DO NOT CIRCULATE — NOT FOR PUBLICATION OR CITATION

Genomics in 2K10 and Beyond: Enhancing Foundational Knowledge and Applications to Improving Human Health

CONFIDENTIAL DRAFT: July 1, 2010

DRAFT

Other Candidate Titles:

Genomics in 2K10 and Beyond: Enhancing Foundational Knowledge and Applications in Genomic Medicine

Genomics in 2K10 and Beyond: Enhancing Foundational Knowledge and Charting a Course for Genomic Medicine

Genomics in 2K10 and Beyond: En Route to the Era of Genomic Medicine

Genomic Opportunities: Enhancing Foundational Knowledge and Applications to Improving Human Health

---

On Behalf of the National Human Genome Research Institute

Introduction

Genomics as a field has a rich history (if history is an appropriate term for a field that is less than 25 years old!) of periodically assessing itself and its relationship to the rest of biomedical research, and of ambitiously looking for future opportunities. During the Human Genome Project (1990-2003), this self-examination yielded a series of five-year, highly focused strategic plans ( ) for achieving the grail of a complete human genome sequence ( ). With that accomplished, in 2003, National Human Genome Research Institute (NHGRI) generated a broader, more open-ended vision for the future of genomics research ( ) that succeeded the more focused earlier plans.

The ideas developed in 2003 have proven to be impressively robust, and there has been significant progress on much of what was discussed in that document. At the same time, many of the objectives identified then are still relevant seven years later. And in the interim, remarkable advances in genomic technology have continued unabated, indeed often at an accelerated pace. For example, large-scale DNA sequencing is at least 100-fold more efficient now compared to the end of the Human Genome Project, genetic variation can be studied at unprecedented resolution in human and essentially any organism of interest, and new approaches are available for the study of gene expression, intracellular biomolecular interactions, biological pathways, and numerous other aspects of cellular and organismal biology. The scientific community has embraced genomics and other large-scale approaches to research, and is using them to study an amazingly broad range of problems across biomedicine and beyond, including many non-medical applications in agriculture, energy, and environmental sciences. And remarkably, genomic applications are poised to become part of medical practice.

With this background and a continued sense of wonder, NHGRI has again extensively engaged a broad swath of the scientific community (genome.gov/Planning) with the objective of identifying— in 2010 terms and with a 2010 perspective— where the field of genomics is going, what major challenges the field will face, how such challenges might be surmounted, and the key theoretical and practical issues that must be addressed for genomics to make its maximum contribution to human health and society. Given the mandate of NHGRI and the broader National Institutes of Health to improve human health, our strategic planning process has emphasized human biology and health applications, but we are well-aware that genomic opportunities extend beyond the clinical arena.

The promise offered by genomics for improving medical care and health promotion has been prominently described (some would even say overhyped) for the better part of two decades. While there are already some notable examples of direct effects on clinical practice ( ), it is fair to say that the realization of improved medical care based on genomic information (i.e., genomic medicine) is still primarily in the future. For our strategic planning process, we have found it constructive to consider two broad ways in which genomics can be used for healthcare applications. On one hand, the development of a deeper knowledge and understanding of biological principles and phenomena has historically been a successful approach for unraveling the biological basis of disease, leading to many examples of new diagnostic, therapeutic, and prevention strategies; examples range from the germ theory of disease to the development of targeted cancer

therapies, such as Gleevec™ ( ). Genomics will be an essential contributor to improving our understanding of the biology of cells and organisms, and will therefore undoubtedly be at heart of many new approaches for disease identification and management. At the same time, many major medical advances have been made in the absence of a deep molecular understanding of the disease state. Through both the use of advanced technologies for observing biological phenomena and the establishment of new correlations between genotypes and phenotypes, diagnostic, prognostic, preventive, and even therapeutic strategies can be developed even without a deep understanding of the underlying disease; examples range from correlations of drug response with specific genetic variants ( ) and the stratification of tumors based on molecular signatures ( ). In thinking about the promise of genomic medicine, it is useful to consider this fundamental distinction and which genomic-based approaches are best-suited for addressing a particular problem.

In articulating our future vision for genomics below, we first describe the continued need to gain foundational knowledge of the genome and how it functions; this includes the continued development of technologies for studying genomes and for acquiring a fundamental understanding about the genomic basis of human disease. We then discuss how those advances offer real and compelling opportunities for applying genomics to improve human health in ways that could only begin to be appreciated at the end of the Human Genome Project. But first, it is important to set a broader context.

Looking Back Before Looking Forward

We, of course, did not start from scratch in thinking about the future of genomics. In 2003, for example, we "*speculate[d] about potential revolutionary technical developments that might enhance research and clinical applications in a fashion that would rewrite entire approaches to biomedicine*," calling them 'Quantum Leaps' ( ). Remarkably, for at least two of these, developments have taken them out of the realm of 'creative dreaming' and into that of realistic expectations:

- "*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*." With current technology, 6,000 individuals can be analyzed for 5 million variants at an estimated cost of about $6.8 million, down from about $1.7 billion in 2003

- "*The ability to sequence DNA at a cost that is lower by four to five orders of magnitude than that at the end of the Human Genome Project, allowing a human genome to be sequenced for $1,000 or less*." A conservative current estimate for sequencing a human genome is $20,000-40,000, down from about $50,000,000 in 2003.

We organized our thinking in 2003 around the concept of Grand Challenges— "*bold, ambitious research targets for the scientific community*" ( ). Work on most of these has commenced, and for many, we now have a better sense of just how grand those challenges are and what technological and other advances are needed to meet them. For example, we have made considerable progress toward having "***…****a detailed*

understanding of the heritable variation in the human genome" (Grand Challenge I-3) ( ) from the HapMap project ( ), 1000 Genomes project ( ), and studies of structural variation in the human genome ( ). Yet, the question of 'missing heritability' (or the 'dark matter of the genome') remains a major unsolved problem ( ). As another example, considerable progress has been made toward "*[c]omprehensively identify[ing] the structural and functional components encoded in the human genome*" (Grand Challenge I-1) ( ), with a large number of human cell types being assayed for a broad range of functional genomic elements by the ENCODE consortium ( ). However, to acquire a comprehensive catalog of functional elements in the human genome, a much larger number of cell types and assays would be needed, which will require new strategies that increase data generation and analysis by two to three orders of magnitude.

In other instances, little progress has been made (e.g., "*[d]evelop genome-based tools that improve the health of all*"; Grand Challenge II-6); such challenges remain, being augmented by news ones that have subsequently emerged. In the strategic plan described here, we have not gone back to an organization based on Grand Challenges. Rather, we have adopted the idea of illustrating our vision for genomic opportunities through the brief description of 'Signature Projects,' efforts with discrete but compelling goals that capitalize on foundational knowledge of the genome in applying genomics en route to addressing important medical problems.

We have, however, retained one organizational concept from the 2003 document ( )—that of 'cross-cutting elements.' These are areas that are so broadly relevant across the entire spectrum of genomics research, that they deserve special attention and description. Here, we emphasize three such areas, falling under the general labels of computational biology, education and training, and societal implications of genomics. The relevance of these to all aspects of genomics research and its applications will remain unabated in the future. Each is featured in a text box below; further, specific examples of their application are mentioned throughout. We emphasize their importance here because they are simply fundamental to genomics.

Finally, a word about limitations. We have tried to limit ourselves to considering the structure and function of nucleic acids as the core of genomics. Obviously, since a major function of nucleic acids (or the genome) is to encode proteins (or the proteome), a clear distinction is often hard to make. However, it is important not to assign too much to one field. One of the hallmarks of genomics is comprehensiveness, and the value of this concept in biological research has been pragmatically recognized by the creation and proliferation of '-omics' disciplines (proteomics, metabolomics, glycomics, and so forth). Although comprehensiveness is one of its hallmarks, genomics is not all of large-scale biology, much less all of biology. And it is also important to recognize that large-scale approaches and projects are by no means the complete future of biomedical research. The importance of the individual investigator has been ratified throughout the history of the biological sciences and will be key to future developments. Indeed, a key motivation for the development of genomics and other –omics disciplines has been to relieve the individual investigator from the having to create their own large data resources and infrastructure, thereby empowering them to pursue more novel and creative research. It thus remains critically important that the data emanating from large-scale genomics projects be accessible to all investigators (see below), and that the tools developed for analyzing genomic data be hardened and made broadly available.

Acquiring Foundational Knowledge of the Human Genome

Since the end of the Human Genome Project, numerous small- and large-scale studies have profoundly advanced our understanding of the human genome and how it functions. Yet many complexities of the human genome remain insufficiently understood, hindering the application of genomics for developing new approaches for disease prevention and management. Therefore, a central focus of genomics must continue to be acquiring fundamental knowledge of the operational intricacies of the human and other genomes. As has been the rule in genomics, this will require the development of new technologies.

Advancing DNA sequencing technologies

Perhaps the most striking advance in genomics since 2003 has been the development of new (often referred to a 'next-generation') DNA sequencing methods ( ). As these have been introduced, two important things have been noted.

First, the cost of genome sequencing has plummeted at a breathtaking rate, and the collective amount of data being produced has exploded, leading to a mushrooming of the opportunities for pursuing studies to understand genome function and to define the genetic basis for health and disease. Further improvements in DNA sequencing technologies will enable personal genome sequencing, bringing new knowledge to individual patients and eventually empowering clinical decision making. But substantial technology advances are still needed to reach that goal. In spite of the remarkable five to six orders of magnitude cost reductions that have already been realized, sequencing a human genome remains much too expensive for most human disease studies and certainly for routine clinical use.

Second, the sequencing accuracy of these new technologies is not yet sufficient for their deployment in a clinical setting. Even in the case of well-understood coding regions (exons), sequencing errors complicate downstream analyses. Meanwhile, the remaining >98% of the genome harbors many (perhaps the majority of) disease-related variants, and our current lack of understanding about this larger portion of the genome makes the current sequencing error rates less tolerable. With further advances will come higher sequencing accuracies; eventually, this will decrease the amount of data redundancy needed for sequencing a human genome, which in turn should lower the currently staggering informatics and data-storage burden.

Some regions of the human genome remain inherently difficult to sequence, even with next-generation DNA sequencing technologies, and 'finishing' to high accuracy any specified genomic region rapidly, accurately, and cost-effectively is not yet routine. Further technological improvements (e.g., much longer read lengths) are needed to allow difficult genomic regions to be readily sequenced and any region to be efficiently finished. Only when we have the ability to sequence genomic DNA at very high accuracy, completeness, and throughput will genome sequencing become a 'clinical grade' activity.

Finally, one could envision clinical applications that would benefit from complete human genome sequencing in hours (or less) rather than weeks or from microbiome

sequencing in under an hour (see Signature Project ?). While speed may be less important for research applications, such rapid 'read outs' in the clinic would be empowering. More widespread knowledge of (and agreement on) the distinguishing technological requirements for the research and clinical genome-analysis markets would accelerate these developments. Similarly, as genome analysis permeates medical practice, it will be imperative to be able to deploy these technologies in locations with minimal infrastructure (e.g., power, computer hardware, and personnel).

The new technologies have also allowed an unprecedented use of DNA sequencing as a tool for detecting and 'counting' molecules and genetic variants in cells. In turn, this has expanded the user community of DNA sequencing instruments and the applications for which DNA sequencing is now the technology of choice. Further empowerment of additional researchers with new applications for inexpensive DNA sequencing should be anticipated and embraced.

Identification of genes underlying specific traits

The most prevalent strategy used in contemporary biological/biomedical research to understand the processes that underlie a trait of interest (including the pathophysiology of disease) is to identify the genes in which genetic variation influences the trait. A deep and broad catalog of human genomic variation that represents the world's populations is, therefore, an essential resource for pursuing such studies in humans, initially as a source of genetic markers for association studies and then as a source of variants for identifying and characterizing candidate genes within associated regions. The development of a catalog of the many types of genetic variation (e.g., SNPs, structural variants, and others) present in the human genome is under way with the HapMap ( ) and, more recently, the 1000 Genomes ( ) projects. The latter has expanded from its original goals, and is now slated to study 2500 individuals from 27 populations around the world. Further analyses will be needed to determine whether this is sufficient to characterize all major human populations or whether additional effort is needed to provide a long-term resource for studying the genetics and population genomics of human disease.

There would also be enormous benefit to cataloging genomic variation in certain non-human species. Such information would have several important uses, such as providing a better understanding of the basic principles of population genomics; it could also potentially lead to new clinical insights and applications. For example, understanding genetic variation in insect disease vectors may help inform the development of new interventions, while knowledge of variation among microbial pathogens may lead to more robust vaccine-design strategies.

In the case of human genetic disease studies, the improving knowledge of genomic variation has accelerated the pace of gene discovery. The last two decades have seen the genetic basis determined for more than 3000 Mendelian (monogenic) diseases and more than 750 genomic loci associated with common (multigenic) diseases, with both lists growing rapidly. These discoveries have implicated new genes and pathways in disease, revealed unexpected genetic connections among diseases, and highlighted the important role of non-coding variants in human disease.

However, for most of the common diseases studied thus far, only a modest fraction of the genetic risk has been accounted for because only a subset of the risk-conferring loci have been identified and, at identified loci, only a subset of the risk-conferring variants have been determined. A number of ideas have been offered to explain how and why we have come to this situation ( ). Most studies have not been powerful enough to detect all variants contributing to a disease; to date, these studies have largely only surveyed the most common genetic variants [e.g., those with minor allele frequencies (MAF) >5-10%; the HapMap project focused on variants with MAF ≥5%, and its successor, the 1000 Genomes project, is currently collecting variants with MAFs down to 1% across the genome and 0.1% in coding regions]. Even for identified disease genes, the spectrum of risk-conferring allelic variants has been incompletely characterized. Furthermore, efforts to discover new genes have been confined largely to a single ethnic group (Europeans), and sample sizes have often been inadequate to achieve genome-wide significance in genetic association studies. Fully characterizing the genetics of common diseases will require identifying the full allelic spectrum [including low-frequency common variants (0.5%-5%), rare genetic variants (<0.5%), and de novo mutations] in sample sets that are diverse and of a size adequate for the task.

To address these limitations, a number of new technological advances are required. These include: (1) DNA sequencing technologies that yield completely assembled human genome sequences, so that all types of genomic variation can be identified; (2) robust tools to evaluate comprehensively the complete catalog of detected genomic variants and to integrate that information with data from genome-wide association studies; (3) routine methods for detecting rare genetic variants in large sample sets, both intensively in genomic regions of interest and extensively across the genome; and (4) analytical approaches for analyzing all available variation data to reliably infer disease-causing loci, genotype-phenotype correlations, interactions among genes and with environmental exposures, and valid clinical predictions.

Understanding the biological basis of diseases and other traits

The outcome of gene identification studies, particularly those involving common diseases and other complex traits, is usually not one or a set of clearly identified causal variants. Linkage disequilibrium limits the resolution in genetic association studies, so that specific genes or variants generally cannot be deemed causal based on the statistical associations alone. And even sequencing studies are often not definitive, generating sets of candidate genes and variants that must then be further analyzed in order to establish a causal relationship. Beyond sequencing efforts, functional analyses are also essential for establishing genotype-phenotype relationships and for understanding the underlying disease biology.

Identifying functional genomic elements. The ENCylopedia Of DNA Elements (ENCODE) project ( ) and, later, the related model organism ENCODE (modENCODE) project ( ) have aimed to develop and use high-throughput approaches for developing a catalog of functional elements in the human and the *C. elegans* and *D. melanogaster* genomes, respectively. These two projects nicely illustrate the status of — and ongoing challenges associated with — current and future efforts that aim to enhance our fundamental knowledge of genome function. For both projects (referred to as

mod/ENCODE for short), existing methods have been adequate for generating useful 'version 1.0' catalogs of functional genomic elements. Of note, the new counting applications of next-generation DNA sequencing technologies (see above) have significantly increased the accuracy, resolution, and output of many of the assays used in the mod/ENCODE projects, such as for analyzing chromatin states (ChIP-Seq) ( ) and RNA (RNA-Seq) ( ). However, only a small fraction of the cell states (e.g., cell types and developmental stages) needed for assembling complete catalogs of functional elements are being assayed by mod/ENCODE. The current technologies are not sufficiently robust, nor are all of the required reagents available, for truly achieving a comprehensive catalog. Revolutionary technological advances (e.g., that increase data production by several orders of magnitude), complete reagent sets (e.g., tagging factors and affinity reagents), and more robust computational tools are needed for fully exploring and understanding functional elements across genomes.

Another limiting factor for mod/ENCODE and related projects is the large number of cells needed for many assays. Because of this requirement, many experiments are performed using either tissue culture cells (which may not accurately reflect in vivo states) or heterogeneous tissue samples (in which sub-tissue-specific patterns may go undetected). In order to determine cell-specific profiles, analytical methods for assaying single cells (or at least very small numbers of cells) are needed, most likely requiring completely novel approaches.

Finally, another opportunity for new insights will come from analyzing genetic variation within functional genomic elements. Integration of genetic variation information with the catalog(s) of functional elements will enable better understanding of the functional consequences of individual nucleotide differences in both coding and non-coding regions of the genome. For example, the National Institutes of Health Common Fund GTEx (Genotype-Tissue Expression) project ( ) aims to map the location in the human genome of sequence variants that quantitatively affect gene expression (eQTLs). Another example of a valuable experimental approach in this area would be systematic studies analyzing how nucleotide changes affect the function of enhancers and other cis-acting regulatory elements in controlling gene expression.

Understanding gene networks and regulatory pathways. In the cell, genes and gene products do not act in isolation, but rather in networked interactions with other (often many) genes and/or gene products. Therefore, understanding the cellular function of a gene product must involve characterizing its role in the networks in which it participates. Accordingly, the analysis of gene interaction networks is a major activity of ongoing biological research and is highly relevant for acquiring a comprehensive understanding of disease phenotypes. The important work in this area is by no means the sole purview of genomics (or any other scientific discipline), and investigator-initiated research, inspired by a range of ideas that cannot practically be articulated here, remains key to advancing the field. However, this is another area in which genomics can make important contributions by generating additional technologies and resources, by stimulating new fruitful areas of inquiry, and by facilitating studies that yield information about specific disease pathways.

Among the needed tools and methods relevant to genomics are high-throughput techniques for (1) genomic manipulation of cell lines and model organisms to enable

better dissection of regulatory element 'grammar'; (2) deciphering the ways in which genes act together in networks to produce phenotypes; (3) studying the dynamics of gene expression, which will lead into the overlap area between genomics and proteomics; and (4) understanding protein localization, modification, and association (protein-protein and protein-nucleic acid). The resources needed include collections of data from high-throughput or systematic studies of gene expression, including cellular responses to genetic changes, external perturbations, and disease, as in the National Institutes of Health Common Fund LINCS (Library of Integrated Network-Based Cellular Signatures) project ( ).

Robust computational tools are needed in this area (as in many other areas of genomics; see Text Box #) for accessing and analyzing large, complex data sets, for generating integrated views of data and analyses, and for developing predictive models. Informatics resources are also needed for maintaining a 'knowledge base' of networks; to be fully successful, this will require changes in scientific practice to ensure regular updates from investigator-initiated studies.

While it is important to approach the genomics of gene networks from the perspective of human disease, studies involving model organisms (e.g., ongoing projects to generate catalogs of variant functional elements and collections of knock-out alleles for all genes), informed by experimental manipulations and evolutionary comparisons, will continue to play a critical role in demonstrating the general principles of gene regulatory pathways. Examples of potential future studies in this area include: (1) Identifying all the genes underlying well-delineated embryonic developmental processes and locating their place in a network. Such information could then be used for comparative studies to understand how the network and its components have evolved and to enhance existing network models; and (2) Beginning with a human disease where there are multiple genes known to have an effect on the phenotype and where the allelic spectrum is known, establish a model system (e.g., model organism or induced pluripotent stem cell) in which the network components can be manipulated and studied in detail to understand and computationally model network architecture and function. Such a study would allow one to relate results to the naturally occurring variants using computational models that predict effects on the network and phenotype.

Pursuing larger-scale studies. In sum, advances in DNA sequencing and genotyping technologies over the past decade, in conjunction with the development of public resources such as the HapMap and 1000 Genomes data sets, have led to the blossoming of genome-wide association studies (genome.gov/26525384) and other genomic approaches that have resulted in the localization of the genetic determinants that contribute to several hundred diseases (both Mendelian and complex) as well as other phenotypes. Meanwhile, the widespread adoption of next-generation DNA sequencing methodologies has put us on the brink of the ability to conduct comprehensive genome-sequencing studies of large cohorts to more robustly define relationships between of genomic variants and clinical phenotypes. As genomic approaches to disease studies become more practical and cost-effective at increased scale, genome analyses should be carried out in both existing and new population-based studies of diseases to integrate DNA sequence with other data types (e.g., RNA expression, methylation, microbiome, environmental factors, and clinical information).

A large-scale prospective study representative of the entire U.S. population that examines all of these factors in broadly consented participants would provide an entirely new level of data about the causes of common diseases (see Signature Project ?). The results would almost certainly reveal new insights about strategies for disease prevention and treatment, and would enhance our ability to predict individual disease risks (see below). The banking of DNA and other biospecimens from such studies, coupled with providing broad access to researchers, would be important for follow-up studies, as accumulating many different types of data on well-defined sample sets make them even more valuable resources. Of course, elucidating the genomic contributions to complex genetic diseases will also require the development of statistical methods for analyzing the massive quantities of data and integrating multiple data types.

Establishing relationships between genotype and phenotype

To date, genomic approaches for studying human diseases and other traits have focused on molecular analyses and the identification of relevant genetic variants. However, the full understanding of disease biology requires establishing the often-complex relationship between causal genomic factors and the phenotypes that they determine. Phenotypic observations are notoriously difficult to compare from study to study. To capitalize on the new opportunities offered by genomics, the scientific community has been simultaneously, and increasingly, focused on improving and standardizing phenotypic measurements, correlating model organism phenotypes with human ones, and developing and implementing methods for large-scale phenotypic analyses.

This last activity is of particular importance as, increasingly, genomic variants are being appreciated within the context of pleiotropy and cellular pathways. In human and population-based studies, large-scale phenotyping efforts are underway in extant studies by standardizing phenotypes across studies (e.g., meta-analyses in genome-wide association studies), adding additional standardized phenotypic measures to the study of existing or new cohorts, and extracting phenotype data from novel sources such as electronic health/medical records. The idea that a single genotype-phenotype association does not occur in isolation, but rather within a rich context of multiple genotype-phenotype associations, is starting to be explored in study designs such as the phenome-wide association study (PheWAS) ( ). Here, individual SNPs from genome-wide association studies are being studied relative to a wide range of phenotypes in an effort to discover novel associations. Such large-scale phenotyping efforts should be expanded to encompass traits and diseases of clinical and public health significance, including drug-response phenotypes.

Similar studies are being pursued with model organisms, as their use offers researchers the opportunity to explore environmental effects (e.g., diet) in a very controlled manner. Many clinical tools (from magnetic resonance imaging to clinical chemistry) have been adapted for use with mice, allowing detection of phenotypes that map easily to human conditions. One potentially informative approach for defining gene function in vivo and understanding correlations between variants and phenotypes involves the development and study of comprehensive collections of null alleles, which are being assembled for several model organisms. The International Mouse Knockout Consortium will soon complete a collection of strains with each of ~17,000 genes knocked out ( ); a similar

effort is being piloted for zebrafish ( ). A systematic effort to examine these animals for phenotypes using a comprehensive and standardized set of assays will yield an encyclopedia-like description of gene function ( ). Pilot studies have already confirmed the previously mentioned observation that a large fraction of genetic variants have pleiotropic effects ( ). Once again, a significant and critical challenge is the development of a computational infrastructure to capture, store, analyze, and integrate phenotype data with the underlying genetic information, both for loss-of-function variants and natural genetic variation. Undoubtedly these and other efforts will identify new candidate genes for human diseases and provide experimental models for more detailed analyses. Furthermore, merging phenotype annotations with knowledge about protein interactions and gene pathways will identify targets for possible therapeutic intervention in human diseases.

Establishing the role of microbes in human health and disease

Microbes co-exist in and on us as part of the ecosystem that is the human body. In some cases, this involves benign resource sharing; for example, there are microbes that derive energy from carbon sources in our diet that humans cannot use. The clinical relevance of microbes is both obvious (infectious diseases) and relatively unexplored (the maintenance of human health). Catalyzed by advances in DNA sequencing technologies and computational biology, there has been a substantial growth of studies aiming to characterize the human microbiome— the bacteria, viruses, and eukaryotic microbes that inhabit the human body.

As part of multiple ongoing initiatives [e.g., the National Institutes of Health Common Fund Human Microbiome Project ( )], investigators are cataloging the microorganisms present at key body sites, and demonstrating relationships between the microbiome and specific human diseases. Such studies are complex with respect to recruitment of human subjects, privacy issues, sample collection, sequencing strategies, genomic and metagenomic data analyses, and data deposition and access. Early results suggest that individuals share almost half of the thousands of bacterial species that reside within our microbiomes, yet they also have unique communities established by their maternal microbiotic inocula and regulated by diet, lifestyle, and other factors. There is also evidence for relationships between discrete alleles in the human genome, specific diseases, and particular microbiome profiles.

It is also becoming increasingly clear that many diseases (e.g., acne, psoriasis, dental disease, and Crohn's disease) are associated with an out-of-balance microbiotic ecosystem. To more fully understand these relationships, studies are needed to generate more complete microbiome characterizations involving larger and more diverse clinical populations. For example, gastrointestinal microbiome studies are usually performed using stool samples, which serve as a convenient proxy for the entire gut because of the difficulty of getting more precise anatomical sampling. Meanwhile, studies correlating microbiome changes with human genotypes are in their infancy; thus, the inclusion of microbiome analyses in future genetic association studies is needed. Additional examinations of the microbiome in animal models, particularly the mouse, are needed, as are tools for handling, analyzing, and interpreting metagenomic sequence data. Fortunately, human microbiome studies are benefiting from collaborations with the environmental metagenomics field, which has extensive

experience in evolutionary biology, microbial ecology, systems biology, and biogeochemistry, all of which are used to understand the functioning of environmental microbiomes.

In its next phase, human microbiome research must move beyond correlative observations to a functional understanding of the microbiome's role in health and disease (see Signature Project ?). New strategies and tools for studying microbiome function (e.g., transcriptomics, proteomics, and metabolomics) are needed. Data demonstrating definitive roles for the microbiome in disease causation, progression, and prevention should lead to the development of new diagnostics and perhaps therapeutics. One can imagine a future in which a patient receives not only an interpretation of their genome sequence, but a microbiome profile and information about its significance as a part of a comprehensive diagnostic work-up (see Signature Project ?).

Human participants in genomics research

To ensure that genomics research findings are broadly shared in an ethical manner, new mechanisms must be developed to enhance the oversight system for protecting human research participants. Processes must be streamlined and perhaps centralized to ensure rigor, consistency, and rationality in the handling of requests for access to human subjects data, and to make it less cumbersome for legitimate investigators to retrieve information from controlled access databases, such as dbGaP ( ). Correspondingly, stricter controls must be implemented to minimize the risk of security breaches by 'rogue scientists' or other unauthorized persons.

New approaches are needed to reconcile the inherent tension between the Office of Human Research Projection's (OHRP's) current interpretation of the term 'human subjects research' (now defined to exclude research on de-identified samples) and the perceived sense of duty by many scientists, Institutional Review Board (IRB) members, and bioethicists to protect the privacy and related interests of participants in genomics studies. The efficacy of existing models of informed consent must be reassessed in light of the trend toward ever-broader data sharing and the growing recognition that genomic data is in some sense inherently identifiable, even if no names are attached. New processes must be developed for obtaining consent from individuals for sharing their data more broadly, thereby maximizing the utility of that data— while maximizing participant comprehension of the potential risks and benefits. New, culturally sensitive community engagement approaches are also needed to promote voluntary and informed research participation by people from diverse populations.

With continued efforts to ensure the routine release of all data associated with genomics research studies (see below), particular attention should be paid to information related to individual research participants, with the establishment of appropriate policies as needed. All such policies must be based on sound empirical data, so research will be needed to systematically assess the impact and effectiveness of data release policies from the standpoint of the research participants as well as the submitting investigators, requesting investigators, and staff responsible for policy implementation. Research is also needed for capturing the early experiences of the participants in various large genomics research initiatives, such as the Personal Genome Project ( ), the Coriell

Personalized Medicine Collaborative ( ), the Marshfield Clinic Biobank ( ), and the U.K. Biobank ( ).

Data release

One of the distinguishing innovations of genomics is the practice of rapid data release, often prior to publication. The release and distribution of data sets reflecting large, broadly useful resources has been a primary objective of many genomic initiatives since the inception of the Human Genome Project. The field has taken an increasingly aggressive approach to data release over the years, with the first meetings primarily dedicated to establishing policies for the release of genomic data being held in 1996 and 1997 ( ). As new issues affecting deposition into publicly available, centralized repositories arose, subsequent meetings were held. In 2003 ( ), the questions of how to deal with the inherent delay attendant upon data analysis (e.g., what are appropriate practice for releasing assemblies of genome sequences) and how to ensure credit for investigators who share their data while, at the same time, rapidly making that data available for broad use. In 2009 ( ), the issue of rapid release of data derived from human subjects was discussed, as was the importance of extending the practice of prepublication data release to other areas of biomedical and biological science.

To facilitate the widespread use of prepublication data while, at the same time, providing appropriate credit to the data producers, the attendees at the 2003 meeting ( ) developed the concept of a 'marker paper' (or 'project description') that provides both a citable source for the planned project as well as details about the project's aims, plans for data release, timetable for publication(s), contact information, and location(s) of available data. Although a handful of marker papers describing large genomics projects have been published to date, these have often not been published until a year or more after the project actually started, and there are many large projects for which a marker paper has never been written. Thus, more recently, the concept was updated by proposing that funding agencies require, or even prepare themselves, a 'project summary' to be deposited in a citable archive, such as *Nature Precedings*, at the start of a project; such an approach would likely avoid some of the problems that have been encountered in trying to publish marker papers in peer-reviewed journals.

With respect to data release itself, there is a continuing need for standardizing and streamlining the approaches for data release at the prepublication stage; some of this can be aided by journal editors. Approaches to international harmonization of data release policies must also be refined. This has been accomplished by several recent international initiatives (e.g., the International Knockout Mouse Consortium, the International Cancer Genomics Consortium, and the International Human Microbiome Consortium), but well-established policies can only be established by agreement among the international funding agencies that support the generation of genomic and other data resources.

Overall, prepublication release of data from large-scale resource-generating projects is a critical aspect in realizing the promises of large-scale biomedical research, and each of the participants in the scientific enterprise— data producers, data users, journals, and funding agencies— has a critical role to play.

Using Genomic Knowledge to Improve Human Health

If conceptualized as a temporal progression from basic research to medical applications, it is fair to say that genomics has succeeded in its initial endeavors and has just started to apply its technologies and strategies to the study of human health and disease. With an increasingly stronger foundational knowledge of genome structure and function, attention is getting acutely focused on the next steps for using genomics to improve human health. While previously on the distant horizon, important successes now appear within striking distance.

The combination of an enhanced understanding of the human genome and growing knowledge about the relationship between genetic variation and physical traits and diseases offers considerable promise for improving the practice of medicine. This potential touches all of the core components of clinical care, including prevention, risk assessment, diagnosis, treatment, and access. At the same time, important, even revolutionary, research advances will not be sufficient to change medical practice in the real world and to ensure the effective integration of genomic information into routine clinical care. Rather, many other critical issues and barriers will need to be addressed by other disciplines. Examples include the development of a suitable infrastructure for genomics-oriented clinical informatics (see Text Box #), appropriate advances in healthcare provider and consumer education (see Text Box #), and effective policies for addressing the many nuances associated with the implementation of genomic medicine. Careful attention must also be paid to the emerging societal implications of these advances (see Text Box #), including the potential impact on health disparities (exacerbating or ameliorating) and the effects on quality of life and healthcare costs. While essential for realizing the contributions of genomics to health improvement, such developments are not completely (or primarily) within the scope of genomics.

In this section, we will first describe ways in which genomic advances can be used to improve health outcomes. We will then discuss some of the larger issues that extend well beyond the world of genomics, but that need to be solved in order for the possibilities offered by genomic medicine to become a reality.

Clinical utility of genotype and other genomic-based data

Reliable markers that can be readily used by physicians and healthcare workers to determine the health and disease status of patients are key to providing effective medical care. Genomic data can potentially provide very informative and valuable markers with clinical utility across a wide range of areas: prevention, such as in determining disease risk; diagnosis, such as in molecular stratification of tumors; and treatment, such as in determining therapeutic effectiveness and the potential for adverse drug response. In many cases, clinical utility will not require a complete understanding of the molecular basis of disease etiology or physiology, but will demand the development of a computational framework in which genomic, phenotypic, and environmental data (including data from electronic health/medical records) are integrated to provide the clinician and patient with the information needed to make the most effective medical decisions tailored to the individual.

Genomic information and disease prevention

Detailed genotypic data or a complete genome sequence of an individual potentially provides information that can be used to assess their risk of developing a disease, whether it is a condition that will manifest pre- or perinatally, in childhood, or later in life. While this potential has been enthusiastically heralded as a benefit of genomic research ( ), until now, it has primarily been useful in assessing the risk for rare Mendelian diseases. For common complex disease, knowledge about the genetic variants acquired to date has generally added little to standard clinical information in predicting disease ( ), and much more work is needed to identify genetic variants and mechanisms that will explain a larger proportion of heritability (see above) and to understand their contributions in the presence of non-genetic risk factors.

This is a complex research area that will require understanding the interaction of genomic factors with the numerous psychological, social, and environmental factors that influence the development and maintenance of health habits contributing to common chronic diseases ( ). The few preventive interventions for changing health habits that have produced significant long-term health improvements (e.g., smoking cessation) have only been implemented at a population level ( ). It is a key, but as-yet untested, assumption that personalized prevention recommendations tailored to an individual based on genomic information will lead to more effective risk-reduction through lifestyle modifications, cumulatively improving the health of the general population.

Evaluating such an approach to disease prevention can occur even without a comprehensive understanding of the complex interactions producing disease or other traits (e.g., behavioral traits). Along with advances in environmental assessments (e.g., personal digital assistants or wearable biosensor dosimeters), genomic data will contribute to increasingly sophisticated studies examining associations between disease and environmental factors. Combining exposure or environmental response data from either observational, natural history-type, or interventional studies with participant genotype data will allow conclusions about environmental effects to be made in the context of, for example, differential sensitivity or resistance on the part of the research participants. Conversely, the results of genetic studies that assess disease risk will be increasingly analyzed in terms of the environmental exposures and other experiences of the participants. Such complex data sets, and the analytical tools that will be needed to interpret them, should help researchers tease apart the interacting effects of genetic predispositions and environmental potentiators, thereby enabling the development of more useful individual predictive risk profiles. Such profiles will include both an inherited component, reflecting both common low-penetrance variants that act in a polygenic fashion and rarer variants that confer a higher disease risk in a monogenic fashion, and a non-inherited component.

The discovery of genomic-based markers that appear to be clinically relevant for assessing disease risk will require robust follow-up studies to evaluate their real-world utility. Large, well-characterized cohorts (see Signature Project ?) will need to be studied in order to assess, for example, the ability of genomic-based markers to predict disease onset (or to permit early detection of occult disease), and subsequently to assess whether such information makes preventive efforts more effective. However, it must be noted that widespread genomic screening of populations is likely to identify

individuals who harbor variants conferring increased risk of life-threatening conditions in whom preventive interventions may be warranted, but the rarity and diversity of such variants will greatly complicate the design and testing of such preventative strategies. Research is also needed to identify opportunities that genomic information presents for improving health outcomes and reducing healthcare costs through preventative strategies.

Genomic information and diagnosis/prognosis

Genomic information offers great potential for improving the diagnosis and subclassification of human disease ( ). The ability to efficiently catalog all genomic variants will lead to establishing the subset of variants that are clinically relevant, although our ability to do the former is currently far greater than the latter. Similarly, our ability to identify epigenetic and proteomic variation comprehensively will likely develop long before we understand the relevance of such variation to human health and disease. Nevertheless, the identification of molecular- (genetic, epigenetic, and proteomic) based subclasses of disease will have enormous clinical relevance in diagnosis because it is likely that such subclassification will reflect the underlying disease etiology. Improvements in the molecular taxonomy of human disease will permit more accurate diagnosis and prognosis by distinguishing different diseases that otherwise share pathophysiological or morphological features. The use of genomic information will similarly inform treatment decisions (see below).

As previously noted for predictive genomic-based markers, discovery of germline or somatic 'genomic signatures' that appear to be clinically relevant for diagnosing a particular disease will require robust follow-up studies that evaluate their utility in real-world clinical settings. Once again, large, well-characterized cohorts (see Signature Project ?) will be needed for studies that assess, for example, the ability of a genomic-based marker to make treatment more effective by providing more accurate diagnostic or prognostic information. The effectiveness of genomic-based classification of patients with existing disease in tailoring interventions and, ultimately, improving outcomes must be demonstrated by research efforts examining relevant and representative patient populations. Once genomically informed interventions are developed, a portfolio of comparative effectiveness research will be needed to evaluate their impact on health outcomes. Ultimately, of course, the clinical utility of genomic information must be demonstrated by a reduction in death and disability, or an improvement in the overall health of patients or even populations.

Genomic information and therapeutic development

The potential therapeutic impact of genomic information can be categorized broadly into three areas: (1) the development of new drugs based on foundational knowledge of the genome; (2) the improvement of drug development by genotypic stratification during clinical trials; and (3) the attainment of higher efficacy and lower adverse event rates in clinical practice due to genotype-guided prescribing of drugs. The latter two are conceptually similar to the use of genomic information for risk assessment, diagnosis, and prognosis, and all of the issues discussed above are relevant to them as well.

With respect to the development of new drugs, the sequence of the human genome and that of human pathogens and their vectors have provided a plethora of new targets for therapeutic development. In general, this approach requires the discovery and validation of the specific protein(s), other cellular component(s), and/or pathway(s) whose malfunction leads to disease, and benefits enormously from an understanding of how molecular malfunction leads to the disease phenotype(s). With that information, candidate therapeutic agents (e.g., small molecules, antibodies, and siRNA) can be identified by high-throughput screening methodologies or can be developed by molecular design technologies. Many of the subsequent steps in the development of an effective therapeutic agent (medicinal chemistry, pharmacokinetics, formulation, and toxicology) do not involve genomics, and thus cannot be expected to be improved by it. Nevertheless, the promise of developing new pharmaceuticals based on genomic-derived knowledge of specific targets and their role in disease has already been demonstrated [e.g., trastuzumab for Her2-positive breast cancer ( ) and imatinib for chronic myelogenous leukemia ( )]; such an approach is becoming increasingly commonplace, particularly for cancer drug development ( ).

Yet, while this approach offers the potential of therapy that is ‘personalized’ based on genotype and therefore presumably more effective than that chosen strictly on the basis of clinical or pathological features, there is a potential difficulty in that a different therapy may be required for each group of patients with a particular genotype. Recent findings in genome-wide association studies suggest that common diseases may in large part be due to rare alleles in different patients, thereby suggesting that even common diseases may need to be treated as ‘rare’ from a therapeutic development point of view. Since the time and cost of developing a drug is essentially identical regardless of the size of the population to be treated, the development of personalized treatments may prove more expensive in aggregate. Economic analyses will be needed to determine the relationship between the additional expense of developing personalized treatments and the potential cost-savings and increased economic lifespan of a more effectively treated patient population.

It is also relevant to note that some of the earliest-discovered human disease genes (e.g., for sickle cell anemia, Huntington disease, and cystic fibrosis) were identified two or more decades ago and, despite intense effort, no effective therapy based on knowledge of the primary genetic defect is yet available for any of them. The difficulty in developing BACE inhibitors for treating Alzheimer's disease despite exquisite genetic information is another striking but not atypical example ( ). In addition, most newly discovered disease genes are not in the ‘druggable’ classes typically targeted for therapeutic development, so that new insights in terms of biology and chemistry are needed for therapeutically manipulating such targets. Lastly, the sheer number (>500,000) of known proteins encoded by human and pathogen genomes, and the thousands now implicated in disease via genetic association and other studies, has generally dispersed characterization efforts and slowed progress on any given potential target.

Finally, it is important to point out that ‘therapeutics’ include not only pharmaceuticals, but also dietary, behavioral, and lifestyle interventions, modification of environmental exposures, and other population-wide or societal interventions, some of which appear to have genotype-specific effects ( ).

Genomic information and clinical trials

Clinical trials of new drugs and other treatments, preventative measures, and diagnostics can be directly improved by the use of genomic information, whether the intervention being studied was or was not developed using genomic information. Genotypic data can be used to better stratify participants, which in turn can increase the statistical power for establishing effectiveness and reducing morbidity, adverse drug reactions, and death with smaller numbers of trial participants. Most dramatically illustrated for cancer therapeutics ( ), the improved efficiency of genotype-stratified clinical trials will likely, to some extent, offset the increased effort required to develop genotype-stratified drugs since the greatest cost of drug development is in the execution of clinical trials.

Genomic information and treatment

Determining the relationship between genomic variation and treatment response will be important for identifying pathways associated with drug metabolism, action, and safety, as well as mechanisms of non-pharmacologic treatments. Correlating disease signatures (as defined by gene-expression profiles, epigenetic changes, and other data) with therapeutic response will improve the targeting of treatments to appropriate patients at appropriate stages of their illness, and the tailoring of treatment for an individual in terms of dose, timing, and monitoring routine. The study of differential responses to treatments among individuals or across time, including idiosyncratic or adverse effects, should allow the identification of additional genetic and other biological, social, behavioral, and environmental markers that may inform clinical decision making.

Validating genomic influences on treatment responses may, in some cases, require the testing of validity and utility in randomized clinical trials (as discussed above), but formal evaluation will not be possible for every variant-treatment association. More indirect lines of evidence will be needed in most cases, such as observational studies of non-randomized interventions in large groups of patients or studies of smaller numbers of more intensively analyzed individuals using detailed pre- and post-treatment phenotyping. Recognizing that individual optimization of treatment would generally be a key goal for genomic-based research of treatment response, the appropriate role for more population-based approaches to treatment selection will have to be defined, particularly in settings where individual genotyping is not practicable.

Attention must be paid to developing a suitable infrastructure for effective implementation of genomic-based individualized treatment, one that ensures rapid availability of valid, cost-effective testing and utilization of the results by appropriate decision-making tools. In addition, clinical practice guidelines will be needed that can be rapidly updated based on new evidence, avoiding undue delays caused by an unreasonably high threshold of evidence.

The impact of treatment approaches based on genomic information should be assessed by clinical studies that analyze outcome, adherence, and intervention costs. Outcome assessments should account for disease improvement/progression as well as patient well-being and quality of life. The development and implementation of effective clinical decision-support systems will be critical for ensuring rapid adoption and appropriate use

of genomic-informed treatment at the point of care. Demonstrating that genomic approaches to individualized treatment are both clinically valid and cost-effective will be critical for wide adoption.

Genomic information and clinical care

The availability of cost-effective genotyping and whole-genome sequencing and their use for preventive, diagnostic, and therapeutic interventions will require new approaches for delivering genetic information to patients and healthcare providers alike. The amount of information to be considered and the fact that data about multiple health conditions will be generated simultaneously will antiquate the current paradigm for providing personalized genetic information. New tools are needed for explaining the significance of relevant genetic variants to patients and their families. Principles need to be established to guide decisions about acquiring genomic information (including testing for rare disease-associated variants and whole-genome sequencing) from either clinical laboratories or direct-to-consumer providers; such efforts should account for economic impact, preventive potential, and therapeutic benefits. Guidance will also be needed to address incidental or unexpected findings derived from whole-genome sequencing, including the early identification of genetic risks for adult-onset disorders.

It will be important to ensure that healthcare professionals accurately interpret genomic data and appropriately convey genetic risk to their patients, including genomic information generated by direct-to-consumer services. This will require significant attention to issues related to genomic medicine literacy among healthcare providers and consumers (see Text Box # and Signature Project ?). An important component will be to determine how genomic information alters health behaviors; for this, it will be valuable to determine the impact that various factors (e.g., family history of a disease, perceived ability to modify risk, and severity of a condition) have on patients' desires to identify and reduce their risk. Ultimately, best practices for ensuring that patients have adequate information, access to appropriate healthcare services, and appropriate follow-up to effectively deal their own genetic information will have to be developed.

While one can envision a future where genomic information will help both doctors and patients make more informed clinical decisions, such a future cannot be realized if there is not a robust clinical informatics system that is readily accessible to primary care providers (see Text Box #). Existing clinical informatics architectures are incapable of storing genome sequence information in a way that can be searched, annotated, and shared across healthcare systems throughout the lifespan of an individual. Thus, new scalable approaches are needed for aggregating, categorizing, validating, and disseminating clinically meaningful genomic information. Such systems must also address public concerns regarding confidentiality, privacy, and security related to the storage and dissemination of personal genomic information.

Finally, in a climate of rising healthcare costs, demonstration of utility looms as a bottleneck to the widespread adoption of genomic medicine. Without evidence of utility, reimbursement for services will lag, and providers will be reluctant to learn about new advances and applications, effectively limiting the dissemination and implementation of genomic technologies in routine clinical practice. The thresholds for evidence of benefit and harm vary dramatically across stakeholders, and defining robust metrics for such

measurements is an important research need. Studies designed to measure patient outcomes (including morbidity and mortality or, minimally, widely accepted surrogate health markers) are critical for gaining wide acceptance. Novel approaches for aggregating large population groups using linked health informatics systems or social media websites may be harnessed for examining the role of genetic variation in patient response.

Using genomic information to reduce health disparities

Most documented causes of health disparities derive from non-genetic factors, and the field of genomics has been cautioned not to 'overemphasize genetics' as a major explanatory factor in health disparities ( ). However, continued genomic research and breakthroughs should, in fact, better inform our understanding of population differences in disease distribution and variable drug response ( ). For example, specific genetic variants along with population differences are associated with increased risk for prostate cancer ( ), kidney disease ( ), and hypertension ( )— diseases with documented prevalence disparities. Thus, while most genomic research will apply broadly to all groups of people, it is also important to study whether there are specific genetic factors that underlie disparate rates of disease incidence or severity among population groups.

In order to be broadly applicable, developments in genetics and genomics research require the involvement of individuals from diverse ancestral populations. In general, genomics research, including the large number of recent genome-wide association studies, has been heavily concentrated on populations of Northern European ancestry who suffer relatively few health disparities ( ). To gain insight into disease etiology and differential distribution of diseases across ethnic groups, it is essential to conduct appropriately designed studies that capture diverse ancestral backgrounds ( ).

To understand, reduce, and ultimately eliminate health disparities will require multidisciplinary research, including the study of social, genetic, and environmental factors that play roles in the pathophysiology of common complex diseases and the variable drug responses that contributes to individual and population differences. An understanding of the interactions among multiple genetic and environmental factors will most likely provide more accurate prediction of disease risk and treatment response. To robustly study gene and environment interactions will require new technologies and standardized collection tools to facilitate the integrated collection and analysis of standardized environmental and clinical data. Additionally, as genomics continues to be applied in healthcare settings, the identification of factors that influence potential barriers to access and utilization of these new technologies will be needed.

Increasing accessibility to genomic medicine

Genomic technologies and advances will only achieve their full potential to improve world health when they become accessible to all humanity. As one example, the current healthcare ecosystem in the U.S., a complex mélange of stakeholders (including patients, providers, insurers/payers, industry, society, and government), determines the accessibility of new medical technologies. The development of novel and effective mechanisms for engaging all relevant stakeholders is needed to maximize the relevance of genomics to medicine and human health. Further, all stakeholders must be

educated regarding genomic advances to be able to weigh their potential benefits and limitations (see below).

Existing healthcare infrastructures are poorly suited for the delivery of genomic medicine to all segments of the population. The current healthcare delivery model is hierarchical and limited in capacity; as genomic-based applications become more broadly relevant to all of healthcare, this model will become increasingly untenable. Optimal models for ensuring that 'best practices' in genomic medicine become available to all at-risk patient populations have yet to be defined. Alternatives to the current system include reliance on non-geneticist healthcare providers guided by informatics support, increased use of telemedicine, and enhanced genomics education for future generations of healthcare providers; all of these must be pursued.

Implications of using genomic information in the healthcare setting

To realize the potential benefit of genomic information in healthcare, new genomic-based health interventions must take advantage of advances in information and communication technologies. To ensure that these new interventions are safe and effective, their implications for consumers, families, communities, and society must be studied, understood, and addressed.

Before genomic information can be fully integrated into healthcare settings, a better understanding is needed about how individuals and their healthcare providers assimilate and use this information. Research is needed to establish best practices for conveying the uncertainties and complexities of genomic-based risk information to individuals and their families, how this information is understood by consumers, and how it influences health behaviors. The impact of personal genomic information on the traditional provider-patient relationship, particularly that generated by direct-to-consumer testing services, also needs to be defined.

As the cost of whole-genome sequencing decreases, it is increasingly likely that an individual's genome sequence will become a routine part of their medical record. The availability of this information, coupled with the development of universal electronic health/medical records and the growing use of social networking technologies, will pose new challenges for the healthcare system. Given the large amount of potentially relevant information contained in an individual's genome sequence, the consequences of a privacy breach could be serious. Research is needed on the appropriate balance between a broadly accessible medical record that includes genomic information and the need for privacy. The potential use of online genomic testing services and social networking sites complicates the protection of genomic information; such issues also need to be explored

The social and economic implications of developing genomic-based diagnostic and therapeutic interventions must also be studied and considered. The possibility for these new interventions to exacerbate or ameliorate current inequities in access to and use of healthcare services needs to be examined. Development of new genomic-based health technologies may divert resources from proven social environment interventions, such as smoking cessation, diet, and exercise; this could be complicated further as the

principles of comparative effectiveness are increasingly used to guide healthcare decisions.

Finally, the current regulatory structures may not be sufficient for the review and oversight of new genomic technologies, and new oversight mechanisms may be needed. Given the likely rapid pace of change associated with the implementation of genomic medicine, addressing these and other emerging issues will require sustained, yet agile, collaborative efforts by the research, regulatory, and healthcare communities.

Educating patients, clinicians, payers, and the public

As noted several times already, education at many levels is going to be key for the successful introduction of genomics into medicine and healthcare (see Text Box #). At the risk of being redundant, we felt it important to emphasize the issue of genomics education (which is actually relevant beyond medicine, as genomics has many non-medical applications as well).

As genomics moves further into routine healthcare, innovative methods are needed to educate medical professionals at the point of care. This must include providing expertise in interpreting genetic/genomic data and making evidence-based decisions. Genomic technologies and their output can be complex, and healthcare providers are universally pressed for time. Research is needed to determine how to efficiently integrate genomics education into existing educational opportunities of all healthcare providers. In addition, the educational approaches must be effective, appropriate, and culturally and linguistically relevant across diverse populations, particularly underserved communities. Clinical decision-support processes are required to integrate genomics into medicine. The challenge will be to develop models that can be implemented at the time, place, and knowledge level needed to provide effective care.

Equally important to well-informed healthcare professionals is a well-informed public that can maximize the utility of genomic medicine and other applications of genomics in society, such as in ancestry determination, agriculture, energy, and environmental remediation. Consumers must be ready to consider the implications of genomic information on themselves and their families. Many have called for a ‘genomically literate’ public, who can make clearly informed decisions. A genomically literate public, including policy makers, will also be necessary for society to successfully deal with the many aspects of life that will be touched by genomics in the coming years and decades. However, research must be conducted to determine what it means to be genomically literate, and what knowledge is necessary to make a truly informed decision when participating in genomic research, when making genomic-based clinical decisions, or when developing and assessing public policy influenced by genomics ( ). Future genomic benefits to society require novel approaches for public education to address the needs of diverse communities in achieving an acceptable level of genomic literacy. One starting point is to look at models developed and evaluated for rare genetic diseases; these likely hold valuable lessons that can be applied to common complex disorders.

To create a holistic approach for enhanced genomic education, appropriate new methods, resources, and tools must be created and evaluated both for their

effectiveness and for the quality of information translated to the healthcare professionals, patients, policy makers, and public. Innovative education research is needed to identify methods that are effective for translating genomic knowledge into practical information for use by all stakeholders in making decisions. Efforts are needed to ensure that diverse racial, ethnic, geographic, education, and socioeconomic communities are engaged throughout the development and evaluation of genomic education methods and technologies, so that all communities can appropriately benefit.

Concluding Comments

To be written

DRAFT

Universal, Cross-Cutting Areas (Text Boxes)

Bioinformatics and Computational Biology

The generation of data is no longer the rate-limiting step for most large genomics projects. Associated with all genomic advances are ever-growing demands with respect to storing, analyzing, integrating, and displaying the generated data. Increasingly, the electronic handling and computational analyses of those data represent the major bottleneck in genomics research.

Genomic studies of all kinds— from large community resource projects to individual laboratory efforts— must proactively plan for their computational needs. While the dominant computational focus in the past was largely related to DNA sequence and gene-expression data, future work will require the assimilation and analysis of many other data types, such as images, proteomes, metabolomes, pathways, electronic health/medical records, clinical phenotypes, and environmental measures.

Genomics requires a computing infrastructure that can accommodate increasingly large and complex data sets. This must include stable and accessible repositories of data and associated metadata that enable users to address many diverse biological questions. Centralized informatics resources that provide consolidated views of many data types are needed to help the scientific community interpret the myriad data sets. Computing needs will likely be met by multiple solutions, including central compute clusters and cloud computing. Finally, as individual laboratories are able to generate large amounts of genomic data, robust analysis tools that can readily be used by non-genomics researchers will be needed.

Compelling biological projects should drive the development of new computational tools, with close linkage between data production and analysis. Such efforts must engage people with diverse expertise, including computer science, statistics, mathematics, epidemiology, bioinformatics, and computational biology. Computational tool should be developed using programming 'best practices' and made freely available to the scientific community.

As greater amounts of diverse data types become available, data integration becomes a significant challenge, but one that is essential for building predictive models of genome function and phenotypes. New paradigms for information integration will be needed to keep pace with the complexity and volume of genomic data; this must include the development and use of ontologies (standard vocabularies) for data annotation.

Visualization is essential for understanding genomic data. In the past, this was largely accomplished by indexing information relative to the one-dimensional length of a genome. However, molecular phenotypes (e.g., methylation, chromatin marks, and RNA expression) differ among cells, tissues, developmental times, and individuals; other data types, such phenotypes and environmental exposures, are based on individuals, not genomes. Such complexities require new approaches for visualizing integrated genomic data.

With the increasing emphasis on translating genomic knowledge to advance human health, there will a greater need for robust medical informatics capabilities, especially in the storage and analysis of clinical data. Medical informaticians and clinicians will be needed on genomics

DRAFT

research teams to aid in the design of studies and the analyses of data. An important issue that will need continual attention is how to assimilate and provide human genomic and phenotypic data in a fashion that allows broad and convenient access for researchers, but protects research participants. Tools will be needed that allow clinicians to diagnose and treat patients based on their personal genomic data, such as DNA sequence, RNA expression, and microbiome information; the development of such tools requires research studies that relate genomic data to molecular classifications of disease and to therapeutic effectiveness.

DRAFT

Education and Training

The full realization of genomic medicine requires a society that is sufficiently educated about the fundamentals of genetics and genomics, including a basic understanding of the potential benefits and harms arising from genomic advances. This will require formal and informal programs to ensure that the educational needs of all— including the general public, healthcare professionals, and basic scientists— are met in a timely fashion.

Genomics education for the general population must occur at all levels (primary, secondary, collegiate, and adult), starting with the learner's earliest intersection with the biological sciences during K-12 education. These programs should promote public understanding and awareness of the role of genomics in human health, medicine, ancestry, forensics, and evolution; they must also include the fundamentals of mathematics, including probability and risk, required to translate personalized genomic information into informed health choices. Broadly increasing the population's appreciation of genomics and its societal implications will enhance the likelihood that wise choices will be made as genomic medicine becomes part of routine clinical care.

The effective application of genomics in healthcare systems will depend on well-trained clinical professionals that can work within multidisciplinary teams. It is critical that a broad range of care providers (e.g., nurses, nutritionists, pharmacists, dentists, physicians, physician assistants, and nurse practitioners) acquire competency in genomics in order to provide services appropriate to their scope of practice. Such skills will need to be developed and nurtured through all stages of training, including incorporation in continuing education by inclusion of genomics content in professional licensing and accrediting processes. Some specialists, such as nurses and physician assistants, have made progress in systematically incorporating genomics in their educational infrastructure over the last decade ( ); the training of physicians has lagged behind, and available evidence suggests a lack of genomics competency among most physicians ( ). Given that many health professional educators themselves are likely not comfortable with teaching genomics, educational efforts might begin with 'train the trainer' approaches that target faculty in key disciplines.

Currently, there are insufficient numbers of healthcare providers with advanced clinical genetics and genomics training (including medical geneticists, genetic counselors, and nurse geneticists) to meet patient demand. Though the most efficient model for delivery of more complex genomic-based clinical services has not been established, it seems evident that the number of trained individuals must be expanded. The emerging complexity of genomic discoveries simply requires more genomics-educated healthcare providers; for example, medical genetics and genetic counseling training programs need to be expanded in numbers and curriculum scope (to include all aspects of genomics). Increasing the numbers of such specialists will likely require a mixture of economic incentives for trainees entering the healthcare work force, as well as active encouragement for the expansion of training opportunities.

Capitalizing on past and future genomic discoveries will require a robust pipeline to create the next generation of genome investigators who will develop new technologies, resources, tools, and paradigms. Their training should encompass genomics, proteomics, and related disciplines at all career levels. Special emphasis should be given to attracting and training

individuals in less traditional disciplines, including bioinformatics/computational biology, statistics, public health, behavioral science, economics, law, and engineering. Translating basic genomic discoveries to health applications will also require the development of a well-trained cohort of clinical investigators versed in the fundamentals of genomics. This will require attention to the genomics content in professional education, including residencies, clinical and postdoctoral fellowships, and graduate programs. Success in these training endeavors should be measured in terms of the quality and timeliness in developing independent investigators.

The broader implications of implementing genomic medicine will raise many questions for scientists, clinicians, and consumers that will require the expertise of those trained in economics, public health, behavioral sciences, and law. In order for the general population to understand and participate in genomic-based healthcare, there must be an intense effort to improve general science literacy of the population, starting at the K-12 level.

DRAFT

Societal Implications of Genomics

While all areas of science have implications for society at large, the societal implications of genomics are especially far-reaching. This has been recognized since the inception of the Human Genome Project, when the U.S. Congress allocated 3% of the National Institutes of Health genomics budgets (an amount subsequently increased to 5%) to support research related to these implications. Thus was born the Ethical, Legal, and Social Implications (ELSI) Program at NHGRI— a program that has spurred the development of related programs in other countries and in other U.S. agencies.

It is now time to anticipate and address the societal implications of genomics in an even more expansive, forward-looking way. The impact of the growing tsunami of genomic information will soon extend well beyond the healthcare arena and touch virtually every area of society, perhaps even calling into question such fundamental issues as what it means to be 'human.' As this occurs, individuals and institutions will face unprecedented challenges to absorb the meaning of the new genomic information and to handle it responsibly.

Some issues in this area relate to the conduct of genomics research itself. For example, what will be the implications of continuing advances in genomic science for concepts of privacy and individual identifiability? What are the implications for informed consent and research governance, especially when genomic data are being shared broadly? When and how should individual genomic research results or incidental findings be returned to research participants?

Other issues will arise from the way genomic information is incorporated into healthcare settings. For example, what will be the implications of integrating individualized genomic-based diagnostic, prevention, and treatment strategies into a rapidly changing healthcare system? How can genomic information best be incorporated into healthcare delivery under the emerging framework of comparative effectiveness? How will healthcare providers and patients understand and interpret genomic information, particularly when obtained from direct-to-consumer genetic services? What regulatory or other oversight reforms are needed in these areas?

Still other issues relate to the way genomic information will be used in non-healthcare settings and to how it will be conceptualized even more broadly. For example, what will be the implications of advances in forensic DNA analysis for law enforcement and for society more broadly? What are the implications of genomic ancestry testing for individuals' understanding of their own identity, of race and ethnicity, and of relationships among human populations? Will increased dissemination of genomic information (especially in areas such as epigenomics and microbiome studies) lead to new conceptions of the 'self' and of relationships between humans and non-humans?

Genomics as a field must renew and strengthen its commitment to support the systematic and rigorous study of these and many other societal implications that the new genomic knowledge is creating. Multidisciplinary research groups that include both genomicists and researchers in social and behavioral sciences, health policy, law, bioethics, and humanities must be fostered. Policy development and regulatory reform must be informed by this research, as should educational efforts targeted at a range of relevant stakeholders.

Candidate Signature Project 1

The Genomic Answer on Cancer

The explosion of large-scale technologies provides the ability to implement a multi-faceted genomic assault on understanding cancer. The compendia of genomic changes in tumors, such as those being generated by The Cancer Genome Atlas (TCGA) and the International Cancer Genomics Consortium (ICGC), are only a beginning. Emerging technologies and data resources will also provide a foundation for new explorations of many fundamental processes in cancer, which will lead to new opportunities to translate genomic discoveries to more personalized and effective approaches to cancer treatment.

Objectives:

- Continue extensive cancer genome sequencing to achieve the power and breadth necessary to develop a comprehensive atlas of mutations in all major cancer types.
- Develop a comprehensive molecular taxonomy to distinguish all sub-types of all major cancer types.
- Develop the resources to pursue genome analyses of rare cancers.
- Use DNA sequencing-based approaches to characterize the genomics of important cancer biology phenomena, such as metastasis, avoidance of immune surveillance, field effects in adjacent tissue, and treatment resistance.
- Develop the computational tools necessary to analyze the full spectrum of diverse genomic data types (e.g., single-nucleotide mutations, genome rearrangements, gene expression and epigenomic alterations, and ChIP-seq results) and to correlate genomic status with cancer-specific phenotypes.
- Develop genotype-specific therapies for 10 important cancers.

DRAFT

Candidate Signature Project 2

Choreography of Gene Regulation

Gene expression is a key determinant of an organism's phenotype, and is tightly regulated yet highly flexible because of complex interactions among proteins (especially transcription factors) and genomic elements (residing both adjacent and distal to genes). Patterns of gene expression are further influenced by genetic variation and the environment. It is not yet possible to predict from genome sequence alone when and where genes are turned on/off throughout an organism's lifetime and how environmental effects influence gene expression. Achieving such a capability will require analysis of larger, more diverse data sets than are currently available.

Objectives:

- Expand the generation of large-scale data sets of 'molecular phenotypes' (e.g., gene expression, chromatin states, transcription factor-binding sites) by obtaining data from many individuals, tissues, disease states, and environmental conditions.
- Develop analytical methods to address questions about how genetic variation and environmental factors influence gene expression.

DRAFT

Candidate Signature Project 3

Evolutionary Origins of the Human Genome

When did each gene and functional genomic element appear during the evolution of the human genome? Comparative genome analyses now focus on finding conserved and/or rapidly changing functional genomic elements. With greater capabilities for DNA sequencing, it should be possible to generate genomic sequences from a sufficient number of organisms to determine the evolutionary history of all genes and functional elements in the human genome.

Objectives:

- Sequence thousands of vertebrate genomes, selected on the basis of phylogenetic relationships, to increase the resolution of detecting conserved elements to as few bases as possible.
- Determine when each gene/functional element first appeared, and make correlations with the appearance of other genomic features.
- Create a description of the human genome as an evolutionary descendent of an ancestral vertebrate genome plus a set of inferences for all subsequent changes in vertebrate phylogeny.

DRAFT

Candidate Signature Project 4

Genomic Origins of Phenotype

A key question for contemporary genetics is how does the genome of an organism determine specific biological properties— i.e., what is the relationship between genotype and phenotype? The increasing capacity for large-scale DNA sequencing offers an opportunity to address this question by allowing extensive correlations to be made between genome sequences and specific phenotypic properties among organisms in a fashion that allows the actual genes (and other genomic elements) underlying these phenotypes to be discerned. This can be pursued at different scales, ranging from those designed to investigate the genomic basis of major phenotypic innovations, such as the four-chambered heart or brain size (by studying organisms over a broad phylogenetic range), to those designed to investigate the genomic basis of phenotypes that differ within a species, such as disease risk (by studying individuals within the species that do and do not display the phenotype).

Objectives:

- Sequence the genomes of multiple organisms displaying a range of different phenotypes.
- Develop analysis tools to identify commonalities in genes and gene networks.
- Develop experimental tools for validating function.

DRAFT

Candidate Signature Project 5

1,000 Diseases Project

There are several thousand Mendelian (monogenic) diseases for which no causative gene has been identified. These disorders collectively affect many thousands of patients for whom improved diagnosis and management could result from gene identification and testing. Furthermore, lessons learned from these relatively distinct and individually rare disorders can be generalized for genetic disorders that are more common.

Objectives:

- Curate the clinical literature and identify 1,000 candidate genetic diseases that represent a spectrum of phenotypes and genetics (recessive, dominant, sporadic, and mosaic).
- Standardize and parallelize informed consent and sample acquisition.
- Generate whole-exome or whole-genome sequences on an appropriate number of samples representing each disease, with multiple samples obtained from distinct families when possible.
- Deposit the data in a widely accessible repository that is open to all qualified investigators for study.
- Develop appropriate new software tools to interrogate, organize, and present sequence data for analyses.

DRAFT

Candidate Signature Project 6

Improving Genomic Literacy

Every individual has an opportunity and the right to be genomically literate to ensure informed decision making in the era of genomic medicine. Needed are accessible tools and resources (via various communication platforms) for understanding genomic information and its relationship to disease risk; these must be multi-lingual and culturally appropriate. Particularly as genomic medicine becomes a reality, this knowledge and improved access will empower and enable the public to make decisions that lead to healthier living.

Objectives:

- Develop a model genomics education program that is targeted toward adult genomic literacy using the public health and K-12 education infrastructure, with an emphasis on reaching both current and future consumers of genomic information.
- Include research, intervention, and evaluation components to monitor changes in knowledge levels across the general public.
- Aim to reduce disparities regarding access to and availability of relevant information.
- Stimulate related education improvements for health care professionals.

DRAFT

Candidate Signature Project 7

Genetic and Environmental Risks throughout Life

The recent success of genome-wide association studies in identifying hundreds of variants associated with a multitude of complex traits demonstrates the potential value of densely characterizing individual genomic variants and relating them to precise phenotypes. Such studies are often constrained, however, by the selected nature of cases and controls, the limited assessment of phenotypes and environmental exposures, and the potential biases in focusing on prevalent disease and relying on participant recall. Large-scale prospective cohort studies avoid many of these problems by assessing incident disease, defining environmental exposures prior to disease onset, and characterizing the evolution of traits over time. A large-scale prospective U.S. cohort study would allow robust studies of genetic and environmental risks throughout life.

Objectives:

- Implement a cohort study that aims to enroll hundreds of thousands of participants that mirrors the U.S. population as a whole with respect to age, ethnicity, geographic ancestry, socioeconomic status, environmental exposures, and urban/rural residence.
- Conduct cost-effective, targeted measures of risk factors, phenotypic traits, and environmental exposures within a 60-90 minute assessment.
- Collect high-quality biospecimens for characterization of genomic variation and environmental exposures using state-of-the-art technologies.
- Facilitate initial assessment and follow-up for disease events through existing electronic health/medical records systems (where feasible).
- Provide for the return of findings to participants and their clinicians (to the degree desired by participants), and for sharing of individual research data with the broad scientific community while maintaining participants' privacy and adhering to their consent specifications.
- Identify genetic and environmental risk factors for 20 complex diseases having the highest morbidity, mortality, and detrimental impact on the quality of life.

DRAFT

Candidate Signature Project 8

Using Genomic Information to Identify Individuals at Very High Disease Risk

Despite the recent identification of hundreds of variants associated with numerous complex diseases and traits, the individual predictive value of these findings beyond standard clinical risk factors is limited. There exists a small subgroup of individuals who carry variants that confer very high risk for a given disease. Although for any specific disease, this number is unlikely to justify population-wide screening, considering multiple conditions simultaneously would identify a much larger number of persons at high risk. Across 40 independent diseases, roughly 90% of the population will be in the top 5% of genetic risk for at least one disease, 33% will be in the top 1%, and 4% will be in the top 0.1%. A clinical study of individuals with genome-wide genotyping data and clinical outcome information could provide a valuable evidence base to learn about risk reduction counseling in response to genetic information.

Objectives:

- Screen a diverse sample of 200,000 individuals with adequate informed consent, family history, and clinical outcome information for genetic variants currently known to be associated with increased disease risk to identify 8,000 persons in the top 0.1% of risk for at least one disease and 66,000 in the top 1%.
- Survey those 200,000 individuals before and after receiving genomic information for personal health behaviors and lifestyle factors, health status, anxiety, and quality of life indicators, and provide broad-based counseling about their results.
- Offer appropriate preventive interventions, probably limited to implementing current prevention guidelines for persons at high risk, where available.
- Assess short-term adherence to and impact of these preventive strategies on non-genetic risk indicators (such as known risk factors) in a subset of participants.
- Educate participating clinicians about the evolving applications of genomics and survey them to assess the impact of genomic information on patient care.

DRAFT

Candidate Signature Project 9

ClinBase

Thousands of genetic variants associated with disease risk and treatment response are already known, yet a robust path for their clinical utility remains poorly established. Clinical bioinformatic tools are needed for capturing and interpreting an increasingly diverse array of genomic variation data (sequence, copy number, structural, expression, epigenetic, and others). Such tools must dynamically interpret information as knowledge evolves, with some activated by physician (e.g., when a drug is prescribed or a test is ordered). Developments in this area would support a vision of the future when significant genomic information about an individual is acquired early in life, and then interpreted and reinterpreted over time as new scientific data and opportunities for improved clinical care arise.

Objectives:

- Develop methodologies for semi-automated extraction and analysis of clinical annotations from existing and evolving databases.
- Design pilot approaches for integrating available genomic variation data into patients' medical records.
- Assess provider needs and preferences for the presentation and organization of genomic variation data.
- Assess patient needs and preferences regarding integrating genomic variation data into their medical records
- Develop mechanisms to maximize patient confidentiality and privacy with respect to genomic variation data.

DRAFT

Candidate Signature Project 10

Health, Disease, and the Human Microbiome

The availability of new genomic technologies, especially for large-scale DNA sequencing, has created the opportunity to make progress toward understanding the role of the complex and intimate relationship between the human host and its resident microbiome in human development, health, and disease, including the selective pressures that drive evolution of the microbiome. To realize this potential, a more complete catalog of the human microbiome gene set, including the genes from plasmids, mobile genetic elements, bacteriophage, and viruses, is needed; such a compendium would facilitate investigations about the cause-and-effect relationships between the composition of the human microbiome and human biology. In the long run, understanding the interactions of the microbiome with the human host will be necessary to develop personalized microbial therapies.

Objectives: *TBD*

DRAFT

Candidate Signature Project 11

The Patient Genome Diagnostic: 30 minutes, $100

In the future, genomic information may be used as routinely as blood chemistry data are utilized today. The optimal 'genome diagnostic' should not require sending the sample to a highly specialized facility that takes days to generate results, but rather should be performed in the healthcare delivery setting with results returned rapidly. Such a diagnostic should include analyses of a patient's genomic DNA, transcriptome, and microbiome.

Objectives: *TBD*

DRAFT

Candidate Signature Project 12

Proteomalyzer

The comprehensive analysis of all peptides and proteins in a tissue sample and their post-translational modifications represents the next needed quantum leap to complement comprehensive nucleic acid analysis for research and clinical applications. The ideal technology would not rely on affinity reagents or labels, but instead provide electronic readout of fingerprints or signatures at sufficient resolution to match a database of validated fingerprints of known proteins.

Objectives: *TBD*

DRAFT