用户使用手册
User Guide

2018 年 5 月 3 日

需求及安装

系统及软件需求

操作系统：CentOS 7.x

软件需求：Java 1.8、glibc 2.14

软件安装

整体目录结构如下：

```
gatk_fpga/
├── bin
├── lib
├── Xilinx_Answer_65444_Linux_Files/
        ├── driver
        ├── etc
        ├── include
        └── tests
```

- 压缩包当中主要有 3 个文件夹，每个文件夹包含文件说明如下：
  - bin/bwa：改进版的 bwa
  - bin/samtools：改进版的 samtools
  - bin/identify_server：认证服务程序
  - bin/gatk：GATK4 版本运行脚本程序
  - lib/gatk_3_6.jar：包含改进加速版的 GATK3.6 版本功能的 jar 包
  - lib/JNILib_3_6：改进加速版的 gatk3.6 所依赖的 JNI 库
  - lib/gatk_4.jar：包含改进加速版的 GATK4 版本功能的 jar 包
  - lib/JNILib_4：改进加速版的 GATK4 版本所依赖的 JNI 库
  - Xilinx_Answer_65444_Linux_Files：FPGA 驱动文件
- 运行软件前需要创建一个用户所有权限的输出目录，用于存放软件运行过程中产生的中间文件及输出文件。根据输入数据及产生的输出文件大小需要将该目录创建在具有足够存储空间的硬盘分区上。
- 运行软件前需要创建一个用户所有权限的临时目录，作为 java 运行所需要的 java.io.tmpdir 路径。

- 软件使用前需要首先开启认证服务，需要执行以下几个步骤：
  - 在运行软件的服务器(可以没有外网访问)设置环境变量到~/.bashrc:

    ```
    export YT_CONNECT_IP=xxx.xxx.xxx.xxx
    export YT_CONNECT_PORT=20006
    ```

    设置完毕后使用命令 source ~/.bashrc 使环境变量的修改立即生效。这里的 xxx.xxx.xxx.xxx 代表的是中间服务器的 IP 地址，这台服务器需要有外网访问，并且能够和运行软件的服务器之间网络连通，PORT 的选择可以选择任意没有被占用的端口，但需要保证与中间服务器设置环境变量中的端口保持一致。
  - 在中间服务器(可以访问外网，同时能够与运行软件的服务器网络连通)上设置环境变量到~/.bashrc:

    ```
    export YT_CONNECT_PORT=20006
    ```

    设置完毕后使用 source ~/.bashrc 使环境变量修改生效。将认证软件复制到中间服务器上，之后后台运行认证软件 identify_server，后台运行方法为：

    ```
    nohup ./identify_server &
    ```

  - 如果运行软件的服务器具有外网访问，那可以直接按照步骤 1 设置环境变量，将 YT_CONNECT_IP 设置为 127.0.0.1，在运行软件的服务器上直接后台运行认证软件 identify_server 即可，不需要另外一台中间服务器。

FPGA 驱动安装

- 编译：分别 cd 进入 driver、include、tests 文件夹，运行 make clean all。
- 安装：cd tests，然后执行 sudo ./load_driver.sh。

功能使用简介

该部分将介绍软件当中包含的各种功能的使用方法及参数介绍。对于 bwa、samtools 这两个工具，不同的加速产品的使用方法相同，对于其他步骤会对不同工具的使用方法分别进行说明。

bwa

使用示例：

```
bwa mem \
    -t NUMBER_THREADS \
    -f HT_FILE \
    -b TABLE_NUMBER_THREADS \
    -R '@RG\tID:READ_GROUP_ID\tSM:case\tLB:caseLib\tPU:runname\tPL:illumina \
    REFERENCE FASTQ_FILE_1 FASTQ_FILE_2 > OUTPUT_FILE
```

参数介绍：

参数	参数类型	描述
NUMBER_THREADS	int	运行 bwa 使用的线程数
HT_FILE	String	输出的 ht 文件的存储路径
TABLE_NUMBER_THREADS	int	用于生成 ht 文件的线程数
READ_GROUP_ID	String	ID 信息，在该软件中，此项必须包含有 tumor、normal、case 等关键字
REFERENCE	String	参考序列 fasta 文件的存储路径
FASTQ_FILE_1	String	输入的第一个 fastq 文件的存储路径
FASTQ_FILE_2	String	输入的第二个 fastq 文件的存储路径
OUTPUT_FILE	String	输出的 sam 文件的存储路径

该软件是对原版 bwa 软件的改进，包含有 bwa 原版软件的所有参数功能，关于 bwa 更详细的使用说明请参考：http://bio-bwa.sourceforge.net/bwa.shtml

compress

使用示例：

```
samtools view \
    -@ NUMBER_THREADS \
    -bS INPUT_FILE > OUTPUT_FILE
```

参数介绍：

参数	参数类型	描述
NUMBER_THREADS	int	运行 samtools 使用的线程数
INPUT_FILE	String	输入的 sam 文件的存储路径
OUTPUT_FILE	String	输出的 bam 文件的存储路径

sort

使用示例：

```
samtools sort \
    -@ NUMBER_THREADS \
    -m MEMORY_NUM \
    -h HT_FILE \
    INPUT_FILE -o OUTPUT_FILE
```

参数介绍：

参数	参数类型	描述
NUMBER_THREADS	int	运行 samtools 使用的线程数
MEMORY_NUM	String	使用的内存大小，例如 1536M
INPUT_FILE	String	输入的 bam 文件的存储路径
OUTPUT_FILE	String	输出的 bam 文件的存储路径

GATK 3.x 加速方案

BaseRecalibrator（软件加速方案）

使用示例：

```
java -d64 -server -XX:+UseParallelGC -XX:ParallelGCThreads=2 \
```

```
-Xms20g -Xmx20g \
-Djava.io.tmpdir=TMP_DIR \
-Djava.library.path=JNI_LIB_PATH \
-jar GATK_JAR \
-T BaseRecalibrator \
-I INPUT_FILE \
-o OUTPUT_TABLE \
-R REFERENCE \
-L TARGET_BED \
-knownSites KNOWN_SITES \
-nct NUMBER_THREADS \
-allowPotentiallyMisencodedQuals \
-rbs 3000
```

参数介绍：

参数	参数类型	描述
TMP_DIR	String	java 临时文件存储路径
JNI_LIB_PATH	String	程序运行所依赖的 JNI 库的存储路径
GATK_JAR	String	包含该功能的 jar 包的存储路径
INPUT_FILE	String	输入的 bam 文件的存储路径
OUTPUT_TABLE	String	输出的 table 文件的存储路径
REFERENCE	String	参考序列 fasta 文件的存储路径
TARGET_BED	String	target bed 文件的存储路径
KNOWN_SITES	String	已知的变异区域对应的 vcf 文件的存储路径（该参数可以添加多个）
NUMBER_THREADS	int	运行 BaseRecalibrator 使用的线程数

该软件是对原版 GATK 3.x 中 BaseRecalibrator 功能的加速改进，支持 BaseRecalibrator 中其他参数，关于 BaseRecalibrator 更详细的使用说明请参考 GATK 官方说明文档。

MuTect2（FPGA 加速方案）

使用示例：

```
java -d64 -server -XX:+UseG1GC -Xms20g -Xmx20g \
    -Djava.io.tmpdir=TMP_DIR \
    -Djava.library.path=JNI_LIB_PATH \
    -jar GATK_JAR \
    -T FastMuTect2 \
    -R REFERENCE \
    -I:tumor TUMOR_INPUT \
    -I:normal NORMAL_INPUT \
    --dbsnp DBSNP \
    --cosmic COSMIC \
    -contamination 0 \
    --max_alt_alleles_in_normal_count 3 \
    --max_alt_alleles_in_normal_qscore_sum 40 \
    --max_alt_allele_in_normal_fraction 0.02 \
    -nct THREADS \
    -dtt THREADS \
    -dt NONE \
    -o OUTPUT_FILE \
    --pair_hmm_implementation FPGA \
    -ntLib PAIRHMM_THREADS \
    -cTable TUMOR_TABLE \
    -nTable NORMAL_TABLE
```

参数介绍：

参数	参数类型	描述
TMP_DIR	String	java 临时文件存储路径
JNI_LIB_PATH	String	程序运行所依赖的 JNI 库存储路径
GATK_JAR	String	包含该功能的 jar 包的存储路径
REFERENCE	String	参考序列 fasta 文件的存储路径

TUMOR_INPUT	String	tumor 的 bam 输入文件的存储路径
NORMAL_INPUT	int	normal 的 bam 输入文件的存储路径
DBSNP	String	SNP 的 vcf 文件的存储路径
COSMIC	String	COSMIC 的 vcf 文件的存储路径
THREADS	int	染色体遍历及处理遍历结果线程数
OUTPUT_FILE	String	MuTect2 的输出 vcf 文件的存储路径
PAIRHMM_THREADS	int	pair-HMM 并行计算的线程数
TUMOR_TABLE	String	输入的 tumor 的 table 文件
NORMAL_TABLE	String	输入的 normal 的 table 文件

若进行计算的输入基因序列的测序深度较深，THREADS 建议设为 1，PAIRHMM_THREADS 建议设为：

$$\text{服务器总逻辑核数} - 2 \times \text{THREADS} - 4$$

若进行计算的输入基因序列的测序深度较浅，THREADS 根据服务器逻辑核数量建议设为 3-6，PAIRHMM_THREADS 建议为：

$$\text{服务器总逻辑核数} - 2 \times \text{THREADS} - 4$$

对于 target sequencing 数据，建议 THREADS 参数设置为 1，PAIRHMM_THREADS 设为：服务器总逻辑核数 - 2 × THREADS - 4。

该软件是对原版 GATK 中 MuTect2 功能的加速改进，支持 MuTect2 中其他参数，关于 MuTect2 更详细的使用说明请参考 GATK 官方说明文档。

MuTect2（软件加速方案）

软件加速方案和 FPGA 加速方案的参数使用除了将-pair_hmm_implementation FPGA 改为-pair_hmm_implementation CPP 之外，其余使用方式都是一致的。

GATK 4 加速方案

GATK4 加速方案使用前首先需要在.bashrc 中添加一个环节变量：

```
export GATK_LOCAL_JAR=~/gatk_fpga/lib/gatk_4.jar
```

该环节变量指明了 gatk4 的 jar 包的所在路径，同时，使用 source ~/.bashrc 使得

环境变量的修改立即生效。

BaseRecalibrator

使用示例：

```
GATK_SHELL --java-options "-Djava.library.path=JNI_LIB_PATH -Xmx32g" \
    BaseRecalibrator \
    -L TARGET_BED \
    -I INPUT_FILE \
    -O OUTPUT_TABLE \
    -R REFERENCE \
    --known-sites KNOWN_SITES
```

参数介绍：

参数	参数类型	描述
GATK_SHELL	String	gatk4 运行脚本路径（bin/gatk）
JNI_LIB_PATH	String	程序运行所依赖的 JNI 库的存储路径（lib/JNILib_4）
INPUT_FILE	String	输入的 bam 文件的存储路径
OUTPUT_TABLE	String	输出的 table 文件的存储路径
REFERENCE	String	参考序列 fasta 文件的存储路径
TARGET_BED	String	target bed 文件的存储路径
KNOWN_SITES	String	已知的变异区域对应的 vcf 文件的存储路径（该参数可以添加多个）

ApplyBQSR

使用示例：

```
GATK_SHELL --java-options "-Djava.library.path=JNI_LIB_PATH -Xmx32g" \
    ApplyBQSR \
    -I INPUT_FILE \
    -bqsr TABLE_FILE \
```

```
    -O OUTPUT_FILE \
    -R REFERENCE
```

参数介绍：

参数	参数类型	描述
GATK_SHELL	String	gatk4 运行脚本路径（bin/gatk）
JNI_LIB_PATH	String	程序运行所依赖的 JNI 库的存储路径（lib/JNILib_4）
INPUT_FILE	String	输入的 bam 文件的存储路径
OUTPUT_FILE	String	输出的 bam 文件的存储路径
REFERENCE	String	参考序列 fasta 文件的存储路径
TABLE_FILE	String	输入的 table 文件的存储路径

MuTect2（FPGA 加速方案）

使用示例：

```
GATK_SHELL --java-options "-Djava.library.path=JNI_LIB_PATH -Xmx32g" \
    Mutect2FPGA \
    -I TUMOR_INPUT \
    -tumor TUMOR_SAMPLE_NAME \
    -I NORMAL_INPUT \
    -normal NORMAL_SAMPLE_NAME \
    -O OUTPUT_FILE \
    -R REFERENCE \
    -NCT NCT_THREADS \
    -L TARGET_BED \
    --max-reads-per-alignment-start 0
```

参数介绍：

参数	参数类型	描述
GATK_SHELL	String	gatk4 运行脚本路径（bin/gatk）

JNI_LIB_PATH	String	程序运行所依赖的 JNI 库存储路径（lib/JNILib_4）
REFERENCE	String	参考序列 fasta 文件的存储路径
TUMOR_INPUT	String	tumor 的 bam 输入文件的存储路径
TUMOR_SAMPLE_NAME	String	tumor 的样本名称
NORMAL_INPUT	String	normal 的 bam 输入文件的存储路径
NORMAL_SAMPLE_NAME	String	normal 的样本名称
OUTPUT_FILE	String	MuTect2 输出 vcf 文件的存储路径
NCT_THREADS	int	匹配 FPGA 性能向 DDR 吞吐数据的线程数
TARGET_BED	String	target bed 文件的存储路径

对于多线程参数 NCT_THREADS，该参数用于匹配 FPGA 性能向 DDR 吞吐数据的多线程参数，对于 Xilinx_ku115 的芯片，一般将该参数设置为 6。

可选参数：
若希望提升运行速度（即测试报告中所提及的 FPGA 加速版 1），可以添加参数-tTable TUMOR_TABLE 和-nTable NORMAL_TABLE，这两个参数的说明同 GATK3.6 FPGA 加速版，为 BaseRecalibrator 步骤所输出的 table 文件，添加这两个参数之后，不需要运行 GATK4 的 ApplyBQSR 步骤，可以直接运行完 BaseRecalibrator 之后就运行 Mutect2FPGA 步骤。不添加这两个参数运行的即为测试报告中所提及的 FPGA 加速版 2。

该软件是对原版 GATK4 中 MuTect2 功能的加速改进，支持 MuTect2 中其他参数，关于 MuTect2 更详细的使用说明请参考 GATK4 官方说明文档。

完整流程示例

GATK 3.x 加速方案

```
# software
gatk=~/gatk_fpga/lib/gatk_3_6.jar
```

```
lib_path=~/gatk_fpga/lib/JNILib_3_6

# data
db=~/data/known_database
dbsnp_del100=$db/dbSNP/dbsnp_138.b37.vcf.gz
mills_1kg=$db/1000G_gold_standard/Mills_and_1000G_gold_standard.indels.b37.vcf.gz
cosmic=$db/COSMIC/b37_cosmic_v73_061615.vcf.gz
reference_file=~/data/reference_sequence/hs37d5.fasta

out_dir=~/outputt
tmp_dir=~/tmp

# bwa
bwa mem -t 12 -M -P -b 4 -f $out_dir/case.ht \
-R '@RG\tID:case\tSM:case\tLB:caseLib\tPU:runname\tPL:illumina' \
${reference_file} ~/input/case_1.fastq.gz ~/input/case_2.fastq.gz > ${out_dir}/case.sam

bwa mem -t 12 -M -P -b 4 -f $out_dir/case.ht \
-R '@RG\tID:case\tSM:case\tLB:caseLib\tPU:runname\tPL:illumina' \
${reference_file} ~/input/normal_1.fastq.gz ~/input/normal_2.fastq.gz > ${out_dir}/normal.sam

# compress
samtools view -@ 12 -bS ${out_dir}/case.sam > ${out_dir}/case.bam
samtools view -@ 12 -bS ${out_dir}/normal.sam > ${out_dir}/normal.bam

# sort
samtools sort -@ 12 -m 1536M -h ${out_dir}/case.ht \
${out_dir}/case.bam -o ${out_dir}/case_sort.bam

samtools sort -@ 12 -m 1536M -h ${out_dir}/normal.ht \
$out_dir/normal.bam -o ${out_dir}/normal_sort.bam
```

```
# index
samtools index ${out_dir}/case_sort.bam      ${out_dir}/case_sort.bam.bai
samtools index ${out_dir}/normal_sort.bam ${out_dir}/normal_sort.bam.bai

# BaseRecalibrator
java -d64 -server -XX:+UseParallelGC -XX:ParallelGCThreads=2 \
      -Xms20g -Xmx20g \
      -Djava.io.tmpdir=${tmp_dir} -Djava.library.path=${lib_path} -jar ${gatk} \
      -T BaseRecalibrator -I ${out_dir}/case_sort_realign.bam \
      -o ${out_dir}/case_sort_realign.grp \
      -R ${reference_file} -knownSites ${mills_1kg} -knownSites ${dbsnp_del100} \
      -knownSites ${cosmic} \
      -nct 12 -allowPotentiallyMisencodedQuals -rbs 3000

java -d64 -server -XX:+UseParallelGC -XX:ParallelGCThreads=2 \
      -Xms20g -Xmx20g \
      -Djava.io.tmpdir=${tmp_dir} -Djava.library.path=${lib_path} \
      -jar ${gatk} \
      -T BaseRecalibrator -I ${out_dir}/case_sort_realign.bam \
      -o ${out_dir}/case_sort_realign.grp \
      -R ${reference_file} -knownSites ${mills_1kg} -knownSites ${dbsnp_del100} \
      -knownSites ${cosmic} \
      -nct 12 -allowPotentiallyMisencodedQuals -rbs 3000

# mutect2
java -d64 -server -XX:+UseG1GC -Xms20g -Xmx20g -Djava.io.tmpdir=${tmp_dir} \
      -Djava.library.path=${lib_path} -jar ${gatk} -T FastMuTect2 \
      -R ${reference_file} \
      -I:tumor ${out_dir}/case_sort_realign.bam \
      -I:normal ${out_dir}/normal_sort_realign.bam \
      --dbsnp ${dbsnp_del100} --cosmic ${cosmic} -contamination 0 \
      --max_alt_alleles_in_normal_count 3 \
```

```
    --max_alt_alleles_in_normal_qscore_sum 40 \
    --max_alt_allele_in_normal_fraction 0.02 -nct 2 -dtt 2 -dt NONE \
    -o ${out_dir}/modifiedM_bqsrInActive2_dtt6nct6ntLib25_turn2.vcf \
    --pair_hmm_implementation FPGA -ntLib 12 \
    -cTable ${out_dir}/case_sort_realign.grp \
    -nTable ${out_dir}/normal_sort_realign.grp
```

GATK4 加速方案（FPGA 加速版 1）

```
# software
export GATK_LOCAL_JAR=~/gatk_fpga/lib/gatk_4.jar
gatk=~/gatk_fpga/bin/gatk
lib_path=~/gatk_fpga/lib/JNILib_4

# data
db=~/data/known_database
dbsnp_del100=$db/dbSNP/dbsnp_138.b37.vcf.gz
mills_1kg=$db/1000G_gold_standard/Mills_and_1000G_gold_standard.indels.b37.vcf.gz
cosmic=$db/COSMIC/b37_cosmic_v73_061615.vcf.gz
reference_file=~/data/reference_sequence/hs37d5.fasta
targetBed=~/data/target.bed

out_dir=~/outputt
tmp_dir=~/tmp

# bwa
bwa mem -t 12 -M -P -b 4 -f $out_dir/case.ht \
-R '@RG\tID:case\tSM:case\tLB:caseLib\tPU:runname\tPL:illumina' \
${reference_file} ~/input/case_1.fastq.gz ~/input/case_2.fastq.gz > ${out_dir}/case.sam

bwa mem -t 12 -M -P -b 4 -f $out_dir/case.ht \
```

```
-R '@RG\tID:case\tSM:case\tLB:caseLib\tPU:runname\tPL:illumina' \
${reference_file} ~/input/normal_1.fastq.gz ~/input/normal_2.fastq.gz > ${out_dir}/normal.sam

# compress
samtools view -@ 12 -bS ${out_dir}/case.sam > ${out_dir}/case.bam
samtools view -@ 12 -bS ${out_dir}/normal.sam > ${out_dir}/normal.bam

# sort
samtools sort -@ 12 -m 1536M -h ${out_dir}/case.ht \
${out_dir}/case.bam -o ${out_dir}/case_sort.bam

samtools sort -@ 12 -m 1536M -h ${out_dir}/normal.ht \
$out_dir/normal.bam -o ${out_dir}/normal_sort.bam

# index
samtools index ${out_dir}/case_sort.bam      ${out_dir}/case_sort.bam.bai
samtools index ${out_dir}/normal_sort.bam ${out_dir}/normal_sort.bam.bai

# BaseRecalibrator
${gatk} --java-options " -Djava.library.path=${lib_path} -Xmx32g" \
      BaseRecalibrator \
      -L ${targetBed} \
      -I ${out_dir}/case_sort.bam \
      -O ${out_dir}/case.table \
      -R ${reference_file} \
      --known-sites ${ dbsnp_del100} \
      --known-sites ${mills_1kg} \
      --known-sites ${cosmic}

${gatk} --java-options " -Djava.library.path=${lib_path} -Xmx32g" \
      BaseRecalibrator \
      -L ${targetBed} \
```

```
    -I ${out_dir}/normal_sort.bam \
    -O ${out_dir}/normal.table \
    -R ${reference_file} \
    --known-sites ${ dbsnp_del100} \
    --known-sites ${mills_1kg} \
    --known-sites ${cosmic}

${gatk} --java-options " -Djava.library.path=${lib_path} -Xmx32g" \
    Mutect2FPGA \
    -R ${reference_file} \
    -L ${targetBed}
    -I ${out_dir}/normal_sort.bam \
    -normal normal \
    -I ${out_dir}/case_sort.bam \
    -tumor case \
    -O ${out_dir}/out.vcf \
    --max-reads-per-alignment-start 0 \
    -tTable ${out_dir}/case.table \
    -nTable ${out_dir}/normal.table
```