**PLANT GENETICS**

# Development of Microsatellite Genetic Markers in Siberian larch (*Larix sibirica* Ledeb.) Based on the *De Novo* Whole Genome Sequencing

**N. V. Oreshkova[1, 2], Yu. A. Putintseva[1], V. V. Sharov[1], D. A. Kuzmin[1], K. V. Krutovsky[1, 3, 4, 5 *]**

*[1]Genome Research and Education Center, Siberian Federal University, 660036 Krasnoyarsk, Russia*

*e-mail: yaputintseva@mail.ru, sharvadim07@yandex.ru, dm.kuzmin@gmail.com, kkrutovsky@gmail.com*

*[2]V. N. Sukachev Institute of Forest, Siberian Branch of Russian Academy of Sciences, 660036 Krasnoyarsk, Russia*

*e-mail: oreshkova@ksc.krasn.ru*

*[3]Georg-August University of Göttingen, 37077 Göttingen, Germany*

*e-mail: konstantin.krutovsky@forst.uni-goettingen.de*

*[4]N. I. Vavilov Institute of General Genetics, Russian Academy of Sciences, 119333 Moscow, Russia*

*e-mail: kkrutovsky@gmail.com*

*[5]Texas A&M University, College Station, TX 77843-2138, USA*

*e-mail: k-krutovsky@tamu.edu*

**Abstract**–This special issue of the journal is devoted to the outstanding population geneticist Yuri Petrovich Altukhov, who paid much attention in his research to the development of molecular genetic markers for population studies. Over the past time markers and methods of their development have undergone significant change. Thanks to modern methods of whole genome sequencing, it has become possible to develop markers of very different types - selectively neutral, as well as functional. Among them, microsatellite markers remain the most informative, convenient, reproducible, relatively inexpensive, and polymorphic. Whole genome sequencing greatly facilitates their

discovery and development. This paper is devoted to the development of new microsatellite markers for a very important species of boreal forest - Siberian larch (*Larix sibirica* Ledeb.). Using a draft assembly of the larch genome, several thousand contigs containing microsatellite loci with di-, tri, tetra- and pentanucleotide motifs were selected. A total of 59 pairs of PCR primers were tested for loci with dinucleotide motifs as the most variable. From them, 11 pairs were finally selected for 11 loci with dinucleotide repeats, which showed a high level of polymorphism and can be used in various population genetic studies and to identify the origin of wood and plant material.
This study was done at the Laboratory of Forest Genomics of the Genome Research and Education Center of the Siberian Federal University with the support of the Department of Forest Genetics and Forest Tree Breeding of the Georg-August University of Göttingen, the Department for Monitoring of Forest Genetic Resources of the Forest Protection Center of the Krasnoyarsk Territory, and the Laboratory of Forest Genetics and Selection of the V. N. Sukachev Institute of Forest of the Siberian Branch of the Russian Academy of Sciences within the framework of the project "Genomics of the key boreal forest conifer species and their major phytopathogens in the Russian Federation" funded by the Government of the Russian Federation (grant no. 14.Y26.31.0004).

*Keywords*: genetic diversity, genome, heterozygosity, Siberian larch, *Larix sibirica*, microsatellite markers, NGS, whole genome sequencing

## INTRODUCTION

The development of molecular genetic markers for the main forest-forming tree species with the help of DNA based methods and their use in scientific research and forestry are extremely important and needed for solving problems of forestry, reforestation and afforestation. To solve these problems, estimates of the level of genetic variability, data on the population structure and differentiation, and effective methods of genetic identification of the wood and plant material origin are required. Among the available genetic markers, nuclear microsatellite loci most fully meet these challenges. These markers are characterized by specificity, reproducibility, codominance, multiple

alleles, high heterozygosity and, moreover, do not require sophisticated equipment for analysis.

However, for Siberian larch (*Larix sibirica* Ledeb.), one of the main forest-forming conifer species in Siberia, such species-specific markers have not been developed till this study. Siberian larch grows in the forest zone of the east and northeast of the European part of Russia, the Urals, Western and Eastern Siberia. Its area stretches from tundra (71°N latitude) on the north to the southern latitudes of Altai and Sayan (46° N) on the south. On the territory of the Russian Federation, larch forests occupy 263 million hectares, about 40% of the forest area of the country (769.8 million hectares).

Previously, markers based on nuclear microsatellite loci developed for other species of this genus were used to analyze the population-genetic variation of *L. sibirica* [1-3]. With the help of these markers, genetic diversity and differentiation were studied in several populations of this species [4, 5]. However, a small number of markers was used in these studies due to poor PCR amplification and the presence of a large number of "null alleles" for many non-species-specific markers.

Thanks to the project on the whole genomic sequencing of conifers in the Laboratory of Forest Genomics of the Siberian Federal University, it has become possible to develop species-specific microsatellite primers for Siberian larch [6-8]. Such markers are more reliable; they allow better population genetic analysis, identification of the timber origin of wood important in the fight against illegal logging, and efficient control for the proper origin of plant material used for reforestation and identification of clonal material.

Therefore, the main objective of the presented study was to develop new highly informative microsatellite genetic markers for *L. sibirica* using data from the whole genome sequencing of this species. To achieve this objective, a computer search for microsatellite loci with high repetitive simple motifs was done in the genomic DNA sequences, oligonucleotide primers were developed, synthesized and tested for the selected loci, a preliminary estimate of allelic diversity was made on two test samples of a Siberian larch population collected in the Republic of Khakassia (Russia), the most promising markers were selected, and multiplex genotyping panels were designed and tested for fragment analysis using the ABI 3130xl Genetic Analyzer with capillary electrophoresis.

## MATERIALS AND METHODS

Sequencing of the Siberian larch genome was done with 74X coverage using the Illumina HiSeq 2000 platform. To select high quality reads and to remove adapter dimers the raw reads were filtered using Trimmomatic [9]. A draft assembly was generated using the CLC Assembly Cell assembler (https://www.qiagenbioinformatics.com). The obtained assembly contained 4.4 million contigs with a total length of ~ 5 Gbp. This assembly was searched for contigs containing microsatellite loci using the GMATo program [10]. The preliminary analysis showed that microsatellite loci with tri-, tetra- and pentanucleotide motifs are much less variable in larch than the loci with dinucleotide motifs (authors' data). Therefore, from all microsatellite loci found, only loci with dinucleotide motifs repeated at least 20 times were selected for the PCR primer design. Primers for the selected microsatellite loci were designed using the WebSat online service [11]. As a result, 59 primers pairs were designed and tested.

Needle samples collected from 100 individual Siberian larch trees in 2014 in two populations (50 trees per population) in the Republic of Khakassia were used in this study. The one population is located in the Shirinsky District of Khakassia near the Shira-Berenjak highway (larch forest with pine on a gentle slope), another - near the Efremkino Village (larch on a steep slope and at its foot). The distance between the populations is 25 km.

DNA from the needles was isolated using the CTAB method [12]. For the PCR, the following reaction mixture was used (example for one sample): 6.8 μl of $H_2O$, 1.5 μl of 10× Solis BioDyne PCR-Buffer (https://www.sbd.ee), 1.5 μl of 25 mM $MCl_2$ from Solis BioDyne, 1 μl of 2.5 mM dNTP from Thermo Scientific (Thermo Fisher Scientific Inc., Waltham, Massachusetts, USA; https://www.thermofisher.com), 1 μl each of forward (F) and reverse (R) primers, 0.2 μl of *Taq* polymerase from Solis BioDyne (5 Units/μl), and 1-2 μl of the DNA sample (50 ng/μl). The amplification program included denaturation as the first step at 94° C for 15 min, then 10 "touchdown" cycles with a 1° C decrease at each cycle: 1 min of denaturation at 94° C, 1 min of annealing at 60° C, 1 min of extension at 72 ° C, then followed by 25 cycles with the same annealing temperature of 50° C, and concluded by the final elongation at 72° C for 20 min.

The testing scheme included, first, the PCR primer annealing tests with four individual DNA samples and staining the PCR products using the Roti®-GelStain dye (Carl Roth GmbH + Co. KG, Karsruhe, Germany; https://www.carlroth.com) following electrophoresis in agarose gel. After the optimization of the PCR amplification conditions, each locus was tested then with 4-8 samples from one population in order to reveal polymorphism. The loci that turned out to be polymorphic were then genotyped in all 100 trees from both populations using the ABI PRISM 3730 sequencer (Applied Biosystems, Thermo Fisher Scientific Inc.). The ABI GeneScan™ 500 LIZ™ dye Size Standard was used as an internal marker for fragment lengths (Applied Biosystems). The visualization of the ABI chromatograms and fragment analysis were done using the GeneMapper 4.0 program (Applied Biosystems).

## RESULTS AND DISCUSSION

Among 59 primer pairs selected for the first test 20 produced no product (for instance loci *Ls_1286838* and *Ls_557345* in Fig. 1), 12 had non-specific amplification and 27 stably amplified supposedly a single-locus PCR product that could be well-genotyped on gels (Table 1).

After the first selection, the forward primer in each of the 27 pairs was labeled either by "blue" (FAM) or "green" (HEX) fluorescent dyes for further testing on the ABI PRISM 3730 sequencer (Table 2). The labeled oligonucleotide primers were synthesized by Sigma (Germany). The trial PCR multiplexes consisting of two or three primer pairs were made taking into account the size of the PCR fragments. Multiplexing was done at the PCR reaction stage by combining two or three different primer pairs in the same PCR reaction and adjusting the total volume by reducing the water portion accordingly. The obtained PCR amplification product was necessarily diluted 50-100 times before electrophoresis.

The testing of polymorphic loci at this stage was carried out using 8-16 samples from each of the two populations. After this testing on a capillary sequencer, additional 9 pairs of primers had to be excluded due to poor or nonspecific amplification, and supposedly a large number of null alleles.

## CONCLUSIONS

Finally, as a result of careful multistage primer testing, 11 reliable microsatellite loci with a high level of polymorphism were selected (Table 3), although there could be null-alleles at three loci. The final testing was carried out using all 100 samples from two populations. It allowed us to compile multiplex panels for the simultaneous genotyping of several loci on the same samples (Table 4). Nucleotide sequences of contigs that contain microsatellite loci and have been used for primer development can be provided by the authors upon request.

Further analysis of natural and artificial Siberian larch populations using the developed markers will allow obtaining reliable quantitative estimates of the parameters of their genetic structure, such as within and between population allelic and genetic diversity, genetic subdivision and differentiation at different hierarchical levels, inbreeding, gene flow, etc.

## ACKNOWLEDGMENTS

The study was done as part of the project “Genomics of the Key Boreal Forest Conifer Species and Their Major Phytopathogens in the Russian Federation” funded by the Government of the Russian Federation (grant no. 14.Y26.31.0004).

## REFERENCES

1. Khasa, D.P., Newton, C.H., Rahman, M.H., et al. Isolation, characterization, and inheritance of microsatellite loci in alpine larch and western larch, *Genome*, 2000, vol. 3, no. 43, pp. 439-448. doi 10.1139/g99-131
2. Isoda, K. and Watanabe, A. Isolation and characterization of microsatellite loci from *Larix kaempferi*, *Mol. Ecol. Notes*, 2006, vol. 6, no. 3, pp. 664-666. doi: 10.1111/j.1471-8286.2006.01291.x
3. Chen, C., Liewlaksaneeyanawin, C., Funda, T. et al. Development and characterization of microsatellite loci in western larch (*Larix occidentalis* Nutt.), *Mol. Ecol. Resour*., 2009, vol. 9, no. 3, pp. 843-845. doi: 10.1111/j.1755-0998.2008.02289.x
4. Oreshkova, N.V. and Belokon, M.M. Assessment of the genetic variation of Siberian larch use microsatellite markers, *Vestnik MSGL - Lesnoy Vestnik*, 2012,

vol. 84, no. 1, pp. 118-122, in Russian (*Орешкова Н.В., Белоконь М.М.* Оценка генетической изменчивости лиственницы сибирской с использованием микросателлитных маркеров // Вестник МГУЛ - Лесной вестник. 2012. Т. 84. № 1. С. 118-122.)

5. Oreshkova, N.V., Belokon, M.M., and Jamiyansuren, S. Genetic Diversity, Population Structure, and Differentiation of Siberian Larch, Gmelin Larch, and Cajander Larch on SSR-Marker Data, *Russian Journal of Genetics*, 2013, vol. 49, no. 2, pp. 178–186.
6. Krutovsky, K.V., Oreshkova, N.V., Putintseva, Y., et al. Preliminary results of *de novo* whole genome sequencing of the Siberian Larch (*Larix sibirica* Ledeb.) and the Siberian Stone Pine (*Pinus sibirica* Du Tour), *Siberian Journal of Forest Science*, 2014, vol. 1, no. 4, pp. 79–83 (in Russian with abstract in English).
7. Oreshkova, N.V., Putintseva, Yu.A., Kuzmin, D.A., et al. Genome sequencing and assembly of Siberian larch (*Larix sibirica* Ledeb.) and Siberian pine (*Pinus sibirica* Du Tour) and preliminary transcriptome data, in *Proceedings of the 4th International Conference on Conservation of Forest Genetic Resources in Siberia*. Barnaul, Russia, 24–29 August, 2015, pp. 127–128.
8. Sadovsky, M.G., Putintseva, Y.A., Birukov, V.V., et al. *De novo* assembly and cluster analysis of Siberian larch transcriptome and genome, *Lecture Notes in Bioinformatics*, 2016, vol. 9656, pp. 455-464.
9. Bolger, A.M., Lohse, M., and Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data, *Bioinformatics*, 2014, vol. 30, no. 15, pp. 2114–2120. doi: 10.1093/bioinformatics/btu17
10. Wang, X., Lu, P., and Luo, Z. GMATo: A novel tool for the identification and analysis of microsatellites in large genomes, *Bioinformation*, 2013, vol. 9, no. 10, pp. 541–544.
11. Martins, W.S. Lucas, D.C.S., Neves, K.F.S., and Bertioli, D.J. WebSat - a web software for microsatellite marker development, *Bioinformation*, 2009, vol. 3, no. 6, pp. 282–283.
12. Devey, M.E., Bell, J.C., Smith, D.N., et al. A genetic linkage map for *Pinus radiata* based on RFLP, RAPD, and microsatellite markers, *Theor. Appl. Genet*., 1996, vol. 92, no. 6, pp. 673–679. doi:10.1007/BF00226088

**Table 1.** PCR primers for 27 microsatellite loci of Siberian larch with the most stable amplification based on the results of initial testing

| Locus | Motif and the number of its repeats in the contig | Nucleotide sequence of the PCR primer | The PCR fragment length, bp |
|---|---|---|---|
| *Ls_798886* | $(AC)_{26}$ | F: ACATGGTGAGGAGTTGAAGGAC<br>R: GGAGACAAAGAGGTAATGTGCC | 150-180 |
| *Ls_955052* | $(CT)_{21}$ | F: CTCTTGTGGTGAAGACATTGGA<br>R: AACCCAAACCCTAGACTTCGAT | 220-280 |
| *Ls_796783* | $(CT)_{21}$ | F: TTCAAGAGGCCAACATCCTTTA<br>R: GACTCAAGCAAACCACAATAATAGC | 160-190 |
| *Ls_19333* | $(TC)_{20}$ | F: TTGGCCTCCCTTACATCTCATA<br>R: CCTCACCTCCAATTCTCCAAG | 230-270 |
| *Ls_440296* | $(TC)_{20}$ | F: AGAGCACCTGAAGACTCCGATA<br>R: AGAAACCCAAAAGGTACAGCAA | 170-190 |
| *Ls_123209* | $(CA)_{21}$ | F: ATGGTGTTTTCTCTTCTTGGGA<br>R: TCTACATCAAATCGCTCATCGT | 250-255 |
| *Ls_250604* | $(GA)_{21}$ | F: AAATTGTAGGAGCCCACTTCAT<br>R: ACCCAATATCTCACCATCGTTT | 170-200 |
| *Ls_318520* | $(TC)_{25}$ | F: CAACAAATCTCCCCAATCACAT<br>R: GAACTGTATGAGGCAACAAGCA | 240-270 |
| *Ls_264438* | $(TC)_{20}$ | F: GGGGAGATCAAATCCACTATCA<br>R: TTGTGCAGGAGTGAAGAAAAGA | 180-200 |
| *Ls_621673* | $(TC)_{23}$ | F: CCCACATGCCAATAAAGGATTA<br>R: TACCATTACCCACCAGGAAGAC | 180-220 |
| *Ls_648755* | $(AT)_{22}$ | F: TCGCATACATAAAGAGACGCAT<br>R: GGAGGCGATTTGCATTTG | 180-220 |
| *Ls_1274831* | $(CT)_{20}$ | F: CCATGTTCAAATTCGGCTAAC<br>R: AGACATTTCGACCCTTGCTAAA | 200-220 |
| *Ls_199446* | $(AG)_{23}$ | F: CAATAAATGGTTGTCTTTGGGG<br>R: GGCCAATGCTTGTCGTTTA | 200-220 |
| *Ls_1106920* | $(AG)_{21}$ | F: TTGGTTGGATTCGATGAACT | 200-220 |

| Locus | Motif and the number of its repeats in the contig | Nucleotide sequence of the PCR primer | The PCR fragment length, bp |
|---|---|---|---|
| | | R: GATACTCTCCCCATTTGGCTAC | |
| *Ls_1089834* | $(AG)_{21}$ | F: GTTTTGTAAGGGTGCGATTTGT<br>R: TTTTACGTGGAAACCCAACTG | 200-220 |
| *Ls_245817* | $(CT)_{20}$ | F: ATCCAATGCCTCACAATTTC<br>R: GGAAACGAGATGTATATTTGGG | 200-240 |
| *Ls_540394* | $(AG)_{20}$ | F: CAAGTTCTCCAAGGCAAGACAT<br>R: CACATCGTATTGTTGGTATCCCT | 200-300 |
| *Ls_66831* | $(CA)_{21}$ | F: TGTCCCTCCATAACCTAAGAATG<br>R: AGCTCAAGGAAAAGACCCTACC | 220-240 |
| *Ls_326284* | $(GT)_{21}$ | F: TTGTATAAGCTCCCTCCAACG<br>R: CCCTTGGGGTTAATAGATTTCA | 220-260 |
| *Ls_396253* | $(AG)_{20}$ | F: CAAGTTCTCCAAGGCAATTCAT<br>R: CACATCGTATTGTTGGTATCCCT | 220-260 |
| *Ls_915025* | $(AT)_{21}$ | F: TTGTTAATTTCCATGCACGC<br>R: GGAATAGGTTTATAGGGCAGTCG | 170-190 |
| *Ls_1287563* | $(CA)_{22}$ | F: GGTTGTACCCTCTTCATTCCTTT<br>R: CCTGTGGATGGGAAATCTATATG | 230-240 |
| *Ls_1295092* | $(CT)_{20}$ | F: CAATCCTTGATCTCTTCATGGT<br>R: CAATCCTTGATCTCTTCATGGT | 210-220 |
| *Ls_12590* | $(TG)_{24}$ | F: AGCATAAGCACACATCATCACA<br>R: TGATACAACCTATGGAAGGCAA | 160-170 |
| *Ls_68475* | $(CT)_{20}$ | F: TGGTATATGTGGTTGTGATGCTC<br>R: GGGAATAGTTAAGGAGGGAAGG | 150-170 |
| *Ls_263105* | $(TG)_{21}$ | F: AAACTTGGTGCTTCAGACGG<br>R: TCACTAGCCTGACATTTGCATC | 220-240 |
| *Ls_304106* | $(GA)_{22}$ | F: ATTAGTGTCCAACTCCTTCCCA<br>R: ATTGGTTCTTTGTTCAAGGGTG | 230-240 |

Note. The annealing temperature was in accordance with the "touchdown" PCR program (60°→50° C); bp - nucleotide base pairs.

**Table 2.** PCR multiplexes tested for 20 primer pairs

| Locus | Fluorescent dye | PCR multiplex |
|---|---|---|
| *Ls_798886* | FAM | 1 |
| *Ls_955052* | FAM | |
| *Ls_796783* | XEX | |
| *Ls_19333* | XEX | |
| *Ls_440296* | FAM | 2 |
| *Ls_123209* | FAM | |
| *Ls_250604* | XEX | |
| *Ls_318520* | XEX | |
| *Ls_264438* | FAM | 3 |
| *Ls_621673* | XEX | |
| *Ls_648755* | FAM | 4 |
| *Ls_1274831* | XEX | |
| *Ls_199446* | FAM | 5 |
| *Ls_1106920* | XEX | |
| *Ls_1089834* | FAM | 6 |
| *Ls_245817* | XEX | |
| *Ls_540394* | FAM | 7 |
| *Ls_66831* | XEX | |
| *Ls_326284* | FAM | 8 |
| *Ls_396253* | XEX | |
| *Ls_915025* | FAM | 9 |
| *Ls_1287563* | XEX | |
| *Ls_1295092* | FAM | 10 |
| *Ls_12590* | XEX | |
| *Ls_68475* | FAM | 11 |
| *Ls_263105* | XEX | |
| *Ls_304106* | FAM | 12 |

**Table 3.** Features of the selected 11 best nuclear microsatellite loci ready for further use in population genetic studies of Siberian larch and to identify the origin of wood and plant material

| Locus | Motif and the number of its repeats in the contig | The PCR fragment length, bp | Number of alleles found* |
|---|---|---|---|
| *Ls_1106920* | $(AG)_{21}$ | 145-263 | 18 |
| *Ls_796783* | $(CT)_{21}$ | 151-177 | 13 |
| *Ls_955052* | $(CT)_{21}$ | 225-347 | 20 |
| *Ls_19333* | $(TC)_{20}$ | 223-255 | 13 |
| *Ls_440296* | $(TC)_{20}$ | 169-243 | 13 |
| *Ls_621673* | $(TC)_{23}$ | 184-218 | 16 |
| *Ls_1089834* | $(AG)_{21}$ | 187-235 | 26 |
| *Ls_1274831*** | $(CT)_{20}$ | 203-229 | 14 |
| *Ls_66831*** | $(CA)_{21}$ | 207-231 | 14 |
| *Ls_915025* | $(AT)_{21}$ | 132-186 | 16 |
| *Ls_12590*** | $(TG)_{24}$ | 151-201 | 23 |

Note. The annealing temperature was in accordance with the "touchdown" PCR program (60°→50° C); bp - nucleotide base pairs. *Based on the genotyping of 100 trees in two Siberian larch populations. **Loci, in which null alleles were found in 2-3 samples.

**Table 4.** Multiplex panels of selected microsatellite loci recommended for Siberian larch

| Locus | Fluorescent dye | PCR multiplex |
|---|---|---|
| *Ls_1106920* | XEX | 1 |
| *Ls_955052* | FAM | |
| *Ls_796783* | XEX | |
| *Ls_19333* | XEX | 2 |
| *Ls_621673* | XEX | |
| *Ls_440296* | FAM | |
| *Ls_1089834* | FAM | 3 |
| *Ls_1274831** | XEX | |
| *Ls_66831** | XEX | 4 |
| *Ls_915025* | FAM | |
| *Ls_12590** | XEX | |

Note. *Loci, in which null alleles were found in 2-3 samples.

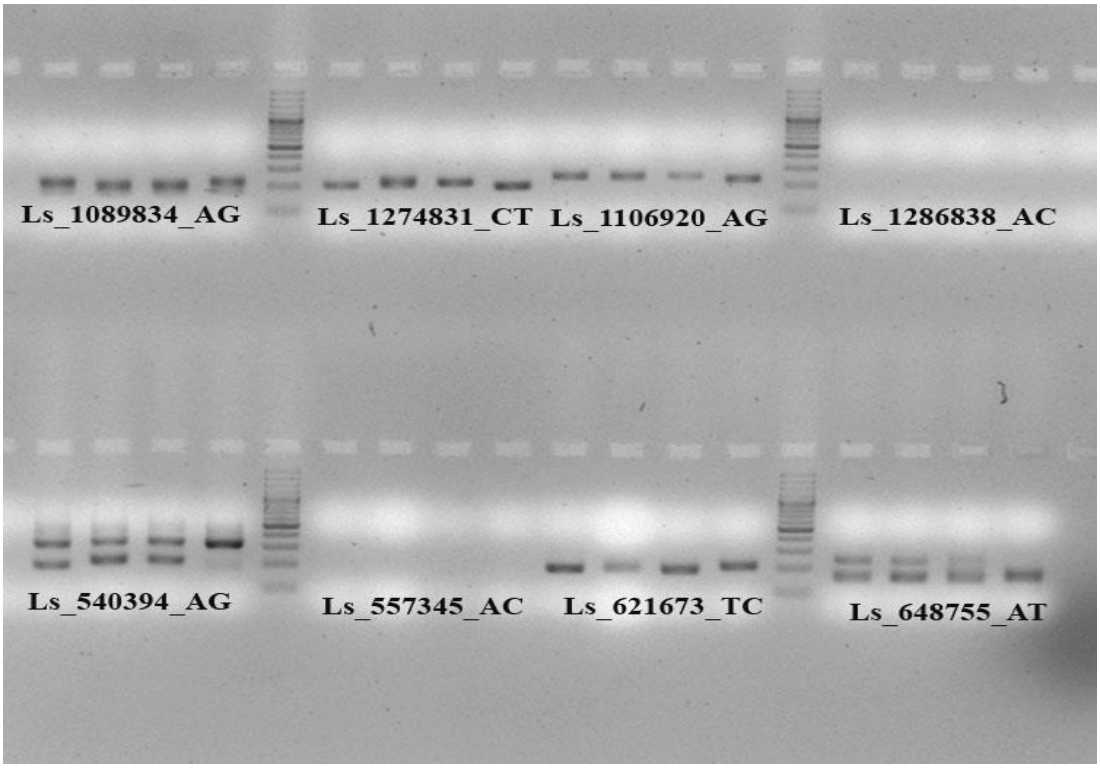


**Fig. 1**. Sample of a polyacrylamide gel stained after electrophoresis of the PCR amplification products amplified by 8 pairs of PCR primers during the initial testing of primers with four samples.