Output description for SRVCNM169 analysis.

Pipeline overview

- FastQC v0.11.8 - read quality control.
- Trimmomatic v.0.33 - adapter and low quality trimming.
- Unicycler v.0.4.6 - prokaryote assembler.
- Quast v4.1 - assemblies quality control
- Kmerfinder v.3.1 - species and contamination determination.
- Snippy v.4.4.0 - haploid variant calling and core genome alignment
- Picard WGSMetrics v1.140 - bam statistics.
- IQTree v.2.1.4 - maximum likelihood phylogeny.

Note:

Depending on the analysis, we could have some ANALYSIS_IDs. This ANALYSIS_IDs are going to be composed of the date of the analysis, and an analysis identification. You can find a README in de ANALYSIS folder with a brief description of the different analysis.

Preprocessing

FastQC

FastQC gives general quality metrics about your reads. It provides information about the quality score distribution across your reads, the per base sequence content (%T/A/G/C). You get information about adapter contamination and other overrepresented sequences.

For further reading and documentation see the FastQC help.

Output directory: `01-fastqc`

- `{sample_id}/{sample_id}_R[12]_fastqc.html`
  - html report. This file can be opened in your favourite web browser (Firefox/chrome preferable) and it contains the different graphs that fastqc calculates for QC.
- `{sample_id}/{sample_id}_R[12]_fastqc`
  - older with fastqc output in plain text.
- `{sample_id}/{sample_id}_R[12]_fastqc.zip`
  - zip file containing the FastQC report, tab-delimited data file and plot images

Trimming

Trimmomatic (1) is used for removal of adapter contamination and trimming of low quality regions. Parameters included for trimming are:

- Nucleotides with phred quality < 10 in 3'end.
- Mean phred quality < 20 in a 4 nucleotide window.
- Read lenght < 50

Results directory: `02-preprocessing`

- Files:
  - `{sample_id}/{sample_id}_R[12]_filtered.fastq.gz`: contains high quality reads with both forward and reverse tags surviving.
  - `{sample_id}/{sample_id}_R[12]_unpaired.fastq.gz`: contains high quality reads with only forward or reverse tags surviving.

  Note:To see how your reads look after trimming, look at the FastQC reports in the 03-preprocQC directory

Table of Contents

Kmerfinder

Kmerfinder (2) is a software used for species indentification and the determination of possible contamination in the sample. We use this software using the bacterial database provided by the developers, and with the "winner takes it all" algorithm. You can check here for a description of the columns in the output.

Output directory: `04-kmerfinder/{sample_id}/`

- `data.json`
  - results in json format.
- `results.spa`
  - results in spa format.
- `results.txt`
  - results in txt format. **This is the format you have to use if you are going to open it with excel.**

NOTE: You can also find in `99-stats` a summary of all samples results (`kmerfinder.csv`).

Assembly

Unicycler

Unicycler (3) is an assembly pipeline for bacterial genomes. It can assemble Illumina-only read sets where it functions as a SPAdes-optimiser.

Output directory: `08-unicycler/{sample_id}`

- `{sample_id}.fasta`: fasta file containing the assembled reads in form of contigs and scaffolds. This is the file we use for annotation and upstream analysis.
- `{sample_id}.gfa`: Graph files for the different assembly optimizations. This files can be used by advanced users with software like Bandage

Annotation and quality control

Prokka

Prokka (4) was used for assembled genome annotation.

Output directory: `00-prokka/{sample_id}`

- `{sample_id}.gff`: This is the master annotation in GFF3 format, containing both sequences and annotations. It can be viewed directly in Artemis or IGV.
- `{sample_id}.gbk`: This is a standard Genbank file derived from the master .gff. If the input to prokka was a multi-FASTA, then this will be a multi-Genbank, with one record for each sequence.
- `{sample_id}.fna`: Nucleotide FASTA file of the input contig sequences.
- `{sample_id}.faa`: Protein FASTA file of the translated CDS sequences.
- `{sample_id}.ffn`: Nucleotide FASTA file of all the prediction transcripts (CDS, rRNA, tRNA, tmRNA, misc_RNA)
- `{sample_id}.sqn`: An ASN1 format "Sequin" file for submission to Genbank. It needs to be edited to set the correct taxonomy, authors, related publication etc.
- `{sample_id}.fsa`: Nucleotide FASTA file of the input contig sequences, used by "tbl2asn" to create the .sqn file. It is mostly the same as the .fna file, but with extra Sequin tags in the sequence description lines.
- `{sample_id}.tbl`: Feature Table file, used by "tbl2asn" to create the .sqn file.
- `{sample_id}.err`: Unacceptable annotations - the NCBI discrepancy report.
- `{sample_id}.log`: Contains all the output that Prokka produced during its run. This is a record of what settings you used, even if the --quiet option was enabled.
- `{sample_id}.txt`: Statistics relating to the annotated features found.

- `{sample_id}.tsv`: Tab-separated file of all features: locus_tag,ftype,len_bp,gene,EC_number,COG,product

QUAST

QUAST (5) evaluates genome assemblies. We compared the reference genome with the contigs and scaffold assemblies. The html results can be opened with any browser (we recommend using Google Chrome).

Output directory: `00-assemblies/quast_results`

- `quast_results/date/report.html`
    - Compressed format of the indexed variants file.
    - The meaning of the different metrics:
    - Contigs (≥ x bp): is total number of contigs of length ≥ x bp.
    - Total length (≥ x bp): is the total number of bases in contigs of length ≥ x bp.
    - Contigs: is the total number of contigs in the assembly.
    - Largest contig: is the length of the longest contig in the assembly.
    - Total length: is the total number of bases in the assembly.
    - Reference length: is the total number of bases in the reference genome.
    - GC (%): is the total number of G and C nucleotides in the assembly, divided by the total length of the assembly.
    - Reference GC (%): is the percentage of G and C nucleotides in the reference genome.
    - N50: is the length for which the collection of all contigs of that length or longer covers at least half an assembly.
    - NG50: is the length for which the collection of all contigs of that length or longer covers at least half the reference genome. This metric is computed only if the reference genome is provided.
    - N75 and NG75: are defined similarly to N50 but with 75 % instead of 50 %.
    - L50 (L75, LG50, LG75) is the number of contigs equal to or longer than N50 (N75, NG50, NG75). In other words, L50, for example, is the minimal number of contigs that cover half the assembly.

Phylogenetic analysis

Snippy

Snippy Snippy finds SNPs between a haploid reference genome and your NGS sequence reads. It will find both substitutions (snps) and insertions/deletions (indels). It can then take a set of Snippy results using the same reference and generate a core SNP alignment (and ultimately a phylogenomic tree).

Output directory: `05-snippy`

- `{sample_id}`
    - `snps.tab`: A simple tab-separated summary of all the variants
    - `snps.csv`: A comma-separated version of the .tab file
    - `snps.html`: A HTML version of the .tab file
    - `snps.vcf`: The final annotated variants in VCF format
    - `snps.bed`: The variants in BED format
    - `snps.gff`: The variants in GFF3 format
    - `snps.bam`: The alignments in BAM format. Includes unmapped, multimapping reads. Excludes duplicates.
    - `snps.bam.bai`: Index for the .bam file
    - `snps.log`: A log file with the commands run and their outputs
    - `snps.aligned.fa`: A version of the reference but with - at position with depth=0 and N for 0 < depth < --mincov (does not have variants)
    - `snps.consensus.fa`: A version of the reference genome with all variants instantiated
    - `snps.consensus.subs.fa`: A version of the reference genome with only substitution variants instantiated

    - `snps.raw.vcf`: The unfiltered variant calls from Freebayes
    - `snps.filt.vcf`: The filtered variant calls from Freebayes
    - `snps.vcf.gz`: Compressed .vcf file via BGZIP

- `core.aln`: A core SNP alignment in the --aformat format (default FASTA)
- `core.full.aln`: A whole genome SNP alignment (includes invariant sites)
- `core.tab`: Tab-separated columnar list of core SNP sites with alleles but NO annotations
- `core.vcf`: Multi-sample VCF file with genotype GT tags for all discovered alleles
- `core.txt`: Tab-separated columnar list of alignment/core-size statistics
- `core.ref.fa`: FASTA version/copy of the --ref
- `core.self_mask.bed`: BED file generated if --mask auto is used.

Output directory: `99-stats`

- `bamstat.csv`: summary file with all bamUtil information in one file for all samples.

Picard WGSMetrics

Picard WGSMetrics provides useful stats for the mapping step.

Output directory: `99-stats`

- `wgsmetrics.csv`: summary file with all Picard WGSMetrics information in one file for all samples.

IQTree

IQTree (6) performs a maximum likelihood phylogeny from a multiple sequence alignment.

Output directory: `03-iqtree`

- `core.full.iqtree.bootstrap.treefile`: phylogenetic tree in newick format with bootstrap values.

REFERENCES

1. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170.
2. Rapid and precise alignment of raw reads against redundant databases with KMA Philip T.L.C. Clausen, Frank M. Aarestrup, Ole Lund.
3. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 2017.
4. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014 Jul 15;30(14):2068-9.
5. Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi and Glenn Tesler. QUAST: quality assessment tool for genome assemblies, Bioinformatics (2013) 29 (8): 1072-1075.
6. B.Q. Minh, H.A. Schmidt, O. Chernomor, D. Schrempf, M.D. Woodhams, A. von Haeseler, R. Lanfear (2020) IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol., 37:1530-1534.