# Genetic Markers Combination Calculation in Wood Samples Identification

Alexey S.Pyataev[1,2], Alexey A. Ibe[2], Elena A. Shilkina[2]

[1]Reshetnev Siberian State University of Science and Technology, Krasnoyarsky Rabochy Av 31, Krasnoyarsk, Russia, 660037

[2]Branch of FBI «Russian Centre of Forest Health» – «Centre of Forest Health of Krasnoyarsk Krai», Akademgorodok 50A building 2, Krasnoyarsk, Russia, 660036

**Abstract**. This paper proposes a method for determining the microsatellite markers combination used to study the genetic structure of a woody plant population using the example of Pinus sylvestris. The developed method optimizes genetic analyzes conduction in the tasks of the forest genetic resources state monitoring.

**Keywords**: genetic structure, microsatellite markers, Pinus sylvestris.

## 1 Introduction

Over the past decade due to the negative economic and environmental consequences of illegal logging increasing attention has been paid to the origin of timber products in the world [1,2]. Statistics of imports and exports conducted by WWF experts in 2008 showed that a significant amount of illegally harvested wood enters the European and Chinese markets from Russia and Eastern Europe. In the Russian Federation only in 2008–2016 period were recorded 197,228 cases of illegal logging, total damage amounted to 104.5 billion rubles, reimbursed - 2.83 billion rubles. (2.7% of the amount of damage assessed). Illegal forests use has been identified in almost all regions of the Russian Federation [3]. In order to prevent these offenses, law enforcement authority officers should be able to conclusively identify the true origin of the wood transported [4].

One of the promising areas of evidence of the timber trade legality is the use of molecular-genetic methods of analysis. These methods are based on the using of genetic markers - microsatellites - varying regions (loci) in nuclear DNA and DNA of organelles (mitochondria and plastids) consisting of tandemly repeated nucleotide sequences. These markers are characterized by a high level of polymorphism and are often found in the genome [5].

The molecular-genetic method compared with the traditional denrochronological identification avoids timely collection of data such as tree age, diameter, height, thickness. The insufficient number or lack of clear annual rings due to wood decay severely limits the use denrochronological identification of wood samples [6].

However, nowadays there are no methods that allow to obtain minimal combinations of molecular primers that give the lowest probability of a coincidence of related multi-focused genotypes. Thus, the task of developing a method for determining the optimal sequence of microsatellite markers used to study the genetic structure of a woody plant population using the example of Scots pine is an urgent one.

## 2 The microsatellite markers combination determination method

Wood samples of pine (*Pinus sylvestris* L.) taken from plantations growing near the Balakhta in the Krasnoyarsk Krai served as a reference sample. The selection cotained 29 samples, which is representative and accepted in the analysis of nuclear codominated nuclear markers. In molecular genetic studies, the average number of samples in the selection less than 30 individuals per cenopopulation [7-8]. Basic methodology for the DNA study relies on the following stages, i.e., DNA extraction (based on the lysis of cell walls), DNA amplification via polymerase chain reaction (PCR) method with specific primers for nuclear microsatellite or organelle alleles, genotyping of the PCR products in automated sequencer, and finally comparison of the DNA profiles obtained for all samples. The wood was thoroughly crushed, homogenized, and the DNA was isolated by the CTAB method [9]. The method is based on the cells breakdown under the cetyltrimethylammonium bromide (CTAB) effect, the removal of proteins using chloroform and the precipitation of DNA with isopropanol. PCR was performed with the use of a commercial kit of reagents "ScreenMix" (JSC "Eurogen", Russia). Amplification was carried out in the thermal cycle T100 Thermal Cycler (BioRad). The amplification products were separated by electrophoresis on 6% polyacrylamide gel using Tris-Borate-EDTA electrode buffer and stained with ethidium bromide. PCR products were visualized by UV using gel documentation system (Figure 1). Data analysis was performed using Vilber Lourmat Bio Capt V. 12.5.0.0 software.

As a result of preliminary work to identify the most polymorphic and stably amplifying loci, 10 microsatellite markers were selected. Table 1 presents the characteristics of recommended nuclear microsatellite loci for Scots pine [10–13].

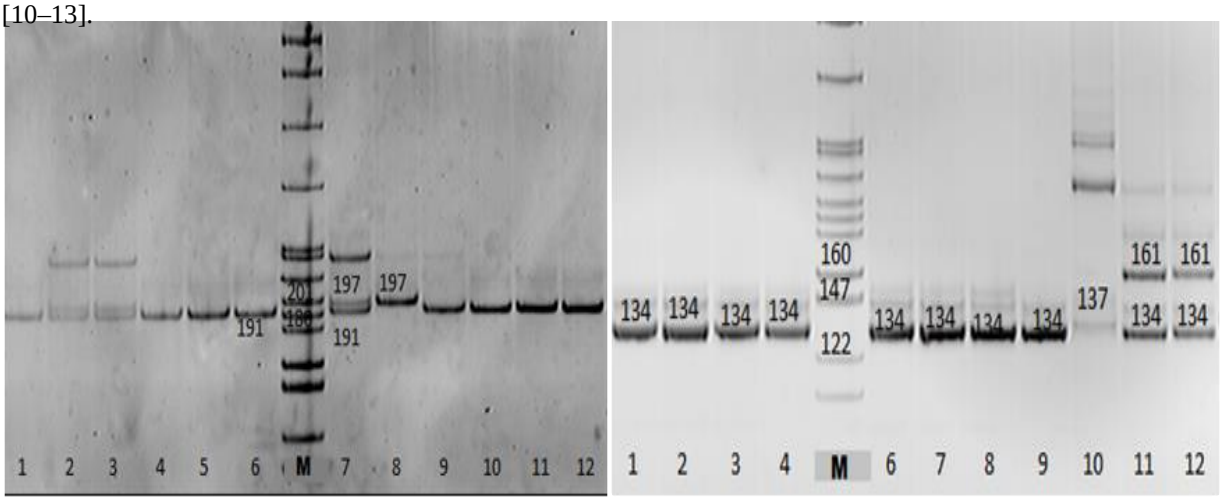


**Figure 1.** The electrophoregrams of nuclear microsatellite loci lw_isotig04306 and PtTx3116 of *Pinus sylvestris* L. The numerals 191, 197, 134, 137, 161 on lw_isotig04306 and PtTx3116 electrophoregrams represent the alleles of amplified DNA fragments. M, molecular-weight size marker

**Table 1**. Nuclear microsatellite locus characteristics which are selected for operation with Scots pine samples.

| Num | Locus | Мотив | Amplicon size, bp | Temperature, °C | Alleles count | Source |
|---|---|---|---|---|---|---|
| 1 | psy119 | (GCT)7 | 315-324 | 55 | 3 | [10] |
| 2 | psy157 | (ACC)7 | 187-202 | 55 | 6 | [10] |
| 3 | PtTx3116 | (TTG)7(TTG)5 | 122-226 | 55 | 8-10 | [11] |
| 4 | PtTx3107 | (CAT)14 | 150-182 | 55-45↓1 | 5-6 | [11] |
| 5 | PtTx4001 | (CA)15 | 201-224 | 60-50↓1 | 4-7 | [11] |
| 6 | lw_isotig21953 | (ATGGG)7 | 208 | 60 | 7 | [12] |
| 7 | PtTx4011 | (CA)20 | 230-284 | 60-50↓1 | 21 | [11] |
| 8 | SPAC11.4 | (AT)5(GT)19 | 130-170 | 65-55↓1 | 38 | [13] |
| 9 | lw_isotig04306 | (TCC)7 | 196 | 55 | 3 | [12] |
| 10 | lw_isotig27940 | (TGGA)5 | 231 | 55 | 3 | [12] |

Currently, the processing of the results of genetic analysis is carried out using proprietary software GenAlEx [14], a free add-in for MS Excel. The results obtained using the GenAlEx program are shown in Table 2. The smallest chance of a coincidence of multilocus genotypes (2.2E-08) is achieved only with 10 genetic markers combination: psy119, psy157, PtTx3116, PtTx3107, PtTx4001, lw_isotig21953, PtTx4011, SPAC11.4, lw_isotig04306, lw_isotig27940. This value indicates a low probability of random genotypes coincidence [15]. The order of genetic markers is presented in table 1 and in figure 2.

**Table 2.** The species identity probability with the same multilocus genotype at increased the genetic markers combination.

| Locus combinations | Identity probability |
|---|---|
| 1 | 1 |

| | |
|---|---|
| 1+2 | 0,76 |
| 1+2+3 | 0,059 |
| 1+2+3+4 | $6{,}3 \cdot 10^{-3}$ |
| 1+2+3+4+5 | $1{,}3 \cdot 10^{-3}$ |
| 1+2+3+4+5+6 | $6{,}6 \cdot 10^{-5}$ |
| 1+2+3+4+5+6+7 | $2{,}3 \cdot 10^{-5}$ |
| 1+2+3+4+5+6+7+8 | $1{,}5 \cdot 10^{-6}$ |
| 1+2+3+4+5+6+7+8+9 | $2{,}3 \cdot 10^{-7}$ |
| 1+2+3+4+5+6+7+8+9+10 | $2{,}2 \cdot 10^{-8}$ |

Genetic analysis conduction of the identity of wood samples determination, using all 10 markers, is very laborious and expensive. To reduce these costs is proposed a method for determining the optimal minimum sequence of markers to eliminate false positive sample identification. The purpose of the method proposed is to find a number and sequence of markers that will give the minimum probability of false-positive sample identification.

The method proposed in this paper is based on the analysis of the alleles pairs occurrence of each locus among the samples of a deliberately unique selection. As a reference selection used samples of Scots pine, taken from natural plantation, growing near the Balakhta in the Krasnoyarsk Krai. The reference selection contains 29 numbered unique samples and initially tested with 10 markers.

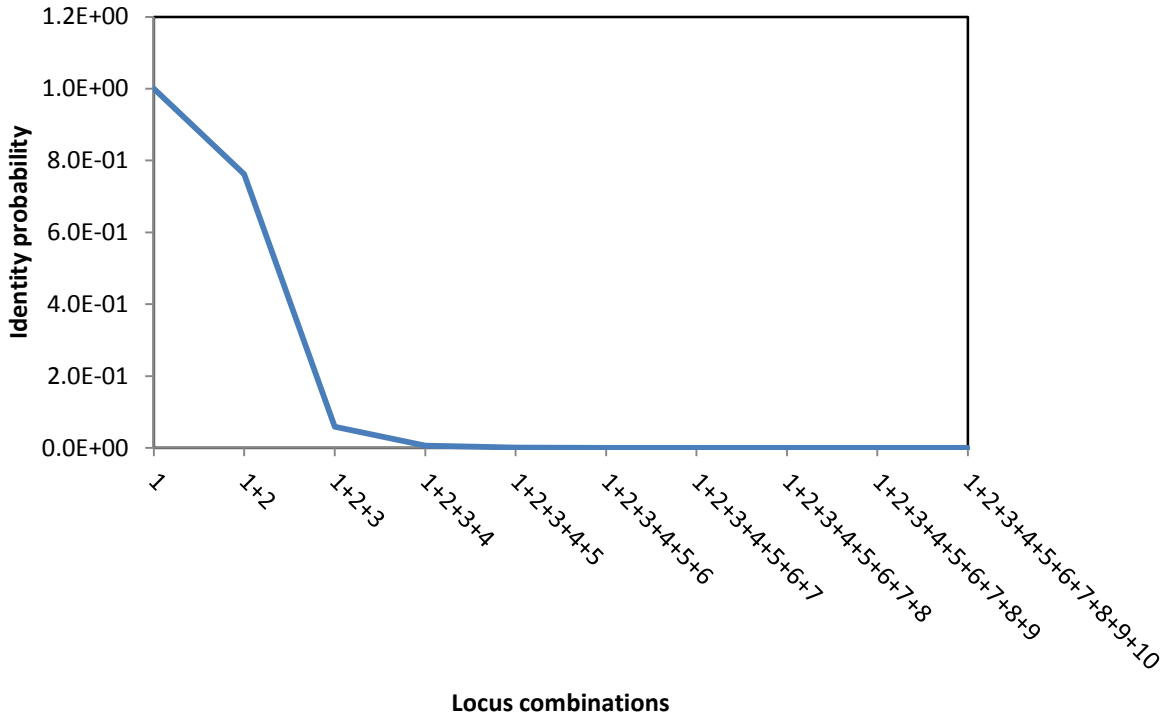


**Figure 2**. The species identity probability with the same multilocus genotype at increase in the combination of genetic markers.

Let us denote

$$M = \{M_i, i = 1..10\},$$

as a set of action markers results. Each marker for each sample gives us a pair of values:

$$M_i = \{m_j = (a_j, b_j) : j = 1..29, a_j \in N, b_j \in N\}.$$

Table 3 shows the molecular weight values of the selection alleles pairs, measured in "bp" (base pair in English-language literature), or in n.p. – nucleotide pairs in the Russian-language literature.

**Table 3.** Allele pairs inside locus by samples.

| № обр. | lw_isotig 21953 | lw_isotig04306 | lw_isotig 27940 | PtTx3107 | … | SPAC 11.4 |
|---|---|---|---|---|---|---|

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 258/258 | 187/187 | 235/255 | 159/165 | … | 146/156 |
| 2 | 223/263 | 187/193 | 247/247 | 153/159 | … | 138/142 |
| 3 | 203/223 | 187/187 | 247/247 | 159/165 | … | 152/152 |
| 4 | 248/258 | 184/187 | 247/247 | 180/180 | … | 142/152 |
| 5 | 258/258 | 187/193 | 247/247 | 159/165 | … | 146/146 |
| 6 | 203/203 | 175/187 | 235/235 | 162/165 | … | 142/146 |
| … | … | … | … | … | … | … |
| 24 | 248/263 | 184/187 | 235/235 | 153/153 | … | 138/146 |
| 25 | 248/248 | 178/187 | 255/255 | 159/168 | … | 146/160 |
| 26 | 223/243 | 193/193 | 239/247 | 159/159 | … | 138/160 |
| 27 | 248/253 | 187/193 | 239/255 | 165/165 | … | 138/142 |
| 28 | 248/253 | 187/187 | 239/255 | 165/165 | … | 138/146 |
| 29 | 203/203 | 193/193 | 255/263 | 162/165 | … | 138/138 |

Selection sample IDs are grouped by the value pairs shown in Table 3 for each marker:

$$G_i = \{g_k(i) = \{m_n\}: k \in N, k < 30, n \in N, n < 30\}.$$

Sample grouping results by allele values on the example of the lw_isotig04306 marker is shown in Table 4.

**Table 4.** Grouped samples by the values of the lw_isotig04306 marker alleles.

| lw__isotig04306 | ids |
|---|---|
| 175/187 | {6,19} |
| 178/187 | {18,25} |
| 181/187 | {16,17,22} |
| 184/184 | {7,10,13} |
| 184/187 | {4,11,12,20,24} |
| 184/193 | {15} |
| 187/187 | {1,3,8,21,23,28} |
| 187/193 | {2,5,9,14,27} |
| 193/193 | {26,29} |

The next stage of the method is to rank the sets of identifiers $G_i$ grouped in pairs from the sets $M_i$ by the number of unique pairs, i.e. in priority $G_i$, in which the maximum number $|g_k(i)| = 1$.

Then there is an iterative process of intersection of sets $G_i$. Each set of grouped genotypes intersects with other sets of grouped genotypes. The first step of the iteration is a pairwise intersection of the sets $G_i$. The results of the intersection of only those sets, which are subsets of a capacity at least two:

$$R = \{R_l = g_m(i) \cap g_n(i): |R_l| > 1; l, m, n, i, j \in N; m, n < 30; i, j \leq 10\}.$$

Then $R_l$ are ranked in cardinality ascending. At the next iterations, $R$ intersects with the remaining $G_i$:

$$R = R \cap G_i.$$

The process continues until the intersection becomes an empty set. Thus, by analyzing a test selectionin a similar way, it is possible to obtain an optimal sequence of markers that uniquely identifies samples of the test species.

For Scots pine samples, taken in natural plantation growing growing near the Balakhta in the Krasnoyarsk Krai, the optimal minimal combination of genetic markers was the sequence lw_isotig21953, SPAC11.4, PtTx3107.

To verify the method, a control group of Scots pine wood from the same plantation was selected and analyzedTable 5 presents the results of grouping samples of the control group according to the values of the lw_isotig04306 marker alleles.

**Table 5.** Control group grouped samples by the values of the lw_isotig04306 marker alleles.

| lw_isotig04306 | ids |
|---|---|

| | |
|---|---|
| 175/187 | {15,16} |
| 178/187 | {24,23} |
| 181/187 | {14,13,22} |
| 184/184 | {9,19,21} |
| 184/187 | {29,10,11,12,20} |
| 184/193 | {28} |
| 187/187 | {26,27,5,6,8,18} |
| 187/193 | {25,3,4,7,17} |
| 193/193 | {1,2} |

Sample analysis of the control group showed the effectiveness of the selected markers sequence. In the case of false positive sample identification, the control group is additionally analyzed with the lw_isotig27940 marker.

The analysis results of the control group of samples were checked by the GenAlEx program with the indicated sequence of markers. The species identity probability with the same multilocus genotype at increase in the combination of genetic markers is given in Table 6 and in Figure 3.

**Table 6.** The species identity probability with the same multilocus genotype at increase in the combination of genetic markers.

| Locus combinations | Identity probability |
|---|---|
| 6 | $5{,}2 \cdot 10^{-2}$ |
| 6+8 | $3{,}4 \cdot 10^{-3}$ |
| 6+8+4 | $3{,}6 \cdot 10^{-4}$ |
| 6+8+4+10 | $3{,}5 \cdot 10^{-5}$ |

The order of genetic markers is presented in table 1.

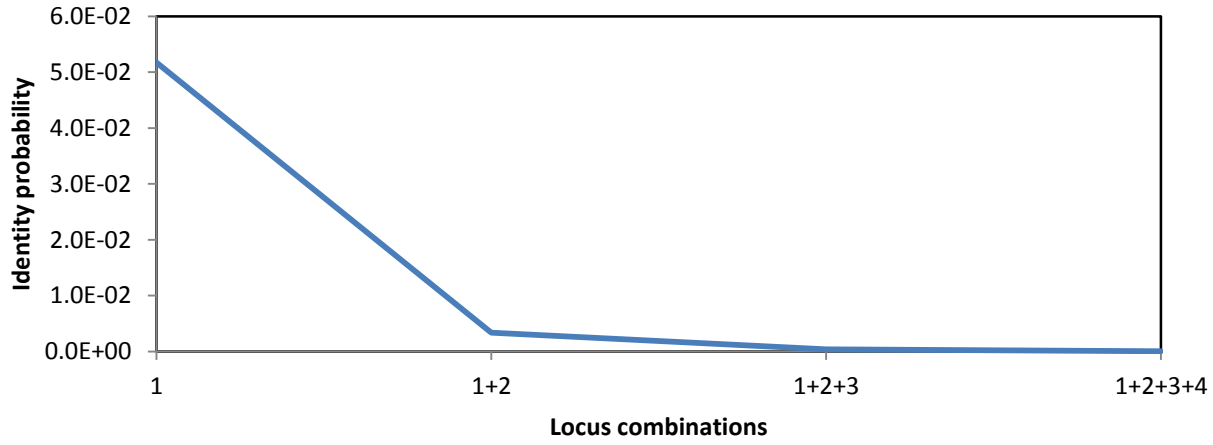


**Figure 3**. The species identity probability with the same multilocus genotype at increase in the combination of genetic markers

Comparing the results of tables 2 and 6 shows that using the sorted marker sequence decrease the probability of a random coincidence of multilocus genotypes. To achieve an acceptable result, the use of only three markers was sufficient.

## 3 Conclusion

To determine the identity of Scots pine samples , taken in natural plantation growing near the Balakhta of the Krasnoyarsk Krai, the sequence of genetic markers {lw_isotig21953, SPAC11.4, PtTx3107} was the minimum optimal combination. The effectiveness of the selected sequence of markers tested on the control group. The proposed

method of the optimal sequence microsatellite markers selection can significantly reduce labor, time and material costs in the wood samples identity determination. In further studies, it is planned to use this algorithm for the marker selection in relation to other sets to similar genetic analysis.

## References


[1] Céline Blanc-Jolivet, Yulai Yanbaev, Birgit Kersten, Bernd Degen. A set of SNP markers for timber tracking of Larix spp. in Europe and Russia // Forestry. 2018. P. 1–15.

[2] WWF World Wide Fund For Nature 2008 Illegal wood for the European market. Frankfurt a M: WWF Germany, P. 43.

[3] Kuzmichev E., Trushina I., Lopatin E. Volumes of Illegal Forest Loggingin Russian Federation // Forestry information. 2018. № 1. С. 63–77.

[4] Latov J.V., Zhavoronkov J.M. Achievements and perspectives of dendro expertise in fighting against illegal felling of forest ranges // Proceedings of Management Academy of the Ministry of the Interior of Russia. 2013. № 4 (28). С. 44-48.

[5] Ilyinov A. A., Raevsky B. V. The current state of Pinus sylvestris L. gene pool in Karelia // Sibirskij Lesnoj Zurnal (Siberian Journal of Forest Science). 2016. N. 5: 45–54

[6] Nowakowska J. A., Oszako T., Tereba A., Konecka A. Forest Tree Species Traced with a DNA-Based Proof for Logging Case in Poland // Evolutionary Biology: Biodiversification from Genotype to Phenotype, DOI 10.1007/978-3-319-19932-0_19.

[7] Nei M. Estimation of average heterozygosity and genetic distance from a small number of individuals // Genetics. 1978 . V. 89. P. 583–590.

[8] Shurkhal A. V., Podogas A. V., Zhivotovsky L. A. Allozyme differentiation in the genus Pinus // Silvae Genetica. 1992. V. 41. P. 105–109.

[9] Devey M. E., Bell J. C., Smith D. N., Neal D. B., Moran G. F. A genetic linkage map for Pinus radiata based on RFLP, RAPD, and microsatellite markers // Theor. Appl. Genet. 1996. V. 92. Iss. 6. P. 673–679.

[10] Sebastiani F., Pinzauti F., Kujala S. T., Gonzalez-Martinez S. C., Vendramin G. G. Novel polymorphic nuclear microsatellite markers for Pinus sylvestris L. // Conservation Genet. Resour. 2012. V.4. Iss. 2. P. 231–234.

[11] Belletti P., Ferrazzini D., Piotti A., Monteleone I., Ducci F. Genetic variation and divergence in Scots pine (Pinus sylvestris L.) within its natural range in Italy // European Journal of Forest Research. 2012. V. 131. Iss. 4. P. 1127–1138.

[12] Fang P., Niu Sh., Yuan H., Li Zh., Zhang Yu., Yuan L., Li W. Development and characterization of 25 EST-SSR markers in Pinus sylvestris var. mongolica (Pinaceae) //Applications in Plant Sciences. 2014. V. 2. Iss. 1. P. 1–4.

[13] Soranzo N., Provan J., Powell W. Characterization of microsatellite loci in Pinus sylvestris L. // Molecular Ecol. 1998. V. 7. P. 1247–1263.

[14] Peakall R., Smouse P. E. GenAlEx v. 6.5: Genetic analysis in Excel. Population genetic software for teaching and research – an update // Bioinformatics. 2012. V. 28. Iss. 19. P. 2537–2539.

[15] Brown S. M. Methods of genome analysis in plants // Ed. P.P. Jauhar. N.-Y., London. Tokyo. 1996. P. 147–159.