**Post-alignment Quality Control on High-Throughput Sequence (HTS) data**

**Aim**

The aim of this standard operating procedure (SOP) is to provide guidance on the quality control (QC) on the alignment after refinement of alignment. The results from this can be used for evaluation of the quality of alignment.

Responsible person: The bioinformatician who performs the analysis

**Tools**

The QCs are performed by the bioinformatics tools: Picard (version v1.74). Picard comprises Java-based command-line utilities that manipulate SAM files.

**Input**

The inputs are BAM file (Input\_BAM\_file) and its index file, which are generated from refinement alignment. The BAM file should be sorted by the chromosome in the order of (1~22, X, Y, M) and genomic coordinates, and contain read group information.

**Procedure**

**1, CollectAlignmentSummaryMetrics**

java -Xmx2g -jar CollectAlignmentSummaryMetrics.jar

MAX\_INSERT\_SIZE=600

INPUT=all.realigned.markDup.baseQreCali.bam

OUTPUT=collectAlignmentSummaryMetrics.txt

REFERENCE\_SEQUENCE=bundle/1.5/b37/human\_g1k\_v37\_decoy.fasta

VALIDATION\_STRINGENCY=LENIENT 2>errCollectAlignmentSummaryMetrics

**2, CollectInsertSizeMetrics**

java -Xmx2g -jar CollectInsertSizeMetrics.jar

HISTOGRAM\_FILE=insertSizeHistogram

INPUT=all.realigned.markDup.baseQreCali.bam

OUTPUT=collectInsertSizeMetrics.txt

REFERENCE\_SEQUENCE=bundle/1.5/b37/human\_g1k\_v37\_decoy.fasta

VALIDATION\_STRINGENCY= LENIENT

2>errCollectInsertSizeMetrics

**3, Preparing target and bait files from agilent capture kit**

* go to website <https://earray.chem.agilent.com/suredesign/search.htm>
* go into the required capture kit and click “download” button on the top right
* extract target and bait regions from designID\_AllTracks.bed file
* create interval file based on the format defined at “3 Intervals of interest” under webpage <http://gatkforums.broadinstitute.org/discussion/1204/what-input-files-does-the-gatk-accept>

**4, CalculateHsMetrics**

java -Xmx2g -jar CalculateHsMetrics.jar

BAIT\_INTERVALS= /Volumes/data.odin/common/captureTechnologies/agilent/SureSelect\_50Mb\_exome\_v11\_hg19/30\_generatedData/agilent.SureSelect\_50Mb\_exome\_v11\_hg19.merged.b37.bait.interval\_list

TARGET\_INTERVALS= /Volumes/data.odin/common/captureTechnologies/agilent/SureSelect\_50Mb\_exome\_v11\_hg19/30\_generatedData/agilent.SureSelect\_50Mb\_exome\_v11\_hg19.merged.b37.target.interval\_list

INPUT=all.realigned.markDup.baseQreCali.bam

OUTPUT=calculateHsMetrics.txt

REFERENCE\_SEQUENCE= bundle/1.5/b37/human\_g1k\_v37\_decoy.fasta

PER\_TARGET\_COVERAGE=perTargetCoverage.txt

VALIDATION\_STRINGENCY=SILENT

2>errCalculateHsMetrics

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

**Output**

1, collectAlignmentSummaryMetrics.txt

High level metrics about the alignment of reads within a SAM file, produced by the CollectAlignmentSummaryMetrics program.

“PCT\_READS\_ALIGNED\_IN\_PAIRS” should be larger than 90%.

2, insertSizeHistogram

Histogram of insertion sizes in the sample, produced by the CollectInsertSizeMetrics program.

3, collectInsertSizeMetrics.txt

statistical distribution of insert size (excluding duplicates), produced by the CollectInsertSizeMetrics program.

“MEDIAN\_INSERT\_SIZE” should be not deviated much from lab fragmentation size.

4, calculateHsMetrics.txt

A set of Hybrid Selection specific metrics from an aligned SAM or BAM file, produced by the CalculateHsMetrics program.

“ON\_BAIT\_BASES”/(OFF\_BAIT\_BASES + NEAR\_BAIT\_BASES) should be between 40% - 70%.

5, perTargetCoverage.txt

GC and mean coverage information for every target, produced by the CalculateHsMetrics program.

6, 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 steps need to be rerun if the processes are not finished. If the quality control is not passed, the comments should be written in the “Bioinformatic comments” under “5-Bioinformatics” in HTS database.

Storage

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

**Reference**

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

Appendix

1, Picard general options

2, Picard options for CollectAlignmentSummaryMetrics

3, Metrics detail for CollectAlignmentSummaryMetrics

4, Picard options for CollectInsertSizeMetrics

5, Metrics detail for CollectInsertSizeMetrics

6, Picard options for CalculateHsMetrics

7, Metrics detail for CalculateHsMetrics