派森诺
Personalbio

解析基因序列·诠释生命密码·改善人类生活

派森诺 | Personalbio

项目结题报告书

项目名称: BSA性状定位分析报告

委托单位: 云南农业大学

制定日期: 2025年02月13日

上海派森诺生物科技股份有限公司
SHANGHAI PERSONALBIO TECHNOLOGY CO.,LTD.

1.1 项目信息

项目信息

Search

项目编号	PN20250103032
开题单号	GPL2025011083
项目类型	BSA性状定位
样本数目	4
测序平台	Illumina NovaSeq
完成日期	2025年2月11日
技术支持	(genome_support@personalbio.cn)

共 6 条

1.2 分析项目及基本要求

分析项目

Search

序号	分析项目	类型	分析
1	数据过滤、整理及质量评估	A	√
2	数据质控	A	√
3	高质量数据获取	A	√
4	参考基因组整理	A	√
5	序列比对	A	√
6	序列比对结果统计	A	√
7	群体SNP检测与注释	A	√
8	群体InDel检测与注释	A	√
9	SNPindex计算及分布	A	√
10	ED计算及分布	A	√

共 14 条 10 条/页 ∨ ‹ 1 2 › 前往 1 页

1.3 实验建库流程

1.3.1 概述

从DNA提取到上机测序，需要严谨、可靠的样本检测、质控流程，在每一环节都对样本质量严格控制，才能确保测序数据的真实可信。



实验建库流程图

1.3.2 基因组总DNA提取

对于各种不同来源的基因组样本，根据过往项目经验，选用最合适的基因组总DNA提取方法，并通过0.8%琼脂糖凝胶电泳检测DNA提取质量，同时采用紫外分光光度计对DNA进行定量。

1.3.3 测序文库制备

以Illumina NovaSeq测序为例，采用Illumina公司的TruSeq DNA PCR-free prep kit试剂的标准建库流程制备测序文库。

1) 首先对提取的DNA采用超声随机打断，进行序列末端修复，通过试剂盒中的End Repair Mix2切除DNA序列5’端的突出碱基，同时添加一个磷酸基团、补齐3’端的缺失碱基；

2) 在DNA序列的3’端添加A碱基以防止DNA片段自连，同时保证目标序列能与测序接头相连（测序接头3’端有一个突出的T碱基）；

3) 在序列5’端添加含有文库特异性标签（即Index序列）的测序接头，使DNA分子能被固定在Flow Cell上；

4) 采用BECKMAN AMPure XP Beads，通过磁珠筛选去除接头自连片段，纯化添加接头后的文库体系；

5) 对上述连上接头的DNA片段进行PCR扩增，从而富集测序文库模板，并采用BECKMAN AMPure XP Beads再次纯化文库富集产物；

6) 通过2%琼脂糖凝胶电泳，对文库做最终的片段选择与纯化。文库插入片段在~450bp。

1.3.4 上机进行高通量测序

1) 上机测序前，需要先对文库在Agilent Bioanalyzer上进行质检，采用Agilent High Sensitivity DNA Kit。合格的文库有且只有单一的峰，且无接头二聚体。

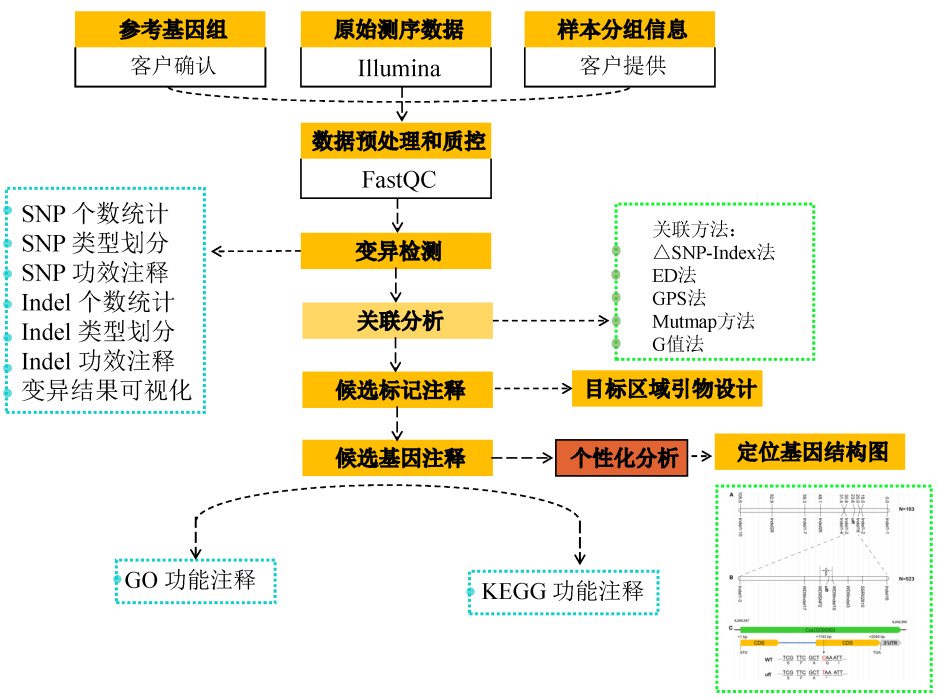
2) 之后，采用Quant-iT PicoGreen dsDNA Assay Kit在Promega QuantiFluor荧光定量系统上对文库进行定量，合格的文库浓度应在2nM以上。

3) 将合格的各上机测序文库（Index序列不可重复）梯度稀释后，根据所需测序量按相应比例混合，并经NaOH变性为单链进行上机测序；

4) 使用NovaSeq测序仪进行2×150bp的双端测序。

1.4 分析流程

首先对高通量测序的原始下机数据根据序列质量进行初步评估与过滤，获得高质量数据。将高质量数据比对到参考基因组，统计比对信息。鉴定变异位点，统计变异的信息，绘制具备论文发表水准的变异位点分布图，全方位、客观地呈现以上分析结果。对变异的信息做注释，并以图表的形式展示相应的注释结果。定位控制性状的候选区间，对候选位点做注释，对相应的候选基因做GO、KEGG注释。并在候选区间内开发分子标记，设计引物。



分析流程图

2.1 测序概述

本项目构建插入片段为400的文库，利用第二代测序技术（Next-Generation Sequencing，NGS），基于Illumina NovaSeq测序平台，对这些文库进行双末端（Paired-end，PE）测序。每个样本构建1个文库，文库的基本情况见表1。

表 1 测序概述

Search

Sample	Insert Size	Sequencing platform	Sequencing Mode
All	400	Illumina NovaSeq	Paired-end, 2×150bp

共 1 条

注:

Sample： 样品名称；

Insert Size： 库插入片段长度；

Sequencing Platform： 测序平台；

Sequencing Mode： 测序模式。

2.2 下机数据统计

分别对每个样品的下机数据进行统计，结果见表2。

表 2 测序数据统计

Search

Sample	Reads_Num	Total_Bases(bp)	N(%)	GC(%)	Q20(%)	Q30(%)
DXY	345,925,474	52,234,746,574	0.00	36.54	98.47	95.79
K	464,744,398	70,176,404,098	0.01	36.69	98.60	96.07
G	504,089,154	76,117,462,254	0.01	36.71	98.54	95.91
J16	369,782,784	55,837,200,384	0.01	36.52	98.57	95.97

共 4 条

注：

Sample： 样品名；

Reads Num.： Reads总数；

Total Bases(bp)： 碱基总数；

N(%)： 模糊碱基所占百分比；

GC(%)： GC含量；

Q20(%)： 碱基识别准确率在99%以上的碱基所占百分比；

Q30(%)： 碱基识别准确率在99.9%以上的碱基所占百分比。

结果目录：1_RawData

2.3 数据质控

2.3.1 碱基质量分布

采用FastQC（[http://www.bioinformatics.babraham.ac.uk/projects/fastqc]）对数据进行质量控制。以其中一个样品R1为例，具体结果分别见图1、2、3、4。

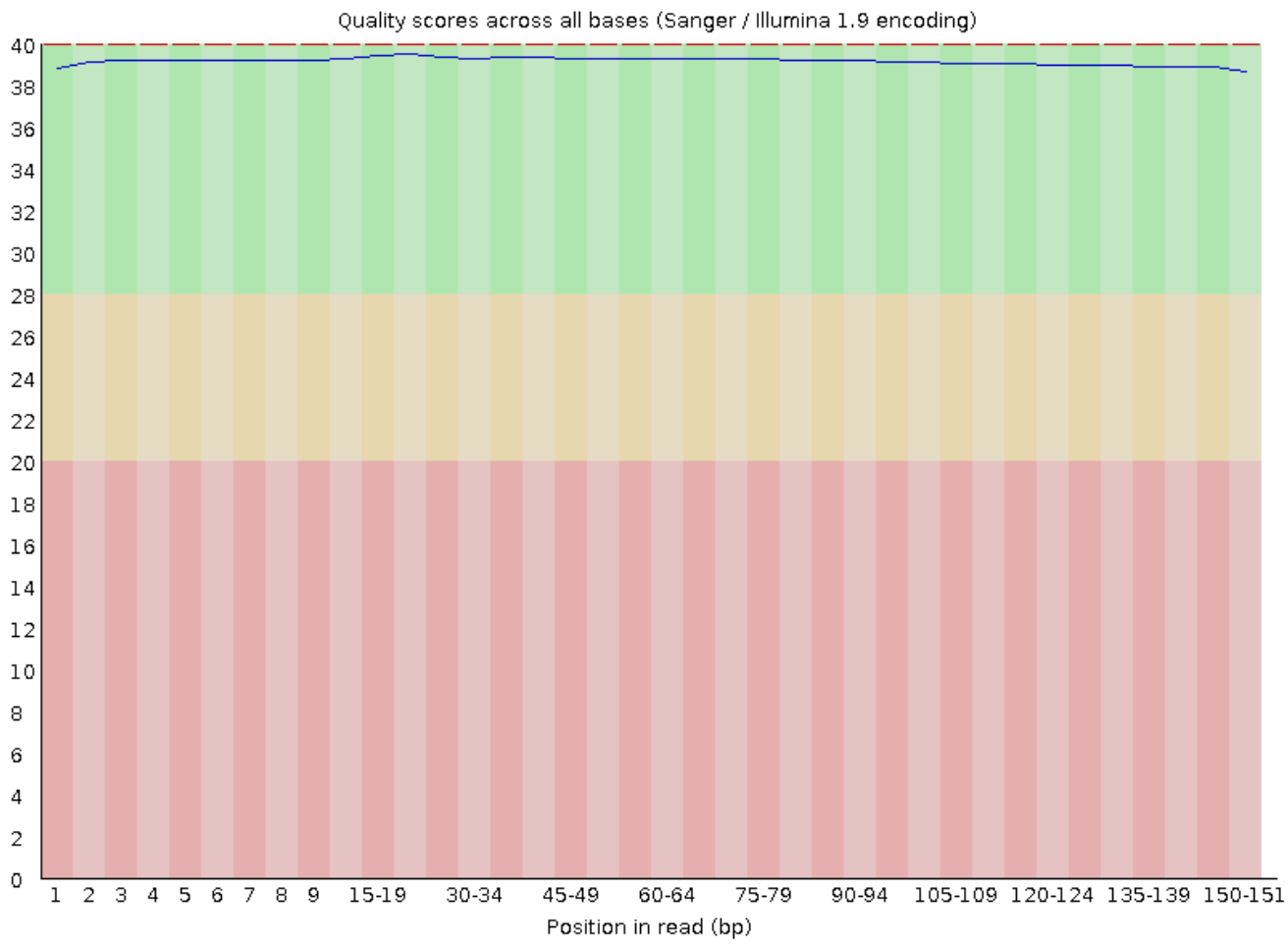


图1 单碱基质量分布图

横坐标是reads碱基位置（5'->3'），纵坐标是所有reads在该位点碱基Q值统计。红线代表中位数，蓝线代表平均数，黄线代表25%-75%区间，触须是10%-90%区间。一般而言，reads的5'端和3'端的碱基质量较低，中间部分的碱基质量较高。从图中可知，本次测序过滤后的数据平均质量很高。

2.3.2 Base content分布

Base content分布主要用于检测测序数据是否存在AT、GC分离现象。通常，这种现象由建库或测序引入，并影响后续的生物信息分析结果。Base content分布结果见图2。结果显示，建库和测序表现出较好的均一度，可以进行后续信息分析。

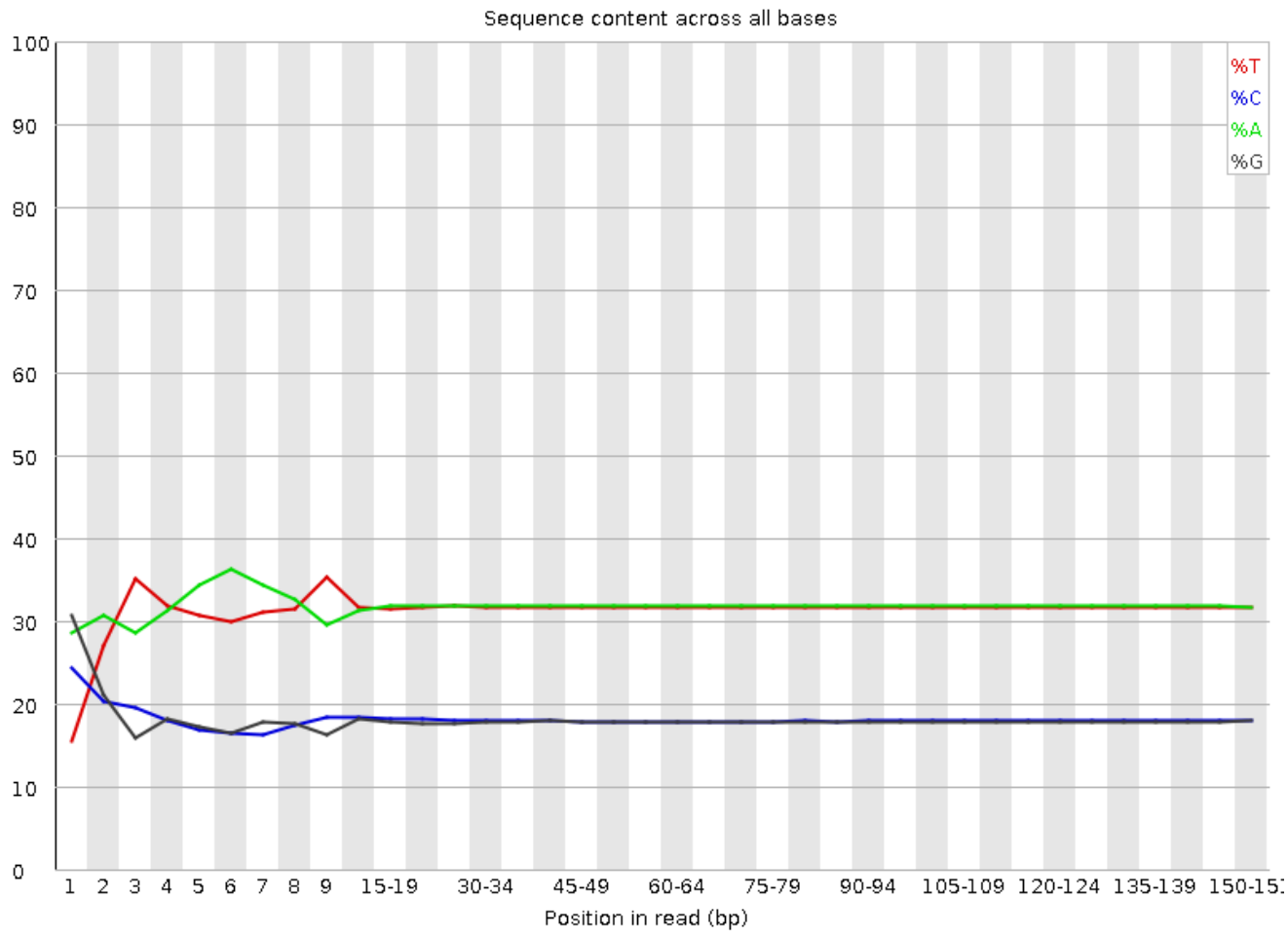


图2 base content分布图

横坐标是reads中碱基位置（5'->3'），纵坐标是该位点某碱基所占的比例统计。对基于Illumina测序平台的基因组测序中，基因组片段化过程中前几个位置的核苷酸组成表现出一定的偏好性，而这种偏好性与测序的物种和实验室环境无关，但会影响基因组测序的均一化程度。除此之外，其余位置四种碱基的出现频率应该是接近的。因此，好的建库和测序样品的四条线应该平行且接近，当部分位置的碱基出现Bias时，则四条线在某些位置纷乱交织，往往提示有overrepresented sequence的污染，当所有位置的碱基比例一致地表现出Bias时，即四条线平行但分开，往往代表文库有Bias（建库过程或本身特点、或者是测序中的系统误差）。

2.3.3 GC分布

GC content分布主要用来检测测序数据的GC分布是否正常。曲线形状的偏差往往提示存在不均一的序列组分，这通常是由于文库的内源或外源污染，通过不同的峰值，大致能够判断污染性质和来源。GC content分布结果见图3。GC content分布结果显示理论分布与实际分布一致，可以进行后续分析。

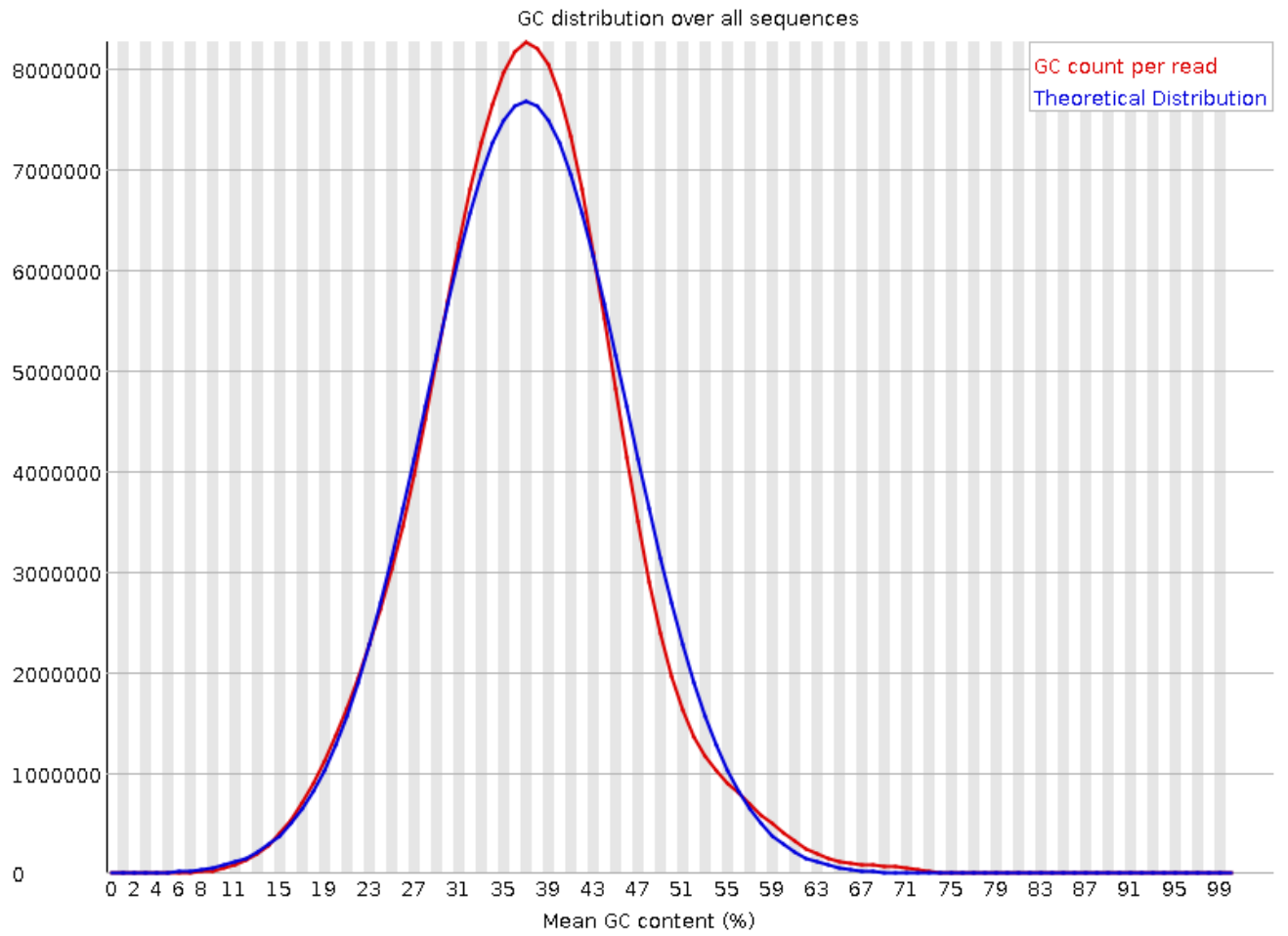


图3 GC content分布图

横坐标是GC比，纵坐标为对应的 reads 数量。红色表示的是实际分布曲线，蓝色的是理论分布曲线（峰值与物种特性有关）。

2.3.4 Sequence base quality

Sequence base quality 主要用来检测测序数据的平均质量分布情况。峰尖代表测序质量，峰宽表示测序整体质量分布，较大峰宽或者前拖尾峰表示测序数据的部分数据质量偏低，峰尖对应的值较低表明整体测序结果较差。本次测序的峰值在Q36左右，表示测序质量较高。Sequence base quality结果见图4。

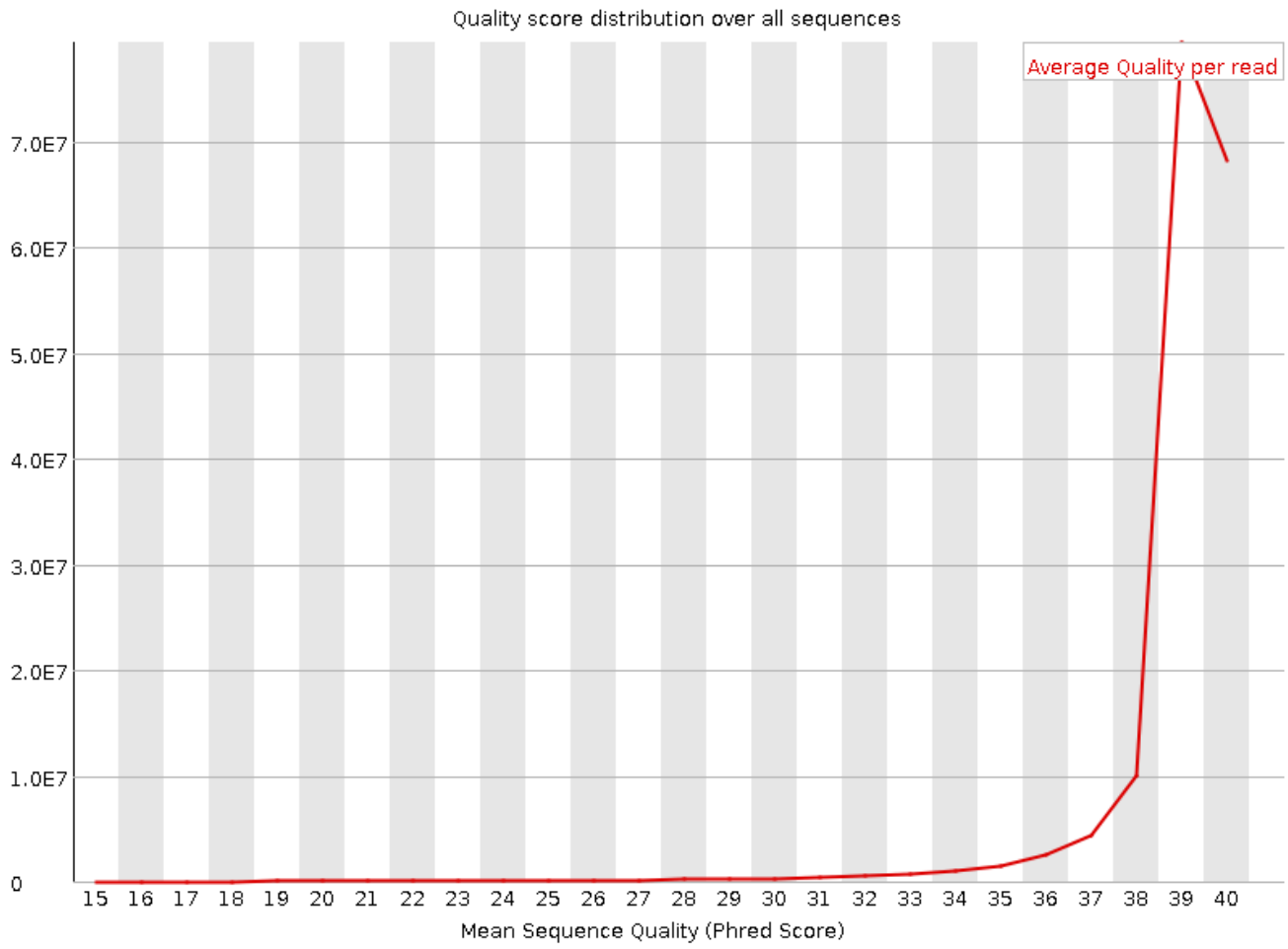


图4 Sequence base quality分布图

碱基质量分布图，横坐标表示reads平均质量，纵坐标为对应平均质量值的reads数目。

结果目录：**1_RawData\fastQC**

2.4 高质量数据获取

测序数据包含一些带接头、低质量的 reads，这些序列会对后续的信息分析造成很大的干扰，为了保证后续信息分析质量，需要对下机数据进行进一步过滤，即将原始下机数据（rawdata）过滤生成高质量序列（high quality data）。使用 fastp (v0.20.0) 采用滑动窗口法对原始数据进行过滤，标准主要包括以下几点：

(1) 接头污染去除，去除 3’ 端的接头污染；

(2) 质量过滤，采用滑动窗口法进行质量过滤，窗口大小设置为 5 bp，步长设置为 1 bp。每一次往前移动一个碱基，取 5 个碱基计算窗口的平均 Q 值，若最后一个碱基的 Q 值 ≤ 2，则仅保留该位置之前的碱基；若窗口的平均 Q 值 ≤ 20，则仅保留该窗口倒数第二个碱基及之前的碱基；

(3) 长度过滤，若双末端中任意一条 reads 的长度 ≤ 50 bp，则去除该双末端 reads。

数据过滤的基本情况见表3。

表 3 高质量数据统计

Search

Sample	HQ_Reads	HQ_Reads(%)	HQ_Data(bp)	HQ_Data(%)	Sequence_Depth
DXY	329,487,224	95.25%	49,419,579,647	94.61%	72.39
K	439,211,186	94.51%	65,899,062,070	93.90%	96.53
G	479,386,964	95.10%	71,852,998,703	94.40%	105.25
J16	351,677,908	95.10%	52,756,866,214	94.48%	77.28

共 4 条

注：

Sample： 样品名；

HQ reads： 高质量 reads数；

HQ reads %: 高质量reads占原始reads的百分比；

HQ Bases(bp)： 高质量 reads 碱基数；

HQ Bases%： 高质量 reads 碱基数占原始碱基总数的百分比；

Q20(%)： 碱基识别准确率在99%以上的碱基所占百分比；

Q30(%)： 碱基识别准确率在99.9%以上的碱基所占百分比。

结果目录：2_HQData

2.5 比对分析

2.5.1 参考基因组整理

根据项目提供参考基因组序列信息对其进行统计，具体的统计结果如表4。

表 4 参考基因组基本信息统计

Search

Type	Number
Min sequence length:	45,475,667
Max sequence length:	88,663,952
Total sequence number:	13
N20:	85,736,662
N20 Number:	2
N50:	61,540,751
N50 Number:	5
N90:	56,760,843
N90 Number:	10
N number:	127,958,425

共 14 条 10 条/页 ∨ ‹ 1 2 › 前往 1 页

注：

Total Sequence number：总序列数量；

Total sequence length (bp)： 基因组长度；

Shortest (bp)：最短序列长度；

Longest (bp)：最长序列长度；

GC content：GC含量；

N number：不确定碱基数量；

N rate：不确定碱基数量占拼接序列长度的比例；

Sequences greater than 1kb：序列长度大于1 kb以上的序列数量；

N20 (bp)：将所有序列按照长度从长到短排列，将序列长度按照该顺序依次相加，当相加的长度达到序列总长度的 20% 时，最后一条序列的长度；

N50 (bp)：将所有序列按照长度从长到短排列，将序列长度按照该顺序依次相加，当相加的长度达到序列总长度的 50% 时，最后一条序列的长度；

N90 (bp)：将所有序列按照长度从长到短排列，将序列长度按照该顺序依次相加，当相加的长度达到序列总长度的 90% 时，最后一条序列的长度

结果目录：0_Refenrence

2.5.2 序列比对

采用bwa（0.7.12-r1039）(Li,H.,and R. Durbin等，2009) mem程序将经过过滤后得到的高质量数据比对到参考基因组上，比对的参数均按照bwamem的默认参数。采用picard 1.107软件 http://www.psc.edu/index.php/user-resources/software/picard对sam文件进行排序并转换为bam文件，采用 “FixMateInformation” 命令确保所有的Paired-end reads信息之间的一致性。测序数据的生成过程涉及到文库的扩增和簇的形成，这两个步骤容易产生一些 Duplicates（即PCR Duplicates和Optical Duplicates），这些Duplicates 不能作为变异检测的证据。采用Picard软件包中的 “MarkDuplicates” 去除Duplicates，即如果多个Paired Reads比对后具有相同的染色体坐标，则仅保留分值最高的Paired Reads。InDel附近的 reads 最容易出现 Mapping 错误，为了尽量减少由于 Mapping 错误导致的 SNP，需要对InDel附近的 reads 重新进行比对，以提高 SNP Calling 的准确性，采用GATK程序中的IndelRealigner命令对所有InDel附近的 reads 进行重新比对，以提高 SNP 预测的准确率。

2.5.3 序列比对结果概述

序列比对结果统计见表5。

表 5 序列比对结果统计

Search

Sample	Total_reads	Mapped_reads	Mapping_rate
J16	362,825,287	360,078,849	99.24%
DXY	340,817,720	338,311,166	99.26%
K	453,782,651	444,458,542	97.95%
G	495,066,943	485,969,074	98.16%

共 4 条

注：

Sample： 样本名；

Mapped reads： 比对至参考基因组上的reads数量（包括单端比对和双端比对）；

Mapped bases: 比对至参考基因组上的总碱基数量（包括单端比对和双端比对）；

Mapping rate： 比对率，比对至参考基因组上的reads数量占总 reads 数量的百分比；

结果目录：3_Mapping\1_Bam

2.5.4 测序深度及覆盖度概述

分别统计每一个样品的平均覆盖深度和覆盖度，结果见表6。

表 6 比对测序深度和覆盖度结果统计

Search

Sample	Coverage ≥1(%)	Coverage ≥4(%)	Coverage ≥10(%)	Coverage ≥20(%)
DXY	96.94%	94.93%	91.81%	84.66%
J16	96.95%	95.08%	92.23%	85.82%
G	98.17%	96.81%	94.62%	90.46%
K	98.20%	96.72%	94.27%	89.23%

共 4 条

注：

Sample_ID： 样品编号；

Avg. Depth (×)：平均测序深度，比对至参考基因组的碱基总数除以基因组大小；

Coverage ≥ 1(%)：参考基因组中至少被 1 条序列覆盖的位点占基因的百分比；

Coverage ≥ 4 (%)：参考基因组中至少被 4 条序列覆盖的位点占基因的百分比；

Coverage ≥ 10 (%)：参考基因组中至少被10条序列覆盖的位点占基因的百分比。

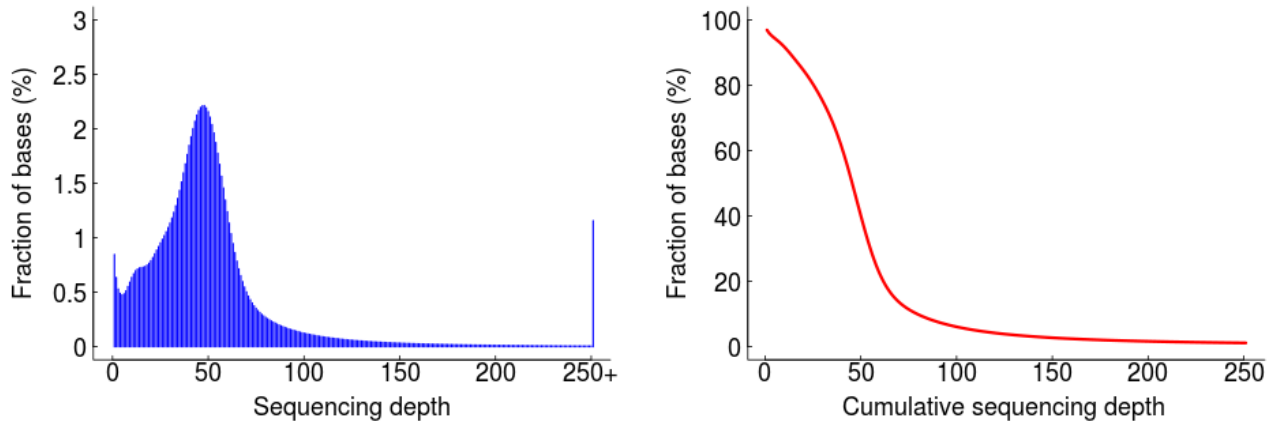


图5 测序深度及覆盖图概述

左图为样品的测序深度分布图，横坐标代表测序深度，纵坐标代表在该测序深度下的碱基占所有碱基的比例。右图为在不同测序深度的累积碱基比例，横坐标代表测序深度，纵坐标代表超过该测序深度的碱基占所有碱基的比例。如测序深度为 1，纵坐标为 95，则代表 95% 的碱基其测序深度大于 1 ×

结果目录：3_Mapping\2_Stat

2.6 SNP分析

2.6.1 SNP检测与分布

SNP(单核苷酸多态性)主要是指在基因组水平上由单个核苷酸的变异所引起的DNA序列多态性，包括单个碱基的转换、颠换等。采用GATK软件（Zhu P Fau He等，2015）来进行SNP的检测，具体操作步骤如下：

(1) InDel附近的alignments 通常不可靠，需要利用已知的InDel信息进行 realignment。分两步：第一步利用GATK 软件包中的RealignerTargetCreator命令，输出一个包含所有可能的InDels的文件；第二步利用IndelRealigner命令对所有InDels附近的reads 进行重新比对，以提高SNP预测的准确率。

(2) 采用UnifiedGenotyper程序获得样品的SNP位点，stand_call_conf设置为30；

为了保证 SNP 位点的可靠性，进一步对获得的SNP位点进行过滤，过滤标准如下：

(1) Fisher test of strand bias (FS) ≤ 60；

(2) HaplotypeScore≤ 13.0；

(3) Mapping Quality (MQ)≥ 40；

(4) Quality Depth(QD) ≥ 2；

(5) ReadPosRankSum≥ -8.0；

(6) MQRankSum> -12.5；

(7) alternative allele called in ≥ 4 reads。

(8) 去除存在缺失的位点。

为了定位目标性状，每组性状相关样本一起call 群体SNP，经过上面条件的过滤对最终得到的群体SNP在每个样本中的数目做统计见表7，具体信息见Total.final_snps.vcf。以滑窗统计SNP的数目，每条染色体上SNP的分布情况见下图6。

表 7 群体SNP在每个样本中的结果统计

Search

Sample	HOM_REF	HET	HOM_ALT
DXY	3,426,988	10,289,194	2,463,925
G	1,452,387	13,265,906	1,461,814
J16	4,505,772	9,324,340	2,349,995
K	1,396,396	13,372,573	1,411,138

共 4 条

注：

Sample：样品名称；

HOM_REF：与参考基因组一致的纯合基因型数目；

HET： 杂合基因型数目；

HOM_ALT：与参考基因组不一致的纯合基因型数量。

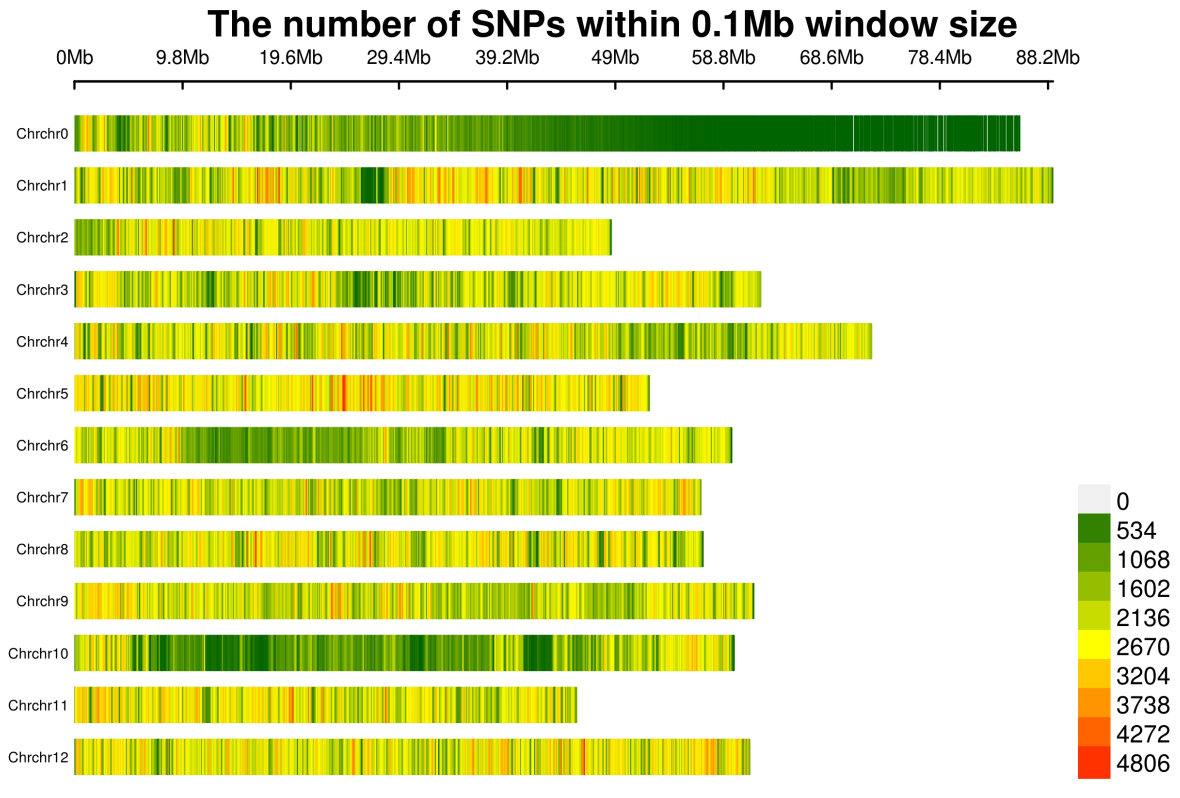


图6 SNP在染色体上分布图

结果目录：4_SNP\1_Detection

2.6.2 SNP注释

采用 ANNOVAR（Wang K, Li M等，2010）软件对SNP 位点进行注释。SNP注释结果见表8。

表 8 SNP注释结果统计

Search

Type	Number	Percentage
exonic total	754,242	4.66
synonymous SNV	355,575	2.2
nonsynonymous SNV	387,450	2.39
stopgain	9,314	0.06
stoploss	1,903	0.01
unknown	0	0.0
splicing	2,451	0.02
ncRNA total	0	0.0
ncRNA_exonic	0	0.0
ncRNA_splicing	0	0.0

共 21 条 10 条/页 ∨ ‹ 1 2 3 › 前往 1 页

注：

Exonic： 外显子区域；

Stopgain： 导致获得终止密码子的SNP；

Stoploss： 导致丢失终止密码子的SNP；

Synonymous： 同义突变；

Nonsynonymous： 非同义突变；

Unknown： 未知功能的位点；

Splicing： splicing junction 2 bp 的区域；

Intronic： 内含子区域；

Intergenic： 基因间区域；

Upstream： 转录起始位点上游 1 kb 的区域；

Downstream： 转录终止位点下游 1 kb 的区域；

Upstream;Downstream： 转录起始位点上游和下游的区域。

结果目录：4_SNP\2_Annotation

2.6.3 变异检测结果可视化

对于比对的bam格式文件，可以使用IGV软件对变异检测结果进行可视化浏览。使用方法参考使用说明文档（IGV使用文档.pdf）。

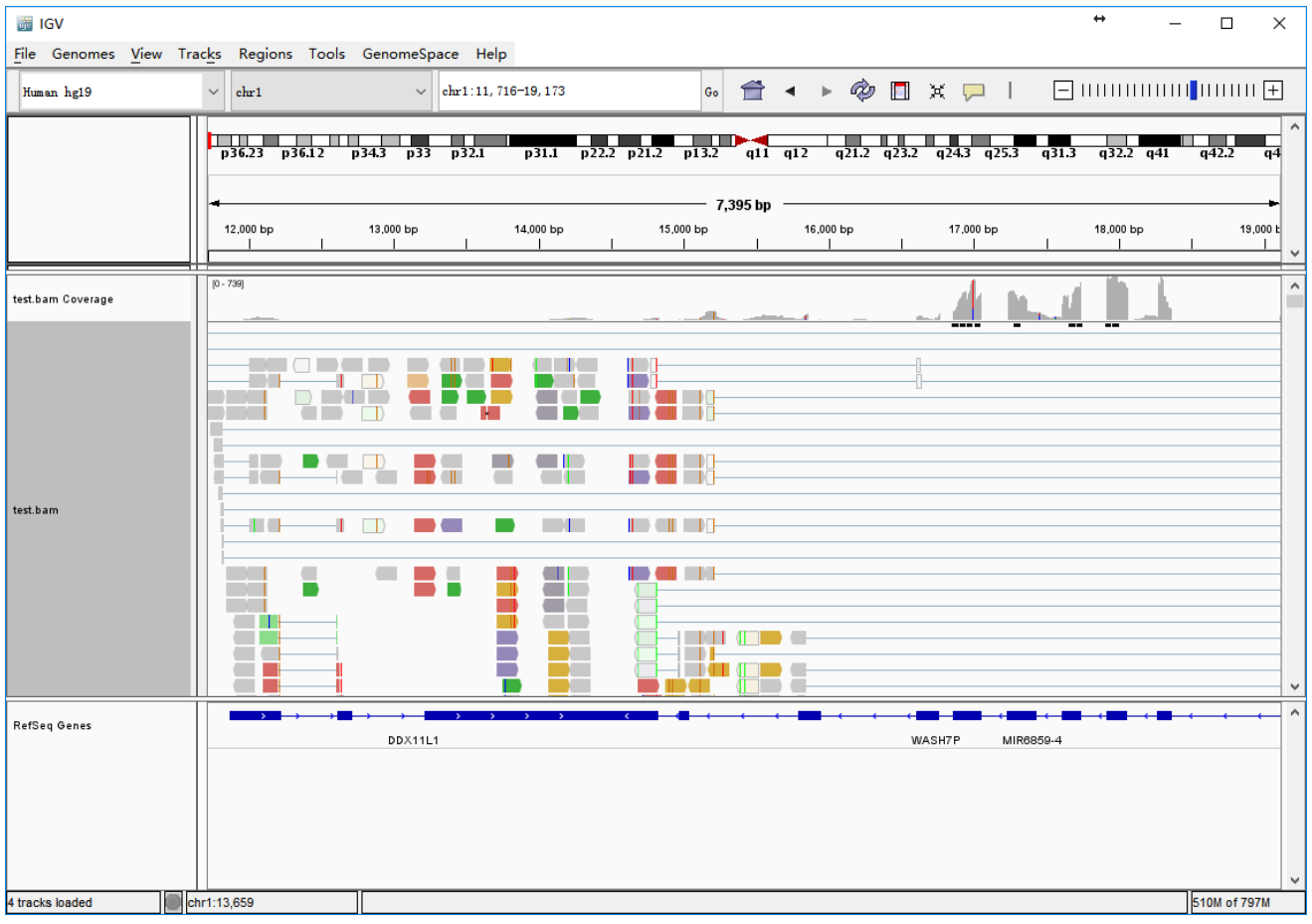


图7 IGV图形界面示意图

结果目录：4_SNP\3_IGV

2.7 InDel分析

2.7.1 InDel检测与长度分布

InDel是小型的Insertion和Deletion的总称。我们采用GenomeAnalysisTK v3.8软件包中的UnifiedGenotyper程序获得样品所有突变位点，stand_call_conf设置为30，再利用SelectVariants命令来提取 InDel 突变信息。

为了保证InDel 结果的可靠性，进一步对 InDel 位点进行过滤，过滤标准如下：

(1) Fisher Test of Strand Bias(FS)≤200；

(2) Read depth（DP）>4;

(3) Quality Depth (QD)≥2；

(4) ReadPosRankSum ≥ -20；

(5) 去除有缺失的位点。

为了定位目标性状，每组性状相关样本一起call 群体InDel，经过上面条件的过滤对最终得到的群体SNP在每个样本中的数目做统计见表9，InDel的长度的分布情况见下图8。具体基因型信息见Total.final_indels.vcf。

表 9 群体InDel在每个样本中的结果统计

Search

Sample	HOM_REF	HET	HOM_ALT
DXY	309,461	1,581,863	537,019
G	63,622	1,984,874	379,847
J16	418,691	1,489,372	520,280
K	62,273	1,999,052	367,018

共 4 条

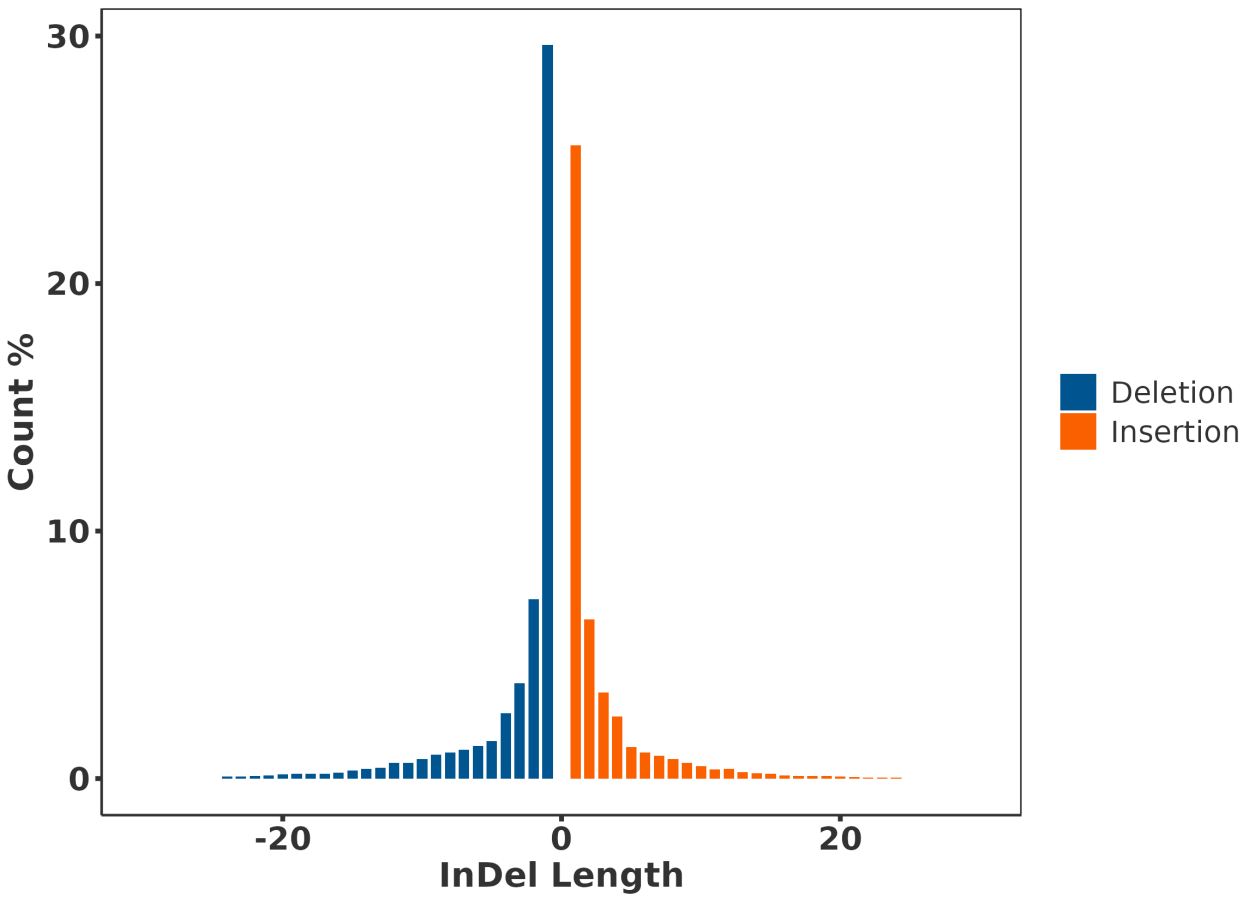


图8 InDel长度分布图

结果目录：5_InDel\1_Detection

2.7.2 InDel注释

采用 ANNOVAR（Wang K, Li M等，2010）软件对InDel位点进行注释。InDel的注释结果见表10。

表 10 InDel注释结果统计

Search

Type	Number	Percentage
exonic total	30,867	1.27
frameshift deletion	9,890	0.41
frameshift insertion	6,783	0.28
nonframeshift deletion	7,354	0.3
nonframeshift insertion	5,940	0.24
nonframeshift substitution	0	0.0
stopgain	724	0.03
stoploss	176	0.01
unknown	0	0.0
splicing	834	0.03

共 25 条 10 条/页 ∨ ‹ 1 2 3 › 前往 1 页

注：

Exonic： 外显子区域；

Stopgain： 导致获得终止密码子的InDel；

Stoploss： 导致丢失终止密码子的InDel；

Frameshift deletion： 移码缺失导致蛋白编码基因的读框改变，长度为3的非整数倍；

Frameshift insertion： 移码插入导致蛋白编码基因的读框改变，长度为 3 的非整数倍；

Nonframeshift deletion： 非移码缺失，不改变蛋白编码基因的读框，长度为 3 的整数倍；

Nonframeshift insertion： 非移码插入，不改变蛋白编码基因的读框，长度为 3 的整数倍；

Splicing： splicing junction 2 bp 的区域；

Intronic：内含子区域；

Intergenic： 基因间区域；

UTR 5’：5’ UTR 区域；

UTR 3’：3’ UTR 区域；

Upstream： 转录起始位点上游 1 kb 的区域；

Downstream： 转录终止位点下游 1 kb 的区域；

Upstream;Downstream： 转录起始位点上游和下游的区域。

结果目录：5_InDel\2_Annotation

2.8 SNP-index 分析

2.8.1 SNP-index 计算

SNP-index 的概念最早由日本学者Takag 等人于 2013 年提出，主要用于筛选与性状基因相关联的候选区域。SNP-index计算原理见图9，基本过程描述

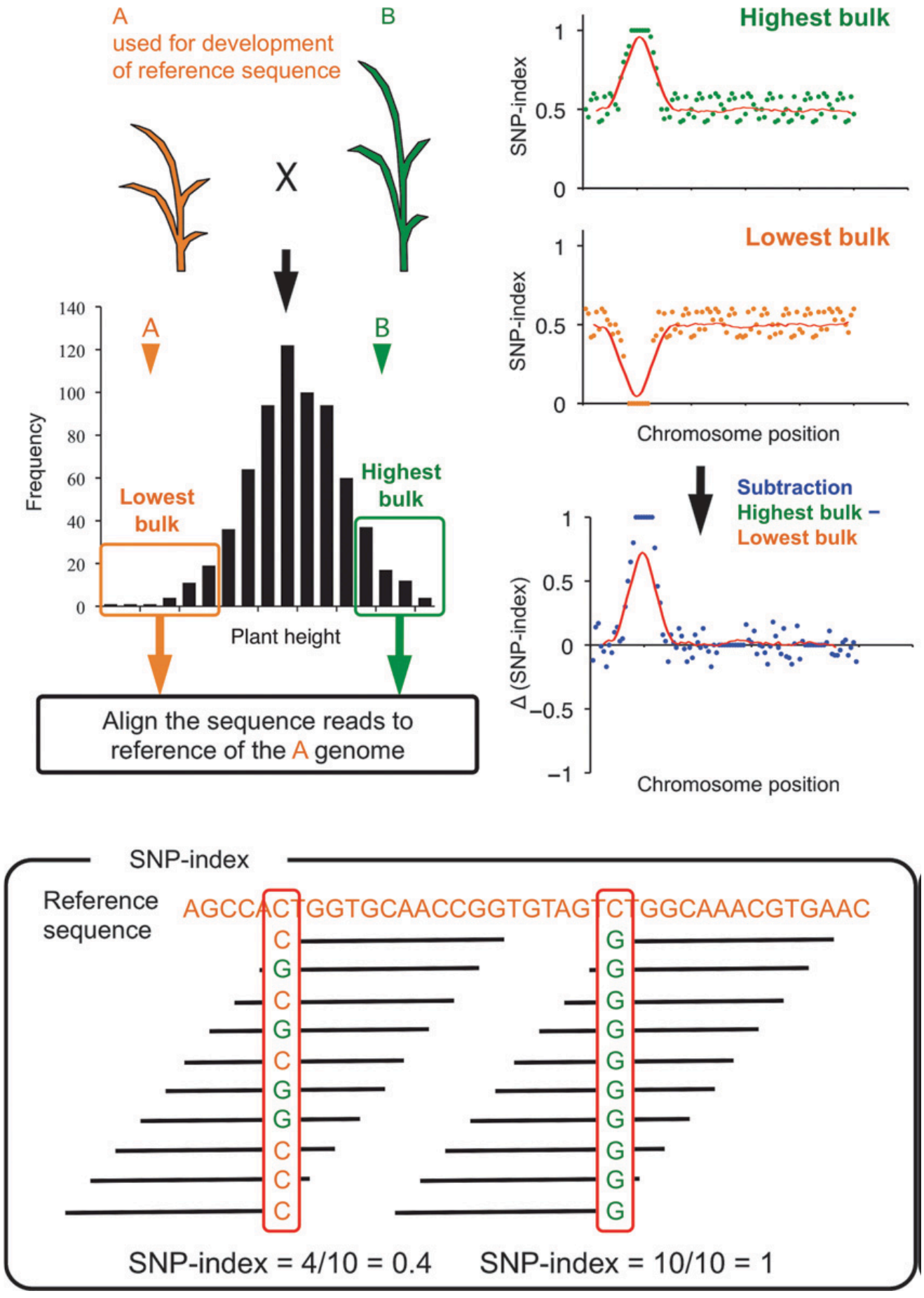


图9 SNP-index原理图

(1) 群体构建：具有显著表型差异的亲本 A 和 B进行杂交，产生 F1 代，F1 代个体进行自交，产生 F2 代个体；

(2) 测序：分别对亲本 A、亲本 B 进行测序；取 F2 代个体具有极端表型的个体分别抽提总 DNA，等量混合，分别对这两个子代池进行测序；

(3) 基因型分析：将亲本 A、亲本 B、子代混池1和子代混池2比对至参考基因组上，筛选 SNP 多态性标记；

(4) SNP-index 计算：假设子代混池1在某一个位点上测序深度为 50，其中，与亲本 A 完全相同的碱基测序深度为 30，与亲本 B 完全相同的碱基测序深度为 20，则子代池1的 SNP-index 为 0.6。分别计算每一个多态性位点的 SNP-index。

针对这个项目，每组性状相关样本一起call 群体SNP，在进行SNP-index计算之前需要对群体SNP做进一步的过滤，过滤步骤如下见表11：

表 11 过滤标准

Search

序号	过滤条件
1	保留亲本1杂合、亲本2纯合的SNP位点
2	去除亲本深度低于8x的点
3	去除子代混池深度低于8x的点
4	筛选出任一混池SNP-index值大于0.3的位点

共 4 条

按照上面表中列举出的标准过滤最终得到SNP-index结果见SNP.index.xls。

2.8.2 SNP-index 分布

为直观反映子代SNP-index在染色体上的分布情况，对SNP-index在染色体上的分布进行作图。图中的红色折线即为以窗口形式展现的SNP-index分布，滑窗分析每个窗口中的SNP-index。计算两个子代混池的 SNP-index 的差值，即得到 △(SNP-index) = SNP-index(极端性状 A) – SNP-index（极端性状 B）。进行10000次置换检验，选取95%（绿色）、99%（橙色）置信水平作为筛选的阈值，图中的红色折线即为以窗口形式展现的△(SNP-index)分布，置信水平99%以上的窗口作为候选区间，两个子代混池和△(SNP-index) 在全基因组的分布结果见图10。两个子代混池和△(SNP-index) 在全基因组的分布圈图的形式见图11。

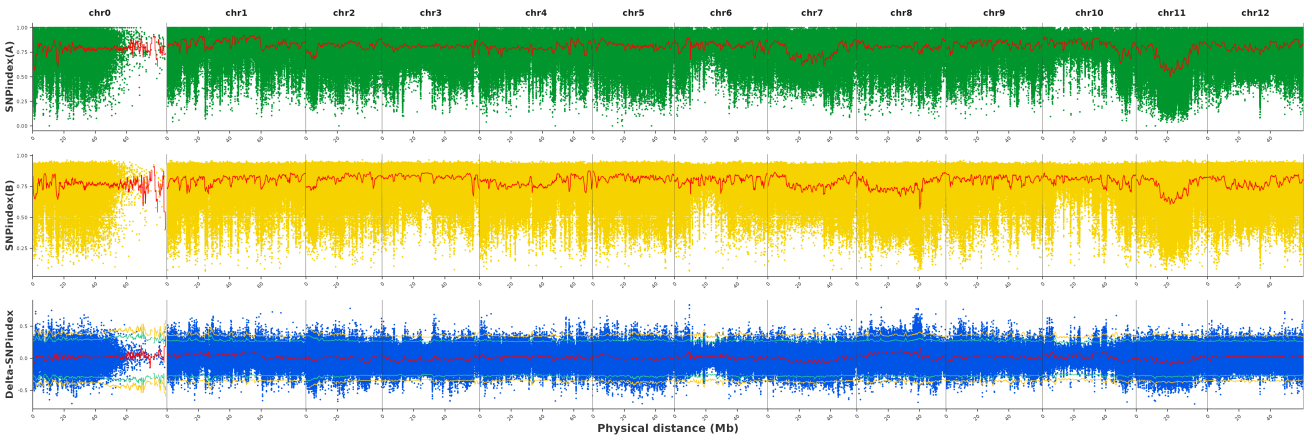


图10 SNP-index 分布图

横坐标为各条染色体的名称及其长度，纵坐标代表SNP-index值，点即为SNP-index，红线为窗口下SNP-index的均值，橙色线为99%置信线，绿色线为95%置信线。

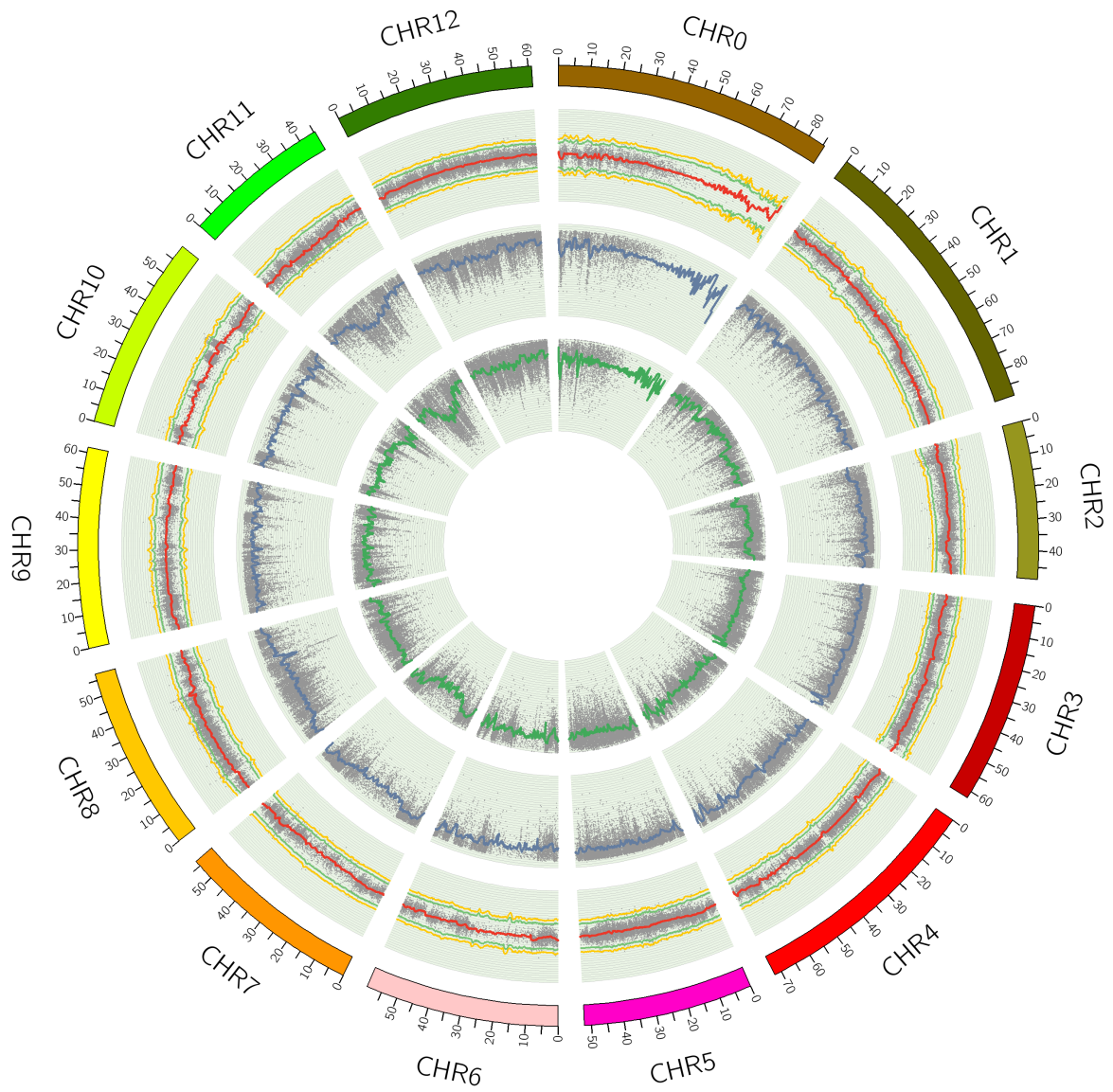


图11 SNP-index 分布圈图

圈图从外到内依次为：I：染色体； II：delta_index分布（绿色：95%置信线，黄色：99%置信线）； III：蓝色线为野生混池的SNP-index分布；IV：绿色线为突变混池的SNP-index分布。

结果目录：6_SNP-index

2.8.3 候选区域

在进行测序前，由于对所有混池样本的性状进行选择，选取极端性状的混池进行测序，即SNP-index 值越大，则该位点的SNP 与目标性状的连锁强度越高，与目标性状的关联度越高；反之，则该位点的SNP 与目标性状的连锁强度越低，与目标性状的关联度越低。

选取超过99百分位数阈值线的区域作为性状相关的候选区域，将这个范围内满足上面筛选条件的SNP挑选出来得到Target.snp.vcf，对这些SNP位点做注释，最终的结果文件见 ***candidate_Region.txt***。关联区域的结果统计见表12。

表12 关联区域统计

Search

Chr	Start	End
-	-	-

共 1 条

注：

Chr： 染色体编号；

Start： 起始位置；

End： 终止位置；

结果目录：6_SNP-index/Target/

2.8.4 候选区域突变基因GO注释

GO是Gene Ontology的缩写。GO的产生主要是为了解决同一基因在不同数据库中定义的混乱性以及不同物种的同一基因在功能定义上的混乱性。它是一个国际标准化的基因功能分类体系，提供了一套动态更新的标准词汇表（controlled vocabulary）来全面描述生物体中基因和基因产物的属性。GO总共有三个ontology，分别描述基因的分子功能（molecular function）、所处的细胞位置（cellular component）、参与的生物学过程（biological process）。每个筛选出的基因对应的GOid及对应功能详见GO/Target.go.annotation.xls。将这些GOid在level2级别进行了汇总并统计，统计结果详见GO/Target_GO_level2.xls。统计结果可视化见图12。

图12. GO注释统计图

结果目录：6_SNP-index/Target/GO

2.8.5 候选区域突变基因的 KEGG功能注释

KEGG（Kyoto Encyclopedia of Genes and Genomes），又称京都基因与基因组百科全书，是基因组破译方面的数据库。KEGG注释的主要目的包括两个：KO（KEGG Ortholog）注释，即将分子网络的相关信息进行跨物种注释；KEGG Pathway注释，即代谢通路注释，获得物种内分子间相互作用和反应的网络。

候选区域非同义突变基因的KO及Pathway注释主要采用KEGG 的KAAS自动化注释系统完成（Moriya Y等，2007），其中基因集选择近源物种，基因KO的判别规则选取 bi-directional best hit（BBH），KO注释的结果见“target.kegg.xls”。KO注释完成后将KO映射到相应的KEGG Pathway通路上， KEGG Pathway统计结果见图13。

图13 KEGG注释统计图

结果目录：6_SNP-index-normal/Target/KEGG

2.9 ED值分析

2.9.1 ED值计算与分布

ED 的概念由美国学者 Jonathon 等人于 2013 年提出，主要用于无亲本测序数据的与性状基因相关联的候选区域的筛选。基于BSA的 ED值计算公式如下：

$$ED = \sqrt{(A_{mut} - A_{wt})^2 + (C_{mut} - C_{wt})^2 + (G_{mut} - G_{wt})^2 + (T_{mut} - T_{wt})^2}$$

ED值计算公式

(ATCG)mut：混池1碱基的频率； (ATCG)wt：混池2碱基的频率

针对这个项目，每组性状相关样本一起call 群体SNP，在进行ED计算之前需要对群体SNP做进一步的过滤，过滤步骤如下见表11：

表 11 过滤标准

Search

序号	过滤条件
1	保留亲本1杂合、亲本2纯合的SNP位点
2	去除亲本深度低于8x的点
3	去除子代混池深度低于8x的点
4	筛选出任一混池SNP-index值大于0.3的位点

共 4 条

为降低差异较小的ED值带来的背景噪声，直观反映子代ED值在染色体上的分布情况，计算ED4（ED值的4次方），对SNP和Indel的ED4值在染色体上的分布分别进行作图。以参考基因组的位置为横坐标，ED4作为纵坐标作图，结果见图10。

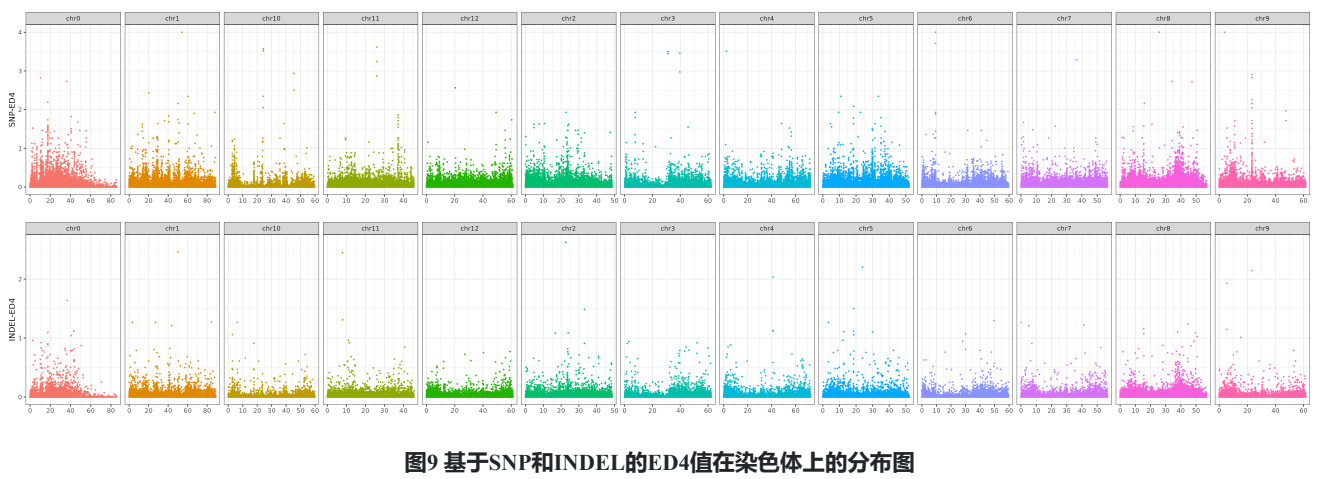


图9 基于SNP和INDEL的ED4值在染色体上的分布图

横坐标为各条染色体的名称及其长度，纵坐标代表ED4值。

结果目录：7_ED

2.9.2 候选区域

ED值反映了两个混池之间基因型频率的差异，在进行测序前，由于对所有混池样本的性状进行选择，选取两个极端性状的混池分别进行测序，ED值越大，则该位点的SNP、Indel 与目标性状的连锁强度越高，与目标性状的关联度越高；反之，则该位点与目标性状的连锁强度越低，与目标性状的关联度越低。为了降低差异较小的ED值带来的背景噪声，计算ED4（ED值的4次方）并对ED4进行LoessFit拟合。选取中位数+3*标准差为阈值，高于阈值的区域作为候选区间。对这些候选位点进行注释。

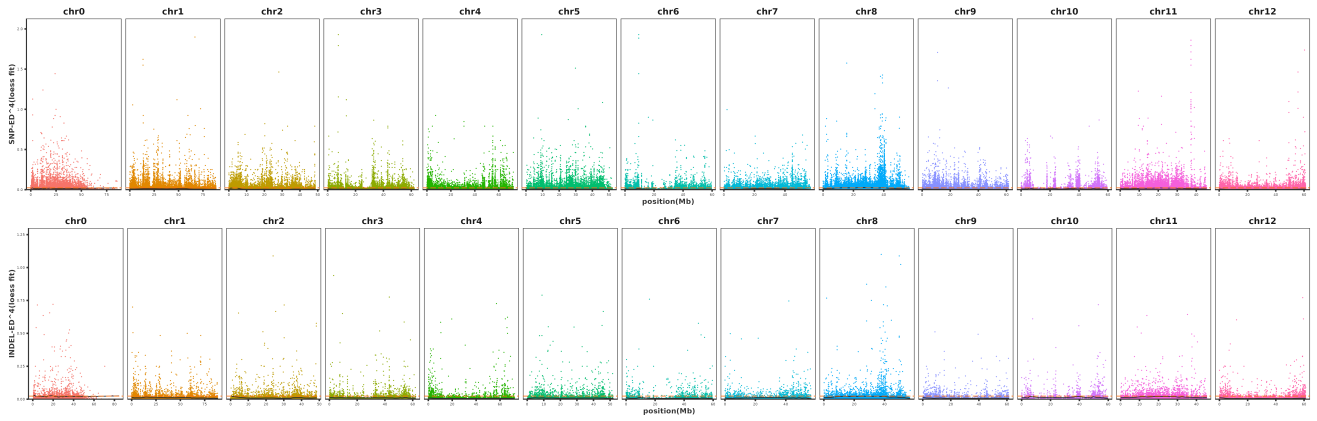


图10 基于SNP和INDEL的ED值在染色体上的分布图

横坐标为各条染色体的名称及其长度，纵坐标代表ED值。

表12 SNP关联区域统计

Search

Chr	Start	End
chr8	12,240,379	16,269,006
chr8	17,007,776	17,912,489
chr8	21,863,944	24,448,303
chr8	25,073,302	29,201,948
chr8	30,064,055	33,375,810
chr8	38,050,531	40,778,959

共 6 条

表13 INDEL关联区域统计

Search

Chr	Start	End
chr0	527,755	2,583,363
chr0	15,516,337	20,434,948
chr0	74,695,894	84,553,878
chr8	35,346,915	36,056,183

共 4 条

注：

Chr： 染色体编号；

Start： 起始位置；

End： 终止位置；

结果目录：7_ED/Target/

2.9.3 候选区域突变基因GO注释

GO是Gene Ontology的缩写。GO的产生主要是为了解决同一基因在不同数据库中定义的混乱性以及不同物种的同一基因在功能定义上的混乱性。它是一个国际标准化的基因功能分类体系，提供了一套动态更新的标准词汇表（controlled vocabulary）来全面描述生物体中基因和基因产物的属性。GO总共有三个ontology，分别描述基因的分子功能（molecular function）、所处的细胞位置（cellular component）、参与的生物学过程（biological process）。每个筛选出的基因对应的GOid及对应功能详见GO/Target.go.annotation.xls。将这些GOid在level2级别进行了汇总并统计，统计结果详见GO/Target_GO_level2.xls。统计结果可视化见图19。

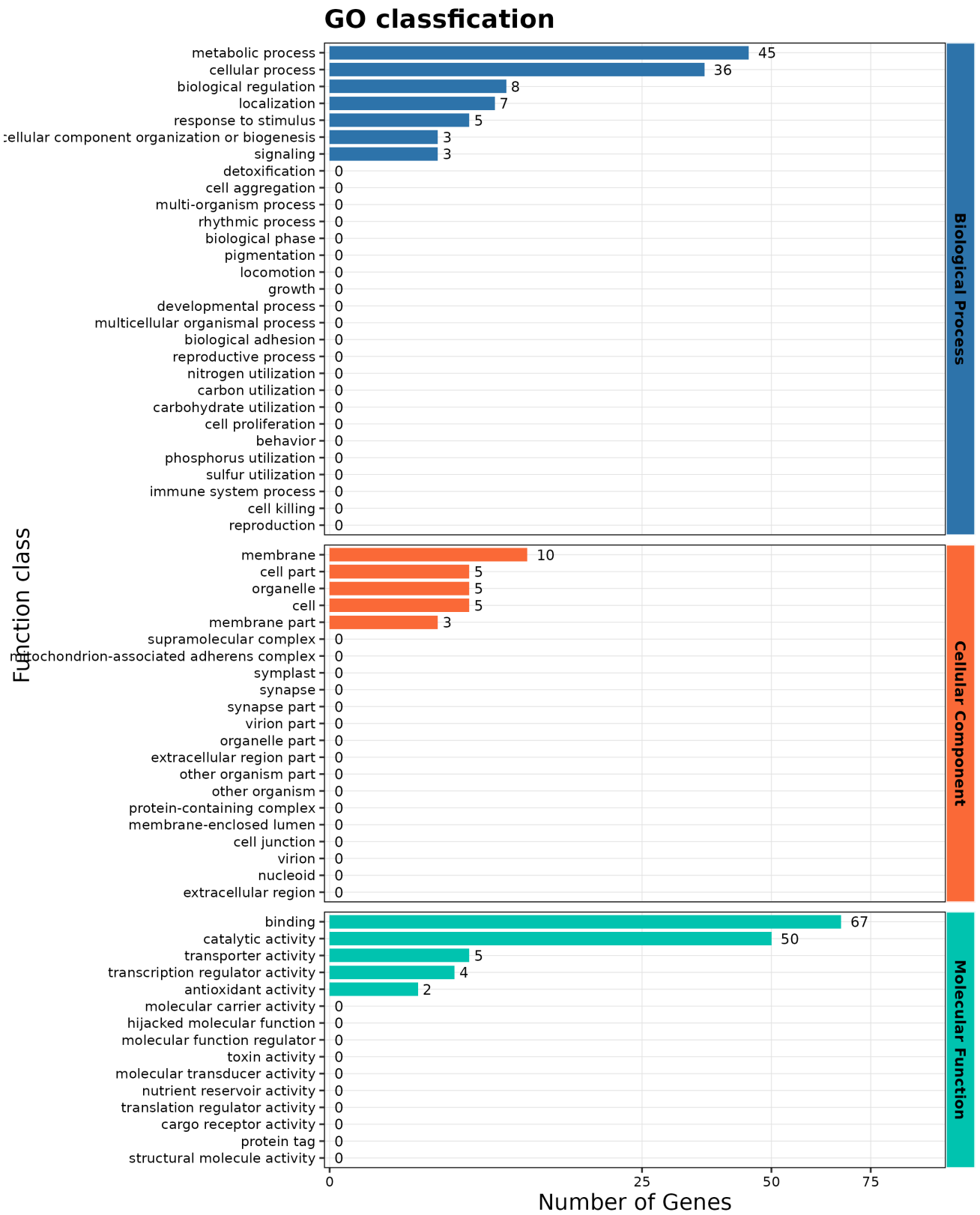


图19. GO注释统计图

结果目录：7_ED/Target/GO

2.9.4 候选区域突变基因的 KEGG功能注释

KEGG（Kyoto Encyclopedia of Genes and Genomes），又称京都基因与基因组百科全书，是基因组破译方面的数据库。KEGG注释的主要目的包括两个：KO（KEGG Ortholog）注释，即将分子网络的相关信息进行跨物种注释；KEGG Pathway注释，即代谢通路注释，获得物种内分子间相互作用和反应的网络。

候选区域非同义突变基因的KO及Pathway注释主要采用KEGG 的KAAS自动化注释系统完成（Moriya Y等，2007），其中基因集选择近源物种，基因KO的判别规则选取 bi-directional best hit（BBH），KO注释的结果见“target.kegg.xls”。KO注释完成后将KO映射到相应的KEGG Pathway通路上， KEGG Pathway统计结果见图20。

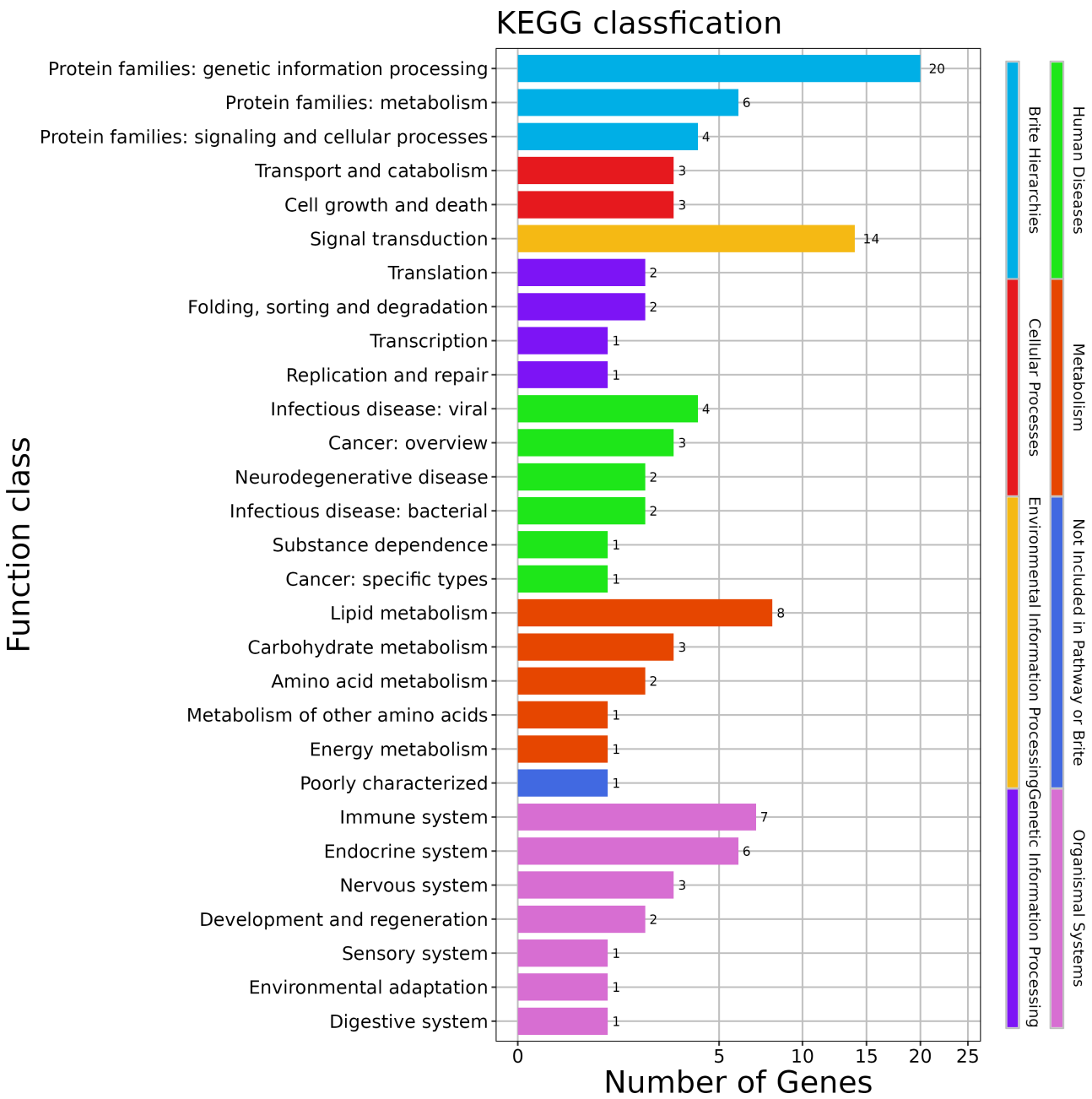


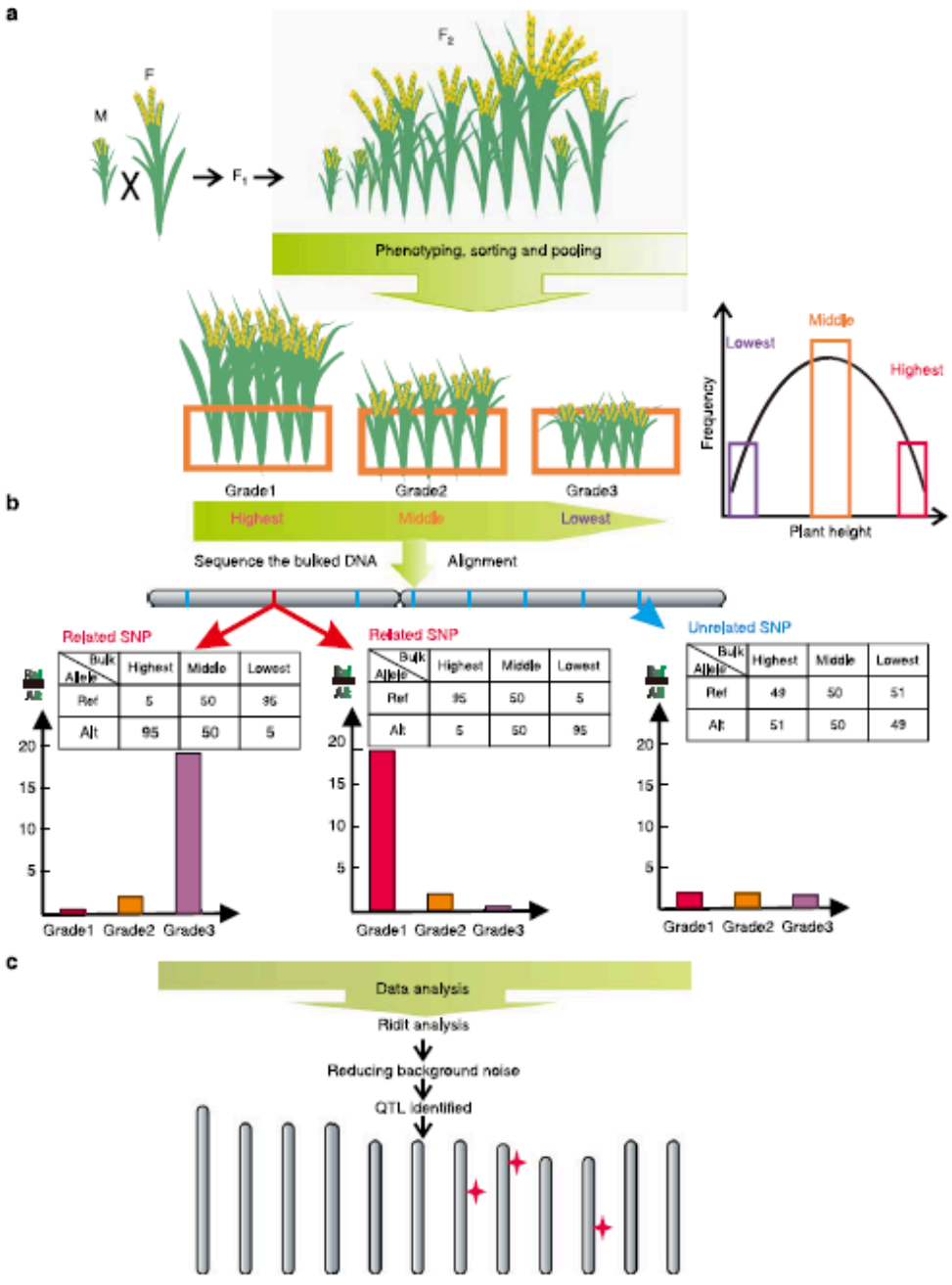
图20 KEGG注释统计图

结果目录：7_ED/Target/KEGG

2.10 GPS分析

2.10.1 GradePool-Seq原理

GradePool-Seq是韩斌研究组开发的一套新的数量性状基因定位方法，该研究提出了一个数量性状基因座（QTL）作图方法，即GradedPool-Seq（GPS），用于快速绘制QTLs图谱，是通过对F2后代的梯度池样本进行全基因组测序，然后进行BSA。从理论上讲，如果一个变异与表型相关的基因密切相关，其p值将相对较小。然而，由于远缘遗传杂交在SNP/InDel信号中存在大量的背景噪声，必须考虑背景噪声的降低。研究中实现的降噪算法是一种非重叠滑动窗口方法，计算在一个确定的基因组区间(~ 400kb)内，统计上显著的超过设置阈值的变异数与总变异数之比。比值最大的区间，即p值显著的SNPs/Indels群所占比例最高。



GPS原理示意图

SNP/InDel采用Radit检验分析在两个混池之间的差异。为了去除假阳性，减少背景噪音，随后统计固定窗口（400kb）内显著SNP/InDel的比值。比值高的地方，即为性状关联区间。

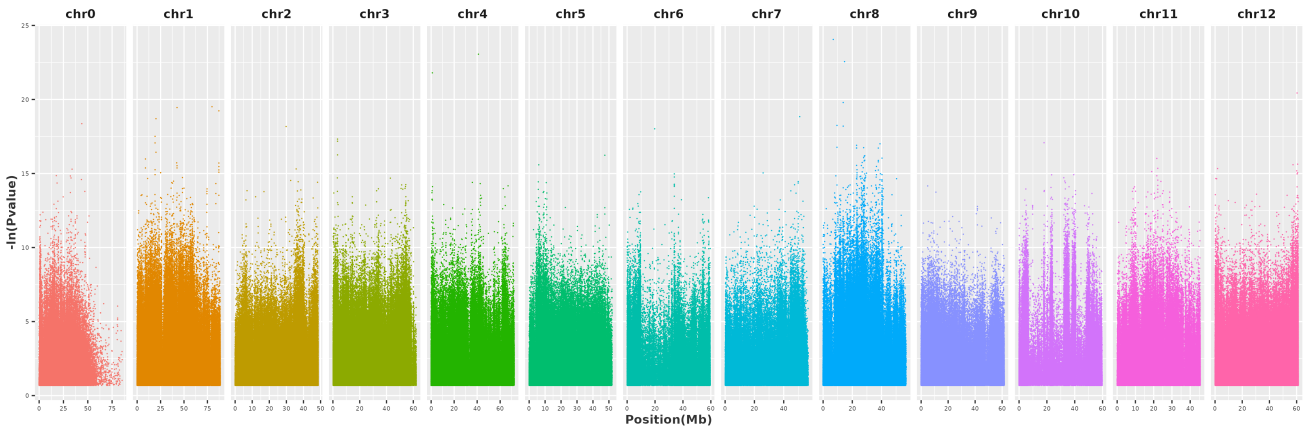
显著SNP/InDel过滤条件：

(1)保留深度在5%到95%的位点；

(2)过滤掉子代中变异与参考基因一样的位点；

2.10.2 GPS鉴定结果

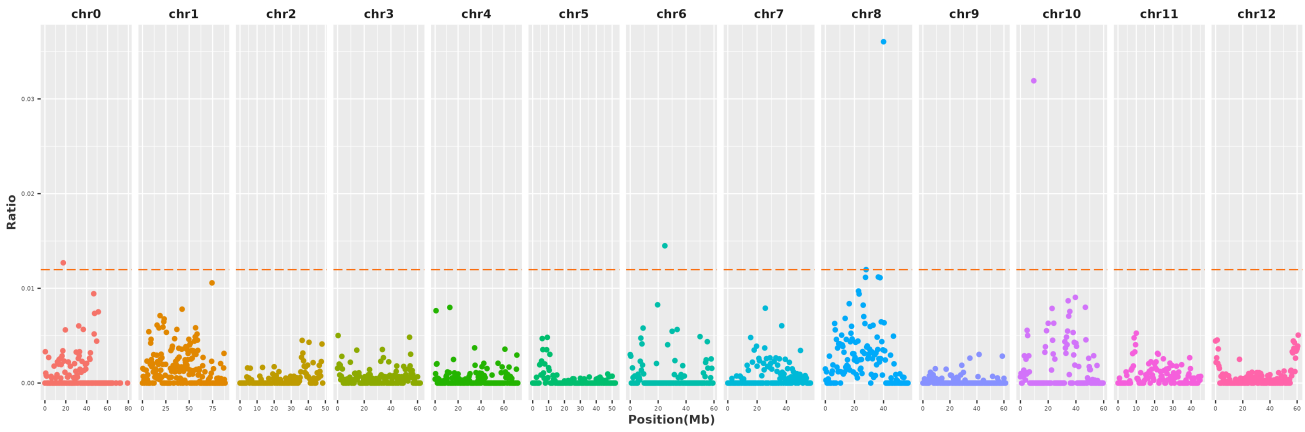
基于GPS方法的程序，对每个变量进行统计检验。Ridit分析，充分检查和分析有序数据，计算其p值，其在染色体上的分布见下图



子代混池P值在染色体上的分布

注:p值图是在降噪算法之前对性状的Ridit分析的结果。纵坐标是染色体上的SNP位置，纵坐标是对应的-ln (pvalue) 值

在采用了降低背景噪声的策略后，结果以比值图的形式呈现，比值高的地方，即为性状关联区间。



子代混池Ridit在染色体上的分布

注：x轴值设置在每个基因组间隔（固定窗口（400kb））的中点，y轴值对应于比值。

结果目录：8_GPS

2.10.3 候选区域

选取比值图中最高的5个窗口作为候选区域，最终的结果文件见***candidate_Region_gps.txt***。关联区域的结果统计见表13。

子代混池Ratio在染色体上的分布

表13 关联区域统计

Search

Chr	Start	End
chr10	9,600,000	9,999,999
chr6	24,800,000	25,199,999
chr0	17,600,000	17,999,999
chr8	28,000,000	28,399,999

共 4 条

注:

Chr： 染色体编号；

Start： 起始位置；

End： 终止位置；

结果目录：8_GPS/Target/

2.10.4 候选区域突变基因GO注释

GO是Gene Ontology的缩写。GO的产生主要是为了解决同一基因在不同数据库中定义的混乱性以及不同物种的同一基因在功能定义上的混乱性。它是一个国际标准化的基因功能分类体系，提供了一套动态更新的标准词汇表（controlled vocabulary）来全面描述生物体中基因和基因产物的属性。GO总共有三个ontology，分别描述基因的分子功能（molecular function）、所处的细胞位置（cellular component）、参与的生物学过程（biological process）。每个筛选出的基因对应的GOid及对应功能详见GO/Target.go.annotation.xls。将这些GOid在level2级别进行了汇总并统计，统计结果详见GO/Target_GO_level2.xls。统计结果可视化见图21。

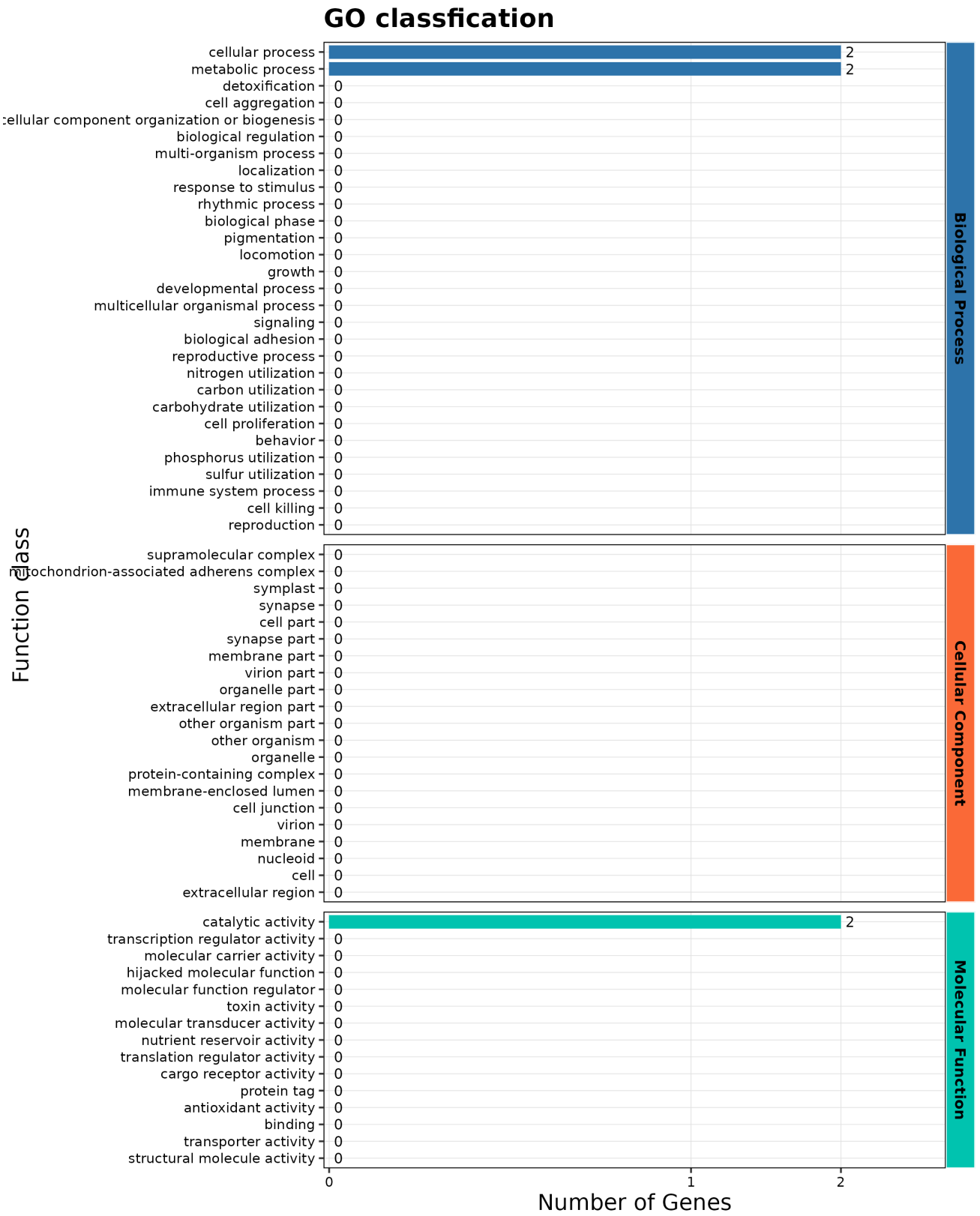


图21. GO注释统计图

结果目录：8_GPS/Target/GO

2.10.5 候选区域突变基因的 KEGG功能注释

KEGG（Kyoto Encyclopedia of Genes and Genomes），又称京都基因与基因组百科全书，是基因组破译方面的数据库。KEGG注释的主要目的包括两个：KO（KEGG Ortholog）注释，即将分子网络的相关信息进行跨物种注释；KEGG Pathway注释，即代谢通路注释，获得物种内分子间相互作用和反应的网络。候选区域非同义突变基因的KO及Pathway注释主要采用KEGG 的KAAS自动化注释系统完成（Moriya Y等，2007），其中基因集选择近源物种，基因KO的判别规则选取 bi-directional best hit（BBH），KO注释的结果见“target.kegg.xls”。KO注释完成后将KO映射到相应的KEGG Pathway通路上， KEGG Pathway统计结果见图22。

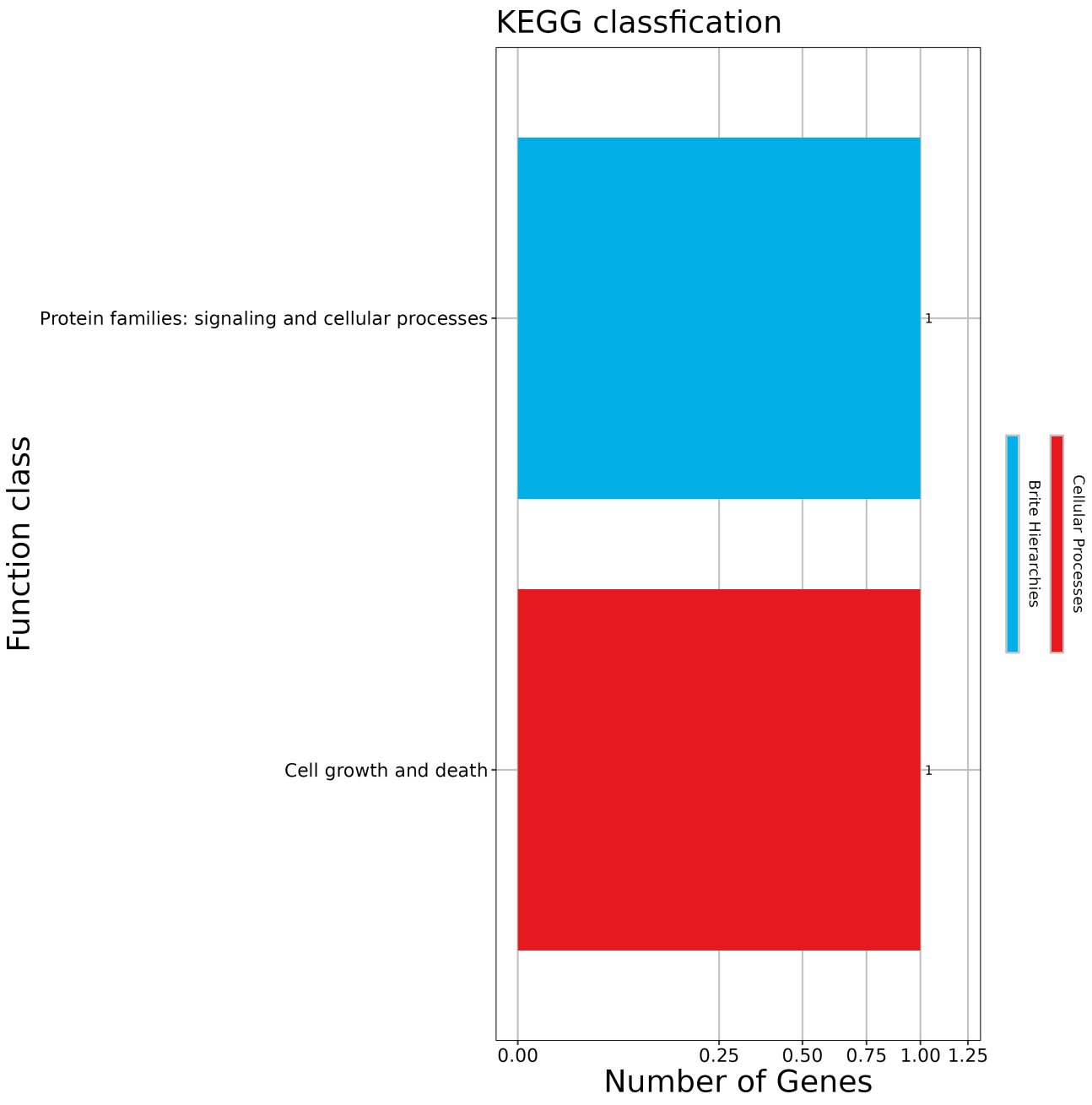


图22 KEGG注释统计图

结果目录：8_GPS/Target/KEGG

2.11 Gprime分析

2.11.1 G 统计值的方法进行定位

利用R语言中的QTLseqr包进行G统计值分析（Magwene et al. 2011），对snp结果进行筛选,过滤步骤如下见表14：

表 14 过滤标准

Search

序号	过滤条件
1	保留亲本1杂合、亲本2纯合的SNP位点
2	去除亲本深度低于8x的点
3	去除子代混池深度低于8x的点

共 3 条

然后利用QTLseqr包中的runGprimeAnalysis函数以2Mb为窗口进行G统计值分析，最后得到的结果见Gprime.tsv文件。

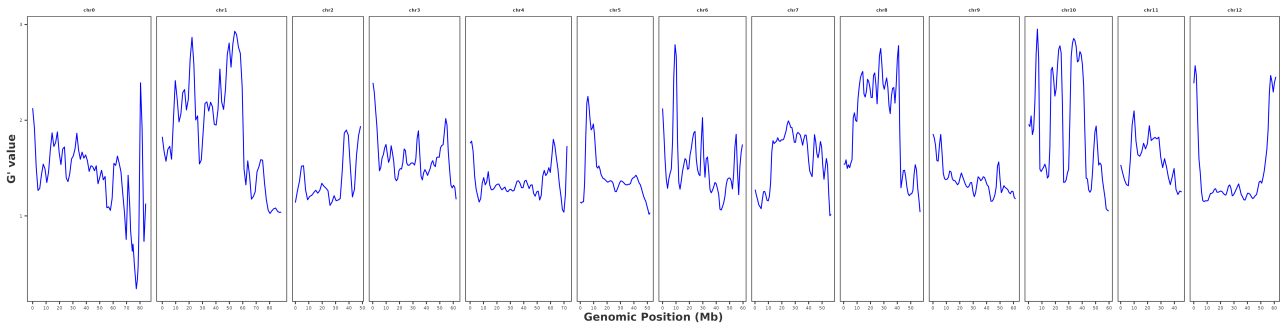


图23 Gprime值分布图

2.11.2 候选区域

选取qvalue值0.05对应的Gprime值作为阈值线确定候选区域。最终的结果文件见**candidate_Region_gprime.txt***。关联区域的结果统计见表15。

表15 关联区域统计

Search

Chr	Start	End
-	-	-

共 1 条

注：

Chr： 染色体编号；

Start： 起始位置；

End： 终止位置；

结果目录：9_Gprime/Target/

2.11.3 候选区域突变基因GO注释

GO是Gene Ontology的缩写。GO的产生主要是为了解决同一基因在不同数据库中定义的混乱性以及不同物种的同一基因在功能定义上的混乱性。它是一个国际标准化的基因功能分类体系，提供了一套动态更新的标准词汇表（controlled vocabulary）来全面描述生物体中基因和基因产物的属性。GO总共有三个ontology，分别描述基因的分子功能（molecular function）、所处的细胞位置（cellular component）、参与的生物学过程（biological process）。每个筛选出的基因对应的GOid及对应功能详见GO/Target.go.annotation.xls。将这些GOid在level2级别进行了汇总并统计，统计结果详见GO/Target_GO_level2.xls。统计结果可视化见图24。

图24 GO注释统计图

结果目录：9_Gprime/Target/GO

2.11.4 候选区域突变基因的 KEGG功能注释

KEGG（Kyoto Encyclopedia of Genes and Genomes），又称京都基因与基因组百科全书，是基因组破译方面的数据库。KEGG注释的主要目的包括两个：KO（KEGG Ortholog）注释，即将分子网络的相关信息进行跨物种注释；KEGG Pathway注释，即代谢通路注释，获得物种内分子间相互作用和反应的网络。

候选区域非同义突变基因的KO及Pathway注释主要采用KEGG 的KAAS自动化注释系统完成（Moriya Y等，2007），其中基因集选择近源物种，基因KO的判别规则选取 bi-directional best hit（BBH），KO注释的结果见“target.kegg.xls”。KO注释完成后将KO映射到相应的KEGG Pathway通路上， KEGG Pathway统计结果见图25。

图25 KEGG注释统计图

结果目录：9_Gprime/Target/KEGG

2.12 LOD smooth 分析（QTGseq）

2.12.1 LOD smooth鉴定结果

利用R语言中的QTGseq包进行smoothLOD方法分析，过滤标准见下表。

表 16 过滤标准

Search

序号	过滤条件
1	保留亲本1杂合、亲本2纯合的SNP位点
2	保留亲本深度大于8x的点
3	保留子代混池深度大于8x的点
4	去掉两个子代池基因型纯合且一致的位点

共 4 条

使用smoothLOD方法鉴定结果见下图。

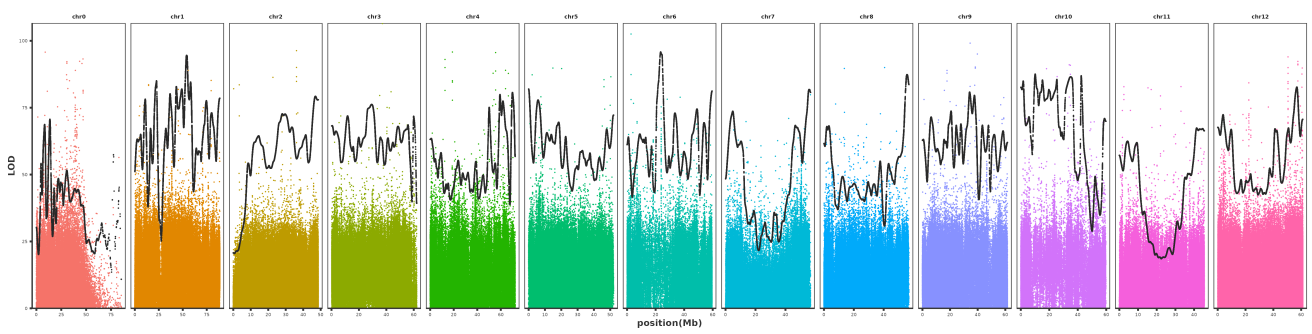


图26 关联区域统计

2.12.2 候选区域

用4M窗口进行LODsmooth计算，提取LODsmooth计算后的所有极大值，最后筛选Pvalue小于等于1e10(-4)且LOD值大于等于3的位点作为性状相关位点，以最终候选位点为中心提取2M区间作为候选区域，最终结果见**candidate_lodsmooth.txt***，关联结果统计见表17。

表 17 性状相关位点

Search

Chr	Start	End
chr0	6,285,694	10,285,694
chr0	11,590,393	15,590,393
chr0	15,826,766	19,826,766
chr0	23,723,423	27,723,423
chr0	28,592,224	32,592,224
chr0	36,474,186	43,956,155
chr0	44,921,188	48,921,188
chr0	54,580,709	63,140,573
chr0	75,484,708	79,484,708
chr0	81,229,424	85,229,424

注：

Chr： 染色体编号；

Start： 起始位置；

End： 终止位置；

结果目录：10_LODSmooth/Target/

2.12.3 候选区域突变基因GO注释

GO是Gene Ontology的缩写。GO的产生主要是为了解决同一基因在不同数据库中定义的混乱性以及不同物种的同一基因在功能定义上的混乱性。它是一个国际标准化的基因功能分类体系，提供了一套动态更新的标准词汇表（controlled vocabulary）来全面描述生物体中基因和基因产物的属性。GO总共有三个ontology，分别描述基因的分子功能（molecular function）、所处的细胞位置（cellular component）、参与的生物学过程（biological process）。每个筛选出的基因对应的GOid及对应功能详见GO/Target.go.annotation.xls。将这些GOid在level2级别进行了汇总并统计，统计结果详见GO/Target_GO_level2.xls。统计结果可视化见图21。

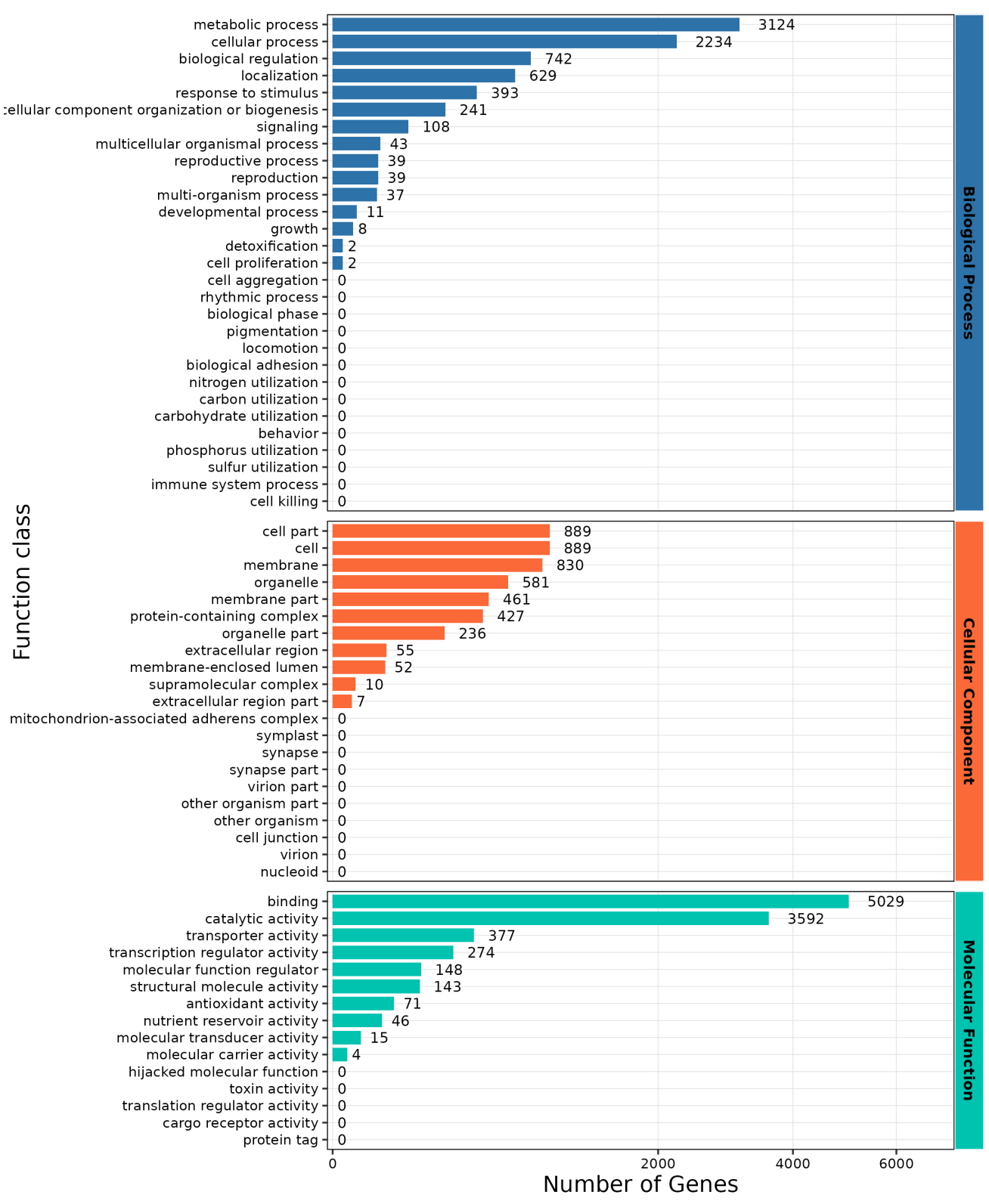


图21. GO注释统计图

结果目录：10_LODSmooth/Target/GO

2.12.4 候选区域突变基因的 KEGG功能注释

KEGG（Kyoto Encyclopedia of Genes and Genomes），又称京都基因与基因组百科全书，是基因组破译方面的数据库。KEGG注释的主要目的包括两个：KO（KEGG Ortholog）注释，即将分子网络的相关信息进行跨物种注释；KEGG Pathway注释，即代谢通路注释，获得物种内分子间相互作用和反应的网络。

候选区域非同义突变基因的KO及Pathway注释主要采用KEGG 的KAAS自动化注释系统完成（Moriya Y等，2007），其中基因集选择近源物种，基因KO的判别规则选取 bi-directional best hit（BBH），KO注释的结果见“target.kegg.xls”。KO注释完成后将KO映射到相应的KEGG Pathway通路上， KEGG Pathway统计结果见图22。

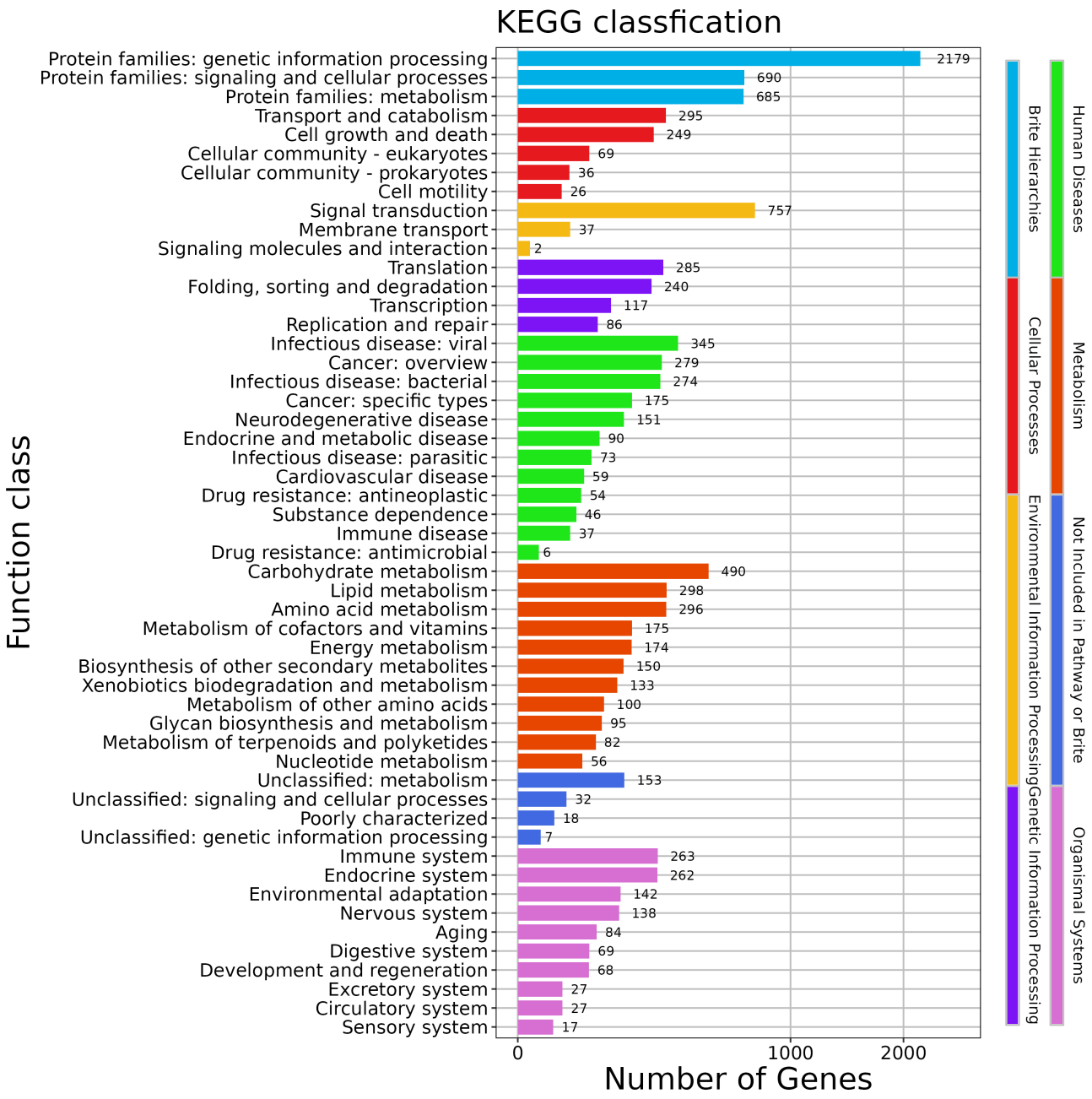


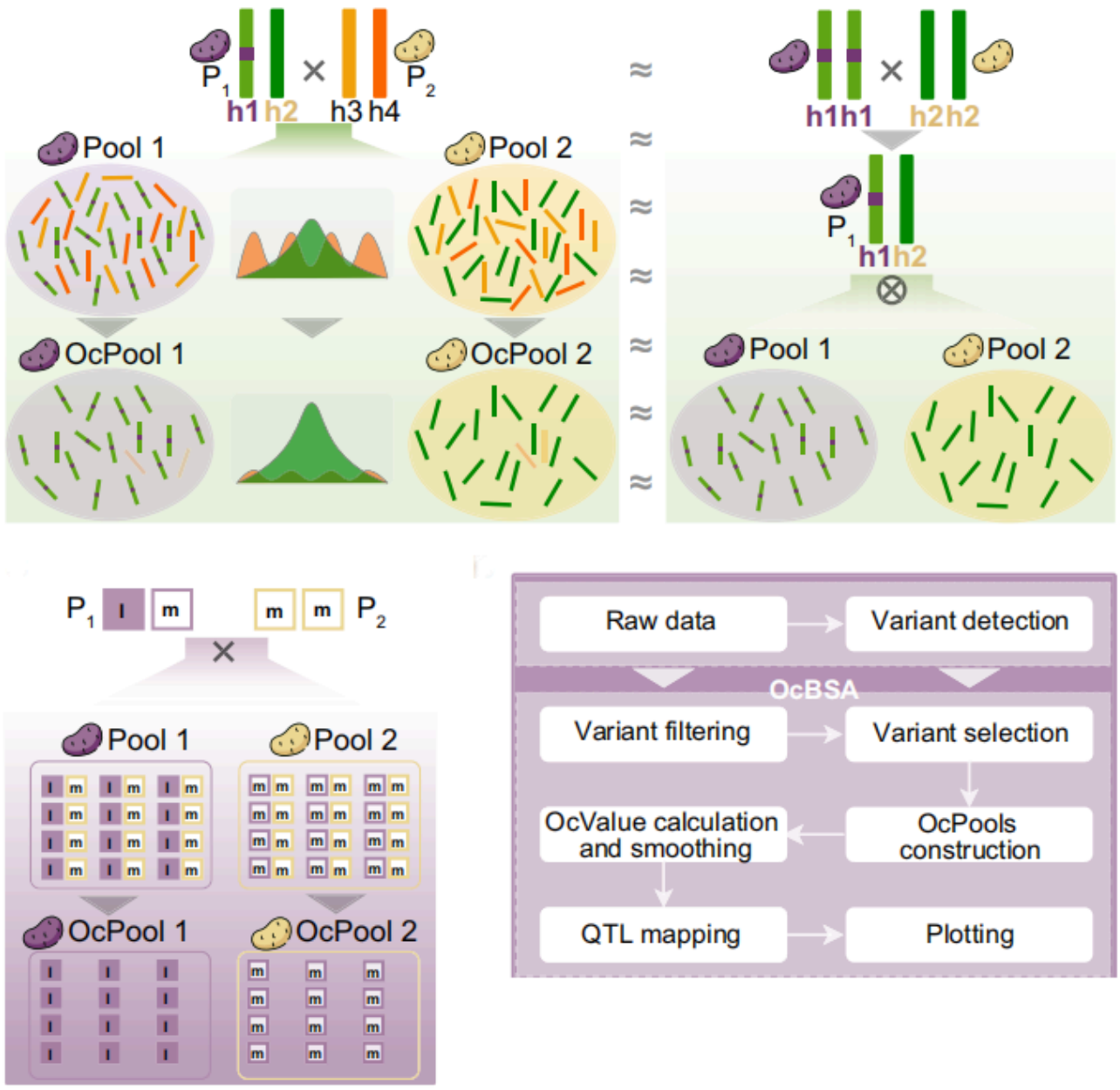
图22 KEGG注释统计图

结果目录：10_LODSmooth/Target/KEGG

2.13 OcBSA分析

2.13.1 OcBSA算法介绍

OcBSA算法如下图模型例示，两个亲本中与目标性状（紫色）分离相关的为显性亲本（P1）。在显性亲本中含有显性基因的单体型为h1，另外一个单体型为h2。隐性亲本（P2）的两个单体型则为h3和h4。在F1后代中获得h1单倍型的单株表现为紫色，获得h2的为黄色；h3和h4对后代的表型没有影响。根据此规律，选取特定标记，在后代混池测序数据中移除隐性亲本的两个单体型h3和h4的信息，得到仅包含单体型h1和h2的两个新的OcPools。新的OcPools类似于F2分离群体的混池，因此，进一步根据亲本P1的两个单体型h1和h2在OcPools中的分离情况可实现目标性状调控基因的定位。



OcBSA示意图

2.13.2 OcBSA分析结果

利用OcBSA工具计算每个位点的OcValue值，再以1M窗口进行拟合得到每个窗口的smoothed_OcValue值，突变位点过滤标准见下表。

表 18 过滤标准

Search

序号	过滤条件
1	保留亲本深度8~500x的点
2	保留子代混池深度8~500x的点

共 2 条

OcBSA结果见下图，图中有3个阈值线，从上到下分别为top0.5%（蓝色），top1%（绿色），top5%（红色）。

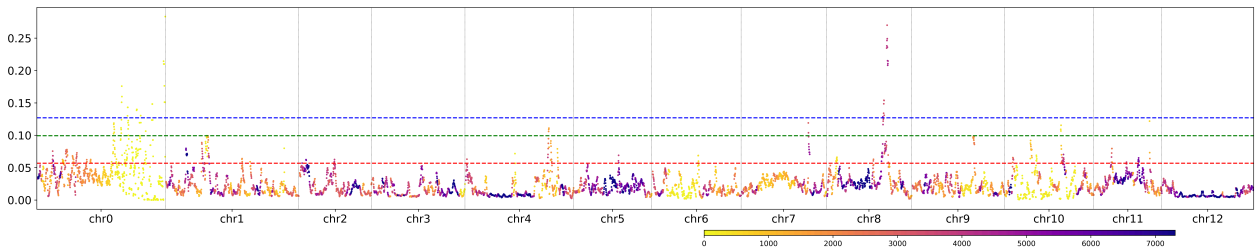


图23 OcBSA结果图

2.13.3 候选区域

利用OcBSA工具计算每个位点的OcValue值，再以1M窗口进行拟合得到每个窗口的smoothed_OcValue值，最终选取smoothed_OcValue值top0.5%作为阈值线作为候选区域，最终结果见**target_ocbsa.txt***，关联结果统计见表17。

表 19 性状相关位点

Search

Chr	Start	End
chr8	36,900,000	38,800,000
chr8	39,400,000	41,300,000
chr0	55,500,000	56,700,000
chr0	59,400,000	60,600,000
chr0	65,400,000	66,600,000
chr0	68,100,000	69,100,000
chr0	75,700,000	76,800,000
chr0	83,200,000	85,300,000

共 8 条

注：

Chr： 染色体编号；

Start： 起始位置；

End： 终止位置；

结果目录：11_OcBSA/Target/

2.13.4 候选区域突变基因GO注释

GO是Gene Ontology的缩写。GO的产生主要是为了解决同一基因在不同数据库中定义的混乱性以及不同物种的同一基因在功能定义上的混乱性。它是一个国际标准化的基因功能分类体系，提供了一套动态更新的标准词汇表（controlled vocabulary）来全面描述生物体中基因和基因产物的属性。GO总共有三个ontology，分别描述基因的分子功能（molecular function）、所处的细胞位置（cellular component）、参与的生物学过程（biological process）。每个筛选出的基因对应的GOid及对应功能详见GO/Target.go.annotation.xls。将这些GOid在level2级别进行了汇总并统计，统计结果详见GO/Target_GO_level2.xls。统计结果可视化见图21。

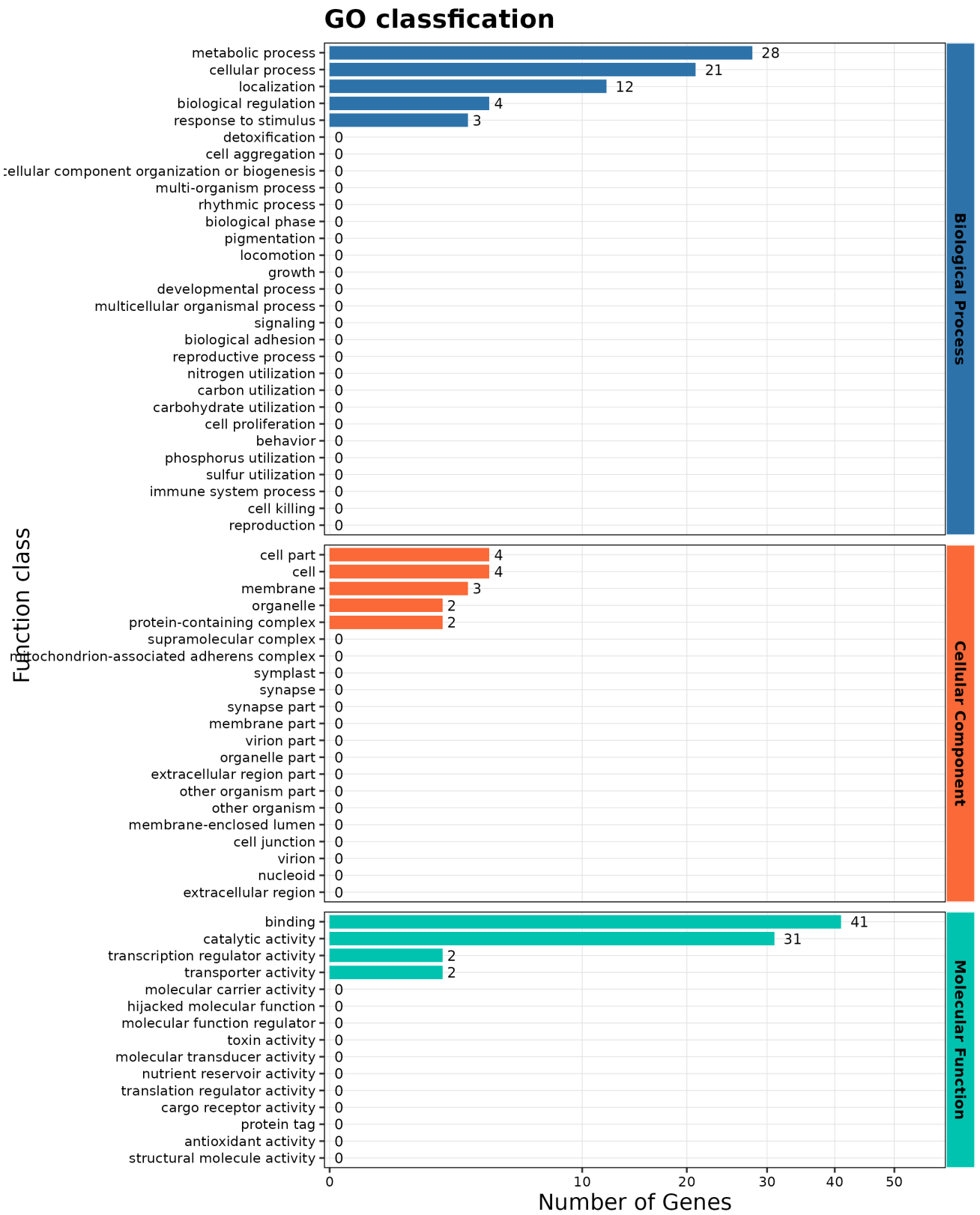


图24 GO注释统计图

结果目录：11_OcBSATarget/GO

2.13.5 候选区域突变基因的 KEGG功能注释

KEGG（Kyoto Encyclopedia of Genes and Genomes），又称京都基因与基因组百科全书，是基因组破译方面的数据库。KEGG注释的主要目的包括两个：KO（KEGG Ortholog）注释，即将分子网络的相关信息进行跨物种注释；KEGG Pathway注释，即代谢通路注释，获得物种内分子间相互作用和反应的网络。

候选区域非同义突变基因的KO及Pathway注释主要采用KEGG 的KAAS自动化注释系统完成（Moriya Y等，2007），其中基因集选择近源物种，基因KO的判别规则选取 bi-directional best hit（BBH），KO注释的结果见“target.kegg.xls”。KO注释完成后将KO映射到相应的KEGG Pathway通路上， KEGG Pathway统计结果见图22。

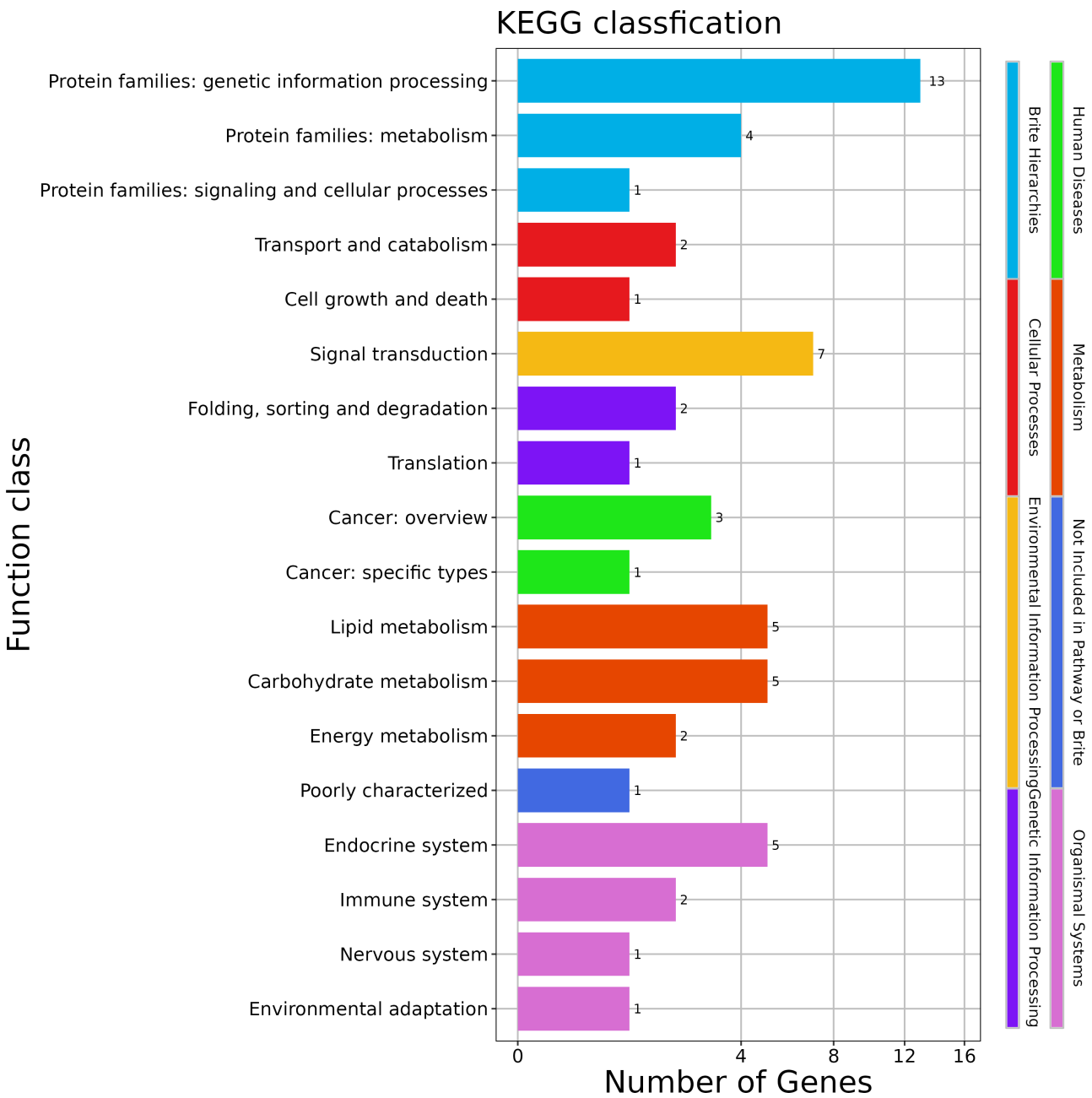


图25 KEGG注释统计图

结果目录：11_OcBSA/Target/KEGG

3.1 参考文献

[1] Schubert, Mikkel, Stinus Lindgreen, and Ludovic Orlando, AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Research Notes 2016,9(1):1-7.

[2] Li, H., and R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009,25(14):1754-60.

[3] Li H., *Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 2009,25, 2078-9.

[4] Zhu P Fau - He, Lingyu, et al.Correction: OTG-snpcaller: An Optimized Pipeline Based on TMAP and GATK for SNP Calling from Ion Torrent Data. Electronic2015,26,1932-6203.

[5] McKenna A, Hanna M, Banks E. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. 2010 GENOMERESEARCH 20:1297-303

[6] Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data, Nucleic Acids Research, 2010,38:e164.

[7] Takaharu Negoro, Kanami Orihara, Tomoko Irahara, Hiroshi Nishiyama, Kanae Hagiwara, Risa Nishida, Hiroki Takagi,Kazue Satoh, YoshikiYamamoto,Shunichi Shimizu,Tamio Hagiwara,Masakazu Ishii,Toshihiro Tanioka1,Yasuko Nakano, Ken Takeda,Isao Yoshimura,Yoji Iikura andTakashi Tobe.Influence of SNPs in cytokine-related genes on the severity of food allergy and atopic eczema in children, 2006 Dec;17(8):583-90

[8] Magwene PM, Willis JH, Kelly JK (2011) The Statistics of Bulk Segregant Analysis Using Next Generation Sequencing. PLOS Computational Biology 7(11): e1002255. doi.org/10.1371/journal.pcbi.1002255

[9] QTLseqr: An R Package for Bulk Segregant Analysis with Next-Generation Sequencing. Plant Genome. 2018 Jul;11(2). doi: 10.3835/plantgenome2018.01.0006

[10] Zhang HW, Wang X, Pan QC, Li P, Li J, Han LQ, Liu YJ, Wang PX, Li DD, Liu Y, Zhang YM, Wang GY, Li L: QTG-seq accelerates QTL fine-mapping through QTL partitioning and whole genome sequencing on bulked segregating samples, Molecular Plant, 2018,DOI:https://doi.org/10.1016/j.molp.2018.12.018.

4.1 名词解释

4.1.1 质量值（Q值）

质量值，简称 Q 值，是碱基读取错误率 p 的取整映射结果。目前，Q 值与 p 值的对应关系有两种计算方式：标准 Sanger 变体计算方式以及早期 Solexa管道（如Illumina Genome Analyzer 软件）计算方式。以标准 Sanger 变体方式为例，Q 值被称为 Phred quality score，它与 p 值之间的对应关系用以下公式计算得到：

Qphred = -10log10**p**

测序错误率与 Q 值的简明对应方式如下：

碱基错误率与质量值对应表

Search

测序错误率	Q 值
5%	13
1%	20
0.1%	30
0.01%	40

共 4 条

不同测序平台得到的 FASTQ 文件中碱基的质量编码采用不同的方案，Q 值与碱基质量的对应关系为：Q 值加上一个偏移数值，得到的结果按照 ASCII 码表转换成对应的字符，参考信息如下所示（http://en.wikipedia.org/wiki/FASTQ_format）。

```
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                               0........9.............................40
                                  3.....9.............................40
 0.2......................26...31.........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

phred_range

如果FASTQ 文件采用 Illumina 1.8+ 版本编码，将所有字符的 ASCII值减去偏移值 33，即可得到碱基的 Q 值。如字符 I 在 ASCII 码表中对应的值为73，减去 33 后得到40，则该字符对应位置的碱基的 Q 值为 40，换算为测序错误率为 0.01%。

4.1.2 SNP

SNP(Single Nucleotide Polymorphism)，即单核苷酸多态性，主要是指在基因组水平上由单个核苷酸的变异所引起的DNA序列多态性，包括单个碱基的转换、颠换等。

4.1.3 InDel

InDel是Insertion和Deletion的总称。主要指一些小片段大小的插入和缺失序列。

4.1.4 CNV

CNV（Copy Number Variation），即拷贝数变异，一般指长度为1 kb以上的基因组大片段的拷贝数增加或者减少

4.1.5 SV

SV（Structural Variation），即染色体结构变异，主要包括大片段的缺失(Deletion, DEL)，插入 (Insertion, INS)，倒位(Inversion, INV)，染色体内易位(intra-chromosomal translocation, ITX)，染色体间易位(Inter-chromosomal translocation, CTX)等。

4.2 文件格式介绍

4.2.1 Fastq格式

样品经过 DNA 抽提、纯化、建库和测序，得到下机数据。下机数据以双端（Paired-end） FASTQ 格式保存，见图 1。FASTQ 格式是一种文本格式，常用于存储生物学序列及其对应的质量分值。FASTQ 格式文件可以采用文本工具（如写字板、UltraEdit等工具）打开，其中第一行为序列名称（以 @ 开头）；第二行为原始序列；第三行为“+”号；不代表任何意义；第四行为第二行序列所对应的碱基质量信息。

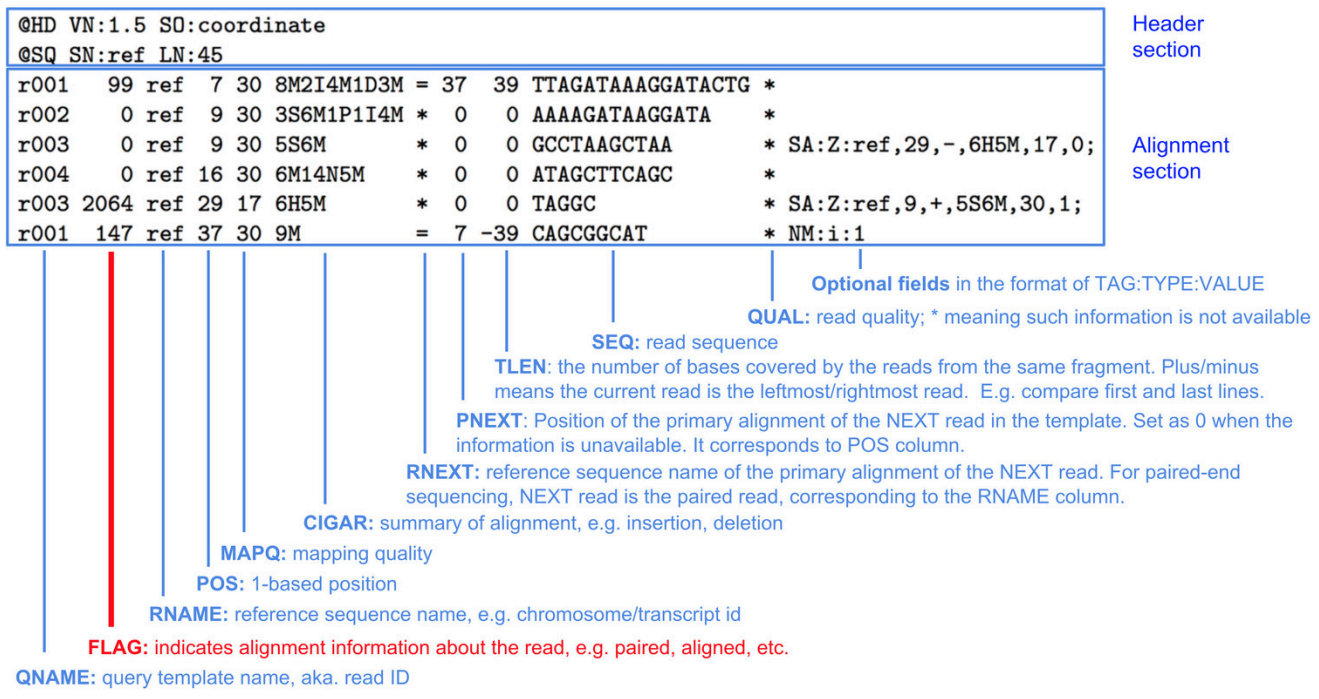
@M00200:111:000000000-A6VNV:1:1101:15594:1337 1:N:0:6

ACGCGGGTATCTAATCCTGTTTGCTCCCCACGCTTTCGCGCCTCAGTGTCAGTTAC+

ABABADBBDFFFGGGFGGGFGGHGBGHGGHGGGGGGGHGGGGGGGGHHGGFBGEG

4.2.2 sam/bam格式

Sam格式文件是一种记录序列比对结果的文本格式文件，以TAB为分割符的文本格式，主要用于测序数据mapping到基因组上的结果表示。bam文件sam文件的二进制文件，占用的磁盘空间比sam文本文件小，需要一些工具才能查看其中的内容，如samtools。主流的比对软件，如bwa和bowtie等，输出的结果文件就是sam/bam格式的。



sam格式文件示例图

sam/bam格式文件包括两部分，以#开头的为注释信息，称为sam/bam文件的头部区，其余为比对结果部分。注释信息部分，都以@开头，用不同的tag表示不同的信息。主要有以下几种tag：

@HD VN:1.0 SO:unsorted（排序类型）

头部区第一行：VN是格式版本；SO表示比对排序的类型，有unknown（default），unsorted，queryname和coordinate几种。

@SQ SN:contig1 LN:9401（序列ID及长度）

参考序列名，这些参考序列决定了比对结果sort的顺序，SN是参考序列名；LN是参考序列长度；每个参考序列为一行。

@RG ID:sample01（样品基本信息）

Read Group。1个sample的测序结果为1个Read Group；该sample可以有多个library的测序结果，可以利用bwa mem -R 加上去这些信息。

例如：@RG ID:ZX1_ID SM:ZX1 LB:PE400 PU:Illumina PL:Miseq

ID：样品的ID号 SM：样品名 LB：文库名 PU：测序仪 PL：测序平台

这些信息可以在形成sam文件时加入，ID是必须要有的后面是否添加看分析要求

@PG ID:bowtie2 PN:bowtie2 VN:2.0.0-beta7（比对所使用的软件及版本）

例如：@PG ID:bwa PN:bwa VN:0.7.12-r1039 CL:bwa sampe -a 400 -f ZX1.sam -r @RG ID:ZX1_ID SM:ZX1 LB:PE400 PU:Illumina PL:Miseq ../0_Reference/Reference_Sequence.fa ZX_HQ_clean_R1.fq.sai ZX_HQ_clean_R2.fq.sai ../2_HQData/ZX_HQ_clean_R1.fq ../2_HQData/ZX_HQ_clean_R2.fq

这里的ID是bwa，PN是bwa，VN是0.7.12-r1039版本。CL可以认为是运行程序@RG是上面RG表示的内容，后面是程序内容

比对结果部分包括11个必须字段和一些可选字段。每个字段的内容如下：

QNAME：比对片段的编号。

FLAG：模板比对的情况，数字表示，每一个数字表示一种比对情况，此处的值是符合情况的数字相加总和。

RNAME：参考序列编号，如果第一部分注释中对SQ-SN进行了定义，这里必须与其保持一致，没有mapping上的序列，是“*”。

POS：比对上的位置（从1开始计数的，没比对上的为0）。

MAPQ：mapping的质量。

CIGAR：比对信息表达，数字加字母表示。如3S6M1P1I60M:前三个碱基被剪切去除了，然后又6个比对上了，接着有一个缺口，一个碱基插入，后面60个比对上了。

RNEXT：另一个片段比对上的参考序列的编号。如果没有另外的片段，此处是“*”，如果是同一片段，用“=”。

PNEXT：另一个片段比对上的位置。若不可用，此处为“0”。

TLEN：template的(比对好的片段)的长度，最左边的为正，最右边的为负。单片段比对上或者不可用时，为“0”。

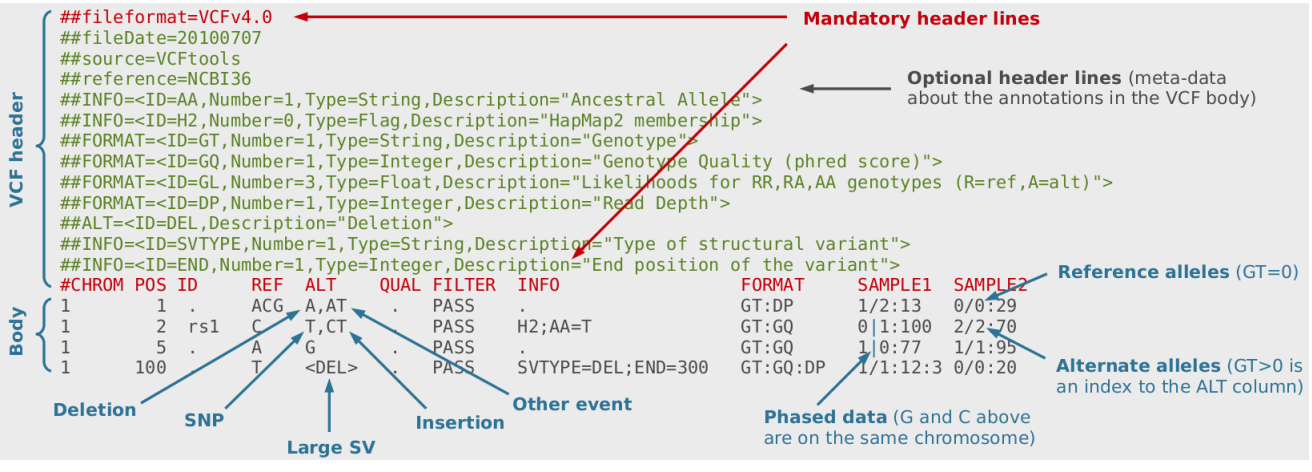
SEQ: 片段的序列信息

QUAL：测序质量信息。

template（参考序列和比对上的序列共同组成的序列为template）的一些信息

4.2.3 VCF格式

VCF格式文件是用于描述变异(包括SNP、InDel和SV等)结果的文本文件。主流的变异检测的软件，如GATK和SAMtools，输出的结果文件就是VCF格式的。不过二者输出的结果格式上有一点细微的差别。



VCF格式文件部分示例图

VCF格式文件包括两部分，以#开头的为表头部分，其余为主体部分。表头部分是对主体部分一些字段的说明。主体部分包括下面几列，其中前9列格式是固定的，后面几列是每个样本的信息。每一列的内容如下：

CHROM 和 POS：代表参考序列名和variant的位置；如果是InDel的话，位置是InDel的第一个碱基位置。

ID：variant的ID。比如在dbSNP中有该SNP的id，则会在此行给出；若没有，则用'.'表示其为一个novel variant。

REF 和 ALT：参考序列的碱基和 Variant的碱基。

QUAL：Phred格式(Phred_scaled)的质量值，表示在该位点存在variant的可能性；该值越高，则variant的可能性越大；计算方法：Phred值 = -10 * log (1-p) p为variant存在的概率; 通过计算公式可以看出值为10的表示错误概率为0.1，该位点为variant的概率为90%。

FILTER：使用上一个QUAL值来进行过滤的话，是不够的。GATK能使用其它的方法来进行过滤，过滤结果中通过则该值为”PASS”;若variant不可靠，则该项不为”PASS”或”.”。

INFO：这一行是variant的详细信息，内容很多，后面单独说明。

FORMAT 和XXX：这两行合起来提供了XXX这个sample的基因型的信息。’ XXX′代表样品名称。

VCF文件中INFO列的信息

该列信息最多了，都是以 “TAG=Value”,并使用”;”分隔的形式。其中很多的注释信息在VCF文件的头部注释中给出。以下是这些TAG的解释

AC(Allele Count) 表示Alternative allele的数目；

AF(Allele Frequency) 表示Alternative allele的频率；

AN(Allele Number) 表示Allele的总数目。

对于1个diploid sample而言：若基因型为0/1，则表示sample为杂合子，Alternative allele数为1(双倍体的sample在该位点只有1个等位基因发生了突变)，Allele的频率为0.5(双倍体的 sample在该位点只有50%的等位基因发生了突变)，总的Allele为2。若基因型1/1，则表示sample为纯合的，Alternative allele数为2，Alternative allele的频率为1，总的Allele为2。

DP：reads覆盖度。是一些reads被过滤掉后的覆盖度。

Dels：Fraction of Reads Containing Spanning Deletions。进行SNP和InDel calling的结果中，有该TAG并且值为0表示该位点为SNP，没有则为InDel。

FS：使用Fisher’s精确检验来检测strand bias而得到的Fhred格式的p值。该值越小越好。一般进行filter的时候，可以设置 FS < 10 ~ 20。

HaplotypeScore：Consistency of the site with at most two segregating haplotypes

InbreedingCoeff：Inbreeding coefficient as estimated from the genotype likelihoods per-sample when compared against the Hard-Weinberg expectation

MLEAC：Maximum likelihood expectation (MLE) for the allele counts (not necessarily the same as the AC), for each ALT allele, in the same order as listed

MLEAF：Maximum likelihood expectation (MLE) for the allele frequency (not necessarily the same as the AF), for each ALT alle in the same order as listed

MQ：RMS Mapping Quality

MQ0：Total Mapping Quality Zero Reads

MQRankSum：Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities

QD：Variant Confidence/Quality by Depth

RPA：Number of times tandem repeat unit is repeated, for each allele (including reference)

RU：Tandem repeat unit (bases)

ReadPosRankSum：Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias

STR：Variant is a short tandem repeat

VCF文件中FORMAT列信息

FORMAT与后面每一列样品的数据一一对应，前者为格式，后者为格式对应的数据。

GT：样品的基因型（genotype）。两个数字中间用’/’分开，这两个数字表示双倍体的sample的基因型。0 表示样品中有ref的allele； 1 表示样品中variant的allele； 2表示有第二个variant的allele。因此： 0/0 表示sample中该位点为纯合的，和ref一致； 0/1 表示sample中该位点为杂合的，有ref和variant两个基因型； 1/1 表示sample中该位点为纯合的，和variant一致。

AD 和 DP：AD(Allele Depth)为sample中每一种allele的reads覆盖度,在diploid中则是用逗号分割的两个值，前者对应ref基因型，后者对应variant基因型； DP（Depth）为sample中该位点的覆盖度。

GQ：基因型的质量值(Genotype Quality)。Phred格式(Phred_scaled)的质量值，表示在该位点该基因型存在的可能性；该值越高，则Genotype的可能性越大；计算方法：Phred值 = -10 * log (1-p) p为基因型存在的概率。

PL：指定的三种基因型的质量值(provieds the likelihoods of the given genotypes)。这三种指定的基因型为(0/0,0/1,1/1)，这三种基因型的概率总和为1。和之前不一致，该值越大，表明为该种基因型的可能性越小。 Phred值 = -10 * log (p) p为基因型存在的概率。

4.3 软件列表

软件列表

Search

软件	功能
fastqc	数据质量评估
fastp	去接头、低质量数据
BWA	序列比对
Picard-tools	格式转化
GATK	变异检测
ANNOVAR	变异注释
CNVnator	鉴定CNV
BreakDancer	鉴定SV

共 8 条