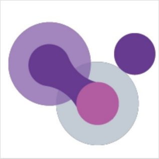

Version 2 ▼

Oct 27, 2020

GHRU (Genomic Surveillance of Antimicrobial Resistance) Retrospective 1 Bioinformatics Methods V.2

Anthony Underwood[1]

[1]Centre for Genomic Pathogen Surveillance

In Development This protocol is published without a DOI.

Anthony Underwood

ABSTRACT

A description of the pipelines used for analysing genome data from the retrospective 1 project of the NIHR Global Health Research Unit (Genomic Surveillance of Antimicrobial Resistance)

All genome sequence data were processed using versioned Nextflow [1] workflows and associated Docker [2] containers covering the foundational analyses of de novo assembly,mapping-based SNP phylogeny, MLST assignation and AMR determinant detection (table 1). The exact steps performed can be derived from examination of the pipeline code but each workflow will be described in brief.

PROTOCOL CITATION

Anthony Underwood 2020. GHRU (Genomic Surveillance of Antimicrobial Resistance) Retrospective 1 Bioinformatics Methods. **protocols.io**
https://protocols.io/view/ghru-genomic-surveillance-of-antimicrobial-resista-bn4pmgvn
Version created by Anthony Underwood

LICENSE

CREATED

Oct 27, 2020

LAST MODIFIED

Oct 27, 2020

PROTOCOL INTEGER ID

43887

ABSTRACT

A description of the pipelines used for analysing genome data from the retrospective 1 project of the NIHR Global Health Research Unit (Genomic Surveillance of Antimicrobial Resistance)

All genome sequence data were processed using versioned Nextflow [1] workflows and associated Docker [2] containers covering the foundational analyses of de novo assembly,mapping-based SNP phylogeny, MLST assignation and AMR determinant detection (table 1). The exact steps performed can be derived from examination of the pipeline code but each workflow will be described in brief.

1 **De novo assembly**

reads trimming and adapter removal using trimmomatic (0.38) [3], read correction using lighter (1.1.1) [4], downsampling to 100x coverage using seqtk (1.3) [5], read merging using flash (1.2.11) [6], assembly using SPAdes (3.12.0) [7]. Quality control was performed using fastqc (0.11.8) [8], multiqc (1.7) [9] and qualifyr (1.4.4) [10]. Species identification was carried out by bactinspector (0.1.3) [11] and contamination checked using Confindr (0.7.2) [12].

2 **Mapping based phylogeny**

Reads were trimmed as described for de novo assembly and mapped to a reference with bwa mem (0.7.17) [13], variants called and filtered using bcf tools (1.9) [14] and the filtering conditions for a low quality position being

'%QUAL<25 || FORMAT/DP<10 || MAX(FORMAT/ADF)<2 || MAX(FORMAT/ADR)<2 || MAX(FORMAT/AD)/SUM(FORMAT/DP)<0.9 || MQ<30 || MQ0F>0.1'. A pseudoalignment where each sample has a base relative to the reference sequence. Missing bases and low quality bases are encoded using the - and N characters. The alignment was used to generate a maximum likelihood tree using iqtree (1.6.8) [15,16] and ultrafast bootstraps (parameters -m GTR+G -alrt 1000 -bb 1000).

3 AMR determinant detection

The ARIBA software (2.14.4) [17] was used to detect acquired genes using the NCBI database [18] (downloaded 2019-10-30) and the pointfinder database (downloaded 2020-12-11) adapted for Ariba [19].

4 MLST detection

The ARIBA software (2.14.4) [17] was used to determine the 7-locus MLST type using the profile and alleles found in the pubmlst database [20,21] (downloaded 2020-12-20).

5 Table 1: Nextflow workflows

Workflow name	Workflow link	Docker hub Container(s) used	Version at publication
De novo assembly	https://gitlab.com/cgps/ghru/pipelines/assembly	bioinformant/ghru-assembly:versionORregistry.gitlab.com/cgps/ghru/pipelines/assembly:version	1.5.5
Mapping based SNP phylogeny	https://gitlab.com/cgps/ghru/pipelines/snp_phylogeny	bioinformant/ghru-snp-phylogeny:versionORregistry.gitlab.com/cgps/ghru/pipelines/snp_phylogeny:version	1.2.2
AMR determinant detection	https://gitlab.com/cgps/ghru/pipelines/dsl2/pipelines/amr_prediction	bioinformant/ghru-amr-prediction:versionORregistry.gitlab.com/cgps/ghru/pipelines/dsl2/pipelines/amr_prediction	1.0
MLST	https://gitlab.com/cgps/ghru/pipelines/dsl2/pipelines/mlst	bioinformant/ghru-mlst:version ORregistry.gitlab.com/cgps/ghru/pipelines/dsl2/pipelines/mlst:version	1.0

6 References

1. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol 2017; 35:316–319.
2. Merkel, D. Docker: lightweight linux containers for consistent development and deployment. 2014; 239.
3. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma Oxf Engl 2014; 30:2114–2120.
4. Song L, Florea L, Langmead B. Lighter: fast and memory-efficient sequencing error correction without counting. Genome Biol 2014; 15:509.
5. Li H. lh3/seqtk. 2020. Available at: https://github.com/lh3/seqtk. Accessed 26 October 2020.
6. Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 2011; 27:2957–2963.
7. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol J Comput Mol Cell Biol 2012; 19:455–477.
8. FastQC A Quality Control tool for High Throughput Sequence Data. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 7 October 2020.
9. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinforma Oxf Engl 2016; 32:3047–3048.
10. Qualifyr. Available at: https://gitlab.com/cgps/qualifyr. Accessed 7 October 2020.
11. BactInspector. Available at: https://gitlab.com/antunderwood/bactinspector. Accessed 26 October 2020.
12. Low AJ, Koziol AG, Manninger PA, Blais B, Carrillo CD. ConFindr: rapid detection of intraspecies and cross-species contamination in bacterial whole-genome sequence data. PeerJ 2019; 7:e6995.

13. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio 2013; Available at: http://arxiv.org/abs/1303.3997. Accessed 26 October 2020.
14. samtools/bcftools. samtools, 2020. Available at: https://github.com/samtools/bcftools. Accessed 26 October 2020.
15. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol 2015; 32:268–274.
16. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol 2018; 35:518–522.
17. Hunt M, Mather AE, Sánchez-Busó L, et al. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. Microb Genomics 2017; 3:e000131.
18. Bacterial Antimicrobial Resistance Reference Gene. Available at: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA313047. Accessed 26 October 2020.
19. ariba_amr_databases. Available at: https://gitlab.com/cgps/ghru/pipelines/data_sources/ariba_amr_databases. Accessed 26 October 2020.
20. Jolley KA, Chan M-S, Maiden MCJ. mlstdbNet - distributed multi-locus sequence typing (MLST) databases. BMC Bioinformatics 2004; 5:86.
21. Jolley KA, Maiden MCJ. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 2010; 11:595.