

ГЕНОМИКА

18 марта 2020, Среда:

- Sanger DNA sequencing
- Next-generation sequencing (NGS) technology
- Major applications of NGS including whole genome *de novo* sequencing, resequencing, target and community (metagenomics) sequencing

23 марта 2020, Понедельник:

- key NCBI resources to find information about genes and genomes
- NCBI sequence database search
- Genbank format: features and sequence annotation
- BLAST search
- Pairwise and Multiple sequence alignment (BioEdit)



NCBI resources, search and retrieval functions

The screenshot shows the NCBI website interface. At the top, there is a search bar with a dropdown menu for 'All Databases' and a 'Search' button. Below the search bar, there is a navigation menu with various categories like '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', and 'Variation'. A 'Sign in to NCBI' link is visible in the top right corner. The main content area features a 'Popular Resources' section with links to PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. Below this is an 'NCBI Announcements' section with several news items. A 'NCBI YouTube channel' banner is also present. At the bottom, there is a footer with navigation links and a 'Write to the Help Desk' link.

- Narrow your search using particular fields: [organism], [author], [subtree], [lineage], etc.

and

- logical Boolean operators: AND, OR, XOR, and NOT

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



NCBI resources, search and retrieval functions

www.ncbi.nlm.nih.gov/Sitemap/samplerecord

Most Visited Getting Started Scholar

NCBI Sample GenBank Record

PubMed Entrez

GenBank Flat File Format

Click on any link in this sample record to see a detailed description of that data element or field
[Resource Guide](#)

LOCUS SCU49845 5028 bp DNA PLN 21-JUN-1999
DEFINITION Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION U49845
VERSION U49845.1 GI:1293613
KEYWORDS .
SOURCE Saccharomyces cerevisiae (baker's yeast)
ORGANISM Saccharomyces cerevisiae
Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE 1 (bases 1 to 5028)
AUTHORS Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
TITLE Cloning and sequence of REV7, a gene whose function is required for DNA damage-induced mutagenesis in Saccharomyces cerevisiae
JOURNAL Yeast 10 (11), 1503-1509 (1994)
PUBMED 7871890
REFERENCE 2 (bases 1 to 5028)
AUTHORS Roemer,T., Madden,K., Chang,J. and Snyder,M.
TITLE Selection of axial growth sites in yeast requires Axl2p, a novel plasma membrane glycoprotein
JOURNAL Genes Dev. 10 (7), 777-793 (1996)
PUBMED 8846915
REFERENCE 3 (bases 1 to 5028)
AUTHORS Roemer,T.
TITLE Direct Submission
JOURNAL Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New Haven, CT, USA
FEATURES
source Location/Qualifiers
1..5028
/organism="Saccharomyces cerevisiae"
/db_xref="taxon:4932"
/chromosome="IX"
/map="9"
CDS
1..206
/codon_start=3
/product="TCP1-beta"
/protein_id="AAA98665.1"
/db_xref="GI:1293614"
/translation="SSIYNGISTSGLDLNGTIADMRQLGIVESYKLRVAVSSASEA
AEVLLRVNDIIRARPRANRQHM"
gene
687..3158
/gene="AXL2"
CDS
687..3158
/gene="AXL2"
/note="plasma membrane glycoprotein"
/codon_start=1
/function="required for axial budding pattern of S.
cerevisiae"
/product="Axl2p"
/protein_id="AAA98666.1"
/db_xref="GI:1293615"
/translation="MTLQISLLLTATISLLHLVVATPYEAYPIGKYPPVARVNESF
TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTPSGEPSSDLLSDANTLLYFN
VILEGTDSDSTSLNNTYQFVVTNRPSISLSSDFNLLALLKRYGYINGRNALKLDFNE
VENVTFDRSMFTNEESIVSYGRSQLYNAPLPNWLFFDSELKFTGTAPVINSIAIAP
TSYSFVLIATDIEGFSAVEVFEEI.VIGAHOLITSTONSLLINVDITGNVSYDLPINLV

- Narrow your search using particular fields: [definition], [organism], [author], [subtree], [lineage], etc. (otherwise, it will search all fields)

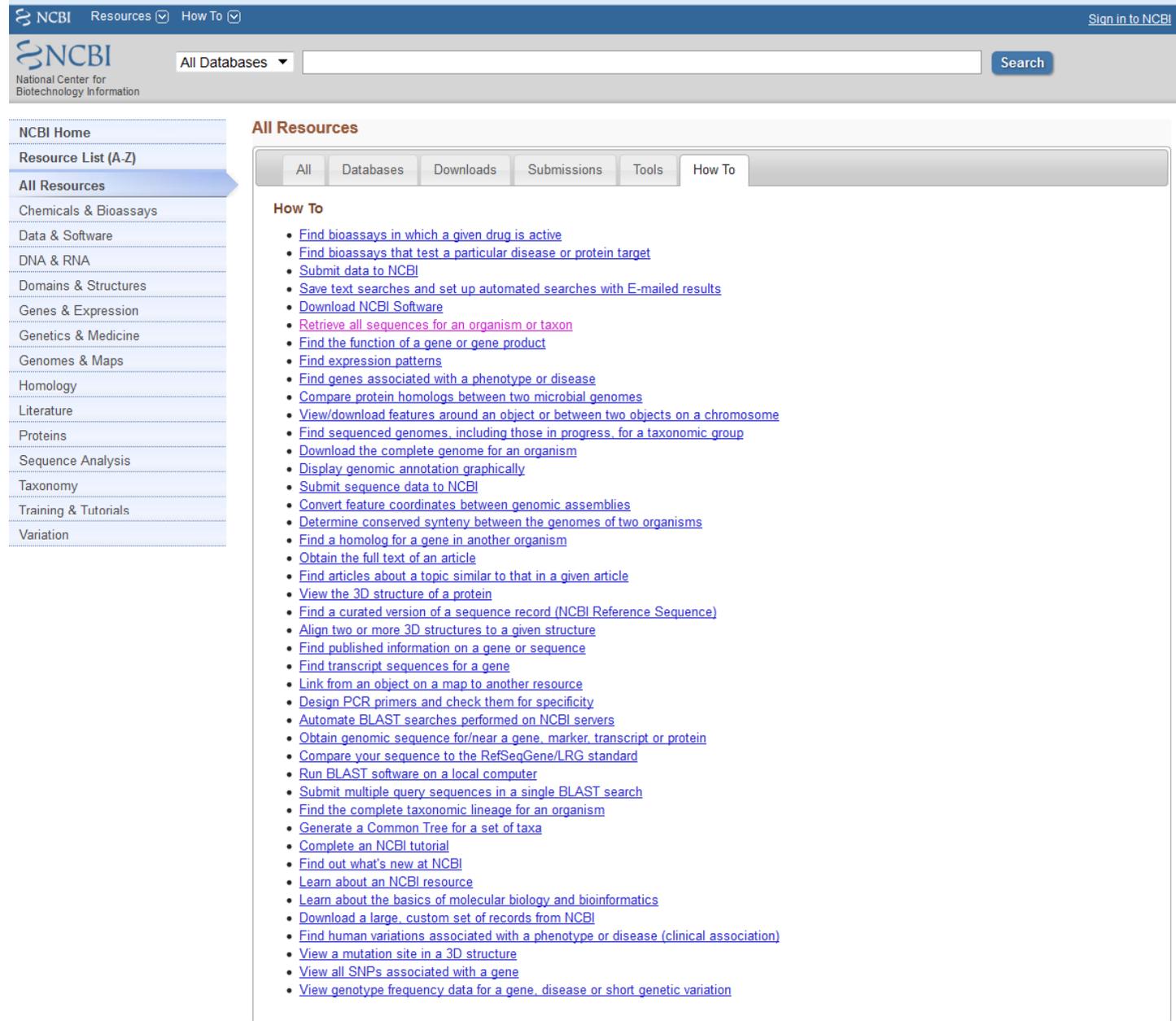
and

- logical Boolean operators: AND, OR, XOR, and NOT

<https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html>



NCBI resources, search and retrieval functions



NCBI Resources How To Sign in to NCBI

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 & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation

All Resources

All Databases Downloads Submissions Tools How To

How To

- [Find bioassays in which a given drug is active](#)
- [Find bioassays that test a particular disease or protein target](#)
- [Submit data to NCBI](#)
- [Save text searches and set up automated searches with E-mailed results](#)
- [Download NCBI Software](#)
- [Retrieve all sequences for an organism or taxon](#)
- [Find the function of a gene or gene product](#)
- [Find expression patterns](#)
- [Find genes associated with a phenotype or disease](#)
- [Compare protein homologs between two microbial genomes](#)
- [View/download features around an object or between two objects on a chromosome](#)
- [Find sequenced genomes, including those in progress, for a taxonomic group](#)
- [Download the complete genome for an organism](#)
- [Display genomic annotation graphically](#)
- [Submit sequence data to NCBI](#)
- [Convert feature coordinates between genomic assemblies](#)
- [Determine conserved synteny between the genomes of two organisms](#)
- [Find a homolog for a gene in another organism](#)
- [Obtain the full text of an article](#)
- [Find articles about a topic similar to that in a given article](#)
- [View the 3D structure of a protein](#)
- [Find a curated version of a sequence record \(NCBI Reference Sequence\)](#)
- [Align two or more 3D structures to a given structure](#)
- [Find published information on a gene or sequence](#)
- [Find transcript sequences for a gene](#)
- [Link from an object on a map to another resource](#)
- [Design PCR primers and check them for specificity](#)
- [Automate BLAST searches performed on NCBI servers](#)
- [Obtain genomic sequence for/near a gene, marker, transcript or protein](#)
- [Compare your sequence to the RefSeqGene/LRG standard](#)
- [Run BLAST software on a local computer](#)
- [Submit multiple query sequences in a single BLAST search](#)
- [Find the complete taxonomic lineage for an organism](#)
- [Generate a Common Tree for a set of taxa](#)
- [Complete an NCBI tutorial](#)
- [Find out what's new at NCBI](#)
- [Learn about an NCBI resource](#)
- [Learn about the basics of molecular biology and bioinformatics](#)
- [Download a large, custom set of records from NCBI](#)
- [Find human variations associated with a phenotype or disease \(clinical association\)](#)
- [View a mutation site in a 3D structure](#)
- [View all SNPs associated with a gene](#)
- [View genotype frequency data for a gene, disease or short genetic variation](#)



NCBI resources, search and retrieval functions

NCBI Resources How To Sign in to NCBI

NCBI National Center for Biotechnology Information

Taxonomy Pinus taeda Search

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

How to: Retrieve all sequences for an organism or taxon

Please note that there is a [YouTube tutorial](#) about this.

Starting with an organism or taxon name...

1. Search the [Taxonomy](#) database with the organism name. Accepted common names usually work at all taxonomic levels. Use the scientific name or formal name if no results are obtained with the common name.
2. Click on the desired taxon name in the results. For terminal taxa - generally subspecies, species, or strains - this link leads directly to the summary page. For higher taxa this link will lead to the Taxonomy Browser showing the lower taxa contained within the higher taxon.
3. If necessary, click on the desired taxon link in the Taxonomy Browser to reach the summary page.
4. The number of records in each database are linked in the Entrez records table on the taxon summary page. Click the linked number of records in the table to retrieve all records from the chosen sequence database (Nucleotide, Nucleotide EST, Nucleotide GSS, Protein).

NCBI Resources How To

Taxonomy Taxonomy Pinus taeda Save search Limits Advanced

Display Settings: Summary

Pinus taeda
(loblolly pine), species, conifers
[Nucleotide](#) [Protein](#)

NCBI Taxonomy Browser

Search for:

Display: 3 levels using filter: none

Pinus taeda

Taxonomy ID: 3352
Genbank common name: **loblolly pine**
Inherited blast name: **conifers**
Rank: species
Genetic code: [Translation table 1 \(Standard\)](#)
Mitochondrial genetic code: [Translation table 1 \(Standard\)](#)
Other names:
authority: **Pinus taeda L.**

[Lineage \(full\)](#)
cellular organisms; Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Coniferophyta; Coniferopsida; Coniferales; Pinaceae; Pinus; Pinus

Entrez records	
Database name	Direct links
Nucleotide	108,590
Nucleotide EST	328,662
Nucleotide GSS	5,393
Protein	30,753
Structure	1
Popset	6,131
GEO Datasets	201
UniGene	17,479
UniSTS	493
PubMed Central	500
SRA Experiments	13
Probe	814
Bio Project	5
Bio Sample	82
GEO Profiles	9,216
Taxonomy	1

Comments and References:

- [FNA - Pinaceae](#)
Thieret JW. Pinaceae Lindley. 1993. In Flora of North America Editorial Committee (Eds.) Flora of North America North of Mexico, Vol. 2: Pteridophytes and Gymnosperms. New York and Oxford. On-line version.
- [Gymnosperm Database](#)
Name verified on date of entry into taxonomy database in: Earle, CJ. 2004 onward. The Gymnosperm Database. (on-line)
- [GRIN \(Apr 16, 2009\)](#)
Name verified on 16 April 2009 in: USDA, ARS, National Genetic Resources Program. Germplasm Resources Information Network - (GRIN) [Online Database]. National Germplasm Resources Laboratory, Beltsville, Maryland.
- [Flora of China - Pinaceae](#)
Liguo Fu, Nan Li, Thomas S. Elias & Robert R. Mill. Pinaceae Lindley. In Wu, Z. Y. & P. H. Raven, eds. July 1999. Flora of China. Vol. 4 (Cycadaceae through Fagaceae). Science Press, Beijing, and Missouri Botanical Garden Press, St. Louis.
- [FNA - Pinus](#)
Kral R. 1993. Pinus Linnaeus, Sp. Pl. 2: 1000. ; Gen Pl. ed. 5. 1753; Gen. Pl. ed. 5, 434, 1754. In Flora of North America Editorial Committee (Eds.) Flora of North America North of Mexico, Vol. 2: Pteridophytes and Gymnosperms. New York and Oxford. On-line version.

Genome information

Genome view: 12 chromosomes

Find: table Highlight all Match case

NCBI resources, search and retrieval functions

NCBI Resources How To Sign in to NCBI

NCBI National Center for Biotechnology Information

All Databases Pseudotsuga menziesii[organism] AND Krutovsky[author] Search

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

Entrez, The Life Sciences Search Engine

Search across databases Pseudotsuga menziesii[organism] AND Krutovsky[author] GO Clear Help

- Result counts displayed in gray indicate one or more terms not found

4 PubMed: biomedical literature citations and abstracts	none Books: online books
5 PubMed Central: free, full text journal articles	none OMIM: online Mendelian Inheritance in Man
none Site Search: NCBI web and FTP sites	

3080 Nucleotide: Core subset of nucleotide sequence records	none dbGaP: genotype and phenotype
18055 EST: Expressed Sequence Tag records	none UniGene: gene-oriented clusters of transcript sequences
none GSS: Genome Survey Sequence records	none CDD: conserved protein domain database
3080 Protein: sequence database	none Clone: integrated data for clone resources
none Genome: whole genome sequences	none UniSTS: markers and mapping data
none Structure: three-dimensional macromolecular structures	139 PopSet: population study data sets
none Taxonomy: organisms in GenBank	none GEO Profiles: expression and molecular abundance profiles
none SNP: short genetic variations	none GEO DataSets: experimental sets of GEO data
none dbVar: Genomic structural variation	none Epigenomics: Epigenetic maps and data sets
none Gene: gene-centered information	none PubChem BioAssay: bioactivity screens of chemical substances
none SRA: Sequence Read Archive	none PubChem Compound: unique small molecule chemical structures
none BioSystems: Pathways and systems of interacting molecules	none PubChem Substance: deposited chemical substance records
none HomoloGene: eukaryotic homology groups	none Protein Clusters: a collection of related protein sequences
7 Probe: sequence-specific reagents	none OMIA: online Mendelian Inheritance in Animals
none BioProject: aggregated biological research project data	1 BioSample: biological material descriptions

none NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections	none MeSH: detailed information about NLM's controlled vocabulary
---	---

Genbank format: features and sequence annotation

NCBI Resources How To

Nucleotide

[Limits](#) [Advanced](#)

[Display Settings:](#) GenBank

[Send:](#)

Pseudotsuga menziesii var. menziesii haplotype Pm-AT1_412m2 alpha tubulin 1 (AT1) gene, complete cds

GenBank: AY832610.1

<http://www.ncbi.nlm.nih.gov/nucore/AY832610.1>

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

[Go to:](#)

LOCUS AY832610 2575 bp DNA linear PLN 26-JAN-2007
DEFINITION Pseudotsuga menziesii var. menziesii haplotype Pm-AT1_412m2 alpha tubulin 1 (AT1) gene, complete cds.
ACCESSION AY832610
VERSION AY832610.1 GI:56481496
KEYWORDS .
SOURCE Pseudotsuga menziesii var. menziesii
ORGANISM [Pseudotsuga menziesii var. menziesii](#)
Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Coniferopsida; Coniferales; Pinaceae; Pseudotsuga.
REFERENCE 1 (bases 1 to 2575)
AUTHORS Krutovsky,K.V. and Neale,D.B.
TITLE Nucleotide diversity and linkage disequilibrium in cold-hardiness- and wood quality-related candidate genes in Douglas fir
JOURNAL Genetics 171 (4), 2029-2041 (2005)
PUBMED [16157674](#)
REFERENCE 2 (bases 1 to 2575)
AUTHORS Krutovsky,K.V. and Neale,D.B.
TITLE Direct Submission
JOURNAL Submitted (19-NOV-2004) Department of Plant Sciences, University of California, Institute of Forest Genetics, Pacific Southwest Research Station, One Shields Avenue, Davis, CA 95616, USA
FEATURES Location/Qualifiers
source 1..2575
/organism="Pseudotsuga menziesii var. menziesii"



Genbank format: features and sequence annotation

```

FEATURES             Location/Qualifiers
     source            1..2575
                        /organism="Pseudotsuga menziesii var. menziesii"
                        /mol_type="genomic DNA"
                        /variety="menziesii"
                        /db_xref="taxon:278161"
                        /haplotype="Pm-AT1_412m2"
                        /dev_stage="haploid megagametophyte"
                        /country="USA: Springfield breeding zone, Oregon"
     gene                <1..>2575
                        /gene="AT1"
     mRNA              join(<1..192,622..856,1267..1637,1826..>2575)
                        /gene="AT1"
                        /product="alpha tubulin 1"
     exon                <1..192
                        /gene="AT1"
                        /number=1
     5'UTR              <1..99
                        /gene="AT1"
     CDS                join(100..192,622..856,1267..1637,1826..2482)
                        /gene="AT1"
                        /codon_start=1
                        /product="alpha tubulin 1"
                        /protein_id="AAV92379.1"
                        /db_xref="GI:56481497"
                        /translation="MRECISIHIGQAGIQVGNACWELYCLEHGIQPDGQMPDQKTVGG
GDDAFNITFFSETGAGKHVPRAVFVDLEPTVIDEVRTGTGRQLFHPEQLISGKEDAANN
FARGHYTIGKEIVDLCLDRIRKLDNCTGLQGFLVFNNAVGGGTGSGLSLLERLSVD
YGKSKLGFVYVSPQVSTSVVEPYNSVLSHSLLEHTDVAVLLDNEAIYDICRRSLD
IERPTYTNLNLRLVSQVSSLTASLRFDGALNVDVTEFQTNLVPYPRIFHMLSSYAPVI
SAEKAYHEQLSVAEITNSAFEPSSMMAKCDPRHGKYMCCCLMYRGDVVPKDVNAAVAT
IKTKRTIQFVDWCPTGFKCGINYPPTVVPGGDLAKVQRAVCMISNSTSVAEVFSRID
HKFDLMYAKRAFVHWYVGEEMEEGEFSEAREDLAALEKDYEEVGAESAEGEGDEGDE
Y"
     exon                622..856
                        /gene="AT1"
                        /number=2
     exon                1267..1637
                        /gene="AT1"
                        /number=3
     exon                1826..>2575
                        /gene="AT1"
                        /number=4
     3'UTR              2483..>2575
                        /gene="AT1"

ORIGIN
1 tctttgcaaa cgccttcttc atccgatttc ctacgtcccg cegtctcttt gctgttttta
61 cctgttttta cgtttttttt cctcttcccg ccagcgaaaa tggagagag cctctcagtc

```



Introduction to BLAST: Basic Local Alignment Search Tool

- widely used search tool to find similar nucleotide or amino acid sequences developed in 1990 and 1997 (Stephen Altschul)
- heuristic approach for performing local alignments through searches of high scoring segment pairs (HSP's) and based on Smith-Waterman algorithm
- inferring function of a query sequence from its similarity with well-studied and annotated sequences
- uses statistics to predict significance of initial matches and to find best local alignments
- provides statistical significance for alignments
- accurate and fast
- a suit of tools (www, standalone, network clients, etc.)



BLAST Concepts for Sequence Similarity Searching

- One sequence by itself is not informative; it must be analyzed by comparative methods against existing sequence databases to develop hypothesis concerning structure and function.
- looks for clusters of nearby or locally dense “similar or homologous” k -tuples
- uses “look-up” tables to shorten search time
- uses larger “word size” than FASTA to accelerate the search process
- performs both Global and Local alignment
- fastest and most frequently used sequence alignment tool



How Does BLAST Really Work?

- The BLAST programs improved the overall speed of searches while retaining good sensitivity (important as databases continue to grow) by breaking the query and database sequences into fragments ("words"), and initially seeking matches between fragments.
- “Word” hits are then extended in either direction in an attempt to generate an alignment with a maximum score value "S".



Alignment

Query sequence: AAC**CG**TTC---T**ACAATTAC**CTAGGC
 ===--=== =-====-=====
Best match sequence: AAC**G**T**TTC****CAG**T**C**CAA**A**T**A**G**C**TAGGC

Hits (+1 per nucleotide): 1×18 (matching nucleotides) = 18

Penalty for mismatching (-2 per nucleotide): -2×5
(mismatching nucleotides) = -10

Penalty for gaps opening (-2 per gap) \times # of gaps + penalty
for extension (-1 per nucleotide): -2×1 (# of gaps) + -1×3
(# of nucleotides in the gap)

Score = $18 \times 1 + 5 \times (-2) + 1 \times (-2) + 3 \times (-1) = 3$



Alignment

Global Alignment:

- compares total length of two sequences
 - [Needleman, S.B. and Wunsch, C.D.](#) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 48(3):443-53(1970)

Local Alignment:

- compares segments of sequences
- finds cases when one sequence is a part of another sequence, or they only match in parts.
 - [Smith, T.F. and Waterman, M.S.](#) Identification of common molecular subsequences. J Mol Biol. 147(1):195-7 (1981)

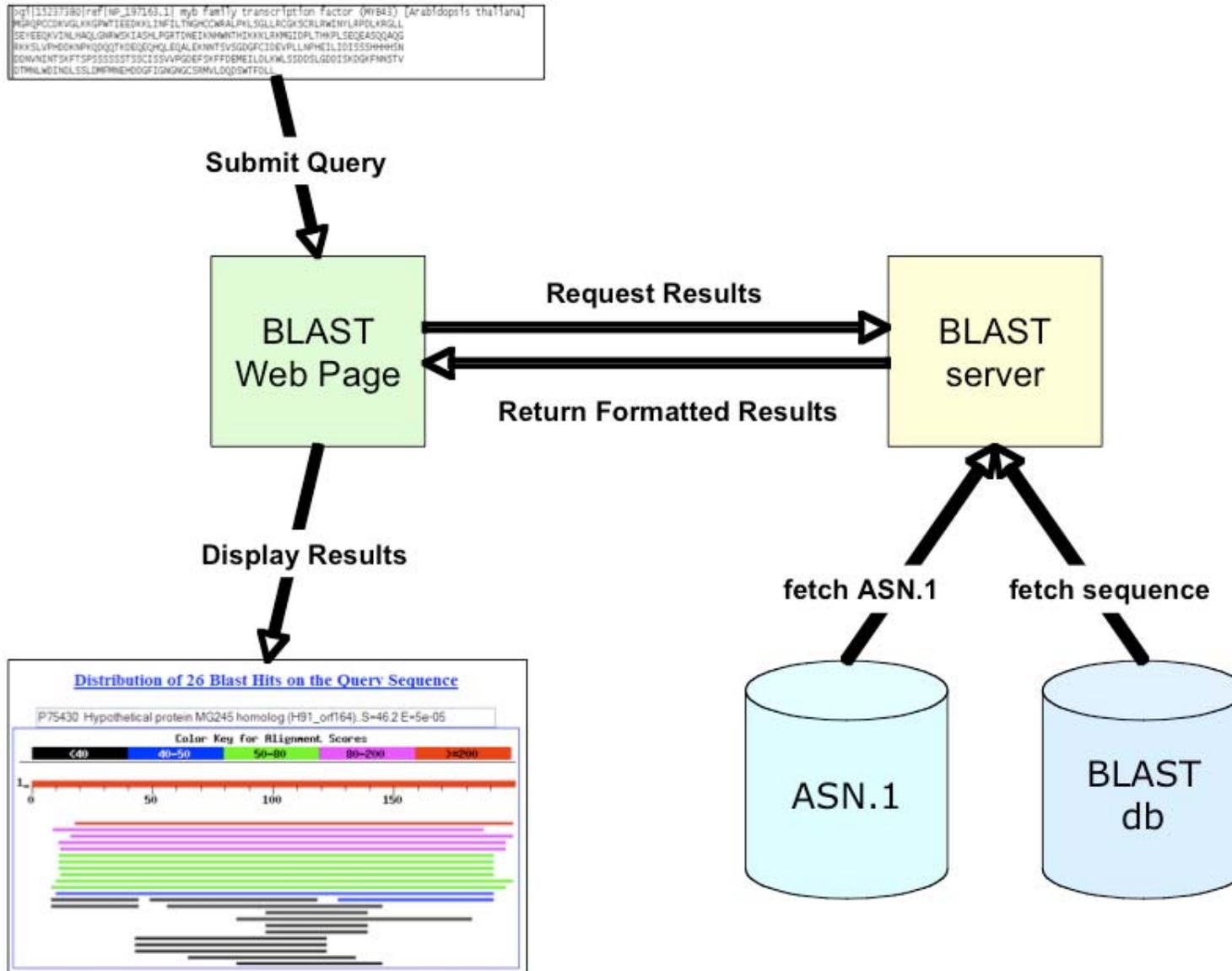


The BLAST algorithm

- The BLAST programs (Basic Local Alignment Search Tools) are a set of sequence comparison algorithms introduced in 1990 that are used to search sequence databases for optimal local alignments to a query.
 - Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) “Basic local alignment search tool.” J. Mol. Biol. 215:403-410.
 - Altschul SF, Madden TL, Schaeffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.” NAR 25:3389-3402.



The BLAST algorithm



BLAST Algorithm

- Scoring of matches done using scoring matrices
- Sequences are split into words (default $n=3$)
 - speed, computational efficiency
- BLAST algorithm extends the initial “seed” hit into an HSP
 - HSP = high scoring segment pair = Local optimal alignment



Sequence Similarity Searching

– The statistics are important

- Discriminating between real and artifactual matches is done using an estimate of probability that the match might occur by chance (expectation values or E-values).
- We'll talk more about the meaning of the scores (S) and e-values (E) that are associated with BLAST hits



What do the Score and the E-value really mean?

- The quality of the alignment is represented by the **Score (S)**.
 - The **score** of an alignment is calculated as the **sum of substitution and gap scores**. Substitution scores for amino acid sequence alignment are given by a look-up substitution matrix (PAM, BLOSUM) whereas gap scores are assigned empirically .
- The significance of each alignment is computed as an **E-value (E)**.
 - **Expectation value**. The number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. The lower the E value, the more significant the score.



Notes on E-values

- Low E-values suggest that sequences are homologous
 - ◉ can't show non-homology
- Statistical significance depends on both the size of the alignments and the size of the sequence database
 - ▶ important consideration for comparing results across different searches
 - ▶ E-value increases as database gets bigger
 - ▶ E-value decreases as alignments get longer



Where does the score (S) come from?

- The quality of each pair-wise alignment is represented as a score (S) and the scores are ranked.
- **Scoring matrices** are used to calculate the score of the alignment base by base (DNA) or amino acid by amino acid (protein).
- **The alignment score is the sum of the scores for all positions together.**



What is a scoring matrix?

- Substitution matrices are used for amino acid alignments.

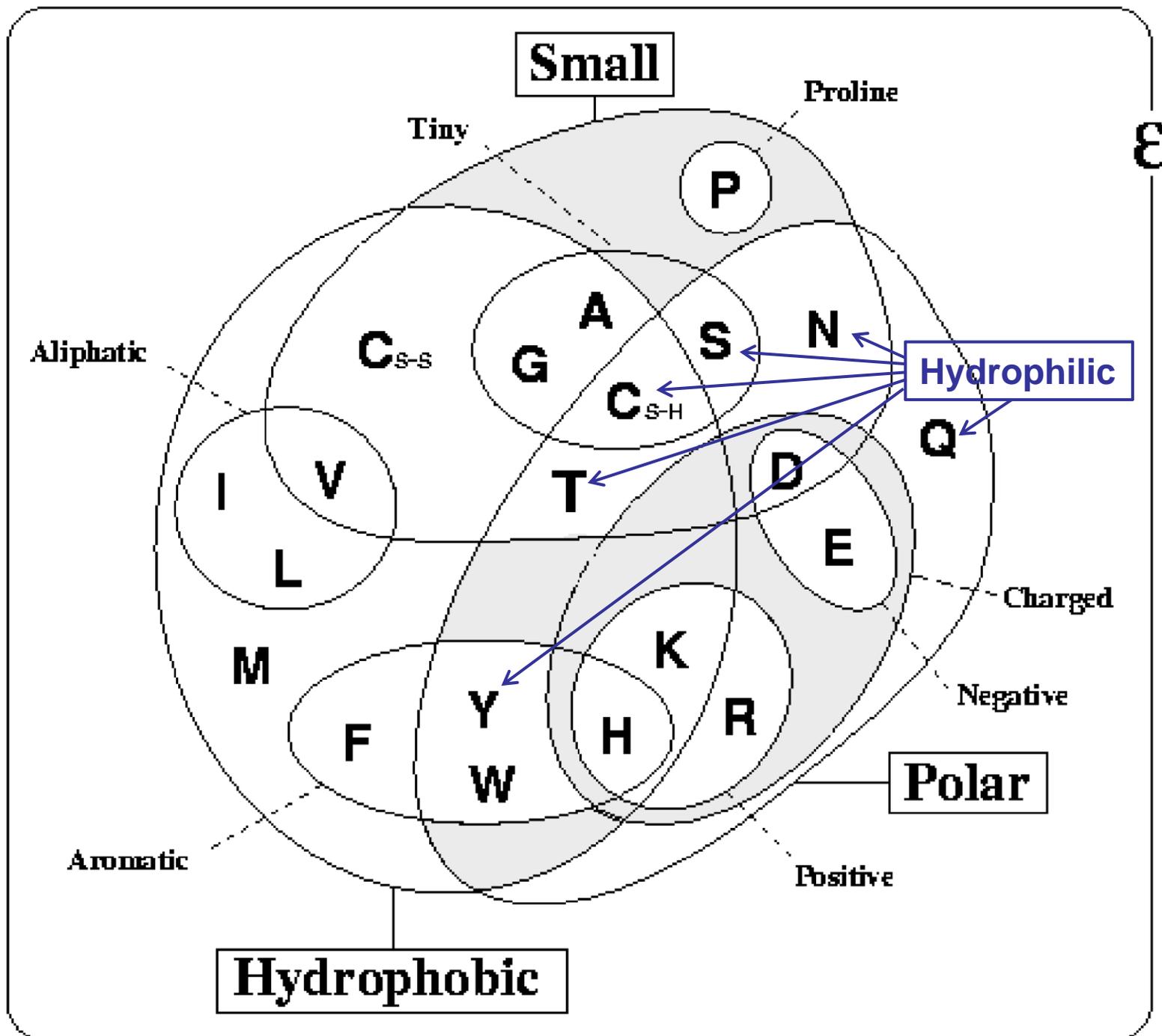
– each possible residue substitution is given a score

	A	C	D	E	F	G	H
alanine Ala A	4	0	-2	-1	-2	0	-2
cysteine Cys C	0	9	-3	-4	-2	-3	-3
aspartic acid Asp D	-2	-3	6	2	-3	-1	-1
glutamic acid Glu E	-1	-4	2	5	-3	-2	0
phenylalanine Phe F	-2	-2	-3	-3	6	-3	-1
glycine Gly G	0	-3	-1	-2	-3	0	-1
histidine His H	-2	-3	-1	0	-1	-1	0

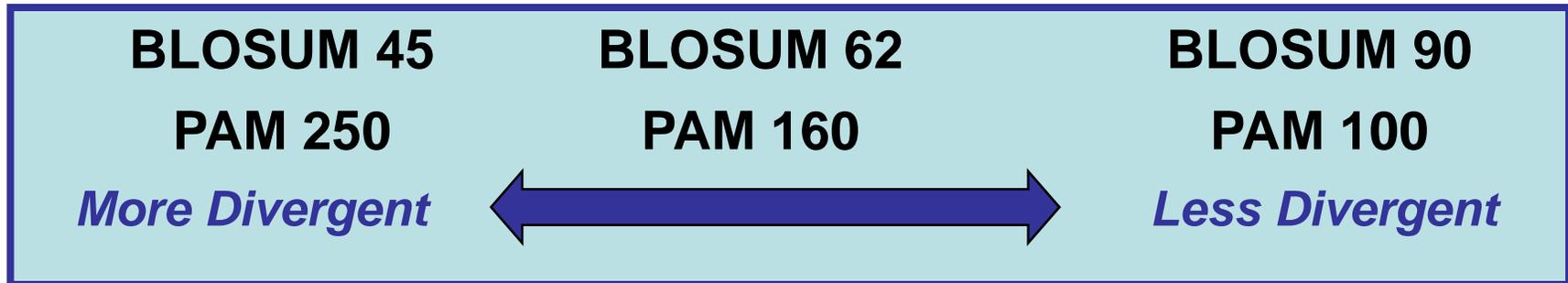
BLOSUM - BLOcks SUBstitution Matrix
(Henikoff & Henikoff 1992)

- A simpler unitary matrix is used for nucleotide pairs (+1 for match, -2 mismatch)





BLOSUM vs. PAM



- **BLOSUM 62 (BLOcks SUBstitution Matrix)** is the default matrix in BLAST 2.0. Though it is tailored for comparisons of moderately distant proteins, it performs well in detecting closer relationships. A search for distant relatives may be more sensitive with a different matrix, such as **PAM (Point Accepted Mutation)**.



Suggested BLAST Cutoffs

- For nucleotide based searches, one should look for hits with E-values of 10^{-6} or less and sequence identity of 70% or more
- For protein based searches, one should look for hits with E-values of 10^{-3} or less and sequence identity of 25% or more

Take Home Message:
Always look at your alignments

Chapter 11 in “Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins” by Andreas D. Baxevanis and B. F. Francis Ouellette (Editors)



Homology: Some Guidelines

- Similarity can be indicative of homology
- Generally, if two sequences are significantly similar over entire length they are likely homologous
- Low complexity regions can be highly similar without being homologous
- Homologous sequences not always highly similar



What BLAST tells you ...

- reports alignments (sometimes surprising)
 - similarities imply evolutionary homology, descent from a common ancestor, but does not always imply similar function
- provides statistical support that alignments are not by chance (E-values)
- provide level of identity and similarity



Phosphatase and tensin homolog (PTEN)

```
>|_gb|AAL08419.1| PTEN [Takifugu rubripes]
Length=412
```

```
Score = 197 bits (501), Expect = 2e-49, Method: Composition-based stats.
Identities = 95/100 (95%), Positives = 98/100 (98%), Gaps = 0/100 (0%)
```

```
Query 2 IVSRNKRRYQEDGFDDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKHNYKI 61
+VSRNKRRYQEDGFDDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKHNYKI
Sbjct 8 MVS RNKRRYQEDGFDDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKHNYKI 67

Query 62 YNLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFKQN 101
YNLCAERHYD AKFNCRVAQYPFEDHNPPQLELIKPF ++
Sbjct 68 YNLCAERHYDAAKFNCRVAQYPFEDHNPPQLELIKPFCE D 107
```

Resulting alignment is called the HSP (high scoring segment pair = local optimal alignment) – more than one HSP per hit possible

```
Score = 83.6 bits (205), Expect = 4e-15, Method: Composition-based stats.
Identities = 60/103 (58%), Positives = 68/103 (66%), Gaps = 32/103 (31%)
```

```
Query 99 KQNKMLKKDKMFHFWVNTFFIPGPEEV-----D 126
KQNKMK+KKDKMFHFWVNTFFIPGPEE +
Sbjct 260 KQNKMMKKDKMFHFWVNTFFIPGPEESRDKLENGAVNNADSQGGVPAPGGQPQSAECRE 319

Query 127 NDKEYLVLTLTkndldkankdkanRYFSPNFKVKLYFTKTVEE 169
+D++YL+LTL+KND DKANKDKANRYFSPNFKVKL F+KTVEE
Sbjct 320 SDRDY LILTL SKNDRDKANKDKANRYFSPNFKVKLCFSKTVEE 362
```

```
>|_gb|AAH93110.1| UG Ptenb protein [Danio rerio]
Length=289
```

```
Score = 197 bits (500), Expect = 2e-49, Method: Composition-based stats.
Identities = 95/99 (95%), Positives = 98/99 (98%), Gaps = 0/99 (0%)
```

```
Query 3 VSRNKRRYQEDGFDDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKHNYKIY 62
VSRNKRRYQEDGFDDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHK+HYKIY
Sbjct 9 VSRNKRRYQEDGFDDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKDHYKIY 68

Query 63 NLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFKQN 101
NLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPF ++
Sbjct 69 NLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFCE D 107
```



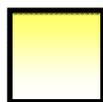
BLAST programs

Program	Description
blastp	Compares an amino acid query sequence against a protein sequence database.
blastn	Compares a nucleotide query sequence against a nucleotide sequence database.
blastx	Compares a nucleotide query sequence translated in all reading frames against a protein sequence database. You could use this option to find potential translation products of an unknown nucleotide sequence.
tblastn	Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
tblastx	Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.



more BLAST programs

Program		Notes
Megablast	Contiguous	Nearly identical sequences
	Discontiguous	Cross-species comparison
Position Specific	PSI-BLAST	Automatically generates a position specific score matrix (PSSM)
	RPS-BLAST	Searches a database of PSI-BLAST PSSMs



nucleotide only



protein only



Other BLAST tools and services

- **MEGABLAST** - for comparison of large sets of long DNA sequences
- **RPS-BLAST** - Conserved Domain Detection
- **BLAST 2 Sequences** - for performing pairwise alignments for 2 chosen sequences
- **Genomic BLAST** - for alignments against select human, microbial or malarial genomes
- **VecScreen** - for detecting cloning vector contamination in sequenced data



BLAST Access

- **NCBI BLAST**

- <https://blast.ncbi.nlm.nih.gov/Blast.cgi>

- **Canadian Bioinformatics Resource: Similarity Searching and Classification**

- https://bioinformatics.ca/links_directory/category/sequence-comparison/similarity-searching-and-classification

- **European Bioinformatics Institute BLAST**

- <http://www.ebi.ac.uk/Tools/sss/ncbiblast/nucleotide.html>



BLAST at the NCBI website

The screenshot shows the NCBI BLAST website. At the top, there is a navigation bar with links for 'Most Visited', 'Getting Started', 'Latest Headlines', and 'Google Scholar'. Below this is the 'BLAST' logo and a search bar. The main content area is titled 'Basic Local Alignment Search Tool' and includes a brief description of BLAST's function. A 'Web BLAST' section offers three main options: 'Nucleotide BLAST' (nucleotide to nucleotide), 'blastx' (translated nucleotide to protein), and 'Protein BLAST' (protein to protein). There is also a 'BLAST Genomes' section with a search input field and buttons for 'Human', 'Mouse', 'Rat', and 'Microbes'. A 'Specialized searches' section lists various tools like SmartBLAST, Primer-BLAST, Global Align, CD-search, GEO, IgBLAST, VecScreen, CDART, Targeted Loci, Multiple Alignment, BioAssay, and MOLE-BLAST, each with a brief description of its purpose.



To learn about a genome of interest, visit NCBI → Taxonomy → TaxBrowser → Genome

Search for *Saccharomyces cerevisiae* as complete name lock Go

Display 0 levels using filter: none

Saccharomyces cerevisiae

Taxonomy ID: 4932
 Genbank common name: **baker's yeast**
 Inherited blast name: **ascomycetes**
 Rank: species
 Genetic code: [Translation table 1 \(Standard\)](#)
 Mitochondrial genetic code: [Translation table 3 \(Yeast Mitochondrial\)](#)
 Other names:
 synonym: **Saccharomyces uvarum var. melibiosus**
 synonym: **Saccharomyces oviformis**
 synonym: **Saccharomyces italicus**
 synonym: **Saccharomyces capensis**
 common name: yeast
 common name: **lager beer yeast**
 common name: **brewer's yeast**
 anamorph: **Candida robusta**
 type material: NRRL Y-12632
 type material: CBS 1171
 type material: ATCC 18824

[Lineage \(full\)](#)
 cellular organisms; Eukaryota; Opisthokonta; Fungi; Dikarya; Ascomycota; saccharomyceta; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomycetes

Entrez records		
Database name	Subtree links	Direct links
Nucleotide	128,703	111,295
Nucleotide EST	34,915	34,915
Nucleotide GSS	7,638	7,638
Protein	654,340	40,184
Structure	3,390	2,461
Genome	1	1
Popset	360	360
Domains	6	3
GEO Datasets	45,859	45,020
PubMed Central	77,159	77,153
Gene	7,060	640
HomoloGene	4,106	4,106
SRA Experiments	47,606	41,848
Probe	12,210	6,655
Assembly	431	254
Bio Project	2,675	2,188
Bio Sample	65,238	60,427
Bio Systems	1,494	1,275
GEO Profiles	1,047,442	1,047,442
PubChem BioAssay	3,076	2,977
Taxonomy	319	1



To learn about a genome of interest, follow the TaxBrowser → Genome links

The screenshot shows the NCBI TaxBrowser interface. At the top, there's a search bar with the query 'txid4932[Organism:exp]'. Below the search bar, there's a notification about testing https on public web servers. The main content area is divided into several sections:

- Saccharomyces cerevisiae (baker's yeast)**: Reference genome: Saccharomyces cerevisiae S288c (assembly R64). Download sequences in FASTA format for genome, transcript, protein. Download genome annotation in GFF, GenBank or tabular format. BLAST against Saccharomyces cerevisiae genome, transcript, protein. All 395 genomes for species: Browse the list. Download sequence and annotation from RefSeq or GenBank.
- Organism Overview**: Genome Assembly and Annotation report [395]; Plasmid Annotation Report [70]; Organelle Annotation Report [153]. ID: 15. Assembly, BioProject, Gene, Components, Protein, PubMed, Taxonomy.
- Summary**: Sequence data: genome assemblies: 395; sequence reads: 106 (See Genome Assembly and Annotation report). Statistics: median total length (Mb): 12.1334, median protein count: 5407, median GC%: 38.3952.
- Publications**: 1. Genome Sequence and Analysis of a Stress-Tolerant, Wild-Derived Strain of Saccharomyces cerevisiae Used in Biofuels Research. McIlwain SJ, et al. G3 (Bethesda) 2016 Jun 1.

On the right side, there are sections for NCBI Resources (Map Viewer), Tools (BLAST Genome), and Search details (txid4932 [Organism:exp]).

Size (in megabases), number of chromosomes are given here



Saccharomyces cerevisiae (... x) +

https://www.ncbi.nlm.nih.gov/genome/ Search

Most Visited Getting Started Latest Headlines Google Scholar

Representative (genome information for reference and representative genomes)

Reference genome: [see all organisms]

- Saccharomyces cerevisiae S288c
 - Submitter: Saccharomyces Genome Database

Loc	Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
Nuc	Chr	I	NC_001133.9	BK006935.2	0.230218	39.3	94	-	4	2	101	1
Nuc	Chr	II	NC_001134.8	BK006936.2	0.813184	38.3	408	-	13	4	425	-
Nuc	Chr	III	NC_001135.5	BK006937.2	0.31662	38.5	163	-	10	4	179	2
Nuc	Chr	IV	NC_001136.10	BK006938.2	1.53	37.9	755	-	28	4	788	1
Nuc	Chr	V	NC_001137.3	BK006939.2	0.576874	38.5	279	-	20	9	309	1
Nuc	Chr	VI	NC_001138.5	BK006940.2	0.270161	38.7	125	-	10	4	140	1
Nuc	Chr	VII	NC_001139.9	BK006941.2	1.09	38.1	530	-	36	10	576	-
Nuc	Chr	VIII	NC_001140.6	BK006934.2	0.562643	38.5	282	-	11	4	297	-
Nuc	Chr	IX	NC_001141.2	BK006942.2	0.439888	38.9	211	-	10	3	230	6
Nuc	Chr	X	NC_001142.9	BK006943.2	0.745751	38.4	359	-	24	6	389	-
Nuc	Chr	XI	NC_001143.9	BK006944.2	0.666816	38.1	313	-	16	5	334	-
Nuc	Chr	XII	NC_001144.5	BK006945.2	1.08	38.5	509	12	21	18	562	2
Nuc	Chr	XIII	NC_001145.3	BK006946.2	0.924431	38.2	461	-	21	15	497	-
Nuc	Chr	XIV	NC_001146.8	BK006947.3	0.784333	38.6	398	-	14	6	418	-
Nuc	Chr	XV	NC_001147.6	BK006948.2	1.09	38.2	536	-	20	11	569	2
Nuc	Chr	XVI	NC_001148.4	BK006949.2	0.948066	38.1	465	-	17	6	490	2
MT	Chr	MT	NC_001224.1	KP283414.1	0.065779	17.1	19	2	24	1	46	-

Chromosomes

Click on chromosome name to open MapViewer

External Resources

Saccharomyces cerevisiae Genome

txid4932[Organism.exp] (1) Genome

Krutovsky[author] (445) Probe

Douglas-fir cDNA library PmlFG_73-6 biosample

Pseudotsuga menziesii[organism:noexp] AND Krutovsky[author] (6) PMC

See more...

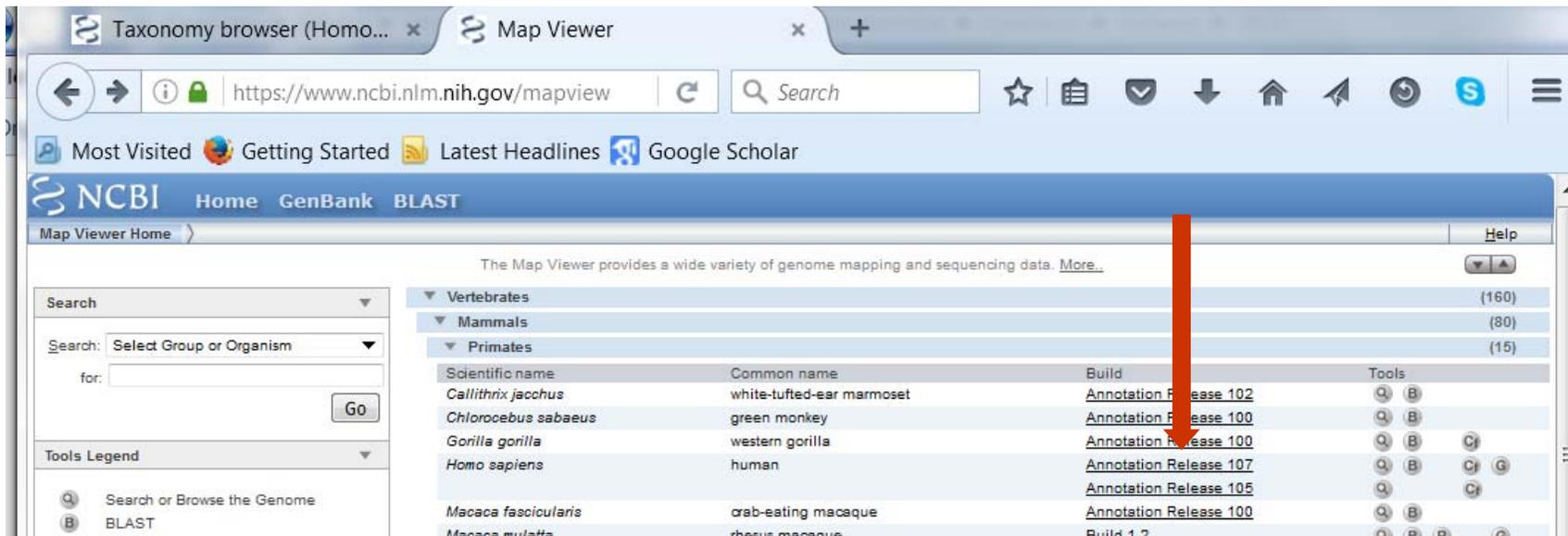


To explore human chromosome 21 at NCBI

→ Find MapViewer

→ Choose Primates

→ Click [Annotation Release 107](#) and, then, chromosome 21



https://www.ncbi.nlm.nih.gov/projects/r

Most Visited Getting Started Latest Headlines Google Scholar

NCBI

NCBI Map Viewer

PubMed Nucleotide Protein Genome Gene Structure PopSet Taxonomy Help

Search for on chromosome(s) Find Advanced Search

Map Viewer

Map Viewer Home
Map Viewer Help
fruit fly Maps Help
FTP

NCBI Resources

Assembly
Gene
Genome

Organism Data in GenBank

EST
Genomic
mRNA
Protein
WGS

***Drosophila melanogaster* (fruit fly) genome view** [BLAST search *Drosophila melanogaster* sequences](#)

Release 5.30 statistics

X 2L 2R 3L 3R 4 Y HT

Lineage: [Eukaryota](#); [Metazoa](#); [Ecdysozoa](#); [Arthropoda](#); [Hexapoda](#); [Insecta](#); [Pterygota](#); [Neoptera](#); [Endopterygota](#); [Diptera](#); [Brachycera](#); [Muscomorpha](#); [Ephydroidea](#); [Drosophilidae](#); [Drosophila](#); [Sophophora](#); [Drosophila melanogaster](#)

July 2014, Release 5.57

This full annotation run includes the following assembly(ies):

- Release 5 (accession [GCF_000001215.2](#))

From early observations of the banding patterns of its polytene chromosomes to current work on mRNA and protein gradients in the developing embryo, *Drosophila melanogaster* has been studied in biology labs for over eighty years. Many of the genes that define the spatial pattern of cell types and body parts have now been identified, along with the regulatory pathways in which they operate. As the majority of these genes have counterparts in higher eukaryotes, the study of the *Drosophila* developmental program provides insight into human development, as well.

D. melanogaster is a member of the melanogaster group of the subgenus *Sophophora* and one of 12 fruit fly species whose genome is being sequenced to completion for use in comparative genomics studies. With previous information on the phylogenetic relationship between these species and the sequence of their genomes, a resource will be created that will provide reagents for studies in molecular evolution, gene function and comparative annotation across the entire *Drosophila* genus. Twenty fruit fly species have been approved for construction of BAC libraries, which will further assist in expanding our knowledge of gene homologies and pathway discovery.

The NCBI Map Viewer provides graphical displays of features on the genome assembly. Map features that can be seen along the sequence include annotated genes and transcripts, Gnomon-predicted gene and transcript models, aligned transcript and genomic sequences, RefSeq scaffolds (the 'Contig' map), the assembly tiling path (the 'Component' map), and more. For some species, additional non-sequence maps such as Genetic maps, Radiation hybrid maps, and others may be available.

References and Credits

- Materials for this presentation have been adapted from the following sources:
 - NCBI HelpDesk - Field Guide Course Materials
 - Bioinformatics: A practical guide to the analysis of genes and proteins
- Strongly recommend BLAST tutorial on NCBI site
 - <http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html>
- Further “Bioinformatics for quantitative geneticists course notes” J. McEwan
 - http://www-personal.une.edu.au/~jvanderw/aabc_materials2004.htm#ModuleC

