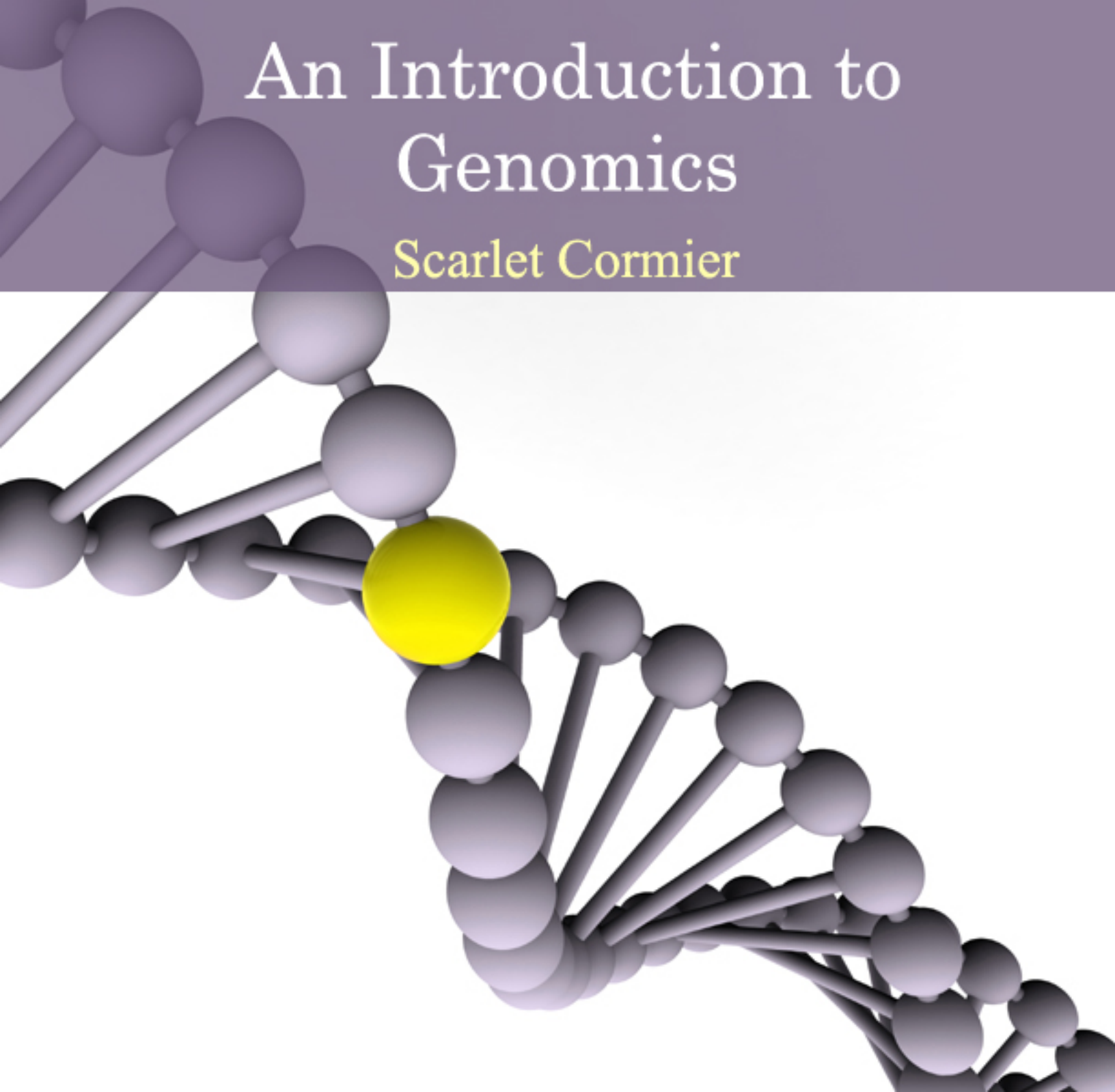
An Introduction to Genomics

Scarlet Cormier



First Edition, 2012

ISBN 978-81-323-3235-0

Published by:
Research World
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter- 1

Genomics

Genomics is a discipline in genetics concerning the study of the genomes of organisms. The field includes intensive efforts to determine the entire DNA sequence of organisms and fine-scale genetic mapping efforts. The field also includes studies of intragenomic phenomena such as heterosis, epistasis, pleiotropy and other interactions between loci and alleles within the genome. In contrast, the investigation of the roles and functions of single genes is a primary focus of molecular biology or genetics and is a common topic of modern medical and biological research. Research of single genes does not fall into the definition of genomics unless the aim of this genetic, pathway, and functional information analysis is to elucidate its effect on, place in, and response to the entire genome's networks.

For the United States Environmental Protection Agency, "the term "genomics" encompasses a broader scope of scientific inquiry associated technologies than when genomics was initially considered. A genome is the sum total of all an individual organism's genes. Thus, genomics is the study of all the genes of a cell, or tissue, at the DNA (genotype), mRNA (transcriptome), or protein (proteome) levels."

History

The first genomes to be sequenced were those of a virus and a mitochondrion, and were done by Fred Sanger. His group established techniques of sequencing, genome mapping, data storage, and bioinformatic analyses in the 1970-1980s. A major branch of genomics is still concerned with sequencing the genomes of various organisms, but the knowledge of full genomes has created the possibility for the field of functional genomics, mainly concerned with patterns of gene expression during various conditions. The most important tools here are microarrays and bioinformatics. Study of the full set of proteins in a cell type or tissue, and the changes during various conditions, is called proteomics. A related concept is materiomics, which is defined as the study of the material properties of biological materials (e.g. hierarchical protein structures and materials, mineralized biological tissues, etc.) and their effect on the macroscopic function and failure in their biological context, linking processes, structure and properties at multiple scales through a materials science approach. The actual term 'genomics' is thought to have been coined by

Dr. Tom Roderick, a geneticist at the Jackson Laboratory (Bar Harbor, ME) over beer at a meeting held in Maryland on the mapping of the human genome in 1986.

In 1972, Walter Fiers and his team at the Laboratory of Molecular Biology of the University of Ghent (Ghent, Belgium) were the first to determine the sequence of a gene: the gene for Bacteriophage MS2 coat protein. In 1976, the team determined the complete nucleotide-sequence of bacteriophage MS2-RNA. The first DNA-based genome to be sequenced in its entirety was that of bacteriophage Φ-X174; (5,368 bp), sequenced by Frederick Sanger in 1977.

The first free-living organism to be sequenced was that of *Haemophilus influenzae* (1.8 Mb) in 1995, and since then genomes are being sequenced at a rapid pace.

As of September 2007, the complete sequence was known of about 1879 viruses, 577 bacterial species and roughly 23 eukaryote organisms, of which about half are fungi. 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 *Drosophila 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 zebrafish *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 species. The mammals dog (*Canis familiaris*), brown rat (*Rattus norvegicus*), mouse (*Mus musculus*), and chimpanzee (*Pan troglodytes*) are all important model animals in medical research.

Human genomics

A rough draft of the human genome was completed by the Human Genome Project in early 2001, creating much fanfare. By 2007 the human sequence was declared "finished" (less than one error in 20,000 bases and all chromosomes assembled). Display of the results of the project required significant bioinformatics resources. The sequence of the human reference assembly can be explored using the UCSC Genome Browser.

Bacteriophage genomics

Bacteriophages have played and continue to play a key role in bacterial genetics and molecular biology. Historically, they were used to define gene structure and gene regulation. Also the first genome to be sequenced was a bacteriophage. However, bacteriophage research did not lead the genomics revolution, which is clearly dominated by bacterial genomics. Only very recently has the study of bacteriophage genomes become prominent, thereby enabling researchers to understand the mechanisms

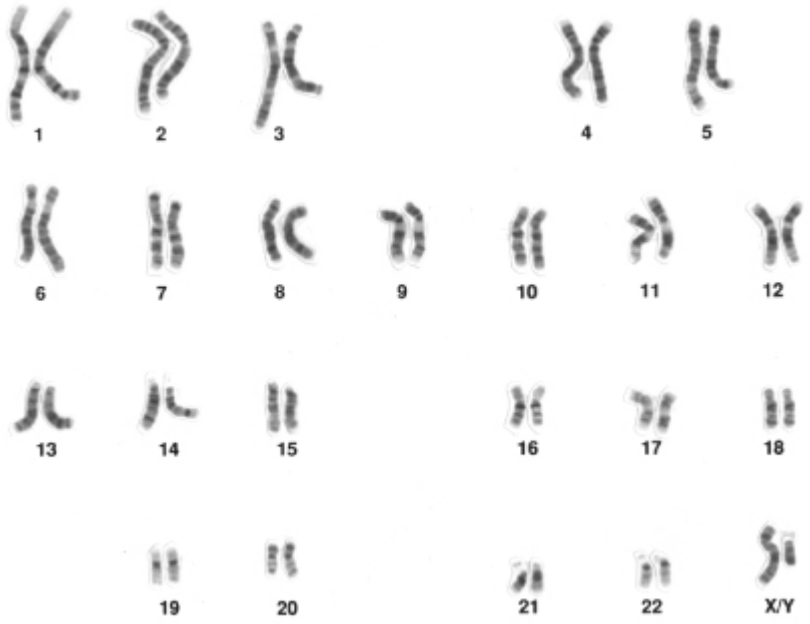
underlying phage evolution. Bacteriophage genome sequences can be obtained through direct sequencing of isolated bacteriophages, but can also be derived as part of microbial genomes. Analysis of bacterial genomes has shown that a substantial amount of microbial DNA consists of prophage sequences and prophage-like elements. A detailed database mining of these sequences offers insights into the role of prophages in shaping the bacterial genome.

Cyanobacteria genomics

At present there are 24 cyanobacteria for which a total genome sequence is available. 15 of these cyanobacteria come from the marine environment. These are six *Prochlorococcus* strains, seven marine *Synechococcus* strains, *Trichodesmium erythraeum* IMS101 and *Crocosphaera watsonii* WH8501. Several studies have demonstrated how these sequences could be used very successfully to infer important ecological and physiological characteristics of marine cyanobacteria. However, there are many more genome projects currently in progress, amongst those there are further *Prochlorococcus* and marine *Synechococcus* isolates, *Acaryochloris* and *Prochloron*, the N_2-fixing filamentous cyanobacteria *Nodularia spumigena*, *Lyngbya aestuarii* and *Lyngbya majuscula*, as well as bacteriophages infecting marine cyanobaceria. Thus, the growing body of genome information can also be tapped in a more general way to address global problems by applying a comparative approach. Some new and exciting examples of progress in this field are the identification of genes for regulatory RNAs, insights into the evolutionary origin of photosynthesis, or estimation of the contribution of horizontal gene transfer to the genomes that have been analyzed.

Chapter- 2

Genome



An image of the 46 chromosomes, making up the diploid genome of human male. (The mitochondrial chromosome is not shown.)

In modern molecular biology and genetics, the **genome** is the entirety of an organism's hereditary information. It is encoded either in DNA or, for many types of virus, in RNA. The genome includes both the genes and the non-coding sequences of the DNA.

Origin of Term

The term was adapted in 1920 by Hans Winkler, Professor of Botany at the University of Hamburg, Germany. In Greek, the word *genome* (γίνομαι) means "I become, I am born, to come into being". The Oxford English Dictionary suggests the name to be a blend of the words ***gene*** and *chromos**ome***. A few related *-ome* words already existed, such as *biome* and *rhizome*, forming a vocabulary into which *genome* fits systematically.

Overview

Some organisms have multiple copies of chromosomes, diploid, triploid, tetraploid and so on. In classical genetics, in a sexually reproducing organism (typically eukarya) the gamete has half of the number of chromosome of the somatic cell and the genome is a full set of chromosomes in a gamete. In haploid organisms, including cells of bacteria, archaea, and in organelles including mitochondria and chloroplasts, or viruses, that similarly contain genes, the single or set of circular and/or linear chains of DNA (or RNA for some viruses), likewise constitute the *genome*. The term genome can be applied specifically to mean that stored on a complete set of *nuclear DNA* (i.e., the "nuclear genome") but can also be applied to that stored within organelles that contain their own DNA, as with the "mitochondrial genome" or the "chloroplast genome". Additionally, the genome can comprise nonchromosomal genetic elements such as viruses, plasmids, and transposable elements.

When people say that the genome of a sexually reproducing species has been "sequenced", typically they are referring to a determination of the sequences of one set of autosomes and one of each type of sex chromosome, which together represent both of the possible sexes. Even in species that exist in only one sex, what is described as "a genome sequence" may be a composite read from the chromosomes of various individuals. In general use, the phrase "genetic makeup" is sometimes used conversationally to mean the genome of a particular individual or organism. The study of the global properties of genomes of related organisms is usually referred to as genomics, which distinguishes it from genetics which generally studies the properties of single genes or groups of genes.

Both the number of base pairs and the number of genes vary widely from one species to another, and there is only a rough correlation between the two (an observation known as the C-value paradox). At present, the highest known number of genes is around 60,000, for the protozoan causing trichomoniasis, almost three times as many as in the human genome.

An analogy to the human genome stored on DNA is that of instructions stored in a book:

- The book (genome) would contain 23 chapters (chromosomes);
- each chapter contains 48 to 250 million letters (A,C,G,T) without spaces;
- Hence, the book contains over 3.2 billion letters total;
- The book fits into a cell nucleus the size of a pinpoint;
- At least one copy of the book (all 23 chapters) is contained in every cell of our body.

Types

Most biological entities that are more complex than a virus sometimes or always carry additional genetic material besides that which resides in their chromosomes. In some contexts, such as sequencing the genome of a pathogenic microbe, "genome" is meant to include information stored on this auxiliary material, which is carried in plasmids. In

such circumstances then, "genome" describes all of the genes and information on non-coding DNA that have the potential to be present.

In eukaryotes such as plants, protozoa and animals, however, "genome" carries the typical connotation of only information on chromosomal DNA. So although these organisms contain chloroplasts and/or mitochondria that have their own DNA, the genetic information contained by DNA within these organelles is not considered part of the genome. In fact, mitochondria are sometimes said to have their own genome often referred to as the "mitochondrial genome". The DNA found within the chloroplast may be referred to as the "plastome".

Genomes and genetic variation

Note that a genome does not capture the genetic diversity or the genetic polymorphism of a species. For example, the human genome sequence in principle could be determined from just half the information on the DNA of one cell from one individual. To learn what variations in genetic information underlie particular traits or diseases requires comparisons across individuals. This point explains the common usage of "genome" (which parallels a common usage of "gene") to refer not to the information in any particular DNA sequence, but to a whole family of sequences that share a biological context.

Although this concept may seem counter intuitive, it is the same concept that says there is no particular shape that is the shape of a cheetah. Cheetahs vary, and so do the sequences of their genomes. Yet both the individual animals and their sequences share commonalities, so one can learn something about cheetahs and "cheetah-ness" from a single example of either.

Sequencing and mapping

The Human Genome Project was organized to map and to sequence the human genome. Other genome projects include mouse, rice, the plant *Arabidopsis thaliana*, the puffer fish, bacteria like E. coli, etc. In 1976, Walter Fiers at the University of Ghent (Belgium) was the first to establish the complete nucleotide sequence of a viral RNA-genome (bacteriophage MS2). The first DNA-genome project to be completed was the Phage Φ-X174, with only 5386 base pairs, which was sequenced by Fred Sanger in 1977. The first bacterial genome to be completed was that of Haemophilus influenzae, completed by a team at The Institute for Genomic Research in 1995.

The development of new technologies has dramatically decreased the difficulty and cost of sequencing, and the number of complete genome sequences is rising rapidly. Among many genome database sites, the one maintained by the US National Institutes of Health is inclusive.

These new technologies open up the prospect of personal genome sequencing as an important diagnostic tool. A major step toward that goal was the completion of the decipherment of the full genome of DNA pioneer James D. Watson in 2007.

Whereas a genome sequence lists the order of every DNA base in a genome, a genome map identifies the landmarks. A genome map is less detailed than a genome sequence and aids in navigating around the genome.

Comparison of different genome sizes

Organism type	Organism	Genome size (base pairs)	mass - in pg	Note
Virus	Bacteriophage MS2	3,569	0.000002	First sequenced RNA-genome
Virus	SV40	5,224		
Virus	Phage Φ-X174	5,386		First sequenced DNA-genome
Virus	HIV	9749		
Virus	Phage λ	48,502		
Virus	Mimivirus	1,181,404		Largest known viral genome
Bacterium	*Haemophilus influenzae*	1,830,000		First genome of a living organism sequenced, July 1995
Bacterium	*Carsonella ruddii*	159,662		Smallest non-viral genome.
Bacterium	*Buchnera aphidicola*	600,000		
Bacterium	*Wigglesworthia glossinidia*	700,000		
Bacterium	*Escherichia coli*	4,600,000		
Bacterium	*Solibacter usitatus* (strain Ellin 6076)	9,970,000		Largest known Bacterial genome
Amoeboid	*Polychaos dubium* (*"Amoeba" dubia*)	670,000,000,000	737	Largest known genome.
Plant	*Arabidopsis thaliana*	157,000,000		First plant genome sequenced, December 2000.
Plant	*Genlisea margaretae*	63,400,000		Smallest recorded flowering plant genome, 2006.
Plant	*Fritillaria assyrica*	130,000,000,000		
Plant	*Populus trichocarpa*	480,000,000		First tree genome sequenced, September

				2006
Plant	*Paris japonica* (Japanese-native, pale-petal)	150,000,000,000	152.23 pg	Largest plant genome known
Moss	*Physcomitrella patens*	480,000,000		First genome of a bryophyte sequenced, January 2008.
Yeast	*Saccharomyces cerevisiae*	12,100,000		
Fungus	*Aspergillus nidulans*	30,000,000		
Nematode	*Caenorhabditis elegans*	100,300,000		First multicellular animal genome sequenced, December 1998
Nematode	*Pratylenchus coffeae*	20,000,000		Smallest animal genome known
Insect	*Drosophila melanogaster* (fruit fly)	130,000,000		
Insect	*Bombyx mori* (silk moth)	530,000,000		
Insect	*Apis mellifera* (honey bee)	236,000,000		
Insect	*Solenopsis invicta* (fire ant)	480,000,000		
Fish	*Tetraodon nigroviridis* (type of puffer fish)	385,000,000		Smallest vertebrate genome known
Mammal	*Homo sapiens*	3,200,000,000	3	
Fish	*Protopterus aethiopicus* (marbled lungfish)	130,000,000,000	143	Largest vertebrate genome known

Note: The DNA from a single (diploid) human cell if the 46 chromosomes were connected end-to-end and straightened, would have a length of ~2 m and a width of ~2.4 nanometers.

Since genomes and their organisms are very complex, one research strategy is to reduce the number of genes in a genome to the bare minimum and still have the organism in question survive. There is experimental work being done on minimal genomes for single cell organisms as well as minimal genomes for multicellular organisms. The work is both *in vivo* and *in silico*.

Genome evolution

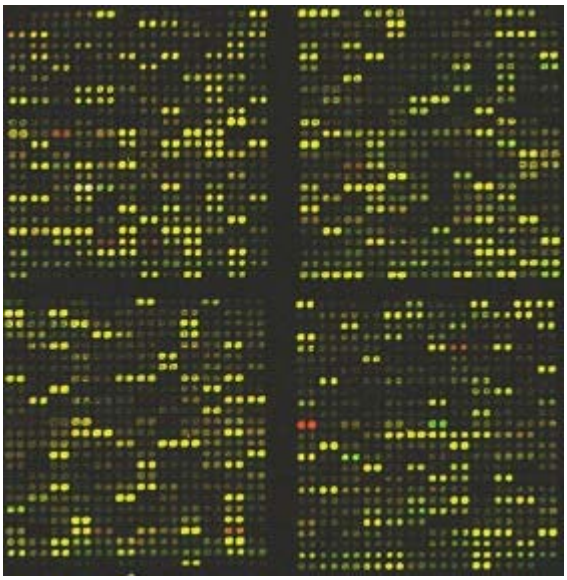
Genomes are more than the sum of an organism's genes and have traits that may be measured and studied without reference to the details of any particular genes and their products. Researchers compare traits such as *chromosome number* (karyotype), genome size, gene order, codon usage bias, and GC-content to determine what mechanisms could have produced the great variety of genomes that exist today.

Duplications play a major role in shaping the genome. Duplications may range from extension of short tandem repeats, to duplication of a cluster of genes, and all the way to duplications of entire chromosomes or even entire genomes. Such duplications are probably fundamental to the creation of genetic novelty.

Horizontal gene transfer is invoked to explain how there is often extreme similarity between small portions of the genomes of two organisms that are otherwise very distantly related. Horizontal gene transfer seems to be common among many microbes. Also, eukaryotic cells seem to have experienced a transfer of some genetic material from their chloroplast and mitochondrial genomes to their nuclear chromosomes.

Chapter- 3

Functional Genomics



A DNA microarray

Functional genomics is a field of molecular biology that attempts to make use of the vast wealth of data produced by genomic projects (such as genome sequencing projects) to describe gene (and protein) functions and interactions. Unlike genomics and proteomics, functional genomics focuses on the dynamic aspects such as gene transcription, translation, and protein-protein interactions, as opposed to the static aspects of the genomic information such as DNA sequence or structures. Functional genomics attempts to answer questions about the function of DNA at the levels of genes, RNA transcripts, and protein products. A key characteristic of functional genomics studies is their genome-wide approach to these questions, generally involving high-throughput methods rather than a more traditional "gene-by-gene" approach.

Goals of functional genomics

The goal of functional genomics is to understand the relationship between an organism's genome and its phenotype. The term functional genomics is often used broadly to refer to the many possible approaches to understanding the properties and function of the entirety of an organism's genes and gene products. This definition is somewhat variable; Gibson and Muse define it as "approaches under development to ascertain the biochemical, cellular, and/or physiological properties of each and every gene product", while Pevsner includes the study of nongenic elements in his definition: "the genome-wide study of the function of DNA (including genes and nongenic elements), as well as the nucleic acid and protein products encoded by DNA". Because of its genome-wide approach, functional genomics requires the use of high-throughput technologies capable of assaying many functions or relationships simultaneously. Functional genomics involves studies of natural variation in genes, RNA, and proteins over time (such as an organism's development) or space (such as its body regions), as well as studies of natural or experimental functional disruptions affecting genes, chromosomes, RNAs, or proteins.

The promise of functional genomics is to expand and synthesize genomic and proteomic knowledge into an understanding of the dynamic properties of an organism at cellular and/or organismal levels. This would provide a more complete picture of how biological function arises from the information encoded in an organism's genome. The possibility of understanding how a particular mutation leads to a given phenotype has important implications for human genetic diseases, as answering these questions could point scientists in the direction of a treatment or cure.

Techniques and applications

Functional genomics includes function-related aspects of the genome itself such as mutation and polymorphism (such as SNP) analysis, as well as measurement of molecular activities. The latter comprise a number of "-omics" such as transcriptomics (gene expression), proteomics (protein expression), phosphoproteomics (a subset of proteomics) and metabolomics. Functional genomics uses mostly multiplex techniques to measure the abundance of many or all gene products such as mRNAs or proteins within a biological sample. Together these measurement modalities endeavor to quantitate the various biological processes and improve our understanding of gene and protein functions and interactions.

At the DNA level

Genetic interaction mapping

Systematic pairwise deletion of genes or inhibition of gene expression can be used to identify genes with related function, even if they do not interact physically. Epistasis refers to the fact that effects for two different gene knockouts may not be additive; that is, the phenotype that results when two genes are inhibited may be different from the sum of the effects of single knockouts.

The ENCODE project

The ENCODE (Encyclopedia of DNA elements) project is an in-depth analysis of the human genome whose goal is to identify all the functional elements of genomic DNA, in both coding and noncoding regions. To this point, only the pilot phase of the study has been completed, involving hundreds of assays performed on 44 regions of known or unknown function comprising 1% of the human genome. Important results include evidence from genomic tiling arrays that most nucleotides are transcribed as coding transcripts, noncoding RNAs, or random transcripts, the discovery of additional transcriptional regulatory sites, further elucidation of chromatin-modifying mechanisms.

At the RNA level: transcriptome profiling

Microarrays

Microarrays measure the amount of mRNA in a sample that corresponds to a given gene or probe DNA sequence. Probe sequences are immobilized on a solid surface and allowed to hybridize with fluorescently-labeled “target” mRNA. The intensity of fluorescence of a spot is proportional to the amount of target sequence that has hybridized to that spot, and therefore to the abundance of that mRNA sequence in the sample. Microarrays allow for identification of candidate genes involved in a given process based on variation between transcript levels for different conditions and shared expression patterns with genes of known function.

SAGE

SAGE (Serial analysis of gene expression) is an alternate method of gene expression analysis based on RNA sequencing rather than hybridization. SAGE relies on the sequencing of 10-17 base pair tags which are unique to each gene. These tags are produced from poly-A mRNA and ligated end-to-end before sequencing. SAGE gives an unbiased measurement of the number of transcripts per cell, since it does not depend on prior knowledge of what transcripts to study (as microarrays do).

At the protein level: protein-protein interactions

Yeast two-hybrid system

A yeast two-hybrid (Y2H) screen tests a "bait" protein against many potential interacting proteins ("prey") to identify physical protein-protein interactions. This system is based on a transcription factor, originally GAL4, whose separate DNA-binding and transcription activation domains are both required in order for the protein to cause transcription of a reporter gene. In a Y2H screen, the "bait" protein is fused to the binding domain of GAL4, and a library of potential "prey" (interacting) proteins is recombinantly expressed in a vector with the activation domain. In vivo interaction of bait and prey proteins in a yeast cell brings the activation and binding domains of GAL4 close enough together to

result in expression of a reporter gene. It is also possible to systematically test a library of bait proteins against a library of prey proteins to identify all possible interactions in a cell.

AP/MS

Affinity purification and mass spectrometry (AP/MS) is able to identify proteins that interact with one another in complexes. Complexes of proteins are allowed to form around a particular "bait" protein. The bait protein is identified using an antibody or a recombinant tag which allows it to be extracted along with any proteins that have formed a complex with it. The proteins are then digested into short peptide fragments and mass spectrometry is used to identify the proteins based on the mass-to-charge ratios of those fragments.

Loss-of-function techniques

Mutagenesis

Gene function can be investigated by systematically "knocking out" genes one by one. This is done by either deletion or disruption of function (such as by insertional mutagenesis) and the resulting organisms are screened for phenotypes that provide clues to the function of the disrupted gene.

RNAi

RNA interference (RNAi) methods can be used to transiently silence or knock down gene expression using ~20 base-pair double-stranded RNA typically delivered by transfection of synthetic ~20-mer short-interfering RNA molecules (siRNAs) or by virally-encoded short-hairpin RNAs (shRNAs). RNAi screens, typically performed in cell culture-based assays or experimental organisms (such as C. elegans) can be used to systematically disrupt nearly every gene in a genome or subsets of genes (sub-genomes); possible functions of disrupted genes can be assigned based on observed phenotypes.

Functional annotations for genes

Genome annotation

Putative genes can be identified by scanning a genome for regions likely to encode proteins, based on characteristics such as long open reading frames, transcriptional initiation sequences, and polyadenylation sites. A sequence identified as a putative gene must be confirmed by further evidence, such as similarity to cDNA or EST sequences from the same organism, similarity of the predicted protein sequence to known proteins, association with promoter sequences, or evidence that mutating the sequence produces an observable phenotype.

Rosetta stone approach

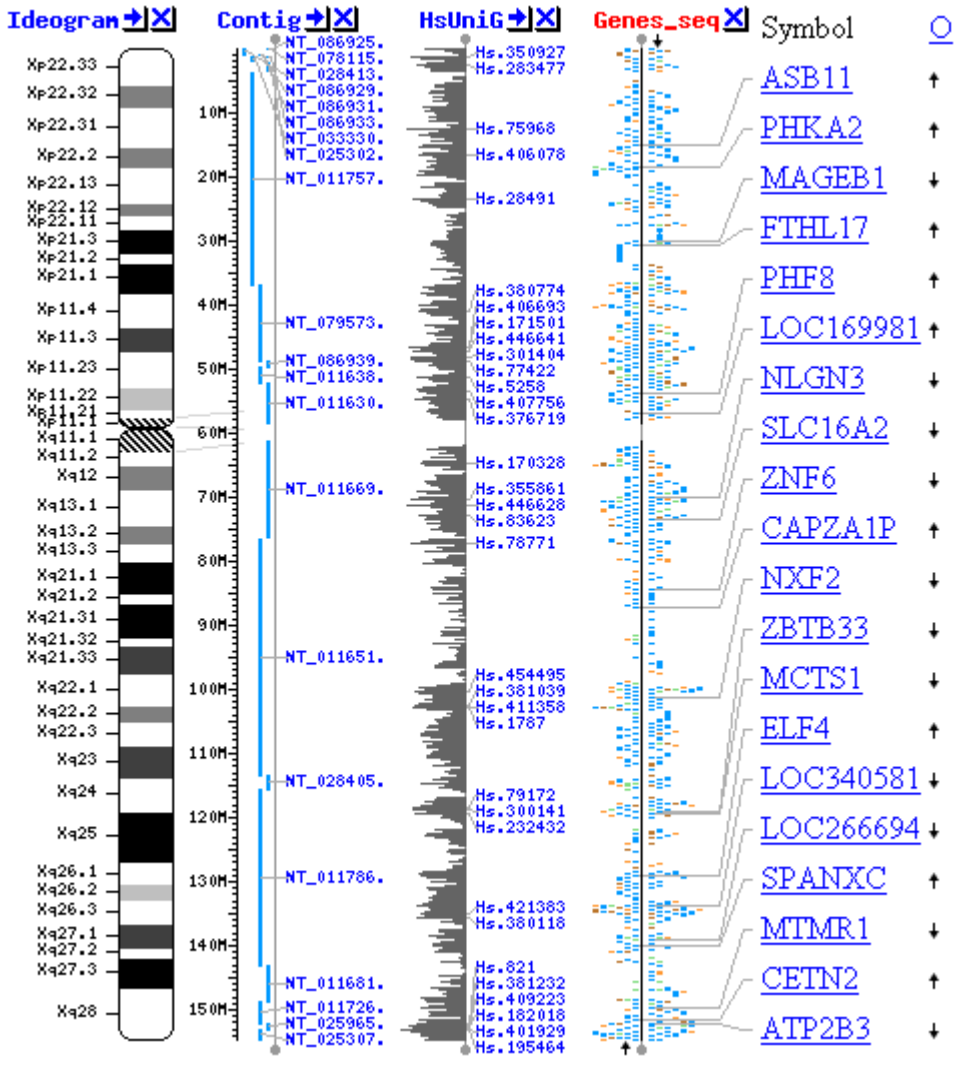
The Rosetta stone approach is a computation method of de novo protein function prediction, based on the hypothesis that some proteins involved in a given physiological process may exist as two separate genes in one organism and as a single gene in another. Genomes are scanned for sequences that are independent in one organism and in a single open reading frame in another. If two genes have fused, it is predicted that they have similar biological functions that make such coregulation advantageous.

Functional genomics and bioinformatics

Because of the large quantity of data produced by these techniques and the desire to find biologically meaningful patterns, bioinformatics is crucial to analysis of functional genomics data. Examples of techniques in this class are data clustering or principal component analysis for unsupervised machine learning (class detection) as well as artificial neural networks or support vector machines for supervised machine learning (class prediction, classification).

Chapter- 4

Bioinformatics



Map of the human X chromosome (from the NCBI website). Assembly of the human genome is one of the greatest achievements of bioinformatics.

Bioinformatics is the application of statistics and computer science to the field of molecular biology.

The term *bioinformatics* was coined by Paulien Hogeweg and Ben Hesper in 1978 for the study of informatic processes in biotic systems. Its primary use since at least the late 1980s has been in genomics and genetics, particularly in those areas of genomics involving large-scale DNA sequencing.

Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques and theory to solve formal and practical problems arising from the management and analysis of biological data.

Over the past few decades rapid developments in genomic and other molecular research technologies and developments in information technologies have combined to produce a tremendous amount of information related to molecular biology. It is the name given to these mathematical and computing approaches used to glean understanding of biological processes.

Common activities in bioinformatics include mapping and analyzing DNA and protein sequences, aligning different DNA and protein sequences to compare them and creating and viewing 3-D models of protein structures.

The primary goal of bioinformatics is to increase the understanding of biological processes. What sets it apart from other approaches, however, is its focus on developing and applying computationally intensive techniques (e.g., pattern recognition, data mining, machine learning algorithms, and visualization) to achieve this goal. Major research efforts in the field include sequence alignment, gene finding, genome assembly, drug design, drug discovery, protein structure alignment, protein structure prediction, prediction of gene expression and protein-protein interactions, genome-wide association studies and the modeling of evolution.

Introduction

Bioinformatics was applied in the creation and maintenance of a database to store biological information at the beginning of the "genomic revolution", such as nucleotide and amino acid sequences. Development of this type of database involved not only design issues but the development of complex interfaces whereby researchers could both access existing data as well as submit new or revised data.

In order to study how normal cellular activities are altered in different disease states, the biological data must be combined to form a comprehensive picture of these activities. Therefore, the field of bioinformatics has evolved such that the most pressing task now involves the analysis and interpretation of various types of data, including nucleotide and amino acid sequences, protein domains, and protein structures. The actual process of analyzing and interpreting data is referred to as computational biology. Important sub-disciplines within bioinformatics and computational biology include:

- the development and implementation of tools that enable efficient access to, and use and management of, various types of information.
- the development of new algorithms (mathematical formulas) and statistics with which to assess relationships among members of large data sets, such as methods to locate a gene within a sequence, predict protein structure and/or function, and cluster protein sequences into families of related sequences.

There are two fundamental ways of modelling a Biological system (e.g. living cell) both coming under Bioinformatic approaches.

- Static
  - Sequences - Proteins, Nucleic acids and Peptides
  - Structures - Proteins, Nucleic acids, Ligands (including metabolites and drugs) and Peptides
  - Interaction data among the above entities including microarray data and Networks of proteins, metabolites
- Dynamic
  - Systems Biology comes under this category including reaction fluxes and variable concentrations of metabolites
  - Multi-Agent Based modelling approaches capturing cellular events such as signalling, transcription and reaction dynamics

A broad sub-category under bioinformatics is structural bioinformatics.

Major research areas

Sequence analysis

Since the Phage Φ-X174 was sequenced in 1977, the DNA sequences of thousands of organisms have been decoded and stored in databases. This sequence information is analyzed to determine genes that encode polypeptides (proteins), RNA genes, regulatory sequences, structural motifs, and repetitive sequences. A comparison of genes within a species or between different species can show similarities between protein functions, or relations between species (the use of molecular systematics to construct phylogenetic trees). With the growing amount of data, it long ago became impractical to analyze DNA sequences manually. Today, computer programs such as BLAST are used daily to search sequences from more than 260 000 organisms, containing over 190 billion nucleotides. These programs can compensate for mutations (exchanged, deleted or inserted bases) in the DNA sequence, in order to identify sequences that are related, but not identical. A variant of this sequence alignment is used in the sequencing process itself. The so-called shotgun sequencing technique (which was used, for example, by The Institute for Genomic Research to sequence the first bacterial genome, *Haemophilus influenzae*) does not produce entire chromosomes, but instead generates the sequences of many thousands of small DNA fragments (ranging from 35 to 900 nucleotides long, depending on the sequencing technology). The ends of these fragments overlap and, when aligned properly by a genome assembly program, can be used to reconstruct the complete genome.

Shotgun sequencing yields sequence data quickly, but the task of assembling the fragments can be quite complicated for larger genomes. For a genome as large as the human genome, it may take many days of CPU time on large-memory, multiprocessor computers to assemble the fragments, and the resulting assembly will usually contain numerous gaps that have to be filled in later. Shotgun sequencing is the method of choice for virtually all genomes sequenced today, and genome assembly algorithms are a critical area of bioinformatics research.

Another aspect of bioinformatics in sequence analysis is annotation, which involves computational gene finding to search for protein-coding genes, RNA genes, and other functional sequences within a genome. Not all of the nucleotides within a genome are part of genes. Within the genome of higher organisms, large parts of the DNA do not serve any obvious purpose. This so-called junk DNA may, however, contain unrecognized functional elements. Bioinformatics helps to bridge the gap between genome and proteome projects — for example, in the use of DNA sequences for protein identification.

Genome annotation

In the context of genomics, **annotation** is the process of marking the genes and other biological features in a DNA sequence. The first genome annotation software system was designed in 1995 by Dr. Owen White, who was part of the team at The Institute for Genomic Research that sequenced and analyzed the first genome of a free-living organism to be decoded, the bacterium *Haemophilus influenzae*. Dr. White built a software system to find the genes (places in the DNA sequence that encode a protein), the transfer RNA, and other features, and to make initial assignments of function to those genes. Most current genome annotation systems work similarly, but the programs available for analysis of genomic DNA are constantly changing and improving.

Computational evolutionary biology

Evolutionary biology is the study of the origin and descent of species, as well as their change over time. Informatics has assisted evolutionary biologists in several key ways; it has enabled researchers to:

- trace the evolution of a large number of organisms by measuring changes in their DNA, rather than through physical taxonomy or physiological observations alone,
- more recently, compare entire genomes, which permits the study of more complex evolutionary events, such as gene duplication, horizontal gene transfer, and the prediction of factors important in bacterial speciation,
- build complex computational models of populations to predict the outcome of the system over time
- track and share information on an increasingly large number of species and organisms

Future work endeavours to reconstruct the now more complex tree of life.

The area of research within computer science that uses genetic algorithms is sometimes confused with computational evolutionary biology, but the two areas are not necessarily related.

Analysis of gene expression

The expression of many genes can be determined by measuring mRNA levels with multiple techniques including microarrays, expressed cDNA sequence tag (EST) sequencing, serial analysis of gene expression (SAGE) tag sequencing, massively parallel signature sequencing (MPSS), or various applications of multiplexed in-situ hybridization. All of these techniques are extremely noise-prone and/or subject to bias in the biological measurement, and a major research area in computational biology involves developing statistical tools to separate signal from noise in high-throughput gene expression studies. Such studies are often used to determine the genes implicated in a disorder: one might compare microarray data from cancerous epithelial cells to data from non-cancerous cells to determine the transcripts that are up-regulated and down-regulated in a particular population of cancer cells.

Analysis of regulation

Regulation is the complex orchestration of events starting with an extracellular signal such as a hormone and leading to an increase or decrease in the activity of one or more proteins. Bioinformatics techniques have been applied to explore various steps in this process. For example, promoter analysis involves the identification and study of sequence motifs in the DNA surrounding the coding region of a gene. These motifs influence the extent to which that region is transcribed into mRNA. Expression data can be used to infer gene regulation: one might compare microarray data from a wide variety of states of an organism to form hypotheses about the genes involved in each state. In a single-cell organism, one might compare stages of the cell cycle, along with various stress conditions (heat shock, starvation, etc.). One can then apply clustering algorithms to that expression data to determine which genes are co-expressed. For example, the upstream regions (promoters) of co-expressed genes can be searched for over-represented regulatory elements.

Analysis of protein expression

Protein microarrays and high throughput (HT) mass spectrometry (MS) can provide a snapshot of the proteins present in a biological sample. Bioinformatics is very much involved in making sense of protein microarray and HT MS data; the former approach faces similar problems as with microarrays targeted at mRNA, the latter involves the problem of matching large amounts of mass data against predicted masses from protein sequence databases, and the complicated statistical analysis of samples where multiple, but incomplete peptides from each protein are detected.

Analysis of mutations in cancer

In cancer, the genomes of affected cells are rearranged in complex or even unpredictable ways. Massive sequencing efforts are used to identify previously unknown point mutations in a variety of genes in cancer. Bioinformaticians continue to produce specialized automated systems to manage the sheer volume of sequence data produced, and they create new algorithms and software to compare the sequencing results to the growing collection of human genome sequences and germline polymorphisms. New physical detection technologies are employed, such as oligonucleotide microarrays to identify chromosomal gains and losses (called comparative genomic hybridization), and single-nucleotide polymorphism arrays to detect known *point mutations*. These detection methods simultaneously measure several hundred thousand sites throughout the genome, and when used in high-throughput to measure thousands of samples, generate terabytes of data per experiment. Again the massive amounts and new types of data generate new opportunities for bioinformaticians. The data is often found to contain considerable variability, or noise, and thus Hidden Markov model and change-point analysis methods are being developed to infer real copy number changes.

Another type of data that requires novel informatics development is the analysis of lesions found to be recurrent among many tumors.

Comparative genomics

The core of comparative genome analysis is the establishment of the correspondence between genes (orthology analysis) or other genomic features in different organisms. It is these intergenomic maps that make it possible to trace the evolutionary processes responsible for the divergence of two genomes. A multitude of evolutionary events acting at various organizational levels shape genome evolution. At the lowest level, point mutations affect individual nucleotides. At a higher level, large chromosomal segments undergo duplication, lateral transfer, inversion, transposition, deletion and insertion. Ultimately, whole genomes are involved in processes of hybridization, polyploidization and endosymbiosis, often leading to rapid speciation. The complexity of genome evolution poses many exciting challenges to developers of mathematical models and algorithms, who have recourse to a spectra of algorithmic, statistical and mathematical techniques, ranging from exact, heuristics, fixed parameter and approximation algorithms for problems based on parsimony models to Markov Chain Monte Carlo algorithms for Bayesian analysis of problems based on probabilistic models.

Many of these studies are based on the homology detection and protein families computation.

Modeling biological systems

Systems biology involves the use of computer simulations of cellular subsystems (such as the networks of metabolites and enzymes which comprise metabolism, signal transduction pathways and gene regulatory networks) to both analyze and visualize the

complex connections of these cellular processes. Artificial life or virtual evolution attempts to understand evolutionary processes via the computer simulation of simple (artificial) life forms.

High-throughput image analysis

Computational technologies are used to accelerate or fully automate the processing, quantification and analysis of large amounts of high-information-content biomedical imagery. Modern image analysis systems augment an observer's ability to make measurements from a large or complex set of images, by improving accuracy, objectivity, or speed. A fully developed analysis system may completely replace the observer. Although these systems are not unique to biomedical imagery, biomedical imaging is becoming more important for both diagnostics and research. Some examples are:

- high-throughput and high-fidelity quantification and sub-cellular localization (high-content screening, cytohistopathology, Bioimage informatics)
- morphometrics
- clinical image analysis and visualization
- determining the real-time air-flow patterns in breathing lungs of living animals
- quantifying occlusion size in real-time imagery from the development of and recovery during arterial injury
- making behavioral observations from extended video recordings of laboratory animals
- infrared measurements for metabolic activity determination
- inferring clone overlaps in DNA mapping, e.g. the Sulston score

Structural Bioinformatic Approaches

Prediction of protein structure

Protein structure prediction is another important application of bioinformatics. The amino acid sequence of a protein, the so-called primary structure, can be easily determined from the sequence on the gene that codes for it. In the vast majority of cases, this primary structure uniquely determines a structure in its native environment. (Of course, there are exceptions, such as the bovine spongiform encephalopathy - aka Mad Cow Disease - prion.) Knowledge of this structure is vital in understanding the function of the protein. For lack of better terms, structural information is usually classified as one of *secondary*, *tertiary* and *quaternary* structure. A viable general solution to such predictions remains an open problem. As of now, most efforts have been directed towards heuristics that work most of the time.

One of the key ideas in bioinformatics is the notion of homology. In the genomic branch of bioinformatics, homology is used to predict the function of a gene: if the sequence of gene *A*, whose function is known, is homologous to the sequence of gene *B,* whose function is unknown, one could infer that B may share A's function. In the structural branch of bioinformatics, homology is used to determine which parts of a protein are

important in structure formation and interaction with other proteins. In a technique called homology modeling, this information is used to predict the structure of a protein once the structure of a homologous protein is known. This currently remains the only way to predict protein structures reliably.

One example of this is the similar protein homology between hemoglobin in humans and the hemoglobin in legumes (leghemoglobin). Both serve the same purpose of transporting oxygen in the organism. Though both of these proteins have completely different amino acid sequences, their protein structures are virtually identical, which reflects their near identical purposes.

Other techniques for predicting protein structure include protein threading and *de novo* (from scratch) physics-based modeling.

Molecular Interaction

Efficient software is available today for studying interactions among proteins, ligands and peptides. Types of interactions most often encountered in the field include - Protein-ligand (including drug), protein-protein and protein-peptide.

Molecular dynamic simulation of movement of atoms about rotatable bonds is the fundamental principle behind computational algorithms, termed **docking algorithms** for studying molecular interactions.

Docking algorithms

In the last two decades, tens of thousands of protein three-dimensional structures have been determined by X-ray crystallography and Protein nuclear magnetic resonance spectroscopy (protein NMR). One central question for the biological scientist is whether it is practical to predict possible protein-protein interactions only based on these 3D shapes, without doing protein-protein interaction experiments. A variety of methods have been developed to tackle the Protein-protein docking problem, though it seems that there is still much work to be done in this field.

Software and tools

Software tools for bioinformatics range from simple command-line tools, to more complex graphical programs and standalone web-services available from various bioinformatics companies or public institutions.

Web services in bioinformatics

SOAP and REST-based interfaces have been developed for a wide variety of bioinformatics applications allowing an application running on one computer in one part of the world to use algorithms, data and computing resources on servers in other parts of

the world. The main advantages derive from the fact that end users do not have to deal with software and database maintenance overheads.

Basic bioinformatics services are classified by the EBI into three categories: SSS (Sequence Search Services), MSA (Multiple Sequence Alignment) and BSA (Biological Sequence Analysis). The availability of these service-oriented bioinformatics resources demonstrate the applicability of web based bioinformatics solutions, and range from a collection of standalone tools with a common data format under a single, standalone or web-based interface, to integrative, distributed and extensible bioinformatics workflow management systems.

Chapter- 5

Proteomics

Robotic preparation of MALDI mass spectrometry samples on a sample carrier.

Proteomics is the large-scale study of proteins, particularly their structures and functions. Proteins are vital parts of living organisms, as they are the main components of the physiological metabolic pathways of cells. The term "proteomics" was first coined in 1997 to make an analogy with genomics, the study of the genes. The word "proteome" is a blend of "**prote**in" and "gen**ome**", and was coined by Marc Wilkins in 1994 while working on the concept as a PhD student. The proteome is the entire complement of

proteins, including the modifications made to a particular set of proteins, produced by an organism or system. This will vary with time and distinct requirements, or stresses, that a cell or organism undergoes.

Complexity of the problem

After genomics, proteomics is considered the next step in the study of biological systems. It is much more complicated than genomics mostly because while an organism's genome is more or less constant, the proteome differs from cell to cell and from time to time. This is because distinct genes are expressed in distinct cell types. This means that even the basic set of proteins which are produced in a cell needs to be determined.

In the past this was done by mRNA analysis, but this was found not to correlate with protein content. It is now known that mRNA is not always translated into protein, and the amount of protein produced for a given amount of mRNA depends on the gene it is transcribed from and on the current physiological state of the cell. Proteomics confirms the presence of the protein and provides a direct measure of the quantity present.

Post-translational modifications

Not only does the translation from mRNA cause differences, many proteins are also subjected to a wide variety of chemical modifications after translation. Many of these post-translational modifications are critical to the protein's function.

Phosphorylation

One such modification is phosphorylation, which happens to many enzymes and structural proteins in the process of cell signaling. The addition of a phosphate to particular amino acids—most commonly serine and threonine mediated by serine/threonine kinases, or more rarely tyrosine mediated by tyrosine kinases—causes a protein to become a target for binding or interacting with a distinct set of other proteins that recognize the phosphorylated domain.

Because protein phosphorylation is one of the most-studied protein modifications, many "proteomic" efforts are geared to determining the set of phosphorylated proteins in a particular cell or tissue-type under particular circumstances. This alerts the scientist to the signaling pathways that may be active in that instance.

Ubiquitination

Ubiquitin is a small protein that can be affixed to certain protein substrates by enzymes called E3 ubiquitin ligases. Determining which proteins are poly-ubiquitinated can be helpful in understanding how protein pathways are regulated. This is therefore an additional legitimate "proteomic" study. Similarly, once it is determined what substrates are ubiquitinated by each ligase, determining the set of ligases expressed in a particular cell type will be helpful.

Additional modifications

Listing all the protein modifications that might be studied in a "Proteomics" project would require a discussion of most of biochemistry; therefore, a short list will serve here to illustrate the complexity of the problem. In addition to phosphorylation and ubiquitination, proteins can be subjected to (among others) methylation, acetylation, glycosylation, oxidation and nitrosylation. Some proteins undergo ALL of these modifications, often in time-dependent combinations, aptly illustrating the potential complexity one has to deal with when studying protein structure and function.

Distinct proteins are made under distinct settings

Even if one is studying a particular cell type, that cell may make different sets of proteins at different times, or under different conditions. Furthermore, as mentioned, any one protein can undergo a wide range of post-translational modifications.

Therefore a "proteomics" study can become quite complex very quickly, even if the object of the study is very restricted. In more ambitious settings, such as when a biomarker for a tumor is sought - when the proteomics scientist is obliged to study sera samples from multiple cancer patients - the amount of complexity that must be dealt with is as great as in any modern biological project.

Limitations to genomic study

Scientists are very interested in proteomics because it gives a much better understanding of an organism than genomics. First, the level of transcription of a gene gives only a rough estimate of its level of expression into a protein. An mRNA produced in abundance may be degraded rapidly or translated inefficiently, resulting in a small amount of protein. Second, as mentioned above many proteins experience post-translational modifications that profoundly affect their activities; for example some proteins are not active until they become phosphorylated. Methods such as phosphoproteomics and glycoproteomics are used to study post-translational modifications. Third, many transcripts give rise to more than one protein, through alternative splicing or alternative post-translational modifications. Fourth, many proteins form complexes with other proteins or RNA molecules, and only function in the presence of these other molecules. Finally, protein degradation rate plays an important role in protein content.

Methods of studying proteins

Determining proteins which are post-translationally modified

One way in which a particular protein can be studied is to develop an antibody which is specific to that modification. For example, there are antibodies which only recognize certain proteins when they are tyrosine-phosphorylated, known as phospho-specific antibodies; also, there are antibodies specific to other modifications. These can be used to determine the set of proteins that have undergone the modification of interest.

For sugar modifications, such as glycosylation of proteins, certain lectins have been discovered which bind sugars. These too can be used.

A more common way to determine post-translational modification of interest is to subject a complex mixture of proteins to electrophoresis in "two-dimensions", which simply means that the proteins are electrophoresed first in one direction, and then in another... this allows small differences in a protein to be visualized by separating a modified protein from its unmodified form. This methodology is known as "two-dimensional gel electrophoresis".

Recently, another approach has been developed called PROTOMAP which combines SDS-PAGE with shotgun proteomics to enable detection of changes in gel-migration such as those caused by proteolysis or post translational modification.

Determining the existence of proteins in complex mixtures

Classically, antibodies to particular proteins or to their modified forms have been used in biochemistry and cell biology studies. These are among the most common tools used by practicing biologists today.

For more quantitative determinations of protein amounts, techniques such as ELISAs can be used.

For proteomic study, more recent techniques such as matrix-assisted laser desorption/ionization (MALDI) have been employed for rapid determination of proteins in particular mixtures and increasingly electrospray ionization (ESI).

Computational methods in studying protein biomarkers

Computational predictive models have shown that extensive and diverse feto-maternal protein trafficking occurs during pregnancy and can be readily detected non-invasively in maternal whole blood. This computational approach circumvented a major limitation, the abundance of maternal proteins interfering with the detection of fetal proteins, to fetal proteomic analysis of maternal blood. Computational models can use fetal gene transcripts previously identified in maternal whole blood to create a comprehensive proteomic network of the term neonate. Such work shows that the fetal proteins detected in pregnant woman’s blood originate from a diverse group of tissues and organs from the developing fetus. The proteomic networks contain many biomarkers that are proxies for development and illustrate the potential clinical application of this technology as a way to monitor normal and abnormal fetal development.

An information theoretic framework has also been introduced for biomarker discovery, integrating biofluid and tissue information. This new approach takes advantage of functional synergy between certain biofluids and tissues with the potential for clinically significant findings not possible if tissues and biofluids were considered individually. By conceptualizing tissue-biofluid as information channels, significant biofluid proxies can

be identified and then used for guided development of clinical diagnostics. Candidate biomarkers are then predicted based on information transfer criteria across the tissue-biofluid channels. Significant biofluid-tissue relationships can be used to prioritize clinical validation of biomarkers.

Establishing protein-protein interactions

Most proteins function in collaboration with other proteins, and one goal of proteomics is to identify which proteins interact. This is especially useful in determining potential partners in cell signaling cascades.

Several methods are available to probe protein-protein interactions. The traditional method is yeast two-hybrid analysis. New methods include protein microarrays, immunoaffinity chromatography followed by mass spectrometry, dual polarisation interferometry, Microscale Thermophoresis and experimental methods such as phage display and computational methods

Practical applications of proteomics

One of the most promising developments to come from the study of human genes and proteins has been the identification of potential new drugs for the treatment of disease. This relies on genome and proteome information to identify proteins associated with a disease, which computer software can then use as targets for new drugs. For example, if a certain protein is implicated in a disease, its 3D structure provides the information to design drugs to interfere with the action of the protein. A molecule that fits the active site of an enzyme, but cannot be released by the enzyme, will inactivate the enzyme. This is the basis of new drug-discovery tools, which aim to find new drugs to inactivate proteins involved in disease. As genetic differences among individuals are found, researchers expect to use these techniques to develop personalized drugs that are more effective for the individual.

Biomarkers

The FDA defines a biomarker as, “A characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention”.

Understanding the proteome, the structure and function of each protein and the complexities of protein-protein interactions will be critical for developing the most effective diagnostic techniques and disease treatments in the future.

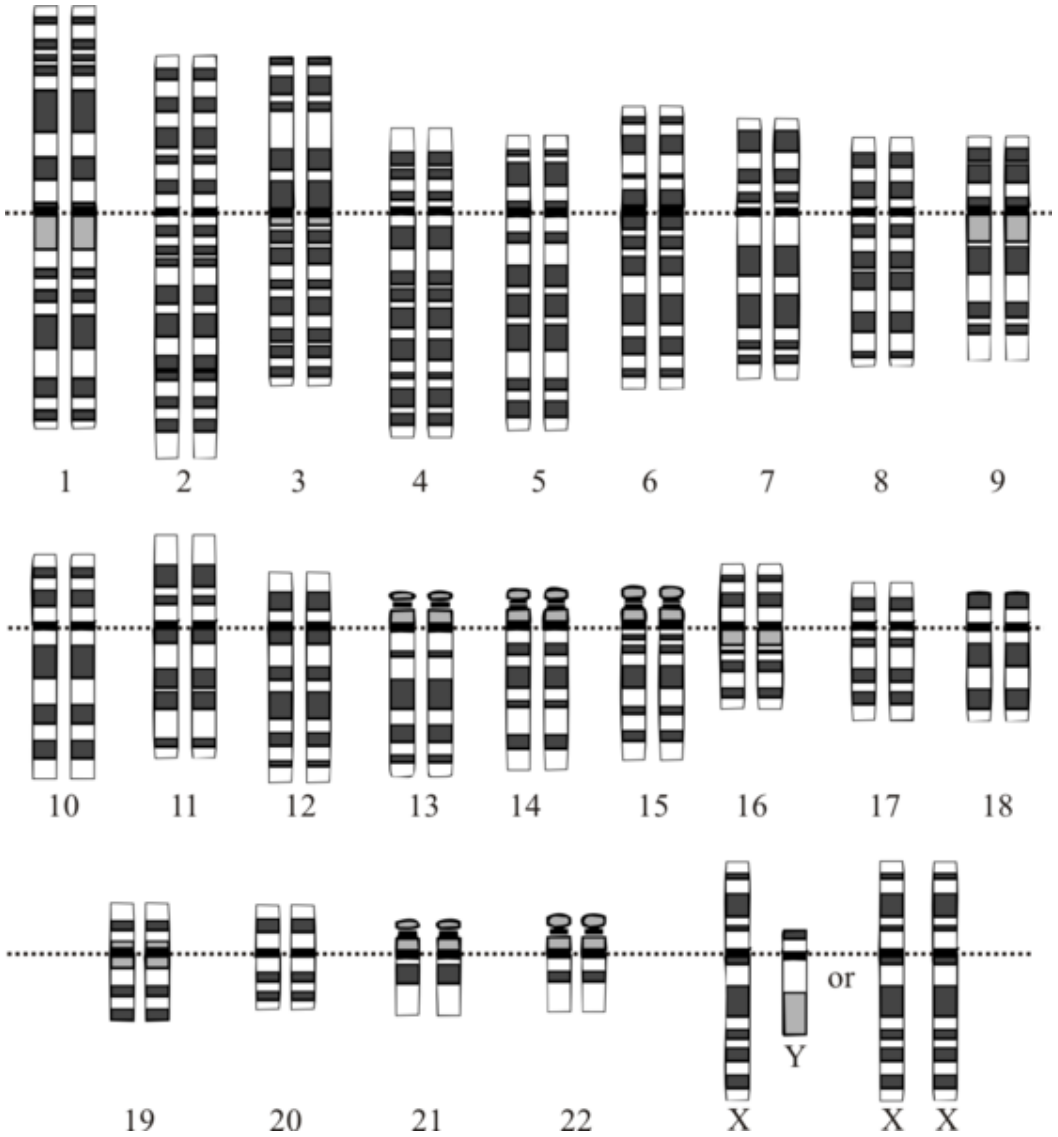
An interesting use of proteomics is using specific protein biomarkers to diagnose disease. A number of techniques allow to test for proteins produced during a particular disease, which helps to diagnose the disease quickly. Techniques include western blot, immunohistochemical staining, enzyme linked immunosorbent assay (ELISA) or mass spectrometry.

Current research methodologies

There are many approaches to attempting to characterize the human proteome, which is estimated to exceed 100,000 unique forms, 25,000 genes plus post-translational modifications.

Chapter- 6

Human Genome



A graphical representation of the normal human karyotype

The **human genome** is the genome of *Homo sapiens*, which is stored on 23 chromosome pairs. 22 of these are autosomal chromosome pairs, while the remaining pair is sex-determining. The haploid human genome occupies a total of just over 3 billion DNA base pairs. The Human Genome Project (HGP) produced a reference sequence of the euchromatic human genome, which is used worldwide in biomedical sciences.

The haploid human genome contains ca. 23,000 protein-coding genes, far fewer than had been expected before its sequencing. In fact, only about 1.5% of the genome codes for proteins, while the rest consists of non-coding RNA genes, regulatory sequences, introns, and noncoding DNA (once known as "junk DNA").

Features

Genes

There are estimated to be between 20,000 and 25,000 human protein-coding genes. The estimate of the number of human genes has been repeatedly revised down as genome sequence quality and gene finding methods have improved. Earlier predictions estimated that human cells have as much as 200,000 genes.

Surprisingly, the number of human genes seems to be less than a factor of two greater than that of many much simpler organisms, such as the roundworm and the fruit fly. However, human cells make extensive use of alternative splicing to produce several different proteins from a single gene, and the human proteome is thought to be much larger than those of the aforementioned organisms. Besides, most human genes have multiple exons, and human introns are frequently much longer than the flanking exons.

Human genes are distributed unevenly across the chromosomes. Each chromosome contains various gene-rich and gene-poor regions, which seem to be correlated with chromosome bands and GC-content. The significance of these nonrandom patterns of gene density is not well understood. In addition to protein coding genes, the human genome contains thousands of RNA genes, including tRNA, ribosomal RNA, microRNA, and other non-coding RNA genes.

Regulatory sequences

The human genome has many different regulatory sequences which are crucial to controlling gene expression. These are typically short sequences that appear near or within genes. A systematic understanding of these regulatory sequences and how they together act as a gene regulatory network is only beginning to emerge from computational, high-throughput expression and comparative genomics studies. Some types of non-coding DNA are genetic "switches" that do not encode proteins, but do regulate when and where genes are expressed.

Identification of regulatory sequences relies in part on evolutionary conservation. The evolutionary branch between the primates and mouse, for example, occurred 70–90

million years ago. So computer comparisons of gene sequences that identify conserved non-coding sequences will be an indication of their importance in duties such as gene regulation.

Another comparative genomic approach to locating regulatory sequences in humans is the gene sequencing of the puffer fish. These vertebrates have essentially the same genes and regulatory gene sequences as humans, but with only one-eighth the noncoding DNA. The compact DNA sequence of the puffer fish makes it much easier to locate the regulatory genes.

Other DNA

Protein-coding sequences (specifically, coding exons) comprise less than 1.5% of the human genome. Aside from genes and known regulatory sequences, the human genome contains vast regions of DNA the function of which, if any, remains unknown. These regions in fact comprise the vast majority, by some estimates 97%, of the human genome size. Much of this is composed of:

Repeat elements

- Tandem repeats
  - Satellite DNA
  - Minisatellite
  - Microsatellite
- Interspersed repeats
  - SINEs
  - LINEs

Transposons

- Retrotransposons
  - LTR
    - Ty1-copia
    - Ty3-gypsy
  - Non-LTR
    - SINEs
    - LINEs
- DNA Transposons

Noncoding DNA

There is also a large amount of sequence that does not fall under any known classification. Much of this sequence may be an evolutionary artifact that serves no present-day purpose, and these regions are collectively referred to as noncoding DNA. These regions were once referred to as "junk" DNA; however, there are a variety of emerging indications that many sequences within are likely to function in ways that are

not fully understood. Recent experiments using microarrays have revealed that a substantial fraction of non-genic DNA is in fact transcribed into RNA, which leads to the possibility that the resulting transcripts may have some unknown function. Also, the evolutionary conservation across the mammalian genomes of much more sequence than can be explained by protein-coding regions indicates that many, and perhaps most, functional elements in the genome remain unknown. The investigation of the vast quantity of sequence information in the human genome whose function remains unknown is currently a major avenue of scientific inquiry. Meanwhile, considering the global genome DNA information as a whole could provide new ways to understand a possible global level function of non coding DNA.

Information content

The 2.9 billion base pairs of the haploid human genome correspond to a maximum of about 691.4 megabytes of data, since every base pair can be coded by 2 bits. However, due to the high degree of redundancy of the human genome, it can be losslessly compressed to roughly 4 megabytes.

The entropy rate of the genome differs significantly between coding and non-coding sequences. It is close to the maximum of 2 bits per base pair for the coding sequences (about 45 million base pairs), but less for the non-coding parts. It ranges between 1.5 and 1.9 bits per base pair for the individual chromosome, except for the Y chromosome, which has an entropy rate below 0.9 bits per base pair.

Sequencing

DNA sequencing determines the order of the nucleotide bases in a genome.

Composite

The Human Genome Project and a parallel project by Celera Genomics each produced and published a haploid human genome sequence, both of which were a composite of the DNA sequence of several individuals.

Personal

A personal genome sequence is a complete sequencing of the chemical base pairs that make up the DNA of a single person. Because medical treatments have different effects on different people because of genetic variations such as single-nucleotide polymorphisms (SNPs), the analysis of personal genomes may lead to personalized medical treatment based on individual genotypes.

The completion of the fifth such map was announced in December 2008. The genome mapped was that of a Korean researcher Seong-Jin Kim. Genome maps had previously been completed for Craig Venter of the U.S. in 2007, James Watson of the U.S. in April

2008, and Yang Huanming of China in November 2008 and Dan Stoicescu in January 2008.

Personal genomes had not been sequenced in the Human Genome Project to protect the identity of volunteers who provided DNA samples. That sequence was derived from the DNA of several volunteers from a diverse population. Another distinction is that the HGP sequence is haploid, however, the sequence maps for Venter and Watson for example are diploid, representing both sets of chromosomes.

Kim's genome had 1.58 million SNPs that had never been reported before and indicates that six out of 10,000 DNA bases are unique to Koreans. Kim's sequence map can be used to assist in building a standard Korean genome, which can then be used to compare the genomes of other Korean individuals for personalized medical treatments.

Mapping

Whereas a genome sequence lists the order of every DNA base in a genome, a genome map identifies the landmarks. A genome map is less detailed than a genome sequence and aids in navigating around the genome.

Variation

An example of a variation map is the HapMap being developed by the International HapMap Project. The HapMap is a haplotype map of the human genome, "which will describe the common patterns of human DNA sequence variation." It catalogs the patterns of small-scale variations in the genome that involve single DNA letters, or bases.

Researchers published the first sequence-based map of large-scale structural variation across the human genome in the journal *Nature* in May 2008. Large-scale structural variations are differences in the genome among people that range from a few thousand to a few million DNA bases; some are gains or losses of stretches of genome sequence and others appear as re-arrangements of stretches of sequence. These variations include differences in the number of copies individuals have of a particular gene, deletions, translocations and inversions.

Variation

Most studies of human genetic variation have focused on single-nucleotide polymorphisms (SNPs), which are substitutions in individual bases along a chromosome. Most analyses estimate that SNPs occur on average somewhere between every 1 in 100 and 1 in 300 base pairs in the euchromatic human genome, although they do not occur at a uniform density. Thus follows the popular statement that "we are all, regardless of race, genetically 99.9% the same", although this would be somewhat qualified by most geneticists. For example, a much larger fraction of the genome is now thought to be involved in copy number variation. A large-scale collaborative effort to catalog SNP

variations in the human genome is being undertaken by the International HapMap Project.

The genomic loci and length of certain types of small repetitive sequences are highly variable from person to person, which is the basis of DNA fingerprinting and DNA paternity testing technologies. The heterochromatic portions of the human genome, which total several hundred million base pairs, are also thought to be quite variable within the human population (they are so repetitive and so long that they cannot be accurately sequenced with current technology). These regions contain few genes, and it is unclear whether any significant phenotypic effect results from typical variation in repeats or heterochromatin.

Most gross genomic mutations in gamete germ cells probably result in inviable embryos; however, a number of human diseases are related to large-scale genomic abnormalities. Down syndrome, Turner Syndrome, and a number of other diseases result from nondisjunction of entire chromosomes. Cancer cells frequently have aneuploidy of chromosomes and chromosome arms, although a cause and effect relationship between aneuploidy and cancer has not been established.

Genetic disorders

Most aspects of human biology involve both genetic (inherited) and non-genetic (environmental) factors. Some inherited variation influences aspects of our biology that are not medical in nature (height, eye color, ability to taste or smell certain compounds, etc.). Moreover, some genetic disorders only cause disease in combination with the appropriate environmental factors (such as diet). With these caveats, genetic disorders may be described as clinically defined diseases caused by genomic DNA sequence variation. In the most straightforward cases, the disorder can be associated with variation in a single gene. For example, cystic fibrosis is caused by mutations in the CFTR gene, and is the most common recessive disorder in caucasian populations with over 1,300 different mutations known. Disease-causing mutations in specific genes are usually severe in terms of gene function, and are fortunately rare, thus genetic disorders are similarly individually rare. However, since there are many genes that can vary to cause genetic disorders, in aggregate they comprise a significant component of known medical conditions, especially in pediatric medicine. Molecularly characterized genetic disorders are those for which the underlying causal gene has been identified, currently there are approximately 2,200 such disorders annotated in the OMIM database.

Studies of genetic disorders are often performed by means of family-based studies. In some instances population based approaches are employed, particularly in the case of so-called founder populations such as those in Finland, French-Canada, Utah, Sardinia, etc. Diagnosis and treatment of genetic disorders are usually performed by a geneticist-physician trained in clinical/medical genetics. The results of the Human Genome Project are likely to provide increased availability of genetic testing for gene-related disorders, and eventually improved treatment. Parents can be screened for hereditary conditions and

counselled on the consequences, the probability it will be inherited, and how to avoid or ameliorate it in their offspring.

As noted above, there are many different kinds of DNA sequence variation, ranging from complete extra or missing chromosomes down to single nucleotide changes. It is generally presumed that much naturally occurring genetic variation in human populations is phenotypically neutral, i.e. has little or no detectable effect on the physiology of the individual (although there may be fractional differences in fitness defined over evolutionary time frames). Genetic disorders can be caused by any or all known types of sequence variation. To molecularly characterize a new genetic disorder, it is necessary to establish a causal link between a particular genomic sequence variant and the clinical disease under investigation. Such studies constitute the realm of human molecular genetics.

With the advent of the Human Genome and International HapMap Project, it has become feasible to explore subtle genetic influences on many common disease conditions such as diabetes, asthma, migraine, schizophrenia, etc. Although some causal links have been made between genomic sequence variants in particular genes and some of these diseases, often with much publicity in the general media, these are usually not considered to be genetic disorders *per se* as their causes are complex, involving many different genetic and environmental factors. Thus there may be disagreement in particular cases whether a specific medical condition should be termed a genetic disorder.

Evolution

Comparative genomics studies of mammalian genomes suggest that approximately 5% of the human genome has been conserved by evolution since the divergence of extant lineages approximately 200 million years ago, containing the vast majority of genes. Intriguingly, since genes and known regulatory sequences probably comprise less than 2% of the genome, this suggests that there may be more unknown functional sequence than known functional sequence. A smaller, yet substantial, fraction of human genes seem to be shared among most known vertebrates. The published chimpanzee genome differs from that of the human genome by 1.23% in direct sequence comparisons. Around 20% of this figure is accounted for by variation within each species, leaving only ~1.06% consistent sequence divergence between humans and chimps at shared genes. This nucleotide by nucleotide difference is dwarfed, however, by the portion of each genome that is not shared, including around 6% of functional genes that are unique to either humans or chimps. In other words, the considerable observable differences between humans and chimps may be due as much or more to genome level variation in the number, function and expression of genes rather than DNA sequence changes in shared genes. On average, a typical human protein-coding gene differs from its chimpanzee ortholog by only two amino acid substitutions; nearly one third of human genes have exactly the same protein translation as their chimpanzee orthologs. A major difference between the two genomes is human chromosome 2, which is equivalent to a fusion product of chimpanzee chromosomes 12 and 13 (later renamed to chromosomes 2A and 2B, respectively).

Humans have undergone an extraordinary loss of olfactory receptor genes during our recent evolution, which explains our relatively crude sense of smell compared to most other mammals. Evolutionary evidence suggests that the emergence of color vision in humans and several other primate species has diminished the need for the sense of smell.

Mitochondrial genome

The human mitochondrial genome, while usually not included when referring to the "human genome", is of tremendous interest to geneticists, since it undoubtedly plays a role in mitochondrial disease. It also sheds light on human evolution; for example, analysis of variation in the human mitochondrial genome has led to the postulation of a recent common ancestor for all humans on the maternal line of descent.

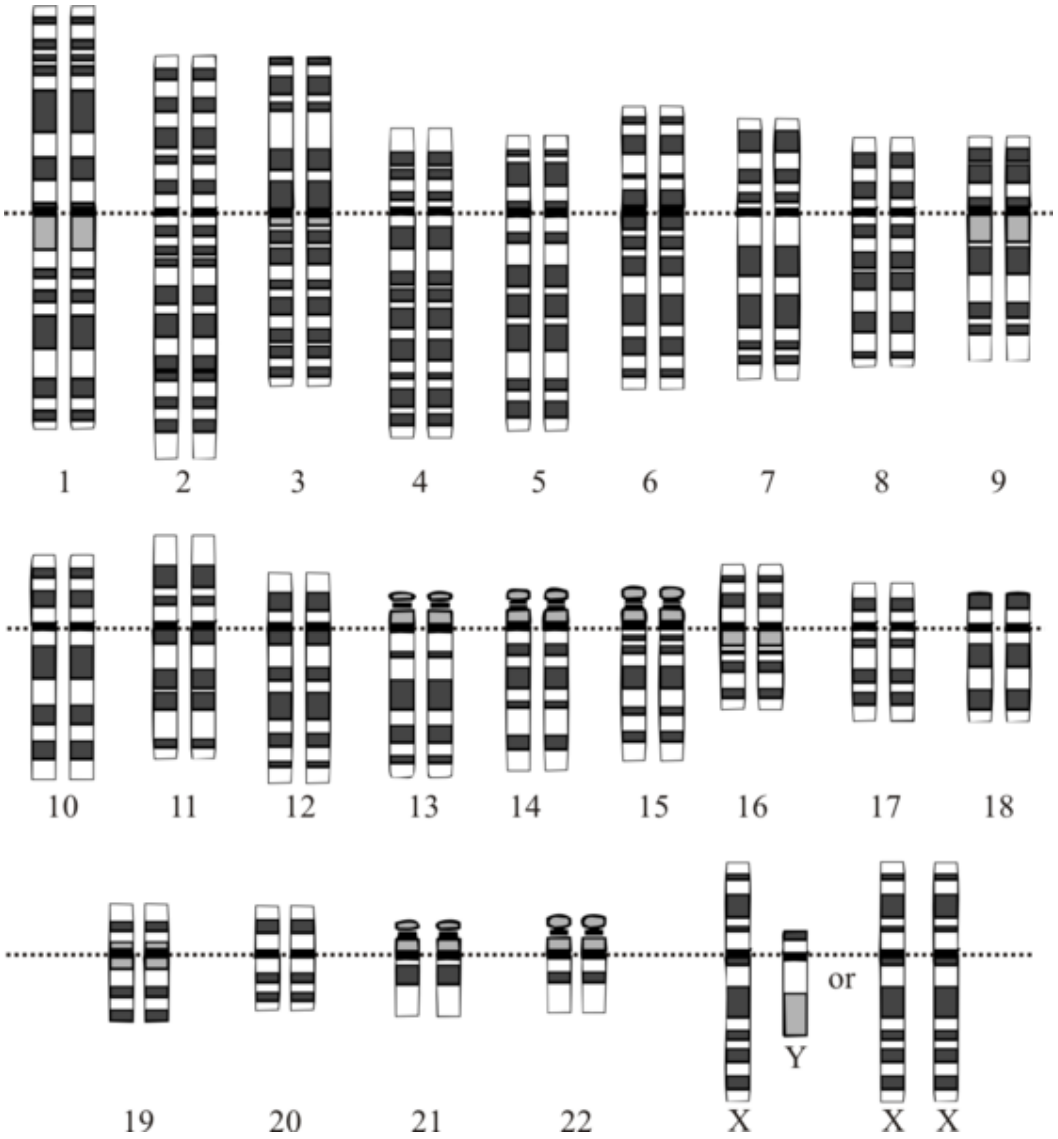
Due to the lack of a system for checking for copying errors, Mitochondrial DNA (mtDNA) has a more rapid rate of variation than nuclear DNA. This 20-fold increase in the mutation rate allows mtDNA to be used for more accurate tracing of maternal ancestry. Studies of mtDNA in populations have allowed ancient migration paths to be traced, such as the migration of Native Americans from Siberia or Polynesians from southeastern Asia. It has also been used to show that there is no trace of Neanderthal DNA in the European gene mixture inherited through purely maternal lineage.

Epigenome

Epigenetics are a variety of features of the human genome that transcend its primary DNA sequence, such as chromatin packaging, histone modifications and DNA methylation, and which are important in regulating gene expression, genome replication and other cellular processes. Epigenetic markers strengthen and weaken transcription of certain genes but do not affect the actual sequence of DNA nucleotides.

Chapter- 7

Human Genetic Variation



A graphical representation of the normal human karyotype

Human genetic variation refers to genetic differences both within and among populations. There may be multiple variants of any given gene in the human population (alleles), leading to polymorphism. Many genes are not polymorphic, meaning that only a single allele is present in the population: that allele is then said to be fixed.

No two humans are genetically identical. Even monozygotic twins, who develop from one zygote, have infrequent genetic differences due to mutations occurring during development and gene copy number variation has been observed. Differences between individuals, even closely related individuals, are the key to techniques such as genetic fingerprinting. Alleles occur at different frequencies in different human populations, with populations that are more geographically and ancestrally remote tending to differ more.

Causes of differences between individuals include the exchange of genes during meiosis and various mutational events. There are at least two reasons why genetic variation exists between populations. Natural selection may confer an adaptive advantage to individuals in a specific environment if an allele provides a competitive advantage. Alleles under selection are likely to occur only in those geographic regions where they confer an advantage. The second main cause of genetic variation is due to the high degree of neutrality of most mutations. Most mutations do not appear to have any selective effect one way or the other on the organism. The main cause is genetic drift, this is the effect of random changes in the gene pool. In humans, founder effect and past small population size (increasing the likelihood of genetic drift) may have had an important influence in neutral differences between populations.

The theory that humans recently migrated out of Africa is sometimes given as an example of this. It has been theorized that the population which migrated out of Africa only represented a small fraction of the genetic variation in Africa, and that this is a contributing cause of the observed lower levels of diversity in all indigenous humans outside of Africa. Generally, more recent neutral polymorphisms caused by mutation are likely to be relatively geographically localized and rare, while older polymorphisms are more likely to be shared by a wider range of human groups. The large majority of observed genetic variation occurs within a population in any geographic region and not between populations in different regions, although it is still usually possible to accurately identify the geographic origins of any individual's ancestors by genetic means.

The study of human genetic variation has both evolutionary significance and medical applications. The study can help scientists understand ancient human population migrations as well as how different human groups are biologically related to one another. From a medical perspective the study of human genetic variation may be important because some disease causing alleles occur at a greater frequency in people from specific geographic regions.

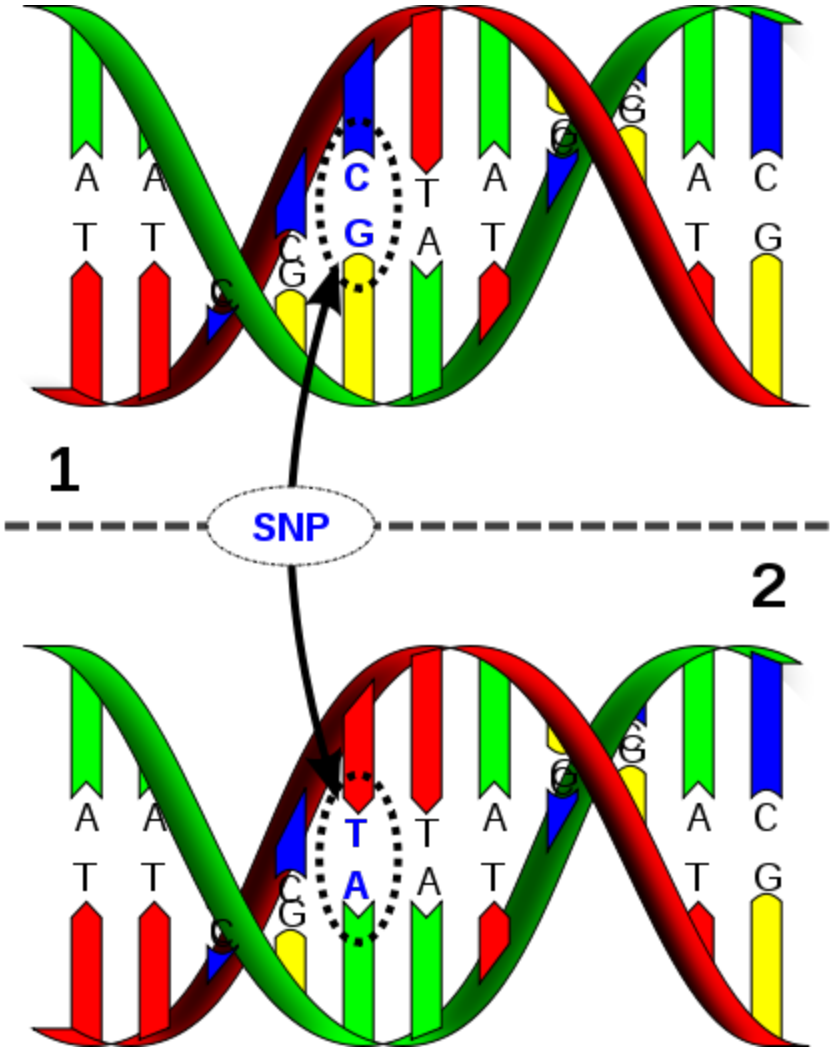
Genetic variation

Genetic variation, variation in alleles of genes, occurs both within and among populations. Genetic variation is important because it provides the “raw material” for natural selection.

Measures of variation

"Genetic variation among individual humans occurs on many different scales, ranging from gross alterations in the human karyotype to single nucleotide changes."

Single nucleotide polymorphisms



DNA molecule 1 differs from DNA molecule 2 at a single base-pair location (a C/T polymorphism).

Nucleotide diversity is based on single mutations called single nucleotide polymorphisms (SNPs). The nucleotide diversity between humans is about 0.1%, which is 1 difference per 1,000 base pairs. A difference of 1 in 1,000 nucleotides between two humans chosen at random amounts to approximately 3 million nucleotide differences since the human genome has about 3 billion nucleotides. Most of these SNPs are neutral but some are functional and influence phenotypic differences between humans through alleles. It is estimated that a total of 10 million SNPs exist in the human population of which at least 1% are functional.

Copy number variation

More recently a better understanding of the structure of the genome has been gained with the publication of two examples of full sequences of an individual's genome. This represents a new development because the Human Genome Project and a parallel project by Celera Genomics produced two haploid sequences, both of which were an amalgamation of sequences from many individuals. Recently the diploid sequences of both Craig Venter and James Watson have been published. Analysis of diploid sequences has shown that non-SNP variation accounts for much more human genetic variation than single nucleotide diversity. This non-SNP variation includes copy number variation and results from deletions, inversions, insertions and duplications. It is estimated that approximately 0.4% of the genomes of unrelated people typically differ with respect to copy number. When copy number variation is included, human to human genetic variation is estimated to be at least 0.5% (99.5% similarity). Copy number variations are inherited but can also arise during development.

Epigenetics

Epigenetics is another type of genetic variation. "This type of variation arises from chemical tags that attach to DNA and affect how it gets read. The chemical tags, called epigenetic markings, act as switches that control how genes can be read." At some alleles, the epigenetic state of the DNA, and associated phenotype, can be inherited transgenerationally.

Genetic variability

Genetic variability is a measure of the tendency of individual genotypes in a population to vary (become different) from one another. Variability is different from genetic diversity, which is the amount of variation seen in a particular population. The variability of a trait describes how much that trait tends to vary in response to environmental and genetic influences.

Clines

In biology, a cline is a term used to describe a continuum of species, populations, races, varieties, or forms of organisms that exhibit gradual phenotypic and/or genetic differences over a geographical area, typically as a result of environmental heterogeneity.

In the scientific study of human genetic variation, a gene cline can be rigorously defined and subjected to quantitative metrics.

Haplogroups

In the study of molecular evolution, a haplogroup is a group of similar haplotypes that share a common ancestor with a single nucleotide polymorphism (SNP) mutation. Haplogroups pertain to deep ancestral origins dating back thousands of years.

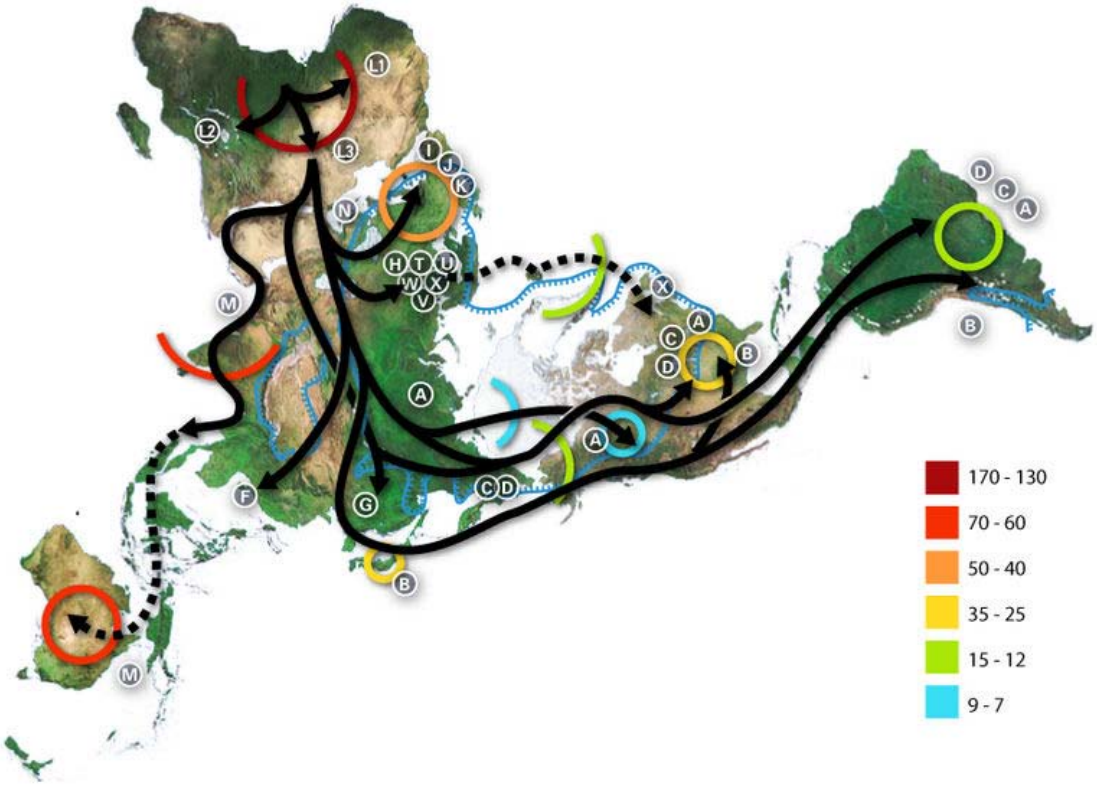
In human genetics, the haplogroups most commonly studied are Y-chromosome (Y-DNA) haplogroups and mitochondrial DNA (mtDNA) haplogroups, both of which can be used to define genetic populations. Y-DNA is passed solely along the patrilineal line, from father to son, while mtDNA is passed down the matrilineal line, from mother to both daughter and son. The Y-DNA and mtDNA may change by chance mutation at each generation.

Variable number tandem repeats

A variable number tandem repeat (VNTR) is a location in a genome where a short nucleotide sequence is organized as a tandem repeat. These can be found on many chromosomes, and often show variations in length between individuals. Each variant acts as an inherited allele, allowing them to be used for personal or parental identification. Their analysis is useful in genetics and biology research, forensics, and DNA fingerprinting.

There are two principal families of VNTRs: microsatellites and minisatellites. The former are repeats of sequences less than about 5 base pairs in length, while the latter involve longer blocks.

History and geographic distribution



Map of the migration of modern humans out of Africa, based on mitochondrial DNA. Colored rings indicate thousand years before present.

A 10-year study published in 2009 analyzed the patterns of variation at 1,327 DNA markers of 121 African populations, 4 African American populations, and 60 non-African populations. The research showed that there is more human genetic diversity in Africa than anywhere else on Earth. The genetic structure of Africans was traced to 14 ancestral population clusters and the ancestral origin of humans was determined to probably be located in southern Africa, near the border of Namibia and South Africa.

Human genetic diversity decreases in native populations with migratory distance from Africa and this is thought to be the result of bottlenecks during human migration, which are events that temporarily reduce population size. It has been shown that variations in skull measurements decrease with distance from Africa at the same rate as the decrease in genetic diversity. These data support the Out of Africa theory over the multiregional origin of modern humans hypothesis. The aforementioned April 2009 study identifies the likely origin of modern human migration as being in southwestern Africa, near the coastal border of Namibia and Angola, and the exit point out of Africa as being in East Africa.

The *recent African origin of modern humans* is the mainstream model describing the origin and early dispersal of anatomically modern humans, *Homo sapiens sapiens*. The

theory is known popularly as the *(Recent) Out-of-Africa* model. The hypothesis originated in the 19th century, with Darwin's *Descent of Man*, but remained speculative until the 1980s when it was corroborated based on a study of present-day mitochondrial DNA, combined with evidence based on physical anthropology of archaic specimens.

According to both genetic and fossil evidence, archaic Homo sapiens evolved to anatomically modern humans solely in Africa, between 200,000 and 100,000 years ago, with members of one branch leaving Africa by 60,000 years ago and over time replacing earlier human populations such as Neanderthals and *Homo erectus*. According to this theory, around the above time frame, one of the African subpopulations went through a process of speciation prohibiting gene flow between African and Eurasian Human populations.

Population genetics

In the field of population genetics, it is believed that the distribution of neutral polymorphisms among contemporary humans reflects human demographic history. It is believed that humans passed through a population bottleneck before a rapid expansion coinciding with migrations out of Africa leading to an African-Eurasian divergence around 100,000 years ago (ca. 5,000 generations), followed by a European-Asian divergence about 40,000 years ago (ca. 2,000 generations). Richard G. Klein, Nicholas Wade and Spencer Wells, among others, have postulated that modern humans did not leave Africa and successfully colonize the rest of the world until as recently as 60,000 - 50,000 years B.P., pushing back the dates for subsequent population splits as well.

The rapid expansion of a previously small population has two important effects on the distribution of genetic variation. First, the so-called founder effect occurs when founder populations bring only a subset of the genetic variation from their ancestral population. Second, as founders become more geographically separated, the probability that two individuals from different founder populations will mate becomes smaller. The effect of this assortative mating is to reduce gene flow between geographical groups, and to increase the genetic distance between groups. The expansion of humans from Africa affected the distribution of genetic variation in two other ways. First, smaller (founder) populations experience greater genetic drift because of increased fluctuations in neutral polymorphisms. Second, new polymorphisms that arose in one group were less likely to be transmitted to other groups as gene flow was restricted.

Our history as a species also has left genetic signals in regional populations. For example, in addition to having higher levels of genetic diversity, populations in Africa tend to have lower amounts of linkage disequilibrium than do populations outside Africa, partly because of the larger size of human populations in Africa over the course of human history and partly because the number of modern humans who left Africa to colonize the rest of the world appears to have been relatively low (Gabriel *et al.* 2002). In contrast, populations that have undergone dramatic size reductions or rapid expansions in the past and populations formed by the mixture of previously separate ancestral groups can have unusually high levels of linkage disequilibrium (Nordborg and Tavare 2002).

Many other geographic, climatic, and historical factors have contributed to the patterns of human genetic variation seen in the world today. For example, population processes associated with colonization, periods of geographic isolation, socially reinforced endogamy, and natural selection all have affected allele frequencies in certain populations (Jorde *et al.* 2000b; Bamshad and Wooding 2003). In general, however, the recency of our common ancestry and continual gene flow among human groups have limited genetic differentiation in our species.

Distribution of variation

The distribution of genetic variants within and among human populations are impossible to describe succinctly because of the difficulty of defining a "population," the clinal nature of variation, and heterogeneity across the genome (Long and Kittles 2003). In general, however, an average of 85% of genetic variation exists within local populations, ~7% is between local populations within the same continent, and ~8% of variation occurs between large groups living on different continents,. (Lewontin 1972; Jorde *et al.* 2000a; Hinds *et al.* 2005). The recent African origin theory for humans would predict that in Africa there exists a great deal more diversity than elsewhere, and that diversity should decrease the further from Africa a population is sampled. Long and Kittles show that indeed, African populations contain about 100% of human genetic diversity, whereas in populations outside of Africa diversity is much reduced, for example in their population from New Guinea only about 70% of human variation is captured.

Phenotypic variation

Faces show phenotypic variation. Some of this is caused by genetic variation.

Sub-Saharan Africa has the most human genetic diversity and the same has been shown to hold true for phenotypic diversity. Phenotype is connected to genotype through gene expression. Genetic diversity decreases smoothly with migratory distance from that region, which many scientists believe to be the origin of modern humans, and that decrease is mirrored by a decrease in phenotypic variation. Skull measurements are an example of a physical attribute whose within-population variation decreases with distance from Africa.

The distribution of many physical traits resembles the distribution of genetic variation within and between human populations (American Association of Physical Anthropologists 1996; Keita and Kittles 1997). For example, ~90% of the variation in human head shapes occurs within continental groups, and ~10% separates groups, with a greater variability of head shape among individuals with recent African ancestors (Relethford 2002).

A prominent exception to the common distribution of physical characteristics within and among groups is skin color. Approximately 10% of the variance in skin color occurs within groups, and ~90% occurs between groups (Relethford 2002). This distribution of skin color and its geographic patterning — with people whose ancestors lived predominantly near the equator having darker skin than those with ancestors who lived predominantly in higher latitudes — indicate that this attribute has been under strong selective pressure. Darker skin appears to be strongly selected for in equatorial regions to prevent sunburn, skin cancer, the photolysis of folate, and damage to sweat glands (Sturm *et al.* 2001; Rees 2003).

A study published in 2007 found that 25% of genes showed different levels of gene expression between populations of European and Asian descent. The primary cause of this difference in gene expression was thought to be SNPs in gene regulatory regions of DNA. Another study published in 2007 found that approximately 83% of genes were expressed at different levels among individuals and about 17% between populations of European and African descent.

Archaic admixture

Interbreeding of Neanderthals and anatomically modern humans during the Middle Paleolithic is a hypothesis. In May 2010, the Neanderthal Genome Project presented genetic evidence that interbreeding did likely take place and that a small but significant portion of Neanderthal admixture is present in the DNA of modern non-African populations.

In December 2010, a study found that between 4% and 6% of the genome of Melanesians (represented by the Papua New Guinean and Bougainville Islander) derives from Denisova hominin - a previously unknown species, which shares common origin with Neanderthals.It was possibly introduced during the early migration of the ancestors of Melanesians into Southeast Asia. This history of interaction suggests that Denisovans once ranged widely over eastern Asia.

Melanesians thus emerge as the most archaic-admixed population, having Denisovan/Neandertal-related admixture of ~8%.

Categorization of the world population

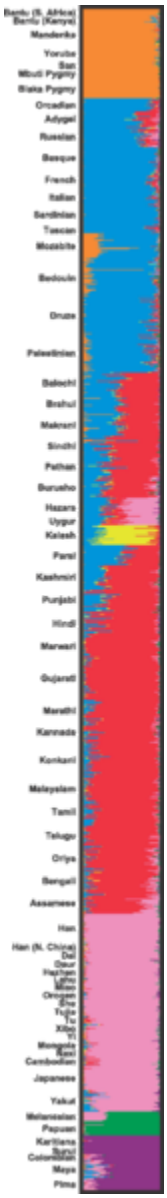


Chart showing human genetic clustering.

New data on human genetic variation has reignited the debate surrounding race. Most of the controversy surrounds the question of how to interpret this new data and whether conclusions based on existing data are sound. A large majority of researchers endorse the view that continental groups do not constitute different subspecies. However, other researchers still debate whether evolutionary lineages should rightly be called "races". These questions are particularly pressing for ancestry related health issues, where self-identified race is often used as an indicator of ancestry.

Although the genetic differences among human groups are relatively small, these differences in certain genes such as duffy, ABCC11, SLC24A5, called ancestry-informative markers (AIMs) nevertheless can be used to reliably situate many individuals within broad, geographically based groupings or self-identified race. For example, computer analyses of hundreds of polymorphic loci sampled in globally distributed populations have revealed the existence of genetic clustering that roughly is associated with groups that historically have occupied large continental and subcontinental regions (Rosenberg *et al.* 2002; Bamshad *et al.* 2003).

Some commentators have argued that these patterns of variation provide a biological justification for the use of traditional racial categories. They argue that the continental clusterings correspond roughly with the division of human beings into sub-Saharan Africans; Europeans, Western Asians, Central Asians, Southern Asians and Northern Africans; Eastern Asians, Southeast Asians, Polynesians and Native Americans; and other inhabitants of Oceania (Melanesians, Micronesians & Australian Aborigines) (Risch *et al.* 2002). Other observers disagree, saying that the same data undercut traditional notions of racial groups (King and Motulsky 2002; Calafell 2003; Tishkoff and Kidd 2004). They point out, for example, that major populations considered races or subgroups within races do not necessarily form their own clusters.

Furthermore, because human genetic variation is clinal, many individuals affiliate with two or more continental groups. Thus, the genetically based "biogeographical ancestry" assigned to any given person generally will be broadly distributed and will be accompanied by sizable uncertainties (Pfaff *et al.* 2004).

In many parts of the world, groups have mixed in such a way that many individuals have relatively recent ancestors from widely separated regions. Although genetic analyses of large numbers of loci can produce estimates of the percentage of a person's ancestors coming from various continental populations (Shriver *et al.* 2003; Bamshad *et al.* 2004), these estimates may assume a false distinctiveness of the parental populations, since human groups have exchanged mates from local to continental scales throughout history (Cavalli-Sforza *et al.* 1994; Hoerder 2002). Even with large numbers of markers, information for estimating admixture proportions of individuals or groups is limited, and estimates typically will have wide confidence intervals (Pfaff *et al.* 2004).

Lewontin's Fallacy

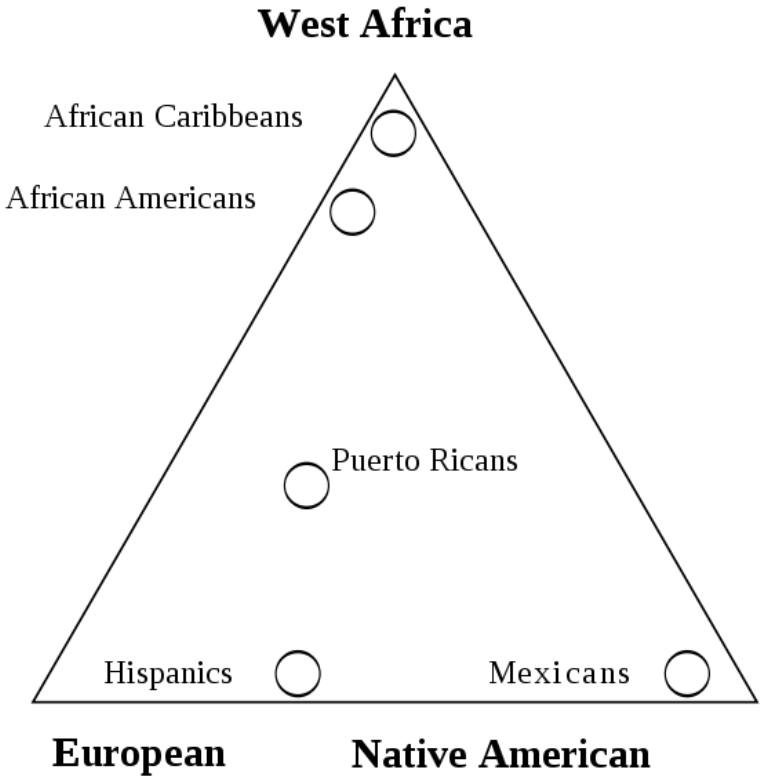
In 2003 A. W. F. Edwards wrote a paper called Lewontin's Fallacy, rebutting the argument that because most genetic variation is within-group, classification of humans is not possible. He claimed that this conclusion ignores the fact that most of the information that distinguishes populations is hidden in the correlation structure of the data and not simply in the variation of the individual factors. Edwards concludes that "It is not true that 'racial classification is ... of virtually no genetic or taxonomic significance' or that 'you can't predict someone’s race by their genes'." Undeterred, in an article titled "Confusions About Human Races" published in 2006, Lewontin maintains that race is no more than a social construct.

Genetic clustering

Genetic data can be used to infer population structure and assign individuals to groups that often correspond with their self-identified geographical ancestry. Recently, Lynn Jorde and Steven Wooding argued that "Analysis of many loci now yields reasonably accurate estimates of genetic similarity among individuals, rather than populations. Clustering of individuals is correlated with geographic origin or ancestry."

Forensic anthropology

Forensic anthropologists can determine race (e.g. Asian, African, or European ancestry) from skeletal remains with a high degree of accuracy by conducting bone analysis. Studies have shown that individual test methods such as midfacial measurements and femur traits can be over 80 percent accurate, and in combination can achieve very high levels of accuracy. The skeletons of mixed-race individuals can, however, exhibit characteristics of more than one racial group. Despite the success of this method with the remains of individuals with ancestry predominantly from a single race, anthropologists, including George W. Gill and C. Loring Brace, disagree on whether race is a valid biological concept.



Triangle plot shows average admixture of five North American ethnic groups. Individuals that self-identify with each group can be found at many locations on the map, but on average groups tend to cluster differently.

Admixture

Miscegenation between two populations reduces the average genetic distance between the populations. During the Age of Discovery which began in the early 15th century, European explorers sailed all around the globe, reaching all the major continents. In the process they came into contact with many populations that had been isolated for thousands of years. It is generally accepted that the Tasmanian aboriginals were the most isolated group on the planet. They were driven to extinction by European explorers, however a number of their descendants survive today as a result of admixture with Europeans. This is an example of how modern migrations have begun to reduce the genetic divergence of the human race.

The demographic composition of the old world has not changed significantly since the age of discovery. However new world demographics were radically changed within a short time following the voyage of Columbus. The colonization of the Americas brought Native Americans into contact with the distant populations of Europe, Africa, and Asia. As a result many countries in the Americas have significant and complex multiracial populations. Furthermore many who identify themselves by only one race still have multiracial ancestry.

Health

Differences in allele frequencies contribute to group differences in the incidence of some monogenic diseases, and they may contribute to differences in the incidence of some common diseases (Risch *et al.* 2002; Burchard *et al.* 2003; Tate and Goldstein 2004). For the monogenic diseases, the frequency of causative alleles usually correlates best with ancestry, whether familial (for example, Ellis-van Creveld syndrome among the Pennsylvania Amish), ethnic (Tay-Sachs disease among Ashkenazi Jewish populations), or geographical (hemoglobinopathies among people with ancestors who lived in malarial regions). To the extent that ancestry corresponds with racial or ethnic groups or subgroups, the incidence of monogenic diseases can differ between groups categorized by race or ethnicity, and health-care professionals typically take these patterns into account in making diagnoses.

Even with common diseases involving numerous genetic variants and environmental factors, investigators point to evidence suggesting the involvement of differentially distributed alleles with small to moderate effects. Frequently cited examples include hypertension (Douglas *et al.* 1996), diabetes (Gower *et al.* 2003), obesity (Fernandez *et al.* 2003), and prostate cancer (Platz *et al.* 2000). However, in none of these cases has allelic variation in a susceptibility gene been shown to account for a significant fraction of the difference in disease prevalence among groups, and the role of genetic factors in generating these differences remains uncertain (Mountain and Risch 2004).

Neil Risch of Stanford University has proposed that self-identified race/ethnic group could be a valid means of categorization in the USA for public health and policy considerations. While a 2002 paper by Noah Rosenberg's group makes a similar claim

"The structure of human populations is relevant in various epidemiological contexts. As a result of variation in frequencies of both genetic and nongenetic risk factors, rates of disease and of such phenotypes as adverse drug response vary across populations. Further, information about a patient's population of origin might provide health care practitioners with information about risk when direct causes of disease are unknown."

Genome projects

Human genome projects are scientific endeavors that determine or study the structure of the human genome. The Human Genome Project was a landmark genome project.

Chapter- 8

Personal Genomics

Personal genomics is a branch of genomics where individual genomes are genotyped and analyzed using bioinformatics tools. It is also related to traditional population genetics. The genotyping stage can have many different experimental approaches including single nucleotide polymorphism (SNP) chips (typically 0.02% of the genome), or partial or full genome sequencing. Once the genotypes are known, there are many bioinformatics analysis tools that can compare individual genomes and find disease association of the genes and loci. The most important aspect of personal genomics is that it may eventually lead to personalized medicine, where patients can take genotype specific drugs for medical treatments.

Personal genomics is not a single individual's vision or invention. Many researchers for decades anticipated this biological branch will eventually arrive with minimum cost of genotyping. Due to the advent of cheap and fast sequencers, full genome personal genomics is becoming a reality. However, there have been active early proponents of personal genomics projects such as George Church in Harvard Medical School.

Genomics used to mean academic research on consensus genomes which have been assembled from many different individuals of a particular species. The personal genomics changes this into customized bioinformatic discovery on individuals.

Use of personal genomics in predictive medicine

Predictive medicine is the use of the information produced by personal genomics techniques when deciding what medical treatments are appropriate for a particular individual.

An example of the use of predictive medicine is pharmacogenomics, in which genetic information can be used to select the most appropriate drug to prescribe to a patient. The drug should be chosen to maximize the probability of obtaining the desired result in the patient and minimize the probability that the patient will experience side effects. It is hoped that genetic information will allow physicians to tailor therapy to a given patient, in order to increase drug efficacy and minimize side effects. There are only a few

examples in which this information is currently useful in clinical practice, but it is anticipated that tailored therapy will emerge rapidly as researchers validate the clinical utility of different pharmacogenomic markers.

Another area in which there is great interest is disease risk prediction based on genetic markers. Researchers in this area have generated a great deal of information through the use of genome-wide association studies. While there is hope that risk information will be useful in providing predictive medicine, most common medical conditions are multifactorial and the actual risk to the individual depends on both genetic and environmental components, both of which are not completely understood at present. Therefore, the clinical utility of personal genomic information is currently limited. It is hoped that with further research, an accurate risk profile might enable individuals to take steps to prevent diseases for which they are at increased risk based on genetics.

Cost of sequencing an individual's genome

There is currently great interest in personal genomics. This is being fuelled by the rapid drop in the cost of sequencing a human genome. This drop in cost is due to the continual development of new, faster, cheaper DNA sequencing technologies such as "next generation DNA sequencing" that may provide access to full genome sequencing so that the entire genetic code of an individual can be deduced all at once.

The National Human Genome Research Institute, part of the U.S. National Institute of Health has set a target to be able to sequence a human-sized genome for US$100,000 by 2009 and US$1,000 by 2014. There is a widespread belief that within 10 years the cost of sequencing a human genome will fall to $1,000.

There are 6 billion base pairs in the diploid human genome. Statistical analysis reveals that a coverage of approximately ten times is required to get coverage of both alleles in 90% human genome from 25 base-pair reads with shotgun sequencing. This means a total of 60 billion base pairs that must be sequenced. An Applied Biosystems SOLiD, Illumina or Helicos sequencing machine can sequence 2 to 10 billion base pairs in each $8,000 to $18,000 run. The purchase cost, personnel costs and data processing costs must also be taken into account. Sequencing a human genome therefore costs approximately $300,000 in 2008.

In 2009, Complete Genomics of Mountain View announced that it would provide full genome sequencing for $5,000, from June 2009. This will only be available to institutions, not individuals.

This cost is still too high for governments to introduce programs into health services to sequence the genomes of all individuals in a country. However, it may be viable when it falls below $1,000, and the cost of sequencing a human genome is dropping rapidly. For example, approximately 1 million babies are born in Canada each year. To sequence all of their genomes would cost approximately $1 billion per year, or just 1% of Canada's total healthcare budget. Given the ethical concerns about presymptomatic genetic testing

of minors, it is likely that personal genomics will first be applied to adults who can provide consent to undergo such testing.

In June 2009, Illumina announced that they were launching their own Personal Full Genome Sequencing Service at a depth of 30X for $48,000 per genome.. Only one year later, in 2010, they cut the price 60% to $19,500. Still too expensive for true commercialization, prices are expected to drop further over the next few years as they realize economies of scale and given the competition with other companies such as Complete Genomics.

Knome's whole genome sequencing approach aims, instead, to read every site in the whole euchromatic portion of a person's genome (roughly 3 billion sites). While significantly more expensive than SNP chip-based genotyping, this approach yields significantly more data, identifying both novel (never-before-seen) and known sequence variants, some of which may be particularly relevant in efforts to understand personal health, as well as ancestry.

Timeline of Personal genomes sequenced

Year	Cost	Personal genomes sequenced	Company
2003	$3,000,000,000	1	Various
2009	$48,000	100	Illumina
2010	$19,500	?	Illumina

Comparative genomics

Comparative genomics analysis is concerned with characterising the differences and similarities between whole genomes. It may be applied to both genomes from individuals from different species or individuals from the same species, generally at lower cost than sequencing from scratch. In personal genomics and personalized medicine, we are concerned with comparing the genomes of different humans. It is likely that many of the techniques which are developed in comparative genomic analysis will be useful in personal genomics and personalized medicine. This includes rare and common Single nucleotide polymorphisms (consisting substituting one base pair by another, for example CATGCCGG to CATGACGG), as well as insertion or deletion of one or many base pairs.

Predictive medicine services already available

At least four companies which offer genome-wide personal genomics services already have gone to market and are selling their services direct to consumer. They are likely to be the first of many. However, the validity of individual risk predictions based on SNPs and the clinical utility of this information is currently questionable.

- deCODEme.com charges $2000 to carry out genotyping of approximately 1 million SNPs and provides risk estimates for 47 diseases as well as ancestry analyses.

- Navigenics, began offering SNP-based genomic risk assessments as of April 2008. Navigenics is medically focused and emphasizes a clinician's and genetic counselor's role in interpreting results. The Health Compass comprehensive genetic test for $999 analyzes your genetic predispositions for a variety of health conditions that meet stringent scientific criteria. Navigenics uses Affymetrix Genome-Wide Human SNP Array 6.0, which genotypes 900,000 SNPs.

- 23andMe sells mail order kits for SNP genotyping. The $199 kit, with $5.00/month subscription, or $499 without a subscription contains everything a consumer needs to take their own saliva sample. The consumer then mails the sample to 23andMe who carry out microarray analysis on it. This provides genotype information for approximately 1,000,000 SNPs. This information is used to estimate the genetic risk of the consumer for 178 diseases and conditions, as well as ancestry analyses.

- Bioresolve describes a similar service to that of 23andMe; however, the Better Business Bureau gave them an "F" reliability rating.

- Knome provides full genome (98% genome) sequencing services for $39,500 for whole genome sequencing and interpretation for consumers. It’s $29,500 for whole genome sequencing and analysis for researchers depending on their requirements.

- HelloGene and HelloGenome personal genome information services describe genotyping and full genome sequencing launched by Theragen in Korea. HelloGenome is the first commercial whole genome sequencing service in Asia while HelloGene is the first in Korea. HelloGene uses Affymetrix SNP chips while HelloGenome uses Solexa machines.

- Illumina, Oxford Nanopore Technologies, Sequenom, Pacific Biosciences, Complete Genomics and 454 Life Sciences are companies focused on commercializing full genome sequencing but are not involved in the predictive medicine (interpretative) side.

Ethical issues

While personalized medicine will certainly be a great asset to healthcare, it opens up several ethical issues which will need to be thought about carefully. No doubt there will be a huge amount of debate concerning the ethics of personalised medicine in the coming years.

Genetic discrimination is discriminating on the grounds of information obtained from an individual's genome. Genetic non-discrimination laws have been enacted in most US states and, at the federal level, by the Genetic Information Nondiscrimination Act (GINA). The GINA legislation prevents discrimination by health insurers and employers but does not apply to life insurance or long-term care insurance.

The likelihood of an individual developing breast cancer is affected by which alleles they have of particular genes. Screening can reveal breast cancer in the early stages, allowing it to be successfully treated. 50% of breast cancers occur in the 12% of the population who are at greatest risk. This poses a very difficult question for health services: Is it ethical to deny somebody free screening for a disease if they are genetically at low (but non-zero) risk of developing that disease?

Other issues

Medical genetics will confront the fact that full sequencing of the genome identifies many polymorphisms that are neutral or harmless. This prospect will create uncertainty in the analysis of individual genomes, particularly in the context of clinical care. Czech medical geneticist Eva Machácková writes: "In some cases it is difficult to distinguish if the detected sequence variant is a causal mutation or a neutral (polymorphic) variation without any effect on phenotype. The interpretation of rare sequence variants of unknown significance detected in disease-causing genes becomes an increasingly important problem."

Chapter- 9

DNA Sequencing

The term **DNA sequencing** refers to sequencing methods for determining the order of the nucleotide bases—adenine, guanine, cytosine, and thymine—in a molecule of DNA.

Knowledge of DNA sequences has become indispensable for basic biological research, other research branches utilizing DNA sequencing, and in numerous applied fields such as diagnostic, biotechnology, forensic biology and biological systematics. The advent of DNA sequencing has significantly accelerated biological research and discovery. The rapid speed of sequencing attained with modern DNA sequencing technology has been instrumental in the sequencing of the human genome, in the Human Genome Project. Related projects, often by scientific collaboration across continents, have generated the complete DNA sequences of many animal, plant, and microbial genomes.

The first DNA sequences were obtained in the early 1970s by academic researchers using laborious methods based on two-dimensional chromatography. Following the development of dye-based sequencing methods with automated analysis, DNA sequencing has become easier and orders of magnitude faster.

History

RNA sequencing was one of the earliest forms of nucleotide sequencing. The major landmark of RNA sequencing is the sequence of the first complete gene and the complete genome of Bacteriophage MS2, identified and published by Walter Fiers and his coworkers at the University of Ghent (Ghent, Belgium), between 1972 and 1976.

Prior to the development of rapid DNA sequencing methods in the early 1970s by Frederick Sanger at the University of Cambridge, in England and Walter Gilbert and Allan Maxam at Harvard, a number of laborious methods were used. For instance, in 1973, Gilbert and Maxam reported the sequence of 24 basepairs using a method known as wandering-spot analysis.

The chain-termination method developed by Sanger and coworkers in 1975 soon became the method of choice, owing to its relative ease and reliability.

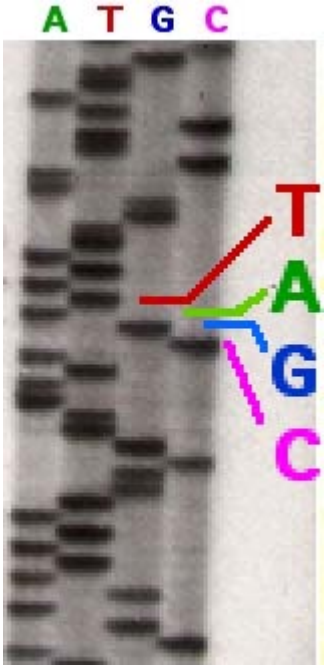
Maxam–Gilbert sequencing

In 1976–1977, Allan Maxam and Walter Gilbert developed a DNA sequencing method based on chemical modification of DNA and subsequent cleavage at specific bases. Although Maxam and Gilbert published their chemical sequencing method two years after the ground-breaking paper of Sanger and Coulson on plus-minus sequencing, Maxam–Gilbert sequencing rapidly became more popular, since purified DNA could be used directly, while the initial Sanger method required that each read start be cloned for production of single-stranded DNA. However, with the improvement of the chain-termination method (see below), Maxam-Gilbert sequencing has fallen out of favour due to its technical complexity prohibiting its use in standard molecular biology kits, extensive use of hazardous chemicals, and difficulties with scale-up.

The method requires radioactive labeling at one 5' end of the DNA (typically by a kinase reaction using gamma-^{32}P ATP) and purification of the DNA fragment to be sequenced. Chemical treatment generates breaks at a small proportion of one or two of the four nucleotide bases in each of four reactions (G, A+G, C, C+T). For example, the purines (A+G) are depurinated using formic acid, the guanines (and to some extent the adenines) are methylated by dimethyl sulfate, and the pyrimidines (C+T) are methylated using hydrazine. The addition of salt (sodium chloride) to the hydrazine reaction inhibits the methylation of thymine for the C-only reaction. The modified DNAs are then cleaved by hot piperidine at the position of the modified base. The concentration of the modifying chemicals is controlled to introduce on average one modification per DNA molecule. Thus a series of labeled fragments is generated, from the radiolabeled end to the first "cut" site in each molecule. The fragments in the four reactions are electrophoresed side by side in denaturing acrylamide gels for size separation. To visualize the fragments, the gel is exposed to X-ray film for autoradiography, yielding a series of dark bands each corresponding to a radiolabeled DNA fragment, from which the sequence may be inferred.

Also sometimes known as "chemical sequencing", this method originated in the study of DNA-protein interactions (DNase I footprinting) and nucleic acid structure, and within these it still has important applications.

Chain-termination methods



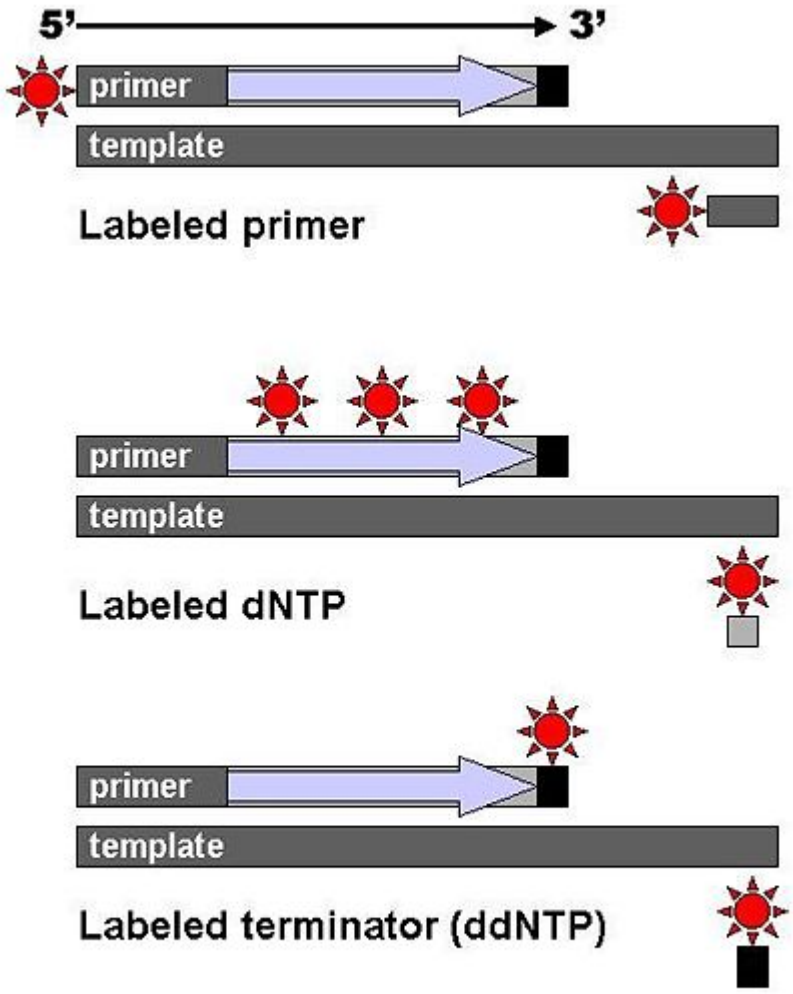
Part of a radioactively labelled sequencing gel

Because the chain-terminator method (or Sanger method after its developer Frederick Sanger) is more efficient and uses fewer toxic chemicals and lower amounts of radioactivity than the method of Maxam and Gilbert, it rapidly became the method of choice. The key principle of the Sanger method was the use of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators.

The classical chain-termination method requires a single-stranded DNA template, a DNA primer, a DNA polymerase, normal deoxynucleotidephosphates (dNTPs), and modified nucleotides (dideoxyNTPs) that terminate DNA strand elongation. These ddNTPs will also be radioactively or fluorescently labelled for detection in automated sequencing machines. The DNA sample is divided into four separate sequencing reactions, containing all four of the standard deoxynucleotides (dATP, dGTP, dCTP and dTTP) and the DNA polymerase. To each reaction is added only one of the four dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP) which are the chain-terminating nucleotides, lacking a 3'-OH group required for the formation of a phosphodiester bond between two nucleotides, thus terminating DNA strand extension and resulting in DNA fragments of varying length.

The newly synthesized and labelled DNA fragments are heat denatured, and separated by size (with a resolution of just one nucleotide) by gel electrophoresis on a denaturing polyacrylamide-urea gel with each of the four reactions run in one of four individual lanes (lanes A, T, G, C); the DNA bands are then visualized by autoradiography or UV light, and the DNA sequence can be directly read off the X-ray film or gel image. In the image on the right, X-ray film was exposed to the gel, and the dark bands correspond to

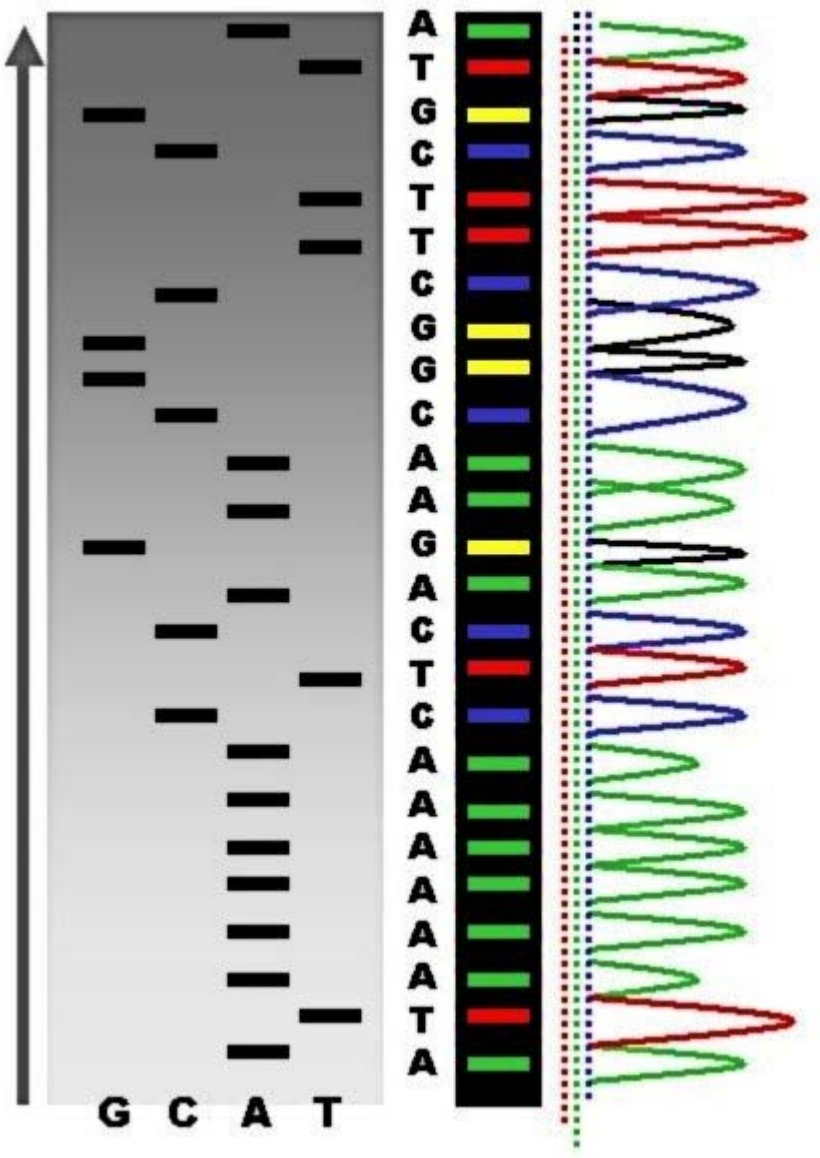
DNA fragments of different lengths. A dark band in a lane indicates a DNA fragment that is the result of chain termination after incorporation of a dideoxynucleotide (ddATP, ddGTP, ddCTP, or ddTTP). The relative positions of the different bands among the four lanes are then used to read (from bottom to top) the DNA sequence.



DNA fragments are labelled with a radioactive or fluorescent tag on the primer (1), in the new DNA strand with a labeled dNTP, or with a labeled ddNTP.

Technical variations of chain-termination sequencing include tagging with nucleotides containing radioactive phosphorus for radiolabelling, or using a primer labeled at the 5' end with a fluorescent dye. Dye-primer sequencing facilitates reading in an optical system for faster and more economical analysis and automation. The later development

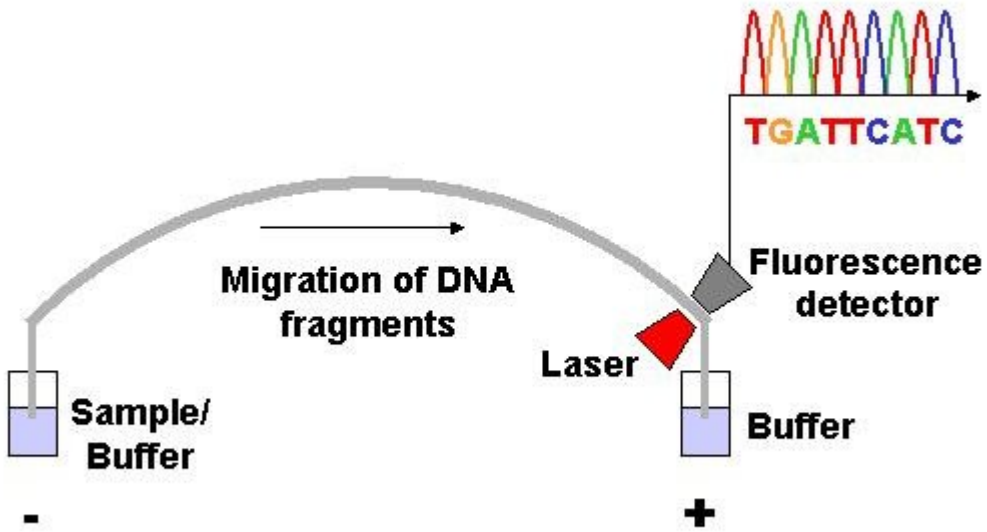
by Leroy Hood and coworkers of fluorescently labeled ddNTPs and primers set the stage for automated, high-throughput DNA sequencing.



Sequence ladder by radioactive sequencing compared to fluorescent peaks

Chain-termination methods have greatly simplified DNA sequencing. For example, chain-termination-based kits are commercially available that contain the reagents needed for sequencing, pre-aliquoted and ready to use. Limitations include non-specific binding of the primer to the DNA, affecting accurate read-out of the DNA sequence, and DNA secondary structures affecting the fidelity of the sequence.

Dye-terminator sequencing



Capillary electrophoresis

Dye-terminator sequencing utilizes labelling of the chain terminator ddNTPs, which permits sequencing in a single reaction, rather than four reactions as in the labelled-primer method. In dye-terminator sequencing, each of the four dideoxynucleotide chain terminators is labelled with fluorescent dyes, each of which emit light at different wavelengths.

Owing to its greater expediency and speed, dye-terminator sequencing is now the mainstay in automated sequencing. Its limitations include dye effects due to differences in the incorporation of the dye-labelled chain terminators into the DNA fragment, resulting in unequal peak heights and shapes in the electronic DNA sequence trace chromatogram after capillary electrophoresis.

This problem has been addressed with the use of modified DNA polymerase enzyme systems and dyes that minimize incorporation variability, as well as methods for eliminating "dye blobs". The dye-terminator sequencing method, along with automated high-throughput DNA sequence analyzers, is now being used for the vast majority of sequencing projects.

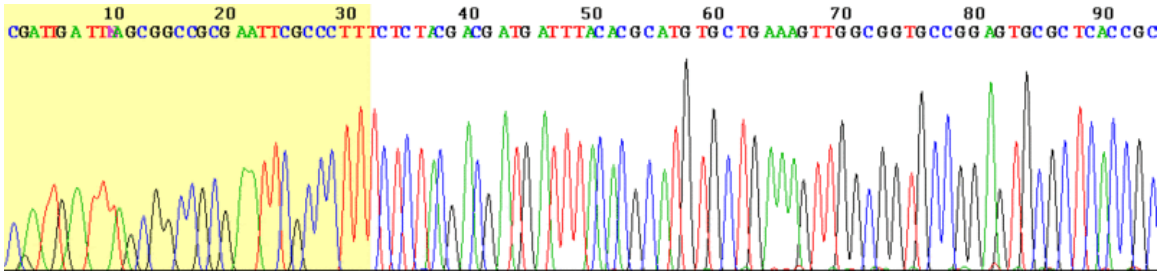
Challenges

Common challenges of DNA sequencing include poor quality in the first 15–40 bases of the sequence and deteriorating quality of sequencing traces after 700–900 bases. Base calling software typically gives an estimate of quality to aid in quality trimming.

In cases where DNA fragments are cloned before sequencing, the resulting sequence may contain parts of the cloning vector. In contrast, PCR-based cloning and emerging sequencing technologies based on pyrosequencing often avoid using cloning vectors. Recently, one-step Sanger sequencing (combined amplification and sequencing) methods such as Ampliseq and SeqSharp have been developed that allow rapid sequencing of target genes without cloning or prior amplification.

Current methods can directly sequence only relatively short (300–1000 nucleotides long) DNA fragments in a single reaction. The main obstacle to sequencing DNA fragments above this size limit is insufficient power of separation for resolving large DNA fragments that differ in length by only one nucleotide. In all cases the use of a primer with a free 5' end is essential.

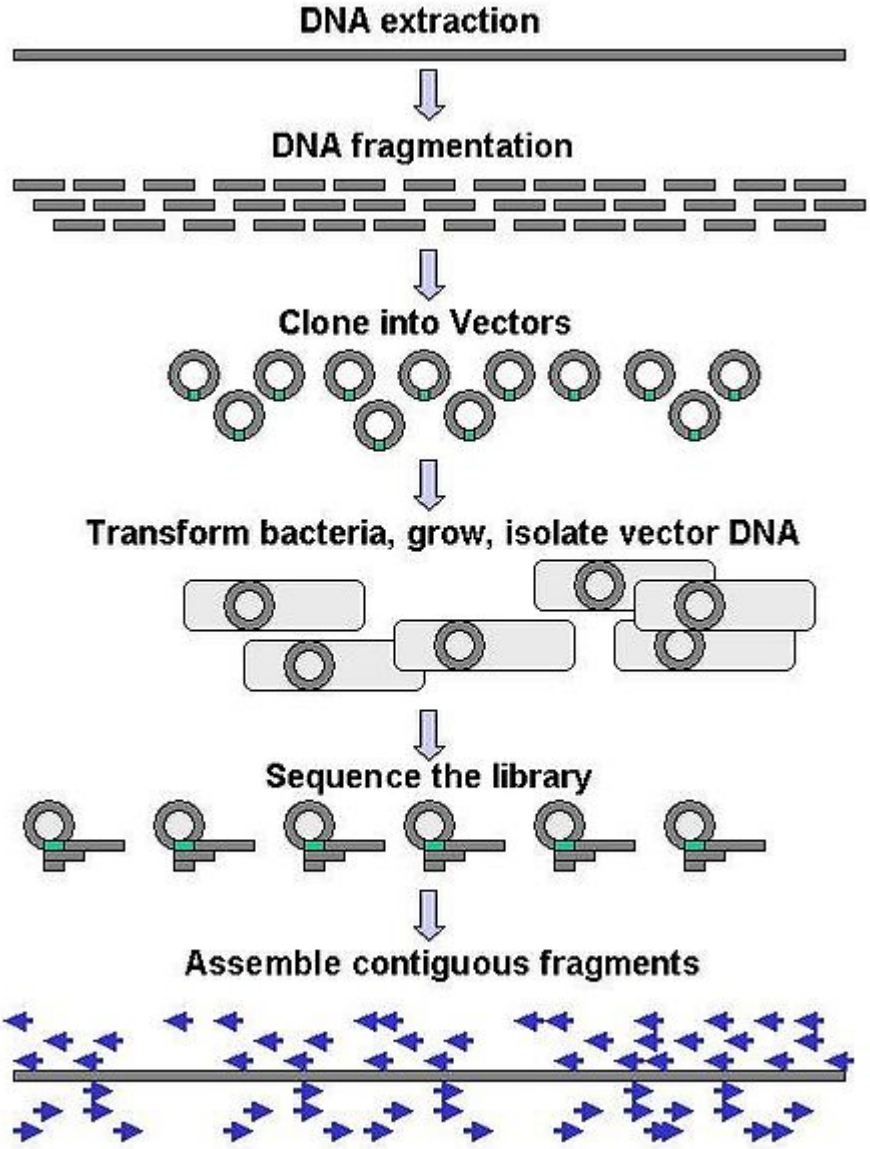
Automation and sample preparation



View of the start of an example dye-terminator read

Automated DNA-sequencing instruments (DNA sequencers) can sequence up to 384 DNA samples in a single batch (run) in up to 24 runs a day. DNA sequencers carry out capillary electrophoresis for size separation, detection and recording of dye fluorescence, and data output as fluorescent peak trace chromatograms. Sequencing reactions by thermocycling, cleanup and re-suspension in a buffer solution before loading onto the sequencer are performed separately. A number of commercial and non-commercial software packages can trim low-quality DNA traces automatically. These programs score the quality of each peak and remove low-quality base peaks (generally located at the ends of the sequence). The accuracy of such algorithms is below visual examination by a human operator, but sufficient for automated processing of large sequence data sets.

Amplification and clonal selection



Genomic DNA is fragmented into random pieces and cloned as a bacterial library. DNA from individual bacterial clones is sequenced and the sequence is assembled by using overlapping DNA regions.

Large-scale sequencing aims at sequencing very long DNA pieces, such as whole chromosomes. Common approaches consist of cutting (with restriction enzymes) or shearing (with mechanical forces) large DNA fragments into shorter DNA fragments. The fragmented DNA is cloned into a DNA vector, and amplified in *Escherichia coli*. Short DNA fragments purified from individual bacterial colonies are individually sequenced and assembled electronically into one long, contiguous sequence.

This method does not require any pre-existing information about the sequence of the DNA and is referred to as *de novo* sequencing. Gaps in the assembled sequence may be filled by primer walking. The different strategies have different tradeoffs in speed and accuracy; *shotgun methods* are often used for sequencing large genomes, but its assembly is complex and difficult, particularly with sequence repeats often causing gaps in genome assembly.

Most sequencing approaches use an *in vitro* cloning step to amplify individual DNA molecules, because their molecular detection methods are not sensitive enough for single molecule sequencing. Emulsion PCR isolates individual DNA molecules along with primer-coated beads in aqueous droplets within an oil phase. Polymerase chain reaction (PCR) then coats each bead with clonal copies of the DNA molecule followed by immobilization for later sequencing. Emulsion PCR is used in the methods by Marguilis et al. (commercialized by 454 Life Sciences), Shendure and Porreca et al. (also known as "Polony sequencing") and SOLiD sequencing, (developed by Agencourt, now Applied Biosystems).

Another method for *in vitro* clonal amplification is *bridge PCR*, where fragments are amplified upon primers attached to a solid surface, used in the Illumina Genome Analyzer. The single-molecule method developed by Stephen Quake's laboratory (later commercialized by Helicos) is an exception: it uses bright fluorophores and laser excitation to detect pyrosequencing events from individual DNA molecules fixed to a surface, eliminating the need for molecular amplification.

High-throughput sequencing

The high demand for low-cost sequencing has driven the development of high-throughput sequencing technologies that parallelize the sequencing process, producing thousands or millions of sequences at once. High-throughput sequencing technologies are intended to lower the cost of DNA sequencing beyond what is possible with standard dye-terminator methods.

Lynx Therapeutics' Massively Parallel Signature Sequencing (MPSS)

The first of the "next-generation" sequencing technologies, MPSS was developed in 1990s at Lynx Therapeutics, a company founded in 1992 by Sidney Brenner and Sam Eletr. MPSS was a bead-based method that used a complex approach of adapter ligation followed by adapter decoding, reading the sequence in increments of four nucleotides; this method made it susceptible to sequence-specific bias or loss of specific sequences. Because the technology was so complex, MPSS was only performed 'in-house' by Lynx Therapeutics and no machines were sold; when the merger with Solexa later lead to the development of sequencing-by-synthesis, a more simple approach with numerous advantages, MPSS became obsolete. However, the essential properties of the MPSS output were typical of later "next-gen" data types, including hundreds of thousands of short DNA sequences. In the case of MPSS, these were typically used for sequencing

cDNA for measurements of gene expression levels. Lynx Therapeutics merged with Solexa in 2004, and this company was later purchased by Illumina.

Polony Sequencing

Polony sequencing, developed in George Church's lab at Harvard, was among the first next-generation sequencing systems used to sequence a full genome in 2005. It combined an in vitro paired-tag library with emulsion PCR, an automated microscope, and ligation-based sequencing chemistry to sequence an E. coli genome at an accuracy of > 99.9999% and a cost approximately 1/10th that of Sanger sequencing. The technology was licensed to Agencourt Biosciences, subsequently spun out into Agencourt Personal Genomics, and ultimately incorporated into the Applied Biosystems SOLiD platform.

454 pyrosequencing

A parallelized version of pyrosequencing was developed by 454 Life Sciences. The method amplifies DNA inside water droplets in an oil solution (emulsion PCR), with each droplet containing a single DNA template attached to a single primer-coated bead that then forms a clonal colony. The sequencing machine contains many picolitre-volume wells each containing a single bead and sequencing enzymes. Pyrosequencing uses luciferase to generate light for detection of the individual nucleotides added to the nascent DNA, and the combined data are used to generate sequence read-outs. This technology provides intermediate read length and price per base compared to Sanger sequencing on one end and Solexa and SOLiD on the other. 454 Life Sciences has since been acquired by Roche Diagnostics.

Illumina (Solexa) sequencing

Solexa, now part of Illumina developed a sequencing technology based on reversible dye-terminators. DNA molecules are first attached to primers on a slide and amplified so that local clonal colonies are formed (bridge amplification). Four types of ddNTPs are added, and non-incorporated nucleotides are washed away. Unlike pyrosequencing, the DNA can only be extended one nucleotide at a time. A camera takes images of the fluorescently labeled nucleotides then the dye along with the terminal 3' blocker is chemically removed from the DNA, allowing a next cycle.

SOLiD sequencing

Applied Biosystems' SOLiD technology employs sequencing by ligation. Here, a pool of all possible oligonucleotides of a fixed length are labeled according to the sequenced position. Oligonucleotides are annealed and ligated; the preferential ligation by DNA ligase for matching sequences results in a signal informative of the nucleotide at that position. Before sequencing, the DNA is amplified by emulsion PCR. The resulting bead, each containing only copies of the same DNA molecule, are deposited on a glass slide. The result is sequences of quantities and lengths comparable to Illumina sequencing.

Future methods

Sequencing by hybridization is a non-enzymatic method that uses a DNA microarray. A single pool of DNA whose sequence is to be determined is fluorescently labeled and hybridized to an array containing known sequences. Strong hybridization signals from a given spot on the array identifies its sequence in the DNA being sequenced. Mass spectrometry may be used to determine mass differences between DNA fragments produced in chain-termination reactions.

DNA sequencing methods currently under development include labeling the DNA polymerase, reading the sequence as a DNA strand transits through nanopores, and microscopy-based techniques, such as AFM or electron microscopy that are used to identify the positions of individual nucleotides within long DNA fragments (>5,000 bp) by nucleotide labeling with heavier elements (e.g., halogens) for visual detection and recording.

In microfluidic Sanger sequencing the entire thermocycling amplification of DNA fragments as well as their separation by electrophoresis is done on a single glass wafer (approximately 10 cm in diameter) thus reducing the reagent usage as well as cost. In some instances researchers have shown that they can increase the throughput of conventional sequencing through the use of microchips. Research will still need to be done in order to make this use of technology effective.

In October 2006, the X Prize Foundation established an initiative to promote the development of full genome sequencing technologies, called the Archon X Prize, intending to award $10 million to "the first Team that can build a device and use it to sequence 100 human genomes within 10 days or less, with an accuracy of no more than one error in every 100,000 bases sequenced, with sequences accurately covering at least 98% of the genome, and at a recurring cost of no more than $10,000 (US) per genome."

Each year NHGRI promotes grants for new research and developments in genomics. 2010 grants and 2011 candidates include continuing work in microfluidic, polony and base-heavy sequencing methodologies

Major landmarks in DNA sequencing

- 1953 Discovery of the structure of the DNA double helix.
- 1972 Development of recombinant DNA technology, which permits isolation of defined fragments of DNA; prior to this, the only accessible samples for sequencing were from bacteriophage or virus DNA.
- 1977 The first complete DNA genome to be sequenced is that of bacteriophage φX174.

- 1977 Allan Maxam and Walter Gilbert publish "DNA sequencing by chemical degradation". Frederick Sanger, independently, publishes "DNA sequencing with chain-terminating inhibitors".

- 1984 Medical Research Council scientists decipher the complete DNA sequence of the Epstein-Barr virus, 170 kb.

- 1986 Leroy E. Hood's laboratory at the California Institute of Technology and Smith announce the first semi-automated DNA sequencing machine.

- 1987 Applied Biosystems markets first automated sequencing machine, the model ABI 370.

- 1990 The U.S. National Institutes of Health (NIH) begins large-scale sequencing trials on *Mycoplasma capricolum*, *Escherichia coli*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae* (at US$0.75/base).

- 1991 Sequencing of human expressed sequence tags begins in Craig Venter's lab, an attempt to capture the coding fraction of the human genome.

- 1995 Craig Venter, Hamilton Smith, and colleagues at The Institute for Genomic Research (TIGR) publish the first complete genome of a free-living organism, the bacterium Haemophilus influenzae. The circular chromosome contains 1,830,137 bases and its publication in the journal Science marks the first use of whole-genome shotgun sequencing, eliminating the need for initial mapping efforts.

- 1996 Pål Nyrén and his student Mostafa Ronaghi at the Royal Institute of Technology in Stockholm publish their method of pyrosequencing

- 1998 Phil Green and Brent Ewing of the University of Washington publish "phred" for sequencer data analysis.

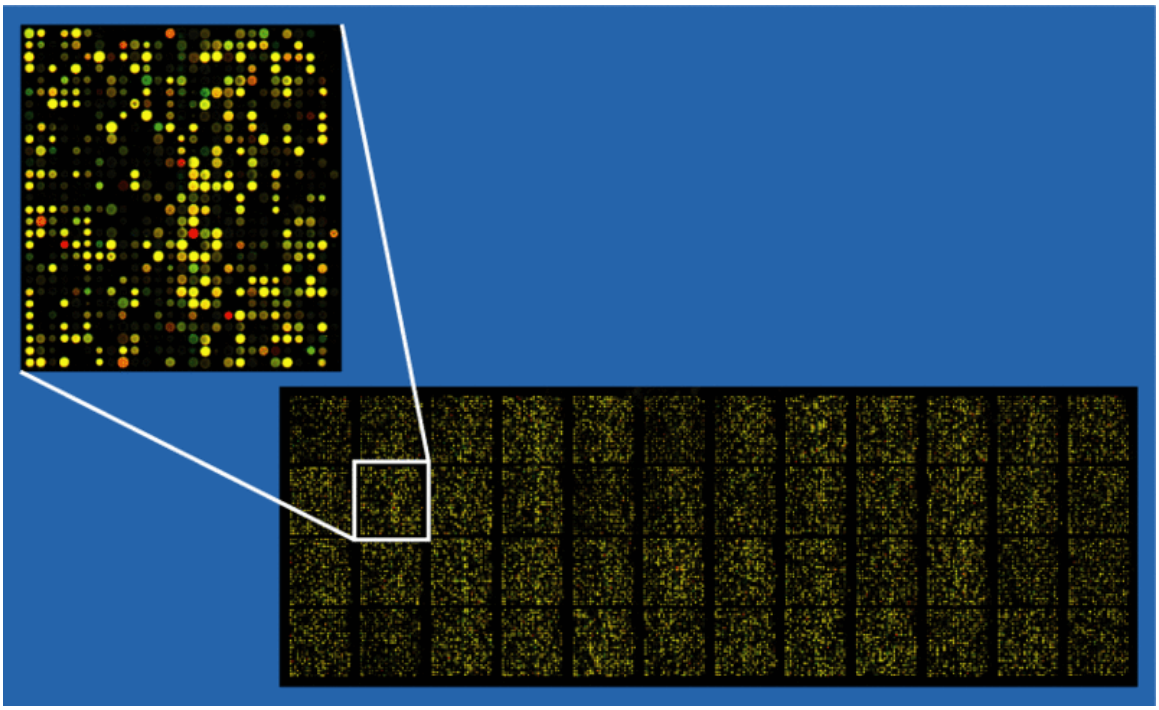
- 2000 Lynx Therapeutics publishes and markets "MPSS" - a parallelized, adapter/ligation-mediated, bead-based sequencing technology, launching "next-generation" sequencing.

- 2001 A draft sequence of the human genome is published.

- 2004 454 Life Sciences markets a parallelized version of pyrosequencing. The first version of their machine reduced sequencing costs 6-fold compared to automated Sanger sequencing, and was the second of a new generation of sequencing technologies, after MPSS.

Chapter- 10

DNA Microarray



Example of an approximately 40,000 probe spotted oligo microarray with enlarged inset to show detail.

A **DNA microarray** is a multiplex technology used in molecular biology. It consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides, called features, each containing picomoles (10^{-12} moles) of a specific DNA sequence, known as *probes* (or *reporters*). These 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 many genetic tests in parallel. Therefore arrays have dramatically accelerated many types of investigation.

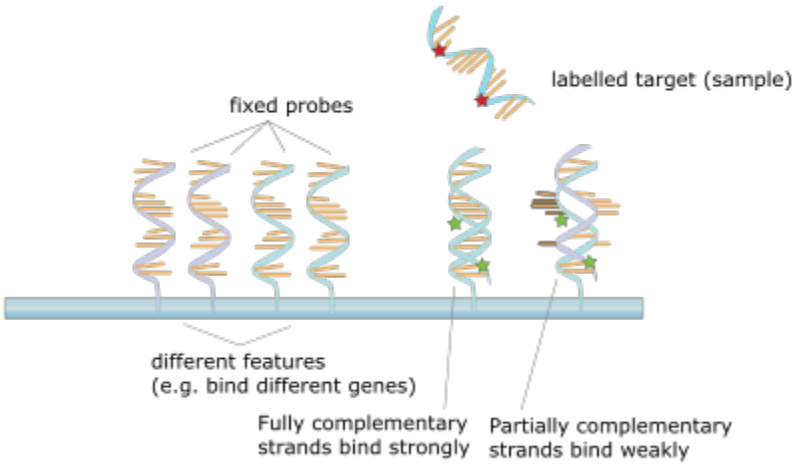
In standard microarrays, the probes are attached via surface engineering to a solid surface by a covalent bond to a chemical matrix (via epoxy-silane, amino-silane, lysine, polyacrylamide or others). The solid surface can be glass or a silicon chip, in which case they are colloquially known as an *Affy chip* when an Affymetrix chip is used. Other microarray platforms, such as Illumina, use microscopic beads, instead of the large solid support. DNA arrays are different from other types of microarray only in that they either measure DNA or use DNA as part of its detection system.

DNA microarrays can be used to measure changes in expression levels, to detect single nucleotide polymorphisms (SNPs), or to genotype or resequence mutant genomes. Microarrays also differ in fabrication, workings, accuracy, efficiency, and cost. Additional factors for microarray experiments are the experimental design and the methods of analyzing the data.

History

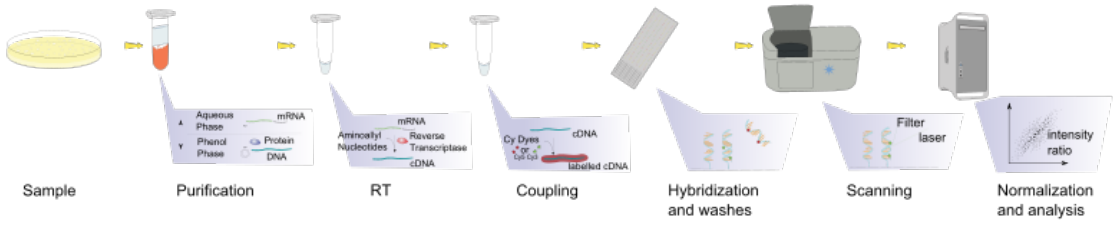
Microarray technology evolved from Southern blotting, where fragmented DNA is attached to a substrate and then probed with a known gene or fragment. Nucleic Acids Res. 1992 Apr 11;20(7):1679-84. Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ. Maskos U, Southern EM. The first reported use of this approach was the analysis of 378 arrayed lysed bacterial colonies each harboring a different sequence which were assayed in multiple replicas for expression of the genes in multiple normal and tumor tissue (Augenlicht and Kobrin, Cancer Research, 42, 1088–1093, 1982). This was expanded to analysis of more than 4000 human sequences with computer driven scanning and image processing for quantitative analysis of the sequences in human colonic tumors and normal tissue (Augenlicht *et al.*, Cancer Research, 47, 6017-6021, 1987) and then to comparison of colonic tissues at different genetic risk (Augenlicht *et al.*, Proceedings National Academy of Sciences, USA, 88, 3286-3289, 1991). The use of a collection of distinct DNAs in arrays for expression profiling was also described in 1987, and the arrayed DNAs were used to identify genes whose expression is modulated by interferon. These early gene arrays were made by spotting cDNAs onto filter paper with a pin-spotting device. The use of miniaturized microarrays for gene expression profiling was first reported in 1995, and a complete eukaryotic genome (*Saccharomyces cerevisiae*) on a microarray was published in 1997.

Principle



hybridization of the target to the probe

The core principle behind microarrays is hybridization between two DNA strands, the property of complementary nucleic acid sequences to specifically pair with each other by forming hydrogen bonds between complementary nucleotide base pairs. A high number of complementary base pairs in a nucleotide sequence means tighter non-covalent bonding between the two strands. After washing off of non-specific bonding sequences, only strongly paired strands will remain hybridized. So fluorescently labeled target sequences that bind to a probe sequence generate a signal that depends on the strength of the hybridization determined by the number of paired bases, the hybridization conditions (such as temperature), and washing after hybridization. Total strength of the signal, from a spot (feature), depends upon the amount of target sample binding to the probes present on that spot. Microarrays use relative quantization in which the intensity of a feature is compared to the intensity of the same feature under a different condition, and the identity of the feature is known by its position. An alternative to microarrays is serial analysis of gene expression, where the transcriptome is sequenced allowing an absolute measurement.



The step required in a microarray experiment

Uses and types

Two Affymetrix chips

Many types of array exist and the broadest distinction is whether they are spatially arranged on a surface or on coded beads:

- The traditional solid-phase array is a collection of orderly microscopic "spots", called features, each with a specific probe attached to a solid surface, such as glass, plastic or silicon biochip (commonly known as a *genome chip*, *DNA chip* or *gene array*). Thousands of them can be placed in known locations on a single DNA microarray.
- The alternative bead array is a collection of microscopic polystyrene beads, each with a specific probe and a ratio of two or more dyes, which do not interfere with the fluorescent dyes used on the target sequence.

DNA microarrays can be used to detect DNA (as in comparative genomic hybridization), or detect RNA (most commonly as cDNA after reverse transcription) that may or may not

be translated into proteins. The process of measuring gene expression via cDNA is called expression analysis or expression profiling.
Applications include:

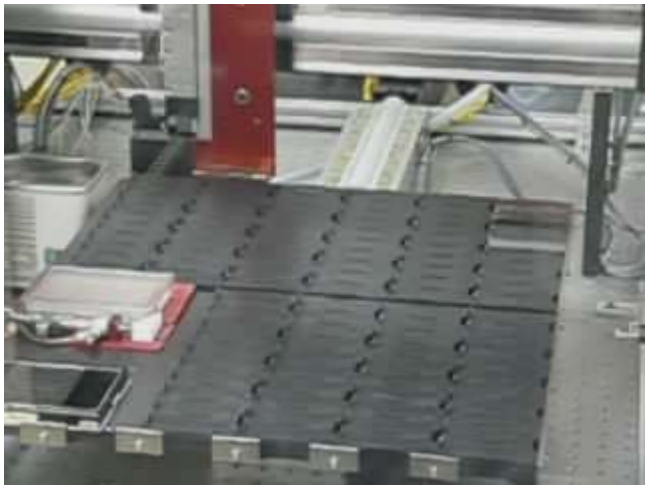
Application or technology	Synopsis
Gene expression profiling	In an mRNA or gene expression profiling experiment the expression levels of thousands of genes are simultaneously monitored to study the effects of certain treatments, diseases, and developmental stages on gene expression. For example, microarray-based gene expression profiling can be used to identify genes whose expression is changed in response to pathogens or other organisms by comparing gene expression in infected to that in uninfected cells or tissues.
Comparative genomic hybridization	Assessing genome content in different cells or closely related organisms.
GeneID	Small microarrays to check IDs of organisms in food and feed (like GMO), mycoplasms in cell culture, or pathogens for disease detection, mostly combining PCR and microarray technology.
Chromatin immunoprecipitation on Chip	DNA sequences bound to a particular protein can be isolated by immunoprecipitating that protein (ChIP), these fragments can be then hybridized to a microarray (such as a tiling array) allowing the determination of protein binding site occupancy throughout the genome. Example protein to immunoprecipitate are histone modifications (H3K27me3, H3K4me2, H3K9me3, etc.), Polycomb-group protein (PRC2:Suz12, PRC1:YY1) and trithorax-group protein (Ash1) to study the epigenetic landscape or RNA Polymerase II to study the transcription landscape.
DamID	Analogously to ChIP, genomic regions bound by a protein of interest can be isolated and used to probe a microarray to determine binding site occupancy. Unlike ChIP, DamID does not require antibodies but makes use of adenine methylation near the protein's binding sites to selectively amplify those regions, introduced by expressing minute amounts of protein of interest fused to bacterial DNA adenine methyltransferase.
SNP detection	Identifying single nucleotide polymorphism among alleles within or between populations. Several applications of microarrays make use of SNP detection, including Genotyping, forensic analysis, measuring predisposition to disease, identifying drug-candidates, evaluating germline mutations in individuals or somatic mutations in cancers, assessing loss of heterozygosity, or genetic linkage analysis.
Alternative splicing detection	An '*exon junction array* design uses probes specific to the expected or potential splice sites of predicted exons for a gene. It

	is of intermediate density, or coverage, to a typical gene expression array (with 1-3 probes per gene) and a genomic tiling array (with hundreds or thousands of probes per gene). It is used to assay the expression of alternative splice forms of a gene. Exon arrays have a different design, employing probes designed to detect each individual exon for known or predicted genes, and can be used for detecting different splicing isoforms.
Fusion genes microarray	A Fusion gene microarray can detect fusion transcripts, *e.g.* from cancer specimens. The principle behind this is building on the alternative splicing microarrays. The oligo design strategy enables combined measurements of chimeric transcript junctions with exon-wise measurements of individual fusion partners.
Tiling array	Genome tiling arrays consist of overlapping probes designed to densely represent a genomic region of interest, sometimes as large as an entire human chromosome. The purpose is to empirically detect expression of transcripts or alternatively splice forms which may not have been previously known or predicted.

Fabrication

Microarrays can be manufactured in different ways, depending on the number of probes under examination, costs, customization requirements, and the type of scientific question being asked. Arrays may have as few as 10 probes or up to 2.1 million micrometre-scale probes from commercial vendors.

Spotted vs. in situ synthesised arrays



A DNA microarray being printed by a robot at the University of Delaware

Microarrays can be fabricated using a variety of technologies, including printing with fine-pointed pins onto glass slides, photolithography using pre-made masks,

photolithography using dynamic micromirror devices, ink-jet printing, or electrochemistry on microelectrode arrays.

In *spotted microarrays*, the probes are oligonucleotides, cDNA or small fragments of PCR products that correspond to mRNAs. The probes are synthesized prior to deposition on the array surface and are then "spotted" onto glass. A common approach utilizes an array of fine pins or needles controlled by a robotic arm that is dipped into wells containing DNA probes and then depositing each probe at designated locations on the array surface. The resulting "grid" of probes represents the nucleic acid profiles of the prepared probes and is ready to receive complementary cDNA or cRNA "targets" derived from experimental or clinical samples. This technique is used by research scientists around the world to produce "in-house" printed microarrays from their own labs. These arrays may be easily customized for each experiment, because researchers can choose the probes and printing locations on the arrays, synthesize the probes in their own lab (or collaborating facility), and spot the arrays. They can then generate their own labeled samples for hybridization, hybridize the samples to the array, and finally scan the arrays with their own equipment. This provides a relatively low-cost microarray that may be customized for each study, and avoids the costs of purchasing often more expensive commercial arrays that may represent vast numbers of genes that are not of interest to the investigator. Publications exist which indicate in-house spotted microarrays may not provide the same level of sensitivity compared to commercial oligonucleotide arrays, possibly owing to the small batch sizes and reduced printing efficiencies when compared to industrial manufactures of oligo arrays.

In *oligonucleotide microarrays*, the probes are short sequences designed to match parts of the sequence of known or predicted open reading frames. Although oligonucleotide probes are often used in "spotted" microarrays, the term "oligonucleotide array" most often refers to a specific technique of manufacturing. Oligonucleotide arrays are produced by printing short oligonucleotide sequences designed to represent a single gene or family of gene splice-variants by synthesizing this sequence directly onto the array surface instead of depositing intact sequences. Sequences may be longer (60-mer probes such as the Agilent design) or shorter (25-mer probes produced by Affymetrix) depending on the desired purpose; longer probes are more specific to individual target genes, shorter probes may be spotted in higher density across the array and are cheaper to manufacture. One technique used to produce oligonucleotide arrays include photolithographic synthesis (Agilent and Affymetrix) on a silica substrate where light and light-sensitive masking agents are used to "build" a sequence one nucleotide at a time across the entire array. Each applicable probe is selectively "unmasked" prior to bathing the array in a solution of a single nucleotide, then a masking reaction takes place and the next set of probes are unmasked in preparation for a different nucleotide exposure. After many repetitions, the sequences of every probe become fully constructed. More recently, Maskless Array Synthesis from NimbleGen Systems has combined flexibility with large numbers of probes.

Two-channel vs. one-channel detection

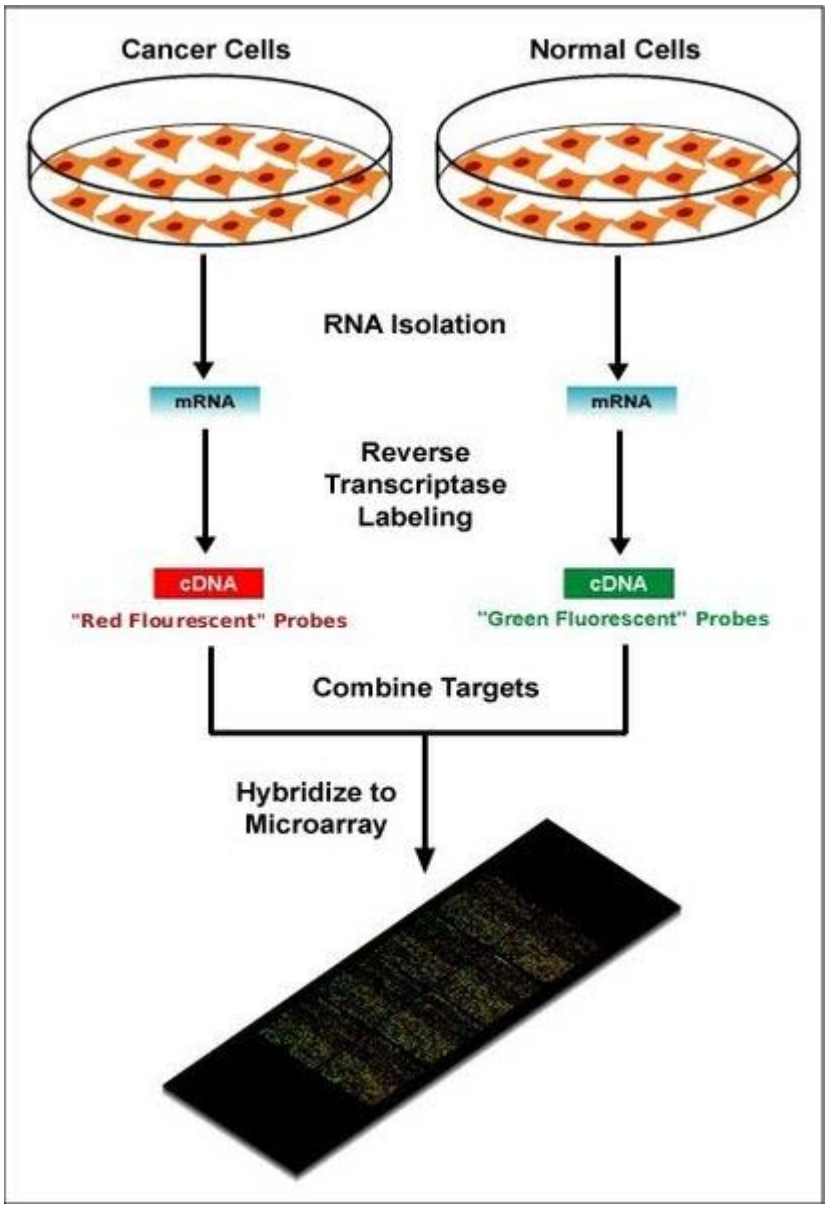


Diagram of typical dual-colour microarray experiment

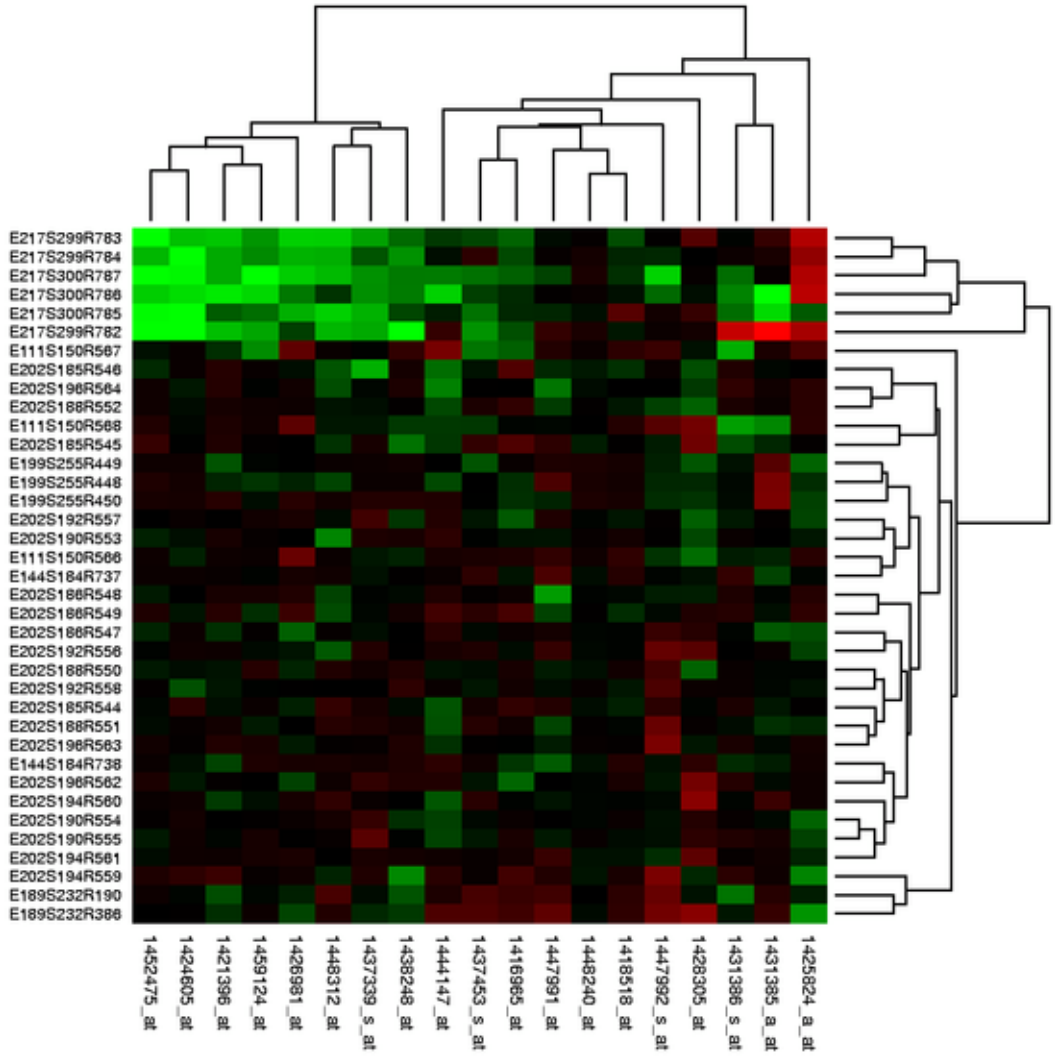
Two-color microarrays or *two-channel microarrays* are typically hybridized with cDNA prepared from two samples to be compared (e.g. diseased tissue versus healthy tissue) and that are labeled with two different fluorophores. Fluorescent dyes commonly used for cDNA labeling include Cy3, which has a fluorescence emission wavelength of 570 nm (corresponding to the green part of the light spectrum), and Cy5 with a fluorescence emission wavelength of 670 nm (corresponding to the red part of the light spectrum). The two Cy-labeled cDNA samples are mixed and hybridized to a single microarray that is then scanned in a microarray scanner to visualize fluorescence of the two fluorophores after excitation with a laser beam of a defined wavelength. Relative intensities of each

fluorophore may then be used in ratio-based analysis to identify up-regulated and down-regulated genes.

Oligonucleotide microarrays often carry control probes designed to hybridize with RNA spike-ins. The degree of hybridization between the spike-ins and the control probes is used to normalize the hybridization measurements for the target probes. Although absolute levels of gene expression may be determined in the two-color array in rare instances, the relative differences in expression among different spots within a sample and between samples is the preferred method of data analysis for the two-color system. Examples of providers for such microarrays includes Agilent with their Dual-Mode platform, Eppendorf with their DualChip platform for colorimetric Silverquant labeling, and TeleChem International with Arrayit.

In *single-channel microarrays* or *one-color microarrays*, the arrays provide intensity data for each probe or probe set indicating a relative level of hybridization with the labeled target. However, they do not truly indicate abundance levels of a gene but rather relative abundance when compared to other samples or conditions when processed in the same experiment. Each RNA molecule encounters protocol and batch-specific bias during amplification, labeling, and hybridization phases of the experiment making comparisons between genes for the same microarray uninformative. The comparison of two conditions for the same gene requires two separate single-dye hybridizations. Several popular single-channel systems are the Affymetrix "Gene Chip", Illumina "Bead Chip", Agilent single-channel arrays, the Applied Microarrays "CodeLink" arrays, and the Eppendorf "DualChip & Silverquant". One strength of the single-dye system lies in the fact that an aberrant sample cannot affect the raw data derived from other samples, because each array chip is exposed to only one sample (as opposed to a two-color system in which a single low-quality sample may drastically impinge on overall data precision even if the other sample was of high quality). Another benefit is that data are more easily compared to arrays from different experiments so long as batch effects have been accounted for. A drawback to the one-color system is that, when compared to the two-color system, twice as many microarrays are needed to compare samples within an experiment.

Microarrays and bioinformatics



Gene expression values from microarray experiments can be represented as heat maps to visualize the result of data analysis.

The advent of inexpensive microarray experiments created several specific bioinformatics challenges:

- the multiple levels of replication in experimental design (Experimental design)
- the number of platforms and independent groups and data format (Standardization)
- the treatment of the data (Statistical analysis)
- accuracy and precision (Relation between probe and gene)
- the sheer volume of data and the ability to share it (Data warehousing)

Experimental design

Due to the biological complexity of gene expression, the considerations of experimental design that are discussed in the expression profiling article are of critical importance if statistically and biologically valid conclusions are to be drawn from the data.

There are three main elements to consider when designing a microarray experiment. First, replication of the biological samples is essential for drawing conclusions from the experiment. Second, technical replicates (two RNA samples obtained from each experimental unit) help to ensure precision and allow for testing differences within treatment groups. The technical replicates may be two independent RNA extractions or two aliquots of the same extraction. Third, spots of each cDNA clone or oligonucleotide are present as replicates (at least duplicates) on the microarray slide, to provide a measure of technical precision in each hybridization. It is critical that information about the sample preparation and handling is discussed, in order to help identify the independent units in the experiment and to avoid inflated estimates of statistical significance.

Standardization

Microarray data is difficult to exchange due to the lack of standardization in platform fabrication, assay protocols, and analysis methods. This presents an interoperability problem in bioinformatics. Various grass-roots open-source projects are trying to ease the exchange and analysis of data produced with non-proprietary chips:

- For example, the "Minimum Information About a Microarray Experiment" (MIAME) checklist helps define the level of detail that should exist and is being adopted by many journals as a requirement for the submission of papers incorporating microarray results. But MIAME does not describe the format for the information, so while many formats can support the MIAME requirements, as of 2007 no format permits verification of complete semantic compliance.
- The "MicroArray Quality Control (MAQC) Project" is being conducted by the US Food and Drug Administration (FDA) to develop standards and quality control metrics which will eventually allow the use of MicroArray data in drug discovery, clinical practice and regulatory decision-making.
- The MGED Society has developed standards for the representation of gene expression experiment results and relevant annotations.

Statistical analysis

Microarray data sets are commonly very large, and analytical precision is influenced by a number of variables. Statistical challenges include taking into account effects of background noise and appropriate normalization of the data. Normalization methods may be suited to specific platforms and, in the case of commercial platforms, the analysis may be proprietary. Algorithms that affect statistical analysis include:

- Image analysis: gridding, spot recognition of the scanned image (segmentation algorithm), removal or marking of poor-quality and low-intensity features (called *flagging*).
- Data processing: background subtraction (based on global or local background), determination of spot intensities and intensity ratios, visualisation of data, and log-transformation of ratios, global or local normalization of intensity ratios.
- Identification of statistically significant changes: t-test, ANOVA, Bayesian method Mann–Whitney test methods tailored to microarray data sets, which take into account multiple comparisons or cluster analysis. These methods assess statistical power based on the variation present in the data and the number of experimental replicates, and can help minimize Type I and type II errors in the analyses.
- Network-based methods: Statistical methods that take the underlying structure of gene networks into account, representing either associative or causative interactions or dependencies among gene products.

Microarray data may require further processing aimed at reducing the dimensionality of the data to aid comprehension and more focused analysis. Other methods permit analysis of data consisting of a low number of biological or technical replicates; for example, the Local Pooled Error (LPE) test pools standard deviations of genes with similar expression levels in an effort to compensate for insufficient replication.

Relation between probe and gene

The relation between a probe and the mRNA that it is expected to detect is not trivial. Some mRNAs may cross-hybridize probes in the array that are supposed to detect another mRNA. In addition, mRNAs may experience amplification bias that is sequence or molecule-specific. Thirdly, probes that are designed to detect the mRNA of a particular gene may be relying on genomic EST information that is incorrectly associated with that gene.

Data warehousing

Microarray data was found to be more useful when compared to other similar datasets. The sheer volume (in bytes), specialized formats (such as MIAME), and curation efforts associated with the datasets require specialized databases to store the data.

Chapter- 11

Epistasis and Functional Genomics

Epistasis refers to genetic interactions in which the mutation of one gene masks the phenotypic effects of a mutation at another locus. Systematic analysis of these epistatic interactions can provide insight into the structure and function of genetic pathways. By examining the phenotypes resulting from pairs of mutations we begin to understand how the function of these genes intersects. Genetic interactions are generally classified as either Positive/Alleviating or Negative/Aggravating. In the case of a positive epistatic interaction, the double mutant exhibits a phenotype which is neutral or improved relative to the phenotype of a single mutant. This phenotypic response occurs when both genes lie within the same pathway. Conversely, negative interactions are characterized by an even stronger defect than would be expected in the case of two single mutations, and in the most extreme cases (synthetic sick/lethal) the double mutation is lethal. This aggravated phenotype arises when genes in compensatory pathways are both knocked out.

High-throughput methods of analyzing these types of interactions have been useful in expanding our knowledge of genetic interactions. Synthetic genetic arrays (SGA), diploid based synthetic lethality analysis on microarrays (dSLAM), and epistatic miniarray profiles (E-MAP) are three important methods which have been developed for the systematic analysis and mapping of genetic interactions. This systematic approach to studying epistasis on a genome wide scale has significant implications for functional genomics. By identifying the negative and positive interactions between an unknown gene and a set genes within a known pathway, these methods can elucidate the function of previously uncharacterized genes within the context of a metabolic or developmental pathway.

Inferring function: alleviating and aggravating mutations

In order to understand how information about epistatic interactions relates to gene pathways, let us consider a simple example of vulval cell differentiation in C. elegans. Cells differentiate from Pn cells to Pn.p cells to VP cells to vulval cells. Mutation of lin-26 blocks differentiation of Pn cells to Pn.p cells. Mutants of lin-36 behave similarly, blocking differentiation at the transition to VP cells. In both cases, the resulting phenotype is marked by an absence of vulval cells as there is an upstream block in the differentiation pathway. A double mutant in which both of these genes have been

disrupted exhibits an equivalent phenotype that is no worse than either single mutant. The upstream disruption at lin-26 masks the phenotypic effect of a mutation at lin-36 in a classic example of an alleviating epistatic interaction.

Aggravating mutations on the other hand give rise to a phenotype which is worse than the cumulative effect of each single mutation. This aggravated phenotype is indicative of two genes in compensatory pathways. In the case of the single mutant a parallel pathway is able to compensate for the loss of the disrupted pathway however, in the case of the double mutant the action of this compensatory pathway is lost as well, resulting in the more dramatic phenotype observed. This relationship has been significantly easier to detect than the more subtle alleviating phenotypes and has been extensively studied in S. cerevisiae through synthetic sick/lethal (SSL) screens which identify double mutants with significantly decreased growth rates.

It should be pointed out that these conclusions from double-mutant analysis, while they apply to many pathways and mutants, are not universal. For example, genes can act in opposite directions in pathways, so that knocking out both produces a near-normal phenotype, while each single mutant is severely affected (in opposite directions). A well-studied example occurs during early development in Drosophila, wherein gene products from the *hunchback* and *nanos* genes are present in the egg, and act in opposite directions to direct anterior-posterior pattern formation. Something similar often happens in signal transduction pathways, where knocking out a negative regulator of the pathway causes a hyper-activation phenotype, while knocking out a positively acting component produces an opposite phenotype. In linear pathways with a single "output", when knockout mutations in two oppositely-acting genes are combined in the same individual, the phenotype of the double mutant is typically the same as the phenotype of the single mutant whose normal gene product acts downstream in the pathway.

Methods of detecting SSL mutants

Synthetic genetic arrays (SGA) and diploid based synthetic lethality analysis of microarrays (dSLAM) are two key methods which have been used to identify synthetic sick lethal mutants and characterize negative epistatic relationships. Sequencing of the entire yeast genome has made it possible to generate a library of knock-out mutants for nearly every gene in the genome. These molecularly bar-coded mutants greatly facilitate high-throughput epistasis studies, as they can be pooled and used to generate the necessary double mutants. Both SGA and dSLAM approaches rely on these yeast knockout strains which are transformed/mated to generate haploid double mutants. Microarray profiling is then used to compare the fitness of these single and double mutants. In the case of SGA, the double mutants examined are haploid and collected after mating with a mutant strain followed by several rounds of selection. dSLAM strains of both single and double mutants originate from the same diploid heterozygote strain (indicated by “diploid” of “dSLAM”). In the case of dSLAM analysis the fitness of single and double mutants is assessed by microarray analysis of a growth competition assay.

Epistatic miniarray profiles (E-MAPs)

In order to develop a richer understanding of genetic interactions, experimental approaches are shifting away from this binary classification of phenotypes as wild type or synthetic lethal. The E-MAP approach is particularly compelling because of its ability to highlight both alleviating and aggravating effects and this capacity is what distinguishes this method from others such as SGA and dSLAM. Furthermore, not only does the E-MAP identify both types of interactions but also recognizes gradations in these interactions and the severity of the masked phenotype, represented by the interaction score applied to each pair of genes.

E-MAPs exploit an SGA approach in order to analyze genetic interactions in a high throughput manner. While the method has been particularly developed for examining epistasis in S. cerevisiae, it could be applied to other model organisms as well. An E-MAP collates data generated from the systematic generation of double mutant strains for a large clearly defined group of genes. Each phenotypic response is quantified by imaging colony size to determine growth rate. This fitness score is compared to the predicted fitness for each single mutant, resulting in a genetic interaction score. Hierarchical clustering of this data to group genes with similar interaction profiles allows for the identification of epistatic relationships between genes with and without known function. By sorting the data in this way, genes known to interact will cluster together alongside genes which exhibit a similar pattern of interactions but whose function has not yet been identified. The E-MAP data is therefore able to place genes into new functions within well characterized pathways. Consider for example E-MAP presented by Collins et al which clusters the transcriptional elongation factor Dst1 alongside components of the mid region of the Mediator complex, which is involved in transcriptional regulation. This suggests a new role for Dst1, functioning in concert with Mediator.

The choice of genes examined within a given E-MAP is critical to achieving fruitful results. It is particularly important that a significant subset of the genes examined have been well established in the literature. These genes are thus able to act as controls for the E-MAP allowing for greater certainty in analyzing the data from uncharacterized genes. Clusters organized by sub-cellular localization and general cellular processes (e.g. cell cycle) have yielded profitable results in S. cerevisiae. Data from protein-protein interaction studies can also provide a useful basis for selecting gene groups for E-MAP data. We would expect genes which exhibit physical interactions to also demonstrate interactions at the genetic level and thus these can serve as adequate controls for E-MAP data. Collins et al (2007) carried out a comparison of E-MAP scores and physical interaction data from large-scale affinity purification methods (AP-MS) and their data demonstrate that an E-MAP approach identifies protein-protein interactions with a specificity equal to that of traditional methods such as AP-MS.

High throughput methods of examining epistatic relationships face difficulties, however as the number of possible gene pairs is extremely large (~20 million in S. cerevisiae) and the estimated density of genetic interactions is quite low. These difficulties can be countered by examining all possible interactions in a single cluster of genes rather than

examining pairs across the whole genome. If well chosen, these functional clusters contain a significantly higher density of genetic interactions than other regions of the genome and thus allows for a higher rate of detection while dramatically decreasing the number of gene pairs to be examined.

Generation of mutant strains: DAmP

Generating data for the E-MAP depends upon the creation of thousands of double mutant strains; a study of 483 alleles, for example, resulted in an E-MAP with ~100,000 distinct double mutant pairs. The generation of libraries of essential gene mutants presents significant difficulties however, as these mutations have a lethal phenotype. Thus, E-MAP studies rely upon strains with intermediate expression levels of these genes. The decreased abundance of messenger RNA perturbation (DAmP) strategy is particularly common for the high-throughput generation of mutants necessary for this kind of analysis and allows for the partial disruption of essential genes without loss of viability. DAmP relies upon the destabilization of mRNA transcripts by integrating an antibiotic selectable marker into the 3'UTR, downstream of the stop codon (figure 2). mRNA's with 3' extended transcripts are rapidly targeted for degrdation and the result is a downregulation of the gene of interest while it remains under the control of its native promoter. In the case of non-essential genes, deletion strains may be used. Tagging at the deletion sites with molecular barcodes, unique 20-bp sequences, allows for the identification and study of relative fitness levels in each mutant strain.

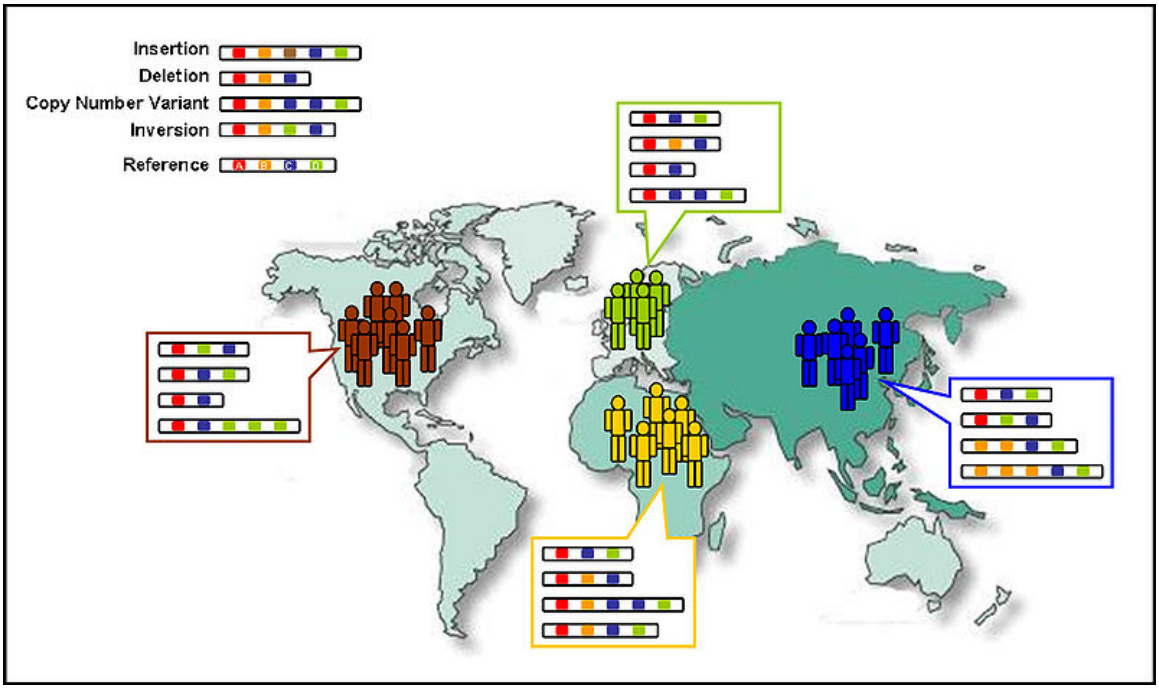
Chapter- 12

1000 Genomes Project

The 1000 Genomes Project, launched in January 2008, is an international research effort to establish by far the most detailed catalogue of human genetic variation. Scientists plan to sequence the genomes of at least one thousand anonymous participants from a number of different ethnic groups within the next three years, using newly developed technologies which are faster and less expensive. In 2010, the project finished its pilot phase, which was described in detail in a publication in Nature. As of late 2010, the project is in its production phase with a target of sequencing upwards of 2000 individuals.

The project unites multidisciplinary research teams from institutes around the world, including the United Kingdom, China and the United States. Each will contribute to the enormous sequence dataset and to a refined human genome map, which will be freely accessible through public databases to the scientific community and the general public alike.

By providing an overview of all genetic variation, not only what is biomedically relevant, the consortium will generate a valuable tool for all fields of natural science, especially in the disciplines of Genetics, Medicine, Pharmacology, Biochemistry and Bioinformatics.



Changes in the number and order of genes (A-D) create genetic diversity within and between populations.

Background

Within the past few decades, advances in human population genetics and comparative genomics have made it possible to gain increasing insight into the nature of genetic diversity. Although, we are just beginning to understand how processes like the random sampling of gametes, structural variations (insertions/ deletions (indels), copy number variations (CNV), retroelements), single-nucleotide polymorphisms (SNPs) and natural selection have shaped the level and pattern of variation within species and also between species.

Human genetic variation

The random sampling of gametes during sexual reproduction leads to genetic drift — a random fluctuation in the population frequency of a trait — in subsequent generations and would result in the loss of all variation in the absence of external influence. It is postulated that the rate of genetic drift is inversely proportional to population size, and that it may be accelerated in specific situations such as bottlenecks, where the population size is reduced for a certain period of time, and by the founder effect (individuals in a population tracing back to a small number of founding individuals).

Anzai et al. demonstrated that indels account for 90.4 % of all observed variations in the sequence of the major histocompatibility locus (MHC) between humans and chimpanzees. After taking multiple indels into consideration, the high degree of genomic similarity between the two species (98.6 % nucleotide sequence identity) drops to only

86.7 %. For example, a large deletion of 95 kilobases (kb) between the loci of the human *MICA* and *MICB* genes, results in a single hybrid chimpanzee *MIC* gene, linking this region to a species-specific handling of several retroviral infections and the resultant susceptibility to various autoimmune diseases. The authors conclude that instead of more subtle SNPs, indels were the driving mechanism in primate speciation.

Besides mutations, SNPs and other structural variants such as CNVs are contributing to the genetic diversity in human populations. Almost 1,500 copy number variable regions, covering around 12% of the genome and containing hundreds of genes, disease loci, functional elements and segmental duplications, have been identified using the HapMap collection, a project describing common patterns of human DNA sequence variation. Although the specific function of CNVs remains elusive, the fact that CNVs span more nucleotide content per genome than SNPs emphasizes the importance of CNVs in genetic diversity and evolution.

Investigating human genomic variations holds great potential for identifying genes that might underlie differences in disease resistance (e.g. MHC region) or drug metabolism.

Natural selection

Natural selection in the evolution of a trait can be divided into three classes. Directional or positive selection refers to a situation where a certain allele has a greater fitness than other alleles, consequently increasing its population frequency (e.g. antibiotic resistance of bacteria). In contrast, stabilizing or negative selection (also known as purifying selection) lowers the frequency or even removes alleles from a population due to disadvantages associated with it with respect to other alleles. Finally, a number of forms of balancing selection exist; those increase genetic variation within a species by being overdominant (heterozygous individuals are fitter than homozygous individuals, e.g. *G6PD*, the gene involved in sickle cell anaemia and malaria resistance) or can vary spatially within a species that inhabits different niches, thus favouring different alleles. Some genomic differences may not affect fitness. Neutral variation, previously thought to be “junk” DNA, is unaffected by natural selection resulting in higher genetic variation at such sites when compared to sites where variation does influence fitness.

It is not fully clear how natural selection has shaped population differences; however, genetic candidate regions under selection have been identified recently. Patterns of DNA polymorphisms can be used to reliably detect signatures of selection and may help to identify genes that might underlie variation in disease resistance or drug metabolism. Barreiro et al. found evidence that negative selection has reduced population differentiation at the amino acid–altering level (particularly in disease-related genes), whereas, positive selection has ensured regional adaptation of human populations by increasing population differentiation in gene regions (mainly nonsynonymous and 5'-untranslated region variants).

It is thought that most complex and Mendelian diseases (except diseases with late onset, assuming that older individuals no longer contribute to the fitness of their offspring) will

have an effect on survival and/or reproduction, thus, genetic factors underlying those diseases should be influenced by natural selection. Gaucher disease (mutations in the *GBA* gene), Crohn's disease (mutation of *NOD2*) and familial hypertrophic cardiomyopathy (mutations in *CMH1*, *CMH2*, *CMH3* and *CMH4*) are all examples of negative selection. These disease mutations are primarily recessive and segregate as expected at a low frequency, supporting the hypothesized negative selection. Few cases have been reported, where disease-causing mutations appear at the high frequencies supported by balanced selection. The most prominent example is mutations of the *G6PD* locus where, if homozygous G6PD enzyme deficiency and consequently sickle-cell anaemia results, but in the heterozygous state are partially protective against malaria. Other possible explanations for segregation of disease alleles at moderate or high frequencies include genetic drift and recent alterations towards positive selection due to environmental changes such as diet or genetic hitch-hiking.

Genome-wide comparative analyses of different human populations, as well as between species (e.g. human versus chimpanzee) are helping us to understand the relationship between diseases and selection and provide evidence of mutations in constrained genes being disproportionally associated with heritable disease phenotypes. Genes implicated in complex disorders tend to be under less negative selection than Mendelian disease genes or non-disease genes.

Project description

Goals

There are two kinds of genetic variants related to disease. The first are rare genetic variants that have a severe effect predominantly on simple traits (e.g. Cystic Fibrosis, Huntington disease). The second, more common, genetic variants have a mild effect and are thought to be implicated in complex traits (e.g. Diabetes, Heart Disease). Between these two types of genetic variants lies a significant gap of knowledge, which the 1000 Genomes Project is designed to address.

The primary goal of this project is to create a complete and detailed catalogue of human genetic variations, which in turn can be used for association studies relating genetic variation to disease. By doing so the consortium aims to discover >95 % of the variants (e.g. SNPs, CNVs, indels) with minor allele frequencies as low as 1% across the genome and 0.1-0.5% in gene regions, as well as to estimate the population frequencies, haplotype backgrounds and linkage disequilibrium patterns of variant alleles.

Secondary goals will include the support of better SNP and probe selection for genotyping platforms in future studies and the improvement of the human reference sequence. Furthermore, the completed database will be a useful tool for studying regions under selection, variation in multiple populations and understanding the underlying processes of mutation and recombination.

Outline

The human genome consists of approximately 3 billion DNA base pairs and is estimated to carry 20,000–25,000 protein coding genes. In designing the study the consortium needed to address several critical issues regarding the project metrics such as technology challenges, proper data quality standards, sequence coverage and data quality.

Over the course of the next three years, scientists at the Sanger Institute, BGI Shenzhen and the National Human Genome Research Institute's Large-Scale Sequencing Network are planning to sequence a minimum of 1,000 human genomes. Due to the large amount of sequence data that need to be generated and analyzed it is possible that other participants may be recruited over time.

Almost 10 billion bases will be sequenced per day over a period of the two year production phase. This equates to more than two human genomes every 24 hours; a groundbreaking capacity. Challenging the leading experts of bioinformatics and statistical genetics, the sequence dataset will comprise 6 trillion DNA bases, 60-fold more sequence data than what has been published in DNA databases over the past 25 years.

To determine the final design of the full project three pilot studies were designed and will be carried out within the first year of the project. The first pilot intends to genotype 180 people of 3 major geographic groups at low coverage (2x). For the second pilot study, the genomes of two nuclear families (both parents and an adult child) are going to be sequenced with deep coverage (20x per genome). The third pilot study involves sequencing the coding regions (exons) of 1,000 genes in 1,000 people with deep coverage (20x).

It has been estimated that the project would likely cost more than $500 million if standard DNA sequencing technologies were used. Therefore, several new technologies (e.g. Solexa, 454, SOLiD) will be applied, lowering the expected costs to $30 million to $50 million. The major support will be provided by the Wellcome Trust Sanger Institute in Hinxton, England; the Beijing Genomics Institute, Shenzhen (BGI Shenzhen), China; and the NHGRI, part of the National Institutes of Health (NIH).

Human genome samples

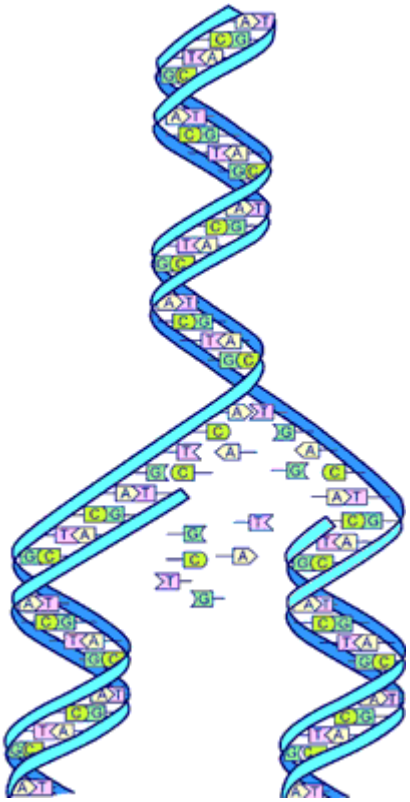
Based on the overall goals for the project, the samples will be chosen to provide power in populations where association studies for common diseases are being carried out. Furthermore, the samples do not need to have medical or phenotype information since the proposed catalogue will be a basic resource on human variation.

For the pilot studies human genome samples from the HapMap collection will be sequenced. It will be useful to focus on samples that have additional data available (such as ENCODE sequence, genome-wide genotypes, fosmid-end sequence, structural variation assays, and gene expression) to be able to compare the results with those from other projects.

Complying with extensive ethical procedures, the 1000 Genomes Project will then use samples from volunteer donors. The following populations will be included in the study: Yoruba in Ibadan, Nigeria; Japanese in Tokyo; Chinese in Beijing; Utah residents with ancestry from northern and western Europe; Luhya in Webuye, Kenya; Maasai in Kinyawa, Kenya; Toscani in Italy; Peruvians in Perú; Gujarati Indians in Houston; Chinese in metropolitan Denver; people of Mexican ancestry in Los Angeles; and people of African ancestry in the southwestern United States.

Chapter- 13

Human Genome Project



DNA Replication

The **Human Genome Project (HGP)** is an international scientific research project with a primary goal of determining the sequence of chemical base pairs which make up DNA and to identify and map the approximately 20,000–25,000 genes of the human genome from both a physical and functional standpoint.

The project began in 1990 and was initially headed by Ari Patrinos, head of the Office of Biological and Environmental Research in the U.S. Department of Energy's Office of Science. Francis Collins directed the National Institutes of Health National Human

Genome Research Institute efforts. A working draft of the genome was announced in 2000 and a complete one in 2003, with further, more detailed analysis still being published. A parallel project was conducted outside of government by the Celera Corporation, which was formally launched in 1998. Most of the government-sponsored sequencing was performed in universities and research centers from the United States, the United Kingdom, Japan, France, Germany, and China. The mapping of human genes is an important step in the development of medicines and other aspects of health care.

While the objective of the Human Genome Project is to understand the genetic makeup of the human species, the project has also focused on several other nonhuman organisms such as *E. coli*, the fruit fly, and the laboratory mouse. It remains one of the largest single investigative projects in modern science.

The Human Genome Project originally aimed to map the nucleotides contained in a human haploid reference genome (more than three billion). Several groups have announced efforts to extend this to diploid human genomes including the International HapMap Project, Applied Biosystems, Perlegen, Illumina, JCVI, Personal Genome Project, and Roche-454.

The "genome" of any given individual (except for identical twins and cloned organisms) is unique; mapping "the human genome" involves sequencing multiple variations of each gene. The project did not study the entire DNA found in human cells; some heterochromatic areas (about 8% of the total genome) remain un-sequenced.

Project

Background

The project began with the culmination of several years of work supported by the United States Department of Energy, in particular workshops in 1984 and 1986 and a subsequent initiative of the US Department of Energy. This 1987 report stated boldly, "The ultimate goal of this initiative is to understand the human genome" and "knowledge of the human as necessary to the continuing progress of medicine and other health sciences as knowledge of human anatomy has been for the present state of medicine." Candidate technologies were already being considered for the proposed undertaking at least as early as 1985.

James D. Watson was head of the National Center for Human Genome Research at the National Institutes of Health (NIH) in the United States starting from 1988. Largely due to his disagreement with his boss, Bernadine Healy, over the issue of patenting genes, Watson was forced to resign in 1992. He was replaced by Francis Collins in April 1993, and the name of the Center was changed to the National Human Genome Research Institute (NHGRI) in 1997.

The $3-billion project was formally founded in 1990 by the United States Department of Energy and the U.S. National Institutes of Health, and was expected to take 15 years. In

addition to the United States, the international consortium comprised geneticists in the United Kingdom, France, Germany, Japan, China, and India.

Due to widespread international cooperation and advances in the field of genomics (especially in sequence analysis), as well as major advances in computing technology, a 'rough draft' of the genome was finished in 2000 (announced jointly by then US president Bill Clinton and the British Prime Minister Tony Blair on June 26, 2000). This first available rough draft assembly of the genome was completed by the UCSC Genome Bioinformatics Group, primarily led by then graduate student Jim Kent. Ongoing sequencing led to the announcement of the essentially complete genome in April 2003, 2 years earlier than planned. In May 2006, another milestone was passed on the way to completion of the project, when the sequence of the last chromosome was published in the journal Nature.

State of completion

There are multiple definitions of the "complete sequence of the human genome". According to some of these definitions, the genome has already been completely sequenced, and according to other definitions, the genome has yet to be completely sequenced. There have been multiple popular press articles reporting that the genome was "complete." The genome has been completely sequenced using the definition employed by the International Human Genome Project. A graphical history of the human genome project shows that most of the human genome was complete by the end of 2003. However, there are a number of regions of the human genome that can be considered unfinished:

- First, the central regions of each chromosome, known as centromeres, are highly repetitive DNA sequences that are difficult to sequence using current technology. The centromeres are millions (possibly tens of millions) of base pairs long, and for the most part these are entirely un-sequenced.
- Second, the ends of the chromosomes, called telomeres, are also highly repetitive, and for most of the 46 chromosome ends these too are incomplete. It is not known precisely how much sequence remains before the telomeres of each chromosome are reached, but as with the centromeres, current technological restraints are prohibitive.
- Third, there are several loci in each individual's genome that contain members of multigene families that are difficult to disentangle with shotgun sequencing methods – these multigene families often encode proteins important for immune functions.
- Other than these regions, there remain a few dozen gaps scattered around the genome, some of them rather large, but there is hope that all these will be closed in the next couple of years.

In summary: the best estimates of total genome size indicate that about 92.3% of the genome has been completed and it is likely that the centromeres and telomeres will remain un-sequenced until new technology is developed that facilitates their sequencing.

Most of the remaining DNA is highly repetitive and unlikely to contain genes, but it cannot be truly known until it is entirely sequenced. Understanding the functions of all the genes and their regulation is far from complete. The roles of junk DNA, the evolution of the genome, the differences between individuals, and many other questions are still the subject of intense interest by laboratories all over the world.

Goals

The sequence of the human DNA is stored in databases available to anyone on the Internet. The U.S. National Center for Biotechnology Information (and sister organizations in Europe and Japan) house the gene sequence in a database known as GenBank, along with sequences of known and hypothetical genes and proteins. Other organizations such as the University of California, Santa Cruz, and Ensembl present additional data and annotation and powerful tools for visualizing and searching it. Computer programs have been developed to analyze the data, because the data itself is difficult to interpret without such programs.

The process of identifying the boundaries between genes and other features in a raw DNA sequence is called genome annotation and is the domain of bioinformatics. While expert biologists make the best annotators, their work proceeds slowly, and computer programs are increasingly used to meet the high-throughput demands of genome sequencing projects. The best current technologies for annotation make use of statistical models that take advantage of parallels between DNA sequences and human language, using concepts from computer science such as formal grammars.

Another, often overlooked, goal of the HGP is the study of its ethical, legal, and social implications. It is important to research these issues and find the most appropriate solutions before they become large dilemmas whose effect will manifest in the form of major political concerns.

All humans have unique gene sequences. Therefore the data published by the HGP does not represent the exact sequence of each and every individual's genome. It is the combined "reference genome" of a small number of anonymous donors. The HGP genome is a scaffold for future work in identifying differences among individuals. Most of the current effort in identifying differences among individuals involves single-nucleotide polymorphisms and the HapMap.

Findings

Key findings of the draft (2001) and complete (2004) genome sequences include

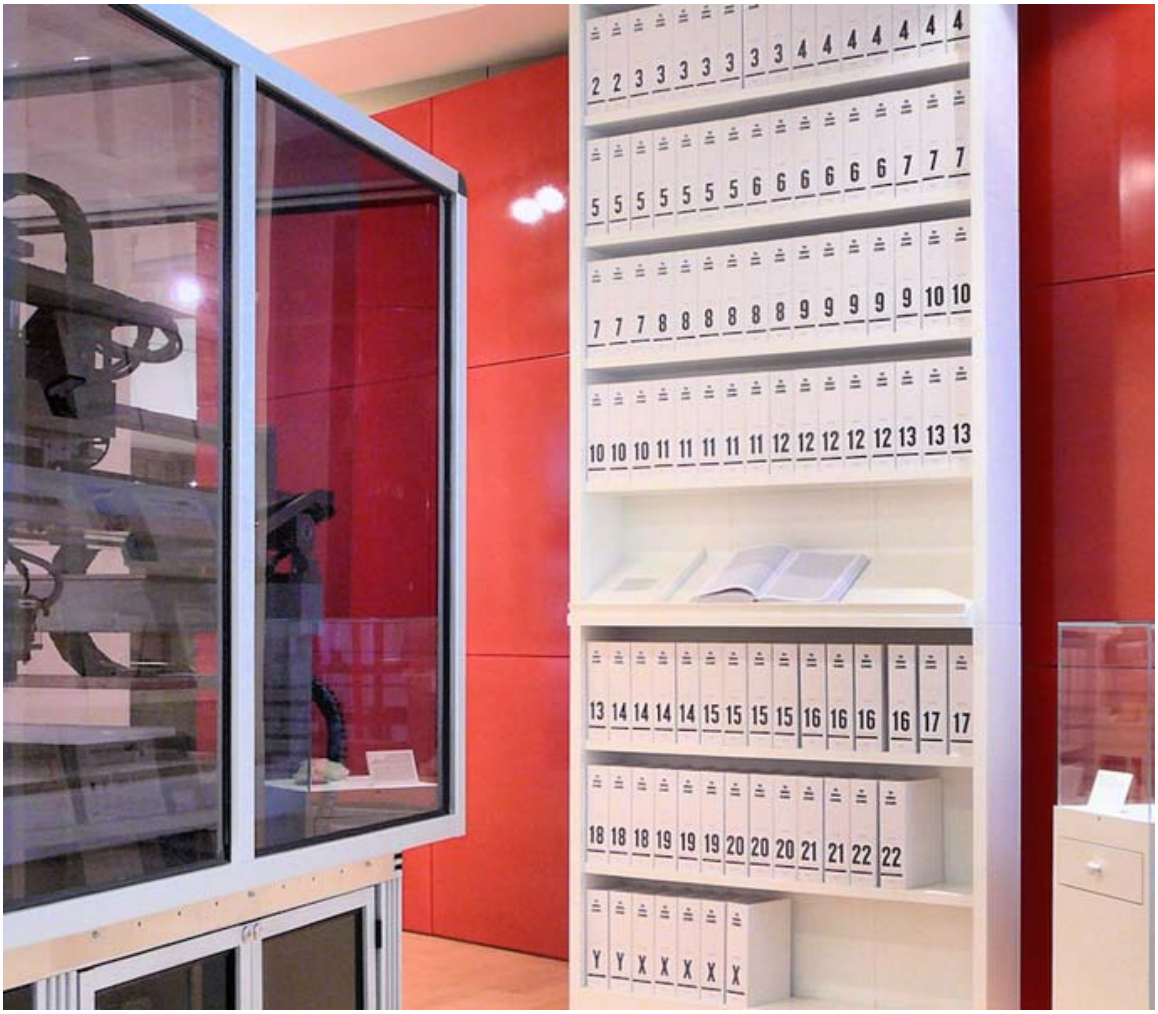
1. There are approximately 20,500 genes in human beings, the same range as in mice and twice that of roundworms. Understanding how these genes express themselves will provide clues to how diseases are caused.

2. Between 1.1% to 1.4% of the genome's sequence codes for proteins

3. The human genome has significantly more segmental duplications (nearly identical, repeated sections of DNA) than other mammalian genomes. These sections may underlie the creation of new primate-specific genes

4. At the time when the draft sequence was published less than 7% of protein families appeared to be vertebrate specific

How it was accomplished



The first printout of the human genome to be presented as a series of books, displayed at the Wellcome Collection, London

The Human Genome Project was started in 1989 with the goal of sequencing and identifying all three billion chemical units in the human genetic instruction set, finding the genetic roots of disease and then developing treatments. With the sequence in hand, the next step was to identify the genetic variants that increase the risk for common diseases like cancer and diabetes.

It was far too expensive at that time to think of sequencing patients' whole genomes. So the National Institutes of Health embraced the idea for a "shortcut", which was to look just at sites on the genome where many people have a variant DNA unit. The theory behind the shortcut was that since the major diseases are common, so too would be the genetic variants that caused them. Natural selection keeps the human genome free of variants that damage health before children are grown, the theory held, but fails against variants that strike later in life, allowing them to become quite common. (In 2002 the National Institutes of Health started a $138 million project called the HapMap to catalog the common variants in European, East Asian and African genomes.)

The genome was broken into smaller pieces; approximately 150,000 base pairs in length. These pieces were then ligated into a type of vector known as "bacterial artificial chromosomes", or BACs, which are derived from bacterial chromosomes which have been genetically engineered. The vectors containing the genes can be inserted into bacteria where they are copied by the bacterial DNA replication machinery. Each of these pieces was then sequenced separately as a small "shotgun" project and then assembled. The larger, 150,000 base pairs go together to create chromosomes. This is known as the "hierarchical shotgun" approach, because the genome is first broken into relatively large chunks, which are then mapped to chromosomes before being selected for sequencing.

Funding came from the US government through the National Institutes of Health in the United States, and a UK charity organization, the Wellcome Trust, as well as numerous other groups from around the world. The funding supported a number of large sequencing centers including those at Whitehead Institute, the Sanger Centre, Washington University in St. Louis, and Baylor College of Medicine.

The Human Genome Project is considered a Mega Project because the human genome has approximately 3.3 billion base-pairs.

If the sequence obtained was to be stored in book form, and if each page contained 1000 base-pairs recorded and each book contained 1000 pages, then 3300 such books would be needed in order to store the complete genome. However, if expressed in units of computer data storage, 3.3 billion base-pairs recorded at 2 bits per pair would equal 786 megabytes of raw data. This is comparable to a fully data loaded CD.

Public versus private approaches

In 1998, a similar, privately funded quest was launched by the American researcher Craig Venter, and his firm Celera Genomics. Venter was a scientist at the NIH during the early 1990s when the project was initiated. The $300,000,000 Celera effort was intended to proceed at a faster pace and at a fraction of the cost of the roughly $3 billion publicly funded project.

Celera used a technique called whole genome shotgun sequencing, employing pairwise end sequencing, which had been used to sequence bacterial genomes of up to six million

base pairs in length, but not for anything nearly as large as the three billion base pair human genome.

Celera initially announced that it would seek patent protection on "only 200–300" genes, but later amended this to seeking "intellectual property protection" on "fully-characterized important structures" amounting to 100–300 targets. The firm eventually filed preliminary ("place-holder") patent applications on 6,500 whole or partial genes. Celera also promised to publish their findings in accordance with the terms of the 1996 "Bermuda Statement," by releasing new data annually (the HGP released its new data daily), although, unlike the publicly funded project, they would not permit free redistribution or scientific use of the data. The publicly funded competitor UC Santa Cruz was compelled to publish the first draft of the human genome before Celera for this reason. On July 7, 2000, the UCSC Genome Bioinformatics Group released a first working draft on the web. The scientific community downloaded one-half trillion bytes of information from the UCSC genome server in the first 24 hours of free and unrestricted access to the first ever assembled blueprint of our human species.

In March 2000, President Clinton announced that the genome sequence could not be patented, and should be made freely available to all researchers. The statement sent Celera's stock plummeting and dragged down the biotechnology-heavy Nasdaq. The biotechnology sector lost about $50 billion in market capitalization in two days.

Although the working draft was announced in June 2000, it was not until February 2001 that Celera and the HGP scientists published details of their drafts. Special issues of *Nature* (which published the publicly funded project's scientific paper) and *Science* (which published Celera's paper) described the methods used to produce the draft sequence and offered analysis of the sequence. These drafts covered about 83% of the genome (90% of the euchromatic regions with 150,000 gaps and the order and orientation of many segments not yet established). In February 2001, at the time of the joint publications, press releases announced that the project had been completed by both groups. Improved drafts were announced in 2003 and 2005, filling in to ≈92% of the sequence currently.

The competition proved to be very good for the project, spurring the public groups to modify their strategy in order to accelerate progress. The rivals at UC Santa Cruz initially agreed to pool their data, but the agreement fell apart when Celera refused to deposit its data in the unrestricted public database GenBank. Celera had incorporated the public data into their genome, but forbade the public effort to use Celera data.

HGP is the most well known of many international genome projects aimed at sequencing the DNA of a specific organism. While the human DNA sequence offers the most tangible benefits, important developments in biology and medicine are predicted as a result of the sequencing of model organisms, including mice, fruit flies, zebrafish, yeast, nematodes, plants, and many microbial organisms and parasites.

In 2004, researchers from the International Human Genome Sequencing Consortium (IHGSC) of the HGP announced a new estimate of 20,000 to 25,000 genes in the human genome. Previously 30,000 to 40,000 had been predicted, while estimates at the start of the project reached up to as high as 2,000,000. The number continues to fluctuate and it is now expected that it will take many years to agree on a precise value for the number of genes in the human genome.

History

In 1976, the genome of the RNA virus Bacteriophage MS2 was the first complete genome to be determined, by Walter Fiers and his team at the University of Ghent (Ghent, Belgium). The idea for the shotgun technique came from the use of an algorithm that combined sequence information from many small fragments of DNA to reconstruct a genome. This technique was pioneered by Frederick Sanger to sequence the genome of the Phage Φ-X174, a virus (bacteriophage) that primarily infects bacteria that was the first fully sequenced genome (DNA-sequence) in 1977. The technique was called shotgun sequencing because the genome was broken into millions of pieces as if it had been blasted with a shotgun. In order to scale up the method, both the sequencing and genome assembly had to be automated, as they were in the 1980s.

Those techniques were shown applicable to sequencing of the first free-living bacterial genome (1.8 million base pairs) of *Haemophilus influenzae* in 1995 and the first animal genome (~100 Mbp) It involved the use of automated sequencers, longer individual sequences using approximately 500 base pairs at that time. Paired sequences separated by a fixed distance of around 2000 base pairs which were critical elements enabling the development of the first genome assembly programs for reconstruction of large regions of genomes (aka 'contigs').

Three years later, in 1998, the announcement by the newly-formed Celera Genomics that it would scale up the pairwise end sequencing method to the human genome was greeted with skepticism in some circles. The shotgun technique breaks the DNA into fragments of various sizes, ranging from 2,000 to 300,000 base pairs in length, forming what is called a DNA "library". Using an automated DNA sequencer the DNA is read in 800bp lengths from both ends of each fragment. Using a complex genome assembly algorithm and a supercomputer, the pieces are combined and the genome can be reconstructed from the millions of short, 800 base pair fragments. The success of both the public and privately funded effort hinged upon a new, more highly automated capillary DNA sequencing machine, called the Applied Biosystems 3700, that ran the DNA sequences through an extremely fine capillary tube rather than a flat gel. Even more critical was the development of a new, larger-scale genome assembly program, which could handle the 30–50 million sequences that would be required to sequence the entire human genome with this method. At the time, such a program did not exist. One of the first major projects at Celera Genomics was the development of this assembler, which was written in parallel with the construction of a large, highly automated genome sequencing factory. Development of the assembler was led by Brian Ramos. The first version of this assembler was demonstrated in 2000, when the Celera team joined forces with Professor

Gerald Rubin to sequence the fruit fly Drosophila melanogaster using the whole-genome shotgun method. At 130 million base pairs, it was at least 10 times larger than any genome previously shotgun assembled. One year later, the Celera team published their assembly of the three billion base pair human genome.

The Human Genome Project was a 13 year old mega project, that was launched in the year 1990 and completed in 2003. This project is closely associated to the branch of biology called Bio-informatics. The human genome project international consortium announced the publication of a draft sequence and analysis of the human genome—the genetic blueprint for the human being. An American company—Celera, led by Craig Venter and the other huge international collaboration of distinguished scientists led by Francis Collins, director, National Human Genome Research Institute, U.S., both published their findings.

This Mega Project is co-ordinated by the U.S. Department of Energy and the National Institute of Health. During the early years of the project, the Wellcome Trust (U.K.) became a major partner, other countries like Japan, Germany, China and France contributed significantly. Already the atlas has revealed some starting facts. The two factors that made this project a success are:

1. Genetic Engineering Techniques, with which it is possible to isolate and clone any segment of DNA.
2. Availability of simple and fast technologies, to determining the DNA sequences.

Being the most complex organisms, human beings were expected to have more than 100,000 genes or combination of DNA that provides commands for every characteristics of the body. Instead their studies show that humans have only 30,000 genes – around the same as mice, three times as many as flies, and only five times more than bacteria. Scientist told that not only are the numbers similar, the genes themselves, baring a few, are alike in mice and men. In a companion volume to the Book of Life, scientists have created a catalogue of 1.4 million single-letter differences, or single-nucleotide polymorphisms (SNPs) – and specified their exact locations in the human genome. This SNP map, the world's largest publicly available catalogue of SNP's, promises to revolutionize both mapping diseases and tracing human history. The sequence information from the consortium has been immediately and freely released to the world, with no restrictions on its use or redistribution. The information is scanned daily by scientists in academia and industry, as well as commercial database companies, providing key information services to bio-technologists. Already, many genes have been identified from the genome sequence, including more than 30 that play a direct role in human diseases. By dating the three millions repeat elements and examining the pattern of interspersed repeats on the Y-chromosome, scientists estimated the relative mutation rates in the X and the Y chromosomes and in the male and the female germ lines. They found that the ratio of mutations in male Vs female is 2:1. Scientists point to several possible reasons for the higher mutation rate in the male germ line, including the fact that there are a greater number of cell divisions involved in the formation of sperm than in the formation of eggs.

Methods

The IHGSC used pair-end sequencing plus whole-genome shotgun mapping of large (≈100 Kbp) plasmid clones and shotgun sequencing of smaller plasmid sub-clones plus a variety of other mapping data to orient and check the assembly of each human chromosome.

The Celera group emphasized the importance of the "whole-genome shotgun" sequencing method, relying on sequence information to orient and locate their fragments within the chromosome. However they used the publicly available data from HGP to assist in the assembly and orientation process, raising concerns that the Celera sequence was not independently derived.

Genome donors

In the IHGSC international public-sector Human Genome Project (HGP), researchers collected blood (female) or sperm (male) samples from a large number of donors. Only a few of many collected samples were processed as DNA resources. Thus the donor identities were protected so neither donors nor scientists could know whose DNA was sequenced. DNA clones from many different libraries were used in the overall project, with most of those libraries being created by Dr. Pieter J. de Jong. It has been informally reported, and is well known in the genomics community, that much of the DNA for the public HGP came from a single anonymous male donor from Buffalo, New York (code name RP11).

HGP scientists used white blood cells from the blood of two male and two female donors (randomly selected from 20 of each) -- each donor yielding a separate DNA library. One of these libraries (RP11) was used considerably more than others, due to quality considerations. One minor technical issue is that male samples contain just over half as much DNA from the sex chromosomes (one X chromosome and one Y chromosome) compared to female samples (which contain two X chromosomes). The other 22 chromosomes (the autosomes) are the same for both genders.

Although the main sequencing phase of the HGP has been completed, studies of DNA variation continue in the International HapMap Project, whose goal is to identify patterns of single-nucleotide polymorphism (SNP) groups (called haplotypes, or "haps"). The DNA samples for the HapMap came from a total of 270 individuals: Yoruba people in Ibadan, Nigeria; Japanese people in Tokyo; Han Chinese in Beijing; and the French Centre d'Etude du Polymorphisms Humain (CEf) resource, which consisted of residents of the United States having ancestry from Western and Northern Europe.

In the Celera Genomics private-sector project, DNA from five different individuals were used for sequencing. The lead scientist of Celera Genomics at that time, Craig Venter, later acknowledged (in a public letter to the journal Science) that his DNA was one of 21 samples in the pool, five of which were selected for use.

On September 4, 2007, a team led by Craig Venter published his complete DNA sequence, unveiling the six-billion-nucleotide genome of a single individual for the first time.

Benefits

The work on interpretation of genome data is still in its initial stages. It is anticipated that detailed knowledge of the human genome will provide new avenues for advances in medicine and biotechnology. Clear practical results of the project emerged even before the work was finished. For example, a number of companies, such as Myriad Genetics started offering easy ways to administer genetic tests that can show predisposition to a variety of illnesses, including breast cancer, disorders of hemostasis, cystic fibrosis, liver diseases and many others. Also, the etiologies for cancers, Alzheimer's disease and other areas of clinical interest are considered likely to benefit from genome information and possibly may lead in the long term to significant advances in their management.

There are also many tangible benefits for biological scientists. For example, a researcher investigating a certain form of cancer may have narrowed down his/her search to a particular gene. By visiting the human genome database on the World Wide Web, this researcher can examine what other scientists have written about this gene, including (potentially) the three-dimensional structure of its product, its function(s), its evolutionary relationships to other human genes, or to genes in mice or yeast or fruit flies, possible detrimental mutations, interactions with other genes, body tissues in which this gene is activated, diseases associated with this gene or other datatypes.

Further, deeper understanding of the disease processes at the level of molecular biology may determine new therapeutic procedures. Given the established importance of DNA in molecular biology and its central role in determining the fundamental operation of cellular processes, it is likely that expanded knowledge in this area will facilitate medical advances in numerous areas of clinical interest that may not have been possible without them.

The analysis of similarities between DNA sequences from different organisms is also opening new avenues in the study of evolution. In many cases, evolutionary questions can now be framed in terms of molecular biology; indeed, many major evolutionary milestones (the emergence of the ribosome and organelles, the development of embryos with body plans, the vertebrate immune system) can be related to the molecular level. Many questions about the similarities and differences between humans and our closest relatives (the primates, and indeed the other mammals) are expected to be illuminated by the data from this project.

The Human Genome Diversity Project (HGDP), spinoff research aimed at mapping the DNA that varies between human ethnic groups, which was rumored to have been halted, actually did continue and to date has yielded new conclusions. In the future, HGDP could possibly expose new data in disease surveillance, human development and anthropology. HGDP could unlock secrets behind and create new strategies for managing the

vulnerability of ethnic groups to certain diseases. It could also show how human populations have adapted to these vulnerabilities.

Advantages of Human Genome Project:

1. Knowledge of the effects of variation of DNA among individuals can revolutionize the ways to diagnose, treat and even prevent a number of diseases that affects the human beings.
2. It provides clues to the understanding of human biology.

Ethical, legal and social issues

The project's goals included not only identifying all of the approximately 24,000 genes in the human genome, but also to address the ethical, legal, and social issues (ELSI) that might arise from the availability of genetic information. Five percent of the annual budget was allocated to address the ELSI arising from the project.

Debra Harry, Executive Director of the U.S group Indigenous Peoples Council on Biocolonialism (IPCB), says that despite a decade of ELSI funding, the burden of genetics education has fallen on the tribes themselves to understand the motives of Human genome project and its potential impacts on their lives. Meanwhile, the government has been busily funding projects studying indigenous groups without any meaningful consultation with the groups.

The main criticism of ELSI is the failure to address the conditions raised by population-based research, especially with regard to unique processes for group decision-making and cultural worldviews. Genetic variation research such as HGP is group population research, but most ethical guidelines, according to Harry, focus on individual rights instead of group rights. She says the research represents a clash of culture: indigenous people's life revolves around collectivity and group decision making whereas the Western culture promotes individuality. Harry suggests that one of the challenges of ethical research is to include respect for collective review and decision making, while also upholding the Western model of individual rights.

Chapter- 14

Structural Genomics

Structural genomics seeks to describe the 3-dimensional structure of every protein encoded by a given genome. This genome-based approach allows for a high-throughput method of structure determination by a combination of experimental and modeling approaches. The principal difference between structural genomics and traditional structural prediction is that structural genomics attempts to determine the structure of every protein encoded by the genome, rather than focusing on one particular protein. What is it that makes it possible to determine the structure of every protein in the genome at once rather than solve the structures one at a time? With full-genome sequences available, structure prediction can be done more quickly through a combination of experimental and modeling approaches, especially because the availability of large number of sequenced genomes and previously-solved protein structures allows scientists to model protein structure on the structures of previously solved homologs.

Because protein structure is closely linked with protein function, the structural genomics has the potential to inform knowledge of protein function. In addition to elucidating protein functions, structural genomics can be used to identify novel protein folds and potential targets for drug discovery. Structural genomics involves taking a large number of approaches to structure determination, including experimental methods using genomic sequences or modeling-based approaches based on sequence or structural homology to a protein of known structure or based on chemical and physical principles for a protein with no homology to any known structure.

As opposed to traditional structural biology, the determination of a protein structure through a structural genomics effort often (but not always) comes before anything is known regarding the protein function. This raises new challenges in structural bioinformatics, i.e. determining protein function from its 3D structure.

Structural genomics emphasizes high throughput determination of protein structures. This is performed in dedicated centers of structural genomics.

While most structural biologists pursue structures of individual proteins or protein groups, specialists in **structural genomics** pursue structures of proteins on a genome wide scale. This implies large scale cloning, expression and purification. One main

advantage of this approach is economy of scale. On the other hand, the scientific value of some resultant structures is at times questioned. A *Science* article from January 2006 analyzes the structural genomics field.

One advantage of structural genomics, such as the Protein Structure Initiative, is that the scientific community gets immediate access to new structures, as well as to reagents such as clones and protein. A disadvantage is that many of these structures are of proteins of unknown function and do not have corresponding publications. This requires new ways of communicating this structural information to the broader research community.

Goals

One goal of structural genomics is to identify novel protein folds. Experimental methods of protein structure determination require proteins that express and/or crystallize well, which may inherently bias the kinds of proteins folds that this experimental data elucidate. A genomic, modeling-based approach such as *ab initio* modeling may be better able to identify novel protein folds than the experimental approaches because they are not limited by experimental constraints.

Protein function depends on 3-D structure and these 3-D structures are more highly-conserved than sequences. Thus, the high-throughput structure determination methods of structural genomics have the potential to inform our understanding of protein functions. This also has potential implications for drug discovery and protein engineering. Furthermore, every protein that is added to the structural database increases the likelihood that the database will include homologous sequences of other unknown proteins. The Protein Structure Initiative (PSI) is a multifaceted effort funded by the National Institutes of Health with various academic and industrial partners that aims to increase knowledge of protein structure using a structural genomics approach and to improve structure-determination methodology.

Methods

Structural genomics takes advantage of completed genome sequences in several ways in order to determine protein structures. The gene sequence of the target protein can also be compared to a known sequence and structural information can then be inferred from the known protein's structure. Structural genomics can be used to predict novel protein folds based on other structural data. Structural genomics can also take modeling-based approach that relies on homology between the unknown protein and a solved protein structure.

de novo methods

Completed genome sequences allow every open reading frame (ORF), the part of a gene that is likely to contain the sequence for the mRNA and protein, to be cloned and expressed as protein. These proteins are then purified and crystallized, and then subjected to one of two types of structure determination: X-ray crystallography and [Nuclear

Magnetic Resonance (NMR). The whole genome sequence allows for the design of every primer required in order to amplify all of the ORFs, clone them into bacteria, and then express them. By using a whole-genome approach to this traditional method of protein structure determination, all of the proteins encoded by the genome can be expressed at once. This approach allows for the structural determination of every protein that is encoded by the genome.

Modelling-based methods

ab initio modeling

This approach uses protein sequence data and the chemical and physical interactions of the encoded amino acids to predict the 3-D structures of proteins with no homology to solved protein structures. One highly successful method for *ab initio* modeling is the Rosetta program, which divides the protein into short segments and arranges short polypeptide chain into a low-energy local conformation. Rosetta is available for commercial use and for non-commercial use through its public program, Robetta.

Sequence-based modeling

This modeling technique compares the gene sequence of an unknown protein with sequences of proteins with known structures. Depending on the degree of similarity between the sequences, the structure of the known protein can be used as a model for solving the structure of the unknown protein. Highly accurate modeling is considered to require at least 50% amino acid sequence identity between the unknown protein and the solved structure. 30-50% sequence identity gives a model of intermediate-accuracy, and sequence identity below 30% gives low-accuracy models. It has been predicted that at least 16,000 protein structures will need to be determined in order for all structural motifs to be represented at least once and thus allowing the structure of any unknown protein to be solved accurately through modeling. One disadvantage of this method, however, is that structure is more conserved than sequence and thus sequence-based modeling may not be the most accurate way to predict protein structures.

Threading

Threading bases structural modeling on fold similarities rather than sequence identity. This method may help identify distantly-related proteins and can be used to infer molecular functions.

Examples of Structural Genomics

There are currently a number of on-going efforts to solve the structures for every protein in a given proteome.

The *Thermotogo maritima* proteome

One current goal of the Joint Center for Structural Genomics (JCSG), a part of the Protein Structure Initiative (PSI) is to solve the structures for all the proteins in *Thermotogo maritima*, a thermophillic bacterium. *T. maritima* was selected as a structural genomics target based on its relatively small genome consisting of 1,877 genes and the hypothesis that the proteins expressed by a thermophilic bacterium would be easier to crystallize.

Lesley *et al* used *Escherichia coli* to express all the open-reading frames (ORFs) of *T. martima*. These proteins were then crystallized and structures were determined for successfully-crystallized proteins using X-ray crystallography. Among other structures, this structural genomics approach allowed for the determination of the structure of the TM0449 protein, which was found to exhibit a novel fold as it did not share structural homology with any known protein.

The *Mycobacterium tuberculosis* proteome

The goal of the TB Structural Genomics Consortium is to determine the structures of potential drug targets in *Mycobacterium tuberculosis*, the bacterium that causes tuberculosis. The development of novel drug therapies against tuberculosis are particularly important given the growing problem of multi-drug-resistant tuberculosis.

The fully sequenced genome of *M. tuberculosis* has allowed scientists to clone many of these protein targets into expression vectors for purification and structure determination by X-ray crystallography. Studies have identified a number of target proteins for structure determination, including extracellular proteins that may be involved in pathogenesis, iron-regulatory proteins, current drug targets, and proteins predicted to have novel folds. So far, structures have been determined for 708 of the proteins encoded by *M. tuberculosis*.