Rapid sequencing and characterization of bacterial genomes

Training component of the 2019 Inspire challenge project : Rapid genomic detection of aquaculture pathogens (CGIAR Platform for Big Data in Agriculture)

NGS data report

Link to data: as per email sent

DNA Extraction

Bacterial resuspension was spinned down followed by the removal of supernatant (ethanol) via decantation. DNA extraction was performed as per the method of Sokolov et al (Sokolov 2000) with some modifications. The pellet was resuspended in 500 µL of lysis buffer (50 mM NaCl, 50 mM Tris-HCl pH8, 50 mM EDTA, 2% SDS) and incubated at 60°C for 30 minutes. Then, 3 µL RNAse A (10 mg/mL) was added to the lysate followed by incubation at room temperature for 10 minutes. Salting out was performed via the addition of 50 µL (0.1x vol) saturated KCl at 4°C for 5 minutes. The lysate was subsequently extracted once with equal volume of chloroform to remove remaining proteins. The aqueous layer containing the DNA was mixed with an equal volume of isopropanol and 20 µL of SPRI bead to promote binding of DNA onto the solid carboxylated layer (Oberacker et al. 2019). After incubation at room temperature for 10 minutes, the mixture was placed on a magnetic rack for 2 minutes followed by the removal of supernatant. The bound magnetic bead was washed twice with 75% ethanol. DNA elution from the bead was performed by the resuspension of the bead with 100 µL of TE buffer followed by incubation at 50°C for 5 minutes.

Illumina library preparation and genome sequencing

Approximately 100 ng of DNA as measured by Qubit was fragmented to 350 bp using a Bioruptor followed by NEB Ultra II library preparation kit for Illumina according to the manufacturer's instructions (NEB, Ipswich, MA). Sequencing was performed on a NovaSEQ6000 (Illumina, San Diego, CA) generating approximately 1 gb of paired-end data (2x150 bp) for each sample.

Nanopore library preparation and genome sequencing

Approximately 400 ng of DNA as measured by Qubit was fragmented with the Nanopore rapid barcoding kit according to the manufacturer's instructions (Oxford Nanopore, UK). The samples were pooled and sequenced on a Nanopore Flonge flow cell. Basecalling of the fast5 file used Guppy v4.4.1 (high accuracy mode).

De novo assembly - Illumina

Raw Illumina paired-end reads were trimmed with fastp v0.21 (S. Chen et al. 2018) to remove low-quality bases and Illumina adapter sequences. The trimmed reads were subsequently used for *de novo* assembly in SPAdes v3.15.0 (--isolate setting) (Bankevich et al. 2012). Contigs smaller than 500 bp representing mostly sequencing artefact were removed and the filtered assembly was used for subsequent analysis.

Hybrid *De novo* assembly - Nanopore and Illumina

Raw nanopore reads were quality- and length-filtered to retain reads longer than 2,000 bp with qscore of 7 or higher. The filtered Nanopore were subsequently used in combination with the Illumina reads for hybrid assembly using Unicycler (default setting) (Wick et al. 2017). Contigs smaller than 500 bp representing mostly sequencing artefact were removed and the filtered assembly was used for subsequent analysis.

Assessment of Genome Assembly

Genome assembly statistics were generated using QUAST (Gurevich et al. 2013). Assessment of the genome completeness used BUSCO5 (Simão et al. 2015) that identified conserved microbial single copy genes as listed in the bacteria_odb10 database.

Taxonomic classification

Ribosomal RNA-containing contigs were identified and its corresponding rRNA genes (5S, 16S and 23S rRNA) were extracted using barrnap into a single fasta file that can be traditionally used to BLAST against the NCBI microbial 16S database. In addition, a more advanced and likely more accurate genome-based classification was also performed using kmerfinder v3 which assigns species-level classification based on a combination of unique DNA signatures (kmers) in the assembly (Hasman et al. 2014).

***In-silico* MLST, identification of AMR genes and virulence factors**

Subject to the availability of the species in the database, an *in-silico* MLST was performed on the assembled genome using the open-source mlst tool that will perform nucleotide similarity search against the pubmlst database (Jolley, Bray, and Maiden 2018). Abricate was employed to perform a BLAST-based nucleotide similarity search of the assembled genome against the curated NCBI AMR (Feldgarden et al. 2021) and virulence factor database (L. Chen et al. 2005). Gene region exhibiting more than 90% identity to the database were included in the report.

Data description

Within the report folder. The most important data will be the raw sequencing data. They have the .fastq.gz extension. For Nanopore, it will have the .nanopore.fastq.gz extension and Illumina data being paired-end will consist of two files with the .R1.fastq.gz and .R2.fastq.gz extension. Genome assembly is in the *.fasta format. Illumina-only assembly was performed with SPADES and hence it has the .spades.fasta extension. Hybrid assembly done with unicycler will have the .unicycler.fasta extension. Additional extensions behind other files are self-explanatory. For example, the *.quast file will be the output of QUAST analysis.

Folder structure of zip file

```
SAMPLE.report/
├── SAMPLE.Illumina_SequencingStats.txt
├── SAMPLE.nanopore.fastq.gz
├── SAMPLE.nanopore.sequencingstat.txt
├── SAMPLE.R1.fastq.gz
├── SAMPLE.R2.fastq.gz
├── SAMPLE.spades.busco.txt
├── SAMPLE.spades.fasta
├── SAMPLE.spades.quast
│   ├── basic_stats
│   │   ├── coverage_histogram.pdf
│   │   ├── cumulative_plot.pdf
│   │   ├── GC_content_plot.pdf
│   │   ├── Nx_plot.pdf
│   │   ├── SAMPLE-spades_coverage_histogram.pdf
│   │   └── SAMPLE.spades_GC_content_plot.pdf
│   ├── icarus.html
│   ├── icarus_viewers
│   │   └── contig_size_viewer.html
│   ├── quast.log
│   ├── report.html
│   ├── report.pdf
│   ├── report.tex
│   ├── report.tsv
│   ├── report.txt
│   ├── transposed_report.tex
│   ├── transposed_report.tsv
│   └── transposed_report.txt
├── SAMPLE.spades.resfinder.txt
├── SAMPLE.spades.rRNAseq.fna
├── SAMPLE.spades.virulencefactor.txt
├── SAMPLE.species_assign.txt
├── SAMPLE.unicycler.busco.txt
├── SAMPLE.unicycler.fasta
├── SAMPLE.unicycler.quast
│   ├── basic_stats
│   │   ├── cumulative_plot.pdf
│   │   ├── GC_content_plot.pdf
│   │   ├── Nx_plot.pdf
│   │   └── SAMPLE.unicycler_GC_content_plot.pdf
│   ├── icarus.html
│   ├── icarus_viewers
│   │   └── contig_size_viewer.html
│   ├── quast.log
│   ├── report.html
│   ├── report.pdf
│   ├── report.tex
│   ├── report.tsv
│   ├── report.txt
│   ├── transposed_report.tex
│   ├── transposed_report.tsv
│   └── transposed_report.txt
├── SAMPLE.unicycler.resfinder.txt
├── SAMPLE.unicycler.rRNAseq.fna
└── SAMPLE.unicycler.virulencefactor.txt
```

Acknowledgements

This work was undertaken as part of the CGIAR Research Program on Fish Agri-Food Systems (FISH) led by WorldFish, and the CGIAR Big Data Platform Inspire Challenge 2019 led by CIAT and IFPRI. These programmes are supported by contributors to the CGIAR Trust Fund. The funders provided support in the form of salary, laboratory consumables and analytical costs, but did not have any additional role in the study design, data collection and analysis. Processing of the samples including gDNA extraction, sequencing and data analysis were undertaken by GeneSEQ and WorldFish.

References

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, et al. 2012. "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing." *Journal of Computational Biology* 19 (5): 455–77. https://doi.org/10.1089/cmb.2012.0021.

Chen, Lihong, Jian Yang, Jun Yu, Zhijian Yao, Lilian Sun, Yan Shen, and Qi Jin. 2005. "VFDB: A Reference Database for Bacterial Virulence Factors." *Nucleic Acids Research* 33 (Database issue): D325-328. https://doi.org/10.1093/nar/gki008.

Chen, Shifu, Yanqing Zhou, Yaru Chen, and Jia Gu. 2018. "Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor." *Bioinformatics* 34 (17): i884–90. https://doi.org/10.1093/bioinformatics/bty560.

Feldgarden, Michael, Vyacheslav Brover, Narjol Gonzalez-Escalona, Jonathan G. Frye, Julie Haendiges, Daniel H. Haft, Maria Hoffmann, et al. 2021. "AMRFinderPlus and the Reference Gene Catalog Facilitate Examination of the Genomic Links among Antimicrobial Resistance, Stress Response, and Virulence." *Scientific Reports* 11 (1): 12728. https://doi.org/10.1038/s41598-021-91456-0.

Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. "QUAST: Quality Assessment Tool for Genome Assemblies." *Bioinformatics (Oxford, England)* 29 (8): 1072–75. https://doi.org/10.1093/bioinformatics/btt086.

Hasman, Henrik, Dhany Saputra, Thomas Sicheritz-Ponten, Ole Lund, Christina Aaby Svendsen, Niels Frimodt-Møller, and Frank M. Aarestrup. 2014. "Rapid Whole-Genome Sequencing for Detection and Characterization of Microorganisms Directly from Clinical Samples." *Journal of Clinical Microbiology* 52 (1): 139–46. https://doi.org/10.1128/JCM.02452-13.

Jolley, Keith A., James E. Bray, and Martin C. J. Maiden. 2018. "Open-Access Bacterial Population Genomics: BIGSdb Software, the PubMLST.Org Website and Their Applications." *Wellcome Open Research* 3: 124. https://doi.org/10.12688/wellcomeopenres.14826.1.

Oberacker, Phil, Peter Stepper, Donna M. Bond, Sven Höhn, Jule Focken, Vivien Meyer, Luca Schelle, et al. 2019. "Bio-On-Magnetic-Beads (BOMB): Open Platform for High-Throughput Nucleic Acid Extraction and Manipulation." *PLOS Biology* 17 (1): e3000107. https://doi.org/10.1371/journal.pbio.3000107.

Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. "BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs." *Bioinformatics* 31 (19): 3210–12. https://doi.org/10.1093/bioinformatics/btv351.

Sokolov, Eugene P. 2000. "An Improved Method for DNA Isolation from Mucopolysaccharide-Rich Molluscan Tissues." *Journal of Molluscan Studies* 66 (4): 573–75. https://doi.org/10.1093/mollus/66.4.573.

Wick, Ryan R., Louise M. Judd, Claire L. Gorrie, and Kathryn E. Holt. 2017. "Unicycler: Resolving Bacterial Genome Assemblies from Short and Long Sequencing Reads." *PLOS Computational Biology* 13 (6): e1005595. https://doi.org/10.1371/journal.pcbi.1005595.