Using NCBI Resources for Gene Discovery

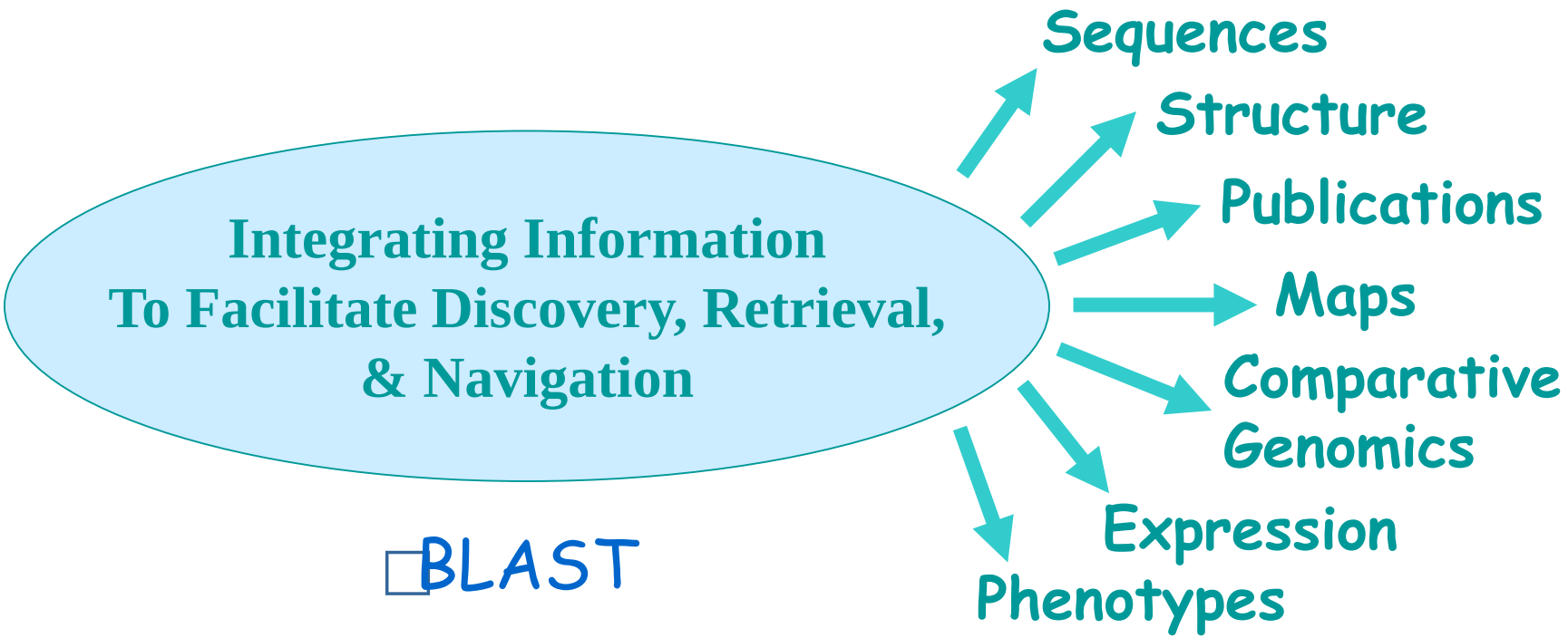
Kim D. Pruitt

Transcriptome 2002

Fundamental resources

- **Primary Databases - GenBank, dbEST, dbSTS, PubMed**
  - Archival - original data submissions
  - Database staff organize, but don't add additional information
- **Derivative Databases - RefSeq, LocusLink, UniGene, Map Viewer**
  - Curated/expert review
    - compilation and correction of data
  - Computationally Derived
  - Combinations

Infrastructure



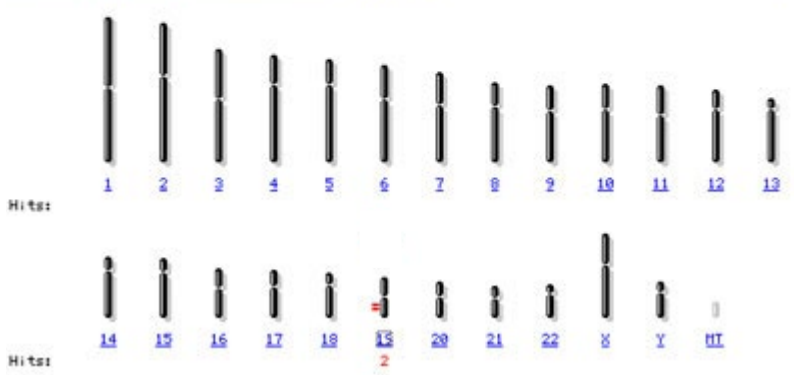
- BLAST
- Entrez
- Displays
- Navigation – extensive crosslinking

Increasing Discovery Space

RefSeq
genomes·transcripts·proteins

LocusLink

NCBI Map Viewer



NCBI Reference Sequences (RefSeq)

Genome Oriented Resource

- A sequence for each macromolecule of Central Dogma
- Linked on a residue by residue basis
- Objectively non-redundant and comprehensive

Curated Resource

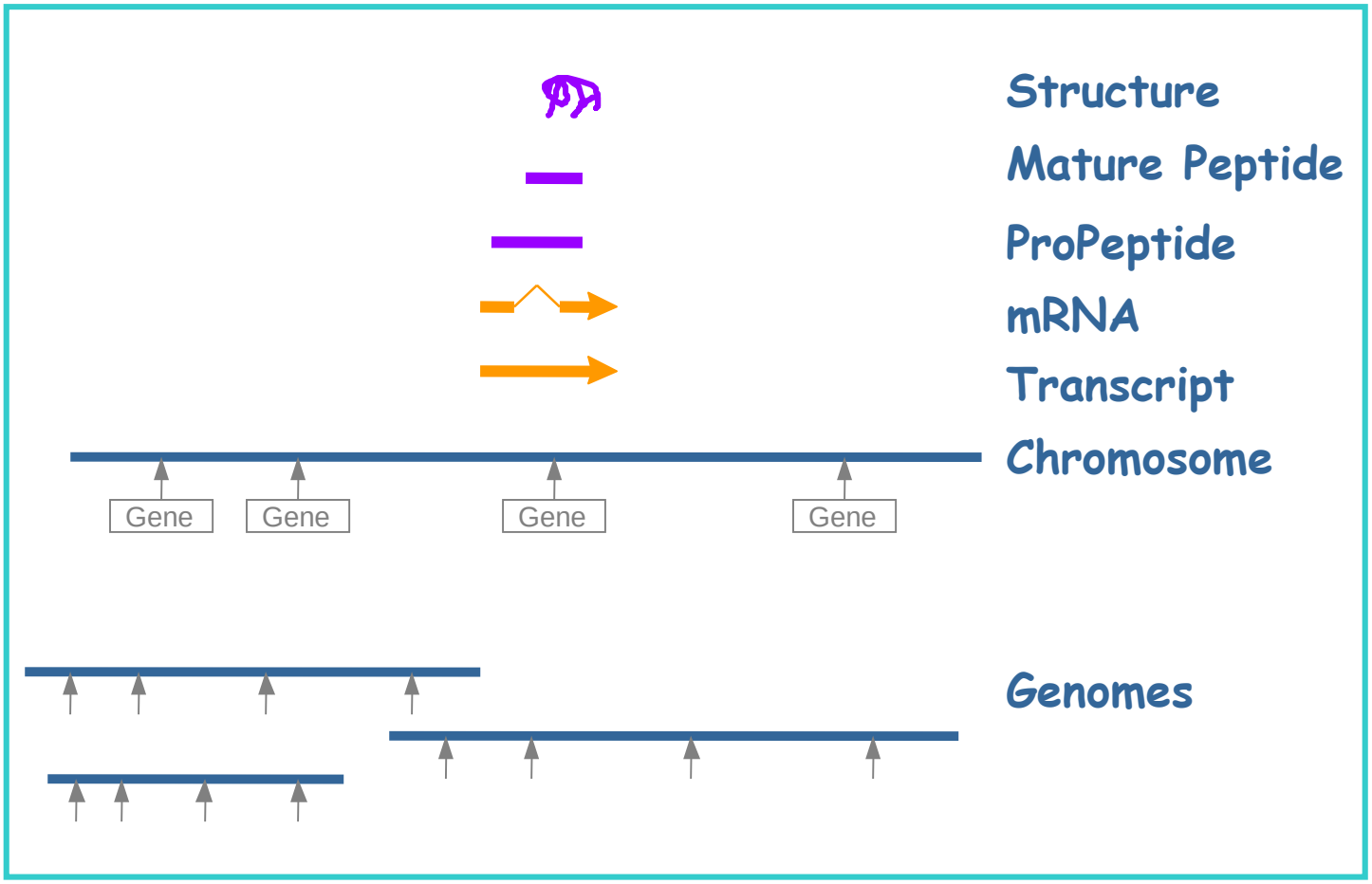
- Authoritative source by genome
- Derivative of GenBank but corrected, merged, extended
- Publicly distributed

Reagents for Genome Annotation and Analysis

Substrate for Functional Genomics

The Basic Model

A framework to anchor other information…



Disease

Function

Development

Pathways

Regulation

Expression

Genetics

Organisms

Populations

RefSeq: Scope

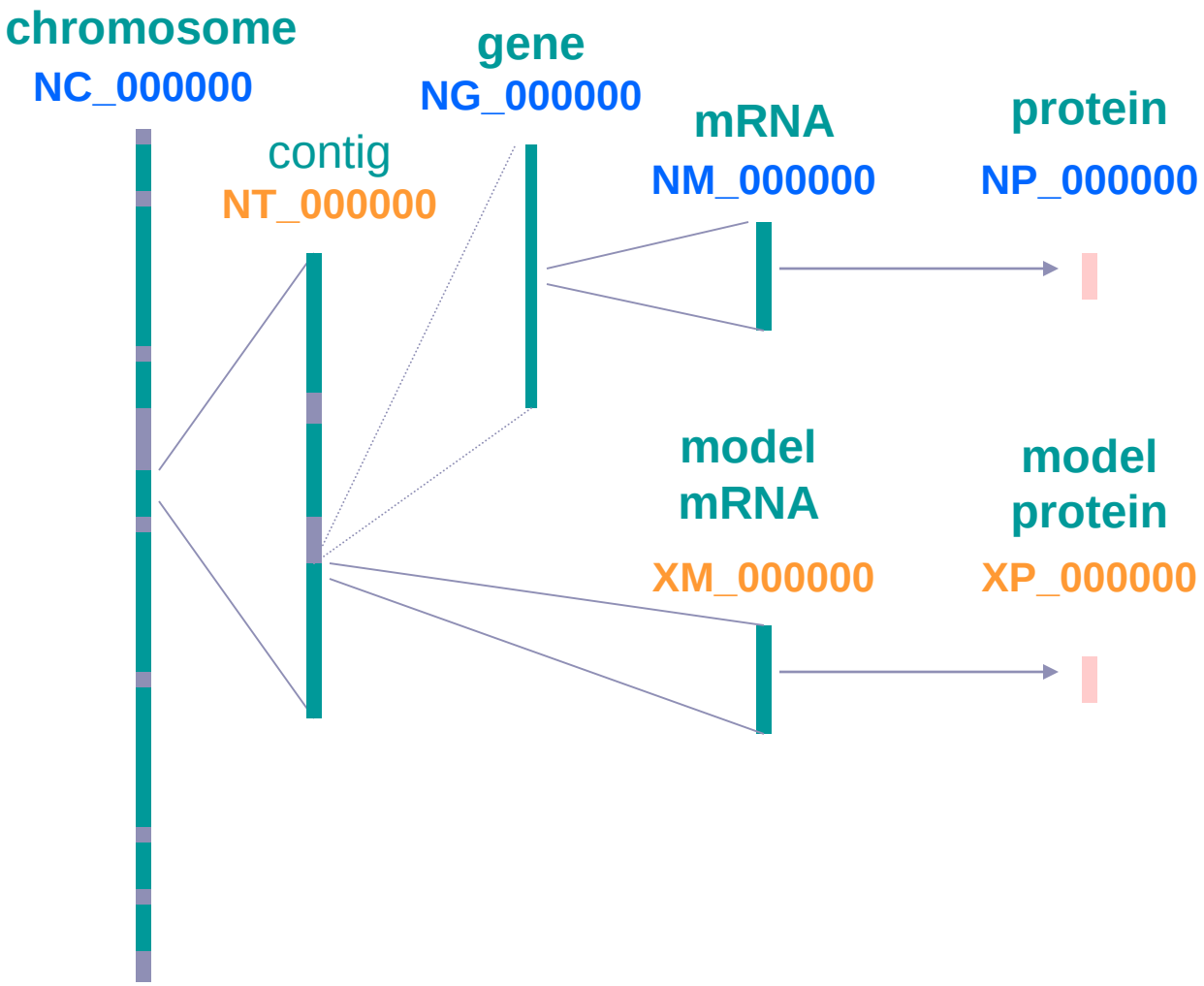
The NCBI Reference Sequence (RefSeq) project provides non-redundant sequence data including bacterial and viral genomes, mitochondrion, chromosomes, constructed genomic contigs, transcripts, and proteins.

Source	*Molecule*	*Organism*
Microbial Genomes		
Complete Genomes		
Genome Build Pipeline	Genomic Transcript Protein	Human Mouse
Collaboration	Genomic Transcript Protein	Yeast Drosophila Arabidopsis Miscellaneous other
LocusLink-based	Genomic Transcript Protein	Human Mouse Rat Drosophila Zebrafish

RefSeq as a protein database over 280,000 proteins

RefSeq: Products

Goal: One sequence entry for each naturally occurring molecule

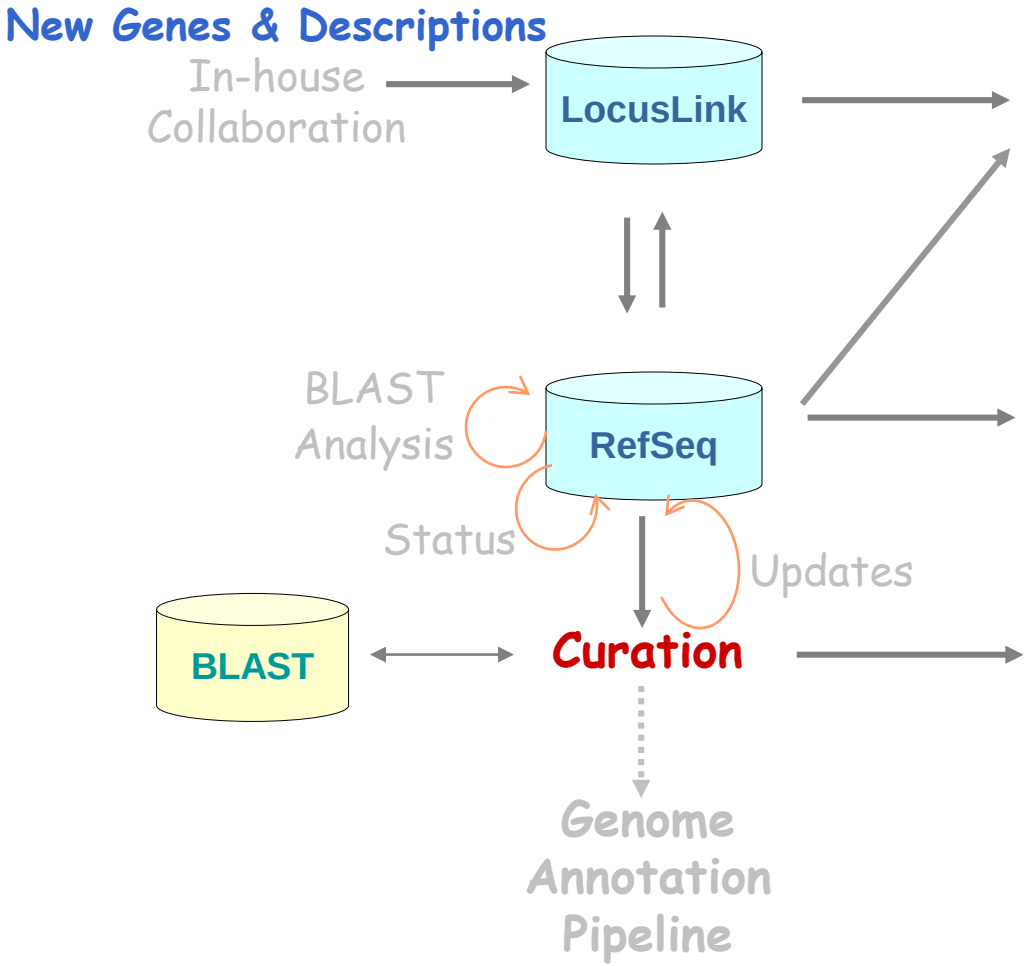


Multiple products for one gene are instantiated as separate RefSeqs with the same LocusID.

Key:
curated
calculated

Process Flow

Database support, automated steps, manual curation



Public Release:

- LocusLink Web Site
- Provisional & Predicted Records (Transcripts & Proteins)
- Reviewed Records (Genomic Regions, Transcripts, Proteins)

Accessible in:

- BLAST
- BLink
- Entrez
- FTP
- LocusLink

RefSeq: Curation

Reviewed Records:
Human - 4,685 accessions (3,445 genes)
Fly - 1,423 accessions (from FlyBase)
Mouse,Rat - 45 accessions

Time

Simple cases:
1-2 days

Large gene families:
months

Review
- Sequence
- Database data
- Literature

Sequence Editing
- Correct errors
- Extend UTRs
- Splice Variants
- Annotate Features

Final Check
- Quality Control

Public

New Type of RefSeq: Genomic Regions

Why?

Correct Assembly through Duplications, Paralogous Gene Clusters

Optimize Annotation in Gene Clusters

Used in Genome Annotation Process

Maintenance

New Genes:
- GenBank
- UniGene
- Genome Annotation
- Collaboration
- e-mail

Updates:
- GenBank updates
- Collaboration
- Ongoing curation
- Genome Annotation
- e-mail

We welcome feedback, suggestions, collaborations

Why Look For RefSeqs?

Enhanced Discovery Space:
- What do we already known?
- Predicted RefSeqs – where do we need to know more?
- Genome Annotation Products (Model RefSeqs)

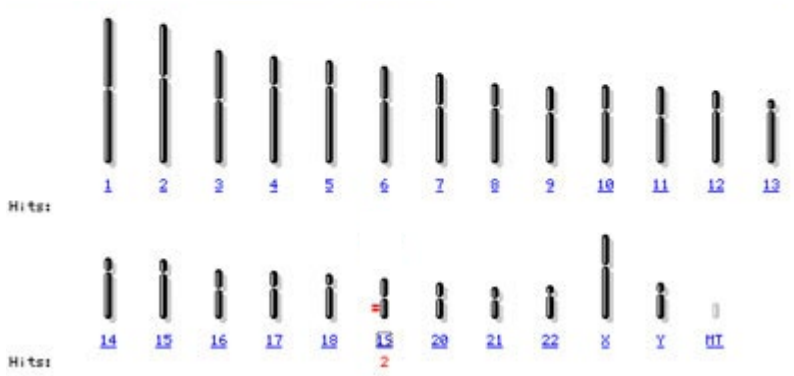
Analysis:
- transcript, protein, annotation, gene index

Increasing Discovery Space

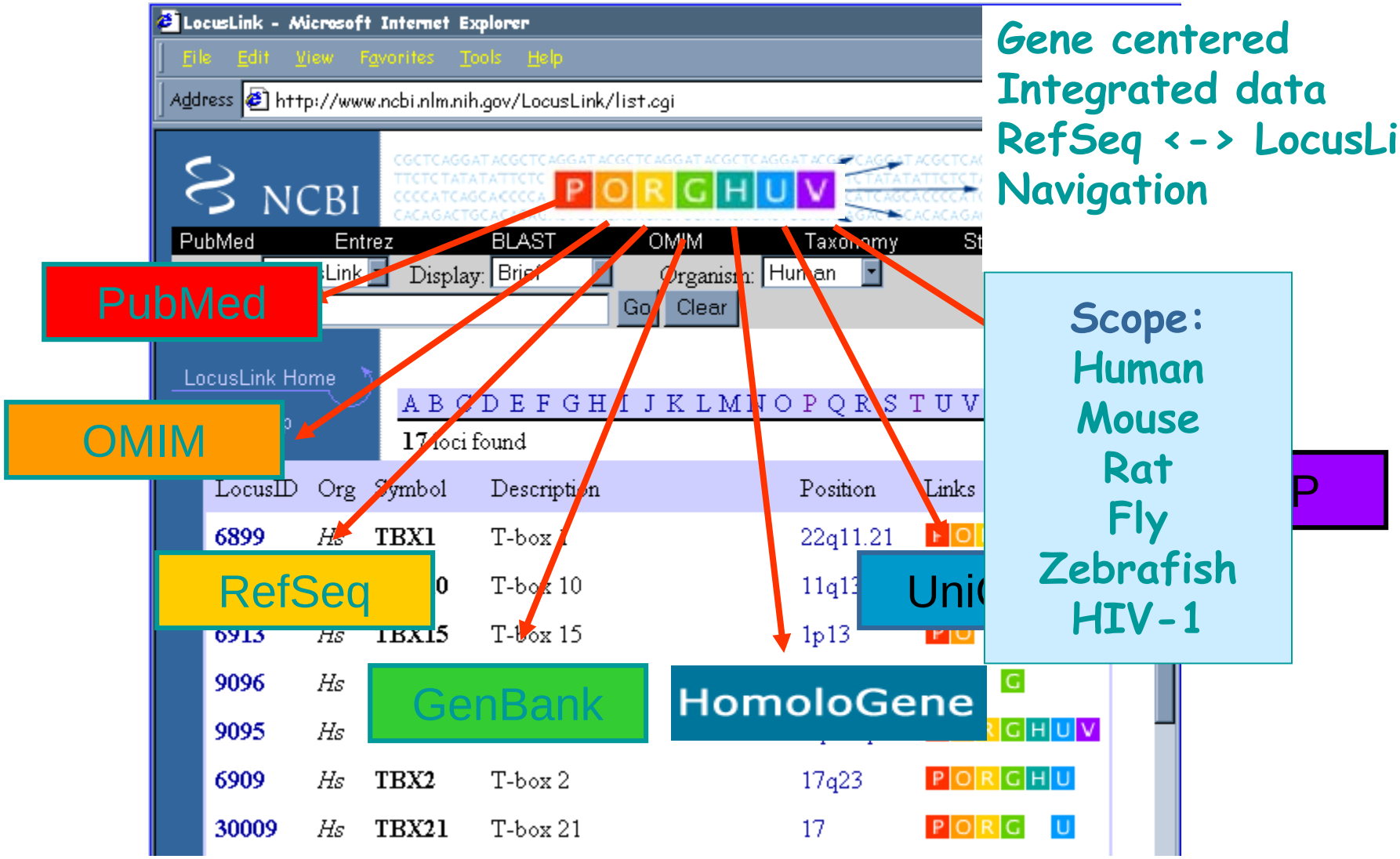
RefSeq
genomes·transcripts·proteins

LocusLink

NCBI Map Viewer

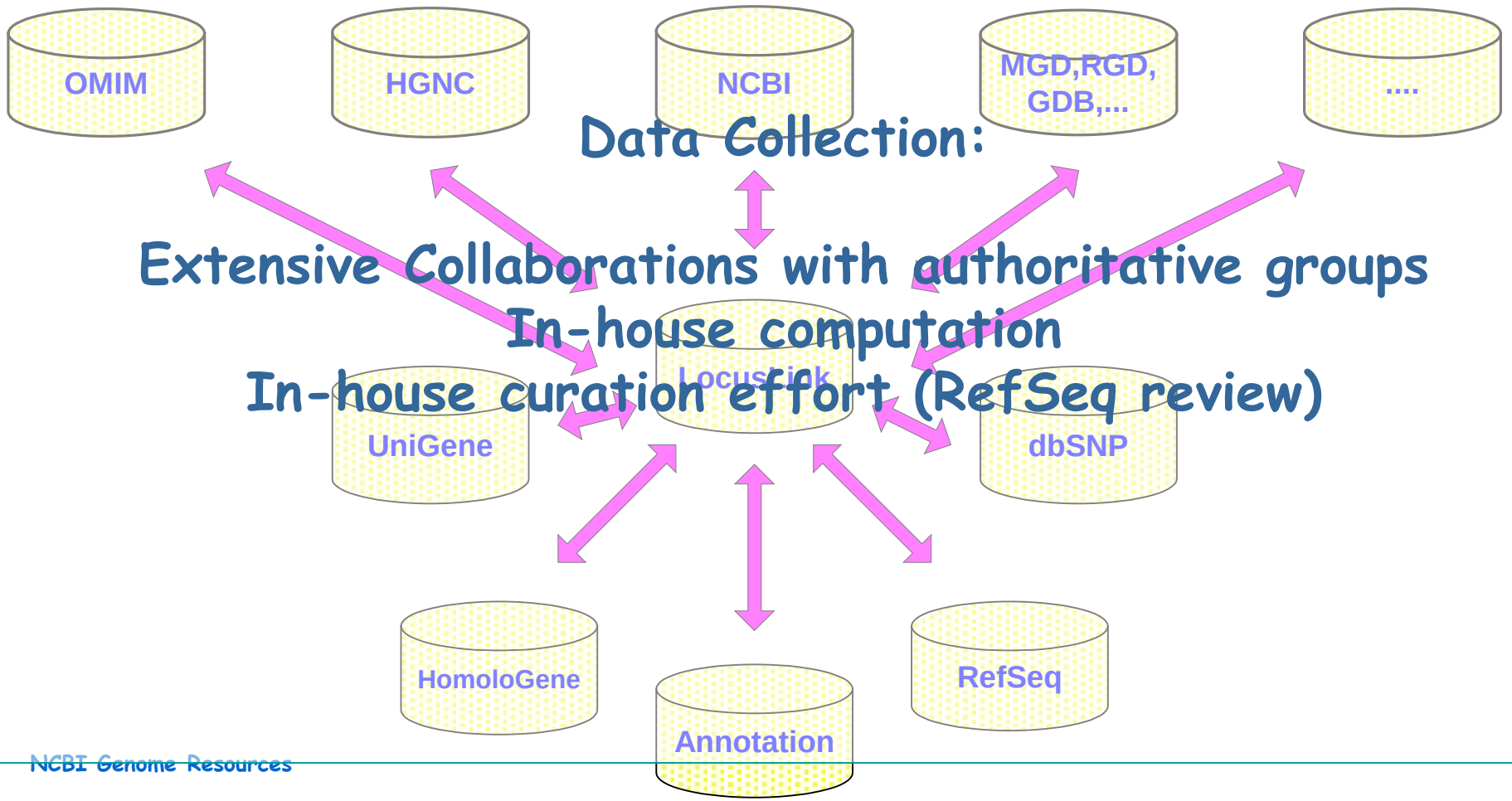


LocusLink


LocusLink - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.ncbi.nlm.nih.gov/LocusLink/list.cgi
NCBI
P O R G H U V
PubMed Entrez BLAST OMIM Taxonomy
Display: Brief
Organism: Human
Go Clear
LocusLink Home
A B C D E F G H I J K L M N O P Q R S T U V
17 loci found
LocusID Org Symbol Description Position Links
6899 Hs TBX1 T-box 1 22q11.21
T-box 10 11q13
6913 Hs TBX15 T-box 15 1p13
9096 Hs
9095 Hs
6909 Hs TBX2 T-box 2 17q23
30009 Hs TBX21 T-box 21 17
PubMed
OMIM
RefSeq
GenBank
HomoloGene
Gene centered
Integrated data
RefSeq <-> LocusLink
Navigation
Scope:
Human
Mouse
Rat
Fly
Zebrafish
HIV-1

LocusLink: Maintenance

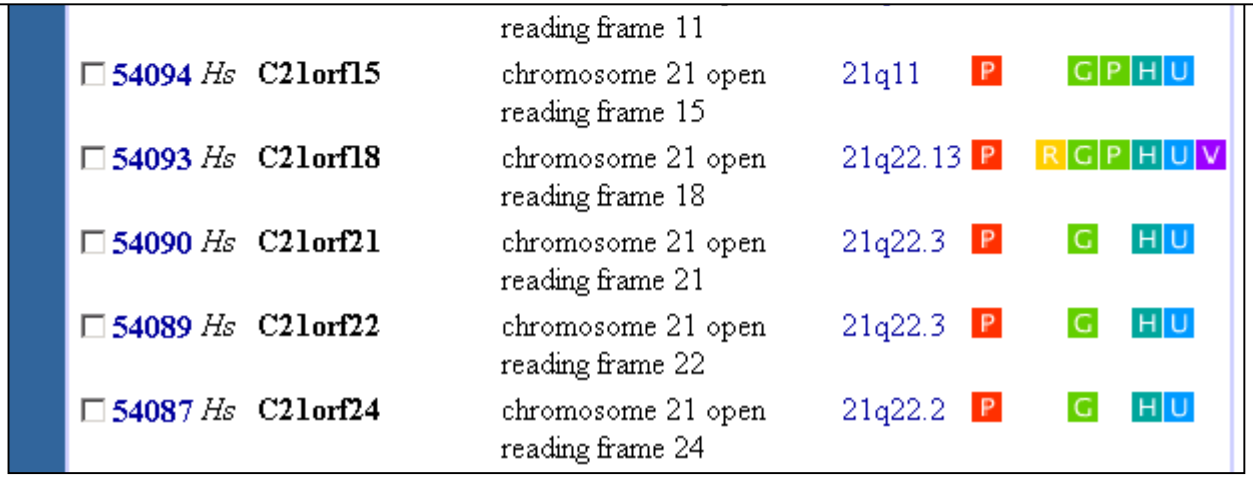
MIM numbers phenotypes Sequence links New genes	Symbols Descriptions Sequence links New Genes	Protein names Database links Sequence links New Genes	Symbols Descriptions Sequence links New Genes	Collaborators (Sequence links) More genomes



LocusLink: Discovery

Find novel uncharacterized genes on a finished chromosome
QUERY= 21[chr] NOT has_omim AND has_homol AND type_gene_protein

AND predicted
AND model
AND provisional
AND C21orf* OR MGC*

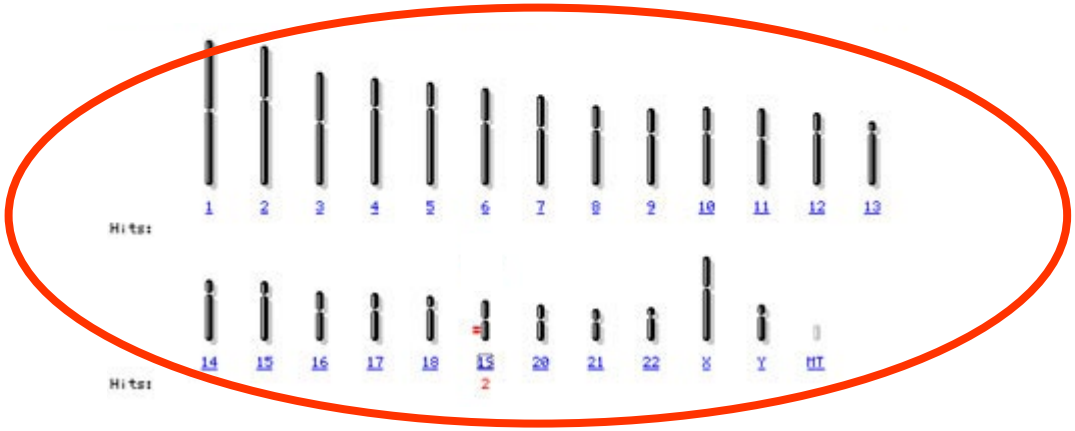


Increasing Discovery Space

RefSeq
genomes·transcripts·proteins

LocusLink

NCBI Map Viewer



Map Viewer

What?

Genome Assembly
Genome Annotation
Integrated map data (genetic, cytogenetic, RH)

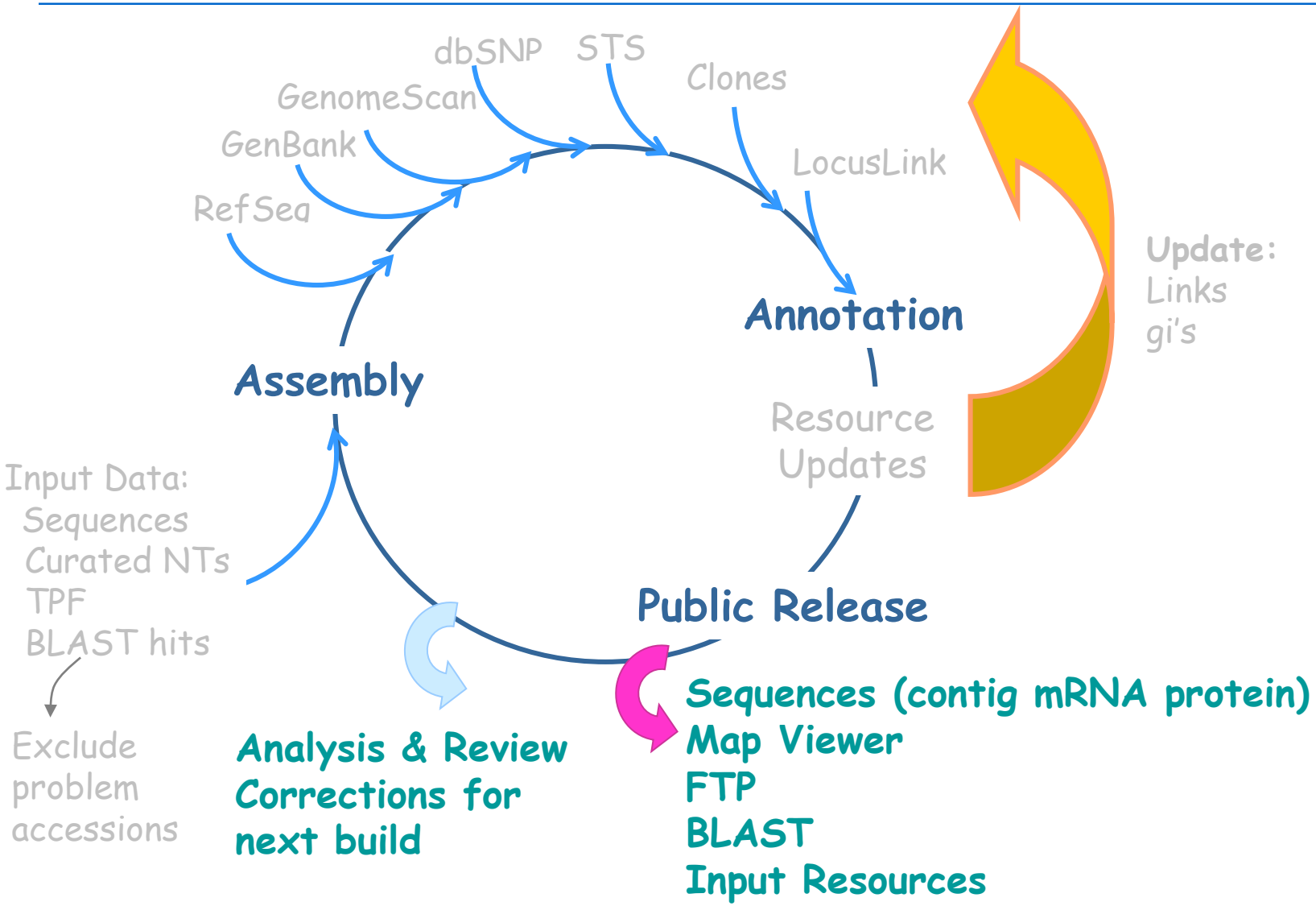
Scope?

Human
Drosophila
Mouse
(model genomes)

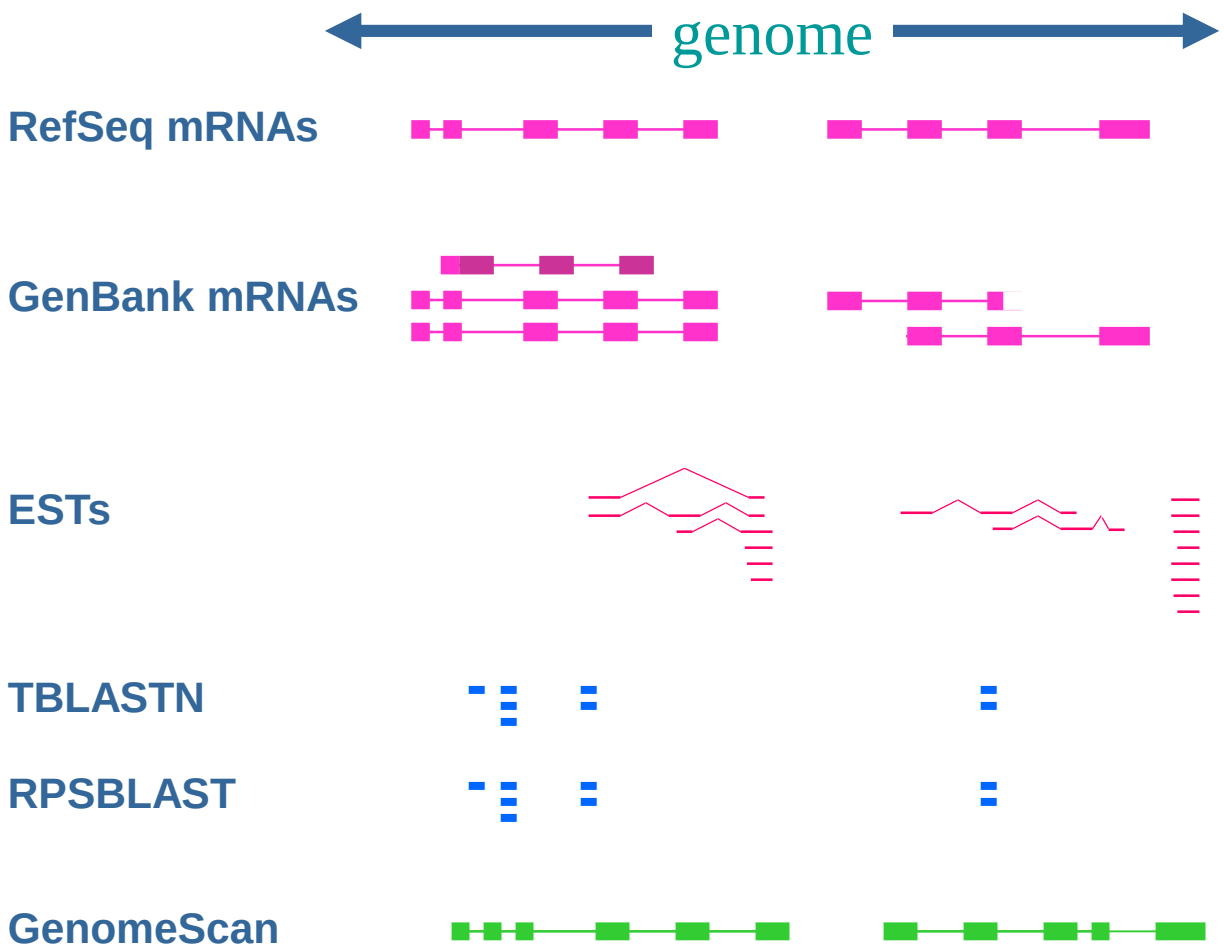
Why?

Facilitate discovery (genes, variation …)
Facilitate navigation
Facilitate use of genomic sequence information

Genome Build Process



RefSeq: a reagent for Annotation



Potential Problems:
- Gene Families
- Partial sequences
- Chimeric
- Intron read-through
- Linker
- Vector
- Wrong organism

RefSeq Advantages:
- Separate Gene Families
- Not Partial
- Means to correct problem sequences

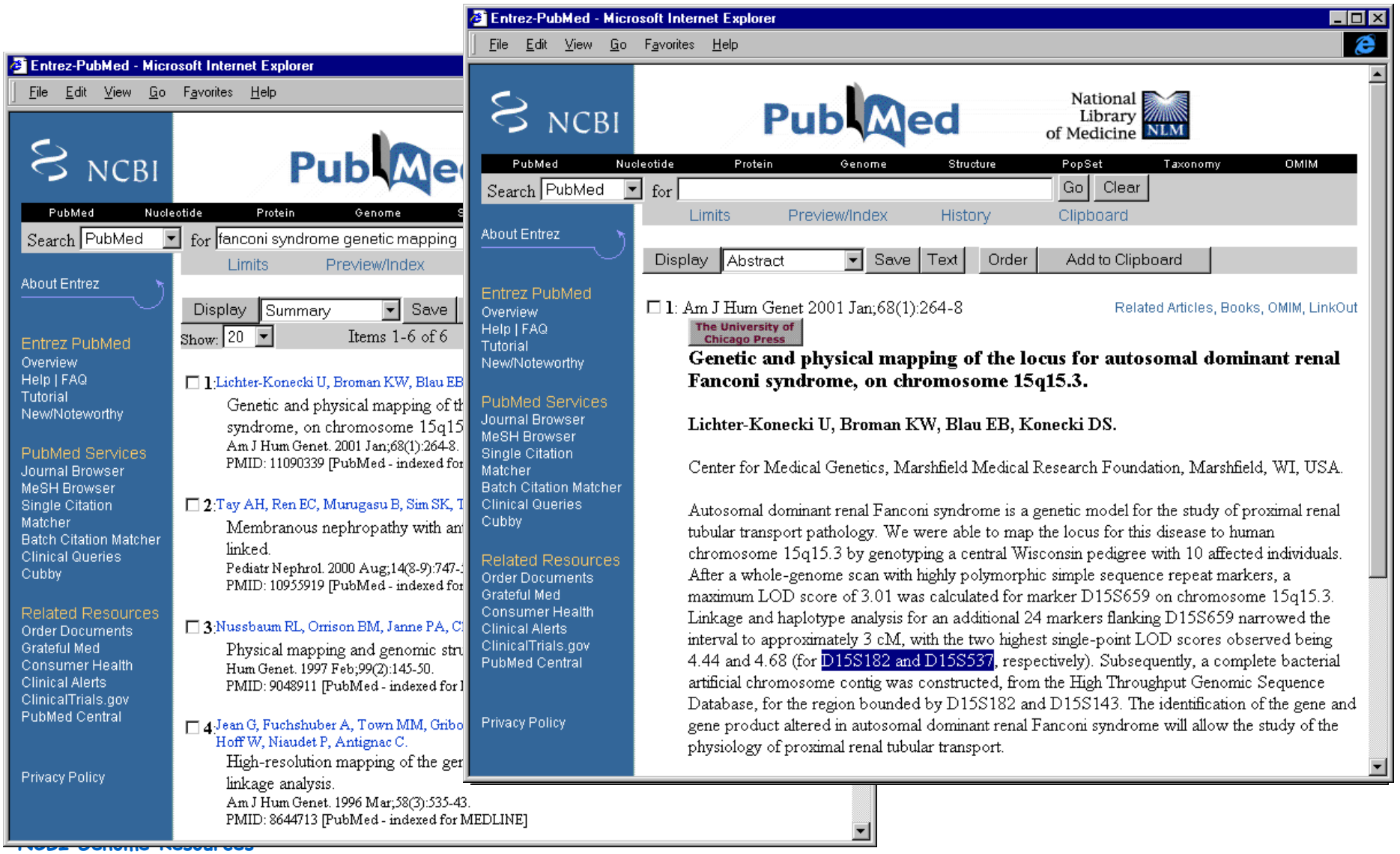
Quality Control:
RefSeq review results in excluding problem GenBank sequences from annotation pipeline

What questions can be asked?

- What genes (markers, SNPs) are between 2 markers?
- What BAC clones are available on Xq28?
- Where are there serine kinases?
- I've cloned gene xyz in my favorite organism. What is related in human?
- What is the evidence that there is a gene at position n?
- I have found a phenotype of interest between markers x and y; what is known about this region?

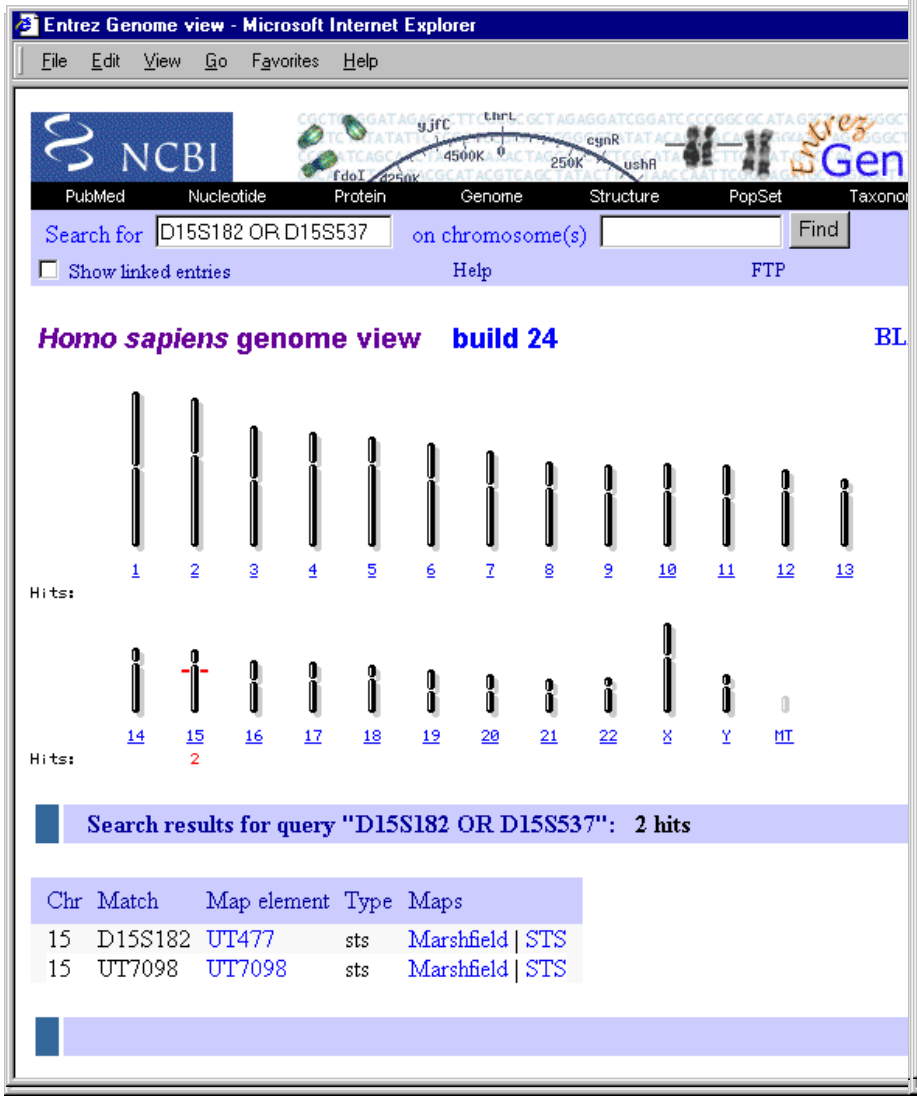
Fanconi Syndrome Genetic Mapping

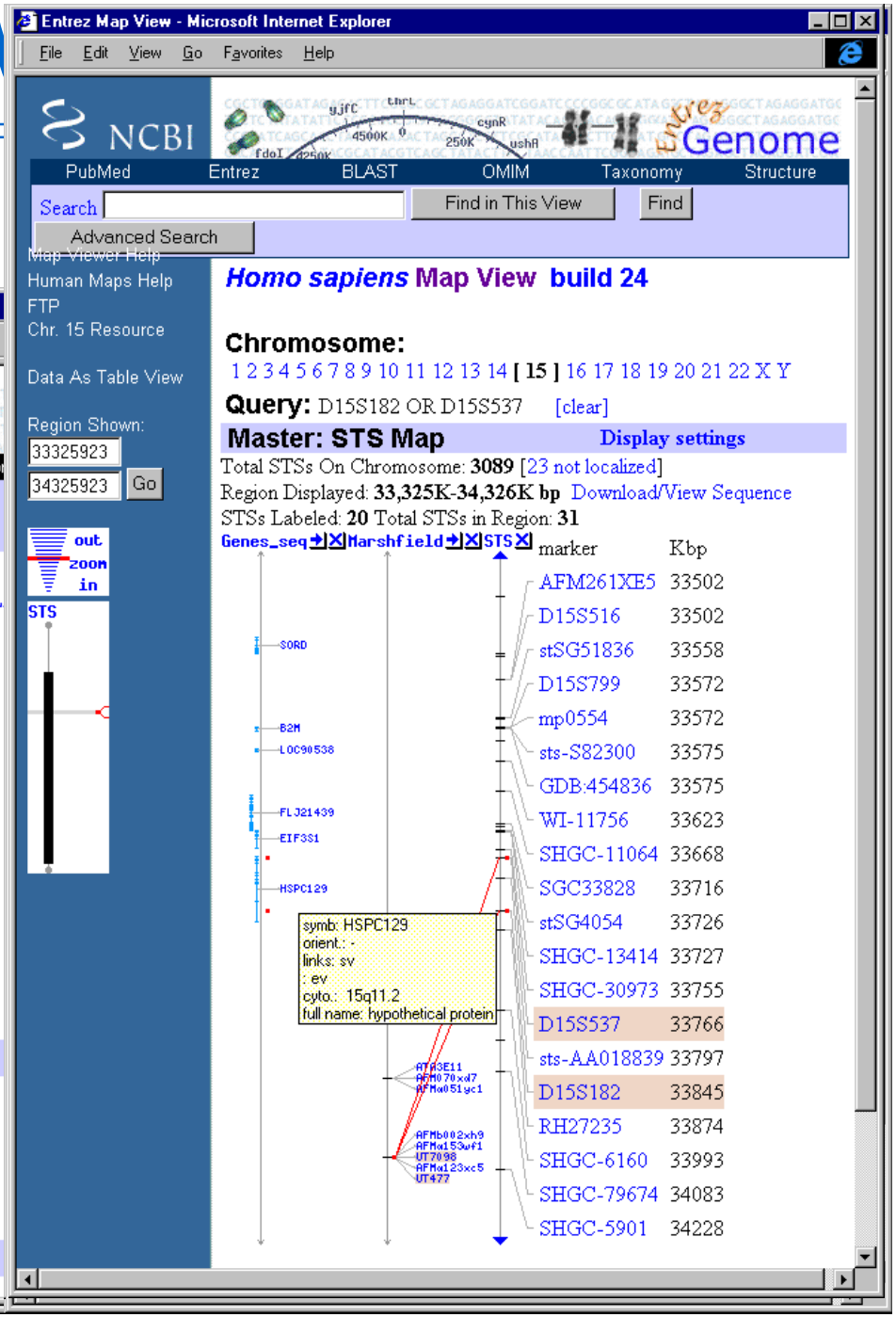
Pathology in proximal renal tubular transport



Query Map V

Look for genes in the region





NCBI

NCBI Sequence Viewer - Microsoft Ir

gil7705461 view - Microsoft Ir

File Edit View Go Favorit

BLAST Pu

Query: gi|7705461 hypothetical
Matching gi: 6841480
Lineage: Eukaryota; Metazoa; C

Best hits | Common Ti

116 BLAST hits to 14 uniq

0 Archaea | 0 Bacteria

Keep only

466 aa

Blast Result - Microsoft Internet Explorer

File Edit View Go Favorites Help

NCBI

Blast 2 Sequences results

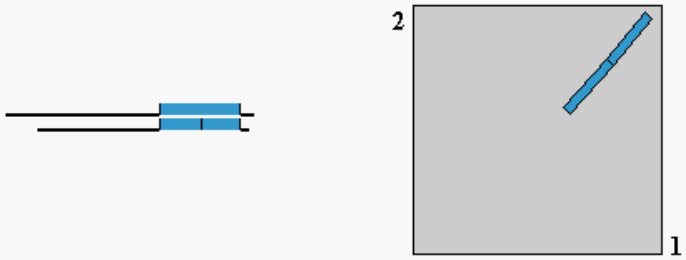
PubMed Entrez BLAST OMIM Taxonomy Structure

BLAST 2 SEQUENCES RESULTS VERSION BLASTP 2.1.2 [Oct-19-2000]

Matrix BLOSUM62 gap open: 11 gap extension: 1
x_dropoff: 50 expect: 300.00 wordsize: 3 Filter Align

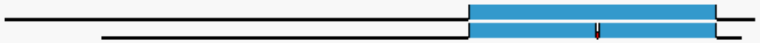
Sequence 1 gi 7705461 **Length** 466 (1 .. 466)

Sequence 2 gi 6323047 **Length** 397 (1 .. 397)



NOTE:The statistics (bitscore and expect value) is calculated based on the size of nr database

```
Score =  159 bits (399), Expect = 5e-38
Identities = 79/153 (51%), Positives = 104/153 (67%), Gaps = 1/153 (0%)
```



```
Query: 290 LVLDLDETLVHCSLNELEDAALTFPVLFQDVIYQVYVRLRPFFREFLERMSQMYEIILFT 349
           L+LDLDETLVH S   +  A    PV   D ++ VYV  RP   EFL R+SQ+YE+++FT
Sbjct: 230 LILDLDETLVHSSFKYMHSADFVLPVEIDDQVHNVYVIKRPGVDEFLNRVSQLYEVVVFT 289

Query: 350 ASKKVYADKLLNILDPKKQLVRHRLFREHCVCVQGNYIKDLNILGRDLSKTIIIDNSPQA 409
           AS   YA+ LL+ LDP    + HRLFRE C   +GNYIK+L+ +GR LS+TII+DNSP +
Sbjct: 290 ASVSRYANPLLDTLDPNGT-IHHRLFREACYNYEGNYIKNLSQIGRPLSETIILDNSPAS 348

Query: 410 FAYQLSNGIPIESWFMDKNDNELLKLIPFLEKL 442
           + +   + +PI SWF D +DNELL +IP LE L
Sbjct: 349 YIFHPQHAVPISSWFSDTHDNELLDIIPLLEDL 381
```

Internet

onomy Help

etical proteins [Caenorhabditi

ccharomyces cerev
otein YLR019w - y
myces cerevisiae]
factor isoform T
human

erevisiae)

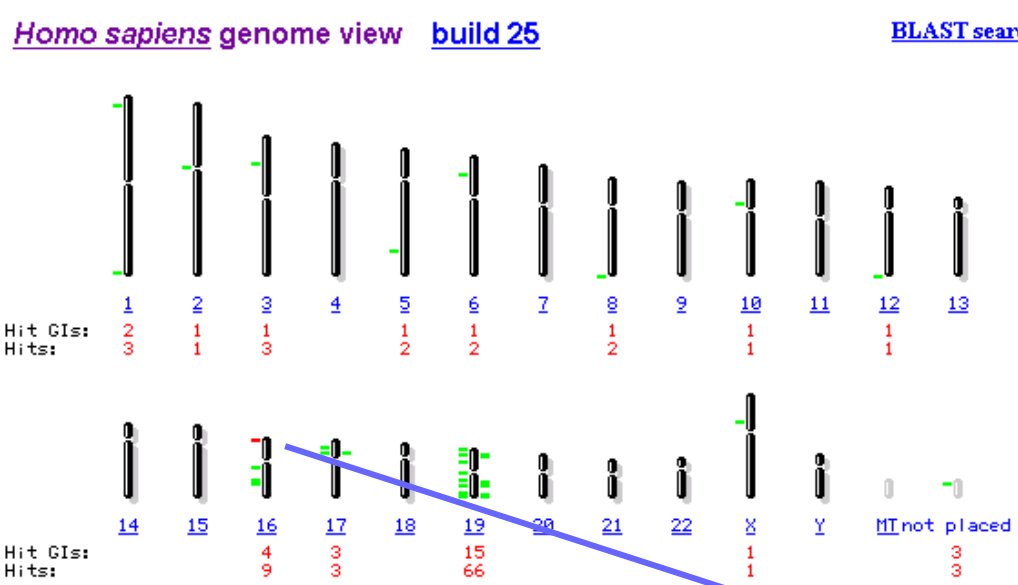
racting factor [Gallus gallus]

erevisiae)

ns]

NCBI Genome Resources

BLAST Queries: genomic distribution of matches



BLAST search results: 35 BLAST hits found
gi|4508012|ref|NM_003454.1| Homo sapiens zinc finger protein 200 (ZNF200), mRNA

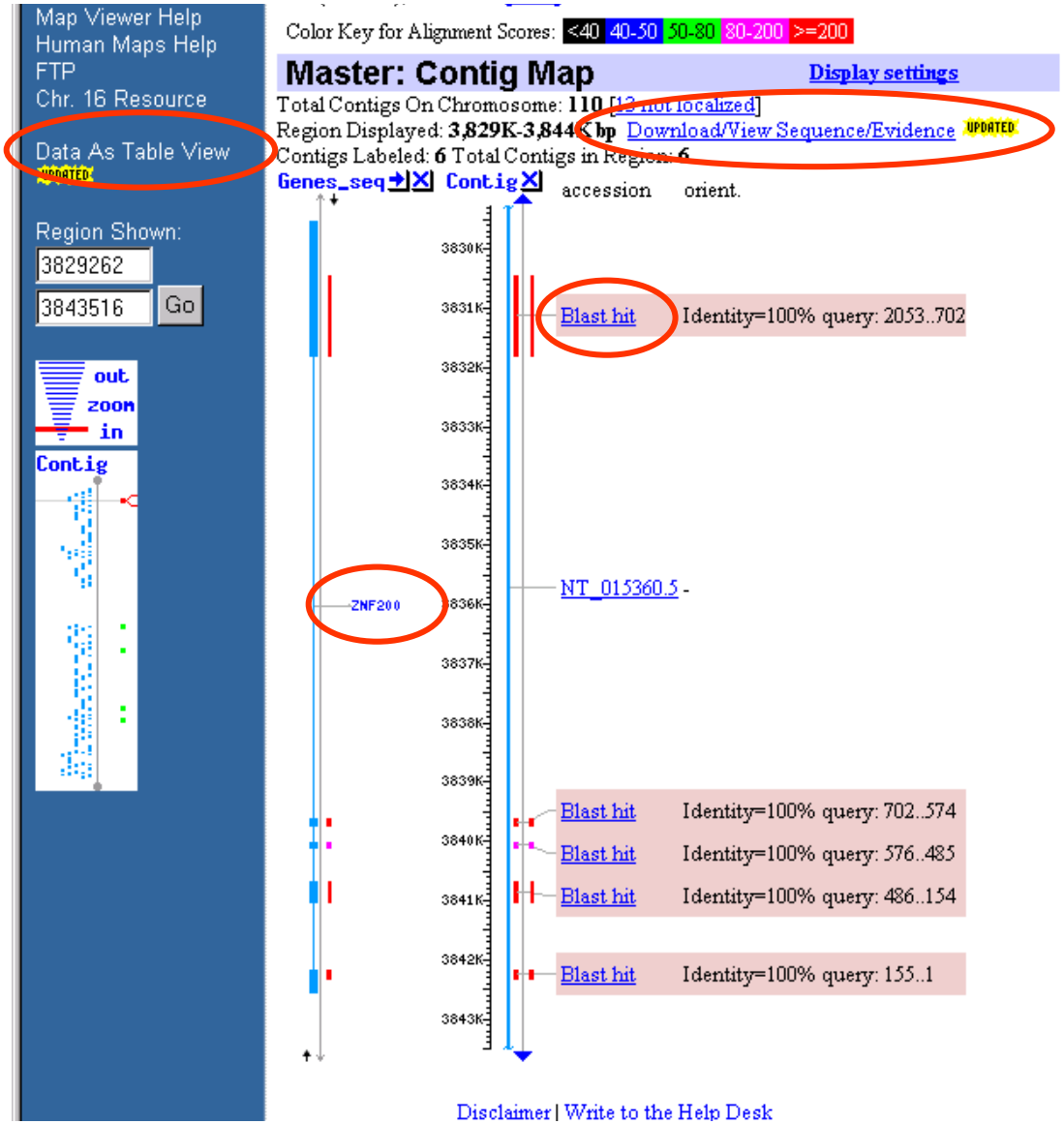
Sort results by score

Chr	Hit GI	Hits	Score	E value	Map element	
1	15297148	**2**	54.0	3e-04	NT_004536	Homo sapiens chromosome 1 working draft sequence segment
1	15296234	1	50.1	0.005	NT_004359	Homo sapiens chromosome 1 working draft sequence segment
2	15294983	1	60.0	5e-06	NT_022300	Homo sapiens chromosome 2 working draft sequence segment
3	15296047	**3**	54.0	3e-04	NT_005563	Homo sapiens chromosome 3 working draft sequence segment
5	15297054	**2**	56.0	8e-05	NT_006951	Homo sapiens chromosome 5 working draft sequence segment
6	15305200	**2**	50.1	0.005	NT_007592	Homo sapiens chromosome 6 working draft sequence segment
8	15299853	**2**	52.0	0.001	NT_007997	Homo sapiens chromosome 8 working draft sequence segment
10	15298811	1	58.0	2e-05	NT_008583	Homo sapiens chromosome 10 working draft sequence segment
12	15302779	1	54.0	3e-04	NT_009455	Homo sapiens chromosome 12 working draft sequence segment
16	15318559	**5**	2680	0.0	NT_015360	Homo sapiens chromosome 16 working draft sequence segment

Result from a BLAST query of a zinc finger protein

Best match

Review the alignment



A click away:

- Alignments (BLAST hit)
- Gene Description

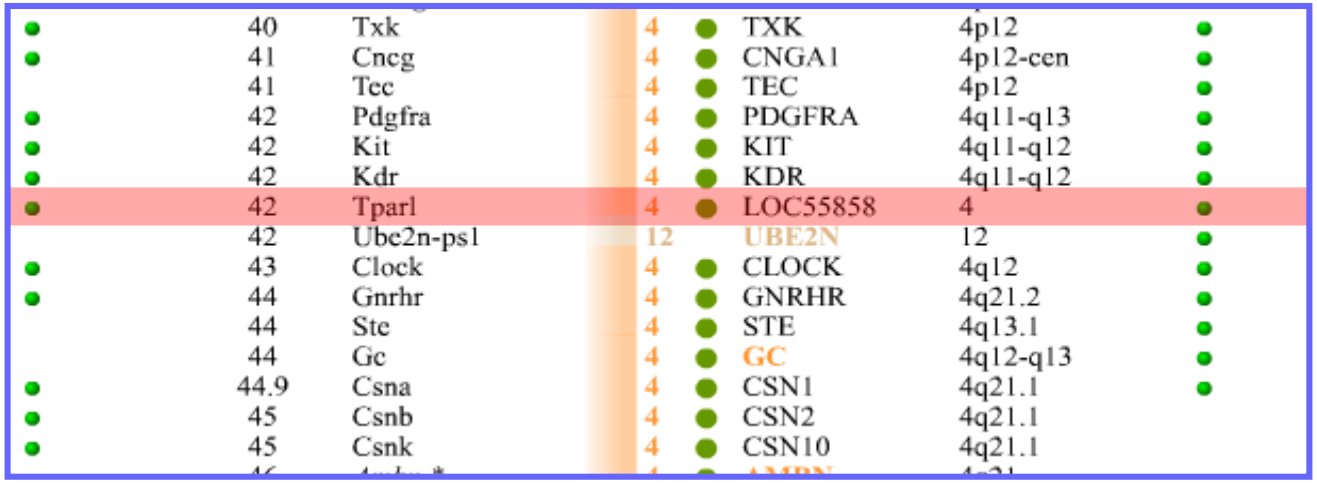
- Report of all features in the region
- Sequence in the region
- other mRNAs aligning in the region

- Homology Maps
- Model Maker - Define your own gene model based on alignments in the region

Homology Maps

Genes in regions of conserved synteny

	4p12	TXK	5	Txk	40	
	4p12	TEC	5	Tec	41	
	4	KIAA1458	5	5033405K12Rik		
	4	KIAA0826	5	2310004H21Rik		
	4q12	SGCB	5	Sgcb		
	4	DKFZP586K0717	5	1300019H17Rik		
	4q11	CHIC2	5	Chic2		
	4q11-q13	PDGFRA	5	Pdgfra	42	
	4q11-q12	KIT	5	Kit	42	
	4q11-q12	KDR	5	Kdr	42	
	4	LOC55858	5	Tparl	42	
	4q12	CLOCK	5	Clock	43	
	4q12-q13.3	SEC3	5	2810407P21Rik		
	4q12	IGFBP7	5	Igfbp7		
	4q12-q13.3	REST	5	Rest		
	4q11-q12	*SMAP31* *		1110018K11Rik		
	4	LEC3	5	5430402I23Rik		
	4	EPHA5	5	Epha5		
	4q13.3	BRDG1	5	Brdg1-pending		

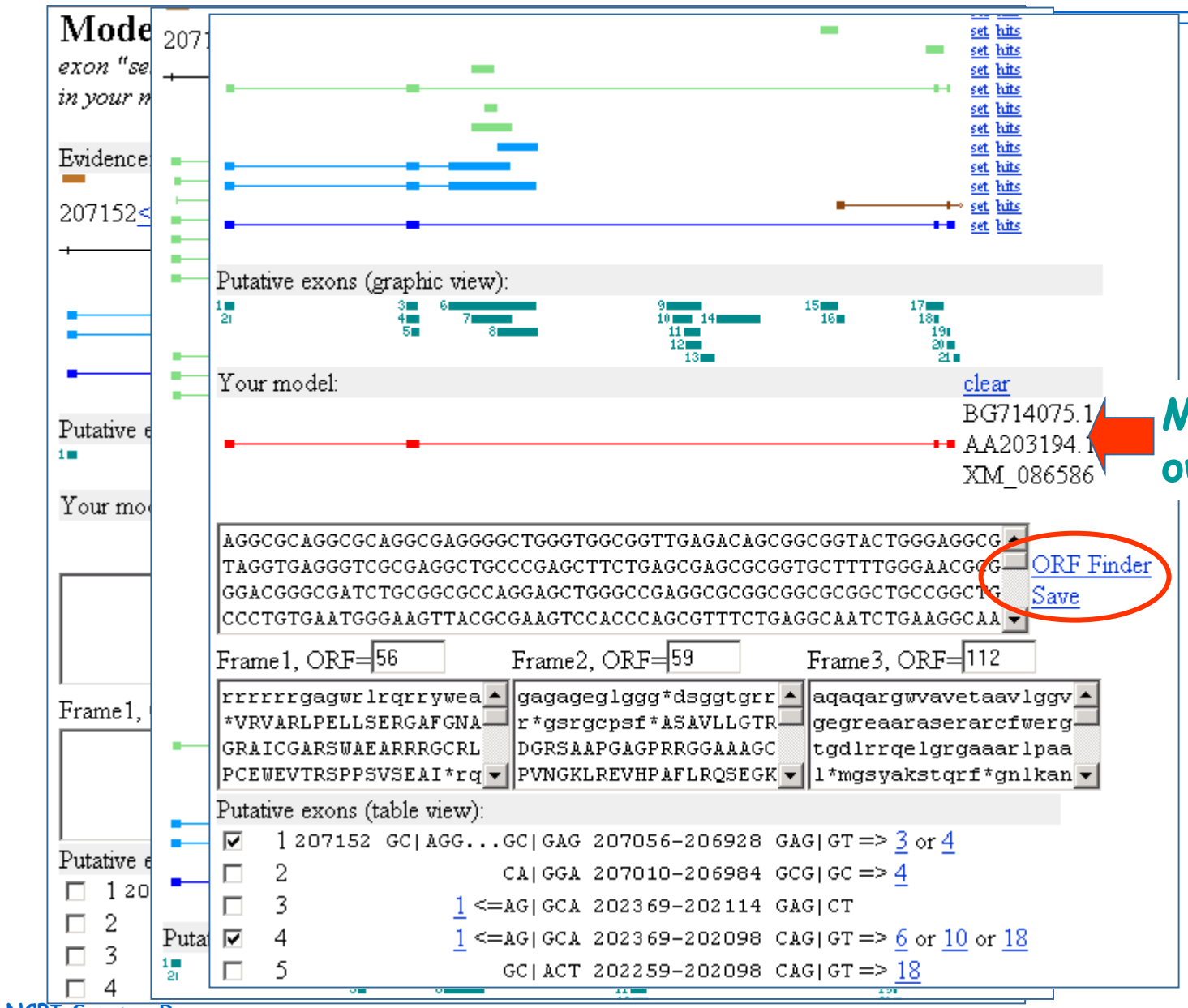


	40	Txk	4	TXK	4p12	
	41	Cncg	4	CNGA1	4p12-cen	
	41	Tec	4	TEC	4p12	
	42	Pdgfra	4	PDGFRA	4q11-q13	
	42	Kit	4	KIT	4q11-q12	
	42	Kdr	4	KDR	4q11-q12	
	42	Tparl	4	LOC55858	4	
	42	Ube2n-ps1	12	UBE2N	12	
	43	Clock	4	CLOCK	4q12	
	44	Gnrhr	4	GNRHR	4q21.2	
	44	Ste	4	STE	4q13.1	
	44	Gc	4	GC	4q12-q13	
	44.9	Csna	4	CSN1	4q21.1	
	45	Csnb	4	CSN2	4q21.1	
	45	Csnk	4	CSN10	4q21.1	

Anchored by human gene order

Anchored by mouse gene order

Map Viewer: Model Maker



Increasing Discovery Space

Entrez

GenBank

RefSeq
genomes·transcripts·proteins

PubMed

Expression (GEO)

LocusLink

UniGene

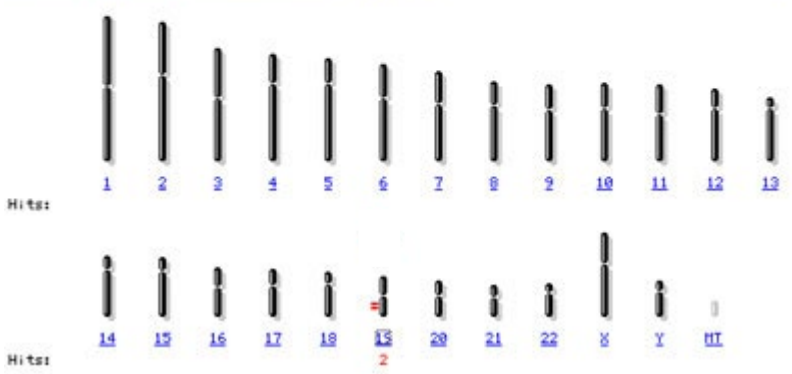
Clones (Clone Registry)

HomoloGene

NCBI Map Viewer

Markers (UniSTS)

OMIM



Variation (dbSNP)

Homology Maps

Acknowledgments

RefSeq Curator Staff
BLAST Team
Entrez Team
NCBI Service Desk Staff

Collaborators:
Human Gene Nomenclature Committee
OMIM Staff
The Jackson Laboratory
Rat Genome Database

Genome Build Team:
Richa Agarwala
Hsiu-Chuan Chen
Slava Chetvernin
Deanna Church
Olga Ermolaeva
Renata Geer
Wratko Hlavina
Wonhee Jang
Jonathan Kans
Ken Katz
Paul Kitts
David Lipman
Adam Lowe
Donna Maglott
Jim Ostell
Kim Pruitt
Sergey Resenchuk
Victor Sapojnikov
Greg Schuler
Steve Sherry
Andrei Shkeda
Tatiana Tatusova
Lukas Wagner
Sarah Wheelan

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