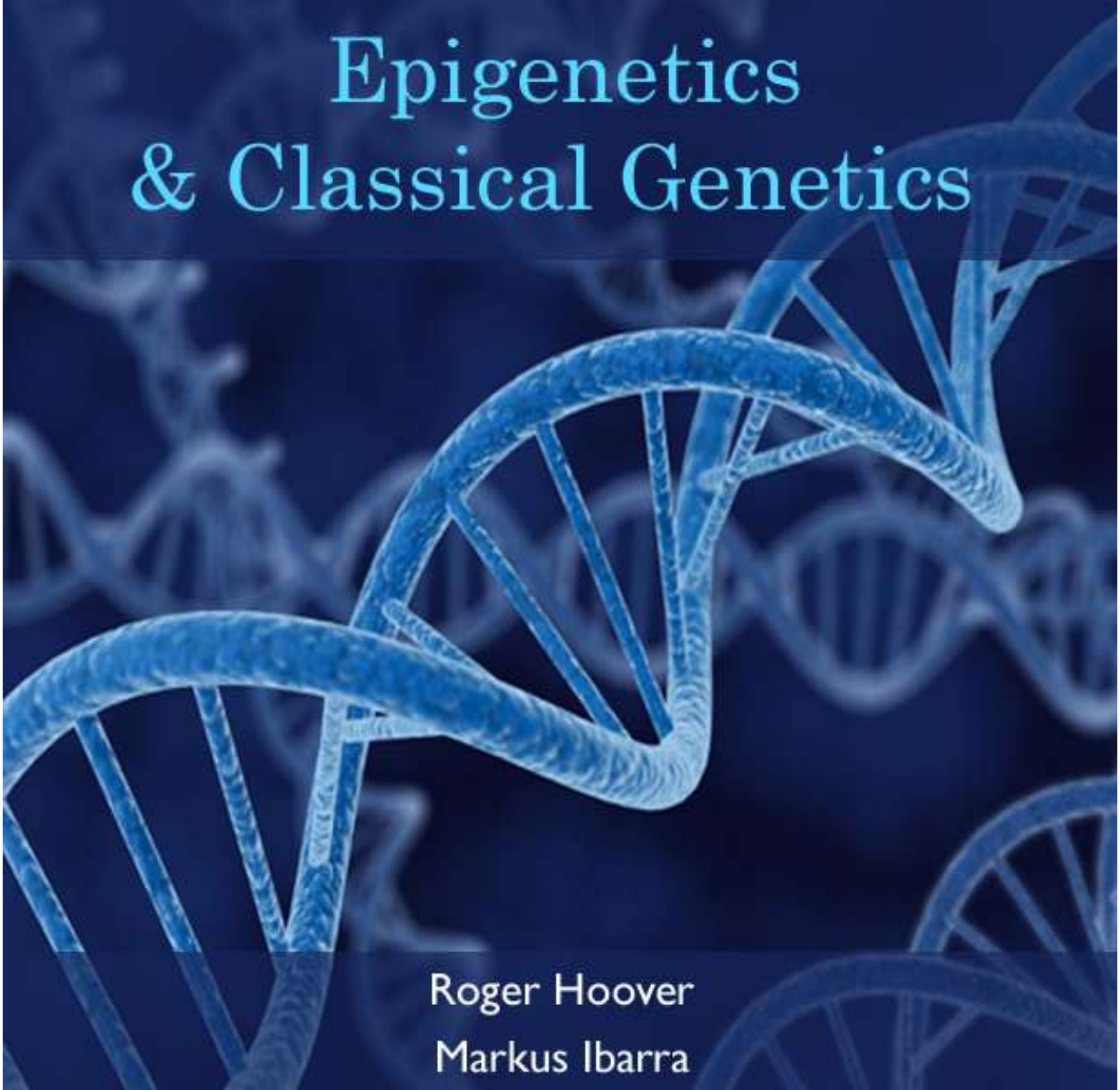
Epigenetics & Classical Genetics



Roger Hoover
Markus Ibarra

First Edition, 2012

ISBN 978-81-323-0703-7

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

Table of Contents

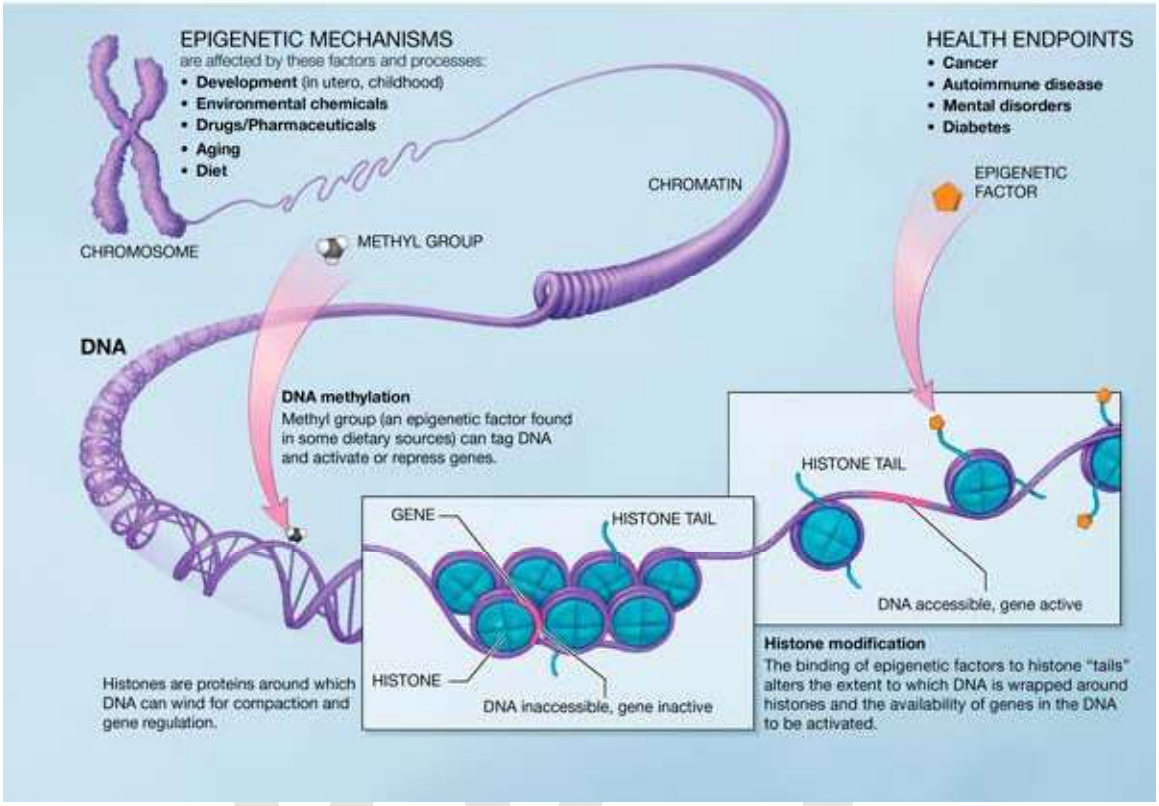
Chapter 1

Epigenetics

In biology, and specifically genetics, **epigenetics** is the study of heritable changes in phenotype (appearance) or gene expression caused by mechanisms other than changes in the underlying DNA sequence, hence the name *epi-* (Greek: *επί-* over, above) *-genetics*. These changes may remain through cell divisions for the remainder of the cell's life and may also last for multiple generations. However, there is no change in the underlying DNA sequence of the organism; instead, non-genetic factors cause the organism's genes to behave (or "express themselves") differently.

One example of epigenetic changes in eukaryotic biology is the process of cellular differentiation. During morphogenesis, totipotent stem cells become the various pluripotent cell lines of the embryo which in turn become fully differentiated cells. In other words, a single fertilized egg cell – the zygote – changes into the many cell types including neurons, muscle cells, epithelium, blood vessels etc. as it continues to divide. It does so by activating some genes while inhibiting others.

Etymology and definitions



Epigenetic mechanisms

Epigenetics (as in "epigenetic landscape") was coined by C. H. Waddington in 1942 as a portmanteau of the words *genetics* and *epigenesis*. *Epigenesis* is an old word which has more recently been used to describe the differentiation of cells from their initial totipotent state in embryonic development. When Waddington coined the term the physical nature of genes and their role in heredity was not known; he used it as a conceptual model of how genes might interact with their surroundings to produce a phenotype.

Robin Holliday defined epigenetics as "the study of the mechanisms of temporal and spatial control of gene activity during the development of complex organisms." Thus *epigenetic* can be used to describe anything other than DNA sequence that influences the development of an organism.

The modern usage of the word in scientific discourse is more narrow, referring to heritable traits (over rounds of cell division and sometimes transgenerationally) that do not involve changes to the underlying DNA sequence. The Greek prefix *epi-* in *epigenetics* implies features that are "on top of" or "in addition to" genetics; thus *epigenetic* traits exist on top of or in addition to the traditional molecular basis for inheritance.

The similarity of the word to "genetics" has generated many parallel usages. The "epigenome" is a parallel to the word "genome", and refers to the overall epigenetic state of a cell. The phrase "genetic code" has also been adapted—the "epigenetic code" has been used to describe the set of epigenetic features that create different phenotypes in different cells. Taken to its extreme, the "epigenetic code" could represent the total state of the cell, with the position of each molecule accounted for in an *epigenomic map*, a diagrammatic representation of the gene expression, DNA methylation and histone modification status of a particular genomic region. More typically, the term is used in reference to systematic efforts to measure specific, relevant forms of epigenetic information such as the histone code or DNA methylation patterns.

The psychologist Erik Erikson used the term *epigenetic* in his theory of psychosocial development. That usage, however, is of primarily historical interest.

Molecular basis of epigenetics

The molecular basis of epigenetics is complex. It involves modifications of the activation of certain genes, but not the basic structure of DNA. Additionally, the chromatin proteins associated with DNA may be activated or silenced. This accounts for why the differentiated cells in a multi-cellular organism express only the genes that are necessary for their own activity. Epigenetic changes are preserved when cells divide. Most epigenetic changes only occur within the course of one individual organism's lifetime, but, if a mutation in the DNA has been caused in sperm or egg cell that results in fertilization, then some epigenetic changes are inherited from one generation to the next. This raises the question of whether or not epigenetic changes in an organism can alter the basic structure of its DNA, a form of Lamarckism.

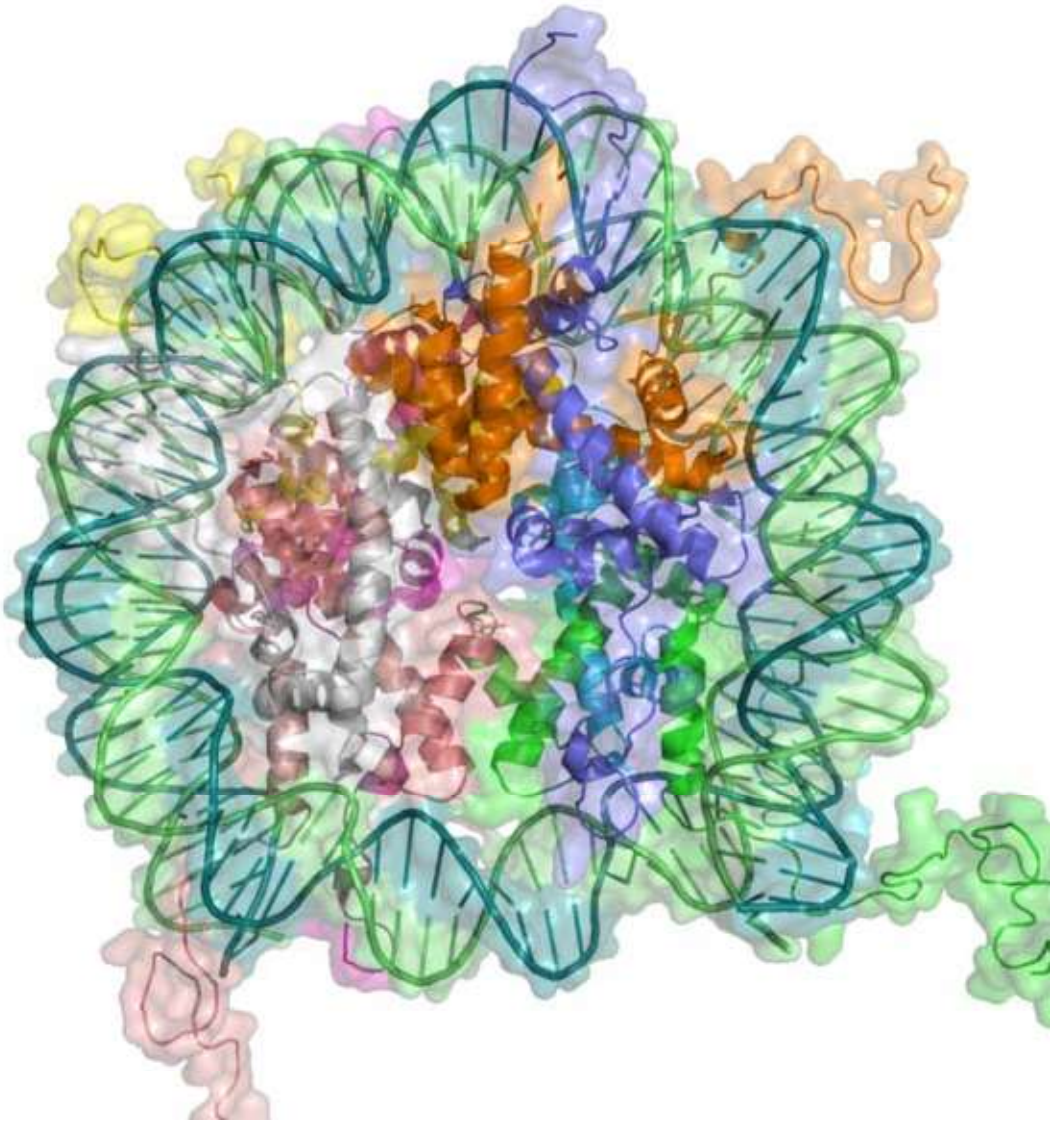
Specific epigenetic processes include paramutation, bookmarking, imprinting, gene silencing, X chromosome inactivation, position effect, reprogramming, transvection, maternal effects, the progress of carcinogenesis, many effects of teratogens, regulation of histone modifications and heterochromatin, and technical limitations affecting parthenogenesis and cloning.

Epigenetic research uses a wide range of molecular biologic techniques to further our understanding of epigenetic phenomena, including chromatin immunoprecipitation (together with its large-scale variants ChIP-on-chip and ChIP-seq), fluorescent in situ hybridization, methylation-sensitive restriction enzymes, DNA adenine methyltransferase identification (DamID) and bisulfite sequencing. Furthermore, the use of bioinformatic methods is playing an increasing role (computational epigenetics).

Mechanisms

Several types of epigenetic inheritance systems may play a role in what has become known as cell memory:

DNA methylation and chromatin remodeling



DNA associates with histone proteins to form chromatin.

Because the phenotype of a cell or individual is affected by which of its genes are transcribed, heritable transcription states can give rise to epigenetic effects. There are several layers of regulation of gene expression. One way that genes are regulated is through the remodeling of chromatin. Chromatin is the complex of DNA and the histone proteins with which it associates. Histone proteins are little spheres that DNA wraps around. If the way that DNA is wrapped around the histones changes, gene expression can change as well. Chromatin remodeling is accomplished through two main mechanisms:

1. The first way is post translational modification of the amino acids that make up histone proteins. Histone proteins are made up of long chains of amino acids. If

the amino acids that are in the chain are changed, the shape of the histone sphere might be modified. DNA is not completely unwound during replication. It is possible, then, that the modified histones may be carried into each new copy of the DNA. Once there, these histones may act as templates, initiating the surrounding new histones to be shaped in the new manner. By altering the shape of the histones around it, these modified histones would ensure that a differentiated cell would stay differentiated, and not convert back into being a stem cell.

2. The second way is the addition of methyl groups to the DNA, mostly at CpG sites, to convert cytosine to 5-methylcytosine. 5-Methylcytosine performs much like a regular cytosine, pairing up with a guanine. However, some areas of genome are methylated more heavily than others and highly methylated areas tend to be less transcriptionally active, through a mechanism not fully understood. Methylation of cytosines can also persist from the germ line of one of the parents into the zygote, marking the chromosome as being inherited from this parent (genetic imprinting).

The way that the cells stay differentiated in the case of DNA methylation is clearer to us than it is in the case of histone shape. Basically, certain enzymes (such as DNMT1) have a higher affinity for the methylated cytosine. If this enzyme reaches a "hemimethylated" portion of DNA (where methylcytosine is in only one of the two DNA strands) the enzyme will methylate the other half.

Although histone modifications occur throughout the entire sequence, the unstructured N-termini of histones (called histone tails) are particularly highly modified. These modifications include acetylation, methylation, ubiquitylation, phosphorylation and sumoylation. Acetylation is the most highly studied of these modifications. For example, acetylation of the K14 and K9 lysines of the tail of histone H3 by histone acetyltransferase enzymes (HATs) is generally correlated with transcriptional competence.

One mode of thinking is that this tendency of acetylation to be associated with "active" transcription is biophysical in nature. Because it normally has a positively charged nitrogen at its end, lysine can bind the negatively charged phosphates of the DNA backbone. The acetylation event converts the positively charged amine group on the side chain into a neutral amide linkage. This removes the positive charge, thus loosening the DNA from the histone. When this occurs, complexes like SWI/SNF and other transcriptional factors can bind to the DNA and allow transcription to occur. This is the "cis" model of epigenetic function. In other words, changes to the histone tails have a direct affect on the DNA itself.

Another model of epigenetic function is the "trans" model. In this model changes to the histone tails act indirectly on the DNA. For example, lysine acetylation may create a binding site for chromatin modifying enzymes (and basal transcription machinery as well). This Chromatin Remodeler can then cause changes to the state of the chromatin. Indeed, the bromodomain — a protein segment (domain) that specifically binds acetyl-

lysine — is found in many enzymes that help activate transcription, including the SWI/SNF complex (on the protein polybromo). It may be that acetylation acts in this and the previous way to aid in transcriptional activation.

The idea that modifications act as docking modules for related factors is borne out by histone methylation as well. Methylation of lysine 9 of histone H3 has long been associated with constitutively transcriptionally silent chromatin (constitutive heterochromatin). It has been determined that a chromodomain (a domain that specifically binds methyl-lysine) in the transcriptionally repressive protein HP1 recruits HP1 to K9 methylated regions. One example that seems to refute this biophysical model for acetylation is that tri-methylation of histone H3 at lysine 4 is strongly associated with (and required for full) transcriptional activation. Tri-methylation in this case would introduce a fixed positive charge on the tail.

It has been shown that the histone lysine methyltransferase (KMT) is responsible for this methylation activity in the pattern of histones H3 & H4. This enzyme utilizes a catalytically active site called the SET domain (Supressor of variegation, Enhancer of zeste, Trithorax). The SET domain is a 130-amino acid sequence involved in modulating gene activities. This domain has been demonstrated to bind to the histone tail and causes the methylation of the histone.

Differing histone modifications are likely to function in differing ways; acetylation at one position is likely to function differently than acetylation at another position. Also, multiple modifications may occur at the same time, and these modifications may work together to change the behavior of the nucleosome. The idea that multiple dynamic modifications regulate gene transcription in a systematic and reproducible way is called the histone code.

DNA methylation frequently occurs in repeated sequences, and helps to suppress the expression and mobility of 'transposable elements': Because 5-methylcytosine is chemically very similar to thymidine, CpG sites are frequently mutated and become rare in the genome, except at CpG islands where they remain unmethylated. Epigenetic changes of this type thus have the potential to direct increased frequencies of permanent genetic mutation. DNA methylation patterns are known to be established and modified in response to environmental factors by a complex interplay of at least three independent DNA methyltransferases, DNMT1, DNMT3A and DNMT3B, the loss of any of which is lethal in mice. DNMT1 is the most abundant methyltransferase in somatic cells, localizes to replication foci, has a 10–40-fold preference for hemimethylated DNA and interacts with the proliferating cell nuclear antigen (PCNA). By preferentially modifying hemimethylated DNA, DNMT1 transfers patterns of methylation to a newly synthesized strand after DNA replication, and therefore is often referred to as the ‘maintenance' methyltransferase. DNMT1 is essential for proper embryonic development, imprinting and X-inactivation.

Histones H3 and H4 can also be manipulated through demethylation using histone lysine demethylase (KDM). This recently identifited enzyme has a catalytically active site

called the Jumonji domain (JmjC). The demethylation occurs when JmjC utilizes multiple cofactors to hydroxylate the methyl group, thereby removing it. JmjC is capable of demethylating mono-, di-, and tri-methylated substrates. .

Chromosomal regions can adopt stable and heritable alternative states resulting in bistable gene expression without changes to the DNA sequence. Epigenetic control is often associated with alternative covalent modifications of histones. The stability and heritability of states of larger chromosomal regions are often thought to involve positive feedback where modified nucleosomes recruit enzymes that similarly modify nearby nucleosomes. A simplified stochastic model for this type of epigenetics is found here .

Because DNA methylation and chromatin remodeling play such a central role in many types of epigenic inheritance, the word "epigenetics" is sometimes used as a synonym for these processes. However, this can be misleading. Chromatin remodeling is not always inherited, and not all epigenetic inheritance involves chromatin remodeling.

It has been suggested that the histone code could be mediated by the effect of small RNAs. The recent discovery and characterization of a vast array of small (21- to 26-nt), non-coding RNAs suggests that there is an RNA component, possibly involved in epigenetic gene regulation. Small interfering RNAs can modulate transcriptional gene expression via epigenetic modulation of targeted promoters.

RNA transcripts and their encoded proteins

Sometimes a gene, after being turned on, transcribes a product that (either directly or indirectly) maintains the activity of that gene. For example, Hnf4 and MyoD enhance the transcription of many liver- and muscle-specific genes, respectively, including their own, through the transcription factor activity of the proteins they encode. RNA signalling includes differential recruitment of a hierarchy of generic chromatin modifying complexes and DNA methyltransferases to specific loci by RNAs during differentiation and development. Other epigenetic changes are mediated by the production of different splice forms of RNA, or by formation of double-stranded RNA (RNAi). Descendants of the cell in which the gene was turned on will inherit this activity, even if the original stimulus for gene-activation is no longer present. These genes are most often turned on or off by signal transduction, although in some systems where syncytia or gap junctions are important, RNA may spread directly to other cells or nuclei by diffusion. A large amount of RNA and protein is contributed to the zygote by the mother during oogenesis or via nurse cells, resulting in maternal effect phenotypes. A smaller quantity of sperm RNA is transmitted from the father, but there is recent evidence that this epigenetic information can lead to visible changes in several generations of offspring.

Prions

Prions are infectious forms of proteins. Proteins generally fold into discrete units which perform distinct cellular functions, but some proteins are also capable of forming an infectious conformational state known as a prion. Although often viewed in the context of

infectious disease, prions are more loosely defined by their ability to catalytically convert other native state versions of the same protein to an infectious conformational state. It is in this latter sense that they can be viewed as epigenetic agents capable of inducing a phenotypic change without a modification of the genome.

Fungal prions are considered epigenetic because the infectious phenotype caused by the prion can be inherited without modification of the genome. PSI+ and URE3, discovered in yeast in 1965 and 1971, are the two best studied of this type of prion. Prions can have a phenotypic effect through the sequestration of protein in aggregates, thereby reducing that protein's activity. In PSI+ cells, the loss of the Sup35 protein (which is involved in termination of translation) causes ribosomes to have a higher rate of read-through of stop codons, an effect which results in suppression of nonsense mutations in other genes. The ability of Sup35 to form prions may be a conserved trait. It could confer an adaptive advantage by giving cells the ability to switch into a PSI+ state and express dormant genetic features normally terminated by premature stop codon mutations.

Structural inheritance systems

In ciliates such as *Tetrahymena* and *Paramecium*, genetically identical cells show heritable differences in the patterns of ciliary rows on their cell surface. Experimentally altered patterns can be transmitted to daughter cells. It seems existing structures act as templates for new structures. The mechanisms of such inheritance are unclear, but reasons exist to assume that multicellular organisms also use existing cell structures to assemble new ones.

Functions and consequences

Development

Somatic epigenetic inheritance, particularly through DNA methylation and chromatin remodeling, is very important in the development of multicellular eukaryotic organisms. The genome sequence is static (with some notable exceptions), but cells differentiate into many different types, which perform different functions, and respond differently to the environment and intercellular signalling. Thus, as individuals develop, morphogens activate or silence genes in an epigenetically heritable fashion, giving cells a "memory". In mammals, most cells terminally differentiate, with only stem cells retaining the ability to differentiate into several cell types ("totipotency" and "multipotency"). In mammals, some stem cells continue producing new differentiated cells throughout life, but mammals are not able to respond to loss of some tissues, for example, the inability to regenerate limbs, which some other animals are capable of. Unlike animals, plant cells do not terminally differentiate, remaining totipotent with the ability to give rise to a new individual plant. While plants do utilise many of the same epigenetic mechanisms as animals, such as chromatin remodeling, it has been hypothesised that plant cells do not have "memories", resetting their gene expression patterns at each cell division using positional information from the environment and surrounding cells to determine their fate.

Medicine

Epigenetics has many and varied potential medical applications. Congenital genetic disease is well understood, and it is also clear that epigenetics can play a role, for example, in the case of Angelman syndrome and Prader-Willi syndrome. These are normal genetic diseases caused by gene deletions or inactivation of the genes, but are unusually common because individuals are essentially hemizygous because of genomic imprinting, and therefore a single gene knock out is sufficient to cause the disease, where most cases would require both copies to be knocked out.

Evolution

Although epigenetics in multicellular organisms is generally thought to be a mechanism involved in differentiation, with epigenetic patterns "reset" when organisms reproduce, there have been some observations of transgenerational epigenetic inheritance (e.g., the phenomenon of paramutation observed in maize). Although most of these multigenerational epigenetic traits are gradually lost over several generations, the possibility remains that multigenerational epigenetics could be another aspect to evolution and adaptation. A sequestered germ line or Weismann barrier is specific to animals, and epigenetic inheritance is expected to be far more common in plants and microbes. These effects may require enhancements to the standard conceptual framework of the modern evolutionary synthesis.

Epigenetic features may play a role in short-term adaptation of species by allowing for reversible phenotype variability. The modification of epigenetic features associated with a region of DNA allows organisms, on a multigenerational time scale, to switch between phenotypes that express and repress that particular gene. When the DNA sequence of the region is not mutated, this change is reversible. It has also been speculated that organisms may take advantage of differential mutation rates associated with epigenetic features to control the mutation rates of particular genes. Interestingly, recent analysis have suggested that members of the APOBEC family of cytosine deaminases are capable of simultaneously mediating genetic and epigenetic inheritance using similar molecular mechanisms.

Epigenetic changes have also been observed to occur in response to environmental exposure—for example, mice given some dietary supplements have epigenetic changes affecting expression of the agouti gene, which affects their fur color, weight, and propensity to develop cancer.

More than 100 cases of transgenerational epigenetic inheritance phenomena have been reported in a wide range of organisms, including prokaryotes, plants, and animals.

Epigenetic effects in humans

Genomic imprinting and related disorders

Some human disorders are associated with genomic imprinting, a phenomenon in mammals where the father and mother contribute different epigenetic patterns for specific genomic loci in their germ cells. The best-known case of imprinting in human disorders is that of Angelman syndrome and Prader-Willi syndrome—both can be produced by the same genetic mutation, chromosome 15q partial deletion, and the particular syndrome that will develop depends on whether the mutation is inherited from the child's mother or from their father. This is due to the presence of genomic imprinting in the region. Beckwith-Wiedemann syndrome is also associated with genomic imprinting, often caused by abnormalities in maternal genomic imprinting of a region on chromosome 11.

Transgenerational epigenetic observations

Marcus Pembrey and colleagues also observed in the Överkalix study that the paternal (but not maternal) grandsons of Swedish boys who were exposed during preadolescence to famine in the 19th century were less likely to die of cardiovascular disease; if food was plentiful then diabetes mortality in the grandchildren increased, suggesting that this was a transgenerational epigenetic inheritance. The opposite effect was observed for females—the paternal (but not maternal) granddaughters of women who experienced famine while in the womb (and therefore while their eggs were being formed) lived shorter lives on average.

Cancer and developmental abnormalities

A variety of compounds are considered as epigenetic carcinogens—they result in an increased incidence of tumors, but they do not show mutagen activity (toxic compounds or pathogens that cause tumors incident to increased regeneration should also be excluded). Examples include diethylstilbestrol, arsenite, hexachlorobenzene, and nickel compounds.

Many teratogens exert specific effects on the fetus by epigenetic mechanisms. While epigenetic effects may preserve the effect of a teratogen such as diethylstilbestrol throughout the life of an affected child, the possibility of birth defects resulting from exposure of fathers or in second and succeeding generations of offspring has generally been rejected on theoretical grounds and for lack of evidence. However, a range of male-mediated abnormalities have been demonstrated, and more are likely to exist. FDA label information for Vidaza(tm), a formulation of 5-azacitidine (an unmethylatable analog of cytidine that causes hypomethylation when incorporated into DNA) states that "men should be advised not to father a child" while using the drug, citing evidence in treated male mice of reduced fertility, increased embryo loss, and abnormal embryo development. In rats, endocrine differences were observed in offspring of males exposed to morphine. In mice, second generation effects of diethylstilbesterol have been described occurring by epigenetic mechanisms.

Recent studies have shown that the Mixed Lineage Leukemia (MLL) gene causes leukemia by rearranging and fusing with other genes in different chromosomes, which is a process under epigenetic control.

Other investigations have concluded that alterations in histone acetylation and DNA methylation occur in various genes influencing prostate cancer.

In 2008, the National Institutes of Health announced that $190 million had been earmarked for epigenetics research over the next five years. In announcing the funding, government officials noted that epigenetics has the potential to explain mechanisms of aging, human development, and the origins of cancer, heart disease, mental illness, as well as several other conditions. Some investigators, like Randy Jirtle, PhD, of Duke University Medical Center, think epigenetics may ultimately turn out to have a greater role in disease than genetics.

DNA methylation in cancer

DNA methylation is an important regulator of gene transcription and a large body of evidence has demonstrated that aberrant DNA methylation is associated with unscheduled gene silencing, and the genes with high levels of 5-methylcytosine in their promoter region are transcriptionally silent. DNA methylation is essential during embryonic development, and in somatic cells, patterns of DNA methylation are generally transmitted to daughter cells with a high fidelity. Aberrant DNA methylation patterns have been associated with a large number of human malignancies and found in two distinct forms: hypermethylation and hypomethylation compared to normal tissue. Hypermethylation is one of the major epigenetic modifications that repress transcription via promoter region of tumour suppressor genes. Hypermethylation typically occurs at CpG islands in the promoter region and is associated with gene inactivation. Global hypomethylation has also been implicated in the development and progression of cancer through different mechanisms.

Variant histones H2A in cancer

The histone variants of the H2A family are highly conserved in mammals, playing critical roles in regulating many nuclear processes by altering chromatin structure. One of the key H2A variants, H2A.X, marks DNA damage, facilitating the recruitment of DNA repair proteins to restore genomic integrity. Another variant, H2A.Z, plays an important role in both gene activation and repression. A high level of H2A.Z expression is ubiquitously detected in many cancers and is significantly associated with cellular proliferation and genomic instability.

Cancer Treatment

Current research has shown that epigenetic pharmaceuticals could be a putative replacement or adjuvant therapy for currently accepted treatment methods such as radiation and chemotherapy, or could enhance the effects of these current treatments. It

has been shown that the epigenetic control of the proto-onco regions and the tumor suppressor sequences by conformational changes in histones directly affects the formation and progression of cancer Epigenetics also has the factor of reversibility, a characteristic that other cancer treatments do not offer.

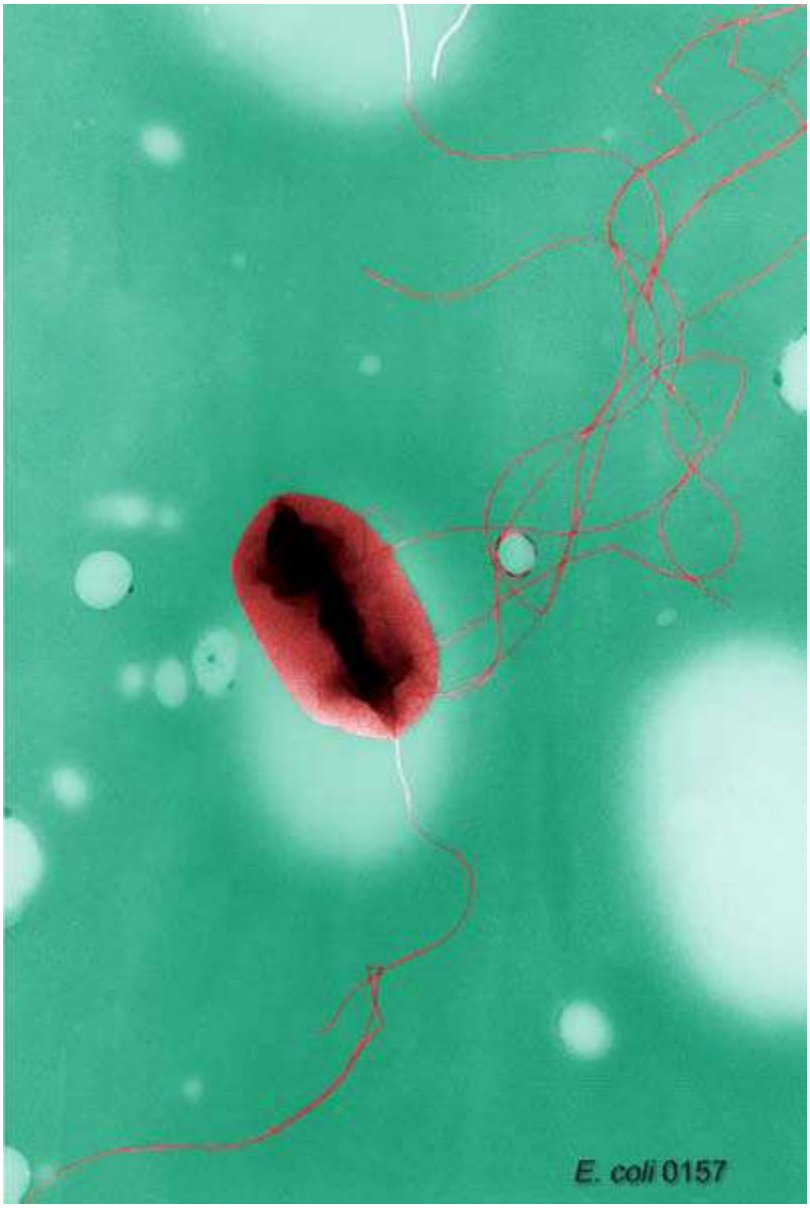
Drug development has mainly focused on Histone Acetyltransferase (HAT) and Histone Deactylase (HDAC), including the introduction of the new pharmaceutical Vorinostat, a HDAC inhibitor, to the market. HDAC specifically has been shown to play an integral role in the progression of oral squamous cancer

Current front-runner candidates for new drug targets are Histone Lysine Methyltransferases (KMT) and Protein Arginine Methyltransferases (PRMT).

Twin studies

Recent studies involving both dizygotic and monozygotic twins have produced some evidence of epigenetic influence in humans.

Epigenetics in microorganisms



Escherichia coli bacteria

Bacteria make widespread use of postreplicative DNA methylation for the epigenetic control of DNA-protein interactions. Bacteria make use of DNA adenine methylation (rather than DNA cytosine methylation) as an epigenetic signal. DNA adenine methylation is important in bacteria virulence in organisms such as *Escherichia coli*, *Salmonella, Vibrio, Yersinia, Haemophilus*, and *Brucella*. In *Alphaproteobacteria*, methylation of adenine regulates the cell cycle and couples gene transcription to DNA replication. In *Gammaproteobacteria*, adenine methylation provides signals for DNA replication, chromosome segregation, mismatch repair, packaging of bacteriophage, transposase activity and regulation of gene expression.

The filamentous fungus *Neurospora crassa* is a prominent model system for understanding the control and function of cytosine methylation. In this organisms, DNA methylation is associated with relics of a genome defense system called RIP (repeat-induced point mutation) and silences gene expression by inhibiting transcription elongation.

The yeast prion PSI is generated by a conformational change of a translation termination factor, which is then inherited by daughter cells. This can provide a survival advantage under adverse conditions. This is an example of epigenetic regulation enabling unicellular organisms to respond rapidly to environmental stress. Prions can be viewed as epigenetic agents capable of inducing a phenotypic change without modification of the genome.

Chapter 2

Transgenerational Epigenetics

Epigenetic inheritance is the transmittance of information from one generation to the next that affects the traits of offspring without alteration of the primary structure of DNA (i.e. the sequence of nucleotides) or from environmental cues. The term “epigenetic inheritance” is used to describe both cell-cell and organism-organism information transfer, while **transgenerational epigenetics** typically refers only to the latter. Although these two levels of epigenetic inheritance are equivalent in unicellular organisms, they may have distinct mechanisms and evolutionary distinctions in multicellular organisms.

Cloned mice with different DNA methylation patterns causing kinks in the tail of one but not the other.

Four general categories of epigenetic modification are known: 1) self-sustaining metabolic loops, in which a mRNA or protein product of a gene stimulates transcription of the gene (e.g. *Wor1* gene in *Candida albicans*), 2) structural templating in which structures are replicated using a template or scaffold structure on the parent (e.g. prions, proteins that replicate by changing the structure of normal proteins to match their own), 3) chromatin marks, in which methyl or acetyl groups bind to DNA nucleotides or histones thereby altering gene expression patterns (e.g. *Lcyc* gene in *Linaria vulgaris* described below), and 4) RNA silencing, in which small RNA strands interfere (RNAi) with the transcription of DNA or translation of mRNA (known only from a few studies, mostly in *Caenorhabditis elegans*).

For some epigenetically influenced traits, the epigenetic marks can be induced by the environment and some marks are heritable, leading some to view epigenetics as a relaxation of the rejection of Lamarckian evolution of acquired characters.

Major Controversies in the History of the Inheritance

Humans have recognized that traits of the parents are often seen in offspring. This insight led to the practical application of selective breeding of plants and animals, eventually leading to domestication, but did not address the central question of inheritance: how are these traits conserved between generations, and what causes variation?

Blending vs. Particulate Inheritance

Addressing these related questions, scientists during the time of the Enlightenment largely argued for the blending hypothesis, in which parental traits were homogenized in the offspring much like buckets of different colored paint being mixed together. Critics of Charles Darwin's *On the Origin of Species*, pointed out that under this scheme of inheritance, variation would quickly be swamped by the majority phenotype. In the paint bucket analogy, this would be seen by mixing two colors together and then mixing the resulting color with only one of the parent colors 20 times; the rare variant color would quickly fade.

Unknown to most of the European scientific community, a monk by the name of Gregor Mendel had resolved the question of how traits are conserved between generations through breeding experiments with pea plants. Charles Darwin thus did not know of Mendel's proposed "particulate inheritance" in which traits were not blended but passed to offspring in discrete units that we now call genes. Darwin came to reject the blending hypothesis even though his ideas and Mendel's were not unified until the 1930s, a period referred to as the Modern Synthesis.

Inheritance of Innate vs. Acquired Characteristics

In his 1809 book, *Philosophie Zoologique*, Jean-Baptiste Lamarck recognized that each species experiences a unique set of challenges due to its form and environment. Thus, he proposed that the characters used most often would accumulate a “nervous fluid.” Such

acquired accumulations would then be transmitted to the individual's offspring. In modern terms, a nervous fluid transmitted to offspring would be a form of epigenetic inheritance.

Lamarckism, as this body of thought became known, was the standard explanation for change in species over time when Charles Darwin and Alfred Russel Wallace co-proposed a theory of evolution by natural selection in 1859. Responding to Darwin and Wallace's theory, a revised neo-Lamarckism attracted a small following of biologists, though the Lamarckian zeal was quenched in large part due to Weismann's famous experiment in which he cut off the tails of mice over several successive generations without having any effect on tail length. Thus the emergent consensus that acquired characteristics could not be inherited became canon.

Origin of Epigenetics and Revision of the Modern Synthesis

Non-genetic variation and inheritance, however, proved to be quite common. Concurrent to the Modern Synthesis (unifying Mendelian genetics and natural selection), C. H. Waddington was working to unify developmental biology and genetics. In so doing, he coined the word "epigenetic" to represent the ordered differentiation of embryonic cells into functionally distinct cell types despite having identical primary structure of their DNA. Waddington's epigenetics was sporadically discussed, becoming more of a catch-all for puzzling non-genetic heritable characters rather than advancing the body of inquiry. Consequently, the definition of Waddington's word has itself evolved, broadening beyond the subset of developmentally signaled, inherited cell specialization.

Does epigenetic inheritance compromise the foundation of the Modern Synthesis? Outlining the Central Dogma of Molecular Biology, Francis Crick succinctly stated, "DNA is held in a configuration by histone[s] so that it can act as a passive template for the simultaneous synthesis of RNA and protein[s]. *None* of the detailed "information" is in the histone (italic added for emphasis). However, he closes the article stating, "this scheme *explains the majority* of the present experimental results!" (italic added for emphasis). Indeed the emergence of epigenetic inheritance (in addition to advances in the study of evolutionary-development, phenotypic plasticity, evolvability, and systems biology) has strained the current framework of the Modern Synthesis and prompted the re-examination of previously dismissed evolutionary mechanisms.

Origin and Inheritance of Epigenes

Epigenetic variation may take one of four general forms. Others may yet be elucidated, but currently self-sustaining feedback loops, spatial templating, chromatin marking, and RNA-mediated pathways modify epigenes at the level of individual cells. Epigenetic variation within multicellular organisms may be endogenous, generated by cell-cell signaling (e.g. during cell differentiation early in development), or exogenous, a cellular response to environmental cues.

Removal vs. retention of epigenetic marks

In sexually reproducing organisms, much of the epigenetic modification within cells is reset during meiosis (e.g. marks at the FLC locus controlling plant vernalization), though some epigenetic responses have been shown to be conserved (e.g. transposon methylation in plants). Differential inheritance of epigenetic marks due to underlying maternal or paternal biases in removal or retention mechanisms may lead to the assignment of epigenetic causation to some parent of origin effects in animals and plants.

Removal of epigenetic marks

In mammals, male and female gametes join during fertilization in different cell cycle states and with different configuration of the genome. The epigenetic marks of the male are rapidly diluted. First, the protamines associated with male DNA are replaced with histones from the female's cytoplasm, most of which are acetylated due to either higher abundance of acetylated histones in the female's cytoplasm or through preferential binding of the male DNA to acetylated histones. Second, male DNA is systematically demethylated in many organisms, however the mechanism and functional outcome of this process has yet to be elucidated.

Recognition of the importance of epigenetic programming to the establishment and fixation of cell line identity during early embryogenesis has recently stimulated interest in artificial removal of epigenetic programming. Epigenetic manipulations may allow for restoration of totipotency in stem cells or cells more generally, thus generalizing regenerative medicine.

Retention of epigenetic marks

Cellular mechanisms may allow for co-transmission of some epigenetic marks. During replication, DNA polymerases working on the leading and lagging strands are coupled by the DNA processivity factor proliferating cell nuclear antigen (PCNA), which has also been implicated in patterning and strand crosstalk that allows for copy fidelity of epigenetic marks. Work on histone modification copy fidelity has remained in the model phase, but early efforts suggest that modifications of new histones are patterned on those of the old histones and that new and old histones randomly assort between the two daughter DNA strands. With respect to transfer to the next generation, many marks are removed as described above. Emerging studies are finding patterns of epigenetic conservation across generations. For instance, centromeric satellites resist demethylation. The mechanism responsible for this conservation is not known, though some evidence suggests that methylation of histones may contribute.

Decay of epigenetic marks

Whereas the mutation rate in a given 100 base gene may be 10^{-7} per generation, epigenes may “mutate” several times per generation or may be fixed for many generations. This raises the question: are changes in epigene frequencies is evolution? Rapidly decaying

epigenetic effects on phenotypes (i.e. lasting less than three generations) may explain some of the residual variation in phenotypes after genotype and environment are accounted for. However, distinguishing these short-term effects from the effects of the maternal environment on early onotgeny remains a challenge.

Contribution to Phenotypes

The relative importance of genetic and epigenetic inheritance is subject to debate. Though hundreds of examples of epigenetic modification of phenotypes have been published, few studies have been conducted outside of the laboratory setting. Therefore, the interactions of genes and epigenes with the environment cannot be inferred despite the central role of environment in natural selection. Experimental methodologies for manipulating epigenetic mechanisms are nascient (e.g.) and will need rigorous demonstration before studies explicitly testing the relative contributions of genotype, environment, and epigenotype are feasible.

Effects on Fitness

Epigenetic inheritance may only affect fitness if it predictably alters a trait under selection. Evidence has been forwarded that environmental stimuli are important agents in the alteration of epigenes. Ironically, Darwinian evolution may act on these neo-Lamarckian acquired characteristics as well as the cellular mechanisms producing them (e.g. methyltransferase genes). Epigenetic inheritance may confer a fitness benefit to organisms that deal with environmental changes at intermediate timescales. Short-cycling changes are likely to have DNA-encoded regulatory processes, as the probability of the offspring needing to respond to changes multiple times during their lifespans is high. On the other end, natural selection will act on populations experiencing changes on longer-cycling environmental changes. In these cases, if epigenetic priming of the next generation is deleterious to fitness over most of the interval (e.g. misinformation about the environment), these genotypes and epigenotypes will be lost. For intermediate time cycles, the probability of the offspring encountering a similar environment is sufficiently high without substantial selective pressure on individuals lacking a genetic architecture capable of responding to the environment. Naturally, the absolute lengths of short, intermediate, and long environmental cycles will depend on the trait, the length of epigenetic memory, and the generation time of the organism. Much of the interpretation of epigenetic fitness effects centers on the hypothesis that epigenes are important contributors to phenotypes, which remains to be resolved.

Deleterious effects

Inherited epigenetic marks may be important for regulating important components of fitness. In plants, for instance, the *Lcyc* gene in *Linaria vulgaris* controls the symmetry of the flower. Linnaeus first described radially symmetric mutants, which arises when *Lcyc* is heavily methylated. Given the importance of floral shape to pollinators, methylation of *Lcyc* homologues (e.g. *CYCLOIDEA*) may have deleterious effects on plant fitness. In animals, numerous studies have shown that inherited epigenetic marks can increase

susceptibility to disease. Transgenerational epigenetic influences are also suggested to contribute to disease, especially cancer, in humans. Tumor methylation patterns in gene promotors have been shown to correlate positively with familial history of cancer. Furthermore, methylation of the *MSH2* gene is correlated with early-onset colorectal and endometrial cancers.

Putatively adaptive effects

Experimentally demethylated seeds of the model organism *Arabidopsis thaliana* have significantly higher mortality, stunted growth, delayed flowing, and lower fruit set, indicating that epigenes may increase fitness. Furthermore, environmentally induced epigenetic responses to stree have been shown to be inherited and positively correlated with fitness. In animals, communal nesting changes mouse behavior increasing parental care regimes and social abilities that are hypothesized to increase offspring survival and access to resources (such as food and mates), respectively.

Macroevolutionary Patterns

Inherited epigenetic effects on phenotypes have been documented in bacteria, protists, fungi, plants, and animals. Though no systematic study of epigenetic inheritance has been conducted (most focus on model organisms), there is preliminary evidence that this mode of inheritance is more important in plants than in animals. The early differentiation of animal germlines is likely to preclude epigenetic marking occurring later in development, while in plants and fungi somatic cells may be incorporated into the germ line.

Life history patterns may also contribute to the occurrence of epigenetic inheritance. Sessile organisms, those with low dispersal capability, and those with simple behavior may benefit most from conveying information to their offspring via epigenetic pathways. Geographic patterns may also emerge, where highly variable and highly conserved environments might host fewer species with important epigenetic inheritance.

Chapter 3

Genomic Imprinting

Genomic imprinting is a genetic phenomenon by which certain genes are expressed in a parent-of-origin-specific manner. It is an inheritance process independent of the classical Mendelian inheritance. Imprinted genes are either expressed **only** from the allele inherited from the mother (e.g. *H19* or *CDKN1C*), or in other instances from the allele inherited from the father (e.g. *IGF-2*). Forms of genomic imprinting have been demonstrated in insects, mammals and flowering plants.

Genomic imprinting is an epigenetic process that involves methylation and histone modifications in order to achieve monoallelic gene expression without altering the genetic sequence. These epigenetic marks are established in the germline and are maintained throughout all somatic cells of an organism.

Appropriate expression of imprinted genes is important for normal development, with numerous genetic diseases associated with imprinting defects including Beckwith-Wiedemann syndrome, Silver-Russell Syndrome, Angelman Syndrome and Prader-Willi Syndrome.

Overview

In diploid organisms, somatic cells possess two copies of the genome. Each autosomal gene is therefore represented by two copies, or alleles, with one copy inherited from each parent at fertilisation. For the vast majority of autosomal genes, expression occurs from both alleles simultaneously. In mammals, however, a small proportion (<1%) of genes are imprinted, meaning that gene expression occurs from only one allele. The expressed allele is dependent upon its parental origin. For example, the gene encoding Insulin-like growth factor 2 (IGF2/Igf2) is only expressed from the allele inherited from the father.

The phrase "imprinting" was first used to describe events in the insect *Pseudococcus nipae*. In Pseudococcids or mealybugs (Homoptera, Coccoidea) both the male and female develop from a fertilised egg. In females, all chromosomes remain euchromatic and functional. In embryos destined to become males, one haploid set of chromosomes becomes heterochromatinised after the sixth cleavage division and remains so in most

tissues; males are thus functionally haploid. In insects, imprinting describes the silencing of the paternal genome in males, and thus is involved in sex determination. In mammals, genomic imprinting describes the processes involved in introducing functional inequality between two parental alleles of a gene.

Imprinted genes in mammals

That imprinting might be a feature of mammalian development was suggested in breeding experiments in mice carrying reciprocal translocations. Nucleus transplantation experiments in mouse zygotes in the early 1980s confirmed that normal development requires the contribution of both the maternal and paternal genomes. The vast majority of mouse parthenogenones/gynogenones (with two maternal or egg genomes) and androgenones (with two paternal or sperm genomes) die at, or before, the blastocyst/implantation stage. In the rare instances that they develop to postimplantation stages, gynogenetic embryos show better embryonic development relative to placental development, while for androgenones, the reverse is true. Nevertheless, for the latter, only a few have been described.

Parthenogenetic/gynogenetic embryos have twice the normal expression level of maternally derived genes, and lack expression of paternally expressed genes, while the reverse is true for androgenetic embryos. It is now known that there are at least 80 imprinted genes in humans and mice, many of which are involved in embryonic and placental growth and development. Various methods have been used to identify imprinted genes. In swine, Bischoff et al 2009 compared transcriptional profiles using short-oligonucleotide microarrays (Affymetrix Porcine GeneChip) to survey differentially expressed genes between parthenotes (2 maternal genomes) and control fetuses (1 maternal, 1 paternal genome) An intriguing study surveying the transcriptome of murine brain tissues revealed over 1300 imprinted gene loci (approximately 10-fold more than previously reported) by Illumina RNA-sequencing (RNA-Seq) technology from F1 hybrids resulting from reciprocal crosses.

No naturally occurring cases of parthenogenesis exist in mammals because of imprinted genes. Experimental manipulation of a paternal methylation imprint controlling the Igf2 gene has, however, recently allowed the creation of rare individual mice with two maternal sets of chromosomes - but this is not a true parthenogenone. Hybrid offspring of two species may exhibit unusual growth due to the novel combination of imprinted genes.

Genetic mapping of imprinted genes

At the same time as the generation of the gynogenetic and androgenetic embryos discussed above, mouse embryos were also being generated that contained only small regions that were derived from either a paternal or maternal source. The generation of a series of such uniparental disomies, which together span the entire genome, allowed the creation of an imprinting map. Those regions which when inherited from a single parent result in a discernible phenotype contain imprinted gene(s). Further research showed that within these regions there were often numerous imprinted genes. Around 80% of

imprinted genes are found in clusters such as these, called imprinted domains, suggesting a level of co-ordinated control. More recently, genome-wide screens to identify imprinted genes have used differential expression of mRNAs from control fetuses and parthenogenetic or androgenetic fetuses hybridized to expression arrays, allele-specific gene expression using SNP genotyping arrays, transcriptome sequencing, in silico prediction pipelines...to name a few.

Imprinting mechanisms

Imprinting is a dynamic process. It must be possible to erase and re-establish the imprint through each generation. The nature of the imprint must therefore be epigenetic (modifications to the structure of the DNA rather than the sequence). In germline cells the imprint is erased, and then re-established according to the sex of the individual; i.e. in the developing sperm (during spermatogenesis), a paternal imprint is established, whereas in developing oocytes (oogenesis), a maternal imprint is established. This process of erasure and reprogramming is necessary such that the current imprinting status is relevant to the sex of the individual. In both plants and mammals there are two major mechanisms that are involved in establishing the imprint; these are DNA methylation and histone modifications.

Regulation

The grouping of imprinted genes within clusters allows them to share common regulatory elements, such as non-coding RNAs and differentially methylated regions (DMRs). When these regulatory elements control the imprinting of one or more genes, they are known as imprinting control regions (ICR). The expression of non-coding RNAs, such as *Air* on mouse chromosome 17 and KCNQ1OT1 on human chromosome 11p15.5, have been shown to be essential for the imprinting of genes in their corresponding regions.

Differentially methylated regions are generally segments of DNA rich in cytosine and guanine nucleotides, with the cytosine nucleotides methylated on one copy but not on the other. Contrary to expectation, methylation does not necessarily mean silencing; instead, the effect of methylation depends upon the default state of the region.

Functions of imprinted genes

The control of expression of specific genes by genomic imprinting is unique to therian mammals (placental mammals and marsupials) and flowering plants. Imprinting of whole chromosomes has been reported in mealybugs. and a fungus gnat (*Sciara*). It has also been established that X-chromosome inactivation occurs in an imprinted manner in the extra-embryonic tissues of mice, where it is always the paternal X-chromosome which is silenced.

The majority of imprinted genes in mammals have been found to have roles in the control of embryonic growth and development, including development of the placenta. Other

imprinted genes are involved in post-natal development, with roles affecting suckling and metabolism.

Theories on the origins of imprinting

A widely accepted hypothesis for the evolution of genomic imprinting is the "parental conflict hypothesis." Also known as the kinship theory of genomic imprinting, this hypothesis states that the inequality between parental genomes due to imprinting is a result of the differing interests of each parent in terms of the evolutionary fitness of their genes. The father's genes that encode for imprinting gain greater fitness through the success of the offspring, at the expense of the mother. The mother's evolutionary imperative is often to conserve resources for her own survival while providing sufficient nourishment to current and subsequent litters. Accordingly, paternally expressed genes tend to be growth promoting whereas maternally expressed genes tend to be growth limiting. In support of this hypothesis, genomic imprinting has been found in all placental mammals, where post-fertilisation offspring resource consumption at the expense of the mother is high; it has not been found in oviparous birds or monotremes (a class of oviparous mammals) where there is relatively little post-fertilisation resource transfer and therefore less parental conflict.

However, our understanding of the molecular mechanisms behind genomic imprinting show that it is the maternal genome that controls much of the imprinting of both its own and the paternally-derived genes in the zygote, making it difficult to explain why the maternal genes would willingly relinquish their dominance to that of the paternally-derived genes in light of the conflict hypothesis. Several other hypotheses that propose a coadaptive reason for the evolution of genomic imprinting have been proposed.

Others have approached their study of the origins of genomic imprinting from a different side, arguing that natural selection is operating on the role of epigenetic marks as machinery for homologous chromosome recognition during meiosis, rather than on their role in differential expression. This argument centers on the existence of epigenetic effects on chromosomes that do not directly affect gene expression, but do depend on which parent the chromosome originated from. This group of epigenetic changes that depend on the chromosome's parent of origin (including both those that affect gene expression and those that do not) are called parental origin effects, and include phenomena such as paternal X inactivation in the marsupials, nonrandom parental chromatid distribution in the ferns, and even mating type switching in yeast. This diversity in organisms that show parental origin effects has prompted theorists to place the evolutionary origin of genomic imprinting before the last common ancestor of plants and animals, over a billion years ago.

Natural selection for genomic imprinting requires genetic variation in a population. A hypothesis for the origin of this genetic variation states that the host-defense system responsible for silencing foreign DNA elements, such as genes of viral origin, mistakenly silenced genes whose silencing turned out to be beneficial for the organism. There appears to be an over-representation of retrotransposed genes, that is to say genes that are

inserted into the genome by viruses, among imprinted genes. It has also been postulated that if the retrotransposed gene is inserted close to another imprinted gene, it may just acquire this imprint.

Problems associated with imprinting

Imprinting may cause problems in cloning, with clones having DNA that is not methylated in the correct position. It is possible that this is due to a lack of time for reprogramming to be completely achieved. When a nucleus is added to an egg during somatic cell nuclear transfer, the egg starts dividing in minutes, as compared to the days or months it takes for reprogramming during embryonic development. If time is the responsible factor, it may be possible to delay cell division in clones, giving time for proper reprogramming to occur.

An allele of the "callipyge" (from the Greek for "beautiful buttocks"), or CLPG, gene in sheep produces large buttocks consisting of muscle with very little fat. The large-buttocked phenotype only occurs when the allele is present on the copy of chromosome 18 inherited from a sheep's father and is *not* on the copy of chromosome 18 inherited from that sheep's mother.

Examples

Prader-Willi/Angelman

The first imprinted genetic disorders to be described in humans were the reciprocally inherited Prader-Willi syndrome and Angelman syndrome. Both syndromes are associated with loss of the chromosomal region 15q11-13 (band 11 of the long arm of chromosome 15). This region contains the paternally expressed genes (SNRPN and NDN) and the maternally expressed gene (UBE3A).

- Paternal inheritance of a deletion of this region is associated with Prader-Willi syndrome (characterised by hypotonia, obesity, and hypogonadism).
- Maternal inheritance of the same deletion is associated with Angelman syndrome (characterised by epilepsy, tremors, and a perpetually smiling facial expression).

NOEY2

NOEY2 is a paternally expressed imprinted gene located on chromosome 1 in humans. Loss of NOEY2 expression is linked to an increased risk of ovarian and breast cancers; in 41% of breast and ovarian cancers the protein transcribed by NOEY2 is not expressed, suggesting that it functions as a tumor suppressor gene Therefore, if a person inherits both chromosomes from the mother, the gene will not be expressed and the individual is put at a greater risk for breast and ovarian cancer.

Other

Other conditions involving imprinting include Beckwith-Wiedemann syndrome, Silver-Russell syndrome, and pseudohypoparathyroidism.

Transient neonatal diabetes mellitus can also involve imprinting.

Imprinted genes in plants

A similar imprinting phenomenon has also been described in flowering plants (angiosperms). During fertilisation of the egg cell, a second, separate fertilization event gives rise to the endosperm, an extraembryonic structure that nourishes the embryo in a manner analogous to the mammalian placenta. Unlike the embryo, the endosperm is often formed from the fusion of two maternal cells with a male gamete. This results in a triploid genome. The uneven ratio of maternal to paternal genomes appears to be critical for seed development. Some genes are found to be expressed from both maternal genomes while others are expressed exclusively from the lone paternal copy.

Chapter 4

Methylated DNA Immunoprecipitation

Methylated DNA immunoprecipitation (MeDIP or mDIP) is a large-scale (chromosome- or genome-wide) technique that is used to enrich for methylated DNA sequences. It consists of isolating methylated DNA fragments via an antibody raised against 5-methylcytosine (5mC). This technique was first described by Weber M. *et al.* and has helped pave the way for viable methylome-level assessment efforts, as the purified fraction of methylated DNA can be input to high-throughput DNA detection methods such as high-resolution DNA microarrays (MeDIP-chip) or next-generation sequencing (MeDIP-seq). Nonetheless, understanding of the methylome remains rudimentary; its study is complicated by the fact that, like other epigenetic properties, patterns vary from cell-type to cell-type.

Background

DNA methylation, referring to the reversible methylation of the 5' position of cytosine by methyltransferases, is a major epigenetic modification in multicellular organisms. In mammals, this modification primarily occurs at CpG sites, which in turn tend to cluster in regions called CpG islands. There is a small fraction of CpG islands that can overlap or be in close proximity to promoter regions of transcription start sites. The modification may also occur at other sites, but methylation at either of these sites can repress gene expression by either interfering with the binding of transcription factors or modifying chromatin structure to a repressive state.

Disease condition studies have largely fueled the effort in understanding the role of DNA methylation. Currently, the major research interest lies in investigating disease conditions such as cancer to identify regions of the DNA that has undergone extensive methylation changes. The genes contained in these regions are of functional interest as they may offer a mechanistic explanation to the underlying genetic causes of a disease. For instance, the abnormal methylation pattern of cancer cells was initially shown to be mechanism through which tumor suppressor-like genes are silenced, although it was later observed that a much broader range of gene types are affected.

Other technologies

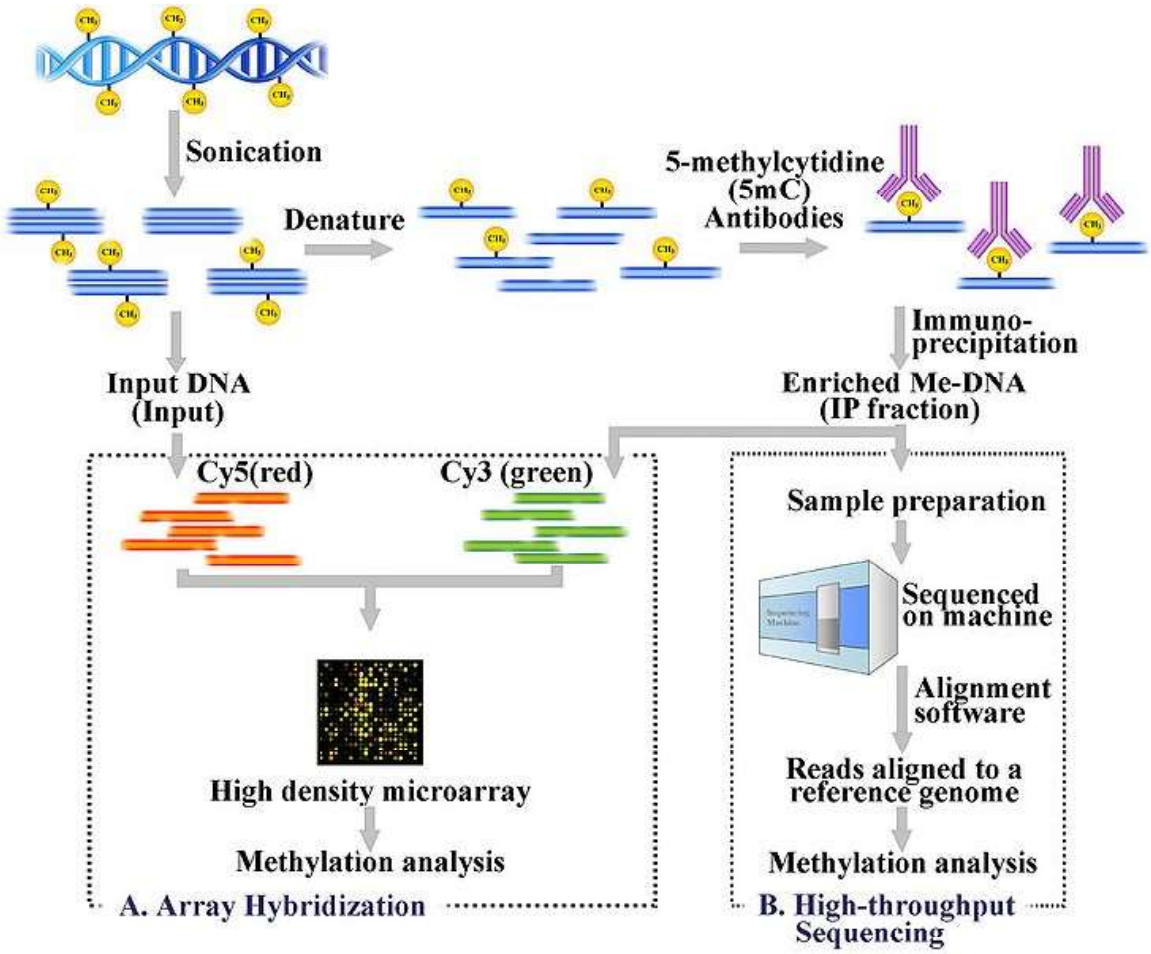
There are two approaches to methylation analysis: typing and profiling technologies. Typing technologies are targeted towards a small number of loci across many samples, and involve the use of techniques such as PCR, restriction enzymes, and mass spectrometry. Profiling technologies such as MeDIP are targeted towards a genome- or methylome-wide level assessment of methylation; this includes restriction landmark genomic scanning (RLGS), and bisulfite conversion-based methods, which rely on the treatment of DNA with bisulfite to convert unmethylated cytosine residues to uracil.

Limitations of other technologies

Other methods mapping and profiling the methylome have been effective but are not without their limitations that can affect resolution, level of throughput, or experimental variations. For instance, RLGS is limited by the number of restriction sites in genome that can be targets for the restriction enzyme; typically, a maximum of ~4100 landmarks can be assessed. Bisulfite sequencing-based methods, despite possible single-nucleotide resolution, have a drawback: the conversion of unmethylated cytosine to uracil can be unstable. In addition, when bisulfite conversion is coupled with DNA microarrays to detect bisulfite converted sites, the reduced sequence complexity of DNA is a problem. Microarrays capable of comprehensively profiling the whole-genome become difficult to design as fewer unique probes are available.

Methods

The following sections outline the method of MeDIP coupled with either high-resolution array hybridization or high-throughput sequencing. Each DNA detection method will also briefly describe post-laboratory processing and analysis. Different post-processing of the raw data is required depending on the technology used to identify the methylated sequences. This is analogous to data generated using ChIP-chip and ChIP-seq.

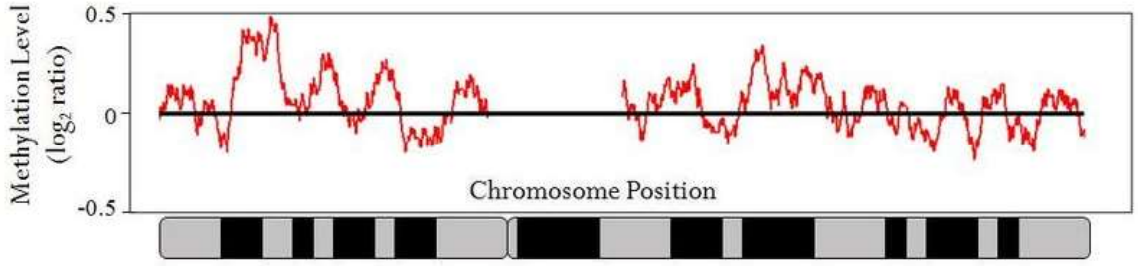


Workflow overview of the MeDIP procedure. MeDIP procedure is followed by array-hybridization (A) or high-throughput/next generation sequencing (B)

Methylated DNA immunoprecipitation (MeDIP)

Genomic DNA is extracted (DNA extraction) from the cells and purified. The purified DNA is then subjected to sonication to shear it into random fragments. This sonication process is quick, simple, and avoids restriction enzyme biases. The resulting fragments range from 300 to 1000 base pairs (bp) in length, although they are typically between 400 and 600 bp. The short length of these fragments are important in obtaining adequate resolution, improving the efficiency of the downstream step in immunoprecipitation, and reducing fragment-length effects or biases. Also, the size of the fragment affects the binding of 5-methyl-cytidine (5mC) antibody because the antibody needs more than just a single 5mC for efficient binding. To further improve binding affinity of the antibodies, the DNA fragments are denatured to produce single-stranded DNA. Following denaturation, the DNA is incubated with monoclonal 5mC antibodies. The classical immunoprecipitation technique is then applied: magnetic beads conjugated to anti-mouse-IgG are used to bind the anti-5mC antibodies, and unbound DNA is removed in the supernatant. To purify the DNA, proteinase K is added to digest the antibodies and release the DNA, which can be collected and prepared for DNA detection.

MeDIP and array-based hybridization (MeDIP-chip)



Simulated data to visualize typical analysis using MeDIP-chip.

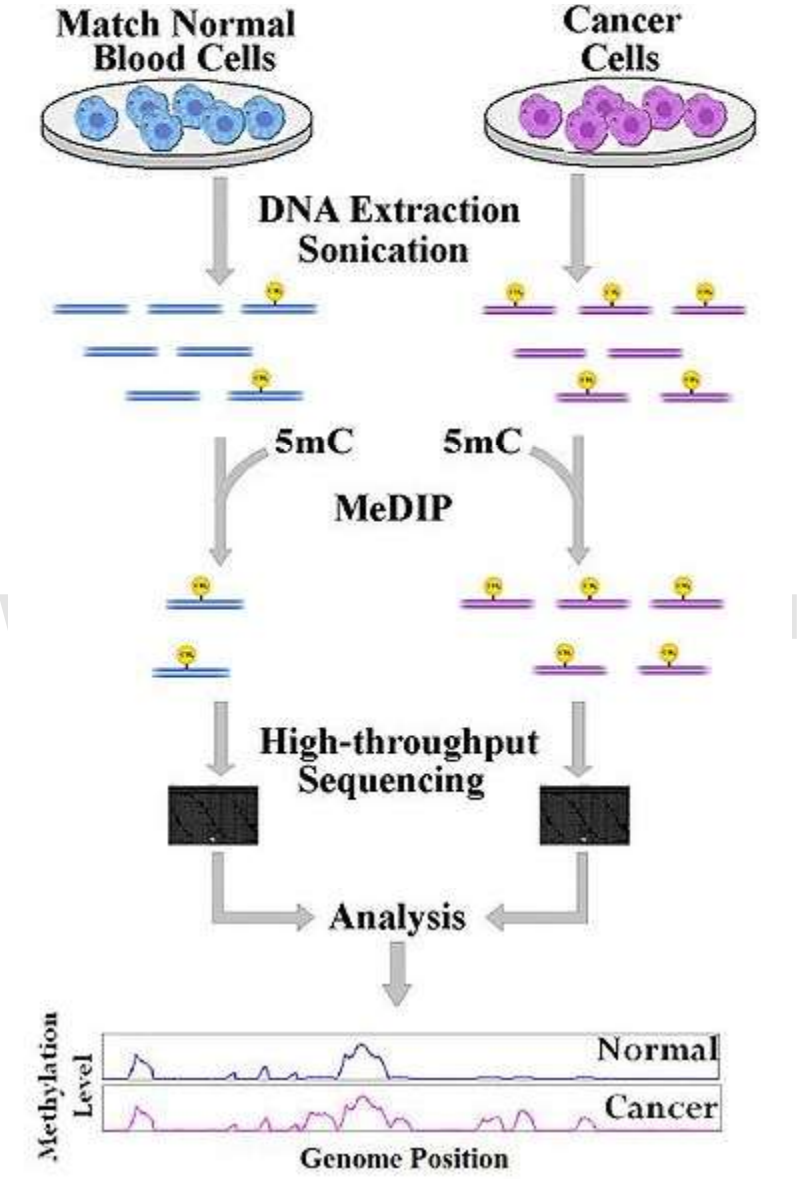
A fraction of the input DNA obtained after the sonication step above is labeled with cyanine-5 (Cy5; red) deoxy-cytosine-triphosphate while the methylated DNA, enriched after the immunoprecipitation step, is labeled with cyanine-3 (Cy3; green). The labeled DNA samples are cohybridized on a 2-channel, high-density genomic microarray to probe for presence and relative quantities. The purpose of this comparison is identify sequences that show significant differential expression, thereby confirming the sequence of interest is enriched. Array-based identification of MeDIP sequences are limited to the array design. As a result, the resolution is restricted to the probes in the array design. There are additional standard steps required in signal processing to correct for hybridization issues such as noise, as is the case with most array technologies.

MeDIP and high-throughput sequencing (MeDIP-seq)

The MeDIP-seq approach, i.e. the coupling of MeDIP with next generation, short-read sequencing technologies such as 454, Illumina (company) (Solexa), and SoLiD (Applied Biosystems), was first described by Down *et al.* in 2008. The high-throughput sequencing of the methylated DNA fragments produces a large number of short reads (36-50bp or 400 bp, depending on the technology). The short reads are aligned to a reference genome using alignment software such as Mapping and Assembly with Quality (Maq), which uses a Bayesian approach, along with base and mapping qualities to model error probabilities for the alignments. The reads can then be extended to represent the ~400 to 700 bp fragments from the sonication step. The coverage of these extended reads can be used to estimate the methylation level of the region. A genome browser such as Ensembl can also be used to visualize the data.

Validation of the approach to assess quality and accuracy of the data can be done with quantitative PCR. This is done by comparing a sequence from the MeDIP sample against an unmethylated control sequence. The samples are then run on a gel and the band intensities are compared. The relative intensity serves as the guide for finding enrichment. The results can also be compared with MeDIP-chip results to help determine coverage needed.

Downstream bioinformatics analysis



Simple workflow demonstrating a typical experiment using MeDIP-seq.

The DNA methylation level estimations can be confounded by varying densities of methylated CpG sites across the genome when observing data generated by MeDIP. This can be problematic for analyzing CpG-poor (lower density) regions. One reason for this density issue is its effect on the efficiency of immunoprecipitation. In their study, Down *et al.* developed a tool to estimate absolute methylation levels from data generated by MeDIP by modeling the density of methylated CpG sites. This tool is called Bayesian tool for methylation analysis (Batman). The study reports the coverage of ~90% of all CpG sites in promoters, gene-coding regions, islands, and regulatory elements where methylation levels can be estimated; this is almost 20 times better coverage than any previous methods.

Studies using MeDIP-seq or MeDIP-chip are both genome-wide approaches that have the common aim of obtaining the functional mapping of the methylome. Once regions of DNA methylation are identified, a number of bioinformatics analyses can be applied to answer certain biological questions. One obvious step is to investigate genes contained in these regions and investigate the functional significance of their repression. For example, silencing of tumour-suppressor genes in cancer can be attributed to DNA methylation. By identifying mutational events leading to hypermethylation and subsequent repression of known tumour-suppressor genes, one can more specifically characterize the contributing factors to the cause of the disease. Alternatively, one can identify genes that are known to be normally methylated but, as a result of some mutation event, is no longer silenced.

Also, one can try and investigate and identify whether some epigenetic regulator has been affected such as DNA methyltransferase (DNMT); in these cases, enrichment may be more limited.

Limitations of MeDIP

Limitations to take note when using MeDIP are typical experimental factors. This includes the quality and cross-reactivity of 5mC antibodies used in the procedure. Furthermore, DNA detection methods (i.e. array hybridization and high-throughput sequencing) typically involve well established limitations. Particularly for array-based procedures, as mentioned above, sequences being analyzed are limited to the specific array design used.

Most typical limitations to high-throughput, next generation sequencing apply. The problem of alignment accuracy to repetitive regions in the genome will result in less accurate analysis of methylation in those regions. Also, as was mentioned above, short reads (e.g. 36-50bp from an Illumina Genome Analyzer) represent a part of a sheared fragment when aligned to the genome; therefore, the exact methylation site can fall anywhere within a window that is a function of the fragment size. In this respect, bisulfite sequencing has much higher resolution (down to a single CpG site; single nucleotide level). However, this level of resolution may not be required for most applications, as the methylation status of CpG sites within < 1000 bp has been shown to be significantly correlated.

Applications of MeDIP

- Weber *et al.* 2005 determined that the inactive X-chromosome in females is hypermethylated on a chromosome wide level using MeDIP coupled with microarray.

- Keshet *et al.* 2006 performed a study on colon and prostate cancer cells using MeDIP-chip. The result is a genome-wide analysis of genes lying in hypermethylated regions as well as conclude that there is an instructive mechanism of de novo methylation in cancer cells.

- Zhang *et al.* 2006 obtained a high resolution methylome mapping in Arabidopsis using MeDIP-chip.
- Novak *et al.* 2006 used the MeDIP-chip approach to investigate human breast cancer for methylation associated silencing and observed the inactivation of the HOXA gene cluster

Chapter 5

Bisulfite Sequencing

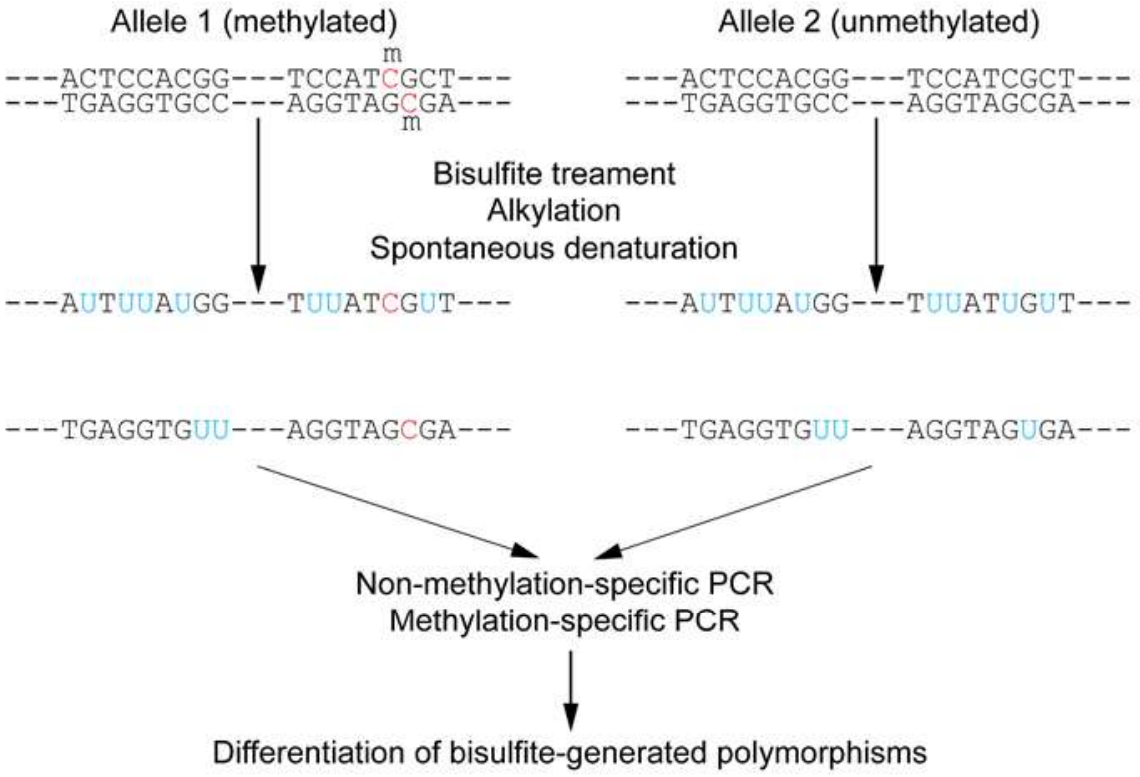


Figure 1: Outline of bisulfite conversion of sample sequence of genomic DNA. Nucleotides in blue are unmethylated cytosines converted to uracils by bisulfite, while red nucleotides are 5-methylcytosines resistant to conversion.

Bisulfite sequencing is the use of bisulfite treatment of DNA to determine its pattern of methylation. DNA methylation was the first discovered epigenetic mark, and remains the most studied. In animals it predominantly involves the addition of a methyl group to the carbon-5 position of cytosine residues of the dinucleotide CpG, and is implicated in repression of transcriptional activity.

Treatment of DNA with bisulfite converts cytosine residues to uracil, but leaves 5-methylcytosine residues unaffected. Thus, bisulfite treatment introduces specific changes in the DNA sequence that depend on the methylation status of individual cytosine residues, yielding single- nucleotide resolution information about the methylation status of a segment of DNA. Various analyses can be performed on the altered sequence to retrieve this information. The objective of this analysis is therefore reduced to differentiating between single nucleotide polymorphisms (cytosines and thymidine) resulting from bisulfite conversion (Figure 1).

Methods

Bisulfite sequencing applies routine sequencing methods on bisulfite-treated genomic DNA to determine methylation status at CpG dinucleotides. Other non-sequencing strategies are also employed to interrogate the methylation at specific loci or at a genome-wide level. All strategies assume that bisulfite-induced conversion of unmethylated cytosines to uracil is complete, and this serves as the basis of all subsequent techniques. Ideally, the method used would determine the methylation status separately for each allele. Alternative methods to bisulfite sequencing include Combined Bisulfite Restriction Analysis and methylated DNA immunoprecipitation (MeDIP).

Methodologies to analyze bisulfite-treated DNA are continuously being developed. To summarize these rapidly evolving methologies, numerous review articles have been written.

The methodologies can be generally divided into strategies based on methylation-specific PCR (MSP) (Figure 3), and strategies employing polymerase chain reaction (PCR) performed under non-methylation-specific conditions (Figure 2). Microarray-based methods use PCR based on non-methylation-specific conditions also.

Non-methylation-specific PCR based methods

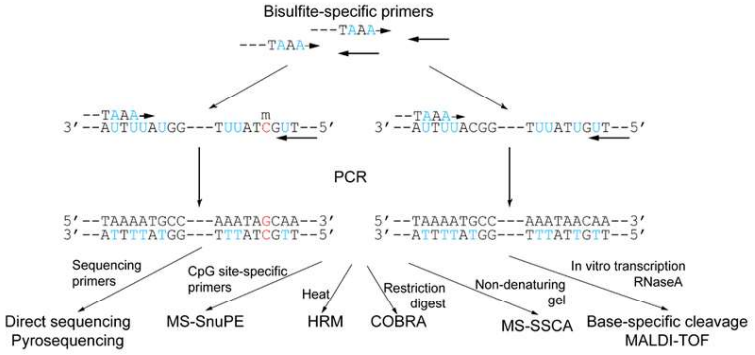


Figure 2: DNA methylation analysis methods not based on methylation-specific PCR. Following bisulfite conversion, the genomic DNA is amplified with PCR that does not discriminate between methylated and non-methylated sequences. The numerous methods available are then used to make the discrimination based on the changes within the amplicon as a result of bisulfite conversion.

Direct sequencing

The first reported method of methylation analysis using bisulfite-treated DNA utilized PCR and standard dideoxynucleotide DNA sequencing to directly determine the nucleotides resistant to bisulfite conversion. Primers are designed to be strand-specific as well as bisulfite-specific (i.e., primers containing non-CpG cytosines such that they are not complementary to non-bisulfite-treated DNA), flanking (but not involving) the methylation site of interest. Therefore, it will amplify both methylated and unmethylated sequences, in contrast to methylation-specific PCR. All sites of unmethylated cytosines are displayed as thymines in the resulting amplified sequence of the sense strand, and as adenines in the amplified antisense strand. This technique required cloning of the PCR product prior to sequencing for adequate sensitivity, and therefore was a very labour-intensive method unsuitable for higher throughput. Alternatively, nested PCR methods can be used to enhance the product for sequencing.

All subsequent DNA methylation analysis techniques using bisulfite-treated DNA is based on this report by Frommer et al. (Figure 2). Although most other modalities are not true sequencing-based techniques, the term "bisulfite sequencing" is often used to describe bisulfite-conversion DNA methylation analysis techniques in general.

Pyrosequencing

Pyrosequencing has also been used to analyze bisulfite-treated DNA without using methylation-specific PCR. Following PCR amplification of the region of interest, Pyrosequencing is used to determine the bisulfite-converted sequence of specific CpG sites in the region. The ratio of C-to-T at individual sites can be determined quantitatively based on the amount of C and T incorporation during the sequence extension. The main limitation of this method is the cost of the technology. However, Pyrosequencing does well allow for extension to high-throughput screening methods.

A further improvement to this technique was recently described by Wong et al., which uses allele-specific primers that incorporate single-nucleotide polymorphisms into the sequence of the sequencing primer, thus allowing for separate analysis of maternal and paternal alleles. This technique is of particular usefulness for genomic imprinting analysis.

Methylation-sensitive single-strand conformation analysis (MS-SSCA)

This method is based on the single-strand conformation polymorphism analysis (SSCA) method developed for single-nucleotide polymorphism (SNP) analysis. SSCA differentiates between single-stranded DNA fragments of identical size but distinct sequence based on differential migration in non-denaturating electrophoresis. In MS-SSCA, this is used to distinguish between bisulfite-treated, PCR-amplified regions containing the CpG sites of interest. Although SSCA lacks sensitivity when only a single nucleotide difference is present, bisulfite treatment frequently makes a number of C-to-T conversions in most regions of interest, and the resulting sensitivity approaches 100%.

MS-SSCA also provides semi-quantitative analysis of the degree of DNA methylation based on the ratio of band intensities. However, this method is designed to assess all CpG sites as a whole in the region of interest rather than individual methylation sites.

High resolution melting analysis (HRM)

A further method to differentiate converted from unconverted bisulfite-treated DNA is using high-resolution melting analysis (HRM), a real-time PCR-based technique initially designed to distinguish SNPs. The PCR amplicons are analyzed directly by temperature ramping and resulting liberation of an intercalating fluorescent dye during melting. The degree of methylation, as represented by the C-to-T content in the amplicon, determines the rapidity of melting and consequent release of the dye. This method allows direct quantitation in a single-tube assay, but assesses methylation in the amplified region as a whole rather than at specific CpG sites.

Methylation-sensitive single-nucleotide primer extension (MS-SnuPE)

MS-SnuPE employs the primer extension method initially designed for analyzing single-nucleotide polymorphisms. DNA is bisulfite-converted, and bisulfite-specific primers are annealed to the sequence up to the base pair immediately before the CpG of interest. The primer is allowed to extend one base pair into the C (or T) using DNA polymerase terminating dideoxynucleotides, and the ratio of C to T is determined quantitatively.

A number of methods can be used to determine this C:T ratio. At the beginning, MS-SnuPE relied on radioactive ddNTPs as the reporter of the primer extension. Fluorescence-based methods or Pyrosequencing can also be used. However, matrix-assisted laser desorption ionization/time-of-flight (MALDI-TOF) mass spectrometry analysis to differentiate between the two polymorphic primer extension products can be used, in essence, based on the GOOD assay designed for SNP genotyping. Ion pair reverse-phase high-performance liquid chromatography (IP-RP-HPLC) has also been used to distinguish primer extension products.

Base-specific cleavage/MALDI-TOF

A recently described method by Ehrich et al. further takes advantage of bisulfite-conversions by adding a base-specific cleavage step to enhance the information gained from the nucleotide changes. By first using in vitro transcription of the region of interest into RNA (by adding an RNA polymerase promoter site to the PCR primer in the initial amplification), RNase A can be used to cleave the RNA transcript at base-specific sites. As RNase A cleaves RNA specifically at cytosine and uracil ribonucleotides, base-specificity is achieved by adding incorporating cleavage-resistant dTTP when cytosine-specific (C-specific) cleavage is desired, and incorporating dCTP when uracil-specific (U-specific) cleavage is desired. The cleaved fragments can then be analyzed by MALDI-TOF. Bisulfite treatment results in either introduction/removal of cleavage sites by C-to-U conversions or shift in fragment mass by G-to-A conversions in the amplified reverse strand. C-specific cleavage will cut specifically at all methylated CpG sites. By analyzing

the sizes of the resulting fragments, it is possible to determine the specific pattern of DNA methylation of CpG sites within the region, rather than determining the extent of methylation of the region as a whole. This method demonstrated efficacy for high-throughput screening, allowing for interrogation of numerous CpG sites in multiple tissues in a cost-efficient manner.

Methylation-specific PCR (MSP)

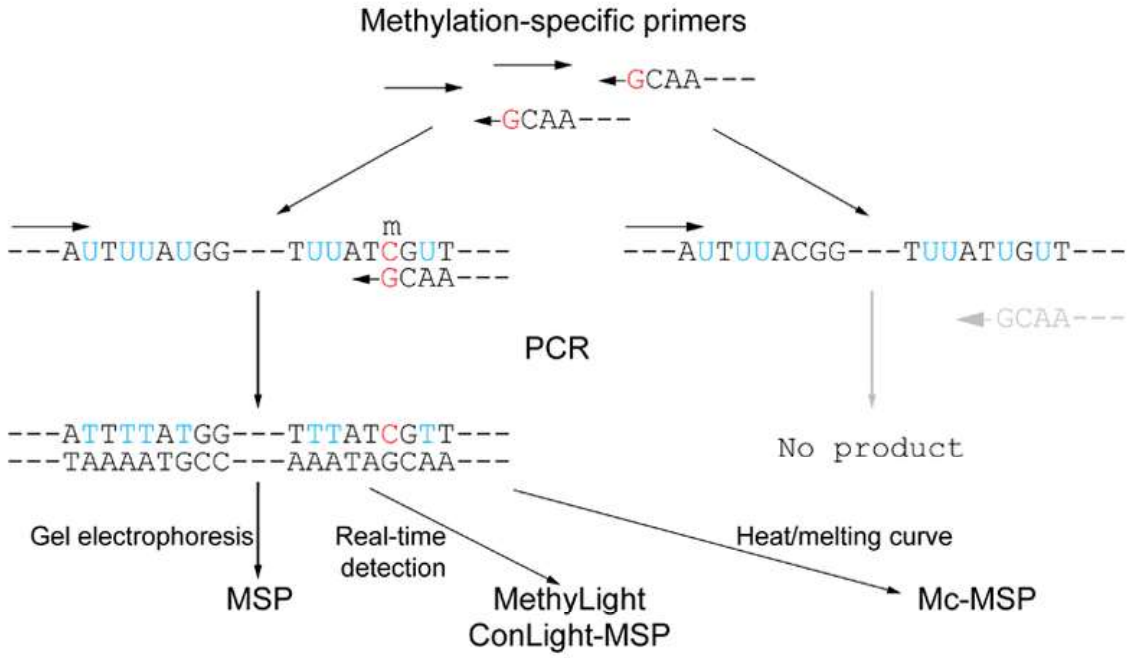


Figure 3: Methylation-specific PCR is a sensitive method to discriminately amplify and detect a methylated region of interest using methylated-specific primers on bisulfite-converted genomic DNA. Such primers will anneal only to sequences that are methylated, and thus containing 5-methylcytosines that are resistant to conversion by bisulfite. In alternative fashion, unmethylated-specific primers can be used.

This alternative method of methylation analysis also uses bisulfite-treated DNA but avoids the need to sequence the area of interest. Instead, primer pairs are designed themselves to be "methylated-specific" by including sequences complementing only unconverted 5-methylcytosines, or, on the converse, "unmethylated-specific", complementing thymines converted from unmethylated cytosines. Methylation is determined by the ability of the specific primer to achieve amplification. This method is particularly useful to interrogate CpG islands with possibly high methylation density, as increased numbers of CpG pairs in the primer increase the specificity of the assay. Placing the CpG pair at the 3'-end of the primer also improves the sensitivity. The initial report using MSP described sufficient sensitivity to detect methylation of 0.1% of alleles. In general, MSP and its related protocols are considered to be the most sensitive when interrogating the methylation status at a specific locus.

The MethyLight method is based on MSP, but provides a quantitative analysis using real-time PCR. Methylated-specific primers are used, and a methylated-specific fluorescence reporter probe is also used that anneals to the amplified region. In alternative fashion, the primers or probe can be designed without methylation specificity if discrimination is needed between the CpG pairs within the involved sequences. Quantitation is made in reference to a methylated reference DNA. A modification to this protocol to increase the specificity of the PCR for successfully bisulfite-converted DNA (ConLight-MSP) uses an additional probe to bisulfite-unconverted DNA to quantify this non-specific amplification.

Further methodology using MSP-amplified DNA analyzes the products using melting-curve analysis (Mc-MSP). This method amplifies bisulfite-converted DNA with both methylated-specific and unmethylated-specific primers, and determines the quantitative ratio of the two products by comparing the differential peaks generated in a melting-curve analysis. A high-resolution melting analysis method that uses both real-time quantification and melting analysis has been introduced, in particular, for sensitive detection of low-level methylation

Microarray-based methods

Microarray-based methods are a logical extension of the technologies available to analyze bisulfite-treated DNA to allow for genome-wide analysis of methylation. Oligonucleotide microarrays are designed using oligonucleotide pairs targeting CpG sites of interest, with one complementary to the unaltered methylated sequence, and the other to the C-to-U-converted unmethylated sequence. The oligonucleotides are also bisulfite-specific to prevent binding to DNA incompletely converted by bisulfite. The Illumina Methylation Assay is one such assay that applies the bisulfite sequencing technology on a microarray level to generate genome-wide methylation data.

Limitations

Incomplete conversion

Bisulfite sequencing relies on the conversion of every single unmethylated cytosine residue to uracil. If conversion is incomplete, the subsequent analysis will incorrectly interpret the unconverted unmethylated cytosines as methylated cytosines, resulting in false positive results for methylation. Only cytosines in single-stranded DNA are susceptible to attack by bisulfite, therefore denaturation of the DNA undergoing analysis is critical. It is important to ensure that reaction parameters such as temperature and salt concentration are suitable to maintain the DNA in a single-stranded conformation and allow for complete conversion. Embedding the DNA in agarose gel has been reported to improve the rate of conversion by keeping strands of DNA physically separate.

Degradation of DNA during bisulfite treatment

A major challenge in bisulfite sequencing is the degradation of DNA that takes place concurrently with the conversion. The conditions necessary for complete conversion, such as long incubation times, elevated temperature, and high bisulfite concentration, can lead to the degradation of about 90% of the incubated DNA. Given that the starting amount of DNA is often limited, such extensive degradation can be problematic. The degradation occurs as depurinations resulting in random strand breaks. Therefore the longer the desired PCR amplicon, the more limited the number of intact template molecules will likely be. This could lead to the failure of the PCR amplification, or the loss of quantitatively accurate information on methylation levels resulting from the limited sampling of template molecules. Thus, it is important to assess the amount of DNA degradation resulting from the reaction conditions employed, and consider how this will affect the desired amplicon. Techniques can also be used to minimize DNA degradation, such as cycling the incubation temperature.

Other concerns

A potentially significant problem following bisulfite treatment is incomplete desulfonation of pyrimidine residues due to inadequate alkalization of the solution. This may inhibit some DNA polymerases, rendering subsequent PCR difficult. However, this situation can be avoided by monitoring the pH of the solution to ensure that desulphonation will be complete.

A final concern is that bisulfite treatment greatly reduces the level of complexity in the sample, which can be problematic if multiple PCR reactions are to be performed (2006). Primer design is more difficult, and inappropriate cross-hybridization is more frequent.

Applications: genome-wide methylation analysis

The advances in bisulfite sequencing have led to the possibility of applying them at a genome-wide scale, where, previously, global measure of DNA methylation was feasible only using other techniques, such as Restriction landmark genomic scanning. The mapping of the human epigenome is seen by many scientists as the logical follow-up to the completion of the Human Genome Project. This epigenomic information will be important in understanding how the function of the genetic sequence is implemented and regulated. Since the epigenome is less stable than the genome, it is thought to be important in gene-environment interactions.

Epigenomic mapping is inherently more complex than genome sequencing, however, since the epigenome is much more variable than the genome. While an individual has only one genome, one's epigenome varies with age, differs between tissues, is altered by environmental factors, and shows aberrations in diseases. Such rich epigenomic mapping, however, representing different ages, tissue types, and disease states, would yield valuable information on the normal function of epigenetic marks as well as the mechanisms leading to aging and disease.

Direct benefits of epigenomic mapping include probable advances in cloning technology. It is believed that failures to produce cloned animals with normal viability and lifespan result from inappropriate patterns of epigenetic marks. Also, aberrant methylation patterns are well characterized in many cancers. Global hypomethylation results in decreased genomic stability, while local hypermethylation of tumour suppressor gene promoters often accounts for their loss of function. Specific patterns of methylation are indicative of specific cancer types, have prognostic value, and can help to guide the best course of treatment.

Large-scale epigenome mapping efforts are under way around the world and have been organized under the Human Epigenome Project. This is based on a multi-tiered strategy, whereby bisulfite sequencing is used to obtain high-resolution methylation profiles for a limited number of reference epigenomes, while less thorough analysis is performed on a wider spectrum of samples. This approach is intended to maximize the insight gained from a given amount of resources, as high-resolution genome-wide mapping remains a costly undertaking.

Chapter 6

DNA Methylation

DNA methylation involves the addition of a methyl group to the 5 position of the cytosine pyrimidine ring or the number 6 nitrogen of the adenine purine ring (cytosine and adenine are two of the four bases of DNA). This modification can be inherited through cell division. DNA methylation is typically removed during zygote formation and re-established through successive cell divisions during development although the latest research shows that hydroxylation of methyl group occurs rather than complete removal of methyl groups in zygote . DNA methylation is a crucial part of normal organismal development and cellular differentiation in higher organisms. DNA methylation stably alters the gene expression pattern in cells such that cells can "remember where they have been" or decrease gene expression; for example, cells programmed to be pancreatic islets during embryonic development remain pancreatic islets throughout the life of the organism without continuing signals telling them that they need to remain islets. In addition, DNA methylation suppresses the expression of viral genes and other deleterious elements that have been incorporated into the genome of the host over time. DNA methylation also forms the basis of chromatin structure, which enables cells to form the myriad characteristics necessary for multicellular life from a single immutable sequence of DNA. DNA methylation also plays a crucial role in the development of nearly all types of cancer.

DNA methylation involves the addition of a methyl group to DNA — for example, to the number 5 carbon of the cytosine pyrimidine ring — in this case with the specific effect of reducing gene expression. DNA methylation at the 5 position of cytosine has been found in every vertebrate examined. In adult somatic tissues, DNA methylation typically occurs in a CpG dinucleotide context; non-CpG methylation is prevalent in embryonic stem cells.

In mammals

DNA methylation is essential for normal development and is associated with a number of key processes including genomic imprinting, X-chromosome inactivation, suppression of repetitive elements, and carcinogenesis.

Between 60% and 90% of all CpGs are methylated in mammals. Methylated C residues spontaneously deaminate to form T residues; hence CpG dinucleotides steadily mutate to TpG dinucleotides, which is evidenced by the under-representation of CpG dinucleotides in the human genome (they occur at only 21% of the expected frequency). (On the other hand, spontaneous deamination of unmethylated C residues gives rise to U residues, a mutation that is quickly recognized and repaired by the cell.)

Unmethylated CpGs are often grouped in clusters called *CpG islands*, which are present in the 5' regulatory regions of many genes. In many disease processes, such as cancer, gene promoter CpG islands acquire abnormal hypermethylation, which results in transcriptional silencing that can be inherited by daughter cells following cell division. Alterations of DNA methylation have been recognized as an important component of cancer development. Hypomethylation, in general, arises earlier and is linked to chromosomal instability and loss of imprinting, whereas hypermethylation is associated with promoters and can arise secondary to gene (oncogene suppressor) silencing, but might be a target for epigenetic therapy.

DNA methylation may affect the transcription of genes in two ways. First, the methylation of DNA itself may physically impede the binding of transcriptional proteins to the gene, and second, and likely more important, methylated DNA may be bound by proteins known as methyl-CpG-binding domain proteins (MBDs). MBD proteins then recruit additional proteins to the locus, such as histone deacetylases and other chromatin remodeling proteins that can modify histones, thereby forming compact, inactive chromatin, termed silent chromatin. This link between DNA methylation and chromatin structure is very important. In particular, loss of methyl-CpG-binding protein 2 (MeCP2) has been implicated in Rett syndrome; and methyl-CpG-binding domain protein 2 (MBD2) mediates the transcriptional silencing of hypermethylated genes in cancer.

Research has suggested that long-term memory storage in humans may be regulated by DNA methylation.

DNA methylation in cancer

DNA methylation is an important regulator of gene transcription and a large body of evidence has demonstrated that aberrant DNA methylation is associated with unscheduled gene silencing, and the genes with high levels of 5-methylcytosine in their promoter region are transcriptionally silent. DNA methylation is essential during embryonic development, and in somatic cells, patterns of DNA methylation are generally transmitted to daughter cells with a high fidelity. Aberrant DNA methylation patterns have been associated with a large number of human malignancies and found in two distinct forms: hypermethylation and hypomethylation compared to normal tissue. Hypermethylation is one of the major epigenetic modifications that repress transcription via promoter region of tumour suppressor genes. Hypermethylation typically occurs at CpG islands in the promoter region and is associated with gene inactivation. Global hypomethylation has also been implicated in the development and progression of cancer through different mechanisms.

DNA methyltransferases

In mammalian cells, DNA methylation occurs mainly at the C5 position of CpG dinucleotides and is carried out by two general classes of enzymatic activities – maintenance methylation and *de novo* methylation.

Maintenance methylation activity is necessary to preserve DNA methylation after every cellular DNA replication cycle. Without the DNA methyltransferase (DNMT), the replication machinery itself would produce daughter strands that are unmethylated and, over time, would lead to passive demethylation. DNMT1 is the proposed maintenance methyltransferase that is responsible for copying DNA methylation patterns to the daughter strands during DNA replication. Mouse models with both copies of DNMT1 deleted are embryonic lethal at approximately day 9, due to the requirement of DNMT1 activity for development in mammalian cells.

It is thought that DNMT3a and DNMT3b are the *de novo* methyltransferases that set up DNA methylation patterns early in development. DNMT3L is a protein that is homologous to the other DNMT3s but has no catalytic activity. Instead, DNMT3L assists the *de novo* methyltransferases by increasing their ability to bind to DNA and stimulating their activity. Finally, DNMT2 (TRDMT1) has been identified as a DNA methyltransferase homolog, containing all 10 sequence motifs common to all DNA methyltransferases; however, DNMT2 (TRDMT1) does not methylate DNA but instead methylates cytosine-38 in the anticodon loop of aspartic acid transfer RNA.

Since many tumor suppressor genes are silenced by DNA methylation during carcinogenesis, there have been attempts to re-express these genes by inhibiting the DNMTs. 5-Aza-2'-deoxycytidine (decitabine) is a nucleoside analog that inhibits DNMTs by trapping them in a covalent complex on DNA by preventing the β-elimination step of catalysis, thus resulting in the enzymes' degradation. However, for decitabine to be active, it must be incorporated into the genome of the cell, which can cause mutations in the daughter cells if the cell does not die. In addition, decitabine is toxic to the bone marrow, which limits the size of its therapeutic window. These pitfalls have led to the development of antisense RNA therapies that target the DNMTs by degrading their mRNAs and preventing their translation. However, it is currently unclear whether targeting DNMT1 alone is sufficient to reactivate tumor suppressor genes silenced by DNA methylation.

In plants

Significant progress has been made in understanding DNA methylation in the model plant *Arabidopsis thaliana*. DNA methylation in plants differs from that of mammals: while DNA methylation in mammals mainly occurs on the cytosine nucleotide in a CpG site, in plants the cytosine can be methylated at CpG, CpHpG, and CpHpH sites, where H represents any nucleotide but guanine.

The principal *Arabidopsis* DNA methyltransferase enzymes, which transfer and covalently attach methyl groups onto DNA, are DRM2, MET1, and CMT3. Both the DRM2 and MET1 proteins share significant homology to the mammalian methyltransferases DNMT3 and DNMT1, respectively, whereas the CMT3 protein is unique to the plant kingdom. There are currently two classes of DNA methyltransferases: 1) the *de novo* class, or enzymes that create new methylation marks on the DNA; and 2) a maintenance class that recognizes the methylation marks on the parental strand of DNA and transfers new methylation to the daughters strands after DNA replication. DRM2 is the only enzyme that has been implicated as a *de novo* DNA methyltransferase. DRM2 has also been shown, along with MET1 and CMT3 to be involved in maintaining methylation marks through DNA replication. Other DNA methyltransferases are expressed in plants but have no known function.

It is not clear how the cell determines the locations of *de novo* DNA methylation, but evidence suggests that, for many (though not all) locations, RNA-directed DNA methylation (RdDM) is involved. In RdDM, specific RNA transcripts are produced from a genomic DNA template, and this RNA forms secondary structures called double-stranded RNA molecules. The double-stranded RNAs, through either the small interfering RNA (siRNA) or microRNA (miRNA) pathways direct de-novo DNA methylation of the original genomic location that produced the RNA. This sort of mechanism is thought to be important in cellular defense against RNA viruses and/or transposons, both of which often form a double-stranded RNA that can be mutagenic to the host genome. By methylating their genomic locations, through an as yet poorly-understood mechanism, they are shut off and are no longer active in the cell, protecting the genome from their mutagenic effect.

In fungi

It can be seen that many fungi have low levels (0.1 to 0.5%) of cytosine methylation, whereas other fungi have as much as 5% of the genome methylated.

This value seems to vary both among species and among isolates of the same species. There is also evidence that DNA methylation may be involved in state-specific control of gene expression in fungi.

Although brewers' yeast (*Saccharomyces*) and fission yeast (*Schizosaccharomyces*) have very little DNA methylation, the model filamentous fungus *Neurospora crassa* has a well-characterized methylation system. Several genes control methylation in *Neurospora* and mutation of the DNA methyl transferase, *dim-2*, eliminates all DNA methylation but does not affect growth or sexual reproduction. While the *Neurospora* genome has very little repeated DNA, half of the methylation occurs in repeated DNA including transposon relics and centromeric DNA. The ability to evaluate other important phenomena in a DNA methylase-deficient genetic background makes *Neurospora* an important system in which to study DNA methylation.

In bacteria

Adenine or cytosine methylation is part of the restriction modification system of many bacteria, in which specific DNA sequences are methylated periodically throughout the genome. A methylase is the enzyme that recognizes a specific sequence and methylates one of the bases in or near that sequence. Foreign DNAs (which are not methylated in this manner) that are introduced into the cell are degraded by sequence-specific restriction enzymes and cleaved. Bacterial genomic DNA is not recognized by these restriction enzymes. The methylation of native DNA acts as a sort of primitive immune system, allowing the bacteria to protect themselves from infection by bacteriophage.

E. coli DNA adenine methyltransferase (Dam) is an enzyme of ~32 kDa that does not belong to a restriction/modification system. The target recognition sequence for *E. coli* Dam is GATC, as the methylation occurs at the N6 position of the adenine in this sequence (G meATC). The three base pairs flanking each side of this site also influence DNA–Dam binding. Dam plays several key roles in bacterial processes, including mismatch repair, the timing of DNA replication, and gene expression. As a result of DNA replication, the status of GATC sites in the *E. coli* genome changes from fully methylated to hemimethylated. This is because adenine introduced into the new DNA strand is unmethylated. Re-methylation occurs within two to four seconds, during which time replication errors in the new strand are repaired. Methylation, or its absence, is the marker that allows the repair apparatus of the cell to differentiate between the template and nascent strands. It has been shown that altering Dam activity in bacteria results in increased spontaneous mutation rate. Bacterial viability is compromised in dam mutants that also lack certain other DNA repair enzymes, providing further evidence for the role of Dam in DNA repair.

One region of the DNA that keeps its hemimethylated status for longer is the origin of replication, which has an abundance of GATC sites. This is central to the bacterial mechanism for timing DNA replication. SeqA binds to the origin of replication, sequestering it and thus preventing methylation. Because hemimethylated origins of replication are inactive, this mechanism limits DNA replication to once per cell cycle.

Expression of certain genes, for example those coding for pilus expression in *E. coli*, is regulated by the methylation of GATC sites in the promoter region of the gene operon. The cells' environmental conditions just after DNA replication determine whether Dam is blocked from methylating a region proximal to or distal from the promoter region. Once the pattern of methylation has been created, the pilus gene transcription is locked in the on or off position until the DNA is again replicated. In *E. coli*, these pilus operons have important roles in virulence in urinary tract infections. It has been proposed that inhibitors of Dam may function as antibiotics.

On the other hand DNA cytosine methylase targets CCAGG and CCTGG sites to methylate cytosine at the C5 position (C meC(A/T)GG). The other methylase enzyme, EcoKI, causes mehtylation of adenine in the sequences AAC(N6A)GTGC and GCAC(N6A)GTT.

Most strains used by molecular biologists are derivatives of K-12, and possess both Dam and Dcm, but there are commercially available strains which possess dam-/dcm- activity. In fact, it is possible to unmethylate the DNA extracted from dam+/dcm+ strains by transforming into dam-/dcm- strains. This would help digest sequences that are not being recognized by methylation-sensitive restriction enzymes.

Detection

DNA methylation can be detected by the following assays currently used in scientific research:

- Methylation-Specific PCR (MSP), which is based on a chemical reaction of sodium bisulfite with DNA that converts unmethylated cytosines of CpG dinucleotides to uracil or UpG, followed by traditional PCR. However, methylated cytosines will not be converted in this process, and primers are designed to overlap the CpG site of interest, which allows one to determine methylation status as methylated or unmethylated.
- The HELP assay, which is based on restriction enzymes' differential ability to recognize and cleave methylated and unmethylated CpG DNA sites.
- ChIP-on-chip assays, which is based on the ability of commercially prepared antibodies to bind to DNA methylation-associated proteins like MCP2.
- Restriction landmark genomic scanning, a complicated and now rarely-used assay based upon restriction enzymes' differential recognition of methylated and unmethylated CpG sites; the assay is similar in concept to the HELP assay.
- Methylated DNA immunoprecipitation (MeDIP), analogous to chromatin immunoprecipitation, immunoprecipitation is used to isolate methylated DNA fragments for input into DNA detection methods such as DNA microarrays (MeDIP-chip) or DNA sequencing (MeDIP-seq).
- Molecular break light assay for DNA adenine methyltransferase activity – an assay that relies on the specificity of the restriction enzyme DpnI for fully methylated (adenine methylation) GATC sites in an oligonucleotide labeled with a fluorophore and quencher. The adenine methyltransferase methylates the oligonucleotide making it a substrate for DpnI. Cutting of the oligonucleotide by DpnI gives rise to a fluorescence increase.

Chapter 7

Nutriepigenomics

Nutriepigenomics is the study of food nutrients and their effects on human health through epigenetic modifications. There is now considerable evidence that nutritional imbalances during gestation and lactation are linked to non-communicable diseases, such as obesity, cardiovascular disease, diabetes, hypertension, and cancer. If metabolic disturbances occur during critical time windows of development, the resulting epigenetic alterations can lead to permanent changes in tissue and organ structure or function and predispose individuals to disease

Overview

Epigenetics relates to heritable changes in gene function that occur independently of alterations in primary DNA sequence. Two major epigenetic mechanisms implicated in nutriepigenomics are DNA methylation and histone modification. DNA methylation in gene promoter regions usually results in gene silencing and influences gene expression. While this form of gene silencing is extremely important in development and cellular differentiation, aberrant DNA methylation can be detrimental and has been linked to various disease processes, such as cancer . The methyl groups used in DNA methylation are often derived from dietary sources, such as folate and choline, and explains why diet can have a significant impact on methylation patterns and gene expression . Gene silencing can also be reinforced through the recruitment of histone deacetylases to decrease transcriptional activation. Conversely, histone acetylation induces transcriptional activation to increase gene expression. Dietary components can influence these epigenetic events, thereby altering gene expression and disturbing functions such as appetite control, metabolic balance and fuel utilization .

Various genetic sequences can be targeted for epigenetic modification. A transcriptome-wide analysis in mice found that a protein-restricted (PR) diet during gestation resulted in differential gene expression in approximately 1% of the fetal genes analyzed (235/22,690). Specifically, increased expression was seen in genes involved in the p53 pathway, apoptosis, negative regulators of cell metabolism, and genes related to epigenetic control . Additional studies have investigated the effect of a PR-diet in rats and

found changes in promoter methylation of both the glucocorticoid receptor and peroxisome proliferator-activated receptor (PPAR) . Altered expression of these receptors can result in elevated blood glucose levels and affect lipid and carbohydrate metabolism . Feeding a PR-diet to pregnant and/or lactating mice also increased expression of glucokinase, acetyl-CoA carboxylase, PPARα, and acyl-CoA oxidase . Changes in expression were reportedly due to epigenetic regulation of either the gene promoter itself, or promoters of transcription factors that regulate gene expression. Additional genes that have been shown, either by in vitro or in vivo studies, to be regulated by epigenetic mechanisms include leptin, SOCS3, glucose transporter (GLUT)-4, POMC, 11-β-hydroxysteroid dehydrogenase type 2 and corticotrophin releasing hormone. Epigenetic modification of these genes may lead to "metabolic programming" of the fetus and result in long-term changes in metabolism and energy homeostasis .

Nutriepigenomics and Development

The period of development in which the nutritional imbalance occurs is very important in determining which disease-related genes will be affected. Different organs have critical developmental stages, and the time point at which they are compromised will predispose individuals to specific diseases . Epigenetic modifications that occur during development may not be expressed until later in life depending on the function of the gene . While the majority of studies implicate prenatal and perinatal periods as critical time windows, some research has shown that nutritional intake during adulthood can also affect the epigenome.

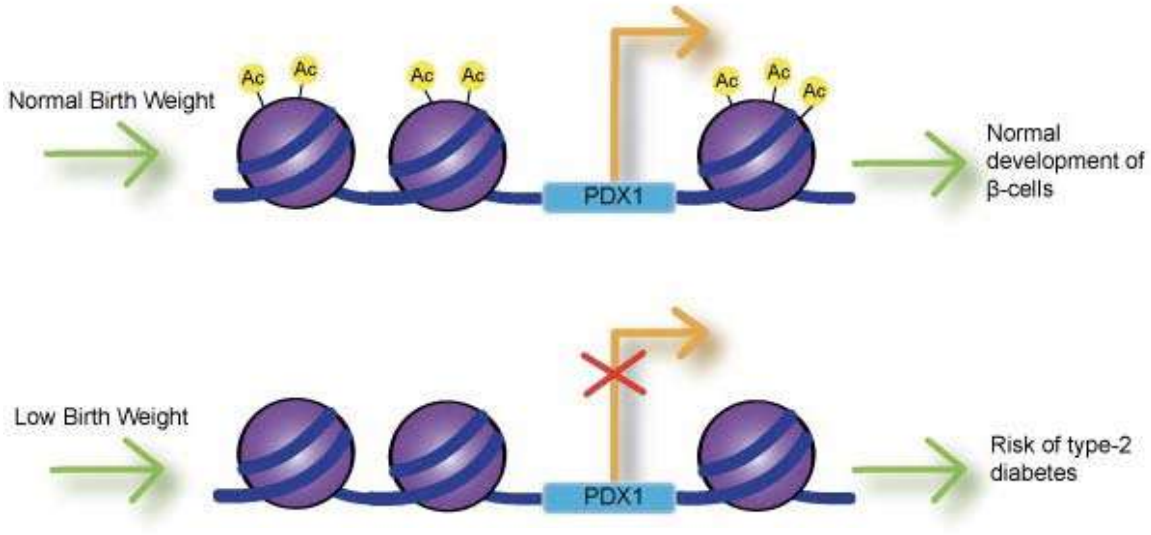
Prenatal

Developmental plasticity is a term used to describe the process in which fetuses adapt to their environment. Environmental cues, including dietary components, present in the in utero environment can induce significant changes in the expression of the genome through epigenetic modifications . Fetal developmental plastic responses can cause changes in lean body mass, endocrinology, blood flow and vascular loading, and lead to increased risk of various diseases in adulthood.

Low Birth Weight

Fetal exposure to calcium, folate, magnesium, high or low protein, and zinc have all been associated with birth weight. Numerous studies have investigated the link between birth weight and risk of disease and have found that low birth weight is significantly associated with coronary heart disease, stroke and type-2 diabetes. Most importantly, these associations occurred after adjusting for lifestyle factors, implying a genetic basis for onset of disease . Impaired insulin secretion is associated with low birth weight and can lead to insulin resistance as babies accumulate body fat . Studies using intrauterine growth retarded (IUGR) rats have found that growth inhibition can lead to decreased expression of Pdx1 transcription factor, which is essential for differentiation and function of pancreatic beta cells . Decreased histone acetylation at the proximal promoter of Pdx1

is responsible for reduced Pdx1 expression and subsequently results in a cascade of histone deacetylation and methylation events that can result in type-2 diabetes.



Development of type 2 diabetes following intrauterine growth retardation in rats is associated with progressive epigenetic silencing of Pdx1

Obesity

Obesity during pregnancy and high-fat maternal diets both show strong associations with obesity in offspring. As the number of overweight reproductive-age women increases, the number of overweight children and infants also increases . It has been postulated that maternal obesity causes an accumulation of fat in fetal adipose tissue (adiposity) and predisposes babies for obesity in childhood and adulthood . Animal studies have shown that maternal overnutrition may impact brain development and cause disruptions to programming of the hypothalamus. Offspring that were exposed to a high-fat or high-caloric maternal diet had increased levels of insulin, glucose and leptin. It is hypothesized that these elevations are due to disturbances in the complex neuronal network that includes the neuropeptide Y (NPY) and proopiomelanocortin (POMC) pathways . This altered neuronal signaling can consequently impact food-intake behavior and lead to diet-induced obesity in adulthood. While epigenetic modifications are most likely involved in the development of obesity, the specific target genes have yet to be identified. Genes involved in adipogenesis, such as fibroblast growth-factor-2, phosphatase and tensin homologue, cyclin-dependent kinase inhibitor 1A and oestrogen receptor-alpha, possess multiple CpG islands in their promoter sites and may act as epigenetic targets. Furthermore, it has been shown that prenatal exposure to a hypomethylating agent, such as bisphenol A (BPA), is associated with increased body weight and suggests modified DNA methylation as a mechanism for increasing susceptibility to obesity.

Folate

It has long been realized that maternal folate intake during pregnancy is linked to fetal development and growth, and can reduce the risk of serious birth defects. Folate is a source of S-adenosyl methionine (SAM), which is used to supply DNA methyltransferases with methyl groups. Therefore, changes in folate supply have a substantial effect on DNA methylation patterns. Low levels of folate are associated with an increased risk of preterm delivery, poor growth of the placenta and uterus, and intrauterine growth retardation. Several complex diseases, including cancer, cardiovascular diseases and autism have also been linked to maternal folate status. Based on animal studies it has been hypothesized that reduced folate intake could increase the risk of neural tube defects by reducing the amount of methylayed DNA during cranial neural tube closure. Recently it was discovered that folate protection from congenital heart defects is linked to epigenetics and Wnt signaling. Multiple environmental factors target the Wnt signaling pathway during embryogenesis and can cause misregulation of the pathway. Folic acid metabolism generates SAM, thereby altering the methylation states of histones H3K9, H3K4, and H3K27 and genetically altering Wnt signaling.

Perinatal

Another critical developmental time window is the perinatal period, which refers to the time period immediately before and after birth. It has been shown that maternal diet in late pregnancy and an infant's diet in the beginning weeks can all have significant impacts on gene expression. Therefore, perinatal nutrition refers to both late-stage in utero nutrition and lactation.

Bone Health

Bone mass and the development of osteoporosis have been studied in relation to perinatal nutrition. An important factor to consider when investigating perinatal nutrition is whether the baby was breast-fed or formula-fed. Studies have shown that breast-fed babies have increased bone mass compared to those were not breast-fed, and that this small increase in bone mass during a period of critical development could potentially program the skeleton to continue along a "healthy" growth trajectory . It has also been shown that maternal vitamin D insufficiency during late pregnancy is associated with reduced bone size and mineral mass in late childhood. Peak bone mass has shown to be a good predictor of risk of fracture and osteoporosis, with even a small increase in peak bone mass resulting in a much lower risk of bone fracture . Research shows that genetic markers explain only a small proportion of variation in bone mass and risk of fracture. Therefore, healthy bone programming is most likely influenced by various epigenetic mechanisms, such as imprinting of the growth promoting genes IGF-2, or changes to the hypothalamic-pituitary-adrenal axis (HPA).

Neurodevelopment

Imbalances in maternal nutrition can also have a significant effect on fetal neurodevelopment. Brain development occurs most rapidly during fetal development and infancy, and research has shown that exposure to certain environmental conditions can have long lasting effects on cognition. Specifically, n-3 fatty acids, iodine, iron and choline have been shown to influence brain development and impact cognitive ability and behavior. The greatest evidence for a link between nutrition and neurodevelopment comes from studies that show low birth weight associated with low IQ and increased risk of schizophrenia. Several studies suggest that breast-feeding promotes long-term neurodevelopment by providing the nutrients necessary for proper brain development. A study in mice showed that choline-deficient diets during the late gestation period impaired fetal brain development, including decreased cell proliferation and reduced visual-spatial and auditory memory. These cognitive changes appeared to be due to altered histone and DNA methylation patterns in the fetal hippocampus, thus providing a link between maternal nutrition, epigenetics, and early brain development.

Type-1 Diabetes

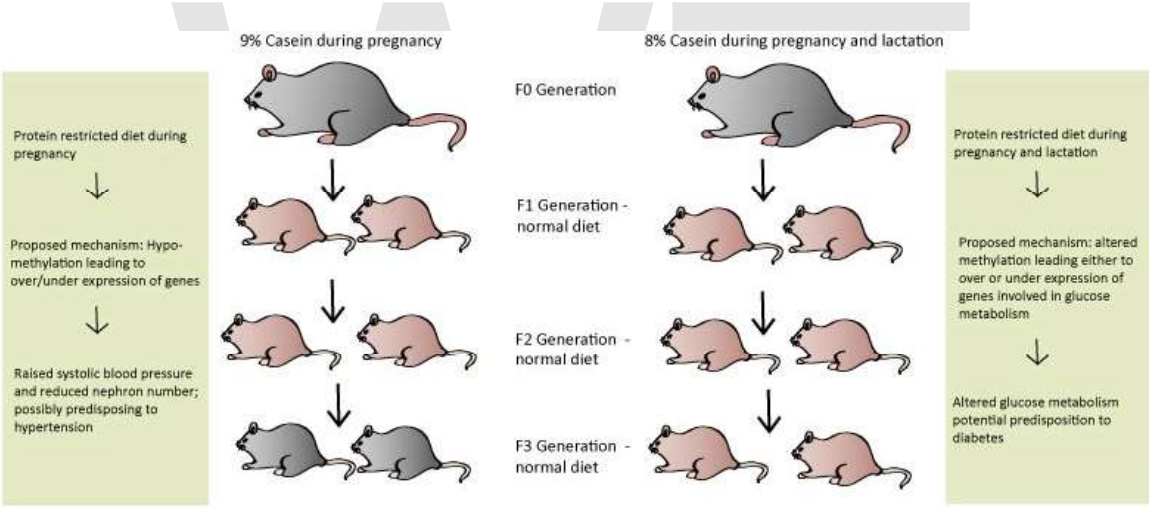
It has been postulated that breast-feeding may also protect against type-1 diabetes, with research showing that formula-fed infants are at an increased risk of developing islet autoantibodies. Individuals with type-1 diabetes experience a pre-clinical diabetes phase characterized by autoimmunity against pancreatic islets. The introduction of certain foods in the first few months of life, such as berries and cereal, is significantly associated with increased risk of islet autoantibody development compared to babies who are exposed to solid foods later in life. While the pathogenesis behind development of autoantibodies remains largely unknown, it is very probable that an epigenetic link exists between perinatal diet and risk of type-1 diabetes.

Adulthood

The majority of research in nutriepigenomics has focused on nutritional imbalances during gestation and lactation periods. However, foods that are consumed during adulthood can also impact gene expression and disease pathogenesis. Cancer is the disease most commonly associated with adult nutrition and epigenetic modifications. DNA hypomethelation promotes cancer progression by allowing increased gene transcription, while hypermethylation can silence tumor suppressor genes and further promote uncontrolled cell division and tumor formation. Compounds found in foods, such as genistein and tea polyphenols, are able to regulate DNA methyltransferases and histone acetylation in cultured cancer cells and may provide protection against certain types of cancer . Other dietary compounds, such as diallyl disulfide present in garlic and sulforaphane present in cruciferous vegetables, have been associated with cancer prevention in clinical trials . This is due to their ability to inhibit histone deacetylase (HDAC) enzymes and prevent silencing of important regulatory genes.

Transgenerational effects

Many believe epigenetic regulation is cleared during the fertilization process, yet more evidence for transgenerational effects (TGEs) are being revealed. These TGEs take place when the epigenetic regulatory patterns are not sufficiently erased during fertilization, possibly due to nutrition levels in previous generations. Later generations may be affected from caloric and protein restriction, high-fat interventions and endocrine disruption in earlier generations. Differences within the nutritional behavior of the maternal rat are believed to cause malprogramming in the F1 generation and may then be passed to subsequent generations. Maternal rats fed a PR-diet during the entire length of pregnancy led to metabolic-related problems in the F1 and F2 generations, even with normal nutrition during the F1 pregnancy. These effects have also been seen in the F3 generation depending on the length of protein restriction. If protein restriction occurred solely during pregnancy, the F1 and F2 offspring had higher systolic blood pressure and lower nephron numbers, possibly predisposing them to hypertension. Altered glucose utilization was detected in the grand-offspring of maternal rats fed a PR-diet during pregnancy and lactation, potentially resulting in diabetes later on in life



Transgenerational effects of maternal protein restriction

Protein-restriction in the F0 generation led to hypomethylation of promoters involved in metabolism in the F1 and F2 generations, even though the F1 pregnant rat was given a normal diet . The exact mechanism of this situation has yet to be elucidated; however, direct transmission is a distinct possibility, meaning the epigenetic marks were preserved during spermatogenesis and oogenesis, when they are normally erased.

Models used in nutriepigenomic studies

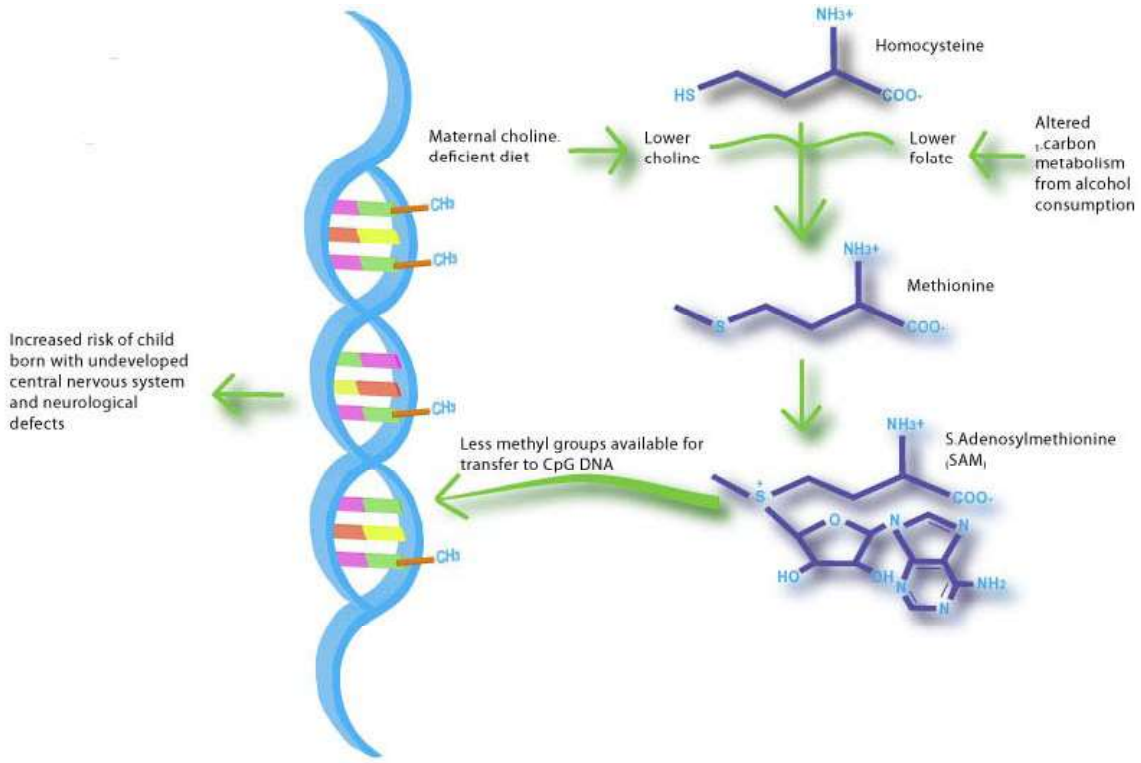
Most research to date use common rodent models to investigate the role of nutrition on phenotype . Popular areas to investigate include IUGR studies, whereby rodents, and sometimes sheep, are subjected to a variety of nutritional conditions. A model for studying IUGR in rodent was developed by Simmons et al (2010) and is used to

investigate type II diabetes. The maternal rats have their uterine arteries ligated, causing altered use of glucose and insulin in the fetus and can therefore serve as a model for diabetes. These growth-retarded rats were found to be highly similar to human fetuses, as they both display symptoms such as lowered glucose and insulin levels. Gestational diabetes may also be studied through chemical induction using streptozotocin treatment of the pregnant rats. Streptozotocin can cause destruction of the beta cells within the pancreas depending on the concentration given.

The predominant means of investigating nutriepigenetics involves varying the nutritional conditions to which a subject is exposed to and monitoring the effects thereafter. Restricting caloric and protein intake are the two most common methods . A pregnant rodent may have their caloric intake reduced up to 30-50% of normal intake. Protein restricted rodents are given 8-9% casein, as opposed to control rats that are fed 20% casein. Micronutrients, such as zinc and iron, may also be restricted to investigate the effects on offspring. Additionally, rats fed diets lacking or including methyl donors are often used to study the effects of diet on epigenomics, as variations within the methylation of DNA are common means of silencing or expressing genes . Supplementing maternal rats with folic acid, vitamin B12, choline and betaine leads to increased levels of DNA methylation at CpG sites and causes a coat color change . This is an example of epigenetically modifiable loci called a “metastable epiallele”, of which only a few have been identified. The above is an example of the “agouti” gene locus, whereby the insertion of a transposable element upstream to the Agouti gene is hypermethylated from the supplementation and causes a change in the rat’s coat color. Diets containing higher carbohydrate and fat content attempt to mimic typical Western-style diets may also be used in nutriepigenetic studies . Another method used is “catch-up”, where offspring of rats born to mothers subjected to various diets are subsequently cross-fostered to mothers fed normal diets .

Future directions

The possibilities of utilizing nutriepigenomics for intervention are quite expansive. This can include preventative therapies, such as providing an optimal regime for nutrition during pregnancy and lactation . It is already common place for pregnant mothers to supplement their diets with choline and folate to prevent the development of neurological disabilities in the fetus.



The nutriepigenetic pathway of maternal choline-deficient diets helps to elucidate the development of fetal alcohol syndrome

A highly specific diet, termed an “EpiG diet”, may be employed for an individual believed to be at higher risk of developing a metabolic disorder . These diets may include supplementation with methyl donors, such as folate. There are also many other natural compounds, such as resveratrol, curcumin and green tea that have been termed “epigenetic modifiers”, as they have anti-cancer capabilities in addition to being used as treatments for metabolic diseases . However, the functions of these compounds still require long-term studies to evaluate their effect over time.

There also exists potential for therapeutic treatments that may correct metabolic disorders, such as type II diabetes . Components of garlic and cruciferous vegetables are known to possess HDAC inhibitors that modify the acetylation of histone proteins and may contain a protection against cancer . These same compounds have also been implicated in irritable bowel syndrome (IBS) and colon cancer, as they may modify the histones normally implicated in these diseases .

Elucidation of disease pathways is another future direction for nutriepigenomic studies. For example, choline-deficient diets and alcohol metabolism during pregnancy may have very similar metabolic pathways . Therefore, animal studies using choline-restricted diets may assist in investigations of fetal alcohol spectrum disorders.

When compared to studies of maternal transmission, investigations into the role of paternal diets are lacking. A review demonstrated the nutrition of both parents do in fact play a role in determining the health of their offspring . A germ-line study reported paternal rats fed a high-fat diet led to insulin dysfunction in the F1 offspring . While this likely occurs via epigenetic modifications similar to those postulated in the maternal diets, the exact mechanism remains to be defined. Assessing the role of epigenetic mechanisms may be easier using paternal inheritance, as sperm transmits epigenetic and genetic information, whereas the female cells also transmit mitochondrial DNA .

Chapter 8

Paramutation and Sex-Determination System

Paramutation

In epigenetics, **paramutation** is an interaction between two alleles of a single locus, resulting in a heritable change of one allele that is induced by the other allele. Paramutation violates Mendel's first law, which states that in the process of the formation of the gametes (egg or sperm) the allelic pairs separate, one going to each gamete, and that each allele remains completely uninfluenced by the other. In paramutation an allele in one generation heritably affects the other allele in future generations, even if the allele causing the change is itself not transmitted. What may be transmitted are patterns of DNA methylation or RNAs such as piRNAs, siRNAs, miRNAs or other regulatory RNAs. Through proper breeding, paramutation can result in isogenic sibling plants with drastically different phenotypes.

Paramutation was first discovered and studied in maize (*Zea mays*) by R.A. Brink at the University of Wisconsin–Madison in the 1950s. Brink noticed that specific weakly expressed alleles of the *red1 (r1)* locus in maize, which encodes a transcription factor that confers red pigment to corn kernels, can heritably change specific strongly expressed alleles to a weaker expression state. The weaker expression state adopted by the changed allele is heritable and can, in turn, change the expression state of other active alleles in a process termed **secondary paramutation**. Brink showed that the influence of the paramutagenic allele could persist for many generations.

Interestingly, paramutation can result in a single allele of a gene controlling a spectrum of phenotypes. At *r1* in maize, for example, the weaker expression state adopted by an allele following paramutation can range from completely colorless to nearly fully-colored kernels. This is an exception to the general observation that continuous variation is controlled by many genes.

Allelic interactions similar to paramutation have since been reported in other organisms, including tomato, pea, and mice.

The molecular basis of paramutation is being unraveled, almost exclusively in maize. Paramutation may share common mechanisms with other epigenetic phenomena, such as gene silencing and genomic imprinting. In maize, paramutation seems to share many traits with the well understood RNA-directed DNA-methylation pathway in *Arabidopsis thaliana*, even though it has never been observed in the famous model plant. Alleman (2006) reported that, in maize, "paramutation is RNA-directed. Stability of the chromatin states associated with paramutation and transposon silencing requires the *mop1* gene, which encodes an RNA-dependent RNA polymerase." Exactly how the RNA produced by this polymerase causes paramutation in maize is not yet understood, but like other epigenetic changes, it involves the covalent modification of DNA and/or the DNA-bound histone proteins without changing the DNA sequence itself.

Sex-determination system

A **sex-determination system** is a biological system that determines the development of sexual characteristics in an organism. Most sexual organisms have two sexes. In many cases, sex determination is genetic: males and females have different alleles or even different genes that specify their sexual morphology. In animals, this is often accompanied by chromosomal differences. In other cases, sex is determined by environmental variables (such as temperature) or social variables (the size of an organism relative to other members of its population). The details of some sex-determination systems are not yet fully understood.

Chromosomal determination

XX/XY sex chromosomes

The **XX/XY sex-determination system** is the most familiar sex-determination systems, as it is found in human beings, most other mammals, as well as some insects. However, at least one monotreme, the platypus, presents a particular sex determination scheme that in some ways resembles that of the ZW sex chromosomes of birds, and also lacks the SRY gene, whereas some rodents, such as several Arvicolinae (voles and lemmings), are also noted for their unusual sex determination systems. The platypus has ten sex chromosomes; males have an XYXYXYXYXY pattern while females have ten X chromosomes. Although it is an XY system, the platypus' sex chromosomes share no homologues with eutherian sex chromosomes. Instead, homologues with eutherian sex chromosomes lie on the platypus chromosome 6, which means that the eutherian sex chromosomes were autosomes at the time that the monotremes diverged from the therian mammals (marsupials and eutherian mammals). However, homologues to the avian DMRT1 gene on platypus sex chromosomes X3 and X5 suggest that it is possible the sex-determining gene for the platypus is the same one that is involved in bird sex-

determination. However, more research must be conducted in order to determine the exact sex determining gene of the platypus.

In the XY sex-determination system, females have two of the same kind of sex chromosome (XX), while males have two distinct sex chromosomes (XY). Some species (including humans) have a gene SRY on the Y chromosome that determines maleness; others (such as the fruit fly) use the presence of two X chromosomes to determine femaleness. The XY sex chromosomes are different in shape and size from each other unlike the autosomes, and are termed allosomes.

XX/X0 sex determination

In this variant of the XY system, females have two copies of the sex chromosome (XX) but males have only one (X0). The *0* denotes the absence of a second sex chromosome. This system is observed in a number of insects, including the grasshoppers and crickets of order Orthoptera and in cockroaches (order Blattodea).

The nematode *C. elegans* is male with one sex chromosome (X0); with a pair of chromosomes (XX) it is a hermaphrodite.

ZW sex chromosomes

The **ZW sex-determination system** is found in birds and some insects and other organisms. The ZW sex-determination system is reversed compared to the XY system: females have two different kinds of chromosomes (ZW), and males have two of the same kind of chromosomes (ZZ). In the chicken, this was found to be dependent on the expression of DMRT1.

Haplodiploidy

Haplodiploidy is found in insects belonging to Hymenoptera, such as ants and bees. Unfertilized eggs develop into haploid individuals, which are the males. Diploid individuals are generally female but may be sterile males. Thus, if a queen bee mates with one drone, her daughters share ¾ of their genes with each other, not ½ as in the XY and ZW systems. This is believed to be significant for the development of eusociality, as it increases the significance of kin selection. This is common also in wasps that are parasitic and in the male greenflies.

Non-genetic sex-determination systems

Many other sex-determination systems exist. In some species of reptiles, including alligators, some turtles, and the tuatara, sex is determined by the temperature at which the egg is incubated. Some species, such as some snails, practice sex change: adults start out male, then become female. In tropical clown fish, the dominant individual in a group becomes female while the other ones are male, and blue wrasse fish are the reverse. In the

marine worm *Bonellia viridis*, larvae become males if they make physical contact with the female, and females if they end up on the bare sea floor.

Some species, however, have no sex-determination system. Hermaphrodites include the common earthworm and certain species of snails. A few species of fish, reptiles, and insects reproduce by parthenogenesis and are female altogether.

In some arthropods, sex is determined by infection, as when Bacteria of the genus *Wolbachia* alter their sexuality; some species consist entirely of ZZ individuals, with sex determined by the presence of *Wolbachia*.

Other unusual systems:

- Swordtail fish
- The Chironomus midge species
- The Platypus has 10 sex chromosomes but lacks the mammalian sex-determining gene SRY, meaning that the process of sex determination in the Platypus remains unknown.

Chapter 9

Soft Inheritance, Structural Inheritance and Testis Determining Factor

Soft inheritance

Soft inheritance is the term coined by Ernst Mayr to include such ideas as Lamarckism, that an organism can pass on characteristics that it acquired during its lifetime to its offspring. It contrasts with modern ideas of inheritance, which Mayr called hard inheritance. Since Mendel, modern genetics has held that the hereditary material is impervious to environmental influences (except, of course, mutagenic effects). In soft inheritance "the genetic basis of characters could be modified either by direct induction by the environment, or by use and disuse, or by an intrinsic failure of constancy, and that this modified genotype was then transmitted to the next generation." Concepts of soft inheritance are usually associated with the ideas of Lamarck and Geoffroy. The concept of hard inheritance holds sway today.

One of the first statements in favour of hard inheritance was made by the English surgeon William Lawrence in 1819. His ideas on heredity were many years ahead of their time, as this extract shows: "The offspring inherit only [their parents'] connate peculiarities and not any of the acquired qualities". This is as clear a rejection of soft inheritance as one can find. However, Lawrence qualified it by including the origin of birth defects owing to influences on the mother (an old folk superstition). So Mayr places Wilhelm His, Sr. in 1874 as the first unqualified rejection of soft inheritance. August Weismann, in 1883, gave a comprehensive denial of Lamarkism (soft inheritance) and with his distinction between germ and soma provided a general ideology of hard inheritance which survives to the present day.

Recent work in plants and mammals on the role of the environment on epigenetic modifications of DNA have led to the argument that inherited epigenetic variation is a kind of soft inheritance.

Structural inheritance

Structural inheritance or **cortical inheritance** is the transmission of a trait in a living organism by a self-perpetuating spatial structures. This is in contrast to the transmission of digital information such as is found in DNA sequences, which accounts for the vast majority of known genetic variation.

Examples of structural inheritance include the propagation of prions, the infectious proteins of diseases such as scrapie (in sheep and goats), bovine spongiform encephalopathy ('mad cow disease') and Creutzfeld-Jakob disease (although the protein-only hypothesis of prion transmission has been considered contentious until recently.) Prions based on heritable protein structure also exist in yeast. Structural inheritance has also been seen in the orientation of cilia in protozoans such as *Paramecium* and *Tetrahymena*, and 'handedness' of the spiral of the cell in *Tetrahymena*, and shells of snails. Some organelles also have structural inheritance, such as the centriole, and the cell itself (defined by the plasma membrane) may also be an example of structural inheritance.

Testis determining factor

Testis-determining factor (**TDF**) is a general term for the gene (or product thereof) that results in maleness in humans and some other species.

Certain genes cause chemical reactions that result in the development of testes. Embryos are gonadally identical, regardless of genetic sex, until a certain point in development; then the testis-determining factor causes male sex organs to develop, whereas lack of this factor will cause the embryo to develop as physically female.

The TDF factor is encoded by the SRY gene located in the Y chromosome. It is a DNA-binding protein that enhances other transcription factors, or is a transcription factor itself. Its expression directly or indirectly causes the development of primary sex cords, which will later develop to seminiferous tubules. These cords form in the central part of the yet-undifferentiated gonad, turning it into a testis. The testis then starts secreting testosterone and the Mullerian Inhibiting Substance.

Older texts discuss the role of the HY antigen in the control of testicular development, which was later disproven.

Role in disease

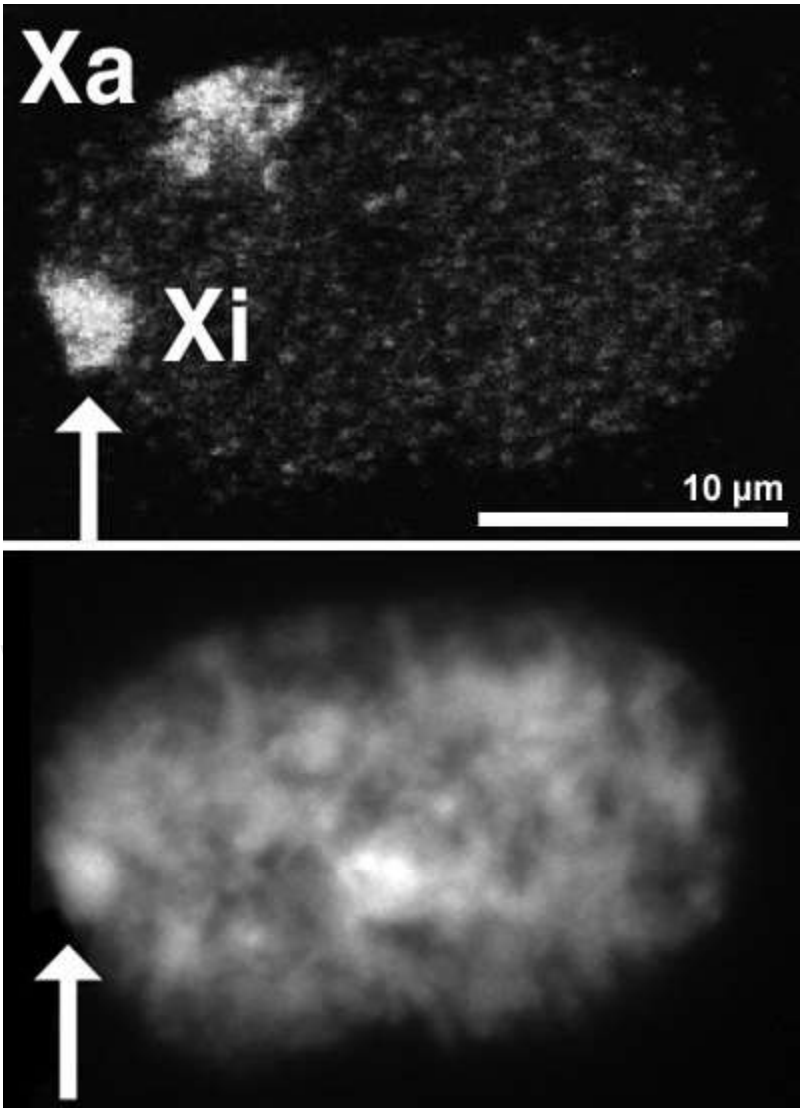
The TDF gene has some interesting implications. The genetic recombination of Crossing over can cause the gene to be transferred on to the X chromosome. In this case, the X chromosome will initiate testis development; so, regardless of whether the person has a Y chromosome, the person will turn into a male. Though everything else will be developed as if it were a female (other sex-related alleles), the apparent sex will be male (a syndrome known as XX male syndrome).

On the converse, such a cross-over event also can result in a Y chromosome that is missing the Sex-Determining Region (SRY), which contains the TDF, replaced with the corresponding sequence from the end of the X chromosome. Individuals that inherit this Y chromosome will develop as females, despite having the normal male chromosomal set of one X and one Y. This is called Swyer syndrome (46XY, genotypic male but phenotypic female).

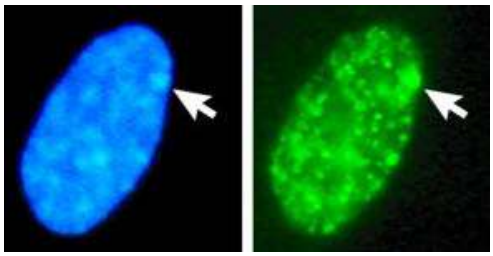
Chapter 10

X-Inactivation

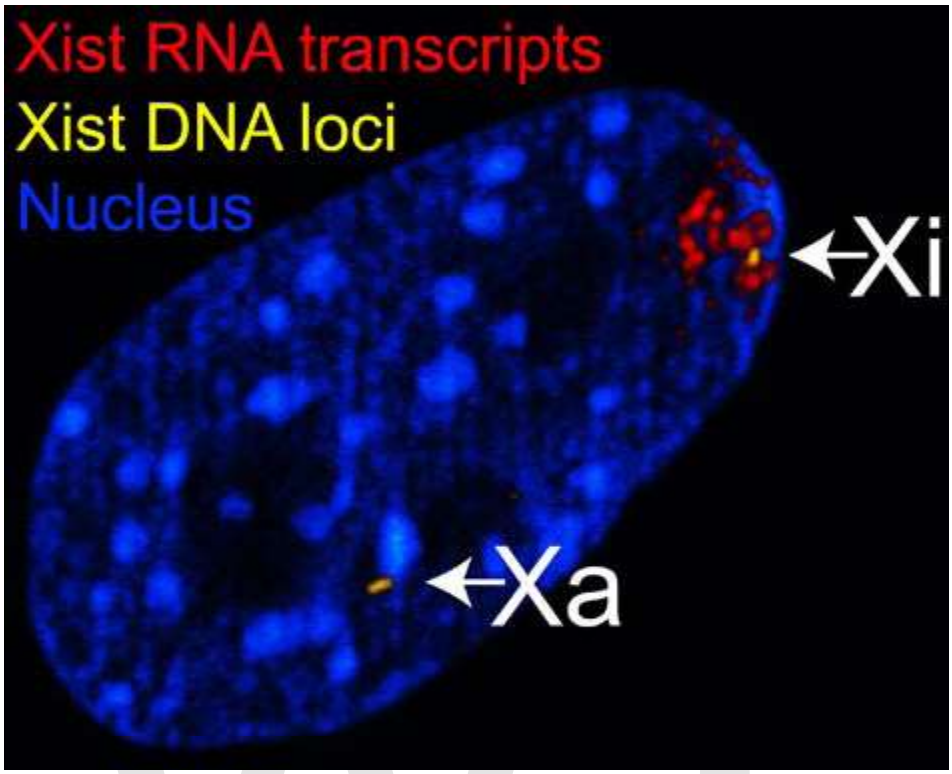
The coloration of tortoiseshell cats is a visible manifestation of X-inactivation. The black and orange alleles of a fur coloration gene reside on the X chromosome. For any given patch of fur, the inactivation of an X chromosome that carries one gene results in the fur color of the other, active gene.



Nucleus of a female cell. Top: Both X-chromosomes are detected, by FISH. Bottom: The same nucleus stained with a DNA stain (DAPI). The Barr body is indicated by the arrow, it identifies the inactive X (Xi).



An interphase female human fibroblast cell. Arrows point to sex chromatin on DNA (DAPI) in cell nucleus(left), and to the corresponding X chromatin (right).
Left: DNA (DAPI)-stained nucleus. Arrow indicates the location of Barr body(Xi). Right: DNA associated histones protein detected



The figure shows confocal microscopy images from a combined RNA-DNA FISH experiment for Xist in fibroblast cells from adult female mouse, demonstrating that Xist RNA is coating only one of the X-chromosomes. RNA FISH signals from Xist RNA are shown in red color, marking the inactive X-chromosome (Xi). DNA FISH signals from Xist loci are shown in yellow color, marking both active and inactive X-chromosomes (Xa, Xi). The nucleus (DAPI-stained) is shown in blue color. The figure is adapted from:

X-inactivation (also called **lyonization**) is a process by which one of the two copies of the X chromosome present in female mammals is inactivated. The inactive X chromosome is silenced by packaging into transcriptionally inactive heterochromatin. X-inactivation occurs so that the female, with two X chromosomes, does not have twice as many X chromosome gene products as the male, which only possess a single copy of the X chromosome. The choice of which X chromosome will be inactivated is random in placental mammals such as mice and humans, but once an X chromosome is inactivated it will remain inactive throughout the lifetime of the cell and its descendants in the organism. Unlike the random X-inactivation in placental mammals, inactivation in marsupials applies exclusively to the paternally derived X chromosome.

History

In 1959 Susumu Ohno showed that the two X-chromosomes of mammals were different: one appeared like the autosomes; the other was condensed and heterochromatic. This finding suggested, independently to two groups of investigators, that one of the X-

chromosomes underwent inactivation. In 1961, Mary Lyon proposed the random inactivation of one female X chromosome to explain the mottled phenotype of female mice heterozygous for coat color genes. The Lyon hypothesis also accounted for the findings that one copy of the X chromosome in female cells was highly condensed, and that mice with only one copy of the X chromosome developed as infertile females. Ernest Beutler, studying heterozygous females for Glucose-6-phosphate dehydrogenase (G6PD) deficiency, independently proposed that there were two red cell populations of erythrocytes in such heterozygotes: deficient cells and normal cells, depending on whether the inactivated X chromosome contains the normal or defective G6PD allele.

Mechanism

Timing

All mouse cells undergo an early, imprinted inactivation of the paternally-derived X chromosome in two-cell or four-cell stage embryos. The extraembryonic tissues (which give rise to the placenta and other tissues supporting the embryo) retain this early imprinted inactivation, and thus only the maternal X chromosome is active in these tissues.

In the early blastocyst, this initial, imprinted X-inactivation is reversed in the cells of the inner cell mass (which give rise to the embryo), and in these cells both X chromosomes become active again. Each of these cells then independently and randomly inactivates one copy of the X chromosome. This inactivation event is irreversible during the lifetime of the cell, so all the descendants of a cell which inactivated a particular X chromosome will also inactivate that same chromosome. This phenomenon, which can be observed in the coloration of calico cats when females are heterozygous for the X-linked gene, should not be confused with mosaicism, which is a term that specifically refers to differences in the genotype of various cell populations in the same individual; X-inactivation, which is an epigenetic change that results in a different phenotype, is *not* a change at the genotypic level. For an individual cell or lineage the inactivation is therefore skewed or 'non-random' this can give rise to mild symptoms in female 'carriers' of X-linked genetic disorders.

X-inactivation is reversed in the female germline, so that all oocytes contain an active X chromosome.

Selection of one active X chromosome

Normal females possess two X chromosomes, and in any given cell one chromosome will be active (designated as Xa) and one will be inactive (Xi). However, studies of individuals with extra copies of the X chromosome show that in cells with more than two X chromosomes there is still only one Xa, and all the remaining X chromosomes are inactivated. This indicates that the default state of the X chromosome in females is inactivation, but one X chromosome is always selected to remain active.

It is hypothesized that there is an autosomally-encoded 'blocking factor' which binds to the X chromosome and prevents its inactivation. The model postulates that there is a limiting blocking factor, so once the available blocking factor molecule binds to one X chromosome the remaining X chromosome(s) are not protected from inactivation. This model is supported by the existence of a single Xa in cells with many X chromosomes and by the existence of two active X chromosomes in cell lines with twice the normal number of autosomes.

Sequences at the **X inactivation center** (**XIC**), present on the X chromosome, control the silencing of the X chromosome. The hypothetical blocking factor is predicted to bind to sequences within the XIC.

Chromosomal component

The X-inactivation center (XIC) on the X chromosome is necessary and sufficient to cause X-inactivation. Chromosomal translocations which place the XIC on an autosome lead to inactivation of the autosome, and X chromosomes lacking the XIC are not inactivated.

The XIC contains two non-translated RNA genes, Xist and Tsix, which are involved in X-inactivation. The XIC also contains binding sites for both known and unknown regulatory proteins.

Xist and Tsix RNAs

The X-inactive specific transcript (Xist) gene encodes a large non-coding RNA that is responsible for mediating the specific silencing of the X chromosome from which it is transcribed. The inactive X chromosome is coated by Xist RNA, whereas the Xa is not. The Xist gene is the only gene which is expressed from the Xi but not from the Xa. X chromosomes which lack the Xist gene cannot be inactivated. Artificially placing and expressing the Xist gene on another chromosome leads to silencing of that chromosome.

Prior to inactivation, both X chromosomes weakly express Xist RNA from the Xist gene. During the inactivation process, the future Xa ceases to express Xist, whereas the future Xi dramatically increases Xist RNA production. On the future Xi, the Xist RNA progressively coats the chromosome, spreading out from the XIC; the Xist RNA does not localize to the Xa. The silencing of genes along the Xi occurs soon after coating by Xist RNA.

Like Xist, the Tsix gene encodes a large RNA which is not believed to encode a protein. The Tsix RNA is transcribed antisense to Xist, meaning that the Tsix gene overlaps the Xist gene and is transcribed on the opposite strand of DNA from the Xist gene. Tsix is a negative regulator of Xist; X chromosomes lacking Tsix expression (and thus having high levels of Xist transcription) are inactivated much more frequently than normal chromosomes.

Like Xist, prior to inactivation, both X chromosomes weakly express Tsix RNA from the Tsix gene. Upon the onset of X-inactivation, the future Xi ceases to express Tsix RNA (and increases Xist expression), whereas Xa continues to express Tsix for several days.

Silencing

The inactive X chromosome does not express the majority of its genes, unlike the active X chromosome. This is due to the silencing of the Xi by repressive heterochromatin, which coats the Xi DNA and prevents the expression of most genes.

Compared to the Xa, the Xi has high levels of DNA methylation, low levels of histone acetylation, low levels of histone H3 lysine-4 methylation, and high levels of histone H3 lysine-9 methylation, all of which are associated with gene silencing. Additionally, a histone variant called macroH2A is exclusively found on nucleosomes along the Xi.

Barr bodies

DNA packaged in heterochromatin, such as the Xi, is more condensed than DNA packaged in euchromatin, such as the Xa. The inactive X forms a discrete body within the nucleus called a Barr body. The Barr body is generally located on the periphery of the nucleus, is late replicating within the cell cycle, and, as it contains the Xi, contains heterochromatin modifications and the Xist RNA.

Expressed genes on the inactive X chromosome

A fraction of the genes along the X chromosome escape inactivation on the Xi. The Xist gene is expressed at high levels on the Xi and is not expressed on the Xa. Other genes are expressed equally from the Xa and Xi; mice contain few genes which escape silencing whereas up to a quarter of human X chromosome genes are expressed from the Xi. Many of these genes occur in clusters.

Many of the genes which escape inactivation are present along regions of the X chromosome which, unlike the majority of the X chromosome, contain genes also present on the Y chromosome. These regions are termed pseudoautosomal regions, as individuals of either sex will receive two copies of every gene in these regions (like an autosome), unlike the majority of genes along the sex chromosomes. Since individuals of either sex will receive two copies of every gene in a pseudoautosomal region, no dosage compensation is needed for females, so it is postulated that these regions of DNA have evolved mechanisms to escape X-inactivation. The genes of pseudoautosomal regions of the Xi do not have the typical modifications of the Xi and have little Xist RNA bound.

The existence of genes along the inactive X which are not silenced explains the defects in humans with abnormal numbers of the X chromosome, such as Turner syndrome (X0) or Klinefelter syndrome (XXY). Theoretically, X-inactivation should eliminate the differences in gene dosage between affected individuals and individuals with a normal

chromosome complement, but in affected individuals the dosage of these non-silenced genes will differ as they escape X-inactivation.

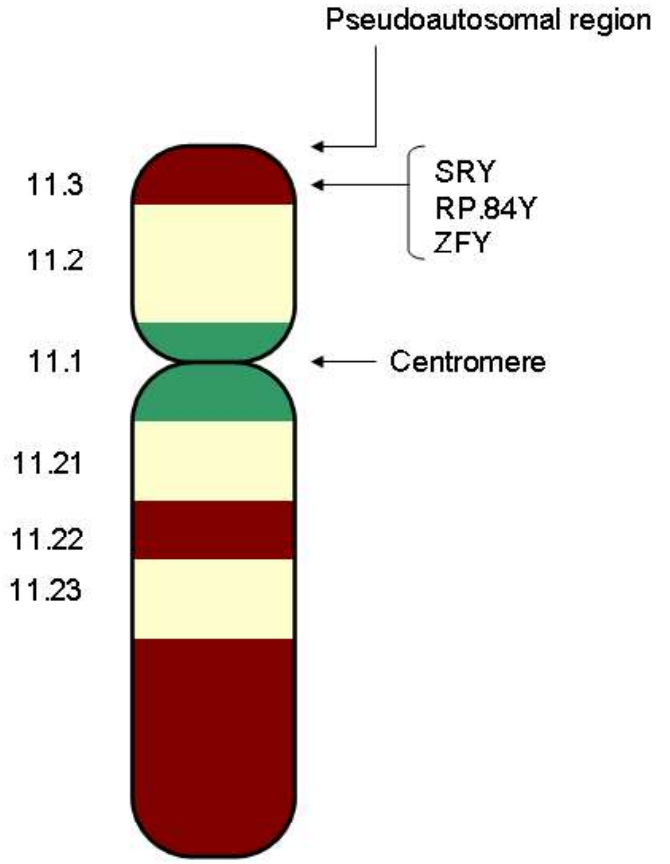
The precise mechanisms that control escape from X-inactivation are not known, but silenced and escape regions have been shown to have distinct chromatin marks. It has been suggested that escape from X-inactivation might be mediated by expression of long non-coding RNA (lncRNA) within the escaping chromosomal domains.

Uses in experimental biology

Stanley Michael Gartler used X chromosome inactivation to demonstrate the clonal origin of cancers. Examining normal tissues and tumors from females heterozygous for isoenzymes of the sex-linked G6PD gene demonstrated that tumor cells from such individuals express only one form of G6PD, whereas normal tissues are composed of a nearly equal mixture of cells expressing the two different phenotypes. This pattern suggests that a single cell, and not a population, grows into a cancer.

Chapter 11

Sex Determination and Differentiation (Human)



The Human Y Chromosome showing the SRY gene which codes for a protein regulating sexual differentiation.

Human sex refers to the processes by which an individual becomes either a male or female during development.

The Jost Paradigm

Under typical circumstances, the sex of an individual will be determined and expressed through the following mechanisms:

- Chromosomal Sex (genetic): Presence or absence of Y chromosome
- Gonadal Sex (Primary Sex Determination): Controlled by presence or absence of testis determining factor (TDF)
- Phenotypic Sex (Secondary Sex Differentiation): Determined by the hormonal products produced by the gonads.

Sex determination

Sex determination at the chromosome level

For the majority of individuals, sex determination is as simple as the presence or absence of a Y chromosome. Those individuals with a Y chromosome (including XXY, XXXY, etc.) will develop into males, and those without one will become female. Some individuals, however, will undergo what is referred to as primary sex reversal, whereby the X and Y chromosomes [cross over] and exchange genetic material. This relatively rare occurrence (approximately 1 in 20000 births) can lead to males with two X chromosomes and females with a Y chromosome.

Testis determining gene

During the late 1980s and early 1990s, coinciding with the mapping of the human genome, researchers began to look for the specific gene on the Y chromosome that, up until then, had been known as the testis determining factor (TDF). Through the study of individuals that underwent primary sex reversal (that is, XX males and XY females), researchers determined that the TDF must lie on the Y chromosome in a location that would permit its exchange to the X chromosome during cross over. In 1985, Dr. David C. Page published an article in Nature boldly stating that the TDF was the ZFY gene on the Y chromosome. However, Dr. MS Palmer later discovered a ZFY analogue on the X chromosome, providing evidence that ZFY was in fact not the TDF. Eventually, Dr. Peter Koopman was able to prove that the SRY gene is the TDF from studies on XX males.. The following evidence further supports this claim:

- SRY is Y specific, and there is no analogue on the X chromosome.
- SRY is deleted or mutated in XY females
- It undergoes expression within the testis at the time of testis differentiation.
- Its sequence suggests that its protein has a DNA binding motif because it has high homology to an 80 amino acid long DNA binding region (HMG box).

SRY a repressor?

Recently, it has been suggested by some that the SRY gene acts as a repressor or inhibitor of another gene, "Z", that is involved in female development. Previously, it was stated that the SRY sequence suggests the presence of a DNA binding motif. Also, the idea that SRY is a repressor is further supported by the fact that a small percentage of sex reversal cases cannot be explained by the absence of SRY and could be due to a mutation in some gene "Z" that prevents the binding of SRY and its subsequent antagonist action.

Other sex determination genes

- DAX1: Exerts its effects early on in development. There is some debate over what its role is in the development of testis. It is a candidate for gene "Z".
- SOX9: mutations in this gene cause severe dwarfism, and a bone disorder called campomelic dysplasia, which occurs in many sex reversed males.

Sex differentiation

Sex differentiation refers to the expression of phenotypic attributes specific to the sex of an individual. While gonad development is a result of the presence or absence of the sex determination gene SRY on the Y chromosome, sex differentiation is determined by the hormonal products produced by the gonads.

Testosterone

In the 1930s, Alfred Jost determined that the presence of testosterone was required for Wolffian duct development in the female rabbit.

Müllerian inhibiting substance

Jost also observed that while testosterone was required for Wolffian duct development, the regression of the Müllerian duct was due to another substance. This was later determined to be Müllerian inhibiting substance (MIS), a 140 kD dimeric glycoprotein that is produced by sertoli cells. MIS blocks the development of Müllerian ducts, promoting their regression.

5-alpha dihydrotestosterone (DHT)

Testosterone is converted to the more potent DHT by 5-alpha reductase. DHT is necessary to exert androgenic effects farther from the site of testosterone production, where the concentrations of testosterone are too low to have any potency. A 5-alpha reductase deficiency results in an androgen disorder characterized by female phenotype or severely undervirilized male phenotype with development of the epididymis, vas deferens, seminal vesicle, and ejaculatory duct, but also a pseudovagina.

Pathologies

The following disorders are caused by a malfunction in the sex determination and differentiation process:

- Congenital Adrenal Hyperplasia - Inability of adrenal to produce sufficient cortisol, leading to increased production of testosterone resulting in severe masculinization of 46 XX females.
- Persistent Müllerian Duct Syndrome - A rare type of pseudohermaphroditism that occurs in 46 XY males, caused by either a mutation in the Müllerian inhibiting substance (MIS) gene, on 19p13, or its type II receptor, 12q13. Results in a retention of Müllerian ducts (persistance of rudimentary uterus and fallopian tubes in otherwise normally virilized males), unilateral or bilateral undescended testes and sometimes causes infertility.
- Male Pseudohermaphroditism - Failure of androgen production or inadequate androgen response, which can cause incomplete masculinization in XY males. Varies from mild failure of masculinization with undescended testes to complete sex reversal and female phenotype (Androgen insensitivity syndrome)
- Swyer syndrome. A form of complete gonadal dysgenesis, mostly due to mutations in the first step of sex determination; the SRY genes.

Chapter 12

Genetic Linkage

Genetic linkage is the tendency of certain loci or alleles to be inherited together. Genetic loci that are physically close to one another on the same chromosome tend to stay together during meiosis, and are thus genetically *linked*.

Background

At the beginning of normal meiosis, a chromosome pair (made up of a chromosome from the mother and a chromosome from the father) intertwine and exchange sections or fragments of chromosome. The pair then breaks apart to form two chromosomes with a new combination of genes that differs from the combination supplied by the parents. Through this process of recombining genes, organisms can produce offspring with new combinations of maternal and paternal traits that may contribute to or enhance survival.

This recombination of genes, called the crossing over of DNA, can cause alleles previously on the same chromosome to be separated and end up in different daughter cells. The further the two alleles are apart, the greater the chance that a cross-over event may occur between them, and the greater the chance that the alleles are separated.

The relative distance between two genes can be calculated by taking the offspring of an organism showing two linked genetic traits, and finding the percentage of the offspring where the two traits do not run together. The higher the percentage of descendants that does not show both traits, the farther apart on the chromosome the two genes are. Genes for which this percentage is lower than 50% are typically thought to be linked.

Genetic linkage can also be understood by looking at the relationships among phenotypes. Among individuals of an experimental population or species, some phenotypes or traits can occur randomly with respect to one another, or with some correlation with respect to one another.

The former is known as independent assortment. Today, scientists understand that independent assortment occurs when the genes affecting the phenotypes are found on

different chromosomes or separated by a great enough distance on the same chromosome that recombination occurs at least half of the time.

The latter is known as genetic linkage. This occurs as an exception to independent assortment, and develops when genes appear near one another on the same chromosome. This phenomenon causes the genes to usually be inherited as a single unit. Genes inherited in this way are said to be linked, and are referred to as "linkage groups". For example, in fruit flies, the genes affecting eye color and wing length are inherited together because they appear on the same chromosome.

Discovery

Genetic linkage was first discovered by the British geneticists William Bateson and Reginald Punnett shortly after Mendel's laws were rediscovered. The understanding of genetic linkage was expanded by the work of Thomas Hunt Morgan. Morgan's observation that the amount of crossing over between linked genes differs led to the idea that crossover frequency might indicate the distance separating genes on the chromosome.

Alfred Sturtevant, a student of Morgan's, first developed genetic maps, also known as linkage maps. Sturtevant proposed that the greater the distance between linked genes, the greater the chance that non-sister chromatids would cross over in the region between the genes. By working out the number of recombinants it is possible to obtain a measure for the distance between the genes. This distance is called a **genetic map unit (m.u.)**, or a **centimorgan** and is defined as the distance between genes for which one product of meiosis in 100 is recombinant. A **recombinant frequency** (RF) of 1 % is equivalent to 1 m.u. But this equivalence is only a good approximate for small percentages; the largest percentage of recombinants cannot exceed 50%, which would be the situation where the two genes are at the extreme opposite ends of the same chromosomes. In this situation, any crossover events would result in an exchange of genes, but only an odd number of crossover events (a 50-50 chance between even and odd number of crossover events) would result in a recombinant product of meiotic crossover. A statistical interpretation of this is through the Haldane mapping function or the Kosambi mapping function, among others. A linkage map is created by finding the map distances between a number of traits that are present on the same chromosome, ideally avoiding having significant gaps between traits to avoid the inaccuracies that will occur due to the possibility of multiple recombination events.

Linkage map

A linkage map is a genetic map of a species or experimental population that shows the position of its known genes or genetic markers relative to each other in terms of recombination frequency, rather than as specific physical distance along each chromosome. Linkage mapping is critical for identifying the location of genes that cause genetic diseases.

A genetic map is a map based on the frequencies of recombination between markers during crossover of homologous chromosomes. The greater the frequency of recombination (segregation) between two genetic markers, the farther apart they are assumed to be. Conversely, the lower the frequency of recombination between the markers, the smaller the physical distance between them. Historically, the markers originally used were detectable phenotypes (enzyme production, eye color) derived from coding DNA sequences; eventually, confirmed or assumed noncoding DNA sequences such as microsatellites or those generating restriction fragment length polymorphisms (RFLPs) have been used.

Genetic maps help researchers to locate other markers, such as other genes by testing for genetic linkage of the already known markers.

A genetic map is **not** a physical map (such as a radiation reduced hybrid map) or gene map.

LOD score method for estimating recombination frequency

The **LOD score** (logarithm (base 10) of odds), developed by Newton E. Morton, is a statistical test often used for linkage analysis in human, animal, and plant populations. The LOD score compares the likelihood of obtaining the test data if the two loci are indeed linked, to the likelihood of observing the same data purely by chance. Positive LOD scores favor the presence of linkage, whereas negative LOD scores indicate that linkage is less likely. Computerized LOD score analysis is a simple way to analyze complex family pedigrees in order to determine the linkage between Mendelian traits (or between a trait and a marker, or two markers).

The method is described in greater detail by Strachan and Read . Briefly, it works as follows:

1. Establish a pedigree
2. Make a number of estimates of recombination frequency
3. Calculate a LOD score for each estimate
4. The estimate with the highest LOD score will be considered the best estimate

The LOD score is calculated as follows:

$$LOD = Z = \log_{10} \frac{\text{probability of birth sequence with a given linkage value}}{\text{probability of birth sequence with no linkage}}$$
$$= \log_{10} \frac{(1-\theta)^{NR} \times \theta^{R}}{0.5^{(NR+R)}}$$

NR denotes the number of non-recombinant offspring, and R denotes the number of recombinant offspring. The reason 0.5 is used in the denominator is that any alleles that are completely unlinked (e.g. alleles on separate chromosomes) have a 50% chance of recombination, due to independent assortment.

Theta is the recombinant fraction, it is equal to R / (NR + R)

In practice, LOD scores are looked up in a table which lists LOD scores for various standard pedigrees and various values of recombination frequency.

By convention, a LOD score greater than 3.0 is considered evidence for linkage. A LOD score of +3 indicates 1000 to 1 odds that the linkage being observed did not occur by chance. On the other hand, a LOD score less than -2.0 is considered evidence to exclude linkage. Although it is very unlikely that a LOD score of 3 would be obtained from a single pedigree, the mathematical properties of the test allow data from a number of pedigrees to be combined by summing the LOD scores. It is important to keep in mind that this traditional cutoff of LOD>+3 is an arbitrary one and that the difference between certain types of linkage studies, particularly analyses of complex genetic traits with hundreds of markers, these criteria should probably be modified to a somewhat higher cutoff.

Recombination frequency

Recombination frequency is a measure of genetic linkage and is used in the creation of a genetic linkage map. Recombination frequency (θ) is the frequency that a single chromosomal crossover will take place between two genes during meiosis. A centimorgan (cM) is a unit that describes a recombination frequency of 1%. In this way we can measure the genetic distance between two loci, based upon their recombination frequency. This is a good estimate of the real distance. Double crossovers would turn into no recombination. In this case we cannot tell if crossovers took place. If the loci we're analysing are very close (less than 7 cM) a double crossover is very unlikely. When distances become higher, the likelihood of a double crossover increases. As the likelihood of a double crossover increases we systematically underestimate the genetic distance between two loci.

During meiosis, chromosomes assort randomly into gametes, such that the segregation of alleles of one gene is independent of alleles of another gene. This is stated in Mendel's Second Law and is known as **the law of independent assortment**. The law of independent assortment always holds true for genes that are located on different chromosomes, but for genes that are on the same chromosome, it does not always hold true.

As an example of independent assortment, consider the crossing of the pure-bred homozygote parental strain with genotype *AABB* with a different pure-bred strain with genotype *aabb*. A and a and B and b represent the alleles of genes A and B. Crossing these homozygous parental strains will result in F1 generation offspring with genotype AaBb. The F1 offspring AaBb produces gametes that are *AB*, *Ab*, *aB*, and *ab* with equal frequencies (25%) because the alleles of gene A assort independently of the alleles for gene B during meiosis. Note that 2 of the 4 gametes (50 %)—*Ab* and *aB*—were not present in the parental generation. These gametes represent **recombinant gametes**. Recombinant gametes are those gametes that differ from both of the haploid gametes that

made up the diploid cell. In this example, the recombination frequency is 50% since 2 of the 4 gametes were recombinant gametes.

The recombination frequency will be 50% when two genes are located on different chromosomes or when they are widely separated on the same chromosome. This is a consequence of independent assortment.

When two genes are close together on the same chromosome, they do not assort independently and are said to be linked. Whereas genes located on different chromosomes assort independently and have a recombination frequency of 50%, linked genes have a recombination frequency that is less than 50%.

As an example of linkage, consider the classic experiment by William Bateson and Reginald Punnett. They were interested in trait inheritance in the sweet pea and were studying two genes—the gene for flower colour (*P*, purple, and *p*, red) and the gene affecting the shape of pollen grains (*L*, long, and *l*, round). They crossed the pure lines *PPLL* and *ppll* and then self-crossed the resulting *PpLl* lines. According to Mendelian genetics, the expected phenotypes would occur in a 9:3:3:1 ratio of PL:Pl:pL:pl. To their surprise, they observed an increased frequency of PL and pl and a decreased frequency of Pl and pL (see table below).

Bateson and Punnett experiment

Phenotype and genotype	Observed	Expected from 9:3:3:1 ratio
Purple, long (*PpLl*)	284	216
Purple, round (*Ppll*)	21	72
Red, long (*ppLl*)	21	72
Red, round (*ppll*)	55	24

Their experiment revealed **linkage** between the *P* and *L* alleles and the *p* and *l* alleles. The frequency of *P* occurring together with *L* and with *p* occurring together with *l* is greater than that of the recombinant *Pl* and *pL*. The recombination frequency cannot be computed directly from this experiment, but intuitively it is less than 50%.

The progeny in this case received two dominant alleles linked on one chromosome (referred to as **coupling** or **cis arrangement**). However, after crossover, some progeny could have received one parental chromosome with a dominant allele for one trait (eg Purple) linked to a recessive allele for a second trait (eg round) with the opposite being true for the other parental chromosome (eg red and Long). This is referred to as **repulsion** or a **trans arrangement**. The phenotype here would still be purple and long but a test cross of this individual with the recessive parent would produce progeny with much greater proportion of the two crossover phenotypes. While such a problem may not seem likely from this example, unfavorable repulsion linkages do appear when breeding for disease resistance in some crops.

When two genes are located on the same chromosome, the chance of a crossover producing recombination between the genes is related to the distance between the two genes. Thus, the use of recombination frequencies has been used to develop **linkage maps** or **genetic maps**.

However, it is important to note that recombination frequency tends to underestimate the distance between two linked genes. This is because as the two genes are located further apart, the chance of double or even number of crossovers between them also increases. Double or even number of crossovers between the two genes results in them being cosegregated to the same gamete, yielding a parental progeny instead of the expected recombinant progeny.

Meiosis Indicators

With very large pedigrees or with very dense genetic marker data, such as from whole-genome sequencing, it is possible to precisely locate and quantify recombinations. With this type of genetic analysis, a meiosis indicator is assigned to each position of the genome for each meiosis in a pedigree. The indicator indicates which copy of the parental chromosome contributes to the transmitted gamete at that position. For example, if the allele from the 'first' copy of the parental chromosome is transmitted, a '0' might be assigned to that meiosis. If the allele from the 'second' copy of the parental chromosome is transmitted, a '1' would be assigned to that meiosis. The two alleles in the parent came, one each, from two grandparents. These indicators are then used to determine identical-by-descent (IBD) states or inheritance states, which are in turn used to identify genes responsible for diseases and phenotypes.

Chapter 13

Dominance (Genetics)

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

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

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

Diploid / haploid

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

Chromosomes, genes, and alleles

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

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

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

Homozygous, heterozygous

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

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

Which trait is *dominant*?

The terms *dominant* and *recessive* refer to the interaction of alleles in producing the phenotype of the heterozygote. If there are two alternative phenotypes, by definition the phenotype exhibited by the heterozygote is called "dominant" and the "hidden" phenotype is called "recessive." The key concept of dominance is that the heterozygote is phenotypically *identical* to one of the two homozygotes. That trait corresponding to the dominant allele may then be called the 'dominant' trait.

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

Nomenclature

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

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

Relationship to other genetic concepts

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

Multiple alleles

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

Incomplete and semi-dominance

Complete dominance occurs when the phenotype of the heterozygote is completely indistinguishable from that of the dominant homozygote. This is frequently not the case. Incomplete dominance occurs when the phenotype of the heterozygous genotype is an intermediate of the phenotypes of the homozygous genotypes. For example, the snapdragon flower color is either homozygous for red or white. When the red

homozygous flower is paired with the white homozygous flower, the result yields a pink snapdragon flower. The pink snapdragon is the result of incomplete dominance.

Co-dominance

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

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

Autosomal *versus* sex-linked dominance

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

Epistasis

Epistasis ["*epi* + *stasis* = to sit on top"] is an interaction between genotypes at two *different* gene loci, which sometimes resembles a dominance interaction at a single locus. Epistasis modifies the characteristic 9:3:3:1 ratio expected for two non-epistatic genes. Most genetic systems involve complex epistatic interactions among multiple gene loci. For two loci, 14 classes of epistatic interactions are recognized. As an example of ***recessive epistasis***, one gene locus may determine whether a flower pigment is yellow

(**AA** or **Aa**) or green (**aa**), while another locus determines whether the pigment is produced (**BB** or **Bb**) or not (**bb**). In a **bb** plant, the flowers will be white, irrespective of the genotype of the other locus as **AA**, **Aa**, or **aa**. The **b** allele is *not* dominant to the **A** allele: the **B** locus shows ***recessive epistasis*** to the **A** locus, because the **B** locus when homozygous for the recessive allele (**bb**) suppresses phenotypic expression of the **A** locus. In a cross between two **AaBb** plants, this produces a characteristic **9:3:4** ratio, in this case of yellow : green : white flowers.

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

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

Molecular mechanisms

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

Dominance typically occurs when one of the two alleles is non-functional at the molecular level, that is, it is not transcribed or else does not produce a protein product. This can be the result of a mutation that alters the DNA sequence of the allele. An organism homozygous for the non-functional allele will generally show a distinctive

phenotype, due to the absence of the protein product. For example, in humans and other organisms, the unpigmented skin of the albino phenotype results when an individual is homozygous for an allele that prevents synthesis of the skin pigment protein melanin. It is important to understand that it is not the lack of function that allows the allele to be described as recessive: this is the interaction with the alternative allele in the heterozygote. Three general types of interaction are possible:

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

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

Dominant and recessive genetic diseases in humans

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

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

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

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

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

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

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

History

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

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

Chapter 14

Epistasis and Genetic Screen

Epistasis

Epistasis is the phenomenon where the effects of one gene are modified by one or several other genes, which are sometimes called **modifier genes**. The gene whose phenotype is expressed is called **epistatic**, while the phenotype altered or suppressed is called **hypostatic**. Epistasis can be contrasted with dominance, which is an interaction between alleles at the same gene locus. Epistasis is often studied in relation to Quantitative Trait Loci (QTL) and polygenic inheritance.

In general, the fitness increment of any one allele depends in a complicated way on many other alleles; but, because of the way that the science of population genetics was developed, evolutionary scientists tend to think of epistasis as the exception to the rule. In the first models of natural selection devised in the early 20th century, each gene was considered to make its own characteristic contribution to fitness, against an average background of other genes. Some introductory college courses still teach population genetics this way.

Epistasis and **genetic interaction** refer to different aspects of the same phenomenon. The term **epistasis** is widely used in population genetics and refers especially to the statistical properties of the phenomenon, and does not necessarily imply biochemical interaction between gene products. However, in general epistasis is used to denote the departure from 'independence' of the effects of different genetic loci. Confusion often arises due to the varied interpretation of 'independence' between different branches of biology.

Examples of tightly linked genes having epistatic effects on fitness are found in supergenes and the human major histocompatibility complex genes. The effect can occur directly at the genomic level, where one gene could code for a protein preventing transcription of the other gene. Alternatively, the effect can occur at the phenotypic level. For example, the gene causing albinism would hide the gene controlling color of a person's hair. In another example, a gene coding for a widow's peak would be hidden by a

gene causing baldness. Fitness epistasis (where the affected trait is fitness) is one cause of linkage disequilibrium.

Studying genetic interactions can reveal gene function, the nature of the mutations, functional redundancy, and protein interactions. Because protein complexes are responsible for most biological functions, genetic interactions are a powerful tool.

Classification by fitness or trait value

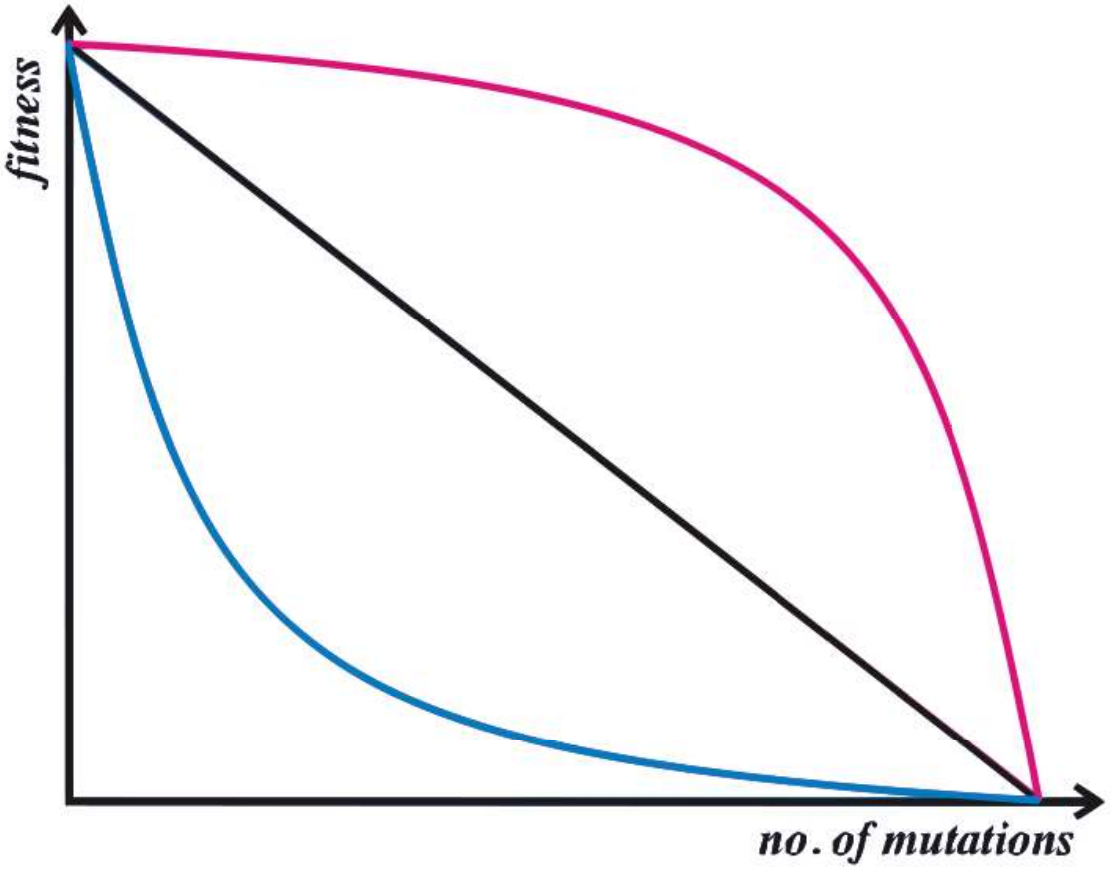


Diagram illustrating different relationships between numbers of mutations and fitness. *Synergistic* epistasis is the blue line - each mutation has a disproportionately large effect on the organism's fitness. *Antagonistic* epistasis is the red line.

Two-locus epistatic interactions can be either synergistic (enhancing the effectiveness) or antagonistic (reducing the activity). In the example of a haploid organism with genotypes (at two loci) *AB*, *Ab*, *aB* or *ab*, we can think of the following trait values where higher values suggest greater expression of the characteristic (the exact values are simply given as examples):

	AB	*Ab*	*aB*	*ab*
No epistasis (additive across loci)	2	1	1	0

Synergistic epistasis	3	1	1	0
Antagonistic epistasis	1	1	1	0

Hence, we can classify thus:

Trait values	**Type of epistasis**
$AB = Ab + aB - ab$	No epistasis, additive inheritance
$AB > Ab + aB - ab$	Synergistic epistasis
$AB < Ab + aB - ab$	Antagonistic epistasis

Understanding whether the majority of genetic interactions are synergistic or antagonistic will help solve such problems as the evolution of sex.

Epistasis and sex

Negative epistasis and sex are thought to be intimately correlated. Experimentally, this idea has been tested in using digital simulations of asexual and sexual populations. Over time, sexual populations move towards more negative epistasis, or the lowering of fitness by two interacting alleles. It is thought that negative epistasis allows individuals carrying the interacting deleterious mutations to be removed from the populations efficiently. This removes those alleles from the population, resulting in an overall more fit population. This hypothesis was proposed by Alexey Kondrashov, and is sometimes known as the *deterministic mutation hypothesis* and has also been tested using artificial gene networks.

However, the evidence for this hypothesis has not always been straightforward and the model proposed by Kondrashov has been criticized for assuming mutation parameters far from real world observations. In addition, in those tests which used artificial gene networks, negative epistasis is only found in more densely connected networks, whereas empirical evidence indicates that natural gene networks are sparsely connected, and theory shows that selection for robustness will favor more sparsely connected and minimally complex networks.

Functional or mechanistic classification

- **Genetic suppression** - the double mutant has a less severe phenotype than either single mutant. [This term can also apply to a case where the double mutant has a phenotype intermediate between those of the single mutants, in which case the more severe single mutant phenotype is "suppressed" by the other mutation or genetic condition. For example, in a diploid organism, a hypomorphic (or partial loss-of-function) mutant phenotype can be suppressed by knocking out one copy of a gene that acts oppositely in the same pathway. In this case, the second gene is described as a "dominant suppressor" of the hypomorphic mutant; "dominant" because the effect is seen when one wild-type copy of the suppressor gene is present. For most genes, the phenotype of the heterozygous suppressor mutation

by itself would be wild type (because most genes are not haplo-insufficient), so that the double mutant (suppressed) phenotype is intermediate between those of the single mutants.]

- **Genetic enhancement** - the double mutant has a more severe phenotype than one predicted by the additive effects of the single mutants.
- **Synthetic lethality** or **unlinked non-complementation** - two mutations fail to complement and yet do not map to the same locus.
- **Intragenic complementation**, **allelic complementation**, or **interallelic complementation** - two mutations map to the same locus, yet the two alleles complement in the heteroallelic diploid. Causes of intragenic complementation include:
    - homology effects such as transvection, where, for example, an enhancer from one allele acts in *trans* to activate transcription from the promoter of the second allele.
    - trans-splicing of two mutant RNA molecules to produce a functional RNA.
    - At the protein level, another possibility involves proteins that normally function as dimers. In a heteroallelic diploid, two different abnormal proteins could form a functional dimer if each can compensate for the lack of function in the other.

Genetic screen

A **genetic screen** (often shortened to **screen**) is a procedure or test to identify and select individuals who possess a phenotype of interest. A genetic screen for new genes is often referred to as **forward genetics** as opposed to **reverse genetics**, the term for identifying mutant alleles in genes that are already known. Mutant alleles that are not tagged for rapid cloning are mapped and cloned by **positional cloning**.

Creating a mutant population

Since unusual alleles and phenotypes are rare, geneticists expose the individuals that are to be screened to a mutagen, such as a chemical or radiation, which generates mutations in their chromosomes. The use of mutagens enables "saturation screens" one of the first of which was performed by Nobel laureates Christiane Nüsslein-Volhard and Eric Wieschaus. A saturation screen is performed to uncover every gene that is involved in a particular phenotype in a given species. This is done by screening and mapping genes until no new genes are found. Mutagens such as random DNA insertions by transformation or active transposons can also be used to generate new mutants. These techniques have the advantage of tagging the new alleles with a known molecular (DNA) marker that can facilitate the rapid identification of the gene.

Types of screen

A **basic screen** involves looking for a phenotype of interest in the mutated population. One might screen for obvious phenotypes such as fruit flies with no wings or an *Arabidopsis* flower with no petals.

More subtle is a **temperature sensitive screen** that involves temperature shifts to enhance the mutant phenotype. A population grown at low temperature would have a normal phenotype, however, the mutation in the particular gene would make it unstable at a higher temperature. A screen for temperature sensitivity in fruit flies, for example, might involve raising the temperature in the cage until some flies faint, then opening a portal to let the others escape. Individuals selected in a screen are liable to carry an unusual version of a gene involved in the phenotype of interest. An advantage of alleles found in this type of screen is that the mutant phenotype is conditional and can be activated by simply raising the temperature. A null mutation in such a gene may be lethal to the embryo and such mutants would be missed in a basic screen.

An **enhancer/suppressor screen** is the most sophisticated type of genetic screen. In this case a mutagenised population has an allele of a gene that leads to a weak mutant phenotype in the biological process of interest. For example, with regard to fruit fly wing development, a weak allele may have small abnormal wings whereas a strong/null allele would have no wings. In this sensitised background it is possible to discover new mutants that either enhance the phenotype (small wings to no wings) or suppress the phenotype (small wings to normal wings). Such a screen has two advantages. First, new genes identified in the screen are often involved in the same biological process as the weak allele in the genetic background, in this case wing formation. Second, due to genetic redundancy, the mutant genes discovered may not have a visible phenotype of their own. In a more basic screen these would not be discovered, however, in the sensitised genetic background a visible phenotype is clear.

Mapping mutants

By the classical genetics approach, a researcher would then locate (map) the gene on its chromosome by crossbreeding with individuals that carry other unusual traits and collecting statistics on how frequently the two traits are inherited together. Classical geneticists would have used phenotypic traits to map the new mutant alleles. With the advent of genomic sequences for model systems such as *Drosophila, Arabidopsis* and *C. elegans* many SNPs have now been identified that can be used as traits for mapping. SNPs are the preferred traits for mapping since they are very frequent, on the order of one difference per 1000 base pairs, between different varieties of organism.

Positional cloning

Positional cloning is a method of gene identification in which a gene for a specific phenotype is identified, with only its approximate chromosomal location (but not the function) known, also known as the candidate region. Initially, the candidate region can

be defined using techniques such as linkage analysis, and positional cloning is then used to narrow the candidate region until the gene and its mutations are found. Positional cloning typically involves the isolation of partially overlapping DNA segments from genomic libraries to progress along the chromosome toward a specific gene. During the course of positional cloning, one needs to determine whether the DNA segment currently under consideration is part of the gene.

Tests used for this purpose include cross-species hybridization, identification of unmethylated CpG islands, exon trapping, direct cDNA selection, computer analysis of DNA sequence, mutation screening in affected individuals, and tests of gene expression. For genomes in which the regions of genetic polymorphisms are known, positional cloning involves identifying polymorphisms that flank the mutation. This process requires that DNA fragments from the closest known genetic marker are progressively cloned and sequenced, getting closer to the mutant allele with each new clone. This process produces a contig map of the locus and is known as chromosome walking. With the completion of genome sequencing projects such as the Human Genome Project, modern positional cloning can use ready-made contigs from the genome sequence databases directly.

For each new DNA clone a polymorphism is identified and tested in the mapping population for its recombination frequency compared to the mutant phenotype. When the DNA clone is at or close to the mutant allele the recombination frequency should be close to zero. If the chromosome walk proceeds through the mutant allele the new polymorphisms will start to show increase in recombination frequency compared to the mutant phenotype. Depending on the size of the mapping population, the mutant allele can be narrowed down to a small region (<30 Kb). Sequence comparison between wild type and mutant DNA in that region is then required to locate the DNA mutation that causes the phenotypic difference.

Modern positional cloning can more directly extract information from genomic sequencing projects and existing data by analyzing the genes in the candidate region. Potential disease genes from the candidate region can then be prioritized, potentially reducing the amount of work involved. Genes with expression patterns consistent with the disease phenotype, showing a (putative) function related to the phenotype, or homologous to another gene linked to the phenotype are all priority candidates. Generalization of positional cloning techniques in this manner is also known as positional gene discovery.

Positional cloning is an effective method to isolate disease genes in an unbiased manner, and has been used to identify disease genes for Duchenne Muscular Dystrophy, Huntington's and Cystic Fibrosis. However, complications in the analysis arise if the disease exhibits locus heterogeneity.

Chapter 15

Haplotype and Introgression

Haplotype

A **haplotype** in genetics is a combination of alleles (DNA sequences) at different places (loci) on the chromosome that are transmitted together. A haplotype may be one locus, several loci, or an entire chromosome depending on the number of recombination events that have occurred between a given set of loci.

In a second meaning, haplotype is a set of single-nucleotide polymorphisms (SNPs) on a single chromosome of a chromosome pair that are statistically associated. It is thought that these associations, and the identification of a few alleles of a haplotype block, can unambiguously identify all other polymorphic sites in its region. Such information is very valuable for investigating the genetics behind common diseases, and has been investigated in the human species by the International HapMap Project.

Many genetic testing companies use the term 'haplotype' to refer to an individual collection of short tandem repeat (STR) allele mutations within a genetic segment, while using the term 'haplogroup' to refer to the SNP/unique-event polymorphism (UEP) mutations which represents the clade to which a collection of potential haplotypes belong.

Haplotype resolution

An organism's genotype may not uniquely define its haplotype. For example, consider a diploid organism and two bi-allelic loci on the same chromosome such as single-nucleotide polymorphisms (SNPs). The first locus has alleles *A* and *T* with three possible genotypes *AA*, *AT*, and *TT*, the second locus having *G* and *C*, again giving three possible genotypes *GG*, *GC*, and *CC*. For a given individual, there are therefore nine possible configurations for the genotypes at these two loci, as shown in the Punnett square below, which shows the possible genotypes that an individual may carry and the corresponding haplotypes that these resolve to. For individuals that are homozygous at one or both loci,

it is clear what the haplotypes are; it is only when an individual is heterozygous at both loci that the gametic phase is ambiguous.

	AA	**AT**	**TT**
GG	AG AG	AG TG or AG TC	TG TG
GC	AG AC	AC TG	TG TC
CC	AC AC	AC TC	TC TC

The only unequivocal method of resolving phase ambiguity is by sequencing. However, it is possible to estimate the probability of a particular haplotype when phase is ambiguous using a sample of individuals.

Given the genotypes for a number of individuals, the haplotypes can be inferred by haplotype resolution or haplotype phasing techniques. These methods work by applying the observation that certain haplotypes are common in certain genomic regions. Therefore, given a set of possible haplotype resolutions, these methods choose those that use fewer different haplotypes overall. The specifics of these methods vary - some are based on combinatorial approaches (e.g., parsimony), whereas others use likelihood functions based on different models and assumptions such as the Hardy-Weinberg principle, the coalescent theory model, or perfect phylogeny. These models are combined with optimization algorithms such as expectation-maximization algorithm (EM), Markov chain Monte Carlo (MCMC), or hidden Markov models (HMM).

Microfluidic whole genome haplotyping is a technique for the physical separation of individual chromosomes from a metaphase cell followed by direct resolution of the haplotype for each allele.

Y-DNA haplotypes from genealogical DNA tests

Unlike other chromosomes, Y chromosomes do not come in pairs. Every human male has only one copy of that chromosome. This means that there is no lottery as to which copy to inherit, and also (for most of the chromosome) no shuffling between copies by recombination; so, unlike autosomal haplotypes, there is therefore effectively no randomisation of the Y-chromosome haplotype between generations, and a human male should largely share the same Y chromosome as his father, give or take a few mutations.

In particular, the Y-DNA that is the numbered results of a Y-DNA genealogical DNA test should match, barring mutations. Within genealogical and popular discussion, this is sometimes referred to as the "DNA signature" of a particular male human, or of his paternal bloodline.

UEP results (SNP results)

Unique-event polymorphisms (UEPs) like SNPs represent haplogroups. STRs represent haplotypes. The results that make up the full Y-DNA haplotype from the Y chromosome DNA test can be divided into two parts: the results for UEPs, sometimes loosely called the SNP results as most UEPs are single-nucleotide polymorphisms, and the results for microsatellite short tandem repeat sequences (Y-STRs).

The UEP results reflect the inheritance of events it is believed can be assumed to have happened only once in all human history. These can be used to directly identify the individual's Y-DNA haplogroup, his place on the broad family tree of the whole of humanity. Different Y-DNA haplogroups identify genetic populations which are often intricately geographically oriented, reflecting the migrations of current individuals' direct patrilineal ancestors tens of thousands of years ago.

Y-STR haplotypes

The other possible part of the genetic results is the **Y-STR haplotype**, the set of results from the Y-STR markers tested.

Unlike the UEPs, the Y-STRs mutate much more easily, which gives them much more resolution to distinguish recent genealogy. But it also means that, rather than the population of descendants of a genetic event all sharing the *same* result, the Y-STR haplotypes are likely to have spread apart, to form a *cluster* of more or less similar results. Typically, this cluster will have a definite most probable center, the **modal haplotype** (presumably close to the haplotype of the original founding event), and also a **haplotype diversity** — the degree to which it has become spread out. The further in the past the defining event occurred, and the more that subsequent population growth occurred early, the greater the haplotype diversity for a particular number of descendants will be. On the other hand, if the haplotype diversity is smaller for a particular number of descendants, this may indicate a more recent common ancestor, or that a population expansion has occurred more recently.

It is important to note that, unlike for UEPs, there is no guarantee that two individuals with a similar Y-STR haplotype will necessarily share a similar ancestry. There is no uniqueness about Y-STR events. Instead, the clusters of Y-STR haplotype results inheriting from different events and different histories all tend to overlap.

Thus, although sometimes a Y-STR haplotype may be directly indicative of a particular Y-DNA haplogroup, it is in most cases a long time since the haplogroups' defining events, so typically the cluster of Y-STR haplotype results associated with descendents of that event has become rather broad, and will tend to significantly overlap the (similarly broad) clusters of Y-STR haplotypes associated with other haplogroups, making it impossible to predict with absolute certainty to which Y-DNA haplogroup a Y-STR haplotype would point. All that can be done from the Y-STRs, if the UEPs are not

actually tested, is to predict probabilities for haplogroup ancestry (as this online program does), but not certainties.

A similar scenario exists for surnames. A cluster of similar Y-STR haplotypes may indicate a shared common ancestor, with an identifiable modal haplotype, but only if the cluster is sufficiently distinct from what may have arisen by chance from different individuals historically having adopted the same name independently. This may require the typing of quite an extensive haplotype to establish, which has fuelled DNA testing companies to offer ever-larger sets of markers - 24 then 37 then 67, and perhaps soon even more.

Plausibly establishing relatedness between different surnames data-mined from a database is significantly harder, because now it must be established not that a *randomly-selected* member of the population is unlikely to have such a close match by accident, but rather that the *very nearest* member of the population in question, chosen purposely from the population for that very reason, would even under those circumstances be unlikely to match by accident. This is for the foreseeable future likely to be impossible, except in special cases where there is further information to drastically limit the size of that population of candidates under consideration.

Introgression

Introgression, also known as **introgressive hybridization**, in genetics (particularly plant genetics), is the movement of a gene (gene flow) from one species into the gene pool of another by repeated backcrossing an interspecific hybrid with one of its parent species. Purposeful introgression is a long-term process; it may take many hybrid generations before the backcrossing occurs.

Introgression is an important source of genetic variation in natural populations and major cause of speciation in the sympatric mode. It can have important effects on the dynamics of hybrid zones, speciation and adaptive radiation. There is evidence that introgression is a ubiquitous phenomenon in plants, in animals, and even in humans, where it may have introduced the microcephalin D allele.

Introgression differs from the simple hybridization. Introgression results in a complex mixture of parental genes, while simple hybridization results in a more uniform mixture, which in the first generation will be an even mix of two parental species. Natural introgression does not have the human direct interference while the exotic introgression is induced intentionally (as for instance genetically modified organisms) or not (human activities affecting local races of crop, human disturbances like in introducing weeds).

An example of introgression is that of a transgene from a transgenic plant to a wild relative as the result of a successful hybridization leading to intentional or unintentional

"genetic pollution". Another important example has been studied by Arnold & Bennett 1993: irises species from southern Louisiana.

An **introgression line** (abbreviation: IL) in plant molecular biology is a line of a crop species that contains genetic material derived from a similar species, for example a "wild" relative. An example of a collection of ILs (called *IL-Library*) is the use of chromosome fragments from *Solanum pennellii* (a wild variety of tomato) introgressed in *Solanum lycopersicum* (the cultivated tomato). The lines of an IL-Library covers usually the complete genome of the donor. Introgression lines allow the study of quantitative trait loci, but also the creation of new varieties by introducing exotic traits.

Chapter 16

Monohybrid Cross

Monohybrid Cross - a method of finding out the inheritance pattern of a trait between two single organisms.

A **monohybrid cross** is a cross between parents who are heterozygous at one locus; for example, Bb x Bb. Example: B = brown. b = blue. BB = Dark Brown. Bb = Brown (not blue). bb = Blue.

Monohybrid inheritance is the inheritance of a single characteristic. The different forms of the characteristic are usually controlled by different alleles of the same gene. For example, a monohybrid cross between two pure-breeding plants (homozygous for their respective traits), one with yellow seeds (the dominant trait) and one with green seeds (the recessive trait), would be expected to produce an F1 (first) generation with only yellow seeds because the allele for yellow seeds is dominant to that of green. A monohybrid cross compares only one trait.

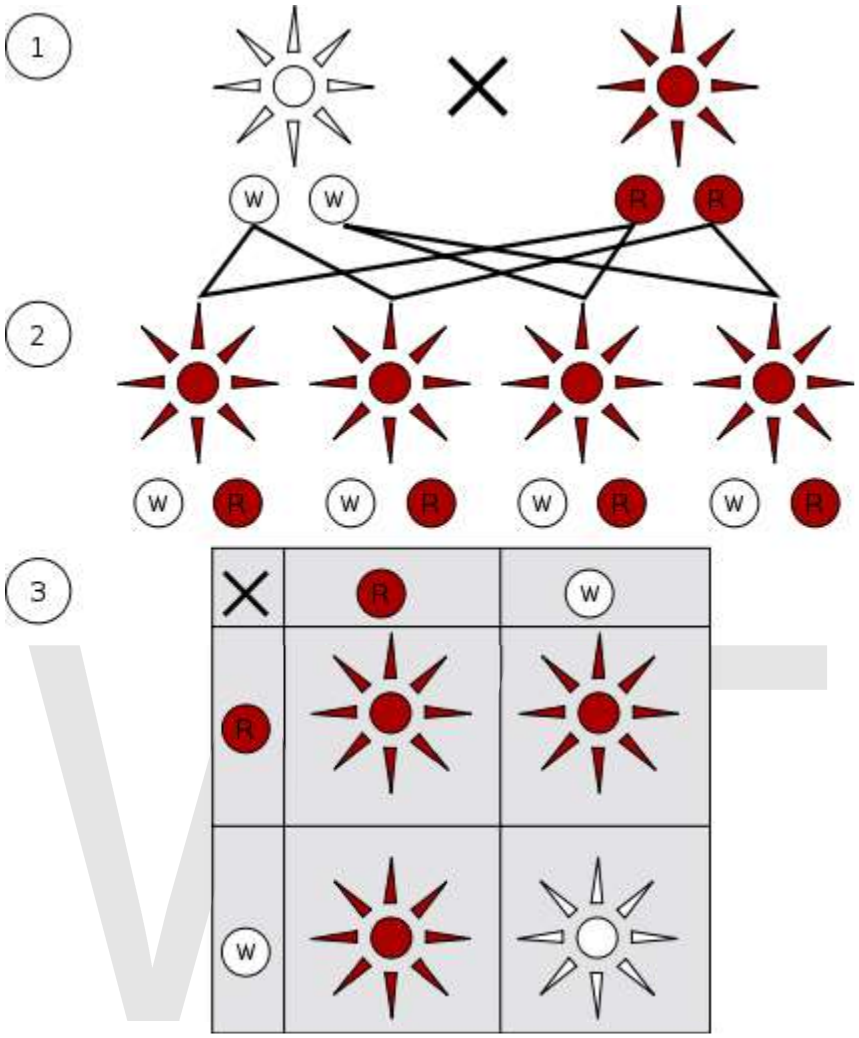


Figure 1 : Inheritance pattern of dominant (red) and recessive (white) phenotypes when each parent (1) is homozygous for either the dominant or recessive trait. All members of the F_1 generation are heterozygous and share the same dominant phenotype (2), while the F_2 generation exhibits a 3:1 ratio of dominant to recessive phenotypes (3).

Usage of Monohybrid Cross

Generally, the monohybrid cross is used to determine the F2 generation from a pair of homozygous grandparents (one grandparent dominant, the other recessive) which results in an F1 generation that are all heterozygous. Crossing two heterozygous parents from the F1 generation results in an F2 generation that produces a 75% chance for the appearance of the dominant phenotype, of which two-thirds are heterozygous, and a 25% chance for the appearance of the recessive phenotype. This cross was originally used by biologist, Gregor Mendel, who crossed two pea plants to obtain a hybrid variety, discovering the possible changes in phenotypes of various alleles.

Introduction

Gregor Mendel (1822–1884) was an Austrian monk who discovered the basic rules of inheritance. From 1858 to 1866, he bred garden peas in his monastery garden and analyzed the offspring of these matings. The garden pea was good choice of experimental organism because: many varieties were available that bred true for clear-cut, qualitative traits like seed texture (round vs wrinkled) seed color (green vs yellow) flower color (white vs purple) tall vs dwarf growth habit and three others that also varied in a qualitative - rather than quantitative - way. peas are normally self-pollinated because the stamens and carpels are enclosed within the petals. By removing the stamens from unripe flowers, Mendel could brush pollen from another variety on the carpels when they ripened.

The results

All the peas produced in the second or hybrid generation were round.

Our interpretation

All the peas of this F1 generation have an Rr genotype. All the haploid sperm and eggs produced by meiosis received one chromosome 7. All the zygotes received one R allele (from the round parent) and one r allele (from the wrinkled parent). Because the round trait is dominant, the phenotype of all the seeds was round.

		P gametes (round parent)	
		R	**R**
P gametes (wrinkled parent)	**r**	Rr	Rr
	r	Rr	Rr

The second cross

Mendel then allowed his hybrid peas to self-pollinate. The wrinkled trait — which had disappeared in his hybrid generation — reappeared in 25% of the new crop of peas.

Interpretation

Random union of equal numbers of R and r gametes produced an F2 generation with 25% RR and 50% Rr - both with the round phenotype - and 25% rr with the wrinkled phenotype.

		F1 gametes	
		R	**r**
F1 gametes	**R**	RR	Rr
	r	Rr	rr

The third cross

Mendel then allowed some of each phenotype in the F2 generation to self-pollinate. His results: All the wrinkled seeds in the F2 generation produced only wrinkled seeds in the F3. One-third (193/565) of the round F1 seeds produced only round seeds in the F3 generation, but two-thirds (372/565) of them produced both types of seeds in the F3 and - once again - in a 3:1 ratio.

Interpretation

One-third of the round seeds and all of the wrinkled seeds in the F2 generation were homozygous and produced only seeds of the same phenotype.

But two thirds of the round seeds in the F2 were heterozygous and their self-pollination produced both phenotypes in the ratio of a typical F1 cross.

Phenotype ratios are approximate The union of sperm and eggs is random. As the size of the sample gets larger, however, chance deviations become minimized and the ratios approach the theoretical predictions more closely. The table shows the actual seed production by ten of Mendel's F1 plants. While his individual plants deviated widely from the expected 3:1 ratio, the group as a whole approached it quite closely.

Round	Wrinkled
45	12
27	8
24	7
19	16
32	11
26	6
88	24
22	10
28	6
25	7
Total: 336	**Total: 107**

Mendel's Hypothesis

To explain his results, Mendel formulated a hypothesis that included the following: In the organism there is a pair of factors that controls the appearance of a given characteristic. (We call them genes.) The organism inherits these factors from its parents, one from each. Each is transmitted from generation to generation as a discrete, unchanging unit. (The wrinkled seeds in the F2 generation were no less wrinkled than those in the P generation although they had passed through the round-seeded F1 generation.) When the gametes are formed, the factors separate and are distributed as units to each gamete. This statement is often called Mendel's rule of segregation. If an organism has two unlike factors (we call them alleles) for a characteristic, one may be expressed to the total exclusion of the other (dominant vs recessive).

The Testcross: A Test of Mendel's Hypothesis

A good hypothesis meets several standards.

- It should provide an adequate explanation of the observed facts. If two or more hypotheses meet this standard, the simpler one is preferred.
- It should be able to predict new facts. So if a generalization is valid, then certain specific consequences can be deduced from it.

In order to test his hypothesis, Mendel predicted the outcome of a breeding experiment that he had not yet carried out. He crossed heterozygous round peas (Rr) with wrinkled (homozygous, rr) ones. He predicted that in this case one-half of the seeds produced would be round (Rr) and one-half wrinkled (rr).

		F1 gametes	
		R	**r**
P gametes	**r**	Rr	rr
	r	Rr	rr

To a casual observer in the monastery garden, the cross appeared no different from the P cross described above: round-seeded peas being crossed with wrinkled-seeded ones. But Mendel predicted that this time he would produce both round and wrinkled seeds and in a 50:50 ratio. He performed the cross and harvested 106 round peas and 101 wrinkled peas.

This kind of mating is called a testcross. It "tests" the genotype in those cases where two different genotypes (like RR and Rr) produce the same phenotype.

Mendel did not stop here. He went on to cross pea varieties that differed in six other qualitative traits. In every case, the results supported his hypothesis. He crossed peas that differed in two traits. He found that the inheritance of one trait was independent of that of the other and so framed his second rule: the rule of independent assortment. Today, we know this rule does not apply to some genes, due to genetic linkage.

Mendel's rules today

Little attention was paid when Mendel published his findings in 1866. Not until 1900, 34 years later and 16 years after his death, was his work brought to light. By then, three men — working independently — discovered the same principles. So the present remarkable development of genetics dates from only the start of the 20th century.

The discovery of chromosomes — and their behavior during meiosis (2n -> n) and fertilization (n + n -> 2n) — established the structural basis for Mendel's rules.

Although many important exceptions to them have been discovered (like complex dominance and genetic linkage), Mendel's rules still form the foundation upon which the science of genetics rests today.

Monohybrid Cross

Mendel did many mating crosses with pea plants. In this case a true-breeding tall plant was crossed with a true-breeding short plant. All of the plants in the next generation were tall. We now know that the tall allele (T) is dominant to the short allele (t) and the cross was of the form TT x tt.

Chapter 17

Phenotype

Individuals in the mollusk species *Donax variabilis* show diverse coloration and patterning in their phenotypes.

A **phenotype** is any *observable characteristic* or trait of an organism: such as its morphology, development, biochemical or physiological properties, behavior, and products of behavior (such as a bird's nest). Phenotypes result from the expression of an organism's genes as well as the influence of environmental factors and the interactions between the two.

The genotype of an organism is the inherited instructions it carries within its genetic code. Not all organisms with the same genotype look or act the same way because appearance and behavior are modified by environmental and developmental conditions. Similarly, not all organisms that look alike necessarily have the same genotype.

This genotype-phenotype distinction was proposed by Wilhelm Johannsen in 1911 to make clear the difference between an organism's heredity and what that heredity produces. The distinction is similar to that proposed by August Weismann, who distinguished between germ plasm (heredity) and somatic cells (the body). A more modern version is Francis Crick's central dogma of molecular biology.

Difficulties in definition

Despite its seemingly straightforward definition, the concept of the phenotype has some hidden subtleties. First, most of the molecules and structures coded by the genetic material are not visible in the appearance of an organism, yet they are observable (for example by Western blotting) and are thus part of the phenotype. Human blood groups are an example. So, by extension, the term phenotype must include characteristics that can be made visible by some technical procedure. Another extension adds behaviour to the phenotype since behaviours are also observable characteristics. Indeed there is research into the clinical relevance of behavioural phenotypes as they pertain to a range of syndromes. Often, the term "phenotype" is incorrectly used as a shorthand to indicate *phenotypical changes* observed in mutated organisms (most often in connection with knockout mice).

Biston betularia morpha *typica*, the standard light-colored Peppered Moth.

Biston betularia morpha *carbonaria*, the melanic Peppered Moth, illustrating discontinuous variation.

Phenotypic variation

Phenotypic variation (due to underlying heritable genetic variation) is a fundamental prerequisite for evolution by natural selection. It is the living organism as a whole that contributes (or not) to the next generation, so natural selection affects the genetic structure of a population indirectly via the contribution of phenotypes. Without phenotypic variation, there would be no evolution by natural selection.

The interaction between genotype and phenotype has often been conceptualized by the following relationship:

genotype + environment → phenotype

A slightly more nuanced version of the relationships is:

genotype + environment + random-variation → phenotype

Genotypes often have much flexibility in the modification and expression of phenotypes; in many organisms these phenotypes are very different under varying environmental conditions. The plant *Hieracium umbellatum* is found growing in two different habitats in Sweden. One habitat is rocky, sea-side cliffs, where the plants are bushy with broad leaves and expanded inflorescences; the other is among sand dunes where the plants grow prostrate with narrow leaves and compact inflorescences. These habitats alternate along the coast of Sweden and the habitat that the seeds of *Hieracium umbellatum* land in, determine the phenotype that grows.

An example of random variation in *Drosophila* flies is the number of ommatidia, which may vary (randomly) between left and right eyes in a single individual as much as they do between different genotypes overall, or between clones raised in different environments.

The concept of phenotype can be extended to variations below the level of the gene that affect an organism's fitness. For example, silent mutations that do not change the corresponding amino acid sequence of a gene may change the frequency of guanine-cytosine base pairs (GC content). These base pairs have a higher thermal stability than adenine-thymine, a property that might convey, among organisms living in high-temperature environments, a selective advantage on variants enriched in GC content.

The Extended Phenotype

The idea of the phenotype has been generalized by Richard Dawkins in *The Extended Phenotype* to mean all the effects a gene has on the outside world that may influence its chances of being replicated. These can be effects on the organism in which the gene resides, the environment, or other organisms.

For instance, a beaver dam might be considered a phenotype of beaver genes, the same way beavers' powerful incisor teeth are phenotype expressions of their genes. Dawkins also cites the effect of an organism on the behaviour of another organism (such as the devoted nurturing of a cuckoo by a parent clearly of a different species) as an example of the extended phenotype.

Phenome and phenomics

Although a phenotype is the ensemble of observable characteristics displayed by an organism, the word *phenome* is sometimes used to refer to a collection of traits and their simultaneous study as *phenomics*.

Chapter 18

Phenotypic Trait and Punnett Square

Phenotypic trait

A **trait** is a distinct variant of a phenotypic character of an organism that may be inherited, environmentally determined or somewhere in between. For example, eye color is a *character* or abstraction of an attribute, while blue, brown and hazel are *traits*.

Definition

A phenotypic trait is an obvious and observable trait; it is the expression of genes in an observable way. An example of a phenotypic trait is hair color, there are underlying genes that control the hair color, which make up the genotype, but the actual hair color, the part we see, is the phenotype. The phenotype is the physical characteristics of the organism. The phenotype is controlled by the genetic make-up of the organism and the environmental pressures the organism is subject to.

A trait may be any single feature or quantifiable measurement of an organism. However, the most useful traits for genetic analysis are present in different forms in different individuals.

A visible trait is the final product of many molecular and biochemical processes. In most cases, information starts with DNA traveling to RNA and finally to protein (ultimately affecting organism structure and function). This is the Central Dogma of molecular biology as stated by Francis Crick.

This information flow may also be followed through the cell as it travels from the DNA in the nucleus, to the Cytoplasm, to the Ribosomes and the Endoplasmic Reticulum, and finally to the Golgi Apparatus, which may package the final products for export outside the cell.

Cell products are released into the tissue, and organs of an organism, to finally affect the physiology in a way that produces a trait.

Genetic origin of traits in diploid organisms

The heritable unit that may influence a trait is called a gene. A gene is a portion of a chromosome. An important reference point along a chromosome, which is a very long and compacted string of DNA, is the centromere; the distance from a gene to the centromere is referred to as the gene's locus or map location. A chromosomal region known to control a trait while the responsible gene within not being identified is referred to as a quantitative trait locus.

The nucleus of a diploid cell contains two of each chromosome, with homologous (mostly identical) pairs of chromosomes having the same genes at the same loci.

Different phenotypic traits are caused by different forms of genes, or alleles, which arise by mutation in a single individual and are passed on to successive generations.

Mendelian expression of genes in diploid organisms

A gene is only a DNA code sequence; the slightly different variations of that sequence are called alleles. Alleles can be significantly different and produce different product RNAs.

Combinations of different alleles thus go on to generate different traits through the information flow charted above. For example, if the alleles on homologous chromosomes exhibit a "simple dominance" relationship, the trait of the "dominant" allele shows in the phenotype.

Gregor Mendel pioneered modern genetics. His most famous analyses were based on clear-cut traits with simple dominance. He determined that the heritable units, what he called "genes", occurred in pairs and could exhibit linkage. His tool was statistics: long before the molecular model of DNA was introduced by James D. Watson and Francis Crick.

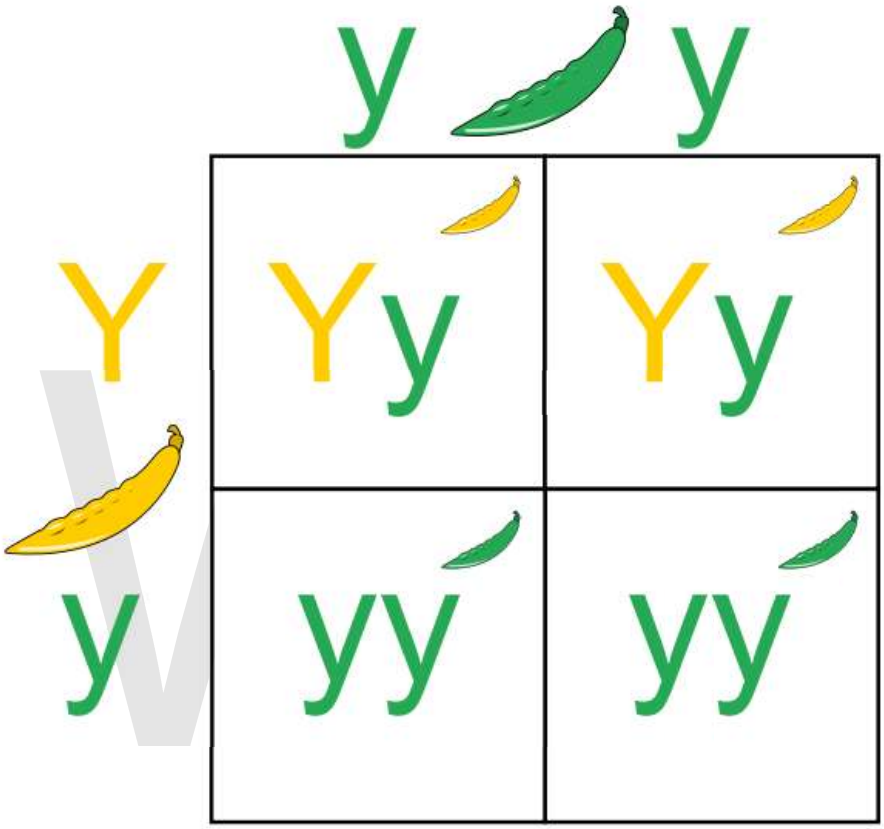
Some examples of Inherited genes include eye color.

Biochemistry of dominance and extensions to expression of traits

The biochemistry of the intermediate proteins determines how they interact in the cell. Therefore, biochemistry predicts how combinations of different alleles will produce varying traits.

Extended expression patterns seen in diploid organisms include facets of incomplete dominance, codominance, and multiple alleles.

Punnett square



A Punnett square showing a typical test cross

The **Punnett square** is a diagram that is used to predict an outcome of a particular cross or breeding experiment. It is named after Reginald C. Punnett, who devised the approach, and is used by biologists to determine the probability of an offspring having a particular genotype. The Punnett square is a summary of every possible combination of one maternal allele with one paternal allele for each gene being studied in the cross.

Monohybrid cross

In this example, both organisms have the genotype *Bb*. They can produce gametes that contain either the *B* or *b* allele. (It is conventional in genetics to use capital letters to indicate dominant alleles and lower-case letters to indicate recessive alleles.) The

probability of an individual offspring having the genotype *BB* is 25%, *Bb* is 50%, and *bb* is 25%.

		Maternal	
		B	**b**
Paternal	**B**	BB	Bb
	b	Bb	bb

It is important to note that Punnett squares give probabilities only for *genotypes*, not *phenotypes*. The way in which the *B* and *b* alleles interact with each other to affect the appearance of the offspring depends on how the gene products (proteins) interact. For classical dominant/recessive genes, like that which determines whether a rat has black hair (*B*) or white hair (*b*), the dominant allele will mask the recessive one. Thus in the example above 75% of the offspring will be black (*BB* or *Bb*) while only 25% will be white (*bb*). The ratio of the phenotypes is 3:1, typical for a monohybrid cross.

Dihybrid cross

More complicated crosses can be made by looking at two or more genes. The Punnett square only works, however, if the genes are independent of each other, which means that having a particular allele of gene X does not imply having a particular allele of gene Y.

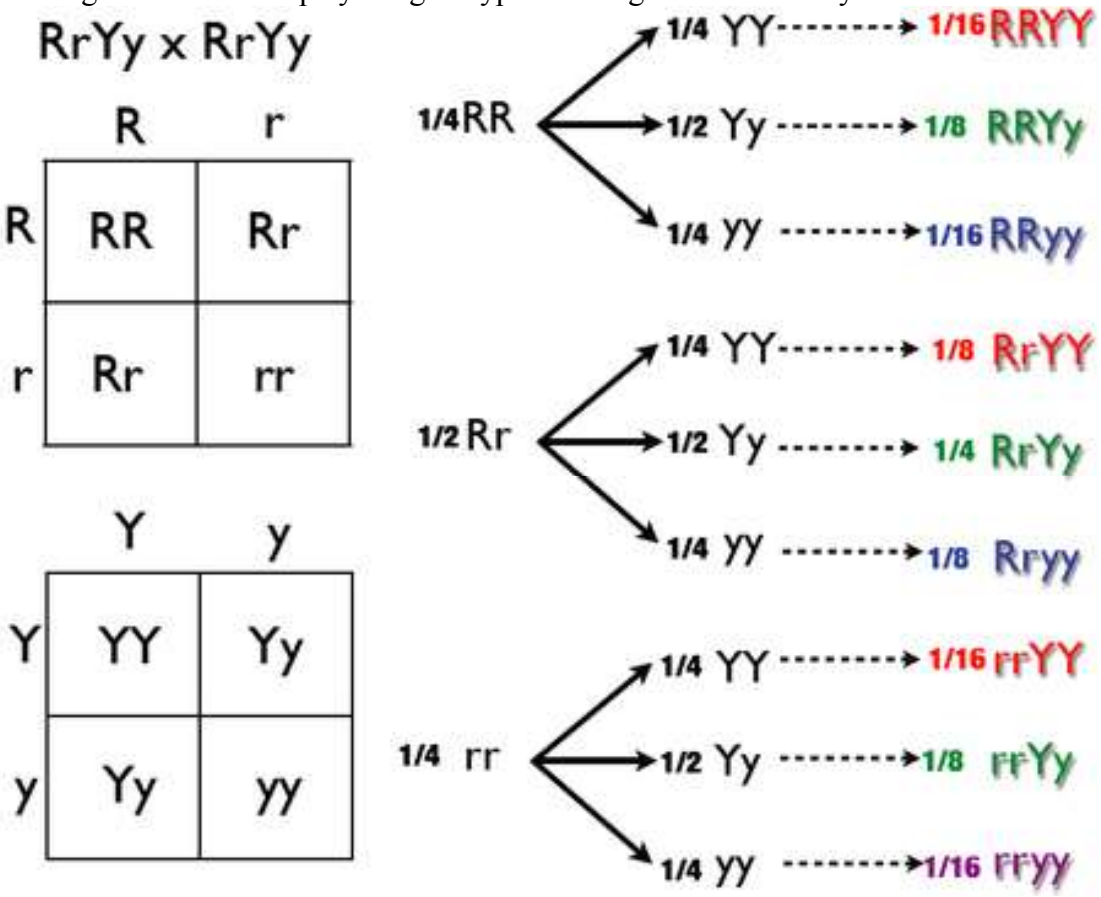
The following example illustrates a dihybrid cross between two heterozygous pea plants. *R* represents the dominant allele for shape (round), while *r* represents the recessive allele (wrinkled). *Y* represents the dominant allele for color (yellow), while *y* represents the recessive allele (green). If each plant has the genotype *RrYy*, and since the alleles for shape and color genes are independent, then they can produce four types of gametes with all possible combinations: *RY*, *Ry*, *rY* and *ry*.

	RY	**Ry**	**rY**	**ry**
RY	RRYY	RRYy	RrYY	RrYy
Ry	RRYy	RRyy	RrYy	Rryy
rY	RrYY	RrYy	rrYY	rrYy
ry	RrYy	Rryy	rrYy	rryy

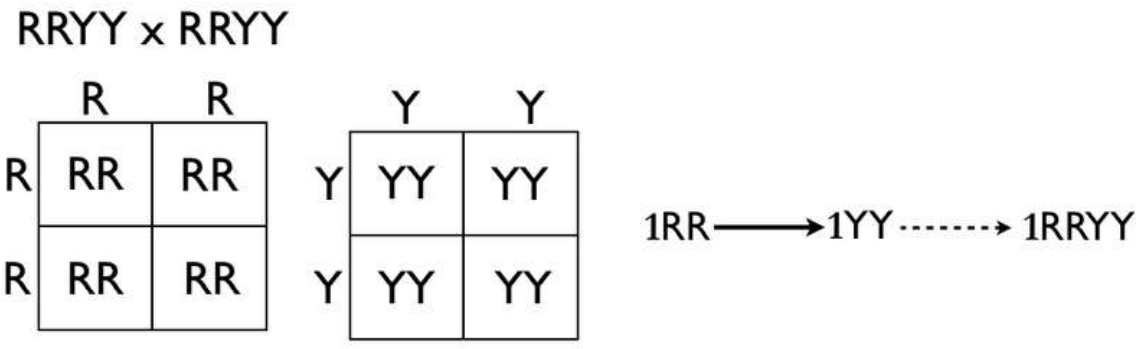
Since dominant traits mask recessive traits, there are nine combinations that have the phenotype round yellow, three that are round green, three that are wrinkled yellow and one that is wrinkled green. The ratio 9:3:3:1 is typical for a dihybrid cross.

Tree method

Another way to solve dihybrid and multihybrid crosses is to use the tree method, although it does not display the genotypes of the gametes correctly.



This method is particularly advantageous when crossing homozygous organisms.



Situations where Punnett squares need to be used with care

The phenotypic ratios of 3:1 and 9:3:3:1 are theoretical predictions based on the assumptions of segregation and independent assortment of alleles. Deviations from expected ratios can occur if any of the following conditions exists:

- the alleles in question are on the same chromosome and linked
- one parent lacks a copy of the gene, e.g. human males have only one X chromosome, from their mother, so only the maternal alleles have an effect on the organism
- the survival rate of different genotypes is not the same, e.g. one combination of alleles may be incompatible with life so that the affected offspring expires *in utero*
- alleles may show incomplete dominance or co-dominance
- there are genetic interactions (epistasis) between alleles of different genes
- the trait is inherited on genetic material from only one parent, e.g. mitochondrial DNA is only inherited from the mother
- the alleles are imprinted

Chapter 19

Quantitative Trait Locus

Quantitative traits refer to phenotypes (characteristics) that vary in degree and can be attributed to polygenic effects, i.e., product of two or more genes, and their environment. **Quantitative trait loci** (QTLs) are stretches of DNA containing or linked to the genes that underlie a quantitative trait. Mapping regions of the genome that contain genes involved in specifying a quantitative trait is done using molecular tags such as AFLP or, more commonly SNPs. This is an early step in identifying and sequencing the actual genes underlying trait variation.

Quantitative traits

Polygenic inheritance, also known as **quantitative** or **multifactorial inheritance** refers to inheritance of a phenotypic characteristic (trait) that is attributable to two or more genes, or the interaction with the environment, or both. Unlike monogenic traits, polygenic traits do not follow patterns of Mendelian inheritance (separated traits). Instead, their phenotypes typically vary along a continuous gradient depicted by a bell curve.

An example of a polygenic trait is human skin color. Many genes factor into determining a person's natural skin color, so modifying only one of those genes changes the color only slightly. Many disorders with genetic components are polygenic, including autism, cancer, diabetes and numerous others. Most phenotypic characteristics are the result of the interaction of multiple genes.

Examples of disease processes generally considered to be results of *multifactorial etiology*:

Congenital malformation

- Cleft palate
- Congenital dislocation of the hip
- Congenital heart defects
- Neural tube defects

- Pyloric stenosis
- Talipes

Adult onset diseases

- Diabetes Mellitus
- Cancer
- Epilepsy
- Glaucoma
- Hypertension
- Ischaemic heart disease
- Manic depression
- Schizophrenia

Multifactorially inherited diseases are said to constitute the majority of genetic disorders affecting humans which will result in hospitalization or special care of some kind.

Multifactorial traits in general

Generally, multifactorial traits outside of illness contribute to what we see as **continuous characteristics** in organisms, such as height, skin color, and body mass. All of these phenotypes are complicated by a great deal of interplay between genes and environment. The continuous distribution of traits such as height and skin colour described above reflects the action of genes that do not quite show typical patterns of dominance and recessiveness. Instead the contributions of each involved locus are thought to be additive. Writers have distinguished this kind of inheritance as *polygenic*, or *quantitative inheritance*.

Thus, due to the nature of polygenic traits, inheritance will not follow the same pattern as a simple monohybrid or dihybrid cross. Polygenic inheritance can be explained as Mendelian inheritance at many loci, resulting in a trait which is normally-distributed. If n is the number of involved loci, then the coefficients of the binomial expansion of $(a + b)^{2n}$ will give the frequency of distribution of all n allele combinations. For a sufficiently high n, this binomial distribution will begin to resemble a normal distribution. From this viewpoint, a disease state will become apparent at one of the tails of the distribution, past some threshold value. Disease states of increasing severity will be expected the further one goes past the threshold and away from the mean.

Heritable disease and multifactorial inheritance

A mutation resulting in a disease state is often recessive, so both alleles must be mutant in order for the disease to be expressed phenotypically. A disease or syndrome may also be the result of the expression of mutant alleles at more than one locus. When more than one gene is involved with or without the presence of environmental triggers, we say that the disease is the result of multifactorial inheritance.

The more genes involved in the cross, the more the distribution of the genotypes will resemble a normal, or Gaussian distribution. This shows that multifactorial inheritance is polygenic, and genetic frequencies can be predicted by way of a polyhybrid Mendelian cross. Phenotypic frequencies are a different matter, especially if they are complicated by environmental factors.

The paradigm of polygenic inheritance as being used to define multifactorial disease has encountered much disagreement. Turnpenny (2004) discusses how simple polygenic inheritance cannot explain some diseases such as the onset of Type I diabetes mellitus, and that in cases such as these, not all genes are thought to make an equal contribution.

The assumption of polygenic inheritance is that all involved loci make an equal contribution to the symptoms of the disease. This should result in a normal curve distribution of genotypes. When it does not, the idea of polygenetic inheritance cannot be supported for that illness.

A cursory look at some examples

Examples of such diseases are not new to medicine. The above examples are well-known examples of diseases having both genetic and environmental components. Other examples involve atopic diseases such as eczema or dermatitis; also spina bifida (open spine) and anencephaly (open skull) are other examples

While schizophrenia is widely believed to be multifactorially genetic by biopsychiatrists, no characteristic genetic markers have been determined with any certainty.

Is it multifactorially heritable?

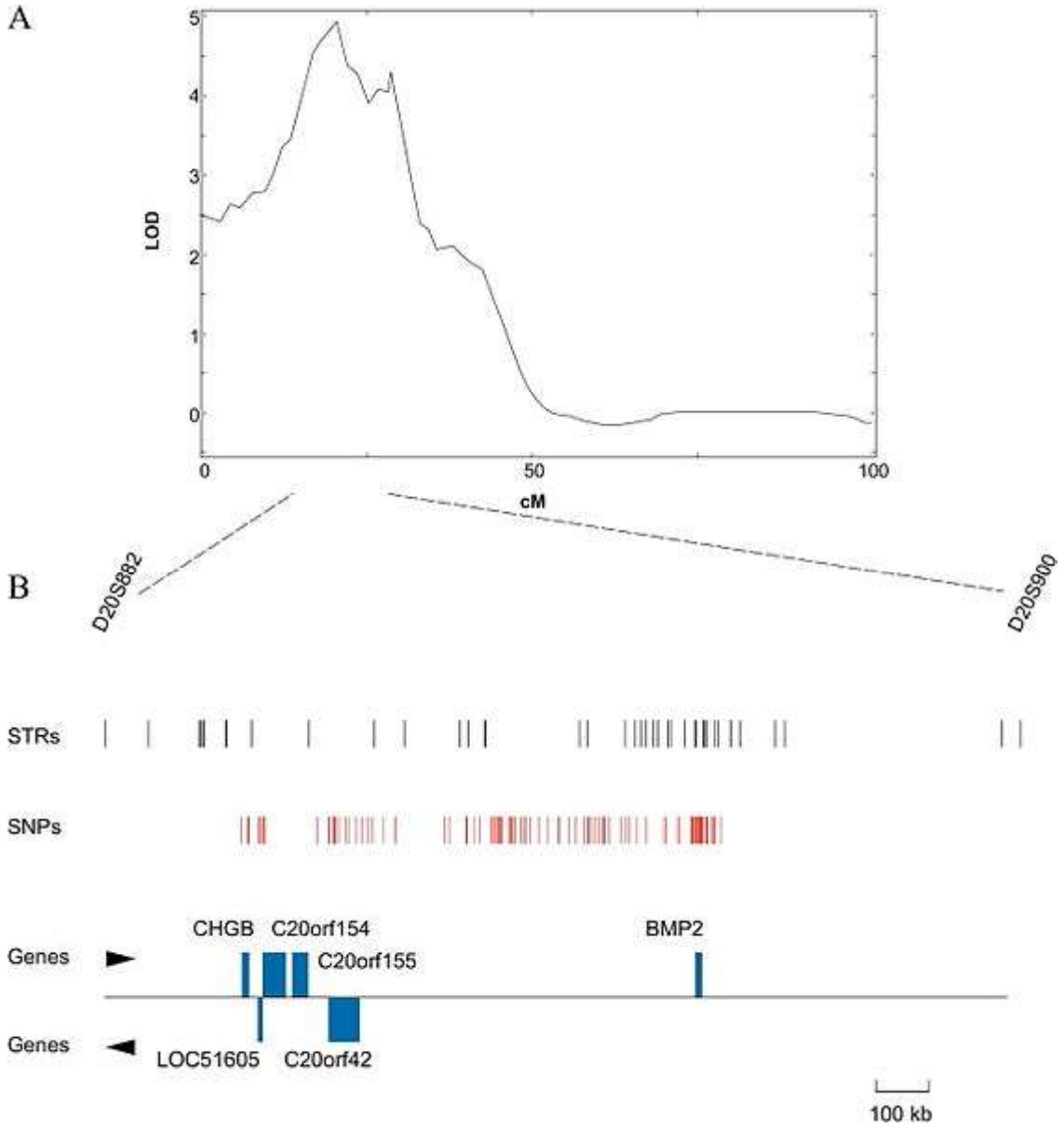
It is difficult to ascertain if any particular disease is multifactorially genetic. If a pedigree chart is taken of the patient's family and relations, and it is shown that the brothers and sisters of the patient have the disease, then there is a strong chance that the disease is genetic and that the patient will also be a genetic carrier. But this is not quite enough. It also needs to be proven that the pattern of inheritance is non-Mendelian. This would require studying dozens, even hundreds of different family pedigrees before a conclusion of multifactorial inheritance is drawn. This often takes several years.

If multifactorial inheritance is indeed the case, then the chance of the patient contracting the disease is reduced if only cousins and more distant relatives have the disease. It must be stated that while multifactorially-inherited disease tends to run in families, inheritance will not follow the same pattern as a simple monohybrid or dihybrid cross.

If a genetic cause is suspected and little else is known about the illness, then it remains to be seen exactly how many genes are involved in the phenotypic expression of the disease. Once that is determined, the question must be answered: if two people have the required genes, why are there differences in expression between them? Generally, what makes the two individuals different are likely to be environmental factors. Due to the involved

nature of genetic investigations needed to determine such inheritance patterns, this is not usually the first avenue of investigation one would choose to determine etiology.

Quantitative trait locus



A QTL for osteoporosis on the human chromosome 20

Typically, QTLs underlie continuous traits (those traits that vary continuously, e.g. height) as opposed to discrete traits (traits that have two or several character values, e.g. red hair in humans, a recessive trait, or smooth vs. wrinkled peas used by Mendel in his experiments).

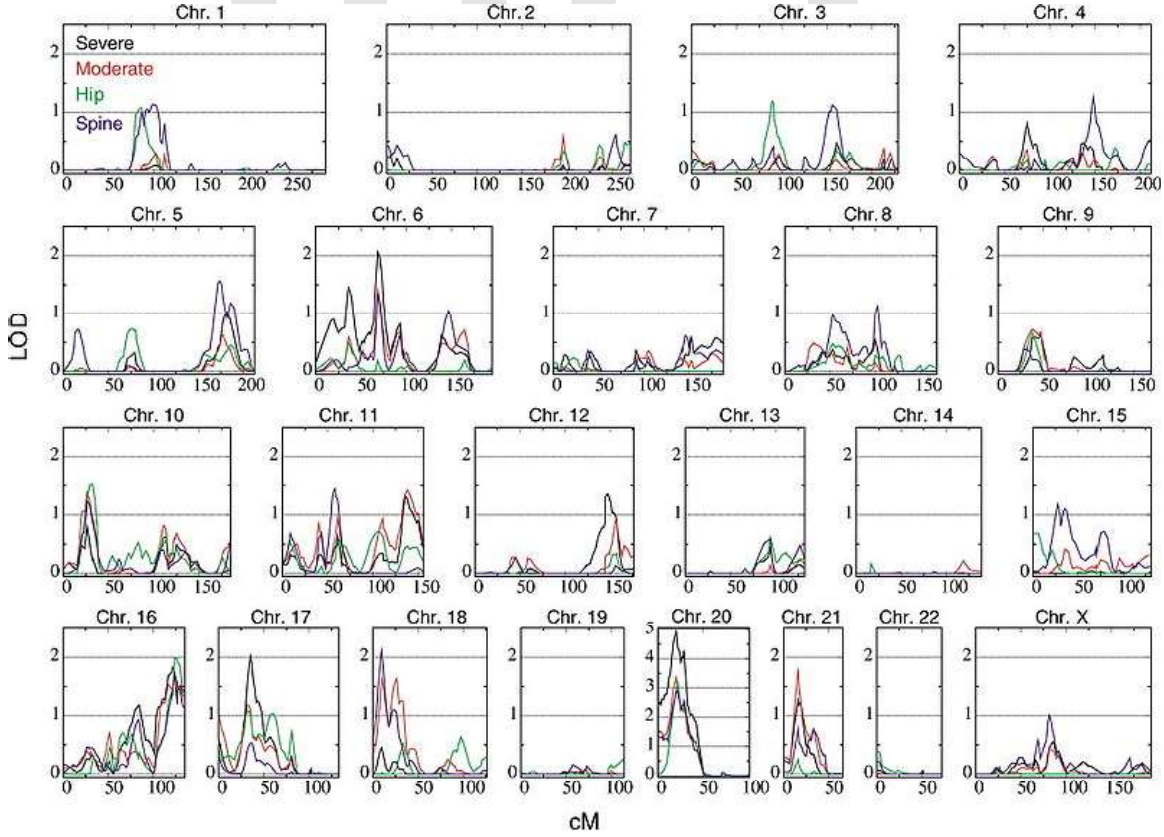
Moreover, a single phenotypic trait is usually determined by many genes. Consequently, many QTLs are associated with a single trait.

A **quantitative trait locus** (**QTL**) is a region of DNA that is associated with a particular phenotypic trait - these QTLs are often found on different chromosomes. Knowing the number of QTLs that explains variation in the phenotypic trait tells us about the genetic architecture of a trait. It may tell us that plant height is controlled by many genes of small effect, or by a few genes of large effect.

Another use of QTLs is to identify candidate genes underlying a trait. Once a region of DNA is identified as contributing to a phenotype, it can be sequenced. The DNA sequence of any genes in this region can then be compared to a database of DNA for genes whose function is already known.

In a recent development, classical QTL analyses are combined with gene expression profiling i.e. by DNA microarrays. Such expression QTLs (eQTLs) describe cis- and trans-controlling elements for the expression of often disease-associated genes. Observed epistatic effects have been found beneficial to identify the gene responsible by a cross-validation of genes within the interacting loci with metabolic pathway- and scientific literature databases.

QTL mapping



Example of a genome-wide scan for QTL of osteoporosis

QTL mapping is the statistical study of the alleles that occur at a locus and the phenotypes (physical forms or traits) that they produce. Because most traits of interest are governed by more than one gene, defining and studying the entire locus of genes related to a trait gives hope of understanding what effect the genotype of an individual might have in the real world.

Statistical analysis is required to demonstrate that different genes interact with one another and to determine whether they produce a significant effect on the phenotype. QTLs identify a particular region of the genome as containing a gene that is associated with the trait being assayed or measured. They are shown as intervals across a chromosome, where the probability of association is plotted for each marker used in the mapping experiment.

The QTL techniques were developed in the late 1980s and can be performed on inbred strains of any species..

To begin, a set of genetic markers must be developed for the species in question. A marker is an identifiable region of variable DNA. Biologists are interested in understanding the genetic basis of phenotypes (physical traits). The aim is to find a marker that is significantly more likely to co-occur with the trait than expected by chance, that is, a marker that has a statistical association with the trait. Ideally, they would be able to find the specific gene or genes in question, but this is a long and difficult undertaking. Instead, they can more readily find regions of DNA that are very close to the genes in question. When a QTL is found, it is often not the actual gene underlying the phenotypic trait, but rather a region of DNA that is closely linked with the gene.

For organisms whose genomes are known, one might now try to exclude genes in the identified region whose function is known with some certainty not to be connected with the trait in question. If the genome is not available, it may be an option to sequence the identified region and determine the putative functions of genes by their similarity to genes with known function, usually in other genomes. This can be done using BLAST, an online tool that allows users to enter a primary sequence and search for similar sequences within the BLAST database of genes from various organisms.

Another interest of statistical geneticists using QTL mapping is to determine the complexity of the genetic architecture underlying a phenotypic trait. For example, they may be interested in knowing whether a phenotype is shaped by many independent loci, or by a few loci, and do those loci interact. This can provide information on how the phenotype may be evolving.

Analysis of variance

The simplest method for QTL mapping is analysis of variance (ANOVA, sometimes called "marker regression") at the marker loci. In this method, in a backcross, one may calculate a t-statistic to compare the averages of the two marker genotype groups. For

other types of crosses (such as the intercross), where there are more than two possible genotypes, one uses a more general form of ANOVA, which provides a so-called F-statistic. The ANOVA approach for QTL mapping has three important weaknesses. First, we do not receive separate estimates of QTL location and QTL effect. QTL location is indicated only by looking at which markers give the greatest differences between genotype group averages, and the apparent QTL effect at a marker will be smaller than the true QTL effect as a result of recombination between the marker and the QTL. Second, we must discard individuals whose genotypes are missing at the marker. Third, when the markers are widely spaced, the QTL may be quite far from all markers, and so the power for QTL detection will decrease.

Interval mapping

Lander and Botstein developed interval mapping, which overcomes the three disadvantages of analysis of variance at marker loci. Interval mapping is currently the most popular approach for QTL mapping in experimental crosses. The method makes use of a genetic map of the typed markers, and, like analysis of variance, assumes the presence of a single QTL. Each location in the genome is posited, one at a time, as the location of the putative QTL....

Composite interval mapping (CIM)

In this method, one performs interval mapping using a subset of marker loci as covariates. These markers serve as proxies for other QTLs to increase the resolution of interval mapping, by accounting for linked QTLs and reducing the residual variation. The key problem with CIM concerns the choice of suitable marker loci to serve as covariates; once these have been chosen, CIM turns the model selection problem into a single-dimensional scan. The choice of marker covariates has not been solved, however. Not surprisingly, the appropriate markers are those closest to the true QTLs, and so if one could find these, the QTL mapping problem would be complete anyway.

Family-pedigree based mapping in plants

Plant geneticists are attempting to incorporate some of the methods pioneered in human genetics. Using family-pedigree based approach has been discussed (Bink et al. 2008). Family-based linkage and association has been successfully implemented (Rosyara et al. 2009)

Chapter 20

Zygosity

Zygosity refers to the similarity of genes for a trait (inherited characteristic) in an organism. If both genes are the same, the organism is homozygous for the trait. If both genes are different, the organism is heterozygous for that trait. If one gene is missing, it is hemizygous, and if both genes are missing, it is nullizygous.

Most eukaryotes have two matching sets of chromosomes, that is, they are diploid. Diploid organisms have the same genes on each of their two sets of chromosomes, except that the sequences of these genes may differ between the two chromosomes in a matching pair and that a few chromosomes may be mismatched as part of a sex-determination system.

The DNA sequence of a gene usually varies from one individual to another. Those variations are called alleles. Some genes have only one allele. Any variation from the DNA sequence of that allele will be fatal in the embryo, and the organism will never survive to be born. But most genes have two or more alleles. The frequency of different alleles varies throughout the population. Some genes may have two alleles with equal distribution. For other genes, one allele may be common, and another allele may be rare. Sometimes, one allele is a disease-causing variation while the other allele is healthy. Sometimes, the different variations in the alleles make no difference at all in the function of the organism.

In diploid organisms, one allele is inherited from the male parent and one from the female parent. Zygosity is a description of whether those two alleles have identical or different DNA sequences.

Types

The words homozygous, heterozygous, and hemizygous are used to describe the genotype of a diploid organism at a single locus on the DNA. *Homozygous* describes a genotype consisting of two identical alleles at a given locus, *heterozygous* describes a genotype consisting of two different alleles at a locus, *hemizygous* describes a genotype consisting of only a single copy of a particular gene in an otherwise diploid organism,

and *nullizygous* refers to an otherwise diploid organism in which both copies of the gene are missing.

Homozygous

A cell is said to be homozygous for a particular gene when identical alleles of the gene are present on both homologous chromosomes. The cell or organism in question is called a *homozygote*. True breeding organisms are always homozygous for the traits that are to be held constant.

An individual that is *homozygous dominant* for a particular trait carries two copies of the allele that codes for the dominant trait. This allele, often called the "dominant allele", is normally represented by a capital letter (such as "P" for the dominant allele producing purple flowers in pea plants). When an organism is homozygous dominant for a particular trait, the genotype is represented by a doubling of the symbol for that trait, such as "PP".

An individual that is *homozygous recessive* for a particular trait carries two copies of the allele that codes for the recessive trait. This allele, often called the "recessive allele", is usually represented by the lowercase form of the letter used for the corresponding dominant trait (such as, with reference to the example above, "p" for the recessive allele producing white flowers in pea plants). The genotype of an organism that is homozygous recessive for a particular trait is represented by a doubling of the appropriate letter, such as "pp".

Heterozygous

A diploid organism is heterozygous at a gene locus when its cells contain two different alleles of a gene. Heterozygous genotypes are represented by a capital letter (representing the dominant allele) and a lowercase letter (representing the recessive allele), such as "Rr" or "Ss". The capital letter is usually written first.

If the trait in question is determined by simple (complete) dominance, a heterozygote will express only the trait coded by the dominant allele and the trait coded by the recessive allele will not be present. In more complex dominance schemes the results of heterozygosity can be more complex.

Hemizygous

A chromosome in a diploid organism is hemizygous when only one copy is present. The cell or organism is called a *hemizygote*. Hemizygosity is observed when one copy of a gene is deleted, or in the heterogametic sex when a gene is located on a sex chromosome. For organisms in which the male is heterogametic, such as humans, almost all X-linked genes are hemizygous in males with normal chromosomes because they have only one X chromosome and few of the same genes are on the Y chromosome. In a more extreme example, male honeybees (known as drones) are completely hemizygous organisms. They develop from unfertilized eggs and their entire genome is haploid, unlike female

honeybees, which are diploid. Transgenic mice generated through exogenous DNA microinjection of an embryo's pronucleus are also hemizygous, and can later be bred to homozygosity to reduce the need to confirm genotype of each litter.

Nullizygous

A nullizygous organism carries two mutant alleles for the same gene. The mutant alleles are both complete loss-of-function or 'null' alleles, so homozygous null and nullizygous are synonymous. The mutant cell or organism is called a *nullizygote*.

Autozygous and allozygous

Zygosity may also refer to the origin(s) of the alleles in a genotype. When the two alleles at a locus originate from a common ancestor by way of nonrandom mating (inbreeding), the genotype is said to be *autozygous*. This is also known as being "identical by descent", or IBD. When the two alleles come (at least to the extent that the descent can be traced) from completely different sources, as is the case in most normal, random mating, the genotype is called *allozygous*. This is known as being "identical by state", or IBS.

Because the alleles of autozygous genotypes come from the same source, they are always homozygous, but allozygous genotypes may be homozygous too. All heterozygous genotypes are, by definition, allozygous because they contain two completely different alleles. Hemizygous and nullizygous genotypes do not contain enough alleles to allow for comparison of sources, so this classification is irrelevant for them.

Monozygotic and dizygotic twins

As discussed above, "zygosity" can be used in the context of a specific genetic locus (example). In addition, the word "zygosity" may also be used to describe the genetic similarity or dissimilarity of twins. Identical twins are **monozygotic**, meaning that they develop from one zygote that splits and forms two embryos. Fraternal twins are **dizygotic** because they develop from two separate eggs that are fertilized by two separate sperm.

In some cases the term "zygosity" is used in the context of a single chromosome.

Population genetics

In population genetics, the concept of heterozygosity is commonly extended to refer to the population as a whole, i.e., the fraction of individuals in a population that are heterozygous for a particular locus. It can also refer to the fraction of loci within an individual that are heterozygous.

Typically, the observed(H_o) and expected(H_e) heterozygosities are compared, defined as follows for diploid individuals in a population:

Observed

$$H_o = \frac{\sum_{i=1}^{n} (1 \text{ if } a_{i1} \neq a_{i2})}{n}$$

where n is the number of individuals in the population, and a_{i1},a_{i2} are the alleles of individual i at the target locus.

Expected

$$H_e = 1 - \sum_{i=1}^{m} (f_i)^2$$

where m is the number of alleles at the target locus, and f_i is the allele frequency of the i^{th} allele at the target locus.

Chapter 21

Microfluidic Whole Genome Haplotyping

Microfluidic whole genome haplotyping is a technique for the physical separation of individual chromosomes from a metaphase cell followed by direct resolution of the haplotype for each allele.

Background

Whole genome haplotyping

Whole genome haplotyping is the process of resolving personal haplotypes on a whole genome basis. Current methods of next generation sequencing are capable of identifying heterozygous loci, but they are not well suited to identify which polymorphisms exist on the same (in cis) or allelic (in trans) strand of DNA. Haplotype information contributes to the understanding of the potential functional effects of variants in cis or in trans. Haplotypes are more frequently resolved by inference through comparison with parental genotypes, or from population samples using statistical computational methods to determine linkage disequilibrium between markers. Direct haplotyping is possible through isolation of chromosomes or chromosome segments. Most molecular biology techniques for haplotyping can accurately determine haplotypes of only a limited region of the genome. Whole genome direct haplotyping involves the resolution of haplotype at the whole genome level, usually through the isolation of individual chromosomes.

Haplotype

A haplotype (haplo: from Ancient Greek ἁπλόος (haplóos, "single, simple") is a contiguous section of closely linked segments of DNA within the larger genome that tend to be inherited together as a unit on a single chromosome. Haplotypes have no defined size and can refer to anything from a few closely linked loci up to an entire chromosome. The term is also used to describe groups of single-nucleotide polymorphisms (SNPs) that are statistically associated. Most of the knowledge of SNP association comes from the effort of the International HapMap Project, which has proved itself a powerful resource in the development of a publicly accessible database of human genetic variation.

Phasing

Phasing is the process of identifying the individual complement of homologous chromosomes. Methods for phasing include pedigree analysis, allele-specific PCR, linkage emulsion PCR haplotype analysis, polony PCR, sperm typing, bacterial artificial chromosome cloning, construction of somatic cell hybrids, atomic force microscopy, among others. Haplotype phasing can also be achieved through computational inference methods.

Microfluidics

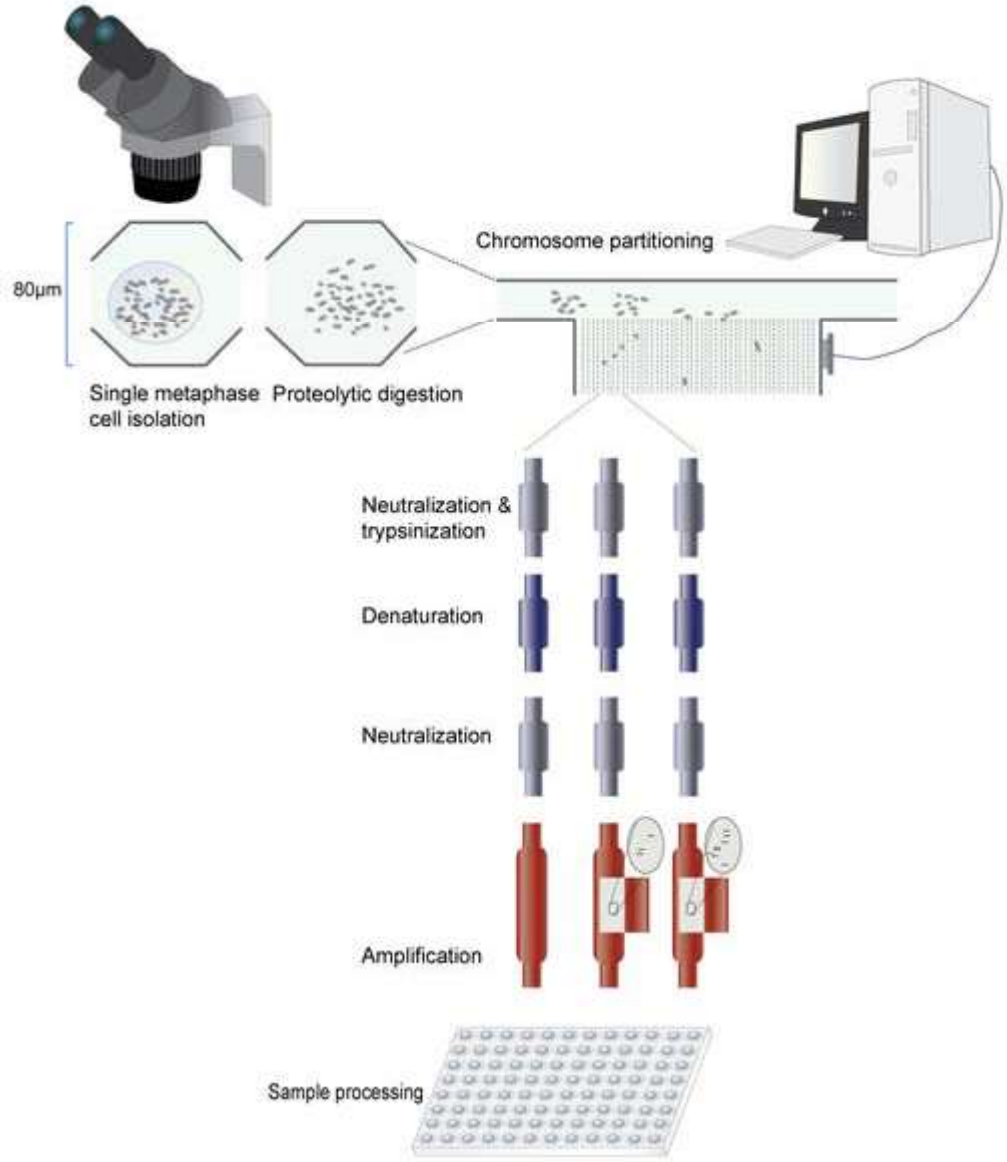
Microfluidics refers to the use of micro-sized channels on a micro-electro-mechanical system (MEMS). Microfluidic channels have a diameter of 10-100μm, making it possible to manipulate and analyze minute volumes. This technology combines engineering, physics, chemistry, biology, and optics. Over the past decades it has revolutionized micro and nanoscale biology, genetics and proteomics. Microfluidic devices can combine several analytical steps into one device. This technology has been coined by some as the "lab on a chip" technology. Most current molecular biology methods use some form of MEMS, including microarray technology and next generation sequencing instruments.

Microfluidic direct deterministic phasing

Principle

Direct deterministic phasing of individual chromosomes can be achieved by isolating single chromosomes for genetic analysis through the use of a microfluidic device.

Methods



Workflow of microfluidic whole genome chromosome isolation and amplification. Not at scale

A single metaphase cell is isolated from solution. The chromosomes are then released from the nucleus, and the cytoplasm is digested enzymatically. Next, the chromosome suspension is directed towards multiple partitioning channels. The chromosomes are physically directed into the partitioning channels using a series of valves. In the first description of this technique, Fan et al. designed a custom-made program (MatLab) to control this process. Once separated, the chromosomes are prepared for amplification by sequential addition and washout of trypsin, denaturation buffer and neutralization solution. The DNA is then ready for further processing. Because of the small amount of DNA, amplification needs to be performed using kits specialized for very small initial

DNA quantities. The amplified DNA is flushed out of the microfluidic device and solubilized by the addition of a buffer. The amplified DNA can now be analyzed by various methods.

Once the chromosomes have been isolated and amplified any molecular haplotyping can be applied as long as the chromosomes remain distinct. This could be accomplished by keeping them physically separated, or identifying each sample by genotyping. Once each chromosome has been identified each pair of homologs can be assorted into one of two haploid genomes.

Applications

Microfluidic direct deterministic phasing allows all the chromosomes to be isolated in the same experiment. This unique feature suggests possible applications within clinical, research and personal genomics realms. Some of the possible clinical applications for this technique include phasing of multiple mutations when parental samples are unavailable, preimplantation genetic diagnosis, prenatal diagnosis and in the characterization of cancer cells.

Whole genome haplotyping through microfluidics will increase the rate of discovery within the HapMap project, and provides an opportunity for corroboration and error detection within the existing database. It will further inform genetic association studies.

As methods for amplification of small amounts of DNA improve, single chromosome sequencing is possible using microfluidics to separate each individual chromosome. A cost-effective approach may be to barcode each individual chromosome and perform parallel resequencing of the entire individual genome. The amplification of each chromosome separately also provides a mechanism to potentially fill in some of the gaps that remain in the human reference genome. Single chromosome sequencing will allow for unmapped sequences to be associated with a single chromosome. Additionally, single chromosome sequencing will be more accurate in the identification of copy number variants and repetitive sequences.

Limitations

As of January 2011, only one publication has described use of this technique. The scientific commons awaits further validation of this method and its efficacy in isolating and amplifying analyzable amounts of DNA. While this method does streamline the process of chromosome isolation, certain parts in the process – such as the initial isolation of a metaphase cell – remain difficult and labour intensive. Other automated techniques for metaphase cell separation would improve throughput. In addition, this method is only applicable to cells in metaphase, which inherently limits the technique to cell types and tissues that undergo mitosis. Single cell analysis does not account for the possibility of mosaicism; therefore, applications in cancer diagnosis and research would necessarily require processing of multiple cells. Finally, since this entire process is based on amplification from a single cell, the accuracy of any genetic analysis is limited to the

ability of commercially available platforms to produce sufficient amounts of unbiased and error free amplicon.

Alternative methods of whole genome haplotyping

Chromosome microdissection

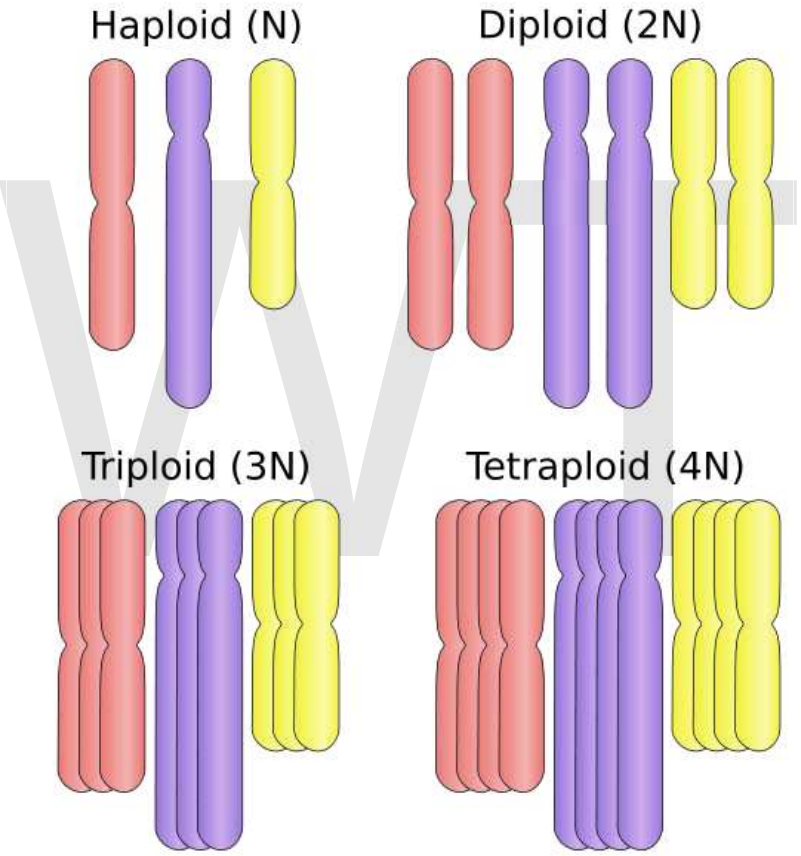
Chromosome microdissection is another process for isolating single chromosomes for genetic analysis. As with the above technique microdissection begins with metaphase cells. The nucleus is lysed mechanically on a glass slide and part of the genetic material is partitioned under microscope. The actual microdissection of genetic material was initially accomplished through the careful use of a fine needle. Today computer-directed lasers are available. The genomic area isolated can range from part of a single chromosome, up to several chromosomes. To accomplish whole genome haplotyping the microdissected genomic section is amplified and genotyped or sequenced. Like with the microfluidic technique, specialized amplification platforms are necessary to address the problem of a small initial DNA sample.

Large insert cloning

Randomly partitioning a complete diploid fosmid library into various pools of equal size presents an alternative method for haplotype phasing. In the proof of principle description of this technique 115 pools were created containing ~5000 unique clones from the original fosmid library. Each of these pools contained roughly 3% of the genome. Between the 3% in each pool and the fact that each clone is a random sampling of the diploid genome, 99.1% of the time each pool contains DNA from a single homolog. Amplification and analysis of each pool provide haplotype resolution limited only by the size of the fosmid insert.

Chapter 22

Polyploid



This image shows haploid (single), diploid (double), triploid (triple), and tetraploid (quadruple) sets of chromosomes. Triploid and tetraploid chromosomes are examples of polyploidy.

Polyploid is a term used to describe cells and organisms containing more than two paired (homologous) sets of chromosomes. Most species are diploid, meaning they have two sets of chromosomes — one set inherited from each parent. However **polyploidy** is found in some organisms and is especially common in plants. In addition, polyploidy also occurs in some tissues of animals who are otherwise diploid, such as human muscle

tissues. This is known as **endopolyploidy**. (Monoploid organisms also occur; a monoploid has only one set of chromosomes.)

Polyploidy refers to a numerical change in a whole set of chromosomes. Organisms in which a particular chromosome, or chromosome segment, is under- or overrepresented are said to be **aneuploid** (from the Greek words meaning "not," "good," and "fold"). Therefore the distinction between aneuploidy and polyploidy is that aneuploidy refers to a numerical change in part of the chromosome set, whereas polyploidy refers to a numerical change in the whole set of chromosomes.

Polyploidy may occur due to abnormal cell division, either during mitosis, or commonly during metaphase I in meiosis.

Polyploidy occurs in some animals, such as goldfish, salmon, and salamanders, but is especially common among ferns and flowering plants, including both wild and cultivated species. Wheat, for example, after millennia of hybridization and modification by humans, has strains that are **diploid** (two sets of chromosomes), **tetraploid** (four sets of chromosomes) with the common name of durum or macaroni wheat, and **hexaploid** (six sets of chromosomes) with the common name of bread wheat. Many agriculturally important plants of the genus *Brassica* are also tetraploids. Polyploidization is a mechanism of sympatric speciation because polyploids are usually unable to interbreed with their diploid ancestors.

Polyploidy can be induced in plants and cell cultures by some chemicals: the best known is colchicine, which can result in chromosome doubling, though its use may have other less obvious consequences as well. Oryzalin also will double the existing chromosome content.

Polyploid types

Polyploid types are labeled according to the number of chromosome sets in the nucleus:

- **triploid** (three sets; 3x), for example seedless watermelons, common in the phylum Tardigrada
- **tetraploid** (four sets; 4x), for example Salmonidae fish
- **pentaploid** (five sets; 5x), for example Kenai Birch (*Betula papyrifera* var. *kenaica*)
- **hexaploid** (six sets; 6x), for example wheat, kiwifruit
- **octaploid** (eight sets; 8x), for example *Acipenser* (genus of sturgeon fish)
- **decaploid** (ten sets; 10x), for example certain strawberries
- **dodecaploid** (twelve sets; 12x), for example the plant *Celosia argentea* and the amphibian *Xenopus ruwenzoriensis*

Polyploidy in animals (non-human)

Examples in animals are more common in the 'lower' forms such as flatworms, leeches, and brine shrimp. Polyploid animals are often sterile, so they often reproduce by parthenogenesis. Polyploid lizards are also quite common and parthenogenetic. Polyploid mole salamanders (mostly triploids) are all female and reproduce by kleptogenesis, "stealing" spermatophores from diploid males of related species to trigger egg development but not incorporating the males' DNA into the offspring. While mammalian liver cells are polyploid, rare instances of polyploid mammals are known, but most often result in prenatal death.

One of the few known exceptions to this 'rule' is an octodontid rodent of Argentina's harsh desert regions, known as the Plains Viscacha-Rat (*Tympanoctomys barrerae*). This rodent is not a rat, but kin to guinea pigs and chinchillas. Its "new" diploid [2n] number is 102 and so its cells are roughly twice normal size. Its closest living relation is *Octomys mimax*, the Andean Viscacha-Rat of the same family, whose 2n = 56. It is surmised that an *Octomys*-like ancestor produced tetraploid (i.e., 4n = 112) offspring that were, by virtue of their doubled chromosomes, reproductively isolated from their parents; but that these likely survived the ordinarily catastrophic effects of polyploidy in mammals by shedding (via translocation or some similar mechanism) the "extra" set of sex chromosomes gained at this doubling. (The closely related Golden Vizcacha Rat, 2n = 96, is thought to have arisen via roughly the same process).

Polyploidy in humans

True polyploidy rarely occurs in humans, although it occurs in some tissues (especially in the liver). Aneuploidy is more common.

Polyploidy occurs in humans in the form of triploidy, with 69 chromosomes (sometimes called 69,XXX), and tetraploidy with 92 chromosomes (sometimes called 92,XXXX). Triploidy, usually due to polyspermy, occurs in about 2–3% of all human pregnancies and ~15% of miscarriages. The vast majority of triploid conceptions end as miscarriage and those that do survive to term typically die shortly after birth. In some cases survival past birth may occur longer if there is mixoploidy with both a diploid and a triploid cell population present.

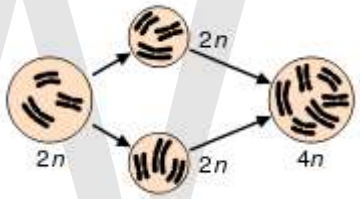
Triploidy may be the result of either digyny (the extra haploid set is from the mother) or diandry (the extra haploid set is from the father). Diandry is mostly caused by reduplication of the paternal haploid set from a single sperm, but may also be the consequence of dispermic (two sperm) fertilization of the egg. Digyny is most commonly caused by either failure of one meiotic division during oogenesis leading to a diploid oocyte or failure to extrude one polar body from the oocyte. Diandry appears to predominate among early miscarriages while digyny predominates among triploidy that survives into the fetal period. However, among early miscarriages, digyny is also more common in those cases <8.5 weeks gestational age or those in which an embryo is present. There are also two distinct phenotypes in triploid placentas and fetuses that are

dependent on the origin of the extra haploid set. In digyny there is typically an asymmetric poorly grown fetus, with marked adrenal hypoplasia and a very small placenta. In diandry, a partial hydatidiform mole develops. These parent-of-origin effects reflect the effects of genomic imprinting.

Complete tetraploidy is more rarely diagnosed than triploidy, but is observed in 1–2% of early miscarriages. However, some tetraploid cells are commonly found in chromosome analysis at prenatal diagnosis and these are generally considered 'harmless'. It is not clear whether these tetraploid cells simply tend to arise during *in vitro* cell culture or whether they are also present in placental cells *in vivo*. There are, at any rate, very few clinical reports of fetuses/infants diagnosed with tetraploidy mosaicism.

Mixoploidy is quite commonly observed in human preimplantation embryos and includes haploid/diploid as well as diploid/tetraploid mixed cell populations. It is unknown whether these embryos fail to implant and are therefore rarely detected in ongoing pregnancies or if there is simply a selective process favoring the diploid cells.

Polyploidy in plants



Speciation via polyploidy: A diploid cell undergoes failed meiosis, producing diploid gametes, which self-fertilize to produce a tetraploid zygote.

Polyploidy is pervasive in plants and some estimates suggest that 30–80% of living plant species are polyploid, and many lineages show evidence of ancient polyploidy (paleopolyploidy) in their genomes. Huge explosions in angiosperm species diversity appear to have coincided with the timing of ancient genome duplications shared by many species. It has been established that 15% of angiosperm and 31% of fern speciation events are accompanied by ploidy increase. Polyploid plants can arise spontaneously in nature by several mechanisms, including meiotic or mitotic failures, and fusion of unreduced (2n) gametes. Both autopolyploids (e.g. potato) and allopolyploids (e.g. canola, wheat, cotton) can be found among both wild and domesticated plant species. Most polyploids display heterosis relative to their parental species, and may display novel variation or morphologies that may contribute to the processes of speciation and eco-niche exploitation. The mechanisms leading to novel variation in newly formed allopolyploids may include gene dosage effects (resulting from more numerous copies of genome content), the reunion of divergent gene regulatory hierarchies, chromosomal rearrangements, and epigenetic remodeling, all of which affect gene content and/or expression levels. Many of these rapid changes may contribute to reproductive isolation and speciation.

Lomatia tasmanica is an extremely rare Tasmanian shrub which is triploid and sterile, and reproduction is entirely vegetative with all plants having the same genetic structure.

There are few naturally occurring polyploid conifers. One example is the giant tree *Sequoia sempervirens* or Coast Redwood which is a hexaploid (6x) with 66 chromosomes (2n = 6x = 66), although the origin is unclear.

Polyploid crops

Polyploid plants tend to be larger and better at flourishing in early succession habitats such as farm fields. In the breeding of crops, the tallest and best thriving plants are selected for. Thus, many crops (and agricultural weeds) may have unintentionally been bred to a higher level of ploidy.

The induction of polyploidy is a common technique to overcome the sterility of a hybrid species during plant breeding. For example, Triticale is the hybrid of wheat (*Triticum turgidum*) and rye (*Secale cereale*). It combines sought-after characteristics of the parents, but the initial hybrids are sterile. After polyploidization, the hybrid becomes fertile and can thus be further propagated to become triticale.

In some situations polyploid crops are preferred because they are sterile. For example many seedless fruit varieties are seedless as a result of polyploidy. Such crops are propagated using asexual techniques such as grafting.

Polyploidy in crop plants is most commonly induced by treating seeds with the chemical colchicine.

Examples of polyploid crops

- Triploid crops: apple, banana, citrus, ginger, watermelon
- Tetraploid crops: apple, durum or macaroni wheat, cotton, potato, cabbage, leek, tobacco, peanut, kinnow, Pelargonium
- Hexaploid crops: chrysanthemum, bread wheat, triticale, oat, kiwifruit
- Octaploid crops: strawberry, dahlia, pansies, sugar cane

Some crops are found in a variety of ploidies: tulips and lilies are commonly found as both diploid and as triploid; daylilies (Hemerocallis cultivars) are available as either diploid or tetraploid; apples and kinnows can be diploid, triploid, or tetraploid.

Terminology

Autopolyploidy

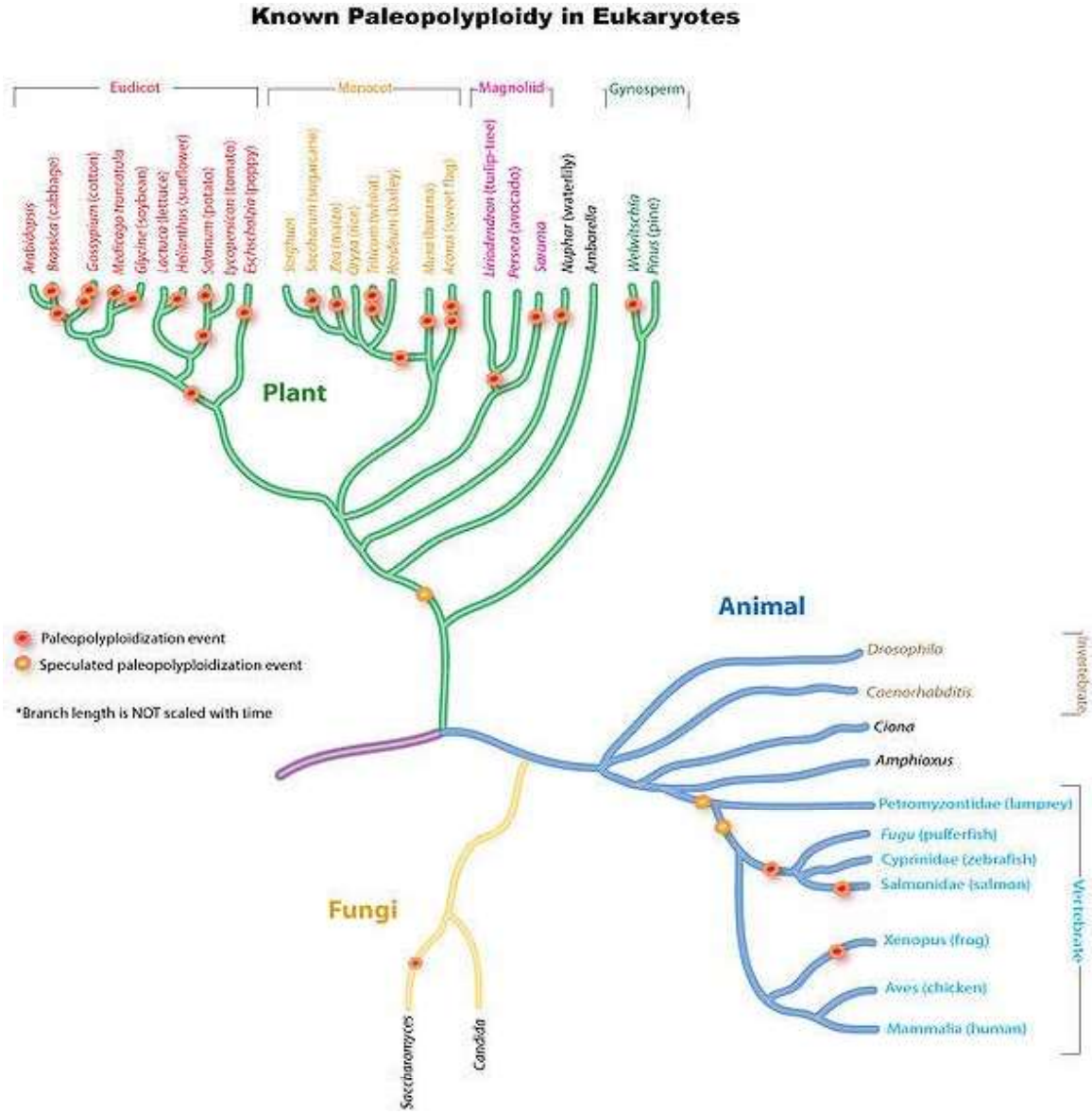
Autopolyploids are polyploids with multiple chromosome sets derived from a single species. Autopolyploids can arise from a spontaneous, naturally occurring genome doubling, like the potato. Others might form following fusion of 2n gametes (unreduced

gametes). Bananas and apples can be found as autotriploids. Autopolyploid plants typically display polysomic inheritance, and are therefore often infertile and propagated clonally perfect.

Allopolyploidy

Allopolyploids are polyploids with chromosomes derived from different species. Precisely it is the result of doubling of chromosome number in an F1 hybrid. *Triticale* is an example of an allopolyploid, having six chromosome sets, allohexaploid, four from wheat (*Triticum turgidum*) and two from rye (*Secale cereale*). *Amphidiploid* is another word for an allopolyploid. Some of the best examples of allopolyploids come from the Brassicas, and the Triangle of U describes the relationships among the three common diploid Brassicas (*B. oleracea, B. rapa,* and *B. nigra*) and three allotetraploids (*B. napus, B. juncea,* and *B. carinata*) derived from hybridization among the diploids.

Paleopolyploidy



This phylogenetic tree shows the relationship between the best-documented instances of paleopolyploidy in eukaryotes.

Ancient genome duplications probably occurred in the evolutionary history of all life. Duplication events that occurred long ago in the history of various evolutionary lineages can be difficult to detect because of subsequent diploidization (such that a polyploid starts to behave cytogenetically as a diploid over time) as mutations and gene translations gradually make one copy of each chromosome unlike its other copy.

In many cases, these events can be inferred only through comparing sequenced genomes. Examples of unexpected but recently confirmed ancient genome duplications include baker's yeast (*Saccharomyces cerevisiae*), mustard weed/thale cress (*Arabidopsis*

thaliana), rice (*Oryza sativa*), and an early evolutionary ancestor of the vertebrates (which includes the human lineage) and another near the origin of the teleost fishes. Angiosperms (flowering plants) have paleopolyploidy in their ancestry. All eukaryotes probably have experienced a polyploidy event at some point in their evolutionary history.

Karyotype

A karyotype is the characteristic chromosome complement of a eukaryote species. The preparation and study of karyotypes is part of cytology and, more specifically, cytogenetics.

Although the replication and transcription of DNA is highly standardized in eukaryotes, the same cannot be said for their karotypes, which are highly variable between species in chromosome number and in detailed organization despite being constructed out of the same macromolecules. In some cases there is even significant variation within species. This variation provides the basis for a range of studies in what might be called evolutionary cytology.

Paralogous

The term is used to describe the relationship among duplicated genes or portions of chromosomes that derived from a common ancestral DNA. Paralogous segments of DNA may arise spontaneously by errors during DNA replication, copy and paste transposons, or whole genome duplications.

Homologous

The term is used to describe the relationship of similar chromosomes that pair at mitosis and meiosis. In a diploid, one homolog is derived from the male parent (sperm) and one is derived from the female parent (egg). During meiosis and gametogenesis, homologous chromosomes pair and exchange genetic material by recombination, leading to the production of sperm or eggs with chromosome haplotypes containing novel genetic variation.

Homoeologous

The term *homoeologous*, also spelled *homeologous*, is used to describe the relationship of similar chromosomes or parts of chromosomes brought together following inter-species hybridization and allopolyploidization, and whose relationship was completely homologous in an ancestral species. In allopolyploids, the homologous chromosomes within each parental sub-genome should pair faithfully during meiosis, leading to disomic inheritance; however in some allopolyploids, the homoeologous chromosomes of the parental genomes may be nearly as similar to one another as the homologous chromosomes, leading to tetrasomic inheritance (four chromosomes pairing at meiosis), intergenomic recombination, and reduced fertility.

Example of homoeologous chromosomes

Durum wheat is the result of the inter-species hybridization of two diploid grass species *Triticum urartu* and *Aegilops speltoides*. Both the diploid ancestors had two sets of 7 chromosomes, which were similar in terms of size and genes contained on them. Durum wheat contains two sets of chromosomes derived from *Triticum urartu* and two sets of chromosomes derived from *Aegilops speltoides*. Each chromosome pair derived from the *Triticum urartu* parent is **homoeologous** to the opposite chromosome pair derived from the *Aegilops speltoides* parent, though each chromosome pair unto itself is **homologous**.