International Journal of Agriculture, Environment and Food Sciences
e-ISSN : 2618-5946
www.jaefs.com
JAEFS
edit PUBLISHING

Research Article
Int J Agric Environ Food Sci 6 (1): 135-140 (2022)

Evaluation of genetic structure of pistachio through whole genome resequencing

Harun Karci **Salih Kafkas***

Çukurova University, Agriculture Faculty, Horticulture Department, Adana, Türkiye

*Corresponding Author: skafkas@cu.edu.tr

Citation
Karci, H., Kafkas, S. (2022). evaluation of genetic structure of pistachio through whole genome resequencing. Journal of Agriculture, Environment and Food Sciences, 6 (1), 135-140.
Doi
https://doi.org/10.31015/jaefs.2022.1.18

Received: 04 January 2022
Accepted: 14 March 2022
Published Online: 30 March 2022
Revised: 02 April 2022

Year: 2022
Volume: 6
Issue: 1 (March)
Pages: 135-140

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-NC) license https://creativecommons.org/licenses/by-nc/4.0/

Copyright © 2022
International Journal of Agriculture, Environment and Food Sciences; Edit Publishing, Diyarbakır, Türkiye.

Available online
http://www.jaefs.com
https://dergipark.org.tr/jaefs

Abstract

Pistachio (*Pistacia vera* L.) is the only edible and cultivated species. Pistachio is the only economically importance and dioecious species in the genus *Pistacia*. There are basic problems in pistachio breeding such as dioecious flower structure, long juvenile period and alternate bearing. These problems can be overcome not with classical breeding methods, but with modern molecular breeding methods. In this study, very high numbers of single nucleotide polymorphism (SNP), insertion/deletion (InDel), structural variants (SV) and copy number variation (CNV) were determined by using the next generation sequencing data of the pistachio genotype obtained with 15x sequencing coverage. A total of 1,785,235 SNP, 260,683 InDel, 5,227 SV and 1,914 CNV variants identified in PvF217 pistachio genotype. The variant density was calculated as one variant per 292 base. The distribution of the obtained variants to the Siirt reference genome was obtained. In addition, all variants were annotated to the reference genome and exonic and genomic variants were described using Annovar. These data will be used to consist of a molecular database in pistachio breeding for DNA fingerprinting, discovering unique cultivar specific alleles and to identify quantitative trait loci related to important nut traits.

Keywords

Pistachio, Resequencing, Genome, SNP

Introduction

Pistachio (*Pistacia vera* L.) takes place within the genus *Pistacia* and Anacardiaceae family. The genus *Pistacia* included at least 11 species (Kafkas, 2006) and pistachio is the only with economic value. *P. vera* is dioecious flower structure and it pollinates with wind with some exceptions (Kafkas et al. 2000). Its ploidy level is diploid and haploid chromosome number is n=15 (Kafkas, 2019). Pistachio has originated in central Asia, later spread from its origin to Mediterranean region of Southern Europe, Middle east, North Africa, China, United States and Australia (Kafkas, 2019).

The primary problems in pistachio breeding are long juvenile period, dioecious blooming nature and alternate bearing event (Gündeşli et al., 2019). The description of the traits and development of loci associated with traits in pistachio breeding are time-consuming processes such as the characterization of genetic resources and evaluation of segregation population. However, development of markers linked with pest and disease resistance, nut quality traits, alternate bearing and yield in pistachio were not performed any studies (Gündeşli et al., 2021). Overcoming these limitation factors in breeding programs can be only possible molecular approaches using next generation technologies (NGS) (Kafkas, 2019).

Although, simple sequence repeats (SSRs) markers have been utilized in different molecular characterization studies of different species such as apricot, apple, pear, pistachio and quince (Hormaza, 2002; Potts et al., 2012; Fan et al., 2013; Zaloglu et al., 2015; Guney et al., 2019), they are not adequate for marker assisted selection in breeding programs. Until now, different molecular marker systems have been improved and utilized in genetic characterization and mapping of pistachio (Kafkas et al., 2006; Motalebipour et al., 2016, Khodaiminjan et al., 2018; Karcı et al., 2020; Karcı et al., 2022 (unpublished)). However, only

several molecular markers based on SNP developed by (Kafkas et al., 2015, Khodaeiminjan et al., 2017) have been used efficiently for marker assisted selection (MAS) according to sexes in early stage of plant development in pistachio, to date (Kafkas et al., 2017).

Although, there are many genetic characterization studies based on different molecular markers such as random amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), inter simple sequence repeat (ISSR), and SSR in pistachio (Kafkas et al., 2006; Zaloglu et al., 2015; Motalebipour et al., 2016; Karcı et al., 2020), there is no markers associated with pest or disease resistance, yield and nut quality characters in pistachio exception of sex markers developed using RAD-seq technology (Kafkas et al., 2015). Recently, Kafkas et al. (2022) studied over the pistachio sex chromosome and they identified ZW sex choromosome using whole genome resequencing (WGR) technology. About 12.6 Mb W specific female genomic region were identified in chromosome 14. The researchers also represented that high chromosomal assembled of female Siirt cv. and male Bagyolu cv. genomes.

The main objectives of the current study, to consist of the workflow of the variant detection on the resequencing data and database of the pistachio such as SNP array for future revealing genes or QTL regions associated complex and important pistachio traits.

Materials and Methods

Plant Material and DNA Extraction

Pistachio genotype (PvF217) was used in this study and pistachio leaf sample was collected from Çukurova University, at the Research and Experimental area of Agriculture Faculty in Adana province of Turkey.

Total genomic DNA was isolated from fresh young leaves by the CTAB method described by Doyle and Doyle (1987) with some modifications (Kafkas et al., 2006b). Qubit Fluorometer (Invitrogen) was used to quantify the isolated DNAs, followed by diluting them to 10 ng/µl for SSR-PCR reactions, and then the samples were stored at -20 ºC for further analysis.

Resequencing and Variant Calling Analysis

The resequencing was performed in Illumina Hi-seq 2500 and genome sequencing coverage is 15x. The clean pair end (PE) reads were obtained once low quality and adaptors cleaned with Trimmomatic (Bolger et al., 2014). The clean reads obtained from each genotype will be mapped to the reference genome of Siirt cultivar with the Bowtie2 (Langmead and Salzberg, 2012) program mem option. After the clean reads are aligned to reference Siirt genome (Kafkas et al., 2022; unpublished) using SAMtools (Li et al., 2009) view option, duplicate reads were eliminated with Picard tools. SNP and Indel loci were determined using the GATK (Genome Analysis Toolkit) program HaplotypeCallerSpark option with hard filtering option (QD < 2.0, QUAL < 30.0, SOR > 3.0, FS > 60.0, MQRankSum < -12.5, ReadPosRankSum < -8.0) (McKenna et al., 2010). Structure variant and copy number variation analysis was performed using Delly program with call and cnv options (Tobias et al., 2012). The detected variants were annotated to Siirt genome gff3 file for getting the information of variants using the Annovar program (Wang and Hakonarson, 2010).

Results and Discussion

Mapping and removing PCR duplicates

Approximately 12,5 Gb raw PE reads were generated from PvF217 pistachio genotype in Illumina using Hi-seq 2500. A total of 279,718 (0.34%) low quality reads and 5,588,708 (6.72%) polluted reads were removed and a total of 11,578,693,200 clean reads were obtained with 92.85%. A total of 77,102,620 of 77,191,288 clean reads were mapped to reference Siirt genome and the mapping rate was computed as 99.89 % in samtools flagstat option. The mapped reads were filtered according to mapping quality rate (MQ=30) and the rest of the reads were detected as 54,525,620 reads.

The remaining mapped reads were still PCR duplicates that reasons to detect false positive variants on the genomes. Thus, Picard tools marked the 5,537,304 (7.18%) PCR duplicates. As a result, a total of 48,988,316 mapped reads were ready for variant detection (Table 1).

Table 1. The raw, clean, low-quality, polluted, mapped, duplicates and remaining reads numbers and percentage of pistachio genotype

Reads	Numbers
Raw Bases Number	12,470,888,100
Clean Bases Number	11,578,693,200
Low-quality Reads Number	279,718
Low-quality Reads Rate(%)	0.34
Adapter Polluted Reads Number	5,588,708
Adapter Polluted Reads Rate(%)	6.72
Mapped Reads Number	77,102,620
Mapped Reads Rate (%)	99.89
Remaining Reads Number	54,525,620
Remaining Reads Rate (%)	70.64
Duplicates	5,537,304
Duplicates Rate (%)	99.08
Remaining Reads Number	48,988,316
Remaining Reads Rate (%)	876.56

Detection of variants and distribution of the variants through the pistachio genome

Variant detection was carried out using two different variant calling program in linux terminal. Firstly, SNP, InDel were detected using GATK, SV and CNV bam files were utilized for detection variants. A total of 1,785,235 SNP and 260,683 InDel were mined in bam format file of PvF217 pistachio genotype using HaplotypeCallerSpark command in linux bash script. The distribution of the SNP and InDel variants were identified according to chromosomes and scaffolds of the pistachio genotype. Although, the most abundance SNP variants were detected on chr13, the least chromosome density in SNP loci was on chr2. The chromosome 13 has the most abundance InDel variants with 22,250 loci, and the scaffolds have the least resolution InDel loci (Table 2, Figure 1).

The SV and CNV variants were calculated as 5,227 and 1,914, respectively. The scaffolds and chr13 were identified to have high density genomic regions in both SV and CNV loci, while the least abundance variants were located on chr2 (Table 2, Figure 1).

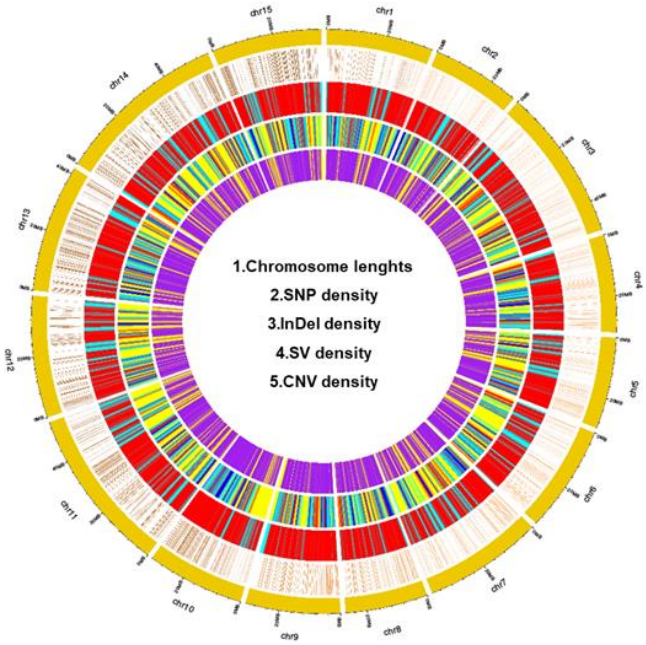

Figure 1. The chromosome lengths, SNP, InDel, SV and CNV variants (from outside to inside) distribution of the pistachio PvF217 genotype

The SV variants were classified according to types such as deletion (DEL), duplication (DUP), insertion (INS), invertion (INV) and translocation (TRA) and the number of these detected loci were calculated as 2,908, 595, 247, 418 and 1,059, respectively (Figure 2).

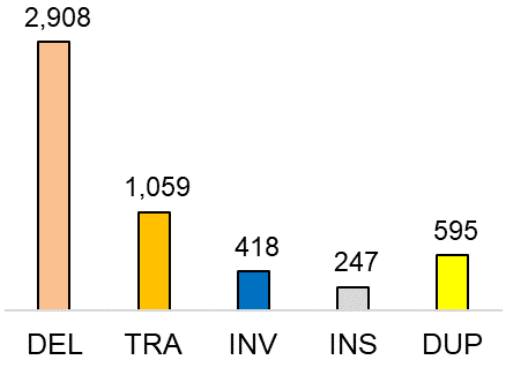

Figure 2. The types of structural variants of the PvF217 pistachio genotype

Genome annotation of variants

The genome annotation was performed using Perl script in Annovar program. Totally, 1,785,235 SNP loci annotated to Siirt reference genome (Kafkas et al., 2022). The most variants were identified on intergenic genomic regions and the number of these variants were computed as 1,335,662. Similarly, the most abundance variants in InDel, SV and CNV were detected as intergenic variants 172,415, 3,103 and 1,511, respectively (Table 3).

Although a total of 73,957 SNP variants were determined on the genic regions, only 4,987 InDel variants were found in genic regions such as frameshift, nonframeshift, stopgain, stoploss (Figure 3).

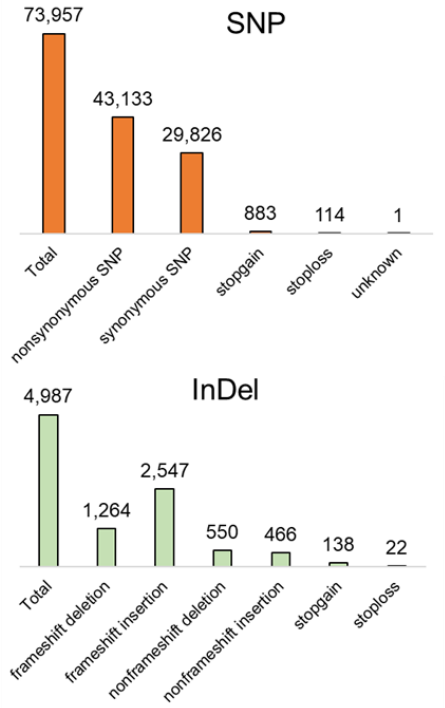

Figure 3. The number of the exonic SNP and InDel variants in detected of whole genome of PvF217

Table 2. The SNP, InDel, SV and CNV variants distribution of the PvF217 whole genome

Chromosomes	SNP	InDel	SV	CNV
chr1	132,979	20,555	300	65
chr2	76,409	12,442	194	48
chr3	131,382	20,136	292	121
chr4	113,709	16,355	260	80
chr5	90,033	13,618	207	74
chr6	96,404	15,427	235	81
chr7	126,309	16,848	357	105
chr8	96,685	14,900	353	66
chr9	100,351	15,927	254	79
chr10	103,469	16,377	332	92
chr11	129,906	18,754	327	100
chr12	116,307	14,831	309	130
chr13	179,803	22,250	574	180
chr14	107,718	16,322	298	104
chr15	99,599	14,043	350	133
scaffold	84,172	11,898	585	456
Total	**1,785,235**	**260,683**	**5,227**	**1,914**

Table 3. The pistachio genotype (PvF217) genome annotation results belonging to SNP, InDel, SV and CNV variants

Genomic regions	SNP	InDel	SV	CNV
UTR5	8,136	2,770	34	4
UTR3	11,559	2,845	21	10
UTR5;UTR3	20	9	1	0
exonic	73,957	4,987	911	59
splicing	571	202	11	2
exonic;splicing	3	1	0	0
upstream	91,591	20,337	337	100
downstream	82,569	17,164	323	86
upstream;downstream	11,356	2,728	54	8
intronic	169,788	37,221	432	134
intergenic	1,335,662	172,415	3,103	1,511
ncRNA_exonic	23	4	0	0
Total	**1,785,235**	**260,683**	**5,227**	**1,914**

The loci detected on the genic chromosome play important roles in the construction of the phenotypes of pistachio. The obtained results demonstrated that these variants can be associated with complex and governing from polygenes traits in pistachio.

There are some the limitation factors of the pistachio breeding such as dioecious character, long juvenile period and alternate bearing (Gündeşli, 2020a,b). The construction of the pistachio germplasm database is required in order to encountered like these breeding problems. However, preference of sequencing platforms is very important for large scale database in pistachio. Because, restriction-site associated DNA sequencing (RAD-seq), diversity arrays technology sequencing (DarT-seq) and GBS (Genotyping by Sequencing) utilize the restriction enzymes and the variants were identified by comparing restricted genome fragments between individuals. The cut genomic fragments cannot be used in other studies. However, resequencing NGS data can be used in order to generate SNP array for genome wide association studies and understood more deeply the complex agronomical important traits in pistachio. Recently, many resequencing findings reported in cucumber (Liu at al., 2021), camelina (Li et al., 2021) and hemp (Ren et al., 2021). On the other hand, the sex regions can be determined using resequencing data in QTL-seq that rapid identification of the sex regions in guinea (Tamiru et al., 2017).

To date, there are no studies related with development of the markers for marker assisted selections in pistachio breeding exception of sex markers developed by Kafkas et al., (2015) and Khodaieminjan et al., (2017). Thus, the obtained results and NGS data in this study can be used for future marker assisted breeding programs in order to develop unique and rare alleles for cultivar fingerprinting; detect the markers associated with nut quality traits, pest and disease resistance, phenological traits, morphological traits in pistachio.

Conclusions

In the present study, 15x resequencing data of pistachio genotype were analyzed in linux terminal using different variant calling program. A robust circus plot was produced and distribution of the SNP, InDel, SV and CNV variants were illustrated in through whole genome pistachio. A total of 2,053,059 variants were detected and 79,914 variants were identified in exonic regions. The resequencing data allowed the allelic variations that can be applied for identifying genes useful to pistachio breeding programs. The consisted of the workflow variants detection can be applied for other cultivars and genotypes. The presented data will be useful in cultivar fingerprinting, germplasm characterization, phylogenetic studies, association and QTL mapping studies in pistachio.

Compliance with Ethical Standards

Conflict of interest

The authors declared that for this research article, they have no actual, potential or perceived conflict of interest.

Author contribution

The contribution of the authors to the present study is equal. All the authors read and approved the final manuscript. All the authors verify that the Text, Figures, and Tables are original and that they have not been published before.

Ethical approval

Not applicable.

Funding

This study was financially supported by The Scientific and Technological Research Council of Turkey (Project No: TUBITAK 118O938)

Data availability

Not applicable.

Consent for publication

Not applicable.

Acknowledgements

The authors thank The Scientific and Technological Research Council of Turkey for financial support.
This article was generated from a PhD thesis.

References

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120. https://doi.org/10.1093/bioinformatics/btu170

Doyle, J. J., & Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue (No. RESEARCH).

Fan, L., Zhang, M. Y., Liu, Q. Z., Li, L. T., Song, Y., Wang, L. F., ... & Wu, J. (2013). Transferability of newly developed pear SSR markers to other Rosaceae species. *Plant Molecular Biology Reporter*, *31*(6), 1271-1282. https://doi.org/10.1007/s11105-013-0586-z

Gündeşli, M. A., Kafkas, S., Zarifikhosroshahi, M., & Kafkas, N. E. (2019). Role of endogenous polyamines in the alternate bearing phenomenon in pistachio. *Turkish Journal of Agriculture and Forestry*, *43*(3), 265-274. https://dx.doi.org/10.3906/tar-1807-74

Gündeşli, M.A. (2020a). Endogenous Gibberellins and Abscisic acid-metabolites: their role for flower bud abscission and embryo development in pistachio. *Turkish Journal of Agriculture and Forestry*. 44(3), 290-300. doi:10.3906/tar-1910-46

Gündeşli, M.A. (2020b). Determination of Sugar contents, Total Phenol and Antioxidant Activity of various parts 'Uzun' pistachio cultivar (*Pistacia vera* L.). *International Journal of Agriculture Environment and Food Sciences*, 4(1), 52-58. https://doi.org/10.31015/jaefs.2020.1.8

Güney, M., Kafkas, S., Keles, H., Aras, S., & Ercişli, S. (2018). Characterization of hawthorn (Crataegus spp.) genotypes by SSR markers. *Physiology and Molecular Biology of Plants*, *24*(6), 1221-1230. https://doi.org/10.1007/s12298-018-0604-6

Hormaza, J. I. (2002). Molecular characterization and similarity relationships among apricot (Prunus armeniaca L.) genotypes using simple sequence repeats. *Theoretical and Applied Genetics*, *104*(2), 321-328. https://doi.org/10.1007/s001220100684

Kafkas, S. (2019) Advances in breeding of pistachio. Chapter. Burleigh Dodds Science Publishing Limited. Doi.10.19103/AS.2018.0042.17

Kafkas, S. (2006). Phylogenetic analysis of the genus Pistacia by AFLP markers. *Plant Systematics and Evolution*, *262*(1), 113-124. https://doi.org/10.1007/s00606-006-0460-7

Kafkas, S., Ozkan, H., Ak, B. E., Acar, I., Atli, H. S., & Koyuncu, S. (2006). Detecting DNA polymorphism and genetic diversity in a wide pistachio germplasm: Comparison of AFLP, ISSR, and RAPD markers. *Journal of the American Society for Horticultural Science*, *131*(4), 522-529. https://doi.org/10.21273/JASHS.131.4.522

Kafkas, S., (2022) The pistachio genomes provide ınsights into nut tree domestication and zw sex chromosome evolution. (Unpublished).

Kafkas, S., Gozel, H., Karcı, H., Bozkurt, H., Paizila, A., Topçu, H., ... & Uzun, M. (2017, November). Marker-assisted cultivar breeding in pistachio. In *VII International Symposium on Almonds and Pistachios 1219* (pp. 63-66). 10.17660/ActaHortic.2018.1219.11

Kafkas, S., Khodaeiaminjan, M., Güney, M., & Kafkas, E. (2015). Identification of sex-linked SNP markers using RAD sequencing suggests ZW/ZZ sex determination in Pistacia vera L. *BMC genomics*, *16*(1), 1-11. https://doi.org/10.1186/s12864-015-1326-6

Kafkas, S., Perl-Treves, R., & Kaska, N. (2000). Unusual Pistacia atlantica Desf.(Anacardiaceae) monoecious sex type in the Yunt Mountains of the Manisa Province of Turkey. *Israel Journal of Plant Sciences*, *48*(4), 277-280. doi: 10.1560/UFCU-7LF6-T0A3-UXWY

Karcı, H., Paizila, A., Güney, M., Zhaanbaev, M., & Kafkas, S. (2022). Revealing Genetic Diversity, Population Structure and Cultivar-Specific SSR Alleles in Pistachio Using SSR Markers. (Unpublished).

Karcı, H., Paizila, A., Topçu, H., Ilikçioğlu, E., & Kafkas, S. (2020). Transcriptome Sequencing and Development of Novel Genic SSR Markers From Pistacia vera L. Frontiers in Genetics, 1021. https://doi.org/10.3389/fgene.2020.01021

Khodaeiaminjan, M., Kafkas, E., Güney, M., & Kafkas, S. (2017). Development and linkage mapping of novel sex-linked markers for marker-assisted cultivar breeding in pistachio (Pistacia vera L.). *Molecular Breeding*, *37*(8), 1-9. https://doi.org/10.1007/s11032-017-0705-x

Khodaeiaminjan, M., Kafkas, S., Motalebipour, E. Z., & Coban, N. (2018). In silico polymorphic novel SSR marker development and the first SSR-based genetic linkage map in pistachio. *Tree Genetics & Genomes*, *14*(4), 1-14. https://doi.org/10.1007/s11295-018-1259-8

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), 357-359. https://doi.org/10.1038/nmeth.1923

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Li, H., Hu, X., Lovell, J. T., Grabowski, P. P., Mamidi, S., Chen, C., ... & Lu, C. (2021). Genetic dissection of natural variation in oilseed traits of camelina by whole-genome resequencing and QTL mapping. *The plant genome*, *14*(2), e20110. https://doi.org/10.1002/tpg2.20110

Liu, X., Gu, X., Lu, H., Liu, P., Miao, H., Bai, Y., & Zhang, S. (2021). Identification of novel loci and candidate genes for resistance to powdery mildew in a resequenced cucumber germplasm. *Genes*, *12*(4), 584. https://doi.org/10.3390/genes12040584

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, *20*(9), 1297-1303. http://www.genome.org/cgi/doi/10.1101/gr.107524.110.

Potts, S. M., Han, Y., Khan, M. A., Kushad, M. M., Rayburn, A. L., & Korban, S. S. (2012). Genetic diversity and characterization of a core collection of Malus germplasm using simple sequence repeats (SSRs). *Plant Molecular Biology Reporter*, *30*(4), 827-837. https://doi.org/10.1007/s11105-011-0399-x

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, *28*(18), i333-i339. https://doi.org/10.1093/bioinformatics/bts378

Ren, G., Zhang, X., Li, Y., Ridout, K., Serrano-Serrano, M. L., Yang, Y., ... & Fumagalli, L. (2021). Large-scale whole-genome resequencing unravels the domestication history of Cannabis sativa. *Science advances*, *7*(29), eabg2286. DOI: 10.1126/sciadv.abg2286

Tamiru, M., Natsume, S., Takagi, H., White, B., Yaegashi, H., Shimizu, M., ... & Terauchi, R. (2017). Genome sequencing of the staple food crop white Guinea yam enables the development of a molecular marker for sex determination. *BMC biology*, *15*(1), 1-20. https://doi.org/10.1186/s12915-017-0419-x

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, *38*(16), e164-e164. https://doi.org/10.1093/nar/gkq603

Zaloğlu, S., Kafkas, S., Doğan, Y., & Güney, M. (2015). Development and characterization of SSR markers from pistachio (Pistacia vera L.) and their transferability to eight Pistacia species. *Scientia Horticulturae*, *189*, 94-103. http://dx.doi.org/10.1016/j.scienta.2015.04.006

Ziya Motalebipour, E., Kafkas, S., Khodaeiaminjan, M., Çoban, N., & Gözel, H. (2016). Genome survey of pistachio (Pistacia vera L.) by next generation sequencing: development of novel SSR markers and genetic diversity in Pistacia species. *BMC genomics*, *17*(1), 1-14. https://doi.org/10.1186/s12864-016-3359-x