CSIR Scientists at the Indian Institute of Genomics and Integrative Biology (IGIB), have successfully mapped the first complete and entirely Indian genome. The breakthrough holds potential to usher in the era of affordable and predictive healthcare.
WITH the first sequencing of the entire genome of a human being, India has finally made its way into an extremely elite Scientific Club. The only countries until now that had mapped the entire genetic material of a human being—the Human Genome—are the USA, UK, China, Canada and Korea. This giant step forward was made possible by CSIR scientists working at the Indian Institute of Genomics and Integrative Biology (IGIB), Delhi who used genetic matter sourced from a healthy male to present the first complete and entirely Indian genome.

Rajesh S. Gokhale, Director, IGIB has expressed the desire to sequence ten other Indian genomes shortly.

This un-named Indian male is from Jharkhand and is in his early fifties, is 167 cm tall and weighs a modest 52 kg. But then, this is only the beginning. Rajesh S. Gokhale, Director, IGIB has expressed the desire to sequence ten other Indian genomes shortly.

The first Indian human genome sequencing (determining the exact order of the base pairs in a segment of DNA) not only marks a national milestone but also sets the stage for India’s entry into personal genomics, opening up new possibilities in disease diagnostics and treatment. Of course, it could take more than a decade before the fruits of this labour really start flowing in (see Interview with Dr Rajesh Gokhale). This would also open up new opportunities for affordable healthcare in India.

Getting on to the Job

The project had been conceived by Prof. Samir K. Brahmachari, now Director-General, CSIR while he was at the IGIB. The team that made the breakthrough possible worked under the leadership of Dr. Rajesh S. Gokhale, Director-IGIB. Dr. Sridhar Sivasubbu and Dr. Vinod Scaria, Scientists at IGIB were the primary guides, and six students of IGIB made up the rest of the team that did the sequencing.

The team at IGIB generated over 51 Gigabases of data using next-generation sequencing technology that enables massively parallel sequencing of millions of genomic fragments of 76 base pairs. This involves handling a large amount of data and capability to use advanced computational tools and techniques, to retrieve the nucleotide sequence of each of these small fragments of DNA. These small fragments are then mapped back to the reference genome, akin to solving a huge jigsaw puzzle.

The IGIB team used advanced software algorithms for this number crunching exercise and used the CSIR Supercomputing Facility at IGIB to perform this humongous task. The CSIR Supercomputing Facility at IGIB has a cluster of computers that can deliver a processing power estimated at a whopping Four teraflop (Four trillion floating-point operations per second), which is roughly 10,000 times more than an average desktop priced at Rs 30,000.
The Indian genome has revealed a large number of hitherto unknown variations in the genome including single nucleotide changes and insertions/deletions. Some patterns of variations have been found, which in the western population have been associated with colorectal cancer. However, as Dr. Sivasubbu hastens to point out, it is too early to categorically state that this association will hold true for Indian populations also. In fact, it is actually too early to start hazarding any guesses. According to Dr. Gokhale, establishing the sequence is only the “first step”. The computational analyses of the genes will take much longer... perhaps even a year.

All genetic variations are important because not only is variation the key to evolution, it is also very often the key to changed physiological processes that may be manifested as diseases. Even a minor change in the base-pair sequence could actually translate into a crucial change in some important protein.

Understanding the functional role of these variations will enhance our understanding of human genetic variability and also, as a bonus to pure knowledge, serve as a starting point for pinpointing disease-associated variations. This then, would provide the take-off platform to usher in the era of affordable yet predictive healthcare. It is indeed a matter of great promise that thanks to genome sequencing, researchers can locate a gene suspected of involvement in an inherited disease in a few days and not years as used to be the case earlier.

The importance of the genome study becomes apparent when we realize that there are now over 1000 genetic tests for human conditions. These tests have value in predicting individual risks for disease. Also, at least 350 biotechnology-based products resulting from the Human Genome studies are in clinical studies.

Only about 14 individual genomes have been sequenced worldwide—most famously, the genomes of Craig Venter (HuRef) and James Watson. Just about a couple of years ago (11 October 2007), the Beijing Genomics Institute, China announced the completion of the first diploid genome sequence of a Han Chinese. The genome, named YH, marks the start of YanHuang Project, which aims to sequence 100 Chinese individuals in three years.

With the first decoding of the human genome, India is now in the select league of countries who have demonstrated the capability to sequence and assemble a complete Human Genome.

Genome Sequencing: From Then to Now
Genome sequencing has had a long history culminating in the Indian high point of December 2009. It is a major achievement but scientists have trod a long path to get here. The complete genomes of 1150 organisms (http://www.genomesonline.org/gold.cgi), both animals and plants, have been sequenced since 1995. Many other genome projects are still on.
The Genome News Network provides details of many of these organisms beginning, appropriately enough, with A for the primitive Archaean *Aeropyrum pernix* and ending with Y for the bacteria *Yersinia pestis*. Many would say that since the last word is far from having been said it is apt that the list is not an A-Z one. Be that as it may, there is little doubting the fact that the common man feels more emotional empathy towards news of the decoding of the human genome although that means a backbreaking sequencing of 3.1 billion base pairs.

The first Human Genome Sequencing initiative was conceived as early as 1984. Sequencing of a human genome requires high computational capability and technological know-how in handling sophisticated machines and analyzing huge volume of data. However, scientists did not hesitate in formally launching the Human Genome Project in the 1990s even though technologies and machinery were no match for their sophisticated counterparts today. It would be a 13-year effort coordinated by the U.S. Department of Energy and the National Institutes of Health and would end in a nail-biting race between two groups.

The goals of the Human Genome Project were to:
- identify all the approximately 20,000-25,000 genes in human DNA,
- determine the sequences of the 3 billion chemical base pairs that make up human DNA,
- store this information in databases,
- improve tools for data analysis,
- transfer related technologies to the private sector, and
- address the ethical, legal, and social issues (ELSI) that may arise from the project.

On 4 September 2007, Prof. Brahmachari, DG-CSIR, went on record saying that it would be, “a boon for countries with rich genetic resources”.

On 26 June 2000, the International Human Genome Sequencing Consortium announced the rough draft of the human genome sequence. The draft sequence covered 90% of the genome at an error rate of one in 1,000 base pairs, but there were more than 150,000 gaps and only 28% of the genome had reached the finished standard. In April 2003, came the fine-tuned version of the human genome sequence. It had less than 400 gaps and 99% of the genome was finished with an accuracy rate of less than one error every 10,000 base pairs.
There is a sense of inexorable progress in the journey of the first vertebrate group—from the fishes to the ultimate vertebrate, the human beings. So too there is poetic inevitability in the fact that the team at IGIB, earlier this year, came out with the complete genome of the Zebrafish (Danio rerio) and followed it up by the time of the year’s ending with the complete genome of an Indian.

The Zebrafish is a popularly used organism for modelling human diseases. The first sequence and analysis of the human genome was published in Science in 2001 by Dr. Venter and his colleagues at Celera Genomics. The publicly funded genome project also published its version of the human genome at the same time in the journal Nature. There was global scientific jubilation at reading, for the first time, the complete genetic blueprint of a human being.

The cost of genome sequencing, however, was still prohibitively high. Just six years ago it cost a billion US dollars to sequence a genome. However, that did not deter scientists from speculating the path research would take if it became possible to do the complete sequencing of an individual under $1,000. The journal Nature Genetics even posed it as the “Question of the Year” in 2007.

On 4 September 2007, Prof. Brahmachari, DG-CSIR, went on record saying that it would be, “a boon for countries with rich genetic resources”. His reasoning was that Indian ethnic groups comprising about one-sixth of the humanity, with large family sizes,
The Zebrafish genome was the right project to attempt before tackling the human genome. It is just half the size of the human genome and gave the scientists at IGIB enough confidence to double their target genes for their next project. CSIR has generously funded FishMap (http://fishmap.igib.res.in)—a unified and centralized resource for storage, retrieval, and display of genomic information of Zebrafish and which is an integral resource for community participation. There is also a shared genomics resource for community annotation of the Zebrafish genome (Zebrafish GenomeWiki: http://fishwiki.igib.res.in) where members of the Zebrafish community can annotate, comment and edit existing data sets using the Zebrafish GenomeWiki portal.

According to Prof. Brahmachari, full genome sequencing could be undertaken of:“(i) 20 samples from 25 distinctly different populations, to map genetic variation; (ii) 10 pairs of identical twins (five male and five female) from each of these 25 populations, to map post-zygotic repeat instability and copy-number variation; (iii) 1,000 naturally aborted foetuses (ethically acceptable as research subjects in India) that do not show cytogenetically detectable chromosomal abnormalities, in order to identify lethal deletions, insertions, duplications or other mutations, thereby identifying genomic regions and pathways essential for development; and (iv) 500 healthy (octogenarian) individuals taking no medication, drawn from 25 different ethnic groups, who will serve as ‘super-normal’ controls in case-control studies to identify disease genes. All the variation observed in this set is to be treated as neutral variation for healthy living.” Within two years the $1,000- price tag does not seem absurdly low.

In August 2009, it was reported that engineer Stephen R. Quake of the Stanford University had invented a new technology for decoding DNA. He calls it the Heliscope Single Molecule Sequencer and has used it to decode his own genome for less than $50,000. He went on record to say that the low cost “will democratize access to the fruits of the genome revolution” by enabling almost any organisation/institute to decode whole human genomes. Budget would not be a major bottleneck any more.

Interestingly, the Indian genome was decoded at a cost of Rs. 15-20 lakhs (~$ 30,000), which is substantially lower than $50,000. Of course this does not include the cost of high-throughput sequencers, supercomputers, other infrastructure or manpower. The dream is now to bring down the costs such that every person finds it an affordable option. In fact, Dr. Gokhale says that the day is not too far when banks may give loans to enable one to get one’s genome decoded. Compare this situation to the close to 3 billion US dollars in expenditure when the first genome was decoded less than a decade back.

The time taken to decode genomes is also being whittled down. While the first human genome sequencing took over a decade, CSIR completed decoding the Indian genome in 45 days and Dr. Quake’s machine can sequence a human genome in four weeks with a staff of just three people. Things can only get better.
The IGIB team used advanced software algorithms for this number crunching exercise and used the CSIR Supercomputing Facility at IGIB to perform this humongous task. The CSIR Supercomputing Facility at IGIB has a cluster of computers that can deliver a processing power estimated at a whopping Four trillion floating-point operations per second speed.

**Human Genome Variation**

However, despite the fact that as human beings we are undoubtedly made using the same blueprint and the genetic sequences of different people are remarkably similar, there are variations too. Roughly at about one in every 1,200 bases, on average, the sequences differ between two individuals. There are over 3 million differences at the basic genetic level between two unrelated individuals. This is the perfect example of unity in diversity—we are the same but we are pretty different too, different enough to be unique.

Variations at the level of single nucleotides are commonly referred to as single nucleotide polymorphisms (SNPs). Most variations arise randomly and do not affect survival. The HuRef analysis, for example, found 3.2 million SNPs and nearly one million non-SNP variants. Genetic variations hold the key to adaptations and sometimes, survival! The importance of structural variation to human health and common genetic disease is now widely accepted. Complex disorders are a consequence of the cumulative effect of a large number of variations that independently have small effects that are individually neither necessary nor sufficient to cause the disease. Of course, environment, diet and lifestyle are often triggering factors.

The quest then became one of finding variations and linking these to diseases/disease-predisposition.

**Haplotype Project**

Haplotypes are groups of linked variations along the chromosomes. The HapMap is a catalogue of common genetic variants that occur in human beings. The International HapMap Project is a global partnership of Canada, China, Japan, Nigeria, UK and USA. This multi-country effort aims to identify and catalogue genetic similarities and differences in human beings. The goal is to compare the genetic sequences of different individuals to identify chromosomal regions where genetic variants are shared. By identifying most of the approximately 10 million SNP variants estimated to occur in the human genome, the International HapMap Project is identifying the basis for a large fraction of the genetic diversity in the human species.

**Human Genome Structural Variation Project**

There are also other human genome variation projects that provide data access and resources. This includes the NHGRI Structural Variation Project and the Genomic Structural Variation Consortium. There are also several projects that involve whole genome sequencing of individuals, the data from which can be mined for structural variants. These include the HuRef project, James Watson’s Personal Genome Sequence and YanHuang. The database of Genomic Variants is perhaps the most comprehensive database for the deposition, retrieval and visualization of phenotypically normal human structural variation.

In May 2008, *Nature* reported the high-resolution integrated genetic mapping and sequencing of structural variation from eight human genomes. These eight individuals were part of the first phase of the Human Genome Structural Variation Project and included four individuals of Yoruba Nigerian ethnicity and four individuals of non-African ethnicity. It studied genetic variation on an intermediate scale—particularly insertions, deletions and inversions affecting from a few thousand to a few million base pairs.

**Indian Genome Variation Project**

India fired the first salvo with respect to mining the secondary data with the CSIR-led Indian Genome Variation project. Spearheaded by Prof. Samir Brahmachari, this was a huge project involving over half a dozen Institutes and more than 150 CSIR scientists. The CSIR research laboratories included IGIB, Centre for Cellular and Molecular Biology (CCMB), Hyderabad, Indian Institute of Chemical Biology (IICB), Kolkata, Central Drug Research Institute (CDRI), Lucknow, Industrial Toxicological Research Centre (ITRC), Lucknow, and the Institute of Microbial Technology (IMTECH), Chandigarh.

Besides, the Indian Statistical Institute (ISI), Kolkata, Anthropological Survey of India and The Centre for Genomic Applications (TCGA), a public-private partnership of CSIR were also closely involved in the study. The study aimed at studying the diversity of 1000 biomedically important and pharmaco-genetically relevant genes in normal, healthy individuals in Indian populations. This, in the second phase, followed up with a genome-wide analysis of about 50,000 SNPs.

The Indian Genome Variation project successfully completed mapping the genetic diversity of the population and it is a huge melting pot of human diversity brought about by wave upon wave of human migrations followed by admixtures with the existing populations.

This project provides the baseline data for other projects seeking to provide answers to vexing questions such as which genetic marker(s) predisposes an individual to a certain genetic disease/disorder, how do certain groups of people respond to a given drug. It focused on actions that can be initiated with the current levels of knowledge with disorders where the number of genes involved is known, for example, eye/muscular problems, diabetes and drug metabolism/response genes.
The first Indian human genome sequencing not only marks a national milestone but also sets the stage for India’s entry into personal genomics, opening up new possibilities in disease diagnostics and treatment.

Mapping Asian Genetic Diversity
Perhaps the success of the Indian study paved the way for the larger initiative of mapping the human genetic diversity in Asia, which is the world’s largest continent. Approximately 4 billion people (60% of the world’s population) live here. Not surprisingly, the genetic diversity of Asia is staggering…it is a macrocosm of India.

It is generally agreed that the original ancestral population spread out from Africa to ultimately colonise the world. Of course, this did not happen in one day. Nor was there just one journey out of Africa. However, it is generally believed that modern humans evolved in Africa and spread across the world, adapting locally to the pressures of the climates, food sources and pathogens. Those that could adapt, survived. They left behind children to carry on their genetic heritage. Those that could not, died. Their direct genetic trail (apart from rare fossils) vanished.

However, just as a wake can provide a clue about the direction a ship has taken even after the craft can no longer be seen, certain telltale signs allow scientists to pick up cold-case evidence and empower them to do the tracking. “Marker” genes can help establish a connection between two geographically separated groups of people. Mutated/disease causing genes can also allow scientists to track backwards to a healthy population from which the broken gene may have originated. This also establishes “patterns” that can be studied to reveal relationships between two populations.

Certain populations may be said to be “more at risk” than others. So can individuals. Even between siblings, one may be genetically more predisposed than the other to a certain disease. Even when it comes to establishing genetic differences between two unrelated individuals, scientists have their ways to tell. Tracking genetic variations through human migrations provide clues to evolution of diseases and phenotypes.

Now, for the first time, CSIR scientists working at IGIB have worked shoulder to shoulder with the Human Genome Organisation’s Pan-Asian SNP Consortium to study Southeast Asian (SEA) and East Asian (EA) populations. 1,928 unrelated individuals representing 73 populations from 10 countries and 10 linguistic lineages from mainland China, India, Indonesia, Japan, Malaysia, the Philippines, Singapore, South Korea, Taiwan and Thailand were studied. Genotyping was performed at eight different genotyping centres. Genotyping data analysis and quality control was centralized to standardize results.

The study conducted within and between the different populations in the Asian continent showed that populations from the same linguistic group tend to cluster together. This is to say that there is considerable relatedness within ethnic/linguistic groups and genetic ancestry correlates with linguistic affiliations as well as geography.

There was also clear evidence that there was a south-to-north migration of East Asians, originating from Southeast Asia, with the majority of East Asian gene pool being derived from Southeast Asia. There was a later divergence of non-Negrito-Asian/Asian Negritos than that of Asian/European. There was a clear increase in genetic diversity from northern to southern latitudes. The study also suggested that there was a single initial entry of modern humans into the continent of Asia, instead of multiple inflows from both southern and northern routes. This indicates that Southeast Asia was the major geographic source of East Asian and North Asian populations.

The results mean different things in different contexts. While the research paper has great anthropological significance, there is another more critical use for the data presented here. Mapping relatedness with global populations reveals history of human migrations no doubt, but it also has medical relevance such as design of genetic association studies and for population-level risk assessment. The results are important in the context of public health and the study of disease distribution. The tremendous genetic diversity of Indian populations means that personalized medicines that have been developed for populations elsewhere may not be of any direct relevance in India. By contrast, it is safe to say that any pharmaceutical house seeking a specific target population will surely find one here.

As Prof. Samir Brahmchari puts it, “We have breached political and ideological boundaries to show that the people of Asia are linked by a unifying genetic thread.”

Leapfrogging
India has bridged the divide in the field of Genomics relatively swiftly. In the 1990s to 2003 when the human Genome Project was going on, India was constrained to remain a bystander due to a lack of resources. When the HapMap project was launched in 2002 as a multi-country effort aimed at identifying and cataloguing genetic similarities and differences in human beings, India chose not to join countries such as Canada, China, Japan, Nigeria (for sample), UK and USA. Instead India chose to independently launch the Indian Genome Variation Consortia to study a large number of samples on smaller number of highly informative markers.

This ultimately resulted in the creation of the Indian Genome Variation database. Next came the decoding of the entire genome of an anonymous Indian, which placed India in an elite scientific club of a few selected nations that had already decoded the genomes of individuals. With the pan-Asia study, India has forged strong Asian collaborative partnerships. This is the first time that scientists representing 60% of the world’s population have themselves carried out the work relating to their own genetic pools. This confident coming of age unequivocally proves that the technology divide has been completely bridged.

Dr Sukanya Datta is Scientist-NISCAIR, posted with DG’s Technical Cell, CSIR Hqrs. Raff Marg, New Delhi