WGS_Pipeline 1.0 使用文档

一. 流程运行环境的搭建

本流程调用了一系列分析软件，用户需要安装和配置这些软件，列表如下：

表 1. 流程所需软件列表

软件名称	功能	下载地址
Bwa	读句比对	http://sourceforge.net/projects/bio-bwa/files/
Samtools	SAM 文件处理	http://sourceforge.net/projects/samtools/files/
Picard	质控	http://sourceforge.net/projects/picard/files/
Queue	矫正比对分值	http://www.broadinstitute.org/gatk/download
GATK	检测变异	http://www.broadinstitute.org/gatk/download
SnpEff &SnpSift	过滤/注释变异	http://snpeff.sourceforge.net/download.html
Annovar	注释变异	http://www.openbioinformatics.org/annovar/annovar_download.html

软件的安装请参照各软件相关文档，具体操作请联系集群管理员。另外，需要注意的是，有些软件需要额外下载一些数据库才能正常运行，包括如下内容：

- **下载参考序列：**

  参考序列 FTP 站地址：ftp-trace.ncbi.nih.gov（匿名登录）

  参考序列文件路径（使用 1000Genomes 计划 phase2 的参考基因组 hs37d5.fa）：

  /1000genomes/ftp/technical/reference/phase2_reference_assembly_sequence

  BWA 软件需要使用**建立了索引**的参考序列，处理代码如下：

  bwa index -a bwtsw hs37d5.fa

- **下载参考 dbsnp 数据库：**

  参考 dbsnp FTP 站地址：ftp-trace.ncbi.nih.gov（匿名登录）

  参考 dbsnp 文件路径（使用 build 137）：/snp/organisms/human_9606/VCF

- **下载 GATK 变异校正（VariantRecalibrator）所需文件：**

  ftp 站地址：ftp.broadinstitute.org（用户名：gsapubftp-anonymous；密码：空）

  路径：/bundle/2.5/b37

  需要下载的文件：

  hapmap_3.3.b37.sites.vcf

  1000G_omni2.5.b37.sites.vcf

  dbsnp_135.b37.vcf

  Mills_and_1000G_gold_standard.indels.b37.sites.vcf

- **下载 Annovar 所需注释数据库：**

  Annovar 软件注释所需数据库较多，这里不再细述，下载完毕请放在同一文件夹。详见：http://www.openbioinformatics.org/annovar/annovar_db.html。

除了主要软件外，还需要一些其他**辅助工具**，这些工具无需安装，均**已包含在本流程的文件夹**中，列表如下：

表 2. 流程所需其他辅助工具

其它工具	备注
Vcf2Bco	VCF→Bco 的格式转换，原版本源自 SVA[1]，为了适用于 VCF4.1，我们对其进行了略微修改。（对 VCF4.0，请使用原版本）
Filter_RemoveLowQual	用于质控。bad quality 标准：DP<5、MQ<20、VQSLOD<0、supported reads<3，用户可根据需要另作修改。
Filter_myFilter	根据用户自定义，进一步筛选用户感兴趣的变异。高级用户可以根据需要进行修改，普通用户可使用我们的提供的默认文件。
DataProcessingPipeline.scala	本工具源自 GATK，由于一定原因目前已停止更新和公开下载，用户可以在 https://github.com/broadgsa/gatk/blob/master/public/scala/qscript/org/broadinstitute/sting/queue/qscripts/GATKResourcesBundle.scala 下载其最新版；但为了避免版本不同可能带来的问题，推荐用户使用我们提供的版本。
UnifiedGenotyper.scala	源自 GATK https://github.com/lifengtian/NGS/blob/master/QUEUE/UnifiedGenotyper.scala。推荐使用我们提供的版本

注 1：Ge D, Ruzzo EK, Shianna KV, He M, Pelak K, Heinzen EL, Need AC, Cirulli ET, Maia JM, Dickson SP, Zhu M, Singh A, Allen AS, Goldstein DB. SVA: software for annotating and visualizing sequenced human genomes. Bioinformatics, 2011, 27(14), 1998-2000.

为了实现并行计算，请确保计算机集群中**安装了集群管理软件 SGE**，具体安装事宜请联系集群管理员。

注：pipeline 代码中使用 qsub 命令时，没有指明任务要投放到的队列，使用的是默认队列。如果需要指定队列，请修改 wgs_pipeline.pl 文件代码中的 qsub 命令。

二．流程的自动化运行

填写配置文件

（如果使用命令行运行程序，则必须先填写配置文件；如果使用图形界面，则不用填写配置文件，图形界面会自动生成配置文件。）

本流程所涉及的常用配置均包含在一个配置文件中。每一个样本对应一个配置文件。用户通过修改配置文件即可完成对流程中软件、输入输出、参数等的设置。

配置文件为纯文本文件，分为 project config 和 system config 两个部分：project config 指出了数据的输入和输出目录、参考数据库路径等；system config 指出了第一步中安装软件的位置和运行参数。每行为一个配置项目，其中行首为“#”表明该行为注释信息，程序会略过。用户需要按照指定的项目名称填写，各配置项目的书写格式为：

【项目名】+【制表符 tab 分隔】+【项目值】

表 3. 配置项目列表

配置项目名称	描述
Sample_Name	样本名称
Input_Folder	输入目录：Fastq 文件所在路径
Output_Folder	输出目录
Email	用户邮箱，用于流程运行完毕后接收邮件提醒
Pipeline_Path	流程主程序 wgs_pipeline.pl 所在目录
bwa_threads	参数（见表 5）
bwa_trim	参数（见表 5）
Picard_MarkDuplicates_maxSequences	参数（见表 5）
Picard_MergeBam_validationStringency	参数（见表 5）
Picard_MergeBam_assumeSorted	参数（见表 5）
Picard_MergeBam_useThreading	参数（见表 5）
Picard_MarkDuplicates_removeDuplicates	参数（见表 5）
Picard_MarkDuplicates_assumeSorted	参数（见表 5）
Bwa	Bwa 软件目录
Samtools	Samtools 软件目录
Picard_MergeBam [注]	Picard 中 MergeSamFiles.jar 所在目录
Picard_MarkDuplicates [注]	Picard 中 MarkDuplicates.jar 所在目录
Picard_CollectSummary [注]	Picard 中 CollectAlignmentSummaryMetrics.jar 所在目录
Queue [注]	Queue.jar 所在目录
Queue_DataProcessingPipeline	DataProcessingPipeline.scala 所在目录
Queue_UnifiedGenotyper	UnifiedGenotyper.scala 所在目录
GATK [注]	GenomeAnalysisTK.jar 所在目录
GATK_SnpRecalHapmap	hapmap_3.3.b37.sites.vcf 所在目录
GATK_SnpRecalOmni	1000G_omni2.5.b37.sites.vcf 所在目录
GATK_SnpRecalDbsnp	dbsnp_135.b37.vcf 所在目录
GATK_IndelRecalMills	Mills_and_1000G_gold_standard.indels.b37.sites.vcf 所在目录
SnpSift [注]	snpSift.jar 所在目录
SnpEff [注]	snpEff.jar 所在目录
Annovar_ConvertAnn	Annovar 的 convert2annovar.pl 所在目录
Annovar_Annotation	Annovar 的 summarize_annovar.pl 所在目录
Annovar_db	Annovar 注释数据库所在目录
Reference_Sequence	参考序列数据库所在目录
Reference_Dbsnp	参考 dbsnp 数据库所在目录
Vcf2Bco [注]	Vcf2Bco.jar 所在目录
Filter_RemoveLowQual	Filter_RemoveLowQual 所在目录
Filter_myFilter	Filter_myFilter 所在目录

注：项目文件问 jar 文件时，请在其路径前添加 java 绝对路径和虚拟机运行参数具体如下：

表 4. jar 配置项目示例

jar 配置项目名称	jar 配置项目值示例（红色部分为额外添加值，使用时请注意修改 java 路径）
Picard_MergeBam	/usr/java/latest/bin/java -Xmx16g -jar /data/software/bin/picard-tools-1.73/MergeSamFiles.jar
Picard_MarkDuplicates	/usr/java/latest/bin/java -Xmx16g -jar /data/software/bin/picard-tools-1.73/MarkDuplicates.jar
Picard_CollectSummary	/usr/java/latest/bin/java -Xmx16g -jar /data/software/bin/picard-tools-1.73/CollectAlignmentSummaryMetrics.jar
Queue	/usr/java/latest/bin/java -Xmx4g -Djava.io.tmpdir=tmp -jar /data/software/bin/Queue-2.5-2/Queue.jar
GATK	/usr/java/latest/bin/java -Xmx16g -jar /data/software/bin/GenomeAnalysisTK-2.5-2/GenomeAnalysisTK.jar
SnpSift	/usr/java/latest/bin/java -Xmx16g -jar /data/software/bin/snpEff_v3_0/snpSift.jar
SnpEff	/usr/java/latest/bin/java -Xmx16g -jar /data/software/bin/snpEff_2_0_5/snpEff.jar
Vcf2Bco	/usr/java/latest/bin/java -Xmx16g -jar /data/software/bin/bchr/wk.jar

配置文件示例：

```
#=== WGS Pipeline1.0 ===

#=== project config ===
Sample_Name test_zsy
Input_Folder /wa/ugoodlfy/old/graduation_project/sampleTest/rawDataTest/
Output_FolderD:\
Email siyaozhang1@gmail.com
bwa_threads 8
bwa_trim 15
Picard_MarkDuplicates_maxSequencesForDiskReadEndsMap 5000000
Picard_MergeBam_validationStringency LENIENT
Picard_MergeBam_assumeSorted true
Picard_MergeBam_useThreading true
Picard_MarkDuplicates_removeDuplicates true
Picard_MarkDuplicates_assumeSorted true

#=== project config ===
Bwa /data/software/bin/bwa-0.6.2/bwa
Samtools /data/software/bin/samtools-0.1.18/samtools
Picard_MergeBam /usr/java/latest/bin/java -Xmx16g -jar /data/software/bin/picard-tools-1.73/MergeSamFiles.jar
Picard_MarkDuplicates /usr/java/latest/bin/java -Xmx16g -jar
/data/software/bin/picard-tools-1.73/MarkDuplicates.jar
Picard_CollectSummary /usr/java/latest/bin/java -Xmx16g -jar
```

/data/software/bin/picard-tools-1.73/CollectAlignmentSummaryMetrics.jar

Queue /usr/java/latest/bin/java -Xmx4g -Djava.io.tmpdir=tmp -jar /data/software/bin/Queue-2.5-2/Queue.jar

Queue_DataProcessingPipeline /wa/ugoodlfy/old/Queue_work/DataProcessingPipeline.scala

Queue_UnifiedGenotyper /wa/ugoodlfy/old/graduation_project/project_wgs_pipeline/UnifiedGenotyper.scala

GATK /usr/java/latest/bin/java -Xmx16g -jar /data/software/bin/GenomeAnalysisTK-2.5-2/GenomeAnalysisTK.jar

GATK_SnpRecalHapmap /db/picard/broad_b37/hapmap_3.3.b37.sites.vcf

GATK_SnpRecalOmni /db/picard/broad_b37/1000G_omni2.5.b37.sites.vcf

GATK_SnpRecalDbsnp /db/picard/broad_b37/dbsnp_135.b37.vcf

GATK_IndelRecalMills /db/picard/broad_b37/Mills_and_1000G_gold_standard.indels.b37.sites.vcf

SnpSift /usr/java/latest/bin/java -Xmx16g -jar /data/software/bin/snpEff_v3_0/snpSift.jar

SnpEff /usr/java/latest/bin/java -Xmx16g -jar /data/software/bin/snpEff_2_0_5/snpEff.jar

Annovar_ConvertAnn /data/software/bin/annovar/convert2annovar.pl

Annovar_Annotation /data/software/bin/annovar/summarize_annovar.pl

Annovar_db /data/software/bin/annovar/humandb

Reference_Sequence /db/database/1000genomes_phase2/hs37d5.fasta

Reference_Dbsnp /db/database/ncbi/dbSNP/dbSnp_b137.vcf

Vcf2Bco /usr/java/latest/bin/java -Xmx16g -jar /data/software/bin/bchr/wk.jar

Filter_RemoveLowQual /data/software/bin/bchr/removeLowQual.perl

Filter_myFilter /data/software/bin/bchr/myfilter.perl

注：考虑到有的用户在计算节点未设置相关软件的环境变量，因此我们规定用户需要在 system config 中指定各个软件的绝对路径，从而保证流程的顺利运行。

使用命令行

填写完配置文件，即可通过命令行自动化运行流程。在管理节点上，运行代码即可：

perl wgs_pipeline.pl –c yourSample.config

由于运行时间较长，建议把任务放到后台运行，命令如下：

nohup perl wgs_pipeline.pl –c yourSample.config &

在命令行会实时显示当前流程运行进度，当显示“All Finished”即表示流程运行完毕，同时用户也会收到相应的邮件提醒。

注：如果 wgs_pipeline.pl 和 yourSample.config 不在当前运行目录下，请注意在 wgs_pipeline.pl 和 yourSample.config 前加上其所在路径。

使用图形界面

用图形界面操作会更加直观。打开 runner.jar 出现以下界面：

图 1. 图形界面“project config”面板

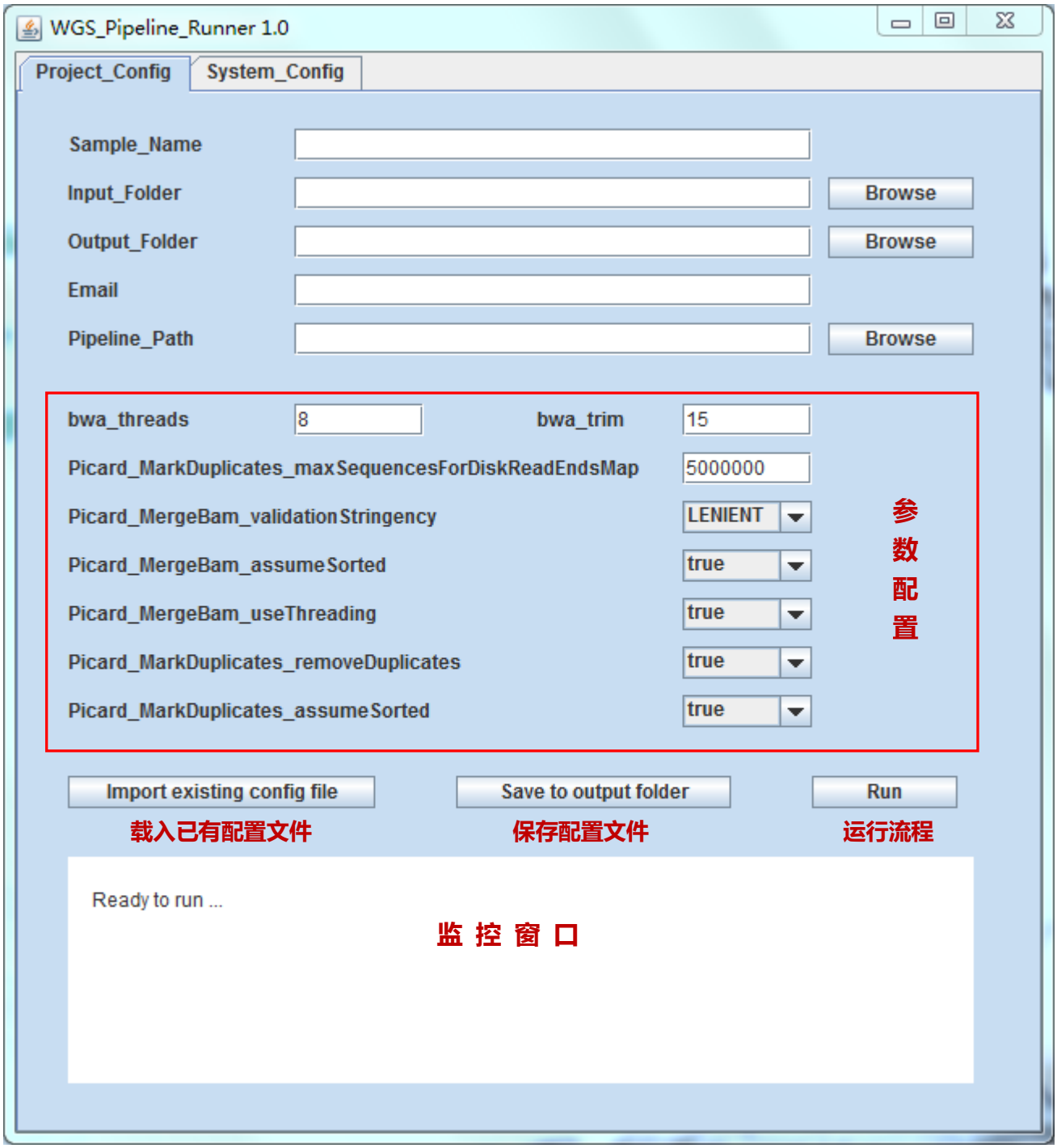
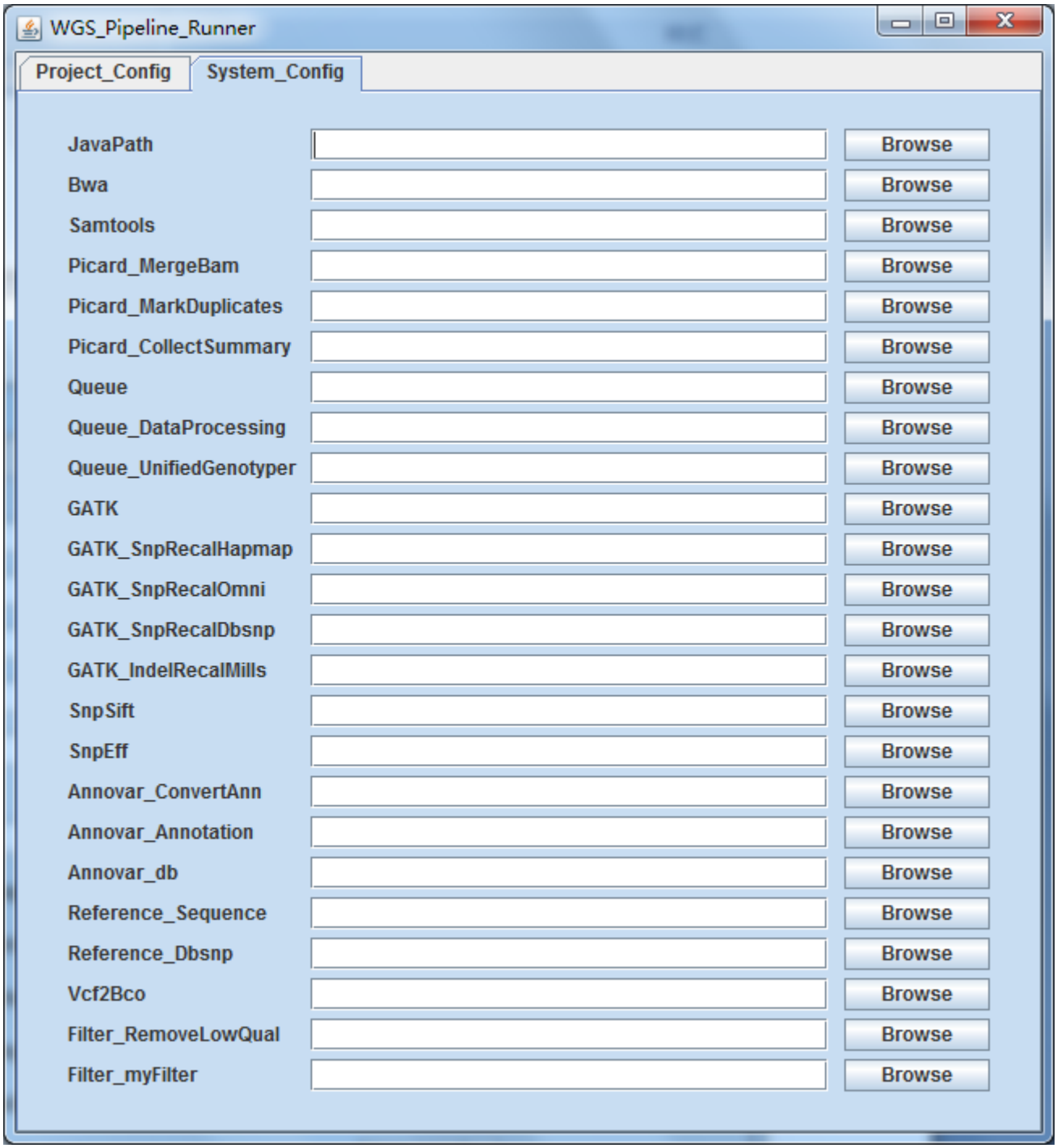


图 1. 图形界面“project config”面板



填写：用户需要在 Project_Config 和 System_Config 面板，按照“填写配置文件”部分的说明，分别填写相应配置项目。Project_Config 面板含参数配置部分，高级用户可以根据需要进行修改。如果想使用已有配置文件，可以点击“Import existing config file”按钮载入已有配置文件，并可以作进一步修改。我们提供的参数列表如下：

表 5. 可修改的参数列表

参数名称	默认值	描述
bwa_threads (-t)	8	线程个数（多线程模式）
bwa_trim (-q)	15	用于 reads 修剪（triming）
Picard_MarkDuplicates_maxSequencesForDiskReadEndsMap	5000000	ReadEnds will always be spilled to disk.

Picard_MergeBam_validationStringency	LENIENT	验证 Sam 文件中 reads 的严格性，候选{STRICT,LENIENT,SILENT}
Picard_MergeBam_assumeSorted	true	验证输入文件是否已排序，候选{true,false}
Picard_MergeBam_useThreading	true	是否开启线程，候选{true,false}
Picard_MarkDuplicates_removeDuplicates	true	是否将 Duplicates 写入到输出文件，候选{true,false}
Picard_MarkDuplicates_assumeSorted	true	验证输入文件是否已排序，候选{true,false}

保存：填写完配置项目，请务必点击“Save to Output_Folder”按钮，将配置文件保存至输出目录，以方便程序调用。

运行：保存配置文件后，点击“Run”即可运行全部流程。主面板 Project_Config 的下方是监控窗口，用户可以实时监控流程的运行进度，当窗口显示“All done.”即表示流程运行完毕，同时用户也会收到相应的邮件提醒。

结果： 自动化运行完全部流程，我们会得到一系列结果，这些包括 BWA align 读句定位生成的 sai 文件，BWA sampe 整合 pair-end 信息得到的 sam 文件，Samtools convert 转换 sam 得到的 bam 文件，Samtools sort 对 bam 文件排序得到的 sorted.bam 文件，Picard rmdup 去除重复得到的 sample_duprmed.bam 文件，GATK UG 和 GATK VQSR 得到的一些列 raw.vcf 文件，Filter 过滤后得到的 filtered.vcf 文件，以及 Annotation 注释后的 csv 变异文件。此外还给出了一个包含对实验数据质量评价的 summary 文件。用户请根据自己需要作进一步分析。

我们将对本流程进行长期维护。