重测序BSA项目结题报告

客户单位: _____

联系人: _____

联系电话: _____

传真: _____

报告日期: _____

天根生化科技（北京）有限公司

目录

1 项目概况

1.1 项目基本信息

(1)样品信息:

样品	编号
亲本 1（父本）	P
亲本 2（母本）	M
混池 1	B1
混池 2	B2

注：编号：对样品的编号，实验建库和后续信息分析均使用该编号。

混池规模：30+30；群体类型：F2群体；研究性状：水稻千粒重

(2) 参考基因组信息：

根据水稻的基因组大小以及 GC 含量等信息，最终选取日本晴水稻基因组作为参考基因组。

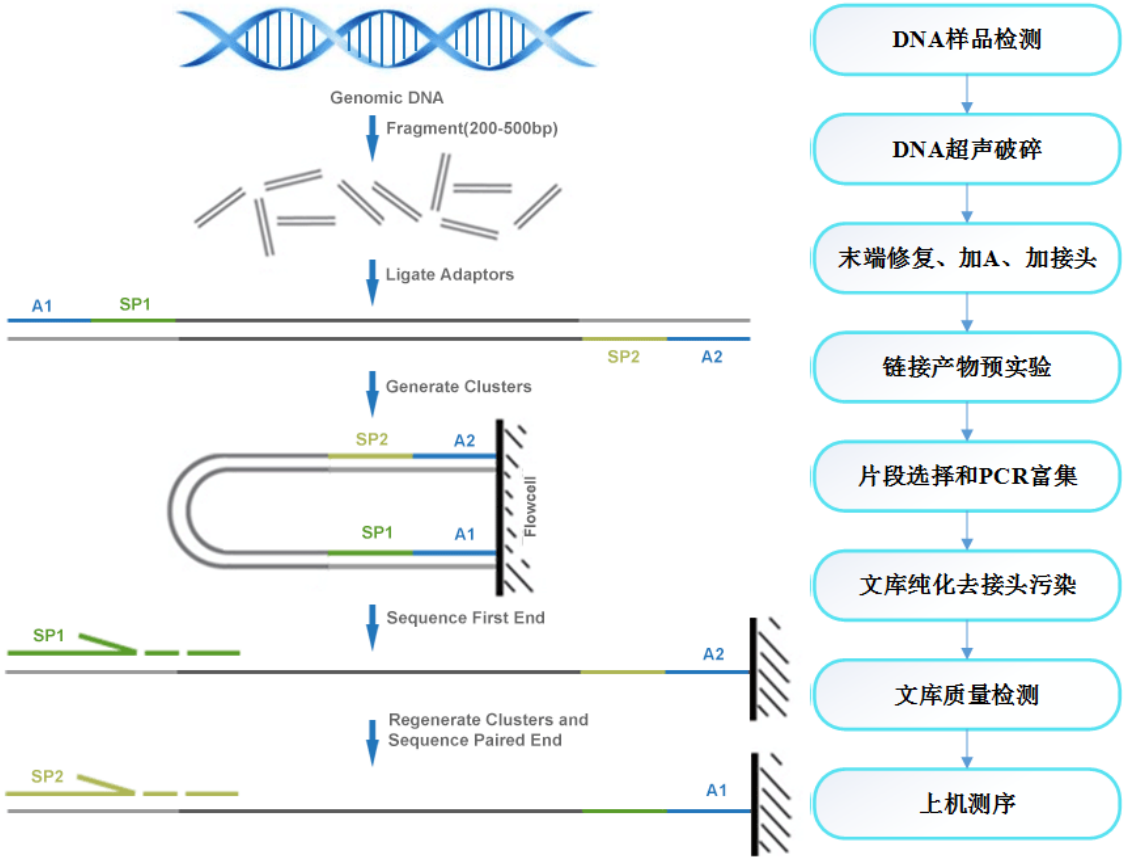
具体信息如下所示：

1. 测序物种信息：水稻（*Oryza sativa*），实际基因组大小为 419.8 Mb，GC 含量为 45.67%；
2. 参考物种信息：日本晴水稻（*Oryza sativa*indica）[1]基因组，组装出的基因组大小为 374.3 Mb，GC 含量为 43.56%，ScaffoldN50 为 500Kb，，该基因组组装到染色体水平，有基因注释信息，版本号为 v7.0，下载地址：http://rapdb.dna.affrc.go.jp/。

2 项目流程

2.1 实验流程

实验流程按照Illumina公司提供的标准protocol执行，包括样品检测、文库构建、文库质量检测和上机测序，具体流程如下:



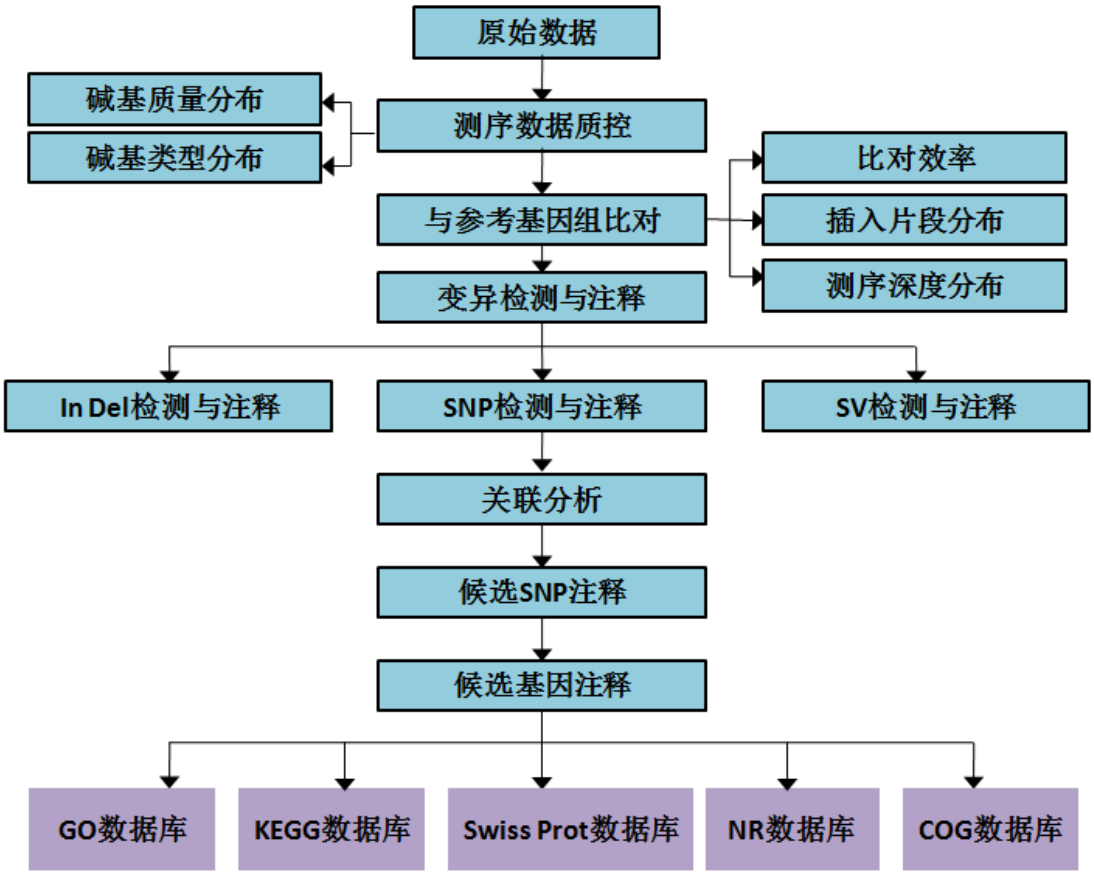
实验流程图

样品检测合格后，用超声破碎的方法将DNA随机打断成350bp的片段，DNA片段经末端修复、3'端加A、加测序接头、纯化、PCR扩增完成测序文库的构建。文库经质检合格后通过Illumina HiSeq^{TM}4000进行测序。

2.2 信息分析流程

信息分析的内容包括：数据质控（去除接头和低质量数据）、与参考基因组比对、变异检测与注释（SNP、InDel）、关联分析、候选SNP及候选基因的注释。

重测序BSA生物信息分析具体流程如下图所示：



重测序BSA生物信息分析流程图

3 生物信息学分析

3.1 测序数据质控

3.1.1 原始数据介绍

高通量测序（如Illunima HiSeq 4000等测序平台）得到的原始图像数据文件，经碱基识别（Base Calling）分析转化为原始测序序列（Sequenced Reads），我们称之为Raw Data或Raw Reads，结果以FASTQ（简称为fq）文件格式存储，其中包含测序序列（Reads）的序列信息以及其对应的测序质量信息。测序样品中真实数据随机截取结果如下：

```
@HWI-7001455:110: C3B41ACXX:4:1101:1401:2163 1:N:0: TAAGGC
CTCTCTCCTATCTTTCCAACCATCTGATAACACCGAACATCCATATTGAGCCCACACTTCTTGATGATCTTTCAATATTTTATGAT
+
CCCFFFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJHHHHHHHFFFFFFFFFEEEEEE
```

FASTQ格式文件中每个Read由四行描述，其中第一行以“@”开头，随后为Illumina测序识别符（Sequence Identifiers）和描述文字（选择性部分）；第二行是碱基序列；第三行以“+”开头，随后为Illumina测序识别符（选择性部分）；第四行是对应序列的测序质量。

Illumina测序识别符（Sequence Identifiers）详细信息见如下：

Illumina 测序标识详细信息

HWI-7001455	Unique instrument name
110	Run ID
C3B41ACXX	Flowcell ID
4	Flowcell lane
1101	Tile number within the flowcell lane
1401	'x'-coordinate of the cluster within the tile
2163	'y'-coordinate of the cluster within the tile
1	Member of a pair, 1 or 2 (paired-end or mate-pair reads only)
N	Y if the read fails filter (read is bad), N otherwise
0	0 when none of the control bits are on, otherwise it is an even number
TAAGGC	Index sequence

通过使用第四行中每个字符对应的ASCII值进行计算，即得到对应第二行碱基的测序质量值。如果测序错误率用e表示，Illunima HiSeq 4000的碱基质量值用Qphred

表示，则有下列关系：

$$Q_{phred} = -10\log_{10}(e)$$

Illunima Casava 1.8版本测序错误率与测序质量值简明对应关系如下表所示：

测序错误率	测序质量值	对应字符
5%	13	.
1%	20	5
0.1%	30	?
0.01%	40	I

碱基识别（Base Calling）分析软件：Illunima Casava 1.8版本

测序参数：双端测序(Paired end，PE)

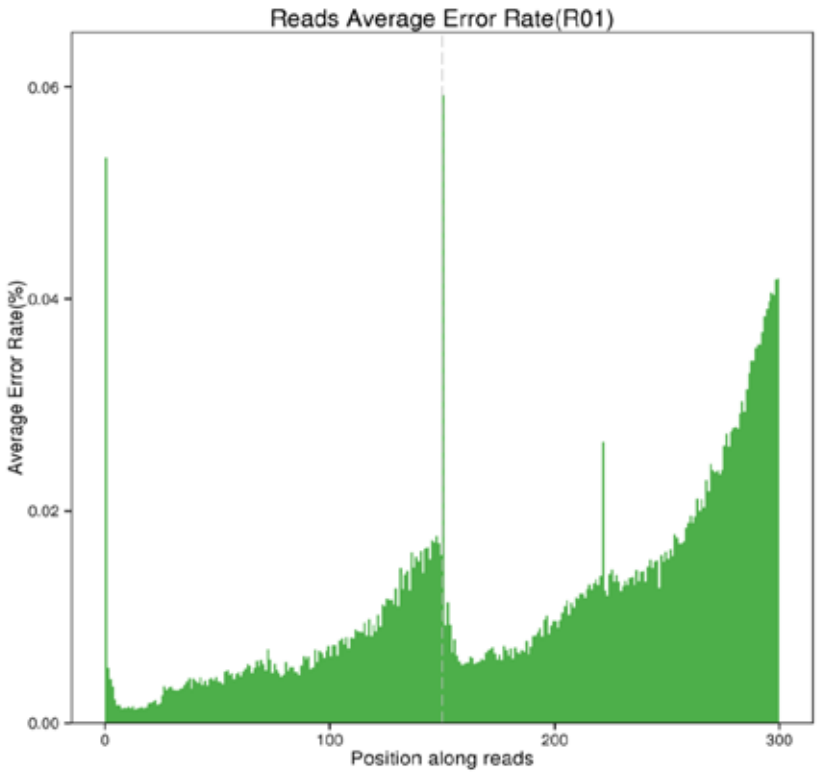
测序序列读长：151bp

3.1.2 碱基测序质量分布

每个碱基测序错误率是通过测序 Phred数值（Phred score，Qphred）得到，而Phred数值是在碱基识别（Base Calling）过程通过一种预测碱基判别发生错误概率模型计算得到的，对应关系如下表所显示：

Phred分值	不正确的碱基识别	碱基正确识别率
10	1/10	90%
20	1/100	99%
30	1/1000	99.9%
40	1/10000	99.99%

在Hiseq4000测序系统测序时，首先会对文库进行芯片制备，目的是将文库DNA模板固定到芯片上，在固定DNA模板的过程中，每个DNA分子会形成一个簇，一个簇就是一个测序位点，在进行固定过程中极少量的簇与簇之间物理位置会发生重叠，在测序时，测序软件通过前4个碱基对这些重叠的点进行分析和识别，将这些重叠点位置分开，保证每个点测到的是一个DNA分子，因此测序序列5′端前几个碱基的错误率相对较高。另外测序错误率会随着测序序列（Sequenced Reads）的长度的增加而升高，这是由于测序过程中化学试剂的消耗而导致的。因此在进行碱基测序质量分布分析时，样品的碱基质量分布在前4个碱基和后十几个碱基的质量值会低于中间测序碱基，但其质量值都高于Q30，根据质量值和错误率的关系，我们将质量值转换成错误率，绘制错误率分布图如下：



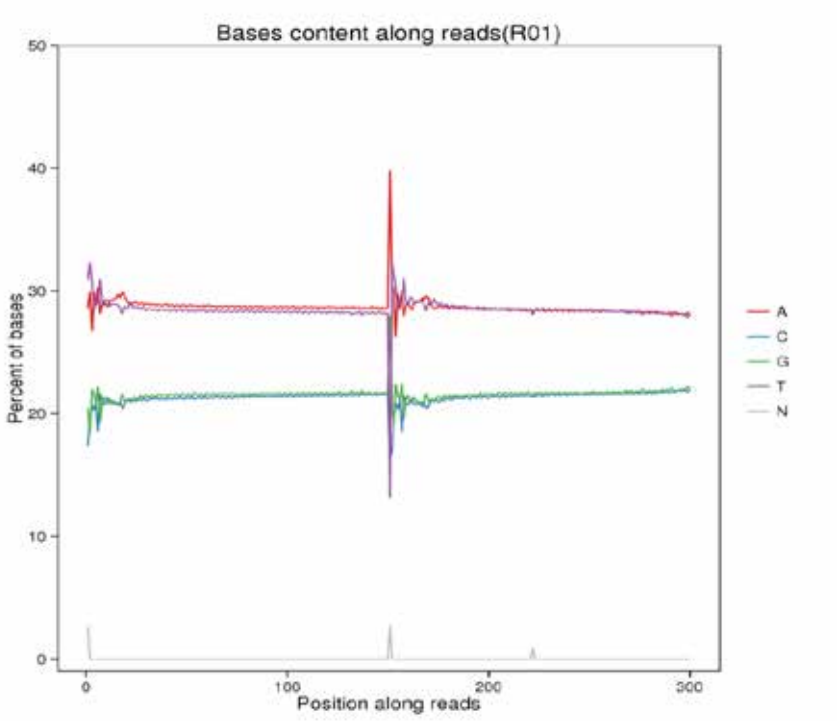
样品P碱基错误率分布

注：横坐标为 reads 的碱基位置，纵坐标为单碱基错误率，前 151bp 为双端测序序列的第一端测序 reads 的错误率分布情况，后 151bp 为另一端测序 reads 的错误率分布情况。

3.1.3 碱基类型分布

碱基类型分布用于检测有无 AT、GC 分离现象，这种分离现象可能是建库测序过程中差异扩增引起的，直接影响到后续的分析。高通量测序的序列为基因组随机打断后的 DNA 片段，位点在整个基因组的分布是近似均匀的，同时根据碱基互补配对的原则，A 与 T 和 C 与 G 的含量分别是一致的。由于测序仪器本身的局限性，前几个碱基的 A/T 和 C/G 含量可能存在着一定波动。

样品各碱基比例分布如下所示：



样品P各碱基比例分布

注：横坐标为reads的碱基位置，纵坐标为碱基所占的比例；不同颜色代表不同的碱基类型，绿色代表碱基G，蓝色代表碱基C，红色代表碱基A，紫色代表碱基T，灰色代表测序中识别不出的碱基N。前151bp为双端测序序列的第一端测序reads的碱基分布，后151bp为另一端测序reads的碱基分布。每个cycle代表测序的每个碱基，如第一cycle即表示该项目所有测序reads在第一个碱基的A、T、G、C、N的分布情况。

该图的结果显示AT、CG碱基基本不发生分离，且曲线较平缓，说明测序结果正常。

3.1.4 低质量数据过滤

测序得到的原始测序序列（Sequenced Reads）或者Raw Reads，里面含有带接头的、低质量的Reads，为了保证信息分析质量，对Raw Reads进行过滤，得到Clean Reads，用于后续信息分析。

数据过滤的主要步骤如下：

(1) 去除带接头（adapter）的reads。

(2) 若一条reads上N（未能确定出具体的碱基类型）的比例大于10%，则过滤掉该Pair-end reads。

(3) 去除低质量reads（质量值Q≤xx的碱基数占整条read的50%以上）。

数据过滤统计结果见下表：

数据过滤统计表

ID	Raw_Reads	Adapter_Related（%）	Inferior_percent (%)	Clean_Reads
P				
M				
B1				
B2				

注： ID：项目样品的编号；Raw_Reads：原始测序reads数；Adapter_Related：含接头被过滤的reads比例；Inferior_percent：N含量超过10%的reads和质量值低于10的碱基超过50%的reads比例；Clean_Reads：过滤后剩余的reads数。

3.1.5 测序数据统计

各样品测序产出数据评估结果如下表所示：

样品测序数据评估统计

ID	Raw_Reads	Clean_Reads	Clean_Base	Q30(%)	GC(%)
P					
M					
B1					
B2					

注：ID：项目样品的编号；Raw_Reads：原始测序reads数目，以四行为一个单位，统计Pair-end序列的个数；Clean_Reads：过滤后的reads数，计算方法同Raw Reads； Clean_Bases：过滤后的碱基数，Clean Reads数乘以序列长度；Q30(%)：质量值大于等于30的碱基占总碱基数的百分比；GC(%)：样品GC含量，即G和C类型的碱基占总碱基的百分比。

3.2 与参考基因组比对统计

重测序获得的测序reads需要重新定位到参考基因组上，才可以进行后续变异分析。bwa[2]软件主要用于二代高通量测序（如Illunima HiSeq 4000等测序平台）得到的短序列与参考基因组的比对。通过比对定位Clean Reads在参考基因组上的位置，统计各样品的测序深度、基因组覆盖度等信息，并进行变异的检测。

3.2.1 比对结果统计

样品的比对结果见下表:

比对结果统计

ID	Total_reads	Mapped(%)	Properly_mapped(%)
P			
M			
B1			
B2			

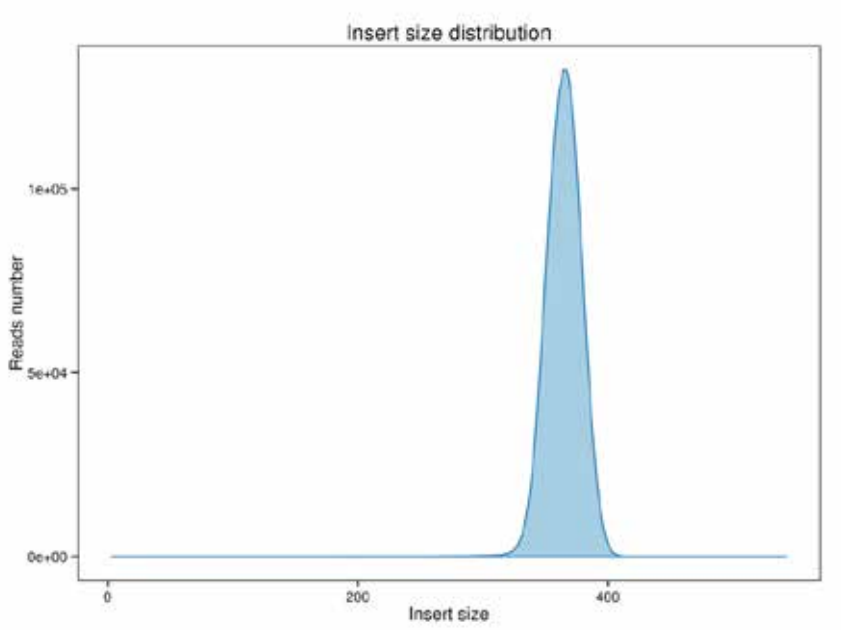
注： ID：项目样品的编号；Total_Reads：Total_Reads数，双端分别统计，即read1和read2记为2条reads；Mapped(%)：定位到参考基因组的Clean Reads数占所有Clean Reads数的百分比；Properly mapped：双端测序序列均定位到参考基因组上且距离符合测序片段的长度分布。

3.2.2 插入片段分布统计

通过检测双端序列在参考基因组上的起止位置，可以得到样品DNA打断后得到的测序片段的实际大小，即插入片段大小(Insert Size)，是信息分析时的一个重要参数。插入片段大小的分布一般符合正态分布，且只有一个单峰，Insert Size分布图可

以展示各个样品的插入片段的长度分布情况。

每个样品测序数据插入片段大小分布的分析使用picard软件工具包中CollectInsertSizeMetric.jar软件实现。



样品P插入片段分布图

注：横坐标为插入片段长度，纵坐标为其对应的reads数。

由上图可知，插入片段长度分布符合正态分布，说明测序数据文库构建无异常。

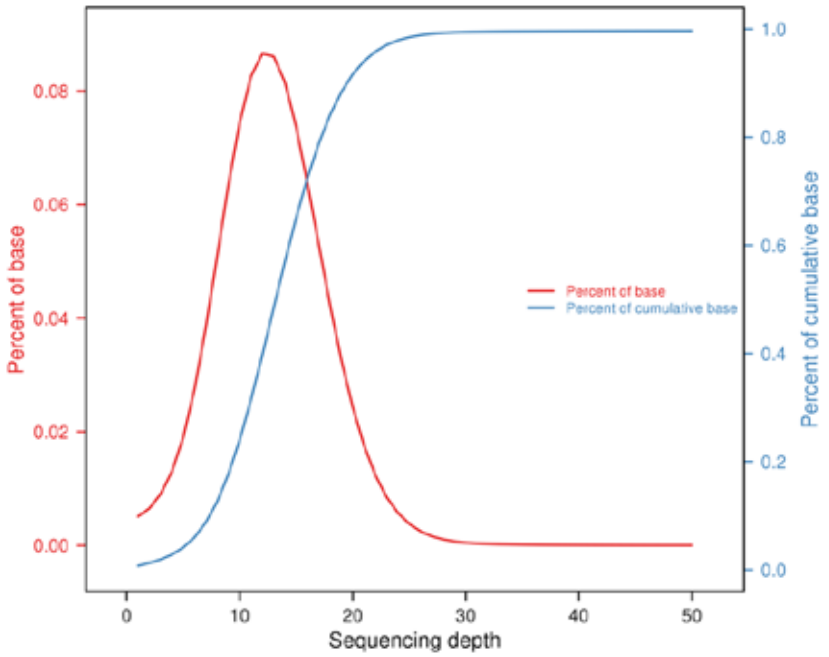
3.2.3 深度分布统计

Reads定位到参考基因组后，可以统计参考基因组上碱基的覆盖情况。参考基因组上被reads覆盖到的碱基数占基因组的百分比称为基因组覆盖度；碱基上覆盖的reads数为覆盖深度。

基因组覆盖度可以反映参考基因组上变异检测的完整性，覆盖到的区域越多，可以检测到的变异位点也越多。覆盖度主要受测序深度以及样品与参考基因组亲缘关系远近的影响。

基因组的覆盖深度会影响变异检测的准确性，在覆盖深度较高的区域（非重复序列区），变异检测的准确性也越高。另外，若基因组上碱基的覆盖深度分布较均匀，也说明测序随机性较好。

样品的碱基覆盖深度分布曲线和覆盖度分布曲线见下图：



样品P的深度分布

注：上图反映了测序深度分布的基本情况，横坐标为测序深度，左纵坐标为该深度对应的碱基所占百分比，对应红色曲线，右纵坐标为该深度及以下的碱基所占百分比，对应蓝色曲线。

各样品的平均覆盖深度和各深度对应的基因组覆盖比例如下表所示：

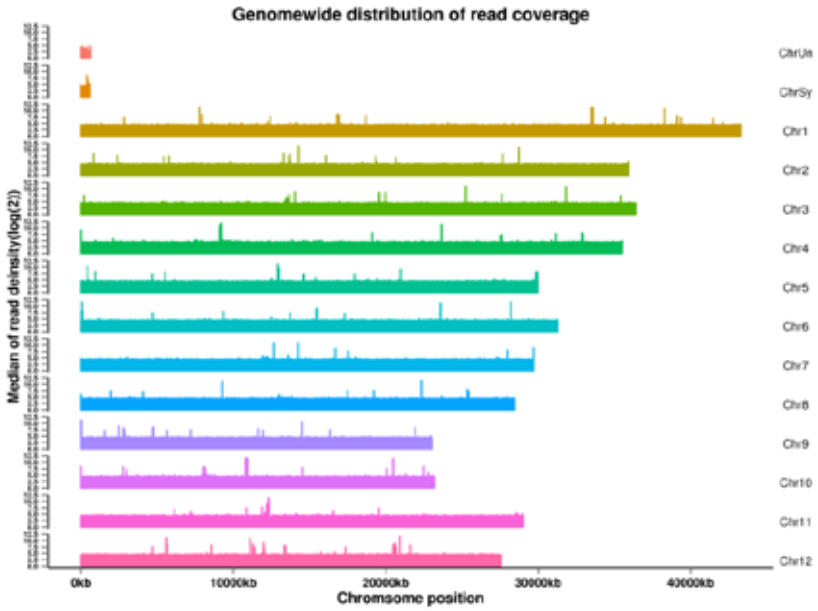
样品覆盖深度和覆盖度比例统计

ID	Ave_depth	Cov_ratio_1X(%)	Cov_ratio_5X(%)	Cov_ratio_10X(%)
P				
M				
B1				
B2				

注： ID：项目样品的编号；Ave-depth：样品平均覆盖深度；Cov_ratio：覆盖深度在给定深度及以上的碱基数占参考基因组总碱基数的比例。

由上表可知，此项目基因组平均覆盖深度约为X，基因组覆盖度约为XX%（至少覆盖1X）。

根据染色体各位点的覆盖深度情况进行作图，若覆盖深度在染色体上的分布比较均匀，则可以认为测序随机性比较好。样品的染色体覆盖深度分布见下图：



样品P染色体覆盖深度分布图

注：横坐标为染色体位置，纵坐标为染色体上对应位置的覆盖深度取对数（log2）得到的值。

由上图可以看出基因组被覆盖的较均匀，说明测序随机性较好。图上深度不均一的地方可能是由于重复序列、PCR偏好性引起的。

3.3 SNP 检测与注释

3.3.1 样品与参考基因组间 SNP 的检测

SNP的检测主要使用GATK[3]软件工具包实现。根据Clean Reads在参考基因组的定位结果，使用Picard[4]进行去重复（Mark Duplicates）、GATK进行局部重比对(Local Realignment)、碱基质量值校正（Base Recalibration）等预处理，以保证检测得到的SNP准确性，再使用GATK进行单核苷酸多态性（Single Nucleotide Polymorphism，SNP）的检测，过滤，并得到最终的SNP位点集。

主要检测过程如下：

(1) 对于BWA比对得到的结果，使用Picard的Mark Duplicate工具去除重复，屏蔽PCR-duplication的影响。

(2) 使用GATK进行InDel Realignment，即对存在插入缺失比对结果附近的位点进行局部重新比对，校正由于插入缺失引起的比对结果错误。

(3)使用GATK进行碱基质量值再校准（Base Recalibration），对碱基的质量值进

行校正。

(4)使用GATK进行变异检测（variant calling），主要包括SNP和InDel。

(5)对SNP进行严格过滤：snp cluster过滤（5bp内如果有2个SNP则过滤掉），Indel附近SNP过滤（InDel附近5bp内的SNP过滤掉）；和相邻INDEL过滤（两个InDel距离小于10bp过滤掉）[5]。具体流程可参考GATK官方网站的BestPractice：

https://www.broadinstitute.org/gatk/guide/best-practices?bpm=DNAseq#variant-discovery-ovw

变异结果使用vcf文件格式展示。vcf文件包括注释行、标题行和数据行三部分。其中注释行包含文件数据行的INFO和FORMAT列中使用的各种标识符的意义解释，而标题行和数据行包含各样品的变异检测结果信息，格式如下所示：

SNP变异结果信息表

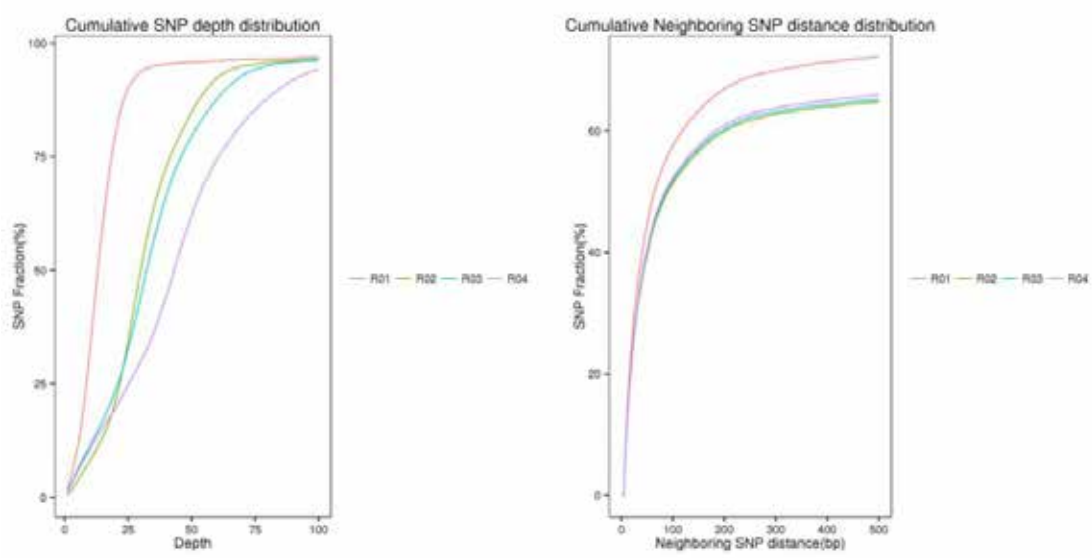
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	P
Chr1	5634	.	G	A	140.84	PASS	[ANNOTATIONS]	GT:AD:DP:GQ:PL	1/1:0,6:6:18:169,18,0
Chr1	30071	.	A	G	141.84	PASS	[ANNOTATIONS]	GT:AD:DP:GQ:PL	1/1:0,6:6:18:170,18,0
Chr1	30478	.	C	T	95.9	PASS	[ANNOTATIONS]	GT:AD:DP:GQ:PL	1/1:0,5:5:15:124,15,0
Chr1	32667	.	A	G	91.03	PASS	[ANNOTATIONS]	GT:AD:DP:GQ:PL	1/1:0,4:4:12:119,12,0

各列意义说明如下：

1	CHROM	Chr1	参考序列的染色体名称
2	POS	5634	参考序列位点坐标
3	ID	.	标识符
4	REF	G	参考序列对应位置碱基
5	ALT	A	SNP位点对应的另外类型的碱基
6	QUAL	140.84	变异位点质量值
7	FILTER	PASS	过滤状态
8	INFO	[ANNOTATIONS]	位点注释信息
9	FORMAT	GT:AD:DP:GQ:PL	基因型信息格式
10	R01	1/1:0,6:6:18:169,18,0	样品的基因型信息

vcf文件的详细说明信息见网页：http://gatkforums.broadinstitute.org/discussion/1268/how-should-i-interpret-vcf-files-produced-by-the-gatk

为了确保样本SNP的可信性，对样本检测的SNP的reads支持数，相邻SNP的距离统计累积分布。



SNP质量分布图

注：左边为 SNP reads 支持数目累积图，右边为相邻 SNP之间的距离累积图。

SNP类型的变异分为转换和颠换两种，同种类型碱基之间突变称为转换（Transition），如嘌呤与嘌呤之间、嘧啶与嘧啶之间的变异，不同类型碱基之间的突变称为颠换（Transversion），如嘌呤与嘧啶之间的变异。一般来说转换比颠换更容易发生，故转换/颠换（Ti/Tv）的比例一般大于1，具体数值和所测物种有关。对于二倍体或者多倍体物种，若同源染色体上的某一SNP位点均为同一种碱基，则该SNP位点称为纯合SNP位点；若同源染色体上的SNP位点包含不同类型的碱基，则该SNP位点称为杂合SNP位点。纯合SNP数量越多，则样品与参考基因组之间差异越大，杂合SNP数量越多，则样品的杂合程度越高，具体结果和样品的材料选择有关。

3.3.2 样品之间 SNP 的检测

根据样品与参考基因组的比对结果，汇总样品之间所有有差异的变异位点，各样品的SNP列表文件格式如下所示：

TIANGEN

各样品SNP列表示意

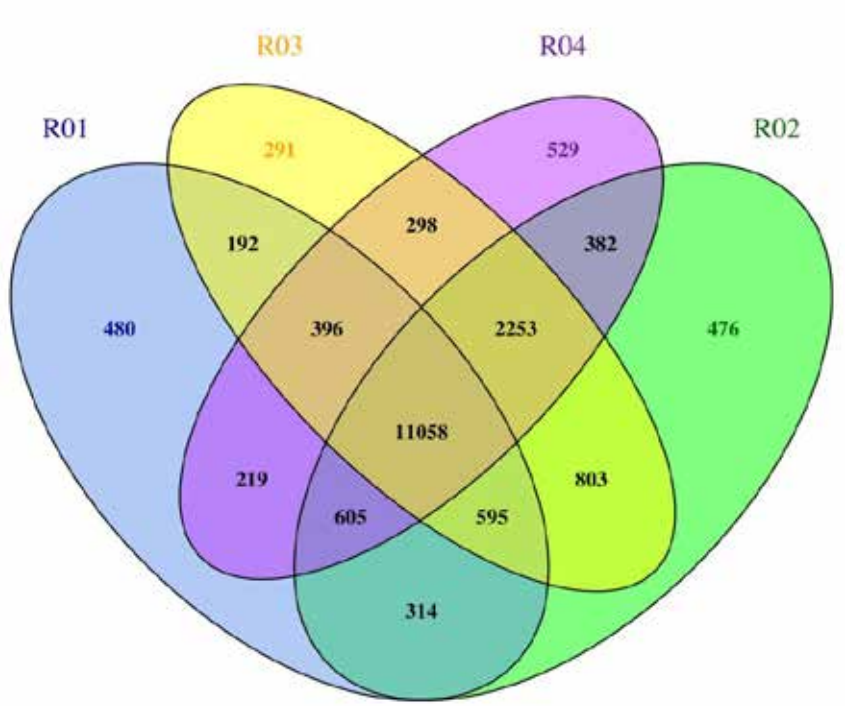
#Chr	Pos	Ref	P	M	B1	B2
chromosome_1	240	C	C	T	C	C
chromosome_1	248	G	G	A	G	G
chromosome_1	422	A	A	T	A	A
chromosome_1	463	C	N	T	C	C
chromosome_1	483	T	N	C	T	T
chromosome_1	631	C	C	T	C	C
chromosome_1	651	T	T	C	T	T

注：Chr：SNP位点所在的染色体名称；Pos：SNP在参考序列的位置；Ref：参考序列的碱基类型；P、M、B1、B2：各样品在该SNP位点对应的碱基类型。

SNP基因型的编码采用标准核苷酸符号，符号表如下所示：

核苷酸代码	意义	核苷酸代码	意义
A	Adenosine	M	A C (aMino group)
C	Cytosine	S	G C (Strong interaction)
G	Guanine	W	A T (Weak interaction)
T	Thymidine	B	G T C (not A) (B comes after A)
U	Uracil	D	G A T (not C) (D comes after C)
R	G A (puRine)	H	A C T (not G) (H comes after G)
Y	T C (pYrimidine)	V	G C A (not T, not U) (V comes after U)
K	G T (Ketone)	N	A G C T (aNy)

样品间SNP的统计结果如下图所示：



样品间SNP统计Venn图

注：变异位点数量venn统计只考虑位置是否相同，不考虑基因型是否相同。

据统计，样品P和M间共有XX个SNP，样品B1和B2间共有XX个SNP。

3.3.3 SNP 结果注释

SnpEff[6]是一款用于注释变异（SNP、Small InDel）和预测变异影响的软件。根据变异位点在参考基因组上的位置以及参考基因组上的基因位置信息，可以得到变异位点在基因组发生的区域（基因间区、基因区或CDS区等），以及变异产生的影响（同义非同义突变等）。

软件可以使用vcf格式文件作为输入和输出。输出结果会在vcf文件的INFO列添加以下字段：

EFF= Effect (Effect_Impact | Functional_Class | Codon_Change | Amino_Acid_Change| Amino_Acid_Length | Gene_Name | Transcript_BioType | Gene_Coding | Transcript_ID | Exon_Rank | Genotype_Number [| ERRORS | WARNINGS])

各标识符说明如下：

类型	意义
Effect	变异所在的区域或类型
Effect impact	变异影响大小（High, Moderate, Low, Modifier）
Functional Class	功能分类（NONE, SILENT, MISSENSE, NONSENSE）
Codon_Change/Distance	编码改变（old_codon/new_codon）或者变异位点到转录本的距离（在基因上下游区域）
Amino_Acid_Change	氨基酸编码改变（原氨基酸类型、位置、改变后氨基酸类型）
Amino_Acid_Length	氨基酸编码蛋白的长度（转录本长度/3）
Gene_Name	基因名
Transcript_BioType	转录本功能
Gene_Coding	编码蛋白(CODING | NON_CODING)
Transcript_ID	转录本ID
Exon/Intron Rank	外显子或内含子位次
Genotype_Number	变异的基因型位次
Warnings/Errors	警告或错误

以上的结果若无法得到，则其对应列为空。具体说明可参见SnpEff的说明文档：

http://snpeff.sourceforge.net/SnpEff_manual.html#output。

本项目样品间的SNP注释具体统计结果如下所示：

SNP注释结果统计

Type		P vs M	B1 vs B2
INTERGENIC			
INTRAGENIC			
INTRON			
UPSTREAM			
DOWNSTREAM			
SPLICE_SITE_ACCEPTOR			
SPLICE_SITE_DONOR			
CDS	START_LOST		
	NON_SYNONYMOUS_START		
	SYNONYMOUS_CODING		
	NON_SYNONYMOUS_CODING		
	SYNONYMOUS_STOP		
	STOP_GAINED		
	STOP_LOST		
Other			

注：Type：SNP所在区域或类型；P、M、B1、B2为各样品相对于参考基因组存在的对应类型的SNP数量，P vs M和B1 vs B2为两个样品间存在的对应类型的SNP数量。

各行意义说明如下表所示：

INTERGENIC	基因间区
INTRAGENIC	基因内（无转录本信息）
INTRON	内含子
UPSTREAM	基因上游区域(5K以内)
DOWNSTREAM	基因下游区域(5K以内)
SPLICE_SITE_ACCEPTOR	剪切供体突变(exon前2bp内)
SPLICE_SITE_DONOR	剪切受体突变(exon后2bp内)
NON_SYNONYMOUS_CODING	非同义编码突变
NON_SYNONYMOUS_START	非同义的起始密码子突变
START_LOST	起始密码子丢失
STOP_GAINED	终止密码子获得
STOP_LOST	终止密码子丢失
SYNONYMOUS_CODING	同义编码突变
SYNONYMOUS_STOP	同义终止密码子突变
Other	由于gff文件中基因信息不完整而无法得到准确的判断

3.4 Small InDel 检测与注释

3.4.1 样品与参考基因组间 Small InDel 的检测

根据样品的Clean Reads在参考基因组上的定位结果，检测样品与参考基因组之间是否存在小片段的插入与缺失（Small InDel：1-5bp）。样品的插入缺失使用GATK检测。Small InDel变异一般比SNP变异少，同样反映了样品与参考基因组之间的差异，并且编码区的InDel会引起移码突变，导致基因功能上的变化。

3.4.2 样品之间 Small InDel 检测

根据样品与参考基因组的Small InDel检测结果，提取样品之间有差异的变异位点，即样品之间的Small InDel变异位点。部分结果如下表所示：

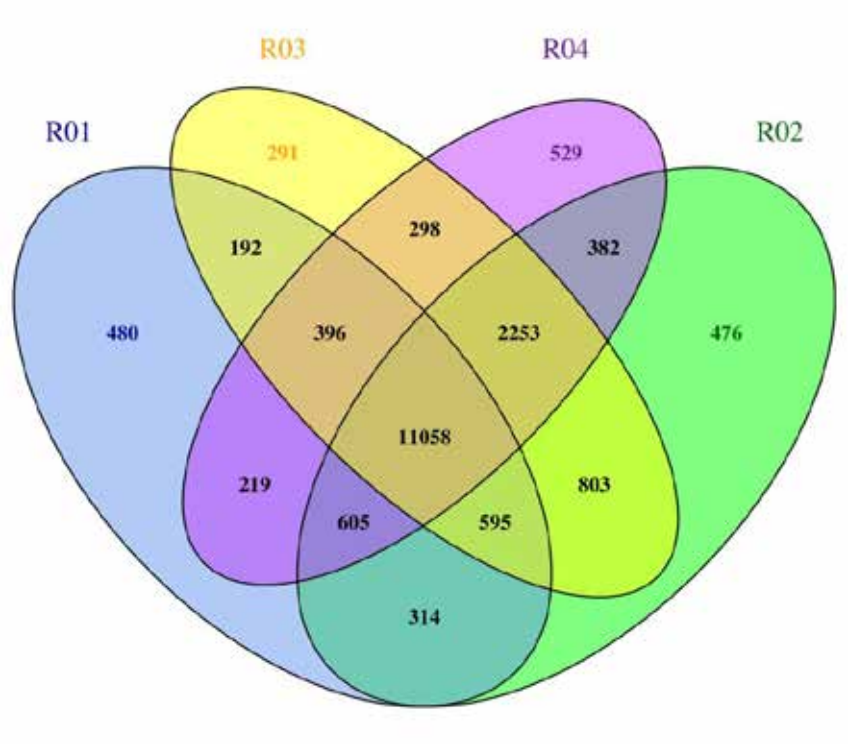
样品测序数据统计

#CHROM	POS	REF	ALT	W	Mut

各列意义说明如下：

列数	标题	示例	意义说明
1	CHROM	LG1	参考序列的染色体名称
2	POS	14890321	参考序列位点坐标
3	REF	C	参考序列对应位置碱基序列
4	ALT	CAT	InDel位点对应的另外类型的碱基序列
5~6	W、Mut	0/0、0/1、1/1、./.	各样品对应的InDel类型（0/0：纯合且与参考基因组一致；0/1：杂合类型；1/1：纯合且与参考基因组不一致；./.：不能确定）

样品间Small InDel的统计结果如下图所示：



样品间Small InDel统计Venn图

注：变异位点数量venn统计只考虑位置是否相同（InDel的起始位置），不考虑基因型是否相同。

据统计，样品P和M间共有XX个InDel，样品B1和B2间共有XX个InDel。

3.4.3 Small InDel 的注释

根据样品检测得到的Small InDel位点在参考基因组上的位置信息，对比参考基因组的基因、CDS位置等信息(一般在gff文件中)，可以注释InDel位点是否发生在基因间区、基因区或CDS区、是否为移码突变等。Small InDel的注释通过SnpEff软件实现。发生移码突变的InDel可能会导致基因功能的改变，具体注释结果见下表：

InDel注释结果统计

Type		P vs M	B1vs B2
INTERGENIC			
INTRAGENIC			
INTRON			
UPSTREAM			
DOWNSTREAM			
SPLICE_SITE_ACCEPTOR			
SPLICE_SITE_DONOR			
CDS	START_LOST		
	FRAME_SHIFT		
	CODON_DELETION		
	CODON_INSERTION		
	EXON_DELETION		
	CODON_CHANGE_PLUS_CODON_DELETION		
	CODON_CHANGE_PLUS_CODON_INSERTION		
	STOP_GAINED		
	STOP_LOST		
Other			

注：Type：InDel所在区域或类型；P、M、B1、B2为各样品相对于参考基因组存在的对应类型的InDel数量，P vs M和B1 vs B2为两个样品间存在的对应类型的InDel数量。

INTERGENIC	基因间区
INTRAGENIC	基因内（无转录本信息）
INTRON	内含子
UPSTREAM	基因上游区域(5K以内)
DOWNSTREAM	基因下游区域(5K以内)
SPLICE_SITE_ACCEPTOR	剪切供体突变(exon前2bp内)
SPLICE_SITE_DONOR	剪切受体突变(exon后2bp内)
CODON_CHANGE_PLUS_CODON_DELETION	非密码子边界上的3的整数倍的删除
CODON_CHANGE_PLUS_CODON_INSERTION	非密码子边界上的3的整数倍的插入
CODON_DELETION	密码子删除(3的整数倍)
CODON_INSERTION	密码子插入(3的整数倍)
EXON_DELETED	整个外显子被删除
FRAME_SHIFT	移码突变(非3的整数倍插入或删除)
START_LOST	起始密码子丢失
STOP_GAINED	终止密码子获得
STOP_LOST	终止密码子丢失
Other	由于gff文件中基因信息不完整而无法得到准确的判断

3.5 关联分析

3.5.1 高质量 SNP 筛选

根据 SNP 检测结果，样品 P 和 M 共筛选到 XX 个 SNP 位点，样品 B1 和 B2 共 XX 个 SNP，在关联分析前，首先对样品 B1 和 B2 间的 XX 个 SNP 进行过滤，过滤标准如下：首先过滤掉有多个基因型的 SNP 位点，其次过滤掉 read 支持度小于 4 的 SNP 位点，再次根据通过亲本的 SNP 信息过滤掉与同表型亲本不同的位点，最终得到高质量的可信 SNP 位点 XX 个。

SNP 过滤统计

	Total SNP	多个等位基因的位点	Read 支持度小于 4 的位点	利用亲本过滤的位点	总过滤位点数	高质量 SNP
B1 vs B2			10704	21550	32254	28005

3.5.2 SNP-index 方法关联结果

SNP-index 是近年来发表的一种通过混池间的基因型频率差异进行标记关联分析的方法[7]，主要是寻找混池之间基因型频率的显著差异，用 Δ(SNP-index)统计。标记 SNP 与性状关联度越强，Δ(SNP-index)越接近于 1。

计算方法简述如下：

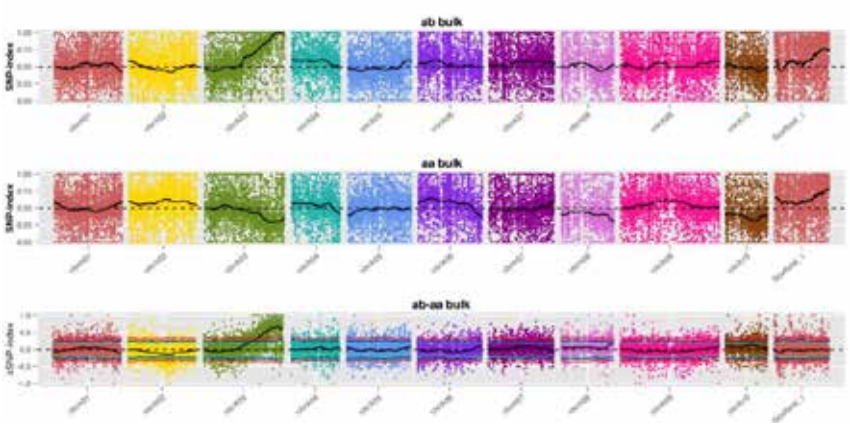
Maa 表示 aa 池来源于母本的深度；Paa 表示 aa 池来源于父本的深度；

Mab 表示 ab 池来源于母本的深度；Pab 表示 ab 池来源于父本的深度；

SNP-index(ab)=Mab/(Pab+Mab)；SNP-index(aa)=Maa/(Paa+Maa)。

Δ(SNP-index)=SNP-index(aa)-SNP-index(ab)。

为了消除假阳性的位点，利用标记在基因组上的位置，可对同一条染色体上标记的 ΔSNP-index 值进行拟合[7]，本项目并采用 DISTANCE 方法对△SNP-index 进行拟合，取每个 SNP 左右距离各 2M 的 SNP 的△SNP-index 的中值作为该位点拟合后的关联值。并根据关联阈值，选择阈值以上的区域作为与性状相关的区域。两个混池分别的 SNP-index 及△SNP-index 的分布如下图所示：



SNP-index 关联值在染色体上的分布

注：横坐标为染色体名称，彩色的点代表计算出来的 SNP-index（或 △SNP-index）值，黑色的线为拟合后的 SNP-index（或 △SNP-index）值。上图是 B1 混池的 SNP-index 值的分布图；中图是 B2 混池的 SNP-index 值的分布图；下图是△SNP-index 值的分布图，其中红色的线代表置信度为 0.99 的阈值线，蓝色的线代表置信度为 0.95 的阈值线，绿色的线代表置信度为 0.90 的阈值线。

根据本项目群体的理论分离比，计算关联阈值为 XX。

根据计算机模拟实验[8]计算结果，当置信度为 0.95 时，定位区域为 158439337-160141497（1.702M）区间内，共得到 XX 个区域，总长度为 XXbp，其中包含非同义突变 SNP 位点的基因共 XX 个，同义突变 SNP 位点的基因共 XX 个。

理论上，目标位点及其附近的连锁位点应趋近于该阈值，因此显著关联的区域附近应该出现一个较高的峰值。但从结果上看，没有超过理论阈值的区域，说明本实验中没有发现显著的定位结果。为了充分利用数据，将阈值降低以寻找比较可能的定位区域，利用拟合后△SNP-index 的 99 百分位数，即 XXX，最终得到可能的定位区域为 158439337-160141497（1.702M），共得到 XX 个区域，总长度为 XXbp，其中包含非同义突变 SNP 位点的基因共 XX 个，同义突变 SNP 位点的基因共 XX 个，移码突变的基因共 XX 个。然而由于未达到理论阈值，这个区域很可能是假阳性区域，需要进一步验证。

关联区域信息统计表

Chromosome ID	Start	End	Size(Mb)	Gene number
Chr02	11,688,755	12,993,319	1.30	148
Total	-	-	1.30	148

注：Chromosome ID：染色体编号；Start：关联区域起始位置；End：关联区域终止位置；Size：关联区域大小，以 Mb 为单位；Gene number：关联区域内的基因数量。

3.5.3 ED 方法关联结果

欧式距离（Euclidean Distance，ED）算法，是利用测序数据寻找混池间存在显著差异标记，并以此评估与性状关联区域的方法[9]。理论上，BSA 项目构建的两个混池间除了目标性状相关位点存在差异，其他位点均趋向于一致，因此非目标位点的 ED 值应趋向于 0。ED 方法的计算公式如下所示，ED 值越大表明该标记在两混池间的差异越大。

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

其中：

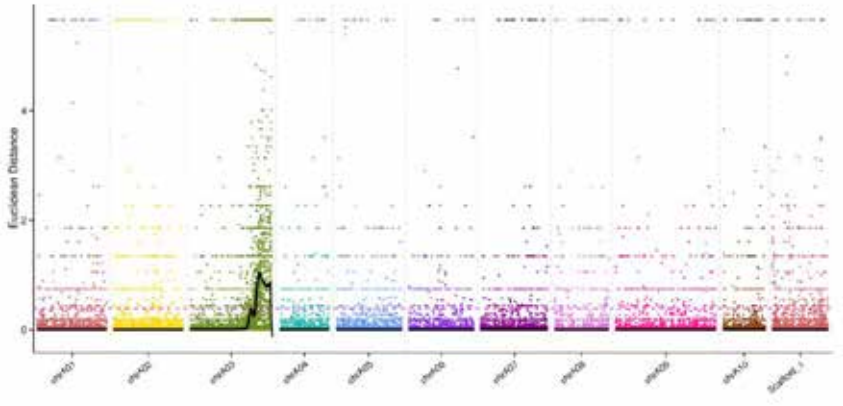
A_{mut} 为 A 碱基在突变混池中的频率，A_{wt} 为 A 碱基在野生型混池中的频率；

C_{mut} 为 C 碱基在突变混池中的频率，C_{wt} 为 C 碱基在野生型混池中的频率；

G_{mut} 为 G 碱基在突变混池中的频率，G_{wt} 为 G 碱基在野生型混池中的频率；

T_{mut} 为 T 碱基在突变混池中的频率，T_{wt} 为 T 碱基在野生型混池中的频率。

在进行分析时，利用两混池间基因型存在差异的 SNP 位点，统计各个碱基在不同混池中的深度，并计算每个位点 ED 值，为消除背景噪音，对原始 ED 值进行乘方处理[9]，本项目取原始 ED 的 4 次方作为关联值以达到消除背景噪音的功能，然后采用局部线性回归 LOESS 方法对 ED 值进行拟合，关联值分布如下图所示：



ED 关联值在染色体上的分布

注：横坐标为染色体名称，彩色的点代表每个 SNP 位点的 ED 值，黑色的线为拟合后的 ED 值，红色的虚线代表显著性关联阈值，ED 值越高，代表该点关联效果越好。

取所有位点拟合值的 median+3SD 作为分析的关联阈值[9]，计算得 XX。

根据关联阈值判定，定位区域为 158439337-160141497（1.702M）区间内。共得到 XX 个区域，总长度为 XXbp，其中包含非同义突变 SNP 位点的基因共 XX 个，同义突变 SNP 位点的基因共 XX 个，移码突变的基因共 XX 个。

理论上，目标位点及其附近的连锁位点应趋近于该阈值，因此显著关联的区域附近应该出现一个较高的峰值。但从结果上看，没有超过理论阈值的区域，说明本实验中没有发现显著的定位结果。为了充分利用数据，将阈值降低以寻找比较可能的定位区域，利用拟合后 ED 的 99 百分位数，即 XXX，最终得到可能的定位区域为 158439337-160141497（1.702M），共得到 XX 个区域，总长度为 XXbp，其中包含非同义突变 SNP 位点的基因共 XX 个，同义突变 SNP 位点的基因共 XX 个，移码突变的基因共 XX 个。然而由于未达到理论阈值，这个区域很可能是假阳性区域，需要进一步验证。

关联区域信息统计表

Chromosome ID	Start	End	Size(Mb)	Gene number
Chr02	11,688,755	12,993,319	1.30	148
Total	-	-	1.30	148

注：Chromosome ID：染色体编号；Start：关联区域起始位置；End：关联区域终止位置；Size：关联区域大小，以 Mb 为单位；Gene number：关联区域内的基因数量。

3.5.4 候选区域筛选

对这两种关联分析方法得到的结果取交集，得到的交集见下表：

关联区域信息统计表

Chromosome ID	Start	End	Size(Mb)	Gene number
Chr02	11,688,755	12,993,319	1.30	148
Total	-	-	1.30	148

注：Chromosome ID：染色体编号；Start：关联区域起始位置；End：关联区域终止位置；Size：关联区域大小，以 Mb 为单位；Gene number：关联区域内的基因数量。

3.6 候选区域的功能注释

3.6.1 候选区域的 SNP 注释

本项目样品间在候选区域内的SNP注释结果见下表：

候选区域内 SNP 注释结果统计

Type		P vs M	B1 vs B2
INTERGENIC			
INTRAGENIC			
INTRON			
UPSTREAM			
DOWNSTREAM			
SPLICE_SITE_ACCEPTOR			
SPLICE_SITE_DONOR			
CDS	START_LOST		
	NON_SYNONYMOUS_START		
	SYNONYMOUS_CODING		
	NON_SYNONYMOUS_CODING		
	SYNONYMOUS_STOP		
	STOP_GAINED		
	STOP_LOST		
Other			

注：Type：SNP 所在区域或类型；P vs M 和 B1 vs B2 为两个样品间在关联区域内存在的对应类型的 SNP 数量。

据统计，亲本间存在非同义突变的 SNP 共 XX 个，这些 SNP 很有可能与性状直接相关，这些 SNP 所在的基因我们称之为非同义突变基因，共 XX 个；混池间存在非同义突变的 SNP 共 XX 个，非同义突变基因共 XX 个。

3.6.2 候选区域的基因注释

应用 BLAST[10]软件对候选区间内的编码基因进行多个数据库（NR[11]、Swiss-Prot、GO[12]、KEGG[13]、COG[14]）的深度注释。通过详细的注释，快速筛选候选基因。候选区域内共注释到 39 个基因，其中在亲本间存在非同义突变基因共注释到 XX 个，注释结果见下表：

候选区域内基因功能注释结果统计

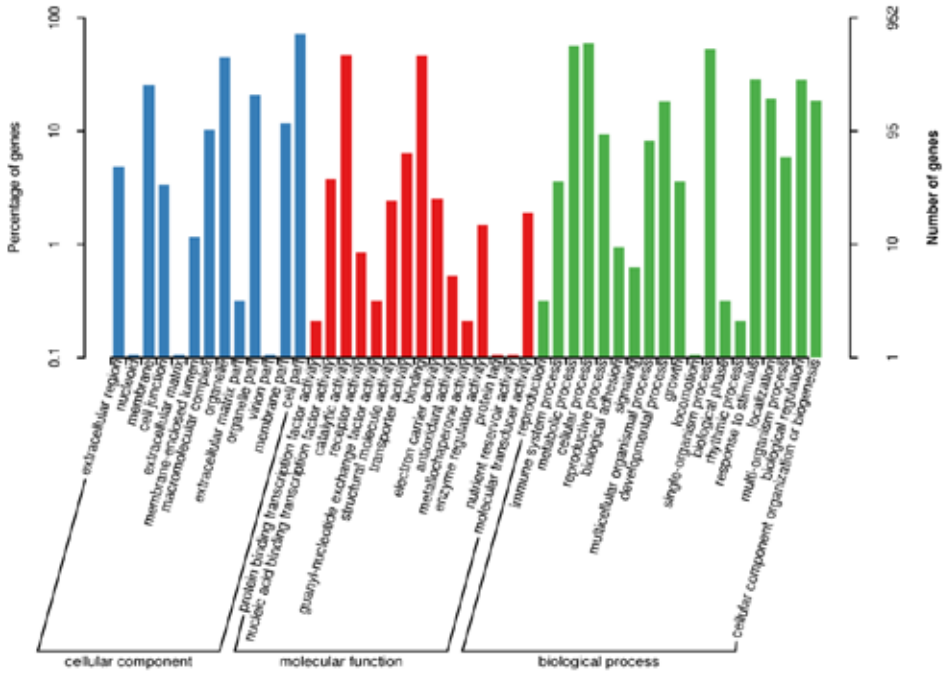
Annotated databases	Gene Num	Non_Syn Gene Num
NR	33	
Swiss-Prot	22	
GO	14	
KEGG	16	
COG	21	
Total	39	

注：Annotated databases：功能注释数据库；Gene Number：在相应数据库有注释信息的候选区域基因数；Non_Syn Gene Num：候选区域内亲本间存在非同义突变的基因数。

3.6.2.1 候选区域内基因的 GO 富集分析

GO数据库是一个结构化的标准生物学注释系统，建立了基因及其产物功能的标准词汇体系，适用于各个物种。该数据库结构分为多个层级，层级越低，节点所代表的功能越具体。通过GO分析并按照Cellular component、Molecular Function、Biological process对基因进行分类。

候选区域内基因 GO 分类统计结果见下图：

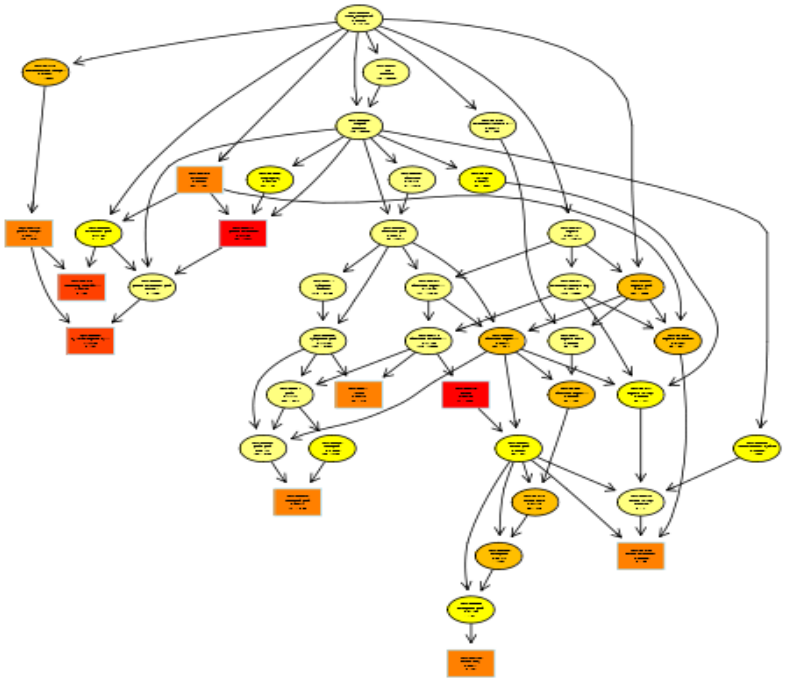


候选区域内基因 GO 注释聚类图

注：横坐标为 GO 各分类内容，纵坐标左边为基因数目所占百分比，右边为基因数目。此图展示的是关联区域内所有基因背景下 GO 各二级功能的基因分类情况。

topGO 有向无环图能直观展示关联区域内基因富集的 GO term 及其层级关系。有向无环图为关联区域内基因 GO 富集分析结果的图形化展示方式，分支代表包含关系，从上至下所定义的功能范围越来越具体。

候选区域内基因的 Cellular component 的 topGO 有向无环图如下：



候选区域内基因的 Cellular componenttopGO 有向无环图分析

注：对每个 GO 节点进行富集，最显著的 10 个节点在图中用方框表示，图中还包含其各层对应关系。每个方框（或椭圆）内给出了该 GO 节点的内容描述和富集显著性值。不同颜色代表不同的富集显著性，颜色越深，显著性越高。

候选区域基因 GO 的富集分析结果见下表：

候选区域内基因的 topGO 富集结果示意表（Cellular component）

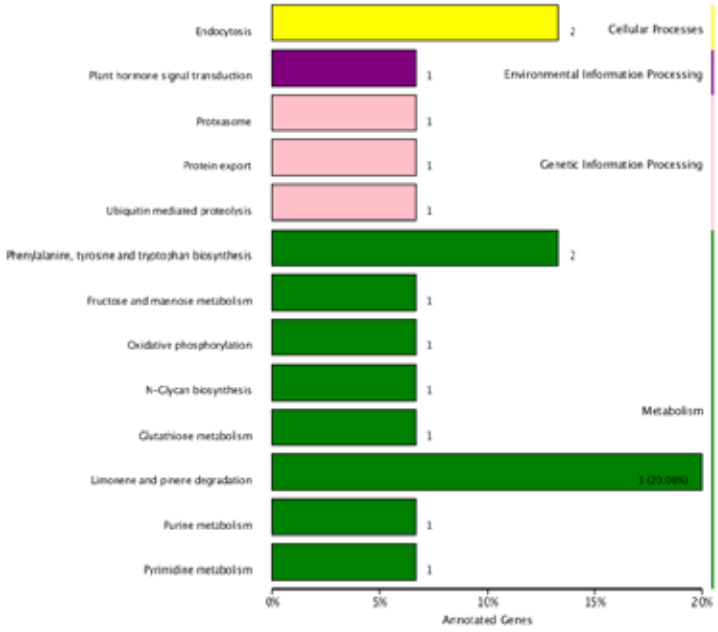
GO.ID	Term	Annotated	Significant	Expected	KS
GO:0005524	ATP binding	2164	54	41.06	2.0e-05
GO:0031418	L-ascorbic acid binding	27	4	0.51	2.5e-05
GO:0005515	protein binding	1076	31	20.42	3.2e-05
GO:0019904	protein domain specific binding	11	0	0.21	0.00038
GO:0045486	naringenin 3-dioxygenase activity	7	4	0.13	0.00043
GO:0016874	ligase activity	472	7	8.96	0.00048
GO:0016853	isomerase activity	238	1	4.52	0.00097
GO:0004478	methionine adenosyltransferase activity	9	0	0.17	0.00158

注：GO.ID：GO 节点的编号；Term：GO 节点名称；Annotated：所有基因注释到该功能的基因数；Significant：DEG 注释到该功能的基因数；Expected：注释到该功能 DEG 数目的期望值；KS：富集节点的显著性统计，KS 值越小，表明富集越显著。

3.6.2.2 候选区域内基因的 KEGG 富集分析

在生物体内，不同基因相互协调来行使生物学功能，不同的基因间相同的作用通路为一个Pathway，基于Pathway分析有助于进一步解读基因的功能。KEGG是关于Pathway的主要公共数据库。

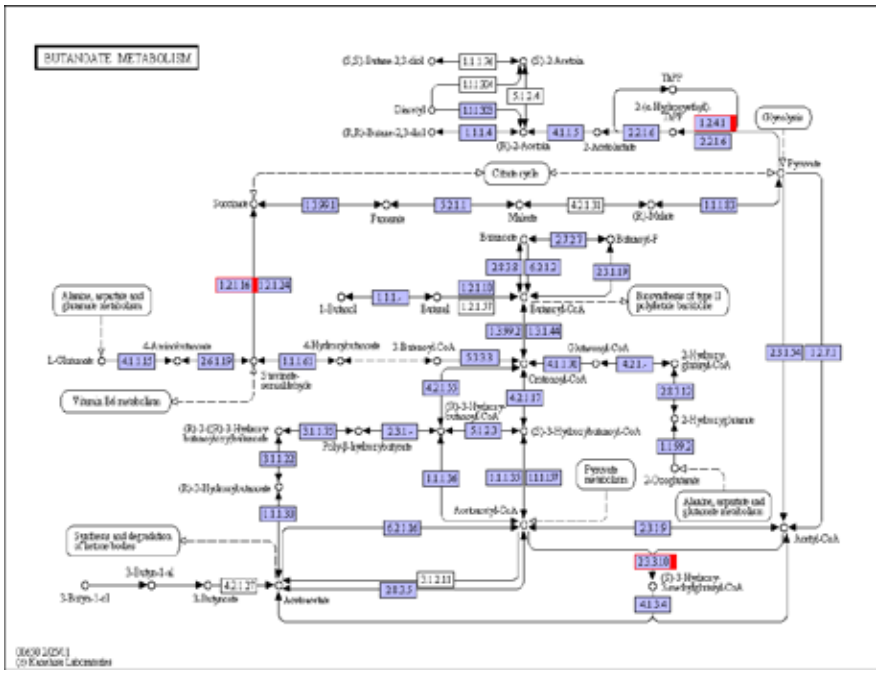
候选区域内基因的KEGG注释结果按照通路类型进行分类，分类图如下：



候选区域内基因的通路分布图

注：横坐标为注释到该通路下的基因个数及其个数占被注释上的基因总数的，纵坐标为 KEGG 代谢通路的名称。

候选区域内基因的代谢通路结果见下图：



候选区域的最优富集通路图

注：红色框标记的为关联区域内的基因，蓝色框代表该通路所需要的所有的酶，说明对应基因与此酶相关，而整个通路是有很多种不同的酶经过复杂的生化反应形成的。关联区域内基因中与此通路相关的均用红色框标

出。

候选区域 KEGG 的富集分析结果见下表：

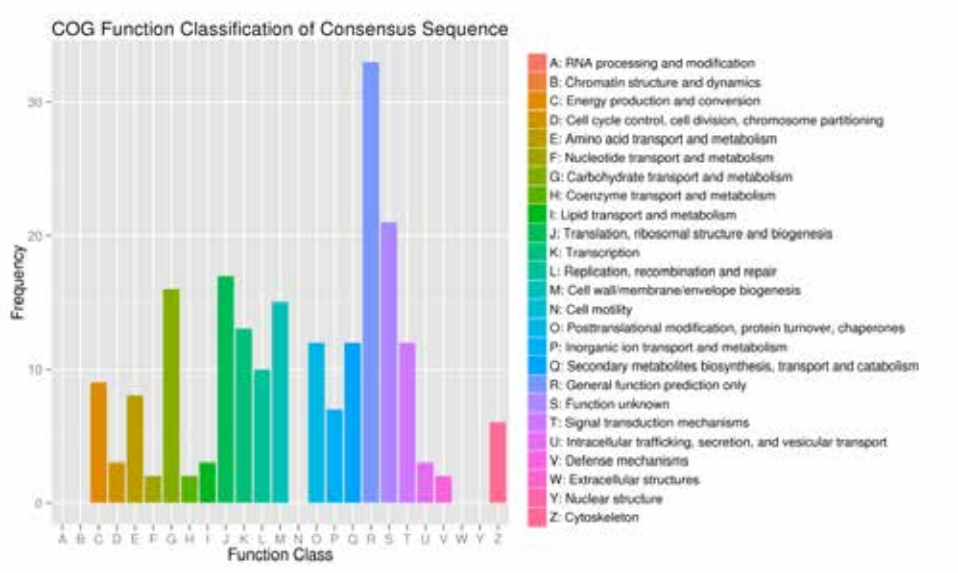
候选区域内基因的 KEGG 富集部分结果

Pathway	KO	Enrichment_Factor	Q_value
Protein processing in endoplasmic reticulum	ko04141	0.20	7.6829e-07
Flavonoid biosynthesis	ko00941	0.14	1.1728e-01
Plant-pathogen interaction	ko04626	0.34	3.0079e-01
Limonene and pinene degradation	ko00903	0.15	5.4617e-01
Porphyrin and chlorophyll metabolism	ko00860	0.22	6.3434e-01
Endocytosis	ko04144	0.35	9.0170e-01
Tyrosine metabolism	ko00350	0.19	1.0000e+00
Pyruvate metabolism	ko00620	0.37	1.0000e+00

注：Pathway：KEGG 通路名称；KO：KEGG 通路编号；Enrichment_Factor：富集因子；Q_value：富集的显著性统计，值越小，富集程度越高。

3.6.2.3 候选区域内基因 COG 分类统计

COG 数据库是基于细菌、藻类、真核生物的系统进化关系构建而成，利用 COG 数据库可以对基因产物进行直系同源分类。关联区域内基因 COG 分类统计结果见下图：

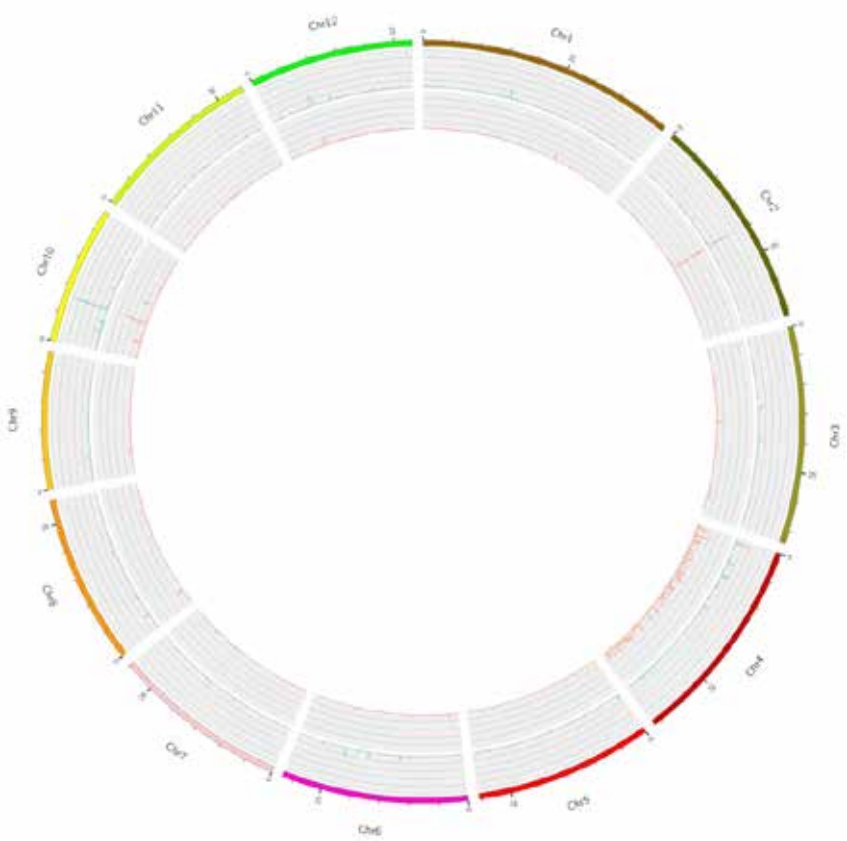


候选区域内基因 COG 注释分类图

注：横坐标为 COG 各分类内容，纵坐标为基因数目。在不同的功能类中，基因所占比例多少反映对应时期和环境下代谢或者生理偏向等内容，可以结合研究对象在各个功能类的分布作出科学的解释。

3.7 结果可视化

结合上述内容，将样品的变异结果及BSA关联分析结果使用circos软件作图，软件网址：http://circos.ca/。亲本间的结果可视化在染色体上的分布见下图：



亲本间结果可视化在染色体上的分布

注：从外到里依次为：第一圈：染色体坐标，第二圈：基因分布，第三圈：SNP、InDel密度分布（红色为SNP分布、蓝色为InDel分布），第四圈：△SNP-index值分布，第五圈：ED值分布。

4 数据下载

4.1 结果文件查看说明

(1) 上传目录中有Readme.txt说明，详细介绍了每个文件所代表的内容。上传的结果数据文件多以文本格式为主(fa文件、txt文件，detail文件，xls文件等)。在Windows系统下查看文件，推荐使用Editplus或 UltraEdit作为文本浏览程序，否则会因文件过大造成死机。在Unix或Linux系统下可以浏览较大的文本文件，用Less等操作命令可以顺利地查看。

(2) 报告文件含有SVG格式的图片文件，SVG是矢量化的图片文件，可以随意放大而不失真。要查看SVG格式的文件，请先安装SVG插件。

【参考文献】

1. Sequencing ProjectInternational Rice Genome. The map-based sequence of the rice genome. Nature436, 793–800 (2005).

2.Li H,Durbin R.Fast and accurate short read alignment with Burrows-Wheeler Transform.Bioinformatics, 2009 25:1754-60

3.McKenna A, Hanna M, Banks E, Sivachenko A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 20:1297-303

4.Picard: http://sourceforge.net/projects/picard/ .(Picard)

5.Joke Reumers, Peter De Rijk, et al. Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. Nature Biotechnology30,61–68(2012)doi:10.1038/nbt.2053

6.Cingolani P, Platts A, Wang le L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.", Fly (Austin). 2012 Apr-Jun;6(2):80-92

7. Fekih, R. et al. MutMap+: Genetic Mapping and Mutant Identification without Crossing in Rice. PLoS One8, 1–10 (2013).

8. Takagi, H. et al. QTL-seq: Rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. Plant J.74, 174–183 (2013).

9. Hill, J. T. et al. MMAPPR: Mutation Mapping Analysis Pipeline for Pooled RNA-seq. Genome Res.23, 687–697 (2013).

10.Stephen F. Altschul, Thomas L. Madden, Alejandro A. Sch?ffer, et al. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. Nucleic Acids Res, 1997 25(17): 3389

11.Yangyang DENG, Jianqi LI, Songfeng WU, et al. Integrated nr Database in Protein

Annotation System and Its Localization. Computer Engineering, 2006 32(5):71-74

12.Michael Ashburner, Catherine A. Ball, Judith A. Blake, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, (25): 25-29

13.Roman L.Tatusov, Michael Y.Galperin, Darren A.Natale, et al. The COG database: a tool for genome_scale analysis of protein functions and evolution. Nucleic Acids Res, 2000_7_1 28(1):33_6

14.Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, et al. The KEGG resource for deciphering the genome. Nucleic Acids Res 2004, (32):D277_D280