Population Characteristics in Forensic Genetics

Faltus, Václav
2007

Population Characteristics in Forensic Genetics

Post-Graduate Student:
MGR. VÁCLAV FALTUS, MSC.
Department of Medical Informatics
Instutite of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague 8, CZ
faltus@euromise.cz

Supervisor:
PROF. RNDR. JANA ZVÁROVÁ, DRSC.
Department of Medical Informatics
Instutite of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague 8, CZ
zvarova@euromise.cz

Field of Study:
Biomedical Informatics

The work was supported by the grant 1M06014 of the Ministry of Education of the Czech Republic.

Abstract

The aim of this paper is to present some methods used in forensic genetics. Forensic genetics is only one part of a wide spectrum of sciences called forensics. It includes identification of victims of natural disasters, mass transportation accidents and industry accidents. It also includes identification of offenders of a crime and determination of paternity. Our current work involves analysis of the genetic data from the Czech population. Therefore and in concordance with other international studies, we focus on methods used in analysis of the genetic profiles of the STR (short tandem repeat) polymorphisms.

Keywords: forensic genetics, identification, STR, statistical methods.

1. Introduction

Within cells, DNA is organized into structures called chromosomes and the set of chromosomes within a cell is called genome. All genes are arranged linearly along the chromosomes. There are 23 pairs of chromosomes in a human body. Almost every cell then contains two sets of chromosomes, one from each parent, 23 chromosomes inherited from mother and 23 from father. One chromosome is the sex chromosome and the others are called autosomal chromosomes. There are areas at chromosomes that we call loci. One locus can contain a gene, part of a gene or only short sequence of nucleotides - letters of the genetic code. Accordingly locus is only an area on particular chromosome described by its unique number. The genes are, what give us lungs, brains, bones, hair-color, allow us to reproduce and think. But there is also plenty of DNA which is proved or believed to have no effect on any processes in our body - the junk DNA.

In DNA there are places (loci) where patterns of two or more nucleotides repeat and where the repeated sequences are adjacent to each other. STR (Short Tandem Repeat) loci are loci, where we observe not very large count of those repetitions. The pattern repetition length (x) usually varies from 2 to 10 letters of the genetic code. The number of adjacent sequences usually varies from 1 to 35. These counts do not need to be integers. If there is only few (say r) first letters (less than n) of the pattern sequence at the end of the loci we write the length of the STR as decimal number $x.r$. Although is it not very common, the different lengths of the STR loci can be called alleles. As it was already mentioned one autosomal chromosome consists of two parts, one from each parent. Therefore there are two corresponding STR loci and one allele is maternal and one paternal. By combining an information from several STR loci we get the genotype of the individual. In forensic genetics this is usually called the genetic profile. You can see one example in Table 1.

locus	a_f	a_s	*locus*	a_f	a_s
D3S1358	17	18	D16S539	12	13
TH01	8	9.3	CSF1PO	10	12
D21S11	28	30	PentaD	10	13
D18S51	13	14	Amelo	X	Y
PentaE	5	10	vWA	16	16
D5S818	8	11	D8S1179	13	13
D13S317	12	13	TPOX	11	11
D7S820	8	10	FGA	21	23.2

Table 1: Example genetic profile

Since the STR loci are very polymorphic, they are sometimes called STR polymorphisms instead of STR loci. This polymorphous nature makes them very useful in forensic genetics. Second characteristics is that the STR loci typically lie in the non-coding region of DNA which makes them the junk DNA. Therefore it is believed that selection pressure does not influence these loci.

The genetics profiles are useful in situations when we are looking for particular individuals. In other situations the individual profile information can be neglected and the database helps to estimate the allele frequencies. Let n denote the total number of alleles existing on one locus. The information from that locus can than be summarized into a Table 2 where g_{ij} $(i, j = 1, \ldots, n, i \leq j)$ are the counts of observed genotypes in the population. Because it is usually not possible to determine which allele was inherited from which parent, the the upper right corner of the table is empty with the genotype counts added to the bottom left corner of the table.

	a_1	a_2	$\cdots$	a_n
a_1	g_{11}			
a_2	g_{12}	g_{22}		
$\vdots$	$\vdots$	$\vdots$	$\ddots$	
a_n	g_{1n}	g_{2n}	$\cdots$	g_{nn}

Table 2: General genotype table

In this paper we start with explaining the terms homo- and heterozygosity. Then we turn to average match probability, discrimination power, polymorphic information content, average exclusion probability and typical paternity index. All statistics presented here will also be demostranted on one locus from [1] and as future work will be implemented into the R package forensic [2]. We take the STR locus denoted $D13S317$. All information form this locus is summarized in Table 3.

	8	9	10	11	12	13	14	15
8	16							
9	20	8						
10	19	3	3					
11	124	64	47	131				
12	69	51	30	192	70			
13	26	23	18	64	61	12		
14	10	7	5	27	24	8	0	
15	0	0	0	0	2	0	0	0

Table 3: STR locus D13S317

2. Population Statistics

Homo- and heterozygosity, average match probability, discrimination power, polymorphic information content, average exclusion probability and typical paternity index are all straightforward statistics used in forensic genetics. They help in planning and performing genetic experiments as well as in national programs for identification of victims and crime offenders.

2.1. Homo- and heterozygosity

An individual is called homozygote when its both alleles of one gene are the same and it is called heterozygote if the alleles differ. When analysing the STR data, the individual is homozygous if it inherited both alleles of the same length. The individual is heterozygous if the allele lengths differ. The proportions of homo- and heterozygous individuals in population is called homo- and heterozygosity.

Let us assume that there are either heterozygotes or homozygotes in the population. Let X be the number of successes (either choosing heterozygote or homozygote from population). Then X might be supposed to be a random variable with binomial distribution taking values $0, 1, \ldots, n_g$, where n_g is the total sample size.

As long as the proportion of heterozygotes (or either homozygotes) is usually not close to 0 or 1 and we expect the sample size n_g be more than 30 with $n_g p$ being more than 5, the normal approximation should perform well. Let $\hat{p} = r/n_g$, where r is the number of successes and n_g number of observed genotypes, be the proportion of successes estimated from the sample. The confidence interval is then

$$\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n_g}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n_g}}\right), \tag{1}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a standard normal distribution $N(0,1)$.

2.1.1 Example: Suppose we have collected genotype data from 1134 people. Suppose the locus of our interest is the $D13S317$ (see Tab.3) and that we observed 894 heterozygous genotypes. The estimated heterozygosity is then

$$\hat{he} = \frac{894}{1134} = 0.7884 \tag{2}$$

with 95% confidence interval being $(0.7634, 0.8118)$. Estimated homozygosity is analogously

$$\hat{ho} = 1 - \hat{he} = \frac{240}{2000} = 0.2116 \tag{3}$$

with 95% confidence interval being $(0.1882, 0.2369)$.

2.2. Average match probability

Let us assume that the innocent suspect is drawn from the same population as the offender but the two are not closely related. DNA profiles of two individuals are declared to be a match if they exhibit identical genotypes. For an polymorphic locus with k alleles the average

match probability [3][4] is

$$\hat{p_m} = \sum_{i=1}^{n} \widehat{p_m^{ho}}(i) + \sum_{i=1,j=1,i<j}^{n} \widehat{p_m^{he}}(i,j), \qquad (4)$$

where n is the number of possible alleles at the locus and $\hat{p}_m^{ho}(i)$ and $\hat{p}_m^{he}(i,j)$ are estimated probabilities of match with individuals being homozygous or heterozygous. Basically it is a sum of weighted probabilities of homozygous and heterozygous genotypes. Since we assume the innocent suspect and the criminal to be drawn from the same population, the weights are simply the same probabilities as the probabilities of drawing the offender genotypes.

$$\begin{aligned} \widehat{p_m^{ho}}(i) &= \omega_i \psi_i \\ \widehat{p_m^{ho}}(i) &= \omega_{ij} \psi_{ij} \end{aligned} \qquad (5)$$

with

$$\begin{aligned} \omega_i &= \hat{p_i}[\theta + (1-\theta)\hat{p_i}] \\ \psi_i &= \frac{[2\theta + (1-\theta)\hat{p_i}][3\theta + (1-\theta)\hat{p_i}]}{(1+\theta)(1+2\theta)} \end{aligned}$$

and

$$\begin{aligned} \omega_{ij} &= 2(1-\theta)\hat{p_i}\hat{p_j} \\ \psi_{ij} &= \frac{2[\theta + (1-\theta)\hat{p_i}][\theta + (1-\theta)\hat{p_j})]}{(1+\theta)(1+2\theta)}, \end{aligned}$$

where $\hat{p_i}$ and $\hat{p_j}$ are the estimated allele frequencies and $i, j = 1, \ldots, n$.

The quantity θ is a number from the interval $[0, 1)$. It is the coancestry coefficient and describes variation in allele proportions among subpopulations. When $\theta = 0$ the whole population is in Hardy-Weinberg equilibrium. Hardy-Weinberg (HW) equilibrium assumes random mating and random segregation of alleles in large population, where there is no genetic drift and mutations occur randomly. For more details about HW equilibrium please see [5] and [6]. For populations such as USA, the recommended value of θ is 0.01 and for small isolated subpopulations it is 0.03.

By taking θ equal to 0 the first formula from (5) simplifies to $\hat{p_i}^4$ which is equivalent of drawing four times the i^{th} allele from very large population. The second formula from (5) then simplifies to $4\hat{p_i}^2\hat{p_j}^2$ and this is equivalent to twice drawing the i^{th} and j^{th} allele in pair.

2.3. Average discrimination power

Average discrimination power [7] is a potential power to differentiate between any two people drawn at random from population. Here we see that it is the exact opposite to the average match probability. The average discrimination power is therefore defined as

$$\hat{p_d} = 1 - \hat{p_m}, \qquad (6)$$

where $\hat{p_m}$ is the estimated match probability from (4).

2.3.1 Example (cont.): The average match probability and average discrimination power for the $D13S317$ locus are shown in Table 4. We can see that with increasing coancestry coefficient the average match probability increases too and the average discrimination power decreases.

	$\theta = 0$	$\theta = 0.01$	$\theta = 0.03$
AMP	0.0780	0.0833	0.0943
ADP	0.9220	0.9167	0.9057

Table 4: Average match probability (AMP) and Average discrimination power (ADP)

2.4. Polymorphic information content

The main contribution of polymorphic information content (PIC) is in the genetic mapping, which plays very important role in genetics and particularly in genetic counseling and research of hereditary diseases or disorders. For more details, please see [8]. Informativeness in this context is represented by the probability that a given offspring of a parent carrying the rare allele at the index locus will allow deduction of the parental genotype at the marker locus. The marker locus is then the polymorphic locus where the informativeness has to be determined.

Polymorphic information content can then be calculated as a sum of the probability of an offspring being informative multiplied by mating frequencies. The probabilities of mating and probabilities of an offspring being informative are given in [8]. The sum of their products is

$$\widehat{pic} = 1 - \sum_{i=1}^{n} \hat{p}_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2\hat{p}_i^2\hat{p}_j^2, \qquad (7)$$

where n is the number of possible alleles at the locus and $\hat{p}_i$ is the estimated frequency of the i^{th} allele ($i = 1, \ldots, n$) at that locus.

The value of PIC varies from 0 to 1. the loci with $PIC \geq 0.5$ are according to (7) called highly informative. Reasonably informative are loci with $0.25 \leq PIC < 0.5$ and only slightly informative are loci with $PIC < 0.25$. Loci with PIC near 1 are most desirable. In here we considered the index and the marker locus to be in nuclear families (one generation). Multigenerational studies will probably allow more extensive computation.

2.4.1 Example (cont.): The estimated PIC for the $D13S317$ polymorphisms is 0.7507 which means that it is highly informative. The other loci from [1] are highly informative too. Their estimated PIC varies from 0.5755 to 0.8832.

2.5. Paternity testing

In paternity testing we first try to exclude men not being fathers of the selected child. After we have only few men left we try to determine if the selected man is the father of the selected child. In the following subsections we will assume that a mother of selected child is undoubtedly known and that there were no mutations occurred at alleles.

2.5.1 Average exclusion probability: The selected man is certainly not a father of the selected child in two cases: The child does not have any allele that could be inherited from selected man. Knowing the allele inherited from mother, the child's second allele does not come from selected man. From previous words we see that the exclusion probability closely corresponds to heterozygosity. The ability to exclude selected man from being father higher as the proportion of heterozygotes in population rises. In remaining cases we can compute the probability of excluding any person from being father. This assumes reliable estimate of allele frequencies.

The calculation is given in [9]. Computation of the expected exclusion probability is not difficult having estimated the heterozygosity (proportion of heterozygotes). Then

$$\begin{aligned} \hat{p}_{exclusion} &= \hat{he}^2[(1-\hat{he}+\hat{he}^2)] \\ &+ \hat{he}^4[\hat{he}(1-\hat{he})], \end{aligned} \tag{8}$$

where $\hat{he}$ is the estimated heterozygosity.

Rougher approximations of this formula are given in [9]. Here they assume the heterozygosity being large enough, or close to 1, and subsequently take $1+\hat{he} \approx 2$. Then

$$\hat{p}_{exclusion} = \hat{he}^2(1-2\hat{he}(1-\hat{he})^2). \tag{9}$$

This approximation is very often used in forensic articles but proves unsatisfying when $\hat{he}$ diverges form 1. Another even rougher approximation is taking $(1-\hat{he}) \approx 0$. This gives

$$\hat{p}_{exclusion} = \hat{he}^2. \tag{10}$$

For multilocus testing let us denote the estimated exclusion probability $\hat{pe}_i$, where the $i = 1, 2, \ldots, n_l$ denotes the i^{th} locus examined and let us assume the n_l loci being independent. The overall probability of exclusion is then

$$\hat{PE} = 1 - [(1-\hat{pe}_1)\ldots(1-\hat{pe}_{n_l})]. \tag{11}$$

2.5.2 Example (cont.): The estimated average exclusion probability and its approximations for the $D13S317$ locus are shown in Table 5.

	acc. to (8)	acc. to (9)	acc. to (10)
AEP	0.5823	0.5776	0.6215

Table 5: Average exclusion probability (AEP)

The most accurate estimate of AEP is 0.5823

2.5.3 Typical paternity index: Typical paternity index [9] is defined as

$$\hat{PI} = \frac{1}{1-\hat{p}_{exclusion}}, \tag{12}$$

where $\hat{p}_{exclusion}$ is the estimated average exclusion probability.

Unfortunately, very often only the roughest approximation of $\hat{p}_{exclusion}$ (10) is taken into account when calculating the typical paternity index. Further taking $1+\hat{he} \approx 2$ leads to

$$\begin{aligned} \hat{PI} &= \frac{1}{(1+\hat{he})(1-\hat{he})} \\ \hat{PI} &= \frac{1}{2(1-\hat{he})}, \end{aligned} \tag{13}$$

where the $\hat{he}$ is the estimated heterozygosity.

Taking the estimate of $\hat{p}_{exclusion}$ (8) we derive

$$\begin{aligned} \hat{PI} &= \frac{1}{1-\hat{he}^2[(1-\hat{he}+\hat{he}^2)]-\hat{he}^4[\hat{he}(1-\hat{he})]} \\ \hat{PI} &= \frac{1}{1-\hat{he}^2+\hat{he}^3-\hat{he}^4-\hat{he}^5+\hat{he}^6}, \end{aligned} \tag{14}$$

where the $\hat{he}$ is the estimated heterozygosity.

2.5.4 Example (cont.): The estimated typical paternity indices and its approximations for the $D13S317$ locus are shown in Table 6. Unfortunately, although the computation of (14) is very simple, some of the approximations is used.

	acc. to (12) & (9)	acc. to (12) & (10)
TPI	2.3675	2.6421

	acc. to (13)	acc. to (14)
TPI	2.3625	2.3939

Table 6: Typical paternity index (TPI)

The most accurate estimate of the TPI is 2.3939.

3. Conclusions

Short overview of population statistics used in forensic genetics was given together with some discussion and references to available literature. As the areas of forensic sciences and forensic genetics are very wide and it is only recently that most of the genotype information is available, there is lots of matters to explore and investigate. In my prospective Ph.D. thesis I would like to focus on statistical aspects of these statistics, testing assumption under which they hold and possibly developing new method for their estimation.

Indispensable is also the development of R [10] - the statistical software and its building blocks - R-packages. Those separate ones usually contain methods and functions for more specific topics of statistics. There are currently two main R-packages available for genetic computations: *genetics* [11] and *gap* [12]. *Genetics* focuses on classes and methods for handling genetic data. *Gap* focuses on data analysis of both population and family data. Even though the R-package *forensic* [2] uses some methods from the two packages mentioned above, its purpose rests on forensic genetics. In future I would like to cooperate with it's authors to develop and include more methods.

References

[1] Halina Šimková, Václav Faltus, Richard Marvan, Tomáš Pexa, Vlastimil Stenzl, Jaroslav Brouček, Ivan Mazura, Jana Zvárová, "Allele frequency data for 17 short tandem repeats in a Czech population sample", *Forensic Science Iternational*, submitted 05/2007.

[2] Miriam Marusiakova (2007), "forensic: Statistical Methods in Forensic Genetics. R package version 0.2.", (Center of Biomedical Informatics, Institute of Computer Science and Academy of Sciences of the Czech Republic).

[3] David J. Balding and Richard A. Nichols, "DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands", *Forensic Science International*, vol. 64, pp. 125–140, 1994.

[4] Guangyun Sun, Stephen T. McGarvey, Riad Bayoumi, Connie J. Mulligan, Ramiro Barrantes, Salmo Raskin, Yixi Zhong, Joshua Akey, Ranajit Chakraborty and Ranjan Deka "Global genetic variation at nine short tandem repeat loci and implications on forensic genetics", *European Journal of Human Genetics*, vol. 11, pp. 39–49, 2003.

[5] Sun Wei Guo and Elizabeth A. Thompson, "Performing the Exact Test of Hardy-Weinberg Proportion for Multiple Alleles", *Biometrics*, vol. 48, pp. 361–372, June 1992.

[6] Mark Huber, Yuguo Chen, Ian Dinwoodie, Adrian Dobra, Mike Nicholas, "Monte Carlo Algorithms for Hardy-Weinberg Proportions", *Biometrics*, vol. 62, pp. 49–53, March 2006.

[7] Sérgio D. J. Pena, "Single-tube single-colour multiplex PCR amplification of 10 polymorphic microsatellites (ALF10): a new powerful tool for DNA profiling", *Pure Appl. Chem.*, vol. 71, pp. 1683–1690, 1999.

[8] David Botstein, Raymond L. White, Mark Skolnick, Ronald W. Davis, "Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms", *Am J Hum Genet*, vol. 32, pp. 314–331, 1980.

[9] Charles Brenner and Jeffrey W. Morris, "Paternity testing calculation in single locus hypervariable DNA probes: Validation and other studies", *Proceeding for The International Symposium on Human Identification*, pp. 21–53, 1989.

[10] R Development Core Team (2007), "R: A language and environment for statistical computing", *R Foundation for Statistical Computing, Vienna, Austria*, ISBN 3-900051-07-0, `www.R-project.org`.

[11] Gregory Warnes and Friedrich Leisch, "genetics: Population Genetics. R package version 1.2.1.".

[12] Jing Hua Zhao in collaboration with other colleagues, and with help from Kurt Hornik and Brian Ripley of the R core development team (2007), "gap: Genetic analysis package. R package version 1.0-11", `www.mrc-epid.cam.ac.uk/Personal/jinghua.zhao/`.