NCBI Public Services

NCBI Molecular Biology Resources

Part 2: Using NCBI BLAST

December 2009

NCBI Public Services

Using BLAST

- Basics of using NCBI BLAST
- Using the new Interface
  - Improved organism and filter options
- New Services
  - Primer BLAST
  - Align 2 Sequences Integration
  - COBALT – protein multiple alignment
- BLAST URL API
- C++ BLAST binaries

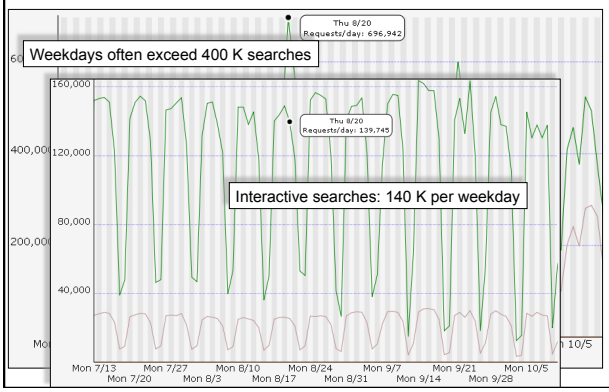
NCBI Public Services

Basic Local Alignment Search Tool

- Widely used similarity search tool
- Heuristic approach based on Smith Waterman algorithm
- Finds best local alignments
- Provides statistical significance
- All combinations (DNA/Protein) query and database.
  - DNA vs DNA
  - DNA translation vs Protein
  - Protein vs Protein
  - Protein vs DNA translation
  - DNA translation vs DNA translation
- www, standalone, and network client

NCBI Public Services

BLAST Activity



NCBI Public Services

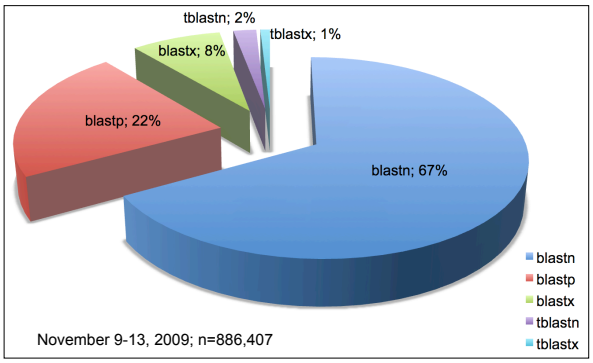
BLAST and BLAST-like programs

NCBI Public Services

- Traditional BLAST (formerly blastall) nucleotide, protein, translations
  - blastn nucleotide query vs. nucleotide database
  - blastp protein query vs. protein database
  - blastx nucleotide query vs. protein database
  - tblastn protein query vs. translated nucleotide database
  - tblastx translated query vs. translated database
- Megablast nucleotide only
  - Contiguous megablast
    - Nearly identical sequences
  - Discontiguous megablast
    - Cross-species comparison
- Position Specific BLAST Programs protein only
  - Position Specific Iterative BLAST (PSI-BLAST)
    - Automatically generates a position specific score matrix (PSSM)
  - Reverse PSI-BLAST (RPS-BLAST)
    - Searches a database of PSI-BLAST PSSMs

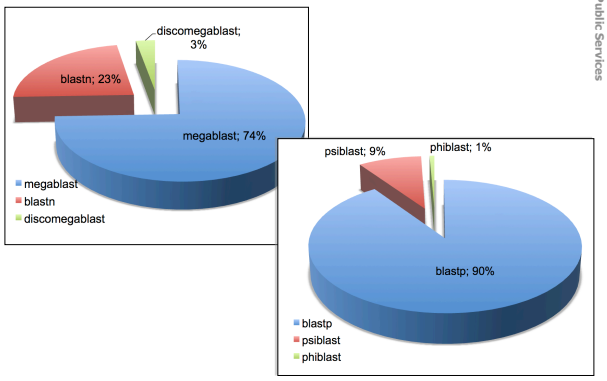
BLAST Searches by Program

NCBI Public Services



Nucleotide and Protein BLAST Programs

NCBI Public Services



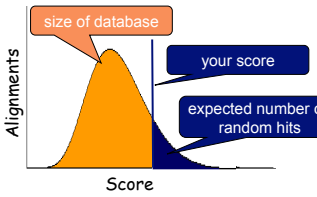
Local Alignment Statistics

NCBI Public Services

High scores of local alignments between two random sequences follow the Extreme Value Distribution

Expect Value

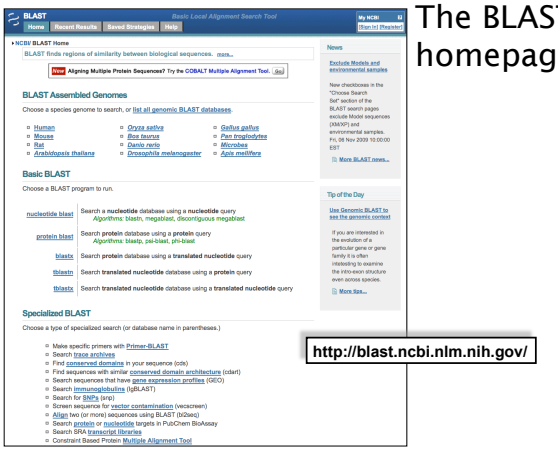
E = number of database hits you expect to find by chance



$E = Kmne^{-\lambda S}$ or $E = mn2^{-S'}$

K = scale for search space
λ = scale for scoring system
S' = bitscore = (λS - lnK)/ln2

(applies to ungapped alignments)



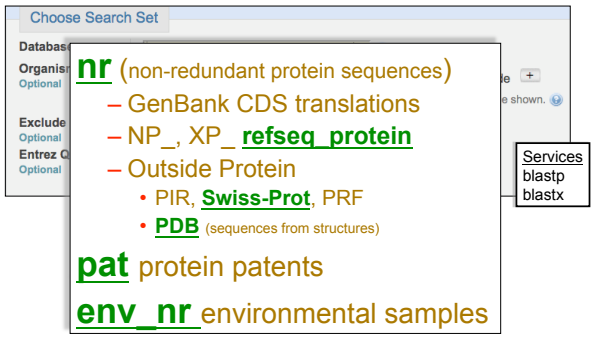
The BLAST homepage

NCBI Public Services

Basic BLAST: Databases

NCBI Public Services

Non-redundant protein



nr (non-redundant protein sequences)
- GenBank CDS translations
- NP_, XP_ **refseq_protein**
- Outside Protein
  - PIR, **Swiss-Prot**, PRF
  - **PDB** (sequences from structures)

pat protein patents

env_nr environmental samples

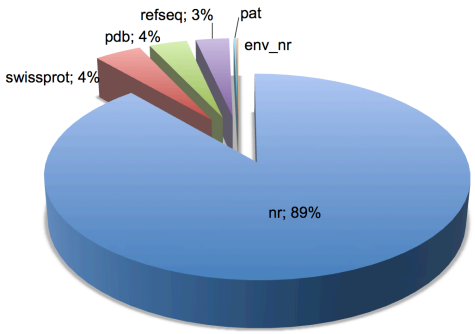
NCBI Public Services

Protein Database Sizes

Database	Sequences	Residues
nr	10,133,783	3,456,922,644
refseq_protein	7,413,069	2,589,005,568
swissprot	430,511	159,291,105
pat	817,680	166,184,433
pdb	44,202	10,171,945

12/04/2009

Protein Database Selection

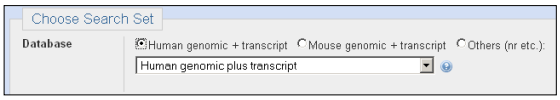


November 9-13, 2009; n=222,791

NCBI Public Services

Nucleotide Databases: Human and Mouse

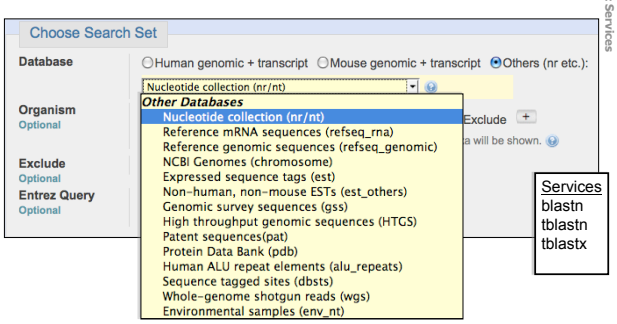
Megablast, blastn service



- Human and mouse genomic and transcript now default
- Separate sections in output for mRNA and genomic
- Direct links to Map Viewer for genomic sequences

NCBI Public Services

Nucleotide Databases: Traditional



NCBI Public Services

Nucleotide Databases: Traditional

Databases are mostly non-overlapping

- **nr (nt)**
  - Traditional GenBank
  - NM_ and XM_ RefSeqs
    - refseq_rna
- **NCBI Genomes**
  - NC_ RefSeqs
  - GenBank Chromosomes
- **dbest**
  - EST Division
    - non-human, non-mouse ests
- **htgs**
  - HTG division
- **gss**
  - GSS division
- **wgs**
  - whole genome shotgun contigs
- **env_nt**
  - environmental samples

NCBI Public Services

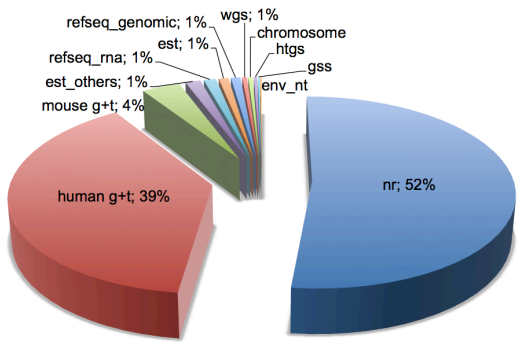
Nucleotide Database Sizes

Database	Sequences	Residues
nr/nt	10,362,162	29,617,088,643
refseq_rna	2,042,538	3,240,301,155
NCBI genomes	10,047	49,094,451,709
est	63,832,451	35,136,825,005
htgs	143,742	24,082,224,044
gss	27,198,629	17,658,377,015
wgs	31,377,631	149,309,157,200
env_nt	17,708,548	7,218,208,433

12/04/2009

NCBI Public Services

Nucleotide Database Selection



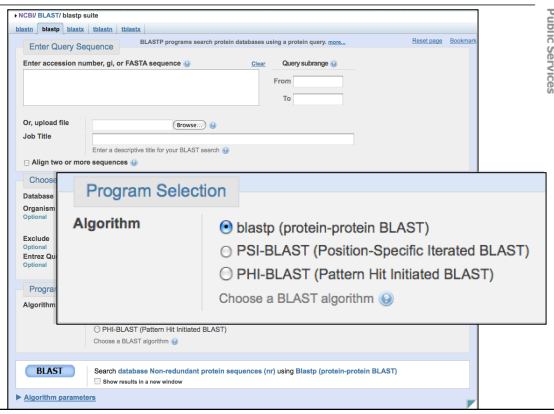
November 9-13, 2009; n=535,836

NCBI Public Services

Using Basic BLAST

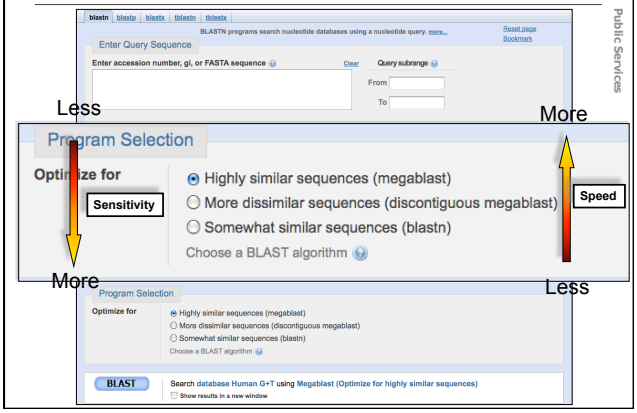
NCBI Public Services

Universal Form: Protein

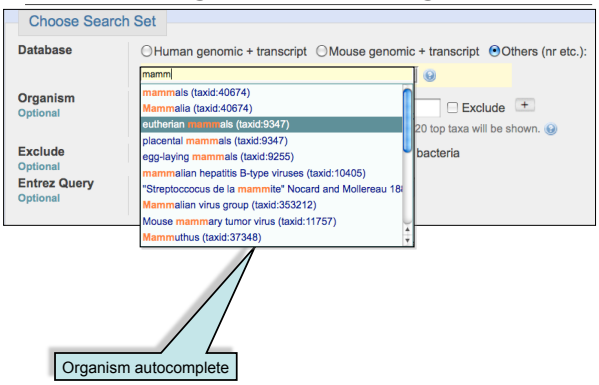


NCBI Public Services

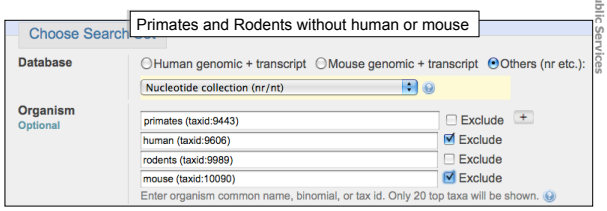
Universal Form: Nucleotide



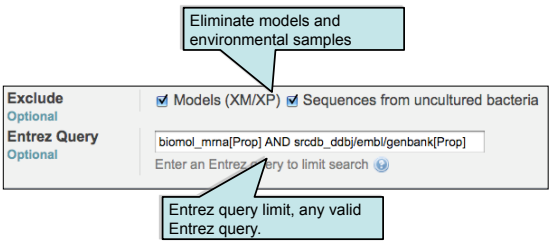
Limiting Database: Organism



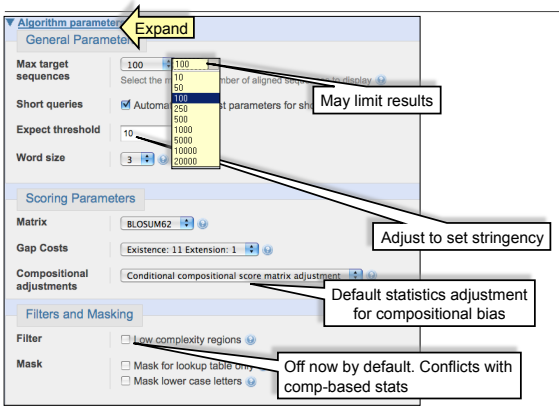
Combining Organisms



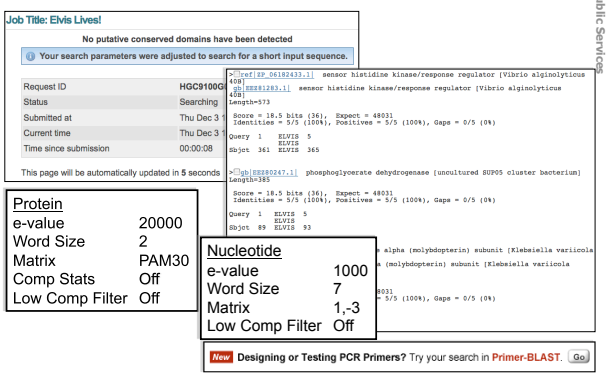
More Limits



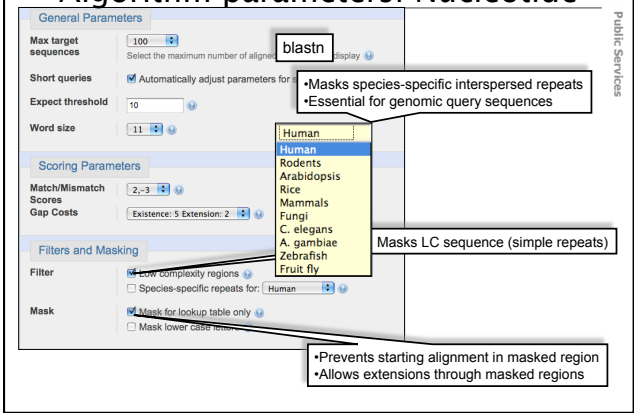
Algorithm parameters: Protein



Automatic Short Sequence Adjustment



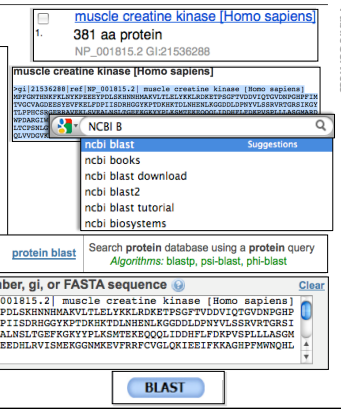
Algorithm parameters: Nucleotide



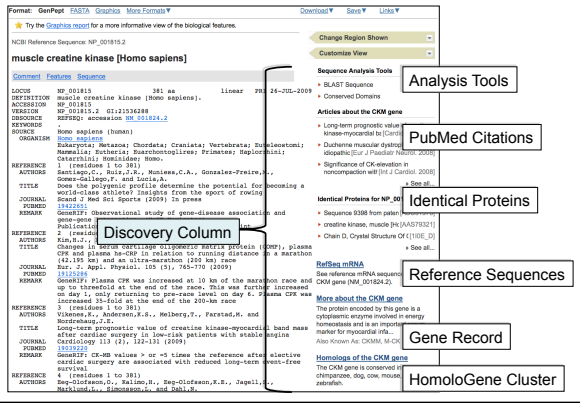
Basic BLAST: Protein

The hard way to run a BLAST Search

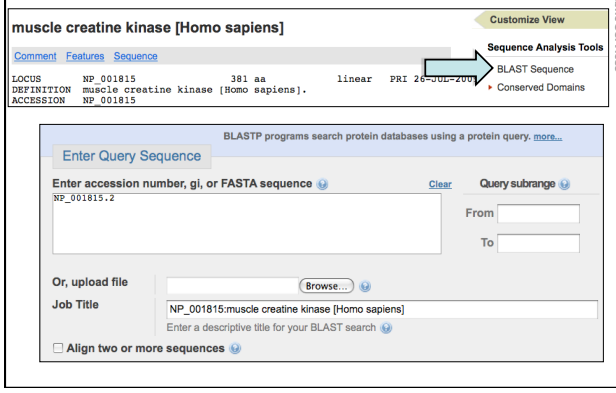
1. Search protein with "Human Muscle Creatine Kinase"
2. Click on summary for NP_001815
3. Change format to FASTA
4. Select sequence
5. Copy sequence
6. Google search "BLAST"
7. Link to NCBI BLAST Homepage
8. Link to Protein BLAST form
9. Paste FASTA sequence into form
10. Click BLAST button



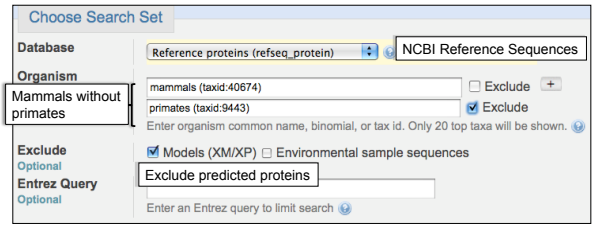
An easier way: Entrez protein record



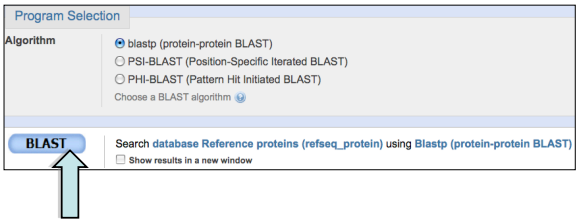
BLAST Ad to BLAST form



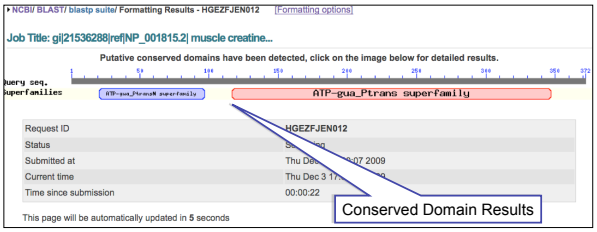
Database and limits



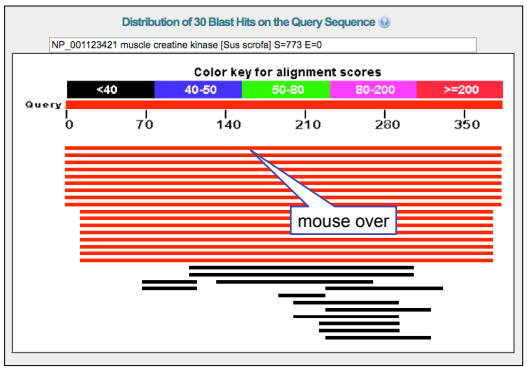
Run Search



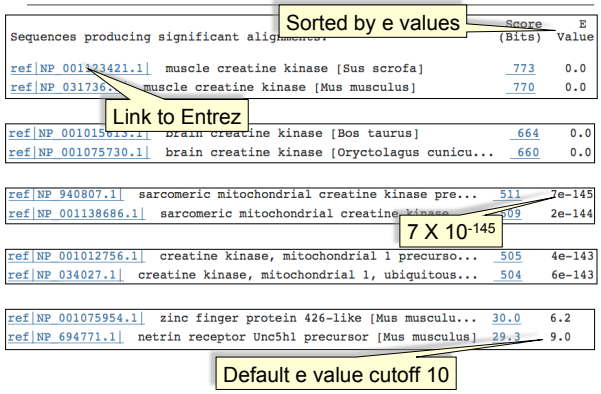
BLAST Formatting Page



BLAST Output: Graphical Overview



BLAST Output: Descriptions



BLAST Output: Alignments

```
>ref|NP_940807.1| UG sarcomeric mitochondrial creatine kinase precursor [Mus musculus]
 ref|NP_001121124.1| UG sarcomeric mitochondrial creatine kinase precursor [Rattus norvegicus]
Length=419

 GENE ID: 76722 Ckmt2 | creatine kinase, mitochondrial 2 [Mus musculus]
(Over 10 PubMed links)

 Score =  511 bits (1316),  Expect = 7e-145, Method: Compositional matrix adjust.
 Identities = 239/361 (66%), Positives = 285/361 (78%), Gaps = 2/361 (0%)

Query  14   YKPEEEYPDLSKHNNHMAKVLTLELYKKLRDKETPSGFTVDDVIQTGVDNPGHPFIMTVG  73
            + P  +YPDL K PIISDRHGGYKP-TDKHKTDLNGVDNPGHPFI TVG
Sbjct  48   FPPSADYPDLRK P+I  RH GY P   KH TDL+GVDNPGHPFIKTVG  107

Query  74   CVAGDP  PV KLRHNGYDPRVMKHPTDLDDDLDPNYVLSSRVR  132
            +  LP  I  RH  P  KH TDL+  +  D  YVLSSRVR
Sbjct  166  PADLFD  KLRHNGYD UMKHPTDLDASKITHGQ-FD YVLSSRVR  166
```

Identical match

positive score (conservative)

Negative or zero

gap

What happens without XP_ filter?

```
Sequences producing significant alignments:                          Score (Bits)  E Value

ref|XP_533641.2|  PREDICTED: similar to Creatine kinase M-type...    772    0.0
ref|XP_537561.2|  PREDICTED: similar to Creatine kinase B-type...    653    0.0
ref|XP_868659.1|  PREDICTED: similar to Creatine kinase B-type...    551    1e-156
ref|NP_001138686.1|  sarcomeric mitochondrial creatine kinase ...    509    6e-144
ref|XP_535451.1|  PREDICTED: similar to creatine kinase, mitoc...    503    4e-142
ref|XP_860111.1|  PREDICTED: similar to creatine kinase, mitoc...    501    2e-141
ref|XP_859467.1|  PREDICTED: similar to creatine kinase, mitoc...    498    2e-140
ref|XP_859430.1|  PREDICTED: similar to creatine kinase, mitoc...    497    3e-140
ref|XP_[illegible]  PREDICTED: [illegible]                           433    8e-121
ref|XP_[illegible]  PREDICTED: [illegible]                           412    2e-114
ref|XP_[illegible]  PREDICTED: [illegible]                           358    3e-98
ref|XP_[illegible]  PREDICTED: [illegible]                           318    2e-86
ref|XP_854438.1|  PREDICTED: similar to Creatine kinase B-type...    239    2e-62
ref|XP_853207.1|  PREDICTED: similar to creatine kinase, brain...    224    6e-58
ref|XP_851000.1|  PREDICTED: similar to creatine kinase, brain...    202    1e-51
ref|XP_853614.1|  PREDICTED: similar to Creatine kinase B-type...    199    1e-50
ref|XP_542948.2|  PREDICTED: similar to creatine kinase, brain...    197    8e-50
ref|XP_860041.1|  PREDICTED: similar to creatine kinase, mitoc...    144    8e-34
ref|XP_546543.2|  PREDICTED: similar to Creatine kinase B-type...    140    1e-32
ref|XP_853217.1|  PREDICTED: similar to Creatine kinase B-type...    103    9e-22
ref|XP_852237.1|  PREDICTED: similar to low-density lipoprotei...    98.6   4e-20
ref|XP_859393.1|  PREDICTED: similar to Creatine kinase, sarco...    97.8   6e-20
ref|XP_849487.1|  PREDICTED: similar to creatine kinase, brain...    95.1   4e-19
ref|XP_544323.2|  PREDICTED: similar to creatine kinase, brain...    74.7   6e-13
ref|XP_851013.1|  PREDICTED: similar to Creatine kinase B-type...    72.8   2e-12
ref|XP_849758.1|  PREDICTED: similar to otoferlin isoform a is...    31.6   6.1
```

Results filtered for domestic dog proteins.
26 additional gene predictions from Dog alone.
Many are extra splice variants predicted by Gnomon.

Other Reports

Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment] NEW

TreeView

Tax BLAST

COBALT extension

TaxBLAST: Taxonomy Reports

Lineage Report

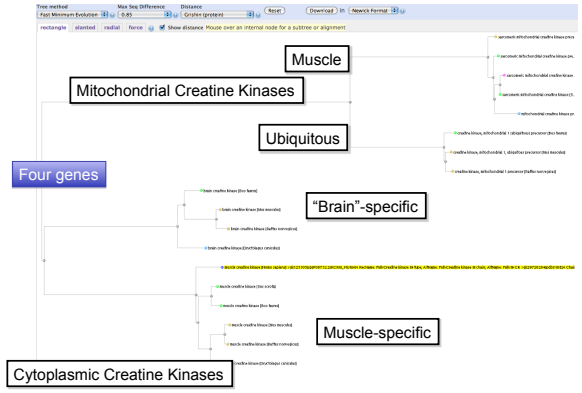
```
Eutheria        [placentals]
. Laurasiatheria  [placentals]
. . Cetartiodactyla [even-toed ungulates]
. . . Sus scrofa (wild boar) --------------   773 2 hits [even-toed ungulates]
. . . Bos taurus (cow) ....................   768 4 hits [even-toed ungulates]
. . Canis lupus familiaris (dogs) ---------   509 1 hit  [carnivores]
. Mus musculus (mouse) --------------------   770 4 hits [rodents]
. Oryctolagus cuniculus (domestic rabbit) .   768 3 hits [rabbits & hares]
. Rattus norvegicus (brown rat) ...........   766 4 hits [rodents]
```

Organism Report

```
 Sus scrofa (wild boar, ...) [even-toed ungulates] taxid 9823
ref|NP_001123421.1| muscle creatine kinase [Sus scrofa]          773  0.0
ref|NP_001038016.1| sarcomeric mitochondrial creatine kina...    509  2e-144

 Mus musculus (mouse) [rodents] taxid 10090
ref|NP_031736.1| muscle creatine kinase [Mus musculus]           770  0.0
ref|NP_067248.1| brain creatine kinase [Mus musculus]            652  0.0
ref|NP_940807.1| sarcomeric mitochondrial creatine kinase ...    511  8e-145
ref|NP_034027.1| creatine kinase, mitochondrial 1, ubiquit...    504  6e-143

 Bos taurus (cow, ...) [even-toed ungulates] taxid 9913
ref|NP_777198.2| muscle creatine kinase [Bos taurus]             768  0.0
ref|NP_001015613.1| brain creatine kinase [Bos taurus]           664  0.0
ref|NP_001029828.1| sarcomeric mitochondrial creatine kina...    508  5e-144
ref|NP_776700.1| creatine kinase, mitochondrial 1 (ubiquit...    504  7e-143

 Oryctolagus cuniculus (domestic rabbit, ...) [rabbits & hares] taxid 9986
ref|NP_001075708.1| muscle creatine kinase [Oryctolagus cu...    768  0.0
ref|NP_001075730.1| brain creatine kinase [Oryctolagus cun...    660  0.0
ref|NP_001156542.1| mitochondrial creatine kinase precurso...    500  1e-141

 Rattus norvegicus (brown rat, ...) [rodents] taxid 10116
ref|NP_036662.1| muscle creatine kinase [Rattus norvegicus]      766  0.0
ref|NP_036661.2| brain creatine kinase [Rattus norvegicus]       650  0.0
ref|NP_001121124.1| sarcomeric mitochondrial creatine kina...    511  8e-145
ref|NP_001012756.1| creatine kinase, mitochondrial 1 precu...    505  4e-143

 Canis lupus familiaris (dogs) [carnivores] taxid 9615
ref|NP_001138686.1| sarcomeric mitochondrial creatine kina...    509  2e-144
```

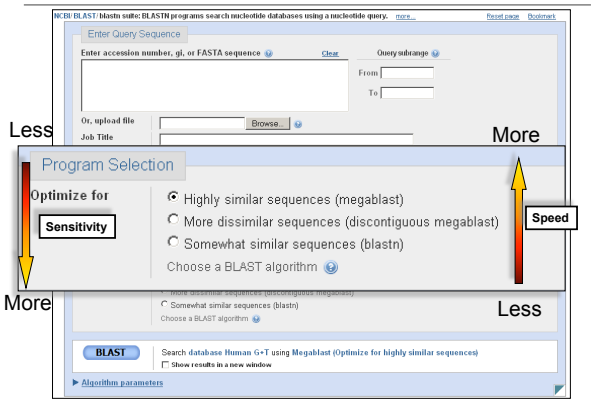
Four genes in each mammal.

TreeView: Distance Tree

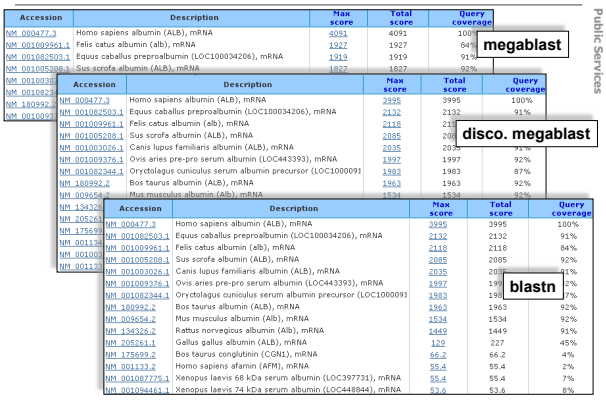


Basic BLAST: Nucleotide

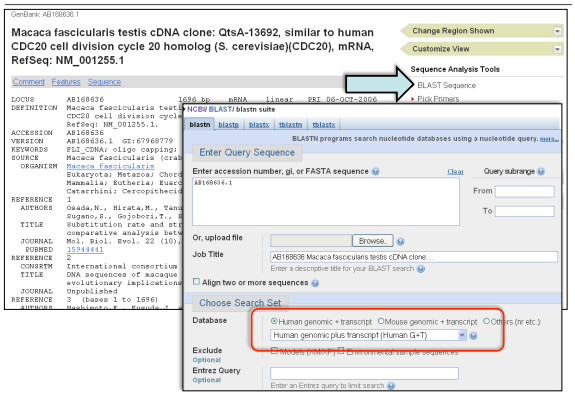
Universal Form: Nucleotide



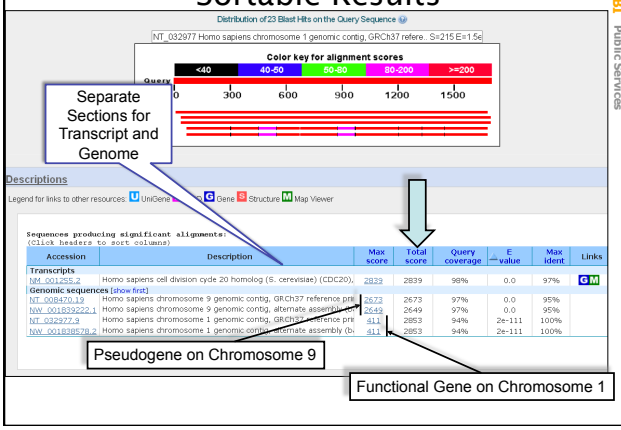
Nucleotide Results: ALB mRNA



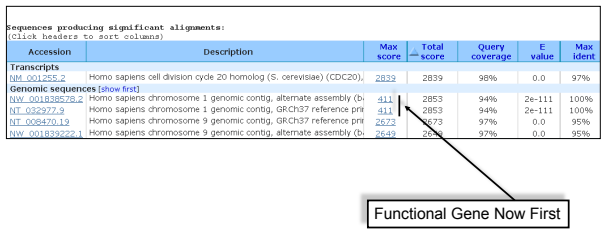
Macaque CDC20 Search



Sortable Results

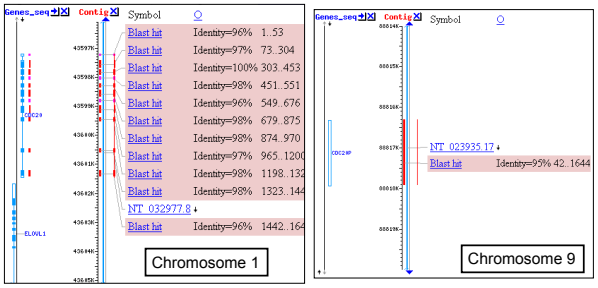


Total Score: All Segments



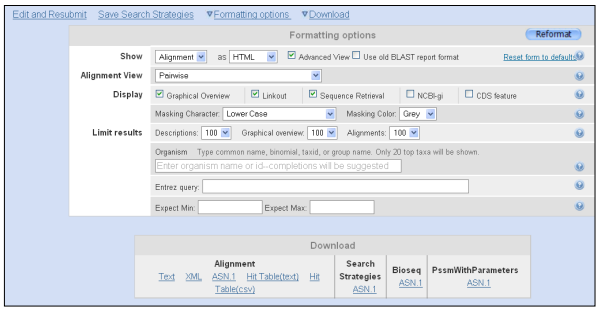
Alignments: Sorting in Exon Order

Links to Map Viewer

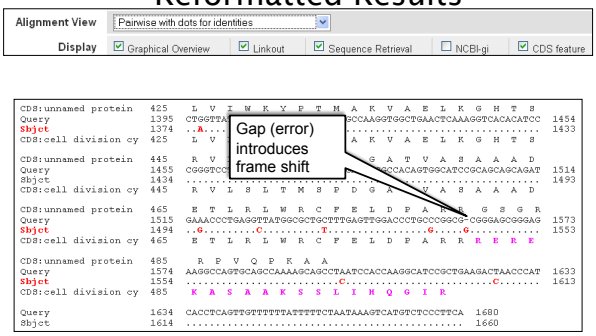


BLAST Formatting Options

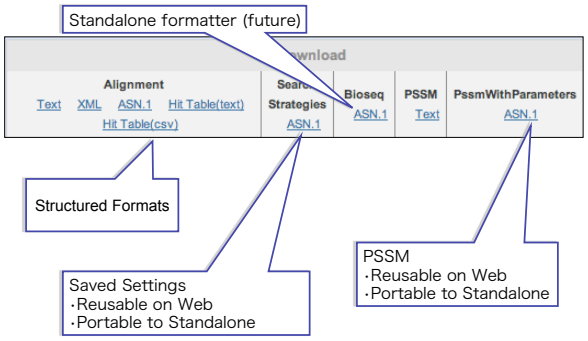
Formatting Page (Now on Results)



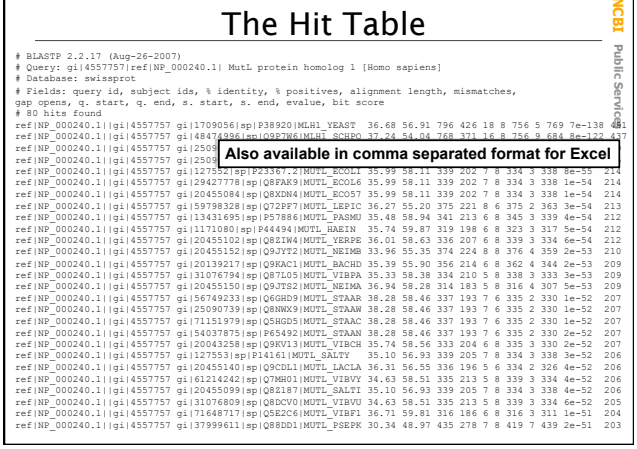
Reformatted Results



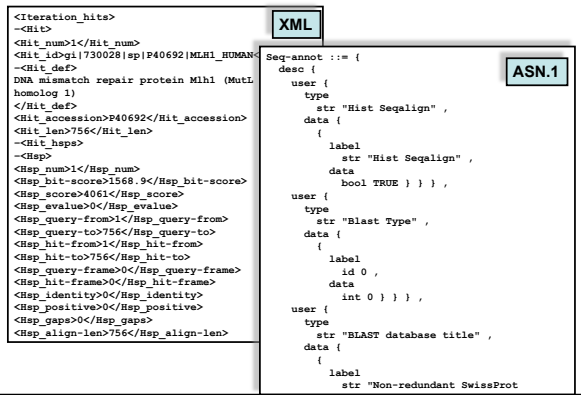
Download Options (Now on Results)



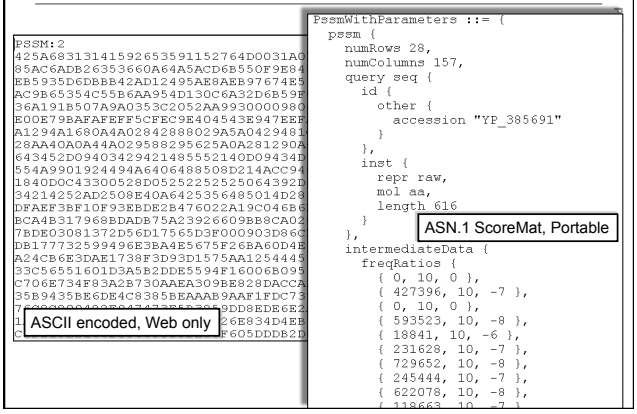
The Hit Table



Structured formats: XML and ASN.1



PSSMs: Restart PSI-BLAST



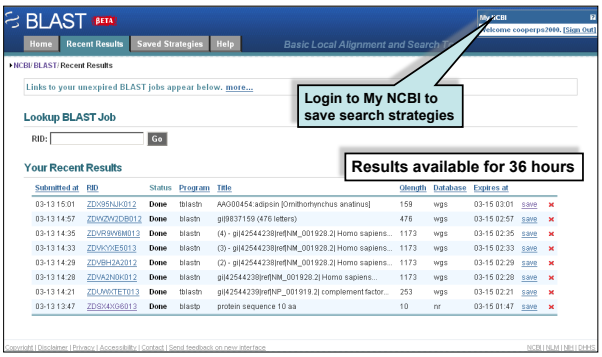
NCBI Public Services

Managing Searches

Recent Results
Saved Strategies

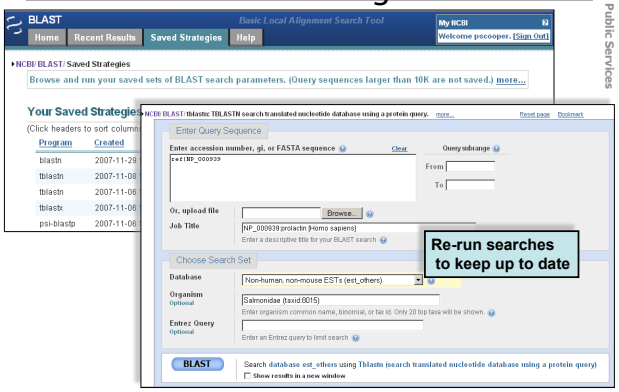
NCBI Public Services

Recent Results



NCBI Public Services

Saved Strategies



NCBI Public Services

Genome and Specialized BLAST

Nucleotide Databases: Human and Mouse

NCBI Public Services

Megablast, blastn service

Choose Search Set

Database: Human genomic + transcript | Mouse genomic + transcript | Others (nr etc.):

Human genomic plus transcript

- Human and mouse genomic and transcript now default
- Separate sections in output for mRNA and genomic
- Direct links to Map Viewer for genomic sequences

Genome BLAST pages

NCBI Public Services

BLAST Assembled Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

Map Viewer Homepage

NCBI Public Services

NCBI Map Viewer

Monocots (13)

Eudicots (21)

Scientific name	Common name	Build	Tools
Arabidopsis thaliana	thale cress	Build 7.0	
Beta vulgaris	beet		
Brassica juncea	brown mustard		
Brassica napus	rapeseed		
Brassica nigra	black mustard		
Brassica oleracea			
Brassica rapa	field mustard		
Capsicum annuum	pepper		
Glycine max	soybean		
Lotus japonicus	birdsfoot deervetch		
Manihot esculenta	cassava		
Medicago sativa	alfalfa		
Phaseolus vulgaris	kidney bean		
Populus trichocarpa	poplar	Build 1.1	
Prunus dulcis	sweet almond		
Solanum lycopersicoides	nightshade		
Solanum lycopersicum	tomato		
Solanum melongena	eggplant		
Solanum peruvianum	Peruvian tomato		
Theobroma cacao	cocoa		
Vigna radiata	mungbean		

Poplar Genome BLAST

NCBI Public Services

NCBI Home > Genomic Biology > BLAST

BLAST Black cottonwood Sequences

Enter an accession, gi, or a sequence in FASTA format:

Or, choose a file to upload

Set subsequence: (optional)

Database: genomic (reference only) 45688 sequences

Program: TBLASTN: Compare a protein sequence against a nucleotide database

genomic (reference only)
HTGS
Non-RefSeq RNA
Non-RefSeq protein
ESTs
Traces-other*
WGS contigs

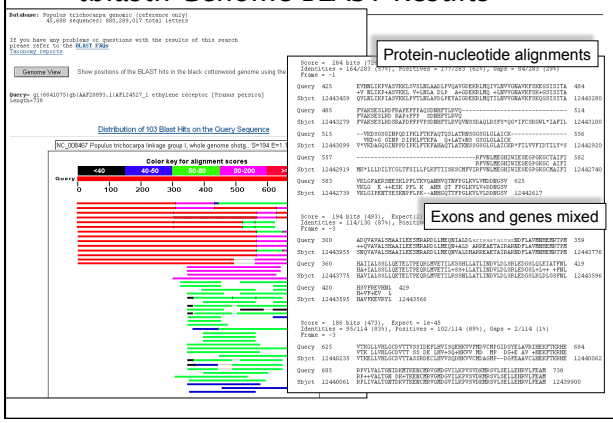
megaBLAST: Compare highly related nucleotide sequences
cross-species megaBLAST: Compare nucleotide sequences for other species to this genome
BLASTN: Compare nucleotide sequences
BLASTP: Compare protein sequences
BLASTX: Compare a nucleotide sequence against a protein database
TBLASTN: Compare a protein sequence against a nucleotide database

Expect: 0.01

Advanced options:

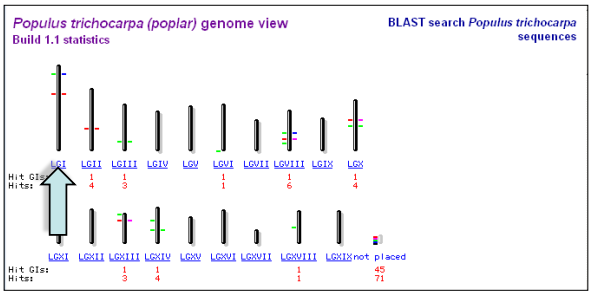
Begin Search | Clear Input

tblastn Genome BLAST Results



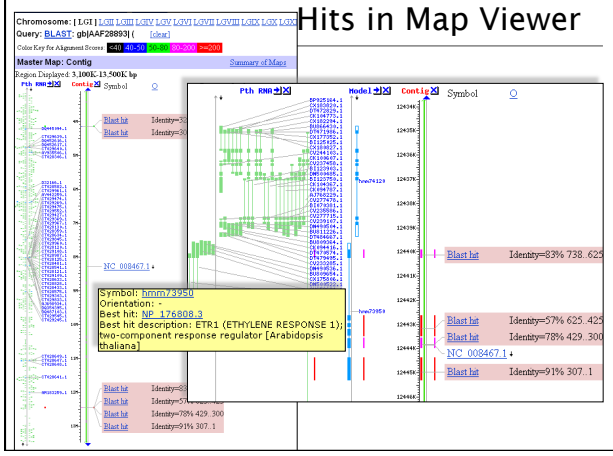
NCBI Public Services

Genomic Context of BLAST Hits



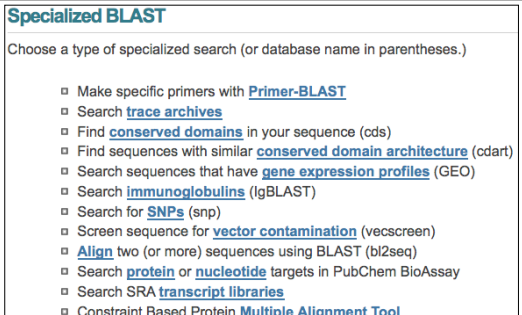
NCBI Public Services

Hits in Map Viewer



NCBI Public Services

Specialized BLAST Pages



NCBI Public Services

BLAST extensions and improvements

NCBI Public Services

- PrimerBlast – primer designer / specificity checker
- COBALT – Protein Multiple Alignment tool
- Integration / expansion of BLAST 2 Sequences

Primer BLAST from Sequence Record

NCBI Public Services

NCBI Reference Sequence: NM_001077.2

Homo sapiens UDP glucuronosyltransferase 2 family, polypeptide B17 (UGT2B17), mRNA

Comment Features Sequence

```
LOCUS       NM_001077                2099 bp    mRNA    linear   PRI 05-PRO-2010
DEFINITION  Homo sapiens UDP glucuronosyltransferase 2 family, poly
            (UGT2B17), mRNA.
ACCESSION   NM_001077 XM_936292
VERSION     NM_001077.2  GI:52851461
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 2099)
  AUTHORS   Strahm,E., Sottas,P.E., Schweizer,C., Saugy,M., Dvorak,J. and
            Saudan,C.
  TITLE     Steroid profiles of professional soccer players: an international
            comparative study
  JOURNAL   Br J Sports Med 43 (14), 1126-1130 (2009)
   PUBMED   19282302
  REMARK    GeneRIF: Observational study of gene-disease association. (HuGE
            Navigator)
REFERENCE   2  (bases 1 to 2099)
  AUTHORS   Ahn,J., Schumacher,F.R., Berndt,S.I., Pfeiffer,R., Albanes,D.,
            Andriole,G.L., Ardanaz,E., Boeing,H., Bueno-de-Mesquita,B.,
            Chanock,S.J., Clavel-Chapelon,F., Diver,W.R., Feigelson,H.S.,
            Gaziano,J.M., Giovannucci,E., Haiman,C.A., Henderson,B.E.,
            Hoover,R.N., Kolonel,L.N., Kraft,P., Ma,J., Le Marchand,L.,
            Overvad,K., Palli,D., Stattin,P., Stampfer,M., Stram,D.O.,
            Thomas,G., Thun,M.J., Travis,R.C., Trichopoulos,D., Virtamo,J.,
```

Change Region Shown

Customize View

Sequence Analysis Tools

- BLAST Sequence
- Pick Primers

Articles about the UGT2B17 gene

- Quantitative trait loci predicting circulating sex steroid hormoe [Hum Mol Genet. 2009]
- Copy-number variations (CNVs) of the human sex steroid metz [Hum Mutat. 2009]
- Characterization of UGTs active against SAHA and associatior [Cancer Res. 2009]

» See all...

RefSeq Protein Product

See the reference protein sequence for UDP-glucuronosyltransferase 2 family, polypeptide B17 precursor (NP_001068.1).

More about the UGT2B17 gene

UGT2B17 belongs to the family of UDP-glucuronosyltransferases (UGTs; EC 2.4.1.17), enzymes that catalyze the transfer of glucuronic acid fro...

Primer BLAST: Template and Primers

NCBI Public Services

PCR Template — Reset page — Save search parameters

Enter accession, gi, or FASTA sequence (A refseq record is preferred) — Clear

NM_001077.2

Or, upload FASTA file — Browse...

Range: From / To — Forward primer — Reverse primer

Primer Parameters

Use my own forward primer (5'->3' on plus strand) — Clear

Use my own reverse primer (5'->3' on minus strand) — Clear

PCR product size: Min 100, Max 1000

# of primers to return: 10

Primer melting temperatures (T_m): Min 57.0, Opt 60.0, Max 63.0, Max T_m difference 3

Please note the recent change in default Tm calculation

Primer BLAST: specificity params

NCBI Public Services

Exon/intron selection

A refseq mRNA sequence as PCR template input is required for options in the section

Exon junction span: No preference

Exon junction match: Exon at 5' side 7, 4; Minimal number of

- No preference
- Primer must span an exon-exon junction
- Primer may not span an exon-exon junction

Intron inclusion: Primer must be separated by at least one intron on the corresponding genomic DNA

Intron length range: Min 1000, Max 1000000

Get Primers

Primer Pair Specificity Checking Parameters

Specificity check: Enable search for primer pairs specific to the intended PCR template

Organism: 9606

Organism-specific search

Enter an organism name, taxonomy id or select from the suggestion list as you type.

Add more organisms

Database: Refseq RNA (refseq_rna)

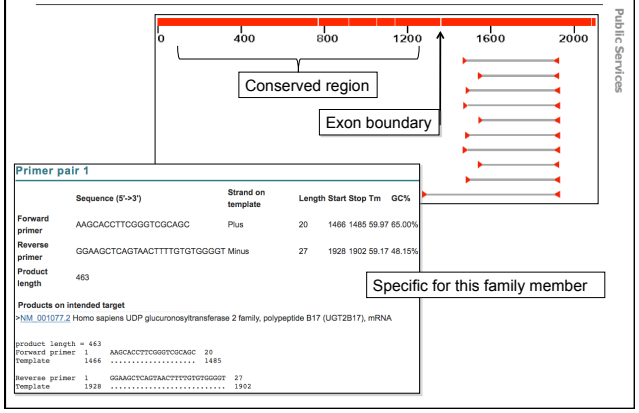
- Refseq RNA (refseq_rna)
- Genome (reference assembly from selected organisms)
- Genome (chromosomes from all organisms)
- nr

Primer specificity stringency: At least 2 total; at least 2 misma

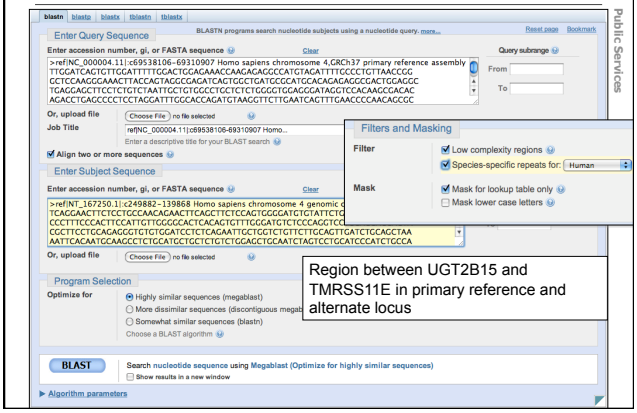
Misprimed product size deviation: 1000

Splice variant handling: Allow primer to ampli

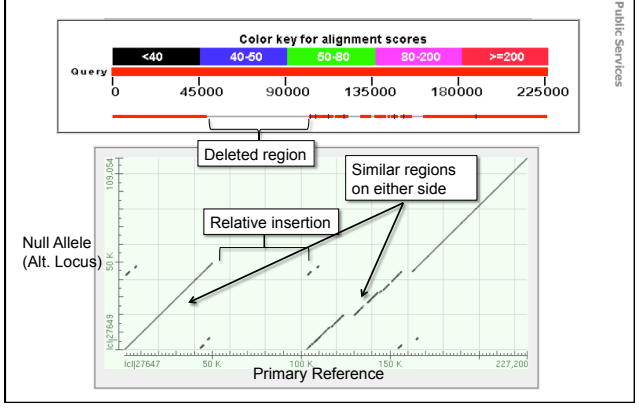
Primer Results



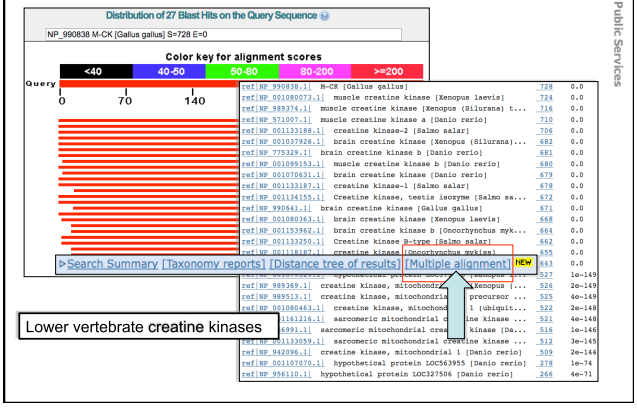
BLAST 2 Sequences



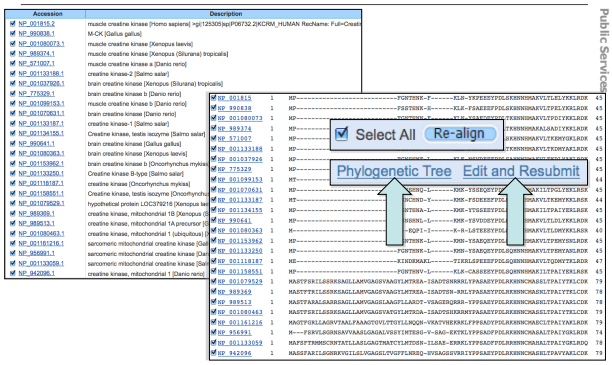
Alignment of reference and null-allele



COBALT Extension of BLAST

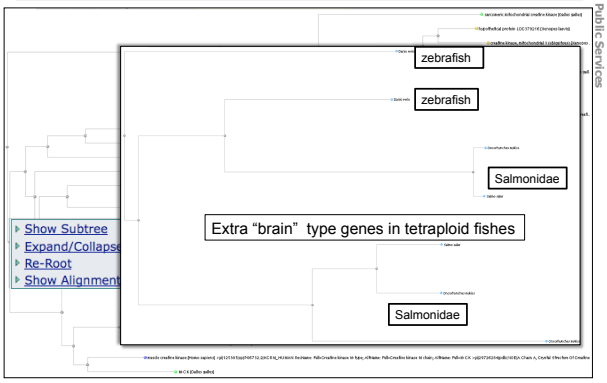


COBALT (Constraint Based Alignment Tool)

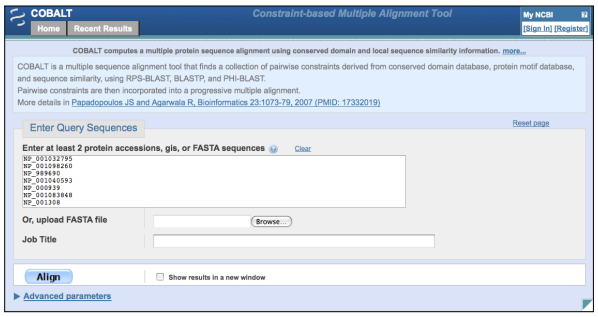


True multiple sequence alignment that uses conserved domain information

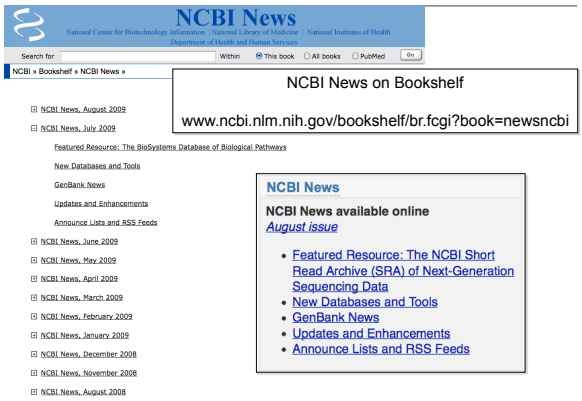
COBALT Tree



COBALT Interface



Keeping up with what's new

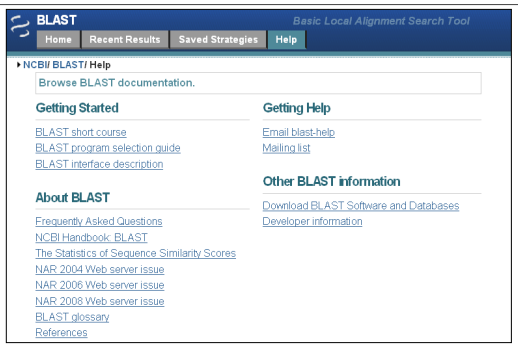


NCBI News on Bookshelf

www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=newsncbi

Getting Help

NCBI Public Services



Service Addresses

NCBI Public Services

- ***General Help*** `info@ncbi.nlm.nih.gov`
- ***BLAST*** `blast-help@ncbi.nlm.nih.gov`

Telephone support: 301- 496- 2475