of immediate data availability outweigh the potential serious errors that raw data may contain and the consequent waste of time, intellectual energy, and resources that the community will necessarily have to expend to bring scientific clarity to the data.

Looking forward, not backward. Fewer than 600 million base pairs of DNA sequence reside in the public databases, most of it redundant. If the human genome is to be completely sequenced over the next 7 to 10 years, then in each of these years the human genome sequencing community must produce accurate genomic sequence data and analysis equivalent to the sum of all DNA sequencing done to date. Such an effort would require the finishing and publication of, on average, 500 million base pairs of sequence each year—the equivalent of the *E. coli* genome being published every 2 days for the next 7 years. The nightly addition of the raw, unedited data to Web sites would double or triple the amount of information to be processed. This scenario considers only the Human Genome Project; however, projects are under way or planned for a large number of other genomes, including mouse, *Drosophila*, plants, parasites, and microbes. Given the enormous scope of the genome project, we feel that the sequencing labs need to focus on ensuring the highest quality data, analysis, and scientific interpretation, made available as soon as practicable upon completion and published in a timely fashion in peer-reviewed journals. In fact, early release of unedited, unfinished data may be detrimental to small molecular biology labs, which do not have the resources or computational tools to deal with the deluge of information.

Despite its tone of fairness, the argument for daily data release suggests indifference toward the intellectual effort that the scientific research community has set as a standard for itself in the publication and release of its work. Scientific custom has held that the scientist should be allowed to communicate to the research community what was achieved and how it was done, to analyze and comment, not only so that careful critical evaluation can be made, but also out of respect for the researcher and the achievement. Some have argued that genome sequencing is different from other scientific pursuits in that it is a public service—but what taxpayer-funded research is not a public service?

We propose that for genome sequencing projects that cannot be completed in 12 to 18 months, finished sequence data (of, for example, BAC clones) should be made available to the wider community immediately, without restriction or delay, as soon as these data have passed a series of rigorous quality control checks and have been annotated. Within a reasonable interval, these releases of data should be followed by complete scientific papers that not only describe the methods used for data generation and analysis, but also attempt to place the data in a broader biological context. We hope that the scientific journals, the scientific community, and the funding agencies will be tolerant or even encouraging of a variety of approaches to data generation, interpretation, and scientific publication to advance this exciting field of genomics.

REFERENCES AND NOTES

1. E. Marshall, *Science* **272**, 477 (1996).
2. The NIH-DOE 6-month policy (NIH-DOE Guidelines for Access to Mapping and Sequencing Data and Material Resources) can be found at http://www.nchgr.nih.gov:80/Grant_info/Funding/Statements/data_release.html.
3. The National Center for Human Genome Research rapid release policy (Policy on Availability and Patenting of Human Genomic DNA Sequence Produced by NCHGR Pilot Projects, 9 April 1996) can be found at http://www.nchgr.nih.gov/Grant_info/Funding/Statements/patenting.html.
4. The Washington University Genome Sequencing Center FTP site (ftp://genome.wustl.edu/pub/gsc1/sequence/README) reads: "This archive site contains prerelease versions of the data, and we ask that you keep the following in mind:

   1) These data should be used for internal research purposes only and not for redistribution. If other researchers request a copy of the sequence, please have them contact the *Caenorhabditis elegans* sequencing project directly.

   2) This is not the final version of the sequence and may contain errors, this is especially true of the data in the shotgun and finishing directories. However, we feel it is of sufficient interest to be prereleased.

   3) Please consult with us before publication of any results related to this sequence until this sequence has been officially released. After it has been released, you are welcome to publish the results of your research as long as Washington University and the *C. elegans* sequencing project are acknowledged appropriately."
5. R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995); C. M. Fraser *et al.*, *ibid.* **270**, 397 (1995); C. J. Bult *et al.*, *ibid.* **273**, 1058 (1996).
6. D. Lipman, personal communication.
7. M. D. Adams *et al.*, *Science* **252**, 1651 (1991); M. D. Adams *et al.*, *Nature* **377** (suppl.), 3 (1995).
8. O. White *et al.*, *Nucleic Acids Res.* **21**, 3829 (1993); C. Anderson, *Science* **259**, 1685 (1993).

The New Genomics: Global Views of Biology

Eric S. Lander

The Human Genome Project was designed as a three-step program to produce genetic maps, physical maps, and, finally, the complete nucleotide sequence map of the human chromosomes. In the past year, the first two milestones have essentially been reached (*1*) and pilot sequencing projects have begun with the aim of increasing speed and efficiency. Although only 1% of the human genome has been sequenced so far, there is growing confidence that the annual production rate can climb over the next 3 years to more than 500 megabases (Mb) worldwide—ensuring that the goal will be comfortably reached by the original, projection of 2005. The mouse, the leading biomedical model system, can likely be sequenced in parallel, although funding has not yet been committed.

With success in sight, thoughts are already turning to what should come next. The answer depends in part on how one understands the significance of the Human Genome Project. Commentators have sought to set the project in historical context by likening it to the Holy Grail, the Manhattan Project, and the moon shot. Each analogy is rich with implications about the appropriate follow-up. However, none of these precedents rings true.

Rather, the Human Genome Project is best understood as the 20th century's version of the discovery and consolidation of the periodic table. In the period from 1869 to 1889, chemists realized that it was possible to systematically enumerate all atoms and to arrange them in an array that captured their similarities and differences. The building blocks of chemistry were rendered finite, and the predictability of matter gave rise to the chemical industry on one hand and the theory of quantum mechanics on the other.

The Human Genome Project aims to produce biology's periodic table—not 100 elements, but 100,000 genes; not a rectangle reflecting electron valences, but a tree structure depicting ancestral and functional affinities among the human genes. The biological periodic table will make it possible to define unique "signatures" for each building block. Just as chemists can recognize atoms by mass and charge alone, biologists will be able to build detectors that allow each gene to be recognized from 20 well-chosen nucleotides or each protein from a distinctive fragment. Molecular biology has tended to

The author is at the Whitehead Institute for Biomedical Research and the Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142, USA. E-mail: lander@genome.wi.mit.edu

examine genes individually. It is now time to gain a global perspective on the cell by asking genome-wide questions, in many cases by studying all 100,000 genes and gene products simultaneously.

With the aim of stimulating ferment, I propose 10 goals for the next phase of genomics. The goals focus on infrastructure, inventories, and technology (rather than specific biological problems or medical applications) because scientific planning fares better at choosing paths than precise destinations. Broad dissemination of the resulting tools will clearly spawn myriad individual projects, with dramatic consequences for understanding life and curing disease.

DNA

1) *Routine re-sequencing of multi-megabase regions of human and mouse DNA.* The human genome will need to be sequenced only once, but it will be re-sequenced thousands of times in order, for example, to unravel the polygenic factors underlying human susceptibilities and predispositions. Geneticists can now trace inheritance patterns to locate chromosomal regions harboring genes for common diseases (including diabetes, hypertension, and schizophrenia), as well as modifier genes in mice that reveal surprising interactions (including suppressors of colon cancer or neurodegeneration). However, the resolution is limited, with regions often exceeding 10 Mb. Rather than assembling the thousands of human families or animal progeny needed to boost genetic resolution, the pragmatic solution in a post-genome world will be to re-sequence the entire region (or perhaps just coding regions within it) from the DNA of a few dozen affected and unaffected individuals to spot the telltale sign of correlated variation (*2*). Similarly, studies of cancer initiation and progression will involve large-scale re-sequencing of all known or suspected cancer-related genes in tumor and normal tissue. Re-sequencing will also provide the ultimate tool for genotyping studies.

Fortunately, sequencing is easier the second time around. Given a DNA sequence, one can build detectors to recognize variations on the theme. For example, Chee *et al.* (*3*) report in this issue the use of massively parallel DNA arrays to resequence by hybridization the 16 kb of the human mitochondrion. If the detector elements could be shrunk from 40 μm to the 1-μm range used in semiconductor manufacture, a set of DNA chips might someday allow re-sequencing of hundreds of megabases in a single hybridization.

2) *Systematic identification of all common variants in human genes.* The human population has vast genetic diversity, with thousands of alleles at most gene loci. Indeed, the typical mutation rate of $\approx 10^{-9}$ per nucleotide implies that every possible single base change occurs about once per generation somewhere in the population of $\approx 6 \times 10^9$ humans. Yet, human diversity is also quite limited in that most genes have only a handful of common variants in their coding regions, with the vast majority of alleles being exceedingly rare. The effective number of alleles (defined as the reciprocal of the chance that two randomly chosen alleles are identical) is rather small, often two or three (*4*). This limited diversity reflects the fact that modern humans are descended from a relatively small population that underwent exponential explosion in evolutionarily recent time (*5*).

Tantalizing examples suggest that common variants may hold the secret to many disease susceptibilities. The apolipoprotein E gene has three major variants (E2, E3, and E4) that explain a large fraction of the risk for Alzheimer's disease, as well as risk for cardiovascular disease. The MTHFR, Factor V, ACE, and CKR-5 genes each have common variants associated with, respectively, homocysteine levels, risk of thrombosis, heart disease, and resistance to HIV.

If the genes are the human "elements," the common variants are the abundant "isotopes." Creating a comprehensive catalog of common variants is a feasible task: most will be encountered by simply re-sequencing the coding regions from 100 random individuals (*6*). (This effort should be distinguished from the Human Genome Diversity project, which seeks to identify rare alleles in far-flung populations in order to reconstruct human evolution and migration.)

The catalog of common variants will transform the search for susceptibility genes through the use of association studies. Association studies test whether a genetic variant increases disease risk by comparing allele frequencies in affecteds and controls (*7*). They are logistically simpler to organize and potentially more powerful than family-based linkage studies, but they have had the practical limitation that one can only test a few guesses rather than being able to systematically scan the genome. In the post-genome world, however, it would be possible to test disease susceptibility against every common variant simultaneously (for example, by genotyping a well-characterized clinical population with a comprehensive DNA array). Notably, testing hundreds of thousands of alternative hypotheses requires only a modest increase (about eightfold) in sample-size (*8*).

Noncoding regions are also important in determining disease risk (as known for the insulin gene in juvenile diabetes). Similar approaches could be used to pinpoint the role of noncoding variation, either by systematic collection of variation in regulatory regions or by use of a dense genetic map to recognize ancestral chromosomal segments in the human population (*9*). Increasingly, population genetics will become a mainstream tool for biomedical research.

3) *Rapid de novo sequencing from other organisms.* Comparative DNA sequencing will unlock the record of 3.5 billion years of evolutionary experimentation. It will not only reveal the precise branches in the tree of life, but will elucidate the timing and character of major evolutionary innovation [including the frequency of invention of truly novel genes and the role of whole-genome duplication events, as have occurred at least twice in the vertebrate lineage (*10*)]. Sequence conservation will provide a powerful way to discern the real functional constraints on genes and gene products. Comparison of related organisms will directly reveal regulatory regions and key architectural features of proteins (*11*). Sequence differences hold the key to understanding how nature generates such diversity of form and function with such an economy of genes—producing, for example, elephants, gazelles, mice, and humans from the same basic mammalian repertoire of genes. The solution will require correlations of sequence variation with temporal and spatial differences in gene expression across organisms. It will likely reveal secrets about the fine-tuning of genes and gene networks, with valuable lessons for human medicine.

Unleashing the full power of DNA sequence comparison will depend on dramatic increases in the efficiency of sequencing. Sequencing of compact bacterial genomes has recently exploded as costs have dropped to below 50 cents per finished base. Labor and reagent costs could, in fact, be slashed through automation and miniaturization (as well as the eventual expiration of a few patents). It should be possible eventually to reach 1 cent per base with conventional automation and perhaps 0.1 cent per base with approaches such as a micro-electromechanical device that incorporates a fully integrated sequencing system on a chip. For the longer term, radical approaches such as single-molecule detection deserve attention.

RNA

4) *Simultaneous monitoring of the expression of all genes.* The mRNA levels sensitively reflect the state of the cell, perhaps uniquely defining cell types, stages, and responses. To decipher the logic of gene reg-

ulation, we should aim to be able to monitor the expression level of all genes simultaneously, with a quantitative sensitivity level of less than one copy per cell, a qualitative sensitivity to distinguish all alternatively spliced forms, and the ability to assay single cells. Recent technological advances in DNA microarrays (*12*) augur well for the eventual feasibility of this goal.

The deluge of data from whole-genome expression studies can be exploited in a variety of ways: (i) *Description*. At the simplest level, straightforward catalogs of the tissue distribution of proteases or G protein–coupled receptors may suggest roles for the gene products. (ii) *Classification*. At a deeper level, global expression profiles may reveal previously unrecognized subtypes among patients, tumors, and drugs—explaining differences in underlying mechanisms, responses to treatment, and side effects. Classification does not require specific knowledge of gene function because expression profiles may serve simply as markers in mathematical cluster analysis. (iii) *Dissecting circuitry*. The greatest challenge will be to decipher the logical circuitry controlling entire developmental or response pathways. The task is perhaps akin to reverse engineering a microprocessor based on recordings from each register. Typical experiments will involve sampling expression throughout a time course after a stimulus or set of related stimuli. In the end, it will include understanding the activation and identifying the target of every transcription factor and assembling the components into a finite list of regulatory modules and sequential cascades. Whole-genome expression studies will require new mathematical tools to analyze the data (*13*) and visualization tools to render them comprehensible.

Spatial localization of gene expression during development will also hold critical clues about function. DNA arrays will not be suitable for such investigations, but high throughput, in situ hybridization to mouse embryos should be feasible.

5) *Generic tools for manipulating cell circuitry*. To understand biological function, it is necessary not just to monitor gene expression, but to disrupt and manipulate it. The coming era will require an arsenal of generic reagents and methods for both germline and transient gene disruption.

For the most tractable genetic systems (yeast, nematodes, and fruit flies), it will be feasible to create a menagerie containing a mutant for every possible gene. These strains will help elucidate function both through direct characterization of the null phenotype and through the identification of interacting genes by screens for suppressors, enhancers, and synthetic lethals. These model organisms are likely to reveal all basic eukaryotic and metazoan functions.

There will be a similar explosion in mouse mutagenesis. Homologous recombination allows disruption of any gene, although a comprehensive program to target every mouse gene seems prohibitively expensive ($10 billion, even before the cost of long-term maintenance). Random mutagenesis with point, deletion, and insertion mutagens will undergo a revival, as global tools for DNA and RNA analysis make it easier to identify mutations and characterize their physiological effects. Interspecies hybrids will also shed light on function, through natural allelic incompatibilities.

Transgenic studies will require not only disruption, but the ability to redesign cellular wiring diagrams. One can envision a growing handbook of reusable modular components, much as architects of electronic circuits use today, that would propel studies ranging from basic research to applied gene therapy.

Transient disruption of gene expression will be more appropriate for the study of many physiological processes and essential for the study of human cells. Focused efforts are needed to improve the efficacy and availability of research reagents for disrupting mRNAs, including anti-sense oligonucleotides and ribozymes. High throughput approaches would advance the state of the art in production and evaluation of such generic tools.

For many aspects of cellular physiology, the most important circuitry occurs at the level not of RNA but of protein. Although generic manipulation of protein function is more difficult than for nucleic acid function, new advances in combinatorial chemistry hold the promise of efficient production of chemical knockout reagents that activate and inactivate proteins.

Protein

6) *Monitoring the level and modification state of all proteins*. The case for global monitoring of RNA applies with equal force to proteins, with the added twist that it will be necessary to follow post-translational modifications, which play key roles in protein circuitry. The concept of global protein monitoring dates back to the use of two-dimensional (2D) protein gel electrophoresis to visualize cellular changes. The approach is conceptually sound but technically limited, in being hard to standardize and automate and capable of detecting only a few thousand of the most abundant proteins. Instead, a future version of the 2D gel might involve an automated system to take total cellular protein, partially separate it chromatographically, proteolytically cleave each fraction, analyze it by mass spectrometry, and recognize the resulting peptides by comparing their signatures to the complete protein database. Modern mass spectrometers can provide sufficient information to allow unique recognition of protein fragments, as well as detection of secondary modifications such as phosphorylation and glycosylation. Alternatively, it may be possible to devise "chip" detectors for proteins.

7) *Systematic catalogs of protein interactions*. One approach to reconstructing cellular machinery and inferring function is to identify protein interactions, because proteins engaged in a common task (such as a signalling cascade or a macromolecular complex) often contact one another. Methods for identifying partners include generic interaction traps (such as the yeast-two hybrid system) and specialized traps (for example, to find proteins targets of kinases or DNA targets of zinc finger proteins). Improvements are needed to ensure that apparent interactions are biologically relevant, by decreasing nonspecific background and exploiting information about temporal and spatial localization of proteins.

Once all proteins are known, it should be possible to assemble comprehensive "interaction maps" of genomes. This has already been accomplished (*14*) for bacteriophage T7 (which has 55 genes), and it should be possible to automate the approach to tackle yeast or even the human genome. Finally, methods for discerning pathways on the basis of principles other than physical interaction should also be explored.

8) *Identification of all basic protein shapes*. Biochemists increasingly believe that it will be possible to infer most protein structures by analyzing their amino acid sequences against a limited compendium of basic shapes (*15*). Although it is not yet clear how rapidly this catalog will approach saturation or whether high throughput, genome-style techniques could accelerate progress, the possibility of a comprehensive approach should be considered.

Society

The societal issues raised by genomics require much more extensive discussion than is possible here, but they must not go unmentioned.

9) *Increased attention to ethical, legal and social issues* (ELSI). As genetic readouts increase in power and decrease in cost, the potential for intrusive applications will skyrocket. Future ELSI efforts will require acute scientific vision to anticipate the problems and propose safeguards.

10) *Public education*. Individuals will be faced with the choice of whether to obtain global views of their own genomes and the

need to interpret the information. At present, the general public has only a rudimentary grasp of genetics and lacks accessible ways to learn more. Although the current ELSI effort includes some educational projects, a distinct and expanded program should be launched that is aimed at education of school children, the general public, physicians, and genetic counselors.

Conclusion

The challenge ahead is to turn the periodic table produced by the era of structural genomics into tools for the coming era of functional genomics. Initial prototypes suggest that the goals are feasible, although more elegant and powerful approaches are surely needed.

It remains to be seen whether the best model to seize the opportunities ahead is a second Human Genome Project, a collection of mini-projects, many investigator-initiated grants, or all of the above. The diverse nature of the goals suggests that multiple efforts will be required rather than a single monolithic project. Some organization will surely be needed: infrastructure is not built by accident.

Whatever the organizational choice, three things are clear. The program should start now, long before the completion of the human sequence; global approaches can already be applied to the complete yeast sequence and can be implemented piecemeal as the human sequence becomes available. The program must include a major commitment to interdisciplinary training of young scientists. Finally, it must ensure broad dissemination of new technologies to every benchtop.

We live in a time of breathtaking transitions in the biological sciences. Molecular genetics has spawned a new revolution every decade and has now brought us to the brink of a global vista on life.

Note added in proof: Statistical aspects of whole-genome association studies are also discussed by Risch and Merikangas (*16*).

REFERENCES AND NOTES

1. T. J. Hudson *et al.*, *Science* **270**, 1945 (1995); C. Dib *et al.*, *Nature* **380**, 152 (1996); G. Shuler *et al.*, *Science* **274**, 540 (1996).
2. Because sites of human polymorphism occur roughly every kilobase, it will not suffice to compare a single affected individual and a single unaffected individual. However, comparison of many individuals will likely implicate the correct gene as having a high proportion of variants in affecteds.
3. M. Chee *et al.*, *Science* **274**, 610 (1996).
4. Although systematic data are limited, it is common experience to find no differences in coding sequence between two random individuals, which indicates a small effective number of alleles.
5. Under classical simplifying assumptions, the effective number of alleles at a locus is expected to be $n = 1 + 4N\mu$, where N is effective population size and μ is mutation rate. The human population was perhaps $N \approx 100{,}000$ individuals roughly 100,000 years ago [F. Ayala, *Science* **270**, 1930 (1995)] and the mutation rate for a typical gene is perhaps $\mu = 2 \times 10^{-6}$. Although only approximate, the model clearly indicates that n is small.
6. Given the limited information about population variation, I am deliberately vague about the precise populations to study. Studies of coding variation in a few dozen genes in several large populations would provide an adequate answer.
7. See, for example, E. S. Lander and N. J. Schork, *Science* **265**, 2037 (1994).
8. A Z score of $Z_1 = 1.96$ standard deviations (SDs) corresponds to the 5% significance level when testing a single variant, whereas $Z_2 = 5.3$ provides the same significance level after correction for testing 500,000 independent hypotheses (five variants in each of 100,000 genes). The required increase in sample size is roughly $(Z_2/Z_1)^2 = 7.3$.
9. The use of genetic markers to recognize chromosomes descended from a common ancestral haplotype, called linkage disequilibrium mapping, is a powerful method for fine-structure mapping of Mendelian traits in isolated populations, such as in Finland [J. Hastbacka *et al.*, *Cell* **78**, 1073 (1994)]. With a sufficiently dense genetic map, ancestral haplotypes should be detectable in larger and older populations. For example, an ancestral mutation introduced into a population 1000 generations ago (≈20,000 years) will reside in a region of about 0.2 cM (or about 200 kb) in which recombination has not altered the ancestral haplotype. The density of polymorphism should permit unique recognition of such a region.
10. P. W. Holland *et al.*, *Development*, suppl. 125 (1994).
11. Efficient approaches will be needed to rapidly isolate orthologous genes from a collection of organisms, at least until sequencing of entire genomes becomes routine.
12. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* **270**, 467 (1995); D. Lockhart *et al.*, *Nature Biotech.*, in press.
13. H. H. McAdams and L. Shapiro, *Science* **269**, 650 (1995).
14. P. L. Bartel *et al.*, *Nature Genet.* **12**, 72 (1996).
15. L. Holm and C. Sander, *Science* **273**, 595 (1996).
16. N. Risch and K. Merikangas, *ibid.*, p. 1516.

17. I thank D. Botstein, M. Chee, F. Collins, G. Duyk, E. Harlow, I. Herskowitz, R. Klausner, D. Lockhart, D. Mack, D. Page, R. Tepper, R. Weinberg, and other colleagues for valuable discussion and comments.