Preprint DOI: Posted on December 11, 2023
https://doi.org/10.1099/acmi.0.000646.v3

RSYD-BASIC: a bioinformatics pipeline for Routine Sequence analYsis and Data processing of BActerial iSolates for cllnical miCrobiology

1.1 Author names

Kat Steinke (ORCiD ID 0000-0003-1951-3001) 1, Karina Gravgaard Thomsen 1, Silje Vermedal Hoegh 1, Sanne Løkkegaard Larsen 1, Karina Kubel Vilhelmsen 1, Thøger Gorm Jensen 1, 2, Marianne Nielsine Skov (ORCiD ID 0000-0002-2619-0864) 1, 2, Thomas Vognbjerg Sydenham (ORCiD ID 0000-0003-1058-2449) 1, 2

1.2 Affiliation(s)

1 Department of Clinical Microbiology, Odense University Hospital - Odense (Denmark)

2 Research Unit of Clinical Microbiology, Faculty of Health, University of Southern Denmark - Odense (Denmark)

1.3 Corresponding author and email address

Kat Steinke; kat.steinke@rsyd.dk

1.4 Keywords

Bioinformatics; whole genome sequencing; clinical microbiology

1.5 Repositories

RSYD-BASIC code: https://gitlab.com/KatSteinke/rsyd-basic

2. Abstract

Background

Whole genome sequencing of bacterial isolates is increasingly becoming routine in clinical microbiology; however, subsequent analysis often needs to be started by a bioinformatician even for comprehensive pipelines. To increase the robustness of our workflow and free up bioinformatician work hours for development and advanced analysis, we aimed to produce a robust, customizable bioinformatic pipeline for bacterial genome assembly and routine analysis results that could be initiated by non-bioinformaticians.

Results

When tested on publicly available sequences, our pipeline yields comparable results in most cases. In routine use, it has already yielded clinically relevant results, allowing us to type a variety of bacterial pathogens isolated in our clinical laboratory and disprove a potential outbreak.

Conclusion

With the RSYD-BASIC pipeline, we present a reads-to-results analysis pipeline operated by non-expert users that greatly eases investigation of potential outbreaks. Results obtained with publicly available sequences are also promising, while underlining the importance of standardized methods.

3. Data summary

The code of the RSYD-BASIC pipeline is available at https://gitlab.com/KatSteinke/rsyd-basic.

GenBank accession numbers and/or PubMLST identifiers of sequences used for the test dataset and the example of combining RSYD-BASIC results with manual investigation are listed in the methods section.

The entire test dataset (reads and metadata files) and analysis results for this dataset are available on Zenodo at https://zenodo.org/record/8344050.

The authors confirm all supporting data, code and protocols have been provided within the article or through supplementary data files.

4. Introduction

Next generation sequencing of bacterial isolates or mixed samples has long been viewed as the next step in clinical diagnostics (1,2). For example, whole genome sequencing (WGS) of bacterial isolates can be used routinely for typing of pathogens including identification of transmission of multi-drug resistant bacteria (3,4), predicting *pathogen* serotypes (5),

improving species identification with possible clinical benefits (6) as well as identifying antimicrobial resistance genes and to some extent predict antimicrobial resistance (7) .

•

However, bioinformatic analyses are required to process and interpret the vast amounts of data generated.

A number of pipelines that can perform “reads-to-results” analyses exist. However, many commercial tools such as Illumina’s BaseSpace suite or 1928 require upload of sequence data, which is not always preferable in regards to patient data privacy. Commercial stand-alone solutions such as SeqSphere+ (8) exist, but often there will be local specific needs for tool implementations and development of in-house solutions to aid analysis. Bioinformatic analysis of genome sequencing data is a fast evolving field. Therefore laboratories frequently choose to implement own bioinformatics solutions. Open source tools that can run locally such as Bactopia (9), often require some familiarity with running programs from the command line. This can pose a challenge for laboratory technicians. Therefore starting the analysis pipeline often depends on a bioinformatician (or, depending on personnel resources, *the* bioinformatician), which makes the process considerably less robust. The combined output from the individual tools in such pipelines may be difficult to interpret (1) and from an operational perspective, building solutions that create outputs tailored for import into the local laboratory information systems can be preferable.

We aimed to develop and implement a customizable user-friendly open source pipeline to allow routine analysis of bacterial whole genome sequencing data in a clinical microbiology laboratory.

5. Results and Discussion

5.1 Workflow description

The RSYD-BASIC pipeline, once set up by a bioinformatician, can be started in two different ways: for routine uses, a “questionnaire”-style interface in the terminal guides through which files need to be supplied (see Figure 1 for an example), with most settings predefined through a default configuration file. For more experienced users, the pipeline can be launched through the command line as well, allowing for more fine-grained configuration.

As input, the pipeline takes a directory containing Illumina reads from one run in fastq

```
### Bacterial isolate genome assembly

# Setup analysis ------------------------------

Type full path or name of Illumina sequencing folder and press
enter:
/data/path/to/test_data/new_dataset/test_data/Alignment_1/test_
dir/Fastq

Output directory will be based on experiment name.

Enter path to Illumina runsheet:
/data/path/to/test_data/runsheet.xlsx

Do you accept /data/path/to/results/ILM_RUN0000_Y20221005_XYZ
as target folder [y/n] y

Running analysis pipeline
```

Figure 1: Example of the pipeline's "questionnaire" mode with a test dataset, showing the pipeline's prompts and the user's input. User input is bolded for clarity in this only; this does not represent a feature of the pipeline.

format, as well as a "run sheet" with metadata for the run, including indications for sequencing (for detailed requirements, see the pipeline's repository). An English translation of the metadata sheet can be found as Supplementary File 1; note that the pipeline currently expects Danish column names as it primarily interfaces with Danish healthcare systems. A report from the laboratory information system (LIS) can be supplied as well, enabling cross-checking of sample numbers and supplied species identification; the location of this report is specified through the configuration file. An example with English explanations is given as supplementary table 1. (Note that these supplementary files cannot be used as direct input to the pipeline without customization of the pipeline; they are included as explanatory for non-Danish speakers.)

The analysis workflow, shown in Figure 2, takes inspiration from other reads-to-results pipelines such as Bactopia, and is implemented using Snakemake (10).

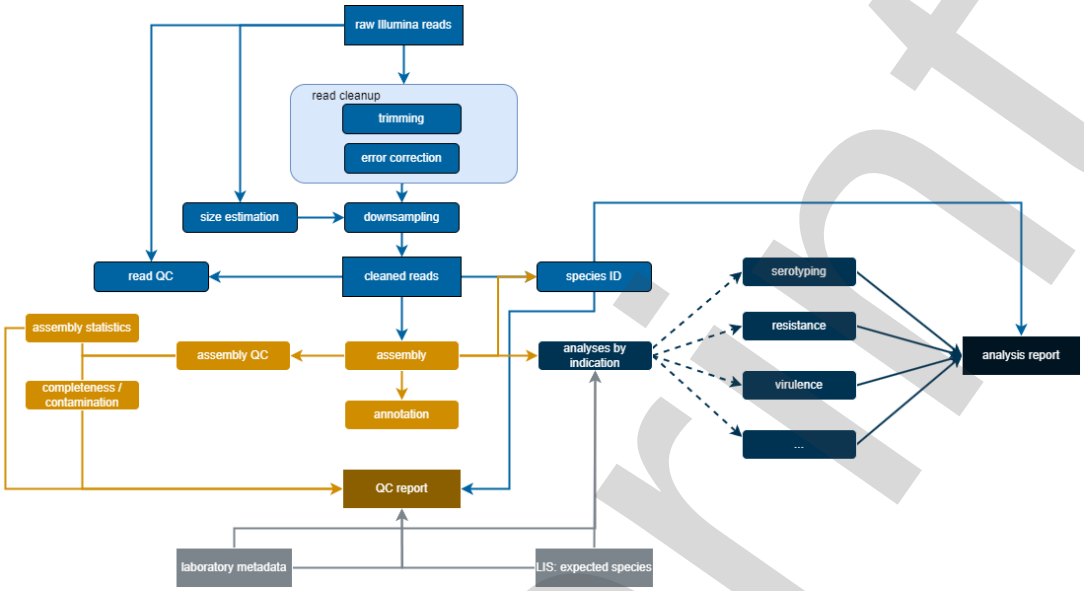

Figure 2: The data flow in the RSYD-BASIC pipeline. Light blue boxes represent operations performed and data obtained from raw Illumina reads; golden boxes represent operations performed on assemblies; dark blue boxes represent specific analyses based on indication; gray boxes represent external data sources. The darkest golden and blue boxes represent the analysis results for the respective inputs.

Read sequences for each sample are initially cleaned by removing the sequencing adapter and any reads belonging to the PhiX sequencing control with bbduk (11). Errors are then corrected with lighter (12). From the raw reads, genome size is estimated using Mash (13); this estimate is then used to downsample the cleaned, corrected reads to a maximum of 100x coverage (14), if required, using reformat.sh from the BBTools suite (11). Finally, quality control reports on the reads before and after cleaning are obtained using FastQC (15).

The cleaned reads are also analyzed with Kraken2 (16) with a database specified by the user (in the case of our use case at the Department of Clinical Microbiology, Odense University (DCM OUH), the Standard-8 database, version 12/9/2022). This allows for investigation of contamination later in the process, as the Kraken output shows the proportion of reads belonging to each organism. Isolate sequences are also compared to databases of RefSeq sequences using mash (13) and sourmash (17).

Most other analyses are performed on assemblies, which are obtained using shovill (18), with skesa as the assembler.

General quality control metrics such as N50, NG50, genome size and amount of contigs are obtained using QUAST (19). Completeness and contamination are estimated using CheckM (20). Finally, species identification is performed using GTDB-Tk (21). QC results as well as the GTDB-Tk species call and the species for which the highest proportion of reads match in

Kraken are reported in a QC-focused result file. If a LIS report has been supplied, any preliminary species ID registered in the LIS as well as the expected genome size for this species (if available) will be shown in the QC results for comparison to aid with detection of laboratory errors.

In addition, clinically relevant properties of the isolates are investigated. By default, this currently encompasses

- Resistance genes (using abritAMR (22))
- Plasmids (using PlasmidFinder (23))
- MLST typing (using mlst (24,25))

If requested in the run sheet, selected virulence or toxin genes can also be reported; these are identified from abritAMR's output as well.

Genomes are annotated using prokka (26).

Species-specific analyses are currently performed only when a LIS report is given; currently, the only species-specific analyses are serovar identification for Salmonella (using SeqSero2 (27) on cleaned reads) and serotyping for E. coli (using SerotypeFinder (28)).

These results are then compiled into an analysis result file.

Finally, QC results are evaluated – this is partially automated, but some results may require manual examination.

After QC evaluation, analysis results may be used further in routine procedures. At DCM OUH, analysis results of sequences that pass quality requirements are entered into the department's LIS. Selected analysis results and metadata for all of the batch's sequences, including QC pass/fail information, are imported into a custom MySQL database, so that sample numbers, file locations etc. can be retrieved, e.g. for outbreak investigations.

At DCM OUH the pipeline is run on a regional compute cluster with 256 cores and 160 GiB of memory.

5.2 Example output

To demonstrate the pipeline's functionality, a small basic test set has been compiled from publicly available Illumina MiSeq reads. The test set consists of *Salmonella* Newport to test serovar detection, *Escherichia coli* for serotyping, *Klebsiella pneumoniae* for resistance and plasmid detection, a *Streptococcus pneumoniae* sample, and a deliberately "bad" sample generated by subsampling the *Strep. pneumoniae* sample.

Key results for the samples are shown and compared with known results in Table 1.

Table 1: Comparison of RSYD-BASIC results to original results for a publicly available test set

Sample number	RSYD-BASIC species call	RSYD-BASIC serovar / serotype	RSYD-BASIC MLST	RSYD-BASIC toxin genes	original species call	original serovar / serotype	original MLST	original toxin genes
1199234567-1	*Salmonella enterica*	Newport	46		*Salmonella enterica*	Newport	46	
1199234567-2	*Escherichia coli*	O22:H8	446	STX1A, STX2C	*Escherichia coli*	O22:H8		stx2
1199234567-3	*Klebsiella pneumoniae*		395		*Klebsiella pneumoniae*		2674	
1199234567-4	*Streptococcus pneumoniae*		416		*Streptococcus pneumoniae*		416	
1199234567-5	NA	NA	NA	NA	*Streptococcus pneumoniae*		416	

The results obtained with the RSYD-BASIC pipeline generally agree with those found in the articles initially describing the sequences analyzed, with two differences. Firstly, no results are obtained for 1199234567-5; this is to be expected, as it deliberately simulates a bad sample. Secondly, the MLST type of the *K. pneumoniae* sample does not match that given in the original article. However, the two types differ only by a single allele (29), and reanalysis of the original assembly deposited in NCBI's database with both the mlst tool (24) in the RSYD-BASIC pipeline and the current version of the MLST tool (30) used by Fursova et al. (31) in the original article yielded MLST type 395 as well. However, the Center for Genomic Epidemiology's MLST tool also allows the use of reads. Therefore, the MLST analysis was repeated using the original reads. The MLST type reported for analysis of the reads was also MLST type 395.

The results obtained from the test dataset using RSYD-BASIC are generally comparable to those found in the articles describing the original sequences, which were obtained with a range of different methods. However, the *K. pneumoniae* sample has a different MLST type than the one originally given in literature. The two types differ by a single allele, with the difference between the two alleles being a single nucleotide (29). Any difference in read filtering, error correction or assembly may therefore have caused the discrepancy – this may also explain the difference between the publicly available sequence (assembled using unicycler) and that used in the original article by Fursova et al. (assembled using SPAdes). This underlines the importance of standardized analysis methods to ensure consistent results.

An English translation of the full result files can be found in supplementary tables 2 (QC report) and 3 (analysis report). Note that the pipeline, due to its current primary use being in Danish healthcare systems, will create this output, including any automated notes, in Danish. The original outputs are available at https://zenodo.org/record/8344050.

5.3 Clinical application

From October 2022 until late March 2023, the pipeline has been used to process sequencing data from 498 individual bacterial isolates in 30 analysis runs. Results from these analyses have been used in various clinical contexts.

In one case, the department of nephrology at OUH noticed an increase in *Staphylococcus aureus* central line-associated blood stream infections (CLABSI) among patients receiving haemodialysis. The DCM was requested to investigate this as a potential outbreak. As CLABSI are included in the prospective routine sequencing and the RSYD-BASIC pipeline performs MLST typing by default, it was immediately possible to conclude that most samples from this department had different sequence types with a distance of multiple alleles to each other. This suggested that the majority of cases were not closely related to each other. Two pairs of isolates showed the same sequence type, of which one pair originated from the same patient. By applying core genome MLST and SNP analysis, it was concluded that the other pair of isolates were too distantly related to represent an outbreak. The applied approach to prospectively disprove the suspected outbreak permitted the department of nephrology to focus on improving patient related infection control practices, rather than performing outbreak investigation and possible screening personnel and environments (32). This example highlights one aspect of the value of prospective WGS for surveillance purposes - the ability to disprove clonal outbreaks. Several reports have been published arguing the cost effectiveness, clinical value and merit for infection control of implementing routine WGS of select bacterial isolates (33,34,35,36). In a future version of RSYD-BASIC automated phylogeny and screening reports for defined surveillance species may be added to support infection control efforts in real time.

6. Conclusions

Routine sequencing is a powerful tool in clinical microbiology, but the vast amounts of data it produces must be analysed to harness its power.

With RSYD-BASIC, we demonstrate a user-friendly open source pipeline that, once set up by a bioinformatician, allows users without in-depth familiarity with the command line to obtain a broad range of clinically relevant results from bacterial isolate sequences.

When tested on publicly available data, RSYD-BASIC reached the same results as the original studies for most samples. However, in one case, a difference of a single nucleotide led to a difference in MLST types; this serves to underline the importance of standardized workflows.

Bioinformatic analysis is often one of the hurdles in implementing truly routine WGS of bacterial isolates. When such analyses can only be performed by bioinformatic experts, this is not only time-consuming, but also carries the risk of one person's absence or illness completely stopping the process. With a pipeline that can be routinely started by laboratory technicians, the laboratory workflow is more robust. Additionally, bioinformaticians are able

to spend more time on in-depth analyses that require their expertise, or on developing and extending bioinformatic tools. The range of information generated by RSYD-BASIC also provides us with a “head start” in outbreak investigations, as more in-depth and computationally expensive analyses can be performed subsequently in a more targeted manner.

7. Methods

7.1 Test dataset acquisition

Existing publicly available reads for species of interest were downloaded from the NCBI’s Short Read Archive (SRA) in SRA Normalized Format (preserving quality scores), using version 2.10.0 of the SRA toolkit (37). Read accessions and sample numbers assigned in the test dataset are shown in Table 2.

To provide a deliberately “failed” sample (sample 1199234567-5), 1000 forward reads and 1000 reverse reads were sampled from read set SRR10955980 using seqtk sample (38).

Table 2: Reads used in the test dataset

Sample number	SRR accession number	Species	Source
1199234567-1	ERR9793822	*Salmonella* Newport	(39)
1199234567-2	SRR7235142	*Escherichia coli*	(40)
1199234567-3	SRR14194623	*Klebsiella pneumoniae*	(31)
1199234567-4	SRR10955980	*Streptococcus pneumoniae*	(41)
1199234567-5	SRR10955980	*Streptococcus pneumoniae*	(41)

7.2 Evaluation of results

Results were compared against those found in the articles originally describing the sequences. Where a discrepancy was found, this was investigated using the original assembly used in the article. This was only the case for *K. pneumoniae*, Both the RSYD-BASIC assembly and the original assembly (GCA_018138665.1) were analyzed using both

the tool used in RSYD-BASIC (mlst, (24,25)) and the tool used by Fursova et al. (31) (MLST, (30)). The latter was also used to analyse the raw reads. For this, the most current database at the time of analysis (version 2023-06-19) was used.

7.3 Sample statistics

Sample statistics were extracted in R using the tidyverse package (42).

7.4 Manual outbreak investigation

All isolates with the same MLST type as identified by mlst in the RSYD-BASIC pipeline were analyzed further with ChewBBACA (43) using the *S. aureus* cgMLST scheme of Leopold et al. (44), processed with ChewBBACA's PrepExternalSchema function. For comparison, unrelated sequences from PubMLST (29) with the same sequence type were added (see Table 3)). Alleles were called with AlleleCall and the results cleaned with ExtractCgMLST at default settings.

A minimum spanning tree was then constructed using the goeBURST Full MST functionality in PhyloViz (45).

SNP analyses were performed using snippy's snippy-multi functionality (46). Sequences were supplied as assemblies, and *S. aureus* SCAID OTT1-2021 (GenBank accession number CP082813.1) was used as the reference sequence.

Table 3: *Staphylococcus aureus* sequences used as background for cgMLST

Database ID	ST	Database	URL
42320	45	PubMLST	https://pubmlst.org/bigsdb?page=info&db=pubmlst_saureus_isolates&set_id=1&id=42320
41852	45	PubMLST	https://pubmlst.org/bigsdb?page=info&db=pubmlst_saureus_isolates&set_id=1&id=41852
41843	45	PubMLST	https://pubmlst.org/bigsdb?page=info&db=pubmlst_saureus_isolates&set_id=1&id=41843
42318	1	PubMLST	https://pubmlst.org/bigsdb?page=info&db=pubmlst_saureus_isolates&set_id=1&id=42318
39207	1	PubMLST	https://pubmlst.org/bigsdb?page=info&db=pubmlst_saureus_isolates&set_id=1&id=39207
39450	1	PubMLST	https://pubmlst.org/bigsdb?page=info&db=pubmlst_saureus_isolates&set_id=1&id=39450
39288	30	PubMLST	https://pubmlst.org/bigsdb?page=info&db=pubmlst_saureus_isolates&set_id=1&id=39288
41833	30	PubMLST	https://pubmlst.org/bigsdb?page=info&db=pubmlst_saureus_isolates&set_id=1&id=41833
41841	30	PubMLST	https://pubmlst.org/bigsdb?page=info&db=pubmlst_saureus_isolates&set_id=1&id=41841

8. Figures and tables

9. Author statements

9.1 Author contributions

Kat Steinke: Conceptualisation, Formal Analysis, Investigation, software, validation, writing – original draft, Visualisation

Karina Gravgaard Thomsen: Conceptualisation, Writing –, review and editing

Silje Vermedal Hoegh: Conceptualisation, Writing –, review and editing

Sanne Løkkegaard Larsen: Conceptualisation, Writing –, review and editing

Karina Kubel Vilhelmsen: Conceptualisation, Writing –, review and editing, Software

Thøger Gorm Jensen: Conceptualisation, Writing –, review and editing

Marianne Skov: Conceptualisation, Writing –, review and editing, Funding acquisition, Supervision

Thomas Vognbjerg Sydenham: Conceptualisation, Writing –, review and editing, Funding acquisition, Project administration, Supervision

9.2 Conflicts of interest

The authors declare that there are no conflicts of interest.

9.3 Funding information

This work was supported by internal funding.

9.4 Ethical approval

No experimental work with humans or animals was performed.

9.5 Consent for publication

The article contains no information that may permit the identification of individuals.

9.6 Acknowledgements

KS wishes to thank Henrik Johansen for his assistance with setting up the cluster version of the pipeline and Martin Vad Møller for additional testing of the pipeline.

10. References

1. Balloux F, Brynildsrud OB, van Dorp L, Shaw LP, Chen H, Harris KA, et al. From Theory to Practice: Translating Whole-Genome Sequencing (WGS) into the Clinic. Trends in Microbiology. 2018 December: p. 1035-1048.

2. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. Nature Reviews Genetics. 2012 August: p. 601-612.

3. Forde BM, Bergh H, Cuddihy T, Hajkowicz K, Hurst T, Playford EG, et al. Clinical Implementation of Routine Whole-genome Sequencing for Hospital Infection Control of Multi-drug Resistant Pathogens. Clinical Infectious Diseases. 2023 February 1: p. e1277-e1284.

4. Werner G, Couto N, Feil EJ, Novais A, Hegstad K, Howden BP, et al. Taking hospital pathogen surveillance to the next level. Microbial Genomics. 2023 April 26.

5. Cooper AL, Low AJ, Koziol AG, Thomas MG, Leclair D, Tamber S, et al. Systematic Evaluation of Whole Genome Sequence-Based Predictions of Salmonella Serotype and Antimicrobial Resistance. Frontiers in Microbiology. 2020 April.

6. Price TK, Realegeno S, Mirasol R, Tsan A, Chandrasekaran S, Garner OB, et al. Validation, Implementation, and Clinical Utility of Whole Genome Sequence-Based Bacterial Identification in the Clinical Microbiology Laboratory. Journal of Molecular Diagnostics. 2021 November: p. 1468-1477.

7. Zankari E, Hasman H, Kaas RS, Seyfarth AM, Agersø Y, Lund O, et al. Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial

susceptibility testing. Journal of Antimicrobial Chemotherapy. 2013 April: p. 771-777.

8. Bletz S, Mellmann A, Rothgänger J, Harmsen D. Ensuring backwards compatibility: traditional genotyping efforts in the era of whole genome sequencing. Clinical Microbiology and Infection. 2015 April: p. 347.e1-347.e4.

9. Petit RA, Read TD. Bactopia: a Flexible Pipeline for Complete Analysis of Bacterial Genomes. mSystems. 2020: p. e00190-20.

10. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH,SV, Forster J, et al. Sustainable data analysis with Snakemake [version 1; peer review: 1 approved, 1 approved with reservations]. F1000Research. 2021 april 2021.

11. Bushnell B. BBMap. [Online]. [cited 2023 03 27. Available from: https://sourceforge.net/projects/bbmap/.

12. Song L, Florea L, Langmead B. Lighter: fast and memory-efficient sequencing error correction without counting. Genome Biology. 2014.

13. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology. 2016 June 20.

14. Desai A, Marwah VS, Yadav A, Jha V, Dhaygude K, Bangar U, et al. Identification of Optimum Sequencing Depth Especially for De Novo Genome Assembly of Small Genomes Using Next Generation Sequencing Data. PLOS ONE. 2013 April: p. e60204.

15. FastQC. [Online]. [cited 2023 03 27. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

16. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biology. 2019 November 28.

17. Brown CT, Irber L. sourmash: a library for MinHash sketching of DNA. Journal of Open Source Software. 2016 September 14: p. 27.

18. Seemann T. Shovill. [Online]. [cited 2023 March 27. Available from: https://github.com/tseemann/shovill.

19. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 2018 July 1: p. i142-i150.

20. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research. 2015 July.

21. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics. 2022 December 1: p. 5315-5316.

22. Sherry NL, Horan KA, Ballard SA, Gonçalves da Silva A, Gorrie CL, Schultz MB, et al. An ISO-certified genomics workflow for identification and surveillance of antimicrobial resistance. Nature Communications. 2023 January 4.

23. Carattoli A, Zankari E, García-Fernández A, Larsen MV, Lund O, Villa L, et al. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrobial Agents and Chemotherapy. 2014 June 12.

24. Seemann T. mlst. [Online]. [cited 2023 March 27. Available from: https://github.com/tseemann/mlst.

25. Jolley KA, Maiden MC. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics. 2010 December 10.

26. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014 July 15: p. 2068-9.

27. Zhang S, den Bakker HC, Li S, Chen J, Dinsmore BA, Lane C, et al. SeqSero2: Rapid and Improved Salmonella Serotype Determination Using Whole-Genome Sequencing Data. Applied and Environmental Microbiology. 2019 November 14.

28. Joensen KG, Tetzschner AMM, Iguchi A, Aarestrup FM, Scheutz F. Rapid and Easy In Silico Serotyping of Escherichia coli Isolates by Use of Whole-Genome Sequencing Data. Journal of Clinical Microbiology. 2015 August: p. 2410-2426.

29. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Research. 2018 September 24.

30. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, et al. Multilocus sequence typing of total-genome-sequenced bacteria. Journal of Clinical Microbiology. 2012 April: p. 1355-61.

31. Fursova NK, Astashkin EI, Ershova ON, Aleksandrova IA, Savin IA, Novikova TS, et al. Multidrug-Resistant Klebsiella pneumoniae Causing Severe Infections in the Neuro-ICU. Antibiotics (Basel). 2021 August 13.

32. Sydenham TV, Thomsen KG, Steinke K, Hoegh SV, Larsen SL, Kubel Vilhelmsen K, et al. Establishing decentralised routine bacterial whole genome sequencing at a clinical microbiology department. 2023 April 15. 33rd ECCMID, the European Congress of Clinical Microbiology and Infectious Diseases ; Conference date: 15-04-2023 Through 18-04-2023.

33. Price V, Ngwira LG, Lewis JM, Baker KS, Peacock SJ, Jauneikaite E, et al. A systematic review of economic evaluations of whole-genome sequencing for the surveillance of bacterial pathogens. Microbial Genomics. 2023 February.

34. Parcell BJ, Gillespie SH, Pettigrew KA, Holden MTG. Clinical perspectives in integrating whole-genome sequencing into the investigation of healthcare and public health outbreaks – hype or help? Journal of Hospital Infection. 2021 March: p. 1-9.

35. Elliott TM, Hare N, Hajkowicz K, Hurst T, Doidge M, Harris PN, et al. Evaluating the economic effects of genomic sequencing of pathogens to prioritise hospital patients competing for isolation beds. Australian Health Review. 2020 October: p. 59-65.

36. Gordon LG, Elliott TM, Forde B, Mitchell B, Russo PL, Paterson DL, et al. Budget impact analysis of routinely using whole-genomic sequencing of six multidrug-resistant bacterial pathogens in Queensland, Australia. BMJ Open. 2021: p. e041968.

37. SRA Toolkit Development Team. SRA Toolkit. [Online]. [cited 2023 August 01. Available from: https://github.com/ncbi/sra-tools.

38. Li H. seqtk. [Online]. [cited 2023 August 01. Available from: https://github.com/lh3/seqtk.

39. Jansson Mörk M, Karamehmedovic N, Hansen A, Nederby Öhd J, Lindblad M, Östlund E, et al. Outbreak of Salmonella Newport linked to imported frozen cooked crayfish in dill brine, Sweden, July to November 2019. Eurosurveillance. 2022 June: p. 2.

40. Wang LYR, Jokinen CC, Laing CR, Johnson RP, Ziebell K, Gannon VPJ. Assessing the genomic relatedness and evolutionary rates of persistent verotoxigenic Escherichia coli serotypes within a closed beef herd in Canada. Microbial Genomics. 2020 June 3: p. e000376.

41. Spanelova P, Jakubu V, Malisova L, Musilek M, Kozakova J, Papagiannitsis CC, et al. Whole genome sequencing of macrolide resistant Streptococcus pneumoniae serotype 19A sequence type 416. BMC Microbiology. 2020 July 25.

42. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the tidyverse. Journal of Open Source Software. 2019: p. 1686.

43. Silva M, Machado MP, Silva DN, Rossi M, Moran-Gilad J, Santos S, et al. chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. Microbial Genomics. 2018 March 15.

44. Leopold SR, Goering RV, Witten A, Harmsen D, Mellmann A. Bacterial whole-genome sequencing revisited: portable, scalable, and standardized analysis for typing and detection of virulence and antibiotic resistance genes. Journal of Clinical Microbiology. 2014 July: p. 2365-70.

45. Nascimento M, Sousa A, Ramirez M, Francisco AP, Carriço JA, Vaz C. PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. Bioinformatics. 2017 January 1: p. 128-129.

46. Seemann T. snippy: Rapid haploid variant calling and core genome alignment. [Online]. [cited 2023 05 30. Available from: https://github.com/tseemann/snippy.

310

311

Dear Dr. Munnoch,
Once again, thank you for the opportunity to submit a revised version of our manuscript "RSYD-BASIC: a bioinformatics pipeline for Routine Sequence analYsis and Data processing of BActerial iSolates for clInical microbiology" to Access Microbiology, and thank you very much for your detailed comments.
We have improved the manuscript according to your feedback; a version with highlighted changes is attached. Please see below a point-by-point response (in blue) to your comments. Any line numbers refer to the version with tracked changes.

Editor comments:

This study would be a valuable contribution to the existing literature.

This is a study that would be of interest to the field and community.

Thank you for your efforts so far. I'm returning the manuscript with similar comments as before but with more detail on addressing some issues. In general, the manuscript is very short and to the point. Much of the information is obviously in the Gitlab page but the point of the publication is more than to advertise this (as this could technically be cited directly). Once the below changes are made, i believe the manuscript will be suitable for official review.

Notes by section:
Author names, typically I would expect this to be a single line.

Thank you for catching this – we have fixed the formatting now.

Abstract:
In general, these consist of between 200 and 250 words. I would encourage you to use this word limit. Abstracts form the basis of how many readers dive into a paper. As it stands, the lack of information within the abstract may indeed be an issue for readers. It should also form similar to a "mini-paper" I.e. it is constructed with introductory material, perhaps some methods, results and your major conclusions/take home message. While generalised, I would expect in your case it to contain more information. This may seem redundant but its compounded by the brevity of the manuscript itself.

We'd originally erred on the side of brevity with this being a short communication, but appreciate the opportunity to go into more detail.

Introduction:
I would include a reference/example where possible for line 54.

Due to the brevity of the manuscript, I would either include examples of metadata sheets with descriptions directly or minimum link specifically to the file (I see there is one in the Gitlab).

Additionally, as the manuscript is in English, so should the metadata sheet etc. I would request all primary resources be in the language of the manuscript where possible.

We have added translated versions of the tables where possible and added explanations to the LIS report to make the example input and output easier to understand. However, these of course currently cannot be used as input or expected as output to the pipeline, as the pipeline interfaces with existing systems that produce Danish output and require Danish input (input forms for metadata sheet creation and our LIS). We agree that internationalized input and output would be ideal and aim to include this in a future version of RSYD-BASIC.

Results:
I would discourage the use of a screen shot for figure 1 and instead use formatted in line text similar to would you would find her in the installation section:
https://github.com/rrwick/Deepbinner

git clone https://github.com/rrwick/Deepbinner.git
pip3 install ./Deepbinner
deepbinner –help

This is partly due to standard formatting approaches (which you have included in the Gitlab repository) but also that screenshot resolution/sizing can be difficult to adjust for readers with additional visual needs.

Thank you for the feedback – we had intended to show the program "in action" but understand the screenshot format is less accessible. We have therefore replaced it with monospaced text intended to be formatted as code, with user input bolded for clarity.

I'd expand on the figure legends where possible, two for example could include information on the colour scheme, why it's important.

We have now expanded on the legends and given additional information on the choice of color scheme.

Due to the nature of the paper, I would encourage expanding some of the description steps from line 94. The information included is enough for people who are to some extent experienced but not those that are likely to be the primary users. "Read cleaning" for example isn't clear unless you have some experience with library generation.

We have now added more detailed information, such as describing the read cleaning process in more detail on lines 125-126 and elaborating on the purpose of read-based analyses on lines 135-136.

This section highlights clearly the benefit of the pipeline, in general, the manuscript

Discussion

Typically, this section is used to place the results/manuscript in context with the literature. Due to the lack of references in the section, it is relatively redundant. I encourage expanding on this section or forming a combined results and discussion section. In both cases, I would expect fairly substantial expansion. For example, lines 174-184, do you have any rationale for why the differences exist? – is this due to versions of software, different software being used in analysis etc.

We have now restructured the results and discussion section by merging them as suggested. We have also addressed the discrepancy in the MLST types for *K. pneumoniae* through closer investigation, which we have detailed in lines 185-188 and discussed in the subsequent paragraph (lines 189-198).

In addition, we have moved our conclusions to a dedicated section (lines 230-249).