**BACKGROUND INFO**

Maybe include some background info about population genetics?

Analysis of genetic diversity allows us to describe populations, selection and some aspects of epidemiology. We often use the number of SNPs and INDELs as a way of determining how diverse individuals are from each other.

**Understand a VCF file**

We will next use the VCF file to examine the genomic diversity of the population. Remember that each row of a VCF file describes one genetic variant in the population, and the columns describe the allele calls in each individual.

What information is present [Variant Call Format - Wikipedia](https://en.wikipedia.org/wiki/Variant_Call_Format)

Uploading the VCF files and obtaining the information?

**Building a database**

**Obtaining info**

* Genomic coordinate
* SNP name 🡪 rs value
* Gene name
* Genomic position
* Genotype frequency
* Allele frequency

Use chr 22? Because it’s the smallest?

Whats the difference between allele frequency and genotype frequency?

**Populations**

* General from ensemble like Europeans, americans etc
* Five populations from International genome (maybe each from different continents?)
* Stats 🡪 within a population and between different populations

**Website**

1. Building forms 🡪 how to put info and get info out, searchbar?

Retrieving SNP information

1. Different pages
2. Homepage
3. Others?
4. Styling part 🡪 CCS etc

**Statistics**

**On python:**

1. **Genetic variation**

* **F-statistics** 🡪 Describes the statistically expected level of heterozygosity in a population; more specifically the expected degree of (usually) a reduction in heterozygosity when compared to Hardy–Weinberg expectation

##################################################################################

**allel.weir\_cockerham\_fst(g, subpops, max\_allele=None, blen=None)**

Compute the variance components from the analyses of variance of allele frequencies according to Weir and Cockerham (1984).

**allel.hudson\_fst(ac1, ac2, fill=nan)**

Calculate the numerator and denominator for Fst estimation using the method of Hudson (1992) elaborated by Bhatia et al. (2013).

Etc [F-statistics — scikit-allel 1.3.3 documentation](https://scikit-allel.readthedocs.io/en/stable/stats/fst.html#allel.weir_cockerham_fst)

##################################################################################

* **Hardy-Weinberg equilibrium** 🡪 The Hardy-Weinberg equilibrium is a principle stating that the genetic variation in a population will remain constant from one generation to the next in the absence of disturbing factors.

**allel.heterozygosity\_observed(g, fill=nan)**

Calculate the rate of observed heterozygosity for each variant.

**allel.heterozygosity\_expected(af, ploidy, fill=nan)**

Calculate the expected rate of heterozygosity for each variant under Hardy-Weinberg equilibrium.

Etc [Hardy-Weinberg equilibrium — scikit-allel 1.3.3 documentation](https://scikit-allel.readthedocs.io/en/stable/stats/hw.html)

##################################################################################

1. **Haploid diversity**

[Selection — scikit-allel 1.3.3 documentation](https://scikit-allel.readthedocs.io/en/stable/stats/selection.html)

Selection

* Integrated haplotype score (IHS)
* Cross-population extended haplotype homozygosity (XPEHH)
* Haplotype diversity, Garud’s H statistics

1. **Test against neutrality?**

* Tajima’D

Others tests/methods

1. Principal components analysis

###########################################################################

**allel.pca(gn, n\_components=10, copy=True, scaler='patterson', ploidy=2)**

Perform principal components analysis of genotype data, via singular value decomposition.

**Notes**

Genotype data should be filtered prior to using this function to remove variants in linkage disequilibrium.

###########################################################################

1. Pairwise distance and ordination

Pairwise distance 🡪

**allel.pairwise\_distance(x, metric, chunked=False, blen=None)**

Compute pairwise distance between individuals (e.g., samples or haplotypes)

###########################################################################

Graphs/plots 🡪 plotly?

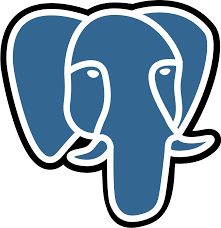
1. Miscellanea? [Miscellanea — scikit-allel 1.3.3 documentation](https://scikit-allel.readthedocs.io/en/stable/stats/misc.html)
2. [Applied Population Genetics (dyerlab.github.io)](https://dyerlab.github.io/applied_population_genetics/population-graphs.html) 🡪 R
3. PopSc package
4. Ploting PCAs?
5. Boxplots, histograms/bar chars

**Temp architecture plan**



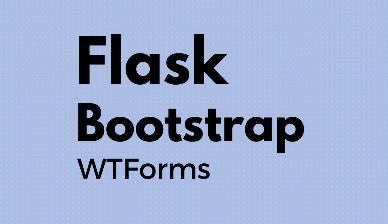
Client

Front-End



Back-End

Database



GET/POST

