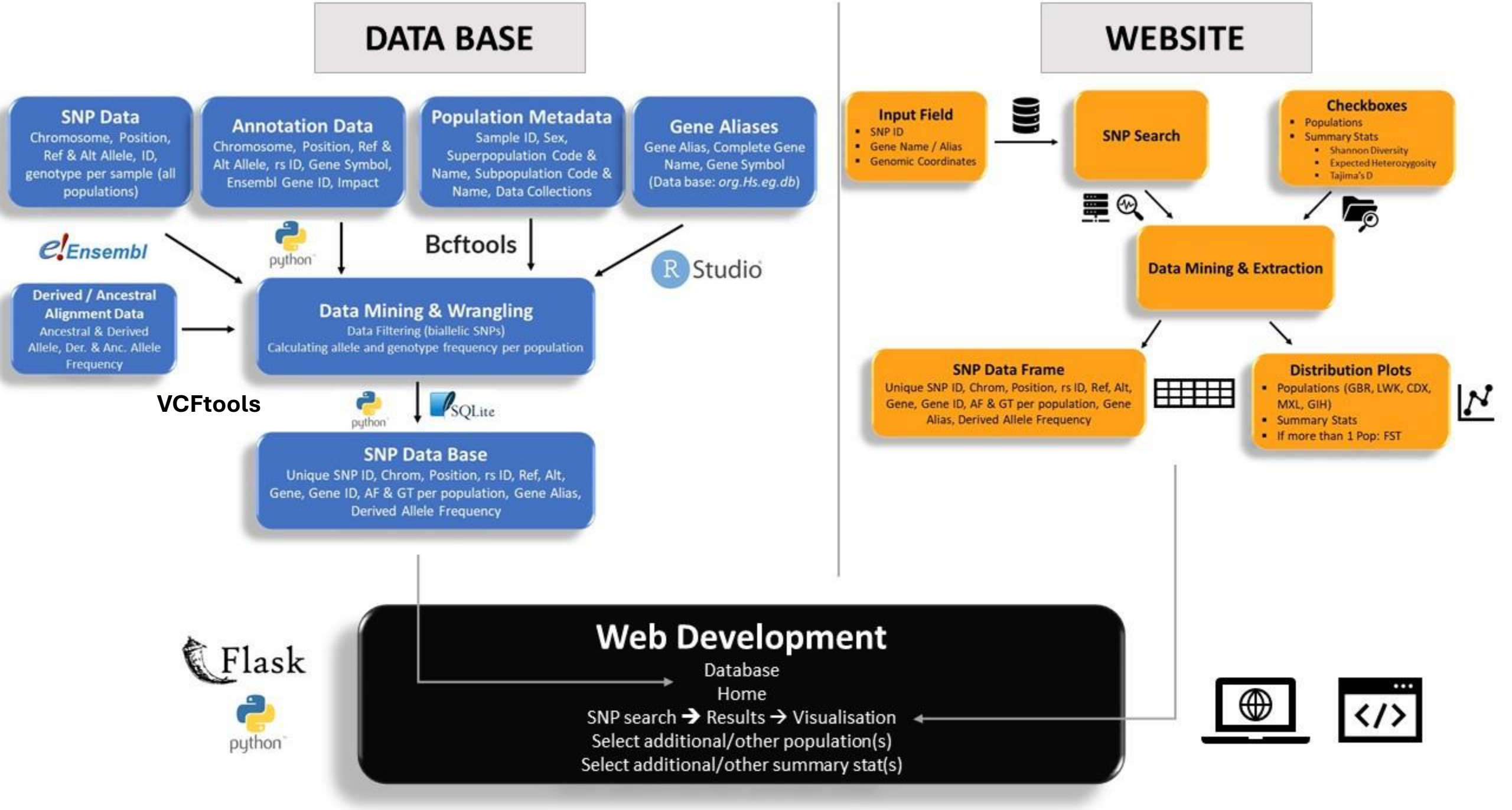
BIO727P GROUP SOFTWARE PROJECT

TEAM CELINE

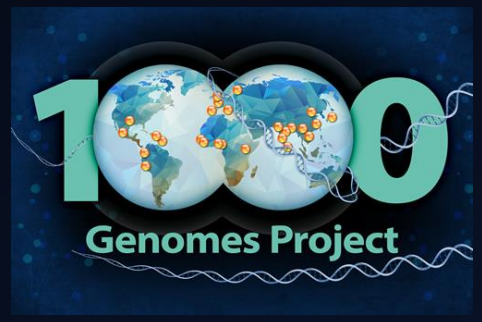
CELINE, AMANAH & GRACIA

04.03.2022

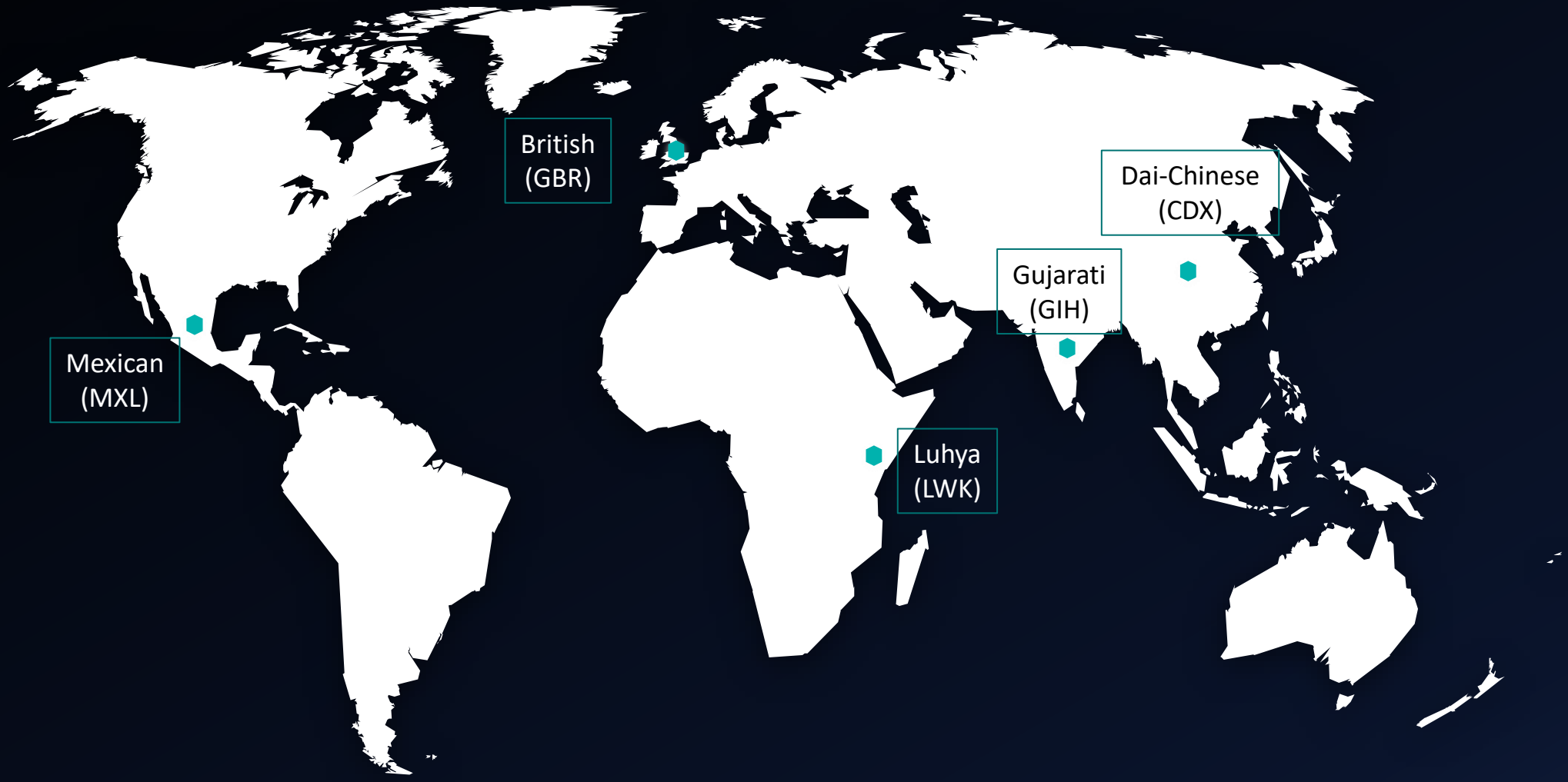

DATA BASE
SNP Data
Chromosome, Position, Ref & Alt Allele, ID, genotype per sample (all populations)
Annotation Data
Chromosome, Position, Ref & Alt Allele, rs ID, Gene Symbol, Ensembl Gene ID, Impact
Population Metadata
Sample ID, Sex, Superpopulation Code & Name, Subpopulation Code & Name, Data Collections
Gene Aliases
Gene Alias, Complete Gene Name, Gene Symbol (Data base: org.Hs.eg.db)
e!Ensembl
python
Bcftools
R Studio
Derived / Ancestral Alignment Data
Ancestral & Derived Allele, Der. & Anc. Allele Frequency
Data Mining & Wrangling
Data Filtering (biallelic SNPs)
Calculating allele and genotype frequency per population
VCFtools
python
SQLite
SNP Data Base
Unique SNP ID, Chrom, Position, rs ID, Ref, Alt, Gene, Gene ID, AF & GT per population, Gene Alias, Derived Allele Frequency
WEBSITE
Input Field
SNP ID
Gene Name / Alias
Genomic Coordinates
SNP Search
Checkboxes
Populations
Summary Stats
Shannon Diversity
Expected Heterozygosity
Tajima's D
Data Mining & Extraction
SNP Data Frame
Unique SNP ID, Chrom, Position, rs ID, Ref, Alt, Gene, Gene ID, AF & GT per population, Gene Alias, Derived Allele Frequency
Distribution Plots
Populations (GBR, LWK, CDX, MXL, GIH)
Summary Stats
If more than 1 Pop: FST
Flask
python
Web Development
Database
Home
SNP search → Results → Visualisation
Select additional/other population(s)
Select additional/other summary stat(s)

Data Collection

- SNP VCF files from Ensembl FTP Server
- Population Sample Data from 1000 Genomes Project
- Gene Alias from `Org.Hs.eg.db` - R Bioconductor package containing genome wide annotation for Human genome



Population Selection



Walk-through for allele and genotype frequency processing

1. Convert numbers to letters

2. Calculate allele counts, frequencies and genotype frequencies per population using *genetics::genotype()* and *summary()* in **R**

3. Frequencies joined to SQL database (see Table 2)

Table 1:

Ref	Alt	Samples (numeric)	Samples (base)
A	G	0|0	A|A
T	C	0|1	T|C
A	T	1|1	T|T

Table 2:

GBR_AF_ref	GBR_AF_alt	GBR_GT 00	GBR_GT 0110	GBR_GT 11
1	0	1	0	0
1	0	1	0	0
1	0	1	0	0

Walk-through for derived allele frequency processing

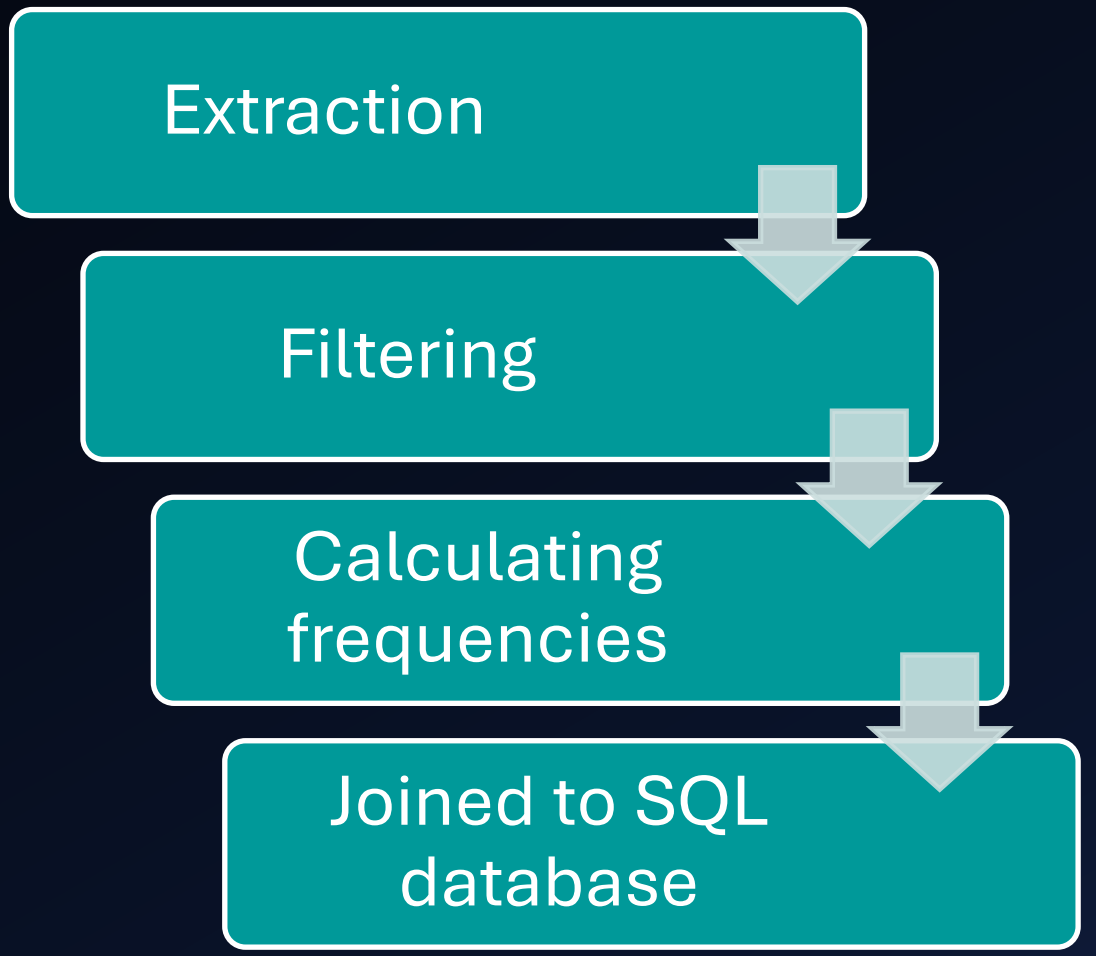
1. Extract position and ancestral allele field from VCF

2. Remove SNPs with missing ancestral allele by position

3. Calculated derived / ancestral allele frequencies

4. Join ancestral and derived alleles to database

Justification: determine the occurrence of a mutated/derived allele in the human population which arose after recent divergence from outgroup



SNP Database

SNP_Data_Table	
Unique_SNP_ID	Gene
Chromosome	Gene ID
Position	AF & AC per population
rsID	GT per population
Ref	Gene Alias
Alt	Derived Allele Frequency

Example of Unique SNP ID:
22 : 50807605 : C : A

Features of Software

- SNP search
  - Browse for SNPs via rs ID, Gene Name (or Alias), or Genomic Coordinates
- Summary statistics selection
  - Shannon Diversity, Expected Heterozygosity and Tajima's D
  - FST Analysis
- Population selection
  - Choose one or more of the 5 provided populations
- Download stats as TXT
  - Download the FST or other stats data frame as a txt file
- Visualise summary stats
  - When choosing summary stats and population, plots are automatically shown

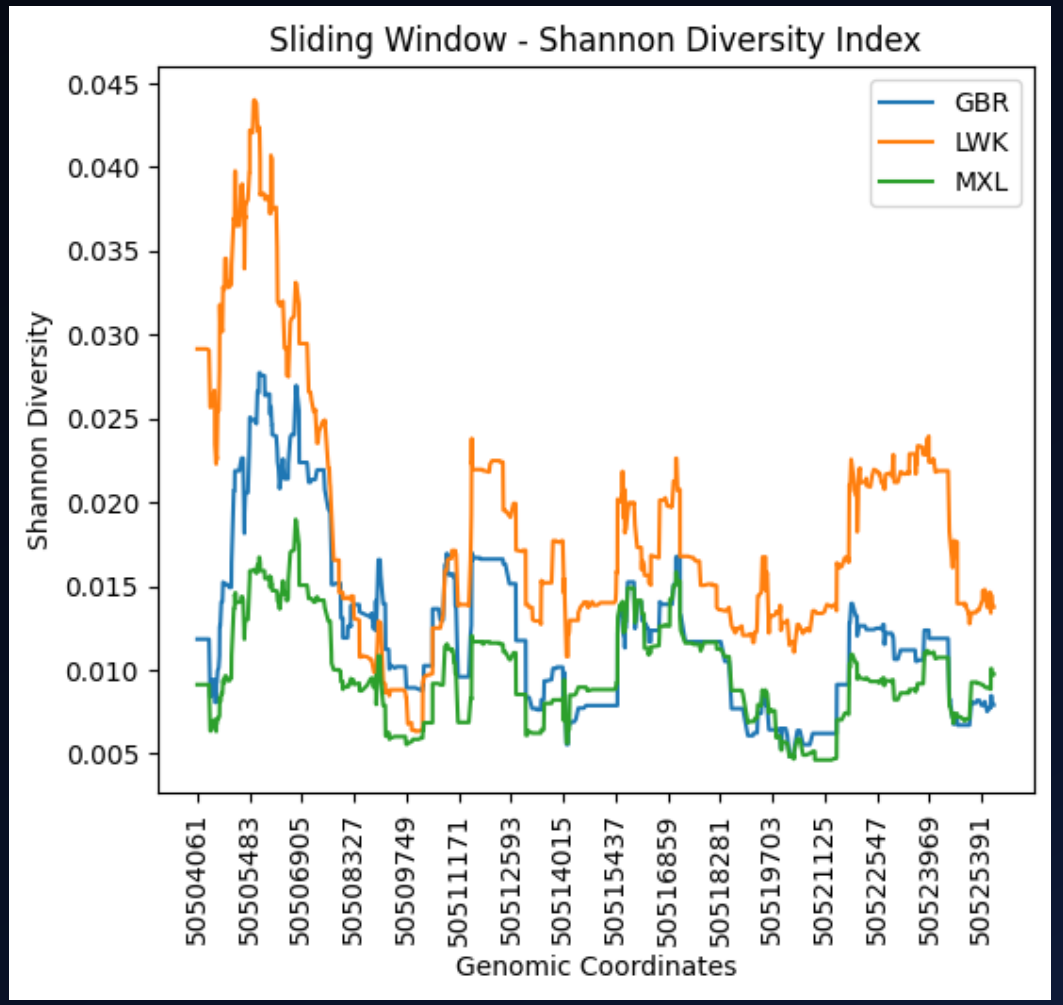
Choice of Summary Statistics

1. Shannon Diversity
   - More effective at describing diversity than using allele counts/richness
   - Sample size
2. Expected Heterozygosity
   - Study the difference in genetic variation in populations affected by urbanisation
   - Physiological responses e.g., disease resistance
3. Tajima's D
   - Infer which loci were affected by natural selection by identifying difference between observed and expected allele frequencies
4. FST
   - Hudson method – not affected by change in sample size and avoid false positive signals

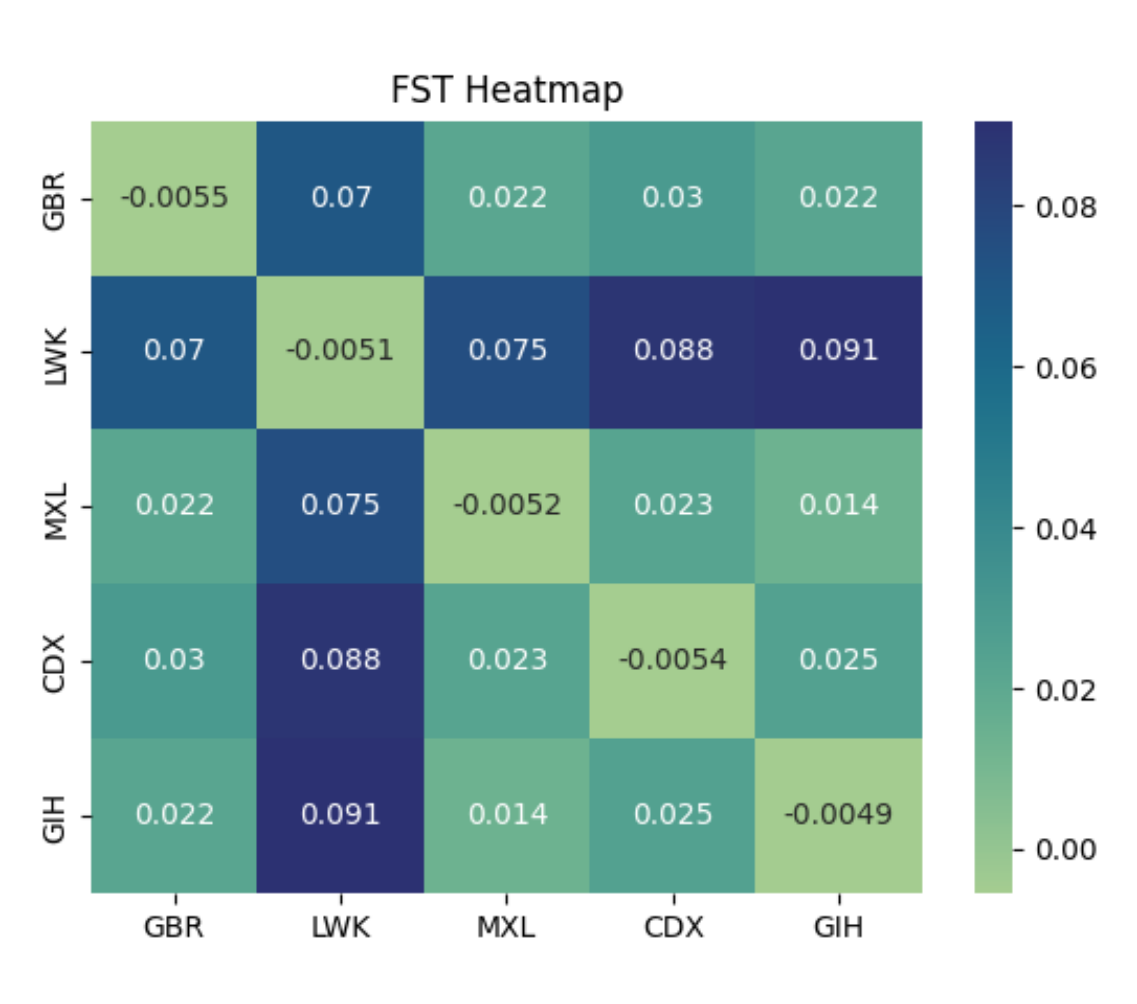
Justification for Visualisation Type

Sliding windows

- X-axis: positions to avoid gaps in graph
- Line graphs commonly used to demonstrate change in variability throughout the region
- Overlapping uniform window size



Justification for Visualisation Type



Heatmap

- FST is calculated pairwise per population combination
- Limited number of comparisons
- Heatmap not too overloaded while easily visualizing which populations are most similar/different to each other

Software Demo

- Celine will show our web application now

Limitations & Opportunities

- Only 5 populations
- Bi-allelic SNPs
- Indels or structural variants ignored
- Uniform window size
- Overlapping windows → less statistical power
- Interactive graphs

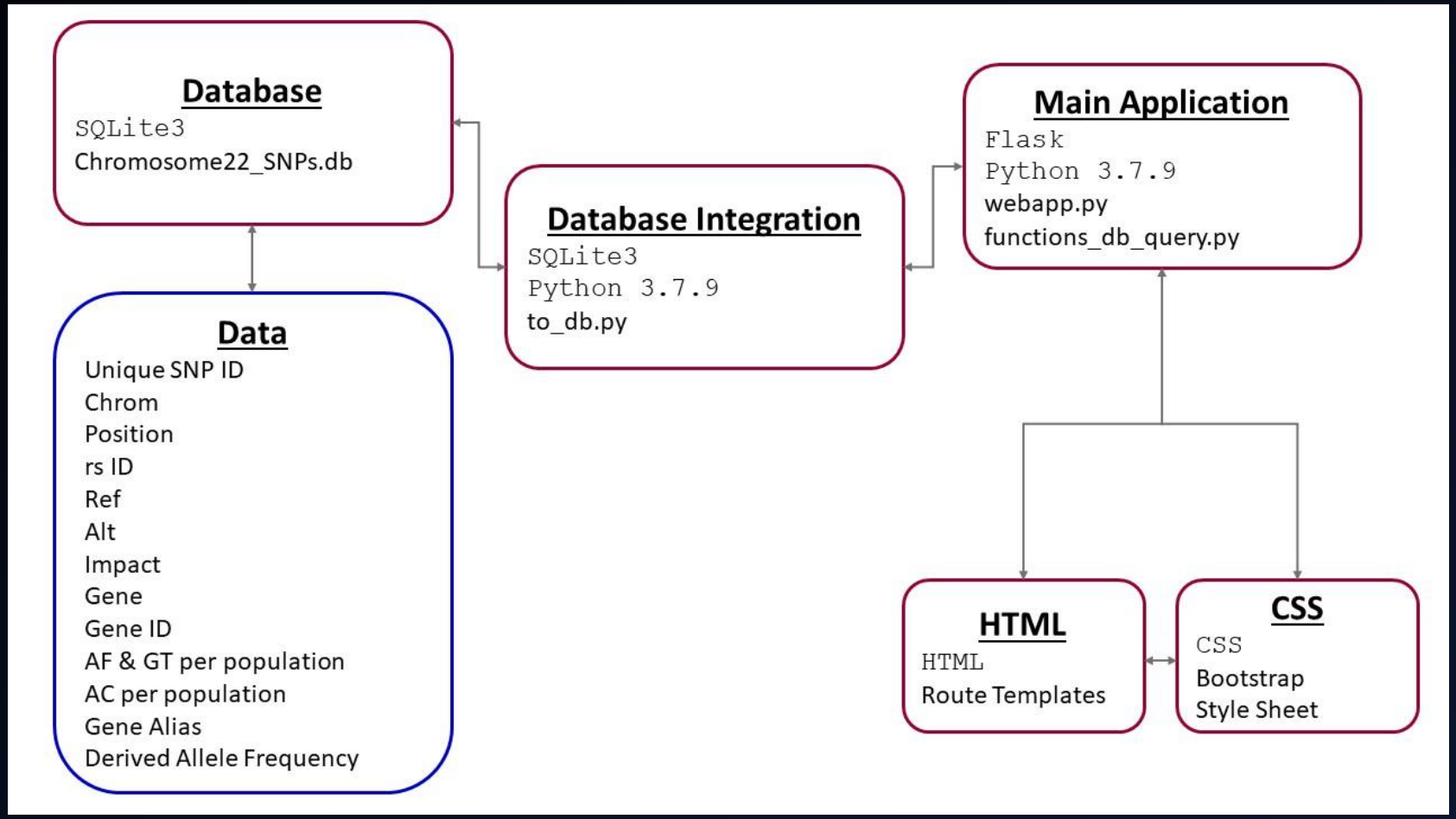
Thank you for listening!

QUESTIONS OR COMMENTS?

Individual Contributions

	Amanah	Celine	Gracia
Implementation 	Data Pre-processing for derived allele frequency (VCFtools)	Web App incl. Flask, HTML / CSS	Data Pre-processing (bcftools) & Data Wrangling
	Genotype & Allele Frequency Calculation	Python to SQL data base	Genotype & Allele Frequency Calculation
	FST statistic	Tajima‘s D and Shannon diversity	Sliding Window
Documentation 	Description of contributed implementation		
	Justification of summary stats	Running the software	Structure / Outline
	Literature research / References	Database connection & query Data visualisation	Transfer documentation into LaTeX

Software Architecture



Data Mining & Wrangling

- Bcftools: pre-process VCF files by population samples
- Vcftools: extract derived allele frequency for SNPs
- R genetics: calculate allele counts, allele & genotype frequency
- Python packages
  - pandas: modify, join and filter data
  - scikit-allele: calculate Tajima's D & FST
  - pandas & math: calculate Shannon Div. & Exp. Heterozygosity