**外显子测序开发结题报告**

(KF-ZZ-XHG-20161226-01)



杭州壹基因

2017-02-10

目 录

[一、产品简介 1](#_Toc474843782)

[1. 项目简介 1](#_Toc474843783)

[2. 样品信息 1](#_Toc474843784)

[二、建库测序流程 2](#_Toc474843785)

[1. DNA样品质检 2](#_Toc474843786)

[2. 文库构建 2](#_Toc474843787)

[3. 库检 3](#_Toc474843788)

[4. 上机测序 3](#_Toc474843789)

[三、生物信息分析流程 4](#_Toc474843790)

[四、基本数据分析 5](#_Toc474843791)

[1. 下机数据 5](#_Toc474843792)

[2. 测序数据质量评估 6](#_Toc474843793)

[2.1 原始测序数据 6](#_Toc474843794)

[2.2 数据过滤 7](#_Toc474843795)

[2.3 测序深度、覆盖度统计 8](#_Toc474843796)

[五、胚系变异检测及分析 9](#_Toc474843797)

[1. Germline SNP检测分析 9](#_Toc474843798)

[2. Germline InDel检测分析 10](#_Toc474843799)

[3. Germline CNV检测分析 11](#_Toc474843800)

[六、Germline mutation注释 12](#_Toc474843801)

[七、高级分析 13](#_Toc474843802)

[1. 易感位点筛查 13](#_Toc474843803)

[2. 突变图谱 15](#_Toc474843804)

[3. SNP所在基因KEGG通路富集分析 16](#_Toc474843805)

[4. 高频CNV分析 18](#_Toc474843806)

[5. 融合基因检测 21](#_Toc474843807)

[6. 已知驱动基因筛选 22](#_Toc474843808)

[附录 24](#_Toc474843809)

[1. 结果文件目录结构 24](#_Toc474843810)

[2. 注释结果格式说明 24](#_Toc474843811)

[3. 项目信息 24](#_Toc474843812)

# 一、产品简介

在人类基因中大约有180,000个外显子，占人类基因组的1%，约30MB。人类基因组的蛋白编码区大约包含85%的致病突变。外显子组测序是指利用序列捕获技术将探针能覆盖到的全基因组外显子区域DNA捕获富集后再进行高通量测序的基因组分析方法。与全基因组重测序相比，外显子组测序费用较低，数据阐释也更为简便，极大地提高了人基因组中外显子区域的研究效率，显著降低研究成本。

## 1. 项目简介

本项目对3个样本进行外显子组测序，进行相应的生物信息学分析，确定遗传变异对个体表型特征的影响，为后续相关研究提供数据支撑。

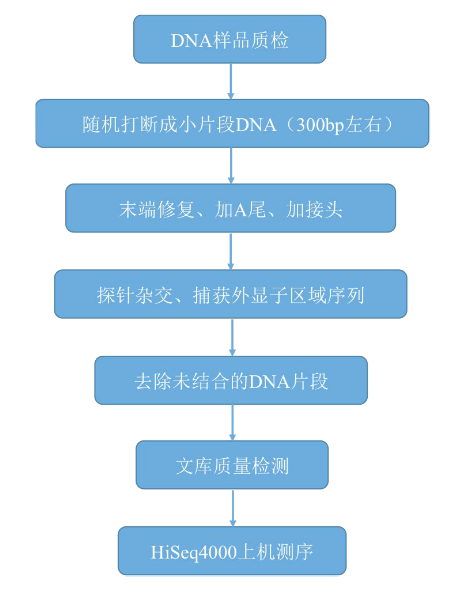
## 2. 样品信息

采用Agilent Sureselect（60M）外显子芯片，基于Illumina测序平台，对样本进行测序，每个样品产生10G的有效数据量。

|  |  |
| --- | --- |
| **表1.1 样本信息表** | |
| **No.** | **Sample ID** | |
| 1 | WGC100873 | |
| 2 | WGC100874 | |
| 3 | WGC100875 | |

# 二、建库测序流程

样品检测、建库、测序每一个环节都会对数据质量和数量产生影响，而数据质量又会直接影响后续信息分析的结果。因此，获得高质量数据是保证生物信息分析正确、全面、可信的前提。为了从源头上保证测序数据的准确性、可靠性，我们对样品检测、建库、测序每一个生产步骤都严格把控，从根本上确保了高质量数据的产出。流程图如下：



**图2.1 建库测序流程图**

## 1. DNA样品质检

壹基因对肿瘤外显子测序DNA样品的检测主要包括 3 种方法：

（1） Nanodrop 检测 DNA 的纯度（OD 260/280比值）。

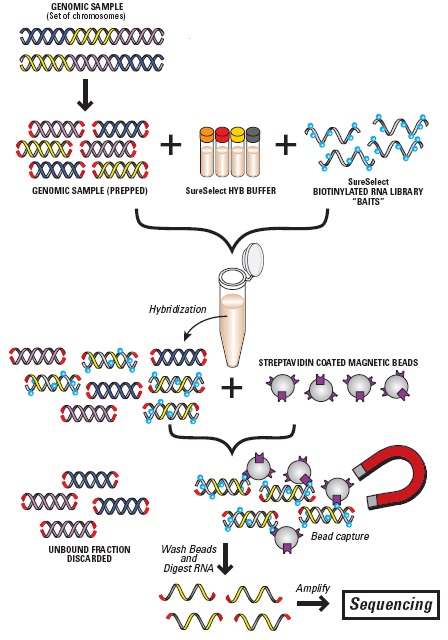
（2） 琼脂糖凝胶电泳分析 DNA 降解程度以及是否有 RNA、蛋白质污染。

（3） Qubit 对 DNA 浓度进行精确定量。其中 OD 值在 1.8~2.0 之间，含量在 1.0 μg 以上的 DNA 样品被用来建库。

## 2. 文库构建

采用Agilent SureSelect Human All Exon V6 的液相芯片捕获系统，对人全外显子区域 DNA进行高效富集，然后在Illumina平台上进行高通量、高深度测序。建库和捕获均严格使用说明书推荐的试剂和耗材，并参照最新优化实验流程进行操作。

实验基本流程：将基因组 DNA 经Covaris破碎仪随机打断成长度为180-280bp的片段， 经末端修复和加 A 尾后在片段两端分别连接上接头制备 DNA文库。带有特异index的文库与生物素标记的探针进行液相杂交，再使用带链霉素的磁珠捕获，洗脱得到基因组中的外显子片段，经PCR线性扩增后进行文库质检，检测合格后进行上机测序。



**图2.2文库构建及捕获流程（Agilent Sureselect）**

## 3. 库检

文库构建完成后，先使用 Qubit 2.0 进行初步定量，随后使用 Agilent 2100 对文库的insert size 进行检测，insert size符合预期后，使用qPCR方法对文库的有效浓度（2 nM）进行准确定量，以保证文库质量。

## 4. 上机测序

根据库检合格文库的有效浓度及数据产出需求进行 Illumina HiSeq4000 平台PE150测序。

PE150（Pair end 150 bp）指高通量双端测序，每端各测150 bp。双端测序是将每条插入片段的两端进行测序的方法，由于插入片段的长度分布已知，双端测序时不仅可以知道片段两端的序列，也能知道这两段序列之间的长度，从而便于后续组装和比对。

# 三、生物信息分析流程

测序结束获得原始测序序列（Raw Data）后，进入信息分析流程，流程分为两个阶段：

1） 测序数据质量评估：主要通过对测序错误率、数据量、比对率等进行统计，评估建库测序是否达到了标准，符合标准则进行后续的分析，否则需重新建库或加测；

2） 变异信息挖掘及分析：将高质量的序列比对到人参考基因组上，检测样本中的变异信息，并对检出的变异进行分析解读。信息分析流程如下：

**测序原始数据**

**数据质量评估**

**与参考基因组比对分析**

**SNV/InDel/CNV检测注释**

**高级信息分析**

**易感基因筛查 突变频谱 通路富集分析 驱动基因 高频CNV计算 融合基因检测**

**图3.1 信息分析流程示意图**

# 四、基本数据分析

## 1. 下机数据

高通量测序（如Illumina HiSeqTM 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。

（注：reads，是一个测序反应能测得的碱基序列。它的长度和测序策略有关，比如Illumina的PE150测序，就是对一条DNA序列测序，得到两端序列，每端得到的就是一个长度为150 bp 的 read）。

FASTQ格式文件中每个 read 由四行描述，如下：

@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG

GCTCTTTGCCCTTCTCGTCGAAAATTGTCTCCTCATTCGAAACTTCTCTGT

+

@@CFFFDEHHHHFIJJJ@FHGIIIEHIIJBHHHIJJEGIIJJIGHIGHCCF

其中第一行和第三行由文件识别标志（sequence identifiers）和读段名（ID）组成（第一行以"@"开头，第三行一般只有“+”，ID可以省略），第二行是碱基序列，第四行为对应第二行碱基的测序质量。 Illumina 测序文件识别标志中的详细信息如下表：

**表4.1 Illumina 测序标识符详细信息**

|  |  |
| --- | --- |
| **HWI-ST1269** | **Unique instrument name** |
| 432 | Run ID |
| H9MN7ADXX | Flowcell ID |
| 1 | Flowcell lane |
| 1101 | Tile number within the flowcell lane |
| 13406 | 'x'-coordinate of the cluster within the tile |
| 2064 | '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 |
| CGGATA | Index sequence |

碱基测序准确率是通过测序Phred数值(Phred score, Qphred)得到，而Phred 数值是在碱基识别(Base Calling)过程中通过一种预测碱基判别发生错误概率模型计算得到的。如果测序错误率用e表示，Illumina测序的碱基质量值用Q表示，则有下列关系：

**Q=-10log10(e)**

此公式可说明，质量值越大测序错误率越低，准确性越高。每个碱基测序准确率是通过测序Phred数值(Phred score, Qphred)得到，而Phred数值是在碱基识别(Base Calling)过程中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：

**表4.2 碱基识别准确率与Phred分值之间的简明对应表**

|  |  |  |
| --- | --- | --- |
| **碱基正确识别率** | **Phred分值** | **测序质量值** |
| 90% | 10 | Q10 |
| 99% | 20 | Q20 |
| 99.9% | 30 | Q30 |
| 99.99% | 40 | Q40 |

注：测序准确率与碱基质量有关，受测序仪本身、测序试剂、样品等多个因素共同影响。

## 2. 测序数据质量评估

### 2.1 原始测序数据

测序数据的产生经过了 DNA 提取、建库、测序多个步骤，这些步骤会产生少许低质量或者无效数据，比如建库阶段会存在建库长度的偏差，测序阶段会出现测序错误的情况。所以需要对下机得到的原始数据进行质量控制，保证后续分析的准确性。测序下机的原始数据的质量信息统计如下:

**表4.3 原始数据质量统计**

| **Sample** | **Size(Gb)** | **G+C(%)** | **N(%)** | **>Q30 (%)** |
| --- | --- | --- | --- | --- |
| WGC100873 | 21 | 52.38 | 0 | 91.86 |
| WGC100874 | 22.2 | 52.55 | 0 | 91.84 |
| WGC100875 | 22.8 | 52.46 | 0 | 91.98 |

注:

(1) Size: 统计原始序列reads总数目；根据 FASTQ 的格式，以四行为一个单位进行统计。

(2) G+C(%): 计算碱基 G 和 C 的数量总和占总的碱基数量的百分比。

(3) N(%): 计算所有reads中包含N的百分比。

(4) >Q30: 计算 Phred 数值大于 30 的碱基占总体碱基的百分比。测序数据质量分布，测序数据的质量主要分布在 Q30(≥90%)以上，这样才能保证后续分析的正常进行。根据测序技术的特点，测序片段末端的碱基质量一般会比前端的低。

### 2.2 数据过滤

测序得到的原始测序数据中会有少量reads包含接头信息、低质量碱基或未检出的碱基，对后续信息分析造成很大干扰。为了保证信息分析质量，必须对raw reads进行精细过滤，得到clean reads，后续分析都基于clean reads。

数据过滤主要包括成对去除以下三种情况的reads及序列：

1. 含接头序列的reads（adapter）；
2. 单端测序read中 N（N 表示无法确定碱基信息）的碱基个数超过该条read碱基总数的5%的 reads；
3. 单端测序read中低质量（质量值低于 30）碱基数超过该条read 长度比例的 50%的reads；

DNA-Seq 接头信息:

* F: GATCGGAAGAGCACACGTCTGAACTCCAGTCAC
* R: AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA

**表4.4 数据过滤情况统计**

| **Sample** | **Adapter Filter (%)** | **Low Qual Filter (%)** | **Undersize Ins Filter (%)** | **Duplicated Filter (%)** | **Size(Gb)** | **>Q30(%)** |
| --- | --- | --- | --- | --- | --- | --- |
| WGC100873 | 29.5698 | 5.10 | 0 | 5.58 | 12.52 | 93.49 |
| WGC100874 | 28.3823 | 5.07 | 0 | 5.99 | 13.46 | 93.44 |
| WGC100875 | 32.8877 | 4.63 | 0 | 5.99 | 12.86 | 93.45 |

注:

(1) Adapter Filter: 因接头而过滤掉的reads数占总reads数的百分比。

(2) Low Qual Filter: 因一条 read 中低质量（ 质量值低于 5）碱基数超过该条 read 长度比例 50%的reads 数占总 raw reads 数的百分比。

(3) Undersize Ins Filter: 因一对 read 间距离短于测序插入片段长度的read 数占总 raw reads 数的百分比。

(4) Duplicated Filter: 因 PCR扩增造成的重复序列read数占总 raw reads 数的百分比。

(5) Size: 统计过滤后序列reads总数目。

(6) >Q30: 计算 Phred 数值大于 30 的碱基占总体碱基的百分比。

### 2.3 测序深度、覆盖度统计

有效测序数据通过 BWA 软件比对到人类参考基因组 (hg19 版本)，得到 BAM 格式的最初的比对结果。再对比对结果进行覆盖度、深度等的统计和捕获效率的统计。通常，人类样本的测序 reads 能达到 95% 以上的比对率；当一个位点的碱基覆盖深度（read depth）达到10 X以上时，则认为该位点处检测出的SNP比较可信。

**表4.5 比对结果**

| **Sample** | **Mapping ratio** | **Capture depth** | **Uniformity** | **Specificity** | **Coverage** |
| --- | --- | --- | --- | --- | --- |
| WGC100873 | 0.9972 | 222 | 0.9992 | 0.8292 | 0.9882 |
| WGC100874 | 0.9980 | 238 | 0.9993 | 0.8266 | 0.9883 |
| WGC100875 | 0.9978 | 246 | 0.9993 | 0.8369 | 0.9885 |

注:

(1) mapping ratio: 比对率, 即测序数据成功比对到参考基因组上的碱基数占总数据量的比例。

(2) capture depth: 捕获区域平均深度，即 capture base / capture len。

(3) uniformity: 均一性，捕获区域中深度大于平均深度10%的碱基数量与捕获区域内所有碱基数量之比。

(4) specificity: 特异性，捕获效率，捕获区域内深度大于平均深度10%的碱基数量与所有碱基数量之比。

(5) coverage: 覆盖度，比对上捕获区域且深度大于平均深度10%的区域与总捕获区域之比。

# 五、胚系变异检测及分析

胚系突变（germline mutation）又叫生殖细胞突变，是来源于精子或卵子这些生殖细胞的突变，因此通常身上所有细胞都带有突变。胚系突变可以遗传，这也是有血缘关系的亲戚，总是或多或少有些像的原因。

## 1. Germline SNP检测分析

SNP全称 Single nucleotide variant,，是指在基因组上由单个核苷酸的替换所引起的变异。我们主要使用 FreeBayes 软件来寻找Germline SNP位点，根据人类参考基因集 hg19 版本使用 snpEff 软件进行类型注释。胚系突变检测在各基因功能元件上的分布如下表所示(下表仅为示例，具体每个样本的 SNP 变异结果见附件文件):

**表5.1 SNP在各基因功能元件上的分布统计**

| **Feature Type** | **Count** |
| --- | --- |
| DOWNSTREAM1 | 16,733 |
| EXON2 | 44,076 |
| INTERGENIC3 | 182,252 |
| INTRON4 | 333,782 |
| SPLICE\_SITE\_ACCEPTOR5 | 69 |
| SPLICE\_SITE\_DONOR6 | 87 |
| SPLICE\_SITE\_REGION7 | 3,873 |
| UPSTREAM8 | 20,041 |
| UTR\_3\_PRIME9 | 9,211 |
| UTR\_5\_PRIME10 | 5,713 |

注：

(1) DOWNSTREAM: 基因下游（5k bases）。

(2) EXON: 外显子。

(3) INTERGENIC: 基因间。

(4) INTRON: 内含子。

(5) SPLICE\_SITE\_ACCEPTOR: 剪切位点接受子。

(6) SPLICE\_SITE\_DONOR: 剪切位点供应子。

(7) SPLICE\_SITE\_REGION: 剪切位点区域。

(8) UPSTREAM: 基因上游（5k bases）。

(9) UTR\_3\_PRIME: 3’端非编码区。

(10) UTR\_5\_PRIME: 5’端非编码区。

## 2. Germline InDel检测分析

编码区或剪接位点处发生的插入缺失都可能会改变蛋白的翻译。移码变异，其插入或缺失的碱基串的长度为 3 的非整数倍，因此可能导致整个阅读框的改变；与非移码变异比较，移码突变对基因功能的影响更大，同时也受到更大的筛选压力。我们利用 FreeBayes 软件检测 Germline InDel 信息，根据人类参考基因集 hg19 版本使用 snpEff 软件进行类型注释。检测到的 InDel 在各基因功能元件上的分布统计如下(下表仅为示例，具体每个样本的 InDel 变异结果见附件)：

**表5.2 InDel在各基因功能元件上的分布统计**

| **Feature Type** | **Count** |
| --- | --- |
| DOWNSTREAM1 | 2,216 |
| EXON2 | 1,398 |
| INTERGENIC3 | 19,011 |
| INTRON4 | 43,157 |
| SPLICE\_SITE\_ACCEPTOR5 | 68 |
| SPLICE\_SITE\_DONOR6 | 73 |
| SPLICE\_SITE\_REGION7 | 582 |
| UPSTREAM8 | 2,687 |
| UTR\_3\_PRIME9 | 1,252 |
| UTR\_5\_PRIME10 | 748 |

注：

(1) DOWNSTREAM: 基因下游（5k bases）。

(2) EXON: 外显子。

(3) INTERGENIC: 基因间。

(4) INTRON: 内含子。

(5) SPLICE\_SITE\_ACCEPTOR: 剪切位点接受子。

(6) SPLICE\_SITE\_DONOR: 剪切位点供应子。

(7) SPLICE\_SITE\_REGION: 剪切位点区域。

(8) UPSTREAM: 基因上游（5k bases）。

(9) UTR\_3\_PRIME: 3’端非编码区。

(10) UTR\_5\_PRIME: 5’端非编码区。

## 3. Germline CNV检测分析

拷贝数变异 ( Copy number variation,CNV ) 表现为基因组上大片段序列的拷贝数增加或减少，是基因组结构变异（Structural variation, SV）的重要组成部分，可分为缺失（Deletion）和重复（Duplication）两种类型，是一种重要的分子机制。CNV能够导致癌症在内的复杂疾病，对于染色体水平的缺失、扩增的研究已经成为肿瘤研究热点。我们基于比对到参考基因组上reads的深度分布情况，采用CNVkit软件检测各样本的拷贝数变化，得到如下的统计结果（具体每个样本的 CNV 变异结果见附件)：

**表5.3 CNV的统计结果**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **gene** | **chr** | **start** | **end** | **copy\_number** | **cnv\_type** |
| PRAMEF12 | 1 | 12837530 | 12837822 | 4.37897 | amplification |

注：

(1) gene: 基因名称。

(2) chr: 染色体。

(3) start: 起始位置。

(4) end: 终止位置。

(5) copy\_number: 拷贝数变异数（<2表示缺失；>2表示重复）。

(6) cnv\_type: 拷贝数变异类型。

# 六、Germline mutation注释

我们进一步使用 Oncotator，GEMINI 等注释软件，调用目前公开的主流数据库（COSMIC、dbSNP、1000Genome等）对检测出的体细胞突变进行详细注释。

1）使用 Refseq、Gencode、Ensemble 等数据库注释变异位点的基因结构，功能类型包括 mRNA、非编码RNA、smallRNA 和 microRNA 等，类型信息包括基因名称、基因ID、转录本ID、碱基突变内容、密码子突变内容、rs ID等；

2）我们调用 SIFT、PolyPhen2、MutationTaster、MutationAssessor、保守区域、LOF、CADD、GERP等数据库信息进行分值输出，从而可评估突变位点的影响；

3）我们还提供了 dbSNP、千人基因组 SNP 数据库 (1000Genome)、CG 数据库、 Hapmap 数据库、Clinvar疾病数据库、COSMIC疾病数据库、OMIM 疾病数据库、HGMD 疾病数据库、CGC 肿瘤数据库，以及 esp6500变异数据库、exac03 突变数据库等的对应注释结果，对变异结果可以进行任何组合的筛选。

**表6.1 数据库注释结果示例**

| **Chr** | **Start** | **End** | **Ref** | **Alt** | **Func.ensGene** | **GeneDetail.ensGene** | **…** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 665058 | 665058 | C | G | intergenic | Dist=1531;dist=12135 | … |

注：

(1) Chr: 变异所在染色体。

(2) Start: 起始位点。

(3) End: 终止位点。

(4) Ref: 参考基因组碱基型。

(5) Alt: 样本基因组碱基型。

(6) Func.ensGene: 对变异位点所在的区域进行注释（ensGene: FASTA sequences for all annotated transcripts in ENSEMBL Known Gene）。

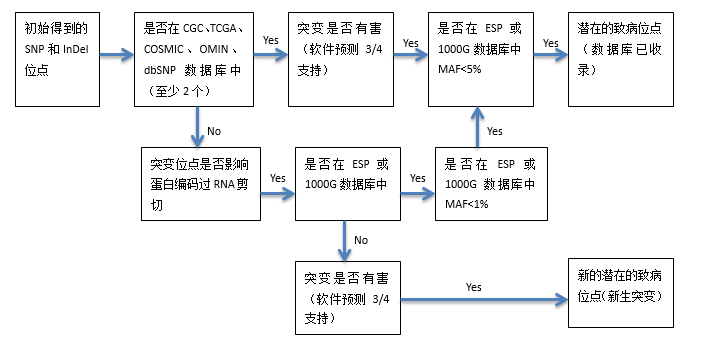
(7) GeneDetail.ensGene: 描述UTR、splicing、ncRNA\_SPLICING或Intergenic区域的变异情况（dist：表示该变异位点距离两侧基因的距离）。

（由于注释信息过多，没有全部展示，如果需要详细注释信息内容和说明，可参照附件文档）

# 七、高级分析

## 1. 易感位点筛查

易感基因（Predisposing gene）是在适宜的环境刺激下能够编码遗传性疾病或获得疾病易感性的基因。我们对于易感基因的筛查流程（见下图）如下：首先过滤测序深度低于10X的位点，其次检索常用数据库（包括dbSNP、1000G、cg、clinvar、cosmic、ESP外显子数据库），再次判断突变位点是否影响蛋白编码和RNA剪切，后续进行等位基因频率过滤（ESP或1000G），最后通过SIFT、Polyphen、MutationTaster、CADD 4种方法评估突变是否有害，最终得到已知的和新生的潜在致病位点。



**图7.1 易感位点筛选流程**

根据上述的筛选条件，得到各样本的易感位点，结果如下（具体每个样本的结果见附件）：

**表7.1 易感位点筛选结果示例**

| **Name** | **Content** |
| --- | --- |
| Gene | AGRN |
| Exonic|Biotype | MISSENSE |
| Chr | 1 |
| Start | 981139 |
| End | 981139 |
| Ref | G |
| Alt | A |
| Func.ensGene | exonic |
| Gene.ensGene | ENSG00000188157 |
| GeneDetail.ensGene | dist=NONE;dist=1831 |
| SIFT\_score | 0.019 |
| SIFT\_pred | D |
| Polyphen2\_HVAR\_score | 0.894 |
| Polyphen2\_HVAR\_pred | P |
| MutationTaster\_score | 1 |
| MutationTaster\_pred | D |
| CADD\_raw | 6.412 |
| CADD\_phred | 29.7 |

注：

(1) Gene:基因名字。

(2) Exonic|Biotype: 突变类型。

(2) Chr: 染色体编号。

(3) Start: 检测到的变异的起始位置。

(4) End：检测到的变异的终止位置。

(5) Ref：参考序列POS上的碱基。

(6) Alt: 发生突变的碱基。

(7) Func.：对变异位点所在的区域进行注释（ensGene: FASTA sequences for all annotated transcripts in ENSEMBL  Known Gene）。

(8) Gene.：对变异位点所在的基因进行注释（ensGene:同上）。

(9) GeneDetail.：描述UTR、splicing、ncRNA\_SPLICING或intergenic区域的变异情况（ensGene：同上；dist：表示该变异位点距离两侧基因的距离）。

(9) SIFT\_score: SIFT分值, 表示该变异对蛋白序列的影响。分值越小越可能“有害”, 表明该SNP导致蛋白结构或功能改变的可能性大。

(10) SIFT\_pred：预测结果, 取值为T或者D。当该变异同时影响多个蛋白序列时，对每条蛋白序列有一个SIFT值，取最小值。D: Deleterious (sift<=0.05)；T: tolerated (sift>0.05)  。

(11) Polyphen2\_HVAR\_score: 利用PolyPhen2基于HumanVar数据库预测该变异对蛋白序列的影响，用于孟德尔遗传病的诊断。Polyphen2\_HVAR\_score是PolyPhen 2分值，数值越大越可能“有害”，表明该SNP导致蛋白结构或功能改变的可能性大。

(12) Polyphen2\_HVAR\_pred: 预测结果，取值为D或P或B。D: Probably damaging (>=0.909)；P: possibly damaging (0.447<=pp2\_hvar<=0.909)；B: benign (pp2\_hvar<=0.446)。

(13) MutationTaster\_score: MutationTaster预测结果，表示该变异对蛋白序列的影响。MutationTaster\_score是MutationTaster分值，取值为 0-1，分值越大，越可能是deleterious。

(14) MutationTaster\_pred: 预测结果，取值为 A、D、N 或者 P。"A" ("disease\_causing\_automatic")；"D" ("disease\_causing")；"N" ("polymorphism")；"P" ("polymorphism\_automatic")。D和N是 MutationTaster根据score进行分类的结果，A和P是利用score结合其他信息得到的(如果 nonsynonymous SNV导致stop-gain, 则被预测为A；如果 nonsynonymous SNV的三个基因型都在 HapMap中有频率，则被预测为 P)。也就是说，MutationTaster的分类结果不仅仅依赖于score值，还会结合其它信息。因此,，A和D都属于deleterious。

(15) CADD\_raw: CADD是一种对SNV、InDel的有害性进行打分的工具，它整合多种信息来注释变异位点的功能；不仅预测编码区变异(包括同义突变和非同义突变的影响)的功能影响，还预测非编码区变异的功能影响。对于SNP，仅对CADD分值排名在前10%的SNP给出分值，“.”表示CADD分值排名不在前10%。

(16) CADD\_phred：CADD\_Phred是转换后的分值；10表示score排名在前10%，20表示前1%，30表示前0.1%，因此，分值要求越低，能保留下来的位点越多。“.”表示CADD\_Phred值小于10。

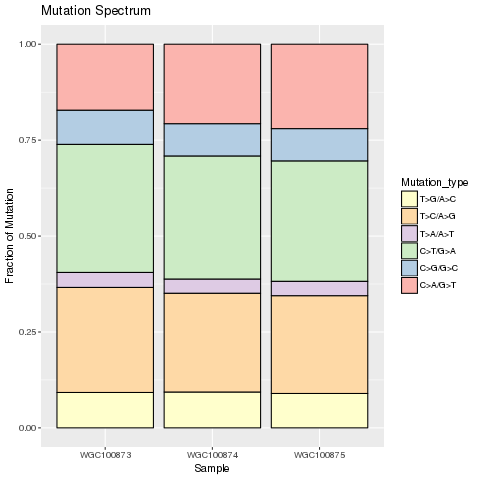
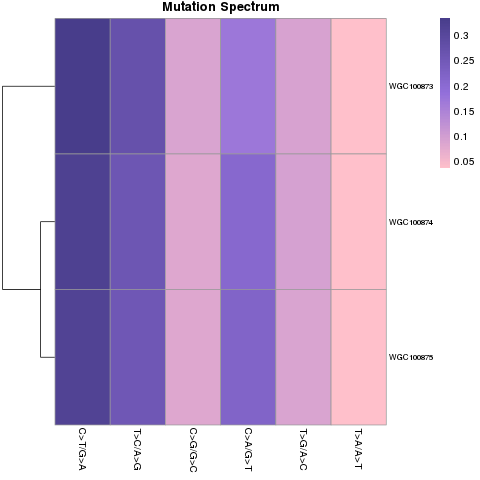
每个样本筛选出的已知和新生的易感位点统计如下：

**表7.2 易感位点的数量统计**

| **Sample** | **潜在致病位点（已知的）** | **新生的潜在致病位点** | |
| --- | --- | --- | --- |
| WGC100873 | 326 | 997 |
| WGC100874 | 357 | 1,281 |
| WGC100875 | 339 | 1,276 |

## 2. 突变图谱

单个碱基的替换一共有6种变异方式：C>A/G>T, C>G/G>C, C>T/G>A, T>A/A>T, T>C/A>G, T>G/A>C，又可以根据发生替换的碱基类别分为两大类：嘌呤和嘧啶之间的替换称为颠换（Transversion）；嘌呤与嘌呤或者嘧啶与嘧啶之间的替换称为转换（Transition）。根据各类型点突变的数量，对肿瘤样本和点突变类型进行聚类分析，我们可以分析出所研究癌症的点突变类型的偏好和各肿瘤样本在点突变水平上的相似程度。分析结果如下：

**图7.2 胚细胞SNV突变频谱柱状图（左图）和热图（右图）**

注：左图表示在不同样本中，不同突变类型出现的频率分布。横坐标为样本名称，纵坐标是各突变类型在样本中的比例，不同颜色代表不同SNV突变类型，可以直观看出，在这些样本中代表A到G的橘色和G到A的绿色占主要部分，说明突变类型最多的为A到G和G到A的颠换。右图为突变频谱的热度图，每一行代表一个样本，每一列代表一种突变类型，颜色越深代表该突变类型在样本中所占比例越高，可以通过样本的突变频谱对样本进行聚类。目前的样本量相对较少，有待进一步扩充，以得到更可靠的结果。

## 3. SNP所在基因KEGG通路富集分析

将snpeff软件检测出的变异所在基因在KEGG (Kyoto Encyclopedia of Genes and Genomes) 数据库中检索，筛选出变异所在基因所在的通路。结果包含两部分内容（仅以某一条代谢路径为例，具体的结果详见附件文件）：

1. 样本中筛选出的所有通路与该通路中包含的变异所在基因列表；
2. 样本中每一条通路的展示图。

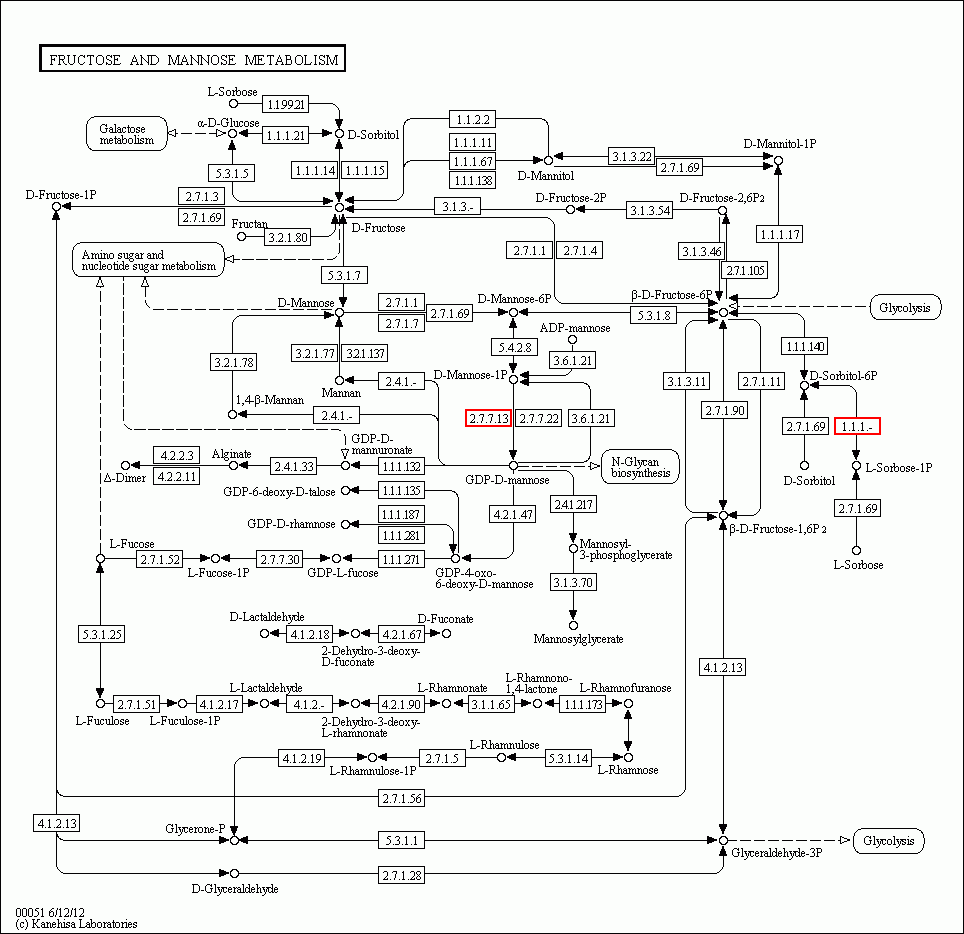
**表7.3 基因的代谢路径和功能注释示例**

| **Name** | **Content** |
| --- | --- |
| Pathway | K04070|Phosphatidylinositol signaling system!K01100|Metabolic pathways!K04144|Endocytosis!K04810|Regulation of actin cytoskeleton !K00562|Inositol phosphate metabolism |
| Gene | ENST00000581097.1|ENSG00000141720.7|OTTHUMG00000133119.2|OTTHUMT00000442251.1|PIP4K2B-002|PIP4K2B|547|UTR5:1-333|CDS:334-547|;ENST00000604912.1|ENSG00000150867.9|OTTHUMG00000017810.3|OTTHUMT00000468421.1|PIP4K2A-007|PIP4K2A|443|UTR5:1-17|CDS:18-443| |

注：

(1) Pathway: KEGG 数据库注释得到的 pathway 显著性富集，以确定基因参与的最主要生化代谢途径和信号转导途径。

(2) Gene：上述通路中所包含的本样本检测出的变异所在基因（多个基因以分号隔开）。



**图7.3 KEGG通路富集分析展示图**

注：上图展示了一条KEGG通路(\*.png)，红色框内为该基因在KEGG数据库中的ID，在网页版的结果（\*.html）中，鼠标悬停时可显示该基因的详细信息，点击时可跳转至官网对应页面。

我们还对3个样本检测出的变异所在基因取交集，最终找到共有的变异所在基因，并筛选出这些基因的代谢通路及注释（详见附件文件）。

**表7.4 变异所在基因和代谢通路的统计**

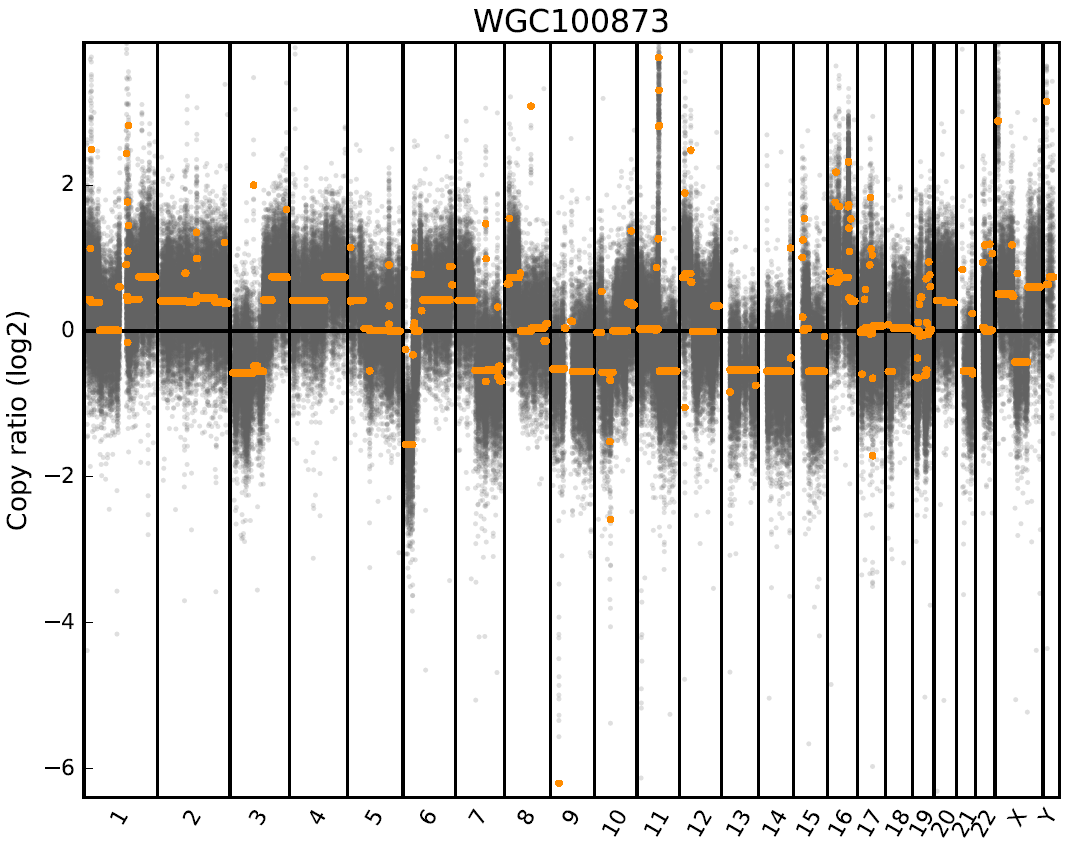
| **Sample** | **Gene(count)** | **Pathway(count)** |
| --- | --- | --- |
| WGC100873 | 6,515 | 1,179 |
| WGC100874 | 6,775 | 1,229 |
| WGC100875 | 6,805 | 1,208 |
| Intersection | 5,343 | 643 |

注：

(1) Intersection: 所有样本取交集。

## 4. 高频CNV分析

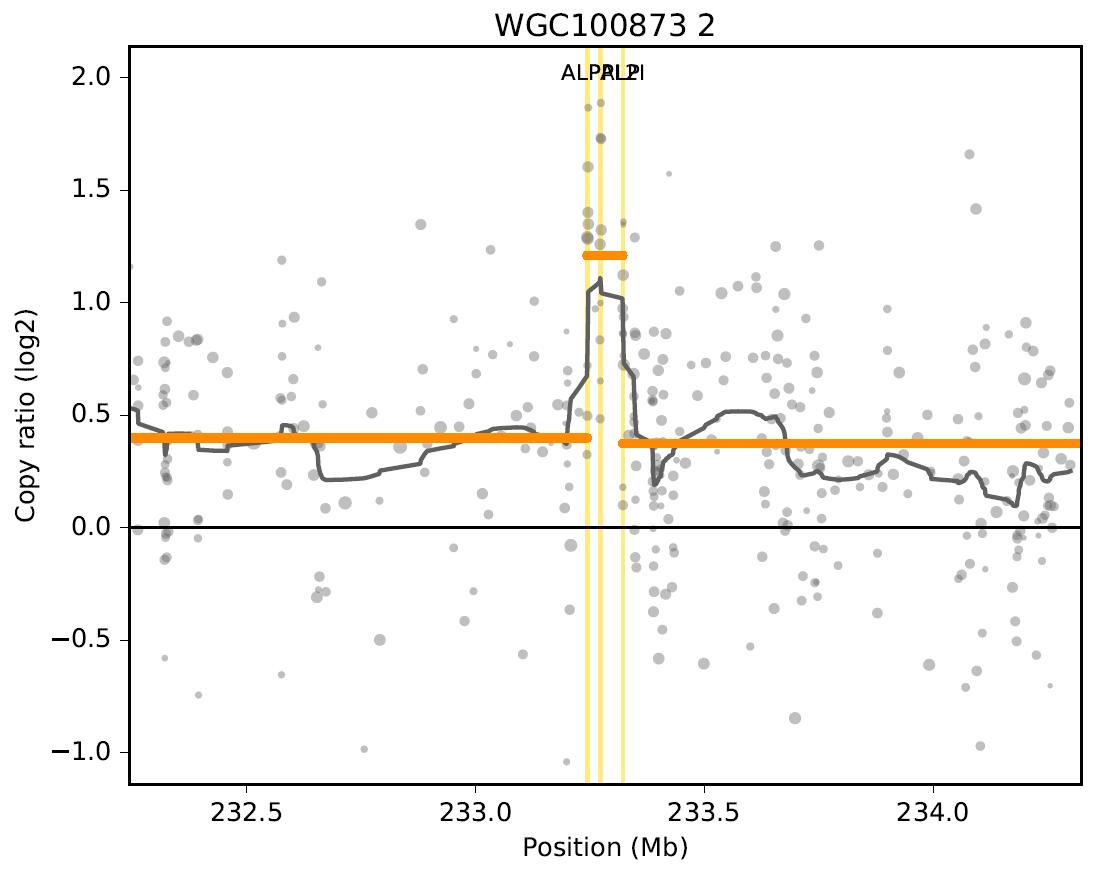
对肿瘤而言，缺失片段可能包含抑癌基因，扩增片段可能存在癌基因。我们使用CNVkit进行CNV的分析，得到以下结果（仅以WGC100873号样本为例，各样本结果见附件）：



**图7.4 CNV在染色体上的分布**

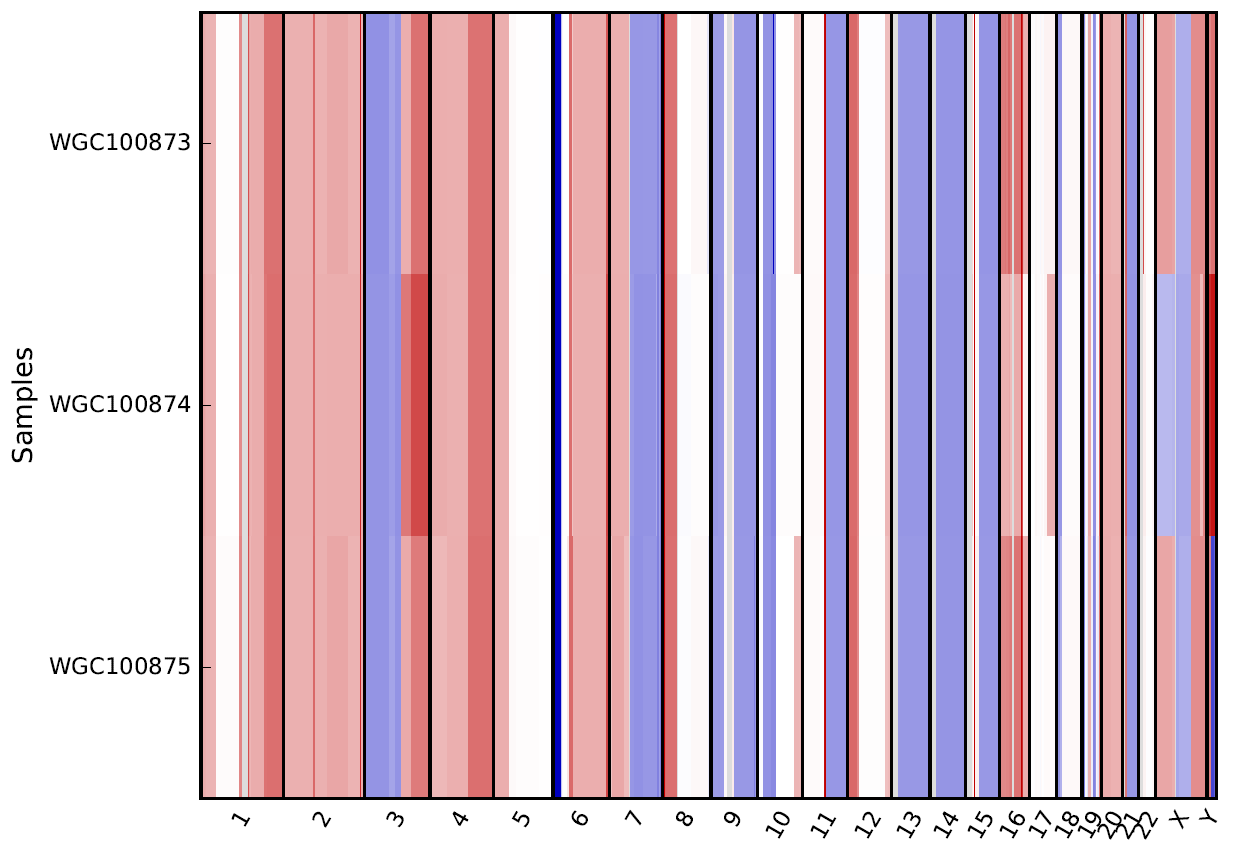
注：上图为CNV在染色体上的分布情况，横坐标为1-22号和X、Y染色体，纵坐标为拷贝数变异的log2值，黑色的点表示真实数据在每个区域（bin）的拷贝数取log2后的值，橙色的点表示在一定范围内富集的黑点拟合后的值，点的大小表示可信度，相对小的点表示较不可靠的bin。以0作为分隔线，大于0的点表示拷贝数增加，小于0的点表示拷贝数减少，围绕在分割线附近的点可视为噪声和不确定性的水平。

我们将每条染色体上的高频CNV所在基因标注如下（仅以一个样本中的一条染色体的结果为例，具体结果见附件）：



**图7.5 2号染色体上的高频CNV所在基因分布图**

注：上图为WGC100873样本在2号染色体上的高频CNV所在基因分布图，横坐标为染色体上的位置，纵坐标为拷贝数变异的log2值，黄色线内的区域为高频cnv所在区域，黄线上部标注了此区域所在的基因名称。



**图7.6 CNV在样本中的分布图**

注：图7.6为CNV在各样本中的分布图，横坐标为1-22号和X、Y染色体，纵坐标为样本编号，红色代表拷贝数增加，蓝色代表拷贝数减少，颜色越深，拷贝数变动的幅度越大。

## 5. 融合基因检测

基因的融合在癌症中非常普遍，也是一些癌症类型的重要特征。一个强启动子与一个下游功能基因（原癌基因）的融合在某些癌症中很普遍。基因融合由两个原本分开的基因或位点融合形成，通常表现为两个或多个基因的编码区首尾相连，至于同一套调控序列（包括启动子、增强子、核糖体结合序列、终止子等）控制之下，构成嵌合基因；融合基因的表达产物为融合蛋白，其功能可能是全新的、与两个融合伙伴都不同的，从而引起基因表达失控导致致癌激活。我们利用factera软件检测样本中的融合位点，结果如下（下表仅为示例，具体结果见附件）：

**表7.5 变融合基因检测结果**

| **Name** | **Content** |
| --- | --- |
| Est\_Type | TRA |
| Region1 | FTH1 |
| Region2 | UBE2E3 |
| Break1 | 11:61735071 |
| Break2 | 2:181738261 |
| Break\_support1 | 42 |
| Break\_support2 | 44 |
| Order1 | NC |
| Order2 | CN |

注：

(1) Est\_Type：结构变异类型（TRA：融合；INV：倒置；DEL：删失；“-”：未确定）。

(2) Region1：最接近断裂点1的基因组区域名称。

(3) Region2：最接近断裂点2的基因组区域名称。

(4) Break1：断裂点1所在的染色体位置。

(5) Break2：断裂点2所在的染色体位置。

(6) Break\_support1：支持断裂点1的reads数。

(7) Break\_support2：支持断裂点2的reads数。

(8) Order1：断裂点1上剪切的reads的方向（CN：剪切的片段在右侧，即断点左侧的片段与其他片段发生融合；NC：剪切的片段在左侧，即断点右侧的片段与其他片段发生融合）。

(9) Order2：断裂点1上剪切的reads的方向（CN：同上；NC：同上）。

## 6. 已知驱动基因筛选

通常癌细胞中的突变分为两种类型：一种对肿瘤的增值扩散有重要影响，使肿瘤获得选择性的生长优势，这样的突变称为驱动突变（Driver mutation）；另外一种对肿瘤增值扩散影响很小或者没有影响的突变称为乘客突变（Passenger mutation）。 检测驱动突变和驱动基因可以帮助我们进一步理解肿瘤形成的分子机理，从而开发更有效的药物来治疗癌症。我们将检测到的变异与已知驱动基因进行比较，筛选出该肿瘤样本中的已知驱动基因。已知驱动基因的来源主要有4个方面：（1）[CGC609](http://cancer.sanger.ac.uk/census)：Cancer Gene Census 列表里所列出的驱动基因；（2）[Bert Vogelstein125](http://www.sciencemag.org/content/339/6127/1546.full)：Bert Vogelstein的论文里125个mut-driver gene；（3）[SMG127](http://www.nature.com/nature/journal/v502/n7471/abs/nature12634.html)：pan-cancer的数据找出的Significantly Mutated Gene；（4）[Comprehensive435](http://www.nature.com/srep/2013/131002/srep02650/full/srep02650.html)：用多种driver检测方法找到的驱动基因。筛选出的已知驱动基因结果如下（下表仅为一个已知驱动基因的示例，具体样本的筛选结果见附件）：

**表7.6 已知驱动基因筛选结果**

| **Name** | **Content** |
| --- | --- |
| Gene | MTOR |
| Chr | 1 |
| Pos | 11264435 |
| Ref | T |
| Alt | G |
| ExonicFunc | . |
| AAChange | . |
| Genome Location | 1:11107485-11259409 |
| Role in Cancer | oncogene |
| Tumour Type ( Somatic ) | endometrial carcinoma; head and neck; clear cell renal cell carcinoma; anaplastic thyroid cancer; urothelial cell carcinoma; central nervous system tumours; testicular germ cell tumours and other tumour types |
| Tumor Type ( Germline ) | . |
| Mutated Tumor Samples ( Bert Vogelstein125 ) | . |
| Cancer Type ( SMG127 ) | KIRC |
| Putative Driver Category ( Comprehensive435 ) | High Confidence Driver |

注：

(1) Gene：变异所在基因名称。

(2) Chr：染色体。

(3) Pos：染色体上的位置。

(4) Ref：参考碱基。

(5) Alt：突变碱基。

(6) ExonicFunc：变异类型。

(7) AAChange：氨基酸改变信息。

(8) Genome Location：变异区域（来自Cancer Gene Census）。

(9) Role in Cancer：癌基因分类（TSG：肿瘤抑制基因；oncogene：致癌基因）。

(10) Tumour Type ( Somatic )：体细胞突变的肿瘤类型（来自cosmic数据库）。

(11) Tumour Type ( Germline )：胚系突变的肿瘤类型（来自cosmic 数据库）。

(12) Mutated Tumor Samples ( Bert Vogelstein125 )：在cosmic数据库中由此突变的肿瘤样本数（来自Bert Vogelstein125）。

(13) Cancer Type ( SMG127 )：癌症类型（来自SMG127）。

(14) Putative Driver Category ( Comprehensive435 )：预测的驱动基因类别（来自Comprehensive435）。

# 附录

## 1. 结果文件目录结构

分析结果包括以下内容:

1.1 qc\_summary\_brief.txt: 数据质量统计的结果。

1.2 var: 包含每个样本检测到的变异位点（包括snp和indel），如sample.var.vcf，和snp和indel在各基因功能元件上的分布统计（网页版，如sample.var.annodb.snp.html）。

1.3 cnv: 包含每个样本检测到的拷贝数变异，如sample.cnv.tsv；每个样本检测到的高频CNV，如sample-scatter.pdf，以及在每条染色体上高频CNV所在基因分布图，如sample.gene.2.pdf。

1.4 anno\_process: 包含每个样本整合多个注释软件后得到的注释信息，如sample.var.join.tsv。

1.5 predisposing\_gene: 包含每个样品筛选出的易感基因（如sample. var.predisposing\_gene.txt）和新生的致病位点（sample .var.predisposing\_new\_gene.txt）的注释信息。

1.6 pathway: 包含所有样本共有的变异所在基因（intersection.gene.txt）、变异所在基因代谢通路（intersection.pathway.txt），每个样本检测到的变异所在基因（如sample.var.gene.uniq.txt）和变异所在基因代谢通路的注释结果（如sample.var.pathway.txt），每个样本中每条通路的展示图（网页版：map\*.html，图片版：map\*.png）。

1.7 fusion\_gene: 包含每个样本检测到的融合基因，如sample. fusion.factera.tsv。

1.8 pathway: 包含所每个样本筛选出的驱动基因，如sample.var.driver\_gene.txt。

## 2. 注释结果格式说明

由于注释结果文件及包含信息较多, 故整理为文档 “注释结果说明.pdf”, 与结果文件一同提供。

## 3. 项目信息

项目负责人:wangxian@1gene.com.cn



电话：0571-87885727

网址：[www.1gene.com.cn](http://www.1gene.com.cn/)

邮箱：[support@1gene.com.cn](mailto:support@1gene.com.cn)

地址：杭州市滨江区聚才路远方中心15楼（310051）