AllTheBacteria - all bacterial genomes assembled, available and searchable

Martin Hunt[1,2,*], Leandro Lima[1,*], Wei Shen[1,3], John Lees[1], Zamin Iqbal[1,4,+]

[1]European Molecular Biology Laboratory - European Bioinformatics Institute, Hinxton, UK
[2]Nuffield Department of Medicine, University of Oxford, Oxford, UK
[3]Institute for Viral Hepatitis, The Second Affiliated Hospital of Chongqing Medical University, China
[4]Milner Centre for Evolution, University of Bath, UK

[*]these authors contributed equally
[+]Corresponding author, email: zi245@bath.ac.uk

Abstract

The bacterial sequence data publicly available at the global DNA archives is a vast source of information on the evolution of bacteria and their mobile elements. However, most of it is either unassembled or inconsistently assembled and QC-ed. This makes it unsuitable for large-scale analyses, and inaccessible for most researchers to use. In 2021 Blackwell et al therefore released a uniformly assembled set of 661,405 genomes, consisting of all publicly available whole genome sequenced bacterial isolate data as of November 2018, along with various search indexes. In this study we extend that dataset by 4.5 years (up to May 2023), tripling the number of genomes. We also expand the scope, as we begin a global collaborative project to generate annotations for different species as desired by different research communities.

In this study we describe the initial v0.1 data release of 1,932,812 assemblies (combining 1,271,428 new assemblies with the 661k dataset). All 1.9 million have been uniformly re-processed for quality criteria and to give taxonomic abundance estimates with respect to the GTDB phylogeny. Using an evolution-informed compression approach, the full set of genomes is just 102Gb in batched xz archives. We also provide multiple search indexes. Finally, we outline plans for future annotations to be provided in further releases.

Introduction

Bacteria are the dominant cellular organisms on the planet, responsible for the functioning of every biome. As sequencing technology improves and becomes more widely accessible, we are seeing a rapid expansion in the breadth and depth of sequencing of the bacterial domain. These genomes bear the imprint of millions of years of evolution and constitute a priceless resource for the understanding of their biology, dynamics and the effect on the ecology of our entire planet.

Bacterial genomes evolve both through "vertical" inheritance, as parents fission into pairs of children, and through multiple modes of horizontal gene transfer including those mediated by viruses and mobile genetic elements such as plasmids and transposons. This has profound implications for their plasticity and for the flexibility of their genomes. Members of a single bacterial species can share as little as 50% of their genomes (the core genome), the rest being

accessory content, present in only a fraction of the genomes of the species. This "optional extra" content consists of fleetingly present content carried by mobile elements typically purged by selection and therefore rarely observed in the population. It also includes valuable cargo providing vital adaptive traits, observed consistently at intermediate frequencies due to balancing selection. For those who seek to explore the fundamental biology of bacteria, and for those working on clinical microbiology and public health, it is of immense value to be able to study the diversity of bacterial genomes and the dynamics of the functional elements they contain.

Unfortunately, genomes available in the public domain are processed inconsistently or not at all, rendering their use for these purposes inaccessible to most researchers. Even when sequence assemblies are available, specific problems include assembly by a range of different tools and settings; variable quality control (QC); and since many are run together in single projects, there are batch effects caused by blocks of genomes all using the same assembly workflow. As a result, these data are not appropriate for large scale analyses, where uncorrected batch artefacts could masquerade as interesting biology when comparing groups. In order to address this for the community, Blackwell et al[1] set out to uniformly assemble, QC and analyse all bacterial isolate whole genome sequence (WGS) raw data available in the ENA as of November 2018. They released 639,981 high-quality assemblies, along with quality control information and fundamental genome-derived statistics – the most important of which was to check the taxonomic abundance within each putatively single isolate dataset to confirm the species label in the submitted ENA metadata, which is not necessarily sequence-derived. In the process they estimated that 8.1% of the species metadata tags in the ENA were incorrect. They also released multiple search indexes with the assemblies: for whole genome comparison (sourmash[2] and ppsketchlib[3]), and for k-mer search (COBS[4]).

Reflecting upon this initiative, the more successful aspects of the Blackwell dataset (abbreviated to "661k") were as follows. It was, to our knowledge, the first uniformly assembled and rigorously QC-ed set of bacterial genomes that set out to encompass all sequenced bacteria. It included assemblies of over 300,000 genomes which had not previously been available (the raw data only had been available). The assemblies and search indexes allowed multiple other studies of plasmids[5, 6], bacterial adaptation[7, 8, 9, 10], and compression/indexing algorithms[11, 12, 13, 14, 15]. However, there were a few limitations. First, the raw data stored at the INSDC has more than doubled since then, and although we realise that keeping up with publicly deposited sequence data is a never-ending task, an update to the dataset would clearly be of great value. Second, the full set of assemblies we produced was almost 1 terabyte in size, even after compression, and the COBS indexes added a further 900Gb - this reduced the accessibility of the data. Third, we wanted to have the taxonomic abundance QC done based on community-supported GTDB[16]. Fourth, we had not provided further useful information on top of the assemblies: gene annotation, species-specific analyses of wide interest (e.g. serotyping, MLST), or built pan-genomes. However to provide all of this was beyond the capacity or expertise of our own research group – to do this properly and best serve the whole community, we realised that we should involve the research communities who focussed on specific genera/species.

We therefore set up this project, named AllTheBacteria, aiming to update the 661k dataset and improve on the above limitations through a community-centric approach. We advertised the project on Twitter/X and the public microbiology bioinformatics Slack channel and gathered colleagues from across the world keen to work together to produce a valuable public resource.

This paper describes the methodology used for the initial release (v0.1) of assemblies, quality control, taxonomic information and initial search indexes. All software pipelines are open source with permissive licenses, available on GitHub. We also describe the communities which

have joined the project and outline plans for future releases. In terms of the data volume reducing accessibility, Břinda et al recently addressed this issue with a general principle called phylogenetic compression[17] – batching data intelligently based on approximate phylogenetic similarity before compressing shrank the 661k assemblies to 20Gb and the indexes to 100Gb. We are able to follow the same approach with this larger dataset.

Methods

Dataset

We downloaded all Illumina bacterial isolate whole genome sequence raw sequence datasets from the ENA as of May 2023. Reads were downloaded using either prefetch/fasterq-dump from the SRA-toolkit (`https://github.com/ncbi/sra-tools`) or enaDataGet (`https://github.com/enasequence/enaBrowserTools`).

Genome Assembly

The genome assembly pipeline (`https://github.com/leoisl/bacterial_assembly_pipeline`) was the same as that used by Blackwell1, a wrapper around Shovill (`https://github.com/tseemann/shovill`) which is itself a wrapper around Spades[18]. However, we used a slightly later commit (bb1346f); the difference between this commit and that used by Blackwell was not sufficient to justify reassembling the 661k.

Taxonomic abundance estimation

Performing taxonomic analysis on isolate data is considerably simpler than on full metagenomic data - we wanted primarily to establish the major species, its relative abundance, and the nature of contaminants. We therefore ran some simulation experiments with mixtures of different species at different abundances (data not shown here) and determined that sylph[19] was more accurate, faster (∼1 minute per sample) and required less RAM (10Gb of RAM for the whole of GTDB) than the tools we used for the 661k (Kraken/Bracken). We therefore used sylph version 0.5.1 with the pre-built GTDB r214 database (`https://storage.googleapis.com/sylph-stuff/v0.3-c200-gtdb-r214.syldb`) and default options.

A species call was made from the "Genome_file" column of the sylph output, using a lookup table generated with TaxonKit[20] using GTDB taxonomy data (`https://github.com/shenwei356/gtdb-taxdump`, v0.4.0). The reads from 3,252 samples resulted in no output from sylph, presumably because there were no matches to the reference database.

Assembly statistics

The program assembly-stats (`https://github.com/sanger-pathogens/assembly-stats`; git commit 7bdb58b) was run on each assembly to gather basic statistics. Assemblies with a total length of less than 100kbp or greater than 15Mbp were excluded.

CheckM

CheckM2[21] version 1.0.1 was run on each assembly, using the default downloaded database uniref100.KO.1.dmnd. We ran "checkm2 predict" with options `--allmodels --database_path --lowmem uniref100.KO.1.dmnd`. 275 samples did not run to completion, stopping with the error message "No DIAMOND annotation was generated". This suggests that the assemblies are of low quality, resulting in very few predicted proteins.

MiniPhy

All assembly FASTA files were compressed using MiniPhy[17] commit 7abe08c (this tool (`https://github.com/karel-brinda/MiniPhy`) has been renamed - it was called mof-compress in the original preprint), which uses intelligent batching of genomes to improve compression. The process has 2 steps: Divide the genomes into approximately equal-sized batches, typically done by species. In our case, the highest-abundance species for each sample was previously determined using sylph (see above), and a CSV file was created mapping the filename to species. Batches were auto-created using the `create_batches.py` script from the MiniPhy repository. MiniPhy was then run on each batch; internally it created an approximate phylogenetic tree and reordered the genomes for better compression. The output is one xz compressed archive per batch.

pp-sketchlib

The high-quality assemblies were sketched at k=14 using pp-sketchlib v2.1.1. This database allows sequence similarity search through computing a Jaccard index, either against all the contents, or sparse queries returning k-nearest neighbours or below a given threshold. We ran `sketchlib sketch -l 2kk_list.txt --kmer 14 -s 1000 -o 2kk_sketch --cpus 32` to use a sketch size of 1000. The resulting HDF5 database of sketches is 4 GB.

Results

This project extends and builds on the 661k dataset, using the same genome assembly pipeline. Thus, we generated 1,271,428 new assemblies, giving a total of 1,932,812 assemblies when combined with the 661k dataset, along with associated taxonomic abundance estimates and quality statistics. We shifted from using the NCBI taxonomy in the 661k project, to using GTDB here, so reprocessed the sequence reads for all samples (including the 661k) in order to get consistent taxonomic estimates. For different use cases, we expect different levels of quality filtering might be needed, but we provide a file list of 1,857,792 "high-quality" assemblies that pass the following criteria: genome size between 100k and 15Mb, no more than 2000 contigs, N50 at least 5000, majority species at above 99% abundance (and the same majority species call for all INSDC sequencing runs from the same sample), CheckM2-completeness at least 90%, and CheckM2-contamination of no more than 5%.

As expected, the data is dominated by the species of high clinical interest - the top 10 species constitute 77% of the high-quality dataset. However, it contains 10,542 different species whereas the 661k contained just 6,997. A comparison of the number of species in the high-quality data set and in the 661k set is shown in Figure 1.

In order to make the data more accessible (i.e. download-able), it was important to compress the assemblies as efficiently as possible, while remaining lossless, and without requiring users to install any special software. Naively gzipping each assembly in its own fasta resulted in a disk usage of 3 Terabytes. Batching by species into sets of 4000 genomes and then using xz as a compressor dropped this by a factor of 10, to around 380Gb. However applying the MiniPhy tool (renamed, previously mof-compress) to do more effective batching before compression with xz, reduced the disk use to 102Gb.

Discussion

This is release 0.1 of AllTheBacteria, the first step in an open community project. Now that the assemblies and quality data are available, we will move to the next phase of the project,

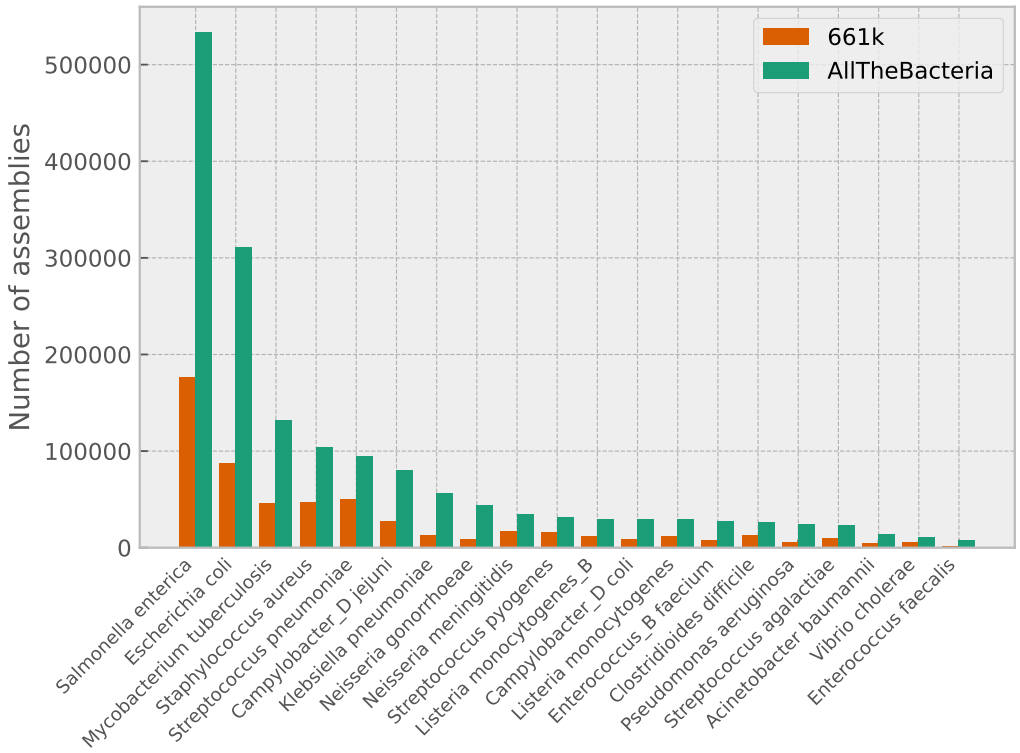

Figure 1: Assembly counts of the 20 most common species in the high quality AllTheBacteria data set, compared with their counts in the 661k set. Species names are from the GTDB.

distributing various work packages to different parts of the community: gene annotation and harmonisation within species, construction of pangenomes (in the standard microbial genomics sense), mobile element and phage analyses, and detection of genes/features of interest to specific research communities. In parallel, there are other indexes we will provide – sourmash, COBS and the snakemake workflow mof-search which combines compressed COBS indexes with minimap to provide full alignment against the full data, on either a laptop or a compute cluster.

We want this data to be of use, and used, and hope for the moment that the assemblies and first search index (ppsketchlib) will be of immediate utility. As we add more products in future releases (more indexes will be first), we will update this preprint.

Data Availability

All assemblies and metadata from Release 0.1 are available here: `https://ftp.ebi.ac.uk/pub/databases/AllTheBacteria/Releases/0.1/`. The assembly pipeline is here: `https://github.com/leoisl/bacterial_assembly_pipeline`.

Author Contributions

Assembly [LL], taxonomic abundance analysis [SW, MH], compression of assemblies and COBs indexes using miniphy [SW], ppsketchlib [JL], all other analyses [MH], planning [LL, MH, SW, JL, ZI], paper writing [ZI, MH, JL].

Acknowledgements

We would like to thank Karel Břinda for help with running MiniPhy, and Daniel Anderson for help with Snakemake. The authorship of this paper is currently very short, as the first phase of this project was completely dependent on the team at EBI/Bath to deliver the assemblies. However many people have volunteered to do future analyses, and their enthusiasm has buoyed us. We would like to thank, for their enthusiasm and probable future contributions: Nabil Fareed-Alikhan, Oliver Schwengers, Laura Carroll, Natacha Couto, Boas van der Putten, Kivumbi Mark Teferi, Sebastian Jaenicke, Conor Meehan, Gultekin Unal, Peter van Heusden, George Bouras, Adrian Cazares, Daniel Cazares, Wendy Figueroa, Michael Hall, Daniel Anderson, Finlay Macguire, Matthew Croxen, Kate Baker, Nick Thomson, Kat Holt, Torsten Seemann and Jo Fothergill.

References

[1] Grace A. Blackwell, Martin Hunt, Kerri M. Malone, Leandro Lima, Gal Horesh, Blaise T. F. Alako, Nicholas R. Thomson, and Zamin Iqbal. Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. *PLOS Biology*, 19(11):e3001421, November 2021.

[2] N. Tessa Pierce, Luiz Irber, Taylor Reiter, Phillip Brooks, and C. Titus Brown. Large-scale sequence comparisons with sourmash. *F1000Research*, 8:1006, July 2019.

[3] John A. Lees, Simon R. Harris, Gerry Tonkin-Hill, Rebecca A. Gladstone, Stephanie W. Lo, Jeffrey N. Weiser, Jukka Corander, Stephen D. Bentley, and Nicholas J. Croucher. Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Research*, 29(2):304–316, February 2019.

[4] Timo Bingmann, Phelim Bradley, Florian Gauger, and Zamin Iqbal. COBS: a Compact Bit-Sliced Signature Index. 2019.

[5] Florent Lassalle, Salah Al-Shalali, Mukhtar Al-Hakimi, Elisabeth Njamkepo, Ismail Mahat Bashir, Matthew J. Dorman, Jean Rauzier, Grace A. Blackwell, Alyce Taylor-Brown, Mathew A. Beale, Adrián Cazares, Ali Abdullah Al-Somainy, Anas Al-Mahbashi, Khaled Almoayed, Mohammed Aldawla, Abdulelah Al-Harazi, Marie-Laure Quilici, François-Xavier Weill, Ghulam Dhabaan, and Nicholas R. Thomson. Genomic epidemiology reveals multidrug resistant plasmid spread between Vibrio cholerae lineages in Yemen. *Nature Microbiology*, 8(10):1787–1798, September 2023.

[6] Ya Hu, Robert A. Moran, Grace A. Blackwell, Alan McNally, and Zhiyong Zong. Fine-Scale Reconstruction of the Evolution of FII-33 Multidrug Resistance Plasmids Enables High-Resolution Genomic Surveillance. *mSystems*, 7(1):e00831–21, February 2022.

[7] Kevin O. Tamadonfar, Gisela Di Venanzio, Jerome S. Pinkner, Karen W. Dodson, Vasilios Kalas, Maxwell I. Zimmerman, Jesus Bazan Villicana, Gregory R. Bowman, Mario F. Feldman, and Scott J. Hultgren. Structure–function correlates of fibrinogen binding by *Acinetobacter* adhesins critical in catheter-associated urinary tract infections. *Proceedings of the National Academy of Sciences*, 120(4):e2212694120, January 2023.

[8] Lewis C. E. Mason, David R. Greig, Lauren A. Cowley, Sally R. Partridge, Elena Martinez, Grace A. Blackwell, Charlotte E. Chong, P. Malaka De Silva, Rebecca J. Bengtsson, Jenny L. Draper, Andrew N. Ginn, Indy Sandaradura, Eby M. Sim, Jonathan R.

Iredell, Vitali Sintchenko, Danielle J. Ingle, Benjamin P. Howden, Sophie Lefèvre, Elisabeth Njamkepo, François-Xavier Weill, Pieter-Jan Ceyssens, Claire Jenkins, and Kate S. Baker. The evolution and international spread of extensively drug resistant Shigella sonnei. *Nature Communications*, 14(1):1983, April 2023.

[9] Michael Biggel, Nadja Jessberger, Jasna Kovac, and Sophia Johler. Recent paradigm shifts in the perception of the role of Bacillus thuringiensis in foodborne disease. *Food Microbiology*, 105:104025, August 2022.

[10] Tracy M Smith, Madison A Youngblom, John F Kernien, Mohamed A Mohamed, Sydney S Fry, Lindsey L Bohr, Tatum D Mortimer, Mary B O'Neill, and Caitlin S Pepperell. Rapid adaptation of a complex trait during experimental evolution of Mycobacterium tuberculosis. *eLife*, 11:e78454, June 2022.

[11] Barış Ekim, Bonnie Berger, and Rayan Chikhi. Minimizer-space de Bruijn graphs: Whole-genome assembly of long reads in minutes on a personal computer. *Cell Systems*, 12(10):958–968.e6, October 2021.

[12] Andrea Cracco and Alexandru I. Tomescu. Extremely fast construction and querying of compacted and colored de Bruijn graphs with GGCAT. *Genome Research*, page genome;gr.277615.122v2, May 2023.

[13] Jamshed Khan, Marek Kokot, Sebastian Deorowicz, and Rob Patro. Scalable, ultra-fast, and low-memory construction of compacted de Bruijn graphs with Cuttlefish 2. *Genome Biology*, 23(1):190, September 2022.

[14] Sebastian Deorowicz, Agnieszka Danek, and Heng Li. AGC: compact representation of assembled genomes with fast queries and updates. *Bioinformatics*, 39(3):btad097, March 2023.

[15] Camille Marchet and Antoine Limasset. Scalable sequence database search using Partitioned Aggregated Bloom Comb-Trees. preprint, Bioinformatics, February 2022.

[16] Donovan H Parks, Maria Chuvochina, Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, 50(D1):D785–D794, January 2022.

[17] Karel Břinda, Leandro Lima, Simone Pignotti, Natalia Quinones-Olvera, Kamil Salikhov, Rayan Chikhi, Gregory Kucherov, Zamin Iqbal, and Michael Baym. Efficient and Robust Search of Microbial Genomes via Phylogenetic Compression. preprint, Bioinformatics, April 2023.

[18] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 19(5):455–477, May 2012.

[19] Jim Shaw and Yun William Yu. Metagenome profiling and containment estimation through abundance-corrected k-mer sketching with sylph. preprint, Bioinformatics, November 2023.

[20] Wei Shen and Hong Ren. TaxonKit: A practical and efficient NCBI taxonomy toolkit. *Journal of Genetics and Genomics*, 48(9):844–850, September 2021.

[21] Alex Chklovski, Donovan H. Parks, Ben J. Woodcroft, and Gene W. Tyson. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nature Methods*, 20(8):1203–1212, August 2023.