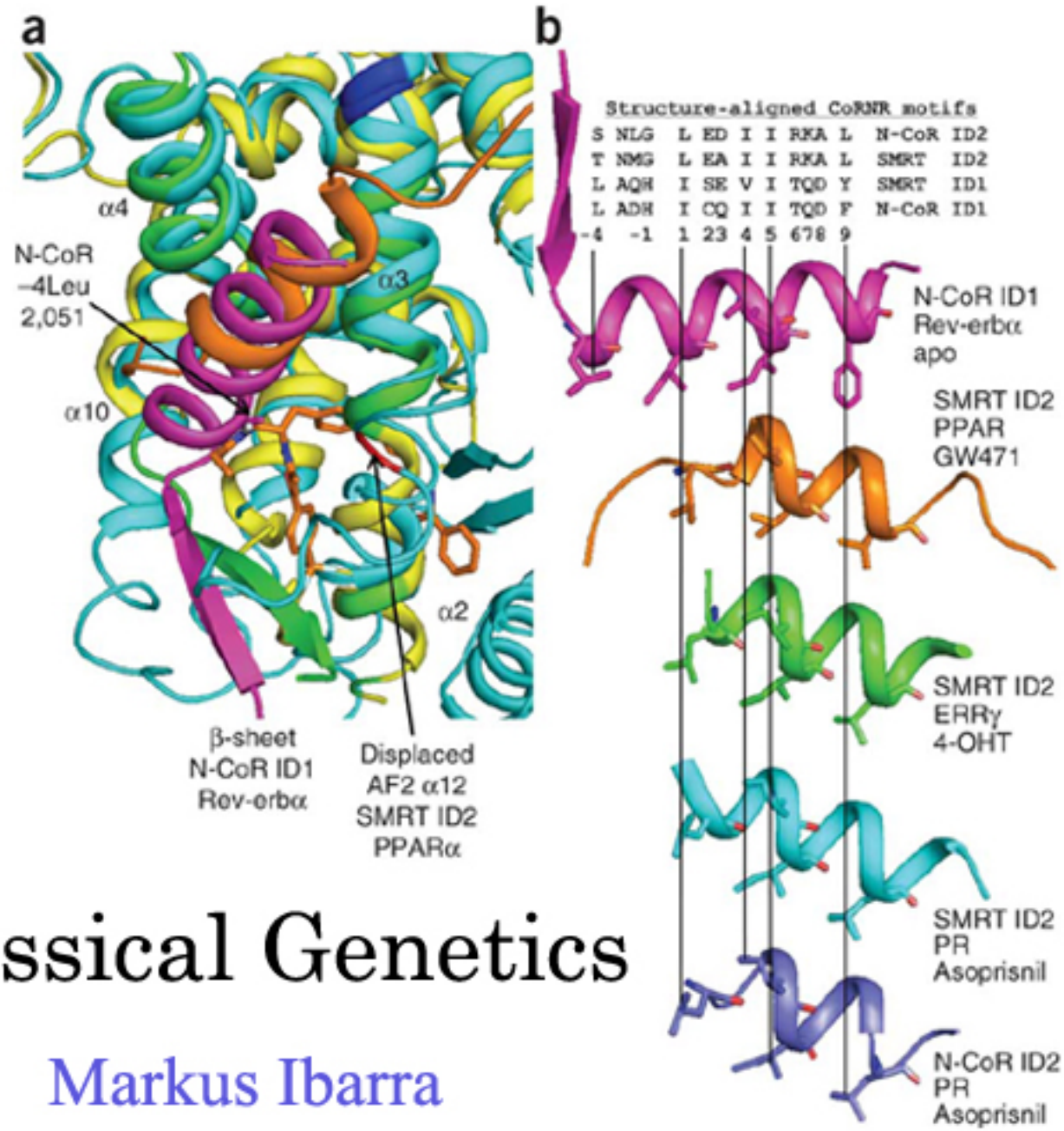



# Classical Genetics

Markus Ibarra

First Edition, 2012

ISBN 978-81-323-4105-5



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

# Table of Contents

# Chapter 1

# 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:

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

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 2

# 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 3

# 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*

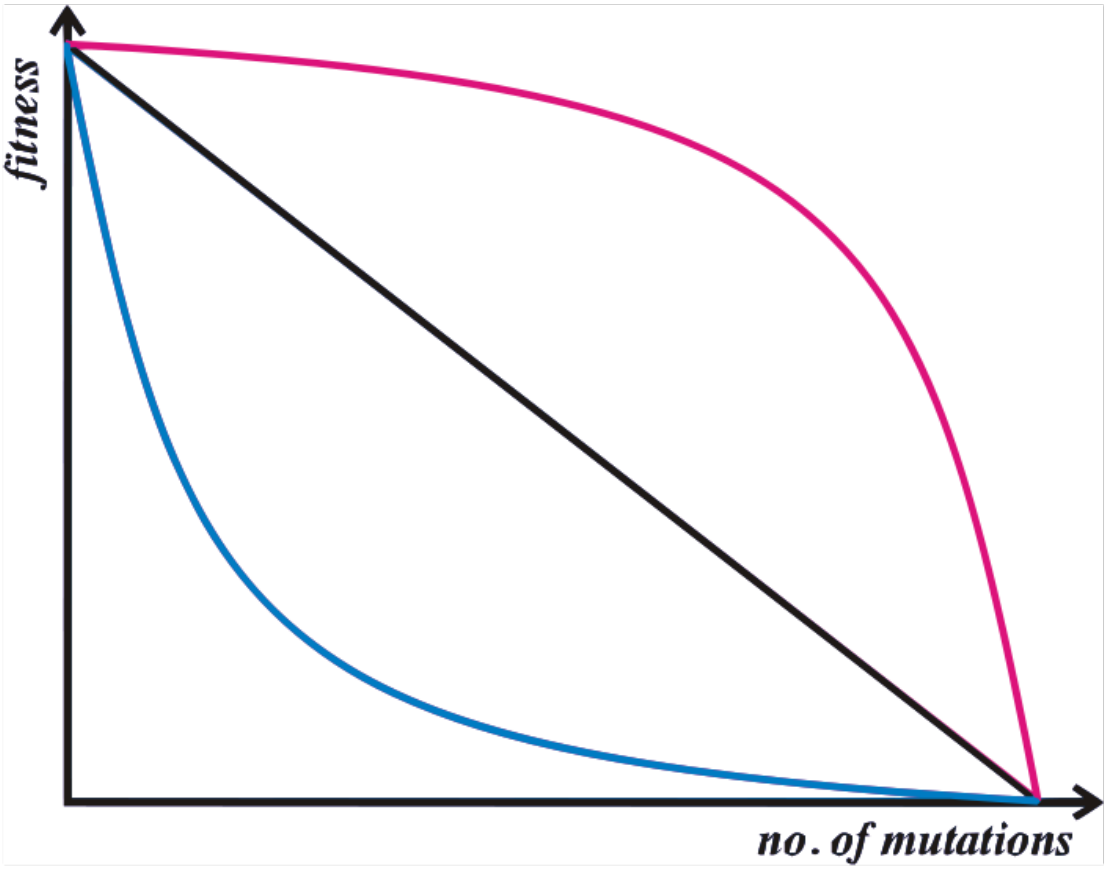



Diagram illustrating different relationships between numbers of mutations and fitness. *Synergistic* epistasis is the blue line - each mutation has a disproportionately large effect on the organism's fitness. *Antagonistic* epistasis is the red line.

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 4

# 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 | TG TG |
| **GC** | AG AC | AG TC or 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 5

# 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.

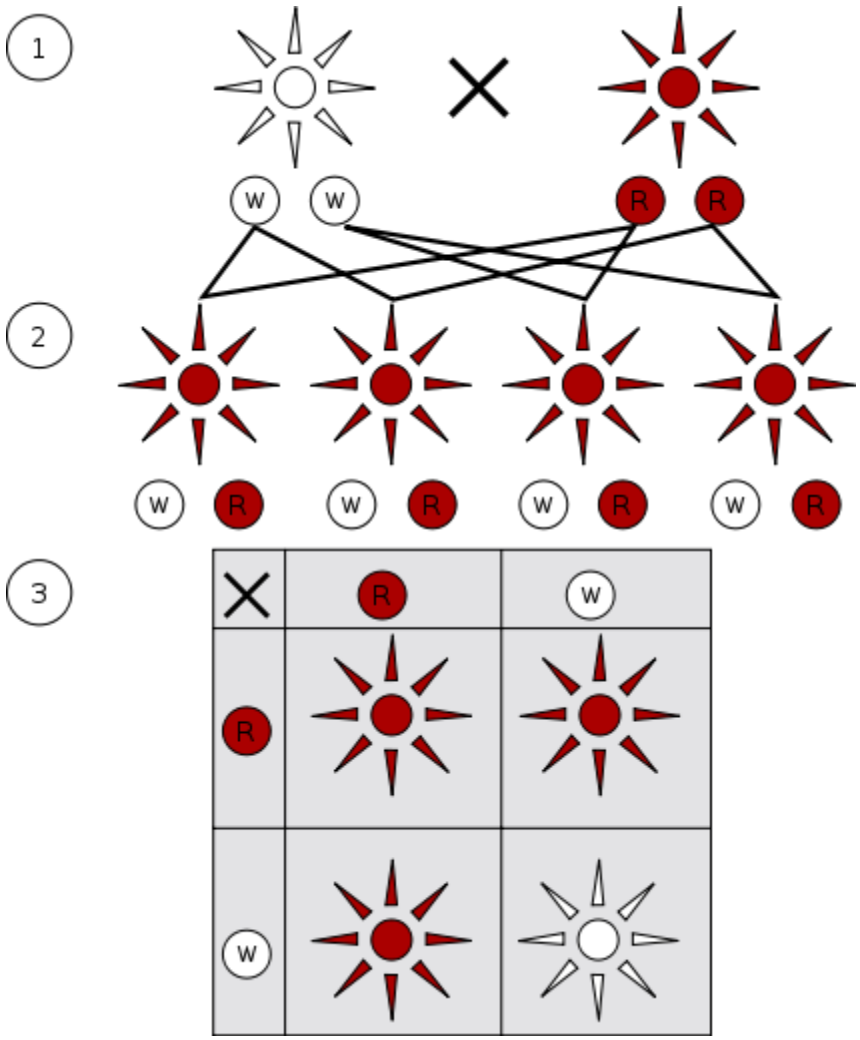



**Figure 1 :** Inheritance pattern of dominant (red) and recessive (white) phenotypes when each parent (1) is homozygous for either the dominant or recessive trait. All members of the $F_1$ generation are heterozygous and share the same dominant phenotype (2), while the $F_2$ generation exhibits a 3:1 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 6

# 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 7

# 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.

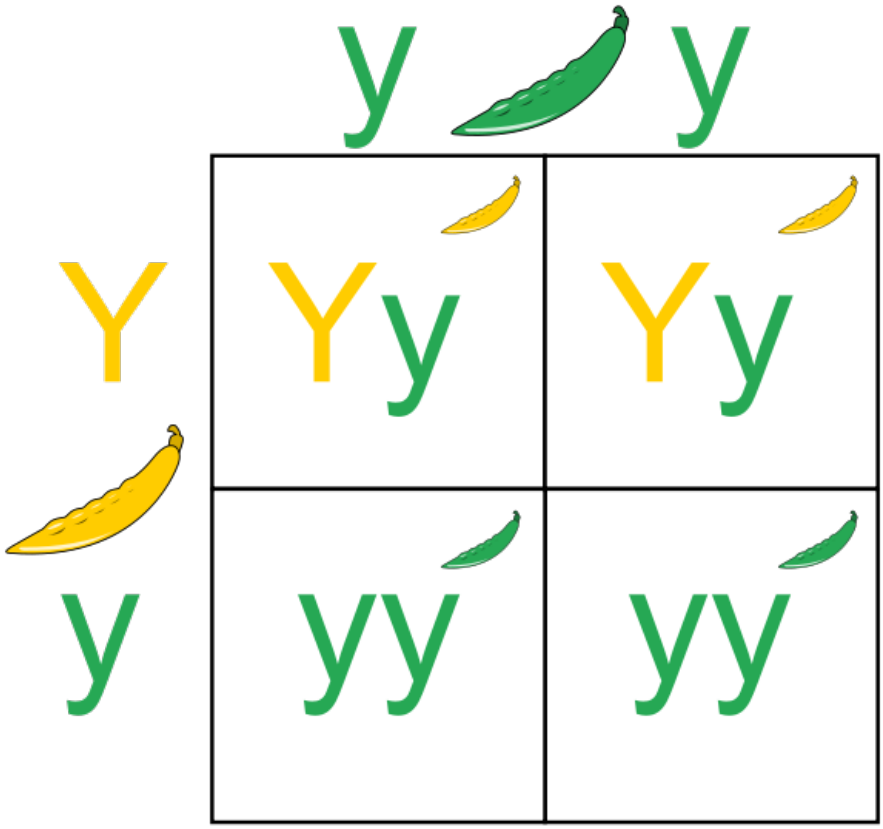
Some examples of Inherited genes include eye color.

## *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.

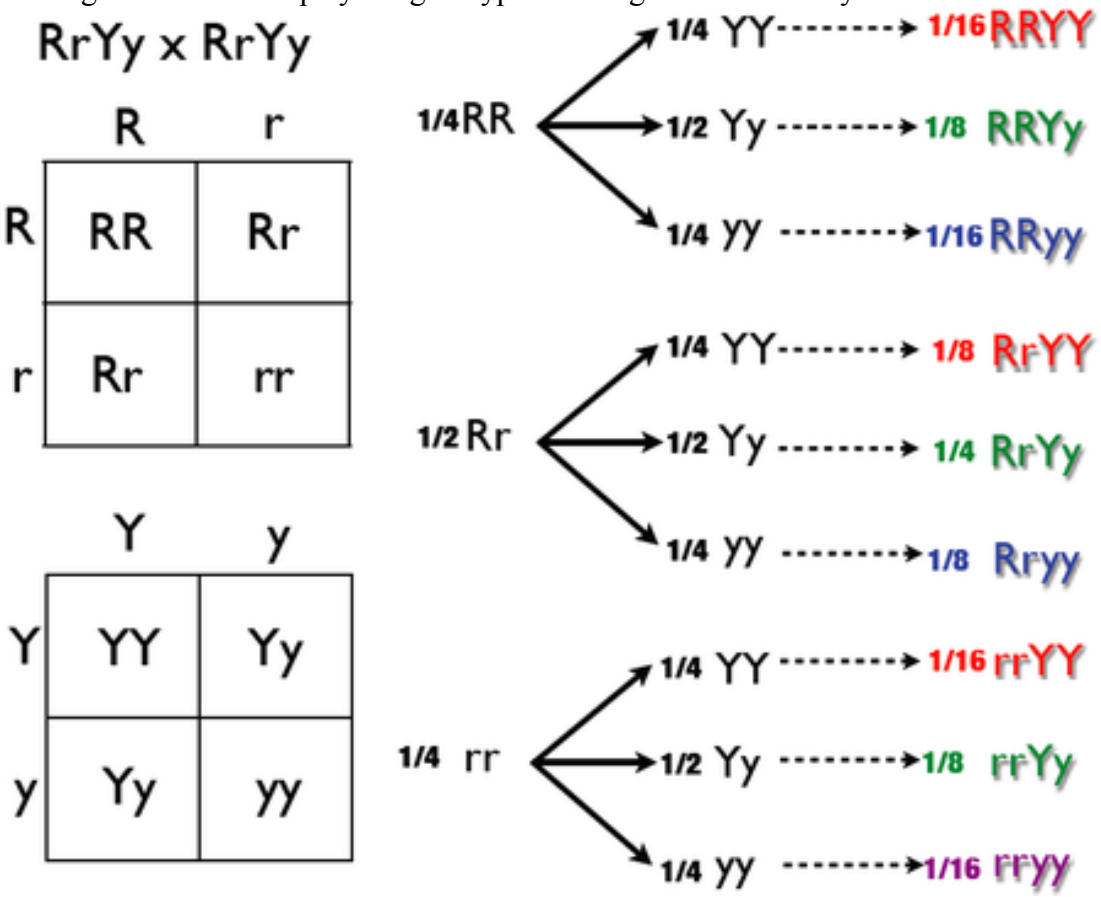
The following example illustrates a dihybrid cross between two heterozygous pea plants. *R* represents the dominant allele for shape (round), while *r* represents the recessive allele (wrinkled). *Y* represents the dominant allele for color (yellow), while *y* represents the recessive allele (green). 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 |

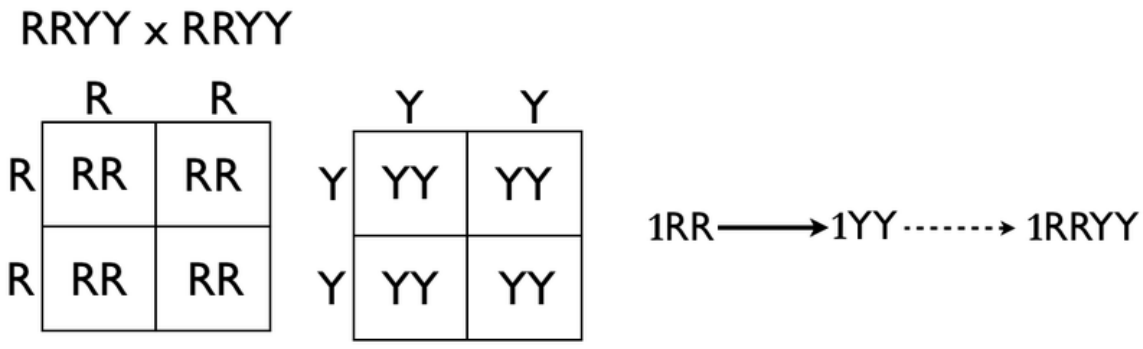
Since dominant traits mask recessive traits, there are nine combinations that have the phenotype round yellow, three that are round green, three that are wrinkled yellow and one that is wrinkled green. The ratio 9:3:3:1 is typical for a dihybrid cross.

## 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 8

# 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?

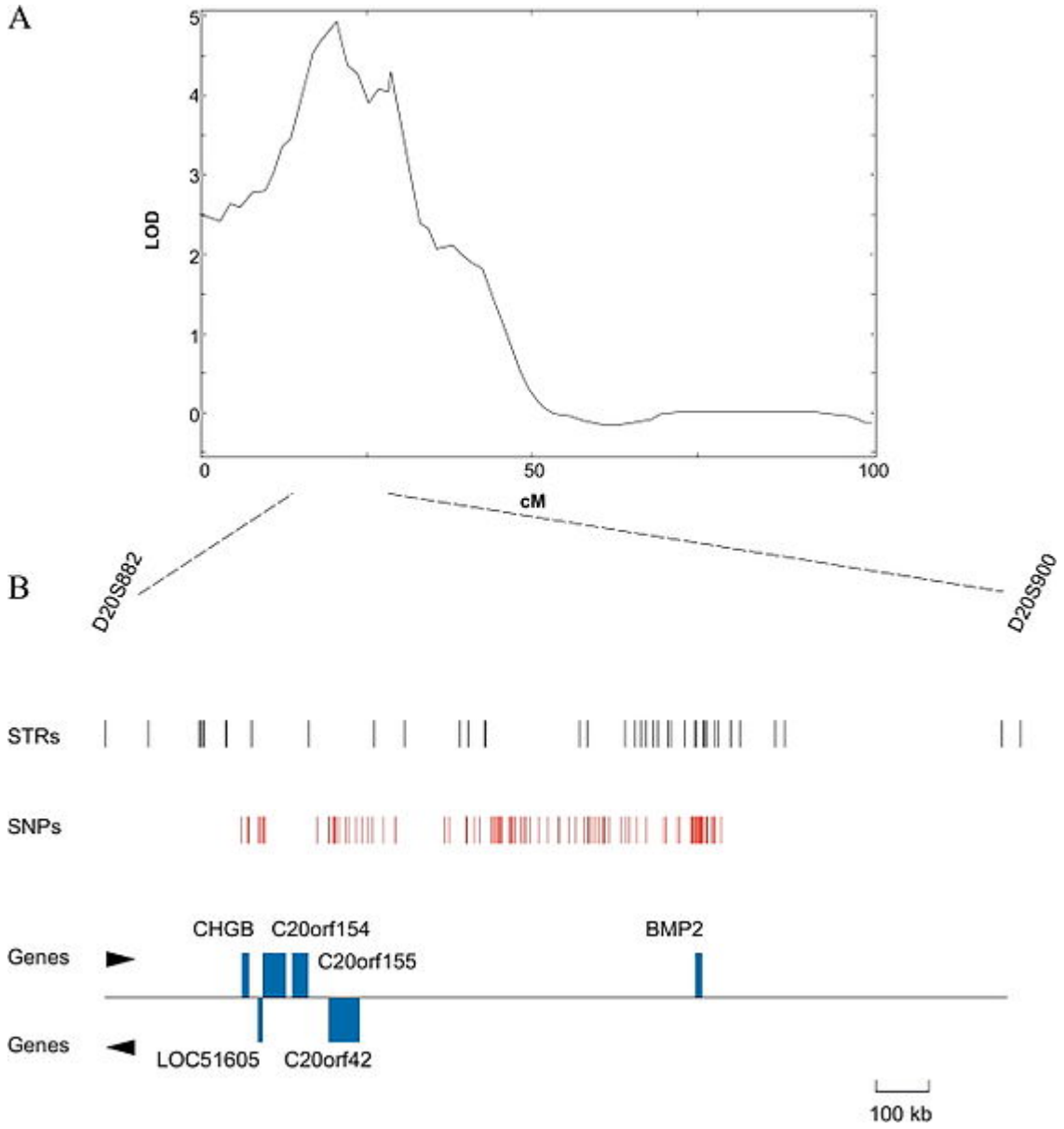
It is difficult to ascertain if any particular disease is multifactorially genetic. If a pedigree chart is taken of the patient's family and relations, and it is shown that the brothers and sisters of the patient have the disease, then there is a strong chance that the disease is genetic and that the 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).

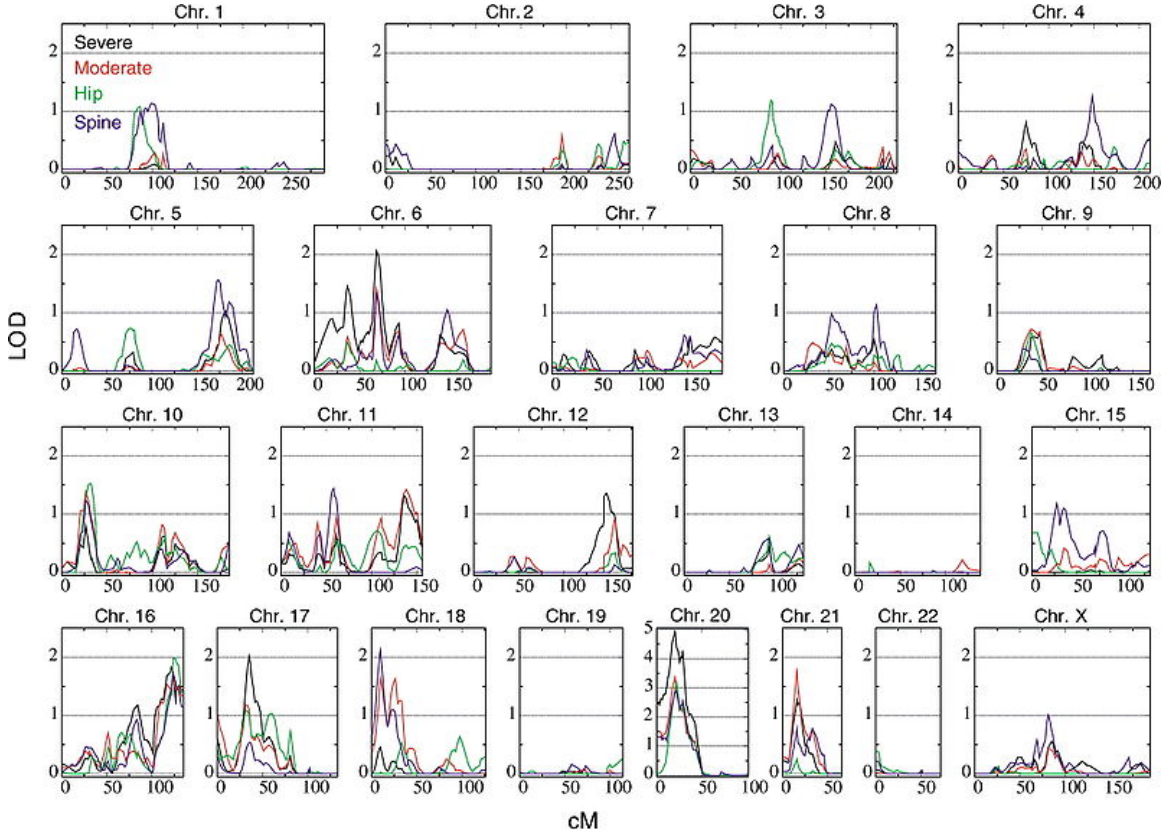
Moreover, a single phenotypic trait is usually determined by many genes. Consequently, many QTLs are associated with a single trait.

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 9

# 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 10

# 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

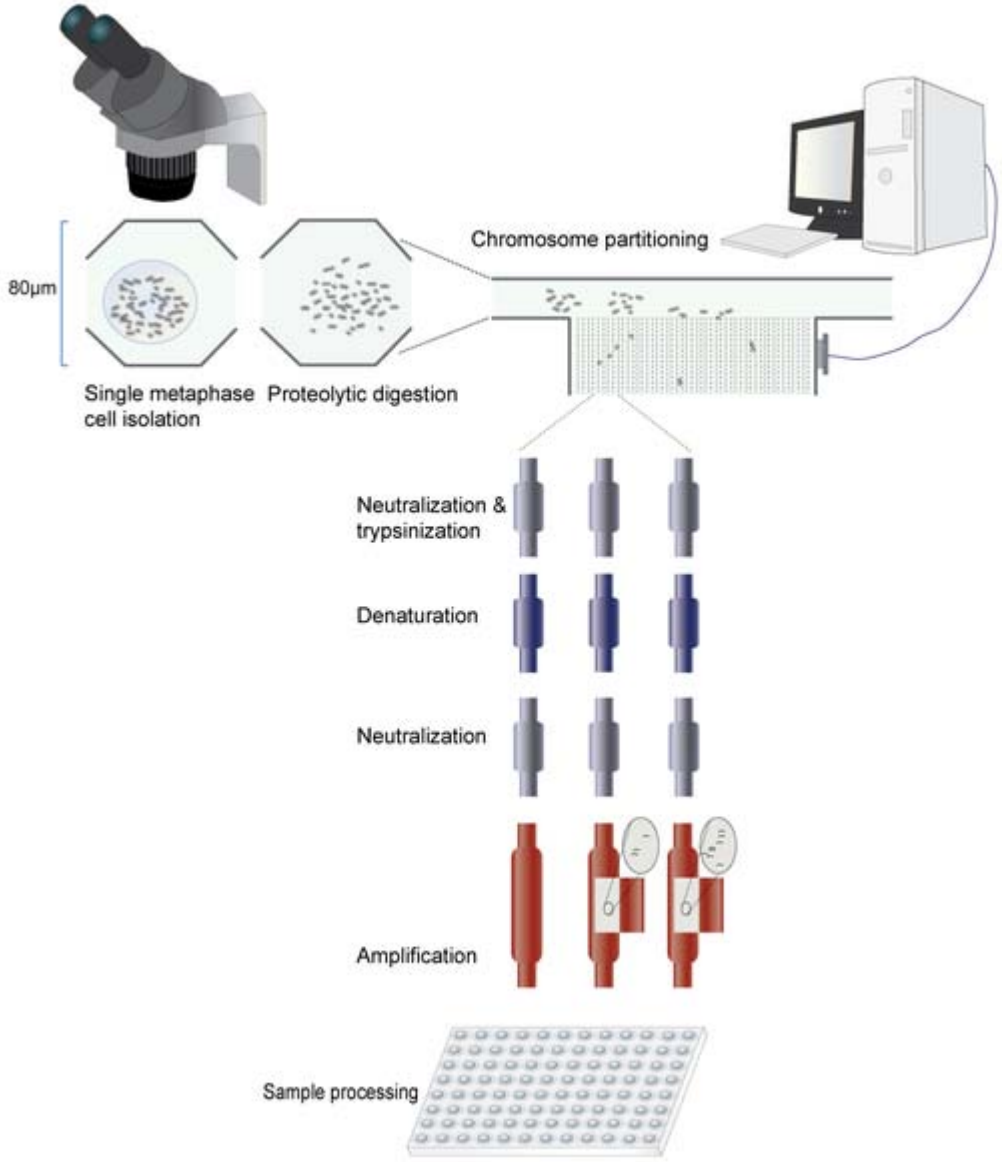
Microfluidics refers to the use of micro-sized channels on a micro-electro-mechanical system (MEMS). Microfluidic channels have a diameter of 10-100μm, making it possible to manipulate and analyze minute volumes. This technology combines engineering, physics, chemistry, biology, and optics. Over the 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

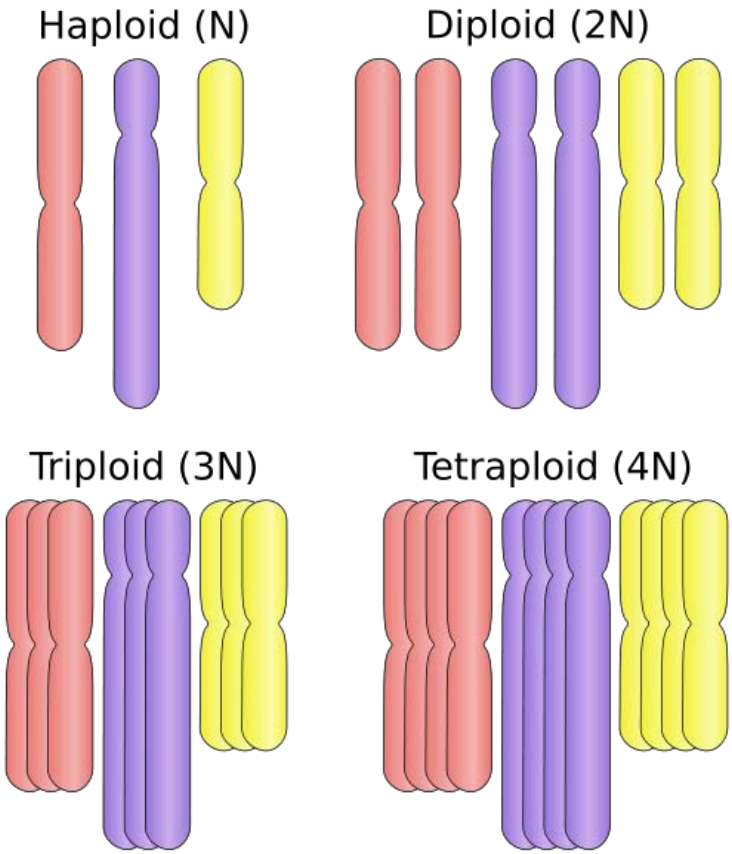
Chromosome microdissection is another process for isolating single chromosomes for genetic analysis. As with the above technique microdissection begins with metaphase cells. The nucleus is lysed mechanically on a glass slide and part of the genetic material is partitioned under microscope. The actual 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 11

# 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.

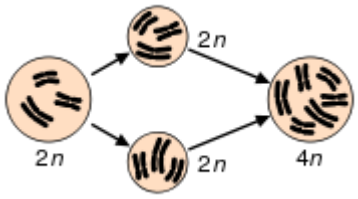
Triploidy may be the result of either digyny (the extra haploid set is from the mother) or diandry (the extra haploid set is from the father). Diandry is mostly caused by reduplication of the paternal haploid set from a single sperm, but may also be the consequence of dispermic (two sperm) fertilization 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

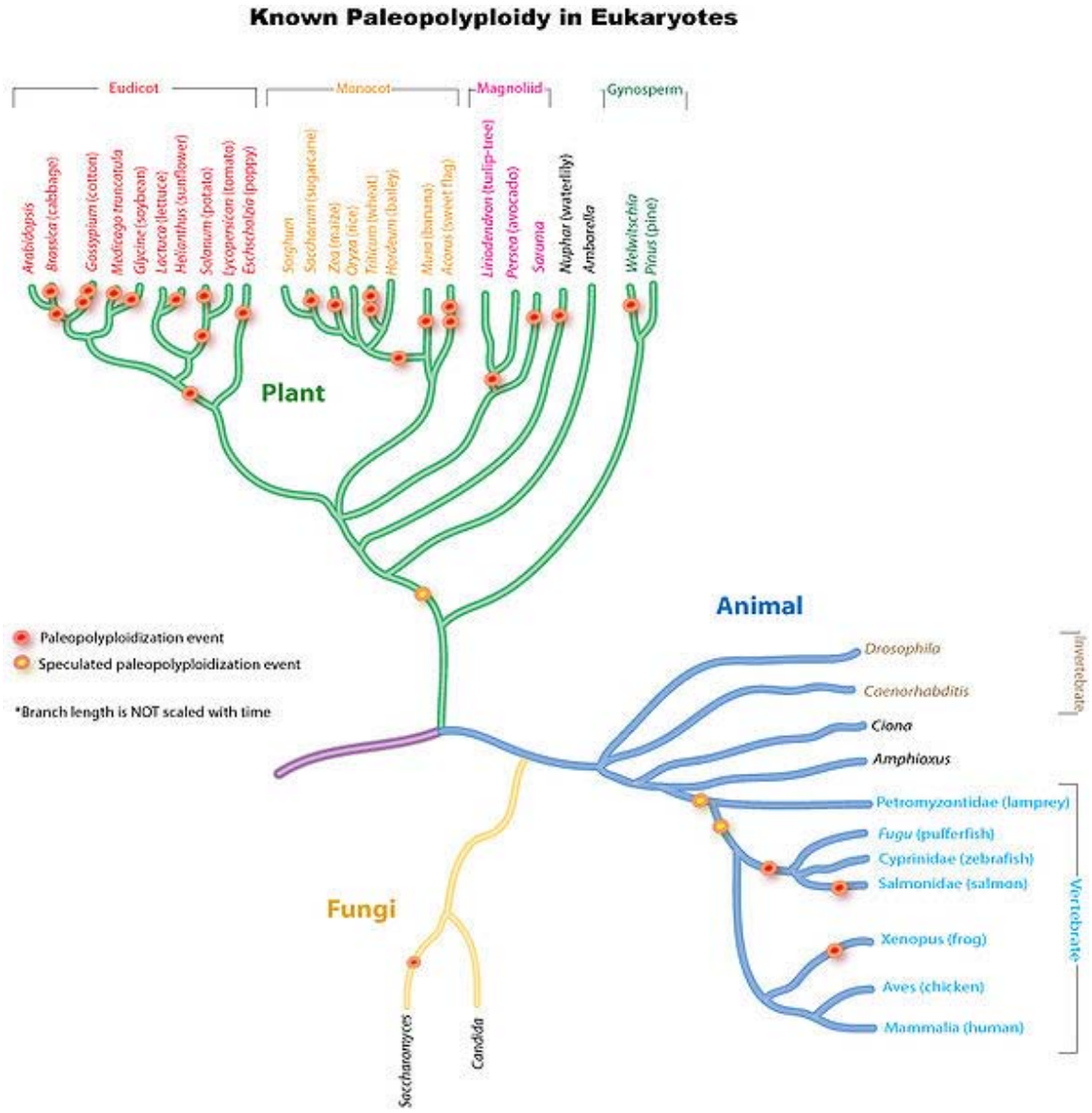
Autopolyploids are polyploids with multiple chromosome sets derived from a single species. Autopolyploids can arise from a spontaneous, naturally occurring genome doubling, like the potato. Others might form following fusion of 2n gametes (unreduced

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**

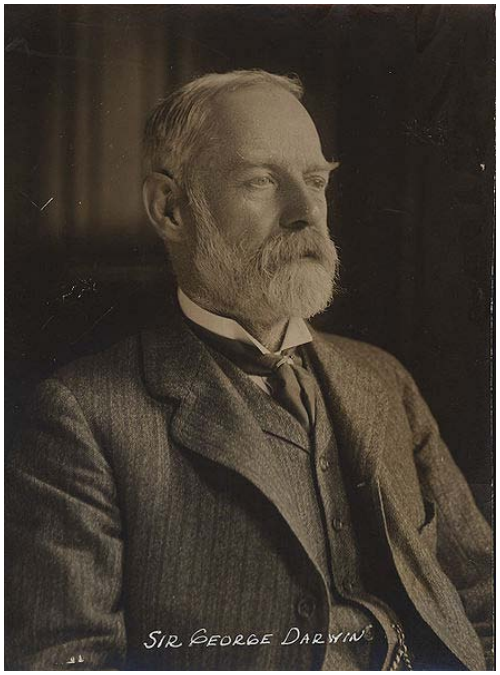
Durum wheat is the result of the inter-species hybridization of two diploid grass species *Triticum urartu* and *Aegilops speltoides*. Both the diploid ancestors had two sets of 7 chromosomes, which were similar in terms of size and genes contained on them. Durum wheat contains two sets of chromosomes 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**.

# Chapter 12

# Genetic Genealogy

**Genetic genealogy** is the application of genetics to traditional genealogy. Genetic genealogy involves the use of genealogical DNA testing to determine the level of genetic relationship between individuals.

## *History*




George Darwin, son of Charles Darwin, was the first to estimate the frequency of first-cousin marriages

The investigation of surnames in genetics can be said to go back to George Darwin, a son of Charles Darwin. In 1875, George Darwin used surnames to estimate the frequency of first-cousin marriages and calculated the expected incidence of marriage between people of the same surname (isonymy). He arrived at a figure between 2.25% and 4.5% for

cousin-marriage in the population of Great Britain, with the upper classes being on the high end and the general rural population on the low end. (His parents, Charles Darwin and Emma Wedgwood, were first cousins.) This simple study was innovative for its era. The next stimulus toward using genetics to study family history had to wait until the 1990s, when certain locations on the Y chromosome were identified as being useful for tracing male-to-male inheritance.

Dr. Karl Skorecki, a Canadian nephrologist of Ashkenazi parentage, noticed that a Sephardic fellow-congregant who was a Kohen like himself had completely different physical features. According to Jewish tradition, all Kohanim are descended from the priest Aaron, brother of Moses. Skorecki reasoned that if Kohanim were indeed the descendants of only one man, they should have a common set of genetic markers and should perhaps preserve some family resemblance to each other.

To test that hypothesis, he contacted Professor Michael Hammer of the University of Arizona, a researcher in molecular genetics and pioneer in Y chromosome research. Their report in the *Nature* in 1997 sent shock waves through the worlds of science and religion. A particular marker was indeed more likely to be present in Jewish men from the priestly tradition than in the general Jewish population. It was apparently true that a common descent had been strictly preserved for thousands of years. Moreover, the data showed that there were very few “non-paternity events”.

The first to test the new methodology in general surname research was Bryan Sykes, a molecular biologist at Oxford University. His study of the Sykes surname obtained valid results by looking at only four markers on the male chromosome. It pointed the way to genetics becoming a valuable assistant in the service of genealogy and history.

In April 2000, Family Tree DNA began offering the first genetic genealogy tests to the public. This offering marked the first time that a personal theory on the Y chromosome could be tested outside of an academic study. Additionally, Sykes’ concept of a surname study, which by this time had been adopted by several other academic researchers outside of Oxford, was expanded into online Surname Projects (an early form of social network) and the effort helped spread knowledge gained through testing to interested genealogists worldwide.

In 2001, Sykes went on to write the popular book *The Seven Daughters of Eve*, which described the seven major haplogroups of European ancestors. In the wake of the book's success, and with the growing availability and affordability of genealogical DNA testing, genetic genealogy as a field began growing rapidly. By 2003, the field of DNA testing of surnames was declared officially to have “arrived” in an article by Jobling and Tyler-Smith in *Nature Reviews Genetics*. The number of firms offering tests, and the number of consumers ordering them, had risen dramatically.

Another milestone in the acceptance of genetic genealogy is the Genographic Project. The Genographic Project is a five-year research study launched in 2005 by the National Geographic Society and IBM, in partnership with the University of Arizona and Family

Tree DNA. Although its goals are primarily anthropological, not genealogical, the project's sale by April 2010 of more than 350,000 of its public participation testing kits, which test the general public for either twelve STR markers on the Y chromosome or mutations on the HVR1 region of the mtDNA, has helped increase the visibility of genetic genealogy.

More state-of-the-art commercial laboratories now recommend testing at least 25 markers, since the more markers tested, the more discriminating and powerful the results will be. A 12-marker STR test is usually not discriminating enough to provide conclusive results for a common surname. Genetic laboratories such as Genebase and Family Tree DNA give the option of testing 67 Y-DNA Markers.

Annual sales of genetic genealogical tests for all companies, including the laboratories that support them, are estimated to be in the area of $60 million (2006).

## *Interpretation*

Since the year 2000, dozens of relevant academic papers have been published, and thousands of private test results organised by surname study groups have been made available on the internet. The comparison of results may be complicated by the fact that some laboratories use different testing methods. Apparently differing results from two sources may in fact be identical, and vice-versa.

## *Uses*

### Paternal and maternal lineages via DNA testing

The two most common types of genetic genealogy tests are Y-DNA (paternal line) and mtDNA (maternal line) genealogical DNA tests. Note that Y chromosome and Y-DNA are used interchangeably here.

These tests involve the comparison of certain sequences of the DNA of pairs of individuals in order to estimate the probability that they share a common ancestor in a genealogical time frame and, through the use of a Bayesian model published by Bruce Walsh, to estimate the number of generations separating the two individuals from their most recent common ancestor or "mrca".

Y-DNA testing involves short tandem repeat (STR) and, sometimes, single nucleotide polymorphism (SNP) testing of the Y-chromosome. The Y-chromosome is present only in males and reveals information on the strict paternal line. These tests can provide insight into the recent (via STRs) and ancient (via SNPs) genetic ancestry. A Y-chromosome STR test will reveal a haplotype, which should be similar among all male descendants of a male ancestor. SNP tests are used to assign people to a paternal haplogroup, which defines a much larger genetic population.

mtDNA testing involves sequencing or testing the HVR-1 region, HVR-2 region or both. An mtDNA test may also include the additional SNPs needed to assign people to a maternal haplogroup—or even include the complete mtDNA.

Either Y-DNA or mtDNA test results can be compared to the results of others via private or public DNA databases.

## Biogeographical and ethnic origins

Additional DNA tests exist for determining biogeographical and ethnic origin, but these tests have less relevance for traditional genealogy.

Genetic genealogy has revealed astonishing links between peoples. For instance, it has shown that the ancient Phoenician people were ancestors of much of the present-day population of the island of Malta. Preliminary results from a study by Pierre Zalloua of the American University of Beirut and Spencer Wells, supported by a grant from National Geographic's Committee for Research and Exploration, were published in the October 2004 issue of *National Geographic*. One of the conclusions is that "more than half of the Y chromosome lineages that we see in today's Maltese population could have come in with the Phoenicians."

## Human migration

Genealogical DNA testing methods are also being used on a longer time scale to trace human migratory patterns. For example, they have been used to determine when the first humans came to North America and what path they followed.

For several years, a number of researchers and laboratories from around the world have been sampling indigenous populations from around the globe in an effort to map historical human migration patterns. Recently, several projects have been created that are aimed at bringing this science to the public. One example, mentioned in History above, is the National Geographic Society's Genographic Project, which aims to map historical human migration patterns by collecting and analyzing DNA samples from over 100,000 people across five continents. Another example is the DNA Clans Genetic Ancestry Analysis, which measures a person's precise genetic connections to indigenous ethnic groups from around the world.

## Typical customers and interest groups

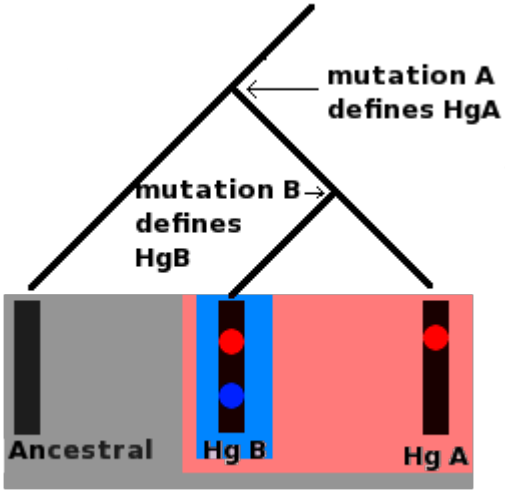
Male DNA testing customers most often start with a Y chromosome test to determine their father's paternal ancestry. Females generally begin with a mitochondrial test to trace their ancient maternal lineage, which males often have tested for the same purpose.

A common consumer goal in purchasing DNA testing services is to acquire quantified, scientific linkage to a specific ancestral group. A compelling example of this motive is found in the expressed desires of some consumers to be proven to have Viking paternal

ancestry. In keeping with this marketplace demand, one British DNA testing service, Oxford Ancestors, offers a Y chromosome test purporting to assess whether given males are of "Viking stock." Those whose DNA falls into the designated haplogroup are issued Viking Descendant certificates by the testing service. The same DNA testing company participated in producing a televised documentary, "The Blood of the Vikings," in conjunction with the BBC, which showed how DNA testing could reveal Viking ancestry.

The RootsWeb Genealogy-DNA Internet discussion group has a membership of 750 subscribers from around the world. Some subscribers have had various DNA tests performed and are seeking advice and guidance in interpreting their results. The list also includes administrators of DNA projects that examine surnames, geographic regions, or ethnic groups. The sophistication of subscribers ranges from expert to novice. In some cases, subscribers have been credited with making useful and novel contributions to knowledge in the field of genetic genealogy.

## *Paternal and maternal DNA lineages*




Ancestral Haplogroup
Haplogroup A (Hg A)
Haplogroup B (Hg B)

All of these molecules are part of the ancestral haplogroup, but at some point in the past a mutation occurred in the ancestral molecule, mutation A, which produced a new lineage; this is haplogroup A and is defined by mutation A. At some more recent point in the past, a new mutation, mutation B, occurred in a person carrying haplogroup A; mutation B defined haplogroup B. Haplogroup B is a subgroup, or subclade of haplogroup A; both haplogrups A and B are subclades of the ancestral haplogroup.

Mitochondria are small organelles that lie in the cytoplasm of eukaryotic cells, such as those of humans. Their primary purpose is to provide energy to the cell. Mitochondria are

thought to be the vestigial remains of symbiotic bacteria that were once free living. One indication that mitochondria were once free living is that they contain a relatively small circular segment of DNA, called mitochondrial DNA (mtDNA). The overwhelming majority of a human's DNA is contained in chromosomes in the nucleus of the cell, but mtDNA is an exception. Individuals inherit their cytoplasm and the organelles it contains exclusively from their mothers, as these are derived from the ovum (egg cell) only, not from the sperm.

When a mutation arises in mtDNA molecule, the mutation is therefore passed in a direct female line of descent. These rare mutations are derived from copying mistakes—when the DNA is copied it is possible that a single mistake occurs in the DNA sequence, an outcome which is called a single nucleotide polymorphism (SNP).

Human Y chromosomes are male-specific sex chromosomes; nearly all humans that possess a Y chromosome will be morphologically male. Y chromosomes are therefore passed from father to son; although Y chromosomes are situated in the cell nucleus, they only recombine with the X chromosome at the ends of the Y chromosome; the vast majority of the Y chromosome (95%) does not recombine. When mutations (SNPs, and STR copying mistakes) arise in the Y chromosome, they are passed down directly from father to son in a direct male line of descent. The Y-DNA and mtDNA therefore share a certain feature: they both pass down unchanged except for mutations.

The other chromosomes, autosomes and X chromosomes in women, share their genetic material (called crossing over leading to recombination) during meiosis (a special type of cell division that occurs for the purposes of sexual reproduction). Effectively this means that the genetic material from these chromosomes gets mixed up in every generation, and so any new mutations are passed down randomly from parents to offspring.

The special feature that both Y-DNA and mtDNA share, above, preserves a "written" record of their mutations because neither DNA gets mixed up or randomized—mutations remain fixed in place on both types of DNA. Furthermore the historical sequence of these mutations can also be inferred. For example, if a set of ten Y chromosomes (derived from ten different men) contains a mutation, A, but only five of these chromosomes contain a second mutation, B, it must be the case that mutation B occurred after mutation A.

Furthermore all ten men who carry the chromosome with mutation A are the direct male line descendants of the same man who was the first to carry this mutation. The first man to carry mutation B was also a direct male line descendant of this man, but is also the direct male line ancestor of all men carrying mutation B. Series of mutations such as this form molecular lineages. Furthermore each SNP mutation may define a set of specific Y chromosomes called a haplogroup.

All men carrying SNP mutation A form a single haplogroup, and all men carrying mutation B are part of this haplogroup, but mutation B (if a SNP) may also define a more recent haplogroup (which is a subgroup or subclade) of its own which men carrying only

mutation A do not belong to. Both mtDNA and Y chromosomes or Y-DNA are grouped into lineages and haplogroups; these are often presented as tree-like diagrams.

## Benefits

Genetic genealogy gives genealogists a means to check or supplement their genealogy results with information obtained via DNA testing. A positive test match with another individual may:

- provide locations for further genealogical research
- help determine ancestral homeland
- discover living relatives
- validate existing research
- confirm or deny suspected connections between families
- prove or disprove theories regarding ancestry
- global culture awareness

## Drawbacks

People who resist testing may cite one of the following concerns:

- Cost
- Quality of testing
- Concerns over privacy issues

Finally, Y-DNA and mtDNA tests each only trace a single lineage (one's father's father's father's etc. lineage or one's mother's mother's mother's etc. lineage). At 10 generations back, an individual has up to 1024 unique ancestors (fewer if ancestor cousins interbred) and a Y-DNA or mtDNA test is only studying one of those ancestors, as well as their descendants and siblings (same sexed siblings for Y-DNA or all siblings for mtDNA). However, most genealogists maintain contact with many cousins (1st, 2nd, 3rd, etc., with different surnames) whose Y-DNA and mtDNA are different, and thus can be encouraged to be tested to find additional ancestral DNA lineages.

## Expected growth

Genetic genealogy is a rapidly growing field. As the cost of testing continues to drop, the number of people being tested continues to increase. The probability of finding a genetic match among the DNA databases should continue to improve. Laboratories and testing firms are engaging in active research and development that will allow for higher confidence intervals and better results interpretation, including historical interpretive reports and customized research.

## *Genetic distance among relatives*

Where the genogram or family tree of individuals is known, it can be used to determine the genetic identity between individuals. It is often described as percentage of genetic identity, referring to the fraction of genome inherited from common ancestors, and not actual genomic identity, which is always approximately 99.9% identical from one human to another.

One method of calculating this genetic similarity is to do an inbreeding calculation by the path or tabular method and then multiply by 2, because any progeny would have a 1 in 2 risk of actually inheriting the identical alleles from both parents. For instance, a brother/sister relation gives 25% risk for two alleles to be identical by descent.

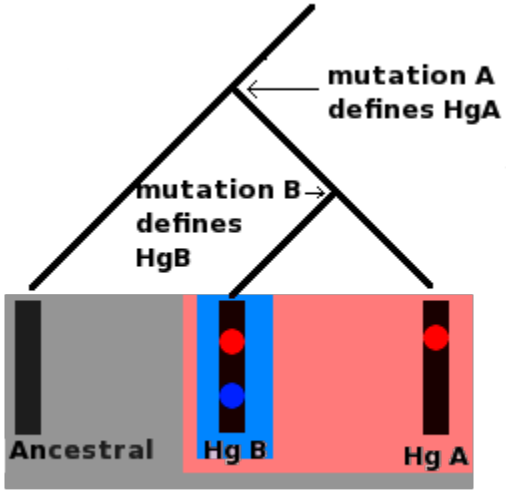
# Chapter 13

# Haplogroup

In the study of molecular evolution, a **haplogroup** is a group of similar haplotypes that share a common ancestor with a single nucleotide polymorphism (SNP) mutation. Because a haplogroup consists of similar haplotypes, this is what makes it possible to predict a haplogroup from haplotypes. An SNP test confirms a haplogroup. Haplogroups are assigned letters of the alphabet, and refinements consist of additional number and letter combinations, for example R1b1. Y-chromosome and mitochondrial DNA haplogroups have different haplogroup designations. Haplogroups pertain to deep ancestral origins dating back thousands of years.

In human genetics, the haplogroups most commonly studied are Y-chromosome (Y-DNA) haplogroups and mitochondrial DNA (mtDNA) haplogroups, both of which can be used to define genetic populations. Y-DNA is passed solely along the patrilineal line, from father to son, while mtDNA is passed down the matrilineal line, from mother to offspring of both sexes. Neither recombines, and thus Y-DNA and mtDNA change only by chance mutation at each generation with no intermixture between parents' genetic material.

## *Haplogroup formation*




Ancestral Haplogroup
Haplogroup A (Hg A)
Haplogroup B (Hg B)

All of these molecules are part of the ancestral haplogroup, but at some point in the past a mutation occurred in the ancestral molecule, mutation A, which produced a new lineage; this is haplogroup A and is defined by mutation A. At some more recent point in the past, a new mutation, mutation B, occurred in a person carrying haplogroup A; mutation B defined haplogroup B. Haplogroup B is a subgroup, or subclade of haplogroup A; both haplogrups A and B are subclades of the ancestral haplogroup.

Mitochondria are small organelles that lie in the cytoplasm of eucaryotic cells, such as those of humans. Their primary purpose is to provide energy to the cell. Mitochondria are thought to be reduced descendants of symbiotic bacteria that were once free living. One indication that mitochondria were once free living is that each contains a circular DNA, called mitochondrial DNA (mtDNA), whose structure is more similar to bacteria than eukaryotic organisms. The overwhelming majority of a human's DNA is contained in the chromosomes in the nucleus of the cell, but mtDNA is an exception.

An individual inherits their cytoplasm and the organelles it contains exclusively from their mother, as these are derived from the ovum (egg cell), sperm only carry chromosomal DNA due to the necessity of maintaining motility. When a mutation arises in mtDNA molecule, the mutation is therefore passed in a direct female line of descent. These mutations are derived from copying mistakes, when the DNA is copied it is possible that a single mistake occurs in the DNA sequence, these single mistakes are called single nucleotide polymorphisms (SNPs).

Human Y chromosomes are male-specific sex chromosomes; nearly all humans that possess a Y chromosome will be morphologically male. Y chromosomes are therefore

passed from father to son; although Y chromosomes are situated in the cell nucleus, they only recombine with the X chromosome at the ends of the Y chromosome; the vast majority of the Y chromosome (95%) does not recombine. When mutations (SNPs) arise in the Y chromosome, they are passed on directly from father to son in a direct male line of descent. The Y chromosome and mtDNA therefore share specific properties.

Other chromosomes, autosomes and X chromosomes in women, share their genetic material (called crossing over leading to recombination) during meiosis (a special type of cell division that occurs for the purposes of sexual reproduction). Effectively this means that the genetic material from these chromosomes gets mixed up in every generation, and so any new mutations are passed down randomly from parents to offspring.

The special feature that both Y chromosomes and mtDNA display is that mutations can accrue along a certain segment of both molecules and these mutations remain fixed in place on the DNA. Furthermore the historical sequence of these mutations can also be inferred. For example, if a set of ten Y chromosomes (derived from ten different men) contains a mutation, A, but only five of these chromosomes contain a second mutation, B, it must be the case that mutation B occurred after mutation A.

Furthermore all ten men who carry the chromosome with mutation A are the direct male line descendants of the same man who was the first person to carry this mutation. The first man to carry mutation B was also a direct male line descendant of this man, but is also the direct male line ancestor of all men carrying mutation B. Series of mutations such as this form molecular lineages. Furthermore each mutation defines a set of specific Y chromosomes called a haplogroup.

All men carrying mutation A form a single haplogroup, all men carrying mutation B are part of this haplogroup, but mutation B also defines a more recent haplogroup (which is a subgroup or subclade) of its own which men carrying only mutation A do not belong to. Both mtDNA and Y chromosomes are grouped into lineages and haplogroups; these are often presented as tree like diagrams.

## *Haplogroup population genetics*

It is usually assumed that there is little natural selection for or against a particular haplotype mutation which has survived to the present day, so apart from mutation rates (which may vary from one marker to another) the main driver of population genetics affecting the proportions of haplotypes in a population is genetic drift — random fluctuation caused by the sampling randomness of which members of the population happen to pass their DNA on to members of the next generation of the appropriate sex.

This causes the prevalence of a particular marker in a population to continue to fluctuate, until it either hits 100%, or falls out of the population entirely. In a large population with efficient mixing the rate of genetic drift for common alleles is very low; however, in a very small interbreeding population the proportions can change much more quickly. The marked geographical variations and concentrations of particular haplotypes and groups of

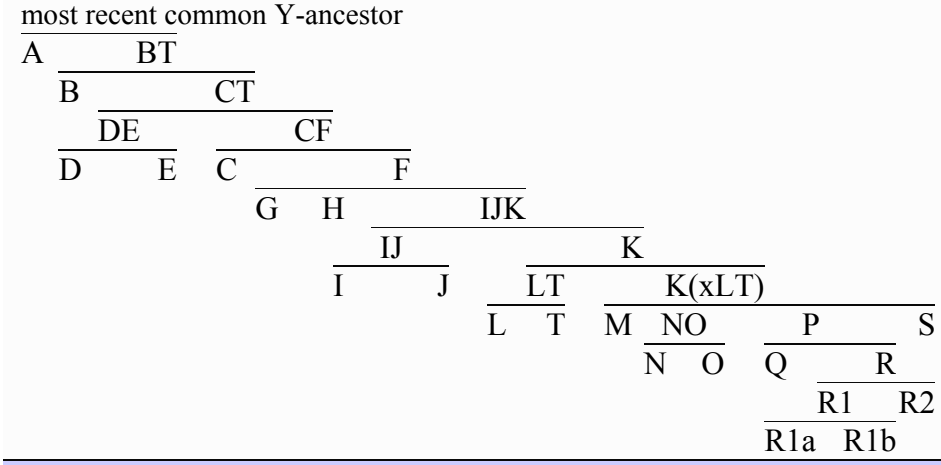
haplotypes therefore witness the distinctive effects of repeated population bottlenecks or founder events followed by population separations and increases.

The lineages which can be traced back from the present will not reflect the full genetic variation of the older population: genetic drift means that some of the variants will have died out. The cost of full Y-DNA and mtDNA sequence tests has limited the availability of data; however, their cost has dropped dramatically in the last decade. Haplotype coalescence times and current geographical prevalences both carry considerable error uncertainties. This is especially troublesome for coalescence times, because most population geneticists still continue (albeit decresing a little bit) to use the "Zhivotovski method", which is heavily criticised by DNA-genealogists for its' falsehood.

## Human Y-chromosome DNA haplogroups

Human Y chromosome DNA (Y-DNA) haplogroups are named from A to T, and are further subdivided using numbers and lower case letters. Y chromosome haplogroup designations are established by the Y Chromosome Consortium.

**Evolutionary tree of Human Y-chromosome DNA (Y-DNA) haplogroups**




Y-DNA by ethnic groups · Famous Y-DNA haplotypes

Y-chromosomal Adam is the name given by researchers to the male who is the most recent common patrilineal (male-lineage) ancestor of all living humans.

Major Y-chromosome haplogroups, and their geographical regions of occurrence (prior to the recent European colonization), include:

**Groups without mutation M168**

- Haplogroup A (M91) (Africa, especially the Khoisan, Ethiopians, and Nilotes)

- Haplogroup B (M60) (Africa, especially the Pygmies and Hadzabe)

**Groups with mutation M168**

(mutation M168 occurred ~50,000 bp)

- Haplogroup C (M130) (Oceania, North/Central/East Asia, North America and significant presence in India)
- Haplogroup F (M89) Oceania, Europe, Asia, North- and South- America
- YAP+ haplogroups
    - Haplogroup DE (M1, M145, M203)
        - Haplogroup D (M174) (Tibet, Japan, the Andaman Islands)
        - Haplogroup E (M96)
            - Haplogroup E1b1a (V38) West Africa and surrounding regions; formerly known as E3a
            - Haplogroup E1b1b (M215) East Africa, North Africa, the Middle East, the Mediterranean, the Balkans; formerly known as E3b

**Groups with mutation M89**

(mutation M89 occurred ~45,000 bp)

- Haplogroup F (P14, M213) (southern India, Sri Lanka, China, Korea)

- Haplogroup G (M201) (present among many ethnic groups in Eurasia, usually at low frequency; most common in the Caucasus, the Iranian plateau, and Anatolia; in Europe mainly in Greece, Italy, Iberia, the Tyrol, Bohemia; extremely rare in Northern Europe)

- Haplogroup H (M69) (India, Sri Lanka, Nepal, and at low frequency in Pakistan, Iran, Central Asia, and Arabia)

- Haplogroup IJK (L15, L16)

**Groups with mutations L15 & L16**

- Haplogroup IJK (L15, L16)
    - Haplogroup IJ (S2, S22)
        - Haplogroup I (M170, P19, M258) (widespread in Europe, found infrequently in parts of the Middle East, and virtually absent elsewhere)
            - Haplogroup I1 (M253, M307, P30, P40) (Northern Europe)
            - Haplogroup I2 (S31) (Central and Southeast Europe, Sardinia)

        - Haplogroup J (M304) (the Middle East, Turkey, Caucasus, Italy, Greece, the Balkans, North and Northeast Africa)
            - Haplogroup J* (Mainly found in Socotra, with a few observations in Pakistan, Oman, Greece, Czechia, and among Turkic peoples)
            - Haplogroup J1 (M267) (Mostly associated with Semitic peoples in the Middle East, Ethiopia, and North Africa, and with Northeast Caucasian peoples in Dagestan; J1 with DYS388=13 is associated with eastern Anatolia)
            - Haplogroup J2 (M172) (Mainly found in West Asia, Central Asia, South Asia, Southern Europe, and North Africa)
    - Haplogroup K (M9, P128, P131, P132)

**Groups with mutation M9**

(mutation M9 occurred ~40,000 bp)

- Haplogroup K
    - Haplogroup LT (L298/P326)
        - Haplogroup L (M11, M20, M22, M61, M185, M295) (South Asia, Central Asia, Southwestern Asia, the Mediterranean)
        - Haplogroup T (M70, M184/USP9Y+3178, M193, M272) (North Africa, Horn of Africa, Southwest Asia, the Mediterranean, South Asia); formerly known as Haplogroup K2
    - Haplogroup K(xLT) (rs2033003/M526)

***Groups with mutation M526***

-
    - Haplogroup M (P256) (New Guinea, Melanesia, eastern Indonesia)
        - Haplogroup NO (M214)
            - Haplogroup N (M231) (northernmost Eurasia, especially among the Uralic peoples)
            - Haplogroup O (M175) (East Asia, Southeast Asia, the South Pacific, South Asia, Central Asia)
                - Haplogroup O1 (MSY2.2)
                - Haplogroup O2 (P31, M268)
                - Haplogroup O3 (M122)
        - Haplogroup P (M45, 92R7, M74/N12) (M45 occurred ~35,000 bp)
            - Haplogroup Q (MEH2, M242, P36) (Occurred ~15,000-20,000 years ago. Found in Asia and the Americas)
                - Haplogroup Q1a3a1 (M3) (North America, Central America, and South America)

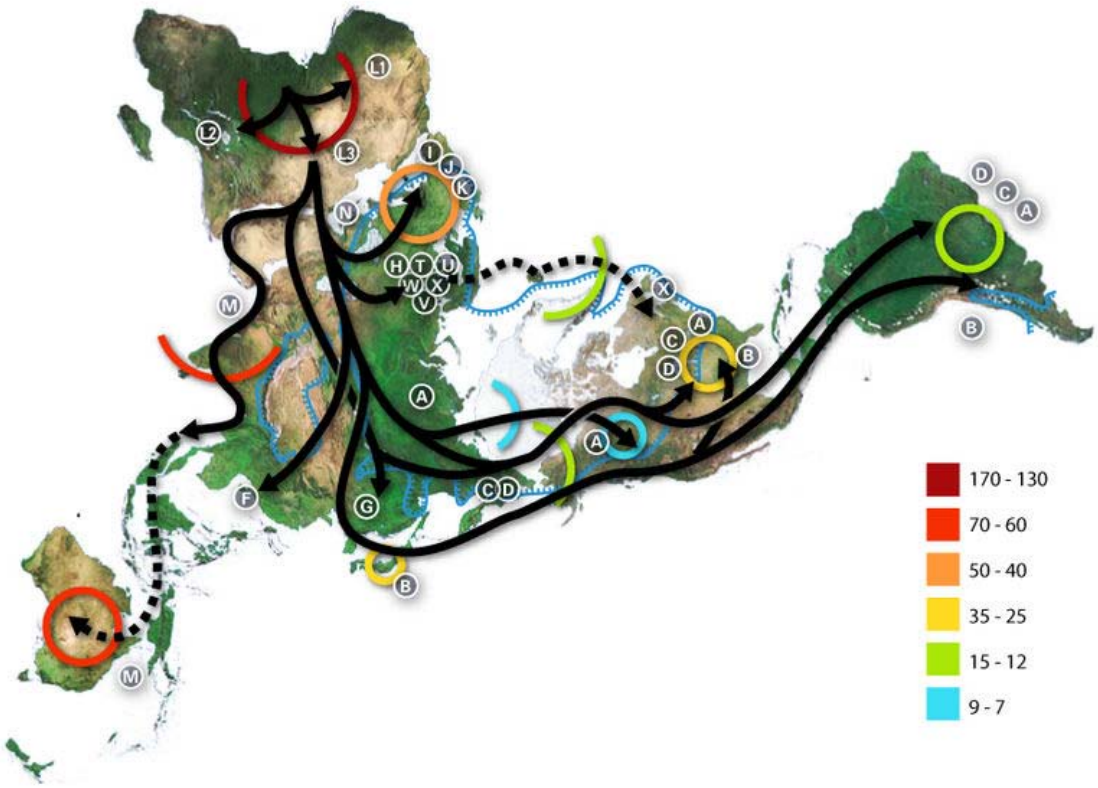
- Haplogroup R (M207)
  - Haplogroup R1 (M173)
    - Haplogroup R1a (M17) (Central Asia, South Asia, and Central, Northern, and Eastern Europe)
    - Haplogroup R1b (M343) (Europe, Caucasus, Central Asia, South Asia, North Africa, Central Africa)
  - Haplogroup R2 (M124) (South Asia, Caucasus, Central Asia)
- Haplogroup S (M230, P202, P204) (New Guinea, Melanesia, eastern Indonesia)

## Human mitochondrial DNA haplogroups

Human mtDNA haplogroups are lettered: A, B, C, CZ, D, E, F, G, H,

HV, I, J, pre-JT, JT, K, L0, L1, L2, L3, L4, L5, L6, M, N, P, Q, R, R0, S, T, U, V, W, X, Y, and Z. The most up-to-date version of the mtDNA tree is maintained by Mannis van Oven on the PhyloTree website.

## Defining populations




Map of human haplotype migration, according to mitochondrial DNA.

Haplogroups can be used to define genetic populations and are often geographically oriented. For example, the following are common divisions for mtDNA haplogroups:

- African: L0, L1, L2, L3, L4, L5, L6
- West Eurasian: H, T, U, V, X, K, I, J, W *(all listed West Eurasian haplogroups are derived from macro-haplogroup N)*
- East Eurasian: A, B, C, D, E, F, G, Y (*note: C, D, E, and G belong to macro-haplogroup M*)
- Native American: A, B, C, D, X
- Australo-Melanesian: P, Q, S

The mitochondrial haplogroups are divided into 3 main groups, which are designated by the 3 sequential letters L, M, N. Humanity first split within the L group between L0 and L1-6. L1-6 gave rise to other L groups, one of which, L3, split into the M and N group. The M group comprises the first wave of human migration out of Africa, following an eastward route along southern coastal areas.

Descendent populations belonging to haplogroup M are found throughout East Africa, Asia, the Americas, and Melanesia, though almost none have been found in Europe. The N group may represent another migration out of Africa, heading northward instead of eastward. Shortly after the migration, the large R group split off from the N.

Haplogroup R consists of two subgroups defined on the basis of their geographical distributions, one found in southeastern Asia and Oceania and the other containing almost all of the modern European populations. Haplogroup N(xR), i.e. mtDNA that belongs to the N group but not to its R subgroup, is typical of Australian aboriginal populations, while also being present at low frequencies among many populations of Eurasia and the Americas.

**The L type** consists of nearly all Africans.

**The M type** consists of:

M1- Ethiopian, Somali and Indian populations. Likely due to much gene flow between the Horn of Africa and the Arabian Peninsula (Saudi Arabia, Yemen, Oman), separated only by a narrow strait between the Red Sea and the Gulf of Aden.

CZ- Many Siberians; branch C- Some Amerindian; branch Z- Many Saami, some Korean, some North Chinese, some Central Asian populations.

D- Some Amerindians, many Siberians and northern East Asians

E- Malay, Borneo, Philippines, Taiwan aborigines, Papua New Guinea

G- Many Northeast Siberians, northern East Asians, and Central Asians

Q- Melanesian, Polynesian, New Guinean populations

**The N type** consists of:

A- Found in some Amerindians, Japanese, and Koreans

I- 10% frequency in Northern, Eastern Europe

S- Some Australian aborigines

W- Some Eastern Europeans, South Asians, and southern East Asians

X- Some Amerindians, Southern Siberians, Southwest Asians, and Southern Europeans

Y- Most Nivkhs and many Ainus; 1% in Southern Siberia

**R**- Large group found within the N type.Populations contained therein can be divided geographically into West Eurasia and East Eurasia. Almost all European populations and a large number of Middle-Eastern population today are contained within this branch. A smaller percentage is contained in other N type groups (See above). Below are **subclades of R**:

B- Some Chinese, Tibetans, Mongolians, Central Asians, Koreans, Amerindians, South Siberians, Japanese, Austronesians

F- Mainly found in southeastern Asia, especially Vietnam; 8.3% in Hvar Island in Croatia.

R0- Found in Arabia and among Ethiopians and Somalis; branch HV (branch H; branch V)- Europe, Western Asia, North Africa;

Pre-JT- Arose in the Levant (modern Lebanon area), found in 25% frequency in Bedouin poupulations; branch JT (branch J; branch T)- North, Eastern Europe, Indus, Mediterranean

U- High frequency in Scandinavia, Baltic countries, Mediterranean

## *Overlap between y-haplogroups and mt-haplogroups*

The ranges of specific y-haplogroups and specific mt-haplogroups overlap, indicating populations that have a specific combination of a y-haplogroup and an mt-haplogroup. Y mutations and mt mutations do not necessarily occur at a similar time, and differential rates of sexual selection between the two genders combined with founder effect and genetic drift can alter the haplogroup composition of a population, so the overlaps are only rough.

## Chapter 14

# F1 Hybrid and Ploidy

# F1 hybrid

**F1 hybrid** is a term used in genetics and selective breeding. **F1** stands for *Filial 1*, the first filial generation seeds/plants or animal offspring resulting from a cross mating of distinctly different parental types. The term is sometimes written with a subscript, as **$F_1$ hybrid.** The offspring of distinctly different parental types produce a new, uniform variety with specific characteristics from either or both parents. In fish breeding, those parents frequently are two closely related fish species, while in plant and animal genetics those parents usually are two inbred lines. Mules are F1 hybrids between horse and donkey. Today, certain domestic hybrid breeds, such as the Savannah cat, are classified by their filial generation number.

Gregor Mendel's groundbreaking work in the 19th century focused on patterns of inheritance and the genetic basis for variation. In his cross-pollination experiments involving two true-breeding, or homozygous, parents, Mendel found that the resulting F1 generation were heterozygous and consistent. The offspring showed a combination of those phenotypes from each of the parents that were genetically dominant. Mendel's discoveries involving the F1 and F2 generation laid the foundation for modern genetics.

## *Production of F1 hybrids*

### In plants

Crossing two genetically different plants produces a hybrid seed (plant) by means of controlled pollination. To produce consistent F1 hybrids, the original cross must be repeated each season. As in the original cross, in plants this is usually done through controlled hand-pollination, and explains why F1 seeds can often be expensive. F1 hybrids can also occur naturally, a prime example being peppermint, which is not a species evolved by cladogenesis or gradual change from a single ancestor, but a sterile

stereotyped hybrid of watermint and spearmint. Unable to produce seeds, it propagates through the vining spread of its own root system.

In agronomy, the term “F1 hybrid” is usually reserved for agricultural cultivars derived from two different parent cultivars, each of which are inbred for a number of generations to the extent that they are almost homozygous. The divergence between the parent lines promotes improved growth and yield characteristics through the phenomenon of heterosis ("hybrid vigour"), whilst the homozygosity of the parent lines ensures a phenotypically uniform F1 generation. Each year, for example, specific tomato "hybrids" are specifically recreated by crossing the two parent heirloom cultivars over again.

Two populations of breeding stock with desired characteristics are subject to inbreeding until the homozygosity of the population exceeds a certain level, usually 90% or more. Typically this requires more than ten generations. After this happens, both populations must be crossed while avoiding self-fertilization. Normally this happens in plants by deactivating or removing male flowers from one population, taking advantage of time differences between male and female flowering or hand-pollinating.

In 1960, 99 percent of all corn planted in the United States, 95 percent of sugar beet, 80 percent of spinach, 80 percent of sunflowers, 62 percent of broccoli, and 60 percent of onions were hybrid. Such figures are probably higher today. Beans and peas are not commercially hybridized because they are automatic pollinators, and hand-pollination is prohibitively expensive.

**F2 hybrid**

While an F2 hybrid, the result of self or cross pollination of an F1, does not have the consistency of the F1 hybrid, it may retain some desirable traits and can be produced more cheaply as no intervention in the pollination is required. Some seed companies offer F2 seed, particularly in bedding plants where consistency is not as critical.

### In animals

Unlike most plants, commonly bred fish species as well as all mammals and birds are not hermaphroditic, and therefore it is impossible to achieve self-fertilization during an F1 cross. F1 crosses in fish can be between two inbred lines or between two closely related fish species, such as cichlid subspecies.. The cross is usually performed by natural or artificial insemination.

## *Advantages*

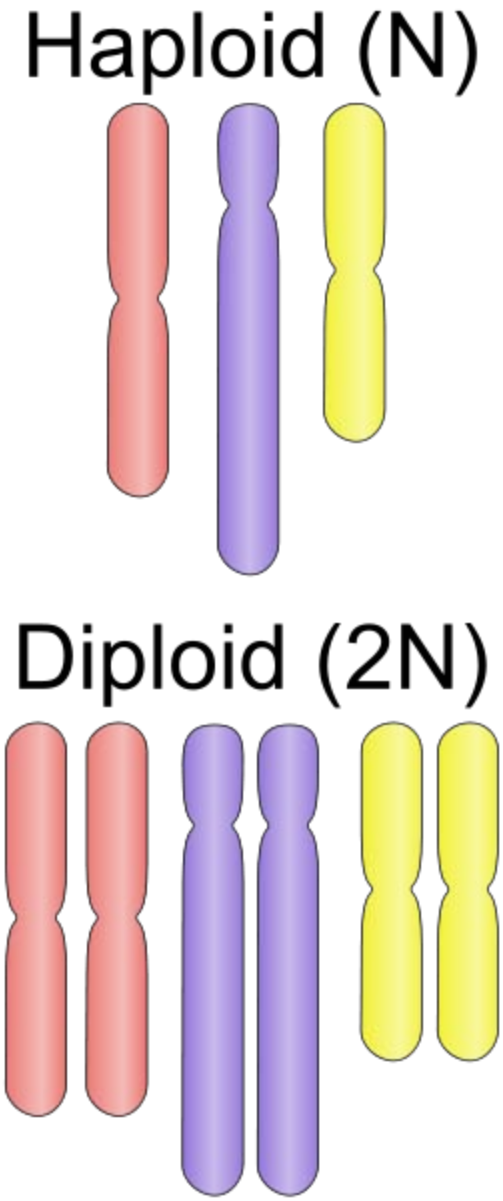
- Homogeneity and predictability - If the parents are homozygous pure lines, there is limited genetic variation between individual F1 plants or animals. This makes their phenotype extremely uniform and thus attractive for mechanical operations and makes it easier to fine-tune the management of the population. Once the

characteristics of the cross are known, repeating this cross will yield exactly the same result.
- Higher performance - As most alleles code for different versions of a protein or enzyme, having two different versions of this allele amounts to having two different versions of the enzyme. This will increase the likelihood of having an optimal version of the enzyme present and reduce the likelihood of a genetic defect. This effect is referred to in genetics as heterosis.
- F1 hybrids can give higher yields than traditional varieties.

## *Disadvantages*

- The main advantage of F1 hybrids in agriculture is also their drawback. When F1 cultivars are used for the breeding of a new generation, their offspring (F2 generation) will vary greatly from one another. Some of the F2 generation will be high in homozygous genes, as found in the weaker parental generation, and these will have a depression in yield and lack the hybrid vigour. From the point of view of a commercial seed producer which does not wish its customers to produce their own seed, this genetic assortment is a desired characteristic.
- Both inbreeding and crossing the lines requires a lot of work, which translates into a much higher seed cost. In general, the higher yield offsets this disadvantage.
- F1 hybrids mature at the same time when raised under the same environmental conditions. This is of interest for modern farmers, because all ripen at the same time and can be harvested by machine. Traditional varieties are often more useful to gardeners because they crop over a longer period of time, avoiding gluts and food shortages.

# Ploidy




Diploid cells have two homologous copies of each chromosome.

**Ploidy** is the number of sets of chromosomes in a biological cell.

Human sex cells (sperm and egg) have one complete set of chromosomes from the male or female parent. Sex cells, also called gametes, combine to produce somatic cells. Somatic cells therefore have twice as many chromosomes. The **haploid number** (n) is the number of chromosomes in a gamete. A somatic cell has twice that many chromosomes (2n).

Humans are **diploid**. A human somatic cell contains 46 chromosomes: 2 complete haploid sets, which make up 23 homologous chromosome pairs. However, many organisms have more than two sets of homologous chromosomes and are called polyploid.

The number of chromosomes in a single (non-homologous) set is called the **monoploid number** (x), and is different from the haploid number (n). Both numbers n, and x, apply to every cell of a given organism. For humans, x = n = 23, which is also written as 2n = 2x = 46. Bread wheat is an organism where x and n differ. It has six sets of chromosomes, two sets from each of three different diploid species that are its distant ancestors. The somatic cells are **hexaploid**, with six sets of chromosomes, 2n = 6x = 42. The gametes are both haploid and **triploid**, with three sets of chromosomes. The monoploid number x = 7, and the haploid number n = 21.

**Tetraploidy** (four sets of chromosomes, 2n = 4x) is common in plants, and also occurs in amphibians, reptiles, and insects.

The Australian bulldog ant, *Myrmecia pilosula*, a haplodiploid species, has n = x = 1, the lowest chromosome number theoretically possible. Haploid individuals of this species have a single chromosome, and diploid individuals have two chromosomes.

**Euploidy** is the state of a cell or organism having an integral multiple of the monoploid number, possibly excluding the sex-determining chromosomes. For example, a human cell has 46 chromosomes, which is an integer multiple of the monoploid number, 23. A human with abnormal, but integral, multiples of this full set (e.g. 69 chromosomes) would also be considered as euploid. **Aneuploidy** is the state of not having euploidy. In humans, examples include having a single extra chromosome (such as Down syndrome), or missing a chromosome (such as Turner syndrome). Aneuploid karyotypes are given names with the suffix *-somy* (rather than *-ploidy*, used for euploid karyotypes), such as trisomy and monosomy.

## *Etymology*

The term *ploidy* is a back-formation from *haploid* and *diploid*. These two terms are from Greek ἁπλόος *haplóos* "single" and διπλόος *diplóos* "double" combined with εἶδος *eîdos* "form" (compare *idol* from Latin *īdōlum*, that from Greek εἴδωλον *eídōlon* derived from εἶδος *eîdos*). The two *haploid* and *diploid* terms were borrowed from German through William Henry Lang's 1908 translation of a 1894 textbook by Eduard Strasburger and colleagues. Strasburger used diploid to refer to an organism with twice the number of chromosomes of a haploid organism, hence "double" and "single".

## *Haploid and monoploid*

As stated above, the **haploid number** (n) is the number of chromosomes in a gamete of an individual, and this is distinct from the monoploid number (x) which is the number of unique chromosomes in a single complete set. Gametes (sperm, and ova) are haploid

cells. The haploid gametes produced by (most) diploid organisms are monoploid, and these can combine to form a diploid zygote. For example, most animals are diploid and produce monoploid gametes.

During meiosis, sex cell precursors have their number of chromosomes halved by randomly "choosing" one homologue, resulting in haploid gametes. Because homologous chromosomes usually differ genetically, gametes usually differ genetically from one another.

All plants and many fungi and algae switch between a haploid and a diploid state (which may be polyploid), with one of the stages emphasized over the other. This is called alternation of generations. Most fungi and algae are haploid during the principal stage of their life cycle.

Male bees, wasps, and ants are haploid organisms because of the way they develop from unfertilized, haploid egg cells.

In humans, the monoploid number (x) equals the haploid number (n), x = n = 23, but in some species (especially plants), these numbers differ. Common wheat has six sets of chromosomes in the somatic cells, derived from its three different ancestral species. The gametes of common wheat are considered as haploid since they contain half the genetic information of somatic cells, but are not monoploid as they still contain three complete sets of chromosomes (n = 3x).

## *Diploid*

**Diploid** (indicated by $2n = 2x$) cells have two homologous copies of each chromosome, usually one from the mother and one from the father. Nearly all mammals are diploid organisms (the viscacha rats *Pipanacoctomys aureus* and *Tympanoctomys barrerae* are the only known exceptions as of 2004), although all individuals have some small fraction of cells that display polyploidy. Human diploid cells have 46 chromosomes and human haploid gametes (egg and sperm) have 23 chromosomes.

Retroviruses that contain two copies of their RNA genome in each viral particle are also said to be diploid. Examples include human foamy virus, human T-lymphotropic virus, and HIV.

## *Haploidisation*

**Haploidisation** (haploidization) is the process of creating a haploid cell (usually from a diploid cell).

A laboratory procedure called haploidisation forces a normal cell to expel half of its chromosomal complement. In mammals this renders this cell chromosomally equal to sperm or egg. This was one of the procedures used by Japanese researchers to produce Kaguya, a fatherless mouse.

Haploidisation sometimes occurs in plants when meiotically reduced cells (usually egg cells) develop by parthenogenesis.

A rare genetic disorder that has occurred in a total of 7 recorded cases is Detrimental Haploidy Syndrome where the somatic cells of the human body are haploid after the first division of cells from fertilisation. As a result of this a human with this syndrome is unfortunately prone to other diseases and unable to reproduce.

## *Polyploidy*

**Polyploidy** is the state where all cells have multiple sets of chromosomes beyond the basic set, for example, in triploids $2n = 3x$, in tetraploids $2n = 4x$. The chromosome sets may be from the same species or from closely related species. In the latter case these are known as allopolyploids (or amphidiploids, which are allopolyploids that behave as if they were normal diploids). Allopolyploids are formed from the hybridization of two separate species. In plants, this probably most often occurs from the pairing of meiotically unreduced gametes, and not by diploid–diploid hybridization followed by chromosome doubling. The so-called Brassica triangle is an example of allopolyploidy, where three different parent species have hybridized in all possible pair combinations to produce three new species.

Polyploidy occurs commonly in plants, but rarely in animals. Even in diploid organisms many somatic cells are polyploid due to a process called endoreduplication where duplication of the genome occurs without mitosis (cell division).

The extreme in polyploidy occurs in the fern-ally genus *Ophioglossum*, the adder's-tongues, in which polyploidy results in chromosome counts in the hundreds, or in at least one case, well over one thousand. Interestingly, these plants seem to have simplified structures in their phenotype.

## *Variable or indefinite ploidy*

Depending on growth conditions, prokaryotes such as bacteria may have a chromosome copy number of 1 to 4, and that number is commonly fractional, counting portions of the chromosome partly replicated at a given time. This is because under exponential growth conditions the cells are able to replicate their DNA faster than they can divide.

## *Mixoploidy*

Mixoploidy refers to the presence of two cell lines, one diploid and one polyploid. Though polyploidy in humans is not viable, mixoploidy has been found in live adults and children. There are two types: diploid-triploid mixoploidy, in which some cells have 46 chromosomes and some have 69, and diploid-tetraploid mixoploidy, in which some cells have 46 and some have 92 chromosomes.

## *Dihaploidy and Polyhaploidy*

Dihaploid and polyhaploid cells are formed by haploidisation of polyploids, i.e., by halving the chromosome constitution.

Dihaploids (which are diploid) are important for selective breeding of tetraploid crop plants (notably potatoes), because selection is faster with diploids than with tetraploids. Tetraploids can be reconstituted from the diploids, for example by somatic fusion.

The term “dihaploid” was coined by Bender to combine in one word the number of genome copies (diploid) and their origin (haploid). The term is well established in this original sense, but it has also been used for doubled monoploids or doubled haploids, which are homozygous and used for genetic research.