DNAseq Pipeline – Single Sample

Analysis Step	GATK Best Practices Pipeline	Sentieon DNAseq
Read alignment	bwa mem -M -R \"@RG\tID:$group\tSM:$sample\tPL:$pl\" -t 32 $fasta $fastq_1 $fastq_2 | samtools view -Sb - >align.bam samtools sort -@ 32 align.bam sorted	bwa mem -M -R \"@RG\tID:$group\tSM:$sample\tPL:$pl\" -t 32 $fasta $fastq_1 $fastq_2 | sentieon util sort -o sorted.bam -t 32 --sam2bam -i –
Sample metrics collection	java -jar ${picard}/CollectAlignmentSummaryMetrics.jar INPUT=sorted.bam OUTPUT=aln_metrics.txt REFERENCE_SEQUENCE=$fasta ADAPTER_SEQUENCE=null VALIDATION_STRINGENCY=SILENT java -jar ${picard}/CollectGcBiasMetrics.jar INPUT=sorted.bam OUTPUT=gc_metrics.txt SUMMARY_OUTPUT=gc_summary.txt CHART_OUTPUT=gcbias.pdf REFERENCE_SEQUENCE=$fasta ASSUME_SORTED=true VALIDATION_STRINGENCY=SILENT java -jar ${picard}/MeanQualityByCycle.jar INPUT=sorted.bam OUTPUT=mq_metrics.txt CHART_OUTPUT=meanq_cycle.pdf REFERENCE_SEQUENCE=$fasta VALIDATION_STRINGENCY=SILENT PF_READS_ONLY=true java -jar ${picard_folder}/QualityScoreDistribution.jar INPUT=sorted.bam OUTPUT=qd_metrics.txt CHART_OUTPUT=qscore_dist.pdf REFERENCE_SEQUENCE=$fasta VALIDATION_STRINGENCY=SILENT PF_READS_ONLY=true java -jar ${picard_folder}/CollectInsertSizeMetrics.jar INPUT=sorted.bam OUTPUT=is_metrics.txt REFERENCE_SEQUENCE=$fasta HISTOGRAM_FILE=is_histogram.txt	sentieon driver -r $fasta -t 32 -i sorted.bam --algo MeanQualityByCycle mq_metrics.txt --algo QualDistribution qd_metrics.txt --algo GCBias --summary gc_summary.txt gc_metrics.txt --algo AlignmentStat aln_metrics.txt --algo InsertSizeMetricAlgo is_metrics.txt sentieon plot metrics -o metrics-report.pdf gc=gc_metrics.txt qd=qd_metrics.txt mq=mq_metrics.txt isize=is_metrics.txt
Duplicate read removal	java -jar ${picard}/MarkDuplicates.jar M=dup_reads I=sorted.bam O=deduped.bam samtools index deduped.bam	sentieon driver -t 32 -i sorted.bam --algo LocusCollector --fun score_info score.txt sentieon driver -t 32 -i sorted.bam --algo Dedup --rmdup --score_info score.txt deduped.bam
Indel realignment	java -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -R $fasta -I deduped.bam -known $Mills -o realigner.intervals java -jar GenomeAnalysisTK.jar -T IndelRealigner -R $fasta -I deduped.bam -known $Mills -targetIntervals realigner.intervals -o realigned.bam	sentieon driver -r $fasta -t 32 -i deduped.bam --algo Realigner -k $Mills realigned.bam
Base quality score recalibration	java -jar GenomeAnalysisTK.jar -T BaseRecalibrator -nct 32 -R $fasta -I realigned.bam -knownSites $dbsnp -knownSites $Mills -o recal.table java -jar GenomeAnalysisTK.jar -T PrintReads -nct 32 -R $fasta -I realigned.bam -BQSR recal.table -o recaled.bam java -jar GenomeAnalysisTK.jar -T BaseRecalibrator -nct 32 -R $fasta -I realigned.bam -knownSites $dbsnp -knownSites $Mills -BQSR recal.table -o after_recal.table java -jar GenomeAnalysisTK.jar -T AnalyzeCovariates -R $fasta -before recal.table -after after_recal.table -plots recal_plots.pdf	sentieon driver -r $fasta -t 32 -i realigned.bam --algo QualCal -k $dbsnp -k $Mills recal_data.table sentieon driver -r $fasta -t 32 -i realigned.bam -q recal_data.table --algo QualCal -k $dbsnp -k $Mills recal_data.table.post --algo ReadWriter recaled.bam sentieon driver -t 32 --algo QualCal --plot --before recal_data.table --after recal_data.table.post recal.csv sentieon plot bqsr -o recal_plots.pdf recal.csv
Variant calling - UnifiedGenotyper	java -jar GenomeAnalysisTK.jar -T UnifiedGenotyper [-L $bed] -nt 32 -R $fasta -I recaled.bam -o UG.vcf	sentieon driver -r $fasta [--interval $bed] -t 32 -i realigned.bam -q recal_data.table --algo Genotyper output-ug.vcf
Variant calling - HaplotypeCaller	java -jar GenomeAnalysisTK.jar -T HaplotypeCaller [-L $bed] -nct 32 -R $fasta -I recaled.bam -o HC.vcf	senteion driver -r $fasta [--interval $bed] -t 32 -i realigned.bam -q recal_data.table --algo Haplotyper output-hc.vcf
Comparison	java -jar RTG.jar vcfeval --baseline=HC.vcf --calls=output-hc.vcf --output=$out_dir --template=$sdf_file [--bed-region $bed] --sample=$sample	

Commands used for testing the DNAseq pipeline on single samples are listed. Arguments supplying the bed file were used with whole-exome sequenced samples. The supplied bed file contained Illumina TruSeq capture targets.

DNAseq Pipeline – Joint Calling

Analysis Step	GATK Best Practices	Sentieon DNAseq
Variant calling – Joint Genotyping	java -jar GenomeAnalysisTK.jar -T GenotypeGVCFs -L chr1 -nt 32 -R $fasta $gvcf_list -o GATK_jointcalling_chr1.vcf.gz -maxAltAlleles 100	sentieon driver -r $fasta --interval chr1 -t 32 --algo GVCFtyper $gvcf_list sentieon_jointcalling_chr1.vcf.gz
Comparison for joint calling per sample	java -jar RTG.jar vcfeval --baseline=GATK_jointcalling_chr1.vcf.gz --calls=sentieon_jointcalling_chr1.vcf.gz --output=$out_dir --template=$sdf_file [--bed-region $bed] --sample=$sample	

Commands used for testing the joint calling with the DNAseq pipeline. For the consistency check, the comparison command was run for all 75 genotyped samples.

Tumor-Normal Pipeline

Analysis Step	MuTect and MuTect2	Sentieon TNseq and TNhaplotyper
Read alignment	bwa mem -R '@RG\tID:$group\tSM:$sample\tPL:$pl' -t 32 -M -K 10000000 $fasta '<bzip2 -dc $fastq_1.bz2' '<bzip2 -dc $fastq_2.bz2' | samtools sort -@ 32 -o sorted.bam -T sorted samtools index sorted.bam	bwa mem -R '@RG\tID:$group\tSM:$sample\tPL:$pl' -t 32 -M -K 10000000 $fasta '<bzip2 -dc $fastq_1.bz2' '<bzip2 -dc $fastq_2.bz2' | sentieon util sort -i - -r $fasta -t 32 -o sorted.bam --sam2bam
Sample metrics collection	java -jar ${picard}/CollectAlignmentSummaryMetrics.jar VALIDATION_STRINGENCY=SILENT INPUT=sorted.bam OUTPUT=aln_metrics.txt REFERENCE_SEQUENCE=$fasta ADAPTER_SEQUENCE=null java -jar ${picard}/CollectGcBiasMetrics.jar VALIDATION_STRINGENCY=SILENT INPUT=sorted.bam OUTPUT=gc_metrics.txt REFERENCE_SEQUENCE=$fasta SUMMARY_OUTPUT=gc_summary.txt CHART_OUTPUT=gc-metrics-report.pdf ASSUME_SORTED=true java -jar ${picard}/MeanQualityByCycle.jar VALIDATION_STRINGENCY=SILENT INPUT=sorted.bam OUTPUT=mq_metrics.txt REFERENCE_SEQUENCE=$fasta CHART_OUTPUT=mq-metrics-report.pdf PF_READS_ONLY=true java -jar ${picard}/QualityScoreDistribution.jar VALIDATION_STRINGENCY=SILENT INPUT=sorted.bam OUTPUT=qd_metrics.txt REFERENCE_SEQUENCE=$fasta CHART_OUTPUT=qd-metrics-report.pdf PF_READS_ONLY=true java -jar ${picard}/CollectInsertSizeMetrics.jar VALIDATION_STRINGENCY=SILENT INPUT=sorted.bam OUTPUT=is_metrics.txt REFERENCE_SEQUENCE=$fasta HISTOGRAM_FILE=metrics-report.pdf	sentieon driver -t 32 -r $fasta -i sorted.bam --algo MeanQualityByCycle mq_metrics.txt --algo InsertSizeMetricAlgo is_metrics.txt --algo QualDistribution qd_metrics.txt --algo GCBias --summary gc_summary.txt gc_metrics.txt --algo AlignmentStat --adapter_seq '' aln_metrics.txt sentieon plot metrics -o metrics-report.pdf gc=gc_metrics.txt qd=qd_metrics.txt mq=mq_metrics.txt isize=is_metrics.txt
Duplicate read removal	java -jar ${picard}/MarkDuplicates.jar VALIDATION_STRINGENCY=SILENT INPUT=sorted.bam OUTPUT=deduped.bam METRICS_FILE=dedup_metrics.txt REMOVE_DUPLICATES=true MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=65536 java -jar ${picard}/BuildBamIndex.jar VALIDATION_STRINGENCY=SILENT INPUT=deduped.bam	sentieon driver -t 32 -r $fasta -i sorted.bam --algo LocusCollector --fun score_info deduped.bam.score.gz sentieon driver -t 32 -r $fasta -i sorted.bam --algo Dedup --score_info deduped.bam.score.gz --rmdup --metrics dedup_metrics.txt deduped.bam
Indel realignment	java -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -R $fasta -I deduped.bam -nt 32 -known $Mills -known $onekg_indels -o realigner.intervals java -jar GenomeAnalysisTK.jar -T IndelRealigner -R $fasta -I deduped.bam -known $Mills -known $onekg_indels --targetIntervals realigner.intervals -o realigned.bam	sentieon driver -t 32 -r $fasta -i deduped.bam --algo Realigner -k $Mills -k $onekg_indels realigned.bam
Base quality score recalibration	java -jar GenomeAnalysisTK.jar -T BaseRecalibrator -R $fasta -I realigned.bam -nct 32 --knownSites $dbsnp --knownSites $Mills --knownSites $onekg_indels -o recal_data.table java -jar GenomeAnalysisTK.jar -T PrintReads -R $fasta -I realigned.bam --BQSR recal_data.table -o recaled.bam java -jar GenomeAnalysisTK.jar -T BaseRecalibrator -R $fasta -I realigned.bam --BQSR recal_data.table -nct 32 --knownSites $dbsnp --knownSites $Mills --knownSites $onekg_indels -o recal_data.table.post java -jar GenomeAnalysisTK.jar -T AnalyzeCovariates -R $fasta -csv recal.csv -before recal_data.table -after recal_data.table.post -plots recal_plots.pdf	sentieon driver -t 32 -r $fasta -i realigned.bam --algo QualCal -k $dbsnp -k $Mills -k $onekg_indels recal_data.table sentieon driver -t 32 -r $fasta -i realigned.bam -q recal_data.table --algo QualCal -k $dbsnp -k $Mills -k $onekg_indels recal_data.table.post sentieon driver --passthru --algo QualCal -k $dbsnp -k $Mills -k $onekg_indels --before recal_data.table --after recal_data.table.post --plot recal.csv sentieon plot bqsr -o recal_plots.pdf recal.csv
Indel joint realignment	java -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -R $fasta -I tumor_recaled.bam -I normal_recaled.bam -nt 32 -known $Mills -known $onekg_indels -o interval.list java -jar GenomeAnalysisTK.jar -T IndelRealigner -R $fasta -I tumor_recaled.bam -I normal_recaled.bam -known $Mills -known $onekg_indels --targetIntervals interval.list -nWayOut _corealigned.bam	sentieon driver -t 32 -r $fasta -i tumor_realigned.bam -i normal_realigned.bam -q tumor_recal_data.table -q normal_recal_data.table --algo Realigner -k $Mills -k $onekg_indels tn_corealigned.bam
Variant calling - MuTect	java -jar mutect-1.1.5.jar -T MuTect -R $fasta --dbsnp $dbsnp -o output-call.stats -vcf mutect.vcf.gz -I:tumor tumor_recaled_corealigned.bam -I:normal normal_recaled_corealigned.bam	sentieon driver -t 32 -r $fasta -i tn_corealigned.bam --algo TNsnv --tumor_sample $tumor_sample_name --normal_sample $normal_sample_name --dbsnp $dbsnp --call_stats_out output-call.stats output-tnsnv.vcf.gz
Variant calling - MuTect2	java -jar GenomeAnalysisTK.jar -T MuTect2 -R $fasta --dbsnp $dbsnp -o mutect2.vcf.gz -I:tumor tumor_recaled_corealigned.bam -I:normal normal_recaled_corealigned.bam -nct 32	sentieon driver -t 32 -r $fasta -i tn_corealigned.bam --algo TNhaplotyper --tumor_sample tumor_sample_name --normal_sample normal_sample_name --dbsnp $dbsnp output-tnhaplotyper.vcf.gz
Comparison	java -jar RTG.jar vcfeval --baseline=mutect2.vcf.gz --calls=tnhaplotyper.vcf.gz --output=$out_dir --template=$sdf_file [--bed-region $bed]	

Commands used in the tumor-normal benchmarking pipelines. The --bed-region argument was used for analysis of the whole-exome sequenced sample. The supplied bed file contained Illumina TruSeq capture targets.