Population Genetics Data Analysis

Ijaz Anwar[1], Tacha Hicks Champod[1], and Prof. Franco Taroni[1]

[1]School of Criminal Justice, University of Lausanne, 1015 Dorigny, Switzerland.

Contents

List of Figures

1 Introduction

This tutorial provides a concise and practical introduction to the population genetic data analyses for the forensic practitioners and students, using the conventional softwares (e.g. Excel spreadsheets) as well as latest data analysis tools (e.g. in R). The participant is expected to have the basic knowledge of R and R Studio and should have installed the required packages.

After this tutorial, the participants will be able to perform;

- Input file preparation
- Forensic statistics
- Population statistics
- Admixture studies
- Population affiliation

In the next section, we will quickly take a look at some basics of the Forensic Biology, Population Genetics and its applications.

2 Forensic Biology

Forensic Biology deals with the crimes where the biological evidence is involved, and the DNA typing technology can be used effectively. The incredible power of DNA technology has been used as an identification tool that brought substantial reformations in criminal justice system and greatly benefited the law enforcement community. DNA analysis has been effective in securing convictions in hundreds of violent crimes, from homicides to assaults. It has also helped to eliminate suspects and has led to the exoneration and release of previously wrongfully convicted individuals [1]. During the last century, the criminal justice system has seen a major advancement due to the development and application of forensic techniques to solve the crime.

Over the past 40 years, the forensic biology field has made great strides. The multilocus probe analysis by restriction fragment length polymorphism (RFLP) analysis developed by Alec Jeffreys [2] explained genetic differences among individuals [3]. DNA typing quickly progressed to the use of single locus variable number of tandem repeat (VNTR) loci by RFLP analysis [4]. Single locus analysis offered greater sensitivity, increased species specificity, and standard statistical interpretations compared with the multilocus approach. VNTR typing was adopted by many crime laboratories in some countries and was the mainstream system of the late 1980's through most of the 1990's.

Human STRs were first reported in 1989 [5]. These STRs were discovered to occur within or between genes along human chromosomes. The STR analysis of the human autosomal DNA provides unique information on the genetic diversity of the populations [6]. STR markers were first described as an effective tool for human identity in the early 1990s [7]. The STR markers used in human identity testing are primarily tetranucleotide repeats [8]. STR typing is more tolerant to the use of degraded DNA templates than other methods of individual identification.

2.1 Forensic statistics

The applicability of STR markers in a certain population depends on "How much informative these markers are for that population". To check this, there are some statistical parameters which describe the efficiency of the chosen STR marker multiplex to be used as a human identification tool.

2.1.1 Allele frequency distribution

It is the frequency distribution of alleles on a certain marker. It describes the variability of the chosen marker to produce maximum number of alleles on a certain locus so that the locus has low match probability.

$$AlleleFreq = \frac{2Nxx+Nxy}{2N}$$

Where Nxx is the number of homozygotes for allele X (XX), and Nxy is the number of heterozygotes containing the allele X and Y. N is the number of samples in the population. Allele frequency can also be calculated by simple count of the proportion of different alleles.

2.1.2 Observed heterozygosity

Observed heterozygosity is a simple measure of the proportion of heterozygotes in the population at a given locus.

$H_o \ = \ \frac{Hets}{N}$

2.1.3 Expected Heterozygosity

It is the proportion of heterozygotes expected under random mating population (Hardy-Weinberg Equilibrium) at a given locus.

$H_e \ = \ 1 - \Sigma pi^2$

where p_i is the allele frequency of the allele i.

The difference will be larger for samples that differ from Hardy–Weinberg proportions markedly.

2.1.4 Profile probability

It is the probability that any two selected persons have the same DNA profile.

$pM \ = \ \Sigma_{i=1}^{n} \ Px_i^2$

Where i is the allele at a locus and Px_i is the frequency of ith allele.

The combined match probability is the product of all match probabilities at each locus, assuming the Hardy-Weinberg Equilibrium.

2.1.5 Power of discrimination

It is simply the inverse of the match probability pM. And can be determined by this formula;

$PD \ = \ 1 - pM$

While the combined power of discrimination is the product of all power of discriminations at each locus.

2.1.6 Power of exclusion

It is the fraction of individuals having a DNA profile that is different from that of a randomly selected individual in a given population. It can be determined by this formula;

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

While the combined power of exclusion is the product of all power of exclusions at each locus.

2.1.7 Paternity Index

It is the probability that the person being tested is the biological father, rather than a randomly selected individual in a given population. It can be obtained by using this formula;

$$PI \;=\; \frac{1}{\left(2\Sigma_{i=1}^{n}\ {pi}^{2}\right)}$$

2.1.8 Polymorphic information content

It is the measure of the usefulness of an STR marker and is determined by the following formula;

$$PIC_i \;=\; 1 - \Sigma_{1}^{n}\ {pi}^{2} \;-\; \left(\Sigma_{1}^{n}\ {pi}^{2}\right)^{2} \;+\; \Sigma_{1}^{n}\ {pi}^{4}$$

where n is the number of alleles and pi is the allele frequency of the ith allele.

3 Population Genetics

Population genetics is the study of variation observed within a population group or among different population groups in terms of allele and genotype frequencies. Great genetic variation exists within groups at the individual nucleotide level. [9]. The modest description of variation is the frequency distribution of genotypes. The most important factor that influences this frequency distribution is the number of heterozygote individuals observed in a population. Over time, isolated populations diverge from one another, each losing heterozygosity due to inbreeding. This variation on the basis of heterozygosity within and among the populations was first explained by Sewall Wright and then these parameters of differentiation were further explained by Bruce Weir for short tandem repeat (STR) markers. These parameters are comprised of co-efficient of inbreeding ($\mathrm{F_{IS}}$ or f), co-efficient of co-ancestry ($\mathrm{F_{ST}}$ or θ) and the co-efficient of relationship ($\mathrm{F_{IT}}$ or F). These indices are well illustrated by Fig 1.

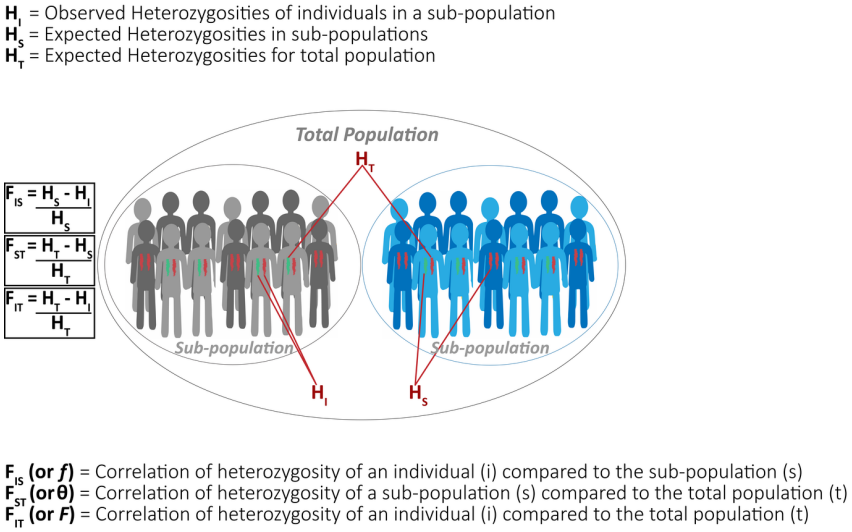

Figure 1: Description of Sewall Wright's F-Statistics and its components. Later on, Weir &Cockerham simplified the notation into f, θ and F.

4 Data Analysis

The participants are required to download the data files and the excel add-in (Genalex file) from the server\escetu\courses\modules\TP génétique\PGDA.

4.1 GenAlEx

GenAlEx - Genetic Analysis in Excel [10] is an Excel add-in that can be freely downloaded from [11]. GenAlEx can be installed into the Excel ribbon or can be used directly into the Excel spreadsheet and provides very easy-to-use functions to handle the genetic data and simple analyses of forensic and population genetics interest. It is also useful to export the genotype data in a variety of other types of input files, that can be directly imported into other softwares (including R).

4.2 R Packages

The participants are advised to install following R packages;

- poppr [12]
- adegenet [13]
- ggplot2 [14]
- dplyr [15]
- FORSTAT (A shiny app available at [16]

4.3 Description of data

The server folder contains three datasets containing DNA profiles collected from the individuals belonging to different populations of three different countries. Here is the description of each dataset;

4.3.1 US Dataset

This dataset contains genotypes of 1036 individuals at 29 autosomal STR loci from four populations living in the United States of America and is available at [17]. This data includes STR profiles from African American (n = 342), Caucasian (n = 361), Hispanic (n = 236) and Asian (n = 97) population.

4.3.2 China Dataset

This dataset contains genotypes from 1814 individuals at 15 autosomal STR loci [18] from Manchu (n = 296), Mongol (n = 507), Kyrgyz (n = 550) and Uzbek (n = 461) populations.

4.3.3 Pakistan Dataset

This dataset contains genotypes from 520 individuals from Punjabi (n = 130), Saraiki (n = 130), Sindhi (n = 130) and Pakhtun (n = 130) populations of Pakistan [19].

4.4 Input file preparation

We will start the data analyses by importing the dataset in GenAlEx , and then some statistical tests will be performed. In order to import the DNA profiles in GenAlEx, the participants are advised to follow the instructions given in Fig 2.

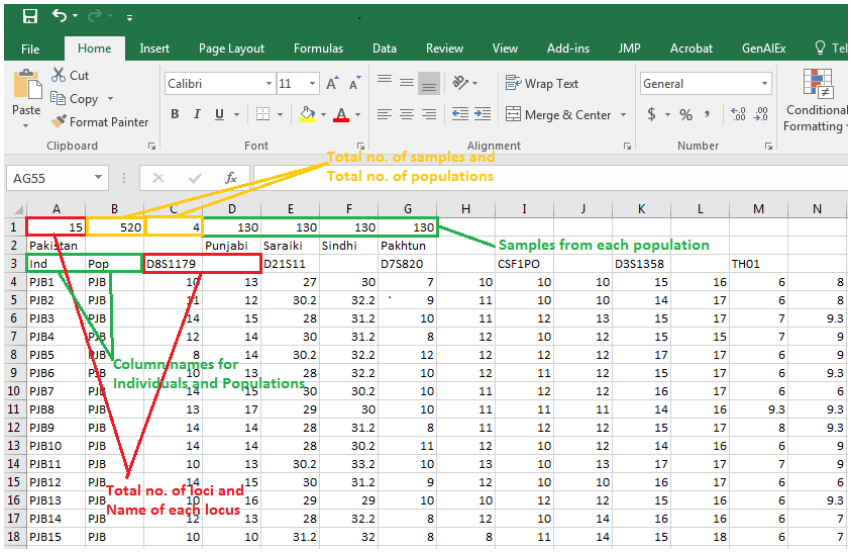

Figure 2: Input File example for GenAlEx spreadsheet.

Here are some explanations;

- The red boxes indicate the total number of loci used and the names of each loci (in two separate columns)
- The yellow boxes indicate the total number of samples collected and number of populations
- The green boxes indicate the identifiers (Ind) for individual names and (Pop) for population name columns
- The top green column indicates number of samples collected from each population.

After the data input, the only formatting to do with the data is to multiply all the alleles with 10, This is done to get rid of the decimal point within the entries. The data will still be good enough to represent the alleles as the number of repeats.

Note: In case, if you face problems in installing or using GenAlEx add-in, then it is recommended to go to properties of the add-in file and Unblock or add it into the trusted sites of Excel.

4.5 Forensic statistics analyses

After formatting the input file, we are ready to perform the statistical analyses on the available datasets. First thing to start with, is the calculation of allele frequencies of all the populations and making separate databases of each population.

4.5.1 Allele frequency calculations

In order to calculate the allele frequencies of each population, we will perform the commands in this sequence;

GenAlEx > *Analysis Options* > *Frequency Based* > *Frequency*

Initiating this analysis, a window will appear, displaying the brief information of the input file. On verifying this information (by clicking OK) another window will pop-up. Please, ensure that the "Frequency by pop" and "Graph All Loci" options are checked, and then click OK. The output of this analysis will give the allele frequency of each locus in all the studied populations.

In order to calculate other statistical parameters of forensic interest, please go to;

Data Management > *Import-Export* > *Export* > *GenePop*

Please, uncheck the "To Text File" and click OK. The participants are required to save this file with a *.gen* extension with the name.

4.5.2 FORSTAT Application

Please go to the homepage of FORSTAT server, available here [20] and upload the saved file with *.gen* extension (Load and view data) and Perform computations (Compute) and once the computations are complete then click Output.

On the left, you will have the access to all the options to perform further detailed tests of forensic importance and explore the data.

In the last option, you will have the possibility of downloading all the statistics and graphs.

4.6 Population statistics analyses

4.6.1 Package installation

Now the participants are requested to open RStudio and install all the required packages, by running this command.

```
install.packages("poppr", "adegenet", "hierfstat", "ggplot2", "dplyr")
```

Activate the installed packages.

```
library("poppr")
library("adegenet")
library("hierfstat")
```

```
library("ggplot2")
library("dplyr")
```

4.6.2 Uploading data

To upload the data into RStudio, we can use both GenAlEx input file as well as the *.gen* (Genepop) file exported from GenAlEx. Here, we will be using GenAlEx file.

```
pk <- read.genalex("PkDataset.csv", ploidy = 2, geo = FALSE,
        region = FALSE, genclone = FALSE, sep = ",", recode = FALSE)
```

To verify whether our input was successful or not, we can check the *class* of the input file.

```
class(pk)
```

Or simply

```
pk
```

It is always better to define colors of your choice in the start, so that a specific color can be used for a certain population at each level of the analysis.

```
myCol <- c("darkblue","red","green","orange")
```

4.6.3 Basic statistics

Now, we will perform our first analysis, that is *basic.stats* from the *hierfstat* package, which will perform all the basic tests of population genetics and you can save the results as an object in the environment and explore the results.

```
bs <- basic.stats(pk, diploid = TRUE, digits = 4)
```

You can check the allele frequencies at each locus for all the populations;

```
bs$pop.freq
```

Co-efficient of inbreeding (F_{IS} or f) at each locus and for each population;

```
bs$Fis
```

Observed Heterozygosities at each locus;

```
bs$Ho
```

All the above said statistics (and other) calculated and summarized per locus;

```
bs$perloc
```

And, Overall statistics;

```
bs$overall
```

4.6.4 Admixture studies

To check whether the studied populations are genetically at a larger distance from each other or they are admixed, we will perform Discriminant Analysis of Principal Components (DAPC) by using *a deg enet* package.

First, just to have a look, what is DAPC and what does it do, we will explore the help file;

```
?dapc
```

Now, we will perform our DAPC, by using the following code and will store it as an object;

```
dapc1 <- dapc(pk, pop=NULL, n.pca=NULL, n.da=NULL, scale=FALSE,
              truenames=TRUE, var.contrib=TRUE, var.loadings=FALSE, pca.info=TRUE,
              pca.select=c("nbEig","percVar"), perc.pca=NULL)
```

Just to explore what a DAPC object is, we will check its structure;

```
dapc1
```

Now, we can plot our graph, and see what is the relevant information given by this graph;

```
scatter(dapc1, cstar=1, col = myCol, pch=20, solid=.4, cex=1.8, scree.pca=TRUE,
        posi.pca="bottomleft", leg=TRUE, txt.leg=c("Punjabi", "Saraiki", "Sindhi", "P
```

You can customize your graph in many different ways to make it more relevant and useful to your needs;

```
scatter(dapc1, scree.da=FALSE,  bg="white",
        pch=20, cstar=0, col=myCol, scree.pca=TRUE,
        posi.pca="bottomright")
```

Further customization;

```
scatter(dapc1, ratio.pca=0.3, bg="white", pch=20,  cell=0,
        cstar=0, col=myCol, solid=.4, cex=1.8, clab=0,
        mstree=TRUE, scree.da=FALSE, scree.pca=TRUE, posi.pca="bottomright",
        leg=TRUE, txt.leg=c("Punjabi", "Saraiki", "Sindhi", "Pakhtun"))
```

Even further exploitations;

```
scatter(dapc1, cell=0, pch=18:23, cstar=0, mstree=TRUE, lwd=2, lty=2)
```

or

```
scatter(dapc1, label.inds = list(air = 2, pch = NA))
```

```
scatter(dapc1, cell=2, pch="", cstar=0, posi.da="top",
        axesel=FALSE, col=terrain.colors(10))
```

Note: Do not forget to change the names of the populations, if you are using a different dataset.

4.6.5 Population affiliation

To calculate the membership probability of each individual with a certain population, we perform *compoplot*, which gives us the idea about the genetic composition of each individual and its probable affiliation with a given population.

```
compoplot(dapc1, posi="bottomleft",
          txt.leg=c("Punjabi", "Saraiki", "Sindhi", "Pakhtun"), lab="",
          ncol=1, xlab="individuals", col=myCol)
```

Note: Do not forget to change the names of the populations, if you are using a different dataset.

At the end, the participants are required to go to the *adegenet* server and explore the possibilities provided by the shiny app.

```
adegenetServer(what = "DAPC")
```

References

[1]Innocence Project - Help us put an end to wrongful convictions!, (n.d.). https://www.innocenceproject.org/.

[2]A.J. Jeffreys, V. Wilson, S.L. Thein, Hypervariable 'minisatellite' regions in human DNA, Nature. 314 (1985) 67–73. https://doi.org/10.1038/314067a0.

[3]A.J. Jeffreys, V. Wilson, S.L. Thein, Individual-specific 'fingerprints' of human DNA, Nature. 316 (1985) 76–79. https://doi.org/10.1038/316076a0.

[4]N. ROYLE, Clustering of hypervariable minisatellites in the proterminal regions of human autosomes, Genomics. 3 (1988) 352–360. https://doi.org/10.1016/0888-7543(88)90127-9.

[5]J.L. Weber, P.E. May, Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction., Am J Hum Genet. 44 (1989) 388–96.

[6]K. Thangaraj, G. Chaubey, V.K. Singh, A.G. Reddy, P.P. Pavate, L. Singh, Genetic Profile of Nine Autosomal STR Loci Among Halakki and Kunabhi Populations of Karnataka India, Journal of Forensic Sciences. 51 (2006) 190–192. https://doi.org/10.1111/j.1556-4029.2005.00038.x.

[7]A.J. Jeffreys, M. Turner, P. Debenham, The efficiency of multilocus DNA fingerprint probes for individualization and establishment of family relationships, determined from extensive casework., Am J Hum Genet. 48 (1991) 824–40.

[8]J.R. Collins, R.M. Stephens, B. Gold, B. Long, M. Dean, S.K. Burt, An exhaustive DNA micro-satellite map of the human genome using high performance computing., Genomics. 82 (2003) 10–9.

[9]G. Barbujani, A. Magagni, E. Minch, L.L. Cavalli-Sforza, An apportionment of human DNA diversity., Proc Natl Acad Sci U S A. 94 (1997) 4516–9.

[10]R. Peakall, P.E. Smouse, GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research–an update, Bioinformatics. 28 (2012) 2537–2539. https://doi.org/10.1093/bioinformatics/bts460.

[11]Welcome to the GenAlEx 6.5 website!, (n.d.). https://biology-assets.anu.edu.au/GenAlEx/Welcome.html.

[12]Z.N. Kamvar, J.F. Tabima, N.J. Grünwald, Poppr: an R package for genetic analysis of populations with clonal partially clonal, and/or sexual reproduction, PeerJ. 2 (2014) e281. https://doi.org/10.7717/peerj.281.

[13]T. Jombart, adegenet: a R package for the multivariate analysis of genetic markers, Bioinformatics. 24 (2008) 1403–1405. https://doi.org/10.1093/bioinformatics/btn129.

[14]H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag New York, 2016.

[15]Package 'dplyr', (n.d.). https://cran.r-project.org/web/packages/dplyr/dplyr.pdf.

[16]P.G. Ristow, M.E. D'Amato, Forensic statistics analysis toolbox (FORSTAT): A streamlined workflow for forensic statistics, Forensic Science International: Genetics Supplement Series. 6 (2017) e52–e54. https://doi.org/10.1016/j.fsigss.2017.09.006.

[17]C.R. Hill, D.L. Duewer, M.C. Kline, M.D. Coble, J.M. Butler, U.S. population data for 29 autosomal STR loci., Forensic Sci Int Genet. 7 (2013) e82–3.

[18]X. Zhan, A. Adnan, Y. Zhou, A. Khan, K. Kasim, D. McNevin, Forensic characterization of 15 autosomal STRs in four populations from Xinjiang China, and genetic relationships with neighboring populations, Scientific Reports. 8 (2018). https://doi.org/10.1038/s41598-018-22975-6.

[19]I. Anwar, S. Hussain, A.U. Rehman, M. Hussain, Genetic variation among the major Pakistani populations based on 15 autosomal STR markers, International Journal of Legal Medicine. (2018). https://doi.org/10.1007/s00414-018-1951-0.

[20]FORSTAT, (n.d.). https://doi.org/https://doi.org/10.1016/j.fsigss.2017.09.006.