## 

## 爱汝康分析技术文档

版本号：V2.0

发布时间：2016.07.28

目录

[1. 运行说明 3](#_Toc2014864168)

[1.1输出 3](#_Toc640489993)

[1.2 运行 4](#_Toc444571716)

[2. 技术路线 5](#_Toc221958977)

[2.1数据评估 6](#_Toc2099014030)

[2.2 比对 6](#_Toc990316699)

[2.3 Bam文件预处理 6](#_Toc1310455558)

[2.4变异检测和注释 8](#_Toc1742217037)

[2.5生成爱汝康报告 9](#_Toc1369385446)

[3.声明 10](#_Toc483963354)

# 运行说明

## 1.1输出

输出文件夹的列表如下：

Analysis

├── Annotation 生成的变异解析结果目录

│   └── Result 变异解析结果目录

│      └── Sample.result.xls 生成的变异解析文件

├── FastQC 生成的序列质量评估结果目录

├── Mapping 与基因组比对文件

│   └─── Summary 生成的统计文件目录

│      └── Sample.summary.txt 统计文件结果

├── Report 生成的报告文件目录

│   └── Sample.pdf 生成的乳腺癌风险筛查报告文件

└── SNP 生成的变异结果目录

└── Filtered 生成的过滤后的结果文件

   └── Sample.sort.recal.ug.filtered.vcf 生成的变异结果文件

输出目录主要包含与参考人类基因组比对结果数据（Mapping目录）、SNP和INDEL检测结果目录（SNP目录）、SNP和INDEL解析结果目录（Annotation目录）、生成报告文件目录（Report目录）。主要包含的文件如下：

Sample.result.xls 生成的变异解析文件

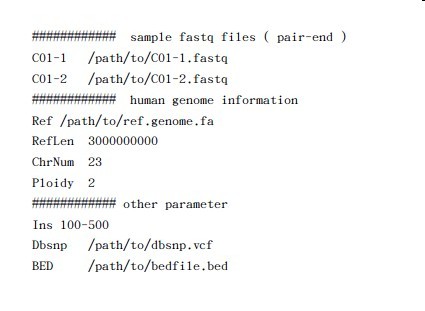
Sample.summary.txt 统计文件结果

Sample.sort.recal.ug.filtered.vcf 生成的变异结果文件.

Sample.pdf 生成的乳腺癌风险筛查报告文件

### 配置文件

基本分析的配置文件包含数据文件、程序所需参数值等，如下所示：



说明：

C01-1: 双端测序fastq reads1文件

C01-2: 双端测序fastq reads2文件

Ref： 参考基因组fa序列文件路径

RefLen： 参考基因组碱基数

ChrNum： 物种的染色体数量

Ploidy： 样本的倍体数

Ins： 插入片段大小

Dbsnp： Dbsnp数据库文件

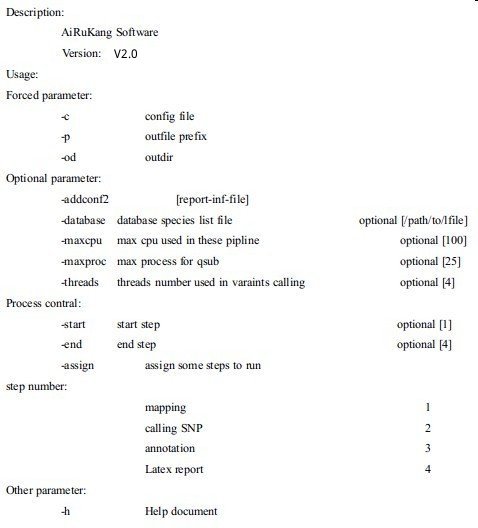
BED： 基因区间文件

## 运行

在shell命令行运行命令：

perl AiRuKang\_Main\_Pair.pl-c 配置文件 -p 输出文件前缀 -od 输出文件夹 -addfile2 受检者信息配置文件

命令行参数参见用法：



通过在配置文件里给出程序所需数据的路径和运行时需要的参数，在主程序运行时只需要输入一个配置文件即可。

软件可分步骤运行，通过 -start、-end来选择起始和结束步骤，或者-assign来选择指定步骤，可以实现对软件运行的控制。

# 2. 技术路线

2.1数据评估

测序数据量，测序数据质量和GC含量的统计。

|  |  |
| --- | --- |
| 软件 | fastq\_qc\_stat |
| 版本 | v1.1 |
| 命令行 | ./fastq\_qc\_stat -a read\_a.fq -b read\_b.fq |
| -a | 双端测序1reads1文件 |
| -b | 双端测序1reads2文件 |

2.2 比对

与基因组比对，基因组覆盖度，基因组覆盖深度统计。

|  |  |
| --- | --- |
| 软件 | bwa |
| 版本 | 0.7.1-r347 |
| 命令行 | bwa mem -M -t 10 ref.genome.fa Sample\_good\_1.fq Sample\_good\_2.fq > sample.bam |
| -M | 将短的比对热点标记 |
| -t | 使用的线程数 |

## 2.3 Bam文件预处理

1. 去除PCR扩增后的重复

|  |  |
| --- | --- |
| 软件 | picard-tools |
| 版本 | 1.94 |
| 命令行 | java -jar MarkDuplicates.jar REMOVE\_DUPLICATES=true INPUT=Sample.sort.bam OUTPUT=Sample.dedup.bam METRICS\_FILE=Sample.dedup.metrics |
| REMOVE\_DUPLICATES | 删除重复的reads |
| INPUT | 输入文件，为排序后的bam文件 |
| OUTPUT | 输出文件，为去重复后的bam文件 |
| METRICS\_FILE | 重复后的数据文件 |

1. 局部重新比对

A)

|  |  |
| --- | --- |
| 软件 | GenomeAnalysisTK |
| 版本 | 3.0.0 |
| 命令行 | java -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -I Sample.dedup.bam -R ref.genome.fa -o Sample.dedup.realn.intervals -L target.bed |
| -T | 使用的分析模块 |
| -I | bam文件 |
| -R | 基因组文件 |
| -L | 捕获区间 |
| -o | 输出的中间文件 |

B)

|  |  |
| --- | --- |
| 命令行 | java -jar GenomeAnalysisTK.jar -T IndelRealigner -I Sample.dedup.bam -R ref.genome.fa -L target.bed -targetIntervals Sample.dedup.realn.intervals -o Sample.dedup.realn.bam |
| -T | 使用的分析模块 |
| -I | bam文件 |
| -targetIntervals | 上一步生成的中间文件 |
| -R | 基因组文件 |
| -L | 捕获区间 |
| -o | 输出的bam文件 |

1. reads的碱基质量值进行重新校正

A)

|  |  |
| --- | --- |
| 软件 | GenomeAnalysisTK |
| 版本 | 3.0.0 |
| 命令行 | java -jar GenomeAnalysisTK.jar -T BaseRecalibrator -R ref.genome.fa -I Sample.dedup.realn.bam -L target.bed -knownSites dbsnp\_138.hg19.vcf -mte -nct 4 -o Sample.dedup.realn.recal.table |
| -T | 使用的分析模块 |
| -I | bam文件 |
| -R | 基因组文件 |
| -L | 捕获区间 |
| -o | 输出的中间文件 |
| -knownSites | 已知变异位点 dbsnp\_138.hg19.vcf |

B)

|  |  |
| --- | --- |
| 命令行 | java GenomeAnalysisTK.jar -T PrintReads -R ref.genome.fa -I Sample.dedup.realn.bam -L target.bed -knownSites dbsnp\_138.hg19.vcf -mte -nct 4 -o Sample.dedup.realn.recal.table -BQSR Sample.dedup.realn.recal.table |
| -T | 使用的分析模块 |
| -I | bam文件 |
| -R | 基因组文件 |
| -L | 捕获区间 |
| -o | 输出的中间文件 |
| -BQSR | 上一步生成的校正文件 |

## 2.4变异检测和注释

SNP、InDel的检测和注释。

1. 变异位点检测

A）

|  |  |
| --- | --- |
| 软件 | samtools |
| 版本 | 1.2.1 |
| 命令行 | samtools mpileup -f ref.genome.fa -l target.bed Sample.dedup.realn.recal.table > mpileup.out |
| mpileup | 使用的分析模块 |
| -l | 捕获区间 |
| mpileup.out | 变异输出文件 |

B）

|  |  |
| --- | --- |
| 软件 | VarScan |
| 版本 | 2.3.9 |
| 命令行 | java -jar VarScan.v2.3.9.jar mpileup2cns mpileup.out --output-vcf 1 -min-var-freq 0.001 > test.vcf |
| --output-vcf | 输出文件格式 |
| -min-var-freq | 输出最小频率阈值0.1% |

1. 变异结果注释

使用Annovar软件包以及爱汝康数据库（V1.5）进行注释

|  |  |
| --- | --- |
| 程序 | Anno\_main.pl |
| 命令行 | perl Anno\_main.pl -Step 1,2,3 -Outdir .Analysis -Sample sample -Prefix prefix |
| -Step | 注释步骤（1.Annovar注释；2.爱汝康数据库;3.注释整合） |
| -Outdir | 输出文件夹 |
| -Sample | 样本名称 |
| -Prefix | 前缀 |

1. 注释结果判定规则

|  |  |  |
| --- | --- | --- |
| 列编号 | 意义 | 判定方法 |
| 第1列 | 阳性/阴性 | 变异位点可在数据库中检出且突变频率>10%为阳性。其余为阴性。 |
| 第2列 | 是否在数据库中检出 | 已在数据库中检出标写数据库，未在数据库用“-”表示。 |
| 第3列 | 是否通过质控 | 位点测序深度>20为Pass，其余为Faild。 |
| 第4列 | 测序深度 | 为实际测序深度 |
| 第5列 | 变异频率 | 为Varscan检测出的变异频率,0.1%一下默认为0%。 |
| 第6列 | 纯合/杂合 | 变异位点情况，若突变频率<10%,默认为纯合 |
| 第7列 | 变异类型 | 根据数据库中注释的非同义突变/非同义突变，其余为其他突变。 |
| 第8列 | 临床意义 | 根据Clinvar中的临床意义判定，若pathogenic为有害突变，benign为无害突变。 |

## 2.5生成爱汝康报告

通过使用perl调用latex实现PDF报告生成。

|  |  |
| --- | --- |
| 程序 | ARK\_tex.pl |
| 命令行 | perl ARK\_tex.pl -Result\_file result.xls -Parameter\_file parameter.txt -Output\_directory ./Report |
| -Result\_file | 注释结果文件 |
| -Parameter\_file | 样本信息文件 |
| -Output\_directory | 输出文件夹 |

# 3.声明

本软件仅对单次样本进行解析，随着数据库的更新完善将解读更进一步的结果。