Big Biological Data

Rayan Chikhi

Institut Pasteur

TUDASTIC 2022

EasyChair submission number 2

Sequence Bioinformatics

Institut Pasteur
Computational Biology Department

Genomes & metagenomes assembly

Algorithms and data structures on k-mers

Sequence search in very large datasets

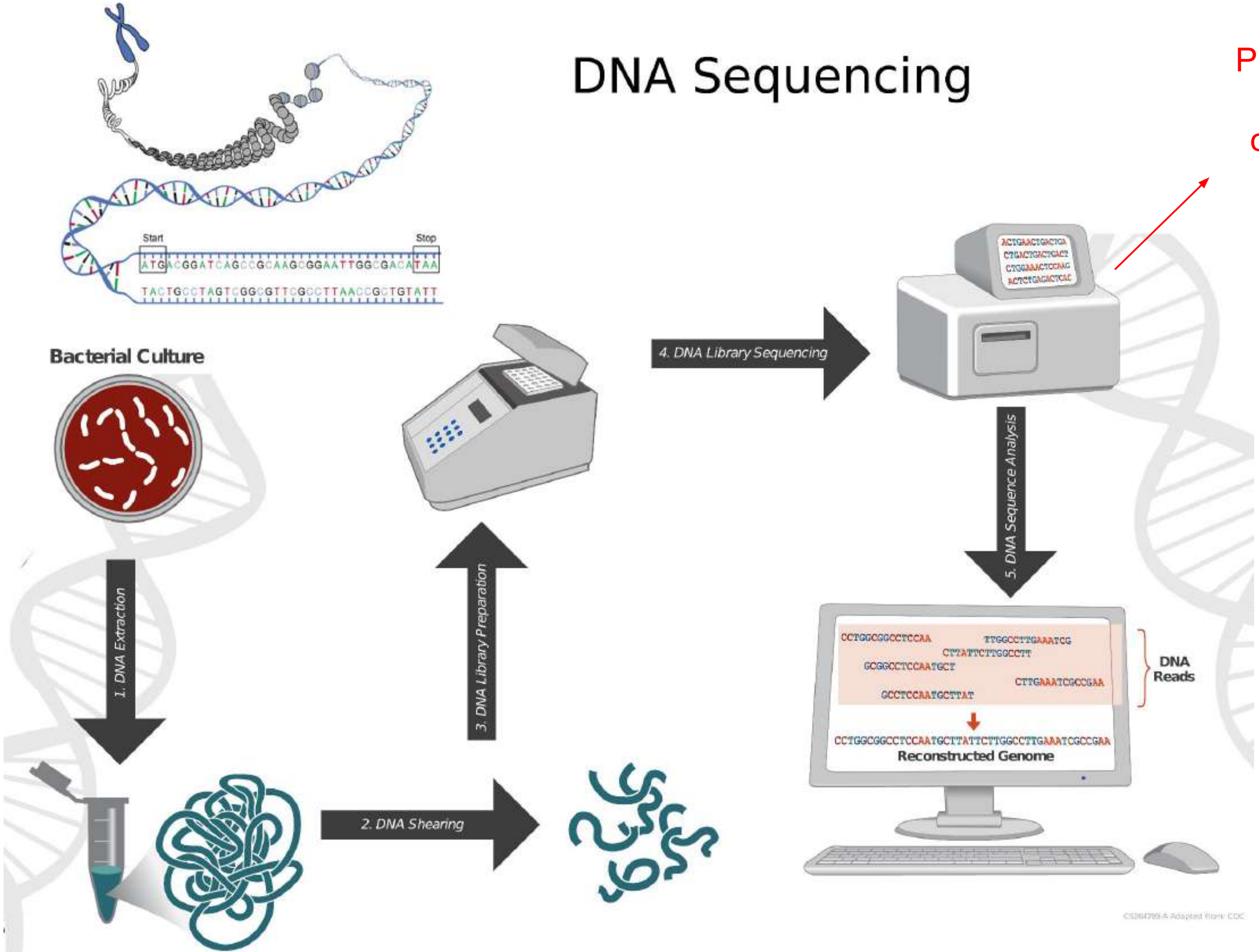
Pangenomics

cnrs PR[AI]RIE PaRis Artificial Intelligence Research InstitutE INSTITUT PASTEUR

Rationale

- Introduction to biological datasets
- From a computer science perspective

DNA Sequencing



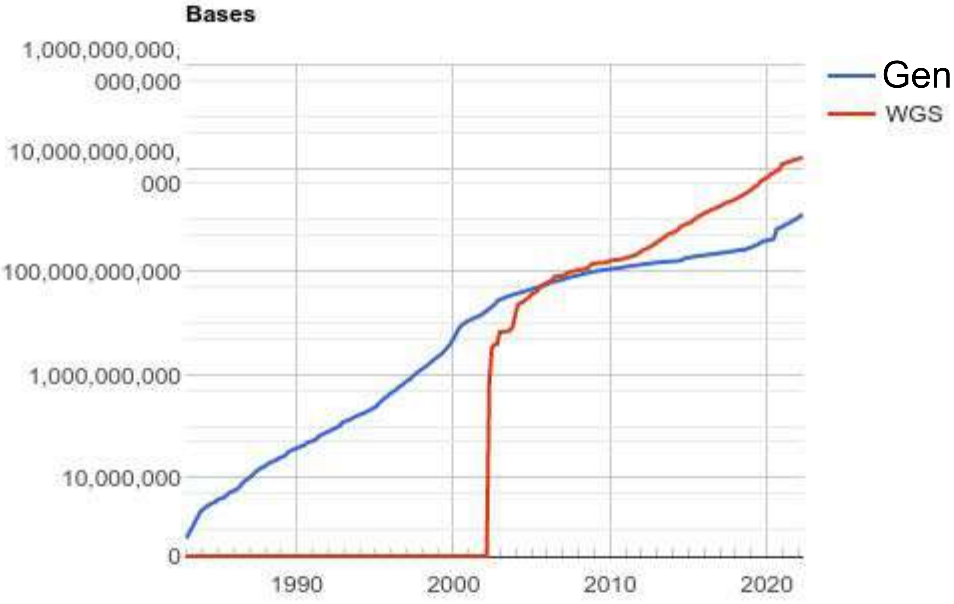
Produces
lots
of data!

2 types of genomic data

- Raw reads
  - Error-prone
  - Many low-abundance distinct k-mers

- Assembled sequences
  - High quality
  - From incomplete gene, plasmids, to complete chromosomes

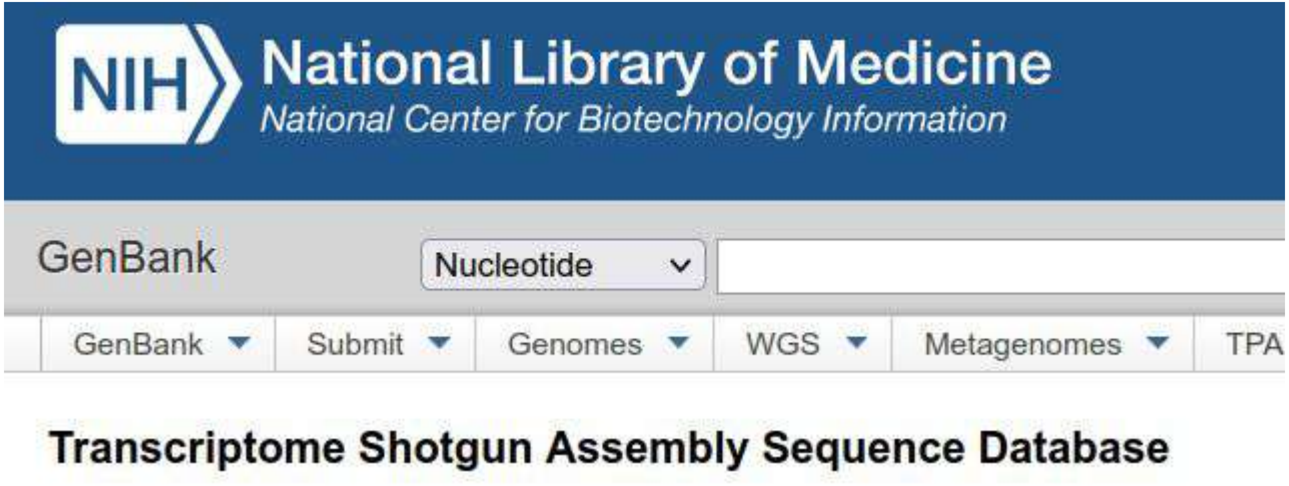
Explosion of available assembled genomes



GenBank : 1 TB, > 200M sequences, 40 year anniv 🎂

(Slide stolen from Antoine)

TSA



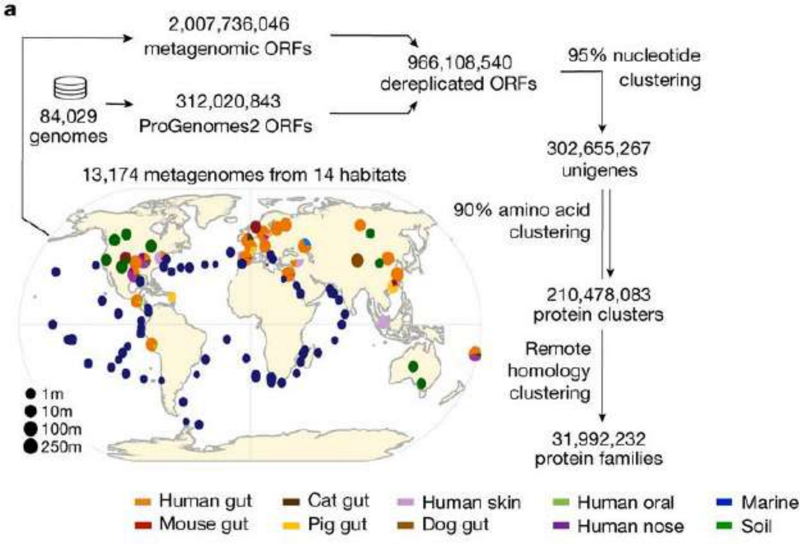
National Library of Medicine
National Center for Biotechnology Information

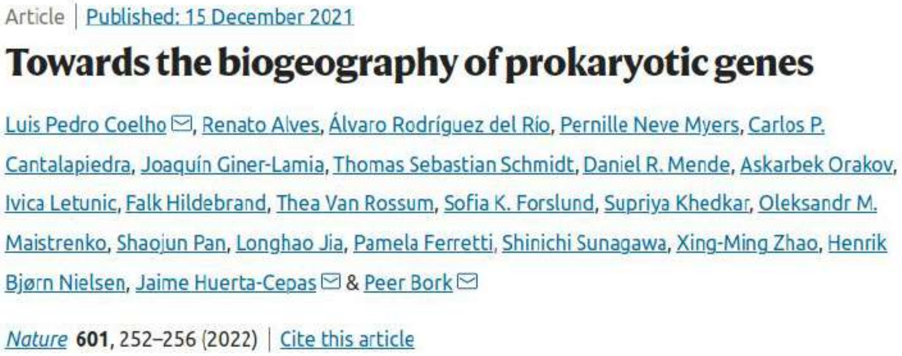
Transcriptome Shotgun Assembly Sequence Database

Type: assemblies
Size: 474 GB (April 2022)
Diversity: high
dBG? ?
# 31-mers: ?
FM-Index? ?

What to mine: inter-species transcripts, RNA viruses, cancer isoforms? (ask Camille)

GMGC (**Global Microbial Gene Catalogue**)





Article | Published: 15 December 2021

Towards the biogeography of prokaryotic genes

Luis Pedro Coelho ✉, Renato Alves, Álvaro Rodríguez del Río, Pernille Neve Myers, Carlos P. Cantalapiedra, Joaquín Giner-Lamia, Thomas Sebastian Schmidt, Daniel R. Mende, Askarbek Orakov, Ivica Letunic, Falk Hildebrand, Thea Van Rossum, Sofia K. Forslund, Supriya Khedkar, Oleksandr M. Maistrenko, Shaojun Pan, Longhao Jia, Pamela Ferretti, Shinichi Sunagawa, Xing-Ming Zhao, Henrik Bjørn Nielsen, Jaime Huerta-Cepas ✉ & Peer Bork ✉

Nature **601**, 252–256 (2022) | Cite this article

Type: assemblies
Size: 124 GB .gz
Diversity: high
dBG? ?
7-mer AA index: done
FM-Index? ?

What to mine: microbial genes, associations to habitat, associations to abundances
(Many examples of mining here: https://www.nature.com/articles/s41586-021-04233-4)

GenBank

GenBank Overview

What is GenBank?

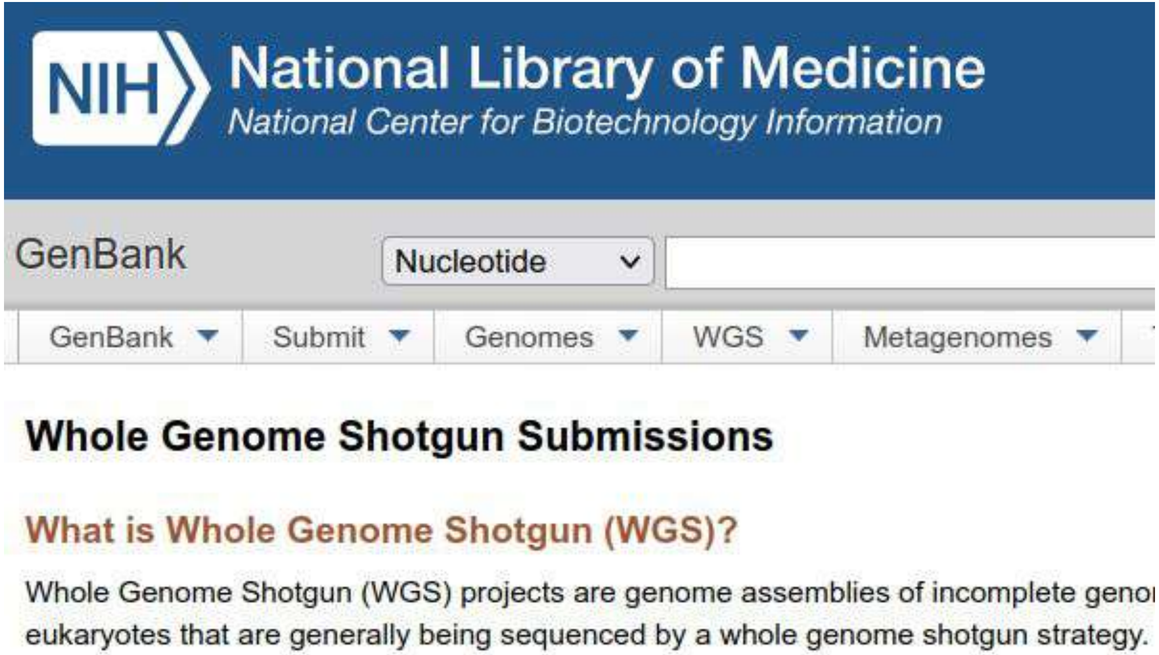
GenBank ® is the NIH genetic sequence database,

Particularity: all sequences are *annotated*

Type: assemblies
Size: 1.2 TB (April 2022)
Diversity: high
dBG? ?
BLAST database: yes

What to mine: genes, association sequences/annotation

WGS


NIH
National Library of Medicine
National Center for Biotechnology Information
GenBank
Nucleotide
GenBank
Submit
Genomes
WGS
Metagenomes

Whole Genome Shotgun Submissions

What is Whole Genome Shotgun (WGS)?

Whole Genome Shotgun (WGS) projects are genome assemblies of incomplete geno
eukaryotes that are generally being sequenced by a whole genome shotgun strategy.

Type: assemblies
Size: 16 TB (April 2022)
Diversity: high
dBG? no
# 31-mers: ?
FM-Index? no

Difference with GenBank: sequences are not necessarily annotated

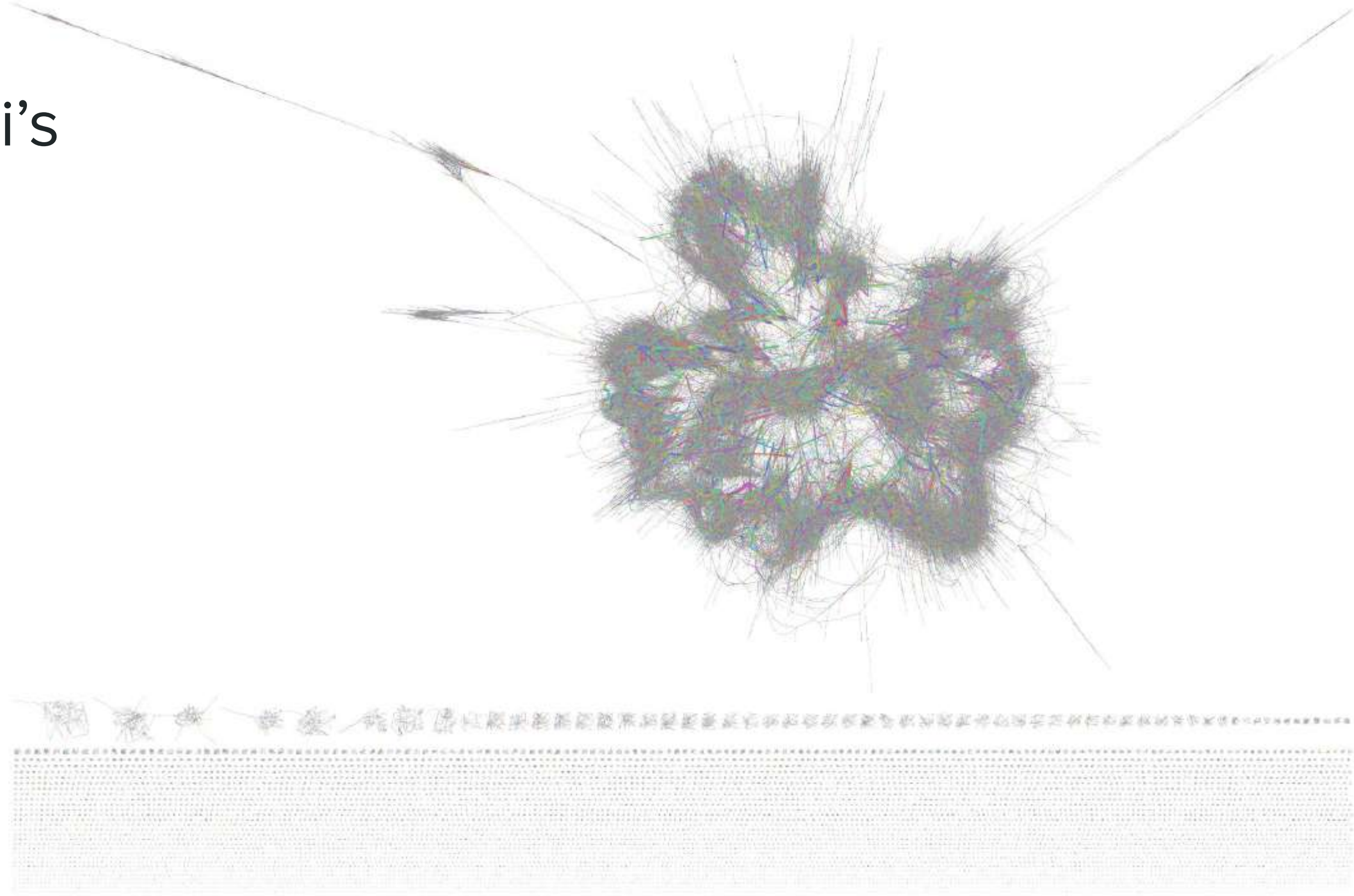
All E.Coli genomes

Type: assemblies
Size: 255 GB .gz
Diversity: low
dBG? feasible
# sequences: 29,865,149
# 10-min-mers (d=0.001): 10,919,184
FM-Index? feasible

Availability: ask me

What to mine: pangenome, HGT, genomic islands, SNPs, ..

167,000 E. coli's mdBG



Many GenBank “divisions”

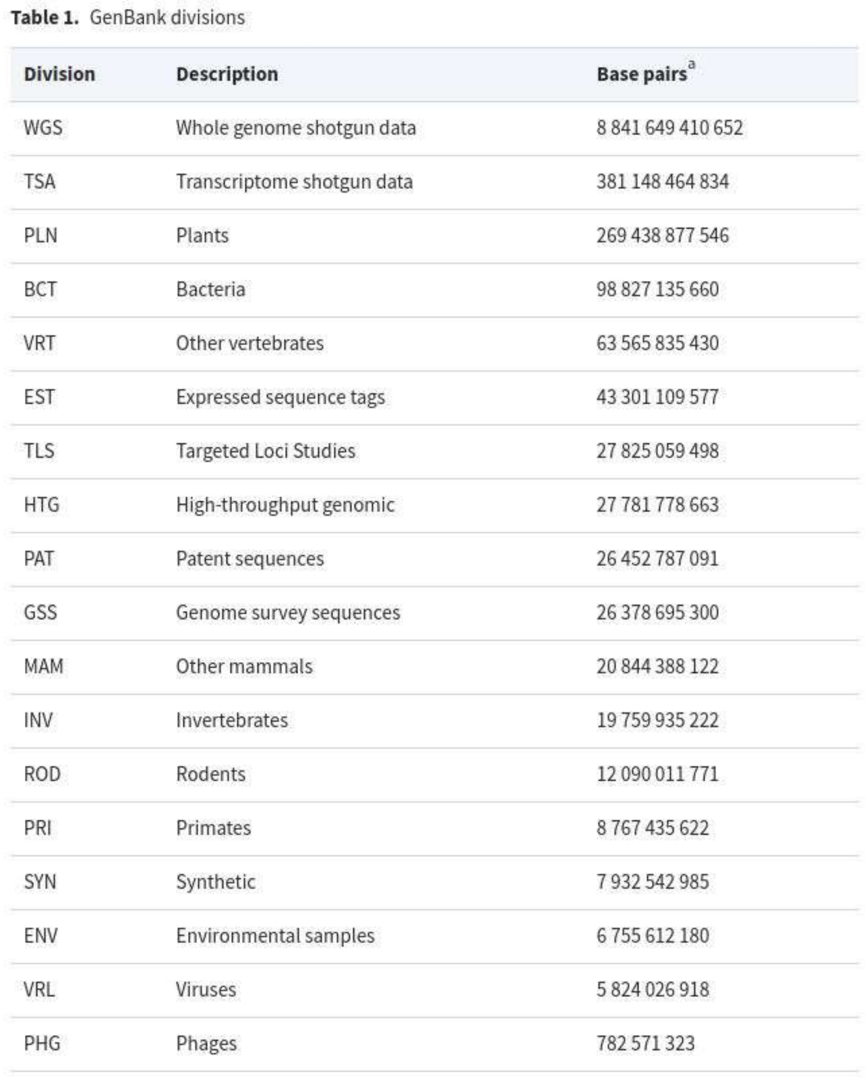
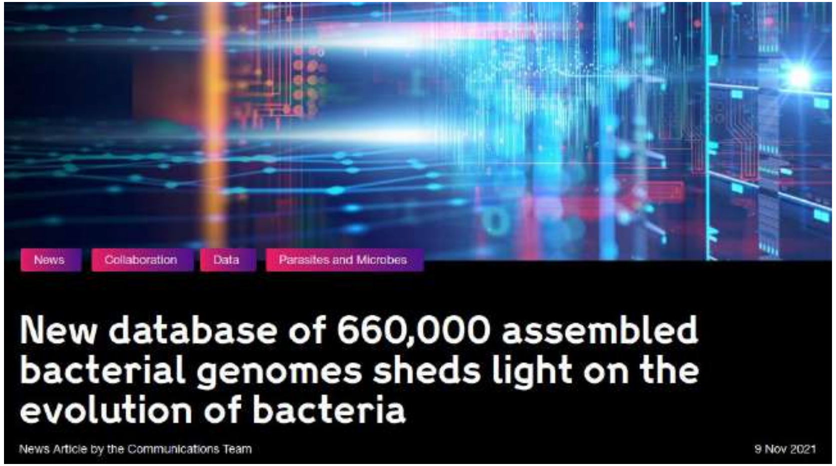


Table 1. GenBank divisions

Division	Description	Base pairs[a]
WGS	Whole genome shotgun data	8 841 649 410 652
TSA	Transcriptome shotgun data	381 148 464 834
PLN	Plants	269 438 877 546
BCT	Bacteria	98 827 135 660
VRT	Other vertebrates	63 565 835 430
EST	Expressed sequence tags	43 301 109 577
TLS	Targeted Loci Studies	27 825 059 498
HTG	High-throughput genomic	27 781 778 663
PAT	Patent sequences	26 452 787 091
GSS	Genome survey sequences	26 378 695 300
MAM	Other mammals	20 844 388 122
INV	Invertebrates	19 759 935 222
ROD	Rodents	12 090 011 771
PRI	Primates	8 767 435 622
SYN	Synthetic	7 932 542 985
ENV	Environmental samples	6 755 612 180
VRL	Viruses	5 824 026 918
PHG	Phages	782 571 323

Source: https://academic.oup.com/nar/article/49/D1/D92/5983623

Blackwell, .., Iqbal’s 661k bacterial genomes collection



New database of 660,000 assembled bacterial genomes sheds light on the evolution of bacteria

News Article by the Communications Team

9 Nov 2021

Type: assemblies
Size: 2.5 TB
Diversity: medium
dBG? yes
# 31-mers: unreported
FM-Index? not yet

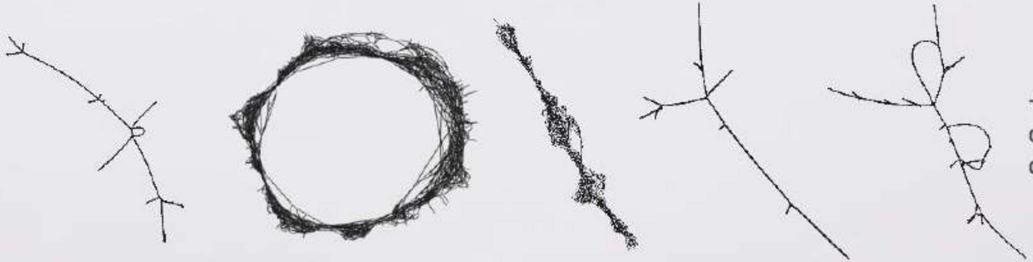
Results: Pangenome graph of 661,405 bacterial genomes

Data from Blackwell et al, 2021:

```
2.9T 661k_assemblies.fa
1.6T 661k_assemblies.fa.lz4
```

```
rust-mdbg -k 10 -l 12 --density 0.001 --minabund 1 661k_assemblies.fa.lz4
```

Largest 5 connected components:



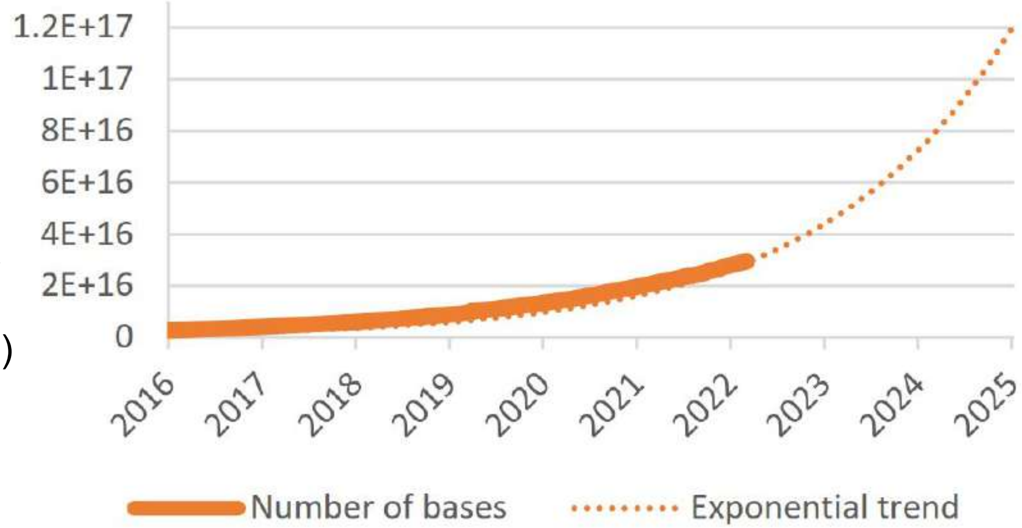
+ 725,820 connected components

Taxons in component	18	22	4	22	10
Dominant species	*Mycobacterium tuberculosis*	*Salmonella enterica*	*Burkholderia gladioli*	*Pseudomonas protegens*	*Cupriavidus alkaliphilus*

SRA

Type: reads
Size: 30 PB
Diversity: extreme
dBG? never
# 31-mers: unreported
FM-Index? never
Index made: STAT

Growth of the Sequence Read Archive



30 petabases
(30 million of gigabases)

YouTube: 100-1000 PB ↗

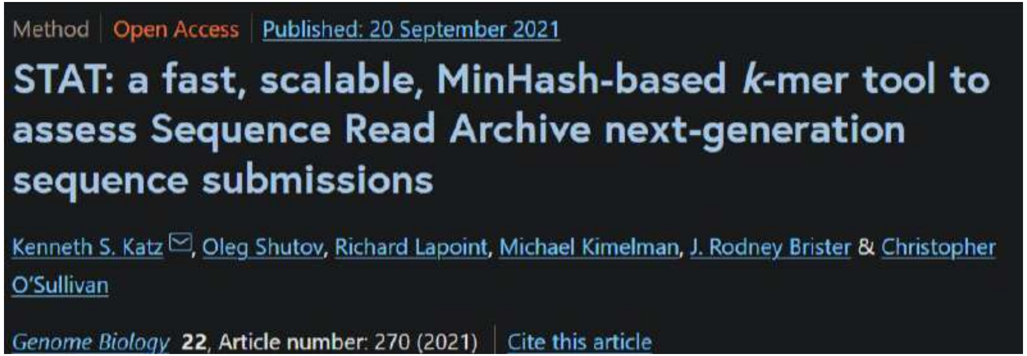
NCBI SRA database : 30 PB

Institut Pasteur: 8 PB

Your laptop: 0.001 PB

NCBI STAT

A recent indexing *tour de force* that it nearly unknown to the community.



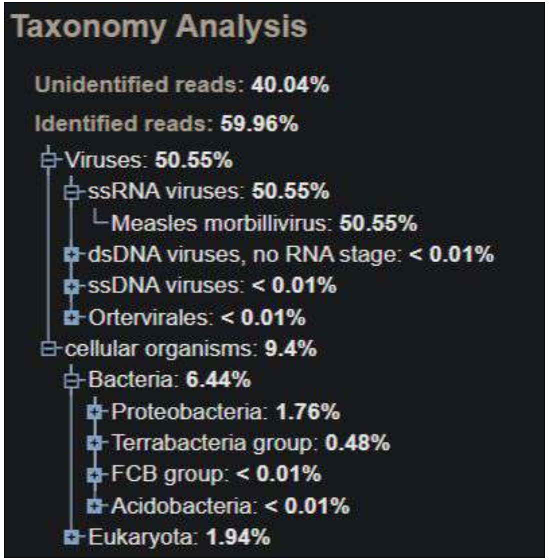
Method | Open Access | Published: 20 September 2021

STAT: a fast, scalable, MinHash-based *k*-mer tool to assess Sequence Read Archive next-generation sequence submissions

Kenneth S. Katz, Oleg Shutov, Richard Lapoint, Michael Kimelman, J. Rodney Brister & Christopher O'Sullivan

Genome Biology **22**, Article number: 270 (2021) | Cite this article

"we have processed more than 27.9 Peta base pairs from runs"



Taxonomy Analysis

Unidentified reads: **40.04%**

Identified reads: **59.96%**

- Viruses: **50.55%**
  - ssRNA viruses: **50.55%**
    - Measles morbillivirus: **50.55%**
  - dsDNA viruses, no RNA stage: **< 0.01%**
  - ssDNA viruses: **< 0.01%**
  - Ortervirales: **< 0.01%**
- cellular organisms: **9.4%**
  - Bacteria: **6.44%**
    - Proteobacteria: **1.76%**
    - Terrabacteria group: **0.48%**
    - FCB group: **< 0.01%**
    - Acidobacteria: **< 0.01%**
  - Eukaryota: **1.94%**

Taxonomic indexing of 32-mer minimizers in 64bp windows
Size: 75 GB (dense version)

The STAT paper is great

Many nuggets:

> *“[..] It is estimated that as little as 30–80 statistically independent single-nucleotide polymorphisms (SNP) can uniquely identify an individual human [..]”*

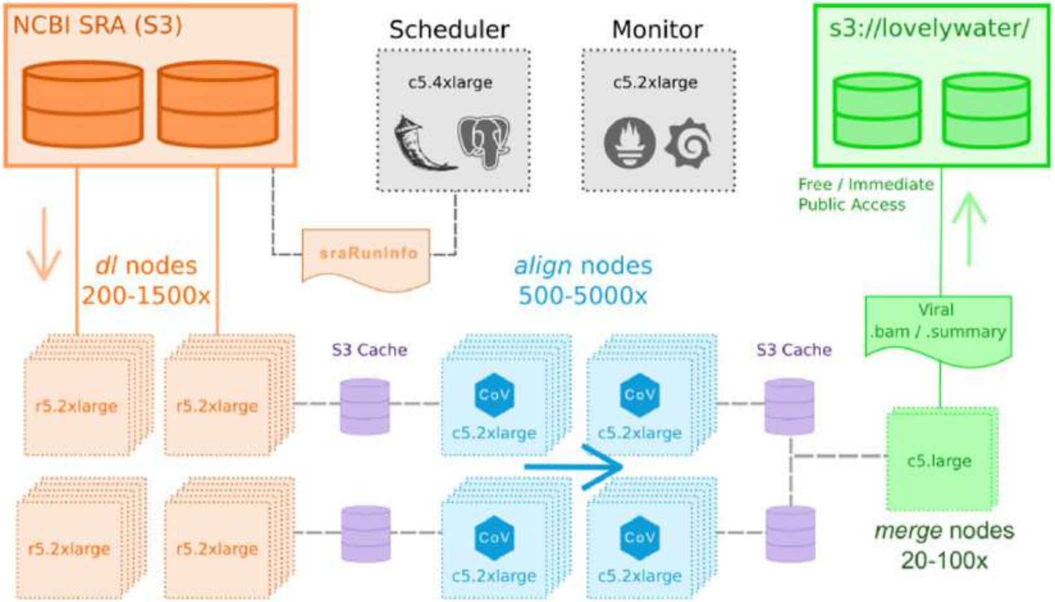
> *“[..] the BLAST® refseq_genomes database is 1.4 terabytes (tb) [..]”*

> *“[..] we released a detection tool containing aligns_to and a Virus “dbs” that allows users to map k-mers found in NGS data to taxa included under Coronaviridae [..]”*

Serratus 2020-2021 assemblies

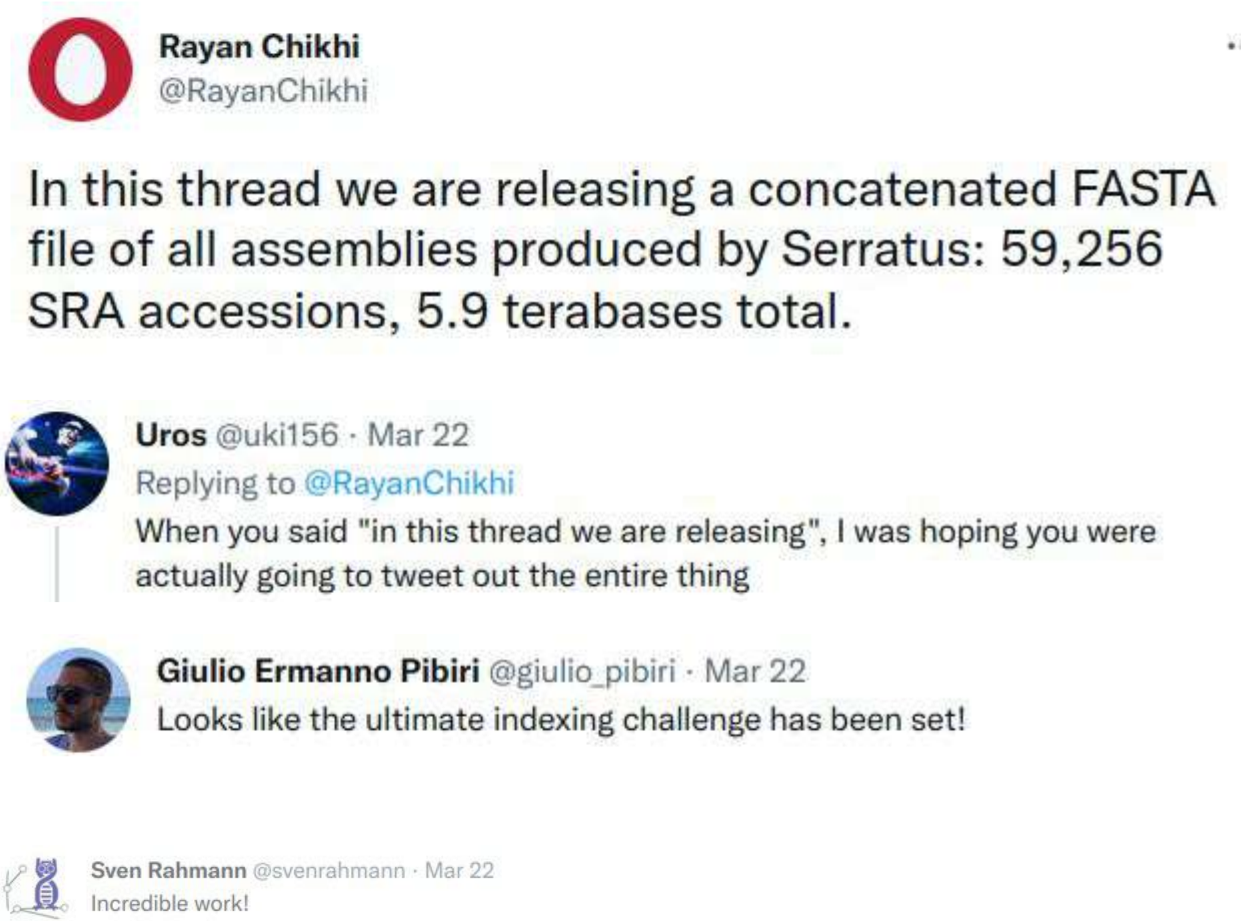
Type: assemblies
Size: 6 TB
Diversity: high
dBG? no
# 31-mers: 1,500 B
FM-Index? no

Serratus architecture



- Aggressively cost-optimized
- Native access to SRA on S3
- Dynamic scaling up to ~22,250s vCPU
- Open Source: GPLv3

SRA also available @ Google Cloud, https://datascience.nih.gov/strides



Rayan Chikhi
@RayanChikhi

In this thread we are releasing a concatenated FASTA file of all assemblies produced by Serratus: 59,256 SRA accessions, 5.9 terabases total.

Uros @uki156 · Mar 22
Replying to @RayanChikhi
When you said "in this thread we are releasing", I was hoping you were actually going to tweet out the entire thing

Giulio Ermanno Pibiri @giulio_pibiri · Mar 22
Looks like the ultimate indexing challenge has been set!

Sven Rahmann @svenrahmann · Mar 22
Incredible work!

(Including Sven in the screenshot)

The “nr” database of BLAST

“The nucleotide collection consists of ***GenBank****+EMBL+DDBJ+PDB+RefSeq sequences, but excludes EST, STS, GSS, WGS, TSA”*

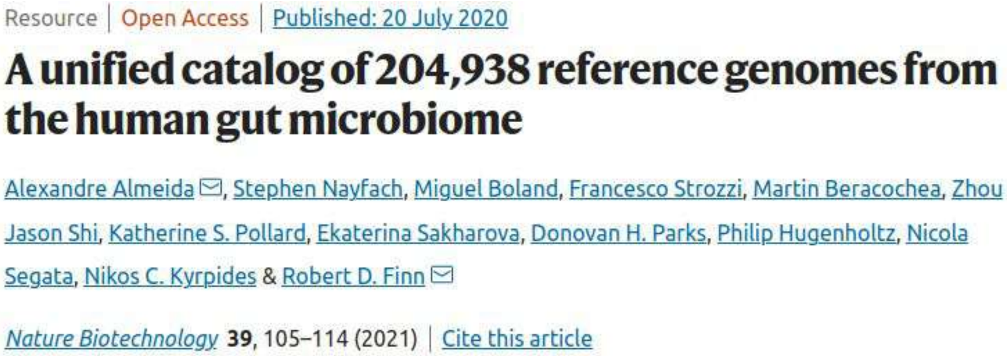
[..] “The database is non-redundant.”

125 GB compressed

ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz

PS: “refseq_genomes” database: 1.5 TB [ref: STAT]

Many others (often metagenomic)

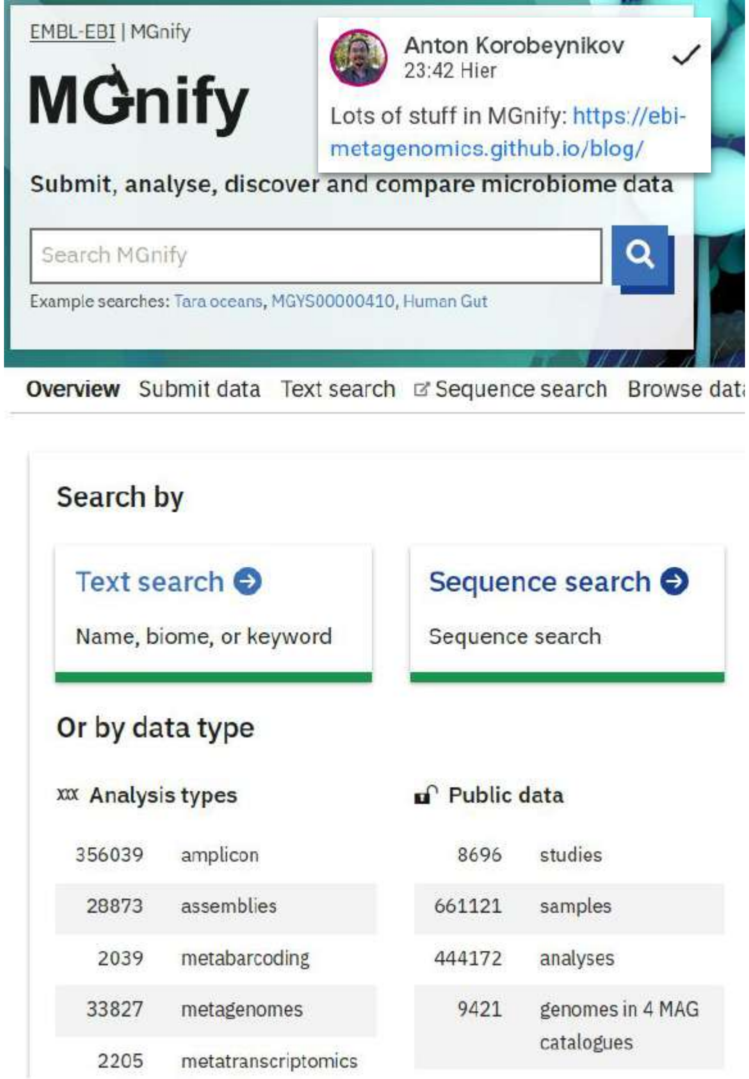


Resource | Open Access | Published: 20 July 2020

A unified catalog of 204,938 reference genomes from the human gut microbiome

Alexandre Almeida, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, Zhou Jason Shi, Katherine S. Pollard, Ekaterina Sakharova, Donovan H. Parks, Philip Hugenholtz, Nicola Segata, Nikos C. Kyrpides & Robert D. Finn

Nature Biotechnology **39**, 105–114 (2021) | Cite this article



Anton Korobeynikov
23:42 Hier

Lots of stuff in MGnify: https://ebi-metagenomics.github.io/blog/

Submit, analyse, discover and compare microbiome data

Example searches: Tara oceans, MGYS00000410, Human Gut

Search by

Text search ➔
Name, biome, or keyword

Sequence search ➔
Sequence search

Or by data type

Analysis types

356039	amplicon
28873	assemblies
2039	metabarcoding
33827	metagenomes
2205	metatranscriptomics

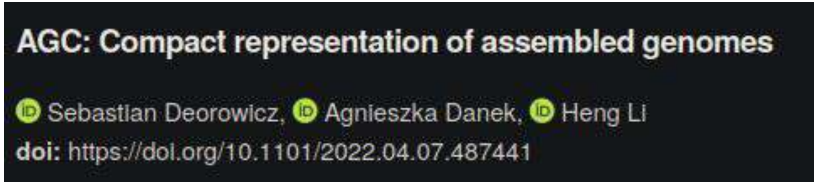
Public data

8696	studies
661121	samples
444172	analyses
9421	genomes in 4 MAG catalogues

MGNify: a database of assemblies of metagenome studies from ENA searchable by metadata

Conclusion

Text indexing community: What can we do with so much data?

- Index it
  - MinHash sketches
  - k-mers
  - k-min-mers
  - BWT
  - r-index
- Compress it
  - gz, xz
  - 

    AGC: Compact representation of assembled genomes

    Sebastian Deorowicz, Agnieszka Danek, Heng Li

    doi: https://doi.org/10.1101/2022.04.07.487441

- And if possible, make biological discoveries from it!