapeutic effects on the biological processes of disease (*49–53*).

The Future

The medicine of today is reactive, with a focus on developing therapies for preexisting diseases, typically late in their progression. Over the next 10 to 20 years, medicine will move toward predictive and preventive modes. New technologies will allow individuals to have the relevant portions of their genomes sequenced, and multiparameter informative molecular diagnostics via blood analysis will become a routine procedure for assessing health and disease status. During this period, there will also be extensive correlations of genetic variations with disease, and this combination of advances will allow for the determination of a probabilistic future health history for each individual.

Preventive medicine will follow as disease-perturbed networks can be used to identify drug targets—first for therapy and later for prevention. Pharmacological intervention will focus on preventing disease-mediated transitions, as well as reversing or terminating those that have occurred. This will require building a fundamental understanding of the systems biology that underlies normal biological and pathological processes, and the development of new technologies that will be required to achieve this goal.

Predictive and preventative medicine will lead naturally to a personalized medicine that will revolutionize health care. Drug companies will have the opportunity for more effective means of drug discovery guided by molecular diagnostics, although the paradigm will shift to partitioning patients with a particular disease into a series of therapeutic windows, each with smaller patient populations but higher therapeutic effectiveness. Health care providers will move from dealing with disease to also promoting wellness (prevention). Finally, the public must be educated as to their roles in a very different type of medicine, as must the physicians who practice it. There will be enormous scientific and engineering challenges to achieve this vision—far greater than those associated with the Human Genome Project. Predictive, preventive, and personalized medicine will transform science, industry, education, and society in ways that we are only beginning to imagine.

References and Notes

1. E. H. Davidson *et al.*, *Science* **295**, 1669 (2002).
2. E. H. Davidson, D. R. McClay, L. Hood, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 1475 (2003).
3. H. Kitano, *Science* **295**, 1662 (2002).
4. U. Alon, *Science* **301**, 1866 (2003).
5. E. V. Rothenberg, E. H. Davidson, in *Innate Immunity*, R. A. B. Ezekowitz, J. A. Hoffman, Eds. (Humana, Totowa, NJ, 2003), pp. 61–88.
6. D. T. Odom *et al.*, *Science* **303**, 1378 (2004).
7. T. Ideker *et al.*, *Science* **292**, 929 (2001).
8. The prediction of classical secretory proteins was based on the existence of signal peptides and the number and position of transmembrane domains in a protein. The prediction of nonclassical secretory proteins (proteins without an N-terminal signal peptide) was based on the algorithm developed by J. D. Bendtsen *et al*. (http://www.cbs.dtu.dk/services/SecretomeP-1.0/).
9. A. D. Weston, N. S. Baliga, R. Bonneau, L. Hood, *Cold Spring Harbor Symp. Quant. Biol.* **68**, 345 (2003).
10. C. A. Klein, *Cell Cycle* **3**, 29 (2004).
11. P. A. Covitz, *Pharmacogenomics J.* **3**, 257 (2003).
12. D. Hanahan, R. A. Weinberg, *Cell* **100**, 57 (2000).
13. S. Brenner *et al.*, *Nature Biotechnol.* **18**, 630 (2000).
14. B. Lin *et al.*, in preparation.
15. B. Lin *et al.*, in preparation.
16. A. G. Utleg *et al.*, in preparation.
17. E. F. Petricoin 3rd *et al.*, *J. Natl. Cancer Inst.* **94**, 1576 (2002).
18. B. L. Adam *et al.*, *Cancer Res.* **62**, 3609 (2002).
19. E. P. Diamandis, *Clin. Chem.* **49**, 1272 (2003).
20. Y. Qu *et al.*, *Clin. Chem.* **48**, 1835 (2002).
21. H. Zhang, X. J. Li, D. B. Martin, R. Aebersold, *Nature Biotechnol.* **21**, 660 (2003).
22. Using this method, Aebersold's group was able to identify 100 peptides in sera that were strongly associated with cancer status in a mouse model for chemically induced skin cancer.
23. J. G. Church, E. A. Stapleton, B. D. Reilly, *Cytometry* **14**, 271 (1993).
24. L. Whetsell, G. Maw, N. Nadon, D. P. Ringer, F. V. Schaefer, *Oncogene* **7**, 2355 (1992).
25. N. L. Simone, R. F. Bonner, J. W. Gillespie, M. R. Emmert-Buck, L. A. Liotta, *Trends Genet.* **14**, 272 (1998).
26. A. Y. Liu *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 10705 (1997).
27. R. A. Craven, R. E. Banks, *Proteomics* **1**, 1200 (2001).
28. A. G. Hadd, D. E. Raymond, J. W. Halliwell, S. C. Jacobson, J. M. Ramsey, *Anal. Chem.* **69**, 3407 (1997).
29. I. Karube, K. Ikebukuro, Y. Murakami, K. Yokoyama, *Ann. N.Y. Acad. Sci.* **750**, 101 (1995).
30. L. C. Waters *et al.*, *Anal. Chem.* **70**, 158 (1998).
31. R. H. Carlson, C. V. Gabel, S. S. Chan, R. H. Austin, J. P. Brody, *Phys. Rev. Lett.* **79**, 2149 (1997).
32. A. Y. Fu, H. P. Chou, C. Spence, F. H. Arnold, S. R. Quake, *Anal. Chem.* **74**, 2451 (2002).
33. J. W. Hong, V. Studer, G. Hang, W. F. Anderson, S. R. Quake, *Nature Biotechnol.* **22**, 435 (2004).
34. J. Fritz *et al.*, *Science* **288**, 316 (2000).
35. Y. Cui, Q. Wei, H. Park, C. M. Lieber, *Science* **293**, 1289 (2001).
36. R. J. Chen *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 4984 (2003).
37. G. Wu *et al.*, *Nature Biotechnol.* **19**, 856 (2001).
38. J. Hahm, C. M. Lieber, *Nano Lett.* **4**, 51 (2004).
39. R. McKendry *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 9783 (2002).
40. Y. Arntz *et al.*, *Nanotechnology* **14**, 86 (2003).
41. N. A. Melosh *et al.*, *Science* **300**, 112 (2003).
42. Y. Bunimovich *et al.*, *Langmuir*, in press.
43. K. B. Lee, S. J. Park, C. A. Mirkin, J. C. Smith, M. Mrksich, *Science* **295**, 1702 (2002).
44. R. Pantoja *et al.*, *Biosensors and Bioelectronics*, in press; published online 6 May 2004; doi.10.1016/j.bios.2004.02.020 (http://dx.doi.org/10.1016/j.bios.2004.02.020).
45. I. Braslavsky, B. Hebert, E. Kartalov, S. R. Quake, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 3960 (2003).
46. S. A. Wickline, G. M. Lanza, *Circulation* **107**, 1092 (2003).
47. G. Choy *et al.*, *Biotechniques* **35**, 1022 (2003).
48. R. Weissleder, U. Mahmood, *Radiology* **219**, 316 (2001).
49. H. R. Herschman, *Science* **302**, 605 (2003).
50. T. F. Massoud, S. S. Gambhir, *Genes Dev.* **17**, 545 (2003).
51. M. E. Phelps, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 9226 (2000).
52. M. E. Phelps, *PET: Molecular Imaging and Biological Applications* (Springer, New York, 2004).
53. K. Shah, A. Jacobs, X. O. Breakefield, R. Weissleder, *Gene Ther.* **11**, 1175 (2004).
54. P. Shannon *et al.*, *Genome Res.* **13**, 2498 (2003).
55. E. Johnston-Halperin *et al.*, *J. Appl. Phys.*, in press.

56. We thank all the members of the NanoSystems Biology Alliance for their insights, expertise, and encouragement. We acknowledge support from the National Cancer Institute, the Army Research Office, and the Department of Energy.