AllTheBacteria – all bacterial genomes assembled, available, and searchable

Martin Hunt[1–4,*], Leandro Lima[1,*], Daniel Anderson[1], George Bouras[5,6], Michael Hall[7,8], Jane Hawkey[9], Oliver Schwengers[10], Wei Shen[1,11], John A. Lees[1,+], Zamin Iqbal[1,12,+]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK
[2]Nuffield Department of Medicine, University of Oxford, Oxford, UK
[3]National Institute of Health Research Oxford Biomedical Research Centre, John Radcliffe Hospital, Headley Way, Oxford, UK
[4]Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, University of Oxford, Oxford, UK
[5]Adelaide Medical School, Faculty of Health and Medical Sciences, The University of Adelaide, Adelaide, Australia
[6]The Department of Surgery – Otolaryngology Head and Neck Surgery, Central Adelaide Local Health Network, Adelaide, Australia
[7]Department of Microbiology and Immunology, The University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Australia
[8]Center for Clinical Research, University of Queensland Centre for Clinical Research, Faculty of Medicine, The University of Queensland, Brisbane, QLD, Australia
[9]Department of Infectious Diseases, School of Translational Medicine, Monash University, Melbourne, Victoria 3004, Australia
[10]Bioinformatics and Systems Biology, Justus Liebig University Giessen, Giessen 35392, Germany
[11]Institute for Viral Hepatitis, The Second Affiliated Hospital of Chongqing Medical University, China
[12]Milner Centre for Evolution, University of Bath, UK

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

Abstract

The bacterial sequence data publicly available via the global DNA archives is a vast potential source of information on the evolution of bacteria. However, most of this sequence data is unassembled, or where assembled was done so with no consistent assembler or quality control. Although this data has great potential, these inconsistencies make it unsuitable for large-scale analyses, and inaccessible for most researchers to reuse. Therefore in our previous effort, we released a uniformly assembled set of 661,405 genomes, consisting of all publicly available whole genome sequenced bacterial isolate data up to a cutoff of November 2018, enriched with various search indexes to make the data easier to sort and use. In this study, we first extend the dataset up to August 2024 with the same consistent assembly pipeline, more than tripling the number of genomes available. We also expand the scope of the dataset beyond genomes, as we begin a global collaborative project to generate annotations, species-specific analyses, evolutionary data, new search indices, and protein structural data. Our collaboration is therefore grass-roots, driven by the needs of different research communities within microbiology.

In this paper, we describe the project as of release 2024-08, comprising 2,440,377 assemblies. All 2.4 million genomes have been uniformly reprocessed for quality criteria and to give taxonomic abundance estimates with respect to the GTDB phylogeny. We further enrich the dataset with sequence

annotations from Bakta, antimicrobial resistance predictions from AMRFinderPlus, and AlphaFold2 protein structure predictions for the 17.7M hypothetical proteins. By applying an evolution-informed compression approach, the full set of genomes is just 130Gb: a reduction of ˜23x compared to compressing individual assemblies. To make the resource as accessible as possible, we also provide multiple search indexes, a method for alignment to the full dataset, and cloud-based access to all the genomes.

The AllTheBacteria data (`https://allthebacteria.org/`) has already been independently used in multiple other analyses – our goal is to make this a self-sustaining community-driven resource, which increases the accessiblity and reuse of bacterial genomes for a large range of purposes.

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 true biological differences when comparing groups. In order to address this for the community, Blackwell et al. (2021) 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 (Pierce et al. 2019) and sketchlib (Lees et al. 2019)), and for k-mer search (COBS (Bingmann et al. 2019)).

The assemblies and search indexes allowed multiple other studies of plasmids (Lassalle et al. 2023; Hu et al. 2022), bacterial adaptation (Tamadonfar et al. 2023; Mason et al. 2023; Biggel et al. 2022; Smith et al. 2022), and compression/indexing algorithms (Ekim, Berger, and Chikhi 2021; Cracco and Tomescu 2023; Khan et al. 2022; Deorowicz, Danek, and Li 2023; Marchet and Limasset 2022). However, there remained 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 1Tb in size, even after compression, and the COBS indexes added a further 900Gb - this high computational demand reduced the accessibility of the data. Third, we wanted to have the taxonomic abundance QC metrics – one of the most important for downstream use – reported based on the community-supported GTDB standard (Parks et al. 2022). 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 bring the expertise of research communities who focussed on specific genera/species inside the project.

This paper describes the methodology we used for assembly, QC, and generation of analyses/products. Our processes are summarised in figure 1. All software pipelines are open source with permissive licenses, available on GitHub.

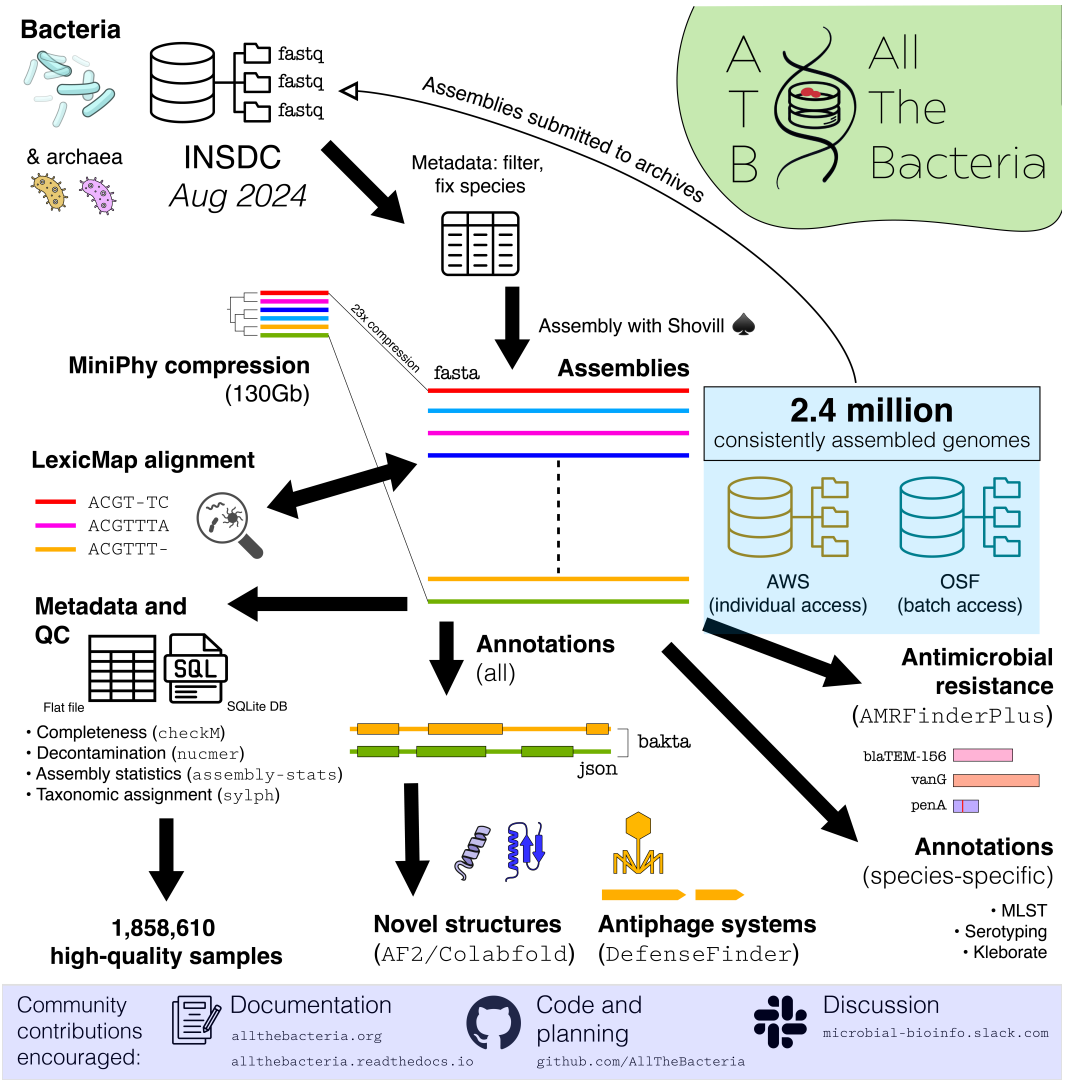

Figure 1: Overview of release 2024-08 of AllTheBacteria, including data sources, analyses, and community aspects of the project.

Methods

Dataset

We downloaded all paired Illumina bacterial isolate whole genome sequence raw sequence metadata from the ENA, using the query `https://www.ebi.ac.uk/ena/portal/api/search?result=read_run&fields=ALL&query=tax_tree(2)&format=tsv`. Samples were processed if they were not in the 661k dataset, and had metadata "instrument_platform" = "ILLUMINA", "library_strategy" == "WGS", "library_source" = "GENOMCIC", and "library_layout" = "PAIRED". The samples were processed in two stages: the first (release 0.2) was from metadata downloaded on June 16th 2023, and then a second round of processing from metadata obtained on August 1st 2024 (incremental release 2024-08).

Genome assembly

The genome assembly pipeline for the 661k dataset used by Blackwell was based around v1.0.4 of Shovill (`https://github.com/tseemann/shovill`) which is a wrapper around Spades (Bankevich et al. 2012). For release 0.2, we refactored and updated the pipeline (`https://github.com/leosl/bacterial_assembly_pipeline`), and used an updated version of Shovill (v1.1.0), however the difference between v1.0.4 and v1.1.0 was minimal and does not impact assembler output, and therefore there was no need to reassemble the existing 661k dataset.

All samples in release 2024-08 were processed using a simple Python script (see `https://github.com/AllTheBacteria/AllTheBacteria/tree/main/reproducibility/All-samples/assembly`), again using Shovill v1.1.0. The script processes one sample, first downloading the reads, then running Sylph, Shovill, and finally removing contigs matching the human genome (as described later).

Taxonomic abundance estimation

Most taxonomic abundance estimation tools are designed for metagenome data which consists of an unknown mix of different taxa. However, in single isolate data, which makes up the entirety of this collection, only a single species is expected to be present, unless the sample is contaminated. Therefore 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 used sylph (Shaw and Yu 2023) version 0.5.1 with the pre-built GTDB r214 database (`https://storage.googleapis.com/sylph-stuff/v0.3-c200-gtdb-r214.syldb`) with default options, which required just 10Gb of RAM and took (~1 minute per sample). Since the 661k dataset had previously been analysed with Kraken/Bracken, we re-downloaded the reads and reprocessed them with sylph.

A species call was made from the "Genome_file" column of the sylph output, using a lookup table generated with TaxonKit (Shen and Ren 2021) 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.

Human decontamination

After assembly, the contigs output by Shovill were matched to the human genome plus HLA sequences using nucmer from version 4.0.0rc1 of the MUMmer package (Marçais et al. 2018). We used the T2T CHM13 version 2 assembly (GCA_009914755.4) of the human genome (Nurk et al. 2022; Rhie et al. 2023). For HLA sequences, we used the file `hla_gen.fasta` from version 3.55.0 of the IPD-IMGT/HLA database (J. Robinson et al. 2000; Dominic J Barker et al. 2023; James Robinson, Dominic

J. Barker, and Marsh 2024). Any contig that had a single match of at least 99% identity and 90% of its length was removed.

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. We found that 21 of the assemblies in the original 661k data set were longer than 15Mbp, and so were removed from our releases, meaning that 661,384 of the samples in AllTheBacteria originate from the 661k dataset.

CheckM

CheckM2 (Chklovski et al. 2023) 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`. 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 (Břinda et al. 2023) commit 7abe08c, which uses intelligent batching of genomes to improve compression. The process has two 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 then compressed with the standard xz tool, to produce one archive file per batch.

sketchlib.rust

The high-quality assemblies were sketched at k=14 using sketchlib.rust v0.1.0. This database allows sequence similarity search through computing a Jaccard index, either against all the contents, sparse queries returning k-nearest neighbours, below a given distance threshold, or against a chosen subset of queries. We ran `sketchlib sketch -f 2kk_list.txt -k 14 -s 1000 -o 2kk_sketch --threads 32` to use a sketch size of 1000. The resulting .skd database of sketches is 4.1 GB, and .skm of metadata is 123 MB.

Antimicrobial resistance detection

Antimicrobial resistance determinants were identified using AMRFinderPlus (Feldgarden et al. 2021) v3.12.8 on all assembly FASTA files with database version v2024-01-31.1. For appropriate species (*Acinetobacter baumannii, Burkholderia cepacia, Burkholderia pseudomallei, Campylobacter jejuni, Campylobacter coli, Citrobacter freundii, Clostridioides difficile, Enterobacter cloacae, Enterobacter asburiae, Enterococcus faecalis, Enterococcus faecium, Enterococcus hirae, Escherichia, Shigella, Klebsiella aerogenes, Klebsiella oxytoca, Klebsiella pneumopniae, Klebsiella pneumopniae species complex, Neisseria gonorrhoeae, Neisseria meningitidis, Pseudomonas aeruginosa, Salmonella, Serratia marcescens, Staphylococcus aureus, Staphylococcus pseudintermedius, Streptococcus agalactiae, Streptococcus mitis, Streptococcus pneumoniae, Streptococcus pyogenes, Vibrio cholerae, Vibrio vulnificus, Vibrio parahaemolyticus*) we used the GTDB species assigned by sylph for the AMRFinderPlus

`--organism` parameter, in accordance with the guidelines at `https://github.com/ncbi/amr/wiki/Running-AMRFinderPlus#--organism-option` (git commit `5f27bbe`), thus incorporating known AMR-informative point mutations. This option was omitted for all other species.

Genome annotation

All assembly FASTA files were annotated using Bakta (Schwengers et al. 2021) v1.9.4 and its full database type v5.1. For consistency and downstream interoperability reasons, the `--keep-contig-headers` option was set; all other parameters were left to default values. To reduce the total amount of genome annotation data (down from 35Tb), a two-step approach was carried out. First, only Bakta JSON result files were kept, as standard output files can be reconstructed using Bakta's `bakta_io` command. Second, all annotated genome files were then grouped into taxonomic batches as explained above and compressed using `xz -9`. The resulting compressed annotation data was thereby reduced in size to 1.5TB.

Protein Structure Prediction of Hypothetical Proteins

Of the 9,319,300,441 proteins in the Bakta annotation output for the 2024-08 release of ATB, we generated protein structures for those that were annotated as hypothetical proteins. After de-duplication of identical sequences, 31,929,327 remained. We then kept only proteins from accessions that passed all quality control checks (as of July 2025), leaving 17,711,165 unique proteins under 3000 amino acids in length. Protein structures were generated using ColabFold (Mirdita et al. 2022) v1.5.5. Multiple sequence alignments were generated with MMSeqs2 (Steinegger and Söding 2017) v15.6f452 using the uniref30 and environmental (i.e. ColabFoldDB) databases. Protein structure predictions were generated using AMD MI250x GPUs on Setonix at the Pawsey Supercomputing Research Centre using the AlphaFold2 (Jumper et al. 2021) ptm model 1 only, with 3 recycles, no templates or relaxation for maximum throughout.

Results

We set up this project, named AllTheBacteria (ATB), aiming to update the 661k dataset and improve on the previous limitations through a community-centric approach. We advertised the project on social media and the public microbiology bioinformatics Slack channel (`microbial-bioinfo.slack.com`) and gathered colleagues from across the world keen to work together to produce a valuable public resource. To avoid the need for unsustainable downloads from public archives, processing of raw read data into assemblies was centralised, and published in phases as results became available. This allowed community members to reuse these assemblies for downstream tasks, and also prototype their pipelines on subsets of the entire dataset. Our community members volunteered analyses based on their research interests and expertise, and when complete we have incorporated these into the public data releases. We also created a github organisation to collate and organise requests for new analysis ideas, and species-specific tools. We currently have over twenty people actively contributing in this way. We are also aware of reuses of the entire dataset in algorithm development, outside of the project community (and outside of microbiology), which was also one of our aims.

This project extends and builds on the 661k dataset, using the same genome assembly pipeline. We generated 1,778,993 new assemblies, giving a total of 2,440,377 assemblies when combined with the 661k dataset, along with associated taxonomic abundance estimates and quality statistics. We made these data available as archives for those that need the entire resource, on the cloud for those that need individual genomes, and on the ENA for long-term archiving and to link to the underlying data and generating studies.

We shifted from using the NCBI taxonomy in the 661k project, to using the community-preferred GTDB here, so reprocessed the sequence reads for all samples (including the 661k) in order to obtain

consistent taxonomic estimates. For different use cases, we expect different levels of quality filtering might be needed, so provide all assemblies and statistics computed from them. For ease of use, we provide a file listing the 2,346,079 assemblies we deem "high-quality": 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 ten species constitute 75% of the high-quality dataset. However, AllTheBacteria expands the number of species from 7,003 to 11,824. A comparison of the number of species in the high-quality data set and in the 661k set is shown in Figure 2.

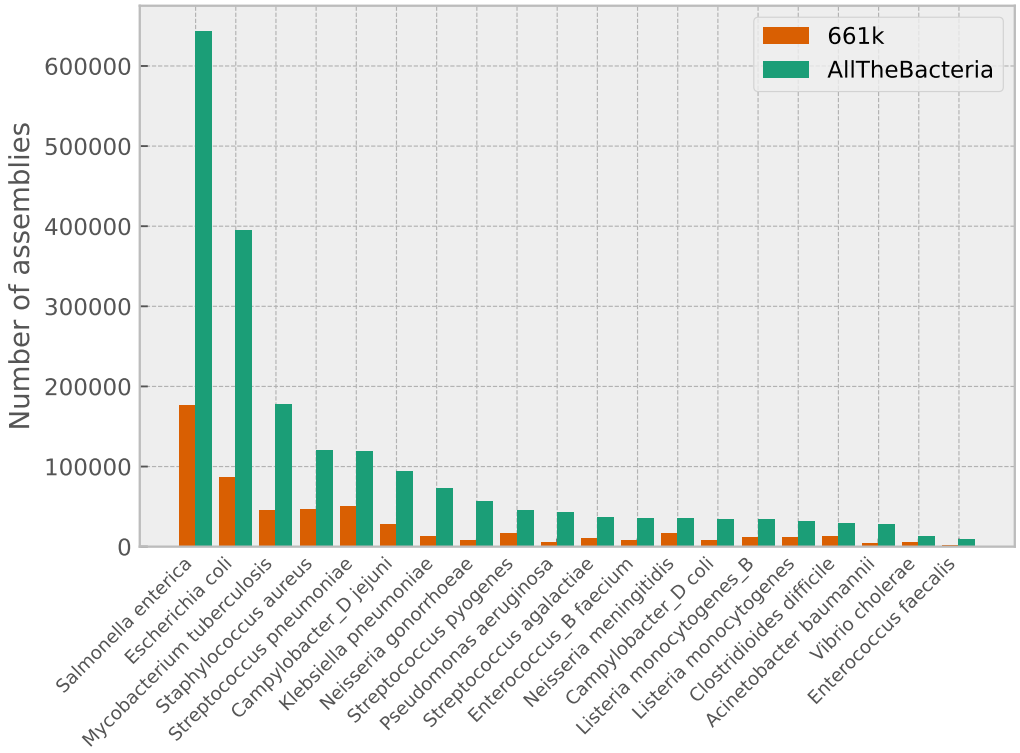

Figure 2: Assembly counts of the 20 most common species in the high quality AllTheBacteria data set, compared with their counts in the 661k set. Species names are from GTDB taxonomy, assigned by Sylph.

In order to make the data more accessible more widely, we particularly focussed on making the dataset searchable, accessible as individual assemblies, and making the entire dataset as small as possible. This decreases demand on network, storage and computing requirements for all downstream users. It was important to losslessly compress the assemblies as efficiently as possible, and without requiring users to install any special software. Naively applying gzip to each assembly in its own fasta file resulted in a disk usage of 3.9 Terabytes. Applying the MiniPhy tool (Břinda et al. (2023)) to intelligently batch sets of assemblies with related sequences, then applying compression with the standard xz tool, reduced the disk use to 130Gb, a 23-fold reduction.

Genome annotations were successfully conducted for 2,438,287 assemblies, resulting in a total of 9,333,646,492 predicted coding sequences (CDS). Among these, 240,369,332 were non-redundant due to their MD5 hash values, and 8,731,879,018 (94%) were exactly identified and linked to UniProt UniRef100 clusters (i.e. exactly identical). 450,878,009 and 66,907,569 were annotated and linked

to UniProt (Consortium 2024) UniRef90 and UniRef50 clusters, respectively. 272,946,667 remaining proteins could not be assigned any functional annotation and were annotated as hypothetical proteins.

Protein structures were generated for 17,708,939 unique proteins annotated as hypothetical by Bakta. Of these, 16,044,432 (91%) were not identical to a representative in the UniProt 202503 release, while 12,004,534 (68%) did not share 90% sequence similarity over 80% coverage (calculated via MMSeqs2) to any UniProt (release 202503) protein. The proteins with a structure prediction generated had an average length of 201 amino acids. 8,619,588 (49%) had a mean pLDDT (predicted Local Distance Difference Test) of at least 70, suggesting a good quality structure prediction.

Discussion

The goal of AllTheBacteria is to generate uniformly assembled, quality controlled, annotated resource encompassing all sequenced bacteria. To sustain and enrich the resource, we have built a community around it who create and share added-value analyses. The current total number of samples we include in AllTheBacteria is 2,440,377, bringing us up to date with all INSDC bacterial (and archaeal) Illumina data up to August 2024. We provide methods to make the data searchable, including a sketchlib search index and LexicMap index. Sketchlib allows finding nearest neighbours, tree-building and epidemiological analyses of the dataset with a much reduced representation. LexicMap (Wei Shen and Iqbal 2024) allows BLAST-like query alignment for sequences at least 500bp in length against the full dataset, with very low RAM requirements (1-2Gb) and extremely quickly – seconds for a rare gene with a few tens of thousand hits, to minutes for a 16s gene which requires alignment to almost every one of the 2 million genomes. The trade-off is that the LexicMap index is relatively large (around 3Tb); since this would be impractical to download, we do not provide it on OSF, but instead recommend downloading the assemblies (130Gb) and recreating the index locally, which is faster. The index is available on AWS, and we provided detailed instructions on how to run searches using cloud resources, so no local compute is required.

Future work our community is working on includes deeper annotation, including of phage and plasmids, harmonisation of gene annotation to provide consistent identifiers within a species, and thereby construction of pangenomes. Our community incorporates researchers focused on more targeted analyses, both focused on individual species and classes of analyses, which will be incorporated in future releases. Examples of ongoing analyses include sequence typing (MLST), serotyping, and antiviral defence system determination (DefenceFinder).

Although they are obviously not bacteria, we have also applied the same assembly process to all (Illumina) sequenced archaea as of July 31st 2024 (n=815), and make the assemblies also available at OSF. Although presently a small dataset, we hope that by showing our tools can be applied to other domains of life we can drive community efforts to improve data reuse and accessibility. Fungal genomes would be an obvious target for future efforts in this space.

At time of writing, AllTheBacteria has already been used for development of new bioinformatic tools (Li 2024; Vicedomini et al. 2025), for discovery of a new biosynthetic gene cluster (McCartney and Hoyles 2025), and to study the global dissemination of a drug resistance gene mediated by a mobile element (Serna et al. 2025). We want these data to be of use - please use them and publish with them. As our collaborative network continues to grow, we envisage generation of progressively more valuable analytic outputs for the research community, as well as triggering innovation in search index methods.

We are continuing to assemble genomes from INSDC as they become available — the next release will contain genomes up to May 2025 – with the ambition to continue this effort, extend to long read sequencing, and fungal genomes.

Data Availability

A website linking to all resources is available at `https://allthebacteria.org/`. Documentation for AllTheBacteria is available at `https://allthebacteria.readthedocs.io/en/latest/`. All data for AllTheBacteria are hosted on the Open Science Framework (OSF) here: `https://osf.io/xv7q9/`. The assembly pipelines for release 0.2 and 2024-08 are at `https://github.com/leoisl/bacterial_assembly_pipeline` and `https://github.com/AllTheBacteria/AllTheBacteria/tree/main/reproducibility/All-samples/assembly`.

In addition, individual assembly FASTA files for each sample are available on AWS with S3 URI of the form `s3://allthebacteria-assemblies/<SAMPLE_ID>.fa.gz` (for example `s3://allthebacteria-assemblies/SAMD00000344.fa.gz`). Assemblies are also available through the ENA, as third-party annotations (TPAs) with ERZ prefixes.

A LexicMap index of the full dataset is also available on AWS (see `https://allthebacteria.org/docs/` for details).

Author Contributions

MH1 refers to Martin Hunt. MH2 refers to Michael Hall. Assembly [LL], taxonomic abundance analysis [SW, MH1], compression of assemblies and COBs indexes using miniphy [SW,MH1], AMR analysis [DA, JH], sketchlib [JL], LexicMap index [WS], gene annotation [OS], protein structure prediction [OS,GB], all other analyses [MH1], planning [LL, MH1, SW, JL, ZI], paper writing [ZI, MH1, JL, DA]. Set-up of Amazon AWS data [MH2, MH1, JL]. Community and website maintenance [MH1, JL, ZI]

Acknowledgements

We would like to thank Karel Břinda for help with running MiniPhy. The authorship of this paper is currently very short, as the first phase of this project was originally dependent on the team at EMBL-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, 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, Finlay Macguire, Matthew Croxen, Kate Baker, Nick Thomson, Kat Holt, Torsten Seemann and Jo Fothergill. We are grateful to Amazon AWS for an AWS Open Data Sponsorship Award which funds hosting of assembly data and a LexicMap index. We gratefully acknowledge provision of computing resources of the de.NBI cloud by the BiGi Service Center (grant W-de.NBI-010). This work was supported with the assistance of resources and services from Phoenix HPC at the University of Adelaide and Pawsey Supercomputing Research Centre, which is supported by the Australian Government. We would like to thank Fabien Voisin and Sarah Beecroft for their assistance in operating ColabFold at scale at Phoenix and Pawsey respectively, with extra acknowledgement to Sarah for containerising ColabFold for use on Setonix's AMD GPUs.

Funding

M.H1., L.L., D.A., W.S., J.A.L. and Z.I. were supported by the European Molecular Biology Laboratory. This work was also supported by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at Oxford University in partnership with the UK Health Security Agency (NIHR200915), and the NIHR Biomedical Research Centre, Oxford. The views expressed are those of the authors and not necessarily those of the

NHS, the NIHR, the Department of Health or the UK Health Security Agency. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Bankevich, Anton et al. (May 2012). "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing". en. In: *Journal of Computational Biology* 19.5, pp. 455–477. ISSN: 1066-5277, 1557-8666. DOI: 10.1089/cmb.2012.0021. URL: http://www.liebertpub.com/doi/10.1089/cmb.2012.0021 (visited on 03/08/2024).

Barker, Dominic J et al. (Jan. 2023). "The IPD-IMGT/HLA Database". en. In: *Nucleic Acids Research* 51.D1, pp. D1053–D1060. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkac1011. URL: https://academic.oup.com/nar/article/51/D1/D1053/6814448 (visited on 11/05/2024).

Biggel, Michael et al. (Aug. 2022). "Recent paradigm shifts in the perception of the role of Bacillus thuringiensis in foodborne disease". en. In: *Food Microbiology* 105, p. 104025. ISSN: 07400020. DOI: 10.1016/j.fm.2022.104025. URL: https://linkinghub.elsevier.com/retrieve/pii/S0740002022000491 (visited on 03/08/2024).

Bingmann, Timo et al. (2019). "COBS: a Compact Bit-Sliced Signature Index". In: DOI: 10.48550/ARXIV.1905.09624. URL: https://arxiv.org/abs/1905.09624 (visited on 03/08/2024).

Blackwell, Grace A. et al. (Nov. 2021). "Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences". en. In: *PLOS Biology* 19.11. Ed. by William P. Hanage, e3001421. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.3001421. URL: https://dx.plos.org/10.1371/journal.pbio.3001421 (visited on 03/08/2024).

Břinda, Karel et al. (Apr. 2023). *Efficient and Robust Search of Microbial Genomes via Phylogenetic Compression*. en. preprint. Bioinformatics. DOI: 10.1101/2023.04.15.536996. URL: http://biorxiv.org/lookup/doi/10.1101/2023.04.15.536996 (visited on 03/08/2024).

Chklovski, Alex et al. (Aug. 2023). "CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning". en. In: *Nature Methods* 20.8, pp. 1203–1212. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-023-01940-w. URL: https://www.nature.com/articles/s41592-023-01940-w (visited on 03/08/2024).

Consortium, The UniProt (Nov. 2024). "UniProt: the Universal Protein Knowledgebase in 2025". In: *Nucleic Acids Research* 53.D1, pp. D609–D617. ISSN: 1362-4962. DOI: 10.1093/nar/gkae1010. eprint: https://academic.oup.com/nar/article-pdf/53/D1/D609/60719276/gkae1010.pdf. URL: https://doi.org/10.1093/nar/gkae1010.

Cracco, Andrea and Alexandru I. Tomescu (May 2023). "Extremely fast construction and querying of compacted and colored de Bruijn graphs with GGCAT". en. In: *Genome Research*, genome, gr.277615.122v2. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.277615.122. URL: http://genome.cshlp.org/lookup/doi/10.1101/gr.277615.122 (visited on 03/08/2024).

Deorowicz, Sebastian, Agnieszka Danek, and Heng Li (Mar. 2023). "AGC: compact representation of assembled genomes with fast queries and updates". en. In: *Bioinformatics* 39.3. Ed. by Tobias Marschall, btad097. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btad097. URL: https://academic.oup.com/bioinformatics/article/doi/10.1093/bioinformatics/btad097/7067744 (visited on 03/08/2024).

Ekim, Barış, Bonnie Berger, and Rayan Chikhi (Oct. 2021). "Minimizer-space de Bruijn graphs: Whole-genome assembly of long reads in minutes on a personal computer". en. In: *Cell Systems* 12.10, 958–968.e6. ISSN: 24054712. DOI: 10.1016/j.cels.2021.08.009. URL: https://linkinghub.elsevier.com/retrieve/pii/S240547122100332X (visited on 03/08/2024).

Feldgarden, Michael et al. (June 2021). "AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence". In: *Scientific Reports* 11.1, p. 12728. DOI: 10.1038/s41598-021-91456-0.

Hu, Ya et al. (Feb. 2022). "Fine-Scale Reconstruction of the Evolution of FII-33 Multidrug Resistance Plasmids Enables High-Resolution Genomic Surveillance". en. In: *mSystems* 7.1. Ed. by Robert G. Beiko, e00831–21. ISSN: 2379-5077. DOI: 10.1128/msystems.00831-21. URL: https://journals.asm.org/doi/10.1128/msystems.00831-21 (visited on 03/08/2024).

Jumper, John et al. (Aug. 2021). "Highly accurate protein structure prediction with AlphaFold". en. In: *Nature* 596.7873, pp. 583–589. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2. URL: https://www.nature.com/articles/s41586-021-03819-2 (visited on 11/25/2023).

Khan, Jamshed et al. (Sept. 2022). "Scalable, ultra-fast, and low-memory construction of compacted de Bruijn graphs with Cuttlefish 2". en. In: *Genome Biology* 23.1, p. 190. ISSN: 1474-760X. DOI: 10.1186/s13059-022-02743-6. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-022-02743-6 (visited on 03/08/2024).

Lassalle, Florent et al. (Sept. 2023). "Genomic epidemiology reveals multidrug resistant plasmid spread between Vibrio cholerae lineages in Yemen". en. In: *Nature Microbiology* 8.10, pp. 1787–1798. ISSN: 2058-5276. DOI: 10.1038/s41564-023-01472-1. URL: https://www.nature.com/articles/s41564-023-01472-1 (visited on 03/08/2024).

Lees, John A. et al. (Feb. 2019). "Fast and flexible bacterial genomic epidemiology with PopPUNK". en. In: *Genome Research* 29.2, pp. 304–316. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.241455.118. URL: http://genome.cshlp.org/lookup/doi/10.1101/gr.241455.118 (visited on 03/08/2024).

Li, Heng (Dec. 2024). "BWT construction and search at the terabase scale". en. In: *Bioinformatics* 20. ISSN: 12. DOI: 10.1093/bioinformatics/btae717. URL: https://academic.oup.com/bioinformatics/article/40/12/btae717/7912338.

Marçais, Guillaume et al. (Jan. 2018). "MUMmer4: A fast and versatile genome alignment system". en. In: *PLOS Computational Biology* 14.1, e1005944. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005944. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005944 (visited on 11/05/2024).

Marchet, Camille and Antoine Limasset (Feb. 2022). *Scalable sequence database search using Partitioned Aggregated Bloom Comb-Trees*. en. preprint. Bioinformatics. DOI: 10.1101/2022.02.11.480089. URL: http://biorxiv.org/lookup/doi/10.1101/2022.02.11.480089 (visited on 03/08/2024).

Mason, Lewis C. E. et al. (Apr. 2023). "The evolution and international spread of extensively drug resistant Shigella sonnei". en. In: *Nature Communications* 14.1, p. 1983. ISSN: 2041-1723. DOI: 10.1038/s41467-023-37672-w. URL: https://www.nature.com/articles/s41467-023-37672-w (visited on 03/08/2024).

McCartney, Anne L. and Lesley Hoyles (Nov. 2025). "Host interactions of bioactive molecules produced by Klebsiella spp". en. In: *Microbiota and Host* 3. ISSN: 1. DOI: 10.1530/MAH-24-0011. URL: https://mah.bioscientifica.com/view/journals/mah/3/1/MAH-24-0011.xml (visited on 05/10/2025).

Mirdita, Milot et al. (June 2022). "ColabFold: making protein folding accessible to all". en. In: *Nature Methods* 19.6, pp. 679–682. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01488-1. URL: https://www.nature.com/articles/s41592-022-01488-1 (visited on 11/25/2023).

Nurk, Sergey et al. (Apr. 2022). "The complete sequence of a human genome". en. In: *Science* 376.6588, pp. 44–53. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.abj6987. URL: https://www.science.org/doi/10.1126/science.abj6987 (visited on 11/05/2024).

Parks, Donovan H et al. (Jan. 2022). "GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy". en. In: *Nucleic Acids Research* 50.D1, pp. D785–D794. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkab776. URL: https://academic.oup.com/nar/article/50/D1/D785/6370255 (visited on 03/08/2024).

Pierce, N. Tessa et al. (July 2019). "Large-scale sequence comparisons with sourmash". en. In: *F1000Research* 8, p. 1006. ISSN: 2046-1402. DOI: 10.12688/f1000research.19675.1. URL: https://f1000research.com/articles/8-1006/v1 (visited on 03/08/2024).

Rhie, Arang et al. (Sept. 2023). "The complete sequence of a human Y chromosome". en. In: *Nature* 621.7978, pp. 344–354. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-023-06457-y. URL: https://www.nature.com/articles/s41586-023-06457-y (visited on 11/05/2024).

Robinson, J. et al. (Mar. 2000). "IMGT/HLA Database – a sequence database for the human major histocompatibility complex". en. In: *Tissue Antigens* 55.3, pp. 280–287. ISSN: 0001-2815, 1399-0039. DOI: 10.1034/j.1399-0039.2000.550314.x. URL: https://onlinelibrary.wiley.com/doi/10.1034/j.1399-0039.2000.550314.x (visited on 11/05/2024).

Robinson, James, Dominic J. Barker, and Steven G. E. Marsh (June 2024). "25 years of the IPD-IMGT/HLA database". en. In: *HLA* 103.6, e15549. ISSN: 2059-2302, 2059-2310. DOI: 10.1111/tan.15549. URL: https://onlinelibrary.wiley.com/doi/10.1111/tan.15549 (visited on 11/05/2024).

Schwengers, Oliver et al. (Nov. 2021). "Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification". en. In: *Microbial genomics* 7.11. ISSN: 2057-5858". DOI: 10.1099/mgen.0.000685.

Serna, Carlos et al. (Nov. 2025). "Global dissemination of npmA mediated pan-aminoglycoside resistance via a mobile genetic element in Gram-positive bacteria". en. In: *Nature Communications* 16. ISSN: 1. DOI: 10.1038/s41467-025-61152-y. URL: https://www.nature.com/articles/s41467-025-61152-y.

Shaw, Jim and Yun William Yu (Nov. 2023). *Metagenome profiling and containment estimation through abundance-corrected k-mer sketching with sylph*. en. preprint. Bioinformatics. DOI: 10.1101/2023.11.20.567879. URL: http://biorxiv.org/lookup/doi/10.1101/2023.11.20.567879 (visited on 03/08/2024).

Shen, Wei and Hong Ren (Sept. 2021). "TaxonKit: A practical and efficient NCBI taxonomy toolkit". en. In: *Journal of Genetics and Genomics* 48.9, pp. 844–850. ISSN: 16738527. DOI: 10.1016/j.jgg.2021.03.006. URL: https://linkinghub.elsevier.com/retrieve/pii/S1673852721000837 (visited on 03/08/2024).

Smith, Tracy M et al. (June 2022). "Rapid adaptation of a complex trait during experimental evolution of Mycobacterium tuberculosis". en. In: *eLife* 11, e78454. ISSN: 2050-084X. DOI: 10.7554/eLife.78454. URL: https://elifesciences.org/articles/78454 (visited on 03/08/2024).

Steinegger, Martin and Johannes Söding (Nov. 2017). "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets". en. In: *Nature Biotechnology* 35.11, pp. 1026–1028. ISSN: 1546-1696. DOI: 10.1038/nbt.3988. URL: https://www.nature.com/articles/nbt.3988 (visited on 11/25/2023).

Tamadonfar, Kevin O. et al. (Jan. 2023). "Structure–function correlates of fibrinogen binding by *Acinetobacter* adhesins critical in catheter-associated urinary tract infections". en. In: *Proceedings of the National Academy of Sciences* 120.4, e2212694120. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2212694120. URL: https://pnas.org/doi/10.1073/pnas.2212694120 (visited on 03/08/2024).

Vicedomini, Riccardo et al. (Feb. 2025). "MUSET: set of utilities for constructing abundance unitig matrices from sequencing data". en. In: *Bioinformatics* 41. ISSN: 3. DOI: 10.1093/bioinformatics/btaf054. URL: https://academic.oup.com/bioinformatics/article/41/3/btaf054/7997265.

Wei Shen, John Lees and Zamin Iqbal (Aug. 2024). *LexicMap: efficient sequence alignment against millions of prokaryotic genomes*. en. DOI: 10.1101/2024.08.30.610459. URL: https://www.biorxiv.org/content/10.1101/2024.08.30.610459v2 (visited on 06/25/2025).