Problems of Forensic Sciences 2007, LXXI, 265–279

ISSN 1230-7483

ON THE SUBTLE SOUTHERN POLISH POPULATION SUBDIVISION

Paulina WOLAŃSKA-NOWAK

Institute of Forensic Research, Krakow, Poland

Abstract

Possible genetic differentiation between a suspect's and the true perpetrator's population may influence the value of DNA evidence. Hence the extent of this possible genetic differentiation of some geographically isolated populations of highlanders from the southern Polish population was the subject of the presented study. Ten microsatellite loci were used to assess the effects of population subdivision on the probability of a chance match between two DNA profiles. Genotype data for the ten analysed loci considered in this study indicated that the total population is in close agreement with the Hardy-Weinberg equilibrium. The mean value of the coancestry coefficient F_{ST}, which is a measure of population subdivision, is 0.0039, which indicates a very subtle differentiation of this population. This value of the F_{ST} parameter has a very limited influence on the frequency of common DNA profiles, but is more marked in the case of rare DNA profiles. However, even in these cases it is not significant for forensic practise, due to the very high discriminatory power of the panel of SGM Plus *loci*.

Key words

DNA evidence; South Poland highlanders; Population subdivision.

Received 8 November 2007; accepted 28 November 2007

1. Introduction

Southern Poland and particularly the Podhale region had less favourable conditions for settlement than other parts of Poland. In the Middle Ages, this area was covered by forest, its plains were swampy and marshy, and the severe climate and poor soils were also discouraging factors. The first settlers came from the valley of the Vistula river and its tributaries – the region of the early cities – Krakow, Bochnia and Sandomierz. Permanent settlements began to develop here in the 13th century. Soon more intensive colonisation of Podhale began, when well-organised Hungarian settlers reached the Spisz region, along the upper part of the Poprad river valley. In the 14th century, numerous German elements from Western Pomerania settled along the Dunajec river in the southern Podhale region. From the south, along the Vah and Orava river valleys, came Valachians (also known as Vlachs). Polish highlanders are usually considered as descendants of Vlachs, who settled this area of the Carpathian Mountains between the 14th and 17th centuries. Analyses of skulls conducted by Polish anthropologists in the 1960s have indicated that they are mostly of Balkan origin. Today, the whole area of the lower Beskidy (*Ziemia Sądecka*) contains two different ethnographic groups: the Sącz Lachy (an old name for the inhabitants of the forests), and mountaineers from Łącko, whose customs, dress, music and dances differ from those of the Lachy. The Lachy's folk culture originated in the heart of Poland, while the culture of the mountaineers from Łącko was influenced by the nomadic Valachian shepherds from the Balkans. Many places scattered along the Poprad River were also settled by Lemks, who are one of four major groups of Ukrainian highlanders inhabiting both sides of the Carpathian Mountains (Fig-

ure 1). During and shortly after World War II, the Lemks who lived within Poland were resettled, partly to the east in the Soviet Union, and partly to western Polish provinces. However, along the eastern bank of the Poprad, there are still abandoned Lemks villages [14]. Hence, the population of the mountainous region of south Poland originated from small and locally isolated populations, which could differ in their genetic composition. Within these small basic units, individuals are often related to each other, due to some shared ancestry and finite sample sizes, which may introduce local levels of inbreeding often resulting in homozygote excess. For forensic purposes, several loci are considered, so hidden population subdivision may result in patterns of linkage disequilibrium, even between physically unlinked loci [10]. Wright [16] proposed that the deviations of genotype frequencies in a subdivided popu- lation may be measured in terms of three parameters: F_{IS}, F_{IT} and F_{ST}, which are called fixation indices: $1 - F_{IT} = (1 - F_{IS})\ (1–F_{ST})$. Nei [11] showed that the above mentioned equation holds true for any situation regardless of the phylogenetic relationships, migration pattern, number of alleles and whether or not there is selection, because they are defined in terms of the present allele and genotype frequencies. Levels of genetic differentiation may be small if one compares large samples of populations, because in this case, subtle differentiation among subgroups of a presumably homogeneous population may cancel each other [10].

Such differentiation can cause problems such as uncertainty in the interpretation of the value of forensic DNA evidence; a particular DNA profile may be more common or rare in one small group of people, because of drift and/or inbreeding. In forensic practice, an STR DNA profile of an individual is a result of typing his/her biological material at several STR polymorphic markers. Hence it could be that even subtle genetic differentiation of the general population may cause potential problems. Some sets of alleles may be more common in a geographically isolated subgroup which contains both the true perpetrator of the crime and the innocent defendant, and therefore forensic assessments which ignore population subdivision may overstate evidential strength of the determined matched profiles [3]. When a match has been obtained between the DNA profile of a defendant and that of a crime scene sample, it is standard practice to report the weight of such evidence in terms of a match probability [4]. Given that the suspect is not the source of a crime scene sample, the match probability represents the probability that another individual, not related to the suspect, in the relevant population would share the matching set of alleles [2].

Calculation of the random match probabilities, which constitutes a measure of the value of DNA evidence, is usually performed under assumption of independence of allelic proportions within and between loci. Balding and Nichols [3] argued that the so called "product rule" could overstate the strength of the DNA evidence by ignoring within-locus correlation arising from the presence of substructure in the general population. They derived a match probability formula to take account of the parameter that measures population differentiation by means of , which reflects the relatedness between the individuals. In practice, it correlates with the above mentioned parameter F_{ST}. The extent of genetic differentiation within the suspect's and possible perpetrator's population may have an important impact on the value of the forensic investigation. In real populations, the genotype frequencies in each subpopulation may not be in agreement with the Hardy-Weinberg equilibrium [11]. The selection of the subpopulations analysed in this study was based on knowledge of the long partial isolation of the above groups, caused by the complicated history of settlement and related cultural aspects. Studying the extent of this possible genetic differentiation of some isolated highland populations from the southern Poland population was the main aim of the presented study. In the research conducted on the subdivision of the southern Polish population, various statistical methods were applied. It also seemed relevant to routine forensic practice to ascertain to what extent applying the coancestry coefficient lowers the value of DNA evidence.

2. Materials and methods

Samples of buccal swabs were obtained from unrelated inhabitants of small villages in southern Poland near to Rzeszów (n = 50), Przemyśl (n = 50), Łomnica (n = 50), Piwniczna (n = 78), Poronin (n = 58) and Zakopane (n = 50). The geographical distance between these analysed groups ranged from about 5 to 200 km (Figure 1). The seventh sample (n = 84) represented a random group of suspects from southern Poland, referred to here after as "random suspects", whose genetic materials were isolated during routine forensic casework.

DNA was extracted with the standard organic method (proteinase K and DTT digestion, followed by phenol-chlorophorm extraction and Microcon 100 concentration). The amounts of DNA in the samples were determined fluorimetrically with Pico Green and

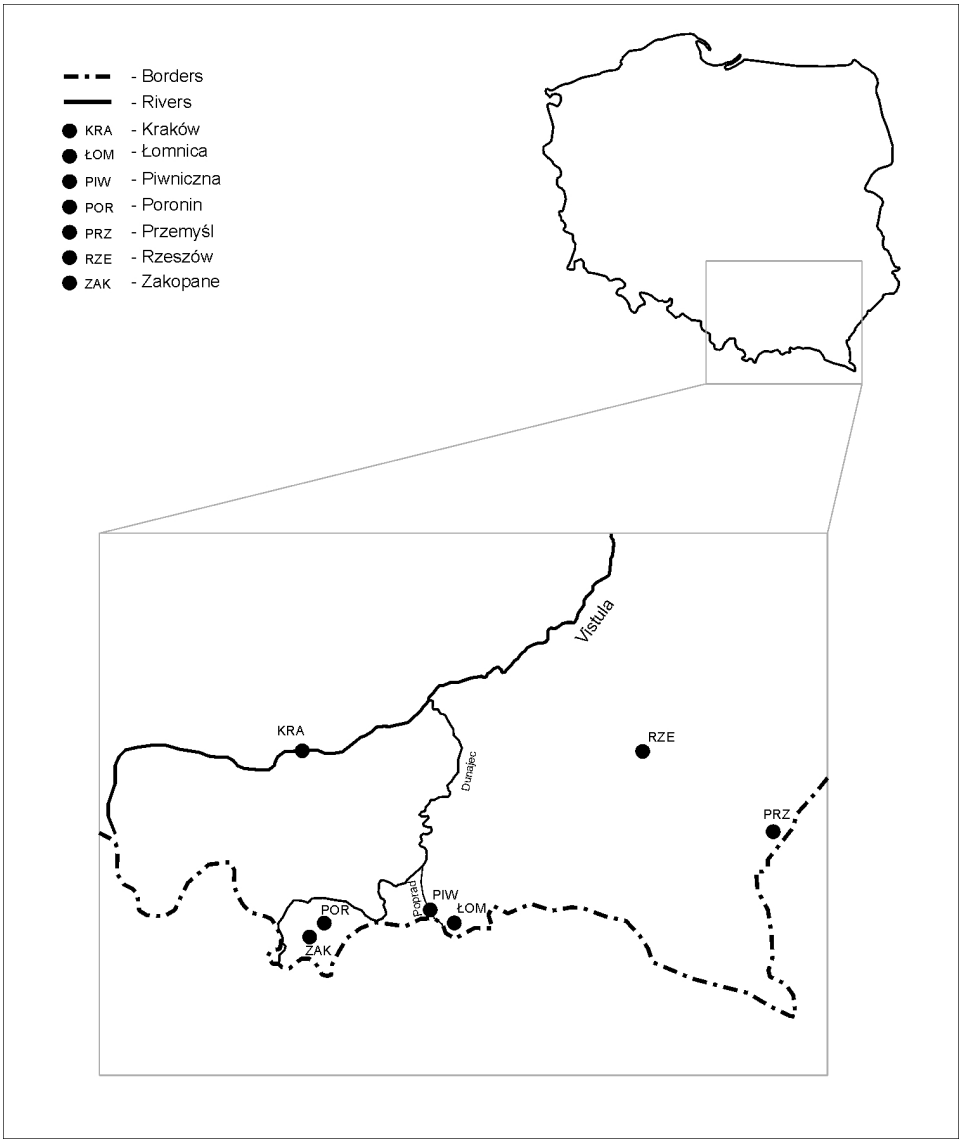

Fig. 1. A map of southern Poland. Dots indicate localisation of analysed groups of individuals.

Fluoroscan Ascent Fl. Amplification of samples was performed with about 1 ng of template DNA in accordance with the manufacturer's instructions [1]. The SGM Plus amplification kit has been optimised to minimise the preferential amplification of smaller alleles and loci. Consequently, the potential for incorrectly assigning homozygosity, which could augment estimates of both consanguinity and substructure, is minimal [7,9,12]. Electrophoresis and typing were done with ABI 310 and Genescan and Genotyper software.

3. Data analysis

The TFPGA program [15] was used for allele frequencies (directly by gene counting) estimation, three estimates of heterozygosity (obtained by direct count of heterozygotes, unbiased heterozygosity and heterozygosity expected under Hardy-Weinberg equilibrium from allelic frequencies), *F*-statistics (the inbreeding coefficients F_{IS}, F_{IT} and F_{ST} calculated according to Weir and Cockerham methods), exact tests for population differentiation (contingency table approach – used to determine significance in differences in allele frequencies among groups of individuals – a Markow Chain Monte Carlo approach was employed that gave approximation of the exact probability of the observed differences in allele frequencies). The Bonferroni correction was used since several tests were performed simultaneously (while a given alpha value may be appropriate for each individual comparison, it is not for the set of all comparisons), so the alpha val-

ues were lowered to account for the number of comparisons being performed.

The mean d^2 values were computed with a simple home made program according to Coltman [8]. Mean d^2 is the squared difference in repeat units between alleles at a microsatellite locus averaged over all loci at which individuals were typed. Mean d^2 has two components: the proportion of loci at which an individual is homozygous (contributing zero to mean d^2), which should correlate with recent inbreeding events and the extent of divergence between alleles at heterozygous loci (contributing to d^2), which should correlate with the degree of outbreeding. Assuming the stepwise mutation model of microsatellite evolution, highly divergent alleles are presumed to have a common ancestor in the more distant evolutionary past than those that are less divergent. Mutations to pre-existing length variants and also no stepwise mutations will interfere with the measurement. However, the greater the number of microsatellite loci used, the more likely the mean d^2 is to reflect ancestral divergence. In a population of randomly mating individuals no variation is expected in the level of ancestral divergence among individuals.

An additional insight into the structure of analysed human population samples was obtained by analysis of a linkage disequilibrium. Allelic association between genotypes in pairs of independent loci, localised on different chromosomes, was examined by a test for congruence between observed and expected genotypic proportions and for *LD* using the GENEPOP software [19]. Assignment of individuals to particular subgroups was performed with the Structure program [18], which implements a model-based clustering method. The model of correlated frequencies was implemented as it turned out to be more effective than the independent frequencies model at detecting subtle population structure.

4. Results and discussion

4.1. Evaluating of fixation indices

The allelic frequency distribution of ten autosomal microsatellite loci in seven population samples living in small villages in southern Poland were analysed. The values of fixation indices, calculated on the basis of alleles proportions in analysed subgroups are presented in Table I.

TABLE I. THE VALUES OF F_{IS}, F_{ST} AND F_{IT} FIXATION INDICES IN POPULATION OF SOUTHERN POLAND

Locus	F_{IT}	F_{ST}	F_{IS}
D3S1358	0.0099	0.0008	0.0091
vWA	0.0277	0.0021	0.0257
D16S539	0.0462	0.0031	0.0432
D2S1338	0.0456	0.0074	0.0384
D8S1179	–0.0472	0.0043	–0.0518
D21S11	0.0244	0.0014	0.023
D18S51	0.039	0.001	0.038
D19S433	–0.0007	0.0042	–0.0049
TH01	–0.025	0.0098	–0.0352
FGA	–0.0085	0.0048	–0.0134
Mean value*	**0.0115**	**0.0039**	**0.0076**
*SD***	0.0098	0.0009	0.0102
Upper***	0.0286	0.0056	0.0254
Lower	–0.0081	0.0023	–0.0133

* – Mean values calculated with the "jacknifing" method throughout all analysed loci; ** – standard deviation of the mean; *** – upper and lower bound of 95% confidence interval calculated with bootstrapping method with 1000 repeats. The mean values for all loci are bolded.

The hierarchical approach incorporating *F* statistics [19] can help to distinguish between two causes of homozygosity excess: hidden substructure and consanguinity. When the population is divided into partially isolated subgroups, individuals from the same subpopulation have an increased probability of sharing a common ancestor and hence an increased probability of homozygosity. This parameter is difficult to estimate: both because of the manner of sample selection, which is absolutely crucial, and the method of estimation of the parameter. The mean value of the F_{ST} parameter in our population is 0.0039 (95% *CI* = 0.0023–0.0056). This indicates a very subtle differentiation of this population. Almost the same result was obtained by Zarrabeitia et al. [27], who analysed micro-geographical population structure in two relatively isolated valleys of Cantabria, a region in northern Spain. The overall F_{ST} value was 0.004. A slightly lower value of 0.003 was obtained for the Byelorussian minority of north-eastern Poland [16].

The value of F_{ST} differs among particular loci. The maximum value in our population was detected for the locus TH01 (0.0098). It is possible that different loci

have a different evolutionary history and hence not the same coancestry coefficients. Probably the mutation rate of the locus is one of the factors that may explain the observed discrepancy. In a previous study [21] performed only on three groups of individuals (living near Krakow, Rzeszów and Poronin) and analysed using three kinds of STR loci (TH01, TPOX and CSF1PO) it was assessed that the mean value of the F_{ST} equals approximately 0.0045. That value and the result obtained in this study are almost the same. However, they are a little higher than the value obtained during a performed comparison of large data sets from the southern Polish population [24] and those from Dolny Śląsk, Białystok and Zielona Góra (data not shown). It is worth adding that the fixation index F_{ST} has not been derived by assuming random mating within subpopulations and therefore it should not be affected by local inbreeding. Its value only depends on local allele frequencies. The correlation between alleles within individuals relative to the alleles of the total population, represented by F_{IT}, has higher values than F_{ST} in our population as a whole (Table I). Hence, the correlation of pairs of alleles within or between individuals is not equivalent. The values of F_{ST} and F_{IT} indicate that there is no random mating within analysed subgroups. The values of indexes F_{IT} and F_{IS} are mutually correlated through analysed loci. Within a subpopulation, the level of inbreeding, as measured by the mean inbreeding coefficient F_{IS}, is small but still higher than zero. Hence, the increased probability of a match between a pair of alleles from one individual compared with pairs drawn from the subpopulation may influence the forensic meaning of the obtained match between two DNA profiles. The average inbreeding value in the studied subpopulations, $F_{IS} > 0$, depends on the number of homozygous loci observed within the DNA profile. The more homozygous the suspect's DNA profile, the more probable that additional matching profiles will be observed within the same subpopulation. On the other hand, individuals heterozygous for many loci are less likely to be observed within this subpopulation.

4.2. The results of mean d^2 calculation

A microsatellite-specific measure, mean d^2, is more powerful in ascertaining the degree of inbreeding and outbreeding than individual heterozygosity. In Table II the mean d^2 value for each examined group is shown. Because of the fact that many examined loci have a lot of irregular allele length (for example loci D21S11, D19S433 and FGA), the difference between alleles corresponds rather to the difference between kinds of reference alleles (from the allelic ladders) than the number of differences in regular repeats. For example: if one individual has in locus D21S11, alleles 29 and 32.2, the assigned difference between them is 4.

The observed number of alleles in each locus in an analysed population sample is a significant parameter which characterises both the level of polymorphism and the markers used. The obtained results indicate that there is a positive correlation between the numbers of observed alleles in particular loci and the values of mean d^2; however, the differences in mean d^2 values in the analysed samples are not statistically sig-

TABLE II. VALUES OF MEAN d^2 IN ANALYSED GROUPS

Locus	*k*	1	2	3	4	5	6	7	s^2	*SD*
D3S1358	8	5.43	5.54	5.89	**5.01**	6.62	5.82	5.3	0.2313	0.4809
vWA	10	3.93	5.18	4.32	4.49	3.08	3.36	4.2	0.4272	0.6536
D16S539	8	3.12	**1.55**	**1.55**	4.75	2.31	2.4	2.49	1.0517	1.0255
D2S1338	12	14.58	12.56	15.5	10.4	12.49	15.06	**5.93**	9.6173	3.1011
D8S1179	11	3.95	5.94	5.94	5.9	5.47	4.51	**2.87**	1.2113	1.1006
D21S11	16	24.34	34.34	26.11	26.95	26.95	**19.09**	27.73	17.799	4.2189
D18S51	16	18.22	12.56	19.84	11.83	15.18	11.29	**10**	11.901	3.4498
D19S433	17	12.9	7.27	7.47	9.81	10.52	9.71	**3.79**	7.2605	2.6945
TH01	6	4.67	**2.92**	3	4.28	6.74	4.82	5.33	1.5175	1.231
FGA	8	26.13	14	21.02	**13.1**	13.82	19.65	18.98	19.473	4.4128

The lowest values of mean d^2 are bolded, *k* – number of alleles in the population, s^2 – variance, *SD* – standard deviation, 1 – random suspects, 2 – Rzeszów, 3 – Przemyśl, 4 – Poronin, 5 – Łomnica, 6 – Piwniczna, 7 – Zakopane.

nificant and seem to be an effect of stochastically explained differences.

4.3. Hardy-Weinberg equilibrium (*HWE*)

Because a departure from *HWE* can be due to population subdivision or inbreeding, the discrepancy between observed and expected genotype frequencies was assessed with the exact probability test [19], which is more powerful than other tests, when the departure from *HWE* is due to specific genotypes. The results of *HWE* analysis in particular subpopulations are illustrated in Table III.

Observed deviations from *HWE* indicate that there may be a real, though weakly expressed, population substructure. The other probable reason for deviations from *HWE* is inbreeding. Interviews with local old people provided information about episodes of marriages between first cousins in small villages occupied by few families. The probability that the offspring of first cousins will inherit two identical by descent (IBD) genes at any locus is 1/16. So this might disturb *HWE* proportions. Hence the presence of consanguinity in ancient times and its persistence into modern times (there have been a lot of incest cases coming to our laboratory in the last ten years) is significant in the context of forensic DNA interpretation [5]. The other reason for deviation from *HWE* may be the presence of many rare alleles with small proportions throughout the analysed groups. Zapata [26] showed that rare alleles with a frequency of less than 3% (usually very short and also those with extreme size) can more strongly influence disequilibrium than others. Such alleles tend to be much more in disequilibrium than the remaining alleles. One can interpret this as a consequence of the high mutation rate of STR loci. Many of the cases of disequilibrium detected between rare alleles might be the result of relatively recent mutational events. Low frequency alleles are more likely to have arisen recently in the population (by the introduction of new mutations) than alleles of moderate frequency, because time is necessary for new alleles to spread. Finally, the combined effect of isolation and of a small effective size may lead to a certain degree of inbreeding and, consequently, to statistically significant departures from *HWE* at microsatellite loci. This seems to be a plausible explanation for presented findings. It is worth noting that the used samples had sufficient sizes, which increased the power of the used tests and provided more opportunities for detecting deviations from *HWE*. Analysis of deviations from *HWE* for each locus revealed that significant deviations arose only for three loci in the Piwniczna sample: D2S138, D18S51 and TH01 (Table III). Finally genotype data for the ten loci in the total population considered in this study are in close agreement with *HWE*, indicating that the substructure of our general population is not detectable.

4.4. Allelic association between pairs of independent loci

To gain further insight into the genetic structure of the analysed population, the linkage disequilibrium was calculated (*LD*) between pairs of loci. The analysis of genotypic linkage disequilibrium may provide

TABLE III. ASSESSED *p*-VALUES FOR HW EXPECTATION IN DIFFERENT SUBPOPULATIONS AND LOCI

Locus	1	2	3	4	5	6	7	Pooled
D3S1358	0.9279	0.4772	0.6612	0.3446	0.5818	0.0873	0.6024	0.9365
vWA	0.0182	0.2350	0.3659	0.3968	0.1722	0.0367	0.2863	0.1014
D16S539	0.6098	0.3989	0.8685	0.5898	0.0719	0.3536	0.6291	0.2280
D2S1338	0.2447	0.0361	0.6800	0.1493	0.2383	**0.0004**	0.0662	0.1664
D8S1179	0.1657	0.6916	0.1422	0.2294	0.2688	0.0371	0.9600	0.911
D21S11	0.3162	0.2208	0.4484	0.5646	0.5826	0.1202	0.4161	0.1303
D18S51	0.3020	0.5838	0.1159	0.4901	0.2053	**0.0000**	0.3323	0.0129
D19S433	0.5212	0.3926	0.3383	0.1819	0.8669	0.3324	0.6729	0.1009
TH01	0.1960	0.1277	0.6435	0.6837	0.2918	**0.0003**	0.0104	0.0135
FGA	0.7384	0.4920	0.4814	0.6038	0.2598	0.3305	0.1409	0.2159

* – The values of $p < 0.005$ are bolded, 1 – random suspects, 2 – Rzeszów, 3 – Przemyśl, 4 – Poronin, 5 – Łomnica, 6 – Piwniczna, 7 – Zakopane. After Bonferroni's adjustment the *p*-value was lowered for each test and equalled 0.005.

useful information on population structure, because a marked heterogeneity in allele frequencies among subpopulations may lead to *LD* between independent loci in the population as a whole [23]. After performing Bonferroni's adjustment for multiple tests there were 3-*p* values in the sample from Zakopane (between D3S1358/D2S338, D3S1358/D19S433 and D21S11/TH01 loci) lower than 0.005. Among 45 possible pairs of loci, this is 6.67% of total comparisons and this is more than expected on the basis of the 5% level of significance. Such gametic disequilibrium could arise due to founder effect, selection, drift or non random mating.

4.5. Exact test of population subdivision

Using the Markov Chain Monte Carlo Method approach with TFPGA software, it was determined that significant differences in allele frequencies really exist among analysed groups of individuals. The obtained results are presented in Table IV.

TABLE IV. THE RESULTS OF THE EXACT TEST OF DIFFERENCES IN ALLELE FREQUENCIES AMONG ANALYSED GROUPS

	1	2	3	4	5	6
2	0.1552					
3	0.2804	0.1104				
4	0.0163	0.0909	0.032			
5	0.0098	**0.0056**	0.3048	**0.0005**		
6	**0.0005**	**0.0055**	0.0103	**0**	**0.0022**	
7	0.3016	0.2169	0.2523	0.0573	0.156	**0.002**

Values of $p < 0.0071$ (after Bonferroni's adjustment) are bolded, 1 – random suspects, 2 – Rzeszów, 3 – Przemyśl, 4 – Poronin, 5 – Łomnica, 6 – Piwniczna, 7 – Zakopane.

As many as seven of the total twenty one comparisons exhibit significant differences in allele frequencies between the seven analysed groups of individuals. The obtained deviations from homogeneity of our total population may be explained by the combined effect of isolation and of small effective size of analysed groups and existence of many rare alleles with low frequencies in particular subpopulations.

4.6. Inferring population substructure by assigning individuals to particular clusters

Analysis of the population of southern Poland was extended by including parameters that specify the subpopulation from which each individual is drawn, following the approach of Prichard, Stephens, and Donnelly [16]. Structure implements a clustering method, based on a Bayesian approach for inferring population structure using genotype data consisting of unlinked markers. It is also very useful in problems where cryptic population structure may be present, as a way of identifying subpopulations. A fairly short run was performed with 1,000,000 permutations using the prior information model, which makes it possible to combine genetic information with data on the geographic sampling location of individuals. A model consisting of four clusters has higher posterior probability than models with other numbers of subgroups. The extent of allele-frequency differences among groups of analysed individuals influences the accuracy of the assignment of cluster individuals to their appropriate subpopulation. As was shown, the range of genetic differentiation is not strongly marked.

4.7. Forensic implications of the obtained results

The discrimination power of the used STR markers calculated on the basis of allele frequencies in the total population of Southern Poland is 5.44 10^{-13}. This means that the probability that two randomly chosen individuals from our population would have the same ten loci DNA profile theoretically equals 1 in nearly two billion.

The implications of this result for forensic interpretation can be illustrated by calculating frequency (in the total population) of two hypothetical DNA profiles which consist of the most rarely and the most commonly observed alleles (Table V). In the case of routine forensic investigations, when a match is obtained between two DNA profiles (one from the crime scene and the other from the suspect) the frequency of a matching DNA profile in the population is calculated [5]. In the simplest cases, it is numerically equivalent to the match probability of these profiles [20]. The match probability quantifies how likely it is to observe a matching STR profile in an alternative suspect, given that the defendant matches the crime scene profile. Assuming that the population structure really exists and taking into account the obtained value of $F_{ST} = 0.0039$, then the probability of matching an alternative suspect from the same subpopulation is slightly elevated. This probability was calculated using the method of Balding and

Nichols [3]. Because the correlations between examined loci are negligible, the probability of a multilocus genotype can be calculated by taking the simple product of single-locus probabilities with an appropriate value of F_{ST}. This kind of calculation takes into account the effects of drift on the allele frequencies at each locus, hence it allows for the linkage disequilibrium that drift has generated between unlinked loci. This treatment avoids the need to assume allelic independence, as well as the need to specify the individual subpopulation.

The hypothesis of equilibrium may be rejected with a sufficiently large sample. However, a forensic scientist may believe that the magnitude of the departure is sufficiently small for the hypothesis of equilibrium, though strictly false, to be adequate for the application at hand. Strong gametic disequilibrium may invalidate this assumption. The obtained data with values of *LD* close to zero indicate that the effect of gametic disequilibrium on the match probabilities involving ten SGM Plus loci may indeed be negligible.

TABLE V. ASSESSED FREQUENCY (*f*) OF TEN LOCI DNA PROFILES

Locus	A – the most common DNA profile	B – the rarest DNA profile
D3S1358	15, 16	13, 19
vWA	17, 18	11, 20
D16S539	11, 12	8, 15
D2S1338	17, 20	15, 26
D8S1179	13, 14	17, 18
D21S11	29, 30	24.2, 36
D18S51	15, 16	13.2, 14.2
D19S433	14, 15	10, 18.2
TH01	6, 9.3	7, 10
FGA	21, 22	21.2, 28
$f\|F_{ST} = 0$	$2.4 \quad 10^{-10}$ $(1 : 4.1 \quad 10^{12})$	$1.0 \quad 10^{-36}$ $(1 : 9.9 \quad 10^{38})$
$f\|F_{ST} = 0{,}004$	$2.8 \quad 10^{-10}$ $(1 : 3.6 \quad 10^{12})$	$9.8 \quad 10^{-34}$ $(1 : 1 \quad 10^{36})$

Headings A and B refer to the most common and rarest alleles in the total population. $f|F_{ST} = 0$ – frequency of DNA profile calculated under assumption of $F_{ST} = 0$, $f|F_{ST} = 0.004$ – frequency of DNA profile calculated under assumption of $F_{ST} = 0.004$.

As is shown in Table V, the obtained value of the coancestry coefficient has a very limited influence on the frequency of a common DNA profile. The influence of the coancestry coefficient on rare DNA profile frequency is more marked, but even in this case it is not forensically significant due to the very high discriminatory power of the panel of SGM Plus loci.

5. Conclusions

In this study, different methods of investigation of the pattern of population subdivision were used. They enabled assessment of different parameters characterising the studied population. The results of our study indicate that the total population is in close agreement with the Hardy-Weinberg Equilibrium. The mean value of the coancestry coefficient F_{ST} is 0.0039, and is relatively higher than the values obtained during analysis of a larger Polish population sample, which suggests a very subtle differentiation of the southern Poland population. However, the obtained value of the F_{ST} parameter has a very limited influence on the frequency of common DNA profiles, but it is more marked in the case of rare DNA profiles. Nevertheless, even in this case it is not significant for forensic practise due to the very high discriminatory power of the panel of SGM Plus loci. On the other hand, application of the population specific coancestry coefficient to calculation of match probability would be appropriate and beneficial for the defendant. Another reason for the weakly expressed differentiation of analysed population could be the low sensitivity of the panel of SGM Plus loci used to detect possible population heterogeneity. Therefore, it seems necessary to perform a further precise population study using more informative markers, such as STRs located on the Y chromosome or appropriate SNPs. The above cited paper by Zarrabaitia et al. indicated that F_{ST} values for Y chromosome STRs were ten times higher than for autosomal STRs [27]. Thus, this result confirms the greater ability of Y-STRs to detect population structure.

Acknowledgements:
I would like to express my gratitude to Dr Wojciech Branicki for his valuable advice and comments imparted during preparation of this manuscript and to Marek Kowalczyk for his kind help in improving a map of southern Poland.

References

1. AmpFlSTR® SGM Plus PCR amplification kit, user's manual, PE Biosystems, 1999.
2. Balding D. J., Donnelly P., Inferring identity from DNA profile evidence, *Proceedings of the National Academy of Science* 1995, 92, 11741–11745.
3. Balding D. J., Nichols R. A., A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity, *Genetica* 1995, 96, 3–12.
4. Balding D. J., Donnelly P., Inference in forensic identification, *Journal of the Royal Statistical Society A* 1995, 158, 21–53.
5. Buckletown J., Triggs C. M., Walsh S. J. [ed.], Forensic DNA evidence interpretation, CRC Press, Boca Raton 2005.
6. Budowle B., Monson K. L., Chakraborty R., 1996, Estimating minimum allele frequencies for DNA profile frequency estimates for PCR-based loci, *International Journal of Legal Medicine* 1996, 108, 173–176.
7. Butler J. M. [ed.], Forensic DNA typing. Biology, technology and genetics of STR markers, Elsevier, Academic Press, Amsterdam 2005.
8. Coltman D. W., Don Bowen W., Wright J. M., Birth weight and neonatal survival of harbour seal pups are positively correlated with genetic variation measured by microsatellites, *Proceedings of the Royal Society of London B* 1998, 265, 803–809.
9. Cotton E. A., Allsop R. F., Guest J. L. [et al.],Validation of the AmpFlSTR SGM Plus system for use in forensic casework, *Forensic Science International* 2000, 112, 151–161.
10. Evett I. W., Weir B. S., Interpreting DNA evidence, [in:] Statistical genetics for forensic scientists, Sinauer Associates, Sunderland 1998.
11. Excoffier L., Handbook of statistical genetics, Balding D. J., Bishop M., Cannings C. [eds.], John Wiley & Sons, Chichester 2003.
12. Gill P., Brinkman B., d'Aloja E [et al.], Considerations from the European DNA profiling group (EDNAP) concerning STR nomenclature, *Forensic Science International* 1997, 87, 185–192.
13. Goldstein D. B., Roemer G. W., Smith D. A. [et al.], The use of microsatellite variation to infer population structure and demographic history in a natural model system, *Genetics* 1999, 151, 797–801.
14. Guzik C., Leśnicki J., Development of rural settlement in Podhale, *Prace Geograficzne* 2003, 112, 161–172.
15. http://herb.bio.nau.edu/~miller (TFPGA – Tools For Population Genetic Analysis).
16. Pepiński W. [et al.], Allele distribution of 15 STR loci in a population sample of Byelorussian minority residing in north-eastern Poland, *Forensic Science International* 2004, 139, 265–267.
17. Presciuttini S., Cerri N., Turrina S. [et al.], Validation of a large Italian database of 15r STR loci, *Forensic Science International* 2006, 156, 266–268.
18. Pritchard J. K., Stephens M., Donnelly P., Inference of population structure using multilocus genotype data, *Genetics* 2000, 155, 945–959.
19. Raymond M. L., Rousset F., An exact test for population differentiation, *Evolution* 1995, 49, 1280–1283.
20. Weir B., Genetic data analysis II, Sinauer Associates Inc., Sunderland 1996.
21. Weir B. S., Cockerham C., Estimating F-statistics for the analysis of population structure, *Evolution* 1984, 38, 1358–1370.
22. Wolańska-Nowak P., Interpretacja wyników ekspertyzy, [w:] Ekspertyza sądowa, Wójcikiewicz J. [red.], Oficyna Wolters Kluwer Business, Warszawa 2007.
23. Wolańska-Nowak P., Application of subpopulation theory to evaluation of DNA evidence, *Forensic Science International* 2000, 113, 63–69.
24. Wolańska-Nowak P., Branicki W., Kupiec T., STR data for SGM Plus and penta E and penta D loci in a population sample from south Poland, *Forensic Science International* 2002, 127, 237–239.
25. Toscanini U., Gusmao L., Berardi G. [et al.], Testing for genetic structure in different urban Argentinian populations, *Forensic Science International* 2007, 165, 35–40.
26. Zapata C., Rodrigues S., Visedo G. [et al.], Spectrum of nonrandom associations between microsatellite loci on human chromosome 11p15, *Genetics* 2001, 158, 1235–1251.
27. Zarrabeitia M. T., Riancho J. A., Lareu M. V. [et al.], Significance of micro-geographical population structure in forensic cases: a Bayesian exploration, *International Journal of Legal Medicine* 2003, 117, 302–305.
28. Zhivotovsky L. A., Suhaib A., Wang W. [et al.], The forensic DNA implications of genetic differentiation between endogamous communities. *Forensic Science International* 2001, 119, 269–272.

Corresponding author
Paulina Wolańska-Nowak
Instytut Ekspertyz Sądowych
ul. Westerplatte 9
PL 31-033 Kraków
e-mail: pwolanska@ies.krakow.pl

Problems of Forensic Sciences 2007, LXXI, 265–279

ANALIZA GENETYCZNA NIEWIELKIEJ SUBSTRUKTURY POPULACJI POLSKI POŁUDNIOWEJ

1. Wstęp

Obszar Polski południowej, a zwłaszcza region Podhala, był mniej sprzyjający dla osadnictwa niż pozostałe obszary Polski. W okresie średniowiecza tereny te porośnięte były lasami, strefa nizinna należała do bagnistych i grząskich, a dodatkowymi niekorzystnymi czynnikami był ostry klimat i jałowe ziemie. Pierwsi osadnicy przybyli na te tereny z obszarów doliny rzeki Wisły i jej dopływów. Stałe osadnictwo zapoczątkowano w 13. wieku. Osadnicy ci przybywali z regionów, na których znajdowały się pierwsze miasta – Kraków, Bochnia i Sandomierz. Wkrótce bardziej zmasowaną kolonizację Podhala rozpoczęło przybycie dobrze zorganizowanych osadników z Węgier, którzy dotarli na Spisz i zasiedlali obszary wzdłuż górnej części doliny rzeki Poprad. W 14. wieku w południowej części Podhala na terenach położonych wzdłuż Dunajca pojawili się liczni osadnicy niemieccy pochodzący z zachodniego Pomorza. Z południa, szlakiem prowadzącym wzdłuż dolin Wagu i Orawy, przybywali Wołosi. Polscy górale są zazwyczaj traktowani jako potomkowie Wołochów, którzy zajęli ten obszar Karpat pomiędzy 14. a 17. wiekiem. Badania czaszek przeprowadzone przez polskich antropologów w latach sześćdziesiątych 20. wieku wykazały, że należą one głównie do osobników pochodzących z regionu Bałkanów. Dziś cały obszar Beskidu Niskiego (Ziemia Sądecka) zamieszkują dwie różne grupy etnograficzne: Lachów Sądeckich (dawna nazwa mieszkańców lasów) oraz górali z Łącka. Ci ostatni niegdyś odróżniali się od Lachów odrębnymi zwyczajami, ubiorem jak również muzyką oraz tańcem. Kultura ludowa Lachów miała swoje polskie korzenie, podczas gdy kultura górali z Łącka pozostawała pod wpływem wołoskich pasterzy z Bałkanów. Wiele miejsc położonych wzdłuż Popradu zostało również zasiedlonych przez Łemków, którzy stanowią jedną z czterech głównych grup ludności zamieszkujących obszary wyżynne Ukrainy po obu stronach Karpat. Podczas II wojny światowej oraz w okresie tuż po jej zakończeniu Łemkowie, którzy zamieszkiwali na terytoriach Polski, zostali przymusowo przesiedleni częściowo na obszary ówczesnego wschodniego Związku Radzieckiego, częściowo na tereny zachodniej Polski. Mimo to, wzdłuż wschodniego brzegu Popradu, wciąż znajdują się ślady porzuconych wiosek łemkowskich [14]. Populacja zamieszkująca obszar górski Polski południowej pochodzi zatem z małych, lokalnie izolowanych populacji, które mogą wykazywać zauważalne różnice genetyczne.

Osoby wchodzące w skład tego typu niewielkich jednostek populacyjnych często wykazują znaczny stopień pokrewieństwa, co wynika ze wspólnego pochodzenia oraz ograniczonej liczebności grupy. W efekcie może dochodzić do pojawiania się regionów o znacznym poziomie endogamii, co ujawnia się poprzez podwyższoną liczbę osobników homozygotycznych. Ukryta substruktura populacji może prowadzić do powstawania nieprzypadkowego wspólnego dziedziczenia alleli, nawet pomiędzy fizycznie niesprzężonymi *loci* genetycznymi [10]. Wright [16] zaproponował pomiar zaburzeń częstości genotypów w populacji posiadającej substrukturę poprzez określenie wartości trzech parametrów: F_{IS}, F_{IT} i F_{ST}, które nazywane są współczynnikami fiksacji (ang. fixation indices) i są ze sobą powiązane, co może zostać wyrażone poprzez równanie: $1 - F_{IT} = (1 - F_{IS})\,(1 - F_{ST})$. Nei [11] wykazał, że równanie to może być stosowane bez względu na zależności filogenetyczne, kierunki migracji, liczbę alleli oraz obecność selekcji naturalnej, ponieważ współczynniki w nim zawarte są definiowane poprzez współczesne częstości alleli i genotypów. Poziomy zróżnicowania genetycznego mogą być małe wówczas, gdy porównywane są duże próby populacyjne. Wynika to z faktu, że subtelne różnice istniejące pomiędzy podgrupami populacji mogą się nawzajem znosić i cała populacja sprawia wrażenie homogennej [10].

Takie zróżnicowanie może sprawiać różnego typu problemy związane z niepewnością co do interpretacji wartości dowodu z badania DNA. Dany profil DNA może być bardziej częsty lub rzadki w przypadku pojedynczej niewielkiej grupy ludzi w związku z istnieniem dryfu genetycznego lub (i) zjawiska endogamii. Indywidualny profil DNA jest uzyskiwany dla potrzeb badań sądowych poprzez analizę próbki biologicznej w zakresie kilkunastu polimorficznych markerów typu STR. Nawet niewielka substruktura populacji generalnej może potencjalnie prowadzić do problemów interpretacyjnych. Wynika to z faktu, że niektóre układy alleli mogą być częstsze w przypadku geograficznie izolowanej grupy, do której należy zarówno winny przestępstwa, jak i niewinny podejrzany. Oznacza to, że obliczenia wartości dowodu prowadzone dla celów sądowych, które nie uwzględniają istniejącej substruktury populacji, mogą zawyżać wartość dowodową wynikającą ze stwierdzonej zgodności profili DNA [3]. Przyjętym sposobem podawania wartości dowodu wynikającego z ujawnionego dopasowania pomiędzy profilami DNA podejrzanego oraz stwierdzonego w próbce pochodzącej z miejsca zdarzenia kryminalnego jest przedstawianie tak zwanego prawdopodobieństwa zgodności [4]. Przyjmując, że podejrza-

ny nie jest źródłem próbki zabezpieczonej na miejscu zdarzenia kryminalnego, prawdopodobieństwo dopasowania oznacza prawdopodobieństwo, że to inna osoba z odpowiedniej populacji, która jest niespokrewniona z podejrzanym, posiada profil DNA zgodny z tym stwierdzonym w dowodowej próbce [2].

Obliczanie prawdopodobieństwa przypadkowej zgodności, które jest miarą wartości dowodu z badania DNA, przeprowadza się zazwyczaj zarówno przy założeniu niezależności proporcji alleli w obrębie lokus, jak i pomiędzy różnymi *loci*. Balding i Nichols [3] wskazywali, że stosowanie tak zwanej „reguły mnożenia" może prowadzić do zawyżenia wartości dowodu z badania DNA w związku z ignorowaniem korelacji pomiędzy allelami w obrębie jednego lokus, która wynika z substruktury istniejącej w populacji generalnej. Wyprowadzili oni wzór służący do obliczania prawdopodobieństwa zgodności, który uwzględnia parametr stanowiący miarę zróżnicowania populacji. Wartość tego parametru jest odzwierciedleniem pokrewieństwa pomiędzy osobnikami. W praktyce jest on skorelowany z przywoływanym powyżej parametrem F_{ST}. Zakres zróżnicowania genetycznego charakterystyczny dla populacji, do której należy podejrzany i prawdopodobny sprawca przestępstwa, może mieć istotny wpływ na wartość dowodu z badania DNA. W rzeczywistych populacjach częstości genotypów charakterystyczne dla poszczególnych subpopulacji mogą nie pozostawać w zgodzie z regułą Hardy'ego-Weinberga [11]. Przy wyborze subpopulacji analizowanych w niniejszej pracy, kierowano się wiedzą na temat ich długotrwałej częściowej izolacji wynikającej ze skomplikowanej historii zasiedlania regionu i związanych z nią aspektów kulturowych. Analiza rozmiarów prawdopodobnego zróżnicowania genetycznego charakterystycznego dla izolowanych społeczności zamieszkujących Polskę południową była podstawowym celem pracy. Badania nad substrukturą populacji Polski południowej przeprowadzono za pomocą szeregu różnych metod statystycznych. Istotnym zagadnieniem był również problem rozmiaru ewentualnego wpływu stosowania współczynnika wsobności na obniżenie wartości dowodu z badania DNA.

2. Materiały i metody

Próbki badawcze w postaci wymazów z jamy ustnej pobrano od niespokrewnionych mieszkańców małych wiosek z terenów Polski południowej: z okolic Rzeszowa (n = 50), Przemyśla (n = 50), Łomnicy (n = 50), Piwnicznej (n = 78), Poronina (n = 58) i Zakopanego (n = 50). Dystans geograficzny pomiędzy miejscami zamieszkiwanymi przez analizowane grupy wynosił od 5 do 200 km. Siódma próba populacyjna (n = 84) zawierała grupę przypadkowych podejrzanych z regionu Polski południowej, których DNA izolowano do prowadzonych spraw sądowych, określanych dalej jako „przypadkowi podejrzani".

DNA izolowano za pomocą standardowej metody organicznej (trawienie struktur komórkowych w obecności proteinazy K i DTT, ekstrakcja mieszaniną fenolowo-chloroformową i zagęszczanie DNA na kolumienkach Microcon 100). Stężenie DNA w badanych próbkach określano metodą fluorymetryczną z zastosowaniem barwnika PicoGreen oraz urządzenia Fluoroscan Ascent Fl. Amplifikację DNA prowadzono zgodnie z protokołem zalecanym przez producenta zestawu SGM Plus, stosując 1 ng matrycy DNA [1]. Zestaw SGM Plus stosowany do amplifikacji DNA został zoptymalizowany przez producenta tak, aby zminimalizować zjawisko preferencyjnej amplifikacji krótszych alleli i *loci*. Zmniejszono przez to możliwość nieprawidłowego oznaczania homozygot, które mogłoby zawyżyć zarówno szacunek poziomu bliskiego pokrewieństwa, jak i substruktury populacji [7, 9, 12]. Rozdział i analizę produktów reakcji PCR prowadzono z zastosowaniem aparatu ABI 310 i programów GeneScan oraz Genotyper.

3. Analiza danych

Program TFPGA [15] stosowano do obliczania częstości alleli (metodą bezpośredniego zliczania obserwacji), trzech przybliżeń heterozygotyczności uzyskanych dzięki zastosowaniu metod: bezpośredniego zliczania heterozygot (heterozygotyczność obserwowana), obliczania heterozygotyczności nieskierowanej i na podstawie częstości alleli przy założeniu równowagi Hardy'ego-Weinberga (heterozygotyczność oczekiwana), współczynników statystycznych F (współczynniki populacyjne F_{IS}, F_{IT}, F_{ST} obliczane zgodnie z metodami zaproponowanymi przez Weira i Cockerhama) oraz testu dokładnego pozwalającego na oszacowanie zróżnicowania populacji (metoda tabeli kontyngencji stosowana w celu określenia znaczenia różnic częstości alleli pomiędzy grupami osobników z wykorzystaniem metody Monte Carlo łańcuchów Markowa, która pozwoliła na przybliżenie dokładnego prawdopodobieństwa obserwowanych różnic w częstościach alleli). Korektę Bonferroniego zastosowano w związku z jednoczesnym wykonywaniem kilkunastu testów (odpowiednia wartość istotności statystycznej alfa zdefiniowana dla pojedynczych obliczeń przestaje być prawidłowa przy jednoczesnym wykonywaniu wielu obliczeń), co doprowadziło do obniżenia wartości alfa proporcjonalnie do liczby przeprowadzonych testów.

Wartości średniej d^2 liczono zgodnie z Coltman [8] za pomocą prostego programu przygotowanego specjalnie dla tego celu. Średnia d^2 oznacza kwadrat różnicy liczby jednostek powtarzalnych pomiędzy allelami lokus mikrosatelitarnego uśredniony dla wszystkich analizowanych *loci*. Średnia d^2 ma dwie składowe: liczbę *loci*,

w których osobnik jest homozygotyczny (brak wpływu na wartość d^2), co powinno być skorelowane z niedawnymi przypadkami endogamii oraz rozmiary dywergencji pomiędzy allelami w *loci* heterozygotycznych (wpływ na wartość d^2), co powinno być skorelowane ze stopniem egzogamii. Zakładając model ewolucji sekwencji mikrosatelitarnych polegający na przybywaniu lub utracie pojedynczych jednostek powtarzalnych (ang. stepwise mutation model), uważa się, że allele wykazujące znaczące różnice w długości są znacznie bardziej odległe ewolucyjnie niż te, które mają bardziej zbliżoną długość. Mutacje powrotne zachodzące w istniejących pierwotnie wariantach długości, jak również mutacje powstające w sposób niezgodny z powyższym modelem, będą zaburzać pomiar. Należy jednak przyjąć, że im wyższa jest liczba *loci* poddawana badaniom, tym parametr d^2 będzie lepiej odzwierciedlał przebieg dywergencji przodków. W przypadku populacji osobników krzyżujących się przypadkowo nie jest spodziewana jakakolwiek zmienność na poziomie ancestralnej dywergencji pomiędzy osobnikami.

Dodatkowy wgląd w strukturę analizowanych prób populacyjnych uzyskano dzięki analizie nierównowagi sprzężeń (ang. linkage disequilibrium). Korelacje alleli pomiędzy parami genotypów *loci* dziedziczonych niezależnie, a zlokalizowanych na różnych chromosomach, analizowano za pomocą testu zgodności pomiędzy obserwowanymi i oczekiwanymi częstościami genotypów oraz testu *LD*. Obydwa testy prowadzono z zastosowaniem programu GENEPOP [19]. Za pomocą programu Structure [18], który pozwala na zastosowanie metody analizy skupień opartej na modelu, przeprowadzono również przyporządkowanie osób do poszczególnych podgrup. Wybrano model skorelowanych częstości, który okazał się bardziej skuteczny w detekcji niewielkiej substruktury populacji niż model niezależnych częstości.

4. Wyniki i dyskusja

4.1. Ocena współczynników fiksacji

Rozkład częstości alleli dziesięciu autosomalnych mikrosatelitarnych *loci* analizowano dla osób należących do siedmiu prób populacyjnych zamieszkujących małe wioski w regionie Polski południowej. Wartości współczynników fiksacji obliczonych na podstawie częstości alleli w analizowanych próbach populacyjnych przedstawiono w tabeli I.

Metoda hierarchiczna z zastosowaniem współczynników statystycznych F [19] może okazać się pomocna w celu ustalenia jednej z dwóch możliwych przyczyn nadmiaru homozygotyczności: ukrytej substruktury populacji lub bliskiego pokrewieństwa pomiędzy badanymi osobnikami. W sytuacji, gdy populacja jest podzielona na częściowo izolowane subpopulacje, ma miejsce podwyższone prawdopodobieństwo, że osobnicy z tej samej subpopulacji będą mieli wspólnych przodków, co ujawnia się w postaci podwyższonego prawdopodobieństwa homozygotyczności. Parametr ten jest jednak trudny do określenia z dwóch powodów: ze względu na sposób selekcji próbek badawczych, który jest niezwykle istotny dla uzyskania prawidłowego wyniku badania oraz samej metody statystycznej umożliwiającej przeprowadzenie takich kalkulacji. Średnia wartość współczynnika F_{ST} w badanej populacji wyniosła 0,0039 (95% CI = 0,0023–0,0056). Wskazuje to na bardzo nieznaczne zróżnicowanie badanej populacji. Prawie taki sam wynik (średnia wartość F_{ST} = 0,004) uzyskali Zarrabeitia i in. [27], którzy analizowali mikrogeograficzną strukturę populacji dwóch względnie izolowanych dolin z regionu Kantabrii w północnej Hiszpanii. Nieznacznie niższą wartość (0,003) uzyskano dla mniejszości białoruskiej zamieszkującej północno-wschodnią Polskę [16].

Wartości F_{ST} różnią się pomiędzy poszczególnymi *loci*. Najwyższą wartość w przypadku badanej populacji uzyskano dla lokus TH01 (0,0098). Nie można wykluczyć, że różne *loci* mają odmienną historię ewolucyjną, co prowadzi do różnic w wartościach współczynnika wsobności. Prawdopodobnie tempo mutacji charakterystyczne dla poszczególnych *loci* jest jednym z czynników, które mogą tłumaczyć zaobserwowane rozbieżności. Poprzednie badania [21] ograniczone do zaledwie trzech grup osób zamieszkujących okolice Krakowa, Rzeszowa i Poronina i trzech *loci* typu STR (TH01, TPOX i CSF1PO) wykazały, że średnia wartość F_{ST} wynosi 0,0045 i jest bardzo zbliżona do uzyskanej w niniejszych, szerszych badaniach. Wartości te są jednak nieco wyższe niż te uzyskane poprzez porównawczą analizę dużych danych populacyjnych dla regionu Polski południowej [24] oraz dla populacji Dolnego Śląska, Białegostoku i Zielonej Góry (dane nieprezentowane). Warto zauważyć, że współczynnik fiksacji F_{ST} nie był wyprowadzony przy założeniu przypadkowego krzyżowania się osobników w obrębie subpopulacji, a zatem na jego wartość nie powinna mieć wpływu zjawisko lokalnej endogamii. Na wartość F_{ST} mają wpływ wyłącznie częstości alleli charakterystyczne dla lokalnej populacji. Korelacja pomiędzy allelami w przypadku poszczególnych osobników i dla całej populacji, wyrażona poprzez wartość F_{IT}, przyjmuje wyższe wartości niż F_{ST} dla całej analizowanej populacji (tabela I). Zatem zjawiska korelacji pomiędzy parami alleli obserwowane u poszczególnych osobników i pomiędzy nimi nie są równoważne. Uzyskane wartości F_{ST} oraz F_{IT} wskazują, że w obrębie analizowanych grup nie można mówić o przypadkowym krzyżowaniu. Wartości współczynników F_{IT} i F_{IS} są wzajemnie skorelowane poprzez analizowane *loci*. W obrębie subpopulacji poziom endogamii mierzony poprzez średni współczynnik endogamii F_{IS}, jest niski, lecz wyższy od zera. Wobec tego podwyższone prawdopodbień-

stwo dopasowania pomiędzy parą alleli jednej osoby w porównaniu do par alleli stwierdzonych w subpopulacji może wpłynąć na sądowe znaczenie uzyskanego dopasowania pomiędzy dwoma profilami DNA. Średnia wartość współczynnika endogamii w badanych subpopulacjach, F_{IS} 0 zależy od liczby homozygotycznych *loci* obserwowanych w profilu DNA. Im bardziej homozygotyczny jest profil DNA podejrzanego, tym wyższe jest prawdopodobieństwo, że w tej samej subpopulacji będą istniały inne zgodne profile DNA. Z drugiej strony osoby heterozygotyczne w wielu *loci* będą się pojawiać w takiej subpopulacji z niższym prawdopodobieństwem.

4.2. Wyniki analizy średniej d^2

Średnia d^2 to parametr specyficzny dla mikrosatelitarnego DNA, który jest bardziej efektywnym narzędziem służącym do szacowania stopnia endo- i egzogamii niż heterozygotyczność charakterystyczna dla danej osoby. W tabeli II pokazano wartości średniej d^2 obliczone dla każdej z analizowanych grup. W związku z tym, że w wielu analizowanych *loci* obserwuje się allele o nieregularnej liczbie jednostek powtarzalnych (na przykład *loci* D21S11, D19S433, FGA), to różnica pomiędzy allelami odpowiada raczej różnicy pomiędzy rodzajami alleli referencyjnych (uwzględnionymi w drabinach allelicznych) niż liczbie różnic w regularnych jednostkach powtarzalnych. Na przykład, jeśli osoba ma w lokus D21S11 allele 29 i 32.2, to różnica im przypisywana wynosi 4.

Liczba alleli obserwowana w poszczególnych *loci* w analizowanej próbie populacyjnej stanowi istotny parametr, który opisuje zarówno poziom polimorfizmu, jak również same markery. Uzyskane wyniki wskazują, że istnieje pozytywna korelacja pomiędzy liczbami obserwowanych alleli w poszczególnych *loci* a wartościami średniej d^2. Różnice wartości średniej d^2 w analizowanych próbach nie są jednak istotne statystycznie i wydaje się, że mogą zostać wyjaśnione poprzez efekty stochastyczne.

4.3. Równowaga Hardy'ego-Weinberga (*HWE*)

Odstępstwo od reguły Hardy'ego-Weinberga może wynikać zarówno z substruktury populacji, jak i endogamii. W związku z tym niezgodność pomiędzy obserwowanymi a oczekiwanymi częstościami genotypów została oszacowana za pomocą testu dokładnego prawdopodobieństwa [19], który jest bardziej efektywny niż inne testy w przypadku, gdy odstępstwo od *HWE* związane jest z występowaniem specyficznych genotypów. Wyniki analizy *HWE* uzyskane dla poszczególnych subpopulacji przedstawiono w tabeli III.

Ujawnione odstępstwa od *HWE* wskazują, że w badanej populacji może istnieć rzeczywista, ale słabo zarysowana substruktura populacji. Inną prawdopodobną przyczyną odstępstwa od *HWE* jest zjawisko endogamii. Rzeczywiście z wywiadu, jaki przeprowadzono ze starszymi osobami zamieszkującymi obszary, na których zbierano próbki, wynika, że dochodziło do epizodów małżeństw pomiędzy pierwszymi kuzynami w małych wioskach zamieszkiwanych przez niewielką liczbę rodzin. Prawdopodobieństwo, że potomstwo pierwszych kuzynów odziedziczy dwa identyczne z pochodzenia (IBD) geny w danym lokus, wynosi 1/16. Może to zaburzyć równowagę Hardy'ego-Weinberga. Wydaje się zatem, że istnienie małżeństw wśród krewnych w przeszłości i utrzymanie się tego zjawiska współcześnie (notowanych jest dużo spraw kazirodczych, w których prowadzone są badania genetyczne), jest istotnym aspektem w kontekście interpretacji wyników badań DNA prowadzonych dla celów sądowych [5]. Inną przyczyną odstępstwa od *HWE* może być obecność wielu rzadkich alleli o niskich częstościach występowania w badanych próbach populacyjnych. Zapata [26] wykazał, że rzadkie allele o częstościach niższych niż 3% (zazwyczaj allele bardzo krótkie lub bardzo długie) mogą silniej wpływać na nierównowagę niż inne. Takie allele zdają się występować znacznie częściej w nierównowadze niż pozostałe warianty genetyczne. Można to wyjaśnić jako konsekwencję szybkiego tempa mutacji *loci* typu STR. Wiele przypadków nierównowagi wykrytej pomiędzy rzadkimi allelami może być efektem relatywnie niedawnych przypadków mutacji. Allele o niskiej częstości z dużym prawdopodobieństwem pojawiły się w populacji niedawno (poprzez zjawisko mutacji) w porównaniu do alleli o średniej częstości, gdyż do ich rozprzestrzenienia się w populacji konieczny jest czas.

Wspólny efekt izolacji i małych rozmiarów populacji może wreszcie prowadzić do pewnego stopnia endogamii, a w konsekwencji statystycznie istotnych odchyleń od *HWE* obserwowanych w *loci* typu STR. Zjawiska te stanowią prawdopodobne wyjaśnienie uzyskanych wyników badań. Warto zauważyć, że badane próby populacyjne miały wystarczającą wielkość, co pozytywnie wpłynęło na wartości stosowanych testów i dostarczyło możliwości do ujawnienia odchyleń od reguły *HWE*. Analiza odchyleń od reguły *HWE* przeprowadzona dla każdego lokus ujawniła znaczące odchylenia zaledwie w przypadku trzech *loci* (D2S138, D18S51 i TH01) w próbie populacyjnej z Piwnicznej. Podsumowując, należy stwierdzić, że dane genotypowe uzyskane dla 10 analizowanych *loci* zasadniczo pozostają w zgodzie z regułą *HWE*, jeśli analizujemy populację w całości. Dowodzi to, że substruktura badanej populacji generalnej jest niemożliwa do oznaczenia.

4.4. Korelacja alleli pomiędzy parami niezależnych loci

Analiza nierównowagi sprzężeń (*LD*) pomiędzy parami analizowanych *loci* była kolejną zastosowaną metodą, która pozwoliła na uzyskanie wglądu w genetyczną strukturę badanej populacji. Analiza nierównowagi sprzężeń może dostarczać użytecznej informacji na temat struktury populacji. Wynika to z faktu, że znaczące różnice w częstościach alleli pomiędzy subpopulacjami mogą być zauważalne, w populacji traktowanej jako całość, jako *LD* pomiędzy niezależnymi *loci* [23]. Po wprowadzeniu poprawki Bonferroniego wynikającej z wykonywania wielu jednoczesnych testów, 3 wartości *p* w próbach z Zakopanego (pomiędzy *loci* D3S1358/D2S338, D3S1358/D19S433 i D21S11/TH01) pozostały istotne statystycznie (na poziomie niższym niż 0,005). Biorąc pod uwagę 45 możliwych par *loci*, wynik ten daje 6,67% wszystkich porównań i jest wyższy niż oczekiwany przy nałożonym 5% progu istotności. Stwierdzona nierównowaga sprzężeń może zostać wytłumaczona poprzez zjawisko efektu założyciela, selekcję, dryft genetyczny lub nieprzypadkowe krzyżowanie się.

4.5. Test dokładny zróżnicowania populacji

Zastosowanie metody Monte Carlo opartej na łańcuchach Markowa wykorzystywanej w programie TFPGA, pozwoliło na wykazanie istotnych różnic w częstościach alleli pomiędzy badanymi grupami populacyjnymi. Uzyskane wyniki przedstawiono w tabeli IV.

Aż 7 spośród 21 analiz porównawczych wykazało znaczące różnice w częstościach alleli pomiędzy siedmioma analizowanymi subpopulacjami. Uzyskane odstępstwa od homogenności analizowanej populacji mogą zostać wytłumaczone poprzez wspólny efekt takich czynników, jak izolacja i niewielkie efektywne rozmiary analizowanych populacji oraz poprzez istnienie wielu rzadkich alleli o niskich częstościach w poszczególnych subpopulacjach.

4.6. Wnioskowanie na temat substruktury populacji poprzez przypisywanie osobników do poszczególnych grup

Analizę populacji Polski południowej rozszerzono o metodę zaproponowaną przez Pricharda i innych [16], pozwalającą na ocenę parametrów, dzięki którym możliwe jest przyporządkowanie każdej osoby do subpopulacji źródłowej. Technika grupowania została wprowadzona w życie dzięki programowi komputerowemu Structure i polega na zastosowaniu metody bayesowskiej, która umożliwia ustalenie struktury populacji na podstawie danych genotypowych pochodzących z analizy niesprzężonych markerów genetycznych. Metoda ta jest szczególnie użyteczna do identyfikacji subpopulacji, gdy w grę wchodzi problem populacji o ukrytej strukturze genetycznej. Zastosowano krótką analizę 1 000 000 permutacji, przyjmując model wykorzystujący informację *a priori*, który umożliwia połączenie danych genetycznych z danymi na temat geograficznej lokalizacji osób, od których pobrano próbki badawcze. W efekcie uzyskano modele zakładające różną liczbę grup, ale najwyższe prawdopodobieństwo uzyskano dla modelu złożonego z czterech grup. Stopień różnic w częstościach alleli występujących pomiędzy analizowanymi grupami wpływa na dokładność, z jaką grupa osób zostaje przypisana do odpowiedniej subpopulacji. Uzyskane dane wskazują, że zróżnicowanie genetyczne pomiędzy badanymi grupami nie jest mocno zaznaczone.

4.7. Sądowe implikacje uzyskanych wyników badań

Siła dyskryminacji analizowanych markerów typu STR obliczona na podstawie częstości alleli w całej badanej populacji Polski południowej wynosi 5,44 10^{-13}. Oznacza to, że prawdopodobieństwo, iż dwie przypadkowo wybrane z tej populacji osoby będą miały ten sam profil DNA w zakresie 10 analizowanych *loci*, jest równe 1 na blisko dwa biliony.

Znaczenie uzyskanych wyników dla interpretacji wyniku z badania DNA prowadzonej dla celów sądowych można zilustrować poprzez obliczenie częstości (w całej populacji) dwóch hipotetycznych profili DNA, które są złożone z najrzadziej i najczęściej obserwowanych w populacji alleli (tabela V). W przypadku rutynowych badań prowadzonych dla celów sądowych, gdy stwierdzana jest zgodność dwóch profili DNA (ujawnionego w próbce pochodzącej z miejsca zdarzenia oraz od podejrzanego), obliczana jest częstość zgodnego profilu DNA w populacji [5]. W najprostszych sprawach uzyskana wartość jest liczbowo równoważna z wartością prawdopodobieństwa zgodności tych profili [20]. Prawdopodobieństwo zgodności pozwala na ocenę, jak bardzo prawdopodobne jest ujawnienie zgodnego profilu DNA w zakresie *loci* STR u innego podejrzanego, biorąc pod uwagę fakt, że profil DNA oskarżonego jest zgodny z profilem DNA oznaczonym w próbce pochodzącej z miejsca zdarzenia kryminalnego. Zakładając, że struktura populacji rzeczywiście istnieje i biorąc pod uwagę obliczoną wartość F_{ST} = 0,0039, prawdopodobieństwo zgodności profilu dowodowego z profilem DNA należącym do innego podejrzanego pochodzącego z tej samej subpopulacji jest nieznacznie podwyższona. Prawdopodobieństwo to obliczono, stosując metodę Baldinga i Nicholsa [3]. W związku z tym, że stwierdzone korelacje pomiędzy badanymi *loci* są mało istotne, prawdopodobieństwo wielolokusowego genotypu może być obliczone poprzez uzyskanie iloczynu prawdopodobieństw dla pojedynczych *loci* z odpowiednią wartością współczynnika F_{ST}.

Taki sposób obliczania prawdopodobieństwa uwzględnia efekt dryfu na częstości alleli w populacji w każdym lokus, a zatem uwzględnia nierównowagę sprzężeń, do której mogło dojść pomiędzy niesprzężonymi *loci* na skutek działania dryfu. Zastosowanie tej metody usuwa potrzebę czynienia założeń co do niezależności dziedziczenia alleli, jak również potrzebę definiowania subpopulacji, do której należy dana osoba.

W sytuacji, gdy mamy do czynienia z wystarczająco dużą próbą populacyjną, może dojść do odrzucenia hipotezy o równowadze *HWE*. Jednak ekspert sądowy może mimo to uważać, że rozmiar odstępstwa od równowagi jest do tego stopnia niewielki, że nie powinien prowadzić do odrzucenia jej zastosowania w praktyce. Założenie to może zostać obalone poprzez stwierdzoną silną nierównowagę sprzężeń. Uzyskane tutaj wartości *LD* na poziomie bliskim zera wskazują, że wpływ nierównowagi sprzężeń na prawdopodobieństwa zgodności przy badaniach prowadzonych z zastosowaniem zestawu SGM Plus może być rzeczywiście pomijany.

Jak pokazano w tabeli V, uzyskana wartość współczynnika wsobności ma bardzo ograniczony wpływ na częstość profilu DNA, który jest częsty w populacji. Wpływ współczynnika wsobności na częstość rzadkich profili DNA jest bardziej znaczący, ale nawet w takich przypadkach nie ma to większego znaczenia sądowego w związku z bardzo wysoką siłą dyskryminacji, jaką zapewnia analiza *loci* wchodzących w skład zestawu SGM Plus.

5. Wnioski

Niniejsza praca przedstawia wyniki badań prowadzonych nad substrukturą populacji z zastosowaniem różnych metod analizy statystycznej. Pozwoliło to na obliczenie różnych parametrów charakteryzujących badaną populację.

Uzyskane wyniki badań wskazują, że cała populacja w niewielkim stopniu odbiega od stanu równowagi Hardy'ego-Weinberga. Średnia wartość współczynnika wsobności $F_{ST} = 0{,}0039$ jest względnie wysoka w porównaniu z wartością uzyskaną poprzez porównanie większej próby populacyjnej z terenu Polski, co wskazuje na bardzo nieznaczne zróżnicowanie populacji Polski południowej mimo bardzo złożonej historii osadnictwa na tych terenach.

Uzyskana wartość parametru F_{ST} ma bardzo ograniczony wpływ na obliczenia częstości profilu DNA, który pojawia się w populacji z wysoką częstością, ale jest bardziej znacząca w przypadku rzadkich profili DNA. Tym niemniej nawet w przypadku rzadkiego profilu DNA nie ma to znaczenia dla badań sądowych ze względu na wysoką siłę dyskryminacji panelu *loci* SGM Plus. Z drugiej jednak strony zastosowanie współczynnika wsobności charakterystycznego dla danej populacji do obliczenia prawdopodobieństwa zgodności byłoby korzystne dla oskarżonego.

Inną przyczyną słabo zaznaczonego zróżnicowania analizowanej populacji może być niska czułość panelu *loci* zawartego w zestawie SGM Plus do detekcji potencjalnej substruktury populacji. Wydaje się zatem, że konieczne jest przeprowadzenie dalszych, precyzyjniejszych badań populacyjnych z zastosowaniem bardziej informatywnych markerów, jak na przykład *loci* STR chromosomu Y lub odpowiedni panel markerów typu SNP. W cytowanej powyżej pracy Zarrabaitia i in. wykazano, że wartości F_{ST} uzyskane dla *loci* STR chromosomu Y były 10-krotnie wyższe niż uzyskane dla *loci* STR autosomalnych. Taki wynik potwierdza wyższą użyteczność markerów ChrY-STR w badaniach nad strukturą populacji.

Podziękowania
Pragnę wyrazić wdzięczność dr Wojciechowi Branickiemu za jego rady i komentarze, które otrzymałam podczas przygotowywanie tej pracy oraz Markowi Kowalczykowi za jego uprzejmą pomoc w udoskonaleniu mapy Polski południowej.