

METHOD

Open Access



BlastFrost: fast querying of 100,000s of bacterial genomes in Bifrost graphs

Nina Luhmann^{1*} , Guillaume Holley² and Mark Achtman¹

*Correspondence:

n.luhmann@warwick.ac.uk

¹Warwick Medical School, University of Warwick, Coventry, UK

Full list of author information is available at the end of the article

Abstract

BlastFrost is a highly efficient method for querying 100,000s of genome assemblies, building on Bifrost, a dynamic data structure for compacted and colored de Bruijn graphs. BlastFrost queries a Bifrost data structure for sequences of interest and extracts local subgraphs, enabling the identification of the presence or absence of individual genes or single nucleotide sequence variants. We show two examples using *Salmonella* genomes: finding within minutes the presence of genes in the SPI-2 pathogenicity island in a collection of 926 genomes and identifying single nucleotide polymorphisms associated with fluoroquinolone resistance in three genes among 190,209 genomes. BlastFrost is available at <https://github.com/nluhmann/BlastFrost/tree/master/data>.

Introduction

Recent advances in DNA sequencing technologies have reduced sequencing costs and hands-on time, and whole-genome sequencing of bacterial pathogens is being routinely performed by public health organizations. The resulting sequence reads and genome assemblies are deposited in the public domain [1–3], enabling comparative analyses of 100,000s of genomes [4, 5] from individual bacterial genera for evolutionary or epidemiological investigations.

New sequencing data are now routinely uploaded to public databases such as the Sequence Read Archives (SRA [6]), which provide ready access to extensive collections of sequencing data for many bacterial genera. Sequences from specific bacterial pathogens are also available as curated collections of genomic assemblies bundled with their meta-data together with dedicated tools for population genomic analyses. Such databases include for example PubMLST [7] and EnteroBase [3].

The analysis of genomic sequences by phylogenetic approaches can yield insights into evolutionary distances for 1000s of bacterial genomes. However, large comparative studies based on sequencing data are limited by computing resources and calculation speed [5]. Even the seemingly simple task of identifying all bacterial strains within a collection that contains a specific antimicrobial resistance gene or other genes of interest is a computational challenge for the large data sets that are currently available. The most popular



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

methods for sequence comparison are BLAST [8] and its successors. However, these alignment-based methods do not scale well for queries of the presence or absence of genes in large data sets. As a result, as recently reviewed by Marchet et al. [9], the alignment step is replaced in some recent software by a k -mer approach [10], in which sets of short sub-sequences of fixed length k are compared between a query and a sequence database. These approaches were implemented because k -mers can directly identify diverse genetic modifications such as single nucleotide polymorphisms (SNPs), insertions or deletions from short read sequences, and do not require assembled genomes or an explicit reference genome.

One recent k -mer-based method, BIGSI, employs a data structure that stores a Bloom filter [11] of k -mers for each genome in a database, and can subsequently index and search very large databases of bacterial and viral sequences [12]. BIGSI queries are very efficient, but the European Nucleotide Archive (ENA) was already so large in 2016 that creating a BIGSI index took months. Furthermore, BIGSI was designed for dealing with genetically diverse collections of data, and other methods and different data structures might be more efficient for creating a query index of sequence data from closely related genomes. A potentially faster approach for the construction of indices would be to index sets of k -mers in a de Bruijn graph [13], where shared k -mers are automatically collapsed into single nodes. Collapsing k -mers that are shared between closely related genomes would decrease both the storage space for the index and the search space for subsequent queries. Recent implementations of such an approach include Mantis [14], Rainbowfish [15], and VARI-Merge [16]. They build joint de Bruijn graphs for multiple genomes, coloring nodes by their source genomes (colored de Bruijn graphs [17]), and can traverse both the shared paths in the graph which represent conserved regions as well as diverging paths which represent variable regions. However, the implementations of these methods do not scale well enough to efficiently handle a modern, large sequence collection [18]. For example, VARI-Merge was benchmarked on a data set of 16,000 *Salmonella* genomes [16], but the *Salmonella* database in EnteroBase already contains > 250,000 genomes.

The recent development of Bifrost [18] introduced a memory efficient, dynamic data structure for indexing either colored or non-colored compacted de Bruijn graphs. It presents a broad range of functions that support querying both sequences and colors, annotating individual vertices, and editing Bifrost graphs while preserving their compaction. The implementation of Bifrost facilitates its ability to rapidly build joint graphs of 100,000s of genomes and permits incremental updates of these large graphs with additional data. However, Bifrost only implements basic k -mer querying. Here, we introduce BlastFrost, a method implemented in C++ for similarity searches in Bifrost graphs by rapid k -mer matching. BlastFrost uses the underlying Bifrost graph structure to extract subgraphs defined by a query, and can thereby efficiently extract sequence variants of the query from a data base of 100,000s of bacterial genomes. Here, we show that BlastFrost performs better than Blast and BIGSI with closely related genomes, and illustrate its features by case studies on the identification of genomic islands and of individual mutations in antimicrobial resistance genes.

Results

Uncompacted de Bruijn graphs of genomic sequences are a popular graph data structure consisting of nodes representing sequences of k -mers within the input genomes. Edges in

the graph represent fixed overlaps of length $k - 1$ between neighboring nodes and can therefore be implicit. Bifrost [18] indexes bacterial genomes in a time and memory efficient implementation of a compacted and colored de Bruijn graph. Here, maximal paths of multiple sequential, non-branching nodes are compacted into single nodes (unitigs) by collapsing the overlaps. In addition, each node is assigned a set of colors representing all input genomes containing the corresponding k -mers of the unitig. We henceforth refer to this particular form of compacted and colored de Bruijn graphs as Bifrost graphs.

BlastFrost relies on Bifrost graphs. As depicted in Fig. 1, we implemented a k -mer search function in BlastFrost which can identify the presence or absence of a query sequence in any of the genomes in a Bifrost graph. The results of that search can be used for subgraph extraction (Fig. 1 bottom) of query matches in order to identify all variants of the query sequence in the Bifrost graph. The following paragraphs provide an overview of the method. Algorithmic details can be found in [Supplemental Material](#).

BlastFrost query search

The input for BlastFrost consists of a Bifrost graph file in GFA format plus an index of the colors of each k -mer in each unitig, pre-computed for a certain k value. We henceforth refer to the genomes indexed in the graph as colors. The input parameters to BlastFrost also include a link to a FASTA file containing one or more query sequences.

For each query sequence, BlastFrost calculates a set of overlapping k -mers, using the same value of k that was used to build the Bifrost graph. This set is then used to search for the query sequence in all genomes in the graph, relying on specific functions from the Bifrost API that determine the presence of each k -mer and its colors in the Bifrost graph. Each query results in a binary sequence for each color of 1s and 0s representing k -mer hits and misses. The k -mer-based search in the graph explicitly assumes that overlapping k -mers of the same color are also contiguous in the underlying genome, which speeds up computation. BlastFrost speeds up computations even further by taking advantage of the fact that Bifrost graphs are compacted into unitigs which encode non-branching nodes as single nodes, and assumes that the color set of a unitig is the same as the individual color sets of each k -mer in that unitig.

A single nucleotide substitution between a query and a color will result in k mismatching k -mers, assuming that the size of k was large enough to avoid random hits in the genomes. The resulting binary sequence would then contain a stretch of k 0's in the binary hit sequence for that query. Deletions are also characterized by runs of 0's that are potentially smaller than k , while insertions and multiple substitutions can lead to longer runs of 0's in the hit sequence. In order to evaluate the significance of k -mer hits between a query and a specific color, we adopted the BLAST approach for computing an E value [19] based on an estimated alignment score, derived from the lengths of 0 runs in the k -mer hits.

To increase the sensitivity of the k -mer-based query, BlastFrost allows additional querying of all k -mers related to a query k -mer by a Hamming distance smaller than or equal to an input parameter d . We refer to this set of additional k -mers as k -mer neighborhoods (Fig. 1). In the following evaluation, we present the necessity for this increased sensitivity, as well as some of the resulting trade-offs.

BlastFrost subgraph extraction

The raw results on k -mer hits from a Bifrost graph are not immediately informative on the

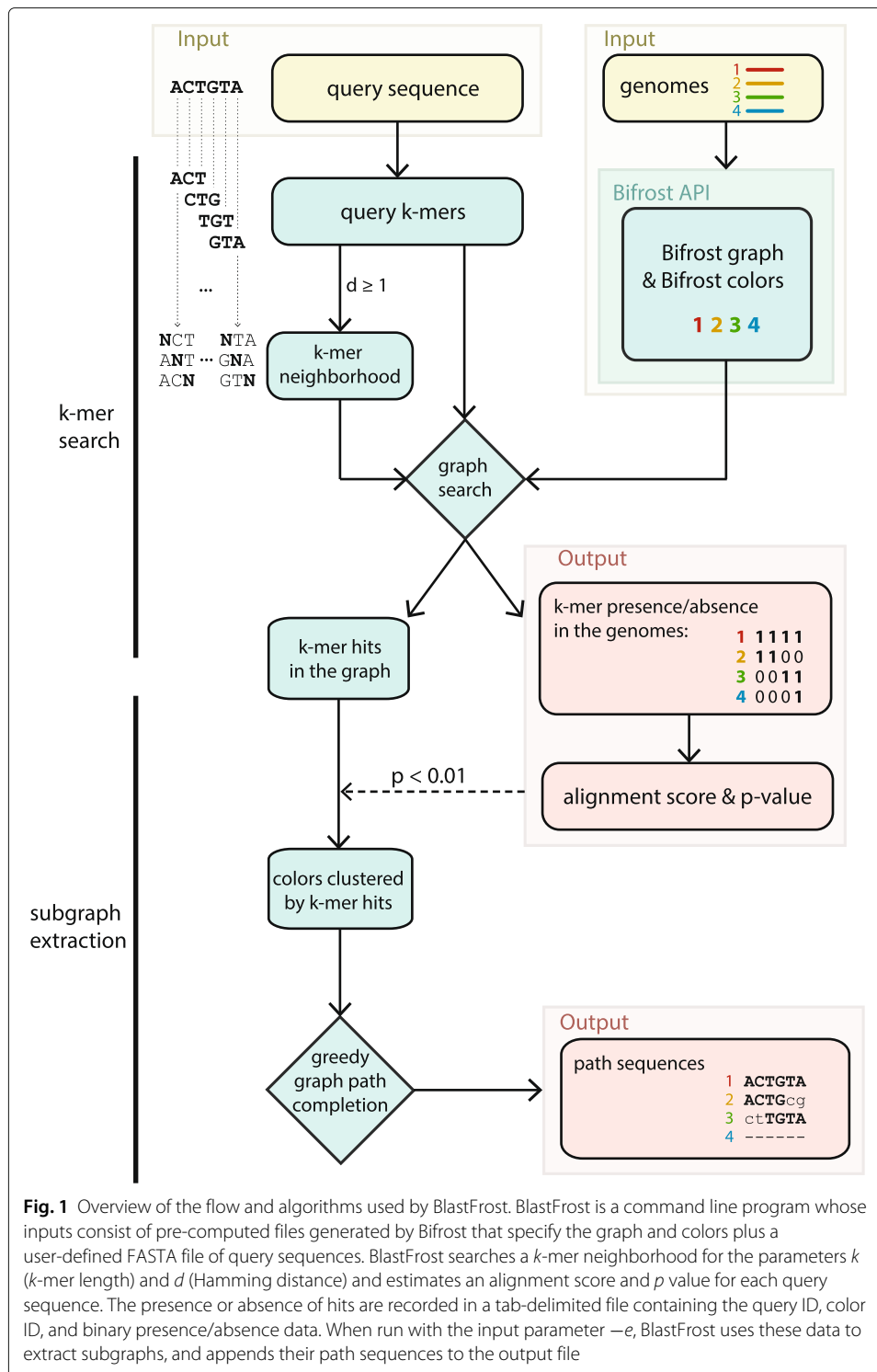


Fig. 1 Overview of the flow and algorithms used by BlastFrost. BlastFrost is a command line program whose inputs consist of pre-computed files generated by Bifrost that specify the graph and colors plus a user-defined FASTA file of query sequences. BlastFrost searches a *k*-mer neighborhood for the parameters *k* (*k*-mer length) and *d* (Hamming distance) and estimates an alignment score and *p* value for each query sequence. The presence or absence of hits are recorded in a tab-delimited file containing the query ID, color ID, and binary presence/absence data. When run with the input parameter *-e*, BlastFrost uses these data to extract subgraphs, and appends their path sequences to the output file

genomic locations of the query hits, the numbers of copies of those query sequences in each genome, or on syntenic relationships. For any specific query, each binary sequence of *k*-mer hits represents a potentially incomplete path of nodes for each color in the graph interrupted by nucleotide changes that were not included in the *k*-mers that are shared between the query and the genome. BlastFrost can account for these potential gaps by

extending the k -mer hit results, and produce a subgraph for each successful k -mer query. Starting from the first unitig in the original k -mer hit list for a specific color, BlastFrost greedily completes a path by traversing non-branching paths of the same color within the graph, i.e., each unitig initially found in the k -mer search is tested for all successor unitigs with the same color. The subgraph is then used to reconstruct the corresponding sub-sequence of each color from the path in the Bifrost graph.

To avoid completing the same paths more than once, BlastFrost clusters colors sharing k -mer hits, completes all their paths simultaneously, and removes colors from those clusters that are absent in intervening unitigs. For each path and its accompanying colors, BlastFrost output the genome sequence in addition to the binary sequence mentioned above. These data allow ready identification of variant positions that distinguish the query from the extracted path sequences.

Evaluation and benchmarks

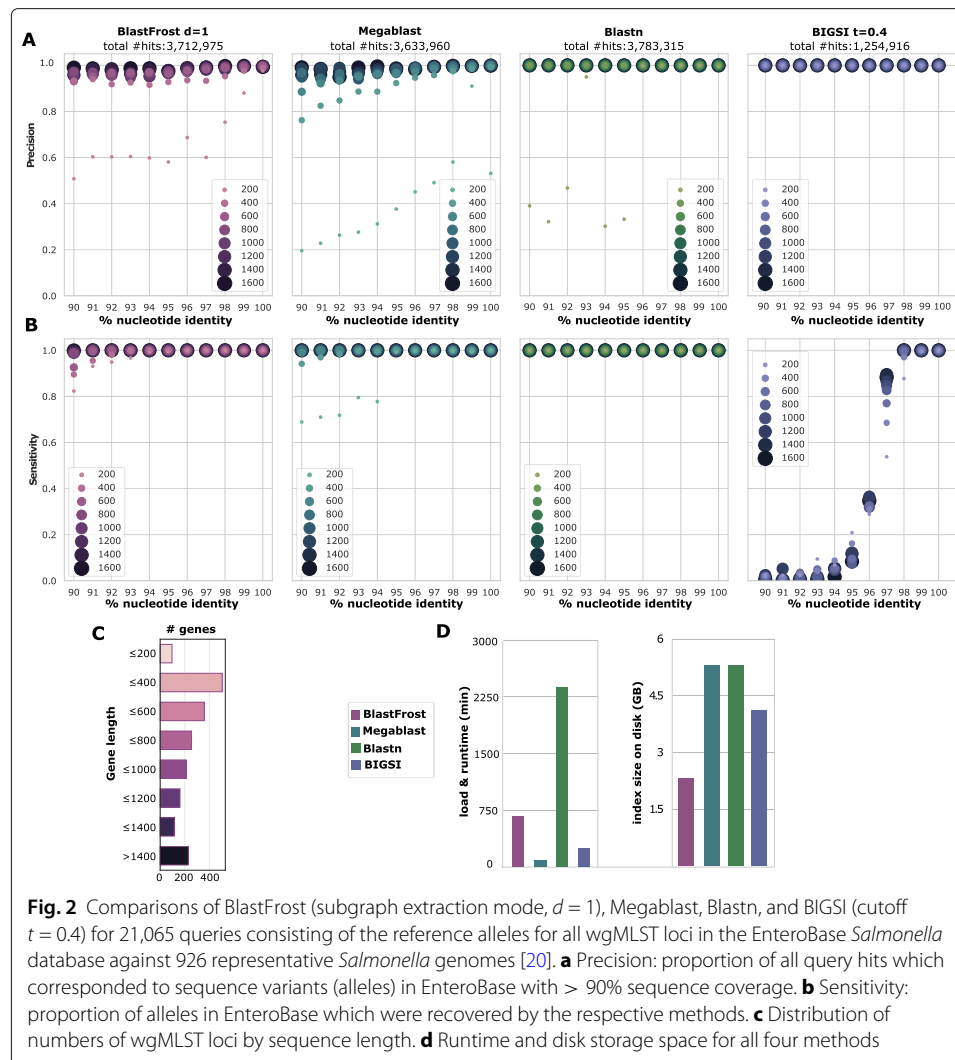
Precision and sensitivity of identifying the presence or absence of all genes in a pan-genome

We used the whole-genome MLST (wgMLST) scheme for the genus *Salmonella* in Enterobase [3, 20] to test the accuracy of BlastFrost for detecting sequence variants of a large number of query sequences in a large number of related genomes. That wgMLST scheme consists of 21,065 single copy orthologs which had been derived from a pan-genome of 537 representative genomes of *Salmonella* with PEPPAN [20, 21]. Enterobase identifies diverse sequence variants of those loci in each assembled genome by combining BLASTN [8] nucleotide and UBLAST [22] amino acid queries, and also scores the absence of significant hits for each genome. Thus, this data set is ideal for testing the efficiency of the detection of presence and absence of multiple genes because both the presence and absence as well as the genome-specific sequence are known for all 21,065 loci in all *Salmonella* genomes in Enterobase.

Bifrost created a graph of 926 representative *Salmonella* genomes from Enterobase [20] in less than 24 min and required less than 5 GB of memory. The graph occupies 2.3 GB of disk space, and it contains more than 33 million unitigs.

We ran BlastFrost (parameter $d = 1$ to support inexact searches) on this graph with 21,065 query sequences, consisting of one representative allele for each locus, and extracted all allelic variants from the corresponding subgraphs.

To calculate precision and sensitivity, we scored extracted sequences that covered at least 90% of the query sequence in a pairwise alignment as being correct. We bin query hits by the nucleotide identity between the query and the Enterobase allele, or the nucleotide identity between the query and the search result if an allele is not stored in Enterobase. We also performed similar analyses with the programs Megablast, which is the default version used by Blast [8], the classical version of BLASTN, and BIGSI [12] with the parameter $t = 0.4$ to support inexact searches. Initial comparisons showed that the precision of all of these methods was very low for genes of less than 200-bp length (Fig. 2a), except for BIGSI which had extremely high precision, and the following description is restricted to genes that were larger than 200 bp. The precision with BlastFrost was at least 95% for alignments with at least 90% nucleotide identity. MegaBlast had somewhat lower precision for genes of less than 400 bp, whereas BLASTN retained very high precision at all levels. We also examined the nature of the false-positive hits by BlastFrost according to our criteria of correct hits to determine whether these were



potentially truly absent in the corresponding genomes. Indeed, all false positives, including genes of < 200 bp length, were found in those genomes with BLASTN, indicating that false positives may have been scored because EnteroBase scoring excludes repetitive DNA elements, including overlapping or duplicated sequences, whereas BlastFrost finds all sequences, including such repetitive DNA. Sensitivity with BlastFrost was 100% down to 94% nucleotide identity and then dropped to $> 94\%$ at 90% identity (Fig. 2b). Megablast and BLASTN yielded almost perfect sensitivity throughout. BIGSI failed dramatically with nucleotide identities below 95%, because it only identified a limited number of hits in total despite a low cutoff parameter t , explaining its superb precision values. In summary, BlastFrost correctly identified all sequence variants down to 90% sequence identity with a gene length of > 200 bp.

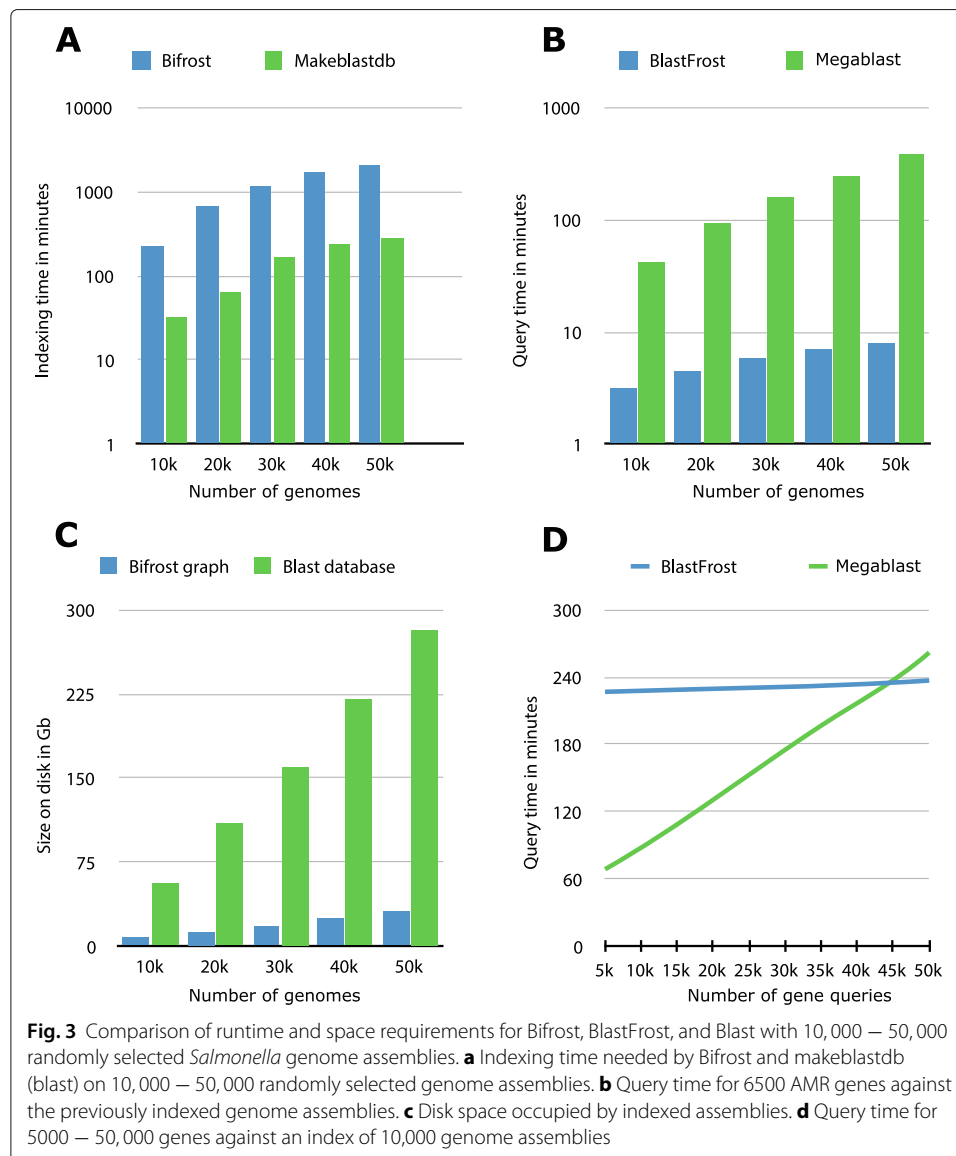
The runtime of Megablast and BIGSI was much faster than that of BlastFrost while the runtime of BLASTN was threefold slower (Fig. 2d). BlastFrost also required the least disk space for the genome indices of all these programs (Fig. 2e).

Benchmarking

The initial results in Fig. 2 indicate that BlastFrost was slower than Megablast and

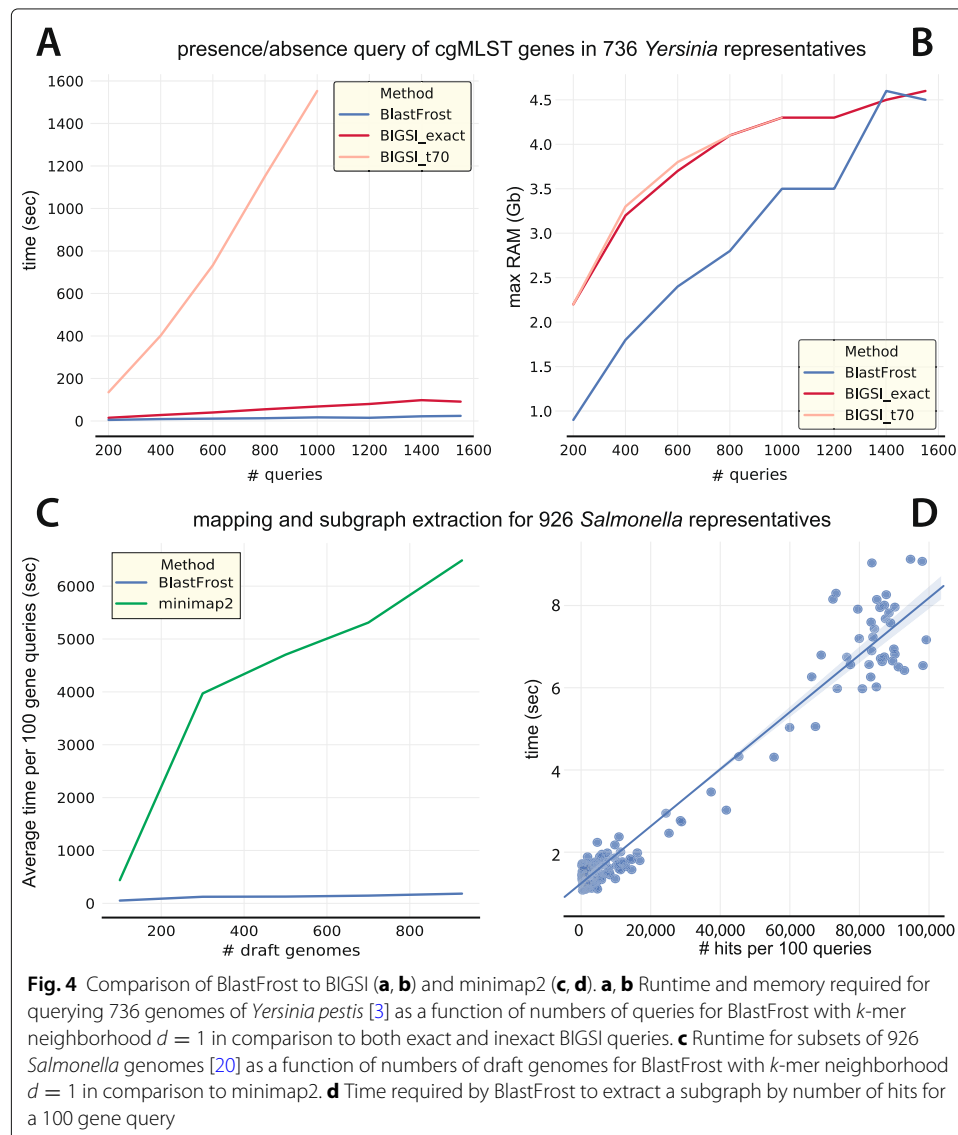
BIGSI. We therefore compared the time and memory requirements between these different methods and minimap2 [23], another widely used software tool, in greater detail to explore their relative strengths.

We compared the times needed by BlastFrost and Bifrost to Megablast for indexing 10,000 – 50,000 *Salmonella* genome assemblies and querying them with multiple individual genes (Fig. 3). The indexing step in BLAST is close to an order of magnitude faster than graph construction by Bifrost (Fig. 3a). However, after loading the graph, BlastFrost is more than tenfold faster than MegaBlast at searching the constructed graph for the presence of 6500 AMR genes (Fig. 3b) because of the extensive indexing information within the Bifrost graph. The Bifrost graph also needs much less disk space than a BLAST database (Fig. 3c). Megablast is clearly faster than BiFrost plus BlastFrost for a single phase of creating an index 10,000 genomes plus a single round of querying up to 50,000 genes (Fig. 3d). Under conditions where a genome database grows continuously and gene queries are conducted repeatedly, Bifrost plus BlastFrost would be much quicker because



Bifrost can rapidly and continuously expand its pre-constructed graph to include additional genomes whereas BLAST needs to calculate a complete index for each additional genome.

We also investigated calling the presence or absence of core genes with BlastFrost and BIGSI [12] on a data set of 736 genomes from representative strains of *Yersinia pestis*, whose genomic diversity is very limited [3]. Creating indices of the 736 genomes was much faster with Bifrost than with BIGSI (Supplemental Material). The query time and maximum RAM usage were measured with BlastFrost and BIGSI for 200 – 1600 random core gene sequences. BlastFrost was timed in an inexact search for k -mer hits with up to one nucleotide mismatch (parameter $d = 1$). For BIGSI, we timed an exact search function as well as an inexact search function for query hits containing at least 70% of the query sequence k -mers (parameter $t = 0.7$). The (inexact) BlastFrost query yielded the same hits as the inexact BIGSI search, but BIGSI was much slower. BlastFrost searches were slightly faster than the exact BIGSI searches (Fig. 4a) and used much less RAM for less than 1200 queries.



Finally, we timed subgraph extraction compared between BlastFrost and minimap2 [23], which is the currently most efficient mapping tool for both short reads as well as chromosome-scale alignments. The average speed needed to extract 100 genes from the wgMLST *Salmonella* scheme described above from was measured across subsets of the 926 representative *Salmonella* genomes. The measurement showed BlastFrost is much faster than minimap2 (Fig. 4c). The time needed for BlastFrost to extract a subgraph is dependent on the number of hits for that query (Fig. 4d), but it achieves a slightly sub-linear growth in time requirement because identical genome segments can be found in multiple genomes within a bacterial genus.

Applications

We took advantage of the large genomics databases available in EnteroBase to demonstrate the abilities of BlastFrost to find variably present genomic elements and to identify single nucleotide variants of individual genes. For genomic elements, we searched the 926 representative genomes of *Salmonella* for known genes in the SPI-2 *Salmonella* pathogenicity island [24–26]. For nucleotide variants, we screened the entire EnteroBase *Salmonella* database for specific substitutions in three genes that are associated with fluoroquinolone resistance in *Salmonella* [27].

Genomic islands

Genomic islands consist of clustered genes from the accessory genome that can be acquired by bacteria through horizontal gene transfer, or which are lost due to gene deletion [28, 29]. Pathogenicity islands are a distinct sub-class of genomic islands, which can range in size from 10 to 200 kb, and encode genes which can contribute to the virulence of the bacteria [26, 30]. SPI-2 is such an island which seems to have been acquired by *Salmonella* after the divergence of *S. bongori* and *S. enterica* from their common ancestor. Subsequently, *S. enterica* split into multiple so-called subspecies [20, 31].

The 44 genes in SPI-2 from *S. enterica* serovar Typhimurium strain LT2 were queried against the Bifrost graph of the 926 representative *Salmonella* genomes described above. Figure 5 shows their distribution according to an exact search (BlastFrost parameter $d = 0$, dark green) and an inexact search (BlastFrost parameter $d = 2$, light green). The inexact search indicated that most of the SPI-2 genes are present in all of the *Salmonella* subspecies and that they were all absent, as expected [30], in *Salmonella bongori*. However, some genes were absent from many or most of the genomes from individual subspecies, or their sequences were too divergent to be detected. This figure also emphasizes the importance of inexact querying because although most SPI-2 genes were identified by the exact search in subspecies I, II, and VI, the inexact search greatly increased the number of SPI-2 genes identified in the other subspecies.

BlastFrost took 111 s to load the Bifrost graph of 926 *Salmonella* genomes into memory and took a further 540 s to search for all SPI-2 genes with the inexact BlastFrost search, for a total of under 11 min.

Nucleotide variants

The subgraph extraction functionality of BlastFrost can also extract known variants of genes involved in antimicrobial resistance or other phenotypes. We downloaded 160,000 *Salmonella* draft assemblies from EnteroBase and created an initial Bifrost graph of those

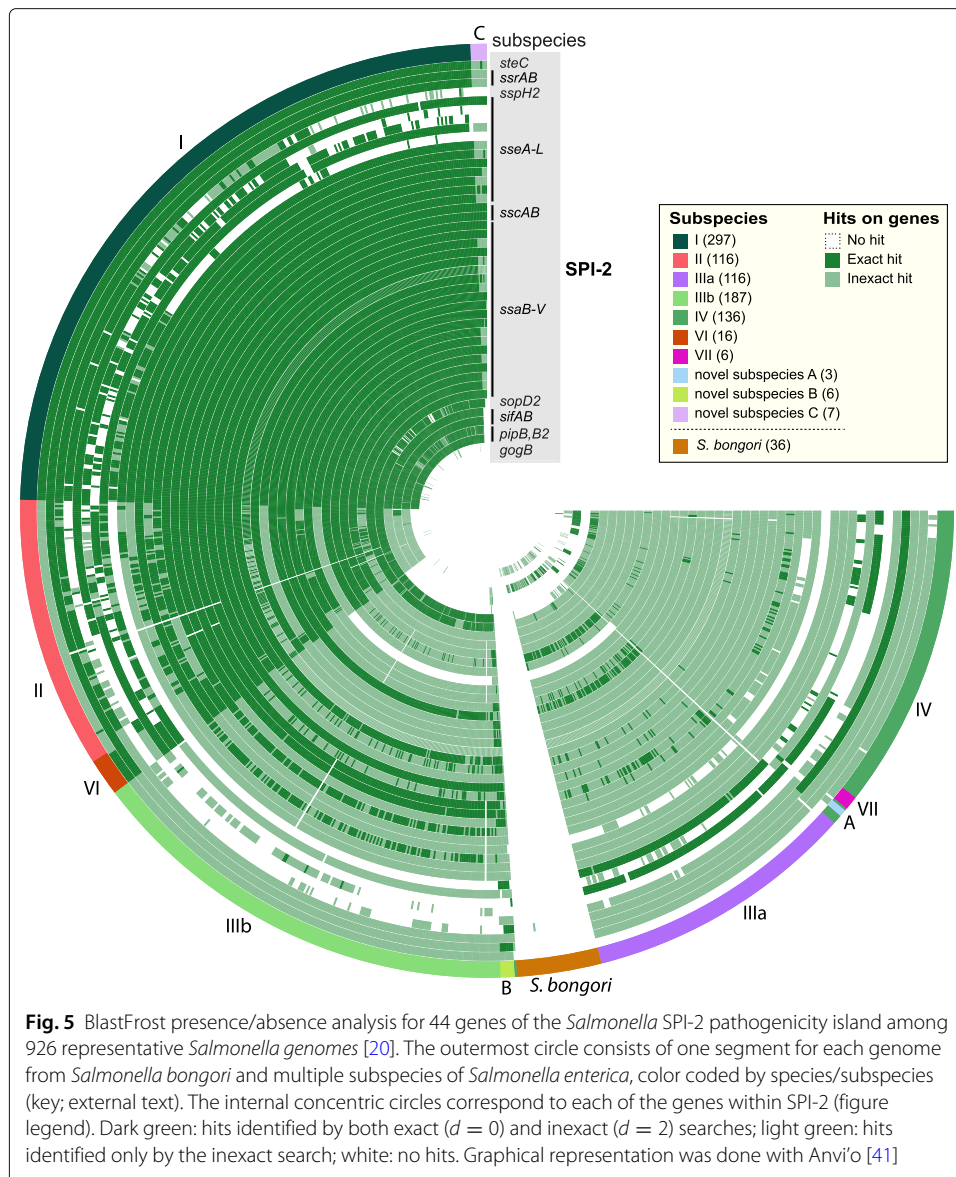
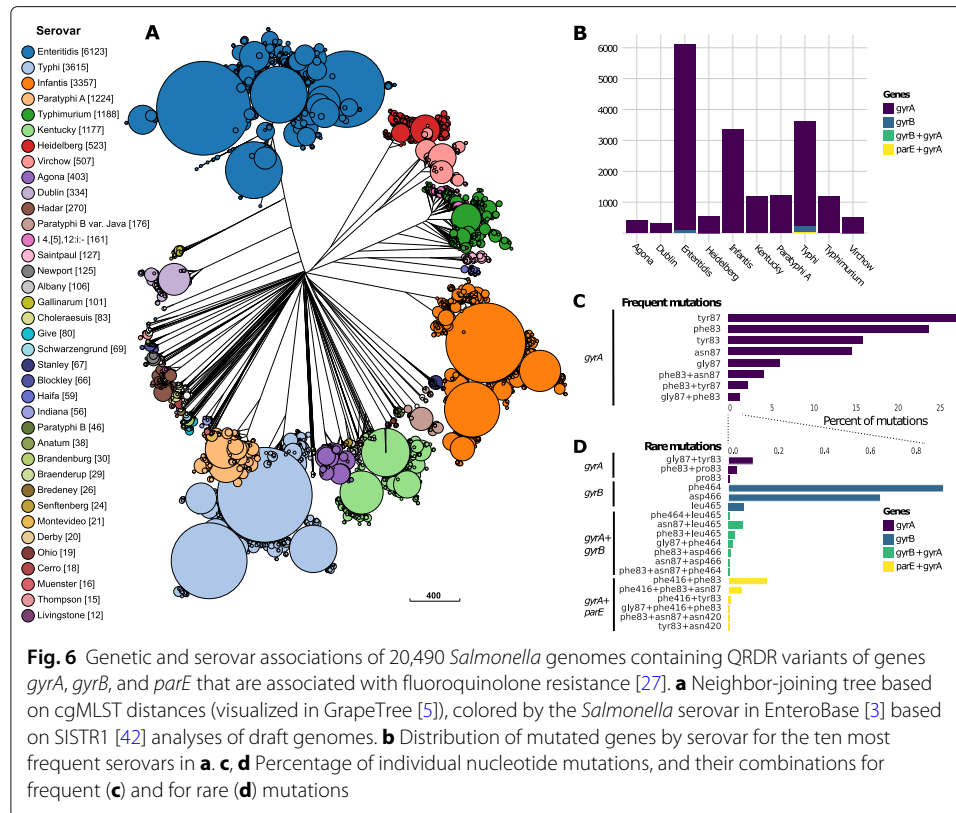


Fig. 5 BlastFrost presence/absence analysis for 44 genes of the *Salmonella* SPI-2 pathogenicity island among 926 representative *Salmonella* genomes [20]. The outermost circle consists of one segment for each genome from *Salmonella bongori* and multiple subspecies of *Salmonella enterica*, color coded by species/subspecies (key; external text). The internal concentric circles correspond to each of the genes within SPI-2 (figure legend). Dark green: hits identified by both exact ($d = 0$) and inexact ($d = 2$) searches; light green: hits identified only by the inexact search; white: no hits. Graphical representation was done with Anvi'o [41]

genomes. This took 4 days and 15 h computation time and required 147 GB of memory. During the course of the investigations in this manuscript, we subsequently updated the Bifrost graph in several iterations, resulting in a final graph containing 190,209 genomes. Updating the Bifrost graph update with 100 additional genomes took about 2.5 h, including the time to load the graph back into memory. The disk size of the final graph of 190,209 genomes is 158.5 GB and it contains 32,692,889 unitigs. We then queried this graph with BlastFrost for a single representative gene sequence from each of the genes *gyrA*, *gyrB*, and *parE*. These genes were chosen because they possess quinolone resistance-determining regions (QRDR) in which individual nucleotide variants can cause reduced susceptibility to fluoroquinolones [27]. The queries resulted in one subgraph per gene, whose sequences were aligned, and scanned for the known nucleotide variants.

Our results showed that 20,490 genomes from multiple serovars (Fig. 6a) contained these QRDR nucleotide variants. Those serovars included common causes of human



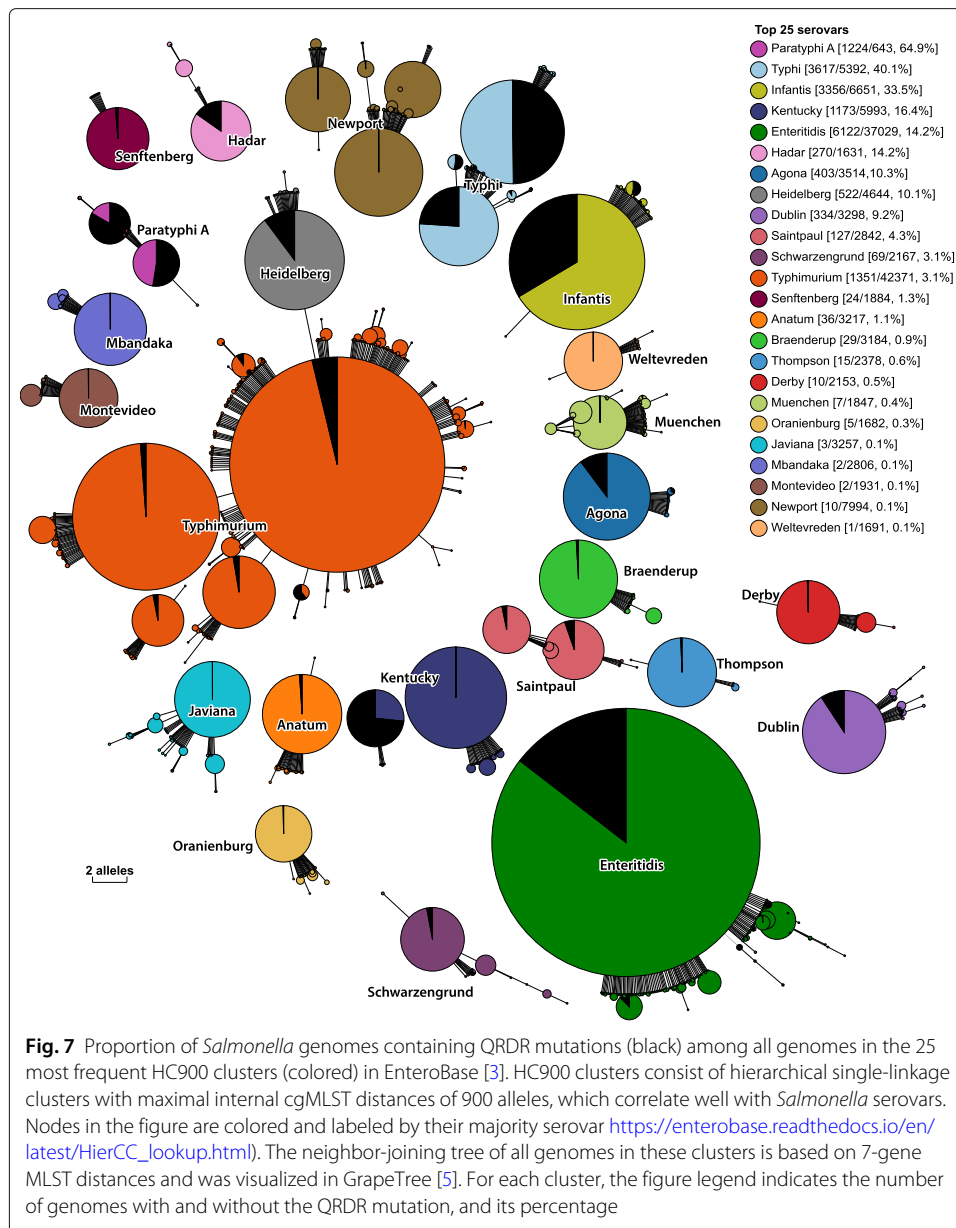
disease, such as Enteritidis, Typhi, and Typhimurium, as well as multiple others that are common in domesticated animals but can cause food-borne gastroenteritis in humans (Fig. 5b). Most of the genomes identified in these BlastFrost queries contain a single nucleotide variant in *gyrA* (89.7%) (Fig. 6c). Variants in *gyrB* (1.8%) and *parE* (0.26%) were also found but they were less common and were normally present together with *gyrA* mutations in the same genomes (Fig. 6c,d).

Most of the *Salmonella* genomes from EnteroBase did not contain these QRDR mutations. The relative proportions of genomes with and without those QRDR mutations are illustrated for common serovars in Fig. 7. Serovars Paratyphi A or Typhi showed the greatest proportions of strains with resistance mutations. Interestingly, almost all fluoroquinolone-resistant strains of serovar Kentucky belong to only one of the two genetic clusters that are associated with this polyphyletic serovar [20, 32].

BlastFrost took 25 min to load the Bifrost graph into memory and 3.5 h to extract all subgraphs using 8 threads. It used a maximum of 160 GB RAM for these analyses.

Discussion and conclusions

BlastFrost implements a highly efficient algorithm for querying de Bruijn graphs, and thereby complements the very computationally efficient Bifrost [18], which calculates compacted and colored graphs that scales to 100,000s of closely related bacterial genomes. Practical applications of the combination of these two methods are also greatly facilitated by the existence of structured sequence databases of closely related bacteria such as EnteroBase [3]. EnteroBase includes genomic assemblies of 100,000s of bacterial strains together with genotypes based on legacy or core genome MLST, which facilitate the



visualization of genetic relationships among the query hits. The combination of Bifrost, BlastFrost, and EnteroBase has the potential to rapidly reveal numerous features of genomic diversity that were previously not readily accessible.

All MLST schemes are inherently limited, because they are based on a fixed selection of genes that were present in an initial, representative set of genomes. However, many bacterial genera are associated with open pan-genomes [33], whose content continues to increase with each additional genome that is sequenced [21], and such novel sequences are not routinely appended to the MLST schemes. Therefore, it is important to emphasize that BlastFrost and Bifrost are not dependent on MLST or on genomic annotations, but can handle any collection of closely related genomic assemblies. BlastFrost can summarize diversity within large, variable regions such as genomic islands. It can also identify

variants of any sequence of interest and subsequently rapidly analyze them to identify single nucleotide polymorphisms.

We note that in theory, we can only guarantee that all query k -mers found in the graph for a specific color are really present in the underlying genome. If BlastFrost is used to query sequences containing repeats longer than the value of k , this could lead to false sequences returned due to the underlying data structure used. Hence, k should be chosen large enough to avoid false-positive sequences [34].

We compared the speed and memory requirements for large genomic data sets between BlastFrost/Bifrost, Megablast, classical BLASTN, and the current state of the art tools BIGSI and minimap2. Computing a Bifrost graph is a costly indexing step that runs much slower than creating a BLAST database. However, BlastFrost can run queries much more quickly than BLAST, which compensates for the extra time needed to create an index if that index is used repeatedly for a large number of queries. As a use case, BlastFrost would be suitable as a web service in support of large databases such as EnteroBase, because it would provide fast search functionalities and could even support comparisons of all genes from numerous whole genomes against a large pan-genome. BlastFrost is not suitable for indexing and querying diverse sequence collections such as RefSeq or SRA, unlike either BIGSI or BLAST. However, for closely related genomes, such as those within a single bacterial genus, BlastFrost is considerably faster than BIGSI and requires less memory for up to 1400 sequence queries. Similarly, BlastFrost is much faster than minimap2 for closely related genomes and also requires less memory. These computational efficiencies did not sacrifice accuracy. BlastFrost has high precision and sensitivity for sequences that are at least 90% identical and over 200 bp in length.

BlastFrost enables the identification of genomic islands or individual nucleotides associated with antimicrobial resistance genes because of the explicit graph data structure in Bifrost which supports graph traversal and extraction of sequences that extend beyond the k -mers that were used for querying. Given a Bifrost graph, genomic islands or nucleotide variants can be identified within 100,000s of genomes in minutes. The Bifrost API freely supports annotation of nodes in the graph, including annotating unitigs with additional data. In future extensions, BlastFrost should be able to extract the local synteny from graphs whose unitigs are annotated with genome coordinates and/or gene annotations. Such information could also be used to reconstruct genomic rearrangements.

BlastFrost is not a general replacement for calling SNPs because its precision suffers with increasing genetic diversity and reduced sequence length. However, it might have the potential for incorporation into approaches to detection of antimicrobial resistance in combinations of databases of AMR genes such as CARD [35] or AMRfinder [36] with genomic sequence collections such as EnteroBase.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-020-02237-3>.

Additional file 1: Supplementary file. Description of data structures and algorithm pseudocode

Additional file 2: Review history

Acknowledgements

This project was supported by the Wellcome Trust (202792/Z/16/Z). GH was supported by the Icelandic Research Fund Project grant number 152399-053. We thank Zheming Zhou, Jane Charlesworth, and Páll Melsted for their helpful feedback during the development of BlastFrost.

Peer review information

Andrew Cosgrove was the primary editor of this manuscript and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 2.

Authors' contributions

NL developed and implemented BlastFrost. NL ran all analyses. GH implemented Bifrost and the Bifrost API suitable for BlastFrost. NL and MA designed the analyses and wrote the manuscript. The authors read and approved the final manuscript.

Authors' information

Twitter handles: @nina_luhmann (Nina Luhmann); @GuillaumeOleSan (Guillaume Holley).

Funding

NL and MA were supported by the Wellcome Trust (202792/Z/16/Z). GH was supported by the Icelandic Research Fund Project grant number 152399-053.

Availability of data and materials

The source code of BlastFrost is available as open source software at <https://github.com/nluhmann/BlastFrost> [37]. The source code is released under a GPL-3.0 License. The exact version used in this paper is archived at Zenodo [38].

All genome assemblies used in the analyses in this paper were downloaded from EnteroBase. For each dataset, a file containing all EnteroBase barcodes for these assemblies has been stored under <https://github.com/nluhmann/BlastFrost/data> [39].

The wgMLST scheme for *Salmonella* in EnteroBase can be downloaded from https://enterobase.warwick.ac.uk/species/senterica/download_data [3].

The sequences for SPI-2 genes were downloaded from the Virulence Factor Database at <http://www.mgc.ac.cn/cgi-bin/VFs/genus.cgi?Genus=Salmonella> [40].

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Warwick Medical School, University of Warwick, Coventry, UK. ²Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, University of Iceland, Reykjavik, Iceland.

Received: 10 February 2020 Accepted: 11 December 2020

Published online: 11 January 2021

References

1. Allard MW, Strain E, Melka D, Bunning K, Musser SM, Brown EW, Timme R. Practical value of food pathogen traceability through building a whole-genome sequencing network and database. *J Clin Microbiol.* 2016;54(8):1975–83.
2. Ribot EM, Freeman M, Hise KB, Gerner-Smidt P. PulseNet: entering the age of next-generation sequencing. *Foodborne Pathog Dis.* 2019;16(7):451–6.
3. Zhou Z, Alikhan N-F, Mohamed K, Fan Y, Achtman M, Brown D, Chattaway M, Dallman T, Delahay R, Kornschober C, et al. The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. *Genome Res.* 2020;30(1):138–52.
4. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics.* 2006;23(1):127–8.
5. Zhou Z, Alikhan N-F, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, Carriço JA, Achtman M. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res.* 2018;28(9):1395–404.
6. Leinonen R, Sugawara H, Shumway M, Collaboration INSD. The sequence read archive. *Nucleic Acids Res.* 2010;39(suppl_1):19–21.
7. Jolley KA, Bray JE, Maiden MC. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res.* 2018;3:124.
8. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403–10.
9. Marchet C, Boucher C, Puglisi SJ, Medvedev P, Salson M, Chikhi R. Data structures based on k-mers for querying large collections of sequencing data sets. *Genome Res.* 2020. <https://doi.org/10.1101/gr.260604.119>.
10. Jaillard M, Lima L, Tournoud M, Mahé P, Van Belkum A, Lacroix V, Jacob L. A fast and agnostic method for bacterial genome-wide association studies: bridging the gap between k-mers and genetic events. *PLoS Genet.* 2018;14(11):1007758.
11. Bloom BH. Space/time trade-offs in hash coding with allowable errors. *Commun ACM.* 1970;13(7):422–6.

12. Bradley P, den Bakker HC, Rocha EP, McVean G, Iqbal Z. Ultrafast search of all deposited bacterial and viral genomic data. *Nat Biotechnol.* 2019;37(2):152.
13. Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci.* 2001;98(17):9748–53.
14. Pandey P, Almodaresi F, Bender MA, Ferdman M, Johnson R, Patro R. Mantis: a fast, small, and exact large-scale sequence-search index. *Cell Syst.* 2018;7(2):201–7.
15. Almodaresi F, Pandey P, Patro R. Rainbowfish: a succinct colored de Bruijn graph representation. In: Proc. of the 17th Workshop on Algorithms in Bioinformatics (WABI'17). Schloss Dagstuhl-Leibniz-Zentrum für Informatik; 2017.
16. Muggli MD, Alipanahi B, Boucher C. Building large updatable colored de Bruijn graphs via merging. *Bioinformatics.* 2019;35(14):i51–i60.
17. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet.* 2012;44(2):226.
18. Holley G, Melsted P. Bifrost: highly parallel construction and indexing of colored and compacted de Bruijn graphs. *Genome Biol.* 2020;21(1):1–20.
19. The statistics of sequence similarity scores. 2020. <https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html>. Accessed 13 Sept 2020.
20. Alikhan N-F, Zhou Z, Sergeant MJ, Achtman M. A genomic overview of the population structure of *Salmonella*. *PLoS Genet.* 2018;14(4):1007261.
21. Zhou Z, Charlesworth J, Achtman M. Accurate reconstruction of bacterial pan- and core genomes with PEPPAN. *Genome Res.* 2020;30(11):1667–79.
22. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics.* 2010;26(19):2460–1.
23. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34(18):3094–100.
24. Wisner A, Desin T, White A, Potter A, Köster W. The *Salmonella* pathogenicity island-1 and -2 encoded type III secretion systems. In: *Salmonella—a diversified superbug*. IntechOpen; 2012.
25. Hensel M. *Salmonella* pathogenicity island 2. *Mol Microbiol.* 2000;36(5):1015–23.
26. Ochman H, Groisman EA. Distribution of pathogenicity islands in *Salmonella* spp. *Infect Immun.* 1996;64(12):5410–2.
27. Song Y, Roumagnac P, Weill F-X, Wain J, Dolecek C, Mazzoni CJ, Holt KE, Achtman M. A multiplex single nucleotide polymorphism typing assay for detecting mutations that result in decreased fluoroquinolone susceptibility in *Salmonella enterica* serovars Typhi and Paratyphi A. *J Antimicrob Chemother.* 2010;65(8):1631–41.
28. Daubin V, Ochman H. Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. *Genome Res.* 2004;14(6):1036–42.
29. Szöllösi GJ, Boussau B, Abby SS, Tannier E, Daubin V. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc Natl Acad Sci.* 2012;109(43):17513–8.
30. Hensel M. Evolution of pathogenicity islands of *Salmonella enterica*. *Int J Med Microbiol.* 2004;294(2–3):95–102.
31. Crisculo A, Issenhuth-Jeanjean S, Didelot X, Thorell K, Hale J, Parkhill J, Thomson NR, Weill F-X, Falush D, Brisse S. The speciation and hybridization history of the genus *Salmonella*. *Microb Genomics.* 2019;5(8):.
32. Achtman M, Wain J, Weill F-X, Nair S, Zhou Z, Sangal V, Krauland MG, Hale JL, Harbottle H, Uesbeck A, et al. Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. *PLoS Pathog.* 2012;8(6):1002776.
33. Medini D, Donati C, Tettelin H, Maignani V, Rappuoli R. The microbial pan-genome. *Curr Opin Genet Dev.* 2005;15(6):589–94.
34. Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics.* 2014;30(1):31–7.
35. Jia B, Raphenya AR, Alcock B, Wagelchner N, Guo P, Tsang KK, Lago BA, Dave BM, Pereira S, Sharma AN, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 2016;44(12):D566–D573.
36. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, Tyson GH, Zhao S, Hsu C-H, McDermott PF, et al. Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrob Agents Chemother.* 2019;63(11):e00483–19.
37. Luhmann N. BlastFrost repository. <https://github.com/nluhmann/BlastFrost>.
38. Luhmann N. Zenodo repository for BlastFrost. <http://doi.org/10.5281/zenodo.4314718>.
39. Luhmann N. Analyses data repository. <https://github.com/nluhmann/BlastFrost/data>.
40. Liu B, Zheng D, Jin Q, Chen L, Yang J. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.* 2018;47(D1):687–92.
41. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ.* 2015;3:1319.
42. Yoshida CE, Kruczkiewicz P, Laing CR, Lingohr EJ, Gannon VP, Nash JH, Taboada EN. The *Salmonella* in silico typing resource (SISTR): an open web-accessible tool for rapidly typing and subtyping draft *Salmonella* genome assemblies. *PLoS ONE.* 2016;11(1):0147101.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.