WGS Pipeline v1.3.8 Documentation

Overview

WGS pipeline is a scalable, portable, and reproducible bioinformatics pipeline to process whole genome sequencing and exome/targeted panel sequencing data. The pipeline currently only has added support for human sequencing data but can certainly be extended to other model systems. Pipeline's summary is explained on Figure 1:

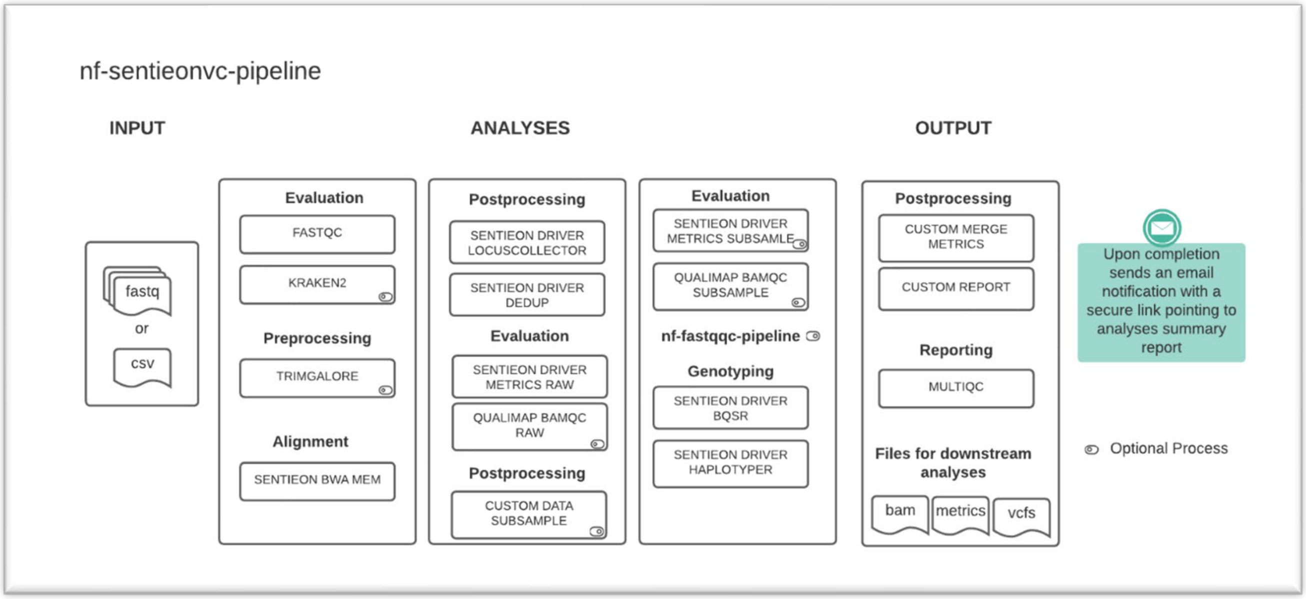


Figure 1. Components of the WGS pipeline

The pipeline is built using Nextflow[1], a workflow tool to run tasks across multiple compute infrastructures in a very portable manner. At BioSkryb, we currently deploy these pipelines at AWS but it is flexible to run locally or other cloud providers. All the processes in the pipeline run in docker containers which makes it easy to reproduce the environment and highly reproducible results. The pipeline takes raw sequencing data in form of fastq files and performs quality control assessments to evaluate the quality of the library build. The pipeline then aligns, removes duplicate reads, base calibrates the reads, all before haplotype calling. There are additional modules users can choose to run as part of their analyses. These modules include joint genotyping to improve upon the variant calling, variant annotation based on public databases such as gnomAD[2], ClinVar[3], and its putative impact on protein, CNV analyses, and Estimating the Allelic Dropouts.

Pipeline Workflow

1. **Trimming**
   Trimming reads by adapter, or custom sequences, improves evaluation and alignment. It is performed using FASTP[4] tool. Input for this step are raw FASTQ files, and output are trimmed FASTQ files.

2. **Alignment**
   In this step reads are aligned to reference genome and additionally processed, to achieve better analysis precision. The input are trimmed FASTQ files, and the main outputs are alignment deduplicated BAM file and its index, and recalibration table. This step consists of several substeps:

   BWA MEM - The BWA[5] MEM is a short read aligner that takes singe/paired-end sequencing data and maps to the reference genome. The input to the tool is a FASTQ file, its output is a BAM[6] file.

   LOCUSCOLLECTOR - The LocusCollector (provided by Sentieon) algorithm collects read information that will be used for removing duplicate reads. The input to the LocusCollector algorithm is a BAM file; its output is the score file indicating which reads are likely duplicates.

   DEDUP - The algorithm (provided by Sentieon)performs the marking/removing of duplicate reads. The input to the Dedup algorithm is a BAM file; its output is the BAM file after removing duplicate reads.

   BQSR - The BQSR (provided by Sentieon) algorithm performs base recalibration. The input is deduplicated BAM file, and the output is recalibration table.

   DATA COMPRESSION – Both primary (FASTQ.gz) and secondary (BAM) files are further compressed resulting in ~60% of data compression. Compression is powered by PetaGene.

3. **Evaluation**
   In this step data evaluation is performed, from two pints of view: from point of reads and from points of aligned reads. To perform evaluation on the optimized way, the first step is to choose reduced subset of reads and aligned reads, in a way which will maintain statistical significance in results. Input of this step are trimmed FASTQ files and deduplicated (or deduplicated and compressed) BAM files, and the outputs are various metrics.

   SEQTK - The seqtk tool (https://github.com/lh3/seqtk) performs read subsampling, provided in FASTQ files, which is represent as output in subsampled FASTQ files.

KRAKEN2[7] - The tool provides a taxonomic classification system using exact k-mer matches to find the least common ancestor to which the sequence matches. The k-mer assignment then informs the classification algorithm. The input to the tools is a FASTQ file, its output is the report (.txt) and unclassified reads (_1FASTQ and _2FASTQ). Since this substep is time and memory consuming, it is set as optional one.

DATA SUBSAMPLE - The tool performs subsampling using samtools based on the user-specified total reads. The input is deduplicated (or deduplicated and compressed) BAM file and the output is subsampled BAM file.

METRICS - The tool evaluates several metrics for the provided alignment data provided as BAM file – (i) WGS Metrics - The algorithm collects metrics related to the coverage and performance of whole-genome sequencing (WGS) experiments. (ii) GC Bias - The algorithm calculates the GC bias in the reference and the sample. (iii) Alignment Stats - The algorithm calculates statistics about the alignment of the reads. (iv) Coverage Metrics - The algorithm calculates the depth coverage of the BAM file. The coverage is aggregated by the interval. The input to the metrics is a deduplication BAM file; its outputs are files containing the metrics data.

QUALIMAP[8] BAMQC - The tool evaluates the quality of the provided alignment data. The input is the deduplication BAM file; its output is the file containing the metrics data.

4. **Variant Calling**

This step performs calling of germline, joint and copy number variants. The inputs are deduplicated (or deduplicated and compressed) BAM files, various variant databases, machine learning model and targeted .bed file. Outputs are variants described in .vcf files and copy number information in .tsv file and various plots.

HAPLOTYPER - Variant calling is performed by using Haplotyper tool, which receives deduplicated (or deduplicated and compressed) BAM file, recalibration table and various variant databases and outputs variants in .vcf file.

JOINT VARIANT CALLING - In order to perform joint variant calling, set of tools is used: GVCFTyper, VARCal, and ApplyVC. The input is .vcf file, and the output is multisample .vcf. file.

DNAScope - Germline calling is performed by using DNAScope tool, which receives deduplicated (or deduplicated and compressed) BAM file and ML model and outputs variants in .vcf file. This is the optional step.

GINKGO[9] - Copy number call tool Ginkgo, is taking the deduplicated (or deduplicated and compressed) BAM file, converts it to .bed file and outputs various plots and .tsv table with copy number information.

5. **Annotation**

   This step performs annotating variants from the previous step, using snpEff and SnpSift tool[10]. The output is annotated .vcf file.

6. **Reporting**

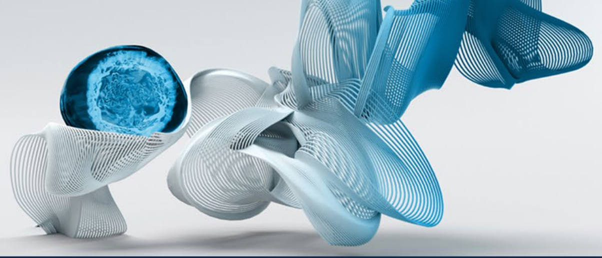
   In this step, all metrics are collected, the custom report about which samples passed evaluation is created, and finally outputs of those substeps are forwarded to MultiQC, to create the final .html report.

   MultiQC - The tool aggregates results of bioinformatics analyses into a single HTML report. The input is the output files from FASTP, KRAKEN2, METRICS, QUALIMAP BAMQC, and FASTQC; its output is HTML report.

Pipeline Parameters

Table 1. Summary of pipeline parameters

Module	Parameter Name	Options	Description
General Pipeline Parameters			
	Mode	wgs (default) exome	Define whether to run the pipeline in WGS mode or Exome/targeted mode
	Exome/Targeted Panel	xGen Exome Hyb Panel v2 (default) TruSight One TWIST	Exome/Targeted panel to use for analyses
	Instrument	NovaSeq NextSeq (default) MiSeq MiniSeq ISeq Other	Instrument used to perform sequencing
	Genome	GRCh38	Reference genome to use for alignment
	Read Length	50 75 (default) 100 150	Cycle used for sequencing
	Ploidy	1 2 (default)	Ploidy number for the samples being processed. Whether the samples are diploid or haploid
Module Parameters			
QC	Toggle	True (default) False	Performs QC using Kraken2 and FastQC to evaluate metagenomic



			contamination and qc checks on your raw sequence data
QC	Subsample Reads	1000000 (default)	Number of paired reads to sample for QC evaluation
Qualimap	Toggle	True False (default)	Qualimap module evaluates the quality of the alignment data
Allelic Dropout	Toggle	True False (default)	Evaluate the allelic dropout rates and uniformity of the coverage. NOTE: Only for HG001/NA12878 samples.
Evaluate Variant Calling	Toggle	True False (default)	Perform benchmarking on variant calling based on ground truth variants. NOTE: Only for HG001/NA12878 samples.
Annotating Variants	Toggle	True False (default)	Perform variant annotation.
Compress BAM/FASTQs	Toggle	True False (default)	Perform compression of the FASTQs and BAMs
DNAScope	Toggle	True False (default)	DNAScope module calls SNPs and small indels. By default, DNAseq is used to call variants but user can also specify DNAScope to call variants.
Joint Genotyping	Toggle	True False (default)	Joint Genotyping module improves variant detection by doing joint analysis and leveraging information from other samples in the analyses
Ginkgo	Toggle	True False (default)	Ginkgo module evaluates for the Copy Number Variation.
Ginkgo	Bin Size (bp)	1000000 (default)	Genomic bin size in bp to use for CNV

Pipeline Output

Table 2. Summary of the key outputs from the pipeline

Analyses Step	Raw Output Name	Data Type	Description

Trimming	**WGS_WF_FastpFull_WF_FASTP**/<biosample name>_1.trim.fastq.gz;
	WGS_WF_FastpFull_WF_FASTP/<biosample name>_2.trim.fastq.gz;
		FASTQ	Trimmed fastq files for each biosample

Alignment	**WGS_WF_SENTIEON_DRIVER_DEDUP_WF_SENTIEON_DRIVER_DEDUP**/<biosample name>.bam	BAM	BAM files with lower quality duplicate reads removed for each biosample. Includes index file for each bam.
Alignment	**WGS_WF_SENTIEON_DRIVER_DEDUP_WF_SENTIEON_DRIVER_DEDUP**/<biosample name>.bam.bai	BAI	BAM index files, necessary for downstream tools for genome viewing
Evaluation	**WGS_WF_SENTIEON_DRIVER_METRICS_WF_SENTIEON_DRIVER_METRICS**/<biosample name>.insertsizemetricalgo.sentieonmetrics.txt	txt	Various metrics about distribution of insert sizes
Evaluation	**WGS_WF_SENTIEON_DRIVER_METRICS_WF_SENTIEON_DRIVER_METRICS**/<biosample name>.gcbias_summary.sentieonmetrics.txt	txt	Coverage Statistics specifically across the genome, but summarized across regions of known GC content for biosample
Evaluation	**WGS_WF_SENTIEON_DRIVER_METRICS_WF_SENTIEON_DRIVER_METRICS**/<biosample name>	txt	Reads mapped and other normalized metrics to each interval (or chromosome) of reference genome for biosample

	.cov_sentieonmetrics.sample_interval_summary.cov_sentieonmetrics.sample_interval_summary		
Evaluation	**WGS_WF_SENTIEON_DRIVER_METRICS_WF_SENTIEON_DRIVER_METRICS**/<biosample name>.wgsmetricsalgo.sentieon metrics.txt	txt	Coverage Statistics across the genome along with counts of positions at each depth for biosample
Variant Calling	**WGS_WF_SENTIEON_DRIVER_HAPLOTYPER_WF_SENTIEON_DRIVER_HAPLOTYPER**/<biosample name>.bqsr.g.vcf.gz	VCF	VCF files for each biosample. Includes index file for each vcf.
Annotation	**WGS_WF_SNPSIFT_ANNOTATION_WF_SNPSIFT_dbNSFP**/MultiSample_predictions.vcf	VCF	Annotated for merged multi-sample VCF for all samples in a project
Reporting	**WGS_WF_MultiQC_WF_MultiQC**/multiqc_report.html	html	MultiQC report providing summary of the key metrics in the analyses

References

1. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).

2. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

3. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).

4. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

5. Li, H. *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. http://arxiv.org/abs/1303.3997 (2013) doi:10.48550/arXiv.1303.3997.

6. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).

7. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).

8. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinforma. Oxf. Engl.* **32**, 292–294 (2016).

9. Garvin, T. *et al.* Interactive analysis and assessment of single-cell copy-number variations. *Nat. Methods* **12**, 1058–1060 (2015).

10. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).