

GMOD User Interface

Kim Pruitt
NCBI



Focus Topics

- How do scientists find named genes and view gene reports?
- How can scientists find information by starting with functional concepts?
- How can scientists retrieve attributes of interest from large datasets (custom reports)?
- How can scientists answer questions about large numbers of genes?
- How can disperse data be integrated to facilitate access and expand the discovery space?



NCBI - organizing principles

- Data is organized by type into different databases
- Databases are cross-linked
- Interface to search across all databases

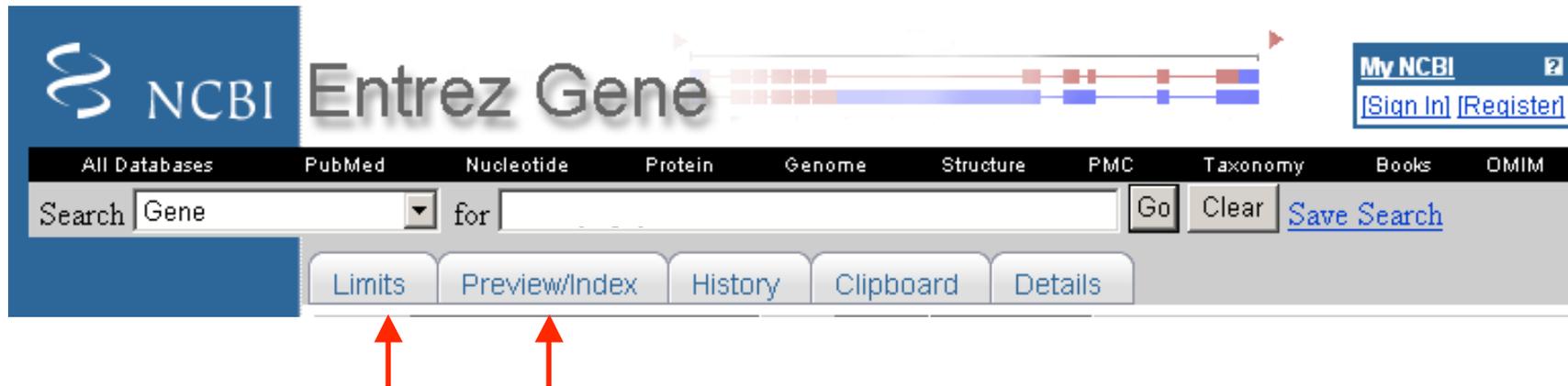


Focus Topics

- How do scientists find named genes and view gene reports?
- How can scientists find information by starting with functional concepts?



Query Gene



- flexible query support
- Query for term (gene symbol; phenotype)
 - Unrestricted (search full record for any match)
 - Restrict the query (search only symbols)



Gene: flexible query support

NCBI Entrez Gene

My NCBI [Sign In] [Register]

All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books OMIM

Search Gene

Query: homeobox AND drosophila[organism]
bsh[Gene Name]
phenotype AND drosophila[organism]
GeneOntology AND drosophila[organism]

Use preview/index tab to explore different options

1: [bsh](#) Links
bsh, isoform B
Chromosome: 2L; **Location:** 38A3-38A3
GeneID: 35266

2: [OdsH](#) Links
Ods-site homeobox [*Drosophila melanogaster*]
Other Aliases: Dmel_CG6352, CG6352, OdsH[mel]
Other Designations: Ods-site homeobox CG6352-PA
Chromosome: X; **Location:** 16D1-16D3
GeneID: 32758

3: [Rx](#) Links
Retinal Homeobox [*Drosophila melanogaster*]
Chromosome: 1; **Location:** 11A1-11A2
GeneID: 32758

Entrez Gene
Home
About
FAQ
Help
Gene Handbook
Statistics
Downloads (FTP)

Mailing Lists
Gene
RefSeq

Feedback
Help Desk
Corrections
About GeneRIFs

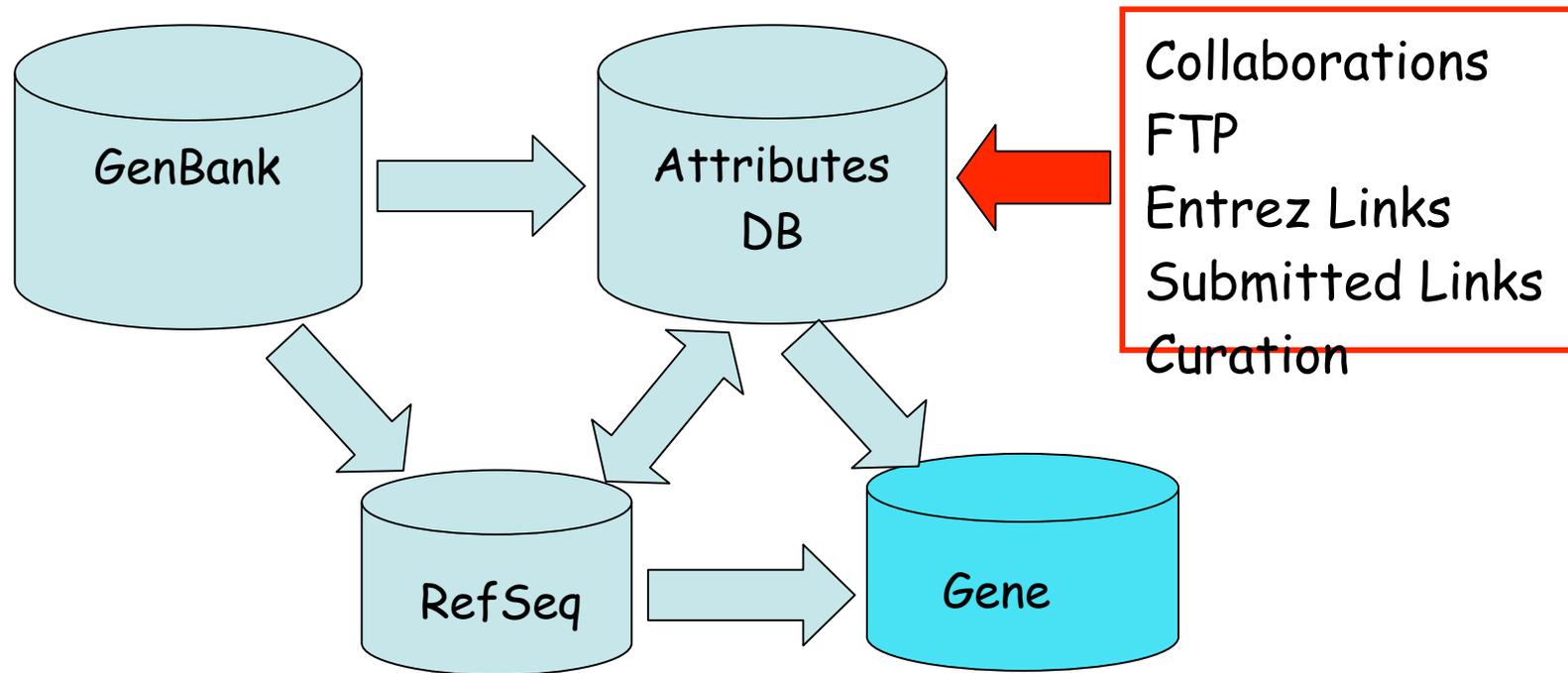
Related Sites
BLAST
Entrez Genome

Entrez Gene

- NCBI's Gene database is the most heavily used db after PubMed (more than nucleotide)
- Gene provides more links to other NCBI databases than any other NCBI resource
- Gene provides more links to external resources than any other NCBI resource



Overview of Gene Data Flow



Entrez Gene



Species: 3,841
Genes: 2,396,221

Content includes:

- Sequences (RefSeq & GenBank)
- Publications & GeneRIFs (References Into Function)
- GO terms
- Interactions & Pathways
- Nomenclature (Symbols & Names)
- Map location & Markers
- Disease & Phenotype names
- Links (related NCBI pages & international web sites)



Gene <-> GMOD

- Genome annotation
 - Annotated reference sequence
 - Gene-2-sequence associations
- Gene attributes
 - Symbols & names; aliases
 - Chromosome or genetic map data
 - Phenotype
 - Publications

FlyBase
HGNC
MGI
RGD
SGD
TAIR
WormBase



Focus Topics

- How can scientists retrieve attributes of interest from large datasets (custom reports)?



Custom reports

- NCBI doesn't currently have a web interface that supports reporting attributes of interest from Gene, or from more than one DB.
- Save data of interest sequentially and combine it manually or via scripting:
 - Full gene reports
 - Map Viewer “Data as Table view” & Download
 - Download sequence from Map Viewer, from nucleotide, from protein
- Use scripting utilities to query and fetch. Use scripts to build custom report.



Annotation data? Gene Table display

- Intron/exon delineation, per annotated variant
- Use 'Send to' option to save as text/file

```
mRNA          bp  exons Protein          aa  exons
NM_001044645.1 596 4      NP_001038110.1 123 4
```

Exon information:
[NM_001044645.1](#) length: 596 bp, number of exons: 4
[NP_001038110.1](#) length: 123 aa, number of exons: 4

EXON		Coding EXON		INTRON	
coords	length	coords	length	coords	length
1 - 100	100 bp	21 - 100	80 bp	101 - 284	184 bp
285 - 461	177 bp	285 - 461	177 bp	462 - 1618	1157 bp
1619 - 1696	78 bp	1619 - 1696	78 bp	1697 - 3119	1423 bp
3120 - 3360	241 bp	3120 - 3154	35 bp		

Display Show Send to



Or - Map Viewer – Data as Table view

[Gallus gallus \(chicken\) Build 2.1](#)

[BLAST Chicken Sequences](#)

Data As Table View

[Download All](#) ¹

RefSeq Transcripts On Sequence

[All Sequence Maps](#)

[next](#)

Region Displayed: **870,500-878,900 bp**

[Download/View Sequence](#)

[Download Data](#)

Total RefSeq Transcripts On Chromosome: **213** [[7 not localized](#)]

RefSeq Transcripts in Region: **3**

start	stop	Accession	Locus	O	Links	Align quality	Description
870999	872124	XM_418185.2	NDUFA7	+	sv pr ev BLink mm	identical	NADH dehydrogenase (ubiquinone)
870999	871009	exon	+	UTR		11 bp	
871010	871060	exon	+	CDS		51 bp	
871061	871168	intron				108 bp	
871169	871218	exon	+	CDS		50 bp	
871219	871328	intron				110 bp	
871329	871478	exon	+	CDS		150 bp	
871479	871778	intron				300 bp	
871779	871875	exon	+	CDS		97 bp	
871876	872124	exon	+	UTR		249 bp	
872990	876349	NM_001044645.1	TVA	+	sv pr ev BLink mm	poor	Tva receptor
872990	873009	exon	+	UTR		20 bp	
873010	873089	exon	+	CDS		80 bp	
873090	873273	intron				184 bp	
873274	873450	exon	+	CDS		177 bp	
873451	874607	intron				1157 bp	
874608	874685	exon	+	CDS		78 bp	



Focus Topics

- How can scientists answer (functional) questions about large numbers of genes?



Via the web site

1. define a Gene query to select the gene set of interest

- As you define the query, use the History tab to retrieve a previous query, or to combine queries
- Save final query definition in MyNCBI to use again later

2. follow calculated links to approach functional questions:

- What conserved domains are found in a gene family?
- What expression data is available in GEO Profiles?
- What variation data is in dbSNP?



Pre-calculated links

The screenshot displays the NCBI Entrez Gene interface. A dropdown menu is open, listing various pre-calculated links for a selected gene. The menu items include:

- Full Report
- Summary
- Brief
- ASN.1
- XML
- Gene Table
- UI List
- LinkOut
- Books Links
- Conserved Domain Links
- Genome Links
- GENSAT Links
- GEO Profile Links
- HomoloGene Links
- Nucleotide Links
- NIH cDNA clone links
- OMIA Links
- OMIM Links
- BioAssay Links
- PMC Links
- Probe Links
- Protein Links
- PubMed Links
- PubMed (GeneRIF) Links
- SNP Links
- Gene Genotype Links
- Taxonomy Links
- UniGene Links
- UniSTS Links

The background interface shows the gene page for *Drosophila melanogaster*. The search bar contains "Gene" and the search results show the gene name and various links. The page is titled "One page." and includes a "Links" section.



Via programming

- NCBI e-utilities: run Entrez queries (eSearch) and retrieve results (eFetch, eLink) from your own scripts



Entrez Programming Utilities

Updated: August 10, 2006

Entrez Programming Utilities are tools that provide access to Entrez data outside of the regular web query interface and may be helpful for retrieving search results for future use in another environment.

Additional information is available in the NCBI Bookshelf Short Courses [Building Customized Data Pipelines Using the Entrez Programming Utilities \(eUtils\)](#) and the NCBI [PowerScripting](#) course.

- [User Requirements](#): Please read for important information on scripting NCBI servers.
- [EInfo](#): Provides field index term counts, last update, and available links for each database.
- [ESearch](#): Searches and retrieves [primary IDs](#) (for use in EFetch, ELink, and ESummary) and term translations and optionally retains results for future use in the user's environment.
- [EPost](#): Posts a file containing a list of [primary IDs](#) for future use in the user's environment to use with subsequent search strategies.
- [ESummary](#): Retrieves document summaries from a list of [primary IDs](#) or from the user's environment.
- [EFetch](#): Retrieves records in the requested format from a list of one or more [primary IDs](#) or from the user's environment.
- [ELink](#): Checks for the existence of an external or Related Articles link from a list of one or more [primary IDs](#). Retrieves primary IDs and relevancy scores for links to Entrez databases or Related Articles; creates a hyperlink to the primary LinkOut provider for a specific ID and database, or lists LinkOut URLs and Attributes for multiple IDs.
- [EGQuery](#): Provides Entrez database counts in XML for a single search using [Global Query](#).
- [ESpell](#): Retrieves spelling suggestions.



Focus Topics

- Many scientists work with several genome websites to answer a research question. How does NCBI facilitate this?
1. Gene integrates information from other NCBI databases and from external resources to centralize information. Gene provides:
 - cross-links between NCBI databases
 - links to external data source
 2. LinkOut: databases can submit LinkOuts to connect their report pages to any Entrez database. Gene reports available LinkOuts.



Extensive data integration and navigation support!

All: 1 Genes Genomes: 1 SNP GeneView: 1

1: CALM2 calmodulin 2 (phosphorylase kinase, delta) [*Gallus gallus*]
 GeneID: 395855 updated 10-Jan-2006

Summary

Gene type: protein coding
 Gene description: calmodulin 2 (p
 RefSeq status: Provisional
 Organism: *Gallus gallus*
 Lineage: Eukaryota; Metazoa; C
 Galliformes; Phasianidae; Phasia
 Gene aliases: CALM2

Genomic regions, transcript
 (minus strand) RefSeq below

Genomic context
 chromosome: 3

Bibliography
 PubMed links
 GeneRIFs:
 1. The presence of multiple conform
 heterogeneity of structure is at least
 targets.

Interactions
 Description
 CALM2 Product
 CaM interacts with myosin V. This
 myosin V.
 NP_990336.1

General gene information

GeneOntology
 Provided by GOA
 Function
 calcium ion binding IEA

General protein information
 Name: calmodulin 2 (phosphoryl

NCBI Reference Sequence
 mRNA Sequence NM_205005
 Source Sequence AF081672
 Product NP_990336 ca
 Conserved Domai
 cd00051: EFh
 Location:
 Location:
 Location:

Related Sequences
 Nucleotide
 mRNA AF081672 AAC3
 mRNA M36167 AAA4
 None O934

Additional Links
 UniGene Gga.4454

All: 1 Genes Genomes: 1 SNP GeneView: 1

1: CALM2 calmodulin 2 (phosphorylase kinase, delta) [*Gallus gallus*]
 GeneID: 395855 updated 10-Jan-2006

Summary

Gene type: protein coding
 Gene description: calmodulin 2 (p
 RefSeq status: Provisional
 Organism: *Gallus gallus*
 Lineage: Eukaryota; Metazoa; C
 Galliformes; Phasianidae; Phasia
 Gene aliases: CALM2

Genomic regions, transcript
 (minus strand) RefSeq below

Genomic context
 chromosome: 3

Bibliography
 PubMed links
 GeneRIFs:
 1. The presence of multiple conform
 heterogeneity of structure is at least
 targets.

Interactions
 Description
 CALM2 Product
 CaM interacts with myosin V. This
 myosin V.
 NP_990336.1

General gene information

GeneOntology
 Provided by GOA
 Function
 calcium ion binding IEA

General protein information
 Name: calmodulin 2 (phosphoryl

NCBI Reference Sequence
 mRNA Sequence NM_205005
 Source Sequence AF081672
 Product NP_990336 ca
 Conserved Domai
 cd00051: EFh
 Location:
 Location:
 Location:

Related Sequences
 Nucleotide
 mRNA AF081672 AAC3
 mRNA M36167 AAA4
 None O934

Additional Links
 UniGene Gga.4454

Genomic regions, transcripts, and products

(minus strand) RefSeq below

NC_006090

23472341] 5' |-----| 3' 23460776]

NM_205005 NP_990336

Bibliography

Gene References into Function (GeneRIF): [Submit](#)

PubMed links

GeneRIFs:

1. The presence of multiple conformations is a physical property of calmodulin (CaM), and it is likely that the heterogeneity of structure is at least partially responsible for the promiscuous ability of CaM to recognize diverse targets. [PubMed](#)

NCBI Reference Sequences (RefSeq)

mRNA Sequence [NM_205005](#)
 Source Sequence [AF081672](#)
 Product [NP_990336](#) calmodulin 2 (phosphorylase kinase, delta)
 Conserved Domains (3) [summary](#)
[cd00051: EFh; EF-hand, calcium binding motif](#)
 Location: 85 - 147 Blast Score: 175
 Location: 12 - 74 Blast Score: 170
 Location: 48 - 111 Blast Score: 112

Related Sequences

Nucleotide	Protein
mRNA AF081672	AAC31608
mRNA M36167	AAA48650
None	O93410

Additional Links

UniGene [Gga.4454](#)

Complex	Source	Pubs
Interaction between bovine CaM and chicken	BIND	PubMed

Evidence Viewer

KEGG
 ModelMaker
 UCSC
 UniGene
 LinkOut

[+](#) Entrez Gene Info
[+](#) Feedback
[+](#) Subscriptions

Map Viewer

[OMIM](#)
[Probe](#)
[RefSeq](#)
[UniGene](#)
[UniSTS](#)

[Feedback](#)

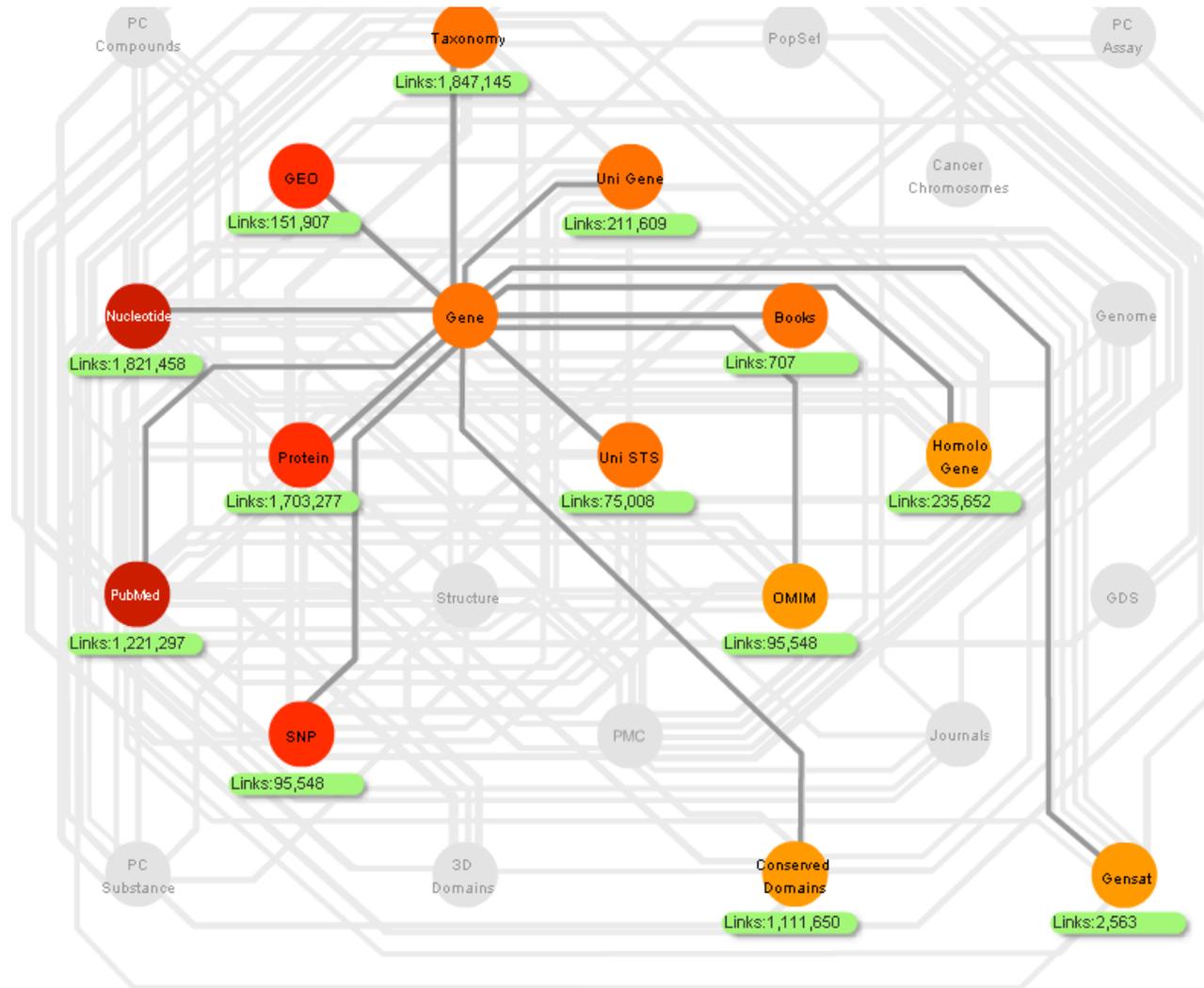
Contact Help Desk
[Submit Correction](#)
[Submit GeneRIF](#)

[Subscriptions](#)

[RefSeq](#)
[Gene](#)
[Map Viewer](#)



Gene links – increasing discovery space



- Links between databases and between records within databases are re-computed daily

- Users can navigate between records based on these links



10,000,000



1,000,000



100,000



10,000



1000

- About interface design
- New design area



Interface design – ask the users early

- What information do you want
- Wireframe design
- User interview presenting design (paper) with scripted questions
 - what information do you see reported;
 - where is information about ‘n’;
 - what information do you get if you follow the link;
- Refine design, re-survey (new) user group





An Integrated Sequence Analysis Application

- Comparative Genome Analysis
- Data import, BLAST analysis, annotation
- Phylogenetic clustering
- Graphical synteny analysis
- Ability to save projects
- Write your own local plug-ins
- Much more!

<http://www.ncbi.nlm.nih.gov/projects/gbench/>



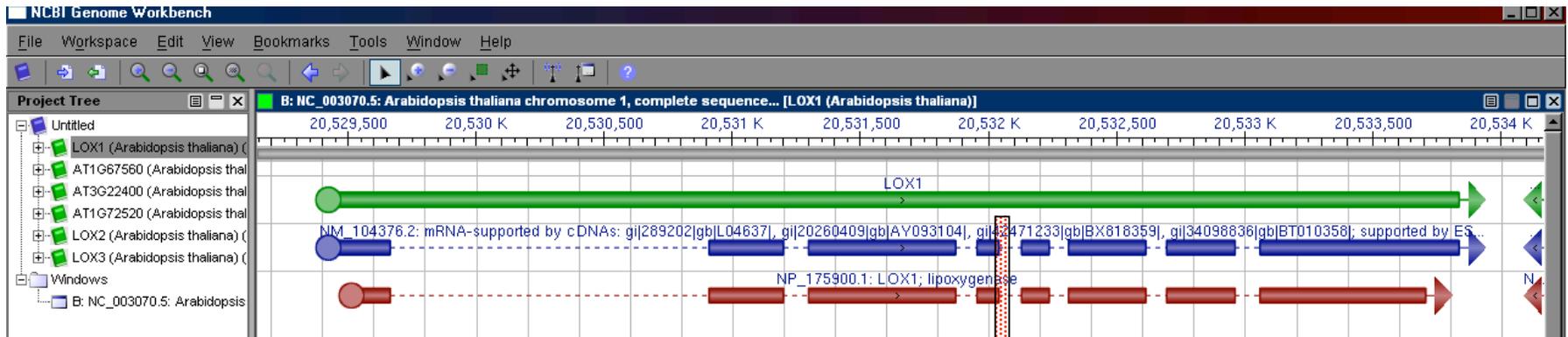


Introduction www.ncbi.nlm.nih.gov/projects/gbench/

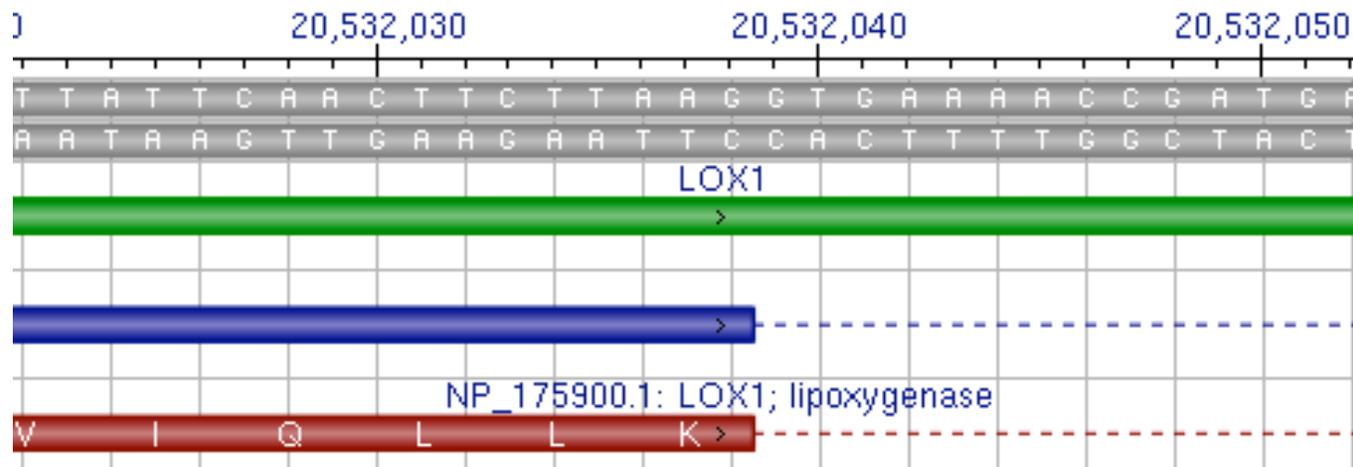
- A robust sequence analysis application
- a stand-alone program
- Windows, Mac, and UNIX-based formats
- Integrates NCBI tools in one interface
 - Entrez queries, mRNA to genomic alignments, BLAST results, multiple sequence alignments, cross-alignments and dot-matrix genome/chromosome comparisons
- Fully scalable (genome to individual nucleotides)
- Import your own data for analysis



At LOX1 - from chromosome to base pair



Gene Level

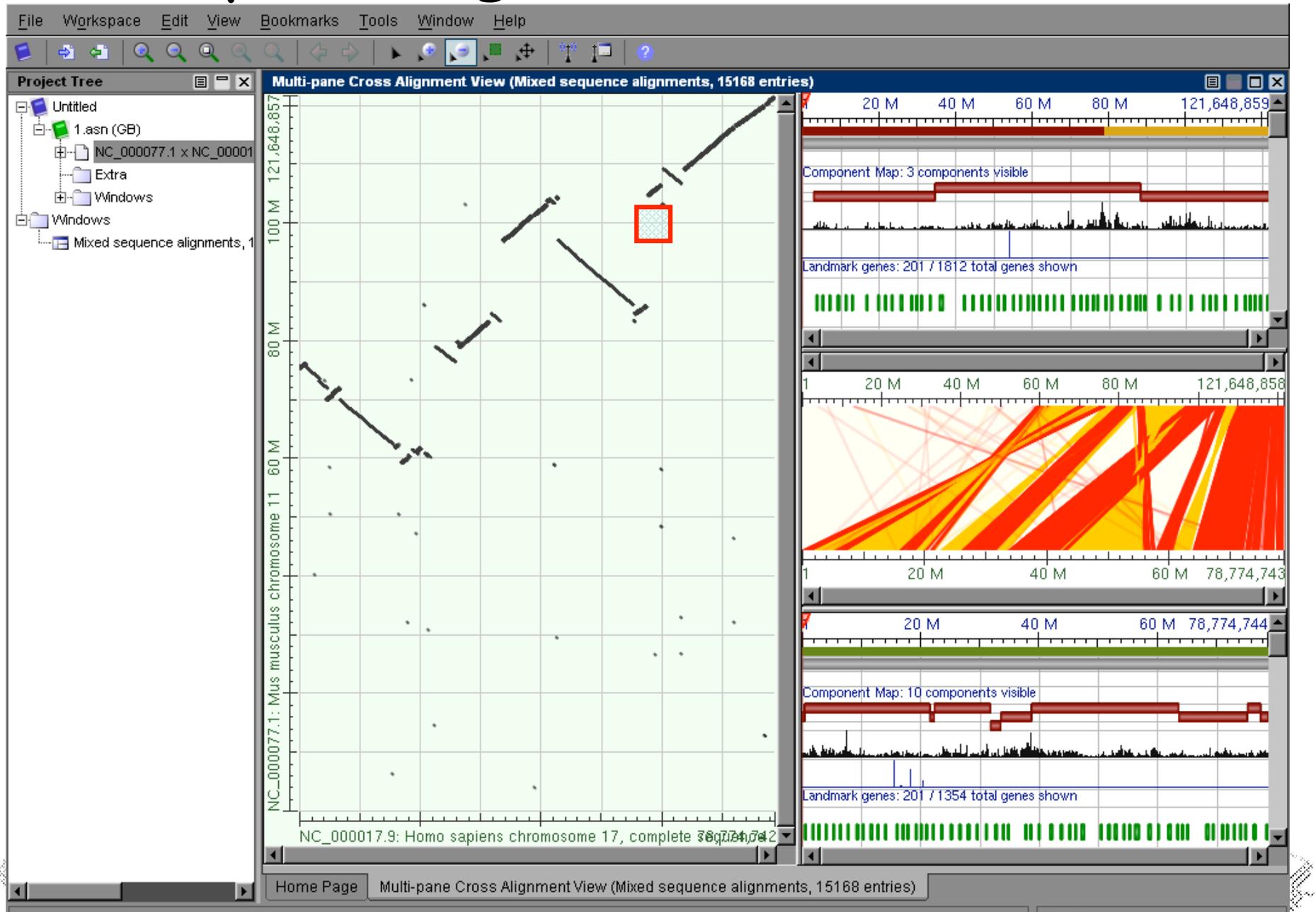


Nucleotide/peptide Level

National Center for Biotechnology Information



Multi-panel alignment view



Multiple alignment view

The screenshot displays the NCBI Genome Workbench interface. The main window shows a multiple sequence alignment of 16 entries. The alignment is presented in a grid format with a scale bar at the top indicating positions from 750 to 870. The sequences are color-coded to show conserved regions. The Project Tree on the left shows the source of the sequences, including LOX1, LOX2, LOX3, and NP_188879.2 (GB).

Description	Mark...	Alignment	Seq...
NP_188879.2	<input type="checkbox"/>	IETCTIIIIWIASALHAAVNFGQYPYAGFLPNRPTVSRRFMP	886
BAD95111.1	<input type="checkbox"/>	SQILTNIWIASGQHAALNFGQYPFGGYVFNRPPLMRRLIP	335
BAD94917.1	<input type="checkbox"/>	IGVVTIIAWVTSCHHAAVNFGQYGYGGYFNPRTTTRIRMPTE	443
AAL32689.1	<input type="checkbox"/>	IGVVTIIAWVTSCHHAAVNFGQYGYGGYFNPRTTTRIRMPTE	
AAG52309.1	<input type="checkbox"/>	SQILTNIWIASGQHAALNFGQYPFGGYVFNRPPLMRRLIP	
AAG51846.1	<input type="checkbox"/>	VSVITTIIWLASAQHAALNFGQYPYGGYVFNRPPLMRRLIP	
CAC19365.1	<input type="checkbox"/>	IETCTIIIIWIASALHAAVNFGQYPYAGFLPNRPTVSRRFMP	
CAC19364.1	<input type="checkbox"/>	VSVITTIIWLASAQHAALNFGQYPYGGYVFNRPPLMRRLIP	
AAG00881.1	<input type="checkbox"/>	VSVITTIIWLASAQHAALNFGQYPYGGYVFNRPPLMRRLIP	
AAF97315.1	<input type="checkbox"/>	VSVITTIIWLASAQHAALNFGQYPYGGYVFNRPPLMRRLIP	
BAB01777.1	<input type="checkbox"/>	IETCTIIIIWIASALHAAVNFGQYPYAGFLPNRPTVSRRFMP	
AAF79461.1	<input type="checkbox"/>	VSVITTIIWLASAQHAALNFGQYPYGGYVFNRPPLMRRLIP	
CAB72152.1	<input type="checkbox"/>	IGVVTIIAWVTSCHHAAVNFGQYGYGGYFNPRTTTRIRMPTE	
CAB56692.1	<input type="checkbox"/>	VSVITTIIWLASAQHAALNFGQYPYGGYVFNRPPLMRRLIP	
AAA32749.1	<input type="checkbox"/>	IGVVTIIAWVTSCHHAAVNFGQYGYGGYFNPRTTTRIRMPTE	
AAA32827.1	<input type="checkbox"/>	VESCTIIIIWIASALHAAVNFGQYPYAGFLPNRPTISRQYMP	



Acknowledgements

Entrez Gene:
Donna Maglott

Web Design:
Mark Johnson

Genome Workbench:
Mike diCuccio

NCBI developers & curators

