Review Article

Indian J Med Res 117, February 2003, pp 43-65

The human genome sequence: impact on health care

M.D. Bashyam & S.E. Hasnain

Centre for DNA Fingerprinting & Diagnostics, Hyderabad, India

Received January 23, 2003

The recent sequencing of the human genome, resulting from two independent global efforts, is poised to revolutionize all aspects of human health. This landmark achievement has also vindicated two different methodologies that can now be used to target other important large genomes. The human genome sequence has revealed several novel/surprising features notably the probable presence of a mere 30-35,000 genes. In depth comparisons have led to classification of protein families and identification of several orthologues and paralogues. Information regarding non-protein coding genes as well as regulatory regions has thrown up several new areas of research. Although still incomplete, the sequence is poised to become a boon to pharmaceutical companies with the promise of delivering several new drug targets. Several ethical concerns have also been raised and need to be addressed in earnest. This review discusses all these aspects and dwells on the possible impact of the human genome sequence on human health, medicine and also health care delivery system.

Key words Ethical issues - human genome - sequencing - significance

Man's unending quest for knowledge and to explore and unfold the mysteries of nature has resulted in several important discoveries that have formed the basis for several remarkable achievements. Scientists have now undertaken yet another path breaking endeavour to define the innermost reaches of the human cell; to unravel the complete information encoded by the human genome, under the umbrella of the Human Genome Project. A private funded effort led by Craig Venter of Celera Genomics has simultaneously been able to decode the human genome sequence. These efforts have now given birth to new endeavours to understand what this sequence means in terms of human health, disease and development. The project has also facilitated rapid advances in the field of engineering and informatics. The availability of the human genome sequence will have a tremendous impact on biomedical sciences and the way medicine is practiced. We can now start asking questions pertaining to human health and disease, how variations in the genome result in disease. The human genome sequence will have (and is already having) a tremendous impact on understanding genetic disorders. The unraveling of genetic loci/genes responsible for important genetic disorders in the pre-genomic era was very slow. Out of the more than 6000 recorded inherited disorders only 75 were traced back to their genes using positional cloning and other approaches. The physical map of the human genome has assisted directly in identifying about 100 more disease causing genes and continuing efforts will possibly identify more disease related loci. Most inherited diseases are rare, but taken together, the more than 6,000 disorders known to result from single altered genes rob millions of healthy and productive lives. Today, little can be done to treat, let alone cure, most of these diseases. But identifying a gene allows scientists to study its structure and characterize the molecular alterations, or mutations, that result in disease. The complete sequencing of the human genome will not only fasten the pace of gene discovery in unigene disorders, but more importantly will help in understanding the molecular basis for multigene disorders. Progress in

understanding the causes of cancer, for example, has taken a leap forward by the recent discovery of cancer genes. The unraveling of the human genome sequence, has also thrown up several ethical issues since the genome sequence available in databases is actually derived from a group of individuals, and several investigators and thinkers have raised the "ethics" of this 'private' information becoming 'public'.

The history

The foundation of this massive human genome sequencing endeavour could be considered to be laid way back in the early part of the 20th century, with the rediscovery of the Mendelian laws (Fig. 1), which has spurred the numerous advances we have made in the field of genetics. All the developments that have laid the foundation for the herculean human genome sequencing effort, are summarized in Fig. 1.

The idea of sequencing the entire human genome was first mooted by the US Department of Energy (DOE) at scientific meetings organized from 1984-86. A US National Research Council committee formally endorsed the project in 1988 in a broader sense by including sequencing of model organisms like bacteria, yeast, worms, flies *etc*. The project was launched in the US as collaboration between the DOE and the National Institutes of Health (NIH). The Office of Genome Research was established in the NIH in 1987, which later came to be known as the National Center for Human Genome Research (NCHGR). By late 1990, countries such as England, France, Japan and the European community also joined the effort. In 1997, it was renamed as the National Human Genome Research Institute (NHGRI), under the leadership of Dr Francis Collins. In 1998, the public funded group, established five high speed sequencing centers *viz*., the Sanger Center, UK; the US Department of Energy Joint Genome Institute in Walnut Creek, California, USA; Washington University School of Medicine in St. Louis, USA; the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts, USA; and Baylor College of Medicine in Houston, Texas, USA. In 1999, the idea of producing a working draft sequence was mooted and approved in order to satisfy the hunger of several thousand scientists worldwide for human genome sequence information. This 'working' draft sequence was released in June 2001. It is expected that the final sequence would be released some time during 2003.

The human genome sequence has been unraveled by two more or less independent efforts. One, the public funded Human Genome Project (to be referred henceforth as HGP), and the other the private funded effort of Celera Genomics. Begun in 1990, the HGP was a 12 yr effort coordinated by the Department of Energy, USA, and the National Institutes of Health, USA under the overall leadership of J D Watson, the Nobel laureate credited with the co-discovery of the structure of DNA. The project was originally planned to last 15 yr, but effective resource management and technological advances accelerated the completion date to 2001. Several important technological developments have activated the genome sequencing efforts. These include developments in the field of DNA sequencing and perhaps more importantly in the field of software development. The DNA sequencing developments include (*i*) four colour fluorescence based sequence detection[1]; (*ii*) dye-labeled terminators[2]; (*iii*) better fluorescent dyes with improved incorporation frequencies[3]; (*iv*) efficient polymerase enzymes suited for cycle sequencing[4]; and (*v*) capillary gel electrophoresis[5]. Simultaneous developments in the field of automation and robotics helped in speeding up the process of sample preparation for sequencing.

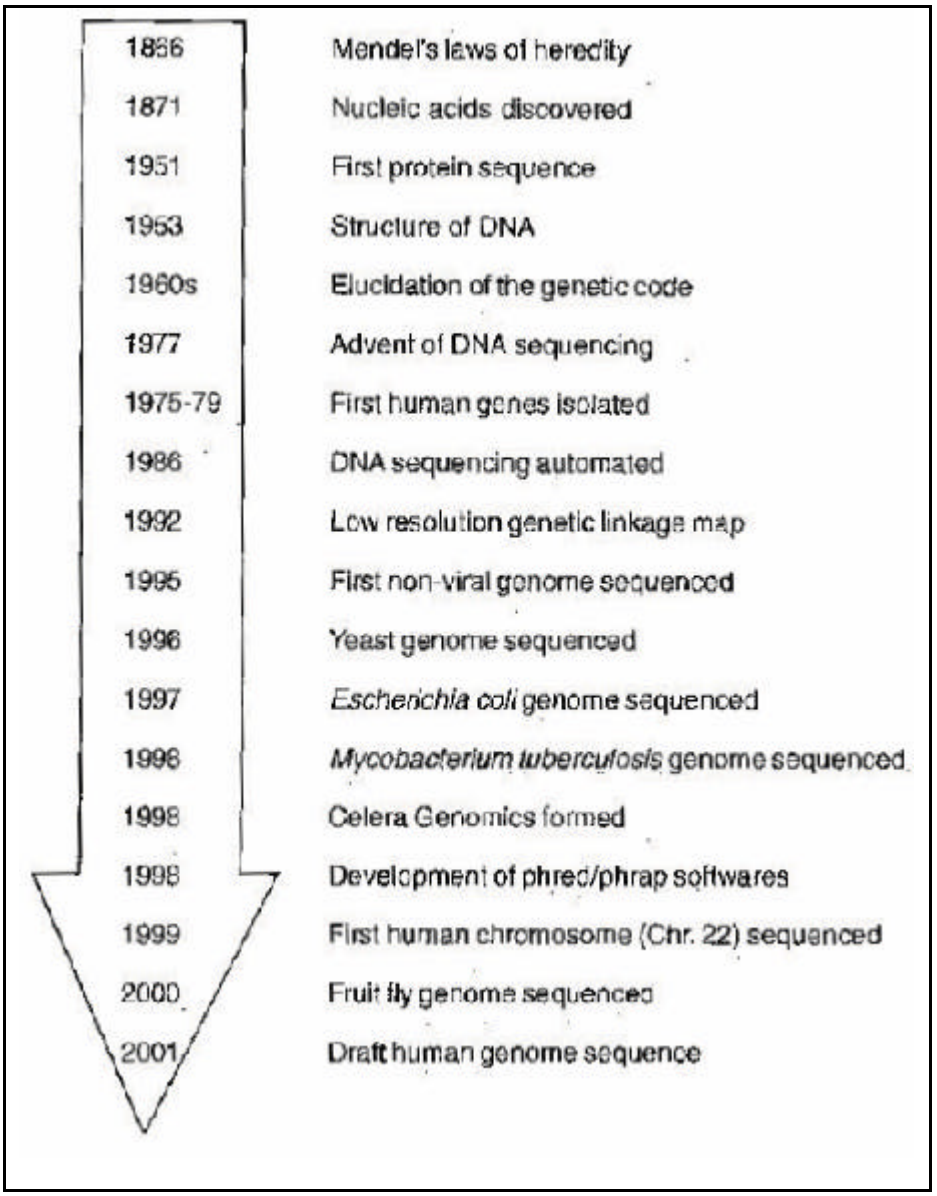

Fig. 1. A time-line showing important discoveries that have laid the foundation for the Human Genome Project. The figure also indicates important milestones in the path to the generation of complete nucleotide sequence of the human genome.

Some of the important improvements in the computing arena included the PHRED software package for assigning a 'base-quality score' in fluorescence based sequencing[6,7]. The PHRAP software was developed to assemble the sequence data generated by using PHRED. Several softwares were also developed for arranging the final sequence data. The amount of sequence deposited in Genbank saw a quantum leap because of the development of these technologies (Fig. 2). These developments also spurred several genome sequencing initiatives and over the years, the total number of finished genomes has shown a tremendous increase, with a phenomenal increase being seen since 1999 (Fig. 3).

The project goals identified in the beginning by the HGP included: (*i*) determine the sequences of the 3 billion chemical base pairs that make up human DNA; (*ii*) identify all the approximately 100-15,000 estimated genes in human DNA; (*iii*) store this information in databases; (*iv*) improve tools for data analysis; (*v*) transfer related technologies to the private sector; and (*vi*) address the ethical, legal, and social issues (ELSI) that may arise from the project.

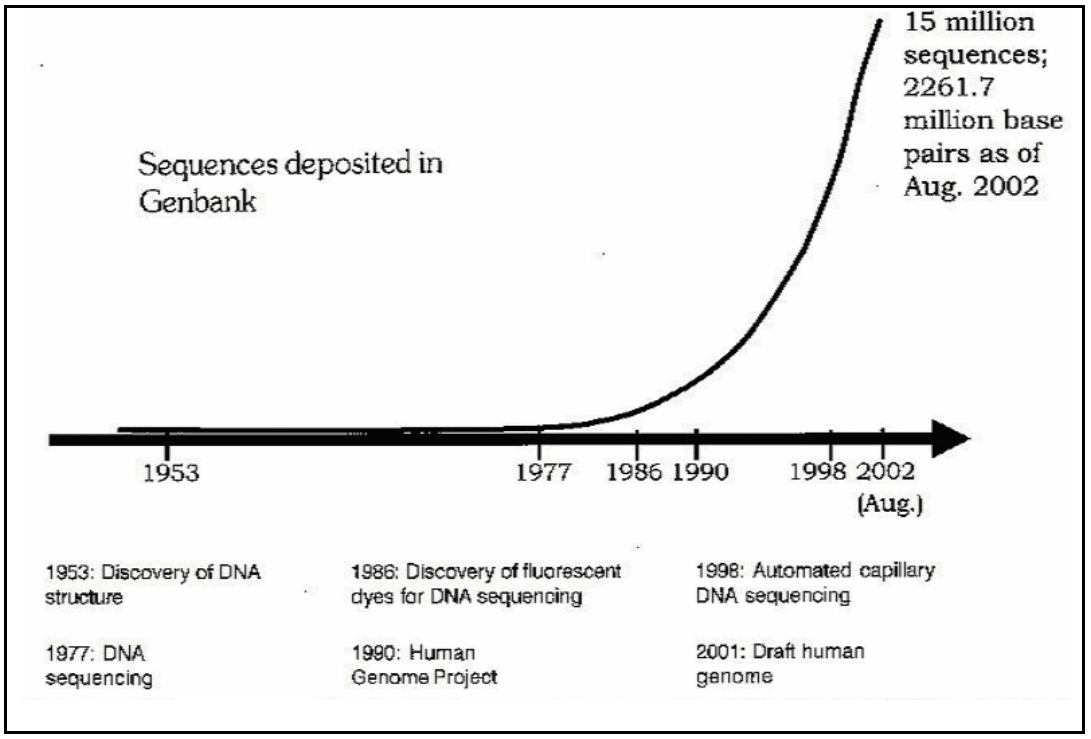

Fig. 2. The pace of deposition of DNA sequences in Genbank. The figure also highlights important technological and other developments that correlate with increase in the amount of sequence deposited.

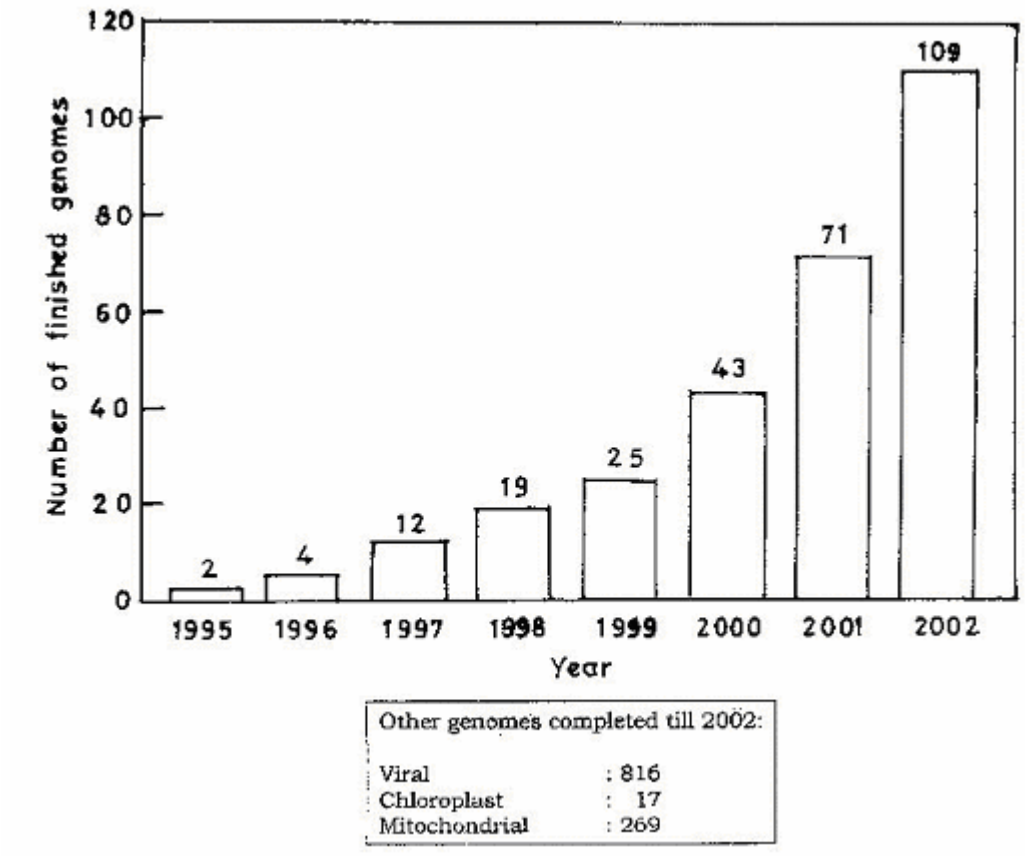

Fig. 3. A time line showing the total number of completed genome sequencing efforts. The histogram includes the genomes of all eukaryotic and prokaryotic organisms, that were completed by the end of 2002. Each bar indicates the total number (which includes the total for previous years). The total number of viral, chloroplast and mitochondrial genome sequences completed till 2002 are shown separately.

Although the private funded effort originated in 1991 at the National Institutes of Health, it got separated from the HGP under the leadership of Craig Venter, through the formation of The Institute of Genome Research (TIGR) and later the Celera Genomics. This review highlights the background which led to this enormous global effort followed by a description of the distinct methodologies followed by the two groups, before describing the salient features of the human genome.

Two distinct approaches

The public funded human genome project arose from two key insights during the mid 1980s, *viz.*, (*i*) that the availability of information on the whole genome sequence would greatly accelerate biomedical research; and (*ii*) that such an effort required a huge amount of infrastructure building.

A programme was launched by Botstein and colleagues[8] to create a human genetic map to locate disease genes based solely on their patterns of inheritance. The sequence of the lambda and phiXl74 bacteriophages[9] provided the initial confidence booster for undertaking this project. A little later in mid 80s work was initiated to create physical maps of genomes of the yeast, *Saccharomyces cerevisiae*[10] and the worm *Caenorhabditis elegans*[11]. Identification of new genes received a shot in the arm by the discovery of the revolutionary technique of expressed sequence tags (ESTs) that involved random shotgun sequencing of cDNA fragments for high throughput gene discovery[12]. The EST method is the principal method of gene discovery used by scientists around the world and has contributed to the discovery of many disease genes including the colon cancer genes[13,14]. ESTs comprise almost 70 per cent of all Genbank accessions.

The human genome project followed the hierarchical shotgun method, illustrated in Fig. 4. The complete sequencing of genomes of model organisms including yeast[15] and *C. elegans*[16] genomes provided evidence that large scale genome sequencing was indeed possible and also validated the two-phase hierarchical paradigm of genome sequencing. In the first step, the genome is divided into appropriately sized fragments and each fragment is sequenced to a high degree of redundancy (by sequencing randomly selected sub fragments). The second step is the finishing phase, in which the final sequence is generated by compiling the sequence data from each fragment, the so-called 'gaps' are closed and ambiguities are resolved. The development of bacterial artificial chromosomes (BACs) was another shot in the arm for the HGP as it became feasible to stably maintain large chunks of human DNA in *Escherichia coli*[17]. Under this scheme, a large set of BAC clones covering the entire genome is generated and independent shotgun sequencing is carried out on each of these clones. Previously constructed genetic and physical maps were used to choose the BAC clones for sequencing specific known regions. Later each BAC clone was restricted with *Hin*dIII and the DNA band pattern for each clone was recorded thus assigning a specific fingerprint to each clone. Then each BAC clone was positioned on the chromosome by anchoring them with the previously determined genetic and physical maps. Techniques such as fluorescent *in situ* hybridization (FISH) and probe hybridization were used for this purpose. New sequence tagged sites (STS) markers were also generated for chromosomal position where previous markers were unavailable. An STS is defined by two short synthetic sequences (typically 20-25 base pairs) designed from a region of sequence that appears as a single copy in the human genome. Clones with minimum overlap were used to reduce redundancy. However, these overlaps, wherever present, proved to be a rich source for the identification of single nucleotide polymorphisms (SNPs). In fact more than 1.4 million SNPs were generated based on overlapping clones alone[18]. Although each collaborating center used a different strategy for sequencing, since all of them used the two common softwares (PHRED and PHRAP), there was no difficulty in combining the data., The overall sequence coverage was about 4.5 fold for the draft sequence output,

i.e., each base was sequenced 4-5 times. The finished human genome sequence stood at 835 Mb in October, 2000, which is more than 25 per cent of the total human genome sequence. The draft sequence released in February, 2001, included 2,212,974 kilo bases of accurate completed sequence, which covered about 63 per cent of the complete genome. As on January 5, 2003, about 98 per cent of the genome has been completed. This effort has provided an estimate of the total human genome size to be approximately 3.3 Gb. The approximate size of each chromosome has also been determined *e.g.*, chromosome 1 - 220 Mb, chromosome 21-35 Mb, X chromosome - 130 Mb, Y chromosome - 20 Mb *etc*.

The private human genome sequencing effort under the leadership of Craig Venter was initiated in NIH. He first showed in 1991 how the revolutionary EST method could be used to quickly (and cheaply) identify genes in the human genome. Adams *et al*[12] ligated the EST approach to automated sequencing. This sparked the beginning of the controversy between Venter and the HGP. The large scale EST generation resulted in development of new algorithms at The Institute for Genomics Research (TIGR) in 1993 to analyze the large volumes of cDNA sequence data[19,20]. In 1997, Venter's lab tested the revolutionary four fluorescent dye technology for DNA sequencing, resulting in automated DNA sequencing. The whole genome shotgun approach for sequencing the human genome was initially ridiculed by most experts. The feasibility of the shotgun approach was validated when the 49 kb bacteriophage genome was sequenced in 1982[9].

In May 1998, Celera Genomics was formed with a stated mission to sequence the entire human genome using a breakthrough DNA analysis technology developed by Applied Biosystems and the TIGR whole genome shotgun method. As a prelude to the human genome, Celera announced it would sequence the fruit fly genome. A formal collaboration was established between Celera and the Berkeley Drosophila Genome Project with the goal of publishing a highly accurate, contiguous genome sequence. A new and advanced assembly algorithm developed by Eugene Myers of Celera was used to assemble the sequence. The 120 Mb euchromatic region of the Drosophila genome was sequenced within 1 yr and published in March 2000[21-23]. This effort revealed that a 5 fold coverage of the genome was enough to yield an authentic output rather than the previously believed 10 fold. Armed with this observation and taking advantage of the sequence data being deposited in the Genbank by the public HGP, Venter's group began their human genome sequencing effort on September 8, 1999 and completed the sequencing on June 17, 2000. The assembly was completed in October 2000.

During the initial stage of sample collection, several concrete steps were taken by the Venter's group to ensure that recommendations of the Helsinki Declaration, for conducting experiments on human subjects, were properly followed. The privacy and confidentiality of the donors was ensured through a two step consent process and a random sample coding system. Given the vast ethnic diversity of *Homo sapiens*, care was taken to ensure that each donor belonged to different ethnic backgrounds including African-American, Chinese, Hispanic, Caucasian *etc.* A total of 21 donors was initially screened. For the final sequencing reactions, only five subjects were used; three females and two males. These included one African-American, one Asian-Chinese, one Hispanic - Mexican, and two Caucasians.

The first important requirement for the whole genome shotgun sequencing is a genomic library of a very high quality which must include the following aspects: (*i*) the library must have equal representation of all parts of the genome; (*ii*) it should be devoid of contaminating DNA from other species, like *Esch. coli*; (*iii*) a very high percentage of clones should actually have an insert; and (*iv*) separate libraries should be made for specific DNA sizes. The Venter's group constructed three libraries, each having insert size in the range of 2 kb, 10 kb, and 50 kb respectively. The sequence obtained from opposite ends of a particular clone were referred to as 'mate pairs'. Knowing that two sequences are derived from the same clone allows these sequences to be linked, even if

the full insert of the clone is unavailable. This is the key to the sequencing strategy called whole genome shotgun (WGS) assembly[24]. Each sequencing reaction was checked for homology with contaminants such as *Esch. coli*, human mitochondrial DNA or the plasmid vector sequences (used for library construction). Only highly accurate sequencing reactions were selected for further analysis.

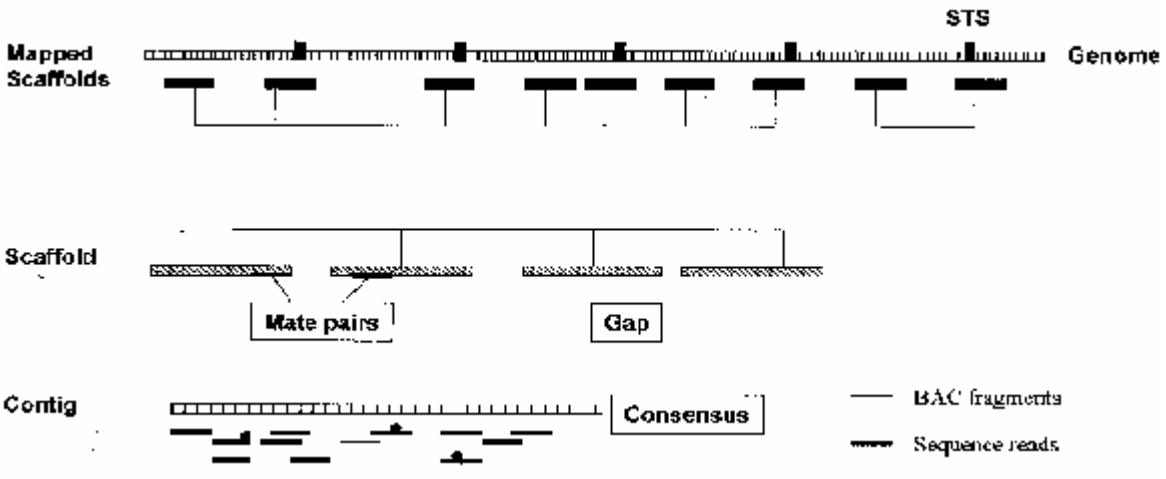

Fig. 4. A depiction of the whole-genome assembly strategy. A contig and the consensus sequence are first created by combining data from overlapping shredded bactig fragments, derived from data deposited in Genbank by the human genome project (HGP), (plain lines) and internally derived sequence reads from five different individuals (solid lines). The consensus represents contigous regions in the human genome, so the term contig. Contigs are combined to produce scaffolds by using mate pair information. The mate pairs help in providing a reasonably accurate measure of the gap between two contigs. Scaffolds are then mapped to the genome with sequence tagged sites (STS) physical map information.

For assembling the final sequence, two sets of data were used. One was generated by the strategy outlined in Fig. 4 using the DNA isolated from the donors. The other data were derived from the sequence generated by the public sequencing effort from the Genbank database. The latter was filtered by removing the contaminating sequences and the synthetic sequence was shredded to generate shotgun data, similar to the ones generated by the sequencing efforts of Celera themselves and then assembled to generate the final sequence. The final sequence thus generated from the synthetic sequence was then compared with the sequence generated in Celera. This approach has been termed the 'whole genome assembly approach'. In another approach, the data from Celera and the HGP were independently partitioned into sets that represented large chromosomal segments, and then shotgun assembly was performed on each set. The success of Venter's efforts has validated the utility of the whole genome shotgun sequencing for large genomes. While the smaller bacterial genomes can be tackled by direct sequencing, larger ones like that of drosophila or humans require the use of genetic markers. The whole genome assembly approach is cheaper than the BAC-based approach. However, the BAC-based approach is more efficient for resolving sequencing anamolies and for determining the sequence of regions that are refractory to DNA sequencing, such as regions high in G+C content. Celera has now turned to the sequencing of the mouse genome, which is of critical importance to biomedical researchers as a model for studies of human biology and medicine.

Important features of the human genome sequence

The coding region of any genome can be divided into two classes *viz.*, protein coding regions and non-protein coding regions. The latter consists of rRNAs; small nucleolar RNAs required for rRNA processing; small nuclear RNAs (components of SNRPS) required for splicing; and tRNAs. There are other species of non-coding RNAs including the telomerase RNA, the signal recognition particle RNA molecules and RNA molecules of novel/unknown function *e.g.*, the *Xist* RNA implicated in X chromosome

inactivation. Computational prediction of such RNA molecules is difficult since it cannot be done based on cDNA information nor by gene-finding algorithms which are based on ORF prediction. A total of 497 tRNA genes have been estimated in the draft genome sequence[18]. In addition, about 324 pseudogenes were also discovered. The total tRNA genes in humans is less than that of the worm but more than that of the fly. The number of tRNA genes does not depend on the complexity of the organism but on the demand for tRNAs in certain tissues or stages of development like in the case of silk worm. The tRNA genes are clustered in specific regions, *e.g.*, almost a whole set of genes is concentrated in a 4 Mb region in chromosome 6[18]. The genes for the large submit and small subunit rRNAs are present as a 44 kb tandem repeat in multiple copies in the short arm of chromosomes 13, 14, 15, 21, and 22[18].

Perhaps the most challenging and definitely the most important task after completion of the draft genome sequence was gene identification, with the ultimate goal of completing the 'periodic table' of life, which will include the complete list of genes and their function. The sequencing of the human genome has now given birth to several new disciplines in the field of genomics *viz*.

(*i*) Functional genomics: Experimental assignment of biological functions to genes; (*ii*) Structural genomics: Elucidation of 3-D structure of gene products, and prediction of protein functions; (*iii*) Proteomics: Studying protein complement of a cell in a specific state/time; (*iv*) Pharmacogenomics: Designing custom drugs; (*v*) Comparative genomics: Studying evolution at the molecular level; (*vi*) Molecular epidemiology: Molecular typing of strains of infectious agents, based on the genome sequence.

It is amazing how human cells can scan through millions of base pairs to accurately detect sequences to be transcribed and also carry out post transcriptional processing as desired. It was a challenge for the two sequencing efforts to accurately predict gene structure in the human genome, to the extent possible. Gene prediction in bacterial genomes is comparatively easier since direct identification of ORFs in smaller genomes is possible. However, prediction in larger eukaryotic genomes is very difficult due to a lower signal-to-noise ratio. This is especially true for the human genome that has a smaller coding area than the worm and the fly genomes. The problem is magnified several fold for the human genome since human genes tend to have very small exons interrupted by large introns. Computer programs for gene prediction are based on available cDNA sequences and/or presence of similar known genes from other organisms. Gene prediction is based on a combination of three simple approaches *viz*., experimental evidence from ESTs and/or cDNA sequences, indirect evidence from previously identified genes from human or other organisms and computational approaches based on information about splice sites, codon bias *etc*. Previous estimates based on reassociation kinetics had put the gene number around 40,000. Walter Gilbert suggested in the mid 80s that the number should be about 1,00,000 based on the ratio of a typical gene to the total size of the genome. Another estimate of 70-80,000 was made based on the density of CpG islands. However, it was later felt that the brain alone would have about 1,00,000 specific transcripts. With the advent of the EST technology, and based on information regarding CpG islands and transcript densities, Incyte Inc. provided an estimate of 142,634 total human genes. Three recent estimates of gene number are (*i*) 35,000 based on EST data in conjunction with chromosome 22 data; (*ii*) 28-35000 based on sequence conservation between human and the puffer fish *Tetraodon nigroviridis* genomes and; (*iii*) 35000 based essentially on data from chromosome 21, 22.

Gene prediction was initiated by the IHGSC based on computational approaches using the Genescan software and then each prediction was confirmed based on EST data and homology with other known genes[18]. In a complementary approach, they first used the EST data that was then extended to use the computational approach. Comparison of results obtained by using these two methods as well as information available on previously determined genes resulted in a considerable amount of overlap. From this

compilation, a non-redundant set was created which has been termed as the Initial Gene Index (IGI), that includes only about 21,000 true genes[18]. The gene prediction estimate was confirmed by comparing with (*i*) the predicted genes for the mouse; (*ii*) the completely sequenced human chromosomes 21 and 22 and; (*iii*) the 'new genes' identified by independent studies. This resulted in an approximate estimate of the total number of genes to be about 32,000. The total gene estimate is about 25,000 for the plant belonging to the mustard family, *Arabidopsis thaliana*, 19-20,000 for the worm, for *Drosophila* it is about 13-14,000 and for the yeast it is about 6000. Humans thus have only double the amount of genes in the worm or the fly. However, the genes in the human genome are spread over a much larger area than the worm or the fly[18].

The Celera group has developed a program called 'Otto' for gene prediction[25]. This program takes into consideration the homology with existing protein sequences (mainly from mouse), and similarity with ESTs or other mRNA data. This software was developed by keeping in mind how a human curator will actually carry out the job of genome annotation. Venter and coworkers have predicted about 30,000 genes, resulting from annotation of 95 per cent of the euchromatic region[25]. This is similar to the number predicted by the other group. However, detailed annotation, comparison with other mammalian genomes such as the mouse (*Mus musculus*) and experimental approaches would be required to be carried out in the near future to determine the exact number.

How do we explain the small gene number? Haldane had predicted way back in 1937 that having too large a number of genes would be deleterious to organisms, since more targets would be available that could be inactivated through mutations, thereby resulting in zygote mortality. Based on estimates of mutation rates, Muller in 1967 estimated the number of human genes to be about 30,000[25]. Several features of the human genome aid in a greater proteome complexity arising out of a moderate increase in gene number. These include the phenomenon of alternate splicing, RNA editing, internal ribosome entry sites, alterations in nature of protein-protein interactions, protein modifications and localization, *etc*. It is predicted that alternative splicing, where the primary transcription product of one gene is spliced in different ways to generate more than one kind of mRNA molecules which then are translated into different protein products, alone will result in almost a five-fold increase in primary protein products.

Alignment of the sequence of known genes with the draft sequence has provided useful information regarding exon-intron boundaries, which is important for designing PCR based diagnostics for molecular genetic disorders[18]. It also helps to map genes in relation to neighboring markers. There is a lot of variation in intron/exon size and gene size in the human genome. Some human genes are very big like the dystrophin gene, which is 2.4 Mb long, and the titin gene, which has 178 exons. The average size of the coding sequence of a human gene is 1340 bp and most exons fall within 50-200 bp[18]. An analysis of smaller exons (less than 19 bp) has revealed a purine bias that might facilitate the splicing process. The average size of introns are more or less conserved in the human, worm and the fly genomes[18]. GC-rich regions in the human genome are usually gene rich with shorter exons whereas AT-rich regions are gene poor with genes having large introns. Conserved splice site sequences were studied by aligning the draft sequence with the known EST sequences[18]. As many as 98 per cent of the genes use the conserved GT dinucleotide at the 5' splice site and the AG dinucleotide at the 3' splice site. Another 0.76 per cent use the related GC-AG rule and 0.1 per cent of the genes follow the rare AT-AC rule[18]. The remaining 1.14 per cent would include mostly sequencing errors. There appears to be a larger frequency of alternative splicing in the human genome compared to the worm and fly genome. As already mentioned, this could be an important explanation for the presence of fewer genes in the genome than was expected.

The knowledge of the human proteome can be determined from the draft sequence based on homologues/analogues from completely sequenced eukaryotic genomes such as

the yeast and partially sequenced genomes of the mouse, wheat *etc*. Such comparisons also yield useful insights into the commonalties and differences between these organisms. Specific information regarding protein domains, domain families, horizontal transfer of genes, mechanism(s) of alternative splicing and post-translational modifications can be elicited. Compared to the worm and the fly proteome (both are invertebrates), the human proteome has a higher proportion of genes falling in the category of the immune system, cytoskeleton, transcription and translation[18]. Humans also have more proteins that fall in the multifunctional category (*i.e.*, proteins with multiple functions) than the other two species. The human genome contains several genes (223 to be exact) that do not have homology with invertebrate proteins but resemble bacterial protein[18]. To rule out the possibility of bacterial contamination, primers based on these sequences were used to perform PCR on freshly isolated human genomic DNA, and successful amplification confirmed them to be genuine human genes[18]. Several of these proteins however share significant homology with proteins from other vertebrate species. There could be two possible explanations for this phenomenon; (*i*) these genes were present in the primordial ancestor and were lost only in the invertebrate lineage or (*ii*) these genes in the vertebrate species (including humans), are a result of horizontal transfer from bacteria. Such horizontal transfers were selected in the vertebrates because they provided some selective advantage. For example, the mitochondrial outer membrane enzyme gene monoamine oxidase, which is a target for same drugs, used in psychiatric disorders, probably, is a result of horizontal transfer from bacteria. A recent report has however challenged the hypothesis of horizontal transfer to explain this phenomenon[26].

The comparison of the human proteome with that of the worm, the fly and the yeast has revealed several orthologous groups between the three. Orthologues are equivalent genes in different species that might have evolved from a common ancestor. Most of the orthologues identified by the human genome project initiative belong to the class of housekeeping proteins that function mainly in metabolism, DNA replication, DNA repair, and translation. Several of the orthologues can be clustered into families. The number of members per family may differ in various species, although they exhibit significant homology. The orthologues between human and fly and human and worm mainly include enzymes known to be conserved through evolution, including enzymes involved in DNA/RNA processing. Other categories include transferases, oxidoreductases, ligases, lyases, isomerases, proteases, molecular chaperons, and GTPases. The protein families that differ significantly from other genomes include those involved in acquired immune function; neural development; inter and intra cellular signaling pathways in homeostasis; and apoptosis. One can appreciate the fact that each of these categories are specific to vertebrates. Clustering of protein sets for each organism has led to the identification of paralogous genes for each species. Paralogues are homologous sequences (sequences that share a common evolutionary ancestor) that diverged by gene duplication (within one organism). The paralogues could differ by having an increased or reduced number of exons, with or without introns. The study of orthologues and paralogues has also resulted in identification of several vertebrate-specific protein families. However, the total fraction of vertebrate-specific protein families identified was only 7 per cent[18]. This means that very few protein domains have originated in the vertebrate lineage with most domains having originated in a common animal ancestor. As expected, only one of the vertebrate-specific protein families represents enzymes, validating the theory of ancient origin of most enzymes[18].

Venter and colleagues have tried to classify genes based on their molecular function. Out of the total genes that they could predict, 41.7 per cent could not be assigned any function[25]. About 10 per cent of genes were predicted to have an enzymatic function, about 12 per cent were classified as signal transducers and another 12-13 per cent as nucleic acid binding proteins or nucleic acid enzymes. Venter's group have also identified several paralogous genes. According to their analyses, majority of intronless paralogues belong to the class of genes involved in the translational process and nuclear

regulation as well as metabolic enzymes[25]. The study of pseudogenes, that arise through such retrotransposition events, or segmental duplications, highlights the evolutionary forces that may result in gene inactivations[25].

The identification of vertebrate-specific gene families has also highlighted important physiological differences between vertebrates and invertebrates. Most of the vertebrate-specific gene families cater to functions such as immunity, nervous system *etc*. Although there appear to be very few novel protein domains that have been discovered in the human genome, there seems to be a far greater extent of innovation in terms of novel vertebrate proteins, which arise by re-shuffling of domains (as they are arranged on a polypeptide chain). The human genome has generated 1.8 times, more novel architecture (from the same existing protein domains) as compared to the worm and the fly, and 5.8 times more than the yeast[18]. Most of this novel architecture is evident in extracellular and transmembrane proteins. Another important feature of the vertebrate protein families is their expansion, as evident from the human genome. Gene duplication has led to the creation of new members in each family. The human genome contains several examples that bear out this fact. There are 30 fibroblast growth factors in humans as opposed to only two each in the fly and the worm[18]. There is a similar expansion in proteins belonging to the TGFβ family, the intermediate filament proteins family, the olfactory receptor family and the immunoglobulin family. Each one of these families are hallmarks of important physiological processes in vertebrates. Therefore, the complexity of the human proteome stems not from a moderately greater number of genes but from a greater innovation in terms of generation of a novel architecture from existing domains, larger protein/domain families and multidomain families. As already stated, other factors like alternative splicing, post-translational modifications along with the ones mentioned previously, together lead to the complexity of the human proteome. Some other mechanisms that may lead to a larger protein diversity include RNA editing[27] and various forms of translational regulations (like the use of internal ribosome entry sites)[28]. A comparative genomics study has revealed that the fly has 177 of the 289 human disease genes examined[29].

In bacteria, genes present in close proximity often encode proteins that function in a common pathway. This is not true for higher organisms. However, conservation of gene location between two diverse organisms suggests their common origin and helps in the identification of disease loci. The presence of linked genes in diverse organisms such as the humans, mice, worm and the fly has been studied[18]. 'Conserved synteny' has been used to describe the presence of at least two genes that reside in the same chromosome in different species. About 180 such chromosomal segments, where the linear order of gene loci has been conserved in human and mice, based on search for orthologous genes, has been discovered. The segments ranged from about 15 to 90 Mb[18].

The genome structure includes several features apart from the coding region. The draft genome sequence has enabled an estimation of the average GC content to be about 41 per cent[18]. There are, as expected, regions (in some cases more than 10 Mb), with a GC content of 36 per cent[18] and similarly some specific regions have a GC content of 47 per cent. One common feature associated with coding regions is the presence of CpG islands located at the 5' end of genes. The methylation of such islands is found to be associated with imprinting and gene inactivation. Computational methods to detect CpG islands have defined them as stretches of DNA more than 200 bp, having a G+C content of more than 50 per cent, with a high frequency of the CG dinucleotide, significantly higher than what can be predicted based on probability[30,31]. The CpG dinucleotide occurs at about one fifth of the 4 per cent frequency which is expected based on probability. The study by Venter and colleagues have confirmed the belief that the CpG islands are usually concentrated close to the first exon of genes[25]. The CpG dinucleotide is grossly misrepresented in the genome, which is expected since the cytosine in CpG is susceptible to methylation followed by deamination, which leads to mutation of Cytosine to Thymine, the most common form of SNP found in the human genome. A search for

CpG islands has identified 28,890 such islands, excluding those that are present in repeat-rich regions[18]. Most of these islands are short, less than 850 bp, and these smaller islands are consistent with the gene regulatory function proposed for them. The density of these islands varies with different chromosomes but correlates with gene density per chromosome. In other words, chromosomes with a higher gene density have a high density of CpG islands[18].

The human genome also gives an opportunity to collate the previously determined genetic and physical map of each chromosome. Such an analysis has revealed that the recombination rates (crossing over frequency) are higher near telomeres (ends of the chromosomes) and lower near centromeres (center of the chromosomes)[18]. Moreover, long chromosome arms have a comparatively low recombination rate compared to short chromosome arms. This also holds true for yeast chromosomes. There could be several explanations for this inverse relationship. In terms of probability, the chance of recombination increases with increase in length of the chromosome arm. Therefore, a higher recombination rate for shorter arms could be a compensatory mechanism for the lower probability of recombination. Another explanation could be the presence of a surveillance mechanism of positive crossover interference that inhibits additional crossovers in regions neighboring a crossover event. Differences in overall recombination rate have been observed between various chromosomes as well as between identical chromosomes from the two sexes[18]. Further analysis is required to determine the mechanism(s) of the observed differential recombination frequencies.

Several organisms that are taxonomically related or have similar gene content, differ greatly in their genome size. This anamoly has been termed as the 'C-value paradox'. The anamoly is due to the presence of large chunks of repetitive DNA in large genomes. About 25 per cent of nuclear DNA is composed of genes and related sequences out of which 1-2 per cent are actual coding regions (exons) and the rest include non-coding regions (introns, 5' and 3' UTRs *etc*.). The remaining 75 per cent includes highly repetitive DNA (40%) and unique or moderately repetitive DNA (35%)[18]. Repeat regions can be of various types and generally include (*i*) pseudogenes, (*ii*) repeat regions present near the centromere and the telomere, (*iii*) simple sequence repeats (SSRs) consisting of mono/di/tri nucleotide repeats such as $(A)_n$ or $(CA)_n$ or $(CAG)_n$, (*iv*) transposon-derived repeats and, (*v*) segmental duplications, arising due to chunks of DNA copied from one region of the genome into another. The repeat regions are actually rich sources of information to understand the process of evolution, mutation and selection. They are important tools for studying population genetics and medical genetics and serve as molecular markers to map chromosomal regions. In medical genetics, their role in mapping disease loci for genetic disorders is extremely important.

Overall, the human genome is very stable, mutations occur at the rate of 10^{-9}. However, the repeat regions are a major source of variation in the human genome, and among these the transposable elements are the most active and also represent a major proportion of the repeat regions. Moreover, the total number of transposons in the human genome appears to be higher compared to the other genomes that have been sequenced. The transposable elements in mammals belong to four major categories namely (*i*) long interspersed repeats (LINES), (*ii*) short interspersed repeats (SINEs), (*iii*) LTR retrotransposons and, (*iv*) DNA transposons. LINEs are full fledged transposons, which include genes for their propagation. SINEs are shorter than LINEs, contain no ORFs and 'piggy back' on LINEs for their propagation. About 5000 full length non-LTR retrotransposons are present out of which about two may still be capable of retrotransposition[18]. Recently, two independent studies have shown that the LINE-1 (L1) retrotransposon, the most common retrotransposons in the human genome, can actually cause genome instability mainly through deletions in transformed cell[32-33]. New Ll retrotranspositions are estimated to occur by creation of specific inversions by twin priming mechanism[34]. The LTR retrotransposons are flanked by long terminal direct repeats (LTR) and the retrotransposons themselves contain genes for protease, reverse

transcriptase (*pol*), RnaseH and integrase. The mechanism of retrotransposition is akin to that of common retroviruses. It is possible that retroviruses arose from such transposons by the acquisition of an envelope (*env*) gene. Most of the LTR retrotransposons in the genome are present as fossils, since they have only the LTR, and have lost the transposon element due to recombination between two LTRs. The DNA transposons resemble typical bacterial transposons, they have terminal inverted repeats which enclose the transposase gene. The transposase gene product binds at the terminal repeats for propagation through a 'cut and paste' mechanism (whereas L1 will move through a 'copy and paste' mechanism). All transposons have different means to ensure their propagation. LINEs and SINEs depend on vertical transfer, whereas DNA transposons depend mainly on horizontal propagation. Since LINEs and SINEs require processing in the nucleus and their encoded protein product binds to their own RNA, they are sustained for a longer time in the genome[18]. However, the DNA transposons do not have any such mechanism and therefore, have a shorter life span in any given genome.

Interesting differences between repeat sequences of human and other genomes have been revealed[18]. Firstly, the euchromatin region of the human genome has a higher density of repeat elements compared to other genomes such as that of the fly, the worm and the mustard weed. Secondly, the transposons in the human genome are more ancient than the ones in the other genomes. Thirdly, the other genomes do not have any dominant families like the LINEs in the human genome[18]. Moreover, unlike the human genome, the genomes of Drosophila and *C. elegans* have a high frequency of short-lived DNA transposons. These features, relating to repeat regions, of the human genome are probably a conserved feature of all mammals. As expected, the density of repeat regions varies in different regions of the genome[18]. Some regions of the genome are repeat-rich whereas the others are devoid of repeats. The Y chromosomes appears to be younger than the other chromosomes, based on the age of the repeat elements present.

The human genome also characteristically contains several simple sequence repeats (SSRs). SSRs are called microsatellites if the repeating unit number is less than three and minisatellites if it is four or more. Trinucleotide repeats are more rare than the dinucleotide repeats. The SSRs are important in the context of mapping of disease loci. Several disease loci have been mapped by using such repeats since they exhibit a great degree of length polymorphism, probably due to slippage during DNA replication. Therefore, a comprehensive cataloguing of such repeats is important in the context of study of novel genetic disorders.

Segmental duplication occurs when a large chunk of the genome (1-200 kb) is duplicated in another region, in the same chromosome (intrachromosomal duplication) or in a different non-homologous chromosome (interchromosomal duplication). Most of such duplications in the human genome have a high degree of sequence identity thereby indicating a more recent origin. The analyses of chromosome 21 and 22 indicate that a high proportion of duplicated regions are located close to the centromere. The region neighboring the telomeres also appears to be rich in duplicated regions. There may exist a mechanism to safeguard against frequent duplications from occurring in other regions of the chromosomes. A recent study by Bailey and coworkers[35] revealed several recent segmental duplications in the human genome, some of which have also been associated with genetic disorders. As seen for retrotranspositions, segmental duplications also could be one mechanism, which would increase the protein repertoire in the human cell.

Significance of human genome sequence

Mapping disease loci

The preferred mode of mapping disease genes used to be through positional cloning[36]. This requires study of genetic markers that are linked to the disease phenotype in affected members in a family, followed by identification of genes present in and around the marker that is linked to the disease (about 1 Mb on both sides of the marker). The availability of the draft sequence has increased the number of polymorphic markers available to facilitate gene mapping studies. These include microsatellite as well as SNP markers. More importantly, the genome sequence has also facilitated direct *in silico* (within the computer) gene identification. Several disease genes (about 30) have already been cloned based on the human genome draft sequence[37-40]. The availability of the genome sequence has also helped in the identification of paralogues of disease genes, which are useful in determining the etiology of related genetic disorders. For example, Presenilin-1 and Presenilin-2; mutations in either of these paralogous genes can cause early onset Alzheimer's disease[41,42]. The paralogues can also be useful for designing potential therapics. The Human genome draft sequence is also proving to be useful to determine the genetic basis for chromosomal deletions occurring due to unequal crossing over between intrachromosomal duplications, for example, the iGeorge/velocardiofacial syndrome region on chromosome 22[43] and the William's-Beuren syndrome recurrent deletion on chromosome 7[44]. The human genome sequence has also resulted in revealing disease loci associated with a class of neurodegenerative diseases known as transmissible spongiform encephalopathies (TSEs) which were earlier shown to be intimately linked with the prion proteins. Recently, by using a combination of bioinformatics and molecular approaches, three genes within a 55 kb interval were identified in the human prion protein locus *viz*., *PRNP, DOPPEL* or *PRND* and *PRNT*[45]. This locus has been implicated in human prion disorders. A 9 cM genome screen on 437 Alzheimer's disease families has revealed 12 additional disease loci on different chromosomes, thus paving the way for a quick identification of genes that may play a major role in the progression of this important disorder[46].

Drug targets

For the past several decades, the pharmaceutical industry has depended upon a limited set of drug targets; a recent compendium has listed only 483 drug targets which account for almost all drugs on the market[47,48]. The draft human genome sequence has however resulted in a 'gold rush' for big pharmaceutical companies, the race to identify new drug targets through high throughput genomic research. Careful analysis of the human genome draft sequence is already beginning to reveal several new drug targets (Table I). A putative homologue of the serotonin receptor has been identified from the long arm of chromosome 11[49]. The gene, named *5-HT3B*, and its product resembles with the previously identified receptor *5-HT3A*. Given the central role of the serotonin pathway in mood disorders and schizophrenia, the discovery of a major new therapeutic target is of considerable importance. Similarly a second receptor for the slow reacting substance of anaphylaxis (named CysLT2), has been identified using the combination of a rat EST and the human genome sequence[50]. As the leukotriene pathway has been a significant target for the development of drugs against asthma, the discovery of this new receptor has important consequences. β-amyloid, whose deposition in senile plaques is the hallmark of Alzheimer's disease, is generated by proteolytic processing of the amyloid precursor protein (APP). One of the enzymes involved is the β-site APP-cleaving enzyme (beta secretase, BACE), which is a transmembrane aspartyl protease. A second enzyme named BACE2 has been identified, based on a computational search of the public human draft genome sequence[51]. The development of antagonists to BACE and BACE2 represents a promising approach to preventing Alzheimer's disease. Similar

searches have yielded several paralogues including putative dopamine receptors, purinergic receptors, insulin-like growth factor receptors *etc*[18]. Many of these could prove to be useful to identify several such paralogues.

Basic research

The draft sequence is expected to provide answers to several unsolved basic questions including those relating to complex phenomenon such as embryonic development, brain function *etc*. The human genome is touted as the master plan for building an organism. But we are still a long way away from understanding how that "master plan" directs construction. Traditionally, developmental geneticists have learned how genes control development by altering a gene and observing the phenotype, in model organisms such as the fruit fly *Drosophila melanogaster*, the nematode worm, and the mouse. Complete genomes (the fly, worm, and human) have simplified the process of locating genes that cause intriguing abnormalities. One unsolved mystery, concerning the polymorphic bitter taste was actually solved based on the draft sequence. The gene was mapped to a specific region where a G-protein coupled receptor was located. The receptor is expressed exclusively in the taste buds and its expression can be regulated in response to bitter substances in cell culture[52-55]. The *Taslr3* gene, which encodes a novel taste receptor, was also identified from the draft sequence and is allelic to the sweet responsiveness locus *Sac* of the mouse[56].

Single Nucleotide Polymorphisms (SNPs) and the birth of predictive medicine

The majority of human DNA polymorphisms are single nucleotide polymorphisms (SNPs), which are single base pair positions in genomic DNA at which different sequence alternatives (alleles) exist in normal individuals in some population(s). They are distinguished from rare sequence variations by a requirement for the least abundant allele to have a frequency of more than 1 per cent in the population. The density of SNPs differ between different genomic regions, but the overall frequency is estimated to be around 1 in 200-600 base pairs, resulting in several million potential single base differences between any two individuals. According to recent studies, there is an estimated 1.4 to 3.1 million SNPs in the human[57,58]. An important fallout of the HGP is the generation of an SNP map of the human genome. The increased prevalence of SNPs in the human genome compared to other types of sequence variations (such as microsatellites) make them ideal for genotyping and genetic dissection of complex traits and disease. SNPs can occur in both coding (gene) and noncoding regions. SNPs have an average spacing of 1.9 kb and 63 per cent of 5 kb intervals contain an SNP[58]. This high prevalence of SNPs within the human genome allows the usage of SNPs as ideal candidate markers. In general, higher the frequency of an SNP allele, older the mutation that produced it; so high-frequency SNPs largely predate human population diversification, and this information can be used in human evolution studies. Most SNPs have no effect on cell function but SNPs that occur in coding regions may underlie differences in our susceptibility to, or protection from, various diseases; in the age of onset and severity of illness; and in the way our bodies respond to treatment[58].

The frequency of occurrence of SNPs in individual autosomes lies in the range of 5.19 to 8.79 differences per 10 kb. The variation in X chromosomes is about 4.69 difference per 10 kb and 1.51 differences per 10 kb lie in the Y chromosome non-recombining region[57]. The lower frequency in the X chromosome compared to antosomes can be easily explained based on the fact that the effective population size for the X chromosome is 3/4th of the autosomes. Similarly, the effective size for the Y chromosomes is 20 per cent that of autosomes[57]. Even within a single chromosome, the nucleotide diversity is not uniformly distributed[18]. That is not surprising since the mutation rate does not display a uniform distribution along the sequence. Some genomic regions have significantly lower or higher diversity than the average. For example, the HLA locus in chromosome 6, which encodes proteins that present antigens to the

immune system, shows very high heterozygosity, which reaches more than 30 SNPs per 10 kb. Presence of a higher GC content correlates with a higher nucleotide diversity, which reflects, at least in part, the high frequency of CpG to TpG mutations arising from the deamination of methylated 5-methylcytosine[57]. Some of the significant uses of SNPs are discussed below:

Mapping of disease genes

The prevalence of SNPs makes them extremely useful for disease associated studies. In several cases, an SNP can be directly linked to a disease. Several diseases are directly caused due to single nucleotide changes such as thalassaemia, muscular dystrophy, phenylketonuria, long QT syndrome, familial hypertrophic cardiomyopathy *etc*. A recent study has led to the identification of novel mutations that cause primary congenital glaucoma (PCG) in Indian pedigrees[59]. This study has also resulted in the development of simple PCR-REA based detection systems for a quick and simple diagnosis of PCG[59]. A similar approach was adopted for other genetic disorders of the eye such as the Axenfeld-Rieger anomaly[60] and retinoblastoma[61]. It is important to determine the disease causing SNPs for common genetic disorders in our population and create a knowledge-based database for the same. This will facilitate the establishment of an effective molecular screening regime for specific genetic disorders based on the common SNPs. The identification of SNPs can also be used to measure linkage disequilibrium within the human genome[62]. Linkage disequilibrium is a measure of the extent to which genomic sequences representing specific allelic combinations (haplotypes) from a common ancestor have persisted over the course of evolution[62]. Study of linkage disequilibrium through SNPs can be used to map disease related genetic loci as well as to answer questions on evolution of human populations[57,63].

Finding genetic loci associated with a disease phenotype is not an easy proposition. This is especially true in the case of complex diseases or traits which result from the cumulative influences of multiple susceptibility factors, some genetic, some environmental. Identifying the level of contribution of each individual disease susceptibility locus/factor toward the disease phenotype is a challenging endeavour. Though positional cloning has been employed successfully to identify genetic loci associated with diseases that follow a simple Mendelian pattern of inheritance[64], genetic mapping displays limited success in defining multiple loci responsible for complex traits, especially loci that have minor contributions to the phenotype[65]. Whereas previous efforts at identifying trait associated genetic loci involved considerable effort to discover polymorphisms across the region of interest, the high prevalence of SNPs significantly facilitates the search for target loci[66]. The main use of the human SNP map will be to dissect the contributions of individual genes to diseases that have a complex, multigene basis, such as central nervous system disorders, cancer and cardiovascular disease[67]. Some advances have already been made in this area. SNPs have recently been implicated in risks for Alzheimer[68]. Martin and co-workers[69] used five SNPs within the *tau* gene to show a strong linkage between this gene and idiopathic Parkinson disease.

Pharmacogenomics

One of the major complications facing drug therapy is the individual variations among patient populations with respect to their responses to drug treatment. Not only do medications vary in their efficacy among individual patients, a drug that is tested and found to be safe in one individual might prove lethal for another. These individual phenotypic differences can be directly attributed to polymorphisms in genes that encode key protein(s) molecules that affect the impact of the drug. Individual genetic variations affect several functional aspects of a drug including drug absorption, distribution, metabolism, excretion and drug target specificity[70]. Understanding the nature of these polymorphisms would allow physicians to customize the selected medication so as to meet the specific treatment requirements of individual patients (pharmacogenomics).

Since SNPs are the most prevalent type of polymorphism within the human genome, individual SNPs could be identified that correlate with the individual variations in drug response[71]. Availability of the entire human genome sequence information as well as extensive knowledge of the SNPs that occur within the human genome, facilitates the process of associating genetic variations with inter-individual phenotypic differences. Pharmacogenomics, the concept of personalized and effective drug selection based on the genotype of a patient, enables the improvement of drug efficacy, safety, as well as the cost-effectiveness of the treatment[71].

Some typical examples of genes that play a role in drug response, identified through pharmacogenetic studies, are given in Table II. One example is the SNP determined in the *Cyp2D6* locus, which affects metabolism of the pain alleviating drug codein. This polymorphism limits the conversion of codeine to its active morphine form and therefore, patients harbouring this polymorphism do not get pain relief when they take the normal dose of codeine. Another example is the presence of a gene variant on both chromosomes that codes for an enzyme thiopurine S-methyltransferase (TMPT), in approximately one in 200 Caucasians. The TMPT metabolizes a class of drugs, thiopurines, used for the treatment of autoimmune diseases such as Crohn's disease and rheumatoid arthritis, renal transplant patients, lymphoblastic leukemia and childhood leukemias. Individuals with this SNP are unable to metabolize the drug leading to a toxic accumulation of the drug in the bloodstream. Members of the CYP3A family are the most abundant CYPs in the human liver and intestine. Their substrates include cyclosporin A, HIV protease inhibitors, calcium channel blockers, and some anti-cancer drugs. The variations in the levels of CYP3A, sometimes upto a 30 fold difference, has been observed for a long time in different populations. Genetic polymorphisms were recently identified in the promoter of several members of this family of genes, which was shown to be responsible for the polymorphic levels in a recent study[72]. This study has also revealed the role of specific SNPs in the coding region of *CYP3A5* that lead to alternative splicing and truncation of the protein and thereby leading to the drastic reduction in protein levels.

The cytochrome P 450 enzymes play a major role in the metabolism of drugs. The SNP in the *CYP2D6* gene is very common in some populations. *CYP2D6* is involved in the metabolism of about 30-40 drugs and SNPs that result in poor metabolism as well as rapid metabolism have been identified. Similarly *CYP2C19* is another gene in which SNPs result in generation of a truncated protein with little or no activity. Sometimes an SNP may even be lethal, The drug warfarin is lethal when given as an overdose, as it may cause excessive bleeding. An SNP in the *CYP2C9* protein, which metabolises this drug, leads to a very low rate of metabolism. If a person harbouring this polymorphism is given a normal dose of warfarin, then it may turn out to be lethal. Therefore, genetic testing before prescribing the drug is very important in such cases. This is an example of the significant effect that pharmacogenomics will have on the future of medical treatment and the way pharmaceuticals are prescribed for each individual. With the advent of lineage independent genotyping through SNP analysis, pharmaceutical companies hope to be able to utilize this new method to develop drugs in a more targeted fashion and to significantly narrow down the number of test subjects by conducting genotyping tests.

Table I. New drug targets

New target gene	Homologue	Role
5-HT3A (Serotonin receptor)	5-HT3A	Schizophrenia
Second receptor for Cysteinyl leukotrienes (CysLT2)	CysLT1	Allergic disorders
BACE2	BACE1	Alzheimer's

Table II. Genes that modulate drug response as a function of SNPs present in them

Locus	Drugs	Role
Drug metabolism		
CYP2D6	Codein, Fluoxetine, Propafenone,	Antidepressants, pain killers
CYP2C9	Warfarin	Anti-coagulant
CYP2C19	Omeprazole	Proton pump inhibitor
Thiopurine S-methyltransferase	6-Mercaptopurine, azathioprine	Anti cancer, immunosuppressor
Dihydropyrimidine dehydrogenase	5-Fluorouracil	Inhibits cell division (anti cancer)
N-acetyltransferase 2	Isoniazid, Hydralazine	Inhibits synthesis of bacterial mycolic acid (anti mycobacterial), Relaxes vascular smooth muscles (used to lower blood pressure)
UDP-glucuronosyltransferase lAl	Irinotecan	Inhibits topoisomerase II (anti cancer)
Drug transport		
MDR1	Digoxin, Fexofenadine	Blocks sodium potassium ion exchange pump (for heart failure), anti histamine (for allergy)
Drug targets		
ACE	Enalapril	Angiotensin converting enzyme inhibitor (for hypertension)
β_2-adrenergic receptor	Albuterol, Terbutaline	Bronchodilator (for asthma)
ALOX5	ABT-761	Lipoxygenase inhibitor (anti cancer)
5-HT2A, 5-HT6	Clozapine	(For schizophrenia)
Disease pathology		
ApoE4	Tacrine	For Alzheimer disease
Stromelysin- 1	Pravastatins	For Alzheimer disease
HERG, KvLQT1, Mink, MiRP1	Terfenadine, anti-arrhythmics	Antihistamine, anti-arrhythmic

SNPs, single nucleotide polymorphisms

The future

Since both the sequencing efforts have resulted in only a draft sequence, it needs to be updated and the gaps need to be filled to result in the complete sequence of the human genome. This will firstly involve complete coverage of the euchromatic region through sequencing of additional clones. The next stage would involve the screening of additional clones in order to close gaps present in the contigs. This will be followed by additional sequencing of heterochromatic regions and repeat-rich regions, although it may prove to be difficult using existing sequencing technologies. Efforts are already underway to identify those segments that are missing in the working draft sequence of the human genome, by constructing a high resolution radiation hybrid map of the human genome draft sequence[73]. The other important endeavour will be to complete the list of proteins encoded by the human genome, that direct our development, appearance, behaviour, talents, and susceptibility to disease[18,25]. The interpretation of the human genome organization will also depend, to a large extent, on genomes of other organisms, which will facilitate the understanding of the functioning, regulation, and origins of our own genes[29,74,75]. The importance of the comparative efforts has already been validated in the fields of anatomy, physiology, and medicine. Decisions and priorities regarding the choice of organisms for whole genome sequencing will ultimately shape biology and influence the potential applications of the completed genome sequence. The other important step will be to generate a complete collection of human cDNAs[76]. This would be followed by cataloguing all possible gene specific alternative splicing patterns.

Completion of the Human Genome Project has enabled researchers to analyze sequence variation types, patterns and their relative frequencies within the human genome. Studying the genetic variation with relation to the human genome allows us to determine the evolutionary history of man and to map the evolutionary relationship shared by the human genome with the genomes of other organisms. But more importantly, this has brought about a change from the previously dominating focus on gene discovery to one aimed at understanding gene function. The emerging transcript profiling techniques are powerful tools that represent a new approach to the elucidation of gene function and the investigation of differences at the level of gene regulation, and can identify genes not previously implicated/known to be involved in a particular biological and/or pathological process. A related development will be the study of regulatory regions within the genome. These regions are important as their action results in differences between the various cell types. Regulatory regions of most genes remain uncharacterized. Comparative genomics has already resulted in characterization of some of these sequences by comparison of orthologous genes[77]. A study of various epigenetic mechanisms such as DNA methylation is another important problem, which should be addressed immediately. In this context, a Human Epigenome Project has been recently launched[78,79].

The development of new genomics and proteomics techniques is enabling researchers to analyze whole genome profile of cells and tissues, instead of investigating one gene at a time. Kim and coworkers[80] have analyzed the relative activity of nearly every worm gene at different developmental stages through several microarray based experiments. Similar work on human genes is following. The important areas that will be benefited by the whole genome approach include embryonic development, differentiation, brain function, stem cell biology *etc*. It is now possible to analyze which genes are switched on in human embryonic stem cells that can develop into any cell type. Following the pattern of gene activity in undifferentiated cells during the development into certain tissue types, could throw light on the molecular features of a specific stem cell.

The brain is the most complex organ in the body and as already mentioned, a large proportion of the total gene pool is thought to be involved in constructing, wiring up, and

maintaining the nervous system. Neuroscientists are now working to decipher the molecular genetic pathways that underlie brain function. The completed genome will also accelerate the search for genes at fault in neurodegenerative diseases. Psychiatrists hope to solve the intricacies of human behaviour based on the human genome sequence. For example, the recent work of Nestler and colleague[81] has identified genes that may be linked to addiction to cocaine, based on differences in the dopamine transporters.

Ethical issues

It is the responsibility of all members of society to ensure that knowledge resulting from scientific research is used responsibly and judiciously. Although, science is often regarded as being neutral, the application and products of research may not be. The way in which we use this knowledge has various ethical implications that must be considered. The HGP has expectedly raised several ethical issues as a fallout of the unraveling of the human genome sequence, pertaining to diagnostics, health care, insurance *etc*.

The National Human Genome Research Institute's (NHGRI) Ethical, Legal and Social Implications (ELSI) Programme was established in 1990 as an integral part of the Human Genome Project. The planners of the Human Genome Project recognized that the information gained from mapping and sequencing the human genome would have profound implications for individuals, families and society. While this information would have the potential to dramatically improve human health, they also realized that it would raise a number of complex ethical, legal and social issues. How should this new genetic information be interpreted and used? Who should have access to it? How can people be protected from the harm that might result from its improper disclosure or use? The ELSI research programme was established to address these issues and has become an integral part of the HGP. ELSI provides a new approach to scientific research by identifying, analyzing and addressing the ethical, legal and social implications of human genetics research at the same time that the basic science is being studied. In this way, problem areas can be identified and solutions developed before scientific information is integrated into health care practice. In January, 1990, the ELSI Working Group issued its first report and defined the function and purpose of the ELSI programme as follows: (*i*) To anticipate and address the implications for individuals and society of mapping and sequencing the human genome; (*ii*) to examine the ethical, legal and social consequences of mapping and sequencing the human genome; (*iii*) to stimulate public discussion of the issues; and (*iv*) to develop policy options that would assure that the information be used to benefit individuals and society.

Several ethical issues were raised right from the beginning of the HGP. One of the initial concerns was whether it was worthwhile to invest so much money to unravel the whole genome, the majority of which consisted of non-coding DNA. However, this issue was solved since scientists argued successfully in favour of complete sequencing given the importance of physical and genetic maps, which would be made available after complete sequencing. Another concern was regarding the discovery of a gene, which might result in a possible genetic disorder, much before the disorder manifests itself. This might result in panic, anxiety and frustration among the general public. However, on the other side, the development of efficient diagnostic methods will make the prognosis and management of the condition easier. It is important that policy makers do not put major blocks in using the human genome sequence data for research purposes.

A survey conducted in 37 countries in 1997, has revealed that invasion of privacy, as a fallout of the HGP, is the most serious concern of professionals. They also agreed that neither employers nor insurance companies should have access to a person's genetic information without his or her consent. It is also important to protect 'medical' information rather than only 'genetic' information. However, certain types of medical information, that includes epidemiology and studies on allele frequency, disease prevalence *etc*., should be freely available to researchers. Privacy concern for such

public health issues is inherently reduced since the data always pertain to a population and not to an individual.

Conclusions: The Book of Life is poised to change the future of medicine

There are several paradigm shifts taking place in the field of biomedical research as a fallout of the human genome sequencing. The goal of the Human Genome Project is to provide scientists with powerful new tools to help them clear the research hurdles that keep them from understanding the molecular essence of several tragic and devastating illnesses, such as diabetes, cardiovascular diseases, immune system disorders, birth defects, schizophrenia, alcoholism, Alzheimer's disease, and manic depression. Functional genomics studies are poised to make two major contributions to the development of more efficient drug therapy. Firstly, functional genomic approaches will aid in the identification of the most suitable targets for drug intervention. As already mentioned the current drug therapy rests on only about 500 odd targets, the emerging number is estimated to be in the range of 5,000 to 10,000 target molecules. Secondly, pharmacogenomics will help in determining why patients respond differently to drugs. The genetic patterns that define these responses are being identified and used to target drugs more effectively during their development. This approach will also allow for individualized drug therapy.

The future of medicine is poised to undergo a sea change leading to an improved diagnostics for diseases, earlier detection of genetic predisposition to diseases, gene therapy and pharmacogenomics for custom drugs. Given the rapid advances in the field of gene therapy, it might become possible to treat genetic disorders by correcting the errors in the gene itself, by replacing the faulty gene with a normal one or by switching off expression of the gene itself. Another important area(s) that will be influenced is anthropology, evolution and human migration. The elucidation of the human genome sequence is already facilitating the study of evolution through germline mutations in lineages, the study of migration of different populations, and comparison of breakpoints in the evolution with ages of populations and historical events. The human genome project will help in understanding the choreography of the intimate dance of gene activity that begins after fertilization of the unicellular egg, culminating in the adult human being made of 10^{14} cells of various types and functions.

References

1. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* 1986; *321* : 674-9.
2. Prober JM, Trainer GL, Dam RJ, Hobbs, FW, Robertson CW, Zagursky RJ, *et al.* A system for rapid DNA sequencing with fluorescent chain terminating dideoxynucleotides. *Science* 1987; *238* : 336-41.
3. Metzker ML, Lu J, Gibbs RA. Electrophoretically uniform fluorescent dyes for automated DNA sequencing. *Science* 1996; *271* : 1420-2.
4. Reeve MA, Fuller CW. A novel thermostable polymerase for DNA sequencing. *Nature* 1995; *376* : 796-7.
5. Luckey JA, Drossman H, Kostichka AJ, Mead DA, D'Cunha J, Norris TB, *et al.* High speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Res* 1990; *18* : 4417-21.
6. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using PHRED. I Accuracy assessment. *Genome Res* 1998; *8* : 175-85.
7. Ewing B, Green P. Base-calling of automated sequencer traces using PHRED II. Error probabilities. *Genome Res* 1998; *8* : 186-94.
8. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 1980; *32* : 314-31.
9. Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB. Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol* 1982; *162* : 729-73.
10. Olson MV, Dutchik JE, Graham MY, Brodeur GM, Helms C, Frank M, *et al.* Random-clone strategy for genomic restriction mapping in yeast. *Proc Natl Acad Sci USA* 1986; *83* : 7826-30.

11. Coulson A, Waterston R, Kiff J, Sulston J, Kohara Y. Genome linking with yeast artificial chromosomes. *Nature* 1988; *335* : 184-6.

12. Adams MD, Soares MB, Kerlavage AR, Fields C, Venter JC. Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nat Genet* 1993; *4* : 373-80.

13. Papadopoulos N, Nicolaides NC, Wei YF, Ruben SM, Carter KC, Rosen CA, *et al.* Mutation of a *mutL* homolog in hereditary colon cancer. *Science* 1994; *263* : 1625-9.

14. Vogelstein B, Kinzler KW. X-rays strike p53 again. *Nature* 1994; *370* : 174- 5.

15. The yeast genome directory. *Nature* 1997; *387* (6632 suppl) : 5.

16. Ainscough R, Bardill S, Barlow K, Basham V, Baynes C, Beard L, *et al.* Genome sequence of the nematode *Caenorhabditis elegans*: a platform for investigating biology. *The C. elegans* sequencing consortium. *Science* 1998; *282* : 2012-18.

17. Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, Tachiiri Y, *et al.* Cloning and stable maintenance of 300 kilobase-pair fragments of human DNA in *Escherichia coli* using an F factor-based vector. *Proc Natl Acad Sci USA* 1992; *89* : 8794-7.

18 International Human Genome Sequencing Consortium (IHGSC). Initial sequencing and analysis of the human genome. *Nature* 2001; *409* : 860-921.

19. Adams MD, Dubnick M, Kerlavage AR, Moreno R, Kelley JM, Utterback TR, *et al.* Sequence identification of 2,375 human brain genes. *Nature* 1992; *355* : 632-4.

20. White O, Dunning T, Sutton G, Adams M, Venter JC, Fields C. A quality control algorithm for DNA sequencing projects. *Nucleic Acids Res* 1993; *21* : 3829-38.

21. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, *et al.* Comparative genomics of the eukaryotes. *Science* 2000; *287* : 2204-15.

22. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, *et al.* A whole-genome assembly of Drosophila. *Science* 2000; *287* : 2196-204.

23. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, *et al*. The genome sequence of *Drosophila melanogaster*. *Science* 2000; *287* : 2185-95.

24. Weber JL, Myers EW. Human whole-genome shotgun sequencing. *Genome Res* 1997; *7* : 401-9.

25. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, *et al.* The sequence of the human genome. *Science* 2001; *291* : 1304-51.

26. Salzberg SL, White O, Peterson J, Eisen JA. Microbial genes in the human genome: lateral transfer or gene loss? *Science* 2001; *292* : 1903-6.

27. Wang Q, Khillan J, Gadue P, Nishikura K. Requirement of the RNA editing deaminase ADAR1 gene for embryonic erythropoiesis. *Science* 2000; *290* : 1765-8.

28. Holcik M, Sonenberg N, Korneluk RG. Internal ribosome initiation of translation and the control of cell death. *Trends Genet* 2000; *16* : 469-73.

29. Rubin GM. The draft sequences. Comparing species. *Nature* 2001; *409* : 820-1 .

30. Larsen F, Gundersen G, Lopez R, Prydz H. CpG islands as gene markers in the human genome. *Genomics* 1992; *13* : 1095-1107.

31. Cross SH, Clark VH, Simmen MW, Bickmore WA, Maroon H, Langford CF, *et al.* CpG island libraries from human chromosomes 18 and 22: landmarks for novel genes. *Mamm Genome* 2000; *11* : 373-83.

32. Gilbert N, Lutz-Prigge S, Moran JV. Genomic deletions created upon LINE-1 retrotransposition. *Cell* 2002; *110* : 315-25.

33. Symer DE, Connelly C, Szak ST, Caputo EM, Cost GJ, Parmigiani G, *et al.* Human LI retrotransposition is associated with genetic instability *in vivo. Cell* 2002; *110* : 327-38.

34. Ostertag EM, Kazazian HH Jr. Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res* 2001; *11* : 2059-65.

35. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, *et al.* Recent segmental duplications in the human genome. *Science* 2002; *297* : 1003-7.

36. Collins, FS. Positional cloning moves from perditional to traditional. *Nat Genet* 1995; *9* : 347-50.

37. Nicole S, Davoine CS, Topaloglu H, Cattolico L, Barral D, Beighton P, *et al.* Perlecan, the major proteoglycan of basement membranes, is altered in patients with Schwartz-Jampel syndrome (chondrodystrophic myotona). *Nat Genet* 2000; *26* : 480-3.

38. Bomont P, Cavalier L, Blondean F, Hamide CB, Belal S, Tazir M, *et al.* The gene encoding gigaxonin, a new member of the cytoskeletal BTB/kelch repeat family, is mutated in giant axonal neuropathy. *Nat Genet* 2000; *26* : 370-4.

39. Pusch CM, Zeitz C, Brandau O, Pesch K, Achatz H, Feil S, *et al.* The complete form of X-linked congenital stationary night blindness is caused by mutations in a gene encoding a leucine-rich repeat protein. *Nat Genet* 2000; *26* : 324-7.
40. The ADHR Consortium. Autosomal dominant hypophosphataemic rickets is associated with mutations in FGF23. *Nat Genet* 2000; *26* : 345-8.
41. Sherrington R, Rogaev El, Liang Y, Rogaeva EA, Levesque G, Ikeda M, *et al.* Cloning of a gene bearing missense mutations in early onset familial Alzheimer's disease. *Nature* 1995; *375* : 754-60.
42. Rogaev El, Sherrington R, Rogaeva EA, Levesque G, Ikeda M, Liang Y, *et al.* Familial Alzheimer's disease in kindreds with missense mutations in a gene on chromosome I related to the Alzheimer's disease type 3 gene. *Nature* 1995; *376* : 775-8.
43. Shaikh TH, Kurahashi H, Saitta SC, O'Hare AM, Hu P, Roe BA, *et al.* Chromosome 22-specific low copy repeats and the 22q 11.2 deletion syndrome: genomic organization and deletion endpoint analysis. *Hum Mol Genet* 2000; *9* : 489-501.
44. Francke, U. Williams-Beuren syndrome: genes and mechanisms. *Hum Mol Genet* 1999; *8* : 1947-54.
45. Makrinou E, Collinge J, Antoniou M. Genomic characterization of the human prion protein (*PrP*) gene locus. *Mamm Genome* 2002; *13* : 696-703.
46. Blacker D, Bertram L, Saunders AJ, Moscarillo TJ, Albert MS, Wiener H, *et al.* Results of a high-resolution genome screen of 437 Alzheimer's disease families. *Hum Mol Genet* 2003; *12* : 23-32.
47. Drews, J. Research and development. Basic science and pharmaceutical innovation. *Nat Biotechnol* 1999; *17* : 406.
48. Drews, J. Drug discovery: a historical perspective. *Science* 2000; *287* : 1960-4.
49. Davies PA, Pistis M, Hanna MC, Peters JA, Lambert JJ, Hales TG, *et al.* The 5-HT_{3B} subunit is a major determinant of serotonin-receptor function. *Nature* 1999; *397* : 359-63.
50. Heise CE, O'Dowd BF, Figueroa DJ, Sawyer N, Nguyen T, Im DS, *et al.* Characterization of the human cysteinyl leukotriene 2 receptor. *J Biol Chem* 2000; *275* : 30531-6.
51. Saunders AJ, Kim TW, Tanzi RE. BACE maps to chromosome 11 and a BACE homolog, BACE2, reside in the obligate Down syndrome region of chromosome 21. *Science* 1999; *286* : 1255-6a.
52. Firestein, S. The good taste of genomics. *Nature* 2000; *404* : 552-3
53. Adler E, Hoon MA, Mueller KL, Chandrashekar J, Ryba NJP, Zuker CS. A novel family of mammalian taste receptors. *Cell* 2000; *100* : 693-702.
54. Chandrashekar J, Mueller KL, Hoon MA, Adler E, Feng L, Guo W, *et al.* T2Rs function as bitter taste receptors. *Cell* 2000; *100* : 703-11.
55. Matsunami H, Montmayeur JP, Buck LB. A family of candidate taste receptors in human and mouse. *Nature* 2000; *404* : 601-4.
56. Max M, Shanker YG, Huang L, Rong M, Liu Z, Campagne F, *et al. Tas1r3*, encoding a new candidate taste receptor, is allelic to the sweet responsiveness locus *Sac. Nat Genet* 2001; *28* : 58-63.
57. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001; *409* : 928-33.
58. Chakravarti A. To a future of genetic medicine. *Nature* 2001; *409* : 822-3.
59. Panicker SG, Reddy ABM, Mandal AK, Ahmed N, Nagarajaram HA, Hasnain SE, *et al.* Identification of novel mutations causing familial primary congenital glaucoma in Indian pedigrees. *Invest Ophthalmol Vis Sci* 2002; *43* : 1358-66.
60. Panicker SG, Sampath S, Mandal AK, Reddy AB, Ahmed N, Hasnain SE. Novel mutation in *FOXC 1* wing region causing Axenfeld-Rieger anomaly. *Invest Ophthalmol Vis Sci* 2002; *43* : 3613-6.
61. Ata-ur-Rasheed M, Vemuganti G, Honavar S, Ahmed N, Hasnain S, Kannabiran C. Mutational analysis of the *RB1* gene in Indian patients with retinoblastoma. *Ophthalmic Genet* 2002; *23* : 121-8.
62. Lander ES, Schork NJ. Genetic dissection of complex traits. *Science* 1994; *265* : 2037-48.
63. Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, *et al.* Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* 1996; *271* : 1380-7.
64. Altmuller J, Palmer LJ, Fischer G, Scherb H, Wjst M. Genome-wide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet* 2001; *69* : 936-50.
65. Peltonen L, Palotie A, Lange K. Use of population isolates for mapping complex traits. *Nat Rev Genet* 2000; *1* : 182-90.
66. Peltonen L, McKusick VA. Dissecting human disease in the postgenomic era. *Science* 2001; *291* : 1224-9.
67. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996; *273* : 1516-7.

68. Emahazion T, Feuk L, Jobs M, Sawyer SL, Fredman D, St Clair D, *et al.* SNP association studies in Alzheimer's disease highlight problems for complex disease analysis. *Trends Genet* 2001; *17* : 407-13.

69. Martin ER, Scott WK, Nance MA, Watts RL, Hubble JP, Koller WC *et al.* Association of single-nucleotide polymorphisms of the *tau* gene with late-onset Parkinson disease. *JAMA* 2001; *286* : 2245-50.

70. Rusnak JM, Kisabeth RM, Herbert DP, McNeil DM. Pharmacogenomics: a clinician's primer on emerging technologies for improved patient care. *Mayo Clin Proc* 2001; *76* : 299-309.

71. Schork NJ, Fallin D, Lanchbury JS. Single nucleotide polymorphisms and the future of genetic epidemiology. *Clin Genet* 2000; *58* : 250-64.

72. Kuehl P, Zhang J, Lin Y, Lamba J, Assem M, Schuetz J, *et al*. Sequence diversity in CYP3A promoters and characterization of the genetic basis of polymorphic CYP3A5 expression. *Nat Genet* 2001; *27* : 383-91.

73. Olivier M, Aggarwal A, Allen J, Almendras AA, Bajorek, ES, Beasley EM, *et al*. A high-resolution radiation hybrid map of the human genome draft sequence. *Science* 2001; *291* : 1298-1302.

74. O'Brien SJ, Menotti-Raymond M, Murphy WJ, Nash WG, Wienberg J, Stanyon R, *et al*. The promise of comparative genomics in mammals. *Science* 1999; *286* : 458-62, 479-81.

75. McConkey EH, Varki A. A primate genome project deserves high priority. *Science* 2000; *289* : 1295-6.

76. Strausberg RL, Feingold EA, Klausner RD, Collins FS. The mammalian gene collection. *Science* 1999; *286* : 455-7.

77. Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE. Human-mouse genome comparisons to locate regulatory sites. *Nat Genet* 2000; *26* : 225-8.

78. Robertson KD, Wolffe AP. DNA methylation in health and disease. *Nat Rev Genet* 2000; *1* : 11-9.

79. Beck S, Olek A, Walter J. From genomics to epigenomics: a loftier view of life. *Nat Biotechnol* 1999; *17* : 1144.

80. Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, *et al.* A gene expression map for *Caenorhabditis elegans*. *Science* 2001; *293* : 2087-92.

81. Nestler EJ, Landsman D. Learning about addiction from the genome. *Nature* 2001; *409* : 834-5.

Reprint requests : Dr S.E. Hasnain, Centre for DNA Fingerprinting and Diagnostics, Nacharam Hyderabad 500076, India