NEU PRESS

Ereğli Tarım Bilimleri Dergisi

Volume: 4 Issue: 1 Year: 2024 Research Article e-ISSN: 2822-4167

Impact of Mark Duplicate Reads During Variant Calling in Next Generation Sequencing (NGS) Data of *Pistacia vera* L.

Harun KARCI[1*] Salih KAFKAS[1]

[1] Çukurova Üniversitesi, Ziraat Fakültesi, Bahçe Bitkileri Bölümü, Adana, Türkiye

Article Info

Received: 03.01.2024
Accepted: 08.03.2024
Published: 28.06.2024

Keywords:
Duplicates,
Variant calling,
NGS,
Pistachio,
Genetic.

ABSTRACT

Pistachio (*Pistacia vera* L.) is a member of Anacardiceae family and the only cultural form of Pistacia species. *P. vera* is a dioecious species and there are a few hermaphrodite and monoecious flower nature within Pistacia. Breeding of the pistachio is quite a long process due to several limiting factors such as dioecious flower habitat, quite long juvenile period and alternate bearing. Recently, pistachio genomes have been released with chromosomal level and genome size was about 600 Mb. In the current paper, Next Generation Sequencing (NGS) data of Siirt cultivar has been analyzed to detect the impact of the ignoring duplicates during variant calling stage. About 5.2 Gb data was utilized for detection of the short InDels and SNPs. The highest mapping rate was exhibited with 99.83% and about 35 million reads was aligned successfully reference map. Mapping quality and read coverage depth filtering were carried out MQ>30 and DP>2, respectively. Totally, 7.18% of the reads represented duplicate reads (2.5 million reads). BAM file without MarkDuplicates (MD) was generated a total of 1,022,161 SNPs and 124,762 InDels, BAM file with MD produced a total of 1,050,788 SNPs and 128,109 InDels. Each VCF files were compared according to positions. Same and different (reference allele same but different alternate alleles) variants at the same positions were recorded separately. In addition, BAM file passing MD stage in variant calling were caused the loss of a total of 42,413 true negative loci (TNL) and the getting of a total 10,439 false positive loci (FPL). Therefore, MD is a significant phase of the variant calling all of the organisms and should be carried out to eliminate of false positive loci. The results of the present study can be beneficial for detection of the variants in the next breeding programs.

Antepfıstığı Yeni Nesil Sekans Verilerinde Varyant Belirleme Aşamasında Duplike Okumaları Belirlemenin Etkisi

Makale Bilgisi

Geliş Tarihi: 03.01.2024
Kabul Tarihi: 08.03.2024
Yayın Tarihi: 28.06.2024

Anahtar Kelimeler:
Duplikeler,
Varyant belirleme,
NGS,
Antepfıstığı,
Genetik.

ÖZET

Antepfıstığı (*Pistacia vera* L.), Anacardiceae familyasının bir üyesi olup Pistacia cinsinin tek kültürel formudur. *P. vera* dioik bitki türlerinden biridir ve Pistacia cinsi içeridinde az sayıda da olsa hermafrodit ve monoik çiçek yapısına sahip genotipler bulunmuştur. Antepfıstığının ıslahı, dioik çiçek yapısı, gençlik döneminin oldukça uzun olması ve periyodisite göstermesi gibi birçok sınırlayıcı faktörden dolayı oldukça uzun sürmektedir. Yakın zamanda kromozom düzeyinde antepfıstığı genomları yayınlanmış ve genom büyüklüğü yaklaşık 600 Mb olarak bildirilmiştir. Bu çalışmada, Siirt çeşidine ait Yeni Nesil Dizileme (NGS) verileri kullanılarak varyant belirleme aşamasında duplike okumaları belirlemenin etkisini tespit etmek amacıyla birçok biyoinformatik analiz yapılmıştır. Kısa InDel'ler ve SNP'lerin tespiti için yaklaşık 5.2 Gb veri kullanılmıştır. Yüzde 99.83 haritalama oranı belirlenmiş ve yaklaşık 35 milyon okuma referans genoma başarıyla haritalanmıştır. Haritalama kalitesi ve okuma derinliği filtrelemeleri sırasıyla MQ>30 ve DP>2 olarak gerçekleştirilmiştir. Tüm okumaların yüzde 7.18'i duplike okumaları (2.50 milyon) oluşturmuştur. MarkDuplicates (MD) işlemi uygulanmayan BAM dosyası kullanılarak yapılan varyant belirleme analizinde toplam 1,022,161 SNP ve 124,762 InDel, MD işlemi gerçekleştirilen BAM dosyası kullanılarak yapılan varyant belirleme analizi sonrasında ise toplam 1,050,788 SNP ve 128,109 InDels tespit edilmiştir. Her iki VCF dosyası genom pozisyonlarına göre karşılaştırılmıştır. Aynı pozisyonlardaki aynı ve farklı (referans aleli aynı ancak farklı alternatif aleller) varyantlar saydırılmış ve ayrı ayrı kaydedilmiştir. Ayrıca MD işlemi yapılmayan BAM dosyasının varyant belirleme işlemlerinde, toplam 42,413 adet gerçek negatif lokus (TNL) kaybına ve toplam 10,439 adet yanlış pozitif lokus (FPL) elde edilmesine neden olmuştur. Sonuç olarak, MD tüm organizmalar için varyant belirleme analizlerinin önemli bir aşaması olmakla birlikte yanlış pozitif lokusların ortadan kaldırılması için de gerçekleştirilmelidir. Elde edilen bu sonuçların, gelecekte yapılacak ıslah programlarında varyant tespitinde faydalı olabileceği düşünülmektedir.

To cite this article:
Karcı, H. & Kafkas, S. (2024). Impact of mark duplicate reads during variant calling in next generation sequencing (NGS) data of *Pistacia vera* L. *Eregli Journal of Agricultural Sciences, 4*(1), 11-18. https://doi.org/10.54498/ETBD.2024.29

***Corresponding author:** Harun KARCI, *karciharun42@gmail.com*

INTRODUCTION

Pistachio takes place in the Anacardiceae family like sumac and cashew species, and the botanic name of the pistachio is *Pistacia vera* L. (Karcı *et al.,* 2022). Pistachio is one of the hard-shelled nuts and this species has an allogamy pollination nature like other several nuts such as walnut, hazelnut and chestnut excluding almond (Vahdati *et al.,* 2021). It has a heterozygous genetic habitat due to its dioecious character.

The highest pistachio producer countries were the USA, Iran and Turkiye according to the Food and Agricultural Organization of the United Nations (FAO) in 2021 (FAOSTAT, 2023). Turkiye has carried out 239,289 t (tons) pistachio in 2022, while 119,355 t production was recorded by Turkish Statistical Institute in 2021.

In pistachio, there are several main limiting factors in breeding such as dioecious flower structure, quite a long juvenile period and alternate bearing (Kafkas *et al.,* 2006). Despite such limiting effects, there was no breeding program with controlled hybridization in pistachio in Turkiye, although many attempts were performed so far. In the last years, the first marker assisted breeding program in pistachio was initiated by Kafkas *et al.* (2017). A huge number of male and female plants constructed by a total of 56 combinations were sown to breeding field in this program. Approximately 10,000 male plants were eliminated with RAD-seq sex markers detected by Kafkas *et al.* (2015). This team also released two chromosomal sequencing of the pistachio male and female genomes, and they have described three genomic invertions into sex chromosome (chr14) using a total of 225 resequencing data belonging to cultivars and genotypes (Kafkas *et al.,* 2023). The researchers stated that many markers associated with agronomical traits can be developed using whole genome of the pistachio, and they would be beneficial for marker assisted selection in pistachio breeding program to facilitate the time of the selection using NGS technology. However, NGS is highly sensitive to technological errors, and the processing of NGS data has become dependent on reliable bioinformatics tools. Thus, the novel bioinformatics and biostatistics approaches are necessary for overcoming the limiting traits in plants.

Next-generation sequencing (NGS) has contributed to detect responsible genes related to diseases, differentially expression genes, reveal the evolution of the species or genus with phylogenetic analysis, perform quantitative trait loci (QTLs) and genome wide association studies (GWAS) analysis with SNP, InDel and structural variants (duplications, translocation, long insertions/deletions, invertions, copy number variants) and release *de novo* sequencing of minor crops (Houston *et al.,* 2012; Vrijenhoek *et al.,* 2015). On the other hand, there are several challenging to study in NGS data such as different genotyping methods, sequencing errors, the lengths of the reads, huge amount of storage problem and novel innovative bioinformatics softwares (Chen *et al.,* 2014). In addition, the sequenced duplicate reads during the PCR amplification of the sequencing cause false positive variants pretending to be aligned to the reference genome at a higher rate. Whereas these reads were due to duplicate reads carried out multiple times DNA fragments.

A few popular and reliable programs such as SAMTools (Li *et al.,* 2009) and Picard (http://broadinstitute.github.io/picard/) remove and mark the duplicates in the BAM file, respectively. SAMTools removes the PCR duplicates by identifying the paired end duplicate reads. It does not recognize the duplicates except the paired end reads, and also does not notice reads mapped to different chromosomes. Picard has MarkDuplicates option and does not remove the duplicates, it just works on the paired reads like SAMTools. Comparing the SAMTools, Picard has a higher potential than SAMTools by determination of the interchromosomal paired reads.

Here, we carried out the variant calling using the Picard MarkDuplicates option and without this option. The comparison of the short indels and SNPs were recorded and discussed with number of loci

each situation to reveal the effects of the marking duplicates in the course of calling.

MATERIALS AND METHODS

Material and Detection of the Variants

The clean reads (150 bp) of Siirt cv. were downloaded from NCBI (National Center for Biotechnology Information) with PRJNA680201 ID number (Kafkas *et al.,* 2023). The sequencing coverage was calculated as 10x when about 600 Mb genome size was considered.

The reads mapped to the *P. vera* reference genome using BWA (v0.7.12) (Li and Durbin, 2009). Then, the Sequence Alignment/Map (SAM/BAM) files were sorted and filtered the quality of the mapping (Q30) using the SAMTOOLS (Li *et al.,* 2009) software. The duplicated reads of the Siirt cultivar were marked using the PICARD software (http://picard.sourceforge.net). Then, read coverage depth was set DP>2 using the Bcftools in both VCF files.

Bcftools v1.9 with 'mpileup and call' commands was used for the list of the variants format (VCFFILE). Short InDels and SNPs were listed from BAM files. The marked duplicates and without marked duplicates vcf files were compared using VCFtools program according to positions. In addition, the genomic positions belonging to both vcf files were extracted for vcf files. Both SNPs and InDels specific to vcf file applied MD and without MD callings were recorded separately. Based on without MD call was considered as "false positive variants", and missing ones were evaluated "true negative variants" using the in-house scripts. These variants were born from without MD call due to misalignment of reads or sequence errors.

The constructed workflow for the variant detection was designed into bash script and final outputs were evaluated in Linux operating system.

RESULTS AND DISCUSSIONS

Whole Genome Re-sequencing and Mapping Stats

Approximately 34.93 million reads data was utilized for detection of the short InDels and SNPs by mapping the reference of the pistachio genome (Kafkas *et al.,* 2023). The used NGS data had been presented with findings of the pistachio genomes with PRJNA680201 NCBI number. The lengths of reads of paired-end was 150 bp and all of clean reads were aligned to reference sequences with a high percentage (99.83%). A total of 34.87 million reads were filtered according to mapping quality threshold (MQ>30). The remaining paired reads was saved as 25.94 million reads used for marking the duplicate reads using option MarkDuplicates (MD) of Picard program. Before the comparison, the first step was initiated with variant calling from 25.94 million reads without MD. Then, the same BAM file included duplicates was processed to mark duplicates into sequences using Picard software. The marked duplicates were 2.50 million reads and it was consisted of the 7.18% of the mapped sequences. Second variant calling with MD was performed using rest of the 23.44 million reads. The comparison of the variants was enjoyed with MD applied BAM and without MD BAM file, and results were detailed in calling section. Karcı and Kafkas (2022) investigated that genetic structure of the pistachio genotype using the whole genome resequencing data and they stated that 77.1 million reads were mapped to reference genome.

Detection of SNPs and InDels and comparison of the MD and without MD VCFs

In the current paper, variant calling was carried out in the pistachio cultivar Siirt using two different approaches to evaluate the impact of the mark duplicates. Firstly, BAM file was just applied Q30 mapping quality filter generated a total of 1,022,161 SNPs and 124,762 InDels, while the normal

process with MD produced a total of 1,050,788 SNPs and 128,109 InDels (Table 1). Although there was a significant difference of the variant numbers, the wave of the mutations was considered by misalignment of the duplicate reads or errors of the sequencing. Since there was an option for difference of the counts of SNP and InDels. In addition, the highest number of SNPs and InDels were determined from Chr11 with 102,580 and 11,803, respectively.

Table 1 *Number of SNP and InDel Variants Detected in VCFs Filtered According to MQ>30 + MarkDuplicates and Only MQ>30*

Chrs	Total			
	MQ>30 + MarkDuplicates		Only MQ>30	
	SNPs	InDels	SNPs	InDels
Chr1	79,017	10,452	77,117	10,213
Chr2	51,513	6,942	50,154	6,784
Chr3	88,479	10,836	86,008	10,558
Chr4	71,064	8,267	69,209	8,036
Chr5	54,372	6,591	52,912	6,409
Chr6	64,071	8,252	62,511	8,062
Chr7	64,250	7,691	62,360	7,454
Chr8	60,092	7,283	58,637	7,124
Chr9	75,654	9,411	73,900	9,215
Chr10	64,944	8,218	63,168	8,019
Chr11	102,580	11,803	99,638	11,522
Chr12	62,855	6,848	60,872	6,669
Chr13	71,966	8,526	69,710	8,111
Chr14	83,222	10,403	80,715	10,158
Chr15	56,709	6,586	55,250	6,428

Each VCF file was compared based on variant positions and position-specific variants were saved in other VCFs. Outputs of the comparison generated two type files such as the same variants at the same positions, the different variants (reference allele same but different alternate alleles) at the same positions and all of these variants were described as common variants at the same positions in both VCFs. The number of variants of chromosomes of Siirt cultivar was given in Table 2. In the present study, the same variants at the same positions were pretty reliable mutations, on the contrary of the different variants at the same positions. Based on misalignment of the duplicates, a lot of false positive loci (FPL) were defined, and the most abundant FPL was created from Chr11 with four SNPs and 157 InDels. On the other hand, true negative loci (TNL) were missed owing to misalignment of the duplicates like FPL. Therefore, it not only introduces the FPL, but also obstructs the TNL. The different variants at the same positions Chr1 were controlled using IGV program and its picture was shared in Figure 1. Especially, it was estimated that such variants were corrected by marking duplicate reads and true variants were identified.

Table 2 *Number of SNPs and InDels Mined from Common Variants at the Same Positions in Both VCFs*

Chrs	No. of common variants at the same positions in both VCFs					
	No. of the same variants at the same positions		No. of the different variants at the same positions		Total (Same+Different)	
	SNPs	InDels	SNPs	InDels	SNPs	InDels
Chr1	76,477	9,883	5	136	76,482	10,019
Chr2	49,660	6,569	5	98	49,665	6,667
Chr3	85,384	10,233	2	149	85,386	10,382
Chr4	68,711	7,782	1	125	68,712	7,907
Chr5	52,462	6,193	2	99	52,464	6,292
Chr6	62,046	7,807	2	103	62,048	7,910
Chr7	61,797	7,210	5	100	61,802	7,310
Chr8	58,243	6,919	1	83	58,244	7,002
Chr9	73,443	8,942	1	122	73,444	9,064
Chr10	62,689	7,788	6	92	62,695	7,880
Chr11	98,763	11,153	4	157	98,767	11,310
Chr12	60,373	6,495	6	61	60,379	6,556
Chr13	69,085	7,861	9	112	69,094	7,973
Chr14	79,960	9,840	1	143	79,961	9,983
Chr15	54,784	6,203	5	94	54,789	6,297

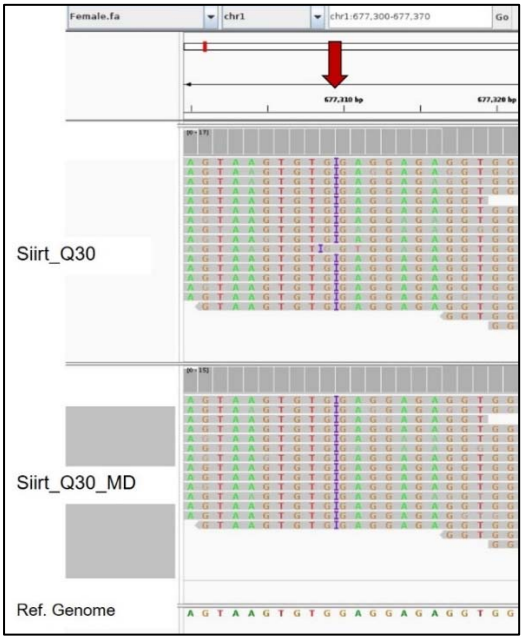

Figure 1 *Mark Duplicates before and after Illustration of the Siirt BAM Files Locus Specific Position in IGV Program*

Mark duplicates option is quite important for getting the TNL and eliminating the FPL. In the

current finding, BAM file passing MD stage in variant calling were caused the loss of a total of 42,413 TNL and the getting of a total 10,439 FPL (Table 3). A total of 36,856 SNPs and 5,557 InDels were TNL lost without MD, while the highest false positive SNPs and InDels were calculated as 871 and 212, respectively. Loss of the TNL was found higher than FPL and the most abundant true negative SNP loci was 3,813. Therefore, the impact of the MD was exhibited on the pistachio NGS data, which means that true negative and false positive loci may be associated with polygenic complex agricultural important traits and the loss of a number of loci was based on only bioinformatic or sequencing errors in this phase. In caution, the standard workflow of bioinformatics should be followed without throwing any steps.

Table 3 *Number of TNL and FPL of Siirt Cultivar*

Chrs	Absent in without MD and Present in MD (Loss of the TNL)		Absent in MD and Present in without MD (Get of the FPL)	
	SNPs	**InDels**	**SNPs**	**InDels**
Chr1	2,535	433	635	194
Chr2	1,848	275	489	117
Chr3	3,093	454	622	176
Chr4	2,352	360	497	129
Chr5	1,908	299	448	117
Chr6	2,023	342	463	152
Chr7	2,448	381	558	144
Chr8	1,848	281	393	122
Chr9	2,210	347	456	151
Chr10	2,249	338	473	139
Chr11	3,813	493	871	212
Chr12	2,476	292	493	113
Chr13	2,872	553	616	138
Chr14	3,261	420	754	175
Chr15	1,920	289	461	131

PCR duplicate elimination constitutes a crucial step in nearly all next-generation sequencing (NGS) variant calling pipelines. This process demands significant time and memory resources, leading to the elimination of varying proportions of reads. The neglecting or discarding PCR duplicates entails overlooking a portion of the generated sequence data. The predominant tools for PCR duplicate marking/removal are Picard and SAMTools and a formal comparison between these two algorithms was performed by Elbert *et al.* (2016). The researchers aimed to detect program potential for removing or ignoring duplicates utilizing Picard versus SAMTools. Although Picard had a pretty potential by censoring duplicates matching inter-chromosomal reads, it has several disadvantages owing to memory requirements and time. Nevertheless, this is not an insoluble reason not to prefer it.

In the realm of next-generation sequencing (NGS) variant calling pipelines, the elimination of PCR duplicates is a widely endorsed practice (Palmer *et al.,* 2023; Kafkas *et al.,* 2023; Karcı and Kafkas, 2022; Zhang *et al.,* 2023). This operation is characterized by its resource-intensive nature in terms of both time and memory (Elbert *et al.,* 2023). The primary algorithm employed for PCR duplicate removal is Picard in terms of its higher potential in working inter-chromosomal regions. Moreover, there exists limited data assessing the necessity of PCR duplicate removal. The main aim of the current paper was

to reveal the meaningful impact of PCR duplicate marking on resultant variant datasets and to discern any false positive in accuracy with Picard.

In another word, eliminating duplicate reads can notably decrease the count of mapped high-quality reads, consequently reducing the average depth of coverage. However, removing duplicates can also reduce read coverage depth, it is one of another possible effect of read depth filtering (Sims *et al.,* 2014).

Consequently, novel biotechnological and bioinformatic approaches and programs are being developed in plant breeding, but the verification of the results of such innovations takes a long time for plant species. Developed novel approaches have different detection methods to identify genes or genomic fragments associated with important agronomic features of plant species. To date, a lot of universal softwares have been developed and remain up to date for the determination of SNP, InDel, SV and CNV loci. However, this process extends to many different populations of each plant species, depending on the flowers, fruits/nuts, trees, leaves and the climatic conditions in which they are cultivated. In order to shorten this period, it is necessary to carry out some comparisons or trials. A long with this study, obtained results confirmed about the necessity of marking/removing duplicate reads.

Author Contribution

Research Design (CRediT 1) Harun Karcı (75%) – Salih Kafkas (25%)
Data Collection (CRediT 2) Harun Karcı (50%) – Salih Kafkas (50%)
Research - Data Analysis - Validation (CRediT 3-4-6-11) Harun Karcı (100%)
Writing the Article (CREDITS 12-13) Harun Karcı (100
Editing and Development of the Text (CRediT 14) Harun Karcı (90%) – Salih Kafkas (10%)
Sample preparation of pistachio and bioinformatic analysis were carried out in Çukurova University, Faculty of Agriculture, Department of Horticulture, Molecular and Bioinformatic Lab.

Funding

This research received no external funding.

Conflict of Interest

The authors declare that they have no conflict of interest.

Sustainable Development Goals

Does not support

REFERENCES

Chen, C., Khaleel, S. S., Huang, H. & Wu, C. H. (2014). Software for pre-processing Illumina next-generation sequencing short read sequences. *Source code for biology and medicine*, *9*(1), 1-11.

Ebbert, M. T., Wadsworth, M. E., Staley, L. A., Hoyt, K. L., Pickett, B., Miller, J. & Ridge, P. G. (2016). Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC bioinformatics*, *17*, 491-500.

Faostat (2023). Statistic Database. http://faostat.fao.org/ (accessed 23.11.19).

Houston, D. D., Elzinga, D. B., Maughan, P. J., Smith, S. M., Kauwe, J. S., Evans, R. P. & Shiozawa, D. K. (2012). Single nucleotide polymorphism discovery in cutthroat trout subspecies using genome reduction, barcoding, and 454 pyro-sequencing. *BMC genomics*, *13*, 1-17.

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).

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.

Kafkas, S., Ozkan, H., Ak, B., Acar, I., Atli, H. & 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).

Karcı, H., Paizila, A., Güney, M., Zhaanbaev, M. & Kafkas, S. (2022). Revealing genetic diversity and population structure in Pistachio (Pistacia vera L.) by SSR markers. *Genetic Resources and Crop Evolution*, *69*(8), 2875-2887.

Karci, H. & Kafkas, S. (2022). Evaluation of genetic structure of pistachio through whole genome resequencing. *International Journal of Agriculture Environment and Food Sciences*, *6*(1), 135-140.

Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics*, *25*(14), 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079.

Palmer, W., Jacygrad, E., Sagayaradj, S., Cavanaugh, K., Han, R., Bertier, L. & Michelmore, R. (2023). Genome assembly and association tests identify interacting loci associated with vigor, precocity, and sex in interspecific pistachio rootstocks. *G3*, *13*(2), jkac317.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, *15*(2), 121-132.

Vahdati, K., Sarikhani, S., Arab, M. M., Leslie, C. A., Dandekar, A. M., Aletà, N. & Mehlenbacher, S. A. (2021). Advances in rootstock breeding of nut trees: objectives and strategies. *Plants*, *10*(11), 2234.

Vrijenhoek, T., Kraaijeveld, K., Elferink, M., De Ligt, J., Kranendonk, E., Santen, G. & Cuppen, E. (2015). Next-generation sequencing-based genome diagnostics across clinical genetics centers: implementation choices and their effects. *European Journal of Human Genetics*, *23*(9), 1142-1150.

Zhang, D. Y., Liu, X. M., Huang, W. J., Wang, Y., Anwarullah, K., Luo, L. F. & Gao, Z. X. (2023). Whole-genome resequencing reveals genetic diversity and signatures of selection in mono-female grass carp (Ctenopharyngodon idella). Aquaculture, 739816.