Biotechnol. Bioprocess Eng. **2000**, 5: 307-312

Perspectives on Functional Genomics

Kyuyoung Song*

Department of Biochemistry, University of Ulsan College of Medicine, 388-1 Poongnap-dong, Songpa-ku, Seoul 138-736, Korea

Abstract As the first assembly of the human genome was announced on June 26, 2000, we have entered post genome era. The genome sequence represents a new starting point for science and medicine with possible impact on research across the life sciences. In this review I tried to offer brief summaries of history and progress of the Human Genome Project and two major challenges ahead, functional genomics and DNA sequence variation research.

Keywords: functional genomics, human genome project, single-nucleotide polymorphisms

INTRODUCTION

The Human Genome Project (HGP) is an international project with its ultimate aim in obtaining the complete nucleotide sequence of the human genome by DNA sequencing. The HGP was conceived in the mid-1980's out of the need for a large-scale project to develop new mutation detection methods. The HGP formally began in October 1990 to discover all of the approximate 100,000 human genes and to determine the three billion bases of the human genome. Initially, the HGP was to last for 15 years, however, rapid technological advances have reduced the expected completion date by at least two years. Furthermore, on June 26, 2000, a rough draft of the entire human genome was completed – well ahead of schedule. It is difficult to imagine what would change as a result of the HGP. In this review, what impact the HGP would have on science and medicine in the future will be discussed.

WHAT IS THE HUMAN GENOME PROJECT?

The genome is defined as the total genetic material contained within the chromosomes of an organism. Humans have 46 chromosomes, 23 from each parent, which represent our genome. The human genome comprises of about three billion base pairs, counting only one of each pair of chromosomes, and encodes about 100,000 proteins. These coding regions make up only about 2% of the genome and the function of the remaining 98% is unknown. The major goals of the 15-year HGP include: construction of high resolution genetic and physical maps of the human genome; determination of the complete nucleotide sequence of human DNA and of the DNAs of selected model organisms; development of capabilities for collecting, storing, distributing and analyzing the data produced; creation of appropriate technologies necessary to achieve these objectives [1,2]. The goals also include analysis of the ethical, legal, and social implications of human genetics research, and training of scientists who would be able to utilize the tools and resources developed through the HGP to pursue biological studies that will improve human health. Over the past ten years excellent progress has been made on the construction of genetic and physical maps of the human genome – ordering polymorphic and monomorphic markers within the genome [3-5]. These maps have assisted in identifying over 100 disease-causing genes. Parallel studies on selected model organisms have been extremely helpful in developing the technology and interpreting human gene function. The sequencing of the DNA of *Escherichia coli, Saccharomyces cerevisiae, Caenorhabditis elegans* and *Drosophila melanogaster* is complete and their complete genome sequences are in GenBank [6-9]. The current emphasis is placed on sequencing mouse genome.

*Corresponding author
Tel: +82-2-2224-4277 Fax: +82-2-2224-4248
e-mail: kysong@www.amc.seoul.kr

FUNCTIONAL GENOMICS

Identification of Coding DNA

From the start of the HGP there has been much debate over whether to go for whole genome sequencing or to focus on the coding DNA sequences. Indeed, even if the entire human genome sequence is available, identification of all the human transcripts is a daunting task that will continue far beyond the projected the year 2003 deadline mainly because much of the DNA does not encode transcripts (coding DNA accounts for a mere 2% of the genome), the coding sequences of most human genes are split into exons, which are separated by noncoding intervening sequences, introns, and many individual human genes undergo alternative splicing and polyadenylation to yield different mRNA sequences

encoding protein isoforms which may be tissue specific. It is possible to identify genes by computer analysis of the DNA sequence using two types of software, homology searches and exon prediction, but the best programs have been only moderately successful. Among various ways of identifying and isolating human coding DNA clones, the expressed sequence tag (EST) approach, which involves the construction of cDNA libraries and sequencing short fragments of cDNA clones from various human tissues, was chosen for the establishment of systematic cDNA resources [10]. ESTs covered only a segment of the gene, however, their utility for gene identification was immediately recognized.

NIH initiated the projects to discover genes from various human tissues. Two examples are CGAP (Cancer Genome Anatomy Project) and BMAP (Brain Molecular Anatomy Project). These two projects have two critical components in common; gene discovery phase and gene expression profiling phase. The gene discovery phase was aimed at cataloging all the genes expressed in normal and cancerous tissues (CGAP) and in various regions of adult and developing brains (BMAP). Construction of cDNA libraries, 3' and 5' EST sequencing and clustering of EST clones into non-redundant sets of unique clones led to the catalog/index of UniGenes expressed in given tissues. Public EST sequencing projects have now contributed over 1.5 million human ESTs to a GenBank division specifically devoted to managing EST sequences (dbEST). The IMAGE (*integrated molecular analysis of gene expression*) consortium (http://image.llnl.gov/) was formed, led by the Lawrence Livermore National Laboratory, and has fostered the development of gridded arrays of cDNA clones and preparation of master arrays of clones representing unique genes, facilitating the widespread use of ESTs among the biomedical research community. The IMAGE consortium clones also provided excellent resources for gene-expression profiling experiments by cDNA microarrays. With the advances of the technologies for obtaining full-length cDNAs, the mammalian gene collection project has launched in 1999 to generate full-length cDNA resources [11].

Gene-Expression Profiling

Knowing the genome sequence or even all the locations of genes of an organism does not guarantee that everything about life at the molecular level will soon be understood. Two major challenges ahead, once the HGP is complete, are to decipher the function of each human gene and to be able to correlate specific molecular variation with phenotype changes.

Functional genomics is the science of understanding how the genome functions through controlling the expression of genes. Once the catalog of UniGenes expressed in given tissues becomes available through various EST projects, the next phase, then, is to examine roles of genes by determining gene expression patterns and monitoring alterations in gene expression profiles during development, metastasis, or other pathological conditions. With advances in high-density DNA microarray technology, it has become possible to screen large numbers of genes to see whether or not they are active under various conditions [12-15]. Microarrays vary in detail, but in essence consist of a dense array of nucleotide sequences, perhaps oligos chemically synthesized on a glass chip, or longer cDNAs robotically spotted. The sequences represent all the genes of a cell or some subset. mRNAs are collected from a cell during some developmental stage, or following an experimental treatment, and they, or cDNAs, are labeled, usually with a fluorescent marker. The labeled molecules are then hybridized to the glass-bound set of cellular DNAs. Expressed sequences glow, so genes that are expressed together are immediately apparent, thus providing a snapshot of gene activity for thousands of genes. This is called gene-expression profiling, and numerous experiments are under way studying how genes are turned on and off in complex plants, pathogens, model organisms, and human cancer cells [16-20]. In the future a complete genome expression analysis for a single cell might eventually be possible. Data from many experiments can be compared and genes that have consistent patterns of activity can be grouped or clustered. Although analytical tools for understanding the information presented on a microarray are still rudimentary, a great deal of effort has been put into applications of various algorithms for clustering similarly expressed sequences [21-23]. It might be possible to hypothesize about unknown genes, by linking their expression patterns to known genes. In this way, genes that characterize a particular cell state, such as malignancy, can be identified – so providing new information about the biology of the cell state.

There has been an expectation that gene-expression profiling will revolutionize cancer diagnosis. The rationale is that tumor behavior is dictated by the expression of thousands of genes, and that microarray analysis should allow the prediction of the behavior and clinical consequences. Recently two reports demonstrated that gene expression information can be used to categorize human cancers with, human acute leukemia [24] and diffuse large B-cell lymphoma (DLBCL) [25] as test cases, in ways that will support therapeutic decisions. However, class prediction entails two challenges: 1) developing algorithms to cluster samples by gene expression patterns, and 2) validation: determining whether those putative classes are meaningful and reflect true biological structure in the data rather than aggregation by chance. An attempt to integrate large databases on gene expression and molecular pharmacology by assessing gene expression profiles in 60 human cancer cell lines has demonstrated how variations in the transcript levels of particular genes relate to mechanisms of drug sensitivity and resistance [26]. A novel molecular characterization of human cancer cell lines frequently used for drug screening and their relationships to tumors in vivo was also established by exploring the variation in expression of approximately 8,000 unique genes among the cell lines [27]. In general microarray experiments

need much observation before formulating hypotheses. The experiments uncover genes, which seem to be important in some regulatory pathway or other, but which are otherwise unknown. At least a tentative hypothesis may be possible, along the lines that genes that behave alike are functionally related, suggesting further experiments and observations.

Functional Genomics Using Model Organisms

Researchers use model organisms such as laboratory mouse, zebrafish, or drosophila to help guide functional genomics. Due to the presence of many valuable animal models of human diseases and the rapid progress of mouse genome sequencing project, mice have been always the major focus. The two most popular approaches are available to generate a mutant mouse on a large scale for such studies. One is based on gene trapping techniques and uses the ability to create mutations in mouse embryonic stem cells on a large scale (http://www.lexgen.com/). The other involves genome-wide, saturation chemical mutagenesis using ethylnitrosourea (ENU), a supermutagen that primarily causes single base changes, to identify and isolate genes with important functions for mammalian biology. Information on a major ENU mutagenesis project and data derived from a systematic screening protocol is available at http://www.har.mrc.ac.uk/mutabase. Mutant animals were then utilized to positionally clone the genes that underlie defects. Advances in microarray technology would make it possible to examine roles of genes in normal physiological as well as pathophysiological states.

Human Genomic Diversity

The central aim of all genetics is to correlate specific molecular variation with phenotype changes. Once the human genome is sequenced, it will become possible to understand the spectrum of genetic variation in the human gene pool and its relation to diseases. Why does one man live to celebrate his hundredth birthday while another succumbs in midlife to cancer or heart disease? And why may one woman's breast cancer be effectively eradicated while another's shows no significant response to the same treatment? The explanations may reside in the cumulative effect of a small number of differences in DNA base sequence called single-nucleotide polymorphisms (SNPs), which underlie individual responses to environment, disease, and medical treatments [2,12].

SNPs are variations of one nucleotide between the DNA sequence of individuals and the most common type of sequence variation. Other variations include a number of base insertions and deletions and sequence repeats (called mini- and microsatellites). The occurrence of SNPs is approximately one every 800 bases throughout the human genome. SNPs are found in genes that encode proteins (called cSNPs) and in noncoding regions. These polymorphisms could be the basis for most human diseases because of their potential functional importance. If the SNP occurs in the protein-coding region of a gene, it could be in the 'wobble' position of a codon and thus not affecting the amino acid sequence of the protein; or it could be in a conserved position and could potentially alter the function of the protein. Some disease-causing mutations are SNPs, for example, single base change in the gene associated with sickle cell anemia. If the SNP results in the formation of stop codon, the protein will be truncated and might or might not be functional. cSNPs that are insertions or deletions can cause shifts in the reading frame of the gene, resulting in a wrong amino acid sequence. Additionally, if the SNP occurs in the control region of the gene, it could affect protein levels through effects on transcription and/or translation.

DNA variations are important in understanding the genetic basis for disease and individual responses to environmental factors, as well as for such normal variations in biological processes as development and aging. Geneticists have long been preoccupied with studying simple single-gene diseases using familial linkage methods [27,28]. However, complex, high-incidence, multigene diseases are thought to require a different approach, for which SNPs can play a key role [2,12,29-32]. Researchers believe that SNPs can be used as a tool for mapping genes in the diseases that have complex patterns of inheritance such as asthma, diabetes, heart disease, schizophrenia and cancer [12,30,31]. SNPs, because they can distinguish the smallest of differences in alleles, also provide a complementary tool to gene expression profiling with microarrays that might be blind to such subtle differences. For SNPs to be truly useful, however, they must be densely arrayed over the genome, in every 5,000 to 10,000 base pairs or so. For this reason, scientists in the public and private sectors have begun to focus their attention on methodically searching for SNPs throughout the human genome [33-38]. The first large-scale survey for SNPs was to identify a total of 3,241 candidate SNPs by a combination of sequencing and high-density variation-detection DNA chips. DNA chips used in this study were prototype genotyping chips, arrays for studying DNA, which could be used for identification and genotyping of mutations and polymorphisms. A genetic map was constructed showing the location of 2,227 of these SNPs [39]. Large scale identification of cSNPs in the genes associated with cardiovascular diseases revealed that the number of cSNPs varies depending on genes, and suggested that it might be difficult to find cSNPs [40,41]. Several million SNPs exist in the human genome, and efforts, both public and private, to generate much denser SNP maps are ongoing. The SNP Consortium was formed in 1999 by joining eleven major pharmaceutical companies and four leading publicly funded genome institutes and started a two-year project which aims to catalogue and detect hundreds of thousands of human SNPs in different DNA samples [37]. SNPs generated in these public projects are freely available from dbSNP, a new database at the NIH National Center for Biotechnology Information. As of August 2000, approximately 804,000 SNPs

had been deposited into public databases (http://hgbase.cgr.ki.se/; http://www.ncbi.nlm.nih.gov/SNP/) compared to 7,000 SNPs in April, 1999. The dense map of SNPs can be used in genome-wide studies of cases and controls designed to identify markers found only in one of those two groups. These markers, in turn, provide valuable clues to the location and identity of complex disease genes.

Pharmacogenomics is a new field emerging from pharmacogenetics (for a review, see [42]). Pharmacogenetics involves genotyping populations to identify SNPs affecting drug metabolism. Before the HGP a relatively small number of drug-metabolizing enzymes was known: cytochrome P450 (CYP) isozymes, N-acetyl transferase (NAT) isozymes, UDP-glucuronosyl transferases and methyl transferases. For most drugs, CYP determines how long and how much of a drug remains in the body. For example, polymorphisms of CYP2D6, one of the 6 CYP isozymes, dictate slow or ultra-rapid metabolizers of antidepressants, antipsychotics, β-blockers and antiarrhythmics and lead to systemic accumulation and toxicity. With the advent of the HGP and automated analysis that can handle up to many samples per day, the goal of pharmacogenomics is to exploit all of the relevant SNP variations to improve drug efficacy and toxicity. It is not for predisposition or predictive testing on the risk of a disease or its prognosis but is rather concerned with genetic effects on drug themselves and with the genetic variances that contribute to the variable effects of drugs in different individuals. Researchers predict that pharmacogenomic methods would reduce clinical development times and costs, reveal new indications for existing drugs and ultimately generate personalized medicines. Also recent advances in genomics will provide an opportunity to expand a range of potential drug targets and lead to changes in the drug discovery paradigm in the future [43].

Other Areas

As we enter the post genomic era, it becomes clear that the really interesting aspects of biology go far beyond assembling genomes and finding genes. Other challenging areas that are not discussed here are proteomics and bioinformatics. Proteomics is the study of proteins using technologies of large-scale protein separation and identification. Current trend on proteomics is available elsewhere [47-49]. A new field of bioinformatics is blossoming [44-46], as the sequencing of the human genome project nears its completion and the amount of data is overwhelming. There are hundreds of databases, of just raw sequences, of expressed sequences, of genetic diseases, of single nucleotide polymorphisms, of chromosomal aberrations associated with cancer, and more. Bioinformatics is developing tools and analytical techniques for understanding large biological data sets, such as genome and protein sequences. New techniques for extracting useful information through data mining of completed genomes continue to be developed. It will be exciting to watch the cooperation between bioinformatics and biology in the coming years.

THE BENEFITS OF THE HUMAN GENOME PROJECT

One of the most exciting prospects for genomic information is its potential influence on medicine [50,51]. It is estimated that the sequence of the human genome should be completely mapped by approximately the year 2003. Already, however, many new techniques developed until now during the period of this project have provided doctors with improved genetic diagnostic and predictive testing. The availability of extensive genetic maps has increased the pace by which different disease genes are localized in the human genome. The techniques have also made it possible to identify susceptible areas of the genome that may be responsible for some disorders, such as diabetes, hypertension and certain forms of cancer that are more complicated and are caused by more than one genetic change. The collection and analysis of DNA samples may, in conjunction with epidemiological evidence, help lead to the identification of genetic factors in some human diseases and eventually to ways to treat or prevent those diseases. The HGP will revolutionize the practice of medicine, providing the means to custom tailor treatments to the needs of each patient, and to prolong healthy life by predicting and preventing diseases.

REFERENCES

[1] Collins, F. S. and D. Galas (1993) A new five-year plan for the U.S. human genome program. *Science* 262: 43-46.

[2] Collins, F. S., A. Patrinos, E. Jordan, A. Chakravarti, R. Gesteland, L. Walters, and the members of the DOE and NIH planning groups (1998) New goals for the U.S. human genome project: 1998-2003. *Science* 282: 582-689.

[3] Hudson, T. J., L. D. Stein, S. S. Gerety, J. Ma, A. B. Castle, J. Silva, D. K. Slonim, R. Baptista, L. Kruglyak, S. H. Xu *et al.* (1995) An STS-based map of the human genome. *Science* 270: 1945-1954.

[4] Dib, C., S. Faure, C. Fizames, D. Samson, N. Drouot, A. Vignal, P. Millasseau, S. Marc, J. Hazan, E. Seboun, M. Lathrop, G. Gyapay, J. Morissette, and J. Weissenbach (1996) A comprehensive genetic map of human genome based on 5,246 microsatellites. *Nature* 380: 152-154.

[5] Schuler, G. D., M. S. Boguski, E. A. Stewart, L. D. Stein, G. Gyapay, K. Rice, R. E. White, P. Rodriguez-Tome, A. Aggarwal, E. Bajorek, S. Bentolila, B. B. Birren, A. Butler, A. B. Castle, N. Chiannilkulchai, A. Chu, C. Clee, S. Cowles, P. J. Day, T. Dibling, N. Drouot, I. Dunham, S. Duprat, C. East, T. J. Hudson *et al.* (1996) A gene map of the human genome. *Science* 274: 540-546.

[6] Blattner, F. R., Plunkett G 3rd, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao (1997) The complete genome sequence of *Escheri-*

chia coli K-12. *Science* 277: 1453-1474.
[7] The yeast genome directory. (1997) *Nature* 387 (Suppl.): 5-105.
[8] The *Caenorhabditis elegans* Sequencing Consortium. (1998) Genome sequence of the nematode *Caenorhabditis elegans*: A platform for investigating biology. *Science* 282: 2012-2018.
[9] The genome sequence of *Drosophila melanogaster*. (2000) *Science* 287: 2185-2195.
[10] Adams, M. D., J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno, A. P. Kerlavage, W. R. McCombie, and J. C. Venter (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252:1651-1656.
[11] Strausberg, R. L., E. A. Feingold, R. D. Klausner, and F. S. Collins (1999) The mammalian gene collection. *Science* 286: 455-457.
[12] Lander, E.S. (1996) The New Genomics: Global Views of Biology. *Science* 274: 536-539.
[13] Khan, J., L. H. Saal, M. L. Bittner, Y. Chen, J. M. Trent, and P. S. Meltzer (1999) Expression profiling in cancer using cDNA microarrays. *Electrophoresis* 20: 223-229.
[14] Lander, E. S. (1999) Array of hope. *Nat. Genet.* 21: 3-4.
[15] Epstein, C. B. and R. A. Butow (2000) Microarray technology - enhanced versatility, persistent challenge. *Curr. Opin. Biotechnol.* 11: 36-41.
[16] DeRisi, J., L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. M. Trent (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* 14: 457-460.
[17] Lockhart, D. J., H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* 14: 1675-1680.
[18] Shelton, D. N., E. Chang, P. S. Whittier, D. Choi, and W. D. Funk (1999) Microarray analysis of replicative senescence. *Curr. Biol.* 9: 939-945.
[19] White, K. P., S. A. Rifkin, P. Hurban, and D. S. Hogness (1999) Microarray analysis of Drosophila development during metamorphosis. *Science* 286: 2179-2184.
[20] Geiss, G. K., R. E. Bumgarner, M. C. An, M. B. Agy, A. B. van't Wout, E. Hammersmark, V. S. Carter, D. Upchurch, J. I. Mullins, and M. G. Katze (2000) Large-scale monitoring of host cell gene expression during HIV-1 infection using cDNA microarrays. *Virology* 266: 8-16.
[21] Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein (1998) Cluster analysis and display of genome-wide expression pattern. *Proc. Natl. Acad. Sci. USA* 95: 14863-14868.
[22] Tamayo, P., D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* 96: 2907-2912.
[23] Alon, U., N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* 96: 6745-6750.
[24] Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286: 531-537.
[25] Alizadeh, A. A., M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshrani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Bostein, P. O. Brown, and L. M. Staudt (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403: 503-511.
[26] Scherf, U., D. T. Ross, M. Waltham, L. H. Smith, J. K. Lee, L. Tanabe, K. W. Kohn, W. C. Reinhold, T. G. Myers, D. T. Andrews, D. A. Scudiero, M. B. Eisen, E. A. Sausville, Y. Pommier, D. Botstein, P. O. Brown, and J. N. Weinstein (2000) A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.* 24: 236-244.
[27] Ross, D. T., U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.* 24: 227-235.
[28] Collins, F. S. (1992) Positional cloning; let's not call it reverse anymore. *Nat. Genet.* 1: 3-6.
[29] Collins, F. S. (1995) Positional cloning moves from perditional to traditional. *Nat. Genet.* 9: 347-350.
[30] Risch, N. and K. Merikangas (1996) The future of genetic studies of complex human diseases. *Science* 273: 1516-1517.
[31] Kruglyak, L. (1997) The use of genetic map of biallelic markers in linkage studies. *Nat. Genet.* 17: 21-24.
[32] Collins, F. S., M. S. Guyer, and A. Chakravarti (1997) Variations on a theme: cataloging human DNA sequence variation. *Science* 278: 1580-1581.
[33] Brookes, A. J. (1999) The essence of SNPs. *Gene* 234: 177-186.
[34] Marshall, E. (1997) 'Playing chicken' over gene markers. *Science* 278: 2046-2048.
[35] Collins, F. S., L. D. Brooks, and A. Chakravarti (1998) A DNA polymorphism discovery resources for research on human genetic variation. *Genome Res.* 8: 1229-1231.
[36] Buetow, K. H., M. N. Edmonson, and A. B. Cassidy (1999) Reliable identification of large numbers of candidate SNPs from public EST data. *Nat. Genet.* 21: 323-325.
[37] Marshall, E. (1999) Drug firms to create public database of genetic mutations. *Science* 284: 406-407.
[38] Marth, G. T., I. Korf, M. D. Yandell, R. T. Yeh, Z. Gu, H. Zakeri, N. O. Stitziel, L. Hillier, P.-Y. Kwok, and W. R. Gish (1999) A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.* 23: 452-456.
[39] Wang, D. G., J. -B. Fan, C. -J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, E.

Hubbell, E. Robinson, M. Mittmann, M. S. Morris, N. Shen, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T. J. Hudson, R. Lipshutz, M. Chee, and E. S. Lander (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280: 1077-1082.

[40] Cargill, M., D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, C. R. Lane, E. P. Lim, N. Kalvanaraman, J. Nemesh, L. Ziaugra, L. Friedland, A. Rolfe, J. Warrington, R. Lipshutz, G. Q. Daley, and E. S. Lander (1999) Characterization of a single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* 22:231-238.

[41] Halushka, M. K., J. -B. Fan, K. Bentley, L. Hsie, N. Shen, A. Weder, R. Cooper, R. Lipshutz, and A. Chakravarti (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* 22: 239-247.

[42] Evans, W. E. and M. V. Relling (1999) Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* 286: 487-491.

[43] Rosamond, J. and A. Allsop (2000) Harnessing the power of the genome in the search for new antibiotics. *Science* 287: 1973-1976.

[44] Spengler, S. J. (2000) Bioinformatics in the information age. *Science* 287: 1221-1223.

[45] Wickware, P. (2000) Next-generation biologists must straddle computation and biology. *Nature* 404: 683-684.

[46] Lee, P. S. and K. H. Lee (2000) Genomic analysis. *Curr. Opin. Biotechnol.* 2: 171-175.

[47] Abbott, A. (1999) A post-genomic challenge : Learning to read patterns of protein synthesis. *Nature* 402: 715-720.

[48] Service, R. F. (2000) Protein data justice for all. *Science* 288: 939-941.

[49] Dutt, M. J. and K. H. Lee (2000) Proteomic analysis. *Curr. Opin. Biotechnol.* 2: 176-179.

[50] Sander, C. (2000) Genomic medicine and the future of health care. *Science* 287: 1977-1978.

[51] Brenner, S. (2000) Genomics. The end of the beginning. *Science* 287: 2173-2174.

[Received April 17, 2000; accepted October 4, 2000]