Alignment Workflow

Genomics and Molecular Pathology Lab

UT Southwestern Medical Center

Author: Brandi Cantarel

Version: 1.0

Effective Date: 06/08/2017

1. Introduction

This SOP describes the Alignment Workflow used for NGS analysis in the UTSW Genomics and Molecular Pathology Lab (GMPL). This pipeline aligns genomic data to the human genome, optimizes this alignment to improve quality and provides quality metrics on the sequence quality. The alignments from this workflow will be used for tumor-only and tumor/normal variation workflow.

2. Requirements

2.1. Data requirements

The pipeline requires as input FASTQ files generated by genome/exome/target panel sequencing. The pipeline also requires the sequence of a reference genome, currently Human GRCh38. The Reference was downloaded and indexed using the commands:

1. bwa.kit/run-gen-ref hs38DH
2. bwa.kit/bwa index hs38DH.fa

3. Procedures

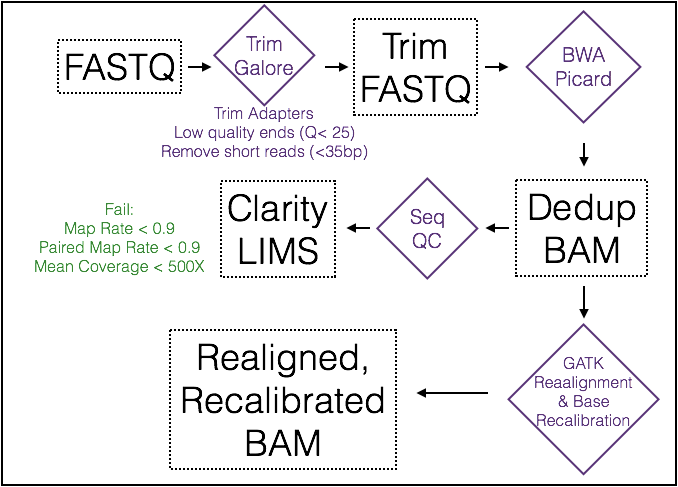


Figure 1: Procedure Overview

3.1. Trim Sequences

Trim the ends of sequences with remaining adapter or quality scores < 25. Remove any sequence less than 35bp after trimming, then generate a file for capturing information about how many sequences were trimmed.

|  |  |
| --- | --- |
| executable | trimgalore/0.4.1 cutadapt/1.9.1 parse\_trimreport.pl |
| input | gzip fastq |
| output | trimmed gzip fastq and report of number of reads remaining post trimming |
| command | >trim\_galore --paired --stringency 3 -q 25 --illumina --gzip --length 35 ${R1} ${R2}  >perl $baseDir/scripts/parse\_trimreport.pl trimreport\_${ct}.txt \*trimming\_report.txt |

3.2. Read Alignment to Human Reference Genome

Align sequencing reads into human reference genome considering ALT contigs, sort alignments, and mark PCR duplicates.

|  |  |
| --- | --- |
| executable | bwakit/0.7.15 samtools/intel/1.3 speedseq/20160506 picard/1.127 bwa/intel/0.7.15 |
| input | gzip fastq, reference fasta, reference ALT.sam |
| output | sorted bam |
| command | >bwa mem -M -t \$SLURM\_CPUS\_ON\_NODE -R '@RG\\tID:${fq1.baseName.split("\\.batch", 2)[0]}\\tLB:tx\\tPL:illumina\\tPU:barcode\\tSM:${fq1.baseName.spl  it("\\.batch", 2)[0]}' ${gatkref} ${fq1} ${fq2}  >sambamba sort -t \$SLURM\_CPUS\_ON\_NODE -o ${pair\_id}.bam output.unsort.bam  >java -Xmx20g -jar \$PICARD/picard.jar MarkDuplicatesWithMateCigar I=output.sort.bam O=${pair\_id}.bam M=${pair\_id}.libcomplex.txt ASSUME\_SORTED=true MINIMUM\_DISTANCE=300 |

3.3. Merger Alignments

Merge Alignments coming from the same sample, with differing barcodes, sort alignment

|  |  |
| --- | --- |
| executable | samtools/intel/1.3 speedseq/20160506 picard/1.127 bwakit/0.7.15 |
| input | gzip fastq, reference fasta, reference ALT.sam |
| output | sorted and duplicate-marked bam, splitters bam, discordants bam, and report of insert size estimation and HLA genotyping |
| command | >sambamba merge -t \$SLURM\_CPUS\_ON\_NODE merge.bam \*.bam  >sambamba sort -t \$SLURM\_CPUS\_ON\_NODE -o output.sort.bam merge.bam  >sambamba sort -N -t \$SLURM\_CPUS\_ON\_NODE -o output.nsort.bam merge.bam  >java -Xmx4g -jar \$PICARD/picard.jar CollectInsertSizeMetrics INPUT=${pair\_id}.bam HISTOGRAM\_FILE=${pair\_id}.hist.ps REFERENCE\_SEQUENCE=${gatkref} OUTPUT=${pair\_id}.hist.txt  >samtools view output.nsort.bam | k8 /cm/shared/apps/bwa/intel/0.7.15/bwakit/bwa-postalt.js -p ${pair\_id}.hla ${index\_path}/hs38DH.fa.alt &> tmp  >run-HLA ${index\_path}.hla > ${index\_path}.hla.top 2> ${index\_path}.hla.log  >touch ${pair\_id}.hla.HLA-dummy.gt  >cat ${pair\_id}.hla.HLA\*.gt | grep ^GT | cut -f2- > ${pair\_id}.hla.all |

3.4. Calculate Metrics of Quality of Alignment

Estimate alignment status values.

|  |  |
| --- | --- |
| executable | bedtools/2.25.0 picard/1.127 samtools/intel/1.3 fastqc/0.11.2 speedseq/20160506 |
| input | sorted and duplicate-marked bam, capture region bed, genome file |
| output | on-target bam, alignment status values including mapping rate, on-target rate, coverage statistics, mean mapping quality and alignment summary metrics |
| command | >fastqc -f bam ${sbam}  >sambamba flagstat -t 30 ${sbam} > ${pair\_id}.flagstat.txt  >sambamba view -t 30 -f bam -L ${capture\_bed} -o ${pair\_id}.ontarget.bam ${sbam}  >sambamba flagstat -t 30 ${pair\_id}.ontarget.bam > ${pair\_id}.ontarget.flagstat.txt  >samtools view -b -q 1 ${pair\_id}.ontarget.bam | bedtools coverage -sorted -hist -g ${index\_path}/genomefile.txt -b stdin -a ${capture\_bed} > ${pair\_id}.mapqualcov.txt  >samtools view -b -F 1024 ${pair\_id}.ontarget.bam | bedtools coverage -sorted -g ${index\_path}/genomefile.txt -a ${capture\_bed} -b stdin -hist | grep ^all > ${pair\_id}.dedupcov.txt  >java -Xmx20g -jar \$PICARD/picard.jar CollectAlignmentSummaryMetrics R=${gatkref} I=${pair\_id}.ontarget.bam OUTPUT=${pair\_id}.alignmentsummarymetrics.txt  >samtools view -F 1024 ${pair\_id}.ontarget.bam | awk '{sum+=\$5} END { print "Mean MAPQ =",sum/NR}' > ${pair\_id}.meanmap.txt |

3.5. Alignment Processing using GATK

3.5.1. Local Realignment around Indels

Realign locally reads to minimize the number of mismatching bases across all the reads.

|  |  |
| --- | --- |
| executable | gatk/3.5 samtools/intel/1.3 |
| input | reference genome fasta, on-target bam, known indel vcf |
| output | realigned bam |
| command | java -Xmx32g -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -known ${knownindel} -R ${gatkref} -o ${pair\_id}.bam.list -I ${dbam} -nt 30 -nct 1  java -Xmx32g -jar GenomeAnalysisTK.jar -I ${dbam} -R ${gatkref} --filter\_mismatching\_base\_and\_quals -T IndelRealigner -targetIntervals ${pair\_id}.bam.list -o ${pair\_id}.realigned.bam |

3.5.2. Base Quality Score Recalibration

Detect systematic errors made by the sequencer when it estimates the quality score of each base call. Build a model of covariation based on the data and a set of known variants, then adjust the base quality scores in the data based on the model.

|  |  |
| --- | --- |
| executable | gatk/3.5 samtools/intel/1.3 |
| input | reference genome fasta, realigned bam, dbsnp vcf |
| output | recalibrated bam |
| command | java -Xmx32g -jar GenomeAnalysisTK.jar -l INFO -R ${gatkref} --knownSites ${dbsnp} -I ${pair\_id}.realigned.bam -T BaseRecalibrator -cov ReadGroupCovariate -cov QualityScoreCovariate -cov CycleCovariate -cov ContextCovariate -o ${pair\_id}.recal\_data.grp -nt 1 -nct 30  java -Xmx32g -jar \$GATK\_JAR -T PrintReads -R ${gatkref} -I ${pair\_id}.realigned.bam -BQSR ${pair\_id}.recal\_data.grp -o ${pair\_id}.final.bam -nt 1 -nct 8 |

3.6. Genome Coverage

Estimate the coverage of bases by exon

|  |  |
| --- | --- |
| executable | bedtools/2.25.0 picard/1.127 samtools/intel/1.3 |
| input | Sorted bam file |
| output | exon coverage statistics |
| command | >bedtools coverage -sorted -hist -g ${index\_path}/genomefile.txt -b ${sbam} -a ${capture\_bed} > ${pair\_id}.covhist.txt  >perl $baseDir/scripts/calculate\_depthcov.pl ${pair\_id}.covhist.txt  >grep ^all ${pair\_id}.covhist.txt > ${pair\_id}.genomecov.txt |

4. Implementation

The Workflow used in this SOP can be downloaded here: <https://git.biohpc.swmed.edu/brandi.cantarel/clinseq_workflows>

Usage:

>nextflow run –with-timeline –w work\_directory alignment.nf –design samplesheet.txt –capture RefGenomeGRCh38.bed –input input\_directory –output output\_directory &>nextflow\_alignment.log &

where the work directory is the folder where the work files for each step will be deposited, the RefGenomeGRCh38.bed is the Human reference genome GRCh38 in bed format, the input directory is the name of the folder where the fastq files reside, the output directory is the name of the folder where the output files will be written, the samplesheet.txt is a tab delimited file that contains the sample id, subject id and fastq files, and the nextflow\_alignment.log is the file that keeps the log of the run and tracks workflow errors.

5. Quality Metrics

Assure nextflow\_alignment.log has no errors. If errors exist point to the folder where error occurred and run sh .command.sh to troubleshoot.

Gather the following metrics from files and assure thresholds are reached. If thresholds are not reached, flag the sample as failed and do not proceed with the sample.

File: Sequence.stats.txt

|  |  |  |  |
| --- | --- | --- | --- |
|  | Lower Threshold Values |  |  |
|  | Tumor | Normal | Saliva |
| Total\_raw | 120,000,000 | 50,000,000 | 30,000,000 |
| Map Rate | 98% | 99% | 95% |
| Proper Pair Rate | 95% | 98% | 90% |
| Percent Ontarget | 45% | 45% | 50% |
| Mean MapQ | 55 | 55 | 55 |
| Median Insert | 180-330 | 180-330 | 180-330 |
| Mean Insert | 200-300 | 200-300 | 200-300 |
| Percent 100x | 95% | 90% | 90% |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Upper Threshold Values |  |  |
|  | Tumor | Normal | Saliva |
| Indel.Rate | 0.001 | 0.001 | 0.001 |
| Error.Rate | 0.005 | 0.003 | 0.0025 |
| Median.Mismatch | 0.05 | 0.003 | 0.0025 |
| Percent.Dups | 0.05% | 0 | 0.02 |

6. Related Documents & References

* <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
  + <http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/>
* <http://bio-bwa.sourceforge.net/>
  + Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754-60.
  + <https://github.com/lh3/seqtk>
  + <https://github.com/lh3/bwa/tree/master/bwakit>
* <http://samtools.sourceforge.net/>
  + Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9.
  + Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011 Nov 1;27(21):2987-93.
* <http://lomereiter.github.io/sambamba/>
  + Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015 Jun 15;31(12):2032-4.
* <https://broadinstitute.github.io/picard/>
* <https://software.broadinstitute.org/gatk/>
  + McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297-303.
  + DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011 May;43(5):491-8.
  + Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43:11.10.1-33.
  + <http://gatkforums.broadinstitute.org/gatk>
* <http://bedtools.readthedocs.io/en/latest/>
  + Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841-2.
* <http://vcftools.sourceforge.net/index.html>
  + Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R; 1000 Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156-8.

7. Revision History