**Refinement of alignment**

Aim

The aim of this standard operating procedure (SOP) is to provide guidance on how to refine alignments generated by bwa/novoalign. This procedure will remove some artifacts from sample preparation, sequencer or aligner (see “Overview of bioinformatics SOPs for exome sequencing” for detail explanation). The result BAM file is cleaned and ready for variant calling.

Tools

The refinement is performed by the bioinformatic tools: GATK (version 2.1-6-g6a46042) and Picard (version 1.74). The local realignment around indels and base quality score recalibration are performed by GATK. GATK is a tool set to handle high-throughput sequencing data for variant calling. The mark duplication is performed by Picard. Picard comprises Java-based command-line utilities that manipulate SAM files.

Input

The inputs are mapped BAM file and its index file, which are generated from alignment step. The BAM file should be sorted by the chromosome (1~22, X, Y, M) and genomic coordinates, and contain read group information.

Procedure

1, Local realignment around indels

The aim of this step is to remove alignment artifacts.

java -Xmx2g -jar GenomeAnalysisTK.jar

-T RealignerTargetCreator

-R bundle/1.5/b37/human\_g1k\_v37\_decoy.fasta

-o all.posiSrt.intervals

-I inPutBamFile

--known bundle/1.5/b37/Mills\_and\_1000G\_gold\_standard.indels.b37.sites.vcf

-nt 3

2>errRealignerTargetCreator > realignerTargetCreatorInfo.txt

java -Xmx4g -jar GenomeAnalysisTK.jar

-T IndelRealigner

-I inPutBamFile

-R bundle/1.5/b37/human\_g1k\_v37\_decoy.fasta

-targetIntervals all.posiSrt.intervals

-o all.realigned.bam

-known bundle/1.5/b37/Mills\_and\_1000G\_gold\_standard.indels.b37.sites.vcf

-compress 0

2>errIndelRealigner > indelRealignerInfo.txt

2, Mark duplicates

The aim of this step is to remove PCR duplicates generated during sample preparation.

java -Xmx2g -jar MarkDuplicates.jar

INPUT=all.realigned.bam

OUTPUT= all.realigned.markDup.bam

METRICS\_FILE= markDup.metrics.txt

CREATE\_INDEX=TRUE

2>errMarkDup

3, Base quality score recalibration

The aim of this step is to remove bias generated when sequencer estimating base quality score.

java -Xmx4g -jar GenomeAnalysisTK.jar

-T BaseRecalibrator

-I all.realigned.markDup.bam

-R bundle/1.5/b37/human\_g1k\_v37\_decoy.fasta

-knownSites bundle/1.5/b37/dbsnp\_135.b37.vcf

-knownSites bundle/1.5/b37/Mills\_and\_1000G\_gold\_standard.indels.b37.sites.vcf

-knownSites bundle/1.5/b37/1000G\_phase1.indels.b37.vcf

-o recal\_data.grp

2>errCountCovariatesPre > countCovariatesPreInfo.txt

java -Xmx1g -jar GenomeAnalysisTK.jar

-T PrintReads

-R bundle/1.5/b37/human\_g1k\_v37\_decoy.fasta

-I all.realigned.markDup.bam

-BQSR recal\_data.grp

-EOQ

-o all.realigned.markDup.baseQreCali.bam

-PF gatkPerformanceLog

2>errPrintReads > printReadsInfo.txt

(See attachments for detail about the options in Picard and GATK).

**Output**

1. Output from base quality score recalibration

The output BAM file (realignMarkdupRecalBam) and its index file (.bai) from base quality score recalibration is the final output from refinement of alignment. They will be used for variant calling.

BAM format is the compressed version of SAM format (see novoalign SOP for detail about the format.)

2. err and Info Outputs

Both err\* and \*Info\* files are logging files for each step. if all processes running successfully:

* All GATK err files should be empty;
* In all GATK info files, there should be a line says “Total runtime …”
* In all Picard err files, there should be a line says “Elapsed time …”

**Variation**

The whole procedure needs to be rerun from the first failed step.

**Storage**

See “SOP for storage and security of high-throughput sequencing data”

**Reference**

1, McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010). **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res.* 20:1297-303.

2, DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, MaKenna A, Fennell T, Kernytsky A, Sivachenko A, Cibulskis K, Gabriel S, Altshuler D and Daly, M (2011). **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nature Genetics.* 43:491-498.

3, Picard: <http://picard.sourceforge.net>

**Appendix**

1, GATK general options

2, Picard general options

3, GATK options for realignerTargetCreator

4, GATK options for indelRealigner

5, GATK options for BaseRecalibrator

6, GATK options for PrintReads

7, Picard options for MarkDuplicates