Polymorphisms of nucleic acids

Nucleic acids (DNA and RNA) are molecules bearing information in the nucleotide sequences and as a whole determine the species. In humans there are about $3x10^9$ nucleotides separated in strings determining **chromosomes**. Each chromosome is composed of different number of nucleotides and homologous chromosome share the highest similarity rate in between them. Polymorphisms can be understood as differences in nucleotide sequences and can have different extent.

Polymorphism is distinguished from mutation just based on its frequency. The mostly accepted threshold is 1%. If there is a variation in DNA sequence at a specific site and if its frequency at a specified population is higher than 1%, one considers it as a polymorphism. Everything with lower presence in population is considered as mutation.

The most common type of polymorphism is called „**single nucleotide polymorphism**“ (or SNP, sometimes read as SNiP) and represents difference of a single base (A,C,T,G) at a defined position of the genome. SNPs are very frequent, their distribution is not homogeneous. They appear more frequently in non-coding regions of the genome. SNPs are also critical for personalized medicine.

Another type of inter-individual variability in DNA sequence is represented different number of a sequence motif repetition. There are several types of repetitions categorized based on length of the motif to be repeated and the distance in between the motif.

SINE and LINE (non-tandemly repeated motifs)

SINE

SINE (https://en.wikipedia.org/wiki/Short_interspersed_nuclear_elements_(SINEs)) stands for „short interspearsed nucleotide elements“. Short in this case stands for motifs 100-700bp long that are present in many copies throughout the genome, but not tandemly next to each other. For example, **Alu element**, 300bp long, is present in about one million copies throughout the human genome (10% of the total nucleotide composition!). SINE represent non-coding elements. The RNA coded by the short-interspersed nuclear element does not code for any protein and recent studies show, that SINE have a strong regulatory influence on gene expression. Aberrantly increased SINE transcription is associated with some human diseases.

LINE

LINE stands for “long interspersed nucleotide elements” and represents motifs around 6-7kb long that are present in around 100 000 copies throughout the genome (almost 20% of genetic material in human nucleus). They represent retrotransposons, elements that are capable of self-recopying in the genome.

Tandemly repeated motifs

A defined region of DNA is tandemly repeated and forms blocks. Number of repetition varies in a population and follows the Mendelian rules of inheritance. We can subdivide this category into several classes based on length of the motif

- **Satellites** -20bp-and longer motifs, frequently found in the region of centromere with up to 50 000 repetitions
- **Minisatellites** – 10-20bp long motif with about 1000 repetitions
- **Microsatellites** – 1-10bp long motif with up to 100 repetitions, the most common type of human polymorphism in the number of repetitions, there are hundreds of those sequences throughout the genome (in coding and non-coding regions) and are widely used as markers in molecular diagnostics. Sometimes they are referred as STR which stands for “short tandem repeats”.

Copy number variation

Copy number variation (CNV) is another perspective how to categorize variation present in genomes. It represents segments of DNA that are variably duplicated or deleted in a population. A typical example is a variant number of duplication of AMY1 gene coding for alpha amylase. CNVs are frequently found in subtelomeric and pericentromeric regions residing LINE elements.All DNA polymorphisms present in population must be somehow evolutionary silent or advantageous in order not to be selected.