




## Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences

Grace A. Blackwell, Martin Hunt, Kerri M. Malone, Leandro Lima, Gal Horesh, Blaise T. F. Alako, Nicholas R. Thomson ,  
Zamin Iqbal  

Published: November 9, 2021 • <https://doi.org/10.1371/journal.pbio.3001421>

Article	Authors	Metrics	Comments	Media Coverage	Peer Review
					

### Abstract

Introduction  
Results  
Discussion  
Methods  
Supporting information  
Acknowledgments  
References

Reader Comments (0)  
Figures

### Abstract

The open sharing of genomic data provides an incredibly rich resource for the study of bacterial evolution and function and even anthropogenic activities such as the widespread use of antimicrobials. However, these data consist of genomes assembled with different tools and levels of quality checking, and of large volumes of completely unprocessed raw sequence data. In both cases, considerable computational effort is required before biological questions can be addressed. Here, we assembled and characterised 661,405 bacterial genomes retrieved from the European Nucleotide Archive (ENA) in November of 2018 using a uniform standardised approach. Of these, 311,006 did not previously have an assembly. We produced a searchable COmpact Bit-sliced Signature (COBS) index, facilitating the easy interrogation of the entire dataset for a specific sequence (e.g., gene, mutation, or plasmid). Additional MinHash and pp-sketch indices support genome-wide comparisons and estimations of genomic distance. Combined, this resource will allow data to be easily subset and searched, phylogenetic relationships between genomes to be quickly elucidated, and hypotheses rapidly generated and tested. We believe that this combination of uniform processing and variety of search/filter functionalities will make this a resource of very wide utility. In terms of diversity within the data, a breakdown of the 639,981 high-quality genomes emphasised the uneven species composition of the ENA/public databases, with just 20 of the total 2,336 species making up 90% of the genomes. The overrepresented species tend to be acute/common human pathogens, aligning with research priorities at different levels from individual interests to funding bodies and national and global public health agencies.

• <https://doi.org/10.1371/journal.pbio.3001421>

# 661K snapshot

## Download and assembly

Paired end reads (Nov 2018)

Source == Genomic

Platform == Illumina

Species estimation

Kraken & Bracken

Assembly

Shovill (SPAdes)

## Quality control

Remove contigs < 200 bp

Quast

Genome size

Number of contigs

CheckM

Completeness

Contamination

## Characterisation

MLST

Phylotype (*E.coli*)

Serovar (*Salmonella*)

AMRFinder

Plasmidfinder

## Searchable

COBS index (~BIGSI)

- DNA searches

minHash sketches

- assembly comparison

pp-sketch index

-distance estimation


# Searchable index: COBS

## K-mer based sequence query

Example: 4-mer

If database contains  
And we query

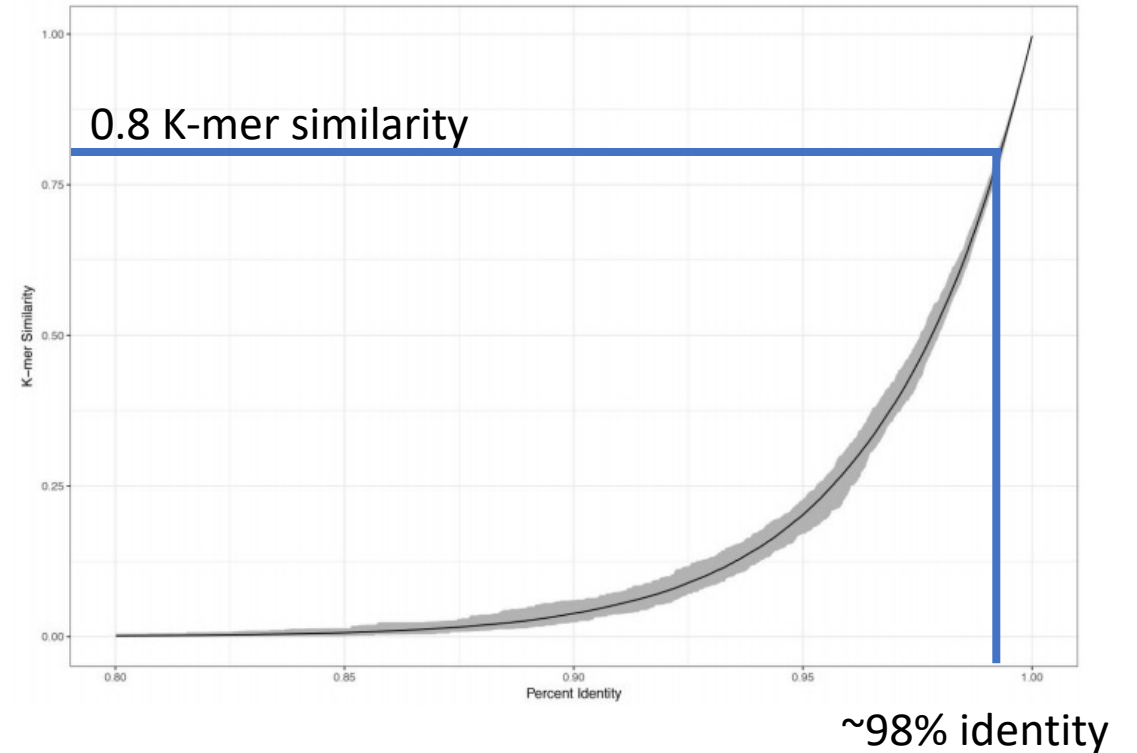
**AAGCGCTTTC**  
**AAGCTCTTTC**



COBS returns:  $3/7 = 43\%$  K-mers present

or

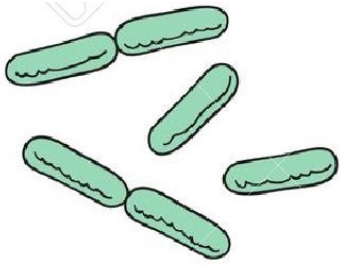
0.43 K-mer similarity



Timo Bingmann *et al.* COBS: a Compact Bit-Sliced Signature Index. In: *26th International Symposium on String Processing and Information Retrieval (SPIRE)*. pages 285-303. Springer. October 2019. preprint arXiv:1905.09624.

Phelim Bradley *et al.* Ultrafast search of all deposited bacterial and viral genomic data. *Nature Biotechnology* 37, 2019.

# Identifying close relatives: minHash index



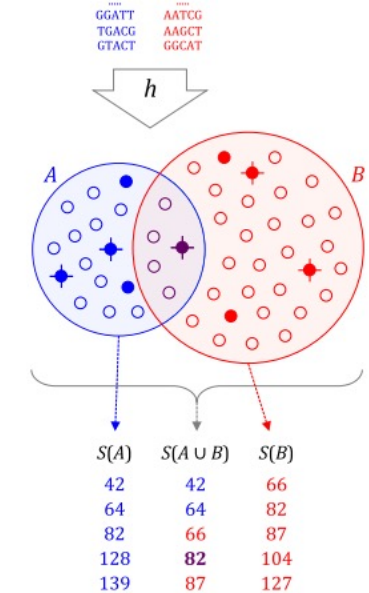
```
[gblackwell@hx-noah-54-02 667_cobs]$ sourmash search A320.fasta.sig 661K_sourmash_index_scaled.sbt.zip -n 20
```

```
== This is sourmash version 3.5.0. ==
== Please cite Brown and Irber (2016), doi:10.21105/joss.00027. ==
```

```
selecting default query k=31.
loaded query: A320.fasta... (k=31, DNA)
loaded 1 databases.
```

```
2502 matches; showing first 20:
similarity match
```

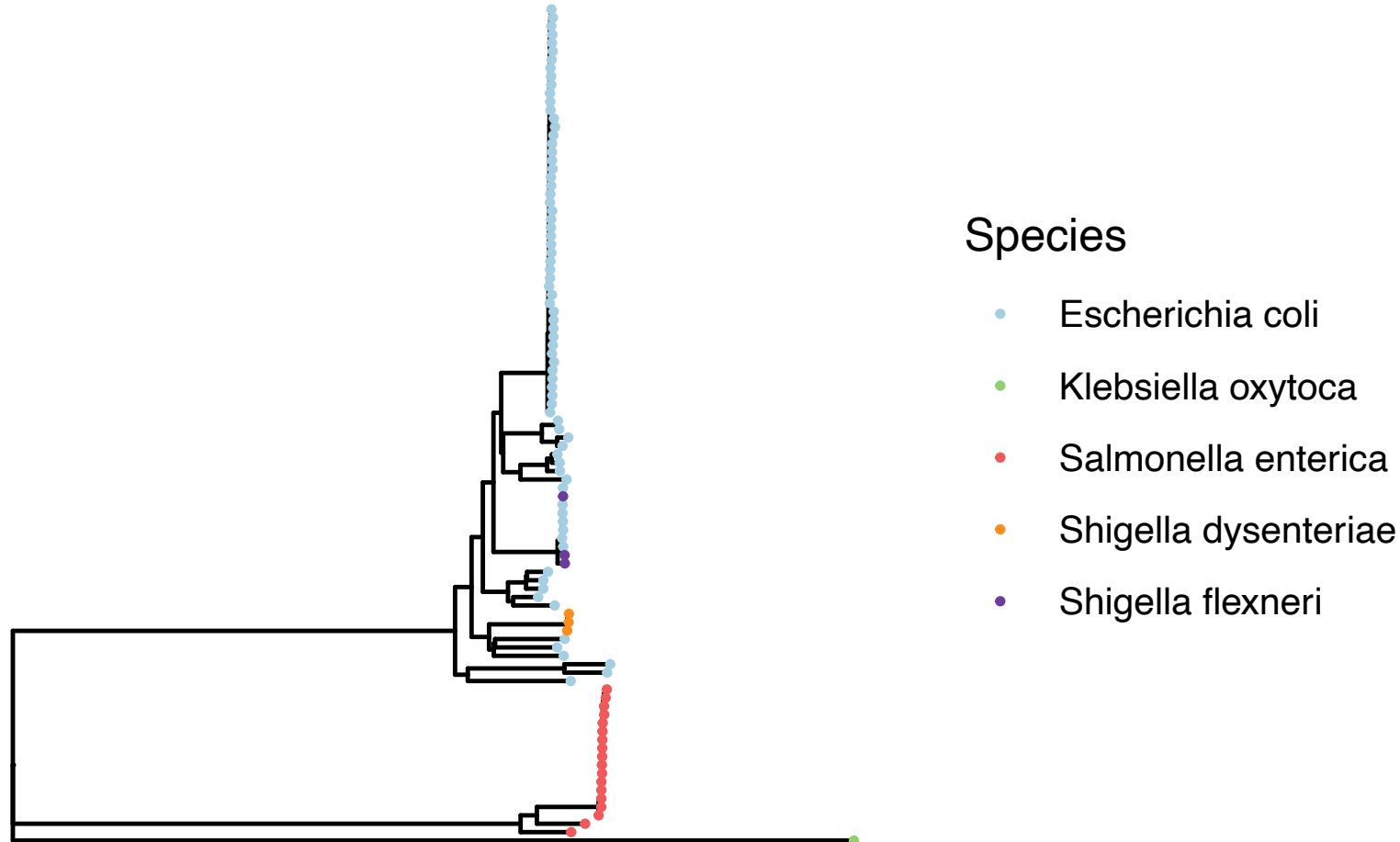
```
-----
99.8% ../split_10K_063_dir/chunk_017/SAMEA54076918.contigs.fa.gz
99.8% ../split_10K_066_dir/chunk_016/SAMEA1466041.contigs.fa.gz
99.6% ../split_10K_039_dir/chunk_054/SAMN01828145.contigs.fa.gz
98.6% ../split_10K_066_dir/chunk_015/SAMEA1466076.contigs.fa.gz
98.2% ../split_10K_066_dir/chunk_015/SAMEA1466068.contigs.fa.gz
97.2% ../split_10K_063_dir/chunk_017/SAMEA54071668.contigs.fa.gz
95.0% ../split_10K_039_dir/chunk_066/SAMN01087912.contigs.fa.gz
94.4% ../split_10K_063_dir/chunk_025/SAMEA1876504.contigs.fa.gz
94.0% ../split_10K_063_dir/chunk_017/SAMEA54070918.contigs.fa.gz
94.0% ../split_10K_058_dir/chunk_096/SAMEA1709897.contigs.fa.gz
93.6% ../split_10K_063_dir/chunk_017/SAMEA54068668.contigs.fa.gz
93.4% ../split_10K_063_dir/chunk_017/SAMEA54070168.contigs.fa.gz
93.4% ../split_10K_063_dir/chunk_017/SAMEA54067918.contigs.fa.gz
92.8% ../split_10K_063_dir/chunk_017/SAMEA54069418.contigs.fa.gz
92.6% ../split_10K_058_dir/chunk_095/SAMEA1709934.contigs.fa.gz
92.4% ../split_10K_065_dir/chunk_069/SAMEA1465991.contigs.fa.gz
92.4% ../split_10K_035_dir/chunk_017/SAMN04555294.contigs.fa.gz
92.4% ../split_10K_035_dir/chunk_018/SAMN04550109.contigs.fa.gz
92.4% ../split_10K_031_dir/chunk_096/SAMN04446463.contigs.fa.gz
92.2% ../split_10K_066_dir/chunk_019/SAMEA1466014.contigs.fa.gz
```



Ondov *et al.* Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology 2016

Brown et al, (2016), sourmash: a library for MinHash sketching of DNA, Journal of Open Source Software, 1(5), 27, doi:10.21105/joss.00027

# Core and accessory distance: pp-sketch index



Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, Corander J, Bentley SD, Croucher NJ. Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Research* **29**:1-13 (2019).

doi:[10.1101/gr.241455.118](https://doi.org/10.1101/gr.241455.118)

# Resource

In [Figshare](#) repository:

- Metadata (both **summarised** and **complete** (127 columns) from ENA)
- QC information
- Characterisation (mlst, serotype, phylotype, AMR and plasmid replicons)

Hosted on the [ftp](#):

- All 661,405 assemblies
- COBS index
- minHash index
- pp-sketch index

# Working with the resource:

Docker://leandroishilima/661k\_query\_indexes:0.0.1

- Container/singularity image with the **version-controlled programs** needed to work with the indexes

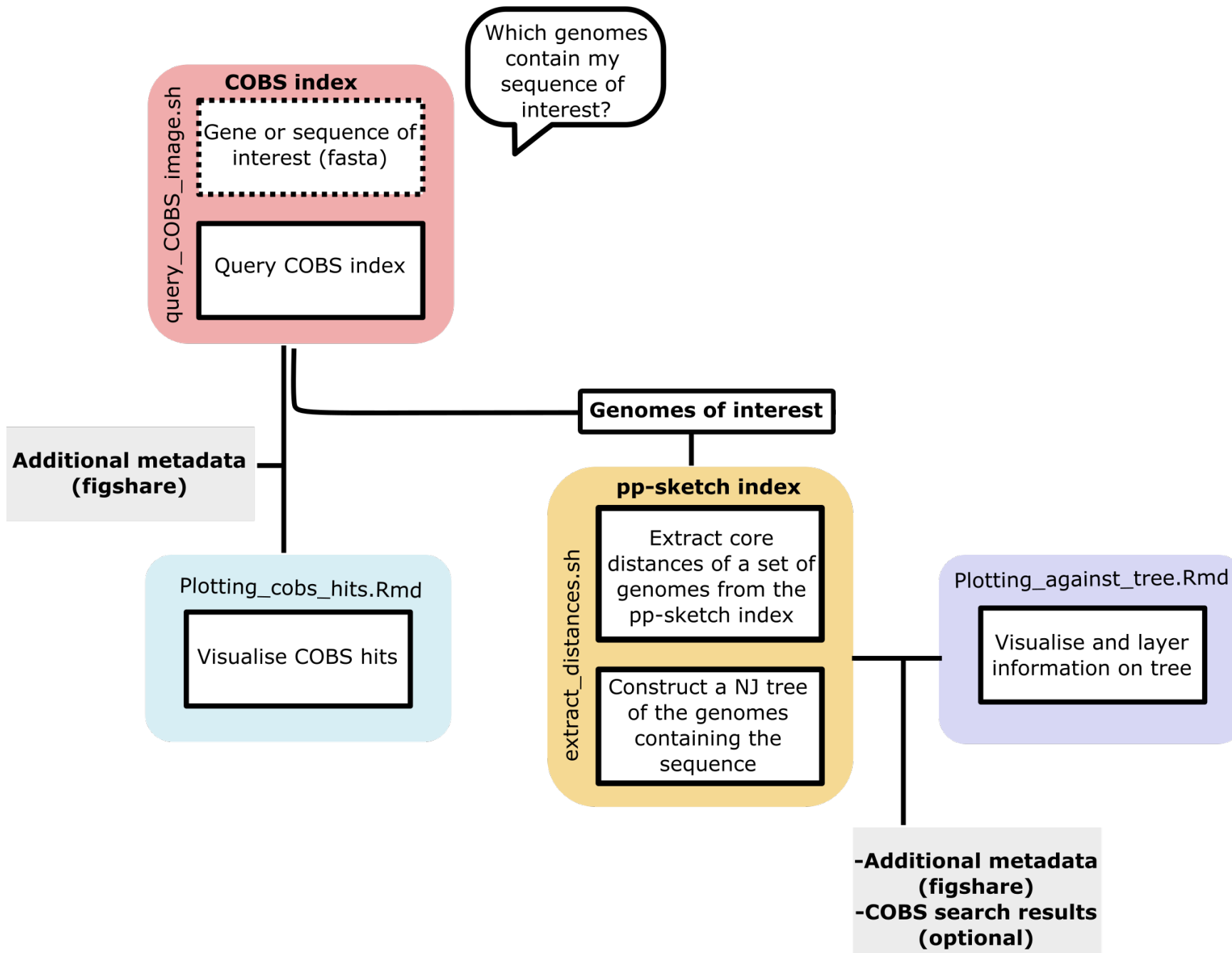
[GitHub](#) repo

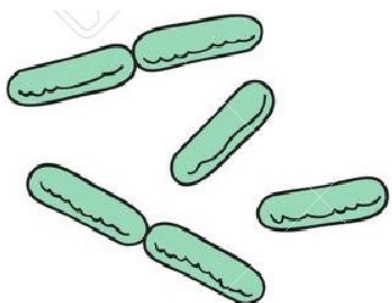
- collection of basic scripts to work with the indexes on sanger clusters
- Rnotebooks for general plotting
- Test files

# Scenario 1: Investigating the distribution of a sequence/gene of interest

- Query COBS index for *bla*<sub>NDM</sub> sequence
- Plot distributions
- Extract core distances from the pp-sketch index and generate tree
- Plot tree and layer information







## Senario 2: What are the most closely-related isolates to mine?

- Sketch and search query genome with sourmash
- Extract core distances from the pp-sketch index and generate tree
- Plot tree and layer information
- Isolate tree node containing genomes with highest similarity

