BMC Genomics

RESEARCH Open Access

A comprehensive catalog of single nucleotide polymorphisms (SNPs) from the black pepper (*Piper nigrum* L.) genome

Hiruni A. Thanthirige[1], Nilni A. Wimalarathna[1] and Anushka M. Wickramasuriya[1*]

Abstract

Background Single nucleotide polymorphisms (SNPs) have emerged as the marker of choice in breeding and genetics, particularly in non-model organisms such as black pepper (*Piper nigrum* L.), a globally recognized spice crop. This study presents a comprehensive catalog of SNPs in the black pepper genome using data from 30 samples obtained from RNA sequencing and restriction site-associated DNA sequencing, retrieved from the Sequence Read Archive, and their consequences at the sequence level.

Results Three SNP calling and filtering pipelines, namely BCFtools, Genome Analysis Toolkit (GATK)-soft filtering, and GATK-hard filtering, were employed. Results revealed 498,128, 396,003, and 312,153 SNPs respectively identified by these pipelines, with 260,026 SNPs commonly detected across all methods. Analysis of SNP distribution across the 45 scaffolds of the black pepper genome showed varying densities, with pseudo-chromosomes Pn25 (0.86 SNPs/kb), Pn8 (0.74 SNPs/kb), and Pn7 (0.72 SNPs/kb) exhibiting the highest densities. Conversely, scaffolds Pn27 to Pn43 exhibited minimal SNP distribution, except Pn45. Approximately 34.80% of SNPs exhibited stronger genetic linkage ($r^2 > 0.7$). Moreover, SNPs predominately mapped to downstream (≈ 32.54%), upstream (≈ 22.52%), and exonic (≈ 16.20%) regions of genes. Transition substitution accounted for the majority (≈ 57.42%) of identified SNPs, resulting in an average transition-to-transversion ratio of 1.36. Notably, 56.09% of SNPs were non-synonymous, with a significant proportion (≈ 53.59%) being missense mutations. Additionally, 12,491 SNPs with high or moderate impacts were identified, particularly in genes associated with secondary metabolism and alkaloid biosynthesis pathways. Furthermore, the expression of 675 genes was potentially influenced by local (cis-acting) SNPs, while 554 genes were affected by distal (trans-acting) SNPs.

Conclusion The findings of the present study underscore the utility of identified SNPs and their targets, especially those impacting important pathways, for future genetic investigations and crop improvement efforts in black pepper. The characterization of SNPs in genes related to secondary metabolism and alkaloid biosynthesis highlights their potential for targeted breeding aimed at enhancing the yield, quality, and resilience of this economically important crop in diverse environmental conditions.

Keywords Black pepper, SNPs, Functional annotation, Synonymous SNPs, Non-synonymous SNPs

*Correspondence:
Anushka M. Wickramasuriya
anushka@pts.cmb.ac.lk
[1]Department of Plant Sciences, Faculty of Science, University of Colombo, Colombo, Sri Lanka

Background

Black pepper (*Piper nigrum* L., 2n = 52), known as the 'King of Spices', is one of the most valuable and widely used spices in the world and is often referred to as 'Black Gold' due to its preeminent status in the global spice trade. Renowned for its characteristic pungency and flavour, primarily due to the alkaloid piperine and volatile oils, black pepper has been a key ingredient in many food preparations for thousands of years [1, 2]. In addition to its culinary uses, black pepper has been used in traditional medicine systems for its potential health benefits as well as in perfumery, as a natural preservation for food, and as an insecticide [3, 4, 5, 6, 7]. These applications are attributed to the various qualities and compounds in black pepper, which offer potential health benefits and functional properties, such as antioxidants, anti-inflammatory, and potential anticancer properties [3, 5, 6].

This perennial woody vine, belonging to the family *Piperaceae*, is believed to have originated from the tropical evergreen forest of the Western Ghats of southern India. From there, black pepper was introduced to other parts of South and Southeast Asia [8]. Today, it is cultivated in many tropical and subtropical regions, including Vietnam, Brazil, Sri Lanka, Malaysia, China, and Indonesia [1, 9]. According to the Food and Agriculture Organization of the United Nations data from 2022, the world's total black pepper harvested area was 689,336 hectares, accounting for a production of 812,674 tons.

Despite being an economically and culturally important crop, the exploitation of its genetic diversity has been relatively limited compared to other crops. Analysis of genetic diversity and population structure of crops is a prerequisite for designing efficient crop breeding and conservation strategies [10, 11]. Previous studies have used molecular markers such as random amplified polymorphic DNA (RAPD [12, 13, 14]), amplified fragment length polymorphism (AFLP [13]), and simple sequence repeats (SSRs [15, 16, 17, 18]) to assess the genetic diversity of black pepper. In recent years, single nucleotide polymorphism (SNP) markers have become widely regarded as the marker of choice for many studies, particularly those focusing on genetic diversity and population structure analysis, marker-trait association studies, marker-assisted breeding, and ecological and evolutionary analyses. This preference is due to their abundance in the genomes, polymorphic nature, and amenability to high-throughput detection platforms [19, 20, 21, 22, 23, 24, 25, 26].

The increasing capacity of next-generation sequencing (NGS) technologies and advances in bioinformatics computing resources have enabled the discovery of SNPs through *de novo* approaches, allowing researchers to identify new genetic variants without prior knowledge of the genome. Furthermore, these technological advances have facilitated the genome-scale discovery of SNPs feasible in various model and nonmodel organisms, including plant species [21]. Several approaches are commonly used to generate sequence data for large-scale detection of variants [27, 28]. Some of these approaches include whole genome sequencing, genotyping-by-sequencing [29], transcriptome sequencing [30, 31, 32, 33], exome sequencing [34], and restriction site-associated DNA sequencing (RAD-seq) [35, 36]. These technologies have been successfully applied in SNP discovery in crops, both with [37] and without reference genome sequences [38].

Over the past years, several software packages have been developed for variant calling from NGS data. Among these, Genome Analysis Toolkit (GATK) HaplotypeCaller [39] and BCFtools mpileup [40] are the most widely utilized variant callers. While many studies have demonstrated GATK's outperformance over BCFtools when analyzing a large number of samples [27, 41, 42, 43], several other studies have reported better performance from BCFtools [44, 45, 46]. In particular, a recent study conducted by Lefouili and Nam in 2022 suggests that BCFtools mpileup may be the first choice over GATK HaplotypeCaller for non-model studies, particularly in insects, as BCFtools mpileup may result in a lower number of false positives than GATK [47]. Additionally, variant callers such as FreeBayes [48], SNVer [49], VarDict [50], and VarScan [51] are also used in studies for the discovery of SNPs.

The availability of a high-quality reference genome for black pepper [52] has enabled scientists to perform comparative genomics and gain deeper insights into the genetic diversity of this valuable crop. However, there remains a lack of information on the frequency, nature, and distribution of SNPs in black pepper. Therefore, this study aimed to address this gap by analyzing SNP marker diversity using publicly available genomic and transcriptomic datasets. The specific objectives were to determine the distribution of SNPs across the black pepper genome, annotate them to assess their potential functional impacts, and exploit gene expression information to identify SNP-gene associations.

Materials and methods

Data collection

For the present study, we used previously published RNA sequencing (RNA-seq) and RAD-seq datasets deposited at the Sequence Read Archive (SRA) database hosted by the National Center for Biotechnology Information (NCBI) [53, 54]. Specifically, we included 14 RAD-seq datasets generated from leaves (BioProject PRJNA1035754) and 16 RNA-seq datasets from various tissues (BioProject PRJNA529760) of black pepper. Detailed information on these datasets is provided in Additional file 1. Additionally, the genome assemblies

and annotation files of the chromosome-scale reference genome of *P. nigrum* [52] were obtained from the Group of Cotton Genetic Improvement (GCGI) at Huazhong Agricultural University [55].

Sequence data preprocessing and mapping

After retrieving data, it is essential to assess the quality of the reads before proceeding into the mapping. The quality of raw sequence reads was assessed using the open-source tool, FastQC v0.2.1 [56] with default parameters. Adapter sequences were trimmed using the Trimmomatic [57]. Following quality-control assessment and trimming, each sample was mapped to the *P. nigrum* reference genome [52] using the Burrows-Wheeler Alignment (BWA [58]). Specifically, we employed the BWM-MEM algorithm, which is an efficient seedling algorithm, with the option -M to flag shorter split hits as secondary. The resulting Sequence Alignment Map (SAM) files were converted to Binary Alignment Map (BAM) format and sorted by genomic coordinates using the *SortSam* module in the Picard toolkit v3.1.1 [59]. Mapping rates were estimated using bamtools [60] to evaluate potential bias in mapping. The *AddOrReplaceReadGroups* module in Picard was employed to modify or add read groups to the sorted BAM files. Finally, these BAM files were indexed using samtools [61].

Variant calling and filtering

In this study, variant calling was performed using two widely used tools: GATK v4.4.0.0 [39] and BCFtools [62]. For the GATK pipeline, the *P. nigrum* reference genome sequence was first indexed using samtools. A sequence dictionary file (.dict) of the reference genome was created using the *CreateSequenceDictionary* module. The *HaplotypeCaller* function was employed to generate a Genomic Variant Call Format (GVCF) file for each sample. These GVCF files were then merged into a single file using the *CombineGVCFs* module. Joint genotyping analysis of all samples was performed with the *GenotypeGVCFs* module, resulting in a Variant Call Format (VCF) file. SNPs were then extracted using the *SelectVariants* module with the following criteria for soft-filtering, hereafter referred to as GATK4 (soft-filtering) pipeline: (i) Phred-scaled P-value for the Fisher's exact test to detect strand bias (FS) > 60.0; (ii) a root mean square of mapping quality across all samples (MQ) < 40.0. Additionally, hard-filtering was performed with the following parameters, hereafter referred to as GATK4 (hard-filtering) pipeline: (i) variant quality by depth (QD) < 2.0; (ii) FS > 60.0; (iii) MQ < 40.0; (iv) strand odds ratio (SOR) > 4.0; (v) U-based z-approximation from the rank sum test for the distance from the end of the reads with the alternate allele (MQRankSum) < -12.5; (vi) U-based z-approximation from the rank sum test for mapping qualities (ReadPosRankSum) < -8.0. Subsequently, VCFtools v0.1.16 [63] was used to retain only SNPs with a minor allele frequency (MAF) of ≥ 0.066, ensuring the allele was present in at least two individuals.

For the BCFtools pipeline, the *mpileup* command was first used to generate a pileup file summarizing the read information at each genomic position. Subsequently, the *call* command was used to call variants/indels from the pileup file, resulting in a VCF file. The *view* command was then used to extract SNPs. The resulting VCF was filtered using the *filter* command with the following criteria: FS > 60.0 and MQ < 40.0. Finally, SNPs with an MAF ≥ 0.066 were retained using VCFtools v0.1.16 for further analysis. Figure 1 provides an overview of the methodology employed in this study for variant calling.

Dendrogram construction

The final filtered VCF files were converted to PHYLIP format using PGDSpider v2.1.1.5 [64] to estimate the relatedness among the samples. The script ascbias.py from the GitHub repository was used to eliminate invariant sites. Maximum likelihood trees were constructed for concatenated SNPs using Randomized Axelerated Maximum Likelihood (RaxML) v8.2.12. The trees were constructed with the general time reversible model of nucleotide substitution with the recommended ascertainment bias correction. The best-scoring trees were visualized using iTOL (Interactive Tree of Life) [65, 66].

Variant annotations

Filtered SNPs discovered by the three pipelines (GATK (soft-filtering), GATK (hard-filtering), and BCFtools) were annotated using SnpEff v5.2 [67]. To annotate SNPs, a custom database was created for black pepper using the following files retrieved from the Hu et al. 2019 [52]: "Piper_nigrum.cds", "Piper_nigrum.pep", "Piper_nigrum.genome.fa", and "Piper_nigrum.gff3". The final VCF files were then annotated using this custom database. Normalized transcript levels of black pepper genes across different tissues and genes associated with alkaloid and secondary metabolism pathways were retrieved from Hu et al. 2019 [52].

Linkage disequilibrium (LD) analysis

To evaluate pairwise LD between SNPs, we calculated pairwise correlation coefficient (r^2) values for bi-allelic SNPs using PLINK (v1.9) [68]. A sliding window approach was applied, specifying a window size of 100 kb (--ld-window-kb 100) and a maximum of 10 SNPs per window (--ld-window 10). The resulting r^2 values were plotted against the physical distances between SNPs (in kilobases) to assess the LD decay. A locally estimated scatterplot smoothing (LOESS) curve was fitted to the data using the geom_smooth function in the R package

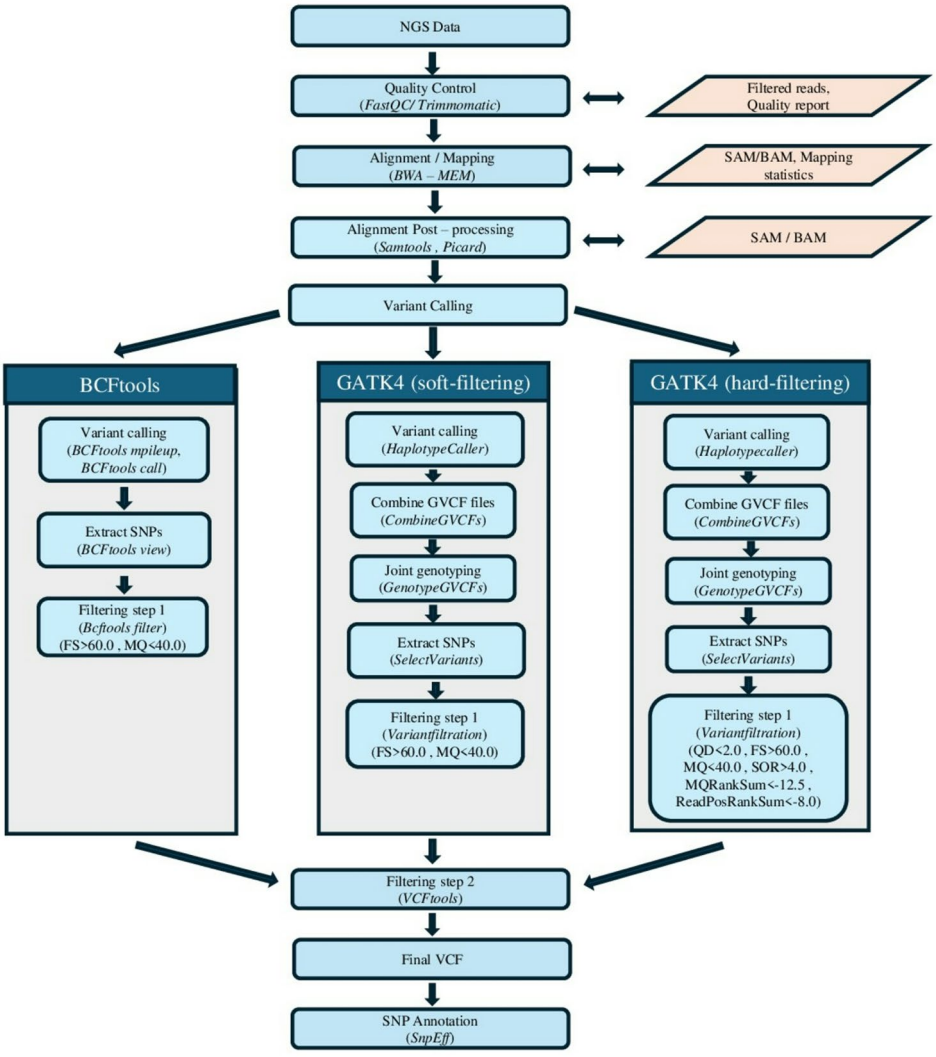

Fig. 1 Overview of the methodology for variant calling in *Piper nigurm*

ggplot2 [69] to visualize the LD decay pattern across the chromosomes. The half-LD decay distance, defined as the physical distance at which the maximum observed average r^2 value decays to 50%, was calculated to estimate the extent of LD decay [70, 71, 72]. In addition, the percentage of variant pairs below the threshold of $r^2 < 0.2$ was calculated for each chromosome using outputs from PLINK.

Expression quantitative trait loci (eQTL) analysis

Normalized gene expression data for various developmental stages of black pepper (SRR8816492, SRR8816486, SRR8816469, SRR8816470, SRR8816488, SRR8816489, SRR8816484, SRR8816485, SRR8816471, SRR8816472, SRR8816474, SRR8816475, SRR8816477, SRR8816478, SRR8816480, and SRR8816482) were retrieved from the MagnoliidsGDB [73, 74]. Genes with nonzero FPKM (Fragments Per Kilobase of transcript per Million mapped reads) values in more than 50% of the accessions with a variance greater than 0.05 were retained for the eQTL analysis.

To select independent SNPs, we pruned SNPs exhibiting high LD using a sliding window of 50 SNPs, a step size of 5 SNPs, and an r^2 threshold of 0.5. eQTL analysis was performed using the Matrix eQTL software (v2.3) [75] to assess the associations between genetic variants and gene expression levels. Both cis- and trans-eQTLs were analysed, with significant thresholds set at 10^{-5} for cis-eQTLs and 10^{-12} for trans-eQTLs. A cis-distance (cisDist) of 10^6 base pairs was used to define cis-eQTLs [76].

Functional annotation of SNP-associated genes was performed using the Functional Analysis module of OmicsBox 3.4 [77, 78]. In brief, protein sequences of target genes were aligned with the NCBI non-redundant protein database using BLASTP. InterProScan was subsequently performed to identify functional protein

domains. Mapping and annotation of Gene Ontology (GO) terms were performed using Blast2GO with default settings (an E-value < 10^{-6} and annotation cut-off = 55). The resulting annotations were classified into three main groups: biological processes, molecular function, and cellular components.

Results

Quality-control assessment and read mapping

The use of high-quality reads is essential for the identification of true variants. In this study, we processed 30 datasets of paired-end sequence reads derived from RNA-seq and RAD-seq. After quality assessment using FastQC and trimming low-quality ends of sequence reads, we retained a total of 1,261,178,896 reads. These reads were then mapped to the black pepper reference genome, yielding mapping rates ranging from 81.84 to 98.81%, with an average mapping rate of 95.06% (Table 1).

Variant calling

For variant calling, we utilized two tools, GATK4 and BCFtools, applying different filtering options. Initially, filtering criteria of FS > 60.0 and MQ < 40.0 were applied, resulting in the retention of 5,219,107 and 6,163,705 SNPs from the GATK4 and BCFtools pipelines, respectively. Subsequent filtering using VCFtools reduced these counts to 396,003 (including 47,152 multi-allelic SNPs) and 498,128 (including 4,668 multi-allelic SNPs) for the GATK4 (soft-filtering) and BCFtools pipelines, respectively. Applying additional filtering with GATK4 (FS > 60.0, MQ < 40.0, QD < 2.0, SOR > 4.0, MQRankSum < -12.5 and ReadPosRankSum < -8.0) yielded 4,355,396 SNPs, which were further filtered down to 312,153 SNPs (including 37,789 multi-allelic SNPs) using VCFtools. A total of 260,026 bi-allelic SNPs were commonly detected across all three pipelines. A comprehensive summary of SNPs across scaffolds Pn1 to Pn45 of the black pepper genome is presented in Table 2.

To further explore the variants captured by RAD-seq and RNA-seq, maximum likelihood-based dendrograms

Table 1 Summary of read mapping statistics

Sequencing method	Sample name	Number of reads	Number of mapped reads	Percentage of mapped reads *(%)
RAD-seq	RADSeq_L1	11,797,966	11,312,242	95.88
	RADSeq_L2	11,576,872	10,746,769	92.83
	RADSeq_L3	11,273,120	9,226,049	81.84
	RADSeq_L4	10,955,876	10,505,325	95.89
	RADSeq_L5	14,283,951	14,113,501	98.81
	RADSeq_L6	13,236,830	12,948,368	97.82
	RADSeq_L7	10,917,273	10,501,409	96.19
	RADSeq_L8	11,736,240	10,194,484	86.86
	RADSeq_L9	16,812,385	15,344,276	91.27
	RADSeq_L10	14,701,799	13,724,218	93.35
	RADSeq_L11	11,176,087	10,571,519	94.59
	RADSeq_L12	10,917,762	10,618,659	97.26
	RADSeq_L13	16,857,697	15,787,988	93.65
	RADSeq_L14	15,268,020	14,496,398	94.95
RNA-seq	RNASeq_R1	72,526,011	69,952,709	96.45
	RNASeq_R2	61,089,624	59,968,012	98.16
	RNASeq_S1	71,972,092	70,011,417	97.28
	RNASeq_S2	61,726,229	60,002,566	97.21
	RNASeq_F1	63,668,082	61,898,613	97.22
	RNASeq_F2	71,737,337	63,457,681	88.46
	RNASeq_L1	78,033,253	76,810,273	98.43
	RNASeq_L2	74,790,187	73,637,708	98.46
	RNASeq_2MAP1	81,879,996	79,395,158	96.97
	RNASeq_2MAP2	69,422,645	66,196,518	95.35
	RNASeq_4MAP1	58,599,249	57,144,443	97.52
	RNASeq_4MAP2	68,899,979	67,155,706	97.47
	RNASeq_6MAP1	51,309,671	49,523,609	96.52
	RNASeq_6MAP2	68,809,951	66,683,083	96.91
	RNASeq_8MAP1	58,110,425	54,850,975	94.39
	RNASeq_8MAP2	67,092,287	62,944,500	93.82

*The mapped percentage is the percentage of reads that were aligned with the black pepper reference genome

Table 2 SNP count per scaffold

Scaffold*	Scaffold length (bp)	Number of SNPs			SNP density (SNP per kb)		
		BCFtools	GATK4 (soft-filtering)	GATK4 (hard-filtering)	BCFtools	GATK4 (soft-filtering)	GATK (hard-filtering)
Pn1	48,451,882	30,058	24,029	18,776	0.62	0.50	0.39
Pn2	43,104,928	28,900	22,970	18,221	0.67	0.53	0.42
Pn3	39,384,597	23,941	18,594	14,889	0.61	0.47	0.38
Pn4	37,189,380	19,524	15,060	11,956	0.52	0.40	0.32
Pn5	36,778,867	28,968	23,112	17,791	0.79	0.63	0.48
Pn6	33,612,360	14,366	11,615	9,061	0.43	0.35	0.27
Pn7	32,737,202	29,134	23,301	18,186	0.89	0.71	0.56
Pn8	32,725,476	29,758	23,922	18,943	0.91	0.73	0.58
Pn9	32,482,239	22,500	18,344	14,259	0.69	0.56	0.44
Pn10	31,231,022	18,203	14,420	11,348	0.58	0.46	0.36
Pn11	29,812,524	19,793	15,345	12,250	0.66	0.51	0.41
Pn12	29,600,758	9,780	7,450	6,115	0.33	0.25	0.21
Pn13	29,427,894	23,902	18,938	14,743	0.81	0.64	0.50
Pn14	29,032,553	17,419	13,906	11,048	0.60	0.48	0.38
Pn15	28,693,427	21,572	17,230	13,774	0.75	0.60	0.48
Pn16	27,948,387	15,911	12,872	10,175	0.57	0.46	0.36
Pn17	25,720,905	15,394	12,027	9,420	0.60	0.47	0.37
Pn18	25,555,373	12,367	9,724	7,668	0.48	0.38	0.30
Pn19	24,808,771	17,505	14,089	11,031	0.71	0.57	0.44
Pn20	24,284,306	16,112	12,720	10,043	0.66	0.52	0.41
Pn21	22,260,265	10,672	8,500	6,610	0.48	0.38	0.30
Pn22	22,167,516	15,629	12,543	9,979	0.71	0.57	0.45
Pn23	20,442,519	12,773	10,164	8,054	0.62	0.50	0.39
Pn24	19,853,247	15,198	12,112	9,575	0.77	0.61	0.48
Pn25	18,224,315	19,164	15,632	12,419	1.05	0.86	0.68
Pn26	14,906,710	9,546	7,360	5,798	0.64	0.49	0.39
Pn27	65,737	-	-	-	-	-	-
Pn28	58,380	-	-	-	-	-	-
Pn29	57,069	-	-	-	-	-	-
Pn30	47,017	5	2	2	0.11	0.04	0.04
Pn31	45,114	-	-	-	-	-	-
Pn32	45,072	-	-	-	-	-	-
Pn33	43,961	-	-	-	-	-	-
Pn34	40,824	5	2	1	0.12	0.05	0.02
Pn35	38,901	7	2	2	0.18	0.05	0.05
Pn36	38,354	-	-	-	-	-	-
Pn37	37,066	-	-	-	-	-	-
Pn38	36,761	-	-	-	-	-	-
Pn39	35,667	-	-	-	-	-	-
Pn40	34,574	-	-	-	-	-	-
Pn41	32,332	-	-	-	-	-	-
Pn42	31,851	-	1	1	-	0.03	0.03
Pn43	30,876	-	-	-	-	-	-
Pn44	30,737	-	-	-	-	-	-
Pn45	30,593	22	17	15	0.72	0.56	0.49
Total		**498,128**	**396,003**	**312,153**			

*Pn1-Pn26 represent pseudo-chromosomes

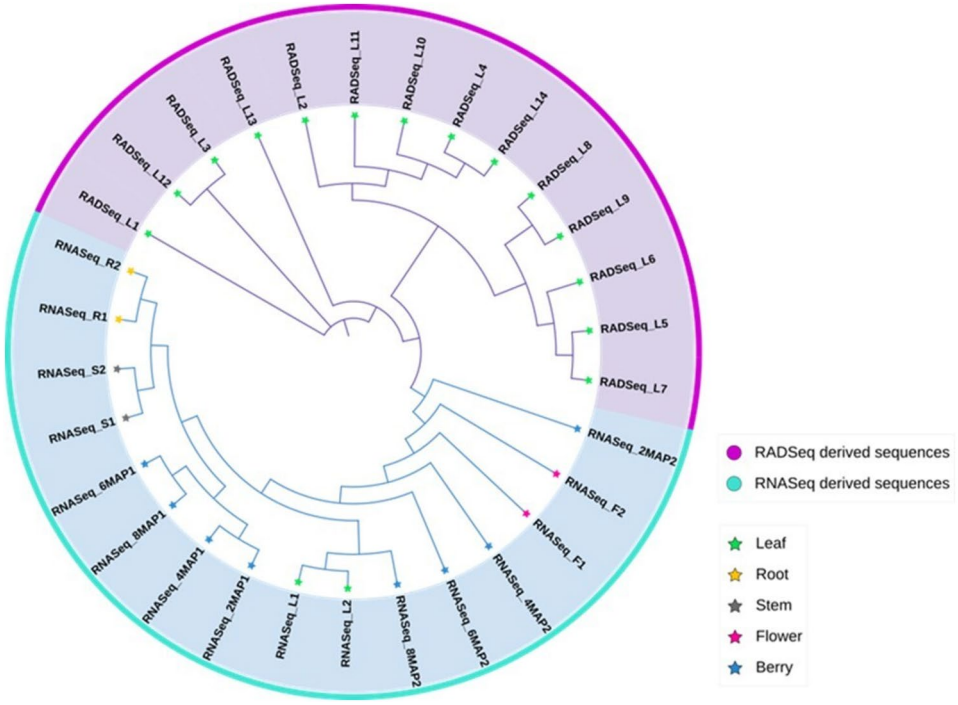

Fig. 2 Dendrogram of 30 black pepper samples based on SNP data derived via GATK4 (hard-filtering)

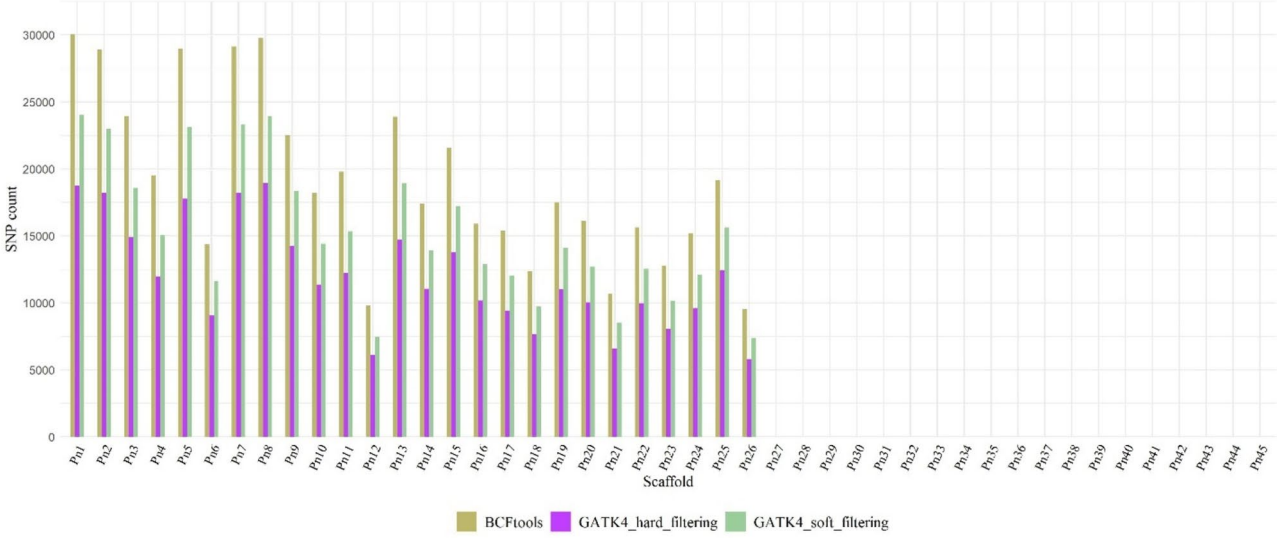

Fig. 3 Distribution of SNPs on the 45 scaffolds of the black pepper genome. Pn1 to Pn26 represent pseudo-chromosomes

were constructed from SNP alignments using RAxML. This analysis revealed a clear separation of the 30 black pepper samples into two well-supported clusters: (i) RAD-seq based samples and (ii) RNA-seq based samples (Fig. 2); only the dendrogram constructed for SNPs detected using GATK4 (hard-filtering) pipeline is shown here. This indicates the presence of potentially distinct SNP populations between the two sequencing techniques.

Filtered SNPs were plotted across scaffolds of the black pepper genome, revealing non-uniform distribution (Table 2; Fig. 3). It was evident that scaffolds Pn1, Pn2, Pn5, Pn7, and Pn8 had the highest number of SNPs. No SNPs were detected in scaffolds Pn27-Pn29, Pn31-Pn33, Pn36-Pn41, and Pn43-Pn44. SNP density ranged between 0.21 and 0.91 in the pseudo-chromosomes. The highest average density of SNPs was observed on pseudo-chromosomes Pn25 (0.86 SNPs/kb), Pn8 (0.74 SNPs/kb), Pn7

(0.72 SNPs/kb), and Pn13 (0.65 SNPs/kb). SNP density across the scaffolds is shown in Fig. 4a-c.

LD analysis

Analysis of the LD among SNPs revealed a pattern of high LD, with average genome-wide LD values of 0.44 and 0.49 for SNPs detected using GATK (hard-filtering) and BCFtools, respectively (Table 3). The average LD among individual chromosomes (Pn1-Pn26) ranged from 0.39 (Pn12) to 0.52 (Pn5). Generally, a lower percentage of SNP marker pairs recorded an r^2 value of less than 0.2 (Table 3; Fig. 5a-f); only 36.56% of the SNP pairs recorded an r^2 value of less than 0.2 for SNPs identified using GATK (hard-filtering), while 31.00% of SNP pairs exhibited r^2 value below 0.2 for SNPs identified using BCFtools. The percentage of SNPs showing complete LD (r^2 = 1.0) ranged from 6.06% on chromosome Pn12 to 7.90% on chromosome Pn13 for SNPs identified using the GATK (hard-filtering) pipeline. In contrast, for the SNPs identified using the BCFtools pipeline, the percentage of SNPs with complete LD (r^2 = 1.0) ranged from 6.32% on chromosome Pn12 to 8.19% on chromosome Pn26 (Table 3).

Furthermore, analysis of the LD decay plots revealed no significant decay across the analysed distance ($r^2 \leq 0.2$), although r^2 decreased rapidly to half of its maximum value (Fig. 5a-b). The half-LD decay distances across the genome were 83.5 kb and 64.5 kb for GATK (hard-filtering) and BCFtools, respectively. Furthermore, half-LD decay was not uniform across chromosomes, ranging from 12.5 (Pn9, Pn14, and Pn20) to 97.5 (Pn12) for SNPs identified using GATK (hard-filtering), and 2.5 (Pn21) to 97.5 (Pn24) for SNPs identified using BCFtools (Table 3). Chromosome-wide LD decay for the SNPs identified using GATK (hard-filtering) is shown in Fig. 6.

Annotation of SNPs

SNPs were annotated using the SnpEff tool, and their distribution across genic and intergenic regions is depicted in Fig. 7. The majority of SNPs were located downstream of genes (32.46 to 32.68%) with an average of 32.54%, and upstream of genes (22.40 to 22.73%) with an average of 22.53%. Following these, SNPs were observed in exonic regions (15.31 to 16.93%) with an average of 16.20%, and intronic regions (14.62 to 15.05%) with an average of 14.82%. Additionally, 12.80 to 13.53% of SNPs were identified in intergenic regions with an average of 13.09%. However, either no or relatively few SNPs were found in the splice acceptor, splice donor, splice region, and 5' untranslated regions (UTR) regions.

Homozygous and heterozygous SNPs

In the context of bi-allelic SNPs, when both alleles are the same, they are referred to as homozygous. Conversely, when the alleles differ, they are referred to as heterozygous SNPs. Figure 8 shows the frequency of homozygous and heterozygous SNPs identified in the samples analyzed. A total of 994,858, 1,067,176, and 1,006,347 homozygous SNPs were identified using the BCFtools, GATK4 (soft-filtering), and GATK4 (hard-filtering) pipelines, respectively. Additionally, 2,299,838, 1,768,314, and 1,688,979 heterozygous SNPs were detected through the BCFtools, GATK4 (soft-filtering), and GATK4 (hard-filtering) pipelines, respectively. The ratio of heterozygous to homozygous SNPs varied from 0.23 to 9.52, with an average of 2.87.

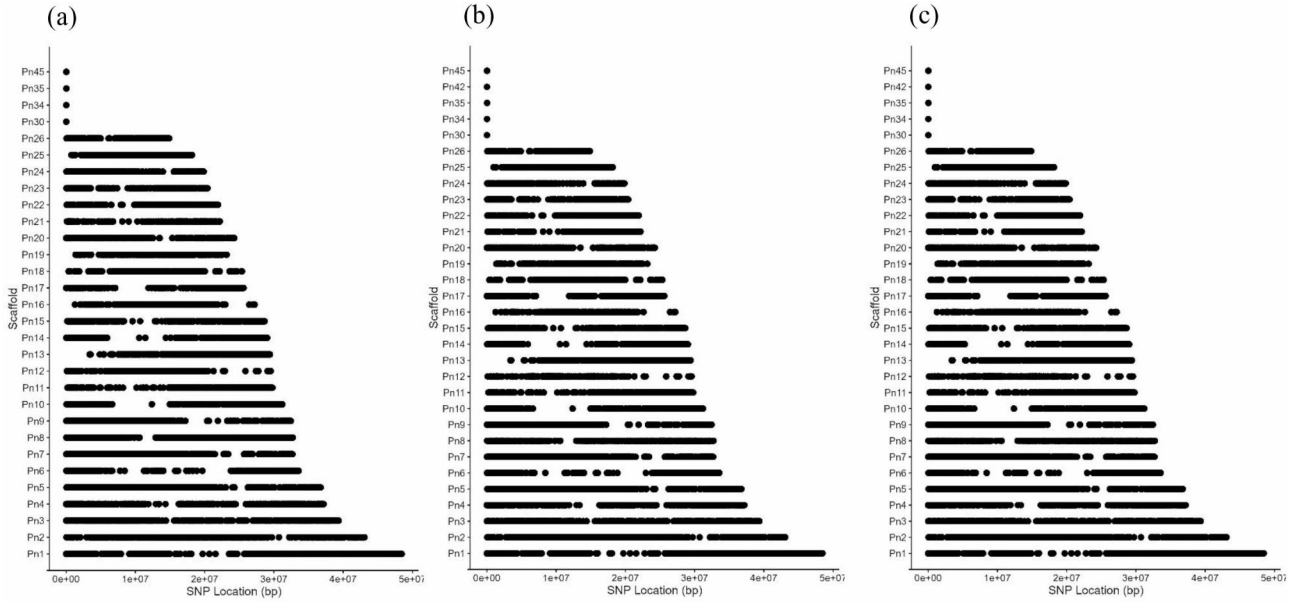

Fig. 4 SNP density plots. (**a**) BCFtools; (**b**) GATK4 (soft-filtering); (**c**) GATK4 (hard-filtering). Scaffolds with no SNPs are not shown

Table 3 Genome-wide and chromosomal scale linkage disequilibrium of SNPs

Scaffold	GATK (hard-filtering)				BCFtools			
	Average LD (r^2)	% $r^2 < 0.2$	% $r^2 = 1$	Half-LD decay distance (kb)	Average LD (r^2)	% $r^2 < 0.2$	% $r^2 = 1$	Half-LD decay distance (kb)
Pn1	0.46	34.66	7.71	18.5	0.51	28.97	8.01	46.5
Pn2	0.45	35.51	7.62	39.5	0.49	30.41	7.72	29.5
Pn3	0.43	37.01	6.86	51.5	0.48	31.60	7.06	30.5
Pn4	0.43	37.75	7.35	32.5	0.48	31.96	7.66	55.5
Pn5	0.47	32.58	7.00	18.5	0.52	27.54	7.83	70.5
Pn6	0.42	39.08	7.11	40.5	0.48	33.17	7.54	33.5
Pn7	0.45	34.95	7.41	69.5	0.50	29.33	7.65	93.5
Pn8	0.43	37.70	7.09	26.5	0.49	31.93	7.41	44.5
Pn9	0.45	35.44	7.48	12.5	0.51	29.47	8.05	29.5
Pn10	0.44	36.46	7.63	23.5	0.50	31.02	7.82	61.5
Pn11	0.44	36.57	7.42	16.5	0.49	30.29	7.60	20.5
Pn12	0.39	41.77	6.06	97.5	0.44	35.26	6.32	5.5
Pn13	0.45	35.54	7.90	32.5	0.50	30.35	8.02	36.5
Pn14	0.43	38.62	7.10	12.5	0.48	32.87	7.22	31.5
Pn15	0.44	36.48	6.81	67.5	0.49	31.78	7.12	55.5
Pn16	0.42	39.20	7.30	94.5	0.48	33.53	7.65	65.5
Pn17	0.44	36.70	7.26	46.5	0.49	32.04	7.68	49.5
Pn18	0.42	38.42	7.08	34.5	0.48	31.38	7.59	92.5
Pn19	0.43	37.95	7.15	88.5	0.48	32.40	7.87	23.5
Pn20	0.43	37.10	7.06	12.5	0.49	31.85	7.49	58.5
Pn21	0.42	39.23	7.61	25.5	0.48	33.56	8.07	2.5
Pn22	0.45	35.59	7.06	73.5	0.49	30.80	7.15	7.5
Pn23	0.43	37.57	6.69	83.5	0.49	30.69	7.41	21.5
Pn24	0.44	36.41	6.91	66.5	0.48	31.08	6.85	97.5
Pn25	0.45	35.10	7.52	23.5	0.50	29.93	7.48	36.5
Pn26	0.43	38.27	7.00	28.5	0.49	32.32	8.19	35.5
WG*	0.44	36.56	7.30	83.5	0.49	31.00	7.59	64.5

*WG: whole genome

Notably, homozygous SNPs were more prevalent in certain RAD-seq samples (i.e., RADSeq_L4, RADSeq_L10, RADSeq_L11, and RADSeq_L14) compared to other samples examined (Fig. 8a). The ratio of heterozygous to homozygous SNPs in RAD-seq samples ranged from 0.23 to 3.69, with an average of 1.77. In addition, all RNA-seq samples exhibited a higher number of heterozygous SNPs compared to homozygous SNPs (Fig. 8b), with the ratio of heterozygous to homozygous SNPs varying from 1.62 to 9.52, and an average of 3.85. Notably, the ratio of heterozygous to homozygous SNPs was higher in RNA-seq samples in the BCFtools pipeline (Fig. 8c).

Transition and transversion SNPs

SNPs can be classified based on nucleotide substitution into transitions (Ts) or transversions (Tv) [79]. Transitions involve a point mutation changing a purine nucleotide to another purine (A↔G) or a pyrimidine nucleotide to another pyrimidine (C↔T). On the other hand, transversions involve substituting a purine for a pyrimidine, or vice versa (C↔G, T↔G, A↔C, A↔T). The quality of the SNP data was analyzed by calculating the ratio of Ts to Tv (Ts/Tv) (Table 4). This ratio serves as a benchmark for evaluating sequencing and SNP data quality in different samples [80]. The average number of Ts and Tv type SNPs were 228,415 (56.81%) and 173,679 (43.19%), respectively, with a Ts/Tv ratio of 1.32. The Ts/Tv ratio observed was relatively lower than the expected ratio of 2.1 and 2.07 reported in whole-genome sequencing for known and novel variants, respectively [80]. However, the ratio was higher than the expected ratio for random substitutions [80]. BCFtools detected a higher number of both Ts and Tv SNPs compared to the GATK4 pipelines. Additionally, a bias towards Ts over Tv was observed in the *P. nigrum* genome. Among the Ts events, the substitution of A↔G (114,527) was the most common, followed by C↔T (113,888) (Table 4), whereas A↔T (54,147) and A↔C (43,410) were the most frequent Tv events.

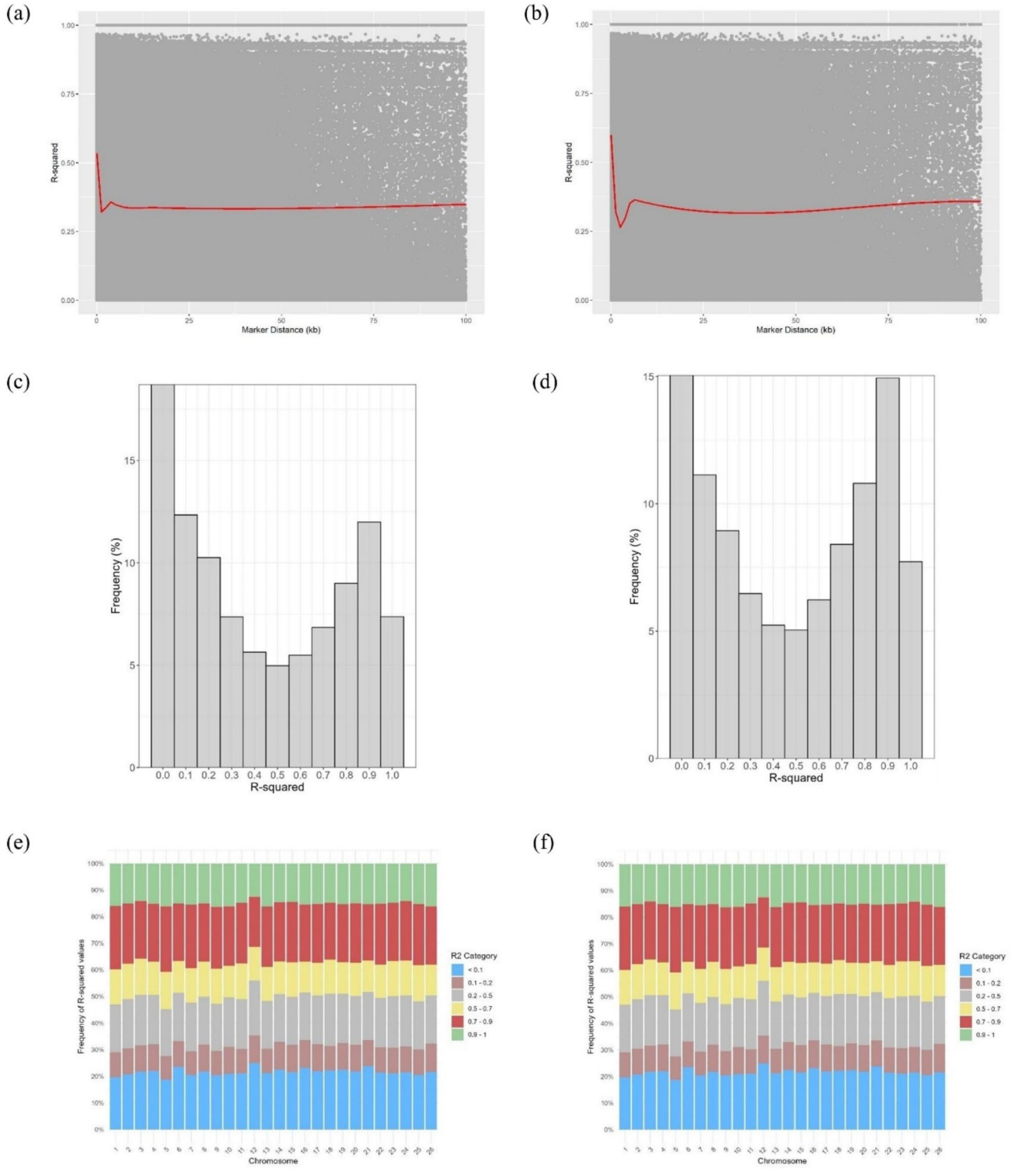

Fig. 5 Estimates of linkage disequilibrium (LD) for the identified SNPs. (**a**) - (**b**) Scatter plots representing the genome-wide LD values over the physical distance. The red curve line represents the LD decay pattern, fitted using nonlinear LOESS regression; (**c**) - (**d**) Distribution of genome-wide LD values; (**e**) - (**f**) Frequency distribution of LD values categorized by chromosomes. Panels (**a**), (**c**), and (**e**) show the analysis of SNPs identified using GATK (hard-filtering), while panels (**b**), (**d**), and (**f**) represent the analysis of SNPs identified using BCFtools

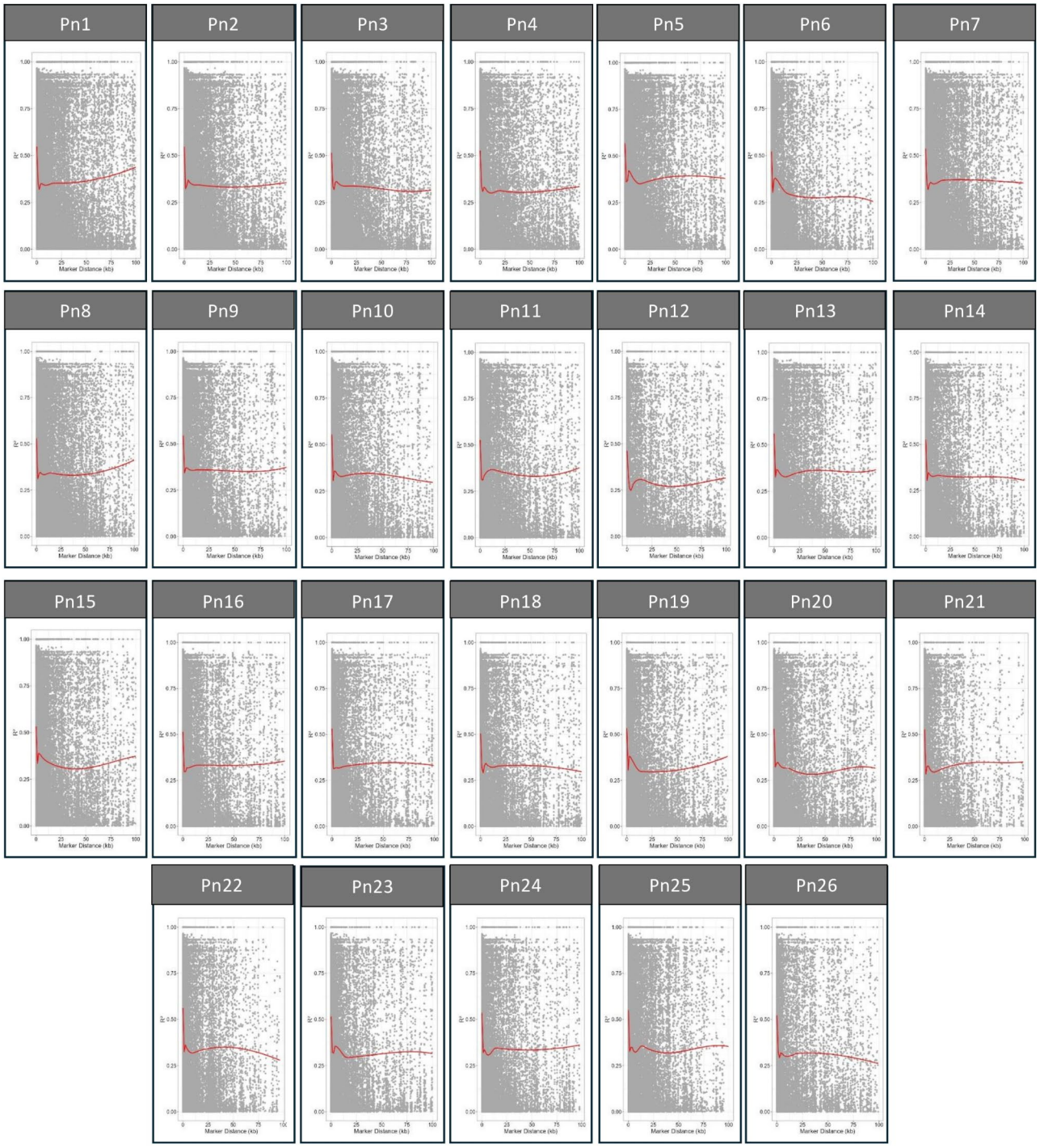

Fig. 6 Scatter plots representing the chromosome-wide linkage disequilibrium (LD) decay for the SNPs identified using GATK (hard-filtering). The red curve line represents the LD decay pattern, fitted using nonlinear LOESS regression. Scatter plots representing chromosome-wide LD-decay for the SNPs identified using BCFtools are provided in Additional file 2

Synonymous and non-synonymous SNPs

Coding SNPs can be functionally categorized into synonymous and non-synonymous SNPs. Synonymous SNPs include silent substitutions, while non-synonymous SNPs include missense and nonsense SNPs. In the present study, missense mutations were found to be the most prevalent, followed by silent and nonsense mutations (Fig. 9). BCFtools identified a total of 96,984 (54.33%) missense, 1,563 (0.88%) nonsense, and 79,946 (44.79%) silent mutations (Fig. 9a). Similarly, GATK4

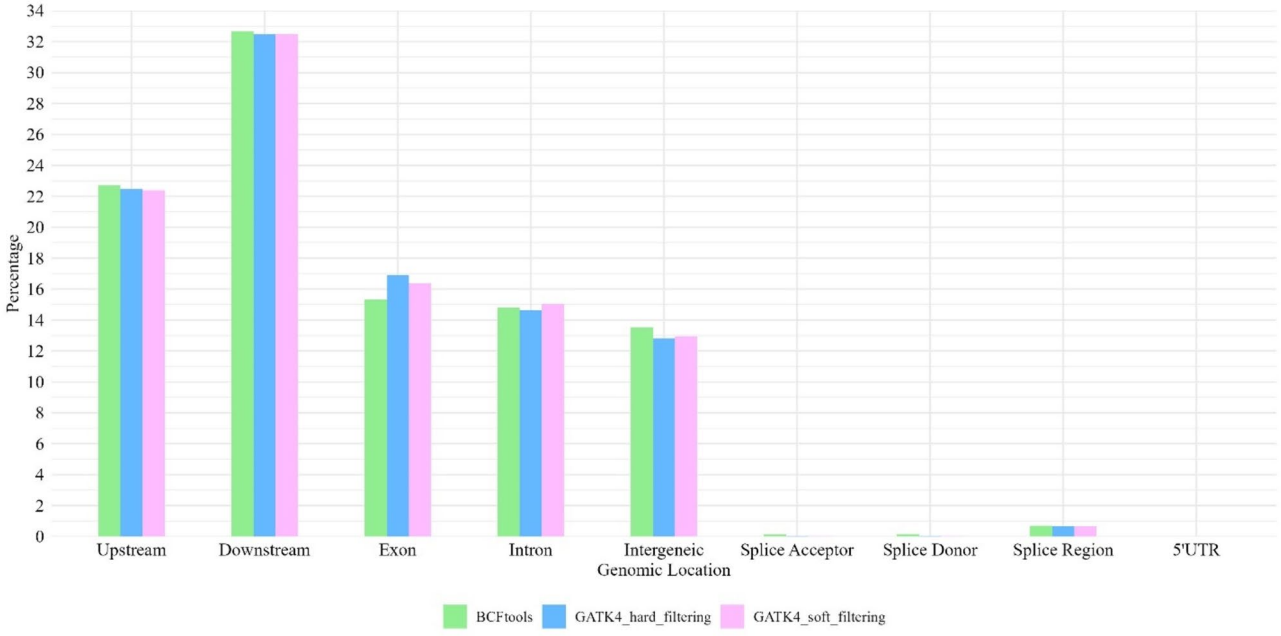

Fig. 7 Genomic annotation of SNPs. UTR: untranslated region

(soft-filtering) identified 79,456 (53.06%) missense, 1,206 (0.81%) nonsense, and 69,072 (46.13%) silent mutations (Fig. 9b). GATK4 (hard-filtering) resulted in a total of 64,879 (53.14%) missense, 997 (0.82%) nonsense, and 56,204 (46.04%) silent mutations (Fig. 9c).

Moreover, we used the SnpEff tool to evaluate the putative impact of the SNPs, categorizing them into three categories: low impact, moderate impact, high impact, and modifier (Fig. 10a-c). A total of 1,322 and 11,169 genes were consistently identified as having high and moderate impact, respectively, across all three SNP calling and filtering pipelines (Additional file 3). SNPs classified as high impact are expected to have a disruptive effect on the gene function due to a gain or loss of stop codons, frameshift variations, splice acceptor/donor variants, and/or loss of start codons in the respective genes [81]. On average 1,044 genes contained stop-gain and 283 genes contained stop-loss variants, potentially leading to significant functional consequences such as protein truncation, loss of function, or degradation of transcripts (Fig. 11a). Interestingly, we observed stop-gain and stop-loss variants in several genes that exhibited significant expression in different stages of black pepper berries compared to other stages, such as root, stem, leaf, and flower. This included Pn1.2104, Pn1.2300, Pn1.3735, Pn2.2864, Pn2.1105, Pn2.883, Pn2.1331, Pn2.1301, Pn3.893, Pn3.4770, Pn5.3086, Pn7.1985, Pn8.2626, Pn8.305, Pn8.631, Pn10.1877, Pn10.1691, Pn11.2427, Pn11.2203, Pn15.786, Pn15.31, Pn16.847, Pn19.897, Pn21.1229, and Pn24.540 (Fig. 12).

Moderate impact SNPs consisted of non-disruptive missense variants that could potentially affect the protein effectiveness due to nucleotide substitutions (Fig. 11b). Further analysis of high and moderate impact SNPs revealed the presence of these types of variants in genes involved in secondary metabolism pathways and alkaloid metabolism pathways (Table 5). Low impact was observed in synonymous variants as well as stop codon retained variants and splice region variants, which are unlikely to change protein function (Fig. 11b). Modifier SNPs typically include non-coding variants (e.g. upstream and downstream gene variants, intergenic variants) that could influence the functionality of respective genes (Fig. 11b).

eQTL analysis

To address the redundancy introduced by high LD in genomic data, SNP markers were pruned prior to conducting eQTL analysis. This pruning, based on their LD levels, retained a total of 61,562 and 38,648 SNPs from the BCFtools and GATK (hard-filtering) pipelines, respectively. Additionally, the filtering of genes based on their expression levels retained 37,093 genes. The association analysis identified a total of 294,055 SNP-gene associations for SNPs detected using the BCFtools pipeline, applying a genome-wide significance threshold ($-\log_{10}$(P-value) > $8.12e^{-7}$) (Fig. 13a). Furthermore, we identified 1,316 significant cis-acting SNPs (cis-SNPs) associated with 1,161 genes (cis-genes), and 2,009 trans-acting SNPs (trans-SNPs) associated with 566 genes (trans-genes).

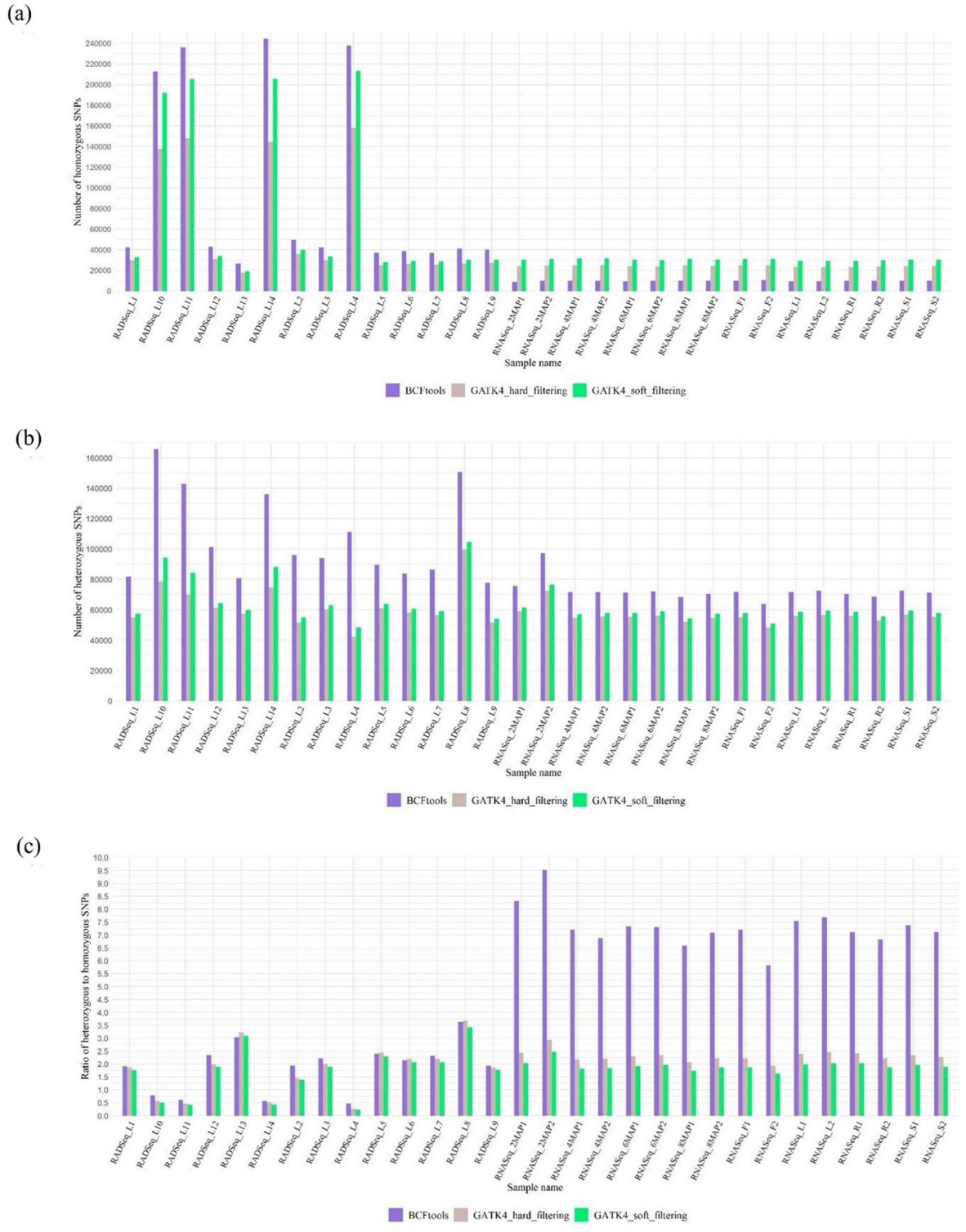

Fig. 8 Frequency of homozygous (**a**), heterozygous (**b**) SNPs identified in the samples, and the ratio of heterozygous to homozygous SNPs per sample (**c**)

Table 4 The frequency of transition and transversion SNPs

Substitution	BCFtools	GATK4 (soft-filtering)	GATK4 (hard-filtering)
Transitions (Ts)	281,040	225,839	178,367
C↔T	140,365	112,609	88,690
A↔G	140,675	113,230	89,677
Transversions (Tv)	**217,088**	**170,164**	**133,786**
C↔G	40,811	31,982	25,518
A↔T	67,884	53,245	41,312
A↔C	54,275	42,441	33,514
G↔T	54,118	42,496	33,442
Ts/Tv ratio	**1.2946**	**1.3272**	**1.3332**

Similarly, the analysis of SNPs detected using GATK (hard-filtering) identified a total of 243,772 SNP-gene associations at the genome-wide significance threshold ($-\log_{10}$(P-value) > $1.29e^{-6}$) (Fig. 13b). Additionally, 851 significant cis-SNPs corresponding to 808 cis-genes, and 1,405 significant trans-SNPs associated with 563 trans-genes were identified.

Furthermore, cis-QTLs were readily visualized as a distinct diagonal pattern, showing correspondence between the genomic positions of eQTLs and their associated transcript loci. In contrast, trans-QTLs displayed a scattered, non-uniform distribution across the genome (Fig. 14a-b).

The intersection of cis- and trans-genes identified through the eQTL analysis of SNPs detected using BCFtools and GATK (hard-filtering) revealed 675 cis-genes and 554 trans-genes, respectively (Additional file 4). GO mapping and annotation of these genes indicated that the majority of shared cis-genes were associated with biological processes, including cellular process (357 genes), response to stimulus (262 genes), and biological regulation (181 genes). In terms of molecular functions, the highest number of genes were associated with binding (277 genes), followed by catalytic activity (221 genes) (Fig. 15a). Similarly, most shared trans-genes were associated with GO terms related to biological processes, such as cellular process (353 genes), response to stimulus (240 genes), biological regulation (148 genes), and molecular functions, such as binding (291 genes), and catalytic activity (226 genes) (Fig. 15b).

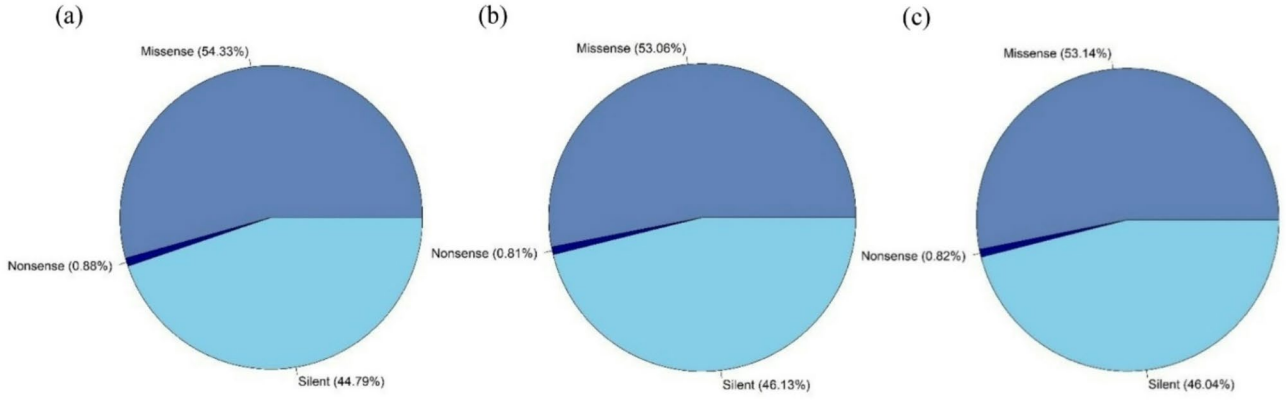

Fig. 9 Number of missense, nonsense, and silent SNPs detected via different SNP calling and filtering pipelines. (**a**) BCFtools; (**b**) GATK4 (soft-filtering); (**c**) GATK4 (hard-filtering)

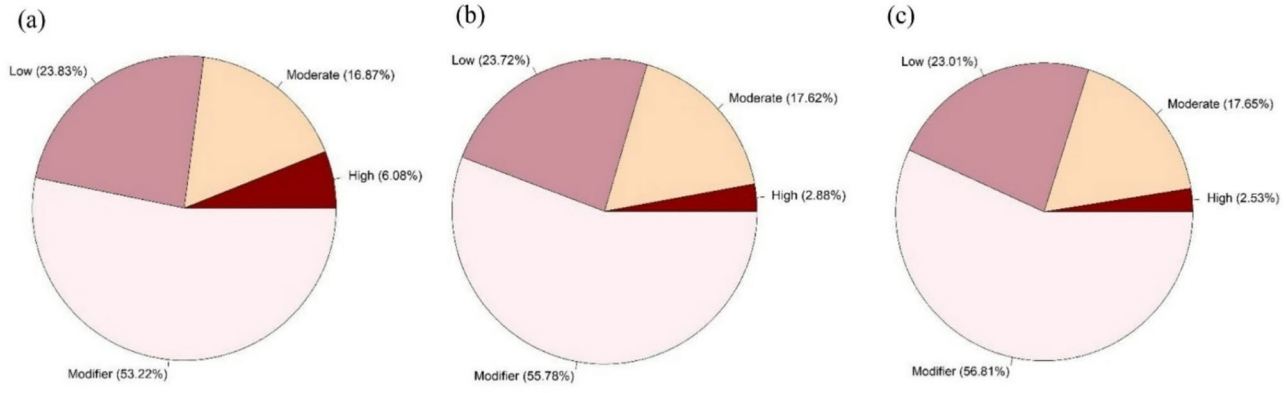

Fig. 10 Percentage contribution of SNPs in high, low, moderate, and modifier type of effects. (**a**) BCFtools; (**b**) GATK4 (soft-filtering); (**c**) GATK4 (hard-filtering)

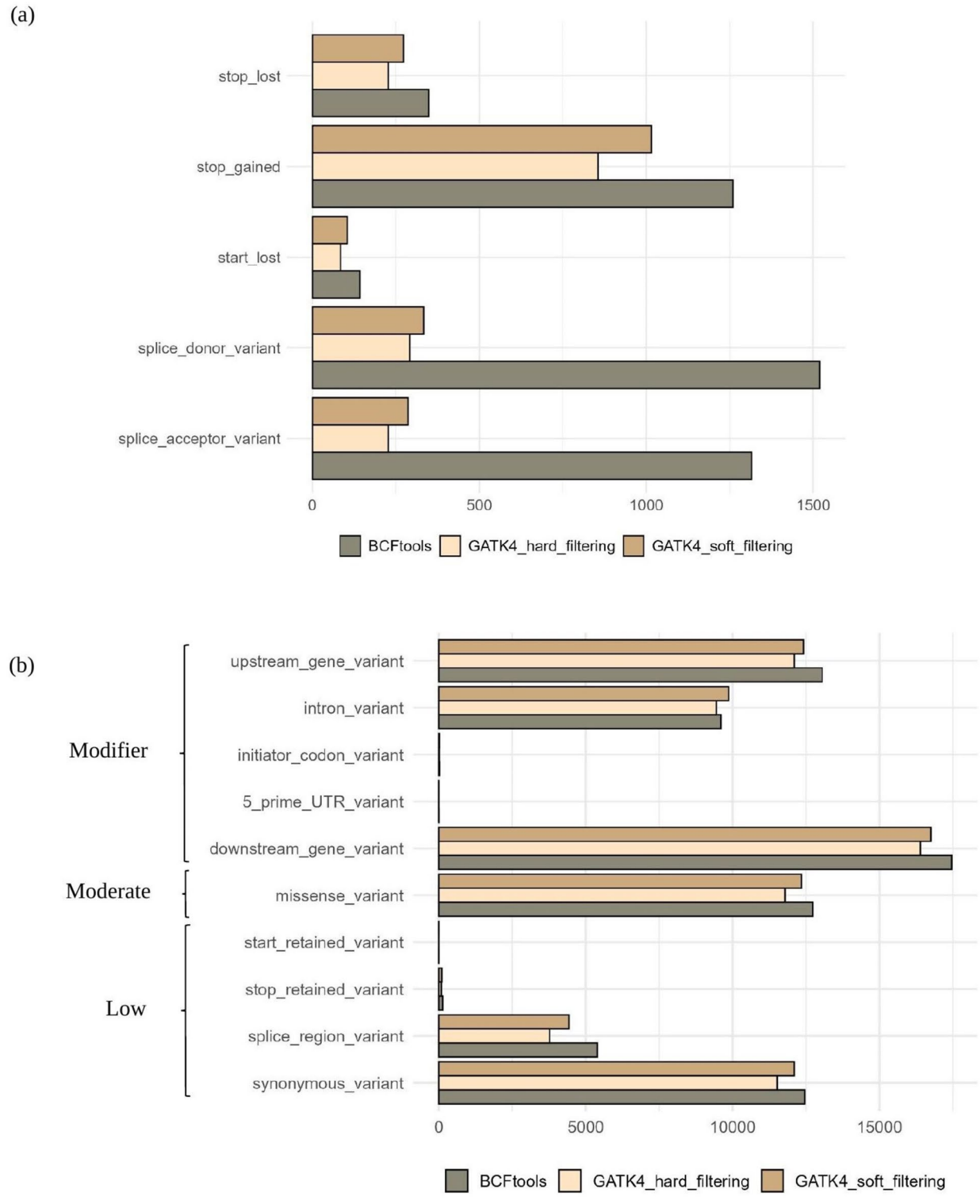

Fig. 11 Types of SNP effects and their distribution. (**a**) High impact SNPs, (**b**) Moderate, low, and modifier SNPs

Discussion

High-density SNPs are the most prevalent and stable molecular genetic markers in eukaryotes. They play a significant role in assessing individual variation, population diversity, and the evolution of plant species [26, 81, 82]. The advent of high throughput sequencing technologies has accelerated the accumulation of sequence data across numerous agriculturally important crops [83, 84, 85, 86,

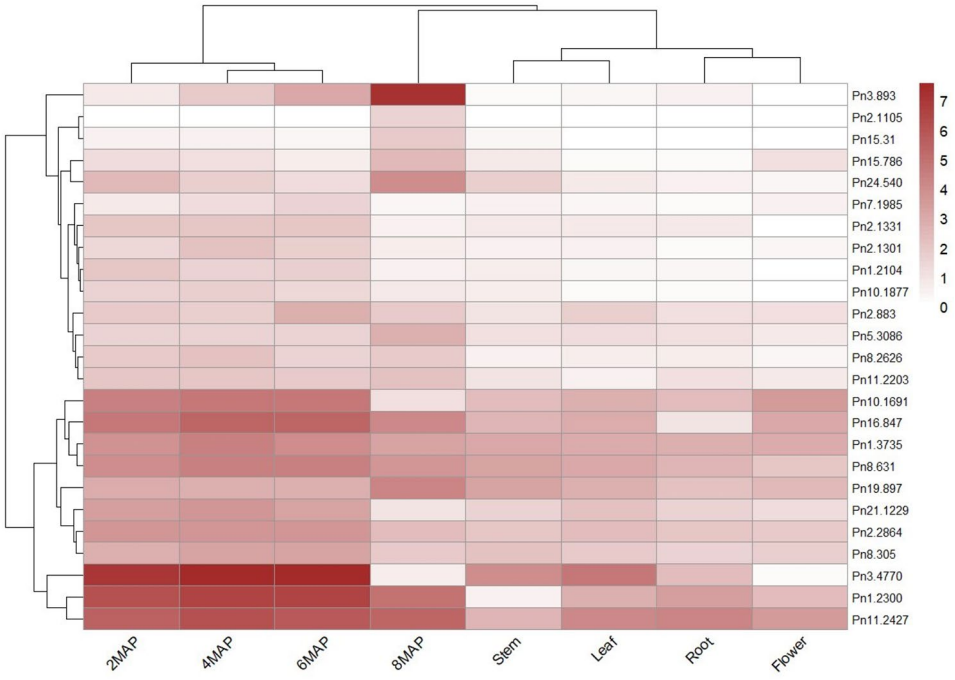

Fig. 12 Expression of selected genes in different stages of black pepper berries (2 months after pollination (MAP), 4MAP, 6MAP, 8MAP), root, stem, leaf, and flower. The color scale on the right represents normalized log2 expression values

Table 5 Genes involved in alkaloid metabolism and secondary metabolism pathways with high or moderate effect SNPs

Pathway	Gene IDs
Alkaloid metabolism	Pn2.2494, Pn3.2263, Pn3.2268, Pn3.2278, Pn4.102, Pn4.3276, Pn5.162, Pn5.2057, Pn5.2058, Pn5.2062, Pn5.2063, Pn6.1978, Pn6.601, Pn7.1477, Pn7.1523, Pn8.619, Pn11.1124, Pn14.1396, Pn14.81, Pn16.214, Pn20.487, Pn22.491, Pn22.492, Pn22.745, Pn24.361, Pn24.364, Pn24.367, Pn24.370, Pn24.371, Pn26.157, Pn51.4
Secondary metabolism	Pn1.1495, Pn1.1976, Pn1.1979, Pn1.1980, Pn1.1982, Pn1.1983, Pn1.1984, Pn1.1985, Pn1.3447, Pn1.3451, Pn1.3454, Pn3.2308, Pn3.3892, Pn3.3894, Pn4.897, Pn4.964, Pn4.965, Pn4.982, Pn4.984, Pn4.2027, Pn4.2030, Pn4.2031, Pn4.2172, Pn4.2177, Pn4.2187, Pn6.1446, Pn6.1456, Pn6.1457, Pn6.1459, Pn6.1460, Pn6.1656, Pn6.1663, Pn6.2497, Pn6.2498, Pn6.2500, Pn6.2501, Pn7.1624, Pn7.1625, Pn7.1626, Pn7.1631, Pn8.264, Pn8.1028, Pn8.1029, Pn10.717, Pn11.267, Pn11.1372, Pn14.1349, Pn14.1353, Pn14.1371, Pn14.1373, Pn15.212, Pn15.213, Pn15.214, Pn15.221, Pn15.222, Pn15.1244, Pn15.1247, Pn17.545, Pn17.615, Pn17.1120, Pn17.1121, Pn17.1721, Pn17.1722, Pn17.1725, Pn19.1165, Pn19.1167, Pn21.925, Pn21.927, Pn21.928, Pn21.931, Pn21.936, Pn21.938, Pn21.950, Pn21.952, Pn21.953, Pn21.954, Pn22.576, Pn22.874, Pn23.254, Pn23.299, Pn23.305, Pn23.341, Pn23.342, Pn23.364, Pn23.365, Pn26.430, Pn26.435, Pn26.436, Pn26.438, Pn26.439

87, 88], presenting opportunities for SNP marker development in non-model crop species, such as black pepper. Despite its economic importance, a comprehensive catalog of SNPs in black pepper has been lacking, hindering genetic diversity exploration and breeding efforts.

Our study utilized publicly available NGS data from RNA-seq and RAD-seq experiments archived in the SRA [53] to identify SNPs in the black pepper genome. Cluster analysis based on SNP data revealed a distinct separation between samples derived from RAD-seq and RNA-seq datasets. This separation can be attributed to the differing methodologies of these sequencing approaches: RAD-seq targets specific genomic regions flanking restriction enzyme cut sites, capturing a subset of genetic variation from both coding and non-coding regions of the genome [36, 89], while RNA-seq focuses on transcribed regions likely to influence phenotypic changes [33, 90]. Thus, understanding these methodological differences is crucial for accurately interpreting genetic variation in studies of population genetics and evolutionary biology.

In our study, we employed two widely used variant callers, BCFtools mpileup and GATK HaplotypeCaller, applying different variant filtering criteria; to ensure a fair comparison, GATK (soft-filtering) was employed with the same filtering criteria as BCFtools (FS > 60.0 and MQ < 40.0). Notably, BCFtools identified a higher number of SNPs compared to GATK (soft-filtering), whereas GATK (hard-filtering) significantly reduced the SNP count. While GATK is generally favored for plant

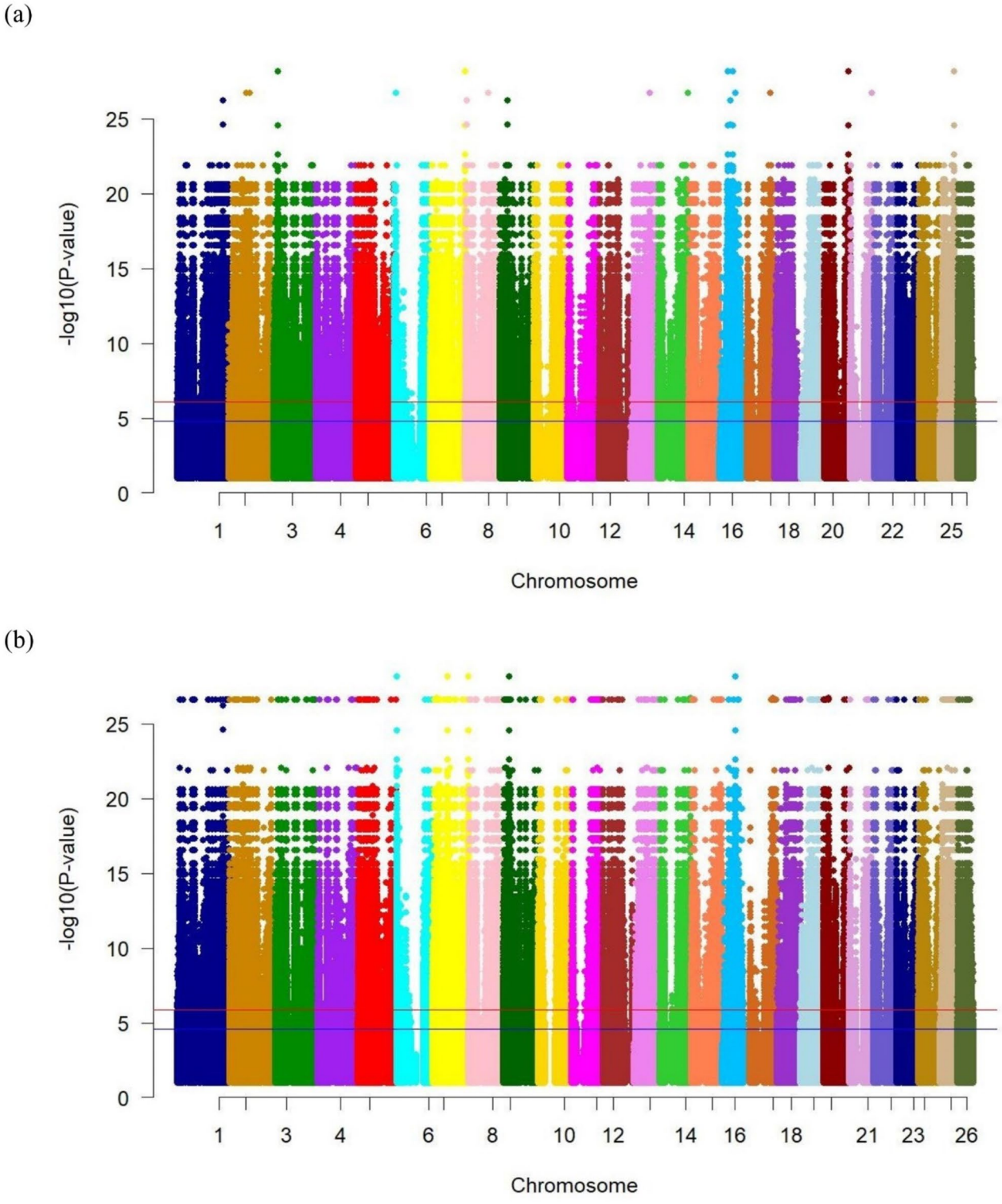

Fig. 13 Manhattan plots of SNPs associated with expressed genes. (**a**) SNPs identified using BCFtools with the blue horizontal line indicating the suggestive threshold of $1.62e^{-5}$ and the red line representing the genome-wide significance threshold of $8.12e^{-7}$; (**b**) SNPs identified using GATK (hard-filtering), where the blue horizontal line represents the suggestive threshold of $2.59e10^{-5}$, and the red line represents the genome-wide significance threshold of $1.29e^{-6}$

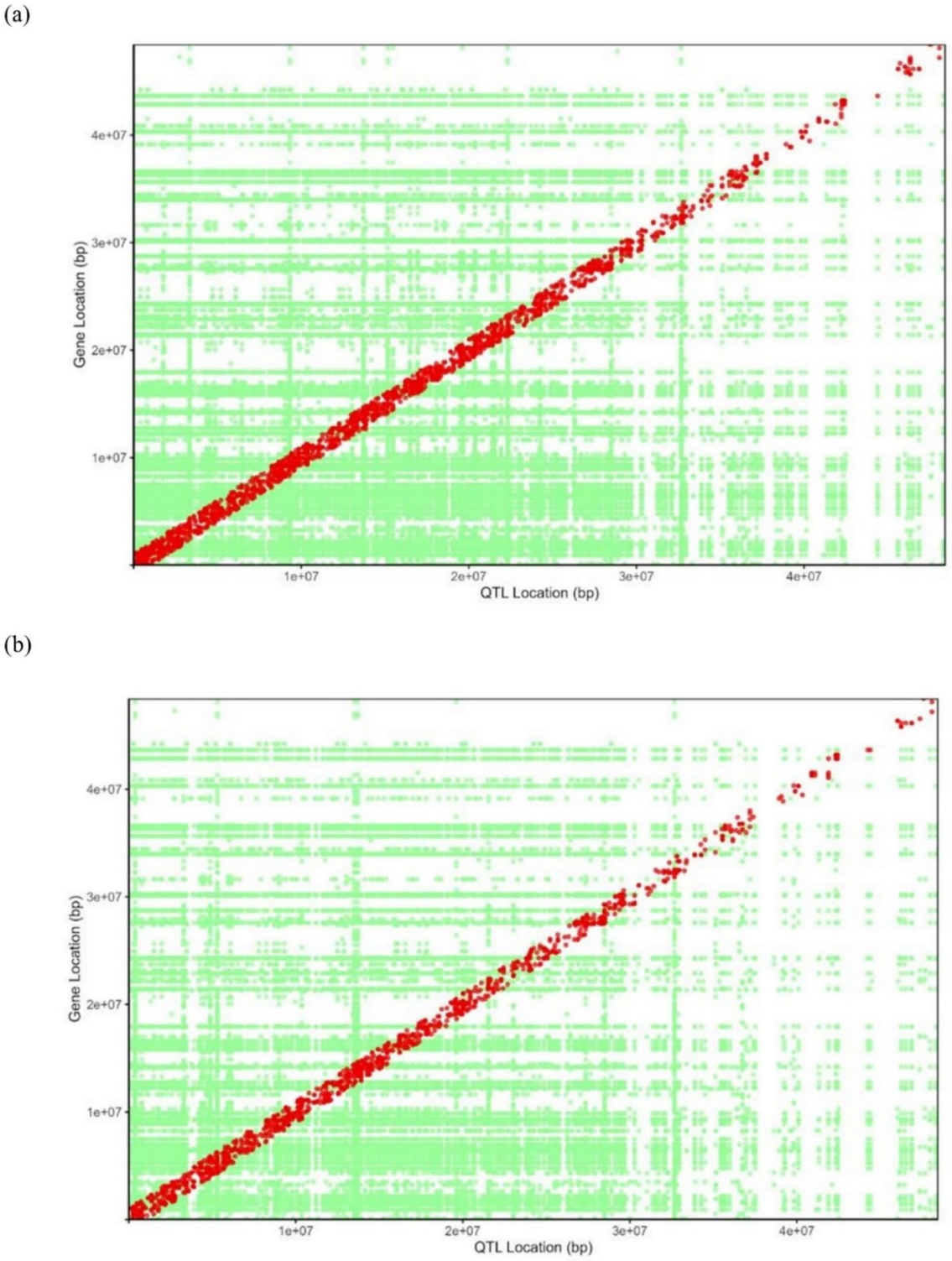

Fig. 14 eQTLs observed across the genome. (**a**) eQTLs identified for SNPs detected using BCFtools; (**b**) eQTLs identified for SNPs detected using GATK (hard-filtering). The genomic position of eQTLs for target genes is shown. The clear diagonal band of red dots represents cis-acting eQTLs, while the off-diagonal green dots represent trans-acting eQTLs

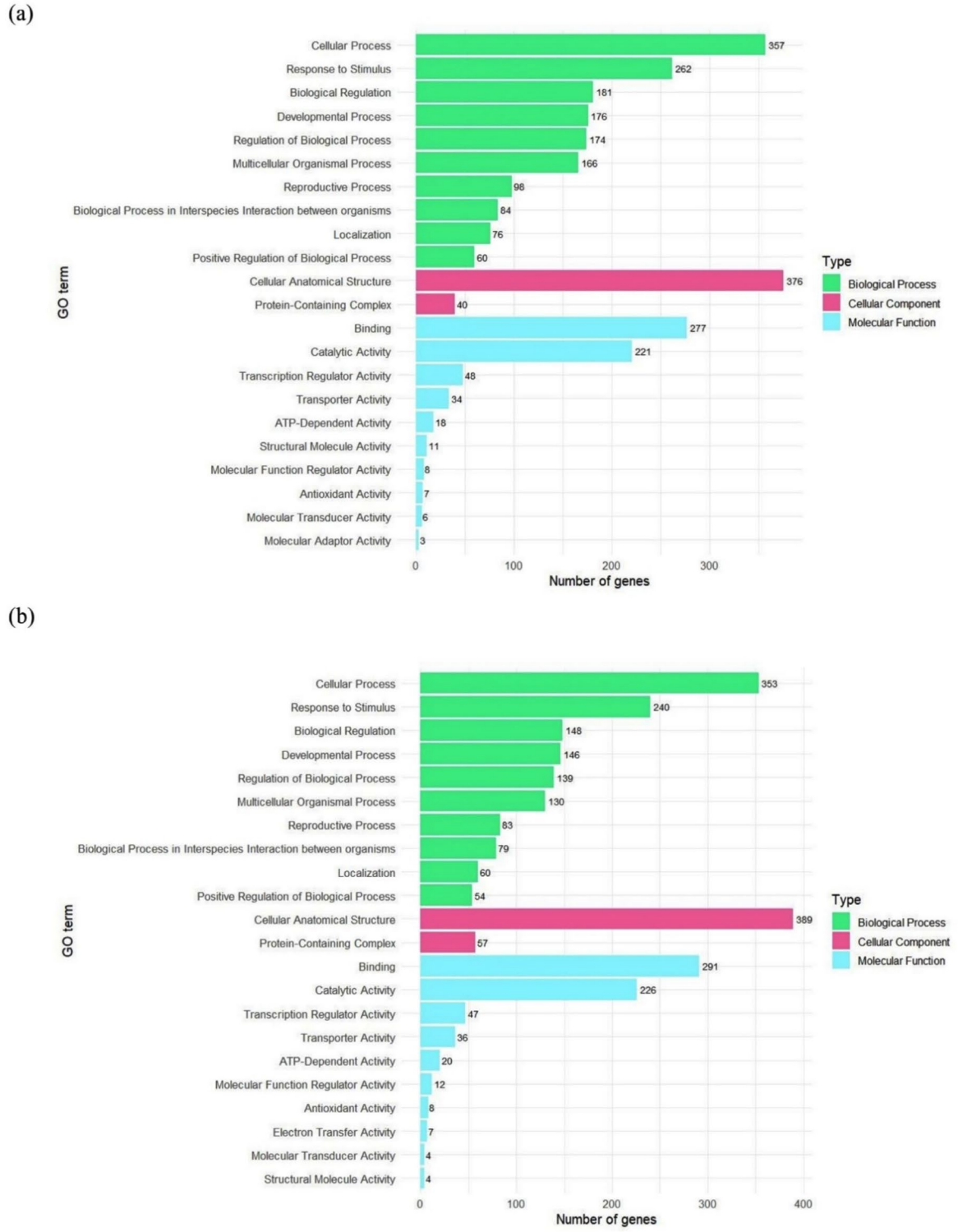

Fig. 15 Top ten Gene Ontology terms (GO) associated with cis-genes (**a**) and trans-genes (**b**). The number next to each bar indicates the number of genes corresponding to the respective GO term

datasets [42, 43], further investigation is necessary to assess the true positive and false positive rates of SNPs and evaluate tool performance in crop genomes.

Several studies have investigated genome-wide SNP coverage in crops such as rice [91], wheat [92], Indian mustard (*Brassica juncea* L.) [93], cotton [94], and soybean [95], revealing diverse SNP density and distribution patterns. For instance, cotton exhibits uneven SNP distribution across its chromosomes, averaging one SNP per 0.5 kb [94], while soybean, with a lower SNP density (one SNP per 41.59 kb), shows similar uneven distribution patterns [95]. In contrast, *B. Juncea* demonstrates a more uniform SNP distribution [93]. In the black pepper reference genome, consisting of 45 scaffolds with pseudochromosomes Pn1 to Pn26, SNP distribution was uneven, averaging approximately one SNP per 3 kb across all scaffolds, with scaffolds Pn27 to Pn45 showing minimal SNP presence. Several factors, including natural genetic diversity, selection pressures favoring specific alleles, variable mutation rates, and methodological limitations, could contribute to the absence or low abundance of SNPs in some of the scaffolds.

The transition-to-transversion (Ts/Tv) ratio, a critical measure of SNP quality [96, 97, 98], was approximately 1.36 in black pepper, indicating a bias towards Ts mutations, a trend observed in other plant genomes [99, 100], such as *Brassica napus* [101], *Hevea brasiliensis* [102], *Camellia sinensis* [103], *Vigna mungo* [104], *Camelina sativa* [105], and *Solanum lycopersicum* [81]. This bias may be influenced by factors such as cytosine methylation levels in the genome [106, 107].

SnpEff serves as a valuable database for predicting the potential impacts of SNPs [67]. Annotation of SNPs using SnpEff revealed that the majority (approximately 83%) were located within gene body regions, with downstream and upstream variants accounting for nearly 54% of SNPs. These variants potentially play significant roles in gene regulation, offering promising targets for breeding programs aimed at enhancing crop traits [79]. However, further studies are needed to validate their functional impacts on gene expression and protein function, as not all variants may be functionally consequential [108, 109]. Notably, we identified nearly 53.58% of missense SNPs, 45.57% of silent SNPs, and a small proportion of nonsense SNPs (0.83%). Missense mutations, which are a type of nonsynonymous SNPs, can cause structural and/or functional alterations in proteins. In contrast, silent mutations are commonly regarded as low-impact variants, as they do not affect protein function. Therefore, the identification of missense SNPs within the coding regions of the black pepper genome, in particular genes associated with alkaloid and secondary metabolism biosynthesis pathways holds particular interest, as it enables the investigation of their potential effects on gene function and phenotype.

The extent of LD between markers and its decay over the genetic distance plays a pivotal role in determining the required number and density of SNP markers for association studies [110, 111, 112]. Self-pollinating plants often exhibit elevated levels of LD due to reduced effective recombination rates [110]. In this study, LD analysis using SNPs identified through BCFtools and GATK (hard-filtering) revealed high levels of LD across the black pepper genome. The observed higher r^2 value suggests a stronger genetic linkage between SNPs, indicating that these loci are more likely to be inherited together due to reduced genetic independence [113]. Since cultivated black pepper is predominantly self-pollinated and propagated by cuttings [114, 115], the elevated LD levels suggest that fewer markers may be sufficient to ensure comprehensive genome coverage for marker-trait association studies [116, 117]. This information is valuable for optimizing marker selection in future association mapping studies.

Expression QTL analysis is a powerful approach for unraveling associations between genetic variants, such as SNPs, and gene expression. This method provides valuable insights into expression-associated SNPs and their corresponding target genes [118, 119]. In our study, we identified at least 675 black pepper genes whose expression is potentially influenced by local SNPs (cis-acting SNPs), while 554 genes are influenced by trans-acting SNPs. Functional annotation of these genes revealed their involvement in key biological processes, including cellular processes, responses to stimuli, biological regulation, as well as molecular functions, such as binding and catalytic activity. This inventory of cis- and trans-eQTLs, along with their associated target genes, serves as an important resource for deepening our understanding of the genetic and regulatory mechanisms underlying gene expression. It also provides a foundation for functional studies aimed at trait improvement in black pepper. While genome-wide association studies (GWAS) remain a powerful approach for identifying associations between genetic variants and phenotypic traits, they often encounter challenges in identifying hub genes important for precision genome editing to improve crop traits [118]. Integrating eQTL analysis with GWAS can address this limitation, providing a robust framework for marker-assisted breeding and facilitating the development of improved black pepper varieties in the future.

Conclusion

The present study presents a comprehensive catalog of genome-wide SNPs within the black pepper genome, accompanied by detailed SNP annotation. This analysis uncovered an average of 402,094 SNPs across both

genomic and transcriptomic datasets, with SNP densities ranging from 0.21 to 0.91 in pseudo-chromosomes. Notably, 260,026 bi-allelic SNPs were consistently identified across multiple SNP calling and filtering pipelines. Furthermore, we identified at least 675 genes potentially influenced by cis-acting SNPs, while 554 genes were affected by trans-acting SNPs. These findings provide a valuable resource for understanding genetic variation within the species, holding significant implications for breeding, conservation, and evolutionary research. Moreover, the substantial number of variants identified in this study forms a foundation for designing genome-wide high-density chips in the future. The availability of such tools has the potential to greatly enhance conservation strategies and breeding efforts aimed at improving black pepper.

Abbreviations

BAM	Binary alignment map
BWA	Burrows-Wheeler aligner
cisDist	Cis-distance
eQTL	Expression quantitative trait loci
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
FS	Fisher's exact test for strand bias
GATK	Genome analysis toolkit
GO	Gene Ontology
GVCF	Genomic variant call format
GWAS	Genome-wide association studies
LD	Linkage disequilibrium
LOESS	Locally estimated scatterplot smoothing
MAF	Minor allele frequency
MAP	Months after pollination
MQ	Mapping quality
MQRankSum	Mapping quality rank sum test
NCBI	National Center for Biological Information
NGS	Next-generation sequencing
QD	Quality by depth
RAD	seq-Restriction site associated DNA sequencing
RaxML	Randomized Axelerated Maximum Likelihood
ReadPosRankSum	Rank sum test for site position
RNA	seq-Ribonucleic acid sequencing
SAM	Sequence alignment map
SNP	Single nucleotide polymorphism
SOR	Strand odds ratio
SRA	Sequence read archive
Ts	Transitions
Tv	Transversions
VCF	Variant call format

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11414-2.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Acknowledgements

The RAD-seq data analysed herein were generated at the Department of Botany and Biodiversity of Research, University of Vienna, Vienna, Austria, through grant P33143-B awarded to Prof. Rosabelle Samuel, and the National Research Council (NRC) of Sri Lanka grant 19-062 awarded to Prof. Tara D. Silva, Department of Plant Sciences, Faculty of Science, University of Colombo and A.M.W. The authors would like to thank Dr. Luiz A. Cauz-Santos and Mr. Dominik Metschina for their technical support.

Author contributions

A.M.W. conceived the study, participated in its design, and assisted in drafting the manuscript. H.A.T. contributed to the study design, performed the analysis and interpreted the data, and drafted the manuscript. N.A.W. generated the RAD-seq data for the analysis. All authors read and approved the final manuscript.

Funding

Not applicable.

Data availability

The RNA-seq and RAD-seq datasets analysed during the current study are publicly available in the Sequence Read Archive (SRA), hosted by the National Center for Biotechnology Information (NCBI), under the BioProject PRJNA529760 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA529760) and PRJNA1035754 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1035754), respectively. The SRA accession numbers of all the samples used in the present study are provided in Additional file 1. The genome assemblies and annotation files of black pepper are publicly available at http://cotton.hzau.edu.cn/EN/Download.htm.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 1 July 2024 / Accepted: 27 February 2025
Published online: 17 March 2025

References

1. Ravindran PN, Kallupurackal JA. Black pepper. Handbook of herbs and spices. Woodhead Publishing; 2012. pp. 86–115.
2. Butt MS, Pasha I, Sultan MT, Randhawa MA, Saeed F, Ahmed W. Black pepper and health claims: a comprehensive treatise. Crit Rev Food Sci Nutr. 2013;53(9):875–86.
3. Hu L, Hao C, Fan R, Wu B, Tan L, Wu H. *De Novo* assembly and characterization of fruit transcriptome in black pepper (*Piper nigrum*). PLoS ONE. 2015;10(6):e0129822.
4. Takooree H, Aumeeruddy MZ, Rengasamy KR, Venugopala KN, Jeewon R, Zengin G, Mahomoodally MF. A systematic review on black pepper (*Piper nigrum* L.): from folk uses to pharmacological applications. Crit Rev Food Sci Nutr. 2019;59(Suppl 1):S210–43.
5. Damanhouri ZA, Ahmad A. A review on therapeutic potential of *Piper nigrum* L. (black pepper): the King of spices. Med Aromatic Plants. 2014;3(3):161.
6. Ahmad N, Fazal H, Abbasi BH, Farooq S, Ali M, Khan MA. Biological role of *Piper nigrum* L. (black pepper): A review. Asian Pac J Trop Biomed. 2012;2(3):S1945–53.
7. Srinivasan K. Black pepper and its pungent principle-piperine: a review of diverse physiological effects. Crit Rev Food Sci Nutr. 2007;47(8):735–48.
8. Hao CY, Rui FAN, Ribeiro MC, Tan LH, Wu HS, Yang JF, Zheng WQ, Huan YU. Modeling the potential geographic distribution of black pepper (*Piper nigrum*) in Asia using GIS tools. J Integr Agric. 2012;11(4):593–9.
9. Zachariah TJ, Parthasarathy VA. Black pepper. Chem Spices. 2008;196:21.
10. Ouborg NJ. Integrating population genetics and conservation biology in the era of genomics. Biol Lett. 2010;6(1):3–6.

11. Turchetto C, Segatto AL, Mäder G, Rodrigues DM, Bonatto SL, Freitas LB. High levels of genetic diversity and population structure in an endemic and rare species: implications for conservation. AoB Plants. 2016;8:plw002.
12. Pradeepkumar T, Karihaloo JL, Archak S, Baldev A. Analysis of genetic diversity in *Piper nigrum* L. using RAPD markers. Genet Resour Crop Evol. 2003;50:469–75.
13. Nazeem PA, Keshavachandran R, Babu TD, Achuthan CR, Girija D, Peter KV. Recent trends in horticultural biotechnology. New India Publishing Agency; 2007. pp. 485–90.
14. Sreedevi M, Syamkumar S, Sasikumar B. Molecular and morphological characterization of new promising black pepper (*Piper nigrum* L.) lines. J Spices Aromatic Crops. 2005;14(1):1–9.
15. Wu BD, Fan R, Hu LS, Wu HS, Hao CY. Genetic diversity in the germplasm of black pepper determined by EST-SSR markers. Genet Mol Res. 2016;15(1) gmr.15018099.
16. Kumari R, Wankhede DP, Bajpai A, Maurya A, Prasad K, Gautam D, Rangan P, Latha M, John KJ, Bhat KV, Gaikwad AB. Genome wide identification and characterization of microsatellite markers in black pepper (*Piper nigrum*): a valuable resource for boosting genomics applications. PLoS ONE. 2019;14(12):e0226002.
17. Joy N, Prasanth VP, Soniya EV. Microsatellite based analysis of genetic diversity of popular black pepper genotypes in South India. Genetica. 2011;139:1033–43.
18. Negi A, Singh K, Jaiswal S, Kokkat JG, Angadi UB, Iquebal MA, Umadevi P, Rai A, Kumar D. Rapid genome-wide location-specific polymorphic SSR marker discovery in black pepper by GBS approach. Front Plant Sci. 2022;13:846937.
19. Mammadov J, Aggarwal R, Buyyarapu R, Kumpatla S. SNP markers and their impact on plant breeding. Int J Plant Genomics. 2012;2012(1):728398.
20. Lai K, Duran C, Berkman PJ, Lorenc MT, Stiller J, Manoli S, Hayden MJ, Forrest KL, Fleury D, Baumann U, Zander M. Single nucleotide polymorphism discovery from wheat next-generation sequence data. Plant Biotechnol J. 2012;10(6):743–9.
21. Kumar S, Banks TW, Cloutier S. SNP discovery through next-generation sequencing and its applications. Int J Plant Genomics. 2012;209:44.
22. Wimalarathna NA, Wickramasuriya AM, Metschina D, Cauz-Santos LA, Bandupriya D, Ariyawansa KGSU, Gopallawa B, Chase MW, Samuel R, Silva TD. Genetic diversity and population structure of *Piper nigrum* (black pepper) accessions based on next-generation SNP markers. PLoS ONE. 2024;19(6):e0305990.
23. Yirgu M, Kebede M, Feyissa T, Lakew B, Woldeyohannes AB, Fikere M. Single nucleotide polymorphism (SNP) markers for genetic diversity and population structure study in Ethiopian barley (*Hordeum vulgare* L.) germplasm. BMC Genomic Data. 2023;24(1):7.
24. Yang X, Tan B, Liu H, Zhu W, Xu L, Wang Y, Fan X, Sha L, Zhang H, Zeng J, Wu D. Genetic diversity and population structure of Asian and European common wheat accessions based on genotyping-by-sequencing. Front Genet. 2020;11:580782.
25. Dube SP, Sibiya J, Kutu F. Genetic diversity and population structure of maize inbred lines using phenotypic traits and single nucleotide polymorphism (SNP) markers. Sci Rep. 2023;13(1):17851.
26. Tang W, Wu T, Ye J, Sun J, Jiang Y, Yu J, Tang J, Chen G, Wang C, Wan J. SNP-based analysis of genetic diversity reveals important alleles associated with seed size in rice. BMC Plant Biol. 2016;16(1):128.
27. Pirooznia M, Kramer M, Parla J, Goes FS, Potash JB, McCombie WR, Zandi PP. Validation and assessment of variant calling pipelines for next-generation sequencing. Hum Genomics. 2014;8(1):14.
28. Yao Z, You FM, N'Diaye A, Knox RE, McCartney C, Hiebert CW, Pozniak C, Xu W. Evaluation of variant calling tools for large plant genome re-sequencing. BMC Bioinformatics. 2020;21(360).
29. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE. 2011;6(5):e19379.
30. Bancroft I, Morgan C, Fraser F, Higgins J, Wells R, Clissold L, Baker D, Long Y, Meng J, Wang X, Liu S. Dissecting the genome of the polyploid crop oilseed rape by transcriptome sequencing. Nat Biotechnol. 2011;29(8):762–6.
31. Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS. SNP discovery via 454 transcriptome sequencing. Plant J. 2007;51(5):910–8.
32. Trick M, Long Y, Meng J, Bancroft I. Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. Plant Biotechnol J. 2009;7(4):334–46.
33. Jehl F, Degalez F, Bernard M, Lecerf F, Lagoutte L, Désert C, Coulée M, Bouchez O, Leroux S, Abasht B, Tixier-Boichard M. RNA-Seq data for reliable SNP detection and genotype calling: interest for coding variant characterization and cis-regulation analysis by allele-specific expression in livestock species. Front Genet. 2021;12:655707.
34. Odumpatta R, Mohanapriya A. Next generation sequencing exome data analysis aids in the discovery of SNP and INDEL patterns in Parkinson's disease. Genomics. 2020;112(5):3722–8.
35. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE. 2008;3(10):e3376.
36. Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA. SNP discovery and genotyping for evolutionary genetics using RAD sequencing. Mol Methods Evolutionary Genet. 2011;157–78.
37. Mammadov JA, Chen W, Ren R, Pai R, Marchione W, Yalçin F, Witsenboer H, Greene TW, Thompson SA, Kumpatla SP. Development of highly polymorphic SNP markers from the complexity reduced portion of maize [*Zea mays* L.] genome for use in marker-assisted breeding. Theor Appl Genet. 2010;121:577–88.
38. Chutimanitsakun Y, Nipper RW, Cuesta-Marcos A, Cistué L, Corey A, Filichkina T, Johnson EA, Hayes PM. Construction and application for QTL analysis of a restriction site associated DNA (RAD) linkage map in barley. BMC Genomics. 2011;12(4).
39. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.
40. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
41. Yi M, Zhao Y, Jia L, He M, Kebebew E, Stephens RM. Performance comparison of SNP detection tools with illumina exome sequencing data-an assessment using both family pedigree information and sample-matched SNP array data. Nucleic Acids Res. 2014;42(12):e101.
42. Wu X, Heffelfinger C, Zhao H, Dellaporta SL. Benchmarking variant identification tools for plant diversity discovery. BMC Genomics. 2019;20(701).
43. Schilbert HM, Rempel A, Pucker B. Comparison of read mapping and variant calling tools for the analysis of plant NGS data. Plants. 2020;9(4):439.
44. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep. 2015;5(1):17875.
45. Kumaran M, Subramanian U, Devarajan B. Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data. BMC Bioinformatics. 2019;20(1):342.
46. Alosaimi S, van Biljon N, Awany D, Thami PK, Defo J, Mugo JW, Bope CD, Mazandu GK, Mulder NJ, Chimusa ER. Simulation of African and non-African low and high coverage whole genome sequence data to assess variant calling approaches. Brief Bioinform. 2021;22(4):bbaa366.
47. Lefouili M, Nam K. The evaluation of Bcftools Mpileup and GATK haplotypecaller for variant calling in non-human species. Sci Rep. 2022;12(1):11331.
48. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. ArXiv Preprint ArXiv:1207.3907. 2012 Jul 17.
49. Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucleic Acids Res. 2011;39(19):e132.
50. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, Johnson J, Dougherty B, Barrett JC, Dry JR. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016;44(11):e108.
51. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009;25(17):2283–5.
52. Hu L, Xu Z, Wang M, Fan R, Yuan D, Wu B, Wu H, Qin X, Yan L, Tan L, Sim S. The chromosome-scale reference genome of black pepper provides insight into Piperine biosynthesis. Nat Commun. 2019;10(1):4702.
53. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. The sequence read archive. Nucleic Acids Res. 2010;39(Suppl 1):D19–21.
54. The Sequence Read Archive. https://www.ncbi.nlm.nih.gov/sra. Accessed 10 July 2023.
55. Group of Cotton Genetic Improvement. https://cotton.hzau.edu.cn/EN/Download.htm. Accessed 8 July 2023.

56. FastQC. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 20 July 2023.
57. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
58. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
59. Picard toolkit. https://github.com/broadinstitute/picard/releases. Accessed 5 September 2023.
60. Bamtools. https://github.com/pezmaster31/bamtools. Accessed 8 Septembe r 2023.
61. Samtools. https://github.com/samtools/samtools. Accessed 10 September 2023.
62. Danecek P, McCarthy SA. BCFtools/csq: haplotype-aware variant consequences. Bioinformatics. 2017;33(13):2037–9.
63. AlbersCornelis A, DePristoMark A, HandsakerRobert E, MarthGabor T, SherryStephen T. The variant call format and vcftools. Bioinformatics. 2011;27(15):2156–8.
64. Lischer HE, Excoffier L. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics. 2012;28(2):298–9.
65. Letunic I, Bork P. Interactive tree of life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics. 2007;23(1):127–8.
66. Letunic I, Bork P. Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021;49(W1):W293–6.
67. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly. 2012;6(2):80–92.
68. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
69. Wickham H. ggplot2: elegant graphics for data analysis. 2nd ed. New York: Springer; 2016.
70. Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet. 2007;39(9):1151–5.
71. Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL, Li MW, He W, Qin N, Wang B, Li J. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet. 2010;42(12):1053–9.
72. Branca A, Paape TD, Zhou P, Briskine R, Farmer AD, Mudge J, Bharti AK, Woodward JE, May GD, Gentzbittel L, Ben C. Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. Proceedings of the National Academy of Sciences. 2011;108(42):E864-70.
73. MagnoliidsGDB. http://www.magnoliadb.com:7777. Accessed 20 January 2025.
74. Chen Y, Yang Z, Chen J, Li P, Zhao X, Huang S, Li Z, Huang S, Luo J, Hu H, Ding Y. MagnoliidsGDB: an integrated functional genomics database for magnoliids. BioRxiv. 2024:2024–08.
75. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28(10):1353–8.
76. Wang T, Niu Q, Zhang T, Zheng X, Li H, Gao X, Chen Y, Gao H, Zhang L, Liu GE, Li J. Cis-eQTL analysis and functional validation of candidate genes for carcass yield traits in beef cattle. Int J Mol Sci. 2022;23(23):15055.
77. Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 2008;36(10):3420–35.
78. BioBam Bioinformatics Solutions. https://www.biobam.com/omicsbox. Accessed 20 January 2025.
79. Guajardo V, Solís S, Almada R, Saski C, Gasic K, Moreno MÁ. Genome-wide SNP identification in *Prunus* rootstocks germplasm collections using genotyping-by-sequencing: phylogenetic analysis, distribution of SNPs and prediction of their effect on gene function. Sci Rep. 2020;10(1):1467.
80. Liu Q, Guo Y, Li J, Long J, Zhang B, Shyr Y. Steps to ensure accuracy in genotype and SNP calling from illumina sequencing data. BMC Genomics. 2012;13(Suppl 8):S8.
81. Bhardwaj A, Dhar YV, Asif MH, Bag SK. In Silico identification of SNP diversity in cultivated and wild tomato species: insight from molecular simulations. Sci Rep. 2016;6(1):38715.
82. Shirasawa K, Fukuoka H, Matsunaga H, Kobayashi Y, Kobayashi I, Hirakawa H, Isobe S, Tabata S. Genome-wide association studies using single nucleotide polymorphism markers developed by re-sequencing of the genomes of cultivated tomato. DNA Res. 2013;20(6):593–603.
83. Rounsley S, Marri PR, Yu Y, He R, Sisneros N, Goicoechea JL, Lee SJ, Angelova A, Kudrna D, Luo M, Affourtit J. De novo next generation sequencing of plant genomes. Rice. 2009;2:35–43.
84. Feuillet C, Leach JE, Rogers J, Schnable PS, Eversole K. Crop genome sequencing: lessons and rationales. Trends Plant Sci. 2011;16(2):77–88.
85. Edwards D, Batley J, Snowdon RJ. Accessing complex crop genomes with next-generation sequencing. Theor Appl Genet. 2013;126(1):1–11.
86. Huang X, Lu T, Han B. Resequencing rice genomes: an emerging new era of rice genomics. Trends Genet. 2013;29(4):225–32.
87. Edwards D, Batley J. Plant genome sequencing: applications for crop improvement. Plant Biotechnol J. 2010;8(1):2–9.
88. Li JY, Wang J, Zeigler RS. The 3,000 rice genomes project: new opportunities and challenges for future rice research. Gigascience. 2014;3:8.
89. Davey JW, Blaxter ML. RADSeq: next-generation population genetics. Brief Funct Genomics. 2010;9(5–6):416–23.
90. Shastry BS. SNP alleles in human disease and evolution. J Hum Genet. 2002;47(11):561–6.
91. Rohilla M, Singh N, Mazumder A, Sen P, Roy P, Chowdhury D, Singh NK, Mondal TK. Genome-wide association studies using 50 K rice genic SNP chip unveil genetic architecture for anaerobic germination of deep-water rice population of Assam, India. Mol Genet Genomics. 2020;295:1211–26.
92. Dadshani S, Mathew B, Ballvora A, Mason AS, Léon J. Detection of breeding signatures in wheat using a linkage disequilibrium-corrected mapping approach. Sci Rep. 2021;11(1):5527.
93. Sandhu SK, Pal L, Kaur J, Bhatia D. Genome wide association studies for yield and its component traits under terminal heat stress in Indian mustard (*Brassica juncea* L). Euphytica. 2019;215(11):188.
94. Wang S, Chen J, Zhang W, Hu Y, Chang L, Fang L, Wang Q, Lv F, Wu H, Si Z, Chen S. Sequence-based ultra-dense genetic and physical maps reveal structural variations of allopolyploid cotton genomes. Genome Biol. 2015;16:1–8.
95. Sun M, Li Y, Zheng J, Wu D, Li C, Li Z, Zang Z, Zhang Y, Fang Q, Li W, Han Y. A nuclear factor YB transcription factor, GmNFYB17, regulates resistance to drought stress in soybean. Int J Mol Sci. 2022;23(13):7242.
96. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas MA, Hanna M, McKenna A. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–8.
97. Batley J, Barker G, O'Sullivan H, Edwards KJ, Edwards D. Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence Tag data. Plant Physiol. 2003;132(1):84–91.
98. Vitte C, Bennetzen JL. Eukaryotic transposable elements and genome evolution special feature: analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proceedings of the National Academy of Science. 2006;103(7):17638–43.
99. Wakeley J. Substitution-rate variation among sites and the estimation of transition bias. Mol Biol Evol. 1994;11(3):436–42.
100. Rosenberg MS, Subramanian S, Kumar S. Patterns of transitional mutation biases within and among mammalian genomes. Mol Biol Evol. 2003;20(6):988–93.
101. Rahman M, Hoque A, Roy J. Linkage disequilibrium and population structure in a core collection of *Brassica napus* (L). PLoS ONE. 2022;17(3):e0250310.
102. Mantello CC, Cardoso-Silva CB, da Silva CC, de Souza LM, Scaloppi Junior EJ, de Souza Gonçalves P, Vicentini R, de Souza AP. *De Novo* assembly and transcriptome analysis of the rubber tree (*Hevea brasiliensis*) and SNP markers development for rubber biosynthesis pathways. PLoS ONE. 2014;9(7):e102665.
103. Yang H, Wei CL, Liu HW, Wu JL, Li ZG, Zhang L, Jian JB, Li YY, Tai YL, Zhang J, Zhang ZZ. Genetic divergence between *Camellia sinensis* and its wild relatives revealed via genome-wide SNPs from RAD sequencing. PLoS ONE. 2016;11(3):e0151424.
104. Raizada A, Souframanien J. Transcriptome sequencing, *de Novo* assembly, characterisation of wild accession of blackgram (*Vigna mungo* Var. *silvestris*) as a rich resource for development of molecular markers and validation of SNPs by high resolution melting (HRM) analysis. BMC Plant Biol. 2019;19:1–6.
105. Luo Z, Brock J, Dyer JM, Kutchan T, Schachtman D, Augustin M, Ge Y, Fahlgren N, Abdel-Haleem H. Genetic diversity and population structure of a *Camelina sativa* spring panel. Front Plant Sci. 2019;10:425924.
106. Shen JC, Rideout WM III, Jones PA. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. Nucleic Acids Res. 1994;22(6):972–6.

107. Zhao H, Li Q, Li J, Zeng C, Hu S, Yu J. The study of neighboring nucleotide composition and transition/transversion bias. Sci China Ser C: Life Sci. 2006;49:395–402.
108. Agrovskii BS, Vorob'ev VV, Gurvich AS, Pokasov VV, Ushakov AN. Intensity fluctuations of pulsed laser radiation during thermal self-interaction in a turbulent medium. Kvantovaya Elektronika. 1980;7(3):545–52.
109. Yates CM, Sternberg MJ. Proteins and domains vary in their tolerance of non-synonymous single nucleotide polymorphisms (nsSNPs). J Mol Biol. 2013;425(8):1274–86.
110. Flint-Garcia SA, Thornsberry JM, Buckler IVES. Structure of linkage disequilibrium in plants. Annu Rev Plant Biol. 2003;54(1):357–74.
111. Otyama PI, Wilkey A, Kulkarni R, Assefa T, Chu Y, Clevenger J, O'Connor DJ, Wright GC, Dezern SW, MacDonald GE, Anglin NL. Evaluation of linkage disequilibrium, population structure, and genetic diversity in the US peanut mini core collection. BMC Genomics. 2019;20:1–7.
112. Vos PG, Paulo MJ, Voorrips RE, Visser RG, van Eck HJ, van Eeuwijk FA. Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. Theor Appl Genet. 2017;130:123–35.
113. Huang Y, Li J, Li W, Han F. Integrative GWAS and eQTL analysis identifies genes associated with resistance to *Vibrio harveyi* infection in yellow drum (*Nibea albiflora*). Front Mar Sci. 2024;11:1435469.
114. Nair RR, Sasikumar B. Polyploidy in a cultivar of black pepper (*Piper nigrum* L.) and its open pollinated progenies. Cytologia. 1993;58(1):27–31.
115. Sasikumar B, George JK, Ravindran PN. Breeding behaviour of black pepper. Indian J Genet Plant Breed. 1992;52(1):17–21.
116. Chao S, Dubcovsky J, Dvorak J, Luo MC, Baenziger SP, Matnyazov R, Clark DR, Talbert LE, Anderson JA, Dreisigacker S, Glover K. Population-and genome-specific patterns of linkage disequilibrium and SNP variation in spring and winter wheat (*Triticum aestivum* L). BMC Genomics. 2010;11:1–7.
117. Uba CU, Oselebe HO, Tesfaye AA, Abtew WG. Association mapping in Bambara groundnut [*Vigna subterranea* (L.) Verdc.] reveals loci associated with agro-morphological traits. BMC Genomics. 2023;24(1):593.
118. Zhao T, Wu H, Wang X, Zhao Y, Wang L, Pan J, Mei H, Han J, Wang S, Lu K, Li M. Integration of eQTL and machine learning to dissect causal genes with pleiotropic effects in genetic regulation networks of seed cotton yield. Cell Rep. 2023;42(9).
119. Martínez-García PJ, Mas-Gómez J, Wegrzyn J, Botía JA. Bioinformatic approach for the discovery of *cis*-eQTL signals during fruit ripening of a woody species as grape (*Vitis vinifera* L). Sci Rep. 2022;12(1):7481.

Publisher's note