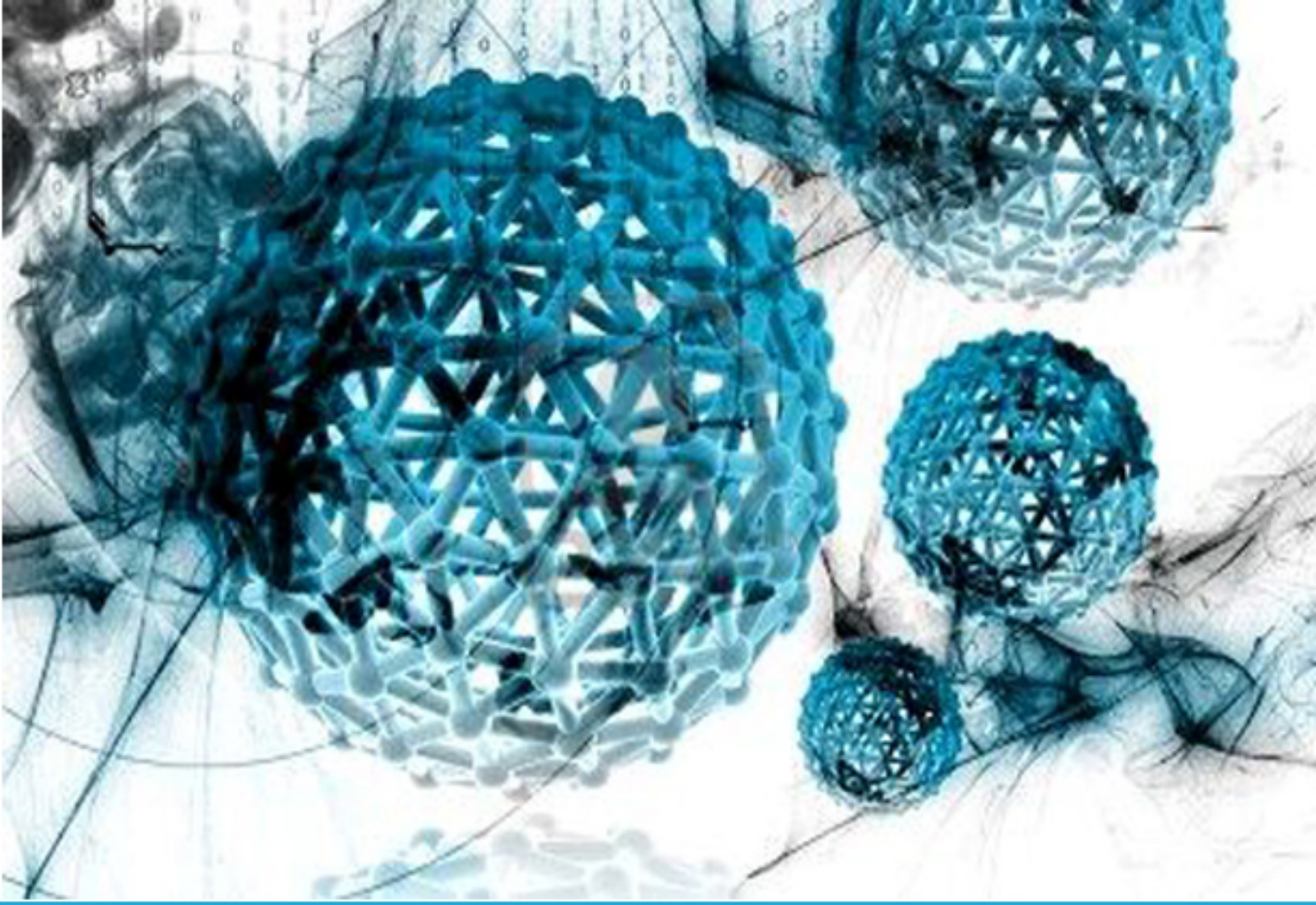


Gene Biology

Lovetta Arndt

First Edition, 2012

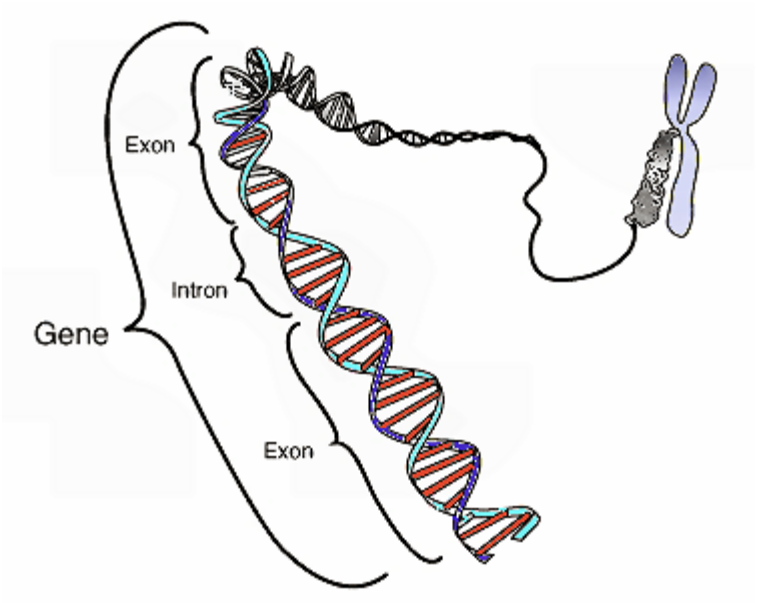
ISBN 978-81-323-4142-0

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

Table of Contents

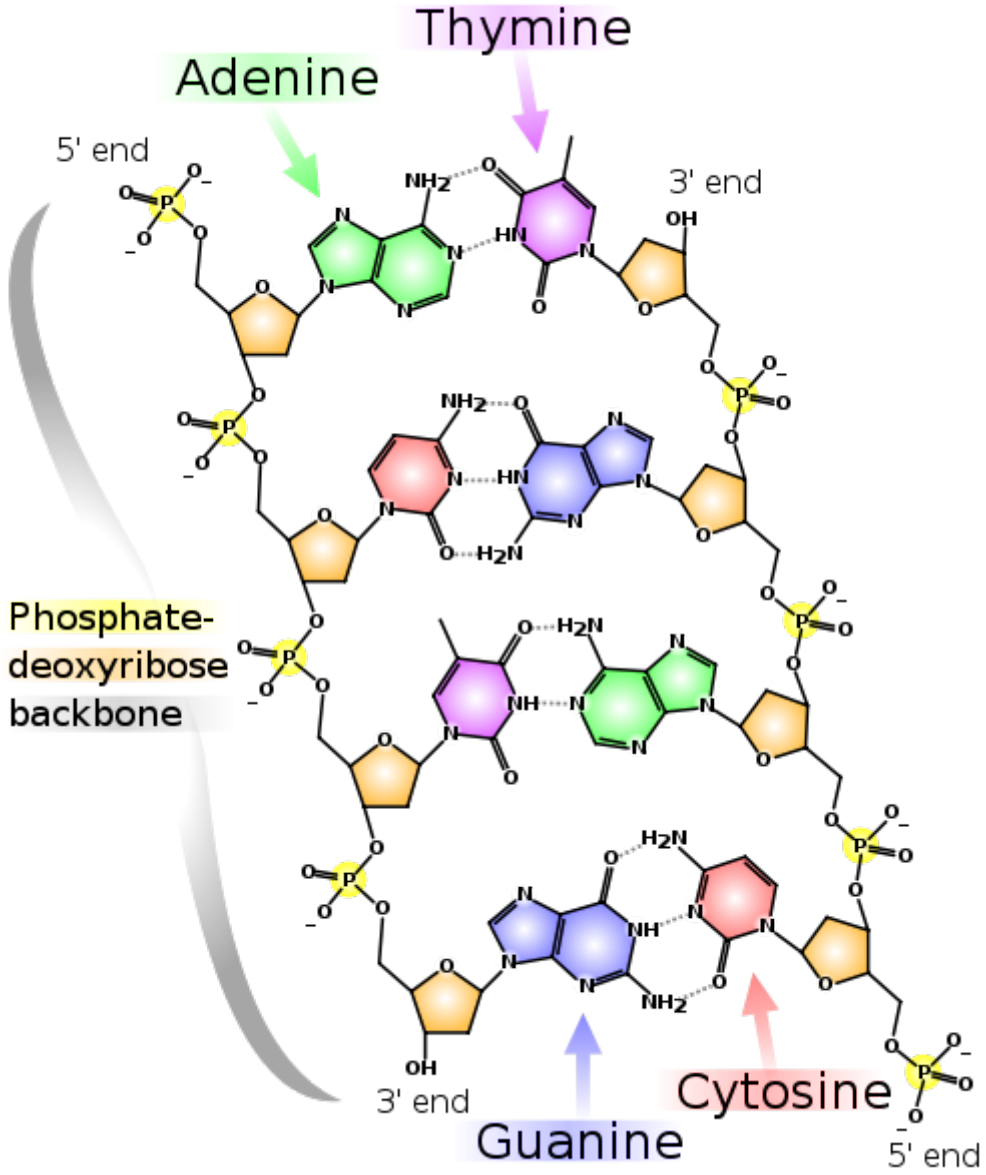
Chapter 1

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

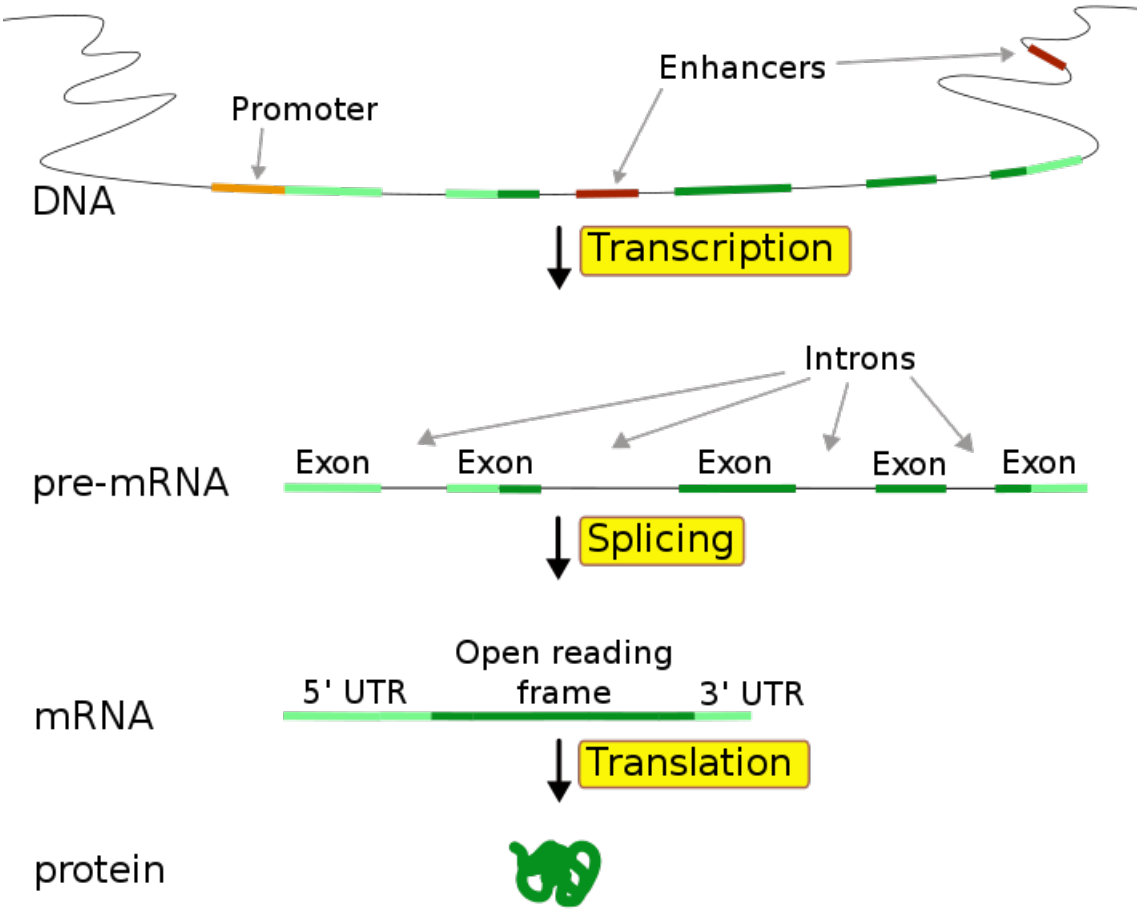


Diagram of the "typical" eukaryotic protein-coding **gene**. Promoters and enhancers determine what portions of the DNA will be transcribed into the precursor mRNA (pre-mRNA). The pre-mRNA is then spliced into messenger RNA (mRNA) which is later translated into protein.

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

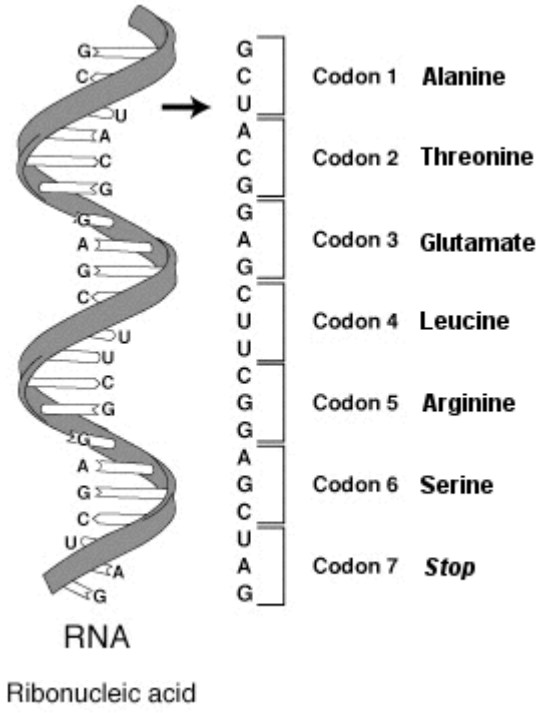
DNA helix on which thousands of genes are encoded. Prokaryotes—bacteria and archaea—typically store their genomes on a single large, circular chromosome, sometimes supplemented by additional small circles of DNA called plasmids, which usually encode only a few genes and are easily transferable between 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.

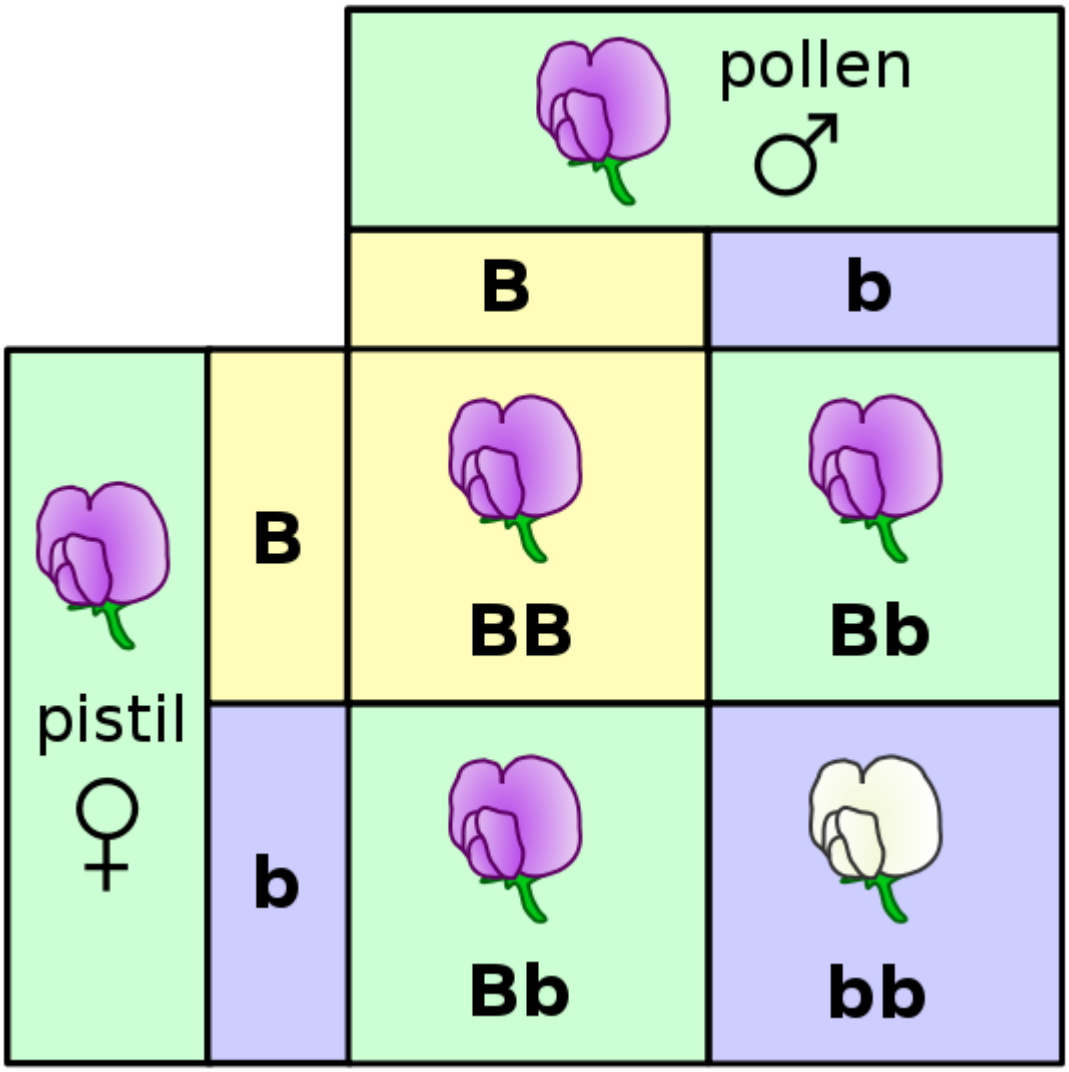
A series of subsequent discoveries led to the realization decades later that chromosomes within cells are the carriers of genetic material, and that they are made of DNA (deoxyribonucleic acid), a polymeric molecule found in all cells on which the 'discrete units' of Mendelian inheritance are encoded. 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

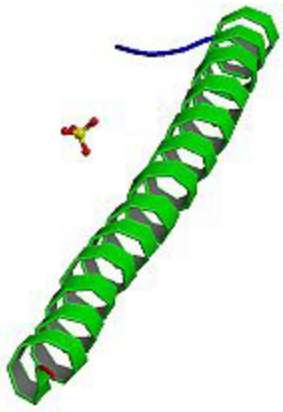
encoding a coherent set of potentially overlapping functional products." This new definition categorizes genes by functional products, whether they be proteins or RNA, rather than specific DNA loci; all regulatory elements of DNA are therefore classified as *gene-associated* regions.

Chapter 2

Adenomatous Polyposis Coli and APC/C Activator Protein CDH1

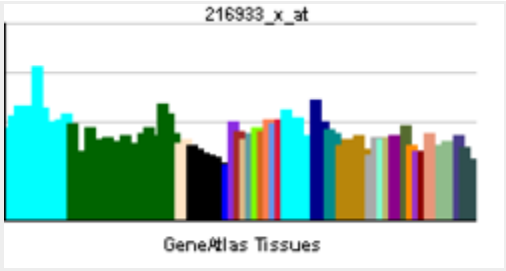
Adenomatous polyposis coli

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.

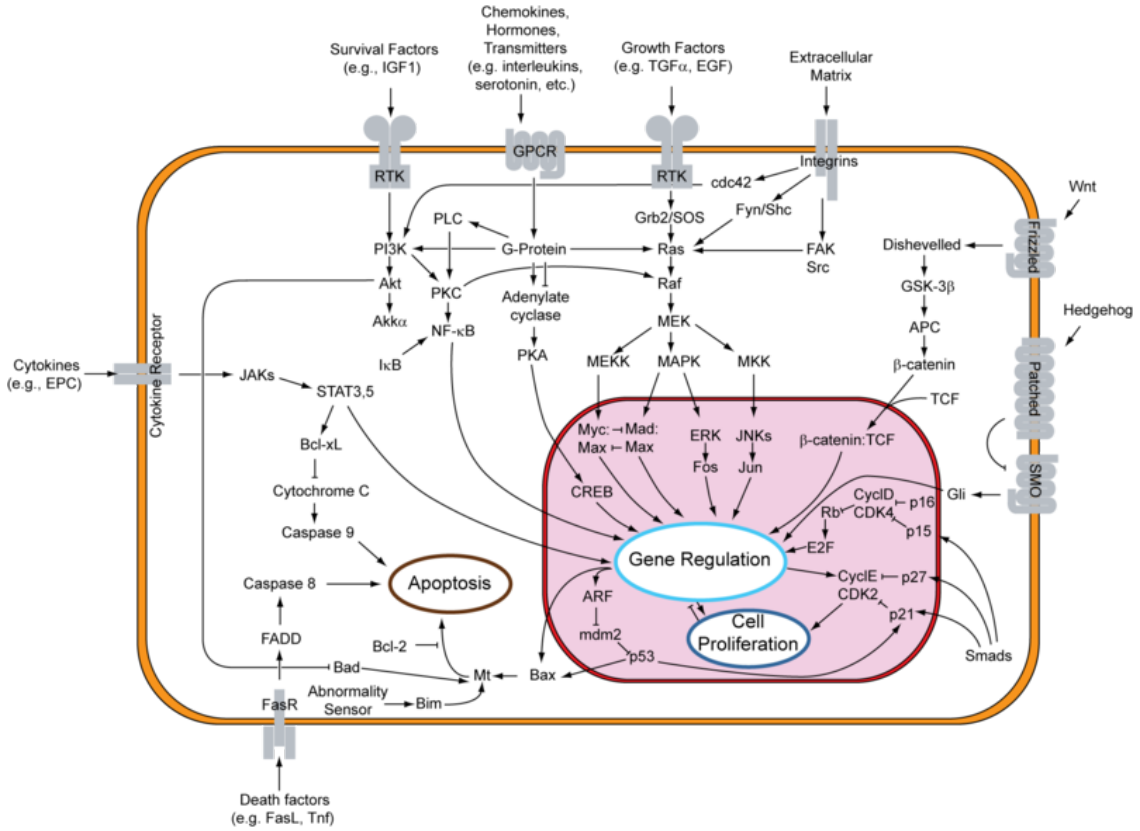
Neurological role

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

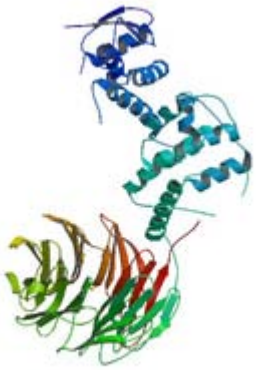
- ARHGEF4,
- AXIN1,
- BUB1,
- CTNNB1,
- CSNK2B,
- CSNK2A1,
- CTNNA1,
- DLG3,
- KIFAP3,
- MAPRE2,
- JUP,
- SIAH1,
- TFAP2A,
- TUBA4A, and
- XPO1.



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

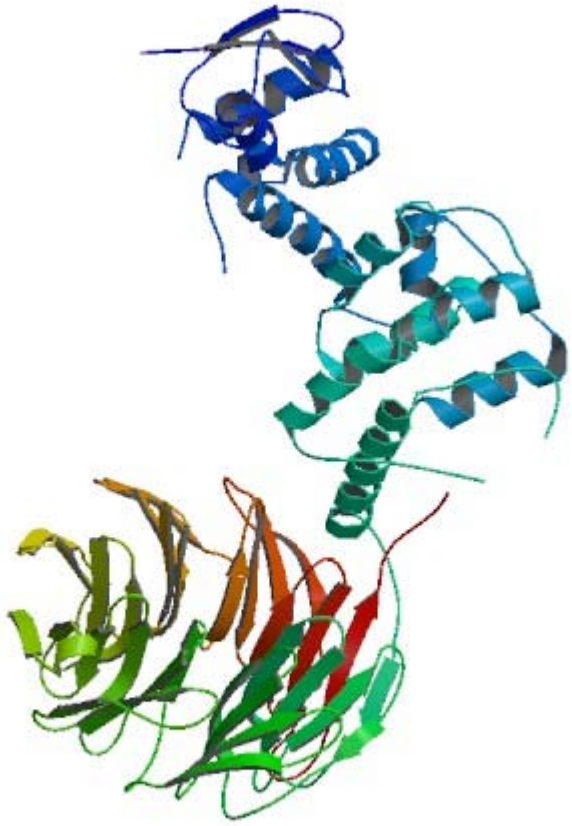


Fig. 1 There is no structure resolved for Cdh1 of *Saccharomyces cerevisiae*. There is a model based on template pdb2ovq, which shows the SCF(Fbw7) ubiquitin ligase complex. Fbw7 is also a WD repeat protein like Cdh1.

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

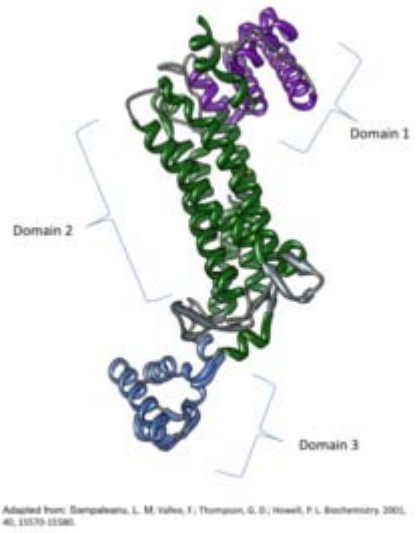
Argininosuccinate lyase

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.

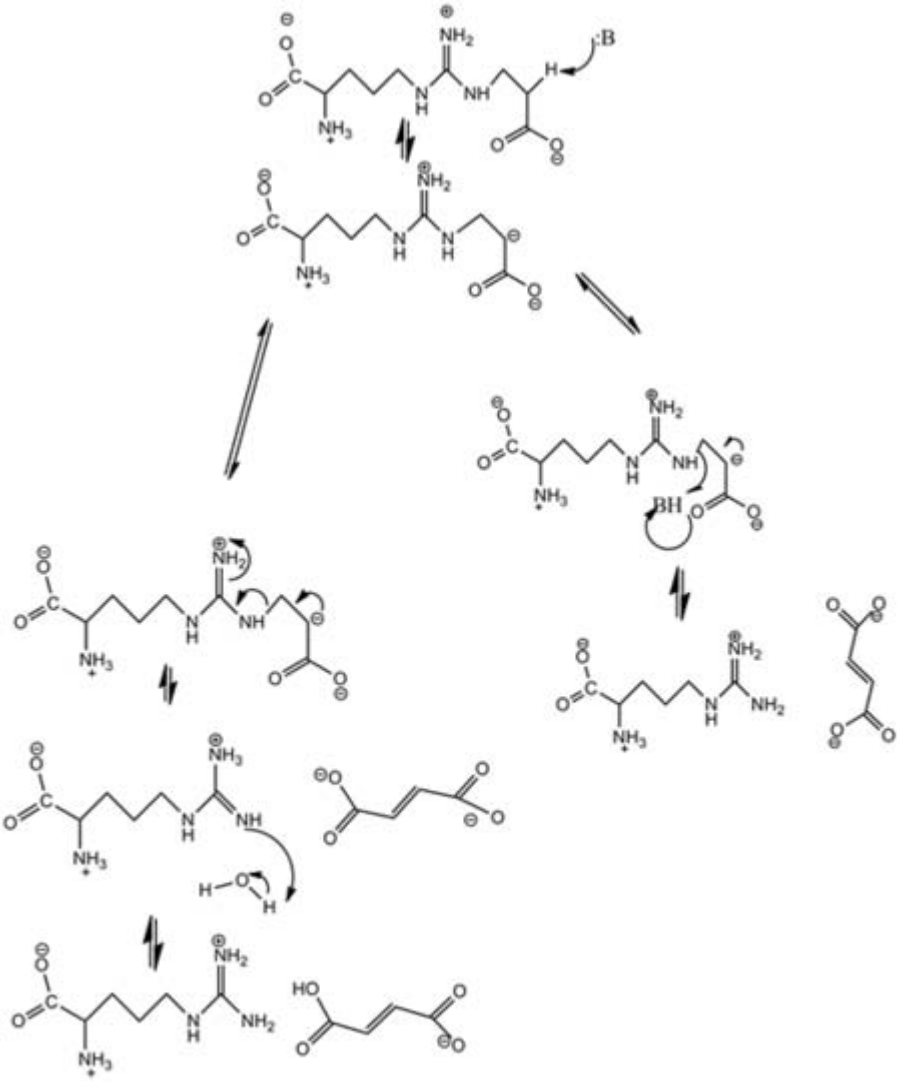
ASL is related to intragenic complementation.

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.

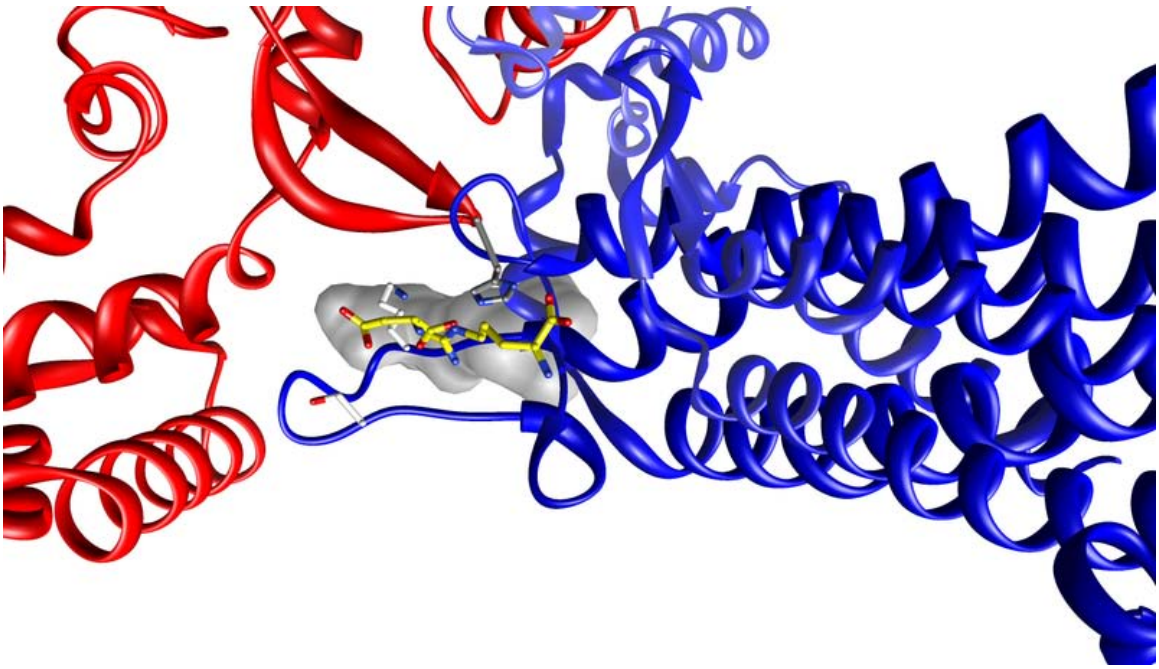
Each monomer in the ASL homotetramer is composed of three structural domains; all three are primarily alpha helical. Domains 1 and 3 are similar in structure as they both consist of helix-turn-helix motifs. Domain 1 of the monomer contains the carboxyl terminus. Domain 2 contains one small beta sheet, 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.

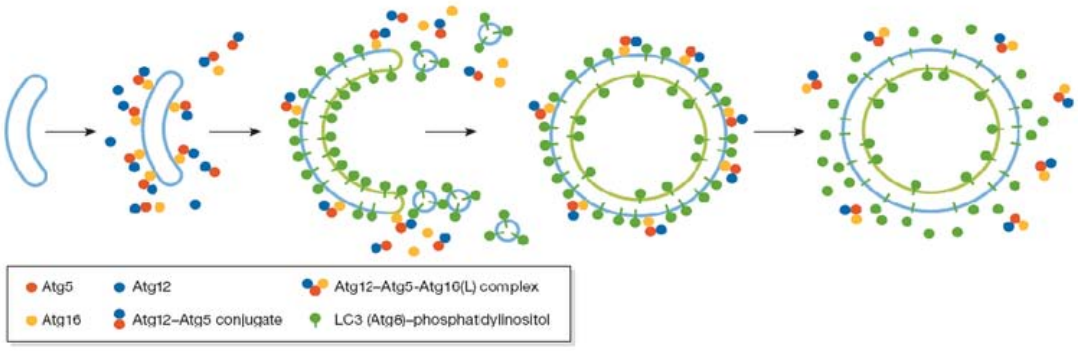


Fig. 1 Atg8 and Atg12 conjugation systems in autophagosome formation

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

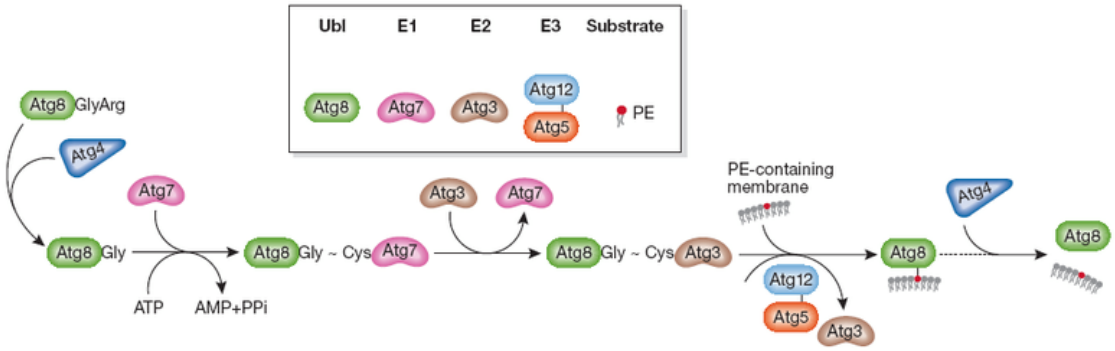


Fig. 2 Post-translational modification and regulatory cycle in yeast

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.

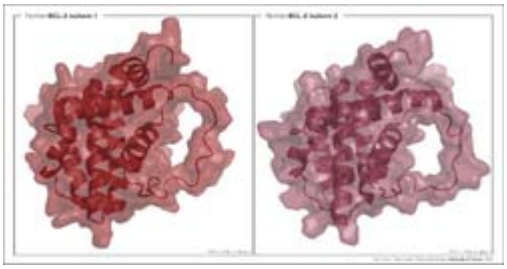
One of them is LC3 (MAP1LC3A), a light chain of the microtubule-associated protein 1 Like Atg8, LC3 needs to be proteolytically cleaved and lipidated to be turned into its active form which can localize to the autophagosomal membrane. Similar to the situation in yeast, the activation process of LC3 is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



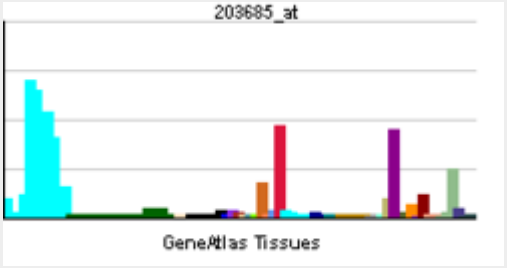
PDB rendering based on 1GJH,1G5M.

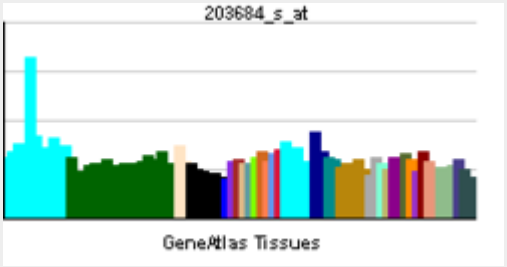
Available structures

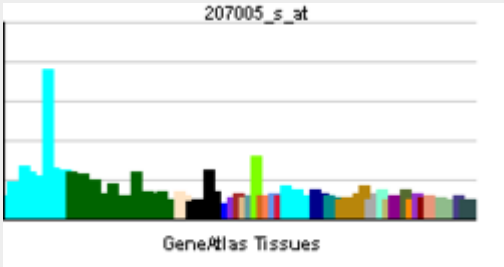
Identifiers	
Symbols	BCL2; Bcl-2
External IDs	OMIM: 151430 MGI: 88138 HomoloGene: 527 GeneCards: BCL2 Gene

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.

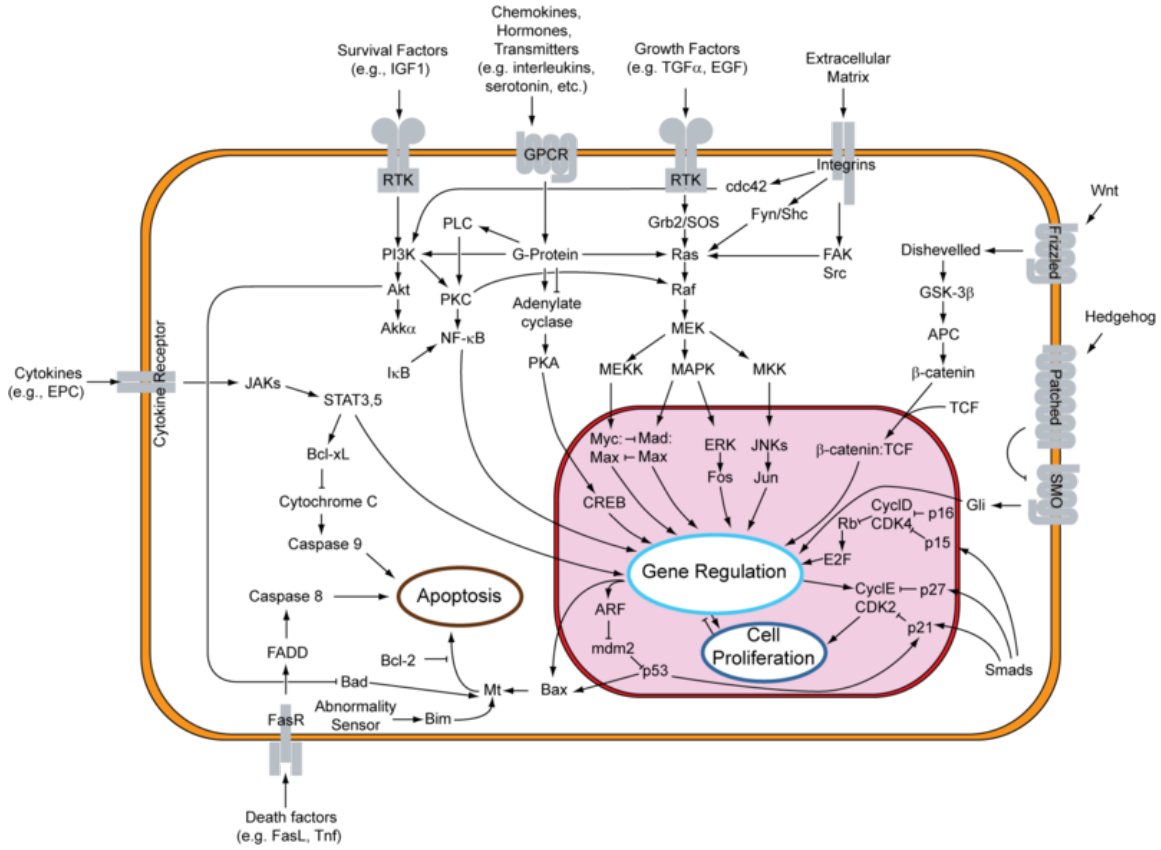
ABT-737

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.

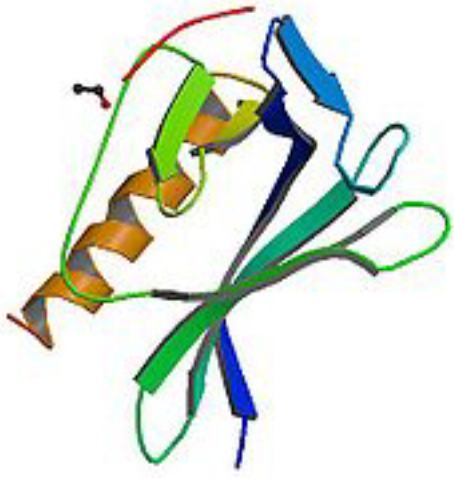
Human BCL-2 genes

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

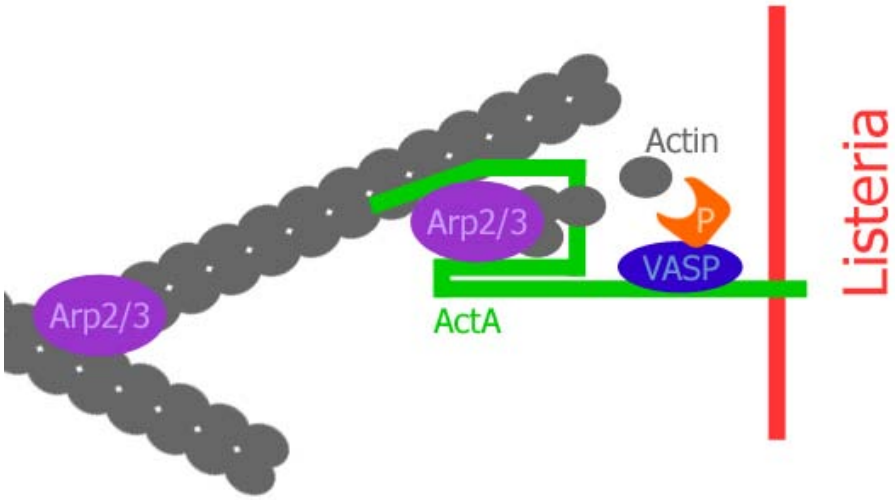


Fig. 1 Actin assembly induced by bacterial protein ActA (shown in green). Mammalian proteins involved in this process are: Profilin (P), Vasodilator-stimulated phosphoprotein (VASP) and actin-related-protein 2 and 3 complex (Arp2/3 complex) as well as actin.

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

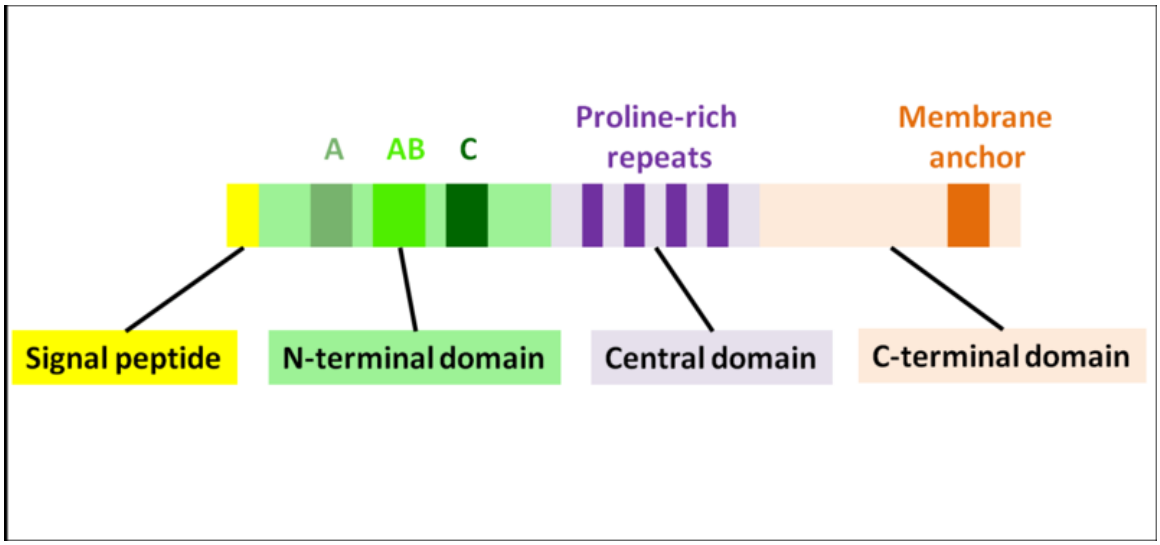


Fig. 2 The ActA protein and its functional domains

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

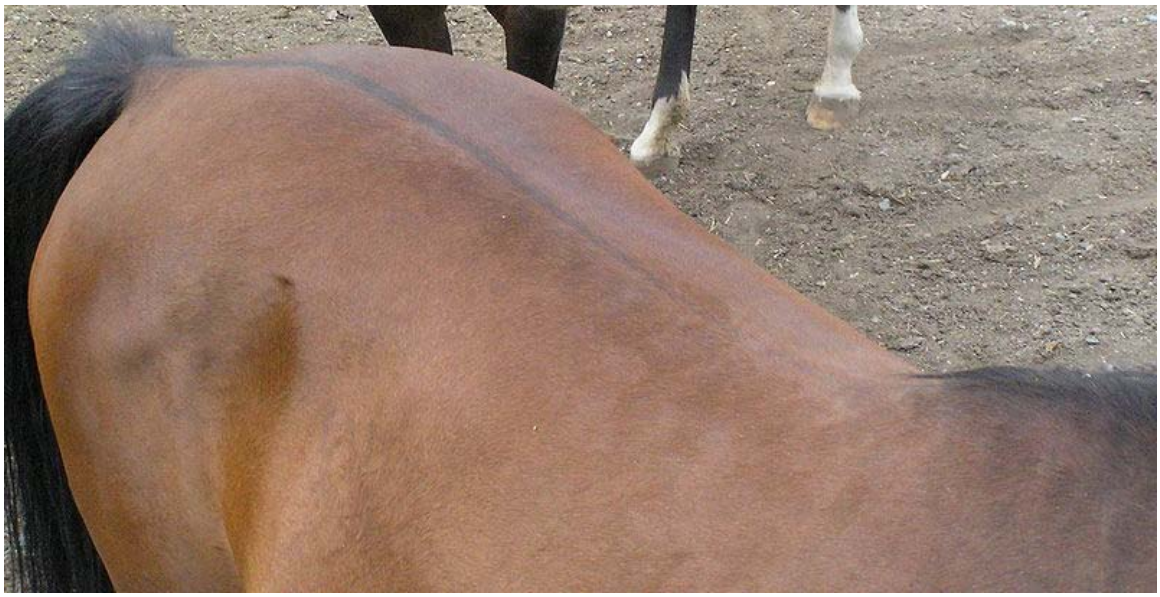
Blue dun, or Grullo.

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.

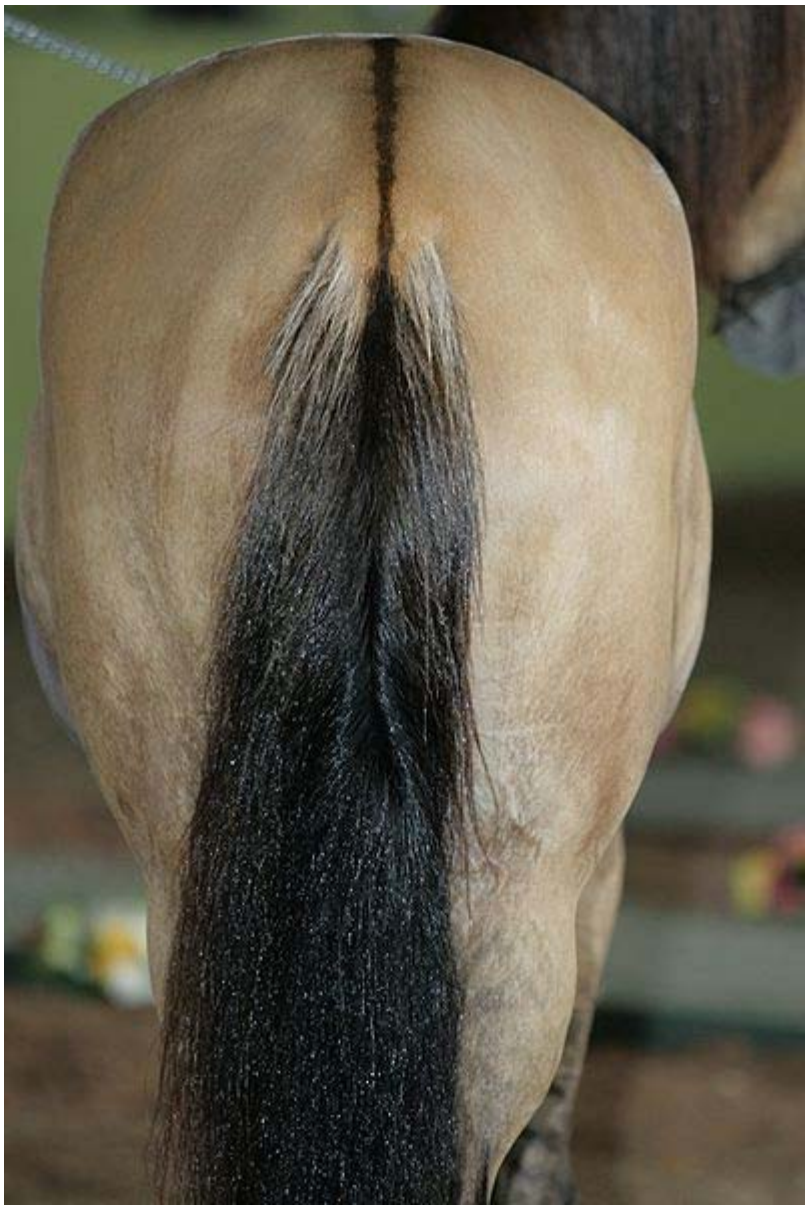
A red dun may also be confused with a perlino, which is genetically a bay horse with two copies of the cream gene, which creates a horse with a cream-colored body but a reddish mane and tail. However, perlinos usually are significantly lighter than a red dun and generally have blue eyes.

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

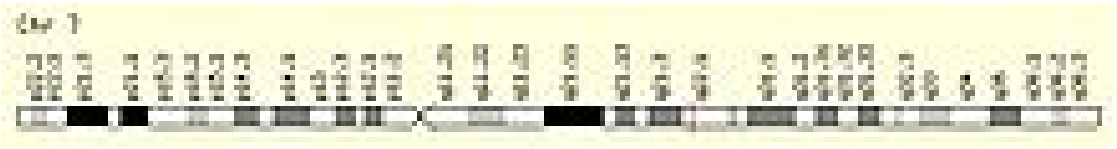
CBLAST Structurally Related Protein. Hermes DNA Transposase. EValue: 1E-6

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.

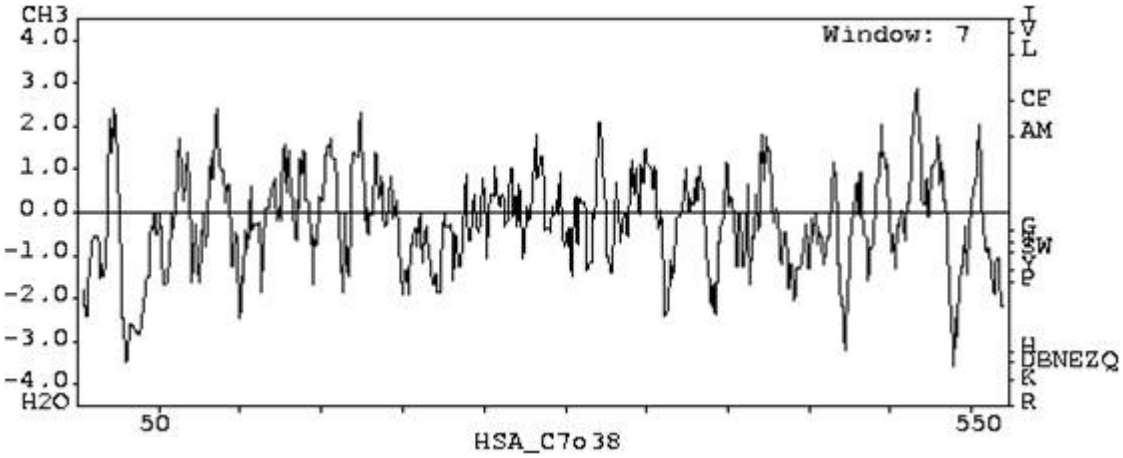
Gene

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.



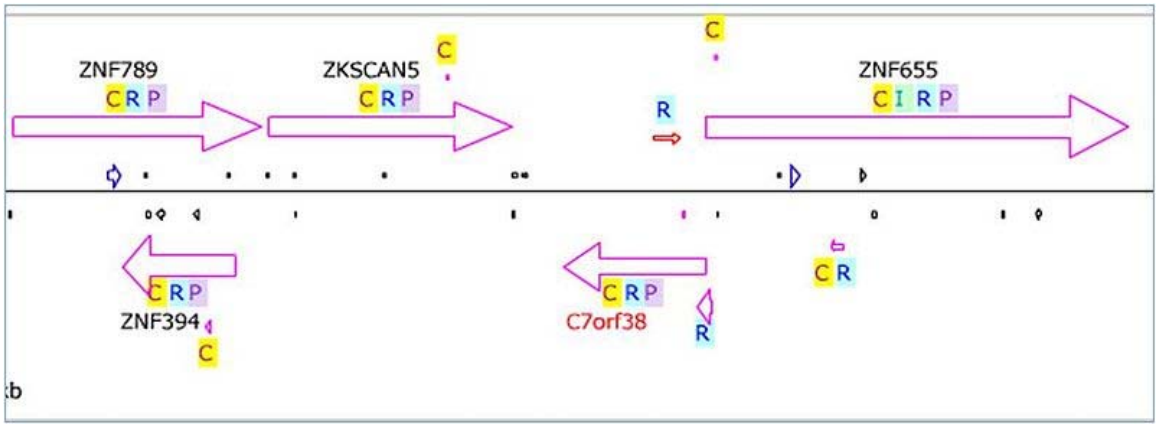
Protein composition

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

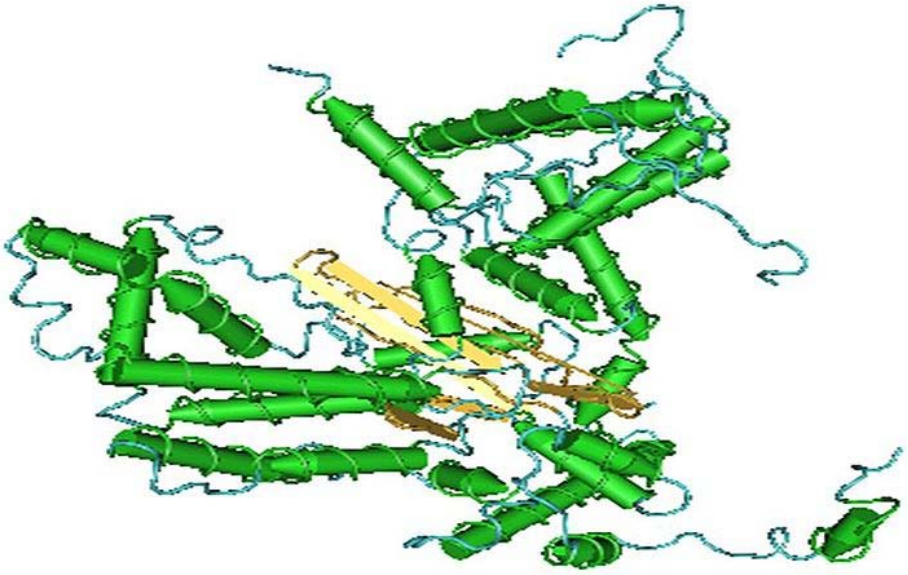
CBLast was used to determine a structurally related protein with experimentally determined structure. The protein Hermes DNA transposase, of the Hermes DBD superfamily, was shown to be structurally similar (Evalue: 1E-6).

hAT Dimerization Domain

Identifiers	
Symbol	hAT
Pfam	PF05699

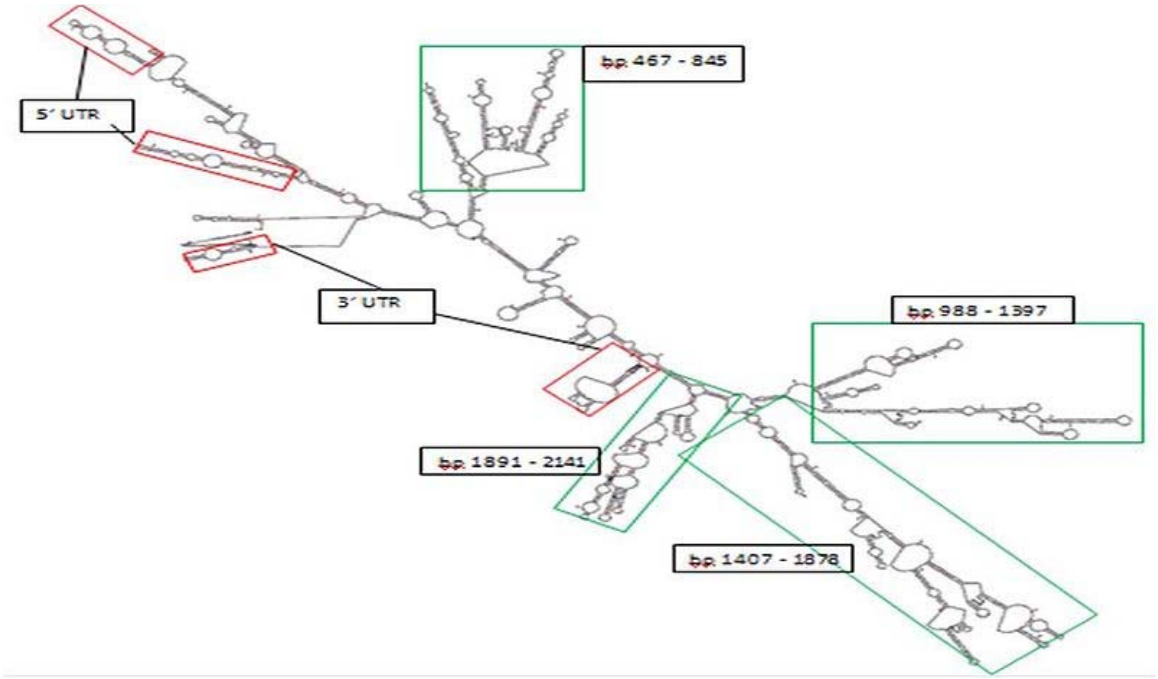
InterPro	IPR008906

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				

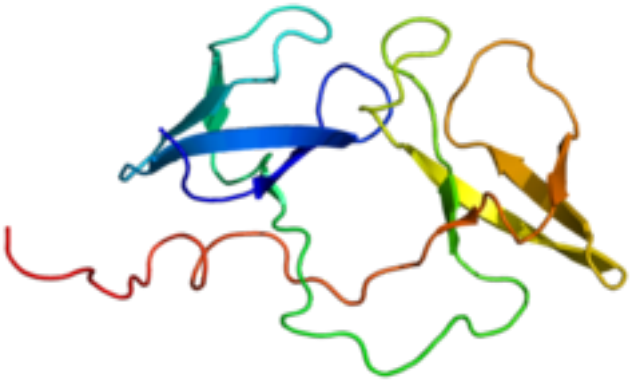
adrenal tumor	0	0/12864
bladder carcinoma	0	0/17758
breast (mammary gland) tumor	10	1/94657
cervical tumor	0	0/34586
chondrosarcoma	0	0/82864
colorectal tumor	0	0/115097
esophageal tumor	0	0/17292
gastrointestinal tumor	0	0/119842
germ cell tumor	0	0/264767
glioma	0	0/107554
head and neck tumor	7	1/137381
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

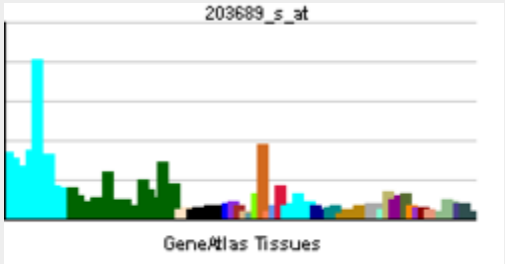
FMR1

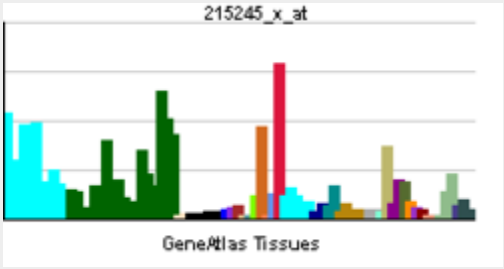
Fragile X mental retardation 1



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	

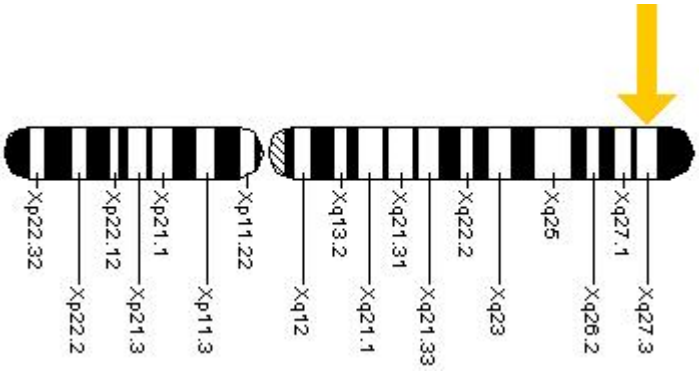




More reference expression data

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

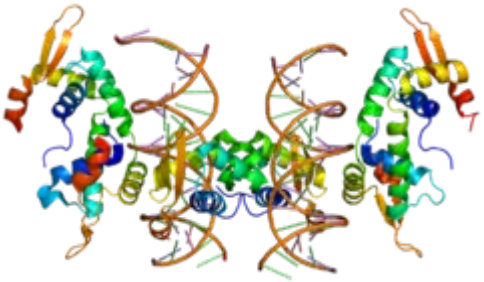
FOXP1 ("forkhead box P1") is a gene that is necessary for the proper development of the brain and lung in mammals. It is a member of the large FOX family of transcription factors.

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.

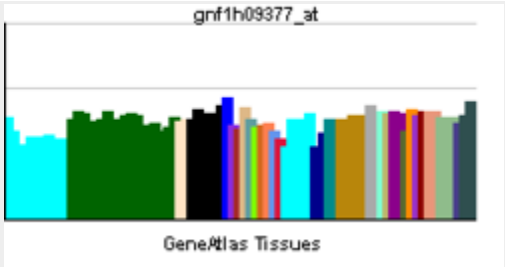
FOXP2

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.

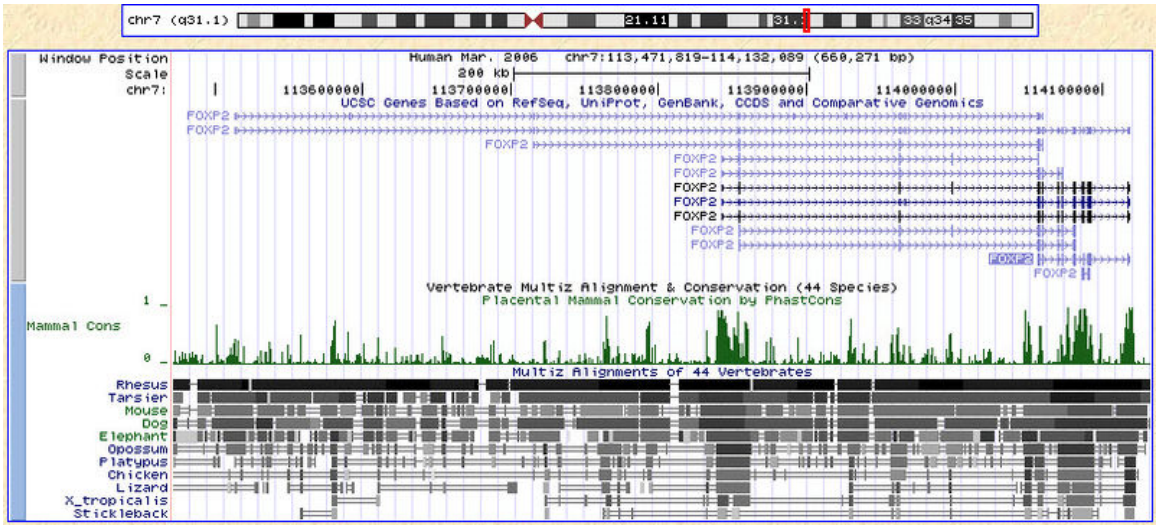
Clinical significance

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.

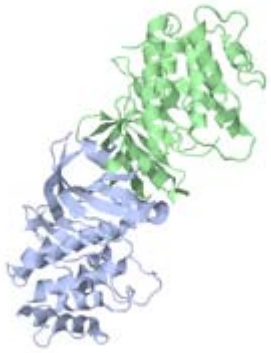
Interactions

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

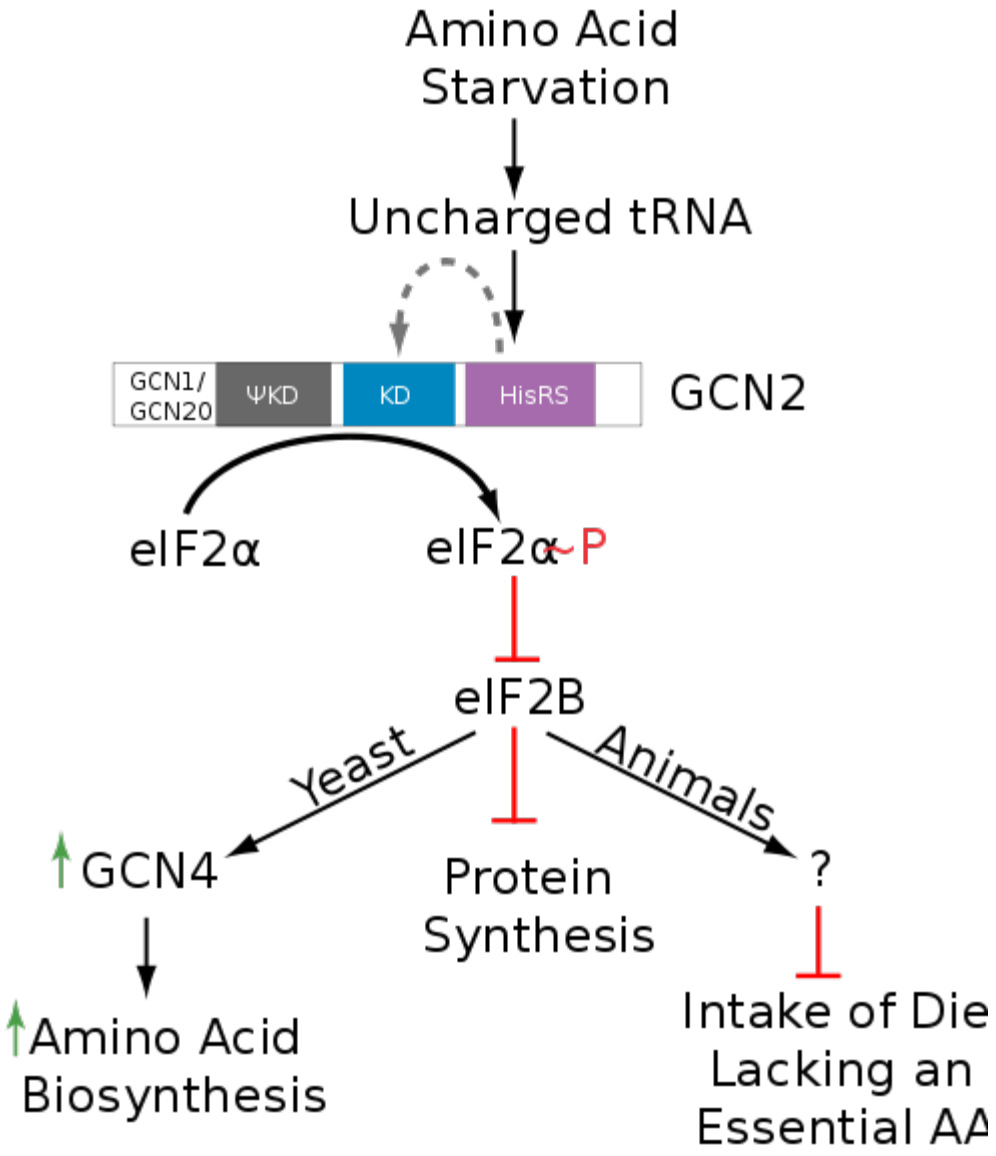


Figure 1:Overview over the functions of GCN2. (GCN1/GCN20=GCN1p/GCN20p binding site; PsiKD = unknown function; KD = Kinase Domain; HisRS = histidyl-tRNA synthetase) Adapted from

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

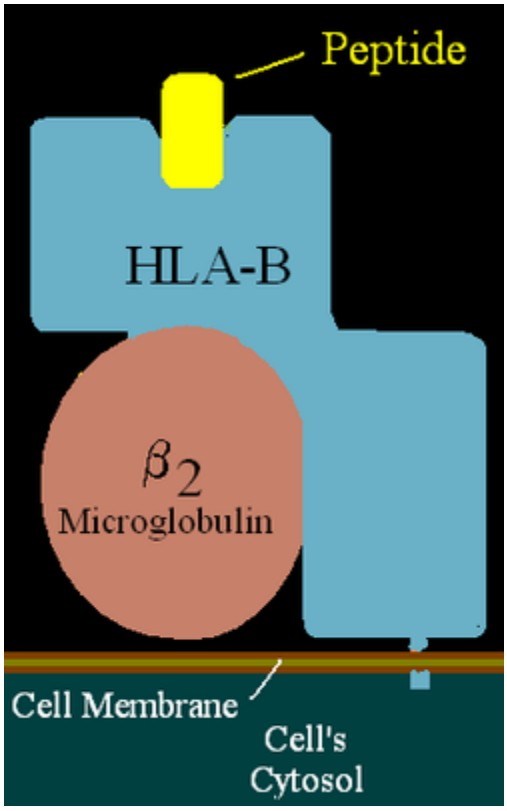
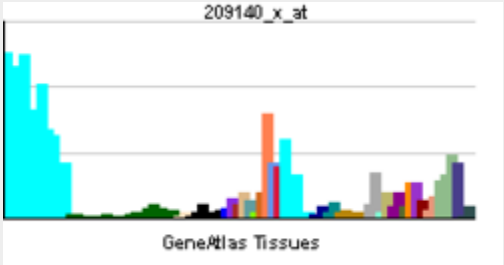


Illustration of HLA-B complexed peptide.

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.

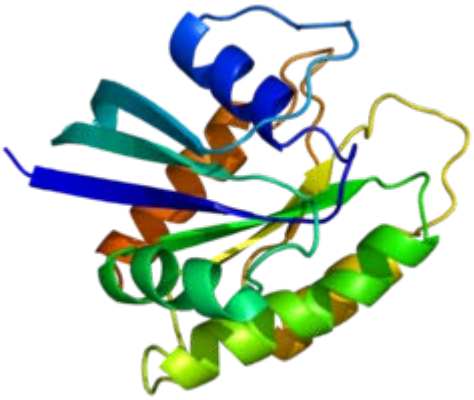
Another version of the HLA-B gene, HLA-B53, has been shown to help protect against severe malaria. HLA-B53 is most common in West African populations, where malaria is a frequent cause of death in children. Researchers suggest that this version of the HLA-B gene may help the immune system respond more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

Chapter 11

HRAS and Huntingtin

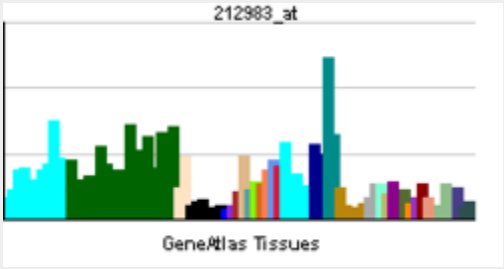
HRAS

V-Ha-ras Harvey rat sarcoma viral oncogene homolog



PDB rendering based on 121p.

Available structures	
Identifiers	
Symbols	HRAS; HRAS1; K-ras; N-ras; RASH1; c-bas/has
External IDs	OMIM: 190020 MGI: 96224 HomoloGene: 55890 GeneCards: HRAS Gene
RNA expression pattern	



Orthologs		
Species	**Human**	**Mouse**
Entrez	3265	15461
Ensembl	ENSG00000174775	ENSMUSG00000025499
UniProt	P01112	O70564
RefSeq (mRNA)	NM_005343	NM_008284
RefSeq (protein)	NP_005334	NP_032310
Location (UCSC)	Chr 11: 0.52 - 0.53 Mb	Chr 7: 141.04 - 141.05 Mb

GTPase HRas also known as **transforming protein p21** is an enzyme that in humans is encoded by the *HRAS* gene. The HRAS gene is located on the short (p) arm of chromosome 11 at position 15.5, from base pair 522,241 to base pair 525,549.

Function

GTPase HRas is involved in regulating cell division in response to growth factor stimulation. Growth factors act by binding cell surface receptors that span the cell's plasma membrane. Once activated, receptor stimulate signal transduction events in the cytoplasm, a process by which proteins and second messengers relay signals from outside the cell to the cell nucleus and instructs the cell to grow or divide. The HRAS protein is a GTPase and is an early player in many signal transduction pathways and is usually associated with cell membranes due to the presence of an isoprenyl group on its C-terminus. HRAS acts as a molecular on/off switch, once it is turned on it recruits and activates proteins necessary for the propagation of the receptor's signal, such as c-Raf and PI 3-kinase. HRAS binds to GTP in the active state and possesses an intrinsic enzymatic activity that cleaves the terminal phosphate of this nucleotide converting it to GDP. Upon conversion of GTP to GDP, HRAS is turned off. The rate of conversion is usually slow but can be sped up dramatically by an accessory protein of the Guanine nucleotide activating protein (GAP) class, for example RasGAP. In turn HRAS can bind to proteins

of the Guanine Nucleotide Exchange Factor (GEF) class, for example SOS1, which forces the release of bound nucleotide. Subsequently, the unbound HRAS is released from the GEF and quickly re-binds available GTP or GDP present in the cytosol. Since GTP is substantially more abundant than GDP, this usually results in HRAS activation. HRAS is in the Ras family, which also includes three other proto-oncogenes: KRAS, RRAS and NRAS. These proteins all are regulated in the same manner and appear to differ largely in their sites of action within the cell.

Clinical significance

Costello syndrome

At least five inherited mutations in the HRAS gene have been identified in people with Costello syndrome. Each of these mutations changes a single protein building block (amino acid) in a critical region of the HRAS protein. The most common mutation replaces the amino acid glycine with the amino acid serine at position 12 (written as Gly12Ser or G12S). The mutations responsible for Costello syndrome lead to the production of an HRAS protein that is permanently active. Instead of triggering cell growth in response to particular signals from outside the cell, the overactive protein directs cells to grow and divide constantly. This uncontrolled cell division can result in the formation of noncancerous and cancerous tumors. Researchers are uncertain how mutations in the HRAS gene cause the other features of Costello syndrome (such as mental retardation, distinctive facial features, and heart problems), but many of the signs and symptoms probably result from cell overgrowth and abnormal cell division.

Bladder cancer

HRAS has been shown to be a proto-oncogene. When mutated, proto-oncogenes have the potential to cause normal cells to become cancerous. Some gene mutations are acquired during a person's lifetime and are present only in certain cells. These changes are called somatic mutations and are not inherited. Somatic mutations in the HRAS gene in bladder cells have been associated with bladder cancer. One specific mutation has been identified in a significant percentage of bladder tumors; this mutation substitutes one protein building block (amino acid) for another amino acid in the HRAS protein. Specifically, the mutation replaces the amino acid glycine with the amino acid valine at position 12 (written as Gly12Val or G12V). The altered HRAS protein is permanently activated within the cell. This overactive protein directs the cell to grow and divide in the absence of outside signals, leading to uncontrolled cell division and the formation of a tumor. Mutations in the HRAS gene also have been associated with the progression of bladder cancer and an increased risk of tumor recurrence after treatment.

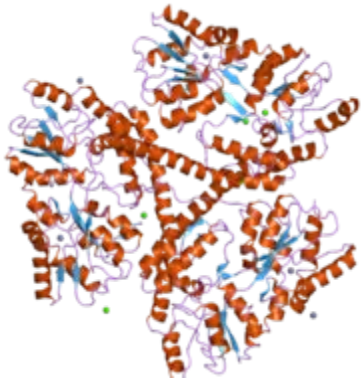
Other cancers

Somatic mutations in the HRAS gene are probably involved in the development of several other types of cancer. These mutations lead to an HRAS protein that is always active and can direct cells to grow and divide without control. Recent studies suggest that

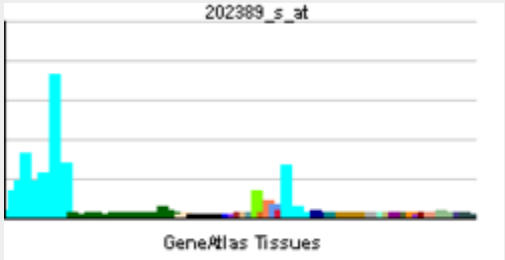
HRAS mutations may be common in thyroid and kidney cancers. The HRAS protein also may be produced at higher levels (overexpressed) in other types of cancer cells.

Huntingtin

Huntingtin (Huntington's disease)



Crystallographic structure of the N-terminal region of the human Huntingtin protein with an artifically attached Maltose-Binding protein used for crystallographic purposes.

Available structures		
Identifiers		
Symbols	HTT; IT15; HD	
External IDs	OMIM: 143100 MGI: 96067 HomoloGene: 1593 GeneCards: HTT Gene	
RNA expression pattern		
		
Orthologs		
Species	**Human**	**Mouse**
Entrez	3064	15194
Ensembl	ENSG00000197386	ENSMUSG00000029104

UniProt	P42858	P42859
RefSeq (mRNA)	NM_002111	NM_010414
RefSeq (protein)	NP_002102	NP_034544
Location (UCSC)	Chr 4: 3.05 - 3.22 Mb	Chr 5: 35.08 - 35.23 Mb

The **Huntington** gene, also called **HTT** or **HD** (**Huntington disease**) gene, is the **IT15** (**"interesting transcript 15"**) gene codes for a protein called the **huntington protein**. The gene and its product are under heavy investigation as part of Huntington's disease clinical research.

It is variable in its structure as there are many polymorphisms of the gene which can lead to variable numbers of glutamine residues present in the protein. In its wild-type (normal) form, it contains 6-35 glutamine residues, however, in individuals affected by Huntington's Disease (an autosomal dominant genetic disorder), it contains greater than 36 glutamine residues (highest reported repeat length is about 250). Its commonly used name is derived from this disease, previously the IT15 label was commonly used.

The mass of huntingtin protein is largely dependent on the number of glutamine residues it has, the predicted mass is around 350 kDa. Normal huntingtin is generally accepted to be 3144 amino acids in size. The exact function of this protein is not known, but it plays an important role in nerve cells. Within cells, huntingtin may be involved in signalling, transporting materials, binding proteins and other structures, and protecting against programmed cell death (apoptosis). The huntingtin protein is required for normal development before birth. It is expressed in many tissues in the body, with the highest levels of expression seen in the brain.

Gene

The 5' end of the HD gene has a sequence of 3 DNA bases, cytosine-adenine-guanine (CAG), coding for the amino acid glutamine, that is repeated multiple times. This region is called a trinucleotide repeat. Normal persons have a CAG repeat count of between 7 and 35 repeats.

The HD gene is located on the short (p) arm of chromosome 4 at position 16.3, from base pair 3,113,411 to base pair 3,282,655.

Protein

Function

The function of huntingtin is unclear. It is essential for development, and absence of huntingtin is lethal in mice. The protein has no sequence homology with other proteins and is highly expressed in neurons and testes in humans and rodents. Huntingtin upregulates the expression of Brain Derived Neurotrophic Factor (BDNF) at the transcription level but the mechanism by which huntingtin regulates gene expression has not been determined. From immunohistochemistry, electron microscopy, and subcellular fractionation studies of the molecule, it has been found that huntingtin is primarily associated with vesicles and microtubules. These appear to indicate a functional role in cytoskeletal anchoring or transport of mitochondria. The Htt protein is involved in vesicle trafficking as it interacts with HIP1, a clathrin binding protein, to mediate endocytosis, the absorption of materials into a cell.

Interactions

Huntingtin has been found to interact directly with at least 19 other proteins, of which 6 are used for transcription, 4 for transport, 3 for cell signalling and 6 others of unknown function (HIP5, HIP11, HIP13, HIP15, HIP16, and CGI-125). Over 100 interacting proteins have been found such as huntingtin-associated protein 1 (HAP1) and huntingtin Interacting Protein 1 (HIP1), these were typically found using two-hybrid screening and confirmed using Immunoprecipitation.

Interacting Protein	PolyQ length dependence	Function
α-adaptin C/HYPJ	Yes	Endocytosis
Akt/PKB	No	Kinase
CBP	Yes	Transcriptional co-activator with acetyltransferase activity
CA150	No	Transcriptional activator
CIP4	Yes	cdc42-dependent signal transduction
CtBP	Yes	Transcription factor
FIP2	Not known	Cell Morphogenesis
Grb2	Not known	Growth factor receptor binding protein
HAP1	Yes	Membrane trafficking
HAP40	Not known	Unknown
HIP1	Yes	Endocytosis, proapoptotic
HIP14/HYP-H	Yes	Trafficking, endocytosis
N-CoR	Yes	Nuclear receptor co-repressor
NF-κB	Not known	Transcription factor
p53	No	Transcription factor

PACSIN1	Yes	Endocytosis, actin cytoskeleton
PSS-95	Yes	Synaptic scaffolding protein
RasGAP	Not known	Ras GTPase activating protein
SH3GL3	Yes	Endocytosis
SIN3A	Yes	Transcriptional repressor
Sp1	Yes	Transcription factor

Huntingtin has also been shown to interact with:

- HIP2,
- MAP3K10,
- OPTN,
- PRPF40A,
- RASA1,
- SETD2,
- TRIP10,
- ZDHHC17.

Clinical significance

Huntington's disease (HD) is caused by a mutation in the Huntingtin gene, where the CAG repeats more than 35 times and is unstable. These expanded repeats lead to production of a huntingtin protein that contains an abnormally long polyglutamine tract at the N-terminal. This makes it part of a class of neurodegenerative disorders known as trinucleotide repeat disorders or polyglutamine disorders. The key sequence which is found in Huntington's disease is a trinucleotide repeat expansion of glutamine residues beginning at the 18th amino acid. In unaffected individuals, this contains between 9 and 35 glutamine residues with no adverse effects. However, 36 or more residues produce an erroneous form of Htt, **mHtt** (standing for mutant Htt). Reduced penetrance is found in counts 36-39.

Enzymes in the cell often cut this elongated protein into fragments. The protein fragments form abnormal clumps, known as neuronal intranuclear inclusions or NIIs, inside nerve cells and may attract other, normal proteins into the clumps. It was once thought that the presence of these clumps played a causal role in Huntington disease. Further research undermined this conclusion by showing that the presence of NIIs actually extended the life of neurons and acted to reduce intracelluar mutant huntingtin in neighboring neurons. Thus, the likelihood of neuronal death can be predicted by accounting for two factors: (1) the length of CAG repeats in the Huntingtin gene and (2) the neuron's exposure to diffuse intracellular mutant huntingtin protein. NIIs (protein clumping) can thereby be construed as a coping mechanism—as opposed to a pathogenic mechanism—to stem neuronal death by decreasing the amount of diffuse huntingtin. This process is particularly likely to occur in the striatum (a part of the brain that coordinates movement) primarily, and the frontal cortex (a part of the brain that controls thinking and emotions).

Classification of the trinucleotide repeat, and resulting disease status, depends on the number of CAG repeats

Repeat count	Classification	Disease status
<28	Normal	Unaffected
28–35	Intermediate	Unaffected
36–40	Reduced Penetrance	+/- Affected
>40	Full Penetrance	Affected

People with 36 to 40 CAG repeats may or may not develop the signs and symptoms of Huntington disease, while people with more than 40 repeats will develop the disorder during a normal lifetime. When there are more than 60 CAG repeats, the person develops a severe form of HD known as Juvenile HD. Therefore, the number of CAG (the sequence coding for the amino acid glutamine) repeats influences the age of onset of the disease. No case of HD has been diagnosed with a count less than 36.

As the altered gene is passed from one generation to the next, the size of the CAG repeat expansion can change; it often increases in size, especially when it is inherited from the father. People with 28 to 35 CAG repeats have not been reported to develop the disorder, but their children are at risk of having the disease if the repeat expansion increases.

Chapter 12

KIAA0090 and MECP2

KIAA0090

KIAA0090 is a human gene coding for a protein of unknown function. KIAA0090 has two aliases OTTHUMP00000002581 and RP1-43E13.1. The gene codes for multiple transcript variants which can localize to different subcellular compartments. KIAA0090 interacts with multiple effector proteins. KIAA0090 contains a conserved COG1520 WD40 like repeat domain thought to be the method of such interaction.

Characterization of the KIAA0090 Gene and its Transcript Products

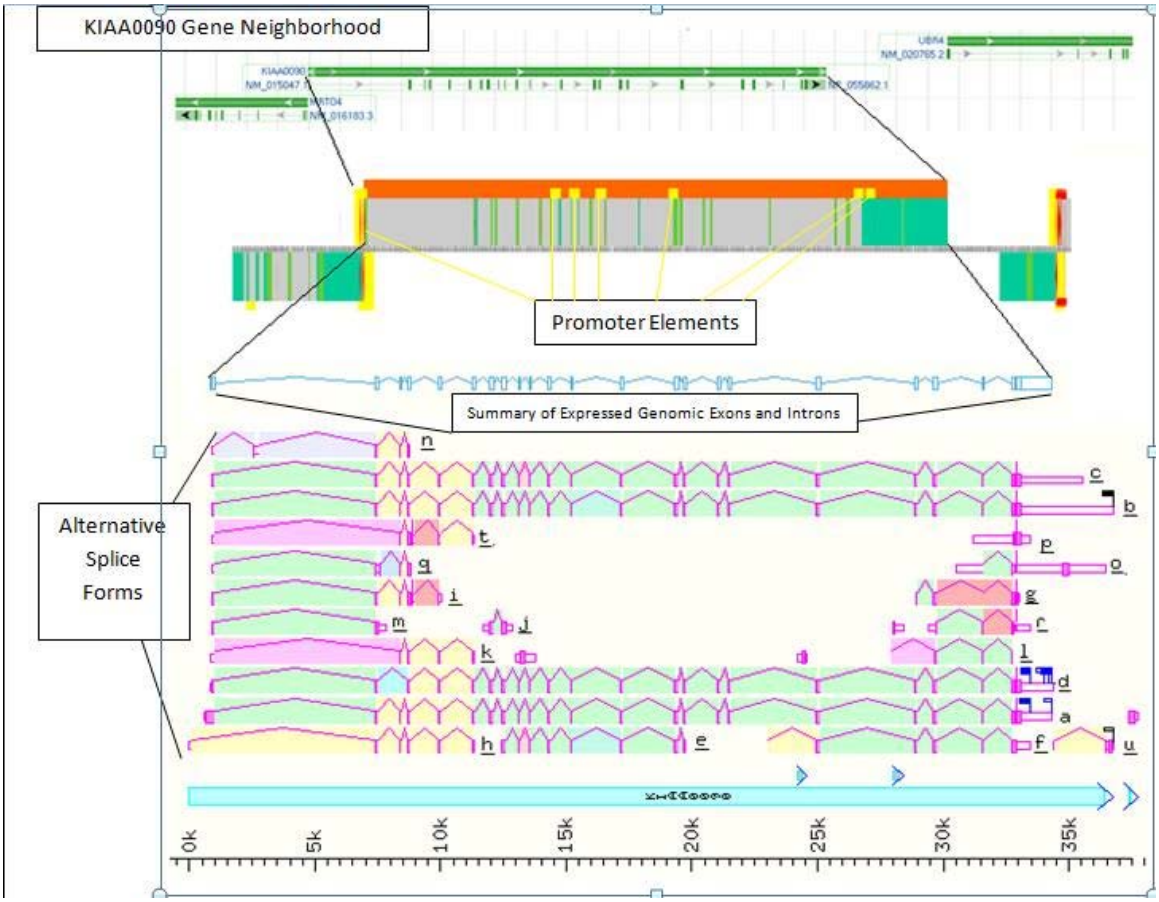


Figure 1: A graphical summary of the KIAA0090 gene neighborhood, regions of promotion, exons/introns, and variant transcript products

KIAA0090 is located on chromosome one in the p arm at location 1p36.132. It covers 36.74 kb, from base pairs 19451486 to 19414744. The gene is composed of 37 gt-at introns/alternative introns with 57 exons expressed in 1 unspliced form of 4253 bp and 20 alternatively spliced forms of varying lengths. The gene has 8 probable promoters. The gene is flanked by UBR4 on its right and MRTO4 on its left. This Information is graphically displayed in **Figure 1**.

Expressed Sequence Tags and isolated cDNA clones indicate KIAA0090 is expressed ubiquitously in low to moderate levels throughout the body. This includes but is not limited to testis, tongue, lung, cerebellum, brain, mammary gland, trachea, placenta, esophageal, salivary gland, brain, hippocampus, amygdale, bone marrow, thalamus, spleen, uterus, thymus, kidney, eye, heart, gall bladder, prostate, liver, parathyroid gland, ovary, stomach, skeletal muscle, colon, pancreas, and skin. Expression of KIAA0090 changes throughout development (embryogenesis, fetal, adult, etc.)and during carcinogenesis. Evidence indicates a correlation between conditions and expression level but no data exists to suggest KIAA0090 is responsible for any disease or stage of development.

The mRNA for this gene codes for 18 protein isoforms6. The remaining 3 splice variants have no evidence supporting their ability to be translated.

Characterization of the KIAA0090 Protein Product

Analysis indicates the KIAA0090 unspliced protein product to be 993 amino acids long with an isoelectric point of 7.418 and a molecular weight of 111765.73 Daltons. The primary structure of this protein contains 4 conserved domains. This includes a signal peptide from position 1 to 22, a COG1520 WD40 like domain, a leucine zipper domain, a DUF1620 domain (domain of unknown function), and a transmembrane domain. These can be viewed in **Figure 2**. Several conserved cysteine residues are present at positions 226,235, 335, 364,449, 581, 675, 925, and 985. Several internal localization signals are also present. Dependent on splice outcome and posttranslational modification, these additional signals indicate the protein could localize to the peroxisome, the plasma membrane, outside the cell, the cytosol, the nucleus, or mitochondria.

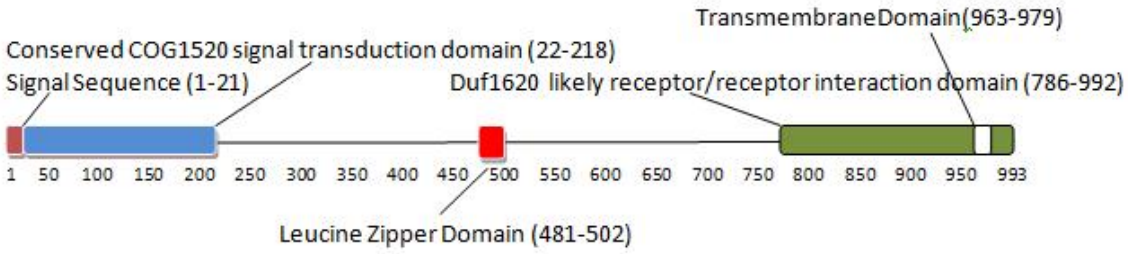


Figure 2: An annotated protein map of KIAA0090 with domain residence and function

Post translational modification of KIAA0090 can occur. 54 possible sites of phosphorylation exist; 33 serines, 10 threonines, and 11 tyrosines. 3 sites of N-linked glycosylation are present at residues 370, 818, and 913 The signal peptide can be cleaved between residues 21 and 22.

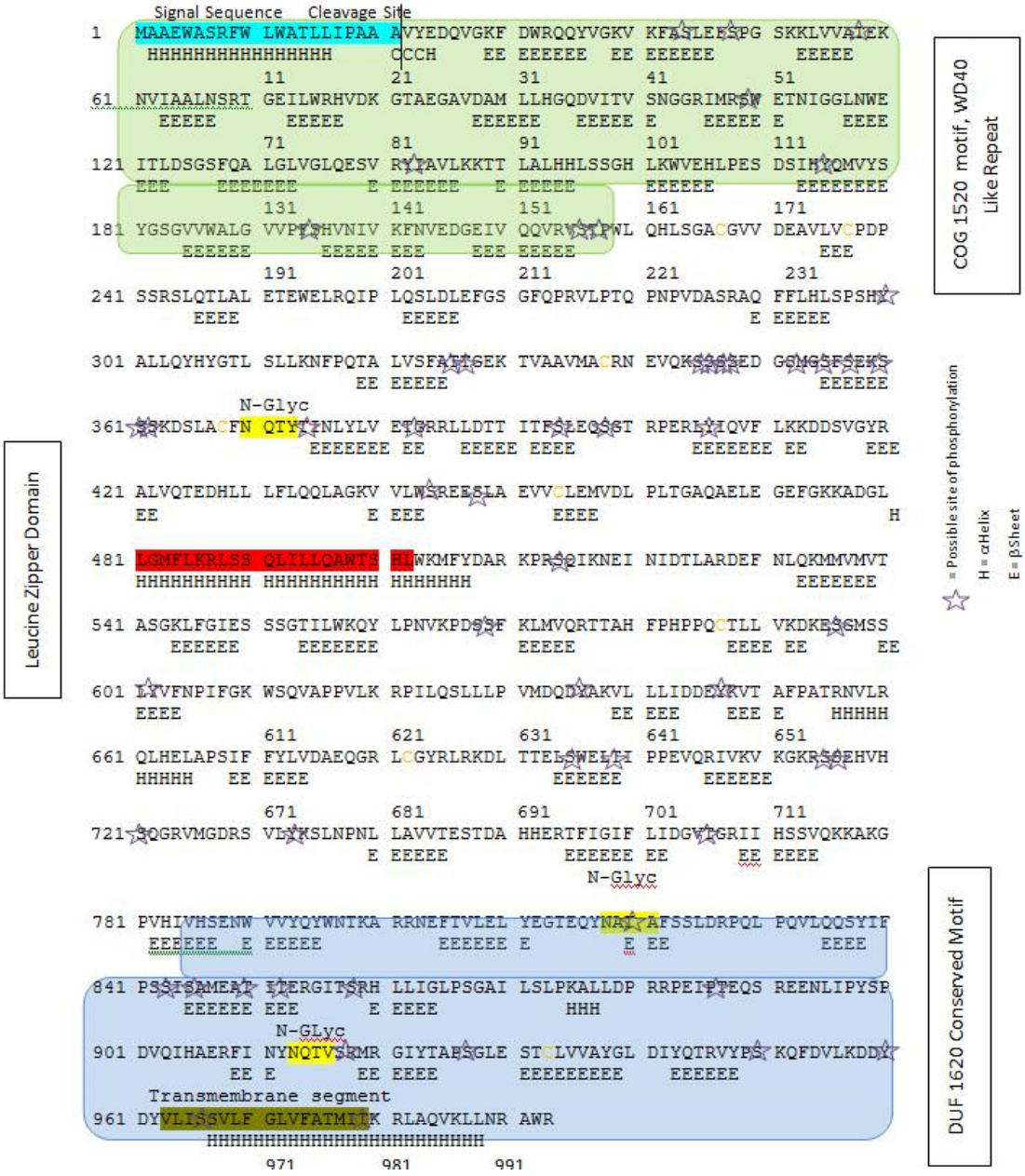


Figure 3: A conceptual translation of KIAA0090 with domains, sites of phosphorylation, sites of N-linked glycosylation, and signal peptide.

This information is graphically displayed in **Figure 3**. Structure beyond the primary remains predicative. Bioinformatic analysis yields consensus data that is also displayed in

Figure 3. The protein is highly conserved throughout Eukaryotes both in multi and single cellular organisms. This includes but is not limited to animals, plants, fungi, and protists.

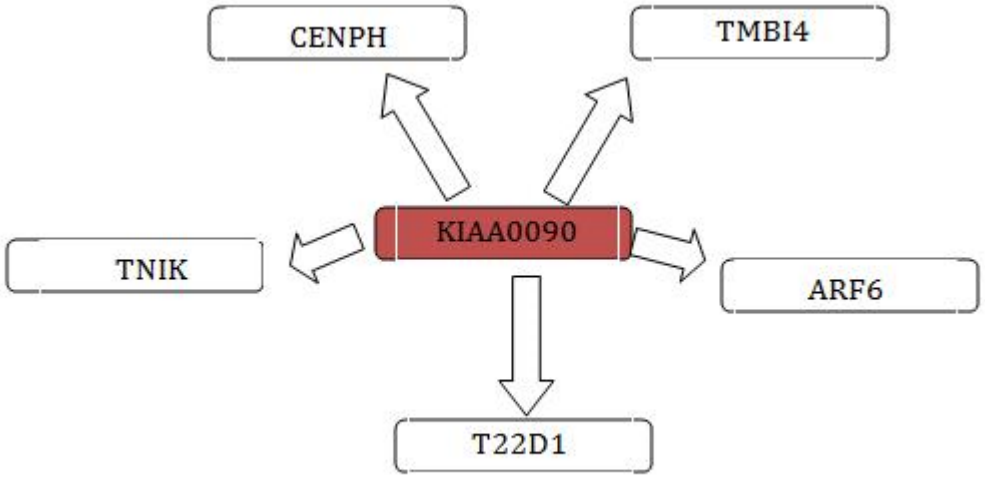
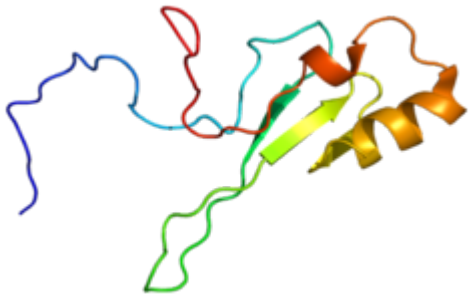


Figure 4: KIAA0090/Protein Interactions

The WD40 like domain COG1520 is KIAA0090s only identified functional effector domain. WD40 containing proteins are signal transducers involved in transduction of signals to binding factors, the centromeres, and other effectors. Coimmunoprecipitation experiments have proven KIAA0090 interaction with these types of proteins; specifically the centromeric protein CENPH, the BAX Inhibitor TMBI4, the ADP ribosylation factor ARF6, the kinase TNIK, and the transcriptional repressor T22D1. The number of splice variants indicates this list is probably not definitive. As further characterization is completed additional interactions would be expected.

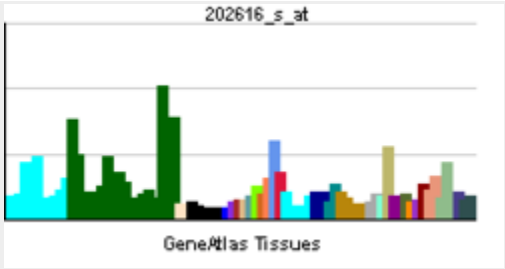
MECP2

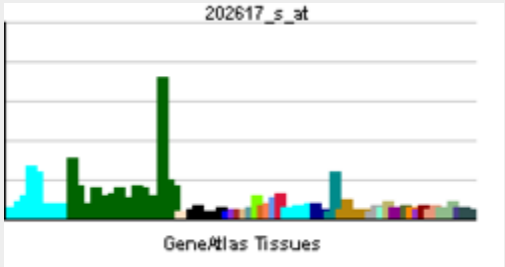
Methyl CpG binding protein 2 (Rett syndrome)

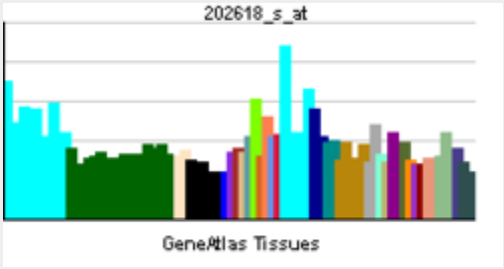


PDB rendering based on 1qk9.

Available structures	
Identifiers	
Symbols	MECP2; RTS; AUTSX3; DKFZp686A24160; MRX16; MRX79; PPMX; RTT
External IDs	OMIM: 300005 MGI: 99918 HomoloGene: 3657 GeneCards: MECP2 Gene
RNA expression pattern	







Orthologs		
Species	**Human**	**Mouse**
Entrez	4204	17257
Ensembl	ENSG00000169057	ENSMUSG00000031393
UniProt	P51608	Q3TYG1
RefSeq (mRNA)	NM_004992	NM_001081979
RefSeq (protein)	NP_004983	NP_001075448
Location (UCSC)	Chr X: 152.94 - 153.02 Mb	Chr X: 70.29 - 70.34 Mb

MECP2 (**methyl CpG binding protein 2 (Rett syndrome)**) is a gene that provides instructions for making its protein product, MECP2, also referred to as MeCP2. MECP2 appears to be essential for the normal function of nerve cells. The protein seems to be particularly important for mature nerve cells, where it is present in high levels. The MeCP2 protein is likely to be involved in turning off ("repressing" or "silencing") several other genes. This prevents the genes from making proteins when they are not needed. Recent work has shown that MeCP2 can also activate other genes. The MECP2 gene is located on the long (q) arm of the X chromosome in band 28 ("Xq28"), from base pair 152,808,110 to base pair 152,878,611.

DNA methylation is the major modification of eukaryotic genomes and plays an essential role in mammalian development. Human proteins MECP2 (this protein), MBD1, MBD2, MBD3, and MBD4 comprise a family of nuclear proteins related by the presence in each of a methyl-CpG binding domain (MBD). Each of these proteins, with the exception of MBD3, is capable of binding specifically to methylated DNA. MECP2, MBD1 and MBD2 can also repress transcription from methylated gene promoters. In contrast to other MBD family members, MECP2 is X-linked and subject to X inactivation. MECP2 is dispensible in stem cells. MECP2 gene mutations are the cause of some cases of Rett syndrome, a progressive neurologic developmental disorder and one of the most common causes of mental retardation in females.

Function

MeCP2 protein is found in all cells in the body, including the brain, acting as a transcriptional repressor and activator, depending on the context. However, the idea that MeCP2 functions as an activator is relatively new and remains controversial. In the brain, it is found in high concentrations in neurons and is associated with maturation of the central nervous system (CNS) and in forming synaptic contacts.

Mechanism of action

The MeCP2 protein binds to forms of DNA that have been methylated. The MeCP2 protein then interacts with other proteins to form a complex that turns off the gene. Methylation is a chemical alteration made to a "cytosine" (C) when it occurs in a particular DNA sequence, "CpG". Many genes have CpG islands, which frequently occur near the beginning of the gene. MECP2 does not bind to these islands in most cases, as they are not methylated. The expression of a few genes may be regulated through methylation of their CpG island, and MECP2 may play a role in a subset of these. Researchers have not yet determined which genes are targeted by the MeCP2 protein, but such genes are probably important for the normal function of the central nervous system. However, the first large-scale mapping of MECP2 binding sites in neurons found that only 6% of the binding sites are in CpG islands, and that 63% of MECP2-bound promoters are actively expressed and only 6% are highly methylated, indicating that MECP2's main function is something other than silencing methylated promoters.

Once bound, MeCP2 will condense the chromatin structure, form a complex with histone deacetylases (HDAC), or block transcription factors directly. More recent studies have demonstrated that MeCP2 may also function as a transcriptional activator, through recruiting the transcription factor CREB1. This was an unexpected finding which suggests that MeCP2 is a key transcriptional regulator with potentially dual roles in gene expression. In fact, the majority of genes that are regulated by MeCP2 appear to be activated rather than repressed. However, it remains controversial whether MeCP2 regulates these genes directly or whether these changes are secondary in nature. Further studies have shown MeCP2 may be able to bind directly to un-methylated DNA in some instances. MeCP2 has been implicated in regulation of imprinted genes and loci that include UBE3A and DLX5.

Structure

MeCP2 is part of a family of methyl-CpG-binding domain proteins (MBD), but possesses its own unique differences which help set it apart from the group. It has two functional domains:

- a methyl-cytosine-binding domain (MBD) composed of 85 amino acids; and
- a transcriptional repression domain (TRD) composed of 104 amino acids

The MBD domain forms a wedge and attaches to the methylated CpG sites on the DNA strands. The TRD region then reacts with SIN3A to recruit histone deacetylases (HDAC). There are also unusual, repetitive sequences found at the carboxyl terminus. This region is closely related to the fork head family, at the amino acid level.

Role in disease

Rett syndrome is caused by mutations in the MECP2 gene. Several types of mutations have been identified in people with Rett syndrome. These mutations include changes in single base pairs (the building material of DNA), insertions or deletions of DNA in the gene, and changes that affect how the gene is processed into a protein. Mutations in the gene alter the structure of the MeCP2 protein or lead to reduced amounts of the protein. As a result, the protein is unable to bind to DNA or turn other genes on or off. Genes that are normally repressed by MeCP2 remain active and continue to make large amounts of particular proteins when they are not needed. Other genes that are normally activated by MeCP2 remain inactive and thus unable to make protein products. This defect probably disrupts the normal functioning of nerve cells, leading to the signs and symptoms of Rett syndrome.

This disease is mainly found in girls with a prevalence of around 1 in every 10,000. Patients are born with very hard to find signs of a disorder, but after about six months to a year and half, speech and motor function capabilities start to decrease. This is followed by seizures, growth retardation and cognitive and motor impairment. This is a X-linked dominant disease that is found predominatley affecting the paternal X chromosome. It has been linked to male lethality, due to its prevalence in females, but in rare cases some males can also be affected by Rett Syndrome.

Mutations in the MECP2 gene have also been identified in people with several other disorders affecting the central nervous system. For example, MECP2 mutations are associated with some cases of moderate to severe X-linked mental retardation. Mutations in the gene have also been found in males with severe brain dysfunction (neonatal encephalopathy) who live only into early childhood. In addition, several people with features of both Rett syndrome and Angelman syndrome (a condition characterized by mental retardation, problems with movement, and inappropriate laughter and excitability) have mutations in the MECP2 gene. Lastly, MECP2 mutations or changes in the gene's activity have been reported in some cases of autism (a developmental disorder that affects communication and social interaction).

More recent studies reported genetic polymorphisms in the MeCP2 genes in patients with systemic lupus erythematosus (SLE). SLE is a systemic autoimmune disease that can affect multiple organs. MeCP2 polymorphisms have been reported so far in European-derived and Asian lupus patients.

The genetic loss of MECP2 has been identified as changing the properties of cells in the locus ceruleus the exclusive source of noradrenergic innervation to the cerebral cortex and hippocampus.

Researchers have concluded that "Because these neurons are a pivotal source of norepinephrine throughout the brainstem and forebrain and are involved in the regulation of diverse functions disrupted in Rett syndrome, such as respiration and cognition, we hypothesize that the locus ceruleus is a critical site at which loss of MECP2 results in CNS dysfunction."

Interactions

MECP2 has been shown to interact with SKI protein and Nuclear receptor co-repressor 1. In neuronal cells the MECP2 mRNA is thought to interact with miR-132, which silences the expression of the protein. This forms part of a homeostatic mechanism that could regulate MECP2 levels in the brain.

Chapter 13

Pseudogene and PTPN11

Pseudogene

Pseudogenes are dysfunctional relatives of known genes that have lost their protein-coding ability or are otherwise no longer expressed in the cell. Although some do not have introns or promoters (these pseudogenes are copied from mRNA and incorporated into the chromosome and are called processed pseudogenes), most have some gene-like features (such as promoters, CpG islands, and splice sites), they are nonetheless considered nonfunctional, due to their lack of protein-coding ability resulting from various genetic disablements (stop codons, frameshifts, or a lack of transcription) or their inability to encode RNA (such as with rRNA pseudogenes). Thus the term, coined in 1977 by Jacq, *et al.*, is composed of the prefix *pseudo*, which means *false*, and the root *gene*, which is the central unit of molecular genetics.

Because pseudogenes are generally thought of as the last stop for genomic material that is to be removed from the genome, they are often labeled as junk DNA. Nonetheless, pseudogenes contain fascinating biological and evolutionary histories within their sequences. This is due to a pseudogene's shared ancestry with a functional gene: in the same way that Darwin thought of two species as possibly having a shared common ancestry followed by millions of years of evolutionary divergence, a pseudogene and its associated functional gene also share a common ancestor and have diverged as separate genetic entities over millions of years.

Properties of pseudogenes

Pseudogenes are characterized by a combination of **homology** to a known gene and **nonfunctionality**. That is, although every pseudogene has a DNA sequence that is similar to some functional gene, they are nonetheless unable to produce functional final products (nonfunctionality). Pseudogenes are quite difficult to identify and characterize in genomes, because the two requirements of homology and nonfunctionality are implied through sequence calculations and alignments rather than biologically proven.

1. Homology is implied by sequence identity between the DNA sequences of the pseudogene and parent gene. After aligning the two sequences, the percentage of identical base pairs is computed. A high sequence identity (usually between 40% and close to 100%) means that it is highly likely that these two sequences diverged from a common ancestral sequence (are homologous), and highly unlikely that these two sequences were independently created.
2. Nonfunctionality can manifest itself in many ways. Normally, a gene must go through several steps in going from a genetic DNA sequence to a fully-functional protein: transcription, pre-mRNA processing, translation, and protein folding are all required parts of this process. If any of these steps fails, then the sequence may be considered nonfunctional. In high-throughput pseudogene identification, the most commonly identified disablements are stop codons and frameshifts, which almost universally prevent the translation of a functional protein product.
3. Pseudogenes for RNA genes are often easier to discover. Many RNA genes occur as multiple copy genes, and pseudogenes are identified through sequence identity and location within the region.

Types and origin of pseudogenes

There are three main types of pseudogenes, all with distinct mechanisms of origin and characteristic features. The classifications of pseudogenes are as follows:

1. **Processed** (or **retrotransposed**) pseudogenes. In higher eukaryotes, particularly mammals, retrotransposition is a fairly common event that has had a huge impact on the composition of the genome. For example, somewhere between 30% - 44% of the human genome consists of repetitive elements such as SINEs and LINEs. In the process of retrotransposition, a portion of the mRNA transcript of a gene is spontaneously reverse transcribed back into DNA and inserted into chromosomal DNA. Although retrotransposons usually create copies of themselves, it has been shown in an *in vitro* system that they can create retrotransposed copies of random genes, too. Once these pseudogenes are inserted back into the genome, they usually contain a poly-A tail, and usually have had their introns spliced out; these are both hallmark features of cDNAs. However, because they are derived from a mature mRNA product, processed pseudogenes also lack the upstream promoters of normal genes; thus, they are considered "dead on arrival", becoming non-functional pseudogenes immediately upon the retrotransposition event. However, occasionally these insertions contribute exons to existing genes and usually via alternatively spliced transcripts. A further characteristic of processed pseudogenes is common truncation of the 5' end relative to the parent sequence, which is a result of the relatively non-processive retrotransposition mechanism that creates processed pseudogenes.
2. **Non-processed** (or **duplicated**) pseudogenes. Gene duplication is another common and important process in the evolution of genomes. A copy of a functional gene may arise as a result of a gene duplication event and subsequently acquire mutations that cause it to become nonfunctional. Duplicated pseudogenes usually have all the same characteristics of genes, including an intact exon-intron

structure and promoter sequences. The loss of a duplicated gene's functionality usually has little effect on an organism's fitness, since an intact functional copy still exists. According to some evolutionary models, shared duplicated pseudogenes indicate the evolutionary relatedness of humans and the other primates.

3. **Disabled** genes, or **unitary** pseudogenes. Various mutations can stop a gene from being successfully transcribed or translated, and a gene may become nonfunctional or deactivated if such a mutation becomes fixed in the population. This is the same mechanism by which non-processed genes become deactivated, but the difference in this case is that the gene was not duplicated before becoming disabled. Normally, such gene deactivation would be unlikely to become fixed in a population, but various population effects, such as genetic drift, a population bottleneck, or in some cases, natural selection, can lead to fixation. The classic example of a unitary pseudogene is the gene that presumably coded the enzyme L-gulono-γ-lactone oxidase (GULO) in primates. In all mammals studied besides primates (except guinea pigs), GULO aids in the biosynthesis of Ascorbic acid (vitamin C), but it exists as a disabled gene (GULOP) in humans and other primates. Another interesting and more recent example of a disabled gene, which links the deactivation of the caspase 12 gene (through a nonsense mutation) to positive selection in humans.

Pseudogenes can complicate molecular genetic studies. For example, a researcher who wants to amplify a gene by PCR may simultaneously amplify a pseudogene that shares similar sequences. This is known as PCR bias or amplification bias. Similarly, pseudogenes are sometimes annotated as genes in genome sequences.

Processed pseudogenes often pose a problem for gene prediction programs, often being misidentified as real genes or exons. It has been proposed that identification of processed pseudogenes can help improve the accuracy of gene prediction methods.

It has also been shown that the parent sequences that give rise to processed pseudogenes lose their coding potential faster than those giving rise to non-processed pseudogenes.

Functional pseudogenes?

By definition, pseudogenes lack a function. However, the classification of pseudogenes generally relies on computational analysis of genomic sequences using complex algorithms. This has led to the incorrect identification of pseudogenes. For example the functional, chimeric gene *jingwei* in *Drosophila* was once thought to be a processed pseudogene.

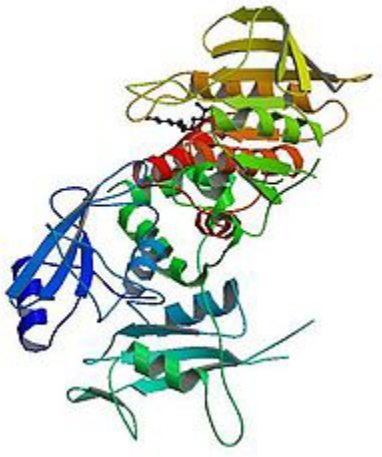
It has been established that quite a few pseudogenes can go through the process of transcription, either if their own promoter is still intact or in some cases using the promoter of a nearby gene; this expression of pseudogenes also appears to be tissue-specific. In 2003, Hirotsune *et al.* identified a retrotransposed pseudogene whose transcript purportedly plays a *trans*-regulatory role in the expression of its homologous

gene, *Makorin1* (MKRN1) and suggested this as a general model under which pseudogenes may play an important biological role. Other researchers have since hypothesized similar roles for other pseudogenes. A bioinformatics analysis has shown that processed pseudogenes can be inserted into introns of annotated genes and be incorporated into alternatively spliced transcripts. Hirotsune's report prompted two molecular biologists to carefully review scientific literature on the subject of pseudogenes. To the surprise of many, they found a number of examples in which pseudogenes play a role in gene regulation and expression, forcing Hirotsune's group to rescind their claim that they were the first to identify pseudogene function. Furthermore, the original findings of Hirotsune *et al.* concerning *Makorin1* have recently been strongly contested; thus, the possibility that some pseudogenes could have important biological functions was disputed. Additionally, University of Chicago and University of Cincinnati scientists reported in 2002 that a processed pseudogene called phosphoglycerate mutase 3 (*PGAM3P*) actually produces a functional protein.

Two 2008 publications in Nature discuss that some endogenous siRNAs are derived from pseudogenes, and thus some pseudogenes play a role in regulating protein-coding transcripts. In June 2010, Nature published an article showing the mRNA levels of tumour suppressor *PTEN* and oncogenic *KRAS* is affected by their pseudogenes *PTENP1* and *KRAS1P*. This discovery demonstrated an miRNA decoy function for pseudogenes and identified their transcripts as biologically active units in tumor biology; thus attributing a novel biological role to expressed pseudogenes, as they can regulate coding gene expression, and reveal a non-coding function for mRNAs in disease progression.

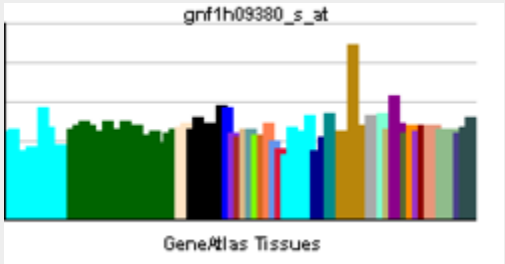
PTPN11

Protein tyrosine phosphatase, non-receptor type 11 (Noonan syndrome 1)



PDB rendering based on 2shp.

Available structures	
Identifiers	
Symbols	PTPN11; BPTP3; CFC; MGC14433; NS1; PTP-1D; PTP2C; SH-PTP2; SH-PTP3; SHP2 ; Syp
External IDs	OMIM: 176876 MGI: 99511 HomoloGene: 2122 GeneCards: PTPN11 Gene
RNA expression pattern	



Orthologs		
Species	**Human**	**Mouse**
Entrez	5781	19247

Ensembl	n/a	ENSMUSG00000043733
UniProt	Q06124	Q05C78
RefSeq (mRNA)	NM_002834	NM_011202
RefSeq (protein)	NP_002825	NP_035332
Location (UCSC)	n/a	Chr 5: 121.39 - 121.45 Mb

PTPN11 is a gene encoding the protein tyrosine phosphatase (PTP) Shp2.

The protein encoded by this gene is a member of the protein tyrosine phosphatase (PTP) family. PTPs are known to be signaling molecules that regulate a variety of cellular processes including cell growth, differentiation, mitotic cycle, and oncogenic transformation. This PTP contains two tandem Src homology-2 domains, which function as phospho-tyrosine binding domains and mediate the interaction of this PTP with its substrates. This PTP is widely expressed in most tissues and plays a regulatory role in various cell signaling events that are important for a diversity of cell functions, such as mitogenic activation, metabolic control, transcription regulation, and cell migration. Mutations in this gene are a cause of Noonan syndrome as well as acute myeloid leukemia.

Structure and function

This phosphatase, along with its paralogue, Shp1, possesses a domain structure that consists of two tandem SH2 domains in its N-terminus followed by a protein tyrosine phosphatase PTP domain. In the inactive state, the N-terminal SH2 domain binds the PTP domain and blocks access of potential substrates to the active site. Thus, Shp2 is auto-inhibited.

Upon binding to target phospho-tyrosyl residues, the N-terminal SH2 domain is released from the PTP domain, catalytically activating the enzyme by releaving this auto-inhibition.

Genetic diseases associated with PTPN11

Missense mutations in the PTPN11 locus are associated with both Noonan syndrome and Leopard syndrome.

It has also been associated with Metachondromatosis.

Noonan syndrome

In the case of Noonan syndrome, mutations are broadly distributed throughout the coding region of the gene but all appear to result in hyper-activated, or unregulated mutant forms of the protein. Most of these mutations disrupt the binding interface between the N-SH2 domain and catalytic core necessary for the enzyme to maintain its auto-inhibited conformation.

Leopard syndrome

The mutations that cause Leopard syndrome are restricted regions affecting the catalytic core of the enzyme producing catalytically impaired Shp2 variants. It is currently unclear how mutations that give rise to mutant variants of Shp2 with biochemically opposite characteristics result in similar human genetic syndromes.

PTPN11 mutations in cancer

Patients with a subset of Noonan syndrome PTPN11 mutations also have a higher prevalence of juvenile myelomonocytic leukemias. Activating Shp2 mutations have also been detected in neuroblastoma, melanoma, acute myeloid leukemia, breast cancer, lung cancer, colorectal cancer. These data suggests that Shp2 may be a proto-oncogene.

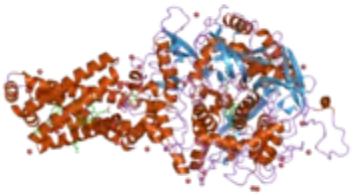
Interactions

PTPN11 has been shown to interact with PLCG2, Janus kinase 2, GAB2, Growth hormone receptor, PTK2B, CD117, Janus kinase 1, CD31, CEACAM1, STAT5A, STAT5B, FRS2, Insulin receptor, GAB1, SOCS3, Glycoprotein 130, Insulin-like growth factor 1 receptor, Grb2, SLAMF1, Epidermal growth factor receptor, PDGFRB, Cbl gene, SOS1, IRS1 and LAIR1.

Chapter 14

SDHB

succinate dehydrogenase complex, subunit B, iron sulfur (Ip)



PDB rendering based on 2h89.

Available structures		
Identifiers		
Symbols	SDHB; SDHIP; SDH1; SDH2; PGL4; IP; FLJ92337; SDH	
External IDs	OMIM: 185470 MGI: 1914930 HomoloGene: 2255 GeneCards: SDHB Gene	
EC number	1.3.5.1	
Orthologs		
Species	**Human**	**Mouse**
Entrez	6390	67680
Ensembl	ENSG00000117118	ENSMUSG00000009863

UniProt	P21912	Q9CQA3
RefSeq (mRNA)	NM_003000	NM_023374
RefSeq (protein)	NP_002991	NP_075863
Location (UCSC)	Chr 1: 17.22 - 17.25 Mb	Chr 4: 140.52 - 140.54 Mb

Succinate dehydrogenase [ubiquinone] iron-sulfur subunit, mitochondrial (SDHB) also known as **iron-sulfur subunit of complex II** (Ip) is a protein that in humans is encoded by the *SDHB* gene.

The succinate dehydrogenase (also called SDH or Complex II) protein complex catalyzes the oxidation of succinate (succinate + ubiquinone => fumarate + ubiquinol). SDHB is one of four protein subunits forming succinate dehydrogenase, the other three being SDHA, SDHC and SDHD. The SDHB subunit is connected to the SDHA subunit on the hydrophilic, catalytic end of the SDH complex. It is also connected to the SDHC/SDHD subunits on the hydrophobic end of the complex anchored in the mitochondrial membrane. The subunit is an iron-sulfur protein with three iron-sulfur clusters. It weighs 30 kDa.

Function of the SDHB protein

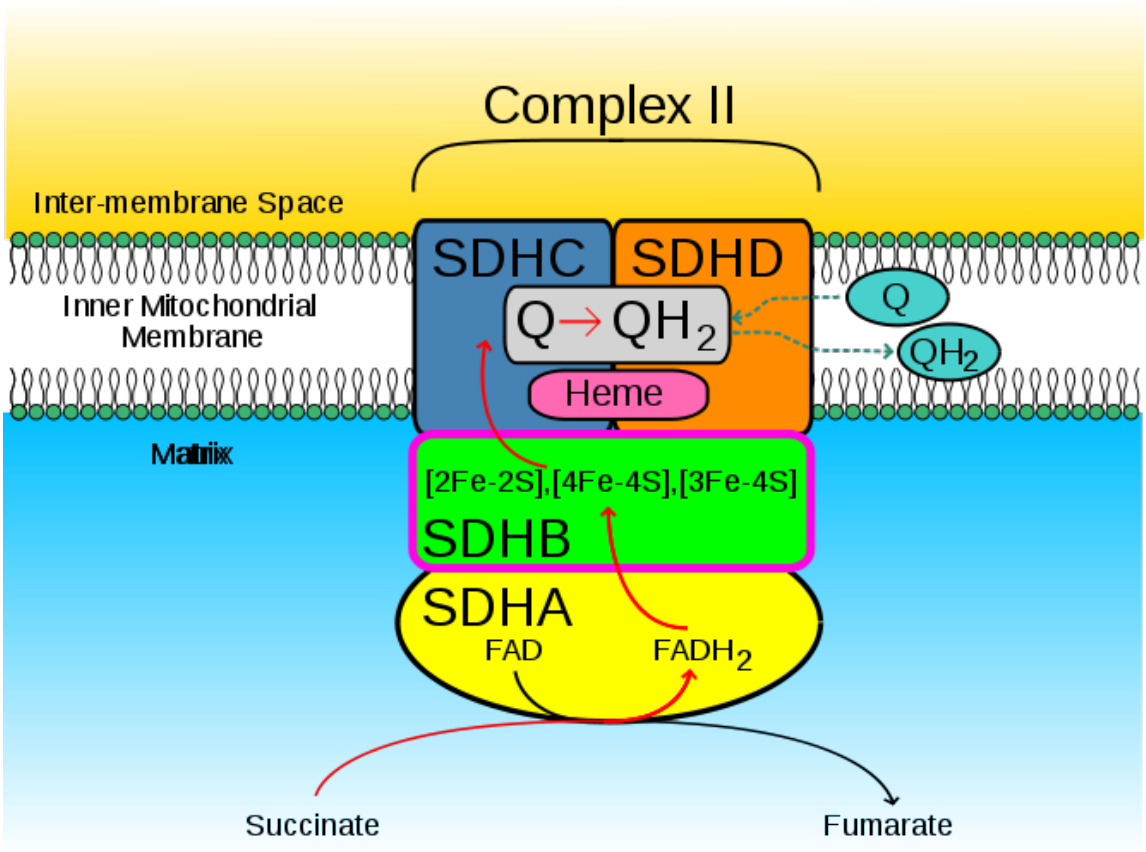


Figure 1: Function of the SDHB protein. Electrons are transferred from the Citric Acid Cycle to the Respiratory Chain. Electron path is shown by red arrows.

The SDH complex is located on the inner membrane of the mitochondria and participates in both the Citric Acid Cycle and Respiratory chain.

SDHB acts as an intermediate in the basic SDH enzyme action shown in Figure 1:

1. SDHA converts succinate to fumarate as part of the Citric Acid Cycle. This reaction also converts FAD to $FADH_2$.
2. Electrons from the $FADH_2$ are transferred to the SDHB subunit iron clusters [2Fe-2S],[4Fe-4S],[3Fe-4S].
3. Finally the electrons are transferred to the Ubiquinone (Q) pool via the SDHC/SDHD subunits.This function is part of the Respiratory chain.

Gene that codes for SDHB

The gene that codes for the SDHB protein is nuclear, not mitchondrial DNA. However, the protein is located in the inner membrane of the mitochondria. The location of the gene

in humans is on the first chromosome at locus p36.1-p35. The gene is coded in 1162 base pairs, partitioned in 8 exons. The expressed protein has 280 amino acids.

Role in Disease

Germline mutations in the gene can cause familial paraganglioma (in old nomenclature, Paraganglioma Type PGL4). The same condition is often called familial pheochromocytoma. Less frequently, renal cell carcinoma can be caused by this mutation.

Paraganlgiomas related to SDHB mutations have a high rate of malignancy. When malignant, treatment is currently the same as for any malignant paraganglioma/pheochromocytoma.

Tumour and Disease Characteristics

Paragangliomas caused by SDHB mutations have several distinguishing characteristics:

1. Malignancy is common, ranging from 38%-83%in carriers with disease. In contrast, tumors caused by SDHD mutations are almost always benign. Sporadic paragangliomas are malignant in less than 10% of cases.
2. Malignant paragangliomas caused by SDHB are usually (perhaps 92%) extra-adrenal. Sporadic pheochromocytomas/paragangliomas are extra-adrenal in less than 10% of cases.
3. The penetrance of the gene is often reported as 77% by age 50 (i.e. 77% of carriers will have at least one tumour by the age of 50). This is likely an overestimate. Currently (2010), families with silent SDHB mutations are being screened to determine the frequency of silent carriers.
4. The average age of onset is approximately the same for SDHB vs non-SDHB related disease (approximately 36 years).

Mutations causing disease have been seen in exons 1 through 7, but not 8. As with the SDHC and SDHD genes, SDHB is a tumor suppressor gene. Note the SDHA gene is not a tumor suppressor gene.

Tumor formation generally follows the Knudson "two hit" hypothesis. The first copy of the gene is mutated in all cells, however the second copy functions normally. When the second copy mutates in a certain cell due to a random event, Loss of Heterozygosity (LOH) occurs and the SDHB protein is no longer produced. Tumor formation then becomes possible.

Given the fundamental nature of the SDH protein in all cellular function, it is not currently understood why only paraganglionic cells are affected. However, the sensitivity of these cells to oxygen levels may play a role.

Disease pathways

The precise pathway leading from SDHB mutation to tumorigenesis is not determined; there are several proposed mechanisms.

Generation of reactive oxygen species

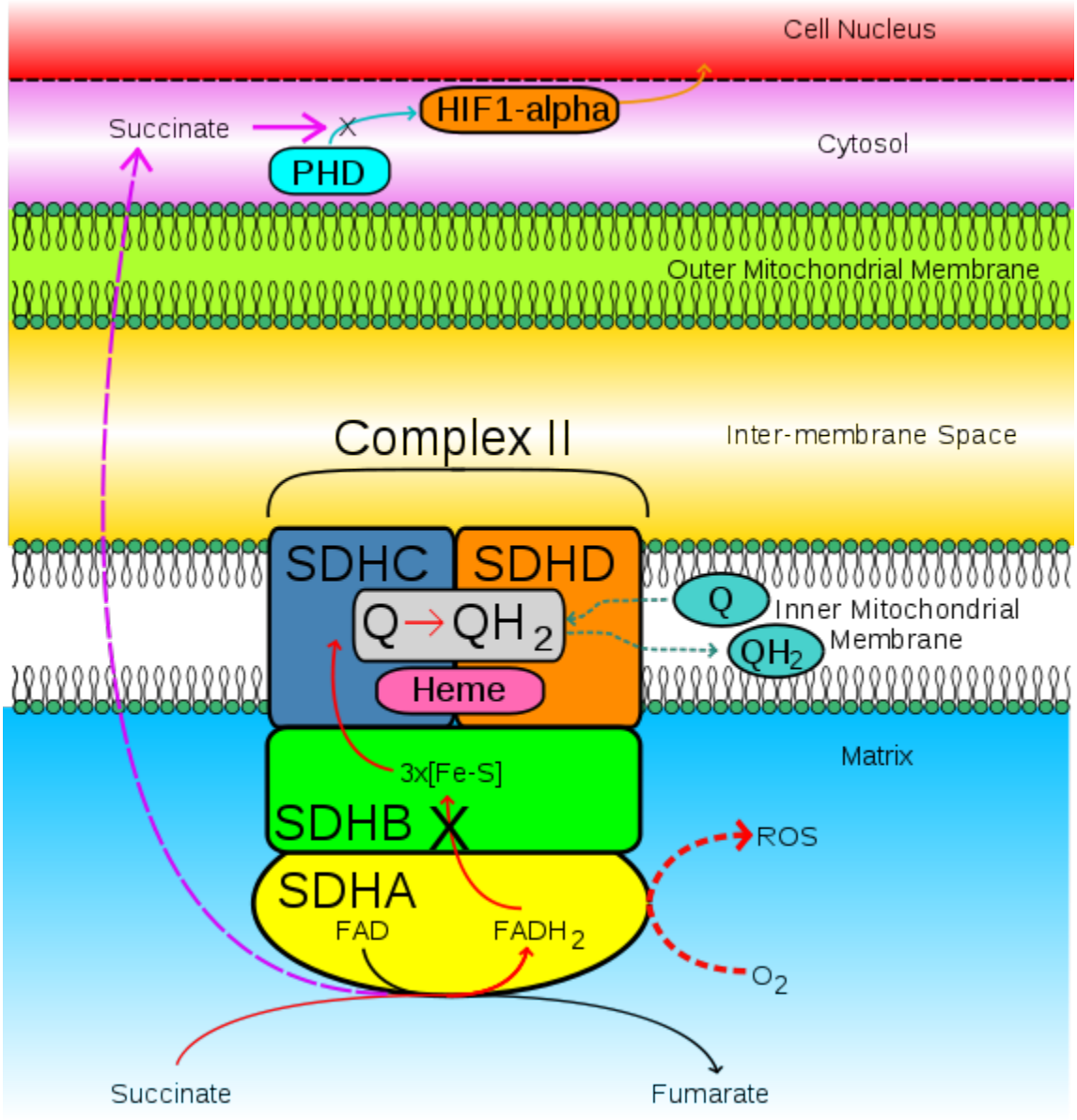


Figure 2: Disease Pathways for SDHB mutations. Electron path during normal function is shown by solid red arrows. Red dashed arrow shows superoxide generation (Pathway 1). Purple dashed arrow shows diffusion of succinate to block PHD (Pathway 2). Black crosses indicate the non-mutated process is blocked.

When succinate-ubiquinone activity is inhibited, electrons that would normally transfer through the SDHB subunit to the Ubiquinone pool are instead transferred to O_2 to create Reactive Oxygen Species (ROS) such as superoxide. The dashed red arrow in Figure 2 shows this. ROS accumulate and stabilize the production of HIF1-α. HIF1-α combines with HIF1-β to form the stable HIF heterodimeric complex, in turn leading to the induction of antiapoptotic genes in the cell nucleus.

Succinate accumulation in the cytosol

SDH inactivation can block the oxidation of succinate, starting a cascade of reactions:

1. The succinate accumulated in the mitochondrial matrix diffuses through the inner and outer mitochondrial membranes to the cytosol (purple dashed arrows in Figure 2).
2. Under normal cellular function, HIF1-α in the cytosol is quickly hydroxylated by prolyl hydroxylase (PHD), shown with the light blue arrow. This process is blocked by the accumulated succinate.
3. HIF1-α stabilizes and passes to the cell nucleus (orange arrow) where it combines with HIF1-β to form an active HIF complex that induces the expression of tumor causing genes.

This pathway raises the possibility of a therapeutic treatment. The build-up of succinate inhibits PHD activity. PHD action normally requires oxygen and alpha-ketoglutarate as cosubstrates and ferrous iron and ascorbate as cofactors. Succinate competes with α-ketoglutarate in binding to the PHD enzyme. Therefore, increasing α-ketoglutarate levels can offset the effect of succinate accumulation.

Normal α-ketoglutarate does not permeate cell walls efficiently, and it is necessary to create a cell permeating derivative (e.g. α-ketoglutarate esters). In-vitro trials show this supplementation approach can reduce HIF1-α levels, and may result in a therapeutic approach to tumours resulting from SDH deficiency.

Impaired developmental apoptosis

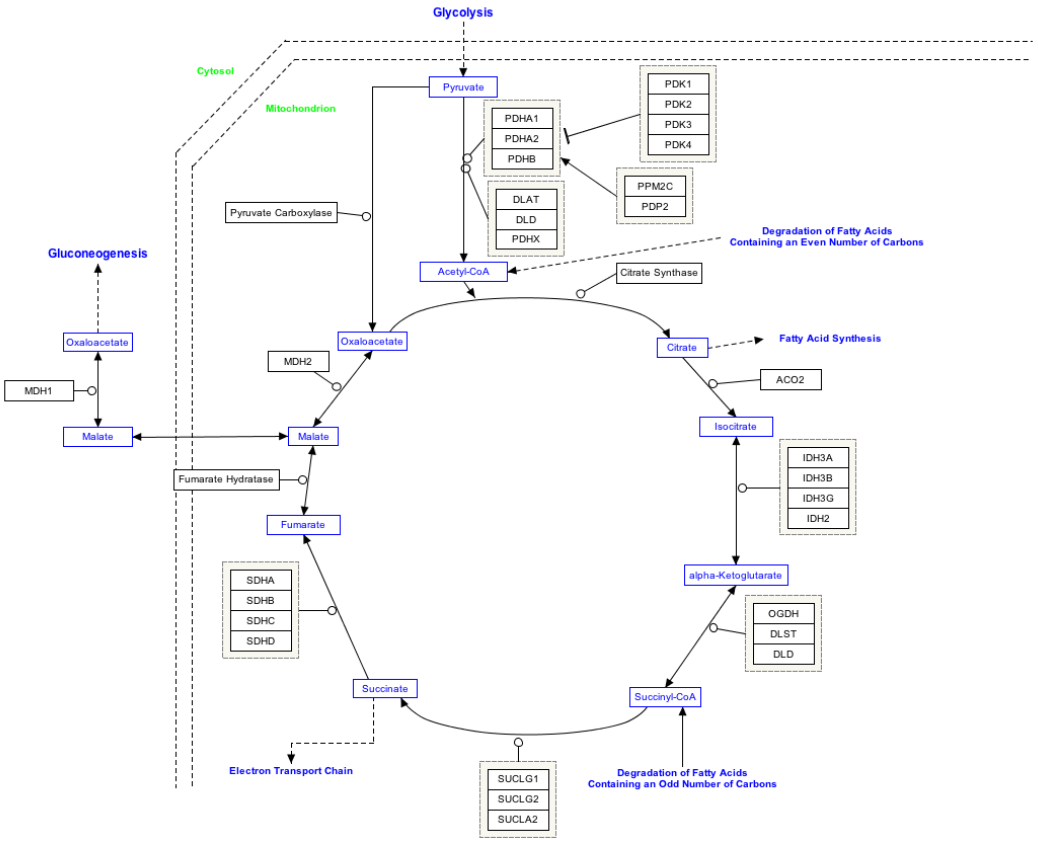
Paraganglionic tissue is derived from the neural crest cells present in an embryo. Abdominal extra-adrenal paraganglionic cells secrete catecholamines that play an important role in fetal development. After birth these cells usually die, a process that is triggered by a decline in nerve growth factor (NGF)which initiates apoptosis (cell death).

This cell death process is mediated by an enzyme called prolyl hydroxylase EglN3. Succinate accumulation caused by SDH inactivation inhibits the prolyl hydroxylase EglN3. The net result is that paranglionic tissue that would normally die after birth remains, and this tissue may be able to trigger paraganglioma/pheochromocytoma later.

Glycolysis upregulation

Inhibition of the Citric Acid Cycle forces the cell to create ATP glycolytically in order to generate its required energy. The induced glycolytic enzymes could potentially block cell apoptosis.

Interactive pathway map



Citric acid cycle edit