## 

## 检益康分析技术文档

版本号：V1.0

发布时间：2016.07.28

目录

[1. 运行说明 3](#_Toc1900066707)

[1.1输出 3](#_Toc377504455)

[1.2 运行 4](#_Toc1798652672)

[2. 技术路线 5](#_Toc1914418021)

[2.1数据评估 6](#_Toc2063717234)

[2.2 比对 6](#_Toc2026043977)

[2.3 Bam文件预处理 6](#_Toc268618527)

[2.4 Calling SNP 8](#_Toc916305134)

[3.声明 9](#_Toc414885740)

# 运行说明

## 1.1输出

输出文件夹的列表如下：

├── 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

└── 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

输出目录主要包含生成的序列质量评估结果目录（FastQC目录）、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）。主要包含的文件如下：

Sample.summary.txt 统计文件结果

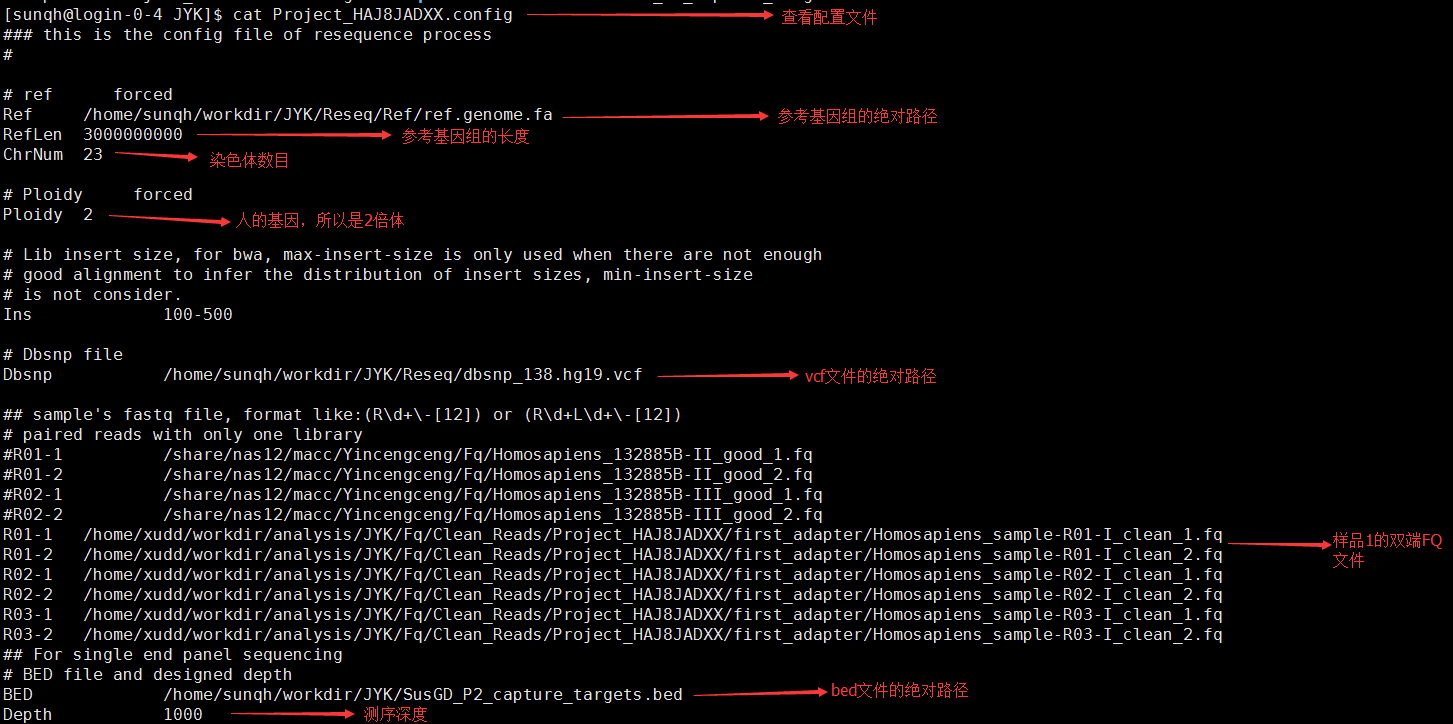
Sample.sort.recal.ug.filtered.vcf 生成的变异结果文件.

用vcf\_xls.pl程序将Result/SNP/Filtered/\*.recal.ug.filtered.vcf转换成xls文件

用main.pl程序将xls文件中所缺的基因补充成完整的xls。

### 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：

R01-1: 双端测序fastq reads1文件

R01-2: 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

Dbsnp： Dbsnp数据库文件

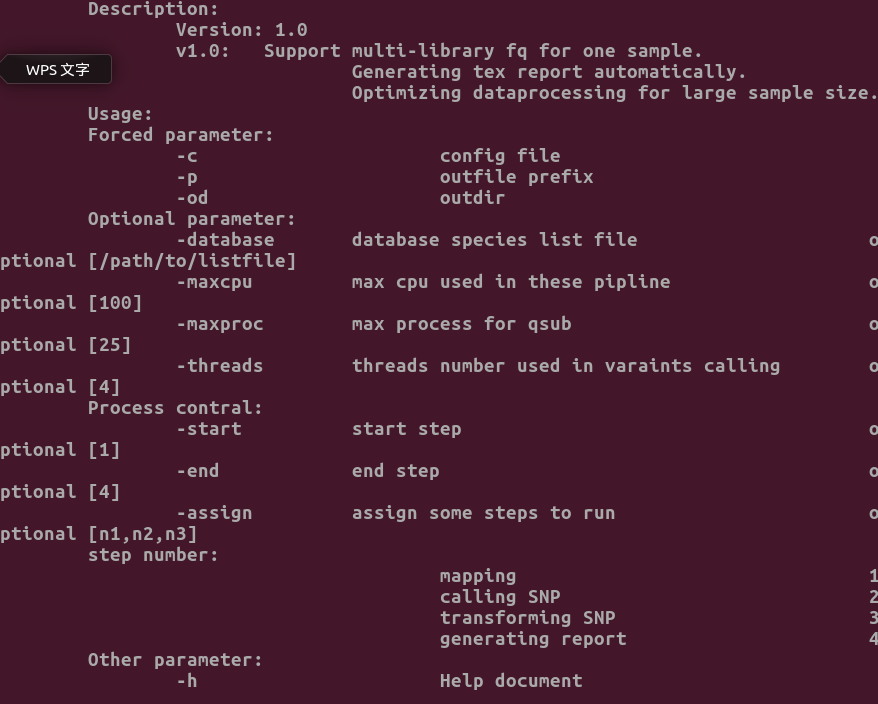
BED： 基因区间文件

## 运行

在shell命令行运行命令：



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。

# 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 Calling SNP

SNP、InDel的检测和注释。

1. 变异位点检测

A）

|  |  |
| --- | --- |
| 软件 | GenomeAnalysisTK |
| 版本 | 3.0.0 |
| 命令行 | java -Xmx20G -jar GenomeAnalysisTK.jar -T UnifiedGenotyper -I Sample.dedup.realn.recal.table.bam -L target.bed -R ref.genome.fa -nct 4 -dt NONE -glm BOTH --dbsnp dbsnp\_138.hg19.vcf -o Sample.dedup.realn.recal.table.vcf -rf BadCigar |
| -T | 使用的分析模块 |
| -I | bam文件 |
| -R | 基因组文件 |
| -L | 捕获区间 |
| -o | 输出文件 |
| -dbsnp | 已知变异位点 dbsnp\_138.hg19.vcf |

B）

|  |  |
| --- | --- |
| 软件 | GenomeAnalysisTK |
| 版本 | 3.0.0 |
| 命令行 | java -jar GenomeAnalysisTK.jar -T SelectVariants -R ref.genome.fa --variant Sample.dedup.realn.recal.table.vcf -o Sample.dedup.realn.recal.table.filter.vcf -select 'DP > 20' |
| -T | 使用的分析模块 |
| -variant | vcf文件 |
| -R | 基因组文件 |
| -o | 输出的vcf文件 |
| -select | 过滤条件 |

# 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。