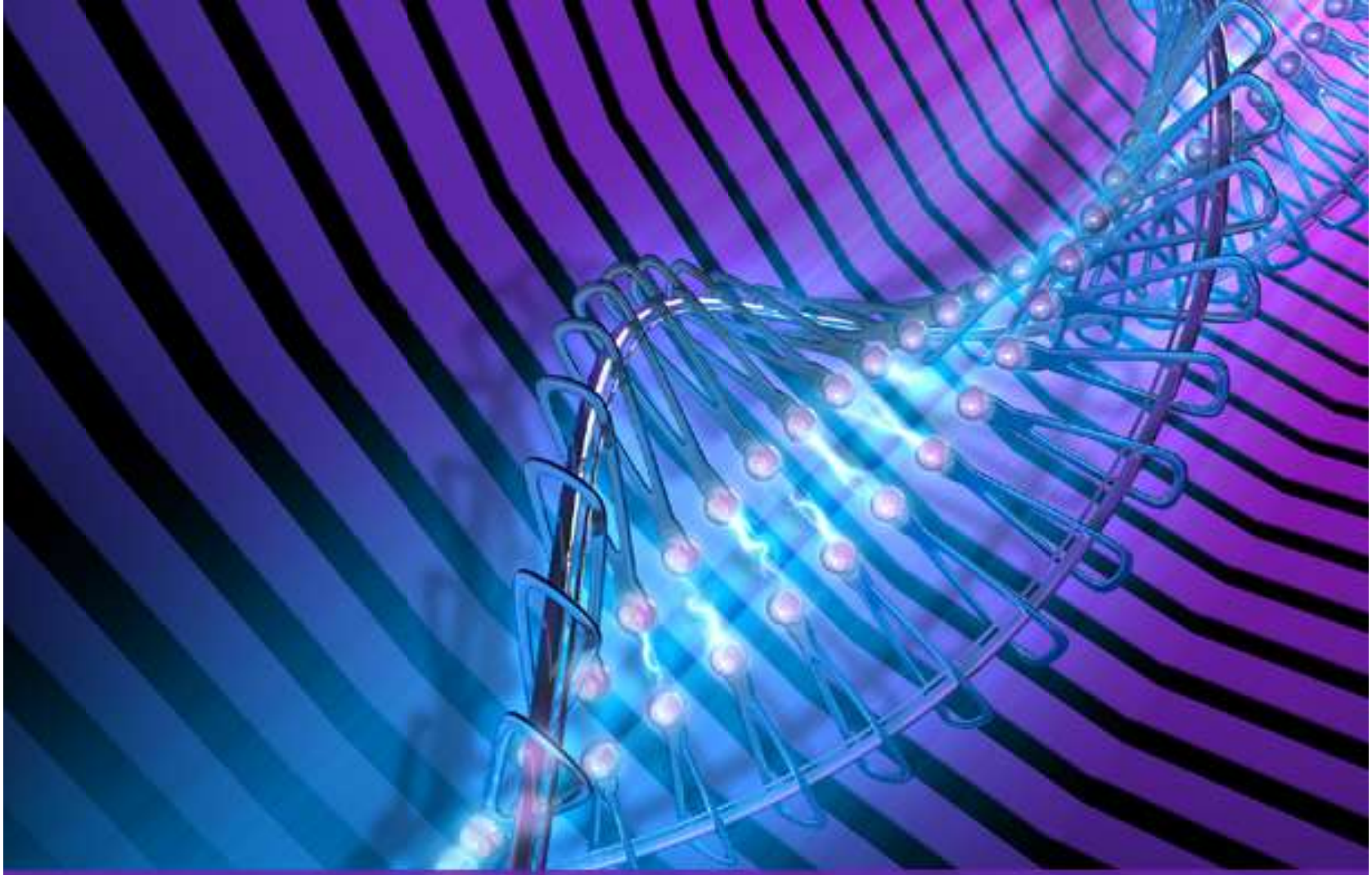


Gene Biology & Genomics

(Concepts, Elements and Applications)

Scarlet Cormier
Lovetta Arndt

First Edition, 2012

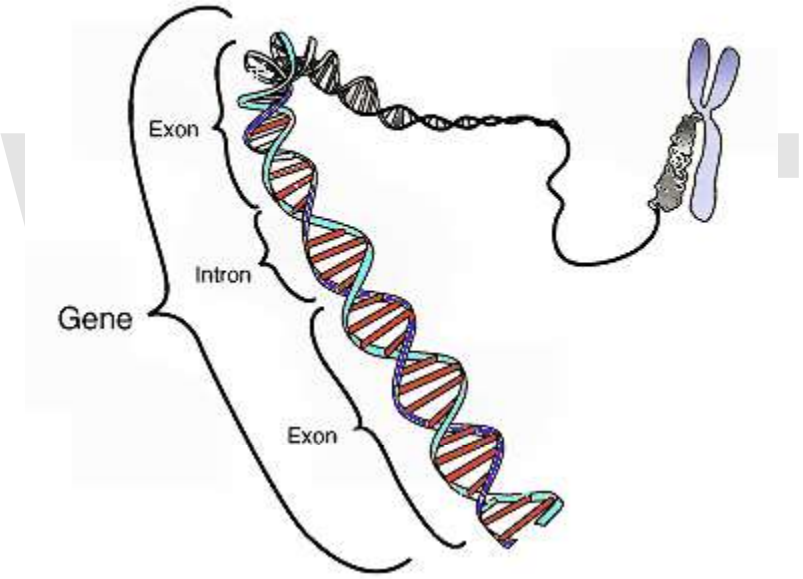
ISBN 978-81-323-0710-5

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

Table of Contents

Chapter- 1

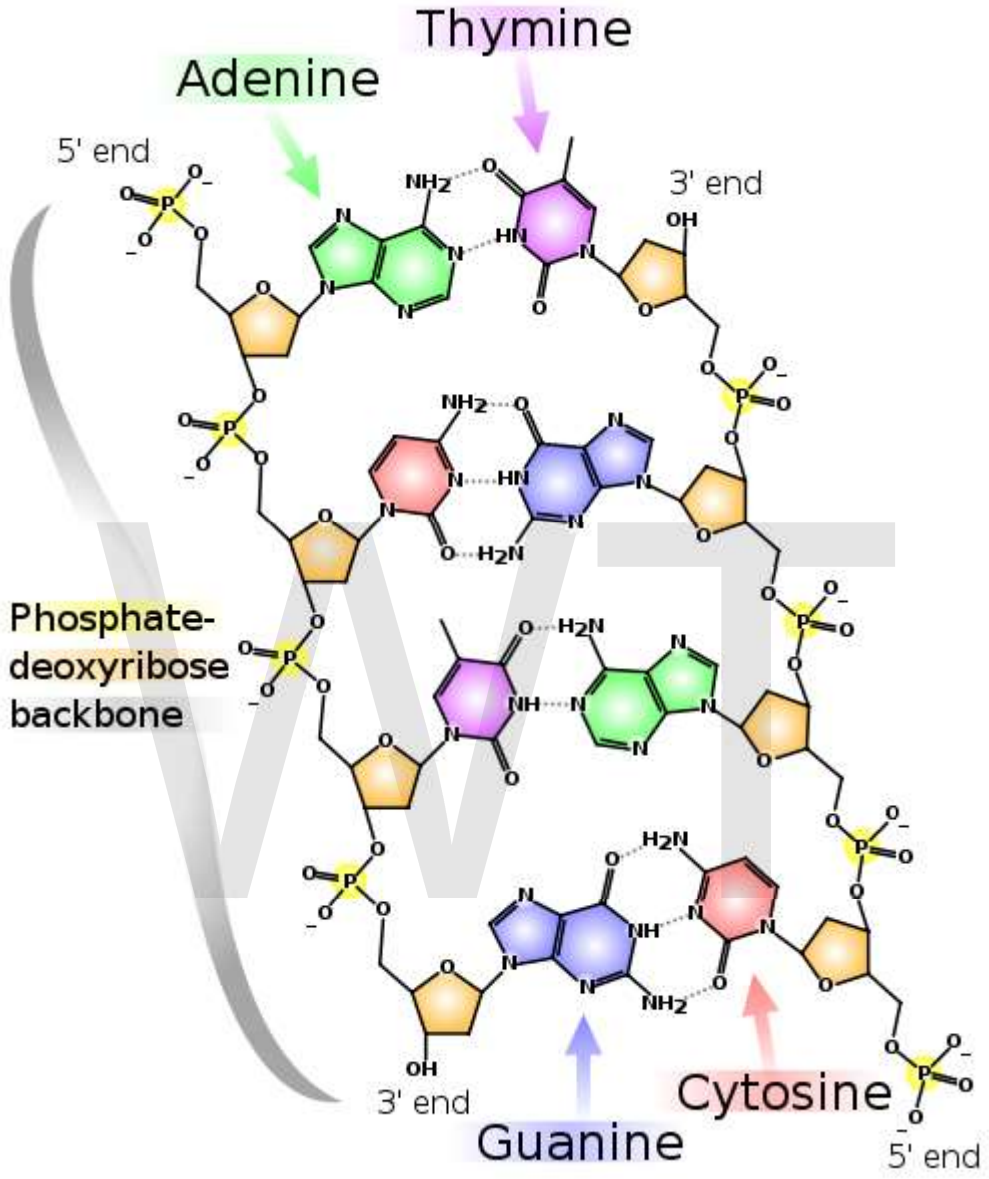
Gene



This stylistic diagram shows a gene in relation to the double helix structure of DNA and to a chromosome (right). The chromosome is X-shaped because it is dividing. Introns are regions often found in eukaryote genes that are removed in the splicing process (after the DNA is transcribed into RNA): Only the exons encode the protein. This diagram labels a region of only 50 or so bases as a gene. In reality, most genes are hundreds of times larger.

A **gene** is a unit of heredity in a living organism. It normally resides on some stretches of DNA and RNA that codes for a type of protein or for an RNA chain that has a function in the organism. Living things depend on genes, as they specify all proteins and functional RNA chains. Genes hold the information to build and maintain an organism's cells and pass genetic traits to offspring, although some organelles (e.g. mitochondria) are self-replicating and are not coded for by the organism's DNA. All organisms have many genes corresponding to many different biological traits, some of which are immediately visible, such as eye color or number of limbs, and some of which are not, such as blood type or

increased risk for specific diseases, or the thousands of basic biochemical processes that comprise life.



The chemical structure of a four-base fragment of a DNA double helix.

A modern working definition of a gene is "*a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and or other functional sequence regions* ". Colloquial usage of the term *gene* (e.g. "good genes", "hair color gene") may actually refer to an allele: a *gene* is the basic instruction, a sequence of nucleic acids (DNA or, in the case of certain viruses RNA), while an *allele* is one variant of that gene. Thus, when the mainstream press refers to "having" a "gene" for a specific trait, this is generally inaccurate. In most cases, all people would have a gene for the trait in question, but certain people will have a specific

allele of that gene, which results in the trait variant. In the simplest case, the phenotypic variation observed may be caused by a single letter of the genetic code - a single nucleotide polymorphism.

Physical definitions

RNA genes and genomes

When proteins are manufactured, the gene is first copied into RNA as an intermediate product. In other cases, the RNA molecules are the actual functional products. For example, RNAs known as ribozymes are capable of enzymatic function, and microRNA has a regulatory role. The DNA sequences from which such RNAs are transcribed are known as RNA genes.

Some viruses store their entire genomes in the form of RNA, and contain no DNA at all. Because they use RNA to store genes, their cellular hosts may synthesize their proteins as soon as they are infected and without the delay in waiting for transcription. On the other hand, RNA retroviruses, such as HIV, require the reverse transcription of their genome from RNA into DNA before their proteins can be synthesized. In 2006, French researchers came across a puzzling example of RNA-mediated inheritance in mouse. Mice with a loss-of-function mutation in the gene Kit have white tails. Offspring of these mutants can have white tails despite having only normal Kit genes. The research team traced this effect back to mutated Kit RNA. While RNA is common as genetic storage material in viruses, in mammals in particular RNA inheritance has been observed very rarely.

Functional structure of a gene

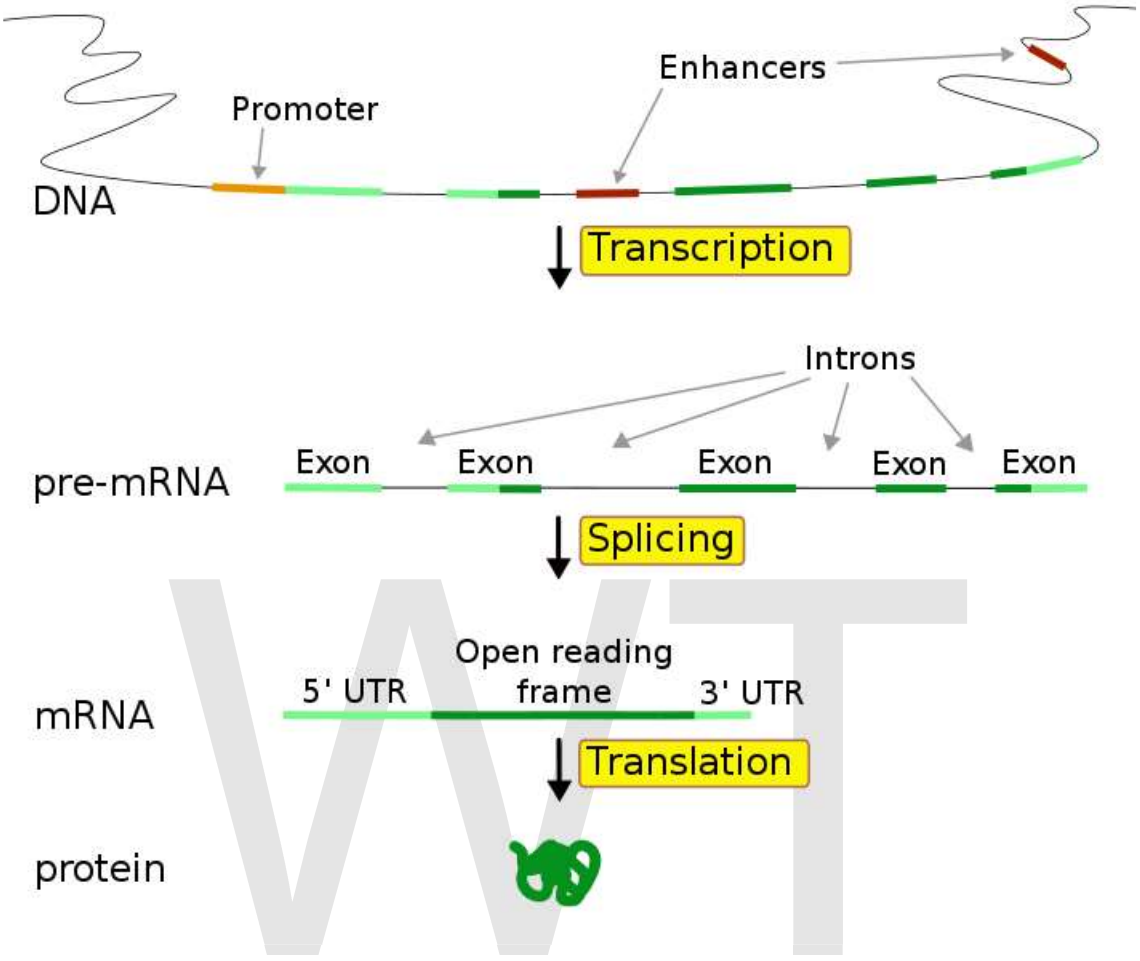


Diagram of the "typical" eukaryotic protein-coding **gene**. Promoters and enhancers determine what portions of the DNA will be transcribed into the precursor mRNA (pre-mRNA). The pre-mRNA is then spliced into messenger RNA (mRNA) which is later translated into protein.

The vast majority of living organisms encode their genes in long strands of DNA. DNA (deoxyribonucleic acid) consists of a chain made from four types of nucleotide subunits, each composed of: a five-carbon sugar (2'-deoxyribose), a phosphate group, and one of the four bases adenine, cytosine, guanine, and thymine. The most common form of DNA in a cell is in a double helix structure, in which two individual DNA strands twist around each other in a right-handed spiral. In this structure, the base pairing rules specify that guanine pairs with cytosine and adenine pairs with thymine. The base pairing between guanine and cytosine forms three hydrogen bonds, whereas the base pairing between adenine and thymine forms two hydrogen bonds. The two strands in a double helix must therefore be *complementary*, that is, their bases must align such that the adenines of one strand are paired with the thymines of the other strand, and so on.

Due to the chemical composition of the pentose residues of the bases, DNA strands have directionality. One end of a DNA polymer contains an exposed hydroxyl group on the deoxyribose; this is known as the 3' end of the molecule. The other end contains an exposed phosphate group; this is the 5' end. The directionality of DNA is vitally important to many cellular processes, since double helices are necessarily directional (a strand running 5'-3' pairs with a complementary strand running 3'-5'), and processes such as DNA replication occur in only one direction. All nucleic acid synthesis in a cell occurs in the 5'-3' direction, because new monomers are added via a dehydration reaction that uses the exposed 3' hydroxyl as a nucleophile.

The expression of genes encoded in DNA begins by transcribing the gene into RNA, a second type of nucleic acid that is very similar to DNA, but whose monomers contain the sugar ribose rather than deoxyribose. RNA also contains the base uracil in place of thymine. RNA molecules are less stable than DNA and are typically single-stranded. Genes that encode proteins are composed of a series of three-nucleotide sequences called codons, which serve as the *words* in the genetic *language*. The genetic code specifies the correspondence during protein translation between codons and amino acids. The genetic code is nearly the same for all known organisms.

All genes have regulatory regions in addition to regions that explicitly code for a protein or RNA product. A regulatory region shared by almost all genes is known as the promoter, which provides a position that is recognized by the transcription machinery when a gene is about to be transcribed and expressed. A gene can have more than one promoter, resulting in RNAs that differ in how far they extend in the 5' end. Although promoter regions have a consensus sequence that is the most common sequence at this position, some genes have "strong" promoters that bind the transcription machinery well, and others have "weak" promoters that bind poorly. These weak promoters usually permit a lower rate of transcription than the strong promoters, because the transcription machinery binds to them and initiates transcription less frequently. Other possible regulatory regions include enhancers, which can compensate for a weak promoter. Most regulatory regions are "upstream"—that is, before or toward the 5' end of the transcription initiation site. Eukaryotic promoter regions are much more complex and difficult to identify than prokaryotic promoters.

Many prokaryotic genes are organized into operons, or groups of genes whose products have related functions and which are transcribed as a unit. By contrast, eukaryotic genes are transcribed only one at a time, but may include long stretches of DNA called introns which are transcribed but never translated into protein (they are spliced out before translation). Splicing can also occur in prokaryotic genes, but is less common than in eukaryotes.

Chromosomes

The total complement of genes in an organism or cell is known as its genome, which may be stored on one or more chromosomes; the region of the chromosome at which a particular gene is located is called its locus. A chromosome consists of a single, very long

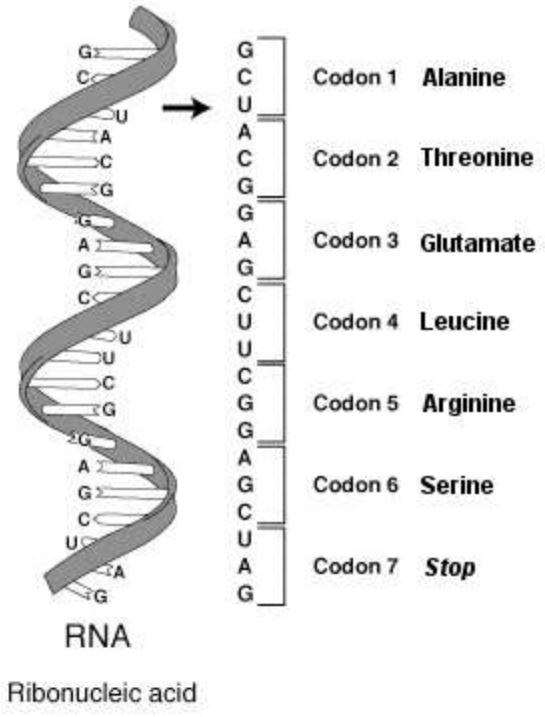
DNA helix on which thousands of genes are encoded. Prokaryotes—bacteria and archaea—typically store their genomes on a single large, circular chromosome, sometimes supplemented by additional small circles of DNA called plasmids, which usually encode only a few genes and are easily transferable between individuals. For example, the genes for antibiotic resistance are usually encoded on bacterial plasmids and can be passed between individual cells, even those of different species, via horizontal gene transfer. Although some simple eukaryotes also possess plasmids with small numbers of genes, the majority of eukaryotic genes are stored on multiple linear chromosomes, which are packed within the nucleus in complex with storage proteins called histones. The manner in which DNA is stored on the histone, as well as chemical modifications of the histone itself, are regulatory mechanisms governing whether a particular region of DNA is accessible for gene expression. The ends of eukaryotic chromosomes are capped by long stretches of repetitive sequences called telomeres, which do not code for any gene product but are present to prevent degradation of coding and regulatory regions during DNA replication. The length of the telomeres tends to decrease each time the genome is replicated in preparation for cell division; the loss of telomeres has been proposed as an explanation for cellular senescence, or the loss of the ability to divide, and by extension for the aging process in organisms.

Whereas the chromosomes of prokaryotes are relatively gene-dense, those of eukaryotes often contain so-called "junk DNA", or regions of DNA that serve no obvious function. Simple single-celled eukaryotes have relatively small amounts of such DNA, whereas the genomes of complex multicellular organisms, including humans, contain an absolute majority of DNA without an identified function. However it now appears that, although protein-coding DNA makes up barely 2% of the human genome, about 80% of the bases in the genome may be expressed, so the term "junk DNA" may be a misnomer.

Gene expression

In all organisms, there are two major steps separating a protein-coding gene from its protein: First, the DNA on which the gene resides must be *transcribed* from DNA to messenger RNA (mRNA); and, second, it must be *translated* from mRNA to protein. RNA-coding genes must still go through the first step, but are not translated into protein. The process of producing a biologically functional molecule of either RNA or protein is called gene expression, and the resulting molecule itself is called a gene product.

Genetic code



Schematic diagram of a single-stranded RNA molecule illustrating the position of three-base codons.

The genetic code is the set of rules by which a gene is translated into a functional protein. Each gene consists of a specific sequence of nucleotides encoded in a DNA (or sometimes RNA) strand; a correspondence between nucleotides, the basic building blocks of genetic material, and amino acids, the basic building blocks of proteins, must be established for genes to be successfully translated into functional proteins. Sets of three nucleotides, known as codons, each correspond to a specific amino acid or to a signal; three codons are known as "stop codons" and, instead of specifying a new amino acid, alert the translation machinery that the end of the gene has been reached. There are 64 possible codons (four possible nucleotides at each of three positions, hence 4^3 possible codons) and only 20 standard amino acids; hence the code is redundant and multiple codons can specify the same amino acid. The correspondence between codons and amino acids is nearly universal among all known living organisms.

Transcription

The process of genetic transcription produces a single-stranded RNA molecule known as messenger RNA, whose nucleotide sequence is complementary to the DNA from which it was transcribed. The DNA strand whose sequence matches that of the RNA is known as the coding strand and the strand from which the RNA was synthesized is the template strand. Transcription is performed by an enzyme called an RNA polymerase, which reads the template strand in the 3' to 5' direction and synthesizes the RNA from 5' to 3'. To

initiate transcription, the polymerase first recognizes and binds a promoter region of the gene. Thus a major mechanism of gene regulation is the blocking or sequestering of the promoter region, either by tight binding by repressor molecules that physically block the polymerase, or by organizing the DNA so that the promoter region is not accessible.

In prokaryotes, transcription occurs in the cytoplasm; for very long transcripts, translation may begin at the 5' end of the RNA while the 3' end is still being transcribed. In eukaryotes, transcription necessarily occurs in the nucleus, where the cell's DNA is sequestered; the RNA molecule produced by the polymerase is known as the primary transcript and must undergo post-transcriptional modifications before being exported to the cytoplasm for translation. The splicing of introns present within the transcribed region is a modification unique to eukaryotes; alternative splicing mechanisms can result in mature transcripts from the same gene having different sequences and thus coding for different proteins. This is a major form of regulation in eukaryotic cells.

Translation

Translation is the process by which a mature mRNA molecule is used as a template for synthesizing a new protein. Translation is carried out by ribosomes, large complexes of RNA and protein responsible for carrying out the chemical reactions to add new amino acids to a growing polypeptide chain by the formation of peptide bonds. The genetic code is read three nucleotides at a time, in units called codons, via interactions with specialized RNA molecules called transfer RNA (tRNA). Each tRNA has three unpaired bases known as the anticodon that are complementary to the codon it reads; the tRNA is also covalently attached to the amino acid specified by the complementary codon. When the tRNA binds to its complementary codon in an mRNA strand, the ribosome ligates its amino acid cargo to the new polypeptide chain, which is synthesized from amino terminus to carboxyl terminus. During and after its synthesis, the new protein must fold to its active three-dimensional structure before it can carry out its cellular function.

DNA replication and inheritance

The growth, development, and reproduction of organisms relies on cell division, or the process by which a single cell divides into two usually identical daughter cells. This requires first making a duplicate copy of every gene in the genome in a process called DNA replication. The copies are made by specialized enzymes known as DNA polymerases, which "read" one strand of the double-helical DNA, known as the template strand, and synthesize a new complementary strand. Because the DNA double helix is held together by base pairing, the sequence of one strand completely specifies the sequence of its complement; hence only one strand needs to be read by the enzyme to produce a faithful copy. The process of DNA replication is semiconservative; that is, the copy of the genome inherited by each daughter cell contains one original and one newly synthesized strand of DNA.

After DNA replication is complete, the cell must physically separate the two copies of the genome and divide into two distinct membrane-bound cells. In prokaryotes - bacteria and

archaea - this usually occurs via a relatively simple process called binary fission, in which each circular genome attaches to the cell membrane and is separated into the daughter cells as the membrane invaginates to split the cytoplasm into two membrane-bound portions. Binary fission is extremely fast compared to the rates of cell division in eukaryotes. Eukaryotic cell division is a more complex process known as the cell cycle; DNA replication occurs during a phase of this cycle known as S phase, whereas the process of segregating chromosomes and splitting the cytoplasm occurs during M phase. In many single-celled eukaryotes such as yeast, reproduction by budding is common, which results in asymmetrical portions of cytoplasm in the two daughter cells.

Molecular inheritance

The duplication and transmission of genetic material from one generation of cells to the next is the basis for molecular inheritance, and the link between the classical and molecular pictures of genes. Organisms inherit the characteristics of their parents because the cells of the offspring contain copies of the genes in their parents' cells. In asexually reproducing organisms, the offspring will be a genetic copy or clone of the parent organism. In sexually reproducing organisms, a specialized form of cell division called meiosis produces cells called gametes or germ cells that are haploid, or contain only one copy of each gene. The gametes produced by females are called eggs or ova, and those produced by males are called sperm. Two gametes fuse to form a fertilized egg, a single cell that once again has a diploid number of genes—each with one copy from the mother and one copy from the father.

During the process of meiotic cell division, an event called genetic recombination or *crossing-over* can sometimes occur, in which a length of DNA on one chromatid is swapped with a length of DNA on the corresponding sister chromatid. This has no effect if the alleles on the chromatids are the same, but results in reassortment of otherwise linked alleles if they are different. The Mendelian principle of independent assortment asserts that each of a parent's two genes for each trait will sort independently into gametes; which allele an organism inherits for one trait is unrelated to which allele it inherits for another trait. This is in fact only true for genes that do not reside on the same chromosome, or are located very far from one another on the same chromosome. The closer two genes lie on the same chromosome, the more closely they will be associated in gametes and the more often they will appear together; genes that are very close are essentially never separated because it is extremely unlikely that a crossover point will occur between them. This is known as genetic linkage.

History

The notion of a gene is evolving with the science of genetics, which began when Gregor Mendel noticed that biological variations are inherited from parent organisms as specific, discrete traits. The biological entity responsible for defining traits was later termed a *gene*, but the biological basis for inheritance remained unknown until DNA was identified as the genetic material in the 1940s. Prior to Mendel's work, the dominant theory of heredity was one of blending inheritance, which proposes that the traits of the

parents blend or mix in a smooth, continuous gradient in the offspring. Although Mendel's work was largely unrecognized after its first publication in 1866, it was rediscovered in 1900 by three European scientists, Hugo de Vries, Carl Correns, and Erich von Tschermak, who had reached similar conclusions from their own research. However, these scientists were not yet aware of the identity of the 'discrete units' on which genetic material resides.

The existence of genes was first suggested by Gregor Mendel (1822–1884), who, in the 1860s, studied inheritance in peaplants (*Pisum sativum*) and hypothesized a factor that conveys traits from parent to offspring. He spent over 10 years of his life on one experiment. Although he did not use the term *gene*, he explained his results in terms of inherited characteristics. Mendel was also the first to hypothesize independent assortment, the distinction between dominant and recessive traits, the distinction between a heterozygote and homozygote, and the difference between what would later be described as genotype (the genetic material of an organism) and phenotype (the visible traits of that organism).

Charles Darwin used the term Gemmule to describe a microscopic unit of inheritance, and what would later become known as Chromosomes had been observed separating out during cell division by Wilhelm Hofmeister as early as 1848. The idea that chromosomes are the carriers of inheritance was expressed in 1883 by Wilhelm Roux. Darwin also coined the word *pangenesis* by (1868). The word pangenesis is made from the Greek words *pan* (a prefix meaning "whole", "encompassing") and *genesis* ("birth") or *genos* ("origin").

Mendel's concept was given a name by Hugo de Vries in 1889, in his book *Intracellular Pangenesis*; although probably unaware of Mendel's work at the time, he coined the term "pangen" for "the smallest particle [representing] one hereditary characteristic". Danish botanist Wilhelm Johannsen coined the word "gene" ("gen" in Danish and German) in 1909 to describe the fundamental physical and functional units of heredity, while the related word genetics was first used by William Bateson in 1905. He derived the word from de Vries' "pangen". In the early 1900s, Mendel's work received renewed attention from scientists. In 1910, Thomas Hunt Morgan showed that genes reside on specific chromosomes. He later showed that genes occupy specific locations on the chromosome. With this knowledge, Morgan and his students began the first chromosomal map of the fruit fly *Drosophila*. In 1928, Frederick Griffith showed that genes could be transferred. In what is now known as Griffith's experiment, injections into a mouse of a deadly strain of bacteria that had been heat-killed transferred genetic information to a safe strain of the same bacteria, killing the mouse.

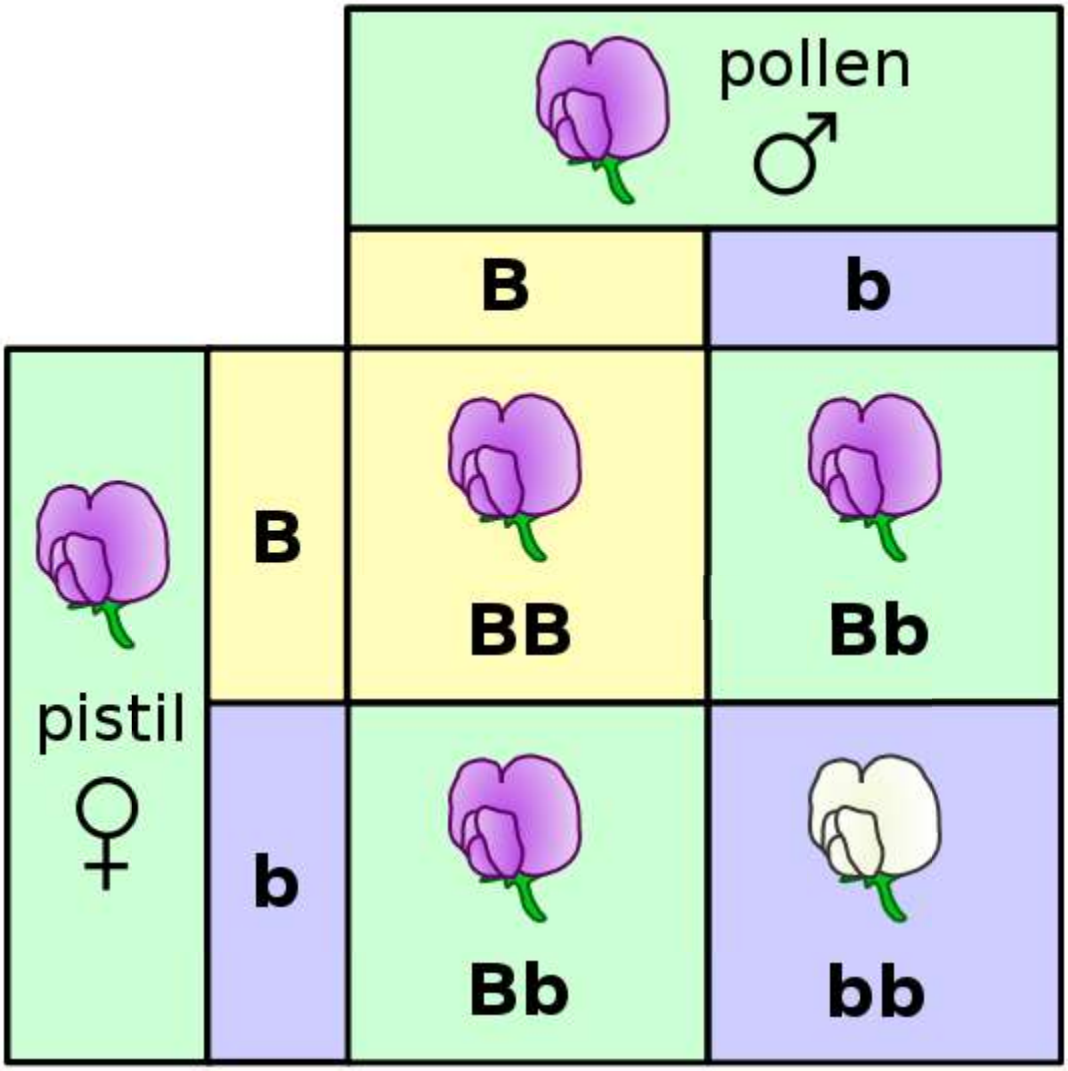
A series of subsequent discoveries led to the realization decades later that chromosomes within cells are the carriers of genetic material, and that they are made of DNA (deoxyribonucleic acid), a polymeric molecule found in all cells on which the 'discrete units' of Mendelian inheritance are encoded. In 1941, George Wells Beadle and Edward Lawrie Tatum showed that mutations in genes caused errors in specific steps in metabolic pathways. This showed that specific genes code for specific proteins, leading to the "one

gene, one enzyme" hypothesis. Oswald Avery, Colin Munro MacLeod, and Maclyn McCarty showed in 1944 that DNA holds the gene's information. In 1953, James D. Watson and Francis Crick demonstrated the molecular structure of DNA. Together, these discoveries established the central dogma of molecular biology, which states that proteins are translated from RNA which is transcribed from DNA. This dogma has since been shown to have exceptions, such as reverse transcription in retroviruses.

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. Richard J. Roberts and Phillip Sharp discovered in 1977 that genes can be split into segments. This led to the idea that one gene can make several proteins. Recently (as of 2003–2006), biological results let the notion of gene appear more slippery. In particular, genes do not seem to sit side by side on DNA like discrete beads. Instead, regions of the DNA producing distinct proteins may overlap, so that the idea emerges that "genes are one long continuum". It was first hypothesized in 1986 by Walter Gilbert that neither DNA nor protein would be required in such a primitive system as that of a very early stage of the earth if RNA could perform as simply a catalyst and genetic information storage processor .

The modern study of genetics at the level of DNA is known as molecular genetics and the synthesis of molecular genetics with traditional Darwinian evolution is known as the modern evolutionary synthesis.

Mendelian inheritance and classical genetics



Crossing between two pea plants heterozygous for purple (B, dominant) and white (b, recessive) blossoms

According to the theory of Mendelian inheritance, variations in phenotype—the observable physical and behavioral characteristics of an organism—are due to variations in genotype, or the organism's particular set of genes, each of which specifies a particular trait. Different forms of a gene, which may give rise to different phenotypes, are known as alleles. Organisms such as the pea plants Mendel worked on, along with many plants and animals, have two alleles for each trait, one inherited from each parent. Alleles may be dominant or recessive; dominant alleles give rise to their corresponding phenotypes when paired with any other allele for the same trait, whereas recessive alleles give rise to their corresponding phenotype only when paired with another copy of the same allele. For example, if the allele specifying tall stems in pea plants is dominant over the allele specifying short stems, then pea plants that inherit one tall allele from one parent and one

short allele from the other parent will also have tall stems. Mendel's work found that alleles assort independently in the production of gametes, or germ cells, ensuring variation in the next generation.

Mutation

DNA replication is for the most part extremely accurate, with an error rate per site of around 10^{-6} to 10^{-10} in eukaryotes. Rare, spontaneous alterations in the base sequence of a particular gene arise from a number of sources, such as errors in DNA replication and the aftermath of DNA damage. These errors are called mutations. The cell contains many DNA repair mechanisms for preventing mutations and maintaining the integrity of the genome; however, in some cases—such as breaks in both DNA strands of a chromosome—repairing the physical damage to the molecule is a higher priority than producing an exact copy. Due to the degeneracy of the genetic code, some mutations in protein-coding genes are *silent*, or produce no change in the amino acid sequence of the protein for which they code; for example, the codons UCU and UUC both code for serine, so the U↔C mutation has no effect on the protein. Mutations that do have phenotypic effects are most often neutral or deleterious to the organism, but sometimes they confer benefits to the organism's fitness.

Mutations propagated to the next generation lead to variations within a species' population. Variants of a single gene are known as alleles, and differences in alleles may give rise to differences in traits. Although it is rare for the variants in a single gene to have clearly distinguishable phenotypic effects, certain well-defined traits are in fact controlled by single genetic loci. A gene's most common allele is called the wild type allele, and rare alleles are called mutants. However, this does not imply that the wild-type allele is the ancestor from which the mutants are descended.

Genome

Chromosomal organization

The total complement of genes in an organism or cell is known as its genome. In prokaryotes, the vast majority of genes are located on a single chromosome of circular DNA, while eukaryotes usually possess multiple individual linear DNA helices packed into dense DNA-protein complexes called chromosomes. Genes that appear together on one chromosome of one species may appear on separate chromosomes in another species. Many species carry more than one copy of their genome within each of their somatic cells. Cells or organisms with only one copy of each chromosome are called haploid; those with two copies are called diploid; and those with more than two copies are called polyploid. The copies of genes on the chromosomes are not necessarily identical. In sexually reproducing organisms, one copy is normally inherited from each parent.

Number of genes

Early estimates of the number of human genes that used expressed sequence tag data put it at 50 000–100 000. Following the sequencing of the human genome and other genomes, it has been found that rather few genes (~20 000 in human, mouse and fly, ~13 000 in roundworm, >46 000 in rice) encode all the proteins in an organism. These protein-coding sequences make up 1–2% of the human genome. A large part of the genome is transcribed however, to introns, retrotransposons and seemingly a large array of noncoding RNAs. Total number of proteins (the Earth's proteome) is estimated to be 5 million sequences.

Genetic and genomic nomenclature

Gene nomenclature has been established by the HUGO Gene Nomenclature Committee (HGNC) for each known human gene in the form of an approved gene name and symbol (short-form abbreviation). All approved symbols are stored in the HGNC Database. Each symbol is unique and each gene is only given one approved gene symbol. This also facilitates electronic data retrieval from publications. In preference each symbol maintains parallel construction in different members of a gene family and can be used in other species, especially the mouse.

Evolutionary concept of a gene

George C. Williams first explicitly advocated the gene-centric view of evolution in his 1966 book *Adaptation and Natural Selection*. He proposed an evolutionary concept of gene to be used when we are talking about natural selection favoring some genes. The definition is: "that which segregates and recombines with appreciable frequency." According to this definition, even an asexual genome could be considered a gene, insofar that it have an appreciable permanency through many generations.

The difference is: the molecular gene *transcribes* as a unit, and the evolutionary gene *inherits* as a unit.

Richard Dawkins' books *The Selfish Gene* (1976) and *The Extended Phenotype* (1982) defended the idea that the gene is the only replicator in living systems. This means that only genes transmit their structure largely intact and are potentially immortal in the form of copies. So, genes should be the unit of selection. In *The Selfish Gene* Dawkins attempts to redefine the word 'gene' to mean "an inheritable unit" instead of the generally accepted definition of "a section of DNA coding for a particular protein". In *River Out of Eden*, Dawkins further refined the idea of gene-centric selection by describing life as a river of compatible genes flowing through geological time. Scoop up a bucket of genes from the river of genes, and we have an organism serving as temporary bodies or survival machines. A river of genes may fork into two branches representing two non-interbreeding species as a result of geographical separation.

Gene targeting and implications

Gene targeting is commonly referred to techniques for altering or disrupting mouse genes and provides the mouse models for studying the roles of individual genes in embryonic development, human disorders, aging and diseases. The mouse models, where one or more of its genes are deactivated or made inoperable, are called knockout mice. Since the first reports in which homologous recombination in embryonic stem cells was used to generate gene-targeted mice, gene targeting has proven to be a powerful means of precisely manipulating the mammalian genome, producing at least ten thousand mutant mouse strains and it is now possible to introduce mutations that can be activated at specific time points, or in specific cells or organs, both during development and in the adult animal.

Gene targeting strategies have been expanded to all kinds of modifications, including point mutations, isoform deletions, mutant allele correction, large pieces of chromosomal DNA insertion and deletion, tissue specific disruption combined with spatial and temporal regulation and so on. It is predicted that the ability to generate mouse models with predictable phenotypes will have a major impact on studies of all phases of development, immunology, neurobiology, oncology, physiology, metabolism, and human diseases. Gene targeting is also in theory applicable to species from which totipotent embryonic stem cells can be established, and therefore may offer a potential to the improvement of domestic animals and plants.

Changing concept

The concept of the gene has changed considerably. From the original definition of a "unit of inheritance", the term evolved to mean a DNA-based unit that can exert its effects on the organism through RNA or protein products. It was also previously believed that one gene makes one protein; this concept was overthrown by the discovery of alternative splicing and trans-splicing.

The definition of a gene is still changing. The first cases of RNA-based inheritance have been discovered in mammals. Evidence is also accumulating that the control regions of a gene do not necessarily have to be close to the coding sequence on the linear molecule or even on the same chromosome. Spilianakis and colleagues discovered that the promoter region of the interferon-gamma gene on chromosome 10 and the regulatory regions of the T(H)2 cytokine locus on chromosome 11 come into close proximity in the nucleus possibly to be jointly regulated.

The concept that genes are clearly delimited is also being eroded. There is evidence for fused proteins stemming from two adjacent genes that can produce two separate protein products. While it is not clear whether these fusion proteins are functional, the phenomenon is more frequent than previously thought. Even more ground-breaking than the discovery of fused genes is the observation that some proteins can be composed of exons from far away regions and even different chromosomes. This new data has led to an updated, and probably tentative, definition of a gene as "a union of genomic sequences

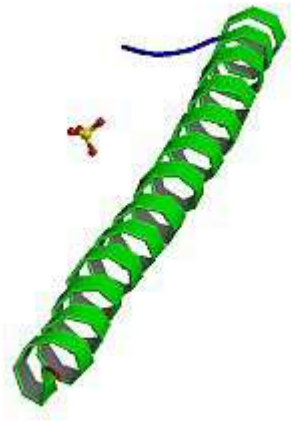
encoding a coherent set of potentially overlapping functional products." This new definition categorizes genes by functional products, whether they be proteins or RNA, rather than specific DNA loci; all regulatory elements of DNA are therefore classified as *gene-associated* regions.

Chapter- 2

Adenomatous Polyposis Coli and APC/C Activator Protein CDH1

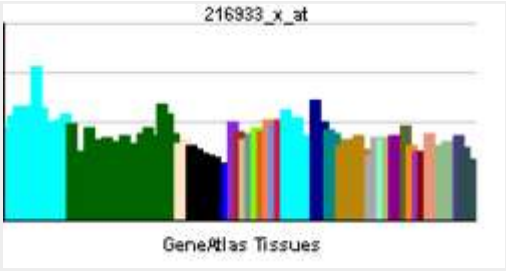
Adenomatous polyposis coli

Adenomatous polyposis coli



PDB rendering based on 1deb.

Available structures	
Identifiers	
Symbols	APC; DP2; DP2.5; DP3; FAP; FPC; GS
External IDs	OMIM: 175100 MGI: 88039 HomoloGene: 30950 GeneCards: APC Gene
Gene Ontology	

RNA expression pattern


Orthologs		
Species	**Human**	**Mouse**
Entrez	324	11789
Ensembl	ENSG00000134982	ENSMUSG00000005871
UniProt	P25054	Q8C9I9
RefSeq (mRNA)	NM_000038	XM_622559
RefSeq (protein)	NP_000029	XP_622559
Location (UCSC)	Chr 5: 112.1 - 112.21 Mb	Chr 18: 34.35 - 34.44 Mb

Adenomatous polyposis coli (APC) also known as **deleted in polyposis 2.5** (DP2.5) is a protein that in humans is encoded by the *APC* gene. Mutations in the *APC* gene may result in colorectal cancer.

APC is classified as a tumor suppressor gene. Tumor suppressor genes prevent the uncontrolled growth of cells that may result in cancerous tumors. The protein made by the *APC* gene plays a critical role in several cellular processes that determine whether a cell may develop into a tumor. The APC protein helps control how often a cell divides, how it attaches to other cells within a tissue, or whether a cell moves within or away from a tissue. This protein also helps ensure that the chromosome number in cells produced through cell division is correct. The APC protein accomplishes these tasks mainly through association with other proteins, especially those that are involved in cell attachment and signaling. The activity of one protein in particular, beta-catenin, is controlled by the APC protein (see: Wnt signaling pathway). Regulation of beta-catenin prevents genes that stimulate cell division from being turned on too often and prevents cell overgrowth.

The human *APC* gene is located on the long (q) arm of chromosome 5 between positions 21 and 22, from base pair 112,118,468 to base pair 112,209,532. The *APC* gene has been shown to contain an internal ribosome entry site. *APC* orthologs have also been identified in all mammals for which complete genome data are available.

The full-length human protein comprises 2843 amino acids with a (predicted) molecular mass of 311646 Da. Most domains of this protein are solved structurally and exhibit high intrinsic disorder and flexibility as a monomer, and a low content of stable secondary structure. Thus it is a member of the intrinsically unstructured proteins. Little is known about the *in vivo* full-length unfolded protein.

Role in cancer

The most common mutation in colon cancer is inactivation of APC. When APC does not have an inactivating mutation, beta catenin does. These mutations can be inherited, or arise sporadically, often as the result of mutations in other genes that produce chromosomal instability. A mutation on APC or β-catenin must be followed by other mutations to become cancerous; however, in carriers of an APC inactivating mutations, the risk of colorectal cancer by age 40 is almost 100%.

Familial adenomatous polyposis (FAP) is caused by mutations in the APC gene. More than 800 mutationsin the APC gene have been identified in families with classic and attenuated types of familial adenomatous polyposis. Most of these mutations cause the production of an APC protein that is abnormally short and nonfunctional. This short protein cannot suppress the cellular overgrowth that leads to the formation of polyps, which can become cancerous. The most common mutation in familial adenomatous polyposis is a deletion of five bases (the building blocks of DNA) in the APC gene. This mutation changes the sequence of amino acids (the building material of proteins) in the resulting APC protein beginning at position 1309.

Another mutation is carried by approximately 6 percent of people of Ashkenazi (eastern and central European) Jewish heritage. This mutation results in the substitution of the amino acid lysine for isoleucine at position 1307 in the APC protein (also written as I1307K or Ile1307Lys). This change was initially thought to be harmless, but has recently been shown to be associated with a 10 to 20 percent increased risk of colon cancer.

Regulation of proliferation

The (Adenomatous Polyposis Coli) APC protein normally builds a complex with glycogen synthase kinase 3-beta(GSK-3β) and axin via interactions with the 20 AA and SAMP repeats. This complex is then able to bind β- catenins in the cytoplasm, that have dissociated from adherens contacts between cells. With the help of casein kinase 1 (CK1), which carries out an initial phosphorylation of β-catenin, GSK-3β is able to phosphorylate β-catenin a second time. This targets β-catenin for ubiquitination and degradation by cellular proteosomes. This prevents it from translocating into the nucleus, where it acts as a transcription factor for proliferation genes. APC is also thought to be targeted to microtubules via the PDZ binding domain, stabilizing them. The deactivation of the APC protein can take place after certain chain reactions in the cytoplasm are started, e.g. through the Wnt signals that destroy the conformation of the complex. In the nucleus it complexes with legless/BCL9, TCF, and Pygo and begins function of an RNA polymerase but for oncogenes.

Mutations

Mutations in APC often occur early on in cancers such as colon cancer. Patients with familial adenomatous polyposis (FAP) have germline mutations, with 95% being nonsense/frameshift mutations leading to premature stop codons. 33% of mutations occur between amino acids 1061-1309. In somatic mutations, over 60% occur within a mutation cluster region (1286-1513), causing loss of axin binding sites in all but 1 of the 20AA repeats. Mutations in APC lead to loss of β-catenin regulation, altered cell migration and chromosome instability.

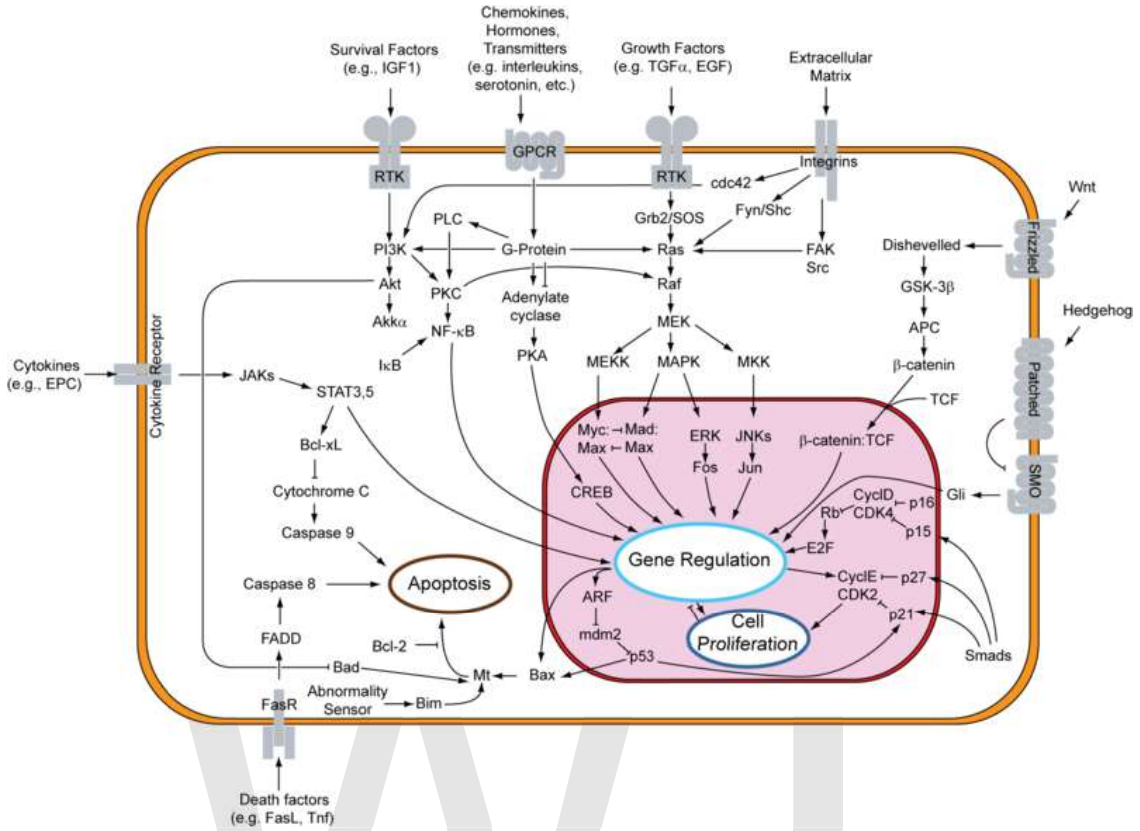
Neurological role

Rosenberg *et al.* found that APC directs cholinergic synapse assembly between neurons, a finding with implications for autonomic neuropathies, for Alzheimer's disease, for age-related hearing loss, and for some forms of epilepsy and schizophrenia.

Interactions

APC (gene) has been shown to interact with

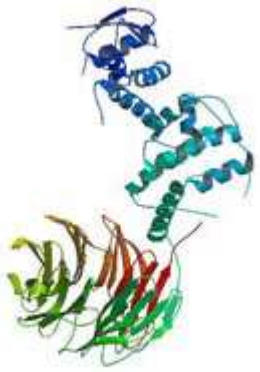
- ARHGEF4,
- AXIN1,
- BUB1,
- CTNNB1,
- CSNK2B,
- CSNK2A1,
- CTNNA1,
- DLG3,
- KIFAP3,
- MAPRE2,
- JUP,
- SIAH1,
- TFAP2A,
- TUBA4A, and
- XPO1.



Overview of signal transduction pathways involved in apoptosis.

APC/C activator protein CDH1

SCF(Fbw7) ubiquitin ligase complex



Identifiers	
Symbol	Cdh1, Hct1
PDB	2ovq
UniProt	P53197

Cdh1 is one of the substrate adaptor protein of the anaphase-promoting complex (APC) in the budding yeast *Saccharomyces cerevisiae*. Functioning as an activator of the APC/C, Cdh1 regulates the activity and substrate specificity of this E3 ubiquitin ligase.

Introduction

Cdh1 plays a pivotal role in controlling cell division at the end of mitosis (telophase) and in the subsequent G1 phase of cell cycle: By recognizing and binding proteins (like mitotic cyclins) which contain a destruction box (D-box) and an additional degradation signal (KEN box), Cdh1 recruits them in a C-box-dependent mechanism to the APC for ubiquination and subsequent proteolysis. Cdh1 is required for the exit of mitosis. Furthermore, it is thought to be a possible target of a BUB2-dependent spindle checkpoint pathway.

Function

The anaphase-promoting complex/cylosome (APC/c) is an ubiquitin E3-ligase complex. Once activated it attaches chains of ubiquitin molecules to its target substrates. These chains are recognised and the substrate is degraded by the Proteasome. Cdh1 is one of the co- activator proteins of APC/c and therefore contributes to the regulation of protein degradation, by providing substrate specificity to the E3-ligase in a cell-cycle regulated manner.

Cdh1 can exist in several forms. It can be phosphorylated by CDKs, which inactivates it and it can be dephosphorylated by Cdc14. In the dephosphorylated form it can interact with APC/c and build the active ligase APC^{Cdh1}.

Suppression of Cdh1 by RNA interference leads to an aberrant accumulation of APC^{Cdh1} target proteins, such as cyclin A and B, the kinase AuroraB, Plk1, Skp2 and Cdc20, another APC/c co- activator.

Stabilising G1-Phase

The main function of Cdh1 is to suppress the re-accumulation of mitotic cyclins and other cell cycle determinants and therefore stabilising the G1-Phase. In early mitose stage it is inactive and only becomes active in the transition from late mitosis to G1.

During the cell cycle Cdk gets activated through cyclins, this leads to the mitotic entry and promotes APC^{Cdc20} activation. APC^{Cdc20} degrades the cyclins, this and the activation of Cdc14 leads to the creation of APC^{Cdh1}. APC^{Cdh1} keeps the cyclin concentration low and the Cdk inactive that maintains the G1-Phase.

G1/S transition

APC^{Cdh1} is thought to prevent premature S-Phase entry by degrading mitotic cyclins in G1 and regulate processes unrelated to the cell cycle. To enter S-Phase APC^{Cdh1} must be inactivated. This is made through degradation of the complex and through posphorylation of Cdh1.

Exit from Mitosis

One characteristic of budding yeast cells exit from mitosis after chromosome segregation is the removal of the mitotic determinants. This requires the inactivation of mitotic CDKs which are inactivated through ubiquitin-dependent pathways. The protein phosphatase Cdc14 dephosphorylates Cdh1 and therefore activates APC^{Cdh1}. As a result the concentration of many APC^{Cdh1} substrates drops down and the cell exit from mitosis.

Cdh1 functions as a tumour suppressor

Cdh1-deficient cells can proliferate but accumulate mitotic errors and have difficulties with cytokinesis.

It has been shown that APC^{Cdh1}-mediated degradation of PIk1 plays an important role in preventing mitosis in cells that have DNA-damage. In healthy cells Cdh1 stays inactive from late G1 to early mitosis. It stays inactive in early mitosis and only becomes active in the transition from late mitosis to G1. A cell that suffers from DNA-damage shows an active Cdh1 already in late G1 and therefore blocks the mitotic entry.

One substrate of APC^{Cdh1} is the transcription factor Ets2, which is activated by the Ras-Raf-MAPK signalling pathway and induces the expression of cyclin D1. This pathway stimulates cell proliferation. It was shown that an increased expression of Ets2 can be associated with various cancer types, in the likes of cervical cancer or oesophageal squamous cell carcinoma.

Fuction of Cdh1 in non- dividing cells

It was shown that APC^{Cdh1} is active in adult brain and liver tissues. It seems that the complex has a function in axongrowth, morphologie and plasticity of synapses as well as in learning and memory.

Structure

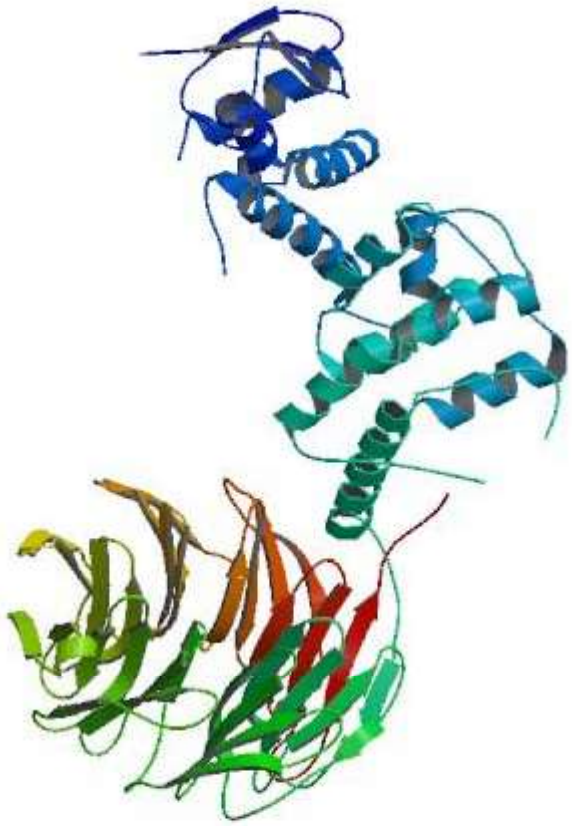


Fig. 1 There is no structure resolved for Cdh1 of *Saccharomyces cerevisiae*. There is a model based on template pdb2ovq, which shows the SCF(Fbw7) ubiquitin ligase complex. Fbw7 is also a WD repeat protein like Cdh1.

The following structural informations are based on the cdh1 protein of *Saccharomyces cerevisiae* also named Hct1. Cdh1 is a cdc20 homolog and is Frizzy-related (*Drosophila*). The protein sequence of cdh1 consists of 566 amino acids and has a molecular weight of 62.8 kDa. Cdh1 comprises different domains important for its proper function, when it interacts with the APC/c complex and the various substrates.

Activation and APC/c Binding

In the N-terminal region at amino acid position 55-61 the cdh1 protein contains a C-Box motif , which is required for the association with the APC/c complex. Especially the residue R56 seems to be important for the binding to APC/c *in vitro* and Cdh1 function *in vivo*.

Cdh1 contains multiple phosphorylation sites for the kinase cdc28. When cdh1 is hyperphosphorytaled, the association of cdh1 to the APC/c is blocked, thus leading to the inactive form of cdh1. Activation can be induced by dephosphorylation through the phosphatase cdc14, which leads to the binding of cdh1 to the APC/c.

Cdh1 as well includes a poly-Ser in the N-terminal region from residue 32-38. In general serine, threonine and tyrosine side chains can act as phosphorylation sites for posttranslational modification. In the cdh1 protein amino acid modifications can be found at residue 156 being a phosphoserine and at residue 157 being a phosphothreonine.

Cdh1 also contains a C-terminal Ile-Arg (IR) dipeptide motif at residue 565 and 566, which is suggested to bind to the Cdc27 subunit of APC.

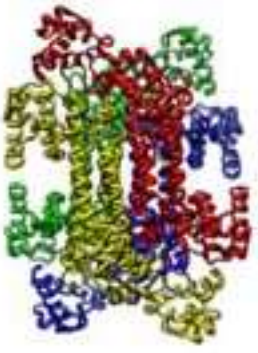
Substrate Binding

Cdh1 has 7 WD repeats, which are located between the middle of the protein and the C-terminal end. They have a conserved core length of about 38 to 43 amino acids, which in general end with tryptophan-aspartic acid (WD). WD repeat proteins are assumed to form a circularized beta propeller structure, which is thought to be essential for the biological function. The WD repeats in cdh1 are suspected to be the binding sites for the APC/c substrates. Thus cdh1 seems to be a sort of linker between the APC/c complex and the substrates. The APC/c substrates contain a D-Box and/or a KEN-Box , which are important for the interaction with cdh1.

Chapter- 3

Argininosuccinate Lyase

Argininosuccinate lyase

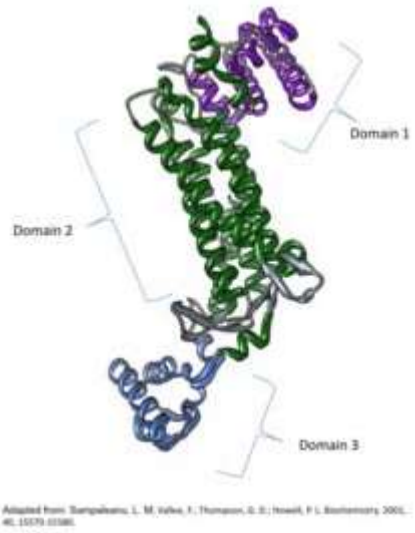


Crystal structure of duck argininosuccinate lyase with bound argininosuccinate.

Identifiers	
EC number	4.3.2.1
CAS number	9027-34-3
Databases	
IntEnz	IntEnz view
BRENDA	BRENDA entry
ExPASy	NiceZyme view
KEGG	KEGG entry
MetaCyc	metabolic pathway
PRIAM	profile
PDB structures	RCSB PDB PDBe PDBsum

Gene Ontology	AmiGO / EGO

Argininosuccinate lyase



Crystallographic structure of the human ASL monomer with labeled domains.

Identifiers	
Symbol	ASL
Entrez	435
HUGO	746
OMIM	608310
RefSeq	NM_000048
UniProt	P04424
Other data	
EC number	4.3.2.1
Locus	Chr. 7 *pter-q22*

ASL (**argininosuccinate lyase**, also known as **argininosuccinase**) is an enzyme that catalyzes the reversible breakdown of argininosuccinate (ASA) producing the amino acid arginine and fumarate. Located in liver cytosol, ASL is the fourth enzyme of the urea cycle and involved in the biosynthesis of arginine in all species and the production of

urea in ureotelic species. Mutations in ASL, resulting low activity of the enzyme, increase levels of urea in the body and result in various side effects.

The ASL gene is located on chromosome 7 between the centromere (junction of the long and short arm) and the long (q) arm at position 11.2, from base pair 64,984,963 to base pair 65,002,090.

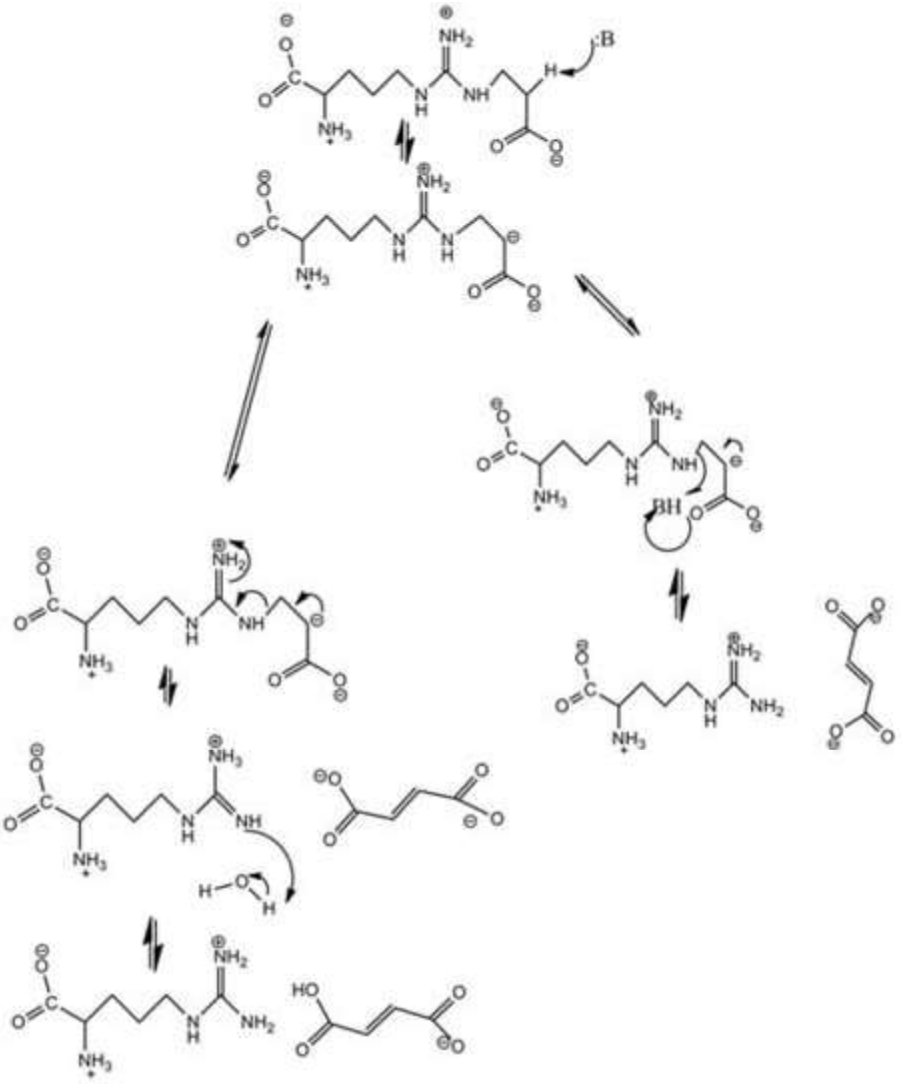
ASL is related to intragenic complementation.

Structure

ASL is composed of four identical monomers; each monomer consisting of a single polypeptide chain between 49-52 kDa, 196-208 kDa for the entire tetrameric enzyme. Each monomer has three highly conserved regions remote from one another, but these regions cluster together in the tetramer to form four active sites. Therefore each ASL homotetramer has four active sites to catalyze the breakdown of argininosuccinate.

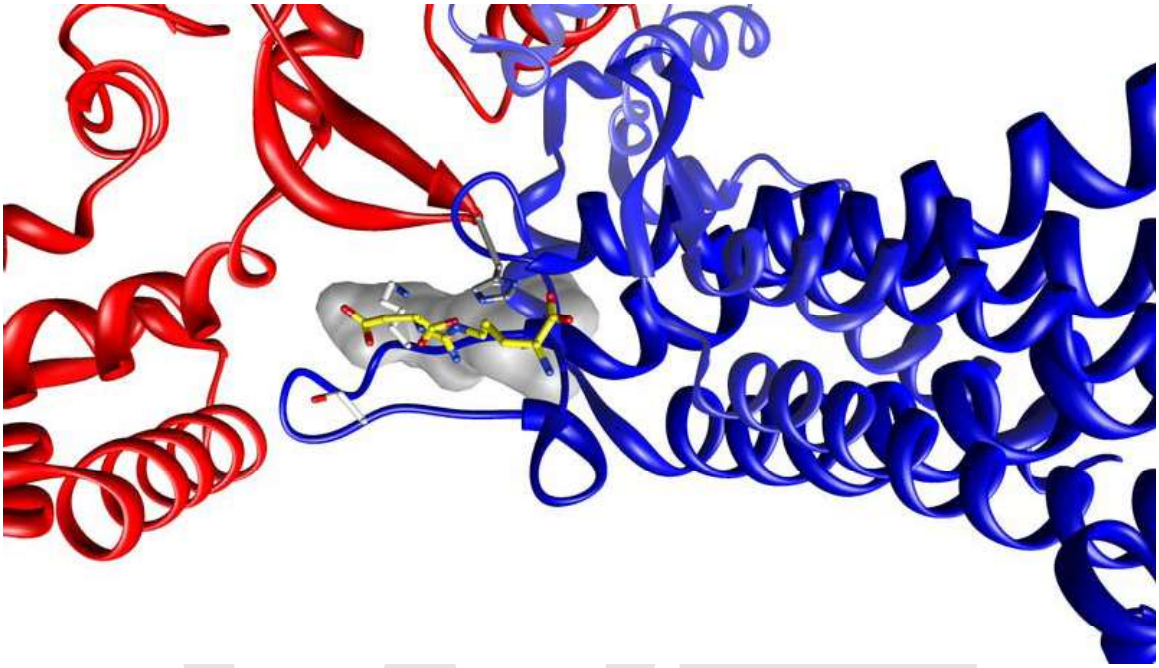
Each monomer in the ASL homotetramer is composed of three structural domains; all three are primarily alpha helical. Domains 1 and 3 are similar in structure as they both consist of helix-turn-helix motifs. Domain 1 of the monomer contains the carboxyl terminus. Domain 2 contains one small beta sheet, nine alpha helices, and the amino terminus. Three of the nine alpha helices on one monomer are engaged mainly in hydrophobic interactions with another monomer to form a dimer. Two dimers then associate by way of alpha helix, one from each monomer, to form a central 20-helix core. The association of all four monomers allows for the catalytic activity at each possible active site.

Mechanism



Sampaleanu, L. M. J. Biol. Chem., 2002, 277, 6, 4166-4175. *(Figure 6)*

Proposed ASL mechanism



Argininosuccinate (in yellow) in the Active Site of ASL

The enzyme's cleavage of the argininosuccinate, to form fumarate and arginine, occurs through an E1cb elimination reaction. The base initiates the reaction by deprotonating the carbon adjacent to the arginine, or leaving group. Recent mutagenic studies of ASL homologues have shown that Histidine 162 or Threonine 161 of ASL is responsible for the proton abstraction of the Cβ, either directly or indirectly through a water molecule. Lysine 289 is thought to stabilize the negatively charged carbanion intermediate. Although there is no consensus of the catalytic acid that donates the proton to the imine functional group of the arginine product, some mutagenesis studies show serine 283 may be involved.

Role in the urea cycle

Ammonia (NH_3) is a toxic substance for many aerobic organisms and must be excreted. Some aquatic organisms release the toxin right directly into their environment, while other ureotelic species must convert their toxic nitrogen waste into non-toxic components, like uric acid or urea, through a series of catalyzed steps better known as the urea cycle. ASL catalyzes the fourth step in the cycle, following the action of argininosuccinate synthetase (ASS) in the liver cytosol. While ASS catalyzes the formation of argininosuccinate from citrulline and aspartate, ASL breaks the newly formed argininosuccinate into L-arginine and fumarate. L-arginine continues through the urea cycle to form urea and orinthine, while fumarate can enter the citric acid cycle.

δ-Crystallin

ASL, δ-crystallin, class II fumarase, aspartase, adenylosuccinase lyase, and 3-carboxy-cis,and cis-muconate lactonizing enzyme are all members of the same homotetrameric superfamily of enzymes, in which most catalyze the same type of elimination reactions where a C-O or C-N bond is broken and fumarate is released as a product. δ-crysatllins are the major structural eye lens water soluble proteins of most birds, reptiles, and some other vertebrates. Within the superfamily, ASL is most closely related to δ-Crystallin in amino acid sequence and in protein fold structure. There are two isoforms of the crystalline, δI and δII. These two isoforms conserve 69% and 71% of the ASL amino acid sequence, respectively, but only the δII isoform retains the same enzymatic activity as ASL. The similarities have led researches to believe that these crystallins have evolved from the recruitment to the lens of preexisting metabolic enzymes, like ASL, by a process called 'gene sharing'. The same gene product functions as a both a lens crystallin and an enzyme in other non-ocular tissues. Comparative studies of the δ-crystallins have been most beneficial for understanding the enzymatic mechanism of the ASL reaction.

Mutations and ASL deficiencies: argininosuccinic aciduria

Mutations in the human ASL gene causes argininosuccinic aciduria, a rare autosomal recessive disorder, and results in deficiencies of the urea cycle. Argininosuccinate lyase is an intermediate enzyme in the urea synthesis pathway and its function is imperative to the continuation of the cycle. A non-functioning enzyme results in patients' accumulation of ammonia, argininosuccinate, and citrulline in the blood, and argininosuccinate is excreted in the urine. Other resulting symptoms include lethargy, vomiting, hypothermia, hyperventilation, hepatomegaly and progressive encephalopathy in infant patients, and abnormal hair growth, hepatic fibrosis, episodic vomiting, growth and developmental delay, in patients experiencing the disorder later in childhood.

ASL is a key enzyme in the conversion of ammonia to urea through the urea cycle. Ammonia builds to toxic levels, resulting in hyperammonemia. Ammonia is toxic in part because it affects the nervous system. There is biochemical evidence that shows rises in ammonia can inhibit glutaminase and therefore limit the rate of synthesis of neurotransmitters such as glutamate, which can explain the developmental delay in argininosuccinic aciduria patients.

One mutation in patients with argininosuccinic aciduria occurs when glutamine 286 is mutated to arginine. The enzyme now has a positively charged arginine in place of a neutrally charged glutamine and studies suggest this change may sterically and/or electrostatically hinder a conformational change necessary for catalysis.

Chapter- 4

ATG8 and Bcl-2

ATG8

autophagy related protein 8

Crystal structure of Atg8

Identifiers	
Symbol	Atg8
Alt. symbols	Apg8, Aut7, Cvt5,
Entrez	852200
PDB	1ugm
UniProt	P38182

Autophagy-related protein 8 (**Atg8**) is a ubiquitin-like protein required for the formation of autophagosomal membranes. The transient conjugation of Atg8 to the autophagosomal membrane through a ubiquitin-like conjugation system is essential for autophagy in

eukaryotes. Even though there are homologues in animals, here we mainly focuses on its role in lower eukaryotes such as Saccharomyces cerevisiae.

Structure

Atg8 is a monomer of 117 aminoacids and a molecular weight of 13,6kDa. It consists of a 5-stranded β-sheet, which is enclosed by two α-helices at one side and one α-helix at the other side and exhibits a conserved GABARAP domain. Even though Atg8 does not show a clear sequence homology to ubiquitin, its crystal structure reveals a conserved ubiquitin-like fold.

Function

In Autophagy

Atg8 is one of the key molecular components involved in autophagy, the cellular process mediating the lysosome/vacuole-dependent turnover of macromolecules and organelles. Autophagy is induced upon nutrient depletion or rapamycin treatment and leads to the response of more than 30 autophagy-related (ATG) genes known so far, including ATG8. How exactly ATG proteins are regulated is still under investigation, but it is clear that all signals reporting on the availability of carbon and nitrogen sources converge on the TOR signalling pathway and that ATG proteins are downstream effectors of this pathway. In case nutrient supplies are sufficient, the TOR signaling pathway hyperphosphorylates certain Atg proteins, thereby inhibiting autophagosome formation. After starvation autophagy is induced through the activation of Atg proteins both on the protein modification and the transciptional level.

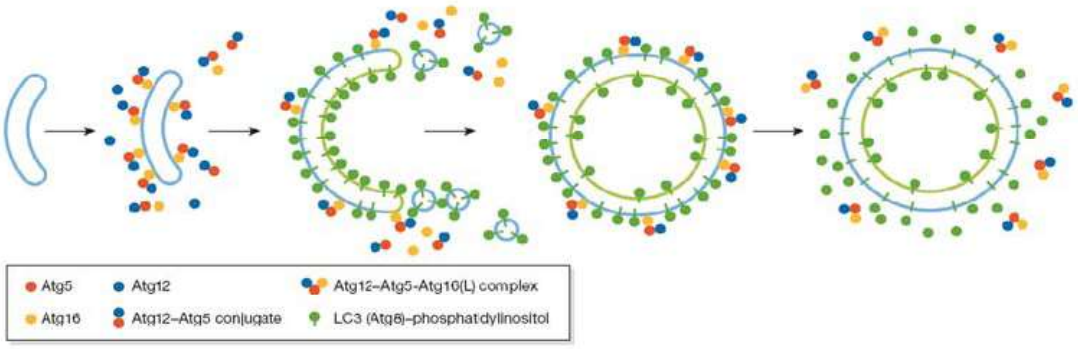


Fig. 1 Atg8 and Atg12 conjugation systems in autophagosome formation

Atg8 is especially important in macroautophagy which is one of three distinct types of autophagy characterized by the formation of double-membrane enclosed vesicles that sequester portions of the cytosol, the so called autophagosomes. The outer membrane of these autophagosomes subsequently fuses with the lysosome/vacuole to release an iner single membrane (autophagic body) destined for degradation. During this process, Atg8 is particularly crucial for autophagosome maturation (lipidation).

Like most Atg proteins, Atg8 is localized in the cytoplasm and at the PAS under nutrient-rich conditions, but becomes membrane-associated in case of autophagy induction. It then localizes to the site of autophagosome nucleation, the phagophore-assembly site (PAS). Nucleation of the phagophore requires the accumulation of a set of Atg proteins and of class III phosphoinositide 3-kinase complexes on the PAS. The subsequent recruitment of Atg8 and other autophagy-related proteins is believed to trigger vesicle expansion in a concerted manner, presumably by providing the driving force for membrane curvature. The transient conjugation of Atg8 to the membrane lipid phosphatidylethanolamine is essential for phagophore expansion as its mutation leads to defects in autophagosome formation. It is distributed symmetrically on both sides of the autophagosome and it is assumed that there is a quantitative correlation between the amount of Atg8 and the vesicle size.

After finishing vesicle expansion, the autophagosome is ready for fusion with the lysosome and Atg8 can either be released from the membrane for recycling (see below) or gets degraded in the autolysosome if left uncleaved.

ATG8 is also required for a different autophagy-related process called the Cytoplasm-to-vacuole-targeting (Cvt) pathway. This yeast-specific process acts constitutively under nutrient-rich conditions and selectively transports hydrolases such as aminopeptidase I to the yeast vacuole. The Cvt pathway also requires Atg8 localised to the PAS for the formation of Cvt vesicles which then fuse with the vacuole to deliver hydrolases necessary for degradation.

Post-translational Modification and Regulatory Cycle

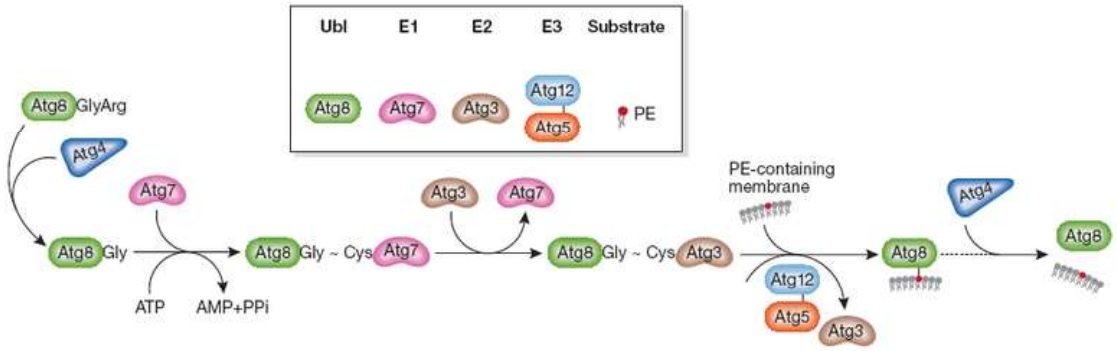


Fig. 2 Post-translational modification and regulatory cycle in yeast

Atg8 exists in a cytoplasmic and in a membrane-associated form. Membrane association is achieved by coupling Atg8 to phosphatidylethanolamine (PE) which is a lipid constituent of plasma membranes. This post-translational modification process, called lipidation, is performed by the Atg8 conjugation system comprising the cysteine protease ATG4 (belonging to the caspase family), as well as the proteins ATG7, ATG3 and the ATG5-ATG12 complex.

The Atg8 conjugation system (Fig.1) works in analogy to the ubiquitination system. However, it is Atg8 itself that represents the ubiquitin-like protein (Ubl) being transferred to PE, while ATG7 functions like an E1 enyzme, ATG3 like an E2 enyzme and the ATG12-ATG5 complex like an E3 ligase

The lipidation process is initiated by an ATG4 dependent post-translational cleavage of the last C-terminal amino acid residue of Atg8. After the cleavage, Atg8 exposes a C-terminal glycine residue (Gly 116) to which PE can then be coupled during the following steps. In the first step, the Gly116 residue of Atg8 binds to a cysteine residue of ATG7 via a thioester bond in an ATP-dependent manner. During the second step, Atg8 is transferred to Atg3 assuming the same type of thioester bondage. Finally, Atg8 is detached from Atg3 and coupled to the hydroxyl head group of PE via an amide bond. This final step was found to be facilitated and stimulated by the ATG5-ATG12 complex.

Both proteins, Atg5 and Atg12 were originally identified as part of another Ubl conjugating system that promotes conjugation of ATG12 to ATG5 via ATG7 and Atg10. This implies, that the ATG12 and the Atg8 conjugation system are actually interdependent.

Mammalian Homologues

In higher eukaryotes Atg8 is not encoded by a single gene as in yeast, but derived from a multigene family. Four of its homologues have already been identified in mammalian cells.

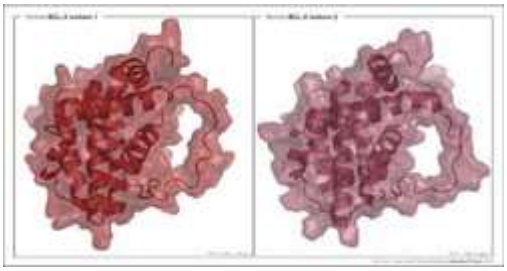
One of them is LC3 (MAP1LC3A), a light chain of the microtubule-associated protein 1 Like Atg8, LC3 needs to be proteolytically cleaved and lipidated to be turned into its active form which can localize to the autophagosomal membrane. Similar to the situation in yeast, the activation process of LC3 is triggered by nutrient depletion, but interestingly also in response to hormones.

Other homologues are the transport factor GATE-16 (Golgi-associated ATPase enhancer of 16 kDa) which plays an important role in intra-golgi vesicular transport by stimulating NSF (N-ethylmaleimide-sensitive factor) ATPase activity and interacting with the Golgi v-SNARE GOS-28, and GABARAP (γ-aminobutyric acid type A receptor associated protein) which facilitates clustering of GABAA receptors in combination with microtubules.

All three proteins are characterized by proteolytic activation processes upon which they get lipidated and localized to the plasma membrane. However, for GATE-16 and GABARAP membrane association seems to be possible even for the non-lipidated forms. Apart from LC3, GABARAP and GATE-16 the most recently but less well characterized mammalian homologue is ATGL8. Little is known about its actual activation process except for its interaction with one of the mammalian ATG4 homologues, hATG4A.

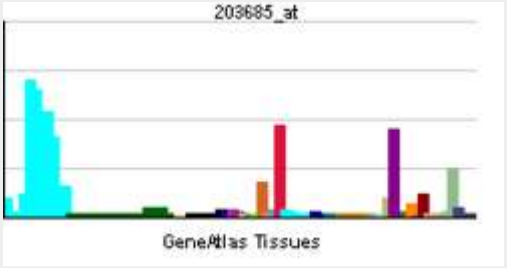
Bcl-2

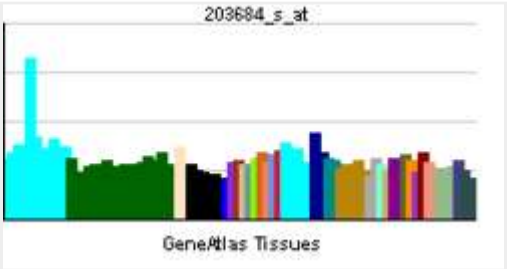
B-cell CLL/lymphoma 2

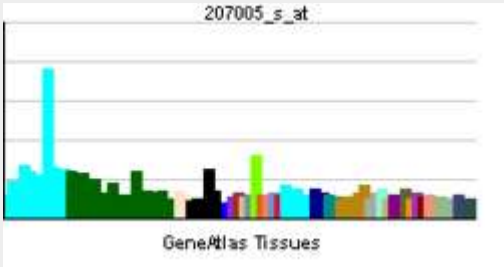


PDB rendering based on 1GJH,1G5M.

Available structures	
Identifiers	
Symbols	BCL2; Bcl-2
External IDs	OMIM: 151430 MGI: 88138 HomoloGene: 527 GeneCards: BCL2 Gene
Gene Ontology	
RNA expression pattern	







Orthologs		
Species	**Human**	**Mouse**
Entrez	596	12043
Ensembl	ENSG00000171791	ENSMUSG00000057329
UniProt	P10415	Q4VBF6
RefSeq (mRNA)	NM_000633	NM_009741
RefSeq (protein)	NP_000624	NP_033871
Location (UCSC)	Chr 18: 58.94 - 59.14 Mb	Chr 1: 108.37 - 108.54 Mb

Bcl-2 (**B-cell lymphoma 2**) is the founding member of the Bcl-2 family of apoptosis regulator proteins encoded by the **BCL2** gene. Bcl-2 derives its name from *B-cell lymphoma 2*, as it is the second member of a range of proteins initially described in chromosomal translocations involving chromosomes 14 and 18 in follicular lymphomas. *Bcl-2* orthologs have been identified in numerous mammals for which complete genome data are available.

Role in disease

The Bcl-2 gene has been implicated in a number of cancers, including melanoma, breast, prostate, and lung carcinomas, as well as schizophrenia and autoimmunity. It is also thought to be involved in resistance to conventional cancer treatment. This supports a role for decreased apoptosis in the pathogenesis of cancer.

Cancer is one of the world's leading causes of death and occurs when the homeostatic balance between cell growth and death is disturbed. Research in cancer biology has discovered that a variety of aberrations in gene expression of anti-apoptotic, pro-apoptotic and BH3-only proteins can contribute to the many forms of the disease. An interesting example can be seen in lymphomas. The over-expression of the anti-apoptotic Bcl-2 protein in lymphocytes alone did not act in an oncogenic manner. But simultaneous

over-expression of Bcl-2 and the proto-oncogene myc may produce aggressive B-cell malignancies including lymphoma. In follicular lymphoma, a chromosomal translocation commonly occurs between the fourteenth and the eighteenth chromosomes—t(14;18)—which places the Bcl-2 gene next to the immunoglobulin heavy chain locus. This fusion gene is deregulated, leading to the transcription of excessively high levels of bcl-2. This decreases the propensity of these cells for undergoing apoptosis.

Apoptosis also plays a very active role in regulating the immune system. When it is functional, it can cause immune unresponsiveness to self-antigens via both central and peripheral tolerance. In the case of defective apoptosis, it may contribute to etiological aspects of autoimmune diseases. The autoimmune disease, type 1 diabetes can be caused by defective apoptosis, which leads to aberrant T cell AICD and defective peripheral tolerance. Due to the fact that dendritic cells (DCs) are the most important antigen presenting cells of the immune system, their activity must be tightly regulated by such mechanisms as apoptosis. Researchers have found that mice containing DCs that are Bim -/-, thus unable to induce effective apoptosis, obtain autoimmune diseases more so than those that have normal DCs. Other studies have shown that the lifespan of DCs may be controlled by factors such as a timer dependent on anti-apoptotic Bcl-2. These investigations illuminate the importance of regulating antigen presentation as dis-regulation can lead to autoimmunity.

Apoptosis plays a very important role in regulating a variety of diseases that have enormous social impacts. For example, schizophrenia is a neurodegenerative disease that may result from an abnormal ratio of pro- and anti-apoptotic factors. There is some evidence that this defective apoptosis may result from abnormal expression of Bcl-2 and increased expression of caspase-3.

Further research into the family of Bcl-2 proteins will provide a more complete picture on how these proteins interact with each other to promote and inhibit apoptosis. An understanding of the mechanisms involved will help discover potential treatments such as inhibitors to target over-expressed proteins that may lead to new therapies in cancer, autoimmune conditions, and neurological diseases.

Targeted therapies

Bcl-2 inhibitors include :

Genasense

An antisense oligonucleotide drug Genasense (G3139) has been developed by Genta Incorporated to target Bcl-2. An antisense DNA or RNA strand is non-coding and complementary to the coding strand (which is the template for producing respectively RNA or protein). An antisense drug is a short sequence of RNA which hybridises with and inactivates mRNA, preventing the protein from being formed.

It was shown that the proliferation of human lymphoma cells (with t(14;18) translocation) could be inhibited by antisense RNA targeted at the start codon region of Bcl-2 mRNA. In vitro studies led to the identification of Genasense, which is complementary to the first 6 codons of Bcl-2 mRNA.

These have shown successful results in Phase I/II trials for lymphoma, and a large Phase III trial was launched in 2004

By the first quarter 2010, Genasense had not received FDA approval due to disappointing results in a melanoma trial. Although safety and efficacy of Genasense have not been established for any use, Genta Incorporated still claims on its website that studies are currently underway to examine the potential role of Genasense in a variety of clinical indications.

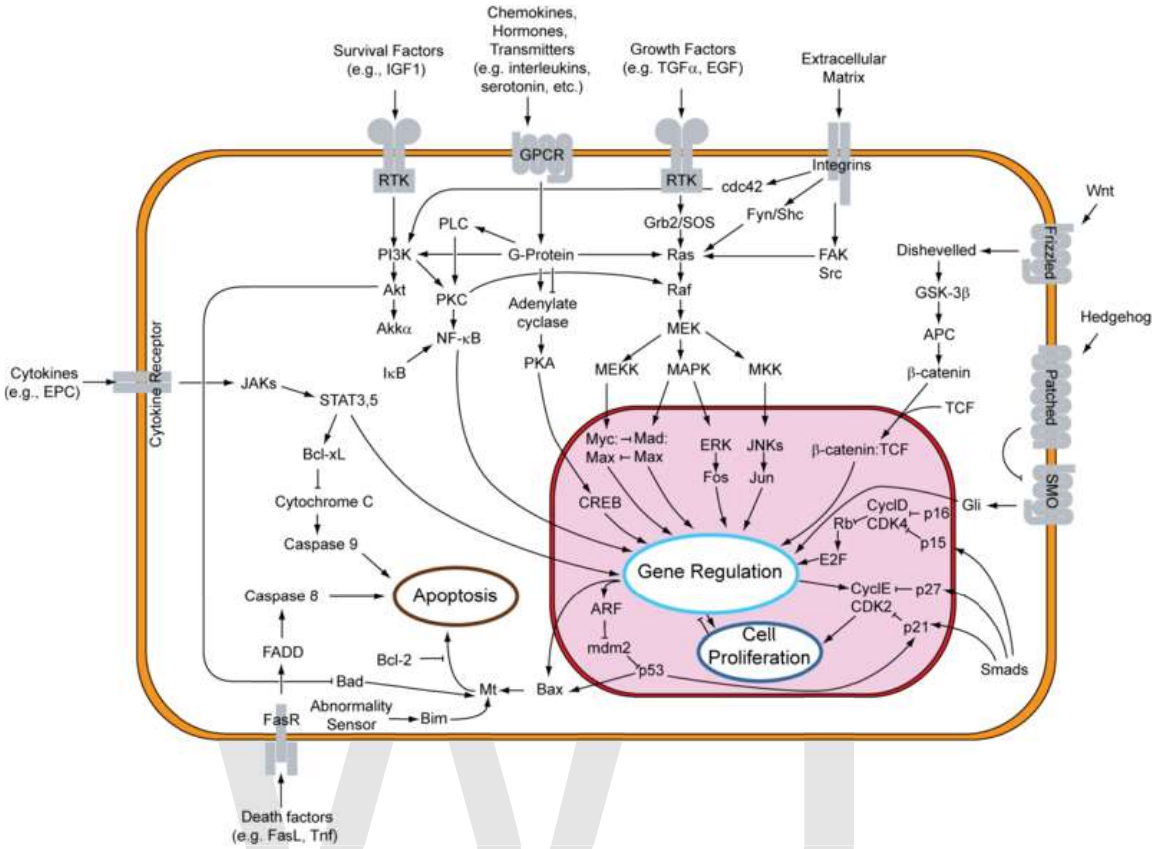
ABT-737

Abbott Laboratories described in the mid-2000s a novel inhibitor of Bcl-2, Bcl-xL and Bcl-w, known as ABT-737. ABT-737 is one among many so-called BH3 mimetic small molecule inhibitors (SMI) targeting Bcl-2 and Bcl-2-related proteins such as Bcl-xL and Bcl-w but not A1 and Mcl-1, which may prove valuable in the therapy of lymphoma and other blood cancers.

Others

- obatoclax (GX15-070) has phase II results for small-cell lung cancer.

Interactions



Overview of signal transduction pathways involved in apoptosis.

Bcl-2 has been shown to interact with RAD9A, BAK1, Reticulon 4, Bcl-2-associated X protein, Caspase 8, BECN1, SOD1, Bcl-2-interacting killer, BH3 interacting domain death agonist, RRAS, C-Raf, BCL2L11, BNIPL, HRK, PSEN1, BMF, BNIP2, BNIP3, Nerve Growth factor IB, BCL2-like 1, Myc, BCAP31, SMN1, CAPN2, PPP2CA, Noxa, Cdk1, TP53BP2, Bcl-2-associated death promoter and IRS1.

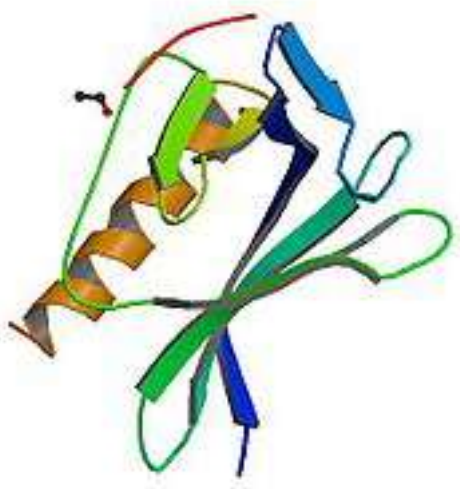
Human BCL-2 genes

BAK; BAK1; BAX; BCL2; BCL2A1; BCL2L1; BCL2L10; BCL2L13; BCL2L14; BCL2L2; BCL2L7P1; BOK; MCL1;

Chapter- 5

Actin Assembly-Inducing Protein

Actin assembly-inducing protein



EVH1 domain-ActA peptide complex

Identifiers	
Symbol	ActA
Entrez	2798121
UniProt	P33379

The **Actin assembly-inducing protein** (**ActA**) is a protein encoded and used by *Listeria monocytogenes* to propel itself through a mammalian host cell. ActA is a bacterial surface protein comprising a membrane-spanning region. In a mammalian cell the bacterial ActA interacts with the Arp2/3 complex and actin monomers to induce actin polymerization on the bacterial surface generating an actin comet tail. The gene encoding ActA is named *actA* or prtB.

Introduction

As soon as *L. monocytogenes* bacteria are ingested by humans, they get internalized into intestinal epithelium cells and rapidly try to escape their internalization vacuole. In the cytosol they start to polymerize actin on their surface by the help of the ActA protein. It has been shown that ActA is not only necessary but also sufficient to induce motility of bacteria in the absence of other bacterial factors.

Discovery of ActA

ActA was discovered by analysing lecithinase-negative Tn*917-lac* Listeria mutants because of the phenotype that they were unable to spread from cell to cell. These mutant bacteria still escaped from the phagosomes as efficiently as wild-type bacteria and multiplied within the infected cells but they were not surrounded by actin like wild-type bacteria. Further analysis showed, that Tn*917-lac* had inserted into *actA*, the second gene of an operon. The third gene of this operon, *plcB*, encodes the *L. monocytogenes* lecithinase. To determine whether *actA* itself, *plcB* or other co-transcribed downstream regions are involved in actin assembly, mutations in the appropriate genes were generated. All mutants except the *actA* mutants were similar to wild-type concerning association with F-actin and cell-cell spreading. Complementation with *actA* restored wild-type phenotype in the *actA* mutants.

Function

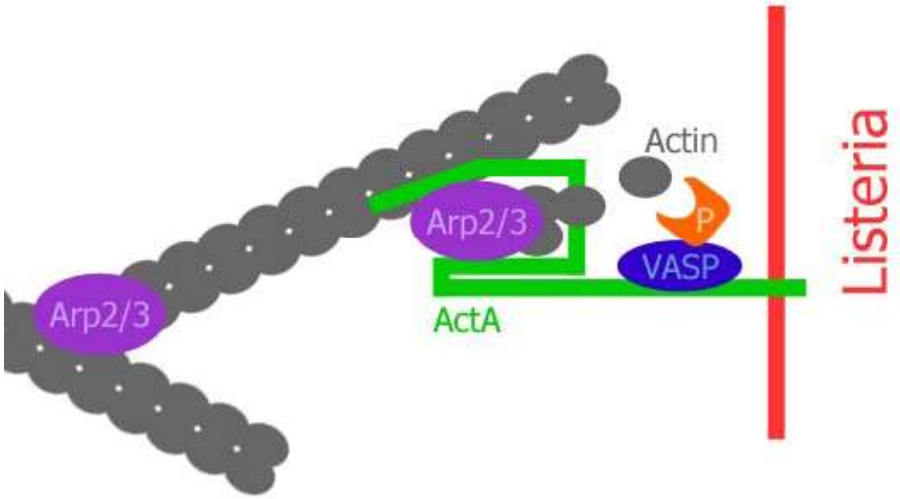


Fig. 1 Actin assembly induced by bacterial protein ActA (shown in green). Mammalian proteins involved in this process are: Profilin (P), Vasodilator-stimulated phosphoprotein (VASP) and actin-related-protein 2 and 3 complex (Arp2/3 complex) as well as actin.

ActA is a protein which acts as a mimic of Wiskott-Aldrich syndrome protein (WASP), a nucleation promoting factor (NPF) present in host cells. NPFs in general recruit and bind to the already in the mammalian cell existing actin-related-protein 2 and 3 complex

(Arp2/3 complex) and induce an activating conformational change of the Arp2/3 complex. Due to this conformational change, NPFs initiate polymerization of a new actin filament at a 70° angle, which leads to the characteristic Y-branched actin structures in the leading edge of motile cells. ActA localizes to the old pole of the bacterium and spans both the bacterial cell membrane and the cell wall, lateral diffusion is inhibited; thus ActA localizes in a polarized and anchored manner on the bacterial surface. Consequently actin polymerization only starts in this region on the surface of the bacterium. Expression of ActA is induced only after entering a mammalian host cell.

Actin filament assembly generates the force that pushes the bacterium in the mammalian host cytoplasm forward. Continuous actin polymerization is sufficient for motility in the cytoplasm and even for infection of adjacent cells.

Research

New data indicates that ActA plays a role also in vacuolar disruption. A deletion mutant of ActA was defective in permeabilizing the vacuole. An 11 amino acid stretch of the N-terminus of the acidic region (32-42) was shown to be important for disruption of the phagosome.

Structure

The primary proteinous product of the *actA* gene consists of 639 amino acids and includes the signal peptide (first N-terminal 29 amino acids) and the ActA chain (C-terminal 610 amino acids). Therefore the sequence of the mature ActA protein consist of 610 amino acids. ActA has a molecular weight of 70,349 Da and is a surface protein.

The ActA chain can be divided into three functional domains (Fig. 2):

- N-terminal domain that is highly charged: amino acid residues 1-234
- central domain with proline-rich repeats: amino acid residues 235-394
- C-terminal domain with a transmembrane domain: amino acid residues 395-610

N-terminal Domain

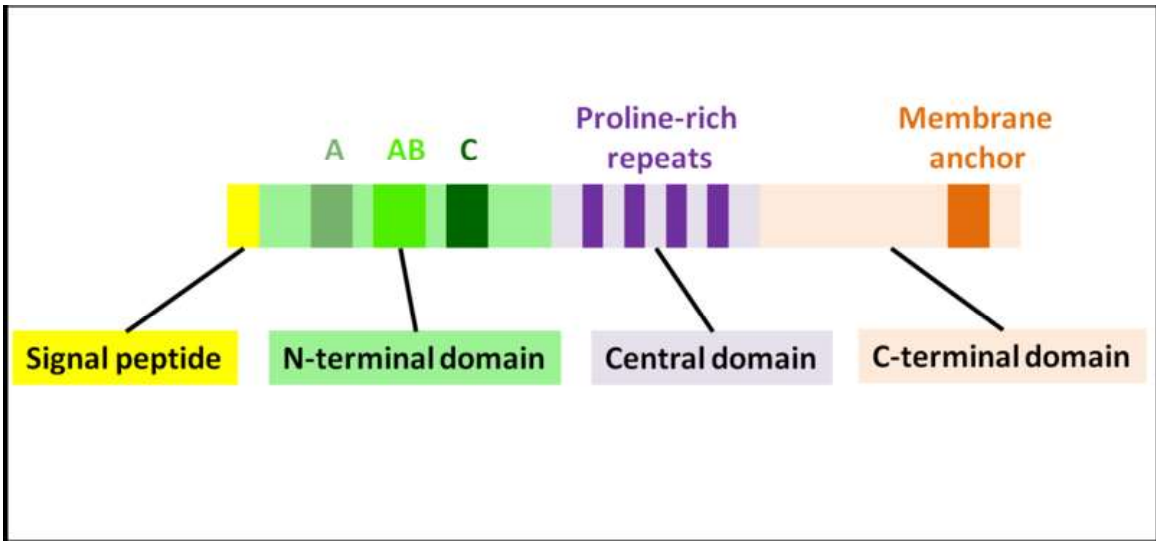


Fig. 2 The ActA protein and its functional domains

The first 156 amino acids of the N-terminal domain consist of three regions (Fig. 2):

- A-region with a stretch of acidic residues: 32-45
- AB-region, an actin monomer-binding region: 59-102
- C-region, a cofilin homology sequence: 145-156

The N-terminal portion of ActA plays an important role in actin polymerization. The domain displays consensus elements present in eukaryotic WASP family NPFs which include an actin monomer-binding region as well as an Arp2/3 binding C (central or cofilin homology) and A (acidic) region. The actin monomer-binding region of ActA has functional properties like the WASP-Homology-2 (WH2) or V domain, but differs in the sequence. Thus in WASP-family NPFs the order of the domains is WH2 followed by C,and then by A, which is not the case in ActA.

Central Domain

The central proline-rich region of ActA is crucial for ensuring efficient bacterial motility. There are four proline-rich repeats containing either FPPPP or FPPIP motifs. These regions mimic those of the host cell cytoskeletal protein zyxin, vinculin and palladin, known to associate with focal adhesions or stress fibers. The vasodilator-stimulated phosphoprotein (VASP) can bind through its Ena/VASP homology 1 domain (EVH1 domain) to the central proline-rich region and recruits profilin, an actin monomer binding protein, which itself promotes polymerization at barbed ends of actin filaments. Furthermore, VASP seems to interact with F-actin through its carboxy-terminal EVH2 domain, which provides a linkage of the bacterium to the tail. This statement is supported by the fact that ActA can bind multiple Ena/VASP proteins simultaneously and has a high affinity between ActA and Ena/VASP. VASP has been shown to reduce the

frequency actin-Y-branches in vitro and thus increases the proportion of filaments which are organized in a parallel alignment in comet tails.

C-terminal Domain

The C-terminal domain of ActA has a hydrophobic region which anchors the protein in the bacterial membrane.

In summary, besides

- the absence of sequence homology in the actin-binding-region and
- an alteration in the sequence of ARP2/3 activating domains typical for WASP-family NPFs (V(WH2)-C-A),
- a major difference between ActA and host NPFs is that ActA does not have elements that bind to regulatory proteins such as Rho family GTPases. This structural difference between ActA and host NPFs can be advantageous for *L. monocytogenes* and its pathogenesis because the actin nucleation activity of *L. monocytogenes* is independent of host regulation.

Analogues

WASP/N-WASP, which is functionally mimicked by ActA, is highly conserved in eukaryotes. It is an important actin-cytoskeleton organizer and is critical for processes such as endocytosis and cell motility. Activated by Cdc42, a Rho-family small GTPase, WASP/N-WASP activates the Arp2/3 complex, which leads to rapid actin polymerization.

Actin-based Motility of other Pathogens

In *Shigella* the protein IcsA activates N-WASP, which in non-infected mammalian cells is activated by the GTPase Cdc42. Active N-WASP/WASP leads to actin polymerization by activating the Arp2/3 complex. In contrast, the *Listeria* ActA protein interacts with and activates directly the Arp2/3 complex.

The *Rickettsia* RickA protein is also able to activate the Arp2/3 complex in a WASP-like manner. In contrast to *Listeria*, the actin filaments are organized in long, unbranched parallel bundles. The Arp2/3 complex is only localized near the bacterial surface and thus it is assumed that a more frequent Arp2/3 complex-independent elongation occurs.

In *Burkholderia pseudomallei* BimA initiates actin polymerization in vitro. It is assumed that intracellular migration of this bacterium functions independently of the Arp2/3 complex.

Chapter- 6

Dun Gene

A bay dun, also called a "classic" or "zebra" dun

A Blue dun, or Grullo

Przewalski's horses. The animal on the left shows the dorsal stripe along its spine, the one on the right shows faint horizontal "zebra" striping on the back of its legs by the knee, both classic examples of "primitive" dun markings.

The **dun gene** is a dilution gene that affects both red and black pigments in the coat color of a horse. The dun gene has the ability to affect the appearance of all black, bay, or chestnut ("red")-based horses to some degree by lightening the base body coat and suppressing the underlying base color to the mane, tail, legs and "primitive markings."

The classic **Dun** is a gray-gold or tan, characterized by a body color ranging from sandy yellow to reddish-brown. Dun horses always have a dark stripe down the middle of their back, a tail and mane darker than the body coat, and usually darker faces and legs. Other duns may appear a light yellowish shade, or a steel gray, depending on the underlying coat color genetics. Manes, tails, primitive markings and other dark areas are usually the shade of the non-diluted base coat color.

The dun allele is a simple dominant, so that the phenotype of a horse with either one copy or two copies of the gene is dun. It has a stronger effect than other dilution genes, such as the silver dapple gene, which acts only on black-based coats, or the cream gene, an

incomplete dominant which must be homozygous to be fully expressed, and when heterozygous is only visible on bay and chestnut coats, and then to a lesser degree.

The dun gene also is characterized by primitive markings which are darker than the body color. Primitive markings include:

- Dorsal stripe (stripe down the center of the back, along the spine), seen almost universally on all duns
- Horizontal striping on the back of forelegs, common on most duns, though at times rather faint
- Shoulder blade stripe, the least commonly-seen of the primitive markings.

Dorsal striping does not guarantee that the horse carries the dun gene. A countershading gene can also produce faint dorsal striping, even in breeds such as the Arabian horse or the Thoroughbred, where the dun gene is not known to be carried in the gene pool. A primary characteristic of the dun gene is the dorsal stripe, and most duns also have visual leg striping. The shoulder stripes are less common and often fainter, but usually visible on horses with a short summer coat.

Taxonomic distribution

Cave painting at Lascaux. Dun is thought to be a primitive trait.

The dun coat color is thought to be a primitive trait in the horse. This is because equines appearing in prehistoric cave paintings are dun and because several closely related species in the genus *Equus* are known to have been dun. These species include both subspecies of *Equus ferus* (the extinct tarpan and the extant but endangered Przewalski's horse), the extinct *Equus lambei*, and the extant onager and kiang.

Shades of dun

Red dun

Blue dun, or Grullo.

The dun gene has a stronger dilution effect on the body than the mane, tail, legs and primitive markings, and so lightens the body coat more. This explains why points on a dun are a shade darker than the coat, or in the case of a "classic" dun, the mane, tail, and legs are often black or only slightly diluted.

- **Dun**, also called **Bay dun** or "zebra" dun. The most common type of dun, has a tan or gold body with black mane, tail and primitive markings. Genetically, the horse has an underlying bay coat color, acted upon by the dun gene.
- **Red dun** horses do not have black points, as there is no black on the horse to be affected. Instead, the points and primitive markings are a darker shade of red than the coat. Genetically, the horse has an underlying chestnut coat color, acted upon by the dun gene. In some places, this is also called a "fox dun."
- **Grullo** or *Grulla*, also called **blue dun** or "mouse" dun, is a smoky, bluish to mouse-brown color and can vary from light to dark. They consistently have black points and they often have a dark or black head, which is an identifying characteristic of this color. The primitive markings are usually all black. Genetically, the horse has an underlying black coat color, acted upon by the dun

gene. Unlike a blue roan, there are no intermingled black and white hairs, and unlike a true gray, which also intermingles light and dark hairs, the color does not change to a lighter shade as the horse ages. With a dun, the hair color itself is one solid shade.

Dun mimics

A countershading stripe, here on a bay horse, is not produced by the dun gene

Since the dun gene, when on a "bay dun" horse, can closely resemble buckskin, in that both colors feature a light-colored coat with a dark mane and tail, classic duns are frequently confused with buckskins. The difference between these two colors is that dun is a tan color, somewhat duller than the more cream or gold buckskin, and duns also possess primitive markings. Some buckskins do show countershading, but it is not related to the primitive markings of dun factor horses.

Genetically, a bay dun is a bay horse with the dun gene that causes the lighter coat color and the primitive markings. A buckskin is bay horse with the addition of the cream gene causing the coat color to be diluted from red to gold, often without primitive markings.

A red dun may also be confused with a perlino, which is genetically a bay horse with two copies of the cream gene, which creates a horse with a cream-colored body but a reddish mane and tail. However, perlinos usually are significantly lighter than a red dun and generally have blue eyes.

To further confuse matters, it is possible for a horse to carry both dun and cream dilution genes; such horses with golden buckskin coloring and a complete set of primitive markings are referred to as a "buckskin dun" or a "dunskin." In the Fjord horse, duns that also carry the creme dilution are called Uls dun or White dun (*ulsblakk*) and Yellow dun

(*gulblakk*) by their respective coat color. On such horses, the light-shaded primitive markings are most noticeable during the summer months when the winter hair sheds.

Countershading is a usually a darker shade of the body color rather than the near-black of primitive markings on bay duns, but it may be harder to differentiate between countershading and a dorsal stripe on light-colored horses such as red duns. In such cases, pedigree analysis, DNA testing, studying possible offspring, and the presence of other primitive markings are used to determine if a horse is a dun.

Breeding and the dun gene

Dorsal stripe and light guard hairs on a dun horse

The three primary dun varieties usually occur in proportion to the occurrence of the corresponding base colors in each particular breed. They are created by the following combinations of the dun gene acting upon an underlying base coat color.

- Red (Chestnut) base + Dun gene= Red Dun.
- Black base + Dun gene= Blue dun, mouse dun or Grullo/Grulla.
- Bay (black base + Agouti gene) + Dun gene= Classic dun, sometimes called "Bay dun" or "Zebra dun".

Other variations result from the interplay of additional genes. For example:

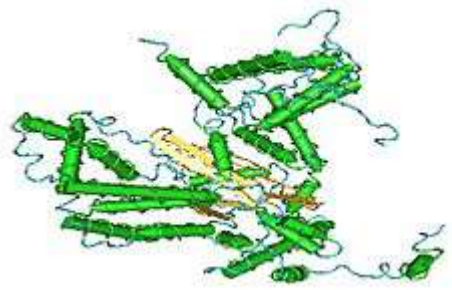
- Chestnut + Dun + cream gene (single copy) = "dunalino" or "palomino dun"
- Bay + Dun + cream gene (single copy) = "dunskin" or "buckskin dun"

A single copy of the cream gene on a black base coat does not lighten black hair, and thus a single copy has no visible effect on a grullo, either. Double copies of the cream gene create very light-colored horses (cremello, perlino and smoky cream). Thus, if a horse with two cream dilution alleles also carries the dun gene, primitive markings are not usually visible to any significant degree.

Chapter- 7

FAM200A

chromosome 7 open reading frame 38



CBLAST Structurally Related Protein. Hermes DNA Transposase. EValue: 1E-6

Identifiers		
Symbols	C7orf38; FLJ36794; DKFZp727G131	
External IDs	HomoloGene: 89159 GeneCards: C7orf38 Gene	
Orthologs		
Species	**Human**	**Mouse**
Entrez	221786	n/a
Ensembl	ENSG00000221909	n/a
UniProt	Q8TCP9	n/a
RefSeq (mRNA)	NM_145111	n/a
RefSeq (protein)	NP_659802	n/a
Location (UCSC)	Chr 7: 98.98 - 98.98 Mb	n/a

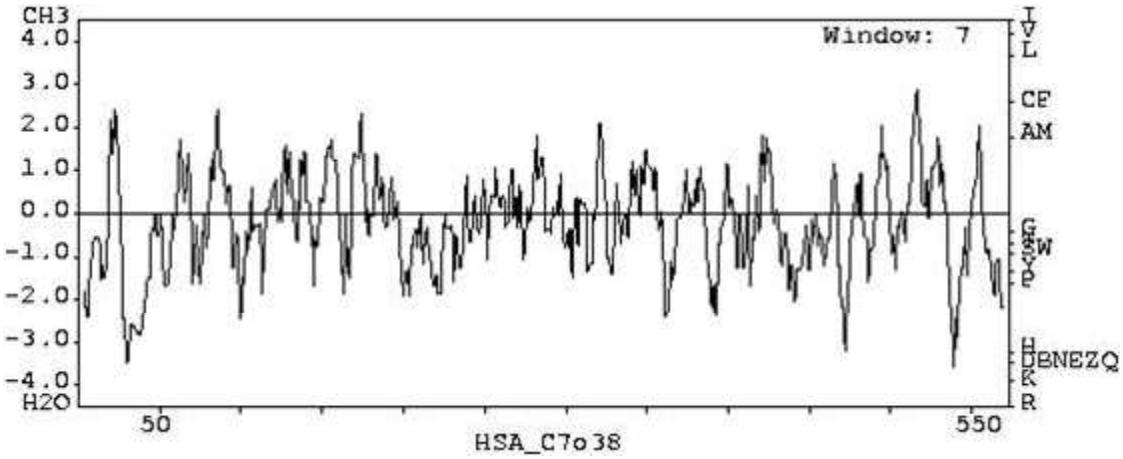
C7orf38 is located on chromosome 7 in the human genome. The gene is expressed in nearly all tissue types at very low levels. Evolutionarily, it can be found throughout the kingdom animalia. While the function of the protein is not fully understood, bioinformatic tools have shown that the protein bares much similarity to zinc finger or transposase proteins. Many of its orthologs, paralogs, and neighboring genes have been shown to possess zinc finger domains. The protein contains a hAT dimerization domain nears its C-terminus. This domain is highly conserved in transposase enzymes.

Gene

C7orf38 is located on Chromosome 7 at q22.1. Its genomic sequence contains 5,612 bp. The predominant transcript contains two exons and is 2,507 bp in length. The translated protein contains 573 amino acids.

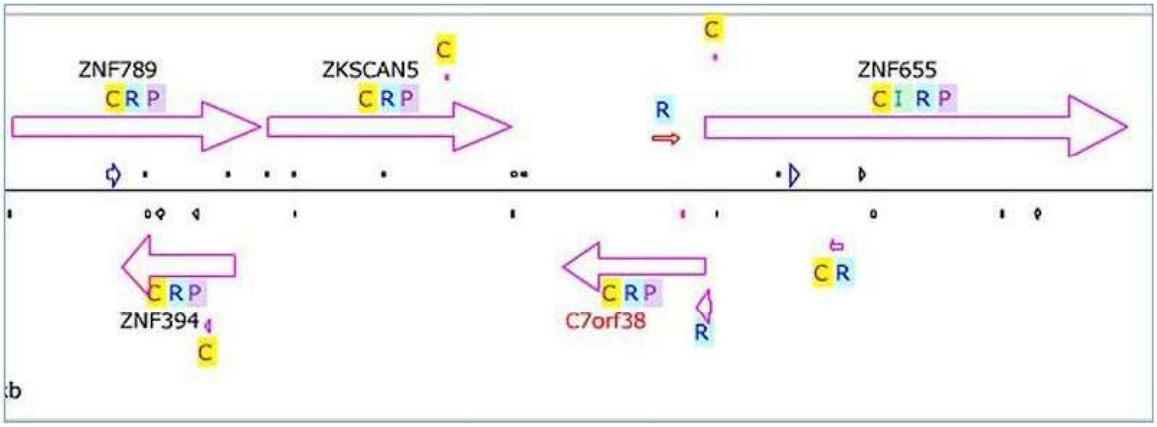
Protein composition

The 573 amino acid protein has a molecular weight of 66,280.05. The isoelectric point was found to occur at a pH of 5.775, about 1.6 pH lower than that of the average human pH. Two deviations from prototypical human proteins are evident. The protein contains less than expected number of glycine residues, and is rich in leucine residues. There are not sections of strong hydrophobicity or hydrophilicity. Thus, it is not predicted to be a transmembrane protein.



Gene neighborhood

The four genes in closest proximity to C7orf38 on chromosome 7 exhibit similar function, many of which are transcription factors.



Name	Orientation	Function
ZNF789	Start: 98,908,451 bp from pter End: 98,923,153 bp from pter Size: 14,703 bases Orientation: plus strand	The gene encodes the zinc finger protein 789. Functionally, the gene has been proposed to participate in regulation of transcription. It is expected to use zinc ion binding.
ZNF394	Start: 98,928,790 bp from pter End: 98,935,813 bp from pter Size: 7,024 bases Orientation : minus strand	The gene encodes zinc finger protein 394. Over expression over ZNF394 inhibits the transchription of c-jun and Ap-1. Suggesting that it is a transcriptional repressor.
ZKSCAN5	Start: 98,940,209 bp from pter End: 98,969,381 bp from pter Size: 29,173 bases Orientation: plus strand	The gene encodes zinc finger with KRAB and SCAN domains 5. This gene encodes a zinc finger protein of the Kruppel family. The protein contains a SCAN box and a KRAB A domain.
ZNF655	Start: 98,993,981 bp from pter End: 99,012,012 bp from pter Size: 18,032 bases Orientation: plus strand	The gene encodes zinc finger protein 655. Numerous alternatively spliced transcripts encoding distinct isoforms have been discovered.
Mihuya	Start: 99,149,738 bp from	The Mihuya gene does not encode a large or

pter End: 99,149,626 bp from pter Size: 112 bases Orientation: plus strand	known functional protein. The antisense relationship to C7orf38 raises the possibility for regulation of expression.

Paralogs

Eight paralogs are found in the human proteome. Similar to the neighboring genes, many of the paralogs function as zinc fingers, or transcription factors.

Name	NCBI Accession Number	Length (AA)	% Identity to C7orf38	% Similarity to C7orf38
hypothetical protein LOC285550	NP_001138663.1	657	79	91
zinc finger MYM-type protein 6	NP_009098.3	1325	38	60
SCAN domain-containing protein 3	NP_443155.1	1325	39	60
zinc finger BED domain-containing protein 5	NP_067034.2	692	35	57
transposon-derived Buster3 transposase-like	NP_071373.2	594	32	53
general transcription factor II-I repeat domain-containing protein 2B	NP_001003795.1	949	25	46
GTF2I repeat domain containing 2	NP_775808.2	949	24	45
EPM2A interacting protein 1	NP_055620.1	607	22	42

Orthologs

Orthologs to C7orf38 can be traced back evolutionarily through plants. The following is not an extensive list of orthologs. It is intended to provide an evolutionary overview of the conservation of C7orf38.

Common Name	Genus & Species	NCBI Accession Number	Length (AA)	% Identity to C7orf38	% Similarity to C7orf38
Chimp	Pan troglodytes	XP_001139775.1	573	99	99
Monkey Macaque	Macaca fascicularis	BAE01234.1	573	96	98
Horse	Equus caballus	XP_001915370.1	573	81	84

Pig	Sus scrofa	XP_001929194	1323	39	61
Cow	Bos taurus	XP_875656.2	1320	38	61
Mouse	Mus musculus	CAM15594.1	1157	37	60
Domestic Dog	Canis lupus familiaris	ABF22701.1	609	37	60
Rat	Rattus Rattus	NP_001102151.1	1249	37	59
Opposum	Monodelphis domestica	XP_001372983.1	608	37	59
Chicken	Gallus Gallus	XP_424913.2	641	37	58
Frog	Xenopus (Silurana) tropicalis	ABF20551.1	656	37	56
Zebra Fish	Danio Rerio	XP_001340213.1	609	37	56
Pea Aphid	Acyrthosiphon pisum	XP_001943527.1	659	36	54
Beatle	Tribolium castaneum	ABF20545.1	599	35	55
Sea Squirt	Ciona intestinalis	XP_002119512.1	524	34	52
Hydra	Hydra magnipapillata	XP_002165429.1	572	29	52
Puffer Fish	Tetraodon nigroviridis	CAF95678.1	539	28	47
Mosquito	Anopheles gambiae	XP_558399.5	591	28	47
Sea Urchin	Strongylocentrotus purpuratus	ABF20546.1	625	27	47
Grass Plant	Sorghum bicolor	XP_002439156.1	524	25	40
Tree Broad Leaf	Populus trichocarpa	XP_002319808.1	788	21	39

Structure

Protein

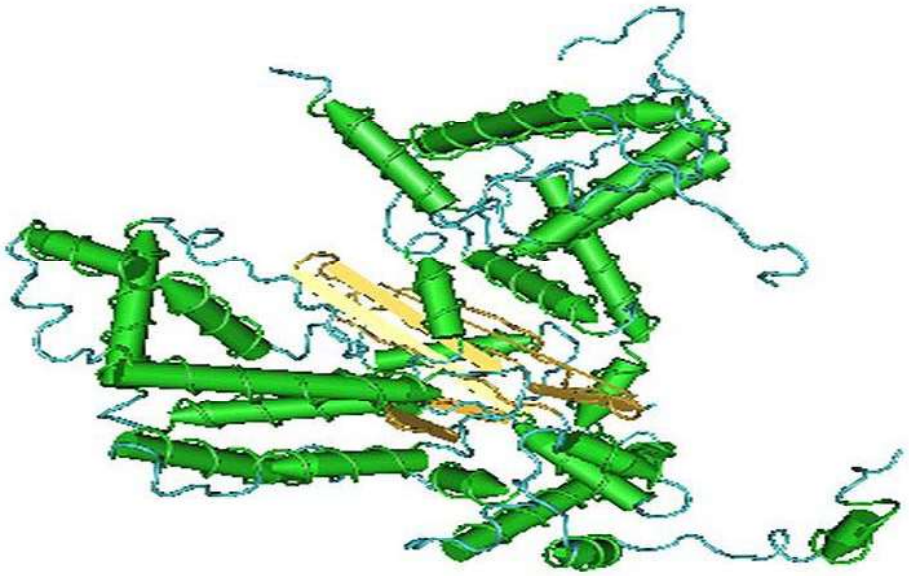
CBLast was used to determine a structurally related protein with experimentally determined structure. The protein Hermes DNA transposase, of the Hermes DBD superfamily, was shown to be structurally similar (Evalue: 1E-6).

hAT Dimerization Domain

Identifiers	
Symbol	hAT
Pfam	PF05699

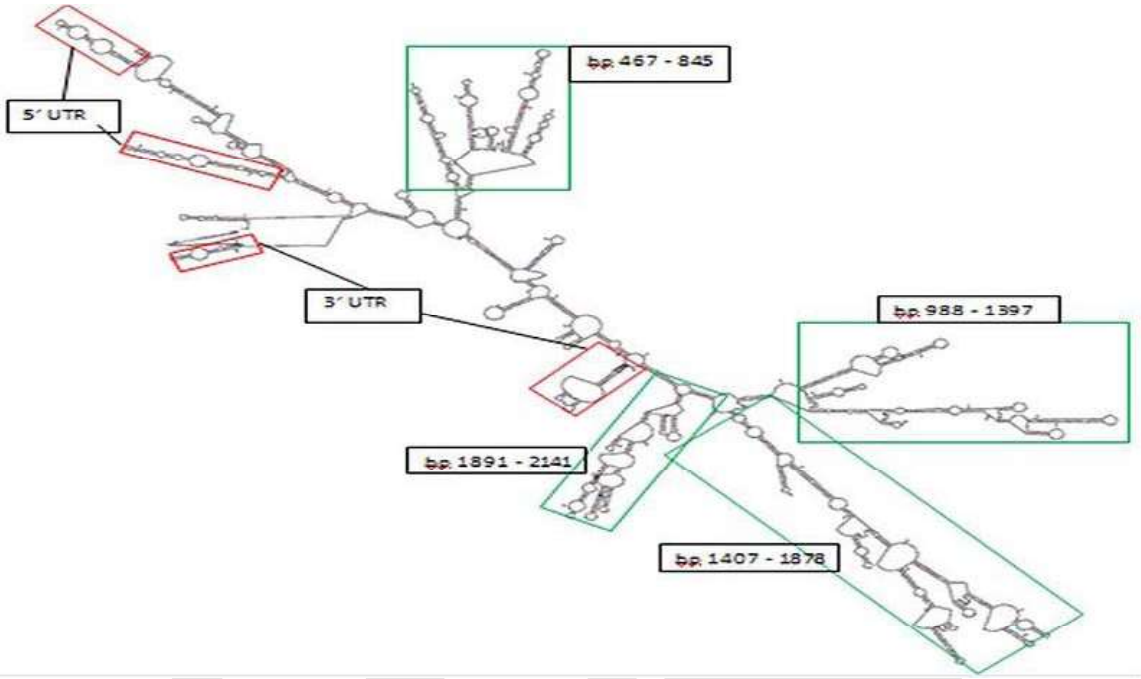
InterPro	IPR008906

The hAT dimerization domain is found at the C-terminus of transposase elements belonging to the Activator superfamily (hAT element superfamily). The isolated dimerization domain forms extremely stable dimers in vitro.



mRNA

The MFOLD program available at Rensselaer BioInformatics Server was used to predict secondary structure of the mature mRNA sequence. The primary sequence of the mRNA secondary structures displayed high levels of conservation in orthologs, suggesting structural importance.



Tissue distribution

The gene appears to be expressed in most tissue types. Very low levels of expression were observed through est profiles, and no deviation was observed between health or developmental states.

adipose tissue	0		0/13157	mouth	14		1/67218
adrenal gland	29		1/33344	muscle	9		1/108172
ascites	0		0/40058	nerve	0		0/15823
bladder	33		1/30128	ovary	9		1/102639
blood	0		0/124115	pancreas	0		0/215277
bone	0		0/71799	parathyroid	0		0/20646
bone marrow	0		0/49119	pharynx	0		0/41509
brain	9		10/1104749	pituitary gland	0		0/16729
cervix	0		0/48491	placenta	10		3/284160
connective tissue	20		3/149585	prostate	10		2/190663
ear	0		0/16341	salivary gland	0		0/20271
embryonic tissue	4		1/215834	skin	14		3/211658
esophagus	0		0/20211	spleen	18		1/54049
eye	4		1/211506	stomach	0		0/97179
heart	22		2/90302	testis	12		4/331401
intestine	0		0/235719	thymus	24		2/81182
kidney	14		3/212558	thyroid	0		0/47953
larynx	0		0/24481	tonsil	0		0/17042
liver	0		0/208419	trachea	0		0/52428
lung	11		4/338185	umbilical cord	0		0/13761
lymph	0		0/44401	uterus	17		4/233964
lymph node	10		1/91914	vascular	0		0/51942
mammary gland	6		1/154501				

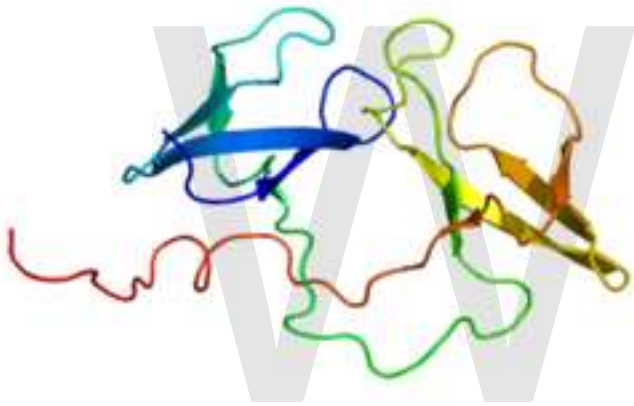
adrenal tumor	0		0/12864
bladder carcinoma	0		0/17758
breast (mammary gland) tumor	10		1/94657
cervical tumor	0		0/34586
chondrosarcoma	0		0/82864
colorectal tumor	0		0/115097
esophageal tumor	0		0/17292
gastrointestinal tumor	0		0/119842
germ cell tumor	0		0/264767
glioma	0		0/107554
head and neck tumor	7		1/137381
kidney tumor	14		1/69385
leukemia	10		1/96638
liver tumor	0		0/96673
lung tumor	0		0/103500
lymphoma	13		1/72055
non-neoplasia	20		2/97518
normal	11		39/3375995
ovarian tumor	12		1/77207
pancreatic tumor	0		0/104956
primitive neuroectodermal tumor...	0		0/126443
prostate cancer	0		0/103685
retinoblastoma	0		0/46512
skin tumor	0		0/125580
soft tissue/muscle tissue tumor	0		0/125854
uterine tumor	22		2/90823

Chapter- 8

FMR1, FOXP1 and FOXP2

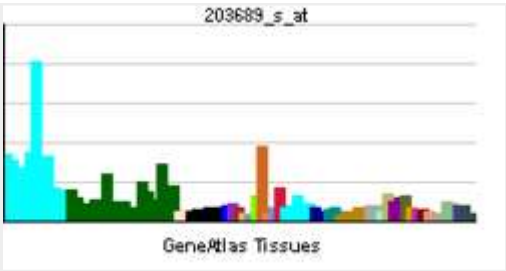
FMR1

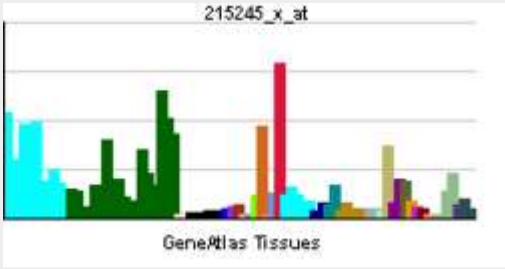
Fragile X mental retardation 1



PDB rendering based on 2bkd.

Available structures	
Identifiers	
Symbols	FMR1; FMRP; FRAXA; MGC87458
External IDs	OMIM: 309550 MGI: 95564 HomoloGene: 1531 GeneCards: FMR1 Gene
RNA expression pattern	

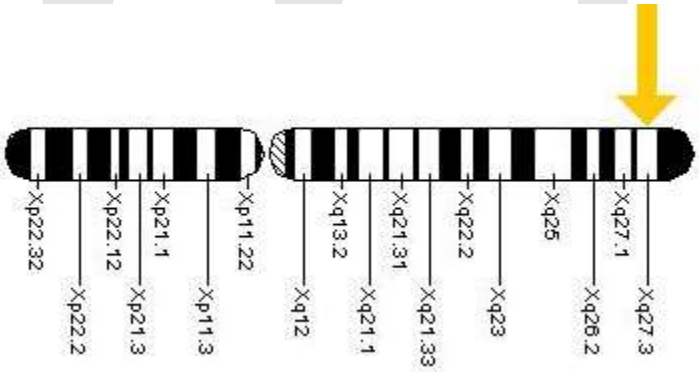




More reference expression data

Orthologs

Species	Human	Mouse
Entrez	2332	14265
Ensembl	ENSG00000102081	ENSMUSG00000000838
UniProt	Q06787	Q6AXB7
RefSeq (mRNA)	NM_002024	XM_990299
RefSeq (protein)	NP_002015	XP_995393
Location (UCSC)	Chr X: 146.8 - 146.84 Mb	Chr X: 64.94 - 64.98 Mb



Location of *FMR1* on the X chromosome.

FMR1 (**fragile X mental retardation 1**) is a human gene that codes for a protein called *fragile X mental retardation protein*, or FMRP. This protein is normally made in many tissues, especially in the brain and testes. It may play a role in the development of synaptic connections between nerve cells in the brain, where cell-to-cell communication occurs. The connections between nerve cells can change and adapt over time in response to experience (a characteristic called synaptic plasticity). FMRP may help regulate synaptic plasticity, which is important for learning and memory.

One region of the *FMR1* gene contains a 3 base *Variable Number Tandem Repeat* (VNTR, or more specifically, a trinucleotide repeat). The sequence *CGG* is repeated a number of times. In most healthy individuals, the number of CGG repeats ranges from fewer than 10 to about 40, with the median at about 29 repeats.

The *FMR1* gene is located on the long (q) arm of the X chromosome at position 27.3, from base pair 146,699,054 to base pair 146,738,156.

Related conditions

Fragile X syndrome: Almost all cases of fragile X syndrome are caused by expansion of the CGG trinucleotide repeat in the *FMR1* gene. In these cases, CGG is abnormally repeated from 200 to more than 1,000 times, which makes this region of the gene unstable. As a result, the *FMR1* gene is methylated, which silences the gene (it is turned off and does not make any protein). Without adequate FMRP, severe learning deficits or mental retardation can develop, along with physical abnormalities seen in fragile X syndrome.

Some, fewer than 1 %, of all cases of fragile X syndrome are caused by mutations that delete part or all of the *FMR1* gene, or change a base pair, leading to a change in one of the amino acids in the gene. These mutations disrupt the 3-dimensional shape of FMRP or prevent the protein from being synthesized, leading to the signs and symptoms of fragile X syndrome.

A CGG sequence in the *FMR1* gene that is repeated about 55 to 200 times is described as a premutation expansion. Men, and probably some women, with this premutation do not have fragile X syndrome, but are at increased risk of developing a disorder known as fragile X-associated tremor/ataxia syndrome (FXTAS). FXTAS is characterized by progressive problems with movement (ataxia), tremor, memory loss, loss of sensation in the lower extremities (peripheral neuropathy) and mental and behavioral changes. The disorder usually develops late in life.

Although most men and women with the premutation are intellectually normal, some of these individuals have mild versions of the physical features seen in fragile X syndrome (such as prominent ears) and may experience emotional problems such as anxiety or depression. About 20 % of women who carry a premutation expansion in the *FMR1* gene experience premature ovarian failure (POF). POF is a loss of ovarian function in women younger than age 40, which can result in infertility. However, as Fragile X is an X-linked recessive disorder, most females with premutation or even full mutation do not exhibit symptoms due to a second, normal X-chromosome.

Researchers have found that some children with a premutation expansion in the *FMR1* gene have learning disabilities, mental retardation, or disorders in the autism spectrum, characterized by deficits in communication and social interaction.

Interactions

FMR1 has been shown to interact with *FXR2*, *CYFIP1*, *CYFIP2*, *NUFIP1*, *FXR1*, and *NUFIP2*.

FOXP1

Forkhead box P1

Identifiers		
Symbols	FOXP1; 12CC4; FLJ23741; HSPC215; MGC12942; MGC88572; MGC99551; QRF1; hFKH1B	
External IDs	OMIM: 605515 MGI: 1914004 HomoloGene: 13092 GeneCards: FOXP1 Gene	
Orthologs		
Species	**Human**	**Mouse**
Entrez	27086	108655
Ensembl	ENSG00000114861	ENSMUSG00000030067
UniProt	Q9H334	Q6P221
RefSeq (mRNA)	NM_001012505	NM_053202
RefSeq (protein)	NP_001012523	NP_444432
Location (UCSC)	Chr 3: 71.09 - 71.72 Mb	Chr 6: 98.9 - 99.23 Mb

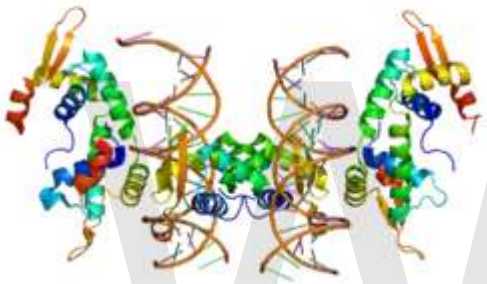
FOXP1 ("forkhead box P1") is a gene that is necessary for the proper development of the brain and lung in mammals. It is a member of the large FOX family of transcription factors.

This gene belongs to subfamily P of the forkhead box (FOX) transcription factor family. Forkhead box transcription factors play important roles in the regulation of tissue- and

cell type-specific gene transcription during both development and adulthood. Forkhead box P1 protein contains both DNA-binding- and protein-protein binding-domains. This gene may act as a tumor suppressor as it is lost in several tumor types and maps to a chromosomal region (3p14.1) reported to contain a tumor suppressor gene(s). Alternative splicing results in multiple transcript variants encoding different isoforms.

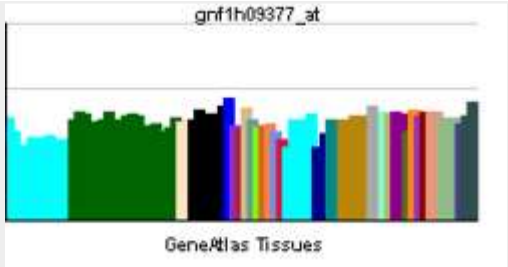
FOXP2

Forkhead box P2



PDB rendering based on 2a07.

Available structures	
Identifiers	
Symbols	FOXP2; CAGH44; DKFZp686H1726; SPCH1; TNRC10
External IDs	OMIM: 605317 MGI: 2148705 HomoloGene: 33482 GeneCards: FOXP2 Gene
RNA expression pattern	



More reference expression data

Orthologs

Species	Human	Mouse
Entrez	93986	114142
Ensembl	ENSG00000128573	ENSMUSG00000029563
UniProt	O15409	Q8BQ27
RefSeq (mRNA)	NM_014491	NM_053242
RefSeq (protein)	NP_055306	NP_444472
Location (UCSC)	Chr 7: 113.84 - 114.12 Mb	Chr 6: 15.14 - 15.39 Mb

Forkhead box protein P2 also known as **FOXP2** is a protein that in humans is encoded by the *FOXP2* gene, located on human chromosome 7 (7q31, at the SPCH1 locus). *FOXP2* orthologs have also been identified in all mammals for which complete genome data are available. The FOXP2 protein contains a forkhead-box DNA-binding domain, making it a member of the FOX group of transcription factors, involved in regulation of gene expression. In addition to this characteristic forkhead-box domain, the protein contains a polyglutamine tract, a zinc finger and a leucine zipper.

In humans, mutations of FOXP2 cause a severe speech and language disorder. Versions of FOXP2 exist in similar forms in distantly related vertebrates; functional studies of the gene in mice and in songbirds indicate that it is important for modulating plasticity of neural circuits. Outside the brain FOXP2 has also been implicated in development of other tissues such as the lung and gut. FOXP2 directly regulates a large number of downstream target genes.

One particular target that is directly downregulated by FOXP2 in human neurons is the CNTNAP2 gene, a member of the neurexin family; variants in this target gene have been associated with common forms of language impairment. Two amino-acid substitutions distinguish the human FOXP2 protein from that found in chimpanzees, but only one of these two changes is unique to humans. Evidence from genetically manipulated mice and human neuronal cell models suggests that these changes affect the neural functions of FOXP2.

Function

FOXP2 is required for proper brain and lung development. Knockout mice with only one functional copy of the FOXP2 gene have significantly reduced vocalizations as pups. Knockout mice with no functional copies of FOXP2 are runted, display abnormalities in

brain regions such as the Purkinje layer, and die an average of 21 days after birth from inadequate lung development.

Different studies of FOXP2 in songbirds suggest that FOXP2 may regulate genes involved in neuroplasticity: During song learning FOXP2 is upregulated in brain regions critical for song learning in young zebra finches. Knockdown of FOXP2 in Area X of the basal ganglia of these birds results in incomplete and inaccurate song imitation. Similarly, in adult canaries higher FOXP2 levels also correlate with song changes. In addition, levels of FOXP2 in adult zebra finches are significantly lower when males direct their song to females than when they sing song in other contexts. Differences between song-learning and non-song-learning birds have been shown to be caused by differences in FOXP2 gene expression, rather than differences in the amino acid sequence of the FOXP2 protein.

FOXP2 also has possible implications in the development of bat echolocation.

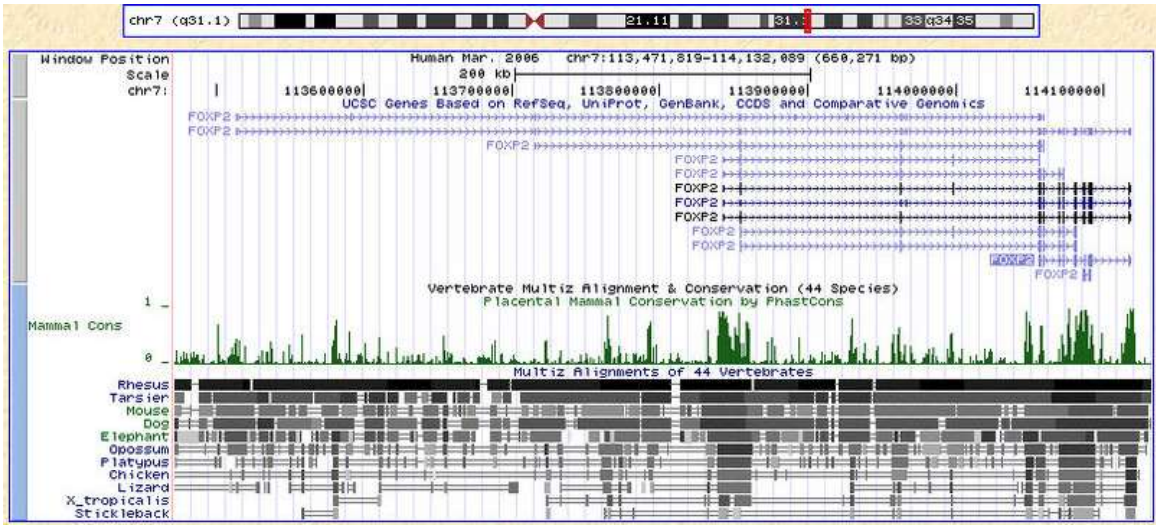
Clinical significance

Several cases of developmental verbal dyspraxia in humans have been linked to mutations in the FOXP2 gene. Such individuals have little or no cognitive handicaps but are unable to correctly perform the coordinated movements required for speech. fMRI analysis of these individuals performing silent verb generation and spoken word repetition tasks showed underactivation of Broca's area and the putamen, brain centers thought to be involved in language tasks. Because of this, FOXP2 has been dubbed the "language gene." People with this mutation also experience symptoms not related to language (not surprisingly, as FOXP2 is known to affect development in other parts of the body as well). Scientists have also looked for associations between FOXP2 and autism and both positive and negative findings have been reported.

There is some evidence that the linguistic impairments associated with a mutation of the FOXP2 gene are not simply the result of a fundamental deficit in motor control. For example:

- the impairments include difficulties in comprehension;
- brain imaging of affected individuals indicates functional abnormalities in language-related cortical and basal/ganglia regions, demonstrating that the problems extend beyond the motor system.

Evolution



Human FOXP2 gene and evolutionary conservation is shown in a multiple alignment (at bottom of figure) in this image from the UCSC Genome Browser. Note that conservation tends to cluster around coding regions (exons).

The FOXP2 protein sequence is generally thought to be highly conserved. Similar FOXP2 proteins can be found in songbirds, fish, and reptiles such as alligators. However, recent studies in bats (chiroptera) has prompted some researchers to conclude that FoxP2 is not well conserved in non-human mammals and write: "We found that contrary to previous reports, FoxP2 is not highly conserved across all nonhuman mammals but is extremely diverse in echolocating bats." Aside from a polyglutamine tract, human FOXP2 differs from chimp FOXP2 by only two amino acids, mouse FOXP2 by only 3 amino acids, and zebra finch FOXP2 by only 7 amino acids. One of the two amino acid difference between human and chimps also arose independently in carnivores and bats. A recent extraction of DNA from Neanderthal bones indicates that Neanderthals had the same version (allele) of the FOXP2 gene as modern humans.

Some researchers have speculated that the two amino acid differences between chimps and humans led to the evolution of language in humans. Others, however, have been unable to find a clear association between species with learned vocalizations and similar mutations in FOXP2. Insertion of both human mutations into mice, whose version of FOXP2 otherwise differs from the human and chimpanzee versions in only one additional base pair, causes changes in vocalizations as well as other behavioral changes, such as a reduction in exploratory tendencies; a reduction in dopamine levels and changes in the morphology of certain nerve cells are also observed. It may also be, based on general observations of development and songbird results, that any difference between humans and non-humans would be due to regulatory sequence divergence (affecting where and when FOXP2 is expressed) rather than the two amino acid differences mentioned above. However the mutation rate of *FOXP*2 is slower in the human lineage than in the lineage before the human-chimpanzee split, and proposed that purifying selection would not have relaxed due to negative deleterious effects. Thus, it was most likely positive selection that

drove the two amino acid differences to fixation in humans, suggesting that differences between humans and non-humans are a result of the two amino acid changes.

Li et al. (2007) found that exons 7 and 17 of *FoxP*2 in bats are highly variable and not as conserved as in other vertebrates. Twenty-two sequences of non-bat eutherian mammals revealed a total number of 20 nonsynonymous mutations in contrast to half that number of bat sequences, which showed 44 nonsynonymous mutations. Interestingly, all cetaceans share three amino acid substitutions, but there are not differences between echolocating and non-echolocating baleen cetaceans. Within bats, however, amino acid variation correlated with different echolocating types. Accelerated evolution in bats is likely due to positive selection on echolocation.

Discovery

The human gene was identified through molecular investigations of an unusual family known as the KE family. Researchers in London discovered that around half of the family members - fifteen individuals across three generations - suffered from severe speech and language deficits. Remarkably, the transmission of the disorder from one generation to the next was consistent with autosomal dominant inheritance i.e. mutation of only a single gene on an autosome (non-sex chromosome) acting in a dominant fashion. This is one of the few known examples of Mendelian (monogenic) inheritance for a disorder affecting speech and language skills, which typically have a complex basis involving multiple genetic risk factors.

In the mid-1990s Oxford scientists began to search for the damaged gene in the KE family, performing a genome-wide scan of DNA samples taken from the affected and unaffected members. This scan confirmed autosomal dominant monogenic inheritance and localized the gene responsible to a small section of chromosome 7. The locus was given the official name "SPCH1" (for speech-and-language-disorder-1) by the Human Genome Nomenclature committee. Mapping and sequencing of the chromosomal region was performed with the aid of bacterial artificial chromosome clones. Around this time, the researchers identified an individual who was unrelated to the KE family, but had a similar type of speech and language disorder. In this case the child, known as CS, carried a chromosomal rearrangement (a translocation) in which part of chromosome 7 had become exchanged with part of chromosome 5. The site of breakage of chromosome 7 was located within the SPCH1 region.

The team went on to pinpoint the precise position of the chromosome-7 breakage in case CS, and found that it lay directly in the middle of a protein-coding gene. Using a combination of bioinformatics and RNA analyses they deciphered the full coding region of the gene, discovering that it encoded a novel member of the forkhead-box (FOX) group of transcription factors. As such, it was assigned with the official name of FOXP2. When the researchers sequenced the FOXP2 gene in the KE family they uncovered a heterozygous point mutation that was shared by all the affected individuals, but absent from unaffected members and a large panel of controls from the general population. This mutation yields an amino-acid substitution at a crucial point of the DNA-binding domain

of the FOXP2 protein, disrupting its function. Further screening of the gene has since identified multiple additional cases of FOXP2 disruption, including different point mutations and chromosomal rearrangements, providing further evidence that damage to one copy of this gene is sufficient to derail speech and language development.

FoxP2 in songbirds

In zebra finch, FoxP2 mRNA expression is observed in structures analogous to FoxP2 rich structures in the human brain including the basal ganglia, cortex (referred to pallium in birds) and the cerebellum. Notably, FoxP2 is expressed in the song control nucleus Area X, which is a basal ganglia-like nucleus dedicated to singing behaviors. In zebra finch, FoxP2 mRNA shows a developmental increased during sensorimotor learning without any change in other FoxP2 enriched structures. Reinforcing the idea that this increase is tied to sensorimotor learning (as opposed to age per se), the canary- which relearns song every year- shows a similar increase in FoxP2 during late summer and early-fall, a time period which corresponds to their yearly sensorimotor learning. Interestingly, blocking this upregulation using lentiviral mediated knockdown in zebra finches impairs song learning and increases variability upon maturation.

Interactions

FOXP2 has been shown to interact with CTBP1.

Chapter- 9

Gcn2

Serine/threonine-protein kinase GCN2

Crystal structure of GCN2

Identifiers	
Symbol	GCN2
Alt. symbols	AAS1
Entrez	851877
PDB	1zyc
UniProt	P15442

GCN2 (general control nonrepressed 2) is a serine/threonine-protein kinase that senses amino acid deficiency through binding to uncharged transfer RNA (tRNA). It plays a key role in modulating amino acid metabolism as a response to nutrient deprivation.

Introduction

GCN2 is the single eukaryotic initiation factor 2α kinase (eIF2α) in *Saccharomyces cerevisiae*. It inactivates eIF2α by phosphorylation under conditions of amino acid deprivation, resulting in repression of general protein synthesis whilst allowing selected mRNA such as GCN4 to be translated due to regions upstream of the coding sequence.

Elevated levels of GCN4 stimulate the expression of amino acid biosynthetic genes, which code for enzymes required to synthesize all 20 major amino acids.

Structure

Protein kinase GCN2 is a multidomain protein and its C-terminus contains a region homologous to histidyl-tRNA synthetase (HisRS) next to the kinase catalytic moiety. This HisRS-like region forms a dimer and dimerization is required for GCN2 function. The crucial contribution to GCN2 function is the promotion of tRNA binding and the stimulation of the kinase domain via physical interaction.

Binding of uncharged tRNA to this synthetase-like domain induces a conformational change in which the GCN2 domains rotate 180° normal to the dimerization surface and thereby transpose from their antiparallel to a parallel orientation. Subsequently the autoinhibited form of GCN2 is activated.

GCN2 autoinhibition results from a conformation that restricts ATP binding. ATP binding induces autophosphorylation of an activation loop which leads to maximal GCN2 kinase activity.

Function

Regulation of translation

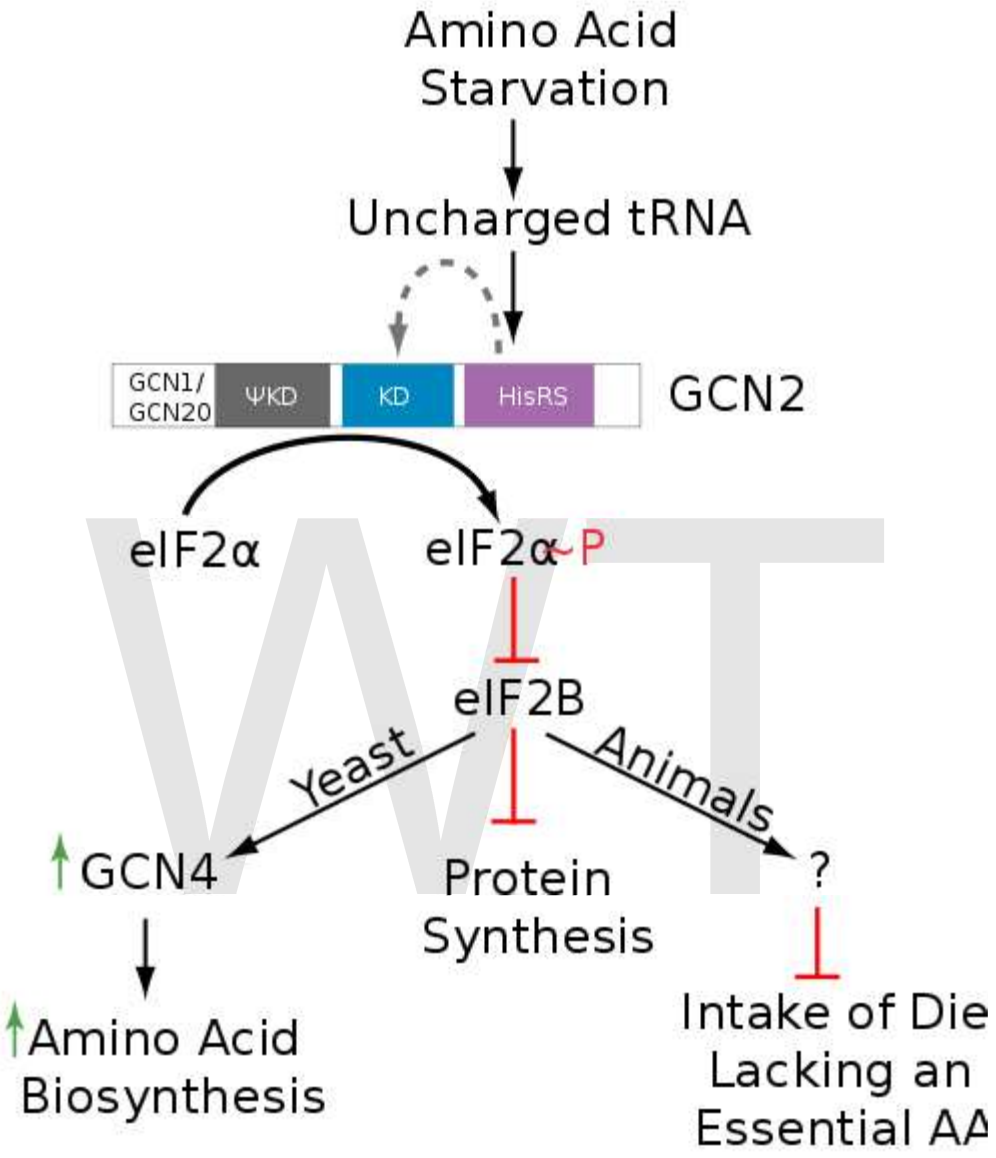


Figure 1:Overview over the functions of GCN2. (GCN1/GCN20=GCN1p/GCN20p binding site; PsiKD = unknown function; KD = Kinase Domain; HisRS = histidyl-tRNA synthetase) Adapted from

GCN2 inhibits general translation by phosphorylation of eIF-2α at serine 51 within 15 min of amino acid deprivation, which then subsequently increases the affinity for its guanine exchange factor EIF2B to sequester eIF-2α leading to reduced formation of the ternary complex (TC) consisting of eIF2, GTP and the initiator Met-tRNA required for translation initiation. eIF2 containing a phosphorylated alpha subunit shows an increased affinity for its only GEF, eIF2B, but eIF2B is only able to exchange GDP with GTP from

unphosphorylated eIF2. So the recycling of eIF2, needed for TC formation, is inhibited by phosphorylation of eIF-2α, which in the end leads to a reduction of global translation rates.

An opposing effect of the reduced availability of TC is the induction of GCN4 expression by translational regulation. Four short ORF's exist in the leader of the GCN4 mRNA. 40S Ribosomal Subunits scanning the mRNA from 5' have TC bound and translate the first upstream ORF (uORF). Under non-starving condition there is enough ternary complex that the subunits rebind it before they reach uORF 4. Translation is again initiated, uORF2,3 or 4 translated and the 40S Subunits subsequently dissociate from GCN4 mRNA. Under starving conditions there is less TC present. Some of the 40S Subunits are not able to rebind TC before they reach uORF 4 but eventually rebind TC before reaching GCN4 coding sequence. Therefore the reduction in TC formation resulting from GCN2 activation by amino acid starvation leads to the induction of GCN4 translation. GCN4 is the primary regulator in response to amino acid starvation, termed general amino acid control (GAAC). It acts as a transcription factor and activates several genes required for amino acid synthesis.

Recently GCN2 has also been implicated in directing eating behavior in mammals by phosphorylating eIF-2α in the anterior Piriform cortex (APC) of the brain. The molecular mechanisms governing this function are not yet known, but a basic zipper transcription factor called ATF4 is a possible candidate. ATF4 is related to GCN4.

Cell Cycle Control

GCN2 also regulates the cell cycle by delaying entry into S phase upon ultraviolet (UV) radiation and exposure to methyl methanesulfonate (MMS). Thereby the cell prevents passing the G1 checkpoint and starting DNA replication when the DNA is damaged. It has been hypothesized, that UV induces nitric oxide synthase activation and $NO^{\cdot}$ production, which leads to the activation of GCN2 and that the cell cycle regulation by GCN2 is independent of eIF2α phosphorylation. Although the causal relationship between GCN2 and cell cycle delay is still under debate, it was suggested that the formation of the pre-replication complex is deferred by GCN2 upon UV-irradiation.

Lipid Metabolism

The absence of essential amino acids causes a downregulation of key components of the lipid synthesis such as the fatty acid synthase. Following leucine-deprivation in mammals, GCN2 decreases the expression of lipogenic genes via SREBP-1c. SREBP-1c actions upon genes regulating fatty-acid and triglyceride synthesis and is reduced by leucine deprivation in the liver in a GCN2-dependant manner.

Regulation

In amino acid replete cells, GCN2 is kept inactive via phosphorylation at serine 577, which is thought to depend on the activity of TORC1. Inactivation of TORC1 by

Rapamycin affects GCN2 and at least partly by dephosphorylation of serine 577. This leads to activation of GCN2 even in amino acid replete cells, probably by increasing the affinity of GCN2 for uncharged tRNA, so that even basal levels permit tRNA binding.

A second stimulatory input to GCN2 is exerted by a complex of GCN1/GCN20. GCN1/GCN20 shows structural similarity to eEF3, a factor important in the binding of tRNA to ribosomes. The GCN1/GCN20 complex physically interacts with GCN2 by binding to its N-terminus. It is thought that GCN1/GCN20 facilitates the transfer of tRNA from the ribosomal A site to the HisRS-like domain of GCN2.

Homologues

There are also GCN2 homologues in *Neurospora crassa*, *Drosophila melanogaster* and mice. Thus, GCN2 may be the most widespread and founding member of the eIF-2α kinase subfamily.

Chapter- 10

HLA-B

Major histocompatibility complex, class I, B

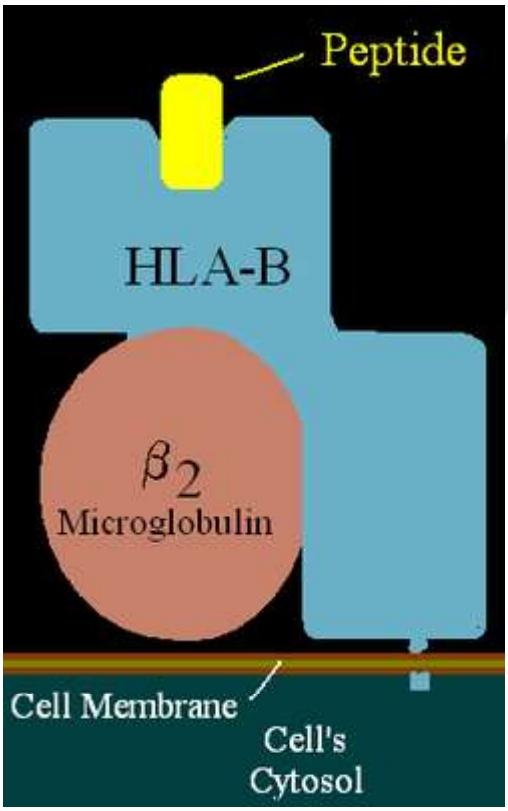
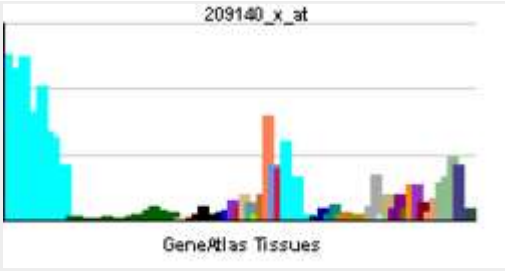


Illustration of HLA-B complexed peptide.

Available structures	
Identifiers	
Symbols	HLA-B; HLA B; SPDA1
External IDs	OMIM: 142830 HomoloGene: 83181 GeneCards: HLA-B Gene

RNA expression pattern



Orthologs

Species	Human	Mouse
Entrez	3106	547349
Ensembl	ENSG00000204523	n/a
UniProt	P01889	n/a
RefSeq (mRNA)	NM_005514	NM_001025208
RefSeq (protein)	NP_005505	NP_001020379
Location (UCSC)	Chr 6: 31.43 - 31.43 Mb	n/a

HLA-B (**major histocompatibility complex, class I, B**) is a human gene that provides instructions for making a protein that plays a critical role in the immune system. HLA-B is part of a family of genes called the human leukocyte antigen (HLA) complex. The HLA complex helps the immune system distinguish the body's own proteins from proteins made by foreign invaders such as viruses and bacteria.

HLA is the human version of the major histocompatibility complex (MHC), a gene family that occurs in many species. Genes in this complex are separated into three basic groups: class I, class II, and class III. In humans, the HLA-B gene and two related genes, HLA-A and HLA-C, are the major genes in MHC class I.

MHC class I genes provide instructions for making proteins that are present on the surface of almost all cells. On the cell surface, these proteins are bound to protein fragments (peptides) that have been exported from within the cell. MHC class I proteins display these peptides to the immune system. If the immune system recognizes the peptides as foreign (such as viral or bacterial peptides), it responds by destroying the infected cell.

The HLA-B gene has many different normal variations, allowing each person's immune system to react to a wide range of foreign invaders. Hundreds of versions (alleles) of HLA-B are known, each of which is given a particular number (such as HLA-B27). Closely related alleles are categorized together; for example, at least 28 very similar alleles are subtypes of HLA-B27. These subtypes are designated as HLA-B*2701 to HLA-B*2728.

The HLA-B gene is located on the short (p) arm of chromosome 6 at position 21.3, from base pair 31,429,845 to base pair 31,432,923.

Related conditions

Serotypes of HLA-B gene products

antigen	-	**Broad antigen**	Split antigens		
B7		**B5**	B51	B52	
B8		**B12**	B44	B45	
B13		**B14**	B64	B65	
B18		**B15**	B62	B63	B70
B27			B72	B75	B77
B35		**B16**	B38	B39	
B37		**B17**	B57	B58	
B41		**B21**	B49	B50	
B42		**B22**	B54	B55	B56
B46		**B40**	B60	B61	
B47					
B48					
B53					
B59					
B67					
B73					
B78					
B81					
B*82					
B*83					

"HLA-" prefix trimmed from serotype names.

Ankylosing spondylitis: A version of the HLA-B gene called HLA-B27 increases the risk of developing ankylosing spondylitis. It is uncertain how HLA-B27 causes this increased risk. Researchers speculate that HLA-B27 may abnormally display to the immune system peptides that trigger arthritis. Other research suggests that joint inflammation characteristic of this disorder may result from improper folding of the HLA-B27 protein

or the presence of abnormal forms of the protein on the cell surface. Although most patients with ankylosing spondylitis have the HLA-B27 variation, many people with this particular variation never develop the disorder. Other genetic and environmental factors are likely to affect the chances of developing ankylosing spondylitis and influence its progression.

HLA-B27 is associated with the spondyloarthropathies, a group of disorders that includes ankylosing spondylitis and other inflammatory joint diseases. Some of these diseases are associated with a common skin condition called psoriasis or chronic inflammatory bowel disorders (Crohn's disease and ulcerative colitis). One of the spondyloarthropathies, reactive arthritis, is typically triggered by bacterial infections of the gastrointestinal or genital tract. Following an infection, affected individuals may develop arthritis, back pain, and eye inflammation. Like ankylosing spondylitis, many factors probably contribute to the development of reactive arthritis and other spondyloarthropathies.

Other disorders: Several variations of the HLA-B gene are associated with adverse reactions to certain drugs. For example, two specific versions of this gene are related to increased drug sensitivity among the Han Chinese population. Individuals who have HLA-B*1502 are more likely to experience a severe skin disorder called Stevens-Johnson syndrome in response to carbamazepine (a drug used to treat seizures). Another version, HLA-B*5801, is associated with an increased risk of severe skin reactions in people treated with allopurinol (a drug used to treat gout, which is a form of arthritis caused by uric acid in the joints).

Among people with human immunodeficiency virus (HIV) infection, a version of HLA-B designated HLA-B*5701 is associated with an extreme sensitivity to abacavir. This drug is a treatment for HIV-1 that slows the spread of the virus in the body. People with abacavir hypersensitivity often develop a fever, chills, rash, upset stomach, and other symptoms when treated with this drug.

Several other variations of the HLA-B gene appear to play a role in the progression of HIV infection to acquired immunodeficiency syndrome (AIDS). AIDS is a disease that damages the immune system, preventing it from effectively defending the body against infections. The signs and symptoms of AIDS may not appear until 10 years or more after infection with HIV. Studies suggest that people with HIV infection who have HLA-B27 or HLA-B57 tend to progress more slowly than usual to AIDS. On the other hand, researchers believe that HIV-positive individuals who have HLA-B35 tend to develop the signs and symptoms of AIDS more quickly than usual. Other factors also influence the progression of HIV to AIDS.

Another version of the HLA-B gene, HLA-B53, has been shown to help protect against severe malaria. HLA-B53 is most common in West African populations, where malaria is a frequent cause of death in children. Researchers suggest that this version of the HLA-B gene may help the immune system respond more effectively to the parasite that causes malaria.

Chapter- 11

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

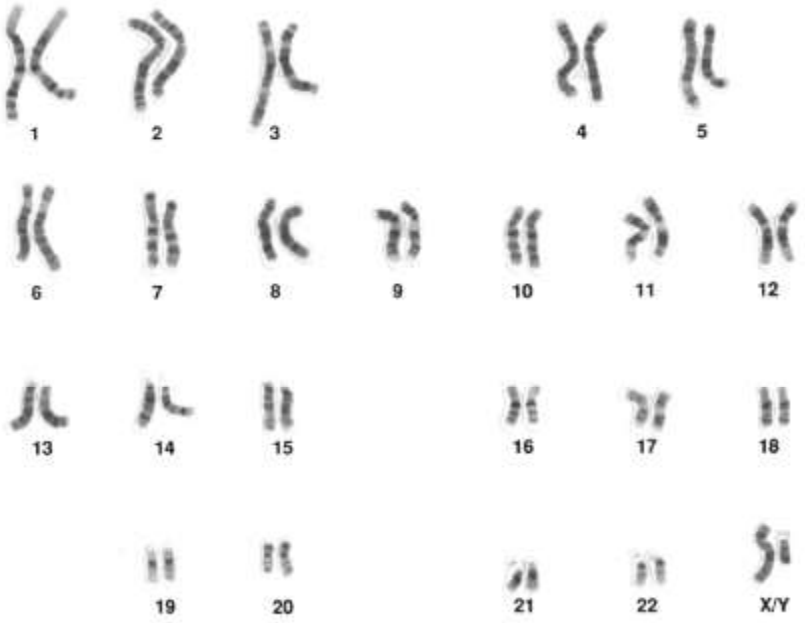
underlying phage evolution. Bacteriophage genome sequences can be obtained through direct sequencing of isolated bacteriophages, but can also be derived as part of microbial genomes. Analysis of bacterial genomes has shown that a substantial amount of microbial DNA consists of prophage sequences and 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- 12

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 ***gen****e* 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

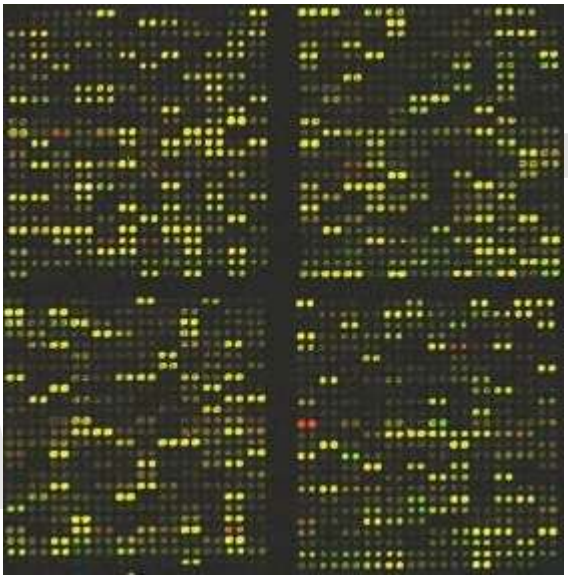
Genomes are more than the sum of an organism's genes and have traits that may be measured and studied without reference to the details of any particular genes and their products. Researchers compare traits such as *chromosome number* (karyotype), genome size, gene order, codon usage bias, and GC-content 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- 13

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

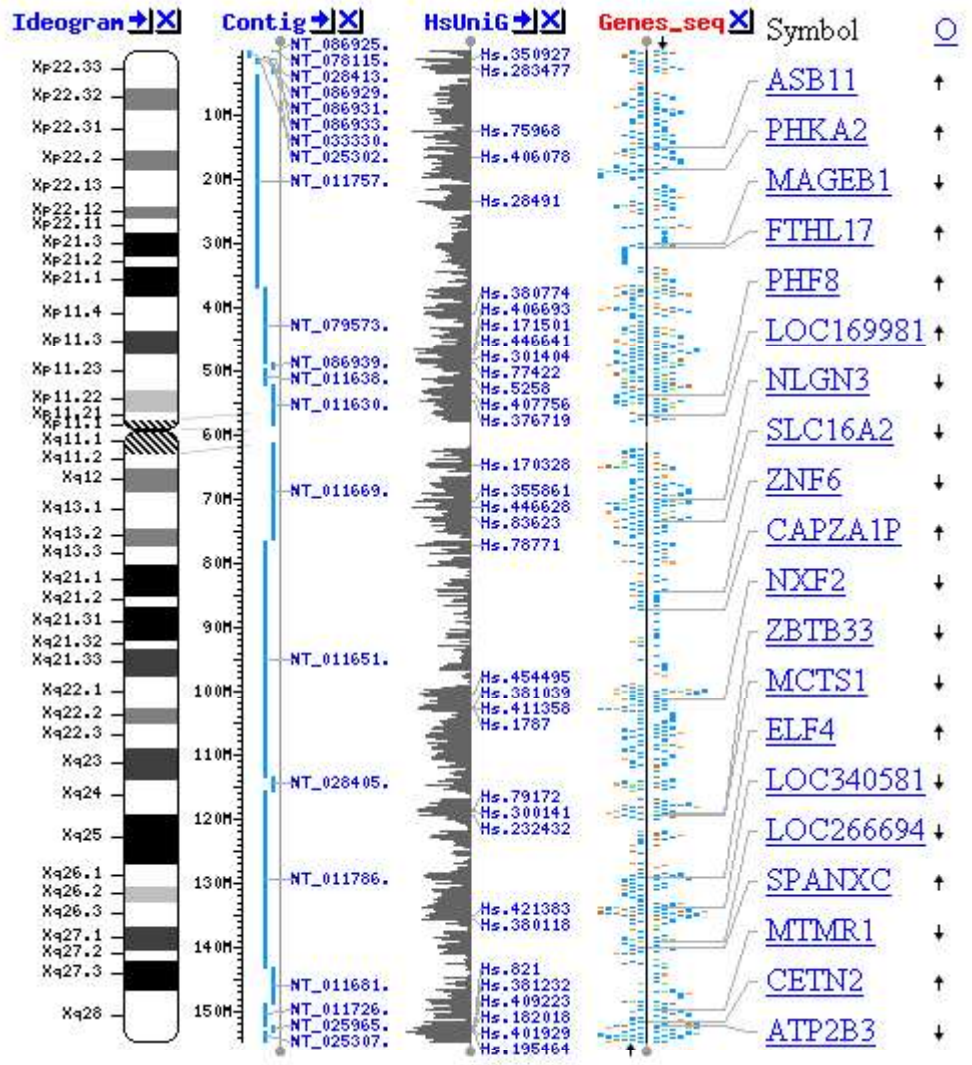
The Rosetta stone approach is a computation method of de novo protein function prediction, based on the hypothesis that some proteins involved in a given physiological process may exist as two separate genes in one organism and as a single gene in another. Genomes are scanned for sequences that are independent 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- 14

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- 15

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.

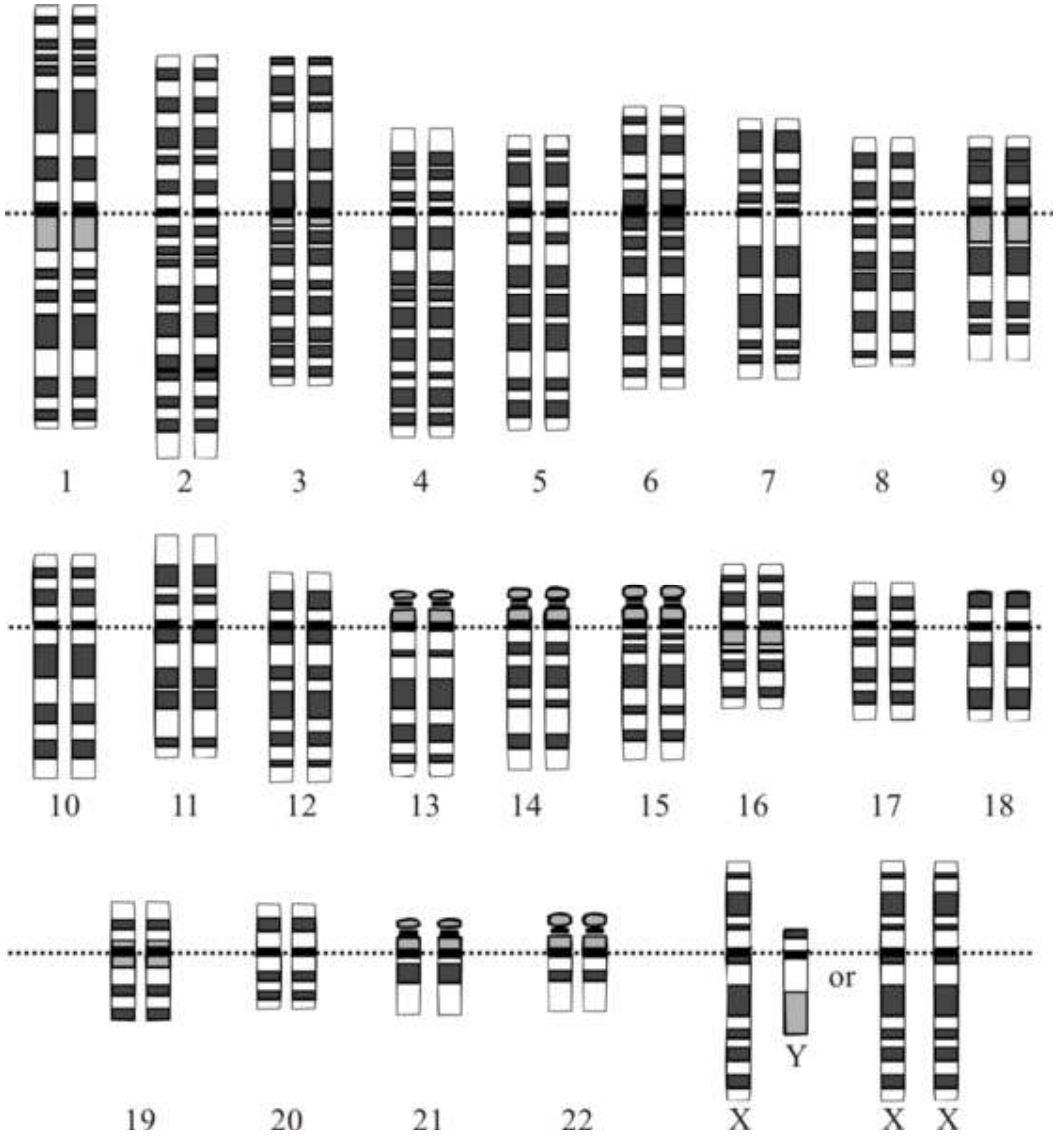
An interesting use of proteomics is using specific protein biomarkers to diagnose disease. A number of techniques allow to test for proteins produced during a particular disease, which helps to diagnose the disease quickly. Techniques include western blot, immunohistochemical staining, enzyme linked 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- 16

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.

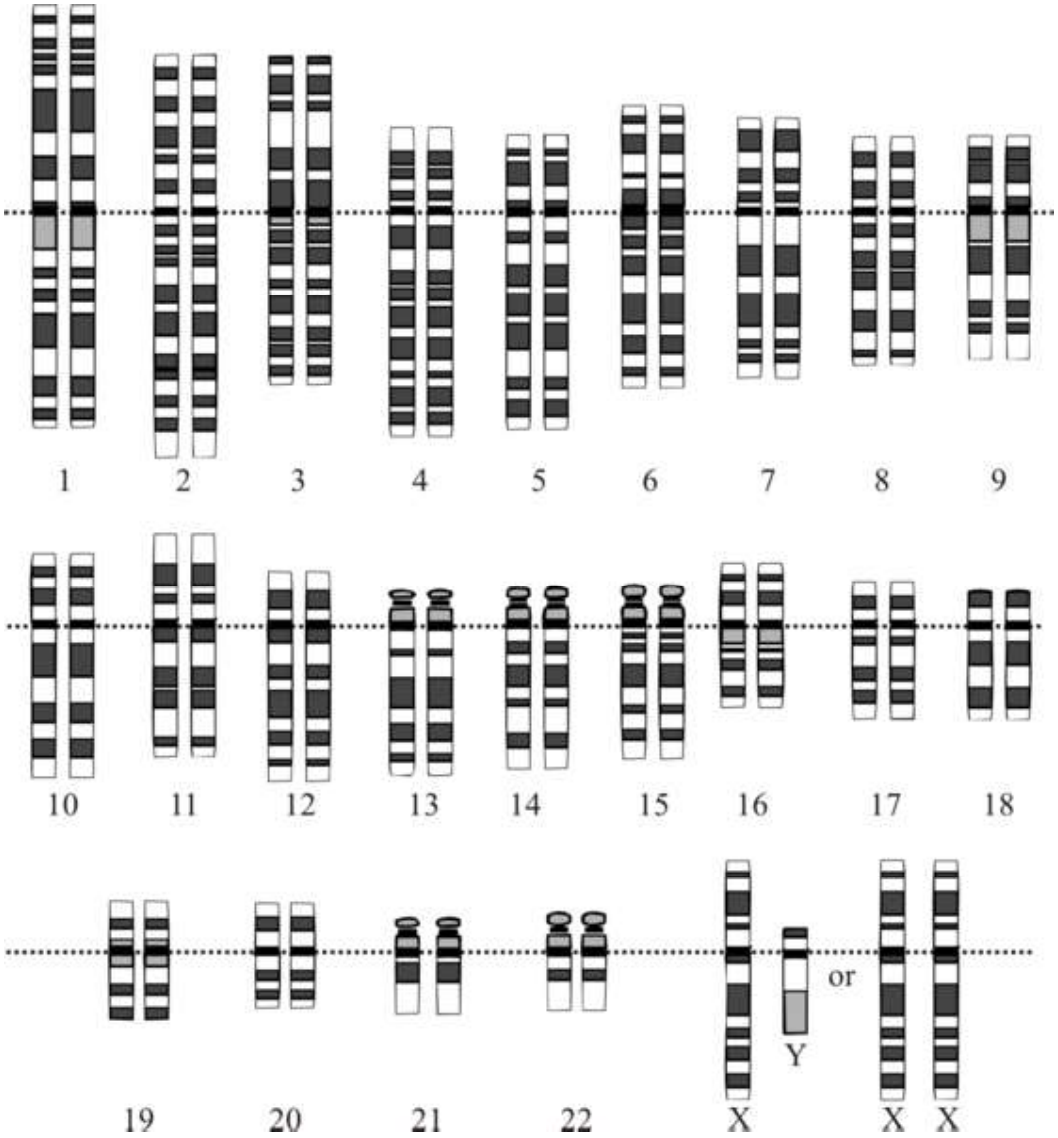
Due to the lack of a system for checking for copying errors, Mitochondrial DNA (mtDNA) has a more rapid rate of variation than nuclear DNA. This 20-fold increase in the mutation rate allows mtDNA to be used for more accurate tracing of maternal ancestry. Studies of mtDNA in populations have allowed ancient 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- 17

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.

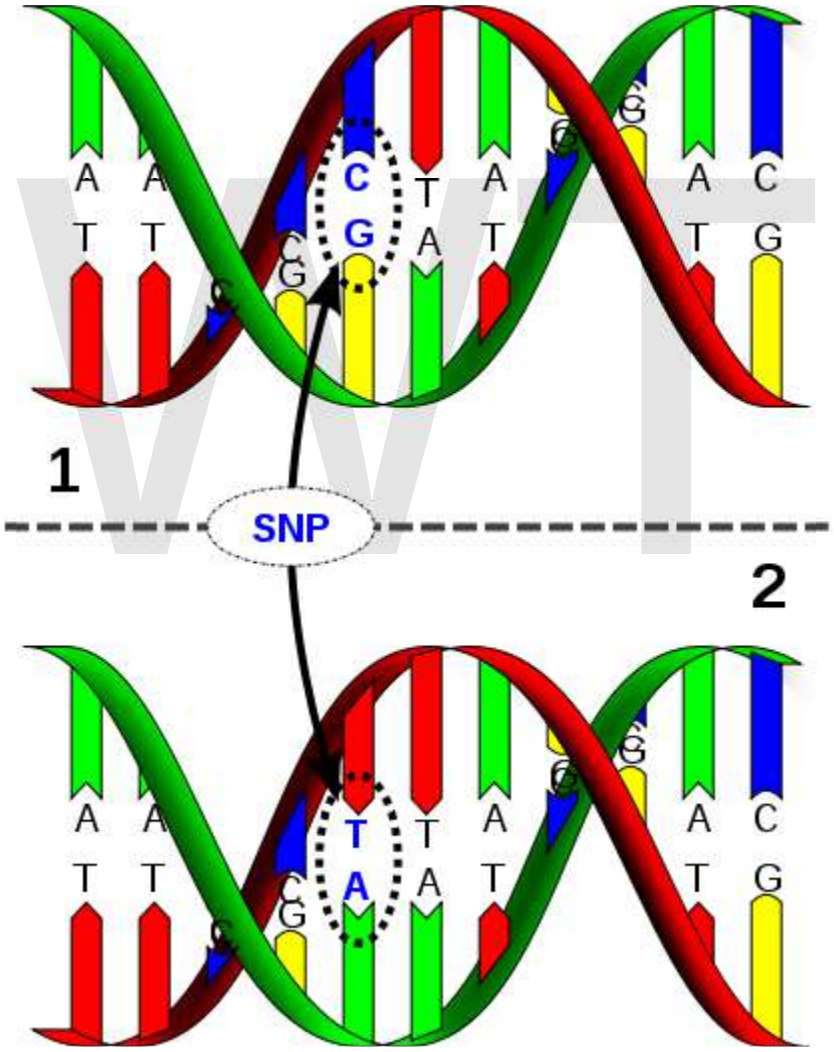
Genetic variation

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.

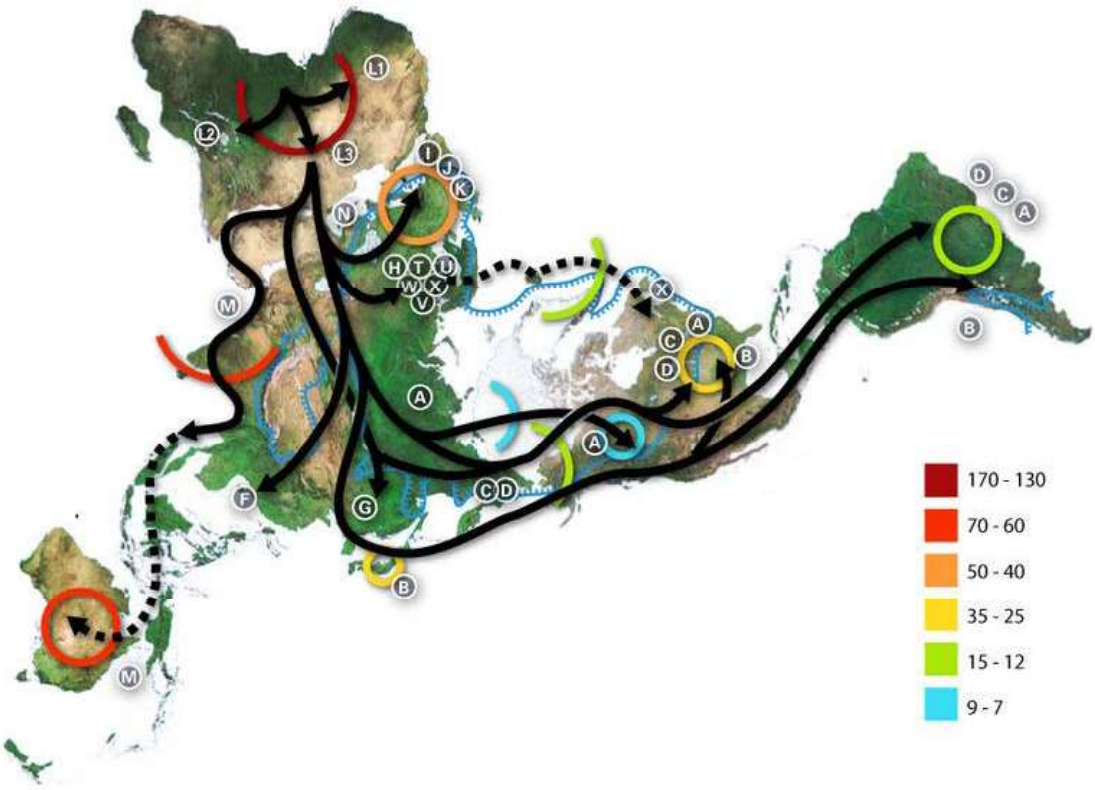
In human genetics, the haplogroups most commonly studied are Y-chromosome (Y-DNA) haplogroups and mitochondrial DNA (mtDNA) haplogroups, both of which can be used to define genetic populations. Y-DNA is passed solely along the patrilineal line, from father to son, while mtDNA is passed down the matrilineal 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

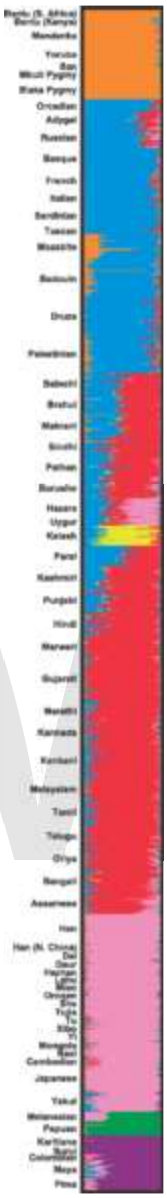


Chart showing human genetic clustering.

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

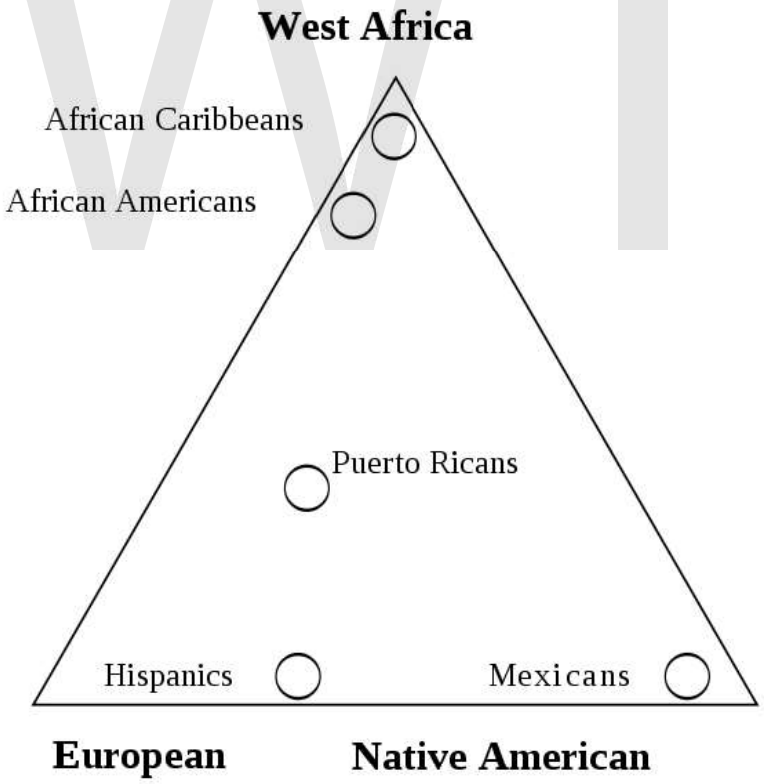
In 2003 A. W. F. Edwards wrote a paper called Lewontin's Fallacy, rebutting the argument that because most genetic variation is within-group, classification of humans is not possible. He claimed that this conclusion ignores the fact that most of the information that distinguishes populations is hidden 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- 18

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 Macháčková 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- 19

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.

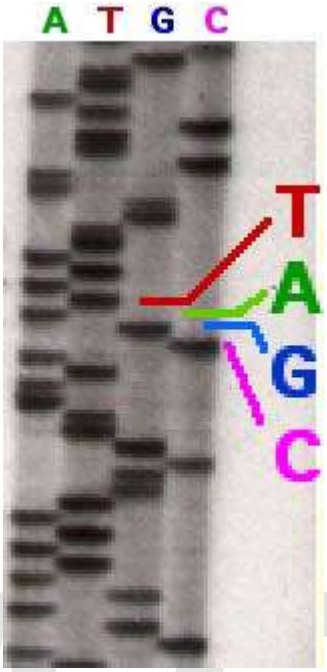
Maxam–Gilbert sequencing

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



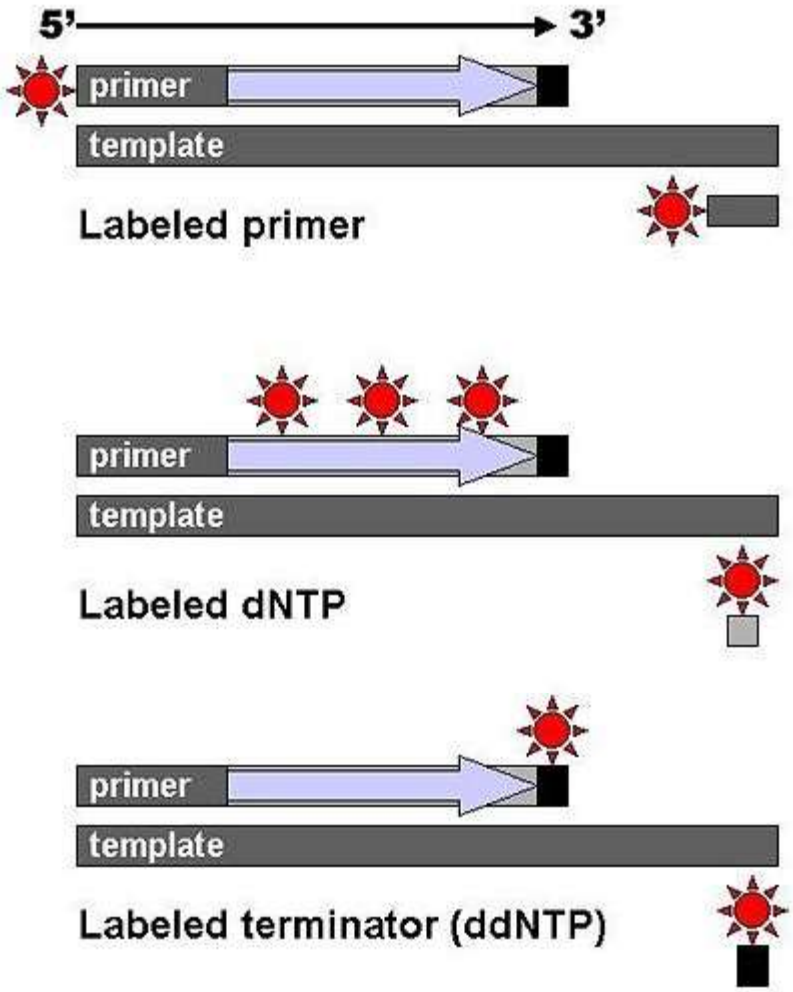
Part of a radioactively labelled sequencing gel

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

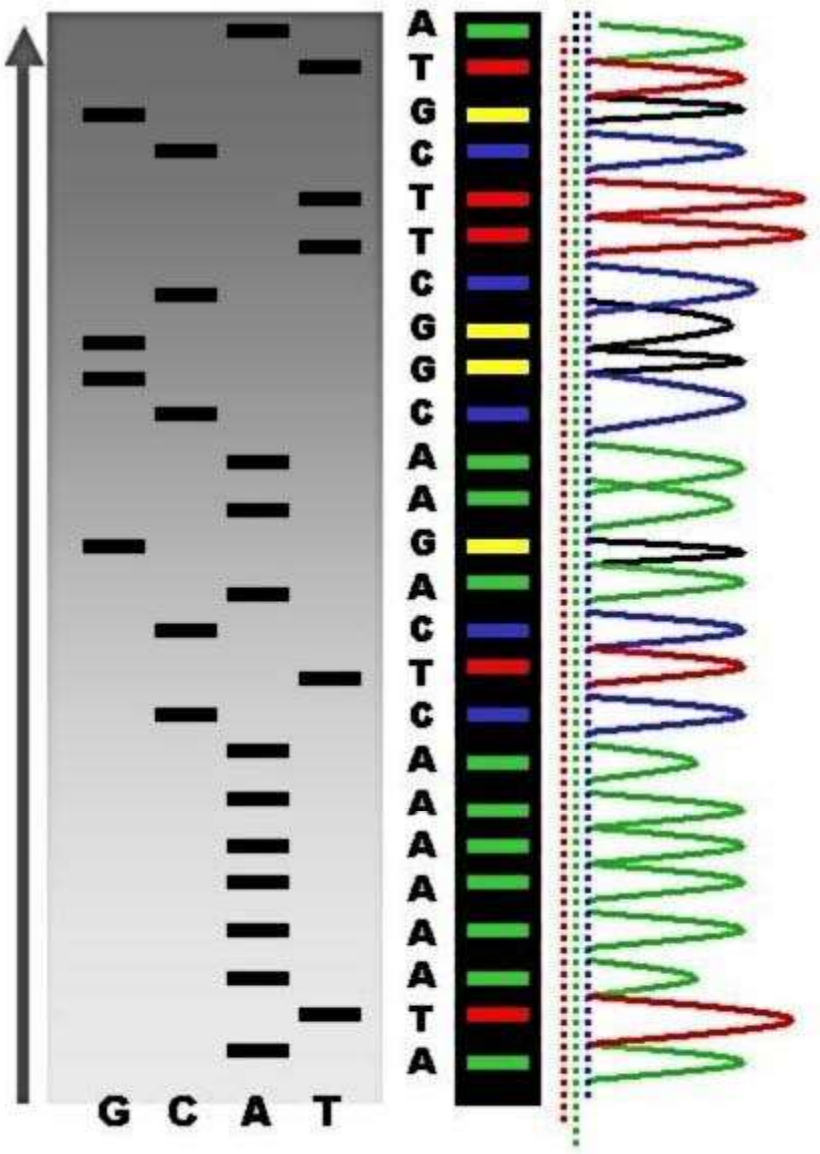
DNA fragments of different lengths. A dark band in a lane indicates a DNA fragment that is the result of chain termination after incorporation of a dideoxynucleotide (ddATP, ddGTP, ddCTP, or ddTTP). The relative positions of the different bands among the four lanes are then used to read (from bottom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

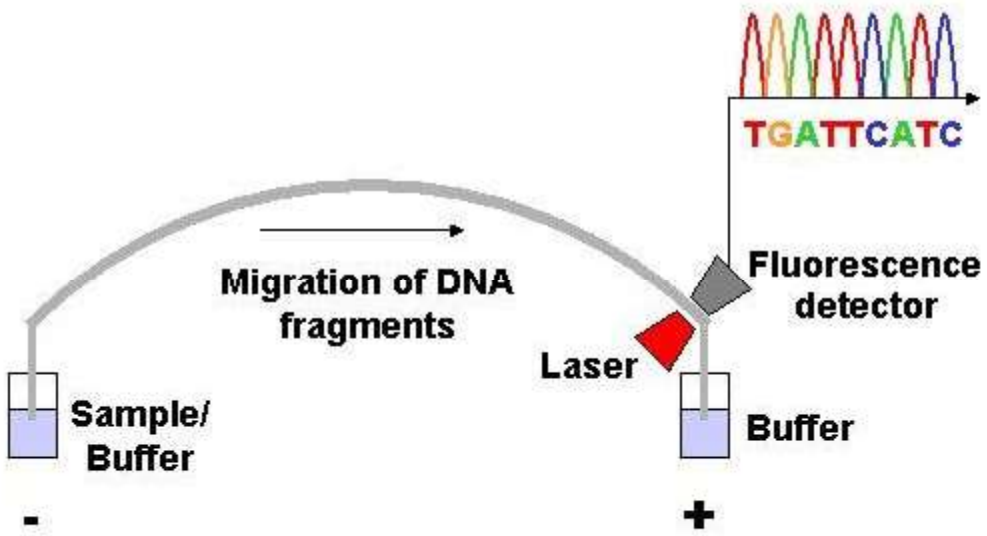
by Leroy Hood and coworkers of fluorescently labeled ddNTPs and primers set the stage for automated, high-throughput DNA sequencing.



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.

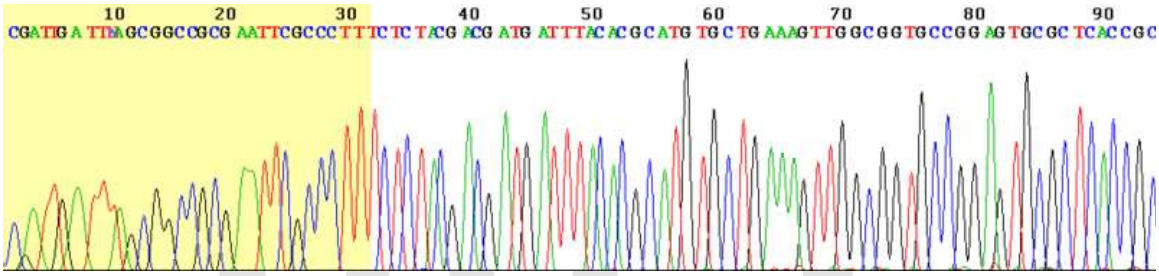
Challenges

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.

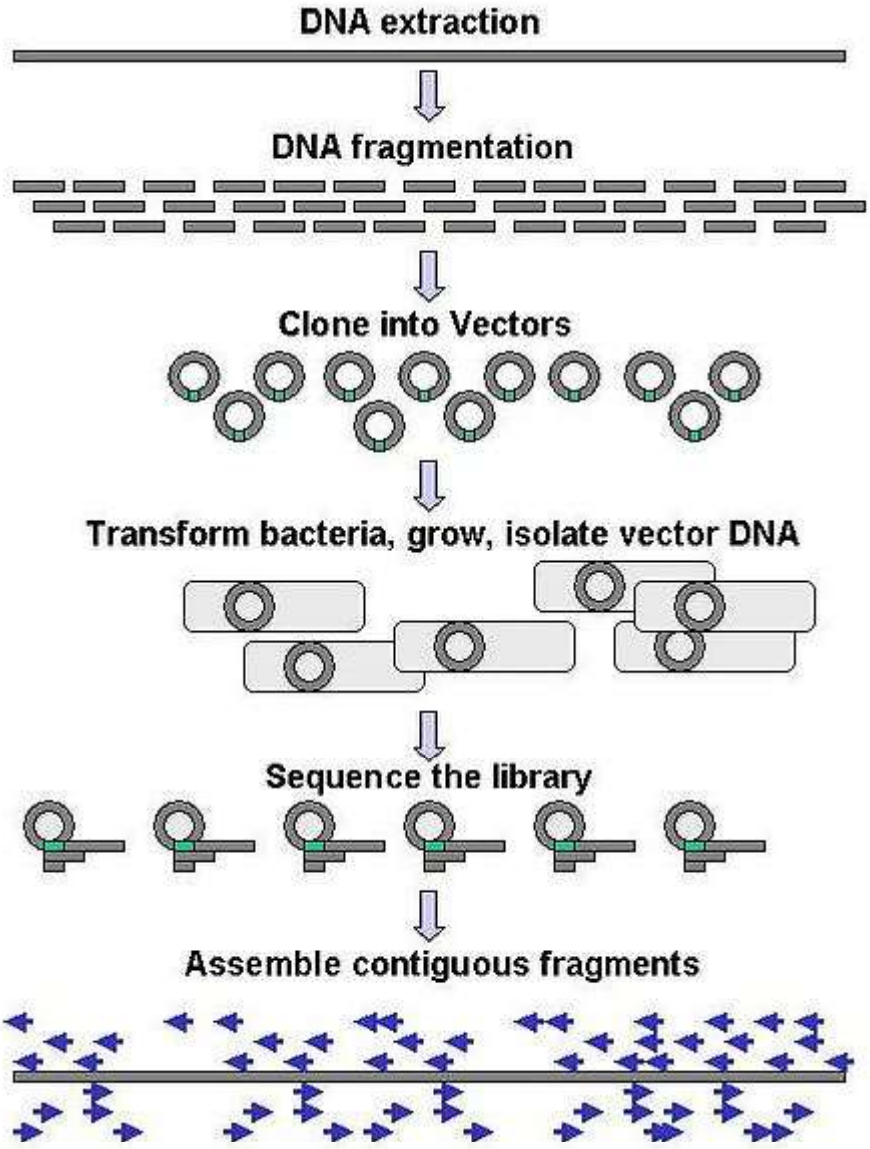
Automation and sample preparation



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.

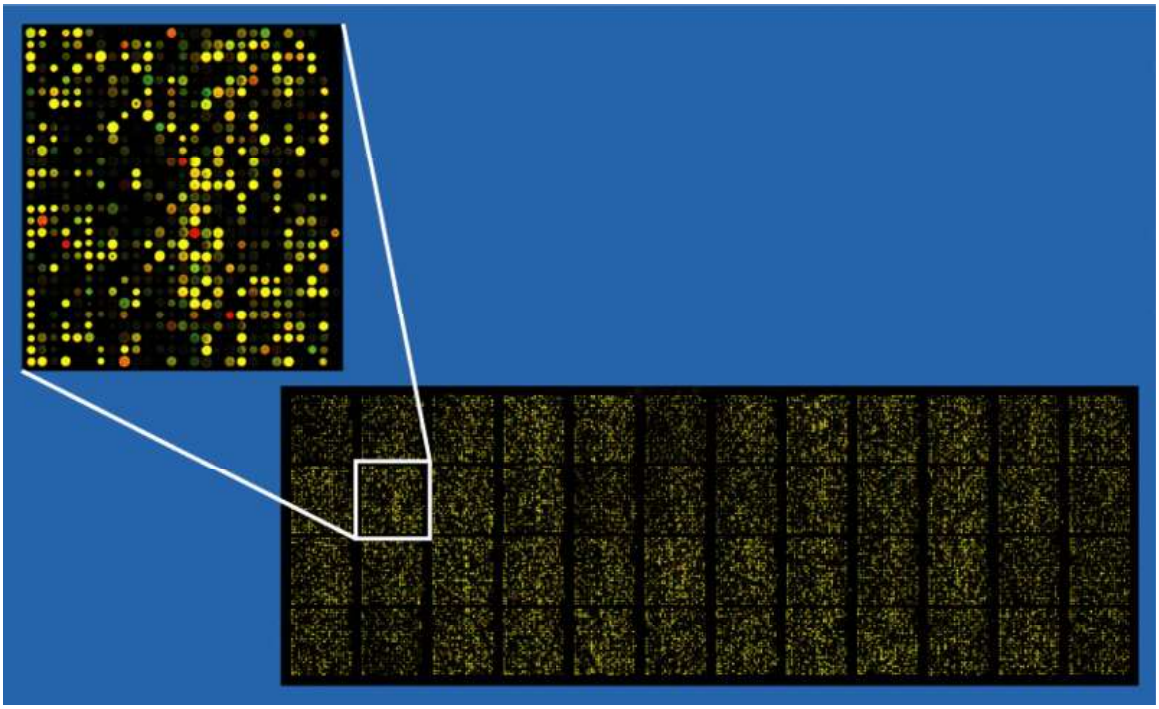
- 2000 Lynx Therapeutics publishes and markets "MPSS" - a parallelized, adapter/ligation-mediated, bead-based sequencing technology, launching "next-generation" sequencing.

- 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- 20

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.

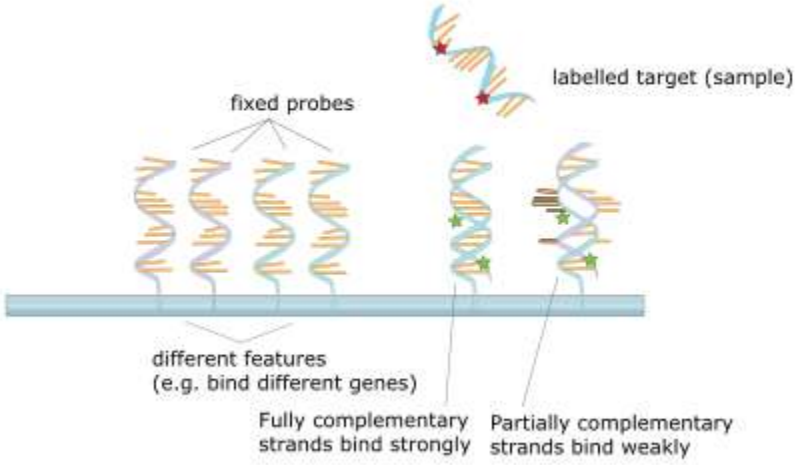
In standard microarrays, the probes are attached via surface engineering to a solid surface by a covalent bond to a chemical matrix (via epoxy-silane, amino-silane, lysine, polyacrylamide or others). The solid surface can be glass or a silicon chip, in which case they are colloquially known as an *Affy 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

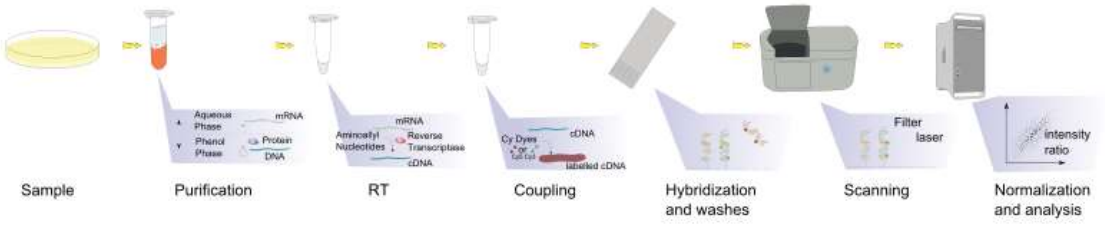
Microarray technology evolved from Southern blotting, where fragmented DNA is attached to a substrate and then probed with a known gene or fragment. Nucleic Acids Res. 1992 Apr 11;20(7):1679-84. Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization 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:

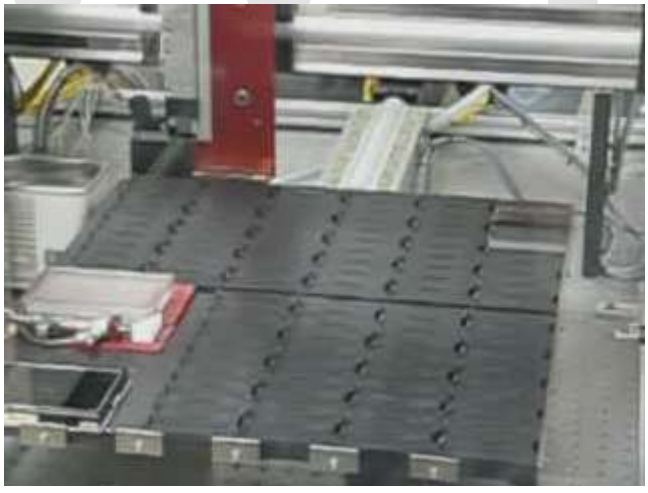
Application or technology	Synopsis
Gene expression profiling	In an mRNA or gene expression profiling experiment the expression levels of thousands of genes are simultaneously monitored to study the effects of certain treatments, diseases, and developmental stages on gene expression. For example, microarray-based 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

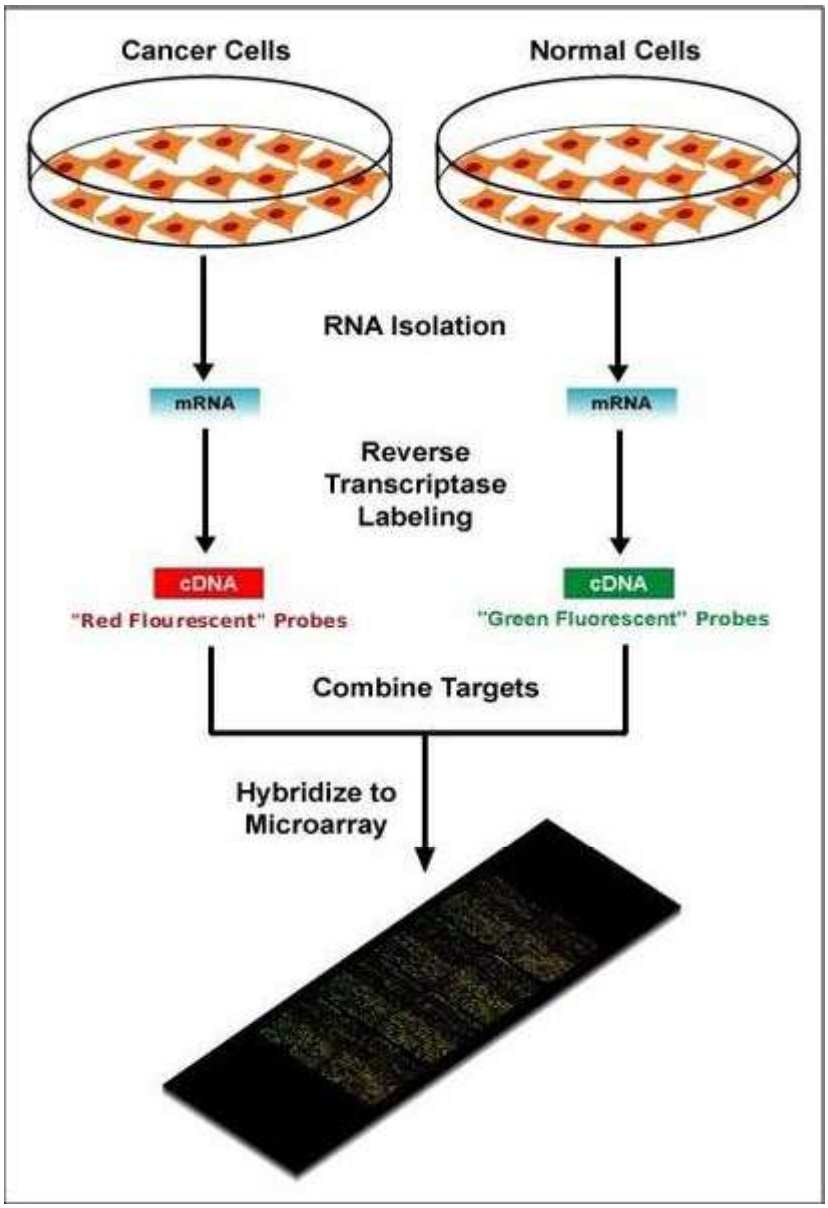


Diagram of typical dual-colour microarray experiment

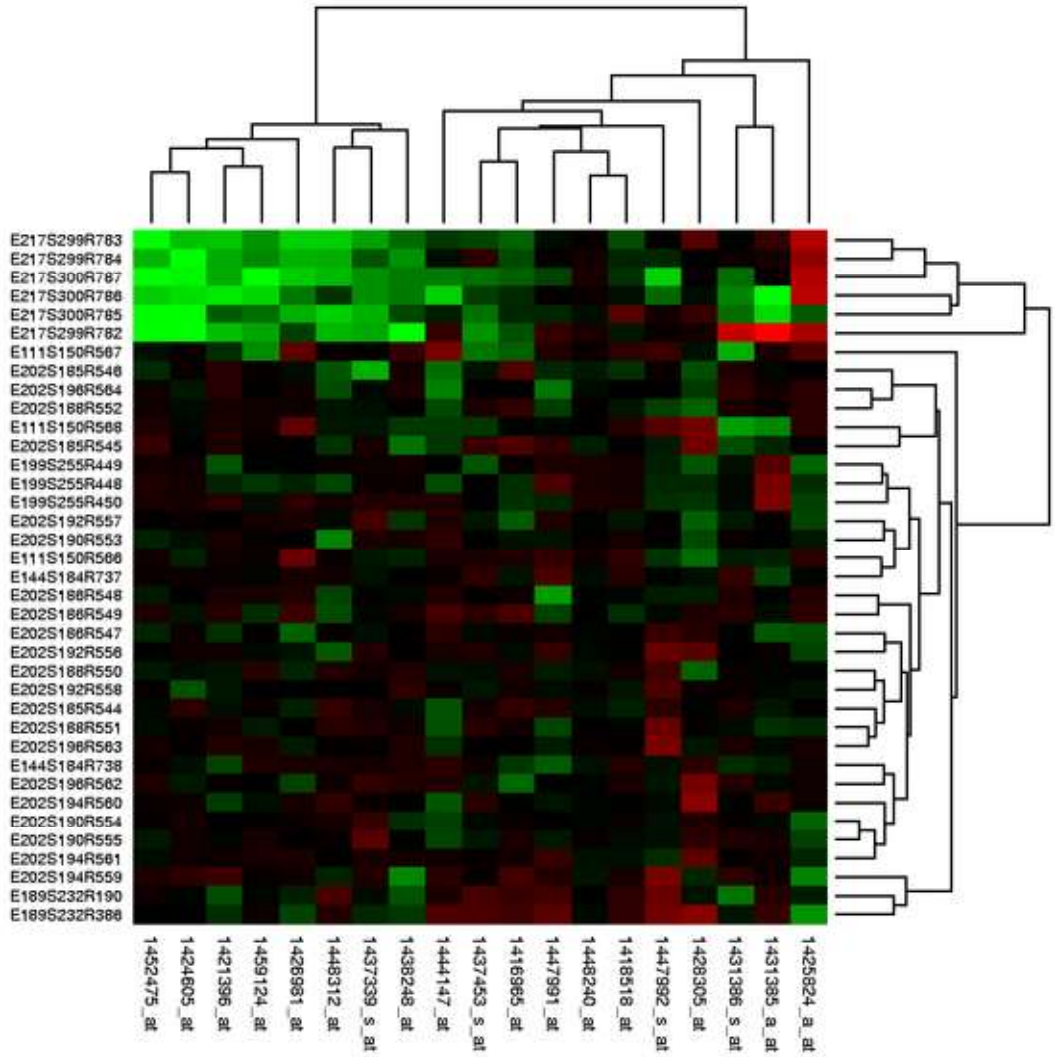
Two-color microarrays or *two-channel microarrays* are typically hybridized with cDNA prepared from two samples to be compared (e.g. diseased tissue versus healthy tissue) and that are labeled with two different fluorophores. Fluorescent dyes commonly used for cDNA labeling include Cy3, which has a fluorescence 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- 21

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.