

Decoding the NCBI Database: A Gateway to Computational Biology and Biomedical/Research

Vivek Chopra^{*1}, Rajesh Kumar¹, Meenam Bhatia², Puneet Bansal¹, Gouri¹, Priyanshi¹, Yukti Kasodniya¹, Priyanka Panwar¹, Nayan Chakraborty¹, Abhishek Bhatt¹, Prerna Kumari¹, Anupriya¹, Shafaque Shamim¹, Kavya Shree¹, Sukaran Malhotra¹, Tarun Rawat¹, Sukanya¹, Amrita¹ and Indrajit Sarkar¹

¹ Department of Botany, Hindu College, University of Delhi, Delhi 110 007, India

² Department of Botany, Daulat Ram College, University of Delhi, Delhi 110 007, India

(Received 17 July, 2025; Accepted 26 September, 2025)

ABSTRACT

Established in 1988, the National Center for Biotechnology Information (NCBI), part of the U.S. National Library of Medicine (NLM) and the National Institutes of Health (NIH), was created to advance resources and tools that support biomedical and genomic research. This paper outlines the main databases and digital tools maintained by NCBI, focusing on their structure, function, and significance in the life sciences and biomedical sciences. Its vast array of resources can be categorized into six key domains: scientific literature, nucleotide and genome sequences, protein and structural data, gene expression and epigenetic profiles, genetic variation and phenotype information, and computational and chemical analysis tools. Since taking over the management of GenBank in 1992, NCBI has worked in close coordination with global partners such as the European Molecular Biology Laboratory (EMBL) and the DNA Data Bank of Japan (DDBJ) to curate and share sequence data. These resources help researchers across the world in retrieving biomedical literature, exploring molecular structures and identifying genetic variants. By integrating vast datasets with genomics, proteomics, and transcriptomics, researchers can conduct advanced studies. The present study provides a compilation of the data, tools, and resources available on NCBI and their applications across various forms of biological sciences.

Key words: *Bioinformatics, Biomedical databases, PubMed, Sequence data, Genomics research*

Introduction

The NCBI, established in 1988 under the National Institutes of Health (NIH), was created to develop advanced information systems for molecular biology. In addition to maintaining the GenBank nucleic acid sequence database—which receives data through international collaborations with DDBJ, EMBL-Bank, and direct submissions from the global research community—NCBI also provides data retrieval systems and computational tools to facilitate

the analysis of GenBank entries and other integrated biological datasets (Sayers *et al.*, 2020). The Figure 1 categorizes the major NCBI Tools and Databases into functional domains.

A brief introduction to various tools and databases is provided in the next section. The most important tools and databases are as follows:

NCBI Tools

Entrez: Entrez(<https://www.ncbi.nlm.nih.gov/Entrez/>) is an integrated system which provides

access to 39 databases with 1.7 billion records via the GQuery Portal. It supports Boolean text searches, data downloads in various formats, and links related records across databases (NCBI, 2024a).

Basic Local Alignment Tool (BLAST): BLAST (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) is a powerful resource used to compare sequence data and retrieve similar sequences from various databases having applications in molecular genetics, molecular biology, bioinformatics, evolutionary biology, medical research and protein biochemistry. BLAST offers flexible search algorithm, accurate statistical output, continuous software enhancements, and high processing speed.

NCBI Databases

Books: Bookshelf (<http://www.ncbi.nlm.nih.gov/books/>) is maintained by NCBI and NLM, and provides access to a wide range of books, documents, and biomedical literature. To search resources, the Browse Tool (<http://www.ncbi.nlm.nih.gov/books/browse/>) is used. The order of different types of resources in Bookshelf by volume is: Reports (83.8%) > Books (10.3%) > Documentation (2.3%) > Databases = Collections (1.8%) (Hoepfner, 2013).

ClinVar: ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar/>) assembles information about genomic variations and their relationship to human health. It helps in variant interpretation by facilitating data sharing across 1,300 organizations worldwide. Data can be submitted via the Submission portal (<https://www.ncbi.nlm.nih.gov/clinvar/submitters/>) in two ways: (a) batch upload of multiple variants and (b) individual submission using Clin Var Submission Wizard (Landrum *et al.*, 2016).

PubChem BioAssay: The database (<http://pubchem.ncbi.nlm.nih.gov/>) includes bioactivity, examination, and researchable descriptions of

chemical substances and assays. Each assay has its own unique accession ID and is organized into both descriptions and results (Wang *et al.*, 2010). This Data can be downloaded through web service, programmatic tools or website. Various services offered by PubChem have been summarized in Table 1 (NCBI, 2024b).

PubChem Substance and Compound: Substance database generally includes the description of samples and stores depositor-contributed information, and unique chemical structures taken from the compound database (<https://www.ncbi.nlm.nih.gov/pccompound>). Primary identifiers are SID (substance ID) and CID (compound ID) (Kim *et al.*, 2016).

Nucleotide Database: The nucleotide database is a collection of genome, gene, and transcript sequence data obtained from many sources like GenBank, RefSeq, TPA (Third Party Annotation), and PDB (Protein Data Bank) (<https://www.ncbi.nlm.nih.gov/nucleotide/>). The Nucleotide database has 146035069 records attained from GenBank, national and international collaboration and internal NCBI or NLM curation (NCBI, 2018).

Online Mendelian Inheritance in Man (OMIM) Database: OMIM is a globally available database hosted on the World Wide Web at (<http://www.ncbi.nlm.nih.gov/omim/>). It is a comprehensive resource of bibliographic information focused on human genes and genetic disorders. OMIM provides full-text summaries outlining gene functions and combined phenotypes.

Protein Database: It (<https://www.ncbi.nlm.nih.gov/protein>) is an extensive repository of computational predictions and experimentally validated data, related to proteins - their sequences, structures, functions, interactions, and translations from annotated coding regions found in GenBank, RefSeq, and

Table 1. List of services offered by PubChem Bioassay

Tools and Resources Description		Url Examples
LinkOut	NCBI Linkout provide service	https://www.ncbi.nlm.nih.gov/projects/linkout/
Structure Ref Seq	3-D Structure reveal function and evolution Curated, non-redundant reference sequence database	https://www.ncbi.nlm.nih.gov/structure https://www.ncbi.nlm.nih.gov/refseq/
Batch Entrez	Retrieval of protein records.	https://www.ncbi.nlm.nih.gov/sites/batchentrez
E-utilities	Automated data access across NCBI	https://www.ncbi.nlm.nih.gov/books/NBK25501/

TPA. It helps in studying protein function, evolutionary relationships, and molecular interactions critical for applications in drug discovery, biotechnology, and molecular biology (Geer *et al.*, 2010a). Details of various tools and resources of protein database have been summarized in Table 2 (Geer *et al.*, 2010a).

Protein Cluster Database: It is a grouping of related protein sequences, or clusters, obtained from plasmid, organelle, and whole genome annotations. (<https://www.ncbi.nlm.nih.gov/proteinclusters/?term=>) It was designed to effectively manage large scale protein data. ProtClustDB organizes protein cluster by sequence similarity, with 7180 curated clusters covering 376,513 proteins and provides an efficient method to aggregate gene and protein annotation (Klimke *et al.*, 2009).

Protein family models: They classify proteins by shared domains, sequence motifs, and evolutionary relationships, often using computational approaches like Hidden Markov Models (HMMs) (Mistry *et al.*, 2021). These models are central to databases such as Pfam and InterPro, supporting large-scale protein annotation and comparative genomics. The application of profile HMMs has significantly enhanced the ability to identify distant homologs with greater sensitivity.

Pop Set: The database (<https://www.ncbi.nlm.nih.gov/popset>) is a collection of affiliated DNA sequences and alignments deduced from population, phylogenetic, mutation, and ecosystem studies that have been submitted to GenBank. PopSet alignments are available on the PopSet record page (NCBI, 2024d). From January 2025, the NCBI PopSet database is no longer available.

Identical Protein Groups: The database (<https://www.ncbi.nlm.nih.gov/ipg/>) consolidates identical protein sequences from multiple sources into single entries, each linked to their nucleotide origins. It al-

lows filtered searches by taxonomy, data source, and group size, enabling precise protein analysis. This makes it a key tool in both proteomics and comparative genomics (NCBI, 2024e). The priority order for selecting the best sequence is: RefSeq > Swiss-Prot > PIR, PDB > GenBank.

Med Gen: MedGen (<https://www.ncbi.nlm.nih.gov/medgen>) compiles comprehensive information on genetic disorders and their associated phenotypes. It includes details such as clinical features, inheritance patterns, and genomic loci. It serves as a valuable tool in biomedical and genetic research (Kitts *et al.*, 2016).

Assembly: The database (<http://www.ncbi.nlm.nih.gov/assembly/>) offers credible accessioning, tracking, and organization of genome sequences into chromosomes and can be categorized into four tiers: contig-level assemblies, scaffold-level assemblies, chromosome-level assemblies, and complete genome assemblies. The NCBI Assembly resource offers searchable access with advanced query options, an efficient organism auto-complete feature, and downloadable genome data via its FTP site (Amberger *et al.*, 2016).

Sequence Read Archive (SRA): It (<https://www.ncbi.nlm.nih.gov/sra>) is a comprehensive repository for high-throughput sequencing data maintained by the NCBI (Robinson, 2025). It acts as a centralized resource for storing and sharing next-generation sequencing (NGS) data generated from various platforms, including Illumina, Ion Torrent, and PacBio (Robinson, 2025). SRA Stores raw sequencing data is easily accessible and has refined hunt capabilities (Sauna and Kimchi-Sarfaty, 2022).

Single Nucleotide Polymorphism (SNP): SNP (<https://www.ncbi.nlm.nih.gov/snp>) is a specialized repository that focuses on cataloging genetic variations at single nucleotide positions in genomic DNA, critical for understanding genetic diversity,

Table 2. Tools and resources of protein database

Service	Description	URL Example
Bioassay search	Search using Entrez	https://www.ncbi.nlm.nih.gov/pcassay/
Bioassay Download	Download Interface	https://pubchem.ncbi.nlm.nih.gov/assay/assaydownload.cgi
Bioassay Classification	Browse the Bioassay Classification tree	https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?p=classification
Bioassay Service Home	Homepage	https://pubchem.ncbi.nlm.nih.gov/assay/
BioAssay FTP	FTP for all records	ftp://ftp.ncbi.nlm.nih.gov/pubchem/Bioassay/

disease susceptibility, and responses to drugs. Researchers can retrieve data from the SRA via online search (<https://www.ncbi.nlm.nih.gov/snp/advanced/>) through the NCBI website, SRA toolkit or Galaxy Platform Integration (NCBI, 2024f).

Structure: The Database (<https://www.ncbi.nlm.nih.gov/structure/?term=>) provides three-dimensional structure of macromolecules, which can be used to examine sequence structure-function relationships (NCBI, 2024g). To visualize these structures and chemical on the web, a WebGL-based viewer, iCn3D (https://www.ncbi.nlm.nih.gov/Structure/icn3d/docs/icn3d_about.html) is used whose source code is openly available on GitHub (<https://github.com/ncbi/icn3d>).

Taxonomy: The Database (<http://www.ncbi.nlm.nih.gov/taxonomy>) offers a standardized classification system for all species found in the public sequence database, currently representing 10% of described species. It serves as the central categorization depository for GenBank, European Nucleotide Archive and DDBJ – members of International Nucleotide Sequence Database Collaboration (INSDC) (Federhen, 2012). It primarily uses Carl Linnaeus classification system across four principal codes.

Conserved Domain Database (CDD): CDD helps annotate protein and nucleotide sequences using domain models from multiple sequence alignments. It identifies domain footprints, functional sites, and structural features to predict protein roles. CDD includes curated data from sources like Pfam, COGs, TIGRFAMs, SMART, and others, totalling over 18,000 models and 4,500 superfamily clusters. Though rich in content, its update frequency is limited by curation demands (Schoch *et al.*, 2020).

Gene Database: The database (www.ncbi.nlm.nih.gov/gene) integrates gene-specific data from RefSeq genomes and other sources, linking sequences to detailed functional and structural information. Each gene entry includes curated and automated data from RefSeq, Gene Ontology, and other databases, with unique identifiers. It offers insights into gene location, expression, variations, phenotypes, and protein products. As of 2014, it covered millions of genes across thousands of taxa, with bacteria being the most abundant. Access is available via Entrez tools, E-Utilities, and FTP (Brown *et al.*, 2015).

Database of Genotypes and Phenotypes (dbGaP):

(<https://www.ncbi.nlm.nih.gov/gap/>) is a comprehensive repository designed to store and share genotype and phenotype data from a variety of studies. dbGaP facilitates access to data linked to specific diseases, traits, and health-related research, enabling researchers to investigate the relationships between genetic variations and phenotypic outcomes. The database prioritizes data privacy and security while providing researchers with valuable insights to promote advancements in personalized medicine and genomic research (Wong *et al.*, 2017). By connecting genotype data with phenotypic information, dbGaP plays a crucial role in enhancing our understanding of genetic influences on health and disease.

Genome: Genome (<https://www.ncbi.nlm.nih.gov/datasets/genome/>) is an important repository that gathers genomic data and annotations of different species and offers a comprehensive interpretation. It offers tools for querying and analysing genomic data, enabling research into genetic variation, disease links, and evolution (Geer *et al.*, 2010b).

BioSample Database: Bio Sample (<http://www.ncbi.nlm.nih.gov/biosample>), contains descriptions of the biological materials being examined in various projects. It includes a cell line, a primary tissue biopsy, an individual organism, or an environmental isolate. Despite the variety of samples, the BioSample database presents new opportunities for improving the collection and standardization of sample descriptions throughout NCBI databases.

BioProject Database: BioProject (<http://www.ncbi.nlm.nih.gov/bioproject>) serves as a comprehensive organization for biological research initiatives and the resulting data is archived in various databases managed by the INSDC. It systematically arranges metadata for research projects expected to generate substantial amounts of data and offers a centralized gateway to access the information once it has been stored in an archival database.

Geo Datasets and Geo Profile Databases: High-throughput technologies, including microarrays, generated very large molecular data sets, the majority of which are stored in the Gene Expression Omnibus (GEO) (<http://www.ncbi.nlm.nih.gov/geo/info/download.html>). GEO is a comprehensive public database that supports sharing, reassessing, and exercise of data. Further, GEO is MIAME compliant to guarantee data integrity and reproducibility. Experimenters can use GEO's internal and external

links, guaranteeing improved data analysis and cooperative research (Edgar *et al.*, 2002).

PubMed Database: PubMed (<https://pubmed.ncbi.nlm.nih.gov/>) is a comprehensive biomedical literature platform maintained by the U.S. National Library of Medicine, widely used for systematic reviews, clinical studies, and academic research. With features like MeSH term indexing and Best Match ranking, it enhances information retrieval (Fiorini *et al.*, 2018). Platforms such as PubMed Central offer free access to complete research articles

MeSH: The database (<https://www.ncbi.nlm.nih.gov/mesh/>), maintained by NCBI, is a thesaurus used for indexing biomedical literature in PubMed. It allows coordinated searches and classification of topics across life sciences which help in retrieval of advanced biomedical information. It also standardizes terminology for more efficient and precise searches.

NLM catalog database: The NLM Catalog (<https://www.ncbi.nlm.nih.gov/nlmcatalog/>) database aids the users to search by title, subject or identifier, which helps in locating authoritative bibliographic records for journals, books and other resources held at the National Library of Medicine. Thus, these records aid in finding trusted publications in biomedicine and health.

Dbvar: Dbvar (<https://www.ncbi.nlm.nih.gov/dbvar/?term=>) contains over 3 million submitted structural variants (SSVs) from 120 human studies including complicated chromosomal rearrangements, insertions, deletions, inversions, translocations, and copy number variations (CNV.) Large structural genomic variants are included in the public database dbvar, which houses millions of entries from different investigations. Efficient data retrieval is facilitated by its browser tools and online search feature (Lappalainen *et al.*, 2013).

Genetic Testing Registry (GTR): GTR (<https://www.ncbi.nlm.nih.gov/gtr/>) is a cross-domain database system with an automatic SQL generator that works in three steps: identifying the relationship between queries and database schemas; generating a translation representation; and predicting SQL using grammar based rules⁵. Submission API is used to help laboratories submit, update, or remove tests efficiently. Fulgent Genetics, with over 19,000 tests (~26% of GTR), is the largest contributor (Sayers *et al.*, 2023).

Biocollection Database: It was designed to curate metadata for natural history collections and link sequence data to voucher specimens. Collections staff manage the hierarchical relationships of the objects within their collections and ensure they are assigned proper Uniform Resource Identifiers (URIs). The

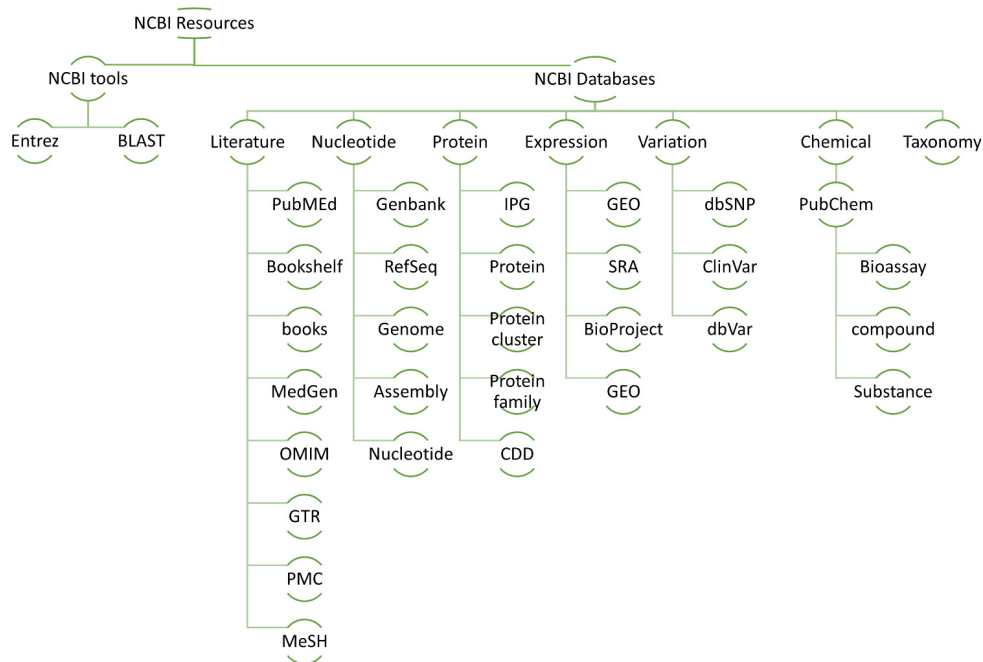


Fig. 1. Categorizing various NCBI Resource

database broadly follows the Darwin Core (DwC) to standardize the usage of data across interconnected databases.

Materials and Methods

This paper explores the resources available to NCBI and assesses their value to a biomedical researcher. The methodology for this study is mixed in nature, having both qualitative and quantitative assessments in order to gain insight into NCBI's resources.

1. **Resource Identification:** This involves systematically identifying and classifying NCBI's diverse resources, including databases (e.g., GenBank, PubMed), tools (e.g., BLAST, Genome Viewer), software, and educational materials. This is done through a thorough review of the NCBI website and relevant literature.
2. **Resource Exploration:** A focused exploration of NCBI databases was conducted using keywords like "PubMed," "BLAST," and "GenBank" across platforms such as PubMed, Google Scholar, and Research Gate. Efforts were made to ensure database-specific relevance and up-to-date information.
3. **Extraction of information** – After getting enough data systematic extraction of data starts using a standardized data extraction form. Entrez system, API and other latest updates were recorded. Main emphasis is on integration of databases with one another and their application in Genomics, Transcriptomics and Proteomic research.
4. Data analyzing and Systematic formatting of Data takes place, here we analyze our data and format our data into pictographic and tabular form.

Results and Discussion

Our study on NCBI tools and databases identified and analysed the composition, functionality, and consolidation of over 35 NCBI databases, covering genomics, proteomics, structural biology, clinical data, and scientific literature.

NCBI arrays an extensive variety of databases that are classified into six main categories:

- A. **Literature Databases** – It includes PubMed, Bookshelf, MedGen, OMIM, Genetic Testing Registry, and PubMed Central. These databases help in providing access to more than 36 million

biomedical articles, abstracts, and citations.

- B. **Nucleotide and Genome Databases** – It includes GenBank, Reference Sequence, Genome, and Assembly. They primarily focus on storing raw nucleotide sequences and whole genome assemblies of organisms.
- C. **Protein and Molecular Databases** – It includes Protein and Conserved Domain Database. They provide protein sequences, domains, and structural annotations.
- D. **Gene Expression, Functional Genomics, and Epigenetics** -It includes Gene Expression Omnibus and Sequence Read Archive. They store and analyse gene expression and sequencing data from experiments.
- E. **Genetic Variation and Phenotypic Data** -It includes dbSNP, ClinVar, and dbVar. The objective is to record information on genetic mutations, variants, and their phenotypic/clinical effects.
- F. **Chemical, Taxonomic, and Computational Tools** -This includes Taxonomy and PubChem. This includes compound databases.

NCBI databases collectively sustain more than 4.3 billion records across diverse branches. GenBank database itself comprises over 97 million base pairs with 93 million entries. DbVar is a structural variant database and contains over 3 million structural variants from 120 human studies which can be retrieved using web-based and command-line tools.

Apart from databases, NCBI comprises more than 80 specialized tools. For example – Sequence alignment tools include BLAST, BLAST+, and Primer-BLAST. Data retrieval and search tools include Entrez, E-utilities, and Genome Data Viewer. But our study mainly focuses on Entrez and BLAST. Entrez links over 40 databases and possesses great cross-linking capacity. It also allows users to save searches and customize filters through their MyNCBI account. BLAST is mainly used for gene identification, evolutionary studies, and mutation analysis. File formats which are supported include FASTA, XML, CSV, ASN.1, and JSON.

The volume of data which NCBI databases hold is huge and is doubling with coming years due to rapid sequencing advances. However, this increase in data creates certain challenges in annotations, maintaining data quality, and versioning of data. Which needs to be taken into account?

Emerging technologies such as machine learning, artificial intelligence, deep learning, robotics, single-

cell omics, advancements in biotechnology, and molecular biology will produce more complex data which necessitates the need for advanced computational infrastructure.

Conclusion

The NCBI databases are considered as important resource that assists a broad spectrum of Biomedical researches and investigations. Each database provides distinct features that contribute to the progression of scientific knowledge. Ongoing enhancement and integration of these tools will bolster research capabilities, stimulate innovation, and ultimately deepen our comprehension of intricate biological systems. Nonetheless, significant deficiencies persist. For example, while platforms such as GenBank and dbSNP provide ample genetic data, they frequently lack connections with clinical information, which impedes translational research.

The research regarding biomolecules, and clinical genetics globally it leads to the generation of data in large quantities and it's maintained by NCBI.

Furthermore, the availability of certain datasets could be improved, particularly for researchers operating in low-resource environments. In addition, as new technologies and methods arise, the databases must adapt to handle broader and additional complicated datasets produced by next-generation sequencing. There is also a demand for better user interfaces and analytical tools that support effective data mining and interpretation.

Acknowledgement

We sincerely thank the Principal of Hindu College, Prof. Anju Srivastava, for providing access to institutional library and digital resources, which were instrumental in compiling the information presented in this review. We also gratefully, acknowledge the support of Prof. Reena Jain, Coordinator, DBT Star College Scheme, for her valuable guidance throughout the preparation of this review.

Conflict of interest: The authors declare that they have no conflict of interest

Author Contributions:

Dr. Vivek Chopra, Dr. Rajesh Kumar and Dr. Meenam Bhatia helps in the conceptualization, guidance, supervision and critical revision of the

paper. He conceived and supervised the study, contributed to the design and interpretation of the paper.

All other students contributed equally, they assisted in data collection, performed database searches, contributed to data organization, helped in proofreading and drafted the paper and manuscript.

References

- Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F. and Hamosh, A. 2016. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* 44(D1): D789-798.
- Brown, G.R., Hem, V., Katz, K.S., Ovetsky, M., Wallin, C., Ermolaeva, O. 2015. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.* 43(D1): D36-42.
- Federhen, S. 2012. The NCBI Taxonomy database. *Nucleic Acids Res.* 40(D1): D136-143.
- Edgar, R., Domrachev, M. and Lash, A.E. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30(1): 207-210.
- Fiorini, N., Canese, K., Starchenko, G., Kireev, E., Kim, W., Miller, V. 2018. Best Match: new relevance search for PubMed. *PLoS Biol.* 16(8):e2005343.
- Geer L.Y., Marchler-Bauer, A., Geer, R.C., Han, L., He, J. and He, S. 2010. The NCBI Bio Systems database. *Nucleic Acids Res.* 38(Database issue): D492-496.
- Hoepfner, M.A. 2013. NCBI Bookshelf: books and documents in life sciences and health care. *Nucleic Acids Res.* 41(Database issue): D1251-12260.
- Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G. and Gindulyte, A. 2016. PubChem Substance and Compound databases. *Nucleic Acids Res.* 44(D1): D1202-213.
- Klimke, W., Agarwala, R., Badretdin, A., Chetvernin, S., Ciuffo, S. and Fedorov, B. 2009. The NCBI Protein Clusters Database. *Nucleic Acids Res.* 37(Database issue): D216-223.
- Kitts, P.A., Church, D.M., Thibaud-Nissen, F., Choi, J., Hem, V. and Sapojnikov, V. 2016. Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res.* 44(D1): D73-80.
- Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S. 2016. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* 44(D1): D862-868.
- Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J.D. and Garner, J. 2013. DbVar and DGVA: public archives for genomic structural variation. *Nucleic Acids Res.* 41(Database issue): D936-941.

- Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L. 2021. Pfam: the protein families database in 2021. *Nucleic Acids Res.* 49(D1): D412-419.
- National Center for Biotechnology Information (NCBI). Entrez Search Interface [Internet]. Bethesda (MD): NCBI; [cited 2024 Sep 3]. Available from: <https://www.ncbi.nlm.nih.gov/Web/Search/entrezfs.html>
- National Center for Biotechnology Information (NCBI). PubChem BioAssay Database [Internet]. Bethesda (MD): National Library of Medicine (US); [cited 2024 Oct 23]. Available from: <https://www.ncbi.nlm.nih.gov/pubchem/bioassay/>
- NCBI Resource Coordinators, 2018. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 46(D1): D8-13.
- National Center for Biotechnology Information (NCBI). Protein [Internet]. Bethesda (MD): National Library of Medicine (US); [cited 2024 Nov 15]. Available from: <https://www.ncbi.nlm.nih.gov/protein/>
- National Center for Biotechnology Information (NCBI). About IPG [Internet]. Bethesda (MD): National Library of Medicine (US); [cited 2024 Oct 29]. Available from: <https://www.ncbi.nlm.nih.gov/ipg/docs/about/>
- National Center for Biotechnology Information (NCBI). MedGen database: LMNB1 gene [Internet]. Bethesda (MD): National Library of Medicine (US); [cited 2024 Oct 9]. Available from: [https://www.ncbi.nlm.nih.gov/medgen?term=lmnb1\[Gene\]](https://www.ncbi.nlm.nih.gov/medgen?term=lmnb1[Gene])
- National Center for Biotechnology Information (NCBI). Article from PubMed Central [Internet]. Bethesda (MD): National Library of Medicine (US); [cited 2024 Oct 16]. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC3078622/>
- National Center for Biotechnology Information (NCBI). MMDB Source Page [Internet]. Bethesda (MD): National Library of Medicine (US); [cited 2024 Oct 11]. Available from: <https://pubchem.ncbi.nlm.nih.gov/source/MMDB>
- Schoch, C.L., Ciufu, S., Domrachev, M., Hotton, C.L., Kannan, S. and Khovanskaya, R. 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)*. 2020: baaa062.
- Sayers, E.W., Beck, J., Brister, J.R., Bolton, E.E., Canese, K. and Comeau, D.C. 2020. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 48(D1): D9-16.
- Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J. and Comeau, D.C. 2023. Database resources of the National Center for Biotechnology Information in 2023. *Nucleic Acids Res.* 51(D1): D29-38.
- Wang, Y., Bolton, E., Dracheva, S., Karapetyan, K., Shoemaker, B.A. and Suzek, T.O. 2010. An overview of the PubChem BioAssay resource. *Nucleic Acids Res.* 38(Database issue): D255-266.
- Wong, K.M., Langlais, K., Tobias GS, Fletcher-Hoppe C, Krasnewich, D. and Leeds, H.S. 2017. The dbGaP data browser: a new tool for browsing dbGaP controlled-access genomic data. *Nucleic Acids Res.* 45(D1): D819-826.
-