Chapter 19

Inference of Ancestry in Forensic Analysis II: Analysis of Genetic Data

Carla Santos, Chris Phillips, A. Gomez-Tato, J. Alvarez-Dios, Ángel Carracedo, and Maria Victoria Lareu

Abstract

Three approaches applicable to the analysis of forensic ancestry-informative marker data—*STRUCTURE*, principal component analysis, and the *Snipper* Bayesian classification system—are reviewed. Detailed step-by-step guidance is provided for adjusting parameter settings in *STRUCTURE* with particular regard to their effect when differentiating populations. Several enhancements to the *Snipper* online forensic classification portal are described, highlighting the added functionality they bring to particular aspects of ancestry-informative SNP analysis in a forensic context.

Key words Genetic ancestry, Reference data, SPSmart browser, Bayesian methods, *STRUCTURE*, *Snipper*, PCA

1 Introduction

Classifying individuals into populations is often useful in population genetics applications. But the definition of populations is commonly subjective, based on linguistic, cultural, or physical characters, as well as the geographical location of sampled individuals. This is a sensible way of incorporating diverse types of information but it may be difficult to know whether a given assignment of individuals to populations based on these subjective criteria matches an assignment in genetic terms. For this reason, it can be useful to confirm that the subjective classifications are consistent with genetic information and hence appropriate for the intended classification regime [1, 2]. A possible approach starts with a set of predefined populations and then classifies individuals

Electronic supplementary material: The online version of this chapter (doi:10.1007/978-1-4939-3597-0_19) contains supplementary material, which is available to authorized users.

William Goodwin (ed.), *Forensic DNA Typing Protocols,* Methods in Molecular Biology, vol. 1420,
DOI 10.1007/978-1-4939-3597-0_19,

of unknown origin into these populations. This involves sampling DNA from members of potential source populations to estimate allele frequencies in each population at a series of unlinked loci. Allele frequencies can be used to compute a set of likelihoods that a given profile of genotypes originates in each population. These likelihoods allow the assignment of individuals of unknown origin to populations based on the highest likelihood ratio [2].

Bayesian population analysis methods infer a simple relationship between the allele frequencies of a population and the allele frequencies observed in the individuals identified as part of that population. An advantage of such methods is that prior information about the samples can be used to progress the analysis. But the ability to differentiate populations in a sample set can be limited when applying a small number of samples and/or markers. Two valid approaches for comparing profiles from forensic casework DNA to reference population data will be considered here: a systematic Bayesian clustering approach (*STRUCTURE* software) and a naïve Bayesian likelihood ratio (LR) based calculator (underlying the *Snipper* web portal). *STRUCTURE* is a flexible approach—different types of markers such as STRs, SNPs, and indels can be readily combined in the same genotype input file (*Snipper* also allows such flexibility but the systems are not yet implemented). However *STRUCTURE* analysis of single profiles, typical of forensic testing, is not so straightforward since the whole set of parental data plus the unknown profile must be re-analyzed in combination each time and this can be both time-consuming and cumbersome to perform for a small number of casework samples in turn. For this reason *Snipper* (http://mathgene.usc.es/snipper/) was developed to provide a simple alternative for making ancestry assignments of single profiles in real time. Both *STRUCTURE* and *Snipper* use a Bayesian approach which, put simply, computes likelihood of membership to each class (in this case ancestry) using the observed frequency of variables in each class (in this case allele frequencies). The difference between both methods lies in how the likelihood is computed (more information about these algorithms is detailed in [2, 3]). Therefore both algorithms require reference data to calculate allele frequencies for comparison to alleles recorded in profiles of unknown origin. In the case of *Snipper*, the reference data allows construction of training sets for calculation of allele frequencies and these can comprise ready-to-use fixed five-population group data (African, European, East Asian, Native American, and Oceanian) already in place for 34 SNPs [3, 4] and/or 46 AIM-indels [5]. It can alternatively consist of end user's own data for any populations and binary marker set where reference genotypes are available, which can then be uploaded as a custom data set. Each algorithm makes the same prior assumption, often untested: that the variables, i.e., the component loci, are independent. For this reason, uniparental data (in

the form of haplotypes where all markers are linked) is not readily incorporated into either analysis system, though *STRUCTURE* has scope for the analysis of linked loci.

As the number of populations increases, the number of dimensions needed to visually represent the pairwise genetic distances also increases. The main idea of multivariate analyses is to help to represent, in a comprehensive way, those multiple dimensions. This is done through the reduction of the dimensionality of a data set composed of a large number of interrelated variables maintaining the maximum proportion of the variation present on that data [1]. Principal component analysis (PCA) is a commonly used multivariate analysis method, especially as an exploratory tool and to summarize genetic similarities and differences between groups of populations. This is possible through the transformation of those variables into a new set of metrics (principal components: PCs) that are not related and can be ordered in a way that the first PCs retain most of the variation present in the original data—the graphical representation of the first two or three PCs summarizes as much of the variation as possible in a comprehensive, graphical way [1, 6]. In the graphics that PCA generates, individuals are represented by points distributed according to their coordinates in two-way or three-way PC comparisons (two or three dimensional plots, respectively). PCA can also be used to represent the relation of an unknown study sample with a set of reference population samples, i.e., the study sample will be represented by a point superimposed onto the PCA plot of the reference population samples.

2 Materials

1. *SPSmart* browser home: http://spsmart.cesga.es/
2. *SPSmart* SNP*for*ID 52-plex and 34-plex variability browser: http://spsmart.cesga.es/snpforid.php
3. Entire genome interface for exploring SNPs (ENGINES) a 1000 Genomes variability browser: http://spsmart.cesga.es/engines.php?dataSet=engines
4. pop.STR: http://spsmart.cesga.es/popstr.php
5. *Snipper* portal: http://mathgene.usc.es/snipper/
6. *STRUCTURE* software: http://pritchardlab.stanford.edu/structure.html
7. *Structure harvester*: http://taylor0.biology.ucla.edu/structureHarvester/#
8. CLUster Matching and Permutation Program (*CLUMPP* software): http://www.stanford.edu/group/rosenberglab/clumpp.html

9. *distruct* software: http://www.stanford.edu/group/rosenberglab/distruct.html
10. For more information about *R* software: http://www.r-project.org/ [7].

3 Methods

3.1 Collection of Ancestry Reference Data with the SPSmart Browser

The statistical analysis of a profile requires reference training sets, i.e., parental populations of interest used to classify casework profiles. Collection of such data previously required locus-by-locus scrutiny of dbSNP or HapMap SNP databases [8], but fortunately SPS (SNPs for Population Studies) makes this task much more straightforward for any number of AIM-SNPs as well as up to 52 STRs in routine forensic use.

SPS comprises the following genomic variability browsers:

- SPSmart home: http://spsmart.cesga.es/ [9].
- SPSmart SNP*for*ID 52-plex and 34-plex variability browser: http://spsmart.cesga.es/snpforid.php [10].
- ENGINES (Entire Genome Interface for Exploring SNPs) a 1000 Genomes variability browser enabling a review of all SNP sites found from 1092 complete genome sequences (1000 Genomes Phase I): http://spsmart.cesga.es/engines.php?dataSet=engines [11].
- pop.STR: http://spsmart.cesga.es/popstr.php [12].

SPSmart [9] is a simple pre-processing engine that includes five different population-based genotype databases: (1) 1000 Genomes Phase I May 2011; (2) HapMap Release #28; (3) Perlegen complete data set; (4) HGDP-CEPH Stanford study; and (5) HGDP-CEPH NIH-Michigan study (*see* **Note 1**). *SPSmart* also generates common population genetics indices such as allele frequencies, heterozygosity, F_{ST}, or *In* (summarized in the downloads tab of each query).

1. Choose the database(s) to search or choose SEARCH in the SNP*for*ID "global map" homepage.
2. Choose the populations to merge into groups by ticking selections up to a maximum of five. If opting to review multiple databases, e.g., HapMap and Perlegen, only one population grouping can be made. Populations are already arranged into sets of genetic diversity based on Rosenberg's original analyses of HGDP-CEPH populations [3, 13], though note that Eurasians are subdivided into European, South Asian, and Middle Eastern subgroups.
3. Add the SNP RefSeq (rs-number) identifiers in the search by SNP window. Search by chromosome region or gene name is also available. Click the "next" button below.

4. Of the filter options presented, MAF is minimum allele frequency and *In* Rosenberg's ancestry informativeness metric [14]. When reviewing SNP data from multiple databases, it is best not to tick option: "Filter SNPs not genotyped on every compared dataset".
5. When SNPs are not found from a query, a *message tab* with the missing rs-numbers appears.
6. In the *downloads* tab, genotype files are available ready to download, copy and paste into Excel (*see* **Note 2**) or notepad. The recommended steps being: download, choose all, copy into Excel, transpose the data into rows = samples and columns = SNPs (in edit menu: "copy" > select new cell > "paste special" > select: "transpose"). This must be completed for each group while taking care to label each set of sample rows with the appropriate description, e.g., African, South Asian, etc. Another option is to query and download all population groups at once—follow the steps previously described in this point for the genotypes and then download the sample list file. This file includes sample, subpopulation, and population group information. Open, select all, and paste in the genotypes Excel file—confirm that the samples are in the same order and remove the duplicated column with sample name (Fig. 1).

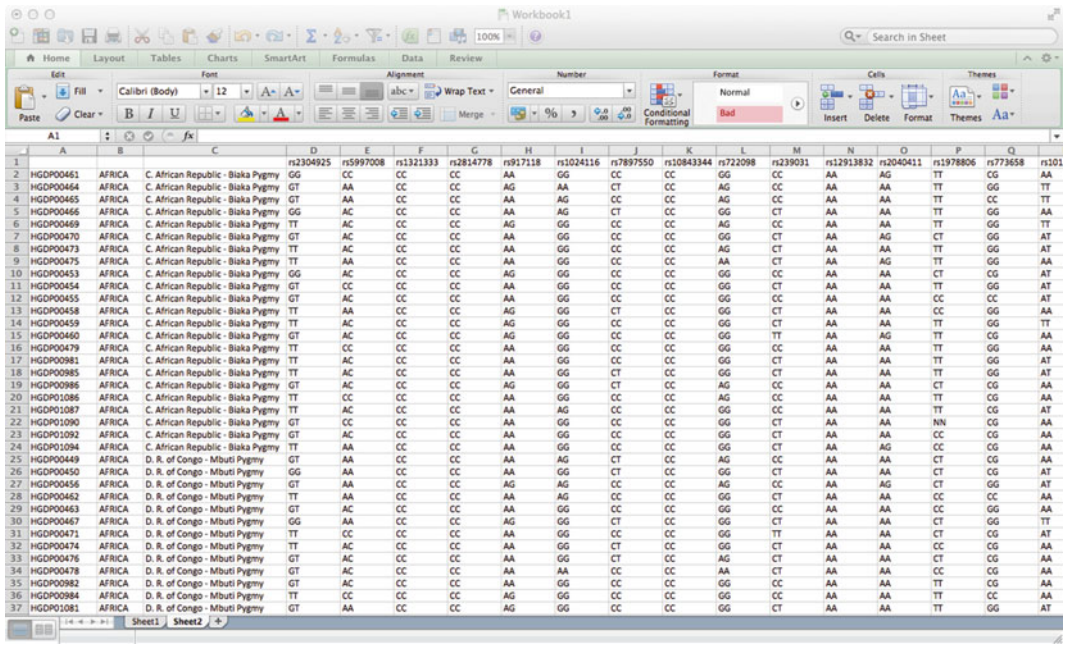

Fig. 1 Example of a reference ancestry genotype data file obtained from *SPSmart*. The data obtained from *SPSmart* was reorganized (original data was transposed so that samples are now organized in rows and markers in columns) and population information (downloaded from *SPSmart* in a separate file) was added

7. The genotypes can be formatted for input to *STRUCTURE*, *Snipper*, or PCA custom data analysis (outlined later). SNaPshot genotypes may need checking against the reference data if they come from different typing platforms, e.g., a SNaPshot C/T SNP may be an A/G SNP in HapMap requiring base inversions of one dataset. For this reason symmetrical SNPs (C/G and A/T) require particular care (*see* **Note 3**).

3.2 STRUCTURE Software

3.2.1 Background on STRUCTURE Analysis

STRUCTURE uses genotypic data of several loci to: (1) infer population structure; (2) identify subpopulations (subsets of samples with distinct allele frequencies); (3) assign individuals to subpopulations (based on probabilities); and (4) study admixture between populations. It uses a population structure model where studied samples represent a mixture of K unknown populations—each characterized by unknown allele frequencies for the loci used and where these are assumed to be in Hardy-Weinberg equilibrium (HWE) and independent (not in linkage disequilibrium). The objective is to classify individuals into K clusters in a way that deviations from HWE and independence are minimized.

Assuming HWE and independence in each subpopulation, the probability that the genotype of an individual belongs to subpopulation k is given by the product of the allele frequencies. Using Bayes rule (*see* **Note 4**) it is possible to calculate the probability of an individual belonging to subpopulation k. If allele frequencies in a population were known in advance, it would be easy to allocate individuals. Equally, if individual allocations were known it would be easy to estimate the frequencies. In practice, we do not know either, but using a Markov Chain Monte Carlo (MCMC) algorithm (*see* **Note 5**) it is possible to obtain a sensible estimate of both.

The estimation of the optimum K value is performed in an independent way: for each simulation a posterior probability value $Pr(K)$ is calculated. In general, for K values under the optimum, $Pr(K)$ is low but it tends to stabilize with higher K values, so a *plateau* is commonly observed. When several K values have similar $Ln\ Pr(K)$ estimates, the smallest of them is usually the most appropriate estimate—generally corresponding to the inflection point of the probability plot. However, it is not always possible to know the real K value, so it is best to choose the lowest value that captures the maximum structure present in the data [15]. Evanno et al. described a method to estimate K based on a second-order change of probability value taking into account the variability of the probability value between different replicates of each K value—*delta K* [16]. This method is implemented in *Structure Harvester* [17] (*see* **Note 6**). However K is not an absolute value, defined values should be carefully considered taking into account any known characteristic of the studied populations. There are several factors that can affect the clustering of the samples: (1) number of markers; (2) number of samples; (3) number of clusters; and (4) allele frequency

correlation. Endogamy or genotyping errors can have the same effects as true population structure as they can simulate linkage disequilibrium in unlinked markers and deviations from HWE.

STRUCTURE has several models for ancestry (Fig. 2, *see* **Note 7**) and allele frequencies (*see* **Note 8**). For more details refer to the *STRUCTURE* manual and articles describing the different models [2, 18–20]. Considering the type of analysis required in a forensic context, the *admixture POPFLAG* ancestry model is appropriate. This combines two important features: consideration of admixture between populations (individuals can have recent ancestors from multiple populations so ancestry membership proportions from each ancestral population can be calculated); and some individuals can be used as a reference to help infer the ancestry of the samples under study. Regarding the allele frequencies model, it is advisable to use the correlated allele frequencies model because it will guarantee that an undetected correlation will be identified without affecting the results should it be absent.

3.2.2 Preparation of a STRUCTURE Input File

Data to be analyzed with *STRUCTURE* needs to be organized in a single matrix (as a text file) where optional information can be considered to complement the genotypic data. Such information should be included in a predefined order and it is important to highlight that only the genotypic data is required for the analysis. We will focus the construction of an input file on the information of greater relevance when analyzing a casework profile (Tables 1 and S1). For more information about constructing *STRUCTURE* input files (especially formatting information on recessive alleles, marker distance, phase information, or phenotype), refer to *STRUCTURE* software manual or to a recent overview [20].

- First line: header line. Headers are only included in the markers columns.
- First column: sample name information that can be an alpha-numerical code which can introduce errors when

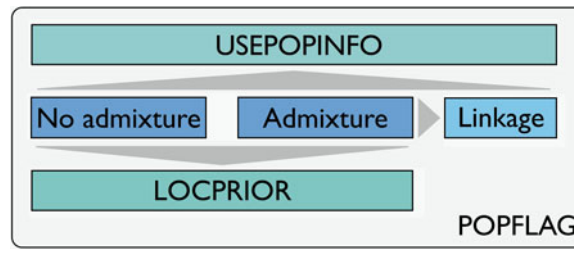

Fig. 2 Schematic representation of *STRUCTURE* ancestry models and their relationship. The central models are *no admixture* and *admixture*; both can be used together with *LOCPRIOR* information. The *admixture* model is the basis for the *linkage* model. All three models (*no admixture*, *admixture*, and *linkage*) can be used in conjunction with the *USEPOPINFO* model. All the above models can be used considering *POPFLAG* information

Table 1
***STRUCTURE* input file format**

						M1	M2	M*n*
S1	1	1	1	Extra	Extra	1	2	1
S1	1	1	1	Extra	Extra	3	4	1
S2	1	1	2	Extra	Extra	3	2	2
S2	1	1	2	Extra	Extra	3	2	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
S10	1	1	3	Extra	Extra	1	4	1
S10	1	1	3	Extra	Extra	1	4	2
S11	2	1	4	Extra	Extra	1	2	2
S11	2	1	4	Extra	Extra	3	4	2
S12	2	1	5	Extra	Extra	3	2	1
S12	2	1	5	Extra	Extra	3	2	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
S20	2	1	6	Extra	Extra	1	4	2
S20	2	1	6	Extra	Extra	1	4	2
S21	*n*	0	7	Extra	Extra	1	2	1
S21	*n*	0	7	Extra	Extra	2	4	1
S22	*n*	0	8	Extra	Extra	2	2	2
S22	*n*	0	8	Extra	Extra	2	2	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
S*n*	*n*	0	9	Extra	Extra	1	4	**-9**
S*n*	*n*	0	9	Extra	Extra	1	4	**-9**

Samples S1…S*n* from populations 1…*n* analyzed with genotypic data from markers M1…M*n*. Samples 1…10 belong to population 1 and can be divided into three locations (1–3). Samples 11…20 belong to population 2 and can be divided into three locations (4–6). Samples 21…S*n* belong to population *n* and can be divided into three locations (7–9). Samples from populations 1 and 2 are reference (*POPFLAG* = 1) and study samples are from population *n* (*POPFLAG* = 0)

running *CLUMPP*, easily solved in *Structure Harvester* (*see* **Note 6**).

- Second column: a numerical code representing the population of origin as defined by the researcher. By default this information is not used by the clustering algorithm but can help organize the output file.
- Third column: *PopFlag* information. This is a Boolean variable where 1 (TRUE) represents the samples that should be used as reference and 0 (FALSE) the casework/study samples.

- Fourth column: *LocPrior* information. A numerical code that denotes subpopulation groups, geographical locations, or other shared characteristic between individuals inside the population groups defined in the second column. This information is used when considering the *LOCPRIOR* ancestry model.
- Any number of extra columns can list useful information for the researcher. For example, as the population and *LocPrior* information is numeric, extra columns with the names can be included as an easy way to cross check data later.
- The following columns include genotypic data for any number of markers (SNPs, indels, or multiallelic markers such as STRs). Genotypes should be coded as numbers. For SNPs, we routinely use *A = 1*, *C = 2*, *G = 3*, and *T = 4*. STRs are already numerically coded but in the case of intermediate alleles the "." should be removed, i.e., *19.3 = 193* (*see* **Note 9**).
- Missing data is usually coded as *-9* but any other code not present in the file can be used.
- Each allele needs to be represented in a separate cell: both alleles in the same line but in different columns or both alleles in the same column but in different lines (we will focus on the latter as shown in Table 1).
- Spaces should not be included.

3.2.3 How to Run STRUCTURE Software

The first stage when running *STRUCTURE* (http://pritchardlab.stanford.edu/structure.html) is to create a new project (File > New project) following four established steps:

- Step 1—project information: name of the project, directory where the project will be saved, and input file.
- Step 2—information of input data set (*see* **Note 10**): number of individuals, ploidy of data, number of markers, and missing data value.
- Steps 3 and 4—format of input data set (*see* **Note 11**): information contained in rows and columns (e.g., row of marker names or individual ID for each individual).

Before creating the project, *STRUCTURE* presents a summary where it is possible to confirm the selected options. If there are no errors, the project opens and the data is visible.

The next stage is to create a new parameter set (Parameter set > New):

- Run length—a *burnin* period of 100,000 is more than sufficient to allow a progressive convergence towards reliable allele frequency estimates in each population and probabilities for membership of individuals to a population. Measurement of the assumed number of populations uses the MCMC estimation and is performed separately from

the *burnin*. About 100,000 MCMC repeats have been shown to provide good ancestry membership proportions estimates. But *burnin* and MCMC repeat number should be adjusted depending on the study objectives and information contained in the data set (*see* **Note 12**).

- Select ancestry model—depending on the study objectives and the data to be analyzed, different ancestry models can be considered. For the *admixture POPFLAG* model select "Use admixture model" under the "Ancestry model" tab and "Update allele frequencies using only individuals with *POPFLAG* = 1 data" under the "Advanced" tab.
- Select allele frequencies model—for the *correlated allele frequencies* model select the "Allele frequencies correlated" option under the "Allele frequencies model" tab.
- Leave the "Compute probability of the data (for estimating *K*)" option under the "Advanced" tab selected so that posterior *Ln Pr*(*K*) values are calculated—those will be used to estimate the optimum *K* value.
- Save the new parameter set with the desired name and confirm the selected options in the summary window that opens after saving. A tree on the left side of the screen will include all the parameter sets created, indicating the one active at the moment.

There are two ways of starting a simulation:

- Run a single *K* value—in the "Parameter set" menu select "Run" and set the assumed number of populations (*K*). This option only allows a single *K* value and replicate at a time.
- Schedule multiple runs—in the "Project" menu select "Start a job". A new window opens—select the parameter set(s) to be analyzed, the *K* values, and the number of iterations for each *K* (*see* **Note 13**). For example, two different parameter sets can be programmed to run from *K* = *2* to *K* = *6*, three replicates for each *K*—this sums up to 30 scheduled runs. This option is advantageous for large projects—a new run starts automatically after the previous one has finished so there is no need for constant attention on the progress of the job (*see* **Note 14**).

3.2.4 STRUCTURE Associated Software

Software associated with *STRUCTURE*, for example, *CLUMPP* [21], contains three algorithms for the alignment of multiple replicate analyses of the same data set which allows the transformation of any number of replicate simulations for each *K* in a single set of data (*see* **Note 15**). Such data is suitable for analysis with another supporting program *distruct* [22] which allows the visualization of the estimated membership coefficients: populations are represented as colors and individuals as bars portioned into colored segments that correspond to membership coefficients in the groups (*see* **Note 16**).

3.2.5 What Information Can Be Obtained from STRUCTURE?

STRUCTURE output files include information on the estimated clusters, i.e., the population groups generated not the input populations. However when the populations are defined in a way that they closely match the calculated clusters, the inferences of the population ancestry membership proportions in each of the predefined clusters can be considered to be the proportions of the input populations. When attempting to classify a population or individual, the use of reference populations closely matching the inferred clusters is important, especially when analyzing admixed samples where it is important to define the contributing parental populations. The ancestry membership proportions for each individual in each cluster are also calculated by *STRUCTURE*.

Allele frequency divergence among populations, average distances (expected heterozygosity) between individuals in the same cluster, mean F_{ST} values, and estimated allele frequencies in each cluster (including estimated ancestral frequencies) are calculated. As a way of quantifying the information given by a particular *STRUCTURE* run and estimating the optimum K value it calculates the estimated probability of the data, the mean likelihood value, and associated variance. And it calculates the mean value of *alpha* (α) as a measure of the relative admixture levels between populations—when $\alpha >> 1$ the individuals are highly admixed; for values of $\alpha << 1$ each individual has its origin mainly in one population (from our experience with the HGDP-CEPH panel of samples, $\alpha < 0.05$—this value varies depending on the population groups considered and the differentiation power of the marker sets used).

The population and individual ancestry membership proportions can be represented in two distinct types of plot:

- A bar plot where each individual of the data set is represented by a vertical line divided into K colored segments proportional to the estimated membership into each of the K inferred clusters. To visualize the bar plot in *STRUCTURE* choose the appropriate result file in the tree on the left side of the window—on the simulation result window menu select Bar plot > Show.
- Each individual is represented as a colored point in a triangle (on the simulation result window menu select Triangle plot > Show). Colors correspond to the population tag in the input file. The estimated ancestry vector for an individual is formed by K components that sum up to 1. This type of plot is particularly useful to represent $K = 3$ data because the vectors can be represented in one triangular plot. For each point, the distance to the triangle vertices gives each of the three components. Individuals located in one of the vertices are completely assigned to the population represented in it.

Despite the advantages of the triangular plot when visualizing $K = 3$ data, bar plots are usually easier to interpret, especially for $K > 3$.

In the case of forensic casework analysis, *STRUCTURE* gives information on the training set (allowing the assessment of the used reference data set—the optimum K value matches the number of reference populations, which are completely differentiated among them) and it also gives us the individual ancestry membership proportions (such information has considerable potential in guiding investigators to more clearly defined suspect pools, this being particularly true when no eyewitness is available or STR profiles fail to match DNA database records). This is illustrated in Subheading 3.5.

3.3 The Snipper Web Portal

3.3.1 Background on Snipper Analysis

The *Snipper* portal includes a straightforward Bayesian system for predicting ancestral origin and estimating the misclassification rate. It uses a set of samples of each population as training sets and assigns individuals to the population that maximizes the posterior probability (maximum likelihood calculation) [3]. The likelihood parameters are estimated from training set allele frequencies assuming HWE and independence for the used loci (*see* **Note 17**).

Snipper was originally designed to provide a real-time ancestry assignment system for 34-plex profiles with reference to default pre-typed AFR-EUR-E ASN training sets and this still represents the simplest approach for assessment of a single casework profile to obtain an immediate overview of ancestry. The portal has been updated to include 34-plex [3, 4] and AIM-indel [5] fixed reference data for five populations groups: AFR-EUR-E ASN-AME-OCE. But the ancestry analyses can be extended beyond the default settings. For example, custom Excel files (including any binary markers that are of interest for the researcher) or frequency based Excel files (helpful when working with STRs or haplotypes) can be used as reference training sets.

A new version of *Snipper* is being prepared (*Snipper App suite version 2.0*) to include new functionalities including turn on/off the HWE assumption; prediction of admixture components; batch analysis (multiple profiles); fine-tuning of a training set; classification of single profiles; and analysis of training sets through multinomial logistic regression (beta version). At the time of writing a publication describing *Snipper 2.0* is in preparation.

3.3.2 Preparation of a Snipper Input File

Careful preparation of the Excel file containing the custom training set profiles and precise matching of unknown profiles to training set data for bases and locus order is important. Therefore it is recommended to sort component SNPs/indels into ascending rs-number order as an aid to data checking.

For *Snipper* analysis using binary markers, an *.xlsx* Excel file (*.xls* can still be used for certain previous options) with sample, population, and genotype information listed (Tables 2 and S2)—how that information is organized is also important so the following considerations should be taken into account:

- Cell 1A indicates the number of samples; cell 1B the number of markers; and cell 1C the number of populations.
- Line 1 (from column D onwards) specifies the marker name (represented by an alpha-numerical code).
- Lines 2–5 can be left empty or can be used to include useful notes (e.g., one of the lines can be used to store the study/casework sample profile and other line can be used

Table 2
***Snipper* input file format**

	A	B	C	D	E	…	XFC	XFD
1	# Samples	# Markers	# Populations	M1	M2	…	M*n*	
2								
3			Profile	AG	TT	…	AC	
4			Concatenate	=D3&E3&…&XFC3				
5								
6	1	P1	S1	AG	CT	…	AA	1
7	2	P1	S2	GG	CC	…	CG	1
8	3	⋮	⋮	⋮	⋮	…	⋮	⋮
9	4	P1	S10	AA	TT	…	AG	1
10	5	P2	S11	AG	CT	…	GG	1
11	6	P2	S12	GG	CC	…	AC	1
12	7	⋮	⋮	⋮	⋮	…	⋮	⋮
13	8	P2	S20	AA	TT	…	CC	1
14	9	P*n*	S21	AG	CT	…	AA	0
15	10	P*n*	S22	GG	CC	…	GG	0
⋮	⋮	⋮	⋮	⋮	⋮	…	⋮	⋮
1048576	*n*	P*n*	S*n*	AA	TT	…	**NN**	0

Samples S1…S*n* from populations 1…P*n* analyzed with genotypic data from markers M1…M*n*. Samples 1…10 belong to population 1; samples 11…20 belong to population 2, and samples 21…S*n* belong to population P*n*. An extra column after the last marker (in this case column XFD) should be included when trying to classify several study samples simultaneously—samples from populations 1 and 2 are reference (labeled as 1) and samples from population P*n* are the unknown study (labeled with 0). Lines 2–5 can be used to include useful information—e.g., when a single profile is being classified it can be included (here in line 3) and concatenated (cell D4)—the concatenated profile can then be copy-pasted directly into *Snipper*

to concatenate that profile—ready for copying and pasting, i.e., if the profile is in line 3 type = *D3&E3&F3…* in the desired cell).

- Column A (from line 6 onwards) has a numeric value that usually represents a sample.
- Column B (from line 6 onwards) has the population names.
- Column C (from line 6 onwards) has the sample names (which can be represented by an alpha-numerical code).
- Column D onwards (from line 6 onwards) includes the genotypes (coded as nucleotide bases—ACGT). Missing data should be coded as NN. Other symbols in the file (e.g. ?, spaces) are not recognized. Triallelic markers can be included in the analysis.
- A new batch analysis option was implemented in *Snipper v2.0* which allows for simultaneous classification of more than one profile. In this case, the input file should be constructed as described in the previous points. An extra column after the last marker (with no headers—start in line 6) needs to be included: training samples are to be marked as 1 and study samples to be classified as 0.

3.3.3 How to Run Snipper

Snipper includes several options to classify individuals and analyze populations. For forensic analysis the two most applicable options are: "*Classification as Europe-East Asia-Africa-America-Oceania (34 SNPs, 46 Indels, or both sets)*" and "*Classification with a custom Excel file of populations*". There is an additional option that works in the same way but allows batch analysis: "*Classification of multiple profiles with a custom Excel file of populations*".

1. The "*Classification as Europe-East Asia-Africa-America-Oceania (34 SNPs, 46 Indels, or both sets)*" option uses fixed training sets and provides a simple system to classify single profiles.
   - Step 1—go to http://mathgene.usc.es/snipper/pop-choosing5groups.html
   - Step 2—choose the marker set from three options: 34-plex SNPs (the original marker set [3] or the revised set [4] can be selected), 46-plex AIM-indels [5], or a combination of 80 binary markers (Indels combined with the revised 34-plex set). SNPs are listed in rs-number order and AIM-indels in electrophoretic order—on the left side links give images listing the marker order in each option.
   - Step 3—choose populations. Three to five main population groups are available (Africa, Europe, East Asia plus America plus Oceania).
   - Step 4—choose the classifier. Four options are now available: naïve Bayesian analysis (considering whether the Hardy-Weinberg principle applies or not), multinomial logistic regression, and genetic distance algorithm.

- Step 5—data input. Depending on the option selected in Step 1, a profile including 34, 46, or 80 markers (68, 92, or 160 bases respectively) should be typed (*see* **Note 18**). As described before, a profile can be built by concatenating data in Excel (using the "&" operand) allowing individual scrutiny of composite genotypes before direct copy-pasting into the query window left of the "*Classify*" button (*see* **Note 19**).

2. "*Classification with a custom Excel file of populations*"—this option allows extension of ancestry analyses beyond the default five-population group comparisons and 34, 46, or 80 binary markers using *Snipper*.
   - Step 1—go to http://mathgene.usc.es/snipper/analysispopfile_new.html
   - Step 2—data input (population). An Excel file prepared as described above (Table 2 without the final column) is uploaded.
   - Step 3—choose classifier. Options as described above.
   - Step 4—data input (individual). A profile string containing the same number of markers in the same order as they appear in the data file uploaded in Step 1 (*see* **Note 18**) is entered in the query window.
3. "*Classification of multiple profiles with a custom Excel file of populations*"—go to http://mathgene.usc.es/snipper/analysismultipleprofiles.html. This option works as above but without the need for individual profile submission. Profiles to be classified are indicated as previously described (Table 2). The multinomial logistic regression classifier function is not currently available for this option.

Snipper also includes an option to analyze training sets to gauge characteristics of the component binary markers—"*Thorough analysis of population data of a custom Excel file*" (http://mathgene.usc.es/snipper/analysispopfile2_new.html). This is useful to assess the informativeness of new candidate AIM binary markers for ancestry inference. After uploading the Excel file of custom data and defining Hardy-Weinberg, choose "*Perform a verbose cross-validation analysis of my population data with the best _ SNPs*" adding the relevant number of markers to assess. Cross-validation removes each component sample in turn, recalculates the allele frequencies in the training set, and then assigns ancestry for the removed profile. The other options "*Try to classify all individuals in the sample*", "*Perform a non verbose cross-validation analysis of my population data*", and "*Compute bootstrap error of my population data*" provide choice of alternative assignment error estimators. Multinomial logistic regression can also be applied to the population data—in this case information given in Step 2 about HWE will be ignored. Once a training set has been assessed for informativeness, users can choose options

2 and 3 described above to compare single or multiple profiles from unknown samples to the custom reference data and assign ancestry in identical fashion to using the fixed training sets.

3.3.4 Evaluating Snipper Output

Results from the analysis of a profile comprise the submitted profile; the assumed classifier; the -log likelihoods (use of -log likelihoods permits easier comparison of the very small likelihood ratio figures normally generated) and percentiles for the training set population groups; the likelihood ratios in verbose format and predicted admixture components and ancestry; a set of plots summarizing the classification; the apparent success of the classification; and a list of the markers in descending order of divergence (*see* **Note 20**). Missing genotypes are flagged in red in the divergence list to allow some assessment of the potential contribution of gaps in the profile, in other words, assignments made with several red markers at the top of the list will be much less reliable than those with gaps at the bottom, although this will be clear from the probabilities obtained. Apparent success measures the rate of correct assignment of training set samples using the markers of the profile. These values are 100 % for a complete set of markers, but drop when significant numbers of gaps occur in the submitted profile (in the case of the three group 34-plex fixed training set this is particularly true for EUR:E ASN comparisons).

3.4 Principal Component Analysis (PCA)

3.4.1 Background on PCA

Principal component analysis or PCA is a multivariate data analysis technique allowing the reduction of dimensionality, i.e., it uses fewer variables, while preserving much of information in the data. Usually two or three principal components are made, constructed as linear combinations of the original variables. Working with only two or three variables allows graphical representation of the data in a 2D plane or 3D graphic, providing fast visual recognition of patterns or clusters. Numerous software packages are available to perform PCA analysis when numerical variables are used. When SNP data is considered, an initial transformation (or recodification) is needed to access this existing software. The next section details SNP data preparation using the statistical package *R*.

3.4.2 Preparation of PCA Input Files

SNP analysis with PCA requires two text files with sample, population, and genotype information. One of the files should include training set data and the other the study samples to be compared. Both files have the same format (Tables 3, S3 and S4)—the system for organizing this data is important so the following considerations should be carefully taken into account:

- The first column includes sample name information in the form of an alpha-numerical code. The column header is "Sample".
- The second column has the populations/groups names. The header is "Population".

Table 3
Principal component analysis SNP input file format

Sample	Population	M_1	M_2	M*n*
S1	P1	AG	CT	AA
S2	P1	GG	CC	CG
⋮	⋮	⋮	⋮	⋮
S10	P1	AA	TT	AG
S11	P2	AG	CT	GG
S12	P2	GG	CC	AC
⋮	⋮	⋮	⋮	⋮
S20	P2	AA	TT	CC
S21	P*n*	AG	CT	AA
S22	P*n*	GG	CC	GG
⋮	⋮	⋮	⋮	⋮
S*n*	P*n*	AA	TT	**NN**

Samples S1…S*n* from populations P1…P*n* analyzed with genotypic data from markers M1…M*n*

- The following columns have genotype data, one marker per column. Each column header will have the corresponding marker name, which can be an alpha-numerical code. Genotypes are coded with nucleotide bases (ACGT) and missing data as NN. Note that markers must be in the same order in both input files.
- Spaces can be included in the file except as part of the genotype data (they will be considered as a new genotype, i.e., TT ≠ T T).
- Triallelic markers can be included in the input file but they will not be considered for the principal components calculation.

3.4.3 Creating a PCA Plot

In this subheading we include *R* scripts that can be used to generate 2D principal components graphics (only if the number of variables (SNPs) is smaller than the number of samples). The script commands can be copied and pasted into the *R* console. With the main focus on graphics, *R* offers a range of options to manipulate data and generate plots that adjust to user needs. In the case of this script, command lines were added to allow changes in color, shape, and size of the symbols representing individuals (*see* **Note 21**).

All text after the # symbol represents notes for the user and will not be computed. A *R* version of the script is included as supplementary in the digital version of this chapter.

```
##############################
### Script to make a 2D PCA ###
##############################

# Important Note: this script can only be used when the number of samples is equal to, or higher than, the
 number of SNPs.


# First open SNPassoc library - this is an association package that allows you to recode SNP data

library(SNPassoc)

TEST<-function(x){try(snp(x,sep=""),silent=TRUE)} # homemade function that detects if a SNP is bi- or tri-
 allelic

##############################
# To read the data from two input files. The computer will prompt for the location of the reference samples
 file and the study samples file, in that order
##############################

RefData<-read.table(file.choose(), sep="\t",header=TRUE,na.strings=c("NN")) # function that reads the
 reference data input file and stores it in the dat object

NRefSamples<-dim(RefData)[1] # get the number of reference samples

StudyData<-read.table(file.choose(), sep="\t",header=TRUE,na.strings=c("NN")) # function that reads the file
 with the study samples and stores it in the StudyData object

NStudySamples<-dim(StudyData)[1] # get the number of study samples

### Because tri-allelics are not considered when making the PCA, they are removed from both data sets using
 the function TEST.

apply(RefData,2,TEST)->RT

deletedSNPs<-which(as.numeric(summary(RT)[,1])==1) # defines which SNPs have to be removed

RefData2<-RefData[,-deletedSNPs] # for the reference samples, only the columns with bi-allelic SNP data are
 kept

StudyData2<-StudyData[,-deletedSNPs] # for the study samples, only the columns with bi-allelic SNP data are
 kept
# this removes tri-allelic SNPs if there were any

##############################
# To recode and typify the data
##############################

ComData<-rbind(RefData2,StudyData2) # combines the reference and study samples in one variable (ComData) -
 this needs to be done prior recoding the SNP data to guarantee the coding uniformity

datSNPT<-apply(ComData,2,function(x) {additive(snp(x,sep=""))}) # the additive function recodes each bi-
 allelic SNP in the data as numeric (0=homozygous for the most frequent allele, 1=heterozygous, 2=
 homozygous for the least frequent allele)

datSNP<-datSNPT[1:NRefSamples,] # after recoding, the reference data set is temporarily isolated into a new
 variable (datSNP) to make some computations

m<-apply(datSNP,2,mean,na.rm=TRUE) # this calculates a vector with the mean value of each "numeric" SNP

s<-apply(datSNP,2,sum,na.rm=TRUE) # this calculates a vector with the number of occurrences of the least
 frequent allele for each SNP

n<-apply(datSNP,2,function(x) sum(!is.na(x))) # this calculates a vector with the number of valid genotypes
 for each SNP (NN genotypes are not considered valid)

p<-s/(2*n) # vector with the frequency of least frequent allele for each SNP

XT<-scale(datSNPT,center=m,scale=sqrt(p*(1-p))) # this typifies the SNPs (to each "numeric" SNP value the
 average is subtracted and then divides by the standard deviation)
```

```
XT[is.na(XT)]<-0 # replaces missing values with 0

X<-XT[1:NRefSamples,] # after recoding and typifying, the reference data set is isolated into a new variable
 (X)

Y<-XT[-c(1:NRefSamples),] # after recoding and typifying, the study data set is isolated into a new variable
 (Y)

if(NStudySamples==1){ # this forces Y to be a matrix if there is only one study sample

              dim(Y)<-c(NStudySamples,length(Y))
             colnames(Y)<-colnames(X)

    }

##########################################
# Some computations before preparing the plot
##########################################

princomp(X, scale = FALSE)->X.PCA # computes the PC of the reference samples

X.PCA$loadings->M # gets the rotation matrix

(Y%*%M)[,1:2]->StudyCoordinates # computes the new coordinates of the study samples


if(NStudySamples==1){StudyCoordinates<-t(StudyCoordinates)} # forces StudyCoordinates to be a matrix

per<-eigen(cov(X))$values/sum(eigen(cov(X))$values) # this calculates the percentage of explained variance
 for each principal component

per<-round(per*100,2) # values are rounded to two decimal positions

# In this part colours in the plot must be chosen (if you have more than three populations)

mycolours<-c("orange","pink","skyblue2") # choose the colours you want to use for each population
 (considering that populations are in alphabetic order). A complete list of colour names can be obtained
 with the command colours() or with the help of the Chart of R colours available at http://research.stowers-
 institute.org/efg/R/Color/Chart/

colours<-as.character(factor(RefData$Population,labels=mycolours)) # population names are converted to the
 corresponding colour name


#####################
# The plot starts here
#####################

quartz() # opens a new graphic display window if you use MacOS
# windows() # is the alternative command for windows OS that opens a new graphic display window if you use a
 Windows PC (remove # here and replace # in front of previous line)

plot(X.PCA$scores[,1:2],col=colours,pch=20,main="put here your plot title",xlab=paste("PC1
 ",per[1],"%",sep=""),ylab=paste("PC2  ",per[2],"%",sep=""),cex=1.5)
# this plots the two first principal components. The plot title, pch and cex values can be changed

legend("topleft",legend=levels(factor(RefData$Population)),col=mycolours,pch=20,cex=0.5,y.intersp=1)
# this adds a legend to the plot. Its position can be changed using "topleft", "topright", "bottomleft" or
 "bottomright". Pch should match the one used in the plot.

### With the next set of commands it is possible to include the study individuals superimposed onto the
 previously plotted principal components graphic.

points(StudyCoordinates,col="black",pch=20,cex=2)
# this estimates the coordinates for the study individuals and plots them onto the previously generated
 principal component graphic. colour, pch and cex can be adjusted.

text(StudyCoordinates,as.character(StudyData[,1]),cex=1,pos=1)
# this adds study individual's descriptors to the plot
```

3.4.4 What Information Can Be Obtained from PCA?

PCA allows the exploration of data sets and shows proximity between individuals. In fact, it is possible to include a casework sample in the PCA plot generated for the reference populations helping to infer, through visual inspection, the most probable classification of that individual (Fig. 3).

3.5 Casework Example of a Custom Ancestry Inference: The 11-M Madrid Bomb Attack

In the 11-M Madrid bomb attack investigation, standard DNA analysis with STRs was supplemented with Y-filer and standard mtDNA analysis in most exhibits. But seven complete STR profiles, originating from five personal items together with a handprint on the handle of the bag containing an undetonated device, failed to match any of the suspects so these DNAs became the focus of specialist genotyping to analyze ancestry, specifically confined to the comparison of European with North African variability. This differentiation can be difficult to achieve for Y-chromosome and mtDNA due to differences in the scope and depth in the databases between European and North African data, so the 34-plex AIM-SNP set was chosen [24].

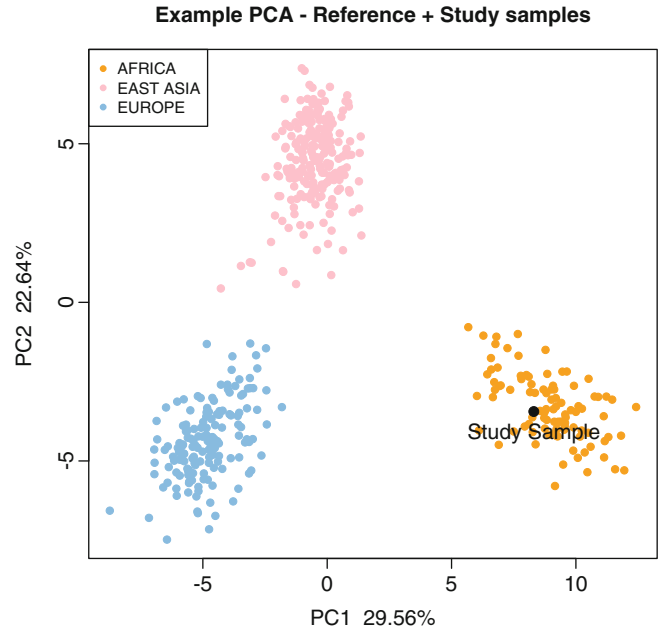

Fig. 3 PCA plot generated using the R script described in Subheading 3.4.3. Three population groups from the HGDP-CEPH panel of samples are used as reference data: Africa (*orange*), Europe (*blue*), and East Asia (*pink*). One study sample was plotted in the reference PCA (*black*). It is possible to infer that the study sample is likely to be African. Both reference and study samples genotypes are supplemented as text files in the online version of this chapter

The approach followed in this case is a good example of the integration in one analysis of the different techniques described in this chapter. Two training sets were made using 48 Moroccans and 48 Spanish from Madrid. Using *Snipper* a cross-validation assessed the accuracy (assignment success/error) and performance (range of likelihoods) of the training sets and to generate pairwise likelihood plots to assess patterns of possible admixture (Fig. 4). Such plots compare two ancestries and enable a simple comparison of the range of likelihoods observed in the unknowns alongside their closest parental population vs. another alternative population. The charts are made in Excel by converting *Snipper* likelihoods to whole numbers (using the =EXP formula in Excel), making each pairwise ratio (in this case, lk Moroccan/lk Spanish) and ranking values in descending order. Charting these with a logarithmic scale provides a simple visual check of the range of divergence between the populations compared as points with varying distances from the midline of 1 (equating to balanced odds of ancestry assignment to either population). The most distant points from the midline represent the strongest assignments. In populations without admixture, points are fully separated; when admixture occurs, a significant proportion of values are close to or cross the midline. Using *STRUCTURE*, admixture patterns were assessed in the training set. Some individuals, corresponding to likelihood ratios between 0.01 and 100, presented admixed ancestry. Considering this information, an area of uncertainty was defined, with individuals falling in that area not assigned to a particular population group.

When comparing PCA (*see* Fig. 2 in [24]), *STRUCTURE*, and *Snipper* results (Fig. 4), they were concordant for all case samples: three were classified as North African, one as European, and three were left unassigned. Those three unassigned profiles probably represent individuals with highly admixed parentage and genomic backgrounds: a reasonable scenario given the proximity of Southern Europe and North Africa.

One 34-plex assignment contradicted the uniparental analysis—although mtDNA and Y-chromosome markers routinely demonstrate informative geographic differentiation, this is not always true when recent gene flow has occurred or populations show strong sex bias (i.e., males are mainly from one population and females from another). The individual inferred to be European from uniparental data gave strong indications to be North African from the 34 SNP genotypes and was later identified by the investigation to be Algerian.

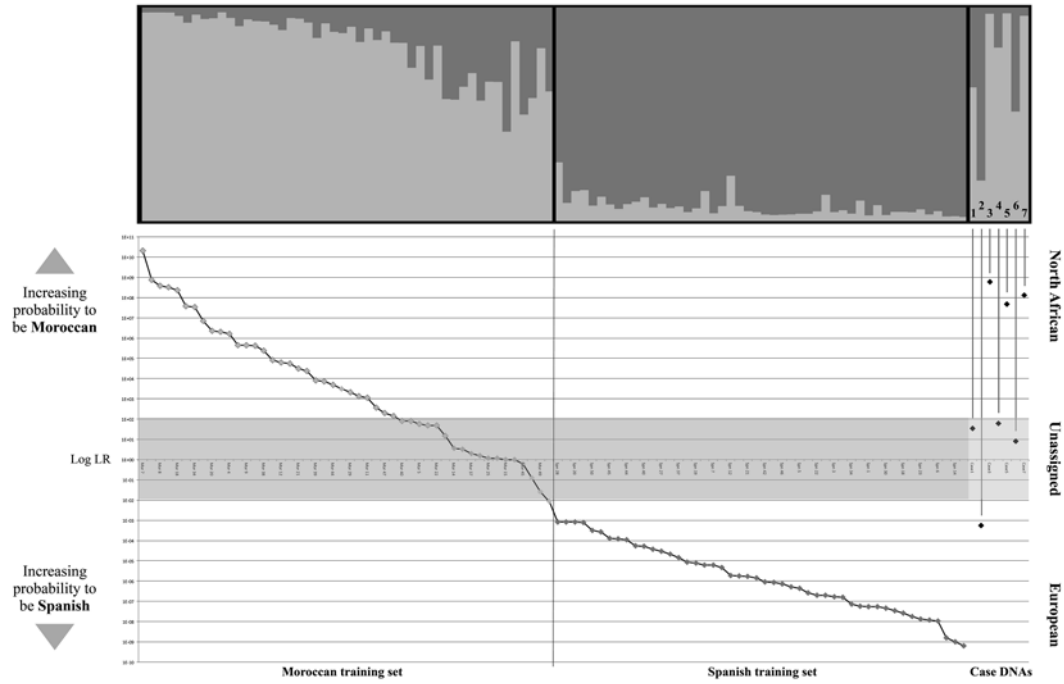

Fig. 4 11-M Madrid bomb attack *STRUCTURE* and *Snipper* analysis results. *STRUCTURE* analysis was performed to compare *Snipper* pairwise plots with an established alternative system of ancestry assessment. *STRUCTURE* runs were performed using the *admixture* ancestry model with 200,000 MCMC steps after a *burnin* of length 200,000. In the *Snipper* pairwise plot, samples are organized from most probable Moroccan to most probable Spanish, defining a separation from likelihood ratios represented on a logarithmic scale with values higher than 1 = higher probability North African and ratios smaller than 1 = higher probability European. Individuals in the *STRUCTURE* plot are in the same order as the *Snipper* pairwise plot, allowing direct comparison of both analyses. Some admixture patterns are present and this helped to establish an uncertainty area (ratios between 0.01 and 100) where individuals would have more probability of being misclassified, so the decision was to leave these unassigned

4 Notes

1. Of all databases included in *SPSmart*, 1000 Genomes and HapMap are of most utility as they have more markers and larger sample sizes (including admixed ancestry populations). This is particularly true for 1000 Genomes—the ENGINES browser allows scrutiny of SNP variation across the whole genome (down to a minor allele frequency of ~1 %) from Africans, Europeans, and East Asians previously used by HapMap. In contrast, the HGDP-CEPH (Human Genome Diversity Panel) samples 1050 individuals with wide currency

in population genetics studies [25, 26]. The geographic coverage is patchy in certain regions but all continents and all major genetic ancestry groups defined by studies of variability are represented. Smallest sampling is 28 Oceanians from two populations and just six San from Namibia. SPSmart provides freely downloadable genotype data from 650,000 SNPs (obtained with Illumina 650K arrays [27]) for each HGDP-CEPH sample in the Stanford University study of this panel.

2. Please note that, despite Excel software is referred as the one to be used, any spreadsheet software such as Numbers in MacOS or the free and open-source OpenOffice Calc or LibreOffice Calc can replace it.
3. Symmetrical base SNPs, comprising an A/C on one strand and a T/G on the other, are a particular problem and source of error when comparing genotypes generated on different platforms or listed in different databases. The *SPSmart* SNP*for*ID browser makes allowance for most base inversions by showing the HapMap (or other) allele frequency summary charts with different allele segments if these differ from the strand interrogated by the 34-plex and 52-plex extension primers. For example, rs2304925 is listed as a SNP*for*ID GT SNP but a HapMap AC SNP and this applies equally to GC or AT SNPs, e.g., rs10141763. The *SPSmart* help file provides a clear and carefully worded guide in the "Symmetrical bases" section. There are four symmetrical bases in the 34-plex: rs773658, rs10141763, rs1335873, and rs16891982. The last of these is the most informative SNP for differentiating component populations within Eurasia so it is particularly important to be clear about differences between SNaPshot and database allele calls for this marker.
4. Bayesian population analysis methods calculate a simple relationship between allele frequencies in a population and allele frequencies observed in the tested individuals. *STRUCTURE* analyzes differences in the distribution of genetic variants between populations through an iterative Bayesian algorithm that tries to group samples into clusters whose members share similar patterns of variation. Bayesian methods have the advantage of allowing the use of prior information about the samples to progress analysis. But the ability to differentiate populations in a sample set can be limited when a small number of samples and/or markers are used.
5. *STRUCTURE* uses an MCMC algorithm that starts by randomly assigning individuals to a predetermined number of K populations. Allele frequencies of each population are estimated

considering the individuals assigned to each population. Individuals are then re-assigned to populations taking into account the estimated frequencies for each population in a process repeated up to 10,000–100,000 times.

6. *Structure Harvester* implements the *delta K* method of Evanno et al. [23] to estimate the optimum *K* value [16, 17]. After running STRUCTURE, pass the zipped folder containing the results files (named *x_run_y_f*, where *x* represents the parameter set name and *y* the run number) to the *Structure Harvester* browser and click Harvest! to start the analysis. Conditions are: a minimum of three sequential *K* values should be included, with more than one replicate for each *K* value (same number of replicates for all *K* values). Posterior probability and *delta K* plots are available to download (Fig. 5)—the optimum *K* is usually the point with the highest *delta K* value or the one which immediately precedes the *Ln Pr*(*K*) plateau. This software is also useful as it automatically generates input files to run *CLUMPP*.
7. There are two basic ancestry models: *no admixture* and *admixture* [2]. The first is used if there is no prior knowledge about the origin of the populations under study or if there is a reason to consider each population as completely discrete. But because admixture between populations is a common characteristic, knowing the approximate median value of the ancestral population proportions for each individual and their populations of origin is very important for the characteriza-

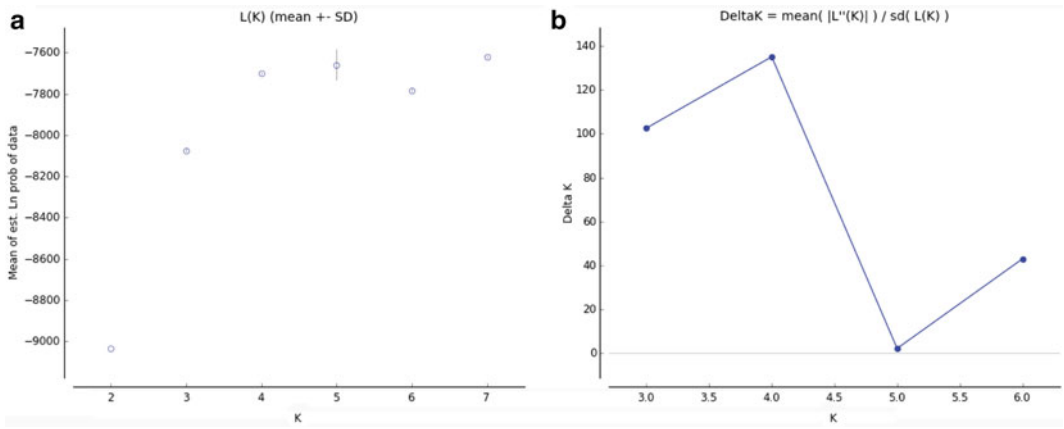

Fig. 5 Example of posterior probability and *delta K* plots obtained with *Structure Harvester* for the same analysis. In this case, the optimum *K* value is 4—the point where the *plateau* in the posterior probability starts with maximum *delta K* value

tion of a study population or, in a forensic context, a casework sample. In this case the *admixture* model is more appropriate. The *LOCPRIOR* option [19] can be used when there is additional sample characteristic data available, e.g., linguistic, geographical, cultural, or phenotypic information. The *LOCPRIOR* parameter is particularly informative when there are weak population structure signals—a situation that can result from using reduced number of markers, small sample sizes, or due to close relationships between populations. It has two main advantages: (1) generally it will not find population structure when this is not present; and (2) it can ignore location information when individual ancestry is not related with it. When admixture LD is present, the *linkage* model [18] (which is based on the *admixture* model) can be applied to obtain more accurate estimates of statistical uncertainty from use of linked markers. Population labels can be used to calculate the probability that each individual originates from the assumed population—individuals with low probabilities can be considered as migrants or having high co-ancestry. This option is included in the *USEPOPINFO* model [2] and should only be used when population labels are well defined beforehand and correspond almost exactly to the groups ultimately defined by the *STRUCTURE* results. The last model considers the specified information about the population of origin of a portion of individuals to help infer the ancestry of other samples with unknown origin: the *POPFLAG* model [2]. This option needs caution as selected samples are treated as the "reference" set (pre-assigned *POPFLAG=1*) meaning allele frequencies estimates are based on a reduced subset of samples and will directly affect the grouping of unknowns (pre-assigned *POPFLAG=0*). This model can be useful when grouping individuals/populations by comparison with very well-defined reference data—this option is particularly useful in the forensic context.

8. There are two allele frequencies models: *independent allele frequencies* and *correlated allele frequencies*. The first is used when frequencies are reasonably different in distinct populations—this implies that knowledge about the correlation level across the population is needed. The second assumes a non-independence level and offers more power to detect distinct populations that are closely related (e.g., Chinese and Japanese)—in the absence of high correlation levels, this model gives the same results as the *independent allele frequencies* model.

9. *STRUCTURE* does not assume a particular mutation process so the scale of the number of repeat units in STRs is not considered (only allele frequencies are important). For this reason there is no need to multiply all other alleles by 10 to compensate the transformation of intermediate alleles (19.3 = 193).
10. To confirm the number of markers and individuals select "Show data file format" showing total lines and columns. As an example, the data file format information for Table 1 would state: one line with m columns (m corresponds to the number of markers) and n lines with $m+6$ columns (four columns with prior information and two with extra information) with $n/2$ individuals (two lines per genotype).
11. When preparing the input file following the example presented in Table 1, there is no need to select the "special format" option because by default *STRUCTURE* assumes genotypes are arranged as two consecutive rows (diploid species) per individual. If both alleles are in the same line but in consecutive columns select "Data file stores data for individuals in a single line".
12. A *burnin* period of 10,000–100,000 is sufficient to observe convergence to an equilibrium point of parameters such as α—the relative admixture levels between populations. To check the variation of the parameter values go to the "Data plot" option in the simulation results window. When excessive variation is observed at the end of the *burnin* period, it is necessary to increase its length. To select an appropriate number of MCMC steps after the *burnin*, it is advisable to perform several simulations for each K value considering different lengths to see if the results are consistent—usually 10,000–100,000 MCMC steps are enough but to obtain precise posterior probability estimates longer simulations might be needed.
13. *STRUCTURE* performs individual analyses for each assumed population number from one up to a reasonable number for the sampling regime—at least three K values more than the number of expected population. If a *plateau* on the posterior probabilities is not reached, larger K values might be needed. Furthermore, clustering algorithms such as the one implemented in *STRUCTURE* can show stochastic variation from the simulations. To diminish their effect, several replicates for each K value should be made (at least three to five replicates advised).
14. Computational times can vary depending on the number of markers and samples to be analyzed, but also on the analysis parameters selected. As a point of reference, running a project

using the example input file supplemented with the online version of this chapter took approximately 3 h 45 min in a computer with a 2.7GHz Intel Core i7 processor. The project included two parameter sets: *admixture* and *admixture POPFLAG*—both were performed through 100,000 burnin steps, 100,000 MCMC repeats, three replicates from $K = 2$ to $K = 6$, and correlated allele frequencies.

15. Independently of the origin of differences between clustering results, a method to deal with replicate results is needed. *CLUMPP* uses replicates of the estimated membership proportion matrices for any *K* number—the result is a set of permutated matrices so that all the replicates have the best correspondence possible. It also generates a matrix that corresponds to the median of the permutated matrices. This is done for the population and individual proportions matrices. Currently the easiest way to prepare input files for *CLUMPP* is with *Structure Harvester* (*see* **Note 6**). Two files are needed: *.indfile* includes individual ancestry membership proportions tables from all replicates per *K* value and *.popfile* includes population ancestry membership proportions tables for all replicates per *K*. In both a blank line separates each table. If the input files are prepared manually take care with the sample name, which must be numeric; if alpha-numerical an error message appears. Both *.indfile* and *.popfile* files, together with *paramfile* and others, must be saved in the software folder together with the executable file. The *paramfile* includes important parameters that must be adjusted: *DATATYPE* defines which data is going to be considered for analysis (0 = individual; 1 = population); *K* is the number of clusters; *C* is the number of individuals or populations (depending on the selected *DATATYPE*); *R* is the number of replicates; *M* is the algorithm used. We recommend *M = 1* so all possible permutations are performed, but with large *K* or *R* values *M = 2* (10,000 random input repeats) is sufficient and for *K* values above 15 *M = 3* is advisable. Metric *S* is the pairwise matrix similarity statistic and we recommend the standardized *G′* (select *S = 2*). It is important to note that to obtain a population and individual mean matrices, two runs are required, adjusting the *.output* file name (no name change overwrites the first run), *DATATYPE* and *C* in between runs. In the Windows version, just execute the *CLUMPP.exe* file—a *cmd* window opens showing the progress of the simulation. In the MacOS version execute the software through the terminal command line: change the directory to the *CLUMPP* folder location (type *cd* > drag the folder into the terminal > ENTER) and then type *./CLUMPP* > ENTER to run the software.

16. A convenient way of visualizing *STRUCTURE* results (especially for $K > 3$) is to show each individual as a straight segment divided into K colors that represents the estimated ancestry membership proportions. *STRUCTURE* gives such bar plots but their format cannot be changed and they only present replicate results for one K value. *Distruct* offers a great variety of options to generate more informative cluster plots. As with *CLUMPP*, *distruct* uses a set of files stored in the same folder of the executable file. The input files include the population Q-matrix (*.popq* file) and the individual Q-matrix (*.indivq* file) obtained directly from *STRUCTURE* (in the case of a single K replicate) or from *CLUMPP*. Files: *.names* and *.languages* define the labels above and below the plot. Both files have the same format: in each line write the population numeric code, space, and preferred name. To define the color of each cluster open the *.perm* file—with a minimum K lines each defining a color. Colors are assigned to each cluster and not to each predefined population, i.e., if population 1 appears in orange and this population is defined in cluster 3 in *.perm*, define "3 orange" and not "1 orange". The *drawparams* file has several modifiable parameters, notably: K, *NUMPOPS* (number of populations) and *NUMINDS* (number of individuals). Remaining parameters adjust graphical aspects of the plot: letter size, distance between text and plot, height of the plot, thickness of the columns representing individuals, thickness of the contour lines, horizontal/vertical orientation, and others (Fig. 6). The "//" symbol indicates that the following text is a comment describing the parameter function and that it will not be used by the software. When computations are complete a *PostScript* (*.ps*) file containing the plot is created To visualize the plot with Windows, specific software such as *GhostView* (freely available online), Acrobat Distiller or Illustrator is needed. With MacOS plots can be opened with preview and exported as a pdf.
17. A new option has been added to Snipper: the ability to apply or not apply the HWE principle. The assumption of HWE when not valid may result in inaccurate genotype frequency estimates and, in turn, an inaccurate classification.
18. Each missing genotype is entered as two Ns per SNP, so only ACGTN characters are permitted. Blank spaces are ignored. Format errors from incorrect bases (either due to incorrect SNP order or inverted bases) are flagged by Snipper with a warning for the relevant SNP position(s).
19. In the supplementary Excel input file included in the online version of this chapter, an example concatenated profile is given. This individual will be classified as African—use Option

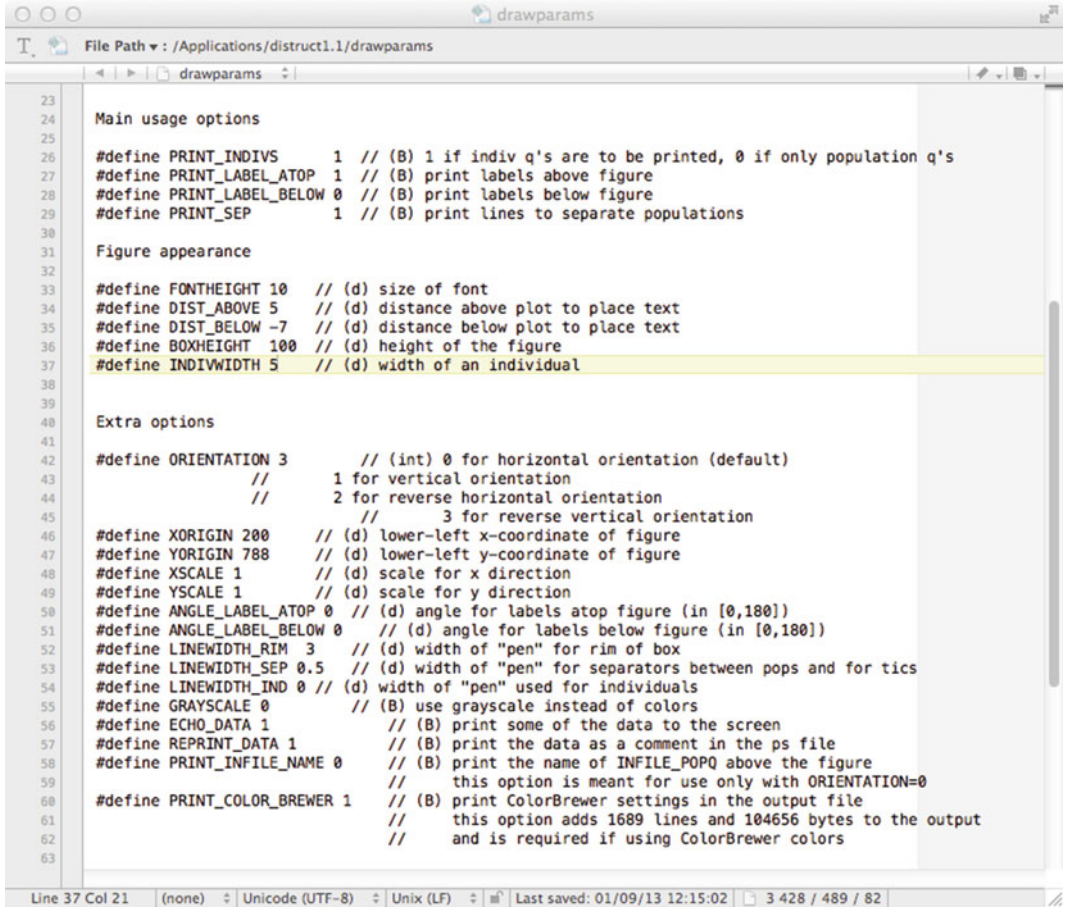

Fig. 6 Example of *distruct* parameters. Considering an output file resembling an A4 sheet lying horizontally (longer side down), and depending on the number of samples and the desired effect, it is worth taking advantage of available space. Reverse horizontal orientation uses the longest side of the virtual sheet. Changing the *XORIGIN* and *YORIGIN* values also helps—for example, *XORIGIN* = 200 moves the plot away from the margin of the sheet and *YORIGIN* = 788 leaves just enough space to separate the plot from the margin without leaving much unused space. The individual bar width (*INDIWIDTH*) can then be adjusted to an appropriate value that allows all the individuals to be included in the virtual sheet

1 or Option 2 in *Snipper* as described above (for Option 2 use the supplementary Excel file as population data input—remove the last two samples (unknown ancestry) and the last column; adjust the number of individuals in cell 1A to 479 and the number of populations in cell 1C to 3). Note that -log(LIKELIHOOD) values are returned, so lower values are better.

20. Divergence is calculated based on the number of populations included in the comparison. For example, on the fixed training

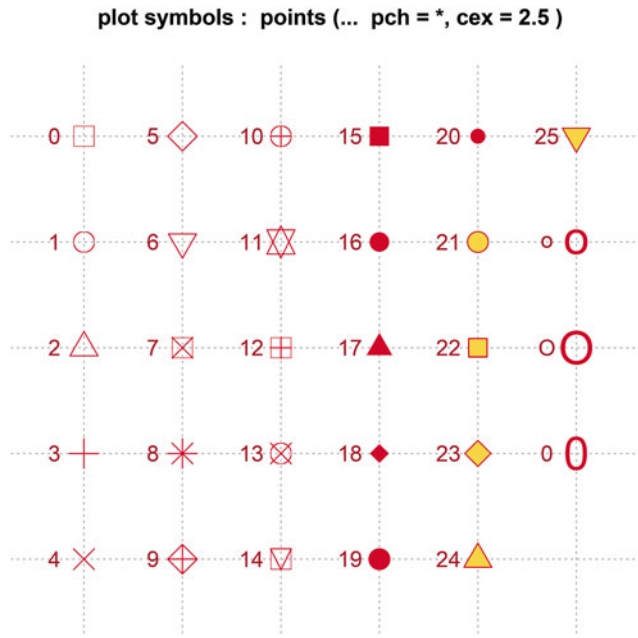

Fig. 7 Symbols available in *R* to define the points shown in plots (*pch* command)

set option, divergence will be calculated based on 3, 4, or 5 groups depending on the option selected in Step 2.

21. When generating PCA plots, it is possible to change graphical parameters to user needs. In the *R* script, command lines are included so the "Population" information can be used to define color of the symbols in plots. In this case, population names are transformed into color names, in population alphabetical order and not input order (a complete list of color available in *R* can be obtained with the command *colours()* (alternatively *colors()*) or with the help of the *Chart of R colours* available at http://research.stowers-institute.org/efg/R/Color/Chart/). The symbol shape can also be changed—for information on available symbols use the *pch* help page by typing *?pch* in R console (Fig. 7). The *cex* command changes the size of the points.
22. If you are using *SNPassoc* package for the first time, you need to install it in *R*. Two options are available: (1) install it from the Package installer option in the *R* console; or (2) download the package zip file from http://cran.r-project.org/web/packages/SNPassoc/index.html and perform a local zip file installation.

References

1. Jobling M, Hollox E, Hurles M et al (2014) Human evolutionary genetics: origins, peoples & disease, 2nd edn. Garland Science - Taylor & Francis Group, New York
2. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155(2): 945–959
3. Phillips C, Salas A, Sánchez JJ et al (2007) Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Sci Int Genet 1(3–4): 273–280
4. Fondevila M, Phillips C, Santos C et al (2013) Revision of the SNPforID 34-plex forensic ancestry test: assay enhancements, standard reference sample genotypes and extended population studies. Forensic Sci Int Genet 7(1):63–74
5. Pereira R, Phillips C, Pinto N et al (2012) Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing. PLoS One 7(1):e29684
6. Jolliffe I (2002) Principal component analysis. Springer, New York
7. R Development Core Team (2011) R: a language and environment for statistical computing. http://www.r-project.org
8. Phillips C (2009) SNP databases. In: Komar AA (ed) Single nucleotide polymorphisms, vol 578, Methods in molecular biology. Humana, New York, pp 43–71
9. Amigo J, Salas A, Phillips C et al (2008) SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access. BMC Bioinformatics 9:428
10. Amigo J, Phillips C, Lareu MV et al (2008) The SNPforID browser: an online tool for query and display of frequency data from the SNPforID project. Int J Legal Med 122(5): 435–440
11. Amigo J, Salas A, Phillips C (2011) ENGINES: exploring single nucleotide variation in entire human genomes. BMC Bioinformatics 12:105
12. Amigo J, Phillips C, Salas A et al (2009) pop.STR—an online population frequency browser for established and new forensic STRs. Forensic Sci Int Genet Suppl Ser 2(1):361–362
13. Rosenberg NA, Pritchard JK, Weber JL et al (2002) Genetic structure of human populations. Science 298(5602):2381–2385
14. Rosenberg NA, Li LM, Ward R et al (2003) Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73(6): 1402–1422
15. Kalinowski ST (2011) The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure. Heredity 106(4):625–632
16. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14(8):2611–2620
17. Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour 4(2):359–361
18. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164(4):1567–1587
19. Hubisz MJ, Falush D, Stephens M et al (2009) Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour 9(5):1322–1332
20. Porras-Hurtado L, Ruiz Y, Santos C et al (2013) An overview of STRUCTURE: applications, parameter settings and supporting software. Front Genet 4:98
21. Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23(14):1801–1806
22. Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. Mol Ecol Notes 4(1):137–138
23. Gonzalez JR, Armengol L, Sole X et al (2007) SNPassoc: an R package to perform whole genome association studies. Bioinformatics 23(5):644–645
24. Phillips C, Prieto L, Fondevila M et al (2009) Ancestry analysis in the 11-M Madrid bomb attack investigation. PLoS One 4(8):e6583
25. Cann HM, de Toma C, Cazes L et al (2002) A human genome diversity cell line panel. Science 296(5566):261–262
26. Rosenberg NA (2006) Standardized subsets of the HGDP-CEPH human genome diversity cell line panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet 70:841–847
27. Li JZ, Absher DM, Tang H et al (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319(5866):1100–1104