**COMPLEX TRAIT GENOMIC ANALYSIS**

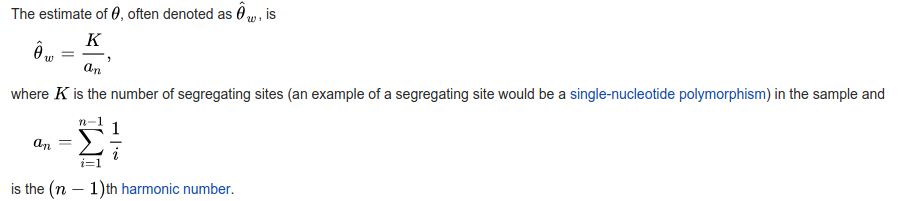
**Population Genomics Basics**

**Metrics of nucleotide diversity**

They require sequence data, because variability is expressed per DNA length. SNP arrays does not allow to infer variability because we do not know the length of DNA and because SNP ascertainment bias. But normally heterozygosity rate in an array and DNA variability are correlated.

There are numerous metrics for DNA variability. The two main ones are Watterson’s estimate () and Tajima’s diversity ().

From Wikipedia



And Tajima’s diversity estimate is the average number of allele differences between all pairs of individuals. In all cases, the estimator must be divided by the number of base pairs analyzed.

Under a neutral model (i.e., only drift, panmixia and constant effective size), both estimators should be the same.

**Toy example:**

Here is a list of 7 aligned sequences

AAATTTTCCGGCA

.............

..T.........G

.............

......A......

.........C...

..T..........

* How many nucleotides?
* Compute  and  diversity

**Real example**

The attached file MC1R\_pigs.fasta contains a list of pig MC1R gene sequences (Fang et al. 2018, Contrasting Mode of Evolution at a Coat Color Locus in Wild and Domestic Pigs, <https://doi.org/10.1371/journal.pgen.1000341>).

A very popular software to analyze sequence data in a small scale is DNASp (<http://www.ub.edu/dnasp/>).

Install the program and upload fasta file.

Compute diversity estimates for the whole set of samples and separately by continent (Europe vs. Asia).

Compare both Tajima's and Watterson's estimators.