KJ-DB-HN160004-06-2016-2162

中南大学 1 个人样本全外显子组建库测序分析结题报告

安诺优达基因科技（北京）有限公司

2016/09/12

目录

一、背景介绍

外显子组（Exome）是一个物种基因组中全部外显子区域的总和，它是基因行使其功能最直接的体现。安诺优达外显子组测序，是基于 Illumina 测序平台，利用序列捕获技术将人全基因组外显子区域 DNA 捕捉并富集后进行高通量测序的基因组分析方法，能够直接发现与蛋白质功能变异相关的遗传突变。相比于全基因组重测序，外显子组测序更加经济、高效。目前，外显子组测序技术已经应用到寻找与各种复杂疾病相关的致病基因和易感基因的研究中。

二、实验流程

本研究采用 Agilent 51M 外显子靶向序列富集系统对人的外显子序列进行捕获。首先，对 DNA 进行纯度、浓度和体积等方面的检测；对符合质量要求的 DNA 建立基因组测序文库，形成片段化的基因组序列，该序列两端包含测序接头；基因组 DNA 文库与 AgilentSureSelect 生物素标记的 RNA 探针进行杂交，通过互补配对原则，目标基因组 DNA 片段与被生物素标记的寡核苷酸探针结合形成杂交复合物；未与液相芯片探针结合的基因组片段被洗脱纯化，利用 PCR 扩增捕获片段；对捕获杂交后的文库进行纯度、浓度检测。最后对检验合格的测序文库进行 Illumina HiSeq 高通量测序。实验流程如下图：

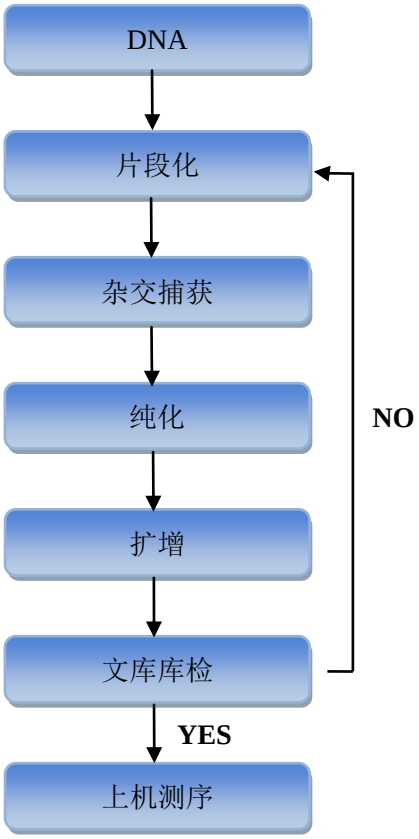

图 1 实验流程图

三、信息分析流程

安诺优达外显子捕获测序信息分析流程主要分为以下部分：
（1）数据处理及质控：将原始下机数据进行过滤并评估测序质量；
（2）比对及质控：将过滤后的数据比对到参考基因组上并对相应指标质控；
（3）Mutation 检测及注释：检测单样本的 SNP、INDEL、LOH 并进行各数据库的注释和分析；
具体技术流程如下图：

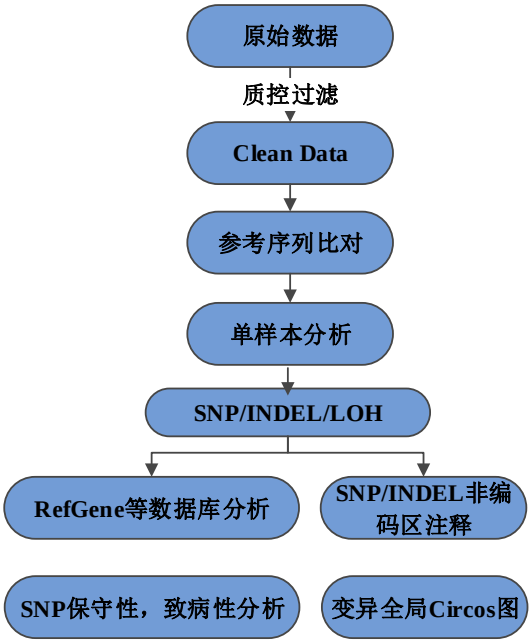

图 2 信息分析流程图

四、标准分析结果

1 样本信息

本报告中共有 1 个样品用于测序信息分析，对应的样本信息见下表：

表 1 样品信息收集表

Sample ID	Sample Name	Sample_Description	Others
D202334	L1667_46	男	-

（1）Sample ID：客户提供的原始样本名称；

（2）Sample Name：分析结果中使用的样本名称；

（3）Sample Description：客户提供的原始样本描述信息；

（4）Others：该样本的其他信息。

2 数据处理及质控

2.1 原始测序数据

高通量测序（Illumina）得到的原始图像数据文件经 CASAVA 碱基识别（Base Calling）分析转化为测序序列（Sequenced Reads），结果以 FASTQ（简称为 fq）文件格式存储，我们称之为 RawReads。

FASTQ 文件包含每条测序序列（Read）的名称、碱基序列以及其对应的测序质量信息。在 FASTQ 格式文件中，每个碱基对应一个碱基质量字符，每个碱基质量字符对应的 ASCII 码值减去 33，即为该碱基的测序质量得分（Phred Quality Score）。不同 Phred Quality Score 代表不同的碱基测序错误率，如 Phred Quality Score 值为 20 和 30 分别表示碱基测序错误率为 1.0%和 0.1%。其中 FASTQ 格式示例如下：

```
@HWI-ST1268:544:H8Y02ADXX:1:1101:1480:2221 1:N:0
TATGGTTGGCTGTTCACGGCCTGGAATTCTCGGGTGCCAAGGAACTCCA
+
1:BDFFDFFHHHHJJJJIJJJJJJIJJJIHIJJJHIDIJJJJJJJIIJJ
```

图 3FASTQ 文件格式示例图

（1）第一行以“@”开头，随后为 Illumina 测序标识别符（Sequence Identifiers）和描述文字（选择性部分）；

（2）第二行是碱基序列；

（3）第三行以“+”开头，随后为 Illumina 测序标识别符（选择性部分）；

（4）第四行是对应碱基的测序质量，该行中每个字符对应的 ASCII 值减去 33，即为对应第二行对应碱基的测序质量值。

2.2 原始数据过滤

测序得到的某些原始下机序列，会含有测序接头序列以及低质量序列，为保证信息分析数据的质量，我们对原始序列进行过滤，得到高质量的 Clean Reads，再进行后续分析。

数据处理步骤如下：

（1）去除接头污染的 Reads（Reads 中接头污染的碱基数大于 5bp。对于双端测序，若一端受到接头污染，则去掉两端的 Reads）；

（2）去除低质量的 Reads（Reads 中质量值 Q≤19 的碱基占总碱基的 50%以上，对于双端测序，若一端为低质量 Reads，则会去掉两端 Reads）；

（3）去除含 N 比例大于 5%的 Reads（对于双端测序，若一端含 N 比例大于 5%，则会去掉两端 Reads）。

数据产出及过滤统计结果见下表：

表 2 数据产出质量情况一览表

Sample	L1667_46
Read Length (bp)	150
Raw Reads	39,450,716
Raw Bases	5,917,607,400
Clean Reads	36,644,028
Clean Reads Rate (%)	92.89
Clean Bases	5,496,604,200
Low-quality Reads	313,834
Low-quality Reads Rate (%)	0.8
Ns Reads	256
Ns Reads Rate (%)	0
Adapter Polluted Reads	2,492,598
Adapter Polluted Reads Rate (%)	6.32
Raw Q30 Bases Rate (%)	91.62
Clean Q30 Bases Rate (%)	91.88

（1）Reads Length (bp)：Reads 长度；

（2）RawReads：原始下机的 Reads 数；

（3）RawBases (bp)：原始下机序列的碱基数；

（4）Clean Reads：过滤后得到的高质量的 Reads 数；

（5）CleanReads Rate (%)：过滤后得到的 Clean Reads 占 RawReads 的比例。这个值越大，说明测序质量越好；

（6）Clean Bases (bp)：过滤后的高质量序列的碱基数；

（7）Low-quality Reads：去除的低质量的 Reads 数（如双端测序一端是低质量，则去除的为双端）；

（8）Low-quality Reads Rate (%)：去掉的低质量的 Reads 占 RawReads 的比例；

（9）Ns Reads：去除的含 N 比例大于 5%的 Reads 数；

（10）Ns Reads Rate (%)：去除的含 N 比例大于 5%的 Reads 占 RawReads 的比例；

（11）Adapter Polluted Reads：去除的接头污染的 Reads 数；

（12）Adapter Polluted Reads Rate(%)：去除的接头污染的 Reads 占 RawReads 的比例；

（13）Raw Q30 Bases Rate (%)：RawReads 中测序质量值大于 30（错误率小于 0.1%）的碱基占总碱基（RawBases）的比例；

（14）Clean Q30 Bases Rate(%)：Clean Reads 中测序质量值大于 30（错误率小于 0.1%）的碱基占总碱基（Clean Bases）的比例。

详细的数据产出及过滤统计请见：

Report/1_FQ/fq_filter.report.xls

RawReads 包含低质量的序列、接头污染的序列、含 N 比例大于 5%的序列，以及 Clean Reads。Clean Reads 所占的比例越高，数据质量越好，根据每种序列所占的比例分别对每个样本绘制饼图，对所有样本绘制柱状图，可以直观地反映数据过滤情况：

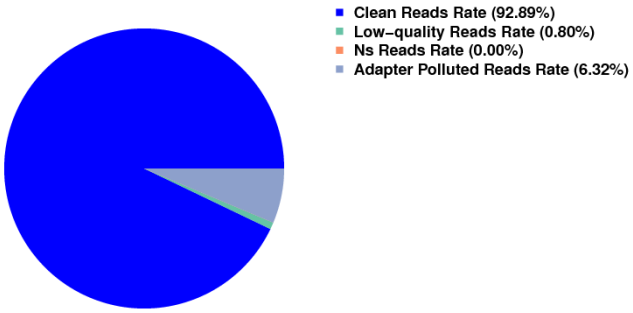

图 4Raw Data 数据组成统计图

横轴表示样品名，纵轴表示不同过滤标准过滤掉的 Reads 的百分比。

详细的统计图请见：

Report/1_FQ/fq_filter.report.xls.filter.png/pdf

2.3 数据量质控

下图为本项目样本的Clean Bases数据量统计图，可以很直观看出这个样本的数据量满足合同要求的数据量：

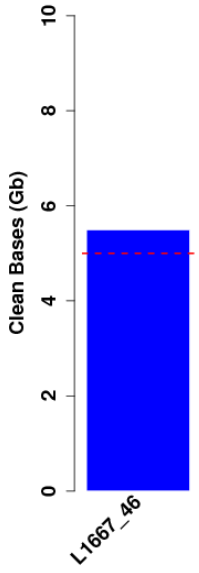

图 5 样品数据量分布图

横坐标是样本名，纵坐标为数据量；红色的虚线标出了合同要求的数据量。

详细的统计图请见：

Report/1_FQ/fq_filter.report.xls.dataSize.png/pdf

2.4 碱基含量分布

以过滤后序列的碱基位置作为横坐标，以每个位置的 ATCGN 碱基的比例作为纵坐标，得到碱基分布图。正常情况下，由于碱基互补配对原则和测序的随机性，在每个测序位置 A

碱基和 T 碱基比例相等，G 碱基和 C 碱基比例相等。

碱基含量分布图如下：

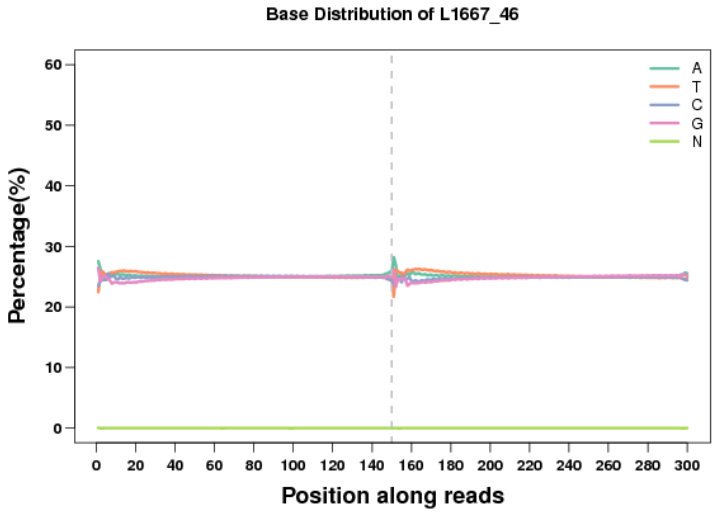

图 6 碱基含量分布图

横轴为 Reads（双端测序则为一对 Reads）的碱基位置，纵轴为单碱基所占的比例，不同颜色代表不同的碱基类型。

详细的统计图请见：

Report/1_FQ/2_CleanData/sampleID/sampleID_Clean.base.pdf/png

2.5 碱基质量值分布

碱基测序质量值反映了测序错误率，质量值越大，错误率越小，它受测序仪本身、测序试剂、样品等多个因素共同影响。碱基测序错误率是利用 Phred 数值（Phred Score，Q_{phred}）通过以下公式转化得到（其中 e 为碱基测序错误率）：

$$Phred = -10\log_{10} e$$

Phred 数值是在碱基识别（Base Calling）过程中，通过一种预测碱基发生错误概率模型计算得到的，对应关系如下表所示：

表 3IlluminaCasava 碱基识别与 Phred 分值之间的简明对应关系表

Phred 数值	碱基测序错误率（e）	碱基正确识别率	Q-Score
10	1/10	90%	Q10
20	1/100	99%	Q20
30	1/1000	99.9%	Q30
40	1/10000	99.99%	Q40

碱基质量值分布具有两个特点：

（1）碱基质量值会随着测序序列（Sequenced Reads）长度的增加而降低，这个特点是 Illumina 高通量测序平台都具有的特征；

（2）前几个碱基因定位的影响，测序质量值较其他位置会低一些。

为了反映 Clean Data 的质量，以保证分析的准确性，以 Q30 碱基百分比作为指标进行统计，Q30 碱基百分比越大说明测序错误率小于 0.1%的碱基在总碱基中的比例越大。下图以所有样本为横轴，以 Q30 百分比作为纵轴直观地反映了过滤后的数据质量：

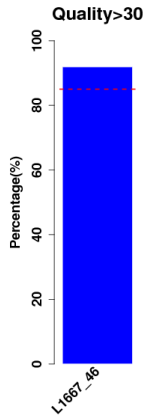

图 7 所有样本 Clean Data 的质量统计图

横坐标是样本名，纵坐标为 Clean Data 中 Q30 百分比，红色的虚线标出了 Q30 质控线。

详细的统计图请见：

Report/1_FQ/fq_filter.report.xls.Q30.png/pdf

为了反映测序过程中测序质量的稳定性，以 Clean Reads 的碱基位置作为横坐标，每个位置的平均测序质量值作为纵坐标，得到每次测序数据质量的分布图：

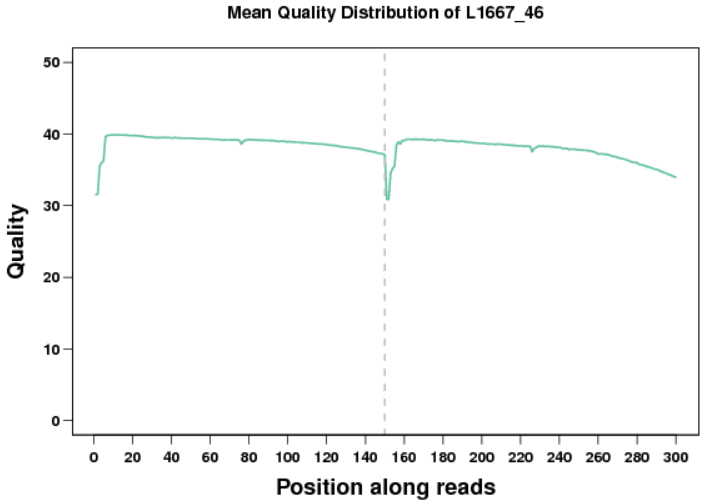

图 8 Clean Reads 碱基平均质量值分布

横坐标为 Clean Reads 的碱基位置，纵坐标为平均碱基质量。

详细的碱基质量分布图请见：

Report/1_FQ/2_CleanData/sampleID/sampleID_Clean.quality.pdf/png

3 比对及质控

3.1 比对信息统计

利用基因组比对软件 BWA[1]，将过滤后的 Clean Reads 比对到参考基因组上，得到 BAM[2] 格式文件。使用 samtools[3]软件对 BAM 进行排序，再利用 Picard[4]标记 Duplication read。统计比对结果，对于外显子测序分析，比对率以及覆盖度指标能反映样本、建库及测序以及参考序列等的质量；Uniq Rate 则反映所测序列的重复度；Duplication Rate 反映建库质量等。

这一系列统计信息都将作为分析的质控指标。具体信息见下表：

表 4 比对数据统计

Sample	L1667_46
Target Region (bp)	51,542,852
Clean Reads	36,727,369
Clean Bases (Mb)	5,500.40
Reads Mapped to Genome	36,635,575
Map Rate (%)	99.75
Reads Mapped to Rarget Region	27,180,872
Capture Specificity (%)	74.19
Duplication Rate (%)	9.26
Uniq Rate (%)	95.88
Uniq Reads Mapped to Target Region	23,774,903
Mean Depth of Target Region	53.52
Coverage of Target Region (%)	96.29
Rate of Nucleotide Mismatch (%)	0.36
Fraction of Target Covered ≥ 4X	93.69
Fraction of Target Covered ≥ 10X	89.77
Fraction of Target Covered ≥ 20X	81.1
Fraction of Target Covered ≥ 30X	70.14
Fraction of Target Covered ≥ 40X	58.74
Uniq Reads Mapped to Flanking Region	15,583,693
Mean Depth of Flanking Region	14.26
Coverage of Flanking Region (%)	84.7
Fraction of Flanking Covered ≥ 4X	62.89

（1）Target Region：基因组目标捕获区域大小；

（2）Clean Reads：过滤后的 Reads 的个数；

（3）Clean Bases (Mb)：过滤后的碱基数；

（4）Reads Mapped to Genome：比对到参考基因组上的 Reads 个数；

（5）Map Rate (%)：比对到参考基因组上的 Reads 的百分比；

（6）Reads Mapped to Target Region：比对到基因组目标捕获区域上的 Reads 个数；

（7）Capture Specificity (%)：比对到基因组目标捕获区域的 Reads 数占比对到基因组 Reads 的比例；

（8）Duplication Rate (%)：去除的 PCR 重复的 Reads 数占 Mapped Reads 的比例；

（9）Uniq Rate (%)：比对到基因组唯一位置的 Reads 占去除 PCR 重复后比对上的 Reads 的比例；

（10）Uniq Reads Mapped to Target Region：基因组目标捕获区域中 Uniq Reads 的数量；

（11）Mean Depth of Target Region：基因组目标捕获区域 Uniq Reads 的平均测序深度，即用于后续分析的数据深度；

（12）Coverage of Target Region (%)：覆盖度，即目标捕获区中有多大比例的区域至少测到了 1 次；

（13）Rate of Nucleotide Mismatch (%)：错配率，即 Reads 与参考序列相比错配碱基占总 Reads 碱基的比例；

（14）Fraction of Target Covered ≥ NX：基因组目标捕获区域深度≥ NX 的碱基覆盖度；

（15）Uniq Reads Mapped to Flanking Region：基因组目标捕获区域的侧翼区（100bp）中 Uniq Reads 的数量；

（16）Mean Depth of Flanking Region：基因组目标捕获区域的侧翼区（100bp）的平均测序深度；

（17）Coverage of Flanking Region (%)：基因组目标捕获区域的侧翼区（100bp）的平均覆盖度；

（18）Fraction of Flanking Region Covered ≥ NX：基因组目标捕获区域的侧翼区（100bp）深度≥ NX 的碱基覆盖度。

详细的统计表请见：

Report/2_MAP/Map.report.xls

Report/2_MAP/sampleID/depthByChr.stat（各染色体的统计信息）

将样本测序序列与基因组参考序列的比对率及目标区域的覆盖度作图，如下：

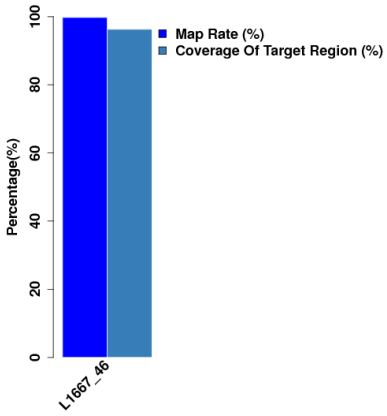

图 9 比对率及覆盖度分析图

横坐标为样本名，纵坐标为比对率及目标区域覆盖度。

详细的统计图请见：

Report/2_MAP/Map.report.png/pdf

3.2 测序深度分布

测序深度的分布可以反映建库测序的均一性及对基因组目标捕获区域的详细覆盖情况。

单碱基深度分布图以测序深度为横坐标，相应深度的位点比例为纵坐标，此处 Fraction of Bases=（特定深度对应的位点数/基因组目标捕获区域的长度），它反映了特定深度下对应的目标捕获区域的覆盖度。下图为选取 Uniq 比对后的测序深度分布图：

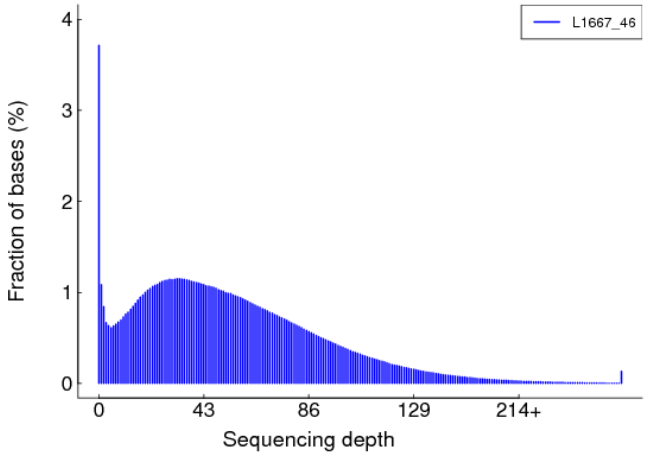

图 10 单碱基深度分布图

单碱基深度分布图以深度为横坐标，相应深度的碱基所占比例为纵坐标。

详细的深度分布图请见：

Report/2_MAP/sampleID/sampleID_histPlot.png/pdf

累积深度分布图以深度为横坐标，大于相应深度的位点比例为纵坐标，此处 Fraction of

Bases=（大于此深度的位点数/目标捕获区域的长度），它反映了大于特定深度下的目标捕获区域的覆盖度。下图为选取 Uniq 比对后的累积深度分布图：

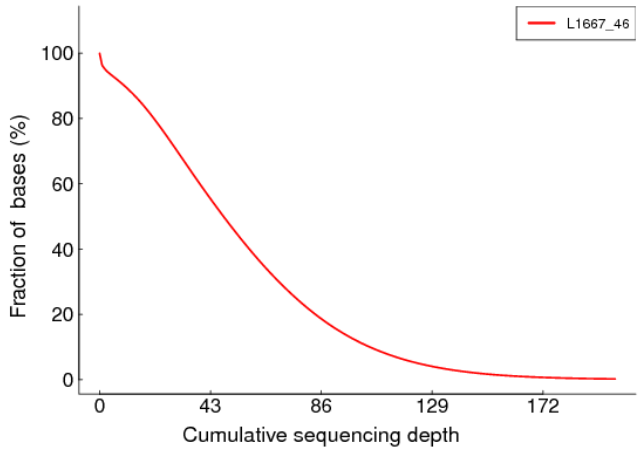

图 11 累积深度分布图

累积深度分布图以深度为横坐标，Fraction of Bases 为纵坐标。

详细的累计深度分布图请见：
Report/2_MAP/sampleID/sampleID_cumuPlot.png/pdf

4 单样本变异检测及注释

在检测变异位点时，通过突变分析软件 GATK[5]检测样本与参考基因组间的变异位点，进行质量过滤，并对其进行注释。

4.1 SNP 检测及注释

在比对到参考基因组序列的基础上，通过突变分析软件 GATK 从中提取全基因组中所有的潜在多态性 SNP 位点，再根据质量值、深度、重复性等因素做进一步的过滤筛选，最终得到高可信度的 SNP 数据集，并对其进行 ANNOVAR[6]注释。

4.1.1 SNP 的 RefSeqGene 数据库注释

Refseq[7]为 NCBI 提供的全方位综合，非冗余的基因和蛋白序列数据库，RefSeqGene 为 RefSeq 的一个子数据库。通过 RefSeqGene 的注释，可以确定 SNP 在基因组不同功能区域的分布，确定是否发生非同义突变，是否影响蛋白功能。

将样本所检测到的 SNP 个数进行统计，作如下柱状图：

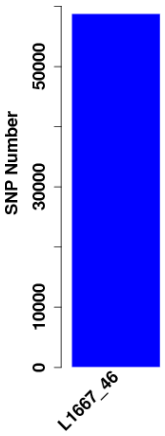

图 12SNP 个数统计图

详细的统计图请见：
Report/3_SNP-INDEL/stat/SNP_num.png/pdf

4.1.1.1 基因组不同区域上 SNP 的分布

检测到的样本 SNP 在基因组不同区域上的分布统计见下表所示：

表 5 基因组不同区域上 SNP 的分布统计

#L1667_46	Number	Percent(%)
Total	58,822	100.00
UTR5	1,354	2.30
UTR3	1,841	3.13
UTR5;UTR3	4	0.01
exonic	17,200	29.24
splicing	67	0.11
exonic;splicing	4	0.01
upstream	517	0.88
downstream	231	0.39
upstream;downstream	28	0.05
intronic	33,723	57.33
intergenic	1,679	2.85
ncRNA_UTR3	0	0
ncRNA_UTR5	0	0
ncRNA_exonic	759	1.29
ncRNA_splicing	2	0.00
ncRNA_intronic	1,412	2.40
Other	1	0.00

（1）Total：基因组中全部的 SNP 数目；

（2）UTR5：发生在基因的 UTR5 的 SNP 数目；

（3）UTR3：发生在基因的 UTR3 的 SNP 数目；

（4）UTR5;UTR3：同时发生在 UTR3 和 UTR5 区域内的 SNP 数目，两个不同的基因；

（5）exonic：发生在外显子区域的 SNP 数目；

（6）splicing：发生在基因剪切区域内（剪切位点上游 2bp，即非 Exonic 区）的 SNP 数目；

（7）exonic;splicing：同时发生在外显子区域和基因剪切区域内的 SNP 数目，两个不同的基因；

（8）upstream：发生在基因上游（1000bp）内的 SNP 数目；

（9）downstream：发生在基因下游（1000bp）内的 SNP 数目；

（10）upstream/downstream：发生在基因上游或者下游（1000bp）内的 SNP 数目，两个不同的基因；

（11）intronic：发生在内含子区域的 SNP 数目；

（12）intergenic：发生在基因间区的 SNP 数目；

（13）NcRNA：没有相关编码注释的 RNA，并非不翻译的 RNA，参见 ANNOVAR 的 Gene 注释说明；子区间注释同上。

（14）Other：是指杂合的 SNP，与参考基因组不一致，所以 Annovar 未得到注释，注释结果中为 NULL。

详细的统计表见：

Report/4_ANNOTATE_SNP-INDEL/stat/sampleID.SNP.genome.region.stat.txt

基因组不同区域的 SNP 分布图见下所示：

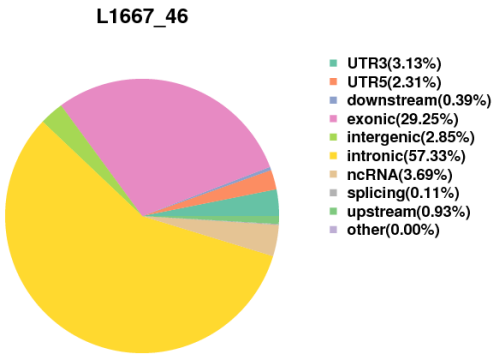

图 13 基因组不同区域的 SNP 分布图

扇形区域为各区间的 SNP 个数百分比；作图时，将相同功能区的合并，如 UTR5;UTR3 归入 UTR5 分类。

详细的统计图

Report/4_ANNOTATE_SNP-INDEL/stat/sampleID.SNP.genome.region.png/pdf

4.1.1.2 外显子区 SNP 功能注释及统计

外显子区域的 SNP 突变可能会影响到氨基酸的编码，进而影响基因功能。将位于外显子区域的突变根据其是否引起氨基酸的改变进行分类注释，如非同义突变、同义突变等，通常非同义突变导致相应氨基酸改变从而使得基因功能发生改变，而 Stopgain 和 Stoploss 导致了终止子的提前出现或缺失，所以也是有害突变。

统计位于外显子区域（exonic）的 SNP 对蛋白质翻译的影响，见下表所示：

表 6 外显子区域的 SNP 的突变类型统计

#L1667_46	Number	Percent(%)
Total	17,200	100.00
nonsynonymous SNV	7,831	45.53
synonymous SNV	8,946	52.01
stopgain	52	0.30
stoploss	6	0.03
unknown	365	2.12

（1）Total：外显子区域的 SNP 数目；

（2）Nonsynonymous SNV：非同义突变，密码子的改变导致编码的氨基酸改变（此处 SNV 同 SNP）；

（3）Synonymous SNV：同义突变，密码子变异为编码同一氨基酸的密码子，核苷酸的改变不引起氨基酸的改变，即不引起基因产物的突变；

（4）Stopgain：密码子的改变导致终止子的出现；

（5）Stoploss：密码子的改变导致终止子的缺失；

（6）Unknown：未知类型。

详细的统计表见：

Report/4_ANNOTATE_SNP-INDEL/stat/sampleID.SNP.ExonicFunc.stat.txt

将样本外显子区的 SNP 功能统计结果作如下图，得到其分布情况：

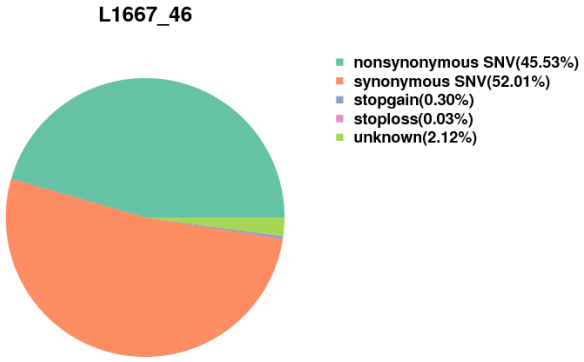

图 14SNP 功能分布图

扇形区域为各类型的 SNP 个数百分比。

详细的统计图请见：

Report/4_ANNOTATE_SNP-INDEL/stat/sampleID.SNP.ExonicFunc.png/pdf

4.1.1.3 基因组和编码区的 SNP 转换/颠换比率

碱基颠换(transversion)是指在碱基置换中嘌呤与嘧啶之间的替代，而转换(transition)则是一个嘌呤被另一个嘌呤，或者是一个嘧啶被另一个嘧啶替代，由于结构原因，转换发生的概率高于颠换发生的概率，转换和颠换的比率可以反映检测到的 SNP 准确性。

在基因组和编码区的 SNP 转换/颠换比率，见下表所示：

表 7 基因组和编码区的 SNP 转换/颠换比率统计

#Sample	TS_genome	TV_genome	TS/TV_genome	TS_exonic	TV_exonic	TS/TV_exonic
L1667_46	41,414	17,408	2.38	13,024	4,176	3.12

（1）TS_genome：全基因组上发生转换的 SNP 数；

（2）TV_geneome：全基因组上发生颠换的 SNP 数；

（3）TS/TV_genome：全基因组上 SNP 转换/颠换比率；

（4）TS_exonic：外显子区域上发生转换的 SNP 数；

（5）TV_exonic：外显子区域发生颠换的 SNP 数；

（6）TS/TV_ exonic：外显子区域上 SNP 转换/颠换比率。

详细的统计表请见：

Report/4_ANNOTATE_SNP-INDEL/stat/sampleID.SNP.TS_TV.stat.txt

4.1.1.4 基因组和编码区的 SNP 的纯合、杂合类型

样品 L1667_46 在基因组和编码区的 SNP 的纯合和杂合类型统计，见下表所示：

表 8 基因组和编码区的 SNP 的纯杂合统计

#L1667_46	Hom_genome	Het_genome	Hom_exonic	Het_exonic
Number	31,477	27,345	8,933	8,267
Percentage(%)	53.51	46.49	51.94	48.06

（1）Hom_genome：全基因组上纯合 SNP 数；

（2）Het_genome：全基因组上杂合 SNP 数；

（3）Hom_exonic：外显子区域上纯合 SNP 数；

（4）Het_exonic：外显子区域上杂合 SNP 数。

详细的统计表请见：

Report/4_ANNOTATE_SNP-INDEL/stat/sampleID.SNP.homo_hete.stat.txt

4.1.1.5 SNP 突变模式分布统计

不同物种、不同环境等影响会导致 SNP 突变模式的不同，通过统计 SNP 突变模式的分布情况，可以得到特定物种、特定类型样本的特有突变模式，从而对该物种或样本有更为全面的了解和分析，SNP 突变模式分布见下表所示：

表 9SNP 突变模式分布

#Sample	L1667_46
T-A	2,898
T-C	20,283
T-G	4,344
C-A	4,285
C-T	21,131
C-G	5,881

（1）T-A：即 T 到 A 的突变（包含反链的 A 到 T 的突变）；

（2）T-C：即 T 到 C 的突变（包含反链的 A 到 G 的突变）；

（3）T-G：即 T 到 G 的突变（包含反链的 A 到 C 的突变）；

（4）C-A：即 C 到 A 的突变（包含反链的 G 到 T 的突变）；

（5）C-T：即 C 到 T 的突变（包含反链的 G 到 A 的突变）；

（6）C-G：即 C 到 G 的突变（包含反链的 G 到 C 的突变）。

详细的统计表请见：

Report/4_ANNOTATE_SNP-INDEL/stat/sampleID.SNP.mutation.pattern.stat.txt

样本的各突变模式的比例作如下饼形图：

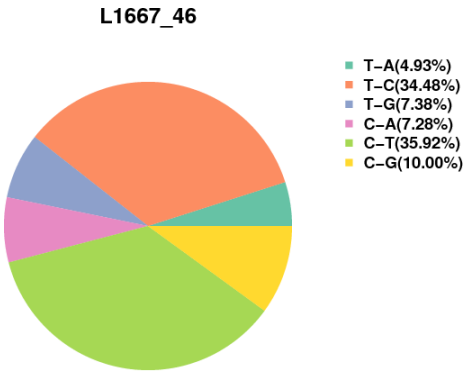

图 155SNP 突变模式分布图

扇形区域为各类型的 SNP 个数百分比。

详细的统计图请见

Report/4_ANNOTATE_SNP-INDEL/stat/sampleID.SNP.mutation.pattern.stat.png/pdf

4.2 INDEL 检测及注释

在比对到参考基因组序列的基础上，通过突变分析软件 GATK 从中提取全基因组中所有的潜在多态性 INDEL 位点，再根据质量值、深度、重复性等因素做进一步的过滤筛选，最终得到高可信度的 INDEL 数据集，并对其进行 ANNOVAR 注释。

4.2.1 INDEL 的 RefSeqGene 数据库注释

RefSeq 为 NCBI 提供的全方位综合，非冗余的基因和蛋白序列数据库，RefSeqGene 为 RefSeq 的一个子数据库。通过 RefSeqGene 的注释，可以确定 INDEL 在基因组不同功能区域的分布，确定是否发生非同义突变，是否影响蛋白功能。

将样本所检测到的 INDEL 个数进行统计，作如下柱状图：

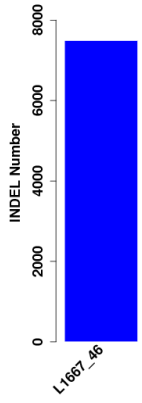

图 166INDEL 个数统计图

详细的统计图请见：

Report/3_SNP-INDEL/stat/INDEL_num.png /pdf

4.2.1.1 基因组不同区域上 INDEL 的分布

INDEL 在基因组不同区域上的分布统计见下表所示：

表 10 基因组不同的 INDEL 分布统计

#L1667_46	Number	Percent(%)
Total	7,680	100.00
UTR5	215	2.80
UTR3	315	4.10
UTR5;UTR3	0	0
exonic	512	6.67
splicing	60	0.78
exonic;splicing	3	0.04
upstream	87	1.13
downstream	30	0.39
upstream;downstream	4	0.05
intronic	5,963	77.64
intergenic	167	2.17
ncRNA_UTR3	0	0
ncRNA_UTR5	0	0
ncRNA_exonic	82	1.07
ncRNA_splicing	0	0
ncRNA_intronic	241	3.14
Other	1	0.01

（1）Total：基因组中全部的 INDEL 数目；

（2）UTR5：发生在基因的 UTR5 的 INDEL 数目；

（3）UTR3：发生在基因的 UTR3 的 INDEL 数目；

（4）UTR5;UTR3：同时发生在 UTR3 和 UTR5 区域内的 INDEL 数目，两个不同的基因；

（5）exonic：发生在外显子区域的 INDEL 数目；

（6）splicing：发生在基因剪切区域内（剪切位点上游 2bp，即非 Exonic 区）的 INDEL 数目；

（7）exonic;splicing：同时发生在外显子区域和基因剪切区域内的 INDEL 数目，两个不同的基因；

（8）upstream：发生在基因上游（1000bp）内的 INDEL 数目；

（9）downstream：发生在基因下游（1000bp）内的 INDEL 数目；

（10）upstream/downstream：发生在基因上游或者下游（1000bp）内的 INDEL 数目，两个不同的基因；

（11）intronic：发生在内含子区域的 INDEL 数目；

（12）intergenic：发生在基因间区的 INDEL 数目；

（13）NcRNA：没有相关编码注释的 RNA，并非不翻译的 RNA，参见 ANNOVAR 的 Gene 注释说明；子区间注释同上。

（14）Other：是指杂合的，与参考基因组不一致，所以 Annovar 未得到注释，注释结果中为 NULL。

详细的统计表见：
Report/4_ANNOTATE_SNP-INDEL/stat/sampleID.INDEL.genome.region.stat.txt

基因组不同区域的 INDEL 分布图见下所示：

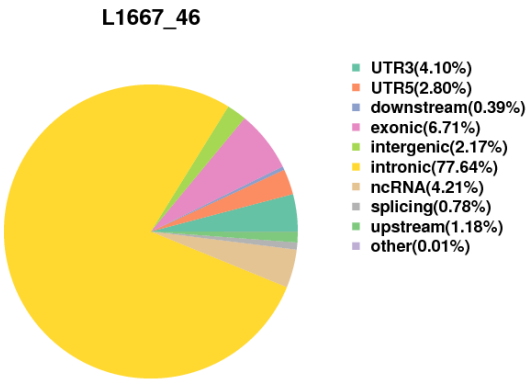

图 177 基因组不同区域的 INDEL 分布图

扇形区域为各区间的 INDEL 个数百分比；作图时，将相同功能区的合并，如 UTR5;UTR3 归入 UTR5 分类。

详细的统计图见：

Report/4_ANNOTATE_SNP-INDEL/stat/sampleID.INDEL.genome.region.stat.png/pdf

4.2.1.2 外显子区 INDEL 功能注释及统计

外显子区域的突变可能会影响到氨基酸的编码，进而影响基因功能。将位于外显子区域的突变根据其是否引起氨基酸的改变进行分类注释，如非同义突变、同义突变等，通常非同义突变导致相应氨基酸改变从而使得基因功能发生改变，而 Stopgain 和 Stoploss 导致了终止子的提前出现或缺失，所以也是有害突变。

统计位于外显子区域（exonic）的对蛋白质翻译的影响，见下表所示：

表 11 外显子区域的的突变类型统计

#L1667_46	Number	Percent(%)
Total	512	100.00
frameshift deletion	59	11.52
frameshift insertion	54	10.55
nonframeshift deletion	170	33.20
nonframeshift insertion	138	26.95
stopgain	3	0.59
stoploss	1	0.20
unknown	87	16.99

（1）Total：外显子区域的数目；

（2）frameshift deletion：缺失移码突变，碱基缺失为非 3 的倍数，造成这位置之后的一系列编码发生移位错误的改变；

（3）frameshift insertion：插入移码突变，碱基插入为非 3 的倍数，造成这位置之后的一系列编码发生移位错误的改变；

（4）nonframeshift deletion：非移码突变，碱基缺失为 3 的倍数；

（5）nonframeshift insertion：非移码突变，碱基插入为 3 的倍数；

（6）Stopgain：密码子的改变导致终止子的出现；

（7）Stoploss：密码子的改变导致终止子的缺失；

（8）Unknown：未知类型。

详细的统计表见

Report/4_ANNOTATE_SNP-INDEL/stat/sampleID.INDEL.ExonicFunc.stat.txt

将样本外显子区的 INDEL 功能统计结果作如下图，得到其分布情况：

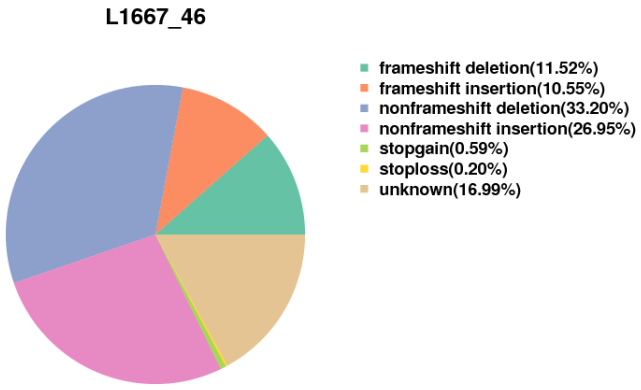

图 188 INDEL 功能分布图

扇形区域为各类型的 INDEL 个数百分比。

详细的统计图见

Report/4_ANNOTATE_SNP-INDEL/stat/sampleID.INDEL.ExonicFunc.stat.png/pdf

4.2.1.3 INDEL 突变模式分布统计

INDEL 的长度的不同会引起对基因组的不同程度的影响，在全基因组及外显子区，其不同长度的 INDEL 的分布有着明显的差异，外显子区因其所需的特有的保守性，3 个碱基的 INDEL 的数量比例较 2 碱基及 4 碱基等的多（3 碱基 INDEL 不容易引起移码）。下表为样品的 INDEL 突变模式的统计：

表 12 INDEL 突变模式统计

#L1667_46	Genome	Exonic
1	3,556	133
2	1,244	32
3	741	173
4	681	9
5	229	4
6	269	54
>6	960	107

（1）Length：INDEL 的长度；

（2）Genome：全基因组上的 INDEL 数；

（3）Exonic：外显子区域的 INDEL 数。

详细的统计表见:

Report/4_ANNOTATE_SNP-INDEL/stat/sampleID.INDEL.length.pattern.stat.txt

将目标区域及外显子区的不同长度的 INDEL 的个数作如下柱形图：

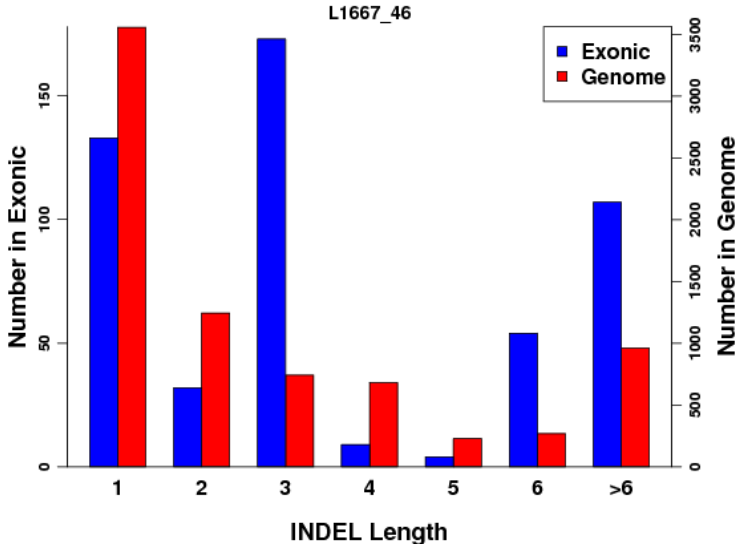

图 19 INDEL 突变模式分布图

详细的统计图见
Report/4_ANNOTATE_SNP-INDEL/stat/sampleID.INDEL.length.pattern.stat.png/pdf

4.3 SNP/INDEL 数据库分析

dbSNP[8]为 NCBI 中的一个数据库(The Database of Short Genetic Variation)。通过此数据库的注释，可以确定检测到的 SNP/INDEL 是否为已知的 SNP/INDEL，并且 NCBI 中提供详细的注释信息。数据库版本为 dbSNP142。

1000 Genome Project[9],为国际千人基因组计划，通过此数据库的注释提供不同人群中的 SNP/INDEL 的次等位基因型频率(MAF)，一般后续的研究基于 MAF<0.01 的 SNP/INDEL。数据库版本为 1000g2015aug。

Exome Aggregation Consortium（ExAC）[10]通过此数据库的注释提供不同人群中的 SNP/INDEL 的 MAF 值，一般后续的研究基于 MAF<0.01 的 SNP/INDEL。数据库版本为 0.3。

Exome Sequencing Project（ESP）[11]，ESP 计划主要是针对心脏，肺和血液疾病，通过对编码蛋白的外显子测序，研究和发现新的致病基因或致病机制。通过此数据库的注释提供 SNP/INDEL 的 MAF 值，一般后续的研究基于 MAF<0.01 的 SNP/INDEL。数据库版本为 esp6500siv2。

cg69 和 cg46 为 Complete Genomics[12]公司所测的 69 个有关系的个体，以及 46 个无关的个体，其中 69 个个体包括 46 个个体，通过此数据库的注释提供 SNP/INDEL 的 MAF 值，一般后续的研究基于 MAF<0.1 的 SNP/INDEL。

以上所有数据库均来自于 ANNOVAR 整理后的数据库 avsnp142 ，popfreq_all_20150413 和 cg69。

样品各数据库统计结果见下表所示：

表 13 SNP 数据库分布统计

DataBase	Number	MAF
dbSNP	58,195	-
Cosmic	499	-
1000G_ALL	56,212	1,344
ExAC_ALL	34,984	1,416

ESP6500siv2_ALL	31,124	1,005
CG46	54,331	5,068
cg69	54,422	5,490

（1）Number：注释到各数库的 SNP 数；

（2）MAF：MAF 值低于阈值的 SNP 数，dbSNP 无 MAF 分析；

详细的统计表见
Report/4_ANNOTATE_SNP-INDEL/stat/sampleID.SNP.database.stat.txt

表 14 INDEL 数据库分布统计

DataBase	Number	MAF
dbSNP	6,811	-
Cosmic	281	-
1000G_ALL	4,992	144
ExAC_ALL	3,461	234
ESP6500siv2_ALL	2,131	58
CG46	2,613	1,542
cg69	2,902	1,838

（1）Number：注释到各数库的 INDEL 数；

（2）MAF：MAF 值低于阈值的 INDEL 数，dbSNP 无 MAF 分析。

详细的统计表见
Report/4_ANNOTATE_SNP-INDEL/stat/sampleID.INDEL.database.stat.txt

数据库注释结果格式说明：

表 15 数据库注释结果说明

Type	Discription
avsnp142	dbSNP142 数据库注释结果,有提供 rs 编号，无“.”表示
PopFreqMax	1000Genome、ESP、ExAC 和 cg46 数据库中的最大的 MAF，无“.”表示
1000G_ALL	1000Genome 所有人群的 MAF，无“.”表示
1000G_AFR	1000Genome 中 African 人群的 MAF，无“.”表示
1000G_AMR	1000Genome 中 Admixed American 人群的 MAF，无“.”表示
1000G_EAS	1000Genome 中 East Asian 人群的 MAF，无“.”表示
1000G_EUR	1000Genome 中 European 人群的 MAF，无“.”表示
1000G_SAS	1000Genome 中 South Asian 人群的 MAF，无“.”表示
ExAC_ALL	ExAC 所有人群的 MAF，无“.”表示
ExAC_AFR	ExAC 中 African 人群人群的 MAF，无“.”表示
ExAC_AMR	ExAC 中 Admixed American 人群人群的 MAF，无“.”表示
ExAC_EAS	ExAC 中 East Asian 人群人群的 MAF，无“.”表示
ExAC_FIN	ExAC 中 Finnish 人群人群的 MAF，无“.”表示
ExAC_NFE	ExAC 中 Non-finnish in European 人群人群的 MAF，无“.”表示
ExAC_OTH	ExAC 中 other 人群人群的 MAF，无“.”表示
ExAC_SAS	ExAC 中 South Asian 人群人群的 MAF，无“.”表示
ESP6500siv2_ALL	ESP 中所有人群的 MAF，无“.”表示
ESP6500siv2_AA	ESP 中 African American 人群的 MAF，无“.”表示
ESP6500siv2_EA	ESP 中 European American 人群的 MAF，无“.”表示
CG46	46 个无关个体的 MAF，无“.”表示
cg69	69 个有关个体的 MAF，无“.”表示

4.4 SNP/INDEL 非编码区域注释

非编码区（Non-coding region），不能够转录为相应信使 RNA，不能指导蛋白质合成（也就是不能编码蛋白质）的区段。非编码区虽然不能编码蛋白质，但对于遗传信息表达是不可缺少的，具有调控遗传信息的作用，启动子（promoter），增强子（enhancer）等调控元件位于非编码区。

ENCODE (Encyclopedia of DNA Elements)[13]，ENCODE 的目标是建立全面的人类基因组功能元件，包括蛋白和 RNA 水平的作用元件，以及任何细胞活动的调控元件。本分析报告使用由 UCSC 数据库汇总的 DNase Clusters[14]和 Txn Factor Chip[15]，DNase Clusters 为 DNaseI hypersensitive 的作用区域，而 Txn Factor Chip 为 transcription factor 的作用区域。

miRBase[16]序列数据库是一个提供包括 miRNA 序列数据、注释、预测基因靶标等信息的全方位数据库，是存储 miRNA 信息最主要的公共数据库之一。通过本数据库的注释提供位于 miRNA 区域的突变信息。

样品非编码区域注释统计结果见下表所示：

表 16SNP 非编码区域注释统计结果

Sample	Number
Total	58,822
DNaseI hypersensitive	22,347
Transcription factor	15,799
miRNA_primary_transcript	29
miRNA	8

（1）Total：检测到的 SNP 总数；

（2）DNaseI hypersensitive：注释到 DNase Clusters 数据库的 SNP 数；

（3）Transcription factor：注释对 Txn Factor Chip 数据库的 SNP 数；

（4）miRNA_primary_transcript：注释到 miRBase 数据库中前体 miRNA 区域的 SNP 数；

（5）miRNA：注释到 miRBase 数据库中成熟 miRNA 区域的 SNP 数。

详细的统计表见：
Report/4_ANNOTATE_SNP-INDEL/stat/sampleID.SNP.ncode.anno.stat.txt

表 17 INDEL 非编码区域注释统计结果

Sample	Number
Total	7,680
DNaseI hypersensitive	2,304
Transcription factor	2,113
miRNA_primary_transcript	2
miRNA	0

（1）Total：检测到的 INDEL 总数；

（2）DNaseI hypersensitive：注释到 DNase Clusters 数据库的 INDEL 数；

（3）Transcription factor：注释对 Txn Factor Chip 数据库的 INDEL 数；

（4）miRNA_primary_transcript：注释到 miRBase 数据库中前体 miRNA 区域的 INDEL 数；

（5）miRNA：注释到 miRBase 数据库中成熟 miRNA 区域的 INDEL 数。

详细的统计表见：
Report/4_ANNOTATE_SNP-INDEL/stat/sampleID.INDEL.ncode.anno.stat.txt

4.5 SNP 保守性预测、致病性分析

dbNSFP[17]数据库是一个针对于非同义突变注释的数据库，主要用于评估氨基酸的保守性以及致病性分析。包括 SIFT scores, PolyPhen2 HDIV scores, PolyPhen2 HVAR scores, LRT scores, MutationTaster scores, MutationAssessor score, FATHMM scores, GERP++ scores, PhyloP scores and SiPhy scores 等，不同的得分值均是基于不同的算法评估保守性和致病性。使用较普遍的为 SIFT 与 PolyPhen2，本报告主要基于这两个值进行后续的筛选，其中 HDIV 适用于分析复杂表型中的罕见突变。

表 18 SIFT 和 PolyPhen2 的保守性与致病性分析

Type	Number	Score
FATHMM_pred	7,438	797
LRT_pred	6,145	708
MetaLR_pred	7,798	87
MetaSVM_pred	7,798	77
MutationAssessor_pred	7,056	792
MutationTaster_pred	7,946	345
Polyphen2_HDIV_pred	7,416	1,574
Polyphen2_HVAR_pred	7,416	1,081
SIFT_pred	7,470	1,274

（1）Number：注释到各数库的 SNP 数；

（2）Score：分值超过有害性阈值的 SNP 数，SIFT 为 D，PolyPhen2 为 P 或 D；

详细的统计图见：
Report/4_ANNOTATE_SNP-INDEL/stat/sampleID.SNP.conservative.stat.txt

每种分值分为原始分值（score）；转化的分值（score_converted,标准化为 0-1，值越大表明越有害）；预测分类（pred）。具体分值类型及预测分类如下表：

表 19 保守性分值表

Score	Categorical Prediction
FATHMM	D: Deleterious; T: Tolerated
LRT	D: Deleterious; N: Neutral; U: Unknown
MutationAssessor	H: high; M: medium; L: low; N: neutral. H/M means functional and L/N means non-functional
MutationTaster	A" ("disease_causing_automatic"); "D" ("disease_causing"); "N" ("polymorphism"); "P" ("polymorphism_automatic"
PolyPhen 2 HDIV	D: Probably damaging (>=0.957), P: possibly damaging (0.453<=pp2_hdiv<=0.956); B: benign(pp2_hdiv<=0.452)
PolyPhen 2 HVar	D: Probably damaging (>=0.909), P: possibly damaging (0.447<=pp2_hdiv<=0.909); B: benign (pp2_hdiv<=0.446)
SIFT	D: Deleterious (sift<=0.05); T: tolerated (sift>0.05)
RadialSVM	D: Deleterious; T: Tolerated
GERP++	higher scores are more deleterious
PhyloP	higher scores are more deleterious
SiPhy	higher scores are more deleterious

dbNSFP 数据库注释结果格式说明见下表所示，由于方法较多，本表只列出部分结果示例，其他方法根据上表中对应保守性分值，结果类似。

表 20 dbNSFP 数据库注释结果格式说明

Type	Discription
SIFT_score	SIFT 预测得到的分值，无“.”表示
SIFT_pred	SIFT 根据保守性阈值判断得到的预测分类，无“.”表示
Polyphen2_HDIV_score	Polyphen2_HDIV 预测得到的分值，无“.”表示
Polyphen2_HDIV_pred	Polyphen2_HDIV 根据保守性阈值判断得到的预测分类，无“.”表示
Polyphen2_HVAR_score	Polyphen2_HVAR 预测得到的分值，无“.”表示
Polyphen2_HVAR_pred	Polyphen2_HVAR 根据保守性阈值判断得到的预测分类，无“.”表示
LRT_score	LRT 预测得到的分值，无“.”表示
LRT_pred	LRT 根据保守性阈值判断得到的预测分类，无“.”表示
MutationTaster_score	MutationTaster 预测得到的分值，无“.”表示
MutationTaster_pred	MutationTaster 根据保守性阈值判断得到的预测分类，无“.”表示

4.6 LOH 检测

LOH（Loss of Heterozygosity）是一个染色体事件，能够引起整个基因及其附近的染色体区域的丢失。

对于大多数基因座中呈现的碱基，个体与个体之间是一致的，然而一小部分基因座包含的碱基（一个或两个）在个体之间是不同的，这些位置叫做单核苷酸多态性（SNPs)，简单的说就是基因中的个别碱基在个体之间是不同的。如果父本和母本的基因组存在 SNPs，那么子代的两个等位基因（一个来源于父本，一个来源于母本）就存在这些 SNPs 的区域，那么这些区域就是杂合的（heterozygous）。然而，如果包含这样区域的其中一个亲本的基因拷贝丢失，就会导致这个区域只有一个拷贝，因此这个区域就丢失了杂合性，即 LOH。简而言之，就是某个来源于父本或母本的基因拷贝如果丢失，那么使得具有 SNP 的区域无法表现出杂合的状态。

长期的细胞遗传学的研究证实，几乎所有的肿瘤细胞都存在染色体片段的非随机性丢失（LOH），这意味着这些丢失的片段中必然包含着某些与肿瘤相关的基因。LOH 反映了遗传性改变的程度。正如 Knudson 的二次打击学说，第一次打击：其中一个抑癌基因发生了点突变而失去功能，通常发生在遗传性肿瘤综合征群中，这类人一出生就有这个突变；第二次打击为发生 LOH，比如 RB1。而其中的机制并不是简单的 deletion，而是有丝分裂重组，基因转变或者 copy-LOH（拷贝数代偿 LOH，即拷贝数不发生变化）。无论哪种机制导致 LOH，最终都是使剩下的那个抑癌基因也失去了功能，这类病人最终发生肿瘤。可见研究 LOH 有助于人类揭示肿瘤发生机制，从而指导用药。

下图展示了样本的 1 号染色体的杂合性缺失情况。

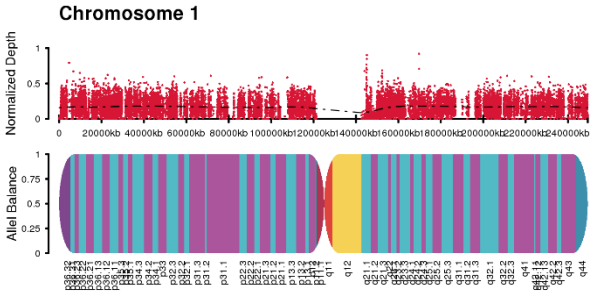

图 19 染色体上的 LOH 分布图

详细的统计图请见：
Report/5_LOH/sampleID/sampleID_LOH_chr*.tif

4.7 变异全局统览

将全基因组检测出的 SNP、InDel、SV 以及深度等信息以圆形图展示出来，便于对整体情况的掌握及对不同样本的直观对比。如下图所示样本的结果，从最外层到最内层依次是：

物种染色体信息；将各染色体按长度比例以不同颜色画出；

（1）从外往内，第一圈为染色体名称。

（2）第二圈为 SNP 密度（1Mb，曲线），并对密度进行了归一化。

（3）第三圈为测序深度，并进行归一化，小于 0.25 绿色，大于 0.75 红色，其他灰色。

（4）第四圈为 INDEL 插入（蓝色）和 INDEL 缺失（黄色）。

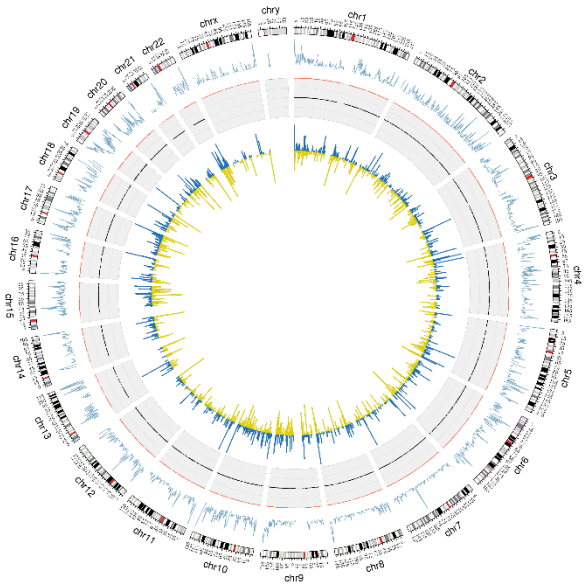

图 201 变异全局统览

详细的统计图：
Report/6_CIRCOS/sampleID.png/pdf

五、参考文献

[1] Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754-1760 (2009).
[2] BAM.http://samtools.github.io/hts-specs/SAMv1.pdf
[3] Li, H. et al. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078-2079 (2009).
[4] Picard.http://broadinstitute.github.io/picard/
[5] McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 20, 1297-1303 (2010).
[6] Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research 38, e164-e164 (2010).
[7] Refseq.http://www.ncbi.nlm.nih.gov/refseq/rsg
[8] Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. Nucleic acids research 29, 308-311 (2001).
[9] Consortium, G. P. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65 (2012).
[10] ExAC.http://exac.broadinstitute.org/
[11] Auer, P. L. et al. Imputation of exome sequence variants into population-based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. The American Journal of Human Genetics 91, 794-808 (2012).
[12] Complete Genomics.http://www.completegenomics.com/public-data/69-genomes/
[13] Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74 (2012).
[14] DNase Clusters. http://genome.ucsc.edu/cgi-bin/hgc?db=hg19&c=chr21&o=33032260&t=33033430&g=wgEncodeRegDnaseClustered&i=58&l=33032260&r=33033430
[15] Txn Factor Chip. http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19;g=wgEncodeRegTfbsClusteredV3
[16] Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic acids research 42, D68-D73 (2014).
[17] Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Human mutation 32, 894-899 (2011).