中国药科大学

本　科　毕　业　论　文

论文题目　基于全基因组测序的线虫功能突变的筛选

英文题目　Causal Mutation Screening in C.elegans based on WGS

专　　业　药学（国家生命科学与技术人才培养基地）

院　　部　生命科学与技术学院

学　　号　14403718

姓　　名　蒋红

指导教师　王旻　教授

课　　题

完成场所　中国药科大学微生物与生化药学实验室

论文工作时间：　2018 年　2　月 至　2018 年　5　月

基于全基因组测序的线虫突变基因的筛选

目录

基于全基因组测序的线虫功能突变的筛选

14403718　蒋红

摘要：目的：筛选导致秀丽隐杆线虫 nrd-3 蛋白定位错误的突变基因。方法：对经 F2 代表型筛选后的野生型亲本和 F2 代突变体全基因组测序（WGS）数据进行质量控制和变异检测。将 F2 代突变体与亲代野生线虫的纯合突变不同的变异位点作为候选位点，从中筛选出高质量的单核苷酸多态性（SNP）标记。定义统计量 R，对 SNP 标记的 R 值作图并枚举选择 R 值最高的区间作为定位间隔，从中注释并鉴定候选位点处的基因。结果：通过对不足 100 只线虫的 WGS 数据的分析，将功能突变定位到线虫 III 号染色体 4-10Mb 区间，从中鉴定出 12 个候选基因，经后续 RNAi 验证，找到功能突变。结论：通过本课题集成的代码对简单模式生物的 WGS 数据分析可大大减少测序及鉴定与验证的工作量，将鉴定功能突变的周期缩短到几周时间。数据和代码：https://github.com/scilavisher/Fastmap。

关键词：正向遗传学；全基因组测序；秀丽隐杆线虫；突变鉴定

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 parental wild typed nematodes from the variation sites of F2 generation phenotypic worms. 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 interval. Then annotate and identify causal mutation in that interval. Results: By analyzing complete genome sequencing data of less than 100 Caenorhabditis elegans, causal mutation was located in the 4-10 Mb region of chromosome III. Twelve candidate genes were identified from the sequences, from which causal mutation was found by subsequent RNAi treat. Conclusion：Analyzing whole genome

sequencing data using my codes can greatly reduce the workload of sequencing and subsequent causal mutation identification and validation. Raw data and codes are available in https://github.com/scilavisher/Fastmap.

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

前言

1. 正向遗传学

遗传学中研究表型的分子机制的主要手段有正向遗传学和反向遗传学。反向遗传学通过在分子水平对特定序列进行敲除、敲入、突变或修饰，观察改变或破坏 DNA 序列对有机体的影响来研究基因功能。它依赖基因组和表达序列标签测序或转录谱产生的序列信息，而没有明确的表型。相反，正向遗传学旨在鉴定特定变异表型背后的序列变化[1]，需要找到定位区间并在其中靶向搜索候选突变，随后根据互补功能实验验证。特定突变表型可通过诱变或 RNA 干扰获得。利用乙基甲磺酸酯（EMS）的化学诱变筛选法可无偏地定位表型突变且能提供稳定的突变体用以进一步研究基因功能。与此相比，大规模 RNAi 筛选不能获得稳定的突变体，不利于进一步分析基因功能。若是 T-DNA 或转座子插入导致的表型变异，理论上通过定位序列标签并分析邻近序列可很快鉴定基因，但化学或放射诱变产生的突变基因定位需要繁琐的作图克隆方法。

定位包含研究基因的区域一般利用与功能基因连锁的遗传标记[1]。使突变体与同种的具有多态性遗传标记的另一品系的野生型个体进行杂交。F1 代自交或杂交产生子二代（F2）中 1/4 带有纯合的致变突变。除非突变显性，或杂合亲代的 RNA 或蛋白足以回复后代突变，便能筛选出具有表型的 F2 突变体。根据孟德尔遗传定律及减数分裂时期染色体随机重组原理，选择的 F2 代突变体的多态位点除非与致变突变连锁，变异碱基的频率应为 1/2；连锁的多态位点与功能突变的物理距离近，相对不易发生重组。因此，出现突变株系碱基多态性频率显著偏离 1/2 的区域即为功能突变所在的基因组区域。

正向遗传学鉴定基因利用的遗传标记分别为限制性片段长度多态性（RFLP），微卫星序列（SSR）和单核苷酸突变（SNP）。SNP 检测的传统方法采用 PCR-单链构象多态性（PCR-SSCP）分析、变性高效液相（dHPLC）等，均需通过凝胶电泳进行分析。尽管正向遗传学方法十分有效，但它往往耗费大量时间，需构建大量品系用于作图，且因为造成表型的突变通常在某些遗传标记的附近，所以只能先定位到较大区域再对其中的位点逐个实验验证。测序可鉴定遗传标记并定量大量 DNA 样本中的等位基因频率，揭示核苷酸序列，而二代测序实现了同时对表型突变的定位与鉴定，将原本需要做一年的工作减少到

只需几周[2]。

2. 测序技术发展

突变指基因组中核苷酸序列的改变，因而用于确定核苷酸顺序的 DNA 测序自然对基因突变检测有重要帮助。但一开始测得的生物大分子是蛋白质和 RNA，且都采用先打断后解析的方法。直到 1968 年，噬菌体 λ 粘性末端的 12 个碱基被引物延伸法测得。1973 年，Gilbert 和 Maxam 将 DNA 序列复制为 RNA 再测序，耗时两年，成功测得乳糖抑制剂结合位点的 24 个碱基。

1976 年，利用与放射标记的碱基之间的距离确定核苷酸顺序的链终止法和化学切断法出现，实现了半天内解析几百个碱基。随后，利用随机克隆后测序，再基于重叠组装的鸟枪法和单链 M13 病毒克隆载体，人们开始能从头组装基因组。1987 年，基于荧光的自动化桑格测序仪出现，该设备一天能测得 1000 个碱基。随着测序数据指数增长和搜索工具如 BLAST 的开发，出现了如 GenBank 的数据中心，测序数据的价值增大，从而进一步激发了数据共享的激情。

2005 年，454 发行了第一款商用二代测序仪。相比一代测序，二代测序技术的优势有：通过体外扩增而非细菌克隆进行复制测序模板；同时平行进行百万计的测序反应而非一支试管一个反应；通过循环生化反应和成像而非测量片段长度进行测序。体外扩增的方法有如桥式扩增，使用固定引物，使扩增产物在模板周围成簇分布；也可在乳滴中 PCR，使每个模板的扩增产物都固定在小珠上；或使用滚环扩增，产生“纳米球”进行测序。经过快速发展期，2012 年以来，二代测序发展速度已放缓，illumina 独占市场优势[3]。如今，虽然二代测序读长依然小于 Sanger 测序，但其准确度已高达 99.9%，且能在两天内测得 10^{12} 数量级的碱基[4]。

理想的测序技术应原位、准确、没有读长限制。自上世纪 80 年代以来，人们为开发优于二代测序的方法不断做出努力，其中以 PacBio 公司的 SMRT 和 Oxford Nanopore Technologies（ONT）的纳米孔测序技术为代表的第三代测序技术实现了单分子测序。PacBio 能实时观察聚合酶合成过程。测序过程无需进行 PCR 扩增，平均读长达到 10Kb-15Kb，且读长不再相等。但三代测序的通量目前仍低于二代测序，错误率约 10%且随机

分布。ONT 利用单链 DNA 通过窄通道时产生的离子流反应链的一级结构。ONT 读长可大于 PacBio，并且其设备小如 USB，方便携带。虽然 ONT 存在如错误并不随机的问题，但发展迅速。

传统的 DNA 微阵列或 NanoString 方法依然可用于突变检测，但目前最常用的检测方法是二代测序中全基因组测序（WGS）和全外显子测序（WES）。虽然 WGS 目前比 WES 昂贵，但其价格下降比 WES 快，甚至在外显子突变检测中也更灵敏[5]，因此最为常用。

通过 WGS，研究人员可以快速获取全基因组上的突变信息，并基于这些信息和功能基因组注释，发现疾病高风险突变，由此开创了疾病基因组学时代。例如，通过 WGS，研究人员筛选到 18 个新的自闭症候选基因[6]，发现 0.6%的 ADCY7 错义突变使患溃疡性结肠炎的风险加倍[7]，高风险的髋骨关节炎与 COMP 和 CHADL 的异常基因型相关[8]，甚至获得了 Y 染色体连锁疾病的系谱[9]。考虑到基因组存在大量间隔重复序列 ，目前 WGS 使用双端测序的居多。WGS 用于变异检测时，一般通过增大测序深度能提高碱基质量值和覆盖度均一度[10]，减少测序错误。但仍有相当多的非测序因素造成假阳性突变，如：插入缺失处的比对错误[11]、参考基因组装错误或研究材料与参考组装的不同或低质量比对引入的 SNP[12]、比对到重复或旁系同源序列或未完整组装的基因组等[13]。检测出假阳性的数量甚至可能多于真实的诱变突变数量，因此需要可靠的过滤方法。

3. 模式生物秀丽隐杆线虫

秀丽隐杆线虫是正向遗传学中经常使用的一种模式生物。秀丽隐杆线虫体型小，约 1-1.5mm。它具有雄性和雌雄同体两种性别，雄虫虽少（0.02%），但可通过与雌雄同体的线虫交配增加到 50%。野生型 N2 秀丽隐杆线虫平均寿命 18-20 天，较高温度下，生命周期和寿命更加缩短，方便杂交传代。其解剖结构清晰，已知雌雄同体成虫有 959 个体细胞和 302 个神经元，透明的身体方便追踪细胞命运或荧光标记蛋白表达。1998 年，秀丽隐杆线虫成为首个测得完整基因组的多细胞生物，其多达 83%的蛋白序列在人中有同源序列，20250 个蛋白编码基因中 38%在人类中具有对应同源基因。因此自 1965 年以来，通过对秀丽隐杆线虫进行 EMS 诱变或 RNAi 处理的遗传筛选不断促进着包括人类生物学中生命过程的研究[14]。

具体而言，线虫遗传中心已注册秀丽隐杆线虫品种达 19 万，包括多种野生型和携带 9000 多种等位基因的秀丽隐杆线虫。秀丽隐杆线虫约表达 28146 种蛋白，基因组 GC 含量约 35.6%，有 5 条常染色体（I-V）和一条性染色体（X）。雌雄同体线虫是二倍体，而雄性只有一条 X 染色体（X/φ）。减数分裂中，除雄性 X 染色体外，同源染色体形成联会复合物，此时可发生交叉，且一般认为分离得到的重组体是单交叉产物，且重组率在环境温度 16-20 度间无显著变化，是用于遗传筛选的基础。

早在十年前人们就已经将 WGS 用于秀丽隐杆线虫功能突变鉴定，但最初 WGS 价格高昂，人们也尝试了如候选区域靶向测序(GIPS)等方法。随着 WGS 价格的降低，近年，秀丽隐杆线虫突变检测几乎都使用 WGS。WGS 能检测出一个线虫的几千个变异[15]，因而需要开发“测序作图”方法以减少候选变异数量。

利用测序作图选定定位区域的方法有很多，对于隐性突变，HA 变异作图[16]通过突变株与具有高度多态性的 CB4856 HA 线虫的远交，利用已知的 HA SNPs 和 WGS 同时对约 10^5个 HA 的 SNPs 与候选突变做连锁分析 。EMS-密度作图通过突变线虫回交，对近等基因系测序并预测突变间的连锁。该方法通过染色体重组边界定义定位区间，且重复回交试验两到三次能提高精确度。变异发现作图（VDM）中变异线虫与未变异亲本间进行回交或远交[17]，可对变异线虫背景中所有 SNPs 作图，提高作图精确度。去除突变线虫同代线虫的突变筛选表型相关基因[18]；或使用同样 EMS 诱导但无表型的线虫进行杂交，并从根据已有数据除去易错位点后进行分析[19]，也得到较好结果。对于显性、半显性和双基因突变，Andy Golden 等也开发出基于 WGS 的定位方法[20]。对于以上方法，利用回归分析（如局部加权回归散点平滑法（LOESS））或贝叶斯网络模型对图谱中的成千的数据点拟合回归线，都能提高作图精确度[21]。定位区域后，可根据致变试剂优先选择最可能的变异类型。如 EMS 作为诱变剂，则 G-A 和 C-T 转换应优先考虑。对于对开放阅读框无明显影响的变异，可在 UCSC Genome Browser 上查证不同物种在推定变异附近的基因组保守性。最后对候选基因进行 Sanger 测序验证后进行下游基因功能鉴定。

4. 本课题研究意义

全基因组测序（WGS）是对秀丽隐杆线虫等模式生物中对引起表型的突变作图的最

最经济快速方法。测序只需几天，成本在两千元内，避免了需要多年时间的基因克隆项目，减少了试剂成本。不仅如此，此方法可用于大规模的遗传筛选，然后对从中获取的许多突变体进行快速测序，从而加深人们对生命过程和基因的理解。此外，表型繁琐的突变体（例如，行为突变体）的作图品系或特定遗传背景的突变品系很难获得，也可通过全基因组测序识别其突变。然而，其数据分析非常复杂，需要专业的生物信息学知识。随着 WGS 成本持续下降并且技术变得普遍，所有使用遗传分析的实验室都需要建立自己的筛选平台。

本课题在实验室对线虫进行杂交和测序的基础上，对测序数据进行分析，以减少表型相关候选位点的个数。建立的方法也可以用于果蝇，拟南芥等模式生物的正向遗传学研究中。

第一章 数据和方法

本实验室已将从 EMS 诱变中分离到的 nrd-3 蛋白定位异常的线虫和野生型线虫进行杂交，得到 F1 代没有表型的线虫，随后让 F1 代线虫进行自交，从 F2 代线虫中挑出 20 到 50 只有表型的线虫，并从亲代线虫中挑取 20-50 只线虫作为对照。将两组线虫破碎提取基因组 DNA 后进行 WGS 测序，分别得到突变型（mutant）和野生型（wt）双端测序文件。本章搭建了从测序数据中找到突变和确定表型相关基因区域的流程。

1.1 数据与工具

本课题需要亲代野生型和子二代挑出的变异线虫测序数据，秀丽隐杆线虫野生型参考基因组数据及变异检测工具。

1.1.1 数据

1) 测序数据：wt 和 mutant 双端测序数据，FASTQ 格式。
2) 参考基因组：秀丽隐杆线虫 Bristol N2 品种 WBcel235 参考基因组文件与注释文件，从 ensemble 下载。

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	-

表 1 N2 strain WBcel235 参考基因组信息

1.1.2 工具

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 统计量

假设建库为 Gamma 抽样过程，测序为泊松抽样过程，理想情况下，F2 代不与假定的表型相关基因连锁的突变碱基频率应服从均值为 0.5 的负二项分布，连锁区域突变碱基频率应趋近 1。本次实验所测基因组池大小未知，无法使用如隐马尔可夫模型推知重组事件数，因而选择定义统计量的方式，通过统计量有效区分连锁和非连锁区域。

令 ref 为测得位点与参考碱基相同的 read 数，alt 为与参考碱基不同的 read 数。考虑到亲代线虫可能并非完全纯合，F1 代突变位点并非全部在一条链上，理想情况下，F2 突变线虫非连锁的碱基测得的 ref 应与 alt 相等；若突变位点与假定的致变突变位点处于互引相，则 alt 大于 ref，且连锁度越大，alt 相比 ref 越大；若处于互斥相，则相反。据此定义统计量 R 为$R = \begin{cases} \max(ref, alt) - 1, & ref \times alt = 0 \\ \frac{alt}{ref}, & alt > ref \\ \frac{ref}{alt}, & ref > alt \end{cases}$，R 值越大，连锁程度越高，并保留变异碱基频率（$\mathrm{V} = \frac{alt}{alt+ref}$）和深度不敏感的标准化突变碱基频数（$\mathrm{NV} = 100 \times \mathrm{V}$）及突变碱基频率差($\mathrm{D} = \frac{|\mathrm{alt}-\mathrm{ref}|}{\mathrm{alt}+\mathrm{ref}}$)用于方法验证。

1.3 流程

将诱变线虫与非亲本线虫杂交，F1 代为杂合体，选择 F1 自交产生的 F2 代具有表型的线虫，则纯合突变会因与造成表型的位点的连锁度出现不同程度的分离，产生等位基因频率的改变。

检测到诱变线虫中的纯合突变可能直接由 illumina 测序错误造成，或属于线虫原本的遗传背景，只有部分由诱变造成，未从诱变产生的突变中鉴定功能突变，应对测序数据进行质控，对检测出的变异进行过滤并通过定义合适的统计量对连锁度进行表征，选择相关基因区域并对其中突变基因进行注释，给出最有可能是造成表型的基因列表。最终，搭建的流程分为数据质控、数据预处理、变异检测、表型相关基因区域的定位和基因注释与进一步筛选五个主要部分。

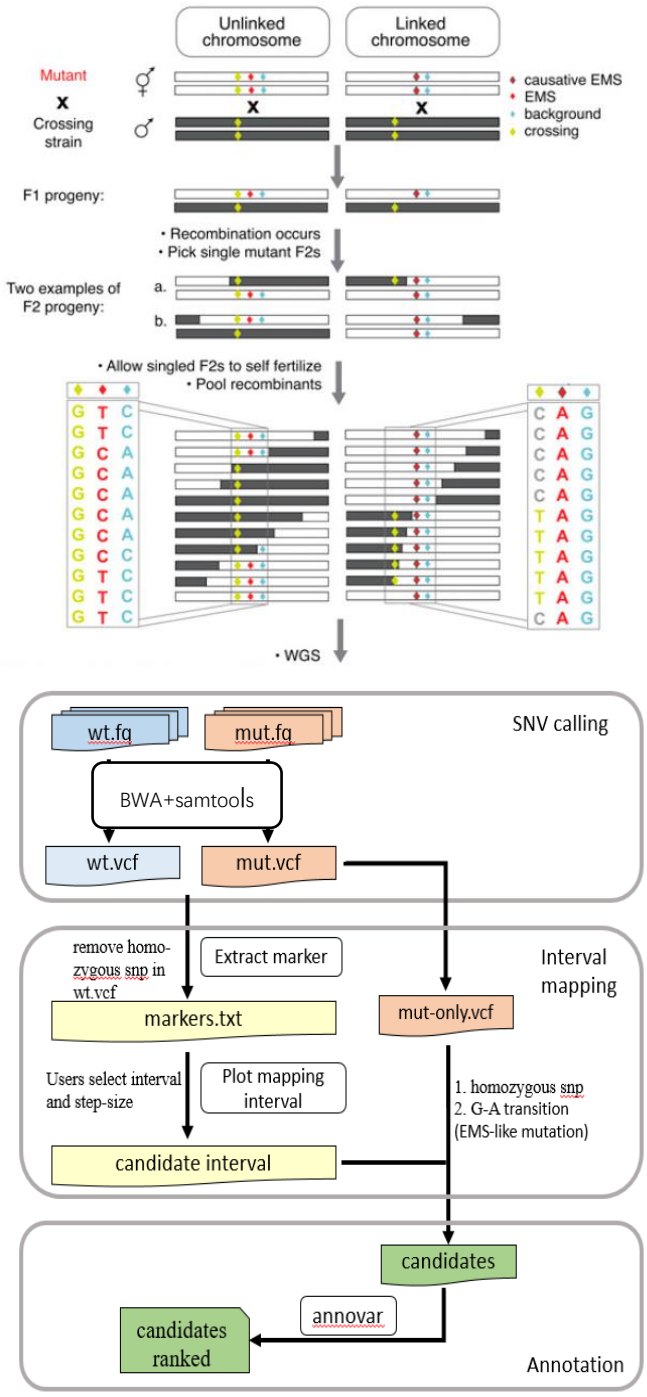


图 1 课题流程

1.3.1 原始数据质控

首先编写脚本检查质量体系。再使用 FastQC 进行数据质控，查看碱基质量值分布、GC 含量分布、N 含量、接头序列、重复率。随后指定接头序列、双端测序、质量体系参数后使用 trimmomatic 去除测序接头和低质量序列，最后再次使用 FastQC 检查质量。

1.3.2 数据预处理

首先使用二代测序序列比对工具 BWA 进行测序数据的比对，将来自原基因组的 reads 比对到线虫参考基因组上，找到其位置并排序。为此，需对参考基因组进行 Burrows Wheeler 变换并添加 FM 索引，以便序列比对以线性复杂度进行搜索和定位，同时使用 samtools 添加 faidx 索引，方便变异检测。然后根据数据质控输出的序列读长信息，选择 bwa mem 比对方法，完成比对后使用 samtools -q 参数筛选比对质量大于 30 的数据，并输出为比原始数据增加位置、比对质量和比对结构信息的 bam 文件。

随后使用 samtools sort 命令进行排序，picard MarkDuplicates 命令标记重复序列。二代测序往往包含 PCR 扩增，以增大打断的 DNA 片段密度，因此初始 DNA 片段密度的不均，DNA 打断和 PCR 过程中的碱基错误及 PCR 反应对模板扩增的倾向性不同，这些非真信号都会在扩增过程中被放大。后续的变异检测依据贝叶斯原理，并不区分重复序列。因而需在 BAM 文件的 FLAG 信息中标记，在变异检测过程忽略重复序列，以消除其带来的检测错误。为随机访问文件中位置，为标记后的文件使用 samtools index 工具添加索引。

数据预处理后目录下应存在 sorted.markdup.bam 文件，sorted.markdup.ban.bai 文件和 markdup_metrics.txt 文件。

1.3.3 变异检测

使用 samtools mpileup 命令进行变异检测，vcftools --remove-indel 命令去除检测出的插入缺失变异。相比 GATK，samtools 在变异检测时默认不丢弃低比对质量的测序数据，避免丢失可能的突变位点；samtools 的 SNP 基因型似然性模型基于测序错误间不独立；samtools 能通过输出调整 MQ 参数来校正质量值，对于插入缺失带来的假单核苷酸突变，应用隐马尔科夫模型得到碱基比对质量（BAQ）来降低影响，如果在次优的比对中，某碱基比对给不同的参考碱基，该 BAQ 值很小，对 SNP 检测的贡献变小。这种方法相比 GATK 使用大量已知变异集基于机器学习算法校正碱基质量值更简便快捷，GATK 更适合人类疾病诊断。

编写脚本处理野生型和突变线虫的 SNPs_only.vcf 文件，突变文件中过滤掉野生型纯合突变，得到所有候选位点。

1.3.4 候选区域选择与注释

首先对检测出的变异进行质控。用 R 中 car 包去除测序深度和检测质量比测序深度异

常的变异位点，得到 SNP 标记。计算 SNP 标记的 R 值并用 R 中 ggplot2 包做的散点图和 loess 平滑曲线，以便选择候选区域。如实验人员有候选区域长度预期，可用枚举法取滑动窗口得到统计量均值最大的区域为候选区域。

为基于基因对候选区域进行注释，首先构建注释文件。用 gtfToGenePred 工具将 gtf 格式的基因组信息文件转换为 GenePred 文件，然后用 retrieve_seq_from_fasta.pl 转换为 FASTA 文件。最后使用 table_annovar.pl 得到注释文件。

1.4 本章小结

本章根据课题目的，设计了用于选择致变突变可能存在的区域的统计量，准备了课题需要的数据集，并搭建了整套流程，建立了工作目录。将质量控制写于 QC.py，数据预处理及变异检测写于 call variation.py，统计量计算、区域筛选与作图写于 main.py，对重复或不同样本进行本课题工作仅需在命令行输入样本名及工作参数顺序执行脚本命令即可完成。

第二章 结果与讨论

根据已定义的统计量和总体流程，编写脚本实现数据质控、数据预处理、变异检测、表型相关基因区域的定位和基因注释。本章展示主要结果并进行讨论。

2.1 突变检测

2.1.1 原始数据质控

检查验证 illumina hiseq2000 使用的碱基测序质量体系为 Phred33，去除测序接头和低质量序列后，数据质量控制的结果显示，所有测序结果的碱基质量值均高于 30；每个读段的质量均值在 39 左右；GC 含量稳定在 35%，与线虫基因组一致，且嘌呤和嘧啶含量分别匹配；没有未知碱基；测序长度 150bp；测序重复水平小于 2 且已去除接头序列,提示测序数据质量很高（表 2-1）。

样本	总序列数	低质量/%	读长/bp	%GC	平均测序深度	平均覆盖度
mut-1	22088348	0	150	35	I:21.6; II:21.5; III:21.5;	I:98.9%; II: 98.8%; III:99.1%;
mut-2	22088348	0	150	35	IV:21.5; V:21.4; X:21.3	IV:98.1%; V:97.4%; X:99.3%
wt-1	15264164	0	150	35	I:54.8; II:53.1; III:53.0;	I: 96.4%; II:96.6%; III:96.8%;
wt-2	15264164	0	150	35	IV:53.4; V:53.7; X:53.9	IV:95.3%; V:95.1%;X:97.6%

表 2-1 测序质量摘要

2.1.2 数据预处理

在保证比对质量大于 30 的情况下，将双端测序数据回贴到参考基因组上，野生型和突变线虫数据均全部回贴到参考基因组上，其中 99.69%成对序列成功配对，0.04%为单端回贴，提示回贴成功。野生型平均测序深度 21.5x,平均覆盖度 96.28%；突变线虫平均测序深度 53.6x，平均覆盖度 98.59%，深度和覆盖度可用于突变检测。比对后进行排序和重复序列标记，发现野生型数据估计文库大小为 430805547，其中 3465 条未配对，配对读段中 1746680 条标记为重复。重复序列占比 12.39%，其中光学重复 1566114 条。突变线虫预计文库大小 629925599，除 7342 条序列未成功配对的读段，有 2401718 条配对重复序列，占比 11.9%，标记重复成功(表 2-2)。

样本	未配对 reads		配对 reads		未比对 reads	重复率	估计文库大小
wt	重复 0	0	重复 1190706	6936918	0	14.65%	230797868
mut	重复 7342	14069	重复 2401718	17860064	0	11.87%	629925599

表 2-2 预处理质量摘要

在本次实验体系中，从图 2-1 可看出，测序深度达到 15x 时，由于重复序列使得数据反映的测序深度小于实验深度，测序深度达到 60x 以上时，增大深度意义不大。

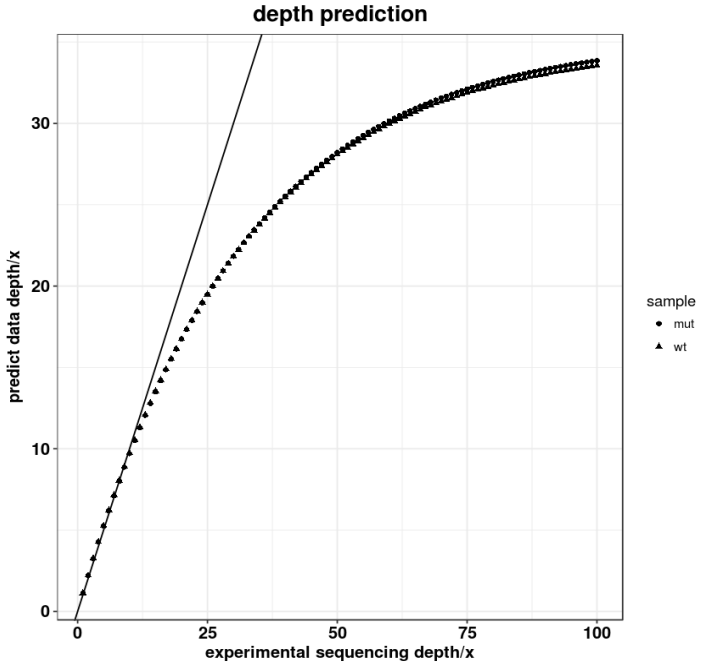


图 2-1 实际与预期测序深度

2.1.3 突变检测

使用 samtools 分别检测到突变线虫和野生型线虫所有突变。突变线虫 15158 个位点过滤掉插入缺失变异后剩余 12795 个单核苷酸变异位点，野生型从 14708 保留了 12336 个位点。得到的单核苷酸突变的中，有 9711 个位点是突变型和野生型共有的，3084 个位点只在突变线虫中检测到。只在野生型线虫中检测到的突变有 2625 个，其中有 1695 个是纯合位点不可能导致表型，除去后得到变异线虫的共 10641 个候选突变位点。这些位点的测序数据影响 R 值，首先进行质量控制。

由测序深度的初步分位信息可知约 90%的测序序列深度在 200x 以内，但有的却超过了 7000x，碱基突变检测质量比深度值 90%的数据在 7 以下，有的却高达 27（附图 1），

它们既是数据上的异常，也与本课题前期实验测序的线虫只数（20-50）不符，因此，需

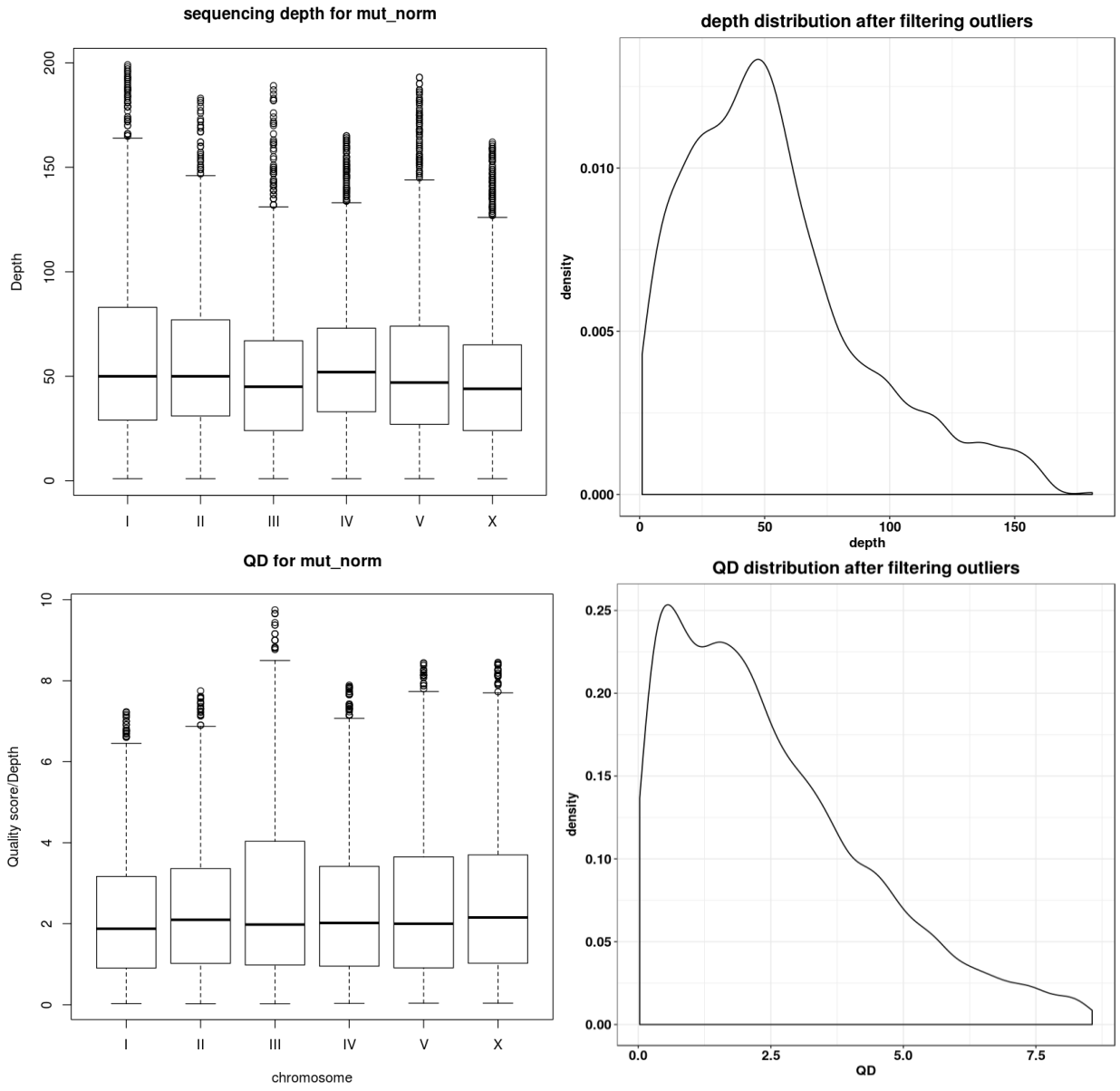


图 2-2 SNP 标记的测序深度与突变检测质量

继续探索数据以筛除异常值。从数据中得到上下四分位数、内限数据和离群值，筛除测序深度大于 199x，碱基质量值比测序深度在大于 9.75 的离群值，从中保留测到多于 10 次的位点，共 86.3%的数据（图 2-2）作为 SNP 标记进行后续区域选择。

2.2 筛选候选基因

2.2.1 候选区域的选取

计算挑选出的 SNP 标记处的 R 值并作散点图，经验性选择数据子集范围（span=0.2）作局部多项式回归拟合(LOESS)线。由图 2-3 反映的连锁程度可以看出表型相关基因位于

三号染色体上，并可选取候选区域范围 4-10Mb（基因重组指数 0.06）。将所选范围内 65 处基因突变按标准质量值排序，其中位于外显子的 A-G 转换非同义单核苷酸突变基因有 12 个（表 2-3），其中 5 个基因编码的蛋白已有报道。随后我们对 12 个基因进行 siRNA 干扰验证其中确有使表型回复的基因。

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

表 2-3 候选基因 *已有编码蛋白功能报道基因

2.2.2 方法验证与讨论

检验所有变异位点测序深度分布服从 Gamma 分布（p-value < 2.2e-16）；除三号染色体外，其他染色体上的突变碱基频率可认为同分布（Kruskal-Wallis rank sum test，p=0.000293），三号染色体显著的高（Kruskal-Wallis rank sum test，p=0.3675），有理由认为表型相关基因位于三号染色体上。

用枚举法对选择区域进行验证。三号染色体总长 13.78MB，从 R 值作图预期找 6MB 的候选区域，对 R 值随位点取 6Mb 的窗口，0.1 的滑窗，找到 R 在窗口内均值最高的正是 4MB-10MB。对所有数据随机等数据量重采样 100 次，得到近 5 成的结果在 4-10MB，超 8 成的结果在 3.5-10MB 之间，认为方法具有足够的稳定性（附图 3）。另外，NV 值（100*variation frequency）大于 48 可认为检测到突变碱基次数并非由碱基频率 0.5 的总体中抽样得到（p=0.05），此次数据的各染色体 NV 值分布于选取区域的 NV 值也并未出现异常（附图 2）。

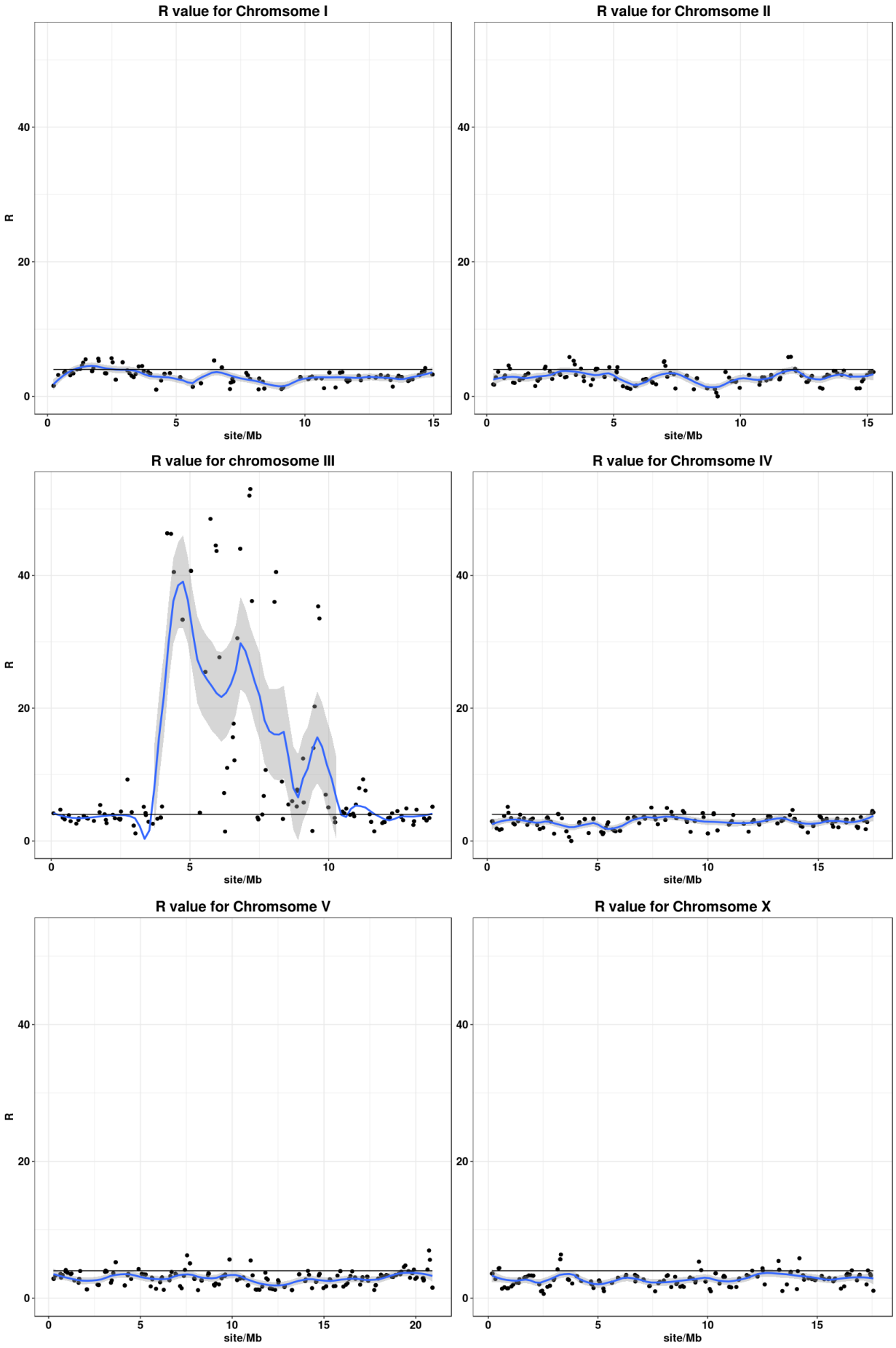

R value for Chromsome I
R value for Chromsome II
R value for chromosome III
R value for Chromsome IV
R value for Chromsome V
R value for Chromsome X
R
site/Mb

图 2-3 R 值及 LOESS 回归（span=0.2） 蓝线：loess 回归线；黑线：R=4

与其他统计量相比，R 值在 alt 或 ref 为 0 时一定程度地受测序深度影响，但其同时考虑了顺式和反式的情形，取值在高度连锁和不连锁时的阶不同，方便更直观地锁定候选区域。但 R 值大小在碱基频率很大的位点间相对不反应碱基频率大小，理想情况下，其他统计量仍以较好的线性关系反应连锁度，但没有拉大非同类数据距离和相对减少同类数据距离的作用。直接采用突变碱基频率作图需要较大程度的 loess 回归，带有较多信息丢失，且分辨率不高，在非理想实验条件，即 F2 代携带由亲代野生型引入的突变的情况下无法正确确定定位区域。综合本课题中不同统计量的结果（附图 4）和其他文献结果[17]，推荐应用 R 值来选择候选区域。

2.3 讨论与展望

本课题目标是找到与假定表型相关基因高连锁的染色质区域，理想情况下亲本线虫是纯合体，F1 线虫杂合，F2 非连锁的位点经高通量测序测到的突变频率应服从 0.5 的不放回抽样($\frac{(alt-ref)^2}{alt+ref} \sim \chi 2(df=1) > 9, p < 0.01$)，实测数据结果与之存在系统偏差(突变碱基频率负二项拟合接近 0.35)，可能由于亲代突变线虫并非完全纯合，但不影响相对高连锁区域的选择。

利用二代测序定位表型相关突变的一个限制在于测序是否覆盖目标突变位点，计算表明，测序深度达到 20x，检测点突变的灵敏度能达到 89%，外显子区突变高达 95%，本课题对突变线虫的测序深度达到 50x，一个编码区可能存在多个突变，应较好地覆盖了表型相关突变。另一个可能的问题是在难以测序或覆盖度低的区域检测出假阳性或假阴性的突变，本课题方法相对稳健，突变检测时未通过质量值进行筛选，扣除的只是分别在野生型和突变型都检测到的纯合突变，保留了突变线虫其余所有突变，降低了遗漏突变的风险。在统计量作图锁定表型相关突变区域时选择了测序深度和碱基质量值比测序深度在数据正常范围内的，保证区域选择准确。但可能存在仅与功能突变连锁不平衡而非物理距离小的情况，但没有集中于一个区域。另外，二代测序最难检测的造成表型的原因是基因重复，很难区分检测到短序列数量的变化是否有基因拷贝数的影响。研究表明多拷贝的转基因相比邻近区的测得的短序列数多很多，但检测基因的单个复制事件依然十分困难，当发现基因的双端测序结果比对到完全不同的基因组区域时，提示可能该处有基因复制发生。本课题待选突变由 EMS 诱变产生，表型不太可能由基因复制造成，数据中也未发现此现象。

最后，功能基因可能位于秀丽隐杆线虫基因组注释错误的区域，本课题注释出的基因在不同线虫基因组中未发现异常，其中经 RNAi 干扰确有使表型回复的基因。综上，可认为本课题成功建立可靠的方法从上万的突变中定位到仅 12 个候选基因，大大减少了互补实验和致变突变定位需要的实验工作量。

正是传统遗传学和快速发展的测序技术的结合促生了鉴定导致表型的基因的新方法。也正是通过对秀丽线虫等模式生物的研究对这些新方法进行概念验证。基于二代测序的方法可以快速对感兴趣的突变定位和克隆，从而革新了正向遗传学筛选。不仅如此，这些方法还给我们提供了各种背景突变和表型突变体。今后，随着三代测序技术的进一步发展和读长的进一步增长，变异和突变鉴定的可靠性有望进一步提高。

2.4 本章小结

本章根据已建立的方法流程，将测序的野生型和突变型线虫的突变位点检测到，从突变线虫变异位点中扣除野生型线虫的纯合突变的位点后的 10641 个位点中，选择了位于三号染色体 4Mb-10Mb 区间的位点，经注释后得到 12 个编码区基因。并经枚举法及概率检验的验证，对比了多种统计量，得到满意的结果。筛选得到的 12 个基因经 RNAi 回复实验，其中找到了表型相关基因。

参考文献

[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]. 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.

致谢

大学四年接近尾声，我要忠心感谢王旻教授和宋晓元教授在实习过程中对我的悉心指导，以及王旻教授作为拔尖计划中我的指导老师两年来在学习生活中的无私帮助。三年来，是王老师的支持使得我能有将想法变为现实的实验平台，有校际及出国学习的机会，也是王老师从导师课到实验室的教导使我受益匪浅。

感谢罗晨老师、虞璐婷师姐、张志远师兄、杜佩师姐、李尤杰师兄和李想师兄在生活和研究中给予的关心和帮助。是他们教给我各种实验和生物医药科研的思考方式，也是他们对我的信任使我迅速成长。

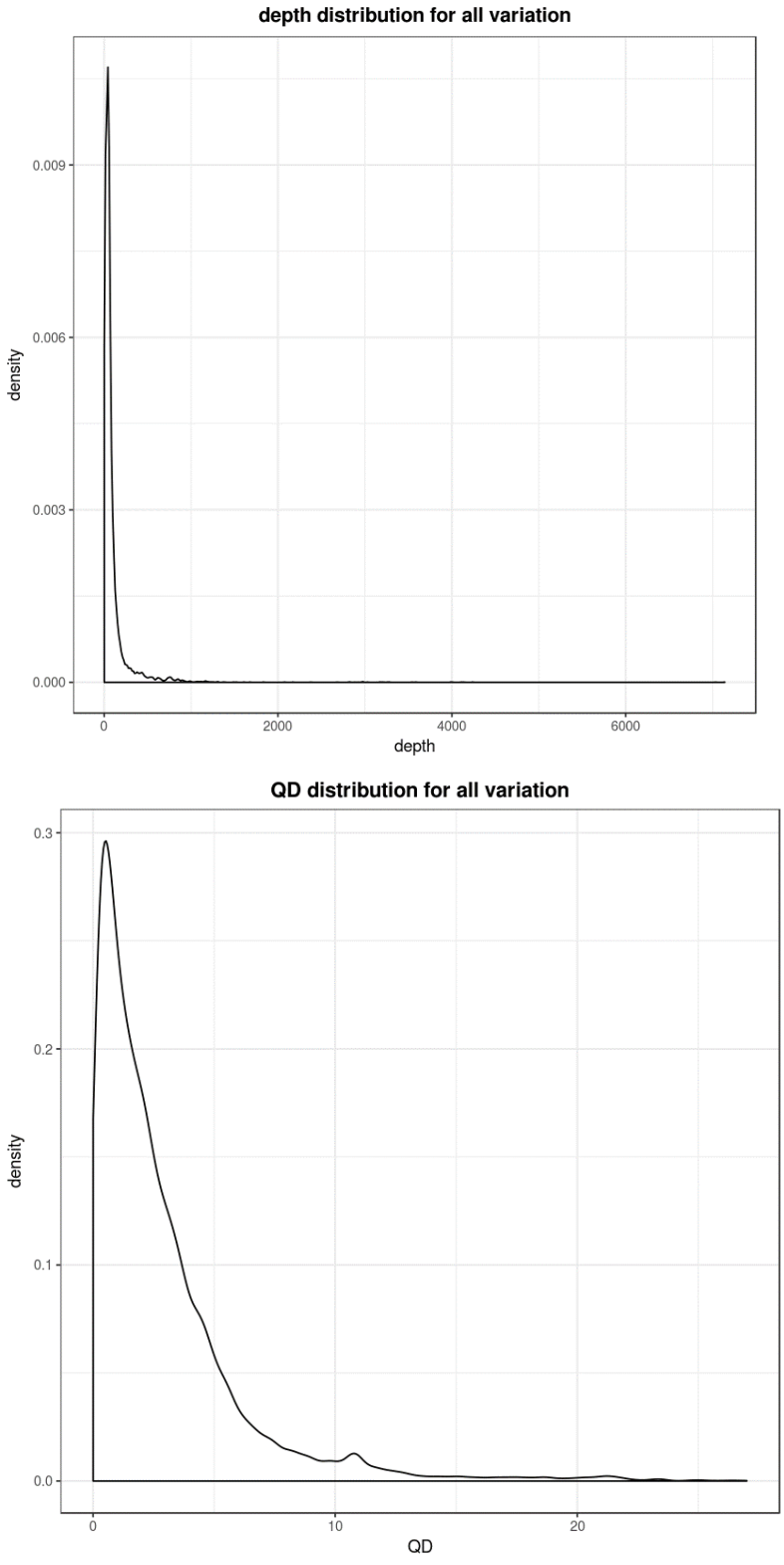
感谢中国科学技术大学宋晓元导师组及 BSC 小组全体成员对我的毕业论文中重难点的建议与帮助，认真为我答疑解惑，及在我毕设期间对我生活上的关心。

感谢李亮、张晶鑫等实习同学对我的大力支持和帮助，我们一起学习，一起提交材料，互帮互助，一起成长。在此对大家表示忠心的感谢。

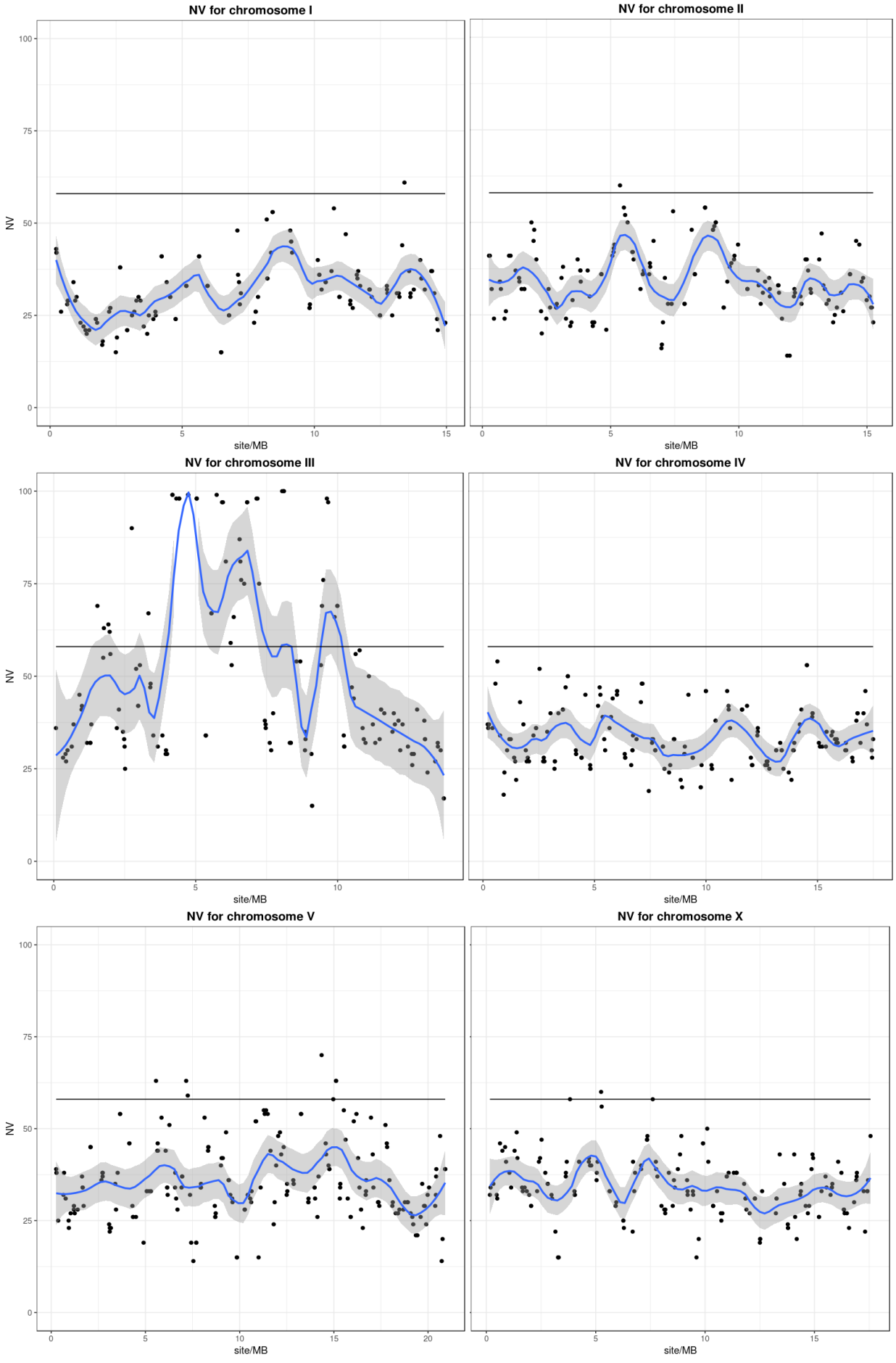
最后，感谢生命科学与技术学院老师们的辛勤付出，本科向您们学习的时间不够多，今后专业上的问题可能还会请教。感谢家人和朋友们的支持，感谢那些和我朝夕相处的同学们的互相激励，让我们共同成长！

附录

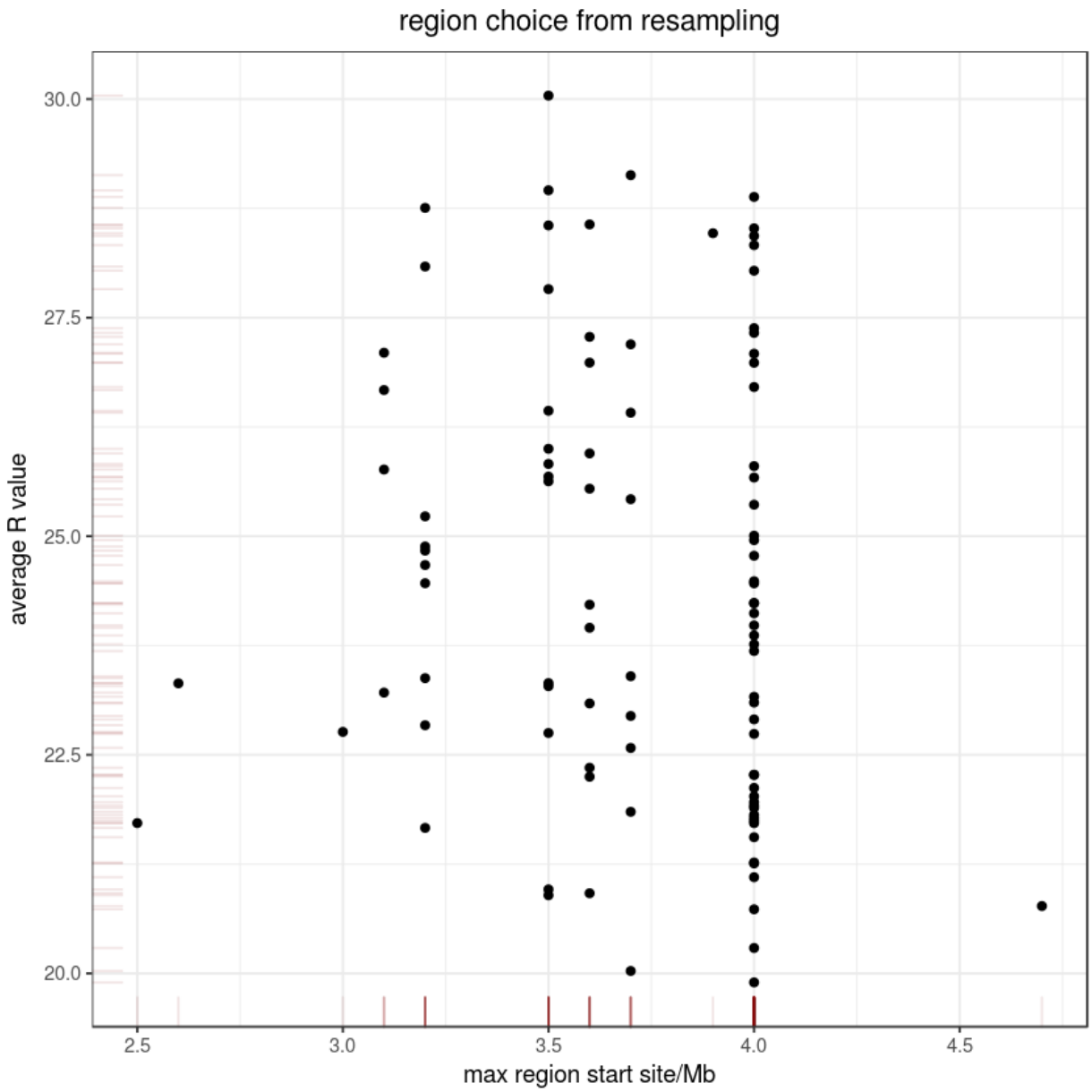
附件 1 QC.rar



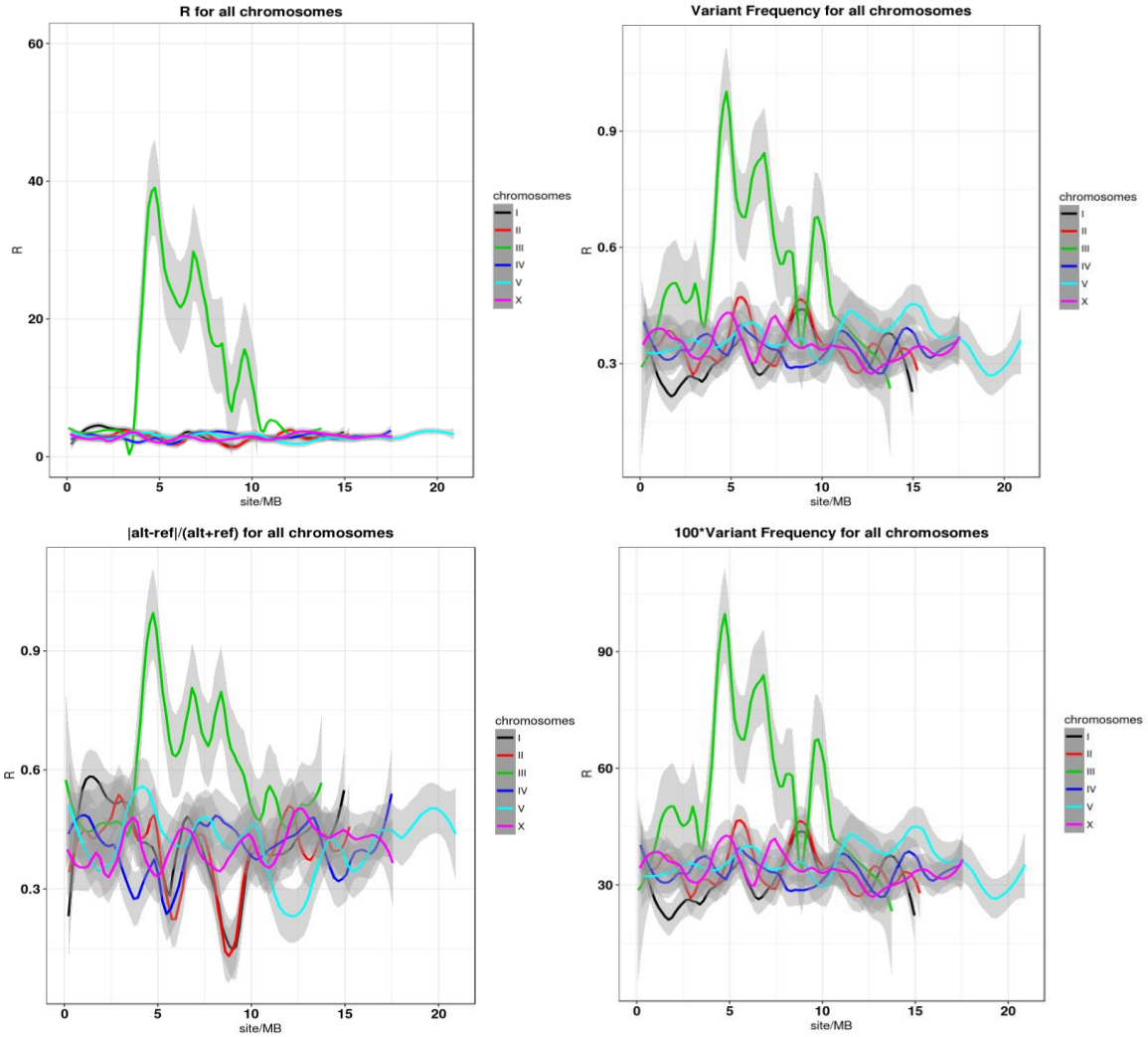
附图 1 变异线虫多态性位点测序深度和碱基质量值比测序深度



附图 2 NV=100*alt/(alt+ref)



附图 3 重采样 100 次突变位点区域在 III 染色体上起始位点


R for all chromosomes
Variant Frequency for all chromosomes
|alt-ref|/(alt+ref) for all chromosomes
100*Variant Frequency for all chromosomes
R
site/MB
chromosomes
I
II
III
IV
V
X

附图 4 不同统计量结果