Plasmid Verification Analysis Report

This report is not a diagnostic / clinical report and is intended for Research Use Only!

Eurofins Project ID: **NG-12345**
Date of Processing: **10 May, 2022**
Pipeline: **Plasmid Verification Pipeline**
Version: **v1.0**

Samples

- Table 1: The samples used in this pipeline:

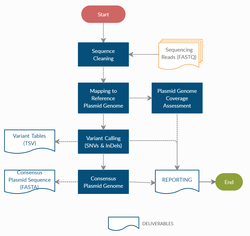
samples
pBR322_10ng_1

Reference Sequence Used

- Table 2: The References used in the analysis:

Reference	Size	Description
pBR322	4361	plasmid cloning vector

Workflow



Quality Control of Raw Sequencing Data

Raw sequencing data are preprocessed to generate clean data for downstream analysis. In this step, quality of raw sequencing is checked and filtered to retain only high quality bases by performing adapter trimming, quality filtering and per-read quality pruning.

Quality is interpreted as the probability of an incorrect base call or, equivalently, the base call accuracy. The quality score is logarithmically based, so a quality score of 10 reflects a base call accuracy of 90%, but a quality score of 20 reflects a base call accuracy of 99% and a quality score of 30 reflects a base call accuracy of 99.9%. These probability values are the results from the base calling algorithm and depend on how much signal was captured for the base incorporation.

Sequencing reads representing reads with quality score at least Q30 is above 90% is of very good quality. For a reasonably good sample source material, according to Illumina specifications, one could expect >75% reads with at least Q30 Phred quality.

Raw sequencing data is processed using fastp[1] software to remove poor quality bases (below Phred Quality 20) using the sliding window approach where in if the average quality of the bases drops below Q20, those bases are removed from the reads; After quality trimming, program checks for presence of any adapters in the reads and removes from the reads. Further, shorter reads which are <30bp length are also removed to retain only high quality sequencing reads for each sample in the analysis. In case of paired-end reads, both the sequencing reads which pass the QC criteria are considered for downstream analysis.

After QC processing, QC metrics such as Q30 reads and GC content can be used to assess the sequencing and sample quality across the samples.

Read Statistics

- Table 3: Sequence Quality Metrics overview. For each sample, the following QC metrics are provided:
  - Sample Name: name of the sample.
  - Total Raw Reads: the total number of raw sequencing reads generated for the sample.
  - Total HQ Reads: the total number of high quality reads after sequence cleaning and filtering.
  - HQ Reads %: High Quality Reads percentage
  - HQ Bases (Q30): Percentage of high quality bases having at least phred quality 30.
  - GC Content: GC content in percentile of high quality sequencing reads.
  - Mean Read Length (bp): Average read length in bp of high quality sequencing reads.

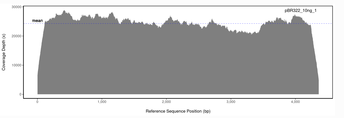
Mapping to Reference

Resulting high quality reads were mapped against the reference sequence to generate alignments using BWA (Burrows-Wheeler Alignment). BWA[6] is a read alignment package that is based on backward search with Burrows-Wheeler Transform (BWT), to efficiently align short sequencing reads against a large reference sequence such as the human genome, allowing mismatches and gaps. Only mapped reads were considered for downstream analysis.

- Table 4: The Mapping statistics are given as follows:
  - Sample Name
  - Total HQ Reads: Total number of high quality reads.
  - Mapped Reads: Number of Reads which map to the reference.
  - Unique Reads: Number of Reads which map uniquely to the reference.

Coverage Analysis

Coverage analysis is done using BEDTools[8] genomeCoverageBed which gives per base report of genome coverage. The per base coverage is then plotted in R[9]



Variant Table

The SNP and InDel calling is done using VarScan2[5]. Allele frequency cut-off used for variant calling is 1%.

- Table 5: Variant Table. For each sample, the following variant summary is provided:
  - CHROM: Chromosome in which the variant is observed.
  - POS: Position at which the variant is observed
  - REF: Reference base
  - ALT: Alt base
  - Allele Freq: Variant allele frequency in percentage
  - Alt Depth: Depth of variant-supporting bases
  - Total Depth: Depth of variant-supporting bases and reference-supporting bases

Consensus sequence

The consensus sequence is generated using BCFtools consensus[4]. The output is generated in FASTA format.

Relevant Programs

- Table 6: The programs/softwares used in this pipeline.

Tool	Version	Description
BCFtools[4]	1.10.2	BCFtools is a set of utilities that manipulate variant calls in the Variant Call Format (VCF) and its binary counterpart BCF
BWA[6]	0.7.17	BWA is a software package for mapping low-divergent sequences against a large reference genome
fastp[1]	0.20.0	Fastp is a tool designed to provide fast all-in-one preprocessing for FastQ files.
RTG_Tools[3]	3.12.1	Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines
sambamba[11]	0.6.8	Sambamba is a high performance modern robust and fast tool (and library), for working with SAM and BAM files.
samtools[7]	1.10.0	SAMtools provide various utilities for manipulating alignments in the SAM format.
Sentieon[10]	202010.02	Sentieon® provides complete solutions for secondary DNA/RNA analysis for a variety of sequencing platforms, including short and long reads.
snpEff[2]	4.3	SnpEff is a genetic variant annotation and effect prediction toolbox.
VarScan[5]	2.2.4	VarScan: variant detection in massively parallel sequencing of individual and pooled samples

Output Files and Descriptions

- Table 7: The deliverables are given as follows,

Tool	Version	Description
BCFtools[4]	1.10.2	BCFtools is a set of utilities that manipulate variant calls in the Variant Call Format (VCF) and its binary counterpart BCF
BWA[6]	0.7.17	BWA is a software package for mapping low-divergent sequences against a large reference genome
fastp[1]	0.20.0	Fastp is a tool designed to provide fast all-in-one preprocessing for FastQ files.
RTG_Tools[3]	3.12.1	Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines
sambamba[11]	0.6.8	Sambamba is a high performance modern robust and fast tool (and library), for working with SAM and BAM files.
samtools[7]	1.10.0	SAMtools provide various utilities for manipulating alignments in the SAM format.
Sentieon[10]	202010.02	Sentieon® provides complete solutions for secondary DNA/RNA analysis for a variety of sequencing platforms, including short and long reads.
snpEff[2]	4.3	SnpEff is a genetic variant annotation and effect prediction toolbox.
VarScan[5]	2.2.4	VarScan: variant detection in massively parallel sequencing of individual and pooled samples

References

[1] Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, 17 (September 2018), i884–i890. DOI: **https://doi.org/10.1093/bioinformatics/bty560**
[2] Pablo Cingolani. 2012. "snpEff: Variant effect prediction".
[3] John G Cleary, Ross Braithwaite, Kurt Gaastra, Brian S Hilbush, Stuart Inglis, Sean A Irvine, Alan Jackson, Richard Littin, Mehul Rathod, David Ware, and others. 2015. Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. *BioRxiv* (2015), 023754.
[4] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, and Heng Li. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* 10, 2 (February 2021). DOI: **https://doi.org/10.1093/gigascience/giab008**
[5] Daniel C Koboldt, Ken Chen, Todd Wylie, David E Larson, Michael D McLellan, Elaine R Mardis, George M Weinstock, Richard K Wilson, and Li Ding. 2009. VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 25, 17 (2009), 2283–2285.
[6] Heng Li and Richard Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* 25, 14 (July 2009), 1754–1760. DOI: **https://doi.org/10.1093/bioinformatics/btp324**
[7] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 16 (2009), 2078–2079. DOI: **https://doi.org/10.1093/bioinformatics/btp352**
[8] Aaron R. Quinlan and Ira M. Hall. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 6 (March 2010), 841–842. DOI: **https://doi.org/10.1093/bioinformatics/btq033**
[9] R Core Team. 2013. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from **http://www.R-project.org/**
[10] Sentieon. 2021. *Sentieon provides complete solutions for secondary dna/rna analysis for a variety of sequencing platforms, including short and long reads*. Sentieon. Retrieved from **https://www.sentieon.com/**
[11] Artem Tarasov, Albert J. Vilella, Edwin Cuppen, Isaac J. Nijman, and Pjotr Prins. 2015. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* (February 2015). DOI: **https://doi.org/10.1093/bioinformatics/btv098**