Basic concepts and terminology

Mutation is the ultimate source of organismal variation. Every newborn human carries several dozen unique mutations in his/her genome. It is thus often brought forward that '*we are all mutants*'.

Mutation: 1) a process that produces a novel allele differing from existing version(s) (or the wild type)
2) the allele that results from a mutational process

Mutant: organism or cell who has received a mutation; often also: whose changed phenotype is attributed to a mutation

Mutations happen somewhere at a **locus** in the genome. A locus is an operational definition of a specific position in the genome and can refer to a single base pair (say position 12,331,156 on chromosome 1 of the human assembly version GRCh38.p8), a stretch of sequence or a specific gene.

Mutations generate novel **alleles**, different character states of the locus. **Haploid** organisms have only one copy of each locus. While at the population-level several alleles can exist, haploid individuals are limited to one allele. **Diploid** organisms inherit one **maternal and one paternal** copy and can thus carry a total of two alleles at a single locus. The combination of alleles is referred to as the **genotype**: **heterozygous** individuals have two different alleles (e.g. Aa), **homozygous** individuals two copies of the same allele (e.g. AA, aa). It is important to note that in diploids (and n-ploids more generally) mutations always first appear in a heterozygous state. Co-inherited allelic combinations of more than one locus are called **haplotypes**. A specific set of haplotype combinations in a diploid are called **diplotypes**.

Mutations can occur in **somatic cells** or in the **germ line**. Somatic mutations can be responsible for **mosaicism** where two or more populations of cells with different genotypes are present in a single individual. While somatic mutations are important for the fate of an individual (e.g. cancer), mutations affecting the germ line are in the center of attention for evolutionary genetics studying **trans-generational processes**. When **genotyping** individuals to assess genetic variation in natural population it is generally assumed that only germ line mutations are quantified. However, if somatic mutations arise at an early stage of development they will affect most cells of the organism and misleadingly suggest germ line mutations.

In contrast to early believes of the saltationists who emphasized the importance of discontinuous, drastic variation ('evolutionary sports') in the generation of novel species, we now know that mutations happen at the level of the individual. Mutation thus introduces genetic **polymorphism** (occurrence of several allelic states) into the population that **segregates** until the novel allele is lost or has reached fixation. The current frequency of an allele in a population is the **allele frequency**; the combination of alleles in (diploid) individuals is called **genotype frequency**. The frequency of each allele can be calculated from the genotype frequency.

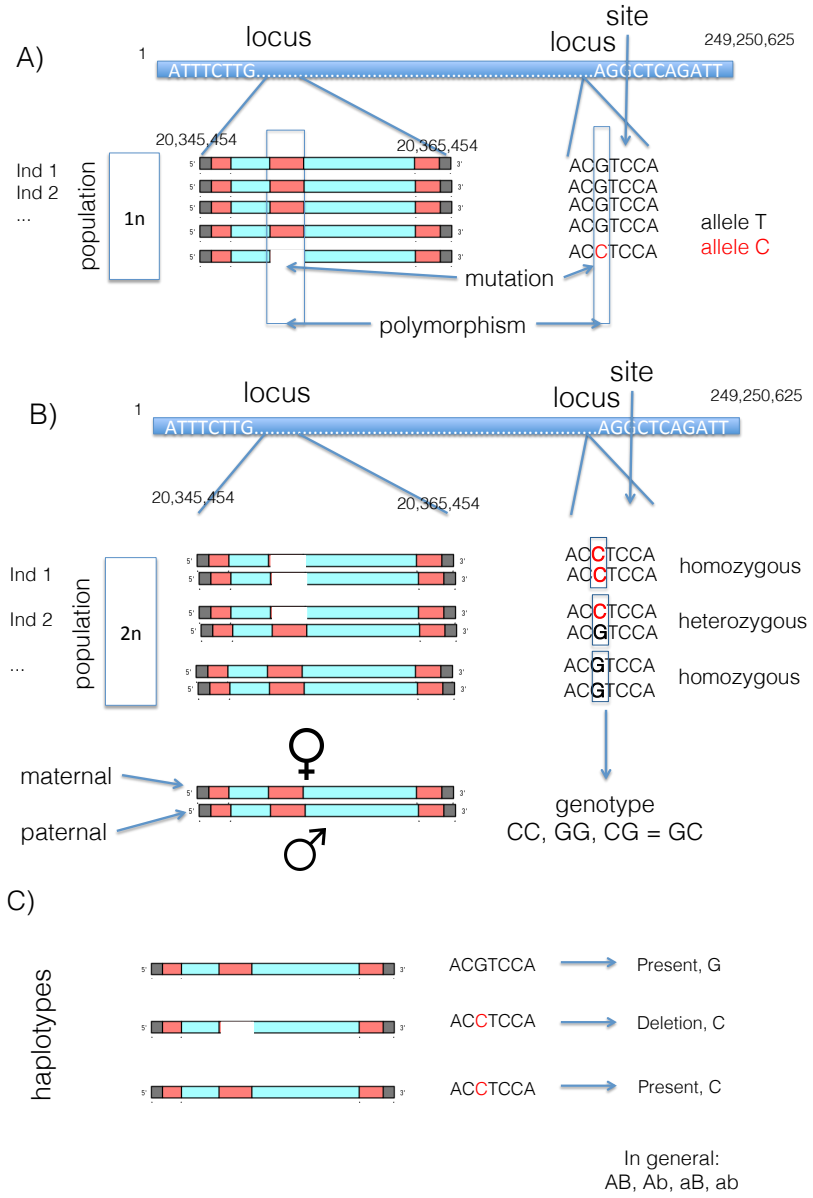

Figure 1: Basic terminology illustrated for A) haploid and B) diploid organisms including the definition of C) haplotypes.

So far, we have only considered the genetic component of mutations. Until recently, mutations could not be studied at the molecular level, but only through their effect on the **phenotype**, the physical or biochemical expression of the genotype. Variation of phenotypic traits can be classified in roughly two categories: discrete and continuous (or **quantitative** variation). In the case of discrete variation, phenotypic differences between individuals can be assigned to a small number of classes (e.g. albinism vs. normal pigmentation) inherited by a single or a few genes. In the case of continuous variation, phenotypes are measured on a metric scale (e.g. body size, IQ) and are inherited by multiple genes. We now know that quantitative differences are also inherited to a large degree. DNA sequence alone, and the polymorphism that goes with it, has no direct implications for evolution. It is the phenotype that interacts with the environment, collects energy, secures reproduction, etc., and is subject to **selection**. The component of the phenotype that is not environmentally determined

(**phenotypic plasticity**) is encoded in the DNA sequence. The phenotype is constituted by its transcription into mRNA, rRNA and other functional RNA sequences followed by translation into proteins, which in turn interact with each other and various RNA molecules. This order of information transfer is generally known as the **central dogma** and establishes the **link between the genotype and the phenotype.** The nature of this link can be described by the **genetic architecture** of the trait. Traits can be encoded by many genes (**poly- or multigenic** trait) or by a single gene (**monogenic** trait). The allelic effect of a gene on a given phenotypic trait is given by its dominance effect, which if often conceptualized by three discrete classes: alleles are said to be **dominant, co-dominant or recessive**.

Classes of mutations

Mutations occur as a by-product either during replication or under the action of a mutagenic agent, such as a chemical or radiation. Several classes of mutations can be distinguished, all of which can – but need not - have severe effects on the phenotype.

Gene mutations

Point mutations: mutation of one base into another. According to the sequence position at which a mutation occurs, one can distinguish between mutations in **non-coding** regions and **protein coding** parts. The protein coding part can be broken down into triplets of nucleotides called **codons**, which determine a single amino acid. The 3rd position of a codon is more variable than the 1st and the 2nd. As proposed by Francis H. C. Crick the 5' base on the tRNA anticodon is not as spatially confined as the other two bases, and could thus have non-standard base pairing during translation. Because of these relaxed conditions the 3rd base pair is usually a **synonymous site**, whereas the 1st and the 2nd sites are **non-synonymous**. However, there are two codons which have no tolerance for nucleotide substitutions: UGG coding for tryptophan and AUG coding for methionine exclusively exhibit non-synonymous sites. **Silent** mutations do not affect the amino acid sequence of a protein (e.g. TCT (Ser) > TCC (Ser)). They may occur in non-coding parts or at synonymous sites. Mutations that change a single amino acid are called **missense** mutations (e.g. TCT (Ser) > GCT (Ala)), mutations resulting in a premature stop codon **nonsense** mutations (e.g . UGC > UGA).

Frameshift mutations: insertion or deletion of a single base or a short stretch of nucleotides that shifts the triplet's reading frame. Always results in a drastically different polypeptide product often introducing stop codons. The gene product is usually nonfunctional.

Chromosomal (or structural) mutations

Duplications/deletions : "**in-dels**" (insertions/deletions) are possible also on larger scales due to a number of mechanisms, such as slippage (e.g. in microsatellites…ATATATAT… > …ATATAT…), **transposable element**s ("jumping genes" that can copy itself from one site in the DNA to another) and unequal crossing-over. Transposable elements promote many structural mutations due to their sequence similarity introducing errors during meiosis. Entire genes can also be duplicated leading to the generation of **paraloguous genes** if maintained over long evolutionary time (e.g. hemoglobin). Gene duplication is thought to be an important processes for the generation of evolutionary novelty, as one gene copy can secure proper function, while the other is free to 'experiment'.

Inversions: reversal of DNA order. Inversions spanning the centromere are called **pericentric**, inversions restricted to one chromosomal arm are called **paracentric.** Both types often produce errors during meiosis.

Translocations : exchange of segments among non-homologous chromosomes.

Fissions/Fusions: one chromosome becomes two or two become one.

Ploidy: change in the number of whole sets of chromosomes (**aneuploidy**, e.g. Trisomy 21 in humans) or even all chromosomes of a genome (**polyploidization** in maize).

Mutation is random

It is a Darwinian principle that generation of variation through **mutation is random** and only selection gives a direction to evolution: mutation provides the raw material and selection chooses from it. The Lamarckian view that mutation and selection are connected has been proven to be incorrect.

" An ideal situation would be if the organism were to respond to the challenge of the changing environment by producing only beneficial mutations where and when needed. But nature has not been kind enough to endow its creations with such a providential ability" *Theodosius Dobzhansky (1951) Genetics and the Origin of Species, 3rd edn, Columbia University Press.*

More precisely: The probability for a mutation having a certain effect on a given phenotype does not depend on whether this effect is beneficial or not. In particular, this means that useful mutations are not produced by exposure to an environment where these mutations are useful (classical experiment: Luria-Delbrück 1943, Lederberg 1952). But it does not mean that the probability for a mutation to be **beneficial** is the same as for it being **deleterious**. In fact, the overwhelming majority of (non-neutral) mutations are deleterious. The reason is the history of previous adaptations has driven the genes to a near optimum that is unlikely to be improved by random change.

Note that epigenetic modifications may possess elements of both random change and directed mutations responding to environmental stimuli.

Rates of mutations

The **mutation rate** (often denoted as **μ**) is a central parameter in evolutionary genetics, as it determines the rate by which novel variation is introduced into a population. Mutation is a **random process** that is often approximated by a **Poisson distribution**.

Assuming a stable mutation rate through time for a group of organisms, mutations can be used like a '**molecular clock**' to estimate the time of divergence between evolutionary lineages. The molecular clock will be introduced later, but note at this points that we can only observe a small number of mutations that have gone to fixation between lineages

(**substitutions ≠ mutations**).

To model the evolution of DNA sequence as a function of mutation, (strong) selection and genetic drift, Motoo Kimura introduced the **infinite sites model**. This model makes the simplifying assumptions that (1) there are an infinite number of sites where mutations can occur, and that (2) every new mutation occurs at a novel site. Under natural conditions and depending on the organism the mutation rate is generally low enough that a locus / site is only affected once by a mutation until it has reached fixation. On a per nucleotide basis they range from 10^{-4} per generation in RNA viruses that lack proof reading mechanisms to 10^{-10} in bacteria and nematodes, with a current estimate in humans of about .3-1 10^{-9} mutations site^{-1} year $^{-1}$ – 0.7-2 10^{-8} mutations site^{-1} generation $^{-1}$or assuming a generation time of 20 years. The large spread of the values is due to different data, models and methods applied; recent estimates tend to home in on lower values. Overall, given generally small mutation rates the assumption of the infinite sites model is generally justified for most organisms. If we consider long time spans in the millions of years homologous sites may experience multiple mutations which can misleadingly suggesting common ancestry (**homoplasy**). Multiple hits including back mutations can be accommodated in nuclear **substitutions models**.

Estimation of the mutation rate
The measurement of mutation rates is difficult. It can be estimated from the rate at which new mutations are detected on the **phenotypic level**. However, there are several problems with this approach:

a) Only a small part of mutations may have a visible effect on the phenotype (even if mutations are not neutral).

b) Bearers of deleterious mutations may die before they can be detected.

c) Depending on the genetic architecture of a trait several mutations may be necessary to result in a phenotypic effect.

Molecular methods allow a more direct assessment of mutation rates on the nucleotide and on the gene level. Rates vary among species, among genes, and even within genes.

a) **Pedigree-based approaches** constitute the shortest timescale at which mutations can be measured. By scoring the alleles of the parents and the offspring, novel mutations that have arisen in the parents' germ line can be directly scored.

b) **Mutation accumulation lines** assess mutations over somewhat larger timescales. New mutations are scored after several generations (generally of bacteria or other microorganisms) have been proliferated in the lab. There is a risk of underestimating mutations that are deleterious to their carrier. Experiments have to be designed to minimize the effect of selection.

c) **Comparative genetic approaches** are an indirect way of measuring mutation rates averaged over many generations by simply counting the number of (base pair) differences between homologous genes relative to the number of generations that have elapsed. This method depends on the **Neutral Theory of Evolution** specifying the per generation mutation

rate equals the proportion of base pairs that differ between two species, divided by twice the number of generations since their common ancestor. This assumption only holds for **selectively neutral** sites, and often it is difficult to estimate the time to the most recent common ancestor (TMRCA) as well as the generation time.

Of importance to either of the methods: since there are a larger number of cell divisions involved in the male sperm line than in the female germ line, more new mutations are transmitted through sperm then through eggs (**male-biased mutation rate**). In male heterogametic species (males: XY) this has the interesting consequence that the X-chromosome is exposed to fewer mutations than the autosomes, in female heterogametic species (females: WZ) it has the opposite effect. The reason for this is that genes on autosomes spend an equal amount of their time in males and females, so that their net mutation rate is the average of the male and female mutation rates. In humans and other male heterogametic species, X-linked genes spend only one-third of their time in males and two-thirds of their time in females. If spermatogenesis is more mutagenic than oogenesis, the X chromosome is thus subjected to a lower mutation rate than the autosomes or the Y chromosome. The reverse is true for Z-linked genes in taxa with female **heterogamety**.

Literature: (Barton et al. 2007; Futuyma 2013)

Barton NH, Briggs DEG, Eisen JA, Goldstein DB, Patel NH. 2007. Evolution. 1st edition. Cold Spring Harbor, N.Y: Cold Spring Harbor Laboratory Press. Chapter 12.

Futuyma DJ. 2013. Evolution. 3rd ed. Sinauer Associates. Chapter 8.