ГЕНОМИКА

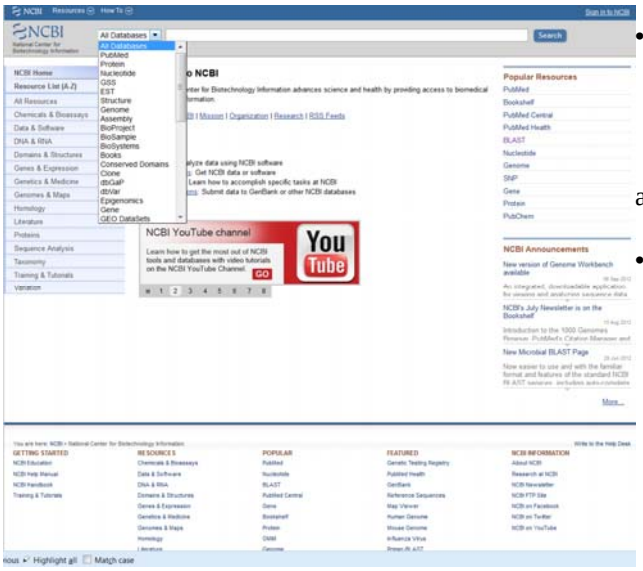
4 апреля 2017, Вторник:

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

5 апреля 2017, Среда:

- 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

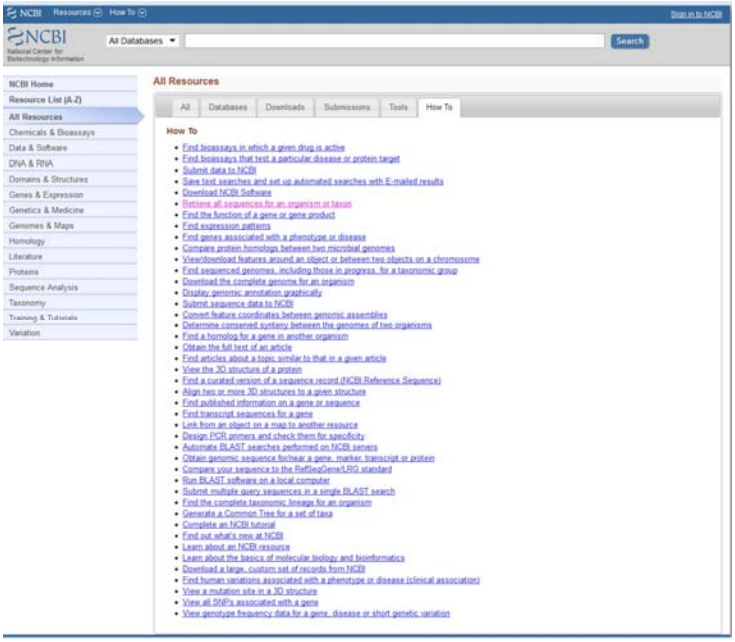
- Narrow your search using particular fields: [definition], [organism], [author], [subtree], [lineage], etc. (otherwise, it will search all fields)

and

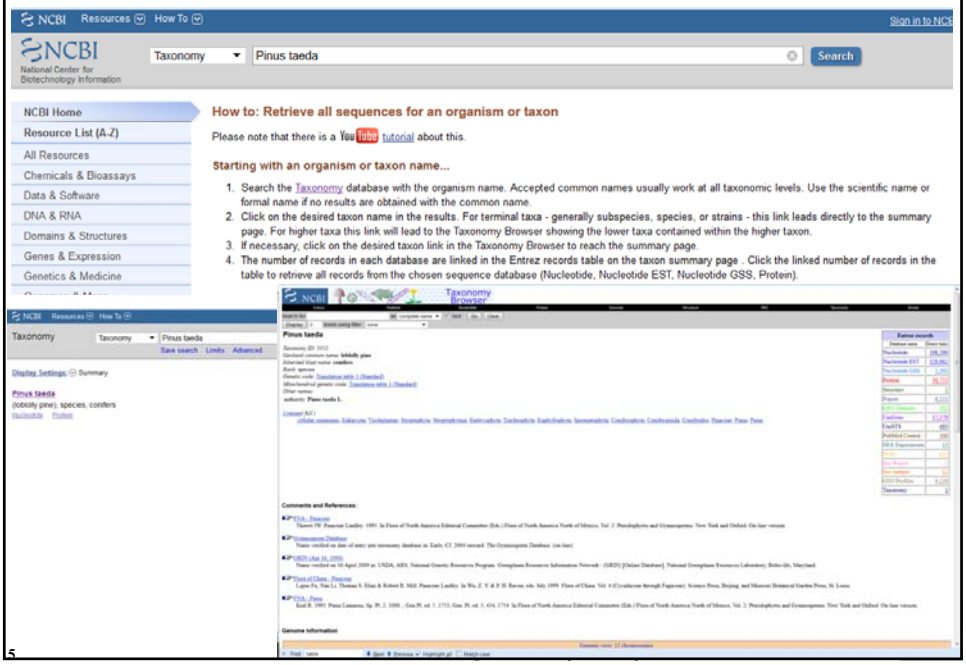
- logical Boolean operators: AND, OR, XOR, and NOT

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

NCBI resources, search and retrieval functions

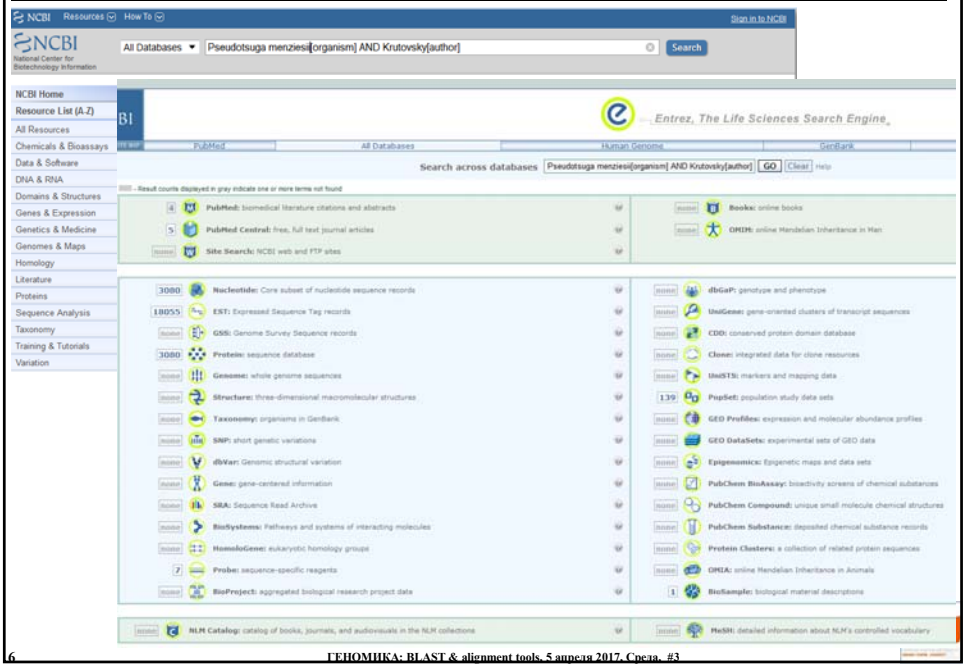


NCBI resources, search and retrieval functions



5

NCBI resources, search and retrieval functions



6

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 ttttgcaaaa cgccttcttc atccgatttc ctacgtcccg ccgttctttt gctgttttta
```

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)

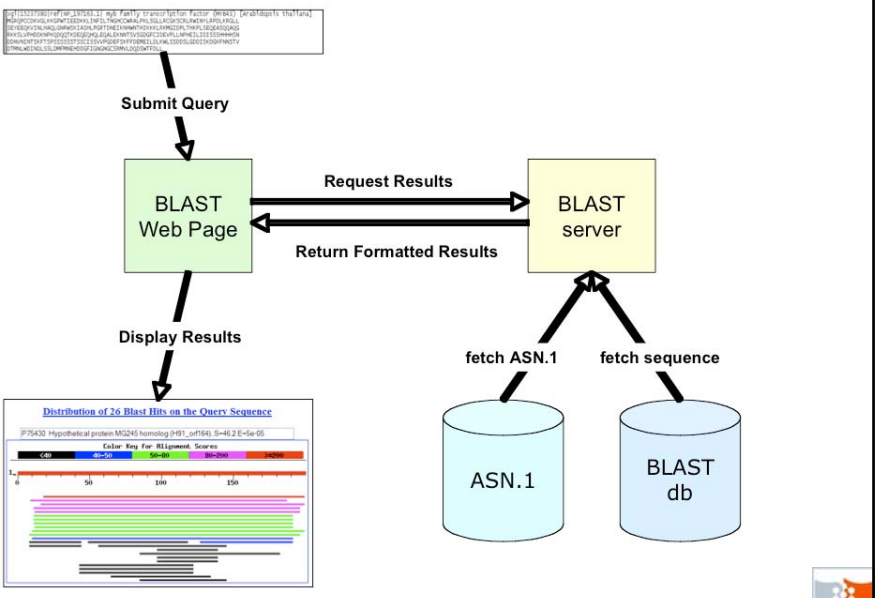
Local Alignment:

- 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

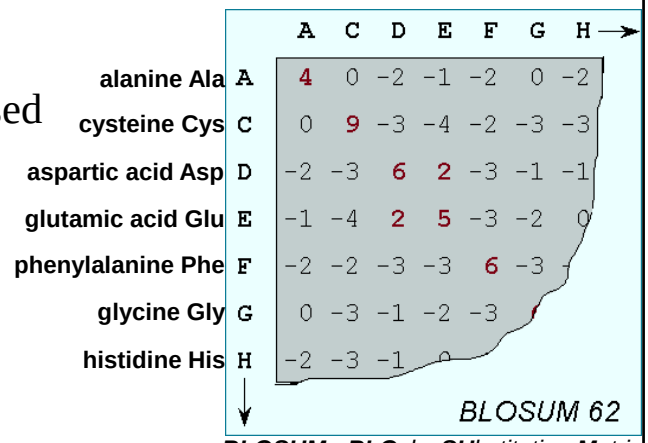
- Low E-values suggest that sequences are homologous
  - can't show non-homology
- Statistical significance depends on both the size of the alignments and the size of the sequence database
  - important consideration for comparing results across different searches
  - E-value increases as database gets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

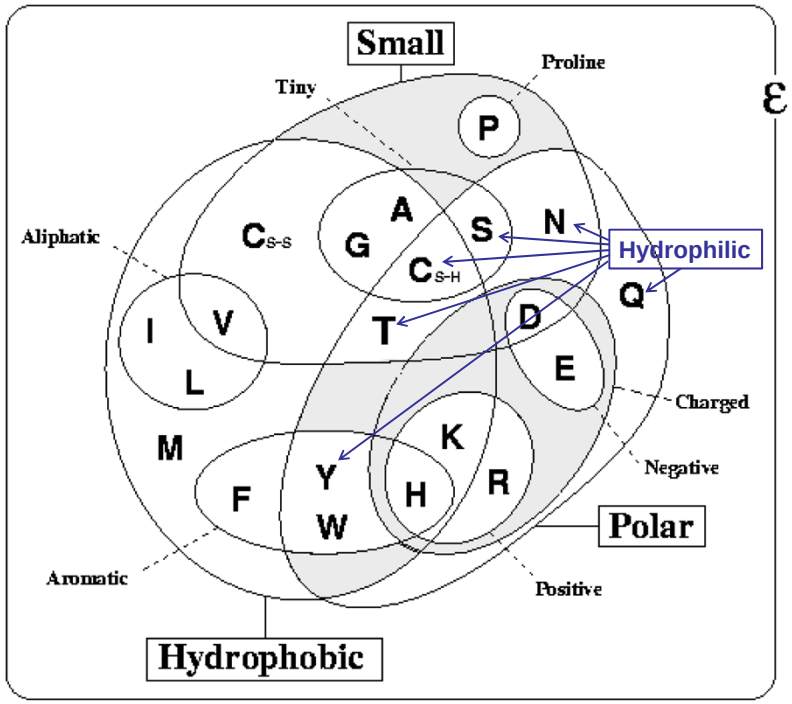
What is a scoring matrix?

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

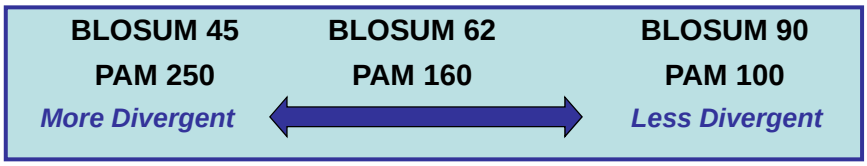


***BLOSUM** - **BLO**cks **SU**bstitution **M**atrix*
(Henikoff & Henikoff 1992)

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



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

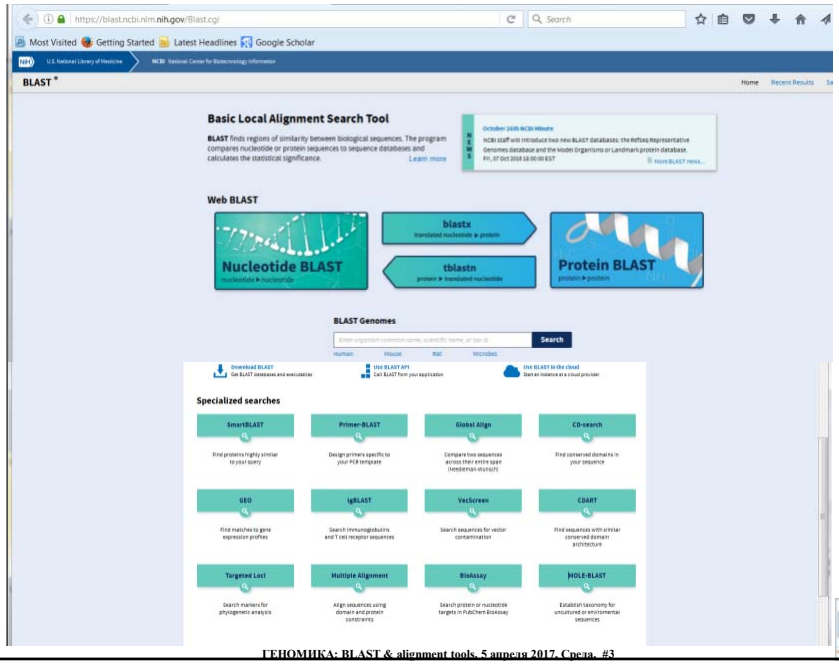
Other BLAST tools and services

- **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

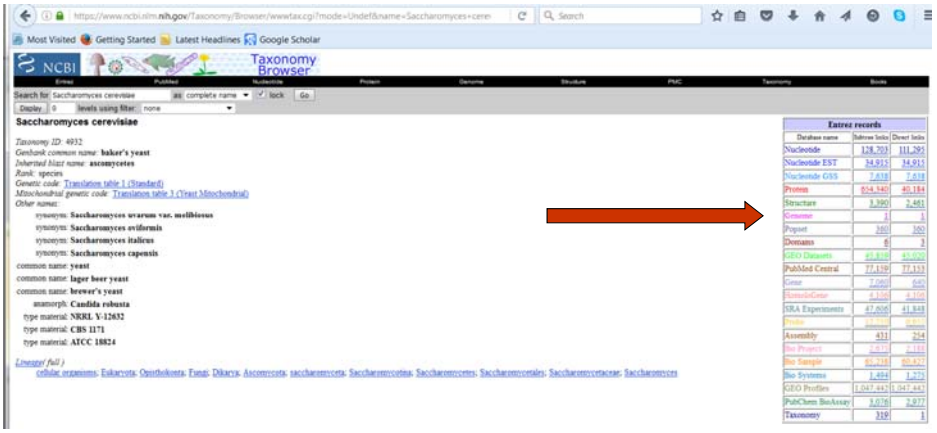
BLAST Access

- **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

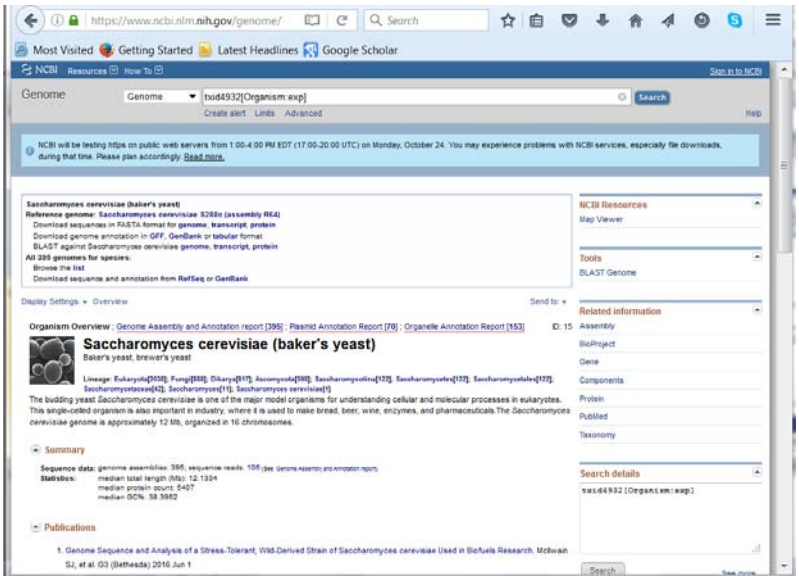
BLAST at the NCBI website



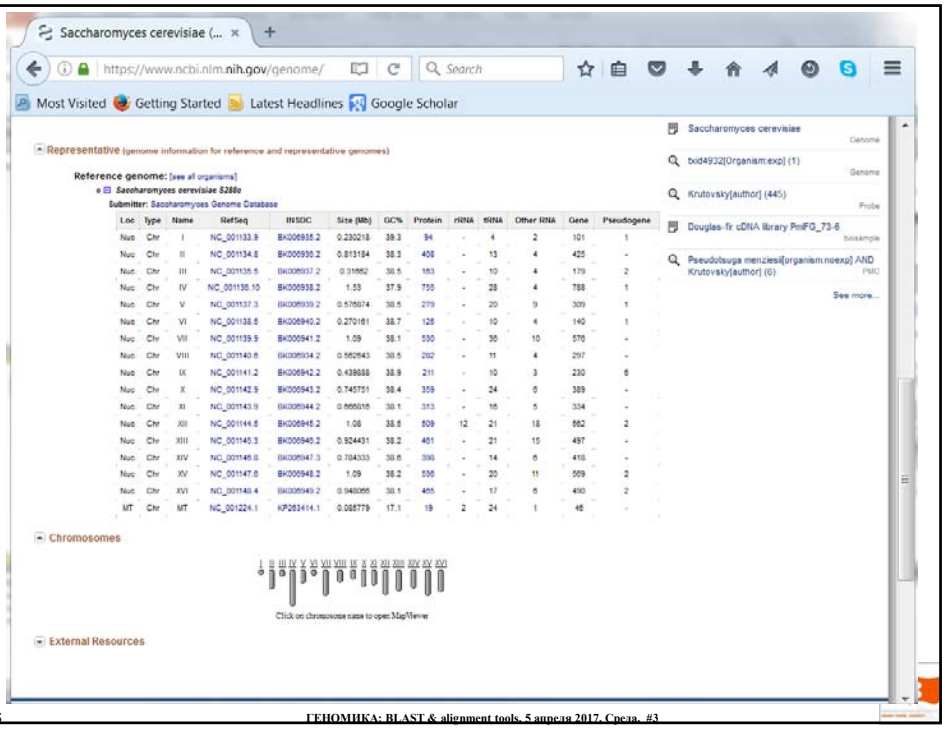
To learn about a genome of interest, visit
NCBI → Taxonomy → TaxBrowser → Genome



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



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

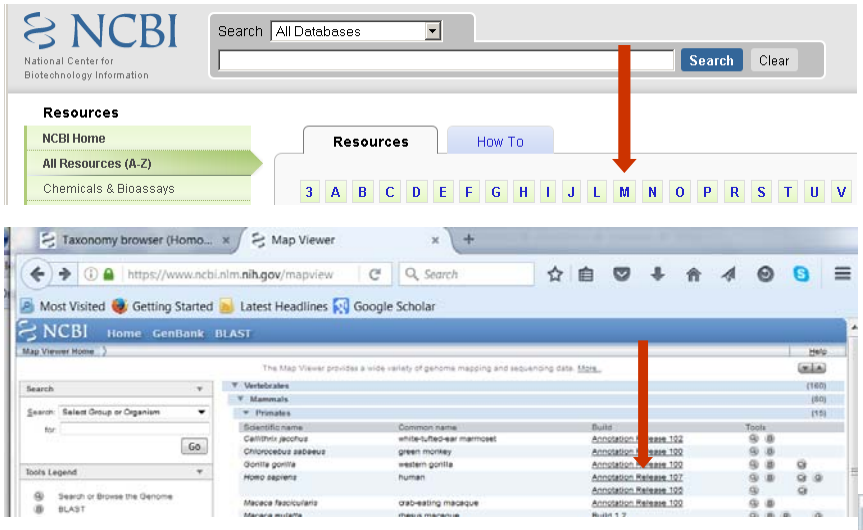


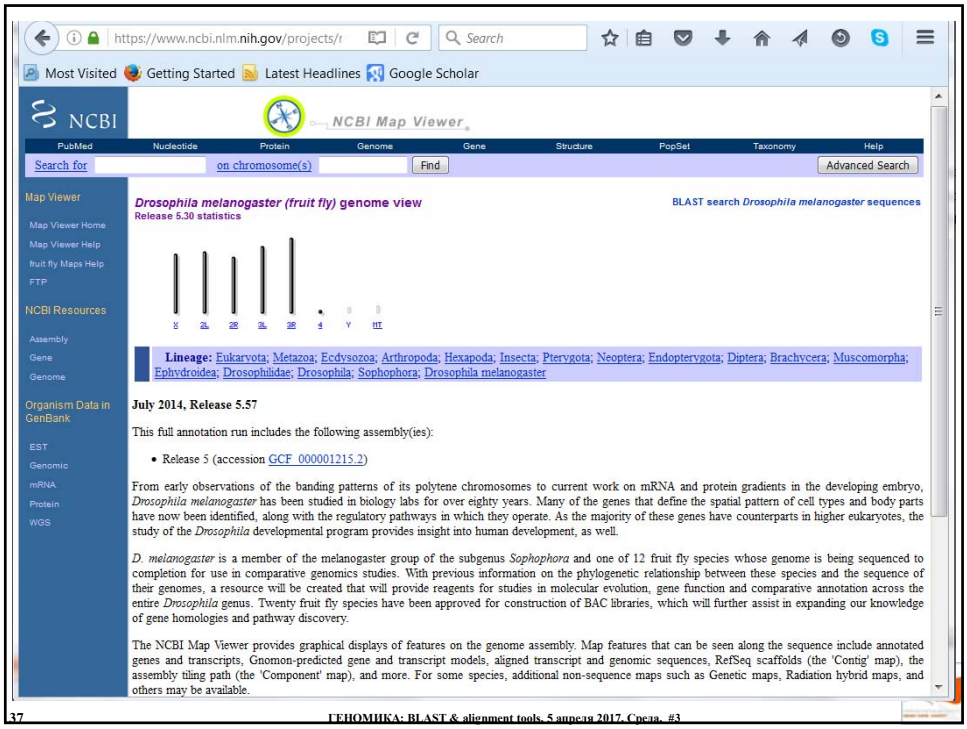
To explore human chromosome 21 at NCBI

→Find MapViewer

→Choose Primates

→Click Annotation Release 107 and, then, chromosome 21





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