The STRAF Book

Alexandre Gouy and Martin Zieger

2021-09-26

Contents

Preface

What is this book?

This is the online version of **The STRAF Book**, which is currently under active development. It is dedicated to the STRAF software, a web application for the analysis of genetic data in forensic practice.

Forensic and population genetics, lost sisters

Genetics has many faces, and forensic and population genetics are two of them. If we were to summarise their respective scopes, we could say that the former is the application of genetics to legal matters, and the latter aims at understanding genetic differences within and between populations, a fundamental matter in evolutionary biology.

Forensic genetics and population genetics have always been tightly linked disciplines. This is likely because quite a number of questions they address are similar. Even though problems in forensics and population genetics seem different, they often correspond to the same question, simply phrased differently.

For example, DNA profiling, used in criminal investigations or parental testing, aims at matching different DNA samples and understanding how related they are in terms of DNA. In population genetics, a common goal is to characterise the genetic diversity of a set of populations, by looking at how related individuals are within and between populations. Both fields aim at **understanding** and **quantifying** the **relatedness** of individuals based on their DNA.

Software and metrics developed in the population genetics for the study of the evolution of species are now used routinely in forensic genetics practice. But forensics is not just *applied population genetics*. The legal implications and unique situations encountered in the forensics world also led to the development of relevant statistical tools and metrics with a more specific purpose.

And then there was STRAF

STRAF was born from the encounter of two scientists: a forensic geneticist and a population geneticist, in 2017, in the beautiful city of Bern (Switzerland). Martin came to visit a population genetics lab, where Alexandre was pursuing his Ph.D. thesis at that time. This encounter led to a fruitful collaboration when they realised that some tools used in population genetics could be leveraged by the forensics community.

This encounter led to a fruitful collaboration when they realised that some tools used in population genetics could be leveraged by the forensics community. The most striking example is the computation of forensics parameters, that describe for example how good are our loci at discriminating samples. These parameters were typically computed using a spreadsheet that had been created by one of the suppliers of assays used to genotype samples. It is the mythical PowerStats v1.2 spreadsheet, allowing to compute forensic statistics and allele frequencies in Microsoft Excel. It has been since then removed from the Internet, and forensic geneticists started sharing this spreadsheet among each other, circulating almost secretly, “under the cloak” as French speakers would say.

As similar operations were done in routine in population genetics, we already had some scripts for the analysis of STR data. Then, after we applied them to an existing dataset, we decided to put everything into a web application so that the forensics community could benefit from it.

A few weeks later, STRAF was born, and after four year, STRAF had become a widely used tool by the forensics community, but not only. It has been used as a support for teaching population genetics, and has been used in evolutionary biology studies.

The positive reception of the software in the community motivated its development over the years until the release of STRAF 2.0 in 2021.

What will you learn?

By reading this book, our hope is that you will:

- Get an overview of common **concepts** in forensic and population genetics
- Learn how to use the **STRAF software** for STR data analysis through **practical applications**
- Be able to **interpret** common metrics and analyses used in forensics practice

Outline

The book is organised as follow:

- We'll start by an **Introduction** to essential forensic and population genetics concepts.
- In **Chapter 1**, we will focus on data, from its generation to its preparation for downstream analysis in STRAF.
- In **Chapter 2**, we will review **forensic parameters** that can be computed in STRAF, and discuss their interpretation.
- In **Chapter 3**, we will review essential population genetics concepts and describe **population genetics indices** that can be computed in STRAF.
- In **Chapter 4**, we will focus on **multivariate statistics** and how they can provide insights into population structure, with a particular focus on Principal Component Analysis (PCA) and Multidimensional Scaling (MDS), two widely used approaches in genetics.
- In **Chapter 5**, we will explain how to compare samples of interest to **reference populations** by loading reference allele frequencies into the software and performing a Multidimensional Scaling analysis.

- In **Chapter 6**, we gather recommendations around potential next analysis steps by presenting STRAF's **file conversion** capabilities and useful methods implemented in **other software**.

Introduction

In this chapter, we will briefly introduce some essential concepts in genetics.

DNA and genetic variation

Each of our cells contains 23 pairs of **chromosomes**, composed of a long **DNA** (deoxyribonucleic acid) molecule. Under this somewhat barbaric name is hiding a simple concept. This molecule is the support of the information used by the body to function and development. This information is encoded by a chain of **nucleotides** of four types that can be referred to using the letters A (Adenine), T (Thymine), C (Cytosine) and G (Guanine).

Developments in biotechnologies enabled the characterisation of the DNA of individuals. These techniques also led to the discovery that this DNA varies between individuals. This **genetic variation**, also called **polymorphism**, can be used to characterise individuals and populations based on their DNA.

Markers of polymorphism

DNA variation can take different forms: it can for example be a **Single Nucleotide Polymorphism** (**SNP**), when a mutation occurs and changes a nucleotide at a given position in the genome. In that case, we would observe different nucleotides in a population at a single position.

There can also be **insertions** and **deletions** (sometimes referred to as **InDels**), of one or multiple nucleotides.

Finally, other markers of genetic variation are **Copy Number Variants** (**CNVs**), when a sequence is repeated a certain number of times. They can contain more or less repetitive units. These units can contain more or less nucleotides. **Short Tandem Repeats** (**STRs**) a type of genetic polymorphism consisting in short sequences from 2 to 7 base pairs that are repeated. The **number of repeats varies** among individuals, therefore characterizing their length can be useful to identify individuals. STRs are still nowadays one of the most commons markers used in forensic genetics.

Polymorphism and forensics

DNA profiling and typing

As DNA varies between individuals, DNA typing became a central element of the forensic scientist toolkit. For example, typical questions forensic genetics aims at answering include:

- What is the probability that a randomly-picked person in a population would match the individual of interest in terms of DNA?
- Which proportion of the population has the same combination of genetic variants as the sample of interest?

The role of population genetics

To answer these questions, it is crucial to first get a good characterisation of **genetic variants** (or **alleles**) frequencies in populations of interest, at different **loci** across the genome. Indeed, these frequencies can vary widely among populations.

In this context, STRAF has been designed to facilitate the analysis of **population data** in forensic genetics.

Digression - Why STRs and not other markers?

Nowadays, it is easier and cheaper to generate whole-genome sequences. One could wonder why STRs are still so popular in forensic practice, and have not been replaced by SNPs that are easily generated from Next-Generation Sequencing (NGS) data.

This is mainly due to the fact that STRs have a high mutation rate, therefore are more diverse in human populations. This explains their high **power of discrimination**.

Furthermore, they can also be used in deciphering mixture components, a very common case in forensics.

Finally, they can be combined in multiplex assays, which is convenient when low amounts of biological material can be recovered.

For all these reasons, STRs remain the dominant marker used in forensic genetics.

Chapter 1

Importing data

In this chapter, we will explain how to prepare the input file containing genotypic data, and how to upload it into STRAF.

1.1 STR data and STRAF input format

STR data consists in various **observations**: the **genotypes** (number of STR repeats) of **each individual**, at **each locus**.

In genetics, we potentially observe two values per individual and per locus, if the markers are diploid, that is two copies are present per sample.

STRAF's input file is a **text** file containing the genotypes of each sample:

- The first column, named **ind**, needs to contain the sample ID
- The second column, , named **pop**, contains the population ID (this column must exist even if a single population is studied)
- The next columns correspond to genotypes: for haploid samples, one column per locus must be reported; for diploid data, two columns per locus (with the

same name)

- Genotypes must be encoded as numbers (STRAF accepts point alleles)
- Missing data (e.g. null alleles) must be indicated with a “0”.

For diploid data, the table should look like this:

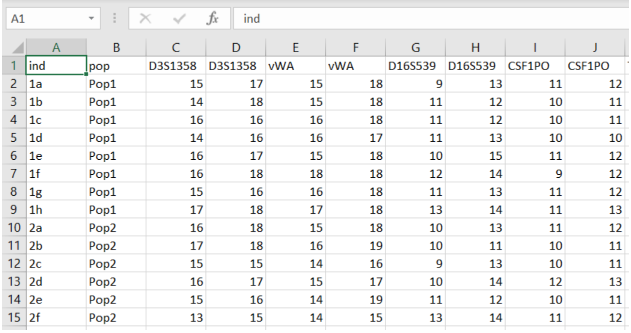
ind	pop	Locus1	Locus1	Locus2	Locus2
A	Bern	12	14	17	17
B	Bern	14	14	13	15.2
C	Lausanne	12	16	15.2	17

For haploid data, the table would look like this:

ind	pop	Locus1	Locus2
A	Bern	12	17
B	Bern	14	13
C	Lausanne	12	15.2

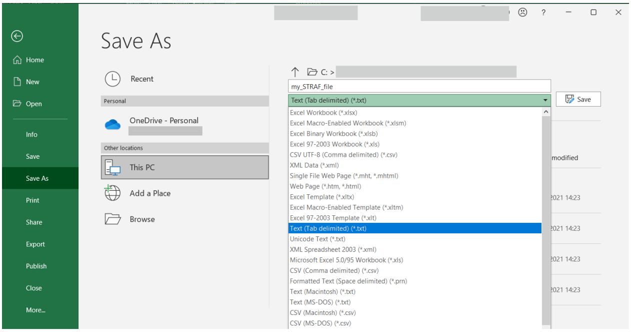
1.2 Generating the input data from Excel

It only takes a few steps to generate an input file in a format that is suitable for use in STRAF. From Excel, for example, we can start from a spreadsheet looking like this:



	A	B	C	D	E	F	G	H	I	J
1	ind	pop	D3S1358	D3S1358	vWA	vWA	D16S539	D16S539	CSF1PO	CSF1PO
2	1a	Pop1	15	17	15	18	9	13	11	12
3	1b	Pop1	14	18	15	18	11	12	10	11
4	1c	Pop1	16	16	16	18	11	12	10	11
5	1d	Pop1	14	16	16	17	11	13	10	10
6	1e	Pop1	16	17	15	18	10	15	11	12
7	1f	Pop1	16	18	18	18	12	14	9	12
8	1g	Pop1	15	16	16	18	11	13	11	12
9	1h	Pop1	17	18	17	18	13	14	11	13
10	2a	Pop2	16	18	15	18	10	13	11	12
11	2b	Pop2	17	18	16	19	10	11	10	11
12	2c	Pop2	15	15	14	16	9	13	10	11
13	2d	Pop2	16	17	15	17	10	14	12	13
14	2e	Pop2	15	16	14	19	11	12	10	11
15	2f	Pop2	13	15	14	15	13	14	11	12

Then, one simply needs to save this table as a tab-delimited text file. This can be achieved by clicking on `Save As` > `Text (Tab-delimited) (*.txt)`


Home
New
Open
Info
Save
Save As
Print
Share
Export
Publish
Close
More...
Save As
Recent
Personal
OneDrive - Personal
Other locations
This PC
Add a Place
Browse
C: >
Save
modified
2021 14:23
2021 14:23
2021 14:23
2021 14:23
Excel Workbook (*.xlsx)
Excel Macro-Enabled Workbook (*.xlsm)
Excel Binary Workbook (*.xlsb)
Excel 97-2003 Workbook (*.xls)
CSV UTF-8 (Comma delimited) (*.csv)
XML Data (*.xml)
Single File Web Page (*.mht, *.mhtml)
Web Page (*.htm, *.html)
Excel Template (*.xltx)
Excel Macro-Enabled Template (*.xltm)
Excel 97-2003 Template (*.xlt)
Text (Tab delimited) (*.txt)
Unicode Text (*.txt)
XML Spreadsheet 2003 (*.xml)
Microsoft Excel 5.0/95 Workbook (*.xls)
CSV (Comma delimited) (*.csv)
Formatted Text (Space delimited) (*.prn)
Text (Macintosh) (*.txt)
Text (MS-DOS) (*.txt)
CSV (Macintosh) (*.csv)
CSV (MS-DOS) (*.csv)

my_STRAF_file

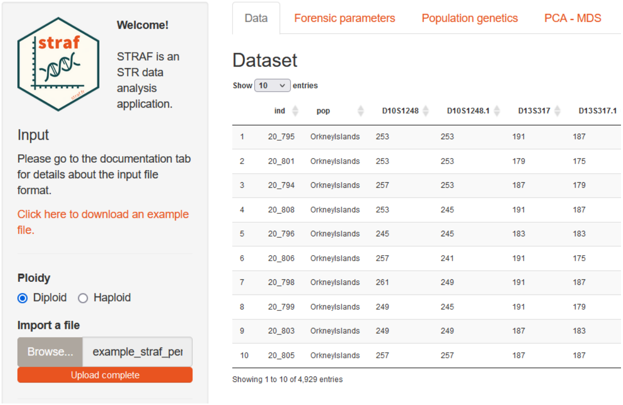
Text (Tab delimited) (*.txt)

1.3 Uploading the data to STRAF

Once the input file has been prepared, it is possible to upload it into STRAF. Alternatively, you can **download an example file** by clicking on the link in the sidebar.

Then, you need to select the **ploidy** of your data (for example, *Diploid* for autosomal markers, and *Haploid* for Y-chromosome markers). After that, you can upload the file by clicking on *Browse* and selecting the file on your computer.

Once the file is uploaded, a preview will be displayed on the right and all the analyses will be available. If not, it is likely that an error has occured. If the error message is not explicit, you can refer to the **Input file checklist** below that gathers common issues with the input file.



Welcome!

STRAF is an STR data analysis application.

Input

Please go to the documentation tab for details about the input file format.

Click here to download an example file.

Ploidy

● Diploid ○ Haploid

Import a file

example_straf_per

Upload complete

Dataset

Show 10 entries

	ind	pop	D10S1248	D10S1248.1	D13S317	D13S317.1
1	20_795	OrkneyIslands	253	253	191	187
2	20_801	OrkneyIslands	253	253	179	175
3	20_794	OrkneyIslands	257	253	187	179
4	20_808	OrkneyIslands	253	245	191	187
5	20_796	OrkneyIslands	245	245	183	183
6	20_806	OrkneyIslands	257	241	191	175
7	20_798	OrkneyIslands	261	249	191	187
8	20_799	OrkneyIslands	249	245	191	179
9	20_803	OrkneyIslands	249	249	187	183
10	20_805	OrkneyIslands	257	257	187	187

Showing 1 to 10 of 4,929 entries

1.4 Common issues

Even though you've been very careful in the generation of STRAF's input file, it is possible that you still run into an error after uploading the file to STRAF. In case STRAF cannot read your input file, we've put together a checklist to identify common issues with the input file.

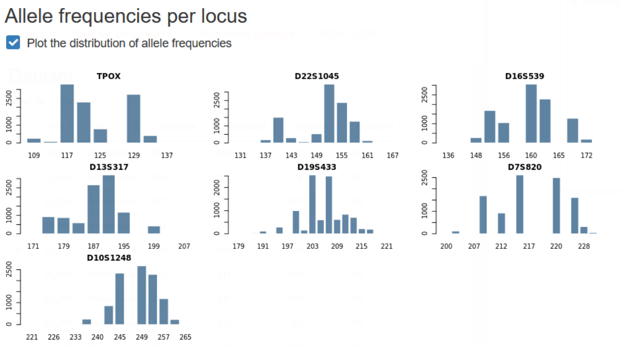
Input file checklist

- Check input parameters in the sidebar: do they actually correspond to the input data?
- Check locus names: are they all different for haploid data? Do both columns for a single locus for diploid data have the exact same name?
- Check that all missing data have been encoded with a "0"
- Try to remove any special characters from sample and locus names
- Check for the presence of empty spaces at the end of each line
- Check if alleles are exclusively encoded with numbers
- Check if values are separated by tabs and not spaces

- Check if the first two columns are names “ind” and “pop”

1.5 Having a first look at the data

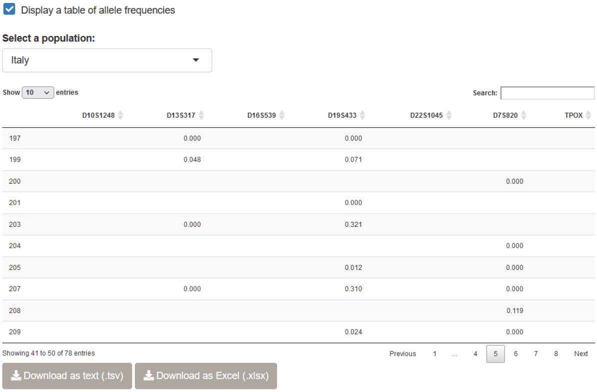
Below the dataset preview, you will be able to generate a plot of allele frequencies at each locus.



Allele frequencies per locus

☑ Plot the distribution of allele frequencies

You can also generate an allele frequency table, which is standard practice when reporting new population data in a forensics journal. You can either download the allele frequencies in a text format (TSV), or as an Excel (XLSX) file. Note that you have the ability to select a specific population using the drop-down menu above the table.



☑ Display a table of allele frequencies

Select a population: Italy

Show 10 entries

Search:

	D10S1248	D13S317	D16S539	D19S433	D22S1045	D7S820	TPOX
197		0.000		0.000			
199		0.048		0.071			
200						0.000	
201				0.000			
203		0.000		0.321			
204						0.000	
205				0.012		0.000	
207		0.000		0.310		0.000	
208						0.119	
209				0.024		0.000	

Showing 41 to 50 of 78 entries

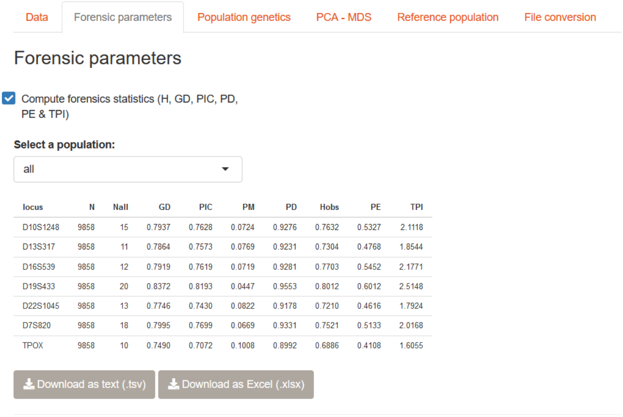
Chapter 2

Forensic parameters

In this chapter, we'll show how to compute forensic parameters using STRAF, and provide details on how they are computed and should be interpreted. We'll introduce a few equations, but please do not be afraid! The goal of this chapter is to translate each of them into plain English.

2.1 How to compute forensic parameters in STRAF

Once your data has bee uploaded, you can go to the **Forensic parameters** tab and check the *Compute forensics statistics* box. The computation will be performed and a table containing the values per locus will be displayed. The computation is done per population and overall, a drop-down menu is present to select the population.



Forensic parameters

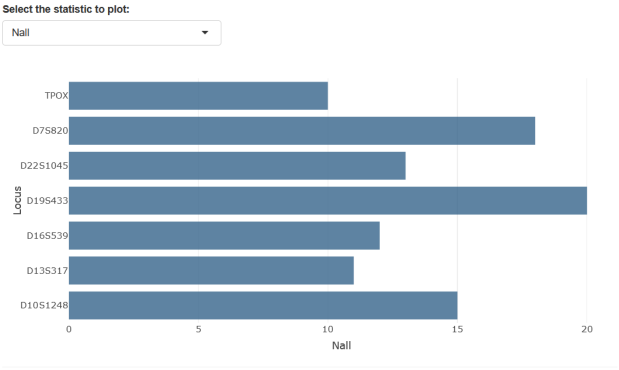
☑ Compute forensics statistics (H, GD, PIC, PD, PE & TPI)

Select a population:

all

locus	N	Nall	GD	PIC	PM	PD	Hobs	PE	TPI
D10S1248	9858	15	0.7937	0.7628	0.0724	0.9276	0.7632	0.5327	2.1118
D13S317	9858	11	0.7864	0.7573	0.0769	0.9231	0.7304	0.4768	1.8544
D16S539	9858	12	0.7919	0.7619	0.0719	0.9281	0.7703	0.5452	2.1771
D19S433	9858	20	0.8372	0.8193	0.0447	0.9553	0.8012	0.6012	2.5148
D22S1045	9858	13	0.7746	0.7430	0.0822	0.9178	0.7210	0.4616	1.7924
D7S820	9858	18	0.7995	0.7699	0.0669	0.9331	0.7521	0.5133	2.0168
TPOX	9858	10	0.7490	0.7072	0.1008	0.8992	0.6886	0.4108	1.6055

Below the forensic parameters table, You can select the metric you would like to represent using the drop-down menu.


Select the statistic to plot:

Nall

2.2 Details on the forensic parameters

2.2.1 Random match probability (PM)

The **Random match probability**, or probability of matching (PM), is defined as the probability of observing a random match between two individuals.

Formula

$$PM = \sum_i (G_i)^2,$$

where G_i is the frequency of the genotype i at a given locus in the population.

Interpretation

Computing PM means calculating, for a given locus, the frequency of each genotypes. Then we take the square of each frequency, i.e. we multiply it by itself. Finally, we sum the values of each genotype.

The intuition behind it is that if we observe a random match in a population when looking at a single locus, it means that our two samples have the same genotype at that locus. In terms of probabilities, sampling a specific genotype in the population has a probability equal to its frequency. And sampling the same genotype a second time (i.e., observing a match), is the probability of sampling this genotype multiplied by itself.

As an example, say the genotype “12-14” has a frequency of 5% in the population, the probability of having a random match between two individuals having the same genotype is 0.05 x 0.05.

To get an overall probability of matching, we sum this over all possible genotypes in our population.

2.3 Power of Discrimination (PD)

The power of discrimination (PD) is defined as the probability of discriminating between two unrelated individuals.

Formula

$$PD = 1 - PM$$

Interpretation

PD is simply 1 - PM. Instead of looking at the probability of matching, we are interested in the probability of “not matching”, i.e. the probability of discrimination.

2.4 Gene diversity

Gene diversity (GD, sometimes simply D), also called **expected heterozygosity** (H_{exp}), is computed using the following estimator:

Formula

$$H_{\text{exp}} = GD = \frac{n}{n-1}\left(1 - \sum_{i=1}^{n}(p_i)^2\right),$$

where n is the number of gene copies sampled and p_i is the frequency of the i^{th} allele in the population.

Interpretation

It is the probability that an individual will be heterozygous at a given locus.

As an example, a value of GD of 0.6 mean that 60% chance of being heterozygote at this locus.

It depends directly on the genetic diversity at this locus, which itself depends on allele frequencies in your population.

Say we have two alleles in a given population, genetic diversity will be higher in a population aith allele frequencies 0.5 and 0.5 than in a population where frequencies are 0.1 and 0.9 (as less heterozygotes can be made with rare alleles). This rationale can be extended to any number of alleles.

2.5 Polymorphism Information Content (PIC)

Formula

The **Polymorphism Information Content** (PIC) is computed as follow:

$$PIC = 1 - \sum_{i=1}^{n} p_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2p_i^2 p_j^2,$$

where p_i and p_j are allele frequencies.

Interpretation

The PIC can be interpreted as:

- the probability that the maternal and paternal alleles of a child are deducible
- or, the probability of being able to deduce which allele a parent has transmitted to the child.

2.6 Power of Exclusion (PE)

Formula

The **power of exclusion** (PE) is defined as:

$$PE = h^2 \left(1 - 2hH^2\right),$$

where h is the proportion of heterozygous individuals and H the proportion of homozygous individuals in the population sample.

Interpretation

The power of exclusion depends on the observed proportions of heterozygous and homozygous individuals in a population. These proportions, multiplied as in the equation above, give the **probability that two persons do not have the same genotype in the population**.

2.7 Typical Paternity Index (TPI)

Finally, the typical paternity index (TPI) reflects the “mean PI for random non-excluded men“ for a given locus.

Formula

Let H be the frequency of homozygotes, then

$$TPI = \frac{1}{2H}$$

Interpretation

Unlike the other parameters, the typical paternity index values are not in the 0-1 range and cannot be interpreted as a probability. It is a ratio of 1 divided by twice the frequency of homozygotes in the population. This is an **odds ratio**, measuring how many times more likely that a possible father is the actual father than a randomly selected man in the population, on average (*typical* case).

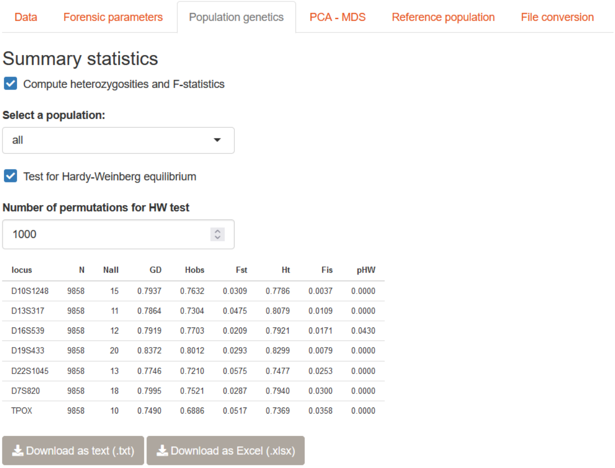
Chapter 3

Population genetics indices

In this chapter, we will see how to compute some population genetics indices in STRAF.

3.1 Computing population genetics parameters in STRAF

Once you have uploaded your genotypes in STRAF, you can go to the **Population genetics** tab to compute relevant population genetics indices. It is also possible to perform a Hardy-Weinberg equilibrium test by checking the relevant box.



Summary statistics

☑ Compute heterozygosities and F-statistics

Select a population: all

☑ Test for Hardy-Weinberg equilibrium

Number of permutations for HW test: 1000

locus	N	Nall	GD	Hobs	Fst	Ht	Fis	pHW
D10S1248	9858	15	0.7937	0.7632	0.0309	0.7786	0.0037	0.0000
D13S317	9858	11	0.7864	0.7304	0.0475	0.8079	0.0109	0.0000
D16S539	9858	12	0.7919	0.7703	0.0209	0.7921	0.0171	0.0430
D19S433	9858	20	0.8372	0.8012	0.0293	0.8299	0.0079	0.0000
D22S1045	9858	13	0.7746	0.7210	0.0575	0.7477	0.0253	0.0000
D7S820	9858	18	0.7995	0.7521	0.0287	0.7940	0.0300	0.0000
TPOX	9858	10	0.7490	0.6886	0.0517	0.7369	0.0358	0.0000

3.2 Details on population genetics indices

3.2.1 Hardy-Weinberg equilibrium

A population is considered at Hardy-Weinberg equilibrium (HWE) when the observed genotypic frequencies are in agreement with the expectations in an “ideal” population, which assumes for example random mating in the population. This is important as the assumptions of the Hardy-Weinberg model allow to derive quantities such as forensic parameters and population genetic indices. Therefore, if some assumptions of the model are violated, conclusions drawn from metrics computed assuming HWE could be challenged.

If a locus presents a significant deviation from HWE, it means that a process is influencing the distribution of allele and genotype frequencies in the population. It could for example be due to **inbreeding**, **hidden population structure**, or **natural selection**.

STRAF reports the **p-value** of a test for HWE. A low p-value indicates a significant deviation from HWE.

3.2.2 Heterozygosities

In STRAF, several measures of **heterozygosity** are computed. They capture different aspects of genetic diversity.

- The **expected heterozygosity** (or **Gene diversity**, H_{exp} or GD) has been defined in the previous chapter.
- The **observed heterozygosity** (H_{obs}) is the proportion of heterozygous genotypes at this locus in the population.
- The **total heterozygosity** (H_T) is the heterozygosity expected if all the individuals in all the subpopulations were behaving as a population at HWE.

3.2.3 F-statistics

STRAF reports two **F-statistics**.

- The F_{IS} is a measure of genetic relatedness within a population. It is sometimes called the **inbreeding coefficient**. High values indicate a high degree of inbreeding.
- The F_{ST} is a measure of genetic differentiation between populations. It takes values between 0 (no differentiation) and 1 (full differentiation).

One concept, multiple estimators.
Several **estimators** of F_{ST} exist (for example, Weir and Cockerham's, Nei's, Hudson's F_{ST}). It's like if each population geneticist had decided to develop their own estimator!
Why is that? In statistics, what we call an **estimator** a metric aiming at estimating a given quantity based on **observed data**. It is important to keep in mind that these estimators rely on a specific **model**, with underlying assumptions. It explains why some estimators are more or less reliable depending on the case and observed data, and each of them has been developed for a different situation. In the case of F_{ST} for example, different estimators assume different demographic models.

3.3 Linkage disequilibrium (LD)

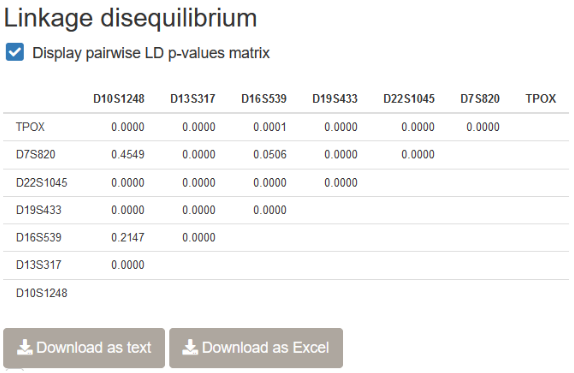
3.3.1 What is linkage disequilibrium?

Linkage disequilibrium is an important quantity to be measured in genetics. It is defined as the **nonindependence of genotypes at distinct loci**. It means that it is more likely to observe the co-occurence of some genotypes at different loci.

Most of the times, this is explained by the physical proximity between loci. If two loci are next to each other on the genome, recombination events between them will be rare and genotypes won't be suffled. Genotypes at these loci will be correlated and linkage disequilibrium will be high. On the other hand, two loci found on two different chromosomes are not expected to show any LD signals as genotypes will be systematically shuffled at each generation.

3.3.2 How to compute LD in STRAF?

It is possible to test for the presence of LD in the dataset you uploaded using STRAF. After checking the *Display pairwise LD p-values matrix*, LD tests between each pair of loci will be performed and p-values will be reported.



Linkage disequilibrium

☑ Display pairwise LD p-values matrix

	D10S1248	D13S317	D16S539	D19S433	D22S1045	D7S820	TPOX
TPOX	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	
D7S820	0.4549	0.0000	0.0506	0.0000	0.0000		
D22S1045	0.0000	0.0000	0.0000	0.0000			
D19S433	0.0000	0.0000	0.0000				
D16S539	0.2147	0.0000					
D13S317	0.0000						
D10S1248							

Important note: Other population genetics software, **Genepop** and **Arlequin**, implement more reliable versions of the LD test that should be preferred. They are currently not implemented in STRAF because of performance limitations. If you need to perform such a test, the **File conversion** utilities (cf. Chapter 6) should facilitate the workflow.

Chapter 4

Multivariate statistics

In this chapter, we will describe the two multivariate statistics methods implemented in STRAF, **Principal Component Analysis** (**PCA**) and **Multidimensional Scaling** (**MDS**).

4.1 Principal Component Analysis (PCA)

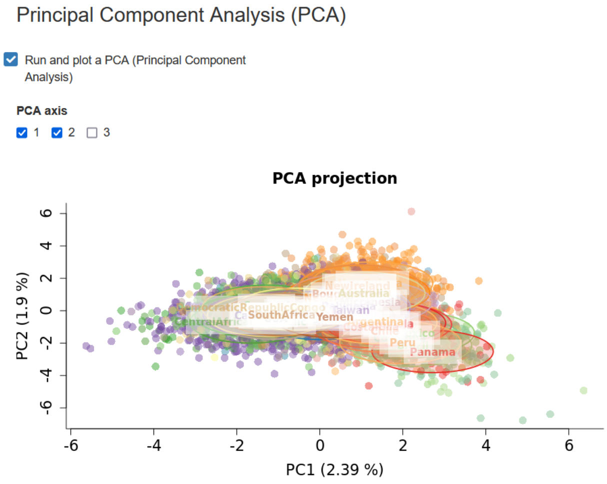
PCA is a method of **dimensionality reduction**. What is does it that it tries to **capture most of the variation** present in the data and **project** it onto a small number of new axes called **principal components** (PCs).

This is a useful method to capture variation from a large number of variables and allows to discover hidden patterns by increasing interpretability.

In our case, if we consider that **each allele at each locus** is a variable, and that our individual observations are the presence / absence of each allele for each sample, we end up with a highly dimensional dataset (we have as many variables as we have alleles!). It gets even worse if you analyse genome sequences, where you can have millions of variables in your dataset. This is definitely not an interpretable dataset if you are not able to easily extract relevant information.

PCA allows to bring most of the variation existing between our samples onto a few axes capturing most of the variation (PCs).

In STRAF, you can perform a PCA by going into the **PCA - MDS** tab and checking the following *Run and plot a PCA* checkbox.



Principal Component Analysis (PCA)

☑ Run and plot a PCA (Principal Component Analysis)

PCA axis

☑ 1 ☑ 2 □ 3

It will trigger the PCA computation and return a graph. You are also able to download to coordinates on different PCs (also called eigenvectors).

Interpreting PCA results

PCA has become a popular tool in population genetics, as the relationships between individuals on the PCA projection tends to reflect their genetic **relatedness**. The closer individuals are on the PCA projection, the more genetically related they tend to be.

It can also be used as a **quality control** tool. For example, if a sample is very far from all the others, including the ones that are part of the same population, it is possible that there is an issue with the data. One would need to inspect the raw data and check if any strange pattern can be found.

It is important to be aware of the influence of non demographic processes on the PCA projection. For example, **imbalanced sample sizes** between populations can drive some patterns. When populations are sampled unevenly, the projection will be **distorted** and distances observed on the projection can be driven by these differences and not by their evolutionary history.

Hence, as multiple processes influence the results, PCA should remain an **exploratory approach** and further analyses should be conducted before drawing any major conclusions on the relationships between populations and individuals.

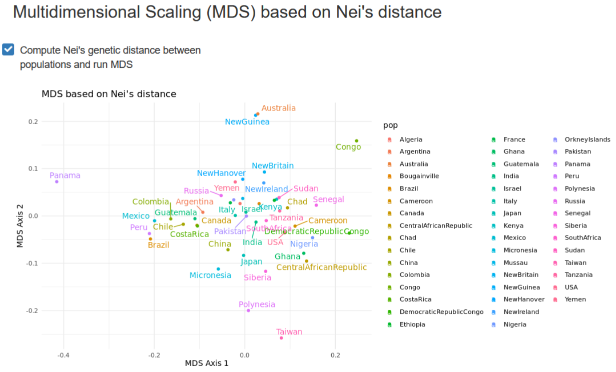
4.2 Multidimensional Scaling (MDS)

An **MDS** is conceptually similar to a PCA. One of the main differences is that it takes as input data in a different format. As PCA uses raw genotypes and can accommodate data at the individual level, **pairwise distances** between data points are used for an MDS.

In forensics practice, it is typically used to compare **populations** and not **individuals**, even though it would be theoretically possible.

As **pairwise distances between populations** are used, and MDS can be based on any genetic distance: one could for example compute pairwise **FSTs**, **Nei's genetic distance**, or any other metric.

Based on these distances, the MDS will **project** the populations onto a **lower-dimensional space**. This projection facilitates the interpretation of relationships between populations.



Multidimensional Scaling (MDS) based on Nei's distance

☑ Compute Nei's genetic distance between populations and run MDS

Interpreting MDS results

Just like with a PCA, populations the **closer** populations are on the MDS projection, the more **genetically related** the tend to be, based on the markers that have been used to compute their pairwise genetic distances. Again, like for the PCA, the MDS should remain an **exploratory approach** and further analyses should be conducted before drawing any major conclusions on the relationships between populations and individuals.

Chapter 5

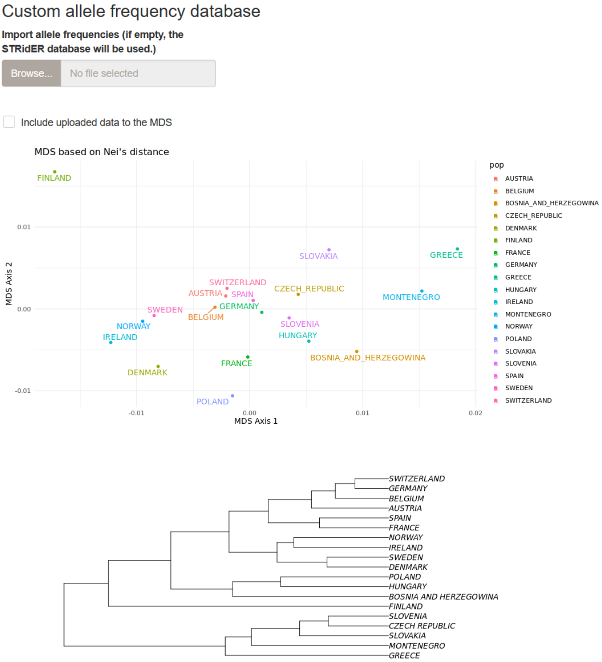
Reference populations analysis

In this chapter, we will see how to compare the uploaded populations to reference populations (based on allele frequencies).

5.1 MDS on reference frequencies

STRAF implements an MDS computation based on allele frequency data. By default, allele frequencies from the STRiDER database are used (loci with less than 10 populations have been excluded). If you are not familiar with the MDS method and interpretation, you can find details in **Chapter 4**.

You can see the output of the MDS in the **Reference population** tab. It includes the MDS projection in two dimensions for 19 European populations, as well as the population tree.

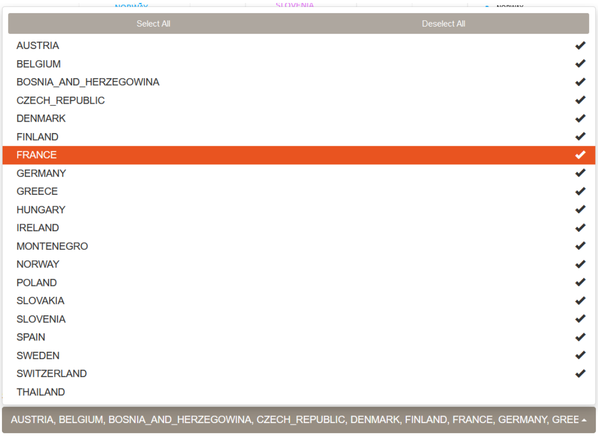


Custom allele frequency database

Import allele frequencies (if empty, the STRidER database will be used.)

□ Include uploaded data to the MDS

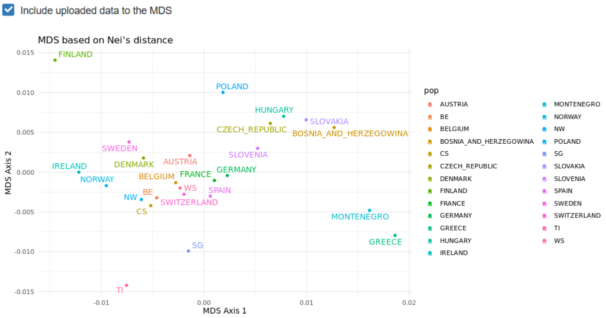
The drop-down menu below the plots allows one to select or unselect populations to be used in the MDS analysis.



AUSTRIA ✓
BELGIUM ✓
BOSNIA_AND_HERZEGOWINA ✓
CZECH_REPUBLIC ✓
DENMARK ✓
FINLAND ✓
FRANCE ✓
GERMANY ✓
GREECE ✓
HUNGARY ✓
IRELAND ✓
MONTENEGRO ✓
NORWAY ✓
POLAND ✓
SLOVAKIA ✓
SLOVENIA ✓
SPAIN ✓
SWEDEN ✓
SWITZERLAND ✓
THAILAND

AUSTRIA, BELGIUM, BOSNIA_AND_HERZEGOWINA, CZECH_REPUBLIC, DENMARK, FINLAND, FRANCE, GERMANY, GREE

Provided that some loci are in common between the reference allele frequencies and the genotypes you uploaded, it is possible to add your own populations to the existing MDS by checking the *Include uploaded data to the MDS* box.


☑ Include uploaded data to the MDS

You will then see your populations added to the MDS plots.

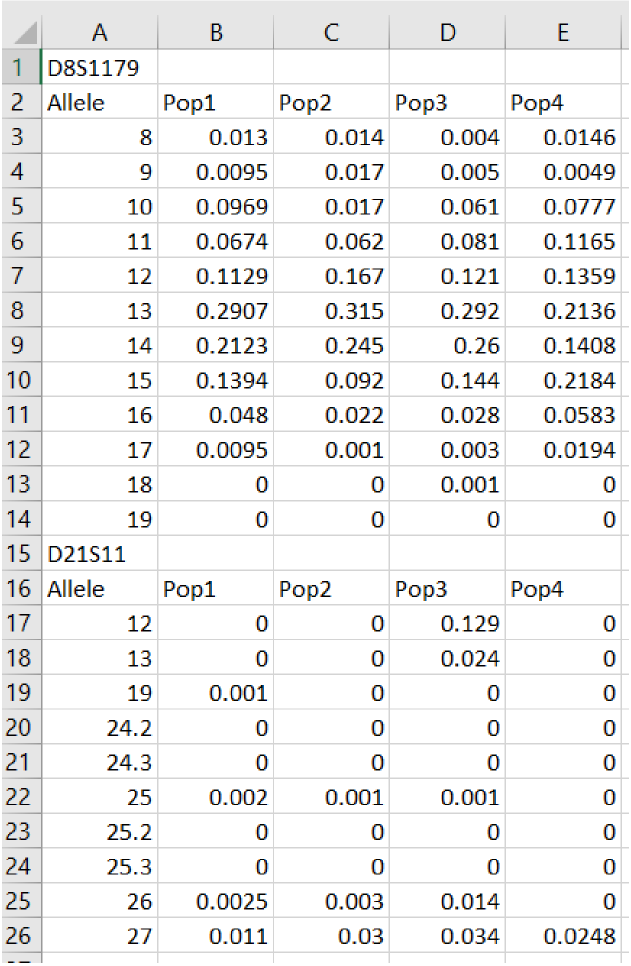
5.2 Preparing a custom allele frequency database

If you want to use other reference databases than STRIDER, for example if you're working with non-European samples or with different loci, it is possible to upload a custom allele frequency database in STRAF.

The data must be formatted as follows:

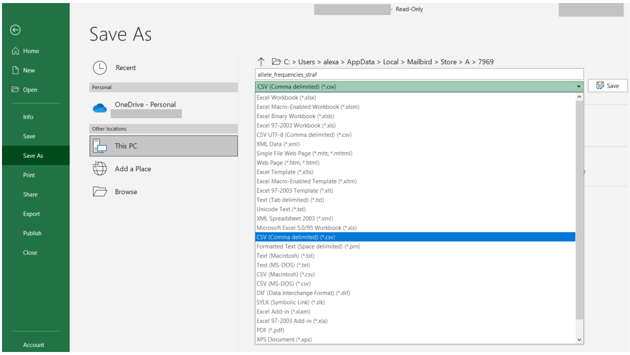
D1S1656			
Allele	Switzerland	Germany	France
9	0.12	0.09	0
10	0.40	0.35	0.28
10.2	0.31	0.41	0.5
11	0.17	0.15	0.22
D2S1338			
Allele	Switzerland	Germany	France
19	0.40	0.38	0.42
20	0.42	0.26	0.28
21	0.31	0.36	0.3

Starting from an Excel file for example, we can start with a spreadsheet that looks like this:



	A	B	C	D	E
1	D8S1179				
2	Allele	Pop1	Pop2	Pop3	Pop4
3	8	0.013	0.014	0.004	0.0146
4	9	0.0095	0.017	0.005	0.0049
5	10	0.0969	0.017	0.061	0.0777
6	11	0.0674	0.062	0.081	0.1165
7	12	0.1129	0.167	0.121	0.1359
8	13	0.2907	0.315	0.292	0.2136
9	14	0.2123	0.245	0.26	0.1408
10	15	0.1394	0.092	0.144	0.2184
11	16	0.048	0.022	0.028	0.0583
12	17	0.0095	0.001	0.003	0.0194
13	18	0	0	0.001	0
14	19	0	0	0	0
15	D21S11				
16	Allele	Pop1	Pop2	Pop3	Pop4
17	12	0	0	0.129	0
18	13	0	0	0.024	0
19	19	0.001	0	0	0
20	24.2	0	0	0	0
21	24.3	0	0	0	0
22	25	0.002	0.001	0.001	0
23	25.2	0	0	0	0
24	25.3	0	0	0	0
25	26	0.0025	0.003	0.014	0
26	27	0.011	0.03	0.034	0.0248

Then, one simply needs to save this table as a CSV (Comma-Separated Values) file. This can be achieved by clicking on `Save As` > `CSV (Comma-delimited) (*.csv)`


[redacted] - Read-Only [redacted]
Home
New
Open
Info
Save
Save As
Print
Share
Export
Publish
Close
Account
Save As
Recent
Personal
OneDrive - Personal
[redacted]
Other locations
This PC
Add a Place
Browse
C: > Users > alexa > AppData > Local > Mailbird > Store > A > 7969
Save
Excel Workbook (*.xlsx)
Excel Macro-Enabled Workbook (*.xlsm)
Excel Binary Workbook (*.xlsb)
Excel 97-2003 Workbook (*.xls)
CSV UTF-8 (Comma delimited) (*.csv)
XML Data (*.xml)
Single File Web Page (*.mht, *.mhtml)
Web Page (*.htm, *.html)
Excel Template (*.xltx)
Excel Macro-Enabled Template (*.xltm)
Excel 97-2003 Template (*.xlt)
Text (Tab delimited) (*.txt)
Unicode Text (*.txt)
XML Spreadsheet 2003 (*.xml)
Microsoft Excel 5.0/95 Workbook (*.xls)
CSV (Comma delimited) (*.csv)
Formatted Text (Space delimited) (*.prn)
Text (Macintosh) (*.txt)
Text (MS-DOS) (*.txt)
CSV (Macintosh) (*.csv)
CSV (MS-DOS) (*.csv)
DIF (Data Interchange Format) (*.dif)
SYLK (Symbolic Link) (*.slk)
Excel Add-in (*.xlam)
Excel 97-2003 Add-in (*.xla)
PDF (*.pdf)
XPS Document (*.xps)

allele_frequencies_straf

CSV (Comma delimited) (*.csv)

Then, you can upload your data and use it as described for the STRIDER database.

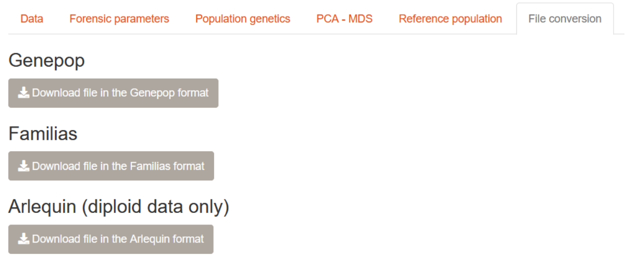
Chapter 6

File conversion

As STRAF is a web application and can be used simultaneously by multiple users, computing resources are limited. Therefore, the most computationally intensive analyses are not available in STRAF. In order to ease the path to other software, file conversion utilities have been implemented. It is possible to convert the input file to the **Genepop**, **Arlequin** and **Familias** formats. They are all available in the **File conversion** tab of the application.

6.1 How to convert a file in STRAF

Once you have imported your file, it is straightforward to conevrt it to another format. You can go to the **File conversion** tab and click on one of the following buttons to download your genotypes in another format.


Data
Forensic parameters
Population genetics
PCA - MDS
Reference population
File conversion
Genepop
Download file in the Genepop format
Familias
Download file in the Familias format
Arlequin (diploid data only)
Download file in the Arlequin format

6.2 Genepop and Arlequin formats

Genepop and **Arlequin** softwares implement several population genetics methods, including ones that are part of standard forensics practice:

- linkage disequilibrium computation
- Hardy-Weinberg tests

STRAF currently implements these computations, however the ones implemented in Genepop are overall more reliable as they can rely on more permutations. They are overall preferable to the HW and LD tests implemented in STRAF.

6.3 Familias

Here a file containing allele frequencies is created. This file can be used in Familias to provide allele frequencies reference.