The human genome: A first look at the blueprint of human inheritance

S. Kumar Singh

The announcement of the successful completion of the human genome project by two groups – the Human Genome consortium and Celera Genomics, on 26 June 2000 marks a watershed in modern day biological research. Within 24 hours of its announcement, parallels have already been drawn between this extraordinary achievement and man's landing on the moon and the invention of the wheel. Basic and applied research in human medicine and welfare are never going to be the same again. Although cynicism and criticism about the usefulness of the finished sequence and the methods adopted abound, there remains no doubt that the knowledge of the complete genetic template of *Homo sapiens* will be a major step forward in the study of human biology.

History

The fascinating story began in the mid 1980s after Fred Sanger had successfully demonstrated in 1977, that double-stranded DNA could be sequenced using the dideoxy sequencing technique[1]. During the mid 80s there was considerable excitement both in scientific as well as public circles about the possibility of sequencing the whole genome of humans. The practicality of such an undertaking was discussed at two meetings, one at Santa Cruz, USA in 1985 (ref. 2) and another a year later at Santa Fe[3]. Realizing the potential of such an effort, two Government agencies of the United States – Department of Energy (DOE) and the National Institute of Health (NIH) in 1988, drafted the first proposal to 'coordinate research and technical activities related to the human genome'. Two major technical advances that made the project feasible were the development of yeast artificial chromosome (YAC) vectors[4] and fluorescence based detection of dideoxy terminated fragments[5] that enabled near-total automation. The project began in earnest in 1990 when the projected goals for the first five years of a fifteen year plan were published[6]. The progress report of that period and final projections was announced five years later in 1998 (ref. 7). The project evolved over the 90s decade and in its final stages had over 16 international centres involved globally (Table 1) in the task of sequencing all the human chromosomes (Figure 1).

The major landmarks during this process were the publication of the human genetic map[8], which was soon fol-

Table 1. Centres involved in the human genome sequencing project. Celera Genomics Inc. independently completed the sequencing

Baylor College of Medicine, Houston, Texas, USA
Genoscope, Evry, France
Keio University, Tokyo, Japan
RIKEN Genomic Sciences Center, Saitama, Japan
University of Washington Genome Center, Seattle, WA, USA
Gesellschaft für Biotechnologische Forschung mbH, Braunschweig, Germany
Washington University Genome Sequencing Center, St. Louis, MO, USA
Institute for Molecular Biotechnology, Jena, Germany
Stanford DNA Sequencing and Technology Development Center, Palo Alto, CA, USA
University of Washington Multimegabase Sequencing Center, Seattle, WA, USA
Genome Therapeutics Corporation, Waltham, MA, USA
The Sanger Centre, Hinxton, UK.
Max Planck Institute for Molecular Genetics, Berlin, Germany
Beijing Human Genome Centre, Institute of Genetics, Chinese Academy of Sciences, Beijing, China
Whitehead Institute for Biomedical Research, MIT, Cambridge, MA, USA
Joint Genome Institute, U.S. Department of Energy, Walnut Creek, CA, USA

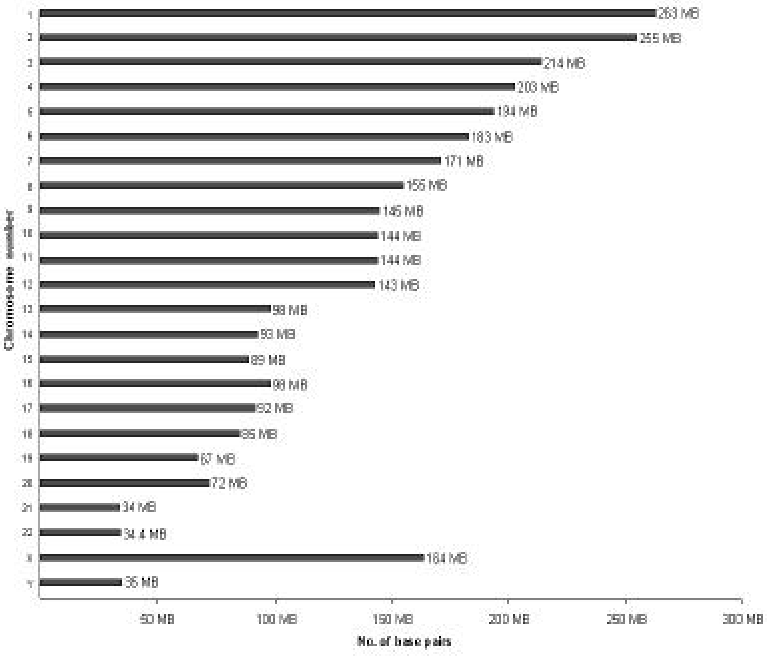

Figure 1. Size distribution of the human chromosomes.

lowed by a sequence tagged sites (STS) based map[9] and then by a transcript map[10]. A breakpoint map highlighting the chromosomal rearrangements that recur frequently[11] and an STS-based radiation hybrid map[12] appeared soon thereafter. A year later Dekoulas *et al.*[13] presented the physical map of the human genome. The first complex locus of substantial size to be successfully mapped and sequenced was that of the Major Histocompatibility Complex present on chromosome VI (ref. 14). The official countdown began late last year with the announcement of the completion of the sequencing of the human chromosome 22 (refs 15, 16). This was followed by the completion in May this year of human chromosome 21 (ref. 17). The rapidity of the whole process can be gauged from the announcement made by the consortium that almost 60% of the genome was sequenced in the past 6 months at the rate of 1000 bases per second! The final draft was compiled from a total of 22 billion bases covering the total human DNA sequence almost 7 fold. The sequencing effort of the publicly-funded international consortium was catalysed dramatically by the entry of a private company, Celera Genomics Inc. into the genome sequence arena. Founded by Craig Venter, the pioneer of the shot-gun sequencing technique, Celera aimed to complete the project within 10 months, between September 1999 and June 2000. The intense race apparently ended in a politically brokered tie, with the two groups jointly announcing the completion of the 'rough draft' on 26th of June 2000. 86% of the total human genome has been evaluated and classified to be highly accurate (average accuracy of 99.9%). Sequencing without evaluation is complete for 97% of the genome.

Whose genome was sequenced?

The most popular answer to this question among genomic intelligentsia appears to be 'it doesn't matter', because more than 99.8% of the 3.2 billion basepairs (bp) between any two humans would be the same. More strikingly (some might not consider it so), we are more than 98% similar to chimpanzees. While sounding evasive, it brings to the fore the scientific question of the importance of the 'other' fraction, which is responsible for all the 'personal' differences. These differences are the major factors that determine everything from the tinge in the colour of our skin to predisposition of certain individuals to dreaded diseases like cancer and Parkinsonism. Hence, it is not surprising that it is this feature which has attracted so much attention among pharmaceutical ventures and is fast becoming the commercial El Dorado of the post genome era.

Indeed, much of the effort in this endeavour has been to map the loci of most of these disease-causing genes so that their genetic underpinnings can be better elucidated. One major effort that needs mention here is the report of painstakingly mapping the numerous chromosomal breakpoints, that are involved in various forms of cancer by Mitelman[11]. Unfortunately, there are still many aspects of innumerable commonly-occurring genetic diseases afflicting mankind that have remained elusive so far. This situation is likely to change dramatically with the availability of the complete sequences of all the chromosomes.

In reality, the 'substrate' for sequencing is obtained by an elaborate process involving collecting DNA samples (blood from female donors and sperm from males) from multiple anonymous donors in strict accordance to the international review board (IRB) protocol[18]. After screening, a few of the samples (~20–50 individuals) are combined and used for library construction followed by sequencing. All through the process total anonymity is maintained about the identity of the donors.

The methodology of sequencing

One of two methodologies has been employed for all the completed genome projects (both prokaryotes and eukaryotic) so far. In the 'Top-down approach', the genome is segregated into smaller segments in a step-wise manner and when the pieces are small enough, they are sequenced after being inserted into vectors. After sequencing, these individual fragments are pieced together by backtracking until their point of origin in a given chromosome is reached. This method has an advantage of modularity, in the sense that any given fragment of a chromosome could be sequenced separately and the whole assembly could be done independent of other regions. This was the approach adopted by the sequencing consortium in its initial stages. Generating genetic maps by plotting the crossover frequencies among different gene loci is the first step in stepwise downsizing of each chromosome. The result is a map with many signposts along the path of the chromosome, which serve as markers to identify a given fragment and to determine the order of different fragments. The lineage of each fragment is tracked downwards up to a point where the fragment is amenable to be sequenced. Arrangement of the sequenced regions is then straightforward.

The bottom-up approach, popularly called as the shotgun procedure[19], involves breaking the genomic DNA into small fragments and sequencing all of them in an unbiased brute force manner. Because of the degeneracy of this process (the total fragments generated typically covers the genome 10 times its original size), many overlaps are present among the fragments. This feature of overlap is exploited in the assembly phase after the sequencing phase is complete. All the sequenced fragments are assembled in one mammoth computational exercise, by matching identifying pairs of sequences among any two fragments. Although prone to artefacts, especially in repeat regions, this procedure has been successfully applied for the sequencing of almost all the microbial genome sequences published so far[20]. This has been made possible mainly due to the increase in accuracy of the algorithms performing the assembly.

In either approach, the actual sequencing is performed in a manner not much different from the dideoxy method proposed by Fred Sanger in 1977. Presently, most of the sequencing is done by specially designed high-speed sequencers, which require little human intervention and have very high throughput. Each of the four dideoxy terminators is tagged with different fluorescence markers, which are read by automated instruments.

The tools of the trade

As the sequencing of a given fragment (inserted into a plasmid, cosmid, YAC, BAC, or PAC) is completed, it is subjected to a set of computational procedures, which align the fragments, do error checking and detect potential protein coding regions. Once a region is predicted to be a coding segment, the features of the gene are analysed (e.g. presence of specific signatures, CpG islands, etc.) and similarity searches run to establish their identity. Some of the programs that are commonly used are listed in Table 2. As can be easily made out, it is an exercise that is highly computation intensive. Celera Genomics Inc. announced that during the assembly involved more than 500×10^{18} (500 million trillion) calculations, making it the biggest exercise in the history of computational biology.

Table 2. Commonly used software tools for sequence assembly, prediction of coding regions, annotation, detecting repeats and splice sites

TRNASCAN	Detects for RNA coding regions
FGENEH	A dynamic programming algorithm that uses linear discriminant functions. Used to identify the presence of genes
GENSCAN	Prediction program for coding regions and exon/intron splice sites
Genquest	Program for sequence assembly, analysis and comparison
MUMmer	A whole genome alignment tool
REPEATMASKER	A repetitive element filter that screens a sequence against a library of repetitive sequences and searches for low complexity regions
Annotator	An interactive genome annotation tool
PSI-BLAST and variants	The most commonly used program for homology searches
Genefinder	Popular tool to detect coding regions and splice sites
Grail XGRAIL/ Grailcnt	Predicts exons, putative genes, detects promoter regions, poly As, CpG islands, similarities in ESTs, and repetitive segments
Net plant gene	Predicts exon/intron splice sites
AAT	Analysis and annotation tool

Table 3. A listing of some of various disease genes that have been mapped on to specific sites on the chromosomes. (Data compiled from genome database; Cancer genome anatomy project website, and NCBI's genome resource site)

Chromosome	Information
Chromosome 1	Chediak–Higashi Syndrome; Charcot-Marie-Tooth Neuropathy-2a; Ductal Breast Cancer; Gaucher disease; Usher Syndrome, type 2.
Chromosome 2	Alstrom Syndrome; Holoprosencephaly 2; Tibial Muscular Dystrophy; Autosomal Recessive Deafness-9; Limb Girdle Muscular Dystrophy 2b.
Chromosome 3	Alkaptonuria; Cornelia de Lange Syndrome; von Hippel-Lindau Syndrome; Acute Myeloid Leukemia
Chromosome 4	Huntington Disease; Juvenile Periodontitis; Wolf Hirschhorn Syndrome; Wolfram Syndrome
Chromosome 5	Cri du chat; Colorectal Cancer; Diastrophic Dysplasia; Treacher Collins Syndrome; Spinal Muscular Atrophy
Chromosome 6	Celiac Disease; Hemochromatosis; Lafora Myoclonus Epilepsy; Spinocerebellar atrophy1.
Chromosome 7	Cystic Fibrosis; Pallister-Hall Syndrome; Pendred Syndrome; Split Hand/Foot Malformation,Type 1 Williams-Beuren Syndrome
Chromosome 8	Lipoprotein lipase: Human;Cohen Syndrome; Langer-Giedion Syndrome; Nijmegen Breakage syndrome; Werner Syndrome
Chromosome 9	Dysautonomia; Friedreich Ataxia; Torsion Dystonia 1; Fanconi Anemia, Type C; Nail-Patella syndrome
Chromosome 10	Cowden Disease; Jackson-Weiss Syndrome; Hermansky-Pudlak Syndrome; Wolman Disease
Chromosome 11	Ataxia-telangiectasia; Beckwith-Wiedemann Syndrome; Bardet-Biedl Syndrome 1; Long QT Syndrome 1; Wilms Tumour Type 1
Chromosome 12	Darier Disease; Dentatorubro-pallidoluysian Atrophy; Noonan Syndrome ; Familial Periodic Fever; Phenylketonuria.
Chromosome 13	Retinoblastoma; Rieger Syndrome, Type 2; Stargardt Disease-2; Wilson Disease.
Chromosome 14	Alzheimer Disease; Graves Disease; Machado–Joseph Disease; Spastic Paraplegia 3A.
Chromosome 15	Amytrophic Lateral Sclerosis 5; Bloom Syndrome; Prader-Willi Syndrome; Tay-Sachs Disease
Chromosome 16	Familial Mediterranean Fever; Fanconi Anemia, Type A; Inflammatory Bowel Disease 1; Polycystic Kidney Disease, adult 1.
Chromosome 17	Breast Cancer 1; Canavan Disease; Miller-Dieker; Lissencephaly; Neurofibromatosis type 1;TP53, Tumor Suppressor Gene.
Chromosome 18	Colorectal Cancer; Holoprosencephaly 4; Niemann-Pick Disease, Type C; Tourette Syndrome
Chromosome 19	Acute T-Cell Leukemia; Diamond-Blackfan Anemia; Myotonic Dystrophy
Chromosome 20	Alagille Syndrome; Corneal Dystrophy, polymorphous SCID, due to ADA deficiency.
Chromosome 21	Holoprosencephaly 1; Trisomy 21 (Down Syndrome); Usher Syndrome, Type1E.
Chromosome 22	The smallest chromosome. DiGeorge syndrome; Cat eye syndrome; Ewing sarcoma; Neurofibromatosis, type 2; Velocardiofacial syndrome
Chromosome X	Adrenoleukodystrophy; Duchenne Muscular Dystrophy; Lowe Syndrome; Norrie Disease; Rett Syndrome.
Chromosome Y	Gonadal dysgenesis; mobius syndrome; prostrate cancer; Adenocarcinoma; Acute myeloid leukemia.

The number of genes

The sizes and a few characteristics of each of the 22 autosomes and the sex chromosomes that are present in normal human cells are listed in Table 3. The list is far from comprehensive and contains only the better-known gene loci. The most popular question often asked has been on the total number of genes that are likely to be coded. After long speculation, realistic numbers are beginning to emerge. Three recent reports are an indication of the unpredictability of the exercise. Two groups[21,22] predict by independent methods, the total number of genes likely to be present, in the range of 33,000–44,000, while another group[23], after analysing the gene indices deposited in the TIGR sequence database arrive at a number in the range of 120,000. Similar reports a few years ago, had suggested numbers ranging from 60,000 to 90,000. Early results from the completed sequence released by the consortium suggest it to be around 38,000. A large proportion of these genes are unclassified and have not been reported before. Till the end of June 2000, the OMIM (Online Mendelian Inheritance in Man)[24] database had 11,741 gene loci mapped onto specific sites on the human chromosomes. It is worthwhile to compare these numbers with those of other organisms both prokaryotes and eukaryotes whose complete genome sequence is available (Table 4).

Table 4. A comparison of the genome sizes and the number of genes present in different organisms sequenced so far

Organism	Genome size (MB)	No. of genes
Haemophilus influenzae	1.83	1,740
Helicobacter pylori	1.66	1,590
Methanococcus janaschii	1.66	1,680
E. coli	4.6	4,288
Yeast (*Saccharomyces ceriviseae*)	13	6,000
Bacillus subtilis	4.2	4,100
Drosophila	139	13,601
Caenorhabditis elegans	97	19,099
Arabidopsis thaliana	100	27,000
Human	3200	33,000–150,000

Implications

Knowing the sequential arrangement of the four letter alphabets that constitutes the genetic language, from one end to the other of all the chromosomes, by itself means very little. Yet, it is this array of sequence that decides the proper co-ordination, communication, and functioning of the 10^{14} (100 trillion) cells that combine to govern our health and well being in its entirety. But to even assume that the knowledge of the complete DNA sequence will provide instant solution to a few of the many biomedical problems would be unwarranted at this stage; even a bit naive. The sequence information will form a starting point at best for scientific investigations. Biochemical phenomena, like mechanistic and molecular details of signal transduction, differential expression of gene products in various tissues during normal development and growth, uncontrolled growth in tumorous tissues among others will still require experimentation in the classical way. Indeed, genomics has little to offer in terms of understanding of phenomena like functioning of the brain during retention of short and long-term memory and human intelligence. The most direct and immediate impact of the availability of the sequence is likely to be in the area of molecular biology and biochemistry where the already well-established protocols of gene cloning and expression will result in rapid characterization of any gene and its protein product once it has been identified.

One important issue that scientists are likely to face (probably already are) is the problem of finding relevant data without getting lost in the digital amazon. This can be a formidable exercise for any scientist uninitiated in the field of 'internet biology'. Currently, many research groups maintain specialized databases of the rapidly generated raw sequence data with varying degrees of user friendliness. A few examples of such databases include mutation databases for P53, Willebrand factor disease, Cystic fibrosis, androgen receptor, factor VII, LDL receptor, *RB1* gene and others. A listing of all such databases can be found at the National Centre for biotechnological Information[24]. These specialized databases are likely to become more comprehensive and many more new ones are likely to appear in the immediate future (see Table 5 for a listing of genome-related websites). Standardized protocols for annotation of the newly-generated sequences and accessing these databases by web-based front end interfaces like browsers is currently an area of intense activity.

Ethical issues

One might be tempted to believe that this genetic blueprint might be the closest mankind has come to grasping the Holy Grail. If current trends are anything to go by, the technologies of the not-so-distant future might well enable mankind to tinker with DNA to the extent of playing God. Humanity has to face the situation of being responsible for tasks, which will increasingly put his own survival as a species at risk. More importantly, there are the lingering questions on the privacy of individual genetic data, fairness in evaluation of genetic facts, consent before use of such data, exploitation of rare genes in individuals, patenting of genetic material, conduct in reporting results in population genetics, genetic testing of modified DNA on individuals among others.

Conclusion

The success of the human genome project has been rightly called as the end of the beginning. It is likely to open many new areas in human biomedical research. Many important insights are likely to be gained in the coming months when the full ramifications of this work become apparent and we begin to understand what it is that we are actually made of. Although the raw text is in place and in the correct order, the grammar and semantics has still to be worked out which will ultimately lead to understanding of its meaning. Even

Table 5. Listing of major genome-related websites

The Sanger Center	http://www.sanger.ac.uk/Info/Intro/
The National Human Genome Research Institute	http://www.nhgri.nih.gov/
The National Centre for Biotechnology Information	http//ncbi.nlm.nih.gov/
The Genome Database.	http://gdbwww.gdb.org/
UK Human Genome Mapping Project Resource Centre (HGMP-RC)	http://www.hgmp.mrc.ac.uk/
The Euchromatin Network	http://www.ncbi.nlm.nih.gov/genemap99/
Primary resource center for German Human Genome Project	http://www.rzpd.de
Human Genome Project Information	http://www.ornl.gov/hgmis/
The Virtual Genome Center	http://alces.med.umn.edu/VGC.html
Kyoto encyclopaedia of genes and genomes	http://www.genome.ad.jp/kegg/
The Human genome catalog	http://genome.ornl.gov/GCat/humanmodel.shtml
Online Mendelian Inheritance in Man (OMIM) A catalog of disease genes and disorders	http://www.ncbi.nlm.nih.gov/omim
The human tumour gene index	http://www.ncbi.nlm.nih.gov/CGAP/hTGI/
The Cancer Chromosome Aberration Project (CCAP)	http://www.ncbi.nlm.nih.gov/CCAP/
The Genetic Annotation Index (GAI): A repository of all polymorphisms associated with cancer	http://lpg.nci.nih.gov/GAI
DbEST The expressed sequence tag database	http://www.ncbi.nlm.nih.gov/dbEST/
UNIGENE A non-redundant set of gene clusters Each cluster has sequences that represent a unique gene	http://www.ncbi.nlm.nih.gov/UniGene/
Celera Genomics. The company that independently sequenced the human genome	http://www.celera.com

by conservative estimates, this is likely to take several years.

Even a monthly visit to this fascinating world is going to be filled with surprises because of the rapidly changing landscapes and signposts. Identification of new disease genes will provide the starting point for development of diagnostic kits, which exploit the subtle differences and abnormalities in these genes. There is definitely a new ray of hope for millions of people who suffer from diseases whose underlying causes are yet to be elucidated at the molecular level. Although much needs to be done in many of these cases, the genome sequence provides a well-illuminated and fertile hunting ground for searching for the culprit genes. As for the present, it is a time to celebrate, now that we know a little more about ourselves.

1. Sanger, F., Nicklen, S. and Coulson, A. R., *Proc. Natl. Acad. Sci. USA*, 1977, **74**, 5463–5467.
2. Sinsheimer, R. L., *Genomics*, 1985, **5**, 954–956.
3. Summary Report of the Santa Fe Workshop, US Department of Energy, Office of Health and Environmental Research, 3–4 March 1986.
4. Burke, D. T. *et al*., *Science*, 1987, **236**, 806–812.
5. Smith, L. M. *et al*., *Nature*, 1986, **321**, 674–679.
6. Collins, F. and Galas, D., *Science*, 1993, **262**, 43–46.
7. Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R. and Walters, L., *Science*, 1998, **282**, 682–689.
8. Dib, C. *et al*., *Nature*, 1996, **380**, 152–154.
9. Hudson, D. J. *et al*., *Science*, 1995, **270**, 1945–1954.
10. Schuler, G. D. *et al*., *Science*, 1996, **274**, 540–546.
11. Mitelman, F. *et al*., *Nature Genet*., 1997, suppl. 417–474.
12. Stewart, E. A. *et al*., *Genome Res*., 1997, **7**, 422–433.
13. Dekoulas, P. *et al*, *Science*, 1998, **282**, 744–746,
14. The MHC sequencing consortium, *Nature*, 1999, **401**, 921–923.
15. Little, P., *Nature*, 1999, **402**, 467–468.
16. Dunham, I. *et al*., *Nature*, 1999, **402**, 489–495.
17. Hattori, M. *et al*., *Nature*, 2000, **405**, 311–319.
18. Executive Summary of Joint NIH-DOE Human Subjects Guidelines, 1996, US National Center for Human Genome Research, and US Department of Energy (NCHGR-DOE) Guidance on Human Subjects Issues in Large-Scale DNA Sequencing. *http://www.ornl.gov/TechResources/ Human_Genome/archive/nchgrdoe.html.*
19. Favello, A., Hillier L, Wilson R. K., *Methods Cell Biol.*, 1995, **48**, 551–569.
20. *http://www.tigr.org* and *http://www.ncbi.nlm.nih.gov/genome* have a comprehensive listing of all the microbial genomes completely sequenced so far.
21. Ewing, B., and Green, P., *Nature Genet*., 2000, **25**, 232–234.
22. Weissanbach, J. *et al*., *Nature Genet*., 2000, **25**, 235–238.
23. Quackenbush, J. *et al*., *Nature Genet*., 2000, **25**, 239–240.
24. *http://www.ncbi.nlm.nih.gov/Omim* McKusick, V. A., Mendelian Inheritance in Man. Catalogs of Human Genes and Genetic Disorders. Johns Hopkins University Press, Baltimore, 1998, 12th edition.

S. Kumar Singh is in the Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India
e-mail: skumar@mbu.iisc.ernet.in