Genomics: Introduction and Applications to Human Health

Moinuddin Ansari and Rehana Abidi

Abstract

Genomics is the study of the collective genetic material in an organism. This discipline is mainly focused on sequencing the DNA in an organism to form a complete picture. By sequencing the entire DNA pattern of an organism, scientists can generate a great deal of information. The genomes of numerous species have been sequenced, from bacteria to humans. The genome of each species is distinctly different, with varying numbers of nucleotides. Within a species, genetic variation may be minimal, but still interesting, because it can explain certain traits or tendencies. Genomics is greatly contributing to our knowledge of human health and understanding of disease. Current genomics research is focused on studies to discover association of genomic variants with disease and use of this knowledge in developing molecular diagnostics.

"We used to think our fate was in our stars. Now we know, in large measure, our fate is in our genes." ***J. D. Watson****, Nobel laureate, Physiology or Medicine (1962).*

1. Introduction

Living organisms may be of two types: prokaryotes and eukaryotes. The prokaryotes are a group of organisms that lack a cell nucleus. They differ from the eukaryotes, which have a cell nucleus. For example, Bacteria are prokaryotes, whereas Humans are eukaryotes. In these cells, the genetic material is the DNA. DNA usually occurs as linear chromosomes in eukaryotes, and circular chromosomes in prokaryotes. The information carried by DNA is held in the sequence of pieces of DNA called genes (fig. 1). The genome is the whole hereditary information of an organism that is encoded in the DNA (or for some viruses, in the RNA).

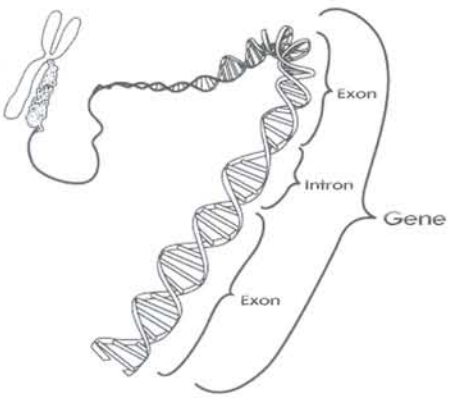

Fig. 1. Gene is a small fraction of DNA

The human genome has approximately 3 billion base pairs of DNA arranged into 46 chromosomes. Genomics is the comprehensive study of the genetic information of a cell or organism.

A gene is a unit of heredity and a region of DNA that influences a particular characteristic in an organism. Genes contain an open reading frame that can be transcribed, as well as regulatory sequences such as promoters and enhancers, which control the transcription of the open reading frame. In many species, only a small fraction of the total sequence of the genome encodes protein. For example, only about 1.5% of the human genome consists of protein-coding exons, with over 50% of human DNA consisting of non-coding repetitive sequences. The reasons for the presence of so much non-coding DNA in eukaryotic genomes and the extraordinary differences in genome size, among species represent a long-standing puzzle. However, DNA sequences that do not code protein may still encode functional non-coding RNA molecules. Among functional RNA molecules, siRNA technology is being applied in many laboratories to assess the roles of genes by loss-of-function phenotype analyses.

2. Genome-oriented Basic Research

(i) Genome Sequencing

The sequence of DNA encodes the necessary information for living things to survive and reproduce. Determining the sequence is therefore useful in fundamental research into why and how organisms live, as well as in applied subjects. Because of the key nature of DNA to living things, knowledge of DNA sequence may be useful in practically any biological research.

The first free-living organism to be sequenced was that of *Haemophilus influenzae* (1.8 Mb), a bacterium in 1995[1], and since then genomes are being sequenced at a rapid pace. As of now, the complete sequence is known of many viruses, bacterial species and eukaryote organisms. Most of the bacteria whose genomes have been completely sequenced are problematic disease-causing agents, such as *Haemophilus influenzae.* Of the other sequenced species, most were chosen because they were well-studied model organisms or promised to become good models. Yeast *(Saccharomyces cerevisiae)* has long been an important model organism for the eukaryotic cell, while the fruit fly *Drosiphila melanogaster* has been a very important tool (notably in early pre-molecular

genetics). The worm *Caenorhabditis elegans* is an often used simple model for multicellular organisms. The zebra fish *Brachydanio rerio* is used for many developmental studies on the molecular level and the flower *Arabidopsis thaliana* is a model organism for flowering plants. The Japanese pufferfish *(Takifugu rubripes)* and the spotted green pufferfish *(Tetraodon nigroviridis)* are interesting because of their small and compact genomes, containing very little non-coding DNA compared to most other species. The mammals like dog *(Canis familiaris)*, brown rat *(Rattus norvegicus)*, mouse *(Mus musculus)*, and chimpanzee *(Pan troglodytes)* are all important model animals in medical research.

(ii) Genome-wide Association Studies

A genome-wide association study (GWAS) is an examination of genetic variation across a given genome, designed to identify genetic associations with observable traits. In human studies, this might include traits such as blood pressure or weight, or why some people get a disease or condition. The completion of the Human Genome Project in 2003 made it possible to find the genetic contributions to common diseases and analyze whole-genome samples for genetic variations that contribute to their onset[2].

These studies normally require two groups of participants: people with the disease (cases) and similar people without disease (controls). After genotyping each participant, the set of markers, such as SNPs (Single Nucleotide Polymorphism), are scanned into computers. Then bioinformatics is applied to survey participants' genomes for markers of genetic variation. If genetic variations are more frequent in people with the disease, the variations are said to be "associated" with the disease.

(iii) DNA Microarray

A DNA microarray is a multiplex technology used in molecular biology and in medicine[3]. It consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides, called features, each containing picomoles of a specific DNA sequence, known as *probes* (or *reporters*). This can be a short section of a gene or other DNA element that are used to hybridize a cDNA or cRNA sample (called *target*) under high-stringency conditions. Probe-target hybridization is usually detected and quantified by detection of fluorophore-, silver-, or chemiluminescence-labeled targets to determine relative abundance of nucleic acid sequences in the target. Since an array can contain tens of thousands of probes, a microarray experiment can accomplish that many genetic tests in parallel. Therefore, arrays have dramatically accelerated many types of investigation (fig. 2).

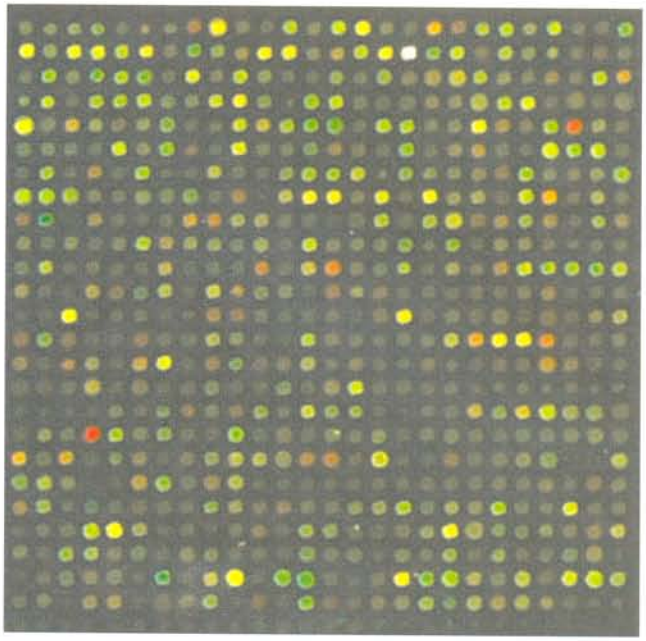

Fig. 2. DNA Microarray allows to study many genes in one experiment

3. Applications

The afore-mentioned genome-oriented basic research has been applied in all aspects of human health. Genomics plays a role in nine of the ten leading causes of death, most notably cancer and heart disease. These diseases are partly the result of how genes interact with environmental and behavioral risk factors, such as diet and physical activity. Also, a large fraction of children's hospitalizations are due to diseases that have genetic components. We shall outline three most important areas below:

(i) Determining genetic contributions to diseases

a. Genes and disease

Genes can have a powerful impact on our health, sometimes directly - through chromosome or single gene disorders - or by influencing our susceptibility to disease. The role of genes in inherited disorders is well understood. For some diseases, one particular gene has such a major effect that mutations in it are said to 'cause' the disease. In most cases, however, there is no major single determinant. Instead, variations in many different genes contribute to disease susceptibility (fig. 3).

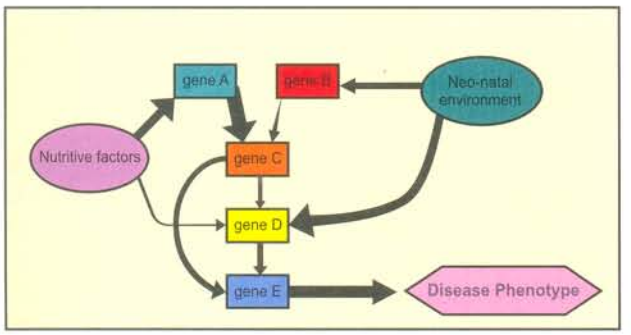

Fig. 3. Different genes can interact to cause disease

However, the extent to which genes contribute to disease varies and much remains to be learned. Advances in understanding the genetic mechanisms behind these diseases enable the development of early diagnostic tests, new treatments, or interventions to prevent disease onset or minimize disease severity. All diseases have a genetic component. Mutations may be inherited or developed in response to environmental stresses such as viruses or toxins. The ultimate goal is to use this information to treat, cure, or if possible, prevent the development of disease[4].

b. Major Types of Genetic Diseases

Many, if not most, diseases have their roots in genes. Genes, through the proteins they encode, determine how efficiently foods and chemicals are metabolized, how effectively toxins are detoxified, and how vigorously infections are targeted. Genetic diseases can be categorized into three major groups: single gene, chromosomal abnormalities, and multifactorial (or complex conditions).

Thousands of diseases are known to be caused by changes in the DNA sequence of single gene. A gene can be changed (mutated) in many ways resulting in an altered protein product that is unable to perform its function. The common gene mutation involves a change or "misspelling" in a single base in the DNA like in sickle cell anemia (fig 4). Other mutations include the loss (deletion) or gain (duplication or insertion) of a single or multiple bases. The altered protein product may still retain some function but at a reduced capacity. In other cases, the protein may be totally disabled by the mutation or gain an entirely new but damaging function. The outcome of a particular mutation depends not only on how it alters a protein's function but also on how vital that particular protein is to survival.

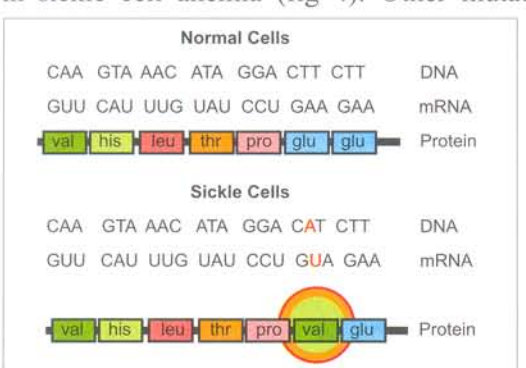

Fig. 4. A single amino acid in the beta chain is altered in sickle cell hemoglobin

In addition, genetic diseases can be caused by larger changes in chromosomes. Chromosomal abnormalities may be either numerical or structural. The most common type of chromosomal abnormality is known as aneuploidy, an abnormal number of chromosomes due to an extra or missing chromosome. A normal karyotype (complete chromosome set) contains 46 chromosomes including an XX (female) or XY (male) sex chromosome pair. Structural chromosomal abnormalities include deletions, duplications, insertions, inversions, or translocations of a chromosome segment (fig. 5).

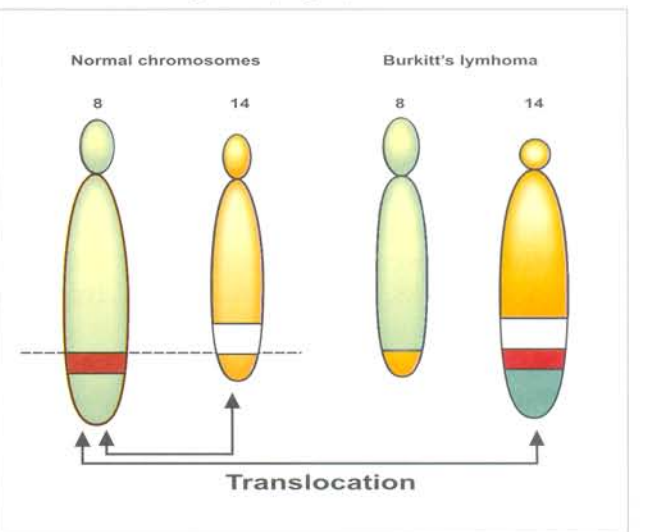

Fig. 5. Reciprocal chromosomal translocations in Burkitt's lymphoma, a tumor of B lymphocyte

Multifactorial diseases are caused by a combination of genetic, behavioral and environmental factors. The underlying etiology of multifactorial diseases is complex and heterogeneous. Examples of these conditions include spina bifida (split spine), diabetes, and heart disease. While multifactorial diseases can recur in families, some mutations can be acquired throughout an individual's lifetime such as in cancer. All genes work in the context of environment and behavior. Alterations in behavior or the environment, such as diet, exercise, exposure to toxic agents, or medications can all have influences on genetic traits.

(ii) Developing biomarkers for Tissue Classification by microarray

DNA microarray technology provides useful tools for profiling global gene expression patterns in different tissue samples say healthy and diseased persons. In such experiments, sufficient numbers of arrays are employed for both the conditions. After the experiment, data is organized properly. First of all, differentially expressed genes between two conditions are identified. One major challenge is the large number of genes relative to the number of samples (arrays). The use of all genes can suppress or reduce the performance of a classification method. Selection of an optimal subset from the differentially expressed genes becomes an important pre step in sample classification. This can be achieved by using a suitable feature selection method. Once significant genes highly correlated with tissue types are identified, all available classification methods may be used for classification of tissue samples. The selected genes for classification after validation may be used as biomarkers[5].

(iii) Use of microarray data for drug discovery

Currently, Systems biology approaches to disease are developed from the idea that disease-perturbed regulatory networks differ from their normal counterparts. Microarray data analyses reveal global changes in gene expression in response to genetic and environmental changes and accordingly, are well suited to construct the normal, disease-perturbed and drug-effected networks, which are useful for drug discovery[6].

4. Promise of Personalized Medical Treatment

Six years after scientists finished decoding the human genome - the genetic instruction book for life - they are starting to take their new knowledge from the research laboratory to the doctor's office and the patient's bedside. Researchers are seeking ways to tailor treatments to individuals - they call it "personalized medicine" - in order to improve patient outcomes and to lower costs in the overburdened health care system. The goal is to deliver the right drug at the right time in the right dose to the right person, and eliminate treatments that do not work. Researchers say that personalized medicine also can reduce unnecessary suffering and expense by minimizing the chance of adverse drug reactions.

Experts caution, however, that it is premature to say that an era of individually customized medicine has arrived. Major scientific and policy hurdles remain before patients can benefit widely from the promises of personalized medicine. Issues of insurance coverage, medical training, privacy and safety remain to be resolved[7].

Costs of decoding the genome have come down proportionately. In 2003, it cost an estimated $300 million to decode the first genome of an individual human. By 2007, the cost per person had come down to $100 million, and by 2008, it was $60,000. The current cost is about $20,000. It is predicted that it soon will be possible to sequence a person's genome in one day. For the first time, this will enable large numbers of patients to be sequenced to get to the bottom of thousands of genetically controlled diseases.

5. Challenges in Translation of Basic Scientific Discoveries into Healthcare

* Physicians feel unprepared to integrate genomics into regular practice. Education of health professionals must be a priority to advance the use of genomics into healthcare. With the rapid advances in genomics research and developing technologies, it will be challenging to keep health professionals informed about the benefits, risks, and limitations of new tools as they become available. Basic genomic literacy is a critical need for patients, physicians, and communities to engage in genomic research, and clinical studies are required to bring about a change in the care paradigms to support clinical genomics applications.

* Consumers are worried about the possible adverse consequences of genetic testing, particularly the privacy issues and discrimination against receiving employment and health insurance. In order for genomic medicine to be integrated into routine clinical practice, associated fears with this type of testing must be put to rest.

* Public-private partnerships will likely be required to generate the evidence base for genomic medicine. These collaborations are desirable because firstly, no single stakeholder group is likely to have sufficient resources or expertise to conduct the necessary studies, and secondly, both will likely benefit from their execution.

* As with any new innovation, genomic testing must be demonstrated to be clinically useful, cost-effective, and of value. Clarity is needed on the drivers of cost effectiveness of genomic technologies.

* The gaps to be filled to bring genomic medicine to fruition are exceedingly complex. There is a need for developing a clear understanding of pathways for translation, the barriers that lie in the translational path, and the strategies to overcome them.

6. A Brief Account of Indian Achievements in Genomics:

The devastating growing population due to high birth rate and blood related marriages favoured in many Indian communities cause prevalence of genetic disorders in India. The demographic factor provided a sound ground to carry out research in the field of Molecular biology and genetics in India. The 15th International Congress of Genetics was held in New Deihi in 1983 under the Presidentship of Dr. M. S. Swaminathan. This event sensitized the science policy makers in India to make investments in genetic research. Accordingly, India has taken important steps in creating the basic infrastructure of molecular biology research in the country, with the Department of Biotechnology (DBT) playing a major role in this endeavour. The DBT has established nine research institutes at various places of the country. These institutes are well equipped with world-class instrumentation and highly competent human resources. Some other organizations such as CSIR, ICAR, ICMR, MOEF, UGC/academic institutes, DAE and some NGOs have also joined hands with the DBT to endorse the research in the field of Molecular biology and genetics.

As a follow up to the 1983 Congress, Genetics Congress Trust was set up, with the main objective of promoting molecular genetics research in India. A symposium was organized by the Genetics Congress Trust in New Delhi during 21-22 January 2004 to discuss how India has fared during the Post-Genetics Congress period[8].

Of late the contributions of Indian scientists in this rapidly developing discipline have been recognized. Indian scientists from Delhi University, CSIR, ICAR etc have added many feathers to India's efforts in this area. Below we summarize the major achievements:

1. Department of Biotechnology (DBT) and Indian Council of Agricultural Research (ICAR) jointly laid down the foundation of the Indian Initiative for Rice Genome Sequencing (IIRGS). The map-based sequence of the rice genome was completed in December 2004 and published in Nature on August 11, 2005[9].

2. Two years ago, a team of Indian scientists from Hyderabad and New Delhi had sequenced the full genome of a harmless bacterium that has been named *Mycobacterium indicus pranii*.

3. Early last year, CSIR had successfully completed the whole genome sequencing of a wild type strain of Zebra fish *(Danio rerio)*. This work marked India's entry into the arena of whole genome sequencing of animals. Zebra fish is popularly used by the scientific community as an organism for modeling human diseases.

4. A recent study published on 24th September 2009, led by scientist from Harvard and the Centre for Cellular and Molecular Biology (a constituent lab under CSIR) in Hyderabad, have determined that a large percentage of Indian population has originated from two sets of relatively few individuals. The primary ancestral groups have been identified as "Ancestral North Indians" (ANI) and "Ancestral South Indians" (ASI). This finding may have implication in health care of Indian population[10].

5. In December 2009, Scientists of Institute of Genomics and Integrative Biology (IGIB), New Delhi have succeeded in completely sequencing the genome of an individual, enabling India to join a league of selected countries: the U.S., the U.K., Canada, China and Korea[11].

Conclusion

The sequencing of the human genome and many microbial genomes has provided new opportunities to study the genetic impact on life processes, leading to development of new technologies that can be translated to clinical practices. Developments of such new technologies take lot of time and resources because of their multidisciplinary nature. We have to take genomic research to its most important end point of improving human health. In India, Science policy makers are giving due importance to Genomics and the field is flourishing satisfactorily[12]. This will change the whole scenario of medical treatment as Caleb Parry, a visionary British physician of the 18th century said, *"It is much more important to know what kind of patient has a disease than to know what kind of disease a patient has."*

Acknowledgements:

We thankfully acknowledge various resources on the internet which we have used in preparing this article. The views and opinions expressed in this article

are those of the authors and do not reflect the views of the organizations where they serve.

References

1. Fleischmann R., Adams M., White O., Clayton R., Kirkness E., Kerlavage A., Bult C., Tomb J., Dougherty B. and Merrick J. Whole genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269: 496-512, (1995).

2. Chakravarti, A. and Little, P. Nature, nurture and human disease. In *50 years of DNA*, eds Clayton, J. and Dennis, C. Nature Publishing Group: 98-100, (2003).

3. Schena M., Shalon D., Davis R. W. and Brown P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270: 368-371, (1995).

4. Rath S. K. Genetic basis of diseases - the bright future. *Drugs and pharmaceuticals current R&D Highlights* 32 (3): 1, (2009).

5. Tao Li, Zhang C. and Ogihara M. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics* 20 (15): 2429-2437, (2004).

6. Chen B. and Cheng W. L. Analyzing microarray data in drug discovery using systems biology. *Expert Opinion on Drug Discovery* 2 (5): 755-768, (2007).

7. www.personlizedmedicine.com

8. Jain H. K. and Sharma R. P. Advancing frontiers of molecular genetics: Where does India stand. *Current Science* 89 (8): 1314-1316, (2005).

9. International Rice Genome Sequencing Project The map-based sequence of the rice genome. *Nature* 436: 793-800, (2005).

10. Reich D., Kumarasamy T., Patterson N., Price A. L. and Singh Lalji, Reconstructing Indian population history. *Nature* 461: 489-494, (2009).

11. CSIR sequences human genome. The Hindu, New Delhi, December 9, 2009.

12. Acharya T., Kumar N. K., Muthuswamy V., Daar S. A. and Singer P. A. Harnessing genomics to improve health in India- an executive course to support genomics policy. *Health Research Policy and Systems* 2:1, (2004).

Moinuddin Ansari, Curator, District Science Centre, NCSM, Purulia.

Dr. Rehana Abidi, Principal Scientist, National Bureau of Fish Genetic Resources, ICAR, Lucknow.