Causal Mutation Screening in C. elegans based on WGS

Abstract: Objectives: To identify recessive mutation which causes nrd-3 protein localizing abnormally in mutagenized nematodes. Methods: After quality control and mutation calling of the whole genome sequencing data, candidate gene list was obtained by subtracting homozygous mutations of the nematodes. quality control and mutation calling of the whole genome sequencing data, candidate gene list was obtained by subtracting homozygous mutations of the Genetic markers were obtained by selecting mutations with normal sequencing depth and QD values. Statistic R was defined and plotted to enumerate and choose the interval with the highest R value as mapping Then annotate and identify causal mutation in that interval. Results: By analyzing complete genome sequencing data of less than 100 Twelve candidate genes were identified from the sequences, from which causal mutation was found in the 4-10 Mb region of chromosome III. Conclusion: Analyzing whole genome sequencing data using my codes can greatly reduce the workload of sequencing and subsequent causal mutation. Raw data and codes are available in https://github.com/ scilavisher/Fastmap.

Keywords: forward genetics; WGS; Caenorhabditis elegans; causal mutation identification

Preface

1. Forward Genetics

The main tools in genetics for studying the molecular mechanisms of phenotypes are forward genetics and reverse genetics. Reverse genetics studies gene function by knocking out, knocking in, mutating or modifying specific sequences at the molecular level and observing the effects of altered or disrupted DNA sequences on the organism. It relies on sequence information generated by sequencing or transcriptional profiling of genomic and expressed sequence tags, without a clear phenotype. In contrast, forward genetics aims to identify the sequence changes behind a specific variant phenotype[1] , which requires locus intervals to be found and candidate mutations to be targeted and searched within them, and subsequently validated against complementary function experiments. Specific mutant phenotypes can be obtained by mutagenesis or RNA interference. Chemical mutagenesis screens using ethyl methanesulfonate (EMS) can unbiasedly target phenotypic mutations and provide stable mutants for further study of gene function. In contrast, large-scale RNAi screens do not provide stable mutants and are not conducive to further analysis of gene function. In the case of phenotypic variants caused by T-DNA or transposon insertion, theoretically the gene can be quickly identified by locating the sequence tag and analysing adjacent sequences, but locating mutated genes by chemical or radioactive mutagenesis requires tedious mapping and cloning methods.

Locating the region containing the gene under study is generally done using genetic markers linked to the functional gene.[1] The mutant is crossed with a wild-type individual of the same strain with a polymorphic genetic marker. The mutant is crossed with a wild-type individual of the same species from another strain with a polymorphic genetic marker, and the F1 generation is selfed or crossed to produce a daughter with a pure mutagenic mutation in 1/4 of the second generation (F2). Unless the mutation is dominant, or the RNA or protein of the heterozygous parent is sufficient to revert the mutation in the progeny, a phenotypic F2 mutant can be screened. According to Mendel's laws of inheritance and the principle of random chromosomal

recombination during meiosis, the polymorphic sites of the selected F2 mutants should have a frequency of 1/2 of the mutated bases unless they are linked to the mutagenic mutation; the linked polymorphic sites are physically close to the functional mutation and are relatively resistant to recombination. Therefore, the region where the frequency of base polymorphisms in the mutant strain is significantly off by 1/2 is the genomic region where the functional mutation is located.

The genetic markers used for forward genetic identification of genes are restriction fragment length polymorphism (RFLP), microsatellite sequence (SSR) and single nucleotide mutation (SNP). traditional methods for SNP detection are PCR-single strand conformation polymorphism (PCR-SSCP) analysis and denaturing high performance liquid phase (dHPLC), all of which require gel electrophoresis. Although forward genetics methods are very effective, they are often time consuming, require the construction of a large number of lines for mapping, and because the mutations responsible for the phenotype are usually in the vicinity of certain genetic markers, they can only be localised to larger areas before being verified experimentally on a site-by-site basis. Sequencing can identify genetic markers and quantify allele frequencies in large numbers of DNA samples, revealing nucleotide sequences, while second-generation sequencing enables simultaneous localisation and identification of phenotypic mutations, reducing what would otherwise take a year of work to just a few weeks.[2] .

2. Sequencing technology development

Mutations are changes in the sequence of nucleotides in the genome, so naturally DNA sequencing, which is used to determine the sequence of nucleotides, is important for mutation detection. However, at first the biological macromolecules measured were proteins and RNAs, and all of them were interrupted and then resolved. It was not until 1968 that the 12 bases of the λ sticky end of the phage were measured by primer extension, and in 1973 Gilbert and Maxam successfully sequenced the 24 bases of the lactose inhibitor binding site by copying the DNA sequence to RNA and then sequencing it over a two-year period.

In 1976, the advent of strand termination and chemical cut-off methods that used distances

from radiolabelled bases to determine nucleotide order enabled the resolution of several hundred bases in half a day. Subsequently, using random cloning followed by sequencing, and then based on overlapping assembly by the birdshot method and single-stranded M13 viral cloning vectors, people began to be able to assemble genomes from scratch. 1987 saw the advent of the automated fluorescence-based Sanger sequencer, a device capable of sequencing 1000 bases a day. With the exponential growth of sequencing data and the development of search tools such as BLAST, data centres such as GenBank emerged and the value of sequencing data increased, further fuelling the passion for data sharing.

In 2005, 454 released the first commercially available second-generation sequencer. The advantages of second-generation sequencing technology over first-generation sequencing include: replication of sequencing templates by in vitro amplification rather than bacterial cloning; millions of sequencing reactions performed in parallel rather than one reaction per tube; and sequencing by cycling biochemical reactions and imaging rather than measuring fragment lengths. In vitro amplification methods include bridge amplification, which uses fixed primers to distribute amplification products in clusters around the template; PCR in emulsion drops, where the amplification product of each template is immobilised on small beads; or rolling loop amplification, which produces 'nanospheres' for sequencing. After a period of rapid growth, the development of second-generation sequencing has slowed down since 2012, with illumina having the sole market advantage[3] . Today, although still smaller than Sanger sequencing, second-generation sequencing is now 99.9% accurate and can sequence within two days 10^{12} orders of magnitude of bases[4] .

The ideal sequencing technology should be in situ, accurate and without read length limitations. Since the 1980s, efforts have been made to develop methods that are superior to second-generation sequencing, with third-generation sequencing technologies such as PacBio's SMRT and Oxford Nanopore Technologies' (ONT) nanopore sequencing technology enabling single-molecule sequencing, with PacBio enabling real-time observation of the polymerase synthesis process. The sequencing process eliminates the need for PCR amplification and

achieves average read lengths of 10Kb-15Kb, which are no longer equal. The ONT takes advantage of the primary structure of the ion flow reaction strand created when single-stranded DNA passes through a narrow channel. read lengths of the ONT can be greater than those of the PacBio and the device is as small as a USB, making it easily portable. Although ONTs suffer from problems such as errors that are not random, they are developing rapidly.

Traditional DNA microarray or NanoString methods are still available for mutation detection, but the most commonly used assays today are whole genome sequencing (WGS) and whole exome sequencing (WES) in second generation sequencing. Although WGS is currently more expensive than WES, it decreases in price faster than WES and is even more sensitive in exon mutation detection[5] and is therefore the most commonly used.

With WGS, researchers can quickly access information about mutations on the whole genome and discover high-risk mutations for disease based on this information and functional genome annotation, thus ushering in the era of disease genomics. For example, through WGS, researchers have screened for 18 new autism candidate genes[6] The discovery of a 0.6% missense mutation in ADCY7 doubled the risk of developing ulcerative colitis[7] The study found that a high risk of hip osteoarthritis was associated with aberrant genotypes of COMP and CHADL[8] A genealogy of Y-chromosome linked disorders was even obtained.[9] . Given the large number of spacer repeats in the genome, WGS currently uses mostly double-end sequencing, and when used for variant detection, WGS generally improves both base quality values and coverage by increasing the depth of sequencing by one degree.[10] This can reduce sequencing errors. However, there are still a number of non-sequencing factors that can cause false positive mutations, e.g. incorrect alignment of insertional deletions[11] However, there are still a number of non-sequencing factors that cause false positive mutations, such as: mis-assignment of insertions, mis-assembly of reference genes or SNPs introduced by different or low-quality alignment of study material to the reference assembly[12] However, there is still a considerable number of non-sequencing factors that can cause false positive mutations, e.g. mis-assignment of insertions, mis-assembled reference genes, or SNPs introduced by different or low-quality alignment of study

material to the reference assembly.[13] The number of false positives detected is even higher. The number of false positives detected may even be greater than the number of true mutations, so reliable filtering methods are needed.

3. Model organism Cryptobacterium hidradenum

The Cryptobacterium histolytica is a model organism often used in forward genetics. It is small, about 1-1.5 mm, and has both male and hermaphroditic sexes, with the males being small (0.02%) but increasing to 50% by mating with hermaphroditic nematodes. The wild type N2 hidrophilous nematode has an average lifespan of 18-20 days, with even shorter life cycles and lifespans at higher temperatures, facilitating hybrid transmission. Its anatomy is well defined, with 959 somatic cells and 302 neurons in known hermaphroditic adults, and its transparent body makes it easy to track cell fate or fluorescent marker protein expression. 1998 saw the first multicellular organism to have its complete genome measured, with up to 83% of its protein sequences homologous in humans and 38% of its 20,250 protein-coding genes corresponding to homologues in humans. The genome of the Worm was the first complete genome to be measured. Thus, since 1965, genetic screening by EMS mutagenesis or RNAi treatment of Cryptobacterium hidradenum has continued to contribute to the study of life processes, including human biology.[14] The genetic screening by EMS mutagenesis or RNAi treatment of Cryptosporidium hidradi has therefore continued to contribute to the study of life processes, including human biology, since 1965.

Specifically, the Nematode Genetics Centre has registered 190,000 species of Cryptobacterium hidradenum, including many wild types and Cryptobacterium hidradenum carrying more than 9,000 alleles. Cryptobacterium histolytica expresses approximately 28,146 proteins, has a genomic GC content of approximately 35.6%, and has five autosomes (I-V) and one sex chromosome (X). Hidrophilic nematodes are diploid, while males have only one X chromosome (X/φ). In meiosis, homologous chromosomes form association complexes, except for the male X chromosome, at which point crossover can occur and the recombinants obtained

by segregation are generally considered to be single-crossover products, and recombination rates do not vary significantly between 16-20 degrees C ambient temperature and are used as the basis for genetic screening.

WGS has been used for the identification of functional mutations in Cryptobacterium histolytica since a decade ago, but initially WGS was expensive and methods such as candidate region targeted sequencing (GIPS) were tried. As the price of WGS has decreased, in recent years WGS has been used almost exclusively for mutation detection in Cryptobacterium hidradenum, which can detect thousands of variants in a single nematode.[15] WGS can detect thousands of variants in a single nematode, thus requiring the development of "sequencing mapping" methods to reduce the number of candidate variants.

There are a number of ways to select localized regions using sequencing mapping, and for recessive mutations, HA variant mapping[16] By backcrossing the mutant strain with the highly polymorphic CB4856 HA nematode, linkage analysis was done using known HA SNPs and WGS for approximately10^5 EMS-density mapping was used to sequence and predict linkage between mutations in near-isogenic lines by backcrossing mutant nematodes. The method defines localisation intervals by chromosomal recombination boundaries, and repeating backcrosses two or three times increases accuracy. Backcrossing or telecrossing between mutant nematodes and unmutated parents in variant discovery mapping (VDM)[17] The VDM can map all SNPs in the mutant nematode background, improving mapping accuracy. Removal of mutations in nematode siblings of mutant nematodes to screen for phenotypically relevant genes[18] The results are also better when crossed with the same EMS-induced nematodes, but without the phenotype, and analysed by removing error-prone loci from the available data.[19] The results were also good. For dominant, semi-dominant and double mutations, Andy Golden et al. also developed a WGS-based localisation method.[20] For these methods, regression analysis (regression analysis) was used. For these methods, regression analysis (e.g. locally weighted regression scatter smoothing (LOESS)) or Bayesian network models can be used to fit regression lines to thousands of data points in the atlas to improve mapping accuracy.[21] The area can then be localised according to

the variability. After locating the region, the most likely type of variation can be prioritised based on the mutagenic agent. If EMS is used as the mutagen, the G-A and C-T transitions should be given priority. For variants that have no obvious effect on the open reading frame, the genomic conservation of different species in the vicinity of the putative variant can be checked on the UCSC Genome Browser. Finally, candidate genes are verified by Sanger sequencing for downstream gene function identification.

4. Significance of this research

Whole genome sequencing (WGS) is the most cost-effective and rapid method for mapping phenotype-causing mutations in model organisms such as Cryptobacterium hidradenum. Sequencing takes only a few days and costs under two thousand dollars, avoiding gene cloning projects that take years and reducing reagent costs. In addition, this method can be used for large-scale genetic screening and then rapid sequencing of the many mutants obtained from it, leading to a deeper understanding of life processes and genes. In addition, mapping lines for phenotypically cumbersome mutants (e.g. behavioural mutants) or mutant lines of specific genetic backgrounds are difficult to obtain and can be identified by whole genome sequencing. However, their data analysis is complex and requires specialist bioinformatics knowledge. As the cost of WGS continues to fall and the technology becomes commonplace, all laboratories using genetic analysis will need to establish their own screening platforms.

This project is based on laboratory crosses and sequencing of nematodes, and analysis of sequencing data to reduce the number of phenotypically related candidate loci. The established method can also be used in forward genetics studies of model organisms such as Drosophila, Arabidopsis and others.

Chapter 1 Data and methods

In our laboratory, nematodes with abnormal localization of nrd-3 protein isolated from EMS mutagenesis have been crossed with wild-type nematodes to obtain F1 generation nematodes without phenotypes. Subsequently, F1 generation nematodes were allowed to self-fertilize and 20 to 50 nematodes with phenotypes only were selected from F2 generation nematodes and 20-50 nematodes were selected from parental nematodes as controls. The two groups of nematodes were shattered to extract genomic DNA and then subjected to WGS sequencing to obtain mutant (mutant) and wild-type (wt) double-end sequencing files, respectively. This chapter sets up the process for finding mutations and identifying phenotypically relevant gene regions from the sequencing data.

1.1 Data and tools

This project requires sequencing data for the parental wild type and variant nematodes singled out in the second generation, reference genomic data for the wild type of Cryptobacterium hidradenum and tools for variant detection.

1.1.1 Data

1) Sequencing data: wt and mutant double-end sequencing data, FASTQ format.
2) Reference genome: Cryptobacterium hidradenum Bristol N2 species WBcel235 reference genome file with annotation file, downloaded from ensemble.

Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
I	NC_003279.8	BX284601.5	15.07	35.7	4,133	6	76	1,221	4,187	160
II	NC_003280.10	NC_003280.10	15.28	36.2	4,704	-	80	1,565	5,195	262
III	BX284602.5	BX284602.5	13.78	35.7	3,690	-	97	1,060	3,826	130
IV	NC_003281.10	NC_003281.10	17.49	34.6	5,155	-	94	16,208	19,688	375
V	BX284603.4	BX284603.4	20.92	35.4	6,659	15	169	2,214	7,832	860
X	NC_003282.8	NC_003282.8	17.72	35.2	3,793	-	305	3,004	5,991	114
MT	BX284604.4	BX284604.4	0.01	23.8	12	2	22	-	12	-

Table 1 N2 strain WBcel235 Reference genome information

1.1.2 Tools

FastQC (v0.11.7) https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Trimmomatic (v0.33) http://www.usadellab.org/cms/index.php?page=trimmomatic

BWA (v0.6) http://bio-bwa.sourceforge.net/

Samtools (v1.4) http://www.htslib.org/doc/samtools-1.4.html

Picard https://broadinstitute.github.io/picard/

R (v3.4.3) http://www.R-project.org/

Python (v2.7) https://www.python.org/

ANNOVAR http://annovar.openbioinformatics.org/en/latest/

1.2 Statistical quantities

Assuming a Gamma sampling process for library construction and a Poisson sampling process for sequencing, ideally, the frequency of mutated bases in the F2 generation that are not linked to the hypothetical phenotype should follow a negative binomial distribution with a mean of 0.5, and the frequency of mutated bases in the linked region should converge to 1. The size of the genomic pool measured in this experiment is unknown, and the number of recombination events cannot be inferred using, for example, a Hidden Markov Model, so the choice is to define a statistic In this experiment, the size of the genomic pool is unknown and the number of recombination events cannot be inferred using a Hidden Markov Model.

Let ref be the number of reads where the measured site is the same as the reference base and alt be the number of reads where the base is different from the reference base. Considering that the parental nematodes may not be completely pure and not all mutant sites in the F1 generation are on one strand, ideally, the ref measured for non-linked bases in the F2 mutant nematode should be equal to the alt; if the mutant site is in the mutually exclusive phase with the presumed mutagenic mutant site, the alt is greater than the ref, and the greater the linkage, the greater the alt compared to the ref; if it is in the mutually exclusive phase, the opposite is true. Accordingly, the statistic R is defined as $R = \begin{cases} \max(ref, alt) - 1, & ref \times alt = 0 \\ \frac{alt}{ref}, & alt > ref \\ \frac{ref}{alt}, & ref > alt \end{cases}$, the larger the value of R, the higher the degree of linkage, and retains the mutant base frequency ($V = \frac{alt}{alt+ref}$) and the depth-insensitive normalized mutant base frequency ($NV = 100 \times V$) and the mutant base frequency difference ($D = \frac{|alt-ref|}{alt+ref}$) for method validation.

1.3 Process

When mutagenic nematodes are crossed with non-parental nematodes and the F1 generation is heterozygous, and the F1 is selected to produce an F2 generation of phenotypic nematodes by self-fertilisation, the pure mutation will segregate to varying degrees due to linkage to the locus causing the phenotype, producing a change in allele frequency.

Pure mutations detected in mutagenic nematodes may be directly caused by illumina sequencing errors or belong to the original genetic background of the nematode, only partially caused by mutagenesis, and functional mutations not identified from mutagenesis-generated mutations, the sequencing data should be quality-controlled, the detected variants filtered and characterised for linkage by defining appropriate statistics, the relevant gene regions selected and the mutations in them The genes are annotated to give a list of genes that are most likely to be responsible for the phenotype. Ultimately, the process is divided into five main parts: data quality control, data pre-processing, variant detection, localisation of phenotype-associated gene regions and gene annotation and further screening.

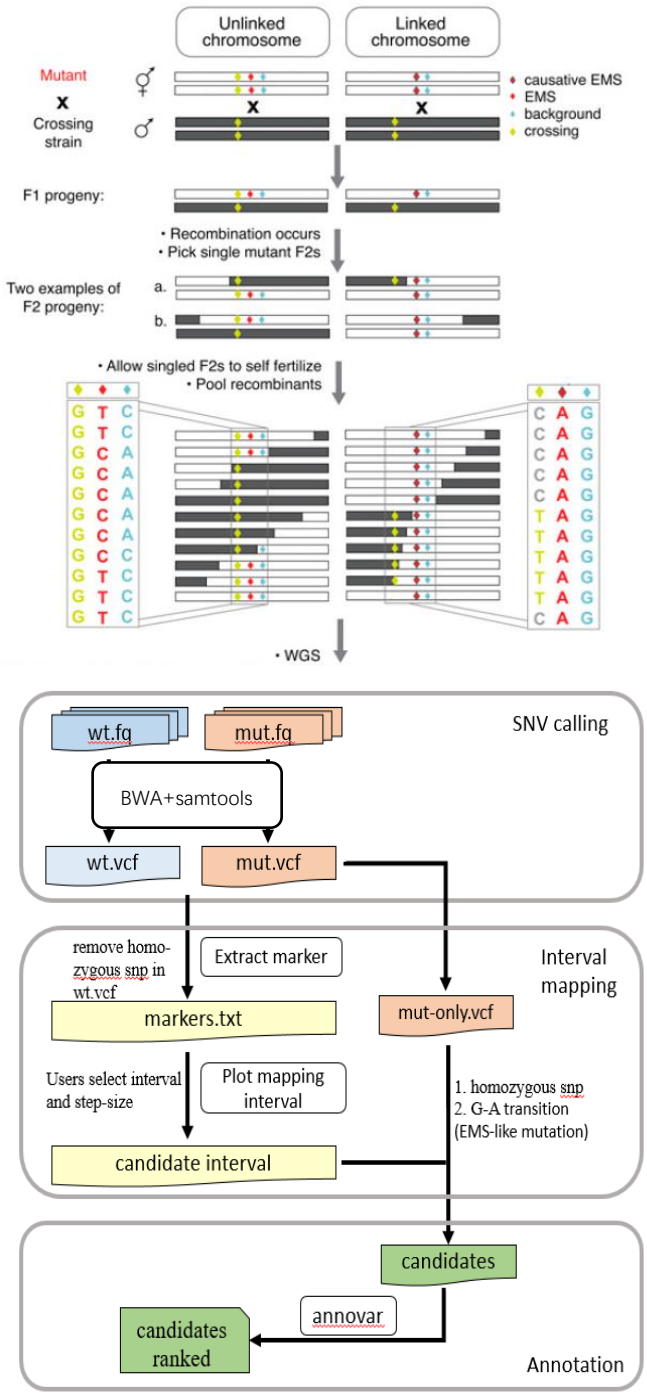


Figure 1 The flow of the subject

1.3.1 Raw data quality control

Scripts were first written to check the quality system. FastQC was then used to QC the data, looking at base quality value distribution, GC content distribution, N content, splice sequences, and repeat rates. After specifying the junction sequence, double-end sequencing, and quality system parameters, the sequencing junction and low quality sequences were removed using

trimmomatic and finally the quality was checked again using FastQC.

1.3.2 Data pre-processing

The sequencing data is first compared using the BWA sequence alignment tool, which compares reads from the original genome to the nematode reference genome to find its position and sequence it. To do this, the reference genome is Burrows Wheeler transformed and FM indexed to allow the sequence alignment to be searched and located with linear complexity, while faidx indexes are added using samtools to facilitate variation detection. The bwa mem alignment method is then selected based on the sequence read length information from the data quality control output. After completion of the alignment, data with an alignment quality greater than 30 are filtered using the samtools -q parameter and output as a bam file with increased position, alignment quality and alignment structure information over the original data.

Subsequent sequencing was performed using the samtools sort command and the picard MarkDuplicates command to mark duplicate sequences. Second generation sequencing often includes PCR amplification to increase the density of interrupted DNA fragments, so uneven initial DNA fragment density, base errors during DNA interruption and PCR and the different propensity of PCR reactions to amplify the template are all non-true signals that are amplified during amplification. Subsequent variant detection is based on Bayesian principles and does not distinguish between repetitive sequences. The FLAG information in the BAM file therefore needs to be flagged and the duplicate sequences ignored during variant detection to eliminate the detection errors they introduce. For random access to positions in the file, the tagged file is indexed using the samtools index tool.

The sorted.markdup.bam file, sorted.markdup.ban.bai file and markdup_metrics.txt file should exist in the directory after data pre-processing.

1.3.3 Variation detection

Use the samtools mpileup command for variant detection and the vcftools --remove-indel command to remove detected insertion-deletion variants. In contrast to GATK, samtools does not discard low quality sequencing data by default during variant detection to avoid losing possible

mutant loci; samtools' SNP genotype likelihood model is based on non-independence between sequencing errors; samtools can correct quality values by adjusting the MQ parameter through the output, and for pseudo-single nucleotide mutations due to insertion deletions, a Hidden Markov If a base is paired to a different reference base in a sub-optimal pairing, the BAQ value is small and the contribution to SNP detection becomes smaller. This approach is simpler and faster than GATK, which uses a large set of known variants based on a machine learning algorithm to correct base quality values, and is more suitable for human disease diagnosis.

Write scripts to process the SNPs_only.vcf files for wild-type and mutant nematodes, with the mutation files filtered out of wild-type pure mutations to obtain all candidate loci.

1.3.4 Candidate area selection and annotation

The detected variants were first subjected to quality control. Variant loci with abnormal sequencing depth and detection quality compared to sequencing depth were removed using the car package in R to obtain SNP markers. The R-value of the SNP markers is calculated using scatter plots and loess smoothing curves made with the ggplot2 package in R in order to select candidate regions. If the experimenter has a candidate region length expectation, the region with the highest mean value of the statistics obtained from the sliding window can be used as a candidate region by enumeration.

To annotate candidate regions based on genes, annotation files were first constructed. The gtfToGenePred tool was used to convert the genomic information file in gtf format to a GenePred file, then retrieve_seq_from_fasta.pl was used to convert to a FASTA file. Finally the annotation file was obtained using table_annovar.pl.

1.4 Summary of this chapter

This chapter designs the statistics used to select regions where mutagenic mutations are likely to be present, prepares the data set needed for the project, and builds the entire process and working directory according to the aims of the project. Quality control is written in QC.py, data pre-processing and variation detection is written in call variation.py, statistic calculation, region screening and mapping is written in main.py, and work on duplicate or different samples is done

by simply entering the sample names and the order of the work parameters on the command line and executing the script commands.

Chapter 2 Results and Discussion

Scripts are written to implement data quality control, data pre-processing, variant detection, localisation of phenotype-associated gene regions and gene annotation based on the defined statistics and overall flow. This chapter presents the main results and provides a discussion.

2.1 Mutation detection

2.1.1 Raw data quality control

Checking the quality system used to validate illumina hiseq2000 for base sequencing was Phred33, and after removing sequencing junctions and low quality sequences, the results of data quality control showed that all sequencing results had base quality values above 30; the average quality of each read segment was around 39; the GC content was stable at 35%, consistent with the nematode genome, and the purine and pyrimidine content were There were no unknown bases; the sequencing length was 150 bp; the sequencing repeat level was less than 2 and the splice sequences were removed, suggesting high quality sequencing data (Table 2-1).

Sample	Total number of sequences	Low quality/%	Read long/bp	%GC	Average sequencing depth	Average coverage
mut-1	22088348	0	150	35	i:21.6; ii:21.5; iii:21.5; iv:21.5; v:21.4; x:21.3	I:98.9%; II: 98.8%; III:99.1%; IV: 98.1%; V:97.4%; X:99.3%
mut-2	22088348	0	150	35		
wt-1	15264164	0	150	35	i:54.8; ii:53.1; iii:53.0; iv:53.4; v:53.7; x:53.9	I: 96.4%; II:96.6%; III:96.8%; IV:95.3%; V:95.1%; X: 97.6%
wt-2	15264164	0	150	35		

Table 2 -1 Summary of sequencing quality

2.1.2 Data pre-processing

Both wild-type and mutant nematode data were back-posted to the reference genome, with

99.69% of paired sequences successfully paired and 0.04% single-ended back-posted, suggesting successful back-posting. The average sequencing depth of the wild type was 21.5x, with an average coverage of 96.28%; the average sequencing depth of the mutant nematode was 53.6x, with an average coverage of 98.59%, and the depth and coverage could be used for mutation detection. Sequencing and duplicate sequence tagging were performed after the comparison, and the estimated library size of the wild-type data was found to be 430805547, of which 3465 were unpaired and 1746680 of the paired read segments were tagged as duplicates. Repeated sequences accounted for 12.39% of the total, including 1566114 optical repeats. The estimated library size for mutant nematodes was 629925599, with 2401718 paired repeats, or 11.9%, in addition to 7342 read segments with unsuccessfully paired sequences, marked as repeats (Table 2-2).

Sample	Unpaired reads		Paired reads		Unmatched reads	Repetition rate	Estimated library size
wt	Repeat 0	0	Repeat 1190706	6936918	0	14.65%	230797868
mut	Repeat 7342	14069	Repeat 2401718	17860064	0	11.87%	629925599

Table 2-2 Summary of pre-treatment quality

In this experimental system, it can be seen from Figure 2-1 that sequencing depths up to 15x make the data reflect a sequencing depth smaller than the experimental depth due to repeated sequences, and increasing the depth is not significant when sequencing depths reach 60x or more.

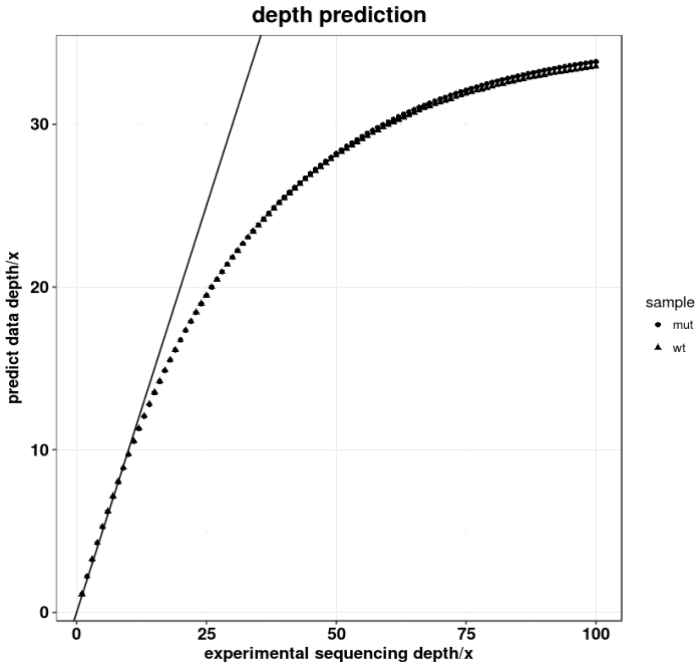


Figure 2 -1 Actual vs. expected sequencing depth

2.1.3 Mutation detection

All mutations were detected separately for mutant and wild-type nematodes using samtools. The mutant nematode had 12,795 single nucleotide variants remaining after filtering out insertional deletion variants at 15,158 loci, and the wild type retained 12,336 loci from 14,708. Of the single nucleotide mutations obtained, 9711 loci were common to both mutant and wild-type, and 3084 loci were detected only in mutant nematodes. Of the 2625 mutations detected only in wild-type nematodes, 1695 were pure loci unlikely to cause the phenotype, and a total of 10,641 candidate mutation loci were removed to obtain mutant nematodes. Sequencing data for these loci affected R-values and were first subjected to quality control.

From the preliminary information on the quantile of sequencing depth it is clear that about 90% of the sequenced sequences are within 200x in depth, but some are over 7000x. 90% of the base mutation detection quality ratio depth values are below 7, but some are as high as 27 (Appendix Figure 1), they are both anomalies in the data and inconsistent with the number of nematodes (20-50) sequenced in the previous experiments of this project, therefore, the data need to be explored further to screen out the outliers. The upper and lower quartiles, inner limit data and outliers were obtained from the data, and outliers with sequencing depths greater than 199x

and base mass values greater than 9.75 than the sequencing depth were screened out, from which

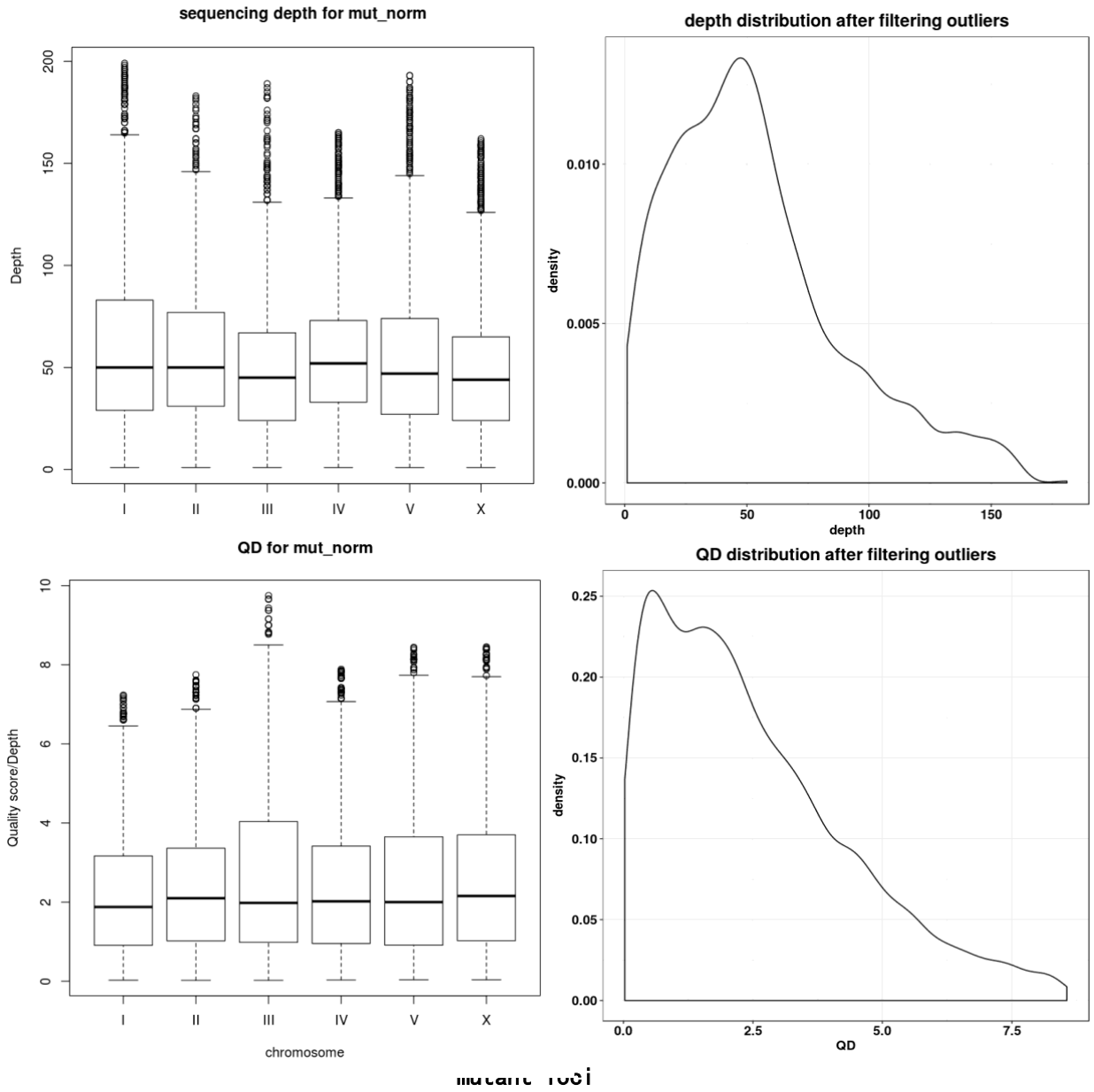


Figure 2-2 Sequencing depth of SNP markers and quality of mutation detection

a total of 86.3% of the loci sequenced more than 10 times to (Figure 2-2) were retained as SNP markers for subsequent region selection.

2.2 Screening for candidate genes

2.2.1 Selection of candidate regions

R values at the selected SNP markers were calculated and scatterplotted, and a range of data subsets (span=0.2) was empirically selected for a local polynomial regression fit (LOESS) line. The degree of linkage reflected in Figures 2-3 shows that the phenotypically related genes are

located on chromosome three and allows the selection of candidate regions ranging from 4-10 Mb (recombination index 0.06). The 65 mutations in the selected range were sorted by standard quality values, of which 12 genes with A-G conversion non-synonymous single nucleotide mutations located in exons (Table 2-3), five of which encode proteins that have been reported. We then performed siRNA interference on the 12 genes to verify that there were indeed genes that revert the phenotype.

Gene	site	ref	alt	R	V
F43C1.5, exon3	4229590	G	A	53	1
R07E5.10a, exon2	4417307	G	A	42	1
T04A8.18, exon2	4710176	G	A	50	0.98
Y32H12A.3, exon4	5364675	G	A	52	1
ZK328.5b, exon8*	6015164	G	A	42	0.98
F23F12.4, exon2*	6499355	G	A	19.5	0.95
F20H11.2, exon7	6597641	G	A	23.5	0.96
B0280.12a, exon8; B0280.12b, exon10*	7142753	G	A	52	0.98
K12H4.3, exon2*	8046583	G	A	36	1
C50C3.8, exon1*	8163304	G	A	45	1
ZK643.5, exon2	8956006	G	A	42	1
ZK1098.2, exon5	9530115	G	A	39	1

Table 2-3 Candidate genes *Protein-coding functions have been reported

2.2.2 Method validation and discussion

The distribution of sequencing depth for all variant loci was tested to obey a Gamma distribution (p-value < 2.2e-16); mutant base frequencies on all chromosomes except chromosome three can be considered identically distributed (Kruskal-Wallis rank sum test, p=0.000293), with chromosome three significantly higher (Kruskal- Wallis rank sum test, p=0.3675), it is reasonable to assume that the phenotype-related genes are located on chromosome three.

Validation of the selected regions was performed using enumeration. The total length of chromosome III was 13.78 MB, and the candidate region of 6 MB was expected to be found from R-value mapping. A window of 6 Mb was taken for R values with loci, and a sliding window of 0.1 was used to find the highest mean value of R within the window of precisely 4 MB-10 MB. 100 resamplings of all data with random equal data volume were obtained, and nearly 50% of the results were in the range of 4-10 MB, and over 80% were in the range of 3.5-10 MB. The method

is considered to be sufficiently stable (Figure 3 attached). In addition, NV values (100*variation frequency) greater than 48 suggest that the number of mutant bases detected was not sampled from the overall base frequency of 0.5 (p=0.05), and the distribution of NV values for each chromosome in the selected regions of this data did not show any abnormalities (Appendix Figure 2).

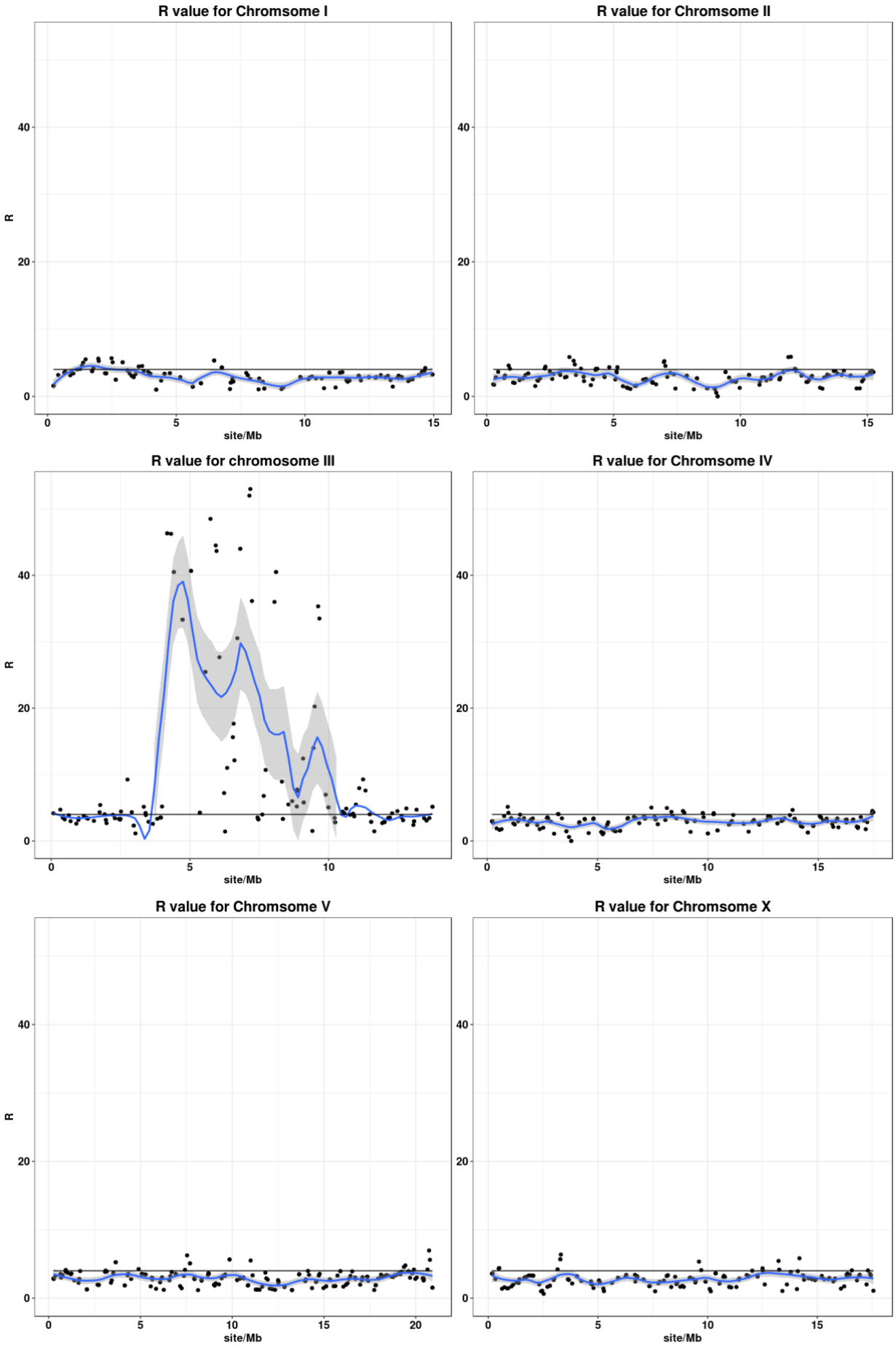

R value for Chromsome I
R value for Chromsome II
R value for chromosome III
R value for Chromsome IV
R value for Chromsome V
R value for Chromsome X
R
site/Mb

Figure 2-3 R-value and LOESS regression (span=0.2) Blue line: loess regression line; black line: R=4

Compared to the other statistics, the R value is somewhat affected by the depth of sequencing when alt or ref is 0, but it takes into account both cis and trans scenarios, taking values of different orders when highly linked and unlinked, facilitating a more intuitive targeting of candidate regions. However, R-value size is relatively unresponsive to base frequency size between loci with large base frequencies, and ideally the other statistics still respond to linkage with good linearity, but do not have the effect of increasing the distance between non-identical data and relatively decreasing the distance between similar data. Direct mapping using mutation base frequencies requires a greater degree of loess regression, carries more information loss and is not high resolution, and does not correctly identify localised regions under non-ideal experimental conditions, i.e. where the F2 generation carries mutations introduced by the parental wild type. Combining the results of the different statistics in this topic (Figure 4 attached) and other literature results[17] , it is recommended to apply R-values to select candidate regions.

2.3 Discussion and outlook

The goal of this project is to find chromatin regions with high linkage to genes associated with the putative phenotype. Ideally, the parental nematodes are purely germline, F1 nematodes are heterozygous, and the mutation frequencies of F2 non-linked loci measured by high-throughput sequencing should obey a non-relaxed sampling of 0.5 ($\frac{(alt-ref)^2}{alt+ref} \sim \chi 2(df = 1) > 9, p < 0.01$), from which the measured data results deviate systematically (negative binomial fit of mutation base frequencies close to 0.35), probably because the parental mutant nematodes are not completely pure, but do not affect the selection of relatively high-interlocking regions.

One limitation of using second generation sequencing to locate phenotype-associated mutations is whether the sequencing covers the target mutation site. Calculations show that sequencing depths up to 20x can achieve a sensitivity of 89% for detecting point mutations, and up to 95% for mutations in exonic regions. The present project sequenced mutant nematodes to a depth of 50x, and multiple mutations may exist in a coding region, which should cover phenotype-associated mutations better. Another possible problem is the detection of false-

positive or false-negative mutations in regions that are difficult to sequence or have low coverage. The method in this study is relatively robust in that mutation detection is not screened by mass values, and only purely conforming mutations detected in both wild-type and mutant types, respectively, are deducted, retaining all the remaining mutations in the mutant nematode and reducing the risk of missing mutations. When targeting phenotype-associated mutation regions in statistical mapping, sequencing depth and base mass ratio sequencing depth were chosen to be within the normal range of the data to ensure accurate region selection. However, there may be instances where there is only an imbalance in linkage with functional mutations rather than a small physical distance, but no concentration in one region. In addition, the most difficult cause of phenotypes to detect by second generation sequencing is gene duplication, and it is difficult to distinguish whether changes in the number of short sequences detected have an effect of gene copy number. Studies have shown that multi-copy transgenes have a much higher number of short sequences measured compared to neighbouring regions, but it is still very difficult to detect individual duplication events in genes, and when double-end sequencing of a gene is found to compare to a completely different genomic region, it suggests that gene duplication may have occurred there. The mutation to be selected for this project was generated by EMS mutagenesis, and it is unlikely that the phenotype was caused by gene duplication, which was also not detected in the data. Finally, functional genes may be located in regions of the Cryptobacterium histolytica genome that have been mis-annotated. The genes annotated in this project were not found to be abnormal in different nematode genomes, and some of them did revert to the phenotype after RNAi interference. In conclusion, it can be concluded that this project has successfully established a reliable method to locate only 12 candidate genes from tens of thousands of mutations, which greatly reduces the experimental workload required for complementation experiments and mutagenic mutation localization.

It is the combination of traditional genetics and rapidly advancing sequencing technologies that has led to new methods for identifying the genes responsible for phenotypes. It is also through the study of model organisms such as the hidrophilic nematode that these new methods are being

proof of concept. Second-generation sequencing-based methods allow rapid localisation and cloning of mutations of interest, thus revolutionising forward genetic screening. In addition, these methods also provide us with a variety of background mutations and phenotypic mutants. In the future, the reliability of variant and mutation identification is expected to increase further with the further development of triple sequencing technologies and the further growth of read lengths.

2.4 Summary of this chapter

In this chapter, based on the established methodological flow, the mutant loci of wild-type and mutant nematodes that were sequenced were detected, and from the 10,641 loci in the mutant nematode variant loci, after subtracting those with pure mutations in wild-type nematodes, the loci located in the 4Mb-10Mb interval of chromosome III were selected and annotated to obtain 12 coding region genes. They were also validated by enumeration method and probability test, and various statistics were compared to obtain satisfactory results. The 12 genes obtained from the screening were subjected to RNAi reversion experiments, in which phenotype-related genes were found.

References

[1] Peters, Cnudde, Gerats. Forward genetics and map-based cloning approaches [J]. Trends in Plant Science, 2003, 8(10): 484-91.

[2] Sun, Schneeberger. SHOREmap v3.0: fast and accurate identification of causal mutations from forward genetic screens [J]. 1940-6029.

[3] Li, Hsieh, Young, et al. Illumina Synthetic Long Read Sequencing Allows Recovery of Missing Sequences even in the "Finished " C. elegans Genome [J]. Scientific Reports, 2015, 5.

[4] Park, Kim. Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing [J]. Int Neurourol J, 2016, 20(Suppl 2): S76-83.

[5] Belkadi, Bolze, Itan, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants [J]. Proc Natl Acad Sci U S A, 2015, 112(17): 5473-8.

[6] Rk, Merico, Bookman, et al. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder [J]. Nat Neurosci, 2017, 20(4): 602-11.

[7] Luo, De, Jostins, et al. Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7 [J]. Nat Genet, 2017, 49(2): 186-92.

[8] Styrkarsdottir, Helgason, Sigurdsson, et al. Whole-genome sequencing identifies rare genotypes in COMP and CHADL associated with high risk of hip osteoarthritis [J]. Nat Genet, 2017, 49(5): 801-5.

[9] Jobling, Tyler. Human Y-chromosome variation in the genome-sequencing era [J]. Nat Rev Genet, 2017, 18(8): 485-97.

[10] Sims, Sudbery, Ilott, et al. Sequencing depth and coverage: key considerations in genomic analyses [J]. Nat Rev Genet, 2014, 15(2): 121-32.

[11] Li, Fan, Tian, et al. The sequence and de novo assembly of the giant panda genome [J]. The sequence and de novo assembly of the giant panda genome [J]. 1476-4687 (Electronic)).

[12] Cheng, Brunner, Kremer, et al. Co-regulation of invected and engrailed by a complex array of regulatory sequences in Drosophila [J]. Developmental biology, 2014, 395(1): 131-43.

[13] Elhaik, Greenspan, Staats, et al. The GenoChip: a new tool for genetic anthropology [J].

1759-6653.

[14] Belkadi, Bolze, Itan, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants [J]. Proc Natl Acad Sci U S A, 2015, 112(17): 5473-8.

[15] Lehrbach, Ji, Sadreyev. Next-Generation Sequencing for Identification of EMS-Induced Mutations in Caenorhabditis elegans [J]. Curr Protoc Mol Biol, 2017.

[16] Jaramillo, Fuchsman, Fabritius, et al. Rapid and Efficient Identification of Caenorhabditis elegans Legacy Mutations Using Hawaiian SNP-Based Mapping and Whole-Genome Sequencing [J]. G3 (Bethesda), 2015, 5(5): 1007-19.

[17] Minevich, Park, Blankenberg, et al. CloudMap: a cloud-based pipeline for analysis of mutant genome sequences [J]. Genetics, 2012, 192(4): 1249-69.

[18] Joseph, Blouin, Fay. Use of a Sibling Subtraction Method for Identifying Causal Mutations in Caenorhabditis elegans by Whole-Genome Sequencing [J]. G3 (Bethesda), 2018, 8(2): 669-78.

[19] Addo, Buescher, Best, et al. Forward Genetics by Sequencing EMS Variation-Induced Inbred Lines [J]. G3 (Bethesda), 2017, 7(2): 413-25.

[20] Smith, Fabritius, Jaramillo, et al. Mapping Challenging Mutations by Whole-Genome Sequencing [J]. G3 (Bethesda), 2016, 6(5): 1297-304.

[21] Edwards, Gifford. High-resolution genetic mapping with pooled sequencing [J]. BMC Bioinformatics, 1471-2105.

Acknowledgements

As my four years at university draw to a close, I would like to thank Professor Wang Min and Professor Song Xiaoyuan for their dedicated guidance during my internship, as well as Professor Wang Min for his selfless help in my studies and life as my supervisor in the Pioneer Programme for the past two years. Over the past three years, it was Mr Wang's support that enabled me to have the platform to turn my ideas into reality, the opportunities to study inter-campus and abroad, and Mr Wang's teaching from tutorials to labs that benefited me greatly.

I would like to thank Ms. Luo Chen, Ms. Yu Luting, Ms. Zhang Zhiyuan, Ms. Du Pei, Ms. Li Youjie and Ms. Li Xiang for their care and help in my life and research. They taught me various experiments and ways of thinking about biomedical research, and their trust in me has enabled me to grow rapidly.

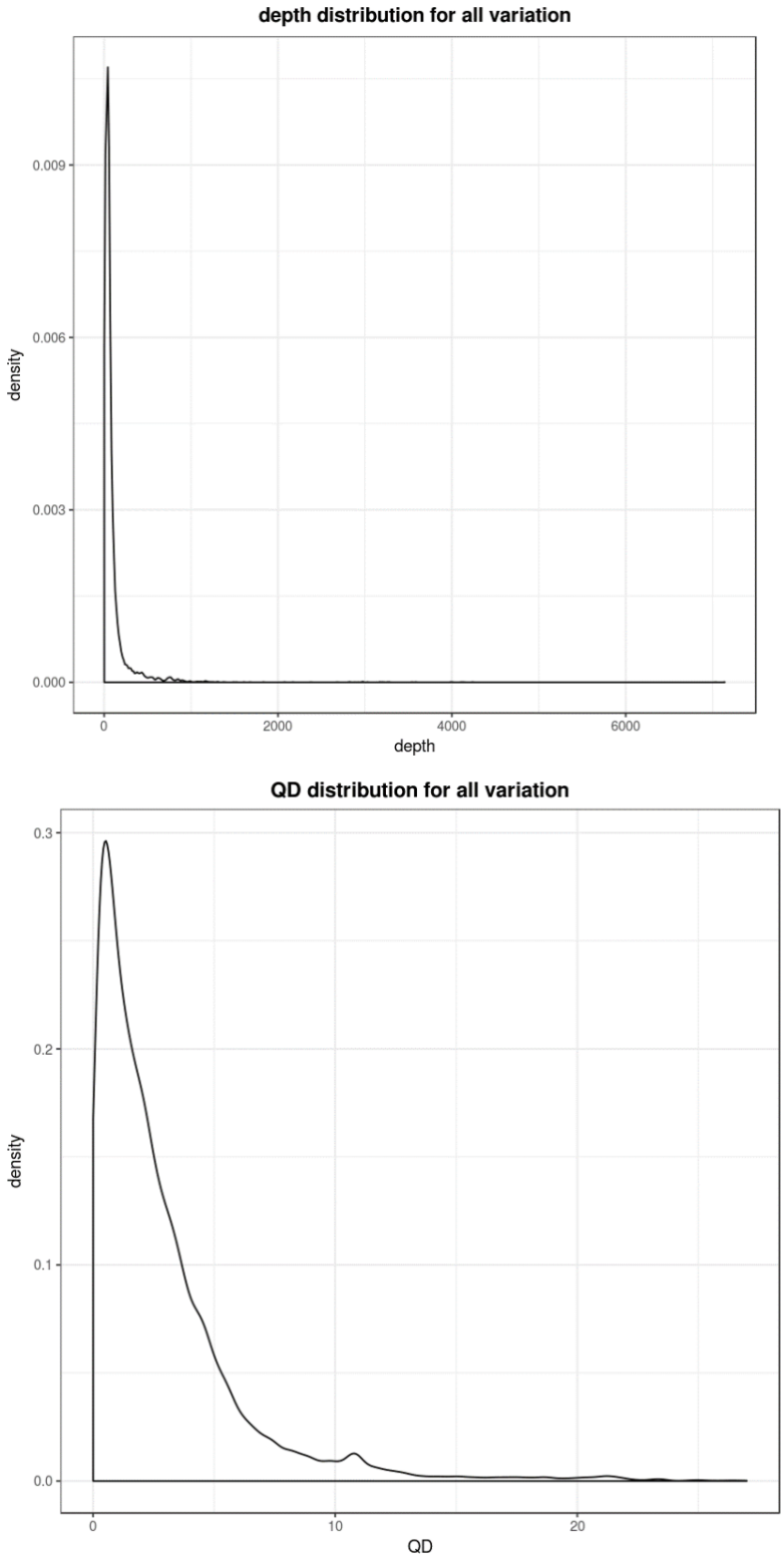
I am grateful to Song Xiaoyuan's supervisory team and all the members of the BSC team at the University of Science and Technology of China for their advice and assistance on the important points of my thesis, their seriousness in answering my questions and solving my problems, and their concern for my life during my final year.

I would like to thank Li Liang, Zhang Jingxin and other internship students for their great support and assistance. We studied together, submitted materials together, helped each other and grew together. I would like to express my loyal thanks to all of you.

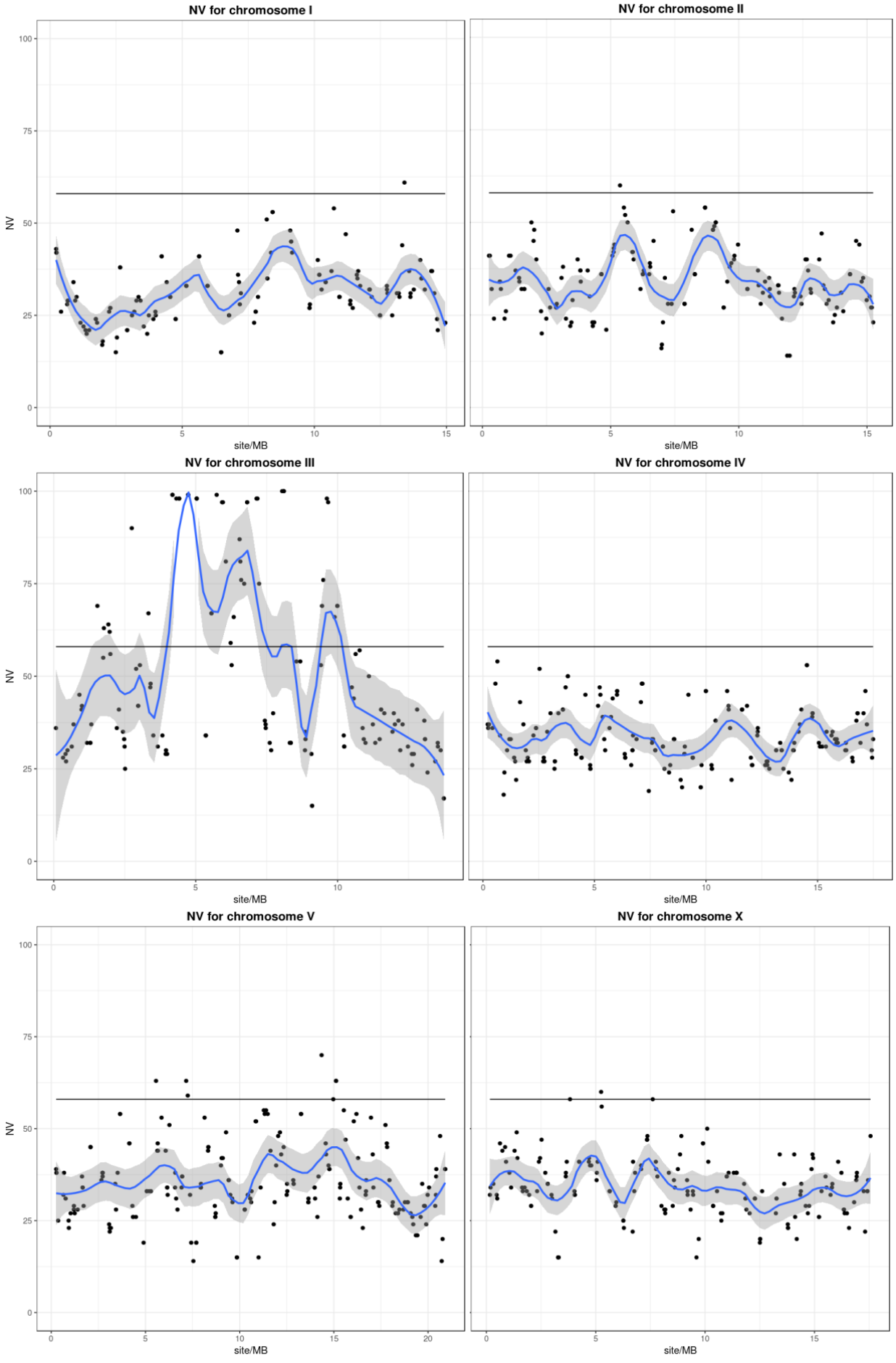
Finally, I would like to thank the teachers in the School of Life Sciences and Technology for their hard work. I did not have enough time to learn from you as an undergraduate, and I may ask for advice on professional issues in the future. Thank you to my family and friends for their support, and to those students I spent time with for inspiring each other to grow together!

Appendix

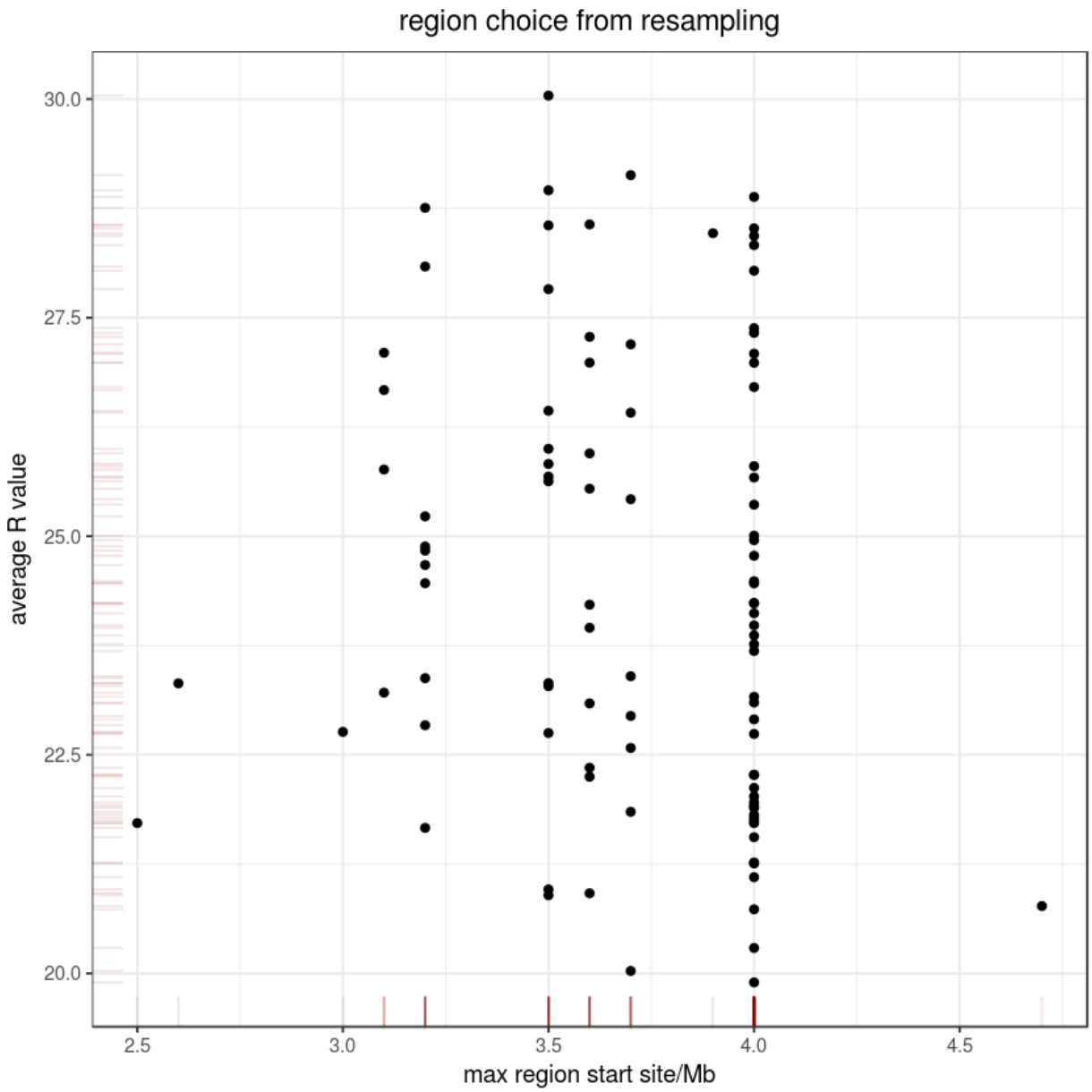
Annex 1 QC.rar



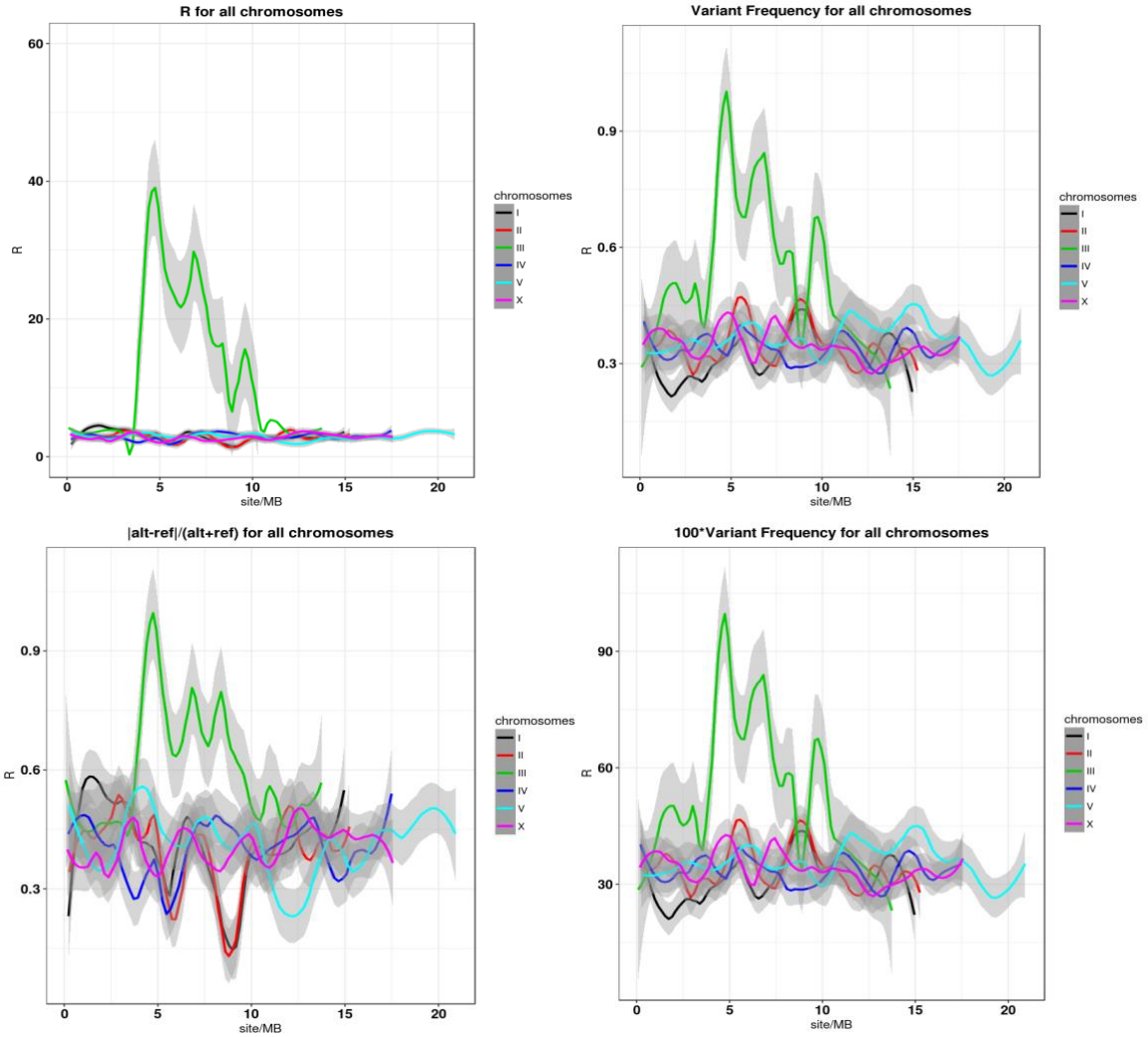
Attachment 1 Sequencing depth and base mass values of polymorphic loci in variant nematodes compared to sequencing depth



Attachment 2 NV=100*alt/(alt+ref)



Attachment 3 Resampling of 100 mutant locus regions at the start site on chromosome III



Attachment 4 Results for different statistics