PLOS Computational Biology

EDUCATION

Ten quick tips for using the NIH Comparative Genomics Resource

Eric S. Tvedte[1], Cecilia Arighi[2], Matthew B. Carson[3], Luke V. Rasmussen[4], Kristi Holmes[3,4], Terence D. Murphy[1]*

1 National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America, **2** Department of Computer and Information Sciences, University of Delaware, Newark, Delaware, United States of America, **3** Galter Health Sciences Library & Learning Center, Northwestern University Feinberg School of Medicine, Chicago, Illinois, United States of America, **4** Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois, United States of America

* murphyte@ncbi.nlm.nih.gov

OPEN ACCESS

Citation: Tvedte ES, Arighi C, Carson MB, Rasmussen LV, Holmes K, Murphy TD (2026) Ten quick tips for using the NIH Comparative Genomics Resource. PLoS Comput Biol 22(2): e1013919. https://doi.org/10.1371/journal.pcbi.1013919

Editor: Patricia M Palagi, SIB: Swiss Institute of Bioinformatics, SWITZERLAND

Published: February 11, 2026

Funding: This work was supported in part by the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health (NIH). The contributions of the NIH author(s) are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views of the NIH

Introduction

The growth of publicly available eukaryotic genomic data has revolutionized both life sciences and biomedical research. By sequencing entire genomes, scientists have accelerated their ability to investigate complex biological systems with greater depth and precision. Genomic data plays a pivotal role in discovering disease-causing mutations, identifying biomarkers for early diagnosis, and uncovering new targets for drug development. Moreover, integrating genomic data with other -omics technologies, such as transcriptomics and proteomics, offers a comprehensive view of cellular functions, paving the way for personalized medicine.

The NIH Comparative Genomics Resource (CGR; https://www.ncbi.nlm.nih.gov/cgr/) [1] led by the National Center for Biotechnology Information (NCBI) is a groundbreaking initiative aimed at enabling researchers to explore and compare genomic data across diverse eukaryotic species. This resource includes genomic data, tools, and community resources to support comparative genomics studies, helping scientists better understand evolutionary relationships and accelerating biomedical research discovery. Its user-friendly platform ensures that data from a wide range of organisms is accessible to researchers worldwide in both human- and machine-readable formats.

The recent "Ten Simple Rules for using public biological data in your research" [2] provides an excellent discussion of the basic principles of using public data. Here, we discuss Ten Quick Tips for utilizing CGR data and tools in genomics research as a guide for both new and experienced users of NCBI resources. CGR resources can be beneficial throughout a project's lifecycle starting with user-generated data (Fig 1). Prioritization of the steps (tips) is up to the researcher.

Tip 1: Clean up the crud

When assembling new genomes, post-assembly curation is typically needed to fix errors. Genome assemblies often contain contaminants (a.k.a. 'crud') from non-target organisms and/or sequencing adaptors. You can identify and remove contaminants

or the U.S. Department of Health and Human Services. This work was supported in part by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), Grant Number UM1TR005121. MBC, LVR, and KH received salary support from NCATS grant UM1TR005121. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

using the NCBI Foreign Contamination Screen (FCS; https://github.com/ncbi/fcs) [3]. The crud itself may be a valuable source of genome data for cobiont organisms that can be assembled to understand potential host-cobiont relationships [4]. And don't forget to document your curation steps and quality metrics to help future researchers assess the quality of your genome assembly [5–7].

Tip 2: Enhance genome quality with annotation

To identify functional elements in the genome sequence, annotation pipelines use ab initio or evidence-based methods or an integration of the two, accounting for repeat content [8]. The NCBI Eukaryotic Genome Annotation Pipeline (EGAP) (https://www.ncbi.nlm.nih.gov/refseq/annotation_euk/process/) has been used for 25 years to generate genome annotations for RefSeq genomes. To support community annotation efforts, you can now use EGAPx (https://github.com/ncbi/egapx), a standalone adaptation of EGAP for annotating genomes that can be readily submitted to GenBank. Whether you use EGAPx or another annotation program, you should submit your genome and annotation to an archival database that is part of the International Nucleotide Sequence Database Consortium (INSDC) [9]; submission promotes data sharing and reuse, and helps ensure your efforts make an impact for years to come.

Tip 3: Download large-scale data and metadata

NCBI GenBank is growing continuously, containing 34 trillion base pairs from over 4.7 billion nucleotide sequences for 581,000 formally described species as of 2025 [10]. Large datasets can increase the confidence in observed patterns. More genomes at different evolutionary distances can be valuable in identifying and confirming the evolutionary context of an observation. NCBI Datasets [11] provides scalable access to gene, genome, and taxonomy information through multiple interfaces (web, command-line, API). A common use case in comparative genomics is to download multiple genomes from a taxonomic group of interest, including one or more reference genomes (see Tip 7) that NCBI has identified as the "best" for a species (https://www.ncbi.nlm.nih.gov/datasets/docs/v2/glossary/). You can use data retrieved from NCBI Datasets in a variety of contexts, such as building catalogs of sequences of a defined area of interest [12], investigating gene variation to understand evolutionary patterns [13], or as a case study for a new bioinformatic tool [14]. All data retrieved through NCBI Datasets includes a comprehensive metadata report. These reports follow documented schemas that provide standardized, structured descriptions of biological data stored at NCBI. Consider the biological questions of interest when choosing the taxonomic scope of the dataset: restricting to closely related taxa enables detection of recent evolutionary events due to high sequence similarity, while including more distantly related taxa allows investigation of deeper evolutionary timescales and can highlight conserved sequences over long periods of time.

Tip 4: Inspect public data for quality issues

Although the growth in public genomic data is mostly beneficial, there are challenges in working with data with inconsistent features and/or quality. In extreme situations,

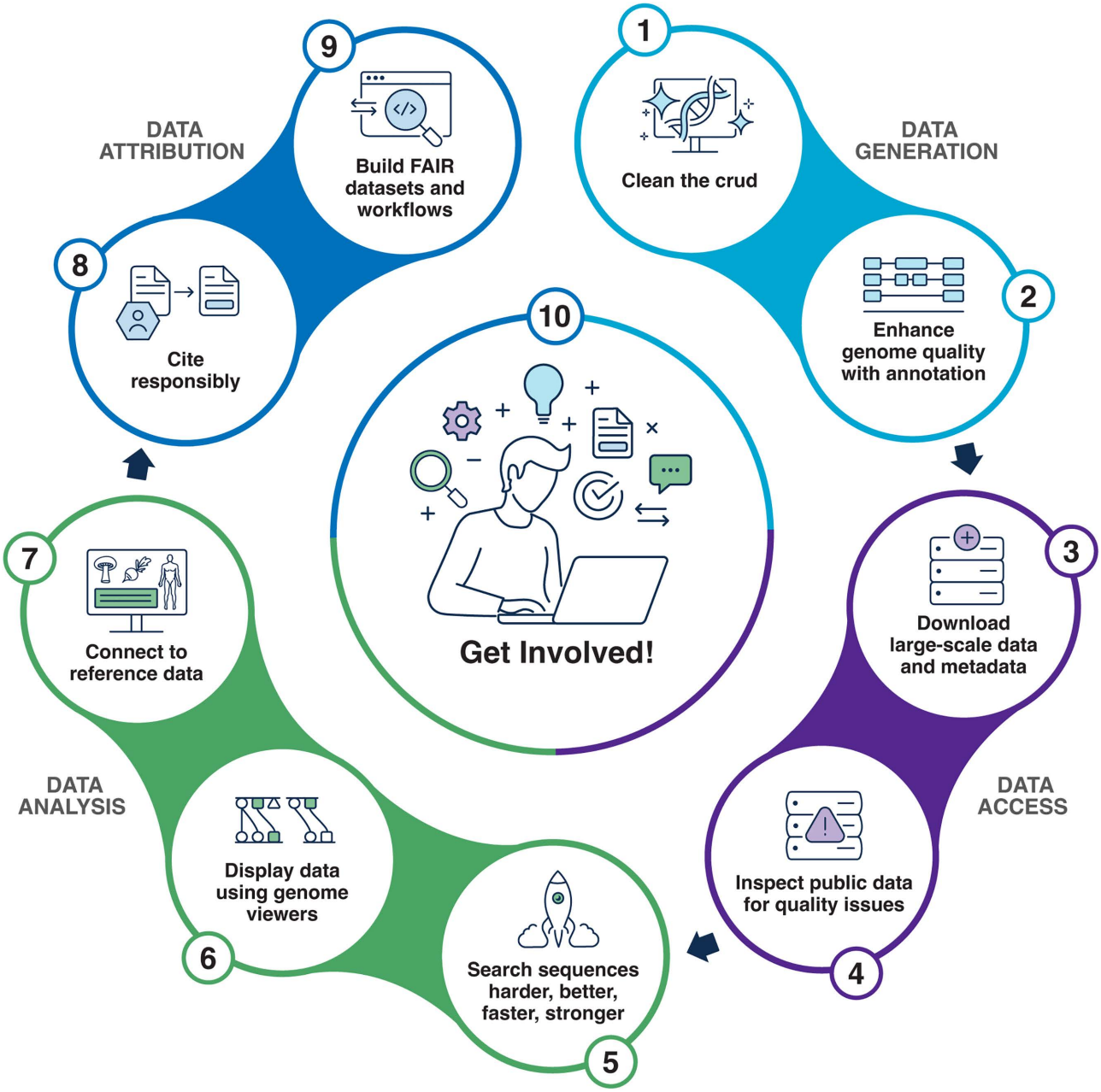

Fig 1. The Ten Quick Tips for conducting genomics research using the NIH Comparative Genomics Resource can utilize NCBI data and tools throughout the lifecycle of a research project.

https://doi.org/10.1371/journal.pcbi.1013919.g001

using incorrect data can actively hinder your biological interpretations of results, so don't assume all publicly-available data is correct. You can use public NCBI FCS reports updated daily (https://ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY_REPORTS/) to select genomes of interest based on your acceptable contamination level threshold or remove sequences assigned as contaminants. The scalable data access of NCBI Datasets can potentially expose systematic errors such as gene family mis-annotations [15]. Filters in NCBI Datasets queries can improve your data quality consistency (e.g., retain chromosome-level genome assemblies, retain annotated genomes, remove atypical/contaminated genomes).

Following data download, you should perform additional quality checks, such as examining the metadata to ensure the data is appropriate for your use case as well as verifying the integrity of downloads via checksums or other means (see Rules 2 and 7 from [2] for additional details). Commonly used quality metrics such as contig N50 and BUSCO [16] scores can also be used to remove low-quality outliers.

Tip 5: Search sequences harder, better, faster, stronger

The default BLAST [17,18] parameters are tailored for a balance between sensitivity and speed, but you might consider adjusting parameters for your specific use case. Examples include the choice of BLAST algorithm (e.g., megablast, discontinuous megablast, and blastn in nucleotide BLAST searches), word size, and filtering on low-complexity regions. It is important to understand the chosen parameters in order to effectively interpret the results; parameters such as max_target_seqs and evalue are commonly misinterpreted [19–21].

New BLAST databases have been developed as part of the CGR project. Understanding how BLAST database size and content can influence search outcomes is crucial for tailoring performance [22]. While the use of larger databases may produce more comprehensive results, sequence redundancy can increase the database size significantly slowing run times without any meaningful change in the results. The BLAST nucleotide database (nt) is currently 2.5 trillion letters and has a doubling rate of less than a year. Conversely, default searches now use the core nucleotide database (core_nt) that is less than half the size of the standard nt database, enabling faster searches with minimal impact on top-hit quality [23]. Default searches of protein databases use the new ClusteredNR database [24], which reduces the redundancy of protein sequences in the standard protein database (nr) into clusters with 90% identity/ 90% length thresholds. Searches using ClusteredNR are faster and provide results across a wider range of organisms and evolutionary distances. Additional options for customizing database content include by source (e.g., RefSeq, WGS, Swiss-Prot), type (rRNA/ITS databases) and organism(s) using tax-ids. Your research goals should help define your search strategies, and by experimenting with BLAST you will be able to design searches to achieve the appropriate balance of match distance, sensitivity, and speed.

Web BLAST searches use a fixed amount of CPU time. If your searches consistently exceed the CPU limit, you can download BLAST databases from NCBI and use BLAST+ executables (https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/) or try ElasticBLAST [25] on the cloud.

Tip 6: Display data using genome viewers

Genome data visualizations can uncover and communicate complex patterns effectively. The NCBI Genome Data Viewer (GDV) [26] is a linear genome browser that can display gene annotation, experimental data (e.g., RNA-seq), and alignments to other genomes to understand how genome sequence correlates to function. Try viewing pairwise whole genome sequence alignments using the NCBI Comparative Genome Viewer (CGV) [27]. CGV views highlight large-scale structural rearrangements such as duplications, deletions, and inversions, allowing you to assess gene synteny and analyze changes that may have contributed to differences in biology or phenotype. When you're ready to publish, GDV and CGV both include high-quality SVG download options to help visualize your story.

If you have a gene or genomic region of interest and want to compare across many species, the new NCBI Multiple Comparative Genome Viewer (MCGV; https://www.ncbi.nlm.nih.gov/mcgv/) can display whole-genome multiple alignments created by the research community. MCGV provides alignment and conservation summaries to help you find locations that are well conserved or divergent.

Tip 7: Connect to reference data

A focal point of the CGR project is the support for high-quality reference sequences for comparative genomics analyses. The NCBI RefSeq resource (https://www.ncbi.nlm.nih.gov/refseq/) is a stable, non-redundant collection of reference sequences for genomes, transcriptomes, and proteins [28–30]. The RefSeq collection is generated from a combination of

automatic processes, manual curation, and community collaboration. As of November 2025 there are over 2200 eukaryote species with whole genomes represented in RefSeq. By using well-established references, you can perform analyses that are understandable and reproducible by the scientific community. Transcriptomics, proteomics, sequence variation, and multiple species alignments can all be informed by annotated references, enabling new biomedical discoveries. You can also use RefSeq sequences as connection points to many other NCBI resources. Try using RefSeq accessions to find gene-centric information on the Gene database [31], search against databases of RefSeq sequences using BLAST, collect orthologs using NCBI Datasets, and visualize annotated features using GDV and CGV.

Tip 8: Cite responsibly

By giving credit to researchers whose data you use in comparative genomics analysis, you promote the sharing of data in open databases [2,32]. Check how the original authors who produced the data and/or code want to be credited [32]. If there is no specific guidance, you should reference a persistent data object identifier (DOI) corresponding to the stored data and/or the original publication where the data was produced. NCBI Genome Datasets pages include links to BioProject metadata as well as links to relevant publications that may contain relevant citations. There is also information on how to cite NCBI services and databases [33,34]. Publications that share and access data in public databases tend to be more highly cited than those that do not [2,35,36], so taking the time to give appropriate credit benefits you as well as the original producer of the data.

Tip 9: Build FAIR datasets and workflows

The FAIR (Findable, Accessible, Interoperable, Reusable) Principles are widely accepted best practices for sharing research output [37,38]. FAIR data enhances discovery, access, and integration with other data by providing an identifiable location, proper context, and guidance for reuse to other researchers. To support FAIR principles using CGR tools and data:

- **Findability:** Use persistent identifiers with rich metadata to improve the findability for newly generated data (Tips 1 and 2) and public data (Tip 3). NCBI assigns permanent identifiers (accession numbers) for all deposited genomic data including BioProjects, assemblies, and annotated proteins which can be used as search strings on NCBI databases. When your workflows use CGR tools, you should provide relevant software versions to enhance the findability of documentation for the original software.
- **Accessibility:** Submitting data and metadata to NCBI enables open and free access using standardized procedures (web, command-line, API). If data types are not supported by NCBI, your data should be downloadable from a general-purpose repository and include documentation for acquiring and incorporating your data into research workflows. Research software that uses CGR tools and data should also be open source.
- **Interoperability:** By preparing data in standard formats (e.g., FASTA/FASTQ, GFF, MAF) alongside rich metadata in JSONL and TSV, you can access and compare datasets from a variety of sources. By reading, writing, and exchanging data using community standard formats, research software can interact with other research software, for example through NCBI Datasets APIs.
- **Reusability:** Enriching your data and software with metadata including relevant attributes appropriate for your data types, licensing information, and references to other data and/or software will support reusability. Reusability can be enhanced using containers (e.g., Docker, Singularity). For example, the NCBI FCS (Tip 1) has been incorporated into multiple genome assembly evaluation workflows [39–41].

It is important to note that reusability does not necessarily guarantee reproducibility. For example, CGR tools can be FAIR compliant but data access (e.g., NCBI Datasets) and data analysis (e.g., BLAST) may return different results over time following updates to data resources. To support reusability and transparency, researchers should report:

- Dataset author, publication year, name, and version used (or date accessed if no version)
- Software author, name, and version used (and citation if available)
- The query itself (including query string if applicable, parameters, or options), execution date of the query, and the context (what was the research question that was answered?)
- A permanent identifier or direct access link to the results if available

Tip 10: Get involved!

Providing feedback is a worthwhile effort; CGR continues to benefit from community input to enhance its tools. Researchers can submit feedback, participate in surveys, or join discussions through forums, workshops, and user meetings. The easiest way to provide feedback is through the yellow "Feedback" button in the lower right-hand corner of CGR and other NCBI webpages.

A wide range of curricula, tutorials, and workshops are freely available (https://www.ncbi.nlm.nih.gov/cgr/) to support use of CGR resources, as are links to published papers which have applied these resources. Individuals can subscribe to the NCBI Insights newsletter (https://ncbiinsights.ncbi.nlm.nih.gov) for CGR-specific content and receive updates on new CGR tools, datasets, and research findings. Email CGR at cgr@nlm.nih.gov to teach a workshop, partner on a webinar, or discuss other ideas you may have to foster information sharing and feedback. By taking advantage of these opportunities, researchers can both contribute to and stay informed about the latest developments in comparative genomics.

Conclusions

The NIH CGR plays a pivotal role in translating comparative genomics into actionable insights for basic science and applied research. By uncovering shared genetic pathways and evolutionary patterns, CGR supports the identification of genes and genetic variations associated with specific traits or diseases. Such insights have profound implications for understanding human health, agriculture, and environmental biology. The integration of community data with NCBI tools enhances the utility of the resource, empowering researchers to generate robust hypotheses and validate findings. As a centralized hub for comparative genomics, CGR fosters collaboration and accelerates discoveries, making it an asset in the pursuit of scientific innovation and the development of solutions to complex biological and medical challenges.

Acknowledgments

We thank Alejandro Sánchez Alvarado, Hannah Carey, Ani Manichaikul, Len Pennacchio, Ken Stuart, Tandy Warnow, and Cathy Wu for discussions contributing to the set of Ten Quick Tips. We thank Nuala O'Leary and Sanjida Rangwala for comments on written content. We thank Leslie Harris for preparing the graphical abstract (Fig 1).

References

1. Bornstein K, Gryan G, Chang ES, Marchler-Bauer A, Schneider VA. The NIH Comparative Genomics Resource: addressing the promises and challenges of comparative genomics on human health. BMC Genomics. 2023;24(1):575. https://doi.org/10.1186/s12864-023-09643-4 PMID: 37759191
2. Oza VH, Whitlock JH, Wilk EJ, Uno-Antonison A, Wilk B, Gajapathy M, et al. Ten simple rules for using public biological data for your research. PLoS Comput Biol. 2023;19(1):e1010749. https://doi.org/10.1371/journal.pcbi.1010749 PMID: 36602970
3. Astashyn A, Tvedte ES, Sweeney D, Sapojnikov V, Bouk N, Joukov V, et al. Rapid and sensitive detection of genome contamination at scale with FCS-GX. Genome Biol. 2024;25(1):60. https://doi.org/10.1186/s13059-024-03198-7 PMID: 38409096
4. Vancaester E, Blaxter ML. MarkerScan: separation and assembly of cobionts sequenced alongside target species in biodiversity genomics projects. Wellcome Open Res. 2024;9:33. https://doi.org/10.12688/wellcomeopenres.20730.1 PMID: 38617467
5. Howe K, Chow W, Collins J, Pelan S, Pointon D-L, Sims Y, et al. Significantly improving the quality of genome assemblies through curation. Gigascience. 2021;10(1):giaa153. https://doi.org/10.1093/gigascience/giaa153 PMID: 33420778

6. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021;592(7856):737–46. https://doi.org/10.1038/s41586-021-03451-0 PMID: 33911273
7. Wang P, Wang F. A proposed metric set for evaluation of genome assembly quality. Trends Genet. 2023;39(3):175–86. https://doi.org/10.1016/j.tig.2022.10.005 PMID: 36402623
8. Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, et al. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. Genome Res. 2024;34(5):769–77. https://doi.org/10.1101/gr.278090.123 PMID: 38866550
9. Karsch-Mizrachi I, Arita M, Burdett T, Cochrane G, Nakamura Y, Pruitt KD, et al. The international nucleotide sequence database collaboration (INSDC): enhancing global participation. Nucleic Acids Res. 2025;53(D1):D62–6. https://doi.org/10.1093/nar/gkae1058 PMID: 39535044
10. Sayers EW, Cavanaugh M, Frisse L, Pruitt KD, Schneider VA, Underwood BA, et al. GenBank 2025 update. Nucleic Acids Res. 2025;53(D1):D56–61. https://doi.org/10.1093/nar/gkae1114 PMID: 39558184
11. O'Leary NA, Cox E, Holmes JB, Anderson WR, Falk R, Hem V, et al. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. Sci Data. 2024;11(1):732. https://doi.org/10.1038/s41597-024-03571-y PMID: 38969627
12. van Valkengoed D, Bryon A, Ros VID, Kupczok A. Insights into diversity, host range, and evolution of iflaviruses in Lepidoptera through transcriptome mining. Virus Evol. 2025;11(1):veaf051. https://doi.org/10.1093/ve/veaf051 PMID: 40755814
13. Debbagh C, Folch G, Jabado-Michaloud J, Giudicelli V, Kossida S. Deciphering *Gorilla gorilla* gorilla immunoglobulin loci in multiple genome assemblies and enrichment of IMGT resources. Front Immunol. 2024;15:1475003. https://doi.org/10.3389/fimmu.2024.1475003 PMID: 39450182
14. Du L, Chen J, Sun D, Zhao K, Zeng Q, Yang N. Krait2: a versatile software for microsatellite investigation, visualization and marker development. BMC Genomics. 2025;26(1):72. https://doi.org/10.1186/s12864-025-11252-2 PMID: 39863857
15. Ticó M, Sullivan E, Guigó R, Mariotti M. Overcoming the widespread flaws in the annotation of vertebrate selenoprotein genes in public databases. bioRxiv. 2024. https://doi.org/2024.10.30.620813
16. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2. https://doi.org/10.1093/bioinformatics/btv351 PMID: 26059717
17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712
18. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402. https://doi.org/10.1093/nar/25.17.3389 PMID: 9254694
19. Shah N, Nute MG, Warnow T, Pop M. Misunderstood parameter of NCBI BLAST impacts the correctness of bioinformatics workflows. Bioinformatics. 2019;35(9):1613–4. https://doi.org/10.1093/bioinformatics/bty833 PMID: 30247621
20. González-Pech RA, Stephens TG, Chan CX. Commonly misunderstood parameters of NCBI BLAST and important considerations for users. Bioinformatics. 2019;35(15):2697–8. https://doi.org/10.1093/bioinformatics/bty1018 PMID: 30541060
21. Madden TL, Busby B, Ye J. Reply to the paper: misunderstood parameters of NCBI BLAST impacts the correctness of bioinformatics workflows. Bioinformatics. 2019;35(15):2699–700. https://doi.org/10.1093/bioinformatics/bty1026 PMID: 30590429
22. National Center for Biotechnology Information. BLAST Homepage and Selected Search Pages; 2019. Available from: https://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_BLASTGuide.pdf
23. Sayers EW, Beck J, Bolton EE, Brister JR, Chan J, Connor R, et al. Database resources of the National Center for Biotechnology Information in 2025. Nucleic Acids Res. 2025;53(D1):D20–9. https://doi.org/10.1093/nar/gkae979 PMID: 39526373
24. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the National Center for Biotechnology Information in 2023. Nucleic Acids Res. 2023;51(D1):D29–38. https://doi.org/10.1093/nar/gkac1032 PMID: 36370100
25. Camacho C, Boratyn GM, Joukov V, Vera Alvarez R, Madden TL. ElasticBLAST: accelerating sequence search via cloud computing. BMC Bioinformatics. 2023;24(1):117. https://doi.org/10.1186/s12859-023-05245-9 PMID: 36967390
26. Rangwala SH, Kuznetsov A, Ananiev V, Asztalos A, Borodin E, Evgeniev V, et al. Accessing NCBI data using the NCBI Sequence Viewer and Genome Data Viewer (GDV). Genome Res. 2021;31(1):159–69. https://doi.org/10.1101/gr.266932.120 PMID: 33239395
27. Rangwala SH, Rudnev DV, Ananiev VV, Oh D-H, Asztalos A, Benica B, et al. The NCBI Comparative Genome Viewer (CGV) is an interactive visualization tool for the analysis of whole-genome eukaryotic alignments. PLoS Biol. 2024;22(5):e3002405. https://doi.org/10.1371/journal.pbio.3002405 PMID: 38713717
28. Pruitt KD, Katz KS, Sicotte H, Maglott DR. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. Trends Genet. 2000;16(1):44–7. https://doi.org/10.1016/s0168-9525(99)01882-x PMID: 10637631
29. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733-45. https://doi.org/10.1093/nar/gkv1189 PMID: 26553804
30. Goldfarb T, Kodali Vamsi K, Pujar S, Brover V, Robbertse B, Farrell Catherine M, et al. NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. Nucleic Acids Research. 2024;53(D1):D243-D57.
31. Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, et al. Gene: a gene-centered information resource at NCBI. Nucleic Acids Res. 2015;43(Database issue):D36-42. https://doi.org/10.1093/nar/gku1055 PMID: 25355515

32. Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, Crosas M, et al. Ten simple rules for the care and feeding of scientific data. PLoS Comput Biol. 2014;10(4):e1003542. https://doi.org/10.1371/journal.pcbi.1003542 PMID: 24763340

33. National Center for Biotechnology Information. How do I cite NCBI services and databases?. Available from: https://support.nlm.nih.gov/kbArticle/?pn=KA-03391

34. Patrias K, Wendling D. Citing medicine: The NLM style guide for authors, editors, and publishers. 2nd ed; 2018. Available from: https://www.ncbi.nlm.nih.gov/books/NBK7256/

35. Piwowar HA, Day RS, Fridsma DB. Sharing detailed research data is associated with increased citation rate. PLoS One. 2007;2(3):e308. https://doi.org/10.1371/journal.pone.0000308 PMID: 17375194

36. Byrd JB, Greene AC, Prasad DV, Jiang X, Greene CS. Responsible, practical genomic data sharing that accelerates research. Nat Rev Genet. 2020;21(10):615–29. https://doi.org/10.1038/s41576-020-0257-5 PMID: 32694666

37. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. https://doi.org/10.1038/sdata.2016.18 PMID: 26978244

38. de Visser C, Johansson LF, Kulkarni P, Mei H, Neerincx P, Joeri van der Velde K, et al. Ten quick tips for building FAIR workflows. PLoS Comput Biol. 2023;19(9):e1011369. https://doi.org/10.1371/journal.pcbi.1011369 PMID: 37768885

39. Rashid U, Wu C, Shiller J, Smith K, Crowhurst R, Davy M, et al. AssemblyQC: a Nextflow pipeline for reproducible reporting of assembly quality. Bioinformatics. 2024;40(8):btae477. https://doi.org/10.1093/bioinformatics/btae477 PMID: 39078114

40. Silva BM, Trindade F d J, Canesin LEC, Souza G, Aleixo A, Nunes GL. Pipeasm: a tool for automated large chromosome-scale genome assembly and evaluation. bioRxiv. 2024. https://doi.org/2024.10.21.598381

41. Obinu L, Booth T, De Weerd H, Trivedi U, Porceddu A. Colora: a Snakemake workflow for complete chromosome-scale de novo genome assembly. Bioinformatics. 2025;41(5):btaf175. https://doi.org/10.1093/bioinformatics/btaf175 PMID: 40238183