Branches and Key Components of Genetics

Marie Linares
Lorena Shuman

First Edition, 2012

ISBN 978-81-323-0685-6

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

Table of Contents

Chapter- 1

Behavioural Genetics, Classical Genetics and Ecological Genetics

Behavioural genetics

Behavioural genetics is the field of study that examines the role of genetics in animal (including human) behaviour. Often associated with the "nature versus nurture" debate, behavioural genetics is highly interdisciplinary, involving contributions from biology, genetics, ethology, psychology, and statistics. Behavioural geneticists study the inheritance of behavioural traits. In humans this often use the twin study or adoption study. In animal studies, breeding, transgenesis, and gene knockout techniques are common; psychiatric genetics is a closely related field.

History

Sir Francis Galton, a nineteenth-century intellectual, is recognized as one of the first behavioural geneticists. Galton, a cousin of Charles Darwin, studied the heritability of human ability, focusing on mental characteristics as well as eminence among close relatives in the English upper-class. In 1869, Galton published his results in *Hereditary Genius*. In his work, Galton "introduced multivariate analysis and paved the way towards modern Bayesian statistics" that are used throughout the sciences—launching what has been dubbed the "Statistical Enlightenment".

Behaviour genetics, *per-se*, gained recognition as a research discipline with the publication in 1960 of the textbook *Behavior Genetics* by J.L. Fuller and W.R. Thompson.

Underscoring the role of evolution in behavioural genetics, Theodosius Dobzhansky was elected the first president of the Behavior Genetics Association in 1972; the BGA bestows the Dobzhansky Award on researchers for their outstanding contributions to the field. In the early 1970s, Lee Ehrman, a doctoral student of Dobzhansky, wrote seminal papers describing the relationship between genotype frequency and mating success in

Drosophila, lending impetus to the pursuit of genetic studies of behaviour in other animals.

Notable behavioural geneticists

Notable behavioural geneticists include Dorret Boomsma, John DeFries, Lindon Eaves, David Fulker, John Hewitt, Kenneth Kendler, John Loehlin, Nick Martin, Gerald McClearn, Robert Plomin, Theodore Reich, who was a pioneer in psychiatric genetics, Hans van Abeelen, Avshalom Caspi, and Steven G. Vandenberg, the founding editor of the journal *Behavior Genetics*.

Journals

Behavioural geneticists are active in a variety of scientific disciplines including biology, medicine, pharmacology, psychiatry, and psychology; thus, behavioural-genetic research is published in a variety of scientific journals, including *Nature* and *Science*. Journals that specifically publish research in behavioural genetics include *Behavior Genetics*, *Molecular Psychiatry*, *Psychiatric Genetics*, *Twin Research and Human Genetics*, and *Genes, Brain and Behavior*.

Classical genetics

Classical genetics consists of the technique and methodologies of genetics that predate the advent of molecular biology. A key discovery of classical genetics in eukaryotes was genetic linkage. The observation that some genes do not segregate independently at meiosis broke the laws of Mendelian inheritance, and provided science with a way to map characteristics to a location on the chromosomes. Linkage maps are still used today, especially in breeding for plant improvement.

After the discovery of the genetic code and such tools of cloning as restriction enzymes, the avenues of investigation open to geneticists were greatly broadened. Some classical genetic ideas have been supplanted with the mechanistic understanding brought by molecular discoveries, but many remain intact and in use. Classical genetics is often contrasted with reverse genetics, and aspects of molecular biology are sometimes referred to as molecular genetics.

Ecological genetics

Ecological genetics is the study of genetics in the context of the interactions among organisms and between the organisms and their environment. While molecular genetics studies the structure and function of genes at a molecular level, ecological genetics (and the related field of population genetics) studies phenotypic evolution in natural populations of organisms. Research in this field is of traits of ecological significance — that is, traits related to fitness, which affect an organism's survival and reproduction (e.g., flowering time, drought tolerance, sex ratio).

Studies are often done on insects and other organisms that have short generation times, and thus evolve at high rates.

History

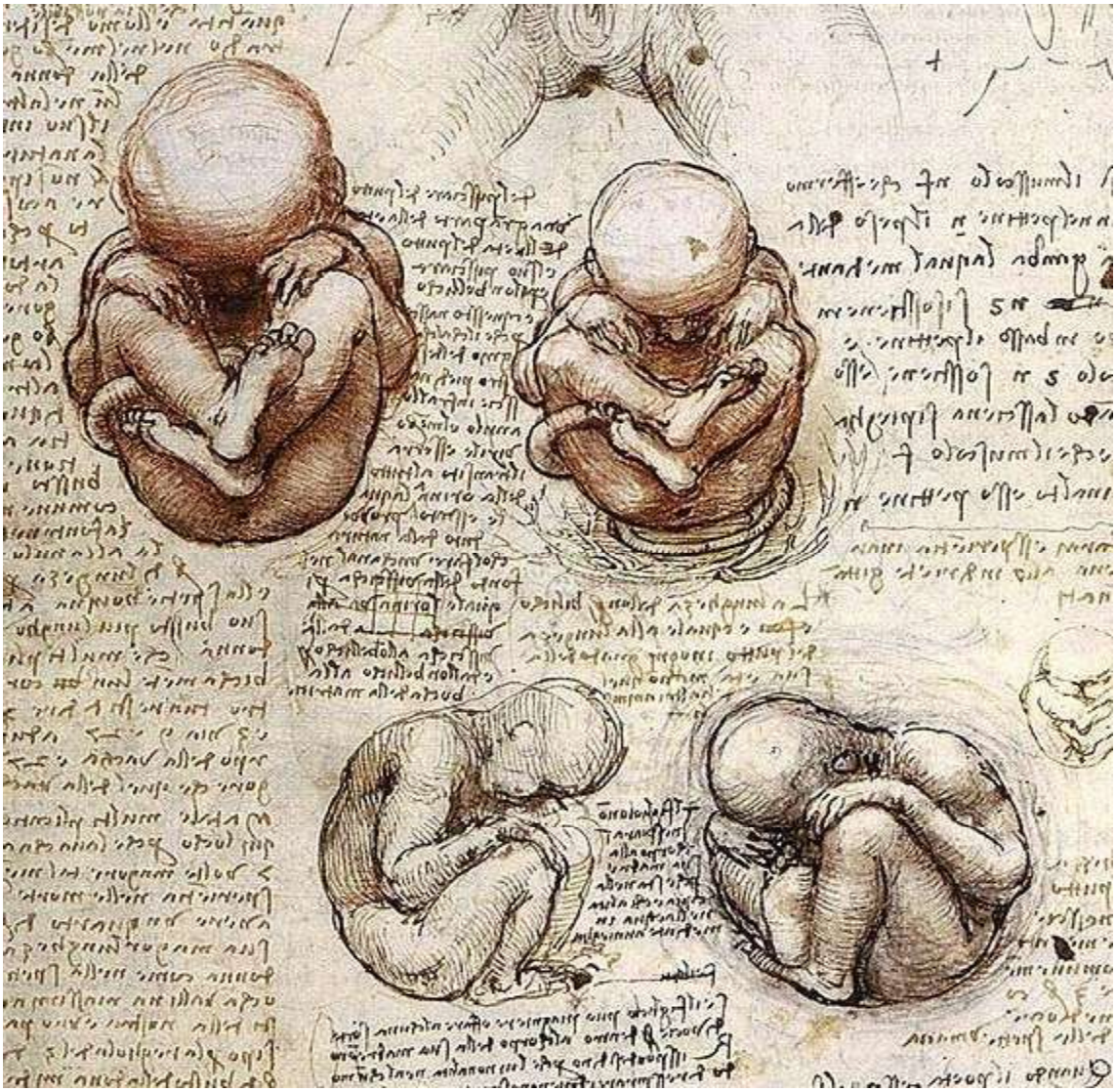
Although work on natural populations had been done previously, it is acknowledged that the field was founded by the English biologist E.B. Ford (1901-1988) in the early 20th century. Ford was taught genetics at Oxford University by Julian Huxley, and started research on the genetics of natural populations in 1924. *Ecological Genetics* is the title of his 1964 'magnum opus' on the subject (4th ed 1975). Other notable ecological geneticists would include Theodosius Dobzhansky who worked on chromosome polymorphism in fruit flies. As a young researcher in Russia, Dobzhansky had been influenced by Sergei Chetverikov, who also deserves to be remembered as a founder of genetics in the field, though his significance was not appreciated until much later.

Philip Sheppard, Cyril Clarke, Bernard Kettlewell and A.J. Cain were all strongly influenced by Ford; their careers date from the post WWII era. Collectively, their work on lepidopterans, and on human blood groups, established the field, and threw light on selection in natural populations where its role had been once doubted.

Work of this kind needs long-term funding, as well as grounding in both ecology and genetics. These are both difficult requirements. Research projects can last longer than a researcher's career; for instance, research into mimicry started 150 years ago, and is still going strongly. Funding of this type of research is still rather erratic, but at least the value of working with natural populations in the field cannot now be doubted.

Chapter- 2

Developmental Biology



"Views of a Fetus in the Womb", Leonardo da Vinci, ca. 1510-1512. The subject of prenatal development is a major subset of developmental biology.

Developmental biology is the study of the process by which organisms grow and develop. Modern developmental biology studies the genetic control of cell growth, differentiation and "morphogenesis", which is the process that gives rise to tissues, organs and anatomy.

Related fields of study

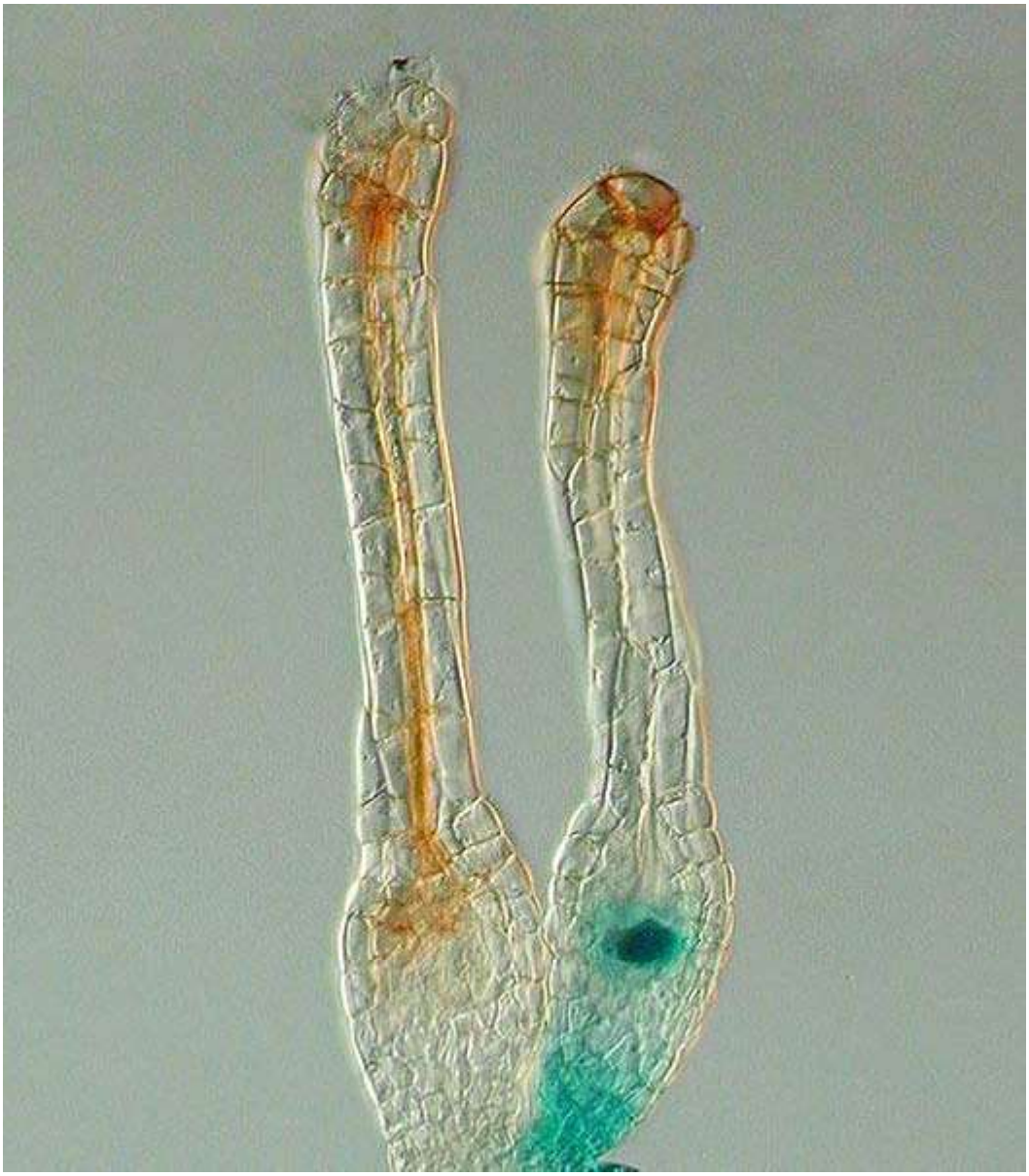
Embryology is a subfield, the study of organisms between the one-cell stage (generally, the zygote) and the end of the embryonic stage. Embryology was originally a more descriptive science until the 20th century. Embryology and developmental biology today deal with the various steps necessary for the correct and complete formation of the body of a living organism.

The related field of evolutionary developmental biology was formed largely in the 1990s and is a synthesis of findings from molecular developmental biology and evolutionary biology which considers the diversity of organismal form in an evolutionary context.

Perspectives

The development of a new life is a spectacular process and represents a masterpiece of temporal and spatial control of gene expression. Developmental genetics studies the effect that genes have in a phenotype, given normal or abnormal epigenetic parameters. The findings of developmental biology can help to understand developmental abnormalities such as chromosomal aberrations that cause Down syndrome. An understanding of the specialization of cells during embryogenesis has provided information on how stem cells specialize into specific tissues and organs. This information has led, for example, to the cloning of specific organs for medical purposes. Another biologically important process that occurs during development is apoptosis—programmed cell death or "suicide." Many developmental models are used to elucidate the physiology and molecular basis of this cellular process. Similarly, a deeper understanding of developmental biology can foster greater progress in the treatment of congenital disorders and diseases, e.g. studying human sex determination can lead to treatment for disorders such as congenital adrenal hyperplasia.

Developmental model organisms



Gene expression pattern determined by histochemical GUS assays in *Physcomitrella patens*. The Polycomb gene FIE is expressed (blue) in unfertilised egg cells of the moss Physcomitrella patens (right) and expression ceases after fertilisation in the developing diploid sporophyte (left). In situ GUS staining of two female sex organs (archegonia) of a transgenic plant expressing a translational fusion of FIE-uidA under control of the native FIE promoter

Often used model organisms in developmental biology include the following:

- Vertebrates

    - Zebrafish *Danio rerio*
    - Medakafish *Oryzias latipes*
    - Fugu (pufferfish) *Takifugu rubripes*
    - Frog *Xenopus laevis*, *Xenopus tropicalis*
    - Chicken *Gallus gallus*
    - Mouse *Mus musculus* (Mammalian embryogenesis)

- Invertebrates
    - Lancelet *Branchiostoma lanceolatum*
    - Ascidian *Ciona intestinalis*
    - Sea urchin *Strongylocentrotus purpuratus*
    - Roundworm *Caenorhabditis elegans*
    - Fruit fly *Drosophila melanogaster* (Drosophila embryogenesis)

- Plants (Plant embryogenesis)
    - *Physcomitrella patens*
    - *Arabidopsis thaliana*
    - Maize
    - Snapdragon *Antirrhinum majus*

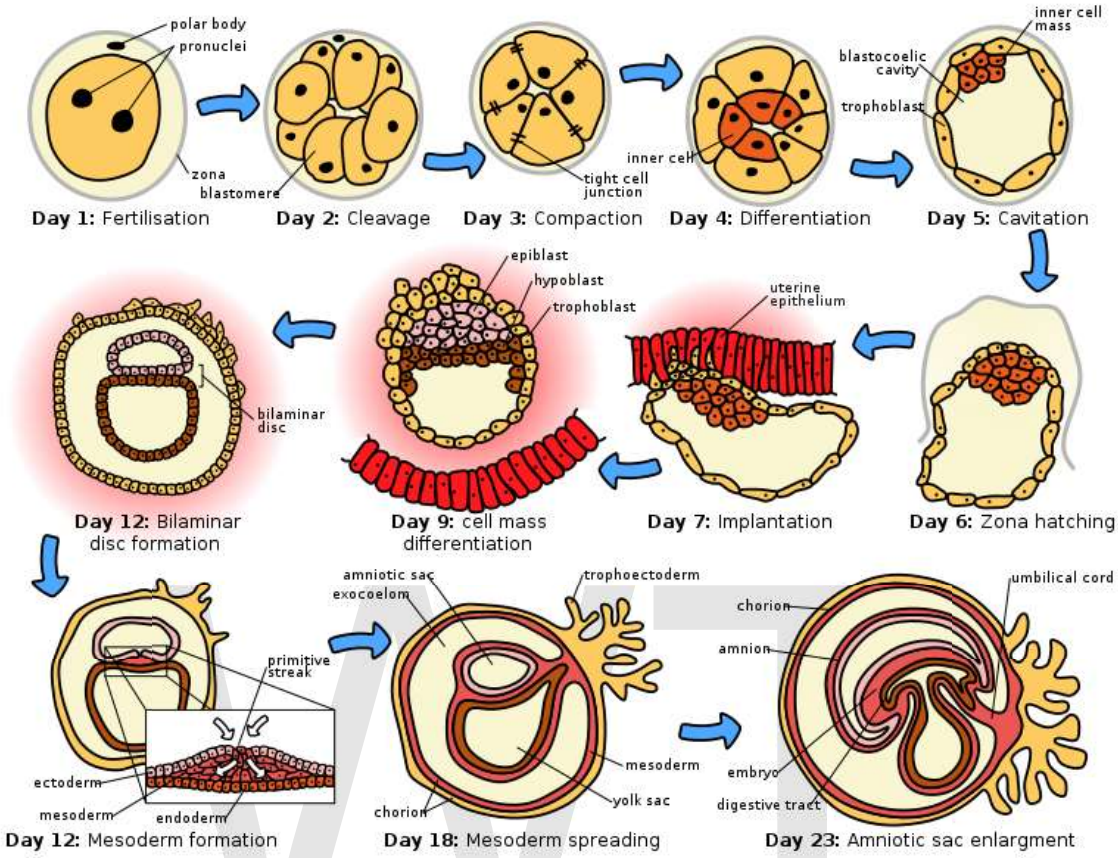
- Other
    - Slime mold *Dictyostelium discoideum*

Studied phenomena

Cell differentiation

Differentiation is the formation of cell types, from what is originally one cell – the zygote or spore. The formation of cell types like nerve cells occurs with a number of intermediary, less differentiated cell types. A cell stays a certain cell type by maintaining a particular pattern of gene expression. This depends on regulatory genes, e.g. for [[transcription factor] name=WolpertDifferentiationmodels>Wolpert L, Beddington R, Jessell T, Lawrence P, Meyerowitz E, Smith J (2002).

Embryonal development



The initial stages of human embryogenesis

Embryogenesis is the step in the life cycle after fertilisation – the development of the embryo, starting from the zygote (fertilised egg). Organisms can differ drastically in the how embryo develops, especially when they belong to different phyla. For example, embryonal development in placental mammals starts with cleavage of the zygote into eight uncommited cells, which then form a ball (morula). The outer cells become the trophectoderm or trophoblast, which will form in combination with maternal uterine endometrial tissue the placenta, needed for fetal nurturing via maternal blood, while inner cells become the inner cell mass that will form all fetal organs (the bridge between these two parts eventually forms the umbilical cord). In contrast, the fruit fly zygote first forms a sausage-shaped syncytium, which is still one cell but with many cell nuclei.

Patterning is important for determining which cells develop into which organs. This is mediated by signaling between adjacent cells by proteins on their surfaces, and by gradients of signaling secreted molecules. An example is retinoic acid, which forms a gradient in the head to tail direction in animals. Retinoic acid enters cells and activates Hox genes in a concentration-dependent manner – Hox genes differ in how much retinoic acid they require for activation and will thus show differential rostral expression boundaries, in a colinear fashion with their genomic order. As Hox genes code for

transcription factors, this causes different activated combinations of both Hox and other genes in discrete anteroposterior transverse segments of the neural tube (neuromeres) and related patterns in surrounding tissues, such as branchial arches, lateral mesoderm, neural crest, skin and endoderm, in the head to tail direction. This is important for e.g. the segmentation of the spine in vertebrates.

Embryonal development does not always proceed correctly, and errors can result in birth defects or miscarriage. Often the reason is genetic (mutation or chromosome abnormality), but there can be environmental influence (like teratogens) or stochastic events. Abnormal development caused by mutation is also of evolutionary interest as it provides a mechanism for changes in body plan.

Growth

Growth is the enlargement of a tissue or organism. Growth continues after the embryonal stage, and occurs through cell proliferation, enlargement of cells or accumulation of extracellular material. In plants, growth results in an adult organism that is strikingly different from the embryo. The proliferating cells tend to be distinct from differentiated cells. In some tissues proliferating cells are restricted to specialised areas, such as the growth plates of bones. But some stem cells migrate to where they are needed, such as mesenchymal stem cells which can migrate from the bone marrow to form e.g. muscle, bone or adipose tissue. The size of an organ frequently determines its growth, as in the case of the liver which grows back to its previous size if a part is removed. Growth factors, such as fibroblast growth factors in the animal embryo and growth hormone in juvenile mammals, also control the extent of growth.

Metamorphosis

Most animals have a larval stage, with a body plan different from that of the adult organism. The larva abrubtly develops into an adult in a process called metamorphosis. For example, caterpillars (butterfly larvae) are specialized for feeding whereas adult butterflies (imagos) are specialised for flight and reproduction. When the caterpillar has grown enough, it turns into an immobile pupa. Here, the imago develops from imaginal discs found inside the larva.

Regeneration

Regeneration is the reactivation of development so that a missing body part grows back. This phenomenon has been studied particularly in salamanders, where the adults can reconstruct a whole limb after it has been amputated. Researchers hope to one day be able to induce regeneration in humans. There is little spontaneous regeneration in adult humans, although the liver is a notable exception. Like for salamanders, the regeneration of the liver involves dedifferentiation of some cells to a more embryonal state.

Developmental systems biology

Computer simulation of multicellular development is a research methodology to understand the function of the very complex processes involved in the development of organisms. This includes simulation of cell signaling, multicell interactions and regulatory genomic networks in development of multicellular structures and processes. *Minimal genomes* for minimal multicellular organisms may pave the way to understand such complex processes *in vivo*.

Chapter- 3

Conservation Genetics

Conservation genetics is an interdisciplinary science that aims to apply genetic methods to the conservation and restoration of biodiversity. Researchers involved in conservation genetics come from a variety of fields including population genetics, molecular ecology, biology, evolutionary biology, and systematics. Genetic diversity is one of the three fundamental levels of biodiversity, so it is directly important in conservation of biodiversity, though genetic factors are also important in the conservation of species and ecosystem diversity. Conservation of genetic variability is important to the overall health of populations because decreased genetic variability leads to increased levels of inbreeding, and reduced fitness

Genetic Diversity

Genetic diversity is the variability of a genes in a species. It can be estimated by the mean levels of heterozygosity in a population, the mean number of alleles per locus, or the percentage of polymorphic loci.

The importance of genetic diversity

If genetic diversity becomes low at many genes of a species, that species becomes increasingly at risk. It has only one possible choice of information at all or nearly all of its genes—in other words, all the individuals are nearly identical. If new pressures (such as environmental disasters) occur, a population with high genetic diversity has a greater chance of having at least some individuals with a genetic makeup that allows them to survive. If genetic diversity is very low, none of the individuals in a population may have the characteristics needed to cope with the new environmental conditions. Such a population could be suddenly wiped out.

The genetic diversity of a species is always open to change. No matter how many variants of a gene are present in a population today, only the variants that survive in the next generation can contribute to species diversity in the future. Once gene variants are lost, they cannot be recovered.

Contributors to extinction

1. Inbreeding and inbreeding elevation which reduces the fitness of populations.
2. The accumulation of deleterious mutations
3. A decrease in frequency of heterozygotes in a population, or heterozygosity, which decreases a species' ability to evolve to deal with change in the environment.
4. Adapting to conditions in captivity
5. Outbreeding depression
6. Fragmented populations
7. Taxonomic uncertainties, which can lead to a reprioritization of conservation efforts
8. Genetic drift as the main evolutionary process, instead of natural selection
9. Management units within species
10. Use of molecular techniques, such as allozymes as molecular markers, to analyze species in depth

Techniques

Specific genetic techniques are used to assess the genetics of a species regarding specific conservation issues as well as general population structure. This analysis can be done in two ways, with current DNA of individuals or historic DNA.

Techniques for analysiing the differences between individuals and populations include

1. Alloenzymes
2. Random Fragment Length Polymorphisms
3. Amplified Fragment Length Polymorphisms
4. Random Amplification of Polymorphic DNA
5. Single strand conformation polymorphism
6. minisatellites
7. microsatellites
8. Single nucleotide polymorphisms
9. Sequence analysis
10. DNA fingerprinting

These different techniques focus on different variable areas of the genomes within animals and plants. The specific information that is required determines which techniques are used and which parts of the genome are analysed. For examplemitochondrial DNA in animals has a high substitution rate, which makes it useful for identifying differences between individuals. However, it is only inherited in the female line, and the mitochondrial genome is relatively small. In plants, the mitochondrial DNA has very high rates of structural mutations, so is rarely used for genetic markers, as the chloroplast genome can be used instead. Other sites in the genome that are subject to high mutation

rates such as the Major Histocompatibility Complex, and the microsatellites and minisatellites are also frequently used.

These techniques can provide information on long-term conservation of genetic diversity and expound demographic and ecological matters such as taxonomy.

Another technique is using historic DNA for genetic analysis. Historic DNA is important because it allows geneticists to understand how species reacted to changes to conditions in the past. This is a key to understanding the reactions of similar species in the future.

Techniques using historic DNA include looking at preserved remains found in museums and caves. Museums are used because there is a wide range of species that are available to scientists all over the world. The problem with museums is that, historical perspectives are important because understanding how species reacted to changes in conditions in the past is a key to understanding reactions of similar species in the future. Evidence found in caves provides a longer perspective and does not disturb the animals.

Another technique that relies on specific genetics of an individual is non invasive monitoring, which uses extracted DNA from organic material that an individual leaves behind, such as a feather. This too avoids disrupting the animals and can provide information about the sex, movement, kinship and diet of an individual.

Other more general techniques can be used to correct genetic factors that lead to extinction and risk of extinction. For example, when minimizing inbreeding and increasing genetic variation multiple steps can be taken. Increasing heterozygosity through immigration, increasing the generational interval through cryopreservation or breeding from older animals, and increasing the effective population size through equalization of family size all helps minimize inbreeding and its effects. Deleterious alleles arise through mutation, however certain recessive ones can become more prevalent due to inbreeding. Deleterious mutations that arise from inbreeding can be removed by purging, or natural selection. Populations raised in captivity with the intent of being reintroduced in the wild suffer from adaptations to captivity.

Inbreeding depression, loss of genetic diversity, and genetic adaptation to captivity are disadvantageous in the wild, and many of these issues can be dealt with through the aforementioned techniques aimed at increasing heterozygosity. In addition creating a captive environment that closely resembles the wild and fragmenting the populations so there is less response to selection also help reduce adaptation to captivity.

Solutions to minimize the factors that lead to extinction and risk of extinction often overlap because the factors themselves overlap. For example, deleterious mutations are added to populations through mutation, however the deleterious mutations conservation biologists are concerned with are ones that are brought about by inbreeding, because those are the ones that can be taken care of by reducing inbreeding. Here the techniques to reduce inbreeding also help decrease the accumulation of deleterious mutations.

Applications

These techniques have wide ranging applications. One application of these specific molecular techniques is in defining species and sub-species of salmonids. Hybridization is an especially important issue in salmonids and this has wide ranging conservation, political, social and economic implications. In Cutthroat Trout mtDNA and alloenzyme analysis, hybridization between native and non-native species was shown to be one of the major factors contributing to the decline in their populations. This led to efforts to remove some hybridized populations so native populations could breed more readily. Cases like these impact everything from the economy of local fishermen to larger companies, such as timber. Specific molecular techniques led to a closer analysis of taxonomic relationships, which is one factor that can lead to extinctions if unclear.

Implications

New technology in conservation genetics has many implications for the future of conservation biology. At the molecular level, new technologies are advancing. Some of these techniques include minisatellites and MHC. These molecular techniques have wider effects from clarifying taxonomic relationships, as in the previous example, to determining the best individuals to reintroduce to a population for recovery by determining kinship. These effects then have consequences that reach even further. Conservation of species has implications for humans in the economic, social, and political realms. In the biological realm increased genotypic diversity has been shown to help ecosystem recovery, as seen in a community of grasses which was able to resist disturbance to grazing geese through greater genotypic diversity. Because species diversity increases ecosystem function, increasing biodiversity through new conservation genetic techniques has wider reaching effects than before.

A short list of studies a conservation geneticist may research include:

1. Phylogenetic classification of species, subspecies, geographic races, and populations, and measures of phylogenetic diversity and uniqueness.
2. Identifying hybrid species, hybridization in natural populations, and assessing the history and extent of introgression between species.
3. Population genetic structure of natural and managed populations, including identification of Evolutionary Significant Units (ESUs) and management units for conservation.
4. Assessing genetic variation within a species or population, including small or endangered populations, and estimates such as effective population size (Ne).
5. Measuring the impact of inbreeding and outbreeding depression, and the relationship between heterozygosity and measures of fitness.
6. Evidence of disrupted mate choice and reproductive strategy in disturbed populations.
7. Forensic applications, especially for the control of trade in endangered species.

8. Practical methods for monitoring and maximizing genetic diversity during captive breeding programs and re-introduction schemes, including mathematical models and case studies.
9. Conservation issues related to the introduction of genetically modified organisms.
10. The interaction between environmental contaminants and the biology and health of an organism, including changes in mutation rates and adaptation to local changes in the environment (e.g. industrial melanism).
11. New techniques for noninvasive genotyping.

Chapter- 4

Genetic Engineering

Genetic engineering, also called **genetic modification**, is the direct human manipulation of an organism's genetic material in a way that does not occur under natural conditions. It involves the use of recombinant DNA techniques, but does not include traditional animal and plant breeding or mutagenesis. Any organism that is generated using these techniques is considered to be a genetically modified organism. The first organisms genetically engineered were bacteria in 1973 and then mice in 1974. Insulin producing bacteria were commercialized in 1982 and genetically modified food has been sold since 1994.

The most common form of genetic engineering involves the insertion of new genetic material at an unspecified location in the host genome. This is accomplished by isolating and copying the genetic material of interest, generating a construct containing all the genetic elements for correct expression, and then inserting this construct into the host organism. Other forms of genetic engineering include gene targeting and knocking out specific genes via engineered nucleases such as zinc finger nucleases or engineered homing endonucleases.

Genetic engineering techniques have been applied in numerous fields including research, biotechnology, and medicine. Medicines such as insulin and human growth hormone are now produced in bacteria, experimental mice such as the oncomouse and the knockout mouse are being used for research purposes and insect resistant and/or herbicide tolerant crops have been commercialized. Genetically engineered plants and animals capable of producing biotechnology drugs more cheaply than current methods (called pharming) are also being developed and in 2009 the FDA approved the sale of the pharmaceutical protein antithrombin produced in the milk of genetically engineered goats.

Definition

Genetic engineering alters the genetic makeup of an organism using techniques that introduce heritable material prepared outside the organism either directly into the host or into a cell that is then fused or hybridized with the host. This involves using recombinant nucleic acid (DNA or RNA) techniques to form new combinations of heritable genetic material followed by the incorporation of that material either indirectly through a vector

system or directly through micro-injection, macro-injection and micro-encapsulation techniques. Genetic engineering does not include traditional animal and plant breeding, in vitro fertilisation, induction of polyploidy, mutagenesis and cell fusion techniques that do not use recombinant nucleic acids or a genetically modified organism in the process. Cloning and stem cell research, although not considered genetic engineering, are closely related and genetic engineering can be used within them. Synthetic biology is an emerging discipline that takes genetic engineering a step further by introducing artificially synthesized genetic material from raw materials into an organism.

If genetic material from another species is added to the host, the resulting organism is called transgenic. If genetic material from the same species or a species that can naturally breed with the host is used the resulting organism is called cisgenic. Genetic engineering can also be used to remove genetic material from the target organism, creating a knock out organism. In Europe genetic modification is synonymous with genetic engineering while within the United States of America it can also refer to conventional breeding methods.

History

Humans have altered the genomes of species for thousands of years through artificial selection and more recently mutagenesis. Genetic engineering as the direct manipulation of DNA by humans outside breeding and mutations has only existed since the 1970s. The term "genetic engineering" was first coined by Jack Williamson in his science fiction novel *Dragon's Island*, published in 1951, one year before DNA's role in heredity was confirmed by Alfred Hershey and Martha Chase, and two years before James Watson and Francis Crick showed that the DNA molecule has a double-helix structure.

In 1972 Paul Berg created the first recombinant DNA molecules by combined DNA from the monkey virus SV40 with that of the lambda virus. In 1973 Herbert Boyer and Stanley Cohen created the first transgenic organism by inserting antibiotic resistance genes into the plasmid of an *E. coli* bacterium. A year later Rudolf Jaenisch created a transgenic mouse by introducing foreign DNA into its embryo, making it the world’s first transgenic animal. In 1976 Genentech, the first genetic engineering company was founded by Herbert Boyer and Robert Swanson and a year later and the company produced a human protein (somatostatin) in *E.coli*. Genentech announced the production of genetically engineered human insulin in 1978. In 1980, the U.S. Supreme Court in the Diamond v. Chakrabarty case ruled that genetically altered life could be patented. The insulin produced by bacteria, branded humulin, was approved for release by the Food and Drug Administration in 1982.

The first field trials of genetically engineered plants occurred in France and the USA in 1986, tobacco plants were engineered to be resistant to herbicides. The People’s Republic of China was the first country to commercialize transgenic plants, introducing a virus-resistant tobacco in 1992. In 1994 Calgene attained approval to commercially release the Flavr Savr tomato, a tomato engineered to have a longer shelf life. In 1994, the European Union approved tobacco engineered to be resistant to the herbicide bromoxynil, making it

the first genetically engineered crop commercialized in Europe. In 1995, Bt Potato was approved safe by the Environmental Protection Agency, making it the first pesticide producing crop to be approved in the USA. In 2009 11 transgenic crops were grown commercially in 25 countries, the largest of which by area grown were the USA, Brazil, Argentina, India, Canada, China, Paraguay and South Africa.

In 2010, scientists at the J. Craig Venter Institute, announced that they had created the first synthetic bacterial genome, and added it to a cell containing no DNA. The resulting bacterium, named Synthia, was the world's first synthetic life form.

Process

Isolating the Gene

Elements of genetic engineering

First, the gene to be inserted into the genetically modified organism must be chosen and isolated. Presently, most genes transferred into plants provide protection against insects or tolerance to herbicides. In animals the majority of genes used are growth hormone genes. Once chosen the genes must be isolated. This typically involves multiplying the gene using polymerase chain reaction (PCR). If the chosen gene or the donor organism's genome has been well studied it may be present in a genetic library. If the DNA sequence is known, but no copies of the gene are available, it can be artificially synthesized. Once isolated, the gene is inserted into a bacterial plasmid.

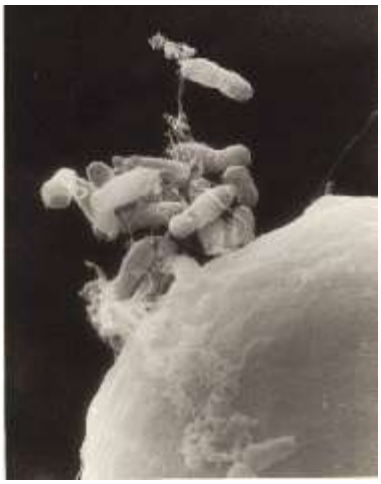
Constructs

The gene to be inserted into the genetically modified organism must be combined with other genetic elements in order for it to work properly. The gene can also be modified at this stage for better expression or effectiveness. As well as the gene to be inserted most constructs contain a promoter and terminator region as well as a selectable marker gene. The promoter region initiates transcription of the gene and can be used to control the location and level of gene expression, while the terminator region ends transcription. The selectable marker, which in most cases confers antibiotic resistance to the organism it is expressed in, is needed to determine which cells are transformed with the new gene. The constructs are made using recombinant DNA techniques, such as restriction digests, ligations and molecular cloning.

Gene Targeting

The most common form of genetic engineering involves inserting new genetic material randomly within the host genome. Other techniques allow new genetic material to be inserted at a specific location in the host genome or generate mutations at desired genomic loci capable of knocking out endogenous genes. The technique of gene targeting uses homologous recombination to target desired changes to a specific endogenous gene. This tends to occur at a relatively low frequency in plants and animals and generally requires the use of selectable markers. The frequency of gene targeting can be greatly enhanced with the use of engineered nucleases such as zinc finger nucleases, engineered homing endonucleases, or nucleases created from TAL effectors. In addition to enhancing gene targeting, engineered nucleases can also be used to introduce mutations at endogenous genes that generate a gene knockout.

Transformation



A. tumefaciens attaching itself to a carrot cell

About 1% of bacteria are naturally able to take up foreign DNA but it can also be induced in other bacteria. Stressing the bacteria for example, with a heat shock or an electric shock, can make the cell membrane permeable to DNA that may then incorporate into their genome or exist as extrachromosomal DNA. DNA is generally inserted into animal cells using microinjection, where it can be injected through the cells nuclear envelope directly into the nucleus or through the use of viral vectors. In plants the DNA is generally inserted using *Agrobacterium*-mediated recombination or biolistics.

In *Agrobacterium*-mediated recombination the plasmid construct must also contain T-DNA. *Agrobacterium* naturally inserts DNA from a tumor inducing plasmid into any susceptible plant's genome it infects, causing crown gall disease. The T-DNA region of this plasmid is responsible for insertion of the DNA. The genes to be inserted are cloned into a binary vector, which contains T-DNA and can be grown in both *E. Coli* and *Agrobacterium*. Once the binary vector is constructed the plasmid is transformed into *Agrobacterium* containing no plasmids and plant cells are infected. The *Agrobacterium* will then naturally insert the genetic material into the plant cells.

In biolistics particles of gold or tungsten are coated with DNA and then shot into young plant cells or plant embryos. Some genetic material will enter the cells and transform them. This method can be used on plants that are not susceptible to *Agrobacterium* infection and also allows transformation of plant plastids. Another transformation method for plant and animal cells is electroporation. Electroporation involves subjecting the plant or animal cell to an electric shock, which can make the cell membrane permeable to plasmid DNA. In some cases the electroporated cells will incorporate the DNA into their genome. Due to the damage caused to the cells and DNA the transformation efficiency of biolistics and electroporation is lower than agrobacterial mediated transformation and microinjection.

Selection

Not all the organism's cells will be transformed with the new genetic material; in most cases a selectable marker is used to differentiate transformed from untransformed cells. If a cell has been successfully transformed with the DNA it will also contain the marker gene. By growing the cells in the presence of an antibiotic or chemical that selects or marks the cells expressing that gene it is possible to separate the transgenic events from the non-transgenic. Another method of screening involves using a DNA probe that will only stick to the inserted gene. A number of strategies have been developed that can remove the selectable marker from the mature transgenic plant.

Regeneration

As often only a single cell is transformed with genetic material the organism must be regrown from that single cell. As bacteria consist of a single cell and reproduce clonally regeneration is not necessary. In plants this is accomplished through the use of tissue culture. Each plant species has different requirements for successful regeneration through tissue culture. If successful an adult plant is produced that contains the transgene in every cell. In animals it is necessary to ensure that the inserted DNA is present in the embryonic stem cells. When the offspring is produced they can be screened for the presence of the gene. All offspring from the first generation will be heterozygous for the inserted gene and must be mated together to produce a homozygous animal.

Confirmation

Further tests using PCR, Southern Blots and Bioassays are needed to confirm that the gene is expressed and functions correctly. The organism's offspring are also tested to ensure that the trait can be inherited and that it follows a Mendelian inheritance pattern.

Applications

Genetic engineering has applications in medicine, research, industry and agriculture and can be used on a wide range of plants, animals and micro organism.

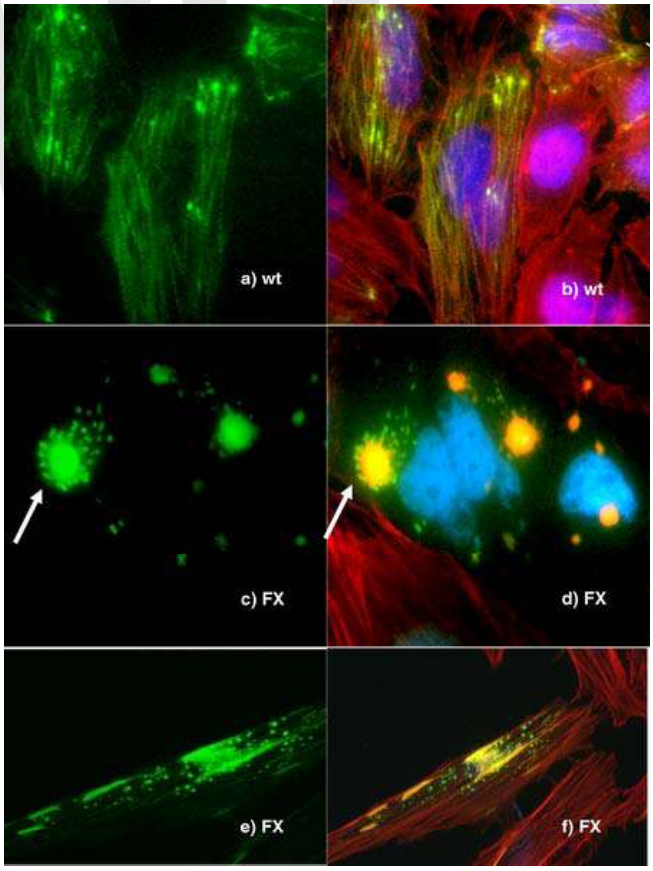
Medicine

In medicine genetic engineering has been used to mass-produce insulin, human growth hormones, follistim (for treating infertility), human albumin, monoclonal antibodies, antihemophilic factors, vaccines and many other drugs. Vaccination generally involves injecting weak live, killed or inactivated forms of viruses or their toxins into the person being immunized. Genetically engineered viruses are being developed that can still confer immunity, but lack the infectious sequences. Mouse hybridomas, cells fused together to create monoclonal antibodies, have been humanised through genetic engineering to create human monoclonal antibodies.

Genetic engineering is used to create animal models of human diseases. Genetically modified mice are the most common genetically engineered animal model. They have been used to study and model cancer (the oncomouse), obesity, heart disease, diabetes, arthritis, substance abuse, anxiety, aging and Parkinson disease. Potential cures can be tested against these mouse models. Also genetically modified pigs have been bred with the aim of increasing the success of pig to human organ transplantation.

Gene therapy is the genetic engineering of humans by replacing defective human genes with functional copies. This can occur in somatic tissue or germline tissue. If the gene is inserted into the germline tissue it can be passed down to that person's descendants. Gene therapy has been used to treat patients suffering from immune deficiencies (notably Severe combined immunodeficiency) and trials have been carried out on other genetic disorders. The success of gene therapy so far has been limited and a patient (Jesse Gelsinger) has died during a clinical trial testing a new treatment. There are also ethical concerns should the technology be used not just for treatment, but for enhancement, modification or alteration of a human beings' appearance, adaptability, intelligence, character or behavior. The distinction between cure and enhancement can also be difficult to establish. Transhumanists consider the enhancement of humans desirable.

Research



Human cells in which some proteins are fused with green fluorescent protein to allow them to be visualized

Knockout mice

Genetic engineering is an important tool for natural scientists. Genes and other genetic information from a wide range of organisms are transformed into bacteria for storage and modification, creating genetically modified bacteria in the process. Bacteria are cheap, easy to grow, clonal, multiply quickly, relatively easy to transform and can be stored at -80°C almost indefinitely. Once a gene is isolated it can be stored inside the bacteria providing an unlimited supply for research.

Organisms are genetically engineered to discover the functions of certain genes. This could be the effect on the phenotype of the organism, where the gene is expressed or what other genes it interacts with. These experiments generally involve loss of function, gain of function, tracking and expression.

- **Loss of function experiments**, such as in a gene knockout experiment, in which an organism is engineered to lack the activity of one or more genes. A knockout experiment involves the creation and manipulation of a DNA construct *in vitro*, which, in a simple knockout, consists of a copy of the desired gene, which has been altered such that it is non-functional. Embryonic stem cells incorporate the altered gene, which replaces the already present functional copy. These stem cells are injected into blastocysts, which are implanted into surrogate mothers. This allows the experimenter to analyze the defects caused by this mutation and thereby determine the role of particular genes. It is used especially frequently in developmental biology. Another method, useful in organisms such as Drosophila (fruit fly), is to induce mutations in a large population and then screen the progeny for the desired mutation. A similar process can be used in both plants and prokaryotes.
- **Gain of function experiments**, the logical counterpart of knockouts. These are sometimes performed in conjunction with knockout experiments to more finely establish the function of the desired gene. The process is much the same as that in knockout engineering, except that the construct is designed to increase the function of the gene, usually by providing extra copies of the gene or inducing synthesis of the protein more frequently.
- **Tracking experiments**, which seek to gain information about the localization and interaction of the desired protein. One way to do this is to replace the wild-type gene with a 'fusion' gene, which is a juxtaposition of the wild-type gene with a reporting element such as green fluorescent protein (GFP) that will allow easy visualization of the products of the genetic modification. While this is a useful technique, the manipulation can destroy the function of the gene, creating secondary effects and possibly calling into question the results of the experiment.

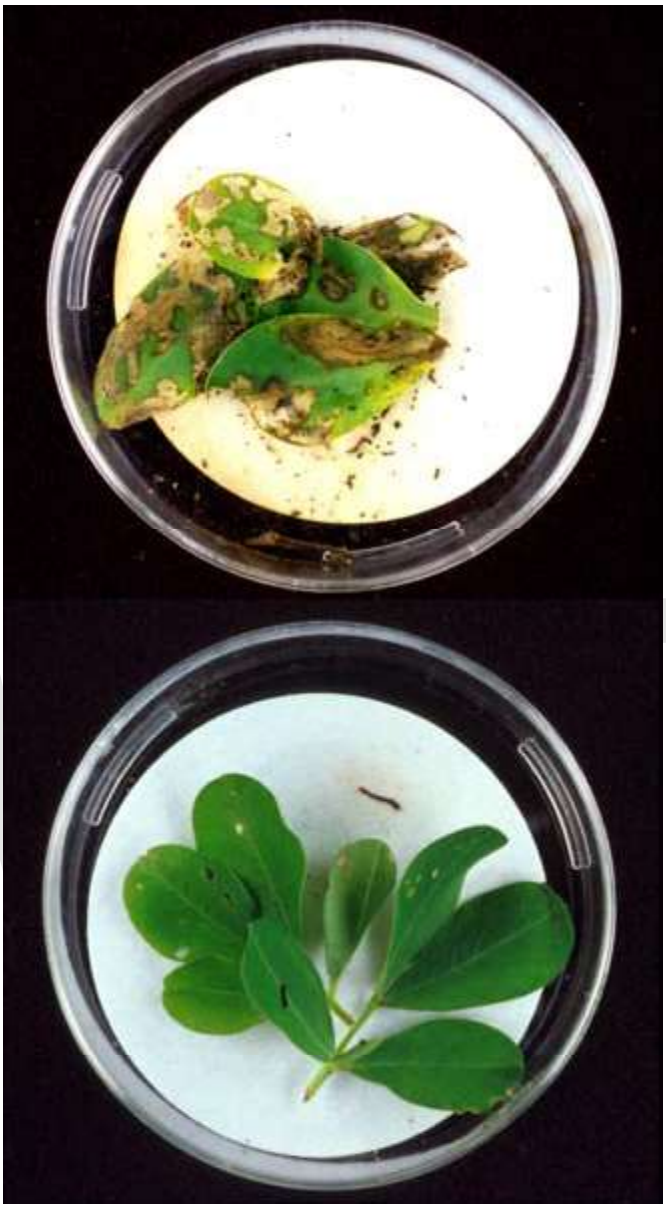
More sophisticated techniques are now in development that can track protein products without mitigating their function, such as the addition of small sequences that will serve as binding motifs to monoclonal antibodies.

- **Expression studies** aim to discover where and when specific proteins are produced. In these experiments, the DNA sequence before the DNA that codes for a protein, known as a gene's promoter, is reintroduced into an organism with the protein coding region replaced by a reporter gene such as GFP or an enzyme that catalyzes the production of a dye. Thus the time and place where a particular protein is produced can be observed. Expression studies can be taken a step further by altering the promoter to find which pieces are crucial for the proper expression of the gene and are actually bound by transcription factor proteins; this process is known as promoter bashing.

Industrial

By engineering genes into bacterial plasmids it is possible to create a biological factory that can produce proteins and enzymes. Some genes do not work well in bacteria, so yeast, a eukaryote, can also be used. Bacteria and yeast factories have been used to produce medicines such as insulin, human growth hormone, and vaccines, supplements such as tryptophan, aid in the production of food (chymosin in cheese making) and fuels. Other applications involving genetically engineered bacteria being investigated involve making the bacteria perform tasks outside their natural cycle, such as cleaning up oil spills, carbon and other toxic waste.

Agriculture



Bt-toxins present in peanut leaves (bottom image) protect it from extensive damage caused by European corn borer larvae (top image).

One of the best-known and controversial applications of genetic engineering is the creation of genetically modified food. There are three generations of genetically modified crops. First generation crops have been commercialized and most provide protection from insects and/or resistance to herbicides. There are also fungal and virus resistant crops developed or in development. They have been developed to make the insect and weed management of crops easier and can indirectly increase crop yield.

The second generation of genetically modified crops being developed aim to directly improve yield by improving salt, cold or drought tolerance and to increase the nutritional

value of the crops. The third generation consists of pharmaceutical crops, crops that contain edible vaccines and other drugs. Some agriculturally important animals have been genetically modified with growth hormones to increase their size while others have been engineered to express drugs and other proteins in their milk.

The genetic engineering of agricultural crops can increase the growth rates and resistance to different diseases caused by pathogens and parasites. This is beneficial as it can greatly increase the production of food sources with the usage of fewer resources that would be required to host the world's growing populations. These modified crops would also reduce the usage of chemicals, such as fertilizers and pesticides, and therefore decrease the severity and frequency of the damages produced by these chemical pollution.

Ethical and safety concerns have been raised around the use of genetically modified food. A major safety concern relates to the human health implications of eating genetically modified food, in particular whether toxic or allergic reactions could occur. Gene flow into related non-transgenic crops, off target effects on beneficial organisms and the impact on biodiversity are important environmental issues. Ethical concerns involve religious issues, corporate control of the food supply, intellectual property rights and the level of labeling needed on genetically modified products.

Other uses

In materials science, a genetically modified virus has been used to construct a more environmentally friendly lithium-ion battery. Some bacteria have been genetically engineered to create black and white photographs while others have potential to be used as sensors by expressing a fluorescent protein under certain environmental conditions. Genetic engineering is also being used to create BioArt and novelty items such as blue roses, and glowing fish.

Opposition and criticism

A 2010 study of Canola found transgenes in 80% of wild (uncultivated or "feral") varieties in North Dakota, meaning 80% of the plants which had established themselves in the area were genetically engineered varieties. The researchers stated that "we found the highest densities of [such transgene-containing] plants near agricultural fields and along major freeways, but we were also finding plants in the middle of nowhere" adding that "over time,..the build-up of different types of herbicide resistance in feral [natural] canola and closely related weeds, like field mustard, could make it more difficult to manage these plants using herbicides."

Chapter- 5

Heritability of IQ

The study of the **heritability of IQ** is a field of research that includes biology, genomics, psychology, philosophy, sociology, and anthropology. Heritability is "an estimate of the genetic and environmental contributions to the *variance* of any phenotypic measure around the mean for a given population." "Heritability refers to the genetic contribution to variance within a population and in a specific environment . . . ; if the environment changes, the heritability measure changes."

Some contend that heritability does not set any limit on how malleable the trait is under changes of environment, because "even highly heritable traits can be strongly manipulated by the environment, so heritability has little if anything to do with controllability.". However, others argue that heritability constrains malleability. The debate about IQ heritability touches on the nature versus nurture divide, and there has been significant controversy in the academic community about it ever since research began in the 19th century.

IQ is a polygenic trait under normal circumstances according to recent research. However, destructive mutation of individual genes associated with development can severely affect intelligence, with Phenylketonuria as an example.

Estimates in the academic research of the heritability of IQ have varied from below 0.5 to a high of 0.9. A 1996 statement by the American Psychological Association gave about .45 for children and about .75 during and after adolescence. A 2004 meta-analysis of reports in *Current Directions in Psychological Science* gave an overall estimate of around .85 for 18-year-olds and older. *The New York Times Magazine* has listed about three quarters as a figure held by the majority of studies.

Methods and results

Heritability calculations

Background

Heritability is defined as the proportion of variance in a trait which is attributable to genotype within a defined population in a specific environment. Heritability takes a value ranging from 0 to 1; a heritability of 1 indicates that all variation in the trait in question is genetic in origin and a heritability of 0 indicates that *none* of the variation is genetic. The determination of many traits can be considered primarily genetic under similar environmental backgrounds. For example, Visscher *et al.* (2006) found that adult height has a heritability estimated at 0.80, when a similar environmental background is present, to control for environment the study only looked at the contribution of heritability to variation within families. The paper stated that "one can never be sure that the estimates are correct, because nature and nurture can be confounded without one knowing it. The authors got around this problem by comparing the similarity between relatives as a function of the exact proportion of genes that they have in common, looking only within families." Other traits have low heritabilities, which indicate a large relative environmental influence. For example, a twin study on the heritability of depression in men calculated it as 0.29, while it was 0.42 for women in the same study.

Heritability for a trait is calculated by measuring how strongly traits covary in people of a given genetic and environmental similarity. The most common method is to consider identical twins reared apart, with any similarities which exists between such twin pairs attributed to genotype. In terms of correlation statistics, this means that theoretically the correlation of tests scores between monozygotic twins would be 1.00 if genetics alone accounted for variation in IQ scores; likewise, siblings and dizygotic twins share on average half of their alleles and the correlation of their scores would be 0.50 if IQ were affected by genes alone. Practically, however, the upper bound of these correlations are given by the reliability of the test, which tends to be 0.90 to 0.95 for typical IQ tests Thus, the actual heritability of IQ will tend to be slightly higher than attained by estimates derived from studies of monozygotic twins, though this effect is small.

In the case of the inheritance of IQ or a certain degree of giftedness, the relatives of probands with a high IQ exhibit a comparably high IQ with a much higher probability than the general population. In 1982, Bouchard and McGue reviewed such correlations reported in 111 original studies in the United States. The mean correlation of IQ scores between monozygotic twins was 0.86, between siblings, 0.47, between half-siblings, 0.31, and between cousins, 0.15. From such data the heritability of IQ was estimated at anywhere between 0.40 and 0.80 in the United States. The reason for this wide margin appeared to be that the heritability of IQ rises through childhood and adolescence, peaking at 0.68 and 0.78 in adults, leaving the overwhelming majority of IQ differences between individuals to be explained genetically.

The finding of rising heritability with age is counter-intuitive; it is reasonable to expect that genetic influences on traits like IQ should become less important as one gains experiences with age. However, that the opposite occurs is well documented. According to work by Robert Plomin, heritability estimates calculated on infant samples are as low as 20%, rising to around 40% in middle childhood, and ultimately as high as 80% in adult samples in the United States. This suggests that the underlying genes actually express themselves by affecting a person's predisposition to build, learn, and develop mental abilities throughout the lifespan.

Estimates and caveats to them

In 2006, *The New York Times Magazine* listed about three quarters as a figure held by the majority of studies, while a 2004 meta-analysis of reports in *Current Directions in Psychological Science* gave an overall estimate of around .85 for 18-year-olds and older. As well, a 1996 statement by the American Psychological Association gave about .45 for children and about .75 during and after adolescence.

The 2006 edition of *Assessing adolescent and adult intelligence* by Alan S. Kaufman and Elizabeth O. Lichtenberger reports correlations of 0.86 for identical twins raised together compared to 0.76 for those raised apart and 0.47 for siblings. A 1994 review in *Behavior Genetics* based on identical/fraternal twin studies found that it is as high as 0.80 in general cognitive ability but it also varies based on the trait, with .60 for verbal tests, .50 for spatial and speed-of-processing tests, and only .40 for memory tests.

There are a number of points to consider when interpreting heritability:

- A high heritability does not mean that the environment has no effect on the development of a trait, or that learning is not involved. Vocabulary size, for example, is very substantially heritable (and highly correlated with general intelligence) although every word in an individual's vocabulary is learned. In a society in which plenty of words are available in everyone's environment, especially for individuals who are motivated to seek them out, the number of words that individuals actually learn depends to a considerable extent on their genetic predispositions.
- A common error is to assume that because something is heritable it is necessarily unchangeable. This is wrong. Heritability does not imply immutability. As previously noted, heritable traits can depend on learning, and they may be subject to other environmental effects as well. The value of heritability can change if the distribution of environments (or genes) in the population is substantially altered. For example, an impoverished or suppressive environment could fail to support the development of a trait, and hence restrict individual variation. This could affect estimates of heritability. Another example is Phenylketonuria which previously caused mental retardation for everyone who had this genetic disorder. Today, this can be prevented by following a modified diet.
- On the other hand, there can be effective environmental changes that do not change heritability at all. If the environment relevant to a given trait improves in a

way that affects all members of the population equally, the mean value of the trait will rise without any change in its heritability (because the differences among individuals in the population will stay the same). This has evidently happened for height: the heritability of stature is high, but average heights continue to increase.
- Even in developed nations, high heritability of a trait within a given group has no necessary implications for the source of a difference between groups.

- In addition to strong evidence for heritability increasing with age, some studies suggest that heritability increases with social class,.Differences among children with higher than average social status are almost entirely due to inherited differences, while among very low social class groups, most of the differences in IQ-scores (at least in children and young adolescents) are attributable to differences between families (shared environment).
- Even among close groups such as families, different individuals—such as siblings—will still experience different environments, which matter in determining intelligence scores.

Test score differences

Intelligence tests measure many important abilities, such as verbal and quantitative reasoning, and can predict socially-relevant outcomes such as academic performance and occupational outcomes. However, intelligence test scores do not reflect all of the intricacies of the everyday meaning of intelligence, so researchers take care to distinguish between IQ test results and intelligence.

Some studies of intelligence tests use statistical methods to extract so-called latent variables from the IQ test scores. One such variable is the general intelligence factor, or *g*, which accounts for most of the differences in IQ test scores between individuals. There are other latent variables in addition to *g*, and IQ tests vary in their ability to measure these latent variables, if they measure them at all. IQ tests scores, while often summarized as a single overall number, are actually multidimensional in nature. Transforming IQ test scores into latent variables is an attempt to find one or dimensions on which to compare IQ test scores.

IQ scores can vary substantially for the same person, even on tests taken at the same age. (IQ score table data and pupil pseudonyms adapted from description of KABC-II norming study cited in Kaufman 2009.)

Pupil	KABC-II	WISC-III	WJ-III
Asher	90	95	111
Brianna	125	110	105
Colin	100	93	101
Danica	116	127	118
Elpha	93	105	93
Fritz	106	105	105

Georgi	95	100	90
Hector	112	113	103
Imelda	104	96	97
Jose	101	99	86
Keoku	81	78	75
Leo	116	124	102

Latent variables are also sometimes called factors or constructs. The construct validity of an IQ test score is a key criteria for judging whether the IQ test score differences are meaningful. Tests which do not measure difference in latent variables for some group are said to have measurement bias. The construct validity of most commonly used IQ tests has been fairly well established within multiple racial-ethnic groups in developing countries such as the United States. That is, test score difference within each racial-ethnic group are valid indicators of differences in latent variables such as *g*. A related question is whether test-score differences between groups are valid. There is a consensus that test score differences between Black and White people in the United States have predictive validity (also called predictive invariance), meaning that test scores predict the same socially-relevant outcomes regardless of the race of the person being tested. To further address this question, three studies using sophisticated statistical techniques have shown that Black-White differences in IQ test scores are not a result of measurement bias (a criterion called measurement invariance).

These studies imply that Black-White IQ differences reflect very general differences in some underlying latent variables, but they are unable to differentiate precisely which latent variables differ under a variety of models. These studies were performed in response to previous investigations which suggested that Black-White IQ differences are primarily differences in *g* in particular.

United States

There are observed differences in average test score achievement between racial-ethnic groups, which vary depending on the populations studied and the type of tests used. Self-defined black and white United States citizens have been the subjects of the greatest number of studies. Black-white average IQ differences appear to increase with age, reaching an average of nearly 17 points by age 24, which is slightly more than one standard deviation. According to James Flynn and others, the overall average black-white gap has reduced by one-third over the course of the 20th century.

For example, the black men inducted into the U.S. armed forces during World War II averaged about 1.5 standard deviations below their white counterparts. This improvement is also reflected in black-white differences on school achievement tests, which have shrunk from about 1.2 to about 0.8 standard deviations. However, these improvements may have stalled for people born after the early 1970s.

The average black-white IQ difference also varies depending on test content. For example, two subsections of the WISC IQ test, known as forward and reverse digit-span, ask children to repeat a long series of numbers either forwards or backwards. The black-white difference on forward digit span is relatively small, while the difference on reverse digit span is relatively large. Across a battery of tests, the size of the black-white gap is correlated with the extent to which the tests measure the psychometric factor *g*, which also accounts for most of the variation in interindividual differences in IQ test performance. Gaps are seen in other tests of cognitive ability or aptitude, including university admission exams such as the SAT and GRE as well as employment tests for corporate settings and the military.

The IQ distributions of other racial and ethnic groups in the United States are less well studied. Hispanic and Native American populations, including Arctic Natives, tend to score worse on average than white populations but better on average than black populations. East Asian populations may score higher on average than white populations in the United States as they do elsewhere. A 1960 study of 1,236 American teenagers calculated six IQ measures for Jews relative to white gentiles. The results found that the relative IQ of American Jews varied from a low of 91.3 (visual reasoning) to a high of 109.7 (mathematics). A recent review by Lynn (2004) used a 10 word vocabulary test to estimate the IQ of American Jews. The population of 150 Jews scored half a standard deviation above the 5,300 white gentiles in verbal IQ.

For each of these populations, there is some evidence that the mixture of ability factors that distinguish individuals are differentially distributed between groups. For example, East Asian populations tend to outscore white populations in performance IQ, whereas the test score differences skew towards higher verbal IQ for Ashkenazi Jew-white differences. However, the mixture of abilities within groups appears to be nearly identical across many ethnic groups. The stability of these differences is also less well studied than black-white differences.

Worldwide

According to Richard Lynn, J. Philippe Rushton, and others, IQ test score differences are observed cross-culturally and around the world. Lynn has published three books summarizing IQ test scores from around the world. The inaccuracy of the cross cultural IQ scores is well documented, but many scholars use the results as an estimate of worldwide IQ scores. Lynn's meta-analysis lists East Asians (105), Europeans (99), Inuit (91), Southeast Asians and Amerindians (87 each), Pacific Islanders (85), South Asians/North Africans (84), Non-Bushmen sub-Saharan Africans (67), Australian Aborigines (62) and Bushmen (54). However, critics point out that different researchers using different data get different results. For example, James Flynn found out that Lynn's comparisons between nations were skewed and that the IQ values should be recalculated.

International achievement test scores, including TIMSS and PISA, have also been used to estimate average IQ worldwide with similar results where data is available.

The very low IQ scores reported for sub-Saharan African populations are especially controversial. For example, Wicherts argues that the average IQ of sub-Saharan Africans is poorly measured and is more likely 78. According to anthropologist Mark Cohen, the frequently reported African mean IQ of 70 is "preposterous". Using Western standards, this would mean that African countries evidencing such a low IQ would be largely dysfunctional. Given that individuals in these countries lead "vibrant artistic, symbolic and spiritual lives", this is, according to Cohen, clearly not the case. Thus, he concludes, the IQ test results from Africa do not reflect actual intelligence levels.

Differences in education, prolonged malnutrition, exposure to toxin, exposure to stress, and exposure to disease are all generally expected to contribute to the lower scores observed in developing countries. However, direct experimental evidence to confirm the role of individual factors is difficult to acquire in most cases because each of these factors tends to also be associated with one another and with unfavorable socioeconomic conditions. In the case of some toxins, such as lead, a negative effect on IQ scores has been established. Two other factors that have as well established negative association with IQ are severely premature birth and severe low birth weight.

Developing nations

Almost all studies on heritability have been in the developed world, mostly in the United States. In developing nations there are many environmental factors affecting IQ which are much less important in developed nations. Examples include nutrition, diseases, environmental toxins, and health care. For example, iodine deficiency causes a fall, in average, of 12 IQ points in China.

Issues in the calculations

Family environment

In the developed world, nearly all personality traits show that, contrary to some expectations, environmental effects actually cause non-related children raised in the same family ("adoptive siblings") to be as different as children raised in different families (Harris, 1998; Plomin & Daniels, 1987). There are some family effects on the IQ of children, accounting for up to a quarter of the variance. However, by adulthood, this correlation disappears, such that adoptive siblings are not more similar in IQ than strangers, while adult full siblings show an IQ correlation of 0.6. Twin studies reinforce this pattern: monozygotic (identical) twins raised separately are highly similar in IQ (0.86), more so than dizygotic (fraternal) twins raised together (0.6) and much more than adoptive siblings (~0.0).

The American Psychological Association's report *Intelligence: Knowns and Unknowns* (1995) states that there is no doubt that normal child development requires a certain minimum level of responsible care. Severely deprived, neglectful, or abusive environments must have negative effects on a great many aspects of development,

including intellectual aspects. Beyond that minimum, however, the role of family experience is in serious dispute.

There is no doubt that such variables as resources of the home and parents' use of language are correlated with children's IQ scores, but such correlations may be mediated by genetic as well as (or instead of) environmental factors. But how much of that variance in IQ results from differences between families, as contrasted with the varying experiences of different children in the same family? Recent twin and adoption studies suggest that while the effect of the family environment is substantial in early childhood, it becomes quite small by late adolescence. These findings suggest that differences in the life styles of families whatever their importance may be for many aspects of children's lives make little long-term difference for the skills measured by intelligence tests. It also stated

"We should note, however, that low-income and non-white families are poorly represented in existing adoption studies as well as in most twin samples. Thus it is not yet clear whether these studies apply to the population as a whole. It remains possible that, across the full range of income and ethnicity, between-family differences have more lasting consequences for psychometric intelligence."

A study of French children adopted between the ages of four and six shows the continuing interplay of nature and nurture. The children came from poor backgrounds with IQs that initially averaged 77, putting them near retardation. Nine years later after adoption, they retook the IQ tests, and all of them did better. The amount they improved was directly related to the adopting family's socioeconomic status. "Children adopted by farmers and laborers had average IQ scores of 85.5; those placed with middle-class families had average scores of 92. The average IQ scores of youngsters placed in well-to-do homes climbed more than 20 points, to 98."

Biased older studies?

Stoolmiller (1999) found that the range restriction of family environments that goes with adoption, that adopting families tend to be more similar on for example SES than the general population, means that role of the shared family environment have been underestimated in previous studies. Corrections for range applied to adoption studies indicate that SE could account for as much as 50% of the variance in IQ. However, the effect of restriction of range on IQ for adoption studies was examined by Matt McGue and colleagues, who write that "restriction in range in parent disinhibitory psychopathology and family SES had no effect on adoptive-sibling correlations [in] IQ".

Eric Turkheimer and colleagues (2003), not using an adoption study, included impoverished US families. Results demonstrated that the proportions of IQ variance attributable to genes and environment vary nonlinearly with SES. They suggest that the role of shared environmental factors may have been underestimated in older studies which often only studied affluent middle class families.

When comparing late 1970s to pre-1963 recorded data, researches DeFries and Plomin found that IQ correlation between parent and child living together fell significantly, from 0.50 to 0.35.

Maternal (fetal) environment

A meta-analysis, by Devlin and colleagues in *Nature* (1997), of 212 previous studies evaluated an alternative model for environmental influence and found that it fits the data better than the 'family-environments' model commonly used. The shared maternal (foetal) environment effects, often assumed to be negligible, account for 20% of covariance between twins and 5% between siblings, and the effects of genes are correspondingly reduced, with two measures of heritability being less than 50%. They argue that the shared maternal environment may explain the striking correlation between the IQs of twins, especially those of adult twins that were reared apart.

Bouchard and McGue reviewed the literature in 2003, arguing that Devlin's conclusions about the magnitude of heritability is not substantially different than previous reports and that their conclusions regarding prenatal effects stands in contradiction to many previous reports. They write that:

Chipuer et al. and Loehlin conclude that the postnatal rather than the prenatal environment is most important. The Devlin et al. (1997a) conclusion that the prenatal environment contributes to twin IQ similarity is especially remarkable given the existence of an extensive empirical literature on prenatal effects. Price (1950), in a comprehensive review published over 50 years ago, argued that almost all MZ twin prenatal effects produced differences rather than similarities. As of 1950 the literature on the topic was so large that the entire bibliography was not published. It was finally published in 1978 with an additional 260 references. At that time Price reiterated his earlier conclusion (Price, 1978).

Dickens and Flynn model

Dickens and Flynn (2001) argue that the arguments regarding the apparent absence of influence from shared family environment should apply equally well to groups separated in time. This is contradicted by the Flynn effect. Changes between generations have happened too quickly to be explained by genetics. This paradox can be explained by observing that heritability estimates include both a direct effect of the genotype on IQ and also indirect effects through which the genotype changes the environment, in turn effecting IQ. That is, those with a higher IQ tend to seek out stimulating environments that further increase IQ. The direct effect can initially have been very small but feedback loops can create large differences in IQ. In their model an environmental stimulus can have a very large effect on IQ, even in adults, but this effect also decays over time unless the stimulus continues. (The model is also adaptable to possible factors, such as nutrition in early childhood, that may cause permanent effects.) The Flynn effect can be explained by a generally more stimulating environment for all people. The authors suggest that programs aiming to increase IQ would be most likely to produce long-term IQ gains if

they taught children how to replicate outside the program the kinds of cognitively demanding experiences that produce IQ gains while they are in the program and motivate them to persist in that replication long after they have left the program.

Regression toward the mean

The heritability of IQ measures the extent to which the IQ of a child is measurably influenced by the IQ of its parents. As IQ is a quantifiable phenotype, one can estimate the expected IQ of child using the equation $\hat{y} = \bar{x} + h^2 \left(\frac{m+f}{2} - \bar{x} \right)$, where

- $\hat{y}$ is the expected IQ of the child,
- $\bar{x}$ is the mean IQ of the population to which the parents belong,
- h^2 is the heritability of IQ,
- m and f are the IQs of the mother and father, respectively.

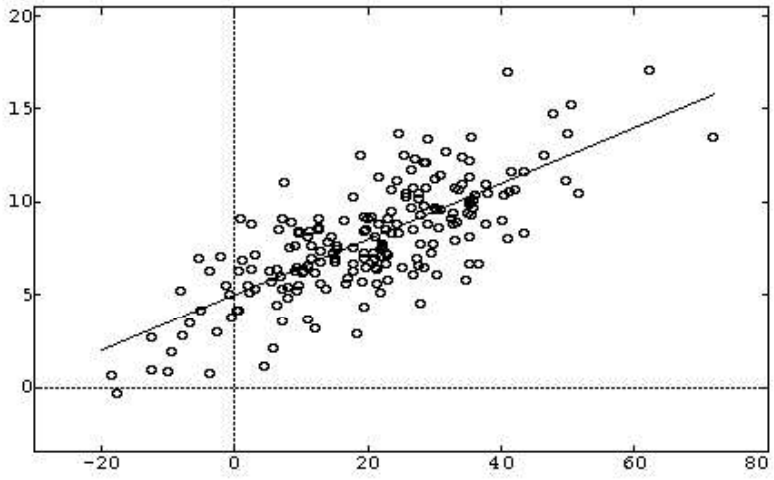


Illustration of linear regression on a data set

The equation asserts that, on average, the IQ of a child tends to the mean IQ of the population. For instance, if the heritability of IQ is 50% and the mean IQ of a population is 100, then a couple with an average IQ of 120 will, on average, have a child with an IQ of 110. Similarly, a couple with an average IQ of 80 will, on average, have a child with an IQ of 90.

It is noted that the above equation relates only statistical averages and is not deterministic. Furthermore, the equation is a general equation based in the inheritance of genetically-based characteristics (in this case, phenotypes), and so it is implicitly assumed that environmental factors are, for the sake of correctly assessing the genetic contribution to IQ, the same across the population.

Operating under the assumption that child and parent are raised in exactly the same environment (unlikely, but usually closer to the truth than in the completely dissimilar environment that the previous equation assumes), h^2 can be replaced by h, which is simply the correlation between parent and offspring IQ. In this case, regression towards the mean is no longer partially caused by environmental differences and therefore only by random genetic variation.

Finally, it is important to note that the expected IQ of the offspring is normally distributed around the mean calculated using the above equation, so in many cases regression towards the mean does not actually occur; as the values are normally distributed, there is a chance that offspring IQ will be more deviant from the mean than that of the parental average.

Historical research

As early as 1869, Francis Galton replaced mere speculations by statistical data through his book, *Hereditary Genius*:

Highly Gifted Men and the Percentage of their Highly Gifted Male Relatives
(classified by occupation and achievement)

	Galton	**Terman**	**Brimhall**		**Weiss**
	%	**%**	**%**	**%**	***n***
Probands	100	84[+]	100	97[+]	1972: *1329* 1994: *357*
Fathers	26	41	29	40	*346*
Brothers	47	–	49	49	*220*
Sons	60	64[*]	–	55	*77*
Grandfathers	14	–	9	9	*681*
Uncles	16	–	13	14	*615*
Nephews	23	–	–	22	*76*
Grandsons	14	–	–	–	–
Greatgrandfathers	0	–	–	4	*1290*
Uncles of the parents	5	–	–	5	*1996*
Cousins	16	–	9[#]	18	*570*
Greatgrandsons	7	–	–	–	–
Cousins of parents	–	–	–	11	*2250*

"+": classified by occupation; 100%, if classified by test
"*": classified only by IQ; classification by occupation gives about 55%; $n = 820$.
"#": some cousins were still too young and did not have full opportunity to become distinguished
"–": no data

Despite the differences in methods and societies, there is a notable parallelism in the published statistics. The ITO-method by Li and Sacks (1954) allows from this set of data the estimation of the underlying number of genes and their allele frequencies.

The inheritance of cognitive deficits

There are many genetic variants known to cause lower IQ. The number of such mutations already known is in the hundreds. For example, an allele of the gene GDI1 is associated with an IQ below 70.

Copy number variation has also been associated with idiopathic learning disability.

There are number of known cases where the homozygotes have severe cognitive deficits and the heterozygotes show a small decrease of IQ. In such cases further alleles are investigated to estimate their influence on IQ. For example, one minor allele of the gene ALDH5A1 is associated with an IQ difference of around 1.5 points.

Interindividual (between individuals) differences in learning ability are also known in mice, dogs and other animals, and the achievements of pure strains can be improved by selective breeding. In such a way also behavior genetics is contributing to our knowledge of the inheritance of mental traits.

The search for specific genes

Unfortunately, most of the research done about the heritability of intelligence have focused on children and young adults. Thus, the role of genetic factors to intelligence is mostly unknown. Many studies attempting to find loci in the genome relating to IQ have had little success. For example, a study by Robert Plomin using groups of around 100 people investigated 1,842 DNA markers in a high-IQ group and in an average-IQ control group. The study used a five-step replication process to eliminate false positives, and no gene met this rigid criterion for replicability.

The failure to find a specific gene associated with IQ indicates that cognitive abilities are very complex and are likely to involve several genes (polygenic). Some estimate that as much as 40% of all genes may contribute to IQ. The more genes that contribute to a trait the more the trait will be continuous instead of discrete. A 2008 study of 500,000 single nucleotide polymorphisms (SNPs) from 7,089 children did not substantially improve on earlier studies. The study did not find any SNPs that accounted for more than 0.5% of the variance in general intelligence.

A 2007 study did find that a gene called FADS2 along with breastfeeding adds about 7 IQ points to those with the "C" version of the gene. Those with the "G" version see no advantage.

There is "a highly significant association" between the CHRM2 gene and intelligence according to a 2006 Dutch family study. The study concluded that there was an

association between the CHRM2 gene on chromosome 7 and Performance IQ, as measured by the Wechsler Adult Intelligence Scale-Revised. The Dutch family study used a sample of 667 individuals from 304 families. A similar association was found independently in the *Minnesota Twin and Family Study* (Comings et al. 2003) and by the Department of Psychiatry at the Washington University. Microcephalin and ASPM are two genes that are associated with brain development. Mutations in these genes are associated with microcephaly, and hence they were initially associated with general intelligence. However recent studies have found no association with general cognitive abilities.

STX1A correlates significantly with intelligence in Williams syndrome patients.

Between-group heritability

Although IQ differences between individuals are shown to have a large genetic component, it does not automatically follow that mean group-level disparities (between-group differences) in IQ can be assumed to have a genetic basis. An analogy, attributed to Richard Lewontin, illustrates this point:

Suppose two handfuls are taken from a sack containing a genetically diverse variety of corn, and each grown under carefully controlled and standardized conditions, except that one batch is lacking in certain nutrients that are supplied to the other. After several weeks, the plants are measured. There is variability of growth within each batch, due to the genetic variability of the corn. Given that the growing conditions are closely controlled, nearly all the variation in the height of the plants within a batch will be due to differences in their genes. Thus, within populations, heritabilities will be very high. Nevertheless, the difference between the two groups is due entirely to an environmental factor - differential nutrition. Lewontin didn't go so far as to have the one set of pots painted white and the other set black, but you get the idea. The point of the example, in any case, is that the causes of between-group differences may in principle be quite different from the causes of within-group variation.

This nurture argument holds an intuitive appeal but is misleading since modern statistical comparisons control for such factors as differences in socio-economic and parental educational backgrounds specifically to eliminate such environmental effects and thereby isolate the genetic influence. In fact, differences in IQ between reportedly discrete genetic groups has been observed, and researchers, such as Arthur Jensen, maintain that environmental differences are too small to account for these differences. They propose, therefore, that genetic differences must provide the primary explanation.

This view is challenged by Peter Schönemann who claims that Arthur Jensen and others routinely confuse the first principal component (PC1) with *g* as Charles Spearman defined it. Schönemann argues that the high IQ heritability estimates reported in the literature derive from restrictive formal models whose underlying assumptions are rarely tested and usually violated by the data. Jensen's view is also rebutted by James Flynn in his book *What Is Intelligence*.

The issue of observed between-group IQ differences is controversial *vis-a-vis* considerations on both the nature of race and the meaning and measurement of intelligence. Important related questions include whether intelligence can be accurately described by a single number, and whether the nature of intelligence is the same across cultures.

Gene-by-Environment Interaction

Researchers have recently begun to empiricaly examine the hypothesis that genetic influences on IQ may depend on environmental inputs.

Eric Turkheimer and colleagues (2003) studied the heritability of IQ in a sample that included a substantial proportion of impoverished US families. Results demonstrated that, in seven-year-old twins, the proportions of IQ variance attributable to genes and environment vary nonlinearly with socioeconomic status. In impoverished families, 60% of the variance in early childhood IQ was accounted for by the shared family environment, and the contribution of genes is close to zero; in affluent families, the result is almost exactly the reverse. They suggest that the role of shared environmental factors may have been underestimated in older studies which often only studied affluent middle class families.

Harden and colleagues (2007) found a similar gene-environment interaction for adolescents. They found that, among higher income families, genetic influences accounted for approximately 55% of the variance in cognitive aptitude and shared environmental influences about 35%. Among lower income families, the proportions were in the reverse direction, 39% genetic and 45% shared environment."

Using a nationally representative sample of young twins, Tucker-Drob and colleagues (2011) found that genetic influences on mental ability emerged over the course of infancy, with larger increases in genetic variance for children living in higher socioeconomic status homes. At 10 months of age genetic influences on mental ability were negligible for children from all socioeconomic backgrounds. By 2 years of age, genes accounted for nearly 50% of the variation in mental ability of children raised in high-SES homes, but genes continued to account for negligible variation in mental ability of children raised in low-SES homes. These results suggest that family resources help to potentiate children's genetic capacities for intellectual development.

A 2007 study by Caspi and colleagues found that a gene called FADS2 along with breastfeeding adds about 7 IQ points to those with the "C" version of the gene. Those with the "G" version see no advantage.

Chapter- 6

Genomics

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

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

History

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

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

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

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

As of September 2007, the complete sequence was known of about 1879 viruses, 577 bacterial species and roughly 23 eukaryote organisms, of which about half are fungi. Most of the bacteria whose genomes have been completely sequenced are problematic disease-causing agents, such as *Haemophilus influenzae*. Of the other sequenced species, most were chosen because they were well-studied model organisms or promised to become good models. Yeast (*Saccharomyces cerevisiae*) has long been an important model organism for the eukaryotic cell, while the fruit fly *Drosophila melanogaster* has been a very important tool (notably in early pre-molecular genetics). The worm *Caenorhabditis elegans* is an often used simple model for multicellular organisms. The zebrafish *Brachydanio rerio* is used for many developmental studies on the molecular level and the flower *Arabidopsis thaliana* is a model organism for flowering plants. The Japanese pufferfish (*Takifugu rubripes*) and the spotted green pufferfish (*Tetraodon nigroviridis*) are interesting because of their small and compact genomes, containing very little non-coding DNA compared to most species. The mammals dog (*Canis familiaris*), brown rat (*Rattus norvegicus*), mouse (*Mus musculus*), and chimpanzee (*Pan troglodytes*) are all important model animals in medical research.

Human genomics

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

Bacteriophage genomics

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

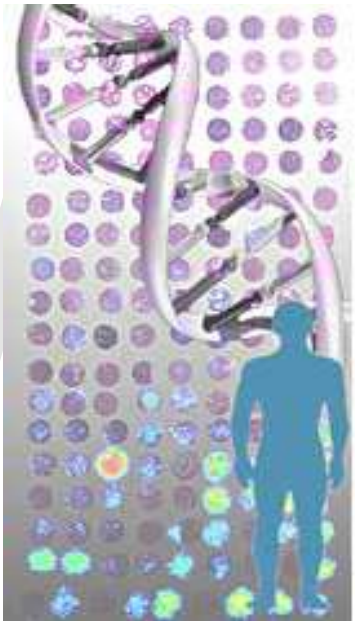
become prominent, thereby enabling researchers to understand the mechanisms underlying phage evolution. Bacteriophage genome sequences can be obtained through direct sequencing of isolated bacteriophages, but can also be derived as part of microbial genomes. Analysis of bacterial genomes has shown that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

Cyanobacteria genomics

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

Chapter- 7

Human Genetics



A small piece of human DNA

Human genetics describes the study of inheritance as it occurs in human beings. Human genetics encompasses a variety of overlapping fields including: classical genetics, cytogenetics, molecular genetics, biochemical genetics, genomics, population genetics, developmental genetics, clinical genetics, and genetic counseling. Genes can be the common factor of the qualities of most human-inherited traits. Study of human genetics can be useful as it can answer questions about human nature, understand the diseases and development of effective disease treatment, and understand genetics of human life.

Genetic differences and inheritance patterns

Inheritance of traits for humans are based upon Gregor Mendel's model of inheritance. Mendel deduced that inheritance depends upon discrete units of inheritance, called factors or genes.

Autosomal dominant inheritance

Autosomal traits are associated with a single gene on an autosome (non-sex chromosome)—they are called "dominant" because a single copy—inherited from either parent—is enough to cause this trait to appear. This often means that one of the parents must also have the same trait, unless it has arisen due to a new mutation. Examples of autosomal dominant traits and disorders are Huntington's disease, and achondroplasia.

Autosomal recessive inheritance

Autosomal recessive traits is one pattern of inheritance for a trait, disease, or disorder to be passed on through families. For a recessive trait or disease to be displayed two copies of the trait or disorder needs to be presented. The trait or gene will be located on a non-sex chromosome. Because it takes two copies of a trait to display a trait, many people can unknowingly be carriers of a disease. From an evolutionary perspective, a recessive disease or trait can remain hidden for several generations before displaying the phenotype. Examples of autosomal recessive disorders are albinism, Cystic Fibrosis, Tay-Sachs disease.

X-linked and Y-linked inheritance

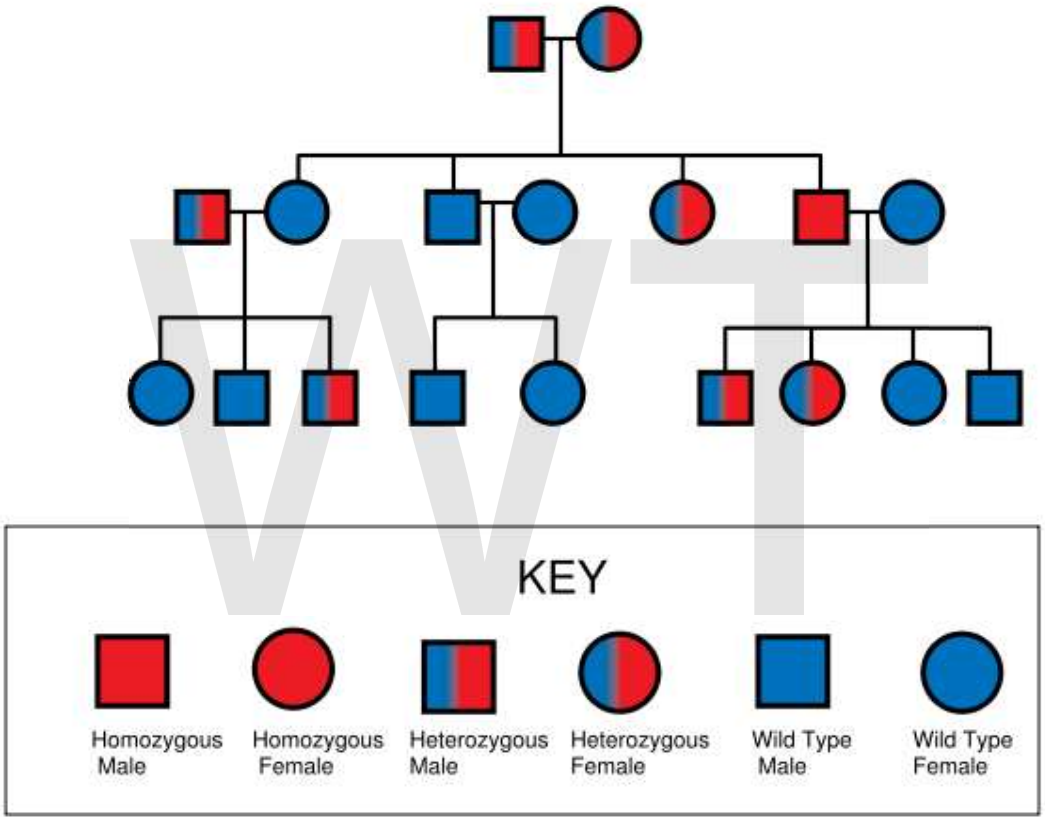
X-linked genes are found on the sex X chromosome. X-linked genes just like autosomal genes have both dominant and recessive types. Recessive X-linked disorders are rarely seen in females and usually only affect males. This is because males inherit their X chromosome and all X-linked genes will be inherited from the maternal side. Fathers only pass on their Y chromosome to their sons, so no X-linked traits will be inherited from father to son. Females express X-linked disorders when they are homozygous for the disorder and become carriers when they are heterozygous. X-linked dominant inheritance will show the same phenotype as a heterozygote and homozygote. Just like X-linked inheritance, there will be a lack of male-to-male inheritance, which makes it distinguishable from autosomal traits. One example of a X-linked trait is Coffin-Lowry syndrome, which is caused by a mutation in ribosomal protein gene. This mutation results in skeletal, craniofacial abnormalities, mental retardation, and short stature.

X chromosomes in females undergo a process known as X inactivation. X inactivation is when one of the two X chromosomes in females is almost completely inactivated. It is important that this process occurs otherwise a woman would produce twice the amount of normal X chromosome proteins. The mechanism for X inactivation will occur during the embryonic stage. For people with disorders like trisomy X, where the genotype has three X chromosomes, X-inactivation will inactivate all X chromosomes until there is only one X chromosome active. X inactivation is not only limited to females, males with Klinefelter syndrome, who have an extra X chromosome, will also undergo X inactivation to have only one completely active X chromosome.

Y-linked inheritance occurs when a gene, trait, or disorder is transferred through the Y chromosome. Since Y chromosomes can only be found in males, Y linked traits are only

passed on from father to son. The testis determining factor, which is located on the Y chromosome, determines the maleness of individuals. Besides the maleness inherited in the Y-chromosome there are no other found Y-linked characteristics.

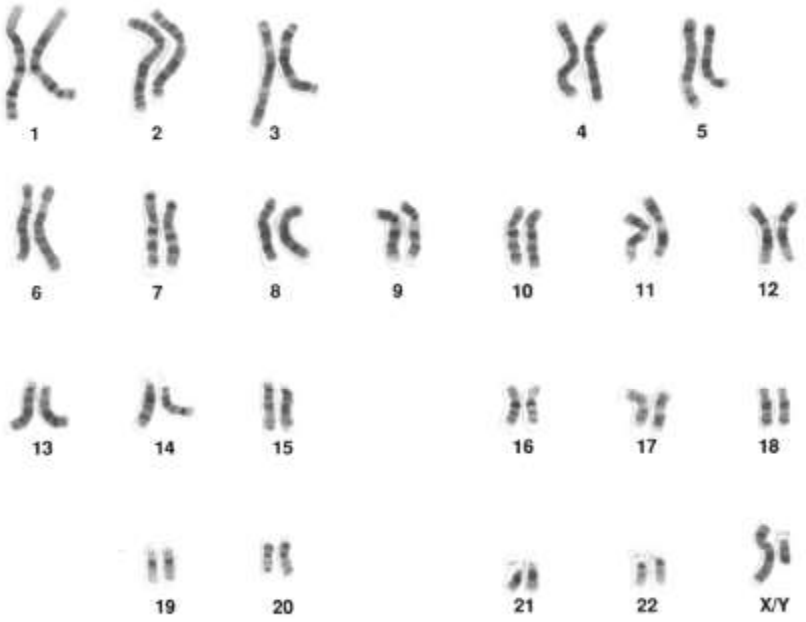
Pedigrees



An example of a family pedigree displaying an autosomal recessive trait

A pedigree is a diagram showing the ancestral relationships and transmission of genetic traits over several generations in a family. Pedigrees are used to help detect many different genetic diseases. A pedigree can also be used to help determine the chances for a parent to produce an offspring with a specific trait. Four different traits can be identified by pedigree chart analysis: autosomal dominant, autosomal recessive, x-linked, or y-linked. Partial penetrance can be shown and calculated form pedigrees. Penetrance is the percentage expressed frequency with which individuals of a given genotype manifest at

least some degree of a specific mutant phenotype associated with a trait. Inbreeding, the mating between closely related organisms of traits can clearly be seen on pedigree charts. Pedigree charts of royal families have a high degree of inbreeding, because it was customary and preferable for royalty to marry another member of royalty. Genetic councilors commonly use pedigrees to help couple determine if the parents will be able to produce healthy children.



A karyotype of a human male, showing 46 chromosomes including XY sex chromosomes

Karyotype

A karyotype is a very useful tool in cytogenetics. A karyotype is picture of all the chromosomes in the metaphase stage arranged according to length and centromere position. A karyotype can also be useful in clinical genetics, due to its ability to diagnose genetic disorders. On a normal karyotype, aneuploidy can be detected by clearly being able to observe any missing or extra chromosomes. Giemsa banding, g-banding, of the karyotype can be used to detect deletions, insertions, duplications, inversions, and translocations. G-banding will stain the chromosomes with light and dark bands unique to each chromosome. A FISH, fluorescent in situ hybridization, can be used to observe deletions, insertions, and translocations. FISH uses fluorescent probes to bind to specific sequences of the chromosomes that will cause the chromosomes to fluoresce a unique color.

Genomics

Genomics refers to the field of genetics concerned with structural and functional studies of the genome. A genome is all the DNA contained within an organism or a cell

including nuclear and mitochondrial DNA. The human genome is the total collection of genes in a human being contained in the human chromosome, composed of over three billion nucleotides. In April 2003, the Human Genome Project was able to sequence all the DNA in the human genome, to discover the human genome was composed around 20,000 protein coding genes.

Population genetics

Population genetics is the branch of evolutionary biology responsible for investigating processes that cause changes in allele and genotype frequencies in populations based upon Mendelian inheritance. Four different forces can influence the frequencies: natural selection, mutation, gene flow (migration), and genetic drift. A population can be defined as a group of interbreeding individuals and their offspring. For human genetics the populations will consist only of the human species. The Hardy-Weinberg principle is a widely used principle to determine allelic and genotype frequencies.

Hardy-Weinberg principle

The Hardy-Weinberg principle states that when no evolution occurs in a population the allele and genotype frequencies do not change from one generation to the next. No evolution refers to no mutation, no gene flow, no natural selection, and no genetic drift. To be in equilibrium two more assumptions need to be made that random mating occurs and there are discrete, non-overlapping generations.

Mitochondrial DNA

In addition to nuclear DNA, humans (like almost all eukaryotes) have mitochondrial DNA. Mitochondria, the "power houses" of a cell, have their own DNA because they are descended from a proteobacterium that merged with eukaryotic cells over 2 billion years ago—an assertion known as the endosymbiotic hypothesis. Mitochondria are inherited from one's mother, and its DNA is frequently used to trace maternal lines of descent. Mitochondrial DNA is only 16kb in length and encodes for 62 genes.

Genes and human characteristics

Genes are a fundamental unit of inheritance. Genes can be defined as a sequence of DNA in the genome that is required for production of a functional product. Genes have both minor and major effects on human characteristics. Human genes have become prominent in the nature versus nurture debate.

Genes and behavior

Genes have a strong influence on human behavior. IQ is largely heritable. However, this has been questioned. The stance that humans inherit substantial behavioral characteristics

is called psychological nativism, compared to the stance that human behavior and culture are virtually entirely constructed (tabula rasa).

In the early 20th century, eugenics was policy in parts of the United States and Europe. The goal was to reduce or eliminate traits that were considered undesirable. One form of eugenics was compulsory sterilization of people deemed mentally unfit. Hitler's eugenics programs turned social consciousness against the practice, and psychological nativism became associated with racism and sexism.

Evolutionary psychology

Evolutionary psychology explains many human behaviors as more or less moderated by genes that evolved in the hunter-gatherer stage of human cultural development.

Genetic disorders

Humans have several genetic diseases, often blamed on rare recessive genes. A few examples of human genetic diseases are: Turner Syndrome, Huntington's disease, Downs Syndrome (in some cases), and sickle cell anemia.

- Cri du Chat syndrome – A disorder caused from a deletion on the short arm of chromosome 5. This deletion results in a phenotype of mental retardation, behavioral problems, and a cat like call. About one in every 50,000 births will have the syndrome.
- Huntington's disease – A neurological disorder caused by a trinucleotide repeat sequence. Huntingtons is an autosomal dominant trait. Most individuals with the disease will first display the phenotype around 40 years of age. The symptoms are jerky uncontrollable movements, mental retardation, and behavioral problems.
- Turner syndrome – A condition that affects females caused by a 45, XO genotype instead of the normal XX genotype. These individuals have only one X chromosome. These individuals are phenotypically female, but will be sterile due to undeveloped ovaries.
- Klinefelter syndrome – A disorder in males caused by the presence of an extra X chromosome. These individuals have a genotype of 47, XXY instead of the normal XY genotype. The symptoms for this syndrome are enlarged breasts, small testes, and sterility.

Human traits with simple inheritance patterns

Dominant	Recessive
Widow's peak	No Widow's peak
Facial Dimples	No Facial Dimples
Able to taste PTC	Unable to taste PTC
Unattached earlobe	Attached earlobe
Cleft chin	No Cleft chin

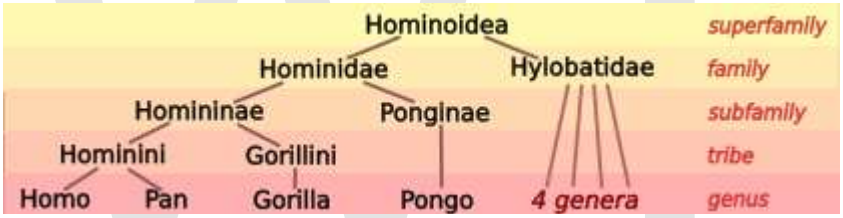
Brunette iris (anatomy)	Blue Iris (anatomy)
Color Vision	Color blindness
Brown Hair	Blonde Hair
normal	turned up nose
Ability to roll tongue (Able to hold tongue in a U shape)	No ability to roll tongue
Normal Pinkies	Crooked Pinkies
Normal Thumb	Hitchhiker's Thumb
Freckles	No Freckles
Wet-type earwax	Dry-type earwax
Curly Hair	Straight Hair

Chapter- 8

Human Evolutionary Genetics

Human evolutionary genetics studies how one human genome differs from the other, the evolutionary past that gave rise to it, and its current effects. Differences between genomes have anthropological, medical and forensic implications and applications. Genetic data can provide important insight into human evolution.

Origin of apes

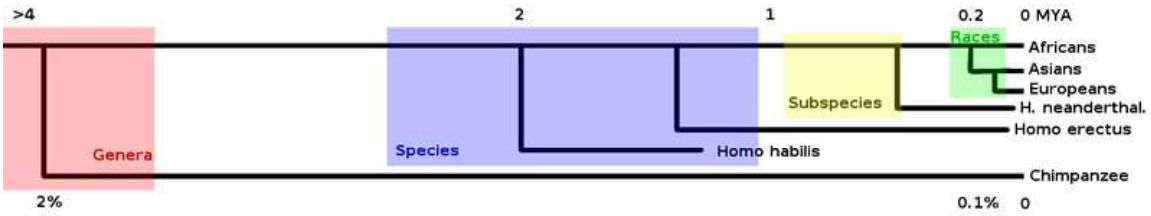


Taxonomic relationships of hominoids

Biologists classify humans, along with only a few other species, as great apes (species in the family Hominidae). The Hominidae include two distinct species of chimpanzee (the bonobo, *Pan paniscus*, and the common chimpanzee, *Pan troglodytes*), two species of gorilla (the western gorilla, *Gorilla gorilla*, and the eastern gorilla, *Gorilla graueri*), and two species of orangutan (the Bornean orangutan, *Pongo pygmaeus*, and the Sumatran orangutan, *Pongo abelii*).

Apes, in turn, belong to the primates order (>400 species). Data from both mitochondrial DNA (mtDNA) and nuclear DNA (nDNA) indicates that primates belong to the group of Euarchontoglires, together with Rodentia, Lagomorpha, Dermoptera, and Scandentia. This is further supported by Alu-like short interspersed nuclear elements (SINEs) which have been found only in members of the Euarchontoglires.

Cladistics



A phylogenetic tree like the one shown above is usually derived from DNA or protein sequences from populations. Often mitochondrial DNA or Y chromosome sequences are used to study ancient human demographics. These single-locus sources of DNA do not recombine and are almost always inherited from a single parent, with only one known exception in mtDNA (Schwartz and Vissing 2002). Individuals from the various continental groups tend to be more similar to one another than to people from other continents. The tree is rooted in the common ancestor of chimpanzees and humans, which is believed to have originated in Africa. Horizontal distance in the diagram corresponds to two things:

1. **Genetic distance**. Given below the diagram, the genetic difference between humans and chimps is less than 2%, or 20 times larger than the variation among modern humans.
2. **Temporal remoteness** of the most recent common ancestor. Rough estimates are given above the diagram, in millions of years. The mitochondrial most recent common ancestor of modern humans lived roughly 200,000 years ago, latest common ancestors of humans and chimps between four and seven million years ago.

Chimpanzees and humans belong to different genera, indicated in red. Formation of species and subspecies is also indicated, and the formation of "races" is indicated in the green rectangle to the right (note that only a very rough representation of human phylogeny is given). Note that vertical distances are not meaningful in this representation.

Speciation of humans and the African apes

The separation of humans from their closest relatives, the African apes (chimpanzees and gorillas), has been studied extensively for more than a century. Five major questions have been addressed:

- Which apes are our closest ancestors?
- When did the separations occur?
- What was the effective population size of the common ancestor before the split?
- Are there traces of population structure (subpopulations) preceding the speciation or partial admixture succeeding it?

- What were the specific events (including fusion of chromosomes 2a and 2b) prior to and subsequent to the separation?

General observations

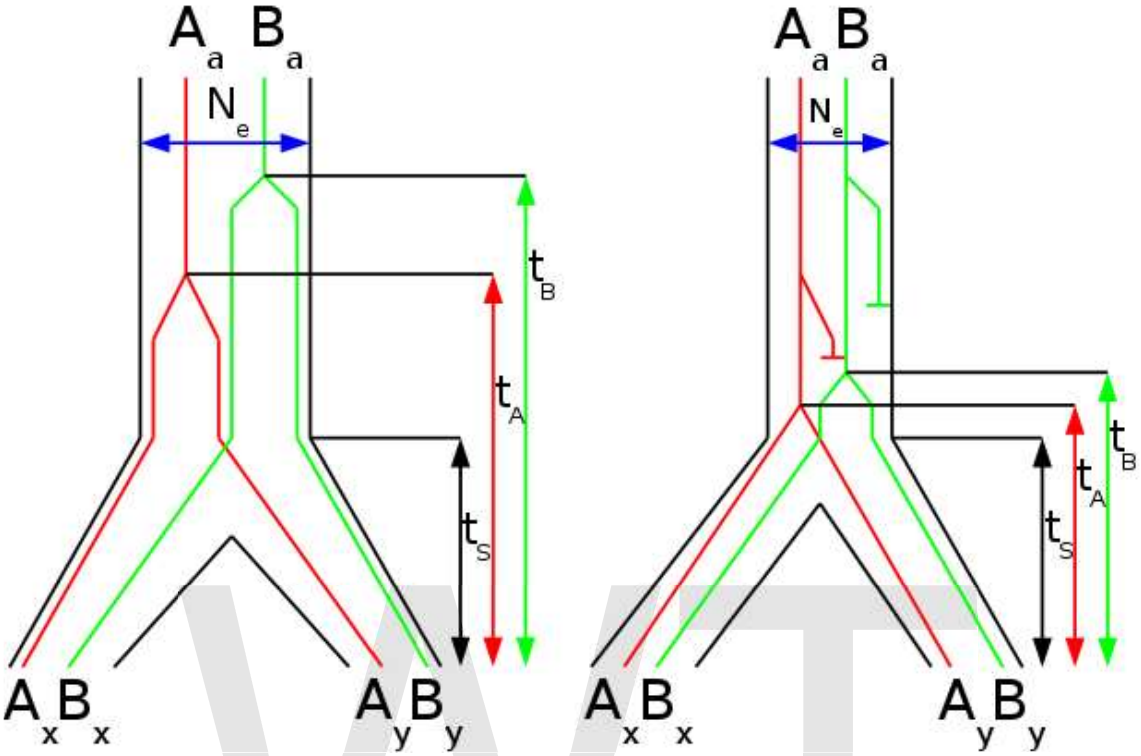
As discussed before, different parts of the genome show different sequence divergence between different hominoids. It has also been shown that the sequence divergence between DNA from humans and chimpanzees varies greatly. For example the sequence divergence varies between 0% to 2.66% between non-coding, non-repetitive genomic regions of humans and chimpanzees. Additionally gene trees, generated by comparative analysis of DNA segments, do not always fit the species tree. Summing up:

- The sequence divergence varies significantly between humans, chimpanzees and gorillas.
- For most DNA sequences, humans and chimpanzees appear to be most closely related, but some point to a human-gorilla or chimpanzee-gorilla clade.
- The human genome has been sequenced, as well as the chimpanzee genome. Humans have 23 pairs of chromosomes, while chimpanzees, gorillas, and orangutans have 24. Human chromosome 2 is a fusion between two chromosomes that remained separate in the other primates.

Divergence times

The divergence time of humans from other apes is of great interest. One of the first molecular studies, published in 1967 measured immunological distances (IDs) between different primates. Basically the study measured the strength of immunological response that an antigen from one species (human albumin) induces in the immune system of another species (human, chimpanzee, gorilla and Old World monkeys). Closely related species should have similar antigens and therefore weaker immunological response to each other's antigens. The immunological response of a species to its own antigens (e.g. human to human) was set to be 1. The ID between humans and gorillas was determined to be 1.09, that between humans and chimpanzees was determined as 1.14. However the distance to six different Old World monkeys was on average 2.46 indicating that the African apes are far closer related to humans than to monkeys. The authors consider the divergence time between Old World monkeys and hominoids to be 30 million years ago (MYA), based on fossil data, and the immunological distance was considered to grow at a constant rate. They concluded that divergence time of humans and the African apes to be roughly ~5 MYA. That was a surprising result. Most scientists at that time thought that humans and great apes diverged much earlier (>15 MYA). The gorilla was, in ID terms, closer to human than to chimpanzees, however the difference was so slight that the trichotomy could not be resolved with certainty. Later studies based on molecular genetics were able to resolve the trichotomy: chimpanzees are phylogenetically closer to humans than to gorillas. However, the divergence times estimated later (using much more sophisticated methods in molecular genetics) do not substantially differ from the very first estimate in 1967.

Divergence times and ancestral effective population size

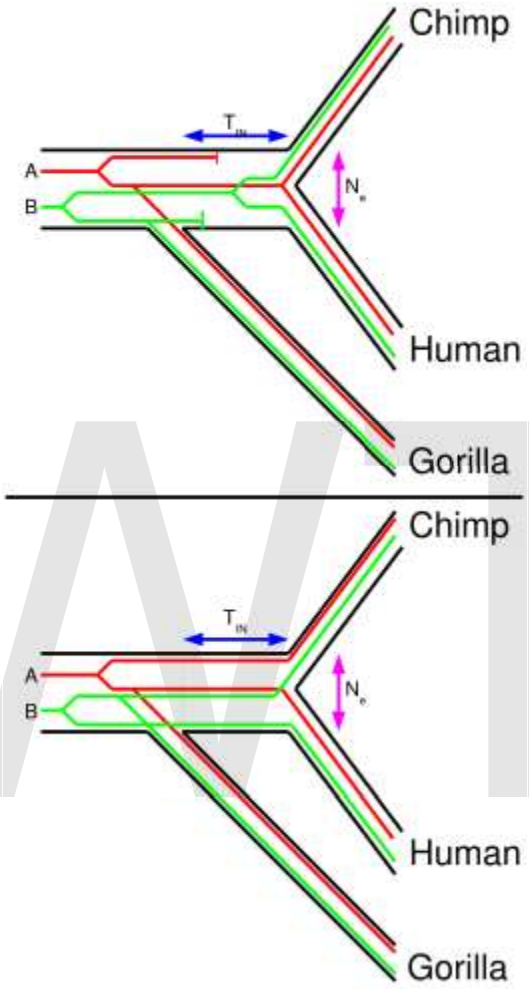


The sequences of the DNA segments diverge earlier than the species. A large effective population size in the ancestral population (left) preserves different variants of the DNA segments (=alleles) for a longer period of time. Therefore, on average, the gene divergence times (t_A for DNA segment A; t_B for DNA segment B) will deviate more from the time the species diverge (t_S) compared to a small ancestral effective population size (right).

Current methods to determine divergence times use DNA sequence alignments and molecular clocks. Usually the molecular clock is calibrated assuming that the orangutan split from the African apes (including humans) 12-16 MYA. Some studies also include some old world monkeys and set the divergence time of them from hominoids to 25-30 MYA. Both calibration points are based on very little fossil data and have been criticized. If these dates are revised, the divergence times estimated from molecular data will change as well. However, the relative divergence times are unlikely to change. Even if we can't tell absolute divergence times exactly, we can be pretty sure that the divergence time between chimpanzees and humans is about sixfold shorter than between chimpanzees (or humans) and monkeys.

One study (Takahata *et al.*, 1995) used 15 DNA sequence from different regions of the genome from human and chimpanzee and 7 DNA sequences from human, chimpanzee and gorilla. They determined that chimpanzees are more closely related to humans than gorillas. Using various statistical methods, they estimated the divergence time human-chimp to be 4.7 MYA and the divergence time between gorillas and humans (and

chimps) to be 7.2 MYA. Additionally they estimated the effective population size of the common ancestor of humans and chimpanzees to be ~100,000. This was somewhat surprising since the present day effective population size of humans is estimate to be only ~10,000. If true that means that the human lineage would have experienced an immense decrease of its effective population size (and thus genetic diversity) in its evolution.



A and B are two different loci. In the upper figure they fit to the species tree. The DNA that is present in today's gorillas diverged earlier from the DNA that is present in today's humans and chimps. Thus both loci should be more similar between human and chimp than between gorilla and chimp or gorilla and human. In the lower graph, locus A has a more recent common ancestor in human and gorilla compared to the chimp sequence. Whereas chimp and gorilla have a more recent common ancestor for locus B. Here the gene trees are incongruent to the species tree.

Another study (Chen & Li, 2001) sequenced 53 non-repetitive, intergenic DNA segments from a human, a chimpanzee, a gorilla, and orangutan. When the DNA sequences were concatenated to a single long sequence, the generated neighbor-joining tree supported the *Homo-Pan* clade with 100% bootstrap (that is that humans and chimpanzees are the

closest related species of the four). When three species are fairly closely related to each other (like human, chimpanzee and gorilla), the trees obtained from DNA sequence data may not be congruent with the tree that represents the speciation (species tree). The shorter internodal time span (T_{IN}) the more common are incongruent gene trees. The effective population size (N_e) of the internodal population determines how long genetic lineages are preserved in the population. A higher effective population size causes more incongruent gene trees. Therefore, if the internodal time span is known, the ancestral effective population size of the common ancestor of humans and chimpanzees can be calculated.

When each segment was analyzed individually, 31 supported the *Homo-Pan* clade, 10 supported the *Homo-Gorilla* clade, and 12 supported the *Pan-Gorilla* clade. Using the molecular clock the authors estimated that gorillas split up first 6.2-8.4 MYA and chimpanzees and humans split up 1.6-2.2 million years later (internodal time span) 4.6-6.2 MYA. The internodal time span is useful to estimate the ancestral effective population size of the common ancestor of humans and chimpanzees.

A parsimonious analysis revealed that 24 loci supported the *Homo-Pan* clade, 7 supported the *Homo-Gorilla* clade, 2 supported the *Pan-Gorilla* clade and 20 gave no resolution. Additionally they took 35 protein coding loci from databases. Of these 12 supported the *Homo-Pan* clade, 3 the *Homo-Gorilla* clade, 4 the *Pan-Gorilla* clade and 16 gave no resolution. Therefore only ~70% of the 52 loci that gave a resolution (33 intergenic, 19 protein coding) support the 'correct' species tree. From the fraction of loci which did not support the species tree and the internodal time span they estimated previously, the effective population of the common ancestor of humans and chimpanzees was estimated to be ~52 000 to 96 000. This value is not as high as that from the first study (Takahata), but still much higher than present day effective population size of humans.

A third study (Yang, 2002) used the same dataset that Chen and Li used but estimated the ancestral effective population of 'only' ~12,000 to 21,000, using a different statistical method.

Genetic differences between humans and other great apes

The alignable sequences within genomes of humans and chimpanzees differ by about 35 million single nucleotide substitutions. Additionally about 3% of the complete genomes differ by deletions, insertions and duplications.

Since mutation rate is relatively constant, roughly one half of these changes occurred in the human lineage. Only a very tiny fraction of those fixed differences gave rise to the different phenotypes of humans and chimpanzees and finding those is a great challenge. The vast majority of the differences are neutral and do not affect the phenotype.

Molecular evolution may act in different ways, through protein evolution, gene loss, differential gene regulation and RNA evolution. All are thought to have played some part in human evolution.

Gene loss

Many different mutations can inactivate a gene, but few will change its function in a specific way. Inactivation mutations will therefore be readily available for selection to act on. Gene loss could thus be a common mechanism of evolutionary adaptation (the "less-is-more" hypothesis).

80 genes were lost in the human lineage after separation from the last common ancestor with the chimpanzee. 36 of those were for olfactory receptors. Genes involved in chemoreception and immune response are overrepresented. Another study estimated that 86 genes had been lost.

Hair keratin gene KRTHAP1

A gene for type I hair keratin was lost in the human lineage. Keratins are a major component of hairs. Humans still have nine functional type I hair keratin genes but the loss of that particular gene may have caused the thinning of human body hair. The gene loss occurred relatively recently in human evolution—less than 240,000 years ago.

Myosin gene MYH16

Stedman *et al.* (2004) stated that the loss of the sarcomeric myosin gene MYH16 in the human lineage led to smaller masticatory muscles. They estimated that the mutation that led to the inactivation (a two base pair deletion) occurred 2.4 million years ago, predating the appearance of *Homo ergaster/erectus* in Africa. The period that followed was marked by a strong increase in cranial capacity, promoting speculation that the loss of the gene may have removed an evolutionary constraint on brain size in the genus *Homo*.

Another estimate for the loss of the MYH16 gene is 5.3 million years ago, long before *Homo* appeared.

Other

- CASPASE12, a cysteinyl aspartate proteinase

Gene addition

Segmental duplications (SDs or LCRs) have had roles in creating new primate genes and shaping human genetic variation.

Selection pressures

Human accelerated regions are areas of the genome that differ between humans and chimpanzees to a greater extent than can be explained by genetic drift over the time since the two species shared a common ancestor. These regions show signs of being subject to natural selection, leading to the evolution of distinctly human traits. Two examples are HAR1F, which is believed to be related to brain development and HAR2 (a.k.a HACNS1) that may have played a role in the development of the opposable thumb.

Genetic differences between humans and Neanderthals

An international group of scientists completed a draft sequence of the Neanderthal genome in May 2010. The results indicate some breeding between humans and Neanderthals as the genomes of non-African humans have 1-4% more in common with Neanderthals than do the genomes of subsaharan Africans. Neanderthals and most humans share a lactose-intolerant variant of the lactase gene that encodes an enzyme that is unable to break down lactose in milk after weaning. Humans and Neanderthals also share the FOXP2 gene variant associated with brain development and with speech in humans, indicating that Neanderthals may have been able to speak. Chimps have two amino acid differences in FOXP2 compared with human and Neanderthal FOXP2.

Sequence divergence between humans and apes

The draft sequence of the common chimpanzee genome published in the summer 2005 showed the regions that are similar enough to be aligned with one another account for 2400 million of the human genome's 3164.7 million bases – that is, 75.8% of the genome. This 75.8% of the human genome is 1.23% different from the chimpanzee genome in single nucleotide polymorphisms (changes of single DNA "letters" in the genome). Another type of difference, called indels (insertions/deletions) account for another ~3 % difference between the alignable sequences. In addition, variation in copy number of large segments (> 20 kb) of similar DNA sequence provides a further 2.7% difference between the two species. Hence the total similarity of the genomes could be as low as about 70%.

The figures above do not take into account differences in the organization of the alignable sequences within the genomes of humans and chimps. Short stretches of alignable sequence may be in very different orders and locations within the two genomes. At present we cannot fully assess the difference in structure of the two genomes, because the human genome was used as a scaffold when the chimpanzee draft genome was assembled. When genomes are sequenced, relatively short sequences of DNA are produced, and these sequences have to be fitted together like a jigsaw puzzle. This requires multiple overlapping reads to accurately assemble the overall sequence. The human genome sequence is relatively accurate, with 8 to 9-fold coverage, but the chimpanzee draft genome only has 3.6-fold coverage. The human genome was sequenced using a hierarchical shotgun method which can deal with duplications and difficult-to-

assemble sequences better than the whole genome shotgun method that was used for the chimpanzee draft genome. The human genome was used as a template for the assembly of the draft chimpanzee genome, on the assumption that the two genomes would be similar.

Almost half of that 1.23% SNP change belongs to the human at 0.53%, whose genetic variance is lower than a chimp, and just over half to the chimp at 0.7%. If we also take into account that random "genetic drift" takes up the bulk of the 0.54% difference, then that percentage difference where SNPs have a potential positive impact on human abilities, is between 0.01% and 0.02%. The bonobo is a sibling species of common chimpanzee and is genetically about as different from humans as are common chimps.

Percentage sequence divergence between humans and other hominids

Locus	Human-Chimp	Human-Gorilla	Human-Orangutan
Alu elements	2	-	-
Non-coding (Chr. Y)	1.68 ± 0.19	2.33 ± 0.2	5.63 ± 0.35
Pseudogenes (autosomal)	1.64 ± 0.10	1.87 ± 0.11	-
Pseudogenes (Chr. X)	1.47 ± 0.17	-	-
Noncoding (autosomal)	1.24 ± 0.07	1.62 ± 0.08	3.08 ± 0.11
Genes (K_s)	1.11	1.48	2.98
Introns	0.93 ± 0.08	1.23 ± 0.09	-
Xq13.3	0.92 ± 0.10	1.42 ± 0.12	3.00 ± 0.18
Subtotal for X chromosome	1.16 ± 0.07	1.47 ± 0.08	-
Genes (K_a)	0.8	0.93	1.96

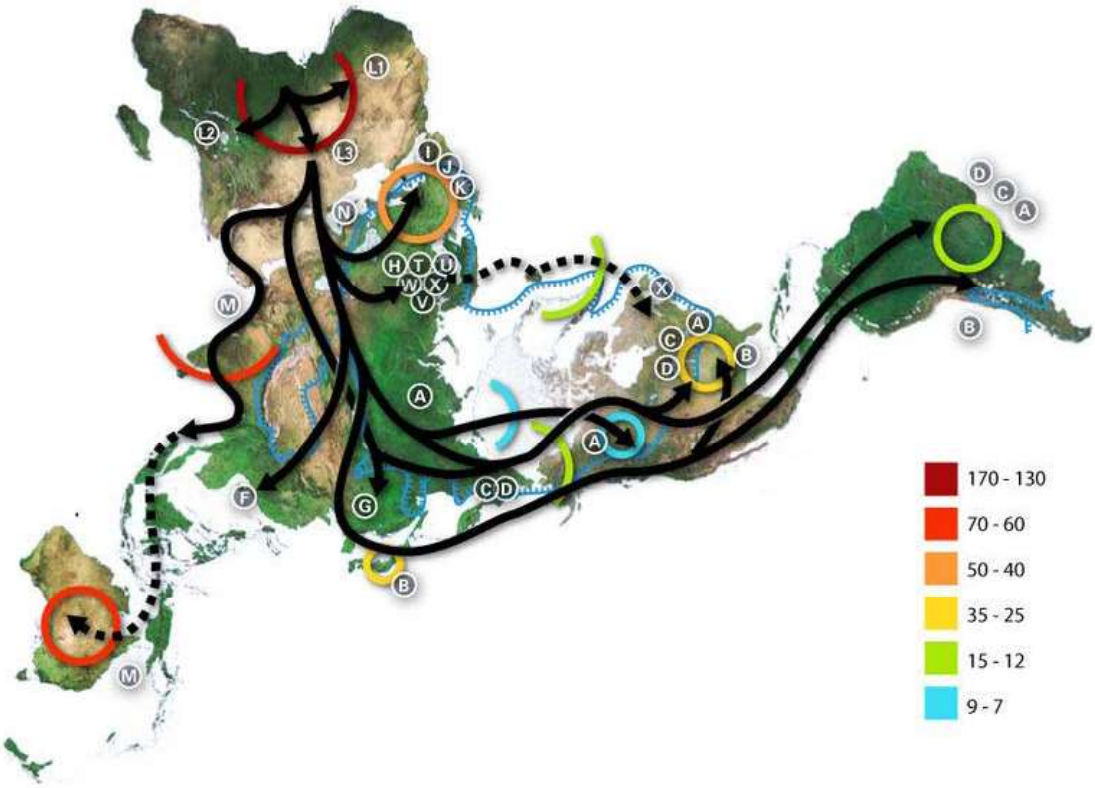
The sequence divergence has generally the following pattern: Human-Chimp < Human-Gorilla << Human-Orangutan, highlighting the close kinship between humans and the African apes. Alu elements diverge quickly due to their high frequency of CpG dinucleotides which mutate roughly 10 times more often than the average nucleotide in the genome. The mutation rate is higher in the male germ line, therefore the divergence in the Y chromosome—which is inherited solely from the father—is higher than in autosomes. The X chromosome is inherited twice as often through the female germ line as through the male germ line and therefore shows slightly lower sequence divergence. The sequence divergence of the Xq13.3 region is surprisingly low between humans and chimpanzees.

Mutations altering the amino acid sequence of proteins (K_a) are the least common. In fact ~29% of all orthologous proteins are identical between human and chimpanzee. The typical protein differs by only two amino acids.

The measures of sequence divergence shown in the table only take the substitutional differences, for example from an A (adenine) to a G (guanine), into account. DNA sequences may however also differ by insertions and deletions (indels) of bases. These are usually stripped from the alignments before the calculation of sequence divergence is

performed. The overall sequence divergence between humans and chimpanzees for example is close to 5% if indels would be included.

Modern humans



Map of the migration of modern humans out of Africa, based on mitochondrial DNA. Coloured rings indicate years before present, in thousands.

Molecular biologists starting with Wesley Brown on mtDNA and Allan Wilson on mtDNA have produced observations relevant to human evolution.

Age of the common ancestor

By estimating the rate at which mutations occur in mtDNA, the age of the common ancestral mtDNA type can be estimated: "the common ancestral mtDNA (type a) links mtDNA types that have diverged by an average of nearly 0.57%. Assuming a rate of 2%-4% per million years, this implies that the common ancestor of all surviving mtDNA types existed 140,000-290,000 years ago." This observation is robust, and this common direct female line ancestor (or mitochondrial most recent common ancestor (mtMRCA)) of all extant humans has become known as Mitochondrial Eve. The observation that the mtMRCA is the direct matrilineal ancestor of all living humans does not mean either that she was the first anatomically modern human, nor that no other female humans lived

concurrently with her. Other women would have lived at the same time and passed nuclear genes down to living humans, but their mitochondrial lineages were lost over time. This could be due to random events such as producing only male children.

African origin for modern humans

There is evidence that modern human mtDNA has an African origin: "We infer from the tree of minimum length... that Africa is a likely source of the human mitochondrial gene pool. This inference comes from the observation that one of the two primary branches leads exclusively to African mtDNAs... while the second primary branch also leads to African mtDNAs... By postulating that the common ancestral mtDNA... was African, we minimize the number of intercontinental migrations needed to account for the geographic distribution of mtDNA types."

The broad study of African genetic diversity headed by Sarah Tishkoff found the San people to express the greatest genetic diversity among the 113 distinct populations sampled, making them one of 14 "ancestral population clusters". The research also located the origin of modern human migration in south-western Africa, near the coastal border of Namibia and Angola.

Y chromosome findings

The Y chromosome is much larger than mtDNA, and is relatively homogeneous; therefore it has taken much longer to find distinct lineages and to analyse them. Conversely, because the Y chromosome is so large by comparison, it holds more genetic information. Y chromosome studies show similar findings to those made with mtDNA. The estimate for the age of the ancestral Y chromosome for all extant Y chromosomes is given at about 70,000 years ago and is also placed in Africa; the individual who contributed this Y chromosomal heritage is sometimes referred to as Y chromosome Adam. The difference in dates between Y chromosome Adam and mitochondrial Eve is usually attributed to a higher extinction rate for Y chromosomes due to greater differential reproductive success between individual men, which means that a small number of very successful men may produce many children, while a larger number of less successful men will produce far fewer children.

Chapter- 9

Human Mitochondrial Genetics

Human mitochondrial genetics is the study of the genetics of the DNA contained in human mitochondria. Mitochondria are small structures in cells that generate energy for the cell to use, and are hence referred to as the "powerhouses" of the cell.

Mitochondrial DNA (mtDNA) is not transmitted through nuclear DNA (nDNA). In humans, as in most multicellular organisms, mitochondrial DNA is inherited only from the mother's ovum.

Mitochondrial inheritance is therefore non-Mendelian, as Mendelian inheritance presumes that half the genetic material of a fertilized egg (zygote) derives from each parent.

Eighty percent of mitochondrial DNA codes for functional mitochondrial proteins, and therefore most mitochondrial DNA mutations lead to functional problems, which may be manifested as muscle disorders (myopathies).

Understanding the genetic mutations that affect mitochondria can help us to understand the inner workings of cells and organisms, as well as helping to suggest methods for successful therapeutic tissue and organ cloning, and to treatments or possibly cures for many devastating muscular disorders.

Mitochondrial function and genome

Because they provide 36 molecules of ATP per glucose molecule in contrast to the 2 ATP molecules produced by glycolysis, mitochondria are essential to all higher organisms for sustaining life. The mitochondrial diseases are genetic disorders carried specifically in mitochondrial DNA; slight problems with any one of the numerous enzymes used by the mitochondria can be devastating to the cell, and in turn, to the organism.

Membrane complexes

The processes carried out by the electron transport chain are mediated by protein complexes (named Complexes I-V, DHO-QO, ETF-QO, and ANT). Complex I, or NADH : coenzyme Q oxidoreductase, uses the energy in NADH to pump protons into the intermembrane space of the mitochondrion, pumping 2 protons per electron and passing 2 electrons via coenzyme Q to complex III or coenzyme Q : cytochrome c oxidoreductase. Complex II or succinate: coenzyme Q oxidoreductase accepts energy from succinate produced in the citric acid cycle and passes it via coenzyme Q to complex III. Complex III pumps 1 protons per electron and passes 1 electron via cytochrome c to complex IV or Cytochrome C : O_2 Oxidoreductase. Complex IV pumps 1 protons into the space between the mitochondrion's two membranes before passing the electron to O_2 which reacts to form water. Complex V (ATP synthase) is a rotary complex which allows approximately (determining the actual number is very difficult) 10 protons to enter the mitochondrial matrix along their concentration gradients. It uses the energy from the gradient to form the bond between ADP and the phosphate group to create ATP. The electron transfer flavoprotein : coenzyme Q oxidoreductase is also an electron-transporting molecule and is involved in the breakdown of fatty acids and amino acids. ANT (adenine nucleotide translocator) is also involved in oxidative phosphorylation as an energy carrying molecule. Each of these eight complexes plays a vital role in the health of the cell and any slight mutation in any one of the proteins that make up these complexes can lead to cell death or stress, which can both in turn lead to a number of diseases.

Genome

Mitochondrial DNA (mtDNA) is present in mitochondria as a circular molecule and in most species codes for 13 or 14 proteins involved in the electron transfer chain, 2 rRNA subunits and 22 tRNA molecules (all necessary for protein synthesis). The number of proteins involved in the electron transfer chain is much larger than 13 or 14, but the others are coded by the nuclear DNA.

In total, the mitochondrion hosts about 3000 proteins, but only about 13 of them are coded on the mitochondrial DNA. Most of the 3000 proteins are involved in a variety of processes other than ATP production, such as porphyrin synthesis. Only about 3% of them code for ATP production proteins. This means most of the genetic information coding for the protein makeup of mitochondria is in chromosomal DNA and is involved in processes other than ATP synthesis. This increases the chances that a mutation that will affect a mitochondrion will occur in chromosomal DNA, which is inherited in a Mendelian pattern. Another result is that a chromosomal mutation will affect a specific tissue due to its specific needs, whether those may be high energy requirements or a need for the catabolism or anabolism of a specific neurotransmitter or nucleic acid. Because several copies of the mitochondrial genome are carried by each mitochondrion (2-10 in humans), mitochondrial mutations can be inherited maternally by mtDNA mutations which are present in mitochondria inside the oocyte before fertilization, or (as stated above) through mutations in the chromosomes.

In humans, the heavy strand of mtDNA carries 28 genes and the light strand of mtDNA carries only 9 genes. Eight of the 9 genes on the light strand code for mitochondrial tRNA molecules. Human mtDNA consists of 16,569 nucleotide pairs. The entire molecule is regulated by only one regulatory region which contains the origins of replication of both heavy and light strands. The entire human mitochondrial DNA molecule has been mapped. The rate of mutation in mtDNA is calculated to be about ten times greater than that of nuclear DNA, possibly due to a paucity of DNA repair mechanisms. This high mutation rate leads to a high variation between mitochondria, not only among different species but even within the same species. It is calculated that if two humans are chosen randomly and their mtDNA is tested, they will have an average of between fifty and seventy different nucleotides. This may not seem like much, but when compared to the total number of nucleotides of a human mitochondrial DNA molecule (16,569), as much as 0.42% of the mtDNA varies between two people.

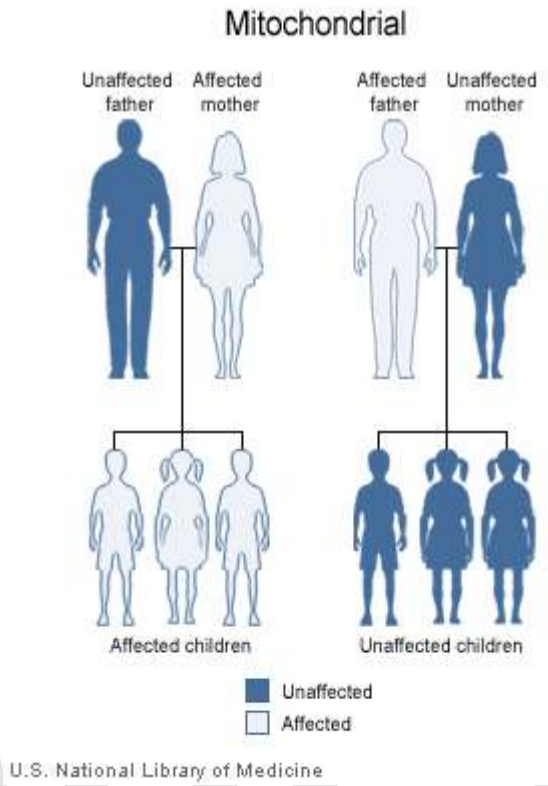
Genetic code variants

The genetic code is, for the most part, universal, with few exceptions: mitochondrial genetics includes some of these. For most organisms the "stop codons" are “UAA”, “UAG”, and “UGA”. In vertebrate mitochondria “AGA” and “AGG” are also stop codons, but not “UGA”, which codes for tryptophan instead. “AUA” codes for isoleucine in most organisms but for methionine in vertebrate mitochondrial mRNA.

There are many other variations among the codes used by other mitochondrial m/tRNA, which happened not to be harmful to their organisms, and which can be used as a tool (along with other mutations among the mtDNA/RNA of different species) to determine relative proximity of common ancestry of related species. (The more related two species are, the more mtDNA/RNA mutations will be the same in their mitochondrial genome).

Using these techniques, it is estimated that the first mitochondria arose around 1.5 billion years ago. A generally accepted hypothesis is that mitochondria originated as an aerobic prokaryote in a symbiotic relationship within an anaerobic eukaryote.

Inheritance patterns



Mitochondrial Inheritance Patterns

Because mitochondrial diseases (diseases due to malfunction of mitochondria) can be inherited both maternally and through chromosomal inheritance, the way in which they are passed on from generation to generation can vary greatly depending on the disease. Mitochondrial genetic mutations that occur in the nuclear DNA can occur in any of the chromosomes (depending on the species). Mutations inherited through the chromosomes can be autosomal dominant or recessive and can also be sex-linked dominant or recessive. Chromosomal inheritance follows normal Mendelian laws, despite the fact that the phenotype of the disease may be masked.

Because of the complex ways in which mitochondrial and nuclear DNA "communicate" and interact, even seemingly simple inheritance is hard to diagnose. A mutation in chromosomal DNA may change a protein that regulates (increases or decreases) the production of another certain protein in the mitochondria or the cytoplasm; this may lead to slight, if any, noticeable symptoms. On the other hand, some devastating mtDNA mutations are easy to diagnose because of their widespread damage to muscular, neural, and/or hepatic tissues (among other high-energy and metabolism-dependent tissues) and because they are present in the mother and all the offspring.

Mitochondrial genome mutations are passed on 100% of the time from mother to all her offspring. The number of affected mtDNA molecules inherited by a specific offspring can vary greatly because

- the mitochondria within the fertilized oocyte is what the new life will have to begin with (in terms of mtDNA),
- the number of affected mitochondria varies from cell (in this case, the fertilized oocyte) to cell depending both on the number it inherited from its mother cell and environmental factors which may favor mutant or wildtype mitochondrial DNA,
- the number of mtDNA molecules in the mitochondria varies from around two to ten.

It is possible, even in twin births, for one baby to receive more than half mutant mtDNA molecules while the other twin may receive only a tiny fraction of mutant mtDNA molecules with respect to wildtype (depending on how the twins divide from each other and how many mutant mitochondria happen to be on each side of the division). In a few cases, some mitochondria or a mitochondrion from the sperm cell enters the oocyte but paternal mitochondria are actively decomposed.

Replication, repair, transcription, and translation

Mitochondrial replication is controlled by nuclear genes and is specifically suited to make as many mitochondria as that particular cell needs at the time. Human mitochondrial DNA (mtDNA) has three promoters, H1, H2, and L (heavy strand 1, heavy strand 2, and light strand promoters). The H2 promoter transcribes almost the entire heavy strand and the L promoter transcribes the entire light strand. The H1 promoter causes the transcription of the two mitochondrial rRNA molecules. When transcription takes place on the heavy strand a polycistronic transcript is created. The light strand produces either small transcripts, which can be used as primers, or one long transcript. The production of primers occurs by processing of light strand transcripts with the Mitochondrial RNase MRP (Mitochondrial RNA Processing). The requirement of transcription to produce primers links the process of transcription to mtDNA replication. Full length transcripts are cut into functional tRNA, rRNA, and mRNA molecules. The process of transcription initiation in mitochondria involves three types of proteins: the mitochondrial RNA polymerase (POLRMT), mitochondrial transcription factor A (TFAM), and mitochondrial transcription factors B1 and B2 (TFB1M, TFB2M). POLRMT, TFAM, and TFB1M or TFB2M assemble at the mitochondrial promoters and begin transcription. The actual molecular events that are involved in initiation are unknown, but these factors make up the basal transcription machinery and have been shown to function in vitro. Mitochondrial translation is still not very well understood. In vitro translations have still not been successful, probably due to the difficulty of isolating sufficient mt mRNA, functional mt rRNA, and possibly because of the complicated changes that the mRNA undergoes before it is translated.

Mitochondrial DNA polymerase

The Mitochondrial DNA Polymerase (Pol gamma, encoded by the POLG gene) is used in the copying of mtDNA during replication. Because the two (heavy and light) strands on the circular mtDNA molecule have different origins of replication, it replicates in a D-loop mode. One strand begins to replicate first, displacing the other strand. This continues

until replication reaches the origin of replication on the other strand, at which point the other strand begins replicating in the opposite direction. This results in two new mtDNA molecules. Each mitochondrion has several copies of the mtDNA molecule and the number of mtDNA molecules is a limiting factor in mitochondrial fission. After the mitochondrion has enough mtDNA, membrane area, and membrane proteins, it can undergo fission (very similar to that which bacteria use) to become two mitochondria. Evidence suggests that mitochondria can also undergo fusion and exchange (in a form of crossover) genetic material among each other. Mitochondria sometimes form large matrices in which fusion, fission, and protein exchanges are constantly occurring. mtDNA shared among mitochondria (despite the fact that they can undergo fusion).

Damage and transcription error

Mitochondrial DNA is susceptible to damage from free oxygen radicals from mistakes that occur during the production of ATP through the electron transport chain. These mistakes can be caused by genetic disorders, cancer, and temperature variations. These radicals can damage mtDNA molecules or change them, making it hard for mitochondrial polymerase to replicate them. Both cases can lead to deletions, rearrangements, and other mutations. Recent evidence has suggested that mitochondria have enzymes that proofread mtDNA and fix mutations that may occur due to free radicals. It is believed that a DNA recombinase found in mammalian cells is also involved in a repairing recombination process. Deletions and mutations due to free radicals have been associated with the aging process. It is believed that radicals cause mutations which lead to mutant proteins, which in turn lead to more radicals. This process takes many years and is associated with some aging processes involved in oxygen-dependent tissues such as brain, heart, muscle, and kidney. Auto-enhancing processes such as these are possible causes of degenerative diseases including Parkinson's, Alzheimer's, and coronary artery disease.

Chromosomally mediated mtDNA replication errors

Because mitochondrial growth and fission are mediated by the nuclear DNA, mutations in nuclear DNA can have a wide array of effects on mtDNA replication. Despite the fact that the loci for some of these mutations have been found on human chromosomes, specific genes and proteins involved have not yet been isolated. Mitochondria need a certain protein to undergo fission. If this protein (made by the nucleus) is not present, the mitochondria grow but they do not divide. This leads to giant, inefficient mitochondria. Mistakes in chromosomal genes or their products can also affect mitochondrial replication more directly by inhibiting mitochondrial polymerase and can even cause mutations in the mtDNA directly and indirectly. Indirect mutations are most often caused by radicals created by defective proteins made from nuclear DNA.

Mitochondrial diseases

Mitochondrial diseases range in severity from asymptomatic to fatal, and are most commonly due to inherited rather than acquired mutations of mitochondrial DNA. A given mitochondrial mutation can cause various diseases depending on the severity of the

problem in the mitochondria and the tissue the affected mitochondria are in. Conversely, several different mutations may present themselves as the same disease. This almost patient-specific characterization of mitochondrial diseases makes them very hard to accurately recognize, diagnose and trace. Some diseases are observable at or even before birth (many causing death) while others do not show themselves until late adulthood (late-onset disorders). This is because the number of mutant versus wildtype mitochondria varies between cells and tissues, and is continuously changing. Because cells have multiple mitochondria, different mitochondria in the same cell can have different variations of the mtDNA. This condition is referred to as heteroplasmy. When a certain tissue reaches a certain ratio of mutant versus wildtype mitochondria, a disease will present itself. The ratio varies from person to person and tissue to tissue (depending on its specific energy, oxygen, and metabolism requirements, and the effects of the specific mutation). Mitochondrial diseases are very numerous and different. Apart from diseases caused by abnormalities in mitochondrial DNA, many diseases are suspected to be associated in part by mitochondrial dysfunctions, such as diabetes mellitus, forms of cancer and cardiovascular disease, lactic acidosis, specific forms of myopathy, osteoporosis, Alzheimer's disease, Parkinsons's disease, stroke, Male infertility and which are also believed to play a role in the aging process.

Chapter- 10

Molecular Genetics

Molecular genetics is the field of biology and genetics that studies the structure and function of genes at a molecular level. The field studies how the genes are transferred from generation to generation. Molecular genetics employs the methods of genetics and molecular biology. It is so-called to differentiate it from other sub fields of genetics such as ecological genetics and population genetics. An important area within molecular genetics is the use of molecular information to determine the patterns of descent, and therefore the correct scientific classification of organisms: this is called molecular systematics.

Along with determining the pattern of descendants, molecular genetics helps in understanding genetic mutations that can cause certain types of diseases. Through utilizing the methods of genetics and molecular biology, molecular genetics discovers the reasons why traits are carried on and how and why some may mutate.

Forward genetics

One of the first tools available to molecular geneticists is the forward genetic screen. The aim of this technique is to identify mutations that produce a certain phenotype. A mutagen is very often used to accelerate this process. Once mutants have been isolated, the mutated gene can be molecularly identified.

Reverse genetics

While forward genetic screens are productive, a more straightforward approach would be to determine the phenotype that results from mutating a given gene. This is called reverse genetics. In some organisms, such as yeast and mice, it is possible to induce the deletion of a particular gene, creating a gene knockout. Alternatives include the random induction of DNA deletions and subsequent selection for deletions in a gene of interest, the application of RNA interference and the creation of transgenic organisms that do not express a gene of interest.

Gene therapy

A mutation in a gene can result in a severe medical condition. A protein encoded by a mutated gene may malfunction and cells that rely on the protein might therefore fail to function properly. This can cause problems for specific tissues or organs, or for the entire body. This might manifest through the course of development (like a cleft palate) or as an abnormal response to stimuli (like a peanut allergy). Conditions related to gene mutations are called genetic disorders. One way to fix such a physiological problem is gene therapy. By adding a corrected copy of the gene, a functional form of the protein can be produced, and affected cells, tissues, and organs may work properly. As opposed to drug-based approaches, gene therapy repairs the underlying genetic defect.

Gene therapy is the process of treating or alleviating diseases by genetically modifying the cells of the affected person, causing the gene to function properly. When a human disease gene has been recognized, molecular genetics tools can be used to explore the process of the gene in both the normal and mutant states. From there, the gene is transferred either in vivo or ex vivo and the body begins to make proteins according to the instructions in the new gene. Gene therapy has to be repeated several times for the infected patient to continually be relieved, however, as repeated cell division and death slowly randomizes the body's ratio of functional-to-mutant genes.

Currently, gene therapy is still being experimented with and products are not approved by the U.S. Food and Drug Administration. There have been several setbacks in the last 15 years that have restricted further developments in gene therapy. As there are unsuccessful attempts, there continue to be a growing number of successful gene therapy transfers which have furthered the research.

Major diseases that can be treated with gene therapy include viral infections, cancers, and inherited disorders, including immune system disorders.

Classical gene therapy

Classical gene therapy is the approach which delivers genes, via a modified virus or "vector" to the appropriate target cells with a goal of attaining optimal expression of the new, introduced gene. Once inside the patient, the expressed genes are intended to produce a product that the patient lacks, kill diseased cells directly by producing a toxin, or activate the immune system to help the killing of diseased cells.

Nonclassical gene therapy

Nonclassical gene therapy inhibits the expression of genes related to pathogenesis, or corrects a genetic defect and restores normal gene expression.

In vivo gene transfer

During In vivo gene transfer, the genes are transferred directly into the tissue of the patient and this can be the only possible option in patients with tissues where individual cells cannot be cultured in vitro in sufficient numbers (e.g. brain cells). Also, in vivo gene transfer is necessary when cultured cells cannot be re-implanted in patients effectively.

Ex vivo gene transfer

During ex vivo gene transfer the cells are cultured outside the body and then the genes are transferred into the cells grown in culture. The cells that have been transformed successfully are expanded by cell culture and then introduced into the patient.

Principles for gene transfer

Classical gene therapies usually require efficient transfer of cloned genes into the disease cells so that the introduced genes are expressed at sufficiently high levels to change the patient's physiology. There are several different physicochemical and biological methods that can be used to transfer genes into human cells. The size of the DNA fragments that can be transferred is very limited, and often the transferred gene is not a conventional gene. Horizontal gene transfer is the transfer of genetic material from one cell to another that is not its offspring. Artificial horizontal gene transfer is a form of genetic engineering.

Techniques in molecular genetics

There are three general techniques used for molecular genetics: amplification, separation and detection, and expression. Specifically used for amplification is polymerase chain reaction, which is an “indispensable tool in a great variety of applications”. In the separation and detection technique DNA and mRNA are isolated from their cells. Gene expression in cells or organisms is done in a place or time that is not normal for that specific gene.

Amplification

There are other methods for amplification besides polymerase chain reaction. Cloning DNA in bacteria is also a way to amplify DNA in genes.

Polymerase chain reaction

The main materials used in polymerase chain reaction are DNA nucleotides, template DNA, primers and Taq polymerase. DNA nucleotides are the base for the new DNA, the template DNA is the specific sequence being amplified, primers are complementary nucleotides that can go on either side of the template DNA, and Taq polymerase is a heat stable enzyme that jump-starts the production of new DNA at the high temperatures

needed for reaction. This technique does not need to use living bacteria or cells; all that is needed is the base sequence of the DNA and materials listed above.

Cloning DNA in bacteria

The word cloning for this type of amplification entails making multiple identical copies of a sequence of DNA. The target DNA sequence is then inserted into a cloning vector. Because this vector originates from a self-replicating virus, plasmid, or higher organism cell when the appropriate size DNA is inserted the "target and vector DNA fragments are then ligated" and create a recombinant DNA molecule. The recombinant DNA molecules are then put into a bacteria strain (usually *E. coli*) which produces several identical copies by transformation. Transformation is the DNA uptake mechanism possessed by bacteria. However, only one recombinant DNA molecule can be cloned within a single bacteria cell, so each clone is of just one DNA insert.

Separation and detection

In separation and detection DNA and mRNA are isolated from cells (the separation) and then detected simply by the isolation. Cell cultures are also grown to provide a constant supply of cells ready for isolation.

Cell cultures

A cell culture for molecular genetics is a culture that is grown in artificial conditions. Some cell types grow well in cultures such a skin cells, but other cells are not as productive in cultures. There are different techniques for each type of cell, some only recently being found to foster growth in stem and nerve cells. Cultures for molecular genetics are frozen in order to preserve all copies of the gene specimen and thawed only when needed. This allows for a steady supply of cells.

DNA isolation

DNA isolation extracts DNA from a cell in a pure form. First, the DNA is separated from cellular components such as proteins, RNA, and lipids. This is done by placing the chosen cells in a tube with a solution that mechanically, chemically, breaks the cells open. This solution contains enzymes, chemicals, and salts that breaks down the cells except for the DNA. It contains enzymes to dissolve proteins, chemicals to destroy all RNA present, and salts to help pull DNA out of the solution.

Next, the DNA is separated from the solution by being spun in a centrifuge, which allows the DNA to collect in the bottom of the tube. After this cycle in the centrifuge the solution is poured off and the DNA is resuspended in a second solution that makes the DNA easy to work with in the future.

This results in a concentrated DNA sample that contains thousands of copies of each gene. For large scale projects such as sequencing the human genome, all this work is done by robots.

mRNA isolation

Expressed DNA that codes for the synthesis of a protein is the final goal for scientists and this expressed DNA is obtained by isolation mRNA (Messenger RNA). First, laboratories use a normal cellular modification of mRNA that adds up to 200 adenine nucleotides to the end of the molecule (poly(A) tail). Once this has been added, the cell is ruptured and its cell contents are exposed to synthetic beads that are coated with thymine string nucleotides. Because Adenine and Thymine pair together in DNA, the poly(A) tail and synthetic beads are attracted to one another, and once they bind in this process the cell components can be washed away without removing the mRNA. Once the mRNA has been isolated, reverse transcriptase is employed to convert it to single-stranded DNA, from which a stable double-stranded DNA is produced using DNA polymerase. Complementary DNA (cDNA) is much more stable than mRNA and so, once the double-stranded DNA has been produced it represents the expressed DNA sequence scientists look for.

The Human Genome Project

The Human Genome Project is a molecular genetics project that began in the 1990s and was projected to take fifteen years to complete. However, because of technological advances the progress of the project was advanced and the project finished in 2003, taking only thirteen years. The project was started by the U.S. Department of Energy and the National Institutes of Health in an effort to reach six set goals. These goals included:

1. identifying 20,000 to 25,000 genes in human DNA (although initial estimate were approximately 100,000 genes),
2. determining sequences of chemical based pairs in human DNA,
3. storing all found information into databases,
4. improving the tools used for data analysis,
5. transferring technologies to private sectors, and
6. addressing the ethical, legal, and social issues (ELSI) that may arise from the projects.

The project was worked on by eighteen different countries including the United States, Japan, France, Germany, and the United Kingdom. The collaborative effort resulted in the discovery of the many benefits of molecular genetics. Discoveries such as molecular medicine, new energy sources and environmental applications, DNA forensics, and livestock breeding, are only a few of the benefits that molecular genetics can provide.

Chapter- 11

Chromosome

A **chromosome** is an organized structure of DNA and protein that is found in cells. It is a single piece of coiled DNA containing many genes, regulatory elements and other nucleotide sequences. Chromosomes also contain DNA-bound proteins, which serve to package the DNA and control its functions. The word *chromosome* comes from the Greek χρῶμα (*chroma*, colour) and σῶμα (*soma*, body) due to their property of being very strongly stained by particular dyes.

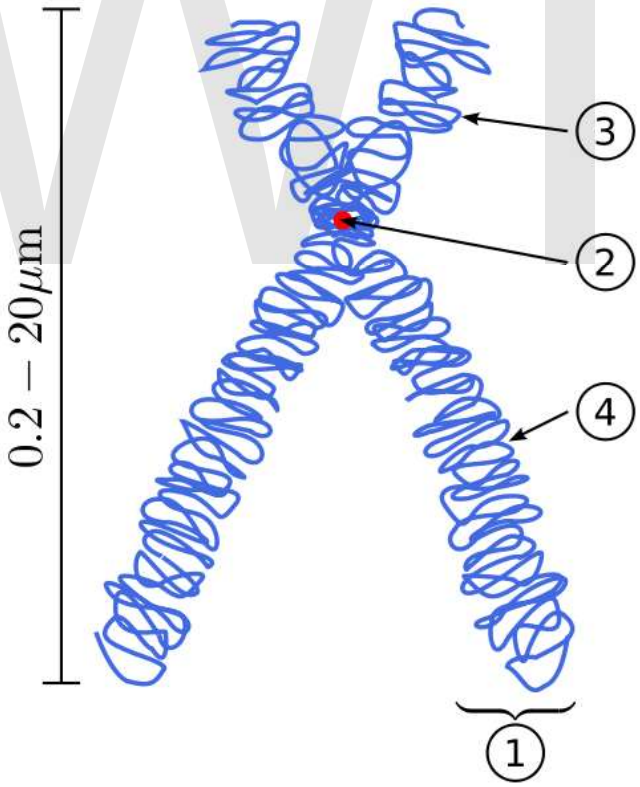


Diagram of a replicated and condensed metaphase eukaryotic chromosome. (1) Chromatid – one of the two identical parts of the chromosome after S phase. (2) Centromere – the point where the two chromatids touch, and where the microtubules attach. (3) Short arm. (4) Long arm.

Chromosomes vary widely between different organisms. The DNA molecule may be circular or linear, and can be composed of 10,000 to 1,000,000,000 nucleotides in a long chain. Typically eukaryotic cells (cells with nuclei) have large linear chromosomes and prokaryotic cells (cells without defined nuclei) have smaller circular chromosomes, although there are many exceptions to this rule. Furthermore, cells may contain more than one type of chromosome; for example, mitochondria in most eukaryotes and chloroplasts in plants have their own small chromosomes.

In eukaryotes, nuclear chromosomes are packaged by proteins into a condensed structure called chromatin. This allows the very long DNA molecules to fit into the cell nucleus. The structure of chromosomes and chromatin varies through the cell cycle. Chromosomes are the essential unit for cellular division and must be replicated, divided, and passed successfully to their daughter cells so as to ensure the genetic diversity and survival of their progeny. Chromosomes may exist as either duplicated or unduplicated—unduplicated chromosomes are single linear strands, whereas duplicated chromosomes (copied during synthesis phase) contain two copies joined by a centromere.

Compaction of the duplicated chromosomes during mitosis and meiosis results in the classic four-arm structure (pictured to the right). Chromosomal recombination plays a vital role in genetic diversity. If these structures are manipulated incorrectly, through processes known as chromosomal instability and translocation, the cell may undergo mitotic catastrophe and die, or it may unexpectedly evade apoptosis leading to the progression of cancer.

In practice "chromosome" is a rather loosely defined term. In prokaryotes and viruses, the term genophore is more appropriate when no chromatin is present. However, a large body of work uses the term chromosome regardless of chromatin content. In prokaryotes DNA is usually arranged as a circle, which is tightly coiled in on itself, sometimes accompanied by one or more smaller, circular DNA molecules called plasmids. These small circular genomes are also found in mitochondria and chloroplasts, reflecting their bacterial origins. The simplest genophores are found in viruses: these DNA or RNA molecules are short linear or circular genophores that often lack structural proteins.

History

Chromosomes as vectors of heredity

In a series of experiments, Theodor Boveri gave the definitive demonstration that chromosomes are the vectors of heredity. His two principles were based upon the *continuity* of chromosomes and the *individuality* of chromosomes. It is the second of these principles that was so original. Boveri was able to test the proposal put forward by Wilhelm Roux, that each chromosome carries a different genetic load, and showed that Roux was right. Upon the rediscovery of Mendel, Boveri was able to point out the connection between the rules of inheritance and the behaviour of the chromosomes. It is interesting to see that Boveri influenced two generations of American cytologists:

Edmund Beecher Wilson, Walter Sutton and Theophilus Painter were all influenced by Boveri (Wilson and Painter actually worked with him).

In his famous textbook *The Cell*, Wilson linked Boveri and Sutton together by the Boveri-Sutton theory. Mayr remarks that the theory was hotly contested by some famous geneticists: William Bateson, Wilhelm Johannsen, Richard Goldschmidt and T.H. Morgan, all of a rather dogmatic turn-of-mind. Eventually complete proof came from chromosome maps in Morgan's own lab.

Chromosomes in eukaryotes

Eukaryotes (cells with nuclei such as those found in plants, yeast, and animals) possess multiple large linear chromosomes contained in the cell's nucleus. Each chromosome has one centromere, with one or two arms projecting from the centromere, although, under most circumstances, these arms are not visible as such. In addition, most eukaryotes have a small circular mitochondrial genome, and some eukaryotes may have additional small circular or linear cytoplasmic chromosomes.

In the nuclear chromosomes of eukaryotes, the uncondensed DNA exists in a semi-ordered structure, where it is wrapped around histones (structural proteins), forming a composite material called chromatin.

Chromatin

Chromatin is the complex of DNA and protein found in the eukaryotic nucleus, which packages chromosomes. The structure of chromatin varies significantly between different stages of the cell cycle, according to the requirements of the DNA.

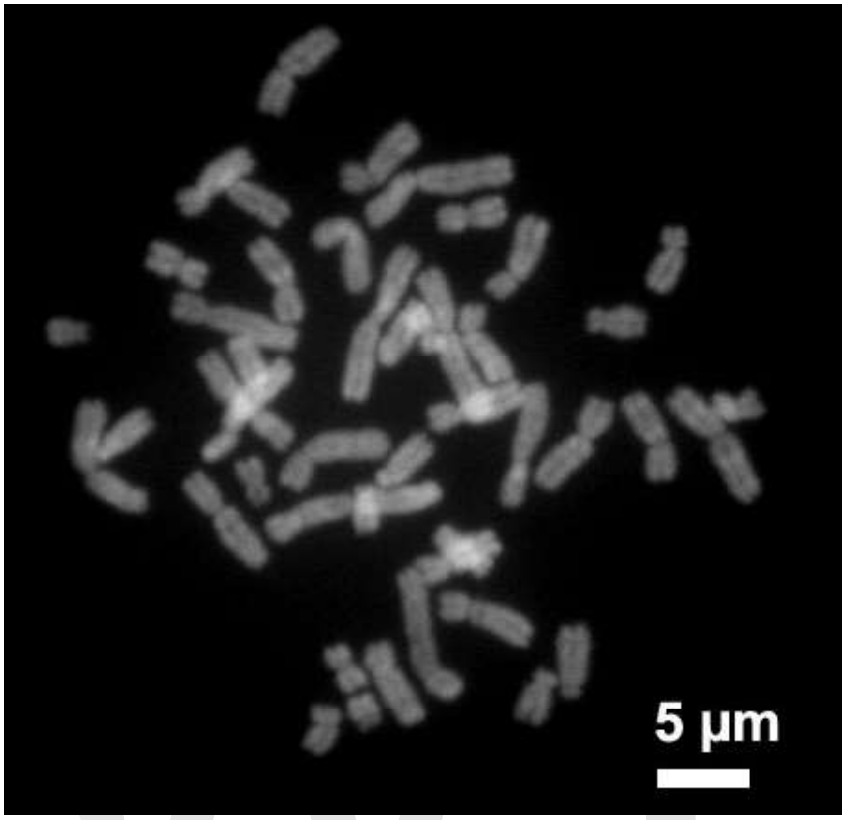
Interphase chromatin

During interphase (the period of the cell cycle where the cell is not dividing), two types of chromatin can be distinguished:

- Euchromatin, which consists of DNA that is active, e.g., being expressed as protein.
- Heterochromatin, which consists of mostly inactive DNA. It seems to serve structural purposes during the chromosomal stages. Heterochromatin can be further distinguished into two types:
    - *Constitutive heterochromatin*, which is never expressed. It is located around the centromere and usually contains repetitive sequences.
    - *Facultative heterochromatin*, which is sometimes expressed.

Individual chromosomes cannot be distinguished at this stage – they appear in the nucleus as a homogeneous tangled mix of DNA and protein.

Metaphase chromatin and division



Human chromosomes during metaphase

In the early stages of mitosis or meiosis (cell division), the chromatin strands become more and more condensed. They cease to function as accessible genetic material (transcription stops) and become a compact transportable form. This compact form makes the individual chromosomes visible, and they form the classic four arm structure, a pair of sister chromatids attached to each other at the centromere. The shorter arms are called *p arms* (from the French *petit*, small) and the longer arms are called *q arms* (*q* follows *p* in the Latin alphabet). This is the only natural context in which individual chromosomes are visible with an optical microscope.

During divisions, long microtubules attach to the centromere and the two opposite ends of the cell. The microtubules then pull the chromatids apart, so that each daughter cell inherits one set of chromatids. Once the cells have divided, the chromatids are uncoiled and can function again as chromatin. In spite of their appearance, chromosomes are structurally highly condensed, which enables these giant DNA structures to be contained within a cell nucleus (Fig. 2).

The self-assembled microtubules form the spindle, which attaches to chromosomes at specialized structures called kinetochores, one of which is present on each sister chromatid. A special DNA base sequence in the region of the kinetochores provides, along with special proteins, longer-lasting attachment in this region.

Chromosomes in prokaryotes

The prokaryotes – bacteria and archaea – typically have a single circular chromosome, but many variations do exist. Most bacteria have a single circular chromosome that can range in size from only 160,000 base pairs in the endosymbiotic bacterium *Candidatus Carsonella ruddii*, to 12,200,000 base pairs in the soil-dwelling bacterium *Sorangium cellulosum*. Spirochaetes of the genus *Borrelia* are a notable exception to this arrangement, with bacteria such as *Borrelia burgdorferi*, the cause of Lyme disease, containing a single linear chromosome.

Structure in sequences

Prokaryotic chromosomes have less sequence-based structure than eukaryotes. Bacteria typically have a single point (the origin of replication) from which replication starts, whereas some archaea contain multiple replication origins. The genes in prokaryotes are often organized in operons, and do not usually contain introns, unlike eukaryotes.

DNA packaging

Prokaryotes do not possess nuclei. Instead, their DNA is organized into a structure called the nucleoid. The nucleoid is a distinct structure and occupies a defined region of the bacterial cell. This structure is, however, dynamic and is maintained and remodeled by the actions of a range of histone-like proteins, which associate with the bacterial chromosome. In archaea, the DNA in chromosomes is even more organized, with the DNA packaged within structures similar to eukaryotic nucleosomes.

Bacterial chromosomes tend to be tethered to the plasma membrane of the bacteria. In molecular biology application, this allows for its isolation from plasmid DNA by centrifugation of lysed bacteria and pelleting of the membranes (and the attached DNA).

Prokaryotic chromosomes and plasmids are, like eukaryotic DNA, generally supercoiled. The DNA must first be released into its relaxed state for access for transcription, regulation, and replication.

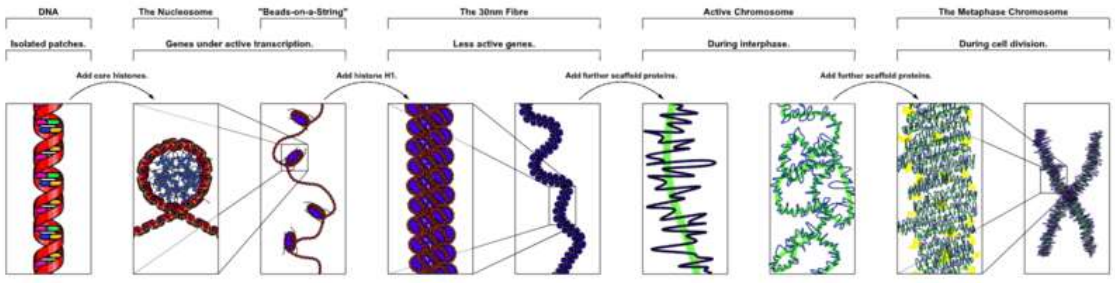


Fig. 2: The major structures in DNA compaction; DNA, the nucleosome, the 10nm "beads-on-a-string" fibre, the 30nm fibre and the metaphase chromosome.

Number of chromosomes in various organisms

Eukaryotes

These tables give the total number of chromosomes (including sex chromosomes) in a cell nucleus. For example, human cells are diploid and have 22 different types of autosome, each present as two copies, and two sex chromosomes. This gives 46 chromosomes in total. Other organisms have more than two copies of their chromosomes, such as bread wheat, which is *hexaploid* and has six copies of seven different chromosomes – 42 chromosomes in total.

Chromosome numbers in some plants

Plant Species	**#**
Arabidopsis thaliana (diploid)	10
Rye (diploid)	14
Maize (diploid or palaeotetraploid)	20
Einkorn wheat (diploid)	14
Durum wheat (tetraploid)	28
Bread wheat (hexaploid)	42
Cultivated tobacco (tetraploid)	48
Adder's Tongue Fern (diploid)	}

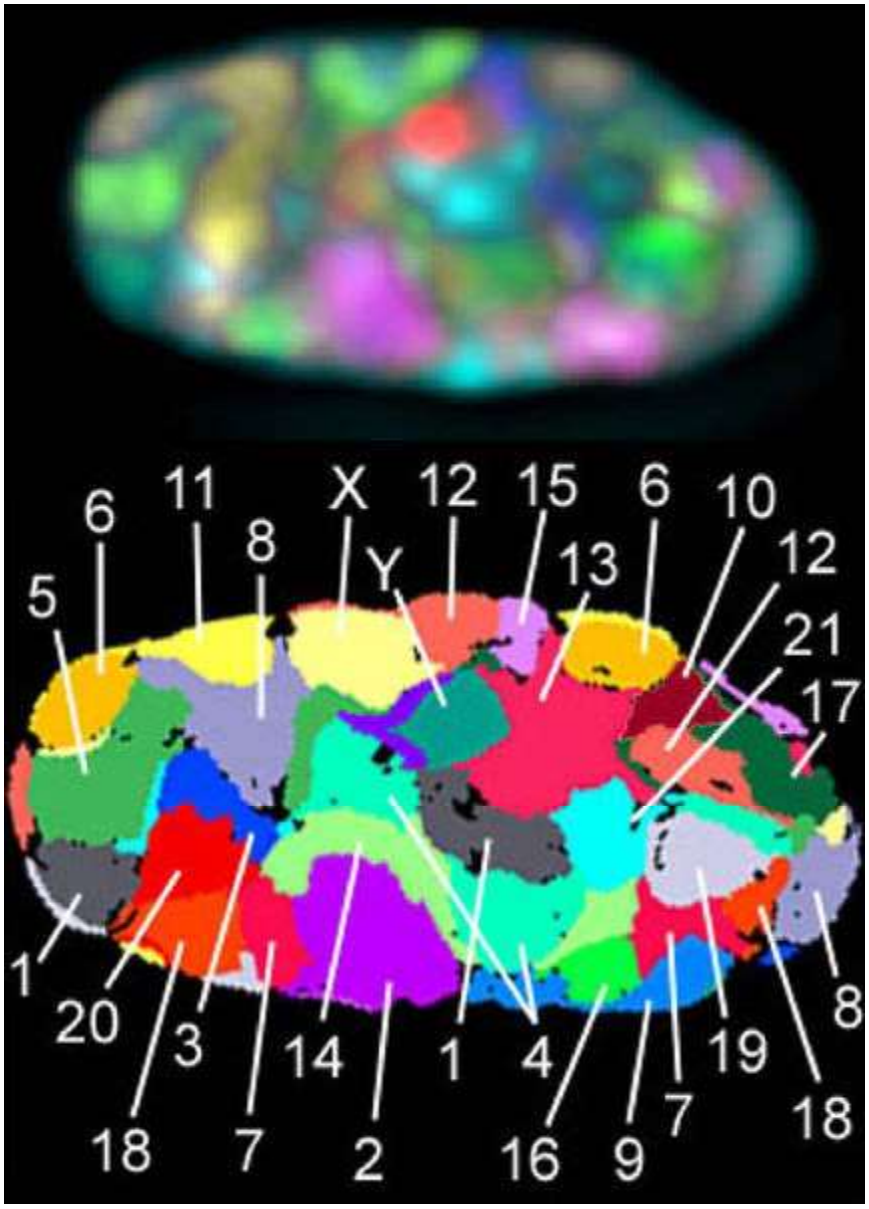
Chromosome numbers (2n) in some animals

Species	#	Species	#
Common fruit fly	8	Guinea Pig	64
Guppy (*poecilia reticulata*)	46	Garden snail	54
Earthworm (*Octodrilus complanatus*)	36	Tibetan fox	36
Domestic cat	38	Domestic pig	38
Laboratory mouse	40	Laboratory rat	42
Rabbit (*Oryctolagus cuniculus*)	44	Syrian hamster	44
Hares	48	Human	46
Gorillas, Chimpanzees	48	Domestic sheep	54
Elephants	56	Cow	60
Donkey	62	Horse	64
Dog	78	Kingfisher	132
Goldfish	100-104	Silkworm	56

Chromosome numbers in other organisms

Species	Large Chromosomes	Intermediate Chromosomes	Microchromosomes
Trypanosoma brucei	11	6	~100
Domestic Pigeon (*Columba livia domestics*)	18	-	59-63
Chicken	8	2 sex chromosomes	60

Normal members of a particular eukaryotic species all have the same number of nuclear chromosomes (see the table). Other eukaryotic chromosomes, i.e., mitochondrial and plasmid-like small chromosomes, are much more variable in number, and there may be thousands of copies per cell.



The 23 human chromosome territories during prometaphase in fibroblast cells

Asexually reproducing species have one set of chromosomes, which are the same in all body cells. However, asexual species can be either haploid or diploid.

Sexually reproducing species have somatic cells (body cells), which are diploid [2n] having two sets of chromosomes, one from the mother and one from the father. Gametes, reproductive cells, are haploid [n]: They have one set of chromosomes. Gametes are produced by meiosis of a diploid germ line cell. During meiosis, the matching chromosomes of father and mother can exchange small parts of themselves (crossover), and thus create new chromosomes that are not inherited solely from either parent. When a male and a female gamete merge (fertilization), a new diploid organism is formed.

Some animal and plant species are polyploid [Xn]: They have more than two sets of homologous chromosomes. Plants important in agriculture such as tobacco or wheat are often polyploid, compared to their ancestral species. Wheat has a haploid number of seven chromosomes, still seen in some cultivars as well as the wild progenitors. The more-common pasta and bread wheats are polyploid, having 28 (tetraploid) and 42 (hexaploid) chromosomes, compared to the 14 (diploid) chromosomes in the wild wheat.

Prokaryotes

Prokaryote species generally have one copy of each major chromosome, but most cells can easily survive with multiple copies. For example, *Buchnera*, a symbiont of aphids has multiple copies of its chromosome, ranging from 10–400 copies per cell. However, in some large bacteria, such as *Epulopiscium fishelsoni* up to 100,000 copies of the chromosome can be present. Plasmids and plasmid-like small chromosomes are, as in eukaryotes, very variable in copy number. The number of plasmids in the cell is almost entirely determined by the rate of division of the plasmid – fast division causes high copy number, and vice versa.

Karyotype

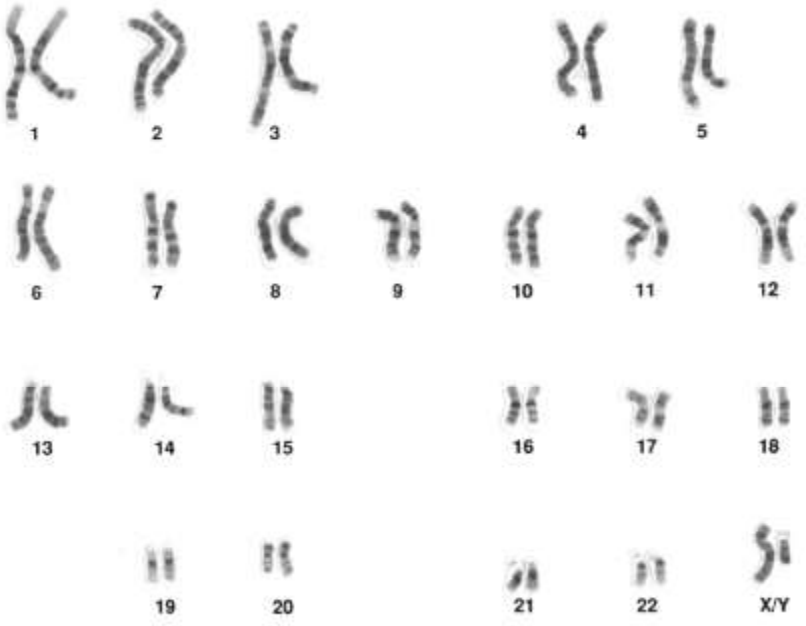


Figure 3: Karyogram of a human male

In general, the **karyotype** is the characteristic chromosome complement of a eukaryote species. The preparation and study of karyotypes is part of cytogenetics.

Although the replication and transcription of DNA is highly standardized in eukaryotes, *the same cannot be said for their karyotypes*, which are often highly variable. There may

be variation between species in chromosome number and in detailed organization. In some cases, there is significant variation within species. Often there is:

1. variation between the two sexes
2. variation between the germ-line and soma (between gametes and the rest of the body)
3. variation between members of a population, due to balanced genetic polymorphism
4. geographical variation between races
5. mosaics or otherwise abnormal individuals.

Also, variation in karyotype may occur during development from the fertilised egg.

The technique of determining the karyotype is usually called *karyotyping*. Cells can be locked part-way through division (in metaphase) in vitro (in a reaction vial) with colchicine. These cells are then stained, photographed, and arranged into a *karyogram*, with the set of chromosomes arranged, autosomes in order of length, and sex chromosomes (here X/Y) at the end: Fig. 3.

Like many sexually reproducing species, humans have special gonosomes (sex chromosomes, in contrast to autosomes). These are XX in females and XY in males.

Historical note

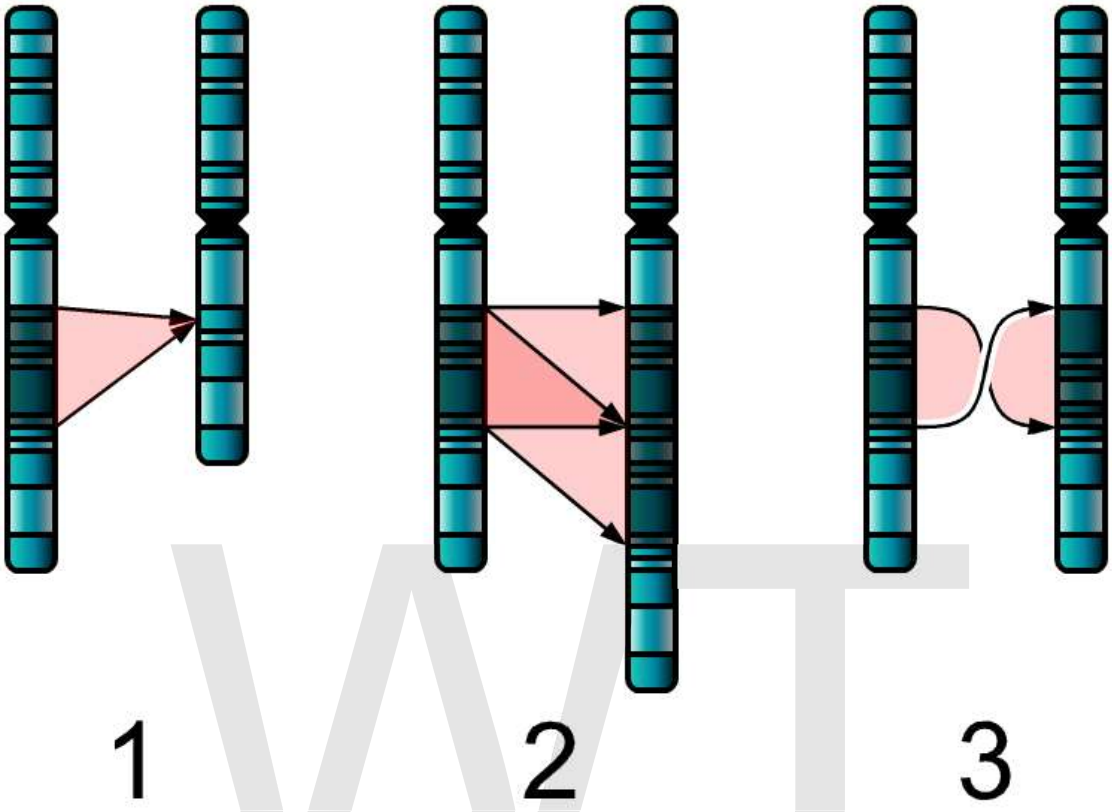
Investigation into the human karyotype took many years to settle the most basic question. How many chromosomes does a normal diploid human cell contain? In 1912, Hans von Winiwarter reported 47 chromosomes in spermatogonia and 48 in oogonia, concluding an XX/XO sex determination mechanism. Painter in 1922 was not certain whether the diploid number of man is 46 or 48, at first favouring 46. He revised his opinion later from 46 to 48, and he correctly insisted on humans having an XX/XY system.

New techniques were needed to definitively solve the problem:

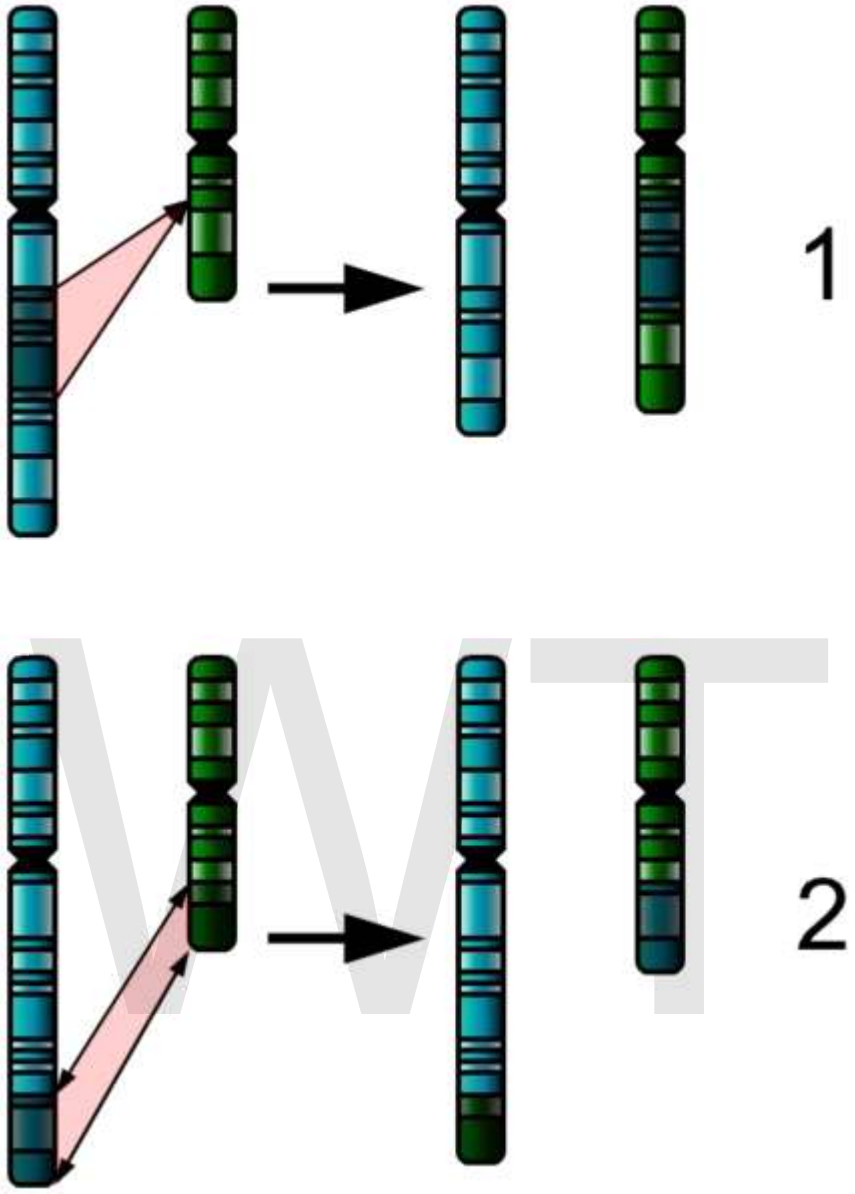
1. Using cells in culture
2. Pretreating cells in a hypotonic solution, which swells them and spreads the chromosomes
3. Arresting mitosis in metaphase by a solution of colchicine
4. Squashing the preparation on the slide forcing the chromosomes into a single plane
5. Cutting up a photomicrograph and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

It took until the mid-1950s for it to become generally accepted that the human karyotype include only 46 chromosomes. Considering the techniques of Winiwarter and Painter, their results were quite remarkable. Chimpanzees (the closest living relatives to modern humans) have 48 chromosomes.

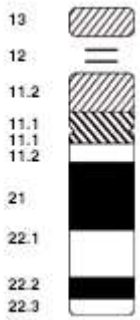
Chromosomal aberrations



The three major single chromosome mutations; deletion (1), duplication (2) and inversion (3)



The two major two-chromosome mutations; insertion (1) and translocation (2)



In Down syndrome, there are three copies of chromosome 21

Chromosomal aberrations are disruptions in the normal chromosomal content of a cell and are a major cause of genetic conditions in humans, such as Down syndrome. Some chromosome abnormalities do not cause disease in carriers, such as translocations, or chromosomal inversions, although they may lead to a higher chance of birthing a child with a chromosome disorder. Abnormal numbers of chromosomes or chromosome sets, aneuploidy, may be lethal or give rise to genetic disorders. Genetic counseling is offered for families that may carry a chromosome rearrangement.

The gain or loss of DNA from chromosomes can lead to a variety of genetic disorders. Human examples include:

- Cri du chat, which is caused by the deletion of part of the short arm of chromosome 5. "Cri du chat" means "cry of the cat" in French, and the condition was so-named because affected babies make high-pitched cries that sound like those of a cat. Affected individuals have wide-set eyes, a small head and jaw, moderate to severe mental health issues, and are very short.
- Down syndrome, usually is caused by an extra copy of chromosome 21 (trisomy 21). Characteristics include decreased muscle tone, stockier build, asymmetrical skull, slanting eyes and mild to moderate developmental disability.
- Edwards syndrome, which is the second-most-common trisomy; Down syndrome is the most common. It is a trisomy of chromosome 18. Symptoms include motor retardation, developmental disability and numerous congenital anomalies causing serious health problems. Ninety percent die in infancy; however, those that live past their first birthday usually are quite healthy thereafter. They have a characteristic clenched hands and overlapping fingers.
- Idic15, abbreviation for Isodicentric 15 on chromosome 15; also called the following names due to various researches, but they all mean the same; IDIC(15), Inverted duplication 15, extra Marker, Inv dup 15, partial tetrasomy 15
- Jacobsen syndrome, also called the terminal 11q deletion disorder. This is a very rare disorder. Those affected have normal intelligence or mild developmental disability, with poor expressive language skills. Most have a bleeding disorder called Paris-Trousseau syndrome.
- Klinefelter's syndrome (XXY). Men with Klinefelter syndrome are usually sterile, and tend to have longer arms and legs and to be taller than their peers. Boys with the syndrome are often shy and quiet, and have a higher incidence of speech delay and dyslexia. During puberty, without testosterone treatment, some of them may develop gynecomastia.
- Patau Syndrome, also called D-Syndrome or trisomy-13. Symptoms are somewhat similar to those of trisomy-18, but they do not have the characteristic hand shape.
- Small supernumerary marker chromosome. This means there is an extra, abnormal chromosome. Features depend on the origin of the extra genetic material. Cat-eye syndrome and isodicentric chromosome 15 syndrome (or Idic15) are both caused by a supernumerary marker chromosome, as is Pallister-Killian syndrome.

- Triple-X syndrome (XXX). XXX girls tend to be tall and thin. They have a higher incidence of dyslexia.
- Turner syndrome (X instead of XX or XY). In Turner syndrome, female sexual characteristics are present but underdeveloped. People with Turner syndrome often have a short stature, low hairline, abnormal eye features and bone development and a "caved-in" appearance to the chest.
- XYY syndrome. XYY boys are usually taller than their siblings. Like XXY boys and XXX girls, they are somewhat more likely to have learning difficulties.
- Wolf-Hirschhorn syndrome, which is caused by partial deletion of the short arm of chromosome 4. It is characterized by severe growth retardation and severe to profound mental health issues.

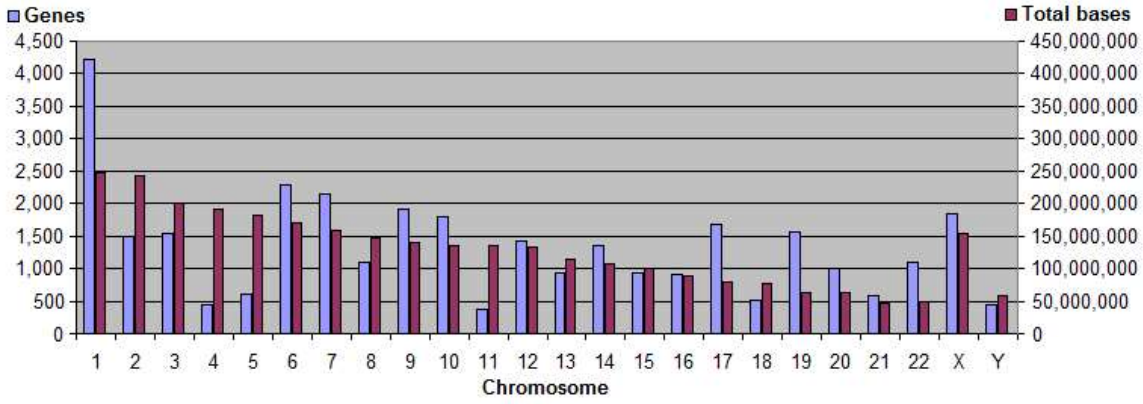
Chromosomal mutations produce changes in whole chromosomes (more than one gene) or in the number of chromosomes present.

- Deletion – loss of part of a chromosome
- Duplication – extra copies of a part of a chromosome
- Inversion – reverse the direction of a part of a chromosome
- Translocation – part of a chromosome breaks off and attaches to another chromosome

Most mutations are neutral – have little or no effect. Chromosomal aberrations are the changes in the structure of chromosomes. It has a great role in evolution. A detailed graphical display of all human chromosomes and the diseases annotated at the correct spot may be found at the Oak Ridge National Laboratory.

Human chromosomes

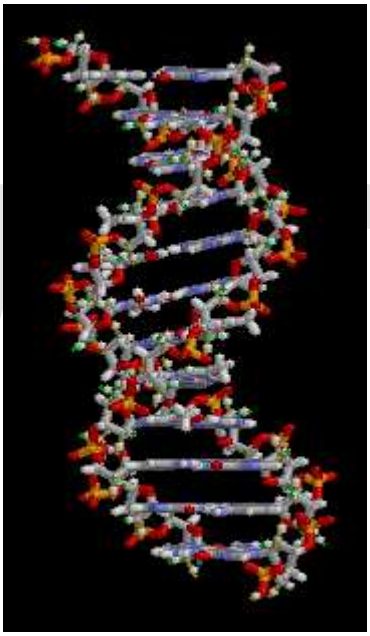
Chromosomes can be divided into two types—autosomes, and sex chromosomes. Certain genetic traits are linked to your sex, and are passed on through the sex chromosomes. The autosomes contain the rest of the genetic hereditary information. All act in the same way during cell division. Human cells have 23 pairs of large linear nuclear chromosomes, (22 pairs of autosomes and one pair of sex chromosomes) giving a total of 46 per cell. In addition to these, human cells have many hundreds of copies of the mitochondrial genome. Sequencing of the human genome has provided a great deal of information about each of the chromosomes. Below is a table compiling statistics for the chromosomes, based on the Sanger Institute's human genome information in the Vertebrate Genome Annotation (VEGA) database. Number of genes is an estimate as it is in part based on gene predictions. Total chromosome length is an estimate as well, based on the estimated size of unsequenced heterochromatin regions.



Chromosome	Genes	Total bases	Sequenced bases
1	4,220	247,199,719	224,999,719
2	1,491	242,751,149	237,712,649
3	1,550	199,446,827	194,704,827
4	446	191,263,063	187,297,063
5	609	180,837,866	177,702,766
6	2,281	170,896,993	167,273,993
7	2,135	158,821,424	154,952,424
8	1,106	146,274,826	142,612,826
9	1,920	140,442,298	120,312,298
10	1,793	135,374,737	131,624,737
11	379	134,452,384	131,130,853
12	1,430	132,289,534	130,303,534
13	924	114,127,980	95,559,980
14	1,347	106,360,585	88,290,585
15	921	100,338,915	81,341,915
16	909	88,822,254	78,884,754
17	1,672	78,654,742	77,800,220
18	519	76,117,153	74,656,155
19	1,555	63,806,651	55,785,651
20	1,008	62,435,965	59,505,254
21	578	46,944,323	34,171,998
22	1,092	49,528,953	34,893,953
X (sex chromosome)	1,846	154,913,754	151,058,754
Y (sex chromosome)	454	57,741,652	25,121,652
Total	**32,185**	**3,079,843,747**	**2,857,698,560**

Chapter- 12

DNA



The structure of part of a DNA double helix

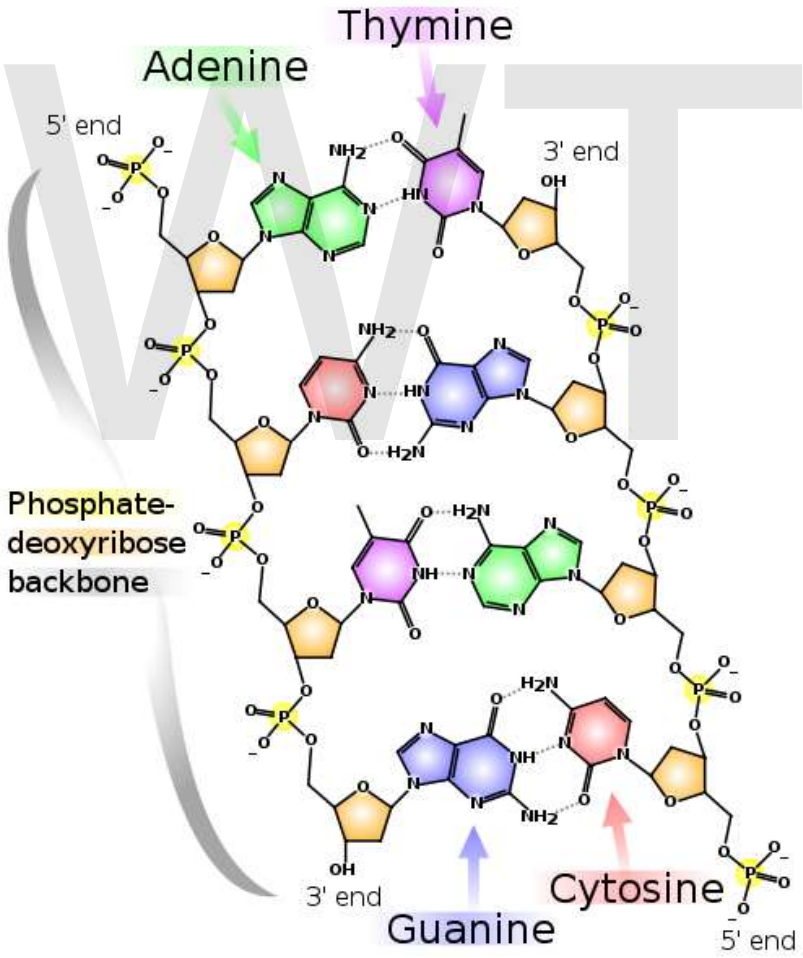
Deoxyribonucleic acid, or **DNA**, is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms (with the exception of RNA viruses). The main role of DNA molecules is the long-term storage of information. DNA is often compared to a set of blueprints, like a recipe or a code, since it contains the instructions needed to construct other components of cells, such as proteins and RNA molecules. The DNA segments that carry this genetic information are called genes, but other DNA sequences have structural purposes, or are involved in regulating the use of this genetic information.

DNA consists of two long polymers of simple units called nucleotides, with backbones made of sugars and phosphate groups joined by ester bonds. These two strands run in opposite directions to each other and are therefore anti-parallel. Attached to each sugar is one of four types of molecules called bases. It is the sequence of these four bases along the backbone that encodes information. This information is read using the genetic code,

which specifies the sequence of the amino acids within proteins. The code is read by copying stretches of DNA into the related nucleic acid RNA, in a process called transcription.

Within cells, DNA is organized into long structures called chromosomes. These chromosomes are duplicated before cells divide, in a process called DNA replication. Eukaryotic organisms (animals, plants, fungi, and protists) store most of their DNA inside the cell nucleus and some of their DNA in organelles, such as mitochondria or chloroplasts. In contrast, prokaryotes (bacteria and archaea) store their DNA only in the cytoplasm. Within the chromosomes, chromatin proteins such as histones compact and organize DNA. These compact structures guide the interactions between DNA and other proteins, helping control which parts of the DNA are transcribed.

Properties

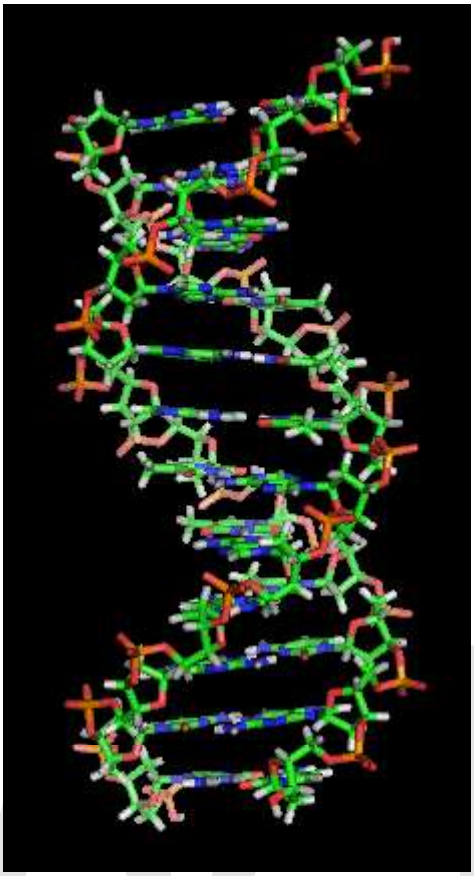


Chemical structure of DNA. Hydrogen bonds shown as dotted lines

DNA is a long polymer made from repeating units called nucleotides. As first discovered by James D. Watson and Francis Crick, the structure of DNA of all species comprises two helical chains each coiled round the same axis, and each with a pitch of 34 Ångströms (3.4 nanometres) and a radius of 10 Ångströms (1.0 nanometres). According to another study, when measured in a particular solution, the DNA chain measured 22 to 26 Ångströms wide (2.2 to 2.6 nanometres), and one nucleotide unit measured 3.3 Å (0.33 nm) long. Although each individual repeating unit is very small, DNA polymers can be very large molecules containing millions of nucleotides. For instance, the largest human chromosome, chromosome number 1, is approximately 220 million base pairs long.

In living organisms, DNA does not usually exist as a single molecule, but instead as a pair of molecules that are held tightly together. These two long strands entwine like vines, in the shape of a double helix. The nucleotide repeats contain both the segment of the backbone of the molecule, which holds the chain together, and a base, which interacts with the other DNA strand in the helix. A base linked to a sugar is called a nucleoside and a base linked to a sugar and one or more phosphate groups is called a nucleotide. If multiple nucleotides are linked together, as in DNA, this polymer is called a polynucleotide.

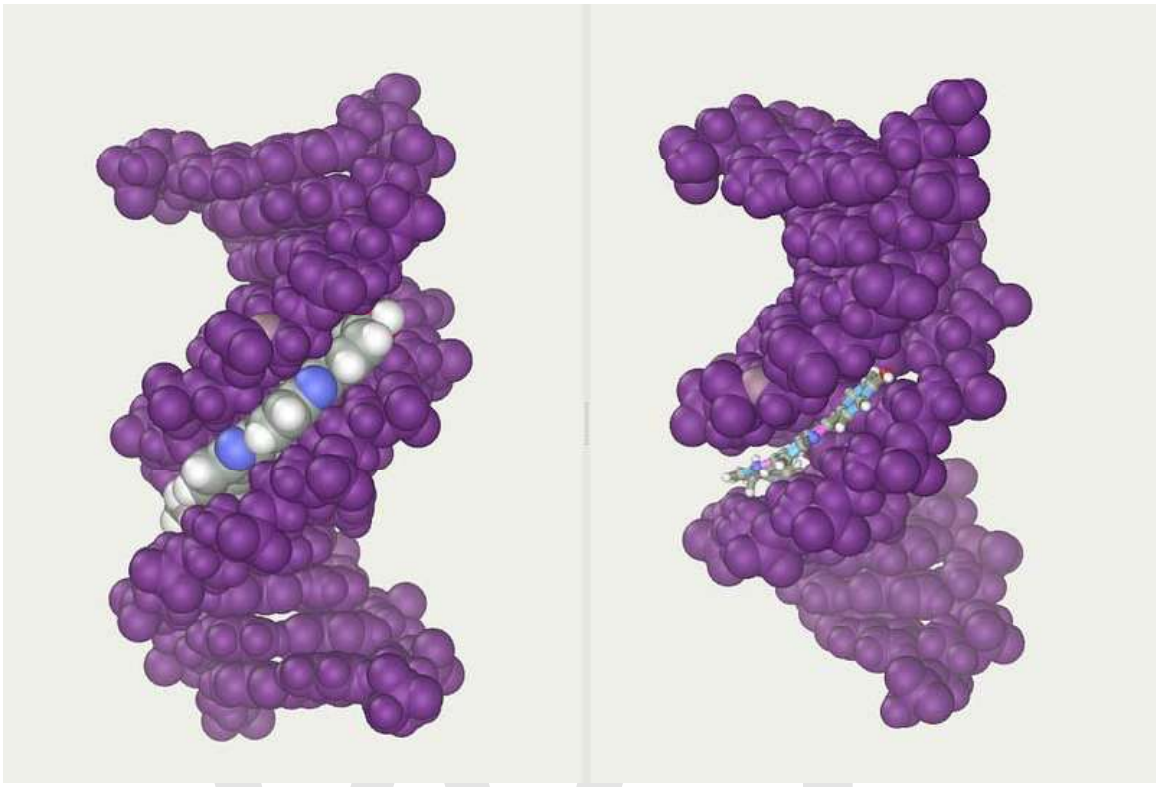
The backbone of the DNA strand is made from alternating phosphate and sugar residues. The sugar in DNA is 2-deoxyribose, which is a pentose (five-carbon) sugar. The sugars are joined together by phosphate groups that form phosphodiester bonds between the third and fifth carbon atoms of adjacent sugar rings. These asymmetric bonds mean a strand of DNA has a direction. In a double helix the direction of the nucleotides in one strand is opposite to their direction in the other strand: the strands are *antiparallel*. The asymmetric ends of DNA strands are called the 5′ (*five prime*) and 3′ (*three prime*) ends, with the 5' end having a terminal phosphate group and the 3' end a terminal hydroxyl group. One major difference between DNA and RNA is the sugar, with the 2-deoxyribose in DNA being replaced by the alternative pentose sugar ribose in RNA.



A section of DNA. The bases lie horizontally between the two spiraling strands.

The DNA double helix is stabilized by hydrogen bonds between the bases attached to the two strands. The four bases found in DNA are adenine (abbreviated A), cytosine (C), guanine (G) and thymine (T). These four bases are attached to the sugar/phosphate to form the complete nucleotide, as shown for adenosine monophosphate.

These bases are classified into two types; adenine and guanine are fused five- and six-membered heterocyclic compounds called purines, while cytosine and thymine are six-membered rings called pyrimidines. A fifth pyrimidine base, called uracil (U), usually takes the place of thymine in RNA and differs from thymine by lacking a methyl group on its ring. Uracil is not usually found in DNA, occurring only as a breakdown product of cytosine. In addition to RNA and DNA, a large number of artificial nucleic acid analogues have also been created to study the proprieties of nucleic acids, or for use in biotechnology.



Major and minor grooves of DNA. Minor groove is a binding site for the dye Hoechst 33258.

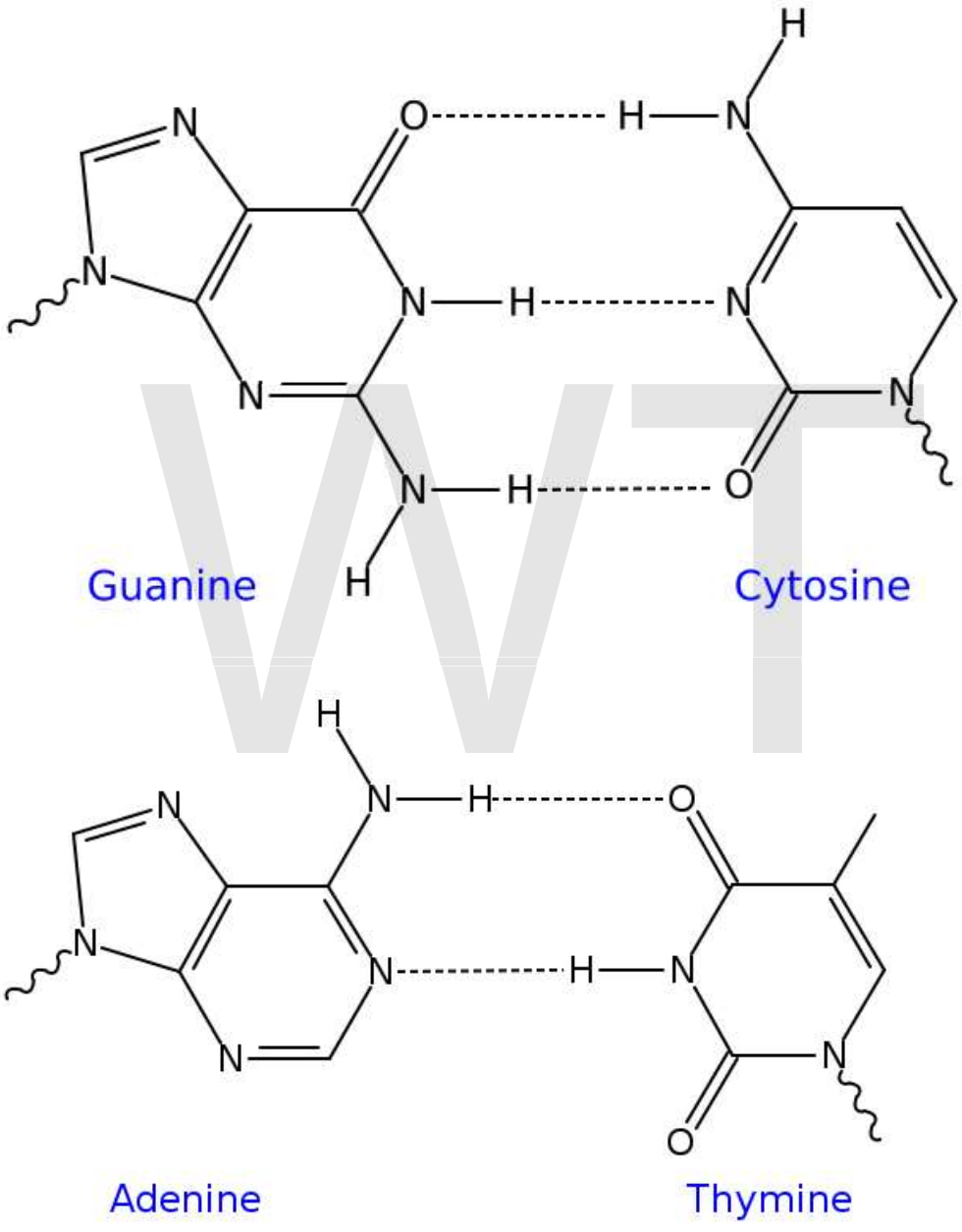
Grooves

Twin helical strands form the DNA backbone. Another double helix may be found by tracing the spaces, or grooves, between the strands. These voids are adjacent to the base pairs and may provide a binding site. As the strands are not directly opposite each other, the grooves are unequally sized. One groove, the major groove, is 22 Å wide and the other, the minor groove, is 12 Å wide. The narrowness of the minor groove means that the edges of the bases are more accessible in the major groove. As a result, proteins like transcription factors that can bind to specific sequences in double-stranded DNA usually make contacts to the sides of the bases exposed in the major groove. This situation varies in unusual conformations of DNA within the cell *(see below)*, but the major and minor grooves are always named to reflect the differences in size that would be seen if the DNA is twisted back into the ordinary B form.

Base pairing

Each type of base on one strand forms a bond with just one type of base on the other strand. This is called complementary base pairing. Here, purines form hydrogen bonds to pyrimidines, with A bonding only to T, and C bonding only to G. This arrangement of two nucleotides binding together across the double helix is called a base pair. As hydrogen bonds are not covalent, they can be broken and rejoined relatively easily. The

two strands of DNA in a double helix can therefore be pulled apart like a zipper, either by a mechanical force or high temperature. As a result of this complementarity, all the information in the double-stranded sequence of a DNA helix is duplicated on each strand, which is vital in DNA replication. Indeed, this reversible and specific interaction between complementary base pairs is critical for all the functions of DNA in living organisms.



Top, a **GC** base pair with three hydrogen bonds. Bottom, an **AT** base pair with two hydrogen bonds. Non-covalent hydrogen bonds between the pairs are shown as dashed lines.

The two types of base pairs form different numbers of hydrogen bonds, AT forming two hydrogen bonds, and GC forming three hydrogen bonds. DNA with high GC-content is more stable than DNA with low GC-content, but contrary to popular belief, this is not due to the extra hydrogen bond of a GC base pair but rather the contribution of stacking interactions (hydrogen bonding merely provides specificity of the pairing, not stability).

As a result, it is both the percentage of GC base pairs and the overall length of a DNA double helix that determine the strength of the association between the two strands of DNA. Long DNA helices with a high GC content have stronger-interacting strands, while short helices with high AT content have weaker-interacting strands. In biology, parts of the DNA double helix that need to separate easily, such as the TATAAT Pribnow box in some promoters, tend to have a high AT content, making the strands easier to pull apart.

In the laboratory, the strength of this interaction can be measured by finding the temperature required to break the hydrogen bonds, their melting temperature (also called T_m value). When all the base pairs in a DNA double helix melt, the strands separate and exist in solution as two entirely independent molecules. These single-stranded DNA molecules (*ssDNA*) have no single common shape, but some conformations are more stable than others.

Sense and antisense

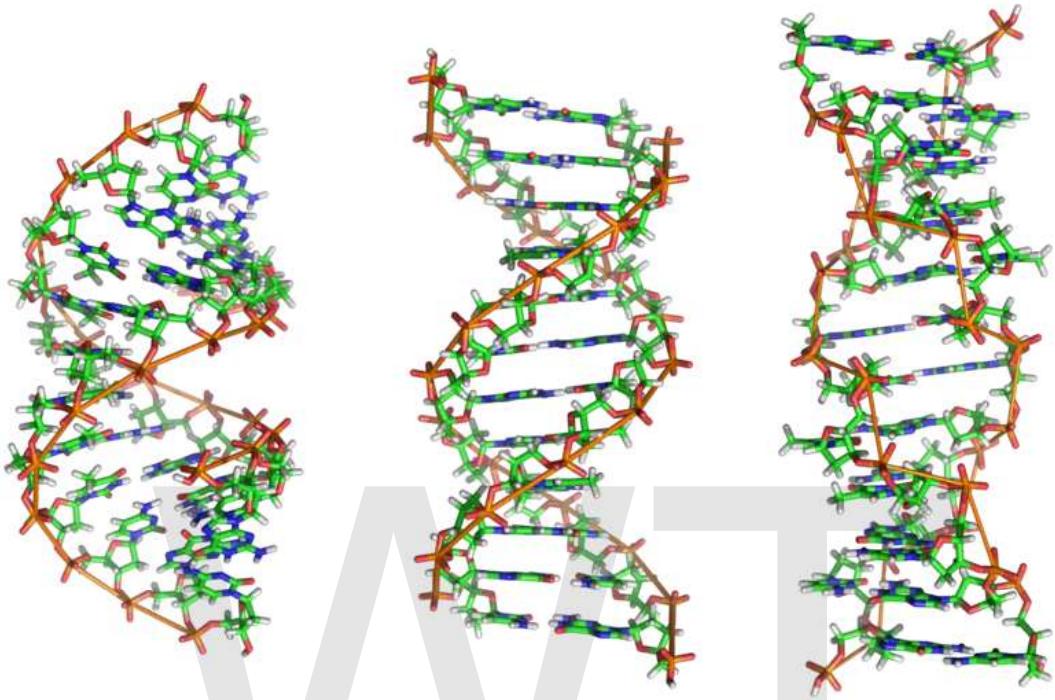
A DNA sequence is called "sense" if its sequence is the same as that of a messenger RNA copy that is translated into protein. The sequence on the opposite strand is called the "antisense" sequence. Both sense and antisense sequences can exist on different parts of the same strand of DNA (i.e. both strands contain both sense and antisense sequences). In both prokaryotes and eukaryotes, antisense RNA sequences are produced, but the functions of these RNAs are not entirely clear. One proposal is that antisense RNAs are involved in regulating gene expression through RNA-RNA base pairing.

A few DNA sequences in prokaryotes and eukaryotes, and more in plasmids and viruses, blur the distinction between sense and antisense strands by having overlapping genes. In these cases, some DNA sequences do double duty, encoding one protein when read along one strand, and a second protein when read in the opposite direction along the other strand. In bacteria, this overlap may be involved in the regulation of gene transcription, while in viruses, overlapping genes increase the amount of information that can be encoded within the small viral genome.

Supercoiling

DNA can be twisted like a rope in a process called DNA supercoiling. With DNA in its "relaxed" state, a strand usually circles the axis of the double helix once every 10.4 base pairs, but if the DNA is twisted the strands become more tightly or more loosely wound. If the DNA is twisted in the direction of the helix, this is positive supercoiling, and the bases are held more tightly together. If they are twisted in the opposite direction, this is negative supercoiling, and the bases come apart more easily. In nature, most DNA has

slight negative supercoiling that is introduced by enzymes called topoisomerases. These enzymes are also needed to relieve the twisting stresses introduced into DNA strands during processes such as transcription and DNA replication.



From left to right, the structures of A, B and Z DNA

Alternate DNA structures

DNA exists in many possible conformations that include A-DNA, B-DNA, and Z-DNA forms, although, only B-DNA and Z-DNA have been directly observed in functional organisms. The conformation that DNA adopts depends on the hydration level, DNA sequence, the amount and direction of supercoiling, chemical modifications of the bases, the type and concentration of metal ions, as well as the presence of polyamines in solution.

The first published reports of A-DNA X-ray diffraction patterns— and also B-DNA used analyses based on Patterson transforms that provided only a limited amount of structural information for oriented fibers of DNA. An alternate analysis was then proposed by Wilkins *et al.*, in 1953, for the *in vivo* B-DNA X-ray diffraction/scattering patterns of highly hydrated DNA fibers in terms of squares of Bessel functions. In the same journal, James D. Watson and Francis Crick presented their molecular modeling analysis of the DNA X-ray diffraction patterns to suggest that the structure was a double-helix.

Although the `B-DNA form' is most common under the conditions found in cells, it is not a well-defined conformation but a family of related DNA conformations that occur at the high hydration levels present in living cells. Their corresponding X-ray diffraction and

scattering patterns are characteristic of molecular paracrystals with a significant degree of disorder.

Compared to B-DNA, the A-DNA form is a wider right-handed spiral, with a shallow, wide minor groove and a narrower, deeper major groove. The A form occurs under non-physiological conditions in partially dehydrated samples of DNA, while in the cell it may be produced in hybrid pairings of DNA and RNA strands, as well as in enzyme-DNA complexes. Segments of DNA where the bases have been chemically modified by methylation may undergo a larger change in conformation and adopt the Z form. Here, the strands turn about the helical axis in a left-handed spiral, the opposite of the more common B form. These unusual structures can be recognized by specific Z-DNA binding proteins and may be involved in the regulation of transcription.

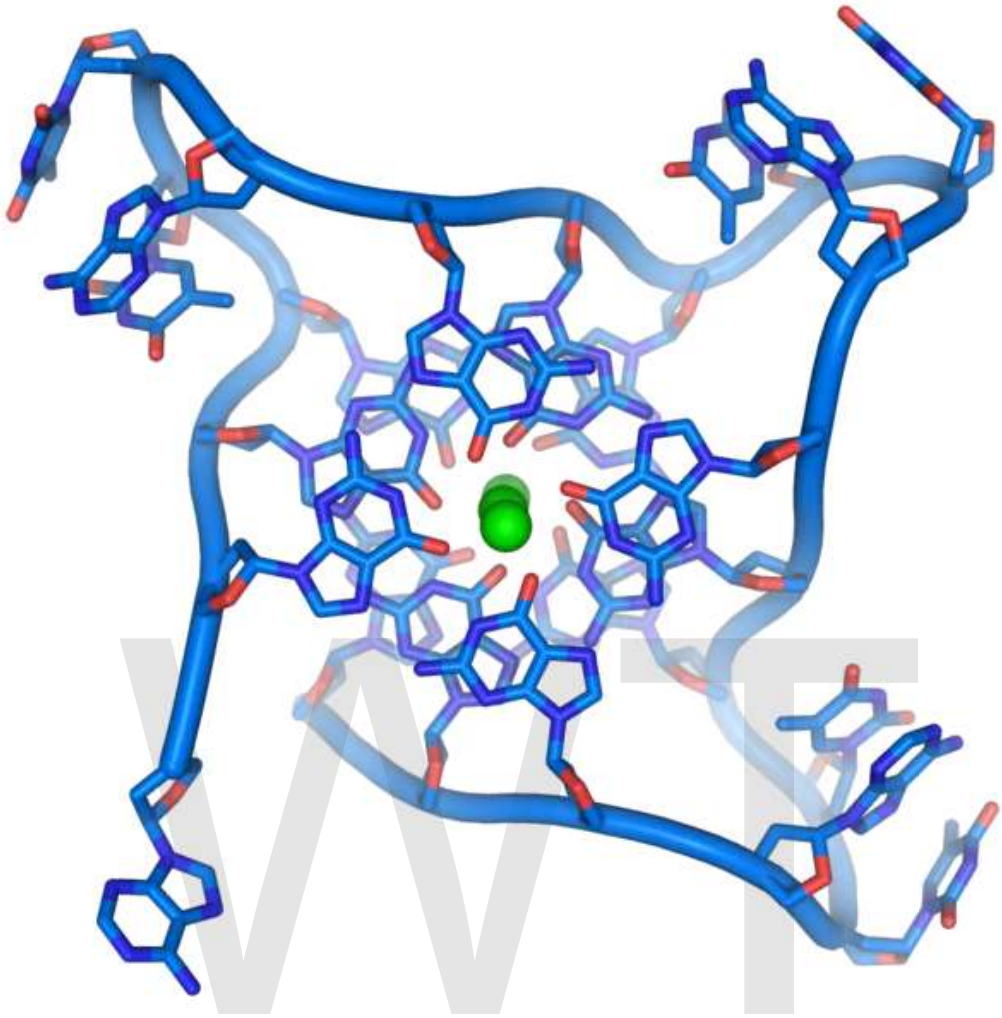
Alternate DNA chemistry

For a number of years exobiologists have proposed the existence of a shadow biosphere, a postulated microbial biosphere of Earth that uses radically different biochemical and molecular processes than currently known life. One of the proposals was the existence of lifeforms that use arsenic instead of phosphorus in DNA.

A December 2010 NASA press conference revealed that the bacterium GFAJ-1, which has evolved in an arsenic-rich environment, is the first terrestrial lifeform found which may have this ability. The bacterium was found in Mono Lake, east of Yosemite National Park. GFAJ-1 is a rod-shaped extremophile bacterium in the family Halomonadaceae that, when starved of phosphorus, may be capable of incorporating the usually poisonous element arsenic in its DNA. This discovery lends weight to the long-standing idea that extraterrestrial life could have a different chemical makeup from life on Earth. The research was carried out by a team led by Felisa Wolfe-Simon, a geomicrobiologist and geobiochemist, a Postdoctoral Fellow of the NASA Astrobiology Institute with Arizona State University.

Quadruplex structures

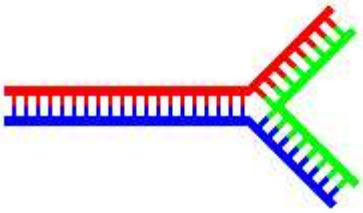
At the ends of the linear chromosomes are specialized regions of DNA called telomeres. The main function of these regions is to allow the cell to replicate chromosome ends using the enzyme telomerase, as the enzymes that normally replicate DNA cannot copy the extreme 3′ ends of chromosomes. These specialized chromosome caps also help protect the DNA ends, and stop the DNA repair systems in the cell from treating them as damage to be corrected. In human cells, telomeres are usually lengths of single-stranded DNA containing several thousand repeats of a simple TTAGGG sequence.



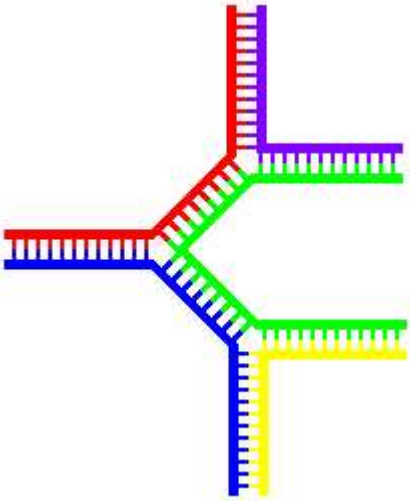
DNA quadruplex formed by telomere repeats. The looped conformation of the DNA backbone is very different from the typical DNA helix.

These guanine-rich sequences may stabilize chromosome ends by forming structures of stacked sets of four-base units, rather than the usual base pairs found in other DNA molecules. Here, four guanine bases form a flat plate and these flat four-base units then stack on top of each other, to form a stable *G-quadruplex* structure. These structures are stabilized by hydrogen bonding between the edges of the bases and chelation of a metal ion in the centre of each four-base unit. Other structures can also be formed, with the central set of four bases coming from either a single strand folded around the bases, or several different parallel strands, each contributing one base to the central structure.

In addition to these stacked structures, telomeres also form large loop structures called telomere loops, or T-loops. Here, the single-stranded DNA curls around in a long circle stabilized by telomere-binding proteins. At the very end of the T-loop, the single-stranded telomere DNA is held onto a region of double-stranded DNA by the telomere strand disrupting the double-helical DNA and base pairing to one of the two strands. This triple-stranded structure is called a displacement loop or D-loop.



Single branch



Multiple branches

Branched DNA can form networks containing multiple branches.

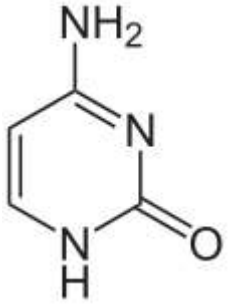
Branched DNA

In DNA fraying occurs when non-complementary regions exist at the end of an otherwise complementary double-strand of DNA. However, branched DNA can occur if a third strand of DNA is introduced and contains adjoining regions able to hybridize with the frayed regions of the pre-existing double-strand. Although the simplest example of branched DNA involves only three strands of DNA, complexes involving additional strands and multiple branches are also possible. Branched DNA can be used in nanotechnology to construct geometric shapes, see the section on uses in technology below.

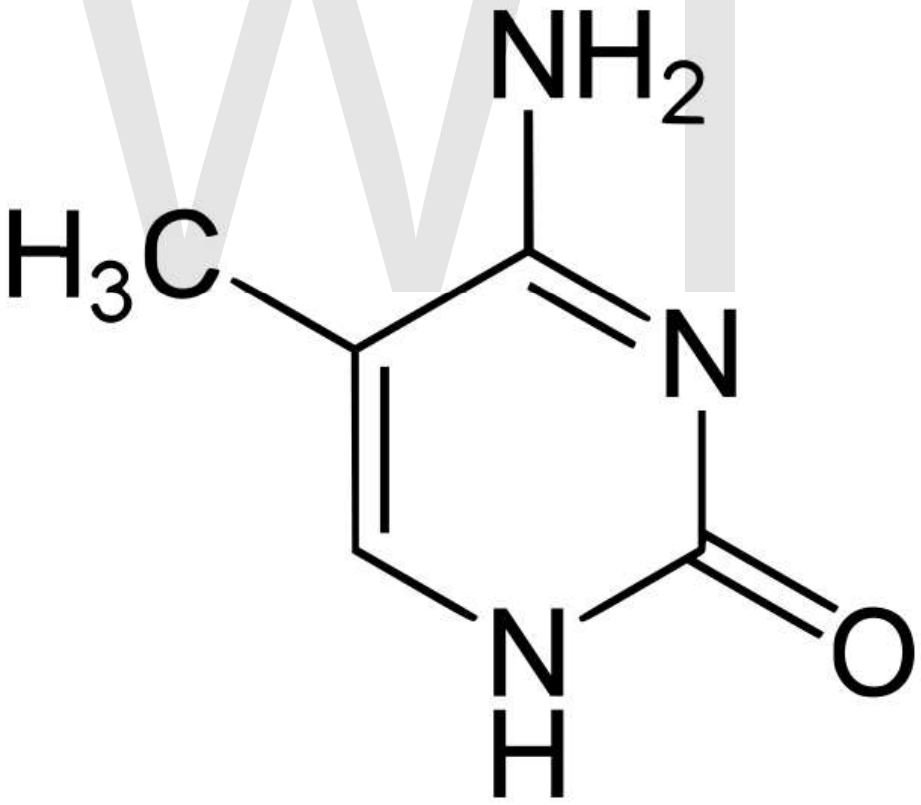
Vibration

DNA may carry out low-frequency collective motion as observed by the Raman spectroscopy and analyzed with a quasi-continuum model.

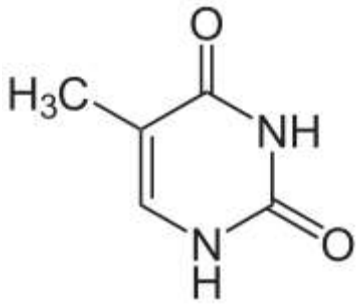
Chemical modifications



cytosine



5-methylcytosine



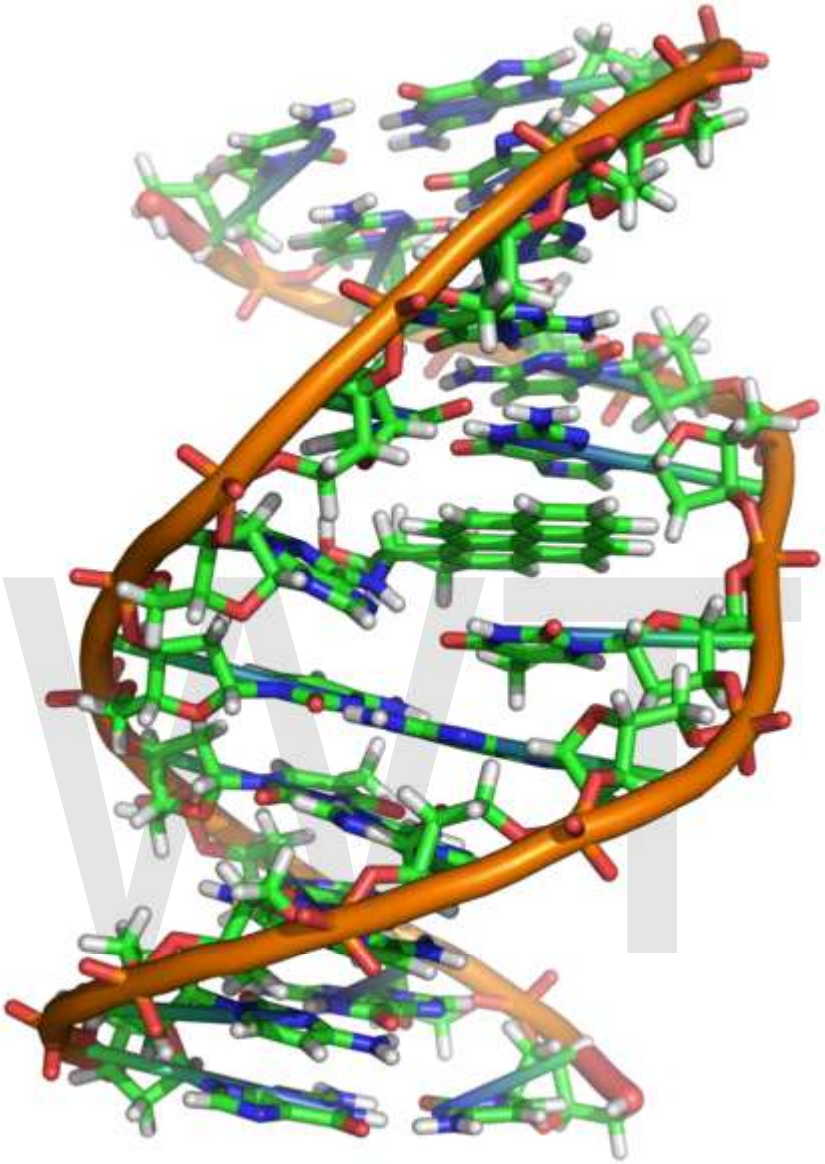
thymine

Structure of cytosine with and without the 5-methyl group. Deamination converts 5-methylcytosine into thymine.

Base modifications

The expression of genes is influenced by how the DNA is packaged in chromosomes, in a structure called chromatin. Base modifications can be involved in packaging, with regions that have low or no gene expression usually containing high levels of methylation of cytosine bases. For example, cytosine methylation, produces 5-methylcytosine, which is important for X-chromosome inactivation. The average level of methylation varies between organisms - the worm *Caenorhabditis elegans* lacks cytosine methylation, while vertebrates have higher levels, with up to 1% of their DNA containing 5-methylcytosine. Despite the importance of 5-methylcytosine, it can deaminate to leave a thymine base, so methylated cytosines are particularly prone to mutations. Other base modifications include adenine methylation in bacteria, the presence of 5-hydroxymethylcytosine in the brain, and the glycosylation of uracil to produce the "J-base" in kinetoplastids.

Damage



A covalent adduct between a metabolically activated form of benzo[*a*]pyrene, the major mutagen in tobacco smoke, and DNA

DNA can be damaged by many sorts of mutagens, which change the DNA sequence. Mutagens include oxidizing agents, alkylating agents and also high-energy electromagnetic radiation such as ultraviolet light and X-rays. The type of DNA damage produced depends on the type of mutagen. For example, UV light can damage DNA by producing thymine dimers, which are cross-links between pyrimidine bases. On the other hand, oxidants such as free radicals or hydrogen peroxide produce multiple forms of damage, including base modifications, particularly of guanosine, and double-strand breaks. A typical human cell contains about 150,000 bases that have suffered oxidative damage. Of these oxidative lesions, the most dangerous are double-strand breaks, as these

are difficult to repair and can produce point mutations, insertions and deletions from the DNA sequence, as well as chromosomal translocations.

Many mutagens fit into the space between two adjacent base pairs, this is called *intercalation*. Most intercalators are aromatic and planar molecules; examples include ethidium bromide, daunomycin, and doxorubicin. In order for an intercalator to fit between base pairs, the bases must separate, distorting the DNA strands by unwinding of the double helix. This inhibits both transcription and DNA replication, causing toxicity and mutations. As a result, DNA intercalators are often carcinogens, and benzo[*a*]pyrene diol epoxide, acridines, aflatoxin and ethidium bromide are well-known examples. Nevertheless, due to their ability to inhibit DNA transcription and replication, other similar toxins are also used in chemotherapy to inhibit rapidly growing cancer cells.

Biological functions

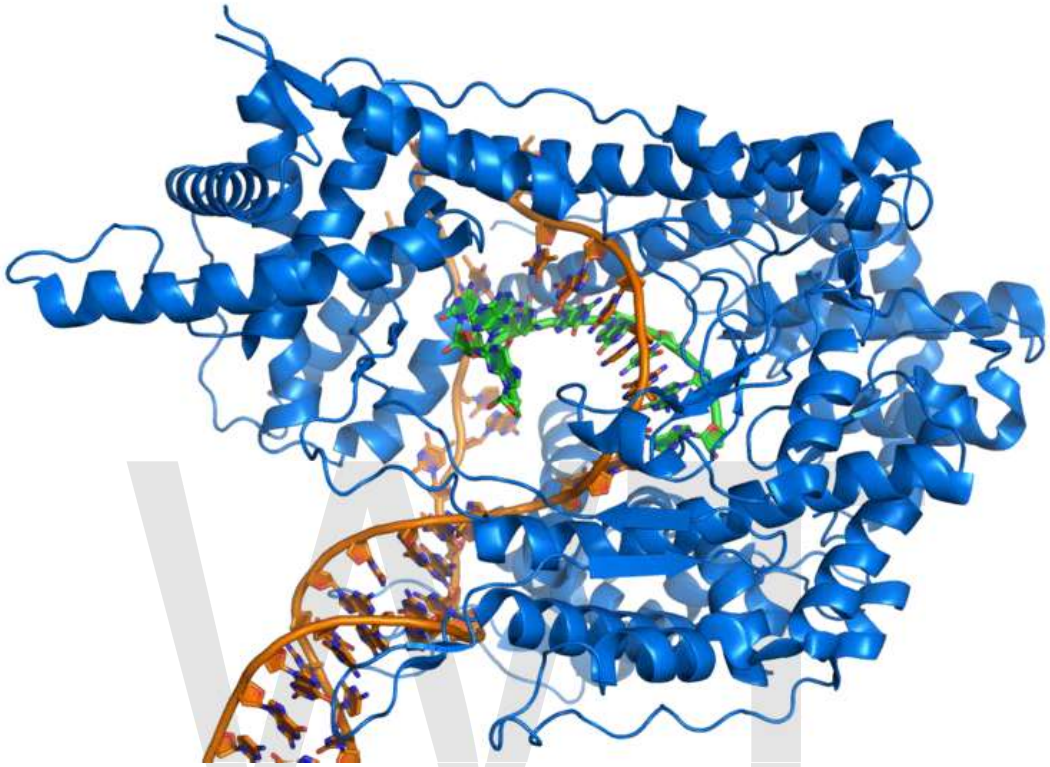
DNA usually occurs as linear chromosomes in eukaryotes, and circular chromosomes in prokaryotes. The set of chromosomes in a cell makes up its genome; the human genome has approximately 3 billion base pairs of DNA arranged into 46 chromosomes. The information carried by DNA is held in the sequence of pieces of DNA called genes. Transmission of genetic information in genes is achieved via complementary base pairing. For example, in transcription, when a cell uses the information in a gene, the DNA sequence is copied into a complementary RNA sequence through the attraction between the DNA and the correct RNA nucleotides. Usually, this RNA copy is then used to make a matching protein sequence in a process called translation, which depends on the same interaction between RNA nucleotides. In alternative fashion, a cell may simply copy its genetic information in a process called DNA replication. The details of these functions are covered in other articles; here we focus on the interactions between DNA and other molecules that mediate the function of the genome.

Genes and genomes

Genomic DNA is tightly and orderly packed in the process called DNA condensation to fit the small available volumes of the cell. In eukaryotes, DNA is located in the cell nucleus, as well as small amounts in mitochondria and chloroplasts. In prokaryotes, the DNA is held within an irregularly shaped body in the cytoplasm called the nucleoid. The genetic information in a genome is held within genes, and the complete set of this information in an organism is called its genotype. A gene is a unit of heredity and is a region of DNA that influences a particular characteristic in an organism. Genes contain an open reading frame that can be transcribed, as well as regulatory sequences such as promoters and enhancers, which control the transcription of the open reading frame.

In many species, only a small fraction of the total sequence of the genome encodes protein. For example, only about 1.5% of the human genome consists of protein-coding exons, with over 50% of human DNA consisting of non-coding repetitive sequences. The reasons for the presence of so much noncoding DNA in eukaryotic genomes and the extraordinary differences in genome size, or *C-value*, among species represent a long-

standing puzzle known as the "C-value enigma". However, DNA sequences that do not code protein may still encode functional non-coding RNA molecules, which are involved in the regulation of gene expression.



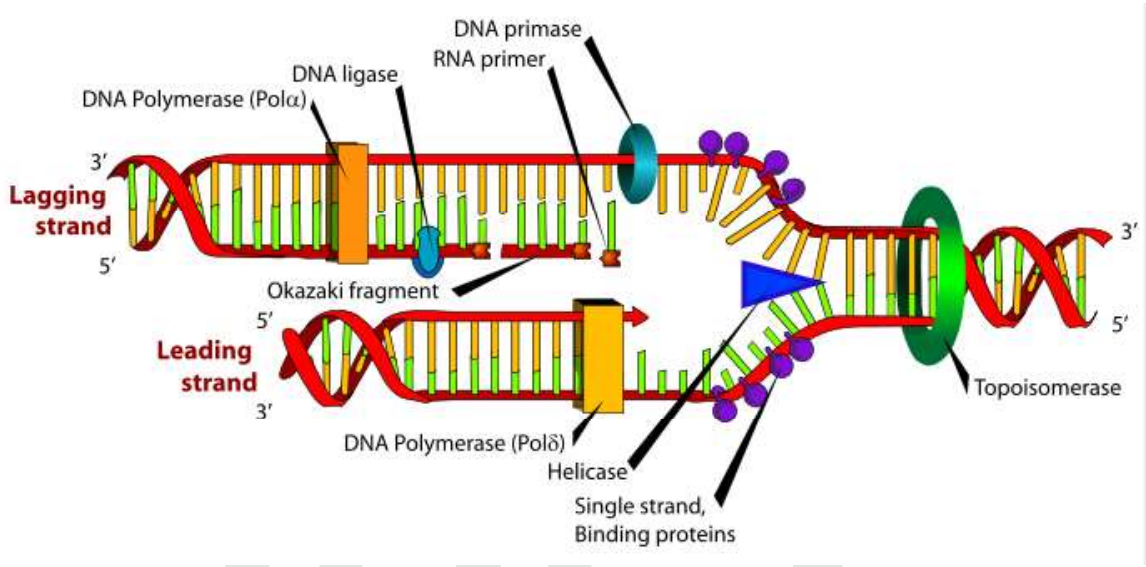
T7 RNA polymerase (blue) producing a mRNA (green) from a DNA template (orange).

Some noncoding DNA sequences play structural roles in chromosomes. Telomeres and centromeres typically contain few genes, but are important for the function and stability of chromosomes. An abundant form of noncoding DNA in humans are pseudogenes, which are copies of genes that have been disabled by mutation. These sequences are usually just molecular fossils, although they can occasionally serve as raw genetic material for the creation of new genes through the process of gene duplication and divergence.

Transcription and translation

A gene is a sequence of DNA that contains genetic information and can influence the phenotype of an organism. Within a gene, the sequence of bases along a DNA strand defines a messenger RNA sequence, which then defines one or more protein sequences. The relationship between the nucleotide sequences of genes and the amino-acid sequences of proteins is determined by the rules of translation, known collectively as the genetic code. The genetic code consists of three-letter 'words' called *codons* formed from a sequence of three nucleotides (e.g. ACT, CAG, TTT).

In transcription, the codons of a gene are copied into messenger RNA by RNA polymerase. This RNA copy is then decoded by a ribosome that reads the RNA sequence by base-pairing the messenger RNA to transfer RNA, which carries amino acids. Since there are 4 bases in 3-letter combinations, there are 64 possible codons (4^3 combinations). These encode the twenty standard amino acids, giving most amino acids more than one possible codon. There are also three 'stop' or 'nonsense' codons signifying the end of the coding region; these are the TAA, TGA and TAG codons.



DNA replication. The double helix is unwound by a helicase and topoisomerase. Next, one DNA polymerase produces the leading strand copy. Another DNA polymerase binds to the lagging strand. This enzyme makes discontinuous segments (called Okazaki fragments) before DNA ligase joins them together.

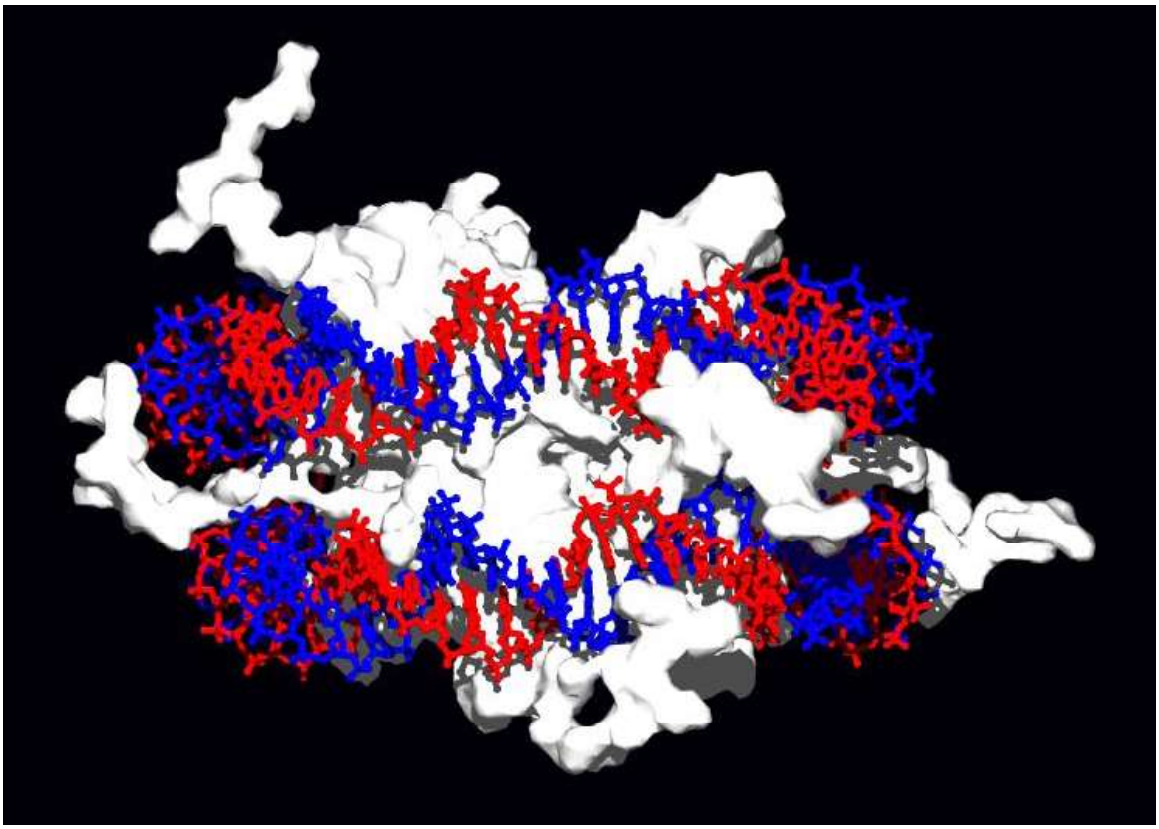
Replication

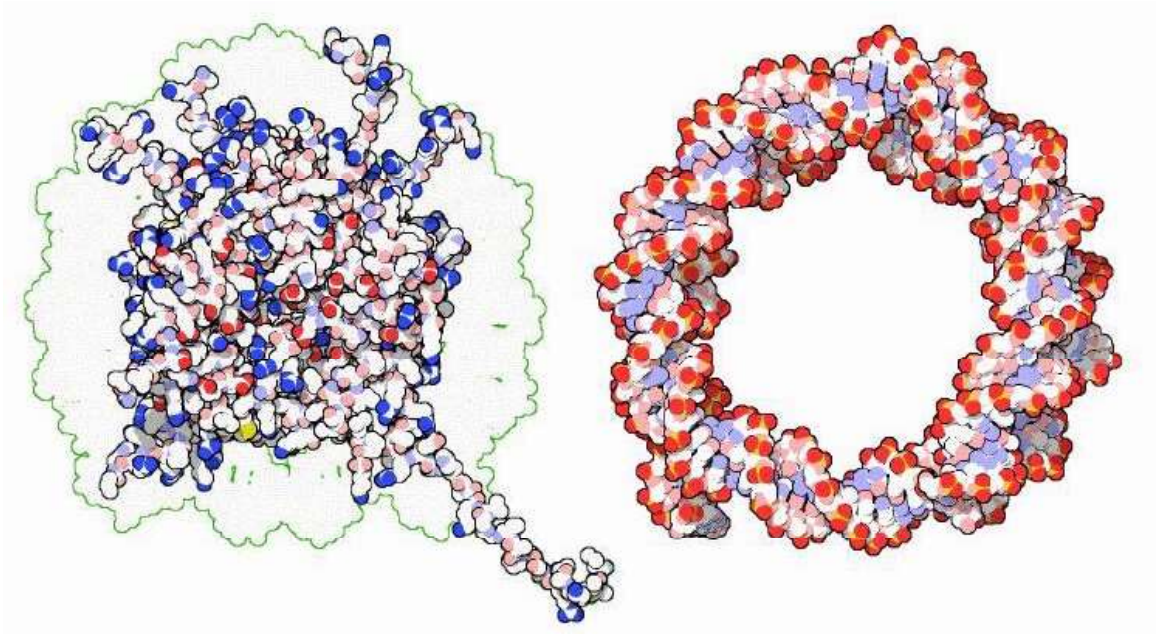
Cell division is essential for an organism to grow, but, when a cell divides, it must replicate the DNA in its genome so that the two daughter cells have the same genetic information as their parent. The double-stranded structure of DNA provides a simple mechanism for DNA replication. Here, the two strands are separated and then each strand's complementary DNA sequence is recreated by an enzyme called DNA polymerase. This enzyme makes the complementary strand by finding the correct base through complementary base pairing, and bonding it onto the original strand. As DNA polymerases can only extend a DNA strand in a 5′ to 3′ direction, different mechanisms are used to copy the antiparallel strands of the double helix. In this way, the base on the old strand dictates which base appears on the new strand, and the cell ends up with a perfect copy of its DNA.

Interactions with proteins

All the functions of DNA depend on interactions with proteins. These protein interactions can be non-specific, or the protein can bind specifically to a single DNA sequence. Enzymes can also bind to DNA and of these, the polymerases that copy the DNA base sequence in transcription and DNA replication are particularly important.

DNA-binding proteins

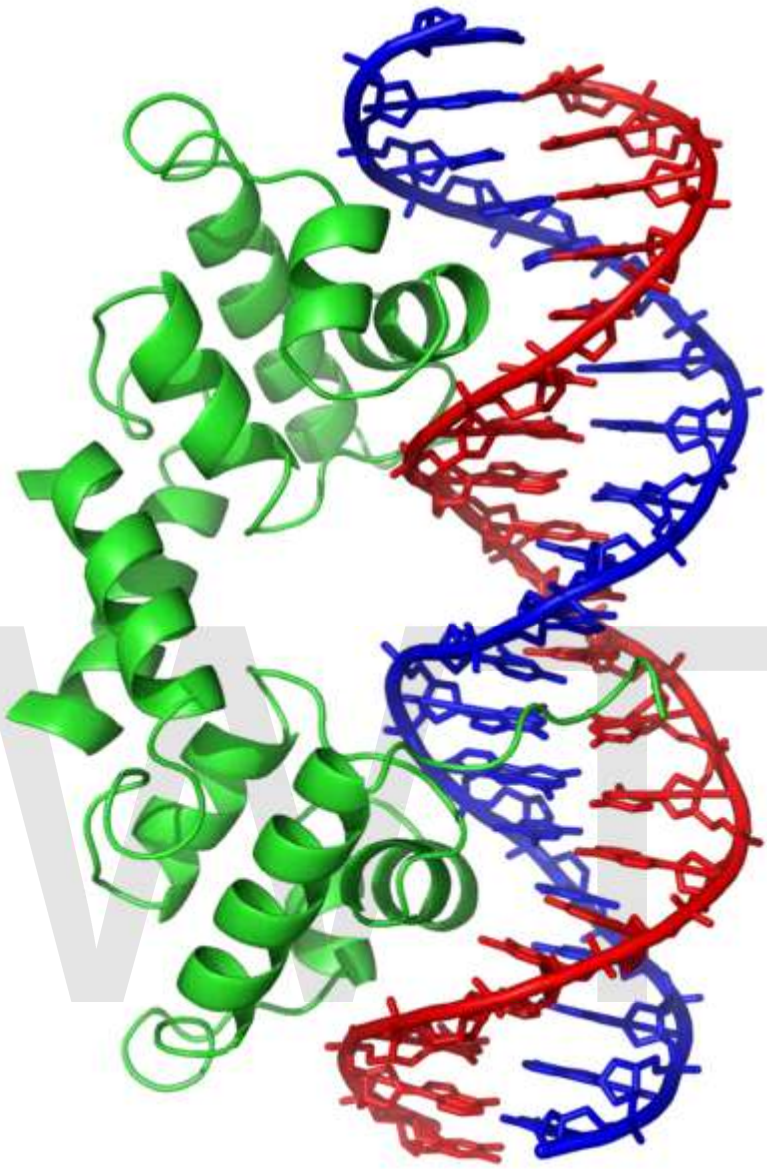




Interaction of DNA with histones (shown in white, top). These proteins' basic amino acids (below left, blue) bind to the acidic phosphate groups on DNA (below right, red).

Structural proteins that bind DNA are well-understood examples of non-specific DNA-protein interactions. Within chromosomes, DNA is held in complexes with structural proteins. These proteins organize the DNA into a compact structure called chromatin. In eukaryotes this structure involves DNA binding to a complex of small basic proteins called histones, while in prokaryotes multiple types of proteins are involved. The histones form a disk-shaped complex called a nucleosome, which contains two complete turns of double-stranded DNA wrapped around its surface. These non-specific interactions are formed through basic residues in the histones making ionic bonds to the acidic sugar-phosphate backbone of the DNA, and are therefore largely independent of the base sequence. Chemical modifications of these basic amino acid residues include methylation, phosphorylation and acetylation. These chemical changes alter the strength of the interaction between the DNA and the histones, making the DNA more or less accessible to transcription factors and changing the rate of transcription. Other non-specific DNA-binding proteins in chromatin include the high-mobility group proteins, which bind to bent or distorted DNA. These proteins are important in bending arrays of nucleosomes and arranging them into the larger structures that make up chromosomes.

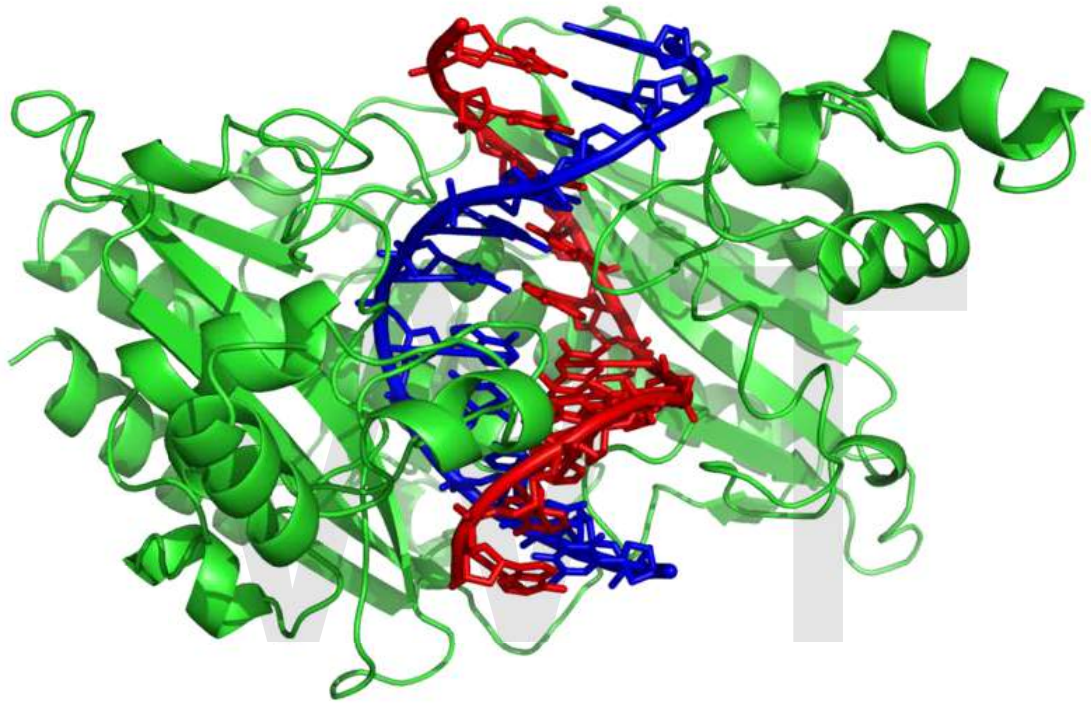
A distinct group of DNA-binding proteins are the DNA-binding proteins that specifically bind single-stranded DNA. In humans, replication protein A is the best-understood member of this family and is used in processes where the double helix is separated, including DNA replication, recombination and DNA repair. These binding proteins seem to stabilize single-stranded DNA and protect it from forming stem-loops or being degraded by nucleases.



The lambda repressor helix-turn-helix transcription factor bound to its DNA target

In contrast, other proteins have evolved to bind to particular DNA sequences. The most intensively studied of these are the various transcription factors, which are proteins that regulate transcription. Each transcription factor binds to one particular set of DNA sequences and activates or inhibits the transcription of genes that have these sequences close to their promoters. The transcription factors do this in two ways. Firstly, they can bind the RNA polymerase responsible for transcription, either directly or through other mediator proteins; this locates the polymerase at the promoter and allows it to begin transcription. Alternatively, transcription factors can bind enzymes that modify the histones at the promoter; this will change the accessibility of the DNA template to the polymerase.

As these DNA targets can occur throughout an organism's genome, changes in the activity of one type of transcription factor can affect thousands of genes. Consequently, these proteins are often the targets of the signal transduction processes that control responses to environmental changes or cellular differentiation and development. The specificity of these transcription factors' interactions with DNA come from the proteins making multiple contacts to the edges of the DNA bases, allowing them to "read" the DNA sequence. Most of these base-interactions are made in the major groove, where the bases are most accessible.



The restriction enzyme EcoRV (green) in a complex with its substrate DNA

DNA-modifying enzymes

Nucleases and ligases

Nucleases are enzymes that cut DNA strands by catalyzing the hydrolysis of the phosphodiester bonds. Nucleases that hydrolyse nucleotides from the ends of DNA strands are called exonucleases, while endonucleases cut within strands. The most frequently used nucleases in molecular biology are the restriction endonucleases, which cut DNA at specific sequences. For instance, the EcoRV enzyme shown to the left recognizes the 6-base sequence 5′-GAT|ATC-3′ and makes a cut at the vertical line. In nature, these enzymes protect bacteria against phage infection by digesting the phage DNA when it enters the bacterial cell, acting as part of the restriction modification

system. In technology, these sequence-specific nucleases are used in molecular cloning and DNA fingerprinting.

Enzymes called DNA ligases can rejoin cut or broken DNA strands. Ligases are particularly important in lagging strand DNA replication, as they join together the short segments of DNA produced at the replication fork into a complete copy of the DNA template. They are also used in DNA repair and genetic recombination.

Topoisomerases and helicases

Topoisomerases are enzymes with both nuclease and ligase activity. These proteins change the amount of supercoiling in DNA. Some of these enzymes work by cutting the DNA helix and allowing one section to rotate, thereby reducing its level of supercoiling; the enzyme then seals the DNA break. Other types of these enzymes are capable of cutting one DNA helix and then passing a second strand of DNA through this break, before rejoining the helix. Topoisomerases are required for many processes involving DNA, such as DNA replication and transcription.

Helicases are proteins that are a type of molecular motor. They use the chemical energy in nucleoside triphosphates, predominantly ATP, to break hydrogen bonds between bases and unwind the DNA double helix into single strands. These enzymes are essential for most processes where enzymes need to access the DNA bases.

Polymerases

Polymerases are enzymes that synthesize polynucleotide chains from nucleoside triphosphates. The sequence of their products are copies of existing polynucleotide chains - which are called *templates*. These enzymes function by adding nucleotides onto the 3′ hydroxyl group of the previous nucleotide in a DNA strand. As a consequence, all polymerases work in a 5′ to 3′ direction. In the active site of these enzymes, the incoming nucleoside triphosphate base-pairs to the template: this allows polymerases to accurately synthesize the complementary strand of their template. Polymerases are classified according to the type of template that they use.

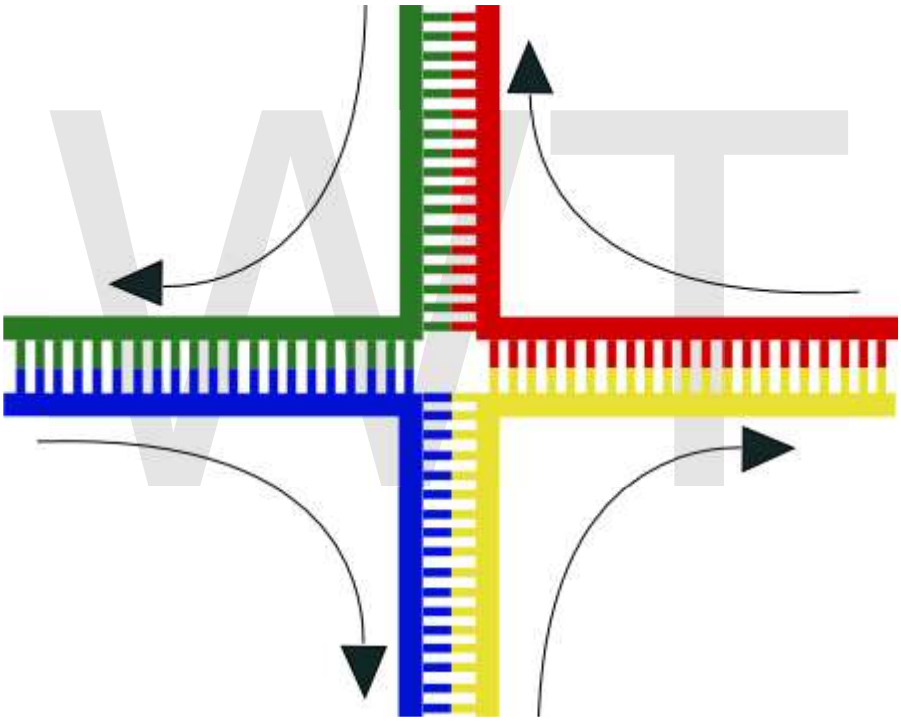
In DNA replication, a DNA-dependent DNA polymerase makes a copy of a DNA sequence. Accuracy is vital in this process, so many of these polymerases have a proofreading activity. Here, the polymerase recognizes the occasional mistakes in the synthesis reaction by the lack of base pairing between the mismatched nucleotides. If a mismatch is detected, a 3′ to 5′ exonuclease activity is activated and the incorrect base removed. In most organisms, DNA polymerases function in a large complex called the replisome that contains multiple accessory subunits, such as the DNA clamp or helicases.

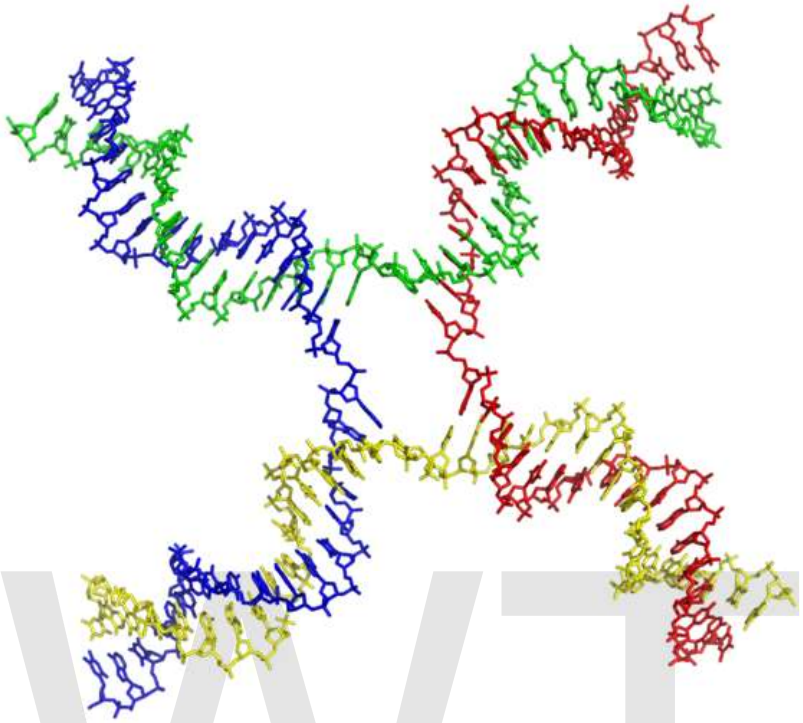
RNA-dependent DNA polymerases are a specialized class of polymerases that copy the sequence of an RNA strand into DNA. They include reverse transcriptase, which is a viral enzyme involved in the infection of cells by retroviruses, and telomerase, which is

required for the replication of telomeres. Telomerase is an unusual polymerase because it contains its own RNA template as part of its structure.

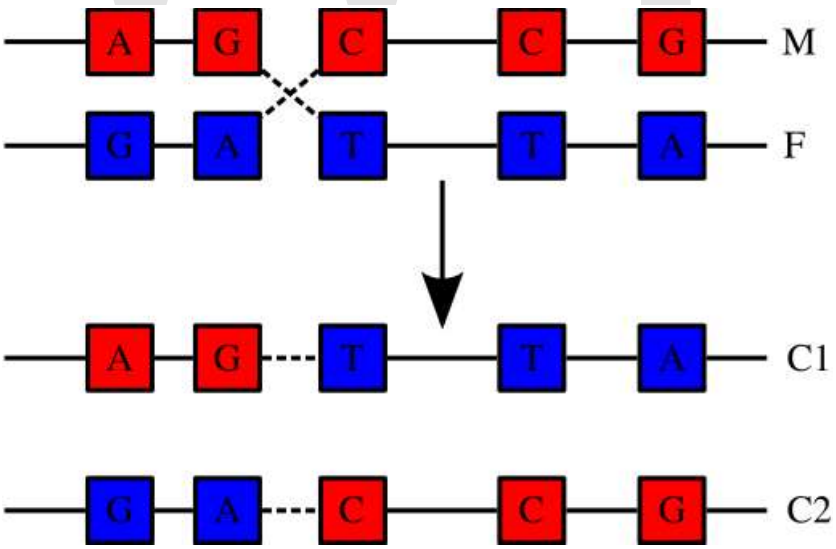
Transcription is carried out by a DNA-dependent RNA polymerase that copies the sequence of a DNA strand into RNA. To begin transcribing a gene, the RNA polymerase binds to a sequence of DNA called a promoter and separates the DNA strands. It then copies the gene sequence into a messenger RNA transcript until it reaches a region of DNA called the terminator, where it halts and detaches from the DNA. As with human DNA-dependent DNA polymerases, RNA polymerase II, the enzyme that transcribes most of the genes in the human genome, operates as part of a large protein complex with multiple regulatory and accessory subunits.

Genetic recombination





Structure of the Holliday junction intermediate in genetic recombination. The four separate DNA strands are coloured red, blue, green and yellow.



Recombination involves the breakage and rejoining of two chromosomes (M and F) to produce two re-arranged chromosomes (C1 and C2).

A DNA helix usually does not interact with other segments of DNA, and in human cells the different chromosomes even occupy separate areas in the nucleus called "chromosome territories". This physical separation of different chromosomes is important for the ability of DNA to function as a stable repository for information, as one of the few times chromosomes interact is during chromosomal crossover when they recombine. Chromosomal crossover is when two DNA helices break, swap a section and then rejoin.

Recombination allows chromosomes to exchange genetic information and produces new combinations of genes, which increases the efficiency of natural selection and can be important in the rapid evolution of new proteins. Genetic recombination can also be involved in DNA repair, particularly in the cell's response to double-strand breaks.

The most common form of chromosomal crossover is homologous recombination, where the two chromosomes involved share very similar sequences. Non-homologous recombination can be damaging to cells, as it can produce chromosomal translocations and genetic abnormalities. The recombination reaction is catalyzed by enzymes known as recombinases, such as RAD51. The first step in recombination is a double-stranded break either caused by an endonuclease or damage to the DNA. A series of steps catalyzed in part by the recombinase then leads to joining of the two helices by at least one Holliday junction, in which a segment of a single strand in each helix is annealed to the complementary strand in the other helix. The Holliday junction is a tetrahedral junction structure that can be moved along the pair of chromosomes, swapping one strand for another. The recombination reaction is then halted by cleavage of the junction and re-ligation of the released DNA.

Evolution

DNA contains the genetic information that allows all modern living things to function, grow and reproduce. However, it is unclear how long in the 4-billion-year history of life DNA has performed this function, as it has been proposed that the earliest forms of life may have used RNA as their genetic material. RNA may have acted as the central part of early cell metabolism as it can both transmit genetic information and carry out catalysis as part of ribozymes. This ancient RNA world where nucleic acid would have been used for both catalysis and genetics may have influenced the evolution of the current genetic code based on four nucleotide bases. This would occur, since the number of different bases in such an organism is a trade-off between a small number of bases increasing replication accuracy and a large number of bases increasing the catalytic efficiency of ribozymes.

However, there is no direct evidence of ancient genetic systems, as recovery of DNA from most fossils is impossible. This is because DNA will survive in the environment for less than one million years and slowly degrades into short fragments in solution. Claims for older DNA have been made, most notably a report of the isolation of a viable bacterium from a salt crystal 250 million years old, but these claims are controversial.

Uses in technology

Genetic engineering

Methods have been developed to purify DNA from organisms, such as phenol-chloroform extraction, and to manipulate it in the laboratory, such as restriction digests and the polymerase chain reaction. Modern biology and biochemistry make intensive use of these techniques in recombinant DNA technology. Recombinant DNA is a man-made DNA sequence that has been assembled from other DNA sequences. They can be transformed into organisms in the form of plasmids or in the appropriate format, by using a viral vector. The genetically modified organisms produced can be used to produce products such as recombinant proteins, used in medical research, or be grown in agriculture.

Forensics

Forensic scientists can use DNA in blood, semen, skin, saliva or hair found at a crime scene to identify a matching DNA of an individual, such as a perpetrator. This process is formally termed DNA profiling, but may also be called "genetic fingerprinting". In DNA profiling, the lengths of variable sections of repetitive DNA, such as short tandem repeats and minisatellites, are compared between people. This method is usually an extremely reliable technique for identifying a matching DNA. However, identification can be complicated if the scene is contaminated with DNA from several people. DNA profiling was developed in 1984 by British geneticist Sir Alec Jeffreys, and first used in forensic science to convict Colin Pitchfork in the 1988 Enderby murders case.

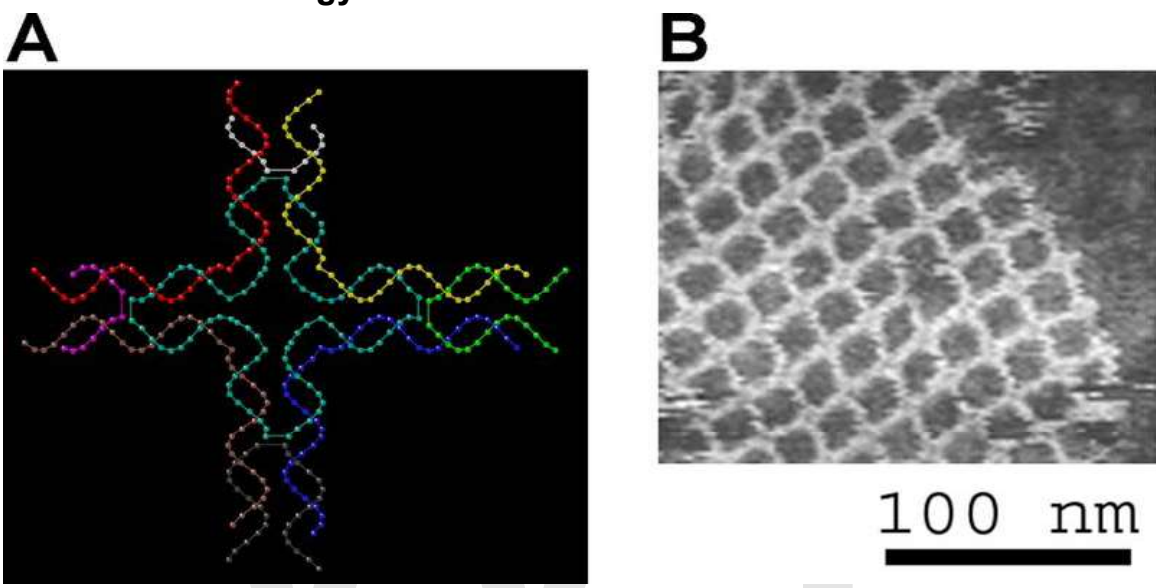
People convicted of certain types of crimes may be required to provide a sample of DNA for a database. This has helped investigators solve old cases where only a DNA sample was obtained from the scene. DNA profiling can also be used to identify victims of mass casualty incidents. On the other hand, many convicted people have been released from prison on the basis of DNA techniques, which were not available when a crime had originally been committed.

Bioinformatics

Bioinformatics involves the manipulation, searching, and data mining of biological data, and this includes DNA sequence data. The development of techniques to store and search DNA sequences have led to widely applied advances in computer science, especially string searching algorithms, machine learning and database theory. String searching or matching algorithms, which find an occurrence of a sequence of letters inside a larger sequence of letters, were developed to search for specific sequences of nucleotides. The DNA sequenced may be aligned with other DNA sequences to identify homologous sequences and locate the specific mutations that make them distinct. These techniques, especially multiple sequence alignment, are used in studying phylogenetic relationships and protein function. Data sets representing entire genomes' worth of DNA sequences, such as those produced by the Human Genome Project, are difficult to use without the

annotations that identify the locations of genes and regulatory elements on each chromosome. Regions of DNA sequence that have the characteristic patterns associated with protein- or RNA-coding genes can be identified by gene finding algorithms, which allow researchers to predict the presence of particular gene products and their possible functions in an organism even before they have been isolated experimentally. Entire genomes may also be compared which can shed light on the evolutionary history of particular organism and permit the examination of complex evolutionary events.

DNA nanotechnology



The DNA structure at left (schematic shown) will self-assemble into the structure visualized by atomic force microscopy at right. DNA nanotechnology is the field that seeks to design nanoscale structures using the molecular recognition properties of DNA molecules.

DNA nanotechnology uses the unique molecular recognition properties of DNA and other nucleic acids to create self-assembling branched DNA complexes with useful properties. DNA is thus used as a structural material rather than as a carrier of biological information. This has led to the creation of two-dimensional periodic lattices (both tile-based as well as using the "DNA origami" method) as well as three-dimensional structures in the shapes of polyhedra. Nanomechanical devices and algorithmic self-assembly have also been demonstrated, and these DNA structures have been used to template the arrangement of other molecules such as gold nanoparticles and streptavidin proteins.

History and anthropology

Because DNA collects mutations over time, which are then inherited, it contains historical information, and, by comparing DNA sequences, geneticists can infer the evolutionary history of organisms, their phylogeny. This field of phylogenetics is a powerful tool in evolutionary biology. If DNA sequences within a species are compared,

population geneticists can learn the history of particular populations. This can be used in studies ranging from ecological genetics to anthropology; For example, DNA evidence is being used to try to identify the Ten Lost Tribes of Israel.

DNA has also been used to look at modern family relationships, such as establishing family relationships between the descendants of Sally Hemings and Thomas Jefferson. This usage is closely related to the use of DNA in criminal investigations detailed above. Indeed, some criminal investigations have been solved when DNA from crime scenes has matched relatives of the guilty individual.

History of DNA research

James D. Watson and Francis Crick (right), co-originators of the double-helix model, with Maclyn McCarty (left).

DNA was first isolated by the Swiss physician Friedrich Miescher who, in 1869, discovered a microscopic substance in the pus of discarded surgical bandages. As it resided in the nuclei of cells, he called it "nuclein". In 1919, Phoebus Levene identified the base, sugar and phosphate nucleotide unit. Levene suggested that DNA consisted of a string of nucleotide units linked together through the phosphate groups. However, Levene

thought the chain was short and the bases repeated in a fixed order. In 1937 William Astbury produced the first X-ray diffraction patterns that showed that DNA had a regular structure.

Raymond Gosling, co-creator of the single X-ray diffraction image

In 1928, Frederick Griffith discovered that traits of the "smooth" form of the *Pneumococcus* could be transferred to the "rough" form of the same bacteria by mixing killed "smooth" bacteria with the live "rough" form. This system provided the first clear suggestion that DNA carries genetic information—the Avery–MacLeod–McCarty experiment—when Oswald Avery, along with coworkers Colin MacLeod and Maclyn McCarty, identified DNA as the transforming principle in 1943. DNA's role in heredity was confirmed in 1952, when Alfred Hershey and Martha Chase in the Hershey–Chase experiment showed that DNA is the genetic material of the T2 phage.

In 1953, James D. Watson and Francis Crick suggested what is now accepted as the first correct double-helix model of DNA structure in the journal *Nature*. Their double-helix, molecular model of DNA was then based on a single X-ray diffraction image (labeled as

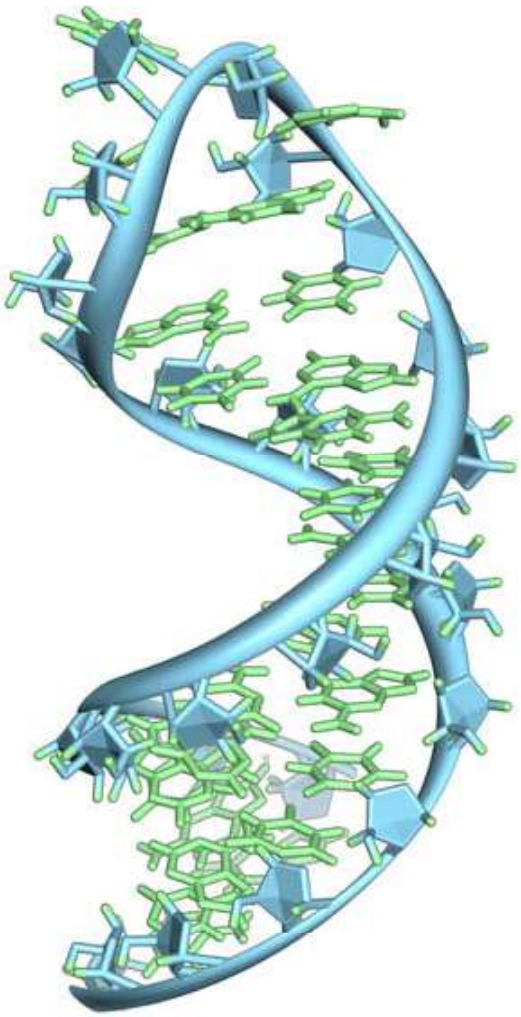
"Photo 51") taken by Rosalind Franklin and Raymond Gosling in May 1952, as well as the information that the DNA bases are paired — also obtained through private communications from Erwin Chargaff in the previous years. Chargaff's rules played a very important role in establishing double-helix configurations for B-DNA as well as A-DNA.

Experimental evidence supporting the Watson and Crick model were published in a series of five articles in the same issue of *Nature*. Of these, Franklin and Gosling's paper was the first publication of their own X-ray diffraction data and original analysis method that partially supported the Watson and Crick model; this issue also contained an article on DNA structure by Maurice Wilkins and two of his colleagues, whose analysis and *in vivo* B-DNA X-ray patterns also supported the presence *in vivo* of the double-helical DNA configurations as proposed by Crick and Watson for their double-helix molecular model of DNA in the previous two pages of *Nature*. In 1962, after Franklin's death, Watson, Crick, and Wilkins jointly received the Nobel Prize in Physiology or Medicine. However, Nobel rules of the time allowed only living recipients, but a vigorous debate continues on who should receive credit for the discovery.

In an influential presentation in 1957, Crick laid out the central dogma of molecular biology, which foretold the relationship between DNA, RNA, and proteins, and articulated the "adaptor hypothesis". Final confirmation of the replication mechanism that was implied by the double-helical structure followed in 1958 through the Meselson–Stahl experiment. Further work by Crick and coworkers showed that the genetic code was based on non-overlapping triplets of bases, called codons, allowing Har Gobind Khorana, Robert W. Holley and Marshall Warren Nirenberg to decipher the genetic code. These findings represent the birth of molecular biology.

Chapter- 13

RNA



A hairpin loop from a pre-mRNA. Highlighted are the nucleobases (green) and the ribose-phosphate backbone (blue).

Ribonucleic acid (**RNA**) is one of the three major macromolecules (along with DNA and proteins) that are essential for all known forms of life.

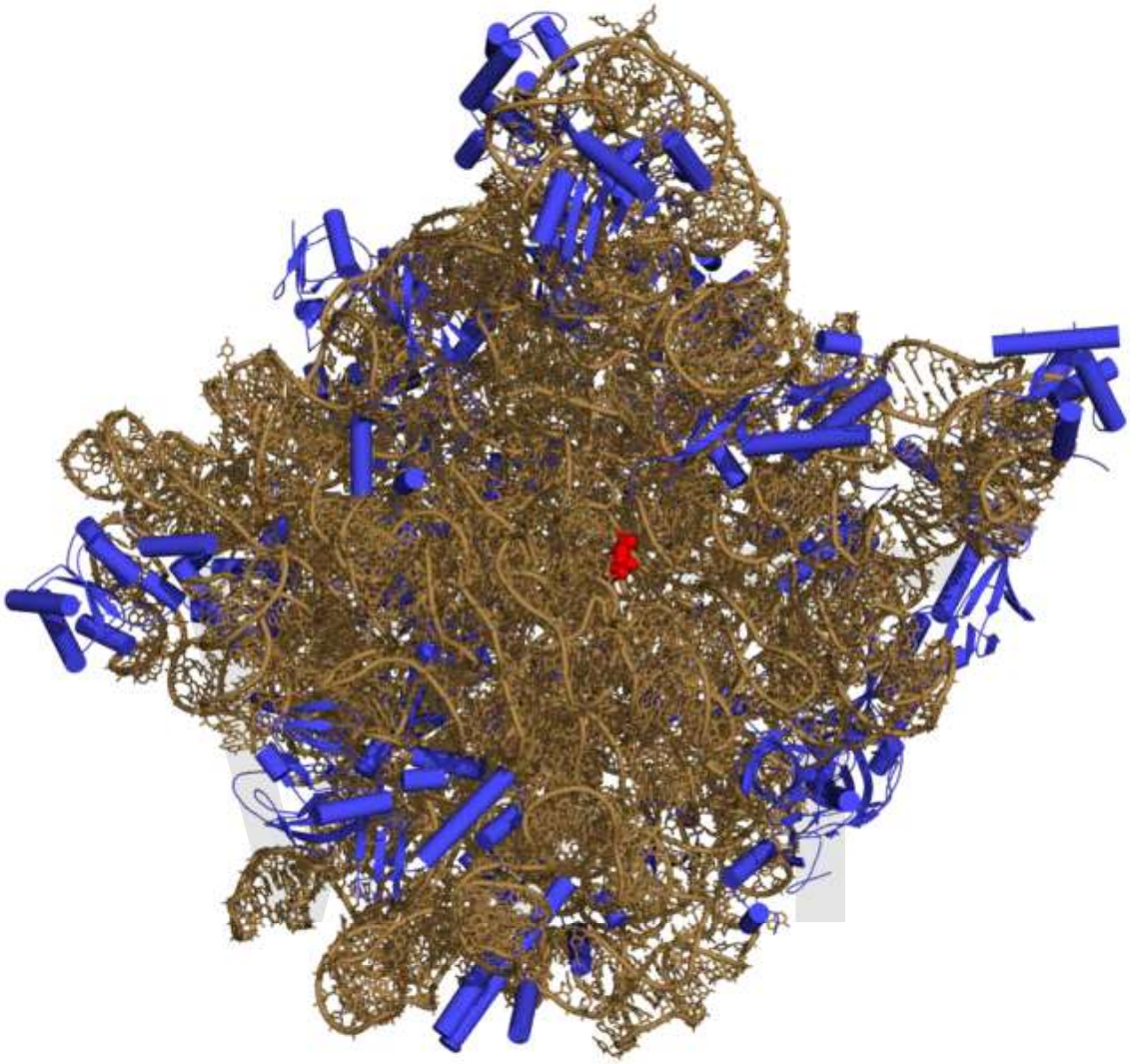
Like DNA, RNA is made up of a long chain of components called nucleotides. Each nucleotide consists of a nucleobase (sometimes called a nitrogenous base), a ribose sugar, and a phosphate group. The sequence of nucleotides allows RNA to encode genetic information. For example, some viruses use RNA instead of DNA as their genetic material, and all organisms use messenger RNA (mRNA) to carry the genetic information that directs the synthesis of proteins.

Like proteins, some RNA molecules play an active role in cells by catalyzing biological reactions, controlling gene expression, or sensing and communicating responses to cellular signals. One of these active processes is protein synthesis, a universal function whereby mRNA molecules direct the assembly of proteins on ribosomes. This process uses transfer RNA (tRNA) molecules to deliver amino acids to the ribosome, where ribosomal RNA (rRNA) links amino acids together to form proteins.

The chemical structure of RNA is very similar to that of DNA, with two differences--(a) RNA contains the sugar ribose while DNA contains the slightly different sugar deoxyribose (a type of ribose that lacks one oxygen atom), and (b) RNA has the nucleobase uracil while DNA contains thymine (uracil and thymine have similar base-pairing properties).

Unlike DNA, most RNA molecules are single-stranded. Single-stranded RNA molecules adopt very complex three-dimensional structures, since they are not restricted to the repetitive double-helical form of double-stranded DNA. RNA is made within living cells by RNA polymerases, enzymes that act to copy a DNA or RNA template into a new RNA strand through processes known as transcription or RNA replication, respectively.

Comparison with DNA



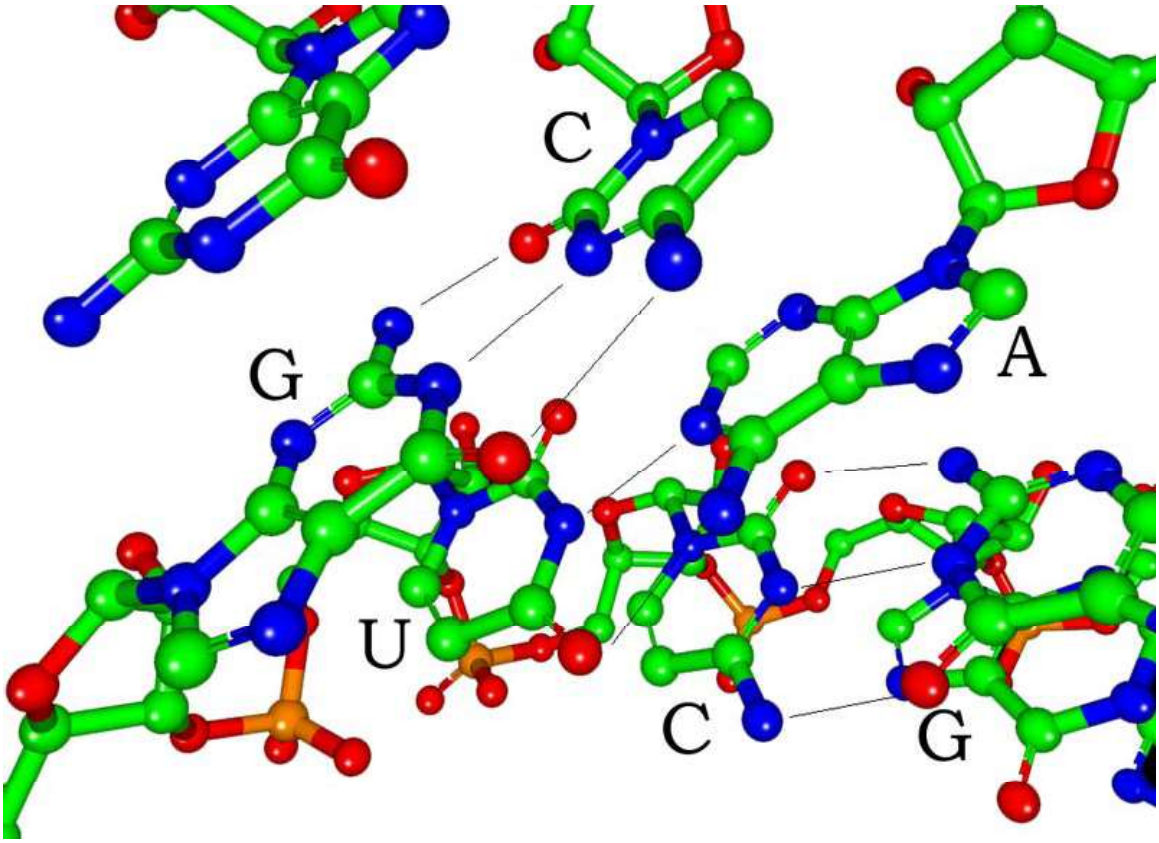
Three-dimensional representation of the 50S ribosomal subunit. RNA is in ochre, protein in blue. The active site is in the middle (red).

RNA and DNA are both nucleic acids, but differ in three main ways. First, unlike DNA, which is, in general, double-stranded, RNA is a single-stranded molecule in many of its biological roles and has a much shorter chain of nucleotides. Second, while DNA contains *deoxyribose*, RNA contains *ribose* (in deoxyribose there is no hydroxyl group attached to the pentose ring in the 2' position). These hydroxyl groups make RNA less stable than DNA because it is more prone to hydrolysis. Third, the complementary base to adenine is not thymine, as it is in DNA, but rather uracil, which is an unmethylated form of thymine.

Like DNA, most biologically active RNAs, including mRNA, tRNA, rRNA, snRNAs, and other non-coding RNAs, contain self-complementary sequences that allow parts of

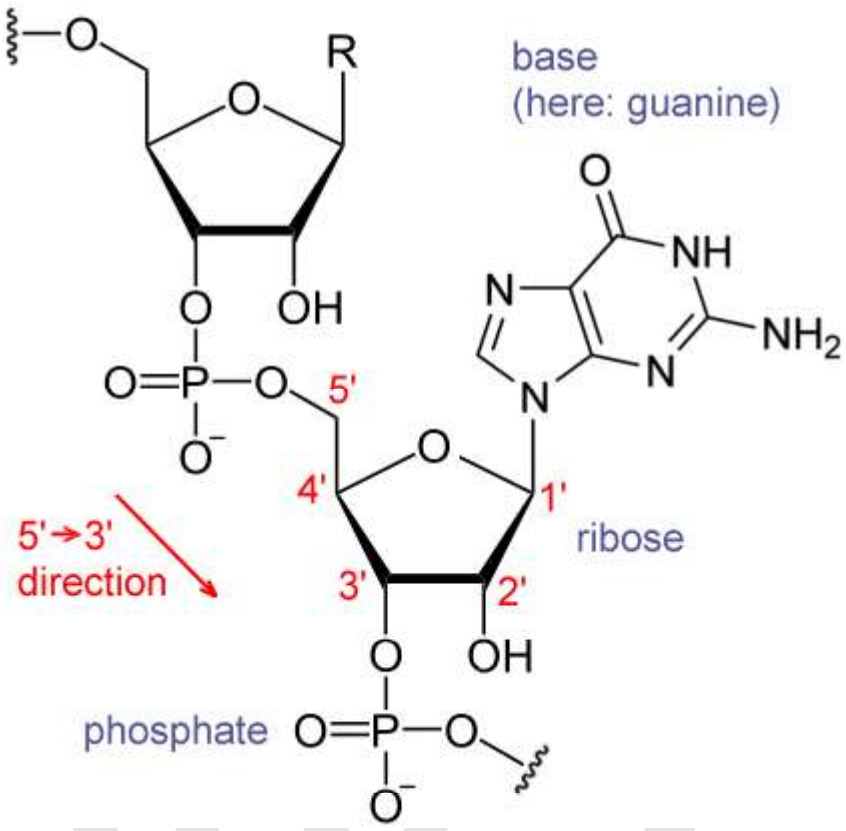
the RNA to fold and pair with itself to form double helices. Structural analysis of these RNAs has revealed that they are highly structured. Unlike DNA, their structures do not consist of long double helices but rather collections of short helices packed together into structures akin to proteins. In this fashion, RNAs can achieve chemical catalysis, like enzymes. For instance, determination of the structure of the ribosome—an enzyme that catalyzes peptide bond formation—revealed that its active site is composed entirely of RNA.

Structure



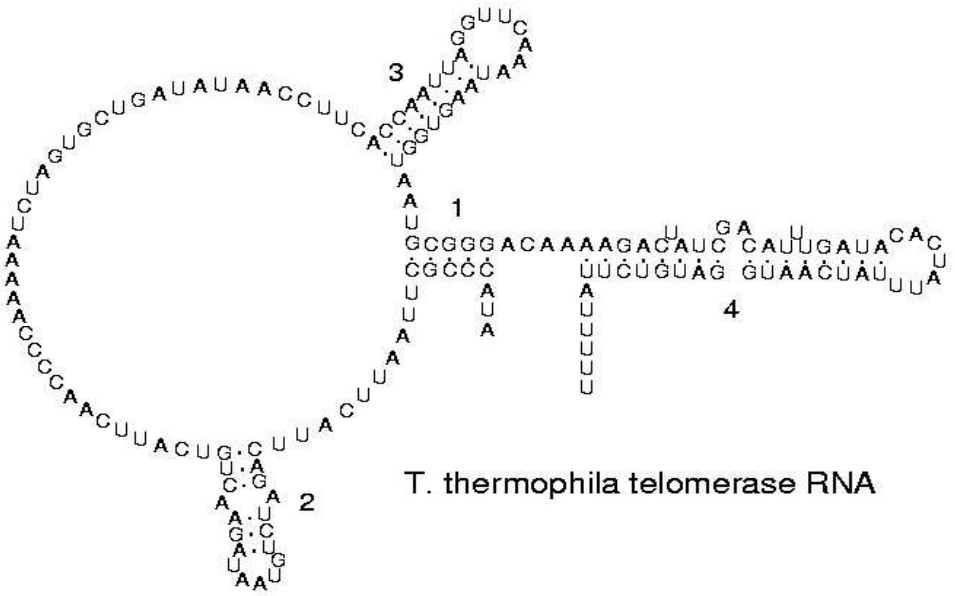
Watson-Crick base pairs in a siRNA (hydrogen atoms are not shown)

Each nucleotide in RNA contains a ribose sugar, with carbons numbered 1' through 5'. A base is attached to the 1' position, in general, adenine (A), cytosine (C), guanine (G), or uracil (U). Adenine and guanine are purines, cytosine, and uracil are pyrimidines. A phosphate group is attached to the 3' position of one ribose and the 5' position of the next. The phosphate groups have a negative charge each at physiological pH, making RNA a charged molecule (polyanion). The bases may form hydrogen bonds between cytosine and guanine, between adenine and uracil and between guanine and uracil. However, other interactions are possible, such as a group of adenine bases binding to each other in a bulge, or the GNRA tetraloop that has a guanine–adenine base-pair.



Chemical structure of RNA

An important structural feature of RNA that distinguishes it from DNA is the presence of a hydroxyl group at the 2' position of the ribose sugar. The presence of this functional group causes the helix to adopt the A-form geometry rather than the B-form most commonly observed in DNA. This results in a very deep and narrow major groove and a shallow and wide minor groove. A second consequence of the presence of the 2'-hydroxyl group is that in conformationally flexible regions of an RNA molecule (that is, not involved in formation of a double helix), it can chemically attack the adjacent phosphodiester bond to cleave the backbone.



Secondary structure of a telomerase RNA

RNA is transcribed with only four bases (adenine, cytosine, guanine and uracil), but these bases and attached sugars can be modified in numerous ways as the RNAs mature. Pseudouridine (Ψ), in which the linkage between uracil and ribose is changed from a C–N bond to a C–C bond, and ribothymidine (T) are found in various places (the most notable ones being in the TΨC loop of tRNA). Another notable modified base is hypoxanthine, a deaminated adenine base whose nucleoside is called inosine (I). Inosine plays a key role in the wobble hypothesis of the genetic code.

There are nearly 100 other naturally occurring modified nucleosides, of which pseudouridine and nucleosides with 2'-O-methylribose are the most common. The specific roles of many of these modifications in RNA are not fully understood. However, it is notable that, in ribosomal RNA, many of the post-transcriptional modifications occur in highly functional regions, such as the peptidyl transferase center and the subunit interface, implying that they are important for normal function.

The functional form of single stranded RNA molecules, just like proteins, frequently requires a specific tertiary structure. The scaffold for this structure is provided by secondary structural elements that are hydrogen bonds within the molecule. This leads to several recognizable "domains" of secondary structure like hairpin loops, bulges, and internal loops. Since RNA is charged, metal ions such as Mg^{2+} are needed to stabilise many secondary and tertiary structures.

Synthesis

Synthesis of RNA is usually catalyzed by an enzyme—RNA polymerase—using DNA as a template, a process known as transcription. Initiation of transcription begins with the

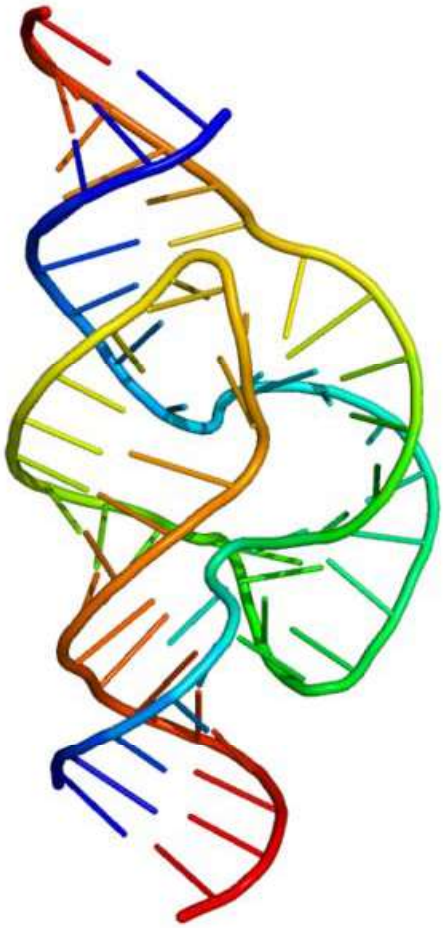
binding of the enzyme to a promoter sequence in the DNA (usually found "upstream" of a gene). The DNA double helix is unwound by the helicase activity of the enzyme. The enzyme then progresses along the template strand in the 3’ to 5’ direction, synthesizing a complementary RNA molecule with elongation occurring in the 5’ to 3’ direction. The DNA sequence also dictates where termination of RNA synthesis will occur.

RNAs are often modified by enzymes after transcription. For example, a poly(A) tail and a 5' cap are added to eukaryotic pre-mRNA and introns are removed by the spliceosome.

There are also a number of RNA-dependent RNA polymerases that use RNA as their template for synthesis of a new strand of RNA. For instance, a number of RNA viruses (such as poliovirus) use this type of enzyme to replicate their genetic material. Also, RNA-dependent RNA polymerase is part of the RNA interference pathway in many organisms.

Types of RNA

Overview



Structure of a hammerhead ribozyme, a ribozyme that cuts RNA

Messenger RNA (mRNA) is the RNA that carries information from DNA to the ribosome, the sites of protein synthesis (translation) in the cell. The coding sequence of the mRNA determines the amino acid sequence in the protein that is produced. Many RNAs do not code for protein however (about 97% of the transcriptional output is non-protein-coding in eukaryotes).

These so-called non-coding RNAs ("ncRNA") can be encoded by their own genes (RNA genes), but can also derive from mRNA introns. The most prominent examples of non-coding RNAs are transfer RNA (tRNA) and ribosomal RNA (rRNA), both of which are involved in the process of translation. There are also non-coding RNAs involved in gene regulation, RNA processing and other roles. Certain RNAs are able to catalyse chemical reactions such as cutting and ligating other RNA molecules, and the catalysis of peptide bond formation in the ribosome; these are known as ribozymes.

In translation

Messenger RNA (mRNA) carries information about a protein sequence to the ribosomes, the protein synthesis factories in the cell. It is coded so that every three nucleotides (a codon) correspond to one amino acid. In eukaryotic cells, once precursor mRNA (pre-mRNA) has been transcribed from DNA, it is processed to mature mRNA. This removes its introns—non-coding sections of the pre-mRNA. The mRNA is then exported from the nucleus to the cytoplasm, where it is bound to ribosomes and translated into its corresponding protein form with the help of tRNA. In prokaryotic cells, which do not have nucleus and cytoplasm compartments, mRNA can bind to ribosomes while it is being transcribed from DNA. After a certain amount of time the message degrades into its component nucleotides with the assistance of ribonucleases.

Transfer RNA (tRNA) is a small RNA chain of about 80 nucleotides that transfers a specific amino acid to a growing polypeptide chain at the ribosomal site of protein synthesis during translation. It has sites for amino acid attachment and an anticodon region for codon recognition that binds to a specific sequence on the messenger RNA chain through hydrogen bonding.

Ribosomal RNA (rRNA) is the catalytic component of the ribosomes. Eukaryotic ribosomes contain four different rRNA molecules: 18S, 5.8S, 28S and 5S rRNA. Three of the rRNA molecules are synthesized in the nucleolus, and one is synthesized elsewhere. In the cytoplasm, ribosomal RNA and protein combine to form a nucleoprotein called a ribosome. The ribosome binds mRNA and carries out protein synthesis. Several ribosomes may be attached to a single mRNA at any time. rRNA is extremely abundant and makes up 80% of the 10 mg/ml RNA found in a typical eukaryotic cytoplasm.

Transfer-messenger RNA (tmRNA) is found in many bacteria and plastids. It tags proteins encoded by mRNAs that lack stop codons for degradation and prevents the ribosome from stalling.

Regulatory RNAs

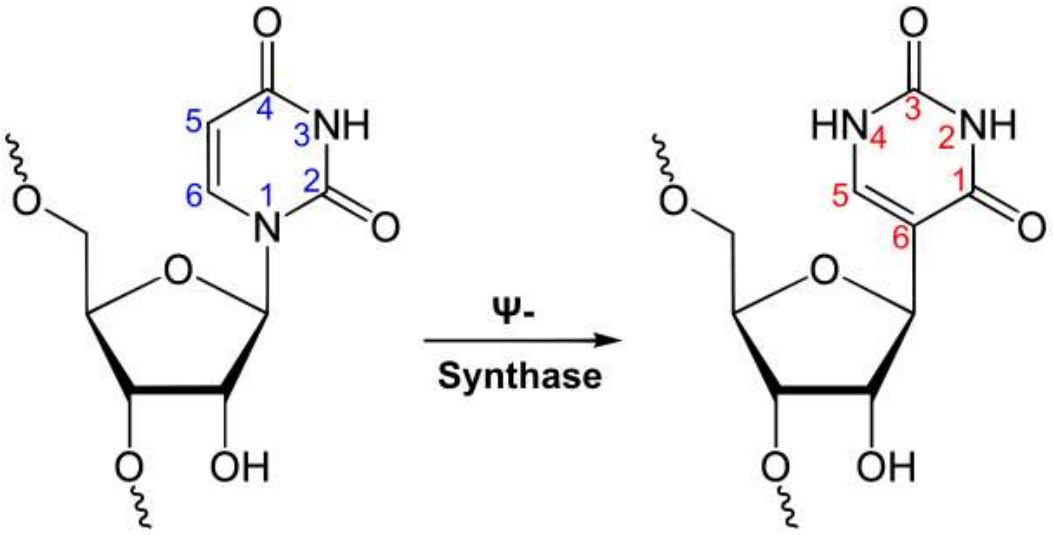
Several types of RNA can downregulate gene expression by being complementary to a part of an mRNA or a gene's DNA. MicroRNAs (miRNA; 21-22 nt) are found in eukaryotes and act through RNA interference (RNAi), where an effector complex of miRNA and enzymes can break down mRNA to which the miRNA is complementary, block the mRNA from being translated, or accelerate its degradation. While small interfering RNAs (siRNA; 20-25 nt) are often produced by breakdown of viral RNA, there are also endogenous sources of siRNAs.

siRNAs act through RNA interference in a fashion similar to miRNAs. Some miRNAs and siRNAs can cause genes they target to be methylated, thereby decreasing or increasing transcription of those genes. Animals have Piwi-interacting RNAs (piRNA; 29-30 nt) which are active in germline cells and are thought to be a defense against transposons and play a role in gametogenesis.

Many prokaryotes have CRISPR RNAs, a regulatory system similar to RNA interference. Antisense RNAs are widespread; most downregulate a gene, but a few are activators of transcription. One way antisense RNA can act is by binding to an mRNA, forming double-stranded RNA that is enzymatically degraded. There are many long noncoding RNAs that regulate genes in eukaryotes, one such RNA is Xist, which coats one X chromosome in female mammals and inactivates it.

An mRNA may contain regulatory elements itself, such as riboswitches, in the 5' untranslated region or 3' untranslated region; these cis-regulatory elements regulate the activity of that mRNA. The untranslated regions can also contain elements that regulate other genes.

In RNA processing



Uridine to pseudouridine is a common RNA modification

Many RNAs are involved in modifying other RNAs. Introns are spliced out of pre-mRNA by spliceosomes, which contain several small nuclear RNAs (snRNA), or the introns can be ribozymes that are spliced by themselves. RNA can also be altered by having its nucleotides modified to other nucleotides than A, C, G and U. In eukaryotes, modifications of RNA nucleotides are generally directed by small nucleolar RNAs (snoRNA; 60-300 nt), found in the nucleolus and cajal bodies. snoRNAs associate with enzymes and guide them to a spot on an RNA by basepairing to that RNA. These enzymes then perform the nucleotide modification. rRNAs and tRNAs are extensively modified, but snRNAs and mRNAs can also be the target of base modification.

RNA genomes

Like DNA, RNA can carry genetic information. RNA viruses have genomes composed of RNA, and a variety of proteins encoded by that genome. The viral genome is replicated by some of those proteins, while other proteins protect the genome as the virus particle moves to a new host cell. Viroids are another group of pathogens, but they consist only of RNA, do not encode any protein and are replicated by a host plant cell's polymerase.

In reverse transcription

Reverse transcribing viruses replicate their genomes by reverse transcribing DNA copies from their RNA; these DNA copies are then transcribed to new RNA. Retrotransposons also spread by copying DNA and RNA from one another, and telomerase contains an RNA that is used as template for building the ends of eukaryotic chromosomes.

Double-stranded RNA

Double-stranded RNA (dsRNA) is RNA with two complementary strands, similar to the DNA found in all cells. dsRNA forms the genetic material of some viruses (double-stranded RNA viruses). Double-stranded RNA such as viral RNA or siRNA can trigger RNA interference in eukaryotes, as well as interferon response in vertebrates.

Key discoveries in RNA biology

Research on RNA has led to many important biological discoveries and numerous Nobel Prizes. Nucleic acids were discovered in 1868 by Friedrich Miescher, who called the material 'nuclein' since it was found in the nucleus. It was later discovered that prokaryotic cells, which do not have a nucleus, also contain nucleic acids. The role of RNA in protein synthesis was suspected already in 1939. Severo Ochoa won the 1959 Nobel Prize in Medicine (shared with Arthur Kornberg) after he discovered an enzyme that can synthesize RNA in the laboratory. Ironically, the enzyme discovered by Ochoa (polynucleotide phosphorylase) was later shown to be responsible for RNA degradation, not RNA synthesis.

The sequence of the 77 nucleotides of a yeast tRNA was found by Robert W. Holley in 1965, winning Holley the 1968 Nobel Prize in Medicine (shared with Har Gobind

Khorana and Marshall Nirenberg). In 1967, Carl Woese hypothesized that RNA might be catalytic and suggested that the earliest forms of life (self-replicating molecules) could have relied on RNA both to carry genetic information and to catalyze biochemical reactions—an RNA world.

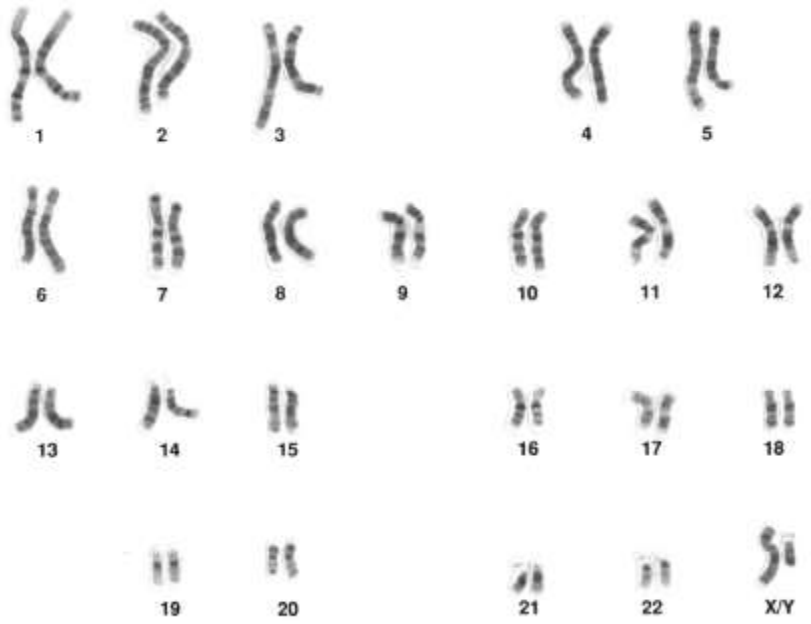
During the early 1970s, retroviruses and reverse transcriptase were discovered, showing for the first time that enzymes could copy RNA into DNA (the opposite of the usual route for transmission of genetic information). For this work, David Baltimore, Renato Dulbecco and Howard Temin were awarded a Nobel Prize in 1975. In 1976, Walter Fiers and his team determined the first complete nucleotide sequence of an RNA virus genome, that of bacteriophage MS2.

In 1977, introns and RNA splicing were discovered in both mammalian viruses and in cellular genes, resulting in a 1993 Nobel to Philip Sharp and Richard Roberts. Catalytic RNA molecules (ribozymes) were discovered in the early 1980s, leading to a 1989 Nobel award to Thomas Cech and Sidney Altman. In 1990 it was found in petunia that introduced genes can silence similar genes of the plant's own, now known to be a result of RNA interference.

At about the same time, 22 nt long RNAs, now called microRNAs, were found to have a role in the development of *C. elegans*. Studies on RNA interference gleaned a Nobel Prize for Andrew Fire and Craig Mello in 2006, and another Nobel was awarded for studies on transcription of RNA to Roger Kornberg in the same year. The discovery of gene regulatory RNAs has led to attempts to develop drugs made of RNA, such as siRNA, to silence genes.

Chapter- 14

Genome



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

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

Origin of Term

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

Overview

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

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

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

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

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

Types

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

include information stored on this auxiliary material, which is carried in plasmids. In such circumstances then, "genome" describes all of the genes and information on non-coding DNA that have the potential to be present.

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

Genomes and genetic variation

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

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

Sequencing and mapping

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

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

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

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

Comparison of different genome sizes

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

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

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

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

Genome evolution

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

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

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

Chapter- 15

Heredity

Heredity is the passing of traits to offspring (from its parent or ancestors). This is the process by which an offspring cell or organism acquires or becomes predisposed to the characteristics of its parent cell or organism. Through heredity, variations exhibited by individuals can accumulate and cause some species to evolve. The study of heredity in biology is called genetics, which includes the field of epigenetics.

History

The ancients had a variety of ideas about heredity: Theophrastus proposed that male flowers caused female flowers to ripen; Hippocrates speculated that "seeds" were produced by various body parts and transmitted to offspring at the time of conception, and Aristotle thought that male and female semen mixed at conception. Aeschylus, in 458 BC, proposed the male as the parent, with the female as a "nurse for the young life sown within her."

Various hereditary mechanisms were envisaged without being properly tested or quantified. These included blending inheritance and the inheritance of acquired traits. Nevertheless, people were able to develop domestic breeds of animals as well as crops through artificial selection. The inheritance of acquired traits also formed a part of early Lamarckian ideas on evolution.

In the 9th century AD, the Afro-Arab writer Al-Jahiz considered the effects of the environment on the likelihood of an animal to survive, and first described the struggle for existence. His ideas on the struggle for existence in the *Book of Animals* have been summarized as follows:

Animals engage in a struggle for existence; for resources, to avoid being eaten and to breed. Environmental factors influence organisms to develop new characteristics to ensure survival, thus transforming into new species. Animals that survive to breed can pass on their successful characteristics to offspring.

In 1000 AD, the Arab physician, Abu al-Qasim al-Zahrawi (known as Albucasis in the West), wrote the first clear description of haemophilia, a hereditary genetic disorder, in his *Al-Tasrif*. In this work, he wrote of an Andalusian family whose males died of bleeding after minor injuries.

During the 18th century, Dutch microscopist Antonie van Leeuwenhoek (1632–1723) discovered "animalcules" in the sperm of humans and other animals. Some scientists speculated they saw a "little man" (homunculus) inside each sperm. These scientists formed a school of thought known as the "spermists." They contended the only contributions of the female to the next generation were the womb in which the homunculus grew, and prenatal influences of the womb. An opposing school of thought, the ovists, believed that the future human was in the egg, and that sperm merely stimulated the growth of the egg. Ovists thought women carried eggs containing boy and girl children, and that the gender of the offspring was determined well before conception.

Types of heredity

Dominant and recessive

An allele is said to be dominant if it is always expressed in the appearance of an organism (phenotype). For example, in peas the allele for green pods, G, is dominant to that for yellow pods, g. Since the allele for green pods is dominant, pea plants with the pair of alleles GG (homozygote) or Gg (heterozygote) will have green pods. The allele for yellow pods is recessive. The effects of this allele are only seen when it is present in both chromosomes, gg (homozygote).

The description of a mode of biological inheritance consists of three main categories:

1. Number of involved loci

- Monogenetic (also called "simple") – one locus
- Oligogenetic – few loci
- Polygenetic – many loci

2. Involved chromosomes

- Autosomal – loci are not situated on a sex chromosome
- Gonosomal – loci are situated on a sex chromosome
  - X-chromosomal – loci are situated on the X chromosome (the more common case)
  - Y-chromosomal – loci are situated on the Y chromosome
- Mitochondrial – loci are situated on the mitochondrial DNA

3. Correlation genotype–phenotype

- Dominant

- Intermediate (also called "codominant")
- Recessive

These three categories are part of every exact description of a mode of inheritance in the above order. Additionally, more specifications may be added as follows:

4. Coincidental and environmental interactions

- Penetrance
  - Complete
  - Incomplete (percentual number)
- Expressivity
  - Invariable
  - Variable
- Heritability (in polygenetic and sometimes also in oligogenetic modes of inheritance)
- Maternal or paternal imprinting phenomena

5. Sex-linked interactions

- Sex-linked inheritance (gonosomal loci)
- Sex-limited phenotype expression (e.g., cryptorchism)
- Inheritance through the maternal line (in case of mitochondrial DNA loci)
- Inheritance through the paternal line (in case of Y-chromosomal loci)

6. Locus–locus interactions

- Epistasis with other loci (e.g., overdominance)
- Gene coupling with other loci
- Homozygotous lethal factors
- Semi-lethal factors

Determination and description of a mode of inheritance is primarily achieved through statistical analysis of pedigree data. In case the involved loci are known, methods of molecular genetics can also be employed.

Charles Darwin: theory of evolution

When Charles Darwin proposed his theory of evolution in 1859, one of its major problems was the lack of an underlying mechanism for heredity. Darwin believed in a mix of blending inheritance and the inheritance of acquired traits (pangenesis). Blending inheritance would lead to uniformity across populations in only a few generations and thus would remove variation from a population on which natural selection could act. This led to Darwin adopting some Lamarckian ideas in later editions of *On the Origin of Species* and his later biological works. Darwin's primary approach to heredity was to outline how it appeared to work (noticing that traits could be inherited which were not

expressed explicitly in the parent at the time of reproduction, that certain traits could be sex-linked, etc.) rather than suggesting mechanisms.

Darwin's initial model of heredity was adopted by, and then heavily modified by, his cousin Francis Galton, who laid the framework for the biometric school of heredity. Galton rejected the aspects of Darwin's pangenesis model which relied on acquired traits.

The inheritance of acquired traits was shown to have little basis in the 1880s when August Weismann cut the tails off many generations of mice and found that their offspring continued to develop tails.

Gregor Mendel: father of modern genetics

The idea of particulate inheritance of genes can be attributed to the Moravian monk Gregor Mendel who published his work on pea plants in 1865. However, his work was not widely known and was rediscovered in 1901. It was initially assumed the Mendelian inheritance only accounted for large (qualitative) differences, such as those seen by Mendel in his pea plants—and the idea of additive effect of (quantitative) genes was not realised until R.A. Fisher's (1918) paper, "The Correlation Between Relatives on the Supposition of Mendelian Inheritance."

Modern development of genetics and heredity

In the 1930s, work by Fisher and others resulted in a combination of Mendelian and biometric schools into the modern evolutionary synthesis. The modern synthesis bridged the gap between experimental geneticists and naturalists; and between both and palaeontologists, stating that:

1. All evolutionary phenomena can be explained in a way consistent with known genetic mechanisms and the observational evidence of naturalists.
2. Evolution is gradual: small genetic changes, recombination ordered by natural selection. Discontinuities amongst species (or other taxa) are explained as originating gradually through geographical separation and extinction (not saltation).
3. Selection is overwhelmingly the main mechanism of change; even slight advantages are important when continued. The object of selection is the phenotype in its surrounding environment. The role of genetic drift is equivocal; though strongly supported initially by Dobzhansky, it was downgraded later as results from ecological genetics were obtained.
4. The primacy of population thinking: the genetic diversity carried in natural populations is a key factor in evolution. The strength of natural selection in the wild was greater than expected; the effect of ecological factors such as niche occupation and the significance of barriers to gene flow are all important.
5. In palaeontology, the ability to explain historical observations by extrapolation from micro to macro-evolution is proposed. Historical contingency means

explanations at different levels may exist. Gradualism does not mean constant rate of change.

The idea that speciation occurs after populations are reproductively isolated has been much debated. In plants, polyploidy must be included in any view of speciation. Formulations such as 'evolution consists primarily of changes in the frequencies of alleles between one generation and another' were proposed rather later. The traditional view is that developmental biology ('evo-devo') played little part in the synthesis, but an account of Gavin de Beer's work by Stephen Jay Gould suggests he may be an exception.

Almost all aspects of the synthesis have been challenged at times, with varying degrees of success. There is no doubt, however, that the synthesis was a great landmark in evolutionary biology. It cleared up many confusions, and was directly responsible for stimulating a great deal of research in the post-World War II era.

Trofim Lysenko however caused a backlash of what is now called Lysenkoism in the Soviet Union when he emphasised Lamarckian ideas on the inheritance of acquired traits. This movement affected agricultural research and led to food shortages in the 1960s and seriously affected the USSR.