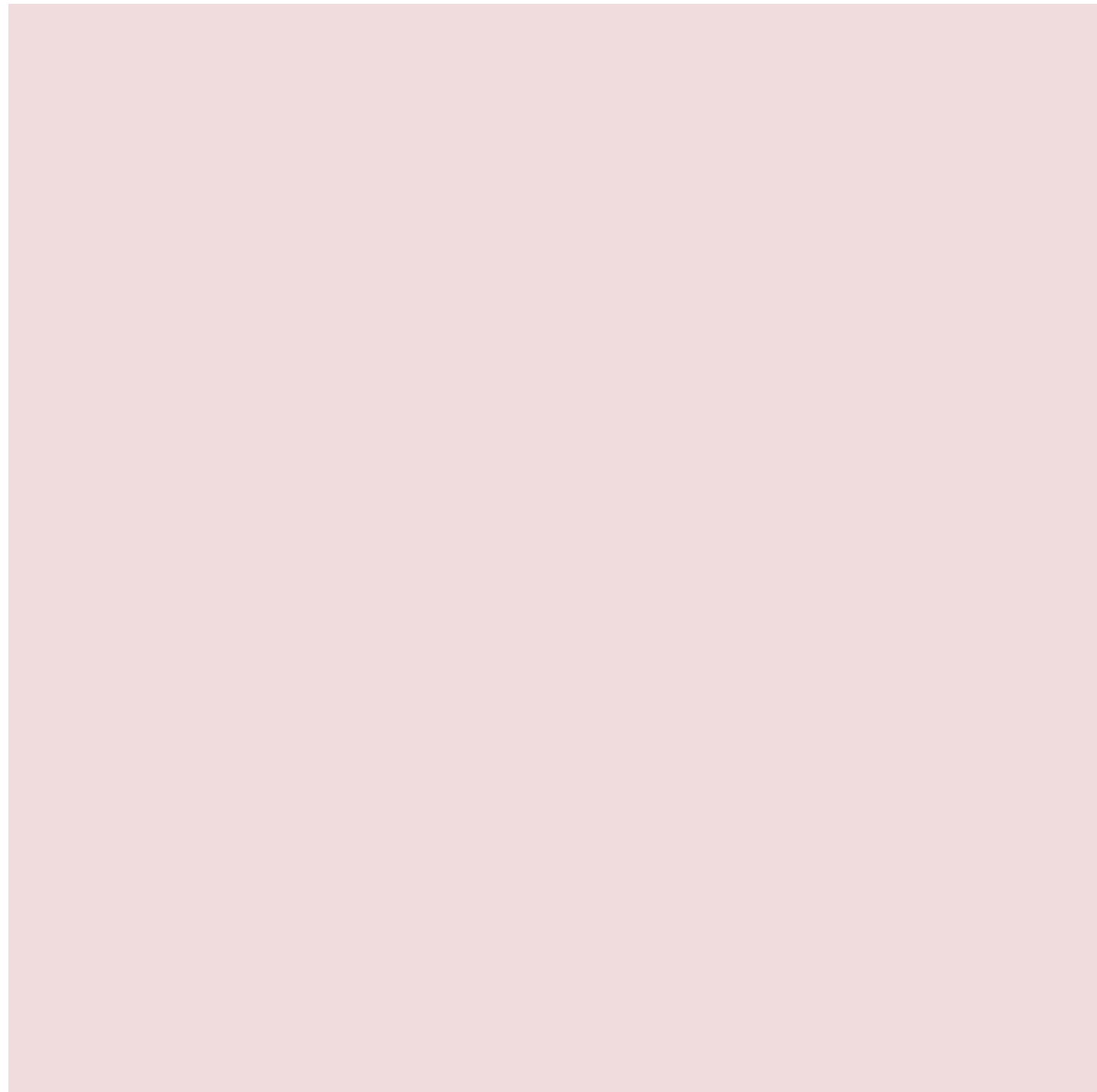# Genetic polymorphism and variation

Auteur :
 VEUILLE Michel, Directeur d'Etudes, Ecole pratique des hautes études, Paris

05-01-2025

Genetic polymorphism is the existence of alternative states of DNA, determining the variation of the organism's higher levels of integration. There are different kinds of genome modifications (mutations). The most studied are nucleotide substitutions in coding and regulatory regions.
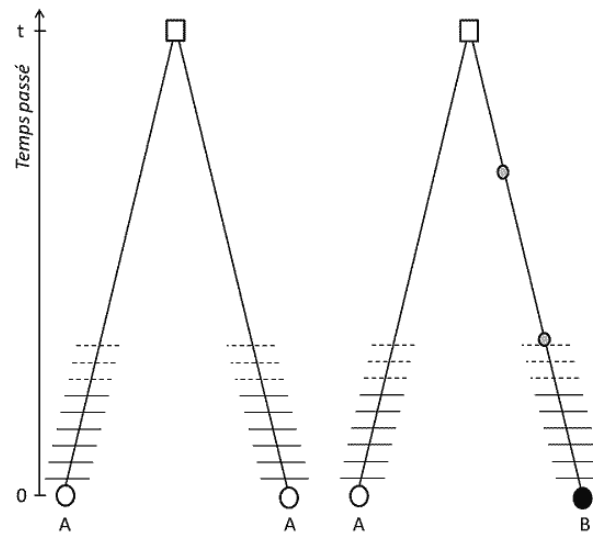
1. Definition

Figure 1. Allelism and homology. Representation of two pairs of genes linked to their last common ancestor (square) by descent lines; horizontal scales represent generations, and grey circles represent mutations in ancestral lines. Two genes are homologous if they have a common ancestor. They are both homologous by position on the chromosome, and homologous by progeny. If they are different (right figure), they are said to be alleles, which means that at least one mutation has occurred since the age (t) of their last common ancestor. Otherwise, they are said to be identical by state (figure on the left). The expectation of the number of mutations between two random genes is Θ =2Tμ, where μ is the mutation rate per unit time (generations) and T is the expectation of t in the population. The probability that two genes randomly selected from the population are alleles, and therefore differ by at least one mutation, is H ≈ Θ/(Θ+1).

Genetic polymorphism is the existence, in a population, of several alternative states of DNA, or **alleles**, at a defined position in the genome, or locus{ind-text}Position of the gene on the chromosome. In population genetics, a set of homologous genes (homology class). Two genes are homologous if they match meiosis. Plural: loci.{end-tooltip}. This definition has several aspects:

(1) First, the trait must be carried by the **chromosomes** and be transmissible.

(2) Then the alleles must be homologous{ind-text}Two chromosomes or two genes are said to be homologous if they match and mutually exclude each other from meiosis.{end-tooltip} for their position in the genome, which excludes certain types of variation that will be discussed later.

(3) But since the trait is transmissible, position homology also implies that the alleles are homologous by descent; if they are different, it is therefore that a **mutation** (at least) has occurred in one of the **lines** that links them to their last common ancestor (Figure 1).

(4) Finally, genetic polymorphism can be defined at the scale of the smallest unit composing DNA: the **nucleotide site**. As a result, each nucleotide variant can determine the polymorphism of the higher levels of biological structure - the individual's gene, protein and phenotype - which become scales at which genetic polymorphism can be described (Figure 2).

Take the example of the ABO blood system, which corresponds to antigenic motives{ind-text}Part of the antigen recognized by an antibody or lymphocyte receptor. Also known as an epitope or antigenic determinant. The same antigen can have several epitopes (identical or different) and thus induce a varied immune response.{end-tooltip} involved in the compatibility during blood transfusions. This is a case of genetic polymorphism present in human populations, involving the protein ABO glycosyl transferase{ind-text}Enzyme allowing the transfer of residues containing sugars to proteins. In the case of the ABO system, glycosyl transferase A and B respectively induce an individual to belong to group A or B. When both glycosyltransferases are present, the individual is of group AB.{end-tooltip} which has three alleles, A, B and O. An individual can have (AA), (AO); (BB), (BO); (OO) or (AB) genotypes. Its phenotype will then be [A],[A],[A]; [B],[B]; [O] or [AB], where we see that A and B "dominate" O (*i.e.* their expression masks that of O), and that they are co-dominant among themselves (the heterozygous between A and B has a recognisable expression, [AB]).

A polymorphism can be described first at the scale of the DNA sequence of the locus encoding a protein. Some polymorphisms are "synonymous", *i.e.* they do not change the amino acid sequence of the protein; they are often the most numerous polymorphisms. The others modify amino acids and are called "replacement" polymorphisms. Replacement polymorphisms of ABO glycosyl transferase are of two types: polymorphisms that change the amino acid but not the antigenic unit (they do not interfere with the ABO phenotype of individuals); and other replacement polymorphisms, which determine the ABO phenotype{end-text}Set of observable characteristics of an individual.{end-tooltip}.
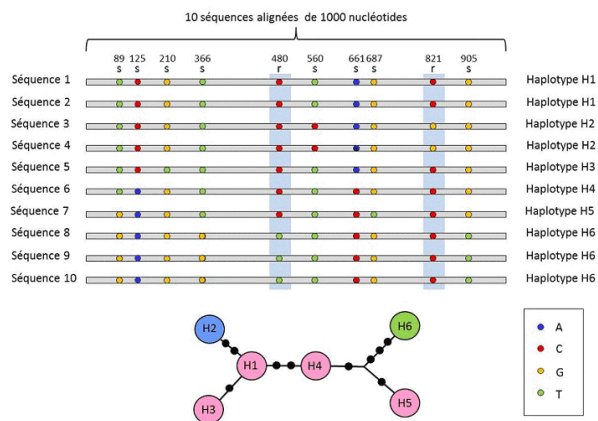
Figure 2. Nucleotide polymorphism and alleles. Example showing the alignment of 10 homologous sequences of 1000 nucleotides encoding a protein. Ten variable sites are color-coded (nucleotides A, C, G, T) with their coordinates on the sequence. They include eight silent sites (s) and two replacement sites (r: amino acid replacement site, highlighted in blue) at positions 480 and 821 (hence three variants of the protein, identified by colour code). There are six different haplotypes (linear arrangements of polymorphic sites, numbered H1 to H6). These haplotypes are grouped according to an unrooted tree (bottom figure) formed by branches connecting the haplotypes together and carrying 10 mutations (black circles) corresponding to the 10 polymorphic sites. The tree is unique because there has been no intra-gene recombination; otherwise there would be as many trees as there are fragments separated by recombination events. In this example, the allelic diversity of the protein is H = 0.42, and the haplotypic diversity H = 0.80; the nucleotide diversity is $\pi = 0.00416$ (definitions and formulas in the text)

Between the nucleotide and allele levels, genetic analyses take into account an intermediate level of description: the linear arrangement of variable sites on the locus. These arrangements, called haplotypes{ind-text}Group of alleles of different loci located on the same chromosome and usually transmitted together. Haplotype is a formed by the contraction of the English phrase haploid genotype, or haploid genotype. All the genes located on the same chromosome and whose alleles segregate together during meiosis constitute a haplotype. These genes are called "genetically linked" {end-tooltip}, are useful in evolutionary genetics because they allow to find the genealogical links between the alleles (Figure 2).

Nucleotide polymorphisms also exist in intergenic regions (outside coding regions). Some affect the areas where gene expression is regulated and therefore have phenotypic expression. Others have no known effect and are referred to as "silent".
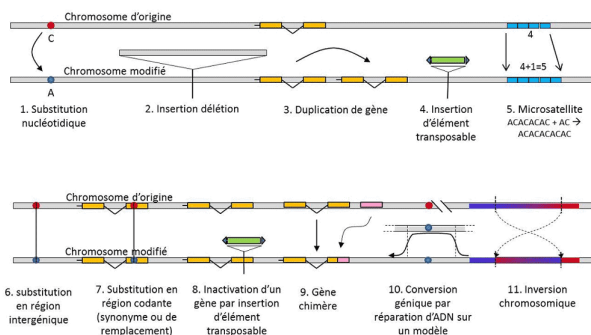

Figure 3. Different types of mutations. 1, nucleotide substitution, affecting a base (A, C, G, or T); 2, insertion or deletion of a DNA fragment; 3, duplication of a gene; 4, insertion of a transposable element; 5, elongation or shortening of a microsatellite locus; 6, substitution in an intergenic region (silent or not); 7, substitution in a coding region (synonymous or amino acid replacement); 8, inactivation of a gene by insertion of a transposable element; 9, creation of a chimeric gene by combination of two coding regions; 10, gene conversion by repair of damaged DNA, using another allele as a model; 11, chromosome inversion.

In addition to nucleotide substitutions, polymorphism for a particular type of mutation, **microsatellites**, refers to variations in the length of an often non-coding DNA fragment (Figure 3) due to a change in the number of repetitions in a short repeated sequence of nucleotides, e.g. CACACACA, or TGTGTG.

Not all genetic variation falls within the definition of genetic polymorphism, since it requires the position homology of

alternatives (Figure 3). Thus, there are repeated sequences obtained by duplicating a gene, for which two repetitions are not homologous per position. Often, we cannot talk about homology for transposable elements either, because they generally change position during cell generations, can multiply and invade the genome. The homology of two transposable elements cannot then be defined. On the other hand, two coding sequences of the same locus, one of which is intact, and the other inactivated by the insertion of a transposable element, correspond well to a homology situation. One day, we will probably consider talking about polymorphism for the modifications of so-called "epigenetic" chromosomes, which are sometimes transmitted over several cell generations, somatic{ind-text}Qualifies non-reproductive cells, or soma. A somatic mutation, which affects a gene of a somatic cell, disappears with the carrier individual.{end-tooltip} or germinal{ind-text}Qualifies gametes. A germline mutation can be transmitted to its offspring.{end-tooltip}, such as methylation. Their importance in the evolution of populations remains to be assessed. The limitation of the notion of genetic polymorphism to certain categories of variation - essentially nucleotide substitutions - comes from the great utility of the latter in exploring the history of natural populations. They lend themselves to mathematical modelling of evolution.

The word **polymorphism** means "several forms". It is opposed to **monomorphism**, which indicates the absence of variation. In the vocabulary of biology, this opposition has long been used, before genetics, to signal the coexistence of several types of individuals in the same species, e.g. caste polymorphism (queen and workers) in social insects; seasonal polymorphism (fur changes) in some Arctic mammals. These cases do not fall under genetic polymorphism, a more recent concept with a narrower and more precise meaning. Continuous variations (e.g. size variation) also do not fall into the category of polymorphism, as they do not present distinct alternatives. However, locus that affect size fall within this definition. We will discuss below the most studied genetic variations because they are the most important for evolution, nucleotide substitutions, and then we will discuss the variation of phenotypes.

2. Measurement

It is not enough to say that a locus is more or less variable without giving a quantitative estimate to this judgment. A researcher can study the variation at different scales. If he is only interested in the alleles of a protein, he will measure the **allelic diversity**, symbolized by '*H*'. If he is interested in DNA diversity, he will measure **nucleotide diversity**, symbolized by 'π'.

The allele diversity *H* is defined as the probability of drawing two different alleles during two sampling with replacement{in-text} Performing a successive draw with delivery of p tokens in an urn containing n tokens, means taking a first token, reading its value, putting it back in the urn, taking a second token, reading its value, putting it back in the urn, etc., until the pth token. This means choosing p objects among n with repetition (you can choose the same object several times) and in order (the order in which you choose the objects is important). The number of successive draws with tokens among n is: n × n × n × ... × n = np.{end-tooltip}. If we call p_i the frequency of the allele of rank i, we show that the probability of picking twice the same allele is $F = \Sigma p_i^2$. The allelic diversity of the sample is then its complement to 1, *i.e.*:

$$H = 1 - \Sigma p_i^2 \ (1)$$

This applies to both protein alleles and haplotypes. This is called **haplotypical diversity**. It can also be called **heterozygosity**, because in the case of a diploid locus, it gives the expected frequency of heterozygotes.

The nucleotide diversity π is the equivalent of *H* at the scale of each nucleotide. It is calculated as the average of the number of nucleotide differences between sequences in a sample taken two by two (δ_{ij}), divided by the length of the DNA fragment into the number of nucleotides *(L)*.

$$\pi = \text{average } _{(\delta_{ij})}/L \ (2) \ (\text{see reference [1]})$$

This value differs between species. In the coding parts of the human genome, two chromosomes randomly selected from the population differ on average by one nucleotide per thousand. In the fruit fly (*Drosophila melanogaster*) this difference is about one in a hundred. The fly is thus ten times more variable than man. When these values are related to the size of the coding regions (about 15,500 genes in *Drosophila*, at least 22,000 in humans), or even genomes (140 million base pairs per haploid genome in flies, about 3,200 million or twenty times more in humans), the number of polymorphic sites is astronomical and results in no one being born of the sexualized generation in any species is, has been and will be genetically identical to another one past, present or future. The information power of DNA polymorphism is immense. In forensics, detectives can identify any suspect from sixteen microsatellite locuses.

3. History

The term **variation** is in the title of the first two chapters of Darwin's book *The Origin of Species* (1859). Darwin introduced this notion into the natural sciences at a time when the laws of biological heredity remained a mystery, and devoted another important book to it, *Variation in Animals and Plants* (1868). Convinced that evolution was about small variations that made little difference to the adaptation of organisms to their living conditions, he attached crucial importance to small quantitative

variations, which prompted his successors (notably Karl Pearson) to found the biometry{ind-text}Science of life measurement. Refers in a very broad sense to the quantitative study of living beings.{end-tooltip}. But the rediscovery of Mendel's laws in 1900 shifted interest to discontinuous variations.
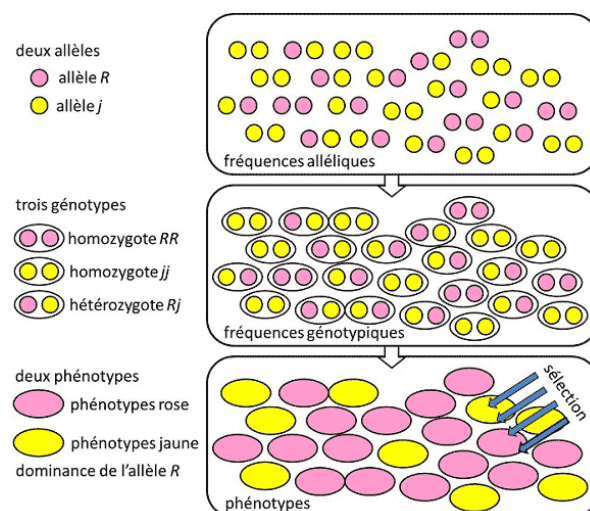

Figure 4. Genes, genotypes and phenotypes: the three structural levels of population genetics are illustrated by a pair of alleles. The allele frequencies are respectively p = 0.40 (Pink allele or R) and q = 0.60 (yellow allele or j), the genotypic frequencies are respectively x = 0.16 (RR homozygotes), y = 0.36 (jj homozygotes) and z = 0.48 (Rj heterozygotes), with p+q = 1 and x+y+z = 1. In this example, the values used for the genotypic frequencies x, y and z are the theoretical values predicted in case of random association of the gametes: these are the so-called Hardy-Weinberg proportions, where the homozygous frequencies are respectively p2 and q2, and that of the heterozygotes of 2pq.

From 1908 to 1930, population genetics was essentially a theoretical discipline seeking to reconcile Darwinism with Mendelian heredity{ind-text}Heredity based on the transmission of a single gene under a dominant, recessive or sex chromosome-linked mode X (or Y). Refers to a genetic trait with simple determinism, by a couple or a small number of couples of genes.{end-tooltip}, and where probabilities played a major role. Genetics is counter-intuitive. It predicts that there is no **reproduction** of the parent in his descendant, because the latter's genotype results from the lottery of the separation of the alleles before the fusion of the two parental half-genomes. It was realized that these are the allelic frequencies{ind-text}Frequency at which the allele of a variant is found in a population. Expressed as a proportion or percentage. The sum of the allele frequencies of all alleles of a gene in a population is therefore by definition equal to 1. In population genetics, allele frequencies represent genetic diversity at the population or species level.{end-tooltip} that are transmitted between generations, not genotypes or phenotypes. These frequencies are more or less stable from one generation to the next and **produce**, beyond the multiple entanglement of filiations between individuals, the same genotypic frequencies{ind-text}Genetic structure of the population. Determined from the allelic frequencies {end-tooltip} (Figure 4). The distribution of population genotypes, associated with variance, is therefore the only predictable factor. Around 1930, the three theoreticians Ronald Fisher, JBS Haldane and Sewall Wright helped to create the notion of the Mendelian population{ind-text}Population whose heredity obeyed Mendel's laws.{end-tooltip} [2]. In this context, evolution necessarily associates three structural levels: the gene, the individual and the population. This three-way association can be summarized by the following formula: population evolution is a change in allele frequency (gene scale, but also group scale), a change that depends on the sorting of phenotypes by selection (individual scale). Empirical population genetics research then developed, but was limited for a long time by our ignorance of how chromosomes work (the structure of DNA was understood in 1953 and its sequencing began in 1977), and had to fall back on the study of the few visible polymorphisms, such as the coloration of ladybug's elytra or the shell of snails (see Focus The great snail debate). There was much debate among geneticists as to whether natural populations were generally polymorphic or monomorphic, and whether polymorphism was in itself beneficial or not. The study of macromolecular scale variation had to wait until 1966 (for proteins) and 1983 (for DNA) before answering these questions.

4. Quantitative variations and heritability

When we look around us, we see physical differences between beings, some complex like facial features, others easy to measure like weight or height. Common sense suggests that they are partly transmissible, although in a way that is difficult to specify. These traits are often polygenic{ind-text}under the dependence of many genes. We are talking about polygenic heredity. Diabetes is a polygenic disease. {end-tooltip}, which means that they are influenced by many locuses. Developmental genetics,

and its application to the comparison of species (**evo-devo**, for "*evolution of development*") unravel the web of complex interactions that make that during development the cells of the body, all with identical genetic baggage, differentiate themselves by epigenesis to give different tissues. It is still unclear how complex characters are built. Measurable phenotypes can be studied using **biometrics**, based on the properties of statistical distributions. One value with useful properties is the variance{ind-text}In a sample of individuals in which a given trait was measured, the variance is the difference between the mean of the square of values and the square of the mean of the values. This measure, which is always positive, indicates the dispersion of individuals. {end-tooltip} [3]. When several independent causes determine the variation of a line, their variances are additive and their sum gives the variance of the line. If they are not independent, the sum of the covariances is added to the sum of the variances. The total variance V_t of a phenotypic trait is such that :

$$V_t = V_{ga} + V_{gd} + V_{gd} + V_{gi} + V_e$$

V_{ga}, additive genetic variance, is the sum of the variances due to each locus independently of the others; V_{gd}, dominant genetic variance, results from the interaction of the two alleles of the same locus; V_{gi}, interaction variance between locuses, results from the fact that the locuses of the same individual interact; V_e, variance of the environment, is assumed here independent of genetic variance.

The genetic variance of dominance is common to two children of the same parents, because they share dominance effects when at a given locus they have received the same alleles from both parents, which happens to them at one in four locus. They do not share these effects with their parents, so they are more "similar" to each other than they are to their parents, despite the fact that they share so many genes with each other. Certainly, a parent may, by chance, have received the same two alleles as one of his children at a given locus. This chance depends on the frequency of alleles in the population. We see that the degree of similarity between a child and his or her parent for a given trait is a public property.

More generally, all components of genetic variance $(V_{ga} + V_{gd} + V_{gi})$ can vary between populations. In breeding lines, a common case is inbreeding, which exhausts genetic variance: as a result, a trait may have the same average in two populations, but have a different genetic variation. The ability of a breeder to select a trait is therefore a characteristic of the line, not of the trait.

The relationship between parents and children is measured by **heritability** h^2. It is defined as the ratio of additive genetic variance to total variance [4]:
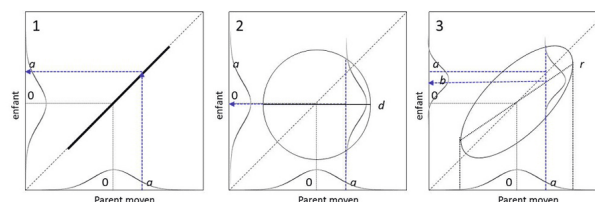
$$h^2 = V_{ga}/V_t$$

Figure 5. Parent-child relationship for a quantitative characteristic: Different theoretical cases of size correlation between parent-mean and child are shown: (1) Correlation of 1, where the children's height is exactly that of the parents, with no difference in mean or variance. Marginal distributions are that of the size of average parents and children. The coordinates of the points (parents, children) are located on the bisector of the coordinate axes. A valuable average parent to generate a valuable child a.
(2) Correlation of 0, where the size of children is independent of that of parents. A valuable average parent has produced a child of indifferent value, within the distribution of children. The coordinates of the points are located on a circle; the expectation of their value is located on a line parallel to the parents' axis (d). (3) Intermediate parent-child correlation between 0 and 1: the coordinates of the points are located on an ellipse. The expectation of their value is no longer located on the bisector of the coordinate axes (which is also the major axis of the distribution ellipse), but on the regression line r. An average parent of value has produced a child of value b (between 0 and a) with a certain variance. It is noted that case (3) is intermediate between cases (1) and (2)

It is shown under various assumptions in Figure 5. Heritability is the ability to determine whether a feature can be selected. Figure 5-3 shows that a breeder will be able, within a crossing generation, to move the average of a character from the value 0 to the value b if he selects individuals with value a. We can show that $b = h^2 a$. To select a trait, it is useful to increase the heritability ratio by standardizing the environment (thus reducing environmental variance), and by placing oneself in the

environmental conditions that favour the appearance of the trait. Geneticist Waddington has shown that, in nature, a change of environment can thus reveal new traits that will then be selected during evolution.

References and notes

Cover image. Snail shells *Cepaea nemoralis & Cepaea hortensis.* André Künzelmann, UFZ.

[1] This formula can be written in a more conventional way as: [latex] \pi =\frac{n(n-1))}{2L}\sum _{i=1}^{n-1}\sum_{j=i+1}^{n}\delta ij[/latex]

[2] Fisher R.A. (1930) *The Genetical Theory of Natural Selection*, Clarendon Press, Oxford; Haldane J.B.S. (1932) *The causes of evolution*. London: Harper & Brothers; Wright S. Evolution *and the Genetics of Populations.* 1, 2, 3, 4; New Edition. University of Chicago Press. 1984 (reprint by the author of his main results).

[3] Being a sample of individuals in whom the studied trait was measured, variance can be defined as the difference between the mean of the square of values and the square of the mean of the values. This measure, which is always positive, indicates the dispersion of individuals.

[4] This is heritability in the strict sense of the term. Heritability in a broad sense is the ratio of the sum of genetic variances to total variance. This formula gives heritability with both parents (the "average parent"); the estimate with only one parent would give $h^2/2$.

L'Encyclopédie de l'environnement est publiée par l'Université Grenoble Alpes - www.univ-grenoble-alpes.fr

Pour citer cet article: **Auteur :** VEUILLE Michel (2025), Genetic polymorphism and variation, Encyclopédie de l'Environnement, [en ligne ISSN 2555-0950] url : http://www.encyclopedie-environnement.org/?p=6896