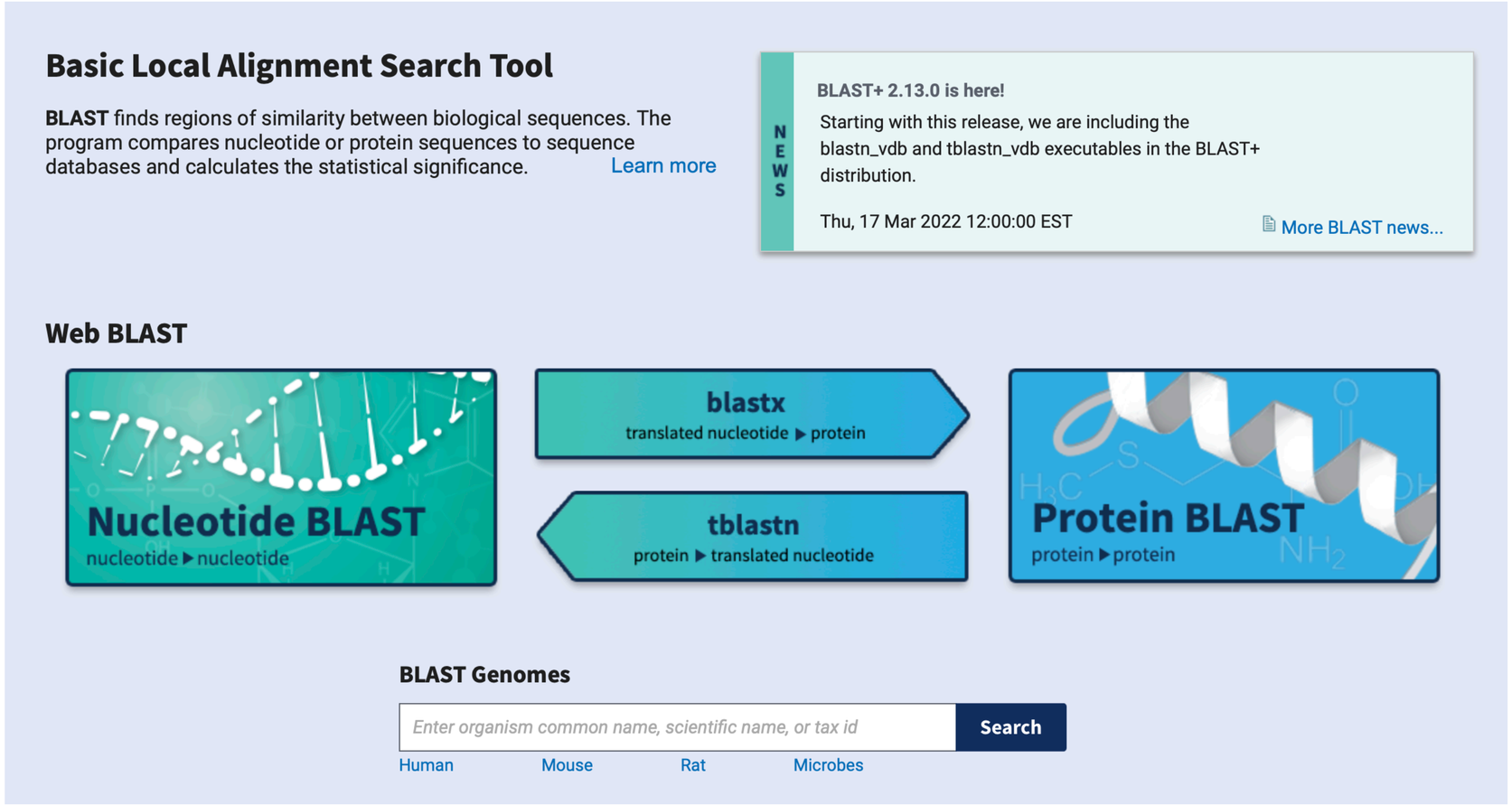
BLAST and Advanced Database Searching



https://blast.ncbi.nlm.nih.gov/Blast.cgi

Review

- Needleman-Wunsch achieves an optimal global alignment (NB that it also presents other alignments that may not be optimal, these score less)
- Smith-Waterman achieves optimal local alignments
- Smith-Waterman is slow. It requires computer space and time proportional to the product of the two sequences being aligned (or the product of a query against an entire database)
- Gotoh (1982) and Myers and Miller (1988) improved the algorithms so both global and local alignment require less time and space
- FASTA and BLAST provide rapid alternatives to S-W.

ian.simpson@ed.ac.uk

Overview

Performing a basic BLAST search
- specify sequence of interest
- select appropriate BLAST program
- selecting database
- set search & formatting parameters

BLAST process
- uses local alignment search strategy
- has three parts: list, scan, extend
- local alignment search statistics and E value
- making sense of raw scores with bit scores
- relationship between E and p values

BLAST search strategies
- principles of BLAST searching
- evaluation of result significance
- handling too many or few results

ian.simpson@ed.ac.uk

Learning Outcomes

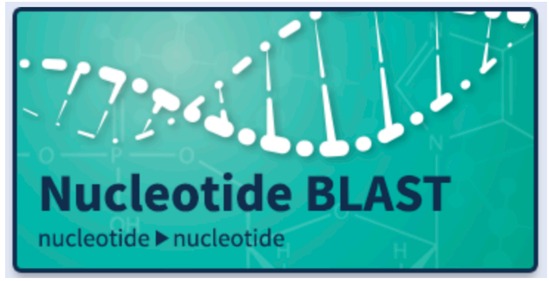
- perform BLAST searches at the NCBI website
- understand how to vary optional BLAST search parameters
- explain the three phases of a BLAST search
  - compile
  - scan/extend
  - trace-back
- define the mathematical relationship between expect values and scores
- strategies for BLAST searching

ian.simpson@ed.ac.uk

BLAST Applications

BLAST

- BLAST (Basic Local Alignment Search Tool) allows rapid sequence comparison of a query sequence against a database
- The BLAST algorithm is fast, accurate, and accessible both via the web and the command line



- identifying orthologs and paralogs
- discovering new genes or proteins
- discovering variants of genes or proteins
- investigating expressed sequence tags (ESTs)
- exploring protein structure and function

ian.simpson@ed.ac.uk

Why Use BLAST?

- BLAST searching is fundamental to understanding the relatedness of any query sequence to other known proteins, RNA, or DNA sequences.
- One way of trying to work out what information an unknown sequence encodes is to assess its similarity to sequences of known function
- As we have seen, sequences are collected into large databases that also annotate those sequences with a wide range of different meta-data
- These databases are huge, e.g. RefSeq currently contains:-
  - 289,333,423 proteins, 56,423,426 transcripts, and 141,099 organisms (October 2023)
- Global alignment of our query sequence against such a large target database is not computationally tractable
- We need to perform a local alignment based search that can balance speed and accuracy
- BLAST - Basic Local Alignment Search Tool
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10. PubMed PMID: 2231712. (in the top 20 most cited papers >70k in WoS, >100k GS)

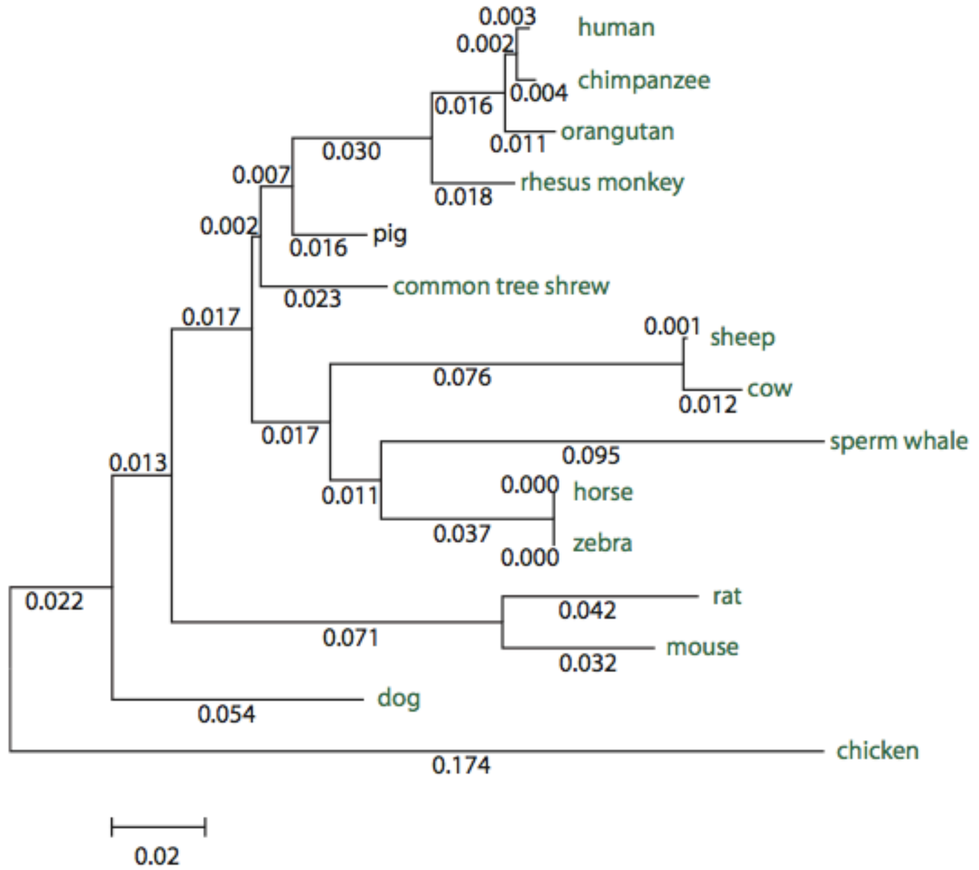
ian.simpson@ed.ac.uk

Types of Homology

Orthologs

Homologous sequences in different species that arose from a common ancestral gene during speciation; may or may not be responsible for a similar function.

Globins in different species



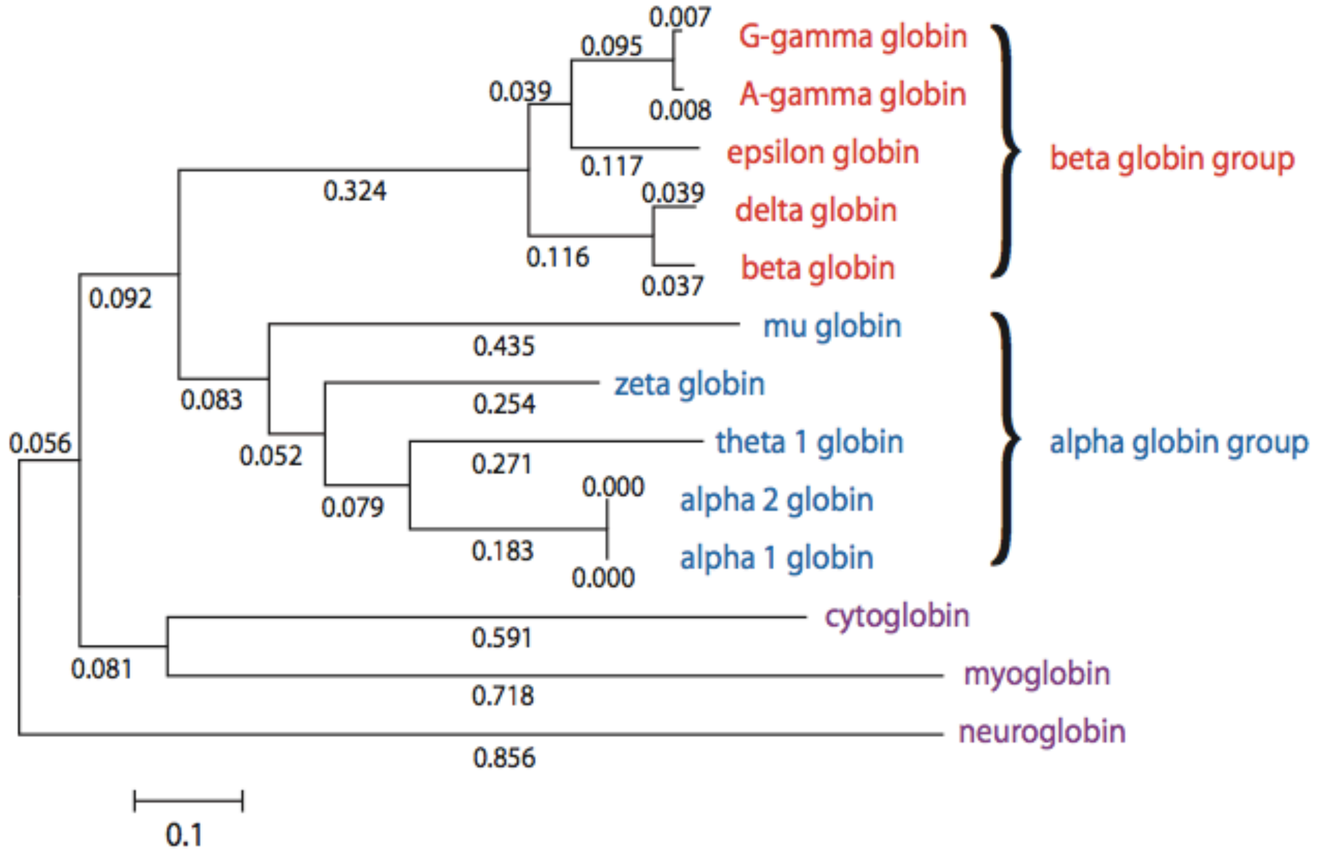
ian.simpson@ed.ac.uk

Types of Homology

Paralogs

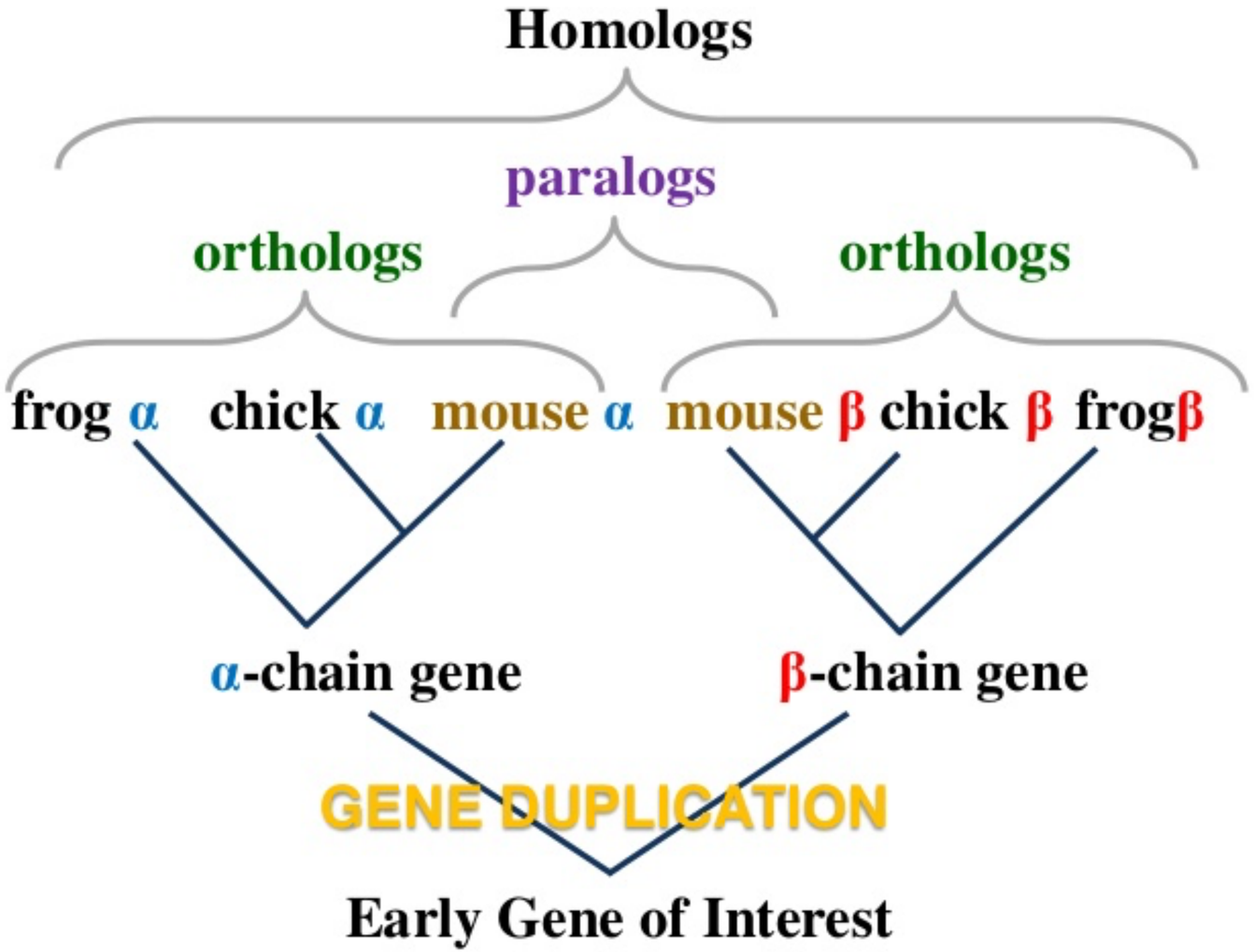
Homologous sequences within a single species that arose by gene duplication.

Globins in humans



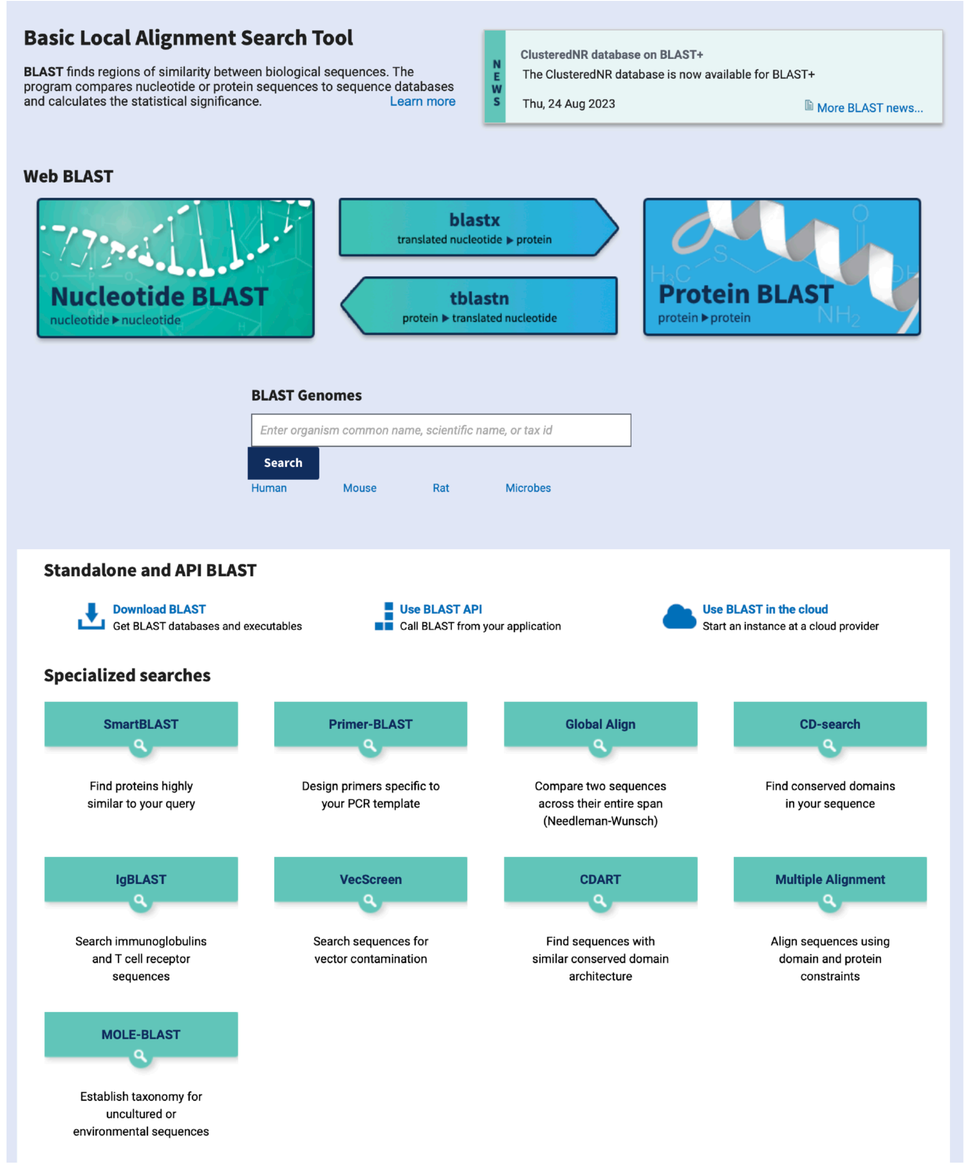
ian.simpson@ed.ac.uk

Orthologs and Paralogs Together



ian.simpson@ed.ac.uk

BLAST - Basic Local Alignment Search Tool



ian.simpson@ed.ac.uk

The BLAST Suite

Program	Query	Number of database searches	Database
blastp	protein	1	protein

Use blastp to compare a protein query to a database of proteins.

Program	Query	Number of database searches	Database
blastn	DNA	1	DNA

Use blastn to compare both strands of a DNA query against a DNA database.

Program	Query	Number of database searches	Database
blastx	DNA	6	protein

Blastx translates a DNA sequence into six protein sequences using all six possible reading frames, and then compares each of these proteins to a protein database.

Program	Query	Number of database searches	Database
tblastn	protein	6	DNA

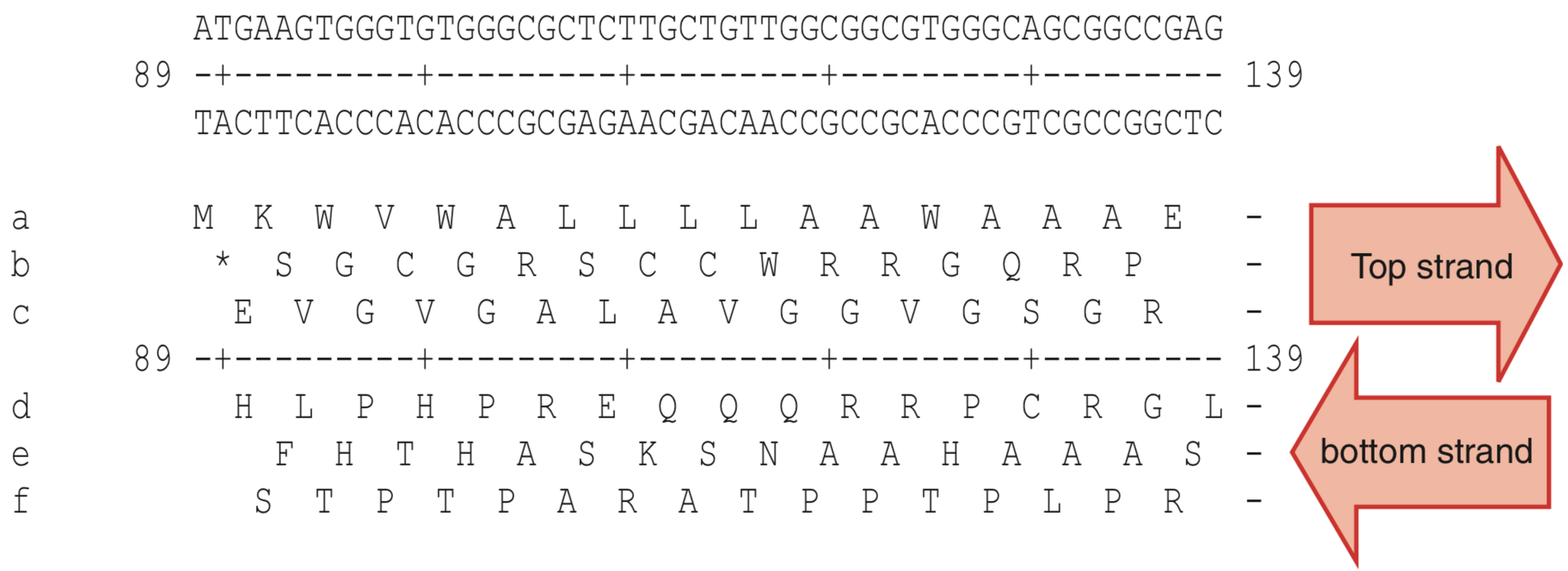
Tblastn is used to translate every DNA sequence in a database into six potential proteins, and then to compare your protein query against each of those translated proteins.

Program	Query	Number of database searches	Database
tblastx	DNA	36	DNA

Tblastx is the most computational intensive BLAST algorithm. It translates DNA from both a query and a database into six potential proteins, then performs 36 protein-protein database searches.

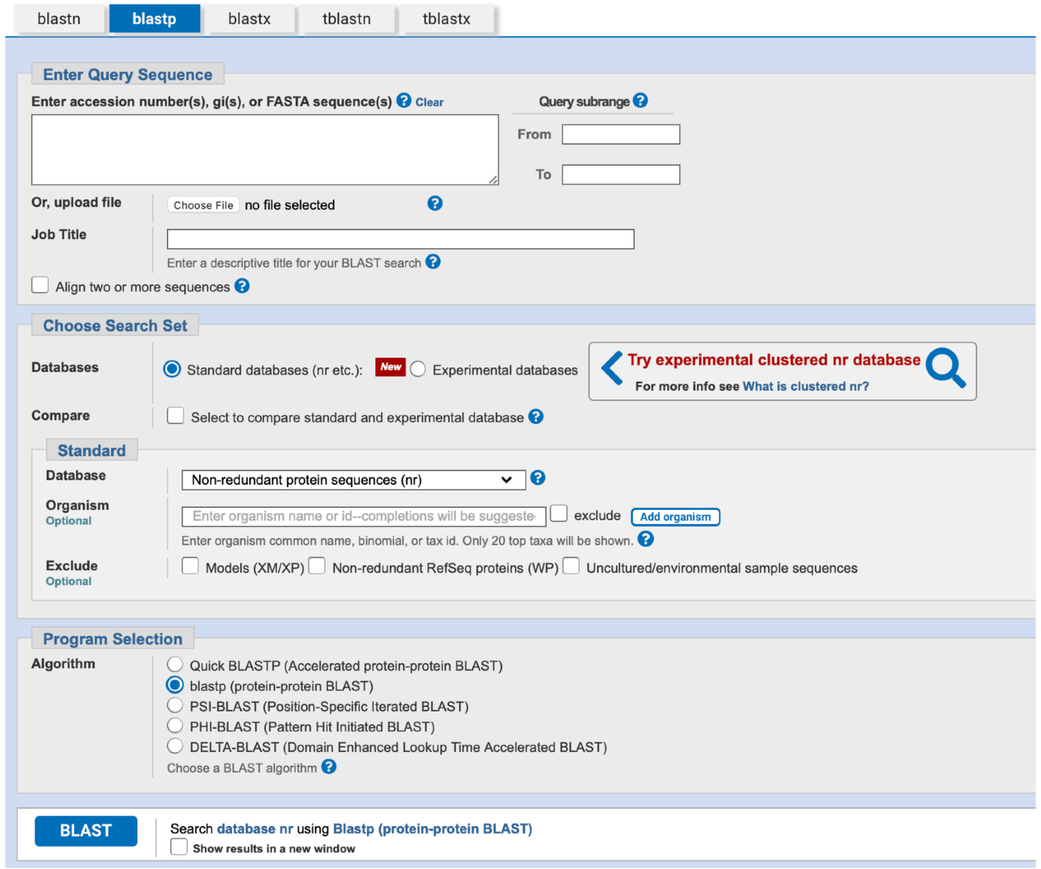
ian.simpson@ed.ac.uk

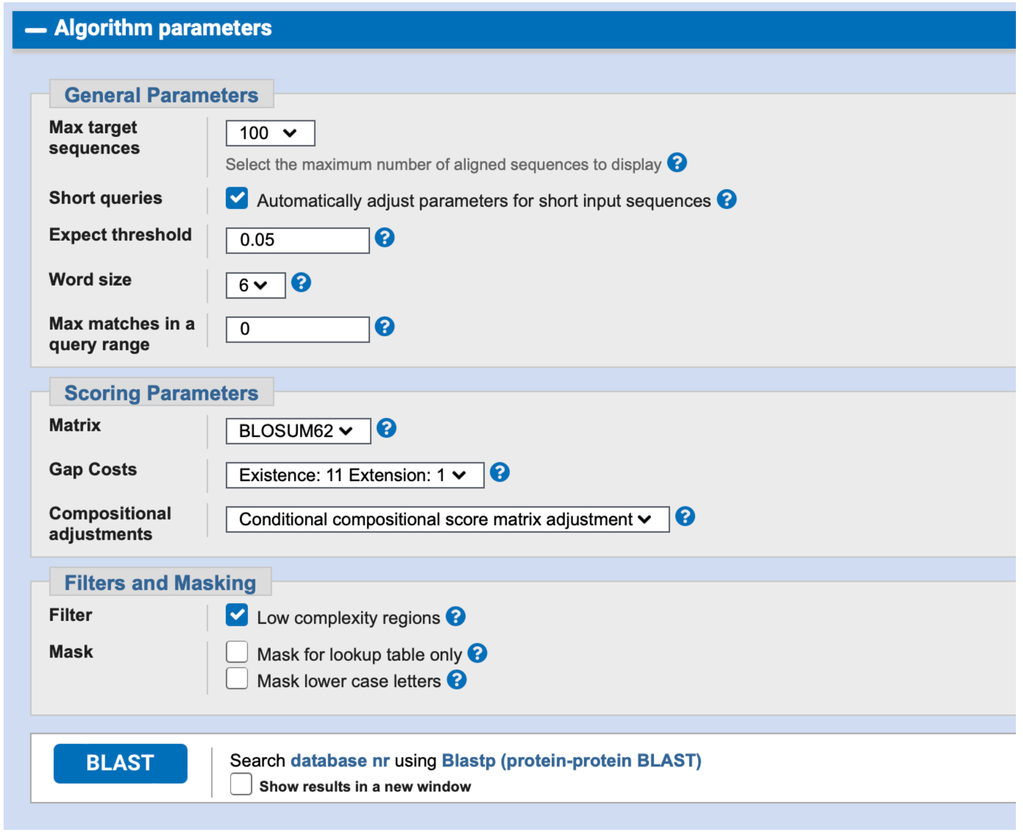
Nucleotide, Protein, Strands & Orientation



ian.simpson@ed.ac.uk

BLASTp Search

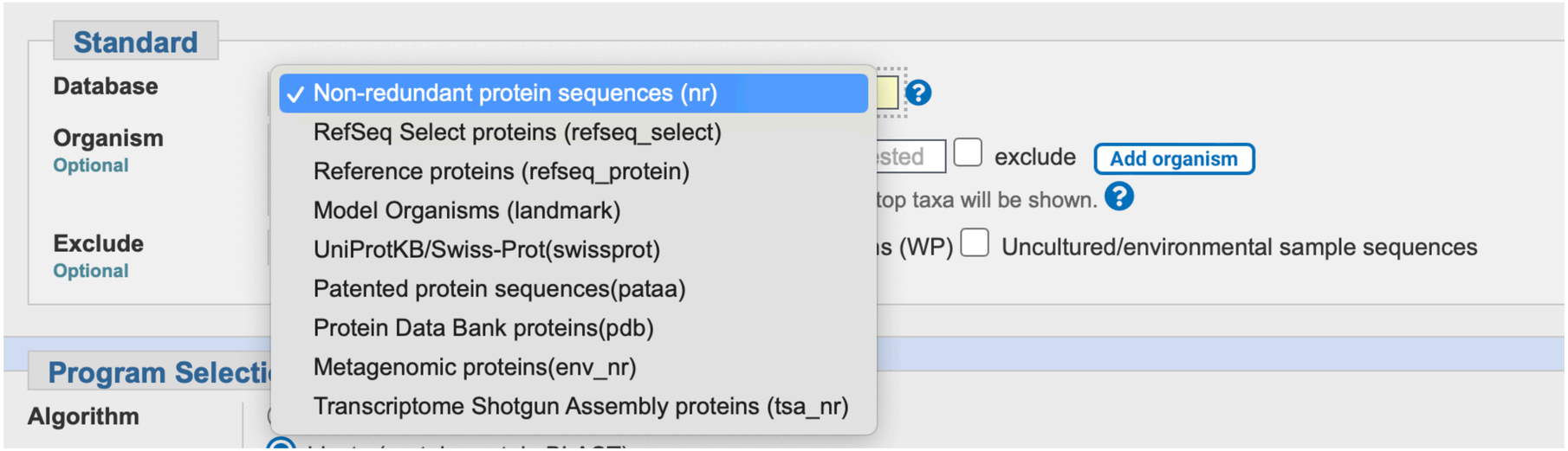




ian.simpson@ed.ac.uk

Optimising Searches

Database Choice - proteins



ian.simpson@ed.ac.uk

NCBI Protein DB Size

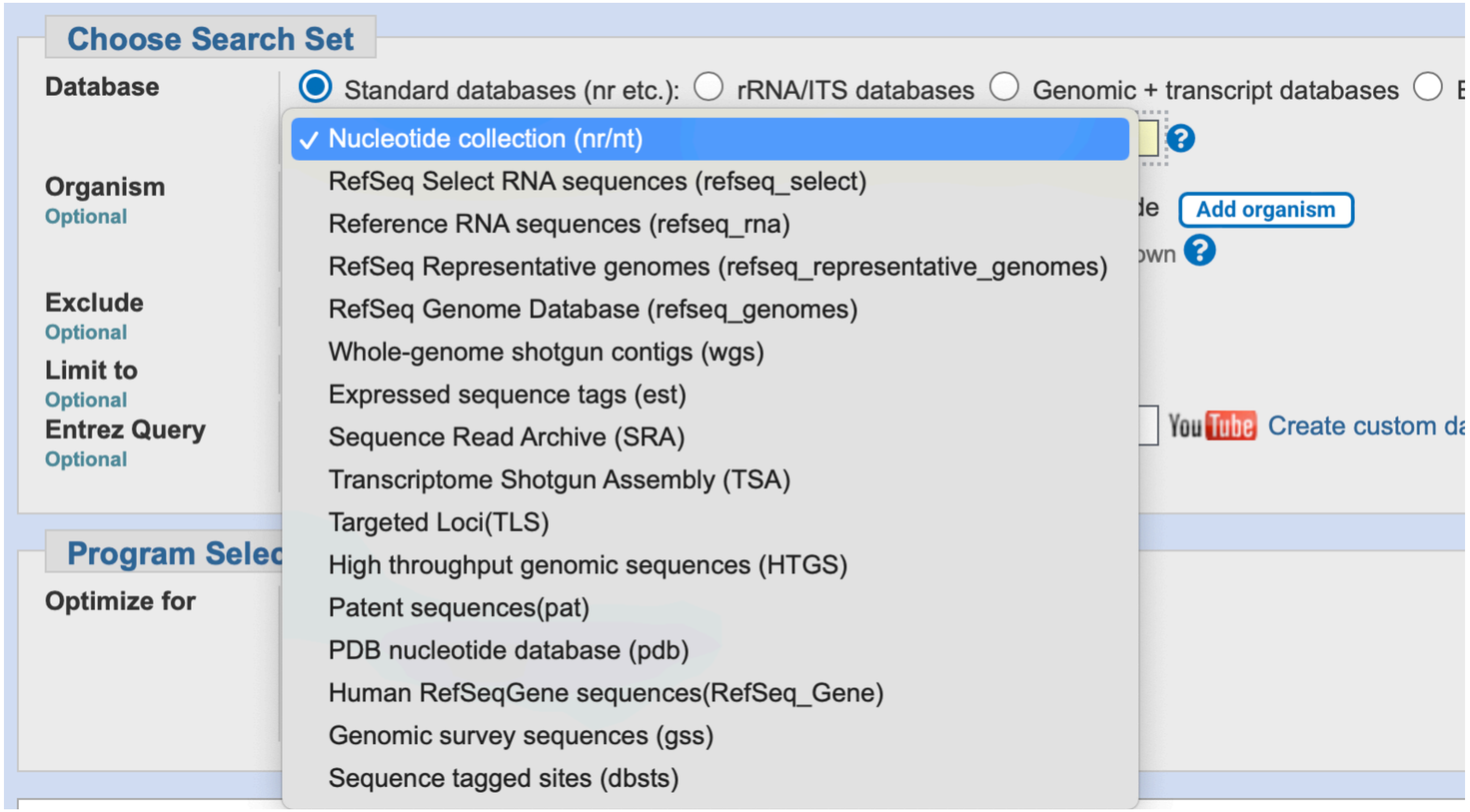
Database	Title	#sequences
nr	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects	622,030,880
RefSeq Select	NCBI RefSeq Select Proteins	65,623,135
Reference proteins	All RefSeq Proteins	283,249,322
UniProtKB/SwissProt	Non-redundant UniProtKB/SwissProt sequences.	482,080

October 2023

ian.simpson@ed.ac.uk

Optimising Searches

Database Choice - nucleotides



ian.simpson@ed.ac.uk

NCBI Nucleotide DB Sizes

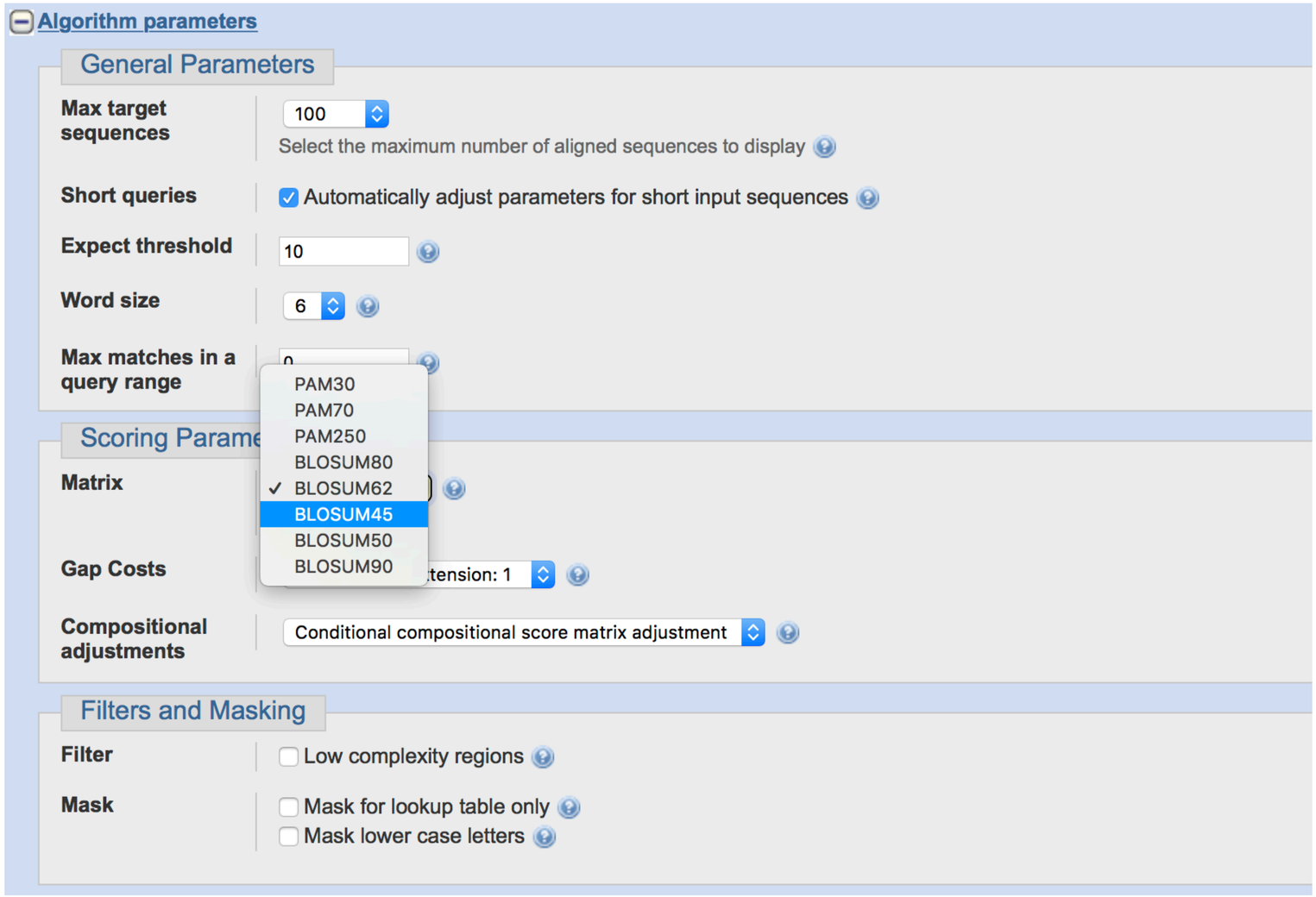
Database	Description	# Sequences
Nucletotide collection (nr/nt)	GenBank+EMBL+DDBJ+PDB+RefSeq sequences, but excludes EST, STS, GSS, WGS, TSA, patent sequences, HTGS and sequences >100Mb.	99,339,860
RefSeqSelect RNA sequences	NCBI RefSeq transcript sequences from **human and mouse**, restricted to the RefSeq Select set with one representative transcript per protein-coding gene.	62,683
RefSeq Genomes	NCBI Refseq genomes across all taxonomy groups. It contains only the top-level sequences; only the longest sequences for any given part of the genomes are included.	44,707,592
Expressed sequence tags (ESTs)	Database of GenBank+EMBL+DDBJ sequences from EST Divisions	77,548,292

October 2023

ian.simpson@ed.ac.uk

Optimising Searches

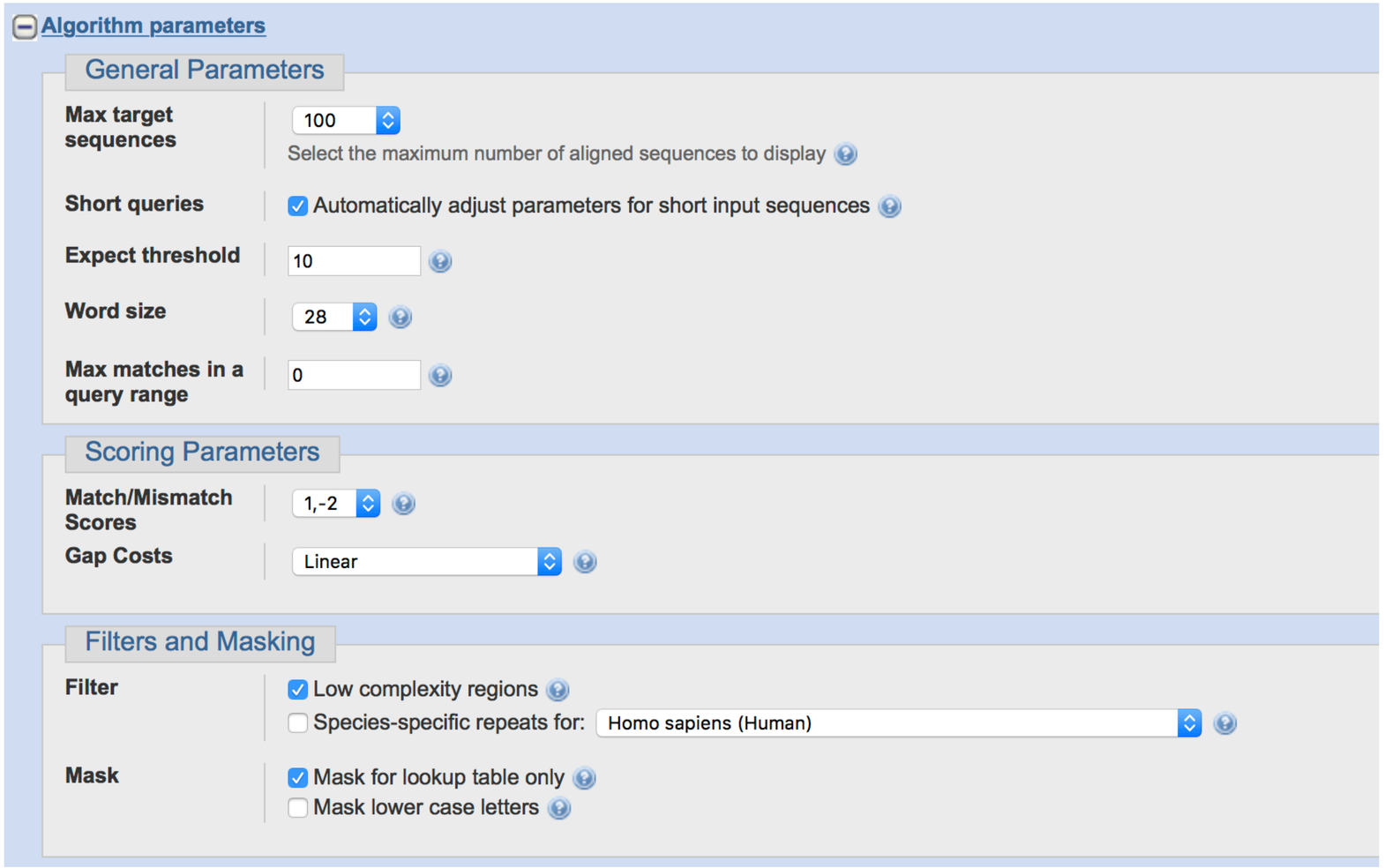
Tuning Parameters



ian.simpson@ed.ac.uk

Optimising Searches

Tuning Parameters



ian.simpson@ed.ac.uk

Parameter Choice Influences Score & E-Value Results

This adjusts substitution matrices based on the composition of the sequences being compared

Yi-Kuo Yu, Stephen F. Altschul, The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions, **Bioinformatics**, Volume 21, Issue 7, April 2005, Pages 902–911, https://doi.org/10.1093/bioinformatics/bti070

Default matrix uses a compositional matrix adjustment

Expect = 0.05

(a) Default: conditional compositional score matrix adjustment

Insulin-like peptide 3 [Drosophila melanogaster]

Sequence ID: ref|NP_648360.2| **Length:** 120 **Number of Matches:** 1

Range 1: 32 to 114 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
31.6 bits(70)	0.050	Compositional matrix adjust.	21/88(24%)	40/88(45%)	12/88(13%)

```
Query  29   HLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQ-  87
            LCG  L E L  +C    + +   T+R  + +   Q++  G       L+ L  + S+Q
Sbjct  32   KLCGRKLPETLSKLCV---YGFNAMTKRTLDPVNFNQID--GFEDRSLLERLLSDSSVQM  86

Query  88   ------KRGIVEQCCTSICSLYQLENYC  109
                  + G+ ++CC   C++ +  YC
Sbjct  87   LKTRRLRDGVFDECCLKSCTMDEVLRYC  114
```

No adjustment

Expect = 0.009

(b) No adjustment (by default, filter low complexity regions)

Insulin-like peptide 3 [Drosophila melanogaster]

Sequence ID: ref|NP_648360.2| **Length:** 120 **Number of Matches:** 1

Range 1: 33 to 114 GenPept Graphics

Score	Expect	Identities	Positives	Gaps
33.5 bits(75)	0.009	21/87(24%)	40/87(45%)	12/87(13%)

```
Query  30   LCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQ--  87
            LCG  L E L  +C    + +   T+R  + +   Q++  G       L+ L  + S+Q
Sbjct  33   LCGRKLPETLSKLCV---YGFNAMTKRTLDPVNFNQID--GFEDRSLLERLLSDSSVQML  87

Query  88   -----KRGIVEQCCTSICSLYQLENYC  109
                 + G+ ++CC   C++ +  YC
Sbjct  88   KTRRLRDGVFDECCLKSCTMDEVLRYC  114
```

This takes into account biases in the amino acid compositions of the sequences being aligned and the databases being queried

Schäffer AA et al. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. **Nucleic Acids Res**. 2001 Jul 15;29(14):2994-3005. doi: 10.1093/nar/29.14.2994.

Composition based statistics

Expect = 1×10^{-4}

(c) Composition-based statistics

Insulin-like peptide 3 [Drosophila melanogaster]

Sequence ID: ref|NP_648360.2| **Length:** 120 **Number of Matches:** 1

Range 1: 33 to 114 GenPept Graphics

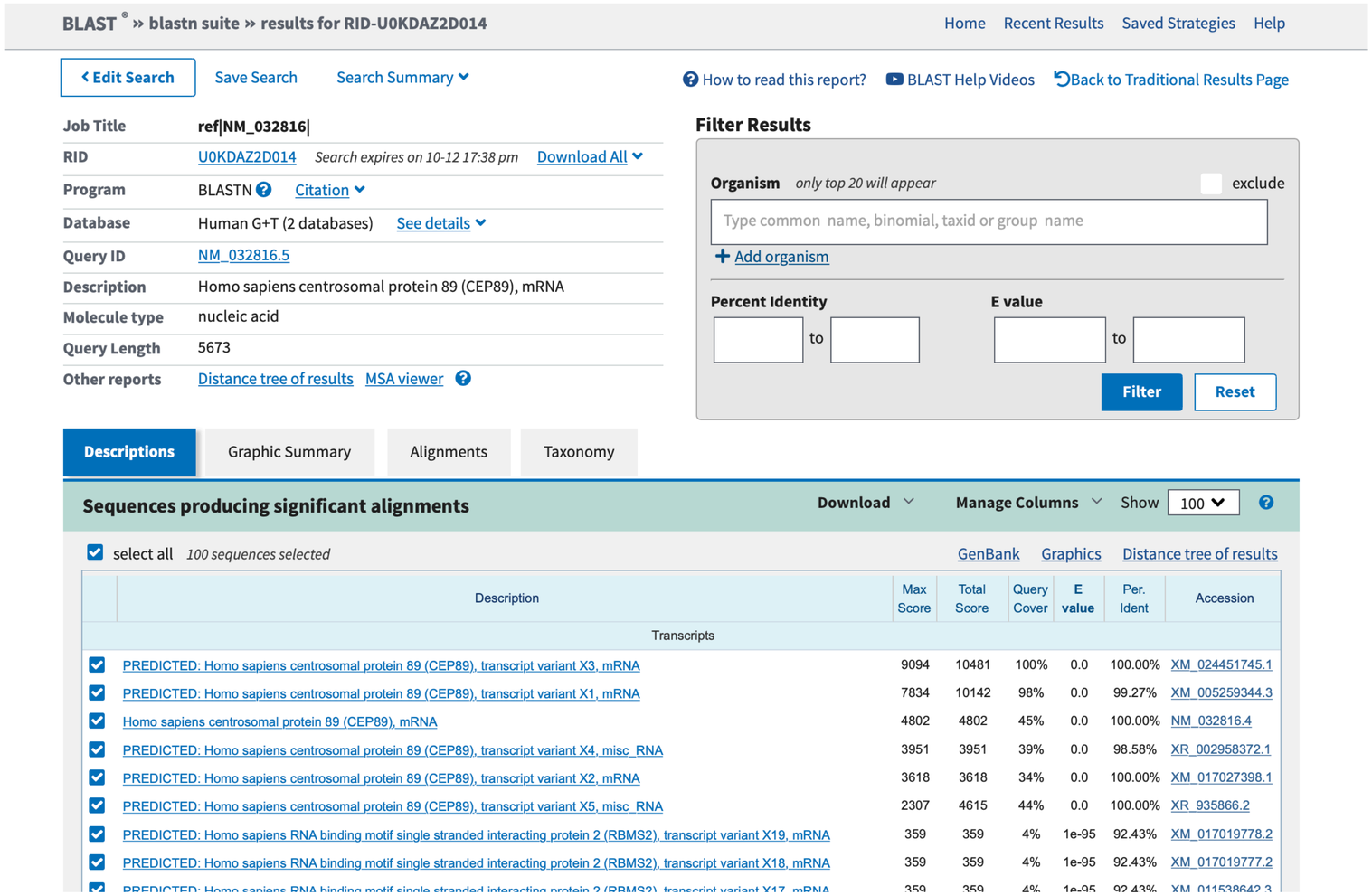
Score	Expect	Method	Identities	Positives	Gaps
30.4 bits(67)	1e-04	Composition-based stats.	21/87(24%)	40/87(45%)	12/87(13%)

```
Query  30   LCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQ--  87
            LCG  L E L  +C    + +   T+R  + +   Q++  G       L+ L  + S+Q
Sbjct  33   LCGRKLPETLSKLCV---YGFNAMTKRTLDPVNFNQID--GFEDRSLLERLLSDSSVQML  87

Query  88   -----KRGIVEQCCTSICSLYQLENYC  109
                 + G+ ++CC   C++ +  YC
Sbjct  88   KTRRLRDGVFDECCLKSCTMDEVLRYC  114
```


ian.simpson@ed.ac.uk

BLAST Result Output



BLAST ® » blastn suite » results for RID-U0KDAZ2D014

Home Recent Results Saved Strategies Help

‹ Edit Search | Save Search | Search Summary

How to read this report? | BLAST Help Videos | Back to Traditional Results Page

Job Title	ref|NM_032816|
RID	U0KDAZ2D014 *Search expires on 10-12 17:38 pm* Download All
Program	BLASTN Citation
Database	Human G+T (2 databases) See details
Query ID	NM_032816.5
Description	Homo sapiens centrosomal protein 89 (CEP89), mRNA
Molecule type	nucleic acid
Query Length	5673
Other reports	Distance tree of results MSA viewer

Filter Results

Organism *only top 20 will appear* exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity to

E value to

Filter Reset

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Manage Columns Show 100

select all *100 sequences selected* GenBank Graphics Distance tree of results

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
Transcripts						
PREDICTED: Homo sapiens centrosomal protein 89 (CEP89), transcript variant X3, mRNA	9094	10481	100%	0.0	100.00%	XM_024451745.1
PREDICTED: Homo sapiens centrosomal protein 89 (CEP89), transcript variant X1, mRNA	7834	10142	98%	0.0	99.27%	XM_005259344.3
Homo sapiens centrosomal protein 89 (CEP89), mRNA	4802	4802	45%	0.0	100.00%	NM_032816.4
PREDICTED: Homo sapiens centrosomal protein 89 (CEP89), transcript variant X4, misc_RNA	3951	3951	39%	0.0	98.58%	XR_002958372.1
PREDICTED: Homo sapiens centrosomal protein 89 (CEP89), transcript variant X2, mRNA	3618	3618	34%	0.0	100.00%	XM_017027398.1
PREDICTED: Homo sapiens centrosomal protein 89 (CEP89), transcript variant X5, misc_RNA	2307	4615	44%	0.0	100.00%	XR_935866.2
PREDICTED: Homo sapiens RNA binding motif single stranded interacting protein 2 (RBMS2), transcript variant X19, mRNA	359	359	4%	1e-95	92.43%	XM_017019778.2
PREDICTED: Homo sapiens RNA binding motif single stranded interacting protein 2 (RBMS2), transcript variant X18, mRNA	359	359	4%	1e-95	92.43%	XM_017019777.2
PREDICTED: Homo sapiens RNA binding motif single stranded interacting protein 2 (RBMS2), transcript variant X17, mRNA	359	359	4%	1e-95	92.43%	XM_011538642.3

View the Search Summary

< Edit Search | Save Search | Search Summary

Search Parameters	
Program	blastn
Word size	28
Expect value	10
Hitlist size	100
Match/Mismatch scores	1,-2
Gapcosts	0,2.5
Low Complexity Filter	Yes
Filter string	L;W -t 45518;m;
Genetic Code	1

Database	
Posted date	Mar 28, 2018 11:35 AM
Number of letters	3,823,040,941
Number of sequences	160,592
Entrez query	None

Karlin-Altschul statistics		
Lambda	1.33271	1.28
K	0.620991	0.46
H	1.12409	0.85

Results Statistics	
Length adjustment	33
Effective length of query	5640
Effective length of database	3817741405
Effective search space	21532061524200
Effective search space used	21532061524200

Search Parameters

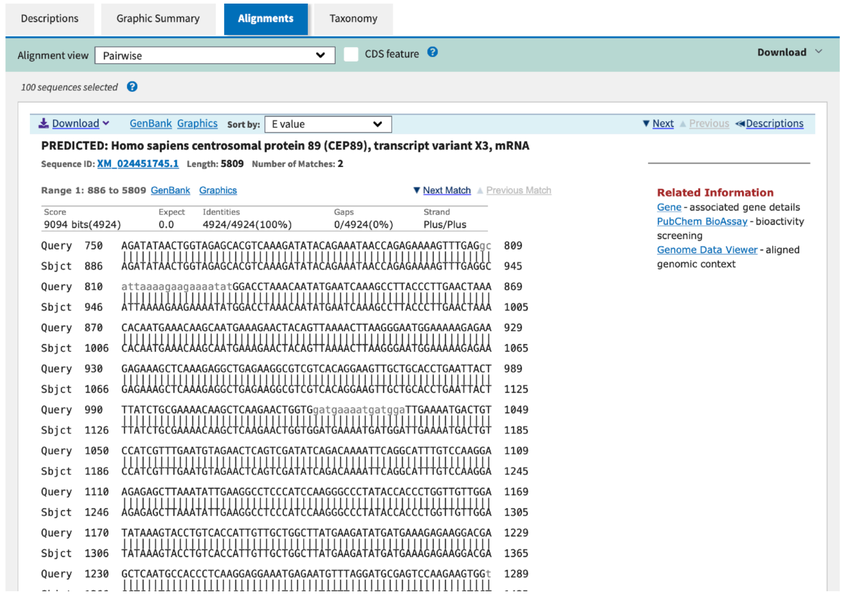
Database Details

Search Space Statistics

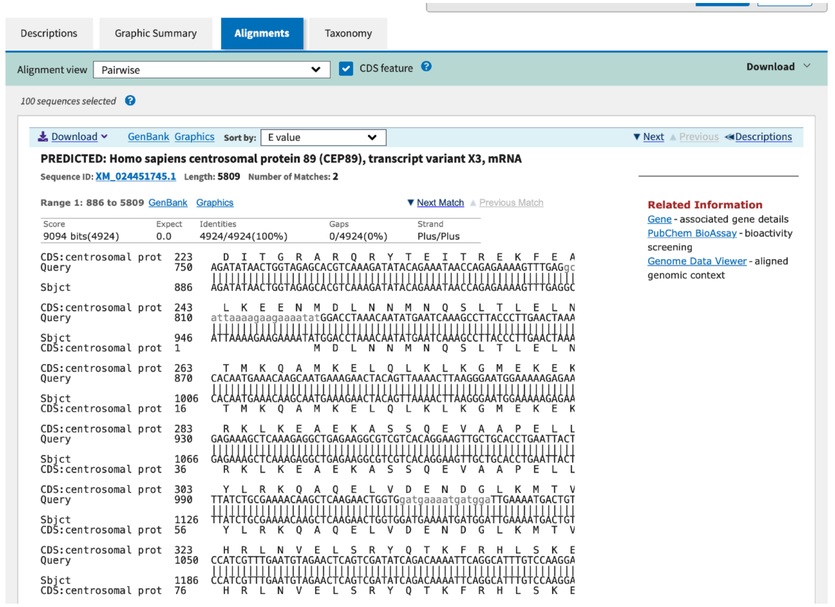
ian.simpson@ed.ac.uk

BLAST Results Viewable in Multiple Formats

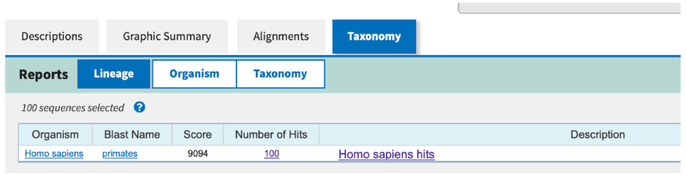
Nucleotide Alignment



Nucleotide Alignment with CDs



Taxonomy



ian.simpson@ed.ac.uk

Fundamentals of BLAST

"*The central idea of the BLAST* algorithm is to confine attention to segment pairs that contain a word pair of length w with a score *of at least T.*"

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10
(Citations - 107,903)

ian.simpson@ed.ac.uk

The Original BLAST Algorithm

Take a given query sequence:

…VTALWGKVNVD…

For w=3 make a list of 3-mers:

VTA, TAL, ALW, **LWG**, WGK…

Setting our Threshold (T) to 12

For each 3-mer in the sequence generate all possible matching pairs and score them using for example BLOSUM62:

So for LWG

```
LWG    4+11+6  = 21
IWG    2+11+6  = 19
MWG    2+11+6  = 19
VWG    1+11+6  = 18
FWG    0+11+6  = 17
AWG    0+11+6  = 17
LWS    4+11+0  = 15
LWM    4+11+0  = 15
LWA    4+11+0  = 15
LYG    4+ 2+6  = 12
LFG    4+ 1+6  = 11
FWS    0+11+0  = 11
AWS   -1+11+0  = 10
...
```

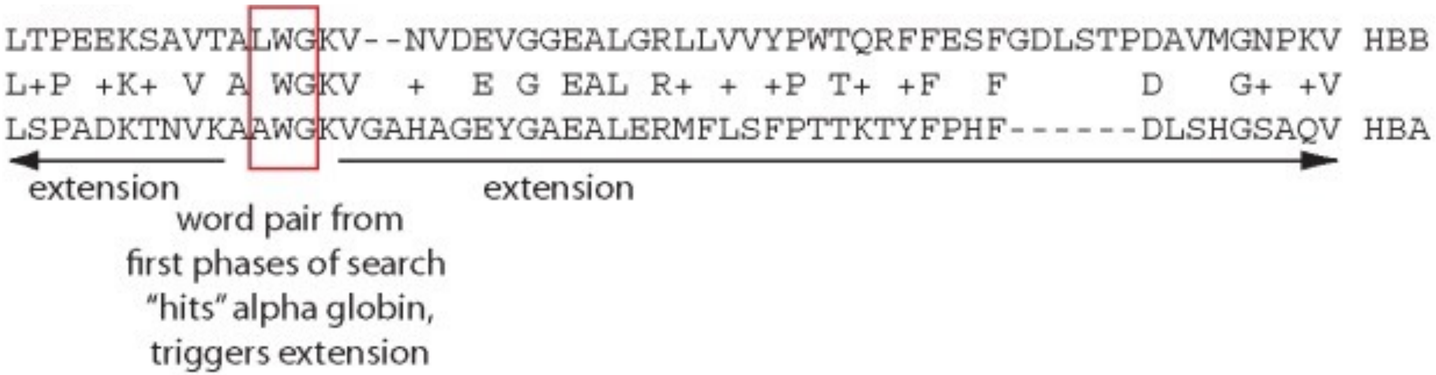
Threshold >=12

NB for BLASTN, the word size is typically 7, 11, or 15 (uses exact match). Changing word size is like changing threshold; w=15 gives fewer matches and is faster than w=11 or w=7. BLASTB, typically uses word size c.5. These can be adjusted

ian.simpson@ed.ac.uk

Scan Target Database for Matches and Extend

- take all words above threshold:
  - (LWG,IWG,MWG,VWG,FWG,AWG,LWS,LWM,LWA & LYG)
- scan the database for entries (“hits”) that match the list
- create a hash-table index of hit locations for each word
- perform gap-free extensions from the hits
- perform gapped extensions from the hits



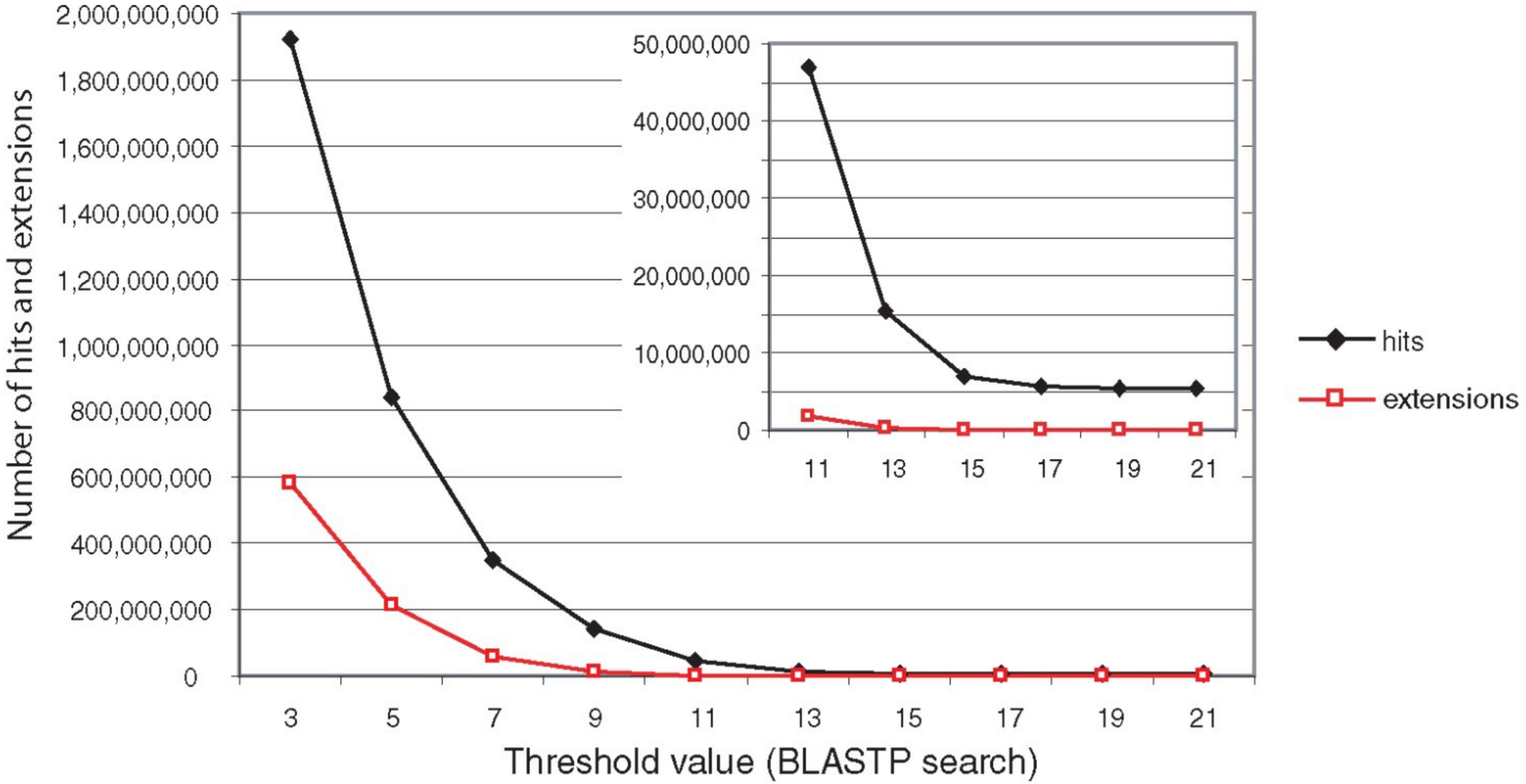
ian.simpson@ed.ac.uk

Perform Traceback to Generate Gapped Alignment

- for each hit-alignment calculate locations of:
  - insertions
  - deletions
  - matches
- apply **composition-based statistics**
  - (uses composition/distribution of sequences being aligned to calculate more accurate E-values)
- generate final gapped-alignment

ian.simpson@ed.ac.uk

Lower Threshold Yields More Hits & Extensions

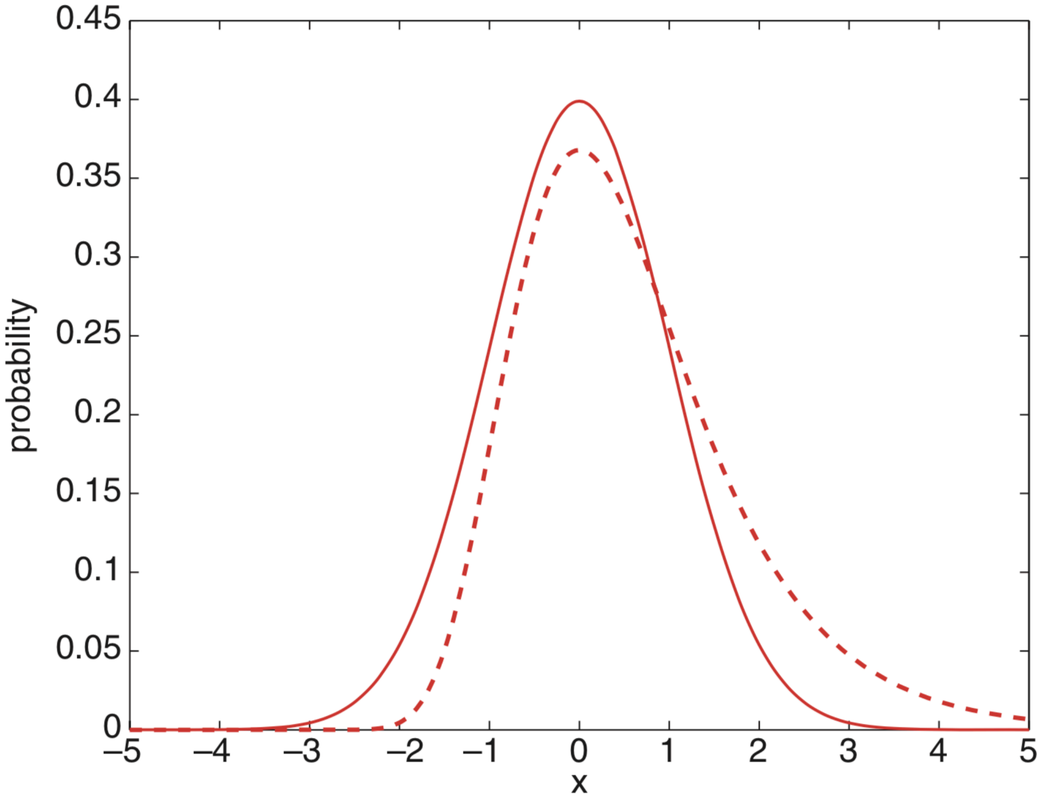


ian.simpson@ed.ac.uk

Interpreting a BLAST Search Result

- it is important to assess the statistical significance of search results
- global alignments - statistics are still very poorly understood
- for local alignments (including BLAST) the statistics are known
  - scores follow an extreme value distribution (EVD) rather than a normal distribution

$$E = Kmne^{-\lambda S}$$



where E is the “expectation”, K derived from search space, m & n are the sequence lengths of any two compared sequences, lambda is the decay constant of the EVD and S is the alignment bit-score

ian.simpson@ed.ac.uk

Properties of the equation $E = Kmne^{-\lambda S}$

- E decreases exponentially with increasing ***S*** (higher S values correspond to better alignments). Very high scores correspond to very low ***E*** values.
- ***E*** for aligning a pair of random sequences must be negative, otherwise long random alignments would acquire great scores
- parameter ***K*** describes the search space (database properties e.g. size, complexity)
- E is the expected number of HSP (highest-scoring pairs) that have an alignment score >= S that you expect to occur by chance in your database search. [looking at the equation as K increases (~database size) then E increases i.e. the larger the database the more high scoring hits you would expect by chance]

ian.simpson@ed.ac.uk

E-values & p-values

The expect value E is the number of alignments with scores greater than or equal to score S that are expected to occur by chance in a given database search.

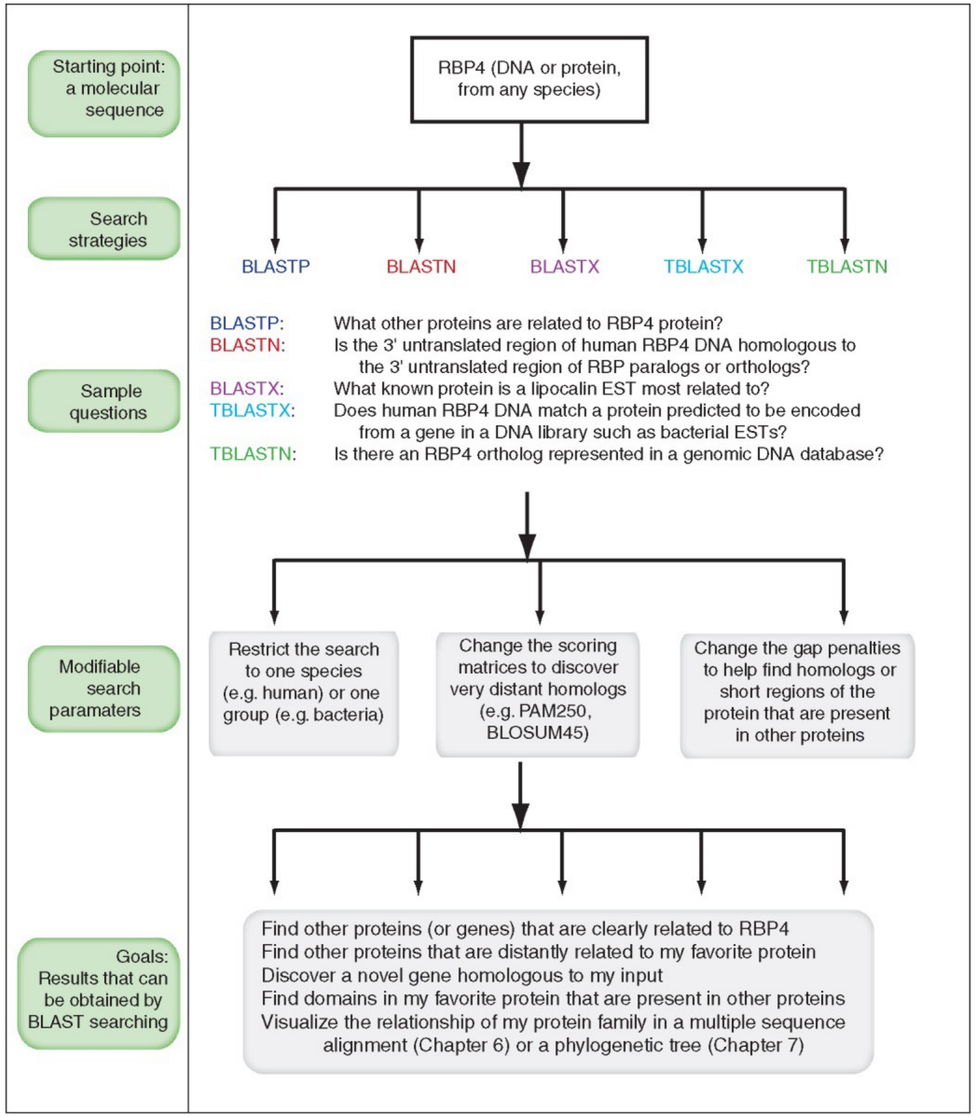
A p-value is a different way of representing the significance of an alignment.

$$p = 1 - e^{-E}$$

E	p
10	0.99995460
5	0.99326205
2	0.86466472
1	0.63212056
0.1	0.09516258
0.05	0.04877058
0.001	0.00099950
0.0001	0.0001000

ian.simpson@ed.ac.uk

BLAST Search Strategies



ian.simpson@ed.ac.uk

Problems Standard BLAST Cannot Solve

- Use human beta globin as a query against human RefSeq proteins, and BLASTP does not “find” human myoglobin. This is because the two proteins are too distantly related. **PSI-BLAST** at NCBI as well as hidden Markov models easily solve this problem

- How can we search using 10,000 base pairs as a query, or even millions of base pairs? Many BLAST-like tools for genomic DNA are available such as **PatternHunter**, **Megablast**, **BLAT**, and **LASTZ**

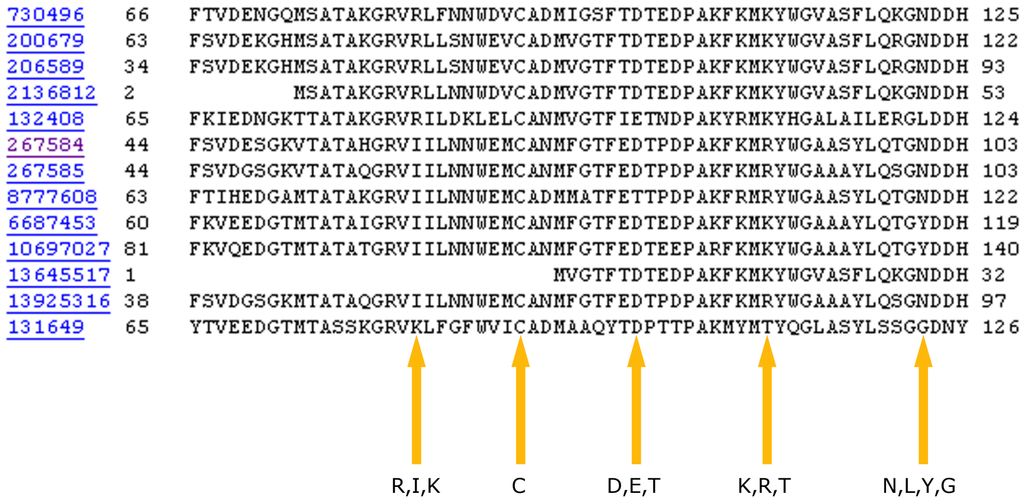
- How can we align tens of millions of short reads to a reference genome (e.g. read alignment)? Specialist algorithms such as **TopHat**, **STAR**, **Callisto**, **Salmon**

ian.simpson@ed.ac.uk

PSI-BLAST - Position specific iterated BLAST

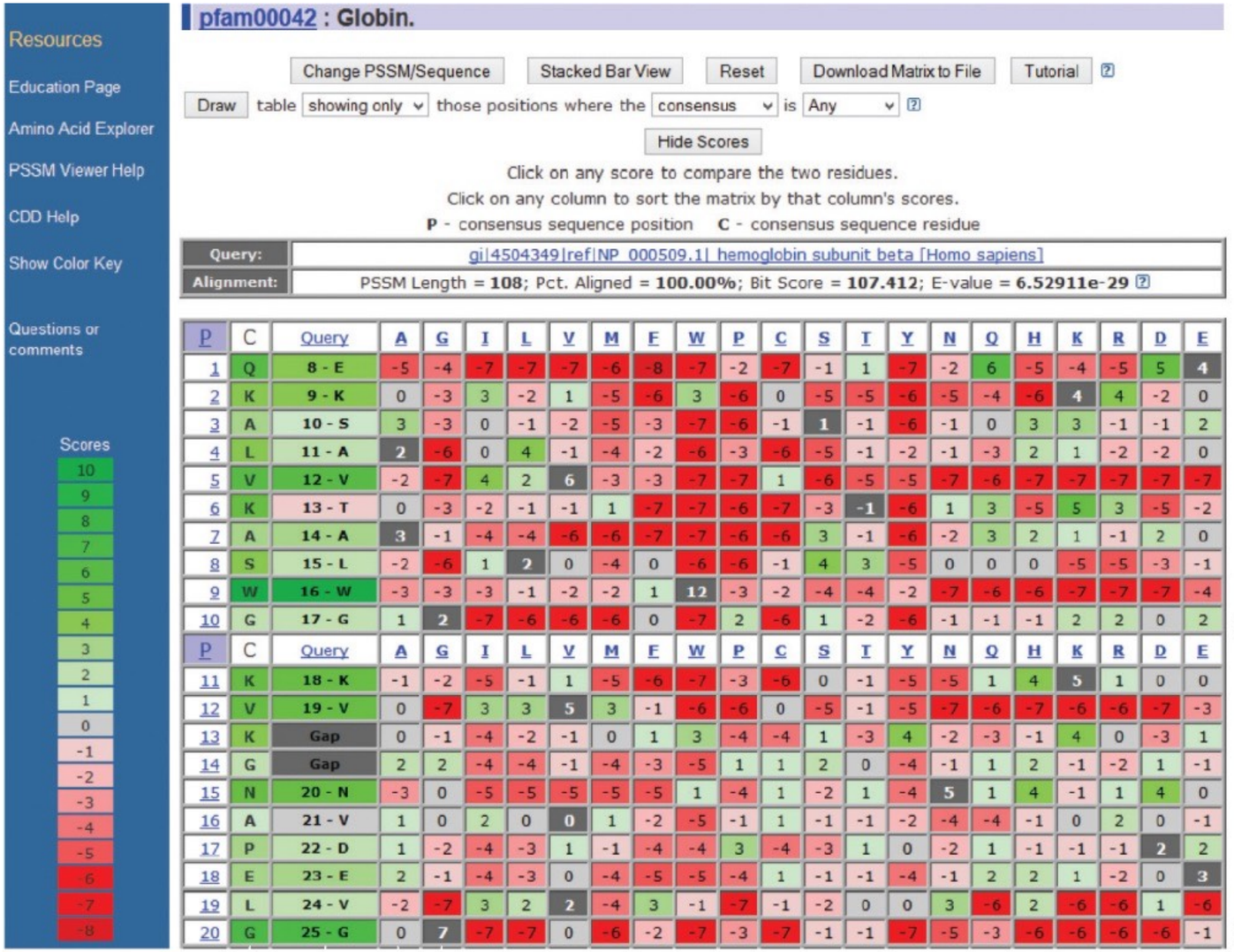
- PSI-BLAST searches deeper into the target database for matches to your query sequence by employing a scoring matrix that is customised to your query sequence

- It first does a BLAST search and from the resulting alignments PSI-BLAST constructs a multiple sequence alignment and then a “profile” or position-specific scoring matrix (PSSM) that describes it
  - (what this means is that depending where in your sequence the match is it may give a different score i.e. for the same aligned pair)



ian.simpson@ed.ac.uk

Position Specific Scoring Matrix (PSSM)



Resources

Education Page

Amino Acid Explorer

PSSM Viewer Help

CDD Help

Show Color Key

Questions or comments

Scores

10 9 8 7 6 5 4 3 2 1 0 -1 -2 -3 -4 -5 6 -7 -8

pfam00042 : Globin.

Change PSSM/Sequence | Stacked Bar View | Reset | Download Matrix to File | Tutorial

Draw table showing only those positions where the consensus is Any

Hide Scores

Click on any score to compare the two residues.

Click on any column to sort the matrix by that column's scores.

P - consensus sequence position **C** - consensus sequence residue

Query:	gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
Alignment:	PSSM Length = **108**; Pct. Aligned = **100.00%**; Bit Score = **107.412**; E-value = **6.52911e-29**

P	C	Query	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
1	Q	8 - E	-5	-4	-7	-7	-7	-6	-8	-7	-2	-7	-1	1	-7	-2	6	-5	-4	-5	5	4
2	K	9 - K	0	-3	3	-2	1	-5	-6	3	-6	0	-5	-5	-6	-5	-4	-6	4	4	-2	0
3	A	10 - S	3	-3	0	-1	-2	-5	-3	-7	-6	-1	1	-1	-6	-1	0	3	3	-1	-1	2
4	L	11 - A	2	-6	0	4	-1	-4	-2	-6	-3	-6	-5	-1	-2	-1	-3	2	1	-2	-2	0
5	V	12 - V	-2	-7	4	2	6	-3	-3	-7	-7	1	-6	-5	-5	-7	-6	-7	-7	-7	-7	-7
6	K	13 - T	0	-3	-2	-1	-1	1	-7	-7	-6	-7	-3	-1	-6	1	3	-5	5	3	-5	-2
7	A	14 - A	3	-1	-4	-4	-6	-6	-7	-7	-6	-6	3	-1	-6	-2	3	2	1	-1	2	0
8	S	15 - L	-2	-6	1	2	0	-4	0	-6	-6	-1	4	3	-5	0	0	0	-5	-5	-3	-1
9	W	16 - W	-3	-3	-3	-1	-2	-2	1	12	-3	-2	-4	-4	-2	-7	-6	-6	-7	-7	-7	-4
10	G	17 - G	1	2	-7	-6	-6	-6	0	-7	2	-6	1	-2	-6	-1	-1	-1	2	2	0	2
P	C	Query	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
11	K	18 - K	-1	-2	-5	-1	1	-5	-6	-7	-3	-6	0	-1	-5	-5	1	4	5	1	0	0
12	V	19 - V	0	-7	3	3	5	3	-1	-6	-6	0	-5	-1	-5	-7	-6	-7	-6	-6	-7	-3
13	K	Gap	0	-1	-4	-2	-1	0	1	3	-4	-4	1	-3	4	-2	-3	-1	4	0	-3	1
14	G	Gap	2	2	-4	-4	-1	-4	-3	-5	1	1	2	0	-4	-1	1	2	-1	-2	1	-1
15	N	20 - N	-3	0	-5	-5	-5	-5	-5	1	-4	1	-2	1	-4	5	1	4	-1	1	4	0
16	A	21 - V	1	0	2	0	0	1	-2	-5	-1	1	-1	-1	-2	-4	-4	-1	0	2	0	-1
17	P	22 - D	1	-2	-4	-3	1	-1	-4	-4	3	-4	-3	1	0	-2	1	-1	-1	-1	2	2
18	E	23 - E	2	-1	-4	-3	0	-4	-5	-5	-4	1	-1	-1	-4	-1	2	2	1	-2	0	3
19	L	24 - V	-2	-7	3	2	2	-4	3	-1	-7	-1	-2	0	0	3	-6	2	-6	-6	1	-6
20	G	25 - G	0	7	-7	-7	0	-6	-2	-7	-3	-7	-1	-1	-7	-5	-3	-6	-6	-6	-6	-1

ian.simpson@ed.ac.uk

PSI-BLAST - Position specific iterated BLAST

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	M	-1	-2	-2	-3	-2	-1	-2	-3	-2	1	2	-2	6	0	-3	-2	-1	-2	-1	1
2	K	-1	1	0	1	-4	2	4	-2	0	-3	-3	3	-2	-4	-1	0	-1	-3	-2	-3
3	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
4	V	0	-3	-3	-4	-1	-3	-3	-4	-4	3	1	-3	1	-1	-3	-2	0	-3	-1	4
5	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
6	A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
7	L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
8	L	-1	-3	-3	-4	-1	-3	-3	-4	-3	2	2	-3	1	3	-3	-2	-1	-2	0	3
9	L	-1	-3	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	2
10	L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
11	A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
12	A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
13	W	-2	-3	-4	-4	-2	-2	-3	-4	-3	1	4	-3	2	1	-3	-3	-2	7	0	0
14	A	3	-2	-1	-2	-1	-1	-2	4	-2	-2	-2	-1	-2	-3	-1	1	-1	-3	-3	-1
15	A	2	-1	0	-1	-2	2	0	2	-1	-3	-3	0	-2	-3	-1	3	0	-3	-2	-2
16	A	4	-2	-1	-2	-1	-1	-1	3	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	-1
...																					
37	S	2	-1	0	-1	-1	0	0	0	-1	-2	-3	0	-2	-3	-1	4	1	-3	-2	-2
38	G	0	-3	-1	-2	-3	-2	-2	6	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4
39	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-3	-2	0
40	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
41	Y	-2	-2	-2	-3	-3	-2	-2	-3	2	-2	-1	-2	-1	3	-3	-2	-2	2	7	-1
42	A	4	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0

note that a given amino acid (such as alanine) in your query protein can receive different scores for matching alanine—depending on the position in the protein

ian.simpson@ed.ac.uk

PSI-BLAST - Position specific iterated BLAST

- The PSSM is used as a query against the database
- PSI-BLAST estimates statistical significance (E values)
- Repeat iteratively, typically 5 times. At each new search, a new profile is used as the query.

After 1st iteration:
Expect = 4e-04
Alignment = 87 aa

After 2nd iteration:
Expect = 1e-36
Alignment = 110 aa

After 3rd iteration:
Expect = 2e-33
Alignment = 146 aa

(a) PSI-BLAST iteration 1 match (human beta globin versus a *C. albicans* globin)

hypothetical protein CaO19.4459 [Candida albicans SC5314]

Sequence ID: refXP_711954.1| Length: 563 Number of Matches: 1

See 1 more title(s)

Range 1: 338 to 424 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
43.5 bits(101)	4e-04	Composition-based stats.	24/87(28%)	42/87(48%)	3/87(3%)

```
Query  59   PKVKAHGKKVLGAFSDGLAHLDNLK---GTFATLSELHCDKLHVDPENFRLLGNVLVCVL  115
            P +K     + G  S  ++ L+NL       A L +LH   L+++  +F+L+G   V
Sbjct  338  PSIKHQAANMAGILSLTISQLENLSILDEYLAKLGKLHSRVLNIEEAHFKLMGEAFVQIF  397

Query  116  AHHFGKEFTPPVQAAYQKVVAGVANAL  142
             FG +FT  ++  + K+   +AN L
Sbjct  398  QERFGSKFTKELENLWIKLYLYIANTL  424
```

(b) PSI-BLAST iteration 2 (human beta globin versus a *C. albicans* globin)

Range 1: 315 to 424 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
136 bits(343)	1e-36	Composition-based stats.	27/110(25%)	48/110(43%)	6/110(5%)

```
Query  39   TQRFFESFG-DLST--PDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLK---GTFATLSEL  92
            +  F      +L +  P      P +K     + G  S  ++ L+NL       A L +L
Sbjct  315  SSLFCRQLYFNLLSKDPTLEKMFPSIKHQAANMAGILSLTISQLENLSILDEYLAKLGKL  374

Query  93   HCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANAL  142
            H   L+++  +F+L+G   V      FG +FT  ++  + K+   +AN L
Sbjct  375  HSRVLNIEEAHFKLMGEAFVQTFQERFGSKFTKELENLWIKLYLYIANTL  424
```

(c) PSI-BLAST iteration 3 (human beta globin versus a *C. albicans* globin)

Range 1: 281 to 426 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
128 bits(321)	2e-33	Composition-based stats.	28/146(19%)	50/146(34%)	6/146(4%)

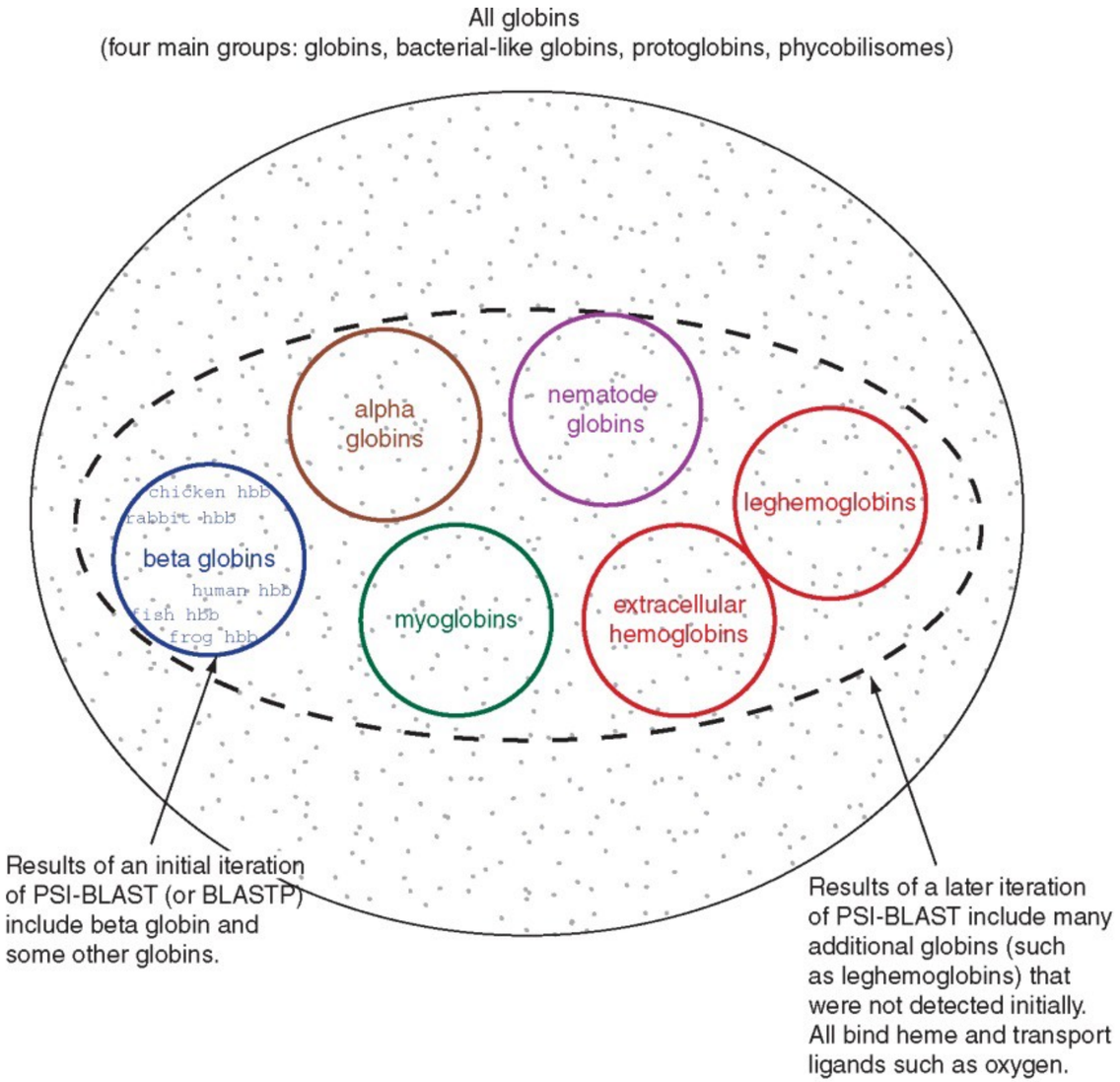
```
Query  5    TPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLS---TPDAVMGNPKV  61
            +          +           + RL      + F           P      P +
Sbjct  281  SRRRIIKRKSSRNVNGSGSTNTNTMTRLDSTTIASSLFCRQLYFNLLSKDPTLEKMFPSI  340

Query  62   KAHGKKVLGAFSDGLAHLDNLK---GTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHH  118
            K     + G  S  ++ L+NL       A L +LH   L+++  +F+L+G   V
Sbjct  341  KHQAANMAGILSLTISQLENLSILDEYLAKLGKLHSRVLNIEEAHFKLMGEAFVQTFQER  400

Query  119  FGKEFTPPVQAAYQKVVAGVANALAH  144
            FG +FT  ++  + K+   +AN L
Sbjct  401  FGSKFTKELENLWIKLYLYIANTLLQ  426
```


ian.simpson@ed.ac.uk

PSI-BLAST increases search sensitivity by detecting homologous matches with relatively low sequence identity



ian.simpson@ed.ac.uk

PSI-BLAST & PSSM Corruption

- Once a match is incorporated into the PSSM it will never be removed and may lead to the inclusion of many other related false positive hits.
- There are three main approaches to removing these false positives
  - Filter biased amino acid regions (this is a BLAST option)
  - lower the expect value threshold to make the search more stringent
  - visually inspect the output from each PSI-BLAST iteration and remove suspicious matches

ian.simpson@ed.ac.uk

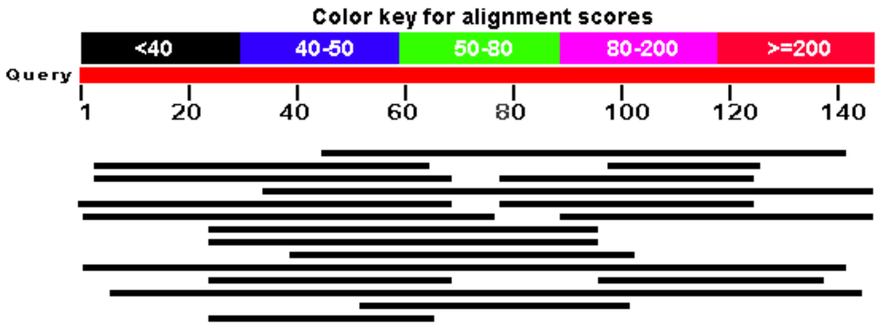
DELTA-BLAST

- In 2012 NCBI introduced DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST) to the family of BLASTP tools
- DELTA-BLAST constructs a PSSM using the results of a Conserved Domain Database (CDD) search, and uses that to search a sequence database
- The results are typically superior to those of PSI-BLAST
- PSI-BLAST creates multiple alignments and position-specific scoring matrices (PSSMs) whereas DELTA-BLAST searches a query against a library of pre-computed PSSMs. One reason DELTA-BLAST outperforms PSI-BLAST is that it results in larger, more complete PSSMs than PSI-BLAST
- One iteration of DELTA-BLAST is recommended.

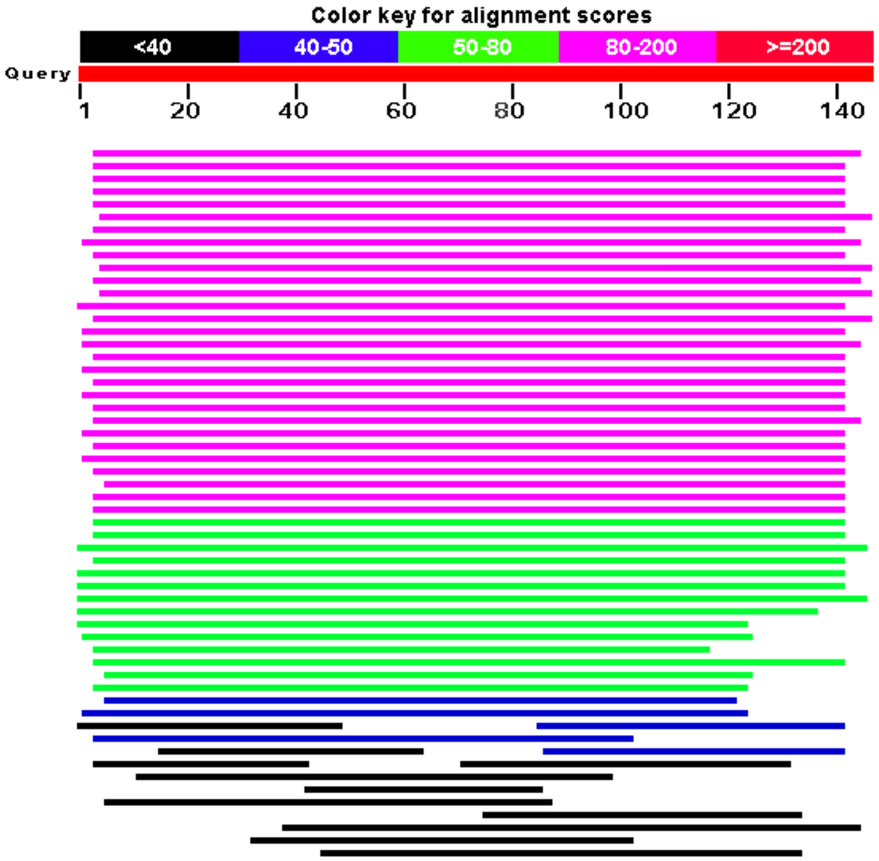
ian.simpson@ed.ac.uk

BLAST vs DELTA-BLAST

BLAST



DELTA-BLAST



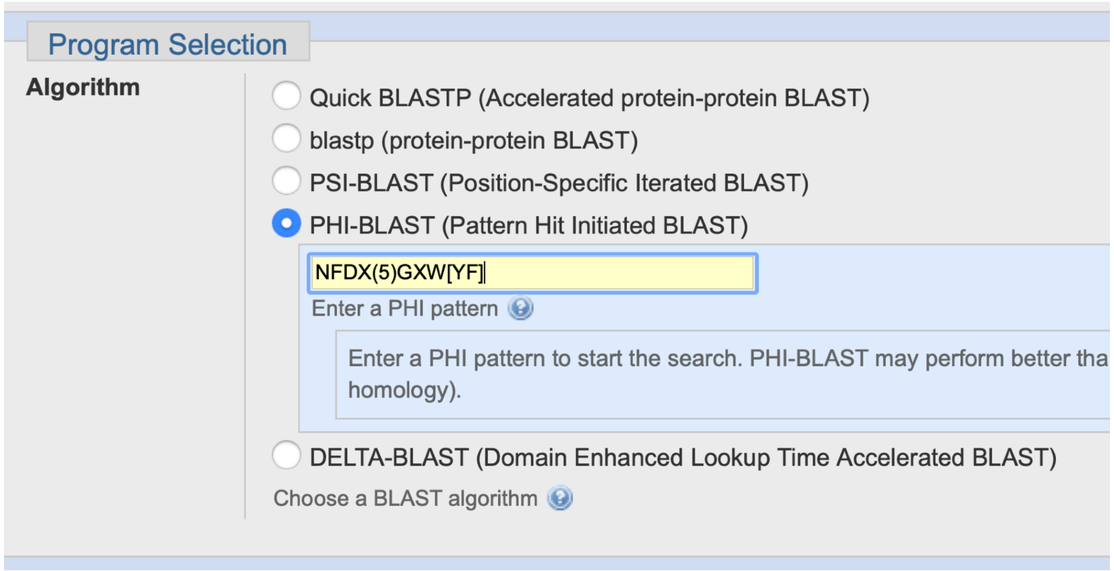
ian.simpson@ed.ac.uk

Assessing DELTA-BLAST & PSI-BLAST

- to assess the performance of BLASTP, PSI-BLAST, DELTA-BLAST or other programs it is necessary to have a “truth” dataset to distinguish true positives, false positives, true negatives, and false negatives.
- perform searches against databases that incorporate structural information to define homology.
- evaluate PSI-BLAST or other programs’ results using a database in which protein structures have been solved and all proteins in a group share < 40% amino acid identity.

ian.simpson@ed.ac.uk

PHI-BLAST - Pattern Hit Initiated BLAST



- sometimes you have a protein query that has a known pattern. You can use PHI-BLAST to include that pattern, which can be user-selected or obtained from a database of such patterns such as PROSITE
- all resulting database matches must include that pattern
- PHI-BLAST is specialised, and is not commonly used but can be very useful.

ian.simpson@ed.ac.uk

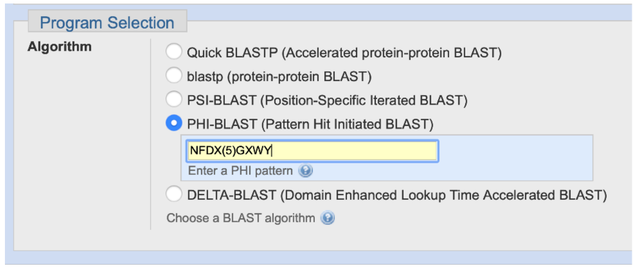
Performing a PHI-BLAST Search

Alignment

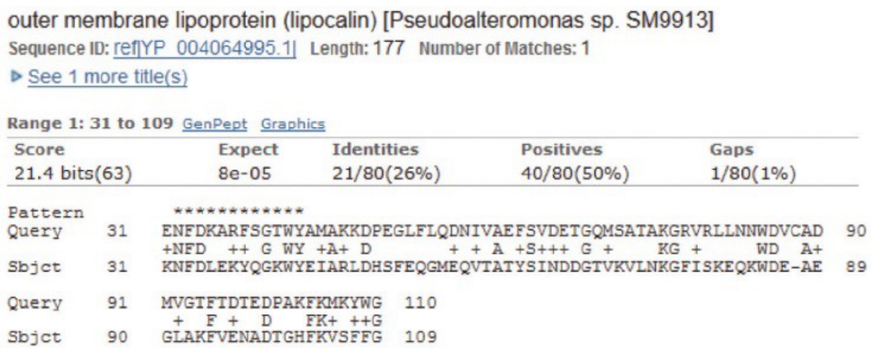
```
MUSCLE (3.8) multiple sequence alignment

NP_006735.2      -MKWVWALLLLAALGSGRAERDCRVSSFRVK--ENFDKARFSGTWYAMAKK
WP_010388720.1   ---MKLAFKTALFITAMFLLSACTSAPEGITPVKNFDLEKYQGKWYEIARL
WP_008992866.1   MKAKNKILIAACAIGLGALLNSCASIPKNAKAVKNFDIDRYLGTWYEIARF
YP_003021245.1   -MKKLSLLLSLLFTG-------CVGIPENVKPVDNFDVHRYLGKWYEIARL
                           :              *   .   .  .***  .: *.**  :*.
```

Pattern Definition



PHI_BLAST Search Result



ian.simpson@ed.ac.uk

BLAST-like Tools for Genomic DNA

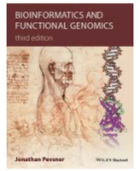
- The analysis of genomic DNA presents special challenges:
  - exons (protein-coding sequence)
  - introns (intervening sequences).
  - sequencing errors or polymorphisms
  - comparison between species (e.g. human and mouse)
- Recently developed tools include:
  - MegaBLAST
  - **BLAT** (BLAST-like alignment tool). BLAT parses an entire genomic DNA database into 11-mers, then searches them against a query (like an inverse BLAST)
  - **SSAHA** uses a similar strategy to BLAT

ian.simpson@ed.ac.uk

Next Week

Week 4 - Multiple Sequence Alignment

If you would like to read ahead then please look at the following chapter, if not we will cover this in next week's lecture
This is available from the Bio1 course "Resource List"



BOOK **Bioinformatics and functional genomics**

Pevsner, Jonathan, 1961-, 3rd ed.., Chichester, West Sussex, UK ; Hoboken, NJ, USA, John Wiley and Sons, Incorporated, 2015

Note: *Read Chapter 6, "Multiple Sequence Alignment".*

Add tags to item

Complete

ian.simpson@ed.ac.uk