NCBI Public Services

A Practical Guide to NCBI BLAST

NCBI Search Services and Tools

NCBI Public Services

- Entrez integrated literature and molecular databases
  - Viewers
    - BLink protein similarities
    - Graphical Sequence Viewer annotation viewer and analysis tool
- BLAST sequence similarity search service
- VAST structure similarity searches
- Tools, special services, standalone software
  - Entrez Utilities Entrez API <ncbi>/books/NBK25501/
  - Standalone BLAST BLAST programs + databases <ncbi>/books/NBK1762/
  - Cn3D 3D structure viewer <ncbi>/Structure/CN3D/cn3d.shtml
  - Genome Workbench sequence analysis / annotation platform <ncbi>/tools/gbench/
  - SRA Utilities <ncbi>/Traces/sra/
    - SRA Run Browser web access
    - SRA toolkit standalone SRA manipulator and client

Topics

- Basics of using NCBI BLAST
  - Motivation, Statistics, Scoring, Search Programs
- Using the Web Interface
- Other Web services
  - COBALT – protein multiple alignment
  - Primer BLAST
  - MOLE-BLAST
- Live Searches

NCBI Public Services

What is BLAST?

- Widely used sequence similarity search tool
- Finds high scoring local alignments between two sequences (protein or DNA)
- Includes a model of score distributions for random local alignments
- Provides statistical significance for alignments

NCBI Public Services

BLAST Fundamentals

NCBI Public Services

- BLAST tells you about non-chance similarities between biological sequences.
- If similarities are not due chance then they must be due to something else!
  - Homology
  - Simple identification
- All BLAST searches begin with a sequence
  - protein or nucleotide
  - experimentally determined or one from database

What BLAST tells you

NCBI Public Services

Here's my sequence.

1. What is it related to? (What does it do?)
   - Homology
   - Function
2. Is it already in the database? (Identification)
   - find the matching sequence in the database
   - organism of origin
3. Where is it located or how is it organized?
   - in a genome
   - other annotation problems
     - comparing sequences
     - looking for frame shifts

BLAST Statistics

NCBI Public Services

```
 Score = 18.5 bits (36),  Expect = 47992
 Identities = 5/5 (100%), Positives = 5/5 (100%), Gaps = 0/5 (0%)

Query  1   ELVIS  5
           ELVIS
Sbjct  8   ELVIS  12
```

- Number of chance alignments = 48 thousand!
- Indistinguishable from chance

The most important statistic: Expect value (e-value)
Expected number of random alignments with a particular score or better

```
 Score = 89.7 bits (204),  Expect = 7e-18
 Identities = 50/103 (49%), Positives = 54/103 (52%), Gaps = 18/103 (17%)

Query  1   MKLLAATVL---LLTICSLEGALVR
           MK L   VL   LL +CSLEGA V
Sbjct  1   MKVL---VLAMVLLCVCSLEGAVVM

Query  54  SPELQAEAKSYFEKSKEQLTPLIKKAGTELVNFLSYFVELGTQ  96
             E   +AK Y E   EQ  P  K   TE       F +L TQ
Sbjct  57  AEEIKTQAKAYLEQANEQFSPIAKRLHTE-------FMDLLTQ  92
```

- Number of chance alignments = 7 X 10^{-18}
- Not due to chance

- The e-value depends directly on the size of the search space (database)
- Search the smallest database likely to contain the sequence of interest

Scoring: Nucleotide

NCBI Public Services

Number of Chance Alignments = 2 X 10^{-73}

```
 Score =  288 bits (318),  Expect = 2e-73
 Identities = 262/325 (81%), Gaps = 8/325 (2%)
 Strand=Plus/Plus

Query  1923   TCAGCCTACCATGAGAATAAGAGAAAGA-AAATGAAGATCAAAAGCTTATTCATCTGTTT  1981
Sbjct  33774  TCAGACTACCCTGAGAATAAGAGAAAGAGAAATGAAGACCTAGA-CTTATCCATCTCTTT  33832
Query  1982   TTGTTTTTCGTTGGTGTAAAGCCAACACCCTGTCTAAAAAACATAAATTTCTTTAATCAT  2041
Sbjct  33833  TTGTTTTTCGTTGGTGTAAAGCCAACACCCTGTCTAAAAAACATAAATTTCTTTAAATAT  33892
Query  2042   TTTGCCTCTTTTCT                                GAATCTAATAGAGTGGT  2100
Sbjct  33893  TTTGCCTCTTTTCTCTGTGCTACAATTAATAAAAAAATGAAAAGAATCTAATTTAATTGT  33952
Query  2101   ACAGCACTGTTA-TT                                            TTCTGTGG  2159
Sbjct  33953  CTATGACTGTTATTT                                            TTCTATGA  34012
Query  2160   AAGTTCCAGTGTTCT                                            GTGGGCTA  2219
Sbjct  34013  AAATTCCACTATTCTCTCTTTCCCTATTTCAATGGAGGACTTCTAGTTCCTTCTGGATTA  34072
Query  2220   AT----TAAATAAATCATTAATACT  2240
Sbjct  34073  ATTGCATAAAAGAAACATTAATACT  34097
```

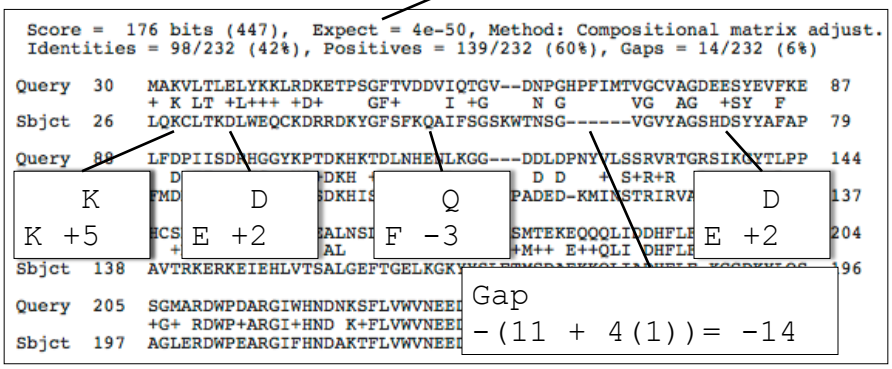
Match=+2

Mismatch=-3

Gap
-(5 + 4(2))= -13

Scoring: Protein

Number of Chance Alignments = 4 X 10^{-50}



Scores from BLOSUM62, a position independent matrix
- Same substitution gets the same score at all positions
- All positions equally likely to change

BLAST Search Programs

Nucleotide Search Programs

NCBI Public Services

- blastn
  - traditional BLAST algorithm
  - most sensitive nucleotide search
- megablast
  - larger word size than blastn
  - different gapping model
    - Contiguous megablast
      - Nearly identical sequences
    - Discontiguous megablast
      - Cross-species comparisons

- Default nucleotide search program
- Best for
  - Identification
  - Same-species annotation

06/06/2016 11

Protein Search Programs

NCBI Public Services

Position Independent scoring

- blastp
- translating searches
  - useful for unannotated protein coding regions
  - six frame translations of query, database or both
    - blastx – translated query
    - tblastn – translated database
    - tblastx – translated query and database

06/06/2016 12

Protein Domains and Position Specific Scoring

NCBI Public Services

Position-specific scoring model

- Multiple alignment based
- Substitution scores depend on the position in the protein.
- Some positions are more important (less likely to change)
- More accurate alignments
- More sensitive at identifying distant homologies
- Better at identifying structural / functional domain

Position-specific Programs
(protein only)

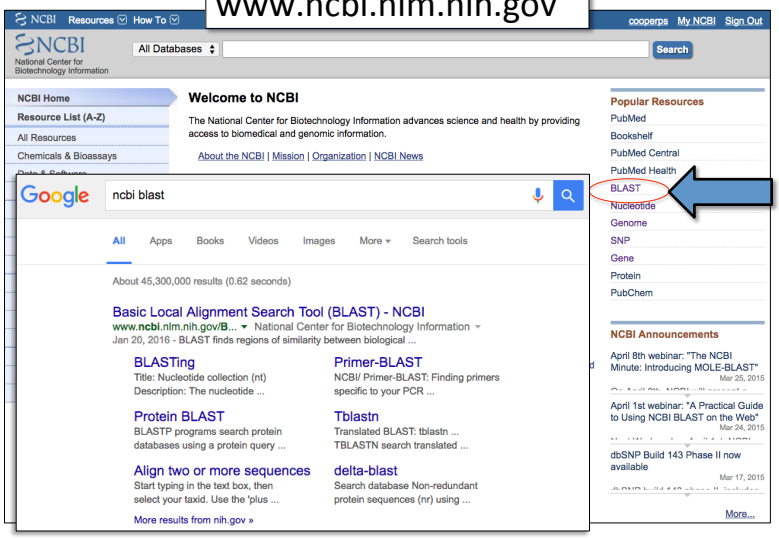
NCBI Public Services

- **Position Specific Iterative BLAST (PSI-BLAST)**
  Automatically generates a position specific score matrix (PSSM) from initial set of BLAST alignments
- **Position-Hit Initiated BLAST (PHI-BLAST)**
  Focuses search around pattern (motif)
- **Domain Enhanced Lookup Time Accelerated (DELTA) BLAST**
  Uses conserved domain PSSM in first round of search
- **Reverse PSI-BLAST (RPS-BLAST)**
  Searches a database of PSI-BLAST PSSMs
  Conserved Domain Database Search
  Quickly identifies type of protein and potential function

  - Runs with all blastp searches at the NCBI
  - Identifies conserved domains in query

Web Access

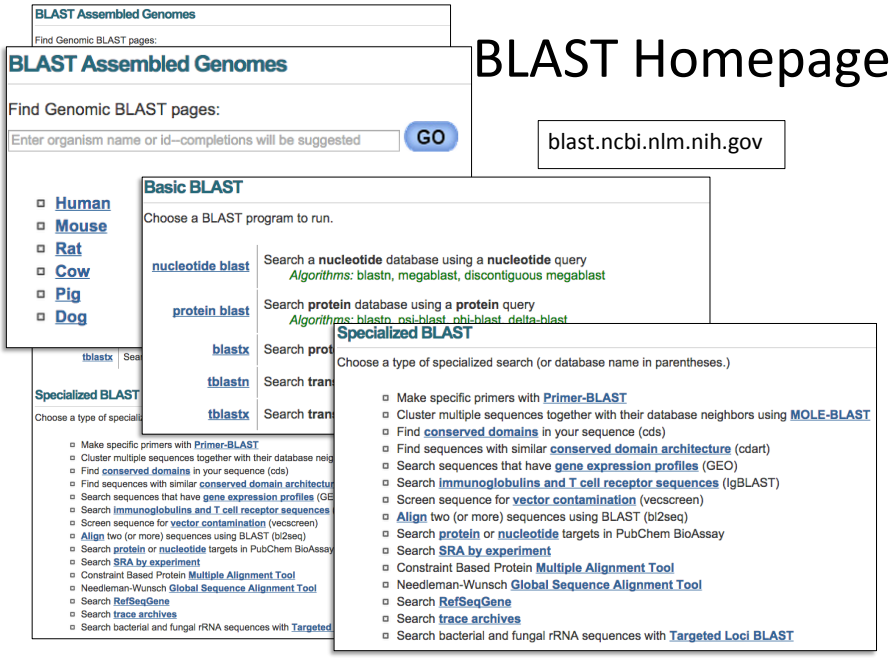
www.ncbi.nlm.nih.gov



NCBI Public Services

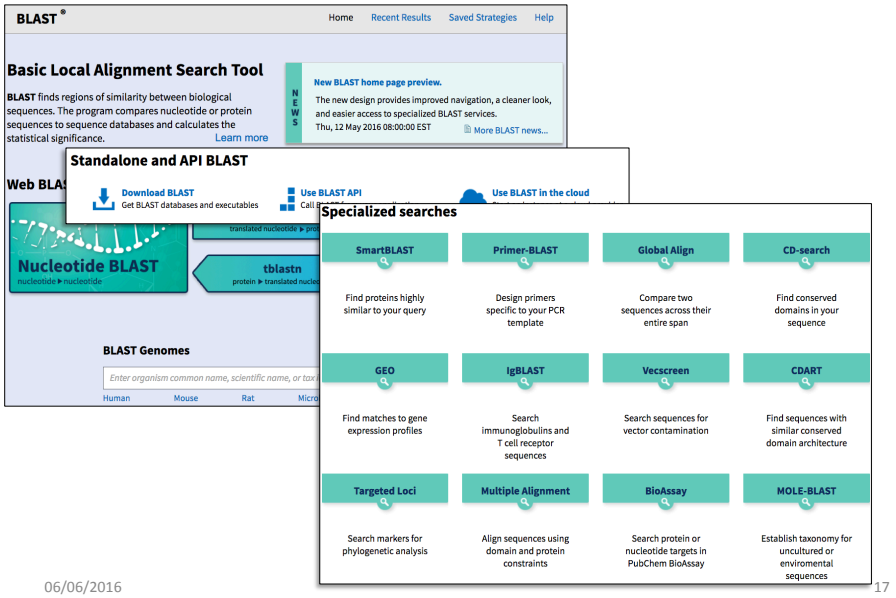
BLAST Homepage

blast.ncbi.nlm.nih.gov



NCBI Public Services

The New BLAST Homepage

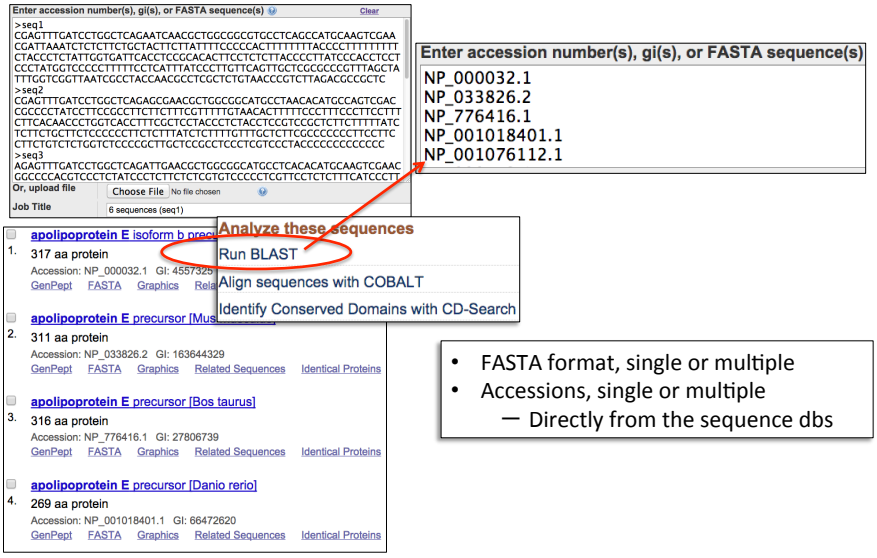


NCBI Public Services

Query Sequences

NCBI Public Services

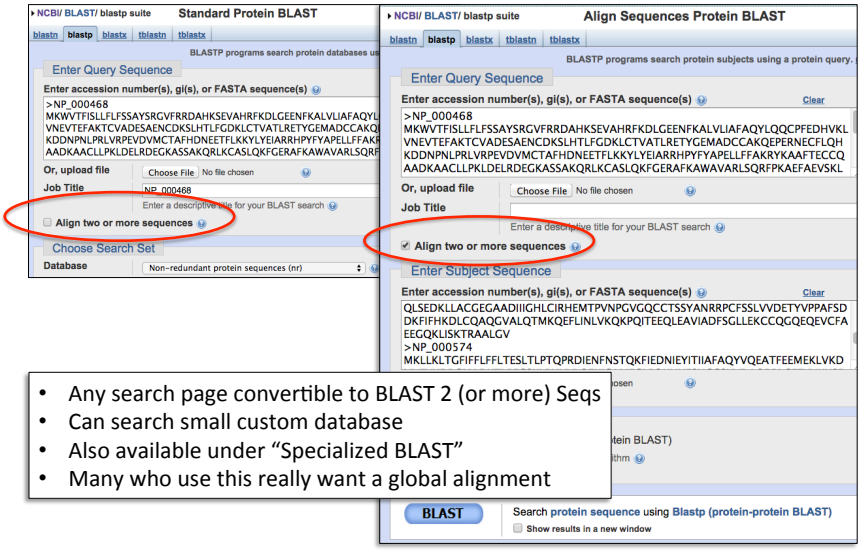
Queries



- FASTA format, single or multiple
- Accessions, single or multiple
  - Directly from the sequence dbs

NCBI Public Services

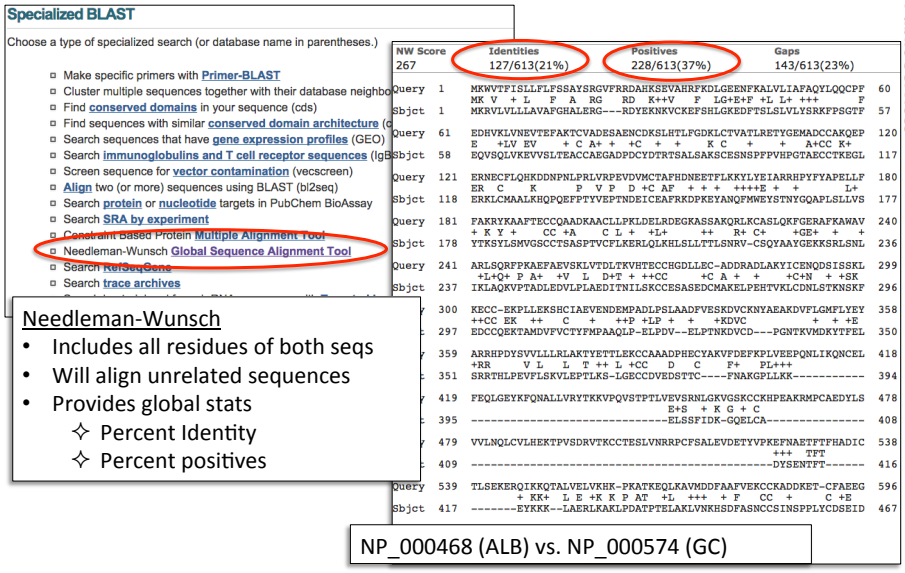
BLAST 2 (or more) Sequences



- Any search page convertible to BLAST 2 (or more) Seqs
- Can search small custom database
- Also available under “Specialized BLAST”
- Many who use this really want a global alignment

NCBI Public Services

Global Alignment Tool

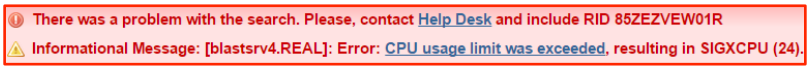


Needleman-Wunsch
- Includes all residues of both seqs
- Will align unrelated sequences
- Provides global stats
  - ✧ Percent Identity
  - ✧ Percent positives

NP_000468 (ALB) vs. NP_000574 (GC)

NCBI Public Services

Restrictions on Web BLAST



- No fixed limit on size / number of queries
- Web BLAST limited by processing time (one hour) for single search
  - – Difficult to estimate processing time
  - – Factors
    - Size and number of queries
    - Size of database
    - Depends on size and complexity of results too

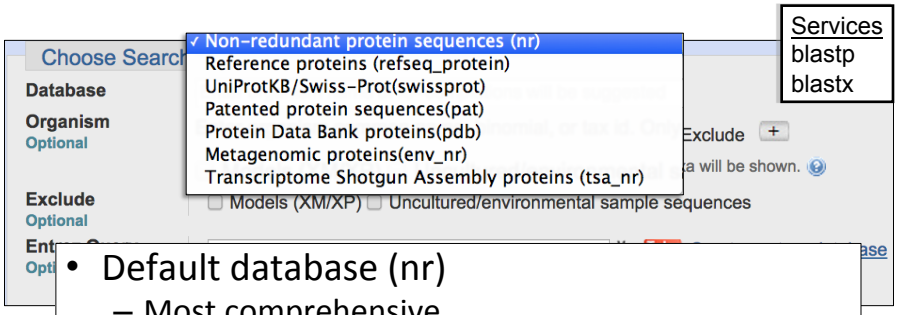
NCBI Public Services

NCBI Public Services

BLAST Databases

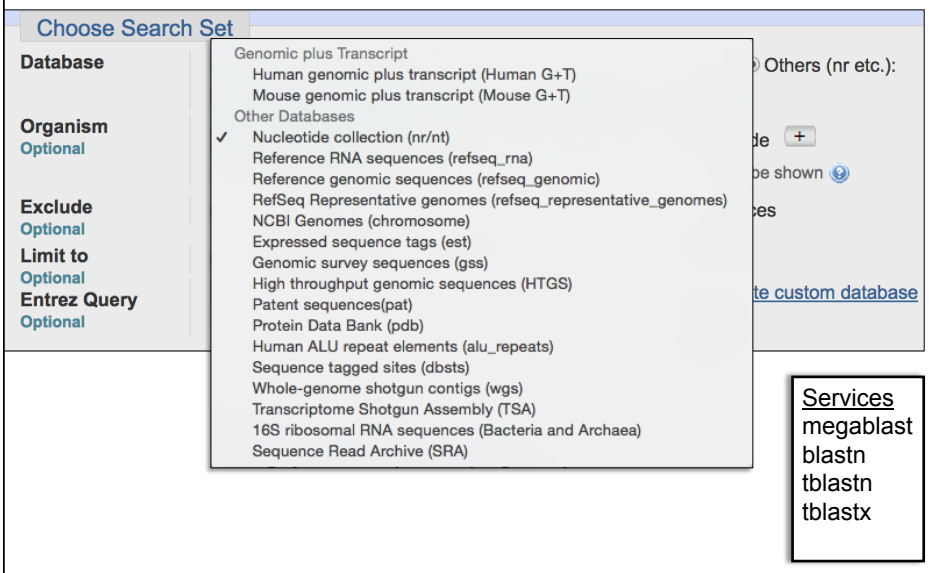
Protein Databases

NCBI Public Services



- Default database (nr)
  - Most comprehensive
  - Useful subsets: RefSeq, Swiss-Prot, PDB
- What's not in nr?
  - US , European and Asian Patents
  - Proteins from metagenomes
  - Proteins from Next-Gen assemblies

Nucleotide Databases



NCBI Public Services

Nucleotide Databases

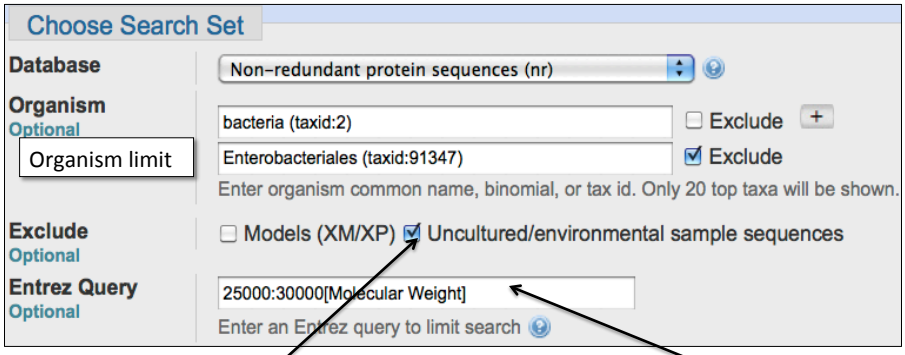
- Default database (nr/nt) is not comprehensive
  - Contains traditional GenBank and RefSeq RNA
  - Useful subsets: RefSeq RNA, 16S rRNA reference sequences
- What is not in nr/nt? The majority of nucleotide data
  - Bulk sequences (EST, GSS, HTGS, STS)
  - RefSeq Genomic Sequences (Chromosome, RefSeq Genomic, RefSeq Representative Genomes)
  - US, European and Asian Patents (pat)
  - Whole Genome Shotgun Contigs (WGS) (Second Largest)
  - Transcriptome Shotgun Assemblies (TSA)
  - Next-Gen RNA-Seq, DNA-Seq Reads (SRA) (Largest set of data)

NCBI Public Services

Limiting Databases

NCBI Public Services

Search the smallest database likely to contain the sequence of interest.



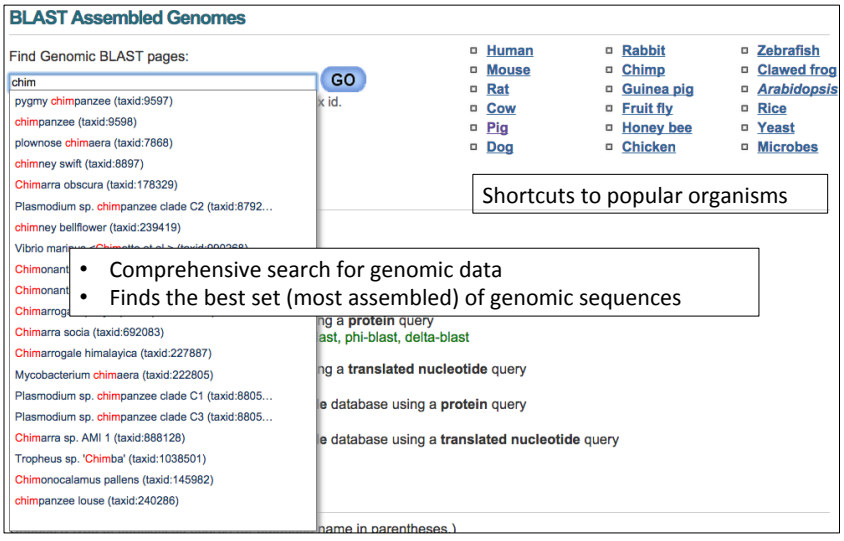
Organism limit

Exclude predicted and uncultured

Limit with Entrez query

Genome Databases

NCBI Public Services



Shortcuts to popular organisms

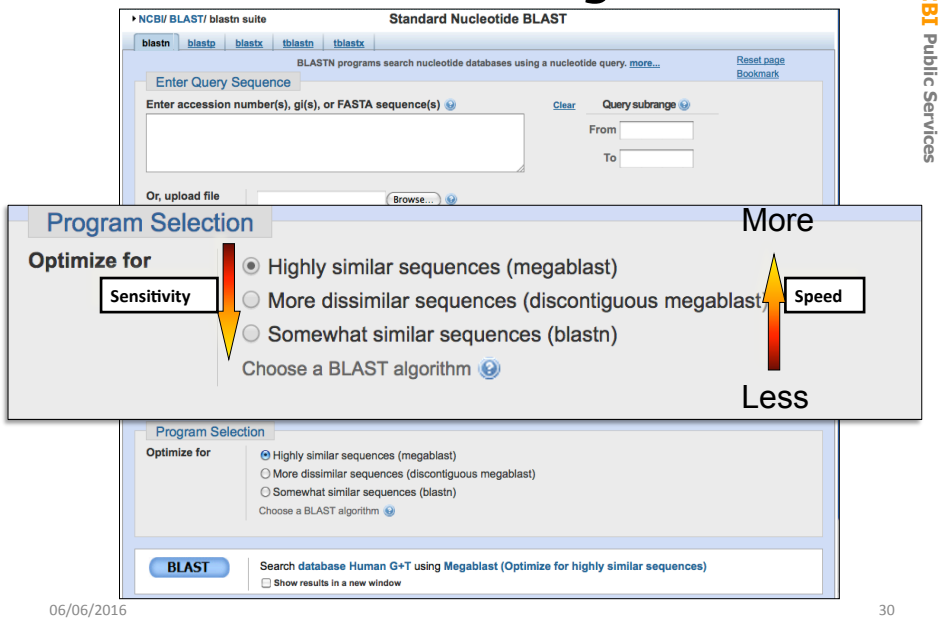
- Comprehensive search for genomic data
- Finds the best set (most assembled) of genomic sequences

NCBI Public Services

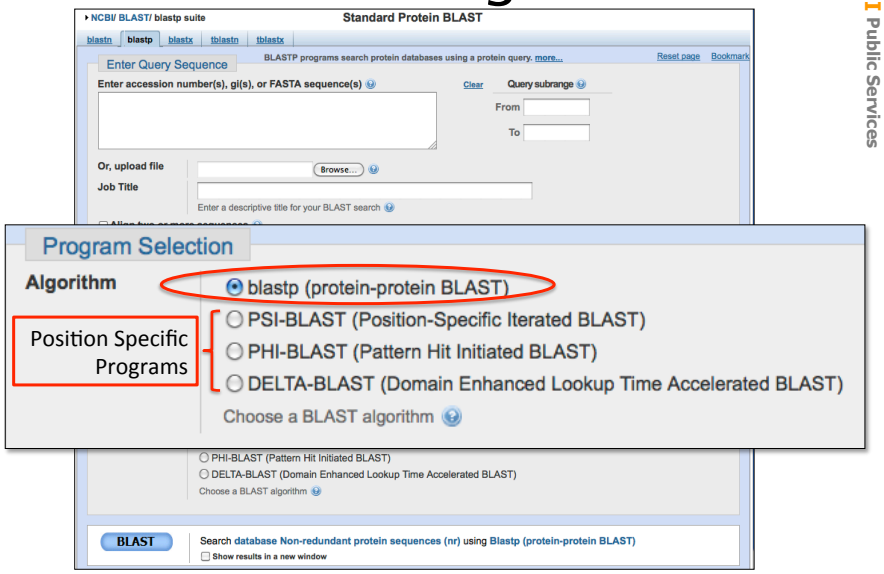
Web Program Selection

NCBI Public Services

Nucleotide Programs

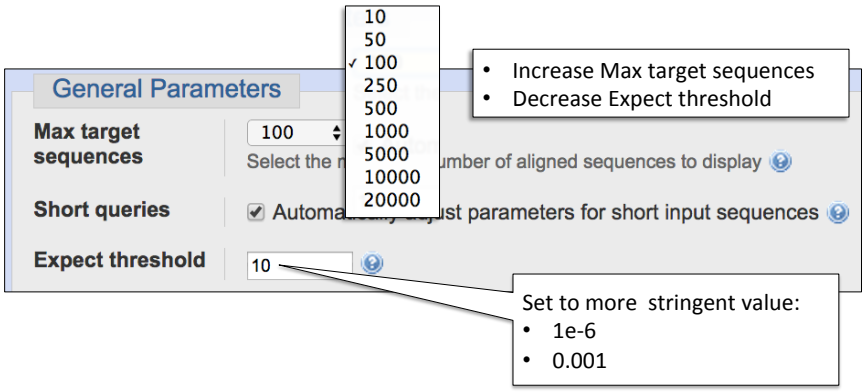


Protein Programs



NCBI Public Services

Algorithm Parameters: General



- Increase Max target sequences
- Decrease Expect threshold

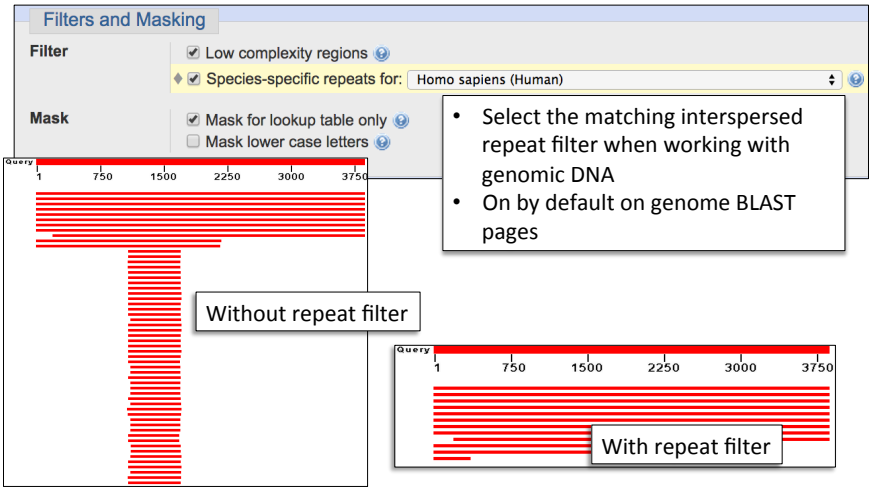
Set to more stringent value:
- 1e-6
- 0.001

Let **Expect threshold** govern output not **Max target** sequences

NCBI Public Services

Nucleotide Repeat Filters

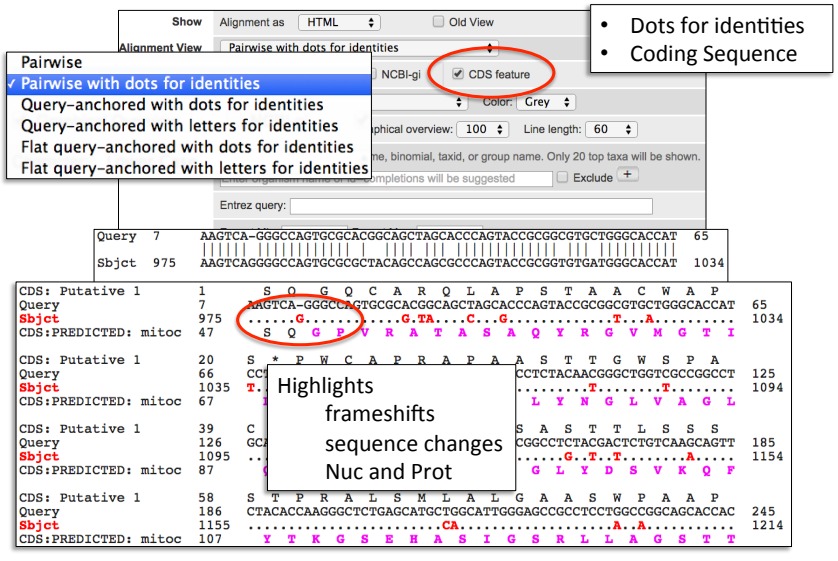
NCBI Public Services



- Select the matching interspersed repeat filter when working with genomic DNA
- On by default on genome BLAST pages

Formatting options

NCBI Public Services



- Dots for identities
- Coding Sequence

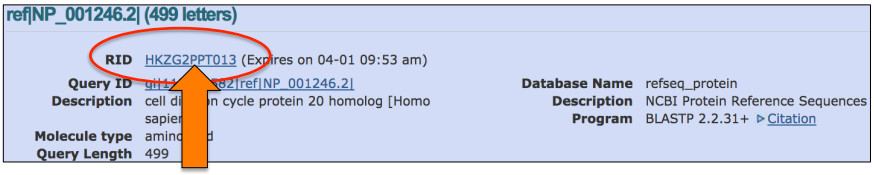
NCBI Public Services

Managing Your Results

NCBI Public Services

The Request ID (RID) is the key

The BLAST Request ID (RID) is the key

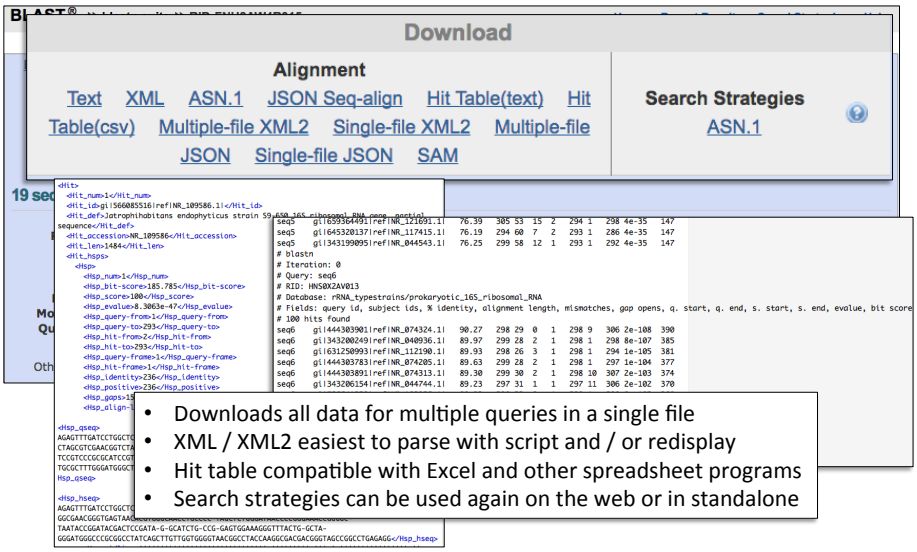


http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Get&RID=HKZG2PPT013

- Uniquely identifies search settings and results
- Persists at NCBI for 36 hours
- View through Recent Results, My NCBI
- Allows sharing results and reformatting
- Send the RID to blast-help@ncbi.nlm.nih.gov to ask about a search

Download Options

NCBI Public Services



- Downloads all data for multiple queries in a single file
- XML / XML2 easiest to parse with script and / or redisplay
- Hit table compatible with Excel and other spreadsheet programs
- Search strategies can be used again on the web or in standalone

NCBI Public Services

Specialized BLAST Services

Nucleotide Services

NCBI Public Services

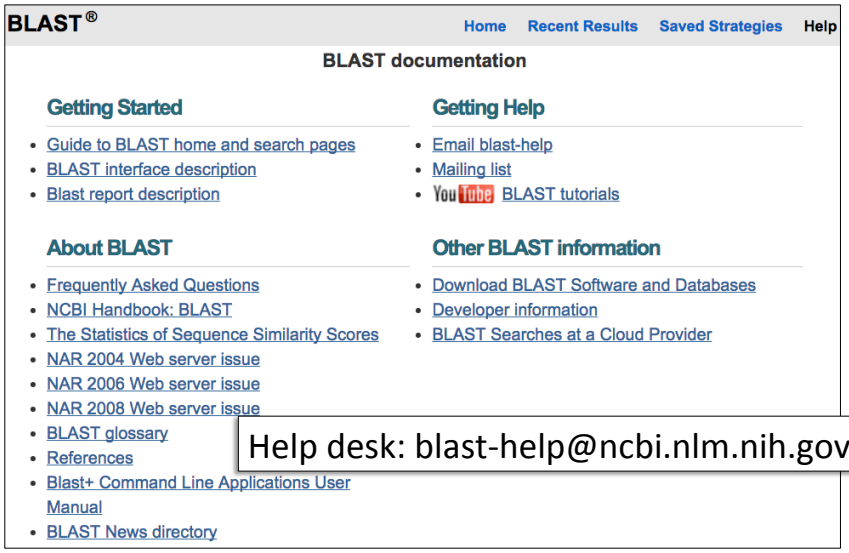
- PrimerBlast
  - primer designer / specificity checker
  - Primer3 primer design
  - Uses RefSeq annotation
    - exon boundaries
    - splice variants
    - SNPs
- MOLE-BLAST
  - Helps identify sources of 16S and other targeted sequences
  - BLAST followed by global multiple alignment
  - Clusters queries plus most similar database sequences
  - Identifies taxonomic units (neighbors)
  - Labels database sequences from type material for accurate ID

Protein Services

NCBI Public Services

- COBALT – Constraint Based Alignment Tool
  - Protein global multiple alignment tool
  - Uses conserved domains to guide alignment
  - Extension to BLAST search
- SmartBLAST – Rapid protein identification tool
  - Uses fast k-mer search
  - Identifies closest match in reference organism database
  - Produces multiple alignment and protein tree
  - Prototype for on-the-fly protein similarity (BLink)

BLAST Help



Help desk: blast-help@ncbi.nlm.nih.gov

NCBI Public Services

More Help Links

- Help Manual: <ncbi>/books/NBK3831/
- Learn: <ncbi>/home/learn.shtml
- Factsheets: <ftp>/pub/factsheets/
- NCBI YouTube: <youtube>/ncbinlm
- NCBI Helpdesks
  - General: info@ncbi.nlm.nih.gov
  - BLAST: blast-help@ncbi.nlm.nih.gov

NCBI Public Services

Web Demonstrations

NCBI Public Services

- Basic BLAST
  - blastp, MLH1
    - COBALT extension
- Genome BLAST
  - blastn, macaque CDC20
    - Formatting options
    - Genome context
- SRA BLAST
  - Melanoma gene expression
- Primer BLAST
  - BRCA1 Exon Primers
- SmartBLAST
  - Finding yeast MLH1