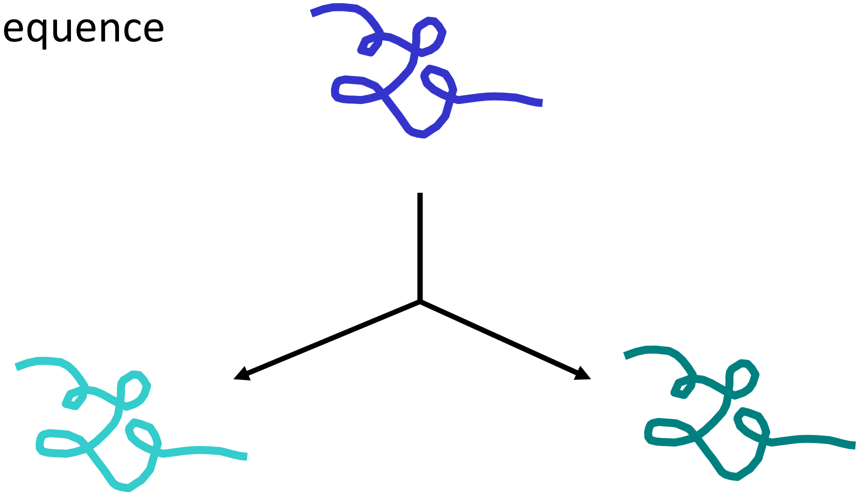
Gene Sequence Analysis

Lecture 2: **Pairwise Alignment**

30/05/2024
Phuc-Loi Luu, PhD

Adapted from Dr. Morgan Langille’s Lecture

Importance of Similarity

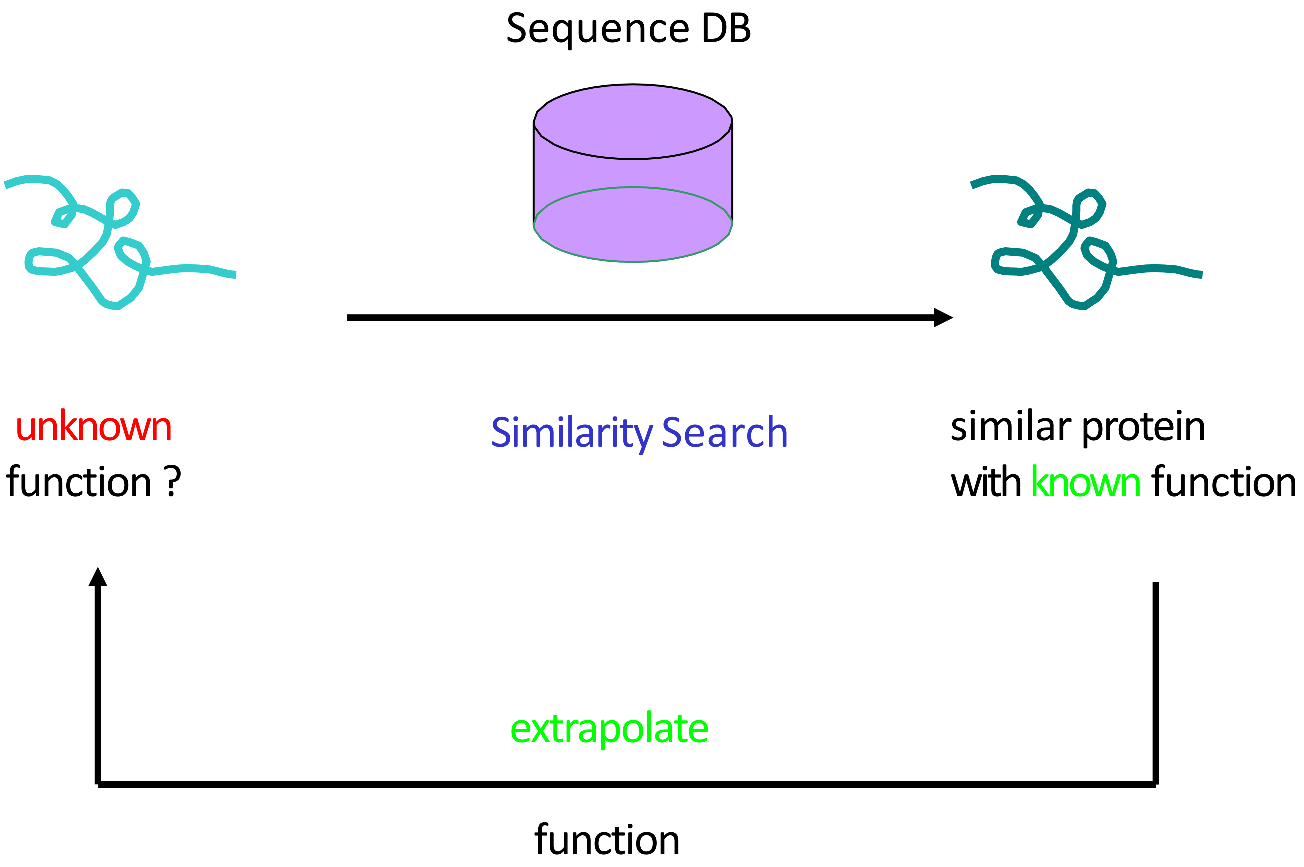


Ancestral protein/gene sequence

Similar (homologous) protein/gene sequences

Similar sequences: probably have the same ancestor, share the same structure, and have a similar biological function

Importance of Similarity



Sequence DB

unknown
function ?

Similarity Search

similar protein
with known function

extrapolate

function

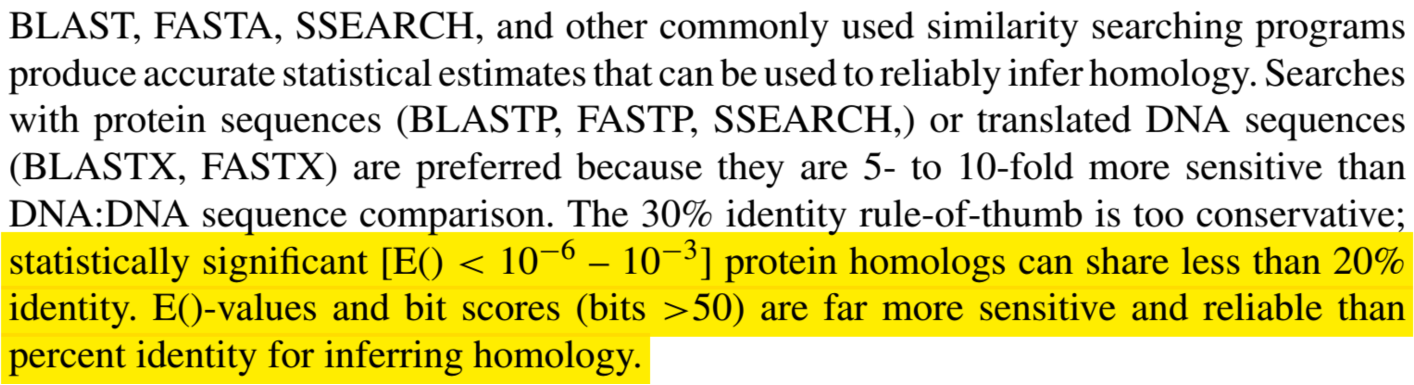
Importance of Similarity

Rule-of-thumb:
If your sequences are more than **100 amino acids** long (or 100 nucleotides long) you can considered them as homologues if **25%** of the **aa** are identical (**70%** of **nucleotide** for DNA). Below this value you enter the twilight zone.

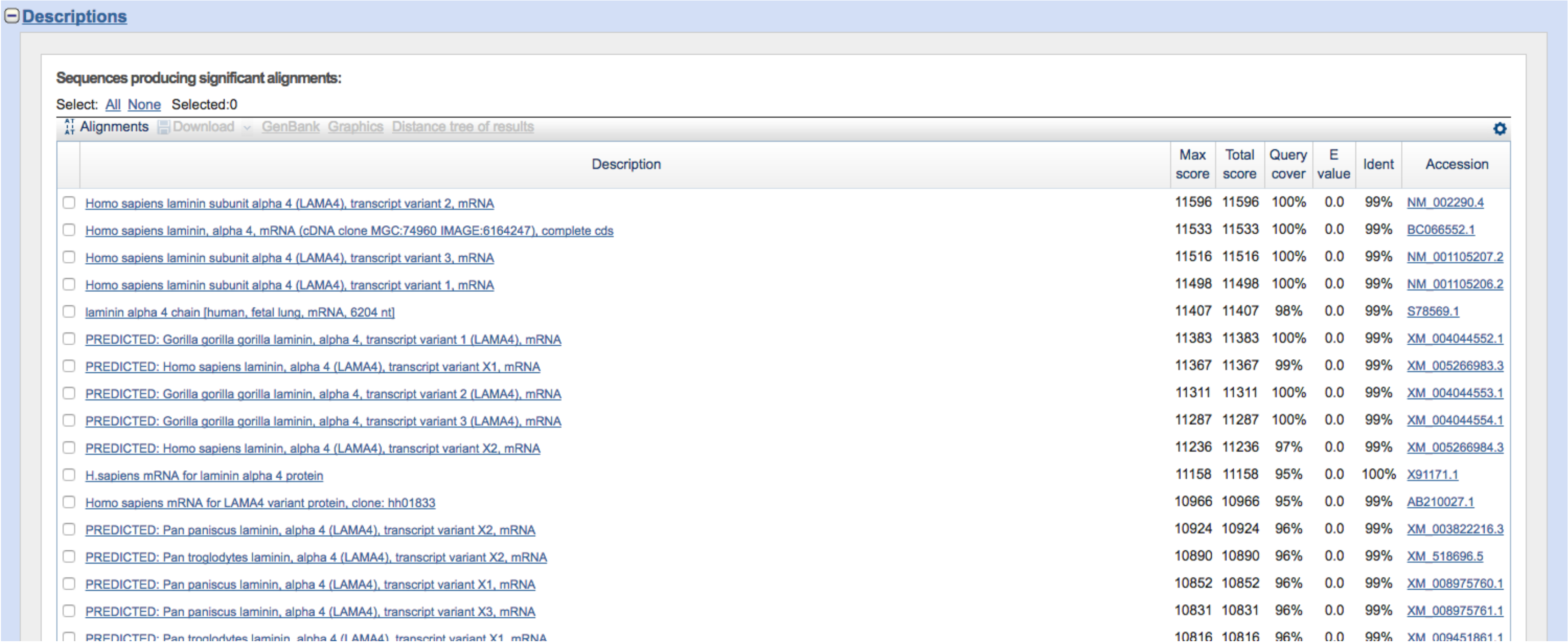
Twilight zone = protein sequence similarity between ~0-20% identity: is not statistically significant, i.e. could have arisen by chance.

Beware:

- E-value (*Expectation value*)
- length of the segments similar between the two sequences
- The number of insertions/deletions



BLAST, FASTA, SSEARCH, and other commonly used similarity searching programs produce accurate statistical estimates that can be used to reliably infer homology. Searches with protein sequences (BLASTP, FASTP, SSEARCH,) or translated DNA sequences (BLASTX, FASTX) are preferred because they are 5- to 10-fold more sensitive than DNA:DNA sequence comparison. The 30% identity rule-of-thumb is too conservative; statistically significant [E() $< 10^{-6} - 10^{-3}$] protein homologs can share less than 20% identity. E()-values and bit scores (bits >50) are far more sensitive and reliable than percent identity for inferring homology.

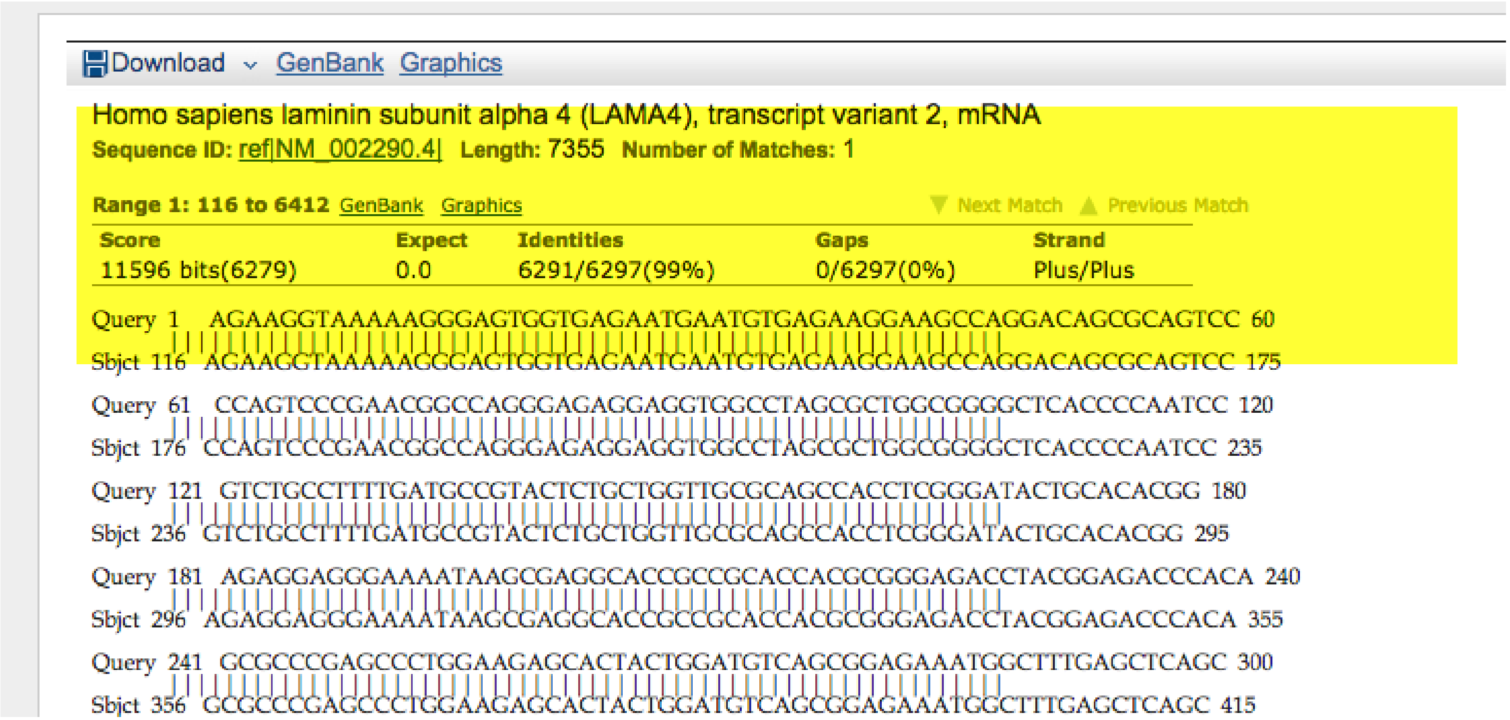


⊖ Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	Homo sapiens laminin subunit alpha 4 (LAMA4), transcript variant 2, mRNA	11596	11596	100%	0.0	99%	NM_002290.4
□	Homo sapiens laminin, alpha 4, mRNA (cDNA clone MGC:74960 IMAGE:6164247), complete cds	11533	11533	100%	0.0	99%	BC066552.1
□	Homo sapiens laminin subunit alpha 4 (LAMA4), transcript variant 3, mRNA	11516	11516	100%	0.0	99%	NM_001105207.2
□	Homo sapiens laminin subunit alpha 4 (LAMA4), transcript variant 1, mRNA	11498	11498	100%	0.0	99%	NM_001105206.2
□	laminin alpha 4 chain [human, fetal lung, mRNA, 6204 nt]	11407	11407	98%	0.0	99%	S78569.1
□	PREDICTED: Gorilla gorilla gorilla laminin, alpha 4, transcript variant 1 (LAMA4), mRNA	11383	11383	100%	0.0	99%	XM_004044552.1
□	PREDICTED: Homo sapiens laminin, alpha 4 (LAMA4), transcript variant X1, mRNA	11367	11367	99%	0.0	99%	XM_005266983.3
□	PREDICTED: Gorilla gorilla gorilla laminin, alpha 4, transcript variant 2 (LAMA4), mRNA	11311	11311	100%	0.0	99%	XM_004044553.1
□	PREDICTED: Gorilla gorilla gorilla laminin, alpha 4, transcript variant 3 (LAMA4), mRNA	11287	11287	100%	0.0	99%	XM_004044554.1
□	PREDICTED: Homo sapiens laminin, alpha 4 (LAMA4), transcript variant X2, mRNA	11236	11236	97%	0.0	99%	XM_005266984.3
□	H.sapiens mRNA for laminin alpha 4 protein	11158	11158	95%	0.0	100%	X91171.1
□	Homo sapiens mRNA for LAMA4 variant protein, clone: hh01833	10966	10966	95%	0.0	99%	AB210027.1
□	PREDICTED: Pan paniscus laminin, alpha 4 (LAMA4), transcript variant X2, mRNA	10924	10924	96%	0.0	99%	XM_003822216.3
□	PREDICTED: Pan troglodytes laminin, alpha 4 (LAMA4), transcript variant X2, mRNA	10890	10890	96%	0.0	99%	XM_518696.5
□	PREDICTED: Pan paniscus laminin, alpha 4 (LAMA4), transcript variant X1, mRNA	10852	10852	96%	0.0	99%	XM_008975760.1
□	PREDICTED: Pan paniscus laminin, alpha 4 (LAMA4), transcript variant X3, mRNA	10831	10831	96%	0.0	99%	XM_008975761.1
□	PREDICTED: Pan troglodytes laminin, alpha 4 (LAMA4), transcript variant X1, mRNA	10816	10816	96%	0.0	99%	XM_009451861.1



Homo sapiens laminin subunit alpha 4 (LAMA4), transcript variant 2, mRNA

Sequence ID: ref|NM_002290.4| Length: 7355 Number of Matches: 1

Range 1: 116 to 6412 GenBank Graphics

Score	Expect	Identities	Gaps	Strand
11596 bits(6279)	0.0	6291/6297(99%)	0/6297(0%)	Plus/Plus

```
Query  1    AGAAGGTAAAAAGGAGTGGTGAGAATGAATGTGAGAAGGAAGCCAGGACAGCGCAGTCC  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  116  AGAAGGTAAAAAGGAGTGGTGAGAATGAATGTGAGAAGGAAGCCAGGACAGCGCAGTCC  175

Query  61   CCAGTCCCGAACGGCCAGGGAGAGGAGGTGGCCTAGCGCTGGCGGGGCTCACCCCAATCC  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  176  CCAGTCCCGAACGGCCAGGGAGAGGAGGTGGCCTAGCGCTGGCGGGGCTCACCCCAATCC  235

Query  121  GTCTGCCTTTTGATGCCGTACTCTGCTGGTTGCGCAGCCACCTCGGGATACTGCACACGG  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  236  GTCTGCCTTTTGATGCCGTACTCTGCTGGTTGCGCAGCCACCTCGGGATACTGCACACGG  295

Query  181  AGAGGAGGGAAAATAAGCGAGGCACCGCCGCACCACGCGGGAGACCTACGGAGACCCACA  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  296  AGAGGAGGGAAAATAAGCGAGGCACCGCCGCACCACGCGGGAGACCTACGGAGACCCACA  355

Query  241  GCGCCCGAGCCCTGGAAGAGCACTACTGGATGTCAGCGGAGAAATGGCTTTGAGCTCAGC  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  356  GCGCCCGAGCCCTGGAAGAGCACTACTGGATGTCAGCGGAGAAATGGCTTTGAGCTCAGC  415
```

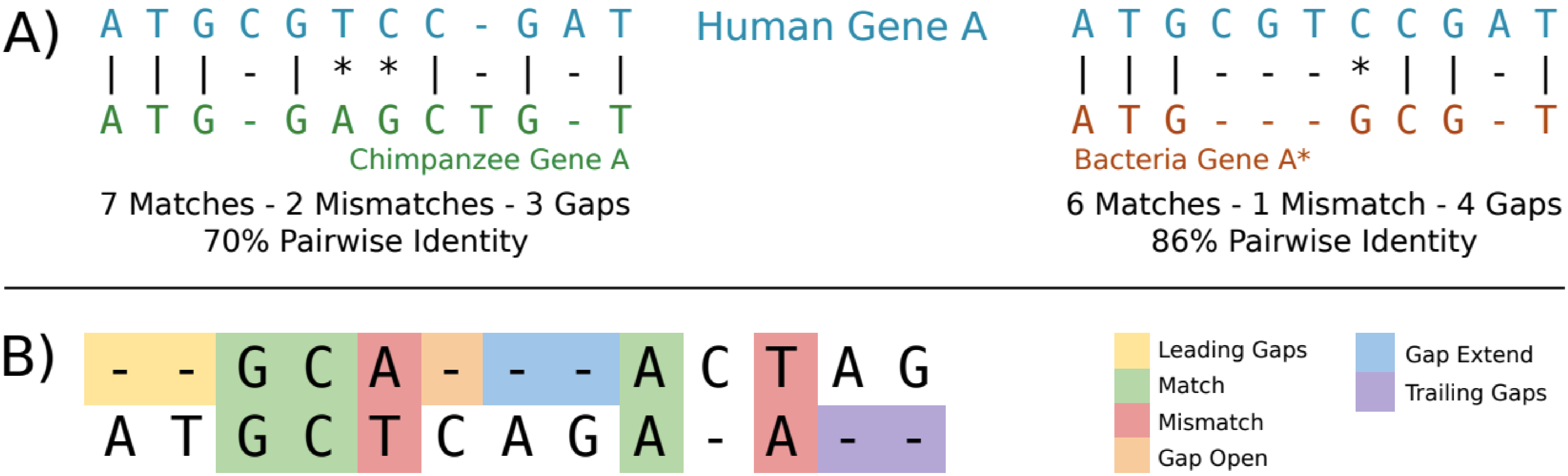
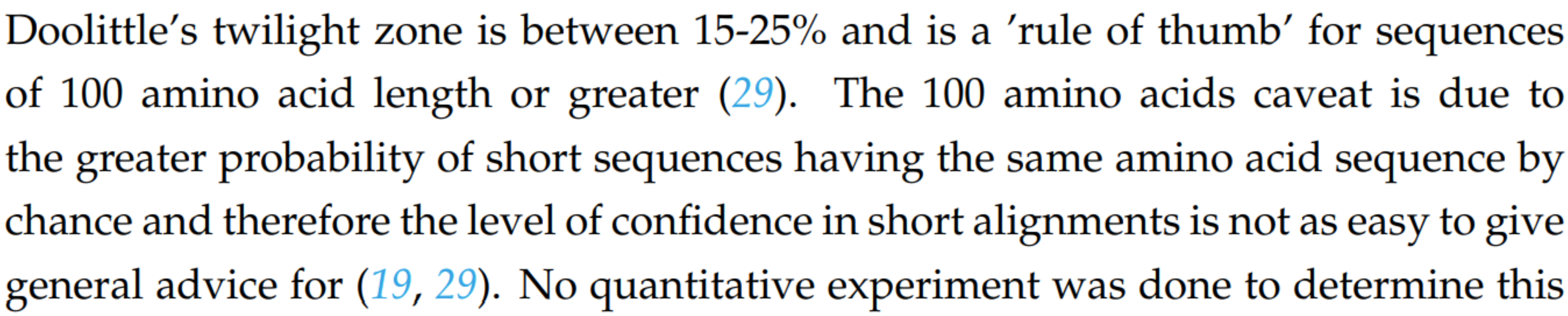
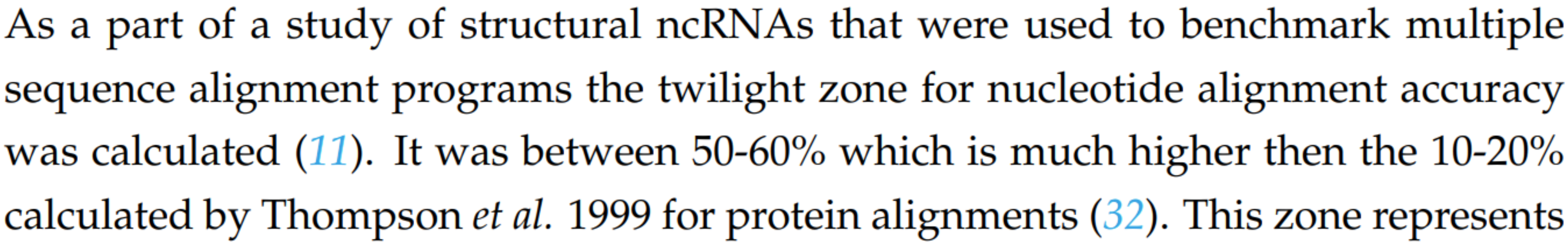


FIGURE 1.2: **Similarity versus homology and alignment scoring:** *A)* Two sequence alignments to Human Gene A highlighting how true homologs (Chimpanzee Gene A) can have similar similarity to non-homologous sequences (Bacteria Gene A*). *B)* A sequence alignment with different reward or penalties highlighted to illustrate how homology search algorithms score alignments based on sequence similarity.

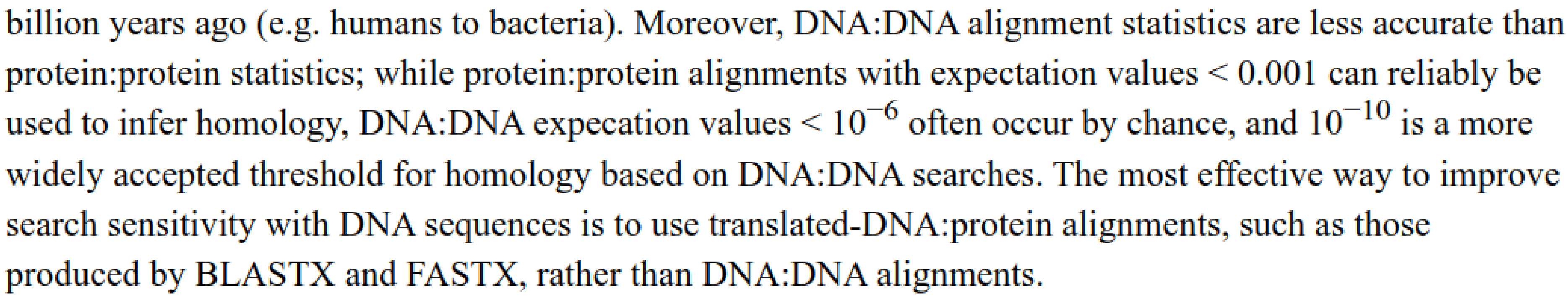
https://ourarchive.otago.ac.nz/bitstream/handle/10523/9302/McGimpseyStephanie2019MSc.pdf



Doolittle's twilight zone is between 15-25% and is a 'rule of thumb' for sequences of 100 amino acid length or greater (29). The 100 amino acids caveat is due to the greater probability of short sequences having the same amino acid sequence by chance and therefore the level of confidence in short alignments is not as easy to give general advice for (19, 29). No quantitative experiment was done to determine this



As a part of a study of structural ncRNAs that were used to benchmark multiple sequence alignment programs the twilight zone for nucleotide alignment accuracy was calculated (11). It was between 50-60% which is much higher then the 10-20% calculated by Thompson *et al.* 1999 for protein alignments (32). This zone represents



billion years ago (e.g. humans to bacteria). Moreover, DNA:DNA alignment statistics are less accurate than protein:protein statistics; while protein:protein alignments with expectation values < 0.001 can reliably be used to infer homology, DNA:DNA expecation values $< 10^{-6}$ often occur by chance, and 10^{-10} is a more widely accepted threshold for homology based on DNA:DNA searches. The most effective way to improve search sensitivity with DNA sequences is to use translated-DNA:protein alignments, such as those produced by BLASTX and FASTX, rather than DNA:DNA alignments.

https://ourarchive.otago.ac.nz/bitstream/handle/10523/9302/McGimpseyStephanie2019MSc.pdf
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3820096/

Outline

- PSSMs/PSI-BLAST
- HMMs/HMMer RNA Alignments
- Genome Alignments
- Assemblers Mappers

Different tools for homology searching

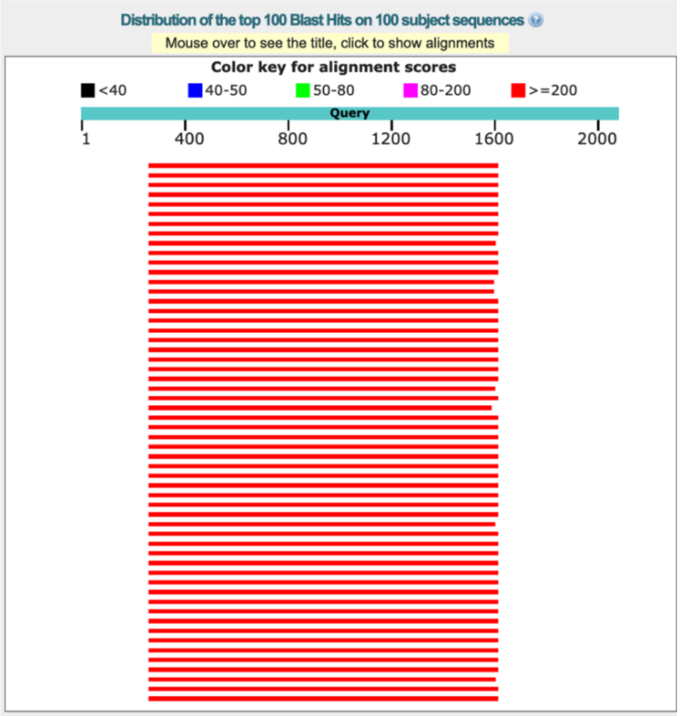
- Searching for protein families
- Aligning genomes
- Looking for RNA genes
- Combining overlapping sequences (assemblers)
- Finding the position of a sequence in a genome

One tool does not do it all

- Blast may give you an answer

    - BUT you could find the answer much quicker or with more precision by using the right tool!

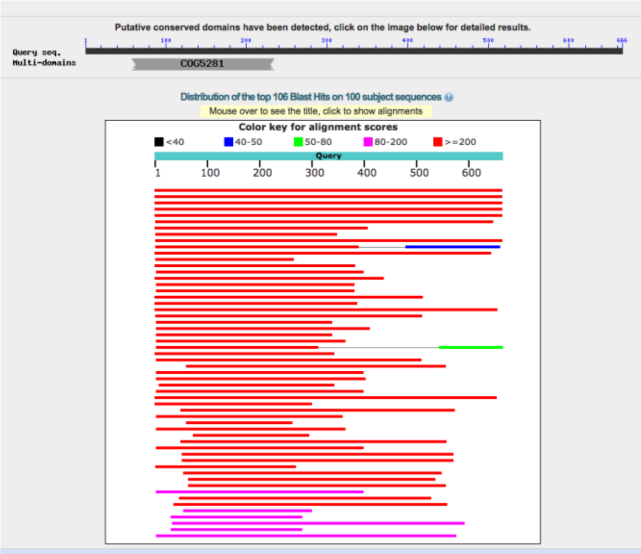
Typical BLAST output



Distribution of the top 100 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments

Color key for alignment scores

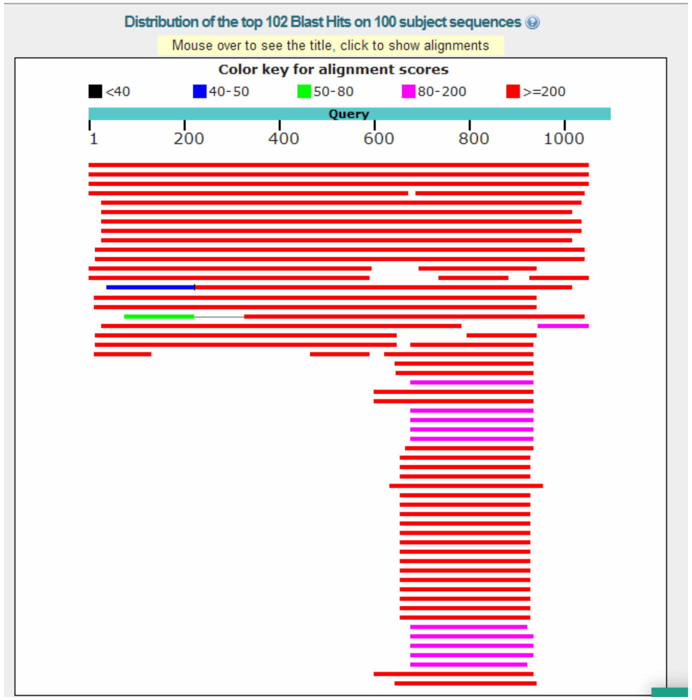


Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of the top 106 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments

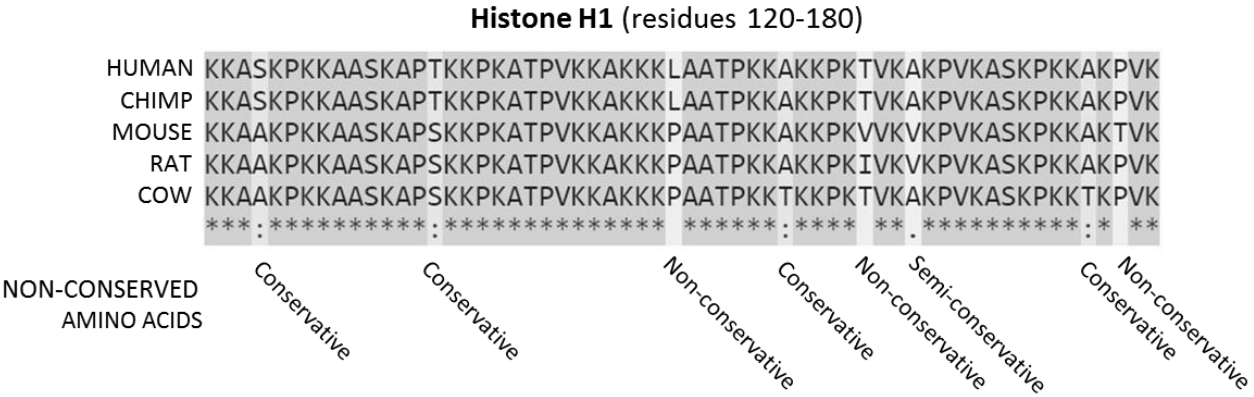
Color key for alignment scores



Distribution of the top 102 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments

Color key for alignment scores

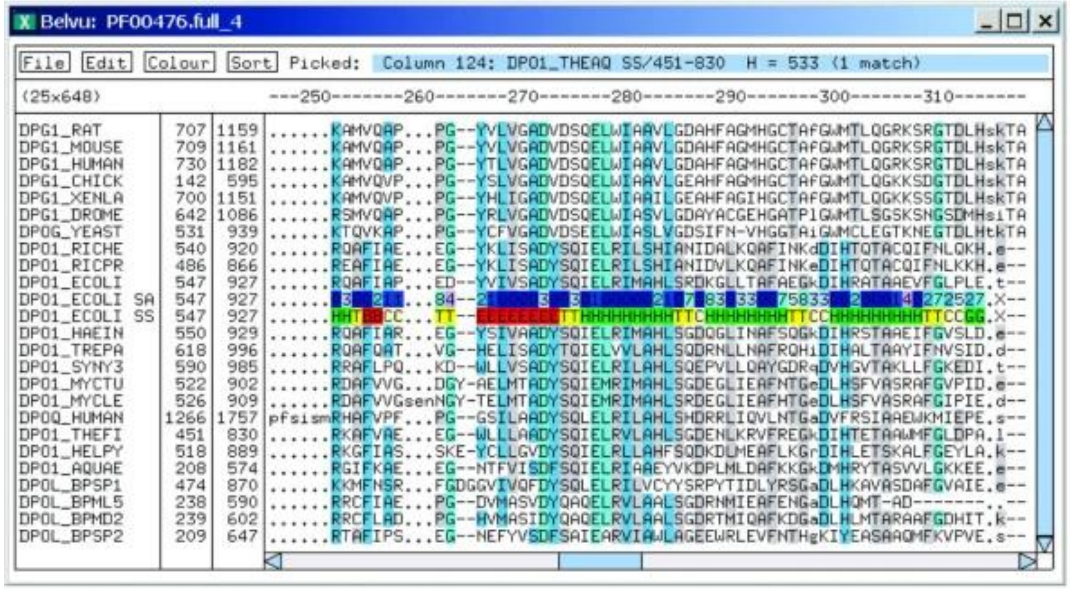


Histone H1 (residues 120-180)

```
HUMAN KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKAKPVKASKPKKAKPVK
CHIMP KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKAKPVKASKPKKAKPVK
MOUSE KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKVVKVKPVKASKPKKAKTVK
  RAT KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKIVKVKPVKASKPKKAKPVK
  COW KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKTKKPKTVKAKPVKASKPKKTKPVK
      ***:**********:************** ******:**** **.**********:* **
```

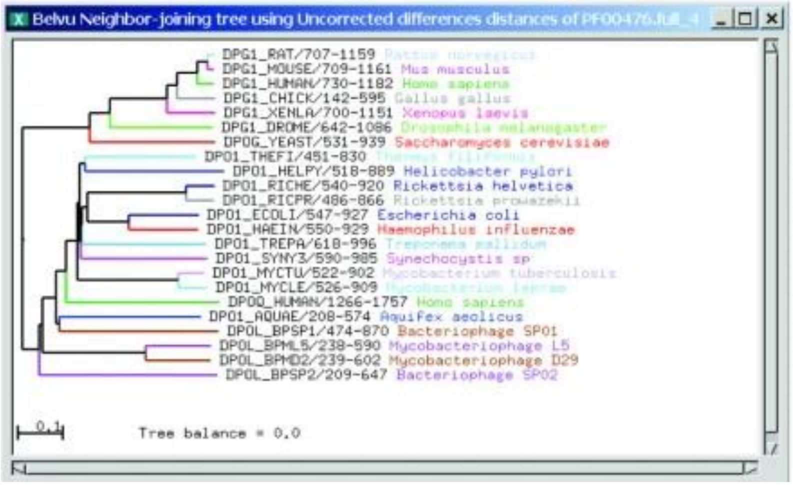
https://en.wikipedia.org/wiki/Sequence_alignment

Distance related proteins



```
Picked:  Column 124: DPO1_THEAQ SS/451-830   H = 533 (1 match)

(25x648)                      ---250--------260--------270--------280--------290--------300--------310-------

DPG1_RAT          707 1159 .......KAMVQAP...PG--YVLVGADVDSQELWIAAVLGDAHFAGMHGCTAFGWMTLQGRKSRGTDLHskTA
DPG1_MOUSE        709 1161 .......KAMVQAP...PG--YVLVGADVDSQELWIAAVLGDAHFAGMHGCTAFGWMTLQGRKSRGTDLHskTA
DPG1_HUMAN        730 1182 .......KAMVQAP...PG--YTLVGADVDSQELWIAAVLGDAHFAGMHGCTAFGWMTLQGRKSRGTDLHskTA
DPG1_CHICK        142  595 .......KAMVQVP...PG--YSLVGADVDSQELWIAAVLGEAHFAGMHGCTAFGWMTLQGKKSDGTDLHskTA
DPG1_XENLA        700 1151 .......KAMVQVP...PG--YHLIGADVDSQELWIAAILGEAHFAGIHGCTAFGWMTLQGKKSSGTDLHskTA
DPG1_DROME        642 1086 .......RSMVQAP...PG--YRLVGADVDSQELWIASVLGDAYACGEHGATPlGWMTLSGSKSNGSDMHsiTA
DPOG_YEAST        531  939 .......KTQVKAP...PG--YCFVGADVDSEELWIASLVGDSIFN-VHGGTAiGWMCLEGTKNEGTDLHtkTA
DPO1_RICHE        540  920 .......RQAFIAE...EG--YKLISADYSQIELRILSHIANIDALKQAFINKdDIHTQTACQIFNLQKH.e--
DPO1_RICPR        486  866 .......REAFIAE...EG--YKLISADYSQIELRILSHIANIDVLKQAFINKeDIHTQTACQIFNLKKH.e--
DPO1_ECOLI        547  927 .......RQAFIAP...ED--YVIVSADYSQIELRIMAHLSRDKGLLTAFAEGkDIHRATAAEVFGLPLE.t--
DPO1_ECOLI SA     547  927 ....... 3  21...84--2     3  3      2  7 83 33  75833  2    4 272527.X--
DPO1_ECOLI SS     547  927 .......HHTEECC...TT--EEEEEEEETTHHHHHHHHHTTCHHHHHHHHHTTCCHHHHHHHHHHTTCCGG.X--
DPO1_HAEIN        550  929 .......RQAFIAR...EG--YSIVAADYSQIELRIMAHLSGDQGLINAFSQGkDIHRSTAAEIFGVSLD.e--
DPO1_TREPA        618  996 .......RQAFQAT...VG--HELISADYTQIELVVLAHLSQDRNLLNAFRQHiDIHALTAAYIFNVSID.d--
DPO1_SYNY3        590  985 .......RRAFLPQ...KD--WLLVSADYSQIELRILAHLSQEPVLLQAYGDRqDVHGVTAKLLFGKEDI.t--
DPO1_MYCTU        522  902 .......RDAFVVG...DGY-AELMTADYSQIEMRIMAHLSGDEGLIEAFNTGeDLHSFVASRAFGVPID.e--
DPO1_MYCLE        526  909 .......RDAFVVGsenNGY-TELMTADYSQIEMRIMAHLSRDEGLIEAFHTGeDLHSFVASRAFGIPIE.d--
DPOQ_HUMAN       1266 1757 pfsismRHAFVPF...PG--GSILAADYSQLELRILAHLSHDRRLIQVLNTGaDVFRSIAAEWKMIEPE.s--
DPO1_THEFI        451  830 .......RKAFVAE...EG--WLLLAADYSQIELRVLAHLSGDENLKRVFREGkDIHTETAAWMFGLDPA.l--
DPO1_HELPY        518  889 .......RKGFIAS...SKE-YCLLGVDYSQIELRLLAHFSQDKDLMEAFLKGrDIHLETSKALFGEYLA.k--
DPO1_AQUAE        208  574 .......RGIFKAE...EG--NTFVISDFSQIELRIAAEYVKDPLMLDAFKKGkDMHRYTASVVLGKKEE.e--
DPOL_BPSP1        474  870 .......KKMFNSR...FGDGGVIVQFDYSQLELRILVCYYSRPYTIDLYRSGaDLHKAVASDAFGVAIE.e--
DPOL_BPML5        238  590 .......RRCFIAE...PG--DVMASVDYQAQELRVLAALSGDRNMIEAFENGaDLHQMT-AD--------..--
DPOL_BPMD2        239  602 .......RRCFLAD...PG--HVMASIDYQAQELRVLAALSGDRTMIQAFKDGaDLHLMTARAAFGDHIT.k--
DPOL_BPSP2        209  647 .......RTAFIPS...EG--NEFYVSDFSAIEARVIAWLAGEEWRLEVFNTHgKIYEASAAQMFKVPVE.s--
```

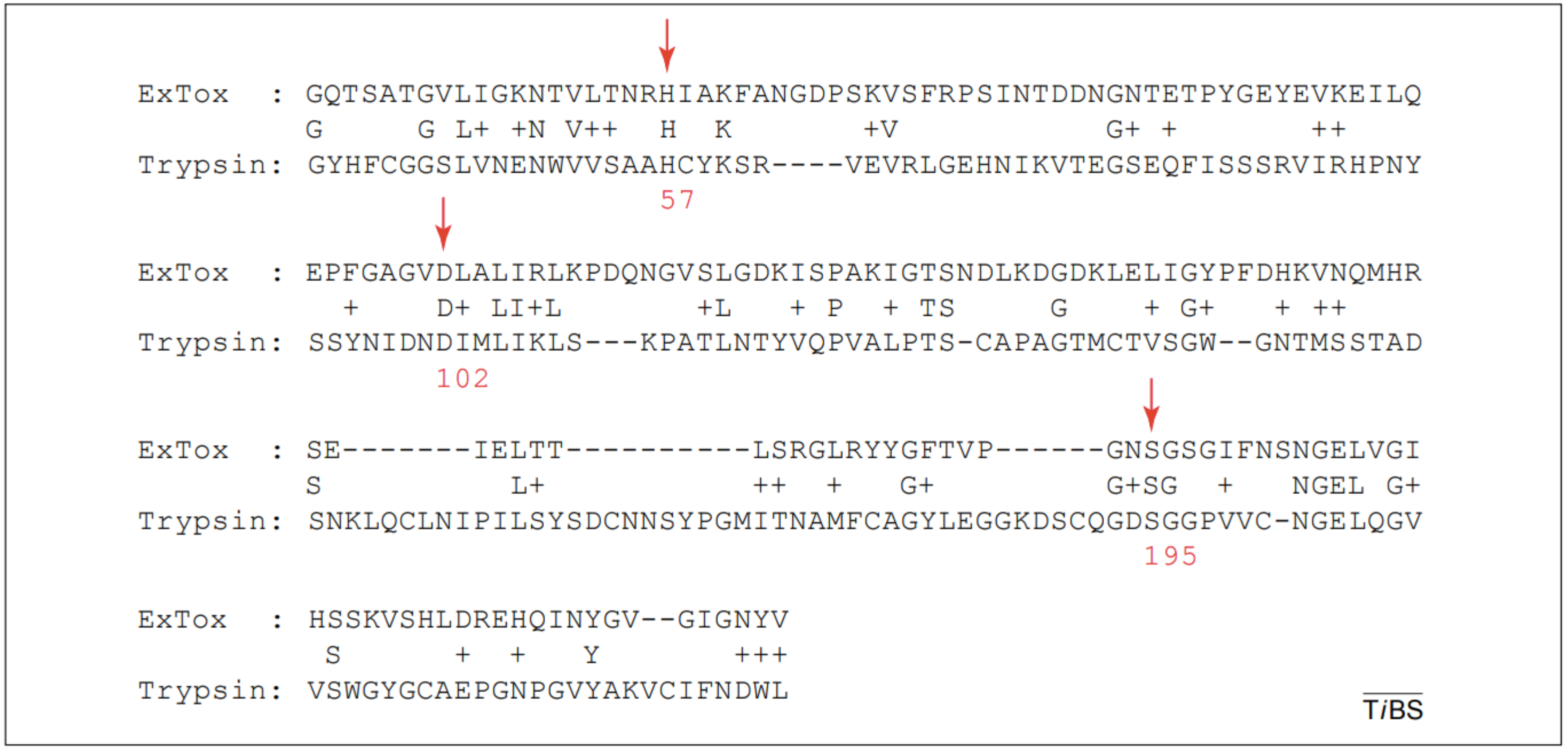
A



Belvu Neighbor-joining tree using Uncorrected differences distances of PF00476.full_4
DPG1_RAT/707-1159 Rattus norvegicus
DPG1_MOUSE/709-1161 Mus musculus
DPG1_HUMAN/730-1182 Homo sapiens
DPG1_CHICK/142-595 Gallus gallus
DPG1_XENLA/700-1151 Xenopus laevis
DPG1_DROME/642-1086 Drosophila melanogaster
DPOG_YEAST/531-939 Saccharomyces cerevisiae
DPO1_THEFI/451-830 Thermus filiformis
DPO1_HELPY/518-889 Helicobacter pylori
DPO1_RICHE/540-920 Rickettsia helvetica
DPO1_RICPR/486-866 Rickettsia prowazekii
DPO1_ECOLI/547-927 Escherichia coli
DPO1_HAEIN/550-929 Haemophilus influenzae
DPO1_TREPA/618-996 Treponema pallidum
DPO1_SYNY3/590-985 Synechocystis sp
DPO1_MYCTU/522-902 Mycobacterium tuberculosis
DPO1_MYCLE/526-909 Mycobacterium leprae
DPOQ_HUMAN/1266-1757 Homo sapiens
DPO1_AQUAE/208-574 Aquifex aeolicus
DPOL_BPSP1/474-870 Bacteriophage SP01
DPOL_BPML5/238-590 Mycobacteriophage L5
DPOL_BPMD2/239-602 Mycobacteriophage D29
DPOL_BPSP2/209-647 Bacteriophage SP02
0.1
Tree balance = 0.0

10.1186/1471-2105-6-108

PSI-BLAST



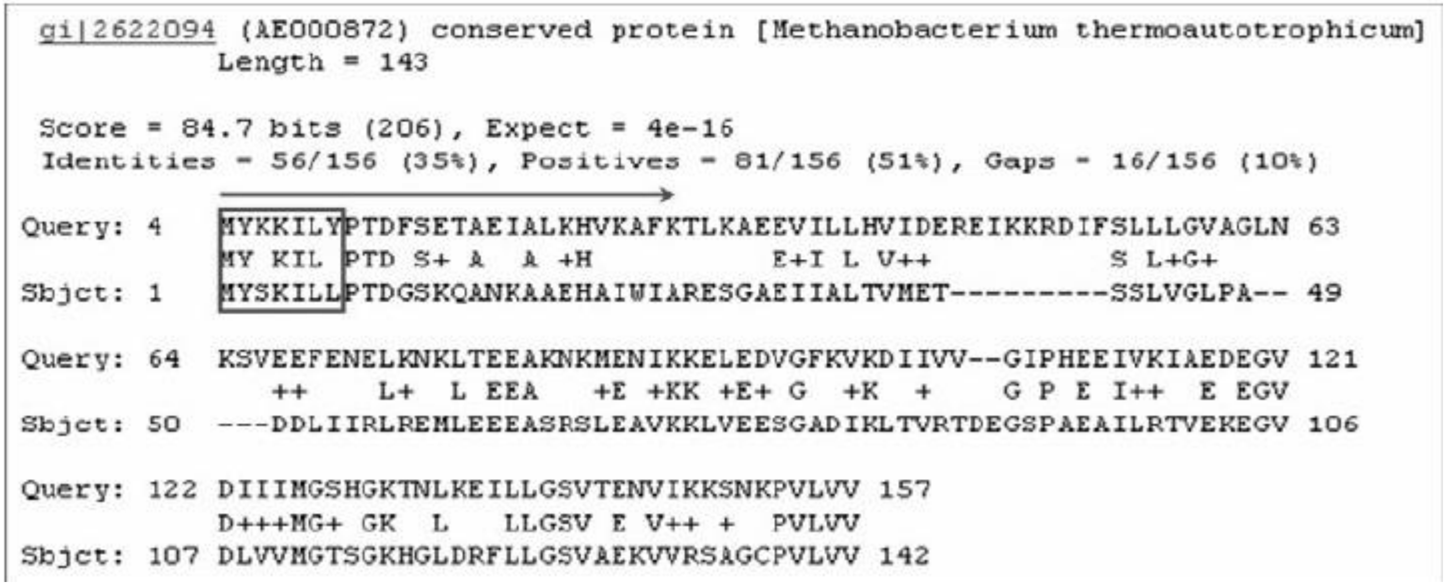
```
                            ↓
ExTox  : GQTSATGVLIGKNTVLTNRHIAKFANGDPSKVSFRPSINTDDNGNTETPYGEYEVKEILQ
         G     G L+ +N V++  H  K       +V           G+ +       ++
Trypsin: GYHFCGGSLVNENWVVSAAHCYKSR----VEVRLGEHNIKVTEGSEQFISSSRVIRHPNY
                            57
                ↓
ExTox  : EPFGAGVDLALIRLKPDQNGVSLGDKISPAKIGTSNDLKDGDKLELIGYPFDHKVNQMHR
           +    D+ LI+L       +L   + P  + TS     G    + G+   + ++
Trypsin: SSYNIDNDIMLIKLS---KPATLNTYVQPVALPTS-CAPAGTMCTVSGW--GNTMSSTAD
                102
                                                      ↓
ExTox  : SE-------IELTT----------LSRGLRYYGFTVP------GNSGSGIFNSNGELVGI
         S          L+           ++  +   G+         G+SG  +   NGEL G+
Trypsin: SNKLQCLNIPILSYSDCNNSYPGMITNAMFCAGYLEGGKDSCQGDSGGPVVC-NGELQGV
                                                      195

ExTox  : HSSKVSHLDREHQINYGV--GIGNYV
          S      +  +   Y       +++
Trypsin: VSWGYGCAEPGNPGVYAKVCIFNDWL
```

TiBS

Fig. 1. Third-iteration PSI–BLAST result from querying exfoliative toxin (ExTox) A from *Staphylococcus aureus* (BAA97652.1) against the non-redundant database. His, Asp and Ser residues are indicated with arrows and numbered as for trypsin [anionic; complexed with the inhibitor benzamidine (1bit)]. The alignment has 15% identity (32/206) and the *E*-value = 6 × 10^{-21}. The threshold *E*-value for inclusion in the profile was 0.005 and the effective search space was 22 926 875 677.

10.1016/S0968-0004(01)02039-4

PSI-BLAST



```
 gi|2622094 (AE000872) conserved protein [Methanobacterium thermoautotrophicum]
           Length = 143

 Score = 84.7 bits (206), Expect = 4e-16
 Identities = 56/156 (35%), Positives = 81/156 (51%), Gaps = 16/156 (10%)

Query: 4   MYKKILYPTDFSETAEIALKHVKAFKTLKAEEVILLHVIDEREIKKRDIFSLLLGVAGLN 63
           MY KIL PTD S+ A  A +H          E+I L V++          S L+G+
Sbjct: 1   MYSKILLPTDGSKQANKAAEHAIWIARESGAEIIALTVMET---------SSLVGLPA-- 49

Query: 64  KSVEEFENELKNKLTEEAKNKMENIKKELEDVGFKVKDIIVV--GIPHEEIVKIAEDEGV 121
              ++    L+  L EEA   +E +KK +E+ G  +K  +    G P E I++  E EGV
Sbjct: 50  ---DDLIIRLREMLEEEASRSLEAVKKLVEESGADIKLTVRTDEGSPAEAILRTVEKEGV 106

Query: 122 DIIIMGSHGKTNLKEILLGSVTENVIKKSNKPVLVV 157
           D+++MG+ GK  L   LLGSV E V++ +  PVLVV
Sbjct: 107 DLVVMGTSGKHGLDRFLLGSVAEKVVRSAGCPVLVV 142
```

An example of high-scoring segment pairs (HSP) found by PSI-BLAST. The first peptide pairs as marked by the box are similar, and we assign the secondary structure element of each amino acid in MYKKILY to its counterpart in MYSKILL.

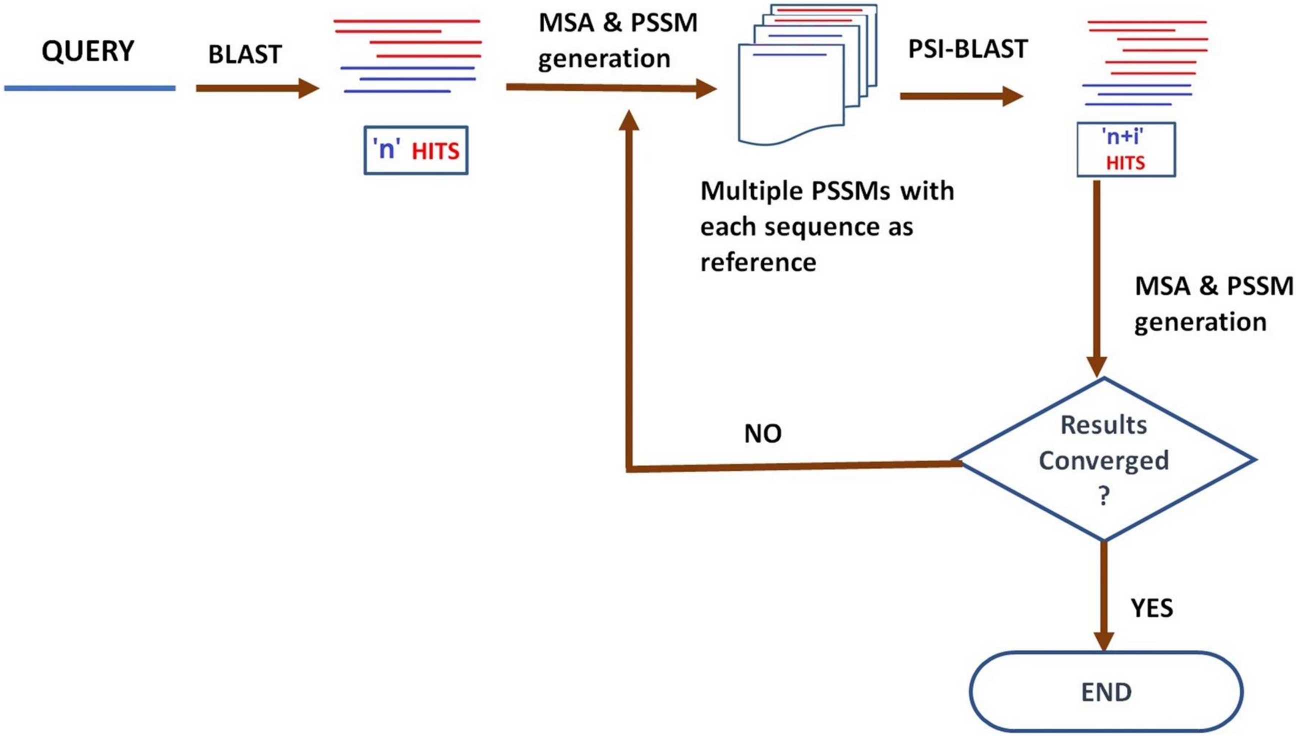
10.1016/S0968-0004(01)02039-4

PSI-BLAST

- Position Specific Iterated – BLAST A
- cycling/iterative method
  - Gives increased sensitivity for detecting **distantly related proteins**
  - Can give insight into functional relationships
  - Very refined statistical methods
- Fast – still based on BLAST methods
- Simple to use

PSI-BLAST

- Essentially we are using intermediate sequences to infer similarity between two sequences that are too dissimilar to link directly.



Profiles & PSSMs Need Multi-sequence Alignment

```
          1                                                50
P43871-1 .......... ...IKKLDSN SIHAIISDIP YGIDYDDWDI LHSNTNSALG
S18997-1 LMSKIYQMDA VDWLKTLENC SVDLFITDPP YESL.EKYRQ IGTTTRLKES
P23192-1 EINKIHQMNC FDFLDQVENK SVQLAVIDPP YNL....... ..........
P29538-1 MDQRLICSNA IKALKNLEEN SIDLIITDPP YNLG.KDY.. ..........
P14751-1 TRHVYDVCDC LDTLAKLPDD SVQLIICDPP YNI....... ..........
P34721-1 KNFNIYQGNC IDFMSHFQDN SIDMIFADPP YFLS.NDG.L TFKNSIIQ..
P50178-1 ENAILVHADS FKLLEKIKPE SMDMIFADPP YFLS.NGG.M SNSGGQIV..
P20590-1 FLNTILKGDC IEKLKTIPNE SIDLIFADPP YFMQ.TEGKL LRTNGDEF..
S43876-1 GPETIIHGDC IEQMNALPEK SVDLIFADPP YNLQ.LGGDL LRPDNSKV..
P28638-1 EAKTIIHGDA LAELKKIPAE SVDLIFADPP YNIG.KNF.. ..........
P23941-1 DLGKLYNGDC LELFKQVPDE NVDTIFADPP FNLD.KEY.. ..........
P14230-1 RSCKIIVGDA REAVQGLDSE IFDCVVTSPP YWGL.RDY.. ..........
P14243-1 NGATLFEGDA LSVLRRLPSG SVRCIVTSPP YWGL.RDY.. ..........
Q04845-1 LNNMLLQGNC AETLKKLPDE SVNLVFTSPP YY........ ..........
S53866-1 WVNDIHEGDA EEVLAELPES SVHMVMTSPP YFGL.RDY.. ..........
P29568-1 .......... ...MNELKDK SINLVVTSPP YPMV.EIWDR LFSELNPKIE

Signature Sequence:                        DPP Y
```

How does PSI-BLAST work?

1) First, a standard blastp is performed

2) The highest scoring hits are used to generate a multiple alignment

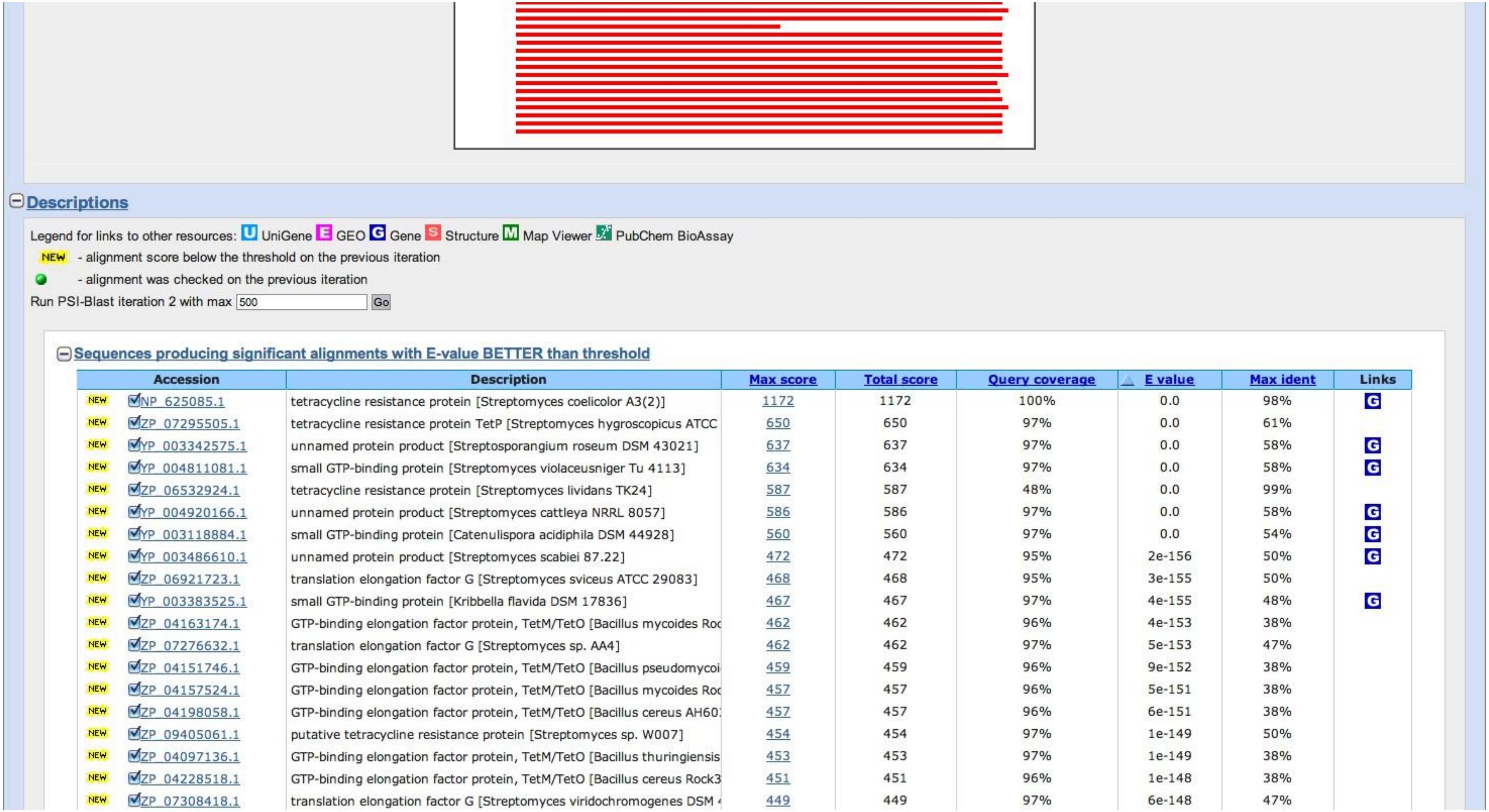
3) A Position Specific Scoring Matrix (PSSM) is generated from the multiple alignment.

- Highly conserved residues get high scores
- Less conserved residues get lower scores
- The PSSM describes the sequence similarity between your query and all significant blastp hits

4) Another similarity search is performed, this time using the new PSSM as the query sequence.

- This PSSM (scoring matrix) is now customized to find sequences that are related to your original query
- Steps 2-4 can be repeated until convergence
- Convergence occurs when no new sequences appear after iteration

PSI-BLAST Example



Descriptions

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay

NEW - alignment score below the threshold on the previous iteration

● - alignment was checked on the previous iteration

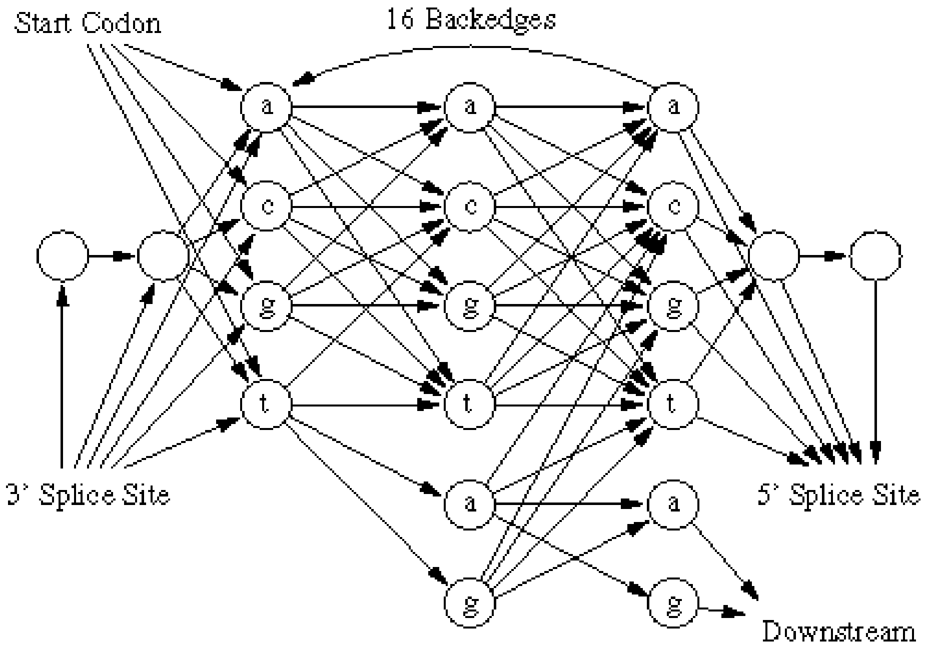
Run PSI-Blast iteration 2 with max 500 Go

Sequences producing significant alignments with E-value BETTER than threshold

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NEW ☑ NP_625085.1	tetracycline resistance protein [Streptomyces coelicolor A3(2)]	1172	1172	100%	0.0	98%	G
NEW ☑ ZP_07295505.1	tetracycline resistance protein TetP [Streptomyces hygroscopicus ATCC	650	650	97%	0.0	61%	
NEW ☑ YP_003342575.1	unnamed protein product [Streptosporangium roseum DSM 43021]	637	637	97%	0.0	58%	G
NEW ☑ YP_004811081.1	small GTP-binding protein [Streptomyces violaceusniger Tu 4113]	634	634	97%	0.0	58%	G
NEW ☑ ZP_06532924.1	tetracycline resistance protein [Streptomyces lividans TK24]	587	587	48%	0.0	99%	
NEW ☑ YP_004920166.1	unnamed protein product [Streptomyces cattleya NRRL 8057]	586	586	97%	0.0	58%	G
NEW ☑ YP_003118884.1	small GTP-binding protein [Catenulispora acidiphila DSM 44928]	560	560	97%	0.0	54%	G
NEW ☑ YP_003486610.1	unnamed protein product [Streptomyces scabiei 87.22]	472	472	95%	2e-156	50%	G
NEW ☑ ZP_06921723.1	translation elongation factor G [Streptomyces sviceus ATCC 29083]	468	468	95%	3e-155	50%	
NEW ☑ YP_003383525.1	small GTP-binding protein [Kribbella flavida DSM 17836]	467	467	97%	4e-155	48%	G
NEW ☑ ZP_04163174.1	GTP-binding elongation factor protein, TetM/TetO [Bacillus mycoides Roc	462	462	96%	4e-153	38%	
NEW ☑ ZP_07276632.1	translation elongation factor G [Streptomyces sp. AA4]	462	462	97%	5e-153	47%	
NEW ☑ ZP_04151746.1	GTP-binding elongation factor protein, TetM/TetO [Bacillus pseudomycoi	459	459	96%	9e-152	38%	
NEW ☑ ZP_04157524.1	GTP-binding elongation factor protein, TetM/TetO [Bacillus mycoides Roc	457	457	96%	5e-151	38%	
NEW ☑ ZP_04198058.1	GTP-binding elongation factor protein, TetM/TetO [Bacillus cereus AH60	457	457	96%	6e-151	38%	
NEW ☑ ZP_09405061.1	putative tetracycline resistance protein [Streptomyces sp. W007]	454	454	97%	1e-149	50%	
NEW ☑ ZP_04097136.1	GTP-binding elongation factor protein, TetM/TetO [Bacillus thuringiensis	453	453	97%	1e-149	38%	
NEW ☑ ZP_04228518.1	GTP-binding elongation factor protein, TetM/TetO [Bacillus cereus Rock3	451	451	96%	1e-148	38%	
NEW ☑ ZP_07308418.1	translation elongation factor G [Streptomyces viridochromogenes DSM	449	449	97%	6e-148	47%	

HMMs & HMMer

- The more powerful way to search for protein families than PSSMs



Hidden Markov Models

Hidden Markov Models in Bioinformatics

- **Used extensively in gene prediction**
- **Used to create Sequence Profiles and to classify sequences into families**
- **Used in Multiple Sequence Alignment**

HMMER 3

- Suite of sequence analysis programs based on HMMs

- Used to build the Pfam database

- Available for free download at
  - http://hmmer.org/

HMMER 3

- HMMER 2 was used for many years
  - Biggest draw back was always speed

- HMMER 3 released in 2011
  - Very fast with comparable speeds to BLAST
  - 100X faster than v2

HMMER Programs

- **hmmbuild** – build a HMM from multiple sequence alignment
- **hmmscan** – searches a query sequence(s) against a database of HMMs (used by PFAM)
- **hmmsearch**– searches a query HMM against a database of sequences (e.g. like psi-blast)
- **phmmer** – search a protein sequence vs a sequence database (e.g. like blastp)

HMMER Search & Software

- http://hmmer.janelia.org/search

- PFAM
  - http://pfam.sanger.ac.uk/

RNA Alignments

- RNA alignments are “special”
- RNA genes often have secondary structures that allow improved searching
- Improved searching is needed since
    - Must search in DNA space (less complex then protein sequences)
    - Often shorter length then proteins

Infernal (RNA Search)

- Infernal is like HMMER
  - Includes use of secondary structure information
  - Uses profile “stochastic context-free grammar”
    - SCFGs vs HMMs
  - “consensus RNA secondary structure profiles”
- Infernal is slow!
- Infernal can be used to search RFAM

RNAseq alignment with Magic-BLAST

Software | Open Access | Published: 25 July 2019

Magic-BLAST, an accurate RNA-seq aligner for long and short reads

Grzegorz M. Boratyn, Jean Thierry-Mieg, Danielle Thierry-Mieg, Ben Busby & Thomas L. Madden ✉

BMC Bioinformatics **20**, Article number: 405 (2019) | Cite this article

10k Accesses | **52** Citations | **51** Altmetric | Metrics

Abstract

Background

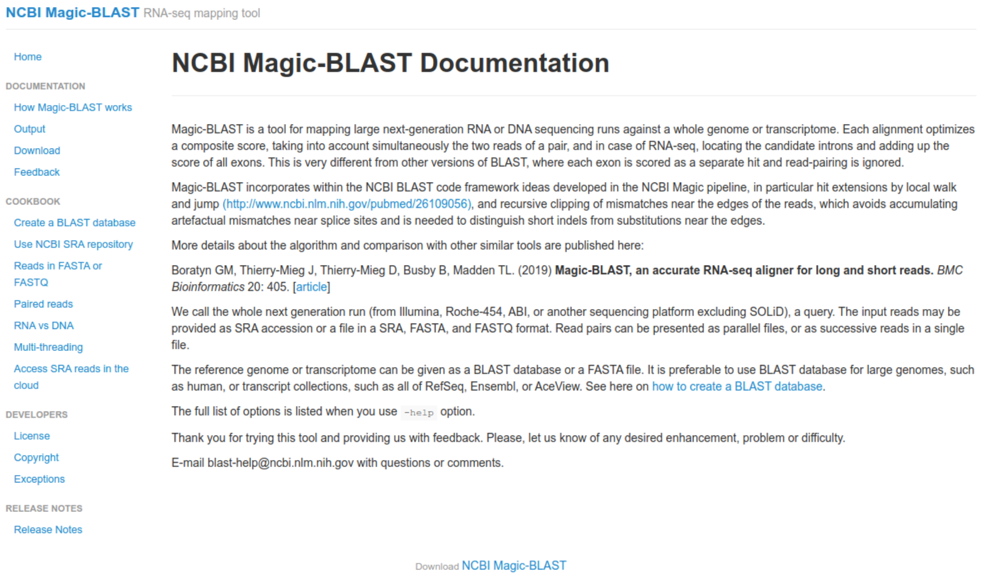
Next-generation sequencing technologies can produce tens of millions of reads, often paired-end, from transcripts or genomes. But few programs can align RNA on the genome and accurately discover introns, especially with long reads. We introduce Magic-BLAST, a new aligner based on ideas from the Magic pipeline.

Results

Magic-BLAST uses innovative techniques that include the optimization of a spliced alignment score and selective masking during seed selection. We evaluate the performance of Magic-BLAST to accurately map short or long sequences and its ability to discover introns on real RNA-seq data sets from PacBio, Roche and Illumina runs, and on six benchmarks, and compare it to other popular aligners. Additionally, we look at alignments of human idealized RefSeq mRNA sequences perfectly matching the genome.

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2996-x

RNAseq alignment with Magic-BLAST



NCBI Magic-BLAST Documentation

Magic-BLAST is a tool for mapping large next-generation RNA or DNA sequencing runs against a whole genome or transcriptome. Each alignment optimizes a composite score, taking into account simultaneously the two reads of a pair, and in case of RNA-seq, locating the candidate introns and adding up the score of all exons. This is very different from other versions of BLAST, where each exon is scored as a separate hit and read-pairing is ignored.

Magic-BLAST incorporates within the NCBI BLAST code framework ideas developed in the NCBI Magic pipeline, in particular hit extensions by local walk and jump (http://www.ncbi.nlm.nih.gov/pubmed/26109056), and recursive clipping of mismatches near the edges of the reads, which avoids accumulating artefactual mismatches near splice sites and is needed to distinguish short indels from substitutions near the edges.

More details about the algorithm and comparison with other similar tools are published here:

Boratyn GM, Thierry-Mieg J, Thierry-Mieg D, Busby B, Madden TL. (2019) **Magic-BLAST, an accurate RNA-seq aligner for long and short reads.** *BMC Bioinformatics* 20: 405. [article]

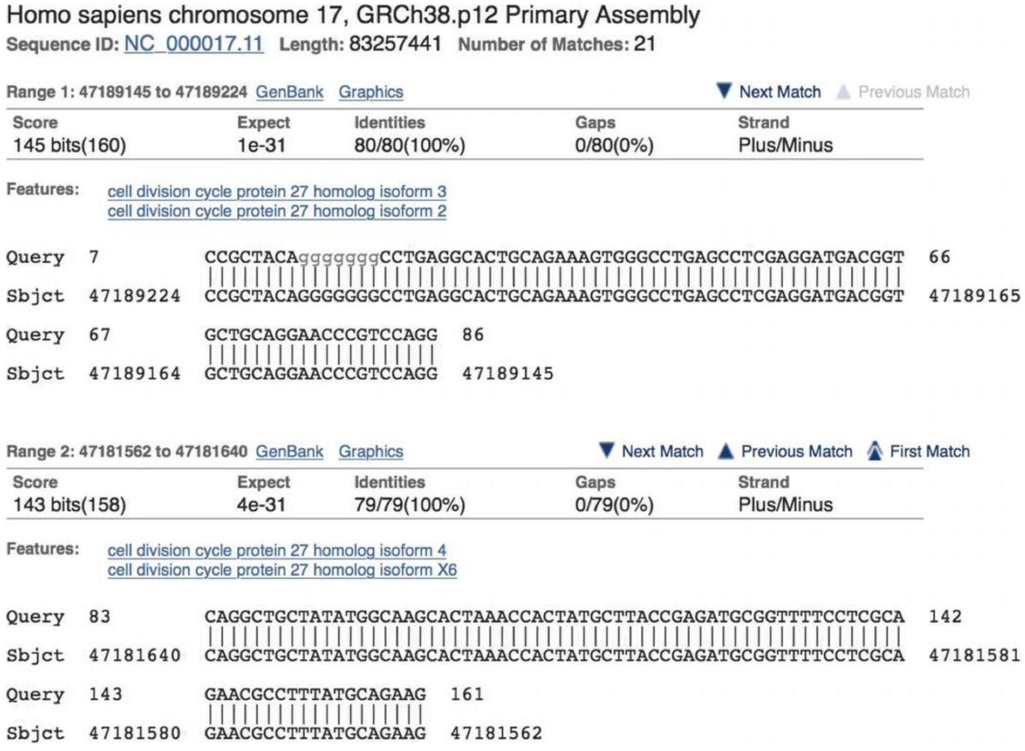
We call the whole next generation run (from Illumina, Roche-454, ABI, or another sequencing platform excluding SOLiD), a query. The input reads may be provided as SRA accession or a file in a SRA, FASTA, and FASTQ format. Read pairs can be presented as parallel files, or as successive reads in a single file.

The reference genome or transcriptome can be given as a BLAST database or a FASTA file. It is preferable to use BLAST database for large genomes, such as human, or transcript collections, such as all of RefSeq, Ensembl, or AceView. See here on how to create a BLAST database.

The full list of options is listed when you use `-help` option.

Thank you for trying this tool and providing us with feedback. Please, let us know of any desired enhancement, problem or difficulty.

E-mail blast-help@ncbi.nlm.nih.gov with questions or comments.



Homo sapiens chromosome 17, GRCh38.p12 Primary Assembly

Sequence ID: NC_000017.11 **Length:** 83257441 **Number of Matches:** 21

Range 1: 47189145 to 47189224 GenBank Graphics

Score	Expect	Identities	Gaps	Strand
145 bits(160)	1e-31	80/80(100%)	0/80(0%)	Plus/Minus

Features: cell division cycle protein 27 homolog isoform 3
cell division cycle protein 27 homolog isoform 2

```
Query  7         CCGCTACAggggggCCTGAGGCACTGCAGAAAGTGGGCCTGAGCCTCGAGGATGACGGT  66
                 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  47189224  CCGCTACAGGGGGGCCTGAGGCACTGCAGAAAGTGGGCCTGAGCCTCGAGGATGACGGT  47189165

Query  67        GCTGCAGGAACCCGTCCAGG  86
                 ||||||||||||||||||||
Sbjct  47189164  GCTGCAGGAACCCGTCCAGG  47189145
```

Range 2: 47181562 to 47181640 GenBank Graphics

Score	Expect	Identities	Gaps	Strand
143 bits(158)	4e-31	79/79(100%)	0/79(0%)	Plus/Minus

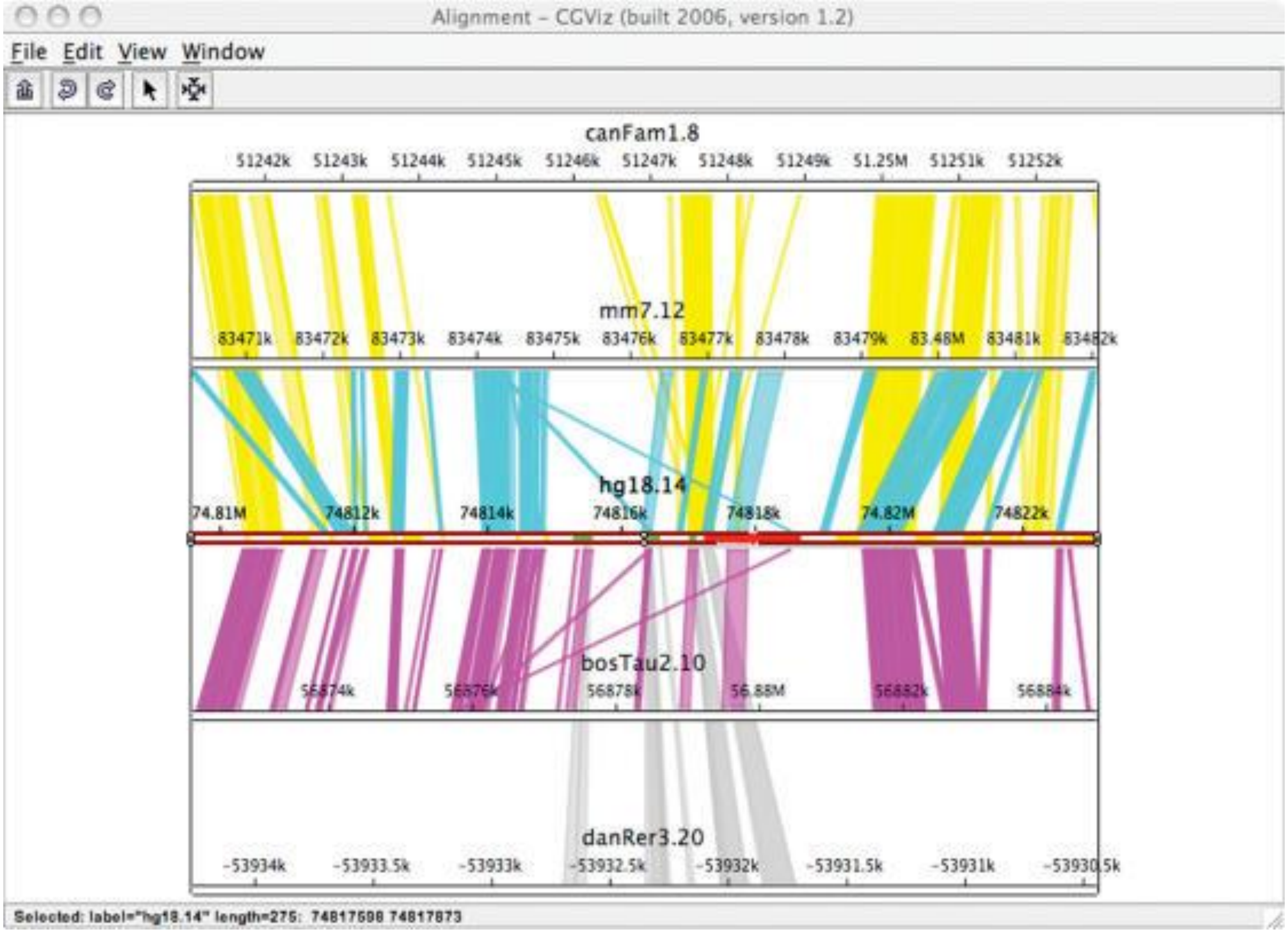
Features: cell division cycle protein 27 homolog isoform 4
cell division cycle protein 27 homolog isoform X6

```
Query  83        CAGGCTGCTATATGGCAAGCACTAAACCACTATGCTTACCGAGATGCGGTTTTCCTCGCA  142
                 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  47181640  CAGGCTGCTATATGGCAAGCACTAAACCACTATGCTTACCGAGATGCGGTTTTCCTCGCA  47181581

Query  143       GAACGCCTTTATGCAGAAG  161
                 |||||||||||||||||||
Sbjct  47181580  GAACGCCTTTATGCAGAAG  47181562
```

https://ncbi.github.io/magicblast/

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2996-x

Genome Alignment

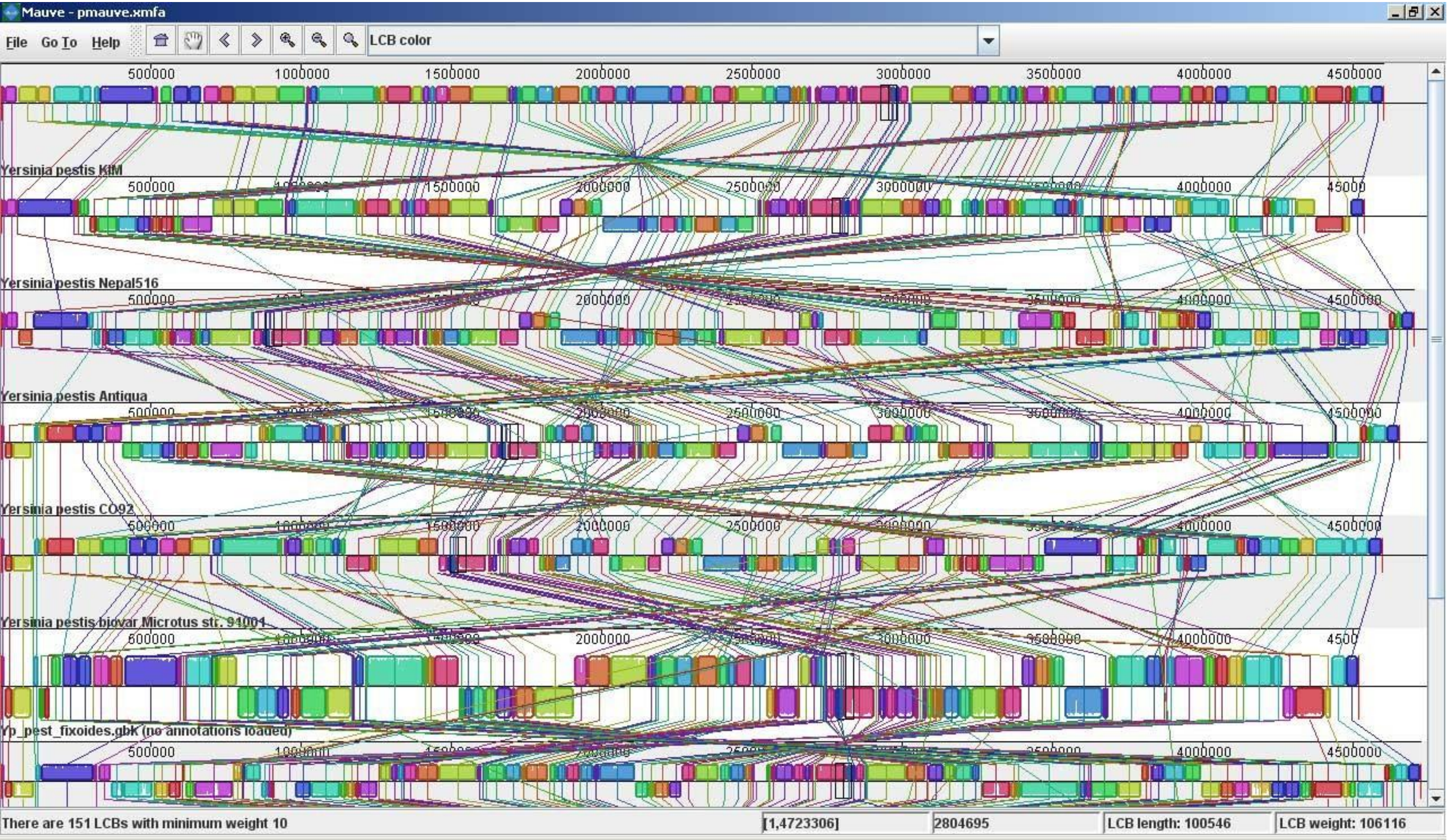

Alignment – CGViz (built 2006, version 1.2)
File Edit View Window
canFam1.8
51242k 51243k 51244k 51245k 51246k 51247k 51248k 51249k 51.25M 51251k 51252k
mm7.12
83471k 83472k 83473k 83474k 83475k 83476k 83477k 83478k 83479k 83.48M 83481k 83482k
hg18.14
74.81M 74812k 74814k 74816k 74818k 74.82M 74822k
bosTau2.10
56874k 56876k 56878k 56.88M 56882k 56884k
danRer3.20
-53934k -53933.5k -53933k -53932.5k -53932k -53931.5k -53931k -53930.5k

Selected: label="hg18.14" length=275: 74817598 74817873

Genome Alignment

- Genome alignment useful for
  - Visualizing genome
    - Rearrangements
    - Insertions/deletions
    - Inversions
  - Annotating genomes
    - Comparing gene annotations across species

Mauve



There are 151 LCBs with minimum weight 10

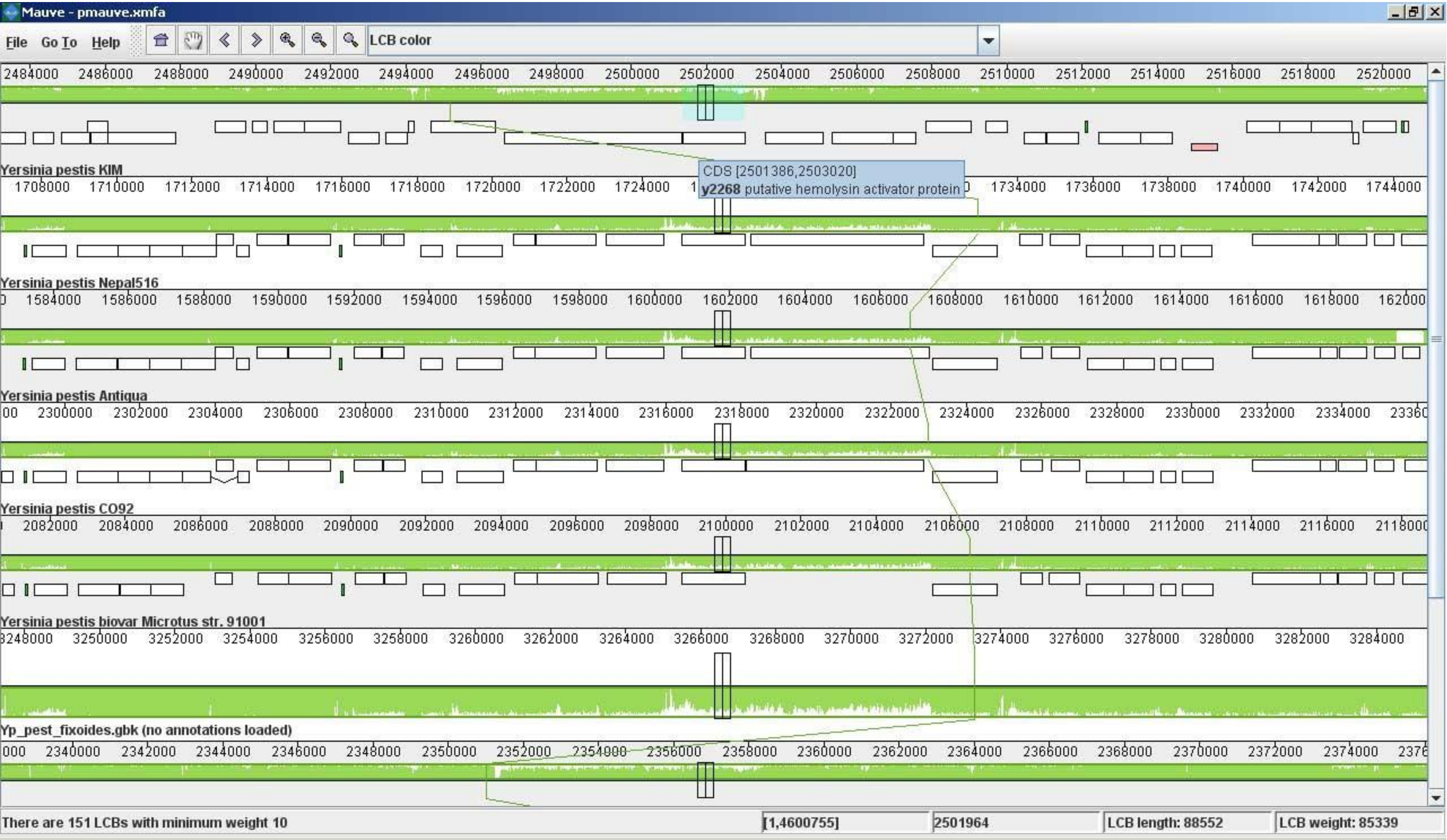
[1,4723306]

2804695

LCB length: 100546

LCB weight: 106116

Mauve (zoomed in)



CDS [2501386,2503020]
y2268 putative hemolysin activator protein

There are 151 LCBs with minimum weight 10

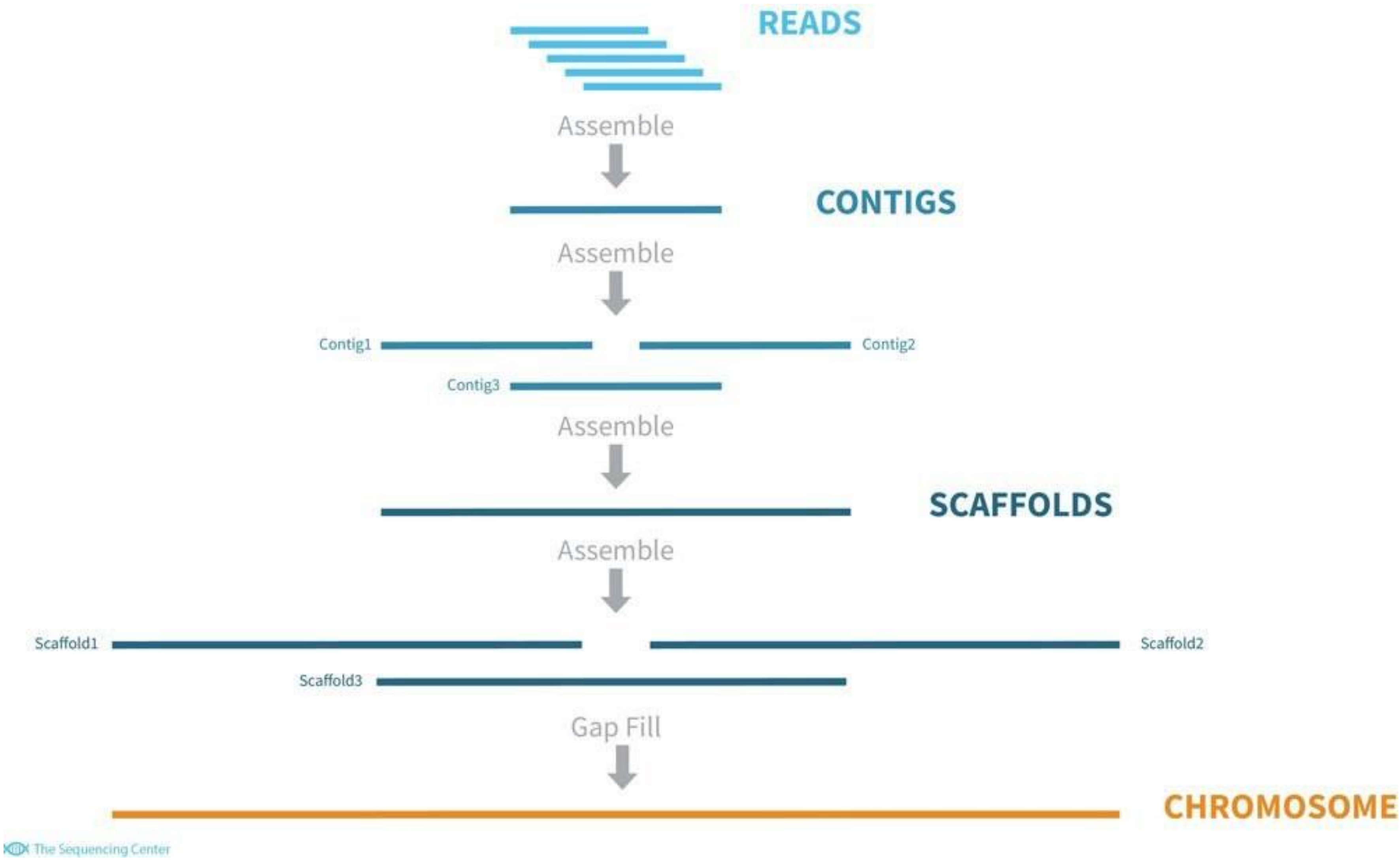
[1,4600755]

2501964

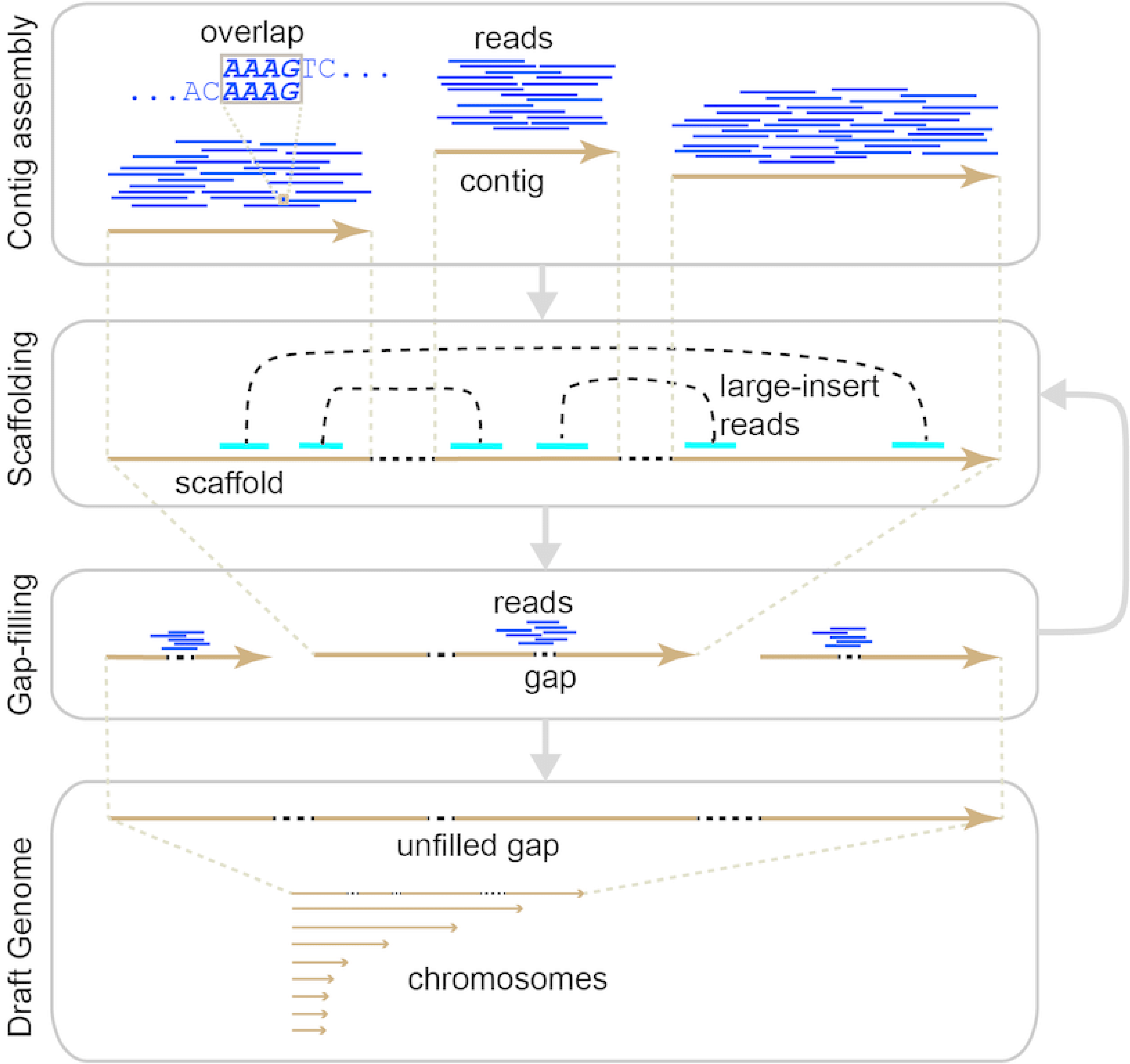
LCB length: 88552

LCB weight: 85339

Assemblers


READS
Assemble
CONTIGS
Assemble
Contig1
Contig2
Contig3
Assemble
SCAFFOLDS
Assemble
Scaffold1
Scaffold2
Scaffold3
Gap Fill
CHROMOSOME
The Sequencing Center

Assemblers

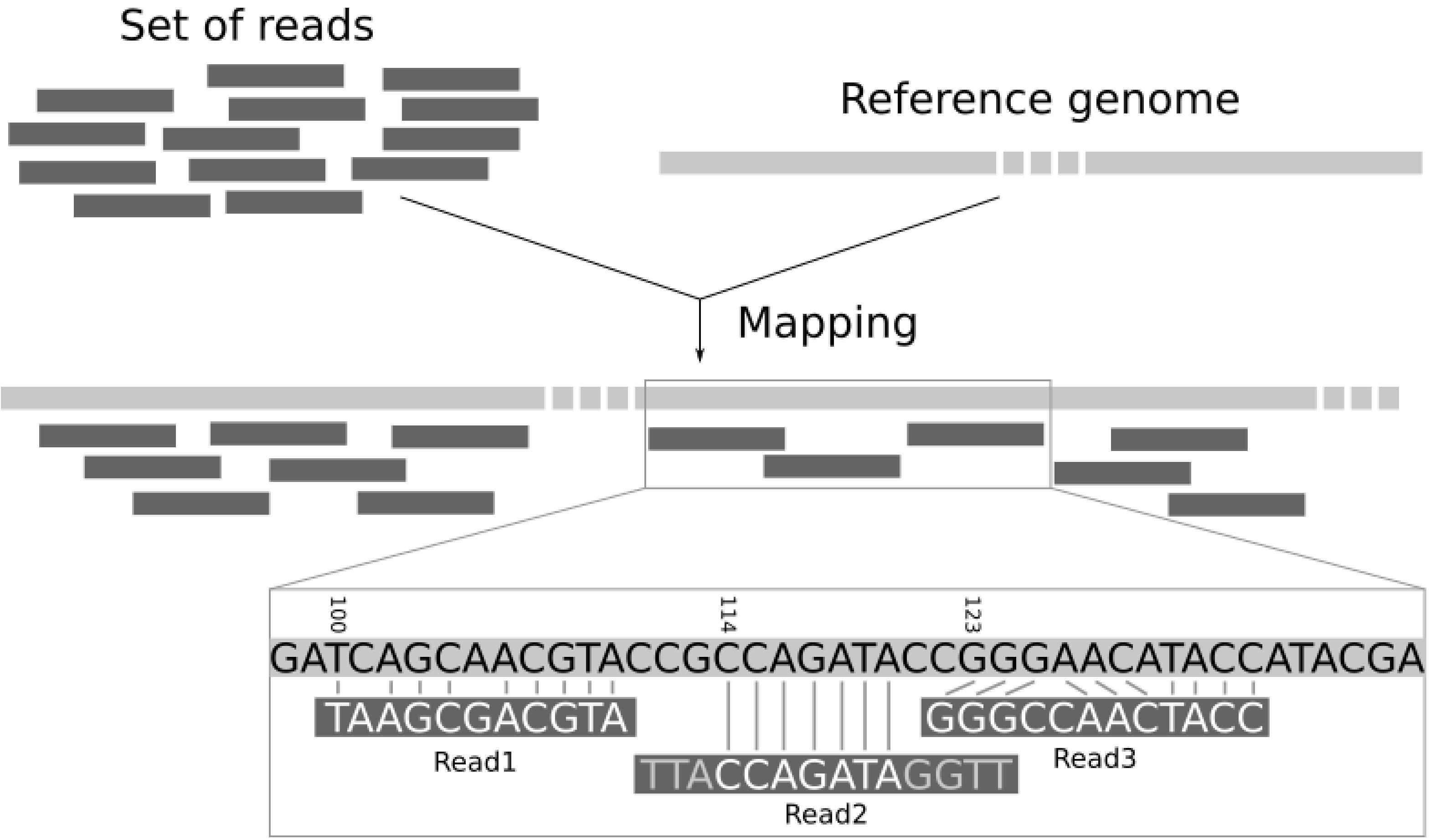


10.1093/BIB/BBW096

Assemblers

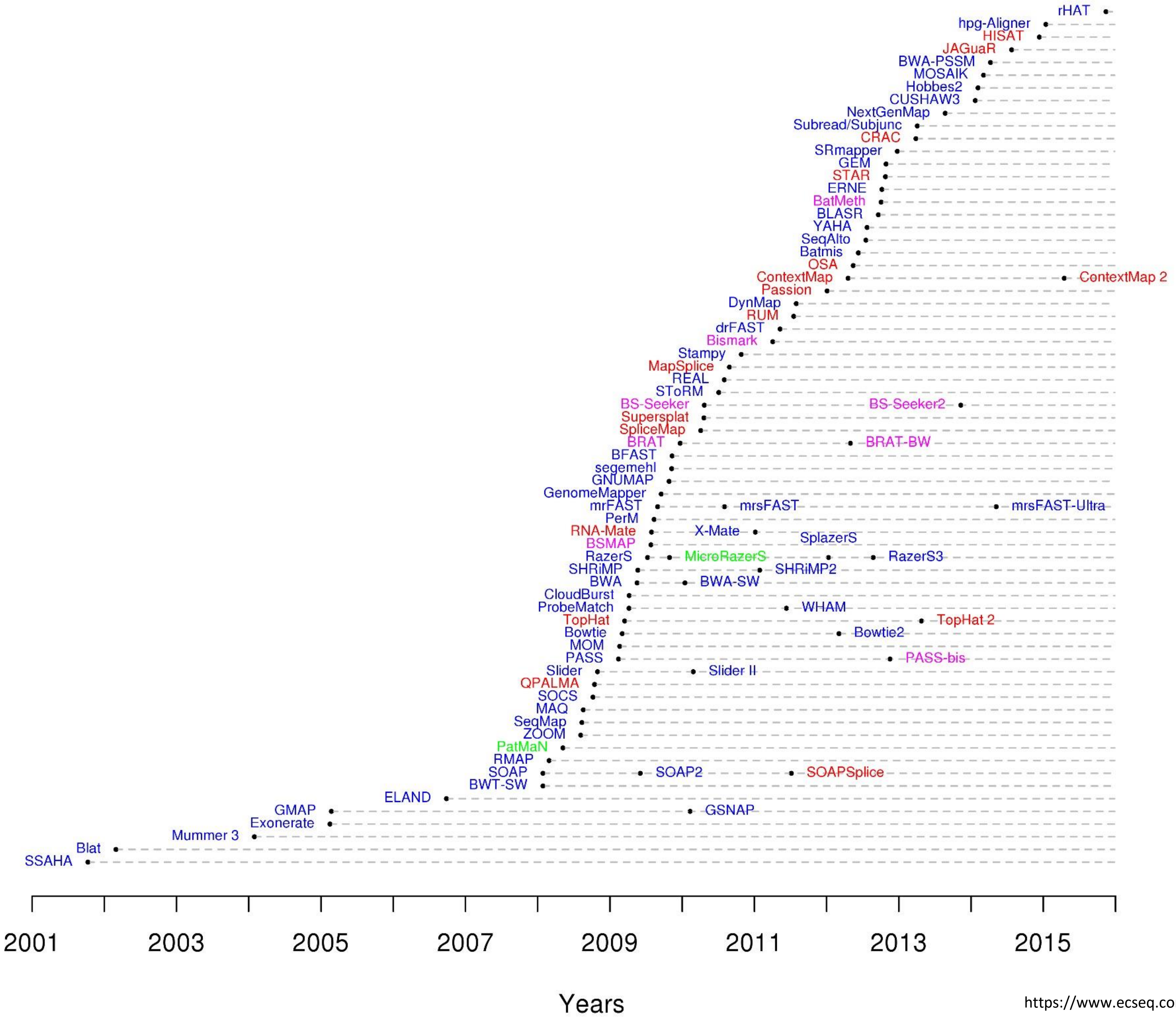
- Assemblers job is to make longer sequences from shorter ones.
- Nothing like homology searching
- Must efficiently compare and join billions of sequences
- Soap-Denovo: http://soap.genomics.org.cn/soapdenovo.html
- Amos: http://sourceforge.net/apps/mediawiki/amos/index.php?title=AMOS
- Many, many, more

Mappers (Aligners)

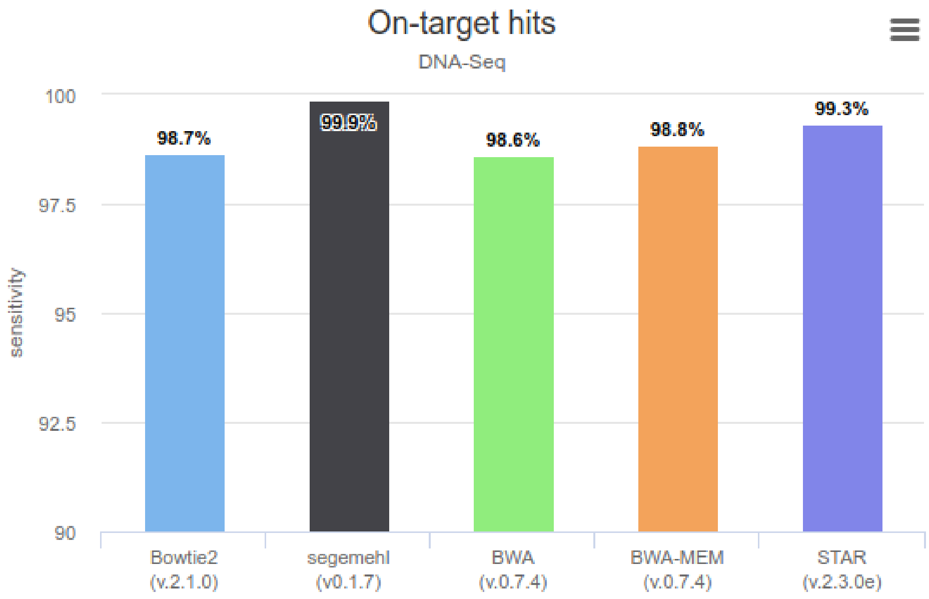


Mappers

- These map a read to a reference genome
- Useful for assembly when a reference genome is already known
  - (think assembly of personal human genomes)
- Identifying SNPs within the same species Very Fast!
- BWA: https://github.com/lh3/bwa
- Bowtie: http://bowtie-bio.sourceforge.net/index.shtml
- Stampy: http://www.well.ox.ac.uk/project-stampy
- Many Others
-

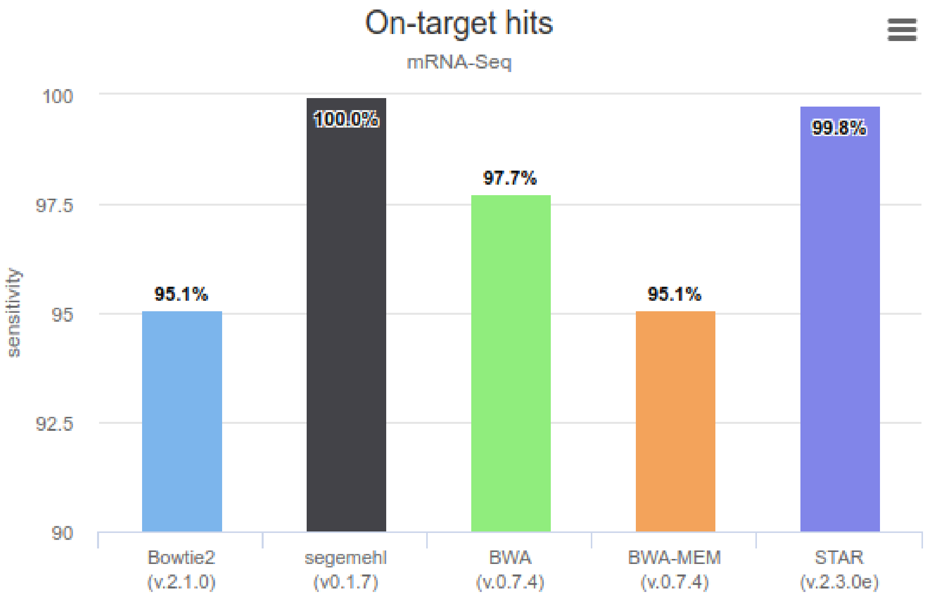


https://www.ecseq.com/support/benchmark



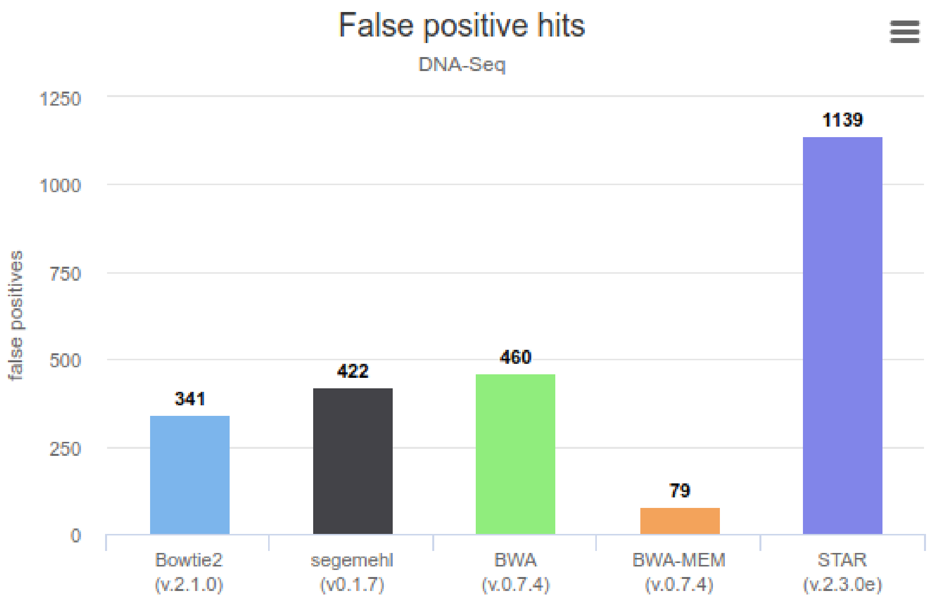
On-target hits

DNA-Seq



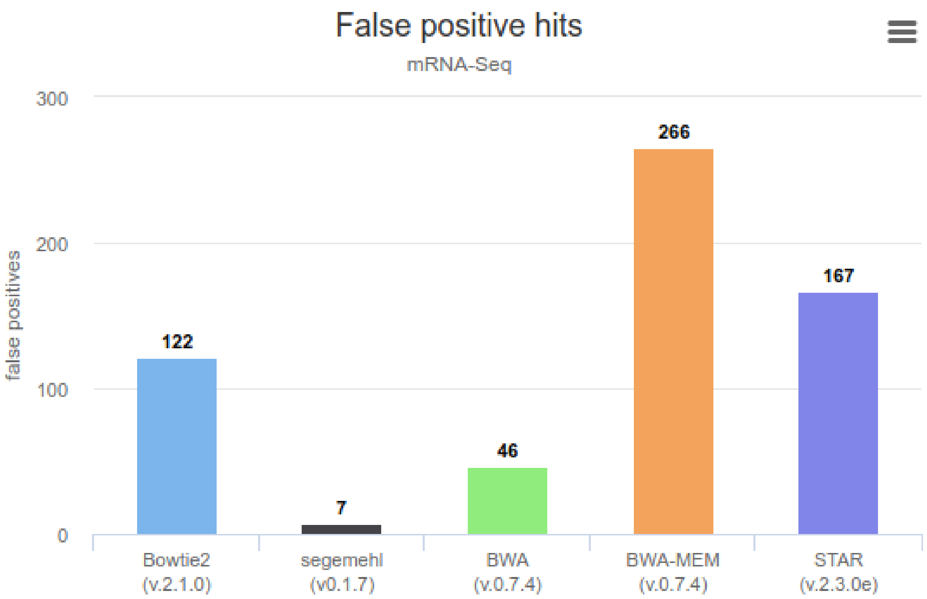
On-target hits

mRNA-Seq



False positive hits

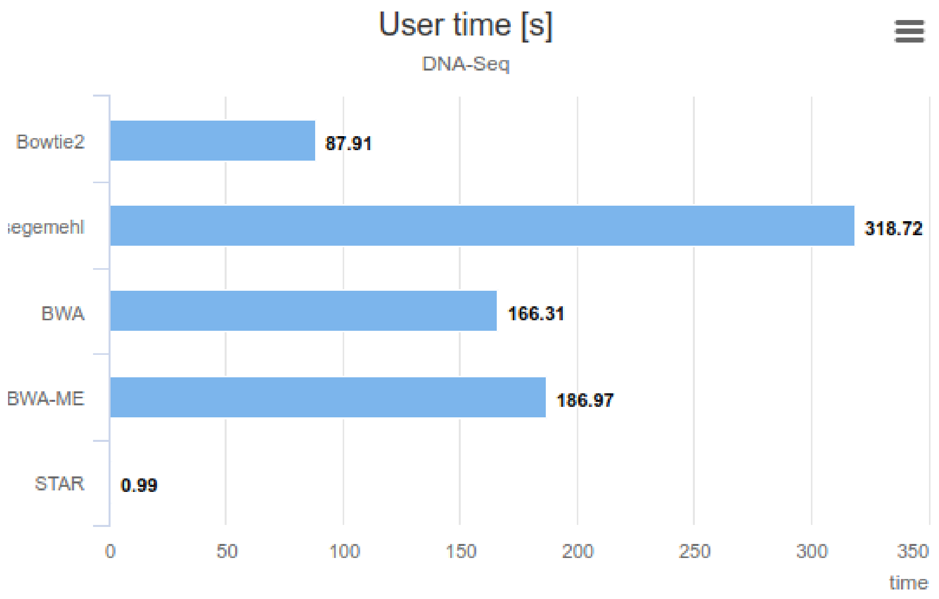
DNA-Seq

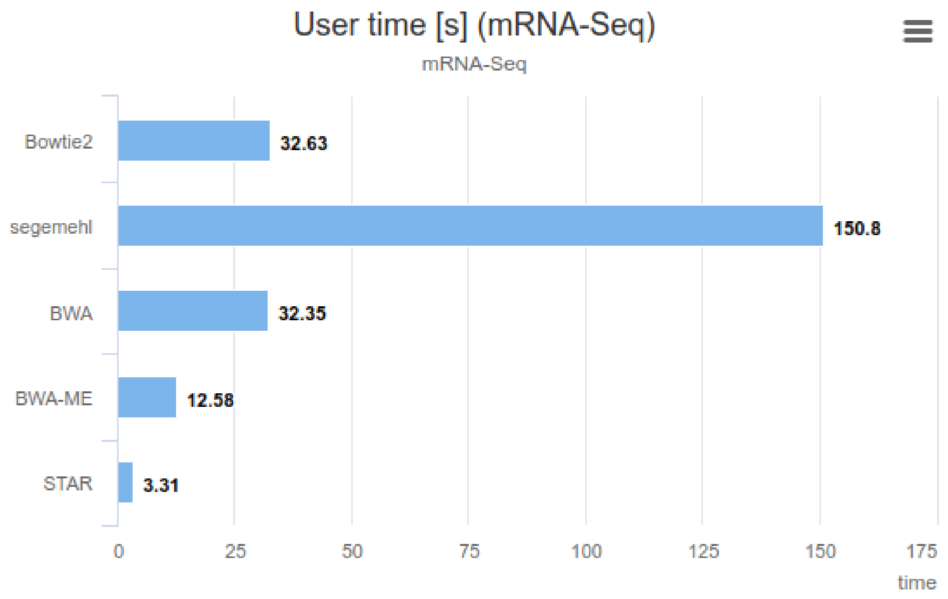


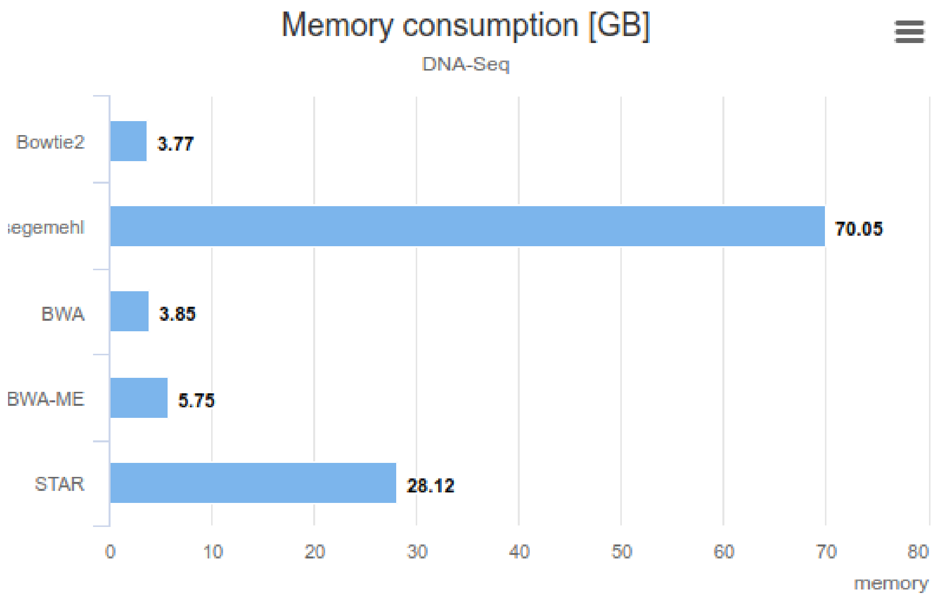
False positive hits

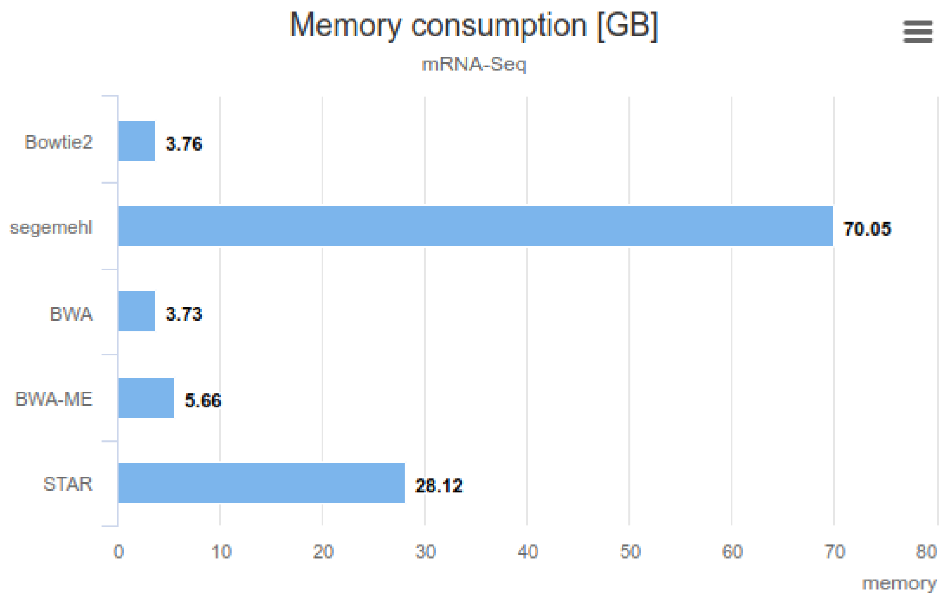
mRNA-Seq

Otto C, Stadler PF, Hoffmann S: 'Lacking alignments? The next generation sequencing mapper segemehl revisited', Bioinformatics. 2014 Jul 1;30(13):1837-43 (2014)



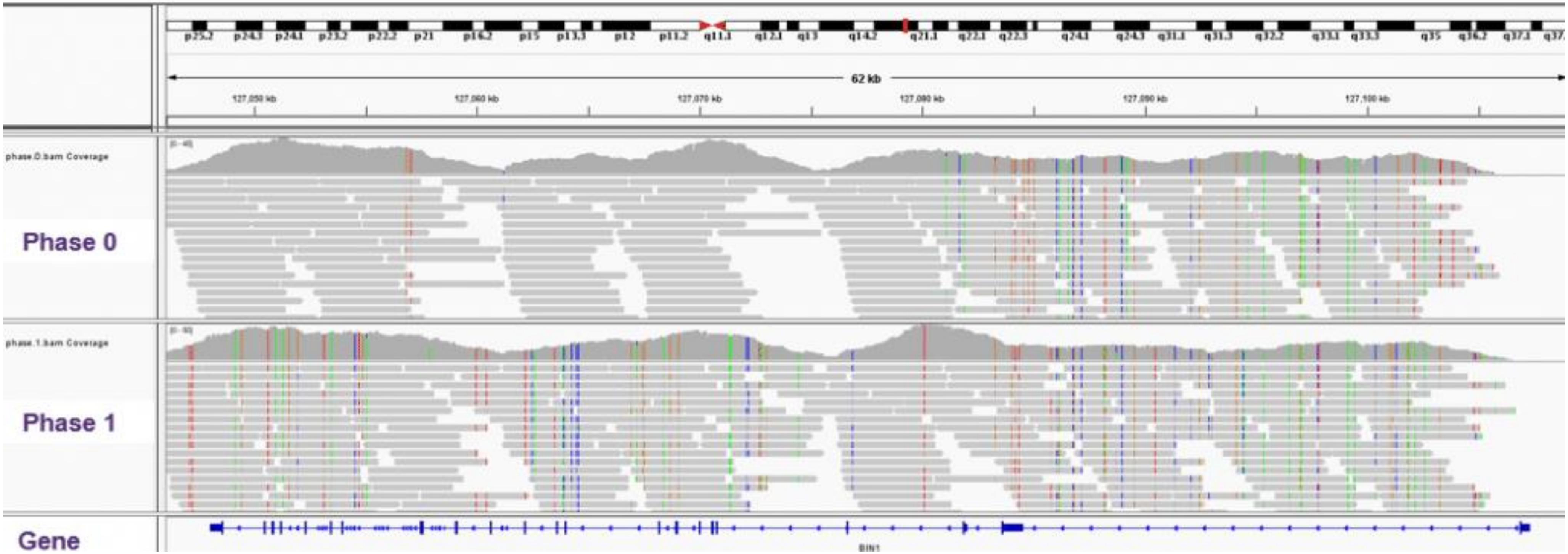






Otto C, Stadler PF, Hoffmann S: 'Lacking alignments? The next generation sequencing mapper segemehl revisited', Bioinformatics. 2014 Jul 1;30(13):1837-43 (2014)

Mappers



IGV snapshot with reads in bam file