PreSeq Pipeline v1.2.0 Documentation

Overview

Making libraries for single-cell sequencing can be complex and expensive. We recommend users to ensure that the single-cell library is uniformly amplified (with low allelic dropouts) by first sequencing using "low-pass" (typically 2M reads per sample). This data is used to estimate the genome coverage were the single-cell libraries be used for high-depth sequencing. Users can then only use the passing libraries for high-depth sequencing. Figure 1 provides an overview of the pipeline processes.

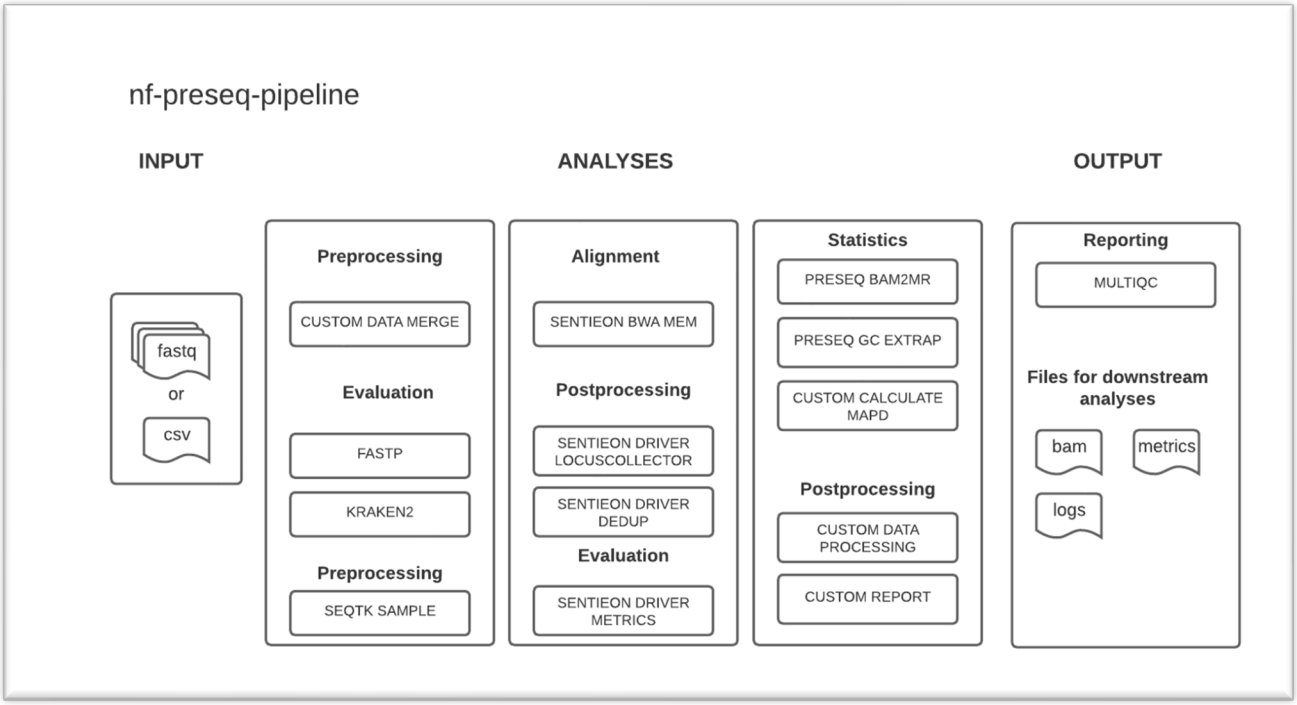

Figure 1. PreSeq pipeline components

Pipeline Workflow

1. **Subsampling**
   Pipeline performs subsampling of the reads using SEQTK (https://github.com/lh3/seqtk) tool. Input for the step are raw FASTQ and output are subsampled FASTQ files
2. **Trimming**

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

3. **Alignment**

In this step reads are aligned to reference genome and additionally processed, in order 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[2] MEM - It 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[3] file.

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

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

4. **Evaluation**

In this step data evaluation is performed, from two pints of view: from point of total 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.

PRESEQ BAM2MR - The tool converts the BAM file to mapped reads to prepare the input for GC EXTRAP. The input is the deduplicated BAM file from DEDUP; its output is the mapped reads.

PRESEQ GC EXTRAP - The tool infers the properties of the behavior under deeper sequencing based upon a small initial sequencing experiment. The input to the tool is the mapped read from BAM2MR; its output is the Preseq scores.

METRICS - The tool evaluates several metrics for the provided alignment data provided as BAM file – (i) GC Bias - The algorithm calculates the GC bias in the reference and the sample. (ii) Alignment Stats - The algorithm calculates statistics about the alignment of the reads.

(iii) Coverage Metrics - The algorithm calculates the depth coverage of the BAM file. The coverage is aggregated by the interval. (iv) Insert Size Metrics – The algorithm calculates the statistical distribution of insert sizes. The input to the metrics is a deduplication BAM file; its outputs are files containing the metrics data.

QUALIMAP[4] 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.

5. **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[5] - The tool aggregates results of bioinformatics analyses into a single HTML report. The input is the output files from FASTP, KRAKEN2[6], METRICS, QUALIMAP BAMQC, and FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/); its output is HTML report.

Pipeline Parameters

Table 1. Summary of pipeline parameters

Module	Parameter Name	Options	Description
General Pipeline Parameters			
	Read Length	50 75 (default) 100 150	Cycle used for sequencing
	Read Sampling	1000 1000000 (default) 2000000	Number of paired reads to sample. NOTE: 1000000 paired reads is equivalent to 2000000 individual reads
	Instrument	NovaSeq NextSeq MiSeq MiniSeq (default)	Instrument used to perform sequencing

		ISeq Other	
	Genome	GRCh38 (default)	Reference genome to use for alignment
Module Parameters			
Kraken2	Toggle	True (default) False	Kraken2 module evaluates the metagenomic contamination.
FastQC	Toggle	True (default) False	FastQC performs qc checks on your raw sequence data.
Qualimap	Toggle	True False (default)	Qualimap module evaluates the quality of the alignment data

Pipeline Output

Table 2. Summary of the key outputs from the pipeline

Analyses Step	Raw Output Name	Data Type	Description
Alignment	**PRESEQ_WF_SENTIEON_DRIVER_DEDUP_WF_SENTIEON_DRIVER_DEDUP/**<biosample name>_sorted.bam	BAM	BAM files with lower quality duplicate reads removed for each biosample. Includes index file for each BAM.
Alignment	**PRESEQ_WF_SENTIEON_DRIVER_DEDUP_WF_SENTIEON_DRIVER_DEDUP**/<biosample name>_sorted.bam.bai	BAI	BAM files index.
Evaluation	**PRESEQ_WF_SENTIEON_DRIVER_METRICS_WF_SENTIEON_DRIVER_METRICS/**<biosample name>.dedup.alignmentstat_sentieonmetrics.txt	txt	Alignment Statistics for biosample
Evaluation	**PRESEQ_WF_SENTIEON_DRIVER_METRICS_WF_SENTIEON_DRI**	txt	Coverage Statistics across the genome along with counts of

	VER_METRICS/<biosample name>. .wgsmetricsalgo.sentieonmetrics.txt		positions at each depth for biosample
Evaluation	**PRESEQ_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	**PRESEQ_WF_SENTIEON_DRIVER_METRICS_WF_SENTIEON_DRIVER_METRICS/**<biosample name>.cov_sentieonmetrics.sample_interval_summary	txt	Reads mapped and other normalized metrics to each interval (or chromosome) of reference genome for biosample
Reporting	**PRESEQ_WF_MULTIQC_WF_MULTIQC/** multiqc_report.html	html	MultiQC report providing summary of the key metrics in the analyses

References

1. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
2. 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.
3. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
4. 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).
5. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

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