Genomics, science and medicine: the future is now

Samuel Broder*

J. Craig Venter

Celera Genomics
45 West Gude Drive
Rockville
MD 20850
USA
*Corresponding author

▼ Basic researchers, clinical investigators, practicing physicians, patients and the general public, now live in a paradigm-shifted world in which the 2.9 billion letter sequence (nucleotide base pairs) of the human genome is available as a fundamental resource for scientific discovery[1,2]. Some findings from the completion of the human genome were expected, confirming knowledge presaged by many decades of research in both human and comparative genetics. Other findings, such as the relatively low gene number and large segmental DNA duplications, were unexpected and startling in their scientific and philosophical implications[1]. In either case, the availability of the human genome sequence will probably have profound implications, first on basic research and then on the practice of medicine.

EST technology for genome sequencing

The journey to this point was not a straight line, nor was it easy in any sense. For us, expressed sequence tags (ESTs) were a crucial starting point[3]. Since recombinant DNA techniques became available in the 1970s, scientists had developed an ability to use cloned DNA, representing a gene of interest, in a wide variety of molecular studies in biology. In the late 1980s, as the Human Genome Project was under discussion, a case was made for a complete genome sequence and a catalog of genes. It is astonishing how quickly access to the complete genome sequence of an organism has become an essential step for any new comprehensive research project. However, only a few years ago, this goal seemed very far away for all but a handful of viruses with very small genomes. Most of the genome sequence projects before 1990 were unavoidably slow and tedious, and targets for achieving even intermediary goals were measured in decades. The EST technology was the first to unleash the full power of an automated random cDNA library sequencing strategy for rapid gene discovery[3,4]. Other efforts lacked both scale and speed. ESTs would help identify new genes by sequencing 300–500 base pairs each of a very large number of cDNAs from a variety of tissues. ESTs could also be used to help map the chromosomal location of genes, recover corresponding genomic sequence, and retrieve complete cDNA clones for further analysis. Perhaps most important of all, ESTs contained enough information to identify an enormous number of genes by similarity searching of electronic databases. When the results were published, the scientific community had the largest collection of human genes in the history of genomic research at that point in time.

Necessity is the mother of invention, and this is no less true for genomics than for other fields. By the mid-1990s, increasingly large numbers of ESTs necessitated the development of computational methods to combine overlapping sequences in a way similar to contig assembly, but with orders-of-magnitudes more data. EST assembly served both to reduce redundancy (multiple copies of the same EST sequence) and to capitalize on it (to create consensus sequences representing up to the full length of the cDNA). The bioinformatics that developed as a consequence of those efforts, in turn, made it possible to explore the entire genomic sequence of a free-living organism.

The first organism targeted was *Haemophilus influenzae* (Ref. 5). This is an important pathogen in its own right, and an elegant model for all of microbiology. The plan was to randomly fragment the bacteria's genomic DNA into small pieces, repeatedly sequence the fragmented DNA until, on average, every nucleotide had been sequenced an appropriate number of times according to a Poisson distribution, and then apply very powerful computational assembly tools (combined with a directed effort to close the remaining gaps) to provide a final fully-assembled complete genome. Along the way, it became necessary to master the advanced automation, robotics and other features of industrial scale DNA sequencing. Since the 1995 publication of *H. influenzae* (Ref. 5), many more genome sequences of free-living organisms have been determined (http://www.tigr.org/tdb/mdb/mdbcomplete.html). The approach is called whole-genome shotgun sequencing, and it formed the basis for our publication of the sequence of first the fruit fly[6] and then the human genome[1].

Shifting the paradigms

Our genomic sequence provides a unique record of who we are and how we evolved as a species, including the fundamental unity of all human beings[7]. The knowledge fostered by understanding the genome might resolve which human characteristics are innate or acquired, as

1359-6446/01/$ – see front matter PII: S1471-1931(01)00175-6

well as the interplay between heredity and environment in defining susceptibility to illness. Such an understanding will make it possible to study how our genomic DNA varies among cohorts of patients, especially the role of such variation in important illnesses and in responses to pharmaceuticals[8–10]. We can also begin new ways of asking fundamental questions regarding complex aspects of the human condition such as language, thought, self-awareness, and higher-order consciousness. The study of the genome and the associated protein content (proteomics) of free-living organisms will eventually make it possible to localize and annotate every human gene, as well as the regulatory elements that control the timing, organ-site specificity, extent of gene expression, protein levels, and the post-translational modifications that define health or illness. For any given physiological process, we will have a new paradigm for addressing its evolution, development, function and mechanism in causing disease, and in affecting the onset and outcome of disease.

Now, we can also more systematically explore curious, and indeed almost mysterious, innate differences between individuals, mediated by epigenetic factors. For example, there are vast numbers of retroelements in mammalian genomes[1,2]. Recent data suggests that somatically active retrotransposons might mediate interference of transcription in neighboring genes, and in some cases this can lead to heritable epigenetic effects. Thus, one can observe variable gene-expressivity in the absence of genetic diversity in the classic sense – a phenomenon illustrated in isogenic agouti (A^{vy}) mice[11]. Such mice can display a variable pattern of yellow fur, obesity, diabetes, and so on, depending on the presence of an active intra-cisternal A particle retrotransposon. Whole genomic databases should significantly help to bring an understanding to these most decidedly non-Mendelian phenomena.

Impact on drug discovery

We have also completed the mouse genome, and rat and dog genome sequencing is currently underway. A number of novel genes have been discovered through this sequencing. These achievements, and the supporting computational biology that has been simultaneously developed, will drive the discovery of new diagnostics and pharmaceuticals in ways that were unimaginable even a few years ago. For the first time, we can utilize the reference DNA sequence for the entire human genome, and the entire set of protein coding genes that total ~30,000, a number smaller than expected. We will have an ever-growing anthology of genomic information from various model organisms that will be essential to modern pharmaceutical discovery and development, and eventually we will have the tools to understand how human complexity is reconciled with relatively small gene numbers.

Target discovery will be accelerated through: (1) interactive programs of protein-based analysis at scale; (2) proteomic analysis of cell compartments in tissues and standardized cell lines; (3) evaluation of post-translational modification and proteolytic processing profiles; (4) true exon-based RNA analysis; (5) DNA variation analyses; (6) high-throughput functional assays; and (7) predictive/molecular toxicology (toxicogenomics), including protein surrogate markers of adverse reactions and efficacy. Sophisticated computational biology tools now make possible a broad examination of gene classes and gene variations (polymorphisms), including the regulatory elements that govern the rate and tissue specificity of gene expression. Comparative genomics will enable significantly more-efficient prediction of gene structure and function and, perhaps more importantly, will enable better use of animal models to define and validate targets for drug development and predict the outcome of clinical trials. Understanding the full range of gene duplications might make it possible to anticipate unintended or 'non-specific' actions of what appear to be 'specific' therapeutic interventions.

Target discovery, lead compound identification, biochemical pharmacology, toxicology, exhaustive literature annotation, and clinical trials, can now be merged with the science of bioinformatics into a powerful and unified machine for discovering and developing new products. This unified process will provide new opportunities for rational small- and large-molecule design, including novel approaches for cancer vaccines, and the reduction of unexpected serious side effects. Fewer pharmaceuticals will be encumbered by 'black box' warnings on the product label or be the subject of market withdrawals.

Taken together, these advances will permit scientists who are versed in the new world of information and computation, to speed the identification of new agents, eliminating products that will probably display toxicity or poor efficacy, and reducing the formidable costs and risks associated with the current paradigms of drug discovery and development. The future is now.

References

1 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351

2 International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921

3 Adams, M.D. *et al.* (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252, 1651–1656

4 Adams, M.D. *et al.* (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* 377 (Suppl.), 3–174

5 Fleischmann, R.D. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–512

6 Adams, M.D. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287, 2185–2195

7 Cavalli-Sforza, L.L. (1998) The DNA revolution in population genetics. *Trends Genet.* 14, 60–65

8 Weber, W.W. (1997) *Pharmacogenetics*, Oxford University Press

9 Broder, S. and Venter, J.C. (2000) Sequencing the entire genomes of free-living organisms: the foundation of pharmacology in the new millennium. *Annu. Rev. Pharmacol. Toxicol.* 40, 97–132

10 Roses, A.D. (2000) Pharmacogenetics and the practice of medicine. *Nature* 405, 857–865

11 Whitelaw, E. and Martin, D.I.K. (2001) Retrotransposons as epigenetic mediators of phenotypic variation in mammals. *Nat. Genet.* 27, 361–365