Whole genomes: the foundation of new biology and medicine

Commentary

Samuel Broder and J Craig Venter

Our genomic DNA sequence provides a unique glimpse of the provenance and evolution of our species, the migration of peoples, and the causation of disease. Understanding the genome may help resolve previously unanswerable questions, including perhaps which human characteristics are innate or acquired. Such an understanding will make it possible to study how genomic DNA sequence varies among populations and among individuals, including the role of such variation in the pathogenesis of important illnesses and responses to pharmaceuticals. The study of the genome and the associated 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 post-translational modifications. For any given physiological process, we will have a new paradigm for addressing its evolution, development, function, and mechanism.

Address
Celera Genomics Corporation, 45 West Gude Drive, Rockville, MD 20850, USA

Current Opinion in Biotechnology 2000, **11**:581–585

0958-1669/00/$ – see front matter

Abbreviation
SNP single nucleotide polymorphism

Introduction

The science and technology for sequencing entire genomes of free-living organisms is a recent development in the history of genetics [1]. The first complete genome for any free-living organism (*Haemophilus influenzae*) was published by Venter and co-workers in 1995 [2], employing a strategy of random whole-genome 'shotgun' sequencing. It therefore became possible to sequence the entire genomes of prokaryotes with great rapidity and efficiency [3]. On the other hand, the feasibility of a comparable sequencing strategy for large, complex eukaryotic organisms was not widely accepted. Enthusiasm was curbed, firstly, by a lack of completely automated, high-throughput DNA sequencing machines and, secondly, by the overall computational resources and algorithms needed to process sequence information from large genomes with treacherously long stretches of repeats or duplicates that could confound the interpretation of the data.

We previously selected *Drosophila melanogaster* as a test system to explore the feasibility of broader whole-genome shotgun sequencing for large and complicated eukaryotic genomes. The decision to sequence the *Drosophila* genome was informed by the unique historical importance of this organism to biology and medicine and by the elegance of the science that has consistently characterized this field for nearly 100 years. At Celera Genomics, we were able to determine the nucleotide sequence of nearly all of the 120 megabase euchromatic of the *Drosophila* genome in collaboration with scientists at the Berkeley Drosophila Genome Project [4–6]. We successfully used newly developed and completely automated DNA-sequencing machines and the whole-genome shotgun sequence method enhanced by powerful assembly algorithms.

More recently, we determined the DNA sequence for the roughly 3.1 billion base pairs in the human genome, and we are nearing the completion of the mouse genome. Such DNA sequence information will probably alter biology and medicine in profound ways.

Whole-genome sequencing

The development of advanced automation, robotics, and computer software for industrial-scale DNA sequencing has proceeded at a remarkable pace. With the successful sequencing of the *H. influenzae* genome in its entirety [2], it became clear that the DNA of complex organisms many megabases in size could be accurately and rapidly sequenced by using a 'shotgun' sequencing strategy [7]. The genome of virtually any free-living organism can now be obtained in this way, providing there is an appropriate allocation of resources.

Generally, a single random DNA-fragment-library is prepared following mechanical or sonic shearing of the entire genomic DNA and inserted in suitable vector systems (e.g. plasmids). The ends of a large number of randomly selected fragments are sequenced from both insert ends until every part of the genome has been sequenced several times on average. For any given average sequence read-length, the number of end sequences needed can be determined by the Lander and Waterman application of Poisson statistics [2]. This number will depend on the goals of the sequencing project and particularly the degree of tolerance for a small number of gaps in the sequence results (i.e. if the tolerance for gaps is low, the number of end sequences must be high). The sequences are then computationally 're-assembled' to provide the complete genome. This strategy is not a theoretical construct or a curiosity useful only for certain small microbial genomics. Rather, the strategy has broad applicability for large, complex genomes.

When an organism reproduces sexually, and thus contains one genome from the mother and one genome from the father, this approach yields an important dividend. Points

of common DNA variation such as single nucleotide polymorphisms (SNPs) become evident even by sequencing one individual.

Whole-genome shotgun sequencing coupled with the advent of completely automated DNA-sequencing machines, such as the new ABI Prism 3700 DNA Analyzer (manufactured by Applied Biosystems, Applera Corp.), has made it possible for a single center to undertake the determination of the reference sequence of complex genomes, including that of the human. Indeed, one can consider that every organism of pharmacological, toxicological or agricultural interest is now plausibly a candidate for whole-genome sequencing. Murine, canine, porcine, caprine, bovine and simian genomes are now within reach. Disciplines not currently viewed as grounded in comparative whole genomics (e.g. organ transplantation and repair) will become so in the near future, because such disciplines will have suitable tools based on computational biology and genomes. Starting with the mouse, and radiating from there, the promise of comparative whole genomics and proteomics extends far beyond our current paradigm of biological and medical research [8].

The challenge of microbial pathogens

Knowing the complete genome of microbial pathogens will accelerate the commercial development of novel pharmaceuticals and biologics to treat pathogenic microbes. We will know how many genes are contained in each pathogen, where they are located within that pathogen's genome, and when two or more evolutionarily similar genes or genetic functions exist in a single microbial genome, thereby creating the potential to confound the search for effective cures. By simultaneously analyzing the microbial genome and the host genome, it will become possible to define those genes that are uniquely important for microbial survival in a given environmental context and those that are redundant or superfluous in that context.

It will also be possible to learn why a given pathogen is virulent in the context of a specific host, when toxic cytokines are activated within a host, whether a given pathogen has evolved proteins with molecular mimicry capable of frustrating host immunity or inducing autoimmunity, and how a pathogen (e.g. the tubercle bacillus) is able to survive in a state of latency, impervious to the host's immune system. The sum total of this information will make it possible to define medications or vaccines that induce specific and effective immunity against the pathogen while minimizing untoward or toxic side effects in the host. This will revitalize the entire private-sector antibiotic and vaccine industry.

Enablement of other data sets including a Universal Protein Library product family – proteomics

The complete genome sequence of an organism, especially the human, is a fundamental dataset that will inform a number of other datasets. Basic researchers will be able to frame new hypotheses for innovative experiments. Pharmaceutical companies will have at their disposal a body of interactive databases to sustain their drug discovery and development pipeline. The complete reference sequence of the human and murine genome will enable proteomics databases — every peptide sequence will have a corresponding sequence in the genome. A library of monoclonal antibodies against medically important epitopes will be available. In addition, it is now possible to analyze cDNA libraries on an unprecedented scale, as a result of having the genome sequence, creating new opportunities for gene discovery and thereby accomplishing the creation of new targets for drug action on a level never considered possible. Gene expression measuring technologies (including those based on microarrays and taq polymerase) will reach their full potential. Thus, for example, commercial high-density gene microarray and similar technologies will contain all human genes, and it will be possible to monitor every splice variant. Comparison of expression of genes between different species will be possible.

Expressed sequence tag (EST) coverage can provide expression data for the more abundant genes. However, the genomic sequence will need to be the starting point for rare genes, precisely those very difficult-to-find genes essential for major advances in basic and applied research. Technologies based on highly reproducible cDNA segmentation (e.g. GeneTag®) will permit very clear profiles of tissue-specific gene expression, and will particularly enable the study of various biological factors on gene expression in the human and model organisms (such as the mouse or the rat). This will eventually make it possible to predict toxic side effects of potential new drugs even before the stage of advanced preclinical toxicology testing. It will also be possible to study tissue-specific promoters and enhancers, providing exciting new targets for small molecules and enhancing the safety and efficacy of gene therapy. The full promise of gene therapy is unlikely ever to be fulfilled without this information.

The research community can look forward to a future in which whole genome datasets provide a foundation for a Universal Protein Library (Proteomics) that will contain the following components.

A proteome index for each key model or target organism, including humans

A dataset of curated gene products (including one example for each identifiable gene product of an organism) will be available. This will encompass a full description of the role and function of the protein product, transcript size and variation, and related clinical information. Closely correlated with the proteomic data, researchers will also eventually be able to access databases that coherently describe complex and potentially confusing post-transcriptional phenomena, such as RNA editing, epigenetic control such as the silencing of gene expression according to parent-of-origin effects (imprinting), post-translational peptide modifications, and other related variables that affect the

size, nucleotide content, or overall synthesis rate of transcripts beyond an analysis of classic genomic sequence.

Paralog datasets

Paralogs are genes that share a common evolutionary history, found within a given organism, in effect through a process of gene duplication. A collection of data encompassing protein families and subfamilies for each key organism will be available, using sequence-clustering tools. Various researchers will, in effect, be able to organize and analyze subsets of gene products, and specifically validated targets, from the proteome index described above, essentially at the touch of a finger (or perhaps more accurately, at the touch of a mouse). An understanding of paralogs can lead to a deeper understanding of the biological 'redundancy' that affects the development of new pharmaceutical agents, particularly in the microbial world. Better insight into the evolution of specified protein domains will have tremendous implications for science.

Ortholog datasets

Orthologs are genes that share a common evolutionary history in two or more organisms. A curated dataset of functional orthologs among different organisms will be available now that the reference human genome is available. Researchers will be able to quickly learn the function and biology of a protein from what is known about othologous proteins. A related concept is synteny, in which genes that exist within chromosomal segments sharing a common evolutionary history can be quickly identified and analyzed for genotype–phenotype correlations (see below).

Synteny datasets

High-resolution syntenic maps between the human and mouse, or other model organisms, will be available to accelerate queries between mapped genes in two or more organisms. The shared evolutionary history of chromosomal fragments between mammalian species will be an invaluable foundation for the assignment of function, and also for validation of pharmaceutical targets for newly discovered genes. Synteny datasets will permit computer-assisted overviews of genes in important chromosomal regions, including efficient analyses of relevant gene knockout effects and other genotype to phenotype correlations. Synteny databases will allow model organisms to 'teach' us how to select genes for further study, including site-directed mutagenesis and gene knockout research. Our preliminary results suggest that there is a higher degree of conservation both within genes and outside conventional genetic regions when one compares the human and mouse genomes, than perhaps expected. The mouse genome will become even more of a fundamental resource for biology and medicine.

Pathway datasets

The scientific community will have at its disposal pathways of metabolic and biologic processes. Linkage of complex pathway information to the diagnosis, prevention, or treatment of important illnesses will become an increasingly powerful tool for creating new diagnostic or drug pipelines. In effect, it will eventually be possible to determine computationally how a gene or protein product influences pathways on a broad front, and how subtle changes in a given pathway affect other distant pathways, in ways that would otherwise be overlooked. It will also be possible to analyze how genes that are strictly speaking not targets for a new drug may contribute to the efficacy or toxicity of the molecule.

Protein interaction data

Databases providing information about empirically proven or theoretically deduced protein–protein interactions will be available. This will be an invaluable tool for new drug design, and may substantially decrease the need for large-scale, empirical drug-screening strategies.

Expression pattern datasets

Databases of gene products having comparable expression patterns will be available. Thus, information using any of several gene expression technologies available now, or in the near future, could be available for rapid analysis and correlations. Tissue expression data obtained from various cDNA libraries in the gene index databases could be readily integrated to design experiments (new drugs as applied variables) and select candidate new agents. Genes that are expressed in coordinated ways in response to defined applied variables could provide a valuable tool for identifying regulatory elements.

Primary nucleotide sequence to tertiary protein structure algorithms

Eventually, the data derived from the reference human genome sequence could be transformed into direct information about the three-dimensional structure of encoded proteins using defined algorithms. To be sure, this particular development is not near, but the availability of whole-genome and proteomic information from several organisms will probably speed the arrival of such algorithms. This could open up a new era in computational drug design based directly on genome sequence.

DNA sequence variation

DNA variation is important to biology and medicine. Indeed, the individuality of members of any complex species arises from the interplay of genetic variation with the environment. The most common form of DNA variation is single nucleotide polymorphisms (SNPs) [9–21]. Put simply, a SNP is the substitution of one purine or pyrimidine nucleotide at a given location in a strand of DNA for another purine or pyrimidine nucleotide. Such substitutions can affect gene function, or they can be neutral. Neutrality is generally inferred if a SNP does not alter protein coding. In practice, this inference can be very wrong, indeed.

SNPs may occur inside or outside of a gene. If they occur within a gene, they may reside in an exon (coding region) or intron

(non-coding) region. SNPs in a coding region (sometimes called cSNPs) can either be synonymous (no amino-acid altering effect) or non-synonymous (amino-acid altering). There is some level of natural selection against amino-acid altering changes [19,20]. The average person would be expected to be heterozygous for roughly 40,000 non-synonymous (amino-acid altering) alleles. In any event, it is very inaccurate to conclude that only SNPs within coding exons can play a direct role in the pathogenesis of important diseases.

SNPs can profoundly affect gene function even if they are at a significant distance upstream of the initiation site for gene transcription. In a recent example, sequence variations in the 5′-flanking region of the leptin gene predicted obesity in women [22]. One needs to keep in mind that enhancers (i.e. the tissue-specific control sequences upon which certain regulatory substances act) may operate over at least 3 kb in either orientation (5′→3′ or 3′→5′) from the start point of transcription. This re-emphasizes the importance of whole genomic information for the future of biological and medical research.

The classic Mendelian model, in which a specific mutation in one gene produces a recognizable disease, may not apply to most common illnesses in our society. It is thought that many common illnesses have a polygenic origin, with several genes (to be more precise, gene variants) playing a comparatively small role individually but with a cumulative effect that leads to a detectable clinical condition or disease. There is considerable interest in using whole-genome association studies, as a tool for identifying genes involved in these common disorders, to detect differences in the frequency of DNA sequence variations between unrelated affected individuals and a control group [23]. It is worth stating that there are still no easy algorithms for interpreting complex genetic interactions and polygenic multi-variant diseases. The tools of industrial scale re-sequencing coupled with powerful computers and software may, however, provide new strategies for solving these problems in the future.

One must keep an open mind in using genetic variation to define new genes and new pathways. Thus, Horikawa *et al.* [24] found that a new gene, *Calpain 10* (CAPN10), not previously known to be related to the regulation of blood glucose concentrations, is associated with Type 2 diabetes mellitus. Moreover, there was a very complex relationship between susceptibility to Type 2 diabetes and SNPs involving *Intron 3*. No existing paradigm for understanding diabetes can easily account for this level of complexity between SNPs in this gene and the phenotype of Type 2 diabetes. Similar considerations apply to many diseases; however, we predict that the new knowledge in genomics will reduce this complexity in ways that significantly improve prevention, diagnosis, and treatment.

Conclusion

Biology and medicine now have the fruits of whole genomics to address many complex problems. Proteomic indices, paralog datasets, ortholog datasets, synteny, computational-metabolic pathway analyses, protein interaction databases, gene expression profiles and SNP databases will fundamentally change how scientists pursue the vital challenge of new and better drugs. This growing knowledge, based on the foundation of the complete human genome sequence, will make it possible for scientists to generate knowledge on a scale previously unimaginable.

Acknowledgements

We wish to thank Beth Hoyle for her excellent editorial assistance in preparing this manuscript.

References

1. Broder S, Venter JC: **Sequencing the entire genomes of free-living organisms: the foundation of pharmacology in the new millennium.** *Annu Rev Pharmacol Toxicol* 2000, **40**:97-132.
2. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb J-F, Doughtery BA, Merrick JM *et al.*: **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd.** *Science* 1995, **269**:496-512.
3. The Institute For Genomic Research website at http://www.tigr.org/
4. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF *et al.*: **The genome sequence of *Drosophila melanogaster.*** *Science* 2000, **287**:2185-2195.
5. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA *et al.*: **A whole-genome assembly of *Drosophila.*** *Science* 2000, **287**:2196-2204.
6. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W *et al.*: **Comparatative genomics of the Eukaryotes.** *Science* 2000, **287**:2204-2215.
7. Venter JC, Adams MD, Sutton GG, Kerlavage AR, Smith HO, Hunkapiller M: **Shotgun sequencing of the human genome.** *Science* 1998, **280**:1540-1542.
8. O'Brien SJ, Menotti-Raymond M, Murphy WJ, Nash WG, Wienberg J, Stanyon R, Copeland NG, Jenkins NA, Womack JE, Marshall Graves JA: **The promise of comparative genomics in mammals.** *Science* 1999, **286**:458-480.
9. Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengård J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF: **Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase.** *Am J Hum Genet* 1998, **63**: 595-612.
10. Kittles RA, Perola M, Peltonen L, Bergen AW, Aragon RA, Virkkunen M, Linnoila M, Goldman D, Long JC: **Dual origins of Finns revealed by Y chromosome haplotype variation.** *Am J Hum Genet* 1998, **62**:117-119.
11. Harris EE, Hey J: **X chromosome evidence for ancient human histories.** *Proc Natl Acad Sci* 2000, **96**:3320-3324.
12. Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengård J, Salomaa V, Vartiainen E, Boerwinkle E, Sing CF: **DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene.** *Nat Genet* 1998, **19**:233-240.
13. Hästabacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E: **Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland.** *Nat Genet* 1992, **2**:204-211.
14. Kruglyak L: **The use of a genetic map of biallelic markers in linkage studies.** *Nat Genet* 1997, **17**:21-24.
15. Risch N, Merikangas K: **The future of genetic studies of complex human diseases.** *Science* 1996, **273**:1516-1517.
16. Shaw SH, Carrasquillo MM, Kashuk C, Puffenberger EG, Chakravarti A: **Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes.** *Genome Res* 1998, **8**:111-123.

17. McKeigue PM: **Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning or parental admixture.** *Am J Hum Genet* 1998, **63**:241-251.

18. Weiss KM: **In search of human variation.** *Genome Res* 1998, **8**:691-697.

19. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Lane CR, Lim EP, Kalayanaraman N, Nemesh J *et al.*: **Characterization of single-nucleotide polymorphisms in coding regions of human genes.** *Nat Genet* 1999, **22**:231-238.

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

21. Landegren U, Nilsson M, Kwok PY: **Reading bits of genetic information: methods for single-nucleotide polymorphism analysis.** *Genome Res* 1998, **8**:769-776.

22. Li W-D, Reed DR, Lee JH, Xu W, Kilker RL. Sodam BR, Price RA: **Sequence variants in the 5′ flanking region of the leptin gene are associated with obesity in women.** *Ann Hum Genet* 1999, **63**:227-234.

23. Kruglyak L: **Prospects for whole-genome linkage disequilibrium mapping of common disease genes.** *Nat Genet* 1999, **22**:139-144.

24. Horikawa Y, Oda N, Cox NJ, Li X, Orho-Melander M, Hara M, Hinokio Y, Lindner TH, Mashima H, Schwarz PE *et al.*: **Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus.** *Nat Genet* 2000, **26**:163-175.