

Revolutionizing Biological Research with the NIH Comparative Genomics Resource (CGR)

Valerie Schneider, Ph.D. 10/26/23

Why We Are Here

In what ways does your work relate to CGR-related resources?

Where do you think CGR might have the greatest impact for your clients?

What types of CGR-engagement opportunities might be most valuable?

Outline

- Intro to Comparative Genomics
- The Value of Research Organisms
- Problem
- CGR Solution
- CGR Impact – Two use cases
- What's Next



What is comparative genomics?

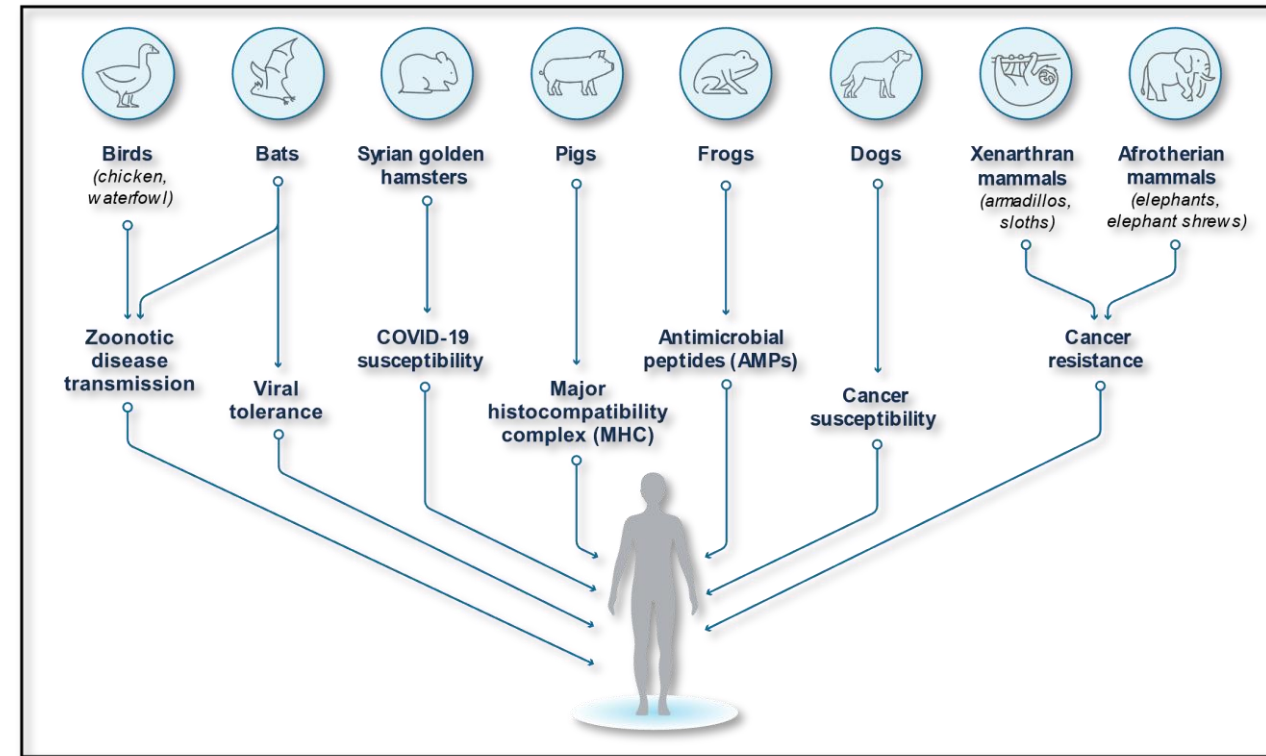


Who does comparative genomics?



The Value of Research Organisms

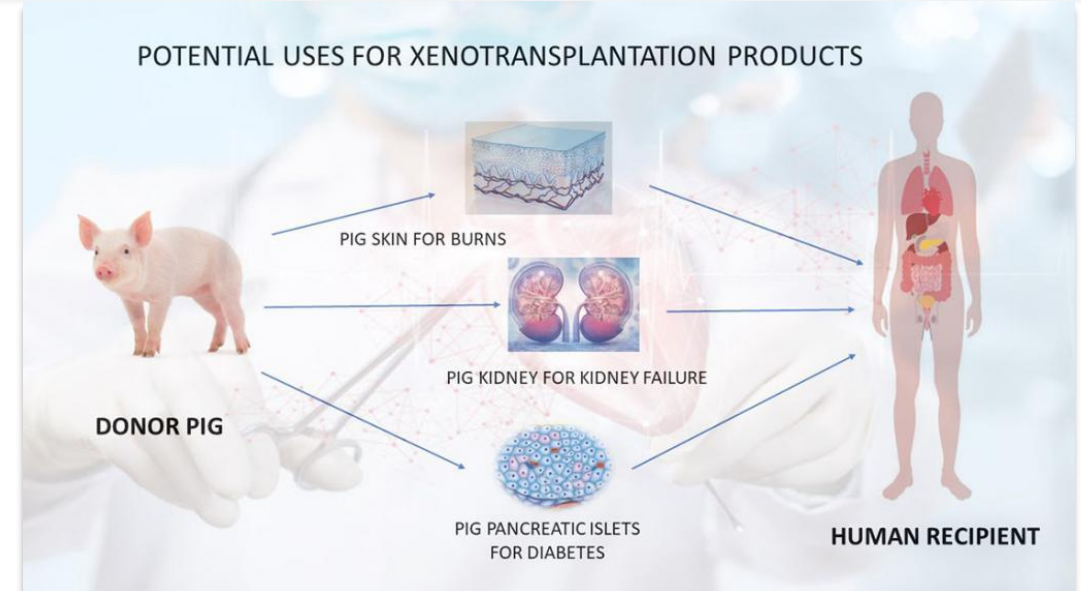
Understand Basic Biological Processes & Human Disease



Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates

Joana Damas^{a,1}, Graham M. Hughes^{b,1}, Kathleen C. Keough^{c,d,1}, Corrie A. Painter^{e,1}, Nicole S. Persky^{f,1}, Marco Corbo^g, Michael Hiller^{g,h,i}, Klaus-Peter Koepfli^j, Andreas R. Pfenning^k, Huabin Zhao^{l,m}, Diane P. Genereuxⁿ, Ross Swofford^o, Katherine S. Pollard^{d,o,p}, Oliver A. Ryder^{q,r}, Martin T. Nweeia^{s,t,u}, Kerstin Lindblad-Toh^{n,v}, Emma C. Teeling^b, Elinor K. Karlsson^{n,w,x}, and Harris A. Lewin^{a,y,z,2}

<https://www.pnas.org/doi/10.1073/pnas.2010146117>

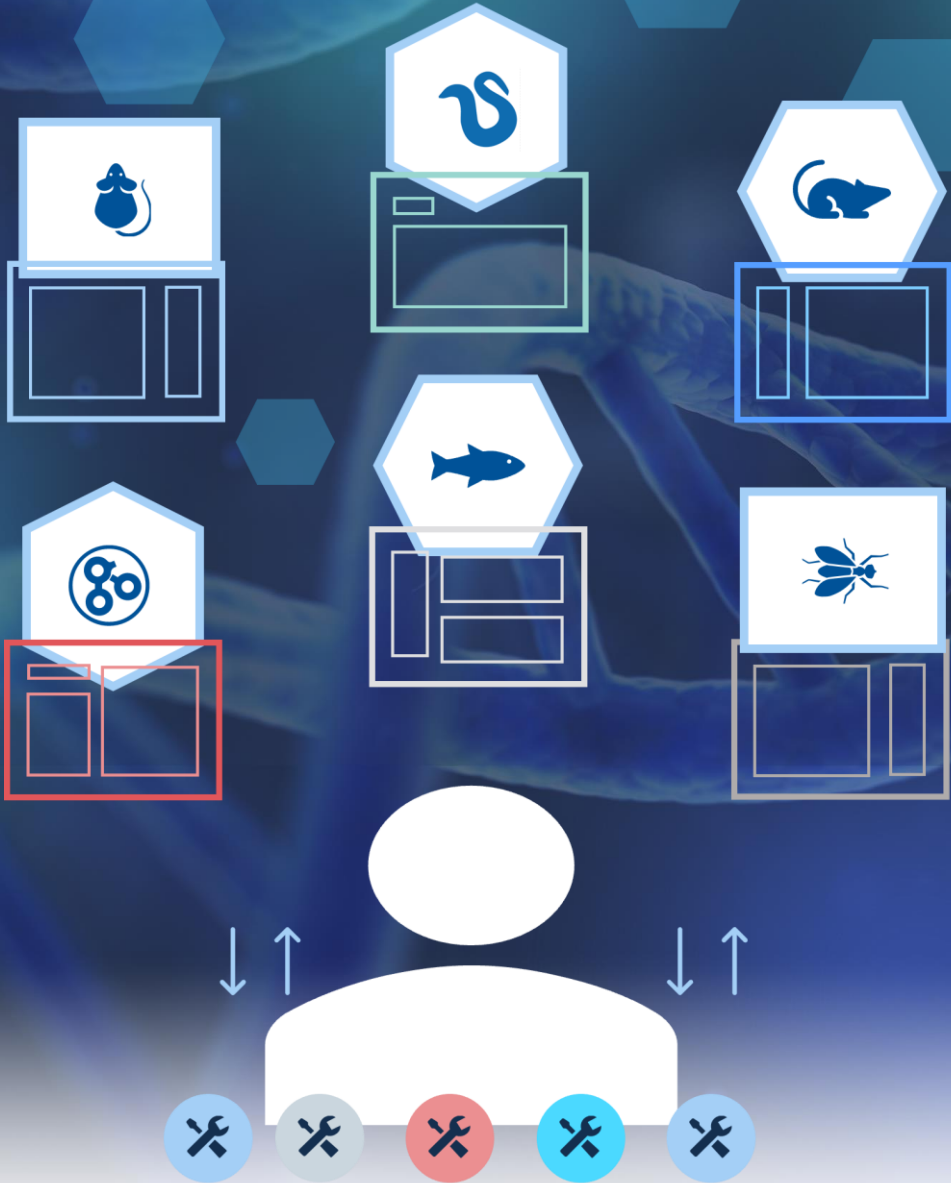


<https://www.fda.gov/vaccines-blood-biologics/xenotransplantation>

Problem

Comparative genomics research faces several limitations and challenges

- **Exponential data growth; variable data quality**
- Multiple different user interfaces
- Limited number of organisms supported
- Siloed data and applications
- Must download data to apply tools
- Limited scalability



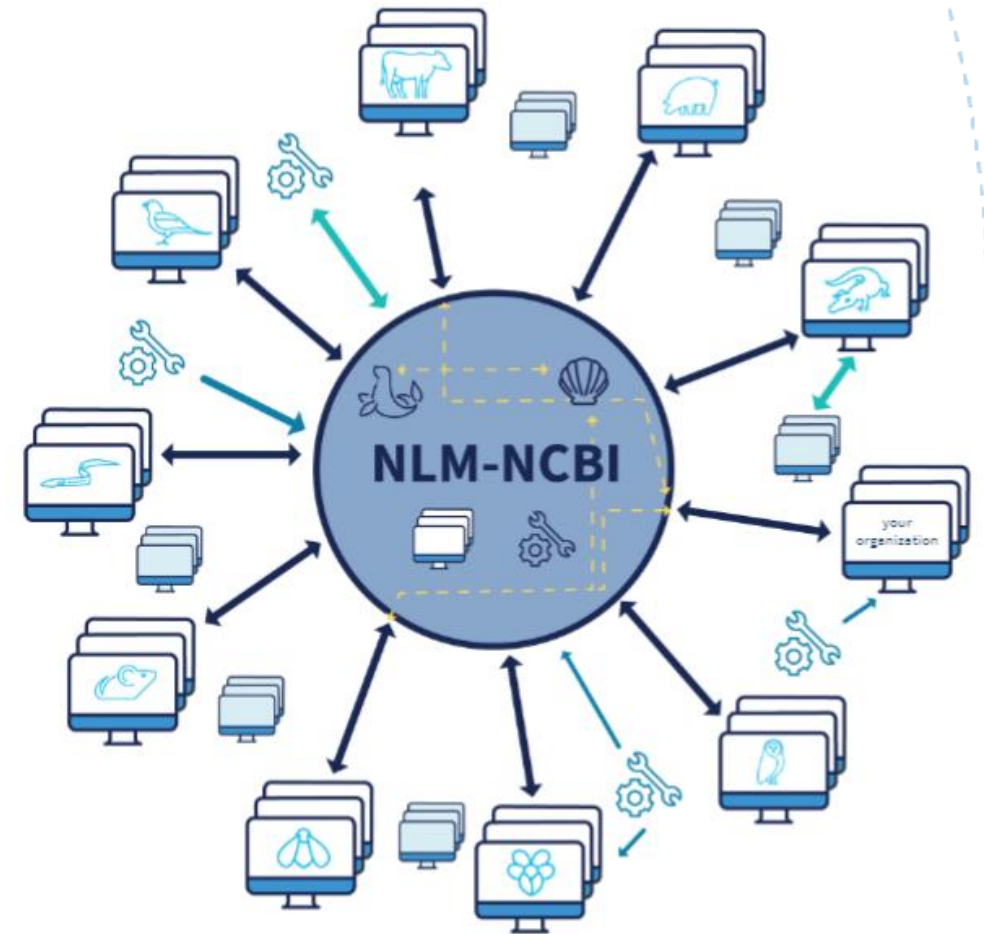
Solution

NIH Comparative Genomics Resource (CGR)

What: CGR **maximizes the impact of eukaryotic research organisms** and their genomic data to biomedical research.

How: CGR facilitates reliable comparative genomics analyses through **community collaboration** and an **NCBI genomics toolkit**. The toolkit includes high-quality data, tools, and **interfaces** for connecting community-provided resources with NCBI.

Outcome: CGR provides you with information and examples about free tools and data so you can **confidently help educators, trainees, researchers, and bioinformaticians** working in comparative genomics.



CGR Components

1 NCBI Toolkit

- Interconnected databases
- Interoperable data and tools

Data Resources

- NCBI Datasets
- Genomes, Genes, Proteins, Expression
- Gene orthology
- Protein architecture

Analysis Tools

- Basic Local Alignment Search Tool (BLAST)
 - ClusteredNR database
- Visualization tools
 - Comparative Genome Viewer (CGV)
 - Multiple Sequence Alignment (MSA) Viewer
 - Genome Data Viewer (GDV)

Data Quality Tools

- Foreign Contamination Screening (FCS) Tool
- Assembled genome QC
- Eukaryotic Annotation Tool (EGAP)

2 Community Collaboration

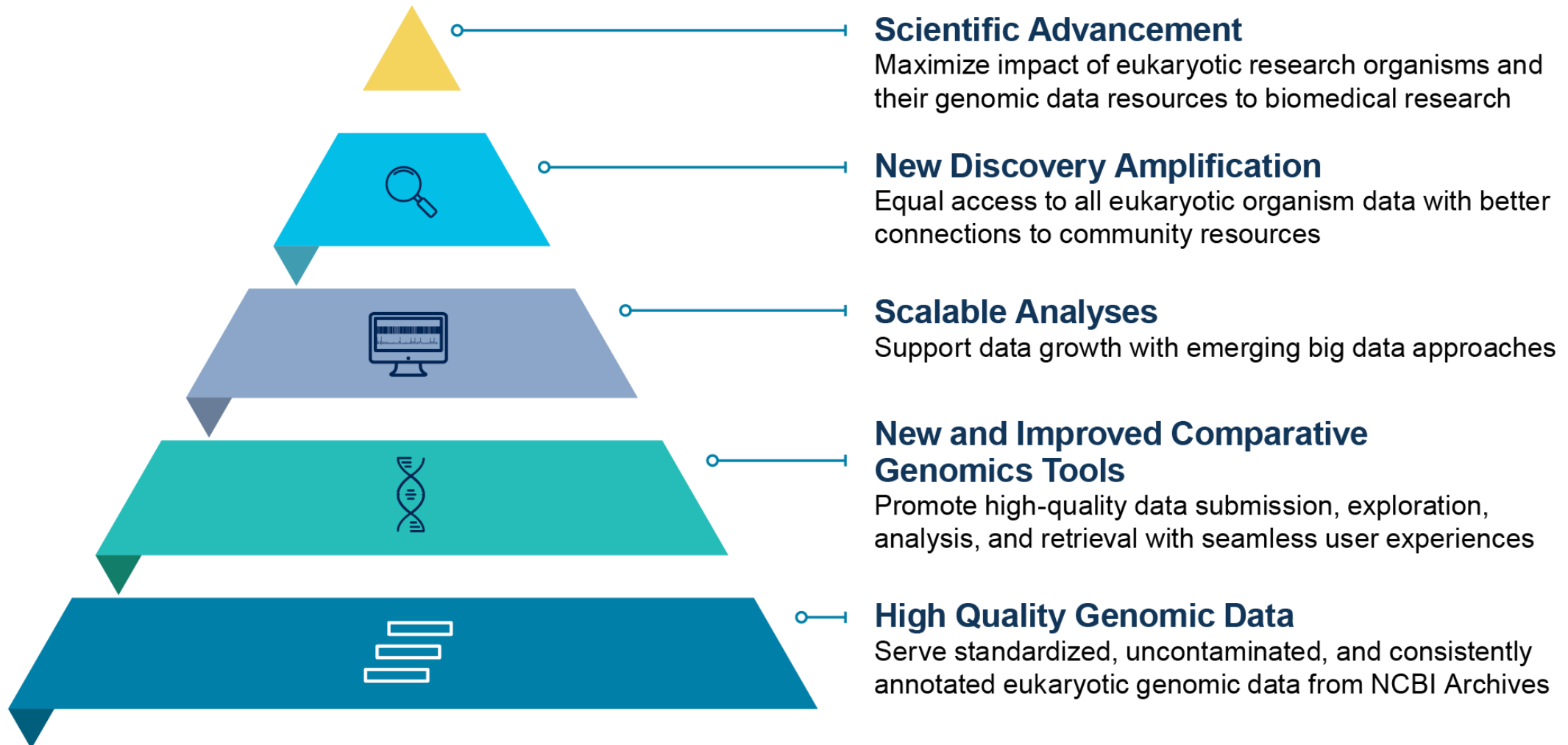
- Connects genome-related data and metadata with the NCBI toolkit
- Informs new developments and improvements

3 Standardized Interfaces

- Connects community content with NCBI content via command line tools, APIs, or resource links



CGR Impact



CGR Impact: Two Case Studies



Finding genomic resources for organisms on NCBI

A public health biologist wants to find and contribute data for an invasive tick species

They need to:

- Identify available genomic data on NCBI for tick species:
 - Datasets Taxonomy
 - Datasets Genome
 - Datasets Command Line
- Improve the quality of their own data:
 - Foreign Contamination Screen
 - Genome Annotation



Haemaphysalis longicornis

Asian longhorned tick

Recently found on the East Coast!

Taxonomy Browser

- Search by common name, species name, higher groups
- Returns table of matches with data availability
- Taxon names link to Taxonomy pages – one stop shop for links to available sequence NCBI data!

Selected taxa

Haemaphysalis longicornis (longhorned tick) ✕ Enter one or more taxonomic names

Taxonomic name	Genomes
▼ Eukaryota (eukaryotes)	34,077
▼ Metazoa (animals)	12,395
▼ Arthropoda (arthropods)	4,797
▼ Arachnida (arachnids)	150
▼ Ixodida (ticks)	44
▼ Ixodidae (hardbacked ticks)	30
▼ Haemaphysalis	4
▼ Haemaphysalis longicornis (longhorned tick)	4

There are four available genome sequences for *H. longicornis* available on NCBI, and 44 for ticks in general.

Taxonomy Page

- Taxonomy-based portal to data: gene expression, raw sequence, overarching projects and more!
- Direct link to designated reference genome
- Links to other CGR resources – annotation table, visualization, BLAST

Database links	
Nucleotide	Protein
All nucleotide sequences 50,367	Protein sequences 27,921
Genomic sequences 2,577	Conserved domains 1
mRNA sequences 47,787	3D structures 2
GEO Datasets	Sequence Read Archive (SRA)
Datasets 0	All SRA experiments 268
Series 3	DNA 192
Samples 37	RNA 76
Platforms 3	
PopSet	Projects and samples
Phylogenetic studies 29	BioProject
Population studies 9	BioSample

Genome
[Browse all 4 genomes](#)

Reference genome
[BIME_HaeL_1.3](#)
TIGMIC Group (2020). Isolate: HaeL-2018.
GenBank GCA_013339765.2

[Download](#)

Summary of data available on NCBI for *H. longicornis*

NCBI Datasets Genome Table: *H. longicornis*

- Summary of available genome sequences. Includes metadata like genome size, quality and annotation availability
- Allows researchers to select most suitable assembly for their research
- Modify visible data using "Select Columns"

Select columns

- GenBank
- Scientific name
- Annotation
- Level
- WGS accession
- Scaffold N50 (kb)
- Sequencing technology
- BioProject
- Genes
- Pseudogenes
- CheckM completeness (%)
- RefSeq
- Modifier
- Size (Mb)
- Release Date
- Contig N50 (kb)
- BUSCO
- Submitter
- BioSample
- Protein-coding
- CheckM marker set
- CheckM contamination (%)

Download ▾ Select columns 4 genomes 1 selected Rows per page 20 ▾

<input type="checkbox"/>	Assembly	RefSeq	Scientific name	Annotation	Size (Mb)	Level	WGS...
<input checked="" type="checkbox"/>	BIME_HaeL_1.3		Haemaphysalis longicorn...	Submitter	2,555	Chromosome	JABSTR...
<input type="checkbox"/>	ASM2241470v1		Haemaphysalis longicorn...		3,156	Scaffold	JADCT01
<input type="checkbox"/>	ASM2984928v1		Haemaphysalis longicorn...		2,477	Contig	JANDBB...
<input type="checkbox"/>	HLAgrLifeRun1		Haemaphysalis longicorn...		7,362	Contig	VFIB01

None of the four *H. longicornis* genomes are curated RefSeq genomes or have RefSeq annotation...What about other species of tick?

What genomic resources do we have for other ticks?

Selected taxa

Haemaphysalis longicornis (longhorned tick) × Enter one or more taxonomic names

Taxonomic name	Genomes
▼ <i>Eukaryota</i> (eukaryotes)	34,077
▼ <i>Metazoa</i> (animals)	12,395
▼ <i>Arthropoda</i> (arthropods)	4,797
▼ <i>Arachnida</i> (arachnids)	150
▼ <i>Ixodida</i> (ticks)	44
▼ <i>Ixodidae</i> (hardbacked ticks)	30
▼ <i>Haemaphysalis</i>	4
<i>Haemaphysalis longicornis</i> (longhorned tick)	4

NCBI Datasets Genome Table: Filtering

- Genome data from related species can be extremely useful to make inferences about our target species
- However, we likely can't analyze all 30 genomes in this family of ticks
- Apply filters for annotation availability and genome quality to select most informative assemblies to study further

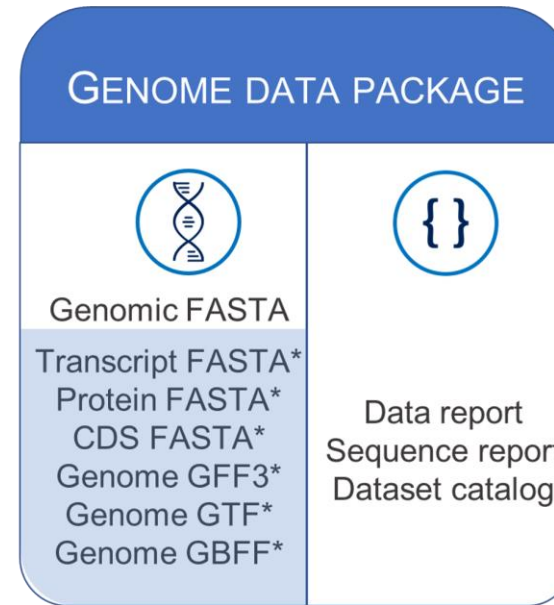
30 genomes assemblies for tick family Ixodidae

The screenshot shows the NCBI Datasets Genome Table filtering interface. At the top, there are three filter buttons: 'Reference', 'RefSeq annotation', and 'scaffold +'. Below this, the 'STATUS' section has several checkboxes: 'Reference genomes' (checked), 'Annotated genomes' (unchecked), 'Annotated by NCBI RefSeq' (checked), 'Annotated by GenBank submitter' (unchecked), 'Exclude atypical genomes' (unchecked), and 'Metagenome-assembled genomes (MAGs)' (unchecked). To the right, the 'SEARCH WITHIN RESULTS' section has a search bar. Below that, the 'ASSEMBLY LEVEL' section has a slider ranging from 'contig' to 'complete', with 'scaffold' selected. The 'YEAR RELEASED' section has a slider ranging from 1980 to 2023, with the start at 1980. A blue arrow points from the '30 genomes assemblies for tick family Ixodidae' text to the top of the interface, and another blue arrow points from the 'YEAR RELEASED' slider to the text below.

Five assemblies that are Refseq annotated and meet our assembly level standards

Genome Data package

- Once you have identified relevant genome assemblies, download bulk data as efficient packages
- Get either actual sequencing and annotation data or metadata for further filtering
- Variety of industry-standard file formats for use in bioinformatics pipelines



Download ▾ Select columns 5 genomes 2 selected Rows per page 20 1-5 of 5 < >

<input type="checkbox"/>	Assembly	Scientific name	Annotation	Level ↑	BUSCO	Action
<input checked="" type="checkbox"/>	BIME_Rsan_1.4	Rhipicephalus sanguineus (brown dog tick)	NCBI RefSeq Submitter	Chromosome	C:96.6%[S:89.4%,D:7.2%],F:0.8%,M:2.6%,n:2934 arachnida_odb10	⋮
<input checked="" type="checkbox"/>	BIME_Rmic_1.3	Rhipicephalus microplus (southern cattle tick)	NCBI RefSeq Submitter	Chromosome	C:95.7%[S:91.5%,D:4.2%],F:0.5%,M:3.8%,n:2934 arachnida_odb10	⋮

After filtering for genome quality and annotation availability, we select two species of related ticks to study further by downloading sequence data and associated metadata right from the Genome Table interface.

NCBI Datasets: Genome Annotation Table

- Download gene, transcript and protein sequences, and metadata
- Tables are available for ~7500 eukaryotic annotated genomes
- Available for both RefSeq and GenBank submitted annotations
- Filter by gene type, gene name, and chromosome or location on the genome

Southern Cattle tick – a better-studied relative!



Genes

Genes annotated on *Rhipicephalus microplus* (southern cattle tick) BIME_Rmic_1.3 (GCF_013339725.1)

Annotation Name: NCBI Annotation Release 100 (November 4, 2020)

Filters evasin

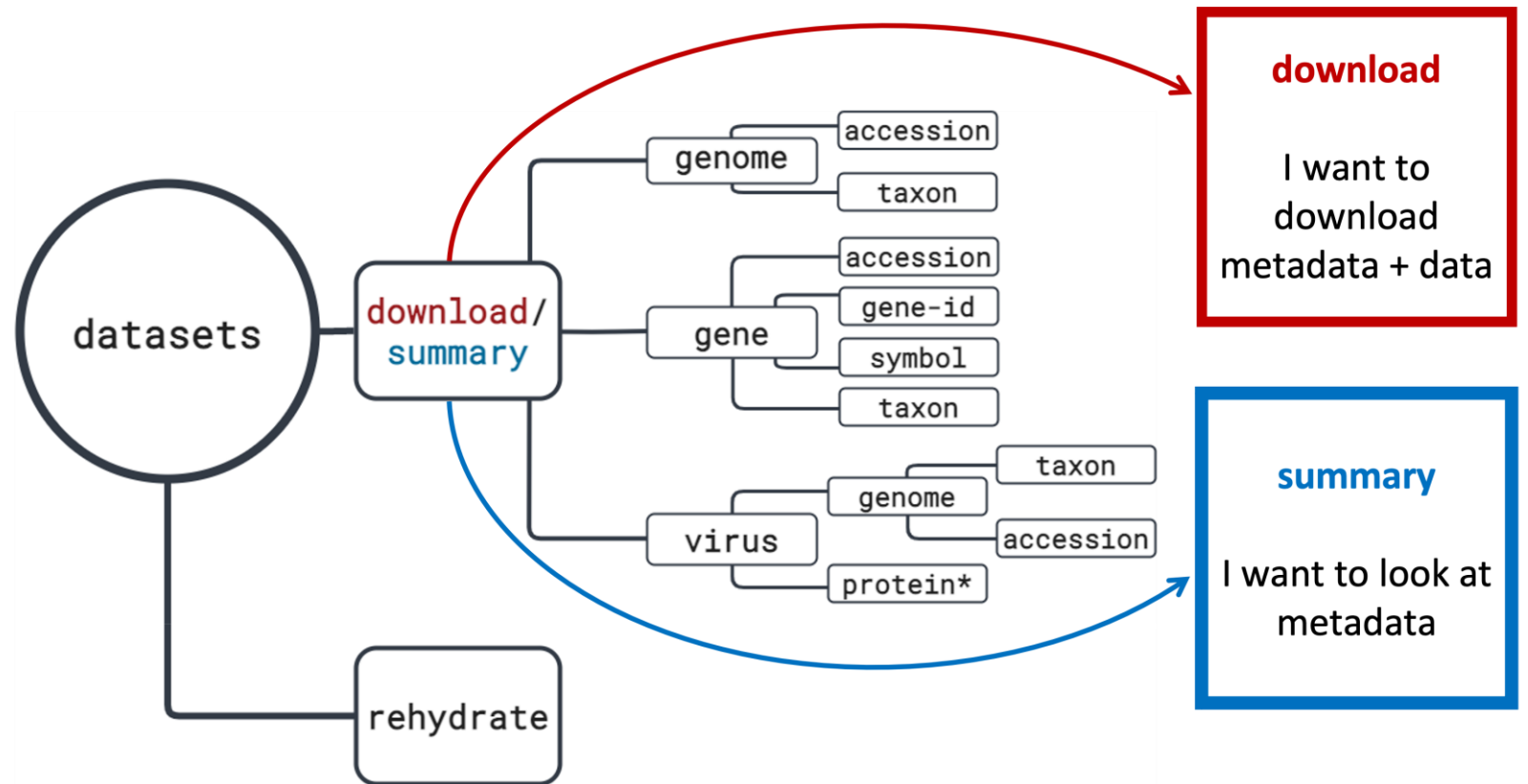
DOWNLOAD Select columns 1 Gene Rows per page 20 1-1 of 1

<input type="checkbox"/>	Genomic location	Chromosome	Orientation	Name	Symbol	Action
<input type="checkbox"/>	NC_051165.1:111928293-112022529	1	minus	evasin-1-like	LOC11915996	⋮

We get one result for searching the *R. microplus* annotation set for *evasin*, a protein in tick saliva that helps evade the host immune system. From here, download data or follow a link to the gene page to learn more!

NCBI Datasets Command Line Interface

- Same information available in the web interfaces
- Look at metadata without downloading large files
- Available in bioinformatics ecosystems like Galaxy
- NCBI Datasets content also available via REST API



Install with **conda**

Galaxy
COMMUNITY HUB

www.ncbi.nlm.nih.gov/datasets/docs/v2/download-and-install/

Using the Comparative Genomics Resource we quickly...

- Found genomic related data for a focal species
- Found genomic related data for related species
- Sorted and filtered those assemblies using metadata
- Sorted and filtered annotated features (genes) using metadata

But what if we want to improve or annotate our own genome assemblies?

- Foreign Contamination Screen
- Eukaryotic Genome Annotation Pipeline

Foreign Contaminant Screening (FCS)

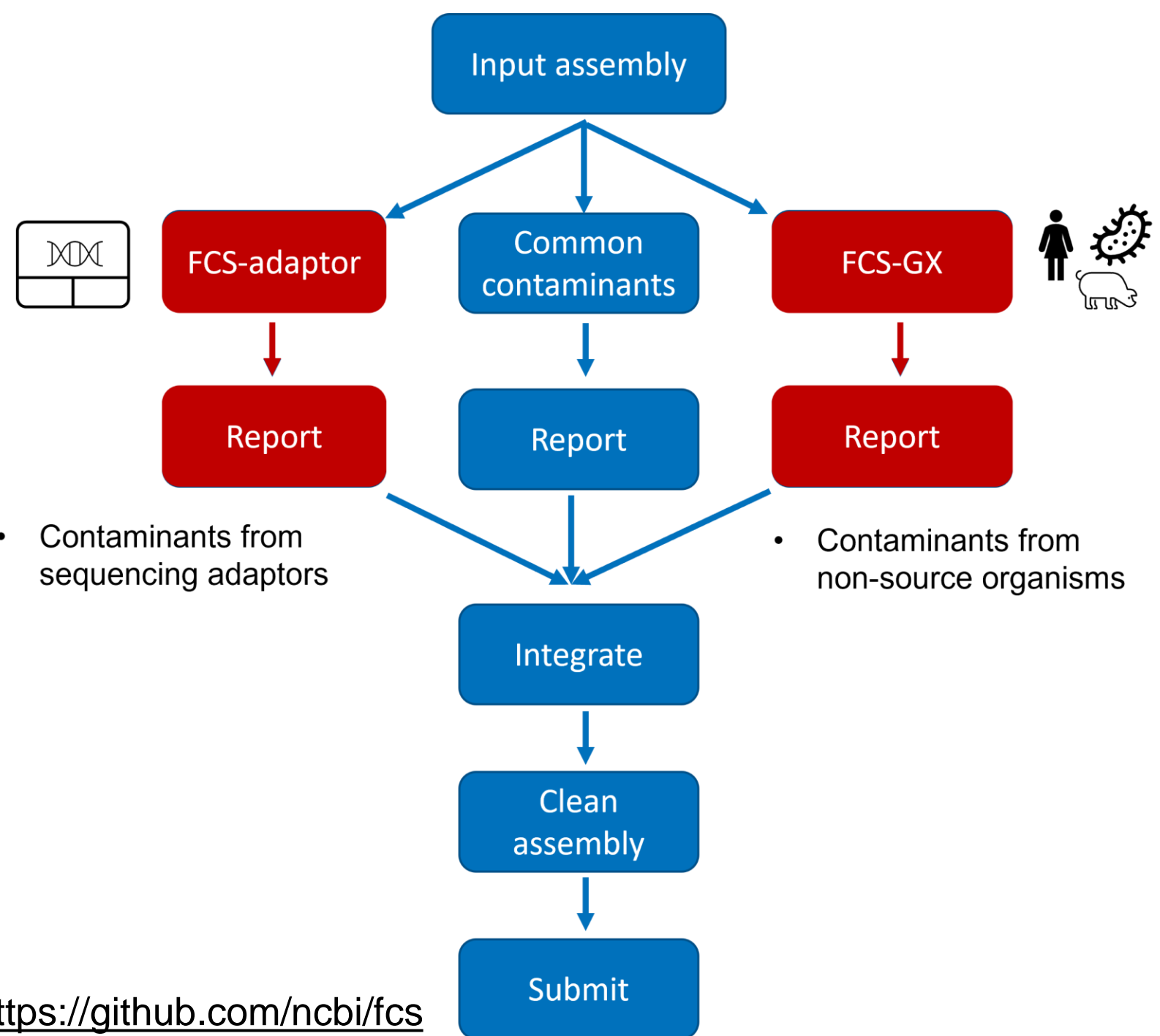
- In 2022, 1 of every 3 eukaryotic genomes submitted to GenBank had detectable contamination

Required User Inputs:

- genome assembly
- NCBI taxonomy identifier

User gets:

- contamination summary report
- actions for cleaning genome
- cleaned genome, contaminants file



EGAP: Eukaryotic Genome Annotation Pipeline

Used by NCBI to annotate >1000 species

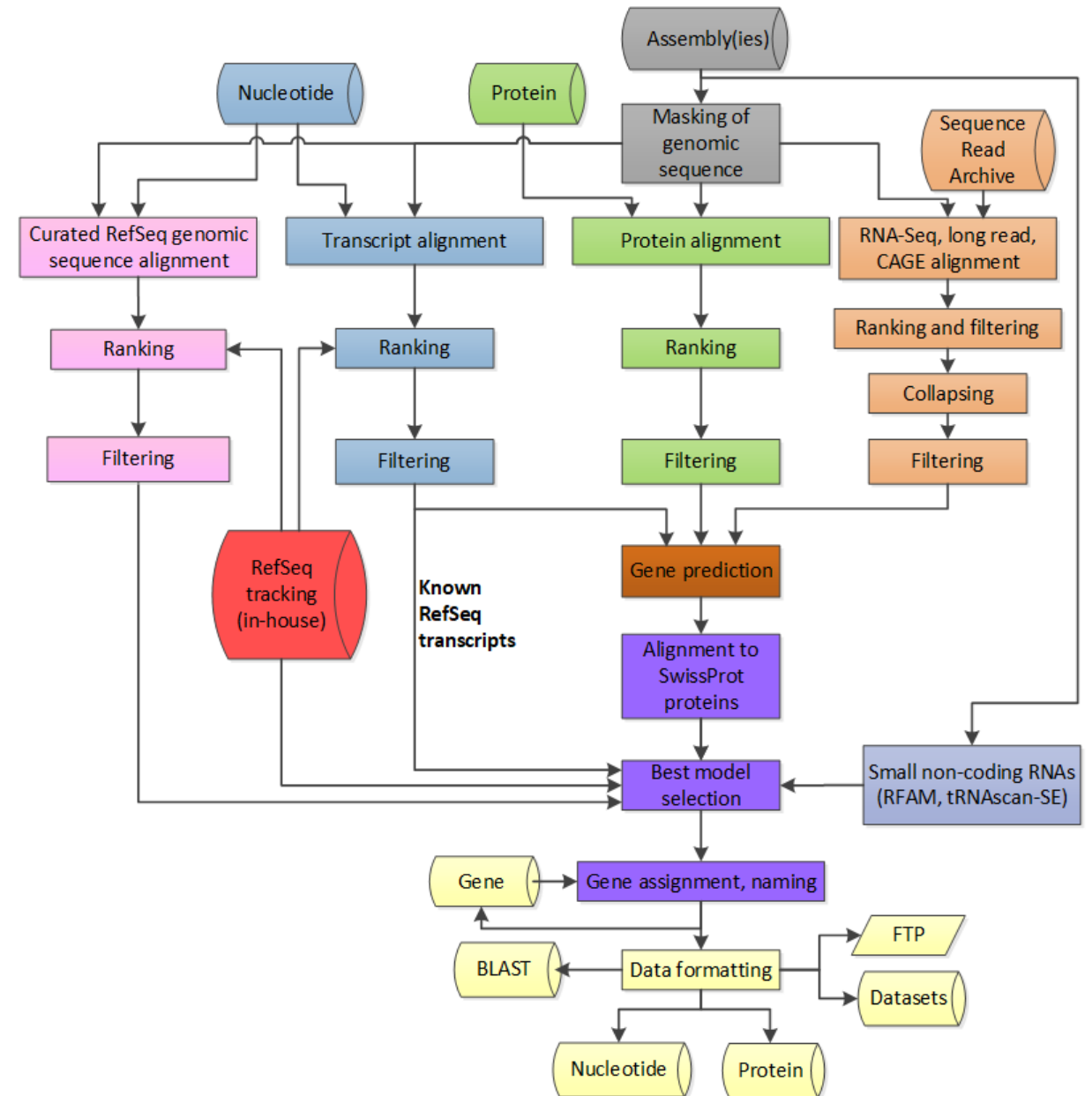
Evidence used for gene prediction:

- ✓ ESTs
- ✓ cDNAs
- ✓ Same and cross-species proteins
- ✓ RNA-Seq
- ✓ PacBio IsoSeq, ONT transcriptomes
- ✓ CAGE

GIVE FEEDBACK!

Cloud-compatible containerized EGAP for public use.

Want to be an alpha tester?



Case Study 1 Summary

Find and contribute genomic data for an invasive tick species

Find Data:

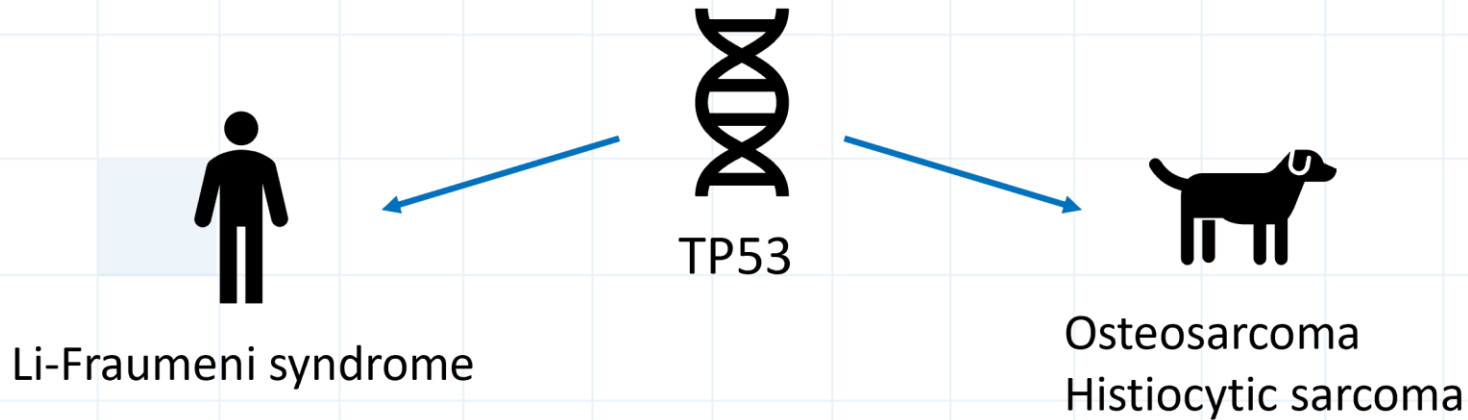
- **Taxonomy Browser**
- **Improved Taxonomy Pages**
- **Datasets Genome Table**
- **Datasets Command Line Interface/API**
- **Datasets Annotation Table**

Improve your data:

- **Foreign Contamination Screen**
- **Eukaryotic Genome Annotation Pipeline**

Case Study 2

Making discoveries in cancers common to humans and dogs



The following are some resources that can help in this research:



NCBI Gene

Multiple Sequence Alignment (MSA) Viewer



Comparative Genome Viewer (CGV)

NCBI Orthologs



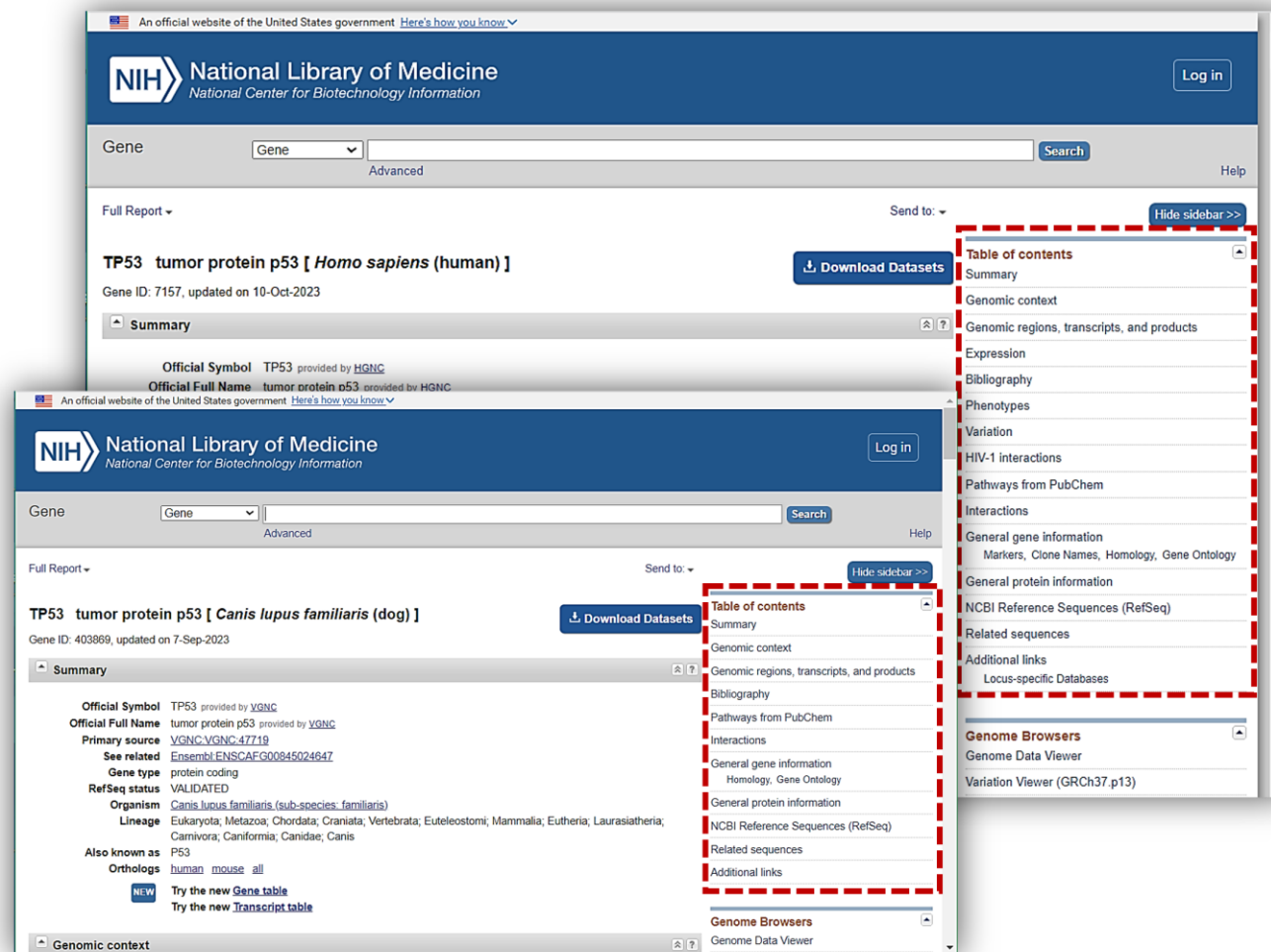
Genome Data Viewer (GDV)

iCn3D



Summary records of an organism's gene-specific information, including sequences, expression data, published literature, functional domains, structures and homologs with *access to more...*

- *Displayed information is aggregated from all relevant NCBI database records and tools*
- *Hyperlinks to related resources such as PubMed, GDV, NCBI Orthologs*
- *Data is accessible via the web, FTP, NCBI Datasets, Eutilities APIs, and the Edirect command-line tool*



The image displays two screenshots of the NCBI Gene website. The top screenshot shows the human TP53 gene record, and the bottom screenshot shows the dog TP53 gene record. Both records include a 'Table of contents' sidebar with various sections like Summary, Genomic context, and Bibliography. The dog record shows more detailed information in the main content area.

Human TP53 Gene Record:

- Gene: TP53 tumor protein p53 [*Homo sapiens* (human)]
- Gene ID: 7157, updated on 10-Oct-2023
- Official Symbol: TP53 provided by HGNC
- Official Full Name: tumor protein p53 provided by HGNC

Dog TP53 Gene Record:

- Gene: TP53 tumor protein p53 [*Canis lupus familiaris* (dog)]
- Gene ID: 403869, updated on 7-Sep-2023
- Official Symbol: TP53 provided by VGNC
- Official Full Name: tumor protein p53 provided by VGNC
- Primary source: VGNC:VGNC:47719
- See related: Ensembl:ENSCAFG00845024647
- Gene type: protein coding
- RefSeq status: VALIDATED
- Organism: *Canis lupus familiaris* (sub-species: familiaris)
- Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Laurasiatheria; Carnivora; Caniformia; Canidae; Canis
- Also known as: P53
- Orthologs: [human](#) [mouse](#) [all](#)

The human TP53 gene record contains has much more information than displayed for the dog version. This data may help to fill in knowledge gaps for this lesser-studied gene.

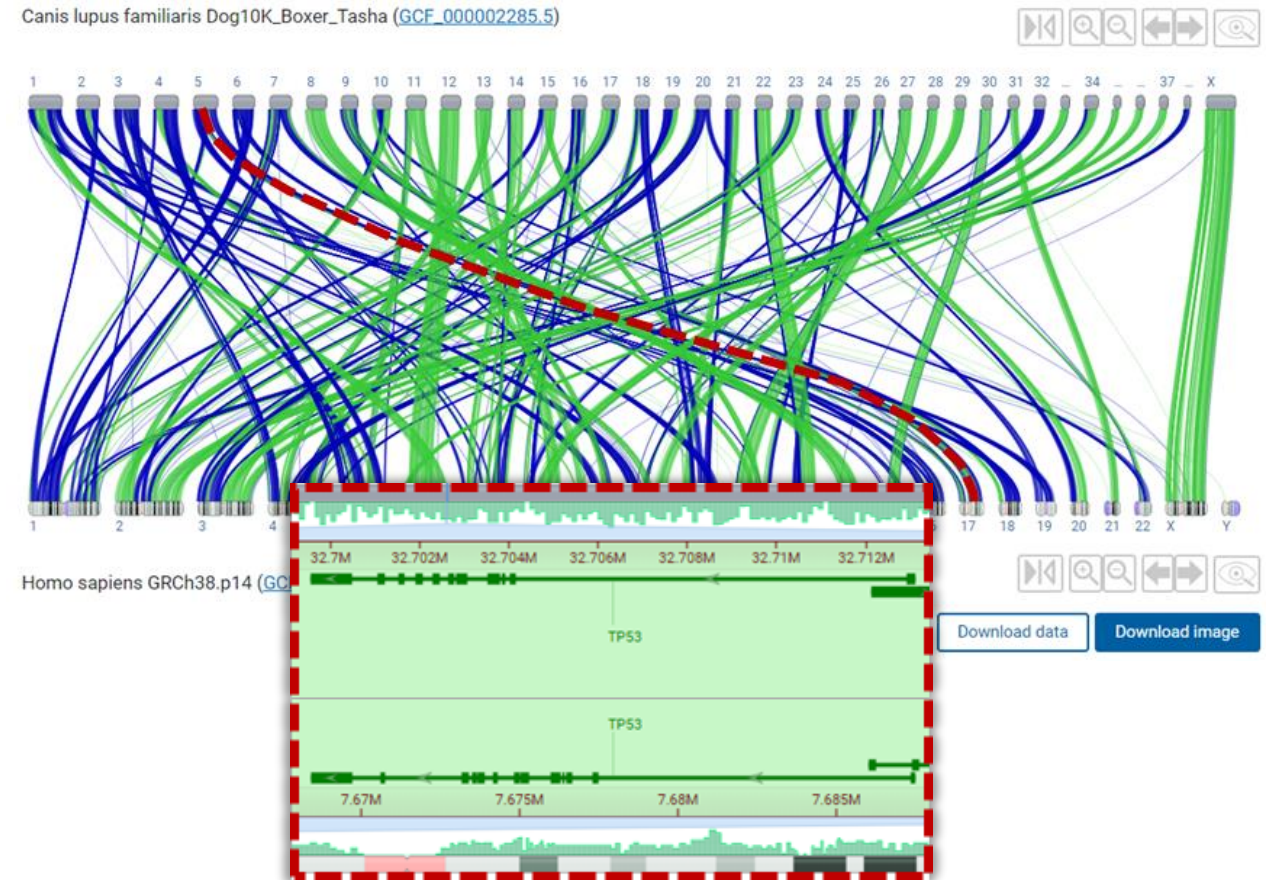


Comparative Genome Viewer (CGV)

Compare two genomes and their gene annotations based on assembly-assembly alignments

- *Zoom to multiple levels*
- *Search by gene symbol or name*
- *Compare gene annotations*

www.ncbi.nlm.nih.gov/genome/cgv



The human and dog TP53 genomic regions appear largely similar but exhibit gene structural differences. This tool can be used as a gateway to do more detailed analysis by facilitating access to other key visualization resources.

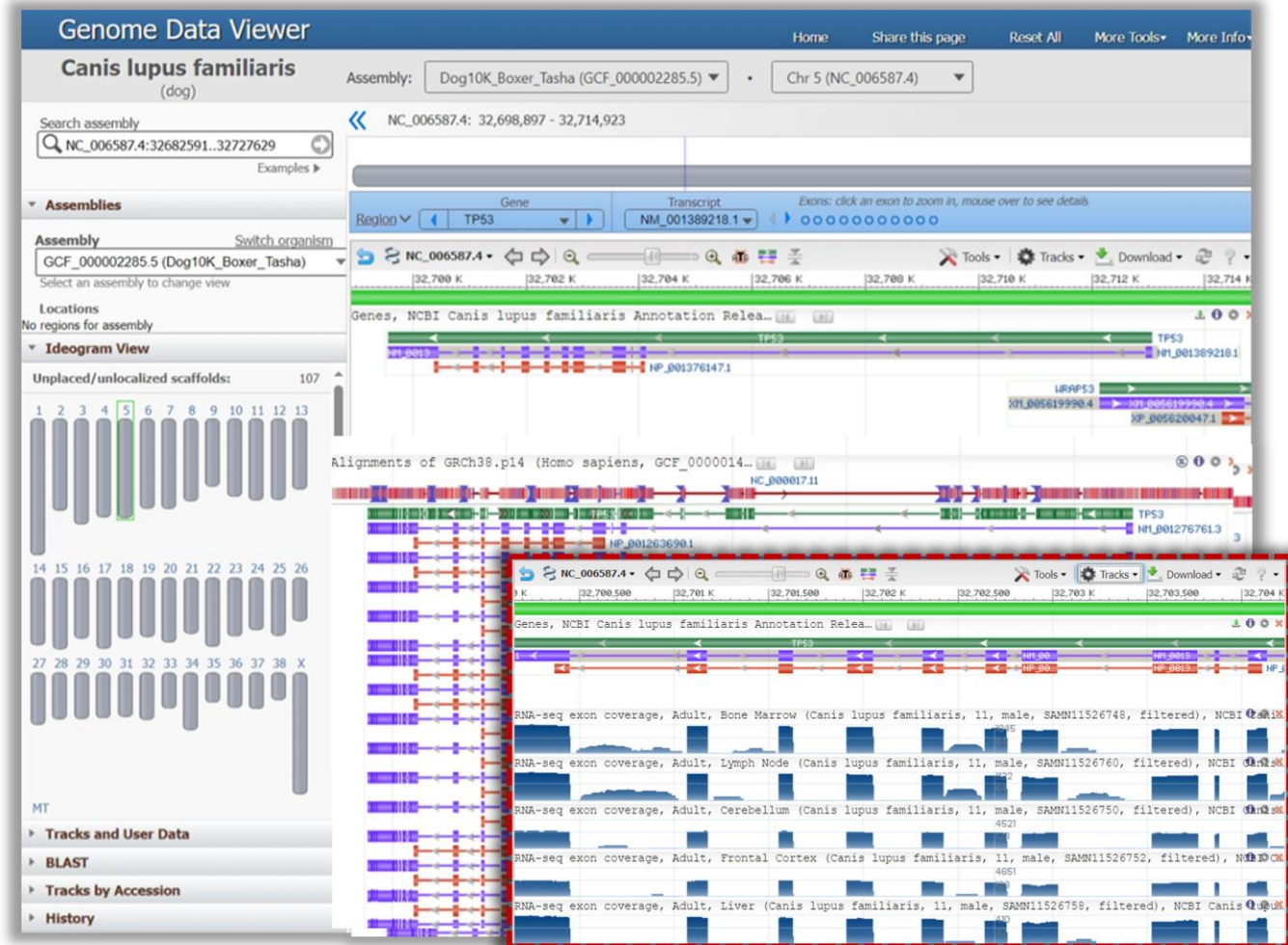


Genome Data Viewer (GDV)

Explore and analyze genomic regions via annotations and alignments.

- *The interactive display enables you to zoom in/out, search for annotations and features, customize the display and download an image or the underlying data.*
- *Explore available NCBI data tracks or upload Track Hub Registry tracks or your own data!*
- *This is a continually developing resource with new data tracks added as new data comes in.*

www.ncbi.nlm.nih.gov/genome/gdv



The human and dog genomic alignments in the TP53 gene region enable direct comparison of differences in known transcript variants. In addition, the GDV browser enables comparisons with other track annotations, including RNA seq expression data for specific tissues.

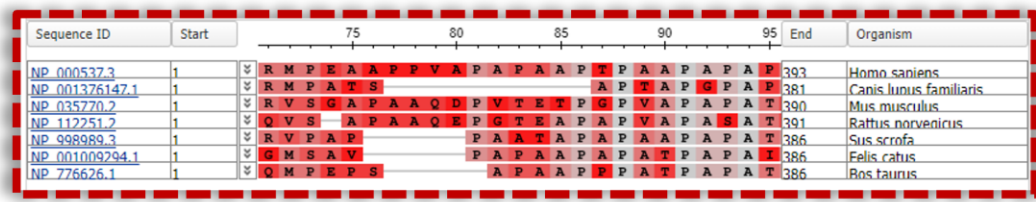
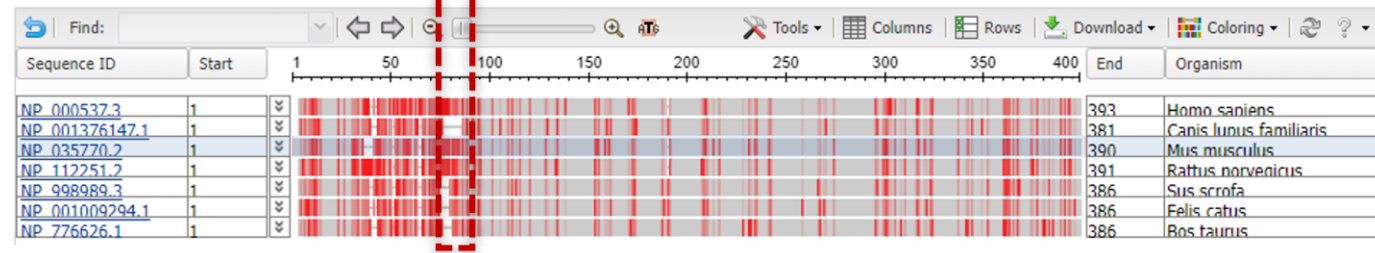


Multiple Sequence Alignment (MSA) Viewer

Compare and examine nucleotide or protein sequence alignments.

- The interactive display enables you to zoom in/out, color residues based on several different schema, customize the display and download an image or the underlying data.*
- This resource can be used to visualize your own sequence alignments or can be accessed from a BLAST search result or a Comparative Genome Viewer (CGV) alignment.*
- The MSA Viewer is also available with an API call for embedding in your own webpage.*

www.ncbi.nlm.nih.gov/tools/msaviewer/



In comparison with the human TP53 protein sequence, the dog and other mammals have a region showing significant sequence diversity. Using MSA Viewer it is easy to zoom in and examine that region.

NCBI Orthologs

Find related genes in other organisms and perform quick sequence comparisons.

- *Select transcript or protein sequences for download or alignment*
- *Refine your results with the taxonomy tree*
- *Drill down to species level information*
- *Access relevant PubMed citations for orthologs*

Learn more: ncbiinsights.ncbi.nlm.nih.gov/2019/04/24/searching-for-orthologous-genes-at-ncbi/

The screenshot shows the NCBI Orthologs interface for the gene TP53. The page title is "TP53 - tumor protein p53". Below the title, there is a description of the gene and its function. The main section is titled "NCBI Orthologs" and shows a search for "TP53" resulting in 421 genes for jawed vertebrates. A taxonomy tree on the left allows for filtering results. A table of results is displayed, with columns for Species, Gene, and a download icon. A "Download data" dialog box is open, showing a list of file types: RefSeq transcripts (FASTA), RefSeq proteins (FASTA), and Tabular data (CSV). The "Download" button is highlighted.

Species	Gene	Download
<input checked="" type="checkbox"/> Homo sapiens human	TP53 tumor protein p53	
<input checked="" type="checkbox"/> Mus musculus house mouse	Trp53 transformation related protein 53	
<input type="checkbox"/> Rattus norvegicus Norway rat	TP53 tumor protein p53	
<input type="checkbox"/> Danio rerio zebrafish	tp53 tumor protein p53	374
<input checked="" type="checkbox"/> Canis lupus familiaris dog	TP53 tumor protein p53	381
<input checked="" type="checkbox"/> Sus scrofa pig	TP53 tumor protein p53	386
<input type="checkbox"/> Gallus gallus chicken	TP53 tumor protein p53	367
<input checked="" type="checkbox"/> Bos taurus cattle	TP53 tumor protein p53	386
<input checked="" type="checkbox"/> Felis catus domestic cat	TP53 tumor protein	386

NCBI's genome annotation pipeline has identified TP53 orthologous sequences for over 400 organisms. In this resource, sets of selected sequences can be quickly aligned or downloaded.



iCn3D

Visualize and map locations of a protein's key sequence residues to its 3D structure, along with NCBI annotations such as the positions of known clinical variants.

- *Interactive display*
- *Customize and download image or the underlying data*
- *Align multiple structures*
- *Source code available on GitHub*

www.ncbi.nlm.nih.gov/Structure/icn3d/icn3d.html

Query: NP_001376147.1; target: 1TSR_A, P53 CORE DOMAIN



Protein 1TSR_A	94	HNYMNCNSSCMGGMNRRP	LIITLEDSSGNLLGRNSFEVRVCACPGRDRRTE	human
BLAST, E: 6.5e-145		HNYMNCNSSCMGGMNRRP	LIITLEDSSGN+LGRNSFEVRVCACPGRDRRTE	
Query: NP_001376147.1	81	HNYMNCNSSCMGGMNRRP	LIITLEDSSGNVLRNSFEVRVCACPGRDRRTE	dog
domain: P53	180 Res	HNYMNCNSSCMGGMNRRP	LIITLEDSSGNLLGRNSFEVRVCACPGRDRRTE	
ClinVar	171 Res	RC.S.D.I.G.D.-.AGL.A.A.I.L.L.K.L.I.P.	V.F.A.Q.G.G.Y.T.C.D.Y.P.E.Q.-.-C.K.G.C.A.R.D.F.A.E.G.E.Q.C.I.K.	
site: DNA binding...	8 Res	-----N-S-----R-----	-----R-CAC--R-----	
site: zinc bindin...	4 Res	-----C--C-----	-----	
site: dimerizatio...	4 Res	-----	-----	

The core DNA binding domain for human TP53 has a sequence very similar to that of the dog. In mapping known ClinVar pathogenic variants to the structure, a similar genetic variant impact may also be predicted for the dog TP53.

Case Study 2 Summary

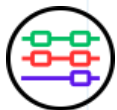
Making discoveries in cancers common to humans and dogs



NCBI Gene: We learned what is known about the dog TP53 gene and what information we might infer from what is known about the well-studied human version.



Comparative Genome Viewer (CGV): We were able to align and explore the synteny for the dog and human TP53 genomic regions.



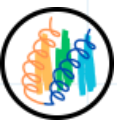
Genome Data Viewer (GDV): We examined the dog and human TP53 gene annotations alongside other annotation tracks including dog tissue-specific RNAseq expression data.



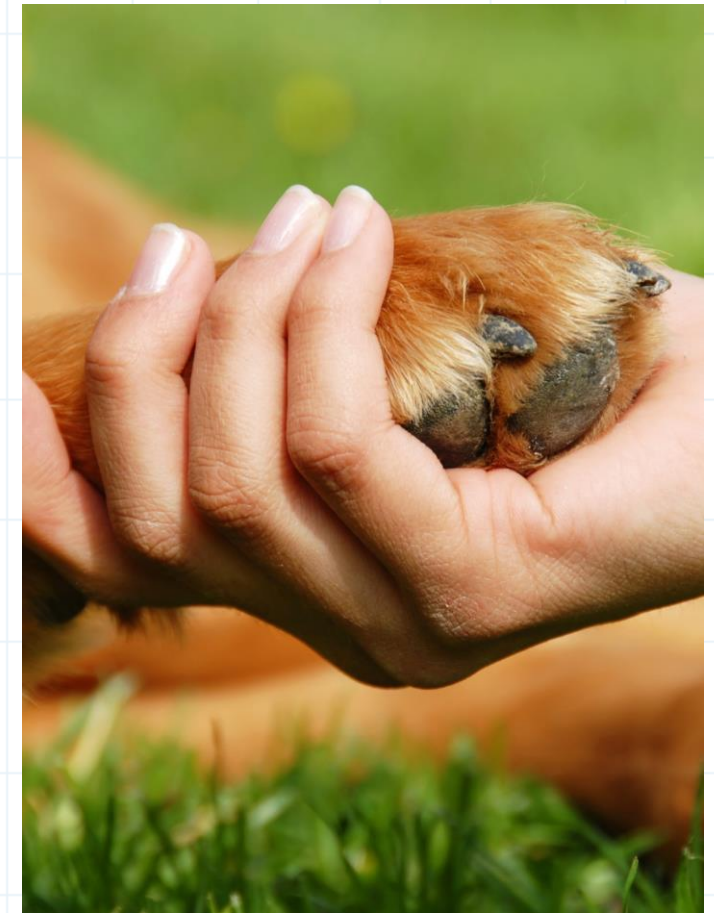
Multiple Sequence Alignment (MSA) Viewer: We were able to explore and directly compare the human, dog and other mammalian TP53 protein sequences.



NCBI Orthologs: We quickly found TP53 orthologous sequences for over 400 organisms and were able to quickly download datasets for organisms we selected, such as human, dog, mouse, rat, and pig.



iCn3D: We were able to interactively visualize the human TP53 protein 3D structure and directly map aligned human and dog sequences and annotations such as known ClinVar pathological clinical variants.



What's next for CGR?

- Ongoing resource improvements based on community feedback
- Making EGAPx publicly available and expanding its taxonomic scope
 - Alpha testers wanted!
- More data available in CGV

How Do I Learn More and Get Involved?



Reach out to us at
cgr@nlm.nih.gov



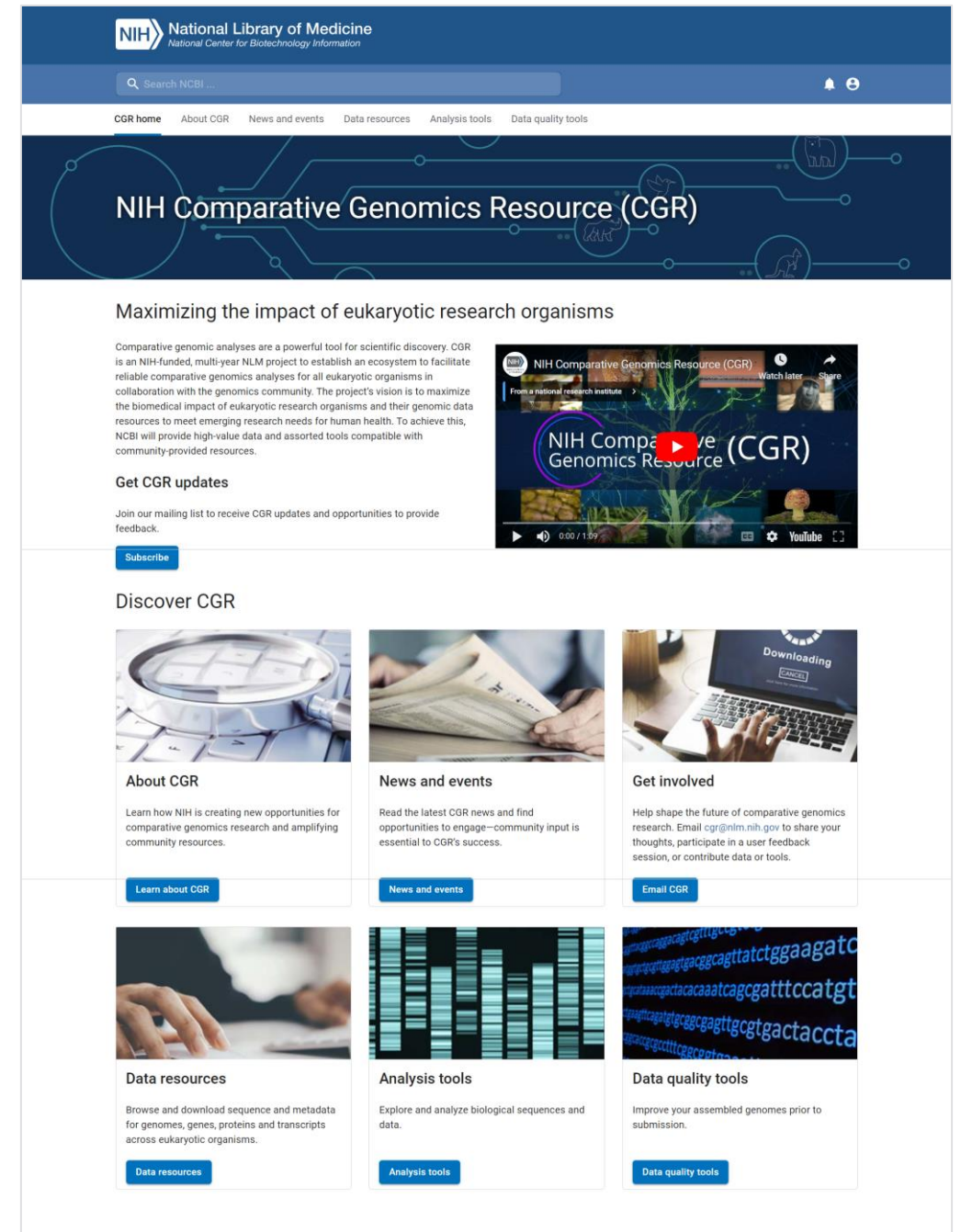
Sign up for our
mailing list -
bit.ly/Subscribe_CGR



Visit the CGR website
ncbi.nlm.nih.gov/cgr
and click the yellow
Feedback button on the
bottom right of the page



Look out for future
meetings, workshops,
webinars, surveys, small
group sessions, user
testing, and interviews to
inform the development
process



NIH National Library of Medicine
National Center for Biotechnology Information

Search NCBI ...

CGR home About CGR News and events Data resources Analysis tools Data quality tools

NIH Comparative Genomics Resource (CGR)

Maximizing the impact of eukaryotic research organisms


Comparative genomic analyses are a powerful tool for scientific discovery. CGR is an NIH-funded, multi-year NLM project to establish an ecosystem to facilitate reliable comparative genomics analyses for all eukaryotic organisms in collaboration with the genomics community. The project's vision is to maximize the biomedical impact of eukaryotic research organisms and their genomic data resources to meet emerging research needs for human health. To achieve this, NCBI will provide high-value data and assorted tools compatible with community-provided resources.

Get CGR updates

Join our mailing list to receive CGR updates and opportunities to provide feedback.

[Subscribe](#)


Discover CGR



About CGR

Learn how NIH is creating new opportunities for comparative genomics research and amplifying community resources.


[Learn about CGR](#)



News and events

Read the latest CGR news and find opportunities to engage—community input is essential to CGR's success.


[News and events](#)



Get involved

Help shape the future of comparative genomics research. Email cgr@nlm.nih.gov to share your thoughts, participate in a user feedback session, or contribute data or tools.

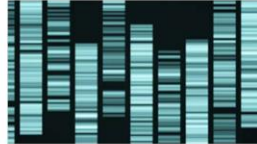
[Email CGR](#)



Data resources

Browse and download sequence and metadata for genomes, genes, proteins and transcripts across eukaryotic organisms.


[Data resources](#)



Analysis tools

Explore and analyze biological sequences and data.

[Analysis tools](#)

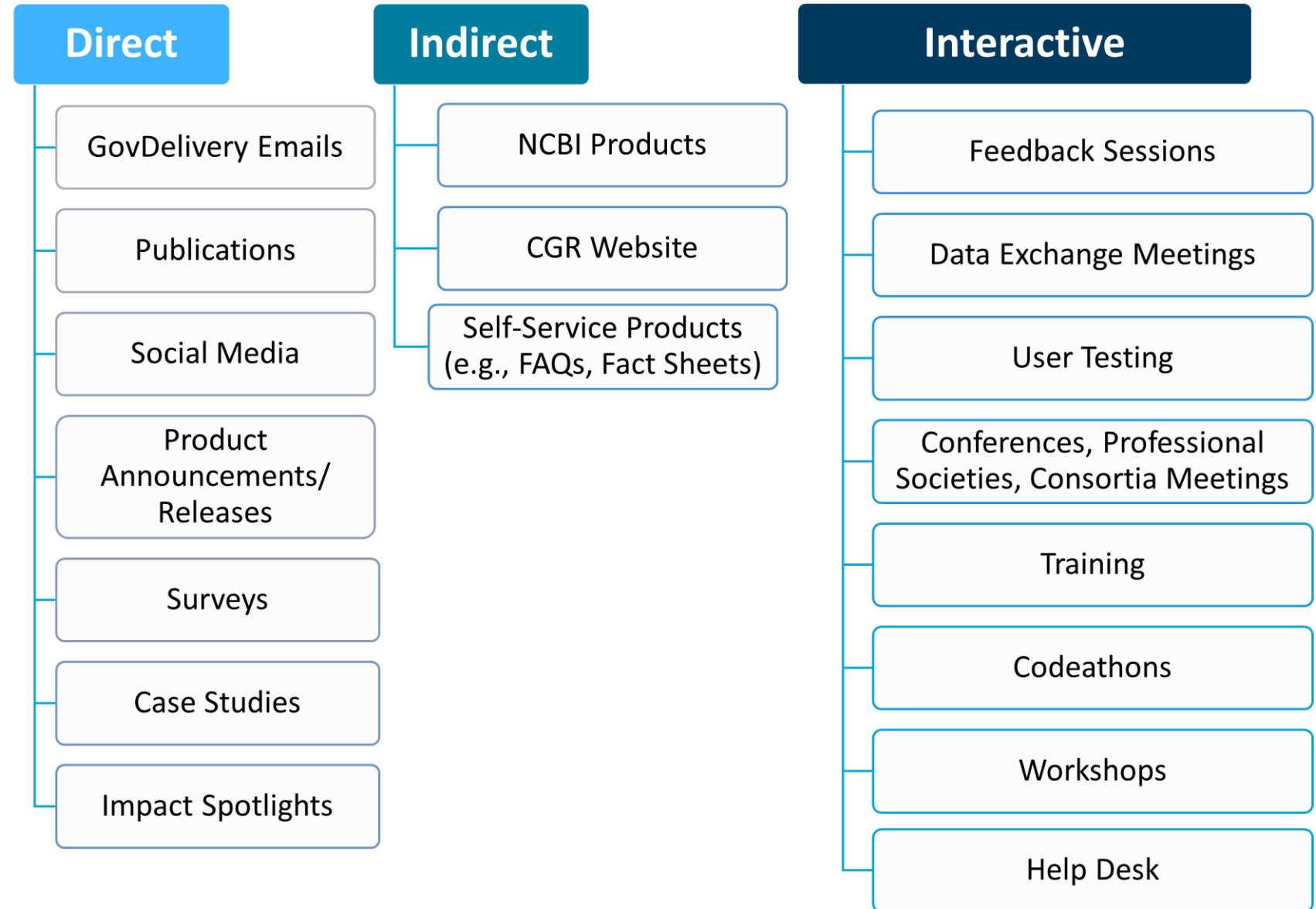


Data quality tools

Improve your assembled genomes prior to submission.

[Data quality tools](#)

Engagement Opportunities



For you today:

CGR Impact Spotlight

The publication below is an example of the types of analyses that could benefit from the National Institutes of Health's (NIH) Comparative Genomics Resource (CGR) facilities and/or comparative genomics analyses for eukaryotic organisms through community collaboration and a National Center for Biotechnology Information (NCBI) genomics toolkit.



Research Summary

Publication: Redmond AG, Pizzarello R, Bakke JN, Dooley H, Shanks T. Evidence for a Highly Complex TNF3 Superfamily in the Jawed Vertebrate Ancestral Lineage. *J Immunol*. 2022 Nov 1;209(10):2173-2178. doi: 10.1093/immunol/kjaa300. Epub 2022 Sep 16. PMID: 36113863.

Topic: Protein family evolution and orthology inference for a particular gene family

Researchers performed a variety of bioinformatic analyses to survey tumor necrosis factors (TNF) in five cartilaginous fishes that occupy a critical phylogenetic position outside bony fish and quadrupeds. Notably, with few exceptions, plants include orthology of all human TNF Superfamilies (TNFSF), suggesting sharks could be used as a model for understanding TNFSF evolution.

Researchers discovered that sharks harbor more than 30 TNFSF genes, which is more than all other vertebrates, due to retention of an ancestral repertoire and lineage-specific expansions. Cytokines of the TNF superfamily are important for immune function and implicated in many human diseases. Researchers discovered the cartilaginous fish immune system may be less primitive than predicted when compared with mammalian systems.

Potential CGR Impact on Research

The following are examples of how CGR resources and capabilities could impact this study.

NCBI Databases: Web interfaces and command line tools would allow for rapid generation of a reference set of bony fish TNFSF sequences, as well as their rapid update as new TNFSF family sequences are sequenced and deposited. NCBI Databases could also be used to produce lists of known TNFSF orthologs.

ESAPs: When released, the ESAPs pipeline will allow researchers to create standardized and high-quality genome annotations and gene predictions, reducing their need to spend time integrating multiple external programs for gene prediction. CGR's common set of tools also makes it easier to reproduce published results.

BLAST with Clusters@NCBI Databases: Researchers could BLAST the TNFSF query set against the new Clusters@NCBI database. This would make it easier to examine other cartilaginous fish sequences available at NCBI, create a gene tree across a wide variety of species, and obtain more sequences for annotating neighboring genes discovered in their synteny analysis. Ultimately, researchers would be able to efficiently focus on clusters containing relevant species and sequences through an intuitive display of taxonomic results.

Similar Genes: Researchers could more deeply explore the evolution and representation of gene families across the tree of life using Similar Genes, which are large NCBI collections of genes related by a combination of calculated orthology and protein architectures.

NIH National Library of Medicine
 Follow us on Twitter @ncbi
 Contact us at cgr@nlin.nih.gov
 Visit us at ncbi.nlm.nih.gov/cgr

Intended Users:	Information Professionals	Researchers
-----------------	---------------------------	-------------

NIH Comparative Genomics Resource (CGR)

NIH Comparative Genomics Resource

Description: The NIH Comparative Genomics Resource (CGR) is a multiyear project intended to maximize the impact of research on eukaryotic (non-bacterial, non-viral organisms such as animals, plants, and fungi) lifeforms and their genomic data. CGR facilitates reliable comparative genomics analyses, including the study of structure, function, evolution, and mapping of eukaryotic genomes. Researchers can compare characteristics of sequenced genomes across different species. Comparative genomics provides insight into evolution and how species change over time, how genes control biological functions, and how gene variants in a single species may contribute to disease. CGR facilitates this through community collaboration and an NCBI Toolkit of interconnected and interoperable data and tools. Its development is led by the National Center for Biotechnology Information.

Popular uses of this product:

Information Professionals	Researchers
<ul style="list-style-type: none"> • Include CGR in subject guides for biology, chemistry, and genetic resources. • Curate FAIR, detailed metadata for genomic research data. • Give feedback on the usability and usefulness of CGR (yellow feedback button). • Promote CGR as a multi-faceted resource for facilitating diverse types of comparative genomics research. 	<ul style="list-style-type: none"> • Download comprehensive genomic data including gene, transcript, protein sequences, and metadata. • Visualize and compare eukaryotic genomes assemblies and annotations. • Use tools to improve the quality of your genome assemblies prior to GenBank submission. • Request NCBI evaluation of your human, mouse, or rat genome assemblies for accuracy, completeness, and correctness. • Share curated data with NCBI to expand and enhance genomic related content. • Give feedback on the usability and usefulness of CGR (yellow feedback button).

Join the CGR community today and revolutionize your research!

Get Involved

Submit your assembled and annotated eukaryotic genomes to contribute to the collection of publicly available data.

Analyze your data with the NCBI Toolkit as part of your comparative genomics workflows.

Identify opportunities to connect your eukaryotic genomics-related data and tools with NCBI resources.


Try out the NCBI Toolkit and let us know what you think.

Provide Feedback

Send an email to cgr@nlin.nih.gov.

Click the yellow Feedback button on the bottom right of our webpages.

Join our mailing list!



bit.ly/Subscribe_CGR

Working to make your eukaryotic comparative genomics research easier!

NIH National Library of Medicine
National Center for Biotechnology Information
ncbi.nlm.nih.gov

The NIH Comparative Genomics Resource
CGR
ncbi.nlm.nih.gov/cgr



ncbi.nlm.nih.gov/cgr

NIH Comparative Genomics Resource (CGR)

Valerie Schneider, Ph.D. 10/26/23

NIH National Library of Medicine
National Center for Biotechnology Information

Outline-2

- Intro to Comparative Genomics
- The Value of Research Organisms
- Problem
- CGR Solution
- CGR Impact – Two use cases
- What's Next



Thank You



NLM-NCBI

Steve Sherry

Terence Murphy

Kim Pruitt

Françoise Thibaud-Nissen

Janet Coleman

Nuala O'Leary

Anatoly Mnev

Sanjida Rangwala

Anne Ketter

Tom Madden

Katya Sukharnikov

Aron Marchler-Bauer

Wratko Hlavina

Rana Morris

Sally Chang

NLM

Patti Brennan

Jodi Nurik

Diane Tuncer

NLM Board of Regents
CGR Working Group

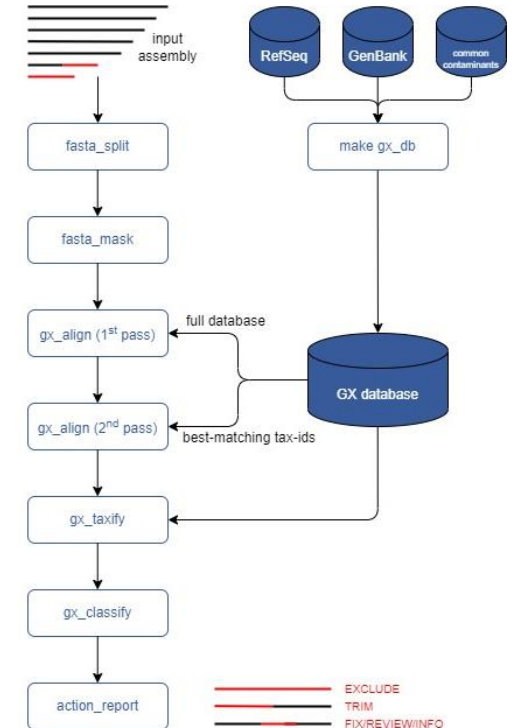
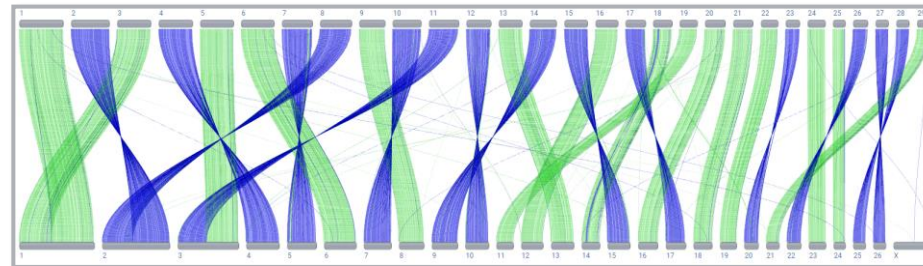
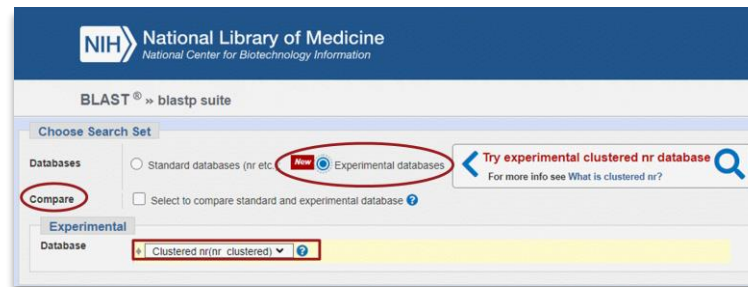
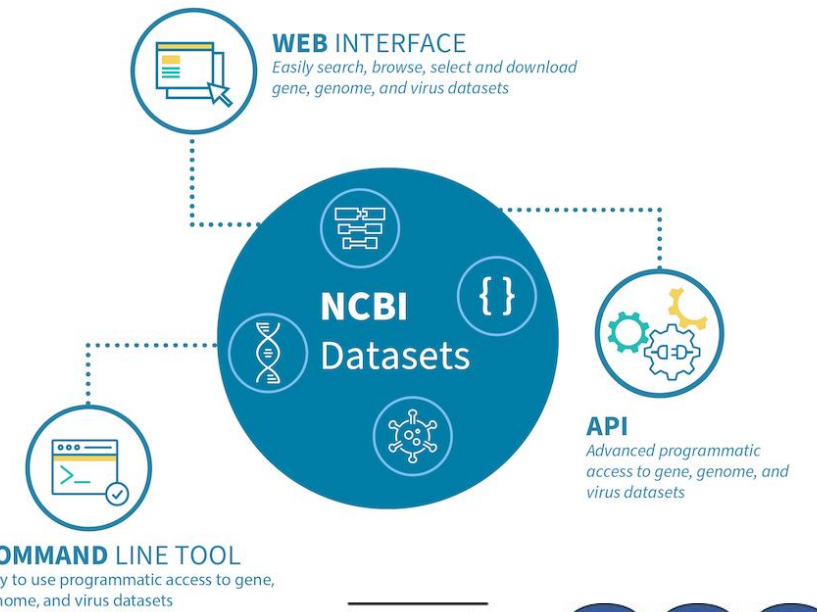
NIH Oversight

NIH CGR Steering Committee

ncbi.nlm.nih.gov/cgr
cgr@nlm.nih.gov

CGR Achievements

- [NCBI Datasets](#)
- [BLAST](#)
- [Comparative Genome Viewer \(CGV\)](#)
- [Foreign Contamination Screen \(FCS\) Tool](#)
- [NCBI Gene](#)
- [Genome Quality Service](#)
- [SPARCLE](#)



How Do I Learn More: Impact Spotlight

- Intro to CGR
- Research Summary
- Potential Impact of CGR to that the research



CGR Impact Spotlight

The publication below is an example of the types of analyses that could benefit from the National Institutes of Health (NIH) Comparative Genomics Resource. CGR facilitates reliable comparative genomics analyses for all eukaryotic organisms through community collaboration and a National Center for Biotechnology Information (NCBI) genomics toolkit.

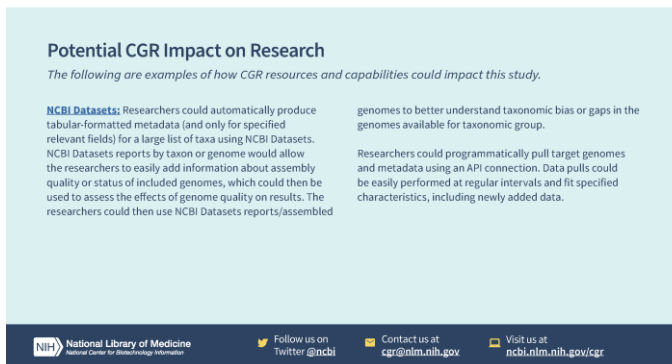


Research Summary

Publication: Bartoszewicz JM, Nasri F, Nowicka M, Renard BY. Detecting DNA of novel fungal pathogens using ResNets and a curated fungi-hosts data collection. *Bioinformatics*. 2022 Sep 16;38(Suppl_2):ii168-ii174. doi: 10.1093/bioinformatics/btac495. PMID: 36124807.

Topic: Bioinformatics resource development

Researchers created a curated database of fungal host-range data linked to publicly available genomes. Using neural networks trained on this data, they tested whether this combination of genomic and host information can be used to predict pathogenicity using both sequence homology and deep-learning approaches. The researchers' database contained over 1400 genomes linked to host and disease phenotype metadata from multiple existing databases and found that their neural networks could accurately detect fungal pathogens in Next Generation Sequencing (NGS) datasets. The trained models predicted pathogenicity and whether a fungus infects humans vs. other hosts. They also developed models with separate classifiers for fungal, viral, and bacterial pathogens.



Potential CGR Impact on Research

The following are examples of how CGR resources and capabilities could impact this study.

NCBI Datasets: Researchers could automatically produce tabular-formatted metadata (and only for specified relevant fields) for a large list of taxa using NCBI Datasets. NCBI Datasets reports by taxon or genome would allow the researchers to easily add information about assembly quality or status of included genomes, which could then be used to assess the effects of genome quality on results. The researchers could then use NCBI Datasets reports/assembled

genomes to better understand taxonomic bias or gaps in the genomes available for taxonomic group.

Researchers could programmatically pull target genomes and metadata using an API connection. Data pulls could be easily performed at regular intervals and fit specified characteristics, including newly added data.

NIH National Library of Medicine
National Center for Biotechnology Information

Follow us on Twitter @ncbi
Contact us at cgr@nlm.nih.gov
Visit us at ncbi.nlm.nih.gov/cgr



CGR Impact Spotlight

The publication below is an example of the types of analyses that could benefit from the National Institutes of Health (NIH) Comparative Genomics Resource. CGR facilitates reliable comparative genomics analyses for all eukaryotic organisms through community collaboration and a National Center for Biotechnology Information (NCBI) genomics toolkit.



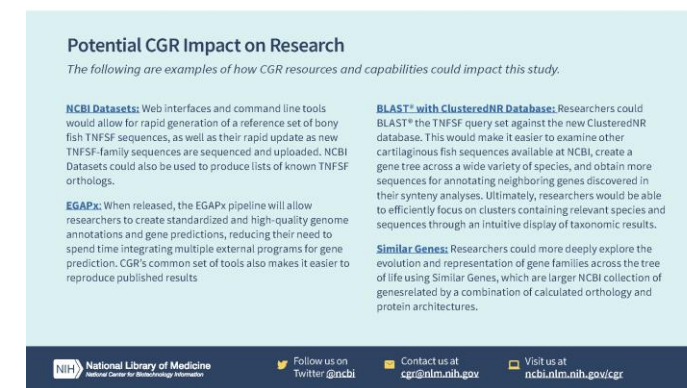
Research Summary

Publication: Redmond AK, Pettinello R, Bakke EK, Dooley H, Shanks. Provide Evidence for a Highly Complex TNFSF Repertoire in the Jawed Vertebrate Ancestor. *J Immunol*. 2022 Nov 1;209(9):1713-1723. doi: 10.4049/jimmunol.2200300. Epub 2022 Sep 16. PMID: 36113883.

Topic: Protein family evolution and orthology inference for a particular gene family

Researchers performed a variety of bioinformatic analyses to survey tumor necrosis factors (TNFs) in five cartilaginous fishes that occupy a critical phylogenetic position outside bony fish and quadrupeds. Notably, with few exceptions, sharks include orthologs of all human TNF Superfamilies (TNFSFs), suggesting sharks could be used as a model for understanding TNFSF evolution.

Researchers discovered that sharks harbor more than 30 TNFSF genes, which is more than all other vertebrates, due to retention of an ancestral repertoire and lineage specific expansions. Cytokines of the TNF superfamily are important for immune function and implicated in many human diseases. Researchers discovered the cartilaginous fish immune system may be less primitive than predicted when compared with mammalian systems.



Potential CGR Impact on Research

The following are examples of how CGR resources and capabilities could impact this study.

NCBI Datasets: Web interfaces and command line tools would allow for rapid generation of a reference set of bony fish TNFSF sequences, as well as their rapid update as new TNFSF family sequences are sequenced and uploaded. NCBI Datasets could also be used to produce lists of known TNFSF orthologs.

EGAPx: When released, the EGAPx pipeline will allow researchers to create standardized and high-quality genome annotations and gene predictions, reducing their need to spend time integrating multiple external programs for gene prediction. CGR's common set of tools also makes it easier to reproduce published results.

BLAST* with ClusteredNR Database: Researchers could BLAST* the TNFSF query set against the new ClusteredNR database. This would make it easier to examine other cartilaginous fish sequences available at NCBI, create a gene tree across a wide variety of species, and obtain more sequences for annotating neighboring genes discovered in their synteny analyses. Ultimately, researchers would be able to efficiently focus on clusters containing relevant species and sequences through an intuitive display of taxonomic results.

Similar Genes: Researchers could more deeply explore the evolution and representation of gene families across the tree of life using Similar Genes, which are larger NCBI collection of genes generated by a combination of calculated orthology and protein architectures.

NIH National Library of Medicine
National Center for Biotechnology Information

Follow us on Twitter @ncbi
Contact us at cgr@nlm.nih.gov
Visit us at ncbi.nlm.nih.gov/cgr