Human Genome Project and After

Meenakshi Lallar and Shubha R Phadke

Department of Medical Genetics, Sanjay Gandhi Post Graduate Institute of Medical Sciences, Lucknow, India

Email: shubharaophadke@gmail.com

"What more powerful form of study of mankind could there be than to read our own instruction book" – Francis S. Collins, Director, NIH.

"Along with Bach's music, Shakespeare's sonnets, and the Apollo Space Program, the Human Genome Project (HGP) is one of those achievements of the human spirit that makes me proud to be a human" – Richard Dawkins, British ethologist & evolutionary biologist.

History

The idea of sequencing the entire human genome was first proposed by the US Department of Energy and others in 1985. There were strong supporters who argued that deciphering the human genome would lead to new understanding and benefits for human health. However, there were determined detractors, who feared such a project would be waste of manpower and money and would provide a meaningless string of letters, with little explanatory power for humans. HGP was officially initiated in the USA as a joint effort of the Department of Energy and the National Institutes of Health (NIH) in 1990 with a 15-year, $3 billion plan for completing the genome sequence. But it was completed much before time in 13 years and with $2.7 billion (Table I).

Principles

Firstly, the HGP welcomed collaborators from any nation in an effort to move beyond borders, to establish an all-inclusive effort aimed at understanding the molecular basis of human life. This was planned to be done using different approaches. The group of publicly funded researchers that eventually assembled (18 countries and more than 200 laboratories) was known as the International Human Genome Sequencing Consortium (IHGSC). It is unfortunate that India was not among the 18 countries. Secondly, HGP worked on the Bermuda principles drafted in 1996. This required that all human genome sequence information, greater than 2 kb, be made freely and publicly available within 24 hours of its assembly. This was usually done by uploading all the sequences on the very same night of sequencing. This provided researchers all around the world, access to HGP data and greatly accelerated the ongoing research. A number of terms and definitions were introduced because of the HGP, some of which are given in Table II.

Goals

The ultimate goal of the Human Genome Project was to decode the exact sequence of all 3.2 billion nucleotide bases that make up the human genome and to identify estimated genes in the human DNA [Collins et al., 1998].

- It strived to achieve coverage of at least 90% of the human Genome in working draft by the end of 2001 and to finish one-third of the human Genome sequence by the end of 2001. HGP also aimed to complete human Genome sequencing by the end of 2005.
- To develop databases to store all the information generated through HGP.
- To develop faster, more efficient sequencing technologies.
- Develop tools for data analysis (Bioinformatics).

  One of the goals of HGP was to decipher the genome of organisms like mice, fruit flies and roundworms. Manipulations on such small organisms are easier and hence experiments based on them, especially breeding provides vital information about developmental and

1990	HGP started; ELSI program founded
1992	Second generation Human Genetic Map developed
1994	HGP's human genetic mapping goal achieved
1995	HGP's human physical mapping goal achieved; First bacterial genome (H. influenzae) sequenced
1996	First human gene map established; Pilot project for Human Genome sequence began in the US; Yeast genome sequenced; Bermuda principles established
1997	E. coli genome sequenced
1998	(C. elegans) genome sequenced; Celera genomics entered HGP race (Private HGP)
1999	Full scale Human Genome sequence began; Sequence of chromosome 22 completed
2000	Draft version of both Public and Private HGP completed; Fruit fly genome sequenced; Executive order issued barring genetic discrimination in US federal workplace
2001	Draft version of HGP published
2002	Draft version of mouse genome sequenced, completed and published; Draft version of rice genome sequenced, completed and published
2003	HGP completed with all its goals achieved

Table 1 Milestones of the Human Genome Project.

BAC	Bacterial artificial chromosome vector carrying a genomic DNA insert, typically 100±200kb
Contig	A contiguous sequence of DNA created by assembling shorter, overlapping sequenced fragments of a chromosome (whether natural or artificial, as in BACs).
Scaffold	The result of connecting contigs by linking information from paired-end reads from plasmids, paired-end reads from BACs, known messenger RNAs or other sources. The contigs in a scaffold are ordered and oriented with respect to one another.
Sequence tagged site	STS stands for sequence tagged sites, a short DNA segment that occurs only once in a genome and whose exact location and order of bases is known.
Genetic map	A genome map in which polymorphic loci are positioned relative to one another on the basis of the frequency with which they recombine during meiosis. The unit of distance is centimorgans (cM), denoting a 1% chance of recombination.
Physical map	A map showing the locations of identifiable markers spaced along the chromosomes. A physical map may be constructed from a set of overlapping clones.
Functional genomics	The study of genomes to determine the biological function of all the genes and their products.
Draft genome sequence	The sequence produced by combining the information from the individual sequenced clones (by creating merged sequence contigs and then employing linking information to create scaffolds) and positioning the sequence along the physical map of the chromosomes.
Methylation	Addition of methyl groups to DNA to suppress gene transcription.
SNP	Single Nucleotide Polymorphism (SNP): A common single-base-pair variation in a DNA sequence.
Haplotype	A specific combination of alleles or sequence variations that are likely to be inherited together.
Genomic library	Contains DNA fragments that represent the entire genome of the organism (coding and non coding).

Table 2 Terms and definitions.

functional genetics which can be applied to human health and diseases. Also, studying different genomes would give us insight into the evolutionary conservation of genes and development of unique genes. It could also lead to understanding which would help in combating human diseases.

- Rapid identification of Single Nucleotide Polymorphisms (SNPs).
- Functional genomics (cDNA clones of human and model organisms).
- Manpower training.
- And lastly, one of the most significant issues addressed were the ethical, legal, and social issues (ELSI) that arose from HGP. Around 3-5% budget of HGP was reserved for ELSI. This is, to date the largest Bioethics program undertaken.

The Genome Wars

Celera Genomics (‘Celera’ is Latin for swiftness) was established in 1998 by the Perkin Elmer Corporation and Craig Venter. Earlier, Craig Venter was head of TIGR, a non-profit genomics research institution where using unique whole genome shotgun sequencing method, he had sequenced the genome of *H. influenza* [Weber & Myers, 1997]. Celera Genomics announced that they would finish the human genome sequencing in three years. The establishment of Celera Genomics heralded a race between the government’s HGP and Celera. Celera began its project later than the HGP, but used a faster method powered by the world’s largest private assemblage of supercomputers. Celera planned for preliminary patents on over 6,000 genes and full patents on a few hundred genes before releasing their sequence. However, a significant portion of the human genome had already been sequenced when Celera entered the field, and thus Celera did not incur any costs in obtaining the existing data, which was freely available to the public from Genbank. Celera sequenced the human genome at a fraction of the cost of the public project, approximately $3 billion of taxpayer dollars versus about $300 million of private funding. In 2001 however, with the intervention of the White House, both Celera and Public HGP officially ”tied” and made the joint official statement of initial HGP draft completion in 2000 (Fig 1).

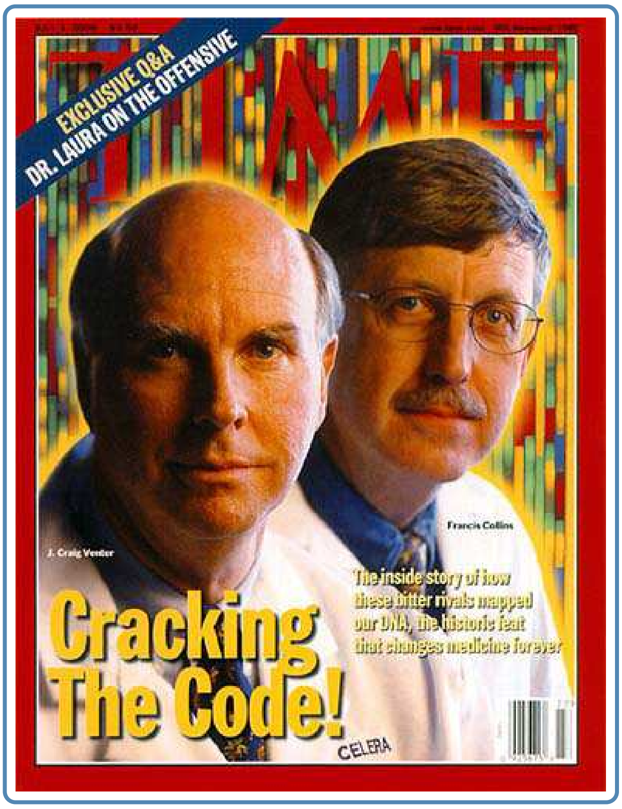
TIME
EXCLUSIVE Q&A
DR. LAURA ON THE OFFENSIVE
J. Craig Venter
Francis Collins
The inside story of how these bitter rivals mapped our DNA, the historic feat that changes medicine forever
Cracking The Code!
CELERA

Figure 1 Time magazine cover showing Craig Venter head of Celera Genomics who led the private HGP and Francis Collins, head of NIH, who led the Public HGP.

Discoveries and inventions that initiated the need and concept of HGP

Several key projects helped to crystallize the HGP. These included:

1. The sequencing of the bacterial viruses, animal virus SV407 and the human mitochondrion between 1977 and 1982. These projects proved the feasibility of assembling small sequence fragments into complete genomes.
2. The human genetic map had been made, which made it possible to locate disease genes of unknown function based solely on their inheritance patterns.
3. The physical maps of clones covering the yeast and worm genomes were made in the mid 1980s, which allowed isolation of genes and regions based solely on their chromosomal position.

4. The development of random shotgun sequencing of complementary DNA fragments for high-throughput gene discovery.

The Public Human Genome Project: Techniques used

Two main phases of the public HGP were:

1. *Shotgun phase* – Hierarchical/ BAC (Bacterial Artificial Chromosome) based/ map based/ clone by clone method (based on insights from yeast and worm studies) - yielded 90% of the human genome sequence in draft form.

2. *Finishing phase* – Year 2000 and onwards, the work mainly involved filling in gaps and resolving DNA sequences in ambiguous areas not obtained during the shotgun phase

- Shotgun phase: Blood samples were taken from anonymous donors and DNA was extracted. The final DNA sample to be sequenced was chosen after multiple phases of blinding, so that the donor and laboratory, both did not know whose DNA was being sequenced.

- STS- tagged BAC Clone Libraries: DNA was broken into 100-200 kb segments which were then combined with bacterial plasmids to form BAC clones. When possible, the DNA fragments within the library vectors were mapped to chromosomal regions by screening for sequence-tagged sites (STSs), which were DNA fragments, usually less than 500 base pairs in length, of known sequence and chromosomal location that could be amplified using PCR. These BAC clones (~830,000) were then transported to different labs all over the world. Different labs were assigned different chromosomes and sections of DNA. In these labs these BAC clones were further divided to even smaller 200 bp fragments to make BAC subclones. From these BAC subclones plasmid and human DNA was separated. Hence, through BAC cloning multiple copies of DNA fragments were obtained.

- Sequencing: The inserts were sequenced using primers matching the vector sequence flanking the genomic DNA insert by using Sanger technique to form Contigs which were arranged into Scaffolds (progressively larger Contigs) using computational analyses by identifying overlaps.

The Human Genome Project relied upon the physical and genetic maps of the human genome established earlier, to map the BAC clones (STS tagged) which served as a platform for generating and analyzing the massive amounts of DNA sequence data that emerged from the shotgun phase (Figure 2).

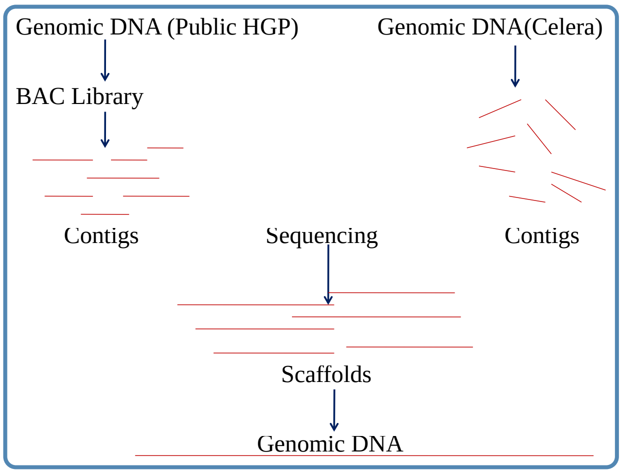

Figure 2 Randomly sequenced fragments aligned by computational analysis, by combining sequences with overlapping ends, into stepwise larger segments.

The Private Human Genome Project (Celera Genomics): Techniques used

Celera used two independent data sets together with two distinct computational approaches to determine the sequence of the human genome. The first data set was generated by Celera using DNA of five anonymous individuals, out of which one DNA was selected blindly. Plasmid clones were made. Sequencing and tracking from both ends of plasmid clones from 2, 10, and 50 kb libraries was done. This generated 27.27 million DNA sequence reads (5.11 fold coverage of the genome). The second data set was obtained from the publicly funded Human Genome Project (BAC Clones); here, Celera "shredded" the Human Genome Project DNA sequence into 550-base-pair sequence reads representing a total of 16.05 million sequence reads. The company then used a whole-genome assembly method and a regional chromosome assembly method to sequence the human genome (Figure 3).

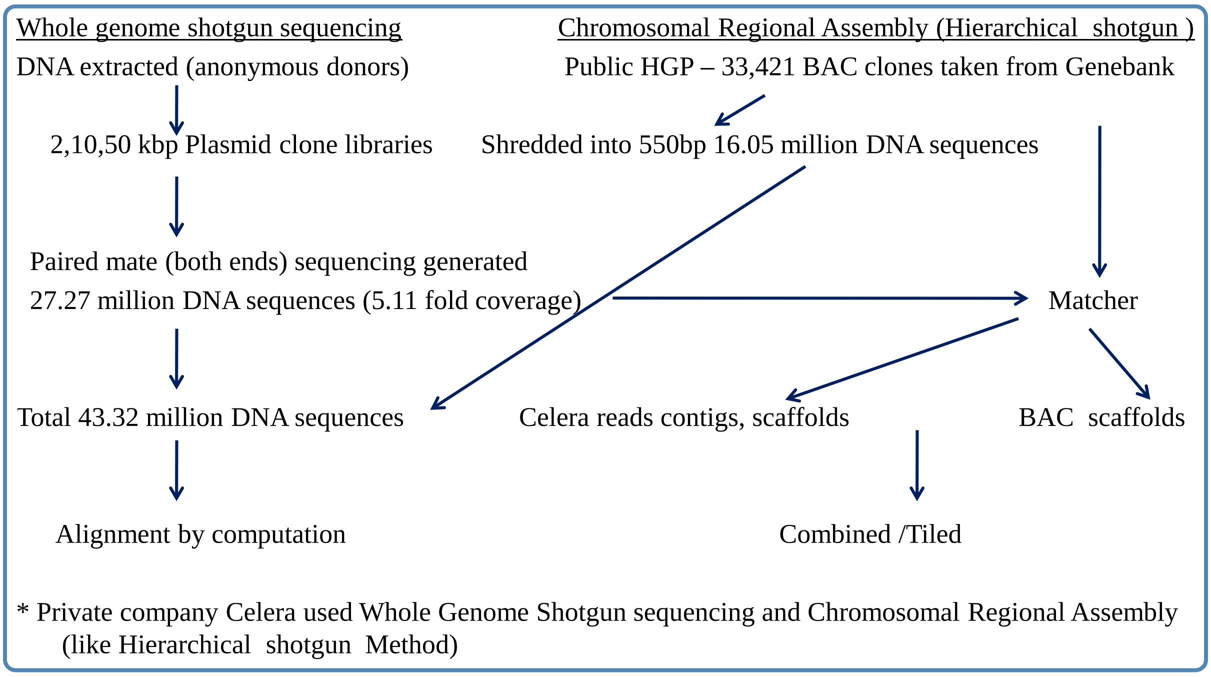
Whole genome shotgun sequencing
DNA extracted (anonymous donors)
2,10,50 kbp Plasmid clone libraries
Paired mate (both ends) sequencing generated
27.27 million DNA sequences (5.11 fold coverage)
Total 43.32 million DNA sequences
Alignment by computation
Chromosomal Regional Assembly (Hierarchical shotgun)
Public HGP – 33,421 BAC clones taken from Genebank
Shredded into 550bp 16.05 million DNA sequences
Matcher
Celera reads contigs, scaffolds
BAC scaffolds
Combined /Tiled
* Private company Celera used Whole Genome Shotgun sequencing and Chromosomal Regional Assembly (like Hierarchical shotgun Method)

Figure 3 Approach used by the Private Human Genome Project.

- Whole-genome assembly method / whole-genome random shotgun method: In the method, Celera generated a massive shotgun library derived from its own DNA sequence data combined with the ”shredded” Human Genome Project DNA sequence data, which together corresponded to a total of 43.32 million sequence reads. Celera then used computational methods and sophisticated algorithms to identify overlapping DNA sequences and to reconstruct the human genome by generating a set of scaffolds.

- Regional chromosome assembly method: Here Celera organized its own data and the Human Genome Project sequence data into the largest possible chromosomal segments, followed by shotgun assembly of the sequence data within each segment (similar to the hierarchical shotgun approach used by the Public HGP). Celera's whole-genome and regional chromosome assembly methods were independent of each other, permitting direct comparison of the data. Celera found that the regional chromosome assembly method was slightly more consistent than the whole-genome assembly method. Many people criticized that although whole genome shotgun method was quick and less expensive, it had more misassembles. This was later refuted as Celera data was in strong agreement with the public HGP data. This whole genome shotgun method eventually paved way for the next generation sequencing technique popular contemporarily.

Human Genome Project: draft completion announcement

The completed HGP Drafts were published separately by the Public HGP (Nature) and the Private Human Genome Project (Science) a day apart on 15th February and 16th February 2001 respectively [IHGSC,2001; Venter et al., 2001]. However, the formal HGP Draft completion announcement was made public on 26 June, 2000 jointly by the two organizations.

Economic impact of the HGP

The $3.8 billion spent on the HGP may well represent the best single investment ever made in science. From 1988 to 2010, HGP added $796 billion in U.S. economic output, $244 billion in personal income for Americans and generated 3.8 million jobyears of employment [Lander, 2011].

Outcome of HGP

Along with sequencing the Human genome the genome sequences of 599 viruses and viroids, 205 naturally occurring plasmids, 185 organelles, 31 eubacteria, seven archaea, one fungus (yeast), two animals (round worm and mouse) and one plant (mustard weed) were identified [IHGSC, 2004].

Sequencing of 99% of euchromatic DNA was finished to 99.99% accuracy. The number of genes in the human genome was estimated to be around 22,000 (~1.5%) and 3.7 million SNPs were mapped (Table III).

Other major feats were:

- Higher resolution physical genetic maps of the human DNA and 15,000 full-length human cDNAs were generated.
- Potential drug targets were established.
- A summary of each chromosome was composed.

- Comparative genomics:

- The basic transcriptional and translational machinery is well known to have been conserved over evolution, from bacteria through to the most complex eukaryotes. Many ribonucleoproteins involved in RNA splicing also appear to be conserved among the animals.
- The Human Genome Project discovered that about 60 percent of genes are conserved between fly and human. Genes in humans were only twice in number as compared to other simpler organisms. But the complexity was explained by the fact that alternative splicing generates a larger number of protein products.
- The euchromatic portion of the human genome has a much higher density of transposable element copies than the euchromatic DNA of the other three organisms (mice, fruit flies and roundworms.)
- The vertically transmitted, long-term residential LINE and SINE elements represent 75% of interspersed repeats in the human genome, but only 5±25% in the other genomes. In contrast, the horizontally transmitted and shorter-lived DNA transposons represent only a small portion of all interspersed repeats in humans (6%) but a much larger fraction in fruitfly, mustard weed and worm (25%, 49% and 87%, respectively).
- The pericentromeric and subtelomeric regions of chromosomes are filled with large recent segmental duplications of sequence from elsewhere in the genome. Segmental duplication is much more frequent in humans than in yeast, fruitfly or worm.

- Bioinformatics: The human genome is 3 billion base pairs long; not only 3 gigabytes of computer data storage space are needed to store this raw data but a huge amount of storage space is required to store the annotations of this data which is being deciphered at an exponential rate. Hence, Bioinformatics has emerged as the need of the hour. It started with sequence alignment programs like FASTA and BLAST in the 1990s. One of the goals of the HGP was to develop Bioinformatics. By the end of HGP, databases like the UCSC browser, Ensembl browser, the NCBI chromosome map etc.

Genome Sequencing	HGP begins-1990	HGP ends-2003	12 years after HGP
Cost to Generate a Human Genome Sequence	~$1 billion	10-50 million	$1000
Human Genome Sequences	0	1	Thousands
Number of Genes with Known Phenotype/ Disease-Causing Mutation	~53	~1474	~9578
Number Phenotypes/Disorders with Known Molecular Basis	~61	~2264	More than 6000
Number of Published Genome-Wide Association Studies (GWAS)	0	0	~1600

Table 3 Consequences of the Human Genome Project.

were formed. Newer programs are being generated to aid gene prediction, sequence alignment and molecular modeling and 3-D visualization of proteins.

- ELSI program: (which was started in 1990) is still ongoing under the flagship of the National Human Genome Research Institute (NHGRI) and it aims to provide for basic and applied research on the ethical, legal and social implications of genetic and genomic research for individuals, families and communities that emerged out of the HGP. It funds and manages studies, workshops, research projects and conferences catering to the social and ethical impact of genomic medicine.

Post Human Genome Project Era

Since the completion of HGP there has been a revolution in the field of genetics and now we have entered the –omics era. Scientists, world over, are studying different microbial, plant, animal, human and cancer genomes. New genomes of different organisms are being coded. Among the most important contemporary human genome related-projects are the International HapMap Project, 1000 Genomes Project, ENCODE project, Human Epigenome Project, 100,000 Genomes Project (UK) and 1,000,000 Genomes Project (US) etc.

Concluding thoughts

The completion of HGP transformed genetics into genomics. High throughput sequencing techniques and microarray technology has made analysis of genome simpler and cheap. This has transformed not only research but patient care as well. Identification of causative genes for monogenic disorders, genome wide association studies for multifactorial disorders and genomic and expression analysis of tumors are important applications of results of the HGP. As this has opened a new exciting era for clinicians and geneticists, the challenges are also apparent. The challenges include the need of functional validation of each new sequence variant identified and better understanding of modifier genes to predict pathogenicity and to develop better understanding of genotype-phenotype correlation. Now we know that the more we learn about the human genome, the more there is to explore. And these words from the HGP Draft article befit our concluding thoughts.

"We shall not cease from exploration. And the end of all our exploring will be to arrive where we started, and know the place for the first time." – T. S. Eliot.

References

1. Collins FS, et al. New goals for the U. S. Human Genome Project: 1998-2003. Science 1998; 282: 682-689.
2. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 2001; 409: 860-921.
3. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004; 431: 931-945.
4. Lander ES. Initial impact of the sequencing of the human genome. Nature 2011; 470.
5. Venter JC, et al. The sequence of the human genome. Science 2001; 291: 1304-1351.
6. Weber JL, Myers EW. Human whole-genome shotgun sequencing. Genome Res 1997; 7: 401-409.