HortiS (2024) 41(3):106-115
http://doi.org/10.16882/HortiS.1538865

RESEARCH PAPER

HORTICULTURAL STUDIES

QTL-Seq: Rapid, Cost-Effective, and Reliable Method for QTL Identification

Yasin TOPCU[1] Manoj SAPKOTA[2] Serkan AYDIN[1]*

[1] Batı Akdeniz Agricultural Research Instiute, 07100, Antalya, Türkiye
[2] Cornell University Cornell Institute of Biotechnology Breeding Insight, Ithaca, New York, USA

Article History
Received 28 June 2024
Accepted 28 August 2024
First Online 7 September 2024

Corresponding Author
E-mail: serkan.aydin@tarimorman.gov.tr

Keywords
Bulk-segregant analysis
Candidate genes
Genetic variations
QTL-mapping
Whole-genome sequencing

Abstract

QTL-seq is a powerful method that integrates whole-genome sequencing (WGS) with bulk-segregant analysis to rapidly and reliably identify quantitative trait loci (QTLs) associated with specific traits. This approach significantly advances traditional QTL mapping by eliminating the need for genome wide DNA markers such as SSR, RFLP, and INDELs, which are typically used in linkage-based QTL mapping. Instead, QTL-seq leverages WGS to detect all genetic variations such as SNPs, Indels, and Structural Variants across the entire genome, providing a comprehensive resource for marker development in marker-assisted selection. The QTL-seq process begins with the creation of genetically diverse mapping populations, such as F_2 or RILs, followed by detailed phenotypic characterization. DNA from plants exhibiting similar phenotypes is pooled into bulk groups and sequenced, allowing for cost-effective and efficient QTL identification. Identified QTLs can be further validated through fine mapping using recombinant screenings and progeny testing, leading to the identification of candidate genes associated with traits of interest. In this study, we outline a user-friendly QTL-seq pipeline, from sequencing to data visualization to demonstrate its practical application. While the manuscript primarily focuses on describing the pipeline, we also conducted a case study analysis with real data to showcase its effectiveness. Our work contributes to the broader understanding of QTL-seq applications and offers practical recommendations for optimizing this method in future breeding programs.

1. Introduction

The current world population of 8.1 billion people as of May 2024 is estimated to reach 9.8 billion by 2050, hence humanity has to find sustainable ways to feed an extra 1.8 billion mouths (UN DESA, 2017). This situation underscores the urgent need for innovative agricultural practices, improved crop varieties with superior yield and resistant to biotic and abiotic stresses. Moreover, the issue is compounded by the gradual reduction in the amount of land available for agriculture (Godfray et al., 2010). In crop plants, many agronomically important traits such as yield, grain size, fruit weight, and plant height are governed by the collective effects of several genes with smaller effects called as quantitative trait loci or QTLs (Falconer, 1996). The QTL-identification is an arduous task yet of paramount importance for genetic enhancement of many important crops. Once these QTLs are identified, the next step is the integration of favorable alleles of QTLs into elite germplasm

mostly via backcrossing with the help of marker assistant selection (Collard and Mackill, 2008; Ribaut and Hoisington, 1998). One of the oldest yet reliable QTL-mapping approaches was linkage-based QTL mapping, in which DNA markers are tightly linked to targeted QTL. However, limitations in linkage mapping such as a restricted number of DNA markers, low marker density across the entire genome, long duration required for developing mapping populations, and difficulty in capturing all the recombination events, presence of heterogeneity in early generations (Abdurakhmonov and Abdukarimov, 2008; Madhusudhana, 2015) prompted researchers to seek alternative rapid, cost-effective, and reliable methods. QTL-seq was introduced by Takagi et al. (2013) more than a decade ago and offered as an alternative tool that may overcome these above-mentioned hurdles. This method simply relies on the advantages of next generation sequencing and bulk segregant analysis (BSA). The BSA method involves selecting individuals with extreme phenotypes from a segregating population, after which the DNA from these selected plants is pooled together into two separate bulks based on the phenotype. Each bulk is expected to be genetically identical within the regions linked to the target trait but different from the other bulk in these regions. This genetic difference between the two bulks is used to identify markers associated with the trait of interest. Essentially, the two pooled DNA samples are genetically identical (monomorphic) except for the regions linked to the trait, where they exhibit genetic dissimilarities (heterozygosity). The advances in whole-genome sequencing have opened a new era for plant breeders. This is mostly because several accessions have been re-sequenced and high-quality reference genomes for many crops such as tomato (Tomato Genome Consortium, 2012), maize (Jiao et al., 2017), rice (Kawahara et al., 2013), soybean (Schmutz et al., 2010), arabidopsis (Cheng et al., 2017) have become available over the past years. Another key component of the QTL-seq is BSA, which is introduced early in 1990s to map a downy mildew resistance in lettuce (Michelmore et al., 1991). In this method, individuals displaying extreme phenotypes are selected from a segregating population, after which the DNAs from these plants are bulked together. Within each pool, the plants are assumed to be genetically identical for a target region, but the pools themselves are dissimilar, variants used for developing markers are polymorphic and highly associated with the trait of interest (Takagi et al., 2013; Wang and Wang, 2023). In other words, two pooled DNA samples exhibit genetic dissimilarities solely within the targeted region, appearing heterozygous and monomorphic for all other regions. Even though BSA offers numerous advantages, genotyping of each marker mostly based on restriction fragment length polymorphism (RFLP) or simple sequence repeat (SSR) for the two bulked DNAs is still a laborious and limiting factor. In contrast to RFLP and SSR commonly used in the past, single nucleotide polymorphisms (SNPs) have numerous advantages due to their abundance, high-throughput genotyping capabilities, cost-effectiveness, and genome-wide distribution (International Rice Genome Sequencing Project, 2005; Nelson et al., 2004; Seeb et al., 2011; Singh et al., 2013). Therefore, BSA equipped with next-generation sequencing is capable of rapid, cost-effective, and reliable QTL mapping in various crops. To date, numerous traits have been mapped and utilized in plant breeding studies. Some of these traits were summarized in Table 1.

The main goal of this research is to present a comprehensive and user-friendly QTL-seq pipeline that encompasses all stages from sequencing to data visualization. By leveraging the methodology and data from Takagi et al. (2013), we aim to provide a clear and practical framework for implementing QTL-seq in plant breeding. Through a detailed case study analysis, we demonstrate the pipeline's effectiveness and offer insights for optimizing this approach, thereby advancing the application of QTL-seq in future breeding programs.

2. Material and Method

2.1. DNA extraction procedures and library preparation for sequencing

The DNA isolation and library preparation determines the success of the following steps. Hence, a high-quality DNA (high molecular weight and contaminant-free such as polysaccharides or phenolics) must be extracted with kits such as DNeasy Plant Mini Kit (Qiagen, Valencia, California, USA), Genomic DNA Purification Kit (Thermo Scientific™ Waltham, Massachusetts, USA), and Quick-DNA Plant/Seed 96 Kit (Zymo Research, Irvine, California, USA). Before NGS library preparation, it is essential to quantify both the quality and quantity of DNA from the selected individuals using NanoDrop ND-1000 spectrophotometer (Thermo Scientific) to ensure that the UV absorbance A260/A280 ratio falls within the range of 1.8 and 2.0 and A260/A230 ratio ≥ 1.5). Moreover, Qubit 2.0 Fluorimeter (Invitrogen, Carlsbad, CA, USA) could also be employed for the same reason. With respect to library preparation, NEBNext Ultra™ II DNA Library Prep Kit (New England Biolabs, USA) in conjunction with barcoded primers from the NEBNext® Multiplex Oligos obtained from Illumina kits (New England Biolabs, USA) could be used.

2.2. Comparative variant analysis

Whole genome sequencing can be performed using platforms such as the Illumina NextSeq 550,

Table 1. Summary of QTL-seq studies.

Crop	Trait of interest	Population size	Generations	QTL interval Mb	Reference
Rice	*Magnaporthe oryzae* (rice blast) resistance	n=241	RILs	Chr 6 2.39 to 4.39	Takagi et al. (2013)
	Seedling vigor	n=531	F_2	Chr 3 36.21 to 37.31	Takagi et al. (2013)
	Salt tolerance	n=199	$F_{2:3}$	Chr 7 20.16 to 24.33	Lei et al. (2020)
	Grain length and weight	n=176	NIL-F_2	Chr 5 15.00 to 20.00	Yaobin et al. (2018)
Cucumber	Early flowering	n=232	F_2	Chr 1 22.86 to 26.31	Lu et al. (2014)
	Pre-harvest sprouting	n=298	F_2	Chr 4 7.30 Mb[a]	Cao et al. (2021)
				Chr 5 0.15 Mb	
Tomato	Heat-tolerance	n=200	F_2	Chr 1 23.80 to 63.52	Wen et al. (2019)
				Chr 2 38.98 to 40.85	
				Chr 7 10.08 to 52.20	
	Fruit weight	n=100	F_2	Chr 1 12.48 to 51.58	Illa-Berenguer et al. (2015)
	Fruit weight	n=100	F_2	Chr 11 49.73 to 51.35	
	Fruit weight	n=200	F_2	Chr 03 60.86 to 61.72	
	Locule number			Chr 2 33.67 to 35.30	
	Locule number			Chr 5 3.25 to 3.98	
	Locule number			Chr 6 41.16 to 43.93	
	Blossom-end rot	n=192	F_2	Chr 3 54.21 to 59.89	Topcu et al. (2021)
				Chr 11 48.13 to 52.12	
	Yellow shoulder disorder	n=192	F_2	Chr 1 21.36 to 55.92	Topcu (2024)
				Chr 4 30.57 to 53.50	
				Chr 11 51.33 to 53.26	
Chickpea	Seed weight	n=221	F_4	Chr 1 0.84 to 0.87	Das et al. (2015)
Groundnut	Rust resistance	n=268	RIL	Chr A03 131.60 to 134.66 Mb	Pandey et al. (2017)
	Late leaf spot resistance	n=268	RIL	Chr A03 131.67–134.65 Mb	
Melon	Stigma color	n=150	F_2	Chr 6 141.48–152.83 cM	Qiao et al. (2021)
				Chr 8 19.71–57.33 cM	
Peanut	Seed weight	n=242	RIL	Chr A05 101.70−111.64 Mb	Wang et al. (2022)
				Chr B02 103.90−111.75 Mb	
				Chr B06 0.30−50.22 Mb	
Maize	Semi-dwarfism	n=533	F_2	Chr 9 111.07 to 124.56 Mb	Chen et al. (2018)
Soybean	Two-seed pod length		BC_3F_{2-n}	Chr03 0.50 to 4.76 Mb	Xie et al. (2021)
				Chr11 3.38 to 7.06 Mb	
				Chr12 9.72 to 11.25 Mb	

[a]Results were given as interval.

1000, and 2000, which utilize paired-end 150 base pairs (bp) (PE150) flow cells. Once sequencing procedure is finished, the raw fastq.gz files can be downloaded directly from the sequencing webpage using the "*wget [option] [URL]*". Before proceeding with further analysis, the FASTQ files are suggested to be filtered and trimmed, which can be done using Trim Galore (version 0.6.5, https://github.com/FelixKrueger/TrimGalore) to ensure a minimum quality value of 28. For this purpose, the following command "trim_galore --paired file_R1.fastq.gz file_R2.fastq.gz --quality 28 --fastqc --stringency 3 --length 60 --illumina" could be used, in which "--quality 28" removes low-quality ends from reads based on the phred score threshold of 28, "--fastqc" runs the FastQC in the default mode on the FastQ files once trimming is completed, "--paired" specifies the paired sequencing files, "--illumina" trims the first 13bp of the Illumina universal adapter 'AGATCGGAAGAGC', "--length 60" discards reads that became shorter than 60bp, "--stringency 3" enables that a minimum of 3 base pairs of the adapter must be present for it to be trimmed. The next step involves aligning the remaining high-quality reads to the reference genome which can be downloaded from public databases using "wget". This reference genome can either be one of the parental accessions to be sequenced along with the bulks or a high-quality reference genome. Before aligning with the bowtie2 (Version 2.4.1) (Langmead and Salzberg, 2012), or SpeedSeq (Chiang et al., 2015), reference genome should be indexed using the "bowtie2-build reference_sequence.fasta index_name" where reference_sequence.fasta is the reference genome fasta file to be indexed, and index_name is the output name. After indexing is done, the aligning can be performed using the following command line "bowtie2 -p 8 n -x index_name -1 file_R1.fastq.gz -

2 file_R2.fastq.gz -S output.sam*"*. In this command line, "-p" is the number (8) of processors/threads used, "-x" is the genome index, "-1 file_R1.fastq.gz" is the file of first paired end read, "-2 file_R2.fastq.gz" is the file of second paired end read, and "-S output.sam" is the output alignment in sam format. Next, the "output.sam" files need to be converted to BAM files using samtools (version 1.16.1) (Li and Durbin, 2009). To achieve this step, the following command line "samtools view -@ 10 -bS output.sam > output.bam" can be utilized. While "-@ 10" defines the number of threads which in this case is 10, -bS defines the output in the BAM format and ignores the compatibility with previous samtools versions. This step is followed by sorting of the bam files using "samtools sort -@ 10 -m 3G output.bam -o output_sorted.bam", in which "-m*"* defines the maximum required memory per thread to be used and "*-o" writes* the final sorted output. Upon indexing the sorted bam files with following command "samtools index output_sorted.bam" Picard tools (Picard version 2.27.5) (https://broadinstitute.github.io/picard/) will be employed to replace read groups and identify duplicate reads. To achieve this step, the following command "java -jar $EBROOTPICARD/picard.jar AddOrReplaceReadGroups --INPUT= output_ sorted.bam --OUTPUT=output_sorted.RG.bam --RGID=4 --RGSM=output --RGLB=output --RGPL=ILLUMINA --RGPU=ignore" and "java -jar $EBROOTPICARD/picard.jar MarkDuplicates INPUT= output_sorted.RG.bam OUTPUT= output_sorted_mkdupl.RG.bam METRICS_FILE= output_sorted_mkduplMetrics.txt*"* can be used. While *"*AddOrReplaceReadGroups" consolidates all the reads in a file under a singular new read-group, "MarkDuplicates" locates, and tags duplicate reads in a BAM-files. The command "*java -jar $EBROOTPICARD/picard.jar*" utilizes Java to run a JAR file named picard.jar, which is located in the directory specified by the environment variable $EBROOTPICARD. In the command lines, "*--INPUT*" shows Input file, "*--OUTPUT*" designates Output file, "*--RGID*" defines Read-Group ID, "*--RGSM*" displays Read-Group sample name, "*--RGLB*" denotes Read-Group library, "*--RGPL*" illustrates Read-Group platform (such as ILLUMINA and SOLID) and finally "*--METRICS_FILE*" specifies the file where metrics about the duplicates will be written. These metrics may contain data such as the count of identified duplicates, their respective locations, and other pertinent statistical information. After completing the previous step, the next step involves indexing the sorted and marked BAM file. This is accomplished by executing the command "*samtools index output_sorted_mkdupl.RG.bam*".

2.3. Variant calling

The variant calling is of utmost importance since QTL-seq heavily depends on the variance between created bulks. Hence, to get reliable results and enhance the accuracy, we must annotate potential insertions/deletions (INDELs) or misalignments accurately. The first step in variant calling pipeline begins with reference genome indexing. The reference genome can be indexed with "SAMtools" developed by Li and Durbin (2009) using the "*Samtools faidx reference_sequence.fa*" command. The INDEL realignment is performed utilizing the Genome Analysis Toolkit (GATK, Version 3.8-1) (McKenna et al., 2010) by following the commands " -T RealignerTargetCreator" which identifies what regions need to be realigned and "-T IndelRealigner" that performs the actual realignment. Both determine false positive SNPs and perform a local realignment in a sequencing dataset. While the first command, "*java -Xmx150g -jar $EBROOTGATK/GenomeAnalysisTK.jar -T RealignerTargetCreator -R reference_sequence.fa -I output_sorted_mkdupl.RG.bam -o output _intervals.list*" creates a list of target intervals for the following step, the second command "*java -Xmx150g –jar$EBROOTGATK/GenomeAnalysis TK.jar -T IndelRealigner -R reference_sequence.fa -I output_sorted_mkdupl.RG.bam -targetIntervals output_intervals.list -o output_realigned_reads .bam*" executes the real realignment of reads based on the target intervals. In both commands abovementioned, "*-Xmx*" defines the memory to be allocated, "*-R"* designates the reference genome to be used, "*-I*" describes the input BAM file containing aligned reads, "*-targetIntervals"* designates the interval file generated from the RealignerTargetCreator step and finally "*-o"* specifies the output file where the information about potential realignment sites will be stored. Before proceeding to the final step of variant calling, the output of the previous command (*output_realigned_reads.bam*) needs to be indexed. The final command in variant calling step utilizes GATK to call haplotypes from aligned reads in the "*output_realigned_reads"* BAM file. The command is "*java -Xmx150g -jar $EBROOTGATK/GenomeAnalysisTK.jar -T HaplotypeCaller -nct 10 -R reference_sequence.fa -I output_realigned_reads.bam –emitRef Confidence GVCF --variant_index_type LINEAR --variant_index_parameter 128000 -o raw_ variants_gvcf.vcf*". In the command line, "*-T HaplotypeCaller"* specifies the tool as HaplotypeCaller, which identifies potential variants. Furthermore, "*-nct 10"* indicates the number of CPU threads to use for parallel execution, "*-R"* refers to reference genome fasta sequence, "*-I*" designates Input BAM file, the "emitRefConfidence" option emits reference confidence scores for each site in the (Genomic Variant Call Format) *GVCF* file, providing information about the likelihood that a particular reference allele is actually present at a given genomic position. "Variant_index_type *LINEAR*" parameter specifies the indexing strategy as LINEAR, meaning that variants are indexed sequentially according to their genomic position for

the output GVCF. The final parameter required in GATK versions older than 3.4 is "*variant_index_parameter 128000*" indicating the size of the bins used in the linear indexing strategy.

2.4. Combining variant callings and filtering

In QTL-seq studies, two representative bulks are typically created to identify genomic regions associated with the trait. In the following command, the variant calls that were previously done for each bulk separately are merged into a single VCF file for the downstream analysis. The command is "*java -Xmx150g -jar $EBROOTGATK/GenomeAnalysisTK.jar -T GenotypeGVCFs -R reference_sequence.fa --variant raw1_variants_gvcf.vcf --variant raw2_variants_gvcf.vcf -nt 10 -o merged.vcf".* While *"-T GenotypeGVCFs"* parameter specifies the tool in GATK being used to perform joint genotyping that involves combining variant calls from multiple samples on GVCF files generated by HaplotypeCaller, "*--variant"* parameter designates which files need to be merged. The following command "*java -Xmx150g -jar $EBROOTGATK/GenomeAnalysisTK.jar -T SelectVariants -R reference_sequence.fa -V merged.vcf -selectType SNP -o SNPs.vcf*" is used to extract SNPs from merged variant calling VCFs, in which "*-T SelectVariants*" indicates the tool being used in GATK that allows selection of specific variants whereas "*-selectType SNP" or "--select-type-to-include SNP"* selects SNP variant from the supplied VCF file, designated by "-V". Once SNPs have been selected, the subsequent steps involve identifying and flagging SNPs with poor quality based on genotype quality, read depth, allele frequency, and various annotation scores, and then filtering them out. This filtering step is crucial to identify high quality SNPs that can be converted into genotyping markers such as KASP (Kompetitive Allele-Specific PCR). To tag low quality SNPs, the following command can be used "*java -Xmx150g -jar $EBROOTGATK/GenomeAnalysisTK.jar -T VariantFiltration -R reference_sequence.fa -V SNPs.vcf --filterExpression "QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0" --filterName "Default_recommended" -o Filtered_snps.vcf".* In the command line, "*-T VariantFiltration*" indicates the tool in GATK being used to filter variants, *"-V SNPs.vcf"* shows the input VCF file containing SNPs that need to be filtered, "*--filterExpression "QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0"* defines the filtering criteria based on QD < 2.0: Variant Quality by Depth (QD) less than 2.0, FS > 60.0: FisherStrand (FS) greater than 60.0, MQ < 40.0: Mapping Quality (MQ) less than 40.0, MQRankSum < -12.5: Mapping Quality Rank Sum Test less than -12.5, ReadPosRankSum < -8.0: Read Position Rank Sum Test less than -8.0. Furthermore, "--filterName "Default_recommended" defines the name of the filter to be applied to variants. The next step involves filtering using VCFtools (version 0.1.16) (Danecek et al., 2011). To keep only high quality SNPs and the following command "vcftools --vcf *Filtered_snps.vcf* --remove-filtered-all --recode --max-missing 1 -c > Filtered_passed_snps.vcf" is performed, in which "vcftools" defines which tools to be used in VCFtools, *"--vcf Filtered_snps.vcf"* specifies the input VCF file containing variants that need to be filtered, *"--remove-filtered-all"* removes all variants that have been flagged as filtered by previous filtering steps, *"--recode"* forces VCFtools to output the filtered variants into a new VCF file as *"Filtered_passed_snps.vcf*" designated in the command line. The last criteria are "*-max-missing 1*" that filters variants where more than one sample has missing data, and "-c" defines the output as compressed VCF files. The steps described above are summarized in Figure 1.

Additionally, a master script detailing each step is provided in Supplemental File 1. Using this script we re-analyzed the QTL-seq data (Takagi et al., 2013) which identified a QTL located in the 2.39 to 4.39 Mb region on chromosome 6, associated with partial resistance to *Magnaporthe oryzae,* the causal agent of rice blast disease in the rice. The final VCF file that shows the SNPs and INDELs between R-bulk (Mainly Nortai-type genomic segments) and S-bulk (Mainly Hitomebore-type genomic segments) was given in (Supplemental File 2).

3. Results and Discussion

The last step in the QTL-seq pipeline is visualizing the SNP allele frequencies or SNP-indexes along the genome and identify QTL regions associated with the trait of interest. This visualization can be done using an R package called QTLseqr (Mansfeld and Grumet, 2018). Since, the R package requires a tabular file format, we need to convert VCF file that has the SNP variants identified between two bulks into tabular format using following command "*java -jar $EBROOTGATK/GenomeAnalysisTK.jar -T VariantsToTable -R reference_sequence.fa -V Filtered_passed_snps.vcf -F CHROM -F POS -F REF -F ALT -GF AD -GF DP -GF GQ -GF PL -o QTL-seqr.table*". While *"-T VariantsToTable"* in the command line designates the tool that converts the variant information from VCF format to a tabular format, "*-R"* defines the reference fasta, "*-V*" specifies the input VCF file containing the filtered SNP variants. Further, *"-F CHROM -F POS -F REF -F ALT"* specifies the components such as chromosome, position, reference allele, and alternate allele of each variant to be included in the output table. Finally, *"-GF AD -GF DP -GF GQ -GF PL"* defines the genotype fields (GF) to be included in the output table such as allelic depths (AD), total read depths (DP), genotype quality (GQ), and

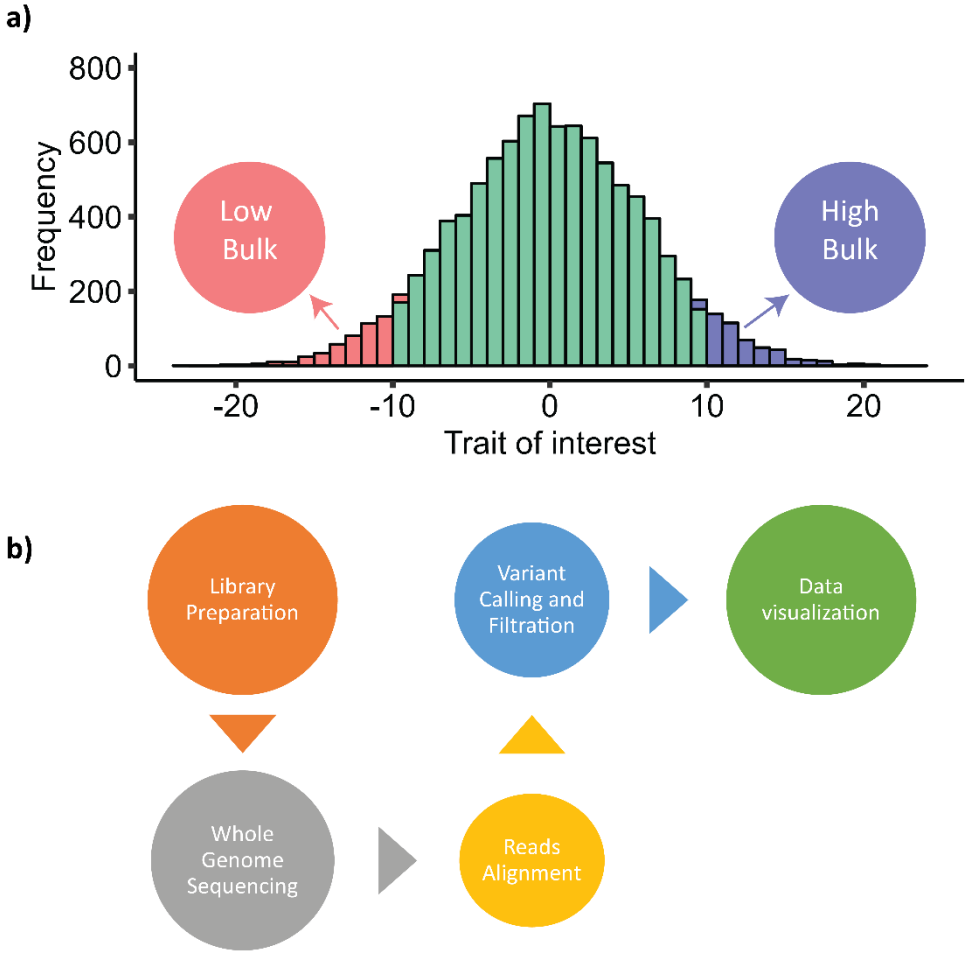

Figure 1. General outline of a QTL-seq script. a) The phenotypic distribution of a hypothetical mapping population. A dataset of 10,000 continuous values was generated using a normal distribution with a mean of 0 and a standard deviation of 6. A seed (set.seed (123)) was used to ensure reproducibility. The 5th percentile of the data defined the lower extreme values (low bulk), while the 95th percentile defined the upper extreme values (high bulk). b) The workflow began with library preparation for each bulk, followed by whole genome sequencing to generate raw reads. These reads were then aligned to a reference genome, and variant calling was used to identify genetic variants (SNPs). The process concluded with data visualization for the analysis and presentation of the results.

phred-scaled likelihoods (PL) for each genotype. The corresponding "*QTL-seqr.table"* file for the rice data was also given as Supplemental File 3. In the QTL-seqr package, further filtering steps can be used based on reference allele frequency, maximum total depth, minimum total depth, sample depth and genotype quality. After desired filtering criteria are met, the "runQTLseqAnalysis() function "can be implemented with some minor changes to original pipeline of (Takagi et al., 2013). The modified "R" script that contains further filtering and QTL-visualization steps was given in Supplemental File 4. We successfully mapped the fungal rice blast disease QTL, *qPi-nor1(t)*, with our script and validated the results obtained by Takagi et al. (2013). The rice blast disease trait, which was used to test our QTL-seq analysis, was estimated to exhibit moderate broad-sense heritability (54.16%) previously (Salleh et al., 2022), underscoring the genetic basis of this trait. The corresponding QTL-seq results were given in Supplemental Figure 1. We identified two QTLs associated with the blast resistance Figure 2. Although the previously identified QTL on chr 6, *qPi-nor1(t)*, was located between 2.39–4.39 Mb ($P < 0.01$), we defined the border of *qPi-nor1(t)* as 2.50- 5.39 ($P < 0.01$) Figure 2a. In addition, we identified another QTL (named as *blast9.1)* on chr9, which locates between 9.28-10.20 ($P < 0.05$) Figure 2b.

The power of next generation sequencing, especially the advances in long and short read sequencing with reduced costs, has opened a new era for QTL mapping and dramatically changed the way of crop breeding practices and genetic studies in various organisms (Varshney et al., 2009; Kim et al., 2016; Varshney et al., 2020). Once more plant genome assemblies along with complete annotations are readily available in plant science, numerous QTL mapping methods have been proposed, and several innovative concepts have been introduced to map QTLs (Bazakos et al., 2017; Wang & Han, 2022). SHOREmap, introduced by Schneeberger et al. (2009) can be seen a corner stone as it was one of the original approaches that

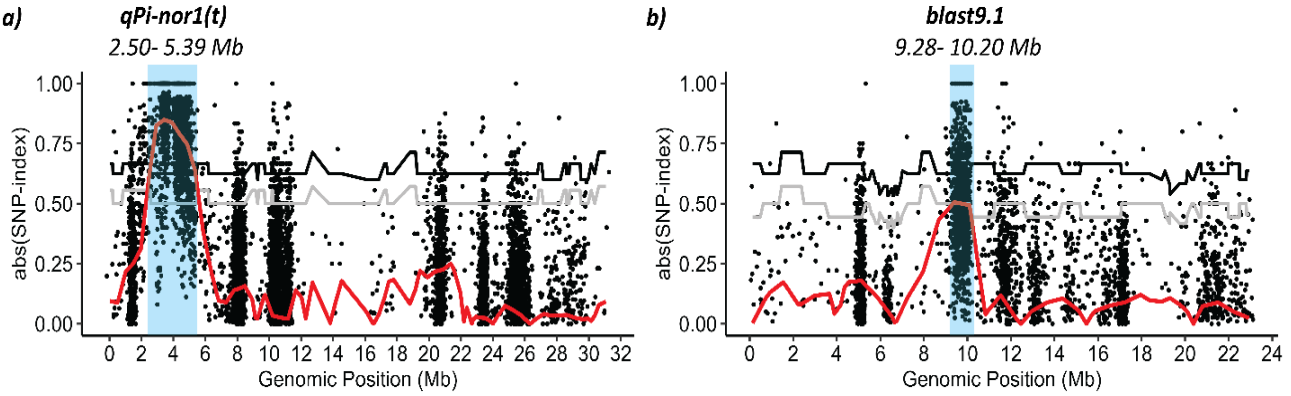

Figure 2. QTL-seq identifies qPi-nor1(t) and blast9.1 QTLs associated with Magnaporthe oryzae (rice blast disease) resistance on a) chr 6 and b) chr 9, respectively. The tricube-smoothed absolute Δ(SNP-index) is shown in red, while confidence intervals of $P < 0.05$ and $P < 0.01$ are depicted in grey and black lines, respectively. The X-axis represents the genomic position in megabases (Mb), and the Y-axis shows the absolute Δ(SNP-index) values. The blue shaded areas on chr6 and chr9 show the QTL-intervals for qPi-nor1(t), and blast9.1 responsible for the rice blast disease.

integrates whole genome resequencing and phenotyping in a large pool of recombinants. Moreover, BSA equipped with genome analysis using microarray-based genotyping or massively parallel sequencing was another pioneering approach that was focusing on mapping of QTLs with minor effects (Ehrenreich et al., 2010). This method, called as Extreme QTL mapping (X-QTL), has three main components. Creating of a large segregating population and selecting progenies from this large mapping population with extreme trait values for comprehensive analysis are of foremost importance for the method (Ehrenreich et al., 2010). The last component is microarray-based genotyping or massively parallel sequencing of pooled allele frequencies. In a similar manner, Next Generation Mapping (NGM) approach, introduced by Austin et al. (2011), detects mutations by sequencing a small pooled F_2 population, without prior knowledge of genetic analysis. Following these ideas, Abe et al. (2012) developed MutMap, a method based on whole-genome resequencing of pooled DNA from a segregating plant population. While MutMap offers significant utility, crop breeding has predominantly relied on QTL breeding, leveraging genetic variations among diverse cultivars and species. Hence, examining QTL variations in natural variants is highly essential for identifying important alleles of genes controlling essential agronomic traits and enhancing breeding efforts. By combining the power of next-generation sequencing with BSA, Takagi et al. (2013) proposed the QTL-seq method as reliable, quick and most importantly cost-effective approach to QTL mapping, leading the way for significant enhancements in crop improvements and sustainable agriculture. Until now, numerous agronomically important traits have been successfully mapped using QTL-seq, and researchers were able to rapidly fine-map and ultimately identify candidate genes in many agronomically important crops (Table 1).

The effectiveness of QTL-seq is mostly determined by the population size, the heritability of the trait, the percentage of plants chosen for each bulk and population structure (e.g., F_2, F_5, NILs or RIL). In addition to these factors, the nature of the trait whether it is governed by single major QTL or many QTLs with minor effects plays a crucial role. Moreover, read depth of the sequencing along with recombination frequency are also important factors. Furthermore, and more importantly, the inheritance of traits, including various forms such as complete dominance, incomplete dominance, overdominance, additive effects, recessive effects, and epistasis, plays a critical role in determining the success of QTL-seq (Takagi et al., 2013). The way these inheritance patterns exhibits in a given population can significantly impact the identification and mapping of QTLs. For example, additive effects allow for a more straightforward association between genotype and phenotype, while dominance and epistasis can complicate QTL detection. Additionally, gene-by-environment interactions (GxE) further influence trait expression, adding another layer of complexity to QTL-seq analysis. These genetic factors, along with the heritability of the traits, precision and depth of sequencing, size of the mapping population, and accuracy in phenotyping, are all crucial components that contribute to the identification of significant QTLs and understanding their effects across various genetic backgrounds and environmental conditions. Based on the previous studies, a minimum population size of 200 is mostly used for QTL mapping, although successful QTL identification has been achieved even with population sizes as small as 100 in tomato (Table 1). The second consideration is the percentage of individuals included in each bulk. Based on a study conducted by Takagi et al. (2013), it was recommended to bulk 10-15% of the population. Furthermore, the appropriate read depth for sequencing largely depends on factors such as the

generation of the population (F_2 vs F_7), genome size of the crop, and the genetic effects under consideration, such as dominance versus complete dominance. For F_2 populations, a minimum read depth of 10x to 20x is recommended, whereas even 5x read depth may suffice in the F_7 generation to detect codominant QTL. However, for QTLs exhibiting a dominance effect, it is advisable to have a read depth of at least 20x or higher in F_2 populations to ensure successful QTL identification (Takagi et al., 2013). Since its conceptualization and widespread adoption of the QTL-seq, several modifications or improvements have been implemented. To accelerate genetic mapping process, Wang et al. (2019) introduced "GradedPool-Seq" approach, in which individuals from F_2 population are assigned into three or more graded groups based on their phenotypic values. Once GradedPool-Seq is compared with the previous methods like MutMap, SHOREmap, Next-Generation Mapping, and QTL-seq, it has several advantages such as high-resolution genetic mapping (~400-kb) and detecting multiple QTLs along with the ability of evaluating multiple phenotypic characters in a single F_2 population. (Wang et al., 2019). "Modified QTL-seq," which is a novel strategy of NGS-BSA application, was introduced by Wang and Wang (2023). The main advantage of this method is multiple comparison analysis, which can effectively speed up QTL mapping for complex traits, thereby accelerating the breeding process in crops (Wang and Wang, 2023). Although QTL-seq and other modified approaches have various advantages, there are still concerns that may hinder successful QTL mapping using these methods (Ott et al., 2011; Slate, 2013; Ashton et al., 2017; Bazakos et al., 2017). These constraints encompass genetic basis of complex traits like epigenetic and epistatic factors, family based experimental designs, pooling errors in BSA, the potential omission of minor QTLs, the influence of environmental interactions, the prevalence of high rates of false positive SNP detection (Flint and Mott, 2001; Mackay, 2001; Clevenger et al., 2018). To address many of these challenges, the size of the mapping population plays a pivotal role as it is related to allele frequency and statistical power (Hamblin et al., 2011; Hong and Park, 2012). Previous studies employing QTL-seq have indicated an average population size of 241, suggesting a reasonable benchmark for future QTL investigations. However, adjustments to the population size should be made based on the specific trait under scrutiny especially to avoid Beavis effect and capture the minor QTL effects (Slate, 2013). Traits with high heritability may tolerate smaller population sizes, whereas traits with lower heritability may benefit from larger population sizes to enhance the detection of minor QTLs and narrow down QTL intervals early in the mapping process (Topcu et al., 2021). To minimize the errors in pooling, the phenotyping should be evaluated in controlled conditions and if it is possible in different environments to minimize the environment effects. Nevertheless, it's important to note that many of these concerns are relevant to other QTL mapping methods as well.

4. Conclusion

In conclusion, the QTL-seq method has demonstrated its effectiveness as a rapid, cost-effective, and reliable approach to QTL mapping across various contexts. This study provides a comprehensive overview of the entire process, from initial DNA isolation to data visualization, offering a valuable pipeline for researchers, particularly in the field of plant breeding.

References

Abdurakhmonov, I.Y., & Abdukarimov, A. (2008). Application of association mapping to understanding the genetic diversity of plant germplasm resources. *International Journal of Plant Genomics*, 2008: 574927.

Abe, A., Kosugi, S., Yoshida, K., Natsume, S., Takagi, H., Kanzaki, H., Matsumura, H., Yoshida, K., Mitsuoka, C., Tamiru, M., Innan, H., Cano, L., Kamoun, S., & Terauchi, R. (2012). Genome sequencing reveals agronomically important loci in rice using MutMap. *Nature Biotechnology*, 30(2): 174-178.

Ashton, D.T., Ritchie, P.A., & Wellenreuther, M. (2017). Fifteen years of quantitative trait loci studies in fish: challenges and future directions. *Molecular Ecology*, 26(6): 1465-1476.

Austin, R.S., Vidaurre, D., Stamatiou, G., Breit, R., Provart, N.J., Bonetta, D., Zhang, J., Fung, P., Gong, Y., Wang, P.W., McCourt, P., & Guttman, D.S. (2011). Next-generation mapping of Arabidopsis genes. *The Plant Journal*, 67(4): 715-725.

Bazakos, C., Hanemian, M., Trontin, C., Jiménez-Gómez, J.M., & Loudet, O. (2017). New strategies and tools in quantitative genetics: How to go from the phenotype to the genotype. *Annual Review of Plant Biology*, 68: 435-455.

Cao, M., Li, S., Deng, Q., Wang, H., & Yang, R. (2021). Identification of a major-effect QTL associated with pre-harvest sprouting in cucumber (*Cucumis sativus* L.) using the QTL-seq method. *BMC Genomics*, 22(1): 249.

Chen, Q., Song, J., Du, W.P., Xu, L.Y., Jiang, Y., Zhang, J., Xiang, X.L., & Yu, G.R. (2018). Identification and genetic mapping for rht-DM, a dominant dwarfing gene in mutant semi-dwarf maize using QTL-seq approach. *Genes & Genomics*, 40(10): 1091-1099.

Cheng, C.Y., Krishnakumar, V., Chan, A.P., Thibaud-Nissen, F., Schobel, S., & Town, C.D. (2017). Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *The Plant Journal*, 89(4): 789-804.

Chiang, C., Layer, R.M., Faust, G.G., Lindberg, M.R., Rose, D.B., Garrison, E.P., Marth, G.T., Quinlan, A.R., & Hall, I.M. (2015). SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat Methods*, 12(10): 966-968.

Clevenger, J., Chu, Y., Chavarro, C., Botton, S., Culbreath, A., Isleib, T.G., Holbrook, C.C., & Ozias-Akins, P. (2018). Mapping late leaf spot resistance in

peanut (*Arachis hypogaea*) using QTL-seq reveals markers for marker-assisted selection. *Frontiers in Plant Science*, 9.

Collard, B.C., & Mackill, D.J. (2008). Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1491): 557-572.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., & Group, G.P.A. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15): 2156-2158.

Das, S., Upadhyaya, H.D., Bajaj, D., Kujur, A., Badoni, S., Laxmi, Kumar, V., Tripathi, S., Gowda, C.L.L., Sharma, S., Singh, S., Tyagi, A.K., & Parida, S.K. (2015). Deploying QTL-seq for rapid delineation of a potential candidate gene underlying major trait-associated QTL in chickpea. *DNA Research*, 22(3): 193-203.

Ehrenreich, I.M., Torabi, N., Jia, Y., Kent, J., Martis, S., Shapiro, J.A., Gresham, D., Caudy, A.A., & Kruglyak, L. (2010). Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature*, 464(7291): 1039-1042.

Falconer, D.S. (1996). *Introduction to quantitative genetics*. Pearson Education, ISBN:8131727408, 9788131727409, 464 p., India.

Flint, J., & Mott, R. (2001). Finding the molecular basis of quatitative traits: successes and pitfalls. *Nature Reviews Genetics*, 2(6): 437-445.

Godfray, H.C.J., Beddington, J.R., Crute, I.R., Haddad, L., Lawrence, D., Muir, J.F., Pretty, J., Robinson, S., Thomas, S.M., & Toulmin, C. (2010). Food security: the challenge of feeding 9 billion people. *Science*, 327(5967): 812-818.

Hamblin, M.T., Buckler, E.S., & Jannink, J.L. (2011). Population genetics of genomics-based crop improvement methods. *Trends in Genetics*, 27(3): 98-106.

Hong, E.P., & Park, J.W. (2012). Sample size and statistical power calculation in genetic association studies. *Genomics Inform*, 10(2): 117-122.

Illa-Berenguer, E., Van Houten, J., Huang, Z., & van der Knaap, E. (2015). Rapid and reliable identification of tomato fruit weight and locule number loci by QTL-seq. *Theoretical and Applied Genetics*, 128(7): 1329-1342.

International Rice Genome Sequencing Project. (2005). The map-based sequence of the rice genome. *Nature*, 436(7052): 793-800.

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M.C., Wang, B., Campbell, M.S., Stein, J.C., Wei, X., & Chin, C.S. (2017). Improved maize reference genome with single-molecule technologies. *Nature*, 546(7659): 524-527.

Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., Ouyang, S., Schwartz, D.C., Tanaka, T., Wu, J., & Zhou, S. (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice*, 6: 1-10.

Kim, C., Guo, H., Kong, W., Chandnani, R., Shuang, L.-S., & Paterson, A.H. (2016). Application of genotyping by sequencing technology to a variety of crop breeding programs. *Plant Science*, 242*:* 14-22.

Langmead, B., & Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4): 357-359.

Lei, L., Zheng, H., Bi, Y., Yang, L., Liu, H., Wang, J., Sun, J., Zhao, H., Li, X., Li, J., Lai, Y., & Zou, D. (2020). Identification of a major QTL and candidate gene analysis of salt tolerance at the bud burst stage in rice (*Oryza sativa* L.) using QTL-Seq and RNA-Seq. *Rice*, 13(1): 55.

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

Lu, H., Lin, T., Klein, J., Wang, S., Qi, J., Zhou, Q., Sun, J., Zhang, Z., Weng, Y., & Huang, S. (2014). QTL-seq identifies an early flowering QTL located near Flowering Locus T in cucumber. *Theoretical and Applied Genetics*, 127(7): 1491-1499.

Mackay, T.F. (2001). The genetic architecture of quantitative traits. *Annual Review of Genetics*, 35(1): 303-339.

Madhusudhana, R. (2015). Linkage mapping. *Sorghum Molecular Breeding*, 47-70.

Mansfeld, B.N., & Grumet, R. (2018). QTLseqr: An R Package for bulk segregant analysis with next-generation sequencing. *The Plant Genome*, 11(2): 180006.

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.

Michelmore, R.W., Paran, I., & Kesseli, R. (1991). Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proceedings of the National Academy of Sciences*, 88(21): 9828-9832.

Nelson, M.R., Marnellos, G., Kammerer, S., Hoyal, C.R., Shi, M.M., Cantor, C.R., & Braun, A. (2004). Large-scale validation of single nucleotide polymorphisms in gene regions. *Genome Research*, 14(8): 1664-1668.

Ott, J., Kamatani, Y., & Lathrop, M. (2011). Family-based designs for genome-wide association studies. *Natural Review Genetics*, 12(7): 465-474.

Pandey, M.K., Khan, A.W., Singh, V.K., Vishwakarma, M.K., Shasidhar, Y., Kumar, V., Garg, V., Bhat, R.S., Chitikineni, A., Janila, P., Guo, B., & Varshney, R.K. (2017). QTL-seq approach identified genomic regions and diagnostic markers for rust and late leaf spot resistance in groundnut (*Arachis hypogaea* L.). *Plant Biotechnology Journal*, 15(8): 927-941.

Qiao, A., Fang, X., Liu, S., Liu, H., Gao, P., & Luan, F. (2021). QTL-seq identifies major quantitative trait loci of stigma color in melon. *Horticultural Plant Journal*, 7(4): 318-326.

Ribaut, J.M., & Hoisington, D. (1998). Marker-assisted selection: new tools and strategies. *Trends in Plant Science*, 3(6): 236-239.

Salleh, S.B., Rafii, M.Y., Ismail, M.R., Ramli, A., Chukwu, S.C., Yusuff, O., & Hasan, N.A. (2022). Genotype-by-environment interaction effects on blast disease severity and genetic diversity of advanced blast-resistant rice lines based on quantitative traits. *Frontiers in Agronomy*, 4.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., Xu, D., Hellsten, U., May, G.D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M.K., Sandhu, D., Valliyodan, B., & Jackson, S.A. (2010). Genome sequence of the palaeopolyploid soybean. *Nature*, 463(7278): 178-183.

Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A.H., Nielsen, K.L., Jørgensen, J.E., Weigel, D., & Andersen, S. U. (2009). SHOREmap: Simultaneous mapping and mutation identification by deep sequencing. *Nature Methods*, 6(8): 550-551.

Seeb, J.E., Carvalho, G., Hauser, L., Naish, K., Roberts, S., & Seeb, L.W. (2011). Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Molecular Ecology Resources*, 11(s1): 1-8.

Singh, N., Choudhury, D.R., Singh, A.K., Kumar, S., Srinivasan, K., Tyagi, R.K., Singh, N.K., & Singh, R. (2013). Comparison of SSR and SNP markers in estimation of genetic diversity and population structure of Indian rice varieties. *Plos One*, 8(12): e84136.

Slate, J. (2013). From beavis to beak color: a simulation study to examine how much qtl mapping can reveal about the genetic architecture of quantitative traits. *Evolution*, 67(5): 1251-1262.

Takagi, H., Abe, A., Yoshida, K., Kosugi, S., Natsume, S., Mitsuoka, C., Uemura, A., Utsushi, H., Tamiru, M., Takuno, S., Innan, H., Cano, L.M., Kamoun, S., & Terauchi, R. (2013). QTL-seq: Rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *The Plant Journal*, 74(1): 174-183.

Tomato Genome Consortium. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485(7400): 635-641.

Topcu, Y. (2024). Elucidating the genetic aspect of yellow shoulder disorder in tomato (*Solanum lycopersicum*) by QTL-seq and linkage mapping. *Scientia Horticulturae*, 332(1): 113225.

Topcu, Y., Sapkota, M., Illa-Berenguer, E., Nambeesan, S.U., & van der Knaap, E. (2021). Identification of blossom-end rot loci using joint QTL-seq and linkage-based QTL mapping in tomato. *Theoretical and Applied Genetics*, 134(9): 2931-2945.

UN DESA. (2017). *United Nations Department of Economic and Social Affairs/Population Division, World population prospects: The 2017 revision, key findings and advance tables (Working Paper No. ESA/P/WP/248)* [Statistics Non-fiction]. United Nations.

Varshney, R.K., Nayak, S.N., May, G.D., & Jackson, S.A. (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in Biotechnology*, 27(9): 522-530.

Varshney, R.K., Sinha, P., Singh, V.K., Kumar, A., Zhang, Q., & Bennetzen, J.L. (2020). 5Gs for crop genetic improvement. *Current Opinion in Plant Biology*, 56: 190-196.

Wang, C., & Han, B. (2022). Twenty years of rice genomics research: From sequencing and functional genomics to quantitative genomics. *Molecular Plant*, 15(4): 593-619.

Wang, C., Tang, S., Zhan, Q., Hou, Q., Zhao, Y., Zhao, Q., Feng, Q., Zhou, C., Lyu, D., Cui, L., Li, Y., Miao, J., Zhu, C., Lu, Y., Wang, Y., Wang, Z., Zhu, J., Shangguan, Y., Gong, J., & Han, B. (2019). Dissecting a heterotic gene through GradedPool-Seq mapping informs a rice-improvement strategy. *Nature Communications*, 10(1), 2982.

Wang, X., & Wang, G. (2023). Application of NGS-BSA and proposal of Modified QTL-seq. *Journal of Plant Biochemistry and Biotechnology*, 32(1): 31-39.

Wang, Z., Yan, L., Chen, Y., Wang, X., Huai, D., Kang, Y., Jiang, H., Liu, K., Lei, Y., & Liao, B. (2022). Detection of a major QTL and development of KASP markers for seed weight by combining QTL-seq, QTL-mapping and RNA-seq in peanut. *Theoretical and Applied Genetics*, 135(5): 1779-1795.

Wen, J., Jiang, F., Weng, Y., Sun, M., Shi, X., Zhou, Y., Yu, L., & Wu, Z. (2019). Identification of heat-tolerance QTLs and high-temperature stress-responsive genes through conventional QTL mapping, QTL-seq and RNA-seq in tomato. *BMC Plant Biology*, 19(1): 398.

Xie, J., Wang, Q., Zhang, Z., Xiong, X., Yang, M., Qi, Z., Xin, D., Zhu, R., Sun, M., Dong, X., Jiang, H., & Chen, Q. (2021). QTL-seq identified QTLs and candidate genes for two-seed pod length and width in soybean (*Glycine max*). *Plant Breeding*, 140(3): 453-463.

Yaobin, Q., Peng, C., Yichen, C., Yue, F., Derun, H., Tingxu, H., Xianjun, S., & Jiezheng, Y. (2018). QTL-Seq identified a major QTL for grain length and weight in rice using near isogenic F2 population. *Rice Science*, 25(3): 121-131.