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 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 ability to characterise the DNA of individuals. These technologies 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 occured and changed 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 are also **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 be and contain more or less repetitive units.

Short Tandem Repeats (**STRs**) a type of genetic polymorphism consisting in short sequences from 2 to 7 base pairs that are repeated a certain number of time. As the number of repeats varies among individuals, characterizing their length can be useful to identify individuals.

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.

Digression - Why STRs and not other markers?

Comparing to whole-genome sequencing. They have a high **power of discrimination** because they have a much higher diversity (their mutation rate is higher than other common markers).

STRAF and population data analysis

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

Chapter 1

Importing data

Work in progress.

1.1 STR data

- Observed values: genotypes for each individual, at each locus.
- Potentially two values observed per individual and per locus, if diploid markers.
- Value = can be anything but tipically correspond, for STR markers, to the length.
- Point alleles

1.2 Input data format

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

1.3 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:

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)`

1.4 Uploading the data to STRAF

Coming soon.

1.5 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"

1.6 Having a first look at the data

Chapter 2

Forensic parameters

In this chapter, we'll introduce a few equations. 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

2.2 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 have a random match in a population by looking at a singly locu, it means that our two samples have the same genotype at that locus. In terms of probability, this will happen as many times as we . The probability of sampling a specific genotype in the population is equal to its frequency.

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

Genetic diversity (GD), or 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

Coming soon.

2.5 Polymorphism Information Content (PIC)

Formula

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

Coming soon.

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

Coming soon.

Chapter 3

Population genetics indices

3.1 Population genetics concepts

Hardy-Weinberg equilibrium

A population is considered at Hardy-Weinberg equilibrium when

Why is it important to check? 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.

- inbreeding
- population structure
- locus is under selection. This is very unlikely that STR loci as they are supposed to evolve neutrally. However, they could be found near loci under selection

In forensics, we need to assume HWE as indices computed (for example, a match probability) would be biased if

- Population structure

Individuals closer from each other are in general more likely to mate with each other.

3.2 How to compute population genetics parameters in STRAF

3.3 Population genetics parameters

- Heterozygosities
- F-statistics: FIS and FST

One concept, multiple estimators.

Several **estimators** of Fst exist (for example, Weir and Cockerham's, Nei's, Hudson's FST). It's like if each population geneticist decided to develop their own estimator! Why is that? In statistics, what we call an **estimator** is. 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.

3.4 What is linkage disequilibrium?

Coming soon.

3.5 How to compute LD in STRAF

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.

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.

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

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.

5.1 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.

5.2 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.

5.3 Familias

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