# AutoMapper

Created by Hui, Yawei, last modified on Nov 02, 2022



## I. INTRODUCTION

The standard mapping of next-generation sequencing data has matured and common practice is available from various large consortia (e.g. TCGA, ICGC) and sequencing centers. To accommodate the needs of more and more embedded computational biologists, the Center for Applied Bioinformatics adopted the common practice and implemented the AutoMapper, a pipeline using the most recent versions of the reference sequence and annotation files. We hope that this will facilitate downstream data integration and cross-group collaboration.

The ultimate goal of the CAB AutoMapper pipeline is to automate the alignment of the raw sequencing data against a given reference genome in an efficient and reliable fashion. The initial inputs expected from end-users are essential information about the sequenced samples in a single project. The final outputs exiting the pipeline are the aligned reads in BAM file format as well as varieties of QC information ready to be put in a summary report and emailed to users.

## II. GENERAL WORKFLOW

The general CAB workflow is depicted in the following flow-chart diagram, which connects the Hartwell Genome Sequencing Facility (GSF), the Shared Resource Management (SRM, v.2), and the St. Jude High-Performance Computing Facility (HPCF). After the initial sample submission by users, the full path of the data processing involves steps:

1. SRM2 sends orders to GSF with samples to be sequenced;

2. Sequenced data are generated, stored, and validated by GSF staff members;

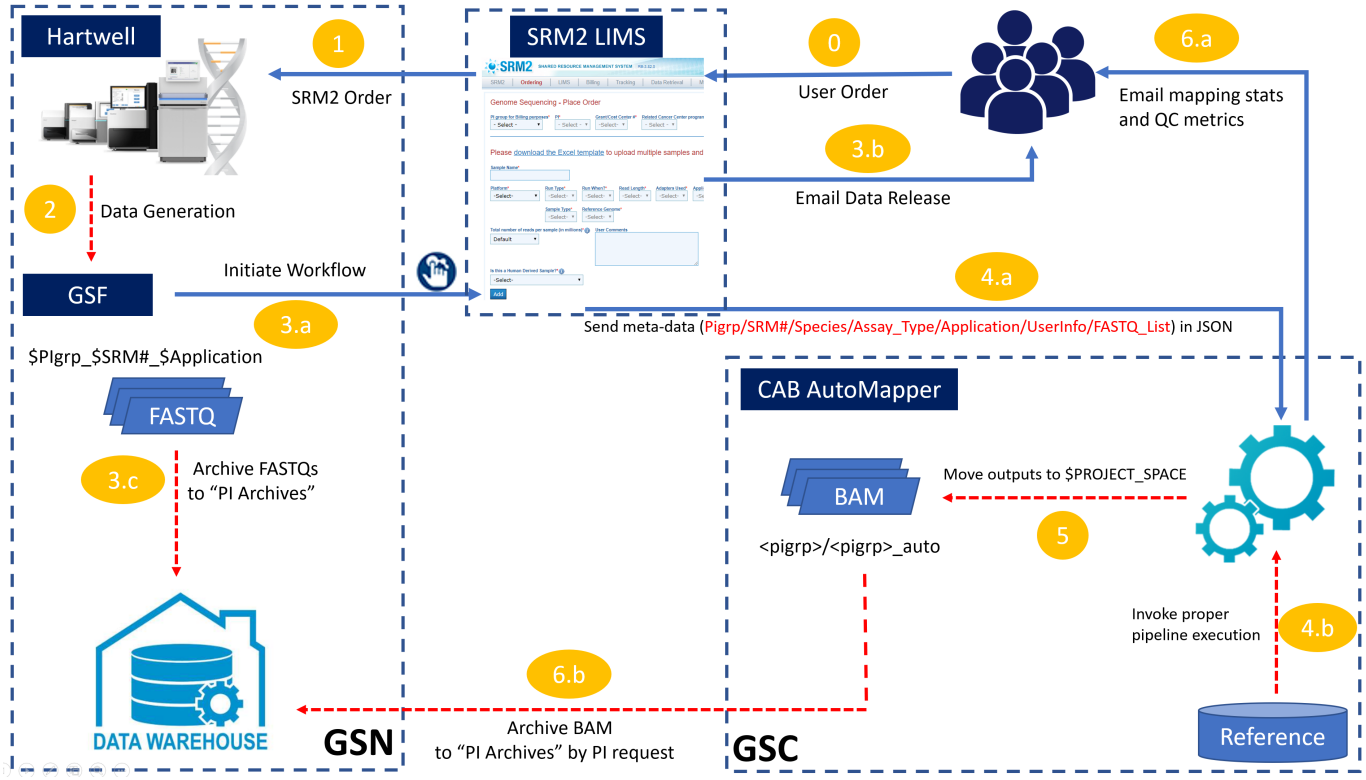
3. Sequenced data are released and put in CAB automation pipeline for processing (either simultaneously or sequentially);

~~c. Archive the sequencing data in FASTQs (under construction);~~

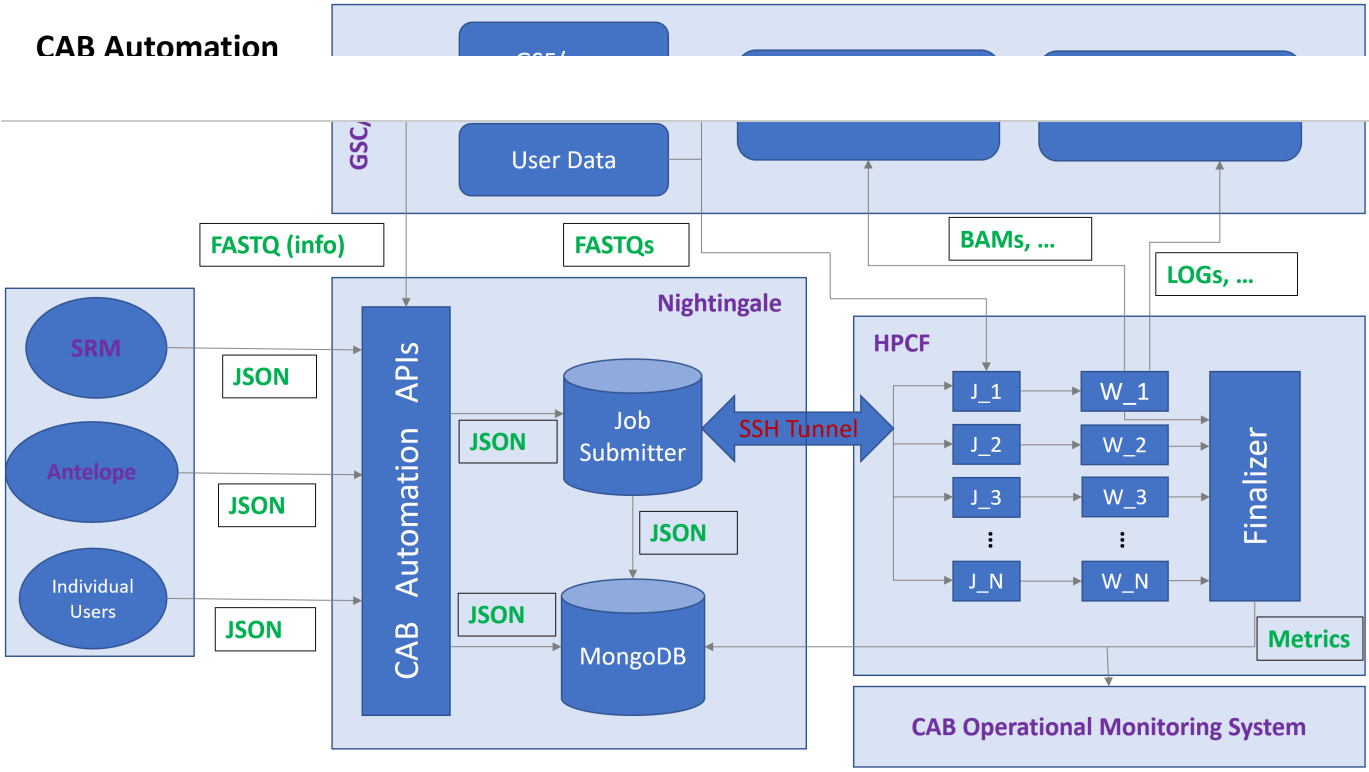
4. Meta-data about the project are received and used to construct the workflow tasks:
   a. JSON payload sent by SRM is parsed by AutoMapper and corresponding LSF job descriptions (in LSF scripts) are sent to HPCF for processing;
   b. Both sequencing data and references are prepared at HPCF for job execution;

5. AutoMapper finishes the jobs and delivers mapping results to proper locations in $PROJECT_SPACE;

6. Wrap-up and Finalize the pipeline processing:
   a. With the final report created at the end of the pipeline, AutoMapper sends user notification of mapping completion;
   b. Archive the mapped data in BAMs (under construction);




Despite the diversity in toolsets used in CAB AutoMapper pipelines and the substantial differences among their inputs/outputs, the workflow management shares a common pattern when dealing with job submissions revolving around a single project which could be uniquely identified with the PI group ($PIgrp) and SRM2 order number ($SRM#). Each sample in the project will be submitted to HPCF as an independent job, i.e. being associated with its own job ID ($JOBID). These jobs will be scheduled to run in different queues depending on what hardware/software platforms are required to carry out the major computational tasks. Upon the completion of a pipeline job (with the status of either success or failure), a wrap-up task follows to complete the necessary I/O operations such as delivering the outputs to the job submitter, archiving various operational records, and clean up the user environment. At the last stage of the AutoMapper operation, the proper report on the overall mapping results for all samples in the project is created, stored with the project's output, and delivered to the job submitter by email (see Appendix B. for an example).



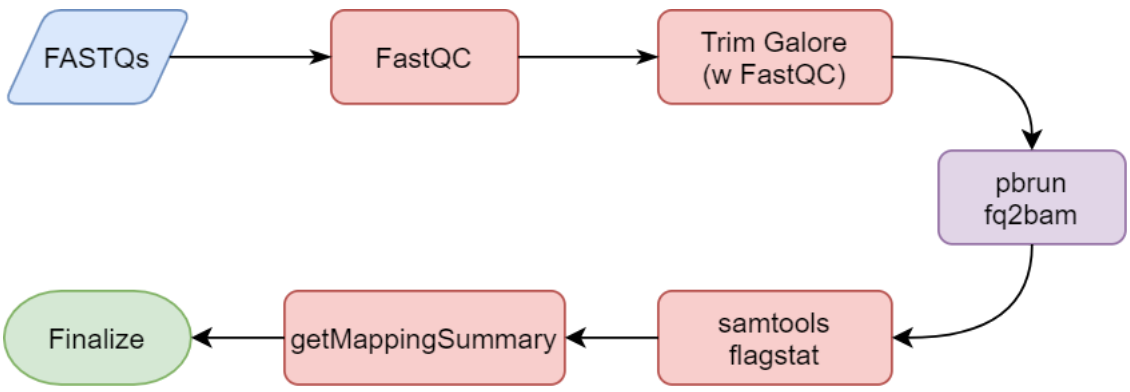

## III. SPECIFICATIONS OF PIPELINES

In designing and implementing pipelines in CAB AutoMapper, a certain set of environment variables are adopted for general-purpose usage. Here is a list of the most frequently used variables in the pipelines and a more comprehensive one could be found in Appendix A.

| NAME | VALUE | DESCRIPTION | EXAMPLE |
|---|---|---|---|
| LSB_MAX_NUM_PROCESSORS | - | The maximum number of processors requested when the job is submitted | 4 |
| PREFIX | - | The prefix as being added in various output files, always taken the value of the sample name | SRR1069943 |
| OUTPUT_DIR | - | The destination directory in which the raw outputs from various pipeline tools are stored, see Extra Notes for details | /research/rgs01/scratch/users/yhui/RNAseq/FQMapping/YFAN-GTex-Supple-UNSTRANDED/SRR1069943/89543094 |
| TMP_DIR | - | The temporary storage location for intermediate I/O created by various pipeline tools, always residing in the scratch space | /research/rgs01/scratch/users/yhui/tmp |

### A. Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq)

**Flowchart -**



**Inputs -**

```
fastqc \
     --outdir ${OUTPUT_DIR}/FASTQC/RAW \
     --threads ${LSB_MAX_NUM_PROCESSORS} \
     --format fastq \
     --quiet  \
     ${OUTPUT_DIR}/${FASTQ}
```

**FASTQC (Paired Ended)** Collapse source

```
fastqc \
     --outdir ${OUTPUT_DIR}/FASTQC/RAW \
     --threads ${LSB_MAX_NUM_PROCESSORS} \
     --format fastq \
     --quiet  \
     ${OUTPUT_DIR}/${FASTQ1} \
     ${OUTPUT_DIR}/${FASTQ2}
```

**TRIM-GALORE (Single Ended)** Collapse source

```
trim_galore \
     --gzip  \
     --clip_R1 15 \
     --cores ${LSB_MAX_NUM_PROCESSORS} \
     --output_dir ${OUTPUT_DIR} \
     --fastqc_args "--outdir ${OUTPUT_DIR}/FASTQC/TRIM" \
     ${OUTPUT_DIR}/${FASTQ}
```

**TRIM-GALORE (Paired Ended)** Collapse source

```
 trim_galore \
      --paired  \
      --gzip  \
      --clip_R1 15 \
      --clip_R2 15 \
      --cores ${LSB_MAX_NUM_PROCESSORS} \
      --output_dir ${OUTPUT_DIR} \
      --fastqc_args "--outdir ${OUTPUT_DIR}/FASTQC/TRIM" \
      ${OUTPUT_DIR}/${FASTQ1} \
      ${OUTPUT_DIR}/${FASTQ2}
```

**PBRUN-FQ2BAM (Single Ended)** Collapse source

```
pbrun fq2bam \
     --ref ${REF_FILE} \
     --out-bam ${OUTPUT_DIR}/${PREFIX}.bam \
     --out-duplicate-metrics ${OUTPUT_DIR}/${PREFIX}.metrics.txt \
     --bwa-options "-K 100000000 -Y -M" \
     --num-gpus 2 \
     --tmp-dir ${TMP_DIR} \
     --in-fq ${TRIM_FASTQ1} "@RG\\tID:${RG_ID}\\tLB:${RG_LB}\\tPL:${RG_PL}\\tSM:${RG_SM}\\tPU:${RG_PU}"
```

**PBRUN-FQ2BAM (Paired Ended)** Collapse source

```
pbrun fq2bam \
     --ref ${REF_FILE} \
     --out-bam ${OUTPUT_DIR}/${PREFIX}.bam \
     --out-duplicate-metrics ${OUTPUT_DIR}/${PREFIX}.metrics.txt \
     --bwa-options "-K 100000000 -Y -M" \
     --num-gpus 2 \
     --tmp-dir ${TMP_DIR} \
     --in-fq ${TRIM_FASTQ1} ${TRIM_FASTQ2} "@RG\\tID:${RG_ID}\\tLB:${RG_LB}\\tPL:${RG_PL}\\tSM:${RG_SM}\\tPU:${RG_PU}"
```

**SAMTOOLS-FLAGSTAT** Collapse source

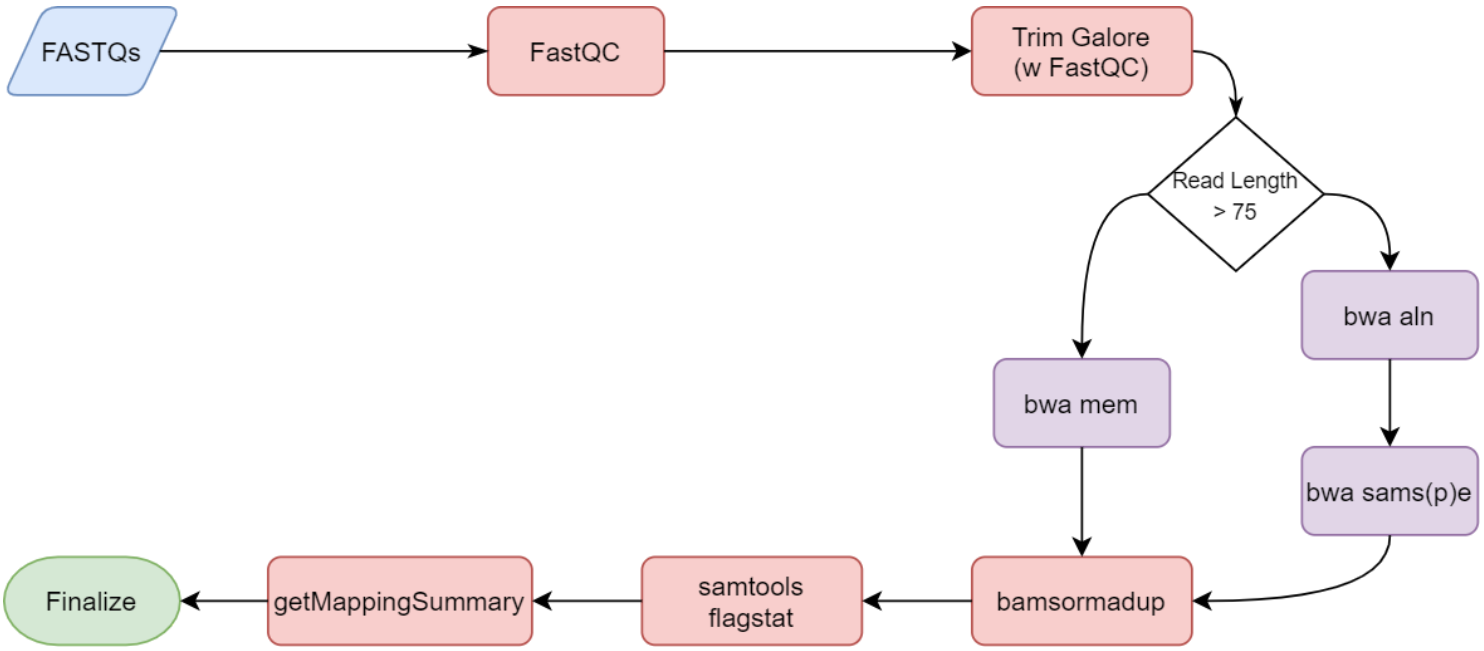
```
samtools flagstat \
     ${OUTPUT_DIR}/${PREFIX}.bam \
```

```
    > ${OUTPUT_DIR}/${PREFIX}.flagstat.txt
```

| | NAME | TYPE | DESCRIPTION |
|---|---|---|---|
| 1 | ${PREFIX}.bam | FILE | BAM file, coordinate-sorted, duplicates marked, created by bamsormadup |
| 2 | ${PREFIX}.bam.bai | FILE | BAM index file, created by bamsormadup |
| 3 | ${PREFIX}.report | FILE | Report on the mapping status and various metrics, created in accord with ATAC-seq mapping criteria |
| 4 | FASTQC/RAW | DIRECTORY | FASTQC report, untrimmed FASTQ files taken as input, created by FastQC |
| 5 | FASTQC/TRIM | DIRECTORY | FASTQC report, trimmed FASTQ files taken as input, created by FastQC |

## B. Chromatin ImmunoPrecipitation followed by sequencing (ChIP-seq)

**Flowchart -**




**Inputs -**

**TRIM-GALORE (Single Ended)** Collapse source

```
trim_galore \
     --gzip  \
     --cores ${LSB_MAX_NUM_PROCESSORS} \
     --output_dir ${OUTPUT_DIR} \
     --fastqc_args "--outdir ${OUTPUT_DIR}/FASTQC/TRIM" \
     ${OUTPUT_DIR}/${FASTQ}
```

**TRIM-GALORE (Paired Ended)** Collapse source

```
trim_galore \
     --paired  \
     --gzip \
     --cores ${LSB_MAX_NUM_PROCESSORS} \
     --output_dir ${OUTPUT_DIR} \
     --fastqc_args "--outdir ${OUTPUT_DIR}/FASTQC/TRIM" \
     ${OUTPUT_DIR}/${FASTQ1} \
     ${OUTPUT_DIR}/${FASTQ2}
```

**BWA-MEM (Single Ended)** Collapse source

```
bwa mem \
     -t ${LSB_MAX_NUM_PROCESSORS} \
     -K 10000000 \
```

```
      -R "@RG\tID:${RG_ID}\tLB:${RG_LB}\tPL:${RG_PL}\tSM:${RG_SM}\tPU:${RG_PU}" \
      ${REF_FILE} \
```

```
      > ${OUTPUT_DIR}/${PREFIX}.unmarked.bam
```

**BWA-MEM (Paired Ended)** Collapse source

```
bwa mem \
      -t ${LSB_MAX_NUM_PROCESSORS} \
      -K 10000000 \
      -R "@RG\tID:${RG_ID}\tLB:${RG_LB}\tPL:${RG_PL}\tSM:${RG_SM}\tPU:${RG_PU}" \
      ${REF_FILE} \
      ${OUTPUT_DIR}/${TRIM_FASTQ1} \
      ${OUTPUT_DIR}/${TRIM_FASTQ2} \
      | samtools view -b - \
      > ${OUTPUT_DIR}/${PREFIX}.unmarked.bam
```

**BAW-ALN** Collapse source

```
bwa aln \
      -t ${LSB_MAX_NUM_PROCESSORS} \
      -f ${OUTPUT_DIR}/${TRIM_FASTQ1}.sai \
      ${REF_FILE} \
      ${OUTPUT_DIR}/${TRIM_FASTQ2}
```

**BWA-SAMSE** Collapse source

```
bwa samse \
      -r "@RG\tID:${RG_ID}\tLB:${RG_LB}\tPL:${RG_PL}\tSM:${RG_SM}\tPU:${RG_PU}" \
      ${REF_FILE} \
      ${OUTPUT_DIR}/${TRIM_FASTQ1}.sai \
      ${OUTPUT_DIR}/${TRIM_FASTQ1} \
      | samtools view -b - \
      > ${OUTPUT_DIR}/${PREFIX}.unmarked.bam
```

**BWA-SAMPE** Collapse source

```
bwa sampe \
      -r "@RG\tID:${RG_ID}\tLB:${RG_LB}\tPL:${RG_PL}\tSM:${RG_SM}\tPU:${RG_PU}" \
      ${REF_FILE} \
      ${OUTPUT_DIR}/${TRIM_FASTQ1}.sai \
      ${OUTPUT_DIR}/${TRIM_FASTQ2}.sai \
      ${OUTPUT_DIR}/${TRIM_FASTQ1} \
      ${OUTPUT_DIR}/${TRIM_FASTQ2} \
      | samtools view -b - \
      > ${OUTPUT_DIR}/${PREFIX}.unmarked.bam
```

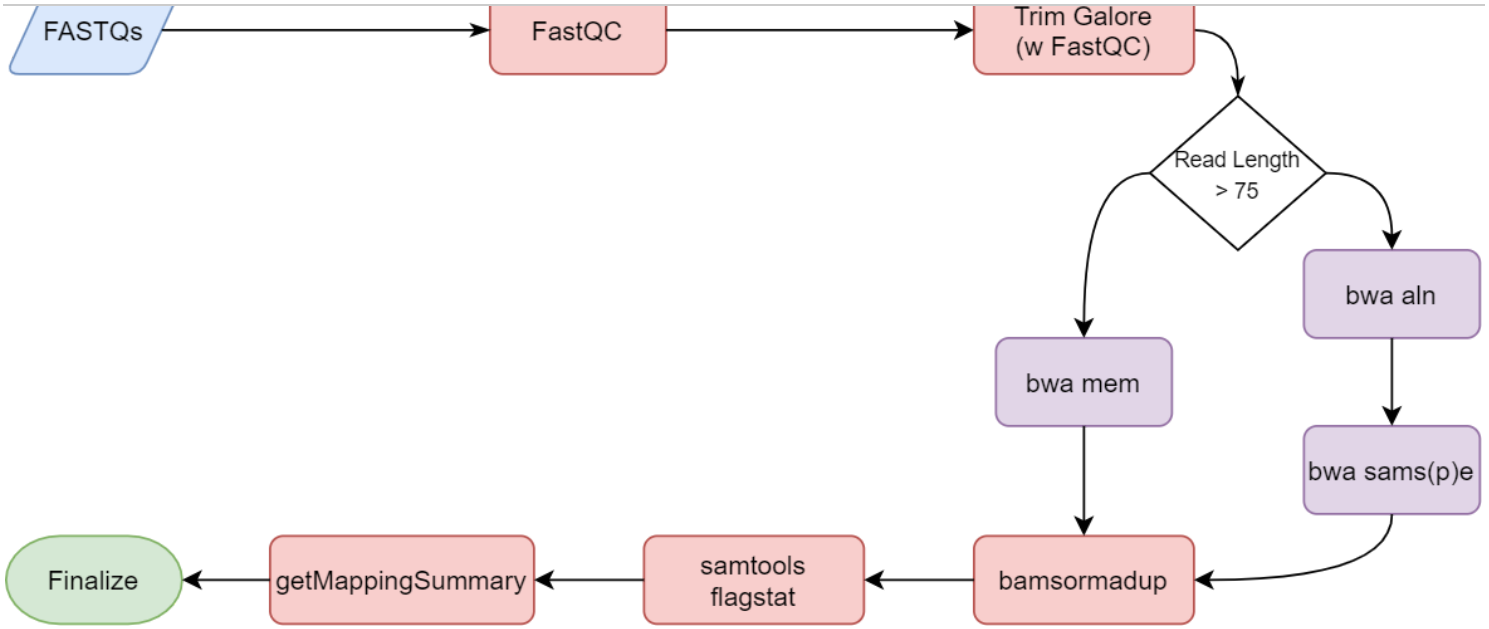
**SAMTOOLS-FLAGSTAT** Collapse source

```
samtools flagstat \
      ${OUTPUT_DIR}/${PREFIX}.bam \
      > ${OUTPUT_DIR}/${PREFIX}.flagstat.txt
```

### Deliverables -

| | NAME | TYPE | DESCRIPTION |
|---|---|---|---|
| 1 | `${PREFIX}.bam` | FILE | BAM file, coordinate-sorted, duplicates marked, created by bamsormadup |
| 2 | `${PREFIX}.bam.bai` | FILE | BAM index file, created by bamsormadup |
| 3 | `${PREFIX}.report` | FILE | Report on the mapping status and various metrics, created in accord with ChIP-seq mapping criteria |
| 4 | FASTQC/RAW | DIRECTORY | FASTQC report, untrimmed FASTQ files taken as input, created by FastQC |
| 5 | FASTQC/TRIM | DIRECTORY | FASTQC report, trimmed FASTQ files taken as input, created by FastQC |

## C. Cleavage Under Targets & Release Using Nuclease (CUT&RUN)




**Inputs -**

**TRIM-GALORE (Single Ended)** Collapse source

```
trim_galore \
     --gzip  \
     --cores ${LSB_MAX_NUM_PROCESSORS} \
     --output_dir ${OUTPUT_DIR} \
     --fastqc_args "--outdir ${OUTPUT_DIR}/FASTQC/TRIM" \
     ${OUTPUT_DIR}/${FASTQ}
```

**TRIM-GALORE (Paired Ended)** Collapse source

```
trim_galore \
     --paired  \
     --gzip \
     --cores ${LSB_MAX_NUM_PROCESSORS} \
     --output_dir ${OUTPUT_DIR} \
     --fastqc_args "--outdir ${OUTPUT_DIR}/FASTQC/TRIM" \
     ${OUTPUT_DIR}/${FASTQ1} \
     ${OUTPUT_DIR}/${FASTQ2}
```

**BWA-MEM (Single Ended)** Collapse source

```
bwa mem \
     -t ${LSB_MAX_NUM_PROCESSORS} \
     -K 10000000 \
     -R "@RG\tID:${RG_ID}\tLB:${RG_LB}\tPL:${RG_PL}\tSM:${RG_SM}\tPU:${RG_PU}" \
     ${REF_FILE} \
     ${OUTPUT_DIR}/${TRIM_FASTQ} \
     | samtools view -b - \
     > ${OUTPUT_DIR}/${PREFIX}.unmarked.bam
```

**BWA-MEM (Paired Ended)** Collapse source

```
bwa mem \
     -t ${LSB_MAX_NUM_PROCESSORS} \
     -K 10000000 \
     -R "@RG\tID:${RG_ID}\tLB:${RG_LB}\tPL:${RG_PL}\tSM:${RG_SM}\tPU:${RG_PU}" \
     ${REF_FILE} \
     ${OUTPUT_DIR}/${TRIM_FASTQ1} \
     ${OUTPUT_DIR}/${TRIM_FASTQ2} \
     | samtools view -b - \
```

```
      > ${OUTPUT_DIR}/${PREFIX}.unmarked.bam
```

```
bwa aln \
     -t ${LSB_MAX_NUM_PROCESSORS} \
     -f ${OUTPUT_DIR}/${TRIM_FASTQ1}.sai \
     ${REF_FILE} \
     ${OUTPUT_DIR}/${TRIM_FASTQ2}
```

**BWA-SAMSE** Collapse source

```

bwa samse \
     -r "@RG\tID:${RG_ID}\tLB:${RG_LB}\tPL:${RG_PL}\tSM:${RG_SM}\tPU:${RG_PU}" \
     ${REF_FILE} \
     ${OUTPUT_DIR}/${TRIM_FASTQ1}.sai \
     ${OUTPUT_DIR}/${TRIM_FASTQ1} \
     | samtools view -b - \
     > ${OUTPUT_DIR}/${PREFIX}.unmarked.bam
```

**BWA-SAMPE** Collapse source

```

bwa sampe \
     -r "@RG\tID:${RG_ID}\tLB:${RG_LB}\tPL:${RG_PL}\tSM:${RG_SM}\tPU:${RG_PU}" \
     ${REF_FILE} \
     ${OUTPUT_DIR}/${TRIM_FASTQ1}.sai \
     ${OUTPUT_DIR}/${TRIM_FASTQ2}.sai \
     ${OUTPUT_DIR}/${TRIM_FASTQ1} \
     ${OUTPUT_DIR}/${TRIM_FASTQ2} \
     | samtools view -b - \
     > ${OUTPUT_DIR}/${PREFIX}.unmarked.bam
```

**SAMTOOLS-FLAGSTAT** Collapse source

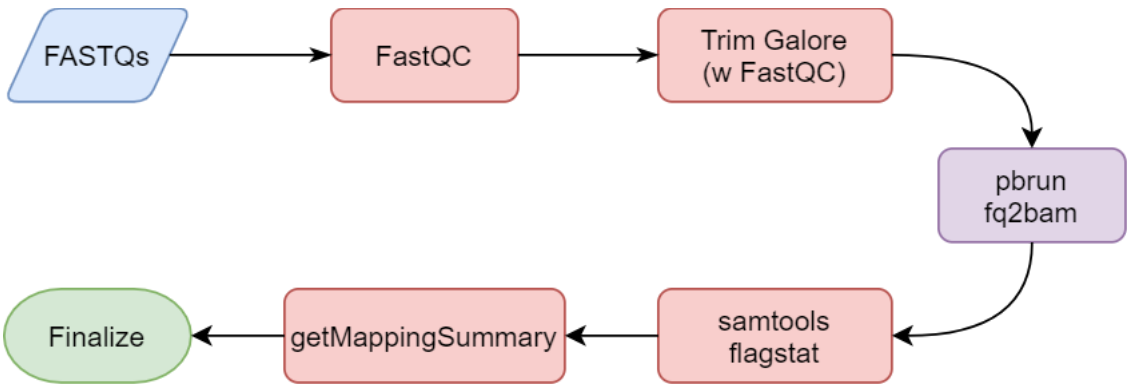
```

samtools flagstat \
     ${OUTPUT_DIR}/${PREFIX}.bam \
     > ${OUTPUT_DIR}/${PREFIX}.flagstat.txt
```

**Deliverables -**

| | NAME | TYPE | DESCRIPTION |
|---|---|---|---|
| 1 | `${PREFIX}.bam` | FILE | BAM file, coordinate-sorted, duplicates marked, created by bamsormadup |
| 2 | `${PREFIX}.bam.bai` | FILE | BAM index file, created by bamsormadup |
| 3 | `${PREFIX}.report` | FILE | Report on the mapping status and various metrics, created in accord with ChIP-seq mapping criteria |
| 4 | FASTQC/RAW | DIRECTORY | FASTQC report, untrimmed FASTQ files taken as input, created by FastQC |
| 5 | FASTQC/TRIM | DIRECTORY | FASTQC report, trimmed FASTQ files taken as input, created by FastQC |

## D. Cleavage Under Targets and Tagmentation (CUT&Tag)

**Flowchart -**



**Inputs -**

```
fastqc \
     --outdir ${OUTPUT_DIR}/FASTQC/RAW \
     --threads ${LSB_MAX_NUM_PROCESSORS} \
     --format fastq \
     --quiet  \
     ${OUTPUT_DIR}/${FASTQ}
```

**FASTQC (Paired Ended)** Collapse source

```
fastqc \
     --outdir ${OUTPUT_DIR}/FASTQC/RAW \
     --threads ${LSB_MAX_NUM_PROCESSORS} \
     --format fastq \
     --quiet  \
     ${OUTPUT_DIR}/${FASTQ1} \
     ${OUTPUT_DIR}/${FASTQ2}
```

**TRIM-GALORE (Single Ended)** Collapse source

```
trim_galore \
     --gzip  \
     --clip_R1 15 \
     --cores ${LSB_MAX_NUM_PROCESSORS} \
     --output_dir ${OUTPUT_DIR} \
     --fastqc_args "--outdir ${OUTPUT_DIR}/FASTQC/TRIM" \
     ${OUTPUT_DIR}/${FASTQ}
```

**TRIM-GALORE (Paired Ended)** Collapse source

```
 trim_galore \
      --paired  \
      --gzip  \
      --clip_R1 15 \
      --clip_R2 15 \
      --cores ${LSB_MAX_NUM_PROCESSORS} \
      --output_dir ${OUTPUT_DIR} \
      --fastqc_args "--outdir ${OUTPUT_DIR}/FASTQC/TRIM" \
      ${OUTPUT_DIR}/${FASTQ1} \
      ${OUTPUT_DIR}/${FASTQ2}
```

**PBRUN-FQ2BAM (Single Ended)** Collapse source

```
pbrun fq2bam \
     --ref ${REF_FILE} \
     --out-bam ${OUTPUT_DIR}/${PREFIX}.bam \
     --out-duplicate-metrics ${OUTPUT_DIR}/${PREFIX}.metrics.txt \
     --bwa-options "-K 100000000 -Y -M" \
     --num-gpus 2 \
     --tmp-dir ${TMP_DIR} \
     --in-fq ${TRIM_FASTQ1} "@RG\\tID:${RG_ID}\\tLB:${RG_LB}\\tPL:${RG_PL}\\tSM:${RG_SM}\\tPU:${RG_PU}"
```

**PBRUN-FQ2BAM (Paired Ended)** Collapse source

```
pbrun fq2bam \
     --ref ${REF_FILE} \
     --out-bam ${OUTPUT_DIR}/${PREFIX}.bam \
     --out-duplicate-metrics ${OUTPUT_DIR}/${PREFIX}.metrics.txt \
     --bwa-options "-K 100000000 -Y -M" \
     --num-gpus 2 \
     --tmp-dir ${TMP_DIR} \
     --in-fq ${TRIM_FASTQ1} ${TRIM_FASTQ2} "@RG\\tID:${RG_ID}\\tLB:${RG_LB}\\tPL:${RG_PL}\\tSM:${RG_SM}\\tPU:${RG_PU}"
```

**SAMTOOLS-FLAGSTAT** Collapse source

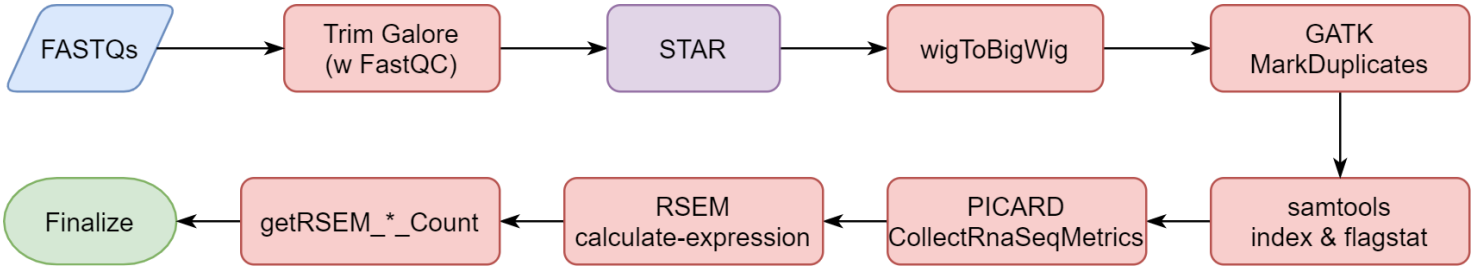
```
samtools flagstat \
     ${OUTPUT_DIR}/${PREFIX}.bam \
```

```
      > ${OUTPUT_DIR}/${PREFIX}.flagstat.txt
```

| | NAME | TYPE | DESCRIPTION |
|---|---|---|---|
| 1 | ${PREFIX}.bam | FILE | BAM file, coordinate-sorted, duplicates marked, created by bamsormadup |
| 2 | ${PREFIX}.bam.bai | FILE | BAM index file, created by bamsormadup |
| 3 | ${PREFIX}.report | FILE | Report on the mapping status and various metrics, created in accord with ATAC-seq mapping criteria |
| 4 | FASTQC/RAW | DIRECTORY | FASTQC report, untrimmed FASTQ files taken as input, created by FastQC |
| 5 | FASTQC/TRIM | DIRECTORY | FASTQC report, trimmed FASTQ files taken as input, created by FastQC |

## E. RNA-sequencing (RNA-seq)

**Flowchart -**




**Inputs -**

**TRIM-GALORE (Single Ended)** Collapse source

```
trim_galore \
     --cores ${LSB_MAX_NUM_PROCESSORS} \
     --output_dir ${OUTPUT_DIR} \
     --fastqc_args "--outdir ${OUTPUT_DIR}/FASTQC/TRIM" \
     ${OUTPUT_DIR}/${FASTQ}
```

**TRIM-GALORE (Paired Ended)** Collapse source

```
trim_galore \
     --paired  \
     --retain_unpaired \
     --cores ${LSB_MAX_NUM_PROCESSORS} \
     --output_dir ${OUTPUT_DIR} \
     --fastqc_args "--outdir ${OUTPUT_DIR}/FASTQC/TRIM" \
     ${OUTPUT_DIR}/${FASTQ1} \
     ${OUTPUT_DIR}/${FASTQ2}
```

**STAR (Single Ended)** Collapse source

```
 STAR \
      --runThreadN ${LSB_MAX_NUM_PROCESSORS} \
      --limitBAMsortRAM ${STAR_BAM_SORT_RAM} \
      --genomeDir ${REF_STAR} \
      --readFilesIn ${OUTPUT_DIR}/${TRIM_FASTQ} \
      --readFilesCommand unpigz -c -p ${LSB_MAX_NUM_PROCESSORS} \
      --outFilterType BySJout \
      --outFilterMultimapNmax 20 \
      --alignSJoverhangMin 8 \
      --alignSJstitchMismatchNmax 5 -1 5 5 \
      --alignSJDBoverhangMin 10 \
      --outFilterMismatchNmax 999 \
      --outFilterMismatchNoverReadLmax 0.04 \
      --alignIntronMin 20 \
      --alignIntronMax 100000 \
      --alignMatesGapMax 100000 \
      --genomeLoad NoSharedMemory \
      --outFileNamePrefix ${PREFIX}.STAR. \
```

```
      --outSAMmapqUnique 60 \
      --outSAMmultNmax 1 \
```

```
      --outSAMunmapped Within \
      --outSAMtype BAM SortedByCoordinate \
      --outReadsUnmapped None \
      --outSAMattrRGline ID:${RG_ID} LB:${RG_LB} PL:${RG_PL} SM:${RG_SM} PU:${RG_PU} \
      --chimSegmentMin 12 \
      --chimJunctionOverhangMin 12 \
      --chimSegmentReadGapMax 3 \
      --chimMultimapNmax 10 \
      --chimMultimapScoreRange 10 \
      --chimNonchimScoreDropMin 10 \
      --chimOutJunctionFormat 1 \
      --chimOutType Junctions WithinBAM SoftClip \
      --quantMode TranscriptomeSAM GeneCounts \
      --twopassMode Basic \
      --peOverlapNbasesMin 12 \
      --peOverlapMMp 0.1 \
      --outWigType wiggle \
      --outWigStrand ${STRAND_TYPE} \
      --outWigNorm RPM
```

**STAR (Paired Ended)** Collapse source

```

 STAR \
      --runThreadN ${LSB_MAX_NUM_PROCESSORS} \
      --limitBAMsortRAM ${STAR_BAM_SORT_RAM} \
      --genomeDir ${REF_STAR} \
      --readFilesIn ${OUTPUT_DIR}/${TRIM_FASTQ1} ${OUTPUT_DIR}/${TRIM_FASTQ2} \
      --readFilesCommand unpigz -c -p ${LSB_MAX_NUM_PROCESSORS} \
      --outFilterType BySJout \
      --outFilterMultimapNmax 20 \
      --alignSJoverhangMin 8 \
      --alignSJstitchMismatchNmax 5 -1 5 5 \
      --alignSJDBoverhangMin 10 \
      --outFilterMismatchNmax 999 \
      --outFilterMismatchNoverReadLmax 0.04 \
      --alignIntronMin 20 \
      --alignIntronMax 100000 \
      --alignMatesGapMax 100000 \
      --genomeLoad NoSharedMemory \
      --outFileNamePrefix ${PREFIX}.STAR. \
      --outSAMmapqUnique 60 \
      --outSAMmultNmax 1 \
      --outSAMstrandField intronMotif \
      --outSAMattributes NH HI AS nM NM MD \
      --outSAMunmapped Within \
      --outSAMtype BAM SortedByCoordinate \
      --outReadsUnmapped None \
      --outSAMattrRGline ID:${RG_ID} LB:${RG_LB} PL:${RG_PL} SM:${RG_SM} PU:${RG_PU} \
      --chimSegmentMin 12 \
      --chimJunctionOverhangMin 12 \
      --chimSegmentReadGapMax 3 \
      --chimMultimapNmax 10 \
      --chimMultimapScoreRange 10 \
      --chimNonchimScoreDropMin 10 \
      --chimOutJunctionFormat 1 \
      --chimOutType Junctions WithinBAM SoftClip \
      --quantMode TranscriptomeSAM GeneCounts \
      --twopassMode Basic \
      --peOverlapNbasesMin 12 \
      --peOverlapMMp 0.1 \
      --outWigType wiggle \
      --outWigStrand ${STRAND_TYPE} \
      --outWigNorm RPM
```

**WIG2BIGWIG** Collapse source

```

wigToBigWig \
     ${OUTPUT_DIR}/${PREFIX}.${SUFFIX_WIG1}.wig \
```

```
      ${REF_STAR}/chrNameLength.txt \
      ${OUTPUT_DIR}/${PREFIX}.${SUFFIX_WIG1}.bw
```

```

gatk MarkDuplicates \
      -I ${OUTPUT_DIR}/${PREFIX}.${SUFFIX_BAM}.bam \
      -O ${OUTPUT_DIR}/${PREFIX}.${SUFFIX_BAM}.marked_dup.bam \
      -M ${OUTPUT_DIR}/${PREFIX}.${SUFFIX_BAM}.marked_dup.metrics.txt \
      --TMP_DIR ${TMP_DIR}
```

**SAMTOOLS-INDEX** Collapse source

```

samtools index \
      ${OUTPUT_DIR}/${PREFIX}.${SUFFIX_BAM}.marked_dup.bam
```

**PICARD-COLLECTRNASEQMETRICS** Collapse source

```

java -jar $PICARD CollectRnaSeqMetrics \
      I=${OUTPUT_DIR}/${PREFIX}.${SUFFIX_BAM}.marked_dup.bam \
      O=${OUTPUT_DIR}/${PREFIX}.${SUFFIX_BAM}.marked_dup.RNA_Metrics \
      REF_FLAT=${REF_STAR}/refFlat.txt \
      STRAND=${STRDTYPE_PICARD} \
      RIBOSOMAL_INTERVALS=${REF_STAR}/rRNA.interval.txt
```

**RSEM-CALCULATE-EXPRESSION (Single Ended)** Collapse source

```

rsem-calculate-expression \
      --num-threads ${LSB_MAX_NUM_PROCESSORS} \
      --no-bam-output  \
      --alignments  \
      --strandedness ${STRAND_TYPE} \
      ${OUTPUT_DIR}/${PREFIX}.${SUFFIX_BAM}.bam \
      ${REF_RSEM} \
      ${PREFIX}.RSEM
```

**RSEM-CALCULATE-EXPRESSION (Paired Ended)** Collapse source

```

rsem-calculate-expression \
      --num-threads ${LSB_MAX_NUM_PROCESSORS} \
      --no-bam-output  \
      --alignments  \
      --paired-end  \
      --strandedness ${STRAND_TYPE} \
      ${OUTPUT_DIR}/${PREFIX}.${SUFFIX_BAM}.bam \
      ${REF_RSEM} \
      ${PREFIX}.RSEM
```

**GET_RSEM_GENE_COUNT** Collapse source

```

getRSEMGeneCount.py \
      -g ${GENE_LIST} \
      -r ${PREFIX}
```

**GET_RSEM_ISOFORMS_COUNT** Collapse source

```

get_rsem_isoforms_count.py \
      -g ${GENE_LIST} \
      -r ${PREFIX}
```

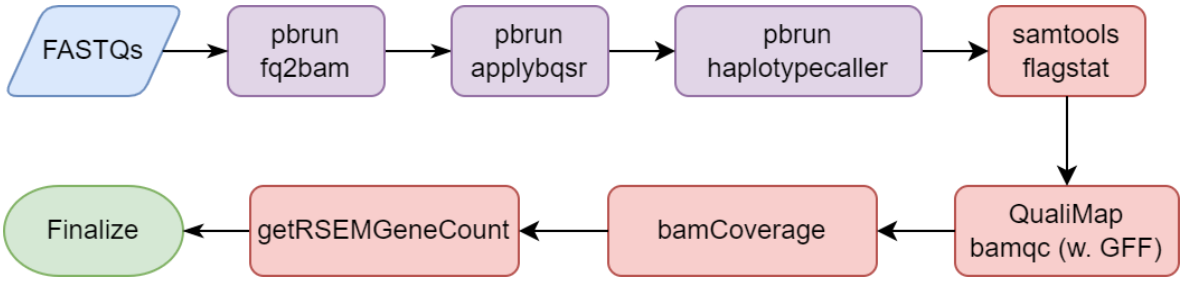
### Deliverables -

| | NAME | TYPE | DESCRIPTION |
|---|---|---|---|
| 1 | `${PREFIX}.report` | FILE | Report on the mapping status and various metrics, created in accord with RNA-seq mapping criteria |

| | NAME | TYPE | DESCRIPTION |
|---|---|---|---|
| 3 | ${PREFIX}.RSEM.genes.results | FILE | File containing gene level expression estimates, created by rsem-calculate-expression |
| 4 | ${PREFIX}.RSEM.isoforms.counts | FILE | A translation table of gene IDs with various RSEM isoform statistics (e.g. count, FPKM, TPM) |
| 5 | ${PREFIX}.RSEM.isoforms.results | FILE | File containing isoform level expression estimates, created by rsem-calculate-expression |
| 6 | ${PREFIX}.STAR.Aligned.sortedByCoord.out.marked_dup.bam | FILE | BAM file, coordinate-sorted, marked-duplication, created by STAR and gatk MarkDuplicates |
| 7 | ${PREFIX}.STAR.Aligned.sortedByCoord.out.marked_dup.bam.bai | FILE | BAM index file, created by samtools |
| 8 | ${PREFIX}.STAR.Aligned.toTranscriptome.out.bam | FILE | BAM file, with alignments translated into transcript coordinates, created by STAR |
| 9 | ${PREFIX}.STAR.Chimeric.out.junction | FILE | File containing chimeric junctions, created by STAR |
| 10 | ${PREFIX}.STAR.ReadsPerGene.out.tab | FILE | File containing counts of number reads per gene, created by STAR |
| 11 | ${PREFIX}.STAR.Signal.UniqueMultiple.str1(2).out.bw | FILE | bigWig file, converted from the wiggle file generated by STAR for Uniquely+Multiple mapped reads, created by wigToBigWig |
| 12 | ${PREFIX}.STAR.Signal.Unique.str1(2).out.bw | FILE | bigWig file, converted from the wiggle file generated by STAR for Uniquely mapped reads only, created by wigToBigWig |
| 13 | ${PREFIX}.STAR.SJ.out.tab | FILE | File containing high confidence collapsed splice junctions, created by STAR |
| 14 | ${PROJECT}_RSEM_gene_count.${UTC_TIME_STAMP}.txt | FILE | Summary table on RSEM gene "count" for all samples in ${PROJECT} |
| 15 | ${PROJECT}_RSEM_gene_FPKM.${UTC_TIME_STAMP}.txt | FILE | Summary table on RSEM gene "FPKM" for all samples in ${PROJECT} |
| 16 | ${PROJECT}_RSEM_gene_TPM.${UTC_TIME_STAMP}.txt | FILE | Summary table on RSEM gene "TPM" for all samples in ${PROJECT} |
| 17 | ${PROJECT}_RSEM_isoform_count.${UTC_TIME_STAMP}.txt | FILE | Summary table on RSEM isoform "count" for all samples in ${PROJECT} |
| 18 | ${PROJECT}_RSEM_isoform_FPKM.${UTC_TIME_STAMP}.txt | FILE | Summary table on RSEM isoform "FPKM" for all samples in ${PROJECT} |
| 19 | ${PROJECT}_RSEM_isoform_TPM.${UTC_TIME_STAMP}.txt | FILE | Summary table on RSEM isoform "TPM" for all samples in ${PROJECT} |

## F. Whole Exome Sequencing (WES)

**Flowchart -**




**Inputs -**

**PBRUN-FQ2BAM** Collapse source

```
--interval-file ${BED_FILE} \
--out-bam ${OUTPUT_DIR}/${PREFIX}.marked_dup.bam \
--out-recal-file ${OUTPUT_DIR}/${PREFIX}.marked_dup.recal.txt \
--out-duplicate-metrics ${OUTPUT_DIR}/${PREFIX}.marked_dup.metrics.txt \
--bwa-options "-K 10000000 -Y" \
--num-gpus 2 \
--tmp-dir ${TMP_DIR} \
--knownSites ${KNOWN_SITES} \
--in-fq ${OUTPUT_DIR}/${FASTQ1} ${OUTPUT_DIR}/${FASTQ2} \
       "@RG\tID:${RG_ID}\tLB:${RG_LB}\tPL:${RG_PL}\tSM:${RG_SM}\tPU:${RG_PU}"
```

**PBRUN-APPLYBQSR** Collapse source

```
pbrun applybqsr \
     --ref ${REF_FILE} \
     --in-bam ${OUTPUT_DIR}/${PREFIX}.marked_dup.bam \
     --in-recal-file ${OUTPUT_DIR}/${PREFIX}.marked_dup.recal.txt \
     --out-bam ${OUTPUT_DIR}/${PREFIX}.marked_dup.recal.bam \
     --num-gpus 2 \
     --tmp-dir ${TMP_DIR}
```

**PBRUN-HAPLOTYPECALLER** Collapse source

```
 pbrun haplotypecaller \
      --ref ${REF_FILE} \
      --interval-file ${BED_FILE} \
      --in-bam ${OUTPUT_DIR}/${PREFIX}.marked_dup.recal.bam \
      --gvcf  \
      --out-variants ${OUTPUT_DIR}/${PREFIX}.haplotype.g.vcf \
      --num-gpus 2 \
      --tmp-dir ${TMP_DIR} \
      --annotation-group StandardAnnotation \
      --annotation-group AS_StandardAnnotation \
      --annotation-group StandardHCAnnotation
```

**SAMTOOLS-FLAGSTAT** Collapse source

```
samtools flagstat \
     ${OUTPUT_DIR}/${PREFIX}.marked_dup.recal.bam \
     > ${OUTPUT_DIR}/${PREFIX}.flagstat.txt
```

**QUALIMAP-BAMQC (with features)** Collapse source

```
 qualimap bamqc \
      -bam ${OUTPUT_DIR}/${PREFIX}.marked_dup.recal.bam \
      -outdir ${OUTPUT_DIR}/BAMQC_WGFF \
      -nt ${LSB_MAX_NUM_PROCESSORS} \
      -nr 500 \
      --feature-file ${FEATURE_FILE} \
      --paint-chromosome-limits  \
      --collect-overlap-pairs  \
      --skip-duplicated  \
      --genome-gc-distr ${SPECIES} \
      --java-mem-size=${JAVAMX}M
```

**BAMCOVERAGE** Collapse source

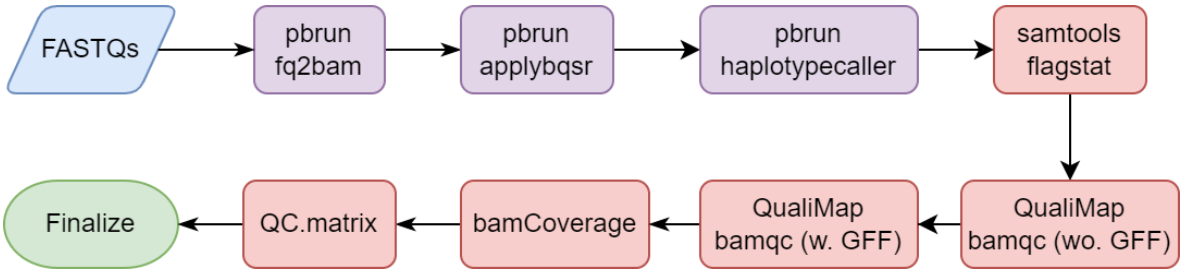
```
bamCoverage \
     --numberOfProcessors ${LSB_MAX_NUM_PROCESSORS} \
     --binSize 50 \
     --bam ${OUTPUT_DIR}/${PREFIX}.marked_dup.recal.bam \
     --outFileName ${OUTPUT_DIR}/${PREFIX}.marked_dup.recal.bw
```

**Deliverables -**

| | | | |
|---|---|---|---|
| 2 | `${PREFIX}.marked_dup.recal.metrics.txt` | FILE | File containing duplicate metrics, created by fq2bam in Parabricks' GPU-accelerated toolbox |
| 3 | `${PREFIX}.marked_dup.recal.bam` | FILE | BAM file, coordinate-sorted, duplicate marked, BQSR re-calibrated, created by tools (fq2bam + applybqsr) in Parabricks' GPU-accelerated toolbox |
| 4 | `${PREFIX}.marked_dup.recal.bam.bai` | FILE | BAM index file, created by fq2bam in Parabricks' GPU-accelerated toolbox |
| 5 | ${PREFIX}.haplotype.g.vcf.gz | FILE | gVCF file, created by haplotypecaller in Parabricks' GPU-accelerated toolbox |
| 6 | ${PREFIX}.marked_dup.bw | FILE | BigWig file, created by bamCoverage |
| 7 | `${PREFIX}.report` | FILE | Report on the mapping status and various metrics, created in accordance with WES mapping criteria |
| 8 | BAMQC_WGFF | DIRECTORY | QUALIMAP report, generated with regions of interest provided in the feature file (GFF/GTF or BED), created by qualimap |

## G. Whole Genome Sequencing (WGS)

**Flowchart -**




**Inputs -**

**PBRUN-FQ2BAM** Collapse source

```
pbrun fq2bam \
     --ref ${REF_FILE} \
     --out-bam ${OUTPUT_DIR}/${PREFIX}.marked_dup.bam \
     --out-recal-file ${OUTPUT_DIR}/${PREFIX}.marked_dup.recal.txt \
     --out-duplicate-metrics ${OUTPUT_DIR}/${PREFIX}.marked_dup.metrics.txt \
     --bwa-options "-K 100000000 -Y" \
     --num-gpus 2 \
     --tmp-dir ${TMP_DIR} \
     --knownSites ${KNOWN_SITES} \
     --in-fq ${OUTPUT_DIR}/${FASTQ1} ${OUTPUT_DIR}/${FASTQ2} \
            "@RG\tID:${RG_ID}\tLB:${RG_LB}\tPL:${RG_PL}\tSM:${RG_SM}\tPU:${RG_PU}"
```

**PBRUN-APPLYBQSR** Collapse source

```
pbrun applybqsr \
     --ref ${REF_FILE} \
     --in-bam ${OUTPUT_DIR}/${PREFIX}.marked_dup.bam \
     --in-recal-file ${OUTPUT_DIR}/${PREFIX}.marked_dup.recal.txt \
     --out-bam ${OUTPUT_DIR}/${PREFIX}.marked_dup.recal.bam \
     --num-gpus 2 \
     --tmp-dir ${TMP_DIR}
```

**PBRUN-HAPLOTYPECALLER** Collapse source

```
pbrun haplotypecaller \
     --ref ${REF_FILE} \
     --in-bam ${OUTPUT_DIR}/${PREFIX}.marked_dup.recal.bam \
     --gvcf  \
     --out-variants ${OUTPUT_DIR}/${PREFIX}.haplotype.g.vcf \
     --num-gpus 2 \
     --tmp-dir ${TMP_DIR} \
     --annotation-group StandardAnnotation \
     --annotation-group AS_StandardAnnotation \
```

```
    --annotation-group StandardHCAnnotation
```

```
samtools flagstat \
    ${OUTPUT_DIR}/${PREFIX}.marked_dup.recal.bam \
    > ${OUTPUT_DIR}/${PREFIX}.flagstat.txt
```

**QUALIMAP-BAMQC (without features)** Collapse source

```

 qualimap bamqc \
      -bam ${OUTPUT_DIR}/${PREFIX}.marked_dup.recal.bam \
      -outdir ${OUTPUT_DIR}/BAMQC_WOGFF \
      -nt ${LSB_MAX_NUM_PROCESSORS} \
      -nr 500 \
      --paint-chromosome-limits  \
      --collect-overlap-pairs  \
      --skip-duplicated  \
      --genome-gc-distr ${SPECIES} \
      --java-mem-size=${JAVAMX}G
```

**QUALIMAP-BAMQC (with features)** Collapse source

```

 qualimap bamqc \
      -bam ${OUTPUT_DIR}/${PREFIX}.marked_dup.recal.bam \
      -outdir ${OUTPUT_DIR}/BAMQC_WGFF \
      -nt ${LSB_MAX_NUM_PROCESSORS} \
      -nr 500 \
      --feature-file ${FEATURE_FILE} \
      --paint-chromosome-limits  \
      --collect-overlap-pairs  \
      --skip-duplicated  \
      --genome-gc-distr ${SPECIES} \
      --java-mem-size=${JAVAMX}G
```

**BAMCOVERAGE** Collapse source

```

bamCoverage \
    --numberOfProcessors ${LSB_MAX_NUM_PROCESSORS} \
    --binSize 50 \
    --bam ${OUTPUT_DIR}/${PREFIX}.marked_dup.recal.bam \
    --outFileName ${OUTPUT_DIR}/${PREFIX}.marked_dup.recal.bw
```

### Deliverables -

| | NAME | TYPE | DESCRIPTION |
|---|---|---|---|
| 1 | `${PREFIX}.flagstat.txt` | FILE | File containing statistics on categories based on bit flags in the FLAG field of BAM files, created by samtools |
| 2 | `${PREFIX}.marked_dup.metrics.txt` | FILE | File containing duplicate metrics, created by fq2bam in Parabricks' GPU-accelerated toolbox |
| 3 | `${PREFIX}.marked_dup.recal.bam` | FILE | BAM file, coordinate-sorted, duplicate marked, BQSR re-calibrated, created by tools (fq2bam + applybqsr) in Parabricks' GPU-accelerated toolbox |
| 4 | `${PREFIX}.marked_dup.recal.bam.bai` | FILE | BAM index file, created by fq2bam in Parabricks' GPU-accelerated toolbox |
| 5 | ${PREFIX}.haplotype.g.vcf.gz | FILE | gVCF file, created by haplotypecaller in Parabricks' GPU-accelerated toolbox |
| 6 | ${PREFIX}.marked_dup.recal.bw | FILE | BigWig file, created by bamCoverage |
| 7 | `${PREFIX}.marked_dup.recal.txt` | FILE | File containing the BQSR report, created by fq2bam in Parabricks' GPU-accelerated toolbox |
| 8 | `${PREFIX}.report` | FILE | Report on the mapping status and various metrics, created in accord with WGS mapping criteria |
| 9 | BAMQC_WGFF | DIRECTORY | QUALIMAP report, generated with regions of interest provided in the feature file (GFF/GTF or BED), created by qualimap |
| 10 | BAMQC_WOGFF | DIRECTORY | QUALIMAP report, generated without regions of interest, created by qualimap |

# IV. REPORT ON MAPPING RESULTS

| ATAC-seq | PASS | WARNING |
|---|---|---|
| MAPPED (%) | ≥ 80 | < 80 |

Recommended next step: check your sample quality or contamination when the mapping rate (MAPPED) % is lower than 80%.

| ChIP-seq | PASS | WARNING |
|---|---|---|
| MAPPED (%) | ≥ 80 | < 80 |

Recommended next step: check your sample quality or contamination when the mapping rate (MAPPED) % is lower than 80%.

| CUT&RUN | PASS | WARNING |
|---|---|---|
| MAPPED (%) | ≥ 80 | < 80 |

Recommended next step: check your sample quality or contamination when the mapping rate (MAPPED) % is lower than 80%.

| CUT&Tag | PASS | WARNING |
|---|---|---|
| MAPPED (%) | ≥ 80 | < 80 |

Recommended next step: check your sample quality or contamination when the mapping rate (MAPPED) % is lower than 80%.

| RNA-seq | PASS | WARNING |
|---|---|---|
| READS_RAW | ≥ 90 M | < 90 M |
| READS_MAPPED | ≥ 60 M | <60 M |
| BASE_RIBOSOMAL (%) | ≤ 15 | > 15 |

See St Jude In-house RNAseq data metrics for the distribution of the historical data.

if you still have furhter questions, please contact us first (cab.helpdesk@stjude.org) before reaching out to Hartwell Center.

| WES | PASS | WARNING |
|---|---|---|
| MAPPED (%) | ≥ 75 | < 75 |
| DUPLICATION (%) | ≤ 45 | > 45 |
| COVERAGE_EXON_20X (%) | ≥ 80 | <80 |

Recommended next steps: we recommend at least 80% of exons covered at least 20X. Please discuss with us (cab.helpdesk@stjude.org) when your samples failed. Note that 45% duplication rate corresponding to 95% of whole exome sequencing data in-house. That is, 95% in house whole exome data has duplication rate less than 45%. See this page for more details.

| WGS | PASS | WARNING |
|---|---|---|
| MAPPED (%) | ≥ 75 | < 75 |
| DUPLICATION (%) | ≤ 20 | > 20 |
| COVERAGE_EXON_20X (%) | ≥ 80 | <80 |

Recommended next steps: we recommend at least 80% of exons covered at least 20X. Please discuss with us (cab.helpdesk@stjude.org) when your samples failed. Note that 20% duplication rate corresponding to 95% of whole genome sequencing data in-house. That is, 95% in house whole exome data has duplication rate less than 20%. See this page for more details.

## B. Definition of Metrics

1. **ATAC-seq**
   - READS
     - RAW - The number of reads in the original FASTQs, directly quoted from the trimming report generated by "trim-galore". [If paried-ended, the number of reads is doubled to include both Read1 and Read2.]
     - TRIMMED - The number of reads used in the mapping process after trimming, directly quoted from the output of "samtools flagstat". For detailed descriptions, please refer to [1].
     - MAPPED
       - TOTAL - The total number of mapped reads, directly quoted from the output of "samtools flagstat". For detailed descriptions, please refer to [1].
       - NON-DUPLICATION - The difference between the total number of mapped reads and the total number of duplication reads, with the latter directly quoted from the output of "samtools flagstat".
   - RATE (%)
     - MAPPED - The mapping rate, directly quoted from the output of "samtools flagstat". For detailed descriptions, please refer to [1].

- DUPLICATION - The percentage fraction of the total number of duplication reads in the total number of mapped reads.

doubled to include both Read1 and Read2.]
- TRIMMED - The number of reads used in the mapping process after trimming, directly quoted from the output of "samtools flagstat". For detailed descriptions, please refer to [1].
- MAPPED
  - TOTAL - The total number of mapped reads, directly quoted from the output of "samtools flagstat". For detailed descriptions, please refer to [1].
  - NON-DUPLICATION - The difference between the total number of mapped reads and the total number of duplication reads, with the latter directly quoted from the output of "samtools flagstat".

- RATE (%)
  - MAPPED - The mapping rate, directly quoted from the output of "samtools flagstat". For detailed descriptions, please refer to [1].
  - DUPLICATION - The percentage fraction of the total number of duplication reads in the total number of mapped reads.

3. **CUT&RUN**
   - READS
     - RAW - The number of reads in the original FASTQs, directly quoted from the trimming report generated by "trim-galore". [If paried-ended, the number of reads is doubled to include both Read1 and Read2.]
     - TRIMMED - The number of reads used in the mapping process after trimming, directly quoted from the output of "samtools flagstat". For detailed descriptions, please refer to [1].
     - MAPPED
       - TOTAL - The total number of mapped reads, directly quoted from the output of "samtools flagstat". For detailed descriptions, please refer to [1].
       - NON-DUPLICATION - The difference between the total number of mapped reads and the total number of duplication reads, with the latter directly quoted from the output of "samtools flagstat".
   - RATE (%)
     - MAPPED - The mapping rate, directly quoted from the output of "samtools flagstat". For detailed descriptions, please refer to [1].
     - DUPLICATION - The percentage fraction of the total number of duplication reads in the total number of mapped reads.
   - FRAGMENTS PROPERLY-PAIRED - Paired-end reads that aligned in opposite orientations(head-to-head) on the same reference sequence (chromosome). The reads may overlap to some extent [2].
4. **CUT&Tag**
   - READS
     - RAW - The number of reads in the original FASTQs, directly quoted from the trimming report generated by "trim-galore". [If paried-ended, the number of reads is doubled to include both Read1 and Read2.]
     - TRIMMED - The number of reads used in the mapping process after trimming, directly quoted from the output of "samtools flagstat". For detailed descriptions, please refer to [1].
     - MAPPED
       - TOTAL - The total number of mapped reads, directly quoted from the output of "samtools flagstat". For detailed descriptions, please refer to [1].
       - NON-DUPLICATION - The difference between the total number of mapped reads and the total number of duplication reads, with the latter directly quoted from the output of "samtools flagstat".
   - RATE (%)
     - MAPPED - The mapping rate, directly quoted from the output of "samtools flagstat". For detailed descriptions, please refer to [1].
     - DUPLICATION - The percentage fraction of the total number of duplication reads in the total number of mapped reads.
   - FRAGMENTS PROPERLY-PAIRED - Paired-end reads that aligned in opposite orientations(head-to-head) on the same reference sequence (chromosome). The reads may overlap to some extent [2].
5. **RNA-seq**
   - READS
     - RAW - The number of reads in the original FASTQs, directly quoted from the trimming report generated by "trim-galore". [If paried-ended, the number of reads is doubled to include both Read1 and Read2.]
     - TRIMMED - The number of reads used in the mapping process after trimming, directly quoted from the final log file of "STAR". For detailed descriptions, please refer to [3].
     - MAPPED - The sum of numbers of reads for 1) uniquely, 2) multiple mapped reads, both directly quoted from the final log file of "STAR". For detailed descriptions, please refer to [3].
     - UNMAPPED - The sum of numbers of reads for 1) too many mismatches, 2) too short, 3) other unmapped reads, all directly quoted from the final log file of "STAR". For detailed descriptions, please refer to [3].
   - RATE (%)
     - MAPPED - The percentage fraction of the total number of MAPPED reads in the total number of TRIMMED reads.
     - UNMAPPED - The percentage fraction of the total number of UNMAPPED reads in the total number of TRIMMED reads.
     - DUPLICATION - The duplication rate, directly quoted from the output of "gatk MarkDuplicates". For detailed descriptions, please refer to [4].
   - BASE (%)
     - RIBOSOMAL - The percentage fraction of the total number of aligned PF bases that are mapped to regions encoding ribosomal RNA, directly quoted from the output of "PICARD CollectRnaSeqMetrics". For detailed descriptions, please refer to [5] and [6].
     - CODING - The percentage fraction of the total number of aligned PF bases that are mapped to protein coding regions of genes, directly quoted from the output of "PICARD CollectRnaSeqMetrics". For detailed descriptions, please refer to [5] and [6].
     - INTRON - The percentage fraction of the total number of aligned PF bases that correspond to gene introns, directly quoted from the output of "PICARD CollectRnaSeqMetrics". For detailed descriptions, please refer to [5] and [6].
     - UTR - The percentage fraction of the total number of aligned PF bases that mapped to untranslated regions (UTR) of genes, directly quoted from the output of "PICARD CollectRnaSeqMetrics". For detailed descriptions, please refer to [5] and [6].
     - INTERGENIC - The percentage fraction of the total number of aligned PF bases that are mapped to intergenic regions of genomic DNA, directly quoted from the output of "PICARD CollectRnaSeqMetrics". For detailed descriptions, please refer to [5] and [6].
6. **WES**
   - READS
     - RAW - The number of reads in the original FASTQs, directly quoted from the output of "samtools flagstat". For detailed descriptions, please refer to [1].
     - MAPPED - The total number of mapped reads, directly quoted from the output of "samtools flagstat". For detailed descriptions, please refer to [1].
   - RATE (%)
     - MAPPED - The mapping rate, directly quoted from the output of "samtools flagstat". For detailed descriptions, please refer to [1].
     - DUPLICATION - The duplication rate, calculated as (UNPAIRED_READ_DUPLICATES + 2 * READ_PAIR_DUPLICATES) / (UNPAIRED_READS_EXAMINED + 2 * READ_PAIRS_EXAMINED), all values used are directly quoted from the MarkDup matrics file generated by "fq2bam".
   - COVERAGE
     - MEAN - The mean coverageData, directly quoted from the output of "qualimap bamqc". For detailed descriptions, please refer to [7].
     - STANDARD - The std coverageData, directly quoted from the output of "qualimap bamqc". For detailed descriptions, please refer to [7].
     - OVERALL (%)
       - Simple Repeat of the contents from EXON (%).
     - EXON (%)
       - 10X - The percentage of reference with a coverageData >= 10X, directly quoted from the output of "qualimap bamqc". For detailed descriptions, please refer to [7].
       - 20X - The percentage of reference with a coverageData >= 20X, directly quoted from the output of "qualimap bamqc". For detailed descriptions, please refer to [7].

        - 30X - The percentage of reference with a coverageData >= 30X, directly quoted from the output of "qualimap bamqc". For detailed descriptions, please refer to [7].
    - MEDIAN_INSERT_SIZE - The median insert size, directly quoted from the output of "qualimap bamqc". For detailed descriptions, please refer to [7].

---

- READS
    - RAW - The number of reads in the original FASTQs, directly quoted from the output of "samtools flagstat". For detailed descriptions, please refer to [1].
    - MAPPED - The total number of mapped reads, directly quoted from the output of "samtools flagstat". For detailed descriptions, please refer to [1].
- RATE (%)
    - MAPPED - The mapping rate, directly quoted from the output of "samtools flagstat". For detailed descriptions, please refer to [1].
    - DUPLICATION - The duplication rate, calculated as (UNPAIRED_READ_DUPLICATES + 2 * READ_PAIR_DUPLICATES) / (UNPAIRED_READS_EXAMINED + 2 * READ_PAIRS_EXAMINED), all values used are directly quoted from the MarkDup matrics file generated by "fq2bam".
- COVERAGE
    - MEAN - The mean coverageData, directly quoted from the output of "qualimap bamqc". For detailed descriptions, please refer to [7].
    - STANDARD - The std coverageData, directly quoted from the output of "qualimap bamqc". For detailed descriptions, please refer to [7].
    - OVERALL (%)
        - 10X - The percentage of reference with a coverageData >= 10X, directly quoted from the output of "qualimap bamqc". For detailed descriptions, please refer to [7].
        - 20X - The percentage of reference with a coverageData >= 20X, directly quoted from the output of "qualimap bamqc". For detailed descriptions, please refer to [7].
        - 30X - The percentage of reference with a coverageData >= 30X, directly quoted from the output of "qualimap bamqc". For detailed descriptions, please refer to [7].
    - EXON (%)
        - 10X - The percentage of reference with a coverageData >= 10X, directly quoted from the output of "qualimap bamqc". For detailed descriptions, please refer to [7].
        - 20X - The percentage of reference with a coverageData >= 20X, directly quoted from the output of "qualimap bamqc". For detailed descriptions, please refer to [7].
        - 30X - The percentage of reference with a coverageData >= 30X, directly quoted from the output of "qualimap bamqc". For detailed descriptions, please refer to [7].
- MEDIAN_INSERT_SIZE - The median insert size, directly quoted from the output of "qualimap bamqc". For detailed descriptions, please refer to [7].
- MEAN_MAPPING_QUALITY - The mean mapping quality, directly quoted from the output of "qualimap bamqc". For detailed descriptions, please refer to [7].

#### Reference


- [1] Samtools Tool: http://www.htslib.org/doc/samtools-flagstat.html
- [2] ATAC-seq Guidelines: https://informatics.fas.harvard.edu/atac-seq-guidelines.html
- [3] STAR Manual: https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf
- [4] GATK Tool: https://gatk.broadinstitute.org/hc/en-us/articles/360037438351-MarkDuplicates-Picard-
- [5] Picard Tool: https://broadinstitute.github.io/picard/command-line-overview.html#CollectRnaSeqMetrics
- [6] Picard Metrics: http://broadinstitute.github.io/picard/picard-metric-definitions.html#RnaSeqMetrics
- [7] Qualimap Manual: http://qualimap.bioinfo.cipf.es/doc_html/analysis.html#bam-qc
- [8] deepTools Manual: https://deeptools.readthedocs.io/en/develop/content/tools/bamCoverage.html


# V. EXTRA NOTES

## A. HPCF Queuing Information

The queue chosen for a specific AutoMapper operation is selected accordingly with the sequencing type of the input data. Currently, the general-purpose queues - "standard" and "cab_auto" - with HPCF SLA support have been used to process sequencing data from ChIP-seq, CUT&RUN, and RNA-seq. To process ATAC-seq, CUT&Tag, and WES/WGS sequencing data which are much large in size comparing to the other three types, CAB AutoMapper utilizes the GPU-based tools in Parabricks' toolbox. Jobs for ATAC-seq, CUT&Tag, and WES/WGS are therefore submitted by activating an application profile that enables the Parabricks capability on GPU nodes.

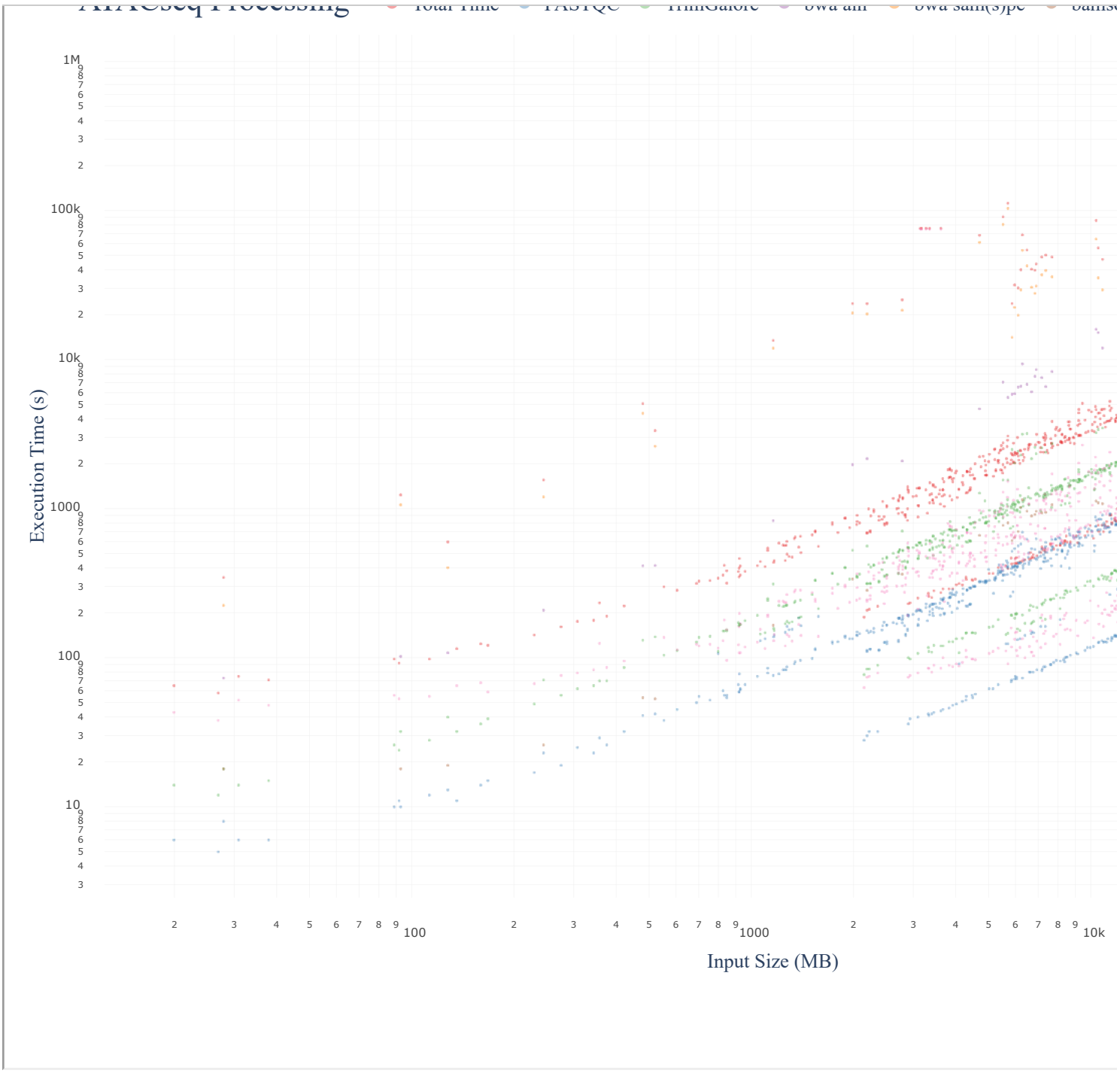
## B. Structure of Output Directory

The execution of the pipeline tasks is organized in such a way that each sample in the project will be run as a single job on a single HPCF node with multi-threading support (if applicable). The outputs of the pipeline are organized as following:

- Output directory - /scratch_space/${USER_ID}/${ASSAY_TYPE}/${JOB_NAME}/${PIGRP}-${SRM_ORDER#}-${SEQUENCING_TYPE}/${SAMPLE_NAME}/${JOB_ID}/;
- For each sample, re-running the pipeline will create under the same ./${SAMPLE_NAME} a different run-time folder with the new LSF job ID;

## C. Performance Metrics

**Period of Operation:** September 23rd, 2019 - November 30th, 2021




# VI. APPENDICES

## A. Environment Variables

| Name | Value | Description | Example |
|---|---|---|---|
| AUTO_DIR | - | Root location for AutoMapper operation | /research/rgs01/applications/hpcf/authorized_apps/cab/Automation |
| REF_DIR | - | Top location for the genreral reference | ${AUTO_DIR}/REF/Homo_sapiens/Gencode_ERCC92/r31 |

| Name | Value | Description | Example |
|---|---|---|---|
| REF_STAR | - | Location for the STAR-prepared references | ${REF_DIR}/STAR-index/2.7 |
| REF_RSEM | - | Location for the RSEM-prepared references | ${REF_DIR}/RSEM-index/v1.3.1/GRCh38.primary_assembly.genome |
| REF_GFF | - | Location for the feature file with regions of interest in GFF/GTF or BED format | ${REF_DIR}/hg38_UCSC_CDS_exons_modif_canonical.bed |
| GENE_LIST | - | Location for the gene ID translation table to assemble statistics from RSEM | ${REF_DIR}/gencode.v31.primary_assembly.annotation.gtf.gene |
| STRDTYPE_STAR | Stranded \| Unstranded | Strandedness information for STAR mapping | Unstranded |
| STRDTYPE_PICARD | NONE \| FIRST_READ_TRANSCRIPTION_STRAND \| SECOND_READ_TRANSCRIPTION_STRAND | Strandedness information for Picard CollectRnaSeqMetrics | NONE |
| STRDTYPE_RSEM | none \| forward \| reverse | Strandedness information for RSEM calculating gene expression | reverse |
| KNOWN_SITES | - | Location for the compressed vcf files for known SNPs and indels | ${REF_DIR}/KNOWN_SITE/Homo_sapiens_assembly38.known_indels.vcf.gz |
| SPECIES | HUMAN \| MOUSE | Species to compare with genome GC distribution | HUMAN |
| JAVAMX | - | Argument to set Java memory heap size | 40G |
| RG_ID | - | Meta-info for the "ID" field in the RG group in BAM file | HNY22DSXX.1 |
| RG_LB | - | Meta-info for the "LB" field in the RG group in BAM file | LIB01 |
| RG_PL | ILLUMINA | Meta-info for the "PL" field in the RG group in BAM file | ILLUMINA |
| RG_SM | - | Meta-info for the "SM" field in the RG group in BAM file | 1778713_DYE2664 |
| RG_PU | - | Meta-info for the "PU" field in the RG group in BAM file | HNY22DSXX.1 |

## B. Example Email for Mapping Report

> **From:** Yawei.Hui@STJUDE.ORG <Yawei.Hui@STJUDE.ORG>
> **Sent:** Friday, December 13, 2019 10:23 PM
> **To:** Hui, Yawei <Yawei.Hui@STJUDE.ORG>
> **Subject:** [CAB DevOps] Automatic Processing: Mapping Completed (HHERZ-163062-STRANDED)
>
> Dear Yawei Hui,
>
> The mapping processes on samples in your project HHERZ-163062-STRANDED were completed. The outputs for successfully mapped samples are delivered to
>
> /research/rgs01/project_space/cab/automapper/common/yhui/HHERZ-163062-STRANDED
>
> A brief report on the mapping results is quoted below.
>
> Please refer to CAB Knowledge Base for general questions on QC metrics. For more questions or further quality control on these samples, please contact CAB Help Desk with "[Transcriptomics]" in your email subject.
>
> Thank you.

**********

| METRIC | PASS | WARNING |
|---|---|---|
| READS_RAW[1] | ≥ 90 M | < 90 M |
| READS_MAPPED[3] | ≥ 60 M | < 60 M |
| BASE_RIBOSOMAL (%) | ≤ 15 | > 15 |

**MAPPING METRICS:**

| INDEX | SAMPLE | STATUS | READS | | | | RATE (%) | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | RAW[1] | TRIMMED[2] | MAPPED[3] | UNMAPPED[4] | MAPPED[3] | UNMAPPED[4] | DUPLICATION |
| 0 | 1713744_SJMMNORM059242_C2-mESC_WT_replicate1_052919 | PASS | 94512214 | 93447320 | 81527636 | 11666112 | 87.24 | 12.48 | 27.76 |
| 1 | 1713745_SJMMNORM059242_C3-mESC_WT_replicate2_052919 | PASS | 126966370 | 126594778 | 110478764 | 15764324 | 87.27 | 12.45 | 41.17 |
| 2 | 1713746_SJMMNORM059242_C4-mESC_Zfp281KO26_replicate1_052919 | PASS | 182556240 | 182446448 | 164007302 | 17937104 | 89.89 | 9.83 | 37.43 |
| 3 | 1713747_SJMMNORM059242_C5-mESC_Zfp281KO26_replicate2_052919 | PASS | 181255054 | 180350084 | 160404376 | 19421678 | 88.94 | 10.77 | 34.82 |

**NOTE 1: Reads included in the original FASTQs.**

**NOTE 2: Reads extracted from the raw reads after adapter trimming by using "trim-galore".**

**NOTE 3: Reads include uniquely and multiple mapped reads while excluding those mapped to too many loci.**

**NOTE 4: Reads include unmapped reads due to 1) too many mismathes; 2) too short; 3) all other.**

No labels

## 9 Comments

Unknown User (jchen4)
What are the versions of STAR and other programs? Thanks!

Hui, Yawei
The current version of STAR used in the AutoMapper RNAseq pipeline is 2.7.5a.

By "other programs", which specific ones are you referring to?

Unknown User (jchen4)
Thanks Yawei for your prompt reply!

Unknown User (jchen4)
For RNA-seq, how the STRDTYPE_* is determined? Is it inferred from sequencing data or SRM order? Thanks!

Hui, Yawei
All information about strandedness in RNAseq mapping is provided in advance by either the HC/SRM or end-users during the submission.

Unknown User (jchen4)
Thank you, @Hui, Yawei and @Wu, Gang . My coworker has a SRM submission with Application field as "RNA-seq Total Stranded", but it is labelled as "unstranded" in the raw fastq folder and Automapper output folder. Perhaps, we can push the data through stranded pipeline? I will send you more information via email.

Wu, Gang
@Unknown User (jchen4) , try this RSeQC

Vegesana, Kasi
Hello,

I just found out about this service. My team recently got back some data from Hartwell, and we ran pseudo-alignment quantification using Salmon. Is it possible for us to put the same raw data in our <pigroup>_auto folder to generate alignments, and counts using RSEM?

Fan, Yiping

yes, we need SRM#, species, strand info from you, please work with @Hui, Yawei