Original Investigation

iDarwin volume 4, pages 4-23

Published on March 13, AS0024 (2024 AD)

GenomeSync: a synchronizable database of genome sequences

Kirill KRYUKOV[1,2*], So NAKAGAWA[3], and Tadashi IMANISHI[3]

[1]Center for Genome Informatics, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, Mishima, Shizuoka, 411-8540, Japan

[2]Bioinformation and DDBJ Center, National Institute of Genetics, Mishima, Shizuoka, 411-8540, Japan

[3]Department of Molecular Life Science, Tokai University School of Medicine, Isehara, Kanagawa, 259-1193, Japan

*Correspondence address: Kirill KRYUKOV, Center for Genome Informatics, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, Mishima, Shizuoka 411-8540, Japan; Email: kirill-kryukov@nig.ac.jp

Email addresses for other authors: So Nakagawa; so@tokai.ac.jp Tadashi Imanishi; imanishi@tokai.ac.jp

Abstract

Background: With the advances of DNA sequencing technology, the number of available genomes of diverse species is rapidly increasing. This vast genomic data forms a basis for numerous applications, including medical, industrial and environmental studies. However, obtaining a taxonomically consistent set of genomes is a still a non-trivial task. Maintaining and keeping this data up to date with changes in both genomes and taxonomy requires substantial time and effort.

Findings: To solve this problem, we have developed and released GenomeSync, a database of genome sequences containing ~850 thousand genomes as of Fabruary, 2024. GenomeSync efficiently stores 13.2 TB of sequence data compressed to 2.8 TB in the Nucleotide Archival Format, which allows quicker transfer and decompression than the usual gzip format. GenomeSync is designed for convenient automatic synchronization of any taxonomic subset using portable command-line operations. GenomeSync also provides the taxonomic structure, and GenomeSync genomes are always consistent with its taxonomy snapshot.

Conclusions: GenomeSync enables us to obtain a taxonomically defined set of genomes and to keep the genomes up-to-date. This enables computational pipelines to conveniently utilize the required genome data, and to interpret the results using the included taxonomy structure. GenomeSync has been maintained since 2015, is regularly updated, and it will be helpful for various studies, including comparative genomics and metagenomics analyses.

Database URL: https://genomesync.org/

Keywords

genome sequence database, synchronization, taxonomy, nucleotide archival format

Introduction

Many biological and medical studies depend on the availability of genome sequences. Comparison between genomes can reveal the evolutionary history of genes and organisms (e.g., Hug et al., 2016; Zhu et al., 2019). In particular, conserved sequence regions among various genomes infer their molecular importance even in the non-coding regions (e.g., Inoue and Saitou, 2021). In addition, metagenomic sequencing data can be analyzed by comparing them with known sequences from a genome database to determine their origins. A large number of genomes available recently can provide unprecedented power to these sequence comparison approaches. However, the rapid growth of genome databases presents a unique challenge of obtaining, managing, and using this massive data.

The traditional approach for obtaining genome sequence data involves navigating the website of a public sequence database such as NCBI Assembly (Sayers, Beck, et al., 2021) or Ensembl (Howe et al., 2021), locating the genomes needed, and downloading them, either manually or semi-automatically. This requires substantial time and effort, and becomes more difficult with a large number of genomes. Moreover, maintaining a previously downloaded set of genomes is also a highly non-trivial task, that includes locating and downloading new genomes, as well as removing obsolete assemblies and replacing them with new ones. As a result, researchers often keep using a fixed set of genomes downloaded years ago, unable to afford to keep up-to-date with currently available genomes which results in missing the potentially available increased accuracy in their investigations. With the number of available genomes growing exponentially (Zhao et al., 2020; https://www.ncbi.nlm.nih.gov/refseq/statistics/), it becomes even harder to justify the cost of maintaining the local genome data.

Taxonomic classification is usually necessary for keeping track of the relationships between diverse organisms and their genomes, and for interpreting the results of genome-based analyses. NCBI Taxonomy (Schoch et al., 2020) is often used for this purpose. However, same like genome data, taxonomy information undergoes frequent changes. Not only are new taxa often introduced, but taxa can also be renamed, merged, split, and moved around in the taxonomic relationship. This frequently creates mismatches between taxonomy and organism names used in genome data. Therefore, keeping taxonomy in sync with the genome data is another non-trivial task necessary when maintaining a local set of genomes.

Most genome databases store sequence data in gzip-compressed FASTA format. This compressed format is known to be sub-optimal in both compactness and speed of extracting the data (Kryukov et al., 2020). This means that downloading and storing gzipped genomes wastes bandwidth and hard disk space compared to more efficient formats. This waste can be significant with large genome databases. The inefficiency manifests not only during the initial downloading of the genome data, but also when distributing the data to computation nodes, and when utilizing it.

The existing systems for downloading or synchronizing genome data, such as NCBI Datasets (2021), entrez-direct (Kans, 2022), and genome_updater (Piro, 2020), can help with some of the issues described above, but they don't solve the entire set of issues. They can help obtaining a selected set of genomes, but don't robustly solve the taxonomic consistency of the downloaded data. Also they may suffer from the inefficiency of compression format. The Genome Taxonomy Database project aims to provide a set of genomes synchronized with taxonomy (Parks et al., 2020), but that project is limited to prokaryotes and not updated frequently.

One particular issue with available public databases is the inclusion of many anomalous genome assemblies. Downloading taxonomically selected set of genomes from NCBI results in many partial, incomplete and other anomalous assemblies. Another issue is the limited representative subsets of genomes, which is limited to species level at the existing databases.

In this study, we addressed these problems by developing the GenomeSync database - a self-consistent genomic dataset providing efficiency and ease of use. GenomeSync is designed for convenient automatic download and synchronization of any taxonomically defined subset using portable command line operations. GenomeSync includes both genome sequences and a matching snapshot of the taxonomy, removing the need for synchronizing between them. Unlike most other databases, GenomeSync stores genome sequence data in Nucleotide Archival Format (NAF) which is efficient in saving disk space and access time when using genome data (Kryukov et al., 2019). GenomeSync mainly contains genome assemblies from NCBI, but it is not limited to only one data source, and includes public genome data from various other databases and repositories.

GenomeSync database

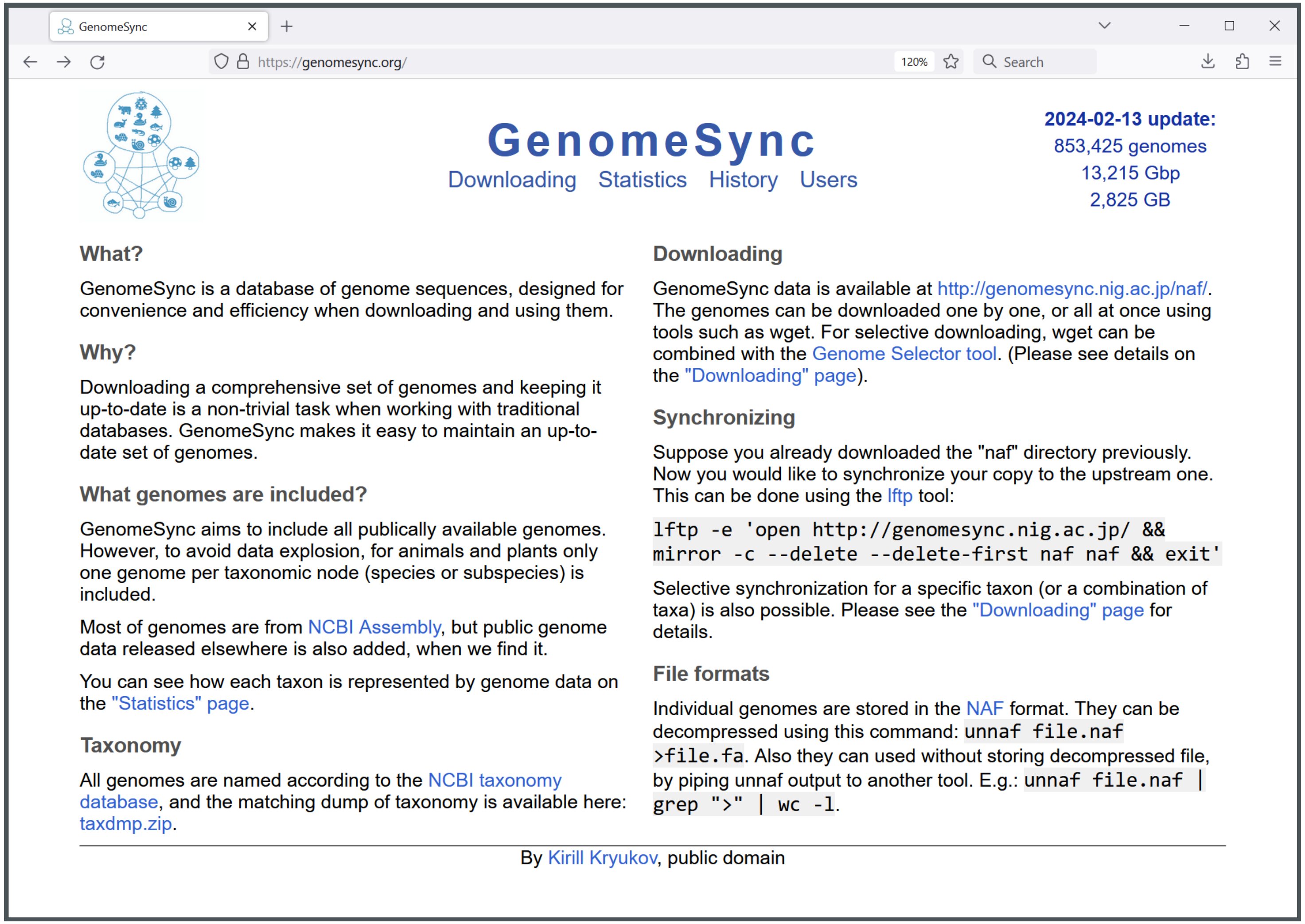
Overview

GenomeSync is an online database of genome sequences (https://genomesync.org/, Fig. 1). The web interface of GenomeSync allows easily discovering what organisms already have assembled genomes, and downloading genome sequences for any taxonomically defined subset. It also allows keeping an already downloaded set of genomes up-to-date, by synchronizing it to the upstream GenomeSync database. All genomes are compressed into the NAF format, improving compactness and speed of use compared to the normally used gzip format (Kryukov et al., 2019; 2020).

Scope

GenomeSync aims to include maximally broad selection of public genome data and to support genome-based investigations. GenomySync includes genome data of all organisms: eukaryotes, prokaryotes and viruses. On the other hand, including all currently existing sequence data would quickly make the dataset enormous, expensive to store and difficult to use. Therefore some line has to be drawn between including all available data and limiting data size to practically usable range.

Fig. 1. Screenshot of the GenomeSync main page (as of February 13, 2024).



GenomeSync

Downloading Statistics History Users

2024-02-13 update:
853,425 genomes
13,215 Gbp
2,825 GB

What?

GenomeSync is a database of genome sequences, designed for convenience and efficiency when downloading and using them.

Why?

Downloading a comprehensive set of genomes and keeping it up-to-date is a non-trivial task when working with traditional databases. GenomeSync makes it easy to maintain an up-to-date set of genomes.

What genomes are included?

GenomeSync aims to include all publically available genomes. However, to avoid data explosion, for animals and plants only one genome per taxonomic node (species or subspecies) is included.

Most of genomes are from NCBI Assembly, but public genome data released elsewhere is also added, when we find it.

You can see how each taxon is represented by genome data on the "Statistics" page.

Taxonomy

All genomes are named according to the NCBI taxonomy database, and the matching dump of taxonomy is available here: taxdmp.zip.

Downloading

GenomeSync data is available at http://genomesync.nig.ac.jp/naf/. The genomes can be downloaded one by one, or all at once using tools such as wget. For selective downloading, wget can be combined with the Genome Selector tool. (Please see details on the "Downloading" page).

Synchronizing

Suppose you already downloaded the "naf" directory previously. Now you would like to synchronize your copy to the upstream one. This can be done using the lftp tool:

```
lftp -e 'open http://genomesync.nig.ac.jp/ &&
mirror -c --delete --delete-first naf naf && exit'
```

Selective synchronization for a specific taxon (or a combination of taxa) is also possible. Please see the "Downloading" page for details.

File formats

Individual genomes are stored in the NAF format. They can be decompressed using this command: `unnaf file.naf >file.fa`. Also they can used without storing decompressed file, by piping unnaf output to another tool. E.g.: `unnaf file.naf | grep ">" | wc -l`.

By Kirill Kryukov, public domain

For GenomeSync, we defined the following criteria for including or excluding the data: i) Only DNA sequences of reasonably complete de-novo assembled genomes are included. Raw sequencing reads, partial assemblies, exomes, individual genes, protein sequences, and unsorted metagenomes are outside of this scope. ii) Only publicly accessible data is included. Genomes from private repositories that have usage restrictions are not used. iii) For plants and animals, only one genome per taxonomic node is included, which typically means one genome per species. This means that, for example, hundreds of available human genomes will not occupy a disproportionately large part of the database. iv) For other organisms, multiple genomes per taxonomic node are included, except tens of thousands of same species genomes from massive multi-isolate bacterial sequencing projects. v) Draft assemblies are included, but genomes that are clearly partial or too small are excluded.

As for iv), GenomySync includes multiple genomes per species (or subspecies) for prokaryotes, fungi, protists and viruses, which is different from iii). This is rooted in the idea of optimally representing sequence diversity. For plants and animals, within species genetic variation is comparatively small, and a single genome may contain most of the sequence characteristic to the organism. In prokaryotes, on the other hand, within species variation is comparatively much larger. Furthermore, in prokaryotes there is a concept of a “core genome” shared by all (or most) members of the species, and an “accessory genome” consisting of genes that are only present in some isolates (Segerman, 2012). Thus multiple individual genomes are necessary to capture the pangenome of a bacterial species.

The described inclusion criteria allow us to keep the overall data size manageable (around 2.8 TB as of February 2024), while maximally representing diverse organisms that already have genome data. Table 1 shows the summary for the number of genomes present in GenomeSync (as of February 19, 2024), as well as numbers of represented species, genera, families and orders.

Data sources

While GenomeSync includes public genomes from any sources, most of the genomes are from GenBank (Sayers, Cavanaugh, et al., 2021) and RefSeq (Li et al., 2021), downloaded via the NCBI Assembly (Sayers, Beck, et al., 2021). Many other sources have been used at various points of time. All GenomeSync data sources are listed in Table 2, including their URLs and citations.

Table 1. Summary of the number of genomes in GenomeSync (as of February 13, 2024), for Viruses, Archaea, Bacteria and Eukaryota. Also shows the number of represented species, genera, families and orders. The percentages show proportion of nodes with genomes among all nodes of particular rank described in the NCBI Taxonomy Database

Taxon	Genomes	Species		Genera		Families		Orders	
Viruses	79,022	30,380	49.71%	2,405	86.48%	243	93.46%	69	95.83%
Archaea	18,410	3,010	23.09%	272	94.44%	100	97.09%	71	93.42%
Bacteria	727,313	72,281	14.72%	4,422	84.73%	822	92.67%	343	95.54%
Eukaryota	28,680	15,572	1.04%	6,719	6.72%	2,359	26.04%	726	52.53%
Total	853,425	121,243	5.80%	13,818	12.75%	3,524	34.19%	1,209	64.00%

Table 2. GenomeSync data sources

Source	URL	Citation
NCBI Taxonomy	`https://www.ncbi.nlm.nih.gov/taxonomy`	Sayers, Beck, et al., 2021
GenBank	`https://www.ncbi.nlm.nih.gov/genbank/`	Sayers, Cavanaugh, et al., 2021
RefSeq	`https://www.ncbi.nlm.nih.gov/refseq/`	Li, O'Neill, et al., 2021
NCBI Assembly	`https://www.ncbi.nlm.nih.gov/assembly/`	Sayers, Beck, et al., 2021
JGI	`https://genome.jgi.doe.gov/portal/`	Nordberg et al., 2014
Ensembl	`https://www.ensembl.org/`	Howe et al., 2021
Phytozome	`https://phytozome.jgi.doe.gov/`	Phytozome, 2021
UCSC	`https://hgdownload.soe.ucsc.edu/`	UCSC, 2021
Zoonomia	`https://zoonomiaproject.org/`	Zoonomia Consortium, 2020
FlySeq	`http://flyseq.org/`	Kim et al., 2021
WormBase	`https://wormbase.org/`	Harris et al., 2020
FlyBase	`https://flybase.org/`	Larkin et al., 2021
VEuPathDB	`https://veupathdb.org/`	Warrenfeltz et al., 2018
FungiDB	`https://fungidb.org/`	Basenko et al., 2018
PineRefSeq	`https://nealelab.ucdavis.edu/pinerefseq/`	PineRefSeq, 2021
TreeGenes	`https://treegenesdb.org/`	Falk et al., 2018
Lepbase	`http://lepbase.org/`	Challis et al., 2016
Vertebrate Genomes Project	`https://vertebrategenomesproject.org/`	Rhie et al., 2021
FernBase	`https://www.fernbase.org/`	Li, Brouwer, et al., 2018
ReefGenomics	`http://reefgenomics.org/`	Liew et al., 2016

Since GenBank and RefSeq are more systematically organized, and provide better support for automatic access, when the same assembly is available in multiple sources, we prefer the GenBank/RefSeq ones. This means that genomes from other sources are often only transiently present in GenomeSync: They are added to GenomeSync when they first become available at some genome repository, then replaced with the corresponding GenBank/RefSeq assemblies when they become available.

Taxonomic structure is obtained from the NCBI Taxonomy database (Schoch et al., 2020). Particular care is taken that all organism names in GenomeSync always match the taxonomy, and genome files and taxonomy data are kept synchronized.

File format

All GenomeSync genomes are stored in the FASTA format compressed into the Nucleotide Archival Format (NAF) (Kryukov et al., 2019). NAF was designed specifically for storing large amounts of DNA sequence data, with a focus on providing a balance between compactness and speed of access. NAF is a reference-free format, which means that each genome stands independent from other genomes, and no reference data is required for decompressing each genome.

Compared to gzip, which is universally used by other genome databases, NAF provides about 1.5-2 times stronger compression and 3-5 times better decompression speed on genome data. On other sequence data, such as collections of virus or gene sequences, NAF achieves up to 10 times better compactness compared to gzip. Previously we benchmarked NAF together with ~50 other compressors on various sequence data, and confirmed that NAF provides optimal combination of compactness and decompression speed on genome data (see details in Kryukov et al., 2020).

Compressor and decompressor for the NAF format are open source and freely available on GitHub at https://github.com/KirillKryukov/naf. They can be installed using commands:

```
git clone --recurse-submodules https://github.com/KirillKryukov/naf.git
cd naf && make && make test && sudo make install
```

After installing, a NAF-compressed file can be decompressed into FASTA format as:

```
unnaf file.naf >file.fa
```

The decompressed data can also be piped into another command, for example, for creating a BLAST database:

```
unnaf file.naf | makeblastdb -dbtype nucl -out file
```

The high speed of the unnaf decompressor means that the genome data can be stored in the NAF format and only extracted when necessary, saving substantial disk space compared to storing it in other formats.

The total size of GenomeSync sequence data (as of February 23, 2024) is 13.65 TB in FASTA format. In the NAF format (one genome per file), this data consumes only 2.845 TB. With gzip compression, the same data would occupy about 4.0 TB. Thus, using NAF compression instead of gzip for GenomeSync saves about 1.2 TB of disk space.

When doing massive genome-based data analysis, often multiple computers are used in parallel. This usually means that genome data has to be copied between machines over the network. Compactness of NAF-compressed GenomeSync data means that the process of distributing this data to calculation nodes occurs faster, saving time each time the genomes have to be distributed or updated.

File names

GenomeSync is structured with the aim of eliminating unnecessary complexity, and providing as immediate and direct access as possible to the sequence data. Each genome is stored in a single NAF file. The files are named using a scheme: “`organism name [source accession date].naf`”. For example, a human genome can be stored in a file named: “`Homo sapiens [refseq GCF_5F000001405.40 2022-02-03].naf`”.

Since the character set used in file names is restricted, we encode all characters that are unsafe in file names with an underscore “`_`” followed by the two-digit hexadecimal ASCII character code. For example, “`Picochlorum sp. 'celeri'`” becomes “`Picochlorum sp. _27celeri_27`”, after encoding single quotation marks “`'`” as “`_27`”. The only characters that are used without such encoding are: small and capital latin letters (a-z, A-Z), digits (0-9), hyphen (-), dot (.). For organism names, additionally space is allowed without encoding. Organism name, source, and accession are encoded separately, then combined together into a file name, so that the square brackets are not altered. Since the underscore character is used by the encoding, this character itself also must be encoded. For example, a genome accession “`GCF_000001405.39`” is recorded as “`GCF_5F000001405.39`” in the file name. Although the encoding uses only two hexadecimal digits (restricting it to codes from 0 to 255), effectively it can represent any Unicode characters, via the UTF-8 encoding.

Organism names are following the taxonomy database. One issue is that sometimes different taxonomic nodes share the same name, and a disambiguation is necessary. In such cases the disambiguation suffix is used in the file name too.

Date part of the file name represents the assembly date, and is obtained from the source database together with the genome. Date is always stored in the ISO 8601 format: YYYY-MM-DD. The modification date of a genome file itself corresponds to when the genome was added to GenomeSync, and does not necessarily match the date in the file name.

Directory structure

GenomeSync contains about 850,000 genomes as of February 2024, each stored in its own file. It would not be a good idea to keep all these files in the same directory, since, depending on the filesystem, this may cause slowdown and instability. Therefore, genomes are stored in about 120 directories, roughly following the taxonomic tree structure. This structure is intended to help avoid directories with too many files. See Suppl. Table 1. for the current (as of February 2024) list of taxa with their corresponding directories. This structure is not fixed, but will continue to change with taxonomy and with continuing proliferation of genome data.

Representative genomes

Many bacterial, fungal and protozoan species have more than one genome in GenomeSync. In some cases, thousands of genomes exist for different strains or isolates of the same species. Sometimes it is useful to select only a smaller subset of genomes, while still maximally covering the various organisms. For this purpose, GenomeSync includes two subsets of genomes: rep and rep2. The rep subset includes representative genomes annotated as such at the NCBI Assembly (https://www.ncbi.nlm.nih.gov/assembly). Additionally, for all eukaryote species that don’t have any genome marked as representative at the NCBI Assembly, the rep subset includes one genome per species. The rep2 subset includes all genomes from the rep subset, plus one genome per species for all still not included species.

When selecting a representative genome out of several genomes available for a species, we try to select a genome with a better assembly quality. We use assembly size and continuity as a proxy for quality. First, for each genome we compute Q as:

$$Q = \frac{N_{ACGT} - 10 \times N_{non-ACGT}}{\sqrt{N_{seq}}}$$

where N_{ACGT} is the number of determined nucleotides in the genome sequence (A, C, G or T), $N_{non\text{-}ACGT}$ is the number of ambiguous nucleotides, denoted with IUPAC nucleotide codes, including "N" (an undetermined nucleotide), and N_{seq} is the number of sequences (either chromosomes, scaffolds, or contigs) in the assembly.

Then we select the genome with the highest Q as representative. This means choosing larger and less fragmented genomes, with fewer unknown nucleotides. Large assembly size does not always mean better quality, but in practice it is a useful proxy measure, because many incomplete and partial assemblies are small and/or very fragmented.

Table 3 shows the number of genomes and FASTA sizes of the two representative subsets, and of the entire GenomeSync data. It also shows the sizes of prokaryote-only selections out of the two subsets and out of the entire database.

Table 3. Number of genomes in representative sets, and their total sizes in FASTA and NAF formats (as of February 13, 2024)

Set	Number of genomes	Size in FASTA format (GB)	Size in NAF format (GB)
rep prokaryotes	18,941	85	20
rep2 prokaryotes	75,398	303	72
all prokaryotes	752,105	2,551	613
rep	34,623	10,385	2,078
rep2	91,080	10,603	2,130
all	860,376	13,654	2,845

Statistics page

GenomeSync website provides a statistics interface webpage at http://genomesync.nig.ac.jp/statistics/ (Fig. 2). It can show what genomes are available for any particular taxon. Taxon names are clickable and lead to a page with statistics for the clicked taxon. In this way, the entire sequenced subset of the taxonomic tree can be navigated (only nodes that have genomes are shown). By default the table includes only immediate child taxa of the selected taxon, but a larger subtree can be shown by tweaking the "Tree depth" option on the page and clicking "Apply".

The statistics table displays the total number of genomes included in GenomeSync for each taxon. It also shows the number of species, genera, families and orders represented by those genomes. When the "Sequenced proportion" option is activated, the table also shows the total number of species, genera, families and orders described in the current taxonomy for each included taxon, as well as percentage of those represented by at least one genome among these total numbers. This data is useful for seeing how thoroughly a particular taxon of interest is represented by genome data. The percentage among known species (or genera, or families) is important for metagenomic applications, as it provides at least some measure of confidence that an unknown sequence can be correctly classified at the species level (or genus, or family level).

The “NAF size” and “FASTA size” options allow inspecting the total size of all genomes for each taxon shown in the table, in NAF and FASTA formats, respectively. It is important to check these sizes before trying to download the genomes.

The “Show genomes” option enables showing the links to individual genomes. Only genomes corresponding to currently visible taxa will be shown, but not those for sub-taxa. E.g., if this option is activated for the view as shown on Fig. 2 (showing the “Cellular organisms” taxon with tree depth 1), no genome links will appear, because the genomes correspond to nodes deeper in the taxonomy tree, and no genome is bound to the “Archaea”, “Bacteria”, or “Eukaryota” taxon directly. When the individual genomes are shown, it includes the accession number, the link to the NCBI Assembly page (if applicable), the download link for the genome file in NAF format, and the size of the NAF file.

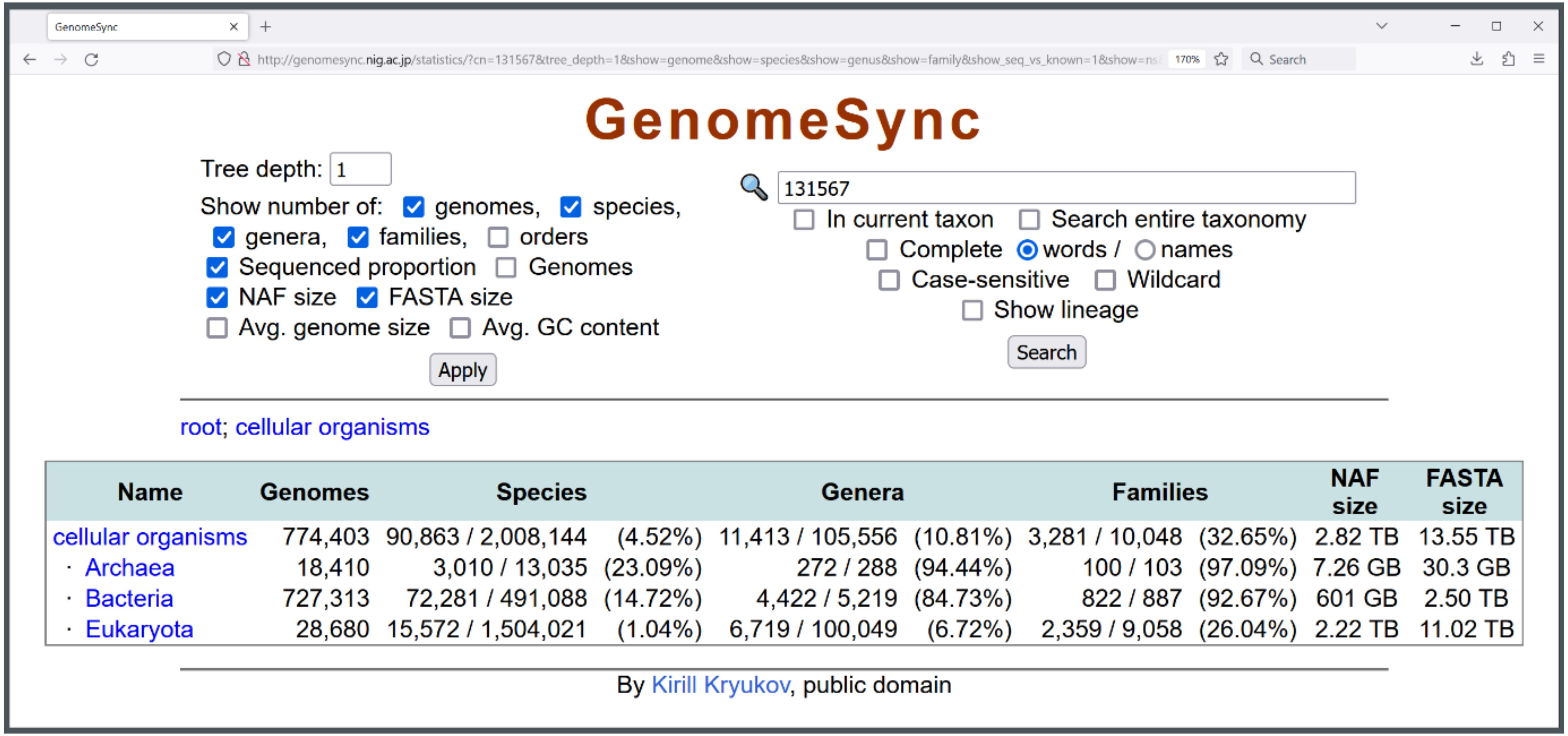
Using those options, the “Statistics” page effectively answers the most common questions related to discovery of genome data. Example questions: “How many mammalian species have genomes?”, “What percentage of fungal genera are represented by genome data?”.

Search by name

The search interface at the top-right of the “Statistics” page allows searching for a taxon by name, or part of a name. Both latin names or common names registered in the NCBI Taxonomy are checked by the search. Marking the “In current taxon” checkbox will restrict the search to the currently selected taxon and its subtree. The “Search entire taxonomy” checkbox switches the search to use the entire taxonomy, rather than just subset with genomes.

By default the search will return names containing the query text anywhere. However, when the “Complete” checkbox is marked, the search will take into account the “words” / ”names” selection. If the “words” option is selected, only entries containing the query as a separate word (not part of a longer word) will be found. If the “names” is selected, only a perfectly matching entry will be shown, if such entry exists.

Fig. 2. Screenshot of the GenomeSync statistics webpage, showing the subtree of depth 1, rooted at the “cellular organisms” taxon (as of February 13, 2024).


GenomeSync
http://genomesync.nig.ac.jp/statistics/?cn=131567&tree_depth=1&show=genome&show=species&show=genus&show=family&show_seq_vs_known=1&show=ns
170%
Search

GenomeSync

Tree depth: 1

Show number of: ☑ genomes, ☑ species, ☑ genera, ☑ families, □ orders
☑ Sequenced proportion □ Genomes
☑ NAF size ☑ FASTA size
□ Avg. genome size □ Avg. GC content

Apply

131567
□ In current taxon □ Search entire taxonomy
□ Complete ● words / ○ names
□ Case-sensitive □ Wildcard
□ Show lineage

Search

root; cellular organisms

Name	Genomes	Species		Genera		Families		NAF size	FASTA size
cellular organisms	774,403	90,863 / 2,008,144	(4.52%)	11,413 / 105,556	(10.81%)	3,281 / 10,048	(32.65%)	2.82 TB	13.55 TB
· Archaea	18,410	3,010 / 13,035	(23.09%)	272 / 288	(94.44%)	100 / 103	(97.09%)	7.26 GB	30.3 GB
· Bacteria	727,313	72,281 / 491,088	(14.72%)	4,422 / 5,219	(84.73%)	822 / 887	(92.67%)	601 GB	2.50 TB
· Eukaryota	28,680	15,572 / 1,504,021	(1.04%)	6,719 / 100,049	(6.72%)	2,359 / 9,058	(26.04%)	2.22 TB	11.02 TB

By Kirill Kryukov, public domain

The “Case-sensitive” checkbox makes the search case-sensitive. By default, the search is case insensitive. The “Wildcard” checkbox activates wildcard search, where each asterisk (*) in the query will match any text. The “Show lineage” checkbox will result in each search result shown together with its entire taxonomic lineage from the root. The results are shown in the alphabetic order. In case if the “Show lineage” option is enabled, the entire lineage is used for alphabetic sorting. With this option, if the species name (or any other name or part of name) used as a query is missing in GenomeSync (does not have any genome assembly yet), it is possible to know what is the nearest taxon present in GenomeSync (taxon that has any genomes). Thus this search aids in genome data discovery even when the exact query is missing genome data, and it can help to locate the available genomes of the closest related organisms.

Downloading individual genomes

The current up-to-date instructions for downloading GenomeSync data are available at the database homepage: https://genomesync.org/. The website instructions take priority when there is any discrepancy between the website and this paper.

The simplest way to download individual genomes from GenomeSync is by opening the GenomeSync data repository (http://genomesync.nig.ac.jp/naf/) in a web browser, then locating and downloading a genome of interest. This method requires first locating the genome, which may be difficult. An easier method is opening the statistics interface (http://genomesync.nig.ac.jp/statistics/), navigating to (or searching for) a taxon of interest, enabling the "Show genomes" option, and then downloading the genomes using the "↓NAF" links.

Another way is to use the GenomeSync Genome Selector tool (http://genomesync.nig.ac.jp/selector/). One or several taxon names can be inputted in this tool, and it will then produce the list of direct links to genome files in the NAF format. Multiple taxon names can be listed using ";;" as a separator. Prepending each taxon name with "(rep)" or "(rep2)" will restrict the selection to the corresponding representative subset. Adding "-" before each name will instruct the tool to exclude this taxon, and all its genomes, from selection. Inclusions and exclusions are processed in the left-to-right order, as they are listed. For example, all reptile genomes can be selected using the "Sauropsida;;-Aves" query. This allows creating complex custom taxonomic selections of genomes.

Downloading genomes in bulk

In many cases, a large set of genomes may be needed. This is difficult and time-consuming when downloading genomes one by one. Thus, GenomeSync also makes it possible to automate the process and download all genomes of a particular taxon in one step. This kind of data collection is usually done prior to some automatic bioinformatic analysis, which normally runs in the Unix environment. Therefore all examples in this paper are shown for a Unix command line, tested on Ubuntu Linux.

Automated downloading of bulk genome data from GenomeSync is done using these command line tools: curl (curl, 2024) and wget (wget, 2024). This is how these tools can be installed on Ubuntu Linux:

```
sudo apt install curl wget
```

To download all genomes contained in GenomeSync, the following command can be used:

```
wget --directory-prefix=./GenomeSync -r -np -N -l inf -nH -A '*.naf' http://genomesync.nig.ac.jp/naf/
```

This command will store the entire collection of genomes to the GenomeSync directory, located under the current directory. The directory will be created if missing. It's important to make sure that there is enough free disk space before trying this.

Downloading genomes for particular taxon is possible by combining this method with the Genome Selector tool. For example, this command will download all mammalian genomes:

```
curl -s 'http://genomesync.nig.ac.jp/selector/?t=Mammalia' | wget -i - --directory-prefix=./GenomeSync -x -N -nH -nv
```

This command will download the genomes into the `./GenomeSync` directory, re-creating the relevant part of the original GenomeSync directory structure inside it. It is possible to download the genomes directly into the destination directory, without creating any other directories, by omitting the `-x` option from the `wget` command.

Downloading representative genomes is possible by adding "(rep)" or "(rep2)" in front of the taxon name in the selector link. For example, this command will download representative prokaryote genomes (those from the "rep" subset):

```
curl -s 'http://genomesync.nig.ac.jp/selector/?t=(rep)Archaea&t=(rep)Bacteria' | wget -i - --directory-prefix=./GenomeSync -x -N -nH -nv
```

It's possible to combine representative and complete selections freely, to construct complex subsets. This example command will download representative bacteria genomes (from the "rep" subset), all archaea genomes, and the human genome:

```
curl -s 'http://genomesync.nig.ac.jp/selector/?t=Archaea&t=(rep)Bacteria&t=Homo sapiens' | wget -i - --directory-prefix=./GenomeSync -x -N -nH -nv
```

Synchronizing

Many new genomes appear every week, and sometimes old assemblies are updated with new builds. Thus, any static collection of genomes gradually becomes more and more out of date. Therefore, it's important to periodically synchronize the local collection of genomes, to make it up-to-date with the upstream GenomeSync database. Synchronization is performed in two steps. First, outdated genomes are deleted. Then new genomes are downloaded. This can be done on any taxonomically specified subset of GenomeSync.

For example, we will consider a scenario where representative archaea genomes were previously downloaded into the `./GenomeSync` directory, and now are going to be updated. The following command will remove the outdated NAF files:

```
(cd GenomeSync/naf; comm -1 -3 <(curl -s 'http://genomesync.nig.ac.jp/selector/?paths=1&t=(rep)Archaea' | sort) <(find * -type f | sort) | xargs -d '\n' rm)
```

This command works by downloading the list of genome file paths from the Genome Selector tool, comparing it with the list of local NAF files, then deleting all files that are missing in the first list, but present in the second one. It may be preferable to execute it part by part, in order to confirm the file lists before deleting the files:

Step 1: Downloading the list of representative archaea genomes from the upstream GenomeSync database, and printing the number of genomes:

```
curl -s 'http://genomesync.nig.ac.jp/selector/?paths=1&t=(rep)Archaea' | sort | tee pathlist-upstream.txt | wc -l
```

Step 2: Making the list of local genome NAF files, and also printing the number of locally installed genomes:

```
(cd GenomeSync/naf; find * -type f) | sort | tee pathlist-local.txt | wc -l
```

Step 3: Comparing the two lists, and saving the list of files to be deleted (files missing in the first list, but present in the second one). Also printing the number of files to be deleted.

```
comm -1 -3 pathlist-upstream.txt pathlist-local.txt | tee files-to-delete.txt | wc -l
```

Step 4. Deleting the files listed in the `files-to-delete.txt` file.

```
cat files-to-delete.txt | (cd GenomeSync/naf; xargs -d '\n' rm)
```

After the outdated files are deleted, the new genomes can be downloaded:

```
curl -s 'http://genomesync.nig.ac.jp/selector/?t=(rep)Archaea' | wget -i - --directory-prefix=./GenomeSync -x -nH -nc -nv
```

If the entire GenomeSync snapshot has to be synchronized, the above method can be used with “root” specified as a taxon.

Verifying

Verifying the integrity of downloaded data is important when working with large datasets. The list of MD5 hashes of NAF files for any selection of genomes can be obtained from the Genome Selector tool, by adding “`naf-md5=1`” to the request. For example, the following command will obtain the list of MD5 hashes for representative Archaea genomes:

```
curl -s 'http://genomesync.nig.ac.jp/selector/?naf-md5=1&t=(rep)Archaea'
```

This list can be saved to a file and used for verifying the genomes (located in the `./GenomeSync` directory):

```
curl -s 'http://genomesync.nig.ac.jp/selector/?naf-md5=1&t=(rep)Archaea' >naf.md5
cat naf.md5 | (cd GenomeSync/naf; md5sum -c --quiet)
```

It’s also possible to verify the files in one step, without saving the list into a file:

```
curl -s 'http://genomesync.nig.ac.jp/selector/?naf-md5=1&t=(rep)Archaea' | (cd
GenomeSync/naf; md5sum -c --quiet)
```

The command can be modified to produce the list of files with mismatching hashes:

```
curl -s 'http://genomesync.nig.ac.jp/selector/?naf-md5=1&t=(rep)Archaea' | (cd
GenomeSync/naf; md5sum -c --quiet 2>/dev/null) | grep -P ': FAILED$' | sed -r 's/:
\sFAILED\s*$//'
```

Also, the command can be extended to immediately delete mismatching files:

```
curl -s 'http://genomesync.nig.ac.jp/selector/?naf-md5=1&t=(rep)Archaea' | (cd
GenomeSync/naf; md5sum -c --quiet 2>/dev/null | grep -P ': FAILED$' | sed -r 's/:
\sFAILED\s*$//' | xargs -d '\n' rm)
```

After the mismatching files have been deleted, the correct files can be re-downloaded for the particular subset of genomes, using commands from the previous sections.

One common scenario is when a set of genomes has been downloaded and verified previously. Now a user is planning to run an update to synchronize the local genomes to the current GenomeSync. In such a case it would be reasonable to verify only newly downloaded files, rather than verifying an entire collection of genomes. This can be done using the following steps, assuming that the previous set of MD5 hashes (corresponding to the current local set of genomes) is stored in "`naf.md5`":

Step 1. Rename the old "`naf.md5`", download the current hashes, and extract those that correspond to the new files:

```
mv naf.md5 naf-old.md5
curl -s 'http://genomesync.nig.ac.jp/selector/?naf-md5=1&t=(rep)Archaea' >naf.md5
comm -1 -3 <(sort naf-old.md5) <(sort naf.md5) >naf-new-only.md5
```

Step 2. Perform the synchronization as described in the "Synchronizing" section above.

Step 3. Verify the newly downloaded genomes using the "`naf-new-only.md5`" file.

Step 4. Remove the files "`naf-old.md5`" and "`naf-new-only.md5`", but keep "`naf.md5`" for future updates.

Update schedule

It's important to have mechanisms ensuring data consistency when working with large datasets. If the upstream database was modified while a user is still downloading the data, the downloaded dataset might be inconsistent. Therefore, it's important to consider the possibility of such update, when performing large scale downloading or synchronization.

When downloading or synchronizing a large number of genomes from GenomeSync, we recommend the following procedure to ensure that the resulting downloaded dataset is self-consistent: 1. Check the date of the latest update on the GenomeSync webpage. 2. Perform downloading or synchronization. 3. Check the date of the latest update on the GenomeSync webpage again. 4. If the date differs from the initial date, indicating that GenomeSync was updated while the download or synchronization was ongoing, perform an additional synchronization step. The update date on GenomeSync webpage is changed as the last step, only after the update is complete. Therefore it provides a robust mechanism for ensuring integrity of the downloaded data.

Taxonomy

Taxonomic classification gives structure to genome data. It is essential for organizing and using the genomes, and for interpreting results of comparisons. GenomeSync uses NCBI Taxonomy (Schoch et al., 2020), and has all genomes named using their corresponding taxonomic names. NCBI Taxonomy undergoes frequent updates, which include adding, deleting, merging and renaming nodes. These changes happen independently of the changes in GenBank and RefSeq databases. Therefore, when downloading data directly from NCBI Taxonomy and GenBank/RefSeq, the genome data and taxonomy are usually not consistent, and require an additional step of resolving conflicts and inconsistencies. GenomeSync avoids these issues, by including a copy of taxonomy data in its updates. Each GenomeSync update provides both genomes and taxonomy, which are 100% consistent to each other. GenomeSync current taxonomy data is available at http://genomesync.nig.ac.jp/taxonomy/taxdmp.zip. The guaranteed consistency between GenomeSync genome names and taxonomy allows focusing on using the genomes, without spending time for harmonizing genomes with taxonomy.

Discussion

We have utilized GenomeSync to identify bacterial pathogen(s) in a given clinical sample. 16S ribosomal RNA (rRNA) genes were enriched and sequenced using MinION developed by Oxford Nanopore Technologies (Mitsuhashi et al., 2017; Nakagawa et al., 2019; Kai et al., 2019; Matsuo et al., 2021; Ohno et al., 2021; Komiya et al., 2022). Nanopore sequencing enables us to obtain almost full-length regions of 16S rRNA genes and analyze the reads even while sequencing is running. For the sequencing data analysis, we utilized GenomeSync, in particular for a representative dataset of bacteria and archaea genomes as well as human genome with blastn (Mitsuhashi et al., 2017; Nakagawa et al., 2019) and/or minimap2 (Nakagawa et al., 2019; Kai et al., 2019; Matsuo et al., 2021; Ohno et al., 2021; Komiya et al., 2022) program. The data analysis system including reference genomes was built on a laptop computer and used in a portable situation (Nakagawa et al., 2019).

In addition, we used GenomeSync to identify viral sequences integrated in eukaryotic genomes (Kryukov et al., 2019a). It is known that various viral sequences were integrated in host genomes, such sequences are called endogenous viral elements (EVEs). Using GenomeSync (as of August 14, 2017), we obtained 4,102 eukaryotic and 7,007 viral genomes. We then conducted one-to-one genome sequence comparison and identified EVEs comprehensively, which is shared in the pEVE website (http://peve.med.u-tokai.ac.jp/, Kryukov et al., 2019a).

GenomeSync has been maintained since February 2015, with updates usually done at least monthly. At the beginning, only 14,720 genomes of 8,578 species were stored in the database. As of February 20, 2024, 860,376 genomes are stored in GenomeSync, representing the increase by a factor of 58 times. Since the rate of accumulation of genome sequences continues to increase, the availability of automatically synchronizable genome datasets will become essential for large scale genome-based analyses. The automation, robustness and compactness provided by GenomeSync will accelerate genome-based discoveries and will be increasingly useful in the future.

Declarations

Availability of data and material

All data is available online at the GenomeSync website: http://genomesync.org/

Competing interests

The authors declare no competing interests.

Funding

This work was supported by the JSPS KAKENHI Grants-in-Aid for Scientific Research (C) (20K06612 to K.K.).

Acknowledgements

We appreciate the financial support of Genesis Healthcare Co.

References

Basenko, E. Y., Pulman, J. A., Shanmugasundram, A., Harb, O. S., Crouch, K., Starns, D., Warrenfeltz, S., Aurrecoechea, C., Stoeckert, C. J., Jr, Kissinger, J. C., Roos, D. S., & Hertz-Fowler, C. (2018). FungiDB: An Integrated Bioinformatic Resource for Fungi and Oomycetes. Journal of fungi (Basel, Switzerland), 4(1), 39.

Challis, R. J., Kumar, S., Dasmahapatra, K. K., Jiggins, C. D., Blaxter, M. (2016). Lepbase: the Lepidopteran genome database. bioRxiv 056994.

curl (2024). curl. https://curl.se/. Accessed 16 February 2024.

Falk, T., Herndon, N., Grau, E., Buehler, S., Richter, P., Zaman, S., Baker, E. M., Ramnath, R., Ficklin, S., Staton, M., Feltus, F. A., Jung, S., Main, D., & Wegrzyn, J. L. (2018). Growing and cultivating the forest genomics database, TreeGenes. Database : the journal of biological databases and curation, 2018, 1–11.

Harris, T. W., Arnaboldi, V., Cain, S., Chan, J., Chen, W. J., Cho, J., Davis, P., Gao, S., Grove, C. A., Kishore, R., Lee, R., Muller, H. M., Nakamura, C., Nuin, P., Paulini, M., Raciti, D., Rodgers, F. H., Russell, M., Schindelman, G., Auken, K. V., … Sternberg, P. W. (2020). WormBase: a modern Model Organism Information Resource. Nucleic acids research, 48(D1), D762–D767.

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaigui, B., Fatima, R., Gall, A., … Flicek, P. (2021). Ensembl 2021. Nucleic acids research, 49(D1), D884–D891.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C., & Banfield, J. F. (2016). A new view of the tree of life. Nature microbiology, 1, 16048.

Inoue, J., & Saitou, N. (2021). dbCNS: A New Database for Conserved Noncoding Sequences. Molecular biology and evolution, 38(4), 1665–1676.

Kai, S., Matsuo, Y., Nakagawa, S., Kryukov, K., Matsukawa, S., Tanaka, H., Iwai, T., Imanishi, T., & Hirota, K. (2019). Rapid bacterial identification by direct PCR amplification of 16S rRNA genes using the MinION™ nanopore sequencer. FEBS open bio, 9(3), 548-557.

Kans J. (2022). Entrez Direct: E-utilities on the Unix Command Line. https://www.ncbi.nlm.nih.gov/books/NBK179288/

Kim, B. Y., Wang, J. R., Miller, D. E., Barmina, O., Delaney, E., Thompson, A., Comeault, A. A., Peede, D., D'Agostino, E. R., Pelaez, J., Aguilar, J. M., Haji, D., Matsunaga, T., Armstrong, E. E., Zych, M., Ogawa, Y., Stamenkovic-Radak, M., Jelic, M., Veselinovic, M. S., Tanaskovic, M., … Petrov, D. A. (2021). Highly contiguous assemblies of 101 drosophilid genomes. eLife, 10, e66405.

Komiya, S., Matsuo, Y., Nakagawa, S., Morimoto, Y., Kryukov, K., Okada, H., & Hirota, K. (2022). MinION, a portable long-read sequencer, enables rapid vaginal microbiota analysis in a clinical setting. BMC medical genomics, 15(1), 68.

Kryukov, K., Ueda ,M. T., Imanishi, T., & Nakagawa, S. (2019). Systematic survey of non-retroviral virus-like elements in eukaryotic genomes. Virus research. 262, 30-36.

Kryukov, K., Ueda, M. T., Nakagawa, S., & Imanishi, T. (2019). Nucleotide Archival Format (NAF) enables efficient lossless reference-free compression of DNA sequences. Bioinformatics, 35(19), 3826–3828.

Kryukov, K., Ueda, M. T., Nakagawa, S., & Imanishi, T. (2020). Sequence Compression Benchmark (SCB) database-A comprehensive evaluation of reference-free compressors for FASTA-formatted sequences. GigaScience, 9(7), giaa072.

Larkin, A., Marygold, S. J., Antonazzo, G., Attrill, H., Dos Santos, G., Garapati, P. V., Goodman, J. L., Gramates, L. S., Millburn, G., Strelets, V. B., Tabone, C. J., Thurmond, J., & FlyBase Consortium (2021). FlyBase: updates to the Drosophila melanogaster knowledge base. Nucleic acids research, 49(D1), D899–D907.

Li, F. W., Brouwer, P., Carretero-Paulet, L., Cheng, S., de Vries, J., Delaux, P. M., Eily, A., Koppers, N., Kuo, L. Y., Li, Z., Simenc, M., Small, I., Wafula, E., Angarita, S., Barker, M. S., Bräutigam, A., dePamphilis, C., Gould, S., Hosmani, P. S., Huang, Y. M., … Pryer, K. M. (2018). Fern genomes elucidate land plant evolution and cyanobacterial symbioses. Nature plants, 4(7), 460–472.

Li, W., O'Neill, K. R., Haft, D. H., DiCuccio, M., Chetvernin, V., Badretdin, A., Coulouris, G., Chitsaz, F., Derbyshire, M. K., Durkin, A. S., Gonzales, N. R., Gwadz, M., Lanczycki, C. J., Song, J. S., Thanki, N., Wang, J., Yamashita, R. A., Yang, M., Zheng, C., Marchler-Bauer, A., … Thibaud-Nissen, F. (2021). RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. Nucleic acids research, 49(D1), D1020–D1028.

Liew, Y. J., Aranda, M., & Voolstra, C. R. (2016). Reefgenomics.Org - a repository for marine genomics data. Database : the journal of biological databases and curation, 2016, baw152.

Matsuo, Y., Komiya, S., Yasumizu, Y., Yasuoka, Y., Mizushima, K., Takagi, T., Kryukov, K., Fukuda, A., Morimoto, Y., Naito, Y., Okada, H., Bono, H., Nakagawa, S., & Hirota, K. (2021). Full-length 16S rRNA gene amplicon analysis of human gut microbiota using MinION™ nanopore sequencing confers species-level resolution. BMC microbiology, 21(1), 35.

Mitsuhashi, S., Kryukov, K., Nakagawa, S., Takeuchi, J. S., Shiraishi, Y., Asano, K., & Imanishi, T. (2017). A portable system for rapid bacterial composition analysis using a nanopore-based sequencer and laptop computer. Scientific reports, 7(1), 5657.

Nakagawa, S., Inoue, S., Kryukov, K., Yamagishi, J., Ohno, A., Hayashida, K., Nakazwe, R., Kalumbi, M., Mwenya, D., Asami, N., Sugimoto, C., Mutengo, M. M., & Imanishi, T. (2019). Rapid sequencing-based diagnosis of infectious bacterial species from meningitis patients in Zambia. Clinical translational immunology 8 (11), e01087.

NCBI Datasets (2021). NCBI Datasets. https://www.ncbi.nlm.nih.gov/datasets/. Accessed 30 June 2021.

Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., Smirnova, T., Grigoriev, I. V., & Dubchak, I. (2014). The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic acids research, 42(Database issue), D26–D31.

Ohno, A., Umezawa, K., Asai, S., Kryukov, K., Nakagawa, S., Miyachi, H., & Imanishi, T. (2021). Rapid profiling of drug-resistant bacteria using DNA-binding dyes and a nanopore-based DNA sequencer. Scientific reports, 11(1), 3436.

Parks, D. H., Chuvochina, M., Chaumeil, P. A., Rinke, C., Mussig, A. J., & Hugenholtz, P. (2020). A complete domain-to-species taxonomy for Bacteria and Archaea. Nature biotechnology, 38(9), 1079–1086.

Phytozome (2021). Phytozome. https://phytozome.jgi.doe.gov/. Accessed 30 June 2021.

PineRefSeq (2021). PineRefSeq. https://nealelab.ucdavis.edu/pinerefseq/. Accessed 30 June 2021.

Piro, V. C. (2020). genome_updater. https://github.com/pirovc/genome_updater. Accessed 30 June 2021.

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., Haase, B., … Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. Nature, 592(7856), 737–746.

Sayers, E. W., Beck, J., Bolton, E. E., Bourexis, D., Brister, J. R., Canese, K., Comeau, D. C., Funk, K., Kim, S., Klimke, W., Marchler-Bauer, A., Landrum, M., Lathrop, S., Lu, Z., Madden, T. L., O'Leary, N., Phan, L., Rangwala, S. H., Schneider, V. A., Skripchenko, Y., … Sherry, S. T. (2021). Database resources of the National Center for Biotechnology Information. Nucleic acids research, 49(D1), D10–D17.

Sayers, E. W., Cavanaugh, M., Clark, K., Pruitt, K. D., Schoch, C. L., Sherry, S. T., & Karsch-Mizrachi, I. (2021). GenBank. Nucleic acids research, 49(D1), D92–D96.

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., & Karsch-Mizrachi, I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database : the journal of biological databases and curation, 2020, baaa062.

Segerman B. (2012). The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories. Frontiers in cellular and infection microbiology, 2, 116.

UCSC (2021). Sequence and Annotation Downloads. https://hgdownload.soe.ucsc.edu/. Accessed 28 June 2021.

Warrenfeltz, S., Basenko, E. Y., Crouch, K., Harb, O. S., Kissinger, J. C., Roos, D. S., Shanmugasundram, A., & Silva-Franco, F. (2018). EuPathDB: The Eukaryotic Pathogen Genomics Database Resource. Methods in molecular biology (Clifton, N.J.), 1757, 69–113.

wget (2024). wget. https://www.gnu.org/software/wget/. Accessed 16 February, 2024.

Zhao, Z., Cristian, A., & Rosen, G. (2020). Keeping up with the genomes: efficient learning of our increasing knowledge of the tree of life. BMC bioinformatics, 21(1), 412.

Zhu, Q., Mai, U., Pfeiffer, W., Janssen, S., Asnicar, F., Sanders, J. G., Belda-Ferre, P., Al-Ghalith, G. A., Kopylova, E., McDonald, D., Kosciolek, T., Yin, J. B., Huang, S., Salam, N., Jiao, J. Y., Wu, Z., Xu, Z. Z., Cantrell, K., Yang, Y., Sayyari, E., … Knight, R. (2019). Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. Nature communications, 10(1), 5477.

Zoonomia Consortium (2020). A comparative genomics multitool for scientific discovery and conservation. Nature, 587(7833), 240–245.

Supplementary Material

Suppl Table 1. GenomeSync directory structure rules (as of February 13, 2024).

https://genomesync.org/Supplementary-Information/TableS1-Directory-structure.pdf

Edited by SAITOU Naruya