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

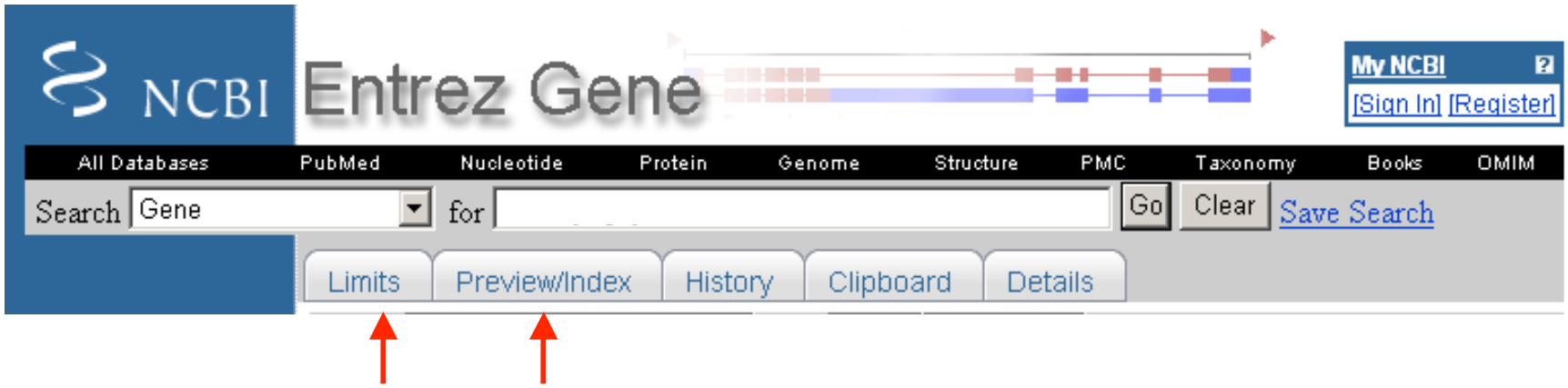
- Data is organized by type into different databases
- Databases are cross-linked
- Interface to search across all databases

Focus Topics

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

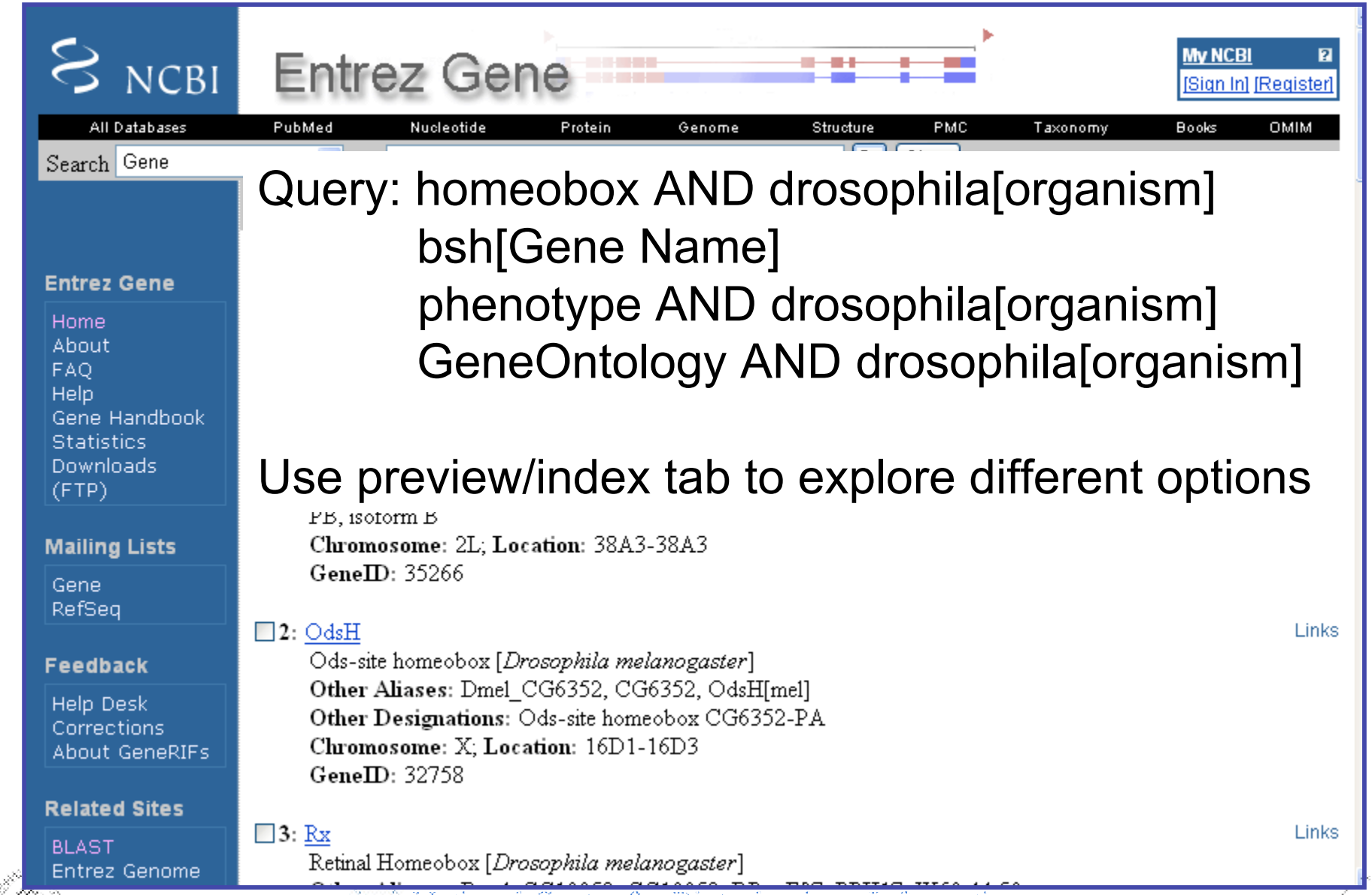
Find Annotated Genes Using Global Query

Query Gene



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

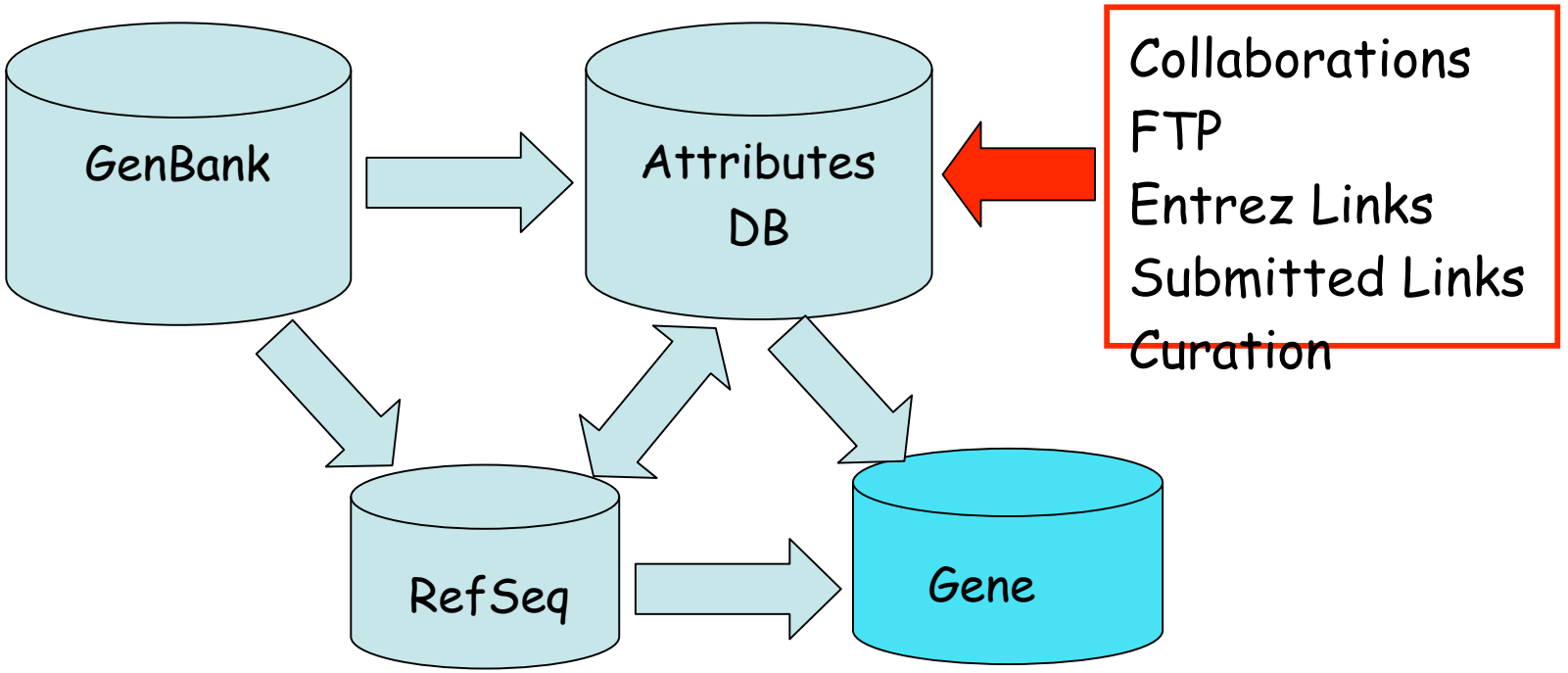
Gene: flexible query support



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 Gene Table Show 5 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

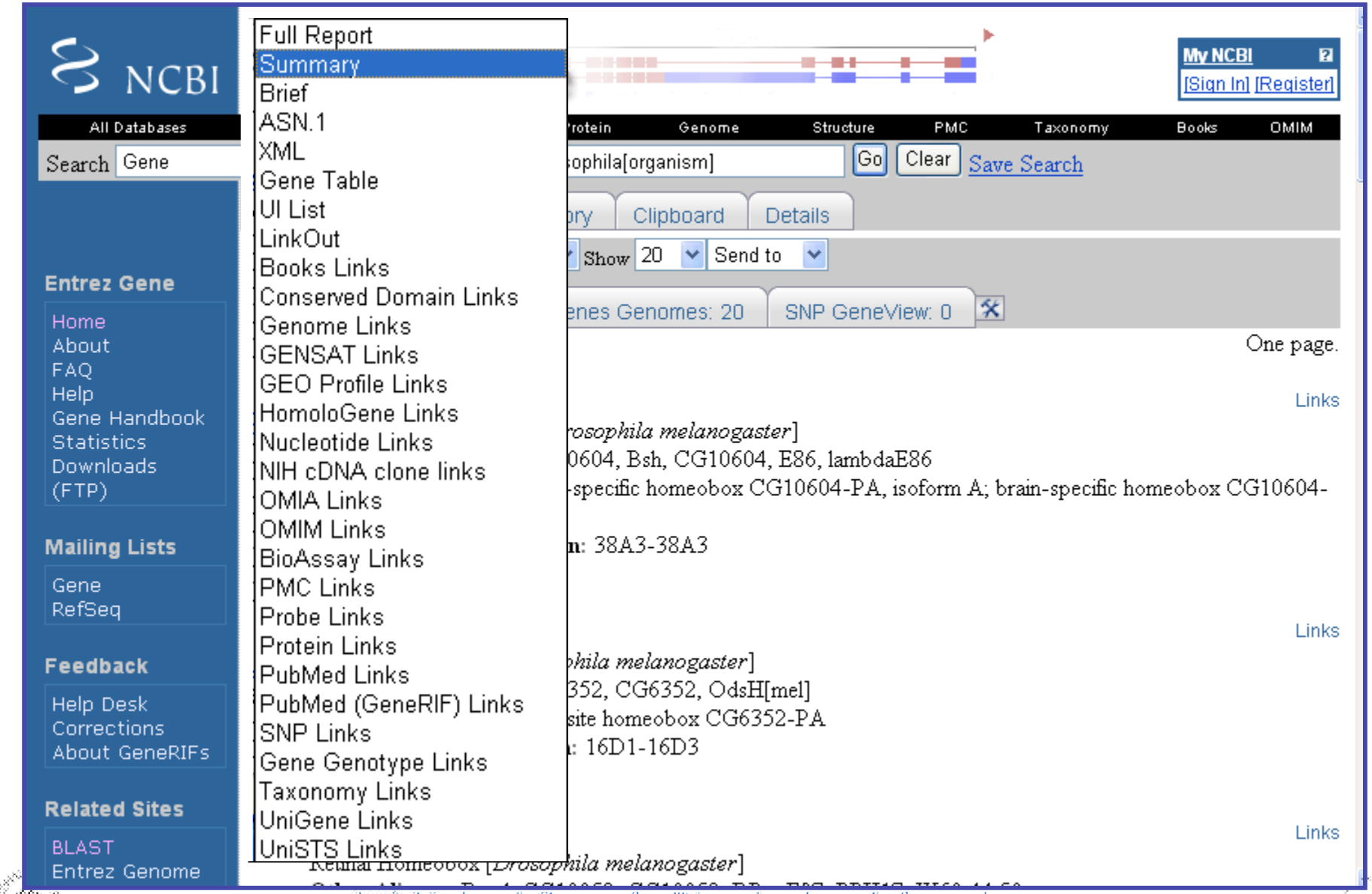
- How can scientists answer (functional) questions about large numbers of genes?

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



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.

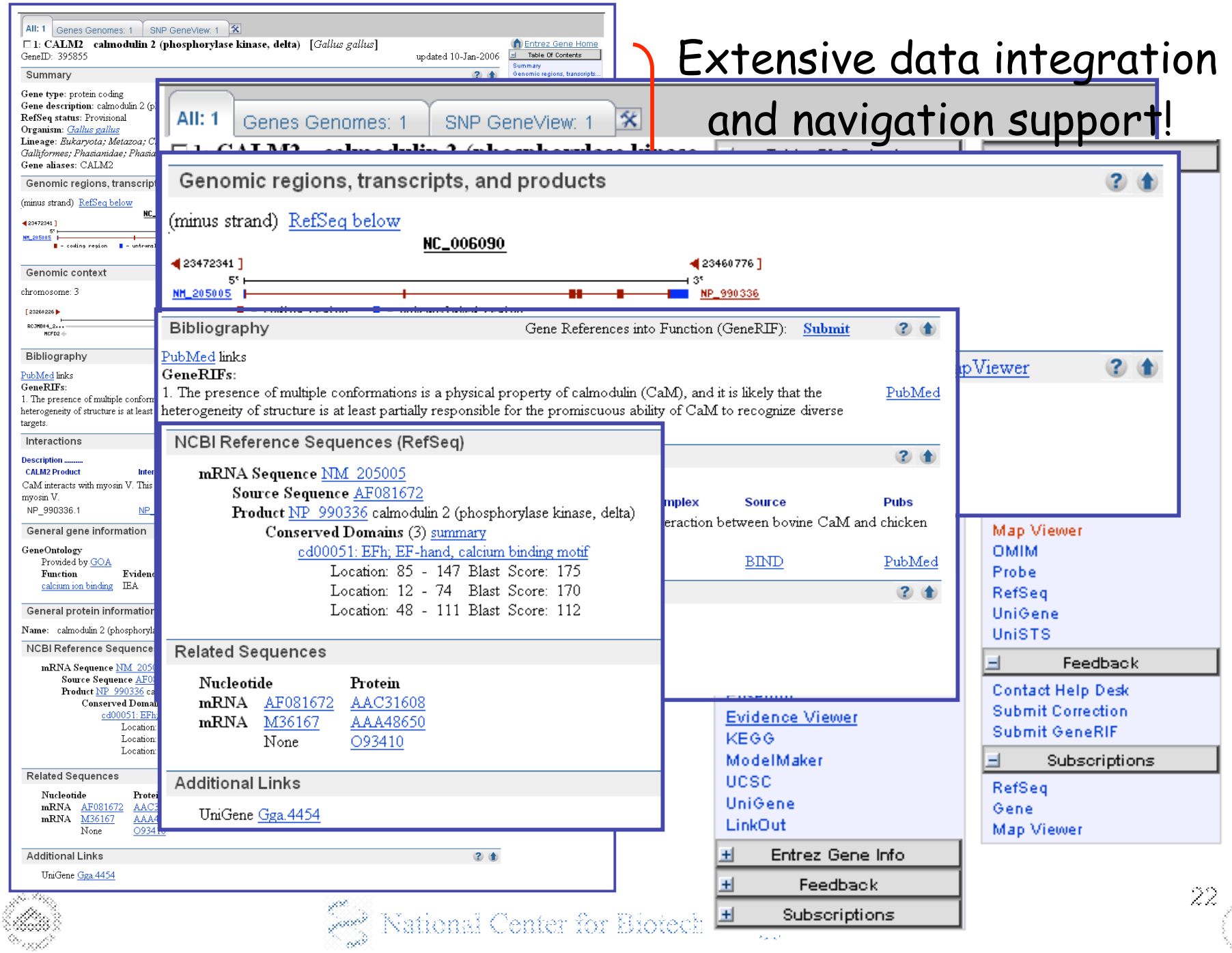
Focus Topics

- Many scientists work with several genome websites to answer a research question. How does NCBI facilitate this?

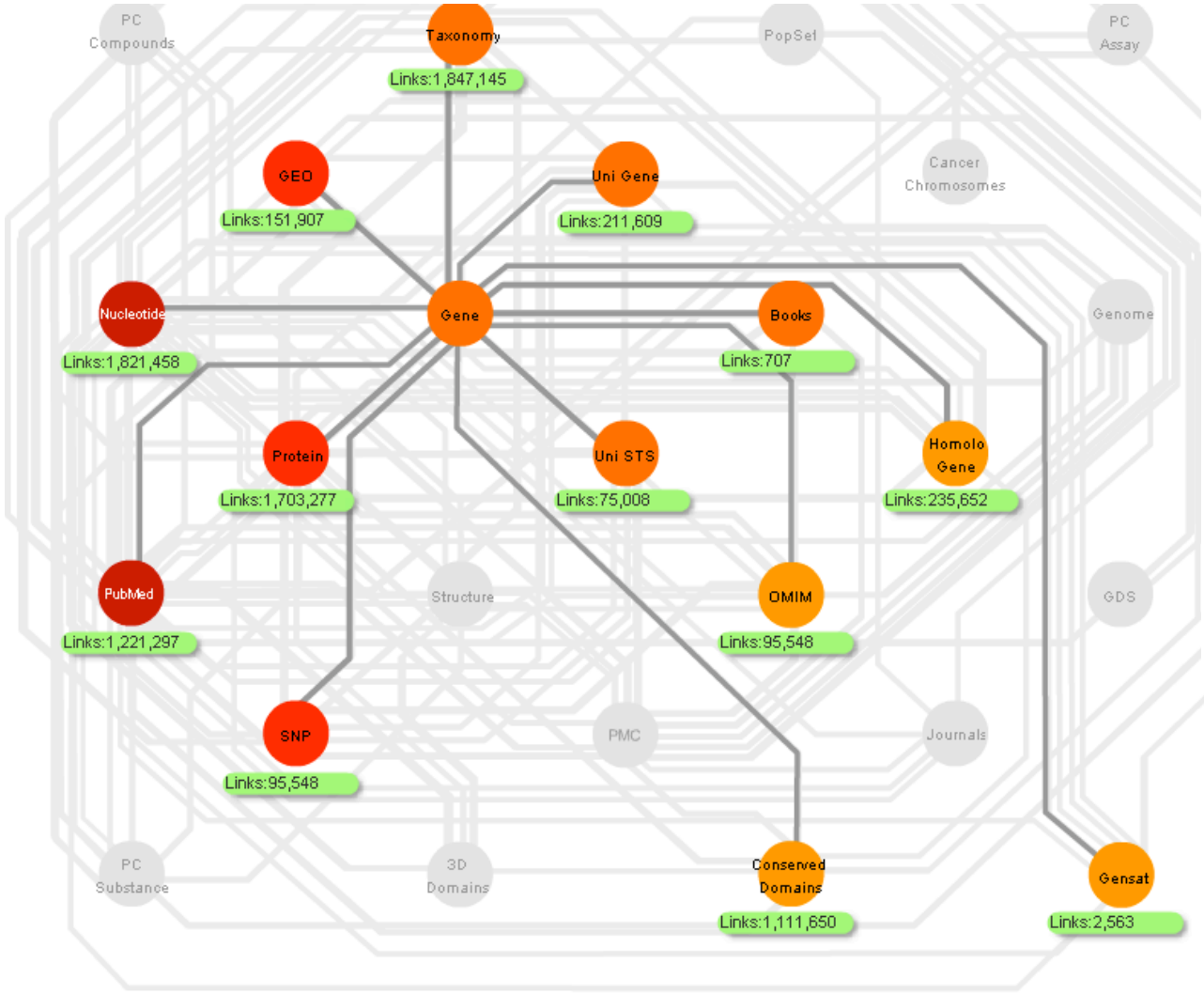
1. Gene integrates information from other NCBI databases and from external resources to centralize information. Gene provides:
   - cross-links between NCBI databases
   - links to external data source

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!



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

- 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

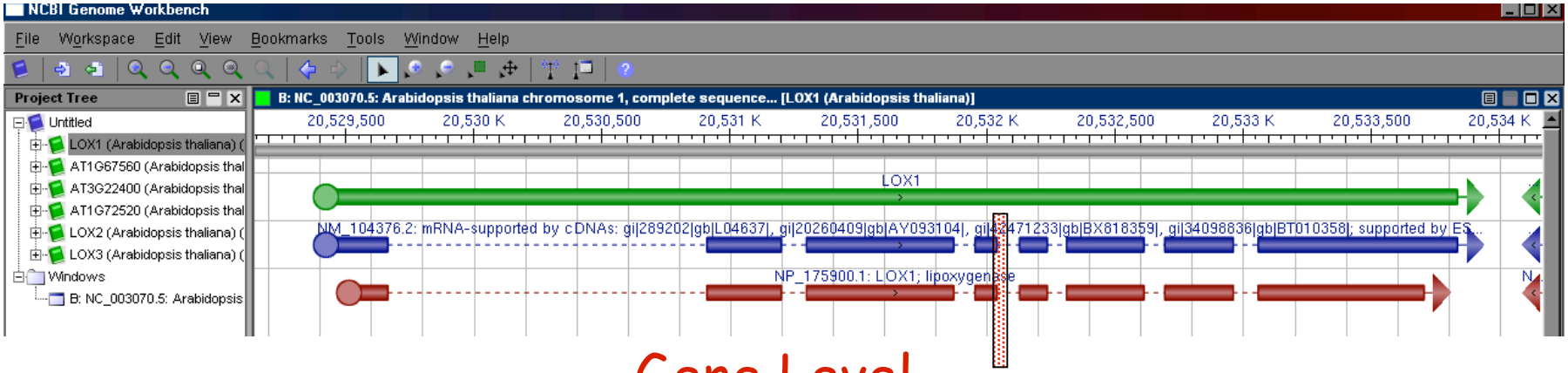
- Comparative Genome Analysis
- Data import, BLAST analysis, annotation
- Phylogenetic clustering
- Graphical synteny analysis
- Ability to save projects
- Write your own local plug-ins
- Much more!

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

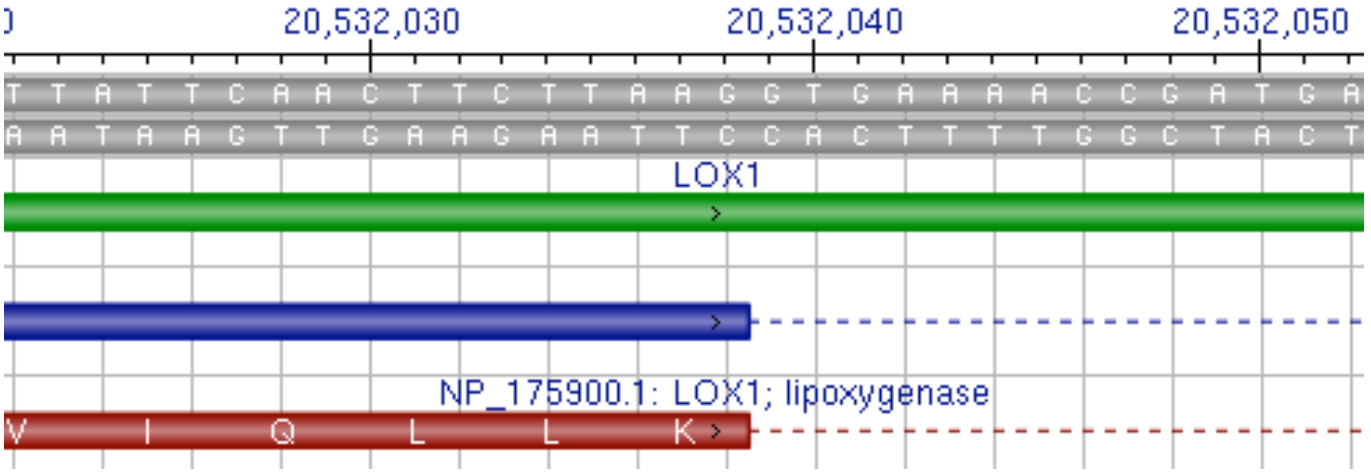
Introduction www.ncbi.nlm.nih.gov/projects/gbench/

- 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

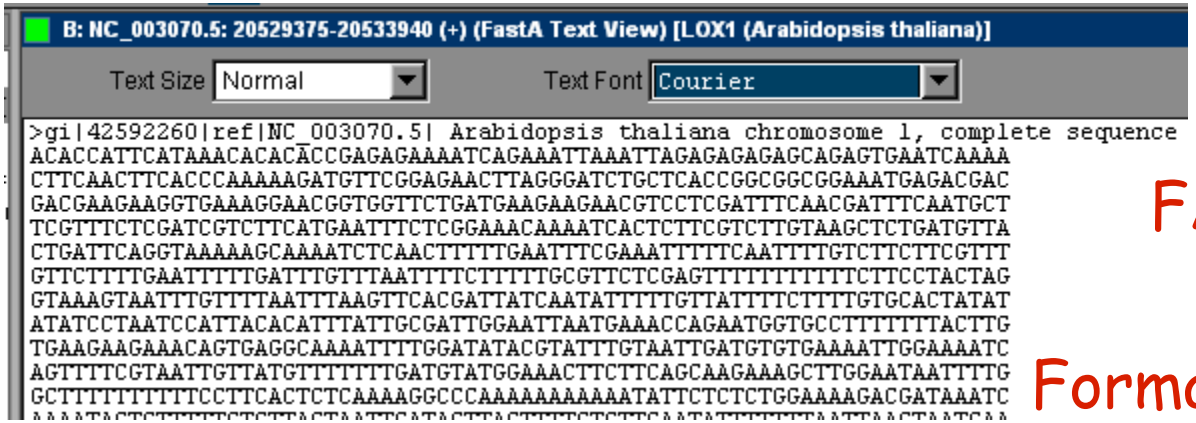


Gene Level



Nucleotide/peptide Level

Other views

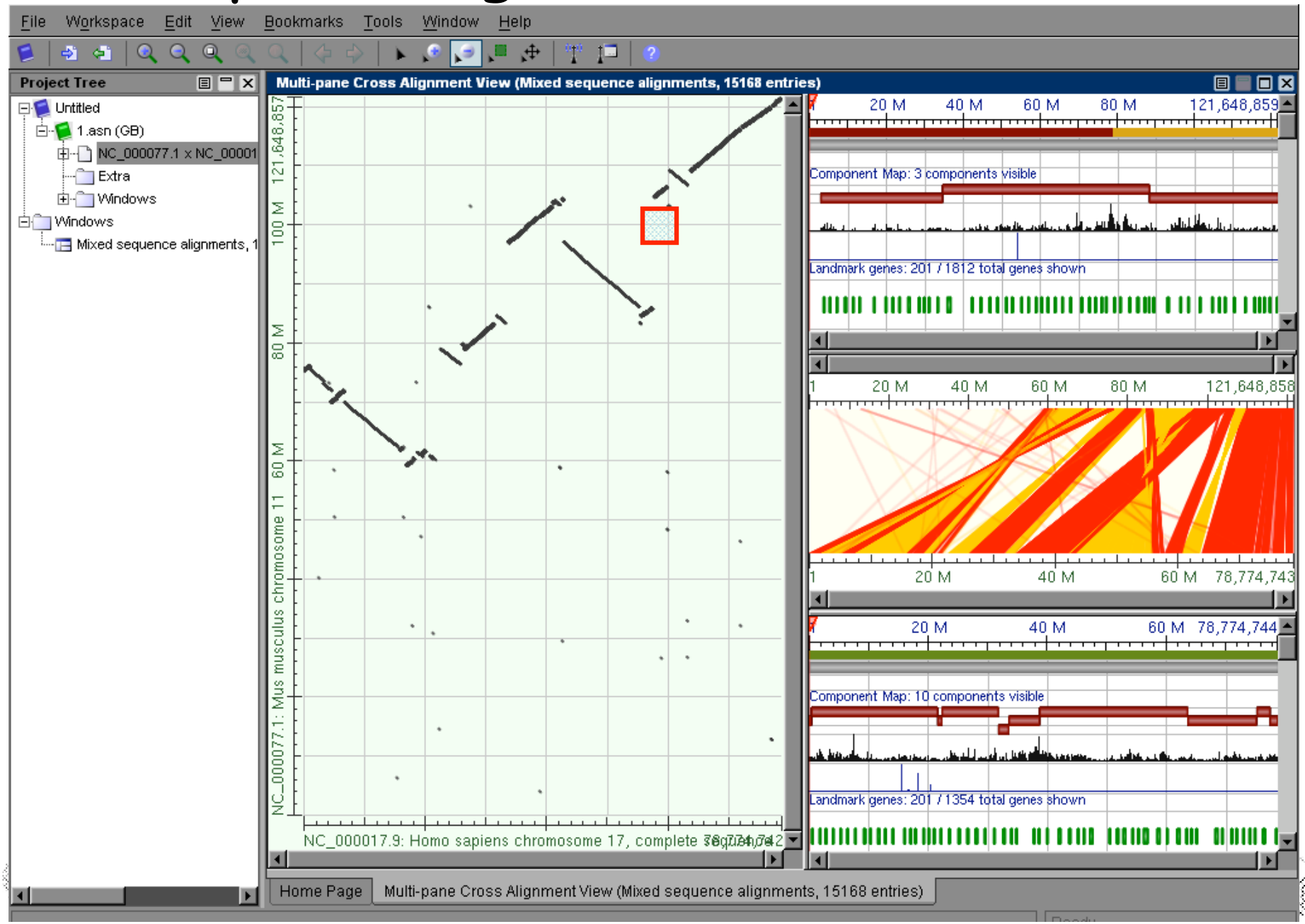


FASTA sequence
Or
Format to highlight CDS, splice sites

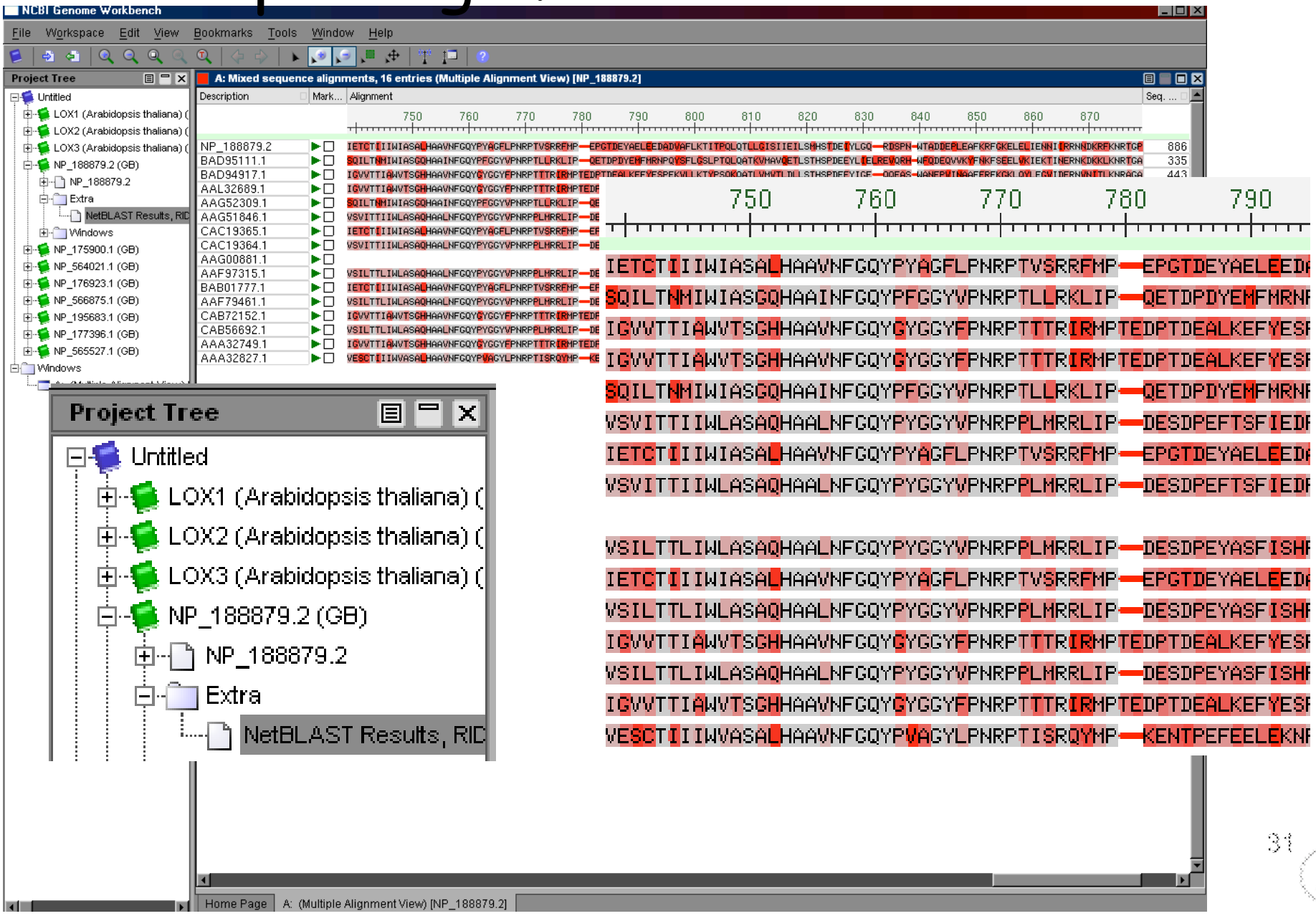
Label	Type	Start	Stop	Length	Strand	Product	Intervals
NM_104374.2	mRNA	205183...	205201...	1835	+	NM_10...	2
NM_001036115.1	mRNA	205183...	205201...	1860	+	NM_00...	3
NM_001036116.1	mRNA	205183...	205201...	1813	+	NM_00...	3
NP_564673.1	CDS	205184...	205198...	1405	+	NP_56...	2
NP_001031192.1	CDS	205184...	205198...	1405	+	NP_00...	2
NP_001031193.1	CDS	205184...	205198...	1405	+	NP_00...	2
PDF1.5	Gene	205207...	205210...	326	+		1
NM_104375.1	mRNA	205207...	205210...	326	+	NM_10...	2
NP_175899.1	CDS	205207...	205210...	326	+	NP_17...	2
LOX1	Gene	205293...	205339...	4566	+		1
NM_104376.2	mRNA	205293...	205339...	4566	+	NM_10...	8
NP_175900.1	CDS	205294...	205338...	4346	+	NP_17...	8
AT1G55030	Gene	205340...	205354...	1352	-		1

Feature table (chromosome)

Multi-panel alignment view



Multiple alignment view



Acknowledgements

Entrez Gene:
Donna Maglott

Web Design:
Mark Johnson

Genome Workbench:
Mike diCuccio

NCBI developers & curators