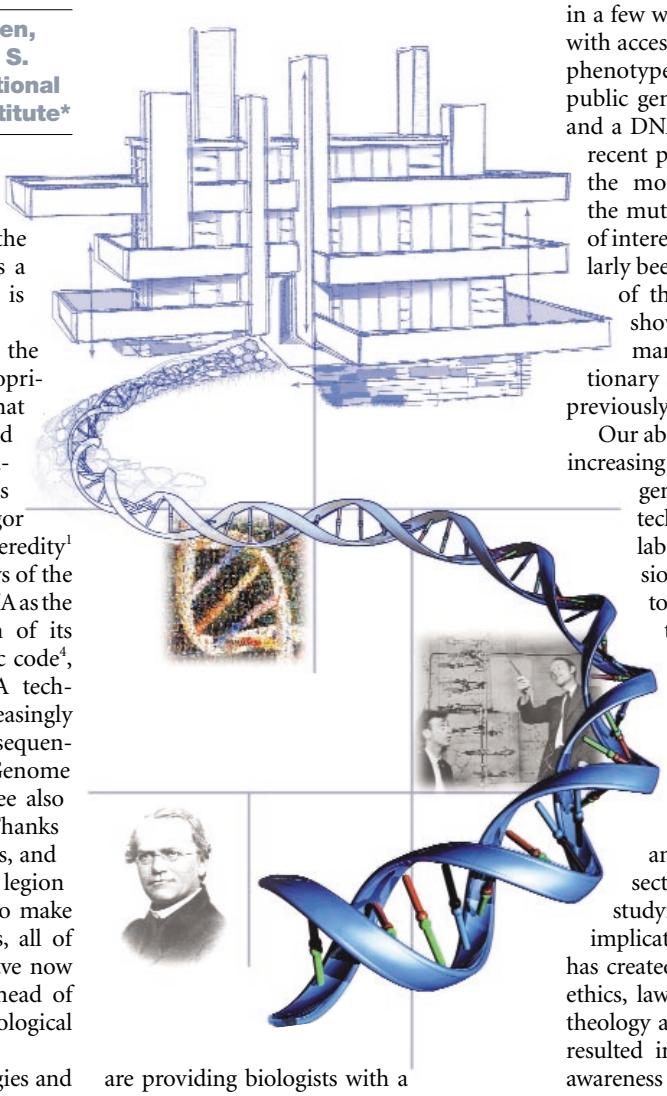
A vision for the future of genomics research

A blueprint for the genomic era.

Francis S. Collins, Eric D. Green, Alan E. Guttmacher and Mark S. Guyer on behalf of the US National Human Genome Research Institute*



DARRYL LEJA

A. BARRINGTON BROWN/SPL

The completion of a high-quality, comprehensive sequence of the human genome, in this fiftieth anniversary year of the discovery of the double-helical structure of DNA, is a landmark event. The genomic era is now a reality.

In contemplating a vision for the future of genomics research, it is appropriate to consider the remarkable path that has brought us here. The rollfold (Figure 1) shows a timeline of landmark accomplishments in genetics and genomics, beginning with Gregor Mendel's discovery of the laws of heredity[1] and their rediscovery in the early days of the twentieth century. Recognition of DNA as the hereditary material[2], determination of its structure[3], elucidation of the genetic code[4], development of recombinant DNA technologies[5,6], and establishment of increasingly automatable methods for DNA sequencing[7–10] set the stage for the Human Genome Project (HGP) to begin in 1990 (see also www.nature.com/nature/DNA50). Thanks to the vision of the original planners, and the creativity and determination of a legion of talented scientists who decided to make this project their overarching focus, all of the initial objectives of the HGP have now been achieved at least two years ahead of expectation, and a revolution in biological research has begun.

The project's new research strategies and experimental technologies have generated a steady stream of ever-larger and more complex genomic data sets that have poured into public databases and have transformed the study of virtually all life processes. The genomic approach of technology development and large-scale generation of community resource data sets has introduced an important new dimension into biological and biomedical research. Interwoven advances in genetics, comparative genomics, high-throughput biochemistry and bioinformatics are providing biologists with a markedly improved repertoire of research tools that will allow the functioning of organisms in health and disease to be analysed and comprehended at an unprecedented level of molecular detail. Genome sequences, the bounded sets of information that guide biological development and function, lie at the heart of this revolution. In short, genomics has become a central and cohesive discipline of biomedical research.

The practical consequences of the emergence of this new field are widely apparent. Identification of the genes responsible for human mendelian diseases, once a herculean task requiring large research teams, many years of hard work, and an uncertain outcome, can now be routinely accomplished in a few weeks by a single graduate student with access to DNA samples and associated phenotypes, an Internet connection to the public genome databases, a thermal cycler and a DNA-sequencing machine. With the recent publication of a draft sequence of the mouse genome[11], identification of the mutations underlying a vast number of interesting mouse phenotypes has similarly been greatly simplified. Comparison of the human and mouse sequences shows that the proportion of the mammalian genome under evolutionary selection is more than twice that previously assumed.

Our ability to explore genome function is increasing in specificity as each subsequent genome is sequenced. Microarray technologies have catapulted many laboratories from studying the expression of one or two genes in a month to studying the expression of tens of thousands of genes in a single afternoon[12]. Clinical opportunities for gene-based pre-symptomatic prediction of illness and adverse drug response are emerging at a rapid pace, and the therapeutic promise of genomics has ushered in an exciting phase of expansion and exploration in the commercial sector[13]. The investment of the HGP in studying the ethical, legal and social implications of these scientific advances has created a talented cohort of scholars in ethics, law, social science, clinical research, theology and public policy, and has already resulted in substantial increases in public awareness and the introduction of significant (but still incomplete) protections against misuses such as genetic discrimination (see www.genome.gov/PolicyEthics).

These accomplishments fulfil the expansive vision articulated in the 1988 report of the National Research Council, *Mapping and Sequencing the Human Genome*[14]. The successful completion of the HGP this year thus represents an opportunity to look forward and offer a blueprint for the future of genomics research over the next several years.

The vision presented here addresses a different world from that reflected in earlier plans published in 1990, 1993 and 1998 (refs 15–17). Those documents addressed the goals of the 1988 report, defining detailed paths towards the development of genome-

*Endorsed by the National Advisory Council for Human Genome Research, whose members are Vickie Yates Brown, David R. Burgess, Wylie Burke, Ronald W. Davis, William M. Gelbart, Eric T. Juengst, Bronya J. Keats, Raju Kucherlapati, Richard P. Lifton, Kim J. Nickerson, Maynard V. Olson, Janet D. Rowley, Robert Tepper, Robert H. Waterston and Tadataka Yamada.

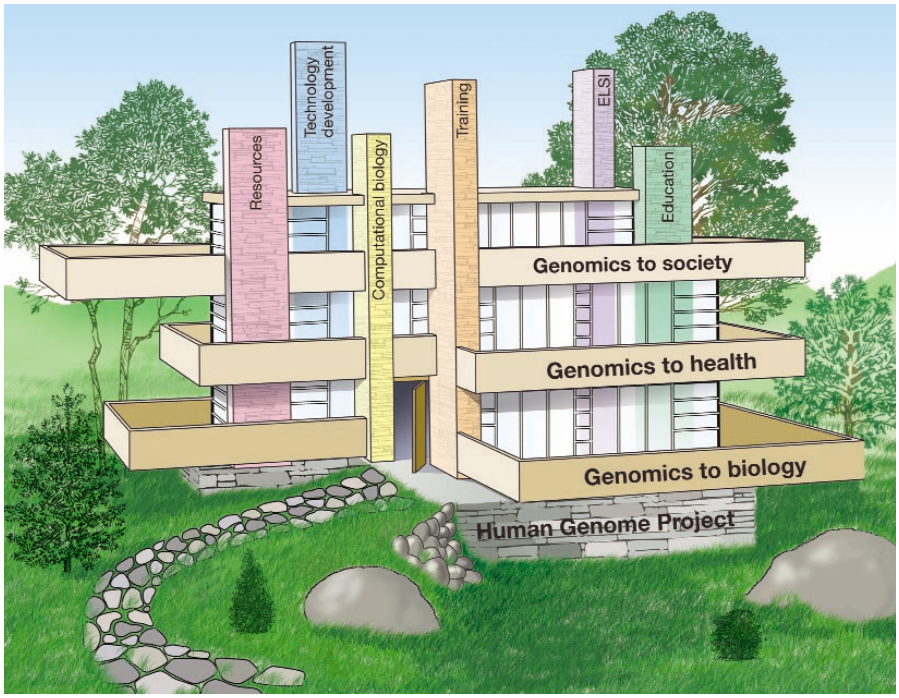


Fig 2 **The future of genomics rests on the foundation of the Human Genome Project.**

analysis technologies, the physical and genetic mapping of genomes, and the sequencing of model organism genomes and, ultimately, the human genome. Now, with the effective completion of these goals, we offer a broader and still more ambitious vision, appropriate for the true dawning of the genomic era. The challenge is to capitalize on the immense potential of the HGP to improve human health and well-being.

The articulation of a new vision is an opportunity to explore transformative new approaches to achieve health benefits. Although genome-based analysis methods are rapidly permeating biomedical research, the challenge of establishing robust paths from genomic information to improved human health remains immense. Current efforts to meet this challenge are largely organized around the study of specific diseases, as exemplified by the missions of the disease-oriented institutes at the US National Institutes of Health (NIH, www.nih.gov) and numerous national and international governmental and charitable organizations that support medical research. The National Human Genome Research Institute (NHGRI), in budget terms a rather small (less than 2%) component of the NIH, will work closely with all these organizations in exploring and supporting these biomedical research capabilities. In addition, we envision a more direct role for both the extramural and intramural programmes of the NHGRI in bringing a genomic approach to the translation of genomic sequence information into health benefits.

The NHGRI brings two unique assets to this challenge. First, it has close ties to a scientific community whose direct role over the past 13 years in bringing about the genomic revolution provides great familiarity with its potential to transform biomedical research. Second, the NHGRI's long-standing mission, to investigate the broadest possible implications of genomics, allows unique flexibility to explore the whole spectrum of human health and disease from the fresh perspective of genome science. By engaging the energetic and interdisciplinary genomics-research community more directly in health-related research and by exploiting the NHGRI's ability to pursue opportunities across all areas of human biology, the institute seeks to participate directly in translating the promises of the HGP into improved human health.

To fully achieve this goal, the NHGRI must also continue in its vigorous support of another of its vital missions — the coupling of its scientific research programme with research into the social consequences of increased availability of new genetic technologies and information. Translating the success of the HGP into medical advances intensifies the need for proactive efforts to ensure that benefits are maximized and harms minimized in the many dimensions of human experience.

A reader's guide

The vision for genomics research detailed here is the outcome of almost two years of intense discussions with hundreds of scientists and members of the public, in more than a dozen workshops and numerous individual consultations (see www.genome.gov/About/Planning). The vision is formulated into three major themes — genomics to biology, genomics to health, and genomics to society — and six crosscutting elements.

We envisage the themes as three floors of a building, firmly resting on the foundation of the HGP (Figure 2). For each theme, we present a series of grand challenges, in the spirit of the proposals put forward for mathematics by David Hilbert at the turn of the twentieth century[18]. These grand challenges are intended to be bold, ambitious research targets for the scientific community. Some can be planned on specific timescales, others are less amenable to that level of precision. We list the grand challenges in an order that makes logical sense, not representing priority. The challenges are broad in sweep, not parochial — some can be led by the NHGRI alone, whereas others will be best pursued in partnership with other organizations. Below, we clarify areas in which the NHGRI intends to play a leading role.

BOX 1 **Resources**

One of the key and distinctive objectives of the Human Genome Project (HGP) has been the generation of large, publicly available, comprehensive sets of reagents and data (scientific resources or 'infrastructure') that, along with other new, powerful technologies, comprise a toolkit for genomics-based research. Genomic maps and sequences are the most obvious examples. Others include databases of sequence variation, clone libraries and collections of anonymous cell lines. The continued generation of such resources is critical, in particular:

- Genome sequences of key mammals, vertebrates, chordates, and invertebrates
- Comprehensive reference sets of coding sequences from key species in various formats, for example, full-length cDNA sequences and corresponding clones, oligonucleotide primers, and microarrays
- Comprehensive collections of knockouts and knock-downs of all genes in selected animals to accelerate the development of models of disease
- Comprehensive reference sets of proteins from key species in various formats, for example in expression vectors, with affinity tags and spotted onto protein chips
- Comprehensive sets of protein affinity reagents
- Databases that integrate sequences with curated information and other large data sets, as well as tools for effective mining of the data
- Cohort populations for studies designed to identify genetic contributors to health and to assess the effect of individual gene variants on disease risk, including a 'healthy' cohort
- Large libraries of small molecules, together with robotic methods to screen them and access to medicinal chemistry for follow-up, to provide investigators easy and affordable access to these tools

The six critically important crosscutting elements are relevant to all three thematic areas. They are: resources (Box 1); technology development (Box 2); computational biology (Box 3); training (Box 4); ethical, legal and social implications (ELSI, Box 5); and education (Box 6). We also stress the critical importance of early, unfettered access to genomic data in achieving maximum public benefit. Finally, we propose a series of 'quantum leaps', achievements that would lead to substantial advances in genomics research and its applications to medicine. Some of these may seem overly bold, but no laws of physics need to be violated to achieve them. Such leaps would have profound implications, just as the dreams of the mid-1980s about the complete sequence of the human genome have been realized in the accomplishments now being celebrated.

I Genomics to biology

Elucidating the structure and function of genomes

The broadly available genome sequences of human and a select set of additional organisms represent foundational information for biology and biomedicine. Embedded within this as-yet poorly understood code are the genetic instructions for the entire repertoire of cellular components, knowledge of which is needed to unravel the complexities of biological systems. Elucidating the structure of genomes and identifying the function of the myriad encoded elements will allow connections to be made between genomics and biology and will, in turn, accelerate the exploration of all realms of the biological sciences.

For this, new conceptual and technological approaches will be needed to:

- Develop a comprehensive and comprehensible catalogue of all of the components encoded in the human genome.
- Determine how the genome-encoded components function in an integrated manner to perform cellular and organismal functions.
- Understand how genomes change and take on new functional roles.

Grand Challenge I-1 Comprehensively identify the structural and functional components encoded in the human genome

Although DNA is relatively simple and well understood chemically, the human genome's structure is extraordinarily complex and its function is poorly understood. Only 1–2% of its bases encode proteins[7], and the full complement of protein-coding sequences still remains to be established. A roughly equivalent amount of the non-coding portion of the genome is under active selection[11], suggesting that it is also functionally important, yet vanishingly little is known about it. It

probably contains the bulk of the regulatory information controlling the expression of the approximately 30,000 protein-coding genes, and myriad other functional elements, such as non-protein-coding genes and the sequence determinants of chromosome dynamics. Even less is known about the function of the roughly half of the genome that consists of highly repetitive sequences or of the remaining non-coding, non-repetitive DNA.

The next phase of genomics is to catalogue, characterize and comprehend the entire set of functional elements encoded in the human and other genomes. Compiling this genome 'parts list' will be an immense challenge. Well-known classes of functional elements, such as protein-coding sequences, still cannot be accurately predicted from sequence information alone. Other types of known functional sequences, such as genetic regulatory elements, are even less well understood; undoubtedly new types remain to be defined, so we must be ready to investigate novel, perhaps unexpected, ways in which DNA sequence can confer function. Similarly, a better understanding of epigenetic changes (for example, methylation and chromatin remodelling) is needed to comprehend the full repertoire of ways in which DNA can encode information.

Comparison of genome sequences from evolutionarily diverse species has emerged as a powerful tool for identifying functionally important genomic elements. Initial analyses of available vertebrate genome sequences[7,11,19] have revealed many previously undiscovered protein-coding sequences. Mammal-to-mammal sequence comparisons have revealed large numbers of homologies in non-coding regions[11], few of which can be defined in functional terms. Further comparisons of sequences derived from multiple species, especially those occupying distinct evolutionary positions, will lead to significant refinements in our understanding of the functional importance of conserved sequences[20]. Thus, the generation of additional genome sequences from several well-chosen species is crucial to the functional characterization of the human genome (Box 1). The generation of such large sequence data sets will benefit from further advances in sequencing technology that yield significant cost reductions (Box 2). The study of sequence variation within species will also be important in defining the functional nature of some sequences (see Grand Challenge I-3).

BOX 2 **Technology development**

The Human Genome Project was aided by several 'breakthrough' technological developments, including Sanger DNA sequencing and its automation, DNA-based genetic markers, large-insert cloning systems and the polymerase chain reaction. During the project, these methods were scaled up and made more efficient by 'evolutionary' advances, such as automation and miniaturization. New technologies, including capillary-based sequencing and methods for genotyping single-nucleotide polymorphisms, have recently been introduced, leading to further improvements in capacity for genomic analyses. Even newer approaches, such as nanotechnology and microfluidics, are being developed, and hold great promise, but further advances are still needed. Some examples are:

- Sequencing and genotyping technologies to reduce costs further and increase access to a wider range of investigators
- Identification and validation of functional elements that do not encode protein
- *In vivo*, real-time monitoring of gene expression and the localization, specificity, modification and activity/kinetics of gene products in all relevant cell types
- Modulation of expression of all gene products using, for example, large-scale mutagenesis, small-molecule inhibitors and knock-down approaches (such as RNA-mediated inhibition)
- Monitoring of the absolute abundance of any protein (including membrane proteins, proteins at low abundance and all modified forms) in any cell
- Improved imaging methods that allow non-invasive molecular phenotyping
- Correlating genetic variation to human health and disease using haplotype information or comprehensive variation information
- Laboratory-based phenotyping, including the use of protein affinity reagents, proteomic approaches and analysis of gene expression
- Linking molecular profiles to biology, particularly pathway biology to disease

Effective identification and analysis of functional genomic elements will require increasingly powerful computational capabilities, including new approaches for tackling ever-growing and increasingly complex data sets and a suitably robust computational infrastructure for housing, accessing and analysing those data sets (Box 3). In parallel, investigators will need to become increasingly adept in dealing with this treasure trove of new information (Box 4). As a better understanding of genome function is gained, refined computational tools for *de novo* prediction of the identity and behaviour of functional elements should emerge[21].

Complementing the computational detection of functional elements will be the generation of additional experimental data by high-throughput methodologies. One example is the production of full-length complementary DNA (cDNA) sequences (see, for example, www.mgc.nci.nih.gov and www.fruitfly.org/EST/full.shtml). Major challenges inherent in programmes to discover genes are the experimental identification and validation of alternate splice forms and messenger RNAs expressed in a highly restricted fashion. Even more challenging is the experimental validation of functional elements that do not encode protein (for example, regulatory regions and non-coding RNA sequences). High-throughput approaches to identify them (Box 2) will be needed to generate the experimental data that will be necessary to develop, confirm and enhance computational methods for detecting functional elements in genomes.

Because current technologies cannot yet identify all functional elements, there is a need for a phased approach in which new methodologies are developed, tested on a pilot scale and finally applied to the entire human genome. Along these lines, the NHGRI recently launched the Encyclopedia of DNA Elements (ENCODE) Project (www.genome.gov/Pages/Research/ENCODE) to identify all the functional elements in the human genome. In a pilot project, systematic strategies for identifying all functionally important genomic elements will be developed and tested using a selected 1% of the human genome. Parallel projects involving well-studied model organisms, for example, yeast, nematode and fruitfly, are ongoing. The lessons learned will serve as the basis for implementing a broader programme for the entire human genome.

Grand Challenge I-2 Elucidate the organization of genetic networks and protein pathways and establish how they contribute to cellular and organismal phenotypes

Genes and gene products do not function independently, but participate in complex, interconnected pathways, networks and molecular systems that, taken together, give rise to the workings of cells, tissues, organs and organisms. Defining these systems and determining their properties and interactions is crucial to understanding how biological systems function. Yet these systems are far more complex than any problem that molecular biology, genetics or genomics has yet approached. On the basis of previous experience, one effective path will begin with the study of relatively simple model organisms[22], such as bacteria and yeast, and then extend the early findings to more complex organisms, such as mouse and human. Alternatively, focusing on a few well-characterized systems in mammals will be a useful test of the approach (see, for example, www.signaling-gateway.org).

Understanding biological pathways, networks and molecular systems will require information from several levels. At the genetic level, the architecture of regulatory interactions will need to be identified in different cell types, requiring, among other things, methods for simultaneously monitoring the expression of all genes in a cell[12]. At the gene-product level, similar techniques that allow *in vivo*, real-time measurement of protein expression, localization, modification and activity/kinetics will be needed (Box 2). It will be important to develop, refine and scale up techniques that modulate gene expression, such as conventional gene-knockout methods[23], newer knock-down approaches[24] and small-molecule inhibitors[25] to establish the temporal and cellular expression pattern of individual proteins and to determine the functions of those proteins. This is a key first step towards assigning all genes and their products to functional pathways.

The ability to monitor all proteins in a cell simultaneously would profoundly improve our ability to understand protein pathways and systems biology. A critical step towards gaining a complete understanding of systems biology will be to take an accurate census of the proteins present in particular cell types under different physiological conditions. This is becoming possible in some model systems, such as microorganisms[26]. It will be a major challenge to catalogue proteins present in low abundance or in membranes. Determining the absolute abundance of each protein, including all modified forms, will be an important next step. A complete interaction map of the proteins in a cell, and their cellular locations, will serve as an atlas for the biological and medical explorations of cellular metabolism[27] (see www.nrcam.uchc.edu, for example). These and other related areas constitute the developing field of proteomics.

BOX 3 **Computational biology**

Computational methods have become intrinsic to modern biological research, and their importance can only increase as large-scale methods for data generation become more prominent, as the amount and complexity of the data increase, and as the questions being addressed become more sophisticated. All future biomedical research will integrate computational and experimental components. New computational capabilities will enable the generation of hypotheses and stimulate the development of experimental approaches to test them. The resulting experimental data will, in turn, be used to generate more refined models that will improve overall understanding and increase opportunities for application to disease. The areas of computational biology critical to the future of genomics research include:

- New approaches to solving problems, such as the identification of different features in a DNA sequence, the analysis of gene expression and regulation, the elucidation of protein structure and protein–protein interactions, the determination of the relationship between genotype and phenotype, and the identification of the patterns of genetic variation in populations and the processes that produced those patterns
- Reusable software modules to facilitate interoperability
- Methods to elucidate the effects of environmental (non-genetic) factors and of gene–environment interactions on health and disease
- New ontologies to describe different data types
- Improved database technologies to facilitate the integration and visualization of different data types, for example, information about pathways, protein structure, gene variation, chemical inhibition and clinical information/phenotypes
- Improved knowledge management systems and the standardization of data sets to allow the coalescence of knowledge across disciplines

Establishing a true understanding of how organized molecular pathways and networks give rise to normal and pathological cellular and organismal phenotypes will require more than large, experimentally derived data sets. Once again, computational investigation will be essential (Box 3), and there will be a greatly increased need for the collection, storage and display of the data in robust databases. By modelling specific pathways and networks, predicting how they affect phenotype, testing hypotheses derived from these models and refining the models based on new experimental data, it should be possible to understand more completely the difference between a 'bag of molecules' and a functioning biological system.

Grand Challenge I-3 Develop a detailed understanding of the heritable variation in the human genome

Genetics seeks to correlate variation in DNA sequence with phenotypic differences (traits). The greatest advances in human genetics have been made for traits associated with variation in a single gene. But most phenotypes, including common diseases and variable responses to pharmacological agents, have a more complex origin, involving the interplay between multiple genetic factors (genes and their products) and non-genetic factors (environmental influences). Unravelling such complexity will require both a complete description of the genetic variation in the human genome and the development of analytical tools for using that information to understand the genetic basis of disease.

Establishing a catalogue of all common variants in the human population, including single-nucleotide polymorphisms (SNPs), small deletions and insertions, and other structural differences, began in earnest several years ago. Many SNPs have been identified[28], and most are publicly available (www.ncbi.nlm.nih.gov/SNP). A public collaboration, the International HapMap Project (www.genome.gov/Pages/Research/HapMap), was formed in 2002 to characterize the patterns of linkage disequilibrium and haplotypes across the human genome and to identify subsets of SNPs that capture most of the information about these patterns of genetic variation to enable large-scale genetic association studies. To reach fruition, such studies need more robust experimental (Box 2) and computational (Box 3) methods that use this new knowledge of human haplotype structure[29].

A comprehensive understanding of genetic variation, both in humans and in model organisms, would facilitate studies to establish relationships between genotype and biological function. The study of particular variants and how they affect the functioning of specific proteins and protein pathways will yield important new insights about physiological

processes in normal and disease states. An enhanced ability to incorporate information about genetic variation into human genetic studies would usher in a new era for investigating the genetic bases of human disease and drug response (see Grand Challenge II-1).

Grand Challenge I-4 Understand evolutionary variation across species and the mechanisms underlying it

The genome is a dynamic structure, continually subjected to modification by the forces of evolution. The genomic variation seen in humans represents only a small glimpse through the larger window of evolution, where hundreds of millions of years of trial-and-error efforts have created today's biosphere of animal, plant and microbial species. A complete elucidation of genome function requires a parallel understanding of the sequence differences across species and the fundamental processes that have sculpted their genomes into the modern-day forms.

The study of inter-species sequence comparisons is important for identifying functional elements in the genome (see Grand Challenge I-1). Beyond this illuminating role, determining the sequence differences between species will provide insight into the distinct anatomical, physiological and developmental features of different organisms, will help to define the genetic basis for speciation and will facilitate the characterization of mutational processes. This last point deserves particular attention, because mutation both drives long-term evolutionary change and is the underlying cause of inherited disease. The recent finding that mutation rates vary widely across the mammalian genome[11] raises numerous questions about the molecular basis for these evolutionary changes. At present, our understanding of DNA mutation and repair, including the important role of environmental factors, is limited.

Genomics will provide the ability to substantively advance insight into evolutionary variation, which will, in turn, yield new insights into the dynamic nature of genomes in a broader evolutionary framework.

Grand Challenge I-5 Develop policy options that facilitate the widespread use of genome information in both research and clinical settings

Realization of the opportunities provided by genomics depends on effective access to the

BOX 4 **Training**

Meeting the scientific, medical and social/ethical challenges now facing genomics will require scientists, clinicians and scholars with the skills to understand biological systems and to use that information effectively for the benefit of humankind. Adequate training capacity will be required to address the following needs:

◆ **Computational skills** As biomedical research is becoming increasingly data intensive, computational capability is increasingly becoming a critical skill.

◆ **Interdisciplinary skills** Although a good start has been made, expanded interactions will be required between the sciences (biology, computer science, physics, mathematics, statistics, chemistry and engineering), between the basic and the clinical sciences, and between the life sciences, the social sciences and the humanities. Such interactions will be needed at the individual level (scientists, clinicians and scholars will need to be able to bring relevant issues, concerns and capabilities from different disciplines to bear on their specific research efforts), at a collaborative level (researchers will need to be able to participate effectively in interdisciplinary research collaborations that bring biology together with many other disciplines) and at the disciplinary level (new disciplines will need to emerge at the interfaces between the traditional disciplines).

◆ **Different perspectives** Individuals from minority or disadvantaged populations are significantly under-represented as both researchers and participants in genomics research. This regrettable circumstance deprives the field of the best and brightest from all backgrounds, narrows the field of questions asked, can lessen sensitivity to cultural concerns in implementing research protocols, and compromises the overall effectiveness of the research. Genomics can learn from successful efforts in training individuals from under-represented populations in other areas of science and health (see, for example, www.genome.gov/Pages/Grants/Policies/ActionPlanGuide).

information (such as data about genes, gene variants, haplotypes, protein structures, small molecules and computational models) by a wide range of potential users, including researchers, commercial enterprises, healthcare providers, patients and the public. Researchers themselves need maximum access to the data as soon as possible (see 'Data release', below). Use of the information for the development of therapeutic and other products necessarily entails consideration of the complex issues of intellectual property (for example, patenting and licensing) and commercialization. The intellectual property practices, laws and regulations that affect genomics must adhere to the principle of maximizing public benefit, but must also be consistent with more general and longer-established intellectual property principles. Further, because genome research is global, international treaties, laws, regulations, practices, belief systems and cultures also come into play.

Without commercialization, most diagnostic and therapeutic advances will not reach the clinical setting, where they can benefit patients. Thus, we need to develop policy options for data access and for patenting, licensing and other intellectual property issues to facilitate the dissemination of genomics data.

II Genomics to health

Translating genome-based knowledge into health benefits

The sequencing of the human genome, along with other recent and expected achievements in genomics, provides an unparalleled opportunity to advance our understanding of the role of genetic factors in human health and disease, to allow more precise definition of the non-genetic factors involved, and to apply this insight rapidly to the prevention, diagnosis and

treatment of disease. The report by the US National Research Council that originally envisioned the HGP was explicit in its expectation that the human genome sequence would lead to improvements in human health, and subsequent five-year plans reaffirmed this view[15–17]. But how this will happen has been less clearly articulated. With the completion of the original goals of the HGP, the time is right to develop and apply large-scale genomic strategies to empower improvements in human health, while anticipating and avoiding potential harm.

Such strategies should enable the research community to achieve the following:

- Identify genes and pathways with a role in health and disease, and determine how they interact with environmental factors.
- Develop, evaluate and apply genome-based diagnostic methods for the prediction of susceptibility to disease, the prediction of drug response, the early detection of illness and the accurate molecular classification of disease.
- Develop and deploy methods that catalyse the translation of genomic information into therapeutic advances.

Grand Challenge II-1 Develop robust strategies for identifying the genetic contributions to disease and drug response

For common diseases, the interplay of multiple genes and multiple non-genetic factors, not a single allele, usually dictates disease susceptibility and response to treatments. Deciphering the role of genes in human health and disease is a formidable problem for many reasons, including impediments to defining biologically valid phenotypes, challenges in identifying and quantifying environmental exposures, technological obstacles to generating sufficient and useful genotypic information, and the difficulties of studying humans. Yet this problem can be solved. Vigorous development of cross-cutting genomic tools to catalyse advances in understanding the genetics of common disease and in pharmacogenomics is needed. Prominent among these will be a detailed haplotype map of the human genome (see Grand Challenge I-3) that can be used for whole-genome association studies of all diseases in all populations, as well as further advances in sequencing and genotyping technology to make such studies feasible (see 'Quantum leaps', below).

More efficient strategies for detecting rare alleles involved in common disease are also needed, as the hypothesis that alleles that increase risk for common diseases are themselves common[30] will probably not be universally true. Computational and experimental methods to detect gene–gene and gene–environment interactions, as well as methods allowing interfacing of a variety of relevant databases, are also required (Box 3). By obtaining unbiased assessments of the relative disease risk that particular gene variants contribute, a large longitudinal population-based cohort study, with collection of extensive clinical information and ongoing follow-up, would be profoundly valuable to the study of all common diseases (Box 1). Already, such projects as the UK Biobank (www.ukbiobank.ac.uk), the Marshfield Clinic's Personalized Medicine Research Project (www.mfldclin.edu/pmrp) and the Estonian Genome Project (www.geenivaramu.ee) seek to provide such resources. But if the multiple population groups in the United States and elsewhere in the world are to benefit fully and fairly from such research (see Grand Challenge II-6), a large population-based cohort study that includes full representation of minority populations is also needed.

BOX 5 Ethical, legal and social implications (ELSI)

Today's genomics research and applications rest on more than a decade of valuable investigation into their ethical, legal and social implications. As the application of genomics to health increases along with its social impact, it becomes ever more important to expand on this work. There is an increasing need for focused ELSI research that directly informs policies and practices. One can envisage a flowering of 'translational ELSI research' that builds on the knowledge gained from prior and forthcoming 'basic ELSI research', which would provide knowledge for direct use by researchers, clinicians, policy-makers and the public. Examples include:

- The development of models of genomics research that use attention to these ELSI issues for enhancing the research, rather than viewing such issues as impediments
- The continued development of appropriate and effective genomics research methods and policies that promote the highest levels of science and of protecting human subjects
- The establishment of crosscutting tools, analogous to the publicly accessible genomic maps and sequence databases that have accelerated other genomics research (examples of such tools might include searchable databases of genomic legislation and policies from around the world, or studies of ELSI aspects of introducing clinical genetic tests)
- The evaluation of new genetic and genomic tests and technologies, and effective oversight of their implementation, to ensure that only those with confirmed clinical validity are used for patient care

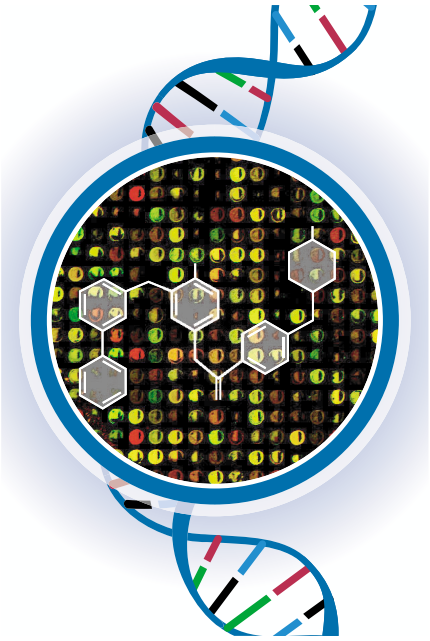
Grand Challenge II-2 Develop strategies to identify gene variants that contribute to good health and resistance to disease

Most human genetic research has traditionally focused on identifying genes that predispose to illness. A relatively unexplored, but important, area of research focuses on the role of genetic factors in maintaining good health. Genomics will facilitate further understanding of this aspect of human biology and allow the identification of gene variants that are important for the maintenance of health, particularly in the presence of known environmental risk factors. One useful research resource would be a 'healthy cohort', a large epidemiologically robust group of individuals (Box 1) with unusually good health, who could be compared with cohorts of individuals with diseases and who could also be intensively studied to reveal alleles protective for conditions such as diabetes, cancer, heart disease and Alzheimer's disease. Another promising approach would be rigorous examination of genetic variants in individuals at high risk for specific diseases who do not develop them, such as sedentary, obese smokers without heart disease or individuals with *HNPCC* mutations who do not develop colon cancer.

Grand Challenge II-3 Develop genome-based approaches to prediction of disease susceptibility and drug response, early detection of illness, and molecular taxonomy of disease states

The discovery of variants that affect risk for disease could potentially be used in individualized preventive medicine — including diet, exercise, lifestyle and pharmaceutical intervention — to maximize the likelihood of staying well. For example, the discovery of variants that correlate with successful outcomes of drug therapy, or with unfortunate side effects, could potentially be rapidly translated into clinical practice. Turning this vision into reality will require the following: (1) unbiased determination of the risk associated with a particular gene variant, often overestimated in initial studies[31]; (2) technological advances to reduce the cost of genotyping (Box 2; see 'Quantum leaps', below); (3) research on whether this kind of personalized genomic information will actually alter health behaviours (see Grand Challenge II-5); (4) oversight of the implementation of genetic tests to ensure that only those with demonstrated clinical validity are applied outside of the research setting (Box 5); and (5) education of healthcare professionals and the public to be well-informed participants in this new form of preventive medicine (Box 6).

The time is right for a focused effort to understand, and potentially to reclassify, all human illnesses on the basis of detailed molecular characterization. Systematic analyses of somatic mutations, epigenetic modifications, gene expression, protein expression and protein modification should allow the definition of a new molecular taxonomy of illness, which would replace our present, largely empirical, classification schemes and advance both disease prevention and treatment. The reclassification of neuromuscular diseases[32] and certain types of cancer[33] provides striking initial examples, but many more such applications are possible.



Such a molecular taxonomy would be the basis for the development of better methods for the early detection of disease, which often allows more effective and less costly treatments. Genomics and other large-scale approaches to biology offer the potential for developing new tools to detect many diseases earlier than is currently feasible. Such 'sentinel' methods might include analysis of gene expression in circulating leukocytes, proteomic analysis of body fluids, and advanced molecular analysis of tissue biopsies. An example would be the analysis of gene expression in peripheral blood leukocytes to predict drug response. A focused effort to use a genomic approach to characterize serum proteins exhaustively in health and disease might also be highly rewarding.

Grand Challenge II-4 Use new understanding of genes and pathways to develop powerful new therapeutic approaches to disease

Pharmaceuticals on the market target fewer than 500 human gene products[34]. Even though not all of the 30,000 or so human protein-coding genes[7] will have products targetable for drug development, this suggests that there is an enormous untapped pool of human gene-based targets for therapeutic intervention. In addition, the new understanding of biological pathways provided by genomics (see Grand Challenge I-2) should contribute even more fundamentally to therapeutic design.

The information needed to determine the therapeutic potential of a gene generally overlaps heavily with the information that reveals its function. The success of imatinib mesylate (Gleevec), an inhibitor of the BCR-ABL tyrosine kinase, in treating chronic myelogenous leukaemia relied on a detailed molecular understanding of the disease's genetic cause[35]. This example offers promise that therapies based on genomic informa-

BOX 6 **Education**

Marked health improvements from integrating genomics into individual and public health care depend on the effective education of health professionals and the public about the interplay of genetic and environmental factors in health and disease. Health professionals must be knowledgeable about genomics to use the outcomes of genomics research effectively. The public must be knowledgeable to make informed decisions about participation in genomics research and to incorporate the findings of such research into their own health care. Both groups must be knowledgeable to engage profitably in discussion and decision-making about the societal implications of genomics.

Promising models for genomics and genetics education exist (see, for example, www.nchpeg.org), but they must be expanded and new models developed. We have entered a unique 'educable era' regarding genomics; health professionals and the public are increasingly interested in learning about genomics, but its widespread application to health is still several years away. For genomics-based health care to be maximally effective once it is widely feasible, and for members of society to make the best decisions about the uses of genomics, we must take advantage now of this unique opportunity to increase understanding. Some examples are:

- Health professionals vary, both individually and by discipline, in the amount and type of genomics education that they require. So multiple models of effective genomics-related education are needed.
- Print, web and video educational products that the public can consume when actively seeking genomic information should be created and made easily available.
- The media are crucial sources of information about genomics and its societal implications. Initiatives to provide the media with greater understanding of genomics are needed.
- High-school students will be both the users of genomic information and the genomics researchers of the future. Especially as they educate all sectors of society, high-school educators need information and materials about genomics and its implications for society, to use in their classrooms.

tion will be particularly effective. Grand Challenge I-1 describes the 'functionation' of the genome, which will increasingly be the critical first step in the development of new therapeutics. But stimulating basic scientists to approach biomedical problems with a genomic attitude is not enough. A therapeutic mindset, lacking in much of academic biomedical research and training, must be explicitly encouraged, and tools developed and provided for its implementation.

A particularly promising example of the gene-based approach to therapeutics is the application of 'chemical genomics'[25]. This strategy uses libraries of small molecules (natural compounds, aptamers or the products of combinatorial chemistry) and high-throughput screening to advance understanding of biological pathways and to identify compounds that act as positive or negative regulators of individual gene products, pathways or cellular phenotypes. Although the pharmaceutical industry applies this approach widely as the first step in drug development, few academic investigators have access to this methodology or are familiar with its use.

Providing such access more broadly, through one or more centralized facilities, could lead to the discovery of a host of useful probes for biological pathways that would serve as new reagents for basic research and/or starting points for the development of new therapeutic agents (the 'hits' from such library screens will generally require medicinal chemistry modifications to yield therapeutically usable compounds).

Also needed are new, more powerful technologies for generating deep molecular libraries, especially ones tagged to allow the ready determination of precise molecular targets. A centralized database of screening results should lead to further important biological insights. Generating molecular probes for exploring the basic biology of health and disease in academic laboratories would not supplant the major role of biopharmaceutical companies in drug development, but could contribute to the start of the drug development pipeline. The private sector would doubtless find many of these molecular probes of interest for further exploration through optimization by medicinal chemistry, target validation, lead compound identification, toxicological studies and, ultimately, clinical trials.

Academic pursuit of this first step in drug development could be particularly valuable for the many rare mendelian diseases, in which often the gene defect is known but the small market size limits the private sector's motivation to shoulder the expense of effective pharmaceutical development. Such translational research in academic laboratories, combined with incentives such as the US Orphan Drug Act, could profoundly increase the availability of effective treatments for rare genetic diseases in the next decade. Further, the development of therapeutic approaches to single-gene disorders might provide valuable insights into applying genomics to reveal the biology of more common disorders and developing more effective treatments for them (in the way that, for example, the search for compounds that target the presenilins has led to general therapeutic strategies for late-onset Alzheimer's disease[36]).

Grand Challenge II-5 Investigate how genetic risk information is conveyed in clinical settings, how that information influences health strategies and behaviours, and how these affect health outcomes and costs

Understanding how genetic factors affect health is often cited as a major goal of genomics, on the assumption that applying such understanding in the clinical setting will improve health. But this assumption actually rests on relatively few examples and data, and more research is needed to provide sufficient guidance about how to use genomic information optimally for improving individual or public health.

Theoretically, the steps by which genetic risk information would lead to improved health are: (1) an individual obtains genome-based information about his/her own health risks; (2) the individual uses this information to develop an individualized prevention or treatment plan; (3) the individual implements that plan; (4) this leads to improved health; and (5) healthcare costs are reduced. Scrutiny of these assumptions is needed, both to test them and to determine how each step could best be accomplished in different clinical settings.

> The non-coding part of the human genome is functionally important, yet little is known about it.

Research is also required that critically evaluates new genetic tests and interventions in terms of parameters such as benefits, access and cost. Such research should be interdisciplinary and use the tools and expertise of many fields, including genomics, health education, health behaviour research, health outcomes research, healthcare delivery analysis, and healthcare economics. Some of these fields have historically paid little attention to genomics, but high-quality research of this sort could provide important guidance in clinical decision-making — as the work of several disciplines has already been helpful in caring for people with an increased risk of colon cancer as a result of mutations in *FAP* or *HNPCC*[37].

Grand Challenge II-6 Develop genome-based tools that improve the health of all

Disparities in health status constitute a significant global issue, but can genome-based approaches to health and disease help to reduce this problem? Social and other environmental factors are major contributors to health disparities; indeed, some would question whether heritable factors have any significant role. But population differences in allele frequencies for some disease-associated variants could be a contributing factor to certain disparities in health status, so incorporating this information into preventive and/or public-health strategies would be beneficial. Research is needed to understand the relationship between genomics and health disparities by rigorously evaluating the diverse contributions of socioeconomic status, culture, discrimination, health behaviours, diet, environmental exposures and genetics.

It is also important to explore applications of genomics in the improvement of health in the developing world (www3.who.int/whosis/genomics/genomics_report.cfm), where both human and non-human genomics will play significant roles. If we take malaria as an example, a better understanding of human genetic factors that influence susceptibility and response to the disease, and to the drugs used to treat it, could have a significant global impact. So too could a better understanding of the malarial parasite itself and of its mosquito vector, which the recently reported genome sequences[38,39] should provide. It will be necessary to determine the appropriate roles of governmental and non-governmental organizations, academic institutions, industry and individuals to ensure that genomics produces clinical benefits for resource-poor nations, and is used to produce robust local research expertise.

To ensure that genomics research benefits all, it will be critical to examine how genomics-based health care is accessed and used. What are the barriers to equitable access, and how can they be removed? This is relevant not only in resource-poor nations, but also in wealthier countries where segments of society, such as indigenous populations, the uninsured, or rural and inner city communities, have traditionally not received adequate health care.

III Genomics to society

Promoting the use of genomics to maximize benefits and minimize harms

Genomics has been at the forefront of giving serious attention, through scholarly research and policy discussions, to the impact of science and technology on society. Although the major benefits to be realized from genomics are in the area of health, as described above, genomics can also contribute to other aspects of society. Just as the HGP and related developments have spawned new areas of research in basic biology and in health, they have also created opportunities for research on social issues, even to the extent of understanding more fully how we define ourselves and each other.

In the next few years, society must not only continue to grapple with numerous questions raised by genomics, but must also formulate and implement policies to address many of them. Unless research provides reliable data and rigorous approaches on which to base such decisions, those policies will be ill-informed and could potentially compromise us all. To be successful, this research must encompass both 'basic' investigations that develop conceptual tools and shared vocabularies, and more 'applied', 'translational' projects that use these tools to explore and define appropriate public-policy options that incorporate diverse points of view.

As it has in the past, such research will continue to have important ramifications for all three major themes of the vision presented here. We now address research that focuses on society itself, more than on biology or health. Such efforts should enable the research community to:

- Analyse the impact of genomics on concepts of race, ethnicity, kinship, individual and group identity, health, disease and 'normality' for traits and behaviours.
- Define policy options, and their potential consequences, for the use of genomic information and for the ethical boundaries around genomics research.

Grand Challenge III-1 Develop policy options for the uses of genomics in medical and non-medical settings

Surveys have repeatedly shown that the public is highly interested in the concept that personal genetic information might

guide them to better health, but is deeply concerned about potential misuses of that information (see www.publicagenda.org/issues/pcc_detail.cfm?issue_type=medical_research&list=7). Topping the list of concerns is the potential for discrimination in health insurance and employment. A significant amount of research on this issue has been done[40], policy options have been published[41–43], and many US states have now passed anti-discrimination legislation (see www.genome.gov/Pages/PolicyEthics/Leg/StateIns and www.genome.gov/Pages/PolicyEthics/Leg/StateEmploy). The US Equal Employment Opportunity Commission has ruled that the Americans with Disabilities Act should apply to discrimination based on predictive genetic information[44], but the legal status of that construct remains in some doubt. Although an executive order protects US government employees against genetic discrimination, this does not apply to other workers. Thus, many observers have concluded that effective federal legislation is needed, and the US Congress is currently considering such a law.

Making certain that genetic tests offered to the public have established clinical validity and usefulness must be a priority for future research and policy making. In the United States, the Secretary's Advisory Committee on Genetic Testing extensively reviewed this area and concluded that further oversight is needed, asking the Food and Drug Administration to review new predictive genetic tests prior to marketing (www4.od.nih.gov/oba/sacgt/reports/oversight_report.pdf). That recommendation has not yet been acted on; meanwhile, numerous websites offering unvalidated genetic tests directly to the public, often combined with the sale of 'nutraceuticals' and other products of highly questionable value, are proliferating.

> It should be possible to understand the difference between a 'bag of molecules' and a biological system.

Many issues currently swirl around the proper conduct of genetic research involving human subjects, and further work is needed to achieve a satisfactory balance between the protection of research participants from harm and the ability to conduct clinical research that benefits society as a whole. Much effort has gone into developing appropriate guidelines for the use of stored tissue specimens (www.georgetown.edu/research/nrcbl/nbac/hbm.pdf), for community consultation when conducting genetic research with identifiable populations (www.nigms.nih.gov/news/reports/community_consultation.html#exec), and for the consent of non-examined family members when conducting pedigree research (www.nih.gov/sigs/bioethics/nih_third_party_rec.html), but confusion still remains for many investigators and institutional review boards.

The use of genomic information is not limited to the arenas of biology and of health, and further research and development of policy options is also needed for the many other applications of such information. The array of additional users is likely to include the life, disability and long-term care insurance industries, the legal system, the military, educational institutions and adoption agencies. Although some of the research informing the medical uses of genomics will be useful in broader settings, dedicated research outside the healthcare sphere is needed to explore the public values that apply to uses of genomics other than for health care and their relationship to specific contextual applications. For example, should genetic information on predisposition to hyperactivity be available in the future to school officials? Or should genetic information about behavioural traits be admissible in criminal or civil proceedings? Genomics also provides greater opportunity to understand ancestral origins of populations and individuals, which raises issues such as whether genetic information should be used for defining membership in a minority group.

Because uses of genomics outside the healthcare setting will involve a significantly broader community of stakeholders, both research and policy development in this area must involve individuals and organizations besides those involved in the medical applications of genomics. But many of the same perspectives essential to research and policy development for the medical uses of genomics are also essential. Both the potential users of non-medical applications of genomics and the

public need education to understand better the nature and limits of genomic information (Box 6) and to grasp the ethical, legal and social implications of its uses outside health care (Box 5).

Grand Challenge III-2 Understand the relationships between genomics, race and ethnicity, and the consequences of uncovering these relationships

Race is a largely non-biological concept confounded by misunderstanding and a long history of prejudice. The relationship of genomics to the concepts of race and ethnicity has to be considered within complex historical and social contexts.

Most variation in the genome is shared between all populations, but certain alleles are more frequent in some populations than in others, largely as a result of history and geography. Use of genetic data to define racial groups, or of racial categories to classify biological traits, is prone to misinterpretation. To minimize such misinterpretation, the biological and sociocultural factors that interrelate genetics with constructs of race and ethnicity need to be better understood and communicated within the next few years.

This will require research on how different individuals and cultures conceive of race, ethnicity, group identity and self-identity, and what role they believe genes or other biological factors have. It will also require a critical examination of how the scientific community understands and uses these concepts in designing research and presenting findings, and of how the media report these. Also necessary is widespread education about the biological meaning and limitations of research findings in this area (Box 6), and the formulation and adoption of public-policy options that protect against genomics-based discrimination or maltreatment (see Grand Challenge III-1).

Grand Challenge III-3 Understand the consequences of uncovering the genomic contributions to human traits and behaviours

Genes influence not only health and disease, but also human traits and behaviours. Science is only beginning to unravel the complicated pathways that underlie such attributes as handedness, cognition, diurnal rhythms and various behavioural characteristics. Too often, research in behavioural genetics, such as that regarding sexual orientation or intelligence, has been poorly designed and its findings have been communicated in a way that oversimplifies and overstates the role of genetic factors. This has caused serious problems for those who have been stigmatized by the suggestion that alleles associated with what some people perceive as 'negative' physiological or behavioural traits are more frequent in certain populations. Given this history and the real potential for recurrence,

it is particularly important to gather sufficient scientifically valid information about genetic and environmental factors to provide a sound understanding of the contributions and interactions between genes and environment in these complex phenotypes.

It is also important that there be robust research to investigate the implications, for both individuals and society, of uncovering any genomic contributions that there may be to traits and behaviours. The field of genomics has a responsibility to consider the social implications of research into the genetic contributions to traits and behaviours, perhaps an even greater responsibility than in other areas where there is less of a history of misunderstanding and stigmatization. Decisions about research in this area are often best made with input from a diverse group of individuals and organizations.

Grand Challenge III-4 Assess how to define the ethical boundaries for uses of genomics

Genetics and genomics can contribute understanding to many areas of biology, health and life. Some of these human applications are controversial, with some members of the public questioning the propriety of their scientific exploration. Although freedom of scientific inquiry has been a cardinal feature of human progress, it is not unbounded. It is important for society to define the appropriate and inappropriate uses of genomics. Conversations between diverse parties based on an accurate and detailed understanding of the relevant science and ethical, legal and social factors will promote the formulation and implementation of effective policies. For instance, in reproductive genetic testing, it is crucial to include perspectives from the disability community. Research should explore how different individuals, cultures and religious traditions view the ethical boundaries for the uses of genomics — for instance, which sets of values determine attitudes towards the appropriateness of applying genomics to such areas as reproductive genetic testing, 'genetic enhancement' and germline gene transfer.

> The time is right to develop and apply large-scale genomic strategies to improve human health.

Implementation: the NHGRI's role

The vision for the future of genomics presented here is broad and deep, and its realization will require the efforts of many. Continuation of the extensive collaboration between scientists and between funding sources that characterized the HGP will be essential. Although the NHGRI intends to participate in all the research areas discussed here, it will need to focus its efforts to use its finite resources as effectively as possible. Thus, it will take a major role in some areas, actively collaborate in others, and have only a supporting role in yet others. The NHGRI's priorities and areas of emphasis will also evolve as milestones are met and new opportunities arise.

The approach that has characterized genomics and led to the success of the HGP — an initial focus on technology development and feasibility studies, followed by pilot efforts to learn how to apply new strategies and technologies efficiently on a larger scale, and then implementation of full-scale production efforts — will continue to be at the heart of the NHGRI's priority-setting process. The following are areas of high interest, not listed in priority order.

Large-scale production of genomic data sets The NHGRI will continue to support genomic sequencing, focusing on the genomes of mammals, vertebrates, chordates and invertebrates; other funders will support the determination of additional genome sequences from microbes and plants. With current technology, the NHGRI could support the determination of as much as 45–60 gigabases of genomic DNA sequence, or the equivalent of 15–20 human genomes, over the next five years. But as the cost of sequencing continues to decrease, the cost/benefit ratio of sequence generation will improve, so that the actual amount of sequencing done will be greatly affected by the development of improved sequencing technology.

The decisions about which genomes to sequence next will be based on the results of comparative analyses that reveal the ability of genomic sequences from unexplored phylogenetic positions to inform the interpreta-

tion of the human sequence and to provide other insights. Finally, the degree to which any new genomic sequence is completed — finished, taken to an advanced draft stage or lightly sampled — will be determined by the use for which the sequence is generated. And, of course, the NHGRI's sequencing programme will maintain close contact with, and take account of the plans and output of, other sequencing programmes, as has happened throughout the HGP.

A second data set ready for production-level effort is the human haplotype map (HapMap). This project, a collaboration between the NHGRI, many other NIH institutes, and four international partners, is scheduled for completion within three years. The outcome of the International HapMap Project will significantly shape the future direction of the NHGRI's research efforts in the area of genetic variation.

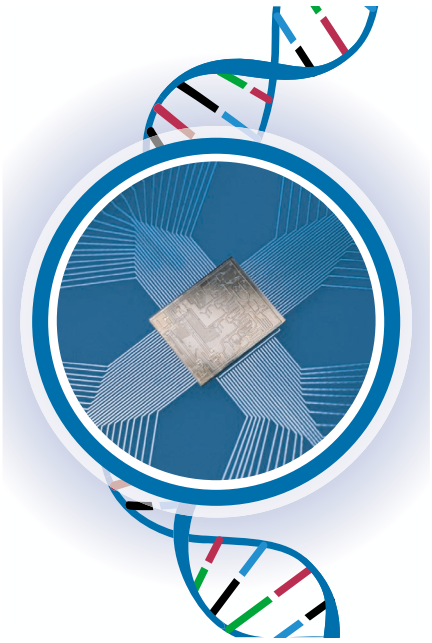
Pilot-scale efforts The NHGRI has initiated the ENCODE Project to begin the development of the human genome 'parts list'. The first phase will address the application and improvement of existing technologies for the large-scale identification of coding sequences, transcription units and other functional elements for which technology is currently available. When the results of the ENCODE Project show evidence of efficacy and affordability at the pilot scale, consideration will be given to implementing the appropriate technologies across the entire human genome.

Technology development Many areas of critical importance to the realization of the genomics-based vision for biomedical research require new technological and methodological developments before pilots and then large-scale approaches can be attempted. Recognizing that technology development is an expensive and high-risk undertaking, the NHGRI is nevertheless committed to supporting and fostering technology development in many of these crucial areas, including the following.

DNA sequencing. There is still great opportunity to reduce the cost and increase the throughput of DNA sequencing, and to make rapid, cheap sequencing available more broadly. Radical reduction of sequencing costs would lead to very different approaches to biomedical research.

Genetic variation. Improved genotyping methods and better mathematical methods are necessary to make effective use of information about the structure of variation in the human genome for identifying the genetic contributions to human diseases and other complex traits.

The genome 'parts list'. Beyond coding sequences and transcriptional units, new computational and experimental approaches are needed to allow the comprehensive determination of all sequence-encoded functional elements in genomes.



Proteomics. In the short term, the NHGRI expects to focus on the development of appropriate, scalable technologies for the comprehensive analysis of proteins and protein machines in human health and in both rare and complex diseases.

Pathways and networks. As a complement to the development of the genome 'parts list' and increasingly effective approaches to proteome analysis, the NHGRI will encourage the development of new technologies that generate a synthetic view of genetic regulatory networks and interacting protein pathways.

Genetic contributions to health, disease and drug response. The NHGRI will place a high priority on creating and applying new crosscutting genomics tools, technologies and strategies needed to identify the genetic bases of medically relevant phenotypes. Research on the genetic contributions to rare and common diseases, and to drug response, will typically involve biological systems and diseases of primary interest to other NIH institutes and other funding organizations. Accordingly, the NHGRI expects that its involvement in this area of research will often be implemented through partnerships and collaborations. The NHGRI is particularly interested in stimulating research approaches to the identification of gene variants that confer disease resistance and other manifestations of 'good health'.

Society must formulate policies to address many of the questions raised by genomics.

Molecular probes, including small molecules and RNA-mediated interference, for exploring basic biology and disease. Exploration of the feasibility of expanding chemical genomics in the academic and public sectors, particularly with regard to the establishment of one or more centralized facilities, will be pursued by the NHGRI in partnership with others.

Databases Another type of community resource for the biological and biomedical research communities is represented by databases (Box 3). But their support represents a potentially significant problem. Funding agencies, reflecting the interest of the research community, tend to prefer to use their research funds to support the generation of new data, and the ongoing need for continued and increasing support for the data archives and robust access to them is often given less attention. Both the scientific community and the funding agencies must recognize that investment in the creation and maintenance of effective databases is as important a component of research funding as data generation. The NHGRI has been a major source of support for several major genetics/genomics-oriented databases, including the Mouse Genome Database (www.informatics.jax.org/mgihome/MGD/aboutMGD.shtml), the *Saccharomyces* Genome Database (genome-www.stanford.edu/Saccharomyces), FlyBase (flybase.bio.indiana.edu), WormBase (www.wormbase.org) and Online Mendelian Inheritance in Man (www.ncbi.nlm.nih.gov/omim). The NHGRI will continue to be a leader in exploring effective solutions to the issues of integrating, displaying and providing access to genomic information.

Ethical, legal and social research The NHGRI's ELSI research activities will increasingly focus on fundamental, widely relevant, societal issues. The community of scholars and researchers working in these social fields, as well as the scope of issues being explored, need to be expanded. The ELSI research community must include individuals from minority and other communities that may be disproportionately affected by the use or misuse of genetic information. New mechanisms for promoting dialogue and collaboration between the ELSI researchers and genomic and clinical researchers need to be developed; such examples might include structural rewards for interdisciplinary research, intensive summer courses or mini-fellowships for cross-training, and the creation of centres of excellence in ELSI studies to allow sustained interdisciplinary collaboration.

Longitudinal population cohort(s) This promising research resource will be so broadly applicable, and will require such

extensive funding that, although the NHGRI might have a supporting role in design and oversight, success will demand the involvement and support of many other funding sources.

Non-genetic factors in health and disease A consequence of an improved definition of the genetic factors underlying human health and disease will be an improvement in the recognition and definition of the environmental and other non-genetic contributions to those traits. This is another area in which the NHGRI will be involved through the development of new strategies and by forming partnerships.

Use of genomic information to improve health care The NHGRI will catalyse collaboration between the diverse scholarly disciplines whose joint efforts will be necessary for research on the best ways for patients and healthcare providers to make effective use of personalized genetic information in the improvement of health. The NHGRI will also strive to ensure that research in this area is informed by, and extends knowledge of, the societal implications of genomics.

Improving the health of all people It will be important for the NHGRI to support research that explores how to ensure that genomic information is used, to the extent that such information is relevant, to reduce global health disparities. That will include a vigorous effort to increase the representation of minorities in the ranks of genomics researchers. But the full solution of the health disparities problem can only come about through a committed and sustained effort by governments, medical systems and society.

Policy development The NHGRI will continue to help facilitate public-policy development in the area of genetic/genomic science. Effective policy development will require attention to those issues for which it could have the greatest impact on the policy agenda and could help to facilitate genomic science. The NHGRI will also focus on issues that would assist the public in benefiting from genomics, such as privacy of genetic information, access to genetics services, direct-to-consumer/providers marketing, patenting and licensing of genetic information, appropriate treatment of human participants in research, and standards, usefulness and quality in genetic testing.

Data release

An important lesson of the HGP has been the benefit of immediately releasing data from large-scale sequencing projects, as embodied in the Bermuda principles (www.gene.ucl.ac.uk/hugo/bermuda.htm). Some other large-scale data production

projects have followed suit (such as those for full-length cDNAs and single-nucleotide polymorphisms), to the benefit of the scientific community. Scientific progress and public benefit will be maximized by early, open and continuing access to large data sets and by ensuring that excellent scientists are attracted to the task of producing more resources of this sort. For this system to continue to work, the producers of community-resource data sets have an obligation to make the results of their efforts rapidly available for free and unrestricted use by the scientific community, and resource users have an obligation to recognize and respect the important contribution made by the scientists who contribute their time and efforts to resource production.

Although these principles have been generally realized in the case of genomic DNA sequencing, they have not been for many other types of community-resource projects (structural biology coordinates or gene expression data, for example). The development of effective systems for achieving the rapid release of data without restrictions and for providing continued widespread access to materials and research tools should be an integral component of the planning and development of new community resources. The scientific community should also develop incentives to support the voluntary release of such data before publication by individual investigators, by appropriately rewarding and protecting the interests of scientists who wish to share their data with the community in such a generous manner.

> Scientific progress will be maximized by early, open and continuing access to large data sets.

Quantum leaps

It is interesting to speculate about potential revolutionary technical developments that might enhance research and clinical applications in a fashion that would rewrite entire approaches to biomedicine. The advent of the polymerase chain reaction, large-insert cloning systems and methods for low-cost, high-throughput DNA sequencing are examples of such advances that have already occurred.

During the course of the NHGRI's planning discussions, other ideas were raised about analogous 'technological leaps' that seem so far off as to be almost fictional but which, if they could be achieved, would revolutionize biomedical research and clinical practice.

The following is not intended to be an exhaustive list, but to provoke creative dreaming:

- the ability to determine a genotype at very low cost, allowing an association study in which 2,000 individuals could be screened with about 400,000 genetic markers for $10,000 or less;
- the ability to sequence DNA at costs that are lower by four to five orders of magnitude than the current cost, allowing a human genome to be sequenced for $1,000 or less;
- the ability to synthesize long DNA molecules at high accuracy for $0.01 per base, allowing the synthesis of gene-sized pieces of DNA of any sequence for between $10 and $10,000;
- the ability to determine the methylation status of all the DNA in a single cell; and
- the ability to monitor the state of all proteins in a single cell in a single experiment.

Conclusions

Preparing a vision for the future of genomics research has been both daunting and exhilarating. The willingness of hundreds of experts to volunteer their boldest and best ideas, to step outside their areas of self-interest and to engage in intense debates about opportunities and priorities, has added a richness and audacity to the outcome that was not fully anticipated when the planning process began. To the extent that this article captures the sense of excitement of the new discipline of genomics, it is to their credit. A complete list of the participants in this planning process can be found at www.genome.gov/About/Vision/Acknowledgements.

A final word is appropriate about the breadth of the vision articulated here. A choice had to be made between portraying a broad view of the future of genomics

research and focusing more narrowly on the specific role of the NHGRI. Recognizing that researchers and the public are more interested in the promise of the field than about the funding source responsible, we have focused here on the broad landscape of scientific opportunity. We have, however, identified the areas that are particularly appropriate for leadership by the NHGRI throughout this article. These are generally research areas that are not specific to a particular disease or organ system, but have broader biomedical and/or social implications. Yet even in those instances, the word 'partnership' appears numerous times intentionally. We expect to have partnerships not only with other public funding sources, such as the other 26 NIH institutes and centres, but also with many other governmental agencies, private foundations and private-sector organizations. Indeed, public–private partnerships, such as the SNP Consortium, the Mouse Sequencing Consortium and the International HapMap Project, provide powerful new models for the generation of public data sets with immediate and far-reaching value. Thus, many of the most exciting opportunities in genomics research cross traditional boundaries of specific disease definitions, classically defined scientific disciplines, funding sources and public versus private enterprise. The new era will flourish best in an environment where such traditional boundaries become ever more porous.

Although the opportunities described here are thought to be highly achievable, the formal initiation of specific programmes will require more detailed analysis. The relative priorities of each component must be addressed in the light of limited resources to support research. The NHGRI plans to release a revised programme announcement and other grant solicitations later this year, providing more specific guidance to extramural researchers about plans for the implementation of this vision. Furthermore, in genomics research, we have learned to expect the unexpected. From past experience, it would be surprising (and rather disappointing) if biological, medical and social contexts did not change in unpredictable ways. That reality requires that this vision be revisited on a regular basis.

In conclusion, the successful completion this month of all of the original goals of the HGP emboldens the launch of a new phase for genomics research, to explore the remarkable landscape of opportunity that now opens up before us. Like Shakespeare, we are inclined to say, "what's past is prologue" (*The Tempest*, Act II, Scene 1). If we, like bold architects, can design and build this unprecedented and noble structure, resting on the firm bedrock foundation of the HGP (Figure 2), then the true promise of genomics research for benefiting humankind can be realized.

"Make no little plans; they have no magic to stir men's blood and probably will themselves not be realized. Make big plans; aim high in hope and work, remembering that a noble, logical diagram once recorded will not die, but long after we are gone will be a living thing, asserting itself with ever-growing insistency" (attributed to Daniel Burnham, architect). ■

Francis S. Collins, Eric D. Green, Alan E. Guttmacher and Mark S. Guyer are at the National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

1. Mendel, G. Versuche über Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines, Abhandlungen, Brünn* **4,** 3–47 (1866).
2. Avery, O. T., MacLeod, C. M. & McCarty, M. Studies of the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* Type III. *J. Exp. Med.* **79,** 137–158 (1944).
3. Watson, J. D. & Crick, F. H. C. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* **171,** 737 (1953).
4. Nirenberg, M. W. The genetic code: II. *Sci. Am.* **208,** 80–94 (1963).
5. Jackson, D. A., Symons, R. H. & Berg, P. Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **69,** 2904–2909 (1972).
6. Cohen, S. N., Chang, A. C., Boyer, H. W. & Helling, R. B. Construction of biologically functional bacterial plasmids *in vitro*. *Proc. Natl Acad. Sci. USA* **70,** 3240–3244 (1973).
7. The International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
8. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94,** 441–448 (1975).
9. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl Acad. Sci. USA* **74,** 560–564 (1977).
10. Smith, L. M. *et al.* Fluorescence detection in automated DNA-sequence analysis. *Nature* **321,** 674–679 (1986).
11. The Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
12. The Chipping Forecast II. *Nature Genet.* **32,** 461–552 (2002).
13. Guttmacher, A. E. & Collins, F. S. Genomic medicine — A primer. *N. Engl. J. Med.* **347,** 1512–1520 (2002).
14. National Research Council. *Mapping and Sequencing the Human Genome* (National Academy Press, Washington DC, 1988).
15. US Department of Health and Human Services, US DOE. *Understanding Our Genetic Inheritance. The US Human Genome Project: The First Five Years.* NIH Publication No. 90-1590 (National Institutes of Health, Bethesda, MD, 1990).
16. Collins, F. & Galas, D. A new five-year plan for the US Human Genome Project. *Science* **262,** 43–46 (1993).
17. Collins, F. S. *et al.* New goals for the US Human Genome Project: 1998–2003. *Science* **282,** 682–689 (1998).
18. Hilbert, D. Mathematical problems. *Bull. Am. Math. Soc.* **8,** 437–479 (1902).
19. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297,** 1301–1310 (2002).
20. Sidow, A. Sequence first. Ask questions later. *Cell* **111,** 13–16 (2002).
21. Zhang, M. Q. Computational prediction of eukaryotic protein-coding genes. *Nature Rev. Genet.* **3,** 698–709 (2002).
22. Banerjee, N. & Zhang, M. X. Functional genomics as applied to mapping transcription regulatory networks. *Curr. Opin. Microbiol.* **5,** 313–317 (2002).
23. Van der Weyden, L., Adams, D. J. & Bradley, A. Tools for targeted manipulation of the mouse genome. *Physiol. Genomics* **11,** 133–164 (2002).
24. Hannon, G. J. RNA interference. *Nature* **418,** 244–251 (2002).
25. Stockwell, B. R. Chemical genetics: Ligand-based discovery of gene function. *Nature Rev. Genet.* **1,** 116–125 (2000).
26. Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415,** 141–147 (2002).
27. Tyson, J. J., Chen, K. & Novak, B. Network dynamics and cell physiology. *Nature Rev. Mol. Cell Biol.* **2,** 908–916 (2001).
28. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409,** 928–933 (2001).
29. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296,** 2225–2229 (2002).
30. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.* **17,** 502–510 (2001).
31. Hirschhorn, J. N., Lohmueller, K., Byrne, E. & Hirschhorn, K. A comprehensive review of genetic association studies. *Genet. Med.* **4,** 45–61 (2002).
32. Wagner, K. R. Genetic diseases of muscle. *Neurol. Clin.* **20,** 645–678 (2002).
33. Golub, T. R. Genomic approaches to the pathogenesis of hematologic malignancy. *Curr. Opin. Hematol.* **8,** 252–261 (2001).
34. Drews, J. & Ryser, S. The role of innovation in drug development. *Nature Biotechnol.* **15,** 1318–1319 (1997).
35. Druker, B. J. Imatinib alone and in combination for chronic myeloid leukemia. *Semin. Hematol.* **40,** 50–8 (2003).
36. Selkoe, D. J. Alzheimer's disease: genes, proteins, and therapy. *Physiol. Rev.* **81,** 741–66 (2001).
37. Lynch, H. T. & de la Chapelle, A. Genomic medicine: hereditary colorectal cancer. *N. Engl. J. Med.* **348,** 919–932 (2003).
38. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419,** 498–511 (2002).
39. Holt, R. A. *et al.* The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298,** 129–149 (2002).
40. Anderlik, M. R. & Rothstein, M. A. Privacy and confidentiality of genetic information: What rules for the new science? *Annu. Rev. Genom. Hum. Genet.* **2,** 401–433 (2001).
41. Hudson, K. L., Rothenberg, K. H., Andrews, L. B., Kahn, M. J. E. & Collins, F. S. Genetic discrimination and health-insurance — An urgent need for reform. *Science* **270,** 391–393 (1995).
42. Rothenberg, K. *et al.* Genetic information and the workplace: Legislative approaches and policy challenges. *Science* **275,** 1755–1757 (1997).
43. Fuller, B. P. *et al.* Policy forum: Ethics — privacy in genetics research. *Science* **285,** 1359–1361 (1999).
44. Miller, P. S. Is there a pink slip in my genes? *J. Health Care Law Policy* **3,** 225–265 (2000).

Acknowledgements

The formulation of this vision could not have happened without the thoughtful and dedicated contributions of a large number of people. The authors were greatly assisted by Kathy Hudson, Elke Jordan, Susan Vasquez, Kris Wetterstrand, Darryl Leja and Robert Nussbaum. A subcommittee of the National Advisory Council for Human Genome Research, including Wylie Burke, William Gelbart, Eric Juengst, Maynard Olson, Robert Tepper and David Valle, provided a critical sounding board for draft versions of this document. We also thank Aravinda Chakravarti, Ellen Wright Clayton, Raynard Kington, Eric Lander, Richard Lifton and Sharon Terry for serving as working-group chairs at the meeting in November 2002 that refined this document. Finally, we thank the hundreds of individuals who participated as workshop planners and/or participants during this 18-month process.