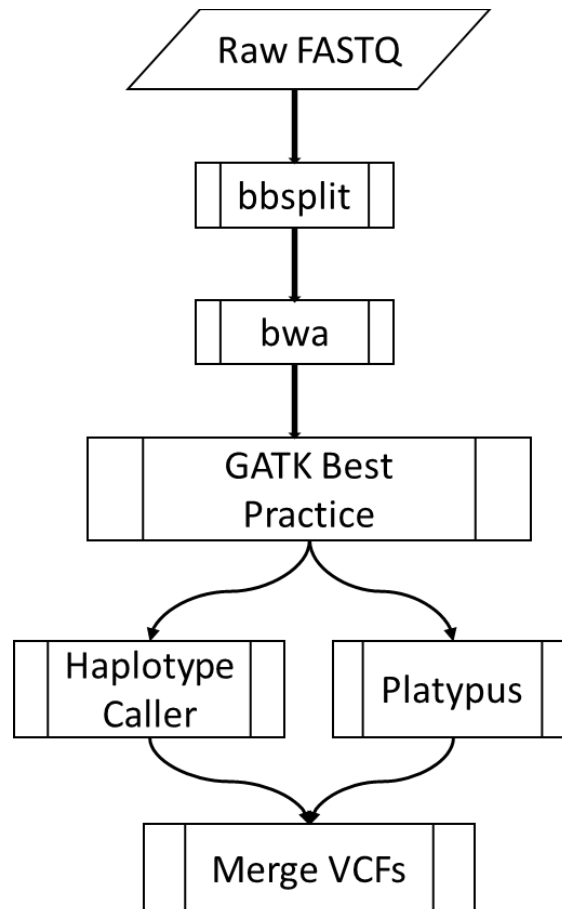


Whole Exome Sequencing Data Analysis Pipeline and Specifications	Document No.:	MCCRD-SOP0011
	Version:	2.0
	Effective Date:	11/01/2017
	Author:	MoCha

1.0 PURPOSE/SCOPE

This Standing Operating Procedure (SOP) describes the pipeline and data analysis specifications for HiSeq PDX Exome Pipeline for Patient-Derived Models used/performed by the Molecular Characterization and Clinical Assay Development Laboratory (MoCha) at the Frederick National Laboratory for Cancer Research. **This SOP is for research purposes only and no clinical samples will be processed using this SOP. Any deviation from this SOP will be noted but will not be formally documented.**

2.0 STANDARD VARIANT CALLING PIPELINE FOR PATIENT-DERIVED MODELS (PDM) EXOME



Whole Exome Sequencing Data Analysis Pipeline and Specifications	Document No.:	MCCRD SOP0011
	Version:	2.0
	Effective Date:	11/01/2017
	Page No.:	Page 2 of 4

3.0 PROCEDURE

- 3.1** FASTQ files are generated with bcl2fastq (version: 2.17.1.14, Illumina). Adaptors are trimmed within this process using the default cutoff of the adapter-stringency option.
- 3.2** PDX Mouse reads are removed from the raw FASTQ files using bbsplit (bbtools v37.36).
- `bbtools bbsplit build=1 -Xmx10g path=<indexPath> in1=Sample.R1.fastq.gz in2=Sample.R2.fastq.gz basename=Sample_%_#.fastq.gz refstats=Sample_stat.txt`
- 3.3** FASTQ files from step 3.2 are mapping against human genome reference hg19 using the BWA (v 0.7.10), resulting sam file is converted to sorted bam file using sentieon(sentieon-genomics/201711.01)
- `bwa mem -M -t <threads> -R '@RG\tID:Sample\tSM:Sample\tLB:Sample\tPL:Illumina' ucsc.hg19.fasta Sample.R1.fastq.gz Sample.R2.fastq.gz | sentieon util sort -o Sample.bwa.bam -t <threads> --sam2bam -i -`
- 3.4** Custom Script is used to assess mapping quality, mean coverage and other QC statistics on the sample.
- Available upon request
- 3.5** GATK Best Practice
- 3.5..1 Mark Duplicates.
- `sentieon driver -t <threads> -i Sample.bwa.bam --algo LocusCollector --fun score_info Sample.markdup.txt.tmp`
 - `sentieon driver -t <threads> -i Sample.bwa.bam --algo Dedup --score_info Sample.markdup.txt.tmp --metrics Sample.markdup.txt Sample.dd.bam`
- 3.5..2 Local Realignment.
- `sentieon driver -t <threads> -r ucsc.hg19.fasta -i Sample.dd.bam --algo Realigner -k {input.phase1} -k {input.mills} Sample.lr.bam`
- 3.5..3 Base Quality Score Recalibration.
- `sentieon driver -t <threads> -r ucsc.hg19.fasta -i Sample.lr.bam --algo QualCal -k {input.phase1} -k {input.mills} Sample.recalibration.matrix.txt`
 - `sentieon driver -t <threads> -r ucsc.hg19.fasta -i Sample.lr.bam -q Sample.recalibration.matrix.txt --algo QualCal -k {input.phase1} -k {input.mills} Sample.recalibrationPost.matrix.txt --algo ReadWriter Sample.bwa.final.bam`

Whole Exome Sequencing Data Analysis Pipeline and Specifications	Document No.:	MCCRD SOP0011
	Version:	2.0
	Effective Date:	11/01/2017
	Page No.:	Page 3 of 4

3.6 Variant Calling

3.6..1 HaplotypeCaller

- sentieon driver -t <threads> -r ucsc.hg19.fasta -i Sample.bwa.final.bam --interval Sureselect.v5.padded.bed --algo Haplotyper -d {input.dbsnp} --min_base_qual 20 --min_map_qual 30 Sample.HC_DNASeq.raw.vcf
- merge 2 calls made on consecutive bases to one
 - Custom script available upon request

3.6..2 Platypus

- platypus callVariants -nCPU=<threads> --bufferSize=1000000 --maxReads=100000000 --bamFiles=Sample.bwa.final.bam --regions=Sureselect.v5.padded.bed --output=Sample.Platypus.raw.vcf --refFile=ucsc.hg19.fasta
- Select PASS Variants
- Remove 3 base-substitutions.

3.7 Merging VCFs from HaplotypeCaller and Platypus

- Recode Tags in the vcf files

3.7..1 Custom perl script available upon request

- GATK Version: 3.4-0
- java -Xmx5g -XX:ParallelGCThreads=<threads> -jar GATK.jar -T CombineVariants -R ucsc.hg19.fasta -dcov 10000 --variant:HC Sample.HC_DNASeq.raw.vcf --variant:Platypus Sample.Platypus.raw.vcf -o Sample.vcf -genotypeMergeOptions PRIORITIZE -priority HC,Platypus -nt <threads>

4.0 REFERENCES

1. Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, 25, 2078-9. PMID: 19505943
2. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA, 2010 *Genome Research* 20:1297-303
3. Justin M Zook, Brad Chapman, Jason Wang, David Mittelman, Oliver Hofmann, Winston Hide & Marc Salit: Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls, *Nature Biotechnology* 32, 246–251 (2014) doi:10.1038/nbt.2835
4. Link to Sentieon tools: <https://www.sentieon.com/>

Whole Exome Sequencing Data Analysis Pipeline and Specifications	Document No.:	MCCRD SOP0011
	Version:	2.0
	Effective Date:	11/01/2017
	Page No.:	Page 4 of 4

5. Link to GATK resource bundle. All files in brackets { } in the SOP are downloaded from <https://software.broadinstitute.org/gatk/download/bundle>
6. Link to Exome/Capture/RNASeq Pipeline on GitHub: https://github.com/FNL-MoCha/nextgenseq_pipeline