Systematic benchmarking of 'all-in-one' microbial SNP calling pipelines

Authors:

Caitlin Falconer[1], Thom Cuddihy[1], Scott A. Beatson[2,4], David L. Paterson[1,2], Patrick NA. Harris[1,2,3], Brian M. Forde[1,2*]

Affiliations:

1. University of Queensland, Faculty of Medicine, UQ Centre for Clinical Research and, QLD, Australia
2. Australian Infectious Disease Research Centre, University of Queensland, Brisbane, QLD, Australia
3. Central Microbiology, Pathology Queensland, Royal Brisbane & Women's Hospital, QLD, Australia
4. School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane, QLD, Australia

***Corresponding author**: B M. Forde; University of Queensland Centre for Clinical Research, Building 71/918 Royal Brisbane & Women's Hospital Campus, Herston, QLD, 4029; **Email:** b.forde@uq.edu.au; T**el**: +61 (0) 7 3346 5474

ORCID: 0000-0002-0476-7714 (CF); 0000-0002-2895-0345 (PNAH); 0000-0002-2264-4785 (BMF); 0000-0002-1806-3283 (SAB); 0000-0003-2079-4437 (DLP);

Keywords: Bacterial genomics, benchmarking, bioinformatics, infectious disease surveillance, single nucleotide polymorphism, whole genome sequencing

Abstract

Clinical and public health microbiology is increasingly utilising whole genome sequencing (WGS) technology and this has lead to the development of a myriad of analysis tools and bioinformatics pipelines. Single nucleotide polymorphism (SNP) analysis is an approach used for strain characterisation and determining isolate relatedness. However, in order to ensure the development of robust methodologies suitable for clinical application of this technology, accurate, reproducible, traceable and benchmarked analysis pipelines are necessary. To date, the approach to benchmarking of these has been largely ad-hoc with new pipelines benchmarked on their own datasets with limited comparisons to previously published pipelines.

In this study, Snpdragon, a fast and accurate SNP calling pipeline is introduced. Written in Nextflow, Snpdragon is capable of handling small to very large and incrementally growing datasets. Snpdragon is benchmarked using previously published datasets against six other all-in-one microbial SNP calling pipelines, Lyveset, Lyveset2, Snippy, SPANDx, BactSNP and Nesoni. The effect of dataset choice on performance measures is demonstrated to highlight some of the issues associated with the current available benchmarking approaches.

The establishment of an agreed upon gold-standard benchmarking process for microbial variant analysis is becoming increasingly important to aid in its robust application, improve transparency of pipeline performance under different settings and direct future improvements and development.

Snpdragon is available at https://github.com/FordeGenomics/SNPdragon.

Impact statement

Whole-genome sequencing has become increasingly popular in infectious disease diagnostics and surveillance. The resolution provided by single nucleotide polymorphism (SNP) analyses provides the highest level of insight into strain characteristics and relatedness. Numerous approaches to SNP analysis have been developed but with no established gold-standard benchmarking approach, choice of bioinformatics pipeline tends to come down to laboratory or researcher preference. To support the clinical application of this technology, accurate, transparent, auditable, reproducible and benchmarked pipelines are necessary. Therefore, Snpdragon has been developed in Nextflow to allow transparency, auditability and reproducibility and has been benchmarked against six other all-in-one pipelines using a number of previously published benchmarking datasets. The variability of performance measures across different datasets is shown and illustrates the need for a robust, fair and uniform approach to benchmarking.

Data Summary

1. Previously sequenced reads for *Escherichia coli* O25b:H4-ST131 strain EC958 are available in BioProject PRJNA362676. BioSample accession numbers for the three benchmarking isolates are:
    - EC958: SAMN06245884
    - MS6573: SAMN06245879
    - MS6574: SAMN06245880
2. Accession numbers for reference genomes against the *E. coli* O25b:H4-ST131 strain EC958 benchmark are detailed in table 2.

3. Simulated benchmarking data previously described by Yoshimura et al. is available at http://platanus.bio.titech.ac.jp/bactsnp (1).
4. Simulated datasets previously described by Bush et al. is available at http://dx.doi.org/10.5287/bodleian:AmNXrjYN8 (2).
5. Real sequencing benchmarking datasets previously described by Bush et al. are available at http://dx.doi.org/10.5287/bodleian:nrmv8k5r8 (2).

Introduction

Microbial whole genome sequencing (WGS) is increasingly being used to support pathogen detection, surveillance, and diagnostics (1, 3). WGS provides the highest level of genomic resolution which allows for the ability to distinguish between closely related isolates and infer potential transmission events (1). Characterisation of isolates at the whole genome level in combination with clinical and epidemiological information can greatly benefit public health microbiology activities (4). There are many examples of the various advantages of WGS in pathogen detection and surveillance and perhaps the most recent is its usefulness in tracking community transmission of SARS-CoV-2 (3, 5). The application of WGS in public health microbiology is now progressing from proof-of-concept to implementation, particularly in food-borne pathogen surveillance and antimicrobial resistant bacterial outbreak detection (6, 7).

Determining isolate relatedness typically involves examining single nucleotide polymorphisms (SNPs). A typical SNP calling workflow includes the following steps: 1. Quality Control; 2. Read mapping; 3. Variant calling; 4. Variant filtering; 5. Downstream analysis (phylogenetic reconstruction and pairwise SNP difference clustering) (figure 1**)**.

By comparing SNPs present within the core genome (the shared region common to all isolates under evaluation), potential transmission events can be identified (8, 9). This information may be used to classify potential outbreak events and when combined with epidemiological data may inform infection prevention and control practices (8). In order to classify isolates as 'related' thresholds based on the number of core SNP differences are applied (10). The selected threshold will depend on various factors including species, strain, and clinical context. Nucleotide mutation rates can vary during different stages of an infection or may be under different selection pressures such as antimicrobial exposure (10). Laboratory processes during

104 culturing (e.g. single colony picks vs sweeps) may also affect the diversity of samples sent for
105 WGS (10). Various studies have used SNP thresholds ranging from 0 for *Yersinia* species to
106 over 35 for *Pseudomonas aeruginosa* and a recently published implementation study applied
107 SNP thresholds of < 16 for multi-drug resistant *Staphylococcus aureus* and < 26 for the other
108 species in the study including vancomycin-resistant *Enterococci*, extended spectrum beta-
109 lactamase (ESBL) producing *Klebsiella pneumoniae* and ESBL-producing *Escherichia coli*
110 (11, 12). Similar cutoffs were also found using a number of different methods such as Poisson
111 distributions (25 SNPs for *E. coli*), within patient maximum diversity (17 SNPs for *E. coli),*
112 with and without recombination SNP distance changes (20 SNPs for *Enterococcus faecium*)
113 and linear mixed models (13 core genome SNPs for methicillin-resistance *Staphylococcus*
114 *aureus*) (13-15). Due to the shortcomings of these hard cut-off approaches, more probabilistic
115 methods are being explored to consider variable mutation rates and incorporating other
116 epidemiological information (10).
117
118 Horizontal gene transfer can also affect apparent SNP level relatedness and masking of
119 prophage and recombination regions has been previously suggested (10, 16). However, there
120 is not yet a consensus on this approach. A recent systematic analysis for real-time genomics
121 based tracking of MDR bacteria in the healthcare environment found masking of prophages
122 had minimal effect while masking of recombination may lead to erroneous conclusions of
123 isolate relatedness (11).
124
125 Other analysis decisions that may impact results include the choice of reference genome. A
126 high quality closed reference genome that is closely related to the isolates of interest can reduce
127 the potential for mis-mapping and maximises the size of the core genome (11). Large diverse

datasets may also reduce the size of the core-genome resulting in fewer sites available for pairwise comparison (11).

Irrespective of the chosen thresholds, references or genome masking approaches, using SNP differences to identify transmission events relies on accurate variant calling. Numerous bioinformatics pipelines are available that implement different approaches for read mapping, variant calling and variant filtering with the aim of maximising the number of true positive SNP calls and minimising false positives and false negatives. BactSNP, Lyveset, Lyveset2, Nesoni, Snippy and SPANDx are all-on-one pipelines targeted at microbial genomics that perform mapping, variant calling, variant filtering as well as various down-stream analyses (table 1) (1, 9, 17-19).

Previous benchmarking studies have been conducted on some of these pipelines. However, there is currently no established 'gold-standard' approach to benchmarking, and this has resulted in benchmarking studies being performed on several different datasets with conflicting performance outcomes, making comparisons between these studies difficult (1, 2). The absence of an established methodology and gold-standard benchmarking approach has been highlighted as a key risk to wide-spread implementation of microbial WGS-based diagnostics and surveillance and may be slowing its adoption in routine public health (7, 16).

In this study, we describe a novel SNP calling pipeline, Snpdragon, which addresses observed limitations in existing methodologies (available at https://github.com/FordeGenomics/SNPdragon). Leveraging datasets previously used to benchmark various microbial SNP calling applications we systematically compare performance of Snpdragon and six all-in-one pipelines BactSNP, Lyveset, Lyveset2, Nesoni,

153 SPANDx and Snippy (1, 9, 17-19). Finally, we highlight the issues surrounding current
154 benchmarking approaches and propose a number of solutions which will become increasingly
155 critical as this technology is integrated into clinical practice.

Methods

Snpdragon

Snpdragon is a SNP calling pipeline implemented in Nextflow and available to be deployed in Docker and Singularity containers (20, 21). It uses BWA-mem for read mapping, Samtools for coverage and Freebayes for variant calling (22-24). Post-filtering of variant calls is performed in a Python program to report high confidence SNPs. Standard SNP filters are applied with setting comparisons to the other all-in-one pipelines detailed in table 1. SNPs not passing filter thresholds are labelled in the output variant call format (vcf) files:

- FAIL_AF: Alternate allele fraction (alternate count/depth) >= 0.5 and < 0.75
- FAIL_AF0.5: Alternate allele fraction < 0.5
- FAIL_DEPTH: Read depth at position < 10
- FAIL_MQM: Mean mapping quality at position < 30
- FAIL_RB: Read balance/strand bias fails if the ratio of alternate alleles on the forward and reverse strands is < 0.05.

Presence/absence matrices and alignment files are generated using the high confidence SNP positions and populated based on all unfiltered SNPs detected in each sample. Optional additional filters include excluding SNPs detected in cliffs. A cliff is classified as a region with a rapid change in aligned read depth and may be the result of sequence anomalies, poor read mapping, repeat regions and breakpoints at positions of large structural rearrangements. The algorithm for the detection of cliffs has been implemented as described in Katz et al. (9). Briefly, a linear trend line for read coverage in window of 10bp is calculated and a region is masked if the slope of the line is >= 3 or <= 3 and the fit of the line (R^2) is >= 0.7. SNPs occurring in high density (which may be the result of mis-mapping or recombination) can also be filtered (25). A sliding window approach is implemented to optionally exclude SNPs occurring at a frequency of 3 or more in a 10bp window.

To optimise memory usage and runtime of the Python program an integer representation of the IUPAC alphabet was developed. Float data types are then used to represent 'SNP addresses' which are a combination of a position and the allele. For example, 1.1 represents position 1 with a base call A. The use of Nextflow also allows for the rapid analysis of very large datasets that may have incremental additions as more isolates are added to a collection, a scenario common to the application of WGS in pathogen surveillance in public health.

Snpdragon produces the following final output files:

- core_snp.fasta – Core SNP multiple sequence alignment (MSA)
- full_snp.fasta – Full SNP MSA including accessory genome (missing positions in each sample denoted with 'N')
- full_aln.fasta – Mutated reference pseudo-genome MSA
- snp_dist.csv – Pairwise SNP distance matrix
- snp_matrix.csv – SNP position matrix (SNP sites by samples)
- core_stats.csv – Number of positions and percent of reference genome coverage for each sample

Intermediate files including all bams, pileups and raw and filtered vcf's are also output but can be optionally cleaned at each step if storage space is a limitation in large analyses.

Benchmarking datasets

Previously published benchmarking datasets are combined to systematically compare the performance of Snpdragon, BactSNP v1.1.0, Lyveset v1.1.4g, Lyveset2 v2.0.1, Nesoni v0.132, SPANDx v4.0.2 and Snippy v4.6.0 (1, 9, 17-19).

EC958

The EC958 dataset consists of three previously isolates of the *E. coli* ST131 strain EC958 (26, 27). These three isolates (EC958, MS6573 and MS6574) are nearly identical with EC958 differing from MS6573 and MS6574 by a single SNP and MS6573 and MS6574 identical. These were mapped to references of decreasing similarity calculated using fastANI which are detailed in table 2 (28).

Yoshimura

The Yoshimura dataset consists of 12 simulated experiments each with 10 samples representing *E. coli, Neiserria meningitidis* and *S. aureus* aligned to increasingly distant reference genomes from 99.9% identity to 97% identity previously described in Yoshimura et al. (1).

Bush-simulated

The Bush-simulated dataset is a collection of 251 isolates from 10 species with SNPs simulated as described in Bush et al. (2). Results from the benchmarking of the six all-in-one pipelines in this study were also combined with expanded benchmarking results from Bush et al. on the 150bp simulated reads (2).

Bush-real

The Bush-real dataset consists of 18 publicly available sequencing experiments. Methods for generating this dataset are described in Bush et al. (2). The ground truth for comparison was previously generated using an intersect of SNP calls using ParSNP and Nucmer (29, 30). SNP calls made by only one of these tools were classified as ambiguous and excluded from benchmarking calculations.

Compute environment

BactSNP, Lyveset, Lyveset2, Nesoni, SPANDx and Snippy were run using default settings with 16 cores and 128G of available RAM. Snpdragon was run using default settings on EC958 and Yoshimura datasets. Three separate results for Snpdragon were generated on the Bush-simulated and Bush-real dataset to benchmark the effect of: 1. excluding SNPs occurring in cliffs and clusters; 2. excluding only SNPs called in cliffs; and 3. including all SNPs irrespective of cliffs and clusters.

Concordance and performance metrics

Pipelines were assessed based on concordance to a 'ground truth' set. True positive (TP), false positive (FP) and false negative (FN) counts were reported and used to calculate recall, precision and F_1 score.

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Recall is a measure of how well all actual (true) positives are captured (at the cost of higher false positives) while precision or the positive predictive value is a measure of how well only true positives are captured (at the cost of false negatives). The F_1 score is the harmonic mean of precision and recall with poorest performance at 0 and the highest score of 1 and is suited to situations where there is a high rate of true negatives, and which are not a relevant measure (i.e. non-variant positions) (31). Pairwise core SNP distance matrices were calculated using snp-dist (32). Run-time and memory usage based on resident set size (RSS) are also reported.

It should be noted that different analyses tools may represent the same variants in different ways. 'Complex' variants and multi-nucleotide polymorphisms (MNPs) are output by some variant calling

257 tools including Freebayes. These can be regularised in the VCF file using vcfallelicprimitives module
258 in vcflib (33). In any case, the pipelines benchmarked in this study reported primitive SNP
259 representations and no additional filtering of the variant calls was performed.
260

Results

Concordance benchmarks

EC958

All pipelines recalled the single true SNP difference between EC958 and MS6573 and MS6574 irrespective of the reference strain. Snpdragon and BactSNP showed high precision reporting only the single known SNP and no false positives irrespective of the reference genome. The other tested pipelines however had poorer performance with more distant reference genomes. Lyveset, Lyveset2, Nesoni, SPANDx and Snippy showed lower F_1 scores due to false positives being reported when using more distant reference genomes (figure 2). Increasing numbers of pairwise SNP differences due to these false positives is shown in supplementary table S1.

Yoshimura

Snpdragon and SPANDx had the highest median combined F_1 score (figure 3 and supplementary table S2**)**. Performance scores declined with increasingly distant reference genomes for all pipelines, though Snippy was most impacted (figure 3). The decline in performance on more distant reference genomes was generally due to a decline in recall (related to increasing numbers of false negative SNP calls) for BactSNP, Lyveset, Lyveset2, Nesoni, Snpdragon and SPANDx (figure 3A). Snippy however showed a decline in precision as a result of higher rates of false positive SNP calls. On the most distant reference genomes tested (97% similarity), recall scores for Lyveset and Lyveset2 were below 0.4 possibly due to too stringent filtering causing higher false negative counts.

Bush-simulated

Combined median F_1 scores were highest for BactSNP and Snpdragon with optional settings to discard SNPs in cliffs followed by Snpdragon with no additional optional filtering (figure 4). Lyveset, Lyveset2 and Snpdragon (with additional filtering to exclude SNPs in cliffs and clusters) showed

poorest performance on this dataset. This largely appears to be related to a decline in recall which was particularly evident on the *Listeria* samples.

In the combined results on with the expanded 150bp simulated dataset from Bush et al. BactSNP showed highest performance based on median F_1 score following by Snpdragon (with settings to exclude only SNPs occurring in cliffs) (figure 5). The Snippy results from Bush et al. could not be replicated, with the results in the paper scoring slightly higher than ours. Nesoni performed similarly to Snippy on this dataset.

Bush-real

BactSNP, Snpdragon (with no additional filtering and with filtering to exclude SNPs occurring in cliffs) and Snippy performed similarly on the Bush-real dataset (figure 6)**.** Lyveset and Lyveset2 showed poorest performance with the lowest median F_1 scores. The decline in performance was generally related to poorer recall, particularly on more distant reference genomes (figure 6A). One sample (rbhstw00167) also scored very poorly on precision in every pipeline.

Computational benchmarks

Snippy had the fastest median runtime on all datasets (figure 7A, figure 8A and figure 9A). Snpdragon was also one of the most rapid on the EC958 and Yoshimura datasets. Runtime for Snpdragon on the Bush-real dataset was mostly affected by whether the additional SNP cluster and cliff finding were used (figure 9A). SPANDx and Lyveset had the highest median runtimes. SPANDx also required the most amount of memory followed by Lyveset2 while the other pipelines tested had lower and generally similar memory requirements across each of the datasets (figure 7B, figure 8B and figure 9B).

Discussion

A newly introduced pipeline Snpdragon and six additional all-in-one pipelines BactSNP, Lyveset, Lyveset2, Nesoni, SPANDx and Snippy were systematically evaluated for performance using a combination of new and previously published benchmarking datasets. Only all-in-one pipelines were included due to the popularity of such pipelines for their ease of use and internal filtering designed to improve accuracy of the reported variant calls. These pipelines were benchmarked not only to evaluate performance but to also explore potential issues in the current benchmarking approaches. The current lack of guidelines for evaluating microbial variant calling pipelines has resulted in diverse and inconsistent approaches in the literature (34). To establish a gold-standard benchmarking approach, real datasets with verified known variants are required (35, 36). For the development of high-quality benchmarking datasets, the following criteria has been proposed (16, 34, 35, 37):

- Relevance: Does the dataset contain the characteristics (variants) of interest
- Representativeness: Does the dataset cover the breadth of possible sample types and features in the study space to establish the stability of the analysis approach
- Non-redundancy: Exclude overlaps and duplications
- Experimentally verified cases: The ground truth is known
- Positive and negative cases: The characteristic under investigation is present and absent in different samples
- Scalability: For testing performance on different dataset sizes
- Reusability: For reproducibility and data sharing

To accurately assess performance of bioinformatics pipelines on any dataset, the ground truth is needed (37). Simulated datasets are attractive for this reason, where features of interest (e.g. SNPs) are introduced in-silico at known positions. However, simulated datasets may not always be representative and may not model all features or potential sources of errors present in real data. Alternatively, using real datasets in benchmarking is problematic as the ground truth is often unknown and instead comparisons are performed against the results of existing methods (35).

The datasets used in this study consisted of a mix of simulated and real data with different characteristics. The EC958 dataset consisted of sequencing data from three almost identical *E. coli* ST131 isolates with a known single SNP difference that had been previously well characterised (26, 27). The Yoshimura dataset was a simulated dataset of 10 samples from three different species with SNPs introduced *in-silico* at known locations and represented both gram-negative and gram-positive bacteria (1). The Bush-simulated and Bush-real datasets were a diverse collection of publicly available isolates and matching closed reference genomes. In the simulated dataset, SNPs were introduced in-silico resulting in ~8000-25000 SNPs per genome with a median distance of ~60-120 bases between SNPs as described previously (2). This represents a much higher SNP rate than the other datasets which were designed to reflect more closely related isolates in an outbreak or transmission event setting. Similarly, the Bush-real dataset consisted of samples with matched closed reference genomes of 87.7% to 99.1% identity with ~8000-13000 SNPs between the sample and the matched reference (2).

For the Bush-simulated and Bush-real datasets, the ground-truth was established by taking an intersection of the results of two assembly-based methods ParSNP and Nucmer (29, 30). While this may be a reasonable approach given the limitations of establishing the known truth for the real datasets, the risk is that the process of benchmarking may become an exercise in concordance with existing methods rather than reflecting true accuracy (35). Using the union of calls may not necessarily reflect true calls if both methods were susceptible to the same biases (34). Additionally, sites were labelled as ambiguous and excluded from benchmarking counts if only one of ParSNP or Nucmer reported a SNP and this may result in under-estimation of false positive rates (38).

This work highlights the difficulties when attempting to interpret different benchmarking studies where the performance of one pipeline on one dataset is not replicated on other datasets and therefore

results may not be generalisable. As has been previously demonstrated, accuracy declined with more distant reference genome, however, the results show some pipelines were more affected than others (39). For example, on EC958 lower F_1 scores were observed for all pipelines except Snpdragon and BactSNP on increasingly distant reference genomes (figure 2). Poorer performance with the other pipelines on this dataset was related to higher rates of false positive SNP calls. The clinical implications of these false positives can be seen in the pairwise core SNP difference matrices (supplementary table S1). In some cases, the number of SNPs reported between these almost identical samples was above the threshold typically used to define isolates as part of a cluster (11). On the Yoshimura dataset, Snippy was the most affected, followed by Lyveset and Lyveset2 by the dis-similarity of the reference genome, but for different reasons. While the precision of Snippy declined due to increasing numbers of false positive SNPs, the recall of Lyveset and Lyveset2 declined due to higher false negative counts (figure 3A). The results on the Bush-simulated and Bush-real datasets however showed the precision of Snippy was less affected by distance to the reference genome (but instead a showed a proportionate decline in recall) (figure 6A). Overall, Snpdragon and BactSNP showed the most stable performance across all datasets and reference types.

The poorer recall across all datasets for Lyveset and Lyveset2 may be related to stricter internal SNP filtering resulting in higher numbers of 'real' SNPs being discarded. Similarly, with the additional filters to exclude SNPs in cliffs and clusters in Snpdragon, a similar decline in recall was observed but only on the Bush-simulated and Bush-real datasets highlighting the difficulties in generalising single benchmarking results across different datasets (figure 4A and 6A).

These results also demonstrated the usefulness of using a variety of benchmarking metrics for comparison. While the F_1 score is useful to report a balance between recall and precision, reporting separate measures provides insight into the underlying causes of the poorer performance (e.g. high false negatives vs high false positives) which varied between pipelines and across datasets.

389

390 The lack of a standardised approach to benchmarking may be slowing implementation of microbial
391 WGS in clinical practice. A criteria for development benchmarking datasets has been proposed by
392 Sarkar et al. and the Global Microbial Identifier (GMI) working group are in ongoing development
393 of an SOP for the validation of benchmarking datasets (35, 40). While simulated datasets are useful,
394 they may not fully represent all characteristics present on real sequencing data that can be potential
395 sources of error and bias. Therefore, building experimentally validated benchmarking datasets such
396 as through Sanger sequencing will be important to generate known ground truths as was done in a
397 recent study comparing several pipelines on *Mycobacterium tuberculosis* (41).

398

399 **Conclusion**

400 This study sought to survey the current landscape of prominent benchmarking studies for the analysis
401 of microbial SNP calling and to comprehensively evaluate a range of all-in-one pipelines. The results
402 highlight the difficulty in comparing results between different benchmarking approaches and the
403 effect of dataset choice. The growing interest in the routine application of microbial WGS for AMR
404 surveillance, outbreak investigation and diagnostics should motivate the development of a gold-
405 standard benchmarking approach.

406

References

1. Yoshimura D, Kajitani R, Gotoh Y, Katahira K, Okuno M, Ogura Y, et al. Evaluation of SNP calling methods for closely related bacterial isolates and a novel high-accuracy pipeline: BactSNP. Microb Genom. 2019;5(5).
2. Bush SJ, Foster D, Eyre DW, Clark EL, De Maio N, Shaw LP, et al. Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism–calling pipelines. GigaScience. 2020;9(2).
3. Black A, MacCannell DR, Sibley TR, Bedford T. Ten recommendations for supporting open pathogen genomic analysis in public health. Nature Medicine. 2020;26(6):832-41.
4. Van Goethem N, Descamps T, Devleesschauwer B, Roosens NHC, Boon NAM, Van Oyen H, et al. Status and potential of bacterial genomics for public health practice: a scoping review. Implementation Science. 2019;14(1):79.
5. 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;11(2):e041968.
6. Lee XJ, Elliott TM, Harris PNA, Douglas J, Henderson B, Watson C, et al. Clinical and Economic Outcomes of Genome Sequencing Availability on Containing a Hospital Outbreak of Resistant Escherichia coli in Australia. Value in Health. 2020;23(8):994-1002.
7. National Microbial Genomics Framework 2019-2022. 2019.
8. Roberts LW, Catchpoole E, Jennison AV, Bergh H, Hume A, Heney C, et al. Genomic analysis of carbapenemase-producing Enterobacteriaceae in Queensland reveals widespread transmission of bla (IMP-4) on an IncHI2 plasmid. Microb Genom. 2020;6(1).
9. Katz LS, Griswold T, Williams-Newkirk AJ, Wagner D, Petkau A, Sieffert C, et al. A Comparative Analysis of the Lyve-SET Phylogenomics Pipeline for Genomic Epidemiology of Foodborne Pathogens. Frontiers in Microbiology. 2017;8(375).
10. Stimson J, Gardy J, Mathema B, Crudu V, Cohen T, Colijn C. Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions. Molecular Biology and Evolution. 2019;36(3):587-603.
11. Gorrie CL, Da Silva AG, Ingle DJ, Higgs C, Seemann T, Stinear TP, et al. Key parameters for genomics-based real-time detection and tracking of multidrug-resistant bacteria: a systematic analysis. The Lancet Microbe. 2021.
12. Schürch AC, Arredondo-Alonso S, Willems RJL, Goering RV. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene-based approaches. Clin Microbiol Infect. 2018;24(4):350-4.

13. Ludden C, Coll F, Gouliouris T, Restif O, Blane B, Blackwell GA, et al. Defining nosocomial transmission of Escherichia coli and antimicrobial resistance genes: a genomic surveillance study. Lancet Microbe. 2021;2(9):e472-e80.
14. Gouliouris T, Coll F, Ludden C, Blane B, Raven KE, Naydenova P, et al. Quantifying acquisition and transmission of Enterococcus faecium using genomic surveillance. Nat Microbiol. 2021;6(1):103-11.
15. Coll F, Raven KE, Knight GM, Blane B, Harrison EM, Leek D, et al. Definition of a genetic relatedness cutoff to exclude recent transmission of meticillin-resistant Staphylococcus aureus: a genomic epidemiology analysis. Lancet Microbe. 2020;1(8):e328-e35.
16. Saltykova A, Mattheus W, Bertrand S, Roosens NHC, Marchal K, De Keersmaecker SCJ. Detailed Evaluation of Data Analysis Tools for Subtyping of Bacterial Isolates Based on Whole Genome Sequencing: Neisseria meningitidis as a Proof of Concept. Frontiers in Microbiology. 2019;10(2897).
17. T S. Snippy: Fast bacterial variant calling from NGS reads. 2015.
18. Sarovich DS, Price EP. SPANDx: a genomics pipeline for comparative analysis of large haploid whole genome re-sequencing datasets. BMC Research Notes. 2014;7(1):618.
19. Victorian-Bioinformatics-Consortium. Nesoni. 2013.
20. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nature Biotechnology. 2017;35(4):316-9.
21. Merkel D. Docker: lightweight Linux containers for consistent development and deployment. Linux Journal. 2014;2014.
22. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
23. Garrison E MG. Haplotype-based variant detection from short-read sequencing. arXiv preprint. 2012;arXiv:1207.3907 [q-bio.GN].
24. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
25. Bush SJ. Generalizable characteristics of false-positive bacterial variant calls. Microbial Genomics. 2021;7(8).
26. Forde BM, Ben Zakour NL, Stanton-Cook M, Phan M-D, Totsika M, Peters KM, et al. The complete genome sequence of Escherichia coli EC958: a high quality reference sequence for the globally disseminated multidrug resistant E. coli O25b:H4-ST131 clone. PLoS One. 2014;9(8):e104400-e.
27. Phan MD, Nhu NTK, Achard MES, Forde BM, Hong KW, Chong TM, et al. Modifications in the pmrB gene are the primary mechanism for the development of chromosomally encoded

resistance to polymyxins in uropathogenic Escherichia coli. J Antimicrob Chemother. 2017;72(10):2729-36.

28. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nature Communications. 2018;9(1):5114.

29. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biology. 2014;15(11):524.

30. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. PLOS Computational Biology. 2018;14(1):e1005944.

31. Takahashi K, Yamamoto K, Kuchiba A, Koyama T. Confidence interval for micro-averaged F1 and macro-averaged F1 scores. Applied Intelligence. 2021.

32. Seeman T KF, Page A. snp-dists.

33. Garrison E, Kronenberg ZN, Dawson ET, Pedersen BS, Prins P. Vcflib and tools for processing the VCF variant call format. bioRxiv. 2021:2021.05.21.445151.

34. Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, et al. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. Front Genet. 2015;6:235-.

35. Sarkar A, Yang Y, Vihinen M. Variation benchmark datasets: update, criteria, quality and applications. Database. 2020;2020.

36. Zhao S, Agafonov O, Azab A, Stokowy T, Hovig E. Accuracy and efficiency of germline variant calling pipelines for human genome data. Scientific Reports. 2020;10(1):20222.

37. Krishnan V, Utiramerur S, Ng Z, Datta S, Snyder MP, Ashley EA. Benchmarking workflows to assess performance and suitability of germline variant calling pipelines in clinical diagnostic assays. BMC Bioinformatics. 2021;22(1):85.

38. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nature Biotechnology. 2014;32(3):246-51.

39. Gorrie CL, Da Silva AG, Ingle DJ, Higgs C, Seemann T, Stinear TP, et al. Key parameters for genomics-based real-time detection and tracking of multidrug-resistant bacteria: a systematic analysis. The Lancet Microbe. 2021;2(11):e575-e83.

40. 12th Global Microbial Identifier Initiative Meeting Report. Global Microbial Identifier; 2019 June 2019.

511 41. Walter KS, Colijn C, Cohen T, Mathema B, Liu Q, Bowers J, et al. Genomic variant-
512 identification methods may alter Mycobacterium tuberculosis transmission inferences. Microb
513 Genom. 2020;6(8).
514 42. Andrews S. FastQC: a quality control tool for high throughput sequence data. Available
515 online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc. 2010.
516

Tables

Table 1. Benchmarked all-in-one variant calling pipelines targeted to analysis of microbial genomic datasets.

Pipeline	Version tested	Release date	Aligner	SNP caller	Default optional caller settings	Additional features	Link	Ref
BactSNP	1.1.0	2018	BWA-mem	Samtools	AF=0.9, Depth=10	Creates assemblies and maps reads back to pseudogenome	https://github.com/IEkAdN/BactSNP	(1)
LyveSet	1.1.4g	2017	SMALT	Varscan	AF=0.75, Depth=10	Optional cliff masking, optional phage masking	https://github.com/lskatz/lyve-SET	(9)
LyveSet2	2.0.1	2018	SMALT	Varscan	AF=0.75, Depth=10	Optional cliff masking, optional phage masking	https://github.com/lskatz/lyve-SET	(9)
Nesoni	0.132	2015	Bowtie2	Freebayes	pvar=0.9		https://github.com/Victorian-Bioinformatics-Consortium/nesoni	(19)
Snippy	4.6.0	2020	BWA-mem	Freebayes	Depth=10, AF=0.9, QUAL=100		https://github.com/tseemann/snippy	(17)
Snpdragon	1.0.0	2022	BWA-mem	Freebayes	MAPQ=10/30, BASEQ=10/10, AF=0.1/0.75, Depth=10/10, strand_balance=0/0.05	Option cliff masking, optional SNP cluster filtering	https://github.com/FordeGenomics/SNPdragon	
SPANDx	4.0.2	2021	BWA-mem	GATK	QualByDepth=10, RMSMAPQ=30, QUAL=30, FS=60	Calls indels, optional SNP cluster filtering	https://github.com/dsarov/SPANDx	(18)

520 Table 2. Reference genomes used in the EC958 benchmarking dataset and the percent identity against

521 the three included samples.

Reference name	Identity (%)	Accession
EC958	100	NZ_HG941718.1
ECJJ1886	99.9	CP006784.1
SE15	99.5	AP009378.1
UTI89	98.3	CP000243.1
IAI39	97.2	CU928164.2
E. coli K12	96.8	U00096.3
SE11	96.7	AP009240.1
Sakai	96.5	BA000007.3

522

Figures

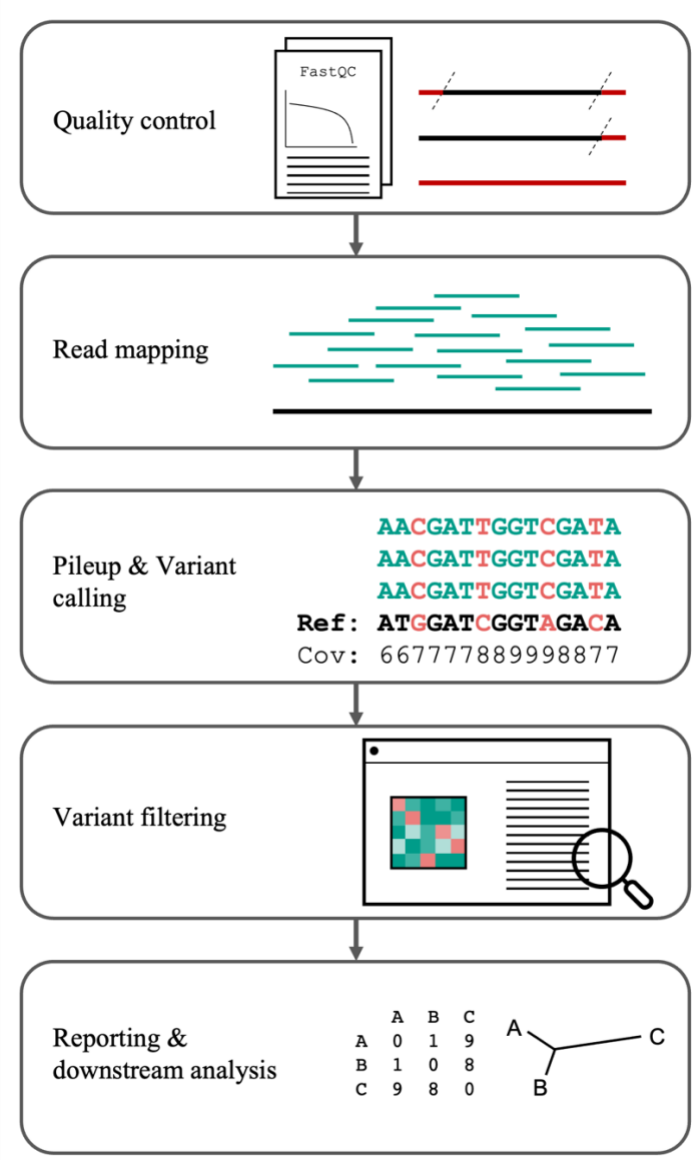


Figure 1. A typical variant calling bioinformatics pipeline. Quality control is measured using FastQC and reads may be trimmed of poor-quality bases (42). Reads are mapped to a chosen reference genome followed by variant calling. Coverage or pileup calculations may also be performed to determine the depth which is the number of reads covering each base in the reference genome. Variant filtering is applied to discard low confidence variant calls based on various measures such as depth, base quality, mapping quality, ratio of the variant to the reference allele (ratio of support/allele fraction) and read bias (only forward or reverse reads reporting a variant). Results are reported in a human readable format in addition to files suitable for other downstream analyses.

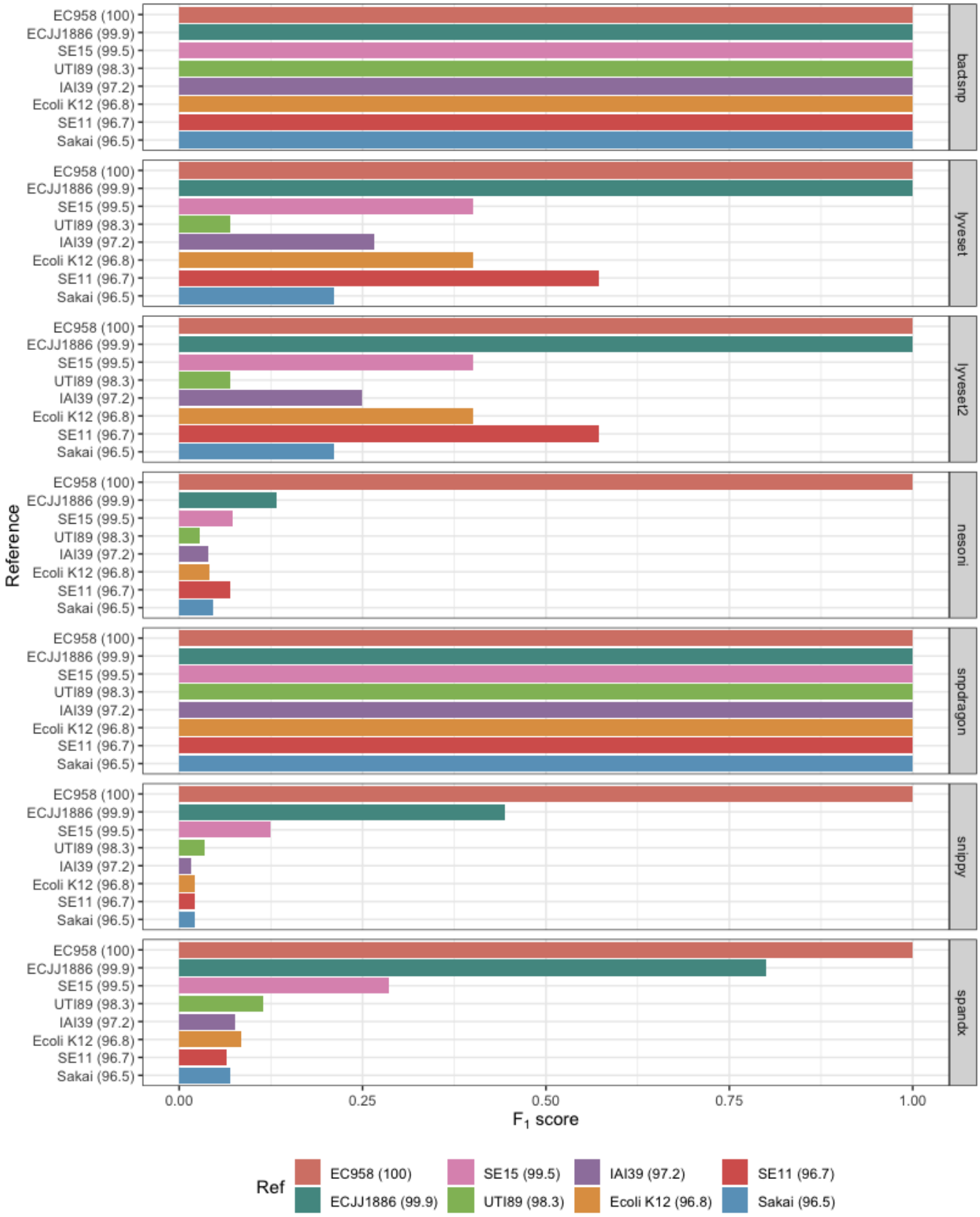


Figure 2. F_1 score for BactSNP, Lyveset, Lyveset2, Nesoni, Snpdragon, Snippy and SPANDx on the EC958 dataset against increasingly distant reference genomes.

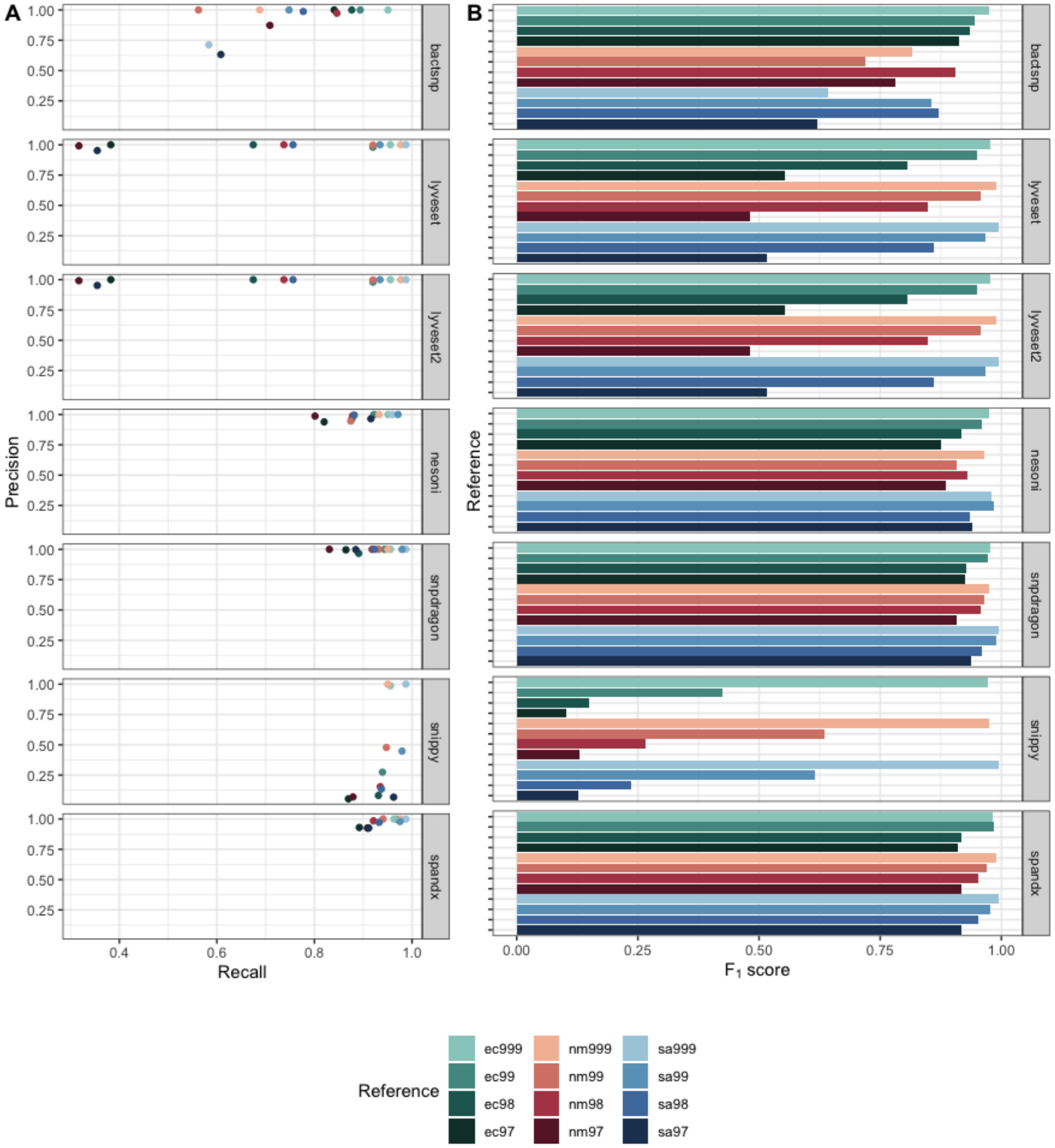


Figure 3. A) Precision vs Recall scatter plot and B) F_1 score on the Yoshimura dataset for each of the pipelines against increasingly distance reference genomes from 99.9% similarity to 97% similarity (1).

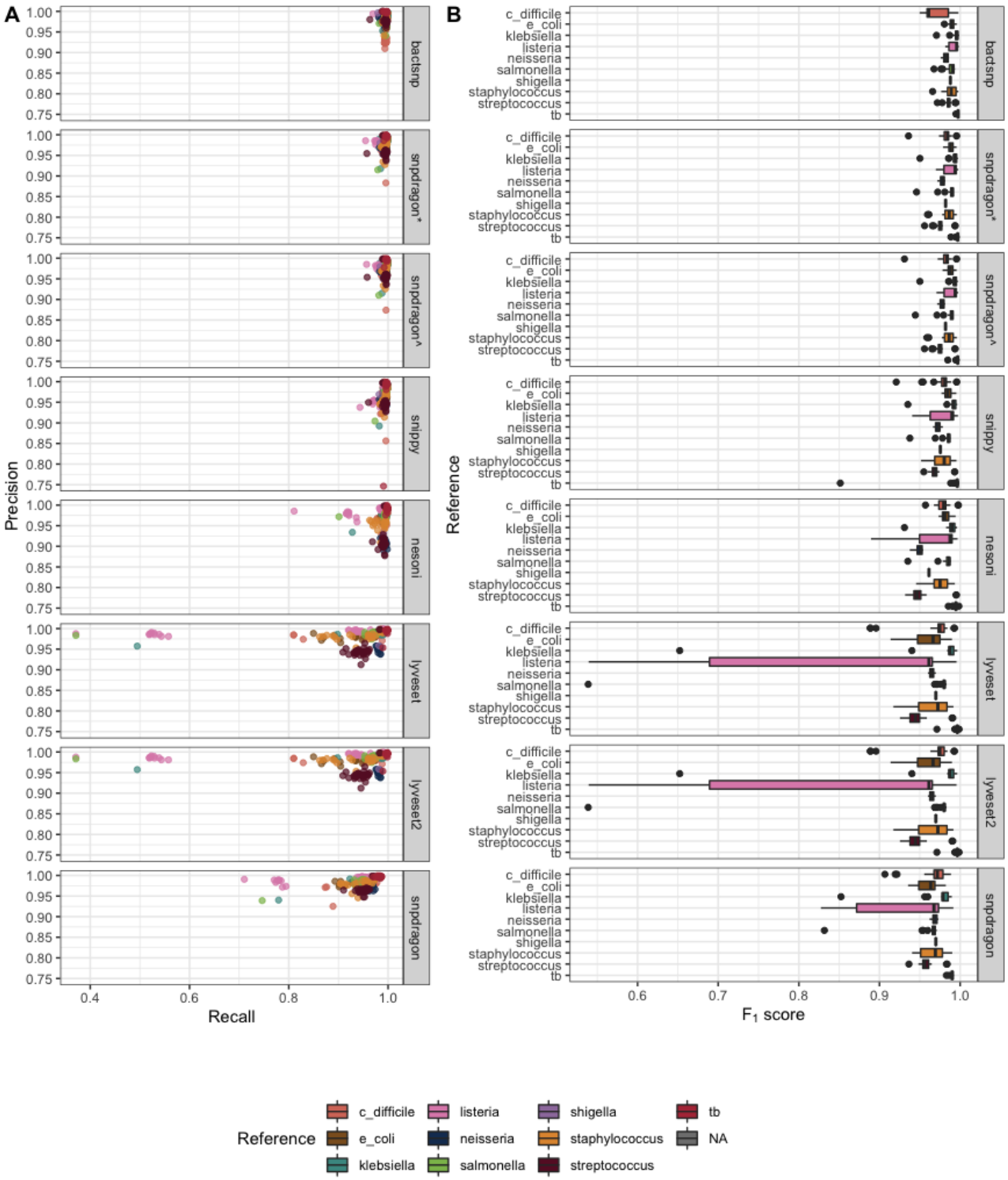
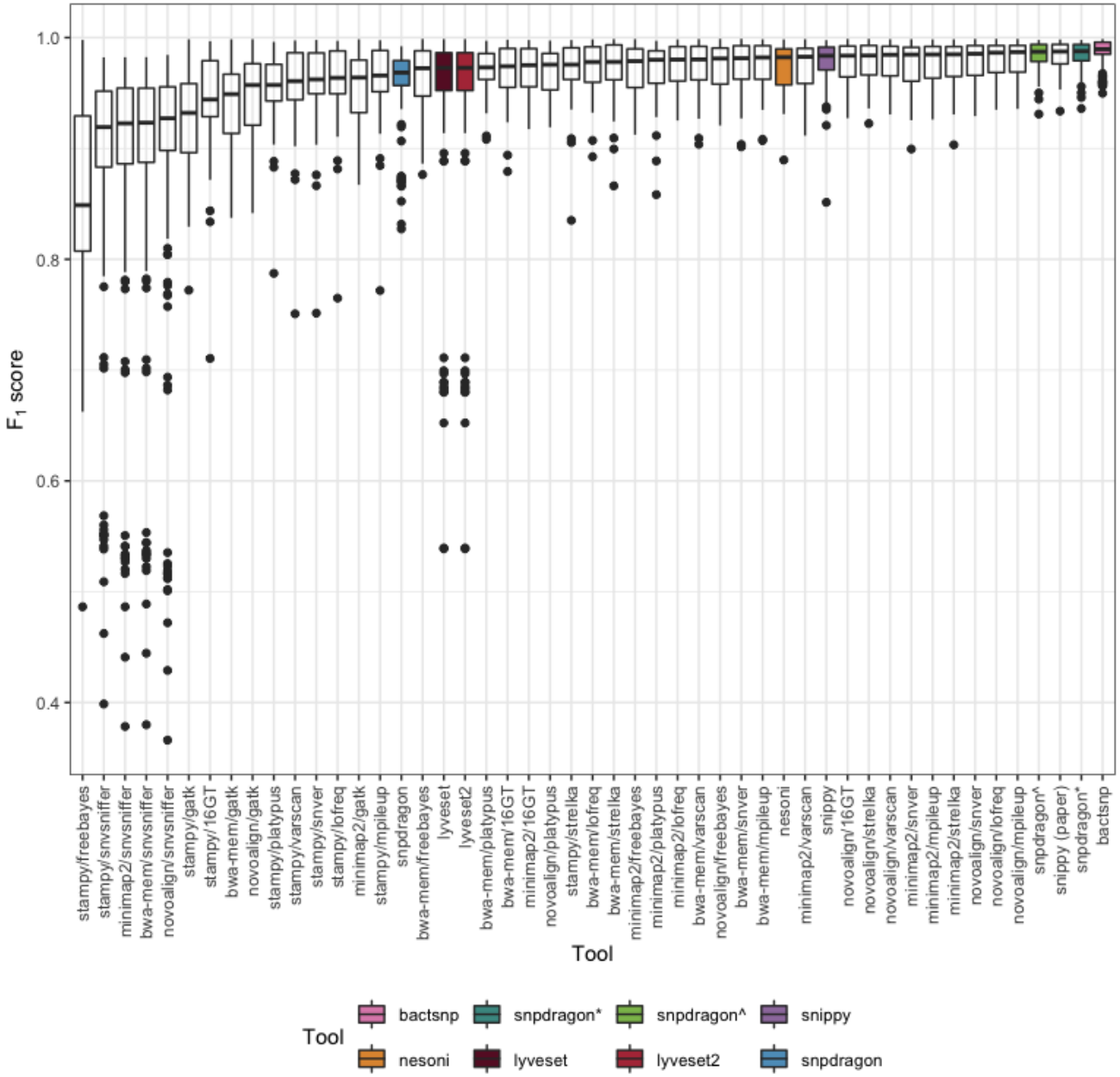


Figure 4. A) Precision vs Recall scatter plot and B) F_1 score boxplot on the Bush-simulated dataset ordered based on median combined F_1 scores (2). Snpdragon = filtering to exclude both SNPs occurring in cliffs and in high density SNP clusters. Snpdragon* = optional filtering settings to exclude SNPs occurring in cliffs. Snpdragon^ = no additional optional filtering settings.



549

550 Figure 5. F_1 scores on combined results of BactSNP, Lyveset, Lyveset2, Nesoni, Snpdragon,

551 Snippy and Bush-et al. supplementary results on the 150bp simulated data (2). Results for the new

552 pipelines analysed in this study are highlighted. Snpdragon = filtering to exclude both SNPs

553 occurring in cliffs and in high density SNP clusters. Snpdragon* = optional filtering settings to

554 exclude SNPs occurring in cliffs. Snpdragon^ = no additional optional filtering settings.

555

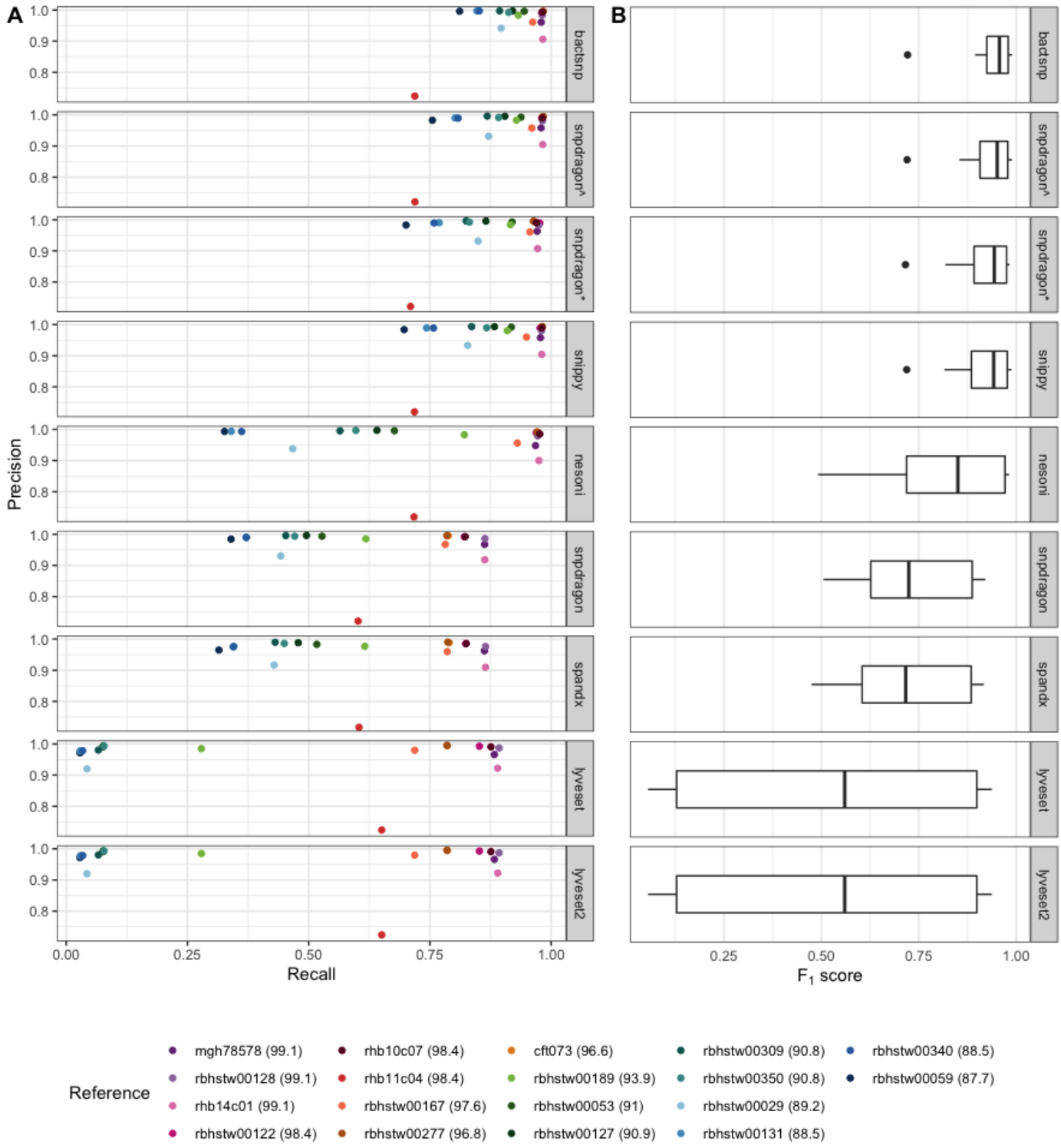


Figure 6. A) Precision vs Recall scatter plot and B) Boxplot of F1 scores on Bush-real dataset ordered by median F_1 score (2). Snpdragon = filtering to exclude both SNPs occurring in cliffs and in high density SNP clusters. Snpdragon* = optional filtering settings to exclude SNPs occurring in cliffs. Snpdragon^ = no additional optional filtering settings.

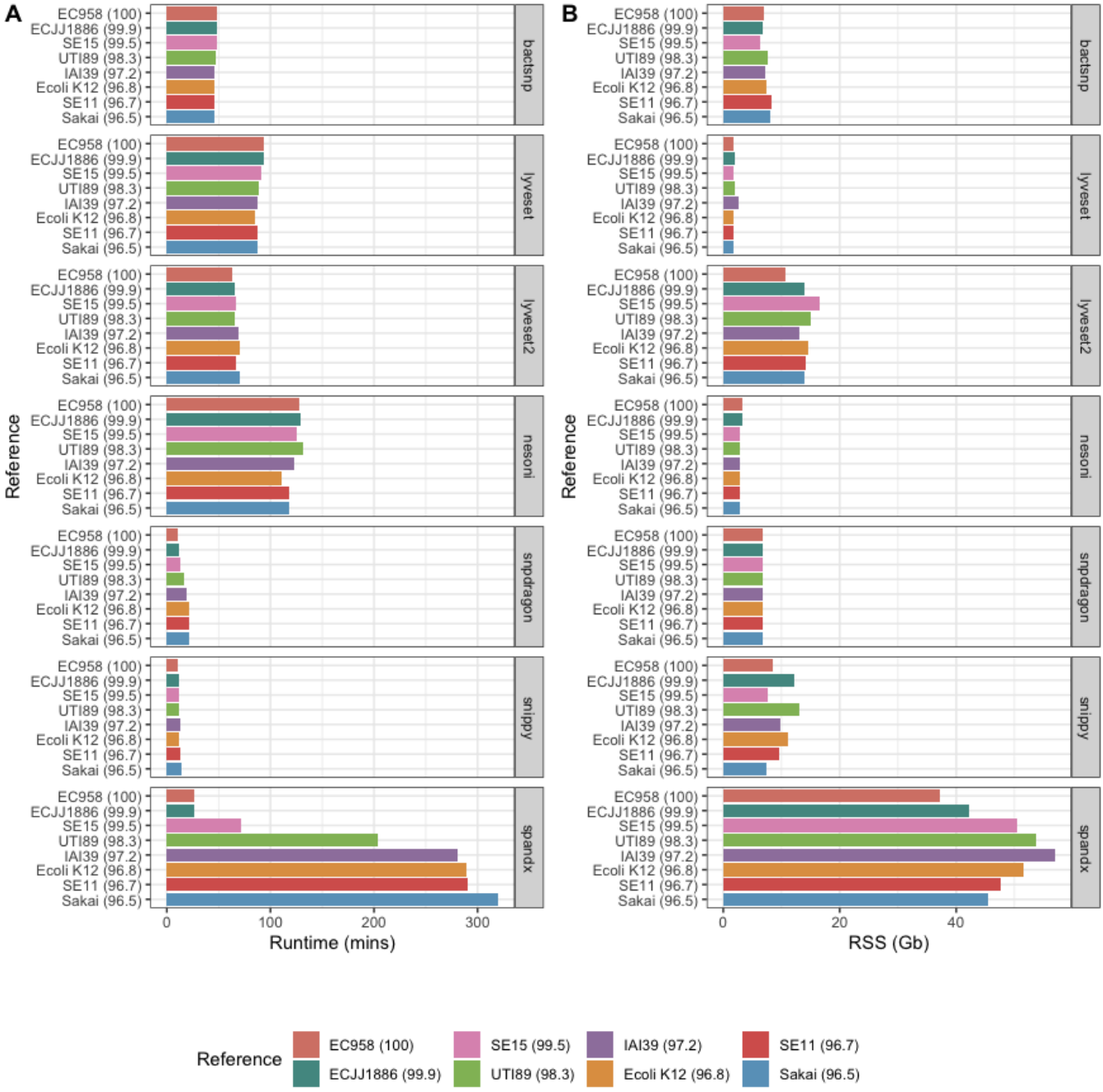


Figure 7. A) Runtime on E. coli ST131 dataset. B) Resident set size (RSS) on EC958 dataset.

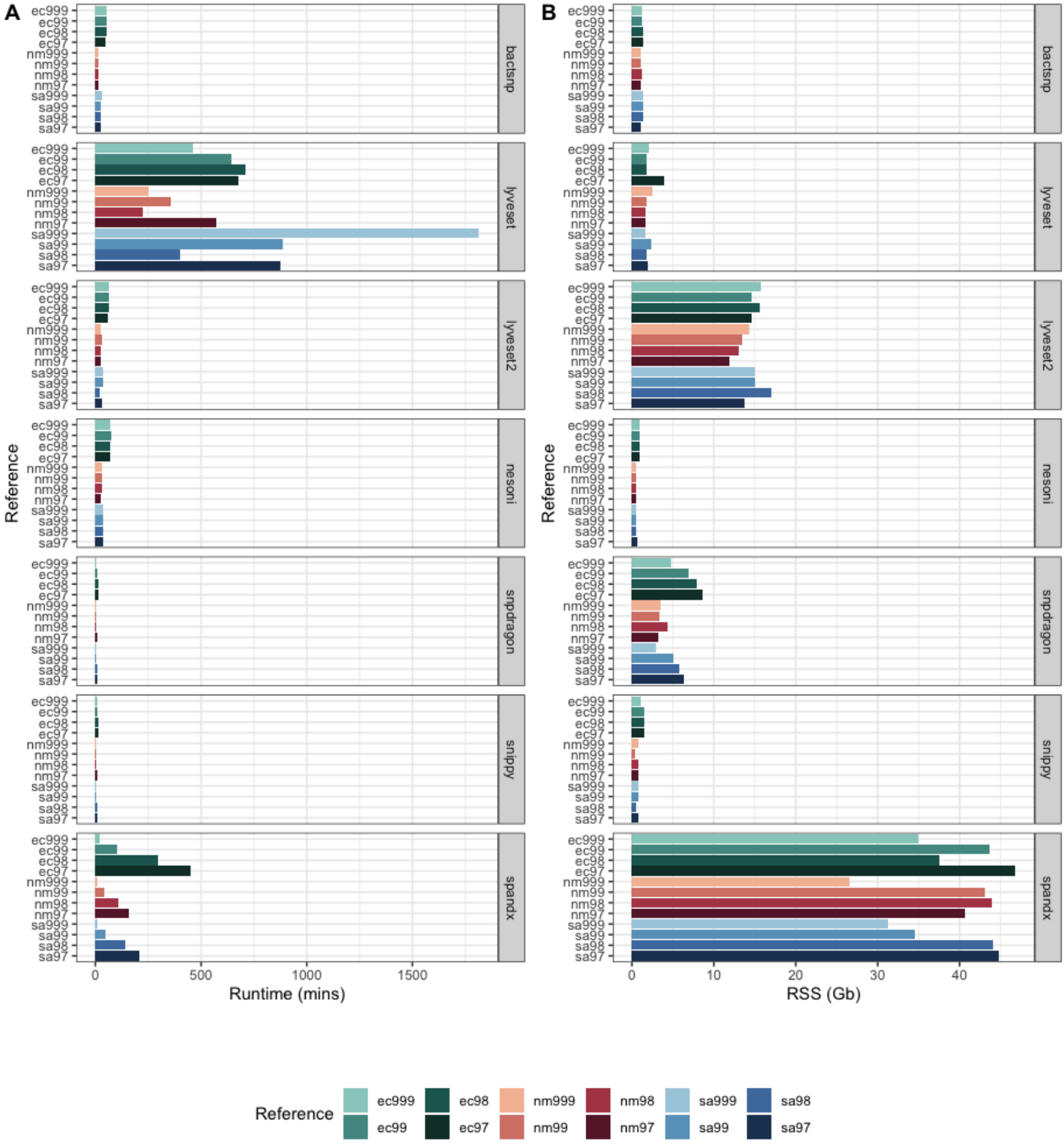


Figure 8. A) Runtime and B) RSS on the Yoshimura dataset (1).

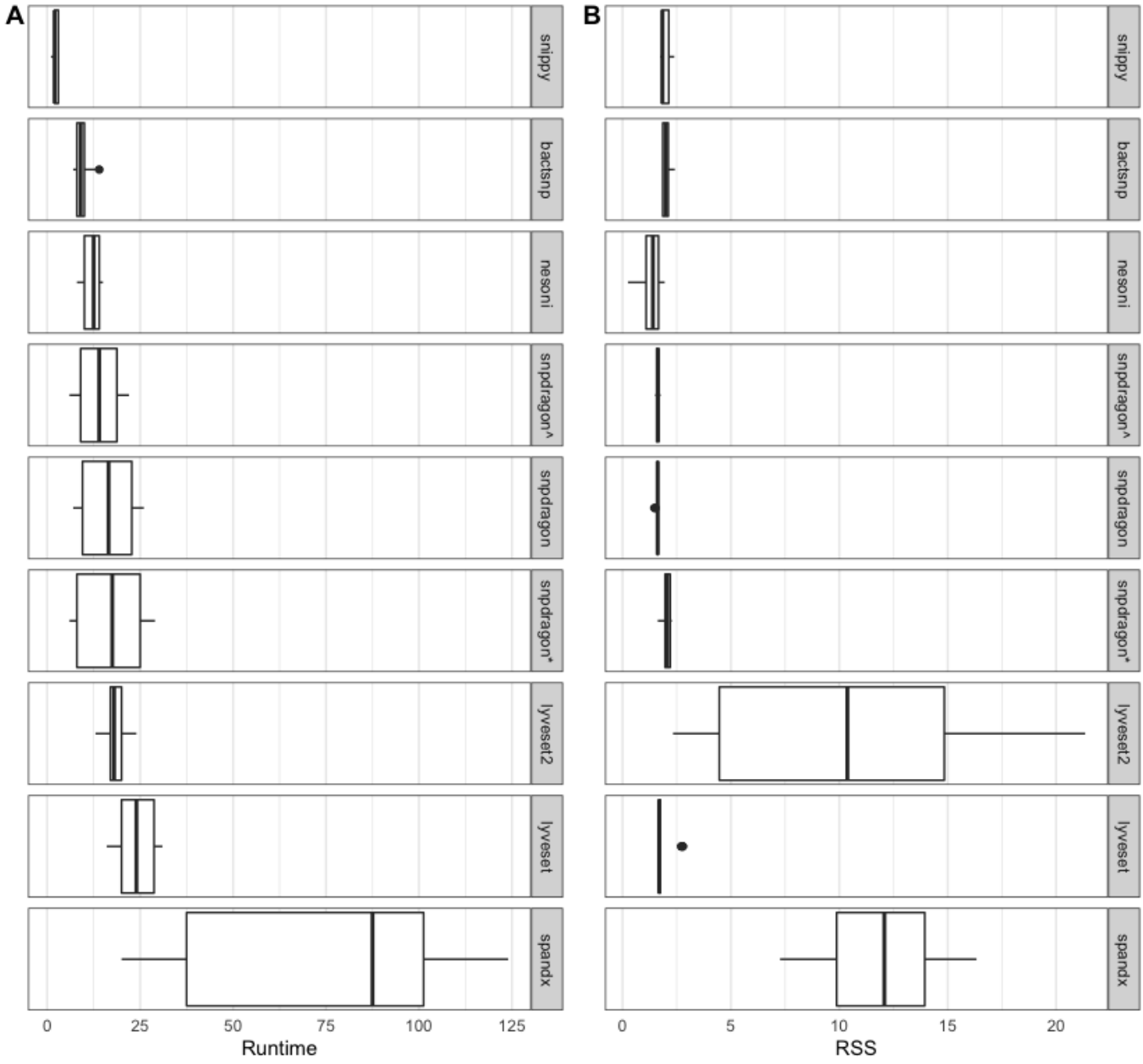


Figure 9. A) Runtime and B) RSS plot on the Bush-real dataset (2).