

Overview of Genomic Databases

Autor: Mateusz Oleszek, nr. 144608

Genomic databases play a crucial role in the field of genetics and genomics, providing researchers with vast repositories of genetic information for analysis and exploration. These databases are comprehensive collections of genomic data, including DNA sequences, gene annotations, genetic variants, and other related information. They serve as valuable resources for studying the structure, function, and evolution of genomes across various species. Genomic databases store data from diverse sources, such as genome sequencing projects, research studies, and public contributions. They enable researchers to access and analyze genetic data on a large scale, facilitating the discovery of genetic variations associated with diseases, the identification of gene functions, and the exploration of genetic relationships between different organisms.

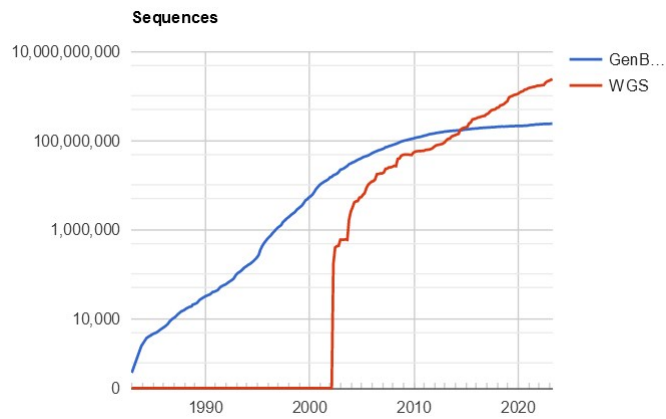
The Los Alamos Sequence Database, established in 1979 and later known as GenBank, served as an early repository for biological sequences. In 1982, GenBank was transferred to the National Center for Biotechnology Information (NCBI), where it currently resides. By the end of 1983, GenBank stored over 2,000 sequences with nearly 1 million base pairs. Concurrently, the International Nucleotide Sequence Database Collaboration (INSDC) was formed by NCBI, EMBL, and DDBJ to manage the growing amount of nucleotide and amino acid sequence data. The INSDC databases have witnessed exponential growth, now housing over 95 billion base pairs.

To handle the vast amount of raw sequence data, specialized databases have been created, including genome browsers, model organism databases, and process-specific databases. The number of genomics databases and tools has significantly increased. Furthermore, it is estimated that there are approximately 3,000 distinct genomic resources, tools, and databases available online.

Sequence Data Repositories:

The International Nucleotide Sequence Database Collaboration (INSDC) comprises GenBank, EMBL, and DDBJ, which accept sequence submissions from researchers worldwide. Each submission receives a unique identification number and is stored in a library. However, the exponential growth of data poses challenges in maintaining accuracy and accessibility. Duplication and incomplete or incorrect information are common issues. Efforts like RefSeq aim to curate and correct sequence data, but the vast amount of data makes it difficult to annotate and organize every entry.

Another challenge is the lack of context and annotation in repository data, making them less useful for research. Developers have repurposed GenBank data to create specialized databases, tailored to specific aesthetics, project history, and community needs. These databases provide better organization and access to genomic data. The main categories of such databases are further discussed in the following sections.



Wzrost ilości danych w bazie GenBank na przestrzeni lat

General Genome Browsers:

To address the research needs of scientists, general genome browsers have emerged as valuable tools. Examples of successful browsers include the UCSC Genome Browser, Ensembl by EBI, and MapViewer by NCBI. These browsers repackage genomic and gene annotation data from databases like GenBank, offering a comprehensive genomic context for specific genome features, such as genes or disease loci. Additionally, these browsers facilitate cross-species comparisons by displaying information in common formats, enabling easier visualization and extraction of data. For instance, users can search for a specific genomic region, like a disease gene, and access visual displays of the corresponding sequence and annotations. These displays are linked to supplementary data and databases for further investigation and also provide links back to the original data sources.

Switch view

Search organisms

Homo sapiens (human)

To view more organisms in the tree, click on nodes that have "+" signs. Press and hold the "+" to expand and reveal all the subgroups. Or, search for an organism using the search box above.

New! Click on Switch view at the top to see another way of navigating genomes.

Homo sapiens (human)

Search in genome

Location, gene or phenotype

Complete: 175,211,747,760,000 7,890,000,000 DNA repeat

Assembly

GRCh38 p14

Browse genome

Compare genomes

Assembly details

Name: GRCh38 p14

RefSeq accession: GCF_000001405.40

GenBank accession: GCA_000001405.29

Submitter: Genome Reference Consortium

Level: Chromosome

Category: Reference genome

Annotation details

Annotation Release: RS_2023_03

Release date: Mar 28, 2023

Tools

All tools

BioMart >

Export custom datasets from Ensembl with this data-mining tool

BLAST/BLAT >

Search our genomes for your DNA or protein sequence

Variant Effect Predictor >

Analyse your own variants and predict the functional consequences of known and unknown variants

Search

All species

for

Go

e.g. BRCA2 or rat 542797383.63627669 or rs699 or coronary heart disease

All genomes

Select a species

Pig breeds

Pig reference genome and 12 additional breeds

View full list of all species

Favourite genomes

Human

GRCh38 p13

384,449,916 GRCh37

Mouse

GRCh38

Zebrafish

GRCh11

Compare genes across species

Find SNPs and other variants for my gene

Gene expression in different tissues

Retrieve gene sequence

UNIVERSITY OF CALIFORNIA SANTA CRUZ

Genomics Institute

UCSC

Genome Browser

Genomes

Genome Browser

Tools

My Data

Projects

Help

About Us

Tools

Genome Browser

Interactively visualize genomic data

BLAT

Rapidly align sequences to the genome

In-Silico PCR

Rapidly align PCR primer pairs to the genome

Table Browser

Download and filter data from the Genome Browser

LiftOver

Convert genome coordinates between assemblies

REST API

Returns data requested in JSON format

Variant Annotation Integrator

Annotate genomic variants

More tools...

Sharing data

Strony wspomnianych genomowych baz danych

Standardized Genome Database Tools: GMOD:

GMOD is an open-source standard database and visualization toolset that promotes standardized querying, browsing, and usage of genome databases across species. Several species- and taxa-specific genome databases have adopted GMOD tools to enhance their annotation, visualization, and query options. The goal is to facilitate research and comparative studies by providing a common framework for database development and data exploration.

Str. 3 / 5

Specific Databases:

Species-specific or taxa-specific genome databases have been developed to provide deeper information about various genomes. These databases are publicly available and often curated. They offer accurate annotations and incorporate species-specific data types. Researchers can access these databases for a more in-depth analysis of specific genomes.

Subject-specific databases focus on specific biological data categories such as protein domains, protein structures, expression data, and genome-wide association studies. These databases serve as valuable resources for researchers working in specific areas of study. However, the large number of these databases can lead to redundancy and lack of integration.

RGD virtual office hours are available by appointment. [Contact us](#) to schedule a time.

Search: Genes, Strains, Ontology & Annotation, Ontomate (Literature), QTL, Orthologs, Genomic Region, All...

Analysis and Visualization: JBrowse Genome Browser, Variant Visualizer, VCMAP Synteny Browser (beta), OLGA Gene List Generator, Disease Portals, Phenotypes and Models.

RGD Video Tutorials: Comparison of Genomic Variants in HRDP Inbred Strains With Two Rat Genome Reference, PhenoMiner Tool, Rat Strain Updates at the Rat Genome Database, Rat Reference Genome mRatBN7.2 Curation, Quantitative phenotype data for HRDP strains at the RGD, Multi-Ontology Enrichment Tool (MOET): a web-based tool for ontology analysis at RGD, RGD: A multispecies resource to explore complex diseases, Molecular Pathways, GA Tool (Gene Annotator), OLGA (Gene list builder and analyzer).

Baza danych genomu szczurów

Solutions to the Current Challenges of Accuracy and Curation:

Ensuring accurate data and efficient management and curation are ongoing challenges for genetic databases. Proposed solutions include education of database biocurators, standardized inclusion of sequence data and references in publications, and community curation. While some databases have embraced community curation, others like GenBank, have concerns about maintaining authoritative repositories. Efforts to incentivize researchers to contribute and improve the reliability of curation are ongoing, even if they are hampered by factors like the reliability of curation.

Conclusion:

Genomic data resources, including sequence data repositories, general genome browsers, species- and taxa-specific databases, and subject-specific databases, have revolutionized the field of bioinformatics. They provide researchers with invaluable tools to explore and analyze genetic information. However, challenges such as maintaining accuracy, accessibility, and organization of data persist.

When it comes to databases containing human genomic data a big problem of privacy and security also emerges, posing both technical and societal challenge. Such as sufficient anonymization, confidentiality and managing access. They require designing the systems from the ground up with privacy in mind. Those concerns have necessitated the removal of large datasets from public databases in the past.

Another obstacle is the overwhelming number of available genomic resources, making it difficult for researchers to locate and effectively utilize them. Despite these challenges, the rewards and opportunities provided by genomic data have been significant, enabling new discoveries and interdisciplinary relationships.

Looking ahead, as the amount of data continues to grow, it is crucial for the scientific community to find solutions, such as advanced training, to ensure that amazing discoveries can be made.

References

The essay was based on "Genomic Data Resources: Challenges and Promises" by Warren C. Lathe III (OpenHelix), Jennifer M. Williams (OpenHelix), Mary E. Mangan (OpenHelix), and Donna Karolchik (University of California, Santa Cruz Genome Bioinformatics Group), published in Nature Education in 2008 (Lathe, Williams, Mangan, & Karolchik, 2008, p. 2).

Wan, Z., Hazel, J.W., Clayton, E.W. et al. Sociotechnical safeguards for genomic data privacy. Nat Rev Genet 23, 429–445 (2022)