Genomics, proteomics, and the new paradigm in biomedical research

Joshua LaBaer, MD, PhD

This article is based on the keynote address that introduced the third biennial Asan Medical Center-Harvard Medical International Symposium "Genetics and Proteomics: Impact on Medicine and Health" that took place in Seoul, Korea, July 3–4, 2001. In his address, the author summarized exciting achievements in the field of genomics and introduced the related and emerging field of proteomics. By using industrialized high-throughput approaches, genomics and proteomics are dramatically accelerating the pace of biological research. They have started a scientific revolution whose impact will range from elucidating the structure of our chromosomes to providing powerful new tools for the study of disease; and from understanding human evolutionary history to novel applications in the medicine of the future. The author's overview highlighted the recent history of the two fields and laid the foundation for the rest of the symposium presentations. ***Genet Med* 2002:4(6, Supplement):2S–9S.**

Key Words: *genomics, proteomics, biomedical research*

The end of the beginning

The announcement of the completion of the draft sequencing of the human genome in the spring of 2001 signaled a watershed event in biology.[1,2] The information provided by the Human Genome Project (HGP), along with the powerful technologies developed in the process of completing it, has already altered dramatically the manner in which biomedical research is performed.

The HGP began in the mid-1980s at the US Department of Energy as the result of a confluence of factors.[3,4] First, from a medical perspective, there was increasing evidence for the importance of genetics in virtually all diseases. It had long been recognized that on one end of a spectrum, diseases like cystic fibrosis are almost entirely caused by a genetic factor. Simply the presence of a particular mutation in both of the alleles of a single gene leads to the disease. On the other end of that spectrum, infectious diseases, such as the acquired immunodeficiency syndrome, are predominantly caused by environmental factors. Yet even for these environmentally caused diseases, genetically determined host factors affect the individual's response to the illness. The majority of human diseases falls somewhere between these extremes. Diseases like diabetes and heart disease comprise a mix of both genetic and environmental factors, and the genetic component is often complex, involving many genes. By the mid-1980s it was clear that DNA sequence could expedite our understanding of the genetic component of disease, help us to understand its pathophysiology, and identify potential targets for treatment.

A second key factor was the development of tools that suggested the feasibility of the project. By the mid-1980s, the Sanger sequencing method,[5] along with some technical improvements in the enzymology of DNA polymerases[6] and the labeling of the nucleotides,[7] had advanced to the point at which sequencing advocates could dream of completing the entire human genome. Even with these tools, however, it was by no means obvious that this was achievable.

After all, if the genome is considered the Book of Life, it is a big book. There are more than 3 billion letters in the human genome. At the time the project was conceived, typical sequencing read lengths were in the 200–300 base range. So with a simple calculation of the number of needed reads and the amount of computing power needed to handle the data (not to mention special technical problems such as repeated sequences), it is not surprising that many argued that it would be an inappropriate use of research money to take on this project and perhaps mere folly altogether.

Nevertheless, the draft version of the human sequence has been completed far ahead of the many decades originally anticipated. More than 20 different major sequencing centers and hundreds of scientists participated in this project. In the final stages of the project, centers were sequencing 24 hours a day, 7 days a week, all across the globe. As Francis Collins put it, "The sun never set on the Human Genome Project."[8]

The HGP was not always a high-throughput sequencing project. That aspect of its operations did not begin until late in the 1990s after an early phase during which genetic maps and technologies were developed that were essential to the high-throughput sequencing that occurred at the end. This may be an important lesson for us as we move forward into the "Proteomic Era." At the beginning, much time will be spent devel-

From the Institute of Proteomics, Harvard Medical School, Boston, Massachusetts.

Joshua LaBaer, MD, PhD, Director, Institute of Proteomics, Harvard Medical School, 250 Longwood Avenue, BCMP, Boston, MA 02115.

Received: August 12, 2002.

Accepted: September 24, 2002.

DOI: 10.1097/01.GIM.0000041504.37108.75

oping new methods, new computational tools, and mapping out the projects that need to be accomplished. Moreover, the HGP has shown us that the technologies themselves are at least as powerful as the data collected. New tools, such as DNA microarrays and transcriptional profiling, serial analysis of gene expression, and haplotype mapping, will prove to revolutionize the way that important genes are identified and diseases are characterized.

Although the data are increasing rapidly, the sequencing of the human genome is not yet complete. As of this printing, more than 98.5% of the genome has been sequenced in draft form, and about 91% has reached "finished" quality (http://www.ncbi.nlm.nih.gov/genome/seq/). Several chromosomes are either completed or are near completed, including 20, 21, 22, 10, 13, 14, 19, 6, and 7.

Lessons from the genome

Although the HGP will impact all aspects of biology, some areas in particular will be directly influenced as indicated in Table 1. Not surprisingly, advances in our understanding of genome structure and function and human evolution are the two disciplines most immediately affected by the genome project. The manuscripts describing the draft sequencing dwelt in these areas at length. Moreover, there has been an unprecedented acceleration in the number of papers published about human evolution and genome structure in the last several years. The sequence is a rich information store that will be mined for years, becoming even more fruitful as additional vertebrate genomes are completed, such as the mouse, the rat, and the dog.

By providing a foundation upon which discovery can occur, the genome sequence is also impacting the daily execution of biomedical research. This is occurring in two key areas: first, genetic approaches are used to create linkages that demonstrate genes that play a role in the etiology of disease; and second, by providing a template that can be used to produce and study the gene products themselves in order to understand the biochemical mechanisms that play a direct role in the causation and treatment of disease.

Finally, on the coming horizon, the genome will have a large effect on the practice of medicine. These profound changes will also demand a careful consideration of the public policy and ethics regarding the use of genetic information.

Table 1
Areas of biology influenced by the Human Genome Project

Lessons from the genome
Genome structure and function
Human evolutionary past
Advances in biomedical research
Genetic approaches
Biochemical and functional approaches
The practice of medicine
Public policy

Genomic structure and function

Among the first lessons learned from the genome sequencing was the level of heterogeneity in the genome. The image from the publication of draft sequencing by the International Human Genome Sequencing Consortium (Fig. 1) is a complex image, but one that illustrates several of these heterogeneous features. This part of the short arm of chromosome 7 is characterized in detail as indicated by the abundance in the line marked "Coverage," which denotes "finished" sequence.[2]

Genetic variation also appears to be heterogeneous across the genome. Single-nucleotide polymorphisms, or SNPs, indicate positions in the genome where there are common single base genetic differences in the population. Peaks in this plot indicate places where SNPs are common in the genome and, as indicated, some areas contain frequent polymorphisms whereas other areas much less so.

Gene density is also heterogeneous in the genome. As many genes have not yet been identified, gene density has been estimated by four different techniques. One method for estimating the presence of genes is to look for regions of homology with the genome of the pufferfish *T. nigroviridis*, indicated with the line marked "Exofish." This animal is evolutionarily distant enough from humans that conserved regions are likely to indicate genes. A second method of gene prediction is to evaluate the frequency of expressed sequence tags (ESTs), which is a measure of RNA species with poly-A tails and thus should correlate with genes in the genome. The line marked "EST" indicates areas where ESTs with at least one intron mapped to genomic DNA. Third, the starts of genes are also marked. The line marked "Genes" indicates the beginnings of genes as predicted by at least one of two gene-predicting software applications called Genie and Ensembl, and the line marked "Known genes" indicates the starts of genes in the RefSeq database. Finally, the dinucleotide CpG occurs much less frequently in the human genome than would be predicted by the known fraction of Cs and Gs. However, areas relatively rich in this dinucleotide, called CpG islands, are found in the genome, often in

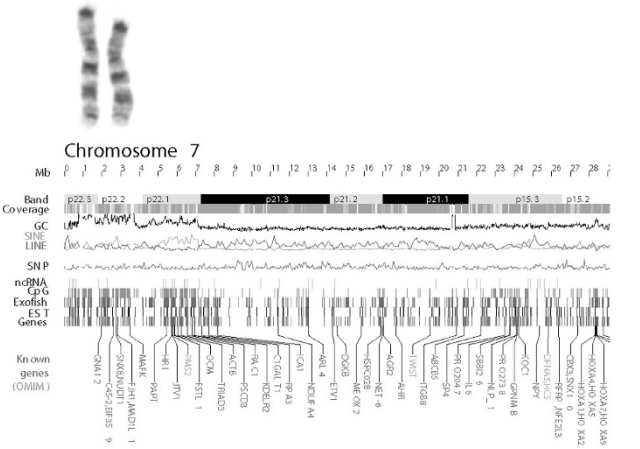

Fig. 1 Features of chromosome 7 from the draft sequence published by the International Human Genome Sequencing Consortium.[2] Reprinted by permission from Nature, copyright 2001, Macmillan Publishers Ltd.

association with the 5' ends of genes. A search of the genome for CpG islands is indicated in the figure by the line marked "CpG." There is good agreement among the four methods showing some areas, such as p22.1, to be very dense in the presence of genes, whereas other areas, such as the area directly to its right, have much fewer predicted genes. Moreover, the notion that gene density correlates with GC content is supported by this analysis of the genome .[2]

The heterogeneity of gene density is evident even at the chromosomal level. Chromosome 21 (Fig. 2), which has been fully sequenced, is not particularly rich in genes compared with a chromosome close in size, chromosome 22 (Fig. 2), which has a much higher density. Down syndrome, or trisomy 21, is one the few situations in which an extra copy of an entire chromosome in humans is compatible with life. Perhaps the reason that trisomy 21 is tolerated is that this chromosome has relatively few genes.[2]

One of the most celebrated surprises of the genome was the number of genes that were predicted from the genome.[9] Although the actual number of genes is still not known, the predicted number is much lower that originally expected (Fig. 3) .[10,11] There was some wonder (almost concern) expressed that the number of genes in an organism as complex as humans or other vertebrates was not significantly larger than that found in the genome of the worm *C. elegans*, which has only a little over 900 cells in the entire organism.

The explanation for this apparent paradox resides at the level of the gene products, not the genes, that is, an organism's complexity is determined by its proteins and their various forms and regulations. At the protein level, humans (and other vertebrates) possess a significant level of diversity. First, humans appear to have many more splice variants per gene than simpler eukaryotes. An analysis of reconstructed mRNAs for chromosomes 22 and 19 suggests that on average there are at least three transcripts per gene in humans compared with the worm, which appears to have about 1.3 splice variants per gene. Taking this into account, the number of different mRNA species in the human may exceed 90,000 compared with fewer than 30,000 in the worm.[2]

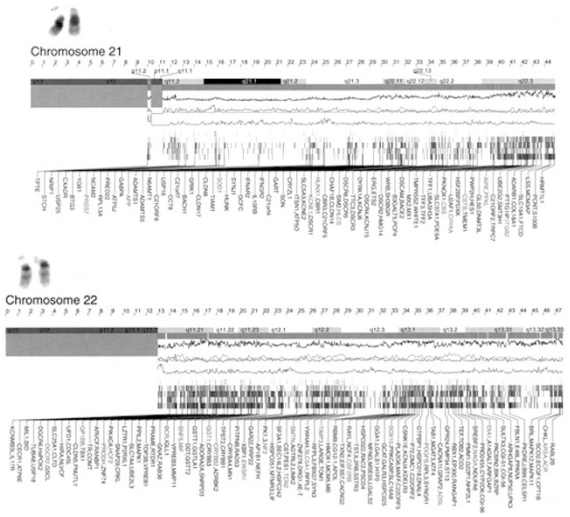

Fig. 2 Variation in gene density on different chromosomes.[2] Although similar in size, estimates for the density of genes on chromosome 21 are much lower than those for chromosome 22. This may partially explain why trisomy 21 is not lethal. Reprinted by permission from Nature, copyright 2001, Macmillan Publishers Ltd.

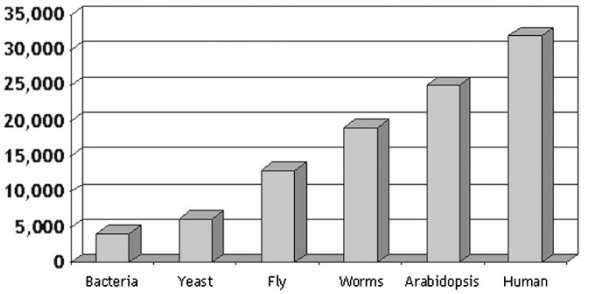

Fig. 3 The predicted number of genes in the human genome in comparison with other genomes.[10,11]

Second, proteins in humans have greater architectural complexity than their counterparts in simpler organisms. Proteins are composed of distinct structural domains, usually ascertained by sequence homology among related proteins both within and between species. These domains often impart particular biochemical functions on proteins that contain them, such as a catalytic activity or the ability to interact with a corresponding domain on another protein. In the tabulation so far (which may be an underestimate), humans have nearly twice as many distinct domain architectures as the worm and fly, and almost 6 times as many as yeast. Moreover, when examining protein homologs over years of evolution, human proteins tend to be more complex and contain more domain architectures per protein than simpler organisms. The availability of a greater selection of domain architectures, along with a more complex assembly of those domains into multidomain proteins, will undoubtedly lead to significantly greater variety of protein function in humans .[12]

Thus, whereas the number of genes may not be dramatically larger in humans and other vertebrates, the complexity of the proteome and its vast catalog of activities is remarkably greater. Moreover, this complexity is likely to increase further when considering the regulation of protein levels and activities by both transcriptional and posttranscriptional mechanisms, and by the ability of proteins to interact with one another in a combinatorial amplification of different activities. Elucidating this regulation and these interactions is one of the great challenges of the Proteomic Era.

Evolutionary past

One of the most fertile areas of research arising from the genome project has been the study of human evolution. This molecular archaeology has exploited the frequency of repeat sequences that appear in the genome (Table 2).[13,14] Long the bane of most genome sequencers, greater than 50% of the human genome is composed of repeat sequences, most of which fall into five classes: transposon derived sequences; inactive,

Table 2
Various types of repeated sequences in the human genome

Repeat sequences in the genome
Transposon-derived
Inactive retroposed copies of cellular genes
Simple sequence repeats, e.g. $(A)_n$, $(CGC)_n$
Segmental duplications, 10–300 kb
Blocks of tandemly repeated sequences, such as centromeres, telomeres, etc.

retroposed copies of cellular genes; simple sequence repeats like multiple As or CGC repeats; large segmental duplications; and blocks of tandomly repeated sequences, such as centromeres and telomeres. These repeat sequences can be dated using several different techniques. In particular, the transposon repeats are useful because transposons must have contained functional elements when they first inserted into the chromosome and therefore their original starting sequence can be deduced. Any changes that are observed in the actual sequence compared with the predicted sequence presumably represent mutations occurring in DNA that is not under selection pressure. Thus the rate of mutation can be used to date the transposons, some of which we now know date back 800 million years. This deep fossil record can tell us much about our history.[15,16] In humans, for example, these nonfunctional sequences disappear very slowly. For unapparent reasons, the human genome retains them for a long time. Also, surprisingly, transposon activity has fallen dramatically in the last 50 million years. There are fewer than predicted new transposons in the human genome during this period.[2]

Some repeat sequences may actually provide an evolutionary advantage. The most common repeat sequences in the genome are called Alu sequences. Historically, Alu repeats have been regarded as irritants by those who have done positional gene cloning because they kept appearing in sequencing runs. Among other reasons, this is why these sequences were referred to as "junk DNA." Alu repeats comprise about 13% of the genome with more than a million and a half copies. They can be dated by their sequence divergence in a manner similar to that used for the transposon elements.

The Alu sequences can then be grouped according to their age and examined for their relative abundance in different parts of the genome (Fig. 4).[2] The Alu sequences are compared with regions of the genome characterized by GC content as a proxy for gene density. (As noted above, chromosomal areas with high GC content tend to have higher gene density.) Whereas Alu sequences that arrived more than 60 million years ago are located predominantly in the GC rich regions, Alu sequences that have arrived in the genome relatively recently, that is, in the last million years, are more abundant in the gene-poor, AT-rich regions. Thus, when new Alu repeats enter the genome, they tend to enter the genome in the gene-poor areas and over time disappear from the genome. However, when an Alu repeat occasionally hits into a gene-rich area, it has a relatively greater tendency to remain. Thus there may be a selective advantage to having an Alu repeat in a gene-rich area. It is not yet clear what that advantage is.

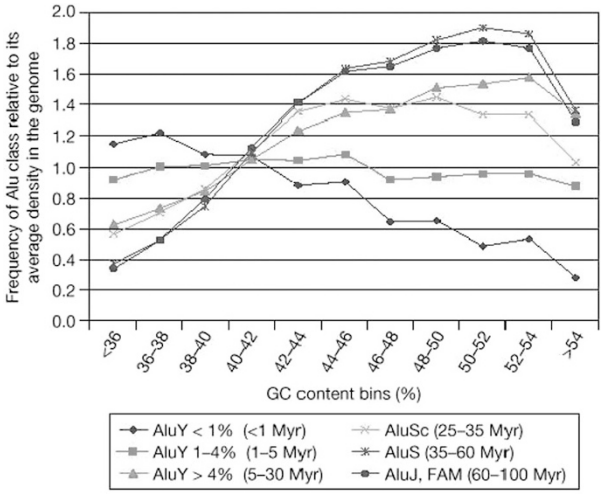

Fig. 4 The age and frequency of Alu sequences in the humane genome.[2] Reprinted by permission from Nature, copyright 2001, Macmillan Publishers Ltd.

The availability of the genome sequence allowed a comparison of the mutation rates between men and women. Because Y chromosomes undergo meiosis only in males, and X chromosomes undergo most of their meioses in females, the relative mutation rates in males and females can be examined by comparing these two chromosomes (Fig. 5). If the mutation rates were the same, then the curve ought to follow a 45-degree angle representing a one-to-one correlation, but it is actually skewed toward male mutations. In fact, there is almost a 2-fold higher mutation rate in men compared with women.[2]

Finally, perhaps the most dramatic evolutionary conclusion from the genome was the level of sequence similarity among

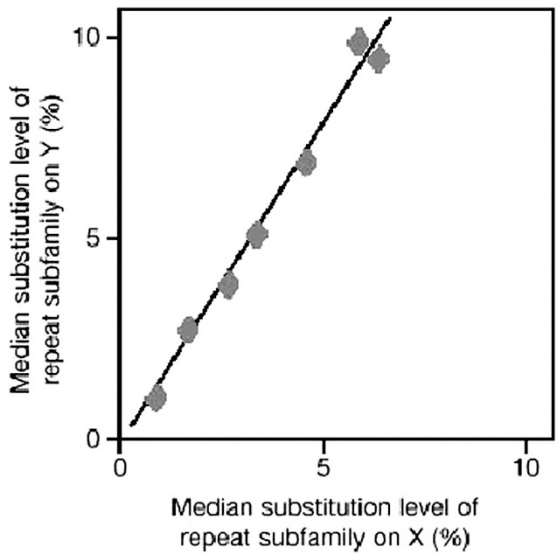

Fig. 5 A comparison of the mutation rates of the X and Y human chromosomes.[2] Reprinted by permission from Nature, copyright 2001, Macmillan Publishers Ltd.

the human species. Humans are 99.9% identical at the nucleotide level. That's a remarkable level of identity in an organism. In fact, most of our genetic differences—primarily measured by looking at SNPs—are shared among all ethnicities and races. Most of these differences date back to a time when humans were in Africa. In fact, the human species is far more alike, one to the other, than most other species on this planet. Most of the current human population appears to have derived from a very small group of common ancestors that expanded very quickly when the population left the African continent.[17,18]

Advances in biomedical research

Since the focus of this conference is the impact of genomics and proteomics on biomedical research, I will provide only a brief overview here. This impact falls into two general areas: medical genetics, and biochemical and functional analysis. The primary goal of medical genetics is to identify genes linked to disease, either directly or because they modify the host response to disease genes (Fig. 6). Proximally, medical genetics facilitates finding individuals at risk for disease and downstream it supplies clues to the discovery of targets for therapeutic intervention. The availability of millions of documented SNPs and eventually the assembly of a haplotype map will provide a framework for the rapid mapping of disease genes. The recent addition of DNA chips into the research toolkit has introduced a powerful method for examining many genes simultaneously for changes in RNA expression in the disease state. These tools are enhanced by the availability of the full sequence for model organisms like the fly and the worm.[19,11] Comparisons made to these well-characterized and experimentally manipulatable organisms help to discern which candidate genes are authentic and enable further experimentation to understand pathophysiology.

The future exploitation of these tools demands the development of important resources, however, which at present are only in their earliest stages. In the next decade there will be a strong need for large collections of tissue and blood samples that are linked to detailed clinical histories. Particularly for diseases that have incomplete penetrance and polygenic etiologies, studies that convincingly link genes to diseases will require large sample sizes. These sample repositories will need to represent a broad cross-section of our population including different age groups, racial backgrounds, geographical distribution, sexes, and perhaps social and economic factors. Moreover, the value of the samples will increase significantly if there are systematically and carefully collected clinical histories attached to them. The creation of such repositories has captured the interest of the private sector, and there are several companies that have begun to assemble these specimens. Although many of these commercial collections will also be available to the academic community, some may be costly, so the public sector must look forward and ensure that it has planned properly for its future need for these specimens.

Perhaps the most exciting application of the knowledge gained from the genome is its ability to expedite and enable the study of the biochemical and functional activity of all proteins. Proteins constitute both the operating machinery and the bricks and mortar of cells. Disease is most often the result of protein malfunction and is, consequently, most often treated by chemicals that modify protein activity. Indeed, it is impossible today to imagine the development of any pharmaceutical without thoroughly understanding the function of the target protein.

The historical approach to drug discovery began by screening for chemicals that caused a particular functional activity, which was then followed by binding experiments to find and identify the receptor, careful biochemistry to understand the protein's activity, and, finally, medicinal chemistry to optimize a good drug. This tedious and slow process has given way to a new paradigm that begins by using various genetic and molecular techniques to identify possible target proteins, a validation process that includes various protein expression and genetic studies to understand their functions in order to select the best target, and then a screening process that searches for small molecule binders to find drug candidates (Fig. 7). Several steps in this new paradigm can occur at very high-throughput levels. There are a number of high-throughput genomics tools available that enable the discovery of potential target proteins including SNP analysis, transcriptional profiling, and two-hybrid system mapping. In addition, the pharmaceutical industry now has combinatorial chemistry, libraries with high levels of complexity, and is capable of screening small molecules at ultra–high-throughput speeds. The challenge for the coming decade will be to develop the tools needed to expedite the process of understanding protein function. A strong emphasis will need to be placed on high-throughput protein expression and functional analysis if we are to achieve the promise of the Proteomic Era.

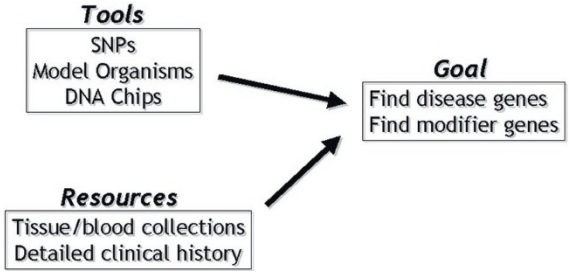

Fig. 6 The goals of medical genetics.

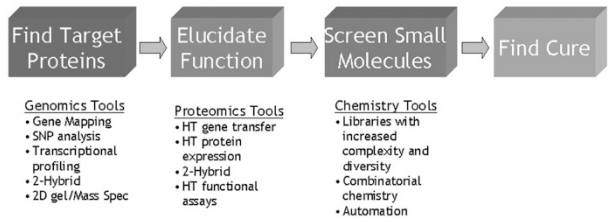

Fig. 7 The new path to drug discovery.

The term *proteomics* can be confusing because its meaning can change depending on who uses it, but it does convey the notion of studying many proteins, perhaps thousands, simultaneously. Proteomics can be separated into two broad classes: abundance-based proteomics and function-based proteomics. The goal of abundance-based proteomics is to identify proteins in different specimens and quantify them with the hopes of identifying some protein that might be increased or decreased in a particular state, such as in a disease tissue compared with a normal one. This can be used to find markers for diseases or even targets for therapy. In function-based proteomics the goal is to express proteins in high-throughput experiments and examine their function. Here the hope is to find proteins that cause a normal state to become diseased (such as causing a normal cell to behave like a cancer cell) or to build databases about the features of proteins (such as where all proteins are localized in the cell or categorizing their enzymatic activities).

To address this challenge, the community will require new and improved reagent sets. To produce and study a protein, the complete coding sequence of the protein must be known and a cloned copy of the DNA must be available. With the availability of the genome sequence, the identification and delineation of the mRNA sequence of nearly all genes can be anticipated in the next few years. An essential next step will be the assembly of a complete physical collection of DNA clones representing the proteome so that the corresponding proteins can be produced and their functions studied. By using cloning tools that allow the rapid transfer of genes from plasmid to plasmid, it is possible to perform large-scale experiments on the functions of many proteins simultaneously. This will lead to databases from which information can be mined about how proteins behave and how they affect disease etiology. Eventually, these data sets will be combined into computerized network models, essentially electronic cells, which allow users to predict the outcome of a cell if it is perturbed in a particular way. Thus the next great challenge in postgenomic biology will be the elucidation of the function of every protein encoded by the genome.

Genomics and medicine

The completion of the genome provides a powerful map for the rapid discovery of genetic markers that predict disease. Although dozens of disease genes have already been discovered more rapidly as a direct result of the genome sequence, the full impact of this will not be felt for some years. Not surprisingly, then, genomics will play an important role in the prediction and prevention of disease. For any given disease, this will depend to some extent on how much of a genetic component is present. In some cases, strong predictions can be made, whereas in others only a mild statistical tendency could be noted. The recent discovery that the genome appears to recombine in large segments will make the discovery of genetic markers easier. Nevertheless, large population studies using SNPs and carefully annotated specimens will be required to find useful markers. Moreover, for many diseases there will likely be a polygenic component and this may require even larger studies.

The value of predicting the predisposition to a disease depends on whether or not an early treatment is available. Alerting an individual that he or she will have early-onset Alzheimer disease could create unfortunate anxiety and angst without offering any substantial benefit. On the other hand, identifying individuals prone to hemochromatosis immediately points to a simple therapy that can virtually prevent the complications of the disease.

One area around which there is both excitement and controversy is pharmacogenomics. This concept recognizes that some individuals are genetically predisposed to respond poorly to a drug or even to have an idiopathic toxic reaction. If markers can be detected to predict these outcomes, doctors can appropriately tailor their therapy to the individual. Individuals will be genetically tested to determine whether or not they are likely to respond to certain drugs. In the example shown (Fig. 8), children who have an arginine at codon 27 in the β-adrenergic receptor for both alleles respond much better to albuterol than if they are homozygous for glycine.[20] Clearly, if oncologists could predict which patients would respond to specific chemotherapies, unnecessary side effects could be avoided and better response rates observed. However, such prediction methods are not yet effective and it would be unfortunate to deny a patient a therapy that otherwise would have been helpful.

An area that has gained particular attention lately is improved disease diagnostics using genomics and, in particular, DNA microarrays. It has become evident that cancers can be characterized beyond what is visible to a pathologist in a microscope. With DNA chips, tumor types can be carefully differentiated with great precision. These arrays can reveal the expression pattern of thousands of genes simultaneously. By applying mathematical tools for analyzing these gene expression patterns and grouping various tumor specimens together that share similar patterns, it is possible to recognize two or more different classes of a tumor that previously appeared to be only a single pathology. Using this approach, different types of lymphomas and leukemias can not only be identified but their likelihood of response to therapy can be predicted.[21,22]

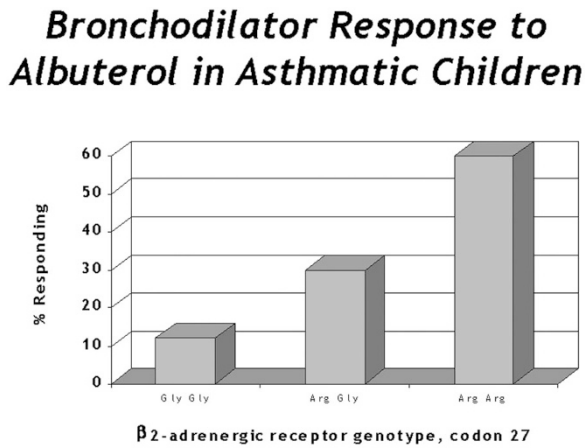

Fig. 8 An example of pharmacogenomics.[20]

The combination of genomics and proteomics will lead to highly specialized treatments arising from an increased understanding of proteins and their function. Perhaps the most exciting example of this is recent approval of STI-571 (Gleevec™). This drug was carefully and specifically designed to inhibit the kinase activity of the product of the *BCR-ABL* gene. This is the protein that results from the chromosome 9:22 translocation in chronic myeloid leukemia. The drug has demonstrated very good response rates in patients and, most important, has shown very limited toxicity. This is truly among the first "designer" drugs and heralds a new era in drug design.

Finally, since other presenters at this conference will talk about gene therapy, let me briefly mention it too. The approach aims to replace a gene that is missing in a patient who is ill without it, or to provide a gene that can replace a function that is otherwise missing. This approach to treatment faces many challenges, including how to deliver the gene to the patient so that it will be expressed appropriately without running any risks of inappropriately infecting or introducing the gene unintentionally into other individuals. At present, this procedure is still very experimental, but success is growing close. In the next decade, methods for doing this reproducibly will likely be achieved.

Public policy and genomics

Among the most important areas in genomics that now demand our attention is how it will interface with public policy (Fig. 9). A number of important questions arise in this area. Who gets access to information about our genes? This is particularly sensitive in countries like the United States, where medical care is largely paid for by private insurance. At the present time, it is not inconceivable that if an insurance company were to obtain information about an individual who had a predisposition to disease, it could cancel his or her insurance. This may not be true in other countries, such as in Korea, where everyone's coverage is guaranteed. But it still raises the question about how individuals can maintain privacy with respect to their medical history or their predisposition to other diseases. Obviously, we cannot choose our genes; therefore, individuals should not face discrimination because of those genes. What constitutes fair use of genetic information? This is something that is coming up now in the United States. Laws are being proposed to prevent unfair use of this genetic information.

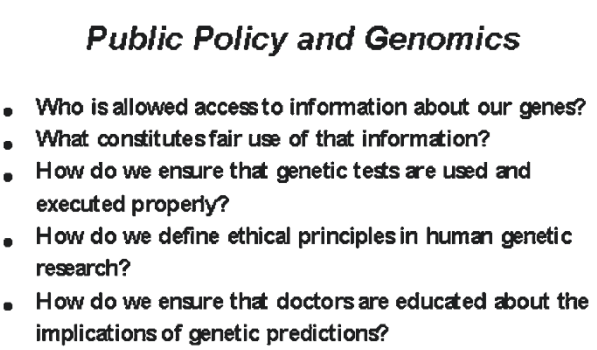

Fig. 9 The important questions that relate to the interface of genomics with public policy.[8]

For doctors, it will be very important to understand how to validate and monitor the use of genetic tests. A common set of standards will need to be set so that a valid test can be separated from one that is not truly predictive. And there must be mechanisms in place to monitor when a test can be clinically used. This in turn also raises issues about ethical principles in human genetic research. Carefully designed guidelines are needed to ensure that research is performed in the most ethical, safe, and effective manner. Physicians will require careful training about the implications of genetic predictions. It would be dangerous to venture down a path of genetic determinism. It must be recognized that whereas genes tell us much about what we are and what illnesses we might someday acquire, they do not tell us everything. They do not always tell when the illness will begin, to what degree it will occur, or even if it will absolutely occur at all. Environment still plays a very large role. Identical twins do not have identical medical histories.

As a world culture, scientists, ethicists, legislators, and lay people need to consider whether limits should be imposed on the use of genetic technology. Is gene therapy acceptable? Most of us would agree that in the treatment of certain illnesses, it would be acceptable. But what about inducing genetic changes because they are desirable—"genetic customization"? If a child with blonde hair or one who will be particularly tall is desired, how many of us would agree to this? Even more extreme, and more than a little frightening to me, might be the use of genetic techniques to change the human gene pool. We do not want to find ourselves in a position where these technologies are possible and the world has not first carefully considered their consequences. We have now entered the Genomic Era; we have reached the end of the beginning. Let us enter it with our eyes wide open and our hopes high.

References

1. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG et al. The sequence of the human genome. *Science* 2001;291:1304–1351.
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J et al. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860–921.
3. Sinsheimer RL. The Santa Cruz Workshop—May 1985. *Genomics* 1989;5:954–956.
4. Palca J. Human genome: Department of Energy on the map. *Nature* 1986;321:371.
5. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, Hutchison CA 3rd, Slocombe PM, Smith M. Nucleotide sequence of bacteriophage Φ X174 DNA. *Nature* 1977;265:687–695.
6. Tabor S, Richardson CC. A single residue in DNA polymerases of the *Escherichia coli* DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. *Proc Natl Acad Sci U S A* 1995;92:6339–6343.
7. Gocayne J, Robinson DA, FitzGerald MG, Chung F-Z, Kerlavage AR, Lentes K-U, Lai J, Wang C-D, Fraser CM, Venter JC. Primary structure of rat cardiac β-adrenergic and muscarine cholinergic receptors obtained by automated DNA sequence analysis: further evidence for a multigene family. *Proc Natl Acad Sci U S A* 1987;84:8296–8300.
8. Collins F. Consequences of the Human Genome Project for medicine and society. Lecture presented at Harvard Medical School, February 20, 2001, Boston.
9. Nicholas Wade. Long-held beliefs are challenged by new human genome analysis. *New York Times* February 12, 2001;sect A:20.
10. Lin X, Kaul S, Rounsley S, Shea TP, Benito M-I, Town CD, Fujii CY, Mason T, Bowman CL, Barnstead M, Feldblyum TV, Buell CR, Ketchum KA, Lee J, Ronning CM, Koo HL, Moffat KS, Cronin LA, Shen M, Pai G, Van Aken S, Umayam L, Tallon LJ, Gill JE, Adams MD, Carrera AJ, Creasy TH, Goodman HM, Somerville CR, Copenhaver GP, Preuss D, Nierman WC, White O, Eisen JA, Salzberg SL, Fraser

CM, Venter JC. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* 1999;402:761–768.
11. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, Cherry JM, Henikoff S, Skupski MP, Misra S, Ashburner M, Birney E, Boguski MS, Brody T, Brokstein P, Celniker SE, Chervitz SA, Coates D, Cravchik A, Gabrielian A, Galle RF, Gelbart WM, George RA, Goldstein LSB, Gong F, Guan P, Harris NL, Hay BA, Hoskins RA, Li J, Li Z, Hynes RO, Jones SJM, Kuehl PM, Lemaitre B, Littleton JT, Morrison DK, Mungall C, O'Farrell PH, Pickeral OK, Shue C, Vosshall LB, Zhang J, Zhao Q, Zheng XH, Zhong F, Zhong W, Gibbs R, Venter JC, Adams MD, Lewis S. Comparative genomics of the eukaryotes. *Science* 2000;287:2204–2215.
12. Koonin EV, Aravind L, Kondrashov AS. The impact of comparative genomics on our understanding of evolution. *Cell* 2000;101:573–576.
13. Smit AFA, Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* 1999;9:657–663.
14. Prak ET, Kazazian HH Jr. Mobile elements and the human genome. *Nat Rev Genet* 2000;1:134–144.
15. Li W-H. Molecular evolution. Sunderland, MA: Sinauer Associates, 1997.
16. Sarich VM, Wilson AC. Generation time and genomic evolution in primates. *Science* 1973;179:1144–1147.
17. Daly MJ, Rioux JD, Scaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. *Nat Genet* 2001;29:229–232.
18. Svante P. The human genome and our view of ourselves. *Science* 2001;291:1219–1220.
19. Wilson R, Ainscough R, Anderson K, Baynes C, Berks M, Bonfield J, Burton J, Connell M, Copsey T, Cooper J, Coulson A, Craxton M, Dear S, Du Z, Durbin R, Favello A, Fraser A, Fulton L, Gardner A, Green P, Hawkins T, Hillier L, Jier M, Johnston L, Jones M, Kershaw J, Kirsten J, Laisster N, Latreille P, Lightning J, Lloyd C, Mortimore B, O'Callaghan M, Parsons J, Percy C, Rifken L, Roopra A, Saunders D, Shownkeen R, Sims M, Smaldon N, Smith A, Smith M, Sonnhammer E, Staden R, Sulston J, Thierry-Mieg J, Thomas K, Vaudin M, Vaughan K, Waterston R, Watson A, Weinstock L, Wilkinson-Sproat J, Wohldman P. 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*. *Nature* 1994;368:32–38.
20. Martinez FD, Graves PE, Baldini M, Solomon S, Erikson R. Association between genetic polymorphisms of the beta2-adrenoceptor and response to albuterol in children with and without a history of wheezing. *J Clin Invest* 1997;100:3184–3188.
21. Alizadeh A, Eisen M, Davis RE, Ma C, Sabet H, Tran T, Powell JI, Yang L, Marti GE, Moore DT, Hudson JR Jr, Chan WC, Greiner T, Weisenburger D, Armitage JO, Lossos I, Levy R, Botstein D, Brown PO, Staudt LM. The lymphochip: a specialized cDNA microarray for the genomic-scale analysis of gene expression in normal and malignant lymphocytes. *Cold Spring Harb Symp Quant Biol* 1999;64:71–78.
22. Golub TR, Slonim DK, TamayoP, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999;286:528–530.