WES 检测报告

2018-08-05

Pan

一、 样本信息

本次分析使用数据为 KPGP（韩国人基因组计划） 编号 KPGP-00245。

表 1-1 样本信息

KPGP 编号	类型	上传时间	性别	检测平台
KPGP-00245	外显子	2015-03-10	男	Illumina

二、 数据质量统计

2.1 测序数据情况汇总

使用 ReSeqTools（He W , et al.）进行统计。

表 2-1 测序情况汇总

Sample Name	Raw reads	Raw data(G)	Depth(x)	Q20(%)	Q30(%)	GC(%)
KPGP-00245	36924736	5.8	63.0	97.35	93.70	48.08

注:

（1） Sample Name：对应样本的编号;

（2） Raw reads：原始序列双端 reads pair 总数（reads 对数）;

（3） Raw data(G)：样本数据量;

（4） Depth：比对后，测序覆盖区域的平均深度;

（5） Q20(%)：计算 phred 数值大于 20 的碱基占总碱基数的比例，数据值为 read1 和 read2 的平均值;

（6） Q30(%)：计算 phred 数值大于 30 的碱基占总碱基数的比例，数据值为 read1 和 read2 的平均值;

（7） GC(%)：计算 G 和 C 的数量占总碱基数的比例，数据值为 read1 和 read2 的平均值。

2.2 测序质量分布图

使用 FastQC（Andrews S, et al.）以及 MultiQC（Ewels P , et al.）统计。

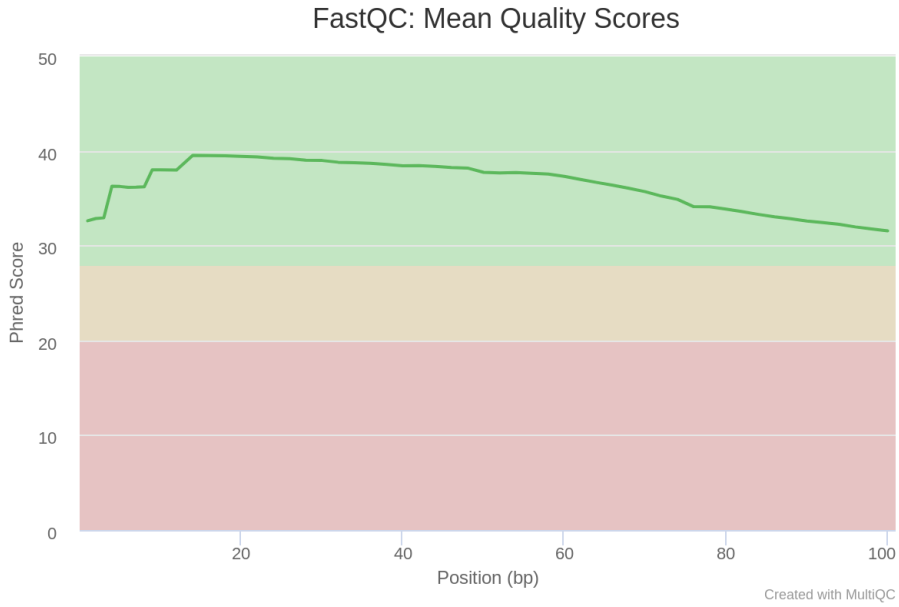

图 2-1 测序质量分布

横坐标为碱基在 reads 中的位置，纵坐标为 phred 得分。

2.3 GC 含量图

使用 FastQC（Andrews S, et al.）以及 MultiQC（Ewels P , et al.）统计。

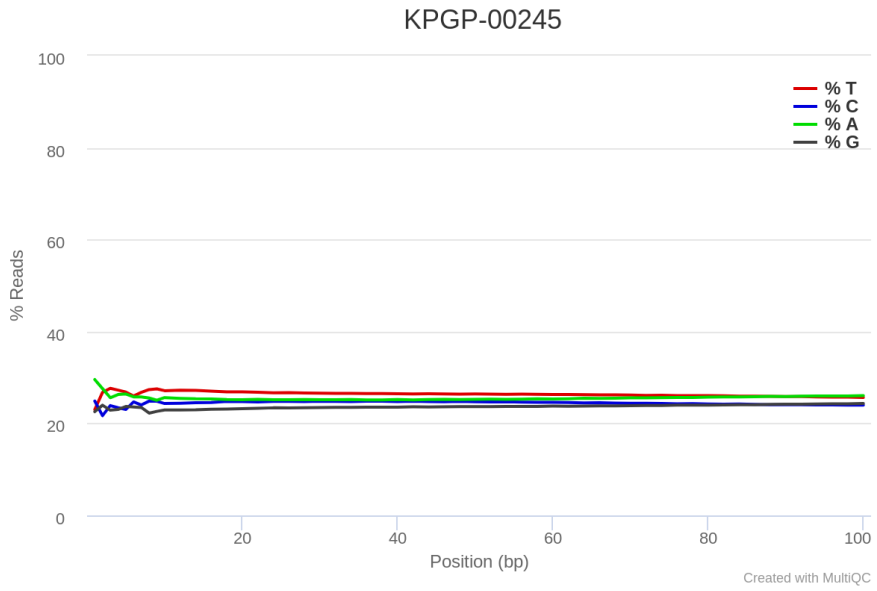

图 2-2 GC 含量图

横坐标为碱基在 reads 中的位置，纵坐标为碱基含量。

2.4 测序深度统计

将测序数据通过 BWA（Li H and Durbin R）与参考基因组（human_g1k_v37_decoy.fasta）进行比对，比对结果使用 samtools（Li H）进行排序，再使用 GATK4 复序列进行标记以及质量校正。将得到的结果使用 QualiMap（García-Alcalde F , et al.）以及 MultiQC（Ewels P , et al.）统计。

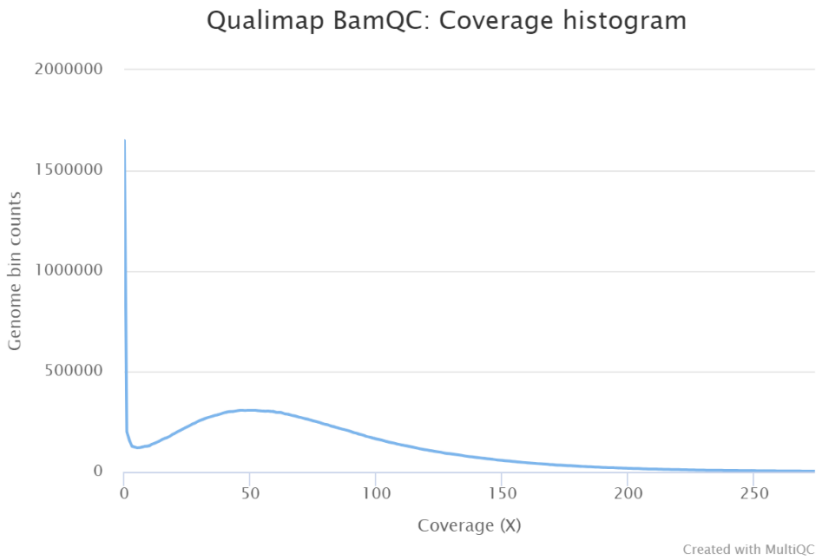

图 2-3 Coverage 图

横坐标为统计平均深度，纵坐标为对应深度得到的 reads 数。

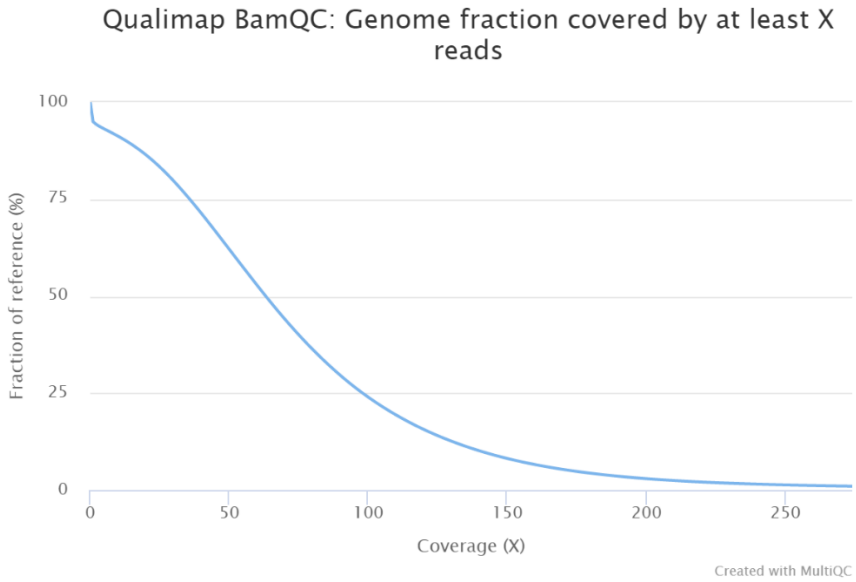

图 2-4 达到深度比例图

横坐标为统计深度，纵坐标为达到对应深度的百分比。

三、 变异检测结果

3.1 SNV、InDel 检测与统计

在比对结果的基础上，使用 GATK4 进行 SNP 位点识别，并使用 Annovar（Kai Wang , et al.）进行结果注释。

表 3-1 SNV、InDel 统计

Sample	exonic	intronic	UTR3	UTR5	intergenic	ncRNA_exonic	ncRNA_intronic	upstream	downstream	splicing	ncRNA_splicing
KPGP-00245	21638	52541	23134	3748	24851	5262	7141	1930	1623	143	29

注：

（1） Sample：样本名称；
（2） exonic：外显子编码区域；
（3） intronic：内含子区域；
（4） UTR3：3'UTR 区域；
（5） UTR5：5'UTR 区域；
（6） intergenic：基因间区域；
（7） ncRNA_exonic：非编码 RNA 外显子区域；
（8） ncRNA_intronic：非编码 RNA 内含子区域；
（9） upstream：转录起始位点上游 1Kb 区域；
（10） downstream：转录起始位点下游 1Kb 区域；
（11） splicing：剪切区域；
（12） ncRNA_splicing：非编码 RNA 剪切区域。

3.2 不同类型变异位点数统计

表 3-2 编码区上不同类型变异位点数

Sample	synonymous	missense	stopgain	stoploss	unknown
KPGP-00245	10785	9944	96	10	437

注：

（1） Sample：样本名称；
（2） synonymous：同义突变；
（3） missense：错义突变；
（4） stopgain：同一碱基发生替换，导致该碱基所在密码子变为终止密码子；
（5） stoploss：同一碱基发生替换，导致该碱基所在终止密码子变为非终止密码子；
（6） unknown：未知功能位点。

3.3 基因型分布统计

表 3-3 基因型分布

Sample	all	het	hom	novel	novel%
KPGP-00245	142194	79298	62896	8147	5.73

注：

（1） Sample：样本名称；

（2） all：所有变异数目；

（3） het：杂合基因型数目；

（4） hom：纯合基因型数目；

（5） novel：未被 dbsnp 注释的变异数目，使用的 dbsnp 版本为 v150（2017-04-04）；

（6） novel%：未被 dbsnp 注释的变异数目占所有变异数目的百分比。

四、 高级分析结果

对位点进行过滤分析。

4.1 Clinvar 提示致病或可能致病的位点

筛选结果共 64 个，此处列出前 5 个。

表 4-1 Clinvar 提示致病的位点

Sample	Chr	Start	End	Ref	Alt	Gene	Info
KPGP-00245	1	152277475	152277475	G	T	FLG	Het
KPGP-00245	3	185237074	185237074	G	T	LIPH	Het
KPGP-00245	5	150227998	150227998	C	T	IRGM	Hom
KPGP-00245	8	11606312	11606312	T	C	GATA4	Hom
KPGP-00245	8	11617240	11617240	A	T	GATA4	Het

4.2 基于低频的筛选

此次筛选选择 exonic 以及 splicing 区域，去除 clinvar 报道良性的位点，去除同义突变位点，再根据 1000g2015_all 选择突变频率小于 0.01 的位点。得到结果 416 个，此处列出前 5 个。

表 4-2 基于低频的筛选

Sample	Chr	Start	End	Ref	Alt	Gene	Info
KPGP-00245	1	1354551	1354551	C	A	ANKRD65	Het
KPGP-00245	1	3800118	3800118	A	G	DFFB	Het
KPGP-00245	1	13414080	13414080	A	C	PRAMEF10;PRAMEF33	Het
KPGP-00245	1	13414081	13414081	G	A	PRAMEF10;PRAMEF33	Het
KPGP-00245	1	13414152	13414152	A	T	PRAMEF10;PRAMEF33	Het

4.3 基于预测软件的筛选

对上述低频结果的软件注释进行分析，选择其中 REVEL（大于 0.45）以及 MCAP（大于 0.025）预测软件认为有害的位点。得到结果 100 个，此处列出前 5 个。

表 4-3 基于预测软件的筛选

Sample	Chr	Start	End	Ref	Alt	Gene	Info
KPGP-00245	2	242035520	242035520	C	G	MTERF4	Het
KPGP-00245	9	88961359	88961359	C	T	ZCCHC6	Het
KPGP-00245	12	6933705	6933705	G	A	GPR162	Het
KPGP-00245	11	6291558	6291558	G	A	CCKBR	Het
KPGP-00245	16	88729481	88729481	G	C	MVD	Het

4.4 基于临床信息的筛选

可以进一步对上述结果进行基于临床信息的筛选。由于数据来源没有提供临床信息，因此无法进行此步骤。

4.5 富集分析

对候选基因进行富集分析。由于未有临床信息，候选基因不能明确寻找，这里使用基于预测软件筛选的基因进行分析。

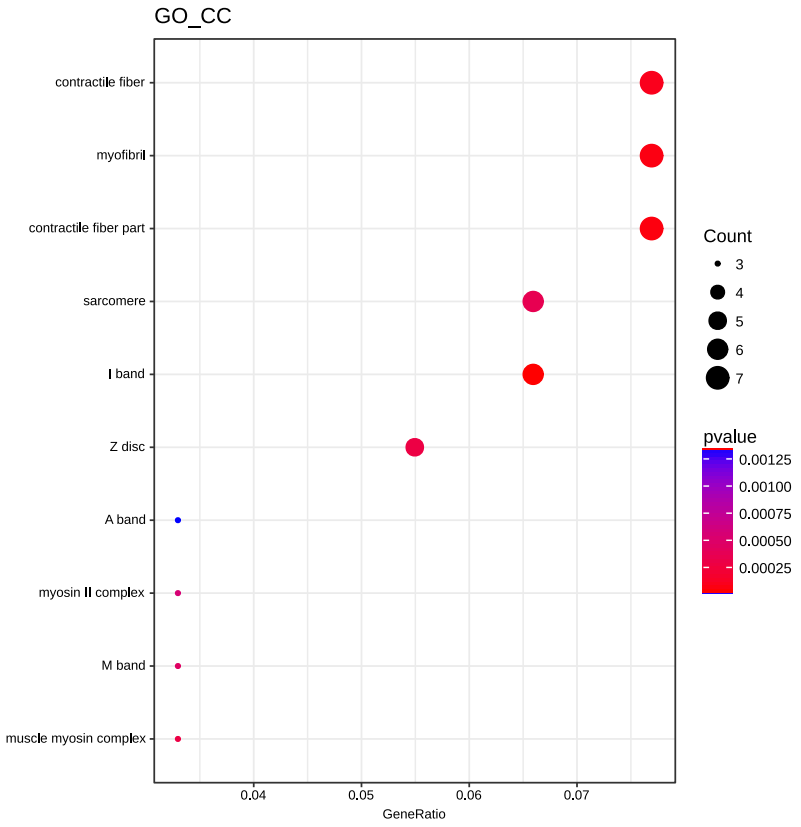

图 4-1 GO 细胞组件富集散点图

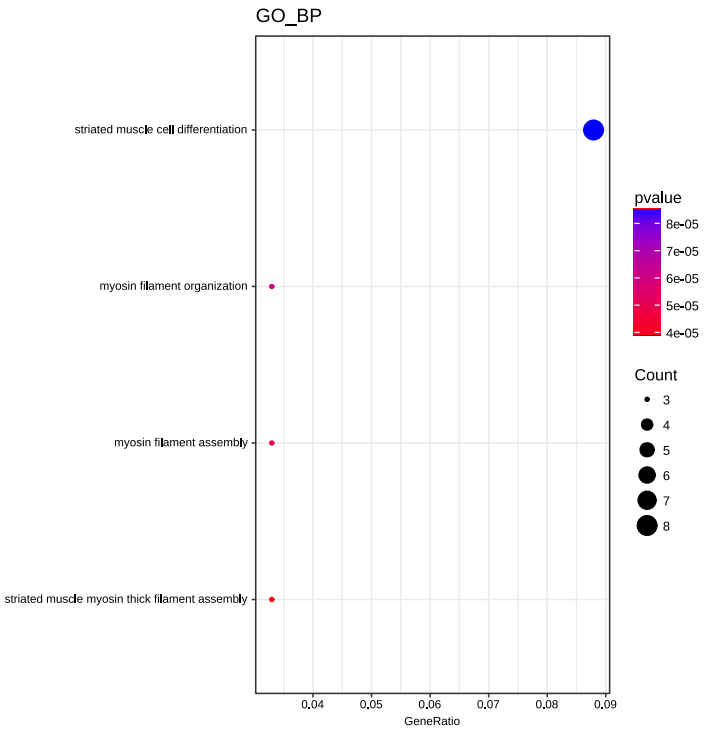

图 4-2 GO 生物学途径富集散点图

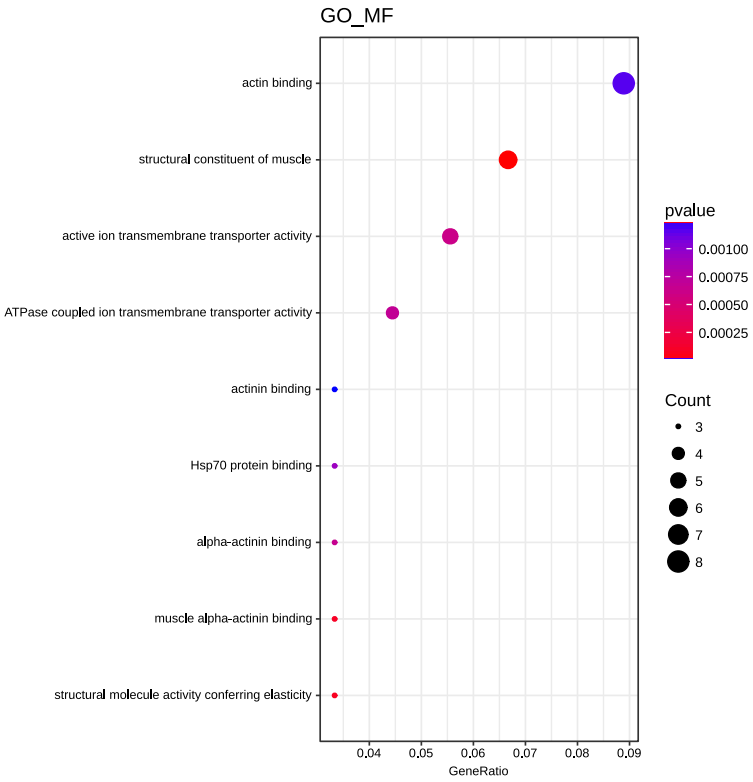

图 4-3 GO 分子功能富集散点图

五、 参考文献

1. An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. Cho YS , et al. Nat Commun. 2016 Nov 24;7:13637. doi: 10.1038/ncomms13637.
2. ReSeqTools: an integrated toolkit for large-scale next-generation sequencing based resequencing analysis. He W , et al. Genet Mol Res. 2013 Dec 4;12(4):6275-83. doi: 10.4238/2013.December.4.15.
3. Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. Leggett RM , et al. Front Genet. 2013 Dec 17;4:288. doi: 10.3389/fgene.2013.00288. Review
4. MultiQC: summarize analysis results for multiple tools and samples in a single report. Ewels P , et al. Bioinformatics. 2016 Oct 1;32(19):3047-8. doi: 10.1093/bioinformatics/btw354. Epub 2016 Jun 16.
5. Fast and accurate short read alignment with Burrows-Wheeler transform. Li H and Durbin R Bioinformatics. 2009 Jul 15;25(14):1754-60. doi: 10.1093/bioinformatics/btp324. Epub 2009 May 18.
6. The Sequence Alignment/Map format and SAMtools. Li H , et al. Bioinformatics. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352. Epub 2009 Jun 8.
7. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. McKenna A , et al. Genome Res. 2010 Sep;20(9):1297-303. doi: 10.1101/gr.107524.110. Epub 2010 Jul 19.
8. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Okonechnikov K , et al. Bioinformatics. 2016 Jan 15;32(2):292-4. doi: 10.1093/bioinformatics/btv566. Epub 2015 Oct 1.
9. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Wang K , et al. Nucleic Acids Res. 2010 Sep;38(16):e164. doi: 10.1093/nar/gkq603. Epub 2010 Jul 3.
10. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Ioannidis NM , et al. Am J Hum Genet. 2016 Oct 6;99(4):877-885. doi: 10.1016/j.ajhg.2016.08.016. Epub 2016 Sep 22.
11. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Jagadeesh KA , et al. Nat Genet. 2016 Dec;48(12):1581-1586. doi: 10.1038/ng.3703. Epub 2016 Oct 24.
12. Gene Ontology Consortium: going forward. Gene Ontology Consortium Nucleic Acids Res. 2015 Jan;43(Database issue):D1049-56. doi: 10.1093/nar/gku1179. Epub 2014 Nov 26.
13. clusterProfiler: an R package for comparing biological themes among gene clusters. Yu G , et al. OMICS. 2012 May;16(5):284-7. doi: 10.1089/omi.2011.0118. Epub 2012 Mar 28.