利用 MGI 平台对大豆进行全基因组重测序分析

许亚茹

（中国农业科学院 作物科学研究所 重大平台中心，北京 100089）

摘要：本研究建立了 MGI 平台全基因重测序的方法。MGI 平台对大豆的全基因进行重测序结果显示，测序数据质量良好，且与参考基因组比对率较高，符合后续分析要求，对其进行 SNP 和 Indel 的变异检测和注释，此结果说明今后可利用 MGI 平台对其它样品进行全基因重测序分析。

关键词：MGI 平台；全基因重测序

Whole genome resequencing analysis of soybeans using the MGI platform

Xu yaru

Abstract: In this study, a method for whole gene resequencing on the MGI platform was established. The results of resequencing the whole genes of soybean by MGI platform showed that the sequencing data was of good quality and had a high comparison rate with the reference genome, which met the requirements of subsequent analysis, and the variation detection and annotation of SNP and Indel were carried out, which indicated that the MGI platform could be used to perform whole gene resequencing analysis on other samples in the future.

Keywords: MGI platform; Whole gene resequencing

1 研究背景

大豆是重要的粮食作物和油料作物，也是人类最主要的植物蛋白来源[1]。我国是野生大豆的发源地，有着极其丰富的大豆种质资源基础，但是育种和产量较其他大豆主产国显得略有不足，究其原因是我国对大豆的研究和发掘力度存在不足，因此，对大豆育成品种的改良势在必行。自 2010 年起，大豆群体水平的重测序也全面开展，在大豆的全基因组变异图谱上也得到了一定的研究进展[2]。本研究利用 MGI 平台对大豆全基因组进行重测序分析，挖掘全基因组水平上的突变。

中国仪器仪表学会

2 实验仪器

主要实验仪器：MGISP-960、MGIDL-T7、DNBSEQ-T7

3 实验结果

3.1 测序数据质量

根据 MGI 平台的测序特点，使用双端测序的数据，要求 Q30 平均比例在 85%以上，可以看出大豆重测序数据 Q30 平均比例在 94.72%以上，说明大豆测序数据质量良好，满足分析要求。

表 1　测序数据统计表

Samples	ID	Clean reads	Clean bases	GC Content	%≥Q20	%≥Q30
P117	P117	169494922	25424238300	36.18%	98.49%	95.27%
P118	P118	166483906	24972585900	36.47%	98.61%	95.70%
P119	P119	186127112	27919066800	35.89%	98.57%	95.61%
P120	P120	192397276	28859591400	36.46%	98.22%	94.72%
P198	P198	141636468	21245470200	37.11%	98.67%	95.84%
P199	P199	169468714	25420307100	36.55%	98.60%	95.66%
P200	P200	155078286	23261742900	37.90%	98.77%	96.14%

样品原始数据碱基质量值可由图 1 看出不存在异常碱基，6 个大豆碱基测序错误率分布均如图 1。

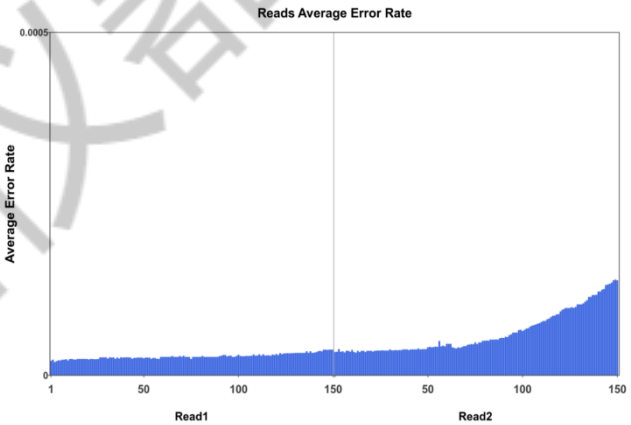

图 1　碱基测序错误率分布图

碱基类型分布检查可用于检测有无 AT、GC 分离现象，若有碱基分离现象可能是测序或

建库所带来的，并会影响后续分析。高通量所测序为基因组随即打断后的 DNA 片段，由于位点在基因组上的分布是近似均匀的，同时，G/C、A/T 含量也是近似均匀的。因此，根据大数定理，在每个测序循环上，GC、AT 含量应当分别相等，且等于基因组的 GC、AT 含量。同样因为重叠等的关系会导致样品前几个碱基 AT、GC 不等波动较大，高于其他测序区段，而其它区段的 GC、AT 含量相等，且分布均匀无分离现象，如图 2 所示。

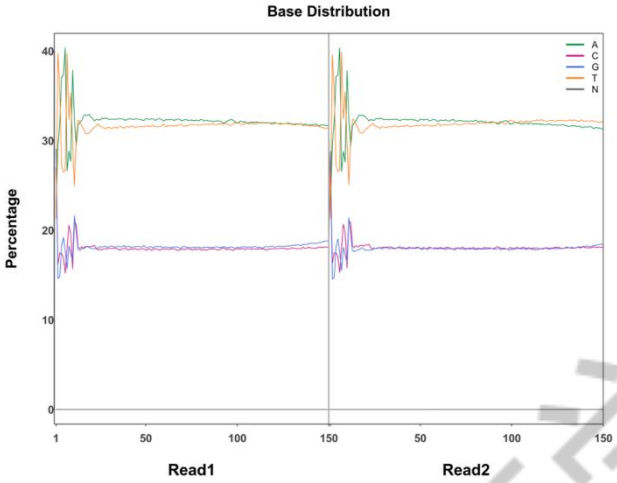

图 2　ATGC 含量分布图

3.2 与参考基因组的序列比对

3.2.1 比对结果

将测序得到的大豆样品与参考基因进行序列比对，bwa 软件主要用于二代高通量测序得到的短序列与参考基因组进行比对，比对结果见表 2，根据比对结果可评估测序数据是否满足后续分析。

表 2　比对效率统计表

Sample_ID	Mapped(%)	Properly_mapped(%)	Averge_depth
P117	99.99%	98.53%	25.44
P118	99.99%	98.55%	24.9
P119	99.99%	98.63%	27.75
P120	99.98%	98.28%	28.58
P198	99.99%	98.58%	21.26
P199	99.98%	98.50%	25
P200	99.99%	98.13%	23.13

将比对到不同染色体的 Reads 进行位置分布统计，绘制 Mapped Reads 在参考基因组上

的覆盖深度分布图，见图 3。

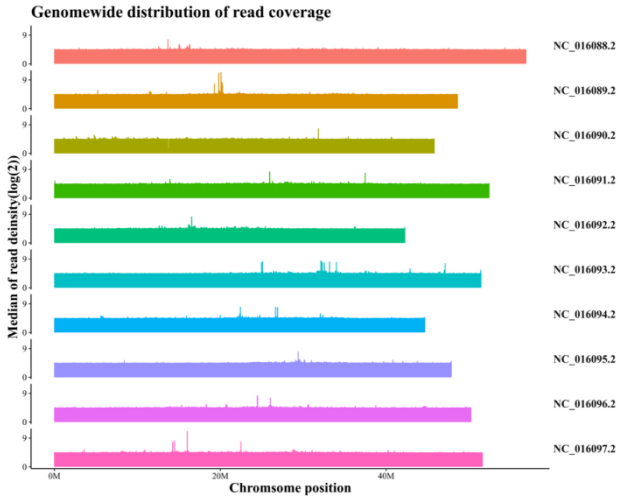

图 3　Mapped Reads 在参考基因组上的位置及覆盖深度分布图

统计 Mapped Reads 在指定的参考基因组不同区域的数目，绘制基因组不同区域样品 Mapped Reads 的分布图，见图 4

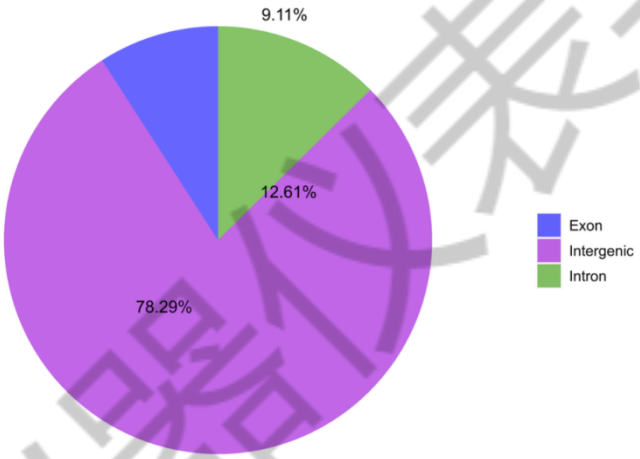

图 4　基因组不同区域 Reads 分布图

3.2.2 插入片段长度检验

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），它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。各样品的插入片段长度模拟分布图见图 5。

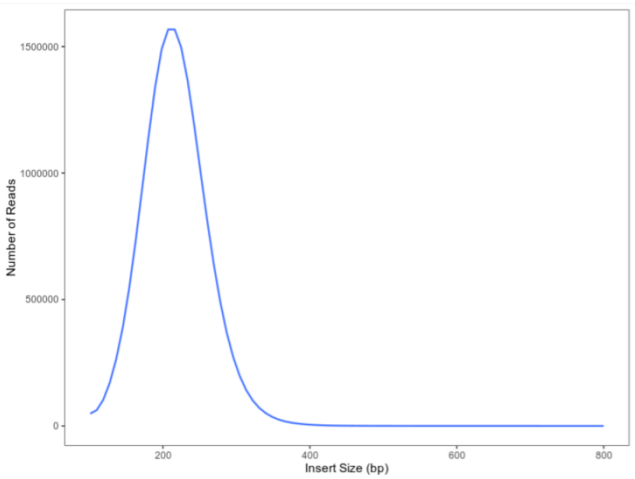

图 5　插入片段长度模拟图

3.2.3 深度分布统计图

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图 6。

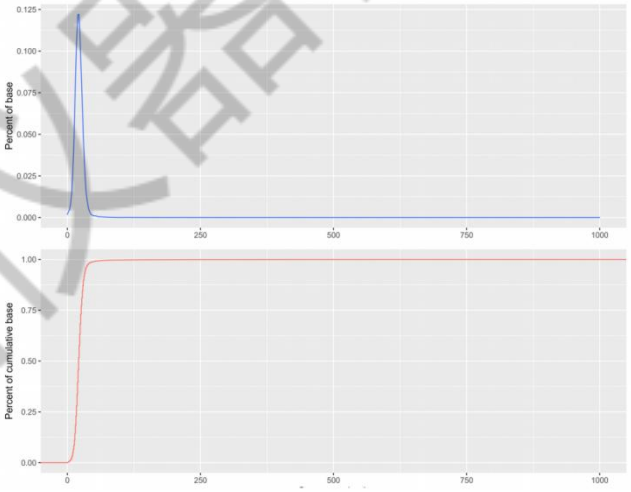

图 6　深度分布统计图

3.3 变异检测

3.3.1 SNP 检测与注释

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，见图 7 和图 8。

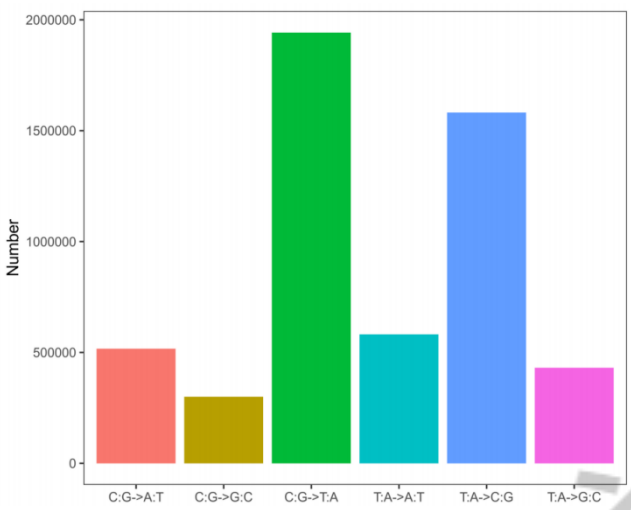

图 7　SNP 突变类型分布图

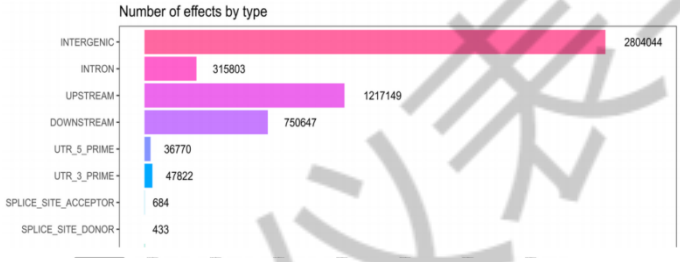

图 8　SNP 注释分类图

3.3.2 Indel 检测与注释

根据所有样品在 CDS 区和全基因范围的 Indel 长度进行统计，其长度分布如图 9。

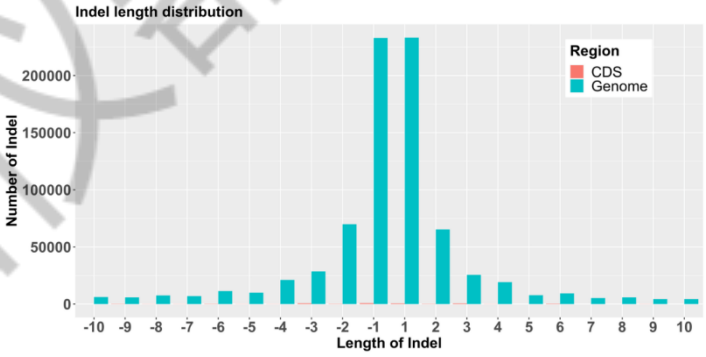

图 9　全基因和编码区 Indel 长度分布图

根据样品检测得到的 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，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。发生移码突变的 Indel 可能会导致基因功能的改变，具体注释结果见图 10。

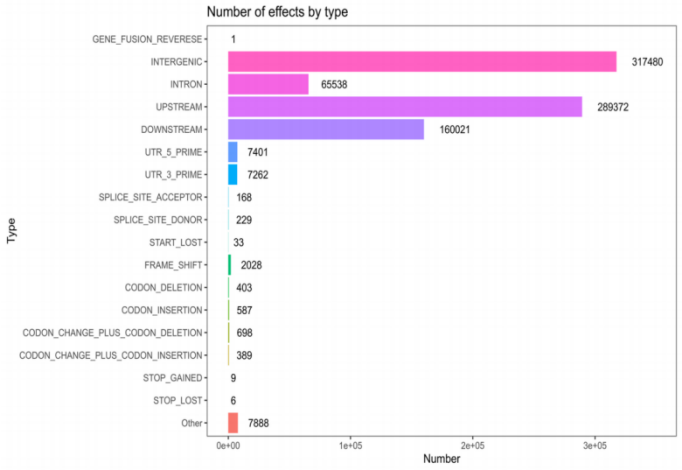

图 10　Indel 注释分类图

4 结论

本文基于 MGI 对大豆进行重基因测序，实验结果可看出，大豆样品测序产出数据良好，与参考基因组序列比对率较高，符合后续分析，对其进行变异检测可得到 SNP 和 Indel 的结果。其它研究表明 MGISEQ-2000 全基因组重测序表现性能稳定、质量可靠，在实际应用上有明显的优势和应用价值[3]。对本次实验说明 MGI 平台对样品进行重测序效果良好，后续可对其它植物进行重测序。

参考文献：

[1]张永芳,钱肖娜,王润梅,等. 不同大豆材料的抗旱性鉴定及耐旱品种筛选[J].作物杂志,2019(5): 41-45.

[2]邬启帆. 基于基因组重测序黄淮海大豆育成品种遗传结构及重要家族遗传基础研究[D]. 南昌大学, 2023.

[3]李伟宁,刘刚,周荣等. MGISEQ-2000、HiSeq 2000 与 NovaSeq 6000 平台全基因组重测序数据的比较分析[J]. 中国畜牧杂志,2021,57(11):156-162.