ГЕНОМИКА

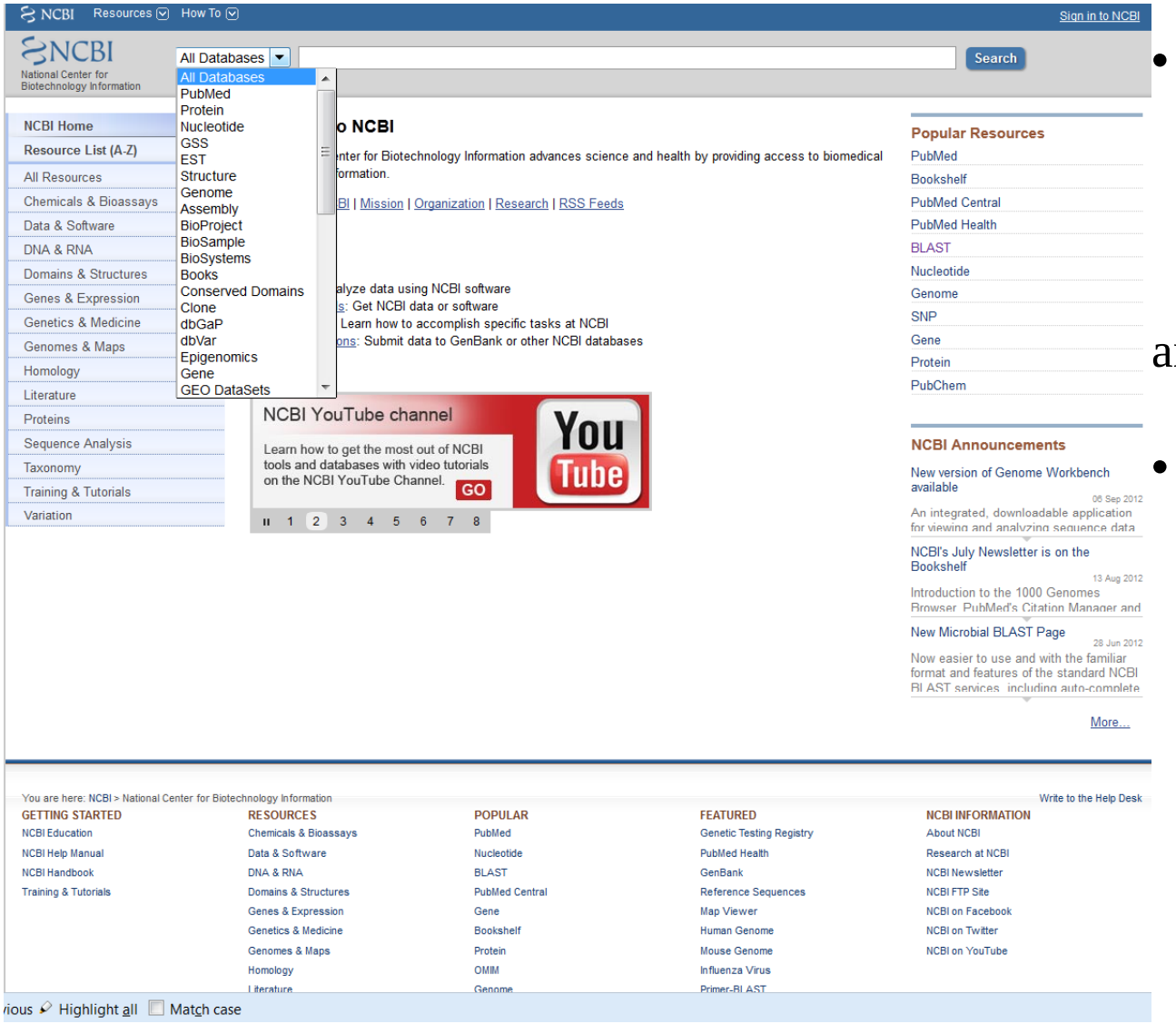
18 марта 2020, Среда:

- 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



- 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 fiel

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             Location/Qualifiers
     source          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="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     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="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF
                     TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLSDANTTLYFN
                     VILEGTDSADSTSLNNTYQFVVTNRPSISLSSDFNLLALLKNYGYTNGKNALKLDPNE
                     VFNVTFDRSMFTNEESIVSYYGRSQLYNAPLPNWLFFDSGELKFTGTAPVINSAIAPE
                     TSYSFVIIATDIEGFSAVEVEFELVIGAHQLTTSIQNSLIINVTDTGNVSYDLPLNYV
```

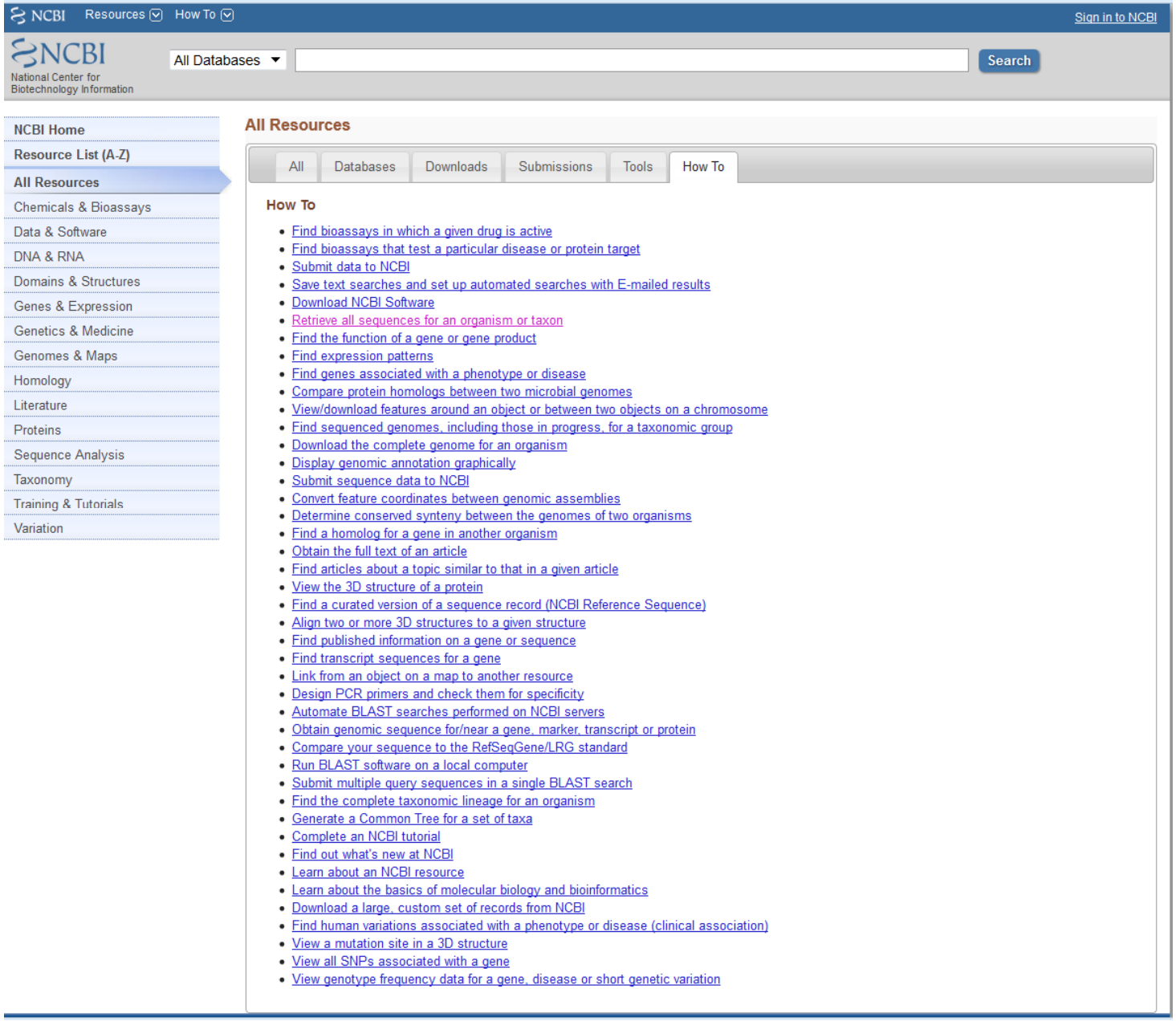
- Narrow your search using particular fields: [definition], [organism], [author], [subtree], [lineage], etc. (otherwise, it will search all fields)

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, search and retrieval functions

NCBI Resources How To Sign in to NCB

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

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search for as complete name lock Go Clear
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	17,379
UniSTS	493
PubMed Central	500
SRA Experiments	13
Probe	814
Bio Project	2
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.

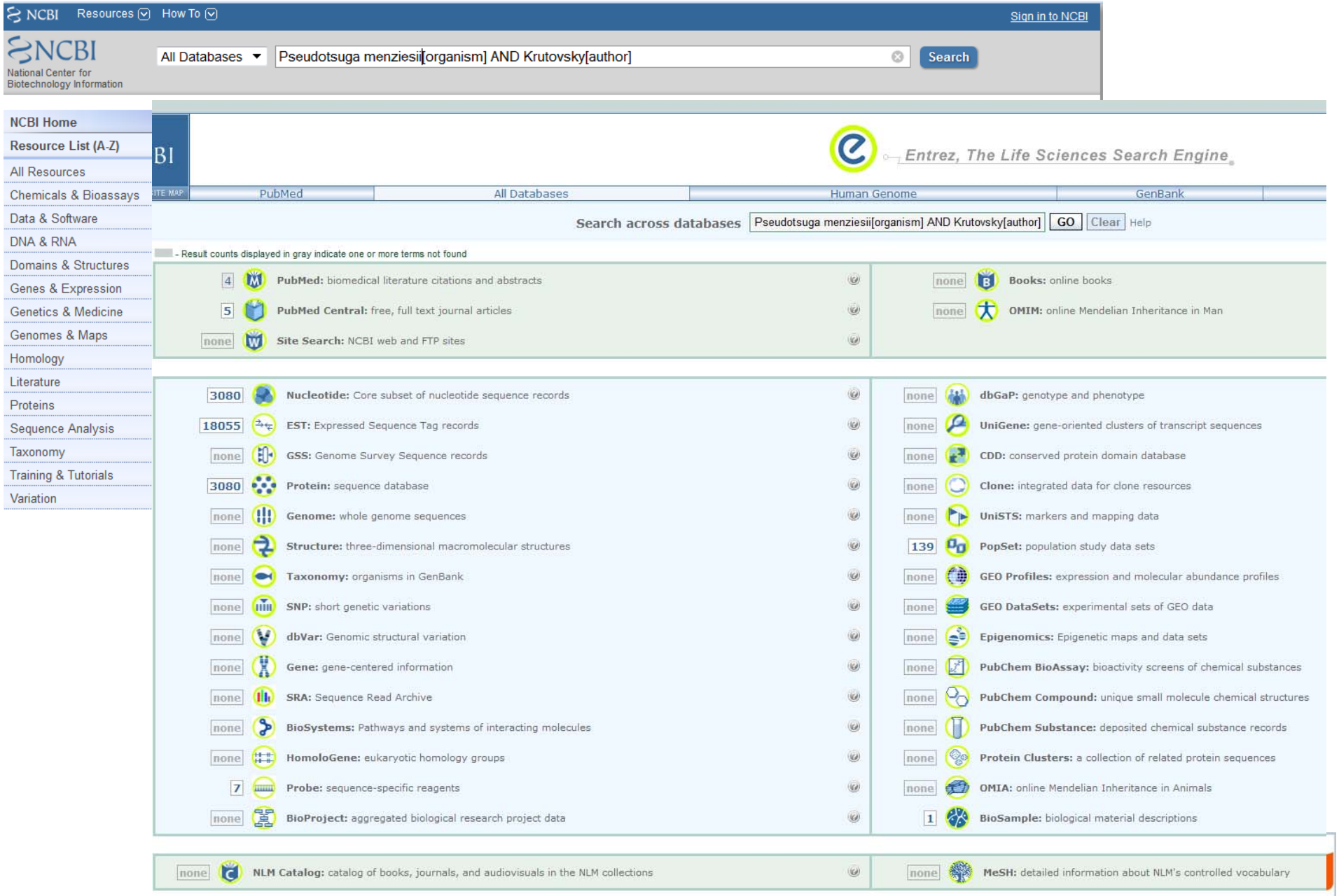
FNA - Pinus
Kral R. 1993. Pinus Linnaeus, Sp. Pl. 2: 1000. ; Gen Pl. ed. 5. 1753; Gen. Pl. ed. 5, 434, 1754. In Flora of North America Editorial Committee (Eds.) Flora of North America North of Mexico, Vol. 2: Pteridophytes and Gymnosperms. New York and Oxford. On-line version.

Genome Information

Genome view: 12 chromosomes

Find: table Next Previous Highlight all Match case

NCBI resources, search and retrieval functions



Genbank format: features and sequence annotation

NCBI Resources How To

Nucleotide 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/nuccore/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="MRECISIHIGQAGIQVGNACWELYCLEHGIQPDGQMPSDKTVGG
                     GDDAFNTFFSETGAGKHVPRAVFVDLEPTVIDEVRTGTYRQLFHPEQLISGKEDAANN
                     FARGHYTIGKEIVDLCLDRIRKLADNCTGLQGFLVFNAVGGGTGSGLGSLLLERLSVD
                     YGKKSKLGFTVYPSPQVSTSVVEPYNSVLSTHSLLEHTDVAVLLDNEAIYDICRRSLD
                     IERPTYTNLNRLVSQVISSLTASLRFDGALNVDVTEFQTNLVPYPRIHFMLSSYAPVI
                     SAEKAYHEQLSVAEITNSAFEPSSMMAKCDPRHGKYMACCLMYRGDVVPKDVNAAVAT
                     IKTKRTIQFVDWCPTGFKCGINYQPPTVVPGGDLAKVQRAVCMISNSTSVAEVFSRID
                     HKFDLMYAKRAFVHWYVGEGMEEGEFSEAREDLAALEKDYEEVGAESAEGEEGDEGDE
                     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 ccgttctttt gctgtttta
       61 ctctgttttg gggtttttgt ccgttgaccg ccagccgaaaa tgagagagtg catctcgatc
```

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:        AACCGTTC---TACAATTACCTAGGC
                       ===--===   =-===-==-======
Best match sequence:   AACGTTTCCAGTCCAAATAGCTAGGC
```

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

Score = 18 ×1 + 5 ×(-2) + 1 × (– 2) + 3 × (– 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)

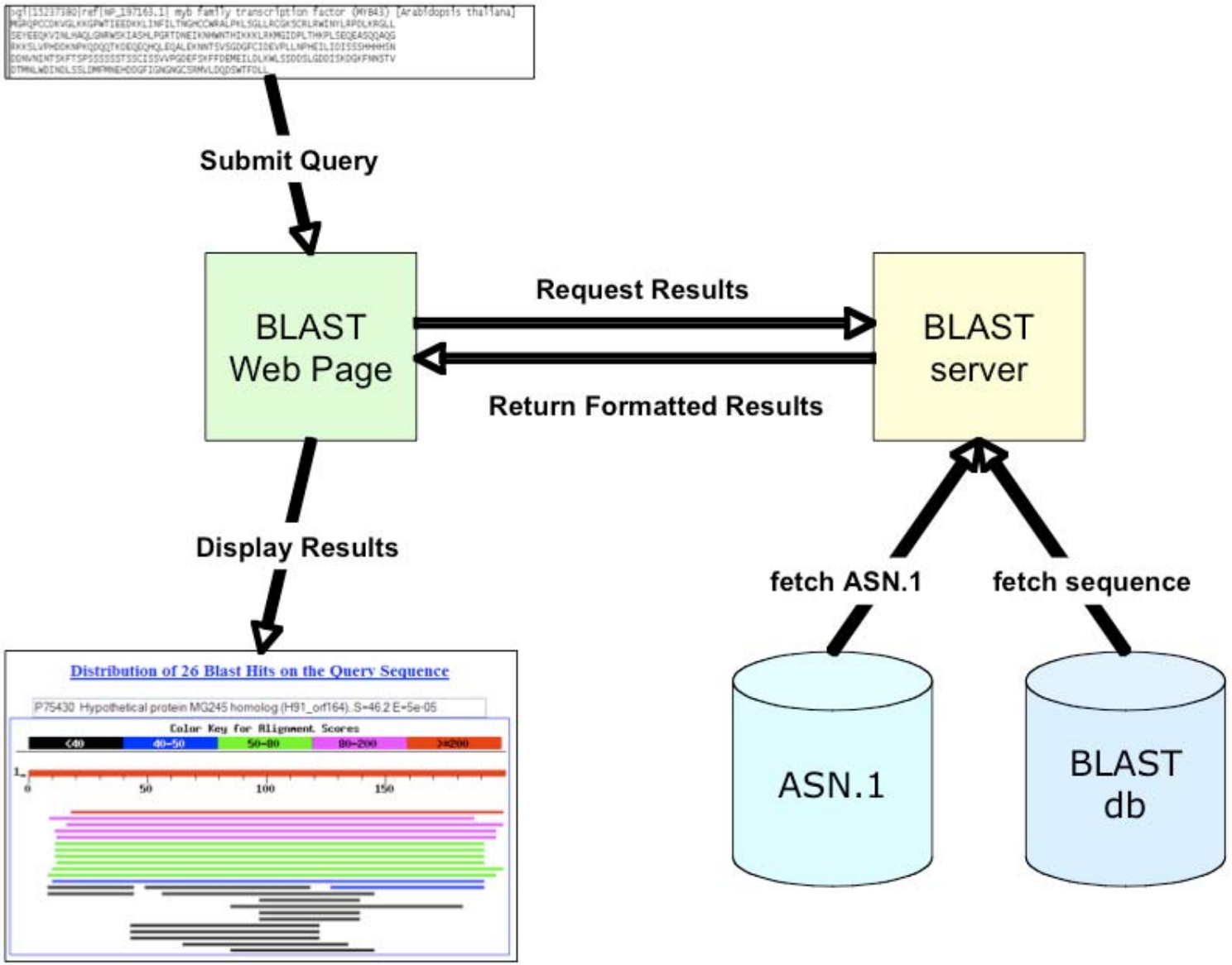
The BLAST algorithm

- 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?

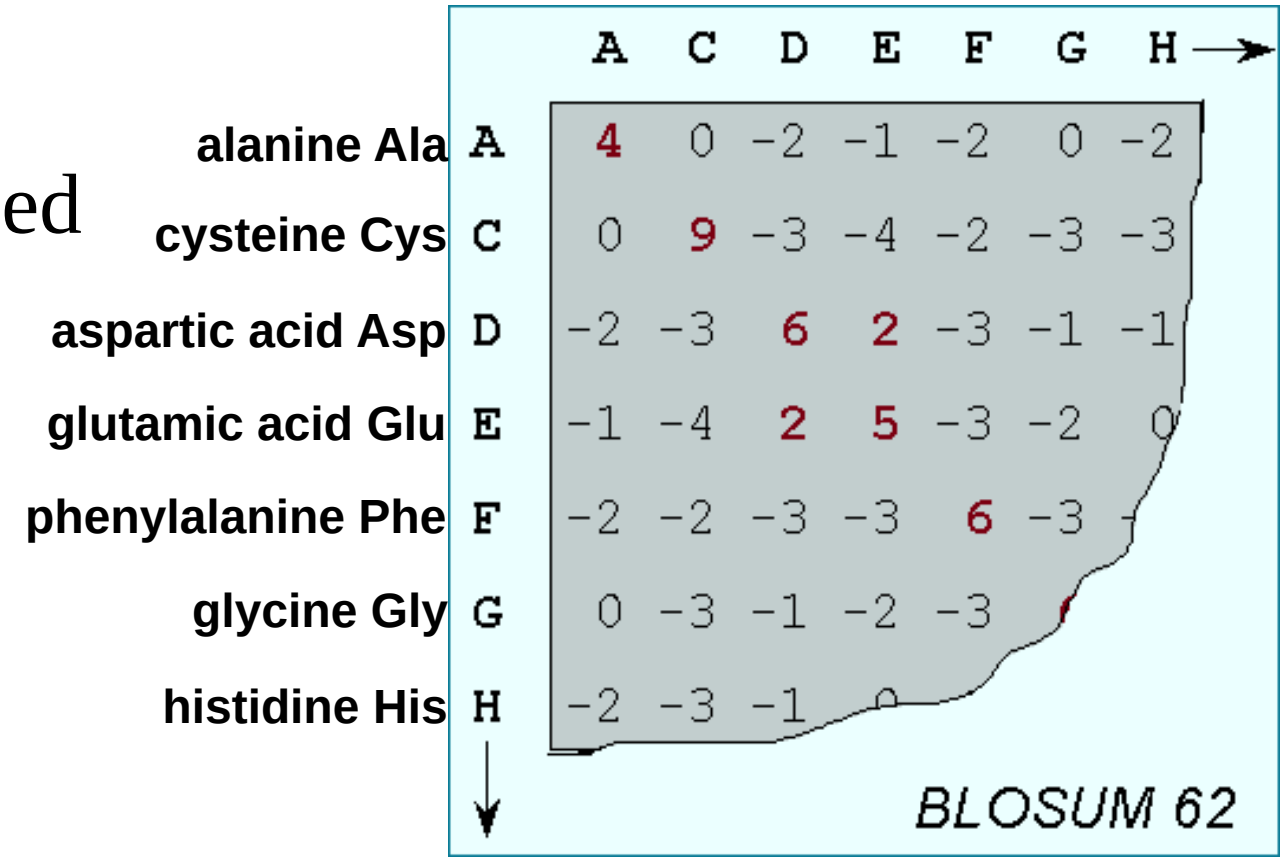
- The quality of each pair-wise alignment is represented as a score (S) and the scores are ranked.

- 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.

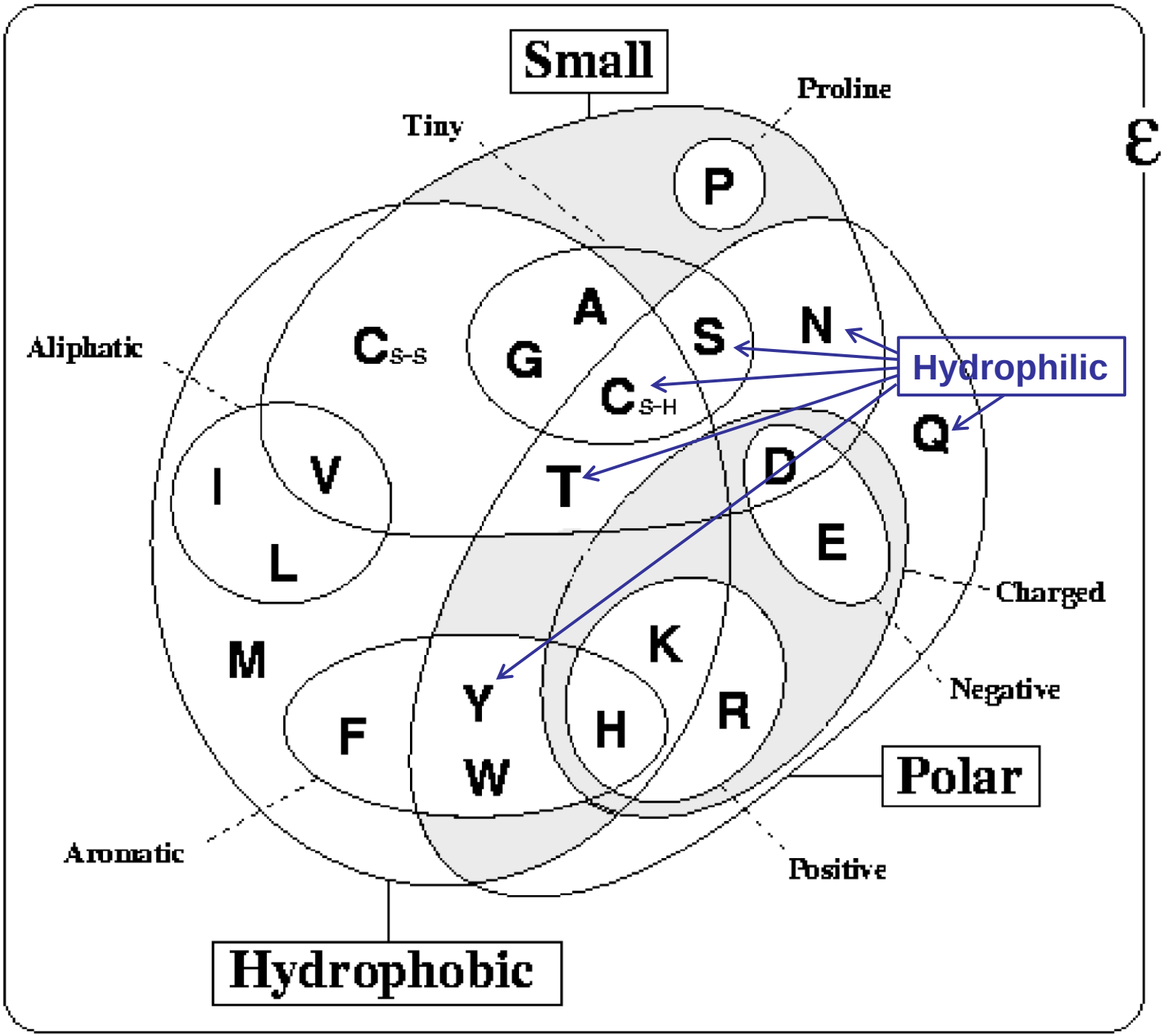
What is a scoring matrix?

- Substitution matrices are used for amino acid alignments.
  - each possible residue substitution is given a score

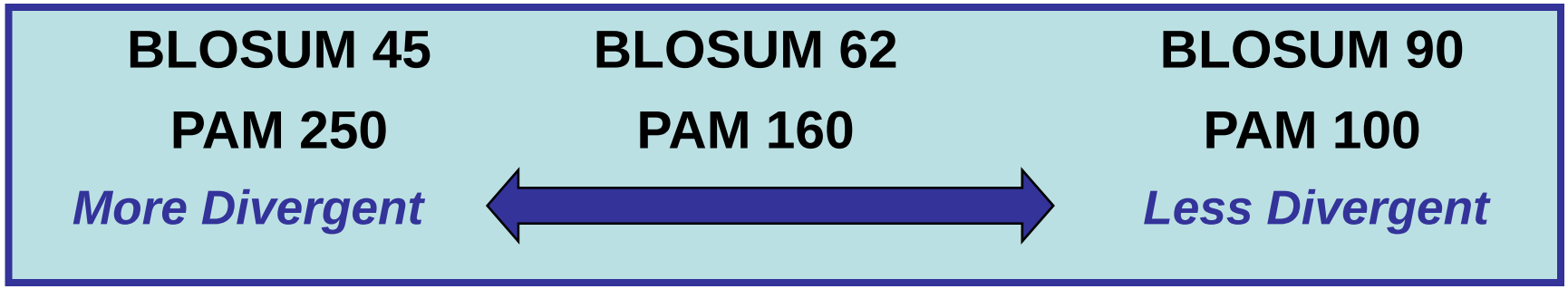


***BLOSUM - BLO**cks **SU**bstitution **M**atrix*
(Henikoff & Henikoff 1992)

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


Small
Proline
Tiny
ε
P
A
G
S
N
Aliphatic
Cs-s
Cs-H
Hydrophilic
Q
D
T
I
V
L
E
Charged
K
M
Y
R
Negative
F
H
W
Polar
Aromatic
Positive
Hydrophobic

BLOSUM vs. PAM



- **BLOSUM 62 (BLO**cks **SU**bstitution **M**atrix) 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 (P**oint **A**ccepted **M**utation**).**

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

```
>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    IVSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKI  61
            +VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKI
Sbjct  8    MVSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKI  67

Query  62   YNLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFKQN  101
            YNLCAERHYD AKFNCRVAQYPFEDHNPPQLELIKPF ++
Sbjct  68   YNLCAERHYDAAKFNCRVAQYPFEDHNPPQLELIKPFCED  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
            KQNKM+KKDKMFHFWVNTFFIPGPEE                                 +
Sbjct  260  KQNKMMKKDKMFHFWVNTFFIPGPEESRDKLENGAVNNADSQQGVPAPGQGQPQSAECRE  319

Query  127  NDKEYLVLTLTkndldkankdkanRYFSPNFKVKLYFTKTVEE  169
            +D++YL+LTL+KND DKANKDKANRYFSPNFKVKL F+KTVEE
Sbjct  320  SDRDYLILTLSKNDRDKANKDKANRYFSPNFKVKLCFSKTVEE  362


>gb|AAH93110.1|  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    VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKIY  62
            VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHK+HYKIY
Sbjct  9    VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKDHYKIY  68

Query  63   NLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFKQN  101
            NLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPF ++
Sbjct  69   NLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFCED  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

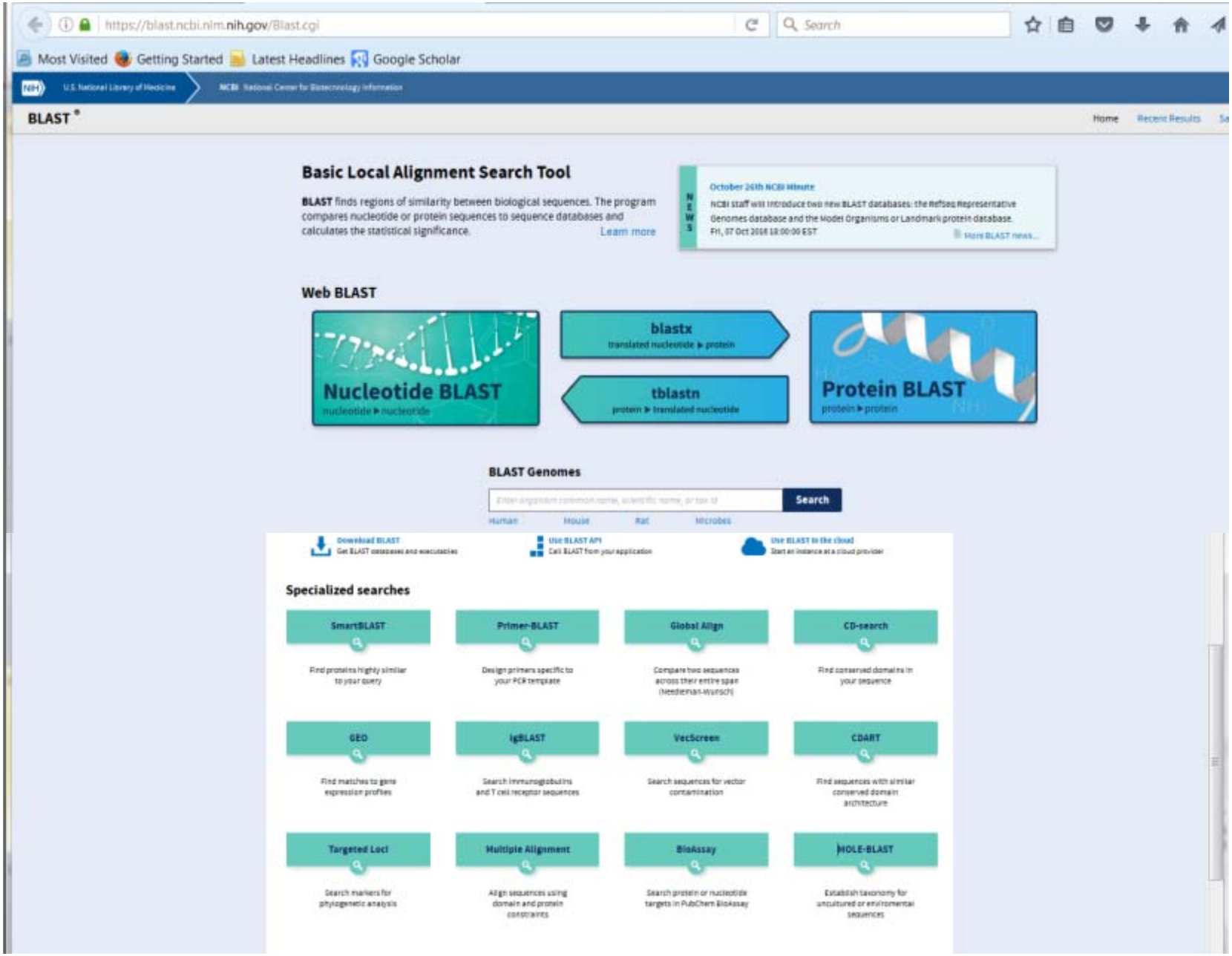
Other BLAST tools and services

- **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

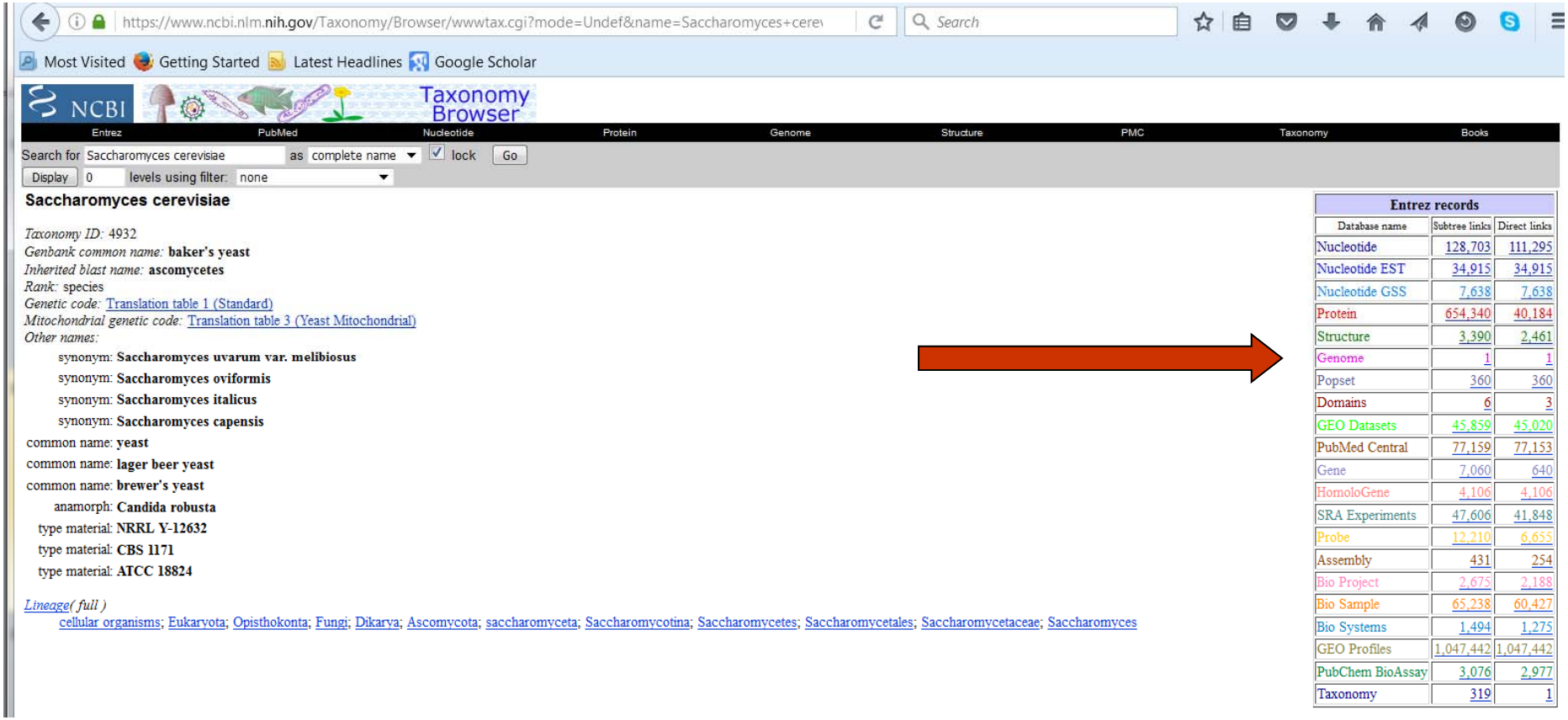
BLAST Access

- **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

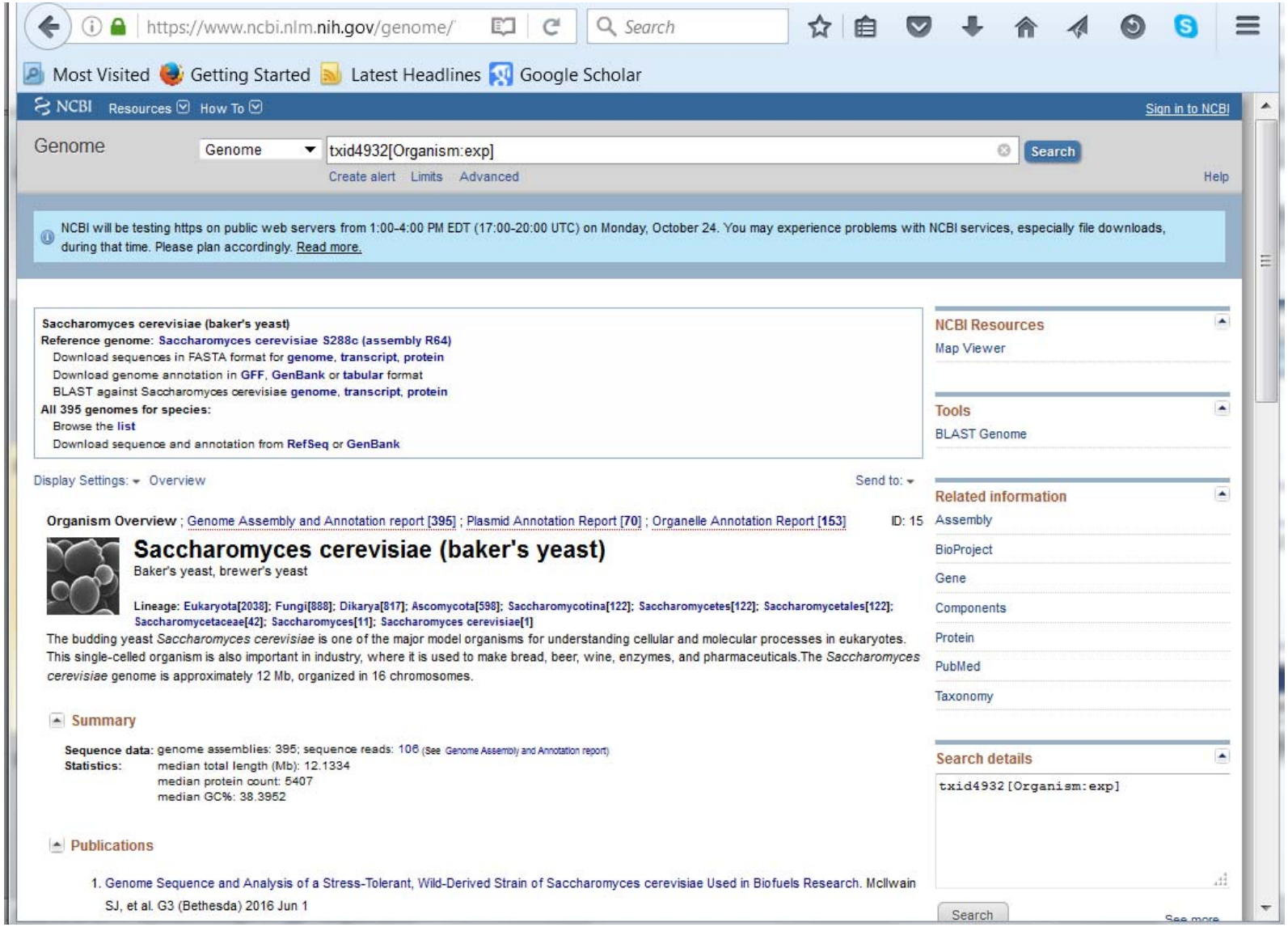
BLAST at the NCBI website



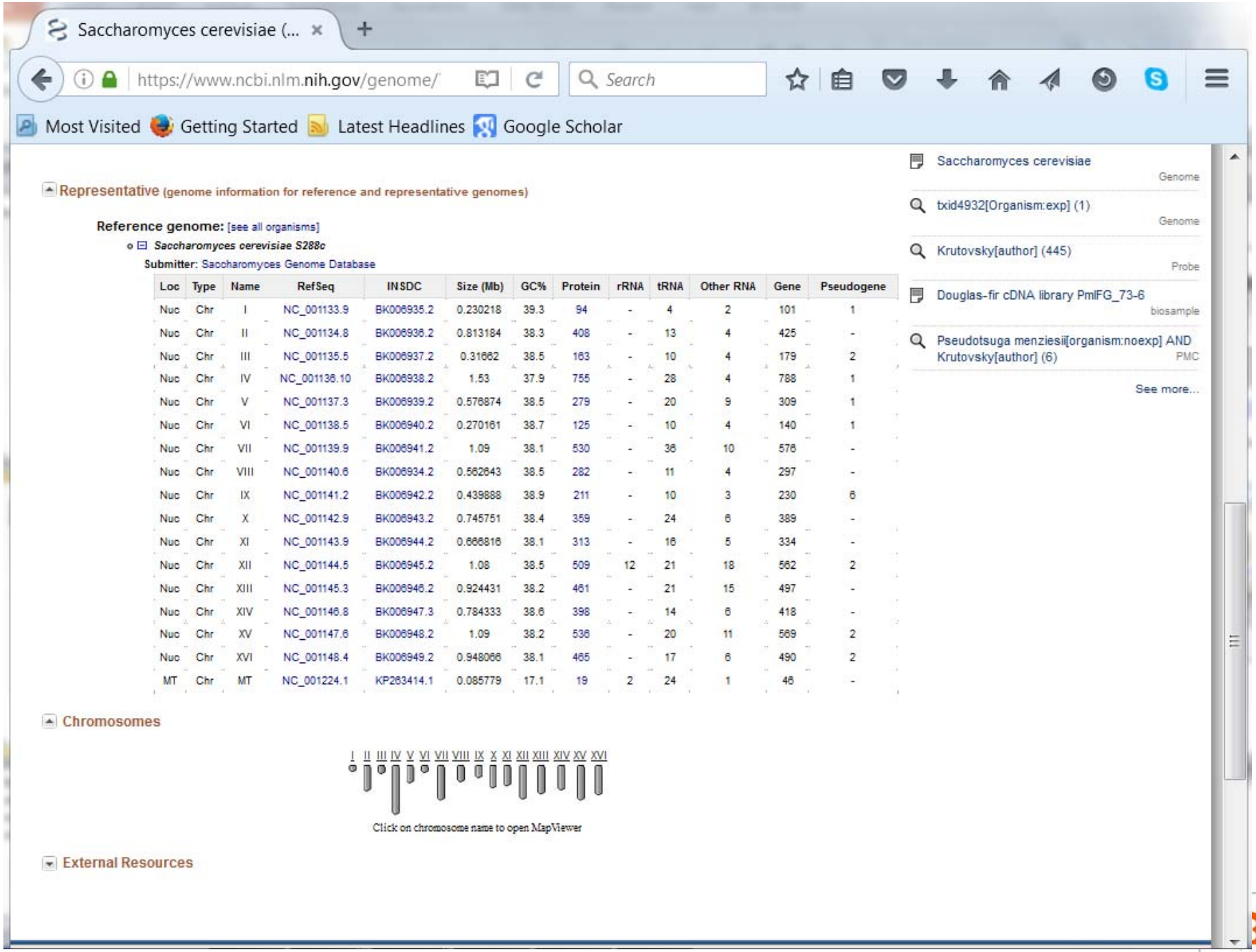
To learn about a genome of interest, visit NCBI → Taxonomy → TaxBrowser → Genome



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



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

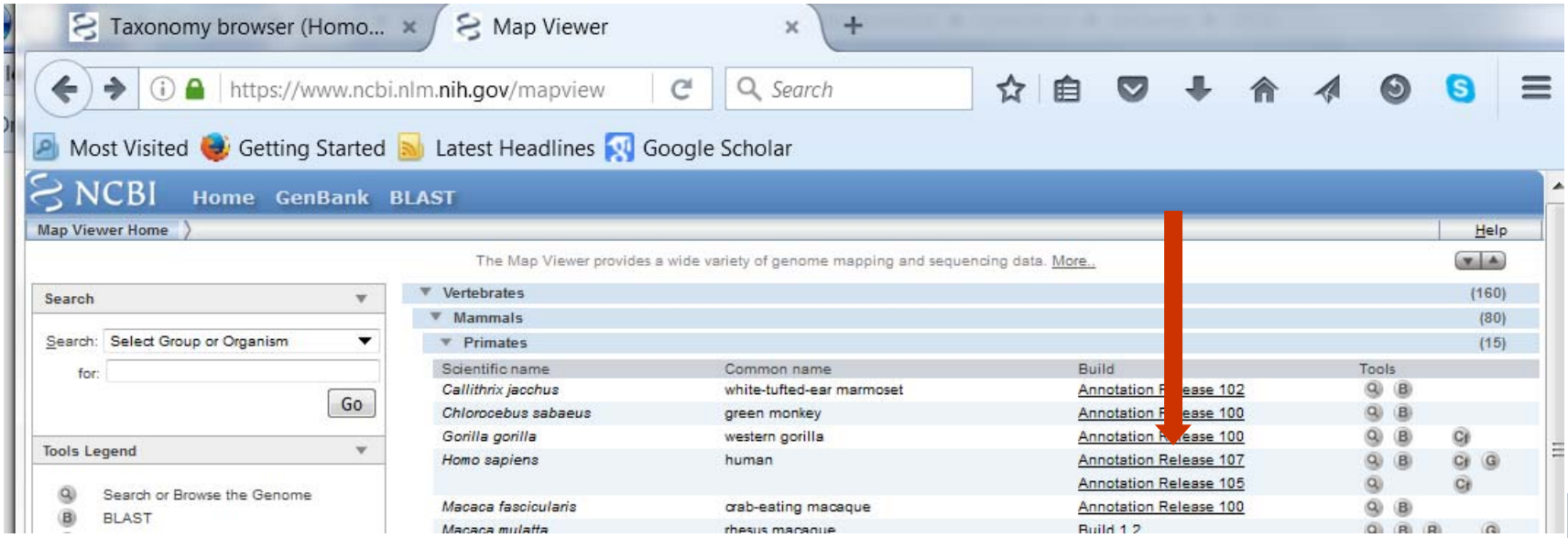

Saccharomyces cerevisiae (...
https://www.ncbi.nlm.nih.gov/genome/
Search
Most Visited
Getting Started
Latest Headlines
Google Scholar
Representative (genome information for reference and representative genomes)
Reference genome: [see all organisms]
Saccharomyces cerevisiae S288c
Submitter: Saccharomyces 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 KP263414.1 0.085779 17.1 19 2 24 1 46 -
Chromosomes
I II III IV V VI VII VIII IX X XI XII XIII XIV XV XVI
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

Search

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

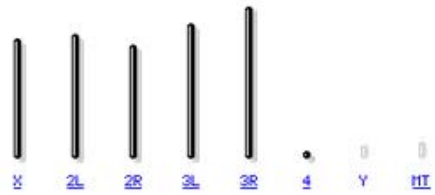
Organism Data in GenBank

- EST
- Genomic
- mRNA
- Protein
- WGS

Drosophila melanogaster (fruit fly) genome view

Release 5.30 statistics

BLAST search *Drosophila melanogaster* sequences



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