Unlocking the Genetic Diversity of Morus indica through Molecular Marker Analysis

Sinsha Vikhin[1], Vidya Niranjan[1*], Lavanya C[1], Spoorthi R Kulkarni[1],

Author 1 Sinsha Vikhin, Department of Biotechnology, RV College of Engineering, Bangalore, Karnataka, India), Visveswaraya Technological University, Belagavi, 590018

Corresponding Author- Vidya Niranjan (vidya.n@rvce.edu.in) Professor and Head of Department of Biotechnology, RV College of Engineering, Bangalore, Karnataka, India, Visveswaraya Technological University, Belagavi, 590018*

Author 3 Lavanya C, Department of Biotechnology, RV College of Engineering, Bangalore, Karnataka, India., Visveswaraya Technological University, Belagavi, 590018

Author 4 Spoorthi R Kulkarni, Department of Biotechnology, RV College of Engineering, Bangalore, Karnataka, India. Visveswaraya Technological University, Belagavi, 590018

Abstract

Mulberry (Morus spp.) is a dioecious and cross-pollinated plant species with significant genetic diversity, yet its potential for genetic improvement has been limited by a lack of specific molecular markers and genetic maps. This manuscript explores the importance of SSRs (Simple Sequence Repeats) and SNPs (Single Nucleotide Polymorphisms) as essential molecular markers for understanding genetic diversity, population structure, and traits of interest within the Morus genus.

The study focuses on Morus indica (mulberry), a versatile plant species crucial to the silk industry in India. Through a combination of genetic linkage mapping techniques, phenotypic characterization, and genotypic analysis, we aim to accomplish several key objectives: mapping genetic linkage, identifying SNPs and SSRs, and developing a specialized microchip for Morus indica.

Our research not only promises to shed light on the genetic basis of observed phenotypes but also offers insights into the genetic diversity and adaptations of Morus indica. The outcomes of this study have broad implications for crop improvement, conservation, and breeding programs, ultimately contributing to the sustainable development of the mulberry industry in India and beyond. By harnessing the power of molecular markers and advanced techniques, we aim to unlock the potential of mulberry for various applications in agriculture and industry.

Key words:

Mulberry, Molecular markers, Single Nucleotide Polymorphism, Simple Sequence Repeats

1.0 INTRODUCTION

1.1 Genetic diversity in Mulberry

The botanical group Morus, commonly referred to as mulberry, is a plant characterized by being dioecious and reliant on cross-pollination. Notably, it serves as the exclusive source of sustenance for the domesticated silkworm, Bombyx mori [1]. The development of mulberry, a tree eudicot plant, is hindered by insufficient knowledge about genetic variation, ploidy-related characteristics, and the relationship between genetic makeup and observable traits. This limits the potential for enhancing mulberry through genetic selection, despite its broad adaptability across the globe and its status as a cross-pollinated species with significant genetic diversity [2]. Mulberry plants exhibit a wide range of ploidy levels, ranging from having a single set of chromosomes (haploid) with 14 chromosomes to having 22 sets of chromosomes (docosaploid) or 308 total chromosomes. In contrast, certain Morus species are diploid, with two sets of chromosomes totalling 28 chromosomes [3]. Genetic diversity in polyploid mulberry plants pertains to the assortment of genetic traits and characteristics found in individuals or populations of mulberry species with multiple sets of chromosomes [4]. Polyploidy refers to the state in which an organism possesses more than two sets of chromosomes, which can arise from different genetic processes [5]. Evaluating genetic diversity in these polyploid mulberry plants involves studying the variations in their genetic composition, encompassing differences in genes, alleles, and other molecular markers. This diversity can arise from different sources, such as mutations, recombination events during reproduction, and historical genetic interactions. It can impact the overall adaptability, resilience, and potential for selective breeding of these plants. The examination of genetic relationships within the mulberry plant holds significant importance within crop enhancement initiatives, and molecular markers are the prevailing tools for comprehending disparities among various Morus species. Employed the Random Amplified Polymorphic DNA (RAPD) technique to illustrate variations among distinct Morus species. Subsequently, the analysis of genetic diversity among mulberry plants cultivated in India utilized Inter-Simple Sequence Repeats (ISSR) [6]. However, the assessment of mulberry's genetic diversity encounters certain limitations, as previously indicated. These include the scarcity of adequate molecular markers specific to mulberry and the absence of codominant markers essential for constructing a genetic map. The latter is a notable obstacle for employing marker-assisted breeding to enhance specific traits [7].

1.2 The Significance of SSR and SNPs in molecular studies

SSR (Simple Sequence Repeats) and SNPs (Single Nucleotide Polymorphisms) are crucial molecular markers utilized in the study of mulberry and other organisms. They play significant roles in understanding genetic diversity, population structure, evolutionary relationships, and traits of interest. Molecular markers have emerged as a pivotal tool. Notably, they have enabled the streamlined analysis of plant varieties and species. In particular, SSRs (Simple Sequence Repeats) stand out due to their numerous advantages. They exhibit high levels of polymorphism, ensuring robust results. Furthermore, SSRs can be automated, requiring only small quantities of DNA [8]. Both SSRs and SNPs are invaluable in mulberry molecular studies. SSRs are well-suited for assessing genetic diversity, population structure, and parentage analysis, while SNPs offer high-resolution insights into genome organization, trait associations, and marker-assisted breeding. Integrating these markers enhances our understanding of mulberry genetics and aids in its improvement and conservation [9].

1.3 The relevance of techniques in *Morus indica*

The application of genetic linkage mapping techniques in Morus indica is highly relevant and offers numerous advantages for crop improvement and the sustainable development of the mulberry industry in India.

Our research aims to accomplish several key objectives: mapping genetic linkage, pinpointing SNPs and SSRs, and developing a specialized microchip for Morus indica (mulberry) based on the differentiation of four distinct phenotypes observed across various geographic areas. We've conducted a comprehensive characterization of these phenotypes, relying on their observable traits for classification. We intend to analyse each sample from the MINDGP database and categorize it into one of the four phenotypes based on its genotypic characteristics [10]. To establish genetic linkage, we've scrutinized the segregation patterns of genetic markers, typically SNPs, with the goal of charting these markers to specific genomic regions. This approach will allow us to identify the genome regions associated with each of the observed phenotypes. Moreover, we've identified SNPs and SSRs distributed throughout the Morus indica genome [11-12]. We've filtered these genetic markers, focusing on those that exhibit uniqueness or a significant correlation with the observed phenotypes. With the list of these distinctive and associated SNPs and SSRs in hand, we are poised to craft a

specialized genotyping microchip or array. This microchip will facilitate the accurate identification of genetic markers relevant to our study. This research can provide a deeper understanding of the genetic diversity and adaptations of Morus indica and may have implications for agriculture, conservation, and breeding programs focused on this species.

2.0 METHODOLOGY

2.1 Retrieval of raw mulberry data

The high-depth sequencing of multiple mulberry accessions via Illumina platform was performed by Jawaharlal University. A total of 21 genotypes were sequenced and deposited in the Morus Indica Genome project [13].

Once the datasets were retrieved, the contigs deposited in the Morus Indica Genome Project were scaffolded and used as the Reference genome.

2.2 Raw data pre-processing

Prior to performing further analysis on the cancer exome reads retrieved, a pre-processing of the raw data was carried out.

2.3 Quality checks of the reads

To determine the quality of the raw data retrieved, a quality assessment was carried out. The quality assessment and pre-processing of raw data was performed using FastQC. FastQC provides a very efficient and a simple way to perform quality control checks on the raw sequence data that arise from high throughput sequencing pipelines and provides a linked set of analyses that can be used to comprehend if the raw data has any problems that must be addressed prior to performing any further analysis [14]. Therefore, for this purpose, FastQC was employed. Quality checks for all 4 Genotypes namely S1, Punjab local, Thailand male, Assama bola of mulberry were performed. This included examining the mean sequence quality per base position, per read, nucleotide content per base position, GC distribution, etc. The exome datasets after performing quality checks were then taken for further analyses.

Raw data pre-processing by FastQC also included the detection of over-represented sequences, which may implicate adaptor or primer contamination. The comprehensive QC report of the raw reads generated by FastQC aids in the determination of the necessity of other pre- processing steps such as base trimming, adaptor clipping or filtering of reads, before alignment.

2.4 Gapped alignment

To map the pre-processed reads to the reference genome with high precision and efficiency, alignment mapping, or gapped alignment of the reads was carried out for all the twenty-cancer exome reads. For this purpose, there are currently several tools meant for the mapping of short reads, which employ the Burrows-Wheeler Transformation (BWT) compression techniques along with the Smith waterman dynamic programming algorithm. BWA (Burrows-Wheeler Aligner) was once such a tool that aligns the short reads and implements the BWT algorithm. In the current study, BWA software package was used for mapping the short reads against a larger reference genome, in this case, the mulberry genome. It utilizes three algorithms for this purpose: BWA-backtrack, BWA-MEM and BWA-SW. BWA discovers an alignment by default within an edit distance 2 to the query sequence, excepting the disallowed gaps that are nearer to the end of the query. It can further be refined and adjusted to find longer gaps, however, at the cost of false alignments and speed. The main purpose for employing BWA for alignment with the reference genome was that it outshines in speed [15-16].

This tool also works on the FM index, which is based off BWT technique and supports local, gapped as well as paired-end alignment to the reference genome. In the current study, the output was obtained in SAM format (Sequence Alignment Map), allowing cross-operation with a variety of other tools such as GATK. This is considered as the first step in comparative genomics studies, which follows the variant calling as the subsequent steps.

2.5 Refining the alignments and conversion by SAM tools

To improve the quality of the alignments and reduce potential false variant calls, the alignment output from the previous step was taken through several refining steps prior to variant calling. Currently, there exists a large number of tools meant for the processing of SAM/BAM files obtained as outputs of alignment to the reference genome. The SAM tools [17], Genome Analysis Toolkit (GATK) [18] and Picard [19] are software by Broad and Sanger Institutes and are employed commonly for operations on SAM/BAM files. In the current study, the cancer exome short reads for all twenty datasets were refined after aligning them with the reference genome. Some of the operations performed on them were sorting the reads, recalibrating the quality scores, marking the PCR and optical duplicates, realignment of the indels as well as filtering of the reads.

A software package for manipulating and analyzing the alignment data present in SAM/BAM format and can convert from other formats of alignments, merge and sort them, erase and remove the PCR duplicates, call the SNPs and other short variants, produce per- position information and display the alignments in a viewer which is text based [20]. The SAM tools toolkit was utilized for conversion of the short-read alignment output from SAM format to BAM. BAM sorting was then carried out followed by BAM sorting and merging. SAM is considered as a generic format of alignment, which is quite easy and simple to work with and versatile to keep most of the information from diverse sequencing platforms. BAM, an equivalent binary representation, has a more compact size and supports the rapid retrieval of alignments in specific regions. By employing positional indexing, sorting and merging, processing on genomic regions without having to load the whole file into the memory is possible. The SAM/BAM format divides the alignment step from the other downstream analyses, permitting an integrated approach to the scrutiny and analysis of genome sequence data. Therefore, the present study employed SAM tools for the conversion of SAM to BAM.

2.6 Variant processing

The preliminary steps for calling of variants include indexing of the reference genome using BWA and SAMtools, assessing the quality of the sequences using FASTQC, removing the adaptors using Trim galore and rechecking the sequence quality prior to merging the BAM alignments. Variant calling and processing are carried out to identify the variants or mutations from the sequenced data.

In the present study, variant processing was first carried out using PICARD and GATK. The Genome Analysis Toolkit (GATK) is a structured framework which is designed to simplify the development of robust and efficient tools of NGS DNA analyses. This tool offers an opulent set of patterns for data access that overarch a majority of analyses, a tool requires. Separating the calculations from the mainstream data management infrastructure allows the user to appropriately optimize the framework of GATK for precision, memory efficiency and stability. Furthermore, GATK allows for the application of tools such as coverage calculators as well as calling of single nucleotide polymorphisms. In the current study, GATK was utilized for removal of PCR duplicates as it enables the users and developers to swiftly and efficiently write NGS tools, the majority of which have been included into huge sequencing projects such as The Cancer Genome Atlas and 1000 Genome Projects [21].

Likewise, PICARD is a collection of command line tools that is commonly utilized for manipulating and modulate the high throughput sequencing data as well as formats such as SAM/BAM/CRAM and VCF, whose file formats are properly defined in the HTS-specs repository [22]. The current study marked the PCR duplicates using PICARD and local re-alignment and base quality recalibration using GATK using appropriate parameters. SortSam, recalibrating the bases and building BAM indexes were also carried out using PICARD. Once the co-variates were analyzed, the variants were called.

2.7 Variant calling

Once the variants were processed, the identified variants were called. Prior to this, the pileup of mapped data and counting the bases were carried out using SAMtools. To identify and extract the SNPs and indels, the variants were then called. For this purpose, GATK was utilized. The total variants were identified, and SNPs were called. The SNPs and indels were extracted first and then filtered. The filtered SNPs and indels were then annotated using a tool called snpEFF [23]. SnpEFF is a tool that annotates the variants/mutations and predicts the effects of the identified genetic variants, including the amino acid changes. A typical snpEFF input includes the predicted variants such as indels, SNPs, insertions or deletions and is in variant call format (VCF). Once the tool analyses, annotates and predicts the effects on known genes, it gives out an output in its default VCF format. In the present study, a thorough analysis of the identified SNPs was carried out and the output obtained was scrutinized. SnpEFF accepts single VCF files and supports over 2500 genomes including plant, animal, mammalian, bacterial and fungal genomes. The advantage of using snpEFF for the purpose of variant calling and annotation is that it is fast and can easily annotate 10,00,000 mutations per minute [24-25]. Another perk of using this tool is that it is easily incorporated with GATK and Galaxy toolkits. Therefore, the present study employed this tool for the purpose of annotating the variants.

2.8 Variant Filtration

Understanding the link between genotype and phenotype is a key scientific challenge, and the ability to anticipate phenotypes based on molecular genotypes is critical in molecular breeding. Whole genome duplications have affected the history of all flowering plants and pose difficulties in determining the link between genotype and phenotype, particularly in polyploid species. Although single nucleotide polymorphisms (SNPs) have become popular tools for genetic mapping, finding and using SNPs in polyploids has proven challenging. We

called SNPs across five peanut genotypes using alternative alignment programme (BWA-mem and Bowtie 2) and variant callers (SAMtools, GATK, and Freebayes) to assess the influence of software choice on these findings. Bowtie 2 and BWA-mem alignments analyzed in SAMtools shared 24.5% concordant SNPs, while SAMtools, GATK, and Freebayes shared 1.4% concordant SNPs. A further investigation of simulated Brassica napus chromosome 1A and 1C genotypes revealed that SAMtools scored the best on Bowtie 2 alignments among the three-software programme. These findings are expected to differ amongst species; hence we suggest a set of best practices for SNP calling in polyploids.

2.9 Identification of SSR Markers

Data Preparation

The genomic data utilized in this study were obtained in FASTA format, providing the foundational sequences for the organism under investigation. The GMATA software [26] was subsequently installed on the local computer system to facilitate the analysis.

Running GMATA

To initiate the SSR identification process, a terminal or command prompt was accessed on the computer. The terminal was navigated to the directory where the GMATA software was installed, ensuring that the tool was readily accessible for analysis.

Identifying SSRs

The cornerstone of SSR identification within GMATA is the MicroSAtellite identification tool (MISA) module. This algorithm systematically detects SSRs within the genomic sequences based on user-defined parameters. Specifically, repeat unit length and the minimum number of repeats were customized to suit the requirements of our investigation.

3.0 RESULTS AND DISCUSSION

3.1 Quality check and preprocessing

The preliminary pre-processing of raw data was carried out successfully for all datasets using FastQC. The overall outcomes showed that per base sequence quality of most of the datasets were represented in green post FastQC check, indicating that the mulberry datasets passed the initial checks.

3.2 Gapped alignment and refining the alignments

The outputs from Bowtie2 generated the alignment in SAM format, allowing inter-operation amongst a large number of tools such as GATK and SAMTools which also implement SAM. Furthermore, the summary files obtained from the software showed that there were no false alignments found. Once the alignments were successfully performed, they were further refined to reduce potential false variant calls and converted from SAM to BAM format using SAMtools. The converted outputs obtained were then found to be in an easy to work format and were used for variant calling.

3.3 Variant Calling

The reads that were aligned was processes for obtaining the variants in vcf format. The details of the variants are shown in the table below.

Table 1 Variant details that are processed

Dataset ID	Genotype name	Total variants	Variants processed
SRR14506990	S1	3,218,628	3,205,581
SRR14506999	Punjab Local	4,658,600	4,624,224
SRR14507001	Assama Bola	3,411,732	3,395,052
SRR14507002	Thailand Male	3,139,443	3,125,916

3.4 Identification of SSR markers

The Morus Indica samples were subjected to Annotation and the potential genes were identified. A total of 75,000 genes were reported and is shown in the link below.

https://docs.google.com/spreadsheets/d/1VUP60MLGJlDg-TKwqoc02Ylp6VwS_Zzb/edit?usp=sharing&ouid=114613627565011743972&rtpof=true&sd=true

The gene wise SSR markers were identified and the results are shown in the Link below.

https://drive.google.com/file/d/1TtKYzqG7-Te63ohZRBcHCtt7izpWUmug/view?usp=sharing

The figure below describes the top distribution of SSR length.

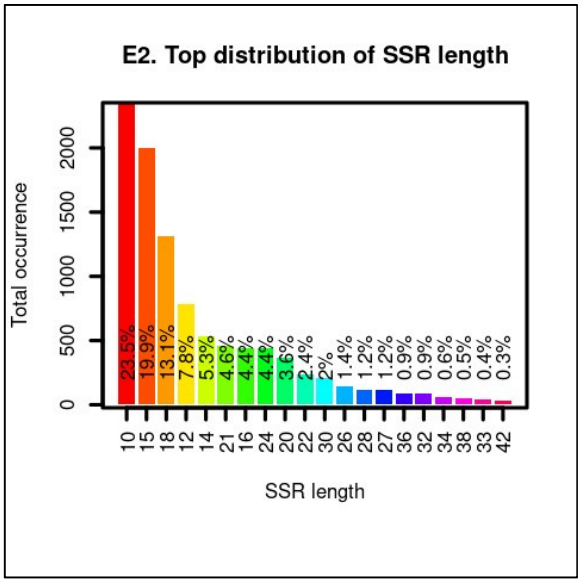

Figure 1 Figure indicating the SSR Length in the Genes of Morus indica

CONCLUSION

In conclusion, the study of Morus, commonly known as mulberry, holds significant importance in the context of crop enhancement, genetic diversity assessment, and the sustainable development of the mulberry industry. Mulberry is a dioecious, cross-pollinated plant with substantial genetic diversity, including various ploidy levels. However, the lack of specific molecular markers and genetic maps has hindered efforts to enhance this versatile plant through selective breeding.

To address these challenges, researchers have turned to molecular markers such as SSRs (Simple Sequence Repeats) and SNPs (Single Nucleotide Polymorphisms). These markers offer valuable tools for studying genetic diversity, population structure, and evolutionary

relationships within Morus species. SSRs, in particular, are known for their polymorphic nature and automation potential, making them a powerful asset in mulberry genetic studies.

The relevance of techniques such as genetic linkage mapping in Morus indica cannot be overstated. By conducting a comprehensive analysis of phenotypes and genotypic characteristics, researchers aim to establish genetic linkage, pinpoint relevant SNPs and SSRs, and develop a specialized microchip. This approach promises to shed light on the genetic basis of observed phenotypes and provide insights into the genetic diversity and adaptations of Morus indica.

Ultimately, the research outlined here has broad implications for the mulberry industry in India and beyond. It can inform strategies for crop improvement, conservation, and breeding programs, leading to the sustainable development of this valuable plant species. By leveraging molecular markers and advanced techniques, we can unlock the potential of mulberry for various applications and ensure its continued contribution to agriculture and industry.

ACKNOWLEDGEMENT

The authors would like to thank Mr. Akshay Uttarkar, Research Associate I, Department of biotechnology, RV College of Engineering for providing valuable suggestions while drafting the article.

CONFLICT OF INTEREST

The authors declare there is no conflict of interest

REFERENCES

1) Awasthi, A.K., Nagaraja, G., Naik, G. et al. Genetic diversity and relationships in mulberry (genus Morus) as revealed by RAPD and ISSR marker assays. BMC Genet 5, 1 (2004). https://doi.org/10.1186/1471-2156-5-1
2) Gnanesh BN, Mondal R, G S A, H B M, Singh P, M R B, P S, Burji SM, T M, V S. Genome size, genetic diversity, and phenotypic variability imply the effect of genetic variation instead of ploidy on trait plasticity in the cross-pollinated tree species of mulberry. PLoS One. 2023 Aug 11;18(8):e0289766. doi: 10.1371/journal.pone.0289766.
3) Nderitu Peris Wangari*, Kinyua Miriam Gacheri, Mutui Mwendwa Theophilus and Ngode Lucas, Use of SSR markers for genetic diversity studies in mulberry

accessions grown in Kenya, International Journal for Biotechnology and Molecular Biology Research, 2013. Doi: : 10.5897/IJBMBR11.057

4) Gnanesh, B. N., Mondal, R., G S, A., H B, M., Singh, P., M R, B., P, S., Burji, S. M., T, M., & V, S. (2023). Genome size, genetic diversity, and phenotypic variability imply the effect of genetic variation instead of ploidy on trait plasticity in the cross-pollinated tree species of mulberry. *PloS one*, *18*(8), e0289766. https://doi.org/10.1371/journal.pone.0289766
5) Madlung, A. Polyploidy and its effect on evolutionary success: old questions revisited with new tools. *Heredity* 110, 99–104 (2013). https://doi.org/10.1038/hdy.2012.79
6) Vijayan, K., Awasthi, A. K., Srivastava, P. P., & Saratchandra, B. (2004). Genetic analysis of Indian mulberry varieties through molecular markers. *Hereditas*, *141*(1), 8–14. https://doi.org/10.1111/j.1601-5223.2004.01813.x
7) Wang, Z. Y., & Tanksley, S. D. (1989). Restriction fragment length polymorphism in Oryza sativa L. *Genome*, *32*(6), 1113-1118.
8) Park, Y. J., Lee, J. K., & Kim, N. S. (2009). Simple sequence repeat polymorphisms (SSRPs) for evaluation of molecular diversity and germplasm classification of minor crops. *Molecules (Basel, Switzerland)*, *14*(11), 4546–4569. https://doi.org/10.3390/molecules14114546
9) Ahmad, A.; Wang, J.-D.; Pan, Y.-B.; Sharif, R.; Gao, S.-J. Development and Use of Simple Sequence Repeats (SSRs) Markers for Sugarcane Breeding and Genetic Studies. *Agronomy* 2018, *8*, 260. https://doi.org/10.3390/agronomy8110260
10) Jain, M., Bansal, J., Rajkumar, M. S., Sharma, N., Khurana, J. P., & Khurana, P. (2022). Draft genome sequence of Indian mulberry (Morus indica) provides a resource for functional and translational genomics. *Genomics*, *114*(3), 110346. https://doi.org/10.1016/j.ygeno.2022.110346
11) Altshuler, D., Daly, M. J., & Lander, E. S. (2008). Genetic mapping in human disease. *Science (New York, N.Y.)*, *322*(5903), 881–888. https://doi.org/10.1126/science.1156409
12) Paterson, A. H., Lander, E. S., Hewitt, J. D., Peterson, S., Lincoln, S. E., & Tanksley, S. D. (1988). Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature*, *335*(6192), 721-726.
13) http://tgsbl.jnu.ac.in/MindGP/
14) https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

15) Youngmok Jung, Dongsu Han, BWA-MEME: BWA-MEM emulated with a machine learning approach, *Bioinformatics*, Volume 38, Issue 9, March 2022, Pages 2404–2413, https://doi.org/10.1093/bioinformatics/btac137

16) Subramaniyan A. et al. (2021) Accelerated seeding for genome sequence alignment with enumerated radix trees. In: *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, pp. 388– 401.

17) Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, Volume 25, Issue 16, August 2009, Pages 2078–2079, https://doi.org/10.1093/bioinformatics/btp352

18) McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

19) Costanza, P., Herzeel, C., & Verachtert, W. (2019). A comparison of three programming languages for a full-fledged next-generation sequencing tool. *BMC bioinformatics*, *20*(1), 301. https://doi.org/10.1186/s12859-019-2903-5

20) Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol10: R25.

21) Koboldt D. C. (2020). Best practices for variant calling in clinical sequencing. *Genome medicine*, *12*(1), 91. https://doi.org/10.1186/s13073-020-00791-w

22) Diouf, I., & Pascual, L. (2021). Multiparental population in crops: methods of development and dissection of genetic traits. *Crop Breeding: Genetic Improvement Methods*, 13-32.

23) Shigemizu, D., Fujimoto, A., Akiyama, S. *et al.* A practical method to detect SNVs and indels from whole genome and exome sequencing data. *Sci Rep* 3, 2161 (2013). https://doi.org/10.1038/srep02161

24) Fuentes-Pardo, A. P., & Ruzzante, D. E. (2017). Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Molecular ecology*, *26*(20), 5369-5406.

25) Agarwala, V., Flannick, J., Sunyaev, S., & Altshuler, D. (2013). Evaluating empirical bounds on complex disease genetic architecture. *Nature Genetics*, **45**, 1418–1427.

26) Wang, X., & Wang, L. (2016). GMATA: an integrated software package for genome-scale SSR mining, marker development and viewing. *Frontiers in plant science*, *7*, 1350.