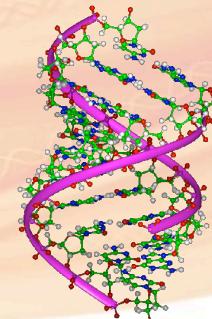


Bioinformatics Computational Methods 1 - BIOL 6308



October 22nd 2013

<http://155.33.203.128/cleslin/home/teaching6308F2013.php>

Last Time

- Why Study Genomes
 - **What About Other Genomic Features**
 - **And now the Genomes...**
- Other Genomic Features
 - Regulatory proteins
 - TF
 - Regulatory regions
 - TFBS
- Consensus
- Genomes
- Gene Ontology
 - Structure of GO
 - How to use GO

NCBI

The National Institutes of Health

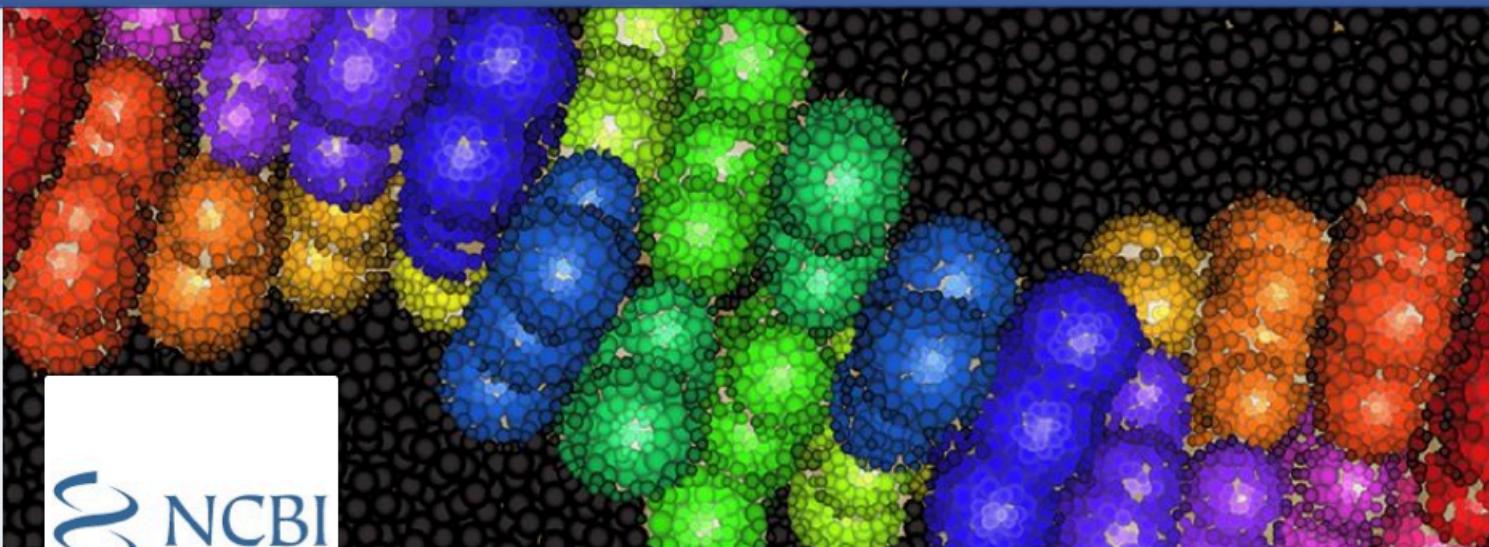


National Center for Biotechnology Information NCBI

- Created by Public Law 100-607 in 1988 as part of National Library of Medicine at NIH to:
 - Create automated systems for knowledge about molecular biology, biochemistry, and genetics
 - Perform research into advanced methods of analyzing and interpreting molecular biology data
 - Enable biotechnology researchers and medical care personnel to use the systems and methods developed
- **NCBI is an important scientific portal**

NCBI has a Facebook Page!!!

NCBI - National Center for Biotechnology Information



NCBI - National Center for Biotechnology Information

73,867 likes · 1,161 talking about this · 513 were here

Medical & Health · Government Organization · Add A C
The National Center for Biotechnology Information (NCBI) advances science and health by providing access to biomedical and genomic information.

About - Suggest an Edit

Photos Likes Policies Events

Like Review

Comment & Privacy Policies

ASHG ANNUAL MEETING San Francisco November 11-15, 2012 American Society of Human Genetics

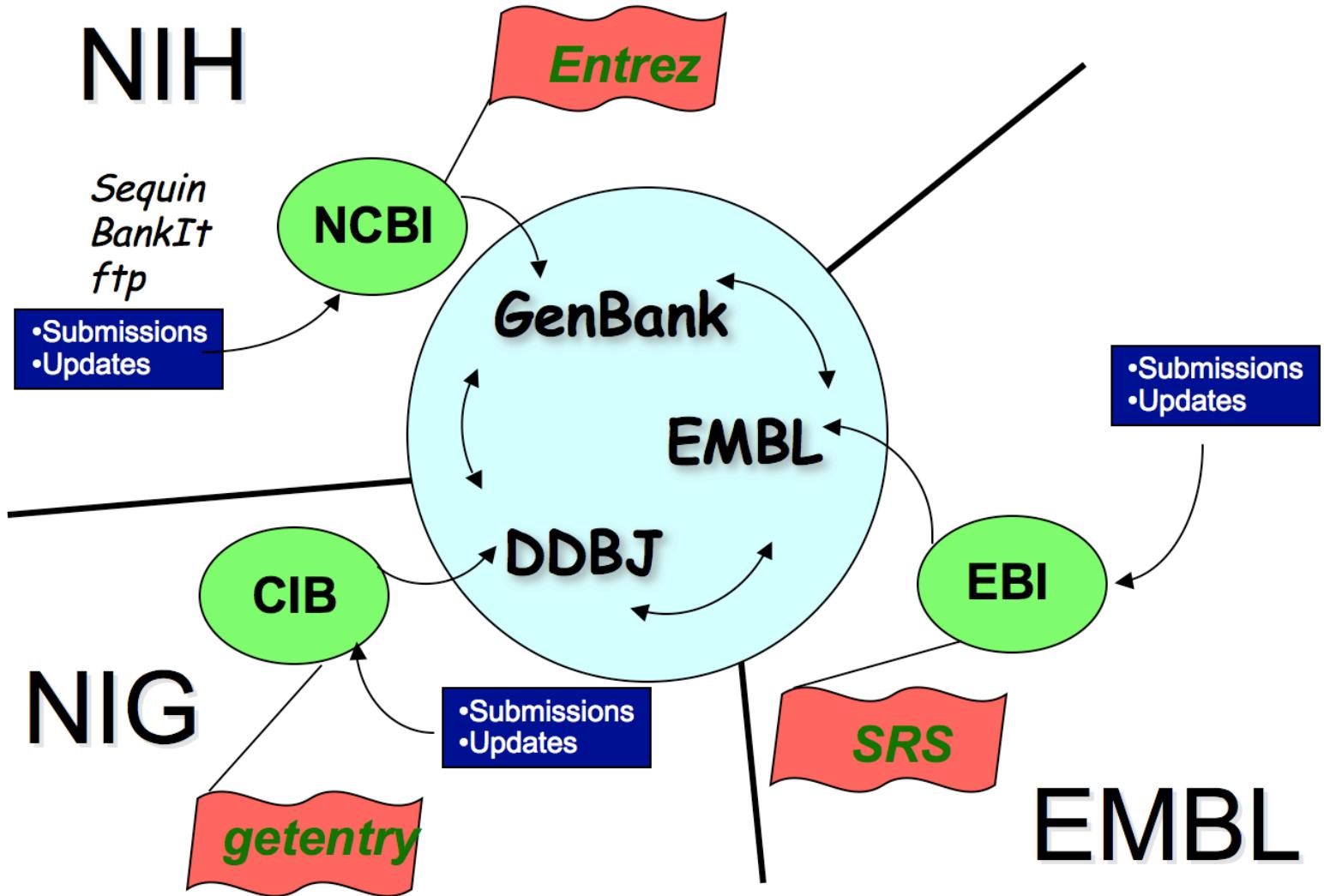
2 ▾

<http://www.facebook.com/ncbi.nlm>

The International Nucleotide Sequence Database Collaboration

- Sequence data shared by INSD
- INSD developed and maintained collaboratively between DDBJ, EMBL, and GenBank for over 18 years
- International Advisory Committee
 - The INSDC advisory board
 - Made up of members of each of the databases advisory bodies
 - Members of committee unanimously endorsed the existing data-sharing policy of the three databases that make up the INSDC
 - <http://www.insdc.org/>

The International Nucleotide Sequence Database Collaboration



Web Access: www.ncbi.nlm.nih.gov

NCBI Resources How To cleslin1 My NCBI Sign Out

NCBI National Center for Biotechnology Information All Databases Search

NCBI Home Resource List (A-Z) All Resources Chemicals & Bioassays Data & Software DNA & RNA Domains & Structures Genes & Expression Genetics

You are here: NCBI > National Center for Biotechnology Information

GETTING STARTED

- NCBI Education
- NCBI Help Manual
- NCBI Handbook
- Training & Tutorials

RESOURCES

- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

POPULAR

- PubMed
- Nucleotide
- BLAST
- PubMed Central
- Gene
- Bookshelf
- Protein
- OMIM
- Genome
- SNP
- Structure

FEATURED

- Genetic Testing Registry
- PubMed Health
- GenBank
- Reference Sequences
- Map Viewer
- Human Genome
- Mouse Genome
- Influenza Virus
- Primer-BLAST
- Sequence Read Archive

NCBI INFORMATION

- About NCBI
- Research at NCBI
- NCBI Newsletter
- NCBI FTP Site
- NCBI on Facebook
- NCBI on Twitter
- NCBI on YouTube

Write to the Help Desk

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS Feeds](#)

Homepage

Popular Resources

PubMed
Bookshelf
PubMed Central
PubMed Health
BLAST
Nucleotide
Genome
SNP

Links!

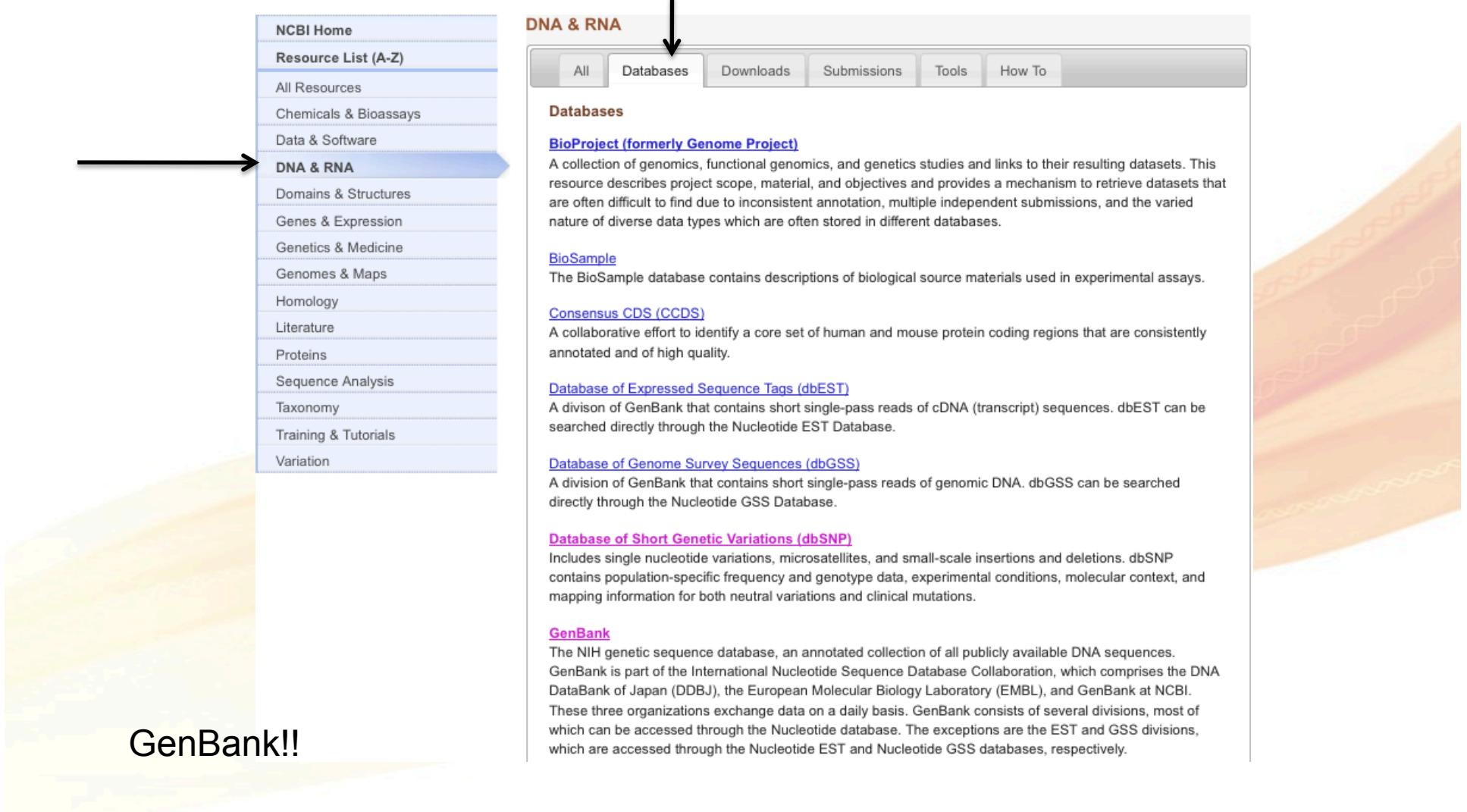
Common footer

NCBI Databases and Services

- GenBank primary sequence database
- Free public access to biomedical literature
 - PubMed free Medline (3.5 million searches per day)
 - PubMed Central full text online access
- Entrez integrated molecular and literature databases
- BLAST highest volume sequence search service
 - 1997 – 38,000 searches per day
 - Now ~ 200 K searches per day
- Software and databases for download

<http://duncan.hull.name/2010/07/27/pubmed-20-million/>

<http://www.ncbi.nlm.nih.gov/guide/all/>



NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

DNA & RNA

All Databases Downloads Submissions Tools How To

Databases

[BioProject \(formerly Genome Project\)](#)
A collection of genomics, functional genomics, and genetics studies and links to their resulting datasets. This resource describes project scope, material, and objectives and provides a mechanism to retrieve datasets that are often difficult to find due to inconsistent annotation, multiple independent submissions, and the varied nature of diverse data types which are often stored in different databases.

[BioSample](#)
The BioSample database contains descriptions of biological source materials used in experimental assays.

[Consensus CDS \(CCDS\)](#)
A collaborative effort to identify a core set of human and mouse protein coding regions that are consistently annotated and of high quality.

[Database of Expressed Sequence Tags \(dbEST\)](#)
A division of GenBank that contains short single-pass reads of cDNA (transcript) sequences. dbEST can be searched directly through the Nucleotide EST Database.

[Database of Genome Survey Sequences \(dbGSS\)](#)
A division of GenBank that contains short single-pass reads of genomic DNA. dbGSS can be searched directly through the Nucleotide GSS Database.

[Database of Short Genetic Variations \(dbSNP\)](#)
Includes single nucleotide variations, microsatellites, and small-scale insertions and deletions. dbSNP contains population-specific frequency and genotype data, experimental conditions, molecular context, and mapping information for both neutral variations and clinical mutations.

[GenBank](#)
The NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis. GenBank consists of several divisions, most of which can be accessed through the Nucleotide database. The exceptions are the EST and GSS divisions, which are accessed through the Nucleotide EST and Nucleotide GSS databases, respectively.

GenBank!!

<http://www.ncbi.nlm.nih.gov/guide/all/>

The screenshot shows the NCBI DNA & RNA Tools page. On the left, a sidebar lists various categories: NCBI Home, Resource List (A-Z), All Resources, Chemicals & Bioassays, Data & Software, DNA & RNA (which is highlighted with a blue background and has a black arrow pointing to it), Domains & Structures, Genes & Expression, Genetics & Medicine, Genomes & Maps, Homology, Literature, Proteins, Sequence Analysis, Taxonomy, Training & Tutorials, and Variation. At the top, there is a navigation bar with tabs: All, Databases, Downloads, Submissions, Tools (which is highlighted with a grey background and has a black arrow pointing to it), and How To. The main content area is titled "DNA & RNA" and contains several sections: "Tools", "Basic Local Alignment Search Tool (BLAST)", "Batch Entrez", "E-Utilities", "Genome BLAST", "Genome Remapping Service", "Genome Workbench", "Open Reading Frame Finder (ORF Finder)", and "Primer-BLAST". Each section provides a brief description of the tool's function.

E-Utilities

DNA & RNA

All Databases Downloads Submissions Tools How To

Tools

Basic Local Alignment Search Tool (BLAST)
Finds regions of local similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as to help identify members of gene families.

Batch Entrez
Allows you to retrieve records from many Entrez databases by uploading a file of GI or accession numbers from the Nucleotide or Protein databases, or a file of unique identifiers from other Entrez databases. Search results can be saved in various formats directly to a local file on your computer.

E-Utilities
Tools that provide access to data within NCBI's Entrez system outside of the regular web query interface. They provide a method of automating Entrez tasks within software applications. Each utility performs a specialized retrieval task, and can be used simply by writing a specially formatted URL.

Genome BLAST
This tool compares nucleotide or protein sequences to genomic sequence databases and calculates the statistical significance of matches using the Basic Local Alignment Search Tool (BLAST) algorithm.

Genome Remapping Service
NCBI's Remap tool allows users to project annotation data and convert locations of features from one genomic assembly to another or to RefSeqGene sequences through a base by base analysis. Options are provided to adjust the stringency of remapping, and summary results are displayed on the web page. Full results can be downloaded for viewing in NCBI's Genome Workbench graphical viewer, and annotation data for the remapped features, as well as summary data, is also available for download.

Genome Workbench
An integrated application for viewing and analyzing sequence data. With Genome Workbench, you can view data in publicly available sequence databases at NCBI, and mix these data with your own data.

Open Reading Frame Finder (ORF Finder)
A graphical analysis tool that finds all open reading frames in a user's sequence or in a sequence already in the database. Sixteen different genetic codes can be used. The deduced amino acid sequence can be saved in various formats and searched against protein databases using BLAST.

Primer-BLAST
The Primer-BLAST tool uses Primer3 to design PCR primers to a sequence template. The potential products are then automatically analyzed with a BLAST search against user specified databases, to check the specificity to the target intended.

<http://www.ncbi.nlm.nih.gov/guide/all/>

A screenshot of the NCBI Site Map page. On the left, a vertical menu titled "Resource List (A-Z)" is shown with many categories like All Resources, Chemicals & Bioassays, Data & Software, etc. An arrow points from this menu to the main content area. The main content area has a header "Site Map" with a downward arrow pointing to a grid of letters (1 through V). Below the grid, a note says "Featured items are in bold." A section labeled "B" contains links to various NCBI services, with "Basic Local Alignment Search Tool (BLAST)" being the featured item.

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

Site Map

1 A B C D E F G H I J L M N O P R S T U V

Featured items are in bold.

B

Basic Local Alignment Search Tool (BLAST)

Batch Entrez

BioAssay Services

BioProject (formerly Genome Project)

BioProject Submission

BioSample

BioSystems

BLAST (Stand-alone)

BLAST Link (BLink)

BLAST Microbial Genomes

BLAST RefSeqGene

BLAST Tutorials and Guides

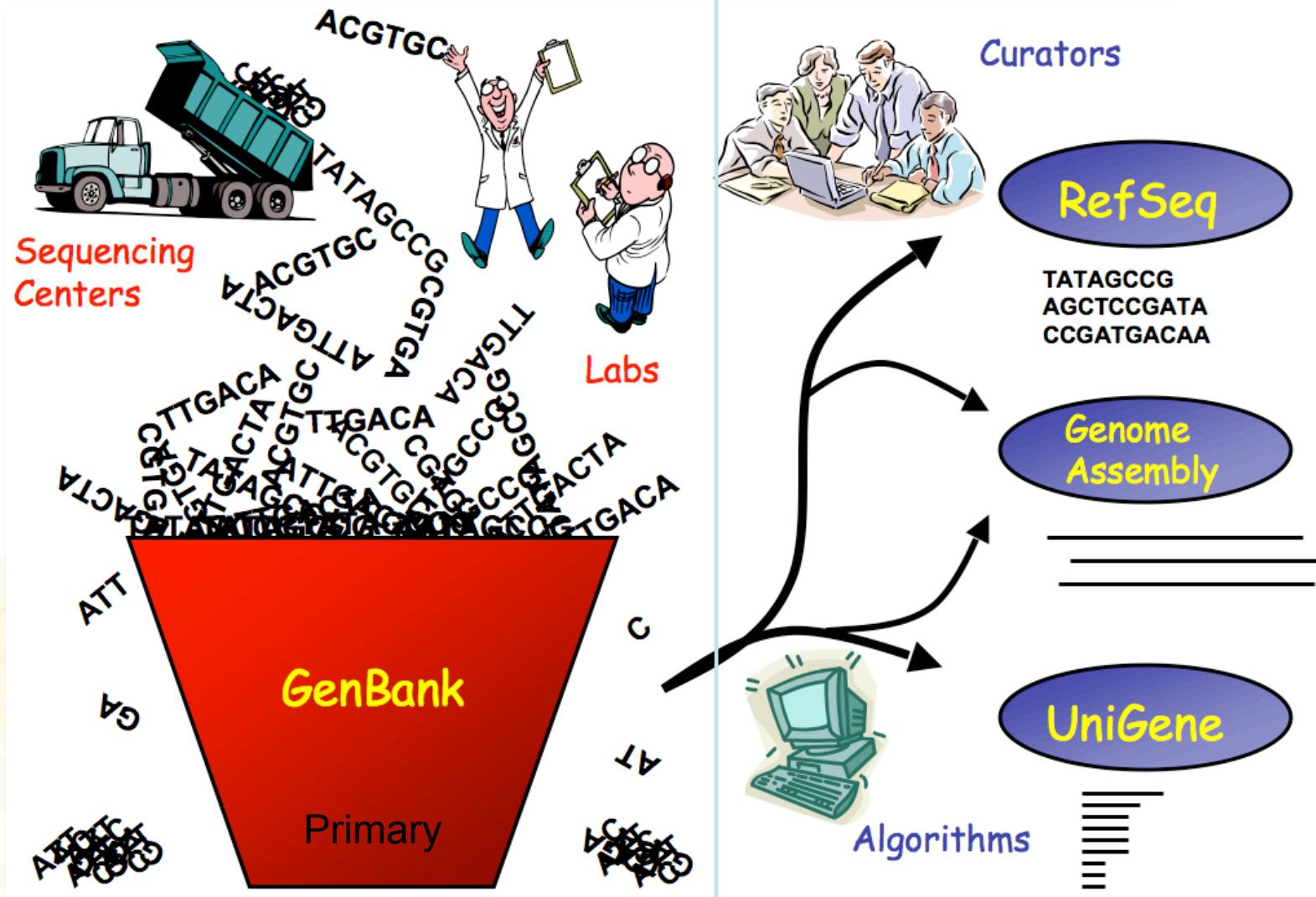
Bookshelf

Batch Entrez

Types of Databases at NCBI

- **Primary Databases**
 - Original submissions by experimentalists
 - Content controlled by the submitter
 - Examples:
 - GenBank
 - SNP
 - GEO
- **Derivative Databases**
 - Built from primary data
 - Content controlled by third party (NCBI)
 - Examples:
 - Refseq
 - RefSNP
 - UniGene
 - NCBI protein, structure and Conserved Domain

Primary Vs Derivative Databases

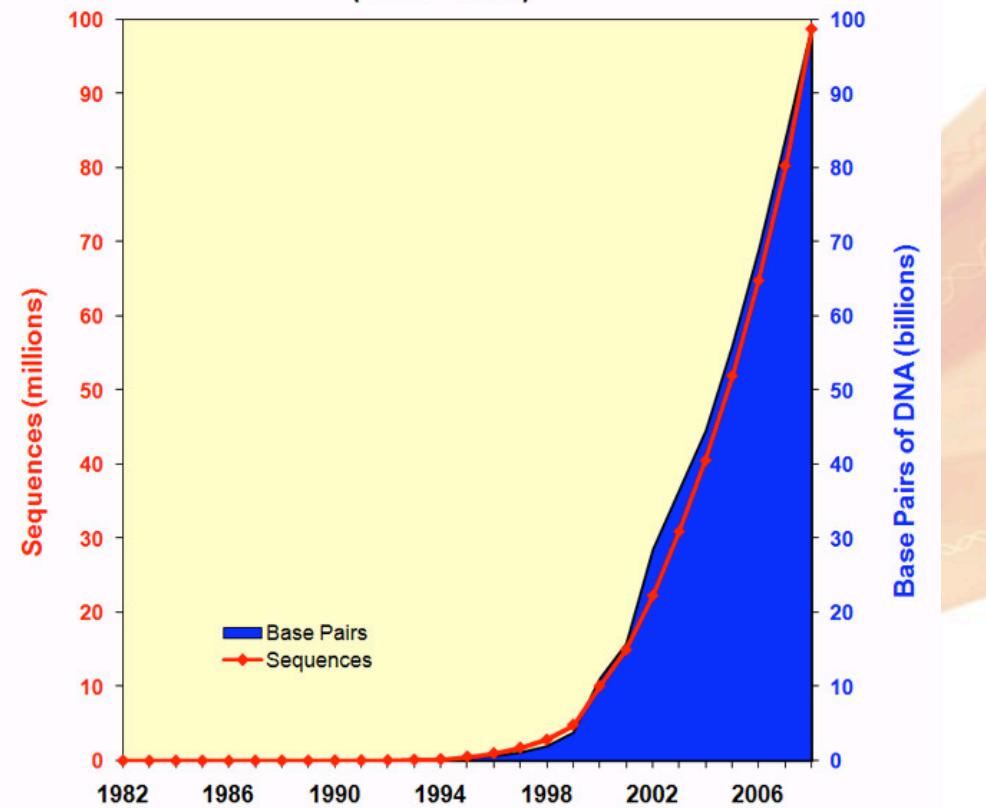


What is GenBank?

- Primary database
- Nucleotide only sequence database
- Archival in nature
 - Historical
 - Reflective of submitter point of view
 - Subjective
 - Redundant
- GenBank Data
 - Direct submissions (traditional [taxonomy] records)
 - Batch (functional) submissions (EST, STS, HTG, etc)
 - ftp accounts (genome data)
- Three collaborating databases
 - GenBank
 - DNA Database of Japan (DDBJ)
 - European Molecular Biology Laboratory (EMBL) Database

Last Year 2012 - Growth of GenBank

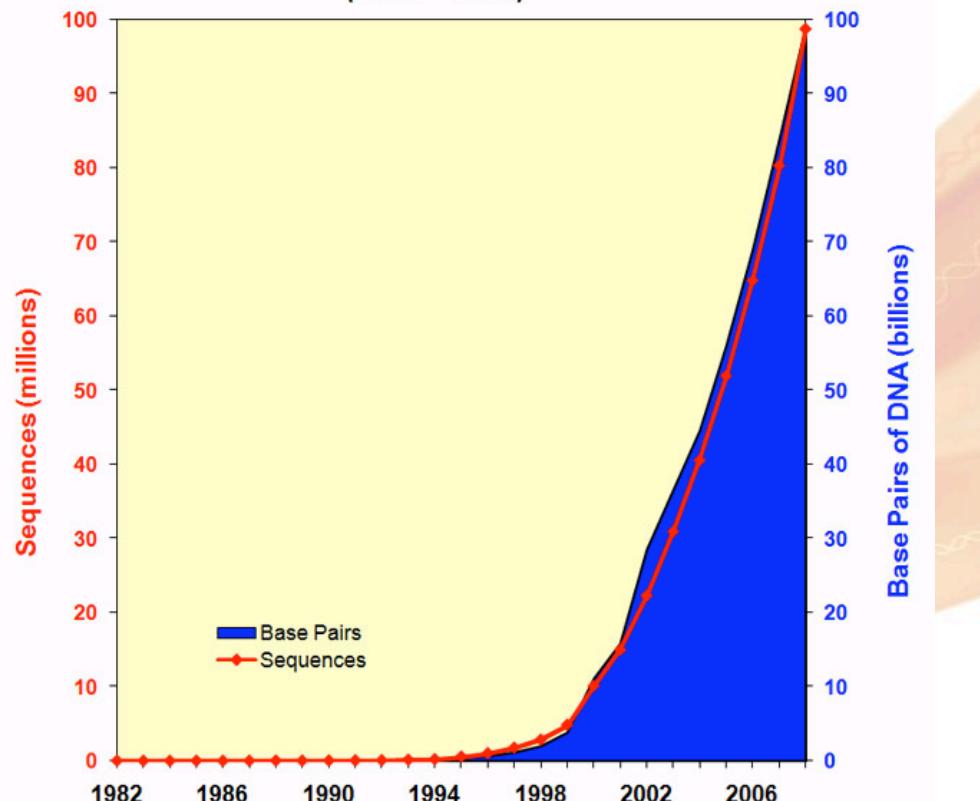
- Get the GBREL.TXT
- Genetic Sequence Data Bank
October 15 2012 NCBI
GenBank Flat File Release **192.0**
Distribution Release Notes
- 157,889,737 loci
- **145,430,961,262 bases**
- **From 157,889,737 reported sequences**



<ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt>

2013 - Growth of GenBank

- Get the GBREL.TXT
- Genetic Sequence Data Bank
October 19 2013 NCBI
GenBank Flat File Release **197.0**
Distribution Release Notes
- 167,295,840 loci
- **154,192,921,011 bases**
- **From 167,295,840 reported sequences**



<ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt>

<ftp://ftp.ncbi.nih.gov/genbank>

Traditional GenBank Record

LOCUS	HSHMLHI	2503 bp	mRNA	linear	PRI	31-MAR-1994
DEFINITION	Human DNA mismatch repair (<i>hmlh1</i>) mRNA, complete cds.					
ACCESSION	U07418					
VERSION	U07418.1 GI:466461					
KEYWORDS	.					
SOURCE	Homo sapiens (human)					
ORGANISM	<u>Homo sapiens</u> Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.					
REFERENCE	1 (bases 1 to 2503)					
AUTHORS	Papadopoulos,N., Nicolaides,N.C., Wei,Y.F., Ruben,S.M., Carter,K.C., Rosen,C.A., Haseltine,W.H., Fleischmann,R.D., Fraser,C.M., Adams,M.D., Venter,J.C., Hamilton,S.R., Petersen,G.M., Watson,P., Lynch,H.T., Peltomaki,P., Mecklin,J.P., Chapelle,A.D., Kinzler,K.W. and Vogelstein,B.					
TITLE	Mutation of a mutL homolog in hereditary colon cancer					
JOURNAL	Science 263 (5153), 1625-1629 (1994)					
MEDLINE	94174309					
PUBMED	8128251					
REFERENCE	2 (bases 1 to 2503)					
AUTHORS	Wei,Y.F.					
TITLE	Direct Submission					
JOURNAL	Submitted (04-MAR-1994) Ying-Fei Wei, Molecular Biology, Human Genome Sciences, Inc., 9620 Medical Center Drive, Rockville, MD 20850, USA					

Traditional GenBank Record

```
LOCUS      HSHMLHI          2503 bp   mRNA    linear    PRI 31-MAR-1994
DEFINITION Human DNA mismatch repair (hmlh1) mRNA, complete cds.
ACCESSION  U07418
VERSION    U07418.1  GI:466461
KEYWORDS   .
SOURCE     Homo sapiens (human)
ORGANISM   Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE  1  (bases 1 to 2503)
```

ACCESSION U07418

VERSION U07418.1 GI:466461

PUBMED [8128251](#)

Version

Tracks changes in sequence

20850, USA

Accession

- Stable
- Reportable
- Universal

GI number

NCBI internal use

Traditional GenBank Record

LOCUS	HSHMLHI	2503 bp	mRNA	linear	PRI 31-MAR-1994
DEFINITION	Human DNA mismatch repair gene		FEATURES	Location/Qualifiers	
ACCESSION	U07418		source	1..2503	
VERSION	U07418.1	GI:4646462		/organism="Homo sapiens"	
KEYWORDS	.			/db_xref="taxon:9606"	
SOURCE	Homo sapiens (human)			/chromosome="3"	
ORGANISM	Homo sapiens			/map="p21"	
	Eukaryota; Metazoa; Mammalia; Eutheria			/tissue_type="gall bladder"	
REFERENCE	1 (bases 1 to 2503)		gene	/dev_stage="adult"	
A				1..2503	
				/gene="hmlh1"	
			CDS	42..2312	
				/gene="hmlh1"	
				/function="DNA mismatch repair"	
				/note="human homolog of E. coli mutL gene product, Swiss-Prot Accession Number P23367"	
				/codon_start=1	
				/protein_id=" AAA17374.1 "	
				/db_xref="GI:466462"	
				/translation="MSFVAGVIRRLLDETVVNRIAAGEVIQRPANA IKE MIENCLDAKSTSIQIVKEGGLKLIQI QDNGTGIRKEDLDIVCERFTTSKLQS FE DLSA I S T Y GFRGE ALASISHVAHVTITTKTADGKCAYRASYSDGKLKA P PK PCAGNQGTQITVEDLFYNTIA TRRKALKNPSEEEYGKILEVVGRYSVHN AGISFSVKKQGETVADVRTLPNASTVDNIRS VFGNAVSRELIEIGCEDKT LA FK MN GYISN AN YSV KKC IF LLF IN HRL VESTSLRKAI ETVYAA YLPK NT H EFL YL SLE I SPQN VDV NVHPTKHEVHF LHE SILERVQQHIESKL LGNSN SSS RMYFTQTLLPGLAGPSGEMVKSTS LTSS STSGSS SDKVYAHQMVR TD SRE QK LDAFLQPLSKPLSS SQPQA IVTEDKTDI SS GRAR QQDE EML PAPA EVA AKN QSLEG DT KGT SEMSE KR GETSSN PRKR HRED SDV EMVED DS RKE MT AACT PRR RI IN LT SVLS LQEEINEQGHEV LREM LHNSFVG CVNPQWALAQHQTKLYLLNNTKLSEELFYQILIY DFANFGV LRLSEPA PLFD LAM LADS PES GWTE EDGP KPEG LAE YIVEFL KKKA EML AD YF SLE IDE EGNL IGPLL IDNYVPPLEG LP I FIL R LATE VN WDEEKE CF E SLS KEC AM FYSIRKQYI SEESTL SGQQ SE VP GSIP NSWK WT VE HIVYK ALR SHIL PP KHFT ED GNI LOLANLPDLYKVF ERC"	

Version
Tracks changes in

20850, USA

Well annotated

Traditional GenBank Record

LOCUS	HSHMLHI	2503 bp	mRNA	linear	PRI	31-MAR-1994				
DEFINITION	Human DNA mismatch repair gene		FEATURES	Location/Qualifiers						
ACCESSION	U07418		source	1..2503						
VERSION	U07418.1	GI:462200		/organism="Homo sapiens"						
KEYWORDS	.			/db_xref="taxon:9606"						
SOURCE	Homo sapiens (human)			/chromosome="X"						
ORGANISM	<u>Homo sapiens</u>			/map="X:1000000..1500000"						
	Eukaryota; Metazoa; Mammalia; Eutheria			BASE COUNT	723 a	539 c	599 g	642 t		
REFERENCE	1 (bases 1 to 2503)			ORIGIN						
				1	gttgaacatc	tagacgttc	ctggcgccaa	aatgtcggtc	gtggcagggg	
				61	ttatttcggcg	gctggacagag	acagtgggtca	accgcatacg	ggcgggggaa	gttatccggc
				121	ggccagctaa	tgttatcaaa	gagatgttgc	agaactgttt	agatgcaaaa	tccacaagta
				181	ttcaagtgat	ttgttaaagag	ggaggccgtt	agttgttca	gatccaagac	atggcaccgg
				241	ggatcggaaa	agaagatctg	gatattgtat	gtgaaagggtt	cactactagt	aaactgcgtt
				301	cctttggggaa	tttagccgtt	atttctaccc	atggctttcg	aggttgaggct	ttggcccgca
				361	taagccatgt	ggctcatgtt	actattacaa	cgaaaacacgc	tgatggaaag	tgtgcataca
				421	gagccatgtt	ctcagatggaa	aaactgttgc	cccctccaa	accatgttgc	ggcaatcaag
				481	ggacccatgtt	cacgttggag	gacccatgtt	acaatcatgc	cacggggaaa	aaagctttaa
				541	aaaaatccaa	tgtttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				601	atgcaggcat	tagtttctca	gttttttttt	tttttttttt	tttttttttt	tttttttttt
				661	tacccaaatgc	cttacccatgtt	gacaattttc	gttttttttt	tttttttttt	tttttttttt
				721	aactttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				781	ccaaatccaa	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				841	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				901	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				961	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				1021	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				1081	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				1141	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				1201	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				1261	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				1321	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				1381	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				1441	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				1501	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				1561	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				1621	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				1681	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				1741	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				1801	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				1861	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				1921	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				1981	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				2041	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				2101	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				2161	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				2221	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				2281	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				2341	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				2401	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt
				2461	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt	tttttttttt

Well annotated

The sequence is the data

What is an Accession Number?

- Label used to identify a sequence
- String of letters and/or numbers that corresponds to a molecular sequence
- Examples (all for retinol-binding protein, RBP4):

X02775
NT_030059
Rs7079946

GenBank genomic DNA sequence
Genomic contig
dbSNP (single nucleotide polymorphism)

DNA

N91759.1
NM_006744

NP_007635
AAC02945
Q28369
1KT7

An expressed sequence tag (1 of 170)
RefSeq DNA sequence (from a transcript)

RefSeq protein
GenBank protein
SwissProt protein
Protein Data Bank structure record

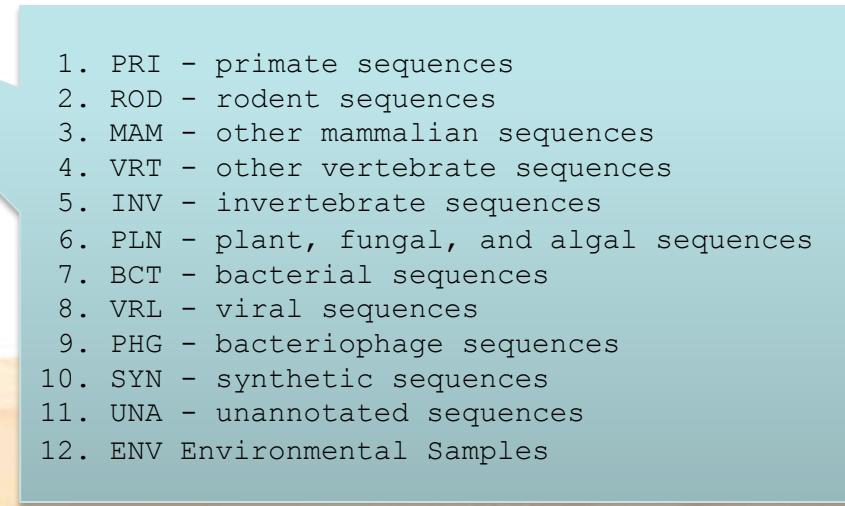
RNA

Protein

- READ: <http://www.ncbi.nlm.nih.gov/Sitemap/sequenceIDs.html>

Organization of GenBank: Traditional (Taxonomy) Divisions

- Records are divided into 20 Divisions
 - 12 Traditional
 - 8 Bulk
- Traditional Divisions:
 - Direct Submissions
 - (Sequin and BankIt)
 - Accurate
 - Well characterized

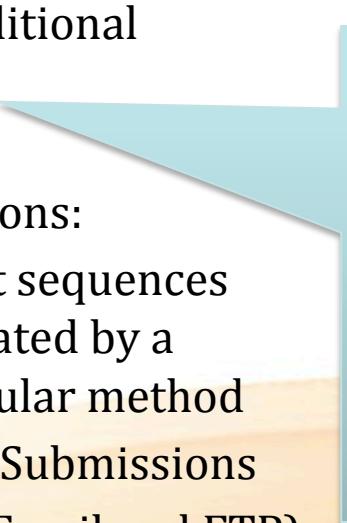
- 
1. PRI - primate sequences
 2. ROD - rodent sequences
 3. MAM - other mammalian sequences
 4. VRT - other vertebrate sequences
 5. INV - invertebrate sequences
 6. PLN - plant, fungal, and algal sequences
 7. BCT - bacterial sequences
 8. VRL - viral sequences
 9. PHG - bacteriophage sequences
 10. SYN - synthetic sequences
 11. UNA - unannotated sequences
 12. ENV Environmental Samples

Entrez query: gbdv_xxx[Properties]

Organization of GenBank: Bulk (Functional) Divisions

- Records are divided into 20 Divisions
 - 12 Traditional
 - 8 Bulk

- Bulk Divisions:
 - collect sequences generated by a particular method
 - Batch Submissions
 - (Email and FTP)
 - Can be Inaccurate
 - Sometimes poorly characterized

- 
1. EST - EST sequences (expressed sequence tags)
 2. PAT - patent sequences
 3. STS - STS sequences (sequence tagged sites)
 4. GSS - GSS sequences (genome survey sequences)
 5. HTG - HTG sequences (high throughput genomic sequences)
 6. HTC - HTC sequences (high throughput cDNA sequences)
 7. WGS - Whole Genome Shotgun (WGS) sequencing projects
 8. TSA - Transcriptome Shotgun Assembly sequences

Entrez query: gbdv_xxx[Properties] e.g. gbdv_est[Properties]

Growth of GenBank Divisions

Division	Description	Release 191 (8/2012)	Annual increase (%) ^a
Taxonomic divisions			
SYN	Synthetic	928 200 038	494.2%
PHG	Phages	84 079 451	34.4%
ENV	Environmental samples	3 374 433 548	32.1%
VRL	Viruses	1 429 464 786	21.1%
BCT	Bacteria	8 439 854 434	21.0%
PLN	Plants	5 481 470 133	15.6%
MAM	Other mammals	863 036 872	6.9%
VRT	Other vertebrates	2 886 594 595	6.7%
PRI	Primates	6 317 656 773	3.3%
UNA	Unannotated	127 803	1.5%
ROD	Rodents	4 435 106 948	0.9%
INV	Invertebrates	2 493 058 927	-1.7%
Functional divisions			
TSA	Transcriptome shotgun data	5 759 588 580	207.3%
WGS	Whole-genome shotgun data	308 196 411 905	47.9%
PAT	Patented sequences	12 118 622 726	8.6%
GSS	Genome survey sequences	21 947 780 105	5.7%
EST	Expressed sequence tags	40 888 051 100	4.8%
HTG	High-throughput genomic	24 359 210 558	0.1%
STS	Sequence tagged sites	636 262 446	0.1%
HTC	High-throughput cDNA	639 165 410	-3.5%
TOTAL	All GenBank sequences	451 278 177 138	33.1%

^aMeasured relative to Release 185 (8/2011).

Current - GenBank Divisions (release 197)

PRI	(46)	Primate
ROD	(31)	Rodent
PLN	(63)	Plant and Fungal
BCT	(106)	Bacteria/Archaea
INV	(35)	Invertebrate
VRT	(31)	Other Vertebrate
VRL	(26)	Viral
MAM	(8)	Mammalian
PHG	(2)	Phage
SYN	(7)	Synthetic
ENV	(62)	Envir. samples
UNA	(1)	Unannotated

EST	(474)	Expressed Sequence Tag
GSS	(278)	Genome Survey Sequence
HTG	(141)	High Throughput Genomic
PAT	(195)	Patent
STS	(20)	Sequence Tagged Site
HTC	(141)	HTC sequences (high throughput cDNA sequences)
TSA	(145)	Transcriptome Shotgun Assembly sequences

“Traditional” (Taxonomy)

- Organized by taxonomy (sort of)
- Direct submissions (Sequin/Bankit)
- Accurate (~1 error per 10,000 bp)
- Well characterized

“Bulk” (Functional)

- Organized by sequence type
- Batch submissions (ftp/email)
- Less accurate
- Poorly characterized

CON* (215) Contigs, virtual

ENV Division

- Environmental sample sequences
- The ENV division of GenBank accommodates sequences obtained via environmental sampling methods in which the source organism is unknown
- Many ENV sequences arise from metagenome samples derived from microbiota in various animal tissues, such as within the gut or skin, or from particular environments, such as freshwater sediment, hot springs or areas of mine drainage
- Environmental sample sequences are generally submitted for whole metagenomic shotgun sequencing experiments or surveys of sequences from targeted genes, like 16S rRNA

GenBank (Functional) Divisions



- **Expressed Sequence Tag**
 - 1st (single) pass single read cDNA
- **Genome Survey Sequence**
 - 1st (single) pass single read gDNA
- **High Throughput Genomic**
 - Incomplete sequences of genomic clones
- **High-Throughput cDNA/mRNA**
 - Similar to EST's but often contain more information
- **Sequence Tagged Site**
 - PCR-based mapping reagents
- **Patent**
 - Sequence which have corresponding patents
- **Transcriptome Shotgun Assembly Sequence Database**
 - Archive of computationally assembled sequences from primary data such as ESTs, traces and Next Generation Sequencing Technologies

*CON

***CON**

CON - Contigs

- Do not contain any sequence data
- Instead, they utilize a CONTIG line type with a join() statement which describes how component sequences can be assembled to form the larger constructed sequence
- CON records are used to represent very long sequences, such as a eukaryotic chromosome, where the sequence is not complete but consists of several contig records with uncharacterized gaps between them
- Rather than listing the sequence itself, CON records contain assembly instructions involving the several component sequences
- An example of such a CON record is CM000663 for human chromosome 1

STS Division

- Is a relatively short, easily PCR-amplified sequence (200 to 500 bp) which can be specifically amplified by PCR and detected in the presence of all other genomic sequences and whose location in the genome is mapped
- <http://www.ncbi.nlm.nih.gov/genome/probe/doc/TechSTS.shtml>

GenBank WGS Projects (The Other Functional Division)

GenBank

WGS

Sequences from WGS sequencing projects are not represented in GenBank release, b/c all WGS project data are made available on a per-project basis

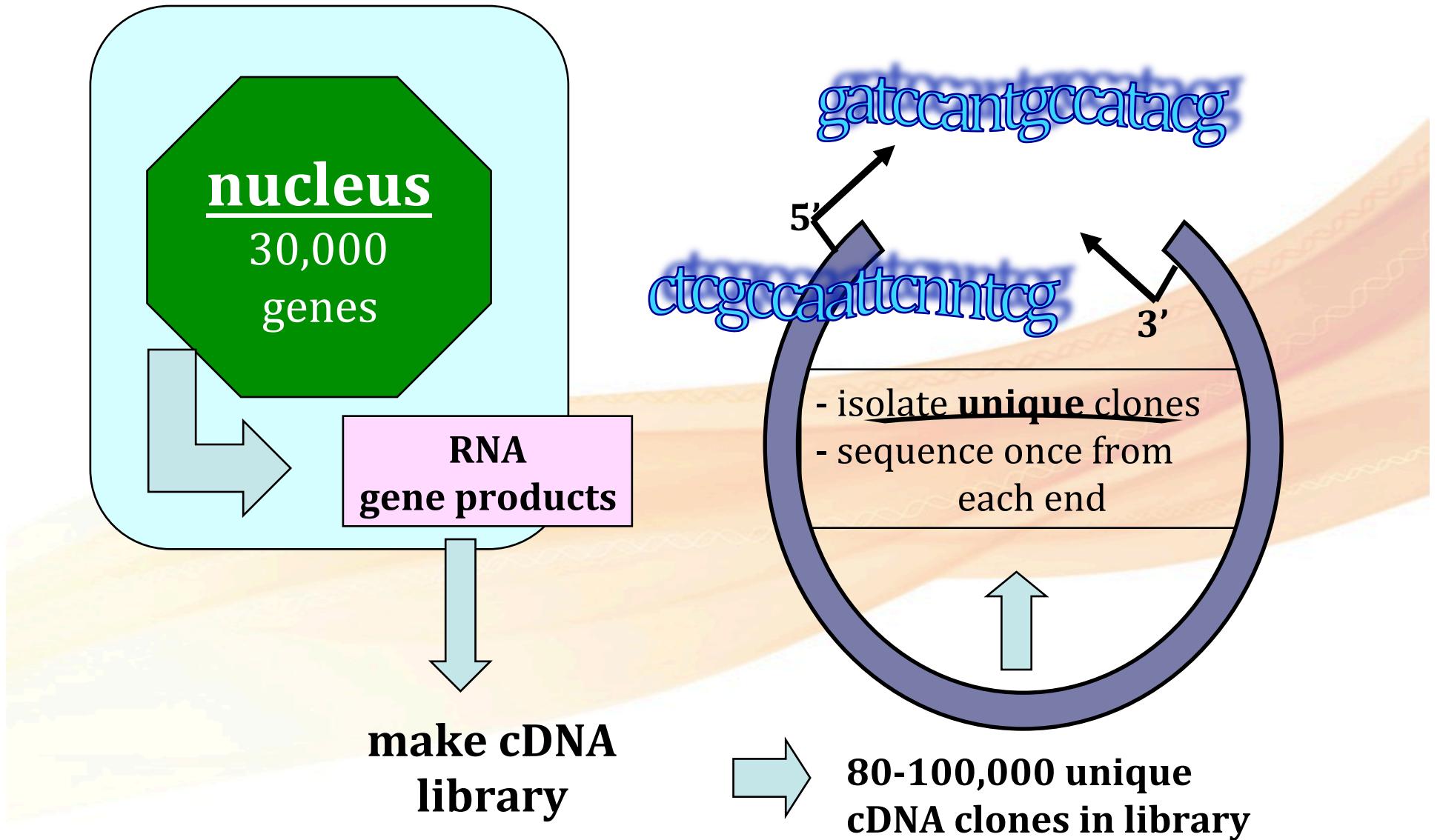
- DDBJ/EMBL/GenBank accepts both complete and incomplete genomes. Whole Genome Shotgun (WGS) sequencing projects are incomplete genomes or incomplete chromosomes that are being sequenced by a whole genome shotgun strategy
- WGS projects may be annotated, but annotation is not required
- The nucleotide data from all WGS projects go into the BLAST wgs database since the fall of 2011
- Proteins from: most WGS projects go into the BLAST nr database
- Proteins from environmental projects are present in either the BLAST nr or env_nr database, depending upon whether that sequence has been identified as a particular organism (nr), or if the organism is not yet known (env_nr)

<http://www.ncbi.nlm.nih.gov/genbank/wgs>

<http://www.ncbi.nlm.nih.gov/projects/WGS/WGSprojectlist.cgi>

<ftp://ftp.ncbi.nih.gov/genbank/wgs/>

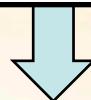
EST Division: Expressed Sequence Tags



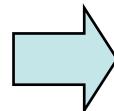
EST Division: Expressed Sequence Tags

```
>IMAGE:275615 5' mRNA sequence
GACAGCATTGGGCCGAGATGTCTCGCTCCGTGGCCTTAGCTGTGCTCGCGCTACTCTCTCTTCTGG
TGGAGGTATCCAGCGTACTCCAAAGATTCAAGGTTACTCACGTACATCCAGCAGAGAATGGAAAGTCAA
TTCCTGAATTGCTATGTGTCTGGGTTCATCCATCCGACATTGAAGTTGACTTACTGAAGAATGGAGA
GAATTGAAAAAGTGGAGCATTCAAGACTTGCTTTCAAGGACTGGTCTTCTATCTCTGTACTAC
TGAATTCACCCCACTGAAAAAGATGAGTATGCCTGCCGTGTTGAACCATGTNGACTTGTACAGNC
AAGTTNAGTTAAGTGGGNATCGAGACATGTAAGGCAGGCATCATGGGAGGTTTGAAGNATGCCGCN
TTGGATTGGGATGAATTCAAATTCTGGTTGCTTGNNTTTTTAATATTGGATATGCTTTG
```

```
>IMAGE:275615 3', mRNA sequence
NNTCAAGTTTATGATTAACTTGTGGAACAAAATAAACAGATTAACCACAAACCATGCCTTA
TTATCAAATGTATAAGANGAAAATATGAATCTTATATGACAAAATGTTCATTCAATTATAACAAATT
AATAATCCTGTCAATNATATTCTAAATTCCCCTAAATTCTAAGCAGAGTATGTAATTGGAGTT
CTTATGCACGCTTAACATCTAACAGCTTGAGTGCAAGAGATTGANGAGTTCAAATCTGACCAAG
GTTGATGTTGGATAAGAGAATTCTCTGCTCCCCACCTCTANGTGCCAGCCCTC
```



make cDNA
library



80-100,000 unique
cDNA clones in library

ESTs in Entrez (query gbdv_est)

▼ Top Organisms [Tree]

[Homo sapiens](#) (8296272)
[Mus musculus](#) (4852144)
[Zea mays](#) (2018634)
[Sus scrofa](#) (1536458)
[Arabidopsis thaliana](#) (1527298)
[Bos taurus](#) (1517145)
[Danio rerio](#) (1481930)
[Glycine max](#) (1422497)
[Xenopus \(Silurana\) tropicalis](#) (1271375)
[Oryza sativa](#) (1248955)
[Ciona intestinalis](#) (1205674)
[Triticum aestivum](#) (1067126)
[Oryza sativa Japonica Group](#) (985283)
[Rattus norvegicus](#) (951258)
[Drosophila melanogaster](#) (821005)
[Xenopus laevis](#) (677806)
[Oryzias latipes](#) (665382)
[Brassica napus](#) (643664)
[Gallus gallus](#) (600075)
[Hordeum vulgare](#) (525775)
All other taxa (31294353)

▼ Top Organisms [Tree]

[Homo sapiens](#) (8314462)
[Mus musculus](#) (4852147)
[Zea mays](#) (2019105)
[Sus scrofa](#) (1621083)
[Bos taurus](#) (1559485)
[Arabidopsis thaliana](#) (1529700)
[Danio rerio](#) (1481936)
[Glycine max](#) (1461436)
[Xenopus \(Silurana\) tropicalis](#) (1271375)
[Oryza sativa](#) (1249965)
[Ciona intestinalis](#) (1205674)
[Rattus norvegicus](#) (1103577)
[Triticum aestivum](#) (1071295)
[Oryza sativa Japonica Group](#) (985283)
[Drosophila melanogaster](#) (821005)
[Xenopus laevis](#) (677806)
[Oryzias latipes](#) (665382)
[Brassica napus](#) (643947)
[Gallus gallus](#) (600423)
[Panicum virgatum](#) (546245)
All other taxa (34512129)

▼ Top Organisms [Tree]

[Homo sapiens](#) (8315272)
[Mus musculus](#) (4853562)
[Zea mays](#) (2019114)
[Sus scrofa](#) (1624141)
[Bos taurus](#) (1559494)
[Arabidopsis thaliana](#) (1529700)
[Danio rerio](#) (1488275)
[Glycine max](#) (1461624)
[Xenopus \(Silurana\) tropicalis](#) (1271375)
[Oryza sativa](#) (1252989)
[Ciona intestinalis](#) (1205674)
[Rattus norvegicus](#) (1103577)
[Triticum aestivum](#) (1073845)
[Oryza sativa Japonica Group](#) (987318)
[Drosophila melanogaster](#) (821005)
[Panicum virgatum](#) (720590)
[Xenopus laevis](#) (677806)
[Oryzias latipes](#) (666891)
[Brassica napus](#) (643937)
[Gallus gallus](#) (600433)
All other taxa (38364698)

▼ Top Organisms [Tree]

[Homo sapiens](#) (8699560)
[Mus musculus](#) (4853570)
[Zea mays](#) (2019137)
[Sus scrofa](#) (1669420)
[Bos taurus](#) (1559495)
[Arabidopsis thaliana](#) (1529700)
[Danio rerio](#) (1488275)
[Glycine max](#) (1461624)
[Triticum aestivum](#) (1286283)
[Xenopus \(Silurana\) tropicalis](#) (1271480)
[Oryza sativa](#) (1253557)
[Ciona intestinalis](#) (1205674)
[Rattus norvegicus](#) (1103577)
[Oryza sativa Japonica Group](#) (987327)
[Drosophila melanogaster](#) (821005)
[Panicum virgatum](#) (720590)
[Xenopus laevis](#) (677911)
[Oryzias latipes](#) (666891)
[Brassica napus](#) (643944)
[Gallus gallus](#) (600434)
All other taxa (40394541)

Less...

2009 - Oct

2010 - Oct

2011 - Oct

2012 - Oct

EST Division

- ESTs continue to be a major source of data for gene expression and annotation studies
- At almost 41 billion base pairs, it remains the largest non-WGS division in GenBank
- EST data are available for download from
- <ftp.ncbi.nlm.nih.gov/repository/dbEST/>
- The data in dbEST are clustered using the BLAST programs to produce the UniGene database (www.ncbi.nlm.nih.gov/unigene)
 - >5.8 million gene-oriented sequence clusters
 - representing 142 organisms

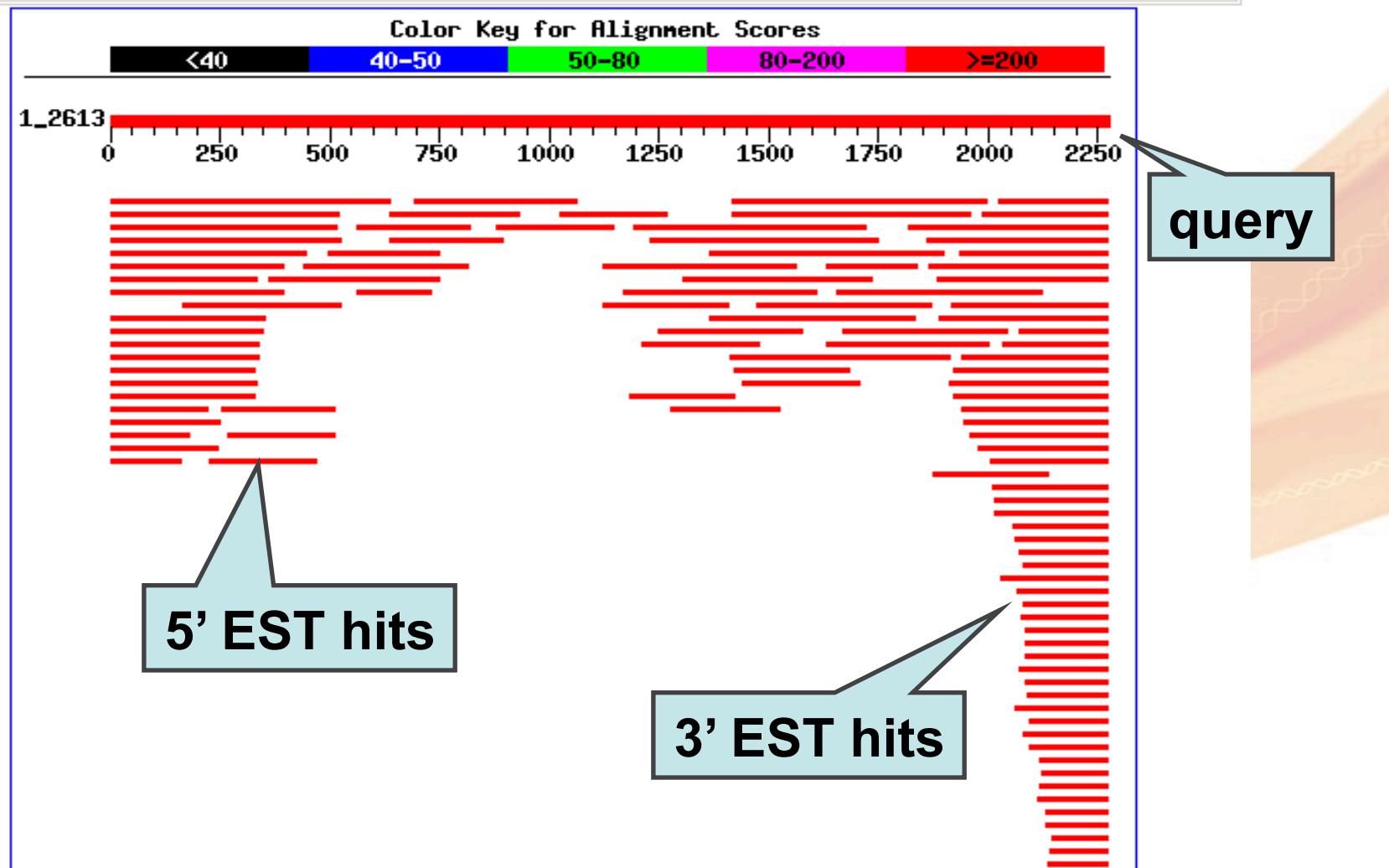
GSS Division

- Similar to the EST division – exception, most of the sequences are genomic in origin, rather than cDNA (mRNA)
- It should be noted that two classes (exon trapped products and gene trapped products) may be derived via a cDNA intermediate
 - Care should be taken when analyzing sequences from either of these classes, as a splicing event could have occurred and the sequence represented in the record may be interrupted when compared to genomic sequence
- The GSS division contains (but is not limited to) the following types of data:
 - random "single pass read" genome survey sequences.
 - cosmid/BAC/YAC end sequences
 - exon trapped genomic sequences
 - Alu PCR sequences
 - transposon-tagged sequences

A Cluster of ESTs

Distribution of 118 Blast Hits on the Query Sequence

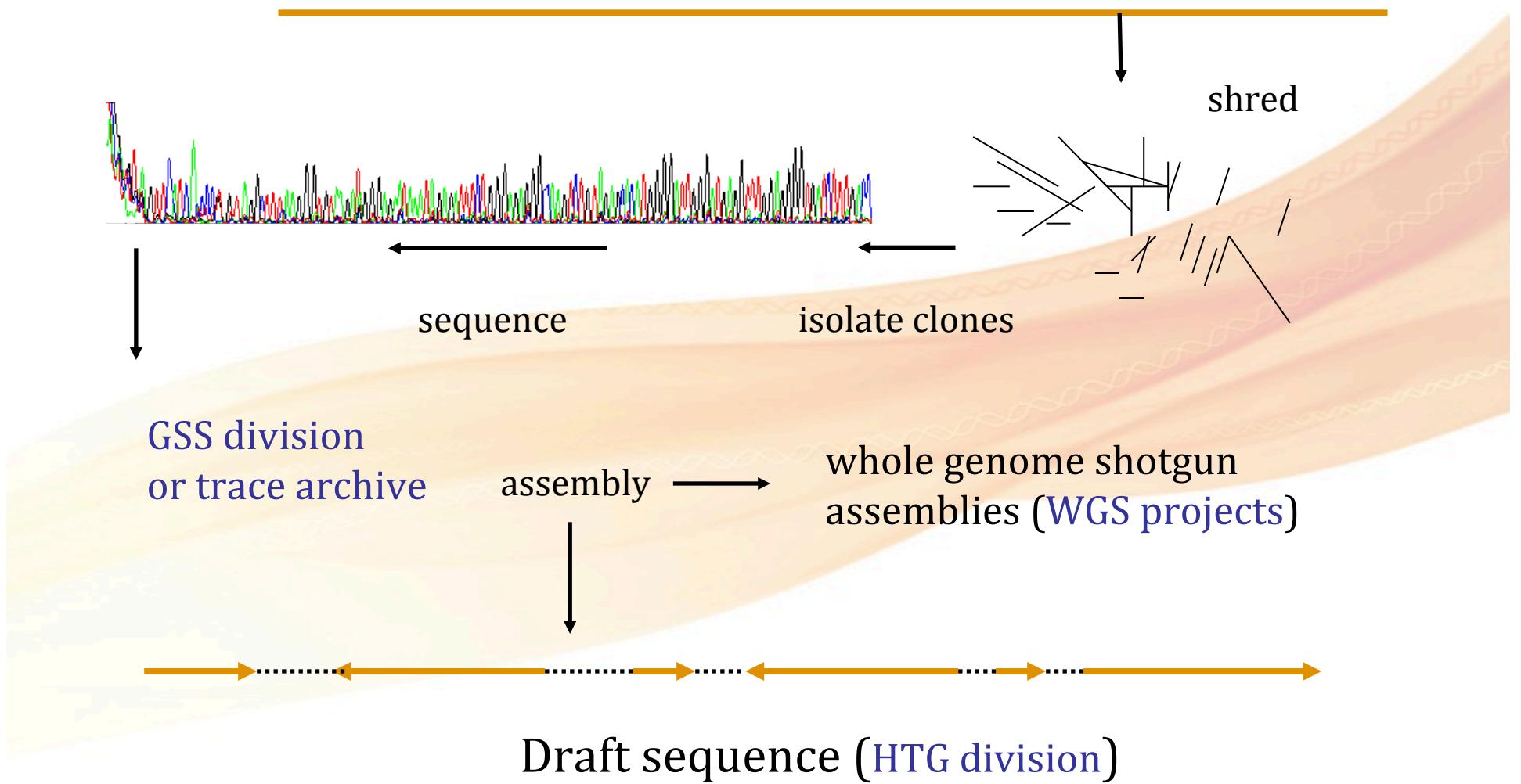
AV823219 AV823219 RAFL5 Arabidopsis thaliana cDNA clone RAFL05-16..S=1225 E=0.0



GenBank Bulk Sequence: EST

GSS, HTG, WGS

Whole BAC insert (or genome)



HTG Division

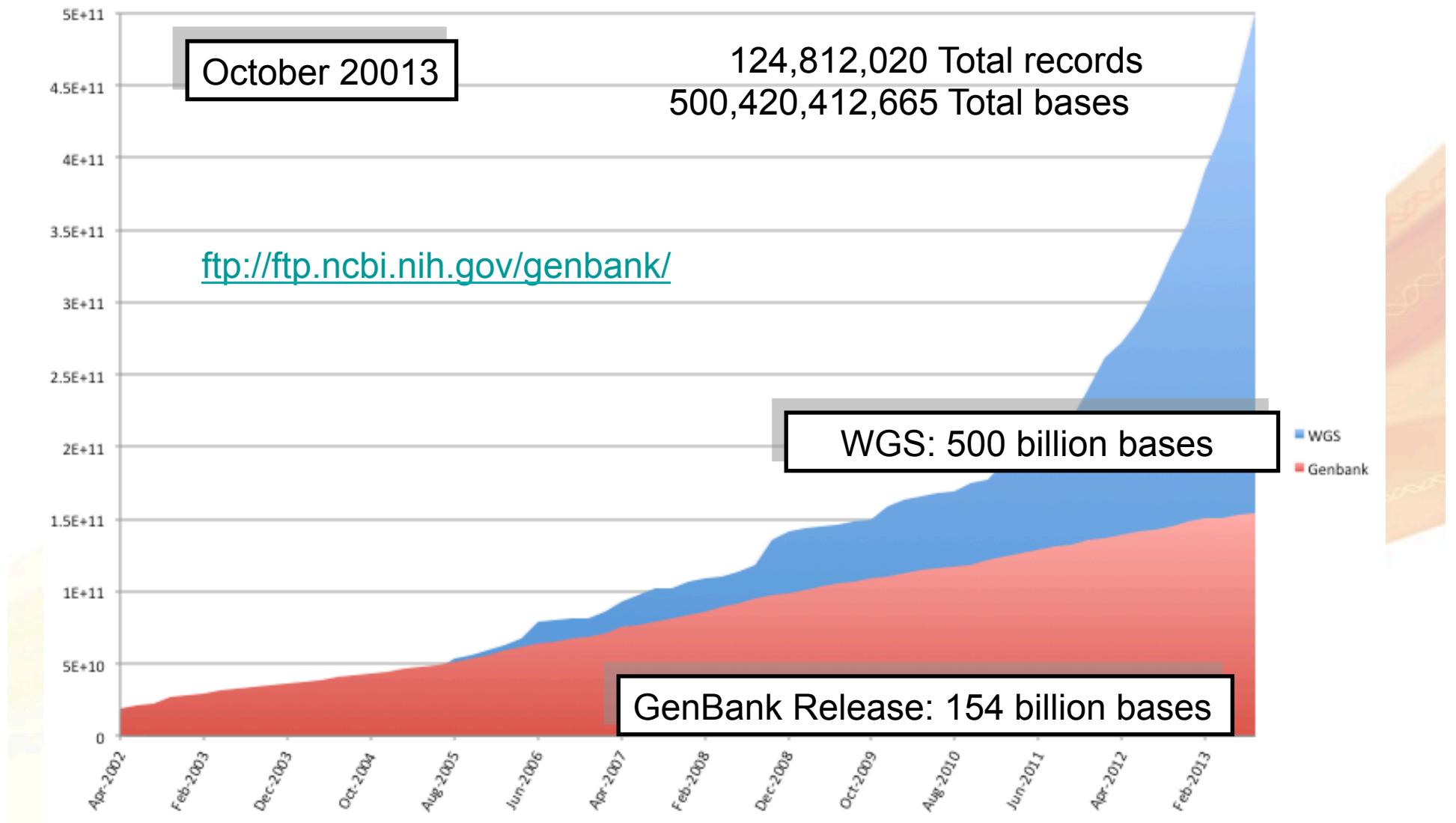
- The HTG division of GenBank (www.ncbi.nlm.nih.gov/genbank/htgs/)
- Contains unfinished large-scale genomic records, which are in transition to a finished state
- These records are designated as belonging to Phases 0 to 3 depending on the quality of the data, with Phase 3 being the finished state
- On reaching Phase 3, HTG records are moved into the appropriate organism division of GenBank

HTC Division

- The HTC division of GenBank contains high- throughput cDNA sequences that are of draft quality but may contain:
 - 5' untranslated regions (UTRs)
 - 3' UTRs, partial coding regions and introns
- HTC sequences that are finished and of high quality are moved to the appropriate organism division of GenBank
- A project generating HTC data is described in (9)*

Kawai,J., Shinagawa,A., Shibata,K., Yoshino,M., Itoh,M.,
Ishii,Y., Arakawa,T., Hara,A., Fukunishi,Y., Konno,H. et al. (2001) Functional annotation of a full-length mouse cDNA collection. *Nature*, 409, 685–690.

The Growth of WGS



Shotgun Assembly Sequences: Genome (WGS)

#	Prefix	Div	BioProject	BioSample	Organism	Contigs			Scaffolds			# of Proteins	Complete	Update Date	Create Date
						#	Mbases	Annot	#	Mbases	Annot				
1	AAAA02	PLN	PRJNA361		Oryza sativa Indica Group	50,231	410.7		3,095	769.7	Y	37,358		2011-02-25	2004-10-21
2	AAAB01	INV	PRJNA1438		Anopheles gambiae str. PEST	69,724	328.9	Y	5	230.5		14,089		2011-05-19	2002-03-22
3	AAAC01	BCT	PRJNA299		Bacillus anthracis str. A2012	1	5.1							2007-01-17	2002-05-09
4	AAAK03	BCT	PRJNA71		Enterococcus faecium DO	163	2.8	Y				2,721		2007-01-17	2004-06-07

- Each WGS project assigned a stable 4-letter WGS accession prefix, which does not change as the project is updated
- Ex. WGS accession number is **XXXX00000000**, project's first assembly version would be **XXXX01000000**, and first contig of that version: **XXXX01000001**
- When there is more sequencing and the genome is reassembled
 - Contigs are submitted as the 02 version of the WGS project
 - No linkage or relationship is expected between the old and new contigs, and the new contigs are given new accession numbers beginning with **XXXX02000001**
 - The 01 contigs are suppressed when the 02 contigs are released

<http://www.ncbi.nlm.nih.gov/Traces/wgs/>

<http://www.ncbi.nlm.nih.gov/genbank/wgs>

TSA Division

- The HTC division of GenBank contains high- throughput cDNA sequences that are of draft quality but may contain:
 - 5' untranslated regions (UTRs)
 - 3' UTRs, partial coding regions and introns
- HTC sequences that are finished and of high quality are moved to the appropriate organism division of GenBank
- A project generating HTC data is described in (9)*

Kawai,J., Shinagawa,A., Shibata,K., Yoshino,M., Itoh,M.,
Ishii,Y., Arakawa,T., Hara,A., Fukunishi,Y., Konno,H. et al. (2001) Functional annotation of a full-length mouse cDNA collection. *Nature*, 409, 685–690.

Shotgun Assembly Sequences: Transcriptome (TSA)

#	Prefix	Div	BioProject	BioSample	Organism	Contigs			Scaffolds			# of Proteins	Complete	Update Date	Create Date
						#	Mbases	Annot	#	Mbases	Annot				
1	GAAA01	VRT	PRJNA77699	SRS283232	Latimeria chalumnae	93,507	72.6							2012-09-27	2012-05-15
2	GAAB01	INV	PRJNA173782		Agrilus planipennis	11,957	6.1							2012-12-10	2012-12-10
3	GAAC01	PLN	PRJNA167330		Camellia sinensis var. sinensis	52,919	23.9							2012-09-25	2012-09-25
4	GAAD01	VRT	PRJNA173219	SRS358712	Misgurnus anguillicaudatus	11,327	9.9							2013-03-11	2013-03-11
5	GAAF01	INV	PRJNA173897		Macrosiphum euphorbiae	550	.7							2012-09-26	2012-09-26

- Each TSA project is assigned a stable 4-letter TSA accession prefix, which does not change as the project is updated
- E.g. TSA accession number is XXXX00000000, then that project's first transcript version would be XXXX01000000, and the first assembly of that version would be XXXX01000001
- When a project is reassembled
 - New assemblies are submitted as the 02 version of the TSA project
 - No linkage or relationship is expected between the old and new assemblies, and the new assemblies are given new accession numbers beginning with XXXX02000001
 - The 01 transcripts are suppressed when the 02 transcripts are released.

<http://www.ncbi.nlm.nih.gov/Traces/wgs/>

<http://www.ncbi.nlm.nih.gov/genbank/tsa>

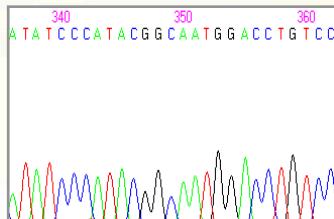
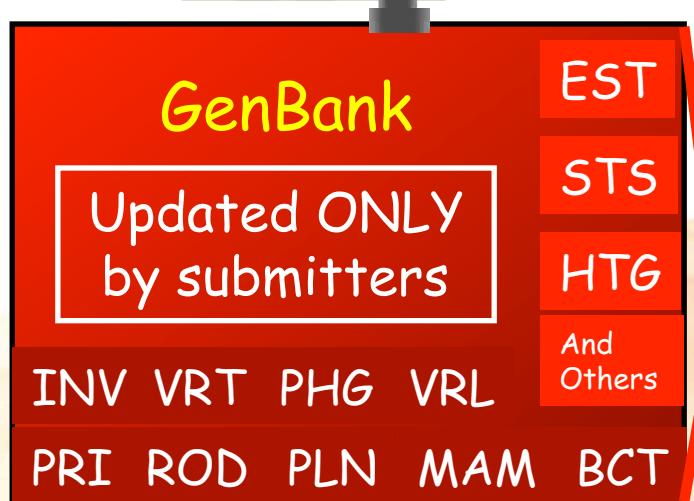
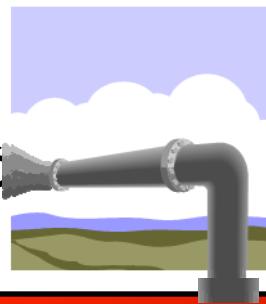
FYI - Obtaining GenBank by FTP

- NCBI distributes GenBank releases in the traditional flat file format as well as in the ASN.1 format used for internal maintenance
- The full bimonthly GenBank release along with the daily updates
 - Which incorporate sequence data from EMBL-Bank and DDBJ
 - Available by anonymous FTP from NCBI at <ftp.ncbi.nlm.nih.gov/genbank>
- GenBank is also available for high-speed download using an Aspera client:
 - www.ncbi.nlm.nih.gov/public/
- Data are partitioned into multiple files; for **release 191**, there are **1852** files requiring **604GB** of uncompressed disk storage
- A script is provided in <ftp.ncbi.nlm.nih.gov/genbank/tools/> to convert a set of daily updates into a cumulative update

Derivative Databases

Sequencing
Centers

CTT
ATT
GCT
TAC
ACT
TAT



Labs

Why Make Reference
Sequences?

RefSeq: NCBI Reference Sequence Database

A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.

Why Make Reference Sequences?

Entrez Nucleotide query:
human[organism] AND lipase[title]



Results: 1 to 20 of 294

<< First < Prev Page 1 Next > Last >>

- [lipoprotein lipase LPL {promoter} \[human, peripheral blood leukocytes, familial combined hyperlipidemia FCHL, mRNA Partial Mutant.](#)

1. [123 nt](#)

123 bp linear mRNA

S78266.1 GI:999330

[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

- [LPL \(Arita\)=lipoprotein lipase \[human, Genomic Mutant, 18 nt\]](#)

2. 18 bp linear DNA

S81338.1 GI:245362

[GenBank](#) [FASTA](#) [Graphics](#)

- [lipoprotein lipase {exon 9} \[human, Genomic Mutant, 18 nt\]](#)

3. 18 bp linear DNA

S75964.1 GI:242490

[GenBank](#) [FASTA](#) [Graphics](#)

....

325 records as of Oct 2013

Entrez Nucleotide query:
human[organism] AND lipase[title]

Results: 8

- [Homo sapiens lipase, endothelial \(LIPG\) gene, complete cds](#)

1. 34,815 bp linear DNA
EU332856.1 GI:166706794
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

34815 bp

- [Homo sapiens lipase, endothelial \(LIPG\), mRNA](#)

2. 4,143 bp linear mRNA
NM_006033.2 GI:62422575
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

4143 bp

- [Homo sapiens lipase, endothelial, mRNA \(cDNA clone MGC:71687 IMAGE:30338950\), complete cds](#)

3. 4,150 bp linear mRNA
BC060825.1 GI:38174525
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

4150 bp

- [Homo sapiens endothelial lipase mRNA, complete cds](#)

4. 3,927 bp linear mRNA
AF118767.1 GI:4836418
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

3927 bp

- [Homo sapiens cDNA FLJ57818 complete cds, highly similar to Endothelial lipase precursor \(EC 3.1.1.3\)](#)

5. 1,760 bp linear mRNA
AK300333.1 GI:194390641
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

1760 bp

- [Homo sapiens cDNA, FLJ96256, highly similar to Homo sapiens lipase, endothelial \(LIPG\), mRNA](#)

6. 1,783 bp linear mRNA
AK315252.1 GI:164694848
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

1783 bp

- [Homo sapiens cDNA FLJ77993 complete cds, highly similar to Homo sapiens endothelial lipase mRNA](#)

7. 1,975 bp linear mRNA
AK291799.1 GI:158257029
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

1975 bp

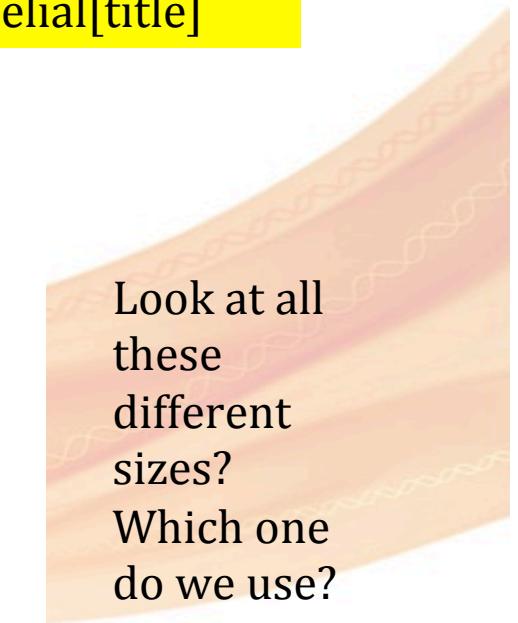
- [ANTISENSE MODULATION OF ENDOTHELIAL LIPASE EXPRESSION](#)

8. 3,927 bp linear DNA
DD265273.1 GI:109245105
[GenBank](#) [FASTA](#) [Graphics](#)

....

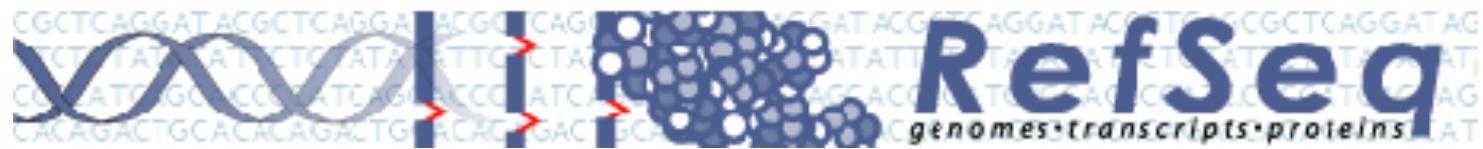
3927 bp

human[organism] AND
lipase[title] AND
endothelial[title]



Look at all
these
different
sizes?
Which one
do we use?

11 records as of Oct 2013



genomes

transcripts

proteins

September 11, 2009: RefSeq Release 37 available for FTP

This release includes:

Proteins: 8,835,796

Organisms: 9,005

Available at: <ftp://ftp.ncbi.nih.gov/refseq/release/>

To receive
incremental
announcements

September 16, 2010: RefSeq Release 43 available for FTP

This release includes:

Proteins: 11,223,078

Organisms: 10,854

Available at: <ftp://ftp.ncbi.nih.gov/refseq/release/>

To receive
incremental
announcements

September 13, 2011: RefSeq Release 49 available for FTP

This release includes:

Proteins: 13,137,813

Organisms: 16,248

Available at:

To receive announcements
of future RefSeq releases and
incremental large updates please subscribe to NCBI's refseq-announce mail list: refseq-announce

September 21, 2012: RefSeq Release 55 available for FTP

This release includes:

Proteins: 17,368,769

Organisms: 17,994

Available at: <ftp://ftp.ncbi.nlm.nih.gov/refseq/release/>

To receive announcements of future RefSeq releases and
incremental large updates please subscribe to NCBI's refseq-announce mail list: refseq-announce

Announcements

September 17, 2013
RefSeq release 61 available for
FTP

This release includes:

Proteins 33,139,114

Organisms 29,414

Available at

<ftp://ftp.ncbi.nlm.nih.gov/refseq/release/>

Documentation [Release Notes](#)

Additional information is available in
[previous announcements](#), follow [NCBI](#)
[on Twitter](#), or subscribe to [NCBI's](#)
[refseq-announce mail list](#).

Growth!!!

<ftp://ftp.ncbi.nih.gov/refseq/release/release-notes/RefSeq-release61.txt>

RefSeq

A subdirectory exists for each sub-section as follows:

Release 55

complete	17994
fungi	332
invertebrate	934
microbial	11368
mitochondrion	3071
plant	285
plasmid	1258
plastid	287
protozoa	157
vertebrate_mammalian	395
vertebrate_other	1474
viral	3011

Release 61

complete	29414
fungi	812
invertebrate	1130
microbial	20859
mitochondrion	3896
plant	363
plasmid	1562
plastid	372
protozoa	182
vertebrate_mammalian	608
vertebrate_other	1830
viral	3629

<ftp://ftp.ncbi.nih.gov/refseq/release/release-notes/RefSeq-release61.txt>

The Reference Sequence – RefSeq

- Aims provide a:
 - Comprehensive
 - Integrated
 - Non-redundant
 - Well-annotated set of sequences,
 - Including genomic DNA, transcripts, and proteins
- Foundation for medical, functional, and diversity studies
- NCBI provides RefSeqs for taxonomically diverse organisms including **eukaryotes, bacteria, and viruses**
- Records are added to the collection as data become publicly available

RefSeq Scope and Access

- Scope
 - NCBI provides RefSeqs for taxonomically diverse organisms including eukaryotes, bacteria, and viruses
 - Additional records are added to the collection as data become publicly available
- **Data Access and Availability**
 - RefSeq is accessible via [BLAST](http://www.ncbi.nlm.nih.gov/blast/), Entrez, and the [NCBI FTP site](http://ftp.ncbi.nlm.nih.gov)
 - Information is also available in Entrez Genomes and Entrez Gene, and for some genomes additional information is available in the Map Viewer
 - Special properties have been defined to facilitate Entrez-based retrieval
 - Next Slide

<http://www.ncbi.nlm.nih.gov/blast/>

[ftp://ftp.ncbi.nlm.nih.gov](http://ftp.ncbi.nlm.nih.gov)

Entrez Queries to Retrieve Sets of RefSeq Records

Query	Accession prefix	RefSeq status retrieved
srcdb_refseq[prop]	All RefSeq accessions	All
srcdb_refseq_known[prop]	NC_, AC_, NG_, NM_, NR_, NP_, AP_	REVIEWED, PROVISIONAL, PREDICTED, INFERRED, and VALIDATED
srcdb_refseq_reviewed[prop]	NC_, AC_, NG_, NM_, NR_, NP_, AP_	REVIEWED
srcdb_refseq_validated[prop]	NC_, NM_, NR_, NP_	VALIDATED
srcdb_refseq_provisional[prop]	NC_, AC_, NG_, NM_, NR_, NP_, AP_	PROVISIONAL
srcdb_refseq_predicted[prop]	NM_, NR_, NP_	PREDICTED
srcdb_refseq_inferred[prop]	AC_, AP_, NM_, NR_, NP_	INFERRED
srcdb_refseq_model[prop]	NT_, NW_, XM_, XR_, XP_, ZP_	Genome annotation models

human[organism] AND srcdb_refseq[prop]

human[organism] AND srcdb_refseq_known[prop]

Reference Sequence: RefSeq

- Every sequence is assigned:
 - A stable accession
 - Version
 - And GI
 - All older versions remain available over time
 - RefSeq accessions have a distinct format
 - **Underscore ("_")** is the primary distinguishing feature of a RefSeq accession
- | <u>Accession</u> | <u>Method</u> | <u>Sequence Type</u> |
|------------------|---------------|--|
| NM_123456 | Mixed | mRNA |
| NM_123456789 | | |
| NP_123456789 | Mixed | protein, from NM_ |
| NR_123456 | Mixed | non-coding RNA |
| XM_123456 | Automated | predicted mRNA |
| XM_123456789 | | |
| XP_123456 | Automated | predicted protein |
| XP_123456789 | | |
| XR_123456 | Automated | predicted non-coding RNA |
| ZP_12345678 | Automated | predicted protein from NZ_ |
| AP_123456 | Mixed | Protein products; alternate protein record |
| YP_123456 | Mixed | Protein products; no corresponding transcript record provided |
| YP_123456789 | | |
| NC_123456 | Mixed | genomic, e.g., chromosomes |
| NG_123455 | Mixed | genomic, incomplete region |
| NT_123456 | Automated | genomic, BAC assembly |
| NW_123456 | Automated | genomic, WGS assembly |
| NW_123456789 | | |
| NZ_ABCD12345678 | Automated | genomic, WGS collection |
| NS_123456 | Automated | genomic, assembly which does not reflect the structure of a real biological molecule |
| AC_123456 | Mixed | genomic, Alternate complete genomic molecule |

<http://www.ncbi.nlm.nih.gov/RefSeq/key.html#accession>

Method

- *Mixed:*
 - Indicates the process flow includes both automated processing and expert review for some of the records
 - Curation analysis may be provided either by NCBI staff or collaborators.
- *Automated:*
 - Indicates records that are not individually reviewed
 - Updates are released in bulk for a genome

COMMENT = Status

- Indicates Status of the record and the GenBank sequence data that was used to provide the record
- May identify a collaboration that supplied the defining sequence information for the genome, gene, or protein
- Level of curation may differ b/t different collaborating groups

DNA topoisomerase III [Pseudoalteromonas flavipulchra JG1]

NCBI Reference Sequence: ZP_11232112.1

[FASTA](#) [Graphics](#)

[Go to:](#)

LOCUS ZP_11232112 640 aa linear BCT 19-OCT-2012

DEFINITION DNA topoisomerase III [Pseudoalteromonas flavipulchra JG1].

ACCESSION ZP_11232112

VERSION ZP_11232112.1 GI:409203909

DBSOURCE REFSEQ: accession [NZ AJMP01000069.1](#)

KEYWORDS .

SOURCE Pseudoalteromonas flavipulchra JG1

ORGANISM [Pseudoalteromonas flavipulchra JG1](#)

Bacteria; Proteobacteria; Gammaproteobacteria; Alteromonadales;
Pseudoalteromonadaceae; Pseudoalteromonas.

WGS REFSEQ: This record is provided to represent a collection of
whole genome shotgun sequences. The reference sequence was derived
from [AJMP01000069](#).

Method: conceptual translation.

COMMENT

Status Keys

Code	Description
MODEL	The RefSeq record is provided by the NCBI Genome Annotation pipeline and is not subject to individual review or revision between annotation runs.
INFERRED	The RefSeq record has been predicted by genome sequence analysis, but it is not yet supported by experimental evidence. The record may be partially supported by homology data.
PREDICTED	The RefSeq record has not yet been subject to individual review, and some aspect of the RefSeq record is predicted.
PROVISIONAL	The RefSeq record has not yet been subject to individual review. The initial sequence-to-gene association has been established by outside collaborators or NCBI staff.
REVIEWED	The RefSeq record has been reviewed by NCBI staff or by a collaborator. The NCBI review process includes assessing available sequence data and the literature. Some RefSeq records may incorporate expanded sequence and annotation information.
VALIDATED	The RefSeq record has undergone an initial review to provide the preferred sequence standard. The record has not yet been subject to final review at which time additional functional information may be provided.
WGS	The RefSeq record is provided to represent a collection of whole genome shotgun sequences. These records are not subject to individual review or revisions between genome updates.

http://www.ncbi.nlm.nih.gov/books/NBK21091/table/ch18.T.refseq_status_codes/?report=objectonly

Why is RefSeq So Useful

- RefSeq provides a biologically **non-redundant set of sequences** for database searching and gene characterization
- Provides a valuable sequence resource for:
 - Similarity searching
 - Gene identification
 - Protein classification
 - Comparative genomics
 - Selection of probes for gene expression
- Acts as molecular "white pages"
 - Provides:
 - a **single**, uniform point of access for searching at the sequence level
 - by **connecting** the results with a diversity of organism-specific databases

Non-redundant Set of Sequences

- Provides an objective and experimentally verifiable definition of "non-redundant" in supplying one example of each natural biomolecule per organism
- Small amount of sequence redundancy introduced from:
 - Close paralogs
 - Alternate splicing products
 - Genome assembly intermediates

Additional RefSeq Benefits

- Updates to reflect current sequence data and biology
- Data validation (next slide)
- Format consistency (next slide)
- Distinct accession series
- Stewardship by NCBI staff and collaborators

Data Validation (1)

- Sequences are validated in several ways
- To confirm that genomic sequence from the region of the mRNA feature really does match the mRNA sequence itself
- That the annotated coding region features really can be translated into the protein sequences they refer to
- Checks for valid ASN.1 format
- Ensures that consistency is maintained in descriptive information (symbols, gene and protein names) between RefSeq and Gene records

Data Validation (2)

- Each molecule is annotated as accurately as possible with:
 - Correct organism name
 - Correct gene symbol for that organism
 - Reasonable names for proteins where possible
- When available, nomenclature provided by official nomenclature groups is used
- **Note that gene symbols are not required or expected to be unique either across species or within a species**

RefSeq: NCBI's Derivative Sequence Database

- Curated transcripts and proteins
 - Reviewed
 - Human, mouse, rat, fruit fly, zebrafish, arabidopsis, microbial genomes, and more
- Model transcripts and proteins
- Assembled Genomes Regions (contigs)
- Chromosome records
 - Human genome
 - microbial
 - organelle
- <ftp://ftp.ncbi.nih.gov/refseq/release>

srcdb_refseq[Properties] = [5,846,009](#) nt, [17,698,431](#) protein, [10,965,669](#) gene
(Oct 2010)

Reference Sequence: RefSeq

<u>Accession</u>	<u>Method</u>	<u>Sequence Type</u>
NM_123456	Mixed	mRNA
NM_123456789		
NP_123456789	Mixed	protein, from NM_
NR_123456	Mixed	non-coding RNA
XM_123456	Automated	predicted mRNA
XM_123456789		
XP_123456	Automated	predicted protein
XP_123456789		
XR_123456	Automated	predicted non-coding RNA
ZP_12345678	Automated	predicted protein from NZ_
AP_123456	Mixed	Protein products; alternate protein record
YP_123456	Mixed	Protein products; no corresponding
YP_123456789		transcript record provided
NC_123456	Mixed	genomic, e.g., chromosomes
NG_123455	Mixed	genomic, incomplete region
NT_123456	Automated	genomic, BAC assembly
NW_123456	Automated	genomic, WGS assembly
NW_123456789		
NZ_ABCD12345678	Automated	genomic, WGS collection
NS_123456	Automated	genomic, assembly which does not reflect the structure of a real biological molecule
AC_123456	Mixed	genomic, Alternate complete genomic molecule

Bold are explained next slides

RefSeq Chromosomes: NC_, NG_ genomic: **Chromosomes** & , Incomplete region

Escherichia coli O157:H7 str. Sakai, complete genome

NCBI Reference Sequence: NC_002695.1

[FASTA](#) [Graphics](#)

Go to:

LOCUS	NC_002695	5498450 bp	DNA	circular	BCT	24-MAY-2011
DEFINITION	Escherichia coli O157:H7 str. Sakai, complete genome.					
ACCESSION	NC_002695					
VERSION	NC_002695.1	GI:	15829254			
DBLINK	Project: 57781					
KEYWORDS	.					
SOURCE	Escherichia coli O157:H7 str. Sakai					
ORGANISM	Escherichia coli O157:H7 str. Sakai Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia.					
REFERENCE	1 (bases 1 to 5498450)					
AUTHORS	Bergholz,T.M., Wick,L.M., Qi,W., Riordan,J.T., Ouellette,L.M. and Whittam,T.S.					
TITLE	Global transcriptional response of Escherichia coli O157:H7 to growth transitions in glucose minimal medium					
JOURNAL	BMC Microbiol. 7, 97 (2007)					
PUBMED	17967175					
REMARK	Publication Status: Online-Only					
REFERENCE	2 (sites)					
AUTHORS	Hayashi,T., Makino,K., Ohnishi,M., Kurokawa,K., Ishii,K., Yokoyama,K., Han,C.G., Ohtsubo,E., Nakayama,K., Murata,T., Tanaka,M., Tobe,T., Iida,T., Takami,H., Honda,T., Sasakawa,C., Ogasawara,N., Yasunaga,T., Kuhara,S., Shiba,T., Hattori,M. and Shinagawa,H.					



RefSeq Chromosomes: NT_, NW_ genomic: BAC assembly & WGS assembly

Mus musculus chromosome 6 genomic contig, alternate assembly (based on MGSCv3)

NCBI Reference Sequence: NW_000272.1

! **Record removed.** This record was removed as a result of standard genome annotation processing. See the genome build documentation at <http://www.ncbi.nlm.nih.gov/genome/guide/build.html> for further information, or contact info@ncbi.nlm.nih.gov.

[FASTA](#) [Graphics](#)

[Go to:](#)

LOCUS	NW_000272	27755676 bp	DNA	linear	CON	27-APR-2006
DEFINITION	Mus musculus chromosome 6 genomic contig, alternate assembly (based on MGSCv3).					
ACCESSION	NW_000272					
VERSION	NW_000272.1	GI:20915425				
KEYWORDS	.					
SOURCE	Mus musculus (house mouse)					
ORGANISM	Mus musculus	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea; Muridae; Murinae; Mus.				
TITLE	Initial sequencing and comparative analysis of the mouse genome					
JOURNAL	Nature 420 (6915), 520-562 (2002)					
PUBMED	12466850					
REFERENCE	2 (sites)					
AUTHORS	Lowe,T.M. and Eddy,S.R.					
TITLE	tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence					
JOURNAL	Nucleic Acids Res. 25 (5), 955-964 (1997)					
PUBMED	9023104					
REMARK	This is the methods paper for tRNAscan-SE.					
COMMENT	GENOME ANNOTATION REFSEQ : Features on this sequence have been produced for build 36 version 1 of the NCBI's genome annotation [see documentation].					
	This DNA sequence represents the MGSCv3, the publicly available Whole Genome Assembly of the mouse genome (C57BL/6J) that was made available in November of 2002.					

Alignment Generated Transcripts: XM_, XP_ Predicted mRNA & Predicted protein

PREDICTED: Mus musculus lipoprotein lipase, transcript variant 3 (Lpl), mRNA

NCBI Reference Sequence: XM_921447.1

 This sequence has been replaced by [NM_008509](#).

[FASTA](#) [Graphics](#)

[Go to:](#)

LOCUS XM_921447 1923 bp mRNA linear ROD 01-DEC-2005
DEFINITION PREDICTED: Mus musculus lipoprotein lipase, transcript variant 3 (Lpl), mRNA.
ACCESSION XM_921447
VERSION XM_921447.1 GI:82917585
KEYWORDS .
SOURCE Mus musculus (house mouse)
ORGANISM [Mus musculus](#)
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
Sciurognathi; Muroidea; Muridae; Murinae; Mus.
COMMENT MODEL [REFSEQ](#): This record is predicted by automated computational analysis. This record is derived from a genomic sequence ([NW_001030897](#)) annotated using gene prediction method: GNOMON, supported by mRNA and EST evidence.
Also see:
[Documentation of NCBI's Annotation Process](#)

[WARNING] On Dec 7, 2005 this sequence was replaced by gi:[6678709](#).

FEATURES Location/Qualifiers
source 1..1923
/organism="Mus musculus"
/mol_type="mRNA"
/strain="mixed"
/db_xref="taxon:[10090](#)"
/chromosome="8"

gene 1..1923
/gene="Lpl"
/note="Derived by automated computational analysis using gene prediction method: GNOMON. Supporting evidence includes similarity to: 74 mRNAs, 785 ESTs, 22 Proteins"
/db_xref="GeneID:[16956](#)"
/db_xref="MGI:[96820](#)"

CDS 234..1658
/gene="Lpl"
/codon_start=1
/product="similar to Lipoprotein lipase precursor (LPL) isoform 3"
/protein_id="[XP_926540.1](#)"
/db_xref="GI:82917586"
/db_xref="GeneID:[16956](#)"
/db_xref="MGI:[96820](#)"

Curated RefSeq Records: NM_, NP_ mRNA & Protein from NM_

Mus musculus lipoprotein lipase (Lpl), mRNA

NCBI Reference Sequence: NM_008509.2

[FASTA](#) [Graphics](#)

[Go to:](#)

LOCUS NM_008509 4049 bp mRNA linear ROD 17-SEP-2011
DEFINITION Mus musculus lipoprotein lipase (Lpl), mRNA.
ACCESSION NM_008509 XM_909793 XM_921447
VERSION NM_008509.2 GI:126723005
KEYWORDS .
SOURCE Mus musculus (house mouse)
ORGANISM [Mus musculus](#)

COMMENT VALIDATED [REFSEQ](#): This record has undergone validation or preliminary review. The reference sequence was derived from [AK152014.1](#), [AK150375.1](#) and [AC158236.2](#).
On Mar 8, 2007 this sequence version replaced gi:[6678709](#).

Publication Note: This RefSeq record includes a subset of the publications that are available for this gene. Please see the Gene record to access additional publications.
COMPLETENESS: complete on the 3' end.

CDS 199..1623
/gene="Lpl"
/EC_number="[3.1.1.34](#)"
/note="O 1-4-5"
/codon_start=1
/product="lipoprotein lipase precursor"
/protein_id="[NP_032535.2](#)"
/db_xref="GI:126723005"
/db_xref="CCDS:[CCDS40357.1](#)"
/db_xref="GeneID:[16956](#)"
/db_xref="MGI:[96820](#)"
/translation="MESKALLLVGVWLQSLTAFRGGVAAADAGRDFSDIESKFALR
TPEDTAEDTCHLIPGLADSVSNCNFHSSKTFVVIIHGWTGMYESWVPKLVAALYKR
EPDSNVIVWDLYRAQQHYPVSAGYTKLVGNNDVARFINWMEEEFNYPLDNVHLGYS
GAHAAGVAGSLTNKKVNRITGLDPAGPNFEYAEAPSRSLSPDDADFVDVLHTFTRGSPG
RSIGIQKPVGHDYIPNGGTQPGCNIGEAIRVIAERGLGDVDQLVKCSHERSIIHLFI
DSLLNEENPSKAYRCNSKEAFEKGCLCSRKNRNNLGYEINKVRAKRSSKMYLKTRS
QMPYKVPHYQVKIHGSTEDGKQHNQAFEISLYGTVAESENIPFTLPEVSTNKTYFL
IYTEVDIGELMMKLKWISDSYFSWPDWSSPSFVIERIRVKAGETQKKVIFCAREKV
SHLQKGKD萨FVKCHDKSLKKSG"

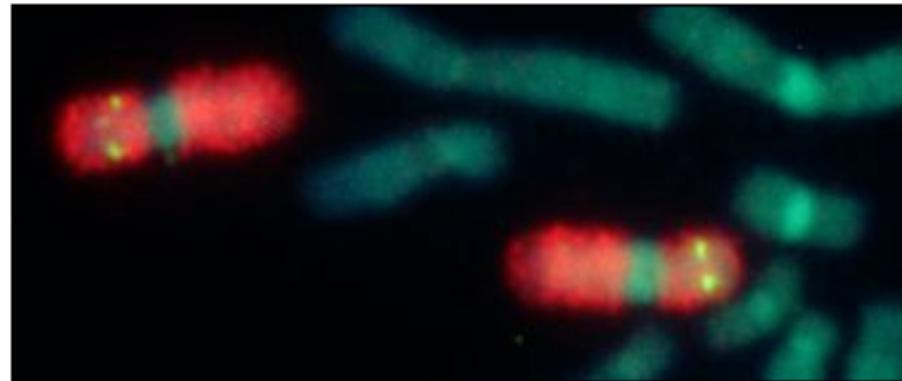
Things to Read

- <ftp://ftp.ncbi.nih.gov/refseq/release/release-notes/RefSeq-release61.txt>
 - 4.1.3 NUCLEOTIDE FEATURE ANNOTATION
 - 4.1.4 PROTEIN FEATURE ANNOTATION
 - 4.2 Tracking Identifiers
 - 4.2.1 GeneID
 - 4.2.2 Transcript ID
 - 4.2.3 Protein ID
 - 4.2.4 Conserved Domain Database (CDD) ID
 - 5.2 RefSeq Distribution Formats
- RefSeq FAQ
 - <http://www.ncbi.nlm.nih.gov/books/NBK50679/>
- Now for the Gene Database

NCBI Gene

Gene integrates information from a wide range of species
A record may include nomenclature, Reference Sequences
(RefSeqs), maps, pathways, variations, phenotypes, and links to
genome-, phenotype-, and locus-specific resources worldwide

2010 = The Gene Database

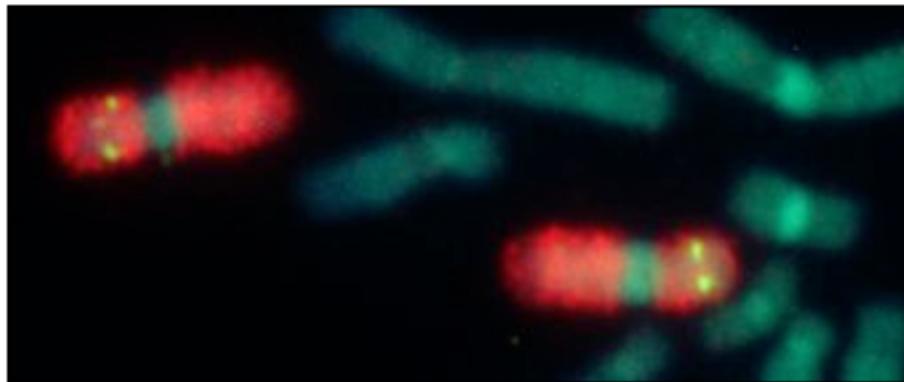


- Gene Centered Information
- Unifies NCBI-annotated and Submitted Genomes
- 8.19 million records for 7806 taxa (Oct 2010)
- Great place to start your searches
 - You'll see with the take home part of this lecture

<http://www.ncbi.nlm.nih.gov/gene>

http://www.ncbi.nlm.nih.gov/projects/Gene/gentrez_stats.cgi

2013 = The Gene Database

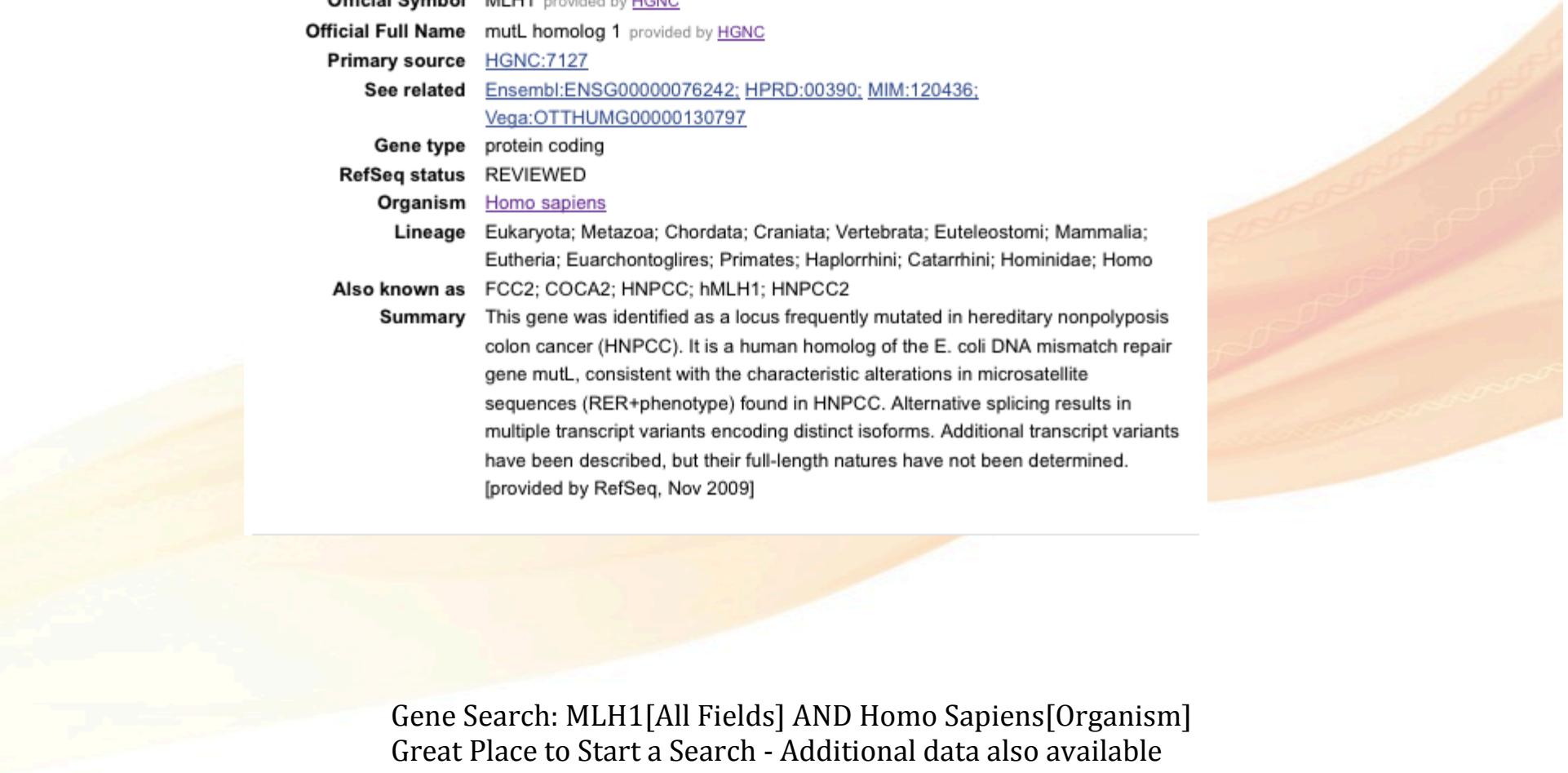


- Gene Centered Information
- Unifies NCBI-annotated and Submitted Genomes
- 13.366 million records for 11306 taxa (Oct 2013)
- Great place to start your searches
 - You'll see with the take home part of this lecture

<http://www.ncbi.nlm.nih.gov/gene>

http://www.ncbi.nlm.nih.gov/projects/Gene/gentrez_stats.cgi

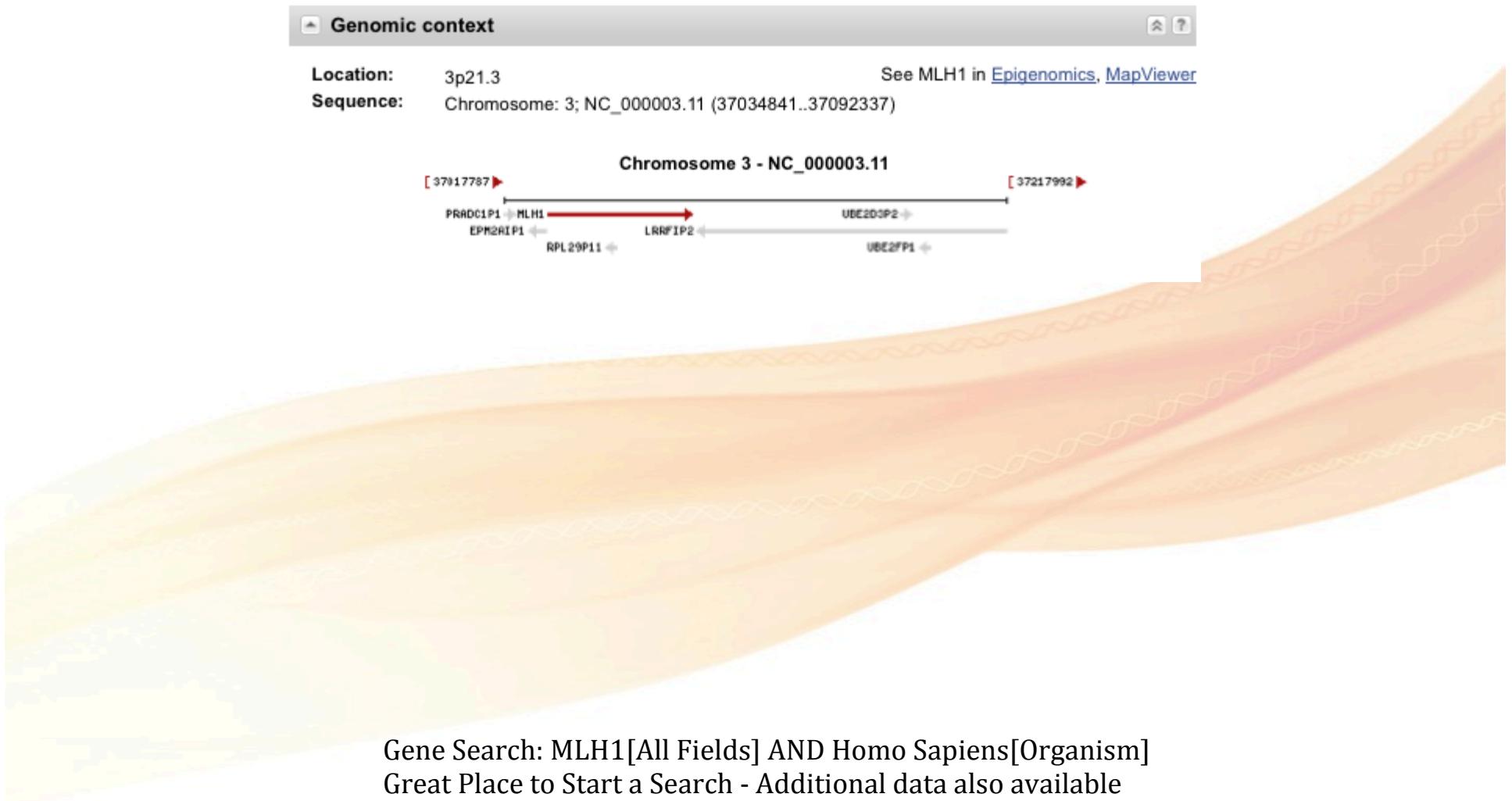
Genes MLH1: One Stop Shopping

A screenshot of a gene information page from a database like HGNC. The title bar says "Summary". The page contains the following data:

Official Symbol	MLH1 provided by HGNC
Official Full Name	mutL homolog 1 provided by HGNC
Primary source	HGNC:7127
See related	Ensembl:ENSG00000076242 ; HPRD:00390 ; MIM:120436 ; Vega:OTTHUMG00000130797
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	FCC2; COCA2; HNPCC; hMLH1; HNPCC2
Summary	This gene was identified as a locus frequently mutated in hereditary nonpolyposis colon cancer (HNPCC). It is a human homolog of the <i>E. coli</i> DNA mismatch repair gene mutL, consistent with the characteristic alterations in microsatellite sequences (RER+phenotype) found in HNPCC. Alternative splicing results in multiple transcript variants encoding distinct isoforms. Additional transcript variants have been described, but their full-length natures have not been determined. [provided by RefSeq, Nov 2009]

Gene Search: MLH1[All Fields] AND Homo Sapiens[Organism]
Great Place to Start a Search - Additional data also available

Genes MLH1: One Stop Shopping

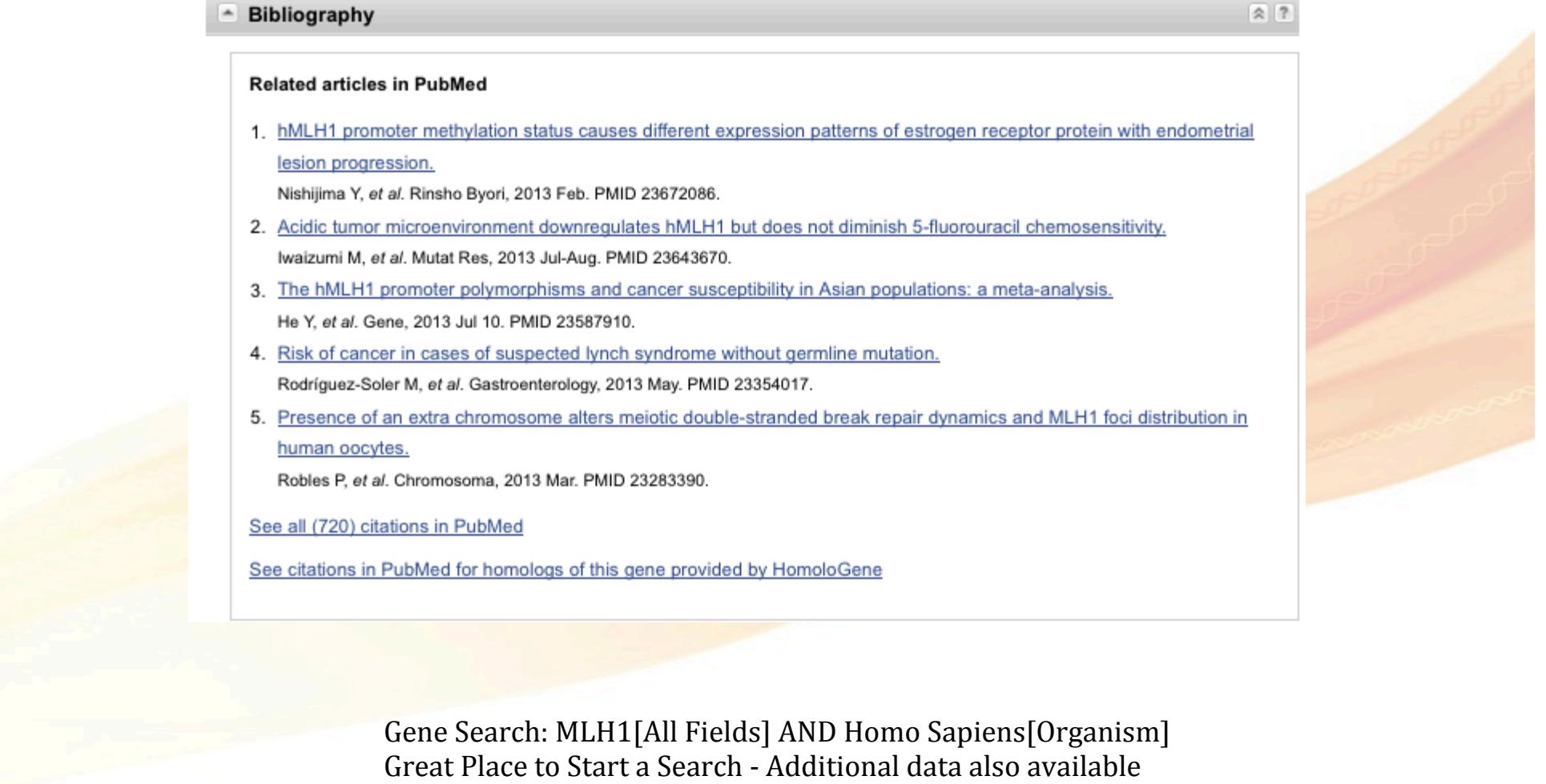




Genes MLH1: One Stop Shopping



Genes MLH1: One Stop Shopping

A screenshot of a computer window titled "Bibliography". The window displays a list of "Related articles in PubMed" for the gene MLH1. The articles are numbered 1 through 5, each with a title, author(s), and PMID. There are also links to "See all (720) citations in PubMed" and "See citations in PubMed for homologs of this gene provided by HomoloGene".

Bibliography

Related articles in PubMed

1. [hMLH1 promoter methylation status causes different expression patterns of estrogen receptor protein with endometrial lesion progression.](#)
Nishijima Y, et al. *Rinsho Byori*, 2013 Feb. PMID 23672086.
2. [Acidic tumor microenvironment downregulates hMLH1 but does not diminish 5-fluorouracil chemosensitivity.](#)
Iwaizumi M, et al. *Mutat Res*, 2013 Jul-Aug. PMID 23643670.
3. [The hMLH1 promoter polymorphisms and cancer susceptibility in Asian populations: a meta-analysis.](#)
He Y, et al. *Gene*, 2013 Jul 10. PMID 23587910.
4. [Risk of cancer in cases of suspected lynch syndrome without germline mutation.](#)
Rodríguez-Soler M, et al. *Gastroenterology*, 2013 May. PMID 23354017.
5. [Presence of an extra chromosome alters meiotic double-stranded break repair dynamics and MLH1 foci distribution in human oocytes.](#)
Robles P, et al. *Chromosoma*, 2013 Mar. PMID 23283390.

[See all \(720\) citations in PubMed](#)

[See citations in PubMed for homologs of this gene provided by HomoloGene](#)

Gene Search: MLH1[All Fields] AND Homo Sapiens[Organism]
Great Place to Start a Search - Additional data also available

NCBI Reference Sequences (RefSeq)

RefSeqs maintained independently of Annotated Genomes

These reference sequences exist independently of genome builds. [Explain](#)

Genomic

NG_007109.2 RefSeqGene

Range 5001..62497
Download [GenBank](#), [FASTA](#), [Sequence Viewer \(Graphics\)](#), [LRG 216](#)

mRNA and Protein(s)

NM_000249.3 → NP_000240.1 DNA mismatch repair protein MLH1 isoform 1

[See proteins identical to NP_000240.1](#)

Status: REVIEWED

Description	Transcript Variant: This variant (1) represents the longest transcript and encodes the longest isoform (1).
Source sequence(s)	BC006850 , CN414955
Consensus CDS	CCDS2663.1
UniProtKB/Swiss-Prot	P40692
Related	ENSP00000231790 , OTTHUMP00000161361 , ENST00000231790 , OTTHUMT00000253337
Conserved Domains (3) <u>summary</u>	
	cd00075 Location:31 – 122 Blast Score: 107
	cd03483 Location:211 – 335 Blast Score: 542
	TIGR00585 Location:6 – 315 Blast Score: 1082



Function	Evidence Code	Pubs
ATP binding	IEA	
ATPase activity	IBA	
contributes to MutSalpha complex binding	IDA	PubMed
guanine/thymine mispair binding	IEA	
protein binding	IPI	PubMed
contributes to protein binding	IPI	PubMed
contributes to single-stranded DNA binding	IDA	PubMed

Process	Evidence Code	Pubs
ATP catabolic process	IBA	
double-strand break repair via nonhomologous end joining	IEA	
intrinsic apoptotic signaling pathway in response to DNA damage	IEA	
isotype switching	IEA	
male meiosis chromosome segregation	IEA	
meiotic metaphase I plate congression	IEA	
mismatch repair	IBA	
negative regulation of mitotic recombination	IEA	
nuclear-transcribed mRNA poly(A) tail shortening	IEA	
oogenesis	IEA	
reciprocal meiotic recombination	IBA	
resolution of meiotic recombination intermediates	IEA	
somatic hypermutation of immunoglobulin genes	IBA	
spermatogenesis	IEA	
spindle midzone assembly involved in meiosis	IEA	
synapsis	IEA	

Component	Evidence Code	Pubs
MutLalpha complex	IBA	
MutLbeta complex	IBA	
chiasma	IBA	
male germ cell nucleus	IEA	
mismatch repair complex	IBA	
nucleus	IC	PubMed
synaptonemal complex	IBA	



Other Data Sources include:
 Phenotypes
 Variation
 Interactions
 Pathways
 General Protein Information
 Related Sequences
 Additional Links

NCBI Protein

The Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB. Protein sequences are the fundamental determinants of biological structure and function.

GenPept Database

- Database of GenBank gene products
 - Part of NCBI Protein Database
 - Contains:
 - a.a. translations from GenBank/EMBL/DDBJ records
 - that have a **coding region feature** annotated on them
- Any nucleotide sequence:
 - With **CDS region** is submitted to the GenBank database
 - Automatically translated into a.a sequence
- Genpept record is created based on the information included in **source record**

GenPept: GenBank CDS translations

FEATURES	Location/Qualifiers
source	1..2484 /organism="Homo sapiens" /mol_type="mRNA" /db_xref="taxon:9606" /chromosome="3" /map="3p22-p23"
gene	1..2484 /gene="M
CDS	22..2292 /gene="M" >gi 463989 gb AAC50285.1 DNA mismatch repair prote... MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKSTSIV... /note="homolog of S. cerevisiae PMS1 (Swiss-Prot Accession Number P14242), S. cerevisiae MLH1 (GenBank Accession Number U07187), Escherichia coli MUTL (Swiss-Prot Accession Number P23367), Salmonella typhimurium MUTL (Swiss-Prot Accession Number P14161) and Streptococcus pneumoniae (Swiss-Prot Accession Number P14160)" /codon_start=1 /product="DNA mismatch repair protein homolog" /protein_id="AAC50285.1" /db_xref="GI:463989" /translation="MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKS TSIQIVKEGLKLIQIQLDNGTGIRKEDLDIVCERFTTSKLQSFEDLASISTYGFRGE ALASISHVAHTITTKTADGKCAYRASYSDGKLKAPPKPCAGNQGTQITVEDLFYNIA TRRKALKNPSEYYGKILEVVGRYSVHNAGISFSVKQGETVADVRTLPNASTVDNIRS"

At Home

- Check out the following
 - http://www.ncbi.nlm.nih.gov/protein/NP_004922
- Note is link to a RefSeq
- Look Over the Features

Protein Sequences from Structures

1B63

Title MUTL COMPLEXED WITH ADPNP

Authors Yang, W.

Primary Citation Ban, C., Junop, M., Yang, W., Transformation of MutL by ATP binding and hydrolysis: a switch in DNA mismatch repair. *Cell* v97 pp.85-97, 1999 [PubMed]

History Deposition 1999-01-20 Release 1999-06-08

Experimental Method Type X-RAY DIFFRACTION Data

Parameters Resolution [Å] R-Value R-Free Space Group
1.90 0.213 (obs.) 0.261 I 2 2 2

Unit Cell Length [Å] a 62.19 b 72.37 c 189.93
Angles [°] alpha 90.00 beta 90.00 gamma 90.00

Molecular Description Polymer: 1 Molecule: MUTL Fragment: ATPASE FRAGMENT Chains: A

Functional Class DNA Mismatch Repair

Source Polymer: 1 Scientific Name: Escherichia coli Expression system: Escherichia coli

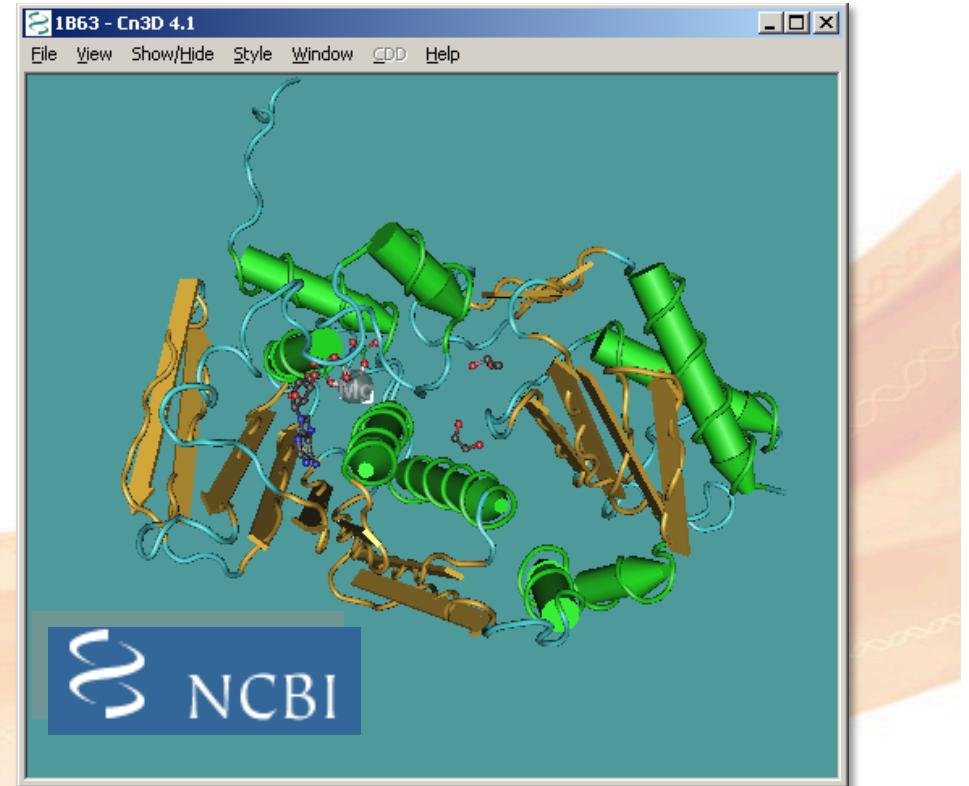
Chemical Component Identifier Name Formula Ligand Structure Ligand Interaction
MG MAGNESIUM ION Mg²⁺ [View] [View]
EDO 1,2-ETHANEDIOL C₂H₆O₂ [View] [View]
ANP PHOSPHOAMINOPHOSPHONIC ACID-ADENYLATE ESTER C₁₀H₁₇N₆O₁₂P₃ [View] [View]

Images and Visualization

Biological Molecule

Display Options: KiNG, Jmol, WebMol, All Images

RCSB PDB PROTEIN DATA BANK



```
>gi|5542073|pdb|1B63|A Chain A, Mutl Complexed With Adpnp
SHMPIQVLPPQLANQIAAGEVVERPASVVKELVENSLDAGATRIDIDIERGGAKLIRIRDNGCGIKKDEL
ALALARHATSKIASLDDLEAIISLGFRGEALASISSLVSRLLTSLTAEQQEAWQAYAEGRDMNVTVKPAA
HPVGTTLEVLDLFYNTPARRKFLRTEKFNHIDEIIRRRIALARFDVTINLSHNGKIVRQYRAVPEGGQK
ERRLGAIICGTAFLEQALAIIEWQHGDLTLRGWVADPNHTTPALAEIQYCYVNNGRMMRDRLINHAIRQACED
KLGADQQPAFVLYLEIDPHQVDVNVPKAHEVRFHQSLVHDFIYQGVLSVLQ
```

The Problems with Genbank and Genpept

- Lot of redundancy
- Same gene could be deposited into the database many times with different name
- **Different version of the same gene** could be **submitted many times** with different accession number
- The features of GenBank record could be chaotic

NCBI UniGene

UniGene computationally identifies transcripts from the same locus; analyzes expression by tissue, age, and health status; and reports related proteins (protEST) and clone resources.

Gene Expression is Regulated in Several Basic Ways

- By region (e.g. brain versus kidney)
 - In development (e.g. fetal versus adult tissue)
 - In dynamic response to environmental signals
 - e.g. immediate-early response genes)
 - In disease states
 - By gene activity
-
- UniGene data come from many cDNA libraries
 - Thus, when you look up a gene in UniGene you get information on its abundance and its regional distribution

What is UniGene?

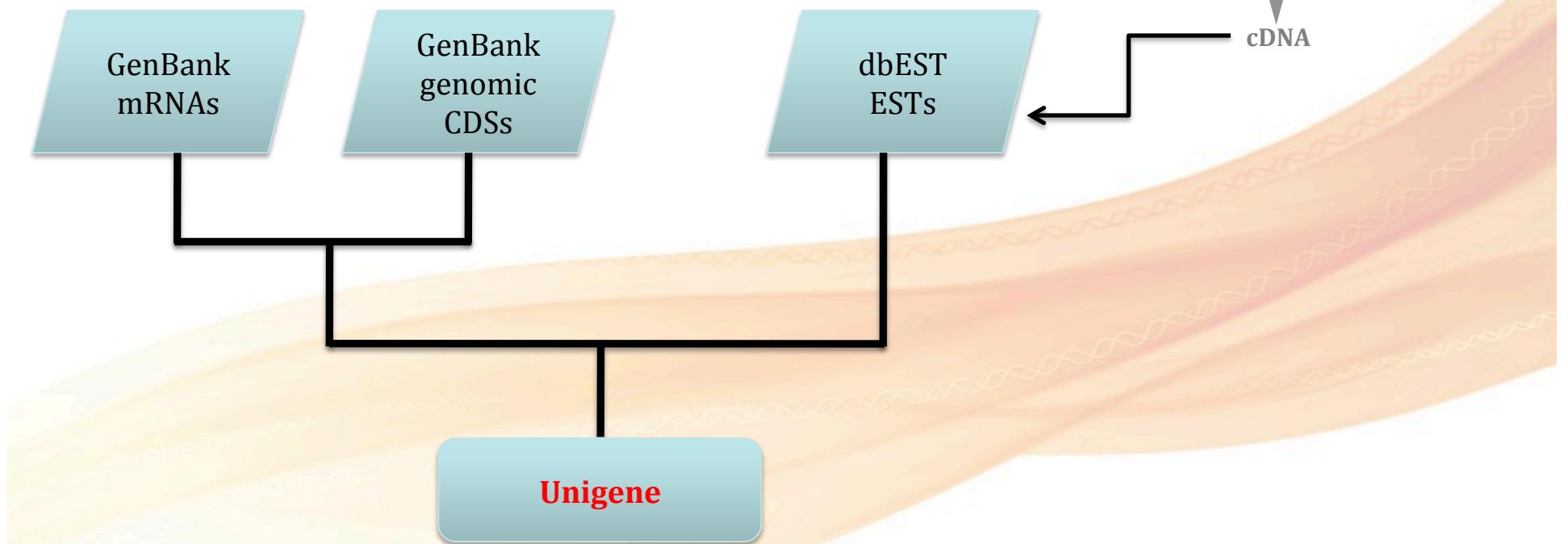
- Ongoing effort at NCBI to cluster EST sequences with traditional gene sequences
- A gene-oriented view of sequence entries
 - Organized view of the **Transcriptome**
- Blast based automated sequence clustering
 - Nonredundant set of gene oriented clusters
 - Each cluster a unique gene
- For each cluster, there is a lot of additional information included
 - Information on tissue types and map locations
 - Includes known genes and uncharacterized ESTs
 - Useful for gene discovery and selection of mapping reagents

<http://www.ncbi.nlm.nih.gov/unigene/>

<http://www.ncbi.nlm.nih.gov/UniGene/help.cgi?item=FAQ>

UniGene

- A non-redundant set of gene-oriented clusters



- Organized view of the Transcriptome

UniGene

Chordata	
Mammalia	
<i>Bos taurus</i> (cow)	43,448
<i>Canis lupus familiaris</i> (dog)	27,853
<i>Equus caballus</i> (horse)	8,348
<i>Homo sapiens</i> (human)	122,726
<i>Macaca fascicularis</i> (crab-eating macaque)	11,951
<i>Macaca mulatta</i> (rhesus monkey)	15,359
<i>Monodelphis domestica</i> (gray short-tailed opossum)	966
<i>Mus musculus</i> (mouse)	79,541
<i>Ornithorhynchus anatinus</i> (platypus)	1,827
<i>Oryctolagus cuniculus</i> (rabbit)	6,576
<i>Ovis aries</i> (sheep)	1,090
Echinodermata	
Echinoidea	
<i>Paracentrotus lividus</i> (common urchin)	8,684
<i>Strongylocentrotus purpuratus</i> (purple sea urchin)	19,639
Arthropoda	
Branchiopoda	
<i>Daphnia pulex</i> (common water flea)	14,190
Insecta	
<i>Acyrtosiphon pisum</i> (pea aphid)	12,891
<i>Aedes aegypti</i> (yellow fever mosquito)	19,345
<i>Anopheles gambiae</i> (African malaria mosquito)	21,387
<i>Apis mellifera</i> (honey bee)	9,758
<i>Bombyx mori</i> (domestic silkworm)	11,198
<i>Culex pipiens</i> (house mosquito)	4,957
<i>Drosophila melanogaster</i> (fruit fly)	17,197
<i>Ixodes scapularis</i> (black-legged tick)	18,161
<i>Tribolium castaneum</i> (red flour beetle)	9,053
Nematoda	
Chromadorea	
<i>Ancylostoma caninum</i> (dog hookworm)	7,394
<i>Caenorhabditis elegans</i> (nematode)	21,662
Platyhelminthes	
Trematoda	
<i>Schistosoma japonicum</i>	9,395
<i>Schistosoma mansoni</i>	10,219
Aves	
Gallus	
Meleagris	
Taeniochelys	
Cephalochordata	
Branchiostoma	
<i>Petromyzon marinus</i> (sea lamprey)	11,069

Chordates

Streptophyta	
Bryopsida	
<i>Physcomitrella patens</i>	20,137
Coniferopsida	
<i>Picea glauca</i> (white spruce)	17,809
<i>Picea sitchensis</i> (Sitka spruce)	16,755
<i>Pinus taeda</i> (loblolly pine)	18,921
Eudicotyledons	
<i>Aquilegia formosa x Aquilegia pubescens</i>	8,063
<i>Arabidopsis thaliana</i> (thale cress)	30,383
<i>Artemisia annua</i> (sweet wormwood)	9,462
<i>Brassica napus</i> (rape)	26,733
<i>Brassica oleracea</i>	5,617
<i>Brassica rapa</i> (field mustard)	14,497
<i>Citrus clementina</i>	6,107
<i>Citrus sinensis</i> (Valencia orange)	15,815
<i>Glycine max</i> (soybean)	30,248
<i>Gossypium hirsutum</i> (upland cotton)	21,743
<i>Gossypium raimondii</i>	3,297
<i>Helianthus annuus</i> (sunflower)	7,846
<i>Lactuca sativa</i> (garden lettuce)	7,940
<i>Lotus japonicus</i>	14,493
<i>Malus x domestica</i> (apple)	16,932
<i>Medicago truncatula</i> (barley)	17,764
<i>Nicotiana tabacum</i> (tobacco)	17,764
<i>Populus tremula x Populus tremuloides</i>	5,957
<i>Populus trichocarpa</i> (western basswood)	5,957
<i>Prunus persica</i> (peach)	6,623
<i>Raphanus raphanistrum</i> (radish)	3,994
<i>Raphanus sativus</i> (radish)	3,994
<i>Solanum lycopersicum</i> (tomato)	5,259
<i>Solanum tuberosum</i> (potato)	5,380
<i>Vigna unguiculata</i> (cowpea)	2,216
<i>Vitis vinifera</i> (wine grape)	4,838
Liliopsida	
<i>Hordeum vulgare</i> (barley)	7,257
<i>Oryza sativa</i> (rice)	7,257
<i>Saccharum officinarum</i> (sugarcane)	7,257
<i>Sorghum bicolor</i> (sorghum)	7,257
<i>Triticum aestivum</i> (Wheat)	7,257
<i>Zea mays</i> (maize)	7,257
Chlorophyta	
Chlorophyceae	
<i>Chlamydomonas reinhardtii</i>	6,778
<i>Volvox carteri</i>	5,656

Plants

Fungi et al.

Unigene Identifier

- **Hs** for [human](#)
- **Mm** for [mouse](#)
- **Rn** for [rat](#)
- **Bt** for [cow](#)
- **Dr** for [zebrafish](#)
- **Dm** for [fruitfly](#)
- **Aga** for [malaria mosquito](#)
- **Xl** for [African clawed frog](#) (*Xenopus laevis*)
- **At** for [thale cress](#) (*Arabidopsis thaliana*)
- **Os** for [rice](#)
- **Ta** for [wheats](#)
- **Zm** for [maize](#)

Examples:

[Mm.213407](#)

[Hs.214142](#)

[At.138](#)

[Ppr.17865](#)

[UniGene build RSS](#)

Should be noted that the procedures for automated sequence clustering are still under development and the results may change from time to time as improvements are made

Unigene is regularly rebuilt.
Therefore, cluster identifiers are not stable gene indices

<ftp://ftp.ncbi.nlm.nih.gov/repository/UniGene/>

Homo sapiens: UniGene Build #236

Sequences Included in UniGene

Known genes are from GenBank 09 Mar 2013

ESTs are from dbEST through 09 Mar 2013

221,081	mRNAs
96	Models
20,073	HTC
1,695,050	EST, 3'reads
4,029,704	EST, 5'reads
1,034,785	EST, other/unknown
7,000,789	total sequences in clusters

Build Method: Genome Based

Alignments between transcript sequences and genomic sequences are used to generate clusters of sequences originating from the same gene.

[More...](#)

Final Number of Clusters (sets)

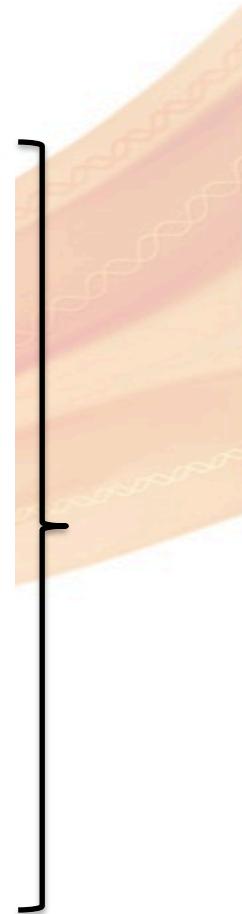
130,056	sets total
34,637	sets contain at least one mRNA
11,925	sets contain at least one HTC sequence
124,724	sets contain at least one EST
29,805	sets contain both mRNAs and ESTs

Hs for human

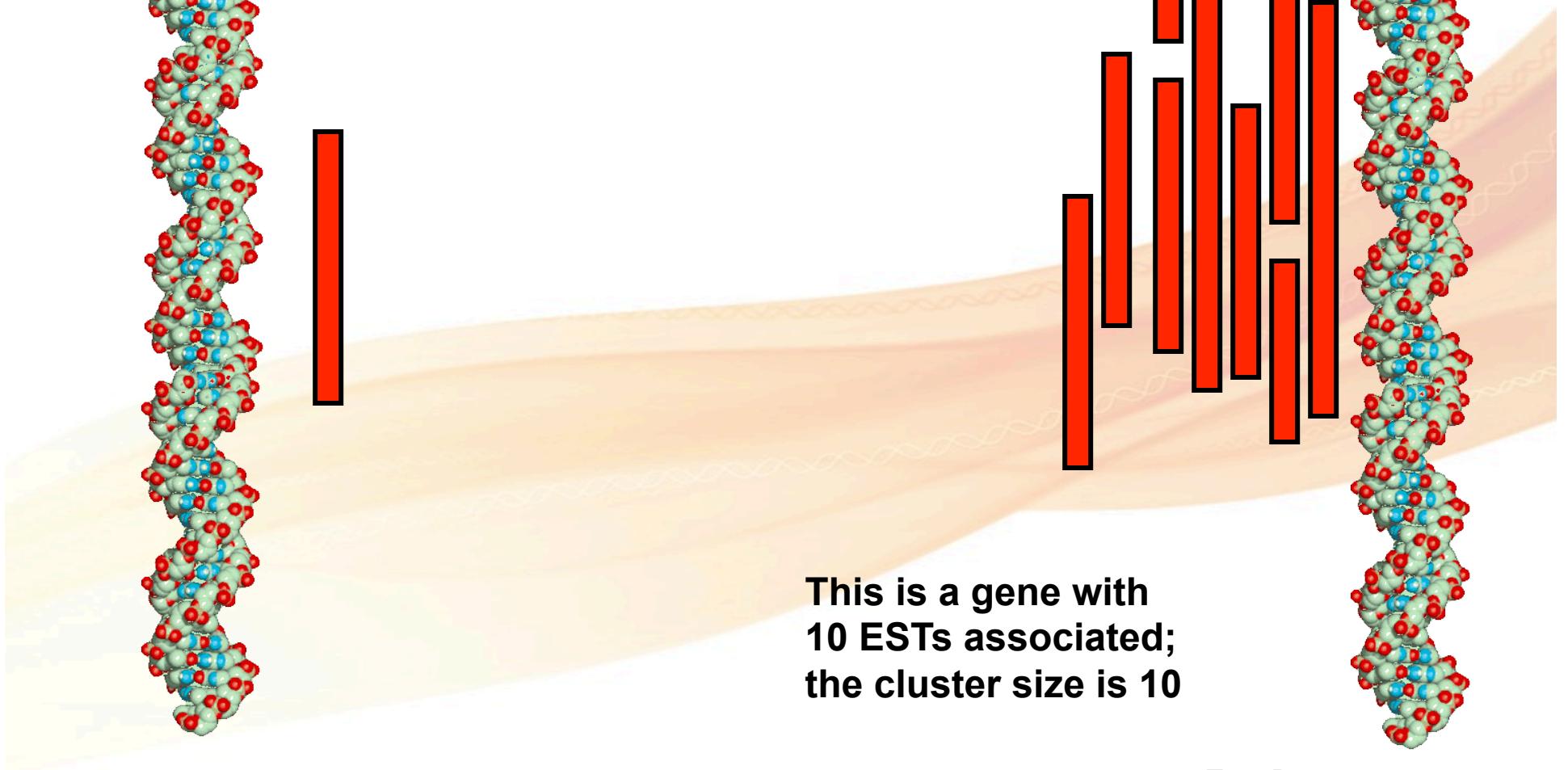
Histogram of cluster sizes for UniGene Hs build 236

Cluster size	Number of clusters
32769-65536	1
16385-32768	4
8193-16384	20
4097-8192	56
2049-4096	224
1025-2048	729
513-1024	1982
257-512	4117
129-256	4557
65-128	3889
33-64	3858
17-32	4815
9-16	7177
5-8	10637
3-4	12760
2	10859
1	64371

sequences



Cluster Sizes in UniGene



**This is a gene with
10 ESTs associated;
the cluster size is 10**

Tom Peavy

Gene Catalog: Fathead Minnow MLH1Cluster

UGID:2056516 UniGene Ppr.17865 *Pimephales promelas* (fathead minnow)

Links

Ppr.17865

Transcribed locus, strongly similar to NP_956953.1 DNA mismatch repair protein
Mlh1 [Danio rerio]

SELECTED PROTEIN SIMILARITIES

Comparison of cluster transcripts with RefSeq proteins. The alignments can suggest function of the cluster.

Best Hits and Hits from model organisms		Species	Id(%)	Len(aa)
NP_956953.1	DNA mismatch repair protein Mlh1	<i>D. rerio</i>	94.4	723

NP_001090545.1 mutL homolog 1,

NP_000240.1 DNA mismatch re

NP_081086.2 DNA mismatch re

XP_001690102.1 mismatch repair p

NP_477022.1 Mlh1

XP_962522.1 hypothetical prote

NP_013890.1 Mlh1p

NP_567345.2 DNA mismatch re

Other

XP_002933455.1 PREDICTED: LO'

XP_001170433.1 PREDICTED: DN

SEQUENCES

Sequences representing this gene; mRNAs, ESTs, and gene predictions supported by transcribed sequences.

EST sequences (10 of 22) [Show all sequences]

DT170650.1	Clone ANNO42204	whole body	3' read P
DT170651.1	Clone ANNO42204	whole body	5' read P
DT182780.1	Clone ANNO49223	whole body	3' read P
DT182781.1	Clone ANNO49223	whole body	5' read P
DT191378.1	Clone ANNO54379	whole body	5' read P
DT206388.1	Clone CAAS13911	brain	3' read PA
DT206389.1	Clone CAAS13911	brain	5' read P
DT225851.1	Clone CAAT2335	brain	3' read PA
DT225852.1	Clone CAAT2335	brain	5' read P
DT237339.1	Clone CAAT9738	brain	3' read P

Download Sequences

ESTs

Key to Symbols

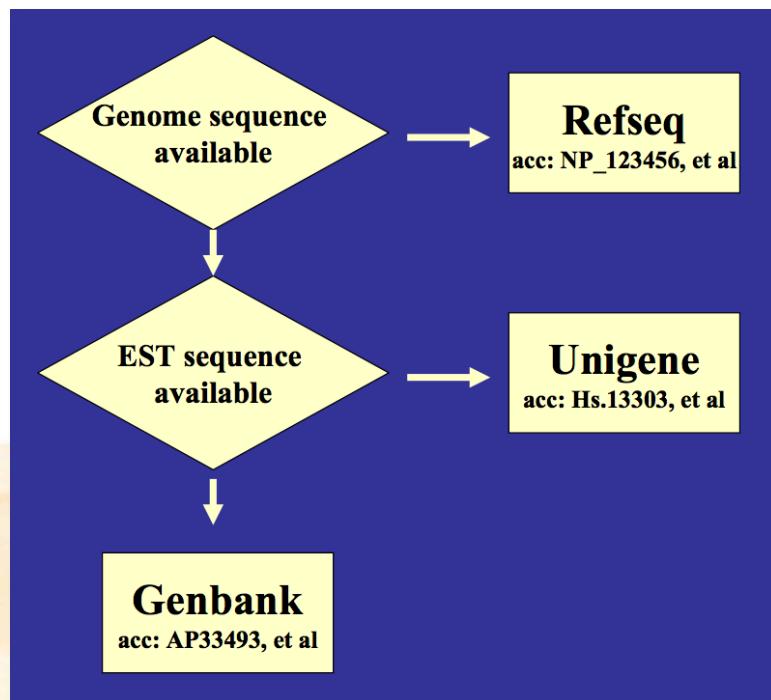
- P Has similarity to known Proteins (after translation)
- A Contains a poly-Adenylation signal
- S Sequence is a Suboptimal member of this cluster
- M Clone is putatively CDS-complete by MGC criteria

<http://www.ncbi.nlm.nih.gov/UniGene/clust.cgi?UGID=2056516&TAXID=90988&SEARCH=Ppr.17865>

FYI – How is Unigene Clustered

- Transcript-Based Build Procedure
 - <http://www.ncbi.nlm.nih.gov/UniGene/help.cgi?item=build1>
- Genome-Based UniGene Build Procedure
 - <http://www.ncbi.nlm.nih.gov/UniGene/help.cgi?item=build2>
- [Pontius JU, Wagner L, Schuler GD. UniGene: a unified view of the transcriptome. In: The NCBI Handbook. Bethesda \(MD\): National Center for Biotechnology Information; 2003.](#)

Refseq? Unigene? Genbank?



Conclusion: UniGene

- UniGene is a useful tool to look up information about expressed genes
- UniGene displays information about:
 - The abundance of a transcript (expressed gene)
 - As well as its regional distribution of expression (e.g. brain vs. liver)



FYI - NCBI - MMDB

MMDB: Molecular Modeling DataBase

- Derived from experimentally determined PDB records
- Value added to PDB records including:
 - Explicit chemical graphs
 - Computationally identified similar 3D structure
 - Links:
 - literature
 - Similar sequences
 - information about chemicals bound to the structure
 - Validation (secondary structure elements)
 - Inclusion of Taxonomy, Citation
 - Conversion to ASN.1 data description language
- Structure neighbors determined by
- Vector Alignment Search Tool (VAST)

<http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml>

Connections in MMDB

- Make it possible to find 3D structures for homologs of a protein sequence of interest then interactively view:
 - Sequence-structure relationships
 - Active sites
 - Bound chemicals
 - Journal articles
- How do you search?
 - If you don't know a PDB accession ID
 - Search [MMDB](#) using the gene and or protein name (e.g. [COMT](#))
 - If you know the PDB
 - Search [MMDB 1VID](#) for example

<http://www.ncbi.nlm.nih.gov/structure/?term=COMT&SITE=NcbiHome&submit.x=0&submit.y=0>

<http://www.ncbi.nlm.nih.gov/structure/?term=1VID&SITE=NcbiHome&submit.x=0&submit.y=0>

Exercise - Links to Explore

- Search [MMDB 1FMK](#) for example
 - Click PDB ID link: [1FMK](#) at the top
 - Click the [View in Cn3D](#), to view the structure on your computer
 - Install Cn3D first
 - Click the Similar Structures: [How many related structures to Domain 1](#)
 - Click the domain image to view the [CDD](#) neighbors (click Show annotation first)
 - Click the [Protein](#) to see the PDB entry in GenPept
 - Scan the file contents
 - Click the Conserved Domains "Domain Families" link to see entry in the [Conserved Domain Database](#)

<http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml>

Human Tyrosine-Protein Kinase C-Src Domains

Crystal Structure Of Human Tyrosine-Protein Kinase C-Src

MMDB ID: 56157 PDB ID: 1FMK [?](#)

PDB deposition date: 1997/1/24 [?](#)
Updated in MMDB: 05/2011 [?](#)

MMDB version: 9 [?](#)
Source organism: Homo sapiens [?](#)
Related structures: total 16264 [?](#)
Inferred interactions: IBIS [?](#)
Experimental method: X-Ray Diffraction [?](#)

First Biological Unit All Biological Units (1) Asymmetric Unit [?](#)

Biological Unit: monomeric; determined by: author [?](#)

Interactions [?](#) Molecular Graphic [?](#) View or Save 3D Structure [?](#)

File Format: Cn3D [?](#)
Display As: 3D structure [?](#)
Data Set: Single 3D structu [?](#)
 View structure
 Download Cn3D

NOTICE
In order to view this biological

Protein
3d Domains
Domain Families
Specific Hits
Super Families
Multidomains

Sequence A
1 75 150 225 300 375 452

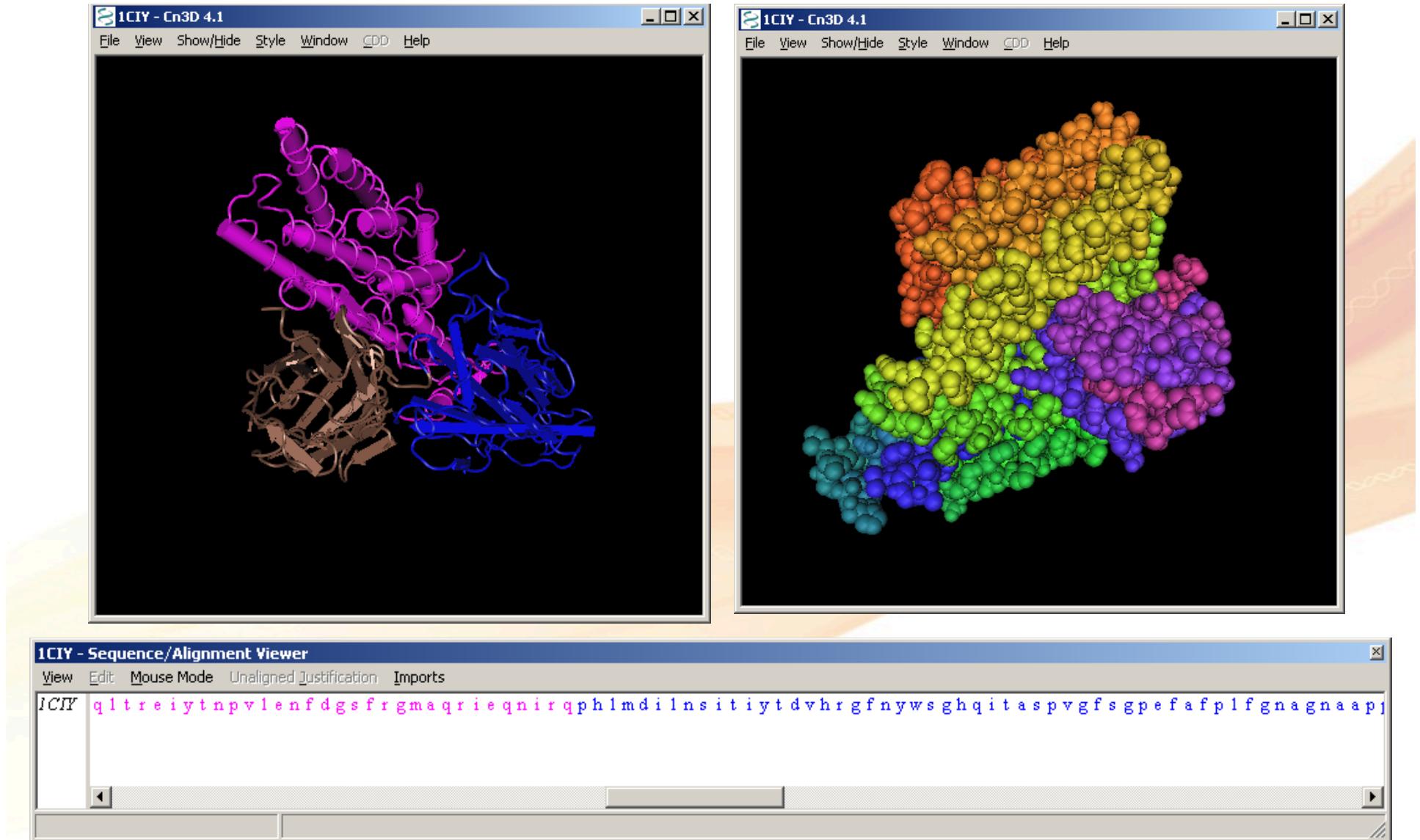
1 2 3 4

SH3 SH2 PTKc_Src_Like
SH3 superfamily SH2 superfamily PKc_like superfamily
PTKc_Src_Like superfamily Pkinase_Tyr

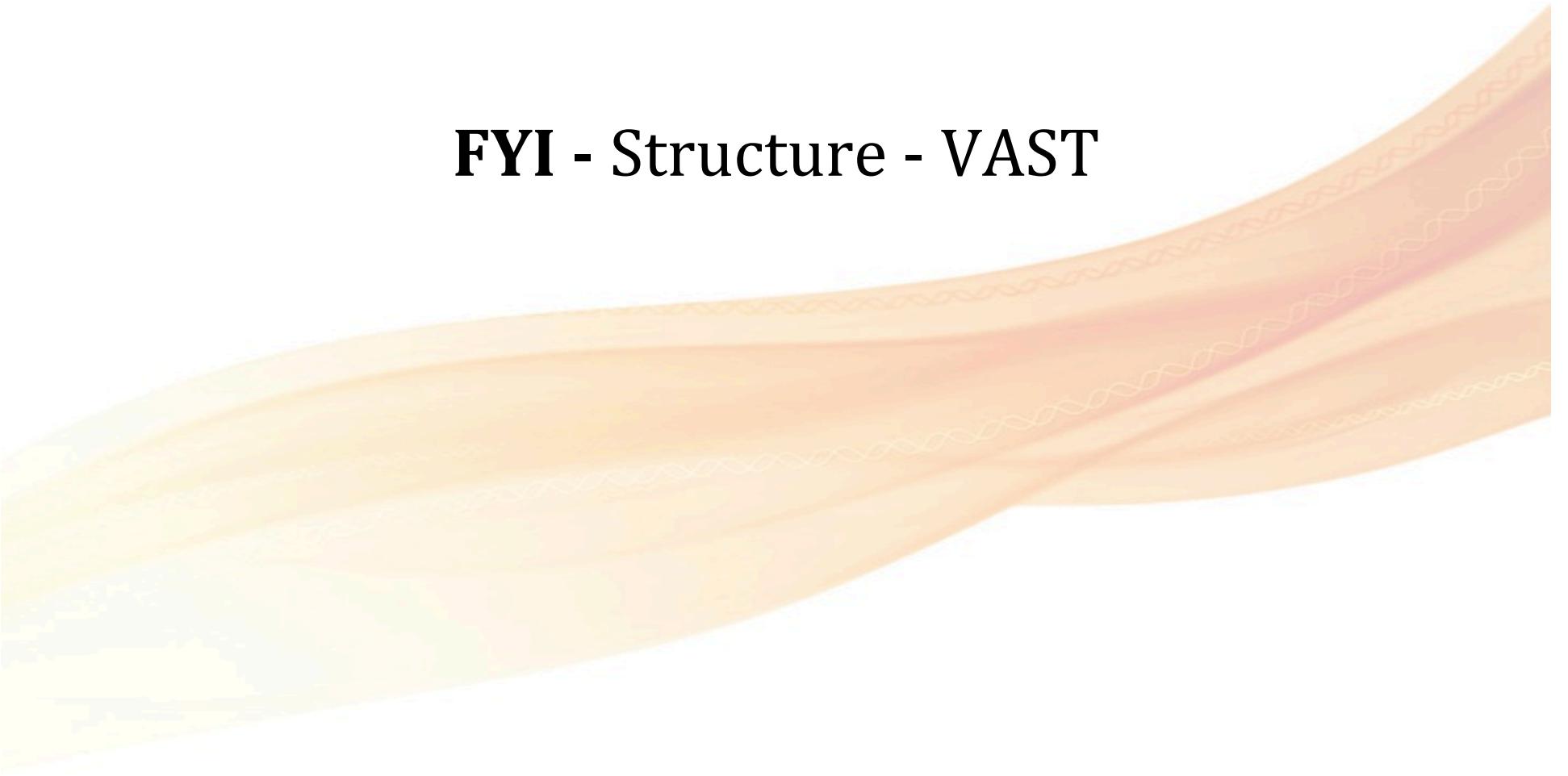
Four 3d domains
Three conserved domains

1
Show annotation ▾

Cn3D 4.1: *Bacillus thuringiensis* Toxin



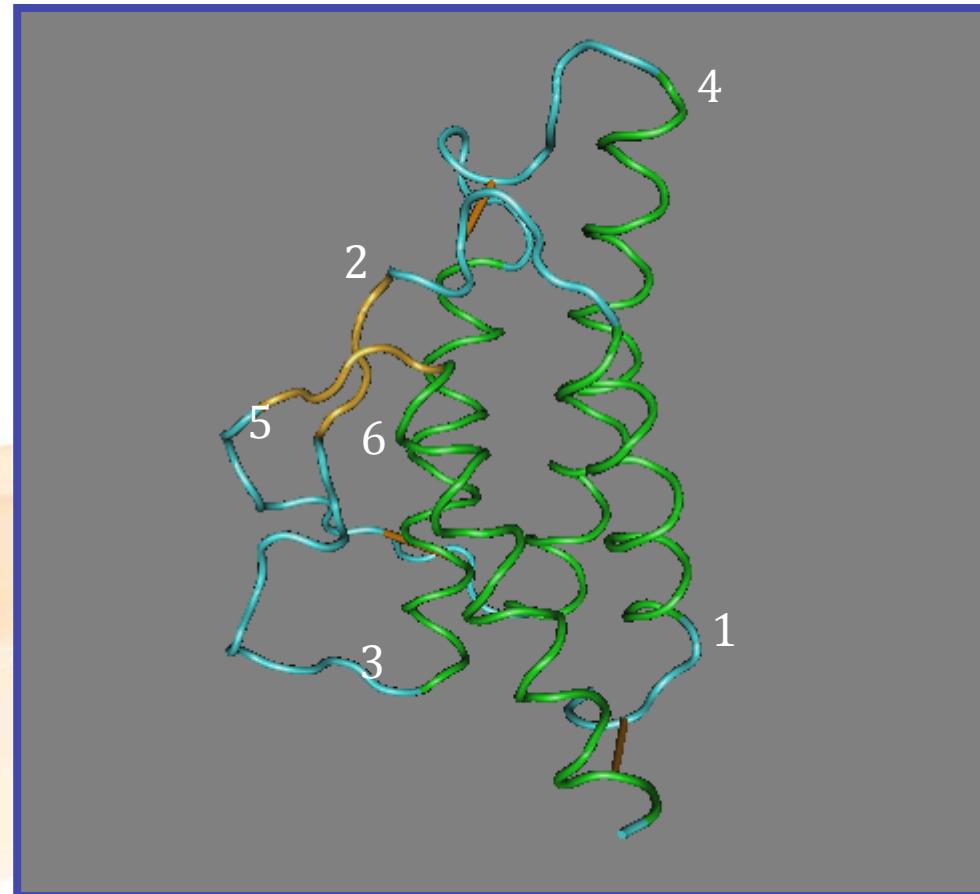
FYI - Structure - VAST



VAST: Related Structures

Vector Alignment Search Tool

For each protein chain



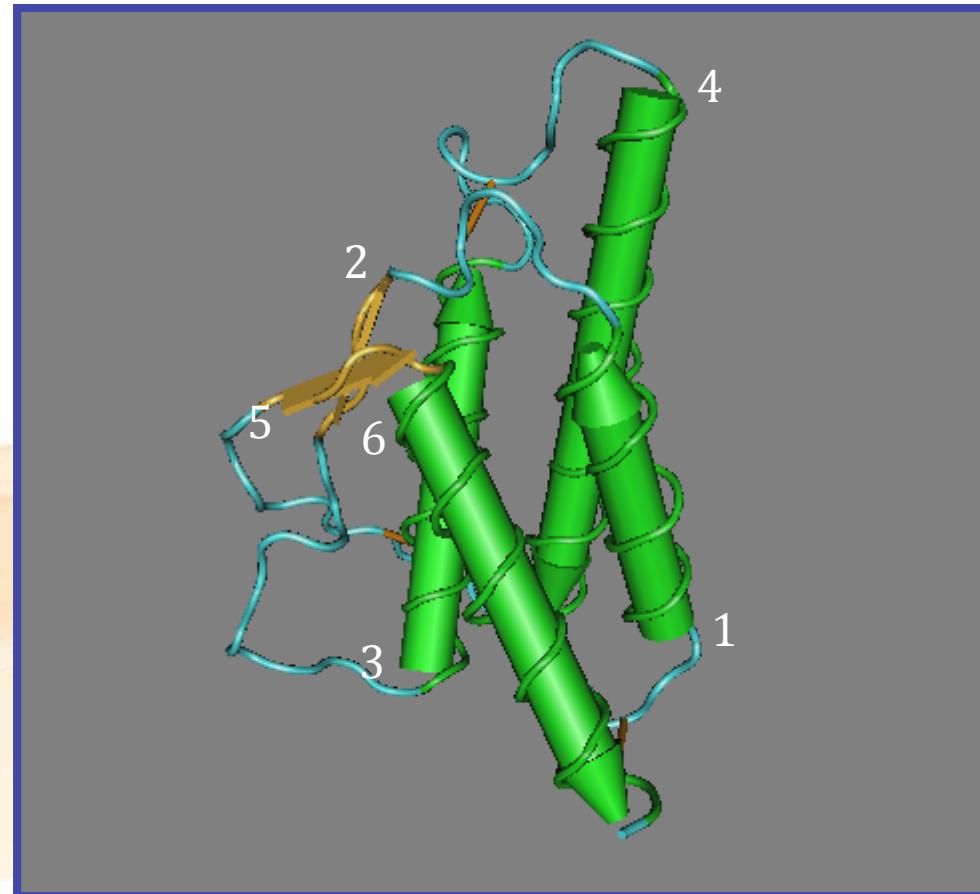
[http://www.ncbi.nlm.nih.gov/Structure/VAST/
vast.shtml](http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml)

Human IL-4

VAST: Related Structures

Vector Alignment Search Tool

For each protein chain
locate SSEs (secondary
structure elements)



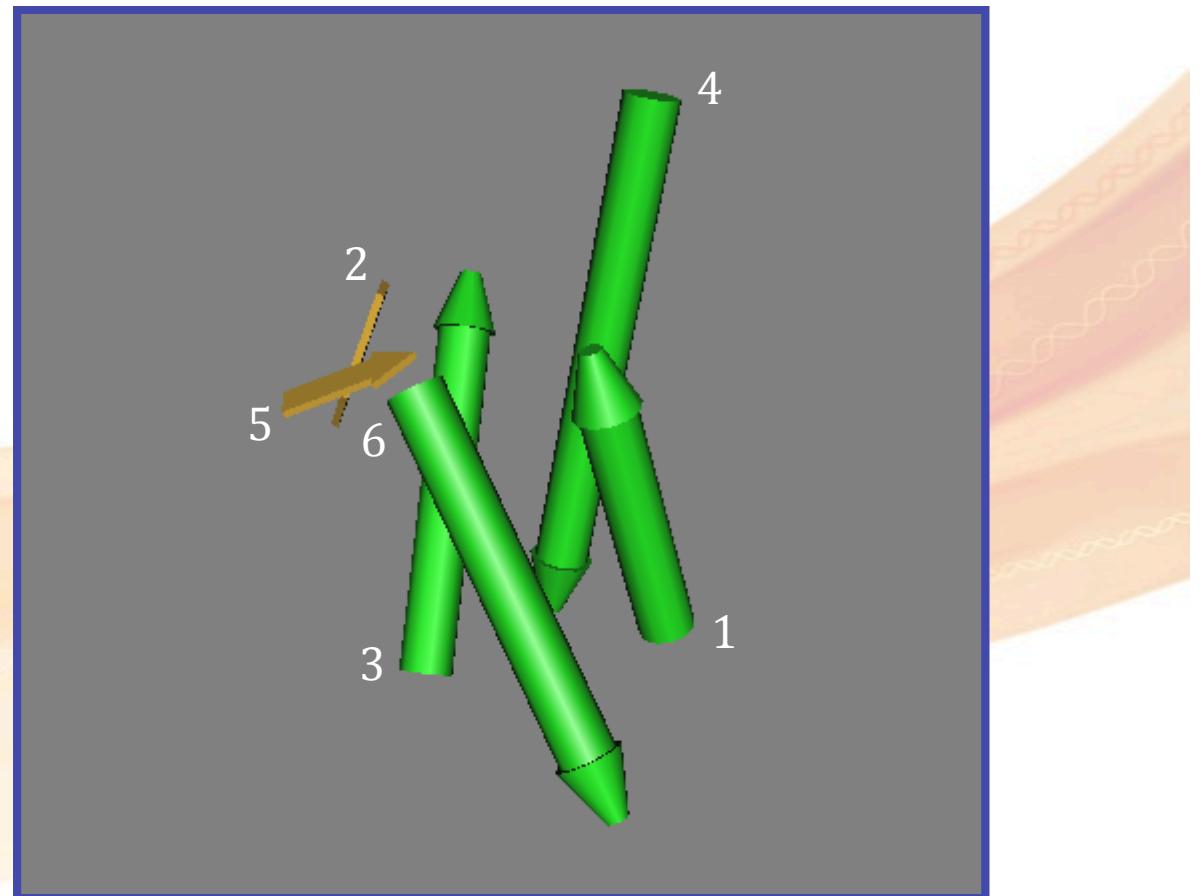
[http://www.ncbi.nlm.nih.gov/Structure/VAST/
vast.shtml](http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml)

Human IL-4

VAST: Related Structures

Vector Alignment Search Tool

For each protein chain
locate SSEs (secondary
structure elements)
and represent them as
individual vectors



[http://www.ncbi.nlm.nih.gov/Structure/VAST/
vast.shtml](http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml)

Human IL-4

VAST: Related Structures

Vector Alignment Search Tool

For each protein chain

locate SSEs (secondary
structure elements)

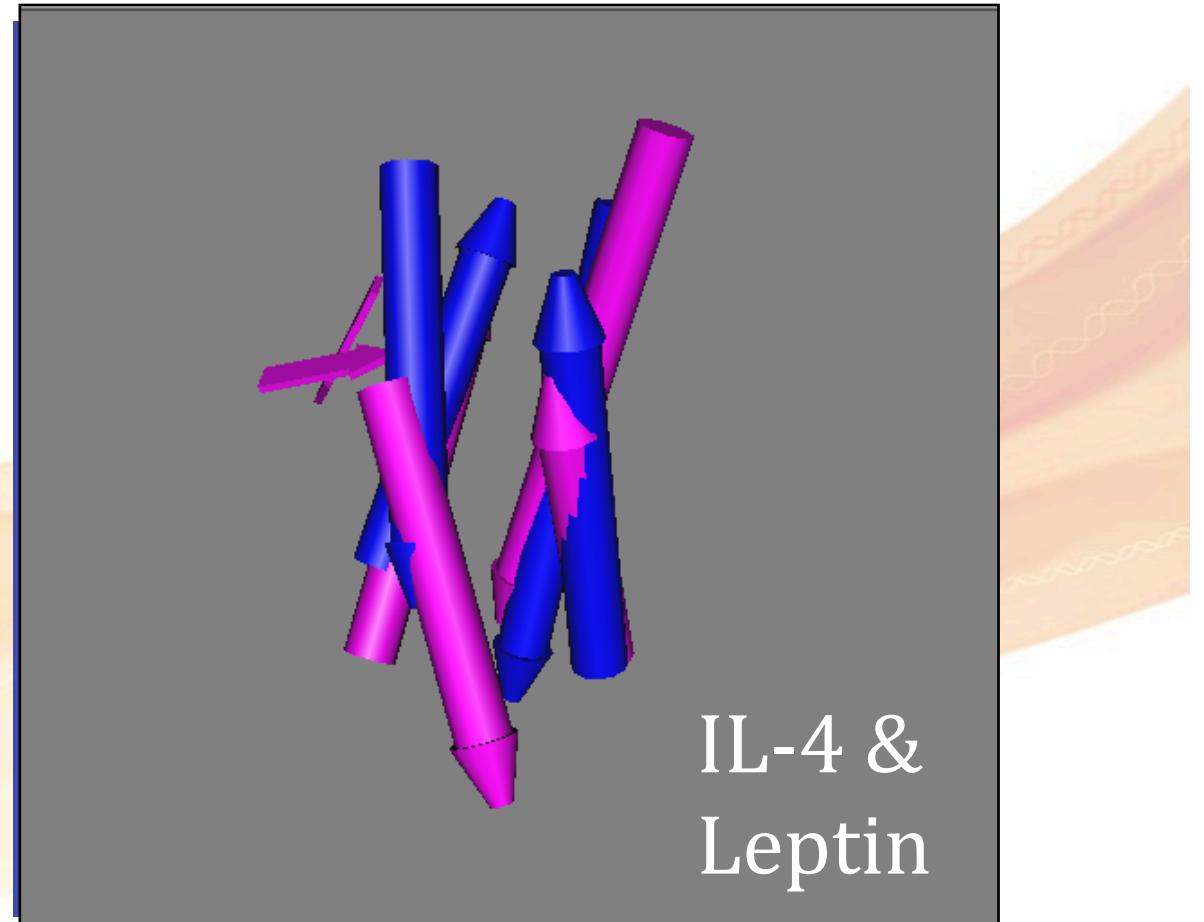
and represent them as
individual vectors

align the vectors

[http://www.ncbi.nlm.nih.gov/Structure/VAST/
vast.shtml](http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml)

Human IL-4

IL-4 &
Leptin



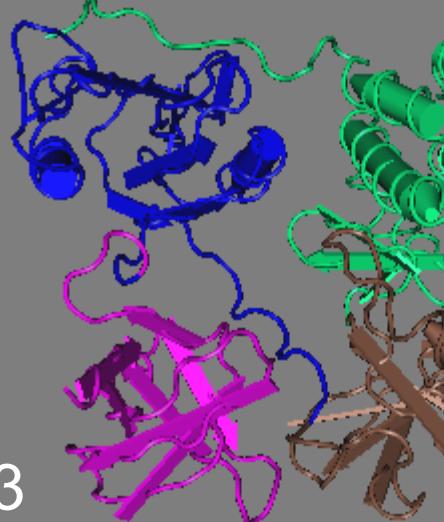
Structure vs Conserved Domain

1FMK - Cn3D 4.1

File View Show/Hide Style Window CDD Help

Conserved phosphotyrosine binding residues

SH2



SH3



SH2
Feature 1

	10	20	30	40	50	60
1AYA_A******
1fmk (query)	#		#			# #
1LKK_A	3	RWFHPNITGVEAENLLltr-gvDGSFLARPsksnpgDFTLSVRRng-----aVTHIKIQN				
2PLD_A	66	EWYFGKITRRESERLLlnaenpRGTFLVREsettkgAYCLSVDfdnakglnVKHYKIRK				
1PIC_A	5	PWFFKNLSRKDAERQLlapgntHGSFLIREsestaGSFSLSVRDfdqnqgevVKHYKIRN				
1CSY_A	10	EWYHASLTRAQAEHMLmrv-prDGAFLVRKrn-epNSYAISFRAeg----kIKHCRVQQ				
1D4T_A	11	TWNVGSSNRNKAENLLrg--krDGTFLVRESS-kqGCYACSVVVdg----eVKHCVINK				
gi 18146618	14	PWFHGKISRESEQIVligsktNGKFLIRARD-nnGSYALCLLHeg----kVLHYRIDK				
gi 1169745	5	AVYHGKISRETGEKLLlat-glDGSYLLRDsesvpGVYCLCVLYhg----yIYTYSQ				
gi 125358	141	DWFFGKIKRTDAEKKLlqagnqTGTFLIREsesqpGNYSLSVREid----tVKHYRIRK				
gi 1174439	115	PWFFGAIIGRSDAEKQLlysenktTGSFLIREsesqkGEFSLSVLDga----vVKHYRIKR				
gi 1536790	143	EWYFGKIGRKDAERQLlspgnpQGAFLIREsettktGAYSLSIRDwdqtrgdhVKHYKIRK				
gi 1174436	121	EWUFGQVKRVDAEKQLmmpfnnLGSFLIRLsdttppGDFSLSVRDid----rVRHYRIKK				
gi 18150844	129	PWYFRKIKRIEAEKLLlpeneHGAFLIRLsesrhNDYSLSVRDgd----tVKHYRIRQ				
gi 3411274	121	EWFLGKIKRVEAEKMLnqsfnnqVGSFLIRDsettpGDFSLSVKDqd----rVRHYRVRR				
gi 125717	118	DWYFGDVGRAEAEEKWLlapgtqSGAFLVRAsstqkNSLSSLRDge----gIKHRYIRT				
gi 729888	116	EWYFKGMSRKEAEKQLlspvnkSGAFMIRDsetmkGCFSLSVRDsg----dtVKHYKIRT				
gi 3721908	125	EWYFGDVKRAEAEEKLmlvrlpSGTFLIRKAetavGNFSLSVRDgd----sVKHYRVRK				
gi 14286176	81	PWFHGKITREQAERLLyp--peTGLFLVREstnypGDYTLCSVCEeg----kVEHYRIMY				
gi 5803171	130	DWYFGKVGRKDAEKKLlapglqKGTFIVRDgeanpGTFSLSVRDydpqkgehIKHRYIRK				
	161	DWUFFENVLRKEADKLLlaeenpRGTFLVRPsehnpNGYSLSVKDwedgrgyhVKHYRIKP				
	83	GULFEGLGRDKAEELLqlpdtkVGSFMIREsetkkGFYSLSVRhr----qVKHYRIFR				

Cn3D

Protein Domains

- What are Protein Domains?
 - Structural Domain
 - Discrete independently folding unit of a protein
 - Conserved Domain (sequence-based)
 - Protein region with recognizable position-specific pattern of sequence conservation
- Sequence-based domains often roughly correspond to structural domains
- Domains often have distinct, identifiable functions

NCBI Conserved Domain Database (CDD)

NCBI's Conserved Domain Database

- Protein annotation resource that consists of a collection of well-annotated multiple sequence alignment models for ancient domains and full-length proteins
- Available as Position-specific scoring matrices (PSSM)
 - For Fast identification of conserved domains in protein sequences via RPS-BLAST
- Sources
 - [SMART](#)
 - [Pfam](#)
 - [COGs](#)
 - NCBI curated domains
 - Structure-informed alignments

<http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>



What is a Conserved Domain

Type 1 Insulin-like Growth Factor Receptor (1IGR),
colored by domain



3D domains are compact structural units identified by purely geometric criteria. Each 3D domain is shown in the same color in both the structure view window and the 3D domains bar graph that shows the span of each domain on the protein chain.

Conserved domains shown here as Domain Families, are recurring units in molecular evolution and appear in protein sequences as conserved blocks of amino acid residues that have distinct functions. Conserved domains serve as building blocks and can be recombined in different arrangements to make proteins with different functions, and often correspond to the 3D domains of a protein structure.

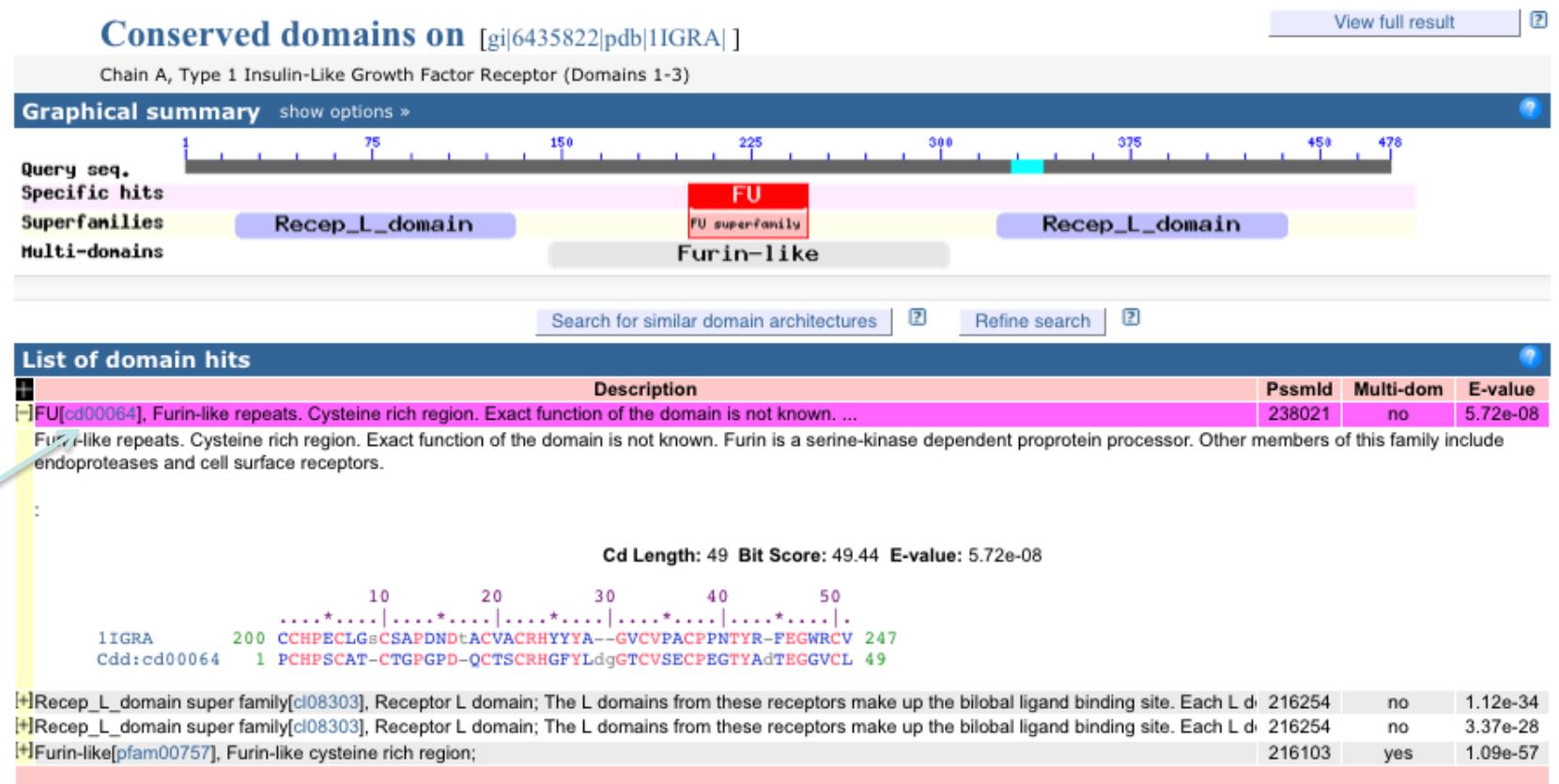
Follow the links in the text below this graphic for additional details and interactive views of the protein structure, conserved domains, and small molecules.

In the live view of the structure record (1IGR, accessible from the text beneath this graphic), click on a conserved domain to view information about its function and the multiple sequence alignment from which the domain model was developed, and to link to other protein sequences that contain the domain.

What is a Conserved Domain @ NCBI

- Distinct functional and/or structural units of a protein
- These two classifications coincide often
- Domains are often identified as recurring (sequence or structure) units
 - May exist in various contexts
 - Previous slide, four "domains" colored in magenta, blue, brown, and green
- In molecular evolution such domains may have been:
 - **Utilized as building blocks**
 - **Recombined in different arrangements** to modulate protein function
- NCBI defines conserved domains - recurring units in molecular evolution, the extents of which can be determined by sequence and structure analysis

Conserved Domain Model



http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi?INPUT_TYPE=precalc&SEQUENCE=6435822

Conserved Domain Model

Sequence Alignment include consensus sequence ?

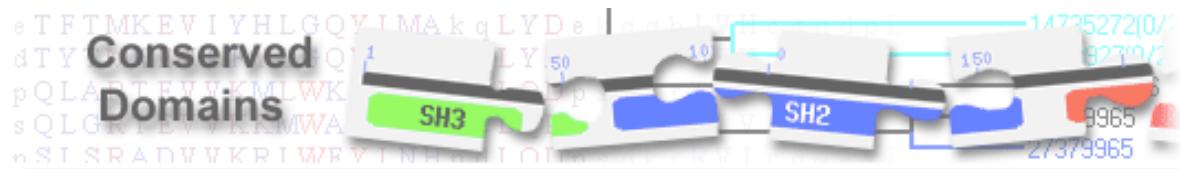
Reformat Format: Hypertext Row Display: up to 20 Color Bits: 2.0 bit Type Selection: the most diverse members

		10	20	30	40	50	60	70	80	
1IGR_A	200	CCHP-----	ECLgSCSApdn-----	dtACwACRhyy-----	yagVCVp--ACPnptyr-----	feg 243				
gi 129542	799	KCHP-----	SCK-KCVDep-----	eKCTvCKegfs-----	largSCIp--DCEPgtfyd-----	sel 842				
gi 232114	1354	KCHH-----	YCK-TCNDag-----	plAcTSCPhsm-----	ldggLCM--ECLssqyyd-----	tts 1397				
gi 423565	1118	DCPE-----	SCL-ICSSa-----	wTC1ACRegft-----	vvhdVCTapkECAaveywd-----	egs 1162				
gi 423565	1023	HCPE-----	RCQ-DCIHe-----	kTCkECMpdff-----	lyndMChr--SPksfyp-----	dm 1063				
gi 4009453	249	SCTDq-----	HCA-FCVAe-----	gTCaKCSSsgfi-----	ldgqNCVksDCKte-----	286				
gi 549130	341	ECTVa-----	NCK-TCNDq-----	gQCqTCNdgy-----	kngdACS--PCHe-----	375				
gi 12643982	741	KCSE-----	NCK-TCTGf-----	hNCtECKgpls-----	lqgsRCsv--TCEDgqff-----	sg 781				
gi 7435769	741	KCSE-----	NCK-TCTEf-----	hNCtECRdgls-----	lqgsRCsv--SCEdgryfn-----	gqdc 784				
gi 17933668	90	G CPLtt-----	ssHCL-RCNKa-----	-GCiKCPmylv-----	tdtrQCVd--QCPagyldqwsa--htefmg	141				
gi 12643811	1391	LCHV-----	NCK-TCHGeg-----	eeDCmECAndiky-----	kqdgrCVt--ECQeghyp-----	dlt 1435				
gi 323072	256	SCGTtnnggiNCG-----	ECTSkesaaragteiTCT-----	KCSSnnlsp-----	plgdACLt--DCPagtavyav-----	gdsg 317				
gi 7514661	962	PCNTp-----	NCK-TCDPkt-----	dneICtKCNdgdy-----	tptnQCVp--DCTaisgyy-----	gdtd 1011				
gi 423565	660	PCPD-----	NCE-LCYNp-----	hICsRCMsgyvi-----	ippnhTCQk-lECRqgefqd-----	sey 705				
gi 126497	695	KCSI-----	SCK-TCSSagr-----	vvqnKC-VCKhveyqp-----	npserICMd--QCPvnsmvp-----	dtnn 747				
gi 1407566	737	ECHP-----	FCT-TCSTgrn-----	kngnGC-VCKfeyhp-----	ndveeICVd--QCPintflipd-----	anesgy 792				
gi 1098514	373	GCHS-----	SCK-DCVTgant-----	seddKC1SCSgdnylkvt-----	ddahsgVCVsasACTsdthftkevadstgsk	438				
gi 115199	217	GCHS-----	SCD-GCTEnamt-----	nqadKtGCKegrylkpe-----	saagqsgACLtaeECTsdkthftre--kagdsk	280				
gi 115199	101	KCNA-----	PCT-ACAGta-----	dKtKCDangaapylkktnpsdptgTCVsavDCQgsagyytd--dsvsda	161					
gi 1098514	121	KCDA-----	TCA-ACSSgl-----	atACTKCEaggatpylkknpgdqgtgTCVskeDCTrdggyyadd--ttdpna	183					

Based on MSA of related protein spanning a variety of organisms to reveal sequence regions that contain the same, or similar patterns of amino acids

<http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?ascbin=8&maxaln=10&seltype=2&uid=238021>

Search with Protein Query



Protein or Nucleotide
Query Sequence

Enter a [protein or nucleotide query](#) as an [accession](#) or [GI](#) number (e.g., AAC50285 or 463989), or as a sequence in [FASTA](#) format, to identify the protein's conserved domains and therefore its putative function:

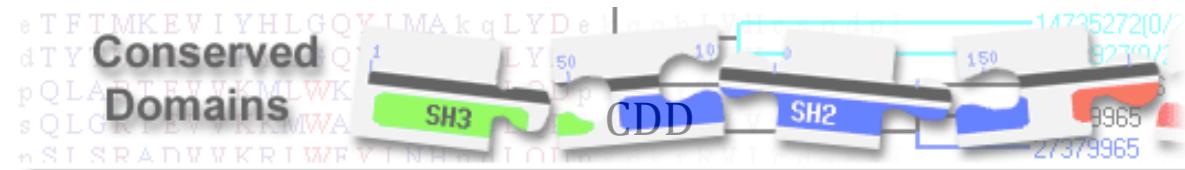
[Submit Query](#)

[Search Database](#)

CDD v3.10 – 44354 PSSMs

```
>gi|45549418|gb|AAS67634.1| ATP7A [Solenodon paradoxus]
IVYQPHLITVEEIKKQIKAVGFPAPIKKQPKYLKLGAIDIERLKNIPVKVSSEGSQQMS
PSSTNDSKVTLTIDGMHCNSCVSNIESALSTLHYVSSIVVSLQNKSIIKYNANSVT
PEILKKAIEAISPGQYRVSITSEVESTNSPSSSSQKAPLNVSQPLTVTWINING
MTCNSCVQSIEGVMSKKAGVKSIQVSLANRNGTVEYDPLLT SPEILRE
```

http://www.ncbi.nlm.nih.gov/Structure/cdd/docs/cdd_search.html

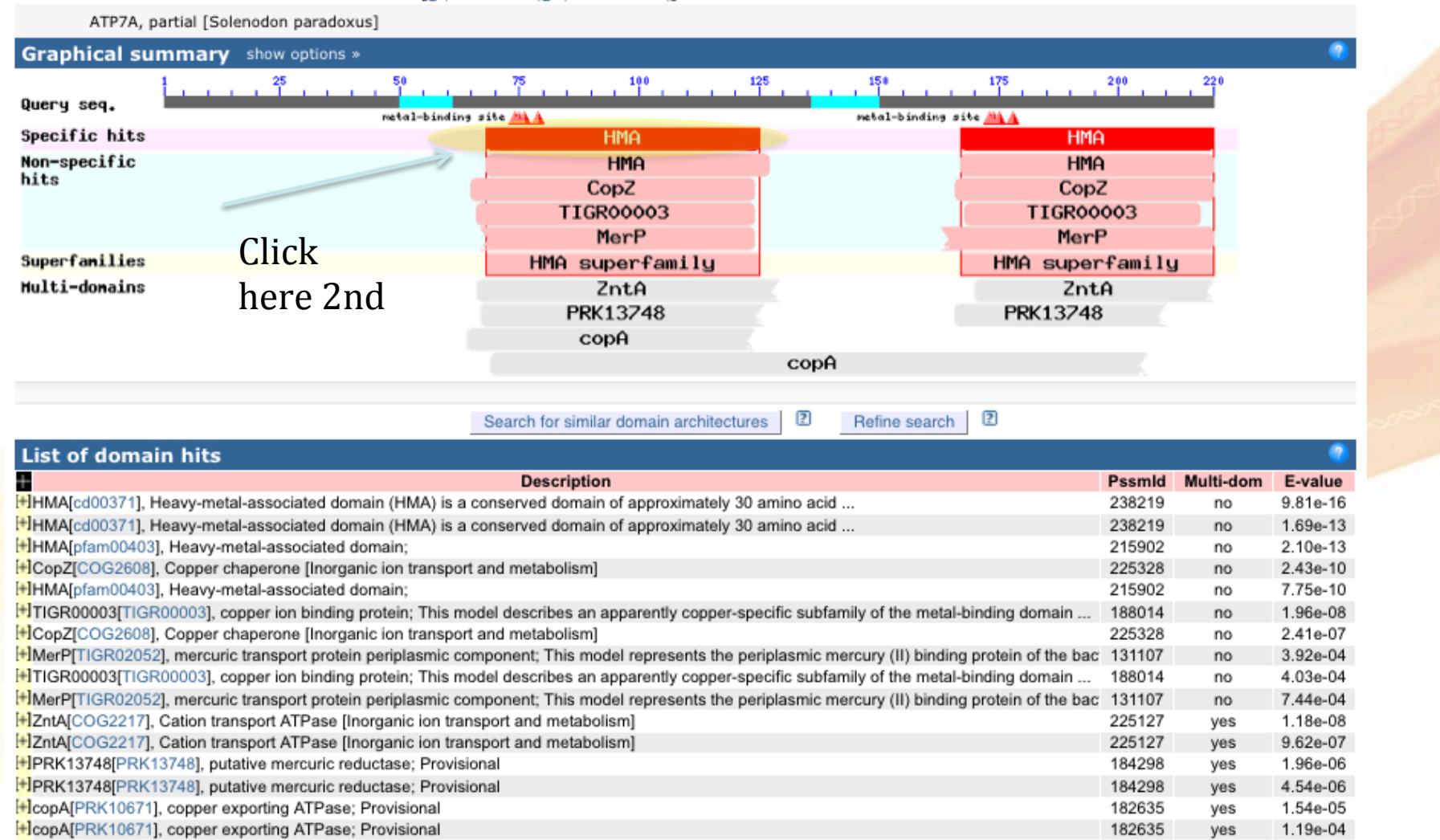


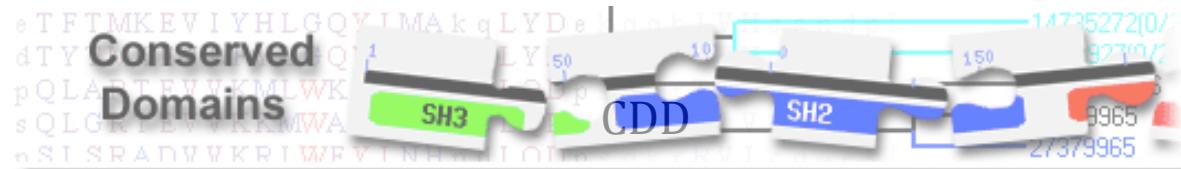
Click here 1st

Previous protein query – Found conserved domains

Conserved domains on [gi|45549418|gb|AAS67634]

[View concise result](#)



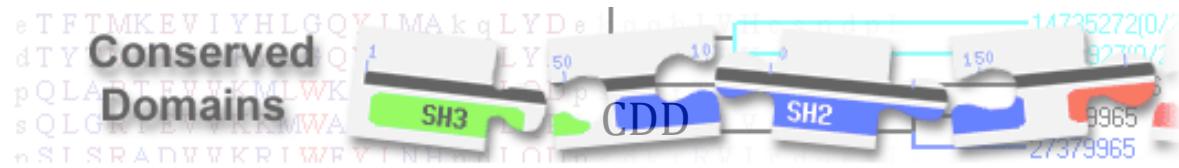


Previous protein query – Found conserved domains

Sequence Alignment include consensus sequence ?

Reformat Format: Hypertext Row Display: up to 20 Color Bits: 2.0 bit Type Selection: the most similar members

	10	20	30	40	50	60	
Feature 1*******
1CPZ_A	3 EFSVKg MSC NH C VARIEEAVGRI S GVkKVKVQLk K A V V K F D e----nv Q ATEIC Q A I N E LG Y QAE 66	### #					
query	68 TLTIDg MHC NSCVSNIESALSTL H YV S IVVSL q nk S AI I KY N an---sv T PEILKK A IE A isp g qy 131						
1AFJ	6 TLAVPg MTC AA C PITVK K ALSK V e G V S KVDVG F ek R AVV T FD d ----ka S V Q KL T K A T D AG P SS 69						
1K0V_A	5 TLQVEg MSC Q H CV K AV E TSV G EL d G V S AV H V N Le a q K DV V SD f ad----kv S V K DI A DA I E D Q G Y D VA 68						
gi_2635862	7 EYVLDg LDC SN C ARK I ENG V K G I k G In G CAVN F a a st T LT V S A d g ke--eqw V T N K V E K V K S ID P H V T 72						
gi_120199	40 ELSVPn A YC G T C IAT E GA R AK p E V R A V N L s sr R V S I V W K eevg g rrt N PCDFL H IA E RG Y Q T H 107						
gi_10764865	17 AVYKV h L H R K C ACD I KKPL L RF q G V q N V DF D le k n E I K V K G i -----e V VI H K Q I E K W SK K V E 78						
gi_3108347	24 EFMVD -MTC EG C VNA V N K LET I e G I e K V E D L s nb V R I L G s -----PV K AM T Q A LE Q T G R K AR 83						
gi_2159998	8 GLSVAg MTC P C STR H VED D ALL L V p G V t R A V D Y p sn K A Q V T G N r-----1 D V S AL V A V G A LG Y G A T 69						
gi_3929319	12 EFMVD -MTC EG C VSA V N S ML K b G V s G V D V D L b s nb L V R V I G S v-----PV K T M L K A L E Q T G R N AR 71						
gi_3122077	39 DLSV s d V H C GG C I S T I E R ALL L T p F V k T A R V N L tar R V T C V Y Q ee <i>iear</i> at D P S K I L G E I N S A GY R AH 106						
gi_2493001	6 NLQ L eg MRC AA C ASS I E R AI K V p G V q SC Q V N fale Q A V V S Y H ge----t T P Q I L T D A V E R AG Y HA R 68						
gi_2493016	102 QLLS g MSC AS C TRV Q N Q LS V p G V t Q A R V N L a e r T AL V M G s -----SP Q DL V Q A VE K AG Y G A E 162						
gi_1703455	260 QLRIDg MHC K C VLN I E E N I G Q L l G V q SI Q V S Len k T A Q V K Y D p s----ct S P V AL Q R A IE A L P PG N F 323						
gi_5759320	95 EFMVD -MTC Q G CVSA V K S KL Q T V e G V k N V D V D L dn q V R I L G s -----PV K T M T E A L E Q T G R K AR 154						
gi_231677	15 AYRVQg FTC AN C AG K FE K N V K Q L s G V e DA K V N F g as K IA V G N a-----T I E E LE K A G A F E N L K V T 75						
gi_3080439	35 ELKVR -MDC D G C V L K I K N S L S SL k G V k T V E I N K qq K V T V S GY a -----D A SK V L K A K A T G K AE 94						
gi_728935	83 LLSVQg MTC G C V S T V T K Q V E G V e S V V V S L v t e E CH V I Y p s----kt T LET A R E M I E D CG F DS N 146						
gi_12643938	403 MLAIAg MTC K C V Q S I E G L I S Q R v G V h Q I S V F l a e g T AV V L Y D p s----rt H PE E LR A VE D MG F EA S 466						
gi_1703455	61 TVRILg MTC Q C V K S I E D R I S N L k G I i S M K V S L eq d S A T V K Y p s----vv C L Q Q V CH Q I G DM G FE A S 124						



Previous protein query – Found conserved domains

Graphical summary [show options »](#) [?](#)

Query seq. 1 25 50 75 100 125 150 175 200 220

Specific hits: metal-binding site ▲▲ HMA metal-binding site ▲▲ HMA

Superfamilies: HMA superfamily HMA superfamily

[Search for similar domain architectures](#) [Refine search](#) [?](#)

List of domain hits [?](#)

	Description	PssmId	Multi-dom	E-value
[+]	HMA[cd00371], Heavy-metal-associated domain (HMA) is a conserved domain of approximately 30 amino acid ... Heavy-metal-associated domain (HMA) is a conserved domain of approximately 30 amino acid residues found in a number of proteins that transport or detoxify heavy metals, for example, the CPx-type heavy metal ATPases and copper chaperones. HMA domain contains two cysteine residues that are important in binding and transfer of metal ions, such as copper, cadmium, cobalt and zinc. In the case of copper, stoichiometry of binding is one Cu ⁺ ion per binding domain. Repeats of the HMA domain in copper chaperone has been associated with Menkes/Wilson disease due to binding of multiple copper ions.	29471	no	1.11e-10
[+]	Cd Length: 63 Bit Score: 61.78 E-value: 1.11e-10 10 20 30 40 50*....*....*....*....*....*.... gi 45549418 68 TLTIDGMHCNSCVSNIESALSTLHYVSSIVSILQNKSAIIKYNAAnSVTPEILKKAIEA 125 Cdd:cd00371 1 ELSVEGMTCAAGCVSKIEKALEKLPGVESVEVDLETGKATVEYDP-EVSPEELLEAIED 57			
[+]	HMA[cd00371], Heavy-metal-associated domain (HMA) is a conserved domain of approximately 30 amino acid ... Heavy-metal-associated domain (HMA) is a conserved domain of approximately 30 amino acid residues found in a number of proteins that transport or detoxify heavy metals, for example, the CPx-type heavy metal ATPases and copper chaperones. HMA domain contains two cysteine residues that are important in binding and transfer of metal ions, such as copper, cadmium, cobalt and zinc. In the case of copper, stoichiometry of binding is one Cu ⁺ ion per binding domain. Repeats of the HMA domain in copper chaperone has been associated with Menkes/Wilson disease due to binding of multiple copper ions.	29471	no	2.86e-09

Go Back
To
Search

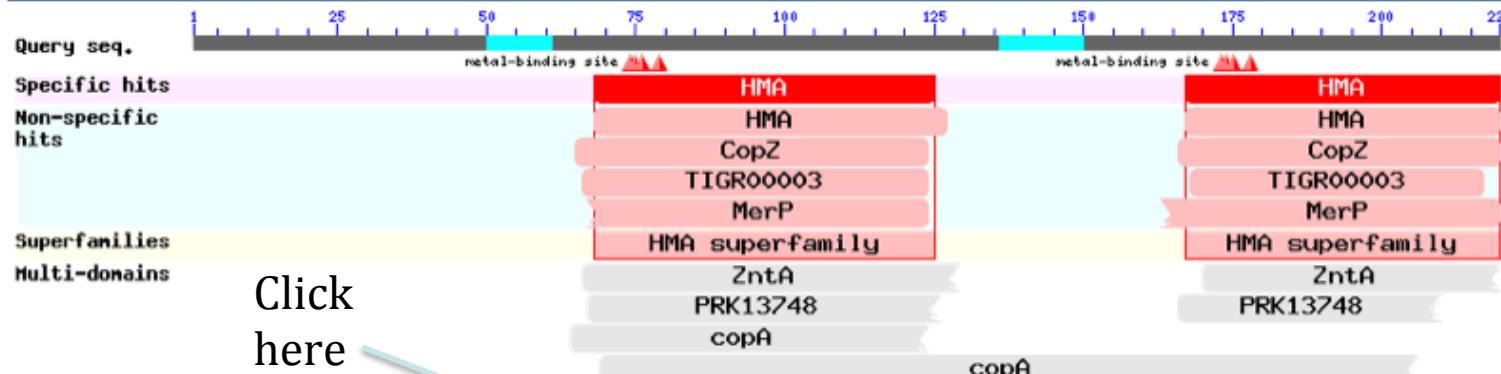
Previous protein query – Found conserved domains

Conserved domains on [gi|45549418|gb|AAS67634]

[View concise result](#)

ATP7A, partial [Solenodon paradoxus]

Graphical summary [show options >](#)



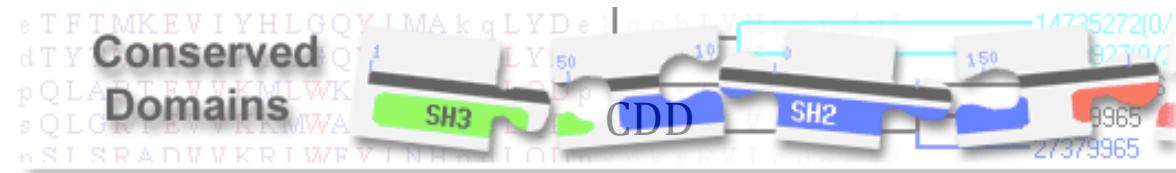
Click
here

[Search for similar domain architectures](#)

[Refine search](#)

List of domain hits

	Description	PssmId	Multi-dom	E-value
[+]	HMA[cd00371], Heavy-metal-associated domain (HMA) is a conserved domain of approximately 30 amino acid ...	238219	no	9.81e-16
[+]	HMA[cd00371], Heavy-metal-associated domain (HMA) is a conserved domain of approximately 30 amino acid ...	238219	no	1.69e-13
[+]	HMA[pfam00403], Heavy-metal-associated domain;	215902	no	2.10e-13
[+]	CopZ[COG2608], Copper chaperone [Inorganic ion transport and metabolism]	225328	no	2.43e-10
[+]	HMA[pfam00403], Heavy-metal-associated domain;	215902	no	7.75e-10
[+]	TIGR00003[TIGR00003], copper ion binding protein; This model describes an apparently copper-specific subfamily of the metal-binding domain ...	188014	no	1.96e-08
[+]	CopZ[COG2608], Copper chaperone [Inorganic ion transport and metabolism]	225328	no	2.41e-07
[+]	MerP[TIGR02052], mercuric transport protein periplasmic component; This model represents the periplasmic mercury (II) binding protein of the bac	131107	no	3.92e-04
[+]	TIGR00003[TIGR00003], copper ion binding protein; This model describes an apparently copper-specific subfamily of the metal-binding domain ...	188014	no	4.03e-04
[+]	MerP[TIGR02052], mercuric transport protein periplasmic component; This model represents the periplasmic mercury (II) binding protein of the bac	131107	no	7.44e-04
[+]	ZntA[COG2217], Cation transport ATPase [Inorganic ion transport and metabolism]	225127	yes	1.18e-08
[+]	ZntA[COG2217], Cation transport ATPase [Inorganic ion transport and metabolism]	225127	yes	9.62e-07
[+]	PRK13748[PRK13748], putative mercuric reductase; Provisional	184298	yes	1.96e-06
[+]	PRK13748[PRK13748], putative mercuric reductase; Provisional	184298	yes	4.54e-06
[+]	copA[PRK10671], copper exporting ATPase; Provisional	182635	yes	1.54e-05
[+]	copA[PRK10671], copper exporting ATPase; Provisional	182635	yes	1.19e-04

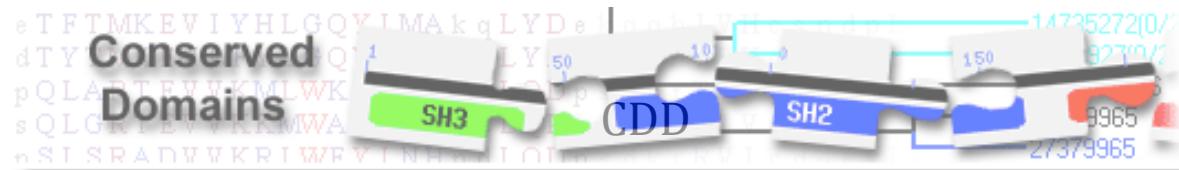


CDART – Conserved Domain Architecture Retrieval Tool

[Query] >gi|45549418|gb|[AAS67634.1]
ATP7A [Solenodon paradoxus]
IVYQPHLITVEEIKKQIKAVGFPFAFIKKQPKYLI
Total architectures: 44

The diagram illustrates the protein architecture of ATP7A. The sequence is represented by a horizontal bar with vertical tick marks at intervals of 125 amino acids. Domains are shown as colored boxes: HMA (green), E1-E2_ATPase (light green), HAD_like (yellow-green), Ribosomal_L2 (purple), NRDB_Rossma (blue), Pyr_redox_dim (light blue), Cu-Zn_Superoxide_Dismutase (orange), Cation_efflux (yellow), and MerT (light green). The HMA domain is present in several locations, including the N-terminus, between 125 and 220, and between 250 and 375. The E1-E2_ATPase domain is a large central region from approximately 250 to 500. The HAD_like domain is located between 500 and 732. Other domains are more sparsely distributed.

Architecture	Description	Taxonomy span	Similarity score	Total nr sequences
ATP7A	cellular organisms	1	1	4118
heavy metal-transporting ATPase	cellular organisms	1	1	3885
zinc/cadmium/mercury/lead-transporting	cellular organisms	1	1	486
heavy metal translocating P-type ATPase	cellular organisms	1	1	457
mercuric ion reductase	Bacteria	1	1	189
copper chaperone for superoxide	Eukaryota	1	1	183
cation-transporting ATPase, P-type	cellular organisms	1	1	158
pit1	Bacteria	1	1	79
ATPase P	cellular organisms	1	1	36
probable copper-transporting ATPase	cellular organisms	1	1	33



CDART – Conserved Domain Architecture Retrieval Tool

ATP7A
taxonomy span: cellular organisms
Similarity score: 1
Total nr sequences: 4118

[Lookup sequences in Entrez](#)

Sequence ID	Organism	Start	End	Length	Action
gi 12699450 AAG47428	ATP7A [Tamias striatus]	50	100	50	domain details >
gi 18418567 NP_567975	heavy metal transport/detoxification	50	100	50	domain details >
gi 12699515 AAG47460	ATP7A [Lama glama]	50	100	50	domain details >
gi 15925076 NP_372610	hypothetical protein SAV2086	50	69	19	domain details >
gi 21283738 NP_646826	hypothetical protein MW2009	50	69	19	domain details >
gi 21165909 AAL47253	ATP7A [Tadarida brasiliensis]	50	100	50	domain details >
gi 4469010 CAB38271	hypothetical protein [Arabidopsis]	50	100	50	domain details >
gi 18413973 NP_568105	heavy metal transport/detoxification	50	100	50	domain details >
gi 17229909 NP_486457	hypothetical protein asl2417 [Nostoc	50	66	16	domain details >
gi 26987326 NP_742751	heavy metal transport/detoxification	50	65	15	domain details >



5 Minutes..

NCBI – dbSNP Short Genetic Variations



Finding SNPs in the Human Genome

- Challenge
 - Identify SNPs that correlate with a particular effect in patients
 - Reliable SNPs could serve as **predictive markers**
 - **Inform our decisions** about numerous aspects of medical care, including:
 - specific diseases
 - effectiveness of various drugs
 - adverse reactions to specific drug
- Scientists **approach** the problem of **identifying, cataloging and characterizing** SNPs in **two** main ways:
 - Genomic approaches
 - Functional approaches

<http://learn.genetics.utah.edu/content/health/pharma/snips/>

Genomic Approaches

Finding SNPs in the Human Genome

- Used by scientists who want to see **the big picture**
- Several large-scale projects - combined the efforts of many institutions:
 - To identify and catalog all of the SNPs in the 3-billion-base pair human genome
 - Each project involves **hundreds** of scientists
 - **Compare genomes** of numerous individuals to identify the differences
 - Comparisons require great deal of computer-powered data analysis
- Scientists sort and catalog results in databases available to anyone over the internet
 - **dbSNP at NCBI**

<http://learn.genetics.utah.edu/content/health/pharma/snips/>

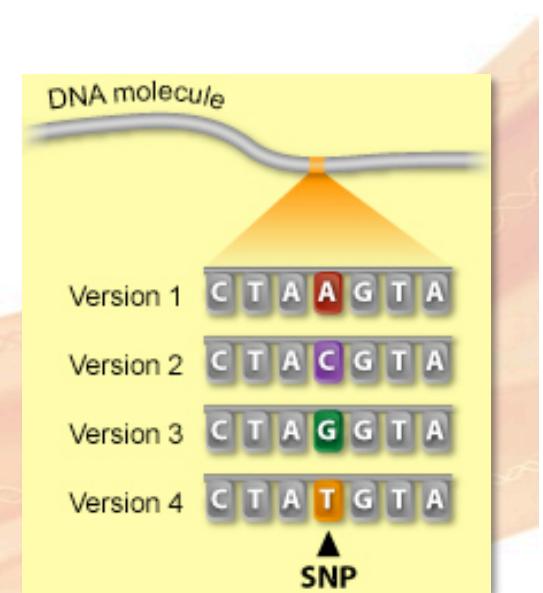
Functional Approaches

Finding SNPs in the Human Genome

- Used by scientists who are interested in a **particular disease or drug response**
 - The biological processes involved in diseases and drug responses are controlled by the activities of many genes
 - **Scientists interested in a particular process select genes known to be involved in the process and examine them in people who:**
 - Have a **response or disease**
 - As well as **those who do not**
- By comparing people's DNA sequences
 - Scientists can identify SNPs that correspond with a particular **function or response**
 - **This is where we will make major changes in the way we live**
 - **Many biotech in the area are going this!**

SNP Quick Reference

- SNPs are single-nucleotide substitutions of one base for another
- Each SNP location in the genome can have up to four versions:
 - A, C, G and T
 - A SNP and its distribution in a population might look like the image right
- Not all single-nucleotide changes are SNPs
 - **To be classified as a SNP:**
 - **Two or more** versions of a **sequence** must each be **present**
 - **>= one percent of the general population**
- SNPs occur throughout the human genome
 - **One** in every **300** nucleotide base pairs
 - **~10 million SNPs** within the **3-billion-nucleotide** human genome
 - So much work to do!

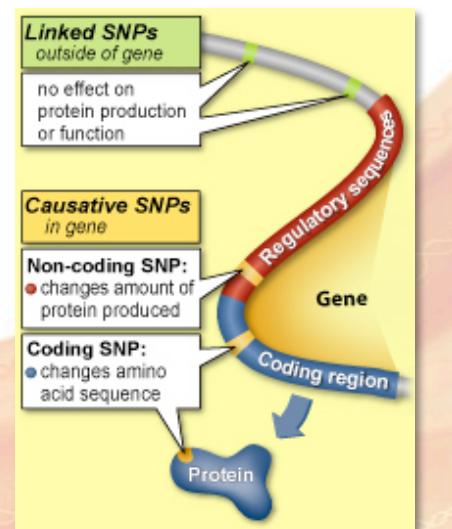


SNPs and Disease-Causing Mutations: Not the Same!

- If you know what a point mutation is, then description of a SNP might sound similar
- Both are single-nucleotide differences in a DNA sequence
- But SNPs should not be confused with disease-causing mutations
 - **First:**
 - SNP - one percent of the general population
 - **No known disease-causing mutation is this common**
 - **Second:**
 - **Most disease-causing mutations** occur within a gene's **coding or regulatory regions**
 - **Affect the function** of the protein encoded by the gene
- Unlike mutations:
 - **SNPs are not necessarily located within genes**
 - Do not always affect the way a protein functions

Linked and Causative SNPs

- **Linked SNPs**
 - **Do not reside** within genes and **do not affect** protein function
 - **Correspond to a particular drug response** or to the risk for getting a certain disease
- **Causative SNPs**
 - **Affect the way a protein functions:**
 - Correlating with a disease or influencing a person's response to medication
 - Causative SNPs come in two forms:
 - **Coding SNPs**
 - Located within coding region of a gene
 - Change amino acid sequence of gene's protein product
 - **Non-coding SNPs**
 - Located within gene's regulatory sequences
 - Change level of gene expression and, therefore, how much RNA and protein is produced



SNPs – Intragenic and Extragenic

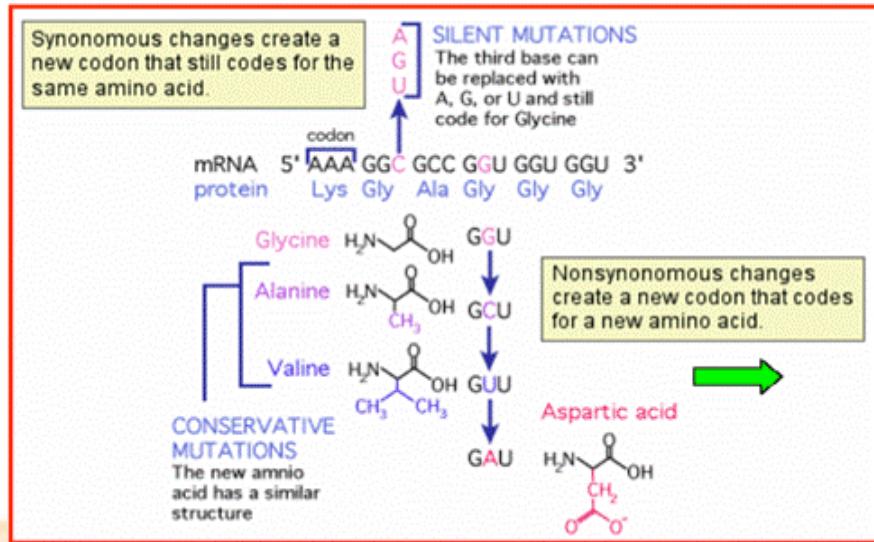
- Intragenic SNPs are often categorized by function
 - Are they in a coding region, an intron, part of the mRNA, outside the mRNA but still in the gene locus (i.e., in the promoter)
- Extragenic SNPs
 - Simply considered ‘genomic’ or might be labeled relative to the nearest gene
 - ie. 5' or 3' to a gene
 - An ‘extragenic’ SNP may affect regulatory regions important in gene expression or other DNA functions such as DNA replication

SNP Functional Categories

- Coding nonsynonymous
 - Missense, nonsense, frame shift
- Coding synonymous
- Intronic
 - splice site
- mRNA utr
 - 5' utr or 3' utr
- (gene) locus region (5' or 3' to the gene)
 - ‘near gene’ usually means within ~2000bp of gene
- genomic/extragenic (distant from any gene)

Coding Nonsynonymous SNPs

- Missense – change an aa
- Nonsense
 - Change an aa to a stop codon
 - Results in a shortened protein
- Frame Shift
 - Are really single-base indels
 - Drop or add one base and the triplet reading frame is thrown out of shift, altering all downstream aa's and usually resulting in an earlier stop codon



SNP Nomenclature

- The Human Genome Variation Society
 - <http://www.hgvs.org/mutnomen/recs.html>
 - Proposed some guidelines for SNP nomenclature, but at the moment, there is minimal consistency
- Different sources will refer to the same SNP in different ways
- While dbSNP identifiers (rs#12345678) are becoming common, they are not required of publishing authors and not used in all cases

SNPs at Base-Pair Level

- The base-pair change is given in various forms:
 - A/C
 - T→G
 - C>T
 - 432G>C
 - T73C
- The HGVS nomenclature recommendations:
 - "c." for a coding DNA sequence (like c.76A>T)
 - "g." for a genomic sequence (like g.476A>T)
 - "m." for a mitochondrial sequence (like m.8993T>C)
 - "r." for an RNA sequence (like r.76a>u)

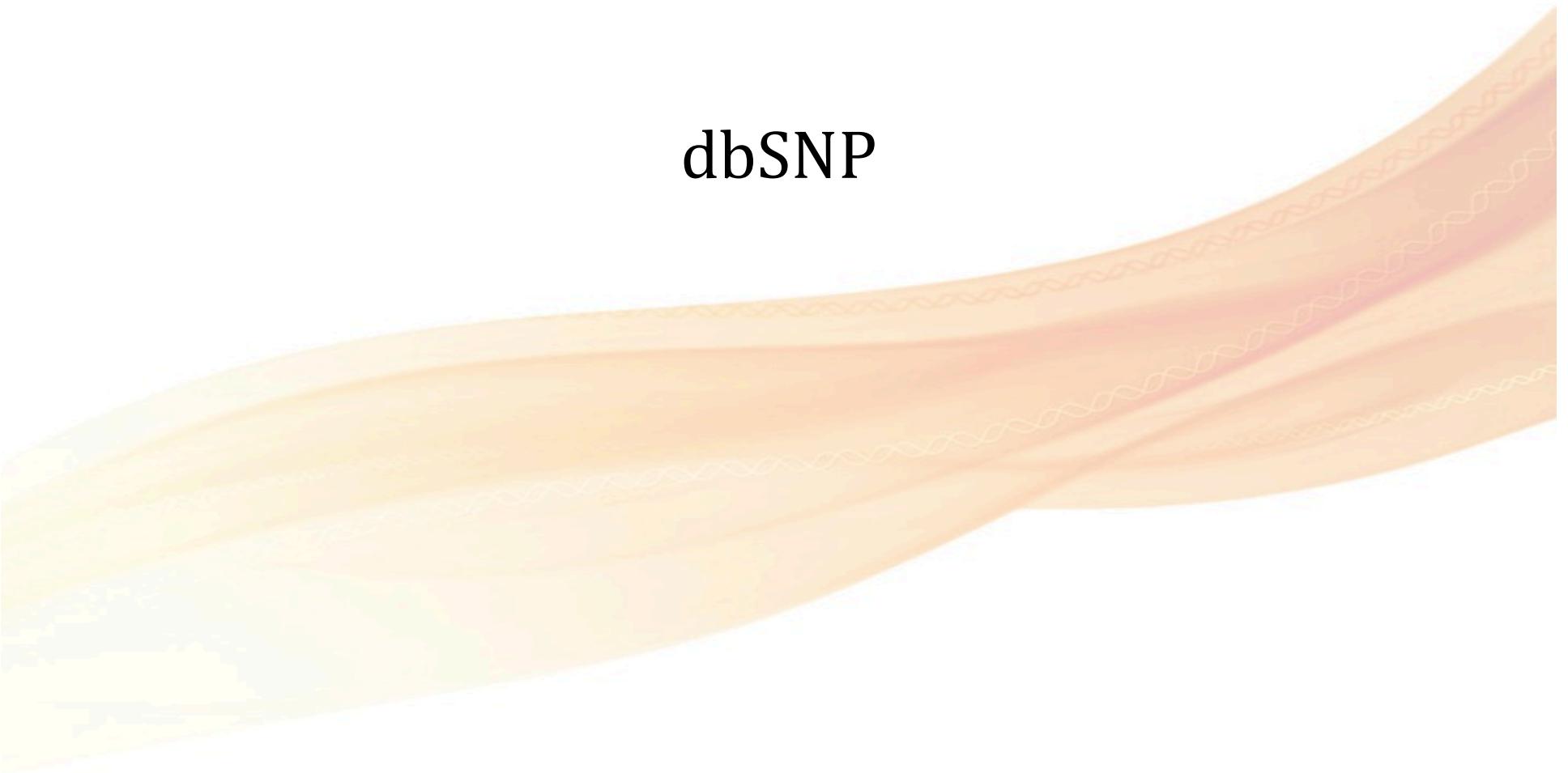
Position, position, position!

- The big issue with SNPs is identifying their location (**numerically**)
- Position can be specified:
 - Number location within a specific sequence
 - Relative to another genetic landmark
 - Start site for a coding region of a gene
 - Start or end of an exon or intron
 - Relative to a marker
 - Position in structure
- Published articles are not always clear on this!!!
- Different resources may use different landmarks/numbering
- Numbering is always relative to the chosen sequence/structure

Coding SNPs

- These are easier because they can be identified by the amino acid position rather than the base-pair position
- Most common nomenclature uses either 3-letter or single amino acid codes:
 - **Asn332Asp OR A95V**
- The HGVS recommendation is similar:
 - "p." for a protein sequence (like p.Lys76Asn)
- Amino Acid (protein) coding sequence positions becoming more consistent, but are not always consistent

dbSNP



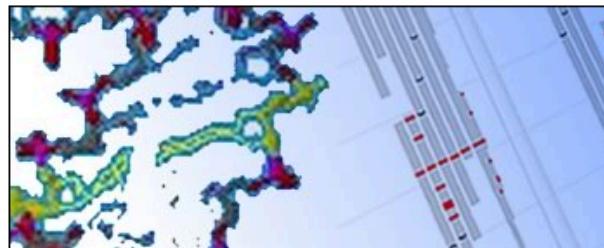
dbSNP

SNP

Limits Advanced

Search

Help



dbSNP

Database of single nucleotide polymorphisms (SNPs) and multiple small-scale variations that include insertions/deletions, microsatellites, and non-polymorphic variants.

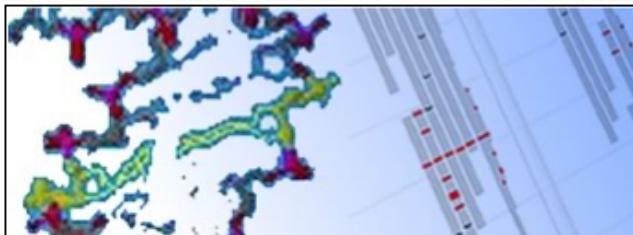
Short Genetic Variations Database of Single Nucleotide Polymorphisms

- Variation
 - Reflects the wide range of dbSNP's variation content
 - Hopefully prevents further misunderstandings about scope of the resource
- dbSNP also stores common and rare variations along with their genotypes and allele frequencies.
- Includes clinically significant variations

Stephen Sherry, PhD

NCBI's dbSNP Database

- International central repository for both **single base nucleotide substitutions** and **short deletion** and **insertion polymorphisms**
- Accepts data submissions from scientists
- Integrated with the NCBI's Entrez system
- Primary Database and Derivative (RefSNP)
- **Single Nucleotide Polymorphisms**
- **Repeat polymorphisms**
- **Insertion-Deletion Polymorphisms**
- **dbSNP build 131 (Mar 25, 2010)**
- **15 Organisms**
- **180,113,082 submissions (submitted SNPs)**
- **validated 24,735,618 reference SNPs**



dbSNP

Database of single nucleotide polymorphisms (SNPs) and multiple small-scale variations that include insertions/deletions, microsatellites, and non-polymorphic variants.

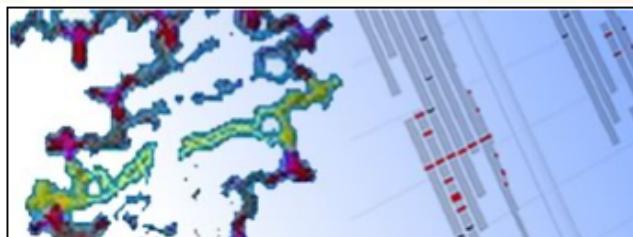
NCBI's dbSNP Database

<http://www.ncbi.nlm.nih.gov/sites/entrez?db=snp>

http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi

- **dbSNP build 138 (Apr 35, 2013)**
- **131 Organisms**
- **505,719,558 submissions (submitted SNPs)**
- **validated 70,316,939 reference SNPs**

Tremendous Growth
in 3 years

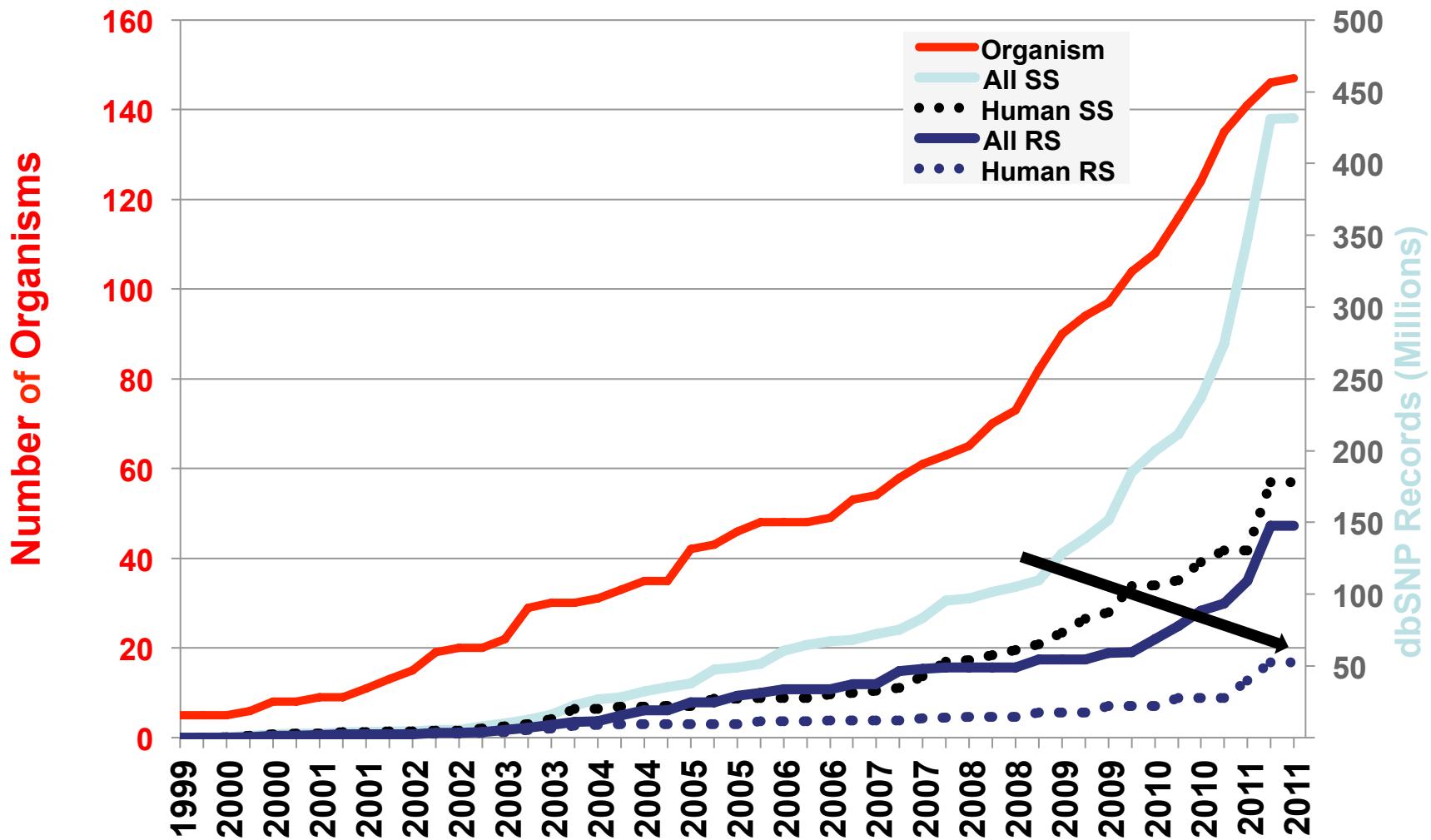


dbSNP

Database of single nucleotide polymorphisms (SNPs) and multiple small-scale variations that include insertions/deletions, microsatellites, and non-polymorphic variants.

dbSNP Content

- Two major classes of content:
- Submitted data, i.e., original observations of sequence variation:
 - Submitted SNPs (SS) with ss# (ss 5586300)
- Computed/curated data:
 - Reference SNP Clusters (Ref SNP) with rs# (rs 4986582)
- Ref SNP clusters are computer-generated and **curated by NCBI staff**
 - Ref SNP Clusters define a non-redundant set of SNPs
 - All individual SNPs:
 - Submitted by a researcher are given a submitter SNP number (ss#)
 - Then redundant (repetitive) submitter SNPs are combined into a RefSNP cluster record
 - With a unique rs#
 - Ref SNP clusters may contain multiple submitted SNPs



505,719,558 submissions (submitted SNPs) - 232,952,851 - Human
validated 70,316,939 reference SNPs - 44,278,189 - Human

Adopted Stephen Sherry, PhD

Types of Genetic Variations

- **Single Nucleotide Polymorphisms (SNP)**

- Single base pair changes

GTCA**T**TCGATT

GTCA**G**TCGATT

- **Indels**

- Small insertion/deletions

CTT---GATC

CTT**ACG**GATC

- **Small variable repeats - microsatellites**

ACGACGACGACGACGACG (6 copies)

ACGACGACGACGACGACGACG (7 copies)

- **Variable Long tandem repeats (can be dozens to hundreds to thousands)**

- **Chromosomal Aberrations: Translocations, Inversions, etc.**

Variation Classes

Abbreviation	Description	Example
SNP	Single Nucleotide Variation	A/G
MNP	Multiple Nucleotide Variation, where all alleles are same length, and length > 1	AT/GA
IN DEL	Deletion/Insertion Variation	-/T
MIXED	Variation has unknown sequence composition but is observed to be heterozygous	HETEROZYGOUS
MICROSATELLITE	Microsatellite/simple sequence repeat	(GACA)3/4/5
NAMED	Allele sequences defined by name tag instead of raw sequence	(Alu)-
NO VARIATION	Submission reports invariant region in surveyed sequence	NOVARIATION
MIXED	Mixture of two or more of the above classes	A/G/-

New SNP Attributes (March 15, 2011)

New attributes have been added to dbSNP to allow searching and filtering of human variation by the following characteristics:

Allele Origin: The rs report summarizes the reported origin(s) of the variant allele asserted by each submitter for the submitted SNP (ss). Current values are germline, somatic, and unknown. Additional attributes will be added in the future release to include:

- not-tested
- tested-inconclusive
- other

FYI

Clinical significance: The significance supplied by the submitter for the submitted SNP (ss).

The supported values are:

- unknown
- untested
- non-pathogenic
- probable-non-pathogenic
- probable-pathogenic
- pathogenic
- other

Global minor allele frequency (MAF): dbSNP is reporting the minor allele frequency for each rs included in a default global population. Since this is being provided to distinguish common polymorphism from rare variants, the MAF is actually the second most frequent allele value. In other words, if there are 3 alleles, with frequencies of 0.50, 0.49, and 0.01, the MAF will be reported as 0.49. The current default global population is 1000Genome phase 1 genotype data from 629 worldwide individuals.

Suspect: Variation suspected to be false positive due to artifacts of the presence of a paralogous sequence in the genome ([Musumeci et al. 2010](#)) ([Sudmant et al. 2010](#)) or evidence suggested sequencing error or computation artifacts.

Stephen Sherry, PhD

RefSNP Summary (Example)

RefSNP	Allele	HGVS Names
Organism: human (Homo sapiens)	<u>Variation Class:</u> SNP: single nucleotide polymorphism	NC_000008.9:g.19857809A>G
Molecule Type: Genomic	<u>RefSNP Alleles:</u> AVG	NG_008855.1:g.21948A>G
Created/Updated in build: 36/132	<u>Allele Origin:</u> G:Germline A:Germline	NM_000237.2:c.953A>G
Map to Genome Build: 37.1	<u>Ancestral Allele:</u> A	NP_000228.1:p.Asn318Ser
Validation Status:	<u>Clinical Source:</u>	NT_167187.1:g.7671675A>G
Citation: PubMed	<u>MAF/MinorAlleleCount:</u> G=0.003/3	
	<u>MAF Source:</u> 1000 Genomes	

A. Allele Origin Indicated as Germline or Somatic for each allele

B. Clinical Effect Click on "VarView" or "OMIM" to view phenotype

C. Global MAF The minor allele (G), frequency (0.003), and allele count (3) is shown

Stephen Sherry, PhD

SNP Gene View (Example)

Region	Chr. position	mRNA pos	dbSNP rs# cluster id	Heterozygosity	Validation	MAF	Allele origin	3D	Clinically Associated	Function	dbSNP allele	Protein residue	Codon pos	Amino acid pos	PubMed
	19805708	476	rs1801177	0.025			0.0175 A: Germline	Yes		missense	A	Asn [N]	1	36	
							A G: Germline	Yes		contig reference	G	Asp [D]	1	36	
	19805815	583	rs11542085	0.024			0.0172	Yes		missense	G	Gln [Q]	3	71	
										contig reference	C	His [H]	3	71	
	19809435	775	rs1121923	0.121			0.0863	Yes		synonymous	A	Val [V]	3	135	
										contig reference	G	Val [V]	3	135	

A. Allele Origin Indicated as Germline or Somatic for each allele
B. Clinical Effect Click on icon under "Clinical Source" to view effect in Variation Viewer
C. Global MAF MAF is shown for the corresponding allele
D. Suspected A red "?" icon is shown for suspected SNP under the "Validation" column

Stephen Sherry, PhD

Functional Terms = Sequence Ontology Standard

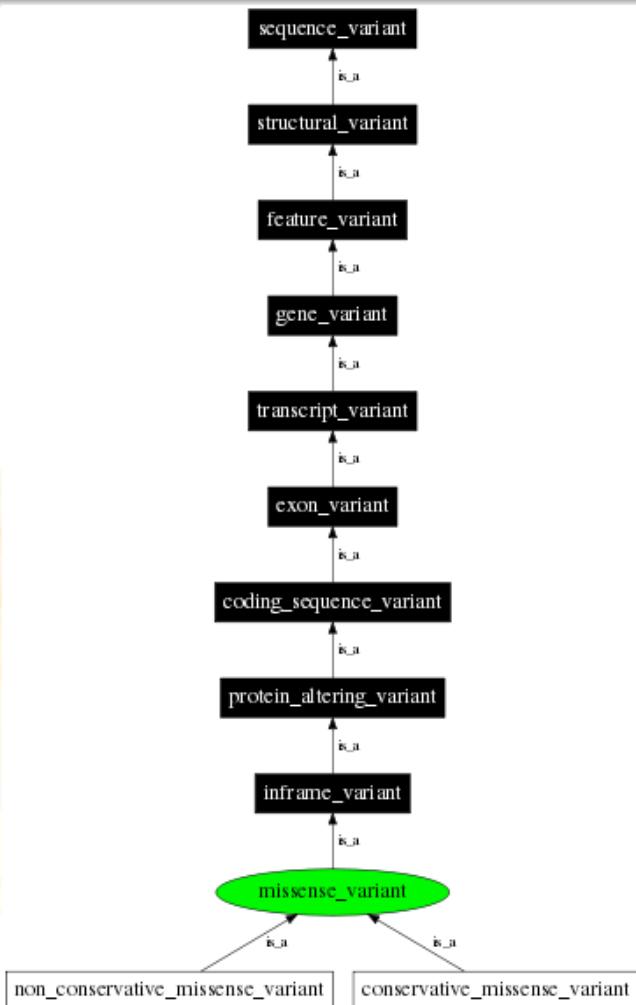
- **Function:** dbSNP's predicted functional effect of variant on RefSeq transcripts
- **Coding - Non-Synonymous** - change in peptide for allele with respect to the reference assembly (**dbSNP terms:** **nonsense**, **missense**, **stop-loss**, **frameshift**, **cds-indel**; Sequence Ontology terms: [stop_gained](http://sequenceontology.org/browser/current_release/term/SO:0001587), [missense_variant](http://sequenceontology.org/browser/current_release/term/SO:0001583), [stop_lost](http://sequenceontology.org/browser/current_release/term/SO:0001578), [frameshift_variant](http://sequenceontology.org/browser/current_release/term/SO:0001589), [inframe_indel](http://sequenceontology.org/browser/current_release/term/SO:0001820))
 - http://sequenceontology.org/browser/current_release/term/SO:0001587
 - http://sequenceontology.org/browser/current_release/term/SO:0001583
 - http://sequenceontology.org/browser/current_release/term/SO:0001578
 - http://sequenceontology.org/browser/current_release/term/SO:0001589
 - http://sequenceontology.org/browser/current_release/term/SO:0001820



MISO

The Sequence Ontology Browser

[Home](#) [Browser](#) [GFF3](#) [Resources](#) [About](#) [Request A Term](#) [Site Map](#)



missense_variant (CURRENT_RELEASE)

SO Accession:	SO:0001583 (SOWiki)
Definition:	A sequence variant, where the change may be longer than 3 bases, and at least one base of a codon is changed resulting in a codon that encodes for a different amino acid.
Synonyms:	missense, missense codon, missense_variant, non synonymous codon, non synonymous variant, non_synonymous_coding, SO:0001584, SO:0001783
DB Xrefs:	EBI: gr SO: ke
Parent:	inframe_variant (SO:0001650)
Children:	non_conservative_missense_variant (SO:0001586) conservative_missense_variant (SO:0001585)

[http://sequenceontology.org/
browser/current_release/term/SO:
0001583](http://sequenceontology.org/browser/current_release/term/SO:0001583)

Functional Terms = Sequence Ontology Standard

- **Function:** dbSNP's predicted functional effect of variant on RefSeq transcripts
- **Locus Region** - variation is 3' to and within 500 bases of a transcript, or is 5' to and within 2000 bases of a transcript (**dbSNP terms:** **near-gene-3, near-gene-5**; Sequence Ontology terms: [downstream_gene_variant](#), [upstream_gene_variant](#))
 - http://sequenceontology.org/browser/current_release/term/SO:0001632
 - http://sequenceontology.org/browser/current_release/term/SO:0001631

Functional Terms = Sequence Ontology Standard

- **Function:** dbSNP's predicted functional effect of variant on RefSeq transcripts
- **Coding - Synonymous** - no change in peptide for allele with respect to the reference assembly (**dbSNP term:** **coding-synon**; Sequence Ontology term: [synonymous variant](#))
 - http://sequenceontology.org/browser/current_release/term/SO:0001819

Functional Terms = Sequence Ontology Standard

- **Function:** dbSNP's predicted functional effect of variant on RefSeq transcripts
- **Untranslated** - variation is in a transcript, but not in a coding region interval
**(dbSNP terms: untranslated-3, untranslated-5; Sequence Ontology terms:
3 prime UTR variant, 5 prime UTR variant)**
 - http://sequenceontology.org/browser/current_release/term/SO:0001624
 - http://sequenceontology.org/browser/current_release/term/SO:0001623

Functional Terms = Sequence Ontology Standard

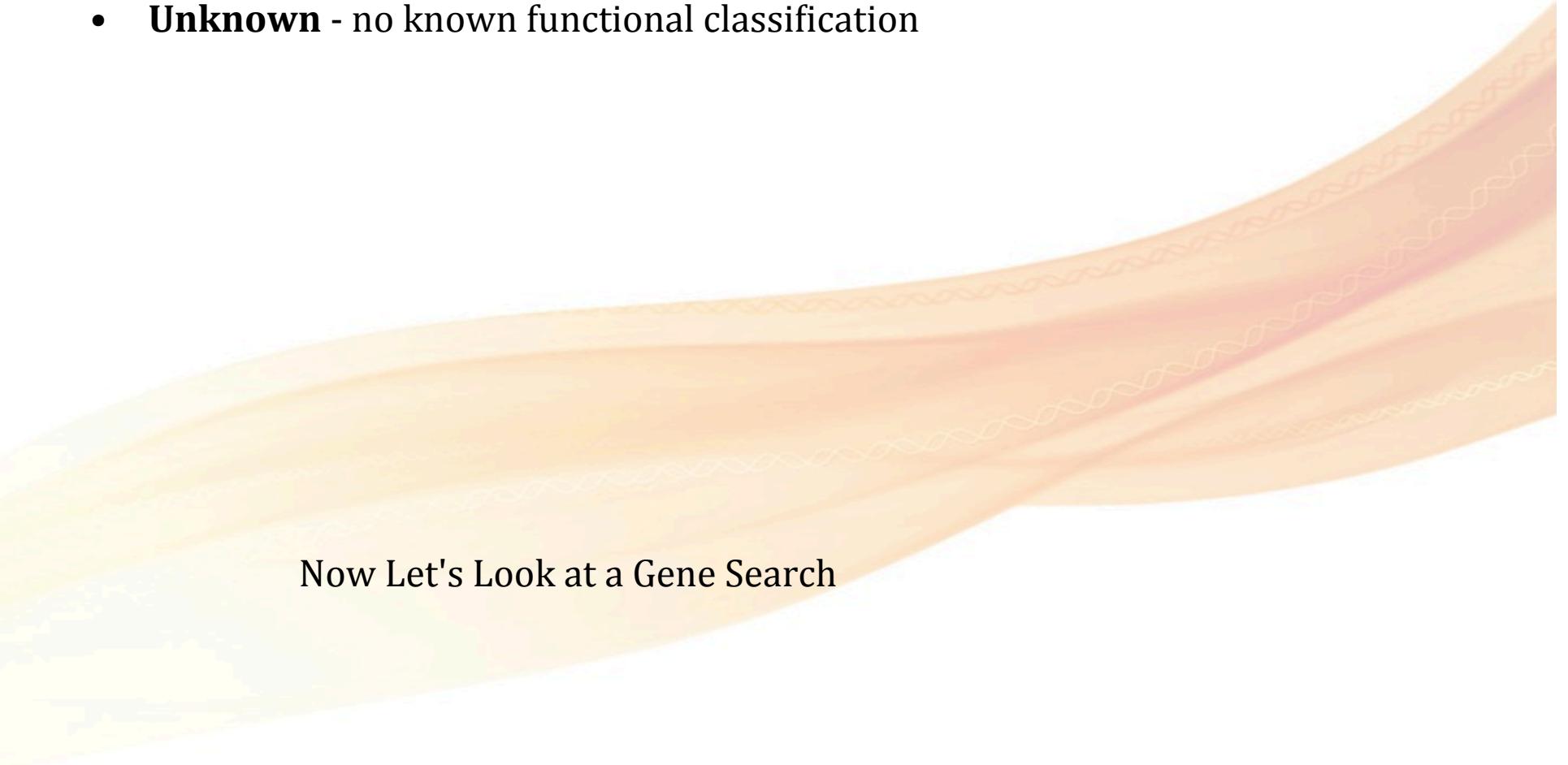
- **Function:** dbSNP's predicted functional effect of variant on RefSeq transcripts
- **Intron** - variation is in an intron, but not in the first two or last two bases of the intron (**dbSNP term: intron**; Sequence Ontology term: [intron_variant](#))
 - http://sequenceontology.org/browser/current_release/term/SO:0001627

Functional Terms = Sequence Ontology Standard

- **Function:** dbSNP's predicted functional effect of variant on RefSeq transcripts
- **Splice Site** - variation is in the first two or last two bases of an intron (**dbSNP terms:** **splice-3, splice-5**; Sequence Ontology terms: [splice acceptor variant](#), [splice donor variant](#))
 - http://sequenceontology.org/browser/current_release/term/SO:0001574
 - http://sequenceontology.org/browser/current_release/term/SO:0001575

Functional Terms = Sequence Ontology Standard

- **Function:** dbSNP's predicted functional effect of variant on RefSeq transcripts
- **Unknown** - no known functional classification



Now Let's Look at a Gene Search

Results: 1 to 20 of 26

<< First < Prev Page of 2 Next > Last >>

Nitric Oxide Synthase 2

- [nos2a – nitric oxide synthase 2a, inducible \[Danio rerio \(zebrafish\)\]](#)

1. nitric oxide synthase 2a, inducible

Official Symbol: **nos2a**

Other Aliases: inosa, nos2, si:dkkey-57m14.4

Other Designations: inducible nitric oxide synthase a; nitric oxide synthase 2, inducible, macrophage

Annotation: Chromosome 5, NC_007116.5 (45400854..45415665)

ID: 404036

- [Nos2a* – nitric oxide synthase 2, inducible, macrophage \[Rattus norvegicus \(Norway rat\)\]](#)

2. nitric oxide synthase 2, inducible, macrophage

Other Aliases: **Nos2a**

Chromosome: 10

 This record was replaced with [GenelID: 24599](#)

ID: 117133

- [Nos2a – nitric oxide synthase 2, inducible, macrophage \[Rattus norvegicus \(Norway rat\)\]](#)

3. nitric oxide synthase 2, inducible, macrophage

Location: 10

 This record was replaced with [GenelID: 24599](#)

ID: 116798

- [NOS2 – nitric oxide synthase 2, inducible \[Homo sapiens \(human\)\]](#)

4. nitric oxide synthase 2, inducible

Official Symbol: NOS2

Other Aliases: HEP-NOS, INOS, NOS, **NOS2A**

Other Designations: NOS type II; NOS, type II; hepatocyte NOS; inducible NO synthase; inducible NOS; nitric oxide synthase 2A (inducible, hepatocytes); nitric oxide synthase, inducible; nitric oxide synthase, macrophage; peptidyl-cysteine S-nitrosylase NOS2

Location: 17q11.2-q12

Annotation: Chromosome 17, NC_000017.10 (26083792..26127555, complement)

MIM: 163730

ID: 4843

[Order cDNA clone](#)

Display Settings: Full Report

NOS2 nitric oxide synthase 2, inducible [*Homo sapiens*]

Gene ID: 4843, updated on 7-Oct-2012

 Summary

Official Symbol	NOS2 provided by HGNC
Official Full Name	nitric oxide synthase 2, inducible provided by HGNC
Primary source	HGNC:7873
See related	Ensembl:ENSG0000007171; HPRD:01225; MIM:163730; Vega:OTTHUMG000001324
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	NOS; INOS; NOS2A; HEP-NOS
Summary	Nitric oxide is a reactive free radical which acts as a biologic mediator in several processes of neurotransmission and antimicrobial and antitumoral activities. This gene encodes a nitric oxide synthase which is expressed in liver and is inducible by a combination of lipopolysaccharides and certain cytokines. Three related pseudogenes are located within the Smith-Magenis syndrome region on chromosome 17. [provided by RefSeq, Jul 2008]

Genomic context

Location: 17q11.2-q12 **Sequence:** Chromosome: 17; NC_000017.10 (26083792..26127555, complement) See NOS2 in [Epigenomics](#), [MapViewer](#)

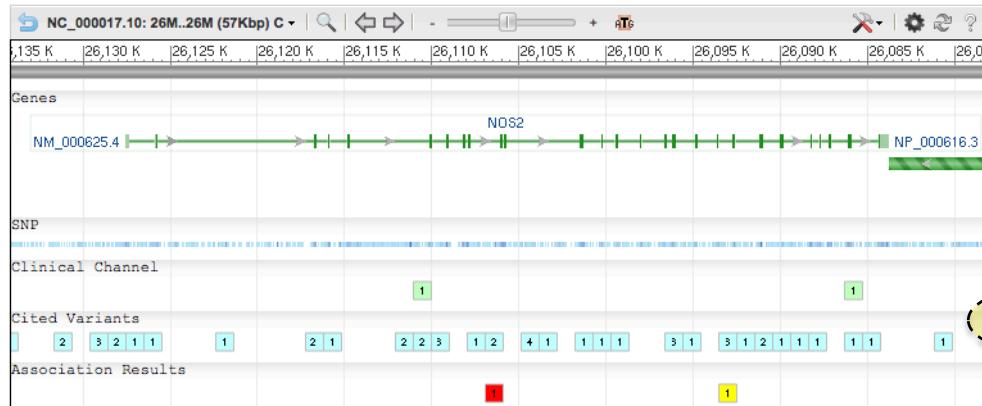


Genomic regions, transcripts, and products

Genomic Sequence NC_000017 chromosome 17 reference GRCh37.p9 Primary Assembly

[Go to reference sequence details](#)

[Go to nucleotide](#) [Graphics](#) [FASTA](#) [GenBank](#)



Send to:

Table of contents

- [Summary](#)
 - [Genomic context](#)
 - [Genomic regions, transcripts, and products](#)
 - [Bibliography](#)
 - [Phenotypes](#)
 - [HIV-1 protein interactions](#)
 - [Interactions](#)
 - [General gene info](#)
 - [General protein info](#)
 - [Reference sequences](#)
 - [Related sequences](#)
 - [Additional links](#)

Related information

- Order cDNA clone
 - 3D structures
 - BioAssay, by Gene target
 - BioAssay, by Protein Target
 - BioProjects
 - BioSystems
 - CCDS
 - Conserved Domains
 - dbVar
 - Full text in PMC
 - GAP
 - Genome
 - GEO Profiles
 - HomoloGene
 - Map Viewer
 - Nucleotide
 - OMIM
 - Probe
 - Protein
 - PubChem Compound
 - PubChem Substance
 - PubMed
 - PubMed (GeneRIF)
 - PubMed (OMIM)
 - RefSeq Proteins
 - RefSeq RNAs
 - RefSeqGene
 - SNP**
 - SNP: GeneView
 - SNP: Genotype
 - SNP: VariantView

Nitric Oxide Synthase 2

Nitric Oxide Synthase 2**

Display Settings: Graphic Summary, 20 per page, Sorted by Default order

Send to:

Results: 1 to 20 of 1249

<< First < Prev Page **1** of 63 Next > Last >>

rs202243635 [Homo sapiens]
1.

GGCCGAGAGATTTAAAGCAGGAATG[A/T]GGCTGAGTTCTCTGCCGCCGGAGCC

17 MapView No VarVu No PubMed GeneView SeqView No 3D No OMIM V G

HGVS Names: [NC_000017.10:g.26127503A>T] [NG_011470.1:g.5053T>A] [NM_000625.4:c.-212T>A] [NT_010799.15:g.864497A>T] [NW_001838430.2:g.4271874T>A]

ID: 202243635

rs202215159 [Homo sapiens]
2.

TACAGTGTGATTAAACTAAAATGCAA[C/T]TCATGTAAATATCTCCATCAAGGAA

17 MapView No VarVu No PubMed GeneView SeqView No 3D No OMIM V G

HGVS Names: [NC_000017.10:g.26083813T>A] [NC_000017.10:g.26083813T>C] [NC_000017.10:g.26083813T>G] [NG_005629.3:g.19111T>A] [NG_005629.3:g.19111T>C] [NG_005629.3:g.19111T>G] [NG_011470.1:g.48743A>C] [NG_011470.1:g.48743A>G] [NG_011470.1:g.48743A>T] [NM_000625.4:c.*459A>C] [NM_000625.4:c.*459A>G] [NM_000625.4:c.*459A>T] [NT_010799.15:g.820807T>A] [NT_010799.15:g.820807T>C] [NT_010799.15:g.820807T>G] [NW_001838430.2:g.4315562A>C] [NW_001838430.2:g.4315562A>G] [NW_001838430.2:g.4315562A>T]

ID: 202215159

**Results are now in different order

Nitric Oxide Synthase 2

Reference SNP(refSNP) Cluster Report: rs202243635

RefSNP	Allele	HGVS Names	Links
Organism: human (<i>Homo sapiens</i>)	SNV: <u>Variation Class:</u> single nucleotide variation	NC_000017.10:g.26127503A>T	
Molecule Type: Genomic	RefSNP Alleles: A/T	NG_011470.1:g.5053T>A	
Created/Updated in build: 137/137	Allele Origin:	NM_000625.4:c.-212T>A	
Map to Genome Build: 37.3	Ancestral Allele: Not available	NT_010799.15:g.864497A>T	
<u>Validation Status:</u>	Clinical Channel: unknown	NW_001838430.2:g.4271874T>A	
	Clinical Significance: NA		
	MAF/MinorAlleleCount: NA		
	MAF Source:		

SNP Details are organized in the following sections:

[GeneView](#) [Map](#) [Submission](#) [Fasta](#) [Resource](#) [Diversity](#) [Validation](#)

Integrated Maps (Hint: click on 'Chr Pos' or 'Contig Pos' column value to see variation in NCBI sequence viewer)

Assembly	Genome Build	Chr	Chr Pos	Contig	Contig Pos	SNP to Chr	Contig allele	Contig to Chr	Neighbor SNP	Map Method
GRCh37.p5	37.3	17	26127503	NT_010799.15	864497	Fwd	A	Fwd	view	blast
HuRef	37.3	17	22334447	NW_001838430.2	4271874	Fwd	T	Rev	view	blast

GeneView

GeneView via analysis of contig annotation: [NOS2](#) nitric oxide synthase 2, inducible

View more variation on this gene (click to hide).

Clinical Source: in gene region cSNP has frequency double hit

[Go](#)

Primary Assembly Mapping

Assembly	SNP to Chr	Chr	Chr position	Contig	Contig position	Allele
GRCh37.p5	Fwd	17	26127503	NT_010799.15	864497	A

RefSeqGene Mapping

RefSeqGene	Gene (ID)	SNP to RefSeqGene	Position	Allele
NG_011470.1	NOS2 (4843)	Rev	5053	T

Nitric Oxide Synthase 2

Gene Model (mRNA alignment) information from genome sequence ↑						
Total gene model (contig mRNA transcript): 6				Contig	Contig Label	List SNP
mRNA	transcript	protein	mRNA orientation			
NM_000625.4	plus strand	NP_000616.3	forward	NT_010799.15	GRCh37.p10	< currently shown
NM_000625.4	minus strand	NP_000616.3	reverse	NT_010799.15	GRCh37.p10	< currently shown
NM_000625.4	plus strand	NP_000616.3	forward	NW_001838430.2	HuRef	View SNP on GeneModel
NM_000625.4	minus strand	NP_000616.3	reverse	NW_001838430.2	HuRef	View SNP on GeneModel
NM_000625.4	plus strand	NP_000616.3	forward	NW_004078092.1	CHM1_1.0	View SNP on GeneModel
NM_000625.4	minus strand	NP_000616.3	reverse	NW_004078092.1	CHM1_1.0	View SNP on GeneModel

Clinical Source
 in gene region
 cSNP
 has frequency
 double hit

gene model Contig Label Contig mRNA protein mRNA orientation transcript SNP count
 (contig mRNA transcript): GRCh37.p10 NT_010799.15 NM_000625.4 NP_000616.3 forward plus strand 329, coding

Region	Chr. position	mRNA pos	dbSNP rs# cluster id	Heterozygosity	Validation	MAF	Allele origin	3D	Clinically Associated	Clinical Significance	Function	dbSNP allele	Protein residue	Codon pos	Amino acid pos	PubMed	Validation status description
	26084276	3722	rs201526410	N.D.							missense	C	Pro [P]	2	1153		Validated by multiple, independent submissions to the refSNP cluster
											contig reference	T	Leu [L]	2	1153		Validated by frequency or genotype data: minor alleles observed in at least two chromosomes.
	26084279	3719	rs199955383	N.D.							missense	T	Val [V]	2	1152		Validated by submitter confirmation
											contig reference	C	Ala [A]	2	1152		All alleles have been observed in at least two chromosomes apiece
	26084286	3712	rs199748579	0.001		0.0005					missense	G	Val [V]	1	1150		Genotyped by HapMap project
											contig reference	A	Met [M]	1	1150		SNP has been sequenced in 1000Genome project.
	26084307	3691	rs149327839	0.000							missense	A	Met [M]	1	1143		Suspect SNPs: SNP suspected from paralogous region (PMID: 21030649). Added to dbSNP on 01/21/2011.
											contig reference	G	Val [V]	1	1143		

Finding SNPs - Entrez SNP Summary

- dbSNP is useful for investigating detailed information on:
 - Small number SNPs
 - And it's good for a picture of the gene
- Entrez SNP is a direct, fast database for querying SNP data
 - Data can be:
 - retrieved in batches for many SNPs
 - “limited” to specific subsets of SNPs
 - And formatted in plain text for easy parsing and manipulation
- More detailed queries can be formed using specific “field tags”
 - For retrieving SNP data

Searching dbSNP

- dbSNP is searched like any other Entrez db
- Specialized fields include:

Field	Tag	Notes
Allele	[Allele]	Uses IUPAC codes for bases
Chromosomal Location	[chrpos]	Uses chromosomal base-pair locations
Contig Position	[ctpos]	Uses contig base-pair locations
Function Class	[Func]	Includes coding synonymous, missense, nonsense, intron, utr, etc.
SNP Class	[SNP_Class]	Includes.snp, indel, mixed

More Complex dbSNP Search

- Retrieve all:
 - synonymous coding reference SNPs
 - for the human norepinephrine transporter gene (*Slc6a2*) from dbSNP
- Utilize the **limits** like we did for lab with the Gene Database
- Search Strategy:

"Homo sapiens"[Organism] AND *Slc6a2*[gene] AND "synonymous codon"[Function_Class]

- Note: To use the (gene name) [gene] field, it is necessary to have the official gene name or gene symbol as per the [Human Gene Nomenclature Committee](#)
 - Entrez Gene can be used to find these
- More on search fields can be found here:
 - <http://www.ncbi.nlm.nih.gov/snp>

More Specific Approach

NCBI Resources How To

dbSNP

SNP

Limits

dbS Database insertion

Organism	Chromosomes	Chromosome Range	Map Weight
<input checked="" type="checkbox"/> Homo sapiens	<input checked="" type="checkbox"/> 1	From _____	<input type="checkbox"/> 1
<input type="checkbox"/> Agelaius phoeniceus	<input type="checkbox"/> 2	To _____	<input type="checkbox"/> 2
<input type="checkbox"/> Alectoris	<input type="checkbox"/> 2a		<input type="checkbox"/> 3-10
<input type="checkbox"/> Alectoris chukar	<input type="checkbox"/> 2b		<input type="checkbox"/> 10+
<input type="checkbox"/> Alectoris rufa	<input type="checkbox"/> 3		
<input type="checkbox"/> Allium cepa	<input type="checkbox"/> 4		
<input type="checkbox"/> Amaranthus caudatus	<input type="checkbox"/> 5		
<input type="checkbox"/> Anopheles funestus	<input type="checkbox"/> 6		
<input type="checkbox"/> Anopheles gambiae	<input type="checkbox"/> 7		
<input type="checkbox"/> Apis mellifera	<input type="checkbox"/> 8		
<input type="checkbox"/> Artemisia tridentata	<input type="checkbox"/> 9		

Function Class	SNP Class	Method Class	Validation Status
<input checked="" type="checkbox"/> coding nonsynonymous	<input type="checkbox"/> het	<input type="checkbox"/> computed	<input type="checkbox"/> by-HapMap
<input checked="" type="checkbox"/> nonsense	<input checked="" type="checkbox"/> in del	<input type="checkbox"/> dhpc	<input checked="" type="checkbox"/> by-1000 Genomes
<input checked="" type="checkbox"/> missense	<input type="checkbox"/> microsatellite	<input type="checkbox"/> hybridize	<input type="checkbox"/> by-cluster
<input checked="" type="checkbox"/> frame shift	<input type="checkbox"/> mixed	<input type="checkbox"/> other	<input type="checkbox"/> by-frequency
<input type="checkbox"/> stop gained	<input type="checkbox"/> mnp	<input type="checkbox"/> rfp	<input type="checkbox"/> by-submitter
<input type="checkbox"/> intron	<input type="checkbox"/> named	<input type="checkbox"/> sequence	<input type="checkbox"/> by-2hit-2allele
<input type="checkbox"/> coding synonymous	<input type="checkbox"/> no variation	<input type="checkbox"/> sscp	<input type="checkbox"/> no-info
<input type="checkbox"/> locus region	<input type="checkbox"/> SNP	<input type="checkbox"/> unknown	<input type="checkbox"/> Suspected
<input type="checkbox"/> mmra utr			<input type="checkbox"/> Paralogous or SND
<input type="checkbox"/> 5' utr			<input type="checkbox"/> other
<input type="checkbox"/> 3' utr			

Variation Allele	Annotation	Heterozygosity	Success Rate
Allele Origin	<input type="checkbox"/> Clinical/LSDB Submissions	<input type="checkbox"/> 0-10	<input type="checkbox"/> 80-85
<input type="checkbox"/> Germline	<input type="checkbox"/> Clinical significance	<input type="checkbox"/> 40-50	<input type="checkbox"/> 85-90
<input type="checkbox"/> Somatic	<input type="checkbox"/> pathogenic	<input type="checkbox"/> 10-20	<input type="checkbox"/> 90-95
IUPAC Variation Allele	<input type="checkbox"/> probable pathogenic	<input type="checkbox"/> 20-30	<input type="checkbox"/> 95+
<input type="checkbox"/> A	<input type="checkbox"/> non pathogenic	<input type="checkbox"/> 30-40	
<input type="checkbox"/> C	<input type="checkbox"/> probable non pathogenic		
<input type="checkbox"/> G	<input type="checkbox"/> unknown		
<input type="checkbox"/> T	<input type="checkbox"/> other		
	<input type="checkbox"/> nucleotide		
	<input type="checkbox"/> OMIM		
	<input type="checkbox"/> protein		

Created Build ID	Updated Build ID	Individual SNP	Minor Allele Freq
<input type="checkbox"/> Current Build ID(137)	<input type="checkbox"/> Current Build ID(137)	<input type="checkbox"/> Venter	Minor Allele Frequency:
<input type="checkbox"/> Last Build ID(136)	<input type="checkbox"/> Last Build ID(136)	<input type="checkbox"/> Watson	>= <input type="checkbox"/> <= <input type="checkbox"/>

Global MAF	Genome Project
Global Minor Allele Frequency:	<input type="checkbox"/> 1000 Genomes
Min <input type="checkbox"/> Max	<input type="checkbox"/> Pilot 1
	<input type="checkbox"/> Pilot 2
	<input type="checkbox"/> Pilot 3
	<input type="checkbox"/> Has Frequency

HapMap Populations
<input type="checkbox"/> ASW
<input type="checkbox"/> CEU
<input type="checkbox"/> CHB
<input type="checkbox"/> CHD
<input type="checkbox"/> GIH

More Specific Approach

dbSNP

SNP

Save search Limits Advanced

Display Settings: Graphic Summary, 20 per page, Sorted by Default order

Format Items per page Sort by

Graphic 5 Default order
 Summary 10 Organism
 FASTA 20 SNP_ID
 FlatFile 50 Success Rate
 Chromosome 100 Heterozygosity
 Report 200 Chromosome Base
Position
 Clinical

Send to:

Change | Remove

Page 1 of 118 Next > Last >>

Apply

NM_052843.2:c.15155G>A] [

NP_001092093.1:p.Arg5052His] [NP_443075.2:p.Arg5052His] [NT_167186.1:g.22027476G>A] [NW_001838544.1:g.528517G>A]

ID: 193302438

rs193291380 [Homo sapiens]

2.

TACGTGTGTATTTCCCCCCCCCAG [A/G] GACTAC

R

>gnl|dbSNP|rs193302438 rs=193302438|pos=201|len=401|taxid=9606|mol="genomic"|class=snp|alleles="A/G"|build=135
TGTGGTGTCTG ACTGGGGTCAG AGAGTGAAAG CGAGAGCTCC TCTGGGGGTG AGCTGGACGA TGCCTTCCGC CGGGCTGCC
GTCGGCTGCA CGGGCTCTC CGCACAAAAA GTCCGGCTGA AGTTTCAGAT GAGGAGCTCT TCCTGAGTGC AGACGAGGGC
CCTGCAGAGC CAGAGGAGCC CGCGGACTGG CAGACATACC

R

CGAAGAGATGG CATTTCATCT GCATCCGTT TGAGGGCTCT ACTGAGGCC GCCAGGGGT AACTCGCTTC CAGGAGATGT
TTGCCACACT GGGCATGGG GTGGAGATCA AGCTGGTGA ACAGGGGCT CGGAGGGTAG AGATGTGCAT CAGCAAAGAG
ACTCCCTGCC CTGTGGTGC TCCAGAGCCA TTGCCCAGCC

R

>gnl|dbSNP|rs193291380 rs=193291380|pos=201|len=401|taxid=9606|mol="genomic"|class=snp|alleles="A/G"|build=135
TACACTGGT TTGTAAATCT TATGCCATCC TCTTCCCTTA GAGTACAGA ATGCATTCGA GAATCTTTGT GTTTAAACA
ATGCTTACT GGACAGTTTC TGCCAGAACAA AGTTCACTT ACTCTCTTC GTTGGAGTCA ATTAAGAAC TCTGCCATG
GCACATTTG TAAATACGTG TGTATTTTC CCCCCCCCAG

R

GACTACTGGC ATTCACTGAT GTGGTCATAG AATTCTCTCC AGAGGGATGG CCATGCCTGG ACCCTGCCA CGCAAATTTG
TATAGGGATG TGATGTTGGA GAACTACAGA AACCTGGTCT CCCTGGGTGA GGATAACTTC AATACACAAT TTATGTATTT
CATACTACGG ATTCCATATT TTTCACTCCT GAAATGTTAC

M

>gnl|dbSNP|rs193290258 rs=193290258|pos=201|len=401|taxid=9606|mol="genomic"|class=snp|alleles="A/C"|build=135
TGCATGGCAT AAAGAGACAG AGATGCTCA CTTCCAGGAT TCTTGAATG TACTTAGAGT CTCTCTCTA GTGGGTCTGA
CCACCTCTT ACTCTGGGCT AAGAGCTGG CAGCATTAG CAGCAGCAGG AACCCCAAGGC AGACAATGGC TGGGATACAG
GTGCTTAGTG ACTCTTCAG AGCTGCCAAG GTGAAGATAG

M

CCCTGGAATG GAAAGTGAGG TCCCTGAGGC TCCTTGGAAAG AACCCAGGAG ATCCAGGAGC TCAGGGCTCT CTAGTTAGG
CTGACCCACT CCAGCCCAAG CCCCCTCATT AAACACCCCC ACCACTCTTA ACCCACATTC CCAAAACTAG CGGCATTCA
CTCTAGCCTC CTGGAGAAAG CAGGACCAGG AAATTAGAAG

FASTA Format and Data Structure for a Submitted SNP (ss) Record

```

define for FASTA records start with ">"
object-type=general
| database name           total length
|                         of sequence
| offset of SNP | Submitter
| unique id   ss# in sequence | SubmitterSNPID
|                           | organism
|                           | molecule
|                           | class of
|                           | alleles
|                           | list of
|                           variation
define: >gnl|dbSNP|ss271 ss=271|pos=51|len=101|handle="DEBNICK" |subid="lp03022" |taxid=9606|mol="Genomic" |class=1|alleles="G/A"
5' sequence: CTGCATCACATGACTGATTCTGTCCATTGGAACAGAGATGATGACTGGT
variation: R
3' sequence: TTACTAAACCTGAGCCCTGGTGTTCATTGTTGATAGGGGGTTGCATTGAT

```

- Notes for ss FASTA records:
 - If variation is a SNP then **appropriate IUPAC nucleotide ambiguity letter** is selected to **represent the reported possible allele states**
 - microsatellite, insertion/deletion:
 - variation is represented as a single "N" on the variation line of the FASTA report
 - If the string of alleles is more than 30 characters, the list of alleles is replaced by the tag "lengthTooLong"

FASTA Format and Data Structure for a Reference SNP (rs) Record

- rs FASTA records do not have the submitter/local SNP ID on the definition line (defline) - since they are clustered data objects constructed from one or more ss records
 - SNP classes defined in: ftp://ftp.ncbi.nih.gov/snp/specs/docsum_3.0.asn
 - Remember the classes?

<ftp://ftp.ncbi.nih.gov/snp/00readme.txt>

ftp://ftp.ncbi.nih.gov/snp/database/README.create_local_dbSNP.txt

<http://www.ncbi.nlm.nih.gov/books/NBK44378/>

NCBI Supports the Public Distribution of dbSNP

- Provide zip-compressed data dumps in four different data formats
- Given this current dynamic environment
 - Periodically release builds and build updates to various organisms
 - Announcement of the build release or update on the dbSNP home site
- Access to the NCBI FTP site is available via the web or anonymous FTP
- The URL/host addresses are:
- World Wide Web: <ftp://ftp.ncbi.nih.gov/snp/>
- Anonymous FTP:
 - `ftp ftp.ncbi.nih.gov`

FTP

```
MacBook-Pro-SSD:~ cleslin$ ftp ftp.ncbi.nih.gov
```

```
Connected to ftp.wip.ncbi.nlm.nih.gov.
```

```
220-
```

```
Warning Notice!
```

You are accessing a U.S. Government information system which includes this computer, network, and all attached devices. This system is for Government-authorized use only. Unauthorized use of this system may result in disciplinary action and civil and criminal penalties. System users have no expectation of privacy regarding any communications or data processed by this system. At any time, the government may monitor, record, or seize any communication or data transiting or stored on this information system.

```
---
```

```
Welcome to the NCBI ftp server! The anonymous access URL is ftp://ftp.ncbi.nlm.nih.gov/
```

Public data may be downloaded by logging in as "anonymous" using your E-mail address as a password.

```
Please see ftp://ftp.ncbi.nlm.nih.gov/README.ftp for hints on large file transfers
```

```
220 FTP Server ready.
```

```
Name (ftp.ncbi.nih.gov:cleslin): anonymous
```

```
331 Anonymous login ok, send your complete email address as your password
```

```
Password:
```

```
230 Anonymous access granted, restrictions apply
```

```
Remote system type is UNIX.
```

```
Using binary mode to transfer files.
```

```
ftp> cd.snp
```

```
250 CWD command successful
```

```
ftp>
```

FTP

```
ftp> dir
229 Entering Extended Passive Mode (|||50308|)
150 Opening ASCII mode data connection for file list
-r--r--r--  1 ftp      anonymous    57152 Mar 23  2012 00readme.txt
dr-xr-xr-x  3 ftp      anonymous    4096 Mar 10  2010 Entrez
-r--r--r--  1 ftp      anonymous     157 Feb  7  2012 FTP_Getting_Started.txt
dr-xr-xr-x  2 ftp      anonymous    4096 Apr 21  2005 HapExampleDir
dr-xr-xr-x  5 ftp      anonymous   53248 Oct 20 16:34 batch
dr-xr-xr-x  9 ftp      anonymous    4096 Mar  5  2010 database
dr-xr-xr-x  2 ftp      anonymous    4096 Aug 29  2003 hs-data-freeze-index
dr-xr-xr-x 11 ftp      anonymous    4096 Feb  8  2012 last_minute_correction
dr-xr-xr-x 113 ftp     anonymous   4096 Feb  8  2012 organisms
dr-xr-xr-x  2 ftp      anonymous    4096 May 10  2010 osiris
dr-xr-xr-x  5 ftp      anonymous    4096 Nov  8  2010 release-notes
dr-xr-xr-x  3 ftp      anonymous    4096 Oct  1  19:33 specs
dr-xr-xr-x  2 ftp      anonymous    4096 Apr 28  2008 status
dr-xr-xr-x 22 ftp     anonymous    4096 Oct 15 16:35 temp
-r--r--r--  1 ftp      anonymous 3170070 May 29 16:14 temp.fromClinVar.30051
dr-xr-xr-x  3 ftp      anonymous    4096 May  1  19:04 test
dr-xr-xr-x  3 ftp      anonymous    4096 Apr 20  2005 webdata
226 Transfer complete
ftp> cd organisms/human_9606/rs_fasta
```

FTP

```
ftp> dir
229 Entering Extended Passive Mode (|||50295|)
150 Opening ASCII mode data connection for file list
-r--r--r-- 1 ftp      anonymous 850305870 Jun  8 22:40 rs_ch1.fas.gz
-r--r--r-- 1 ftp      anonymous 507372005 Jun  8 22:58 rs_ch10.fas.gz
-r--r--r-- 1 ftp      anonymous 515659936 Jun  8 23:11 rs_ch11.fas.gz
-r--r--r-- 1 ftp      anonymous 485748640 Jun  8 23:24 rs_ch12.fas.gz
-r--r--r-- 1 ftp      anonymous 357265542 Jun  8 23:36 rs_ch13.fas.gz
-r--r--r-- 1 ftp      anonymous 322122121 Jun  8 23:46 rs_ch14.fas.gz
-r--r--r-- 1 ftp      anonymous 301860380 Jun  8 23:54 rs_ch15.fas.gz
-r--r--r-- 1 ftp      anonymous 331279473 Jun  9 00:03 rs_ch16.fas.gz
-r--r--r-- 1 ftp      anonymous 289118287 Jun  9 00:11 rs_ch17.fas.gz
-r--r--r-- 1 ftp      anonymous 279678636 Jun  9 00:20 rs_ch18.fas.gz
..
..
-r--r--r-- 1 ftp      anonymous 556362264 Jun  9 03:39 rs_ch8.fas.gz
-r--r--r-- 1 ftp      anonymous 462457170 Jun  9 04:04 rs_ch9.fas.gz
-r--r--r-- 1 ftp      anonymous       72 Jun  9 04:10 rs_chAltOnly.fas.gz
-r--r--r-- 1 ftp      anonymous 161121 Jun  9 04:10 rs_chMT.fas.gz
-r--r--r-- 1 ftp      anonymous 144427703 Jun  9 04:13 rs_chMulti.fas.gz
-r--r--r-- 1 ftp      anonymous 43975102 Jun  9 04:15 rs_chNotOn.fas.gz
-r--r--r-- 1 ftp      anonymous 16415699 Jun  9 04:16 rs_chPAR.fas.gz
-r--r--r-- 1 ftp      anonymous 53189087 Jun  9 04:17 rs_chUn.fas.gz
-r--r--r-- 1 ftp      anonymous 413074086 Jun  9 04:23 rs_chX.fas.gz
-r--r--r-- 1 ftp      anonymous 22522600 Jun  9 04:29 rs_chY.fas.gz
226 Transfer complete
ftp> prompt ←
Interactive mode off.
ftp> mget rs_ch*
```

Turn off the prompting

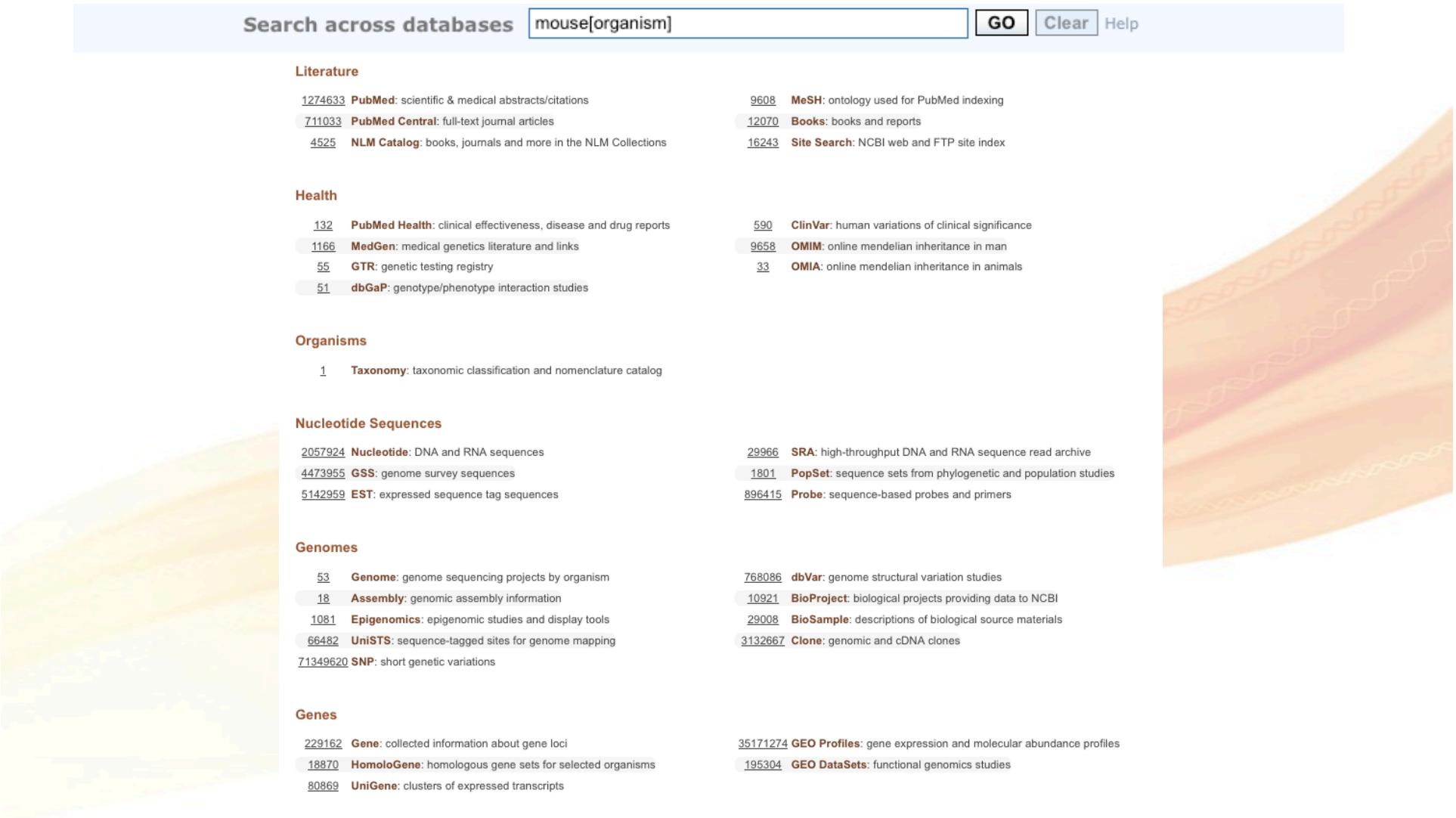
Things to Do for Thursday

- Read
 - Featured Resource: An Expanded Set of Discovery Components in the Entrez System
 - <http://www.ncbi.nlm.nih.gov/books/NBK7039/>
 - Sequence Identifiers: A Historical Note
 - <http://www.ncbi.nlm.nih.gov/Sitemap/sequenceIDs.html>
 - Slide 55 "Things to Read"
- Try FTP
- Do the "Go Through These Slides After Lecture or At Home" – next slide
- Get started on the lab - posted later tonight



Go Through These Slides After Lecture or
At Home

Perform an Organism Search – Across Databases



Search across databases Help

Literature

[1274633 PubMed](#): scientific & medical abstracts/citations
[711033 PubMed Central](#): full-text journal articles
[4525 NLM Catalog](#): books, journals and more in the NLM Collections

[9608 MeSH](#): ontology used for PubMed indexing
[12070 Books](#): books and reports
[16243 Site Search](#): NCBI web and FTP site index

Health

[132 PubMed Health](#): clinical effectiveness, disease and drug reports
[1166 MedGen](#): medical genetics literature and links
[55 GTR](#): genetic testing registry
[51 dbGaP](#): genotype/phenotype interaction studies

[590 ClinVar](#): human variations of clinical significance
[9658 OMIM](#): online mendelian inheritance in man
[33 OMIA](#): online mendelian inheritance in animals

Organisms

[1 Taxonomy](#): taxonomic classification and nomenclature catalog

Nucleotide Sequences

[2057924 Nucleotide](#): DNA and RNA sequences
[4473955 GSS](#): genome survey sequences
[5142959 EST](#): expressed sequence tag sequences

[29966 SRA](#): high-throughput DNA and RNA sequence read archive
[1801 PopSet](#): sequence sets from phylogenetic and population studies
[896415 Probe](#): sequence-based probes and primers

Genomes

[53 Genome](#): genome sequencing projects by organism
[18 Assembly](#): genomic assembly information
[1081 Epigenomics](#): epigenomic studies and display tools
[66482 UniSTS](#): sequence-tagged sites for genome mapping
[71349620 SNP](#): short genetic variations

[768086 dbVar](#): genome structural variation studies
[10921 BioProject](#): biological projects providing data to NCBI
[29008 BioSample](#): descriptions of biological source materials
[3132667 Clone](#): genomic and cDNA clones

Genes

[229162 Gene](#): collected information about gene loci
[18870 HomoloGene](#): homologous gene sets for selected organisms
[80869 UniGene](#): clusters of expressed transcripts

[35171274 GEO Profiles](#): gene expression and molecular abundance profiles
[195304 GEO DataSets](#): functional genomics studies

2013 - Oct

Apolipoprotein E (APOE)

- Important serum lipid transport protein
- Defects implicated in cardiovascular disease and late-onset Alzheimer disease (LOAD).
- Three common isoforms (alleles)

Isoform	Position 112 (130)	Position 158 (176)
e3	Cysteine (C)	Arginine (R)
e4	Arginine (R)	Arginine (R)
e2	Cysteine (C)	Cysteine (C)

The e4 isoform (allele) is associated with increased risk of LOAD

Human APOE Data Goals

- Reference Sequences
 - transcript (mRNA)
 - protein
 - gene (gDNA)
- Sites of Expression
 - UniGene Expression profile
 - GEO profiles
- View genomic assemblies and Maps
 - Comparative Maps
- Disease gene polymorphisms
- Genotypes
 - Reference Genome
 - HuRef (JC Venter genome)
- Homologs in other species
 - HomoloGene
 - BLAST to find Panda homologs

Human APOE Gene and RefSeqs

Gene Database

1. Search All Databases with APOE from NCBI Homepage
2. Retrieve PubMed results
3. Follow Gene Ad to Gene ID 348
4. Follow link to “reference sequence details”

Gene Database

1. Look for the Transcript and protein
 1. NM_000041.2
 2. NP_000032.1
2. Look at the Genomic
 1. RefSeq Gene NG_007084.2
3. Look for the Genome Builds
 1. Reference NC_000019.9
 2. HuRef

Keep this Gene page open for the rest of the searching, right click and "Open Link In New Window" for the rest of the slides

Genome Maps

APO Gene Cluster

1. Gene "Links" menu (on the right) to Map Viewer
2. Maps and Options
 1. Remove all but Gene
3. Zoom out 10X using graphic on right
4. Zoom in 2X by clicking on map
5. What other genes are part of the APO gene cluster?

Comparative Maps

1. Map Viewer "Maps and Options"
 1. Add Chimp and Mouse Gene Maps

Allelic Variants and Disease

OMIM

1. Gene links menu to OMIM
2. Retrieve
APOLIPOPROTEIN E; APOE
3. Click allelic variants
4. Jump to .0016 ALZHEIMER DISEASE 2
[APOE, CYS112ARG]

dbSNP

1. Gene links menu to SNP
2. Click the "Gene View" on once
of the graphic overives
3. Check "Clinical Source" and
click "Refresh"
4. Look for position 130 Cys->Arg
5. Click on rs429358 in table
6. Note HuRef has C allele

Expression Information

UniGene

1. Gene links menu to UniGene
2. Look at the EST profile
3. What is the expression pattern, widespread or confined to a certain organ?

GEO Profiles

1. Gene links menu to GEO profiles
2. Use History to combine link from Gene AND GDS596
3. Predominate expression in liver, some in brain

Homologous Genes / Proteins

HomoloGene

1. Gene links menu to HomoloGene
2. Use “pairwise alignments” device to compare Human and Chimp proteins at position 130
3. Use UniGene portion to find rabbit (*Oryctolagus cuniculus*) homolog
4. What's the Identity?

Using BLAST

1. Gene link to RefSeq protein NP_000032
2. “Run BLAST” from Discovery Column
3. Choose Reference Proteins database
4. Organism limit Giant Panda
5. BLAST button
6. Any homologs found in Panda?

Finding a Structure

Related Structures

1. Gene link to [NP_000032](#)
2. Links menu to Related Structures (Summary)
3. Change to “All similar MMDB” and click “Refresh Display”
4. Link to sequence alignment from 1B68_A
5. Do you see the polymorphism at position 112? Is so what is it?

Structure and Cn3D

1. Click the 1B68 from the "Related Structures" page you are at
2. Click the MMDB ID: 16994
3. Related structures link to structure for 1B68
4. Click Structure View in Cn3D (Cn3D must be installed)
5. Manipulate structure
6. Add side chains using Style -> Edit Global Style
7. Highlight Arg 112