UDC 575.112 **Research article**

Application of MLVA Analysis of Chloroplast Microsatellite Loci for Differentiating Wild and Cultivated *Malus* spp. germplasm

Vladislav A. Shevtsov[1], Oxana N. Khapilina[1], Yulia M. Efremova[1,2], Ainur S. Sarsenova[1], Asem R. Tumenbayeva[1], Saule M. Magzumova[1]

[1] "National Center for Biotechnology" LLP, Astana, Kazakhstan
[2]Kazakh-German University, Almaty, Kazakhstan

Corresponding author: Vladislav A. Shevtsov: xatabadich@gmail.com
Co-authors: (1: OKh) oksfur@mail.ru; (2: YuE) yefremova@dku.kz, (3: AS) sarainur@mail.ru; (4: AT) asem.tumenbaeva2016@mail.ru; (5: SM) magzumovas@list.ru

Received: 10.07.2026 **Accepted:** 17.09.2026 **Published:** 30.09.2026

Copyright: Shevtsov V.A. et al. This article is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/)

Abstract

Background and Aim. The genus *Malus* comprises wild and cultivated apple forms of major importance for breeding, the conservation of genetic resources, and studies of fruit-crop evolution. Although SSR (Simple Sequence Repeat) and SNP markers are widely used to assess genetic diversity in apples, organellar markers remain important because the chloroplast genome predominantly reflects a single maternal lineage and has a simpler, more conserved structure than the nuclear genome. The aim of this study was to investigate the genetic variability of chloroplast loci in representatives of the genus *Malus* using multiple-locus analysis of tandem repeats and to identify differences between wild and cultivated germplasm.

Materials and Methods. Forty-two chloroplast assemblies deposited in GenBank/ENA and one assembly generated in this study were analysed; the chloroplast genome of *Prunus ursina* (NC_068527) served as the outgroup. Microsatellite and VNTR-like loci were identified with MISA, orthologous loci were clustered with custom Python scripts, and primers for the variable loci were designed with Primer3. MLVA profiles were built with MLVA_finder, pairwise distances were calculated under the Stepwise Mutation Model, and a neighbour-joining tree was constructed.

Results. MISA analysis identified 2,918 SSR loci, of which 96.1% were mononucleotide repeats, predominantly poly-A/T. Clustering reduced these to 92 orthologous loci, 48 of which (approximately 52%) varied in repeat number. The eight most informative primer pairs were selected for the cpMLVA panel. The resulting tree showed tight clustering of *M. domestica* samples, more peripheral positions for wild species, and clear separation of the *Prunus* outgroup. BLAST analysis of the complete chloroplast genome confirmed that a record annotated as *M. fusca* (OR897844.1) is taxonomically questionable.

Conclusion. Chloroplast tandem repeats can serve as a compact marker system for differentiating wild and cultivated representatives of the genus *Malus* and for detecting erroneous metadata in public databases.

Keywords: *Malus* spp.; chloroplast genome; cpSSR; tandem repeats; VNTR; MLVA.

Introduction

The genus *Malus* comprises approximately 30–55 species of wild and cultivated apples, distributed primarily across the temperate regions of the Northern Hemisphere, and is one of the most important groups in global fruit-crop breeding [1, 2]. The cultivated apple (*Malus domestica* Borkh.) is the product of prolonged hybridization among *M. sieversii, M. sylvestris, M. orientalis*, and *M. baccata*, with the Central Asian ancestor *M. sieversii*, whose range encompasses the mountainous regions of Kazakhstan, Kyrgyzstan, and neighboring countries, playing a key role in its formation [3]. Whole-genome studies confirm the complex reticulate history of cultivated apples and multiple introgression events from wild species [4].

Wild relatives of *M. domestica* are of interest because they constitute a valuable genetic reservoir of resistance to biotic and abiotic stresses, including *Venturia inaequalis, Erwinia amylovora*, low temperatures, and drought, and are actively used in modern breeding programs [5]. At the same time, the decline of natural *M. sieversii* populations and the erosion of allelic diversity in commercial cultivars make conservation and inventorying high-priority tasks [6]. This is particularly relevant to Kazakhstan, a center of origin of cultivated apples where the most diverse populations of wild apples are still found. Accurate sample identification, including the distinction between wild and cultivated forms and the verification of cultivar identity, is essential for the effective use of genetic collections [7]. Molecular markers (SSR, SNP, and GBS) are widely used for the passporting of *M. domestica* cultivars and the analysis of population structure [4, 7]; however, the applicability of compact, readily reproducible marker systems for the preliminary screening of large collections remains an open question.

The chloroplast genome (plastome) of angiosperms is a circular double-stranded DNA molecule averaging 150-160 kb in size, with a conserved quadripartite structure comprising a large single-copy region (LSC), a small single-copy region (SSC), and a pair of inverted repeats (IRa and IRb) [8]. Despite its relatively small size, the plastome encodes approximately 110-130 genes involved in photosynthesis, transcription, and translation and retains a high degree of structural conservation even at the plant order level [9]. In most angiosperms, including members of the Rosaceae, chloroplasts are maternally inherited, enabling unambiguous tracing of maternal phylogenetic relationships and preventing the mixing of parental alleles during hybridization [10]. The low level of recombination and relatively slow substitution rate compared with the nuclear genome make the plastome a convenient system for phylogeographic and species-specific studies [11]. In addition to point substitutions and indels, plastomes contain abundant microsatellite (cpSSR) and longer tandem (VNTR-like) repeats, predominantly mononucleotide poly-A/T repeats concentrated mainly in the LSC region [12]. These loci have elevated mutation rates owing to replication errors (polymerase slippage) and are suitable for distinguishing closely related taxa and populations [13]. cpSSR markers have proven effective both for plant barcoding [14] and for analyzing the population structure of woody and fruit crops [12, 13]. Taken together, these properties-compactness, conservation of flanking regions, maternal inheritance, and the presence of hypervariable repeats-support viewing the plastome as a natural mini-system of markers suitable for the development of compact assay panels.

Microsatellites (Simple Sequence Repeats, SSRs) are short tandem repeats with monomers 1–6 bp long that are widely distributed in both nuclear and organellar genomes [15]. Variation in repeat number is driven primarily by DNA polymerase replication errors and produces high polymorphism while conserved flanking regions permit the design of primers for PCR amplification [16]. Longer tandem repeats (Variable Number Tandem Repeats, VNTRs; monomers ranging from several to tens of base pairs) follow similar mutational mechanisms and provide comparable discriminatory power [17]. Multiple-locus VNTR/SSR analysis (MLVA) simultaneously assesses repeat numbers at several independent loci, enabling the construction of reproducible allelic profiles and improving the discriminatory power of genotyping closely related samples [18]. The use of several markers also reduces the influence of homoplasy at individual loci and makes the method suitable for clustering, comparative analysis, and the construction of standardized MLVA profile databases [19]. In the present study, this logic was adapted to the *Malus* chloroplast genome: variable cpSSR/VNTR loci were treated not as individual markers but as a compact multilocus panel for assessing maternal plastome lineages and differentiating wild and cultivated germplasm.

The purpose of this study was to investigate the genetic variability of chloroplast loci in representatives of the genus *Malus* using multiple-locus analysis of tandem repeats and to identify differences between wild and cultivated germplasm.

Materials and Methods

DNA Extraction and Sequencing of *Malus* spp.

Genomic DNA was extracted from leaf material according to a protocol developed for plant plastomes [20]. Libraries were prepared using the Nextera XT DNA Library Preparation Kit (Illumina, USA). Sequencing was performed on an Illumina MiSeq platform using the MiSeq Reagent Kit v3 (600 cycles). The quality of raw reads was assessed using FastQC v0.11.15 [21], and aggregate quality assessment across all libraries was performed using MultiQC [22]. Next-generation sequencing (NGS)

was used to analyze *Malus* chloroplast DNA. Chloroplast DNA was isolated according to the protocol of *Chao Shi* et al. [20].

The dataset for MLVA analysis comprised 43 Malus chloroplast assemblies: 42 assemblies deposited in GenBank/ENA and one assembly generated in the present study (Chloroplast-162, shown as Chloroplast-162 in Figure 3). The chloroplast genome of *Prunus ursina* (NC_068527) was used as the outgroup. Thus, the final matrix used for the MLVA tree contained 44 samples in total (43 *Malus* and one Prunus outgroup), which is consistent with Figure 3. The selected samples, with their accession numbers and species assignments, are listed in Appendix 1.

Microsatellite and VNTR-like loci were identified using MISA (MIcroSAtellite identification tool, version 25.08.2020) [23] with the following parameters: definition (unit_size, min_repeats) = 1-10, 2-6, 3-5, 4-5, 5-5, and 6-5; interruptions(max_difference_between_2_SSRs) = 100. The MISA loci from all samples were clustered using custom Python scripts, with loci occurring at homologous positions treated as orthologous clusters. Loci with more than one allele across the dataset were selected as candidates for the cpMLVA panel.

For the selected variable loci, flanking regions were used to design primers with Primer3 [24] (default parameters for a 50–300 bp product, Tm 56–62 °C, and GC content 30–60%). In parallel, a manually selected primer pair was generated for each locus directly from conserved 20-bp flanking regions, followed by evaluation of Tm, GC content, homopolymeric tracts, and dimerization propensity. A total of 48 candidate primer pairs were generated and ranked *in silico* using a composite metric. Based on this ranking, the eight most informative and bioinformatically valid primer pairs were selected for the cpMLVA panel (Table 1).

MLVA profiles were constructed using the MLVA_finder software package (i2bc, https://github.com/i2bc/MLVA_finder), which determines the repeat copy number at each locus from assemblies and generates a sample-by-locus size/repeat-number matrix. The resulting profiles were analysed with the custom Python script mlva_distance_analysis.py, which converts repeat calls to inferred locus sizes, calculates pairwise distances using an SMM-style mean absolute size-distance metric, and constructs a neighbor-joining (NJ) tree with Bio.Phylo DistanceTreeConstructor. Missing values are ignored for pairwise distance calculation and replaced only by a maximum-distance fallback if two samples share no loci; in the final matrix used for Figure 3, no missing marker cells were reported. The display tree was rooted by automatic outgroup detection based only on the distance matrix, and *P. ursina* (NC_068527) was identified as a singleton outgroup. Bootstrap or other branch-support values were not calculated; therefore, the tree is interpreted as an MLVA profile-relatedness visualization. The MLVA distance analysis module and accompanying script are available on GitHub: https://github.com/Vladislav-Shevtsov/MLVA_distance_analysis.

De novo assembly of the *Malus* chloroplast genome generated in this study from Illumina MiSeq data was performed using Geneious Prime v2024.

Results

The sample used for chloroplast DNA extraction was collected on the premises of the “National Center for Biotechnology” LLP on 21 May 2026. The *Malus* spp. chloroplast genome was annotated using Chloe 1.0 (Figure 1). The chloroplast genome was 160,241 bp long; a total of 318 structural elements were annotated, including 131 gene features, 103 CDSs, 45 tRNAs, 8 rRNAs, 29 introns, and 2 repeat region features. The genome contained 112 unique genes: 78 protein-coding genes, 30 tRNA genes, and 4 rRNA genes.

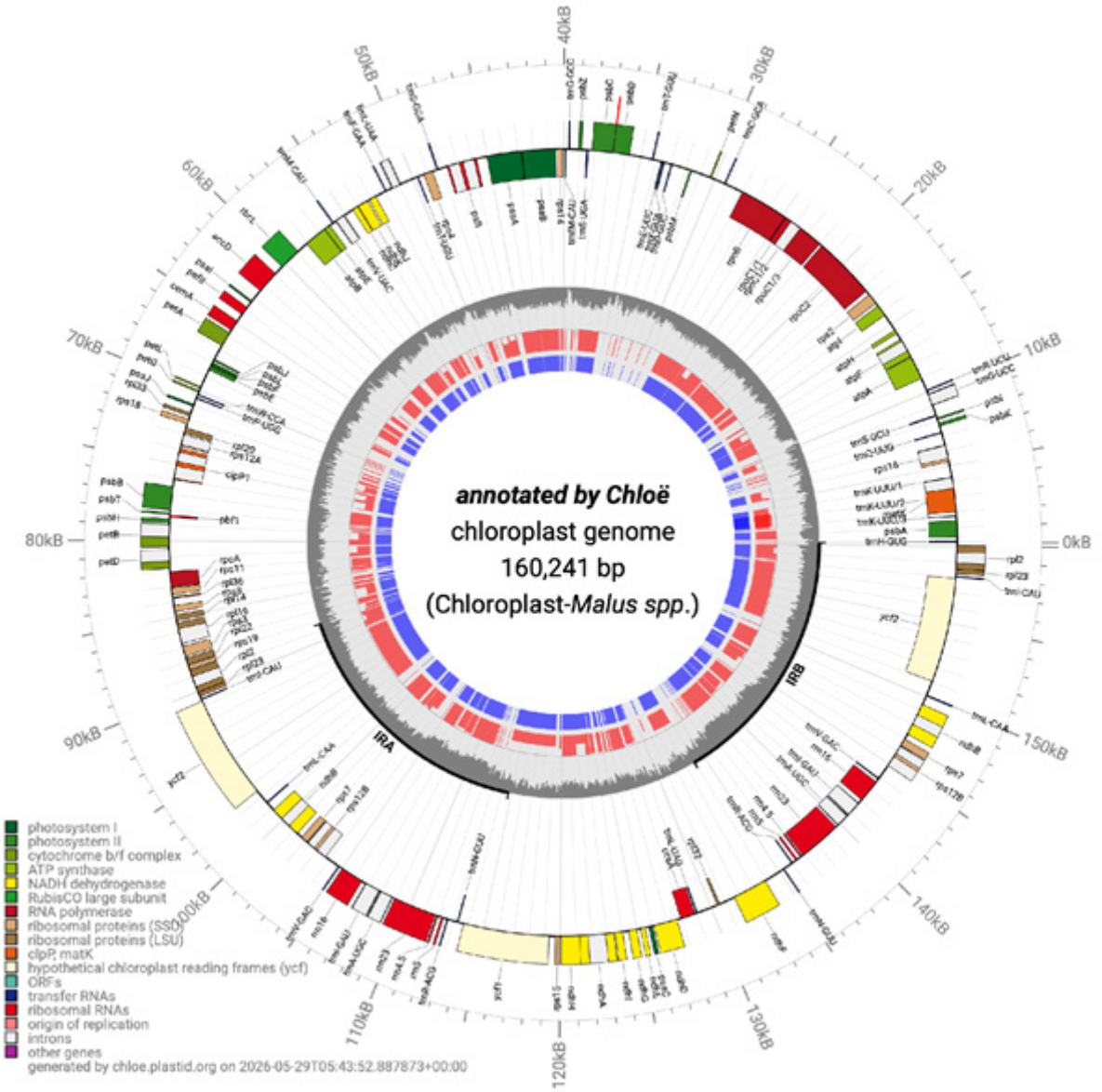


Figure 1. Gene structure in the *Malus* spp. chloroplast genome

The annotated genes included photosystem genes, ATP synthase genes, genes of the NADH dehydrogenase complex, ribosomal proteins, and rRNA/tRNA components. The laboratory component of the study-including DNA extraction, sample preparation, and molecular genetic procedures-was also carried out at the National Center for Biotechnology.

Selection of SSR/VNTR Loci in Chloroplast Genomes

For cpMLVA analysis, 42 public *Malus* chloroplast assemblies (Appendix 1) were selected, representing nine species of the genus: *M. domestica, M. fusca, M. doumeri, M. kansuensis, M. sieboldii, M. sikkimensis, M. rockii, M. baccata*, and one interspecific hybrid form, Malus hybrid. One assembly generated in the present study (Chloroplast-162 in Figure 3) was added to the MLVA profiling dataset. The *Prunus ursina* chloroplast genome (NC_068527.1) was included as an outgroup. MISA analysis of the 42 oriented, full-length public *Malus* chloroplast sequences identified 2,918 SSR loci; 393 loci (approximately 13.5%) occurred in compound configurations. The distribution by monomer length was strongly skewed toward mononucleotide repeats: 2,803 loci (96.1%) had a single-nucleotide monomer (predominantly poly-A/T: 2,699 of 2,803 loci), whereas 115 loci were dinucleotide repeats (predominantly AT/AT) (Figure 2). No tri-, tetra-, penta-, or hexanucleotide repeats were detected under the specified parameters.

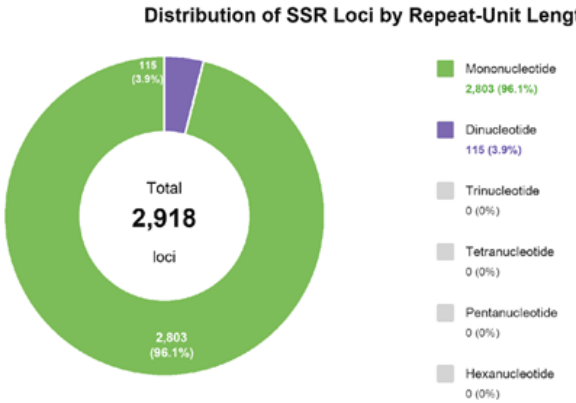


Figure 2. Distribution of SSR loci identified in *Malus* chloroplast genomes by monomer length

Aggregate clustering of loci against the reference assembly MK434916.1 (*M. domestica)* reduced the 2,918 primary MISA signals to 92 nonempty clusters (orthologous loci). Of these, 48 loci (approximately 52%) varied in repeat number, whereas 44 were monomorphic. Based on a combination of informativeness index, presence rate, and flanking-region quality, 48 initial primer pairs (designed with Primer3 or selected manually) were chosen from the variable loci and ranked *in silico*. The final cpMLVA panel comprised the eight most informative and technically valid primer pairs according to Primer3 analysis (Table 1).

Table 1. Eight most informative primer pairs for cpSSR/VNTR loci selected for the *Malus* spp. cpMLVA panel

№	Locus	Forward primer (5′→3′)	Reverse primer (5′→3′)
1	cpSSR012_1bp_62bp_15u	GTAAAATCCTTACATGATGC	GATCCCTTTATCAGCCAAAA
2	cpSSR053_1bp_52bp_12u	GAAAGGTGAAAGCTAAGACG	AATACTTGAACGGTCGATTC
3	cpSSR005_1bp_60bp_13u	CCCACACTTCTTTTCTTCTC	AGGGTACTTCATCTCATTTT
4	cpSSR057_1bp_68bp_10u	TCAGAAAAGGTGGAGGTTTT	CCGTTGAGTTCTTAATAATC
5	cpSSR055_1bp_62bp_11u	GAGTTACTTCGGAAACAAAA	GGAATTGGGAGAATAACAAA
6	cpSSR009_1bp_57bp_13u	ACCATGAGGAAGAACGTAAC	GAACAGTAAAGAGGTAAACA
7	cpSSR059_1bp_58bp_12u	ATGCATACGGTTCAACAAAA	CTACATCTACACATTGATTC
8	cpSSR052_1bp_68bp_11u	ACTCAAATTGTTTCGTTTTGACAGG	TCAAATTCATGAACCGTCCCC

MLVA Clustering

Using the custom script mlva_distance_analysis.py together with SciPy, seaborn, matplotlib, pandas, NumPy, and Bio.Phylo, a matrix of eight cpSSR/VNTR markers was constructed for 44 samples (43 *Malus* samples, including Chloroplast-162, and one *Prunus* outgroup) (Figure 3). The corresponding QC output reported 44 samples, eight MLVA loci, zero missing marker cells, the smm-size distance metric, NJ tree construction, and automatic identification of NC_068527_Prunus_ursina as the singleton outgroup. The MLVA tree showed clear clustering of *M. domestica* samples and separation of wild species and the outgroup.

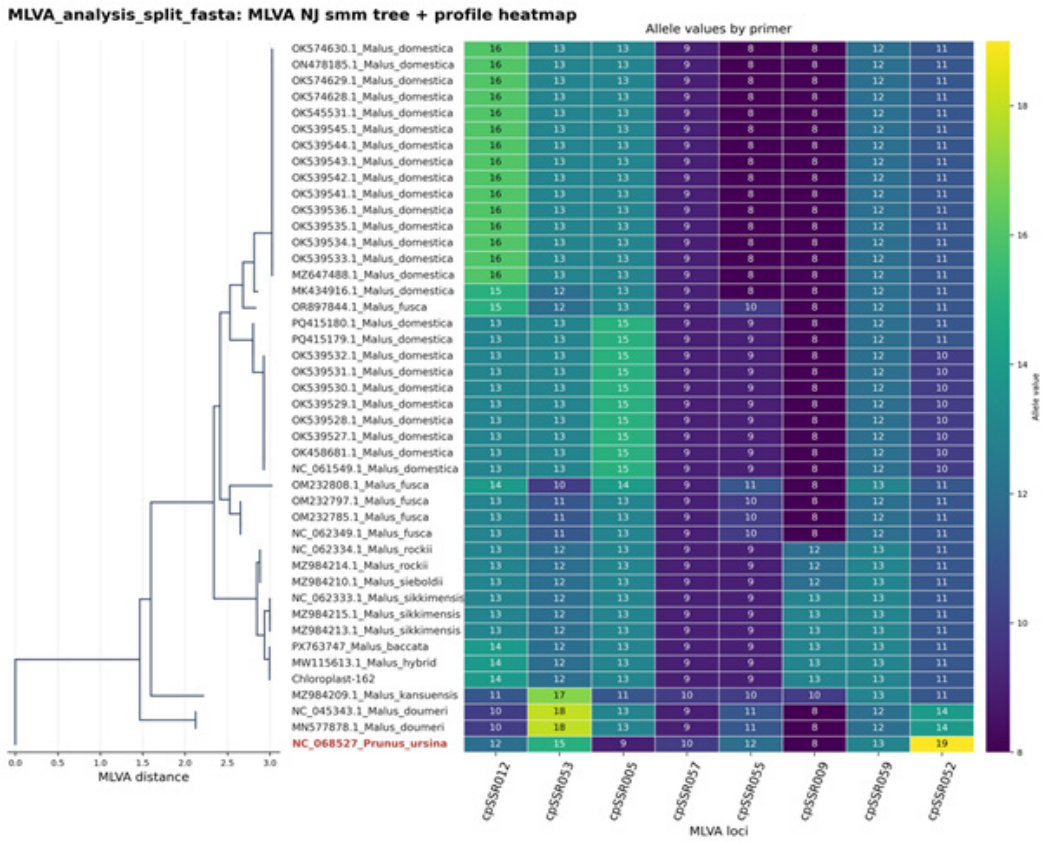


Figure 3. cpSSR/VNTR marker matrix for 44 samples of the genera *Malus* and *Prunus*

Wild *Malus* species occupied more peripheral positions, reflecting the accumulation of differences in their cpMLVA profiles. The outgroup sample *Prunus ursina* was clearly separated from all *Malus* representatives, confirming the informativeness of the selected loci for detecting intergeneric differences.

According to BLAST analysis of the complete chloroplast genome, the sample annotated as *Malus fusca* (OR897844.1) did not show specific similarity to *M. fusca*. When the results were sorted by Total Score, the highest values were obtained for *Malus sylvestris* (7.468e+05), followed by *M. sieversii* (4.589e+05), *M. domestica* (4.130e+05), and *M. sylvestris* (4.127e+05). *M. fusca* records had lower Total Score values, approximately 3.89–3.92e+05, and therefore were not supported as the closest group by this criterion.

When the results were sorted by Percent Identity, the closest matches were other species of the genus *Malus*: *M. hupehensis* (99.83%), *M. sieversii* (99.78%), and *M. micromalus, M. baccata, Malus* sp., and *M. sylvestris* (mostly 99.70–99.77%).

Given that *M. sieversii* is considered the principal ancestor of the cultivated apple *Malus domestica* and that *M. sieversii* accounted for the largest number of species-level matches in the BLAST analysis, the record was most likely submitted to NCBI with a provisional or erroneous species annotation. The position of this sample within the M. domestica cluster in the MLVA tree is consistent with the BLAST results and shows that analysis using eight chloroplast MLVA markers correctly reflects its actual genetic similarity. This finding supports the validity of the MLVA approach for analyzing plant chloroplast genomes and identifying questionable metadata in databases.

Discussion

The results support the applicability of chloroplast MLVA as an independent marker strategy. From a methodological perspective, transferring MLVA logic from bacterial genomes to angiosperm plastomes is reasonable because the plastome is comparable to a bacterial chromosome in size and compactness, undergoes little meaningful recombination, and is maternally inherited. Its cpSSR/VNTR allelic profile

can therefore be treated as a single haploprofile. Compact cpMLVA profiling substantially reduces the amount of data required to differentiate samples. Rather than performing whole-plastome alignment, the MLVA method alone can be applied. This approach is readily transferable among laboratories, is compatible with standard clustering methods used in bacterial MLVA, and may be suitable for building public MLVA-based databases for angiosperm chloroplast genomes.

The main methodological difference from earlier cpSSR work lies in the way the marker set was assembled. Universal cpSSR primer sets for dicotyledons comprise about ten loci designed from conserved plastome regions [13], and their transfer to non-model taxa is frequently accompanied by loss of amplification or of polymorphism [12]. In this work the panel was built de novo within the genus: clustering of the primary MISA signals against the *M. domestica* reference reduced 2,918 raw loci to 92 orthologous positions, of which 48 (approximately 52%) varied in repeat number. The proportion of informative loci is therefore considerably higher than is usually obtained with universal primer sets, because selection was based directly on observed allelic variation within the target group rather than on flanking-region conservation alone.

The topology obtained with the eight-marker panel reproduces the principal features described in previous chloroplast studies of the genus. Using plastid sequence data, *Volk* et al. [6] showed that *M. domestica* cultivars carry a narrow set of chloroplast haplotypes, whereas wild *Malus* species retain markedly greater plastome heterogeneity, which is attributed to a limited number of maternal lineages having entered the domesticated gene pool. The tight cluster of nearly identical cpMLVA profiles formed by the *M. domestica* samples, the more peripheral positions of the wild species, and the clear separation of the *Prunus* outgroup in our tree are in direct agreement with that observation.

The result is also consistent with nuclear-marker and whole-genome analyses describing cultivated apples as the product of reticulate evolution with a dominant *M. sieversii* contribution and repeated introgression from *M. sylvestris* and other wild species [1, 3, 4]. Because the plastome is maternally inherited, cpMLVA resolves only the maternal component of this history; the limited resolution within *M. domestica* should therefore be read as a property of the marker system rather than as an absence of diversity among cultivars, which nuclear SSR and SNP panels resolve well [4, 7].

The result obtained for accession OR897844.1 illustrates a practical application of the panel. This record, deposited as *M. fusca*, fell within the *M. domestica* cluster in the cpMLVA tree, and whole-plastome BLAST comparison returned *M. sylvestris, M. sieversii*, and *M. domestica* rather than *M. fusca* as the closest matches. Discrepancies between the declared taxonomic affiliation of public sequence records and their actual similarity are a recognised problem in plant barcoding, where the reliability of identification depends primarily on the quality of reference-database annotation [14]. cpMLVA does not replace whole-plastome comparison, but it flags such records at a small fraction of the computational cost, and the agreement between the eight-locus profile and the whole-genome BLAST result indicates that the reduced panel retains sufficient signal for this purpose.

Several limitations should be noted. Sampling was uneven: most wild species were represented by single accessions, so intraspecific variation is probably underestimated. Homoplasy at individual loci is also expected, because repeat-number states are limited and identical allele sizes can arise independently; the use of several independent loci mitigates this effect but does not eliminate it, as is well documented for bacterial MLVA schemes [18, 19]. Expanding the panel and increasing the sampling of wild populations, in particular of *M. sieversii* in Kazakhstan, are therefore the next steps.

Potential practical applications of the approach include screening *Malus* collections in centers of origin and commercial germplasm banks, where cpMLVA could rapidly and inexpensively detect divergent maternal lineages, potential introgressants from wild species, and erroneous duplicates.

Conclusion

The 42 public *Malus* chloroplast assemblies screened with MISA contained a rich set of variable tandem repeats, predominantly mononucleotide repeats, located in the LSC region of the plastome. The adapted MLVA-like approach (cpMLVA), implemented using a panel of eight primer pairs and evaluated on 43 *Malus* profiles together with the *Prunus* outgroup, has potential for differentiating wild and cultivated *Malus* representatives.

In addition, the method was further validated by BLAST analysis of the complete chloroplast genome of the disputed *Malus fusca* sample. When ranked by Total Score, the greatest similarity was observed

not with *M. fusca* but with *M. sylvestris* (7.468e+05), *M. sieversii* (4.589e+05), and M. domestica (4.130e+05), whereas *M. fusca* records had lower values. When sorted by Percent Identity, the closest matches also included several *Malus* species with high identity (99.70–99.83%), indicating ambiguity in the original *M. fusca* annotation in NCBI. The position of this sample within the *M. domestica* group in the MLVA tree was consistent with the BLAST results and confirmed the informativeness of the cpMLVA panel for identifying taxonomically questionable records. In our analysis, the eight-locus profile reproduced the principal groups within *Malus* and correctly separated the *Prunus* outgroup, confirming that the panel provides sufficient resolution.

Authors' Contributions

VSh and OKh: Conceptualized and designed the study, developed the cpMLVA workflow, performed the bioinformatic analysis, interpreted the results, and drafted the manuscript. YuE and ASa: Participated in data analysis, interpretation of chloroplast genome variation, sample processing, and validation of the analytical outputs. ATu: Provided laboratory support, data curation, and preparation of materials for analysis. SMa: Participated in chloroplast DNA extraction. All authors have read, reviewed, and approved the final manuscript.

Financing Information

This work was supported by the Ministry of Education of the Republic of Kazakhstan in the framework of program funding for research AP26104633. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1 Cornille, A., Giraud, T., Smulders, M.J.M., Roldán-Ruiz, I., Gladieux, P. (2014). The domestication and evolutionary ecology of apples. *Trends in Genetics*, 30(2), 57–65. DOI: 10.1016/j.tig.2013.10.002.

2 Peter, J. (2020). The *Malus* collection. *Royal Botanical Gardens.* https://www.rbg.ca/the-malus-collection/

3 Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Viola, R. (2010). The genome of the domesticated apple (Malus × domestica Borkh.). *Nature Genetics*, 42(10), 833–839. DOI:10.1038/ng.654.

4 Migicovsky, Z., Gardner, K.M., Richards, C., Chao, C.T., Schwaninger, H.R., Fazio, G., Zhong, G.-Y., Myles, S. (2021). Genomic consequences of apple improvement. *Horticulture Research*, 8, 9. DOI:10.1038/s41438-020-00441-7.

5 Tegtmeier, R., Švara, A., Gritsenko, D., Khan, A. (2025). *Malus sieversii*: a historical, genetic, and conservational perspective of the primary progenitor species of domesticated apples. *Horticulture Research*, 12(1), uhae244. DOI:10.1093/hr/uhae244.

6 Volk, G.M., Henk, A.D., Baldo, A., Fazio, G., Chao, C.T., Richards, C.M. (2015). Chloroplast heterogeneity and historical admixture within the genus *Malus. American Journal of Botany,* 102(7), 1198–1208. DOI:10.3732/ajb.1500095.

7 Liang, W., Dondini, L., De Franceschi, P., Paris, R., Sansavini, S., Tartarini, S. (2015). Genetic diversity, population structure and construction of a core collection of apple cultivars from Italian germplasm. *Plant Molecular Biology Reporter,* 33(3), 458–473. DOI:10.1007/s11105-014-0754-9.

8 Wicke, S., Schneeweiss, G.M., dePamphilis, C.W., Müller, K.F., Quandt, D. (2011). The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Molecular Biology*, 76(3), 273–297. DOI:10.1007/s11103-011-9762-4.

9 Daniell, H., Lin, C.-S., Yu, M., Chang, W.-J. (2016). Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biology*, 17(134), 2–29. DOI:10.1186/s13059-016-1004-2.

10 Mower, J.P., Vickrey, T.L. (2018). Structural diversity among plastid genomes of land plants. *Advances in Botanical Research*, 85, 263–292. DOI: 10.1016/bs.abr.2017.11.013.

11 Provan, J., Powell, W., Hollingsworth, P.M. (2001). Chloroplast microsatellites: new tools for studies in plant ecology and evolution. *Trends in Ecology & Evolution,* 16(3), 142–147. DOI:10.1016/S0169-5347(00)02097-8.

12 Ebert, D., Peakall, R. (2009). Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. *Molecular Ecology Resources*, 9(3), 673–690. DOI:10.1111/j.1755-0998.2008.02319.x.

13 Weising, K., Gardner, R.C. (1999). A set of conserved PCR primers for the analysis of simple sequence repeat polymorphisms in chloroplast genomes of dicotyledonous angiosperms. *Genome,* 42(1), 9–19. DOI:10.1139/g98-104.

14 Hollingsworth, P.M., Graham, S.W., Little, D.P. (2011). Choosing and using a plant DNA barcode. *PLoS ONE*, 6(5), e19254. DOI: 10.1371/journal.pone.0019254.

15 Tautz, D. (1989). Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Research,* 17(16), 6463–6471. DOI:10.1093/nar/17.16.6463.

16 Schlötterer, C. (2000). Evolutionary dynamics of microsatellite DNA. *Chromosoma*, 109(6), 365–371. DOI:10.1007/s004120000089.

17 van Belkum, A., Scherer, S., van Alphen, L., Verbrugh, H. (1998). Short-sequence DNA repeats in prokaryotic genomes. *Microbiology and Molecular Biology Reviews*, 62(2), 275–293. DOI:10.1128/MMBR.62.2.275-293.1998.

18 Lindstedt, B.-A. (2005). Multiple-locus variable number tandem repeats analysis for genetic fingerprinting of pathogenic bacteria. *Electrophoresis*, 26(13), 2567–2582. DOI:10.1002/elps.200500096.

19 Vergnaud, G., Pourcel, C. (2009). Multiple locus variable number of tandem repeats analysis. *Methods in Molecular Biology*, 551, 141–158. DOI:10.1007/978-1-60327-999-4_12.

20 Shi, C., Hu, N., Huang, H., Gao, J., Zhao, Y.J., Gao, L.Z. (2012). An improved chloroplast DNA extraction procedure for whole plastid genome sequencing. *PLoS ONE*, 7(2), e31468. DOI: 10.1371/journal.pone.0031468.

21 Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. *Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom*. http://www.bioinformatics.babraham.ac.uk/projects/fastqc

22 Ewels, P., Magnusson, M., Lundin, S., Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19), 3047–3048. DOI:10.1093/bioinformatics/btw354.

23 Beier, S., Thiel, T., Münch, T., Scholz, U., Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. *Bioinformatics*, 33(16), 2583–2585. DOI:10.1093/bioinformatics/btx198.

24 Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., Rozen, S.G. (2012). Primer3 - new capabilities and interfaces. *Nucleic Acids Research,* 40(15), e115. DOI:10.1093/nar/gks596.

Appendix 1 – Accession numbers of chloroplasts used in the article (42 *Malus* accessions and the *Prunus ursina* outgroup)

MK434916.1, OK458681.1, OK539544.1, MN577878.1, OK539527.1, OK539545.1, MW115613.1, OK539528.1, OK545531.1, MZ647488.1, OK539529.1, OK574628.1, MZ984209.1, OK539530.1, OK574629.1, MZ984210.1, OK539531.1, OK574630.1, MZ984213.1, OK539532.1, OM232785.1, MZ984214.1, OK539533.1, OM232797.1, MZ984215.1, OK539534.1, OM232808.1, NC_045343.1, OK539535.1, OR897844.1, NC_061549.1, OK539536.1, PQ415179.1, NC_062333.1, OK539541.1, PQ415180.1, NC_062334.1, OK539542.1, ON478185.1, NC_062349.1, OK539543.1, PX763747.1, NC_068527.1

Malus spp. жабайы және мәдени гермоплазмасын ажырату үшін хлоропласт микросателлиттік локустарының MLVA-талдауын қолдану

Шевцов В.А.[1], Хапилина О.Н.[1], Ефремова Ю.М.[1,2], Сарсенова А.С.[1], Туменбаева А.Р.[1], Магзумова С.М.[1]

[1]«Ұлттық биотехнология орталығы» ЖШС, Астана, Қазақстан
[2]Қазақстан-Неміс университеті, Алматы, Қазақстан

Түйін

Алғышарттар мен мақсат. *Malus* туысы селекция, генетикалық ресурстарды сақтау және жеміс дақылдарының эволюциясын зерттеу үшін маңызы зор жабайы және мәдени алма формаларын біріктіреді. Алманың генетикалық әртүрлілігін бағалауда SSR және SNP маркерлері кеңінен қолданылғанымен, органеллалық маркерлер өз маңызын сақтайды, өйткені хлоропласт геномы негізінен бірыңғай аналық тұқымқуалаушылық желісін көрсетеді және ядролық геномға қарағанда қарапайым әрі консервативті құрылымға ие. Зерттеудің мақсаты – тандемді қайталанулардың көплокустық талдауы арқылы *Malus* туысы өкілдерінің хлоропласт локустарының генетикалық өзгергіштігін зерттеу және жабайы мен мәдени гермоплазма арасындағы айырмашылықтарды анықтау.

Материалдар мен әдістер. GenBank/ENA дерекқорына орналастырылған 42 хлоропласт жинағы және осы жұмыста алынған бір жинақ талданды; сыртқы топ ретінде *Prunus ursina* (NC_068527) хлоропласт геномы пайдаланылды. Микросателлиттік және VNTR тәрізді локустар MISA бағдарламасымен анықталды, ортологтық локустар Python скрипттерімен кластерленді, өзгермелі локустарға праймерлер Primer3 арқылы жобаланды. MLVA профильдері MLVA_finder пакетімен құрылды, жұптық қашықтықтар Stepwise Mutation Model бойынша есептеліп, neighbor-joining ағашы тұрғызылды.

Нәтижелер. MISA талдауы 2 918 SSR локусын анықтады, олардың 96,1%-ы мононуклеотидті қайталанулар (негізінен поли-A/T) болды. Кластерлеу нәтижесінде 92 ортологтық локус қалды, оның 48-і (шамамен 52%) қайталану санымен ерекшеленді. cpMLVA панеліне ең ақпаратты сегіз праймер жұбы таңдалды. Алынған ағашта *M. domestica* үлгілері тығыз кластер түзді, жабайы түрлер шеткі орындарды иеленді, ал *Prunus* сыртқы тобы айқын бөлінді. Толық хлоропласт геномының BLAST талдауы *M. fusca* (OR897844.1) деп белгіленген жазбаның таксономиялық тұрғыдан күмәнді екенін растады.

Қорытынды. Хлоропласт тандемді қайталанулары *Malus* туысының жабайы және мәдени өкілдерін ажырату үшін, сондай-ақ ашық дерекқорлардағы қате метадеректерді анықтау үшін ықшам маркерлік жүйе бола алады.

Кілт сөздер: *Malus* spp.; хлоропласт геномы; cpSSR; тандемді қайталанулар; VNTR; MLVA.

Применение MLVA-анализа микросателлитных локусов хлоропластов для дифференциации дикорастущего и культурного генофонда *Malus* spp.

Шевцов В.А.[1], Хапилина О.Н.[1], Ефремова Ю.М.[1,2], Сарсенова А.С.[1], Туменбаева А.Р.[1], Магзумова С.М.[1]

[1]ТОО «Национальный центр биотехнологии», Астана, Казахстан
[2]Казахстанско-Немецкий университет, Алматы, Казахстан

Аннотация

Предпосылки и цель. Род *Malus* объединяет дикорастущие и культурные формы яблони, имеющие важное значение для селекции, сохранения генетических ресурсов и изучения эволюции плодовых культур. Несмотря на широкое применение SSR- и SNP-маркеров для оценки генетического разнообразия яблони, органелльные маркеры сохраняют своё значение, поскольку хлоропластный геном преимущественно отражает единую материнскую линию наследования и имеет более простую и консервативную структуру по сравнению с ядерным геномом. Цель

работы – изучить генетическую изменчивость хлоропластных локусов представителей рода *Malus* с помощью мультилокусного анализа тандемных повторов и выявить различия между дикорастущим и культурным генофондом.

Материалы и методы. Проанализированы 42 хлоропластные сборки, депонированные в GenBank/ENA, и одна сборка, полученная в настоящей работе; в качестве внешней группы использован хлоропластный геном *Prunus ursina* (NC_068527). Микросателлитные и VNTR-подобные локусы выявляли с помощью MISA, ортологичные локусы кластеризовали собственными скриптами на Python, праймеры к вариабельным локусам подбирали в Primer3. MLVA-профили строили в пакете MLVA_finder, попарные расстояния рассчитывали по модели ступенчатых мутаций (SMM), после чего строили дерево методом neighbor-joining.

Результаты. Анализ MISA выявил 2 918 SSR-локусов, из которых 96,1% составили мононуклеотидные повторы, преимущественно поли-A/T. Кластеризация свела их к 92 ортологичным локусам, 48 из которых (около 52%) различались по числу повторов. Для панели cpMLVA отобраны восемь наиболее информативных пар праймеров. Полученное дерево показало плотную кластеризацию образцов *M. domestica*, более периферийное положение дикорастущих видов и чёткое отделение внешней группы *Prunus*. BLAST-анализ полного хлоропластного генома подтвердил, что запись, аннотированная как *M. fusca* (OR897844.1), является таксономически спорной.

Заключение. Хлоропластные тандемные повторы могут служить компактной маркерной системой для дифференциации дикорастущих и культурных представителей рода *Malus*, а также для выявления ошибочных метаданных в открытых базах данных.

Ключевые слова: *Malus* spp.; хлоропластный геном; cpSSR; тандемные повторы; VNTR; MLVA.