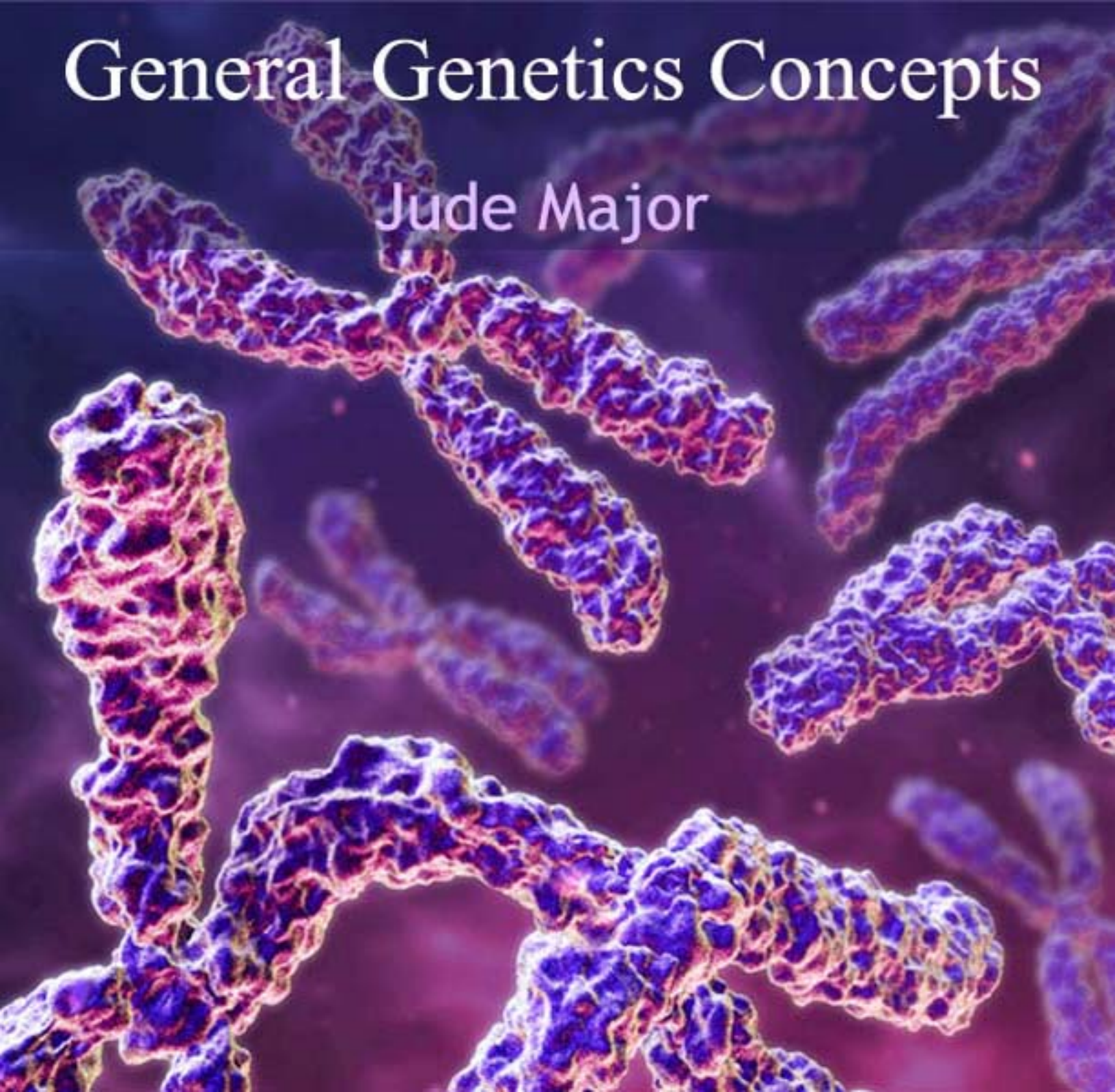
General Genetics Concepts

Jude Major



First Edition, 2012

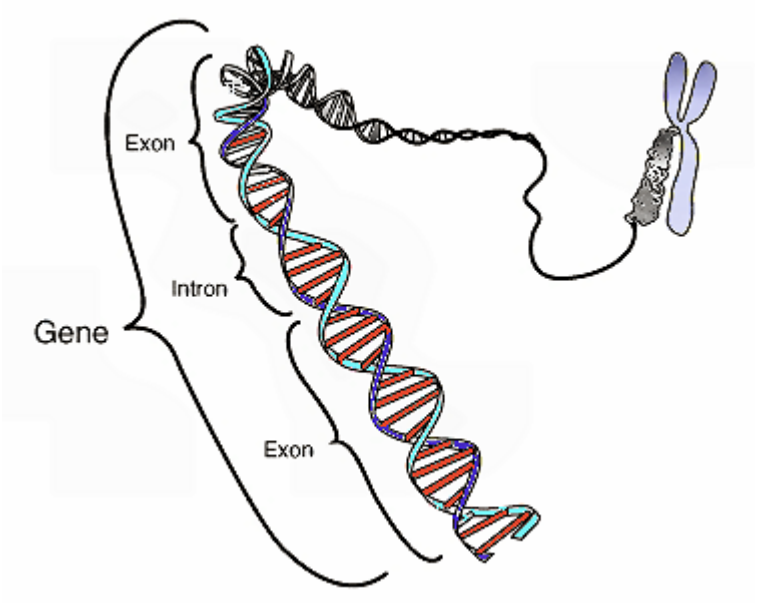
ISBN 978-81-323-3284-8

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

Table of Contents

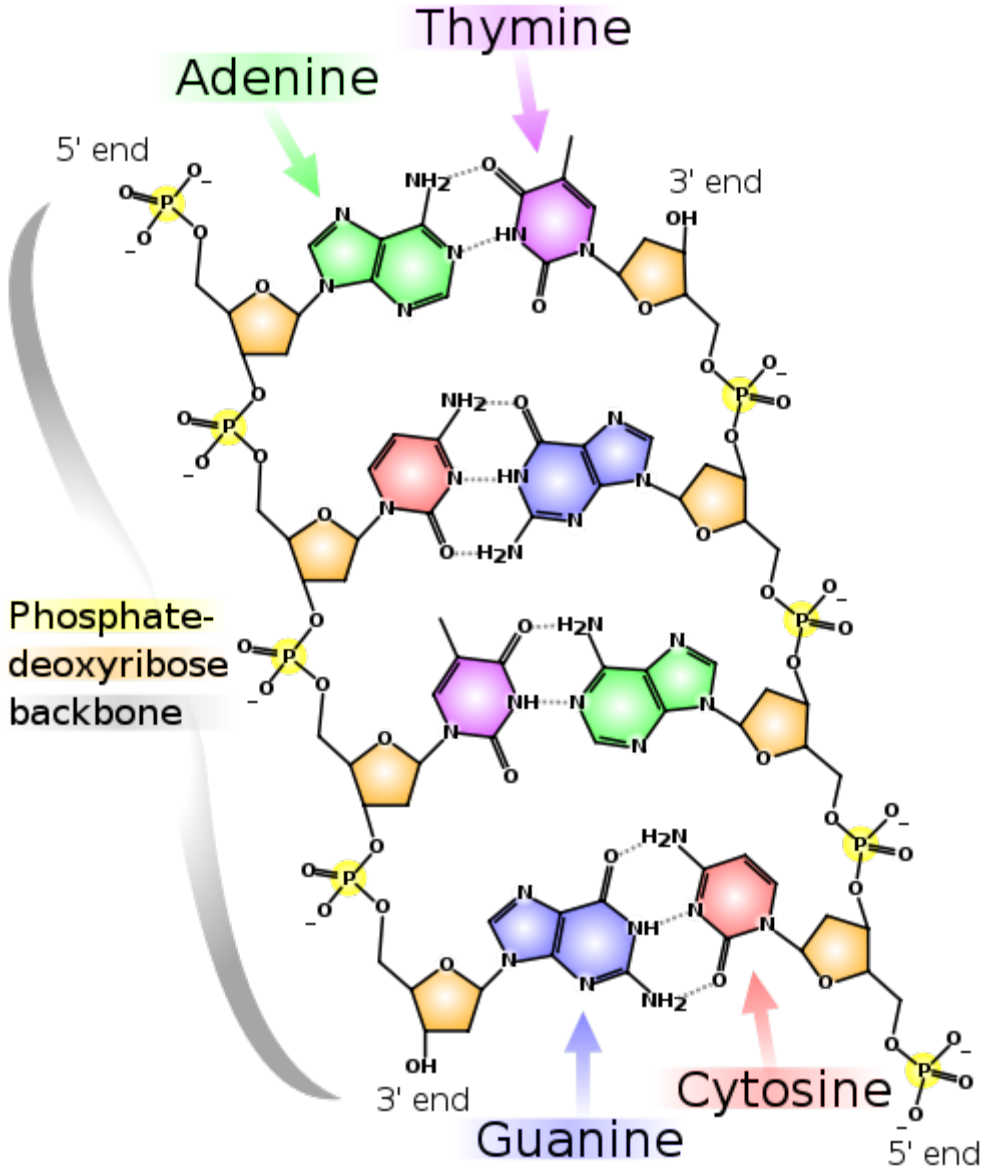
Chapter- 1

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 a stretch of DNA that codes for a type of protein or for an RNA chain that has a function in the organism. All 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 acid (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

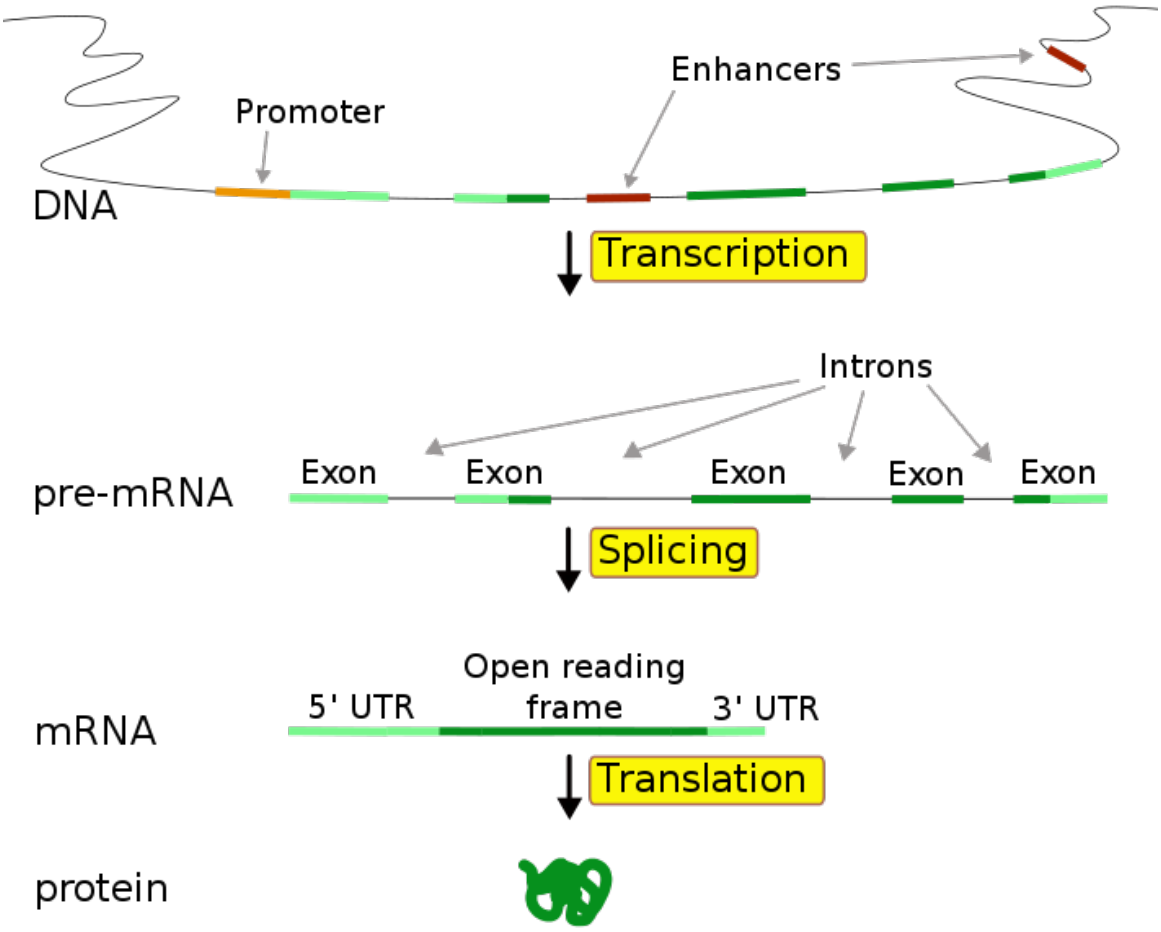


Diagram of the "typical" eukaryotic protein-coding **gene**. Promoters and enhancers determine what portions of the DNA will be transcribed into the precursor mRNA (pre-mRNA). The pre-mRNA is then spliced into messenger RNA (mRNA) which is later translated into protein.

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

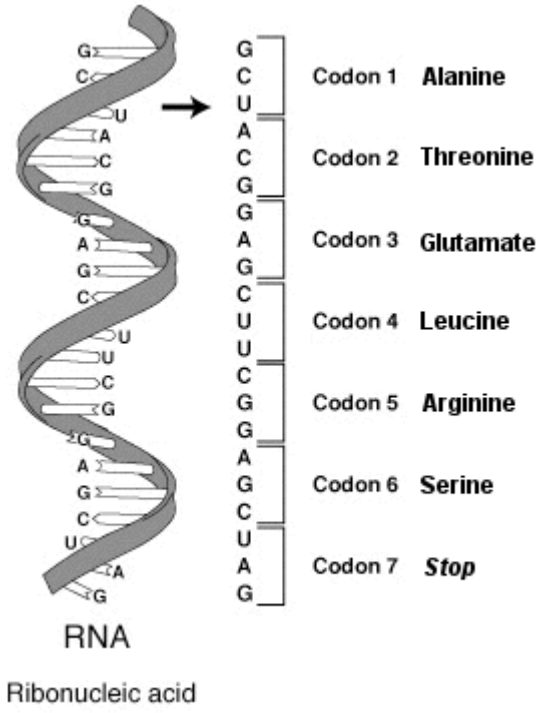
DNA helix on which thousands of genes are encoded. Prokaryotes—bacteria and archaea—typically store their genomes on a single large, circular chromosome, sometimes supplemented by additional small circles of DNA called plasmids, which usually encode only a few genes and are easily transferable between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

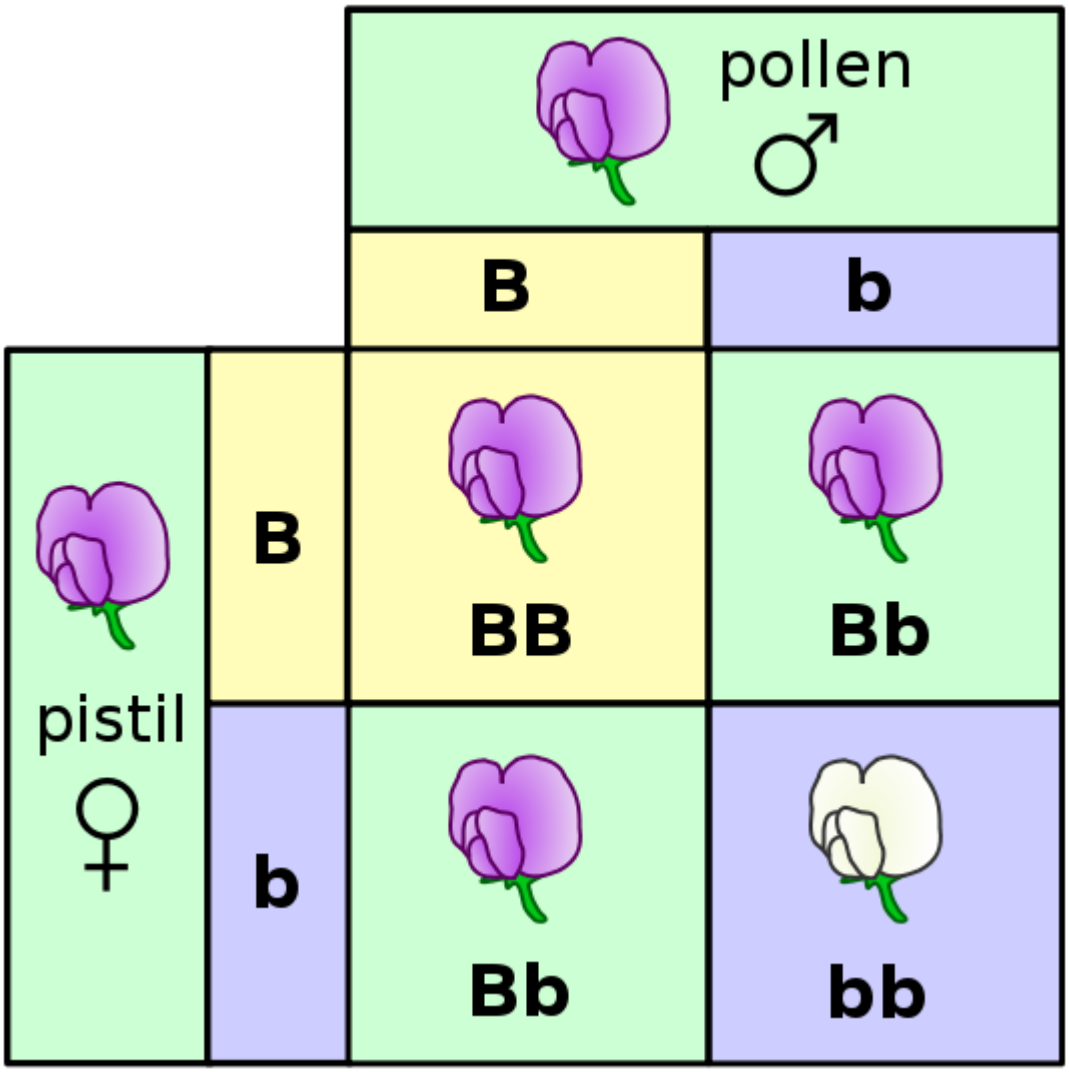
A series of subsequent discoveries led to the realization decades later that chromosomes within cells are the carriers of genetic material, and that they are made of DNA (deoxyribonucleic acid), a polymeric molecule found in all cells on which the 'discrete units' of Mendelian inheritance are encoded.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

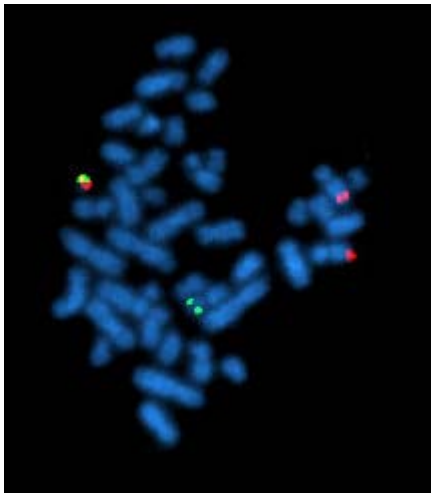
The definition of a gene is still changing. The first cases of RNA-based inheritance have been discovered in mammals. Evidence is also accumulating that the control regions of a gene do not necessarily have to be close to the coding sequence on the linear molecule or even on the same chromosome. Spilianakis 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

Cytogenetics



A metaphase cell positive for the bcr/abl rearrangement using FISH

Cytogenetics is a branch of genetics that is concerned with the study of the structure and function of the cell, especially the chromosomes. It includes routine analysis of G-Banded chromosomes, other cytogenetic banding techniques, as well as molecular cytogenetics such as fluorescent in situ hybridization (FISH) and comparative genomic hybridization (CGH).

History

Early years

Chromosomes were first observed in plant cells by Karl Wilhelm von Nägeli in 1842. Their behavior in animal (salamander) cells was described by Walther Flemming, the discoverer of mitosis, in 1882. The name was coined by another German anatomist, von Waldeyer in 1888.

The next stage took place after the development of genetics in the early 20th century, when it was appreciated that the set of chromosomes (the karyotype) was the carrier of

the genes. Levitsky seems to have been the first to define the karyotype as the phenotypic appearance of the somatic chromosomes, in contrast to their genic contents. Investigation into the human karyotype took many years to settle the most basic question: how many chromosomes does a normal diploid human cell contain? In 1912, Hans von Winiwarter reported 47 chromosomes in spermatogonia and 48 in oogonia, concluding an XX/XO sex determination mechanism. Painter in 1922 was not certain whether the diploid number of man was 46 or 48, at first favoring 46. He revised his opinion later from 46 to 48, and he correctly insisted on man having an XX/XY system. Considering their techniques, these results were quite remarkable.

New techniques were needed to definitively solve the problem:

1. Using cells in culture
2. Pre-treating cells in a hypotonic solution, which swells them and spreads the chromosomes
3. Arresting mitosis in metaphase by a solution of colchicine
4. Squashing the preparation on the slide forcing the chromosomes into a single plane
5. Cutting up a photomicrograph and arranging the result into an indisputable karyogram.

It took until the mid 1950s until it became generally accepted that the karyotype of man included only 46 chromosomes. Rather interestingly, the great apes have 48 chromosomes. Human chromosome 2 was formed by a merger of ancestral chromosomes, reducing the number.

Applications in biology

McClintock's work on maize

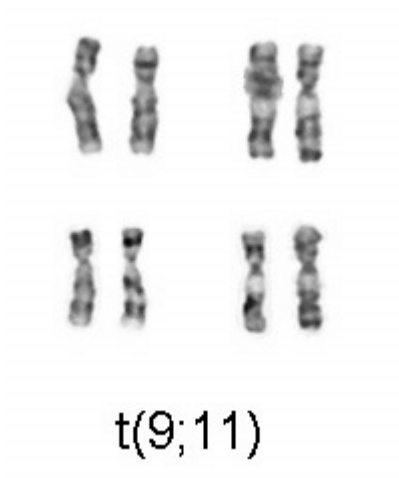
Barbara McClintock began her career as a maize cytogeneticist. In 1931, McClintock and Harriet Creighton demonstrated that cytological recombination of marked chromosomes correlated with recombination of genetic traits (genes). McClintock continued her career in cytogenetics studying the mechanics and inheritance of broken and ring (circular) chromosomes of maize. During her cytogenetic work, McClintock discovered transposons, a find which eventually led to her Nobel Prize in 1983.

Natural populations of Drosophila

In the 1930s, Dobzhansky and his co-workers collected *Drosophila pseudoobscura* and *D. persimilis* from wild populations in California and neighboring states. Using Painter's technique they studied the polytene chromosomes and discovered that the wild populations were polymorphic for chromosomal inversions. All the flies look alike whatever inversions they carry: this is an example of a cryptic polymorphism.

Evidence rapidly accumulated to show that natural selection was responsible. Using a method invented by L'Heretier and Teissier, Dobzhansky bred populations in *population cages*, which enabled feeding, breeding and sampling whilst preventing escape. This had the benefit of eliminating migration as a possible explanation of the results. Stocks containing inversions at a known initial frequency can be maintained in controlled conditions. It was found that the various chromosome types do not fluctuate at random, as they would if selectively neutral, but adjust to certain frequencies at which they become stabilised. By the time Dobzhansky published the third edition of his book in 1951 he was persuaded that the chromosome morphs were being maintained in the population by the selective advantage of the heterozygotes, as with most polymorphisms.

Human abnormalities and medical applications



translocation 9;11 associated with AML

In the event of procedures which allowed easy enumeration of chromosomes, discoveries were quickly made related to aberrant chromosomes or chromosome number. In some congenital disorders, such as Down's syndrome, cytogenetics revealed the nature of the chromosomal defect: a "simple" trisomy. Abnormalities arising from nondysjunction events can cause cells with aneuploidy (additions or deletions of entire chromosomes) in one of the parents or in the fetus. In 1959, Lejeune discovered patients with Down syndrome had an extra copy of chromosome 21. Down syndrome is also referred to as trisomy 21.

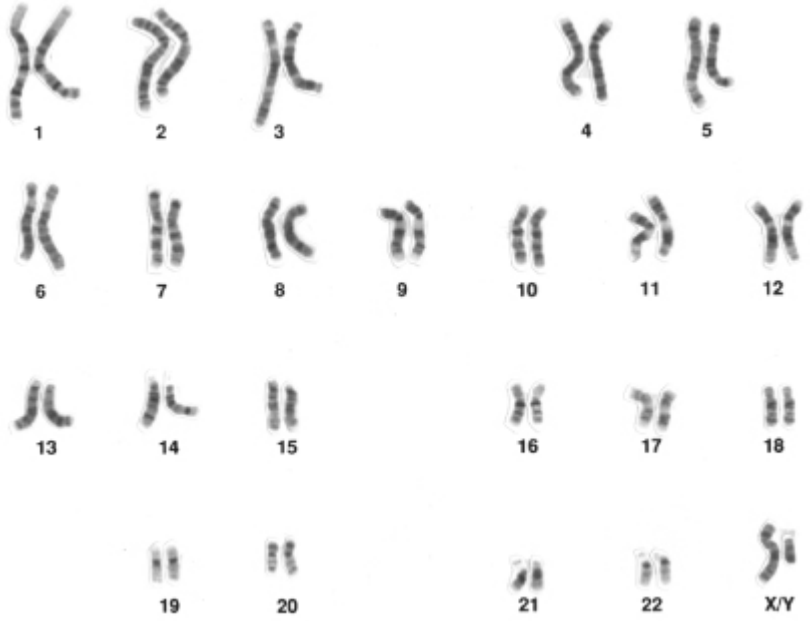
Other numerical abnormalities discovered include sex chromosome abnormalities. An individual with only one sex chromosome (the X) has Turner syndrome, an additional X chromosome in a male, resulting in 47 total chromosomes, has Klinefelter's Syndrome. Many other sex chromosome combinations are compatible with live birth including XXX, XYY, and XXXX. The ability for mammals to tolerate aneuploidies in the sex chromosomes arises from the ability to inactivate them, which is required in normal females to compensate for having two copies of the chromosome. Not all genes on the X

Chromosome are inactivated, which is why there is a phenotypic effect seen in individuals with extra X chromosomes.

Trisomy 13 was associated with Patau's Syndrome and trisomy 18 with Edward's Syndrome.

In 1960, Peter Nowell and David Hungerford discovered a small chromosome in the white blood cells of patients with Chronic myelogenous leukemia (CML). This abnormal chromosome was dubbed the Philadelphia chromosome - as both scientists were doing their research in Philadelphia, Pennsylvania. Thirteen years later, with the development of more advanced techniques, the abnormal chromosome was shown by Janet Rowley to be the result of a translocation of chromosomes 9 and 22. Identification of the Philadelphia chromosome by cytogenetics, in addition to other tests, is used today as a diagnostic for CML.

Advent of banding techniques



Human male karyotype

In the late 1960s, Caspersson developed banding techniques which differentially stain chromosomes. This allows chromosomes of otherwise equal size to be differentiated as well as to elucidate the breakpoints and constituent chromosomes involved in chromosome translocations. Deletions within one chromosome could also now be more specifically named and understood. Deletion syndromes such as DiGeorge syndrome, Prader-Willi syndrome and others were discovered to be caused by deletions in chromosome material.

Diagrams identifying the chromosomes based on the banding patterns are known as **cytogenetic maps**. These maps became the basis for both prenatal and oncological fields to quickly move cytogenetics into the clinical lab where karyotyping allowed scientists to look for chromosomal alterations. Techniques were expanded to allow for culture of free amniocytes recovered from amniotic fluid, and elongation techniques for all culture types that ruins everything.

Beginnings of molecular cytogenetics

In the 1980s, advances were made in molecular cytogenetics. While radioisotope-labeled probes had been hybridized with DNA since 1969, movement was now made in using fluorescent labeled probes. Hybridizing them to chromosomal preparations using existing techniques came to be known as *fluorescent in situ hybridization* (FISH). This change significantly increased the usage of probing techniques as fluorescent labeled probes are safer and can be used almost indefinitely. Further advances in micromanipulation and examination of chromosomes led to the technique of chromosome microdissection whereby aberrations in chromosomal structure could be isolated, cloned and studied in ever greater detail.

Techniques

Routine analysis

Routine chromosome analysis refers to analysis of metaphase chromosomes which have been banded using trypsin followed by Giemsa, Leishmanns, or a mixture of the two. This creates unique banding patterns on the chromosomes. The molecular mechanism and reason for these patterns is unknown, although it likely related to replication timing and chromatin packing.

Several chromosome-banding techniques are used in cytogenetics laboratories. Quinacrine banding (Q-banding) was the first staining method used to produce specific banding patterns. This method requires a fluorescence microscope and is no longer as widely used as Giemsa banding (G-banding). Reverse banding (R-banding) requires heat treatment and reverses the usual white and black pattern that is seen in G-bands and Q-bands. This method is particularly helpful for staining the distal ends of chromosomes. Other staining techniques include C-banding and nucleolar organizing region stains (NOR stains). These latter methods specifically stain certain portions of the chromosome. C-banding stains the constitutive heterochromatin, which usually lies near the centromere, and NOR staining highlights the satellites and stalks of acrocentric chromosomes. High-resolution banding involves the staining of chromosomes during prophase or early metaphase (prometaphase), before they reach maximal condensation. Because prophase and prometaphase chromosomes are more extended than metaphase chromosomes, the number of bands observable for all chromosomes increases from about 300 to 450 to as many as 800. This allows the detection of less obvious abnormalities usually not seen with conventional banding.

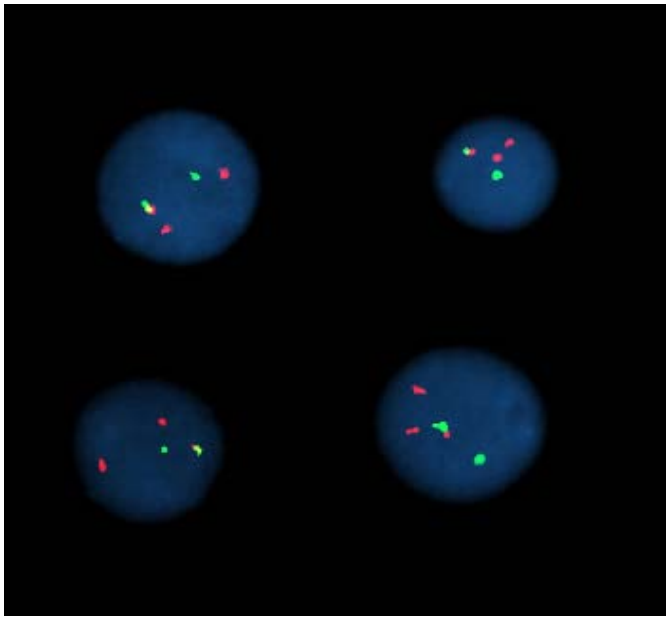
Slide preparation

Cells from bone marrow, blood, amniotic fluid, cord blood, tumor, and tissues (including skin, umbilical cord, liver, and many other organs) can be cultured using standard cell culture techniques in order to increase their number. A mitotic inhibitor (colchicine, colcemid) is then added to the culture. This stops cell division at mitosis which allows an increased yield of mitotic cells for analysis. The cells are then centrifuged and media and mitotic inhibitor is removed, and replaced with a hypotonic solution. This causes the cells to swell so that the chromosomes will spread when added to a slide. After the cells have been allowed to sit in hypotonic, Carnoy's fixative (3:1 methanol to glacial acetic acid) is added. This kills the cells, lyses the red blood cells, and hardens the nuclei of the remaining white blood cells. The cells are generally fixed repeatedly to remove any debris or remaining red blood cells. The cell suspension is then dropped onto specimen slides. After aging the slides in an oven or waiting a few days they are ready for banding and analysis.

Analysis

Analysis of banded chromosomes is done at a microscope by a clinical laboratory specialist in cytogenetics (CLSp(CG)). Generally 20 cells are analyzed which is enough to rule out mosaicism to an acceptable level. The results are summarized and given to a board-certified medical geneticist and a pathologist for review, and to write an interpretation taking into account the patients previous history and other clinical findings. The results are then given out reported in an *International System for Human Cytogenetic Nomenclature 2005* (ISCN2005).

Fluorescent in situ hybridization



Interphase cells positive for a t(9;22) rearrangement

Fluorescent in situ hybridization refers to using fluorescently labeled probe to hybridize to cytogenetic cell preparations.

In addition to standard preparations FISH can also be performed on:

- bone marrow smears
- blood smears
- paraffin embedded tissue preparations
- enzymatically dissociated tissue samples
- uncultured bone marrow
- uncultured amniocytes
- cytospin preparations

Slide preparation

The slide is aged using a salt solution usually consisting of 2X SSC (salt, sodium citrate). The slides are then dehydrated in ethanol, and the probe mixture is added. The sample DNA and the probe DNA are then co-denatured using a heated plate and allowed to re-anneal for at least 4 hours. The slides are then washed to remove excess unbound probe, and counterstained with 4',6-Diamidino-2-phenylindole (DAPI) or propidium iodide.

Analysis

Analysis of FISH specimens is done by fluorescence microscopy by a clinical laboratory specialist in cytogenetics. For oncology generally a large number of interphase cells are scored in order to rule out low level residual disease, generally between 200 and 1000 cells are counted and scored. For congenital problems usually 20 metaphase cells are scored.

Future of cytogenetics

Advances now focus on molecular cytogenetics including automated systems for counting the results of standard FISH preparations and techniques for virtual karyotyping, such as comparative genomic hybridization arrays, CGH and Single nucleotide polymorphism-arrays.

Chapter- 3

Dominance

Dominance in genetics is a relationship between two variant forms (alleles) of a single gene, in which one allele masks the expression of the other in influencing some trait. In the simplest case, if a gene exists in two allelic forms (**A** & **B**), three combinations of alleles (genotypes) are possible: **AA**, **AB**, and **BB**. If **AB** individuals (heterozygotes) show the same form of the trait (phenotype) as **AA** individuals (homozygotes), and **BB** homozygotes show an alternative phenotype, allele **A** is said to *dominate* or *be dominant to* allele **B**, and **B** is said to *be recessive to* **A**.

By convention, dominant alleles are written in uppercase letters, and recessive alleles in lowercase letters. In this example, allele ***B*** is replaced by ***a***. Then, **A** is *dominant* to **a** (and **a** is *recessive* to **A**), the **AA** and **Aa** genotypes have the same phenotype, and the **aa** genotype has a different phenotype.

Background: diploid, chromosomes, genes, loci, & alleles

Diploid / haploid

Most familiar plants, like peas, and familiar animals, like fruit flies and humans, have paired chromosomes, and are described as diploid. One chromosome of each pair is contributed by each parent, one by the female parent in her ova, and one by the male parent in his sperm, which are joined at fertilization. The ova and sperm cells have only one copy of each chromosome and are described as (haploid). Production of haploid gametes occurs through a cell division process called meiosis.

Chromosomes, genes, and alleles

Each chromosome of a matching pair is structurally similar to the other, and each member of a homologous pair has the same genetic material arranged in the same order and physical locations (loci, sing. locus). The genetic material in each chromosome comprises a series of discrete genes that influence various traits. Thus, each gene also has a corresponding homologue, which may exist in different forms: the variant forms are

called alleles. The alleles at the same locus on the two homologous chromosomes may be identical or different.

In popular usage, "gene" and "allele" are often used interchangeably. This produces misunderstandings. Properly, 'gene' refers to a hereditary unit, ordinarily at a fixed position on a chromosome, that influences a particular trait. Genes are now understood to comprise DNA. 'Allele' refers to any of the many particular forms of a gene that any be present in an individual. *E.g.*, it is inaccurate to say "This pea plant has a pair of wrinkled genes", and it is more accurate to say, "This plant has two 'w' alleles for the 'Seed Shape' gene, and will produce wrinkled peas."

Homozygous, heterozygous

If two alleles of a given gene are identical, the organism is called a homozygote and is homozygous with respect to that gene; if instead the two alleles are different, the organism is a heterozygote and is heterozygous. The genetic makeup of an organism, either at a single locus or over all its genes collectively, is called the genotype. The genotype of an organism directly or indirectly affects its molecular, physical, behavioral, and other traits, which individually or collectively are called the phenotype. At heterozygous gene loci, the two alleles interact to produce the phenotype. The simplest form of allele interaction is the one described by Mendel, now called Mendelian, in which the appearance/phenotype caused by one allele is apparent, called dominant, and the appearance/phenotype caused by the other allele is not apparent, called recessive.

In the simplest case, the phenotypic effect of one allele completely masks the other in heterozygous combination; that is, the phenotype produced by the two alleles in heterozygous combination is identical to that produced by one of the two homozygous genotypes. The allele that masks the other is said to be *dominant* to the latter, and the alternative allele is said to be *recessive* to the former.

Which trait is *dominant*?

Dominant and recessive alleles do not necessarily produce 'stronger' and 'weaker' phenotypes, respectively. Rather, the terms simply refer to their interaction in producing the phenotype of the heterozygote. If there are two alternative phenotypes, by definition the phenotype exhibited by the heterozygote is called "dominant" and the "hidden" phenotype is called "recessive." The key concept of dominance is that the heterozygote is phenotypically *identical* to one of the two homozygotes. That trait corresponding to the dominant allele may then be called the 'dominant' trait.

It is critical to understand that dominance is a *genotypic* relationship between *alleles*, as manifested in the phenotype. It is unrelated to the nature of the phenotype itself, e.g., whether it is regarded as 'normal or abnormal,' 'standard or nonstandard,' 'healthy or diseased,' 'stronger or weaker,' or 'more or less' extreme. It is also important to distinguish between the 'round' gene locus, the 'round' allele at that locus, and the 'round' phenotype it

produces. It is inaccurate to say that 'the round gene dominates the wrinkled gene' or that 'round peas dominate wrinkled peas.'

Nomenclature

In genetics, the common convention is that dominant alleles are written as capital letters and recessive alleles as lower-case letters. In the pea example, once the dominance relationships of the two alleles are known, it is possible to designate the dominant allele that produces a round shape by a capital-letter symbol **R**, and the alternative recessive allele that produces a wrinkled shape by a lower-case symbol **r**. The homozygous dominant, heterozygous, and homozygous recessive genotypes are then written **RR**, **Rr**, and **rr**, respectively. It would also possible to designate the two alleles as **W** and **w**, and three genotypes **WW**, **Ww**, and **ww**, the first two of which produced round peas and the third wrinkled peas. Note that the choice or "**R**" or "**W**" as the symbol for the dominant allele does not pre-judge whether the allele causing the 'round' or 'wrinkled' phenotype when homozygous is the dominant one.

Another system of notation designates the gene involved in seed shape as the "*Shp*" gene, which exists in two allelic forms, Shp^R and Shp^w, the dominance relationships of the two being indicated by the case of the superscripts. This system is the standard system in *Drosophila* genetics.

Relationship to other genetic concepts

The concept of dominance is involved with a number of other genetic concepts.

Multiple alleles

Although any individual has at most two different alleles, most genes exist in a large number of allelic forms in the population as a whole. In some cases, the alleles have different effects on the phenotype, and their dominance interactions with each other can be described as a series. For example, the best known human blood groups, the ABO system, comprises three sets of alleles at the $\boldsymbol{I}$ locus, $\boldsymbol{I^A}$, $\boldsymbol{I^B}$, and $\boldsymbol{I^O}$. The first two are dominant to the latter: that is, the **AA** and **AO** genotypes produce indistinguishable blood group phenotypes, called "*Type A*", as do **BB** and **BO**, which produce "*Type B*" blood. In another example, coat color in siamese cats and related breeds is determined by a series of alleles at the albino gene locus ($\boldsymbol{c}$) that produce different levels of pigment and hence different levels of color dilution. Four of these are $\boldsymbol{c^+}$, $\boldsymbol{c^b}$, $\boldsymbol{c^s}$, and $\boldsymbol{c^a}$ (standard, Burmese, siamese, and albino, respectively), where the first allele is completely dominant to the last three, and the last is completely recessive to the first three.

Incomplete and semi-dominance

Complete dominance occurs when the phenotype of the heterozygote is completely indistinguishable from that of the dominant homozygote. This is frequently not the case.

Incomplete dominance occurs when the phenotype of the heterozygous genotype is an intermediate of the phenotypes of the homozygous genotypes. For example, the snapdragon flower color is either homozygous for red or white. When the red homozygous flower is paired with the white homozygous flower, the result yields a pink snapdragon flower. The pink snapdragon is the result of incomplete dominance.

Co-dominance

Co-dominance occurs when the contributions of both alleles are visible in the phenotype. In the **ABO** example, the $\boldsymbol{I^A}$ and $\boldsymbol{I^B}$ alleles are co-dominant in producing the **AB** blood group phenotype, in which both **A**- and **B**-type antigens are made. Another example occurs at the locus for the Beta-globin component of hemoglobin, where the three molecular phenotypes of $\mathbf{Hb^A/Hb^A}$, $\mathbf{Hb^A/Hb^S}$, and $\mathbf{Hb^S/Hb^S}$ are all equally detectable by protein electrophoresis. (The medical condition produced by the heterozygous genotype is called an *incomplete dominant*, see above). For most gene loci at the molecular level, both alleles are expressed co-dominantly, because both are transcribed into RNA.

Co-dominance and incomplete or semi-dominance are not the same phenomenon. For example, pink flowers might be the product of two alleles that produce red and white pigments that become mixed (co-dominance on the pigment level, no dominance on the color level), or the result of one allele that produces the usual amount of red pigment and another non-functional allele that produces no pigment, so as to produce a dilute, intermediate pink color (no dominance at either level).

Autosomal *versus* sex-linked dominance

In humans and other mammal species, sex is determined by two sex chromosomes called the X chromosome and the Y chromosome. Human females are typically **XX**, males are typically **XY**. The remaining pairs of chromosome are found in both sexes and are called autosomes; genetic traits due to loci on these chromosomes are described as autosomal, and may be dominant or recessive. Genetic traits on the **X** and **Y** chromosomes are called sex linked, because they tend to be characteristic of one sex or the other. In practice, the term almost always refers to **X**-linked traits. Females have two copies of every gene locus found on the **X** chromosome, just as for the autosomes, and the same dominance relationships apply. Males however have only one copy of each **X**-chromosome gene locus, and are described as hemizygous for these genes. The **Y** chromosome is much smaller than the **X**, and contains a much smaller set of genes that influence 'maleness', such as the **SRY** gene for testis determining factor. Dominance rules for sex-linked gene loci are determined by their behavior in the female: because the male has only one allele, that allele is always expressed regardless of whether it is dominant or recessive.

Epistasis

Epistasis ["*epi* + *stasis* = to sit on top"] is an interaction between genotypes at two *different* gene loci, which sometimes resembles a dominance interaction at a single locus. Epistasis modifies the characteristic 9:3:3:1 ratio expected for two non-epistatic genes.

Most genetic systems involve complex epistatic interactions among multiple gene loci. For two loci, 14 classes of epistatic interactions are recognized. As an example of ***recessive epistasis***, one gene locus may determine whether a flower pigment is yellow (**AA** or **Aa**) or green (**aa**), while another locus determines whether the pigment is produced (**BB** or **Bb**) or not (**bb**). In a **bb** plant, the flowers will be white, irrespective of the genotype of the other locus as **AA**, **Aa**, or **aa**. The **b** allele is *not* dominant to the **A** allele: the **B** locus shows ***recessive epistasis*** to the **A** locus, because the **B** locus when homozygous for the recessive allele (**bb**) suppresses phenotypic expression of the **A** locus. In a cross between two **AaBb** plants, this produces a characteristic **9:3:4** ratio, in this case of yellow : green : white flowers.

In ***dominant epistasis***, one gene locus may determine yellow and green pigment as in the previous example: **AA** and **Aa** are yellow, and **aa** are green. A second locus determines whether a pigment precursor is produced (**dd**) or not (**DD** or **Dd**). Here, in a **D-** plant, the flowers will be colorless irrespective of the genotype at the ***A*** locus, because of the epistatic effect of the dominant **D** allele. Thus, in a cross between two **AaDd** plants, 3/4 of the plants will be colorless, and the yellow and green phenotypes are expressed only in **dd** plants. This produces a characteristic **12:3:1** ratio of white : yellow : green plants.

Supplementary epistasis occurs when two loci affect the same phenotype. For example, if pigment color is produced by **CC** or **Cc** but not **cc**, and by **DD** or **Dd** but not **dd**, then pigment is produced only in **C-D-** genotypes, and not in any genotype combination with **cc** *or* **dd**. That is, *both* loci must have at least one dominant allele to produce the phenotype. This produces a characteristic ratio **9:7** ratio of unpigmented to pigmented plants.

Molecular mechanisms

The molecular basis of dominance was unknown to Mendel. It is now understood that a gene locus includes a long series (hundreds to thousands) of bases or nucleotides of deoxyribonucleic acid (DNA) at a particular point on a chromosome. The central dogma of molecular biology states that "*DNA makes RNA makes protein*", that is, that DNA is transcribed to make an RNA copy, and RNA is translated to make a protein. In this process, different alleles at a locus may or may not be transcribed, and if transcribed may be translated to slightly different forms of the same protein (called isoforms). Proteins often function as enzymes that catalyze chemical reactions in the cell, which directly or indirectly produce phenotypes. In any diploid organism, the DNA sequences of the two alleles present at any gene locus may be identical (homozygous) or different (heterozygous). Even if the gene locus is heterozygous at the level of the DNA sequence, the proteins made by each allele may be identical. In the absence of any difference between the protein products, neither allele can be said to be dominant (see co-dominance*, below). Even if the two protein products are slightly different (allozymes), it is likely that they produce the same phenotype with respect to enzyme action, and again neither allele can be said to be dominant.*

Dominance typically occurs when one of the two alleles is non-functional at the molecular level, that is, it is not transcribed or else does not produce a protein product. This can be the result of a mutation that alters the DNA sequence of the allele. An organism homozygous for the non-functional allele will generally show a distinctive phenotype, due to the absence of the protein product. For example, in humans and other organisms, the unpigmented skin of the albino phenotype results when an individual is homozygous for an allele that prevents synthesis of the skin pigment protein melanin. It is important to understand that it is not the lack of function that allows the allele to be described as recessive: this is the interaction with the alternative allele in the heterozygote. Three general types of interaction are possible:

1. In the typical case, the single functional allele makes sufficient protein to produce a phenotype identical to that of the homozygote: this is called *haplosufficiency*. For example, suppose the standard amount of enzyme produced in the functional homozygote is 100%, with the two functional alleles contributing 50% each. The single functional allele in the heterozygote produces 50% of the standard amount of enzyme, which is sufficient to produce the standard phenotype. If the heterozygote and the functional-allele homozygote have identical phenotypes, the functional allele is dominant to the non-functional allele. This occurs at the albino gene locus: the heterozygote produces sufficient enzyme to convert the pigment precursor to melanin, and the individual has standard pigmentation.
2. Alternatively, a single functional allele in the heterozygote may produce insufficient gene product for proper function, and the phenotype resembles that of the homozygote for the non-functional allele. This *haploinsufficiency* is much less common: usually the deficiency of gene product results in *incomplete dominance* (below).
3. The intermediate interaction occurs where the heterozygous genotype produces a phenotype intermediate between the two homozygotes. Depending on which of the two homozygotes the heterozygote most resembles, one allele is said to show *incomplete dominance* over the other. For example, in humans the *Hb* gene locus is responsible for the Beta-chain protein (HBB) that is one of the two globin proteins that make up the blood pigment hemoglobin. Many people are homozygous for an allele called Hb^A; some persons carry an alternative allele called Hb^S, either as homozygotes or heterozygotes. The hemoglobin molecules of Hb^S/Hb^S homozygotes undergo a change in shape that distorts the morphology of the red blood cells, and causes a severe, life-threatening form of anemia called sickle-cell anemia. Persons heterozygous Hb^A/Hb^S for this allele have a much less severe form of anemia called sickle-cell trait. Because the disease phenotype of Hb^A/Hb^S heterozygotes is more similar to but not identical to the Hb^A/Hb^A homozygote, the Hb^A allele is said to be *incompletely dominant* to the Hb^S allele.

In some cases, dominance of a non-standard allele results when that allele produces a defective protein that interferes with the proper function of the protein produced by the standard allele. The presence of the defective protein "dominates" the standard protein, and the disease phenotype of the heterozygote more closely resembles that of the homozygote for two variant alleles.

Dominant and recessive genetic diseases in humans

In humans, many genetic traits or diseases are classified simply as "dominant" or "recessive." Especially with respect to so-called recessive diseases, this can oversimplify the underlying molecular basis and lead to misunderstanding of the nature of dominance. For example, the genetic disease phenylketonuria (PKU) results from any of a large number (>60) of alleles at the gene locus for the enzyme phenylalanine hydroxylase (**PAH**). Many of these alleles produce little or no **PAH**, as a result of which the substrate phenylalanine and its metabolic byproducts accumulate in the central nervous system and can cause severe mental retardation if untreated.

The genotypes and phenotypic consequences of interactions among three alleles are shown in the following table:

Genotype	**PAH activity**	**[phe] conc**	**PKU ?**
AA	100%	60 uM	No
AB	30%	120 uM	No
CC	5%	200 ~ 300 uM	Hyperphenylalanemia
BB	0.3%	600 ~ 2400 uM	Yes

In unaffected persons homozygous for a standard functional allele (**AA**), **PAH** activity is standard (100%), and the concentration of phenylalanine in the blood [**phe**] is about 60 uM. In untreated persons homozygous for one of the PKU alleles (**BB**), **PAH** activity is close to zero, [phe] ten to forty times standard, and the individual manifests PKU.

In the **AB** heterozygote, **PAH** activity is only 30% (not 50%) of standard, blood [**phe**] is elevated two-fold, and the person does not manifest PKU. Thus, the **A** allele is dominant to the **B** allele with respect to PKU, but the **B** allele is incompletely dominant to the **A** allele with respect to its molecular effect, determination of **PAH** activity level (0.3% < 30% << 100%). Finally, the **A** allele is an incomplete dominant to **B** with respect to [phe], as 60 uM < 120 uM << 600 uM. Note once more that it is irrelevant to the question of dominance that the recessive allele produces a more extreme [phe] phenotype.

For a third allele **C**, a **CC** homozygote produces a very small amount of **PAH** enzyme, which results in a somewhat elevated level of [**phe**] in the blood, a condition called hyperphenylalanemia, which does not result in mental retardation.

That is, the dominance relationships of any two alleles may vary according to which aspect of the phenotype is under consideration. It is typically more useful to talk about the phenotypic consequences of the allelic interactions involved in any genotype, rather than to try to force them into dominant and recessive categories.

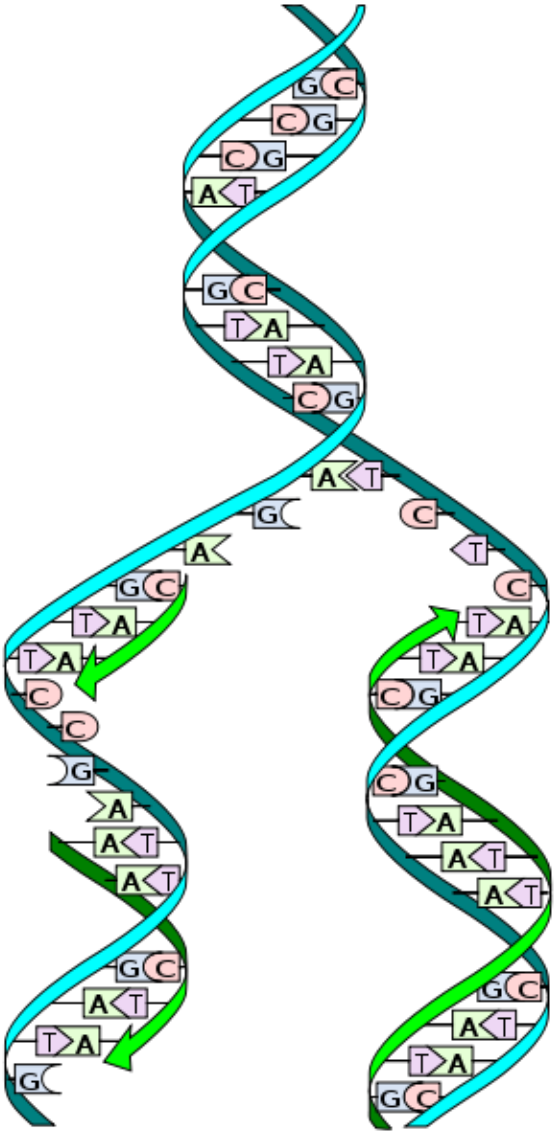
History

The concept of dominance was first described by the "Father of Genetics," Gregor Mendel, in the 1860s. Mendel observed that, for a variety of traits of garden peas having to do with the appearance of seeds, seed pods, and plant appearance, there occurred two discrete phenotypes: round *vs* wrinkled, or yellow *vs* green seeds, red *vs* white flowers, tall *vs* short plants, and so on. When bred separately, the plants always produced the same phenotypes, generation after generation. However, when lines with different phenotypes were crossed (interbred), one and only one of the parental phenotypes showed up in the offspring: green, or round, or red, or tall, and so on. However, when these hybrid plants were crossed, the offspring plants showed the two original phenotypes, in a characteristic **3:1** ratio, with the more common type having the phenotype of the parental hybrid plants. Mendel reasoned that each of the parents in the first cross were homozygotes for different alleles (**AA** and **aa**), that each contributed one allele to the offspring, such that all of these hybrids were heterozygotes (**Aa**), and that one of the two alleles in the hybrid cross ***dominated*** expression of the other: **A** masked **a**. The final cross between two heterozygotes (**Aa** X **Aa**) would produce **AA**, **Aa**, and **aa** offspring in a **1:2:1** *genotype* ratio with the first three classes showing the "**A**" phenotype, and the last showing the "**a**" phenotype, thereby producing the **3:1** *phenotype* ratio.

Mendel did not use the terms gene, allele, phenotype, genotype, homozygote, and heterozygote, all of which were introduced afterward. He did introduce the notation of capital and lowercase letters for dominant and recessive alleles, respectively, still in use today.

Chapter- 4

DNA Replication



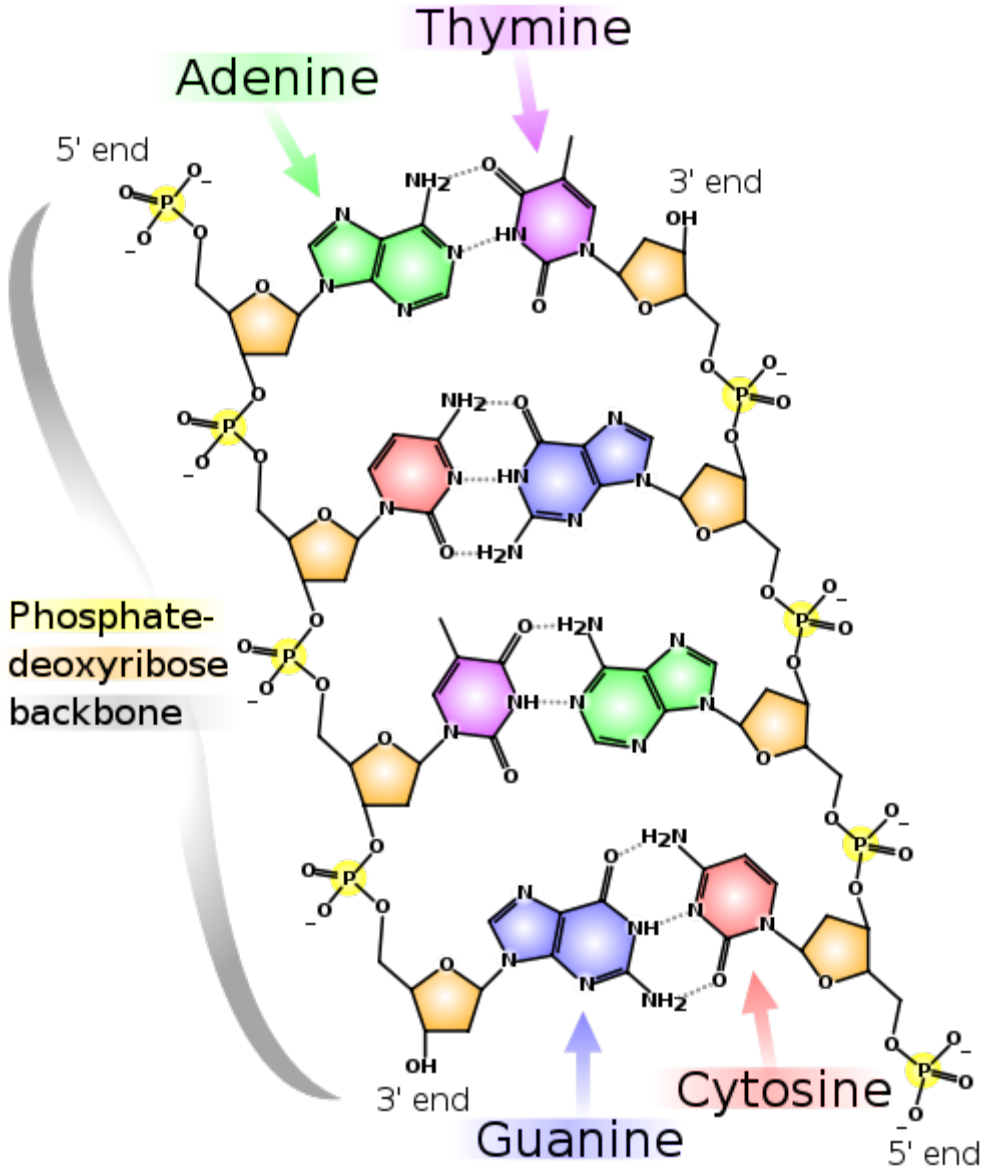
DNA replication. The double helix is unwound and each strand acts as a template. Bases are matched to synthesize the new partner strands.

DNA replication, the basis for biological inheritance, is a fundamental process - that occurs in all living organisms - that copies their DNA. This process is *replication*, in that each strand of the original double-stranded DNA molecule serves as template for the reproduction of the complementary strand. Therefore, following DNA replication, two identical DNA molecules have been produced from a single double-stranded DNA molecule. Cellular proofreading and error toe-checking mechanisms ensure near perfect fidelity for DNA replication.

In a cell, DNA replication begins at specific locations in the genome, called "origins". Unwinding of DNA at the origin, and synthesis of new strands, forms a replication fork. In addition to DNA polymerase, the enzyme that synthesizes the new DNA by adding nucleotides matched to the template strand, a number of other proteins are associated with the fork and assist in the initiation and continuation of DNA synthesis.

DNA replication can also be performed *in vitro* (outside a cell). DNA polymerases, isolated from cells, and artificial DNA primers are used to initiate DNA synthesis at known sequences in a template molecule. The polymerase chain reaction (PCR), a common laboratory technique, employs such artificial synthesis in a cyclic manner to amplify a specific target DNA fragment from a pool of DNA.

DNA structure



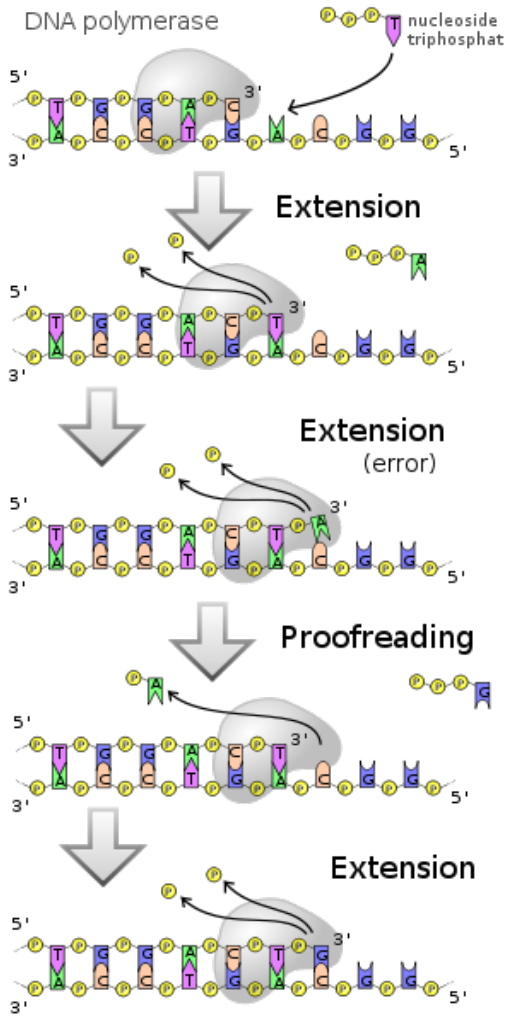
The chemical structure of DNA

DNA usually exists as a double-stranded structure, with both strands coiled together to form the characteristic double-helix. Each single strand of DNA is a chain of four types of nucleotides having the bases: adenine, cytosine, guanine, and thymine. A nucleotide is a mono-, di-, or triphosphate deoxyribonucleoside; that is, a deoxyribose sugar is attached to one, two, or three phosphates. Chemical interaction of these nucleotides forms phosphodiester linkages, creating the phosphate-deoxyribose backbone of the DNA double helix with the bases pointing inward. Nucleotides (bases) are matched between strands through hydrogen bonds to form base pairs. Adenine pairs with thymine and cytosine pairs with guanine.

DNA strands have a directionality, and the different ends of a single strand are called the "3' (three-prime) end" and the "5' (five-prime) end." These terms refer to the carbon atom in deoxyribose to which the next phosphate in the chain attaches. In addition to being complementary, the two strands of DNA are antiparallel: They are orientated in opposite directions. This directionality has consequences in DNA synthesis, because DNA polymerase can synthesize DNA in only one direction by adding nucleotides to the 3' end of a DNA strand.

The pairing of bases in DNA through hydrogen bonding means that the information contained within each strand is redundant. The nucleotides on a single strand can be used to reconstruct nucleotides on a newly synthesized partner strand.

DNA polymerase



DNA polymerases adds nucleotides to the 5' end of a strand of DNA. If a mismatch is accidentally incorporated, the polymerase is inhibited from further extension. Proofreading removes the mismatched nucleotide and extension continues.

DNA polymerases are a family of enzymes that carry out all forms of DNA replication. A DNA polymerase can only extend an existing DNA strand paired with a template strand; it cannot begin the synthesis of a new strand. To begin synthesis of a new strand, a short fragment of DNA or RNA, called a primer, must be created and paired with the template strand before DNA polymerase can synthesize new DNA.

Once a primer pairs with DNA to be replicated, DNA polymerase synthesizes a new strand of DNA by extending the 3' end of an existing nucleotide chain, adding new nucleotides matched to the template strand one at a time via the creation of phosphodiester bonds. The energy for this process of DNA polymerization comes from two of the three total phosphates attached to each unincorporated base. (Free bases with their attached phosphate groups are called nucleoside triphosphates.) When a nucleotide is being added to a growing DNA strand, two of the phosphates are removed and the energy produced creates a phosphodiester (chemical) bond that attaches the remaining phosphate to the growing chain. The energetics of this process also help explain the directionality of synthesis - if DNA were synthesized in the 3' to 5' direction, the energy for the process would come from the 5' end of the growing strand rather than from free nucleotides.

In general, DNA polymerases are extremely accurate, making less than one error for every 10^7 nucleotides added. Even so, some DNA polymerases also have proofreading ability; they can remove nucleotides from the end of a strand in order to correct mismatched bases. If the 5' nucleotide needs to be removed during proofreading, the triphosphate end is lost. Hence, the energy source that usually provides energy to add a new nucleotide is also lost.

DNA replication within the cell

Origins of replication

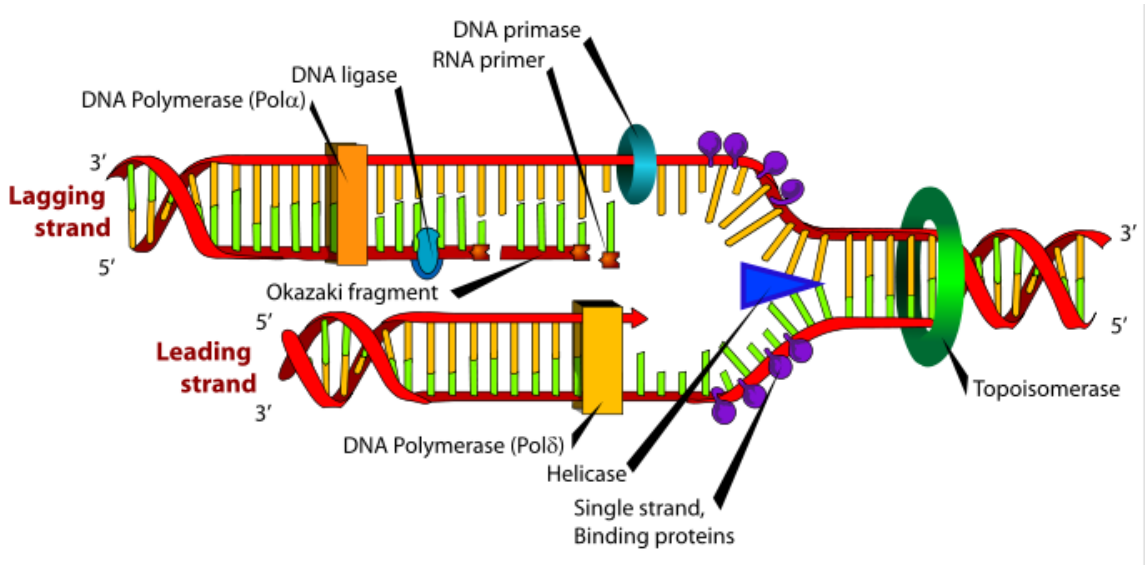
For a cell to divide, it must first replicate its DNA. This process is initiated at particular points within the DNA, known as "origins", which are targeted by proteins that separate the two strands and initiate DNA synthesis. Origins contain DNA sequences recognized by replication initiator proteins (e.g., dnaA in *E coli'* and the Origin Recognition Complex in yeast). These initiator proteins recruit other proteins to separate the two strands and initiate replication forks.

Initiator proteins recruit other proteins to separate the DNA strands at the origin, forming a bubble. Origins tend to be "AT-rich" (rich in adenine and thymine bases) to assist this process, because A-T base pairs have two hydrogen bonds (rather than the three formed in a C-G pair)—in general, strands rich in these nucleotides are easier to separate due the positive relationship between the number of hydrogen bonds and the difficulty of breaking these bonds. Once strands are separated, RNA primers are created on the template strands. To be more specific, the leading strand receives one RNA primer per active origin of replication while the lagging strand receives several; these several fragments of RNA primers found on the lagging strand of DNA are called Okazaki

fragments, named after their discoverer. DNA Polymerase extends the leading strand in one continuous motion and the lagging strand in a discontinuous motion (due to the Okazaki fragments). RNase removes the RNA fragments used to initiate replication by DNA Polymerase, and another DNA Polymerase enters to fill the gaps. When this is complete, a single nick on the leading strand and several nicks on the lagging strand can be found. Ligase works to fill these nicks in, thus completing the newly replicated DNA molecule.

As DNA synthesis continues, the original DNA strands continue to unwind on each side of the bubble, forming 2 replication forks. In bacteria, which have a single origin of replication on their circular chromosome, this process eventually creates a "theta structure" (resembling the Greek letter theta: θ). In contrast, eukaryotes have longer linear chromosomes and initiate replication at multiple origins within these.

The replication fork



Many enzymes are involved in the DNA replication fork

The replication fork is a structure that forms within the nucleus during DNA replication. It is created by helicases, which break the hydrogen bonds holding the two DNA strands together. The resulting structure has two branching "prongs", each one made up of a single strand of DNA. These two strands serve as the template for the leading and lagging strands, which will be created as DNA polymerase matches complementary nucleotides to the templates; The templates may be properly referred to as the leading strand template and the lagging strand template.

Leading strand

The leading strand is the template strand of the DNA double helix so that the replication fork moves along it in the 3' to 5' direction. This allows the new strand synthesized

complementary to it to be synthesized 5' to 3' in the same direction as the movement of the replication fork.

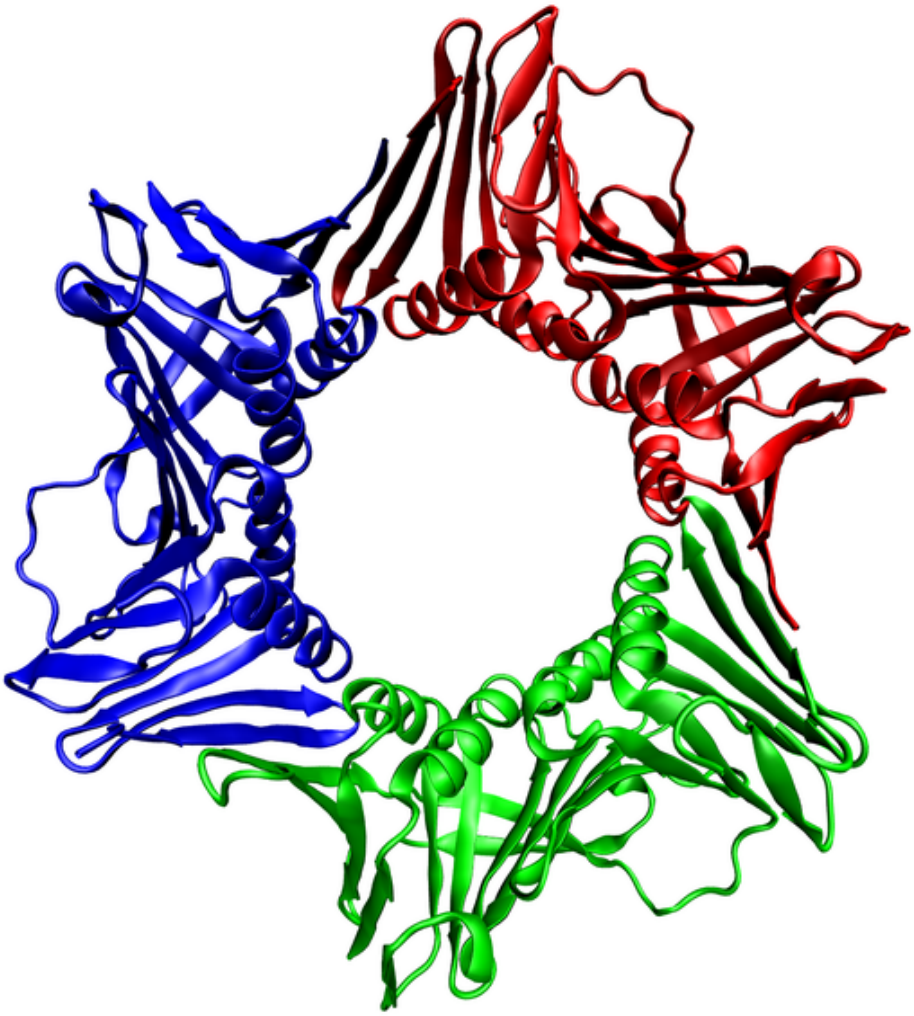
On the leading strand, a polymerase "reads" the DNA and adds nucleotides to it continuously. This polymerase is DNA polymerase III (DNA Pol III) in prokaryotes and presumably Pol ε in eukaryotes.

Lagging strand

The lagging strand is the strand of the template DNA double helix that is oriented so that the replication fork moves along it in a 5' to 3' manner. Because of its orientation, opposite to the working orientation of DNA polymerase III, which moves on a template in a 3' to 5' manner, replication of the lagging strand is more complicated than that of the leading strand.

On the lagging strand, primase "reads" the DNA and adds RNA to it in short, separated segments. In eukaryotes, primase is intrinsic to Pol α. DNA polymerase III or Pol δ lengthens the primed segments, forming Okazaki fragments. Primer removal in eukaryotes is also performed by Pol δ. In prokaryotes, DNA polymerase I "reads" the fragments, removes the RNA using its flap endonuclease domain (RNA primers are removed by 5'-3' exonuclease activity of polymerase I [weaver, 2005], and replaces the RNA nucleotides with DNA nucleotides (this is necessary because RNA and DNA use slightly different kinds of nucleotides). DNA ligase joins the fragments together.

Dynamics at the replication fork



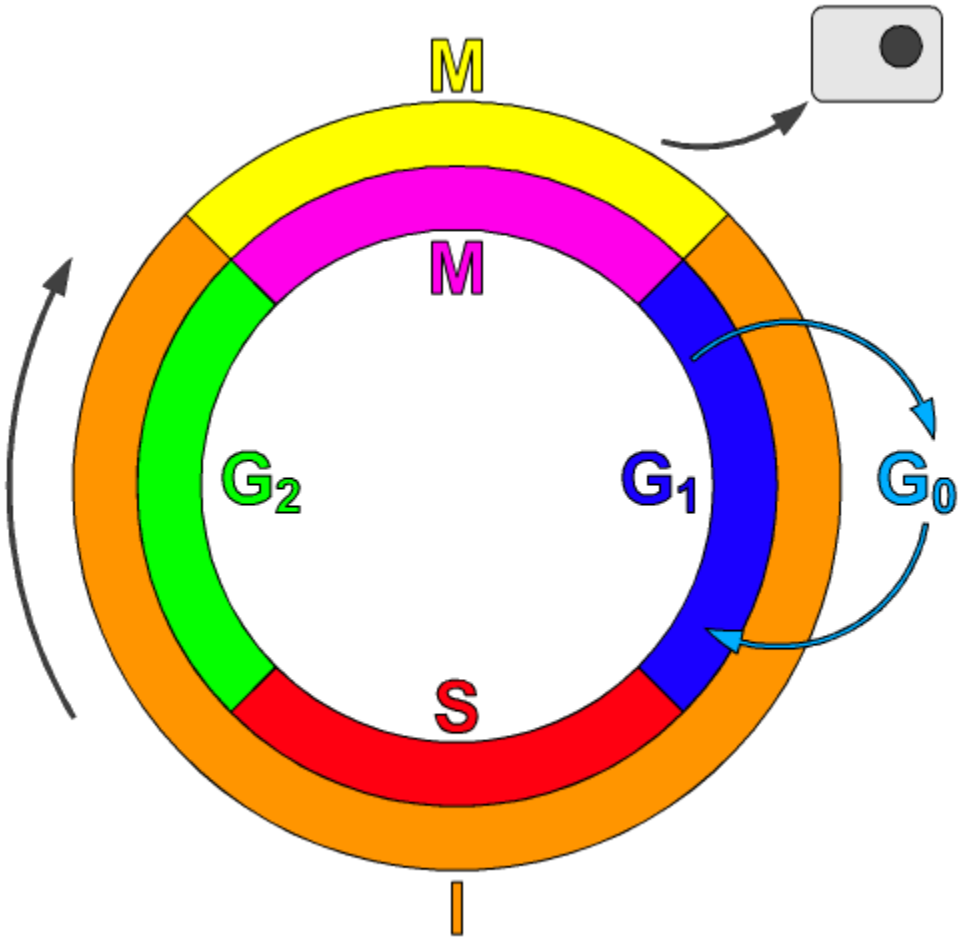
The assembled human DNA clamp, a trimer of the protein PCNA

As helicase unwinds DNA at the replication fork, the DNA ahead is forced to rotate. This process results in a build-up of twists in the DNA ahead. This build-up would form a resistance that would eventually halt the progress of the replication fork. DNA topoisomerases are enzymes that solve these physical problems in the coiling of DNA. Topoisomerase I cuts a single backbone on the DNA, enabling the strands to swivel around each other to remove the build-up of twists. Topoisomerase II cuts both backbones, enabling one double-stranded DNA to pass through another, thereby removing knots and entanglements that can form within and between DNA molecules.

Bare single-stranded DNA has a tendency to fold back upon itself and form secondary structures; these structures can interfere with the movement of DNA polymerase. To prevent this, single-strand binding proteins bind to the DNA until a second strand is synthesized, preventing secondary structure formation.

Clamp proteins form a sliding clamp around DNA, helping the DNA polymerase maintain contact with its template, thereby assisting with processivity. The inner face of the clamp enables DNA to be threaded through it. Once the polymerase reaches the end of the template or detects double-stranded DNA, the sliding clamp undergoes a conformational change that releases the DNA polymerase. Clamp-loading proteins are used to initially load the clamp, recognizing the junction between template and RNA primers.

Regulation of replication



The cell cycle of eukaryotic cells

Eukaryotes

Within eukaryotes, DNA replication is controlled within the context of the cell cycle. As the cell grows and divides, it progresses through stages in the cell cycle; DNA replication occurs during the S phase (**S**ynthesis phase). The progress of the eukaryotic cell through the cycle is controlled by cell cycle checkpoints. Progression through checkpoints is

controlled through complex interactions between various proteins, including cyclins and cyclin-dependent kinases.

The G1/S checkpoint (or restriction checkpoint) regulates whether eukaryotic cells enter the process of DNA replication and subsequent division. Cells that do not proceed through this checkpoint are quiescent in the "G0" stage and do not replicate their DNA.

Replication of chloroplast and mitochondrial genomes occurs independent of the cell cycle, through the process of D-loop replication.

Bacteria

Most bacteria do not go through a well-defined cell cycle and, instead, continuously copy their DNA; during rapid growth, this can result in the concurrent occurrences of multiple rounds of replication. Within *E coli*, the best-characterized bacteria, regulation of DNA replication can be achieved through several mechanisms, including: the hemimethylation and sequestering of the origin sequence, the ratio of ATP to ADP, and the levels of protein DnaA. These all control the process of initiator proteins binding to the origin sequences.

Because *E coli* methylates GATC DNA sequences, DNA synthesis results in hemimethylated sequences. This hemimethylated DNA is recognized by a protein (SeqA), which binds and sequesters the origin sequence; in addition, dnaA (required for initiation of replication) binds less well to hemimethylated DNA. As a result, newly replicated origins are prevented from immediately initiating another round of DNA replication.

ATP builds up when the cell is in a rich medium, triggering DNA replication once the cell has reached a specific size. ATP competes with ADP to bind to DnaA, and the DnaA-ATP complex is able to initiate replication. A certain number of DnaA proteins are also required for DNA replication — each time the origin is copied the number of binding sites for DnaA doubles, requiring the synthesis of more DnaA to enable another initiation of replication.

Termination of replication

Because bacteria have circular chromosomes, termination of replication occurs when the two replication forks meet each other on the opposite end of the parental chromosome. *E coli* regulate this process through the use of termination sequences that, when bound by the Tus protein, enable only one direction of replication fork to pass through. As a result, the replication forks are constrained to always meet within the termination region of the chromosome.

Eukaryotes initiate DNA replication at multiple points in the chromosome, so replication forks meet and terminate at many points in the chromosome; these are not known to be regulated in any particular manner. Because eukaryotes have linear chromosomes, DNA

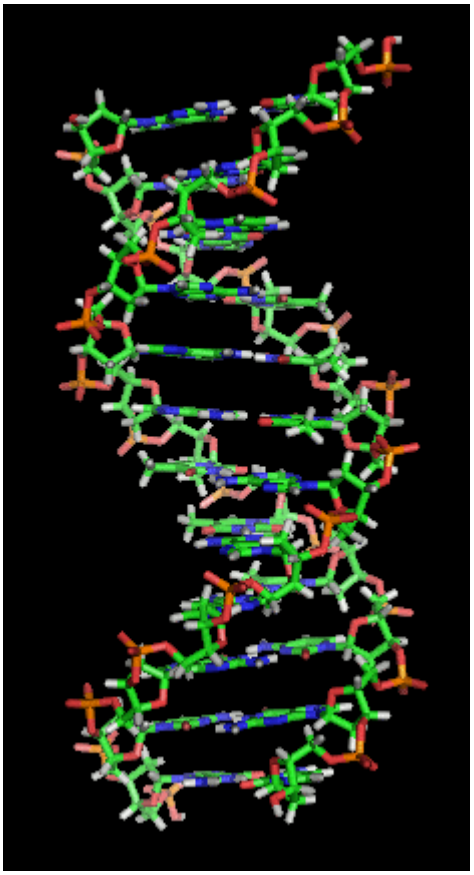
replication is unable to synthesize to the very end of the chromosomes (telomeres), resulting in telomere shortening. This is a normal process in somatic cells — cells are able to divide only a certain number of times before the DNA loss prevents further division. (This is known as the Hayflick limit.) Within the germ cell line, which passes DNA to the next generation, telomerase extends the repetitive sequences of the telomere region to prevent degradation. Telomerase can become mistakenly active in somatic cells, sometimes leading to cancer formation.

Polymerase chain reaction

Researchers commonly replicate DNA *in vitro* using the polymerase chain reaction (PCR). PCR uses a pair of primers to span a target region in template DNA, and then polymerizes partner strands in each direction from these primers using a thermostable DNA polymerase. Repeating this process through multiple cycles produces amplification of the targeted DNA region. At the start of each cycle, the mixture of template and primers is heated, separating the newly synthesized molecule and template. Then, as the mixture cools, both of these become templates for annealing of new primers, and the polymerase extends from these. As a result, the number of copies of the target region doubles each round, increasing exponentially.

Chapter- 5

Nucleic Acid Double Helix



Two complementary regions of nucleic acid molecules will bind and form a double helical structure held together by base pairs.

In molecular biology, the term **double helix** refers to the structure formed by double-stranded molecules of nucleic acids such as DNA and RNA. The double helical structure of a nucleic acid complex arises as a consequence of its secondary structure, and is a fundamental component in determining its tertiary structure.

The DNA double helix is a spiral polymer of nucleic acids, held together by nucleotides which base pair together. In B-DNA, the most common double helical structure, the double helix is right-handed with about 10–10.5 nucleotides per turn. The double helix structure of DNA contains a **major groove** and **minor groove**, the major groove being wider than the minor groove. Given the difference in widths of the major groove and minor groove, many proteins which bind to DNA do so through the wider major groove.

History

The double-helix model of DNA structure was first published in the journal Nature by James D. Watson and Francis Crick in 1953, based upon the crucial X-ray diffraction image of DNA (labeled as "Photo 51") from Rosalind Franklin in 1952, followed by her more clarified DNA image with Raymond Gosling, Maurice Wilkins, Alexander Stokes, and Herbert Wilson, as well as base-pairing chemical and biochemical information by Erwin Chargaff. The previous model was triple-stranded DNA.

Crick, Wilkins, and Watson each received one third of the 1962 Nobel Prize in Physiology or Medicine for their contributions to the discovery. (Franklin, whose breakthrough X-ray diffraction data was used to formulate the DNA structure, died in 1958, and thus was ineligible to be nominated for a Nobel Prize.)

Nucleic acid hybridization

Hybridization is the process of complementary base pairs binding to form a double helix. Melting is the process by which the interactions between the strands of the double helix are broken, separating the two nucleic acid strands. These bonds are weak, easily separated by gentle heating, enzymes, or physical force. Melting occurs preferentially at certain points in the nucleic acid. **T** and **A** rich sequences are more easily melted than **C** and **G** rich regions. Particular base steps are also susceptible to DNA melting, particularly **T A** and **T G** base steps. These mechanical features are reflected by the use of sequences such as **TATAA** at the start of many genes to assist RNA polymerase in melting the DNA for transcription.

Strand separation by gentle heating, as used in PCR, is simple providing the molecules have fewer than about 10,000 base pairs (10 kilobase pairs, or 10 kbp). The intertwining of the DNA strands makes long segments difficult to separate. The cell avoids this problem by allowing its DNA-melting enzymes (helicases) to work concurrently with topoisomerases, which can chemically cleave the phosphate backbone of one of the strands so that it can swivel around the other. Helicases unwind the strands to facilitate the advance of sequence-reading enzymes such as DNA polymerase.

Base pair geometry

The geometry of a base, or base pair step can be characterized by 6 coordinates: Shift, slide, rise, tilt, roll, and twist. These values precisely define the location and orientation

in space of every base or base pair in a nucleic acid molecule relative to its predecessor along the axis of the helix. Together, they characterize the helical structure of the molecule. In regions of DNA or RNA where the "normal" structure is disrupted, the change in these values can be used to describe such disruption.

For each base pair, considered relative to its predecessor, there are the following base pair geometries to consider:

- **Shear**
- **Stretch**
- **Stagger**
- **Buckle**
- **Propeller twist**: rotation of one base with respect to the other in the same base pair.
- **Opening**
- **Shift**: displacement along an axis in the base-pair plane perpendicular to the first, directed from the minor to the major groove.
- **Slide**: displacement along an axis in the plane of the base pair directed from one strand to the other.
- **Rise**: displacement along the helix axis.
- **Tilt**: rotation around this axis.
- **vRoll**: rotation around this axis.
- **Twist**: rotation around the helix axis.
- **vx-displacement**
- **y-displacement**
- **inclination**
- **tip**
- **pitch**: the number of base pairs per complete turn of the helix.

Rise and twist determine the handedness and pitch of the helix. The other coordinates, by contrast, can be zero. Slide and shift are typically small in B-DNA, but are substantial in A- and Z-DNA. Roll and tilt make successive base pairs less parallel, and are typically small. A diagram of these coordinates can be found in 3DNA website.

Note that "tilt" has often been used differently in the scientific literature, referring to the deviation of the first, inter-strand base-pair axis from perpendicularity to the helix axis. This corresponds to slide between a succession of base pairs, and in helix-based coordinates is properly termed "inclination".

DNA helix geometries

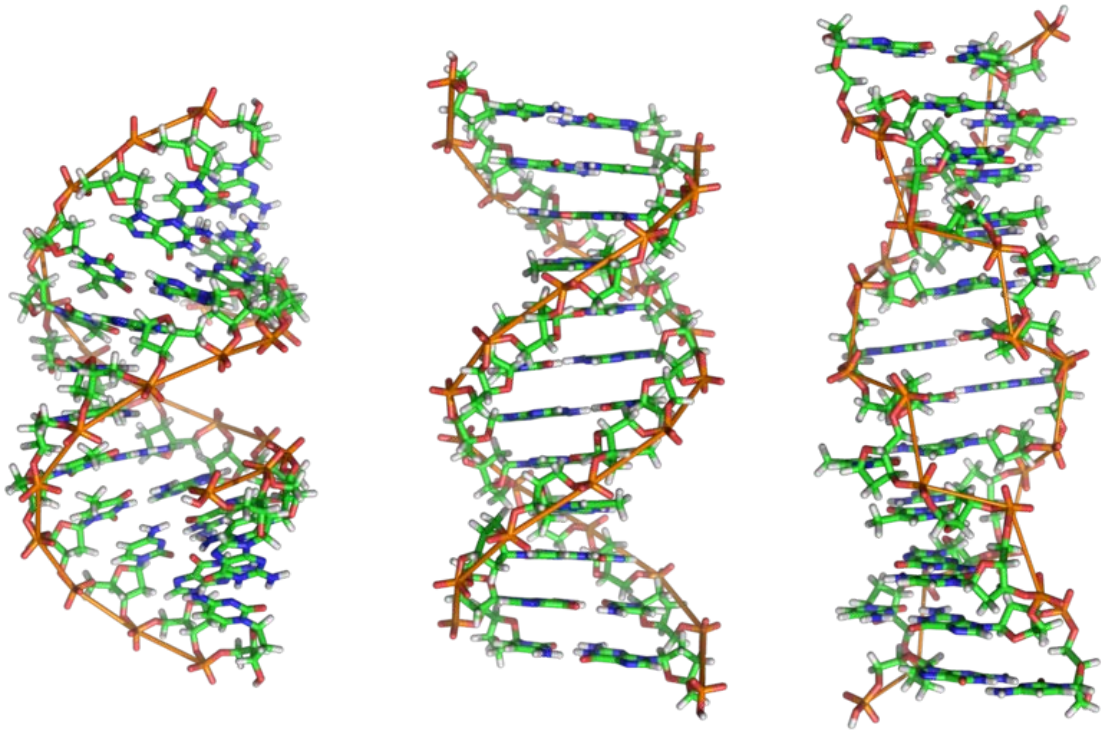
At least three DNA conformations are believed to be found in nature, A-DNA, B-DNA, and Z-DNA. The "B" form described by James D. Watson and Francis Crick is believed to predominate in cells. It is 23.7 Å wide and extends 34 Å per 10 bp of sequence. The double helix makes one complete turn about its axis every 10.4-10.5 base pairs in

solution. This frequency of twist (known as the helical *pitch*) depends largely on stacking forces that each base exerts on its neighbours in the chain.

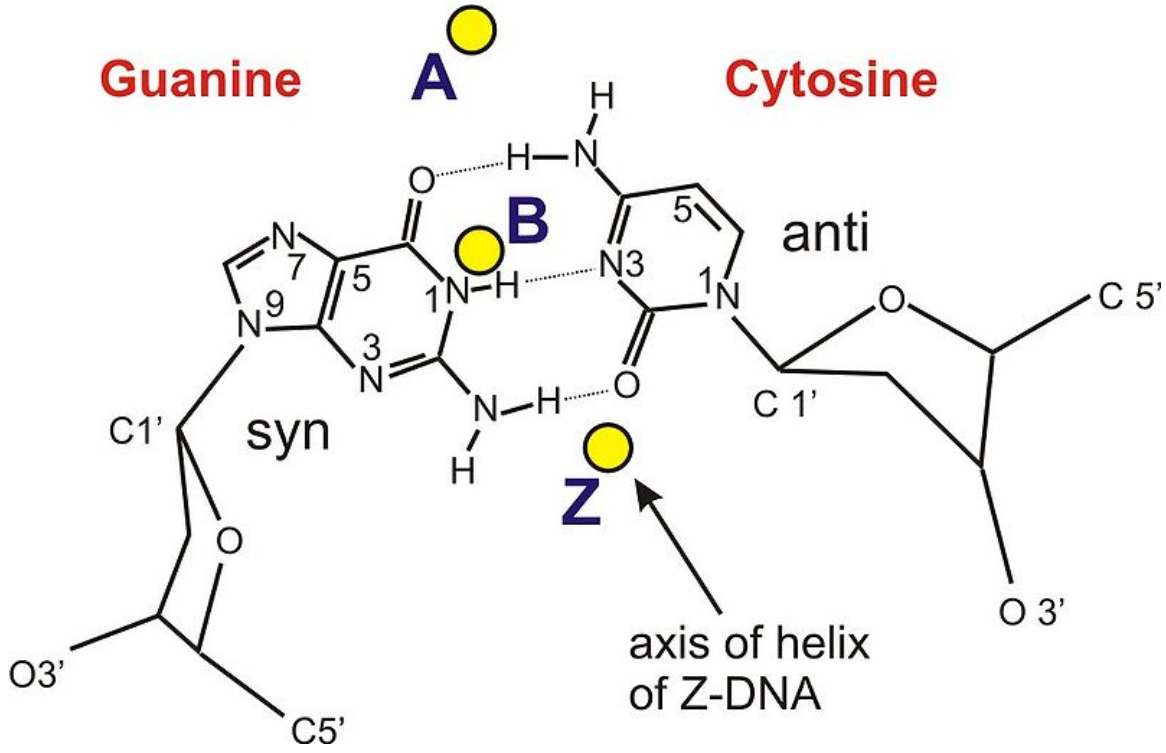
Other conformations are possible; A-DNA, B-DNA, C-DNA, E-DNA, L-DNA (the enantiomeric form of D-DNA), P-DNA, S-DNA, Z-DNA, etc. have been described so far. In fact, only the letters F, Q, U, V, and Y are now available to describe any new DNA structure that may appear in the future. However, most of these forms have been created synthetically and have not been observed in naturally occurring biological systems. Also note the triple-stranded DNA possibility.

A- and Z-DNA

A-DNA and Z-DNA differ significantly in their geometry and dimensions to B-DNA, although still form helical structures. The A form appears likely to occur only in dehydrated samples of DNA, such as those used in crystallographic experiments, and possibly in hybrid pairings of DNA and RNA strands. Segments of DNA that cells have methylated for regulatory purposes may adopt the Z geometry, in which the strands turn about the helical axis the opposite way to A-DNA and B-DNA. There is also evidence of protein-DNA complexes forming Z-DNA structures.



The structures of A-, B-, and Z-DNA



The helix axis of A-, B-, and Z-DNA

Structural features of the three major forms of DNA

Geometry attribute	A-DNA	B-DNA	Z-DNA
Helix sense	right-handed	right-handed	left-handed
Repeating unit	1 bp	1 bp	2 bp
Rotation/bp	32.7°	35.9°	60°/2
bp/turn	11	10.5	12
Inclination of bp to axis	+19°	−1.2°	−9°
Rise/bp along axis	2.3 Å (0.23 nm)	3.32 Å (0.332 nm)	3.8 Å (0.38 nm)
Pitch/turn of helix	28.2 Å (2.82 nm)	33.2 Å (3.32 nm)	45.6 Å (4.56 nm)
Mean propeller twist	+18°	+16°	0°
Glycosyl angle	anti	anti	C: anti, G: syn
Sugar pucker	C3'-endo	C2'-endo	C: C2'-endo, G: C2'-exo
Diameter	23 Å (2.3 nm)	20 Å (2.0 nm)	18 Å (1.8 nm)

Supercoiled DNA

The B form of the DNA helix twists 360° per 10.4-10.5 bp in the absence of torsional strain. But many molecular biological processes can induce torsional strain. A DNA

segment with excess or insufficient helical twisting is referred to, respectively, as positively or negatively "supercoiled". DNA *in vivo* is typically negatively supercoiled, which facilitates the unwinding (melting) of the double-helix required for RNA transcription.

Non-helical forms

Other non-double helical forms of DNA have been described, for example side-by-side (SBS) and triple helical configurations. Single stranded DNA may exist *in statu nascendi* or as thermally induced despiralized DNA.

DNA bending

DNA is a relatively rigid polymer, typically modelled as a worm-like chain. It has three significant degrees of freedom; bending, twisting and compression, each of which cause particular limitations on what is possible with DNA within a cell. Twisting/torsional stiffness is important for the circularisation of DNA and the orientation of DNA bound proteins relative to each other and bending/axial stiffness is important for DNA wrapping and circularisation and protein interactions. Compression/extension is relatively unimportant in the absence of high tension.

Example sequences and their persistence lengths (B DNA)

Sequence	Persistence Length /base pairs
Random	154±10
$(\mathbf{CA})_{repeat}$	133±10
$(\mathbf{CAG})_{repeat}$	124±10
$(\mathbf{TATA})_{repeat}$	137±10

Persistence length/Axial stiffness

DNA in solution does not take a rigid structure but is continually changing conformation due to thermal vibration and collisions with water molecules, which makes classical measures of rigidity impossible. Hence, the bending stiffness of DNA is measured by the persistence length, defined as:

"The length of DNA over which the time-averaged orientation of the polymer becomes uncorrelated by a factor of *e*".

This value may be directly measured using an atomic force microscope to directly image DNA molecules of various lengths. In an aqueous solution, the average persistence length is 46-50 nm or 140-150 base pairs (the diameter of DNA is 2 nm), although can vary significantly. This makes DNA a moderately stiff molecule.

The persistence length of a section of DNA is somewhat dependent on its sequence, and this can cause significant variation. The variation is largely due to base stacking energies and the residues which extend into the minor and major grooves.

Models for DNA bending

Stacking stability of base steps (B DNA)

Step	Stacking ΔG /kcal mol^{-1}
T A	-0.19
T G or **C A**	-0.55
C G	-0.91
A G or **C T**	-1.06
A A or **T T**	-1.11
A T	-1.34
G A or **T C**	-1.43
C C or **G G**	-1.44
A C or **G T**	-1.81
G C	-2.17

The entropic flexibility of DNA is remarkably consistent with standard polymer physics models such as the *Kratky-Porod* worm-like chain model. Consistent with the worm-like chain model is the observation that bending DNA is also described by Hooke's law at very small (sub-piconewton) forces. However for DNA segments less than the persistence length, the bending force is approximately constant and behaviour deviates from the worm-like chain predictions.

This effect results in unusual ease in circularising small DNA molecules and a higher probability of finding highly bent sections of DNA.

Bending preference

DNA molecules often have a preferred direction to bend, ie. anisotropic bending. This is, again, due to the properties of the bases which make up the DNA sequence - a random sequence will have no preferred bend direction, i.e. isotropic bending.

Preferred DNA bend direction is determined by the stability of stacking each base on top of the next. If unstable base stacking steps are always found on one side of the DNA helix then the DNA will preferentially bend away from that direction. As bend angle increases then steric hindrances and ability to roll the residues relative to each other also play a role, especially in the minor groove. **A** and **T** residues will be preferentially be found in the minor grooves on the inside of bends. This effect is particularly seen in DNA-protein binding where tight DNA bending is induced, such as in nucleosome particles.

DNA molecules with exceptional bending preference can become intrinsically bent. This was first observed in trypanosomatid kinetoplast DNA. Typical sequences which cause this contain stretches of 4-6 **T** and **A** residues separated by **G** and **C** rich sections which keep the A and T residues in phase with the minor groove on one side of the molecule. For example:

```
 |                 |                 |                 |
|                 |
G A T T C C C A A A A A T G T C A A A A A A T A G G C A A A A A A T G C
C A A A A A A T C C C A A A C
```

The intrinsically bent structure is induced by the 'propeller twist' of base pairs relative to each other allowing unusual bifurcated Hydrogen-bonds between base steps. At higher temperatures this structure, and so the intrinsic bend, is lost.

All DNA which bends anisotropically has, on average, a longer persistence length and greater axial stiffness. This increased rigidity is required to prevent random bending which would make the molecule act isotropically.

DNA circularization

DNA circularization depends on both the axial (bending) stiffness and torsional (rotational) stiffness of the molecule. For a DNA molecule to successfully circularize it must be long enough to easily bend into the full circle and must have the correct number of bases so the ends are in the correct rotation to allow bonding to occur. The optimum length for circularization of DNA is around 400 base pairs (136 nm), with an integral number of turns of the DNA helix, i.e. multiples of 10.4 base pairs. Having a non integral number of turns presents a significant energy barrier for circularization, for example a 10.4 x 30 = 312 base pair molecule will circularize hundreds of times faster than 10.4 x 30.5 ≈ 317 base pair molecule.

DNA stretching

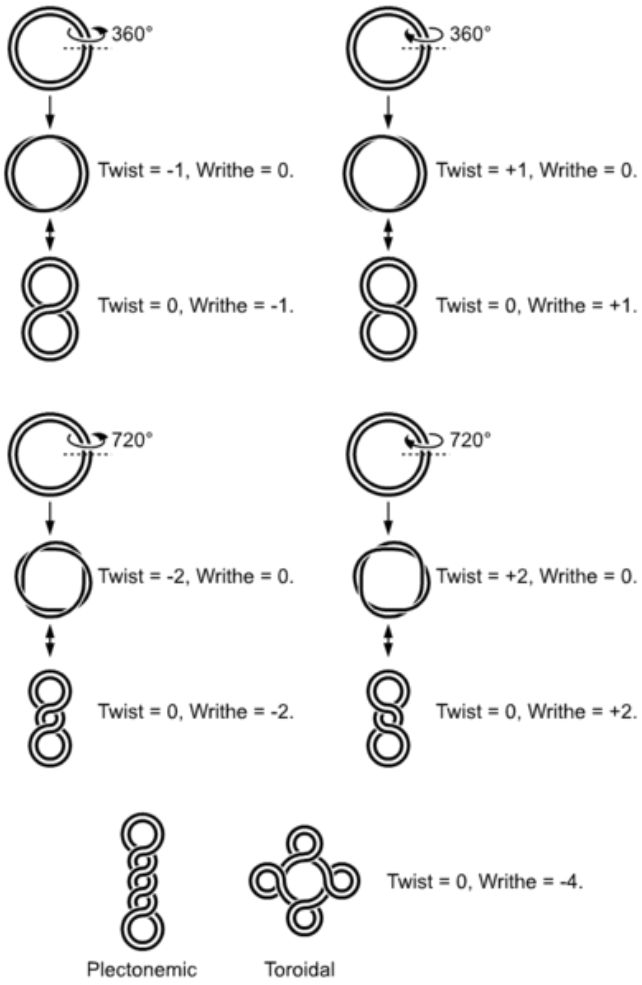
Longer stretches of DNA are entropically elastic under tension. When DNA is in solution, it undergoes continuous structural variations due to the energy available in the thermal bath of the solvent. This is due to the thermal vibration of the molecule combined with continual collisions with water molecules. For entropic reasons, more compact relaxed states are thermally accessible than stretched out states, and so DNA molecules are almost universally found in a tangled relaxed layouts. For this reason, a single molecule of DNA will stretch under a force, straightening it out. Using optical tweezers, the entropic stretching behavior of DNA has been studied and analyzed from a polymer physics perspective, and it has been found that DNA behaves largely like the *Kratky-Porod* worm-like chain model under physiologically accessible energy scales.

Under sufficient tension and positive torque, DNA is thought to undergo a phase transition with the bases splaying outwards and the phosphates moving to the middle.

This proposed structure for overstretched DNA has been called "P-form DNA," in honor of Linus Pauling who originally presented it as a possible structure of DNA

The mechanical properties DNA under compression have not been characterized due to experimental difficulties in preventing the polymer from bending under the compressive force.

DNA topology



Supercoiled structure of circular DNA molecules with low writhe. Note that the helical nature of the DNA duplex is omitted for clarity.

Within the cell most DNA is topologically restricted. DNA is typically found in closed loops (such as plasmids in prokaryotes) which are topologically closed, or as very long molecules whose diffusion coefficients produce effectively topologically closed domains. Linear sections of DNA are also commonly bound to proteins or physical structures (such as membranes) to form closed topological loops.

Francis Crick was one of the first to propose the importance of linking numbers when considering DNA supercoils. In a paper published in 1976, Crick outlined the problem as follows:

In considering supercoils formed by closed double-stranded molecules of DNA certain mathematical concepts, such as the linking number and the twist, are needed. The meaning of these for a closed ribbon is explained and also that of the writhing number of a closed curve. Some simple examples are given, some of which may be relevant to the structure of chromatin.

Analysis of DNA topology uses three values:

L = linking number - the number of times one DNA strand wraps around the other. It is an integer for a closed loop and constant for a closed topological domain.
T = twist - total number of turns in the double stranded DNA helix. This will normally tend to approach the number of turns that a topologically open double stranded DNA helix makes free in solution: number of bases/10.5, assuming there are no intercalating agents (e.g., chloroquine) or other elements modifying the stiffness of the DNA.
W = writhe - number of turns of the double stranded DNA helix around the superhelical axis
$L = T + W$ and $\Delta L = \Delta T + \Delta W$

Any change of T in a closed topological domain must be balanced by a change in W, and vice versa. This results in higher order structure of DNA. A circular DNA molecule with a writhe of 0 will be circular. If the twist of this molecule is subsequently increased or decreased by supercoiling then the writhe will be appropriately altered, making the molecule undergo plectonemic or toroidal superhelical coiling.

When the ends of a piece of double stranded helical DNA are joined so that it forms a circle the strands are topologically knotted. This means the single strands cannot be separated any process that does not involve breaking a strand (such as heating). The task of un-knotting topologically linked strands of DNA falls to enzymes known as topoisomerases. These enzymes are dedicated to un-knotting circular DNA by cleaving one or both strands so that another double or single stranded segment can pass through. This un-knotting is required for the replication of circular DNA and various types of recombination in linear DNA which have similar topological constraints.

The linking number paradox

For many years, the origin of residual supercoiling in eukaryotic genomes remained unclear. This topological puzzle was referred to by some as the "linking number paradox". However, when experimentally determined structures of the nucleosome displayed an overtwisted left-handed wrap of DNA around the histone octamer, this "paradox" was solved.

Chapter- 6

Cloning

The sea anemone, *Anthopleura elegantissima* in process of cloning

Cloning in biology is the process of producing similar populations of genetically identical individuals that occurs in nature when organisms such as bacteria, insects or plants reproduce asexually. Cloning in biotechnology refers to processes used to create copies of DNA fragments (molecular cloning), cells (cell cloning), or organisms. The term also refers to the production of multiple copies of a product such as digital media or software.

The term *clone* is derived from *κλῶνος*, the Greek word for "trunk, branch", referring to the process whereby a new plant can be created from a twig. In horticulture, the spelling *clon* was used until the twentieth century; the final *e* came into use to indicate the vowel is a "long o" instead of a "short o". Since the term entered the popular lexicon in a more general context, the spelling *clone* has been used exclusively.

Molecular cloning

Molecular cloning refers to the process of making multiple molecules. Cloning is commonly used to amplify DNA fragments containing whole genes, but it can also be

used to amplify any DNA sequence such as promoters, non-coding sequences and randomly fragmented DNA. It is used in a wide array of biological experiments and practical applications ranging from genetic fingerprinting to large scale protein production. Occasionally, the term cloning is misleadingly used to refer to the identification of the chromosomal location of a gene associated with a particular phenotype of interest, such as in positional cloning. In practice, localization of the gene to a chromosome or genomic region does not necessarily enable one to isolate or amplify the relevant genomic sequence. To amplify any DNA sequence in a living organism, that sequence must be linked to an origin of replication, which is a sequence of DNA capable of directing the propagation of itself and any linked sequence. However, a number of other features are needed and a variety of specialised cloning vectors (small piece of DNA into which a foreign DNA fragment can be inserted) exist that allow protein expression, tagging, single stranded RNA and DNA production and a host of other manipulations.

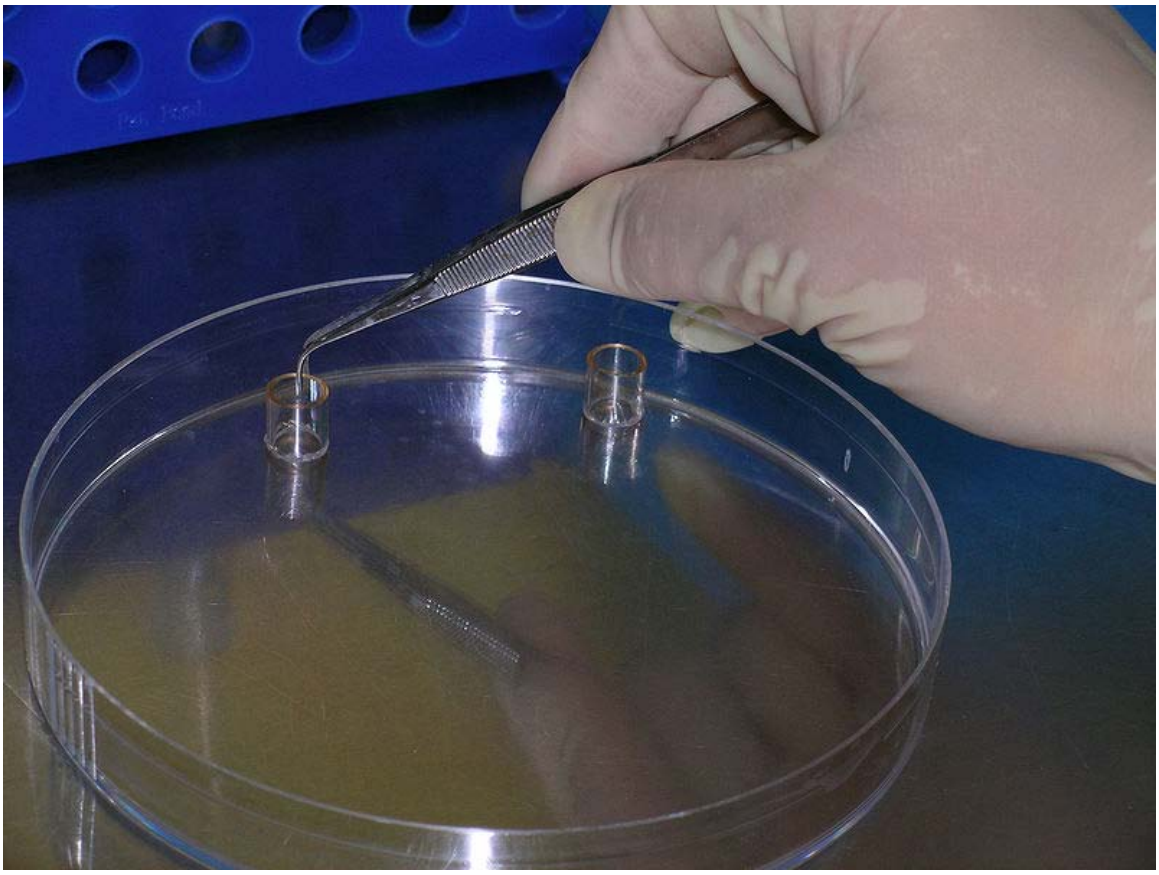
Cloning of any DNA fragment essentially involves four steps

1. fragmentation - breaking apart a strand of DNA
2. ligation - gluing together pieces of DNA in a desired sequence
3. transfection - inserting the newly formed pieces of DNA into cells
4. screening/selection - selecting out the cells that were successfully transfected with the new DNA

Although these steps are invariable among cloning procedures a number of alternative routes can be selected, these are summarized as a 'cloning strategy'.

Initially, the DNA of interest needs to be isolated to provide a DNA segment of suitable size. Subsequently, a ligation procedure is used where the amplified fragment is inserted into a vector (piece of DNA). The vector (which is frequently circular) is linearised using restriction enzymes, and incubated with the fragment of interest under appropriate conditions with an enzyme called DNA ligase. Following ligation the vector with the insert of interest is transfected into cells. A number of alternative techniques are available, such as chemical sensitivation of cells, electroporation, optical injection and biolistics. Finally, the transfected cells are cultured. As the aforementioned procedures are of particularly low efficiency, there is a need to identify the cells that have been successfully transfected with the vector construct containing the desired insertion sequence in the required orientation. Modern cloning vectors include selectable antibiotic resistance markers, which allow only cells in which the vector has been transfected, to grow. Additionally, the cloning vectors may contain colour selection markers, which provide blue/white screening (??-factor complementation) on X-gal medium. Nevertheless, these selection steps do not absolutely guarantee that the DNA insert is present in the cells obtained. Further investigation of the resulting colonies must be required to confirm that cloning was successful. This may be accomplished by means of PCR, restriction fragment analysis and/or DNA sequencing.

Unicellular organisms



Cloning cell-line colonies using cloning rings

Cloning a cell means to derive a population of cells from a single cell. In the case of unicellular organisms such as bacteria and yeast, this process is remarkably simple and essentially only requires the inoculation of the appropriate medium. However, in the case of cell cultures from multi-cellular organisms, cell cloning is an arduous task as these cells will not readily grow in standard media.

A useful tissue culture technique used to clone distinct lineages of cell lines involves the use of cloning rings (cylinders). According to this technique, a single-cell suspension of cells that have been exposed to a mutagenic agent or drug used to drive selection is plated at high dilution to create isolated colonies; each arising from a single and potentially clonal distinct cell. At an early growth stage when colonies consist of only a few of cells, sterile polystyrene rings (cloning rings), which have been dipped in grease are placed over an individual colony and a small amount of trypsin is added. Cloned cells are collected from inside the ring and transferred to a new vessel for further growth.

Cloning in stem cell research

Somatic cell nuclear transfer, known as SCNT, can also be used to create embryos for research or therapeutic purposes. The most likely purpose for this is to produce embryos

for use in stem cell research. This process is also called "research cloning" or "therapeutic cloning." The goal is not to create cloned human beings (called "reproductive cloning"), but rather to harvest stem cells that can be used to study human development and to potentially treat disease. While a clonal human blastocyst has been created, stem cell lines are yet to be isolated from a clonal source.

Organism cloning

Organism cloning (also called reproductive cloning) refers to the procedure of creating a new multicellular organism, genetically identical to another. In essence this form of cloning is an asexual method of reproduction, where fertilization or inter-gamete contact does not take place. Asexual reproduction is a naturally occurring phenomenon in many species, including most plants and some insects. Scientists have made some major achievements with cloning, including the asexual reproduction of sheep and cows. There is a lot of ethical debate over whether or not cloning should be used. However, cloning, or asexual propagation, has been common practice in the horticultural world for hundreds of years.

Horticultural

The term *clone* is used in horticulture to mean all descendants of a single plant, produced by vegetative reproduction or apomixis. Many horticultural plant cultivars are clones, having been derived from a single individual, multiplied by some process other than sexual reproduction. As an example, some European cultivars of grapes represent clones that have been propagated for over two millennia. Other examples are potato and banana. Grafting can be regarded as cloning, since all the shoots and branches coming from the graft are genetically a clone of a single individual, but this particular kind of cloning has not come under ethical scrutiny and is generally treated as an entirely different kind of operation.

Many trees, shrubs, vines, ferns and other herbaceous perennials form clonal colonies. Parts of a large clonal colony often become detached from the parent, termed fragmentation, to form separate individuals. Some plants also form seeds is asexually, termed apomixis, e.g. dandelion.

Parthenogenesis

Clonal derivation exists in nature in some animal species and is referred to as parthenogenesis (reproduction of an organism by itself without a mate). This is an asexual form of reproduction that is only found in females of some insects, crustaceans and lizards. The growth and development occurs without fertilization by a male. In plants, parthenogenesis means the development of an embryo from an unfertilized egg cell, and is a component process of apomixis. In species that use the XY sex-determination system, the offspring will always be female. An example is the "Little Fire Ant" (*Wasmannia auropunctata*), which is native to Central and South America but has spread throughout many tropical environments.

Artificial cloning of organisms

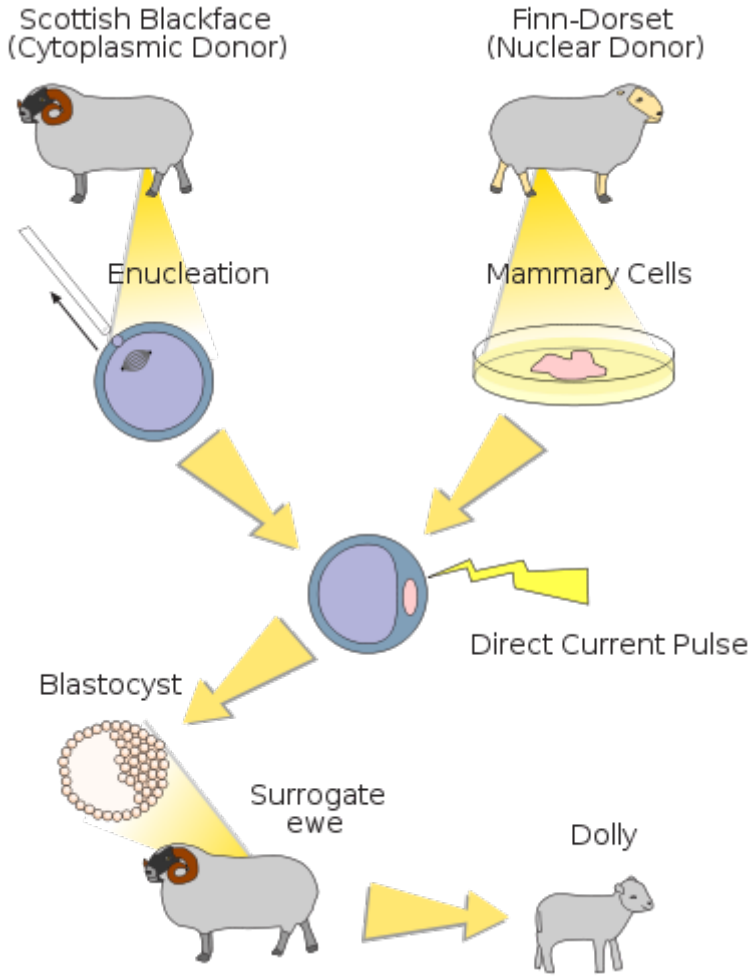
Artificial cloning of organisms may also be called *reproductive cloning*.

Methods

Reproductive cloning generally uses "somatic cell nuclear transfer" (SCNT) to create animals that are genetically identical. This process entails the transfer of a nucleus from a donor adult cell (somatic cell) to an egg that has no nucleus. If the egg begins to divide normally it is transferred into the uterus of the surrogate mother. Such clones are not strictly identical since the somatic cells may contain mutations in their nuclear DNA. Additionally, the mitochondria in the cytoplasm also contains DNA and during SCNT this DNA is wholly from the donor egg, thus the mitochondrial genome is not the same as that of the nucleus donor cell from which it was produced. This may have important implications for cross-species nuclear transfer in which nuclear-mitochondrial incompatibilities may lead to death.

Artificial *embryo splitting* or *embryo twinning* may also be used as a method of cloning, where an embryo is split in the maturation before embryo transfer. It is optimally performed at the 6- to 8-cell stage, where it can be used as an expansion of IVF to increase the number of available embryos. If both embryos are successful, it gives rise to monozygotic (identical) twins.

Dolly the Sheep



Dolly, a Finn-Dorset ewe, was the first mammal to have been successfully cloned from an adult cell. She was cloned at the Roslin Institute in Scotland and lived there from her birth in 1996 until her death in 2003 when she was six. Her stuffed remains were placed at Edinburgh's Royal Museum, part of the National Museums of Scotland.

Dolly was publicly significant because the effort showed that the genetic material from a specific adult cell, programmed to express only a distinct subset of its genes, can be reprogrammed to grow an entirely new organism. Before this demonstration, it had been shown by John Gurdon that nuclei from differentiated cells could give rise to an entire organism after transplantation into an enucleated egg. However, this concept was not yet demonstrated in a mamallian system.

Cloning Dolly the sheep had a low success rate per fertilized egg; she was born after 237 eggs were used to create 29 embryos, which only produced three lambs at birth, only one of which lived. Seventy calves have been created and one third of them died young;

Prometea took 277 attempts. Notably, although the first clones were frogs, no adult cloned frog has yet been produced from a somatic adult nucleus donor cell.

There were early claims that Dolly the Sheep had pathologies resembling accelerated aging. Scientists speculated that Dolly's death in 2003 was related to the shortening of telomeres, DNA-protein complexes that protect the end of linear chromosomes. However, other researchers, including Ian Wilmut who led the team that successfully cloned Dolly, argue that Dolly's early death due to respiratory infection was unrelated to deficiencies with the cloning process.

Water Buffalo

On September 15, 2007, the Philippines announced its development of Southeast Asia's first cloned water buffalo. The Philippine Council for Agriculture, Forestry and Natural Resources Research and Development (PCARRD), under the Department of Science and Technology in Los Baños, Laguna approved this project.

Species cloned

The modern cloning techniques involving nuclear transfer have been successfully performed on several species. Landmark experiments in chronological order:

- Tadpole: (1952) Many scientists questioned whether cloning had actually occurred and unpublished experiments by other labs were not able to reproduce the reported results.
- Carp: (1963) In China, embryologist Tong Dizhou produced the world's first cloned fish by inserting the DNA from a cell of a male carp into an egg from a female carp. He published the findings in a Chinese science journal.
- Mice: (1986) A mouse was the first mammal successfully cloned from an early embryonic cell. Soviet scientists Chaylakhyan, Veprencev, Sviridova, and Nikitin had the mouse "Masha" cloned. Research was published in the magazine "Biofizika" volume XXXII, issue 5 of 1987.
- Sheep: (1996) From early embryonic cells by Steen Willadsen. Megan and Morag cloned from differentiated embryonic cells in June 1995 and Dolly the sheep from a somatic cell in 1997.
- Rhesus Monkey: Tetra (January 2000) from embryo splitting
- Gaur: (2001) was the first endangered species cloned.
- Cattle: Alpha and Beta (males, 2001) and (2005) Brazil
- Cat: CopyCat "CC" (female, late 2001), Little Nicky, 2004, was the first cat cloned for commercial reasons
- Dog: Snuppy, a male Afghan hound was the first cloned dog (2005).
- Rat: Ralph, the first cloned rat (2003)
- Mule: Idaho Gem, a john mule born 4 May 2003, was the first horse-family clone.
- Horse: Prometea, a Haflinger female born 28 May 2003, was the first horse clone.

- Water Buffalo: Samrupa was the first cloned water buffalo. It was born on February 6, 2009, at India's Karnal National Diary Research Institute but died five days later due to lung infection.
- Camel: (2009) Injaz, is the first cloned camel.

Human cloning

Human cloning is the creation of a genetically identical copy of an existing or previously existing human. The term is generally used to refer to *artificial* human cloning; human clones in the form of identical twins are commonplace, with their cloning occurring during the natural process of reproduction. There are two commonly discussed types of human cloning: therapeutic cloning and reproductive cloning. Therapeutic cloning involves cloning adult cells for use in medicine and is an active area of research. Reproductive cloning would involve making cloned humans. A third type of cloning called replacement cloning is a theoretical possibility, and would be a combination of therapeutic and reproductive cloning. Replacement cloning would entail the replacement of an extensively damaged, failed, or failing body through cloning followed by whole or partial brain transplant.

The various forms of human cloning are controversial. There have been numerous demands for all progress in the human cloning field to be halted. Most scientific, governmental and religious organizations oppose reproductive cloning. The American Association for the Advancement of Science (AAAS) and other scientific organizations have made public statements suggesting that human reproductive cloning be banned until safety issues are resolved. Serious ethical concerns have been raised by the future possibility of harvesting organs from clones. Some people have considered the idea of growing organs separately from a human organism - in doing this, a new organ supply could be established without the moral implications of harvesting them from humans. Research is also being done on the idea of growing organs that are biologically acceptable to the human body inside of other organisms, such as pigs or cows, then transplanting them to humans, a form of xenotransplantation.

The first hybrid human clone was created in November 1998, by American Cell Technologies. It was created from a man's leg cell, and a cow's egg whose DNA was removed. It was destroyed after 12 days. Since a normal embryo implants at 14 days, Dr Robert Lanza, ACT's director of tissue engineering, told the Daily Mail newspaper that the embryo could not be seen as a person before 14 days. While making an embryo, which may have resulted in a complete human had it been allowed to come to term, according to ACT: "[ACT's] aim was 'therapeutic cloning' not 'reproductive cloning'"

On January, 2008, Wood and Andrew French, Stemagen's chief scientific officer in California, announced that they successfully created the first 5 mature human embryos using DNA from adult skin cells, aiming to provide a source of viable embryonic stem cells. Dr. Samuel Wood and a colleague donated skin cells, and DNA from those cells was transferred to human eggs. It is not clear if the embryos produced would have been capable of further development, but Dr. Wood stated that if that were possible, using the

technology for reproductive cloning would be both unethical and illegal. The 5 cloned embryos, created in Stemagen Corporation lab, in La Jolla, were destroyed.

Ethical issues of cloning

Because of recent technological advancements, the cloning of animals (and potentially humans) has been an issue. The Catholic Church and many religious organizations oppose all forms of cloning, on the grounds that life begins at conception. Judaism does not equate life with conception and, though some question the wisdom of cloning, Orthodox rabbis generally find no firm reason in Jewish law and ethics to object to cloning. From the standpoint of classical liberalism, concerns also exist regarding the protection of the identity of the individual and the right to protect one's genetic identity.

Gregory Stock is a scientist and outspoken critic against restrictions on cloning research. Bioethicist Gregory Pence also attacks the idea of criminalizing attempts to clone humans.

The social implications of an artificial human production scheme were famously explored in Aldous Huxley's novel *Brave New World.*

On December 28, 2006, the U.S. Food and Drug Administration (FDA) approved the consumption of meat and other products from cloned animals. Cloned-animal products were said to be virtually indistinguishable from the non-cloned animals. Furthermore, companies would not be required to provide labels informing the consumer that the meat comes from a cloned animal.

Critics have raised objections to the FDA's approval of cloned-animal products for human consumption, arguing that the FDA's research was inadequate, inappropriately limited, and of questionable scientific validity. Several consumer-advocate groups are working to encourage a tracking program that would allow consumers to become more aware of cloned-animal products within their food.

Joseph Mendelson, legal director of the Center for Food Safety, said that cloned food still should be labeled since safety and ethical issues about it remain questionable.

Carol Tucker Foreman, director of food policy at the Consumer Federation of America, stated that FDA does not consider the fact that the results of some studies revealed that cloned animals have increased rates of mortality and deformity at birth.

Cloning extinct and endangered species

Cloning, or more precisely, the reconstruction of functional DNA from extinct species has, for decades, been a dream of some scientists. The possible implications of this were dramatized in the best-selling novel by Michael Crichton and high budget Hollywood thriller *Jurassic Park*. In real life, one of the most anticipated targets for cloning was once the Woolly Mammoth, but attempts to extract DNA from frozen mammoths have

been unsuccessful, though a joint Russo-Japanese team is currently working toward this goal.

In 2001, a cow named Bessie gave birth to a cloned Asian gaur, an endangered species, but the calf died after two days. In 2003, a banteng was successfully cloned, followed by three African wildcats from a thawed frozen embryo. These successes provided hope that similar techniques (using surrogate mothers of another species) might be used to clone extinct species. Anticipating this possibility, tissue samples from the last *bucardo* (Pyrenean Ibex) were frozen in liquid nitrogen immediately after it died in 2000. Researchers are also considering cloning endangered species such as the giant panda, ocelot, and cheetah. The "Frozen Zoo" at the San Diego Zoo now stores frozen tissue from the world's rarest and most endangered species.

In 2002, geneticists at the Australian Museum announced that they had replicated DNA of the Thylacine (Tasmanian Tiger), extinct about 65 years previous, using polymerase chain reaction. However, on February 15, 2005 the museum announced that it was stopping the project after tests showed the specimens' DNA had been too badly degraded by the (ethanol) preservative. On 15 May 2005 it was announced that the Thylacine project would be revived, with new participation from researchers in New South Wales and Victoria.

In January 2009, for the first time, an extinct animal, the Pyrenean ibex mentioned above was cloned, at the Centre of Food Technology and Research of Aragon, using the preserved DNA of the skin samples from 2001 and domestic goat egg-cells. (The ibex died shortly after birth due to physical defects in its lungs.) One of the continuing obstacles in the attempt to clone extinct species is the need for nearly perfect DNA. Cloning from a single specimen could not create a viable breeding population in sexually reproducing animals. Furthermore, even if males and females were to be cloned, the question would remain open whether they would be viable at all in the absence of parents that could teach or show them their natural behavior.

Cloning endangered species is a highly ideological issue. Many conservation biologists and environmentalists vehemently oppose cloning endangered species — mainly because they think it may deter donations to help preserve natural habitat and wild animal populations. The "rule-of-thumb" in animal conservation is that, if it is still feasible to conserve habitat and viable wild populations, breeding in captivity should not be undertaken in isolation.

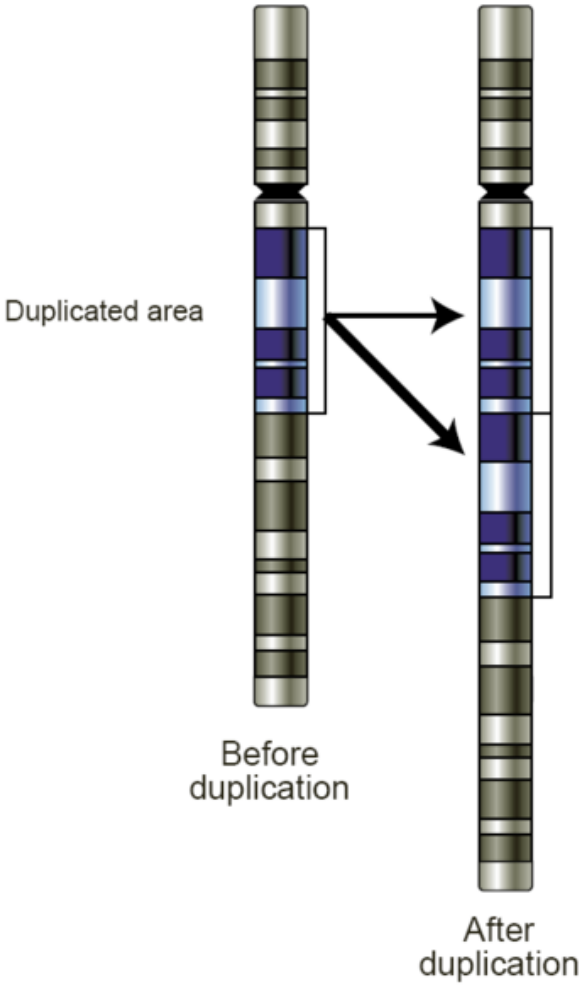
In a 2006 review, David Ehrenfeld concluded that cloning in animal conservation is an experimental technology that, at its state in 2006, could not be expected to work except by pure chance and utterly failed a cost-benefit analysis. Furthermore, he said, it is likely to siphon funds from established and working projects and does not address any of the issues underlying animal extinction (such as habitat destruction, hunting or other overexploitation, and an impoverished gene pool). While cloning technologies are well-established and used on a regular basis in plant conservation, care must be taken to ensure genetic diversity. He concluded:

Vertebrate cloning poses little risk to the environment, but it can consume scarce conservation resources, and its chances of success in preserving species seem poor. To date, the conservation benefits of transgenics and vertebrate cloning remain entirely theoretical, but many of the risks are known and documented. Conservation biologists should devote their research and energies to the established methods of conservation, none of which require transgenics or vertebrate cloning.

Chapter- 7

Gene Duplication and Gene Expression

Gene duplication



Schematic of a region of a chromosome before and after a duplication event

Gene duplication (or **chromosomal duplication** or **gene amplification**) is any duplication of a region of DNA that contains a gene; it may occur as an error in homologous recombination, a retrotransposition event, or duplication of an entire chromosome. The second copy of the gene is often free from selective pressure — that is, mutations of it have no deleterious effects to its host organism. Thus it accumulates mutations faster than a functional single-copy gene, over generations of organisms.

A duplication is the opposite of a deletion. Duplications arise from an event termed unequal crossing-over that occurs during meiosis between misaligned homologous chromosomes. The chance of this happening is a function of the degree of sharing of repetitive elements between two chromosomes. The product of this recombination are a duplication at the site of the exchange and a reciprocal deletion.

Gene duplication as an evolutionary event

Gene duplication is believed to play a major role in evolution; this stance has been held by members of the scientific community for over 100 years. Susumu Ohno was one of the most famous developers of this theory in his classic book *Evolution by gene duplication* (1970). Ohno argued that gene duplication is the most important evolutionary force since the emergence of the universal common ancestor. Major genome duplication events are not uncommon. It is believed that the entire yeast genome underwent duplication about 100 million years ago. Plants are the most prolific genome duplicators. For example, wheat is hexaploid (a kind of polyploid), meaning that it has six copies of its genome.

The duplication of a gene results in an additional copy that is free from selective pressure. One kind of view is that this allows the new copy of the gene to mutate without deleterious consequence to the organism. This freedom from consequences allows for the mutation of novel genes that could potentially increase the fitness of the organism or code for a new function. An example of this is the apparent mutation of a duplicated digestive gene in a family of ice fish into an antifreeze gene.

Another view is the DDC(duplication-degeneration-complementation) model, which assumes that after gene duplication, instead one copy retains the original function, the two copies degenerate complementary subfucntion (called subfunctionalization).

The two genes that exist after a gene duplication event are called paralogs and usually code for proteins with a similar function and/or structure. By contrast, orthologous genes are ones which code for proteins with similar functions but exist in different species, and are created from a speciation event.

It is important (but often difficult) to differentiate between paralogs and orthologs in biological research. Experiments on human gene function can often be carried out on other species if a homolog to a human gene can be found in the genome of that species, but only if the homolog is orthologous. If they are paralogs and resulted from a gene duplication event, their functions are likely to be too different.

The paralogous segments can be repeat sequences with more than 90% sequence similarity. In such cases, they are known as low copy repeats (LCRs) though they are not highly repetitive sequences. They are mostly found in pericentronomic, subtelomeric and interstitial regions of a chromosome. The LCRs, due to their size (>1Kb), similarity, and orientation, are highly susceptible to duplications and deletions. These genomic rearrangements are caused by the mechanism of non-allelic homologous recombination. The resulting genomic variation leads to gene dosage dependent neurological disorders such as Rett-like syndrome and Pelizaeus-Merzbacher disease.

Gene duplication as amplification

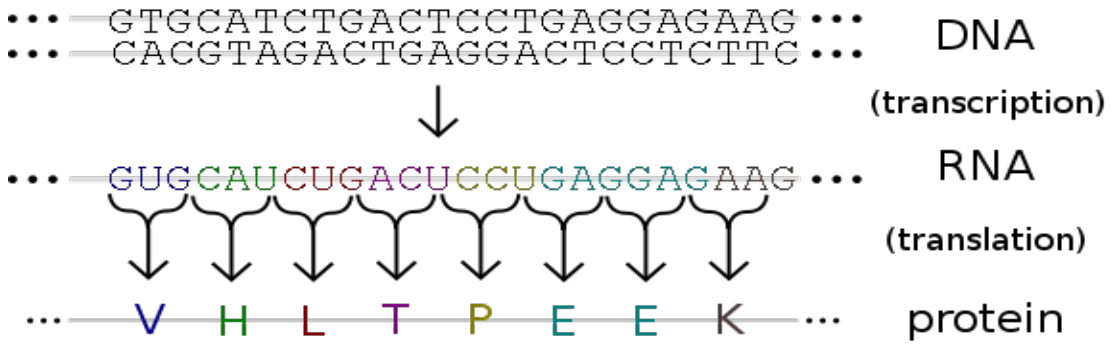
Gene duplication doesn't necessarily constitute a lasting change in a species' genome. In fact, such changes often don't last past the initial host organism. From the perspective of molecular genetics, amplification is one of many ways in which a gene can be overexpressed. Genetic amplification can occur artificially, as with the use of the polymerase chain reaction technique to amplify short strands of DNA *in vitro* using enzymes, or it can occur naturally, as described above. If it's a natural duplication, it can still take place in a somatic cell, rather than a germline cell (which would be necessary for a lasting evolutionary change).

Also, in either event, duplications can be and often are marginally or severely detrimental. For instance, duplications of oncogenes are a common cause of many types of cancer, as is the case with P70-S6 Kinase 1 amplification and breast cancer. In such cases the genetic duplication occurs in a somatic cell and affects only the genome of the cancer cells themselves, not the entire organism, much less any subsequent offspring.

Genomic microarrays detect Duplications

Technologies such as genomic microarrays, also called array comparative genomic hybridization (array CGH), are used to detect chromosomal abnormalities, such as microduplications, in a high throughput fashion from genomic DNA samples. In particular, DNA microarray technology can simultaneously monitor the expression levels of thousands of genes across many treatments or experimental conditions, greatly facilitating the evolutionary studies of gene regulation after gene duplication or speciation.

Gene expression



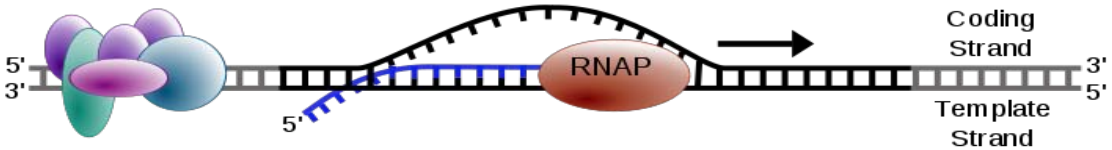
Genes are expressed by being transcribed into RNA, and this transcript may then be translated into protein.

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. These products are often proteins, but in non-protein coding genes such as ribosomal RNA (rRNA) genes or transfer RNA (tRNA) genes, the product is a functional RNA. The process of gene expression is used by all known life - eukaryotes (including multicellular organisms), prokaryotes (bacteria and archaea) and viruses - to generate the macromolecular machinery for life. Several steps in the gene expression process may be modulated, including the transcription, RNA splicing, translation, and post-translational modification of a protein. Gene regulation gives the cell control over structure and function, and is the basis for cellular differentiation, morphogenesis and the versatility and adaptability of any organism. Gene regulation may also serve as a substrate for evolutionary change, since control of the timing, location, and amount of gene expression can have a profound effect on the functions (actions) of the gene in a cell or in a multicellular organism.

In genetics, gene expression is the most fundamental level at which genotype gives rise to the phenotype. The genetic code stored in DNA in form of nucleotide sequence is "interpreted" by gene expression, and the properties of the expression products give rise to the organism's phenotype.

Mechanism

Transcription



The process of transcription is carried out by RNA polymerase (RNAP), uses DNA (black) as a template and produces RNA (blue).

The gene itself is typically a long stretch of DNA which carries genetic information encoded by genetic code. Every molecule of DNA consists of two strands, each of them having 5' and 3' ends oriented in anti-parallel direction. The coding strand contains the genetic information while template strand (non-coding strand) serves as a blueprint for the production of RNA. The production of RNA copies of the DNA is called transcription, and is performed by RNA polymerase, which adds one RNA nucleotide at a time to a growing RNA strand. This RNA is complementary to the template 3' → 5' DNA strand, which is itself complementary to the coding 5' → 3' DNA strand. Therefore, the resulting 5' → 3' RNA strand is identical to the coding DNA strand with the exception that thymines (T) are replaced with uracils (U) in the RNA. A coding DNA strand reading "ATG" is transcribed as "AUG" in RNA.

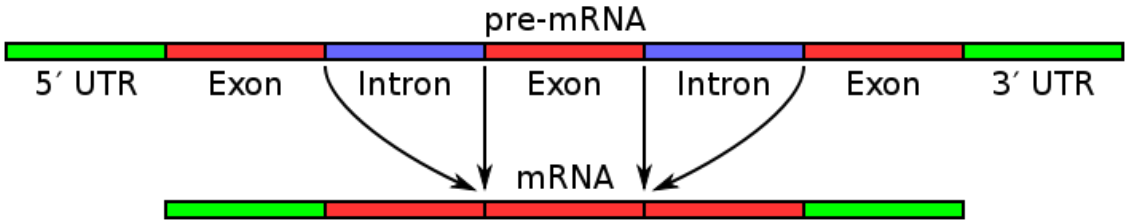
Transcription in prokaryotes is carried out by a single type of RNA polymerase, which needs DNA sequence called Pribnow box and sigma factor (σ factor) to start transcription. In eukaryotes, the transcription is done by three types of RNA polymerases, each of them needs special DNA sequence called promoter and a set of DNA-binding proteins - transcription factors to initiate the process. RNA polymerase I is responsible for transcription of rRNA genes, while RNA polymerase II transcribes all protein-coding genes but also some non-coding RNAs (e.g. snRNAs, snoRNAs or long non-coding RNAs) as well. It contains special part called C-terminal domain (CTD) that is rich of serines, which after being phosphorylated accumulate factors necessary for RNA modification and maturation. RNA polymerase III transcribes 5S rRNA and tRNA genes but also some small non-coding RNA genes (e.g. 7SK). Transcription ends on a special sequence called terminator.

RNA processing

While transcription of prokaryotic protein-coding genes creates messenger RNA (mRNA) which is ready for translation, transcription of eukaryotic genes leaves a primary transcript of RNA (pre-mRNA), which fist has to undergo series of modification to become a mature mRNA.

These include 5' *capping*, which is set of enzymatic reactions that add 7-methylguanosine (m^7G) to the 5' end of pre-mRNA and thus protect the RNA from degradation by exonucleases. The m^7G cap is then bound by cap binding complex heterodimer (CBC20/CBC80) which aids in mRNA export to cytoplasm and also protect the RNA from decapping.

Another modification is 3' *clevage and polyadenylation*. They occur if polyadenylation signal sequence (5'- AAUAAA-3') is present in pre-mRNA,which is usually between protein-coding sequence and terminator. The pre-mRNA is first cleaved and then a series of ~200 adenines (A) are added to form poly(A) tail which protects the RNA from degradation. Poly(A) tail is bound by multiple poly(A)-binding proteins (PABP) necessary for mRNA export and translation re-iniciation.



Simple illustration of exons and introns in pre-mRNA and the formation of mature mRNA by splicing. The UTRs are non-coding parts of exons at the ends of the mRNA.

Very important modification of eukaryotic pre-mRNA is *RNA splicing*. Majority of eukaryotic pre-mRNAs consist of alternating segments called exons and introns. During the process of splicing, RNA-protein catalytical complex known as spliceosome, catalyze two transesterification reactions, which remove intron and release it in form of lariat structure and then splice neighbouring exons together. In certain cases, some introns or exons can be either removed or retained in mature mRNA. This so-called alternative splicing creates series of different transcripts originating from a single gene. Because these transcripts can be potentially translated into different proteins, splicing extends the complexity of eukaryotic gene expression.

Extensive RNA processing may be an evolutionary advantage made possible by the nucleus of eukaryotes. In prokaryotes transcription and translation happen together whilst in eukaryotes the nuclear membrane separates the two processes giving time for RNA processing to occur.

non-coding RNA maturation

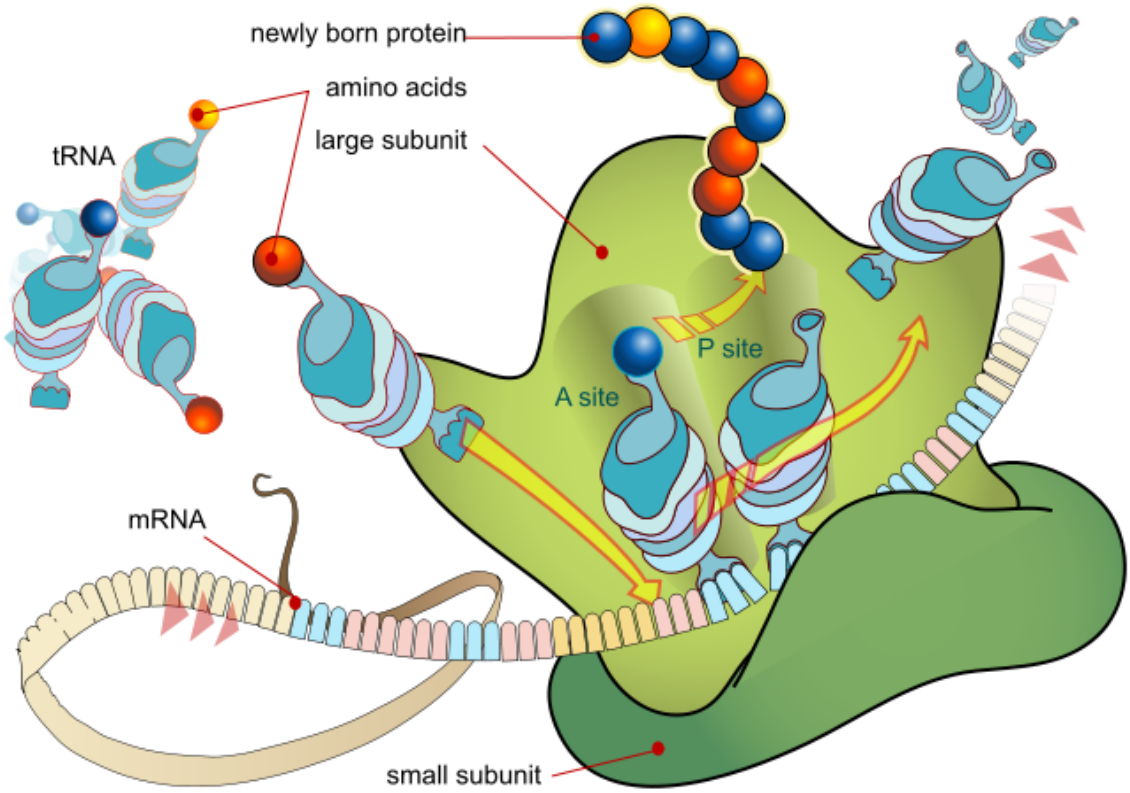
In most organisms non-coding genes (ncRNA) are transcribed as precursors which undergo further processing. In the case of ribosomal RNAs (rRNA), they are often transcribed as a pre-rRNA which contains one or more rRNAs, the pre-rRNA is cleaved and modified (2′-O-methylation and pseudouridine formation) at a specific sites by approximately 150 different small nucleolus-restricted RNA species, called snoRNAs. SnoRNAs associate with proteins, forming snoRNPs. While snoRNA part basepair with the target RNA and thus position the modification to precise site, the protein part performs the catalytical reaction. In eukaryotes, in particular a snoRNP, called RNase MRP cleaves the 45S pre-rRNA into the 28S, 5.8S, and 18S rRNAs. The rRNA and RNA processing factors form large aggregates called the nucleolus.
In the case of transfer RNA (tRNA), for example, the 5' sequence is removed by RNase P, whereas the 3' end is removed by the tRNase Z enzyme and the non-templated 3' CCA tail is added by a nucleotidyl transferase. In the case of micro RNA (miRNA), miRNAs are first transcribed as primary transcripts or pri-miRNA with a cap and poly-A tail and processed to short, 70-nucleotide stem-loop structures known as pre-miRNA in the cell nucleus by the enzymes Drosha and Pasha. After being exported, it is then processed to mature miRNAs in the cytoplasm by interaction with the endonuclease Dicer, which also initiates the formation of the RNA-induced silencing complex (RISC), composed of the Argonaute protein. Even snRNAs and snoRNAs themselves undergo series of

modification before they become part of functional RNP complex. This is done either in the nucleoplasm or in the specialized compartments called Cajal bodies. Their bases are methylated or pseudouridinilated by a group of small Cajal body-specific RNAs (scaRNAs) which are structurally similar to snoRNAs.

RNA export

In eukaryotes most mature RNA must be exported to the cytoplasm from the nucleus. While some RNAs function in the nucleus, many RNAs are transported through the nuclear pores and into the cytosol. Notably this includes all RNA types involved in protein synthesis. In some cases RNAs are additionally transported to a specific part of the cytoplasm, such as a synapse; they are then towed by motor proteins that bind through linker proteins to specific sequences (called "zipcodes") on the RNA.

Translation



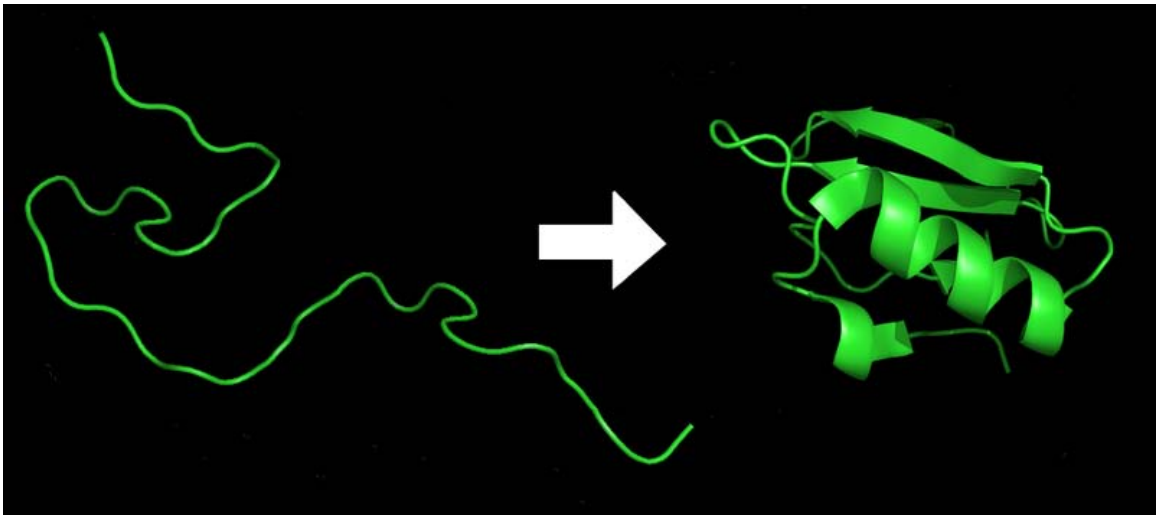
During the translation, tRNA charged with amino acid enters the ribosome and aligns with the correct mRNA triplet. Ribosome then adds amino acid to growing protein chain.

For some RNA (non-coding RNA) the mature RNA is the final gene product. In the case of messenger RNA (mRNA) the RNA is an information carrier coding for the synthesis of one or more proteins. mRNA carrying a single protein sequence (common in eukaryotes) is monocistronic whilst mRNA carrying multiple protein sequences (common in prokaryotes) is known as polycistronic.

Every mRNA consists of three parts - 5' untranslated region (5'UTR), protein-coding region or open reading frame (ORF) and 3' untranslated region (3'UTR). Coding region carries information for protein synthesis encoded by genetic code into form of triplets. Each triplet of nucleotides of the coding region is called codon and corresponds to a binding site complementary to an anticodon triplet in transfer RNA. Transfer RNAs with the same anticodon sequence always carry identical type of amino acid. Amino acids are then chained together by the ribosome according to order of triplets in the coding region. The ribosome helps transfer RNA to bind to messenger RNA and takes the amino acid from each transfer RNA and makes a structure-less protein out of it.

In prokaryotes translation generally occurs at the point of transcription (co-transcriptionally), often using a messenger RNA which is still in the process of being created. In eukaryotes translation can occur in a variety of regions of the cell depending on where the protein being written is supposed to be. Major locations are the cytoplasm for soluble cytoplasmic proteins and the membrane of endoplasmic reticulum for proteins which are for export from the cell or insertion into a cell membrane. Proteins which are supposed to be expressed at the endoplasmic reticulum are recognised part-way through the translation process. This is governed by the signal recognition particle - a protein which binds to the ribosome and directs it to the endoplasmic reticulum when it finds a signal sequence on the growing (nascent) amino acid chain.

Folding



Protein before (left) and after (right) folding

The polypeptide folds into its characteristic and functional three-dimensional structure from random coil. Each protein exists as an unfolded polypeptide or random coil when translated from a sequence of mRNA to a linear chain of amino acids. This polypeptide lacks any developed three-dimensional structure (the left hand side of the neighboring figure). Amino acids interact with each other to produce a well-defined three-dimensional structure, the folded protein (the right hand side of the figure), known as the native state.

The resulting three-dimensional structure is determined by the amino acid sequence (Anfinsen's dogma).

The correct three-dimensional structure is essential to function, although some parts of functional proteins may remain unfolded Failure to fold into the intended shape usually produces inactive proteins with different properties including toxic prions. Several neurodegenerative and other diseases are believed to result from the accumulation of *misfolded* (incorrectly folded) proteins. Many allergies are caused by the folding of the proteins, for the immune system does not produce antibodies for certain protein structures.

Enzymes called chaperones assist the newly formed protein to attain (fold into) the 3-dimensional structure it needs to function. Similarly, RNA chaperones help RNAs attain their functional shapes. Assisting protein folding is one of the main roles of the endoplasmic reticulum in eukaryotes.

Protein transport

Many proteins are destined for other parts of the cell than the cytosol and a wide range of signalling sequences are used to direct proteins to where they are supposed to be. In prokaryotes this is normally a simple process due to limited compartmentalisation of the cell. However in eukaryotes there is a great variety of different targeting processes to ensure the protein arrives at the correct organelle.

Not all proteins remain within the cell and many are exported, for example digestive enzymes, hormones and extracellular matrix proteins. In eukaryotes the export pathway is well developed and the main mechanism for the export of these proteins is translocation to the endoplasmic reticulum, followed by transport via the Golgi apparatus.

Regulation of gene expression

The patchy colours of a tortoiseshell cat are the result of different levels of expression of pigmentation genes in different areas of the skin.

Regulation of gene expression refers to the control of the amount and timing of appearance of the functional product of a gene. Control of expression is vital to allow a cell to produce the gene products it needs when it needs them; in turn this gives cells the flexibility to adapt to a variable environment, external signals, damage to the cell, etc. Some simple examples of where gene expression is important are:

- Control of Insulin expression so it gives a signal for blood glucose regulation
- X chromosome inactivation in female mammals to prevent an "overdose" of the genes it contains.
- Cyclin expression levels control progression through the eukaryotic cell cycle

More generally gene regulation gives the cell control over all structure and function, and is the basis for cellular differentiation, morphogenesis and the versatility and adaptability of any organism.

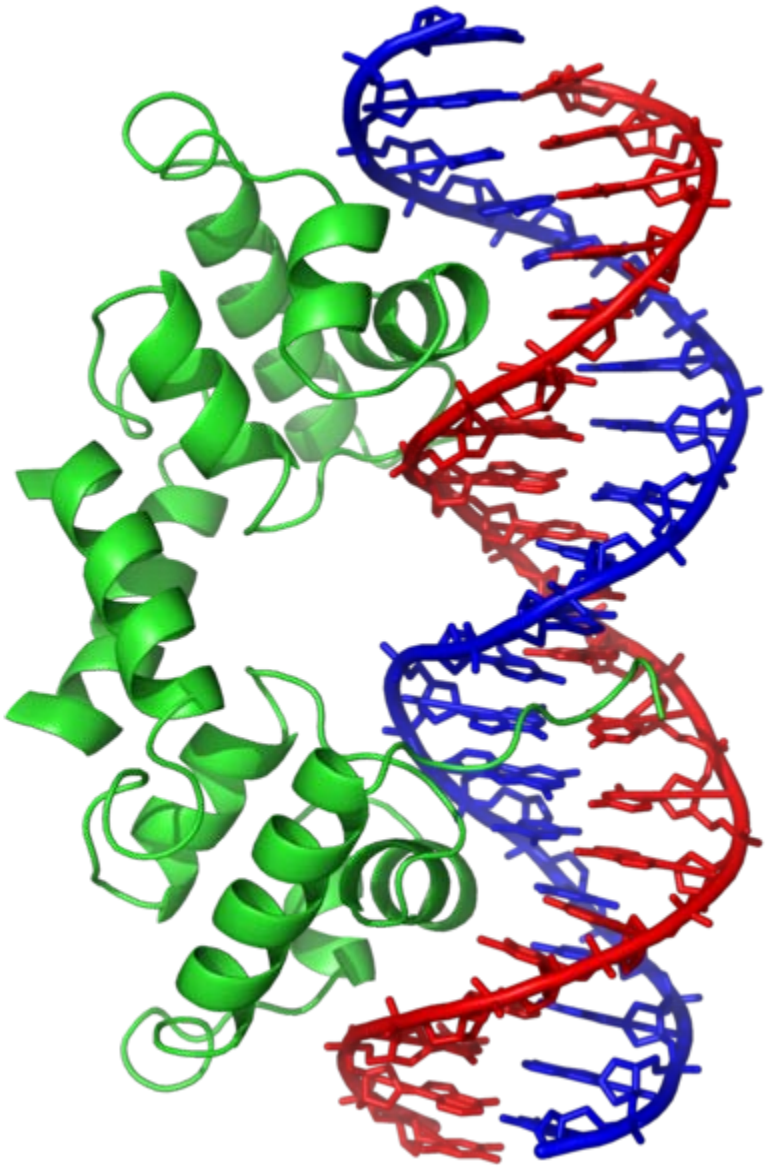
Any step of gene expression may be modulated, from the DNA-RNA transcription step to post-translational modification of a protein. The stability of the final gene product, whether it is RNA or protein, also contributes to the expression level of the gene - an unstable product results in a low expression level. In general gene expression is regulated through changes in the number and type of interactions between molecules that collectively influence transcription of DNA and translation of RNA.

Numerous terms are used to describe types of genes depending on how they are regulated, these include:

- A *constitutive gene* is a gene that is transcribed continually compared to a facultative gene which is only transcribed when needed.
- A *housekeeping gene* is typically a constitutive gene that is transcribed at a relatively constant level. The housekeeping gene's products are typically needed for maintenance of the cell. It is generally assumed that their expression is unaffected by experimental conditions. Examples include actin, GAPDH and ubiquitin.
- A *facultative gene* is a gene which is only transcribed when needed compared to a constitutive gene.
- An *inducible gene* is a gene whose expression is either responsive to environmental change or dependent on the position in the cell cycle.

Transcriptional regulation

Regulation of transcription can be broken down into three main routes of influence; genetic (direct interaction of a control factor with the gene), modulation (interaction of a control factor with the transcription machinery) and epigenetic (non-sequence changes in DNA structure which influence transcription).



The lambda repressor transcription factor (green) binds as a dimer to major groove of DNA target (red and blue) and disables initiation of transcription. From PDB 1LMB.

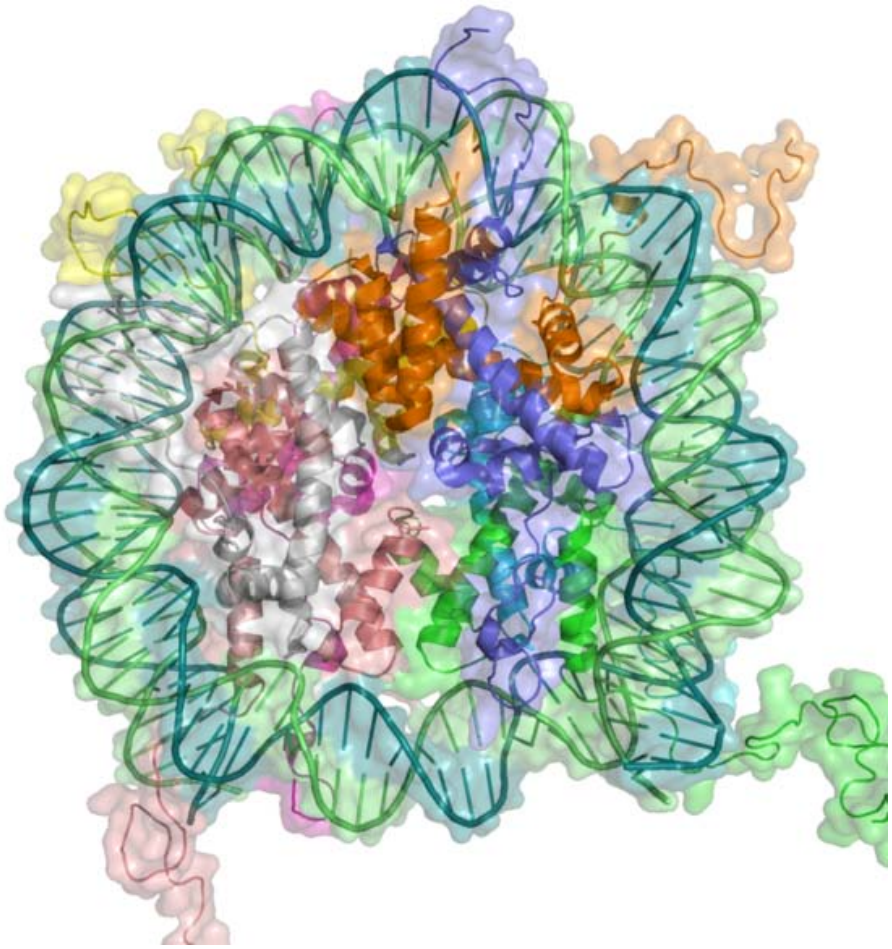
Direct interaction with DNA is the simplest and the most direct method by which a protein can change transcription levels. Genes often have several protein binding sites around the coding region with the specific function of regulating transcription. There are many classes of regulatory DNA binding sites known as enhancers, insulators, repressors and silencers. The mechanisms for regulating transcription are very varied, from blocking key binding sites on the DNA for RNA polymerase to acting as an activator and promoting transcription by assisting RNA polymerase binding.

The activity of transcription factors is further modulated by intracellular signals causing protein post-translational modification including phosphorylated, acetylated, or

glycosylated. These changes influence a transcription factor's ability to bind, directly or indirectly, to promoter DNA, to recruit RNA polymerase, or to favor elongation of a newly synthetized RNA molecule.

The nuclear membrane in eukaryotes allows further regulation of transcription factors by the duration of their presence in the nucleus which is regulated by reversible changes in their structure and by binding of other proteins. Environmental stimuli or endocrine signals may cause modification of regulatory proteins eliciting cascades of intracellular signals, which result in regulation of gene expression.

More recently it has become apparent that there is a huge influence of non-DNA-sequence specific effects on translation. These effects are referred to as epigenetic and involve the higher order structure of DNA, non-sequence specific DNA binding proteins and chemical modification of DNA. In general epigenetic effects alter the accessibility of DNA to proteins and so modulate transcription.



In eukaryotes, DNA is organized in form of nucleosomes. Note how the DNA (blue and green) is tightly wrapped around the protein core made of histone octamer (ribbon coils), restricting access to the DNA. From PDB 1KX5.

DNA methylation is a widespread mechanism for epigenetic influence on gene expression and is seen in bacteria and eukaryotes and has roles in heritable transcription silencing and transcription regulation. In eukaryotes the structure of chromatin, controlled by the histone code, regulates access to DNA with significant impacts on the expression of genes in euchromatin and heterochromatin areas.

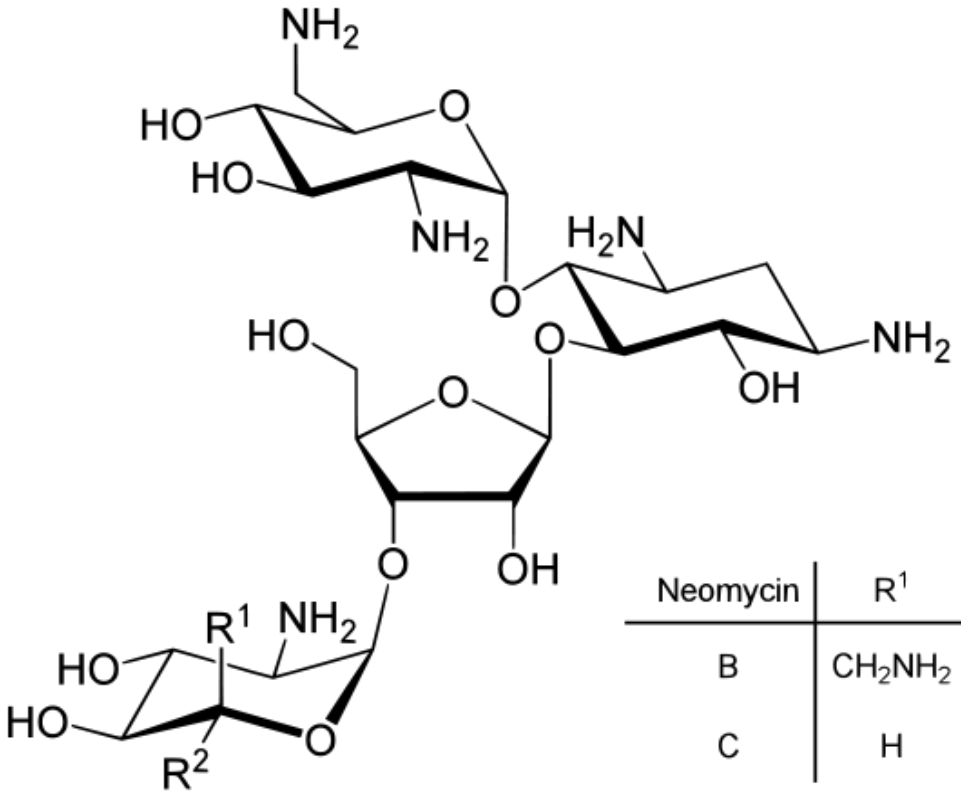
Post-transcriptional regulation

In eukaryotes, where export of RNA is required before translation is possible, nuclear export is thought to provide additional control over gene expression. All transport in and out of the nucleus is via the nuclear pore and transport is controlled by a wide range of importin and exportin proteins.

Expression of a gene coding for a protein is only possible if the messenger RNA carrying the code survives long enough to be translated. In a typical cell an RNA molecule is only stable if specifically protected from degradation. RNA degradation has particular importance in regulation of expression in eukaryotic cells where mRNA has to travel significant distances before being translated. In eukaryotes RNA is stabilised by certain post-transcriptional modifications, particularly the 5' cap and poly-adenylated tail.

Intentional degradation of mRNA is used not just as a defence mechanism from foreign RNA (normally from viruses) but also as a route of mRNA *destabilisation*. If an mRNA molecule has a complementary sequence to a small interfering RNA then it is targeted for destruction via the RNA interference pathway.

Translational regulation



Neomycin	R^1	R^2
B	CH_2NH_2	H
C	H	CH_2NH_2

Neomycin is an example of a small molecule which reduces expression of all protein genes inevitably leading to cell death, thus acts as an antibiotic.

Direct regulation of translation is less prevalent than control of transcription or mRNA stability but is occasionally used. Inhibition of protein translation is a major target for toxins and antibiotics in order to kill a cell by overriding its normal gene expression control. Protein synthesis inhibitors include the antibiotic neomycin and the toxin ricin.

Protein degradation

Once protein synthesis is complete the level of expression of that protein can be reduced by protein degradation. There are major protein degradation pathways in all prokaryotes and eukaryotes of which the proteasome is a common component. An unneeded or damaged protein is often labelled for degradation by addition of ubiquitin.

Measurement

Measuring gene expression is an important part of many life sciences - the ability to quantify the level at which a particular gene is expressed within a cell, tissue or organism can give a huge amount of information. For example measuring gene expression can:

- Identify viral infection of a cell (viral protein expression)
- Determine an individual's susceptibility to cancer (oncogene expression)
- Find if a bacterium is resistant to penicillin (beta-lactamase expression)

Similarly the analysis of the location of expression protein is a powerful tool and this can be done on an organism or cellular scale. Investigation of localisation is particularly important for study of development in multicellular organisms and as an indicator of protein function in single cells. Ideally measurement of expression is done by detecting the final gene product (for many genes this is the protein) however it is often easier to detect one of the precursors, typically mRNA, and infer gene expression level.

mRNA quantification

Levels of mRNA can be quantitatively measured by Northern blotting which gives size and sequence information about the mRNA molecules. A sample of RNA is separated on an agarose gel and hybridized to a radio-labeled RNA probe that is complementary to the target sequence. The radio-labeled RNA is then detected by an autoradiograph. The main problems with Northern blotting stem from the use of radioactive reagents (which make the procedure time consuming and potentially dangerous) and lower quality quantification than more modern methods (due to the fact that quantification is done by measuring band strength in an image of a gel). Northern blotting is, however, still widely used as the additional mRNA size information allows the discrimination of alternately spliced transcripts.

A more modern low-throughput approach for measuring mRNA abundance is reverse transcription quantitative polymerase chain reaction (RT-PCR followed with qPCR). RT-PCR first generates a DNA template from the mRNA by reverse transcription, which is called cDNA. This cDNA template is then used for qPCR where the change in fluorescence of a probe changes as the DNA amplification process progresses. With a carefully constructed standard curve qPCR can produce an absolute measurement such as number of copies of mRNA, typically in units of copies per nanolitre of homogenized tissue or copies per cell. qPCR is very sensitive (detection of a single mRNA molecule is possible), but can be expensive due to the fluorescent probes required.

Northern blots and RT-qPCR are good for detecting whether a single gene is being expressed, but it quickly becomes impractical if many genes within the sample are being studied. Using DNA microarrays transcript levels for many genes at once (expression profiling) can be measured. Recent advances in microarray technology allow for the quantification, on a single array, of transcript levels for every known gene in several organism's genomes, including humans.

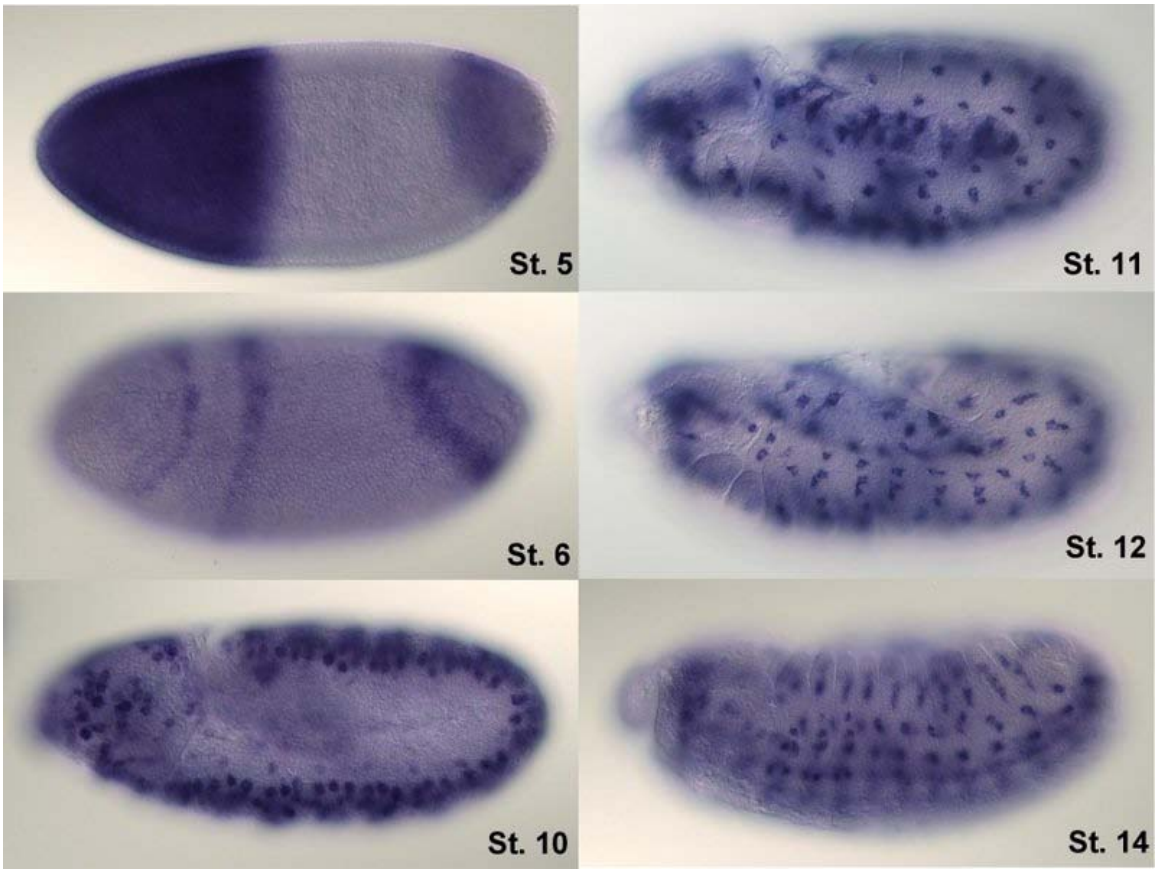
Alternatively "tag based" technologies like Serial analysis of gene expression (SAGE), which can provide a relative measure of the cellular concentration of different mRNAs, can be used. The great advantage of tag-based methods is the "open architecture", allowing for the exact measurement of any transcript, with a known or unknown sequence.

Protein quantification

For genes encoding proteins the expression level can be directly assessed by a number of means with some clear analogies to the techniques for mRNA quantification.

The most commonly used method is to perform a Western blot against the protein of interest - this gives information on the size of the protein in addition to its identity. A sample (often cellular lysate) is separated on a polyacrylamide gel, transferred to a membrane and then probed with an antibody to the protein of interest. The antibody can either be conjugated to a fluorophore or to horseradish peroxidase for imaging and/or quantification. The gel-based nature of this assay makes quantification less accurate but it has the advantage of being able to identify later modifications to the protein, for example proteolysis or ubiquitination, from changes in size.

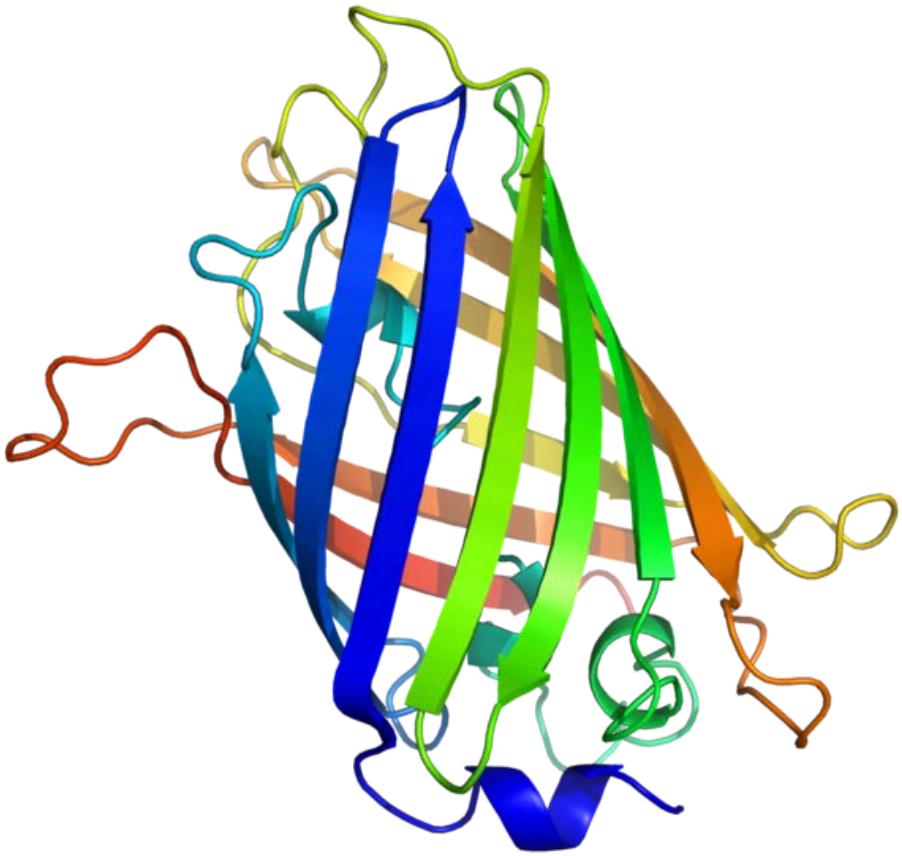
Localisation



In situ-hybridization of Drosophila embryos at different developmental stages for the mRNA responsible for the expression of hunchback. High intensity of blue color marks places with high hunchback mRNA quantity.

Analysis of expression is not limited to only quantification; localisation can also be determined. mRNA can be detected with a suitably labelled complementary mRNA

strand and protein can be detected via labelled antibodies. The probed sample is then observed by microscopy to identify where the mRNA or protein is.



The three-dimensional structure of green fluorescent protein. The residues in the centre of the "barrel" are responsible for production of green light after exposing to higher energetic blue light. From PDB 1EMA.

By replacing the gene with a new version fused a green fluorescent protein (or similar) marker expression may be directly quantified in live cells. This is done by imaging using a fluorescence microscope. It is very difficult to clone a GFP-fused protein into its native location in the genome without affecting expression levels so this method often cannot be used to measure endogenous gene expression. It is, however, widely used to measure the expression of a gene artificially introduced into the cell, for example via an expression vector. It is important to note that by fusing a target protein to a fluorescent reporter the protein's behavior, including its cellular localization and expression level, can be significantly changed.

The enzyme-linked immunosorbent assay works by using antibodies immobilised on a microtiter plate to capture proteins of interest from samples added to the well. Using a detection antibody conjugated to an enzyme or fluorophore the quantity of bound protein

can be accurately measured by fluorometric or colourimetric detection. The detection process is very similar to that of a Western blot, but by avoiding the gel steps more accurate quantification can be achieved.

Expression system

An expression system is a system specifically designed for the production of a gene product of choice. This is normally a protein although may also be RNA, such as tRNA or a ribozyme. An expression system consists of a gene, normally encoded by DNA, and the molecular machinery required to transcribe the DNA into mRNA and translate the mRNA into protein using the reagents provided. In the broadest sense this includes every living cell but the term is more normally used to refer to expression as a laboratory tool. An expression system is therefore often artificial in some manner. Expression systems are, however, a fundamentally natural process. Viruses are an excellent example where they replicate by using the host cell as an expression system for the viral proteins and genome.

In nature

In addition to these biological tools, certain naturally observed configurations of DNA (genes, promoters, enhancers, repressors) and the associated machinery itself are referred to as an expression system. This term is normally used in the case where a gene or set of genes is switched on under well defined conditions. For example the simple repressor switch expression system in Lambda phage and the lac operator system in bacteria. Several natural expression systems are directly used or modified and used for artificial expression systems such as the Tet-on and Tet-off expression system.

Gene networks

Genes have sometimes been regarded as nodes in a network, with inputs being proteins such as transcription factors, and outputs being the level of gene expression. The node itself performs a function, and the operation of these functions have been interpreted as performing a kind of information processing within cell and determine cellular behavior.

Gene networks can also be constructed without formulating an explicit causal model. This is often the case when assembling networks from large expression data sets. Covariation and correlation of expression is computed across a large sample of cases and measurements (often transcriptome or proteome data). The source of variation can be either experimental or natural (observational). There are several ways to construct gene expression networks, but one common approach is to compute a matrix of all pair-wise correlations of expression across conditions, time points, or individuals and convert the matrix (after thresholding at some cut-off value) into a graphical representation in which nodes represent genes, transcripts, or proteins and edges connecting these nodes represent the strength of association.

Techniques and tools

The following experimental techniques are used to measure gene expression and are listed in roughly chronological order, starting with the older, more established technologies. They are divided into two groups based on their degree of multiplexity.

- Low-to-mid-plex techniques:
  - Reporter gene
  - Northern blot
  - Western blot
  - Fluorescent in situ hybridization
  - Reverse transcription PCR

- Higher-plex techniques:
  - SAGE
  - DNA microarray
  - Tiling array
  - RNA-Seq

Chapter- 8

Homologous Recombination

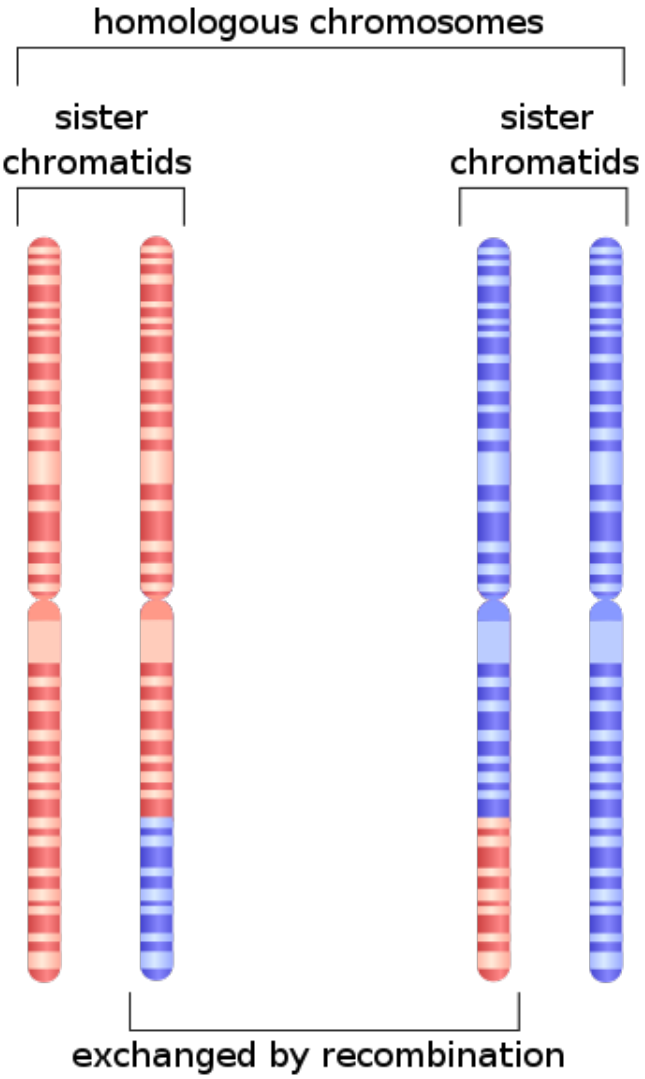


Figure 1. During meiosis, homologous recombination can produce new combinations of genes as shown here between similar but not identical copies of human chromosome 1.

Homologous recombination is a type of genetic recombination in which nucleotide sequences are exchanged between two similar or identical molecules of DNA. It is most widely used by cells to accurately repair harmful breaks that occur on both strands of DNA, known as double-strand breaks. Homologous recombination also produces new combinations of DNA sequences during meiosis, the process by which eukaryotes make gamete cells, like sperm and egg cells in animals. These new combinations of DNA represent genetic variation in offspring, which in turn enables populations to adapt during the course of evolution. Homologous recombination is also used in horizontal gene transfer to exchange genetic material between different strains and species of bacteria and viruses.

Although homologous recombination varies widely among different organisms and cell types, most forms of it involve the same basic steps. After a double-strand break occurs, sections of DNA around the 5' ends of the break are cut away in a process called *resection*. In the *strand invasion* step that follows, an overhanging 3' end of the broken DNA molecule then "invades" a similar or identical DNA molecule that is not broken. After strand invasion, one or two cross-shaped structures called Holliday junctions connect the two DNA molecules. Depending on how the two junctions are cut by enzymes, the type of homologous recombination that occurs in meiosis results in either chromosomal crossover or non-crossover. Homologous recombination that occurs during DNA repair tends to result in non-crossover products, in effect restoring the damaged DNA molecule as it existed before the double-strand break.

Homologous recombination is conserved across all three domains of life as well as viruses, suggesting that it is a nearly universal biological mechanism. The discovery of genes for homologous recombination in protists—a diverse group of eukaryotic microorganisms—has been interpreted as evidence that meiosis emerged early in the evolution of eukaryotes. Since their dysfunction has been strongly associated with increased susceptibility to several types of cancer, the proteins that facilitate homologous recombination are topics of active research. Homologous recombination is also used as a technique in molecular biology for introducing genetic changes into target organisms. The development of gene targeting techniques that rely on homologous recombination was the subject of the 2007 Nobel Prize for Physiology or Medicine.

History and discovery

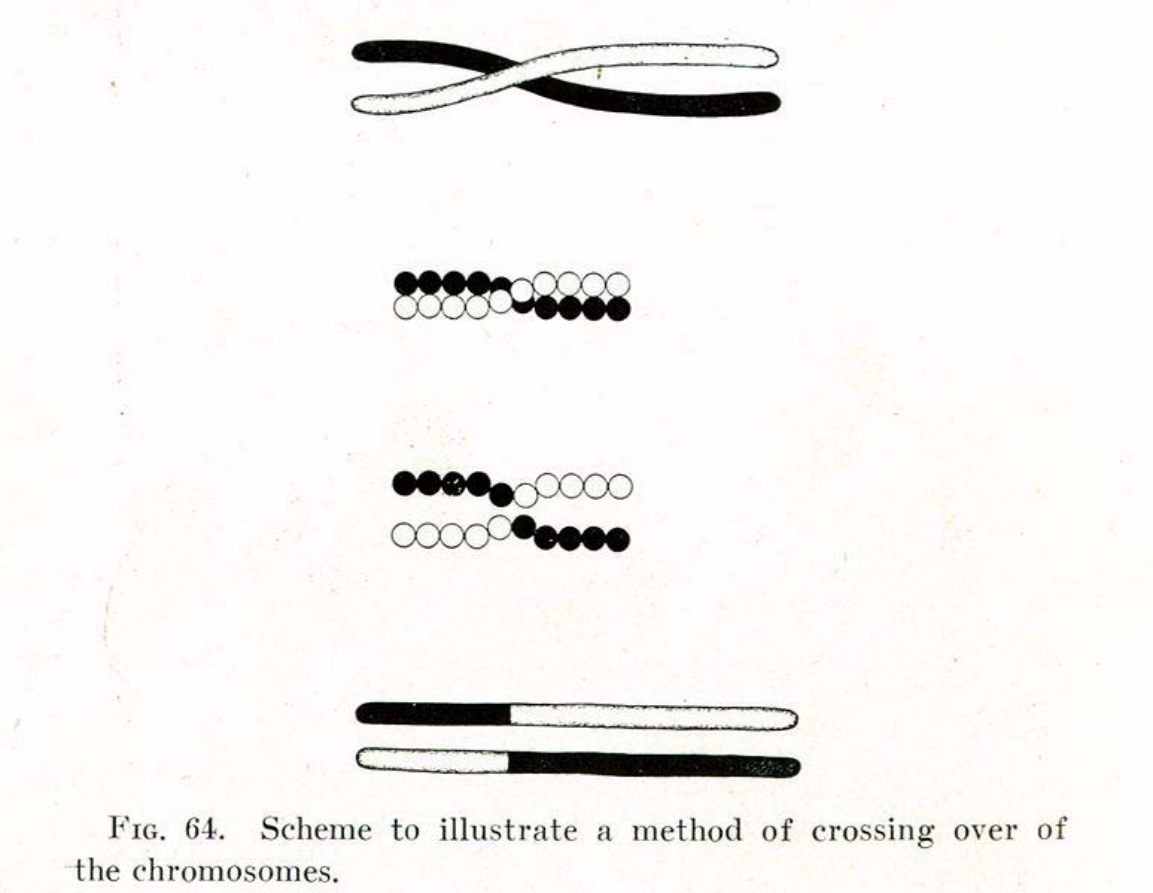


FIG. 64. Scheme to illustrate a method of crossing over of the chromosomes.

Figure 2. An early illustration of crossing over from Thomas Hunt Morgan

In the early 1900s, William Bateson and Reginald Punnett found an exception to one of the principles of inheritance originally described by Gregor Mendel in the 1860s. In contrast to Mendel's notion that traits are independently assorted when passed from parent to child—for example that a cat's hair color and its tail length are inherited independent of each other—Bateson and Punnett showed that certain genes associated with physical traits can be inherited together, or genetically linked. In 1911, after observing that linked traits could on occasion be inherited separately, Thomas Hunt Morgan suggested that "crossovers" can occur between linked genes, where one of the linked genes physically crosses over to a different chromosome. Two decades later, Barbara McClintock and Harriet Creighton demonstrated that chromosomal crossover occurs during meiosis, the process of cell division by which sperm and egg cells are made. Within the same year as McClintock's discovery, Curt Stern showed that crossing over—later called "recombination"—could also occur in somatic cells like white blood cells and skin cells that divide through mitosis.

In 1947, the microbiologist Joshua Lederberg showed that bacteria—which had been assumed to reproduce only asexually through binary fission—are capable of genetic

recombination, which is more similar to sexual reproduction. This work established *E. coli* as a model organism in genetics, and helped Lederberg win the 1958 Nobel Prize in Physiology or Medicine. Building on studies in fungi, in 1964 Robin Holliday proposed a model for recombination in meiosis which introduced key details of how the process can work, including the exchange of material between chromosomes through Holliday junctions. In 1983, Jack Szostak and colleagues presented a model now known as the DSBR pathway, which accounted for observations not explained by the Holliday model. During the next decade, experiments in *Drosophila*, budding yeast and mammalian cells led to the emergence of other models of homologous recombination, called SDSA pathways, which do not always rely on Holliday junctions.

In eukaryotes

Homologous recombination is essential to cell division in eukaryotes like plants, animals, fungi and protists. In cells that divide through mitosis, homologous recombination repairs double-strand breaks in DNA caused by ionizing radiation or DNA-damaging chemicals. Left unrepaired, these double-strand breaks can cause large-scale rearrangement of chromosomes in somatic cells, which can in turn lead to cancer.

In addition to repairing DNA, homologous recombination also helps produce genetic diversity when cells divide in meiosis to become specialized gamete cells—sperm or egg cells in animals, pollen or ovules in plants, and spores in fungi. It does so by facilitating chromosomal crossover, in which regions of similar but not identical DNA are exchanged between homologous chromosomes. This creates new, possibly beneficial combinations of genes, which can give offspring an evolutionary advantage. Chromosomal crossover begins when a protein called Spo11 makes a targeted double-strand break in DNA. The sites of these double-strand breaks often occur at recombination hotspots, regions in chromosomes that are about 1,000–2,000 base pairs in length and have high rates of recombination. The absence of a recombination hotspot between two genes on the same chromosome often means that those genes will be inherited by future generations in equal proportion. This represents linkage between the two genes greater than would be expected from genes that independently assort during meiosis.

Timing within the cell cycle

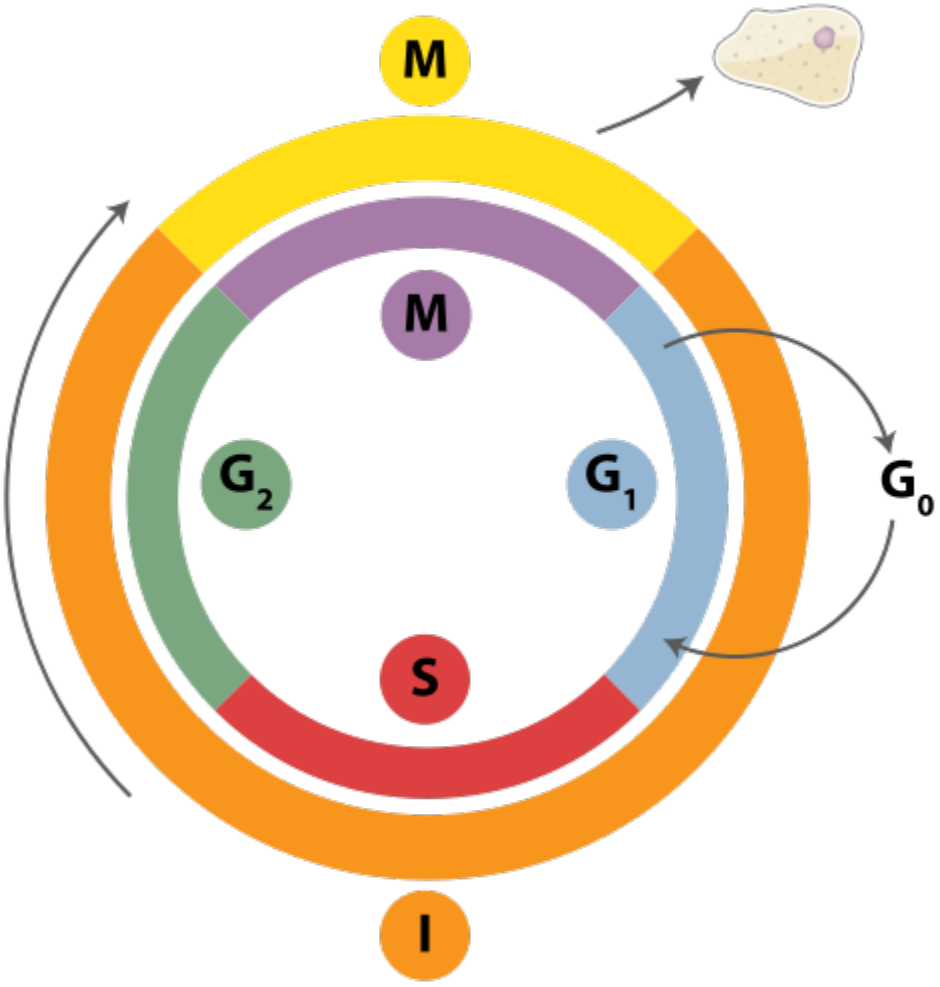


Figure 3. Homologous recombination repairs DNA before the cell enters mitosis (M phase). It occurs only during and shortly after DNA replication, during the S and G_2 phases of the cell cycle.

Double-strand breaks can be repaired through homologous recombination or through non-homologous end joining (NHEJ). NHEJ is a DNA repair mechanism which, unlike homologous recombination, does not require a long homologous sequence to guide repair. Whether homologous recombination or NHEJ is used to repair double-strand breaks is largely determined by the phase of cell cycle. Homologous recombination repairs DNA before the cell enters mitosis (M phase). It occurs during and shortly after DNA replication, in the S and G_2 phases of the cell cycle, when sister chromatids are more easily available. Compared to homologous chromosomes, which are similar to another chromosome but often have different alleles, sister chromatids are an ideal template for homologous recombination because they are an identical copy of a given chromosome. In contrast to homologous recombination, NHEJ is predominant in the G_1 phase of the cell cycle, when the cell is growing but not yet ready to divide. It occurs less frequently after the G_1 phase, but maintains at least some activity throughout the cell

cycle. The mechanisms that regulate homologous recombination and NHEJ throughout the cell cycle vary widely between species.

Cyclin-dependent kinases (CDKs), which modify the activity of other proteins by adding phosphate groups to (that is, phosphorylating) them, are important regulators of homologous recombination in eukaryotes. When DNA replication begins in budding yeast, the cyclin-dependent kinase Cdc28 begins homologous recombination by phosphorylating the Sae2 protein. After being so activated by the addition of a phosphate, Sae2 uses its endonuclease activity to make a clean cut near a double-strand break in DNA. This allows a three-part protein known as the MRX complex to bind to DNA, and begins a series of protein-driven reactions that exchange material between two DNA molecules.

Models

Two primary models for how homologous recombination repairs double-strand breaks in DNA are the DSBR pathway (sometimes called the *double Holliday junction model*) and the synthesis-dependent strand annealing (SDSA) pathway. The two pathways are similar in their first several steps. After a double-strand break occurs, the MRX complex (MRN complex in humans) binds to DNA on either side of the break. Next a resection, in which DNA around the 5' ends of the break is cut back, is carried out in two distinct steps. In the first step of resection, the MRX complex recruits the Sae2 protein. The two proteins then trim back the 5' ends on either side of the break to create short 3' overhangs of single-strand DNA. In the second step, 5'→3' resection is continued by the Sgs1 helicase and the Exo1 and Dna2 nucleases. As a helicase, Sgs1 "unzips" the double-strand DNA, while Exo1 and Dna2's nuclease activity allows them to cut the single-stranded DNA produced by Sgs1.

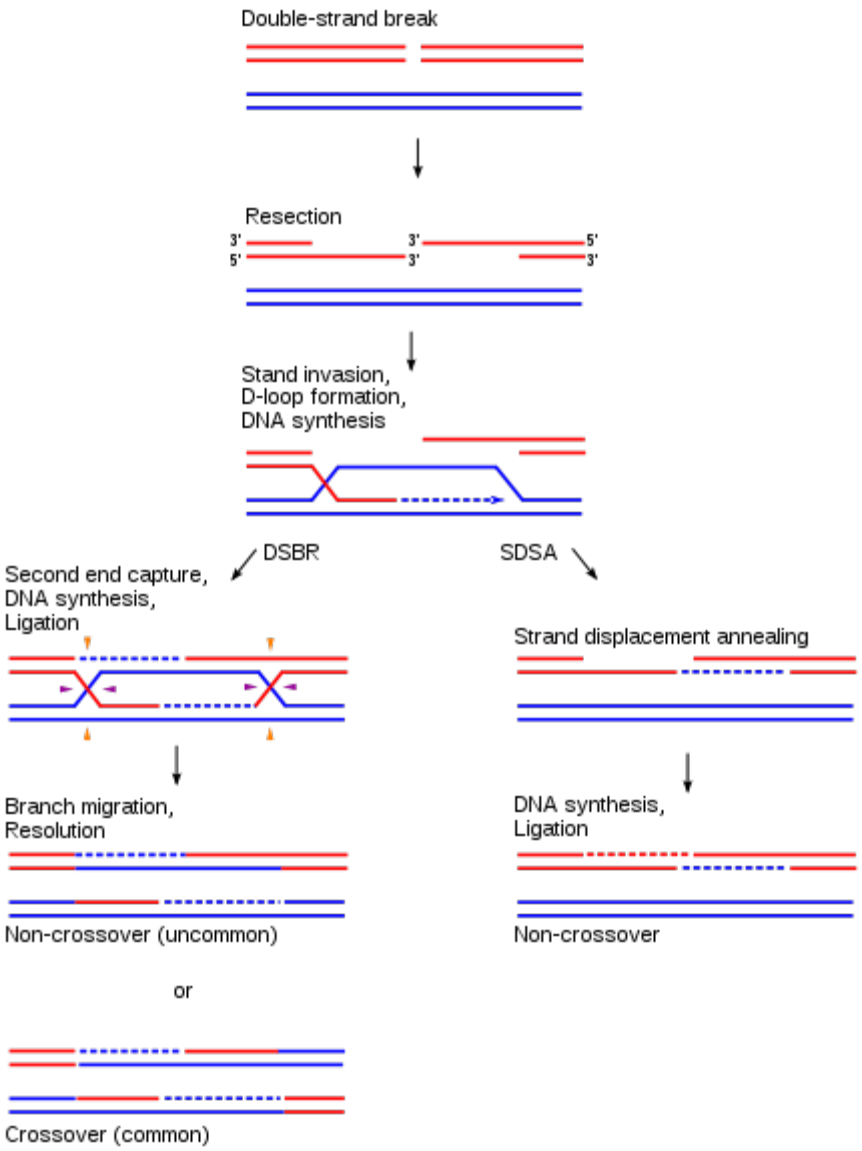


Figure 4. The DSBR and SDSA pathways follow the same initial steps, but diverge thereafter. The DSBR pathway most often results in chromosomal crossover (bottom left), while SDSA always ends with non-crossover products (bottom right).

The RPA protein, which has high affinity for single-stranded DNA, then binds the 3' overhangs. With the help of several other proteins that mediate the process, the Rad51 protein (and Dmc1, in meiosis) then forms a filament of nucleic acid and protein on the single strand of DNA coated with RPA. This nucleoprotein filament then begins searching for DNA sequences similar to that of the 3' overhang. After finding such a sequence, the single-stranded nucleoprotein filament moves into (invades) the similar or identical recipient DNA duplex in a process called *strand invasion*. In cells that divide through mitosis, the recipient DNA duplex is generally a sister chromatid, which is identical to the damaged DNA molecule and provides a template for repair. In meiosis,

however, the recipient DNA tends to be from a similar but not necessarily identical homologous chromosome. A displacement loop (D-loop) is formed during strand invasion between the invading 3' overhang strand and the homologous chromosome. After strand invasion, a DNA polymerase extends the end of the invading 3' strand by synthesizing new DNA. This changes the D-loop to a cross-shaped structure known as a Holliday junction. Following this, more DNA synthesis occurs on the invading strand (i.e., one of the original 3' overhangs), effectively restoring the strand on the homologous chromosome that was displaced during strand invasion.

DSBR pathway

After the stages of resection, strand invasion and DNA synthesis, the DSBR and SDSA pathways become distinct. The DSBR pathway is unique in that the second 3' overhang (which was not involved in strand invasion) also forms a Holliday junction with the homologous chromosome. The double Holliday junctions are then converted into recombination products by nicking endonucleases, a type of restriction endonuclease which cuts only one DNA strand. The DSBR pathway commonly results in crossover, though it can sometimes result in non-crossover products. Because of this tendency for chromosomal crossover, the DSBR pathway is a likely model of how homologous recombination occurs during meiosis.

Whether recombination in the DSBR pathway results in chromosomal crossover is determined by how the double Holliday junction is cut, or "resolved". Chromosomal crossover will occur if one Holliday junction is cut on the crossing strand and the other Holliday junction is cut on the non-crossing strand (in Figure 4, along the horizontal purple arrowheads at one Holliday junction and along the vertical orange arrowheads at the other). Alternatively, if the two Holliday junctions are cut on the crossing strands (along the horizontal purple arrowheads at both Holliday junctions in Figure 4), then chromosomes without crossover will be produced.

SDSA pathway

Homologous recombination via the SDSA pathway occurs in cells that divide through mitosis and results in non-crossover products. In this model, the invading 3' strand is extended along the recipient DNA duplex by a DNA polymerase, and is released as the Holliday junction between the donor and recipient DNA molecules slides in a process called *branch migration*. The newly synthesized 3' end of the invading strand is then able to anneal to the other 3' overhang in the damaged chromosome through complementary base pairing. After the strands anneal, a small flap of DNA can sometimes remain. Any such flaps are removed, and the SDSA pathway finishes with the resealing, also known as *ligation*, of any remaining single-stranded gaps.

SSA pathway

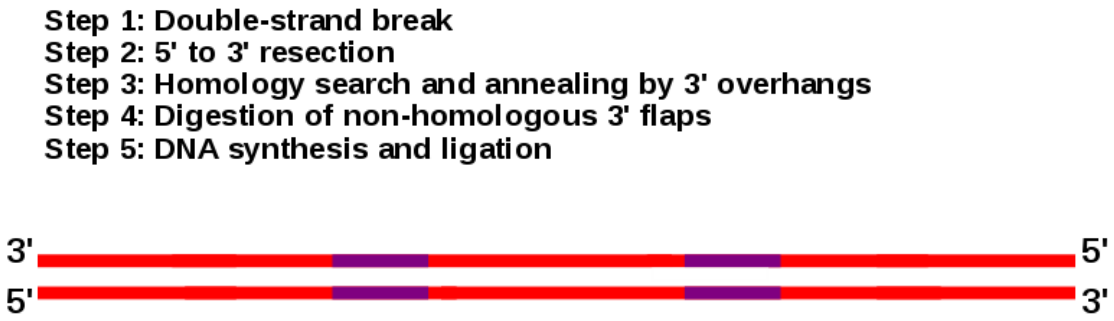


Figure 5. Recombination via the SSA pathway occurs between two repeat elements (purple) on the same DNA duplex, and results in deletions of genetic material.

The single-strand annealing (SSA) pathway of homologous recombination repairs double-strand breaks between two repeat sequences. The SSA pathway is unique in that it does not require a separate similar or identical molecule of DNA, like the DSBR or SDSA pathways of homologous recombination. Instead, the SSA pathway only requires a single DNA duplex, and uses the repeat sequences as the identical sequences that homologous recombination needs for repair. The pathway is relatively simple in concept: after two strands of the same DNA duplex are cut back around the site of the double-strand break, the two resulting 3' overhangs then align and anneal to each other, restoring the DNA as a continuous duplex.

As DNA around the double-strand break is cut back, the single-stranded 3' overhangs being produced are coated with the RPA protein, which prevents the 3' overhangs from sticking to themselves. A protein called Rad52 then binds each of the repeat sequences on either side of the break, and aligns them to enable the two complementary repeat sequences to anneal. After annealing is complete, leftover non-homologous flaps of the 3' overhangs are cut away by a set of nucleases, known as Rad1/Rad10, which are brought to the flaps by the Saw1 and Slx4 proteins. New DNA synthesis fills in any gaps, and ligation restores the DNA duplex as two continuous strands. The DNA sequence between the repeats is always lost, as is one of the two repeats. The SSA pathway is considered mutagenic since it results in such deletions of genetic material.

BIR pathway

During DNA replication, double-strand breaks can sometimes be encountered at replication forks as DNA helicase unzips the template strand. These defects are repaired in the *break-induced replication* (BIR) pathway of homologous recombination. The precise molecular mechanisms of the BIR pathway remain unclear. Three proposed mechanisms have strand invasion as an initial step, but differ in how they model the migration of the D-loop and later phases of recombination.

The BIR pathway can also help to maintain the length of telomeres, regions of DNA at the end of eukaryotic chromosomes, in the absence of (or in cooperation with) telomerase. Without working copies of the telomerase enzyme, telomeres typically shorten with each cycle of mitosis, which eventually blocks cell division and leads to senescence. In budding yeast cells where telomerase have been inactivated through mutations, two types of "survivor" cells have been observed to avoid senescence longer than expected by elongating their telomeres through BIR pathways.

Maintaining telomere length is critical for cell immortalization, a key feature of cancer. Most cancers maintain telomeres by upregulating telomerase. However, in several types of human cancer, a BIR-like pathway helps to sustain some tumors by acting as an alternative mechanism of telomere maintenance. This fact has lead scientists to investigate whether such recombination-based mechanisms of telomere maintenance could thwart anti-cancer drugs like telomerase inhibitors.

In bacteria

Figure 6. Crystal structure of two chains of the RecA protein bound to DNA. A double-strand break and two adjacent 3' overhangs are visible.

Homologous recombination is a major DNA repair process in bacteria. It is also important for producing genetic diversity in bacterial populations, although the process differs substantially from meiotic recombination, which brings about diversity in eukaryotic genomes. Homologous recombination has been most studied and is best understood for *Escherichia coli*. Double-strand DNA breaks in bacteria are repaired by the RecBCD pathway of homologous recombination. Breaks that occur on only one of the two DNA strands, known as single-strand gaps, are thought to be repaired by the RecF pathway. Both the RecBCD and RecF pathways include a series of reactions known as *branch migration*, in which single DNA strands are exchanged between two intercrossed molecules of duplex DNA, and *resolution*, in which those two intercrossed molecules of DNA are cut apart and restored to their normal double-stranded state.

RecBCD pathway

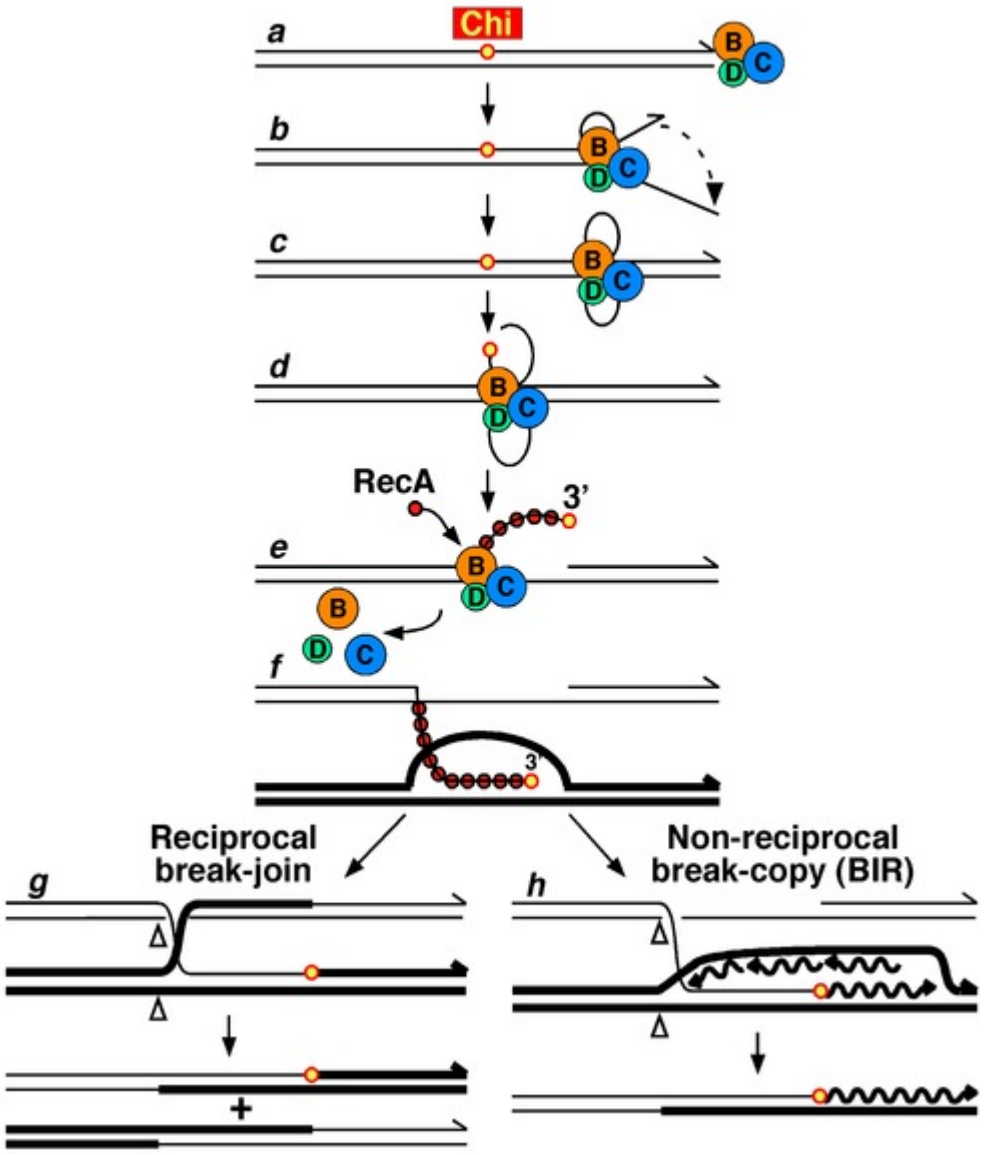


Figure 7A. Molecular model for the RecBCD pathway of recombination. This model is based on reactions of DNA and RecBCD with ATP in excess over Mg^{2+} ions. Step a: RecBCD binds to a double-stranded DNA end. Step b: RecBCD unwinds DNA. RecD is a fast helicase on the 5'-ended strand, and RecB is a slower helicase on the 3'-ended strand (that with an arrowhead). This produces two single-stranded (ss) DNA tails and one ss loop. The loop and tails enlarge as RecBCD moves along the DNA. Step c: The two tails anneal to produce a second ss DNA loop, and both loops move and grow. Step d: Upon reaching the Chi hotspot sequence RecBCD nicks the 3'-ended strand. Further unwinding produces a long 3'-ended ss tail with Chi near its end. Step e: RecBCD loads RecA protein onto the Chi tail. At some undetermined point, the RecBCD subunits disassemble. Step f: The RecA-ssDNA complex invades an intact homologous duplex DNA to produce a D-loop, which can be resolved into intact, recombinant DNA in two

ways. Step g: The D-loop is cut and anneals with the gap in the first DNA to produce a Holliday junction. Resolution of the Holliday junction (cutting, swapping of strands, and ligation) at the open arrowheads by some combination of RuvABC and RecG produces two recombinants of reciprocal type. Step h: The 3' end of the Chi tail primes DNA synthesis, from which a replication fork can be generated. Resolution of the fork at the open arrowheads produces one recombinant (non-reciprocal) DNA, one parental-type DNA, and one DNA fragment.

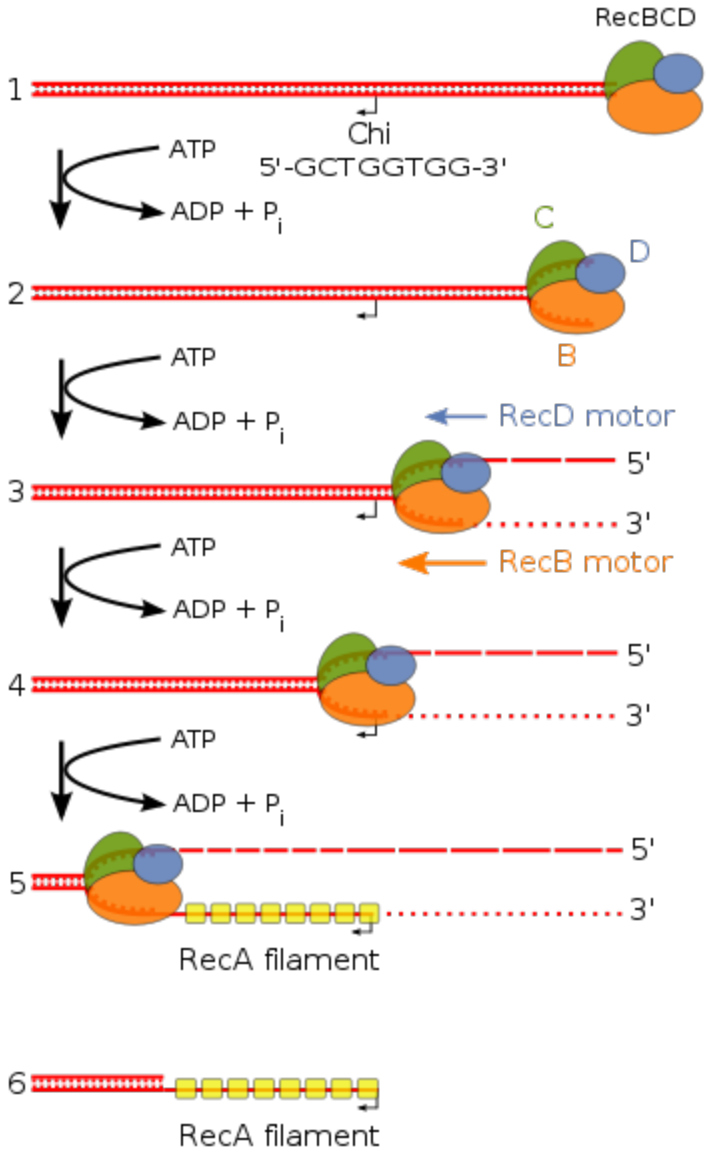


Figure 7B. Beginning of the RecBCD pathway. This model is based on reactions of DNA and RecBCD with Mg^{2+} ions in excess over ATP. Step 1: RecBCD binds to a DNA double strand break. Step 2: RecBCD initiates unwinding of the DNA duplex through ATP-dependent helicase activity. Step 3: RecBCD continues its unwinding and moves down the DNA duplex, cleaving the 3' strand much more frequently than the 5' strand.

Step 4: RecBCD encounters a Chi sequence and stops digesting the 3' strand; cleavage of the 5' strand is significantly increased. Step 5: RecBCD loads RecA onto the 3' strand. Step 6: RecBCD unbinds from the DNA duplex, leaving a RecA nucleoprotein filament on the 3' tail.

The RecBCD pathway is the main recombination pathway used in bacteria to repair double-strand breaks in DNA. These double-strand breaks can be caused by UV light and other radiation, as well as chemical mutagens. Double-strand breaks may also arise by DNA replication through a single-strand nick or gap. Such a situation causes what is known as a collapsed replication fork and is fixed by several pathways of homologous recombination including the RecBCD pathway.

In this pathway, a three-subunit enzyme complex called RecBCD initiates recombination by binding to a blunt or nearly blunt end of a break in double-strand DNA. After RecBCD binds the DNA end, the RecB and RecD subunits begin unzipping the DNA duplex through helicase activity. The RecB subunit also has a nuclease domain, which cuts the single strand of DNA that emerges from the unzipping process. This unzipping continues until RecBCD encounters a specific nucleotide sequence (5'-GCTGGTGG-3') known as a Chi site.

Upon encountering a Chi site, the activity of the RecBCD enzyme changes drastically. DNA unwinding pauses for a few seconds and then resumes at roughly half the initial speed. This is likely because the slower RecB helicase unwinds the DNA after Chi, rather than the faster RecD helicase, which unwinds the DNA before Chi. Recognition of the Chi site also changes the RecBCD enzyme so that it cuts the DNA strand with Chi and begins loading multiple RecA proteins onto the single-stranded DNA with the newly generated 3' end. The resulting RecA-coated nucleoprotein filament then searches out similar sequences of DNA on a homologous chromosome. Upon finding such a sequence, the single-stranded nucleoprotein filament moves into the homologous recipient DNA duplex in a process called *strand invasion*. The invading 3' overhang causes one of the strands of the recipient DNA duplex to be displaced, to form a D-loop. If the D-loop is cut, another swapping of strands forms a cross-shaped structure called a Holliday junction. Resolution of the Holliday junction by some combination of RuvABC or RecG can produce two recombinant DNA molecules with reciprocal genetic types, if the two interacting DNA molecules differ genetically. Alternatively, the invading 3’ end near Chi can prime DNA synthesis and form a replication fork. This type of resolution produces only one type of recombinant (non-reciprocal).

RecF pathway

Bacteria appear to use the RecF pathway of homologous recombination to repair single-strand gaps in DNA. When the RecBCD pathway is inactivated by mutations and additional mutations inactivate the SbcCD and ExoI nucleases, the RecF pathway can also repair DNA double-strand breaks. In the RecF pathway the RecQ helicase unwinds the DNA and the RecJ nuclease degrades the strand with a 5’ end, leaving the strand with the 3’ end intact. RecA protein binds to this strand and is either aided by the RecF, RecO,

and RecR proteins or stabilized by them. The RecA nucleoprotein filament then searches for a homologous DNA and exchanges places with the identical or nearly identical strand in the homologous DNA.

Although the proteins and specific mechanisms involved in their initial phases differ, the two pathways are similar in that they both require single-stranded DNA with a 3’ end and the RecA protein for strand invasion. The pathways are also similar in their phases of *branch migration*, in which the Holliday junction slides in one direction, and *resolution*, in which the Holliday junctions are cleaved apart by enzymes. The alternative, non-reciprocal type of resolution may also occur by either pathway.

Branch migration

Immediately after strand invasion, the Holliday junction moves along the linked DNA during the branch migration process. It is in this movement of the Holliday junction that base pairs between the two homologous DNA duplexes are exchanged. To catalyze branch migration, the RuvA protein first recognizes and binds to the Holliday junction and recruits the RuvB protein to form the RuvAB complex. Two sets of the RuvB protein, which each form a ring-shaped ATPase, are loaded onto opposite sides of the Holliday junction, where they act as twin pumps that provide the force for branch migration. Between those two rings of RuvB, two sets of the RuvA protein assemble in the center of the Holliday junction such that the DNA at the junction is sandwiched between each set of RuvA. The strands of both DNA duplexes—the "donor" and the "recipient" duplexes—are unwound on the surface of RuvA as they are guided by the protein from one duplex to the other.

Resolution

In the resolution phase of recombination, any Holliday junctions formed by the strand invasion process are cut, thereby restoring two separate DNA molecules. This cleavage is done by RuvAB complex interacting with RuvC, which together form the RuvABC complex. RuvC is an endonuclease that cuts the degenerate sequence 5'-(A/T)TT(G/C)-3'. The sequence is found frequently in DNA, about once every 64 nucleotides. Before cutting, RuvC likely gains access to the Holliday junction by displacing one of the two RuvA tetramers covering the DNA there. Recombination results in either "splice" or "patch" products, depending on how RuvC cleaves the Holliday junction. Splice products are crossover products, in which there is a rearrangement of genetic material around the site of recombination. Patch products, on the other hand, are non-crossover products in which there is no such rearrangement and there is only a "patch" of hybrid DNA in the recombination product.

Facilitating genetic transfer

Homologous recombination is an important method of integrating donor DNA into a recipient organism's genome in horizontal gene transfer, the process by which an organism incorporates foreign DNA from another organism without being the offspring

of that organism. Homologous recombination requires incoming DNA to be highly similar to the recipient genome, and so horizontal gene transfer is usually limited to similar bacteria. Studies in several species of bacteria have established that there is a log-linear decrease in recombination frequency with increasing difference in sequence between host and recipient DNA.

In bacterial conjugation, where DNA is transferred between bacteria through direct cell-to-cell contact, homologous recombination helps integrate foreign DNA into the host genome via the RecBCD pathway. The RecBCD enzyme promotes recombination after DNA is converted from single-strand DNA–in which form it originally enters the bacterium–to double-strand DNA during replication. The RecBCD pathway is also essential for the final phase of transduction, a type of horizontal gene transfer in which DNA is transferred from one bacterium to another by a virus. Foreign, bacterial DNA is sometimes misincorporated in the capsid head of bacteriophage virus particles as DNA is packaged into new bacteriophages during viral replication. When these new bacteriophages infect other bacteria, DNA from the previous host bacterium is injected into the new bacterial host as double-strand DNA. The RecBCD enzyme then incorporates this double-strand DNA into the genome of the new bacterial host.

In viruses

Homologous recombination occurs in several groups of viruses. In DNA viruses such as herpesvirus, recombination occurs through a break-and-rejoin mechanism like in bacteria and eukaryotes. There is also evidence for recombination in some RNA viruses, specifically positive-sense ssRNA viruses like retroviruses, picornaviruses, and coronaviruses. There is controversy over whether homologous recombination occurs in negative-sense ssRNA viruses like influenza.

In RNA viruses, homologous recombination can be either precise or imprecise. In the precise type of RNA-RNA recombination, there is no difference between the two parental RNA sequences and the resulting crossover RNA region. Because of this, it is often difficult to determine the location of crossover events between two recombining RNA sequences. In imprecise RNA homologous recombination, the crossover region has some difference with the parental RNA sequences – caused by either addition, deletion, or other modification of nucleotides. The level of precision in crossover is controlled by the sequence context of the two recombining strands of RNA: sequences rich in adenine and uracil decrease crossover precision.

Homologous recombination is important in facilitating viral evolution. For example, if the genomes of two viruses with different disadvantageous mutations undergo recombination, then they may be able to regenerate a fully functional genome. Alternatively, if two similar viruses have infected the same host cell, homologous recombination can allow those two viruses to swap genes and thereby evolve more potent variations of themselves.

Effects of dysfunction

Without proper homologous recombination, chromosomes often incorrectly align for the first phase of cell division in meiosis. This causes chromosomes to fail to properly segregate in a process called nondisjunction. In turn, nondisjunction can cause sperm and ova to have too few or too many chromosomes. Down's syndrome, which is caused by an extra copy of chromosome 21, is one of many abnormalities that result from such a failure of homologous recombination in meiosis.

Deficiencies in homologous recombination have been strongly linked to cancer formation in humans. For example, each of the cancer-related diseases Bloom's syndrome, Werner's syndrome and Rothmund-Thomson syndrome are caused by malfunctioning copies of RecQ helicase genes involved in the regulation of homologous recombination: BLM, WRN and RECQ4, respectively. In the cells of Bloom's syndrome patients, who lack a working copy of the BLM protein, there is an elevated rate of homologous recombination. Experiments in mice deficient in BLM have suggested that the mutation gives rise to cancer through a loss of heterozygosity caused by increased homologous recombination. A loss in heterozygosity refers to the loss of one of two versions—or alleles—of a gene. If one of the lost alleles helps to suppress tumors, like the gene for the retinoblastoma protein for example, then the loss of heterozygosity can lead to cancer.

Decreased rates of homologous recombination cause inefficient DNA repair, which can also lead to cancer. This is the case with BRCA1 and BRCA2, two similar tumor suppressor genes whose malfunctioning has been linked with considerably increased risk for breast and ovarian cancer. Cells missing BRCA1 and BRCA2 have a decreased rate of homologous recombination and increased sensitivity to ionizing radiation, suggesting that decreased homologous recombination leads to increased susceptibility to cancer. Because the only known function of BRCA2 is to help initiate homologous recombination, researchers have speculated that more detailed knowledge of BRCA2's role in homologous recombination may be the key to understanding the causes of breast and ovarian cancer.

Evolutionary origins

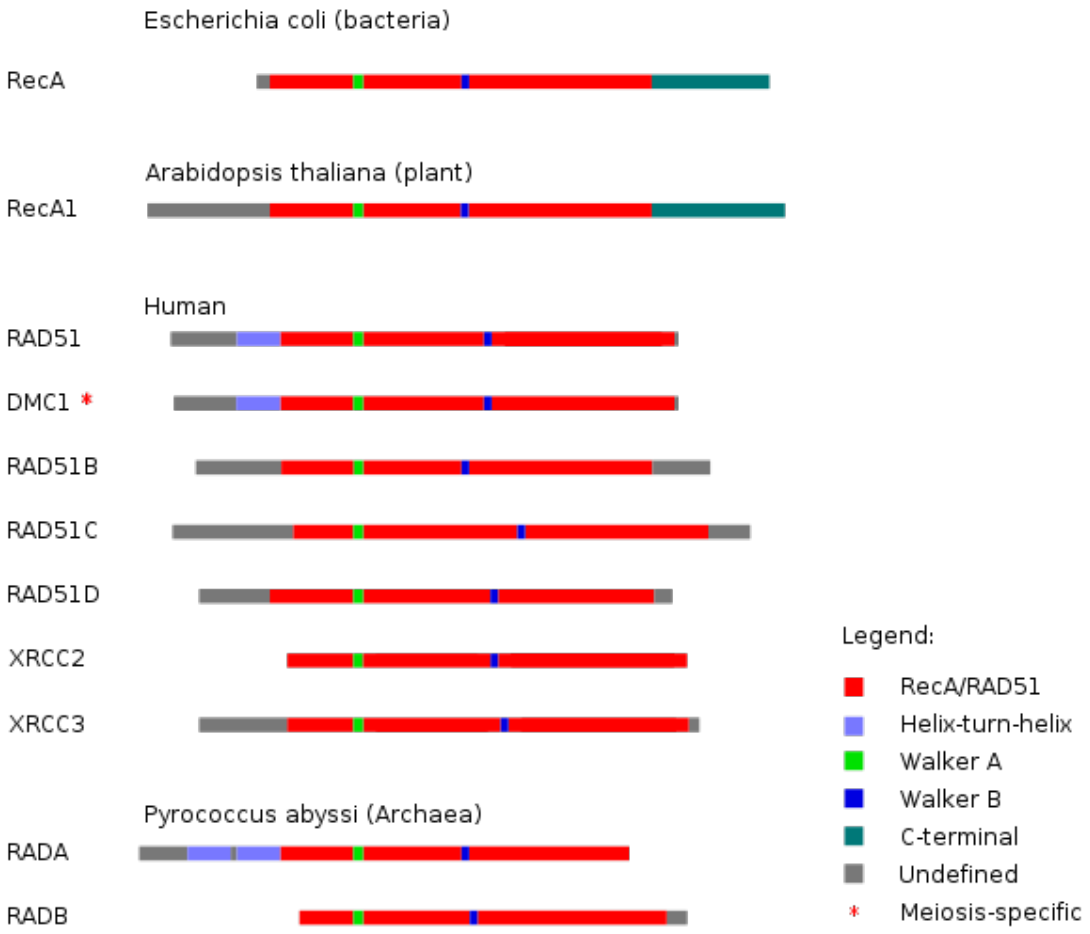


Figure 8. Protein domains in homologous recombination-related proteins are conserved across the three main groups of life: archaea, bacteria and eukaryotes.

Based on the similarity of their amino acid sequences, sets of proteins involved in homologous recombination are thought to share common evolutionary origins. One such set of proteins is the RecA/Rad51 protein family, which includes the RecA protein from bacteria, the Rad51 and Dmc1 proteins from eukaryotes and the RadA and RadB proteins from archaea. These proteins play key roles in the beginning stages of homologous recombination in the organisms that express them. The proteins in the RecA/Rad51 protein family share a long conserved region known as the RecA/Rad51 domain. Within this protein domain are two sequence motifs, Walker A and Walker B. The Walker A and B motifs allow members of the RecA/Rad51 protein family to engage in ATP hydrolysis, which provides energy for the proteins to drive reactions in homologous recombination.

Studies modeling the evolutionary relationships between the Rad51, Dmc1 and RadA proteins indicate that they are monophyletic, or that they share a common molecular ancestor. Within this protein family, Rad51 and Dmc1 are grouped together in a separate clade from RadA. One of the reasons for grouping these three proteins together is that they all possess a modified helix-turn-helix motif, which helps the proteins bind to DNA, toward their N-terminal ends. An ancient gene duplication event of a eukaryotic RecA gene and subsequent mutation has been proposed as a likely origin of the modern RAD51 and DMC1 genes.

The discovery of Dmc1 in several species of *Giardia*, one of the earliest protists to diverge as a eukaryote, suggests that meiotic homologous recombination—and thus meiosis itself—emerged very early in eukaryotic evolution. In addition to research on Dmc1, studies on the Spo11 protein have provided information on the origins of meiotic recombination. Spo11 is a type II topoisomerase that initiates homologous recombination in meiosis by making targeted double-strand breaks in DNA. Phylogenetic trees based on the sequence of genes similar to SPO11 in animals, fungi, plants, protists and archaea have led scientists to believe that the version Spo11 currently in eukaryotes emerged in the last common ancestor of eukaryotes and archaea.

Technological applications

Gene targeting

Figure 9. As a developing embryo, this chimeric mouse had the agouti coat color gene introduced into its DNA via gene targeting. Its offspring are homozygous for the agouti gene.

Many methods for introducing DNA sequences into organisms to create recombinant DNA and genetically modified organisms use the process of homologous recombination. Also called gene targeting, the method is especially common in yeast and mouse genetics. The gene targeting method in knockout mice uses mouse embryonic stem cells to deliver artificial genetic material (mostly of therapeutic interest), which represses the target gene of the mouse by the principle of homologous recombination. The mouse thereby acts as a working model to understand the effects of a specific mammalian gene. In recognition of their discovery of how homologous recombination can be used to

introduce genetic modifications in mice through embryonic stem cells, Mario Capecchi, Martin Evans and Oliver Smithies were awarded the 2007 Nobel Prize for Physiology or Medicine.

Advances in gene targeting technologies which hijack the homologous recombination mechanics of cells are now leading to the development of a new wave of more accurate, isogenic human disease models. These engineered human cell models are thought to more accurately reflect the genetics of human diseases than their mouse model predecessors. This is largely because mutations of interest are introduced into endogenous genes, just as they occur in the real patients, and because they are based on human genomes rather than rat genomes. Furthermore, certain technologies enable the knock-in of a particular mutation rather than just knock-outs associated with older gene targeting technologies.

Protein engineering

Protein engineering with homologous recombination develops chimeric proteins by swapping fragments between two parental proteins. These techniques exploit the fact that recombination can introduce a high degree of sequence diversity while preserving a protein's ability to fold into its tertiary structure, or three-dimensional shape. This stands in contrast to other protein engineering techniques, like random point mutagenesis, in which the probability of maintaining protein function declines exponentially with increasing amino acid substitutions. The chimeras produced by recombination techniques are able to maintain their ability to fold because their swapped parental fragments are structurally and evolutionarily conserved. These recombinable "building blocks" preserve structurally important interactions like points of physical contact between different amino acids in the protein's structure. Computational methods like SCHEMA and statistical coupling analysis can be used to identify structural subunits suitable for recombination.

Techniques that rely on homologous recombination have been used to engineer new proteins. In a study published in 2007, researchers were able to create chimeras of two enzymes involved in the biosynthesis of isoprenoids, a diverse class of compounds including hormones, visual pigments and certain pheromones. The chimeric proteins acquired an ability to catalyze an essential reaction in isoprenoid biosynthesis—one of the most diverse pathways of biosynthesis found in nature—that was absent in the parent proteins. Protein engineering through recombination has also produced chimeric enzymes with new function in members of a group of proteins known as the cytochrome P450 family, which in humans is involved in detoxifying foreign compounds like drugs, food additives and preservatives.

Cancer therapy

In 2009, researchers reported trial results of a cancer therapy that exploits deficiencies in homologous recombination in certain types of cancer. The drug, named olaparib, a PARP1 inhibitor, was shown to shrink or stop the growth of tumors from breast, ovarian and prostate cancers caused by mutations in the BRCA1 or BRCA2 genes. BRCA1 and

BRCA2 are necessary for DNA repair by homologous recombination. When BRCA1 or BRCA2 are absent, another type of DNA repair mechanism called base-excision repair usually compensates for the lack of DNA repair by homologous recombination. However, when the PARP1 protein – which is necessary for base-excision repair – is inhibited, DNA repair is drastically reduced, and the cell dies.

By stopping DNA repair in such a fashion, olaparib applies the concept of synthetic lethality to specifically target cancer cells. An article in the *New England Journal of Medicine* noted that exploiting synthetic lethality was a new direction in anti-cancer drug development. While PARP1 inhibitors represent a novel approach to cancer therapy, researchers have cautioned that they may prove insufficient for treating late-stage, metastatic cancers. Cancer cells can become resistant to a PARP1 inhibitor if they experience deletions of mutations in BRCA2. This undermines the drug's synthetic lethality by restoring cancer cells' ability to repair DNA by homologous recombination.

Chapter- 9

Promoter (Biology)

In genetics, a **promoter** is a region of DNA that facilitates the transcription of a particular gene. Promoters are typically located near the genes they regulate, on the same strand and upstream (towards the 5' region of the sense strand).

Overview

In order for the transcription to take place, the enzyme that synthesizes RNA, known as RNA polymerase, must attach to the DNA near a gene. Promoters contain specific DNA sequences and response elements which provide a secure initial binding site for RNA polymerase and for proteins called transcription factors that recruit RNA polymerase. These transcription factors have specific activator or repressor sequences of corresponding nucleotides that attach to specific promoters and regulate gene expressions.

In bacteria
: the promoter is recognized by RNA polymerase and an associated sigma factor, which in turn are often brought to the promoter DNA by an activator protein binding to its own DNA binding site nearby.

In eukaryotes
: the process is more complicated, and at least seven different factors are necessary for the binding of an RNA polymerase II to the promoter.

Promoters represent critical elements that can work in concert with other regulatory regions (enhancers, silencers, boundary elements/insulators) to direct the level of transcription of a given gene.

Identification of relative location

As promoters are typically immediately adjacent to the gene in question, positions in the promoter are designated relative to the transcriptional start site, where transcription of RNA begins for a particular gene (i.e., positions upstream are negative numbers counting back from -1, for example -100 is a position 100 base pairs upstream).

Promoter elements

- Core promoter - the minimal portion of the promoter required to properly initiate transcription
    - Transcription Start Site (TSS)
    - Approximately -34 bp upstream of the start site
    - A binding site for RNA polymerase
        - RNA polymerase I: transcribes genes encoding ribosomal RNA
        - RNA polymerase II: transcribes genes encoding messenger RNA and certain small nuclear RNAs
        - RNA polymerase III: transcribes genes encoding tRNAs and other small RNAs
    - General transcription factor binding sites
- Proximal promoter - the proximal sequence upstream of the gene that tends to contain primary regulatory elements
    - Approximately -250 bp upstream of the start site
    - Specific transcription factor binding sites
- Distal promoter - the distal sequence upstream of the gene that may contain additional regulatory elements, often with a weaker influence than the proximal promoter
    - Anything further upstream (but not an enhancer or other regulatory region whose influence is positional/orientation independent)
    - Specific transcription factor binding sites

Prokaryotic promoters

In prokaryotes, the promoter consists of two short sequences at -10 and -35 positions *upstream* from the transcription start site. Sigma factors not only help in enhancing RNAP binding to the promoter but also help RNAP target specific genes to transcribe.

- The sequence at -10 is called the Pribnow box, or the -10 element, and usually consists of the six nucleotides TATAAT. The Pribnow box is absolutely essential to start transcription in prokaryotes.
- The other sequence at -35 (the -35 element) usually consists of the seven nucleotides TTGACAT. Its presence allows a very high transcription rate.
- Both of the above consensus sequences, while conserved on average, are not found intact in most promoters. On average only 3 of the 6 base pairs in each consensus sequence is found in any given promoter. No promoter has been identified to date that has intact consensus sequences at both the -10 and -35; artificial promoters with complete conservation of the -10/-35 hexamers has been found to promote RNA chain initiation at very high efficiencies.
- Some promoters contain a UP element (consensus sequence `5'-AAAWWTWTTTTNNNAAANNN-3'`; W = A or T; N = any base) centered at -50; the presence of the -35 element appears to be unimportant for transcription from the UP element-containing promoters.

It should be noted that the above promoter sequences are only recognized by the sigma-70 protein that interacts with the prokaryotic RNA polymerase. Complexes of prokaryotic RNA polymerase with other sigma factors recognize totally different core promoter sequences.

```
   <-- upstream
downstream -->
5'-
XXXXXXXPPPPPXXXXXXPPPPPPXXXXGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGXXXX-3'
           -35       -10       Gene to be transcribed

(note that the optimal spacing between the -35 and -10 sequences is 17
bp)
```

Probability of occurrence of each nucleotide

```
 for -10 sequence
 T    A    T    A    A    T
77%  76%  60%  61%  56%  82%
 for -35 sequence
 T    T    G    A    C    A
69%  79%  61%  56%  54%  54%
```

Eukaryotic promoters

Eukaryotic promoters are extremely diverse and are difficult to characterize. They typically lie upstream of the gene and can have regulatory elements several kilobases away from the transcriptional start site(enhancers). In eukaryotes, the transcriptional complex can cause the DNA to bend back on itself, which allows for placement of regulatory sequences far from the actual site of transcription. Many eukaryotic promoters, between 10 and 20% of all genes, contain a TATA box (sequence TATAAA), which in turn binds a TATA binding protein which assists in the formation of the RNA polymerase transcriptional complex. The TATA box typically lies very close to the transcriptional start site (often within 50 bases).

Eukaryotic promoter regulatory sequences typically bind proteins called transcription factors which are involved in the formation of the transcriptional complex. An example is the E-box (sequence CACGTG), which binds transcription factors in the basic-helix-loop-helix (bHLH) family (e.g. BMAL1-Clock, cMyc).

Detection of promoters

A wide variety of algorithms have been developed to facilitate detection of promoters in genomic sequence, and promoter prediction is a common element of many gene prediction methods. A promoter region is located before the -35 and -10 Consensus sequences. The closer the promoter region is to the consensus sequences the more often transcription of that gene will take place. There is not a set pattern for promoter regions as there are for consensus sequences.

Evolutionary change

A major question in evolutionary biology is how important tinkering with promoter sequences is to evolutionary change, for example, the changes that have occurred in the human lineage after separating from chimps.

Some evolutionary biologists, for example Allan Wilson, have proposed that evolution in promoter or regulatory regions may be more important than changes in coding sequences over such time frames.

A key reason for the importance of promoters is the potential to incorporate endocrine and environmental signals into changes in gene expression: A great variety of changes in the extracellular or intracellular environment may have impact on gene expression, depending on the exact configuration of a given promoter: the combination and arrangement of specific DNA sequences that constitute the promoter defines the exact groups of proteins that can be bound to the promoter, at a given timepoint. Once the cell receives a physiological, pathological, or pharmacological stimulus, a number of cellular proteins are modified biochemically by signal cascades. By changes in structure, specific proteins acquire the capability to enter the nucleus of the cell and bind to promoter DNA, or to other proteins that themselves are already bound to a given promoter. The multi-protein complexes that are formed have the potential to change levels of gene expression. As a result the gene product may increase or decrease inside the cell.

Binding

The binding of a promoter sequence (P) to a sigma factor-RNAP complex (R) is a two-step process:

1. R+P ↔ RP(closed). $K = 10^7$
2. RP(closed) → RP(open). $K = 10^{-2}$

Diseases associated with aberrant promoter function

Though OMIM is a major resource for gathering information on the relationship between mutations and natural variation in gene sequence and susceptibility to hundreds of diseases, it requires a sophisticated search strategy to extract those diseases that are associated with defects in transcriptional control where the promoter is believed to have direct involvement.

This is a list of diseases that evidence suggests have some involvement of promoter malfunction, either through direct mutation of a promoter sequence or mutation in a transcription factor or transcriptional co-activator.

Keep in mind that most diseases are heterogeneous in etiology, meaning that one "disease" is often many different diseases at the molecular level, though the symptoms

exhibited and the response to treatment might be identical. How diseases respond differently to treatment as a result of differences in the underlying molecular origins is partially addressed by the discipline of pharmacogenomics.

Not listed here are the many kinds of cancers that involve aberrant changes in transcriptional regulation owing to the creation of chimeric genes through pathological chromosomal translocation. Importantly, intervention on the number or the structure of promoter-bound proteins is a key to treat a disease without to cause a number of changes in the expression of unrelated genes that share particular elements with the specific gene that is the target of therapy. Such genes, whose change is not desirable, are capable to influence the potential of a cell to become cancerous, and form a tumor.

Canonical sequences and wild-type

The usage of canonical sequence for a promoter is often problematic, and can lead to misunderstandings about promoter sequences. Canonical implies perfect, in some sense.

In the case of a transcription factor binding site, then there may be a single sequence which binds the protein most strongly under specified cellular conditions. This might be called canonical.

However, natural selection may favor less energetic binding as a way of regulating transcriptional output. In this case, we may call the most common sequence in a population, the wild-type sequence. It may not even be the most advantageous sequence to have under prevailing conditions.

Recent evidence also indicates that several genes (including the proto-oncogene c-myc) have G-quadruplex motifs as potential regulatory signals.

Diseases associated with promoter elements

- Asthma
- Beta thalassemia
- Rubinstein-Taybi syndrome

Chapter- 10

Noncoding DNA

In genetics, **noncoding DNA** describes components of an organism's DNA sequences that do not encode for protein sequences. In many eukaryotes, a large percentage of an organism's total genome size is noncoding DNA, although the amount of noncoding DNA, and the proportion of coding versus noncoding DNA varies greatly between species.

Much of this DNA has no known biological function and at one time was sometimes referred to as "**Junk DNA**". However, many types of noncoding DNA sequences do have known biological functions, including the transcriptional and translational regulation of protein-coding sequences. Other noncoding sequences have likely but as-yet undetermined function, an inference from high levels of homology and conservation seen in sequences that do not encode proteins but appear to be under heavy selective pressure.

Fraction of noncoding genomic DNA

The amount of total genomic DNA varies widely between organisms, and the proportion of coding and noncoding DNA within these genomes varies greatly as well. More than 98% of the human genome does not encode protein sequences, including most sequences within introns and most intergenic DNA.

While overall genome size, and by extension the amount of noncoding DNA, are correlated to organism complexity, there are many exceptions. For example, the genome of the unicellular *Polychaos dubium* (formerly known as *Amoeba dubia*) has been reported to contain more than 200 times the amount of DNA in humans. The pufferfish *Takifugu rubripes* genome is only about one eighth the size of the human genome, yet seems to have a comparable number of genes; approximately 90% of the *Takifugu* genome is noncoding DNA and most of the genome size difference appears to lie in the noncoding DNA. The extensive variation in nuclear genome size among eukaryotic species is known as the C-value enigma or C-value paradox.

About 80% of the nucleotide bases in the human genome may be transcribed, but transcription does not necessarily imply function.

Types of noncoding DNA sequences

Noncoding functional RNA

Noncoding RNAs are functional RNA molecules that are not translated into protein. Examples of noncoding RNA include ribosomal RNA, transfer RNA, Piwi-interacting RNA and microRNA.

MicroRNAs are predicted to control the translational activity of approximately 30% of all protein-coding genes in mammals and may be vital components in the progression or treatment of various diseases including cancer, cardiovascular disease, and the immune system response to infection.

Cis-regulatory elements

Cis-regulatory elements are sequences that control the transcription of a gene. Cis-elements may be located in 5' or 3' untranslated regions or within introns. Promoters facilitate the transcription of a particular gene and are typically upstream of the coding region.

Enhancer sequences may exert very distant effects on the transcription levels of genes.

Introns

Introns are non-coding sections of a gene, transcribed into the precursor mRNA sequence, but ultimately removed by RNA splicing during the processing to mature messenger RNA. Many introns appear to be mobile genetic elements.

Studies of group I introns from *Tetrahymena* indicate that some introns appear to be selfish genetic elements, neutral to the host because they remove themselves from flanking exons during RNA processing and do not produce an expression bias between alleles with and without the intron. Some introns do appear to have significant biological function, possibly through ribozyme functionality that may regulate tRNA and rRNA activity as well as protein-coding gene expression, evident in hosts that have become dependent on such introns over long periods of time; for example, the *trnL-intron* is found in all green plants and appears to have been vertically inherited for several billions of years, including more than a billion years within chloroplasts and an additional 2–3 billion years prior in the cyanobacterial ancestors of chloroplasts.

Pseudogenes

Pseudogenes are DNA sequences, related to known genes, that have lost their protein-coding ability or are otherwise no longer expressed in the cell. Pseudogenes arise from retrotransposition or genomic duplication of functional genes, and become "genomic fossils" that are nonfuctional due to mutations that prevent the transcription of the gene,

such as within the gene promoter region, or fatally alter the translation of the gene, such as premature stop codons or frameshifts. Pseudogenes resulting from the retrotransposition of an RNA intermediate are known as processed pseudogenes; pseudogenes that arise from the genomic remains of duplicated genes or residues of inactivated are nonprocessed pseudogenes.

While Dollo's Law suggests that the loss of function in pseudogenes is likely permanent, silenced genes may actually retain function for several million years and can be "reactivated" into protein-coding sequences and a substantial number of pseudogenes are actively transcribed. Because pseudogenes are presumed to evolve without evolutionary constraint, they can serve as a useful model of the type and frequencies various spontaneous genetic mutations.

Repeat sequences, transposons and viral elements

Transposons and retrotransposons are mobile genetic elements. Retrotransposon repeated sequences, which include long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs), account for a large proportion of the genomic sequences in many species. Alu sequences, classified as a short interspersed nuclear element, are the most abundant mobile elements in the human genome. Some examples have been found of SINEs exerting transcriptional control of some protein-encoding genes.

Endogenous retrovirus sequences are the product of reverse transcription of retrovirus genomes into the genomes of germ cells. Mutation within these retro-transcribed sequences can inactivate the viral genome.

Over 8% of the human genome is made up of (mostly decayed) endogenous retrovirus sequences, as part of the over 42% fraction that is recognizably derived of retrotransposons, while another 3% can be identified to be the remains of DNA transposons. Much of the remaining half of the genome that is currently without an explained origin is expected to have found its origin in transposable elements that were active so long ago (> 200 million years) that random mutations have rendered them unrecognizable. Genome size variation in at least two kinds of plants is mostly the result of retrotransposon sequences.

Telomeres

Telomeres are regions of repetitive DNA at the end of a chromosome, which provide protection from chromosomal deterioration during DNA replication.

Functions of noncoding DNA

Many noncoding DNA sequences have important biological functions as indicated by Comparative genomics studies that report some regions of noncoding DNA that are highly conserved, sometimes on time-scales representing hundreds of millions of years,

implying that these noncoding regions are under strong evolutionary pressure and positive selection. For example, in the genomes of humans and mice, which diverged from a common ancestor 65–75 million years ago, protein-coding DNA sequences account for only about 20% of conserved DNA, with the remaining majority of conserved DNA represented in noncoding regions. Linkage mapping often identifies chromosomal regions associated with a disease with no evidence of functional coding variants of genes within the region, suggesting that disease-causing genetic variants lie in the noncoding DNA.

Some noncoding DNA sequences are genetic "switches" that do not encode proteins, but do regulate when and where genes are expressed. According to a comparative study of over 300 prokaryotic and over 30 eukaryotic genomes, eukaryotes appear to require a minimum amount of non-coding DNA. This minimum amount can be predicted using a growth model for regulatory genetic networks, implying that it is required for regulatory purposes. In humans the predicted minimum is about 5% of the total genome.

Some noncoding DNA sequences determine the expression levels of various genes. Other sequences of noncoding DNA determine where transcription factors attach.

Some specific sequences of noncoding DNA may be features essential to chromosome structure, centromere function and homolog recognition in meiosis.

Noncoding DNA and evolution

Shared sequences of apparently non-functional DNA are a major line of evidence for common descent.

Pseudogene sequences appear to accumulate mutations more rapidly than coding sequences due to a loss of selective pressure. This allows for the creation of mutant alleles that incorporate new functions that may be favored by natural selection; thus, pseudogenes can serve as raw material for evolution and can be considered "protogenes".

Junk DNA

Junk DNA, a term that was introduced in 1972 by Susumu Ohno, was a provisional label for the portions of a genome sequence for which no discernible function had been identified. According to a 1980 review in *Nature* by Leslie Orgel and Francis Crick, junk DNA has "little specificity and conveys little or no selective advantage to the organism". The term is currently, however, an outdated concept, being used mainly in popular science and in a colloquial way in scientific publications, and may have slowed research into the biological functions of noncoding DNA. Several lines of evidence indicate that many "junk DNA" sequences are likely to have unidentified functional activity, and other sequences may have had functions in the past.

Still, a significant amount of the sequence of the genomes of eukaryotic organisms currently appears to fall under no existing classification other than "junk". For example, one experiment removed 0.1% of the mouse genome with no detectable effect on the phenotype. This result suggests that the removed DNA was largely nonfunctional. In addition, these sequences are enriched for the heterochromatic histone modification H3K9me3.