The STRAF Book

Alexandre Gouy and Martin Zieger

2021-05-02

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 forensics 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 are the same question, simply phrased differently.

As an example, DNA profiling, used in criminal investigations or parental testing, aims at matching different DNA samples and understanding how related are some samples 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. Hence you can now

imagine why the two fields are linked: they both want to **understand and quantify** the **relatedness** of a set of samples.

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 Bern, Switzerland, Martin Zieger came to visit a population genetics lab, where Alexandre Gouy 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. 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.

STRAF’s story highlights the importance of communication between fields.

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

WORK IN PROGRESS

Essential concepts

- DNA
- Genotypes and phenotypes
- Genetic variation, polymorphism
- Markers of polymorphism
- Short Tandem Repeats

Polymorphism and forensics

- Goals: typing and matching
- Paternity and maternity testing, suspect
- Matching require reference populations
- Allele frequencies known and reported

Digression - Why are STR still so popular?

Comparing to whole-genome sequencing.

Data analysis

- Data analysis in forensic genetics
- STRAF’s scope

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:

- Screen capture Excel

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)`

- Screen capture Save As

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.

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"

Chapter 2

Forensic parameters

WORK IN PROGRESS

In this chapter, we'll introduce a few equations. Do not be afraid! Each of them will be translated to plain English.

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

Coming soon.

2.2 Power of Discrimination (PD)

Formula

$$PD = 1 - PM$$

Interpretation

Coming soon.

2.3 Gene diversity

Genetic diversity (GD), or expected heterozygosity (H_{exp}), is computed using the following estimator:

Formula

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

Interpretation

Coming soon.

2.4 Polymorphism Information Content (PIC)

Polymorphism Information Content (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.

Formula

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

Interpretation

Coming soon.

2.5 Power of Exclusion (PE)

Formula

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

Interpretation

Coming soon.

2.6 Typical Paternity Index (TPI)

Formula

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

Interpretation

Coming soon.

Chapter 3

Population genetics indices

WORK IN PROGRESS

3.1 Population genetics concepts

- Hardy-Weinberg equilibrium
- Population structure

3.2 Indices

- Heterozygosities
- F-statistics
- 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.

Chapter 4

Multivariate statistics

WORK IN PROGRESS

4.1 Principal Component Analysis (PCA)

PCA is a method of dimentionality reduction. What is does it that it captures most of the variation in our data and tries to project it onto a small number of new variables called components.

This is a useful method to capture variation from a large number of variables and allows to discover hidden pattern 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.

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

- PCA plot on Pemberton data.

- PCA plot on Y haplogroups.

Interpreting PCA results

- Beware of the influence of sample size on the results.

4.2 Multidimensional Scaling (MDS)

MDS is conceptually similar to PCA. One of the main differences is that it takes different types of data as input. Pairwise distances between data points. In forensics practice, it is often used to compare populations and not individuals. It could be run on a pairwise FSTs or other genetic distances.

- MDS plot on Pemberton data.

Interpreting MDS results

Chapter 5

File conversion

WORK IN PROGRESS

As STRAF is a web application and can be used simultaneously by multiple users, computing resources are limited. Therefore more 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 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, however the ones implemented in Genepop as they can rely on more permutations and are overall preferable to the HW and LD tests implemented in STRAF.

5.2 Familias

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