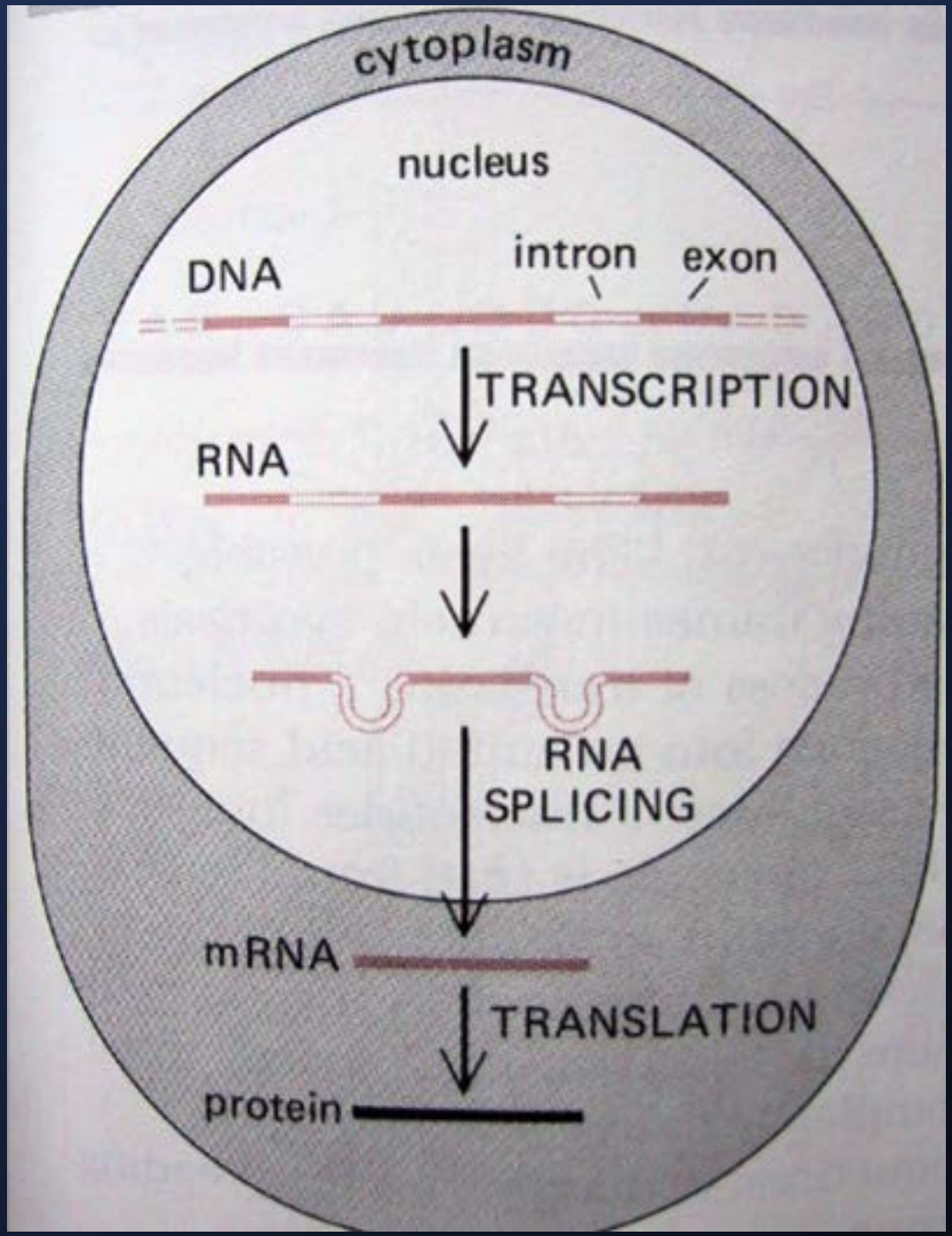
Unknown sequence (http://www.wiley.com/legacy/products/subject/life/bioinformatics/questions_10.html)

ATGGAGAATAGTCTTAGATGTGTTTGGGTACCCAAGCTGGCTTTTGTACTCTTCGGAGCTTCCTTGCTCA
GCGCGCATCTTCAAGTAACCGGTTTTCAAATTAAAGCTTTCACAGCACTGCGCTTCCTCTCAGAACCTTC
TGATGCCGTCACAATGCGGGGAGGAAATGTCCTCCTCGACTGCTCCGCGGAGTCCGACCGAGGAGTTCCA
GTGATCAAGTGGAAGAAAGATGGCATTCATCTGGCCTTGGGAATGGATGAAAGGAAGCAGCAACTTTCAA
ATGGGTCTCTGCTGATACAAAACATACTTCATTCCAGACACCACAAGCCAGATGAGGGACTTTACCAATG
TGAGGCATCTTTAGGAGATTCTGGCTCAATTATTAGTCGGACAGCAAAAGTTGCAGTAGCAGGACCACTG
AGGTTCCTTTCACAGACAGAATCTGTCACAGCCTTCATGGGAGACACAGTGCTACTCAAGTGTGAAGTCA
TTGGGGAGCCCATGCCAACAATCCACTGGCAGAAGAACCAACAAGACCTGACTCCAATCCCAGGTGACTC
CCGAGTGGTGGTCTTGCCCTCTGGAGCATTGCAGATCAGCCGACTCCAACCGGGGGACATTGGAATTTAC
CGATGCTCAGCTCGAAATCCAGCCAGCTCAAGAACAGGAAATGAAGCAGAAGTCAGAATTTTATCAGATC
CAGGACTGCATAGACAGCTGTATTTTCTGCAAAGACCATCCAATGTAGTAGCCATTGAAGGAAAAGATGC
TGTCCTGGAATGTTGTGTTTCTGGCTATCCTCCACCAAGTTTTACCTGGTTACGAGGCGAGGAAGTCATC
CAACTCAGGTCTAAAAAGTATTCTTTATTGGGTGGAAGCAACTTGCTTATCTCCAATGTGACAGATGATG
ACAGTGGAATGTATACCTGTGTTGTCACATATAAAAATGAGAATATTAGTGCCTCTGCAGAGCTCACAGT
CTTGGTTCCGCCATGGTTTTTAAATCATCCTTCCAACCTGTATGCCTATGAAAGCATGGATATTGAGTTT
GAATGTACAGTCTCTGGAAAGCCTGTGCCCACTGTGAATTGGATGAAGAATGGAGATGTGGTCATTCCTA
GTGATTATTTTCAGATAGTGGGAGGAAGCAACTTACGGATACTTGGGGTGGTGAAGTCAGATGAAGGCTT
TTATCAATGTGTGGCTGAAAATGAGGCTGGAAATGCCCAGACCAGTGCACAGCTCATTGTCCCTAAGCCT
GCAATCCCAAGCTCCAGTGTCCTCCCTTCGGCTCCCAGAGATGTGGTCCCTGTCTTGGTTTCCAGCCGAT
TTGTCCGTCTCAGCTGGCGCCCACCTGCAGAAGCGAAAGGGAACATTCAAACTTTCACGGTCTTTTTCTC
CAGAGAAGGTGACAACAGGGAACGAGCATTGAATACAACACAGCCTGGGTCCCTTCAGCTCACTGTGGGA
AACCTGAAGCCAGAAGCCATGTACACCTTTCGAGTTGTGGCTTACAATGAATGGGGACCGGGAGAGAGTT
CTCAACCCATCAAGGTGGCCACACAGCCTGAGTTGCAAGTTCCAGGGCCAGTAGAAAACCTGCAAGCTGT
ATCTACCTCACCTACCTCAATTCTTATTACCTGGGAACCCCCTGCCTATGCAAACGGTCCAGTCCAAGGT
TACAGATTGTTCTGCACTGAGGTGTCCACAGGAAAAGAACAGAATATAGAGGTTGATGGACTATCTTATA
AACTGGAAGGCCTGAAAAAATTCACCGAATATAGTCTTCGATTCTTAGCTTATAATCGCTATGGTCCGGG
CGTCTCTACTGATGATATAACAGTGGTTACACTTTCTGACGTGCCAAGTGCCCCGCCTCAGAACGTCTCC
CTGGAAGTGGTCAATTCAAGAAGTATCAAAGTTAGCTGGCTGCCTCCTCCATCAGGAACACAAAATGGAT
TTATTACCGGCTATAAAATTCGACACAGAAAGACGACCCGCAGGGGTGAGATGGAAACACTGGAGCCAAA
CAACCTCTGGTACCTATTCACAGGACTGGAGAAAGGAAGTCAGTACAGTTTCCAGGTGTCAGCCATGACA

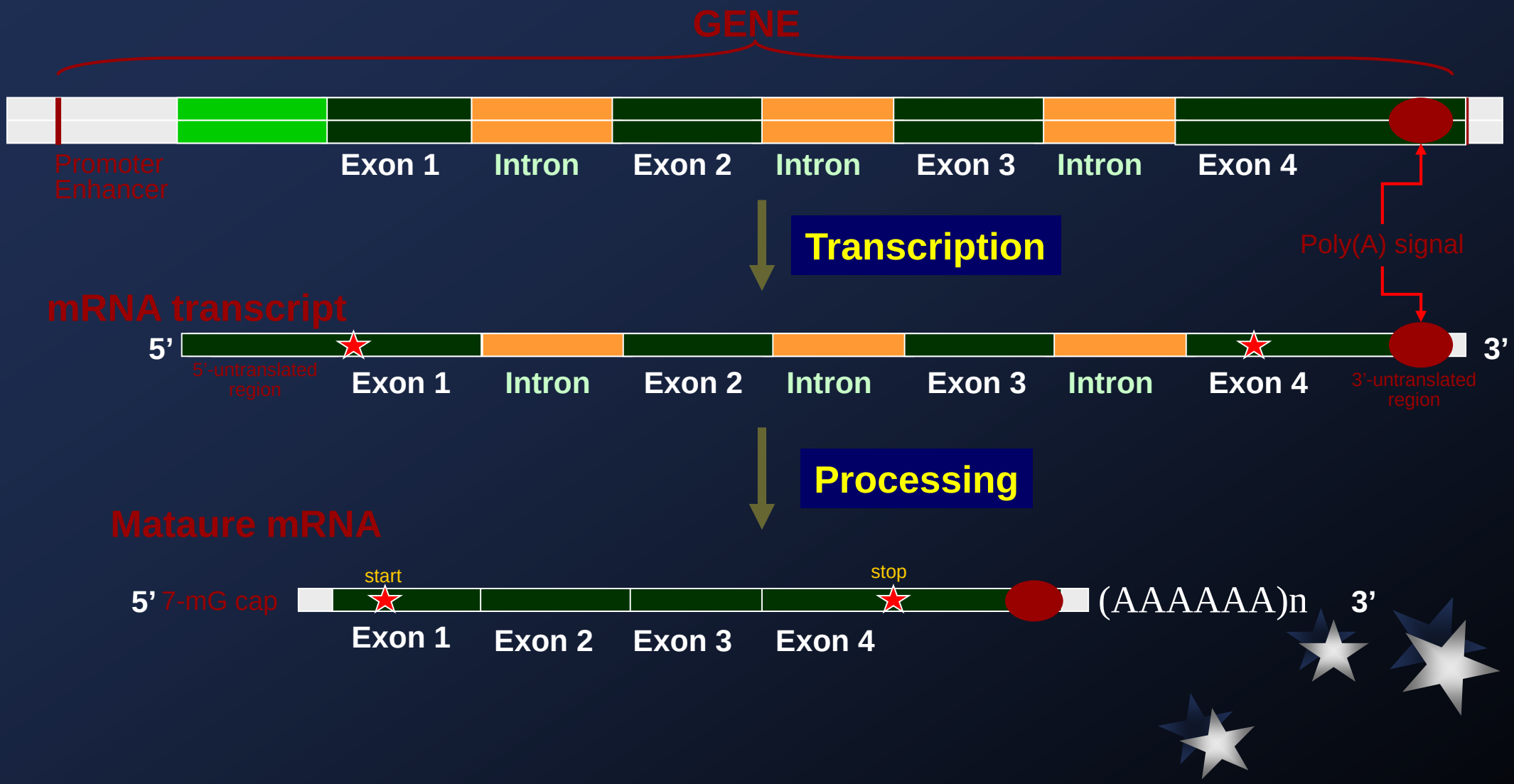
DNA

RNA

Protein



The Organization of an Eukaryotic Gene



Bioinformatics

Simple Mathematics:

Human Genome

3×10^9 bps

Human Genes (1.5% of the genome)

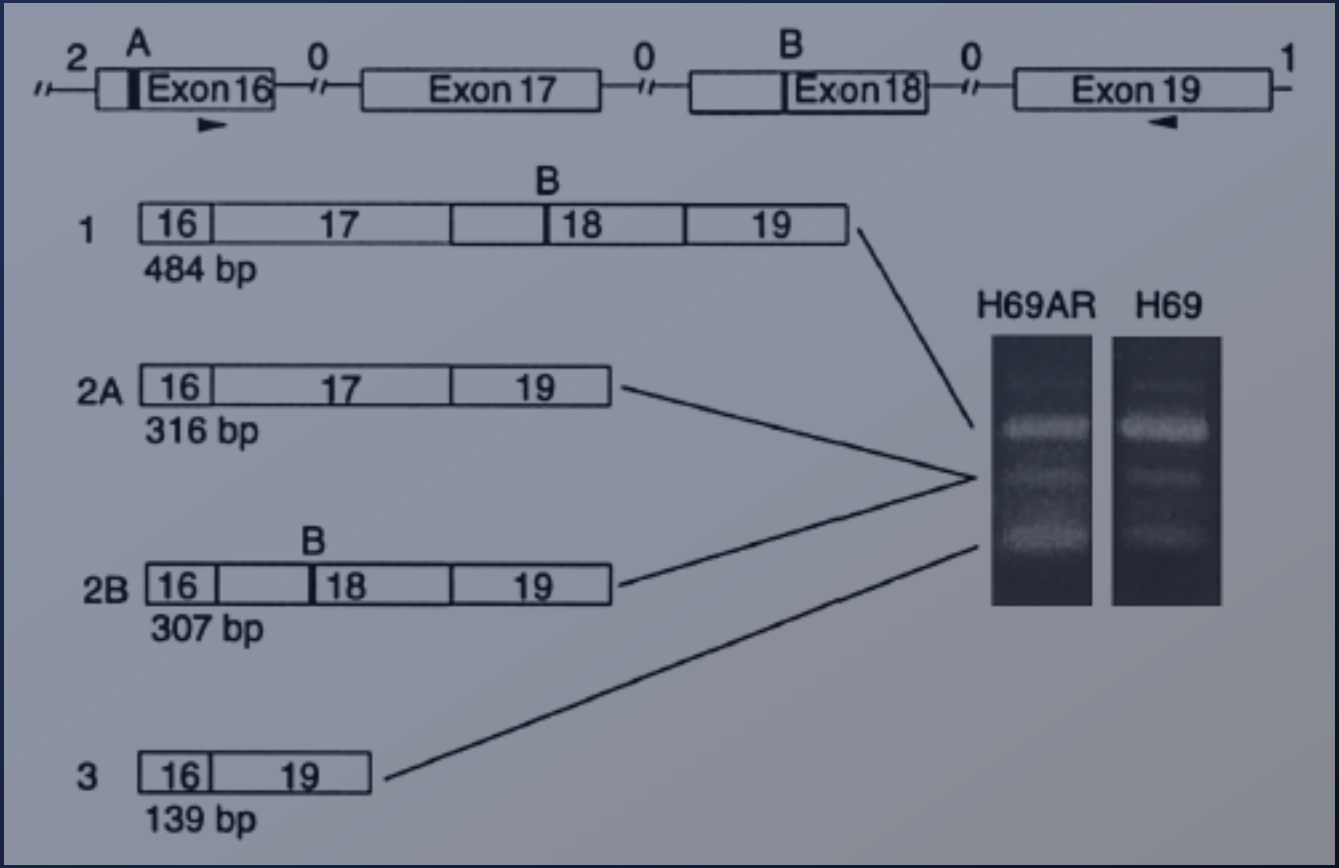
40,000 genes

In a given cell type at a certain stage, it is estimated that around 25 to 50 % of the genes are transcribed or expressed.

10,000 to 20,000 genes



Bioinformatics



40,000 x 35%
x 5~10 splicing=
70,000 ~ 140,000

+

40,000 x 65% =
26,000

96,000 ~ 166,000

Gene identification involves 4 main stages

Find the putative coding region(s) in the sequence

Open reading frame

Find non-coding features of interest in the sequence

CpG islands
Tandemly and dispersed repeats
Promoter regions (TATA box, cap signal, CCAAT-box)
Transcription factors, Poly-A sites

Determine the exon-intron organization

Branch point signal
CT(G,A)A(C,T)

5' and 3' splice sites:
AG/GUAAGU----------PyPyPyPyPyPyPyPy-CAG/G

Identify the gene

motif, signal and pattern
Blast, FASTA
Functional studies

GENE FINDERS

Banbury Cross	*http://igs-server.cnrs-mrs.fr/igs/banbury*
FGENEH	*http://genomic.sanger.ac.uk/gf/gf.shtml*
GeneID	*http://www1.imim.es/geneid.html*
GeneMachine	*http://genome.nhgri.nih.gov/genemachine*
GeneParser	*http://beagle.colorado.edu/_eesnyder/GeneParser.htl*
GENSCAN	*http://genes.mit.edu/GENSCAN.html*
Genotator	*http://www.fruitfly.org/_nomi/genotator/*
GRAIL	*http://compbio.ornl.gov/tools/index.shtml*
GRAIL-EXP	*http://compbio.ornl.gov/grailexp/*
HMMgene	*http://www.cbs.dtu.dk/services/HMMgene/*
MZEF	*http://www.cshl.org/genefinder*
PROCRUSTES	*http://www-hto.usc.edu/software/procrustes*
RepeatMasker	*http://ftp.genome.washington.edu/RM/RepeatMasker.html*
Sputnik	*http://rast.abajian.com/sputnik/*



Bioinformatics

Gene prediction programs:

GRAIL



GENSCAN
Web Server at MIT

```
          \\|//
          (o o)
-. .-.   .-oOOo~(_)~oOOo-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.
||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|
|/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X||
'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-
```

Bioinformatics

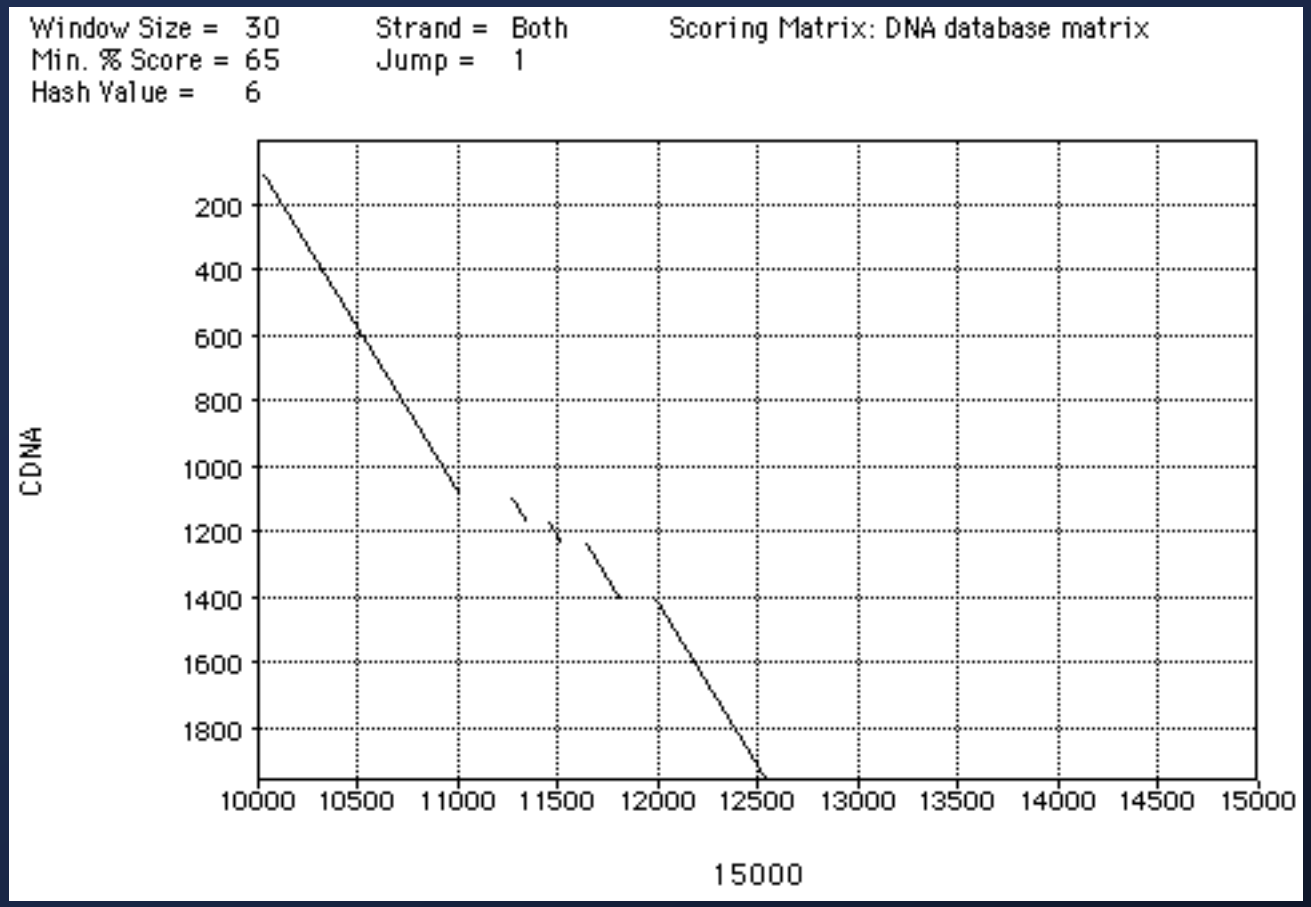
GENSCAN Performance Data

Method	Accuracy per nucleotide Sn	Sp	AC	Accuracy per exon Sn	Sp	(Sn+Sp)/2	ME	WE
GENSCAN	0.93	0.93	0.91	0.78	0.81	0.80	0.09	0.05
FGENEH	0.77	0.85	0.78	0.61	0.61	0.61	0.15	0.11
GeneID	0.63	0.81	0.67	0.44	0.45	0.45	0.28	0.24
GenePa2	0.66	0.79	0.66	0.35	0.39	0.37	0.29	0.17
GenLang	0.72	0.75	0.69	0.50	0.49	0.50	0.21	0.21
GRAILII	0.72	0.84	0.75	0.36	0.41	0.38	0.25	0.10
SORFIND	0.71	0.85	0.73	0.42	0.47	0.45	0.24	0.14
Xpound	0.61	0.82	0.68	0.15	0.17	0.16	0.32	0.13

Bioinformatics

Accuracy as a Function of Exon Length

Length range (bp)	Annotated exons No.	%Exact	%Part	%Miss	Predicted exons No.	%Exact	%Part	%Wrong
<= 24	89	38	8	52	44	77	11	11
25 - 49	163	58	15	25	124	76	6	18
50 - 74	248	70	12	16	204	85	9	6
75 - 99	382	85	8	6	389	84	6	10
100 - 124	351	84	9	7	366	81	8	11
125 - 149	425	88	8	4	460	81	10	7
150 - 174	261	88	9	2	283	81	11	7
175 - 199	167	91	7	2	188	81	12	7
200 - 299	353	90	8	1	390	82	8	8
>= 300	211	66	19	1	204	69	20	10
Total	2650	81	10	8	2678	81	10	9



GRAIL 2

10138 - 11018	+
12608 - 12748	x
13530 - 13923	x

GENSCAN

10138 - 11018	+
11268 - 11341	+
11450 - 11518	+
11644 - 11808	+
11989 - 12144	+
12360 - 12454	x
12608 - 12748	x

FGENES

1880 - 1908	x
5061 - 5175	x
5900 - 6049	x
8317 - 8544	+
10357 - 11018	+
11268 - 11341	+
11450 - 11518	+
11644 - 11864	+
polyA: 12521	+

cDNA and genomic DNA alignment and matrix analysis:

Gene identification involves 4 main stages

Find the putative coding region(s) in the sequence

Open reading frame

Find non-coding features of interest in the sequence

CpG islands
Tandemly and dispersed repeats
Promoter regions (TATA box, cap signal, CCAAT-box)
Transcription factors, Poly-A sites

Determine the exon-intron organization

Branch point signal
CT(G,A)A(C,T)

5' and 3' splice sites:
AG/GUAAGU----------PyPyPyPyPyPyPyPy-CAG/G

Identify the gene

motif, signal and pattern
Blast, FASTA
Functional studies

ORF Finder (Open Reading Frame Finder)

NCBI

Tools
for data mining

GenBank
sequence submission support and software

FTP site
download data and software

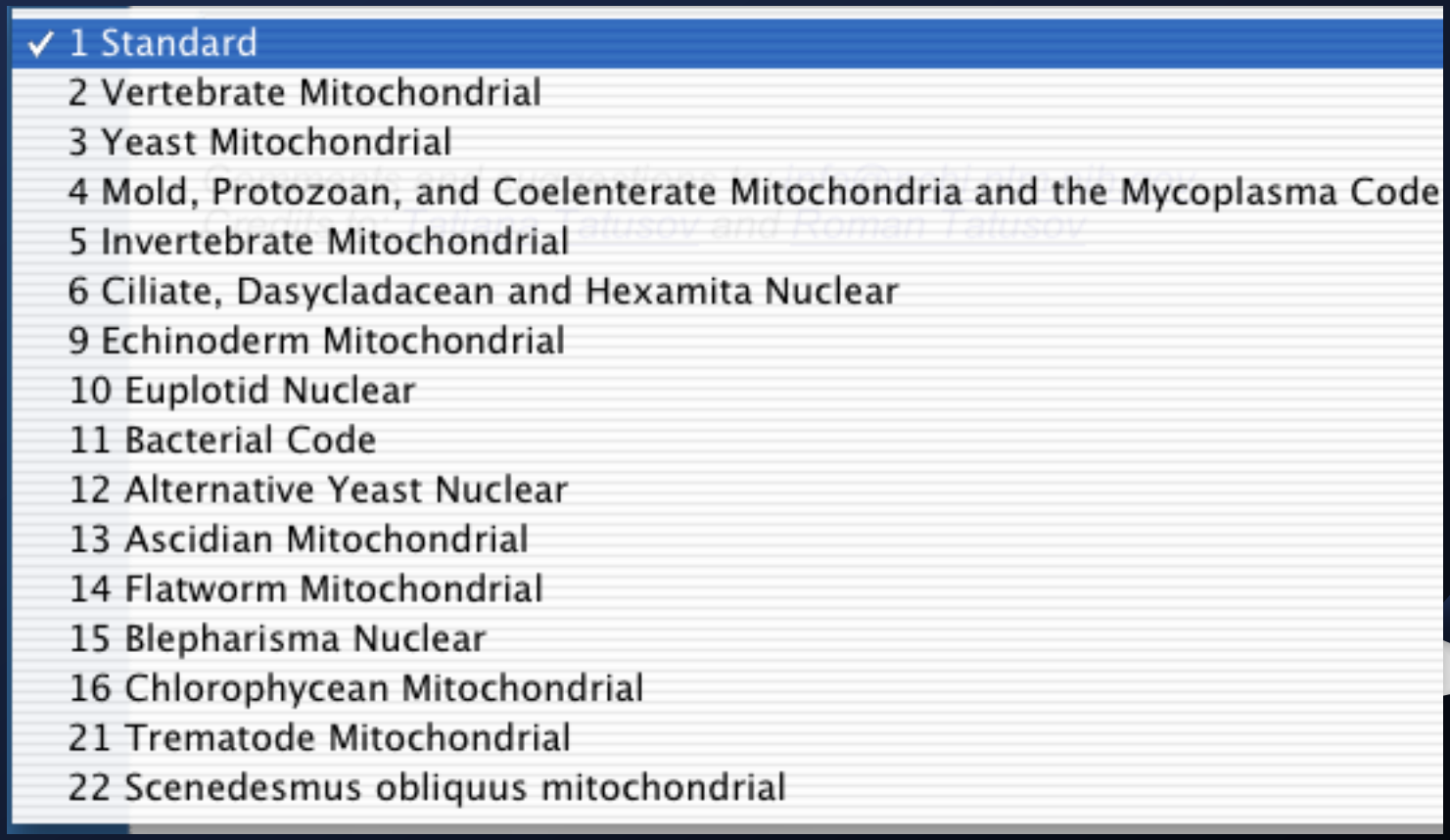
The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database.
This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server. The ORF Finder should be helpful in preparing complete and accurate sequence submissions. It is also packaged with the Sequin sequence submission software.

Enter GI or ACCESSION [ ] OrfFind Clear
or sequence in FASTA format

```
ACCAGTGTACAGGTCACCCATCAGG
ACCACTCAGTTAAGGAAGAT
CCTGAAGCAGTTCAGAAGGAATAAG
CATTCCTTCTTTCACAGGCATCAGG
AATTGTCAAATGATGATTAT
GAGTTCCCTAAACAAAAGCAAAGAT
GCATTTTCACTGCAATGTCAAAGTTT
AGCTGCT
```

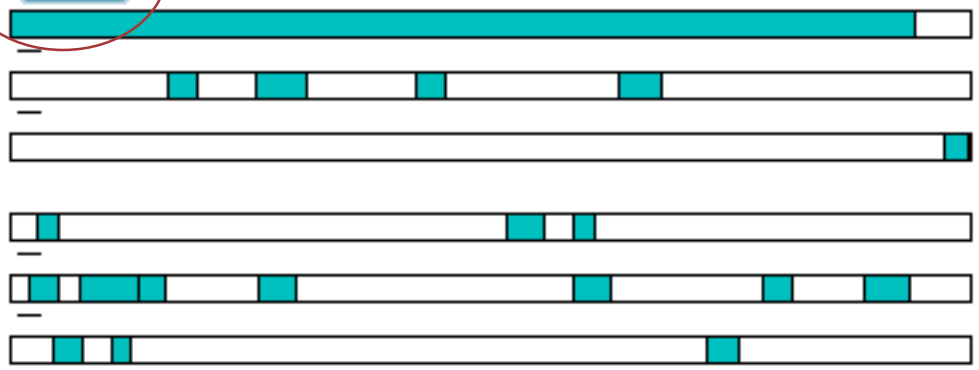
FROM: [ ] **TO:** [ ]

Genetic codes
1 Standard


✓ 1 Standard
2 Vertebrate Mitochondrial
3 Yeast Mitochondrial
4 Mold, Protozoan, and Coelenterate Mitochondria and the Mycoplasma Code
5 Invertebrate Mitochondrial
6 Ciliate, Dasycladacean and Hexamita Nuclear
9 Echinoderm Mitochondrial
10 Euplotid Nuclear
11 Bacterial Code
12 Alternative Yeast Nuclear
13 Ascidian Mitochondrial
14 Flatworm Mitochondrial
15 Blepharisma Nuclear
16 Chlorophycean Mitochondrial
21 Trematode Mitochondrial
22 Scenedesmus obliquus mitochondrial

Anonymous

View | 1 GenBank | Redraw | 100 | SixFrames

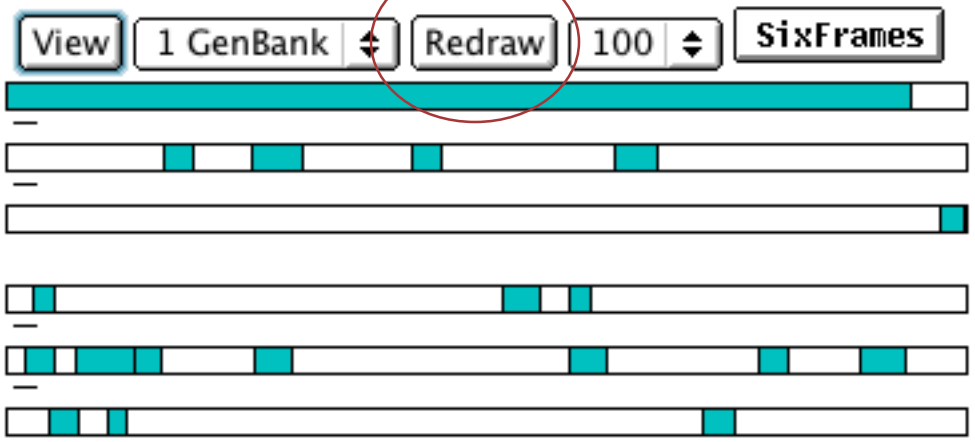


Frame	from	to	Length
+1	1	4344	4344
-2	342	620	279
+2	1181	1432	252
-2	4104	4325	222
+2	2921	3127	207
-1	2386	2571	186
-2	1194	1376	183
-2	2712	2891	180
-3	3344	3505	162
+2	1955	2098	144
-3	215	358	144
-2	96	236	141
+2	767	901	135
-2	3621	3752	132
-2	627	755	129
+3	4482	4607	127
-1	130	240	111
-1	2707	2808	102
-3	488	589	102

```
LOCUS       tmpseq_0                4608 bp            linear       31-OCT-2002
DEFINITION  No definition line found.
ACCESSION   tmpseq_0
VERSION
KEYWORDS    .
SOURCE      Unknown.
  ORGANISM  Unknown.
            Unclassified.
FEATURES             Location/Qualifiers
     source          1..4608
BASE COUNT     1282 a   1179 c   1058 g   1089 t
ORIGIN
        1 atggagaata gtcttagatg tgtttgggta cccaagctgg cttttgtact cttcggagct
       61 tccttgctca gcgcgcatct tcaagtaacc ggttttcaaa ttaaagcttt cacagcactg
      121 cgcttcctct cagaaccttc tgatgccgtc acaatgcggg gaggaaatgt cctcctcgac
      181 tgctccgcgg agtccgaccg aggagttcca gtgatcaagt ggaagaaaga tggcattcat
      241 ctggccttgg gaatggatga aaggaagcag caactttcaa atgggtctct gctgatacaa
      301 aacatacttc attccagaca ccacaagcca gatgagggac tttaccaatg tgaggcatct
      361 ttaggagatt ctggctcaat tattagtcgg acagcaaaag ttgcagtagc aggaccactg
      421 aggttccttt cacagacaga atctgtcaca gccttcatgg gagacacagt gctactcaag
      481 tgtgaagtca ttggggagcc catgccaaca atccactggc agaagaacca acaagacctg
      541 actccaatcc caggtgactc ccgagtggtg gtcttgccct ctggagcatt gcagatcagc
      601 cgactccaac cgggggacat tggaatttac cgatgctcag ctcgaaatcc agccagctca
      661 agaacaggaa atgaagcaga agtcagaatt ttatcagatc caggactgca tagacagctg
      721 tattttctgc aaagaccatc caatgtagta gccattgaag gaaaagatgc tgtcctggaa
      781 tgttgtgttt ctggctatcc tccaccaagt tttacctggt tacgaggcga ggaagtcatc
      841 caactcaggt ctaaaaagta ttctttattg ggtggaagca acttgcttat ctccaatgtg
      901 acagatgatg acagtggaat gtatacctgt gttgtcacat ataaaaatga gaatattagt
      961 gcctctgcag agctcacagt cttggttccg ccatggtttt taaatcatcc ttccaacctg
     1021 tatgcctatg aaagcatgga tattgagttt gaatgtacag tctctggaaa gcctgtgccc
     1081 actgtgaatt ggatgaagaa tggagatgtg gtcattccta gtgattattt tcagatagtg
     1141 ggaggaagca acttacggat acttggggtg gtgaagtcag atgaaggctt ttatcaatgt
     1201 gtggctgaaa atgaggctgg aaatgcccag accagtgcac agctcattgt ccctaagcct
     1261 gcaatcccaa gctccagtgt cctcccttcg gctcccagag atgtggtccc tgtcttggtt
     1321 tccagccgat ttgtccgtct cagctggcgc ccacctgcag aagcgaaagg gaacattcaa
```



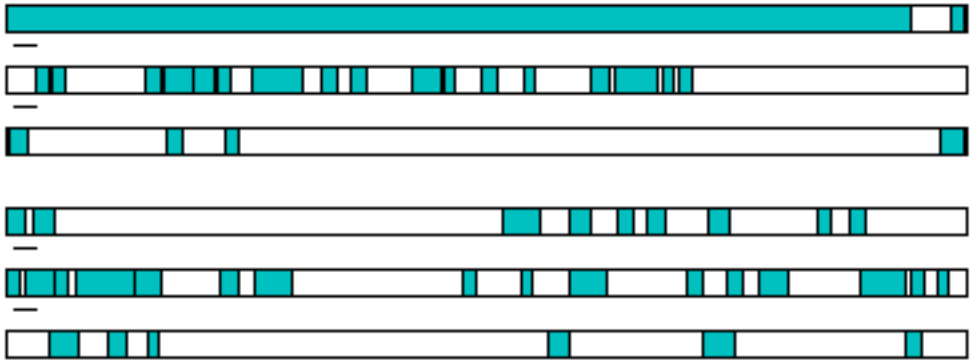
Anonymous



Frame	from	to	Length
+1	1	4344	4344
-2	342	620	279
+2	1181	1432	252
-2	4104	4325	222
+2	2921	3127	207
-1	2386	2571	186
-2	1194	1376	183
-2	2712	2891	180
-3	3344	3505	162
+2	1955	2098	144
-3	215	358	144
-2	96	236	141
+2	767	901	135
-2	3621	3752	132
-2	627	755	129
+3	4482	4607	127
-1	130	240	111
-1	2707	2808	102
-3	488	589	102

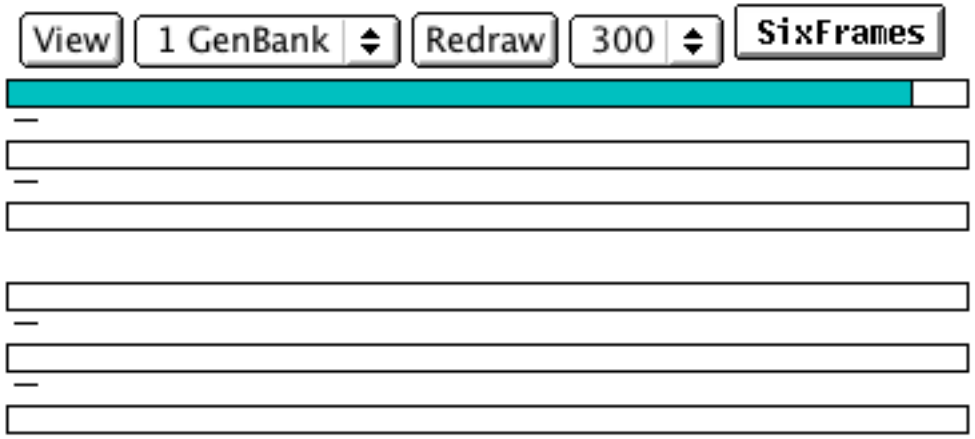
Anonymous

View | 1 GenBank | Redraw | 50 | SixFrames



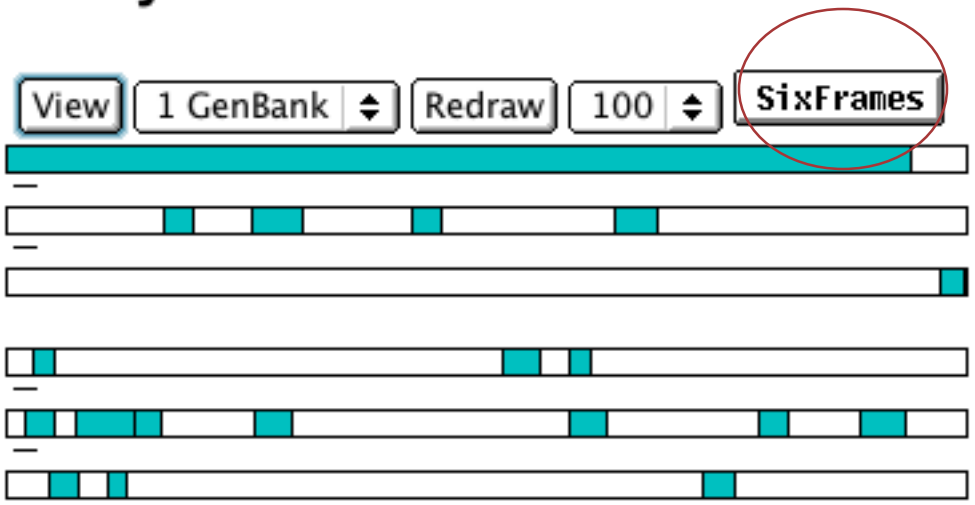
Frame	from	to	Length
+1	1	4344	4344
-2	342	620	279
+2	1181	1432	252
-2	4104	4325	222
+2	2921	3127	207
-1	2386	2571	186
-2	1194	1376	183
-2	2712	2891	180
-3	3344	3505	162
+2	1955	2098	144
-3	215	358	144
-2	96	236	141
+2	767	901	135
-2	3621	3752	132
-2	627	755	129
+3	4482	4607	127
-1	130	240	111
-1	2707	2808	102
-3	488	589	102
+2	905	1003	99
-1	1	99	99
-1	3376	3471	96
-3	2609	2704	96
-1	3079	3171	93
+2	2807	2896	90
-2	1026	1115	90
-1	4048	4134	87
+3	18	104	87

Anonymous



Frame	from	to	Length
+1	1	4344	4344

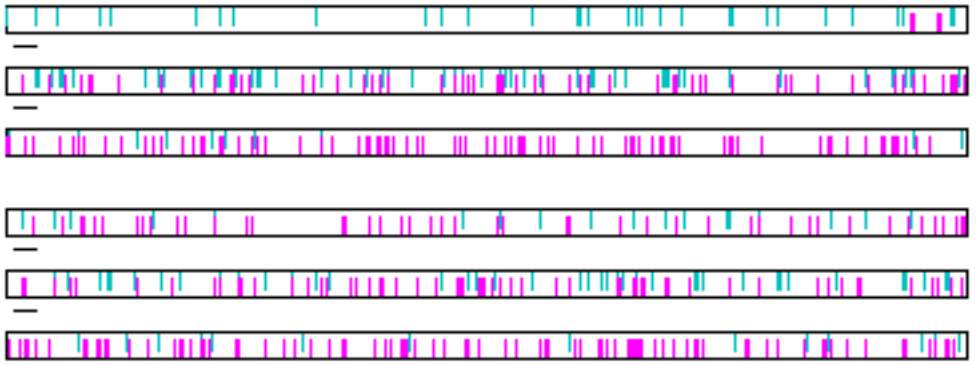
Anonymous

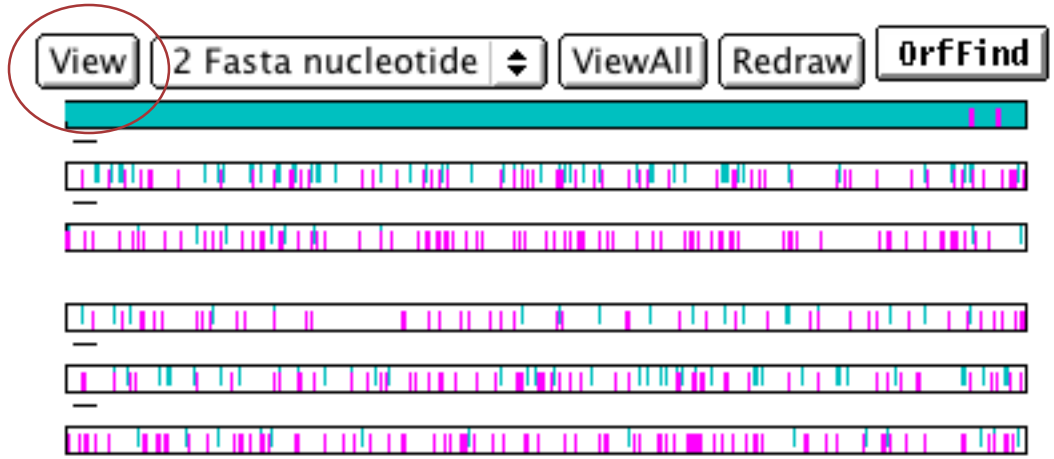


Frame	from	to	Length
+1	1	4344	4344
-2	342	620	279
+2	1181	1432	252
-2	4104	4325	222
+2	2921	3127	207
-1	2386	2571	186
-2	1194	1376	183
-2	2712	2891	180
-3	3344	3505	162
+2	1955	2098	144
-3	215	358	144
-2	96	236	141
+2	767	901	135
-2	3621	3752	132
-2	627	755	129
+3	4482	4607	127
-1	130	240	111
-1	2707	2808	102
-3	488	589	102

Anonymous

View | 2 Fasta nucleotide | ViewAll | Redraw | OrfFind

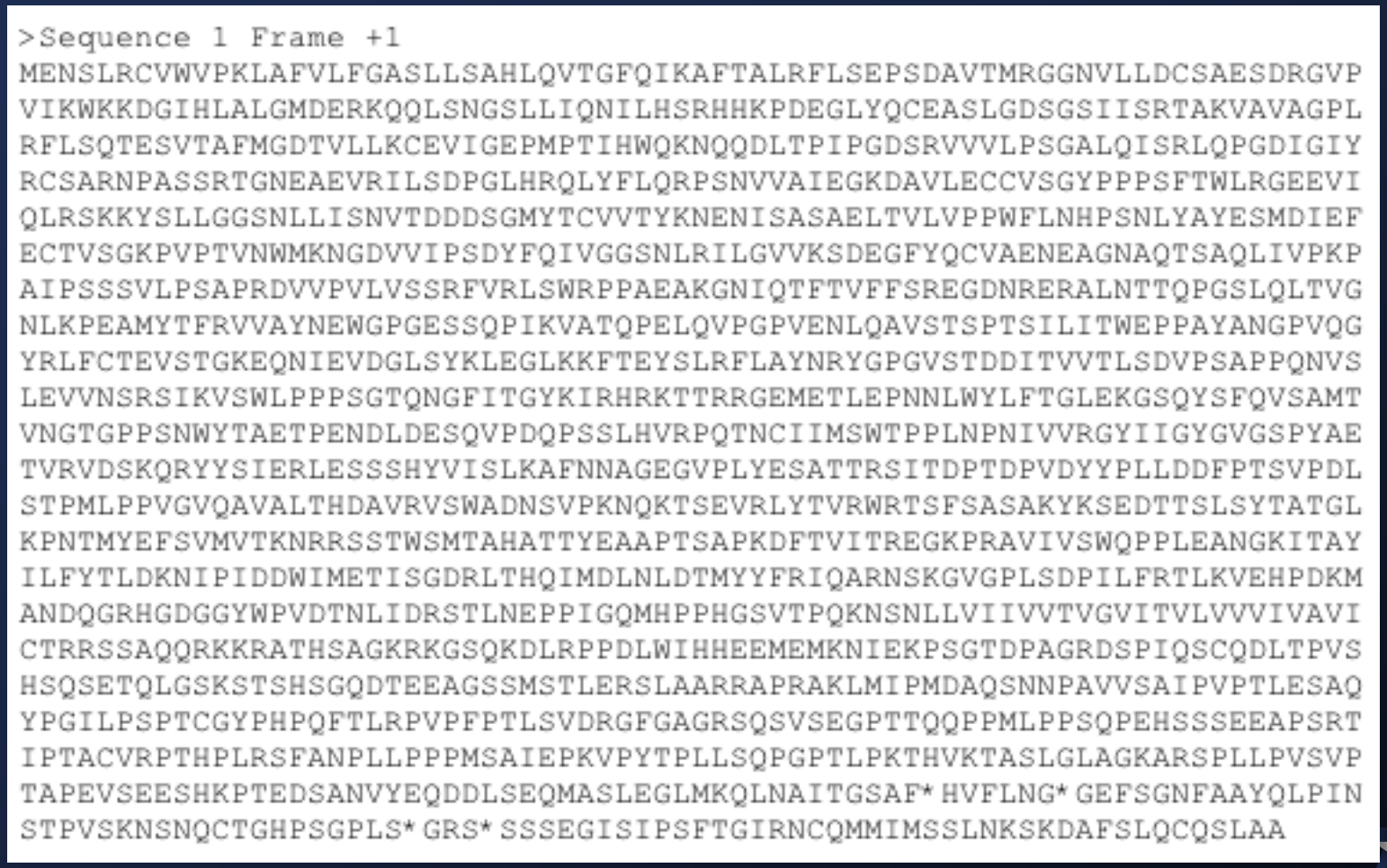




Length: **1535 aa**

Alternative Initiation Codons

```
  1 atggagaatagtcttagatgtgtttgggtacccaagctggcttttgtactcttcggagct
    M  E  N  S  L  R  C  V  W  V  P  K  L  A  F  V  L  F  G  A
 61 tccttgctcagcgcgcatcttcaagtaaccggttttcaaattaaagctttcacagcactg
    S  L  L  S  A  H  L  Q  V  T  G  F  Q  I  K  A  F  T  A  L
121 cgcttcctctcagaaccttctgatgccgtcacaatgcggggaggaaatgtcctcctcgac
    R  F  L  S  E  P  S  D  A  V  T  M  R  G  G  N  V  L  L  D
181 tgctccgcggagtccgaccgaggagttccagtgatcaagtggaagaaagatggcattcat
    C  S  A  E  S  D  R  G  V  P  V  I  K  W  K  K  D  G  I  H
241 ctggccttgggaatggatgaaaggaagcagcaactttcaaatgggtctctgctgatacaa
    L  A  L  G  M  D  E  R  K  Q  Q  L  S  N  G  S  L  L  I  Q
301 aacatacttcattccagacaccacaagccagatgagggactttaccaatgtgaggcatct
    N  I  L  H  S  R  H  H  K  P  D  E  G  L  Y  Q  C  E  A  S
361 ttaggagattctggctcaattattagtcggacagcaaaagttgcagtagcaggaccactg
    L  G  D  S  G  S  I  I  S  R  T  A  K  V  A  V  A  G  P  L
421 aggttcctttcacagacagaatctgtcacagccttcatgggagacacagtgctactcaag
    R  F  L  S  Q  T  E  S  V  T  A  F  M  G  D  T  V  L  L  K
481 tgtgaagtcattggggagcccatgccaacaatccactggcagaagaaccaacaagacctg
    C  E  V  I  G  E  P  M  P  T  I  H  W  Q  K  N  Q  Q  D  L
```


>Sequence 1 Frame +1
MENSLRCVWVPKLAFVLFGASLLSAHLQVTGFQIKAFTALRFLSEPSDAVTMRGGNVLLDCSAESDRGVP
VIKWKKDGIHLALGMDERKQQLSNGSLLIQNILHSRHHKPDEGLYQCEASLGDSGSIISRTAKVAVAGPL
RFLSQTESVTAFMGDTVLLKCEVIGEPMPTIHWQKNQQDLTPIPGDSRVVVLPSGALQISRLQPGDIGIY
RCSARNPASSRTGNEAEVRILSDPGLHRQLYFLQRPSNVVAIEGKDAVLECCVSGYPPPSFTWLRGEEVI
QLRSKKYSLLGGSNLLISNVTDDDSGMYTCVVTYKNENISASAELTVLVPPWFLNHPSNLYAYESMDIEF
ECTVSGKPVPTVNWMKNGDVVIPSDYFQIVGGSNLRILGVVKSDEGFYQCVAENEAGNAQTSAQLIVPKP
AIPSSSVLPSAPRDVVPVLVSSRFVRLSWRPPAEAKGNIQTFTVFFSREGDNRERALNTTQPGSLQLTVG
NLKPEAMYTFRVVAYNEWGPGESSQPIKVATQPELQVPGPVENLQAVSTSPTSILITWEPPAYANGPVQG
YRLFCTEVSTGKEQNIEVDGLSYKLEGLKKFTEYSLRFLAYNRYGPGVSTDDITVVTLSDVPSAPPQNVS
LEVVNSRSIKVSWLPPPSGTQNGFITGYKIRHRKTTRRGEMETLEPNNLWYLFTGLEKGSQYSFQVSAMT
VNGTGPPSNWYTAETPENDLDESQVPDQPSSLHVRPQTNCIIMSWTPPLNPNIVVRGYIIGYGVGSPYAE
TVRVDSKQRYYSIERLESSSHYVISLKAFNNAGEGVPLYESATTRSITDPTDPVDYYPLLDDFPTSVPDL
STPMLPPVGVQAVALTHDAVRVSWADNSVPKNQKTSEVRLYTVRWRTSFSASAKYKSEDTTSLSYTATGL
KPNTMYEFSVMVTKNRRSSTWSMTAHATTYEAAPTSAPKDFTVITREGKPRAVIVSWQPPLEANGKITAY
ILFYTLDKNIPIDDWIMETISGDRLTHQIMDLNLDTMYYFRIQARNSKGVGPLSDPILFRTLKVEHPDKM
ANDQGRHGDGGYWPVDTNLIDRSTLNEPPIGQMHPPHGSVTPQKNSNLLVIIVVTVGVITVLVVVIVAVI
CTRRSSAQQRKKRATHSAGKRKGSQKDLRPPDLWIHHEEMEMKNIEKPSGTDPAGRDSPIQSCQDLTPVS
HSQSETQLGSKSTSHSGQDTEEAGSSMSTLERSLAARRAPRAKLMIPMDAQSNNPAVVSAIPVPTLESAQ
YPGILPSPTCGYPHPQFTLRPVPFPTLSVDRGFGAGRSQSVSEGPTTQQPPMLPPSQPEHSSSEEAPSRT
IPTACVRPTHPLRSFANPLLPPPMSAIEPKVPYTPLLSQPGPTLPKTHVKTASLGLAGKARSPLLPVSVP
TAPEVSEESHKPTEDSANVYEQDDLSEQMASLEGLMKQLNAITGSAF*HVFLNG*GEFSGNFAAYQLPIN
STPVSKNSNQCTGHPSGPLS*GRS*SSSEGISIPSFTGIRNCQMMIMSSLNKSKDAFSLQCQSLAA

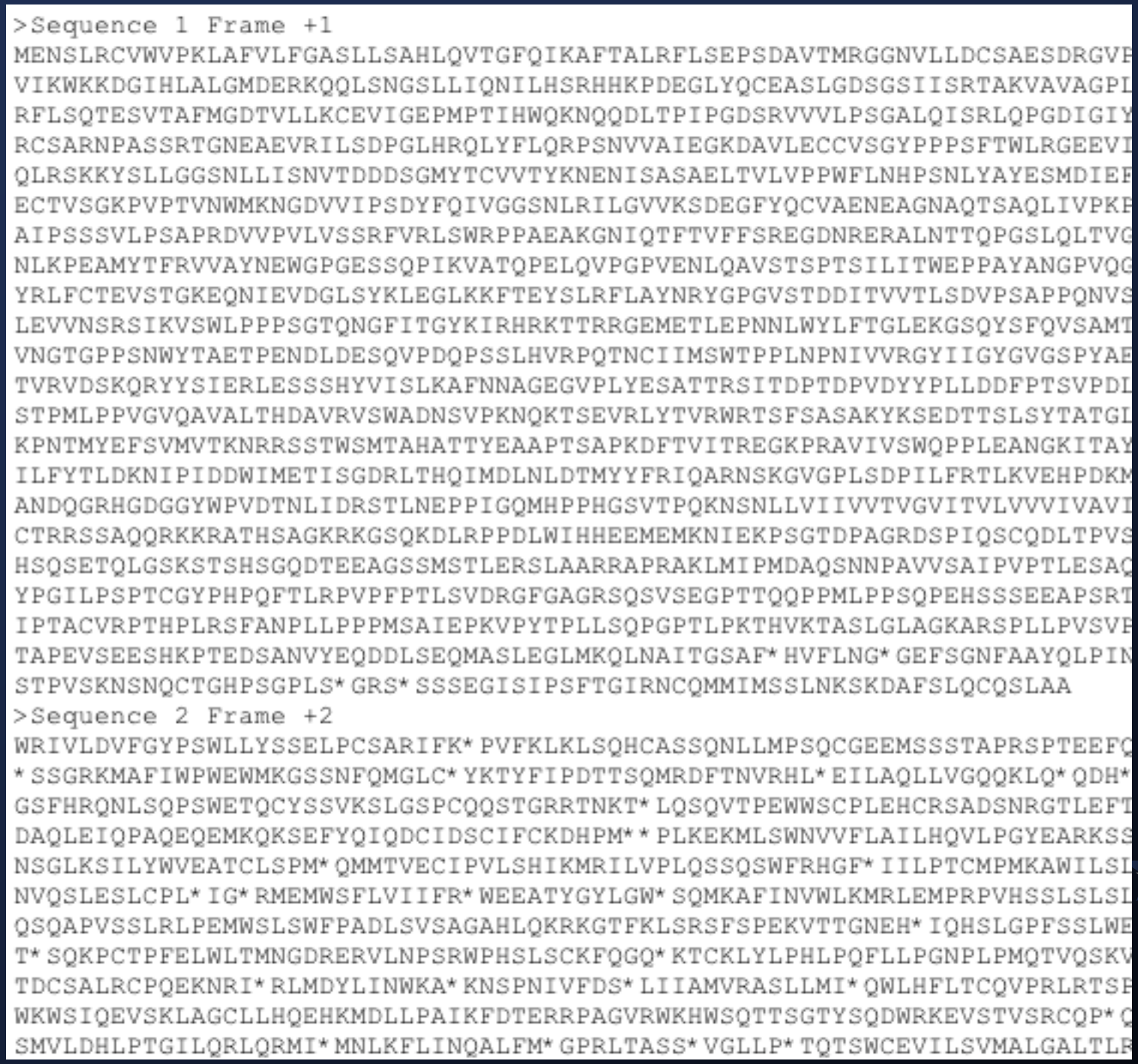


Length: **1535 aa**

Alternative Initiation Codons

```
  1 atggagaatagtcttagatgtgtttgggtacccaagctggcttttgtactcttcggagct
    M  E  N  S  L  R  C  V  W  V  P  K  L  A  F  V  L  F  G  A
 61 tccttgctcagcgcgcatcttcaagtaaccggttttcaaattaaagctttcacagcactg
    S  L  L  S  A  H  L  Q  V  T  G  F  Q  I  K  A  F  T  A  L
121 cgcttcctctcagaaccttctgatgccgtcacaatgcggggaggaaatgtcctcctcgac
    R  F  L  S  E  P  S  D  A  V  T  M  R  G  G  N  V  L  L  D
181 tgctccgcggagtccgaccgaggagttccagtgatcaagtggaagaaagatggcattcat
    C  S  A  E  S  D  R  G  V  P  V  I  K  W  K  K  D  G  I  H
241 ctggccttgggaatggatgaaaggaagcagcaactttcaaatgggtctctgctgatacaa
    L  A  L  G  M  D  E  R  K  Q  Q  L  S  N  G  S  L  L  I  Q
301 aacatacttcattccagacaccacaagccagatgagggactttaccaatgtgaggcatct
    N  I  L  H  S  R  H  H  K  P  D  E  G  L  Y  Q  C  E  A  S
361 ttaggagattctggctcaattattagtcggacagcaaaagttgcagtagcaggaccactg
    L  G  D  S  G  S  I  I  S  R  T  A  K  V  A  V  A  G  P  L
421 aggttcctttcacagacagaatctgtcacagccttcatgggagacacagtgctactcaag
    R  F  L  S  Q  T  E  S  V  T  A  F  M  G  D  T  V  L  L  K
481 tgtgaagtcattggggagcccatgccaacaatccactggcagaagaaccaacaagacctg
    C  E  V  I  G  E  P  M  P  T  I  H  W  Q  K  N  Q  Q  D  L
```

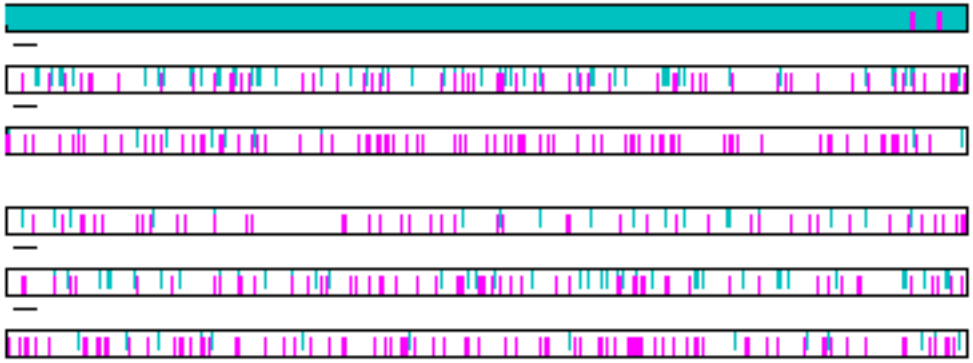



>Sequence 1 Frame +1
MENSLRCVWVPKLAFVLFGASLLSAHLQVTGFQIKAFTALRFLSEPSDAVTMRGGNVLLDCSAESDRGVP
VIKWKKDGIHLALGMDERKQQLSNGSLLIQNILHSRHHKPDEGLYQCEASLGDSGSIISRTAKVAVAGPL
RFLSQTESVTAFMGDTVLLKCEVIGEPMPTIHWQKNQQDLTPIPGDSRVVVLPSGALQISRLQPGDIGIY
RCSARNPASSRTGNEAEVRILSDPGLHRQLYFLQRPSNVVAIEGKDAVLECCVSGYPPPSFTWLRGEEVI
QLRSKKYSLLGGSNLLISNVTDDDSGMYTCVVTYKNENISASAELTVLVPPWFLNHPSNLYAYESMDIEF
ECTVSGKPVPTVNWMKNGDVVIPSDYFQIVGGSNLRILGVVKSDEGFYQCVAENEAGNAQTSAQLIVPKP
AIPSSSVLPSAPRDVVPVLVSSRFVRLSWRPPAEAKGNIQTFTVFFSREGDNRERALNTTQPGSLQLTVG
NLKPEAMYTFRVVAYNEWGPGESSQPIKVATQPELQVPGPVENLQAVSTSPTSILITWEPPAYANGPVQG
YRLFCTEVSTGKEQNIEVDGLSYKLEGLKKFTEYSLRFLAYNRYGPGVSTDDITVVTLSDVPSAPPQNVS
LEVVNSRSIKVSWLPPPSGTQNGFITGYKIRHRKTTRRGEMETLEPNNLWYLFTGLEKGSQYSFQVSAMT
VNGTGPPSNWYTAETPENDLDESQVPDQPSSLHVRPQTNCIIMSWTPPLNPNIVVRGYIIGYGVGSPYAE
TVRVDSKQRYYSIERLESSSHYVISLKAFNNAGEGVPLYESATTRSITDPTDPVDYYPLLDDFPTSVPDL
STPMLPPVGVQAVALTHDAVRVSWADNSVPKNQKTSEVRLYTVRWRTSFSASAKYKSEDTTSLSYTATGL
KPNTMYEFSVMVTKNRRSSTWSMTAHATTYEAAPTSAPKDFTVITREGKPRAVIVSWQPPLEANGKITAY
ILFYTLDKNIPIDDWIMETISGDRLTHQIMDLNLDTMYYFRIQARNSKGVGPLSDPILFRTLKVEHPDKM
ANDQGRHGDGGYWPVDTNLIDRSTLNEPPIGQMHPPHGSVTPQKNSNLLVIIVVTVGVITVLVVVIVAVI
CTRRSSAQQRKKRATHSAGKRKGSQKDLRPPDLWIHHEEMEMKNIEKPSGTDPAGRDSPIQSCQDLTPVS
HSQSETQLGSKSTSHSGQDTEEAGSSMSTLERSLAARRAPRAKLMIPMDAQSNNPAVVSAIPVPTLESAQ
YPGILPSPTCGYPHPQFTLRPVPFPTLSVDRGFGAGRSQSVSEGPTTQQPPMLPPSQPEHSSSEEAPSRT
IPTACVRPTHPLRSFANPLLPPPMSAIEPKVPYTPLLSQPGPTLPKTHVKTASLGLAGKARSPLLPVSVP
TAPEVSEESHKPTEDSANVYEQDDLSEQMASLEGLMKQLNAITGSAF*HVFLNG*GEFSGNFAAYQLPIN
STPVSKNSNQCTGHPSGPLS*GRS*SSSEGISIPSFTGIRNCQMMIMSSLNKSKDAFSLQCQSLAA
>Sequence 2 Frame +2
WRIVLDVFGYPSWLLYSSELPCSARIFK*PVFKLKLSQHCASSQNLLMPSQCGEEMSSSTAPRSPTEEFQ
*SSGRKMAFIWPWEWMKGSSNFQMGLC*YKTYFIPDTTSQMRDFTNVRHL*EILAQLLVGQQKLQ*QDH*
GSFHRQNLSQPSWETQCYSSVKSLGSPCQQSTGRRTNKT*LQSQVTPEWWSCPLEHCRSADSNRGTLEFT
DAQLEIQPAQEQEMKQKSEFYQIQDCIDSCIFCKDHPM**PLKEKMLSWNVVFLAILHQVLPGYEARKSS
NSGLKSILYWVEATCLSPM*QMMTVECIPVLSHIKMRILVPLQSSQSWFRHGF*IILPTCMPMKAWILSL
NVQSLESLCPL*IG*RMEMWSFLVIIFR*WEEATYGYLGW*SQMKAFINVWLKMRLEMPRPVHSSLSLSL
QSQAPVSSLRLPEMWSLSWFPADLSVSAGAHLQKRKGTFKLSRSFSPEKVTTGNEH*IQHSLGPFSSLWE
T*SQKPCTPFELWLTMNGDRERVLNPSRWPHSLSCKFQGQ*KTCKLYLPHLPQFLLPGNPLPMQTVQSKV
TDCSALRCPQEKNRI*RLMDYLINWKA*KNSPNIVFDS*LIIAMVRASLLMI*QWLHFLTCQVPRLRTSP
WKWSIQEVSKLAGCLLHQEHKMDLLPAIKFDTERRPAGVRWKHWSQTTSGTYSQDWRKEVSTVSRCQP*Q
SMVLDHLPTGILQRLQRMI*MNLKFLINQALFM*GPRLTASS*VGLLP*TQTSWCEVILSVMALGALTLR

Anonymous

Program blastp Database nr BLAST with parameters Cognitor

View 3 Fasta protein ViewAll Redraw OrfFind



Length: **1535 aa**

Alternative Initiation Codons

```
  1 atggagaatagtcttagatgtgtttgggtacccaagctggcttttgtactcttcggagct
    M  E  N  S  L  R  C  V  W  V  P  K  L  A  F  V  L  F  G  A
 61 tccttgctcagcgcgcatcttcaagtaaccggttttcaaattaaagctttcacagcactg
    S  L  L  S  A  H  L  Q  V  T  G  F  Q  I  K  A  F  T  A  L
121 cgcttcctctcagaaccttctgatgccgtcacaatgcggggaggaaatgtcctcctcgac
    R  F  L  S  E  P  S  D  A  V  T  M  R  G  G  N  V  L  L  D
```

```
BLASTP 2.2.3 [Apr-24-2002]


Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query=
         (1536 letters)

Database: /usr/people3/tatusov/cog2001/xyva
           108,091 sequences; 33,061,072 total letters


                                                                  Score    E
Sequences producing significant alignments:                       (bits) Value

Cj1279c                                                             42   0.007
BH0916                                                              40   0.019
DR0009                                                              40   0.033
PA1695                                                              37   0.28
jhp0683                                                             37   0.28
HP0746                                                              36   0.36
TP0369                                                              36   0.48
mlr2363                                                             35   0.62
YPL105c                                                             35   0.62
YJR151c                                                             35   0.62
YDL140c                                                             35   0.62
Rv2082                                                              35   0.81
YIR019c                                                             35   0.81
PH0954_2                                                            34   1.4
slr0599                                                             34   1.8
CT868                                                               33   2.4
YFR007w                                                             33   2.4
```



Gene identification involves 4 main stages

Find the putative coding region(s) in the sequence

Open reading frame

Find non-coding features of interest in the sequence

CpG islands
Tandemly and dispersed repeats
Promoter regions (TATA box, cap signal, CCAAT-box)
Transcription factors, Poly-A sites

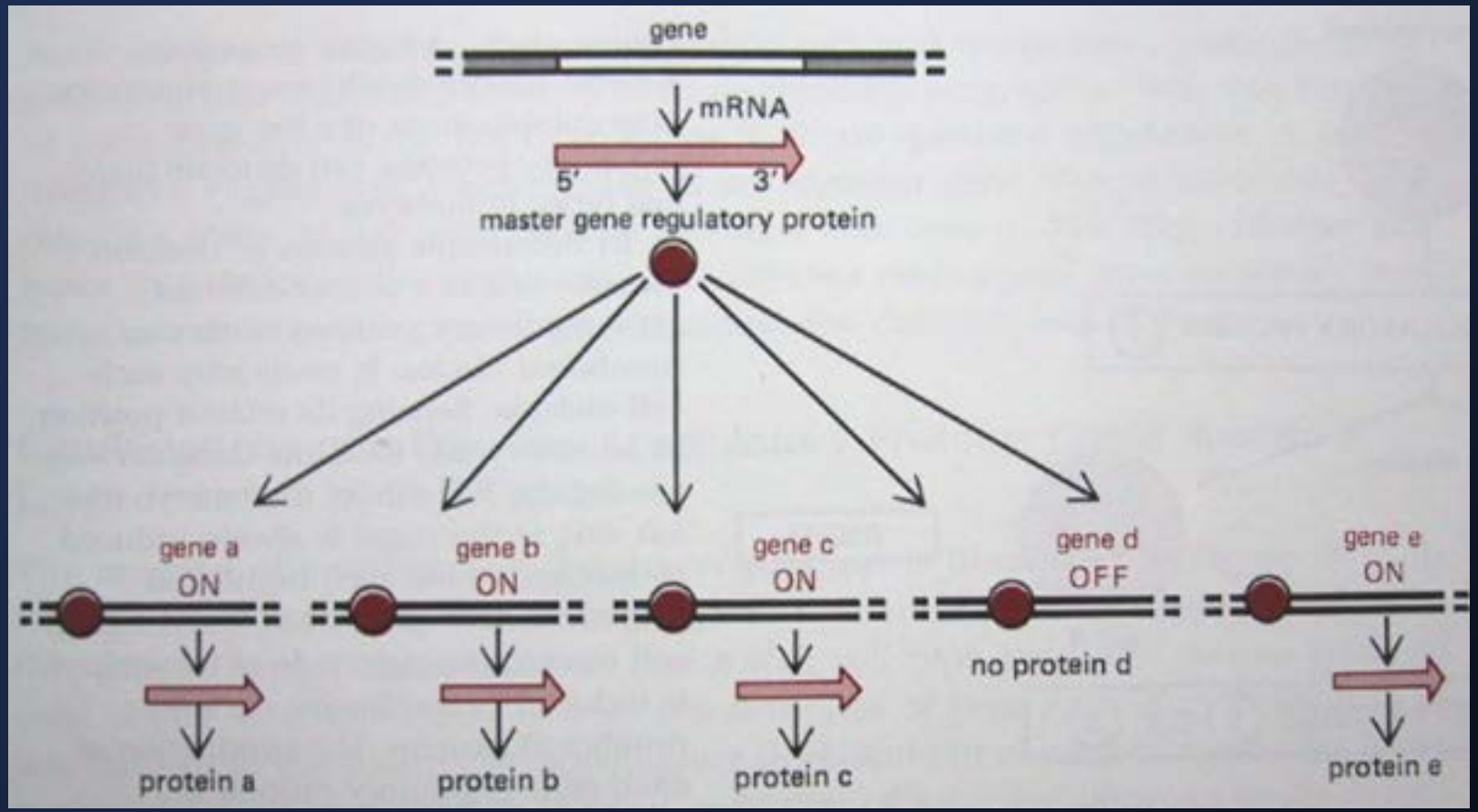
Determine the exon-intron organization

Branch point signal
CT(G,A)A(C,T)

5' and 3' splice sites:
AG/GUAAGU----------PyPyPyPyPyPyPyPy-CAG/G

Identify the gene

motif, signal and pattern
Blast, FASTA
Functional studies


gene
mRNA
5'
3'
master gene regulatory protein
gene a
ON
protein a
gene b
ON
protein b
gene c
ON
protein c
gene d
OFF
no protein d
gene e
ON
protein e

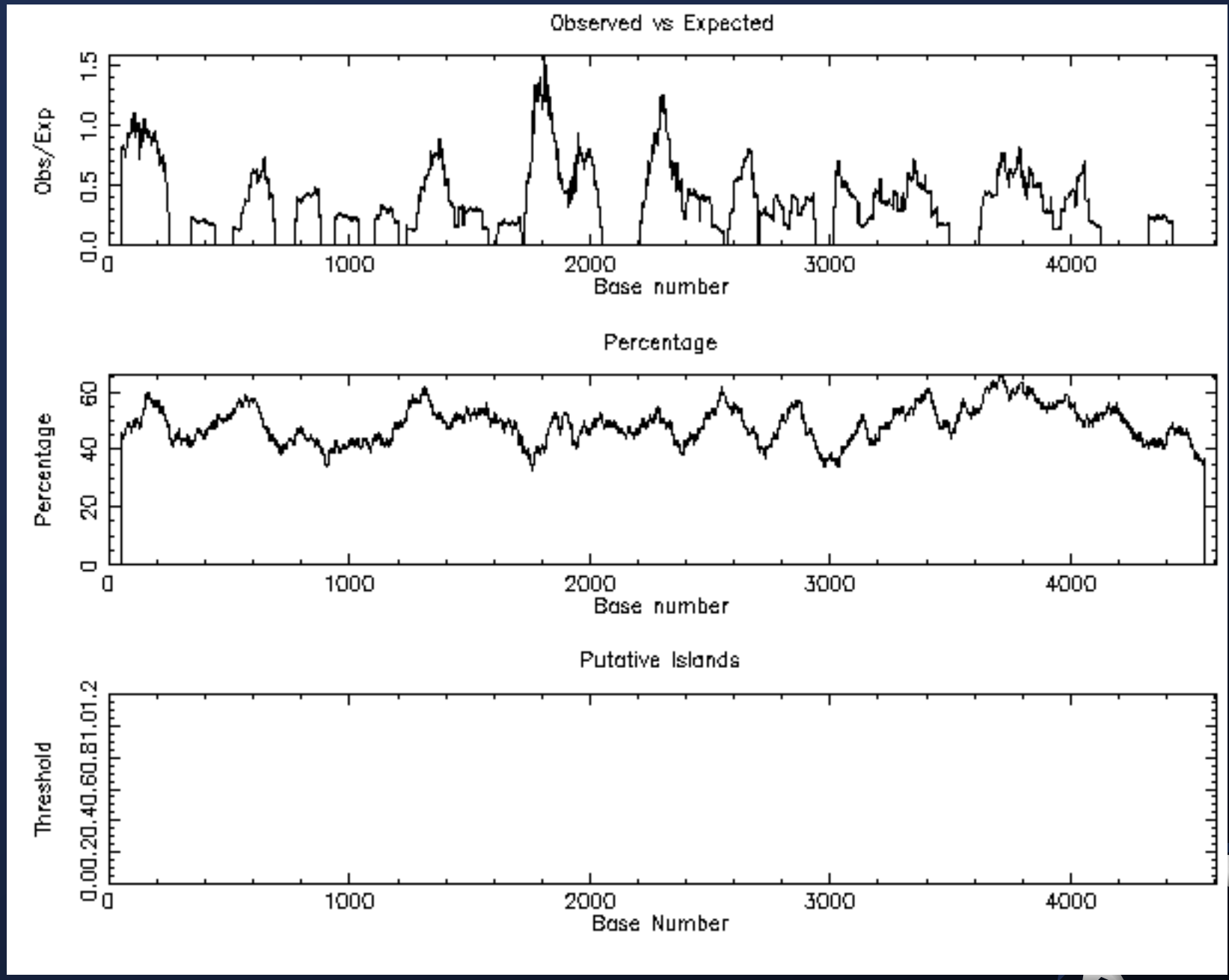
EMBOSS CpGPlot/CpGReport/Isochore

Detection of regions of genomic sequences that are rich in the CpG pattern is important because such regions are resistant to methylation and tend to be associated with genes which are frequently switched on. Regions rich in the CpG pattern are known as CpG islands.
The function of the program cpgplot is to plot CpG rich areas, and cpgreport to report all CpG rich regions.

The nuclear genomes of vertebrates are mosaics of isochores, very long stretches of DNA that are homogeneous in base composition and are compositionally correlated with the coding sequences that they embed. Isochores can be partitioned in a small number of families that cover a range of GC levels. Program isochore plots GC content over a sequence.

Program	Window	Step	Obs/Exp	MinPC	Length	Reverse	Complement
cpgplot	100	1	0.6	50	200	no	no

Enter or Paste a nucleic acid Sequence (at least 100bp) in any format: Help


Observed vs Expected
Obs/Exp
0.0
0.5
1.0
1.5
0
1000
2000
3000
4000
Base number
Percentage
Percentage
0
20
40
60
0
1000
2000
3000
4000
Base number
Putative Islands
Threshold
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
1000
2000
3000
4000
Base Number

http://sullivan.bu.edu/~mfrith/HPD.html

Human Promoter Database

See also DBTSS.

The Database: 2004 transcription start sites

Format

First column
: An accession number for the sequence used to find the transcription start site

Second column
: The accession number of the contig in the NCBI genome assembly on which the transcription start site is located

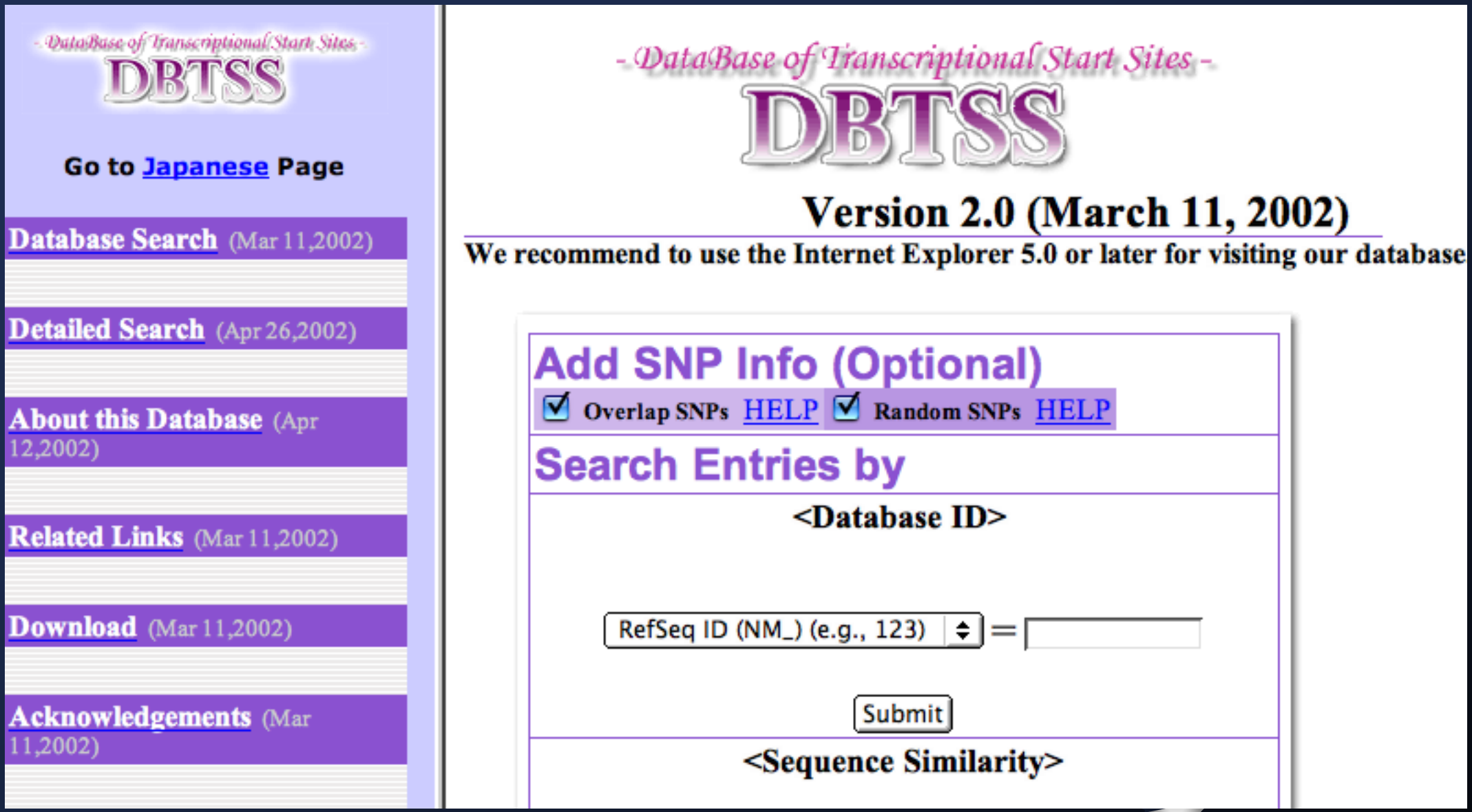
Third column
: The position of the transcription start site in this contig

Fourth column
: The sense strand for transcription

Get promoter sequences from -500 to 100 relative to the transcription start site.
Submit RESET

---

http://www.epd.isb-sib.ch/





TRANSFAC - The Transcription Factor Database

***Release 5.0** public*

The TRANSFAC database is free for users from non-profit organizations only. Users from commercial enterprises have to license the TRANSFAC databases and accompanying programs. Please read the DISCLAIMER !

Previous TRANSFAC releases	Search	Browse	Documentation	Extended Search

How to cite TRANSFAC ?
TRANSFAC-Reports
Classification of Transcription factors

http://transfac.gbf.de/

Research Group
Bioinformatics/ AG Bioinformatik

D-38124 Braunsch
Deutsch

FunSiteP 2.1 Promoter Recognition
Job Submission Form

?

SEARCH TITLE	THRESHOLD default: 1.0 range: 0.2 to 4.0	Use the EXAMPLE
my_sequence		

Enter or Paste one or more Sequence(s) in EMBL, GENBANK, FASTA or free text format:

Max. 30000 characters possible in the input window!

```
TATGAACAGGATG
ATCTGAGTGAACAAATGGCAAGTTTGGAAGGACTCATGAAGCAGCTTAATGCCATCAC
AGGCTCAGCCTT
TTAACATGTATTTCTGAATGGATGAGGTGAATTTTCCGGGAACTTTGCAGCATACCAATT
ACCCATAAAC
AGCACACCTGTGTCCAAGAACTCTAACCAGTGTACAGGTCACCCATCAGGACCACTC
AGTTAAGGAAGAT
CCTGAAGCAGTTCAGAAGGAATAAGCATTCCTTCTTTCACAGGCATCAGGAATTGTCA
AATGATGATTAT
GAGTTCCCTAAACAAAAGCAAAGATGCATTTTCACTGCAATGTCAAAGTTTAGCTGCT
```

Upload a file: Browse... **(should be < 1MB)**

```
 Sequence name:
 ID    /www/transfac/htdocs/programs/

 Sequence length = 4608
 Threshold for promoter weight=1.000000

DIRECT STRAND
 Number of found binding sites = 138
 Number of found potential transcription starts = 6

              Start        Interval
    N           of                          Confidential              Most probable
           transcription                       level      Weight    Bucher classes

    1)          820        820    823        vague      1.03       6.1.5
    2)         3623       3620   3625        vague      1.11       6.1.5
    3)         3708       3708   3711        vague      1.04       6.1.6
    4)         3739       3737   3739        vague      1.09       6.1.4
    5)         3782       3781   3785        moderate   1.40       6.1.6
    6)         3800       3796   3800        vague      1.02       6.1.4 > 6.1.3

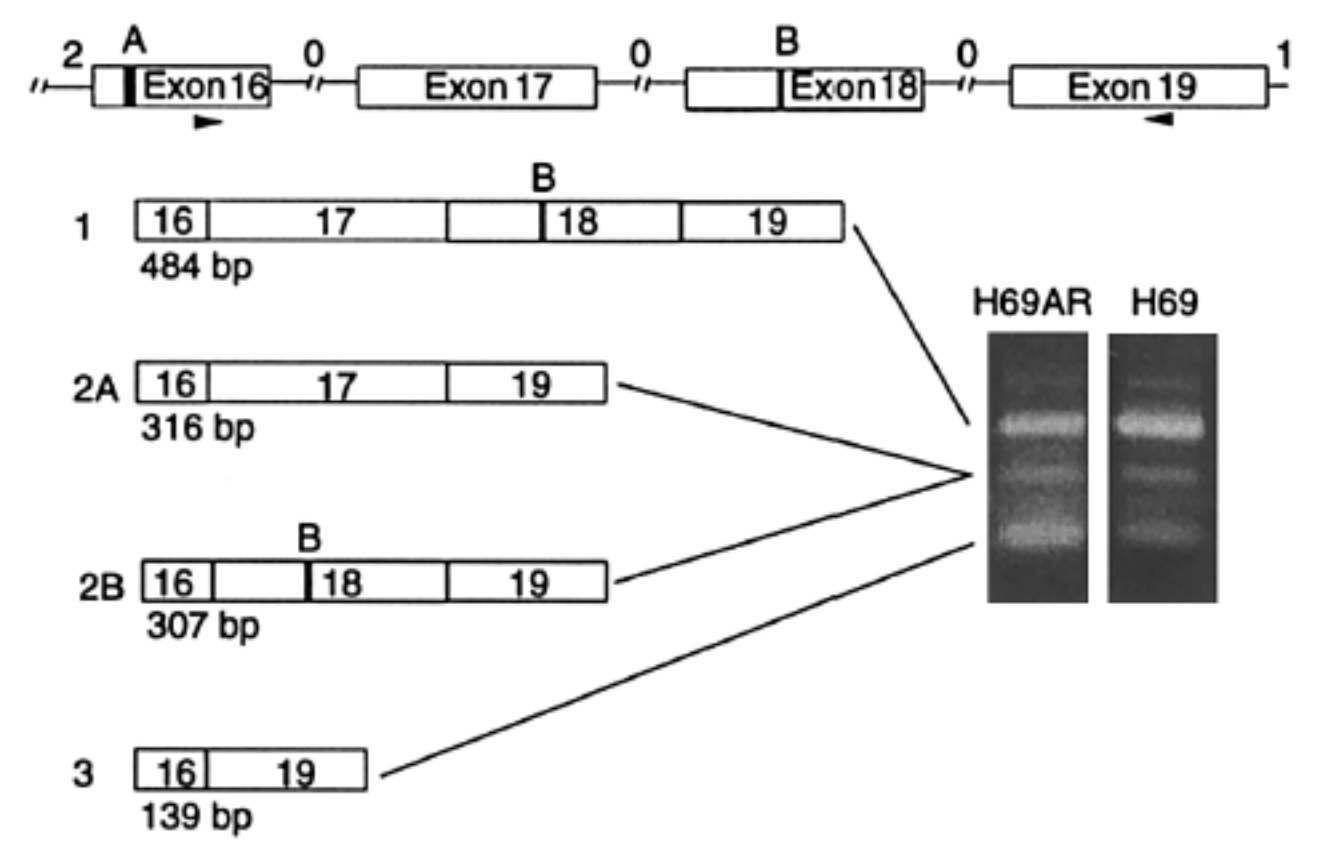
COMPLEMENTARY STRAND
 Number of found binding sites = 91
 Number of found potential transcription starts = 1

              Start        Interval
    N           of                          Confidential              Most probable
           transcription                       level      Weight    Bucher classes

    1)          501        501    504 (C)    vague      1.09       6.1.6
```



Bioinformatics



40,000 x 35%
x 5~10 splicing=
70,000 ~ 140,000

+

40,000 x 65% =
26,000

96,000 ~ 166,000

http://binfo.ym.edu.tw/passdb/index.html

Putative ALternative Splicing Database (PALS db)

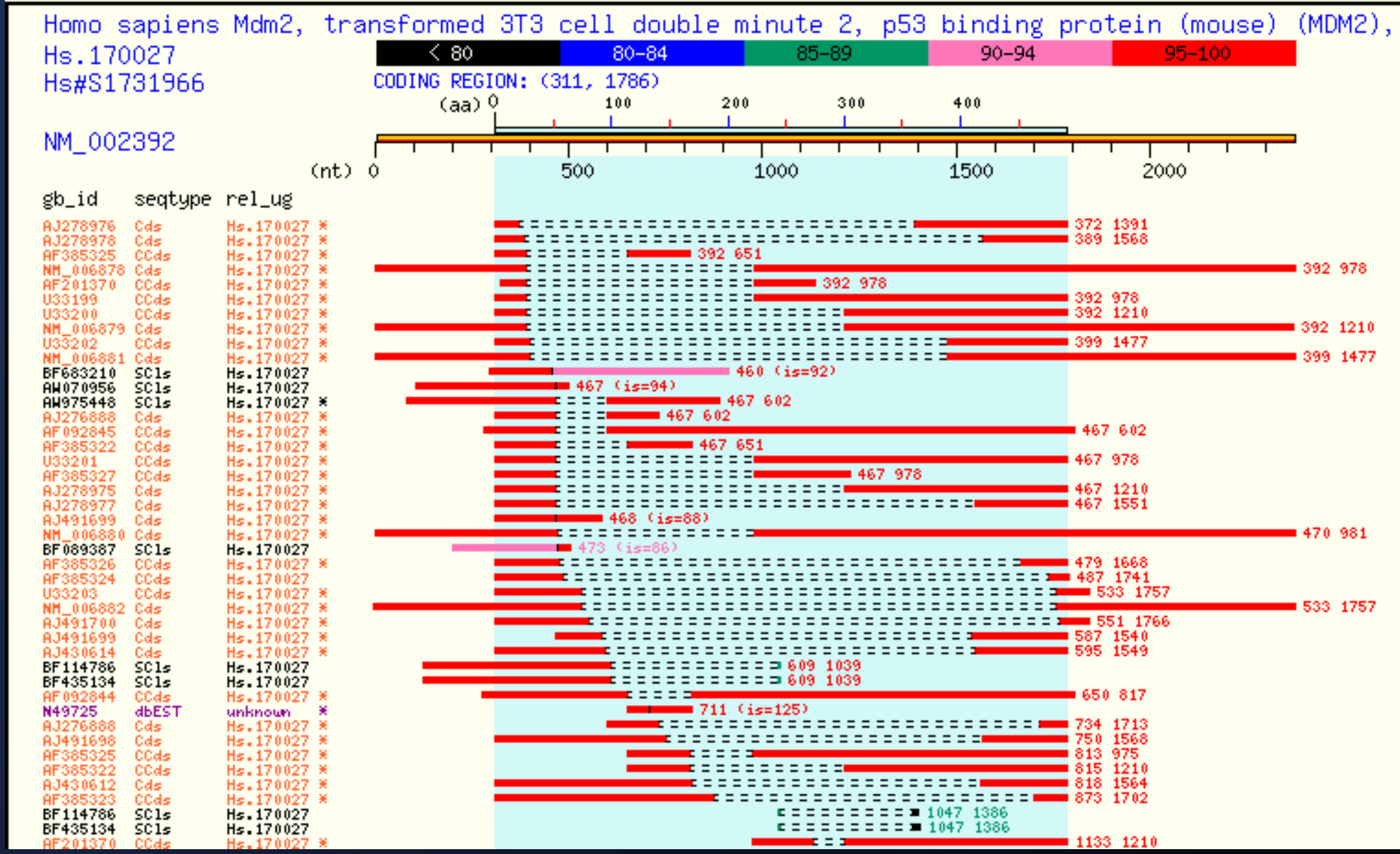
[Introduction] [Rationale] [History] [Statistics] [User's Guide] [Chinese Version] [Mirror Sites] [Supplementary Material]

Introduction of PALS db

Putative **AL**ternative **S**plicing **D**ata**b**ase (PALS db) is a database designed to provide biologists a wide coverage of putative alternative splicing (AS) sites to all known human genes. We used the unique sequences of UniGene clusters as reference sequences to construct PALS db.

To collect the human and mouse EST sequences that are absent in the UniGene database, the reference sequence was used as a query sequence for blast analysis against dbEST. The splicing sites can then be discovered by comparing reference sequence with other sequences in the same UniGene cluster or with related EST sequences from dbEST.

The work of us differs from other available databases in that we tried to cover all known human genes and all available sources of sequences possibly containing AS information. Though genomic sequences are the absolute reference sequences to find AS, half of human genome draft is still unordered and fragmented, which makes alignment more complicated.


Homo sapiens Mdm2, transformed 3T3 cell double minute 2, p53 binding protein (mouse) (MDM2),
Hs.170027
Hs#S1731966
NM_002392
< 80
80-84
85-89
90-94
95-100
CODING REGION: (311, 1786)
(aa) 0 100 200 300 400
(nt) 0 500 1000 1500 2000
gb_id seqtype rel_ug
AJ278976 Cds Hs.170027 * 372 1391
AJ278978 Cds Hs.170027 * 389 1568
AF385325 CCds Hs.170027 * 392 651
NM_006878 Cds Hs.170027 * 392 978
AF201370 CCds Hs.170027 * 392 978
U33199 CCds Hs.170027 * 392 978
U33200 CCds Hs.170027 * 392 1210
NM_006879 Cds Hs.170027 * 392 1210
U33202 CCds Hs.170027 * 399 1477
NM_006881 Cds Hs.170027 * 399 1477
BF683210 SCls Hs.170027 460 (is=92)
AW070956 SCls Hs.170027 467 (is=94)
AW975448 SCls Hs.170027 * 467 602
AJ276888 Cds Hs.170027 * 467 602
AF092845 CCds Hs.170027 * 467 602
AF385322 CCds Hs.170027 * 467 651
U33201 CCds Hs.170027 * 467 978
AF385327 CCds Hs.170027 * 467 978
AJ278975 Cds Hs.170027 * 467 1210
AJ278977 Cds Hs.170027 * 467 1551
AJ491699 Cds Hs.170027 * 468 (is=88)
NM_006880 Cds Hs.170027 * 470 981
BF089387 SCls Hs.170027 473 (is=86)
AF385326 CCds Hs.170027 * 479 1668
AF385324 CCds Hs.170027 487 1741
U33203 CCds Hs.170027 * 533 1757
NM_006882 Cds Hs.170027 * 533 1757
AJ491700 Cds Hs.170027 * 551 1766
AJ491699 Cds Hs.170027 * 587 1540
AJ430614 Cds Hs.170027 * 595 1549
BF114786 SCls Hs.170027 609 1039
BF435134 SCls Hs.170027 609 1039
AF092844 CCds Hs.170027 * 650 817
N49725 dbEST unknown * 711 (is=125)
AJ276888 Cds Hs.170027 * 734 1713
AJ491698 Cds Hs.170027 * 750 1568
AF385325 CCds Hs.170027 * 813 975
AF385322 CCds Hs.170027 * 815 1210
AJ430612 Cds Hs.170027 * 818 1564
AF385323 CCds Hs.170027 * 873 1702
BF114786 SCls Hs.170027 1047 1386
BF435134 SCls Hs.170027 1047 1386
AF201370 CCds Hs.170027 * 1133 1210

http://cgsigma.cshl.org/new_alt_exon_db2/

Alternative Splicing Database

This project is supported by the Human Frontier Science Program. It is also partly supported by a Grant-in-Aid *"Genome Science"* for Scientific Research on Priority Areas from The Ministry of Education, Science, Sports, and Culture in Japan

You are the **2850-th** visitor of this page

- About This Project
- Members of This Project
- Data Entry
- Search Data
- Define Groups
- Analyze Data(Java)
- Predict Exons
- Predict enhancer/silencer: by consensus
- GibbsDNA(multisequence alignment)
- Blast search
- Links to Other Sites



ASAP Search Page

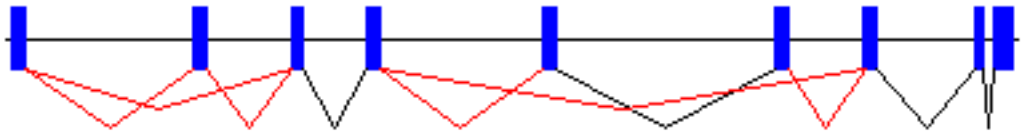
Search by Gene Symbol

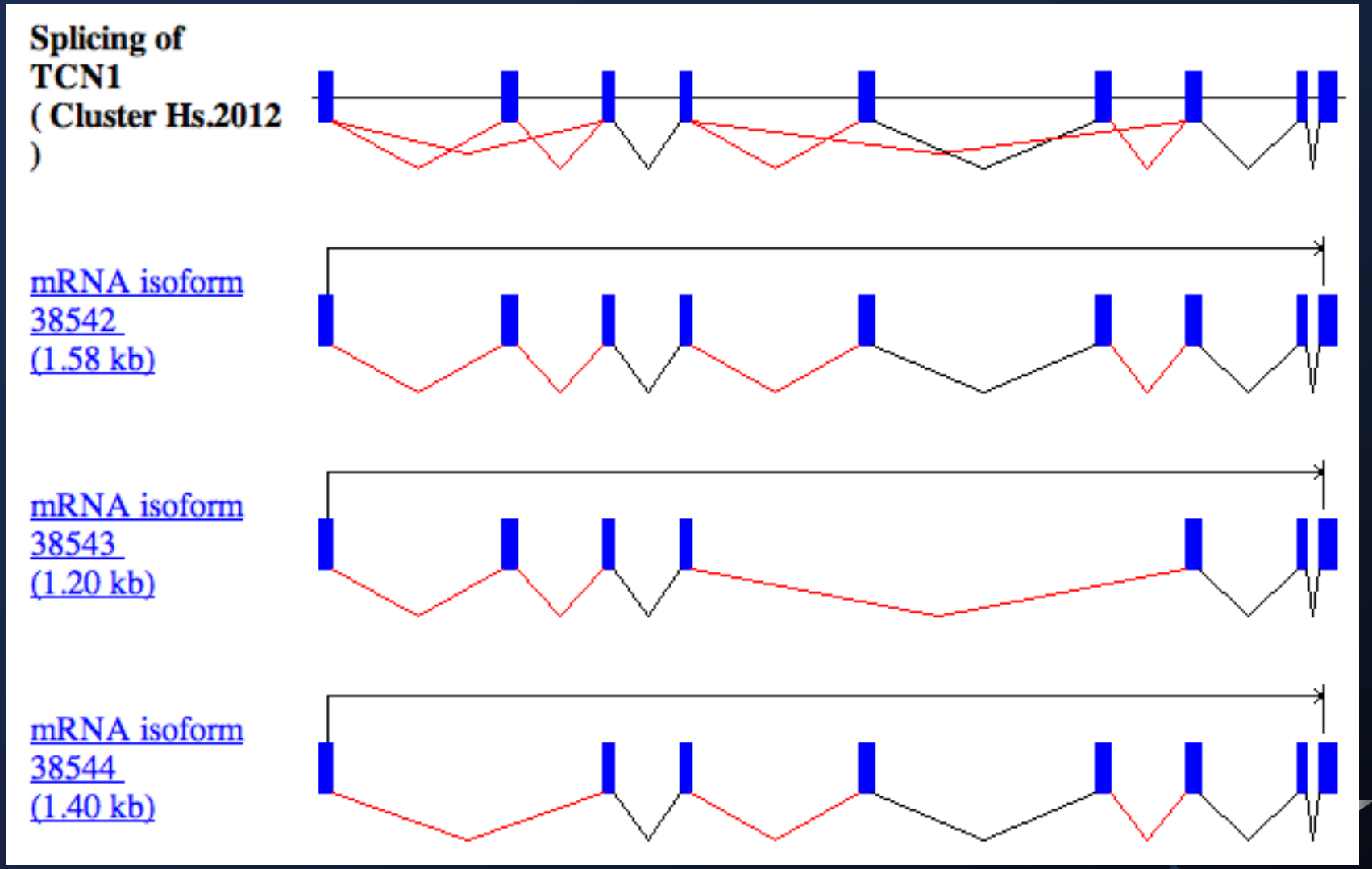
For TCN1 Search

You may search by any of the following items:

- **Gene Symbol:** search by a gene symbol (e.g. TCN1)
- **UniGene Sequence Identifier:** search by a UniGene sequence identifer (e.g. Hs#S3362)
- **UniGene Cluster Identifier:** search by a UniGene cluster identifier (e.g. Hs.2012)
- **Gene Title:** search by a gene title (e.g. transcobalamin I (vitamin B12 binding protein, R binder family))
- **GeneBank Sequence Identifier:** search by a GeneBank sequence identifier (e.g. J05068)

http://www.bioinformatics.ucla.edu/HASDB/

UniGene Cluster Id	Gene	Splicing Overview
Hs.2012 (Graphical View) Table View	transcobalamin I (vitamin B12 binding protein, R binder family) (TCN1)	


Splicing of
TCN1
(Cluster Hs.2012
)
mRNA isoform
38542
(1.58 kb)
mRNA isoform
38543
(1.20 kb)
mRNA isoform
38544
(1.40 kb)

Gene identification involves 4 main stages

Find the putative coding region(s) in the sequence

Open reading frame

Find non-coding features of interest in the sequence

CpG islands
Tandemly and dispersed repeats
Promoter regions (TATA box, cap signal, CCAAT-box)
Transcription factors, Poly-A sites

Determine the exon-intron organization

Branch point signal
CT(G,A)A(C,T)

5' and 3' splice sites:
AG/GUAAGU----------PyPyPyPyPyPyPyPy-CAG/G

Identify the gene

motif, signal and pattern
Blast, FASTA
Functional studies

Ab Initio Gene Discovery

- **Protein coding sequences within a whole genome sequence can be identified using a process known as ab initio gene discovery, in which software that recognizes features common to protein coding transcripts is used.**
  - The existence of long open reading frames
  - Particularly ones for which the codon bias is typical of that observed for the species being studied
  - Proximity of transcriptional and Translational initiation motifs and 3’ polyadenylation site
  - Splicing consensus sequences at putative intron-exon boundaries

Hidden Markov Models

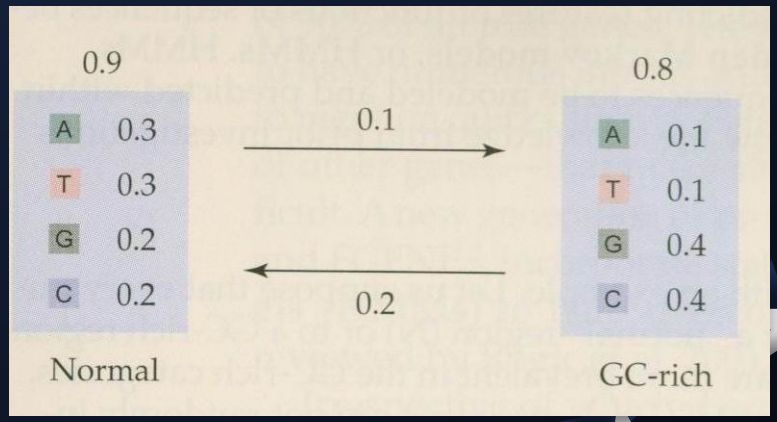
- **An example of the HMM**

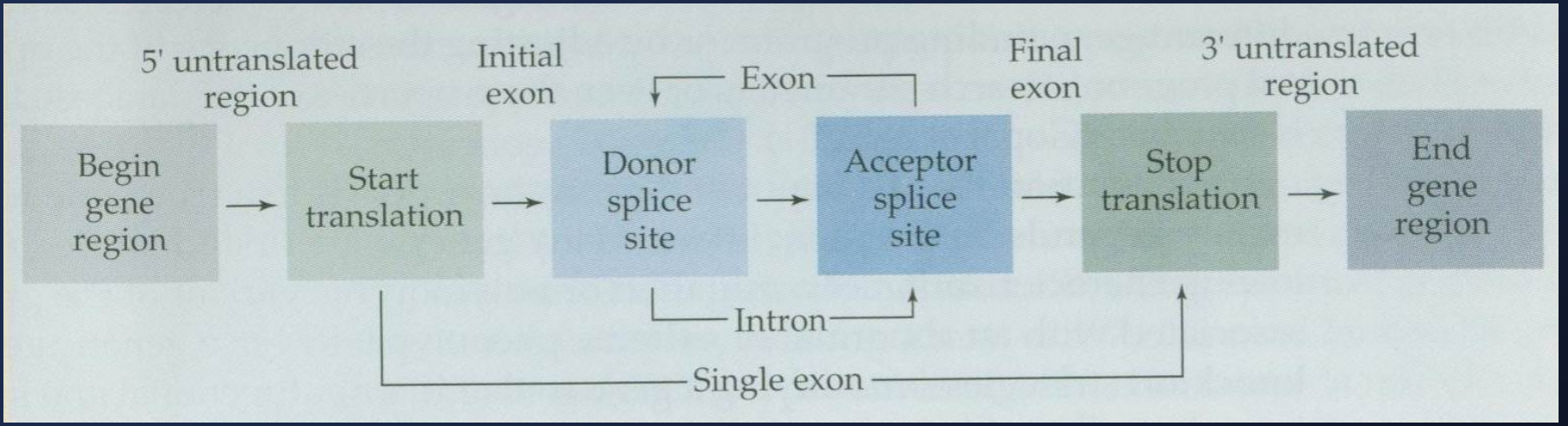
Normal region GC-rich region

NNNNNNNNNRRRRRNNNNNNNNNNNNNNNNNNRRRRRRRRNNN

TTACTTGACGCCAGAAATCTATATTTGGTAACCCGACGCTAA

	A	T	G	C	Mean length
N	0.3	0.3	0.2	0.2	10
R	0.1	0.1	0.4	0.4	5




5' untranslated region
Initial exon
Exon
Final exon
3' untranslated region
Begin gene region
Start translation
Donor splice site
Acceptor splice site
Stop translation
End gene region
Intron
Single exon

The New GENSCAN Web Server at MIT

Identification of complete gene structures in genomic DNA



For information about Genscan, click here

This server provides access to the program Genscan for predicting the locations and exon-intron structures of genes in genomic sequences from a variety of organisms.

News:

http://genes.mit.edu/GENSCAN.html

- **The server is back online. Apologies for the delay.**

Organism: Vertebrate Suboptimal exon cutoff (optional): 1.00

Sequence name (optional):

Print options: Predicted peptides only

Upload your DNA sequence file (one-letter code, upper or lower case, spaces/numbers ignored):

Browse...

Or paste your DNA sequence here (one-letter code, upper or lower case, spaces/numbers ignored):

```
ACAGCCCCTGAAGTGTCTGAGGAGAGCCACAAACCAACAGAGGATTCAGCC
AATGTGTATGAACAGGATG
ATCTGAGTGAACAAATGGCAAGTTTGGAAGGACTCATGAAGCAGCTTAATGC
CATCACAGGCTCAGCCTT
TTAACATGTATTTCTGAATGGATGAGGTGAATTTTCCGGGAACTTTGCAGCATA
CCAATTACCCATAAAC
AGCACACCTGTGTCCAAGAACTCTAACCAGTGTACAGGTCACCCATCAGGA
CCACTCAGTTAAGGAAGAT
CCTGAAGCAGTTCAGAAGGAATAAGCATTCCTTCTTTCACAGGCATCAGGAA
TTGTCAAATGATGATTAT
GAGTTCCCTAAACAAAAGCAAAGATGCATTTTCACTGCAATGTCAAAGTTTAG
CTGCT
```

GENSCAN 1.0 Date run: 31-Oct-102 Time: 15:53:11

Sequence 15:53:10 : 4608 bp : 48.55% C+G : Isochore 2 (43 - 51 C+G%)

Parameter matrix: HumanIso.smat

Predicted genes/exons:

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.01	Sngl	+	1	4344	4344	0	0	75	38	2516	0.868	230.84
1.02	PlyA	+	4561	4566	6							-3.64

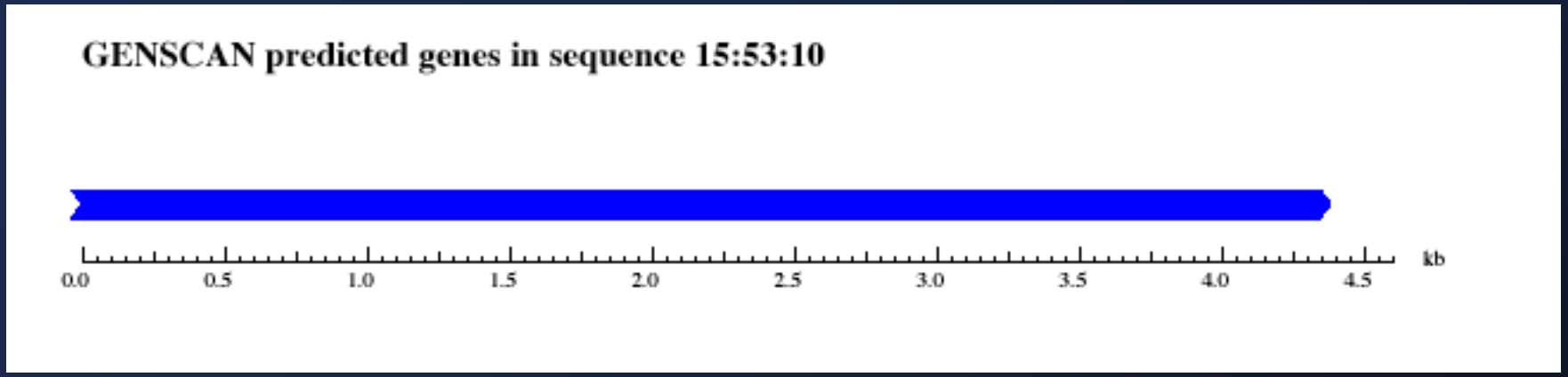
Click here to view a PDF image of the predicted gene(s)

Click here for a PostScript image of the predicted gene(s)

Predicted peptide sequence(s):

>15:53:10|GENSCAN_predicted_peptide_1|1447_aa
MENSLRCVWVPKLAFVLFGASLLSAHLQVTGFQIKAFTALRFLSEPSDAVTMRGGNVLLD
CSAESDRGVPVIKWKKDGIHLALGMDERKQQLSNGSLLIQNILHSRHHKPDEGLYQCEAS
LGDSGSIISRTAKVAVAGPLRFLSQTESVTAFMGDTVLLKCEVIGEPMPTIHWQKNQQDL
TPIPGDSRVVVLPSGALQISRLQPGDIGIYRCSARNPASSRTGNEAEVRILSDPGLHRQL
YFLQRPSNVVAIEGKDAVLECCVSGYPPPSFTWLRGEEVIQLRSKKYSLLGGSNLLISNV
TDDDSGMYTCVVTYKNENISASAELTVLVPPWFLNHPSNLYAYESMDIEFECTVSGKPVP
TVNWMKNGDVVIPSDYFQIVGGSNLRILGVVKSDEGFYQCVAENEAGNAQTSAQLIVPKP
AIPSSSVLPSAPRDVVPVLVSSRFVRLSWRPPAEAKGNIQTFTVFFSREGDNRERALNTT
QPGSLQLTVGNLKPEAMYTFRVVAYNEWGPGESSQPIKVATQPELQVPGPVENLQAVSTS
PTSILITWEPPAYANGPVQGYRLFCTEVSTGKEQNIEVDGLSYKLEGLKKFTEYSLRFLA
YNRYGPGVSTDDITVVTLSDVPSAPPQNVSLEVVNSRSIKVSWLPPPSGTQNGFITGYKI




GENSCAN predicted genes in sequence 15:53:10
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
kb

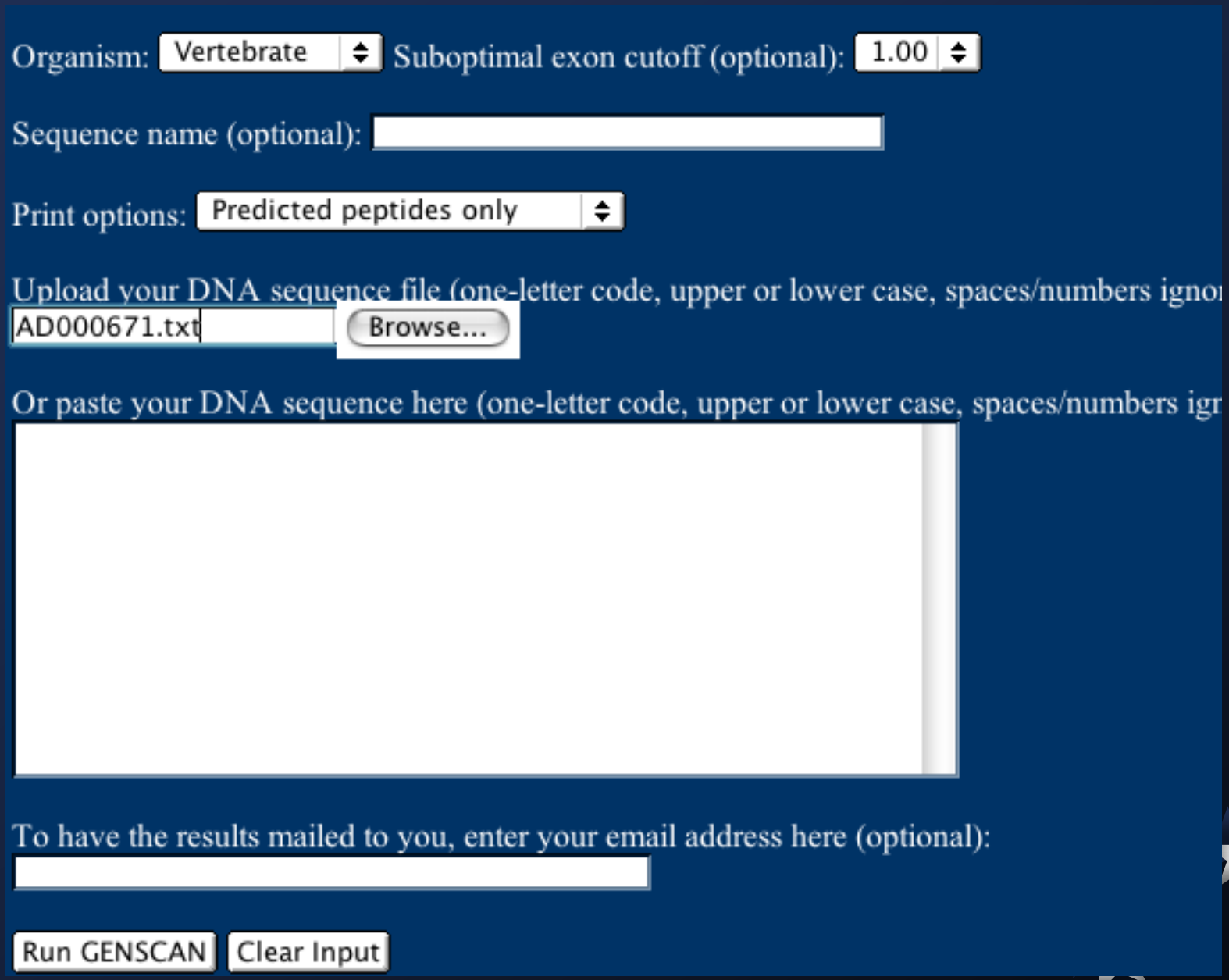
1: AD000671

Homo sapiens DNA from chromosome 19-cosmid f24109 containing HRX2, genomic sequence
gi|1905893|gb|AD000671.1|AD000671[1905893]

Display default Save Text Add to Clipboard Get Subsequence

1: AD000671. Homo sapiens DNA ...[gi:1905893]

```
LOCUS       AD000671               46251 bp    DNA     linear   PRI 10-DEC-1997
DEFINITION  Homo sapiens DNA from chromosome 19-cosmid f24109 containing HRX2,
            genomic sequence.
ACCESSION   AD000671
VERSION     AD000671.1  GI:1905893
KEYWORDS    chromosome 19.
SOURCE      Homo sapiens
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 46251)
  AUTHORS   Lamerdin,J.E.
  TITLE     Direct Submission
  JOURNAL   Submitted (27-NOV-1996) J.E. Lamerdin, Human Genome Center,
            Lawrence Livermore National Laboratory, 7000 East Ave, Livermore,
            CA, USA, 94551 jane@acgt.llnl.gov ow@tornak.llnl.gov
REFERENCE   2  (bases 1 to 46251)
  AUTHORS   Lamerdin,J.E., McCready,P.M., Adamson,A.W., Burkhart-Schultz,K.,
            Garcia,E., Kyle,A., Ramirez,M., Stilwagen,S., Garnes,J.,
            Danganan,L., Bruce,R., Quan,G., Montgomery,M., Ow,D., Kobayashi,A.,
            Olsen,A.O. and Carrano,A.V.
  TITLE     Direct Submission
  JOURNAL   Submitted (27-NOV-1996) Sequence analysis of a 1 Mb region in human
            19q13.1
```


Organism: Vertebrate
Suboptimal exon cutoff (optional): 1.00
Sequence name (optional):
Print options: Predicted peptides only
Upload your DNA sequence file (one-letter code, upper or lower case, spaces/numbers ignor
AD000671.txt
Browse...
Or paste your DNA sequence here (one-letter code, upper or lower case, spaces/numbers igr
To have the results mailed to you, enter your email address here (optional):
Run GENSCAN
Clear Input

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.00	Prom	+	4679	4718	40							0.09
1.01	Init	+	10138	11018	881	0	2	81	105	377	0.994	32.79
1.02	Intr	+	11268	11341	74	0	2	70	115	-4	0.996	0.04
1.03	Intr	+	11450	11518	69	0	0	115	109	0	0.954	4.45
1.04	Intr	+	11644	11808	165	2	0	147	80	32	0.994	8.75
1.05	Intr	+	11989	12144	156	2	0	74	-15	178	0.050	6.59
1.06	Intr	+	12360	12454	95	1	2	25	42	87	0.040	-2.02
1.07	Intr	+	12608	12748	141	1	0	77	44	155	0.047	11.06
1.08	Intr	+	13530	13892	363	2	0	37	78	138	0.030	3.64
1.09	Intr	+	14980	15052	73	0	1	57	105	54	0.665	3.47
1.10	Intr	+	15295	17315	2021	2	2	111	117	903	0.898	83.77
1.11	Intr	+	17870	17983	114	1	0	97	10	142	0.860	8.45
1.12	Intr	+	18079	18229	151	0	1	100	73	159	0.998	15.85
1.13	Intr	+	18506	18785	280	0	1	61	56	221	0.998	13.28
1.14	Intr	+	18958	19013	56	1	2	108	61	89	0.928	7.31
1.15	Intr	+	19314	19517	204	1	0	83	54	266	0.683	22.40
1.16	Intr	+	20147	20241	95	0	2	75	82	112	0.951	9.48
1.17	Intr	+	20499	20597	99	2	0	125	86	48	0.998	9.11
1.18	Intr	+	20730	20843	114	2	0	135	50	63	0.855	8.35
1.19	Intr	+	20989	21159	171	0	0	94	59	122	0.460	10.55
1.20	Intr	+	21233	21328	96	1	0	52	72	77	0.892	3.21
1.21	Intr	+	21704	21863	160	1	1	60	81	38	0.738	0.37
1.22	Intr	+	22666	22779	114	2	0	135	109	168	0.807	24.52
1.23	Intr	+	22948	23132	185	2	2	66	41	233	0.687	16.43
1.24	Intr	+	23208	23297	90	2	0	101	89	129	0.995	14.99
1.25	Intr	+	23391	23495	105	2	0	72	105	137	0.988	14.71
1.26	Intr	+	23608	23681	74	0	2	65	61	127	0.995	6.40
1.27	Intr	+	24284	24399	116	2	2	81	84	61	0.988	5.59
1.28	Intr	+	24495	24586	92	1	2	84	100	57	0.999	6.71
1.29	Intr	+	24669	24806	138	2	0	110	77	218	0.998	24.07
1.30	Intr	+	25477	25635	159	0	0	97	86	180	0.898	19.40
1.31	Intr	+	25852	25972	121	0	1	106	99	252	0.999	28.67



>gi|GENSCAN_predicted_peptide_1|3182_aa
MSLPPIRLPSPYGSDRLVQLAARLRPALCDTLITVGSQEFPAHSLVLAGVSQQLGRRGQW
ALGEGISPSTFAQLLNFVYGESVELQPGELRPLQEAARALGVQSLEEACWRARGDRAKKP
DPGLKKHQEEPEKPSRNPERELGDPGEKQKPEQVSRTGGREQEMLHKHSPPRGRPEMAGA
TQEAQQEQTRSKEKRLQAPVGQRGADGKHGVLTWLRENPGGSEESLRKLPGPLPPAGSLQ
TSVTPRPSWAEAPWLVGGQPALWSILLMPPRYGIPFYHSTPTTGAWQEVWREQRIPLSLN
APKGLWSQNQLASSSPTPGSLPQGPAQLSPGEMEESDQGHTGALATCAGHEDKAGCPPRP
HPPPAPPARSRPYACSVCGKRFSLKHQMETHYRVHTGEKPFSCSLCPQRSRDFSAMTKHL
RTHGAAPYRCSLCGAGCPSLASMQAHMRANSAPEAPQEGRRVPSPGRSRVAEAQASEPYR
RPLVALRPLSSGMDTSSTPEAPRPPQEAAGPEDSVSGHTLRHAGPGLMAAAAGGGSCPGP
GSARGRFPGRPRGAGGGGGRGGRGNGAERVRVALRRGGGATGPGGAEPGEDTALLRLLGL
RRGLRRLRRLWAGPRVQRGRGRGRGRGWGPSRGCVPEEESSDGESDEEEFQGFHSDEDVA
PSSLRSALRSQRGRAPRGRGRKHKTTPLPPPRLADVAPTPPKTPARKRGEEGTERMVQAL
TELLRRAQAPQAPRSRACEPSTPRRSRGRPPGRPAGPCRRKQQAVVVAEAAVTIPKPEPP
PPVVPVKHQTGSWKCKEGPGPGPGTPRRGGQSSRGGRGGRGRGRGGGLPFVIKFVSRAKK
VKMGQLSLGLESGQGQGQHEESWQDVPQRRVGSGQGGSPCWKKQEQKLDDEEEEKKEEEE
KDKEGEEKEERAVAEEMMPAAEKEEAKLPPPPLTPPAPSPPPPLPPPSTSPPPPLCPPPP
PPVSPPPLPSPPPPPAQEEQEESPPPVVPATCSRKRGRPPLTPSQRAEREAARAGPEGTS
PPTPTPSTATGGPPEDSPTVAPKSTTFLKNIRQFIMPVVSARSSRVIKTPRRFMDEDPPK
PPKVEVSPVLRPPITTSPPVPQEPAPVPSPPRAPTPPSTPVPLPEKRRSILREPTFRWTS
LTRELPPPPPAPPPPPAPSPPPAPATSSRRPLLLRAPQFTPSEAHLKIYESVLTPPPLGA
PEAPEPEPPPADDSPAEPEPRAVGRTNHLSLPRFAPVVTTPVKAEVSPHGAPALSNGPQT
QAQLLQPLQALQTQLLPQALPPPQPQLQPPPSPQQMPPLEKARIAGVGSLPLSGVEEKMF
SLLKRAKVQLFKIDQQQQQKVAASMPLSPGGQMEEVAGAVKQISDRGPVRSEDESVEAKR
ERPSGPESPVQGPRIKHVCRHAAVALGQARAMVPEDVPRLSALPLRDRQDLATEDTSSAS
ETESVPSRSRRGKVEAAGPGGESEPTGSGGTLAHTPRRSLPSHHGKKMRMARCGHCRGCL
RVQDCGSCVNCLDKPKFGGPNTKKQCCVYRKCDKIEARKMERLAKKGPRRGAGAGGPREE
VVAHPGPEEQDSLLQRKSARRCVKQRPSYDIFEDSDDSEPGGPPAPRRRTPRENELPLPE
PEEQSRPRKPTLQPVLQLKARRRLDKDALAPGPFASFPNGWTGKQKSPDGVHRVRVDFKE
DCDLENVWLMGGLSVLTSVPGGPPMVCLLCASKGLHELVFCQVCCDPFHPFCLEEAERPL
PQHHDTWCCRRCKFCHVCGRKGRGSKVWAQGLARHLLECERCRHAYHPACLGPSYPTRAT
RKRRHWSSAWFSPNIALLPQICSACVRCKSCGATPGKNWDVEWSGDYSLCPRCTQLYEKG
NYCPICTRCYEDNDYESKMMQCAQCDHWVHAKCEGLSDEDYEILSGLPDSVLYTCGPCAG
AAQPRWREALSGALQGGLRQVLQGLLSSKVVGPLLLCTQCGPDGKQLHPGPCGLQAVSQR
FEDGHYKSVHSFMEDMVGILMRHSEEGETPDRRAGGQMKGLLLKLLESAFGWFDAHDPKY

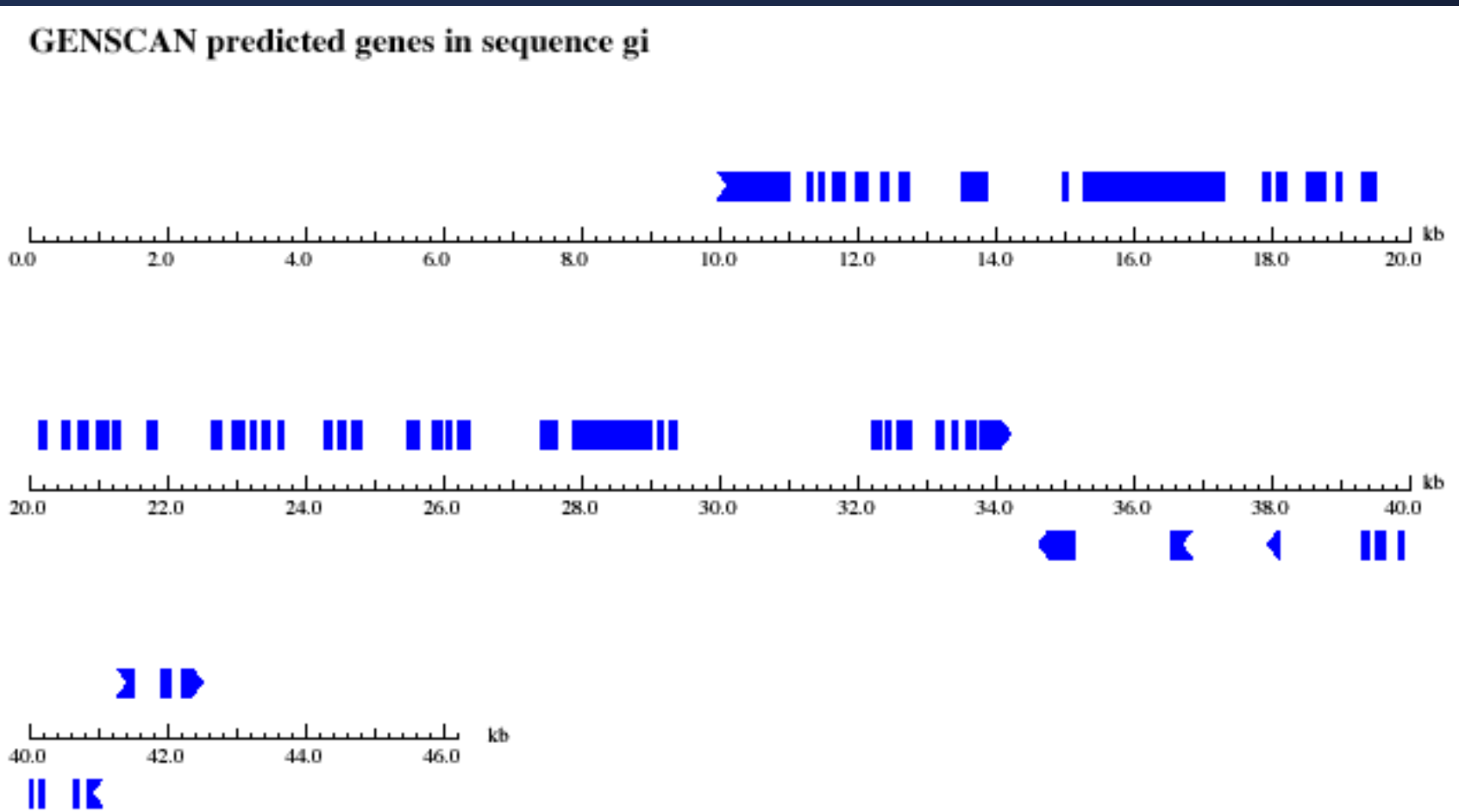
2.03 PlyA -	34784	34779	6							1.05
2.02 Term -	35136	34790	347	1	2	114	41	92	0.784	2.11
2.01 Init -	36690	36534	157	0	1	68	80	261	0.984	21.34
2.00 Prom -	37375	37336	40							-11.21
3.09 PlyA -	38067	38062	6							1.05
3.08 Term -	38129	38108	22	1	1	116	33	4	0.298	-4.23
3.07 Intr -	39414	39319	96	2	0	91	82	52	0.535	4.52
3.06 Intr -	39655	39522	134	1	2	98	39	167	0.690	12.75
3.05 Intr -	39931	39833	99	1	0	108	44	185	0.999	16.91
3.04 Intr -	40050	40001	50	1	2	103	78	70	0.934	6.39
3.03 Intr -	40202	40136	67	2	1	73	74	62	0.978	2.27
3.02 Intr -	40722	40635	88	2	1	119	94	63	0.853	10.47
3.01 Init -	40901	40858	44	2	2	83	52	21	0.853	-2.17
3.00 Prom -	41330	41291	40							-4.61
4.00 Prom +	41369	41408	40							-12.18
4.01 Init +	41441	41501	61	1	1	72	82	130	0.999	12.16
4.02 Intr +	41929	42033	105	2	0	111	68	48	0.973	5.79
4.03 Term +	42218	42357	140	0	2	69	48	145	0.988	7.03
4.04 PlyA +	42493	42498	6							1.05



```
>gi|GENSCAN_predicted_peptide_2|167_aa
MGPGRCLLTALLLLALAPPPEASQYCGRLEYWNPDNKCCSSCLQRFGPPPCPELSSLASQ
PLSRLLDELEVLEELIVLLDPEPGPGGGMAHGTTRHLAARYGLPAAWSTFAYSLRPSRSP
LRALIEMVVAREPSASLGQLGTHLAQLGRADALRVLSKLGSSGVCWA

>gi|GENSCAN_predicted_peptide_3|199_aa
MAEYLASIFGTEKDKVNCSFYFKIGVCRHGDRCSRLHNKPTFSQTIVLLNLYRNPQNTAQ
TADGSHCHVSDVEVQEHYDSFFEEVFTELQEKYGEIEEMNVCDNLGDHLVGNVYVKFRRE
EDGERAVAELSNRWFNGQAVHGELSPVTDFRESCCRQYEMGECTRGGFCNFMHLRPISQN
LQRQLYGRGPRRRLPGSIM

>gi|GENSCAN_predicted_peptide_4|101_aa
MNLERVSNEEKLNLCRKYYLGGFAFLPFLWLVNIFWFFREAFLVPAYTEQSQIKGYVWRS
AVGFLFWVIVLTSWITIFQIYRPRWGALGDYLSFTIPLGTP
```


GENSCAN predicted genes in sequence gi
0.0
2.0
4.0
6.0
8.0
10.0
12.0
14.0
16.0
18.0
20.0
kb
20.0
22.0
24.0
26.0
28.0
30.0
32.0
34.0
36.0
38.0
40.0
kb
40.0
42.0
44.0
46.0
kb

http://compbio.ornl.gov/grailexp/

GrailEXP

Grail Experimental Gene Discovery Suite

Home

FAQ

News

Download

References

Credits/ Acknowledgments

Bugs/Questions/ Comments

allgenes.org

GrailEXP is a software package that predicts exons, genes, promoters, polyas, CpG islands, EST similarities, and repetitive elements within DNA sequence. GrailEXP is used by the Computational Biosciences Section at Oak Ridge National Laboratory to annotate the entire known portion of the human genome (including both finished and draft data).

If you are interested in microbial genome analysis and annotation, you should g to the Generation home page.

Perform Analysis

Select organism: Human (Homo sapiens)

Select output type: Human-Readable Text

GRAILEXP

Grail Experimental Gene Discovery S

Organism: human
Sequence: >gi|1905893|gb|AD000671.1|AD000671 Homo sapiens DNA from chromosome
containing HRX2, genomic sequence
Length: 46251 bases
Estimated Time to Analyze: 19.2502 seconds

Your analysis has been dispatched to the server. The request ID is **1036098954-21618**.

Check results



PERCEVAL Exon Candidates (44 predicted)

Index	Std	Begin	End	Frm	Type	Len	Scr	Quality
1	+	10138	11022	0	Single	885	81	Good
2	+	12608	12711	0	Internal	104	68	Marginal
3	+	13530	13923	0	Initial	394	69	Marginal
4	+	15698	15771	1	Terminal	74	53	Marginal
5	+	16358	16532	2	Internal	175	66	Marginal
6	+	17163	17315	0	Initial	153	88	Good
7	+	17843	17983	0	Internal	141	81	Good
8	+	18079	18229	0	Internal	151	99	Excellent
9	+	18664	18785	0	Initial	122	52	Marginal
10	+	18958	19013	2	Internal	56	92	Excellent
11	+	19371	19517	1	Internal	147	96	Excellent
12	+	19579	19626	0	Internal	48	72	Good
13	+	20147	20292	1	Internal	146	74	Good
14	+	20600	20672	0	Initial	73	42	Marginal
15	+	20989	21177	0	Internal	189	75	Good
16	+	22666	22817	1	Internal	152	95	Excellent
17	+	22948	23132	1	Internal	185	77	Good
18	+	23208	23297	0	Internal	90	100	Excellent
19	+	23391	23495	0	Internal	105	99	Excellent
20	+	23608	23681	0	Internal	74	96	Excellent
21	+	24495	24586	1	Internal	92	100	Excellent
22	+	24669	24806	0	Internal	138	100	Excellent
23	+	25477	25635	0	Internal	159	93	Excellent
24	+	25852	25972	0	Internal	121	100	Excellent
25	+	26048	26130	1	Terminal	83	100	Excellent
26	+	27997	28281	0	Initial	285	79	Good
27	+	28769	28899	1	Internal	131	76	Good
28	+	29259	29382	0	Internal	124	75	Good
29	+	29907	30117	2	Internal	211	66	Marginal
30	+	32200	32337	1	Internal	138	65	Marginal
31	+	32596	32773	1	Internal	178	86	Good
32	+	33146	33262	2	Internal	117	100	Excellent



Gene Finder

This page contains software tools designed to predict putative internal protein coding exons in genomic DNA sequences. Human, mouse and arabidopsis exons are predicted by a program called MZEF (developed by Dr. Michael Zhang). fission yeast exons are predicted by a program called Pombe (developed by Dr. Tim Chen and Dr. Michael Zhang).

- Human
- Mouse
- Arabidopsis
- Fission Yeast

For more information about MZEF click here
To contact Dr. Zhang mzhang@cshl.org
Zhang's Lab Home Page

Human Internal Coding Exon Finder by Michae Zhang

Strand (1 for Watson, 2 for Crick) 1
Overlap (integer) 0
Prior Probability (Between 0.00 and 1.00) .02

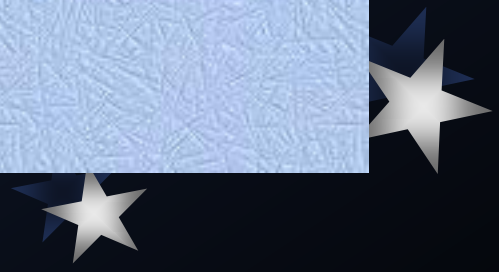
Please paste your sequence in FASTA format (First title line begin with a ">". Max = 200 kb)

Or, give the name of a file in FASTA format containing the sequences here:

AD000671.txt Browse...

Please mail the results to :

Submit Clear



uploaded sequence file: AD000671.txt

Internal coding exons predicted by MZEF
Sequence_length: 46251 G+C_content: 0.568

Coordinates	P	Fr1	Fr2	Fr3	Orf	3ss	Cds	5ss
11464 - 11518	0.666	0.497	0.445	0.360	112	0.522	0.487	0.689
12024 - 12079	0.830	0.379	0.518	0.689	211	0.525	0.585	0.544
13530 - 13892	0.630	0.501	0.481	0.573	221	0.456	0.557	0.648
14980 - 15052	0.786	0.522	0.520	0.356	112	0.515	0.515	0.684
16358 - 16451	0.744	0.434	0.481	0.635	211	0.488	0.569	0.632
17160 - 17315	0.930	0.416	0.500	0.620	211	0.490	0.564	0.702
22666 - 22779	0.994	0.452	0.514	0.625	221	0.546	0.611	0.670
23391 - 23495	0.992	0.384	0.389	0.672	211	0.495	0.569	0.630
24495 - 24586	0.840	0.395	0.571	0.416	111	0.498	0.519	0.634
24669 - 24806	0.976	0.452	0.481	0.619	111	0.543	0.594	0.577
25477 - 25635	0.944	0.632	0.412	0.411	122	0.534	0.549	0.483
25852 - 25972	0.999	0.690	0.395	0.377	122	0.526	0.598	0.613
26048 - 26126	0.788	0.583	0.369	0.325	112	0.552	0.495	0.548
27974 - 28267	0.965	0.613	0.404	0.417	122	0.504	0.548	0.514
28769 - 29018	0.961	0.545	0.414	0.441	122	0.555	0.512	0.651
32200 - 32337	0.908	0.421	0.386	0.592	121	0.519	0.546	0.615
32596 - 32773	0.966	0.434	0.364	0.581	221	0.527	0.525	0.650

HMMgene (v. 1.1)

Prediction of vertebrate and C. elegans genes

One of the World Wide Web Prediction Servers
from the Center for Biological Sequence Analysis

Informatio

This server was set up on Feb. 27, 1998. Please mail krogh@cbs.dtu.dk in case of problems.

Jan. 18, 2002:
Error in the processing of annotation in the sequence file corrected.

Submission of a local file (HTML 3.0 or higher)

Organism:
- (•) Human (and other vertebrates)
- ( ) C. elegans

File in FASTA format

Browse...

Options:
- [ ] Predict signals

best prediction

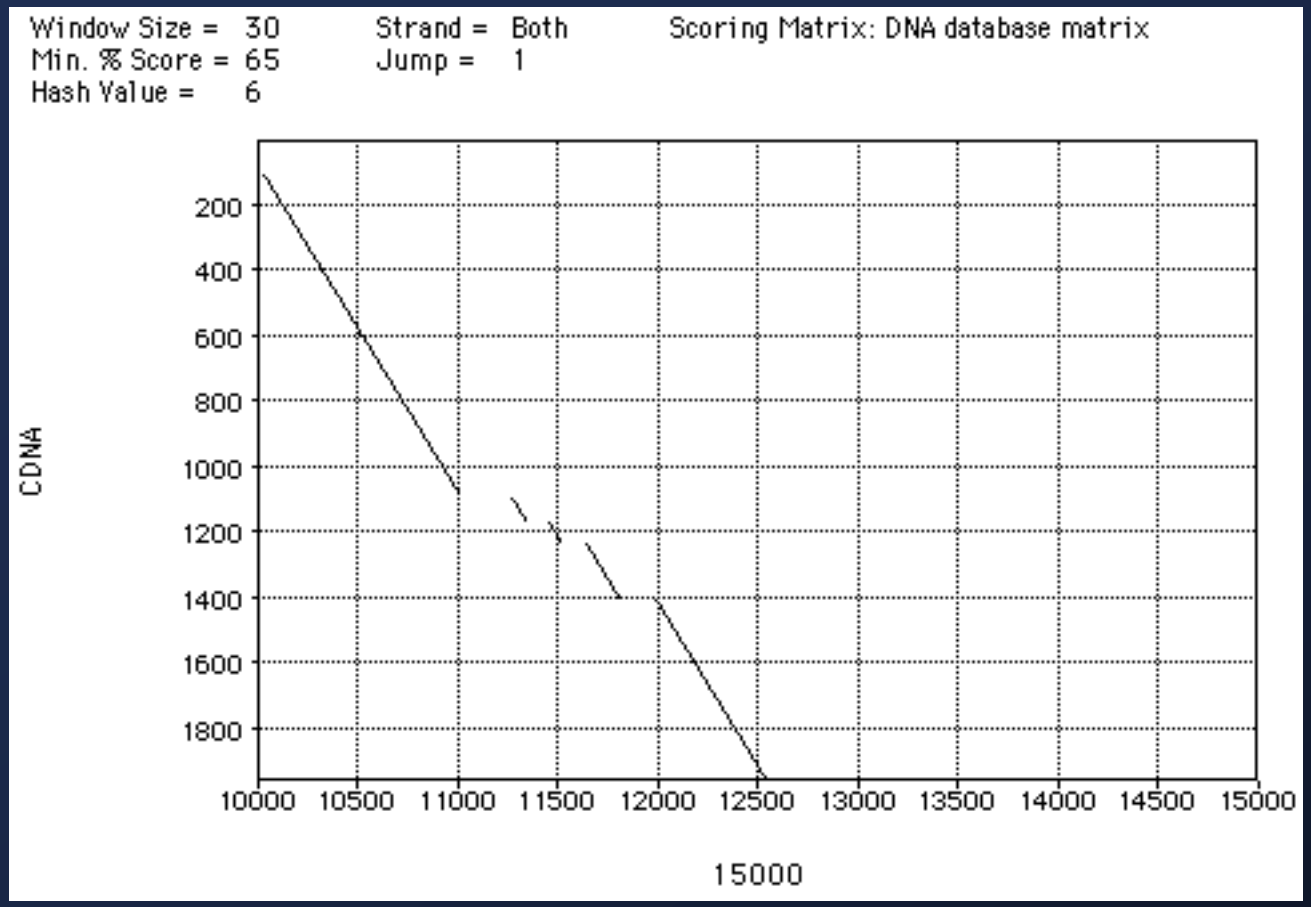
File with annotation (optional)

Browse...

```
## gff-version 1
## date: Thu Oct 31 22:25:46 2002
## HMMgene1.1a (human model sim10gc.C.bsmod)

# SEQ: gi|1905893|gb|AD000671.1|AD000671 46251 (+) A:9981 C:12880 G:13407 T:9983
gi|1905893|gb|AD000671.1|AD000671 HMMgene1.1a firstex 10138 11018 0.969 + 2 bestparse:cds_
gi|1905893|gb|AD000671.1|AD000671 HMMgene1.1a exon_1 11268 11341 0.995 + 1 bestparse:cds_1
gi|1905893|gb|AD000671.1|AD000671 HMMgene1.1a exon_2 11450 11518 0.926 + 1 bestparse:cds_1
gi|1905893|gb|AD000671.1|AD000671 HMMgene1.1a exon_3 11644 11808 0.934 + 1 bestparse:cds_1
gi|1905893|gb|AD000671.1|AD000671 HMMgene1.1a exon_4 15002 15052 0.418 + 1 bestparse:cds_1
gi|1905893|gb|AD000671.1|AD000671 HMMgene1.1a exon_5 15295 15489 0.931 + 1 bestparse:cds_1
gi|1905893|gb|AD000671.1|AD000671 HMMgene1.1a exon_6 16246 16451 0.556 + 0 bestparse:cds_1
gi|1905893|gb|AD000671.1|AD000671 HMMgene1.1a exon_7 16812 17315 0.976 + 0 bestparse:cds_1
gi|1905893|gb|AD000671.1|AD000671 HMMgene1.1a exon_8 17843 17983 0.858 + 0 bestparse:cds_1
gi|1905893|gb|AD000671.1|AD000671 HMMgene1.1a exon_9 18079 18229 0.974 + 1 bestparse:cds_1
gi|1905893|gb|AD000671.1|AD000671 HMMgene1.1a exon_10 19314 19517 0.501 + 1 bestparse:cds_
gi|1905893|gb|AD000671.1|AD000671 HMMgene1.1a exon_11 20147 20241 0.476 + 0 bestparse:cds_
gi|1905893|gb|AD000671.1|AD000671 HMMgene1.1a exon_12 20499 20597 0.999 + 0 bestparse:cds_
gi|1905893|gb|AD000671.1|AD000671 HMMgene1.1a exon_13 20730 20843 0.492 + 0 bestparse:cds_
gi|1905893|gb|AD000671.1|AD000671 HMMgene1.1a exon_14 20989 21177 0.835 + 0 bestparse:cds_
gi|1905893|gb|AD000671.1|AD000671 HMMgene1.1a exon_15 21233 21328 0.603 + 0 bestparse:cds_
gi|1905893|gb|AD000671.1|AD000671 HMMgene1.1a exon_16 21746 21863 0.960 + 1 bestparse:cds_
gi|1905893|gb|AD000671.1|AD000671 HMMgene1.1a exon_17 22666 22779 0.940 + 1 bestparse:cds_
gi|1905893|gb|AD000671.1|AD000671 HMMgene1.1a exon_18 22948 23132 0.444 + 0 bestparse:cds_
gi|1905893|gb|AD000671.1|AD000671 HMMgene1.1a exon_19 23208 23297 0.677 + 0 bestparse:cds_
gi|1905893|gb|AD000671.1|AD000671 HMMgene1.1a exon_20 23391 23495 0.910 + 0 bestparse:cds_
gi|1905893|gb|AD000671.1|AD000671 HMMgene1.1a exon_21 23608 23681 0.881 + 2 bestparse:cds_
```

Predicts	GenScan	GRAIL	MZEF	HMMgen
1	10138-11018	10138-11022	11464-11518	10138-11018
2	11268-11341	12608-12711	12024-12079	11268-11341
3	11450-11518	13530-13923	13530-13892	11450-11518
4	11644-11808	15698-15771	14980-15052	11644-11808
5	11989-12144	16358-16532	16358-16451	15002-15052



GRAIL 2

10138 - 11018	+
12608 - 12748	x
13530 - 13923	x

GENSCAN

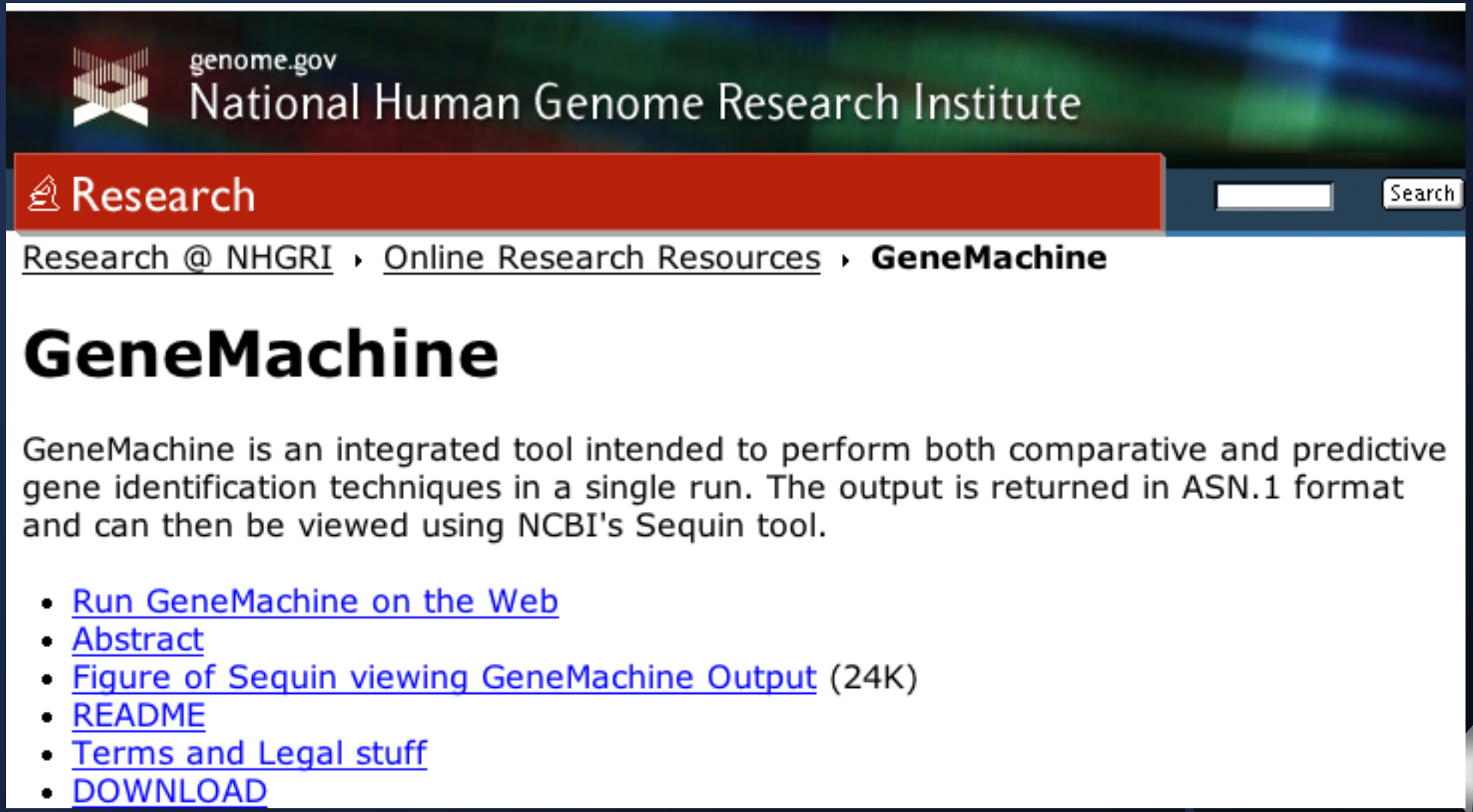
10138 - 11018	+
11268 - 11341	+
11450 - 11518	+
11644 - 11808	+
11989 - 12144	+
12360 - 12454	x
12608 - 12748	x

FGENES

1880 - 1908	x
5061 - 5175	x
5900 - 6049	x
8317 - 8544	+
10357 - 11018	+
11268 - 11341	+
11450 - 11518	+
11644 - 11864	+
polyA: 12521	+

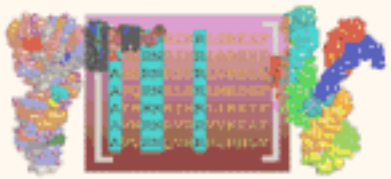
cDNA and genomic DNA alignment and matrix analysis:

http://research.nhgri.nih.gov/genemachine/




File Edit Search Options Misc Annotate
Target Sequence
bigcont
Done
Display Format
Graphic
Style
GeneMachine
Scale
50
00
20000
25000
30000
35000
40000
45000
50000
55000
60000
65000
70000
blastn - est
AI036202
AI036202
AI036202
AU020134
AA119662
AI385991
AA065350
blastn - nr
X57569
L76631
M61099
U31216
AC003997
S48085
S48085
blastx - swissprot
Q13255
P23385
Q14831
P41594
Fgenes
Fgenes
Grail; good
Genscan
exon: MZEF; P=0.657
exon: MZEF; P=0.635
exon: Grail; good
exon: Grail; good
exon: Grail; excellent
exon: Grail; good
exon: MZEF; P=0.990
exon: MZEF; P=0.551
exon: MZEF; P=0.750
exon: MZEF; P=0.938
exon: Grail; excellent
exon: MZEF; P=0.637
exon: Grail; excellent
exon: Grail; excellent
exon: Grail; good
exon: Grail; marginal
exon: Grail; excellent
exon: MZEF; P=0.999
exon: Grail; excellent
exon: Grail; excellent
exon: MZEF; P=0.847
exon: MZEF; P=0.577
exon: Grail; excellent
exon: Grail; excellent

BioEdit



Copyright © 1997-2005
Tom Hall
Ibis Therapeutics
Carlsbad, CA 92008

Biological sequence alignment editor for Win95/98/NT/2K/XP

BioEdit is a biological sequence alignment editor written for Windows 95/98/NT/2000/XP. An intuitive multiple document interface with convenient features makes alignment and manipulation of sequences relatively easy on your desktop computer. Several sequence manipulation and analysis options and links to external anaylsis programs facilitate a working environment which allows you to view and manipulate sequences with simple point-and-click operations.

BioEdit.zip (full installation 12.6 Mb)

Bug fixes / changes

BioEdit General information

BioDoc.pdf (pdf format help doc)

View Screenshots

BioEdit's features include: ***New version is WinXP compatible***

- Several modes of hand alignment
- Automated ClustalW alignment
- Automated Blast searches (local and WWW)
- Plasmid drawing and annotation
- Accessory application configuration
- Restriction mapping
- RNA comparative analysis tools
- Graphical matrix data viewing tools
- Shaded alignment figures
- Translation-based nucleic acid alignment
- ABI trace viewing, editing and printing
- Customizable ... other features

Note: Although BioEdit was recently updated, it is no longer being reliably maintained, and the documentation is out of date and no longer maintained. It is being updated slowly, but there is no guaranteed finish date. Until documentation is complete, play with the menus and see what happens, or email with a question; If you would like to support our effort to create a better alignment editor, please send an email to Carla Kuiken at

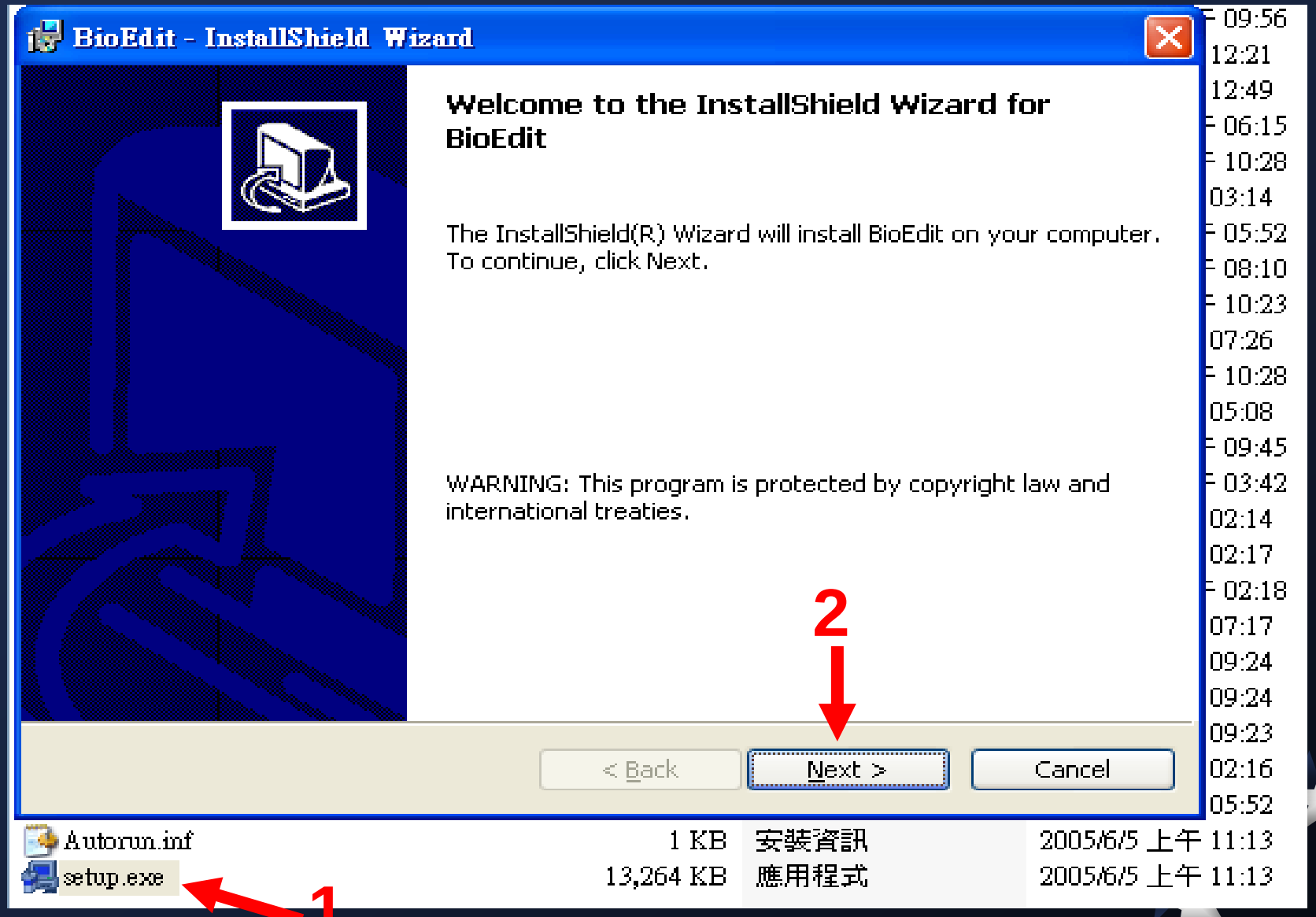
BioEdit introduction

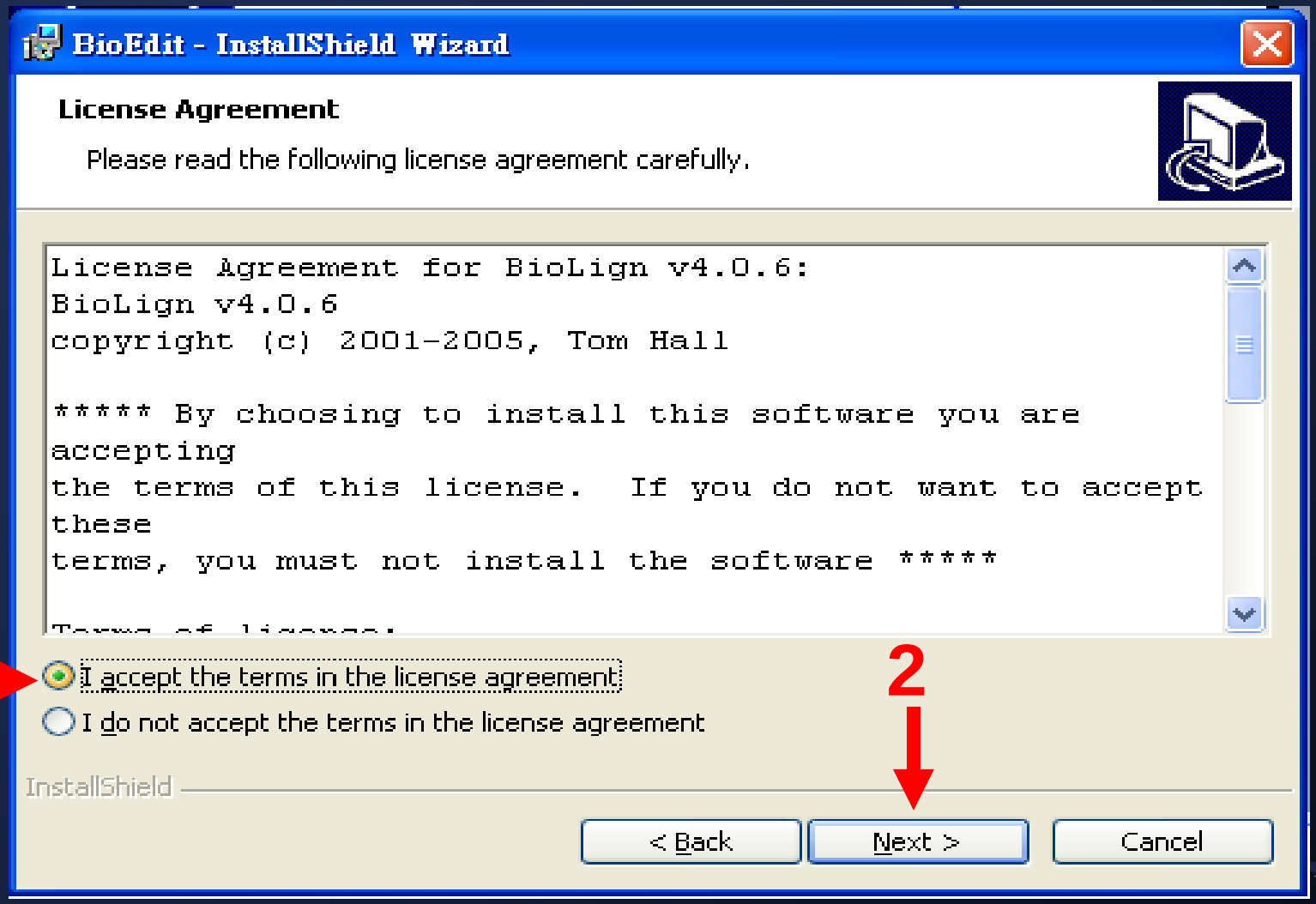
- BioEdit is a mouse-driven, easy-to-use sequence alignment editor and sequence analysis program designed and written by a graduate student who knows how frustrating and time consuming it can be to rely upon word-processors and command-line programs for sequence manipulation.
- BioEdit is intended to supply a single program that can handle most simple sequence and alignment editing and manipulation functions that researchers are likely to do on a daily basis, as well as a few basic sequences analyses.

- http://www.mbio.ncsu.edu/BioEdit/page2.html



1. BioEdit installation





Press "**NEXT**" all the way……..

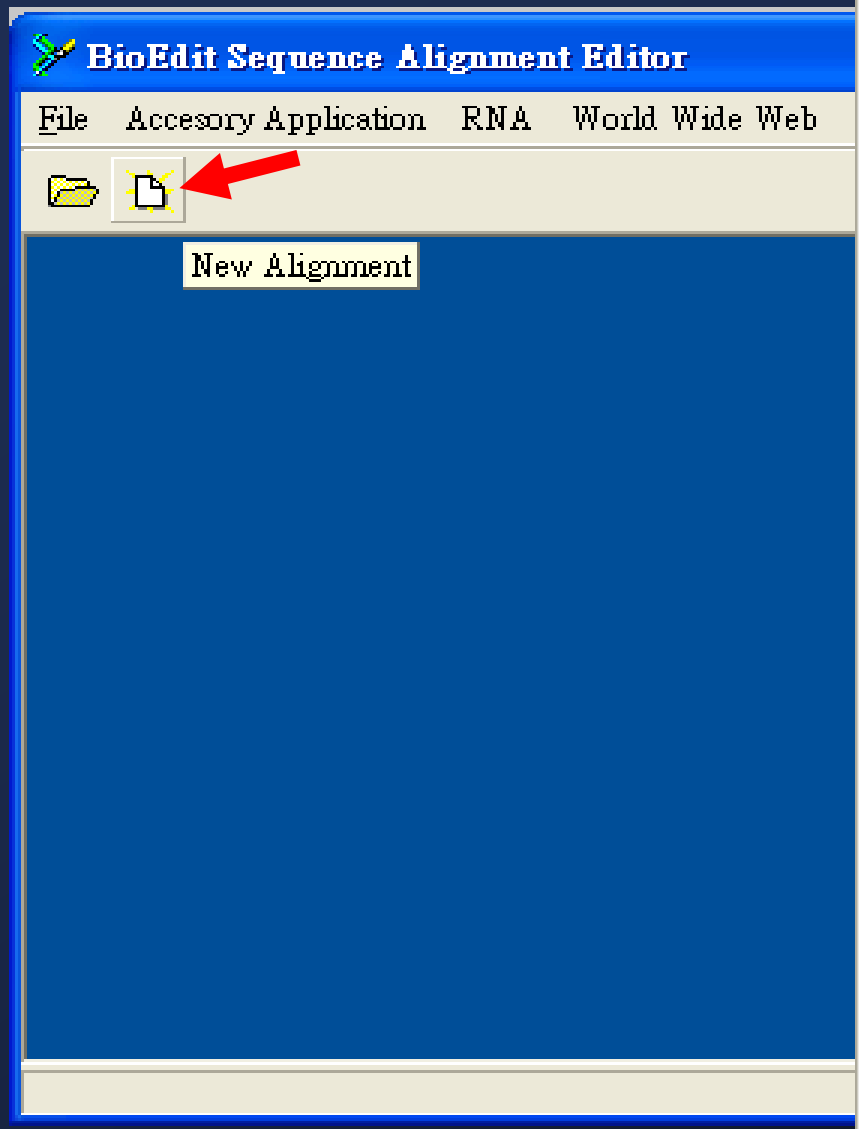
Fasta Format:

- This format contains a one line header followed by lines of sequence data.
- Sequences in fasta formatted files are preceded by a line starting with a " >" symbol.
- The first word on the first line is usually the name of the sequence. The rest of the line is a description of the sequence.
- The remaining lines contain the sequence itself.
- Blank lines in a FASTA file are ignored, and so are spaces in a sequence.
- Fasta files containing multiple sequences are just the same, with one sequence listed right after another. This format is accepted for many multiple sequence alignment programs.

```
>FOSB_MOUSE Protein fosB. 338 bp  |
MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS
GGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT
DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD
LPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTASLFTHSEVQVLGDPFPVVSPSY
TSSFVLTCPEVSAFAGAQRTSGSEQPSDPLNSPSLLAL
>HsRAD51_human
MAMQMQLEANADTSVEEESFGPQPISRLEQCGINANDVKKLEEAGFHTVEAVAYAPKKELINIKGISEAKADKILAEAAK
LVPMGFTTATEFHQRRSEIIQITTGSKELDKLLQGGIETGSITEMFGEFRTGKTQICHTLAVTCQLPIDRGGGEGKAMYI
DTEGTFRPERLLAVAERYGLSGSDVLDNVAYARAFNTDHQTQLLYQASAMMVESRYALLIVDSATALYRTDYSGRGELSA
RQMHLARFLRMLLRLADEFGVAVVITNQVVAQVDGAAMFAADPKKPIGGNIIAHASTTRLYLRKGRGETRICKIYDSPCL
PEAEAMFAINADGVGDAKD
```

Ref: http://www.ebi.ac.uk/help/formats_frame.html

2. Input/output sequence



New Alignment Ctrl+N
Open ... Ctrl+O
New from Clipboard
RNAMotif
New Text
Open As Text
Save Ctrl+S
Save As...
Retrieve sequences from GenBank or GenPept
Copy file name to clipboard
Export
Import
Import from Clipboard
Merge into Alignment based on a Reference Sequence
Merge from Clipboard
Align and merge into alignment using top sequence as an alignment reference
Append Alignment
Go to GenBank
Close

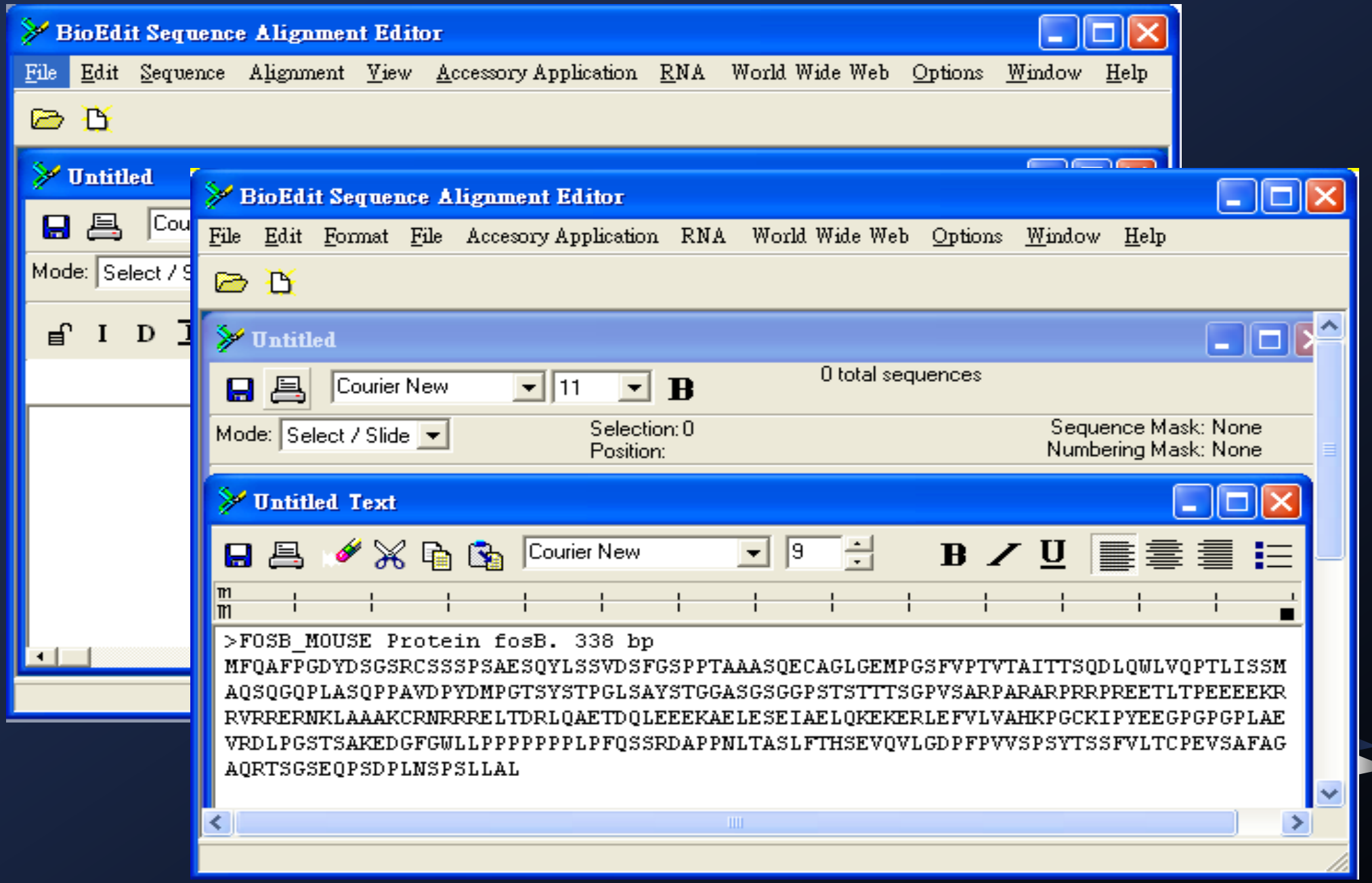
Batch ABI to SCF trace file conversion
Batch Export of Raw Sequence Trace Data

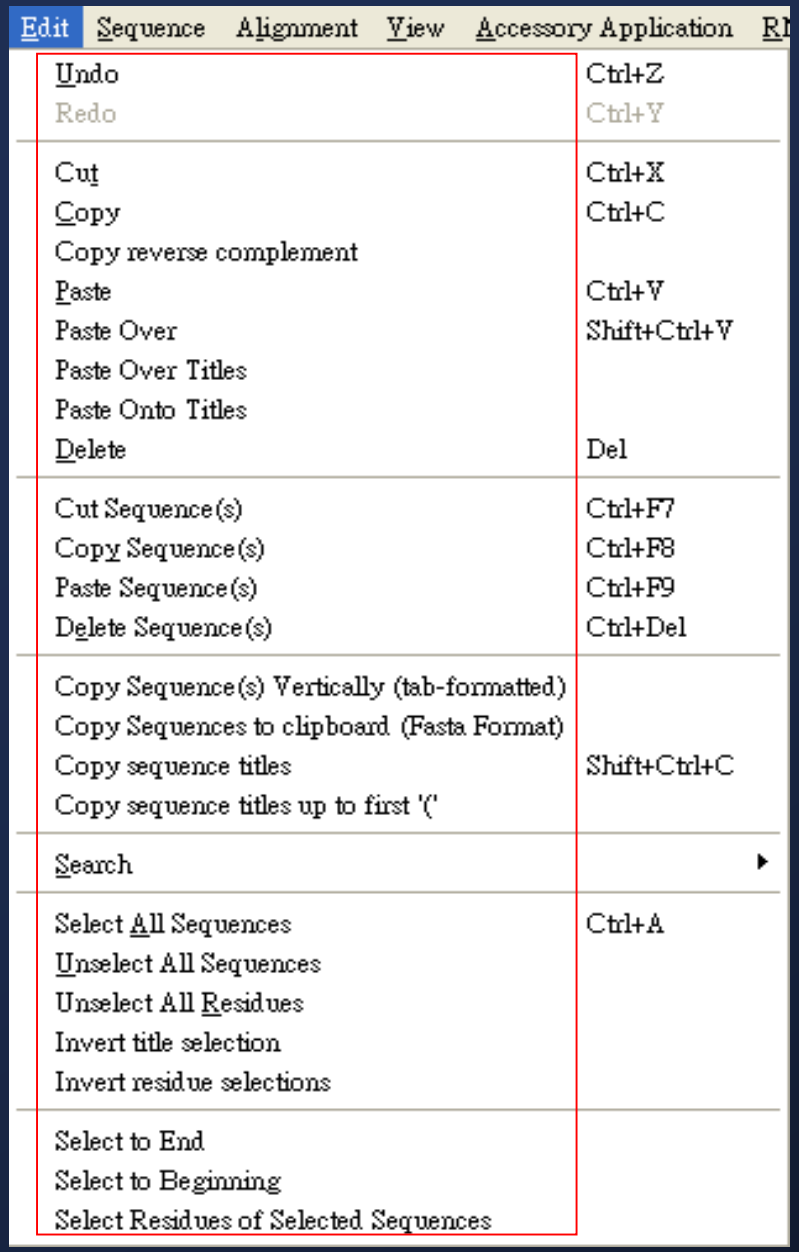
Revert to Saved Ctrl+R

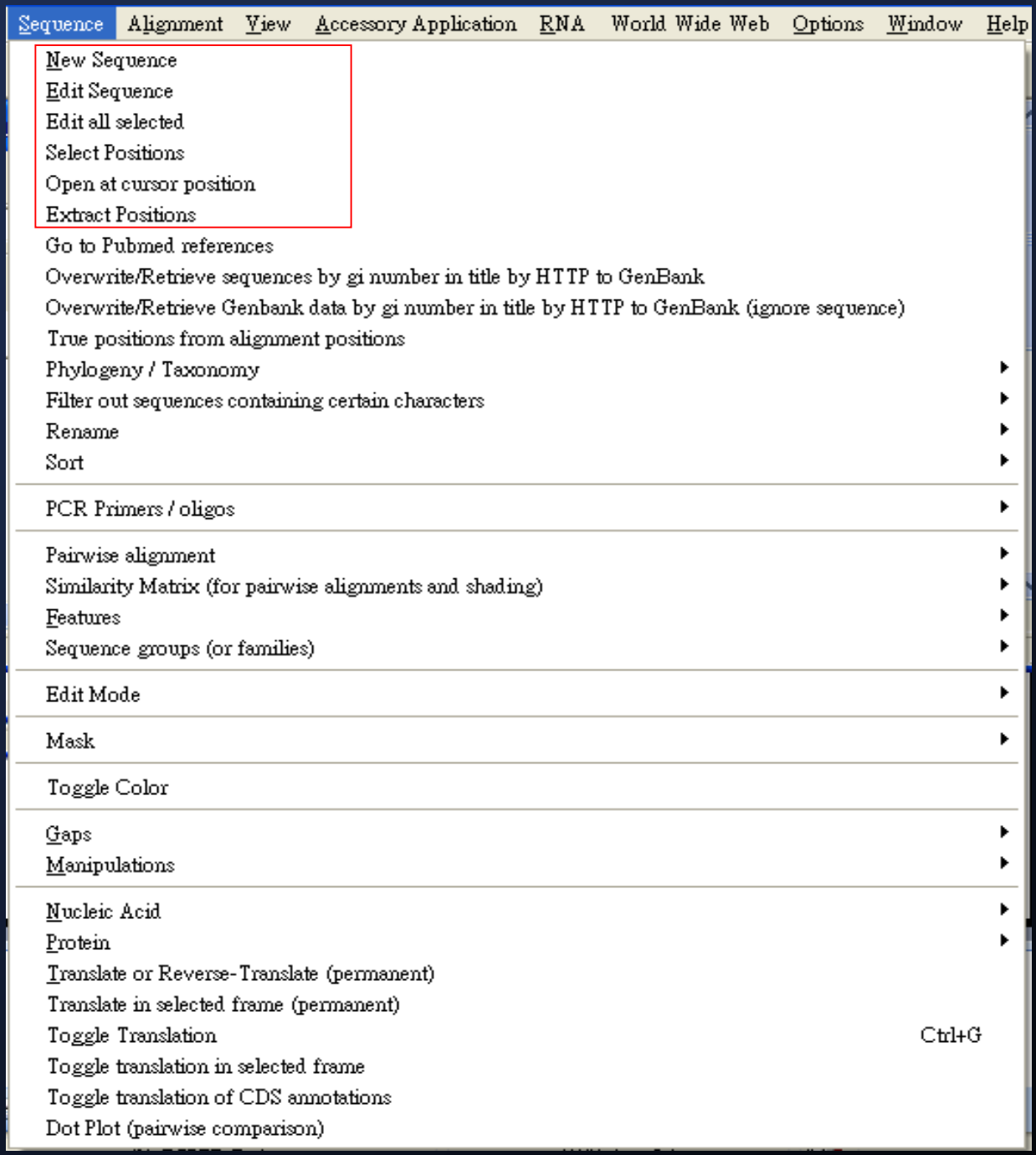
Graphic View
Print Window
Print Alignment as Text Ctrl+P

C:\BioEdit\Temp\~out.tmp
C:\Documents and Settings\kevinwu\桌面\test.fasta

Exit

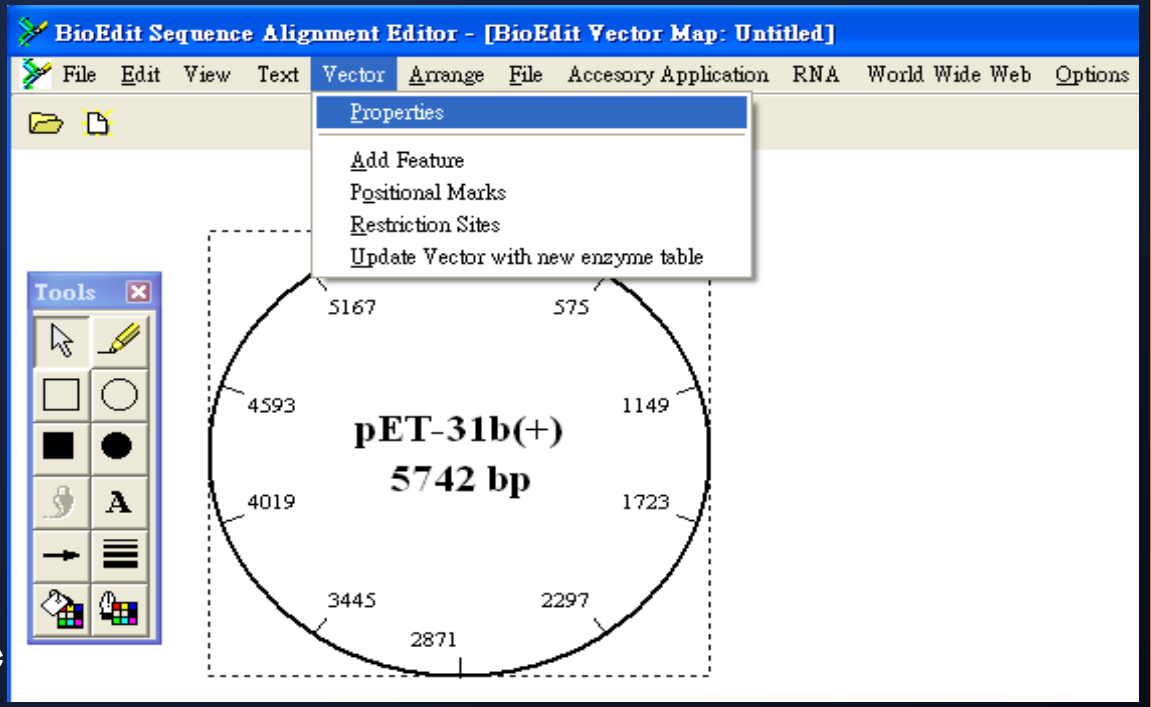

BioEdit Sequence Alignment Editor
File Edit Sequence Alignment View Accessory Application RNA World Wide Web Options Window Help
Untitled
Mode: Select / S
BioEdit Sequence Alignment Editor
File Edit Format File Accesory Application RNA World Wide Web Options Window Help
Untitled
Courier New
11
0 total sequences
Mode: Select / Slide
Selection: 0
Position:
Sequence Mask: None
Numbering Mask: None
Untitled Text
Courier New
9
>FOSB_MOUSE Protein fosB. 338 bp
MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTAITTSQDLQWLVQPTLISSM
AQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGSGGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKR
RVRRERNKLAAAKCRNRRRELTDRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAE
VRDLPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTASLFTHSEVQVLGDPFPVVSPSYTSSFVLTCPEVSAFAG
AQRTSGSEQPSDPLNSPSLLAL


Edit
Sequence
Alignment
View
Accessory Application
Undo Ctrl+Z
Redo Ctrl+Y
Cut Ctrl+X
Copy Ctrl+C
Copy reverse complement
Paste Ctrl+V
Paste Over Shift+Ctrl+V
Paste Over Titles
Paste Onto Titles
Delete Del
Cut Sequence(s) Ctrl+F7
Copy Sequence(s) Ctrl+F8
Paste Sequence(s) Ctrl+F9
Delete Sequence(s) Ctrl+Del
Copy Sequence(s) Vertically (tab-formatted)
Copy Sequences to clipboard (Fasta Format)
Copy sequence titles Shift+Ctrl+C
Copy sequence titles up to first '('
Search
Select All Sequences Ctrl+A
Unselect All Sequences
Unselect All Residues
Invert title selection
Invert residue selections
Select to End
Select to Beginning
Select Residues of Selected Sequences

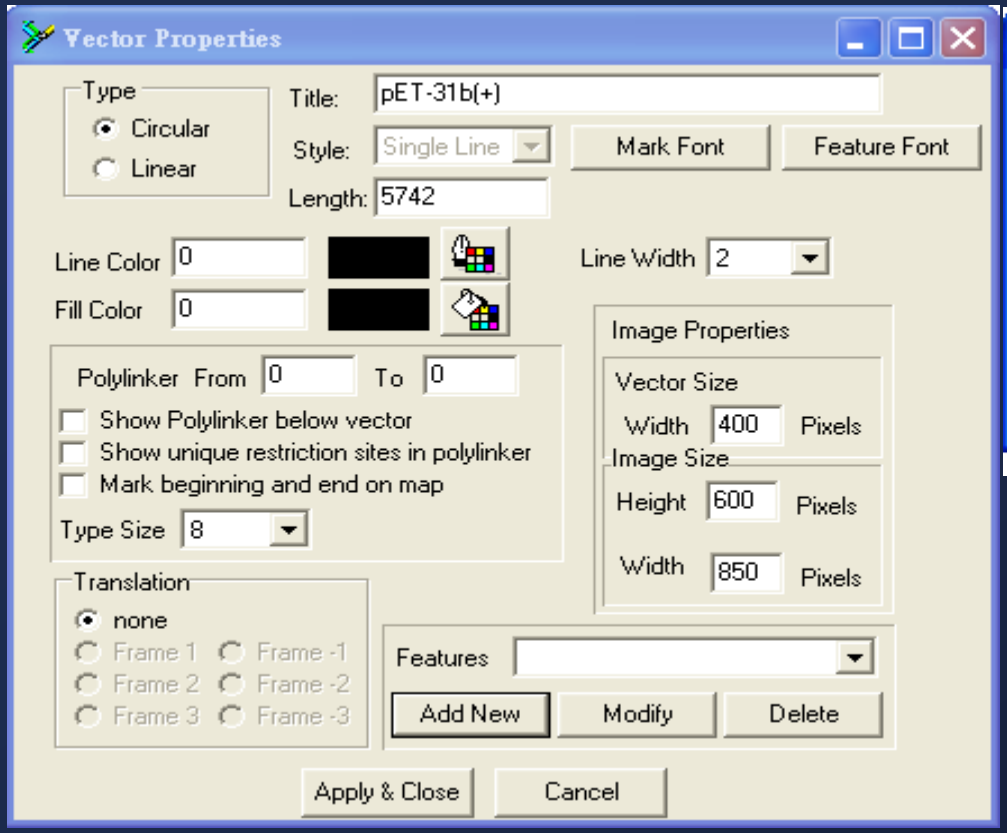

Sequence
Alignment
View
Accessory Application
RNA
World Wide Web
Options
Window
Help
New Sequence
Edit Sequence
Edit all selected
Select Positions
Open at cursor position
Extract Positions
Go to Pubmed references
Overwrite/Retrieve sequences by gi number in title by HTTP to GenBank
Overwrite/Retrieve Genbank data by gi number in title by HTTP to GenBank (ignore sequence)
True positions from alignment positions
Phylogeny / Taxonomy
Filter out sequences containing certain characters
Rename
Sort
PCR Primers / oligos
Pairwise alignment
Similarity Matrix (for pairwise alignments and shading)
Features
Sequence groups (or families)
Edit Mode
Mask
Toggle Color
Gaps
Manipulations
Nucleic Acid
Protein
Translate or Reverse-Translate (permanent)
Translate in selected frame (permanent)
Toggle Translation Ctrl+G
Toggle translation in selected frame
Toggle translation of CDS annotations
Dot Plot (pairwise comparison)

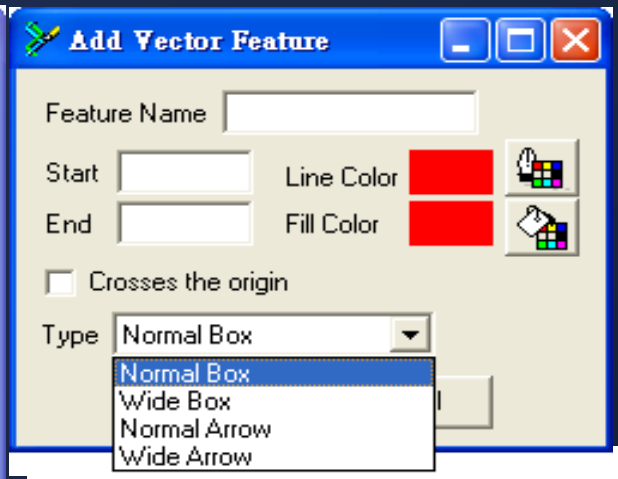
Plasmid Drawing

- 從課程網頁上下載並匯入 **pET31 sequence** 到 **BioEdit** 中
- **[Sequence] > [Nucleic Acid] > [Create Plasmid from Aequence] >**
  - Properties
  - Position Marks
  - Restriction sites

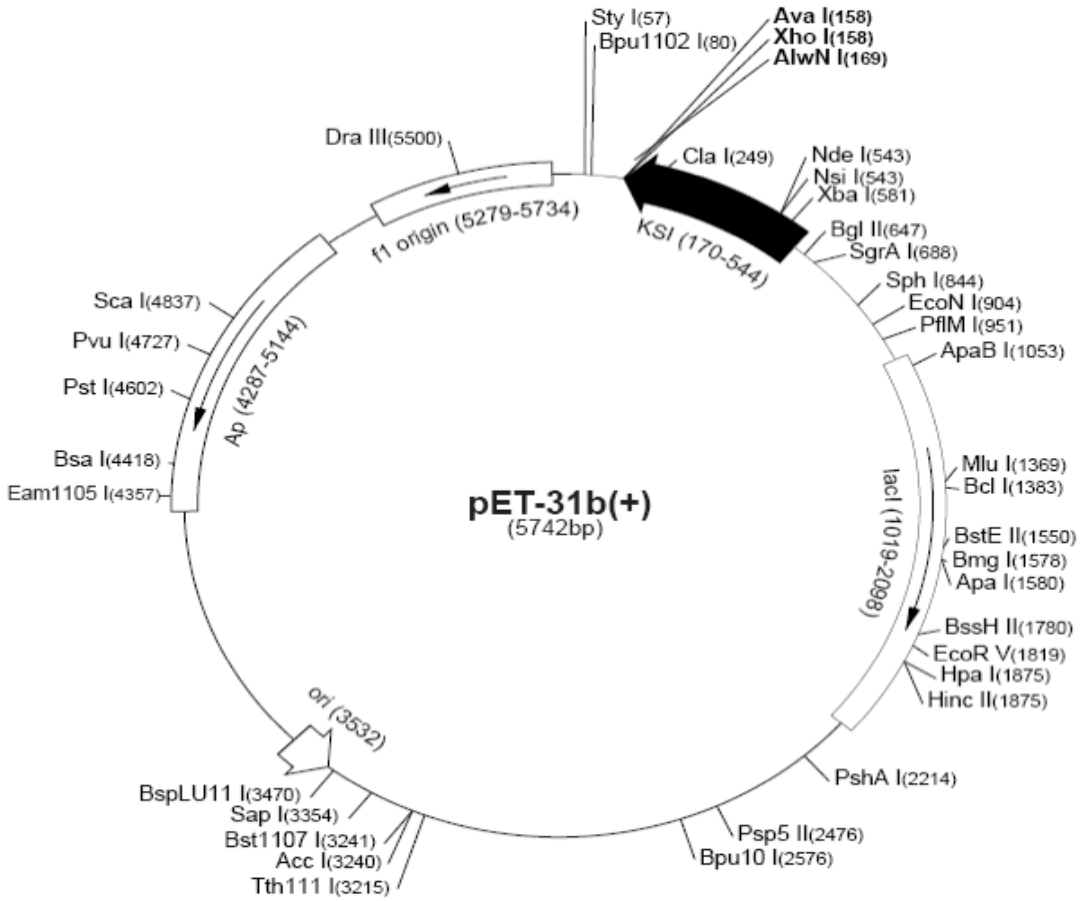


- 存檔 **[File] > [Save as …] > ***
- 匯出圖檔 **[File] > [Export Image …] > *.wmf file**



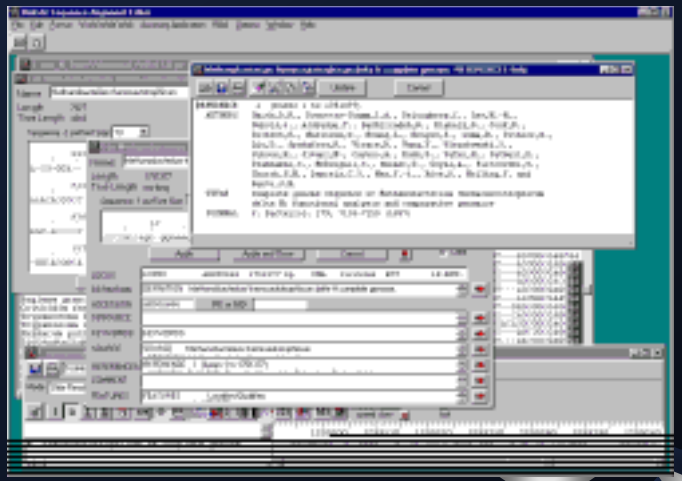


練習加入如右圖所示之 features & restriction sites:





Split view of an alignment of 6205 prokaryotic 16S RNAs showing identities and similaritie



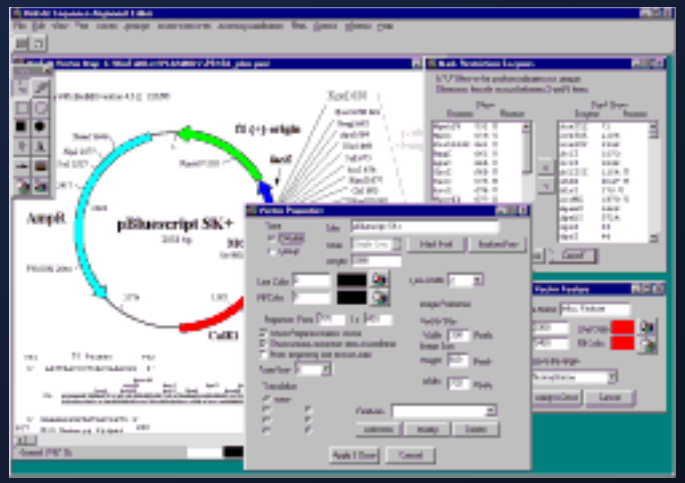
Shows single sequence editing and GenBank info



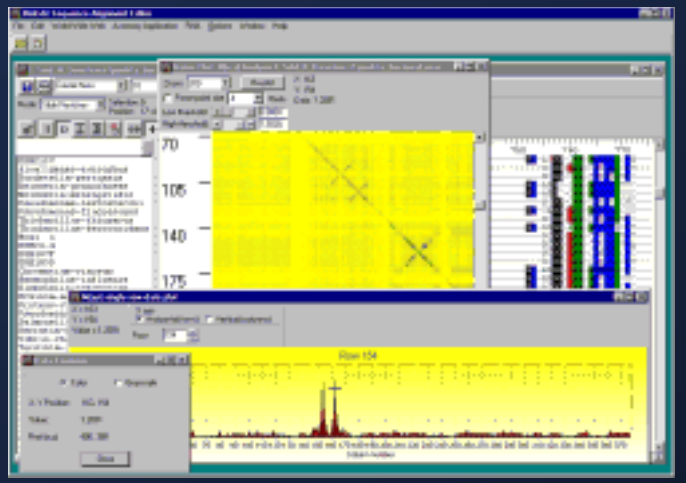
Hydrophobicity plots



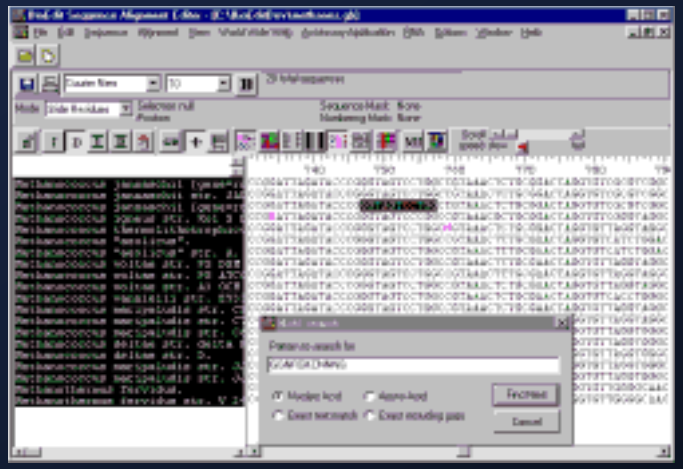
ABI and SCF trace viewing, editing



Plasmid drawing and annotation

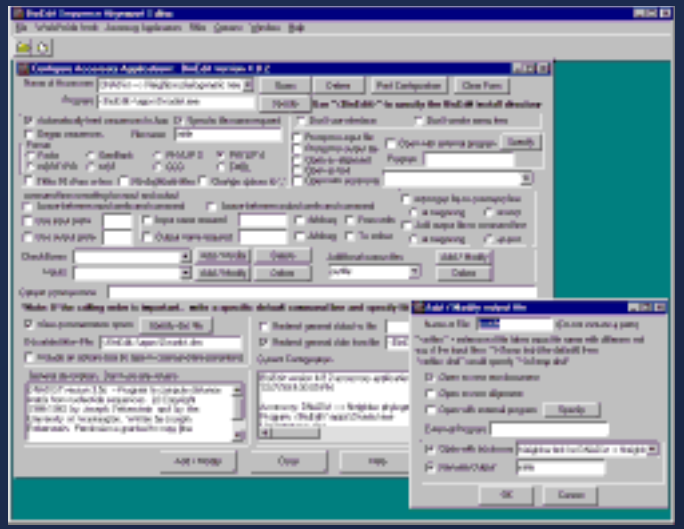


Mutual information analysis and graphical data examination



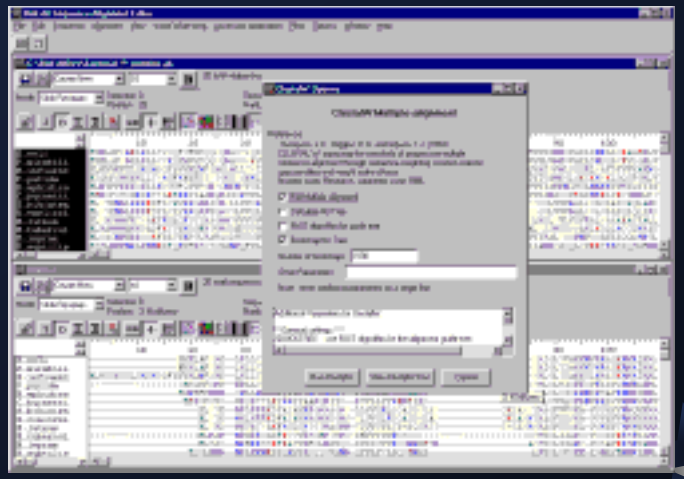
User-defined motif searching

Graphic shaded view of alignment

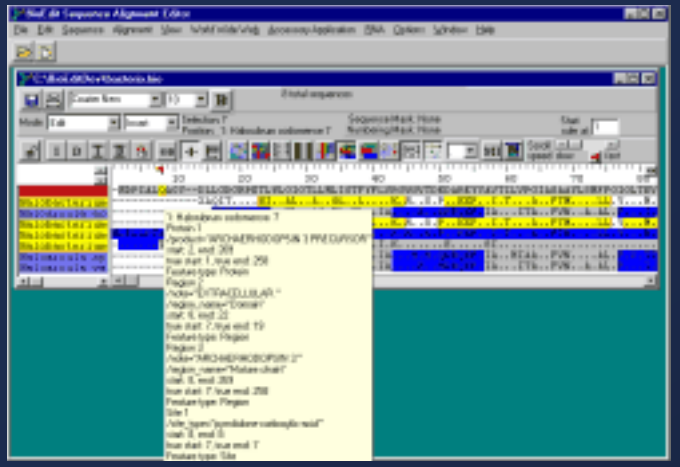




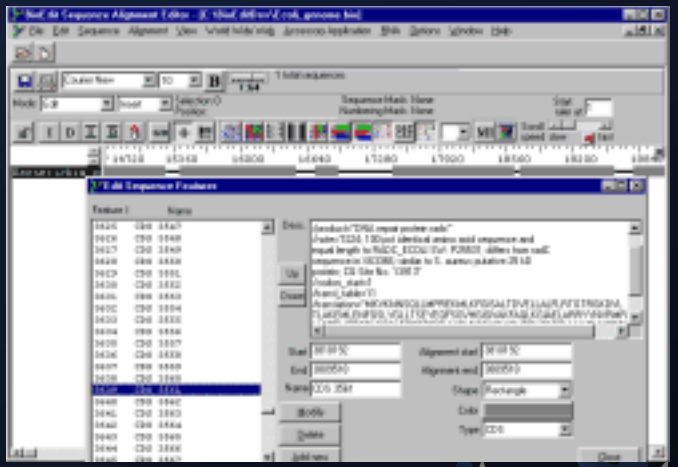
Configuration of external accessory analysis applications



Automated link to ClustalW with command line options available



Dynamic feature annotation shading and information



Easy editing of graphical feature annotations

http://informagen.com/SA/

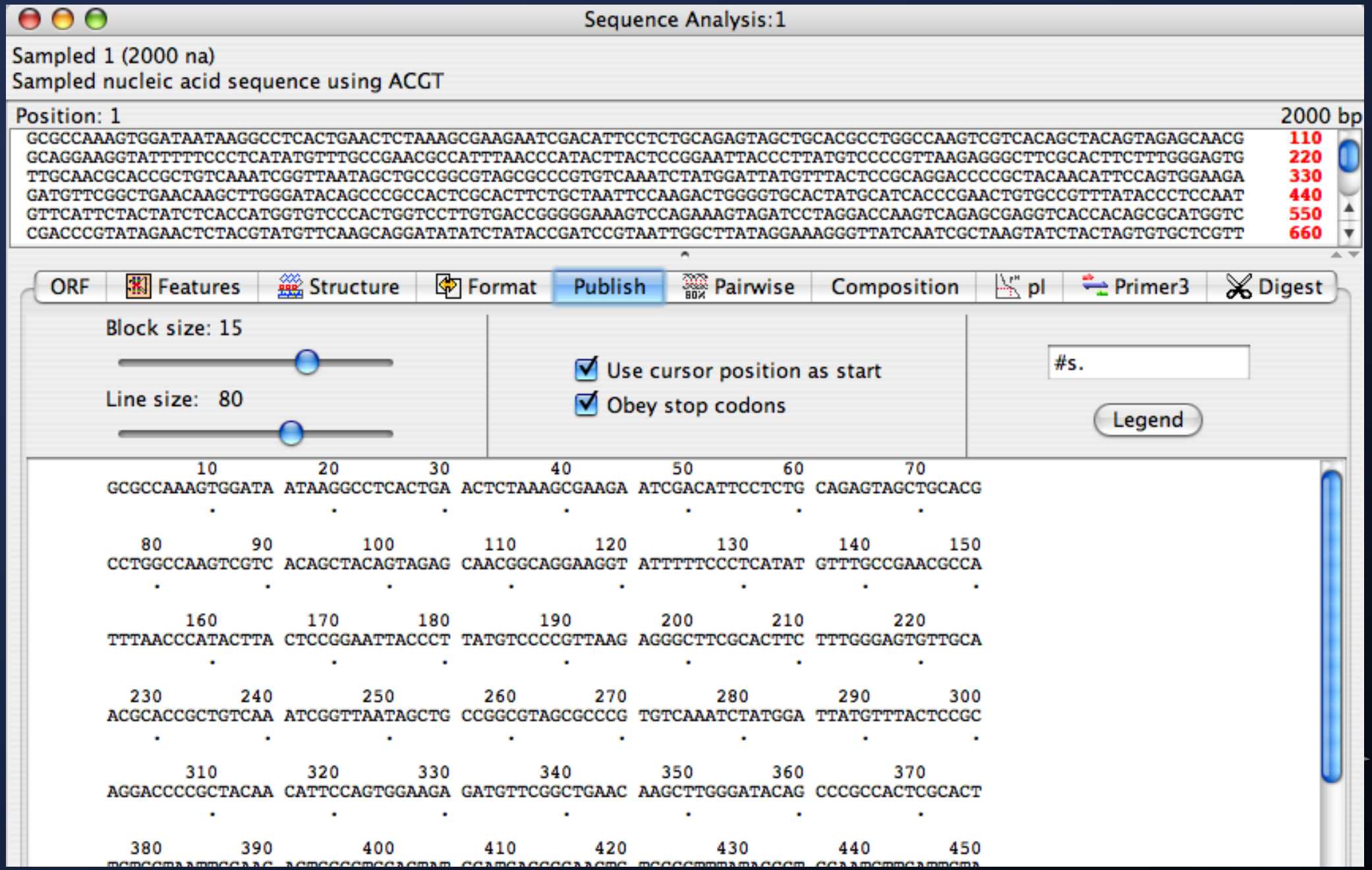


Sequence Analysis

This (Java) application is a collage of coding projects which I have written over the past several years for various clients in work as a bioinformatics consultant. These clients have graciously allowed me to release these works into the public domai as freeware for Macintosh OS X in order to promote the platform and to encourage migration from Classic.

The upper window panel can hold several sequences, which are both editable and selectable. The tabs in the lower analysis panel try to keep up with the current sequence selection to provide immediate feedback. The selection is used in some modu as only the portion being analyzed for other modules i.e. **Digest** is used to determine if enzymes cut in the in or outside of th selection.

Most commonly available sequence formats have been reverse engineered. You can also access a sequences from the NCB via its GID or UID. This currently cannot be done from behind a firewall.


Sequence Analysis:1
Sampled 1 (2000 na)
Sampled nucleic acid sequence using ACGT
Position: 1
2000 bp
GCGCCAAAGTGGATAATAAGGCCTCACTGAACTCTAAAGCGAAGAATCGACATTCCTCTGCAGAGTAGCTGCACGCCTGGCCAAGTCGTCACAGCTACAGTAGAGCAACG 110
GCAGGAAGGTATTTTTCCCTCATATGTTTGCCGAACGCCATTTAACCCATACTTACTCCGGAATTACCCTTATGTCCCCGTTAAGAGGGCTTCGCACTTCTTTGGGAGTG 220
TTGCAACGCACCGCTGTCAAATCGGTTAATAGCTGCCGGCGTAGCGCCCGTGTCAAATCTATGGATTATGTTTACTCCGCAGGACCCCGCTACAACATTCCAGTGGAAGA 330
GATGTTCGGCTGAACAAGCTTGGGATACAGCCCGCCACTCGCACTTCTGCTAATTCCAAGACTGGGGTGCACTATGCATCACCCGAACTGTGCCGTTTATACCCTCCAAT 440
GTTCATTCTACTATCTCACCATGGTGTCCCACTGGTCCTTGTGACCGGGGGAAAGTCCAGAAAGTAGATCCTAGGACCAAGTCAGAGCGAGGTCACCACAGCGCATGGTC 550
CGACCCGTATAGAACTCTACGTATGTTCAAGCAGGATATATCTATACCGATCCGTAATTGGCTTATAGGAAAGGGTTATCAATCGCTAAGTATCTACTAGTGTGCTCGTT 660
ORF
Features
Structure
Format
Publish
Pairwise
Composition
pI
Primer3
Digest
Block size: 15
Line size: 80
Use cursor position as start
Obey stop codons
#s.
Legend
10 20 30 40 50 60 70
GCGCCAAAGTGGATA ATAAGGCCTCACTGA ACTCTAAAGCGAAGA ATCGACATTCCTCTG CAGAGTAGCTGCACG
80 90 100 110 120 130 140 150
CCTGGCCAAGTCGTC ACAGCTACAGTAGAG CAACGGCAGGAAGGT ATTTTTCCCTCATAT GTTTGCCGAACGCCA
160 170 180 190 200 210 220
TTTAACCCATACTTA CTCCGGAATTACCCT TATGTCCCCGTTAAG AGGGCTTCGCACTTC TTTGGGAGTGTTGCA
230 240 250 260 270 280 290 300
ACGCACCGCTGTCAA ATCGGTTAATAGCTG CCGGCGTAGCGCCCG TGTCAAATCTATGGA TTATGTTTACTCCGC
310 320 330 340 350 360 370
AGGACCCCGCTACAA CATTCCAGTGGAAGA GATGTTCGGCTGAAC AAGCTTGGGATACAG CCCGCCACTCGCACT
380 390 400 410 420 430 440 450

Next Generation Sequencing



Cost-effective
Fast
Ultra throughput
Cloning-free
Short reads











NGS Data Format

- **Different NGS platforms produce a variety of data formats**
  - Solexa (Illumina): *.fastq
  - SOLiD (ABI): *.cFasta, *.qual, *.stats
- **Fastq format is more popular.**
- **There are also different flavors of fastq formats**
  - Sanger
  - Solexa
  - Illumina 1.3+, 1.5+, 1.8+



FASTQ and FASTA Formats

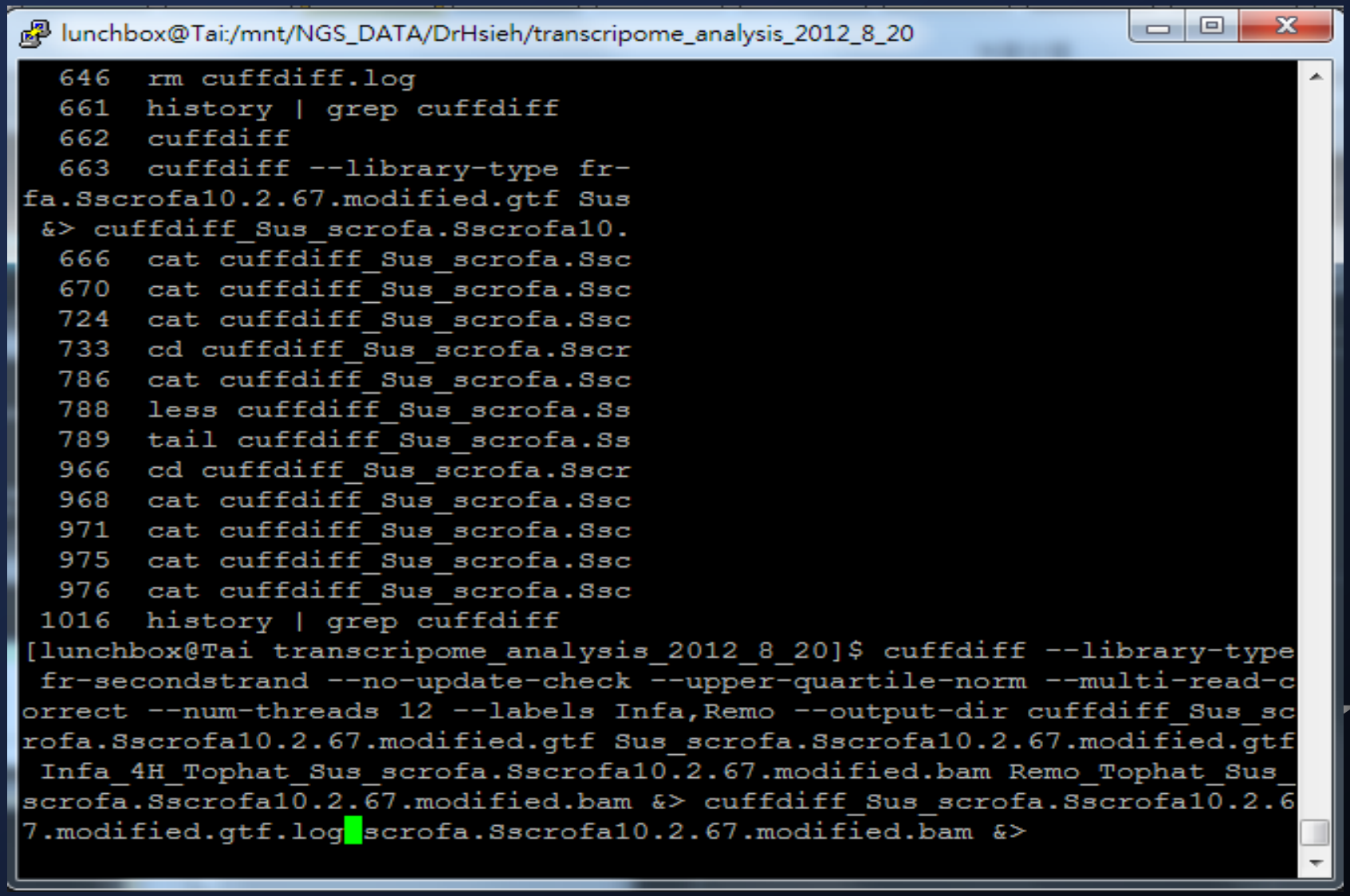
- **FastA format (everybody knows about it)**
  - Header line starts with “>” followed by a sequence ID
  - Sequence (string of nt).
- **FastQ format**
  - First is the sequence (like Fasta but starting with “@”)
  - Then “+” and sequence ID (optional) and in the following line are QVs encoded as single byte ASCII codes
- **Different quality encode variants**
- **Nearly all downstream analysis take FastQ as input sequence**



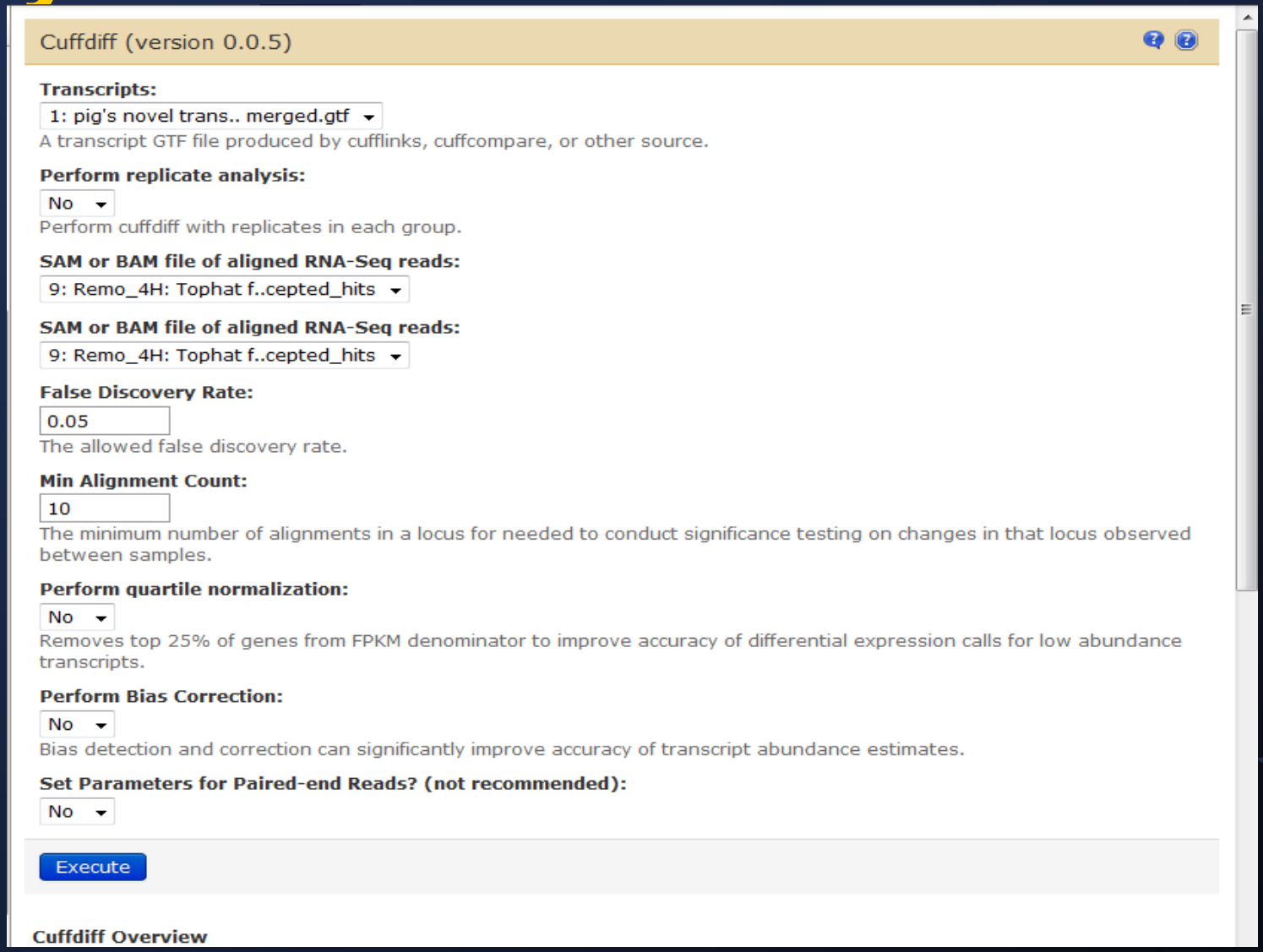
Galaxy Platform



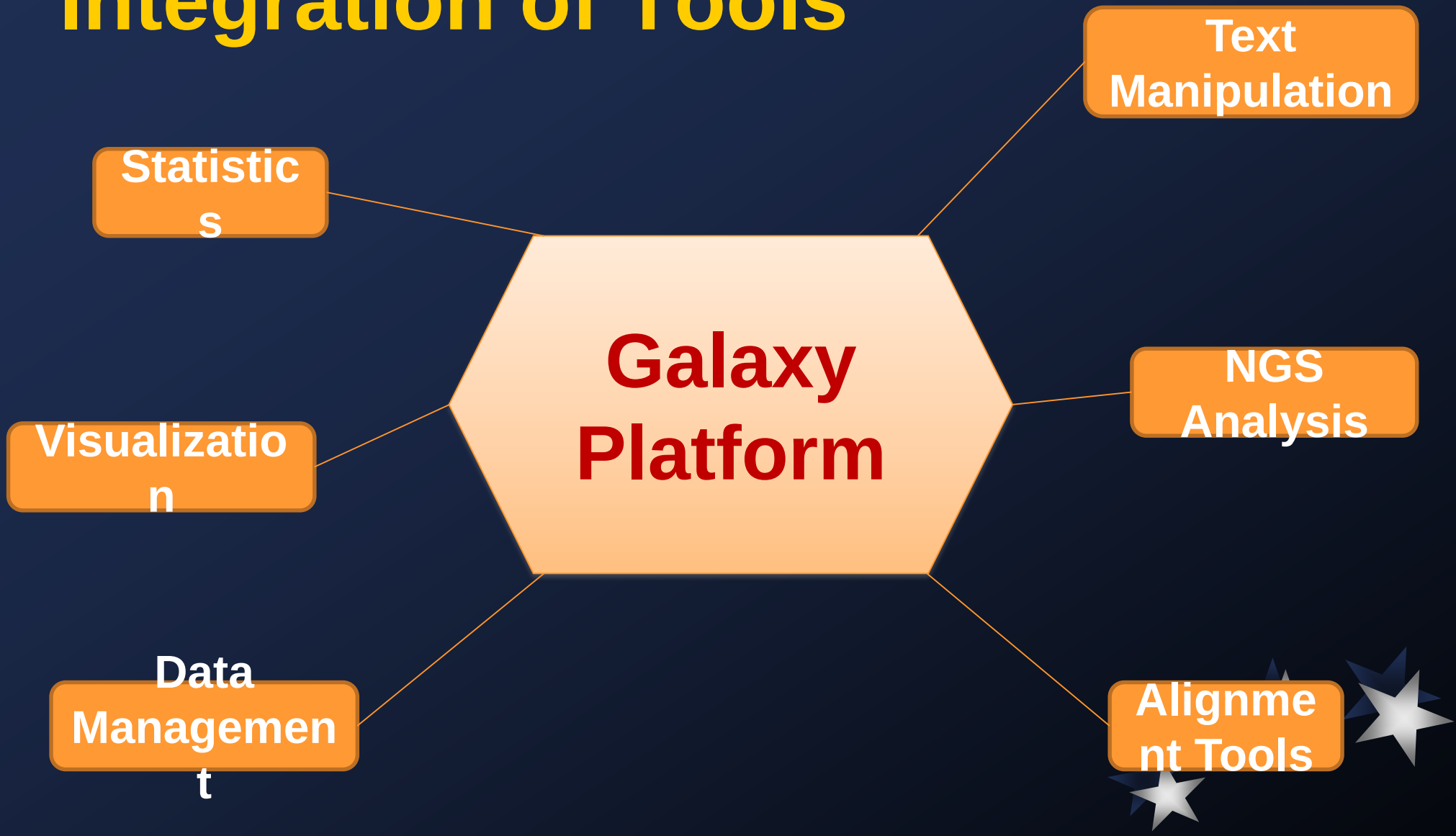
Turn command line into ....



Way better interface



Integration of Tools

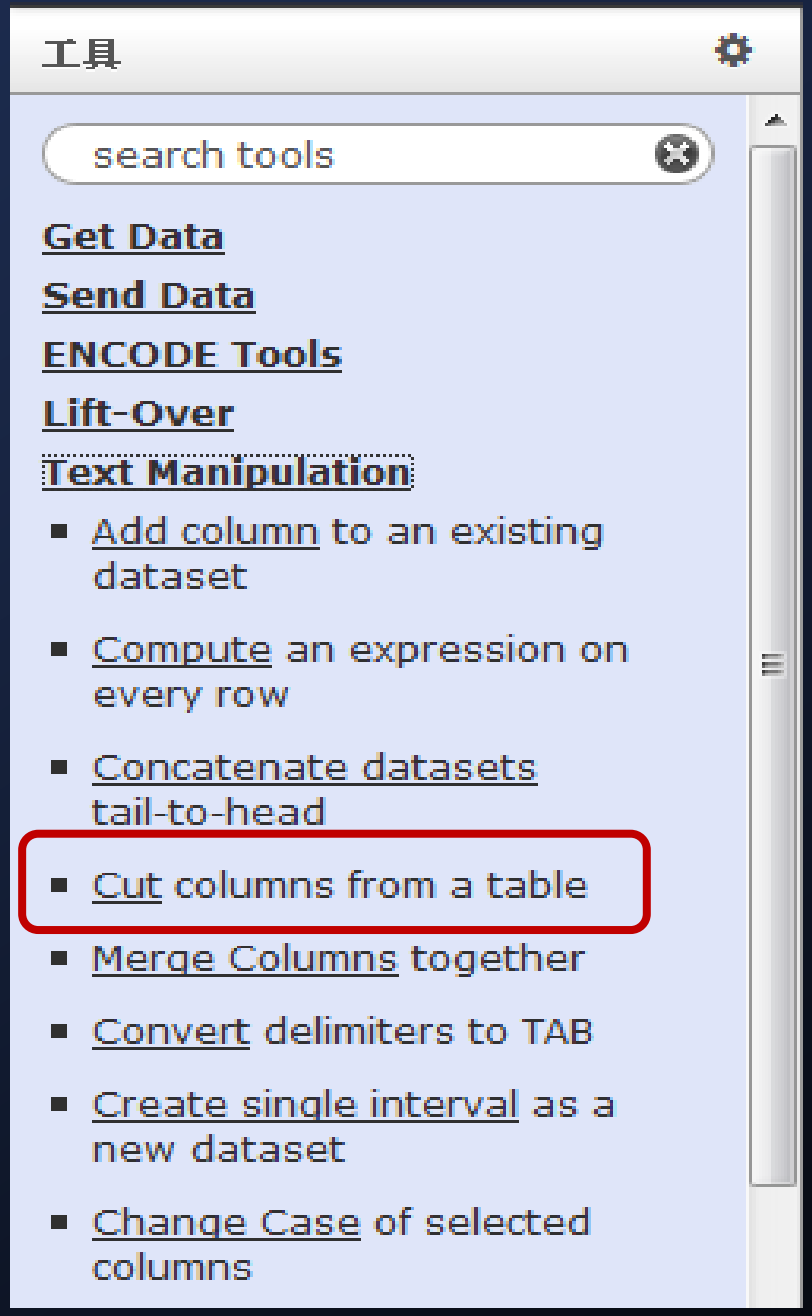


Basic Tutorial: Text Manipulation

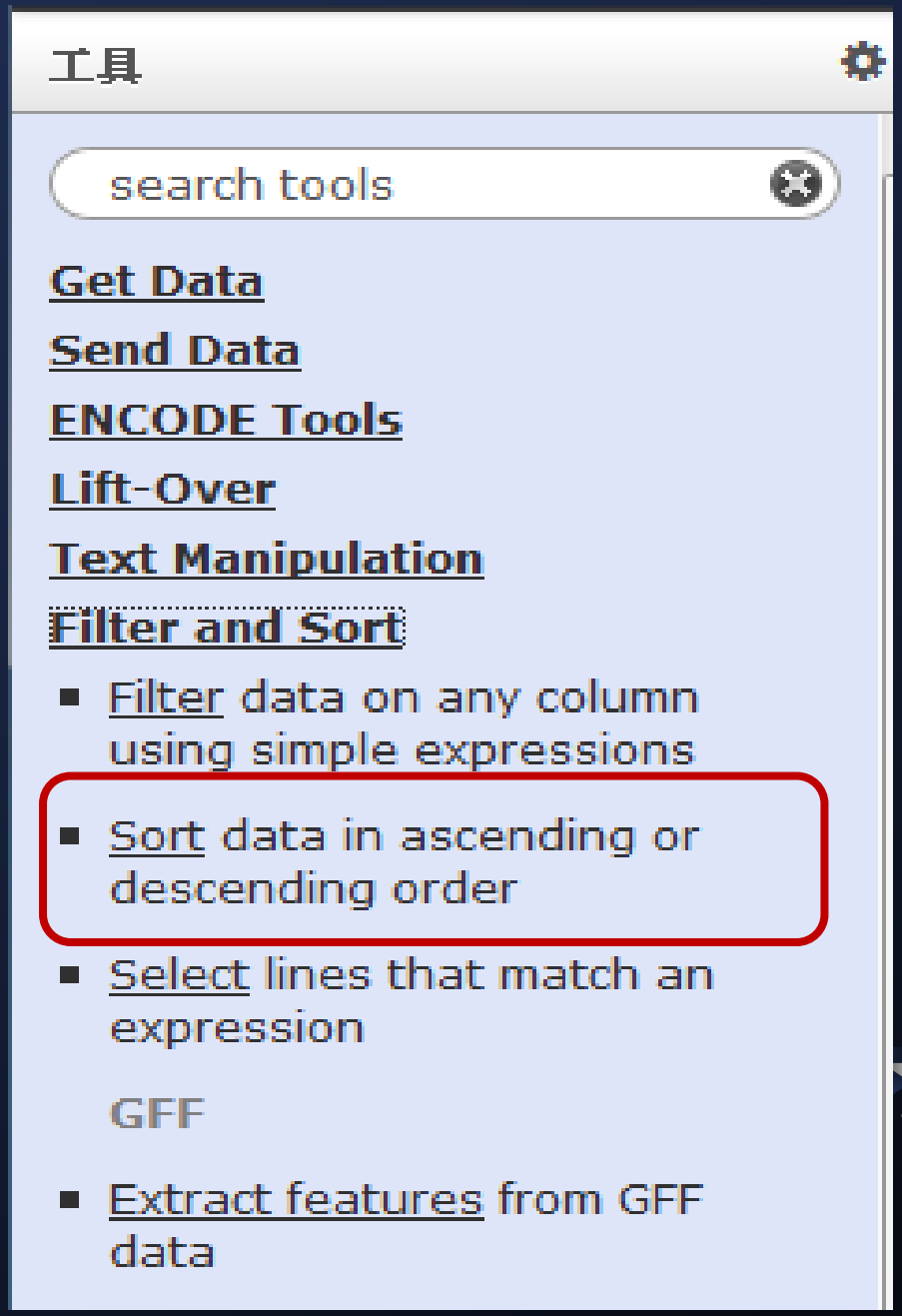
- Cut
- Sort
- Word Count
- Extract first lines of a file
- Format conversion



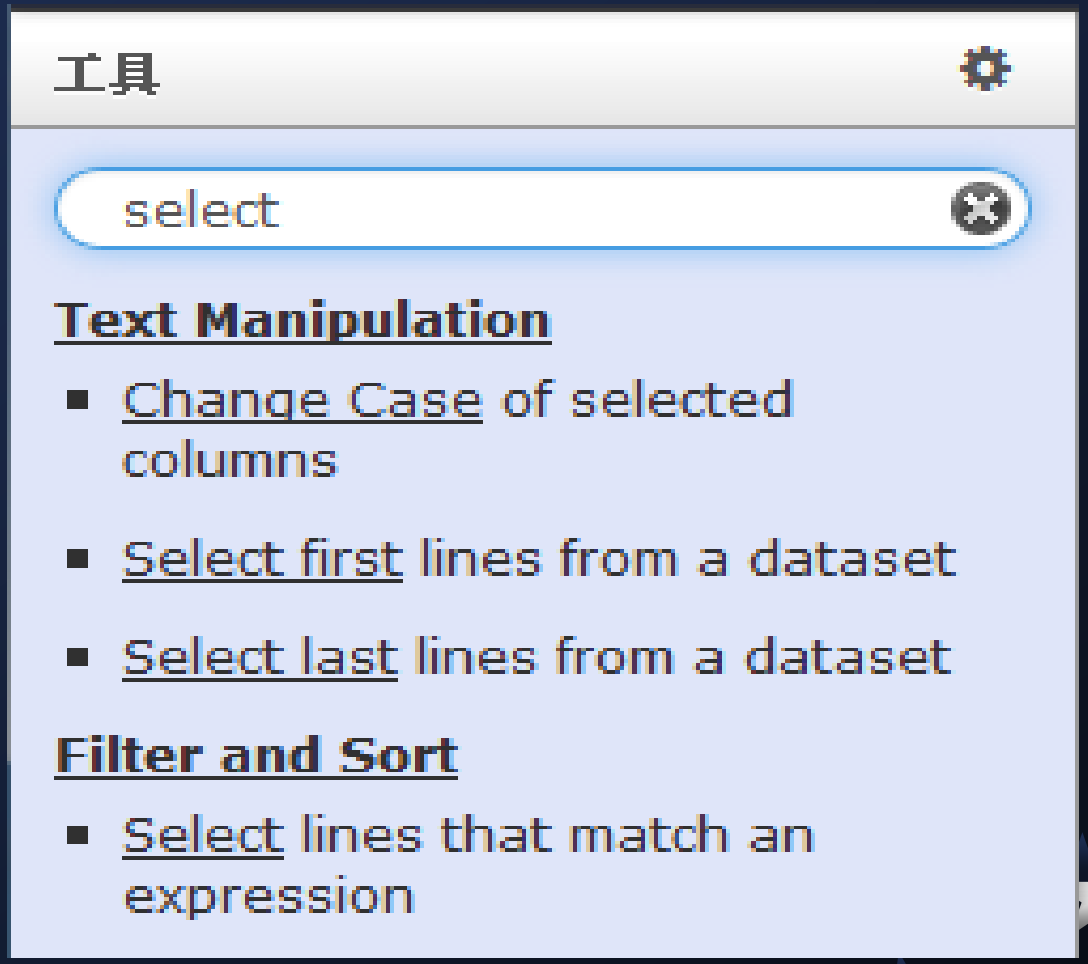
Cut


工具
search tools
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Add column to an existing dataset
Compute an expression on every row
Concatenate datasets tail-to-head
Cut columns from a table
Merge Columns together
Convert delimiters to TAB
Create single interval as a new dataset
Change Case of selected columns

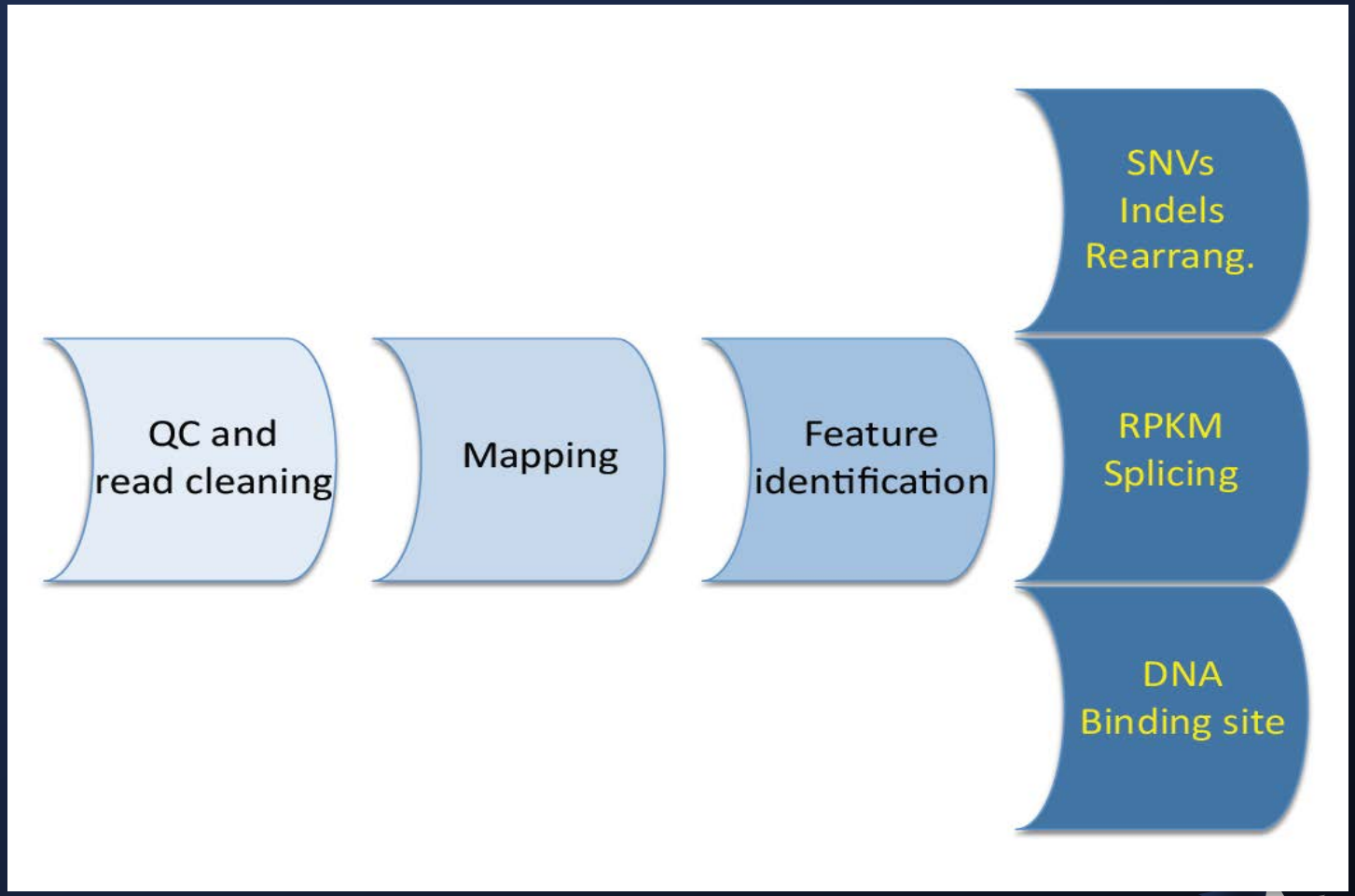
Sort


工具
search tools
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort
Filter data on any column using simple expressions
Sort data in ascending or descending order
Select lines that match an expression
GFF
Extract features from GFF data

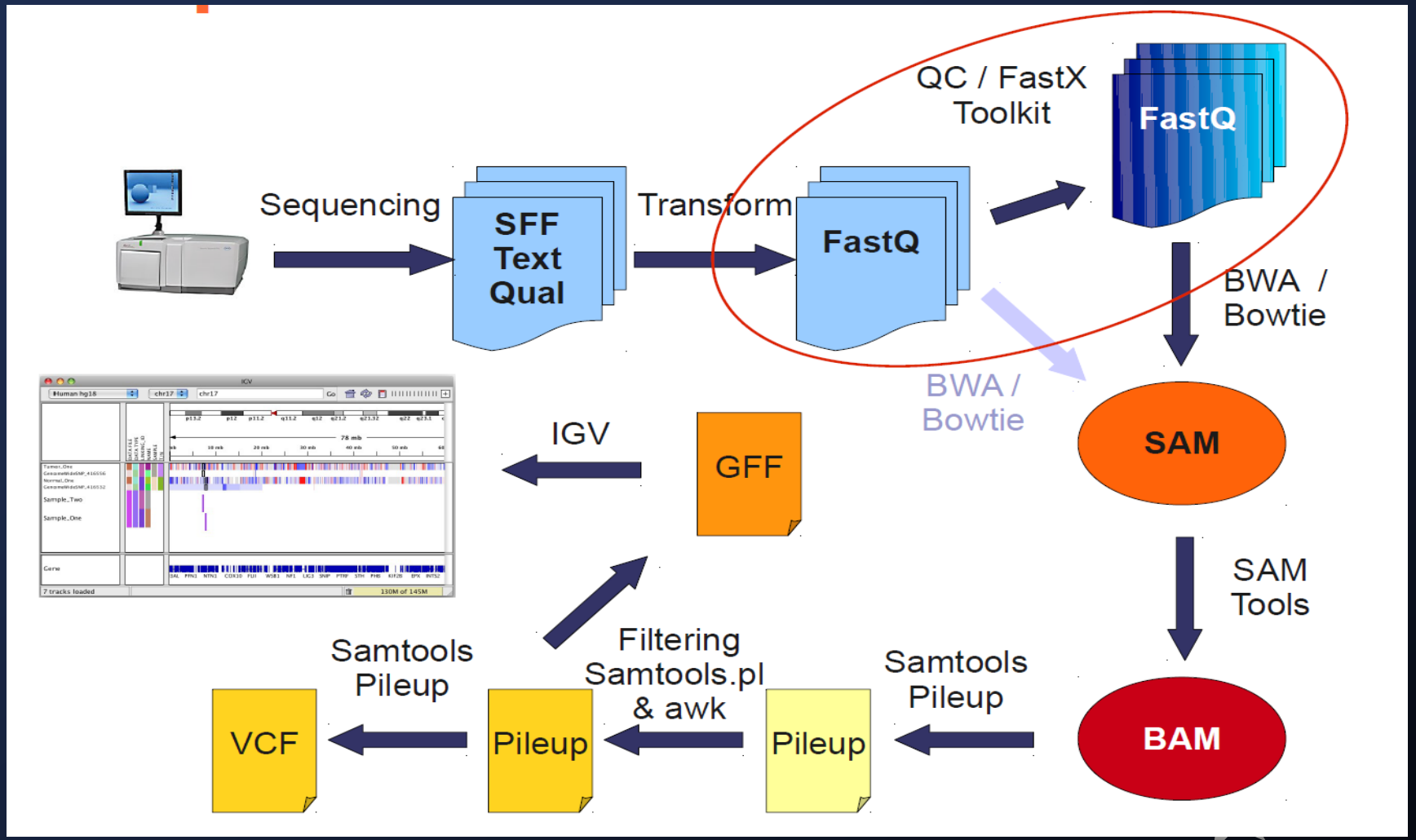
Search tools



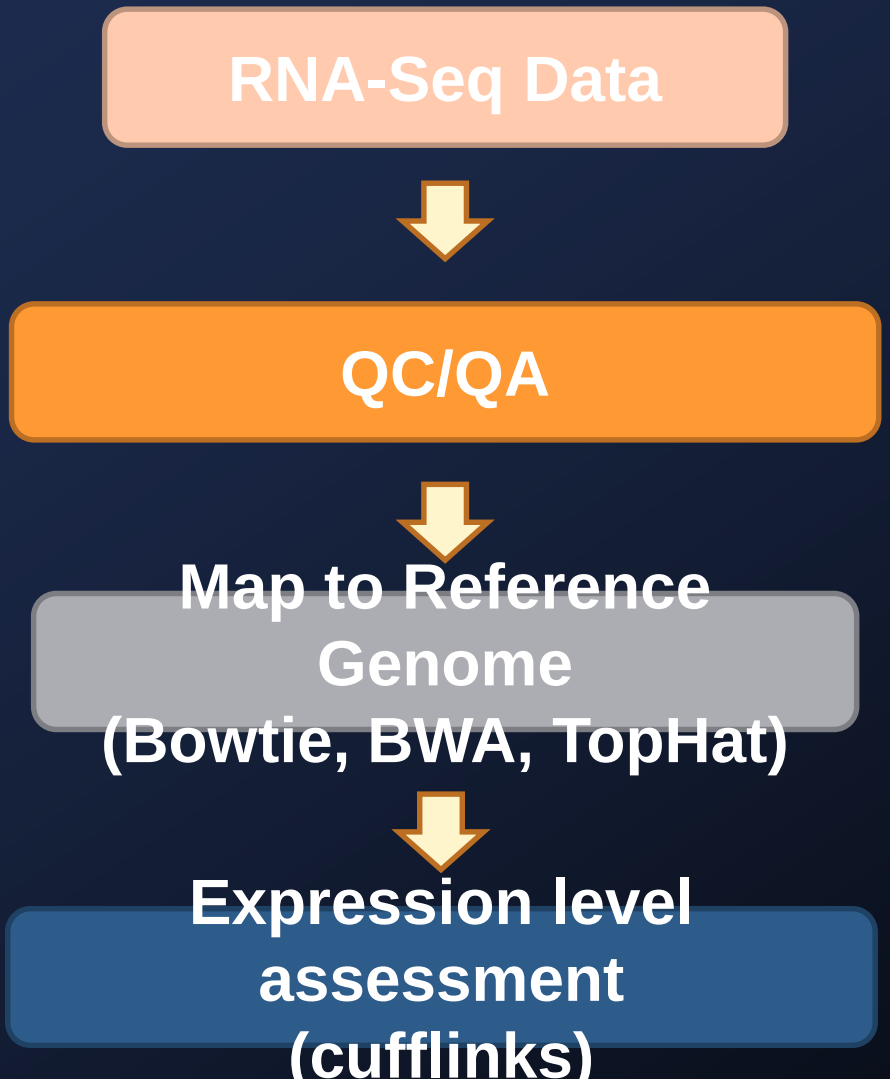
Basic Steps of NGS Data Processing



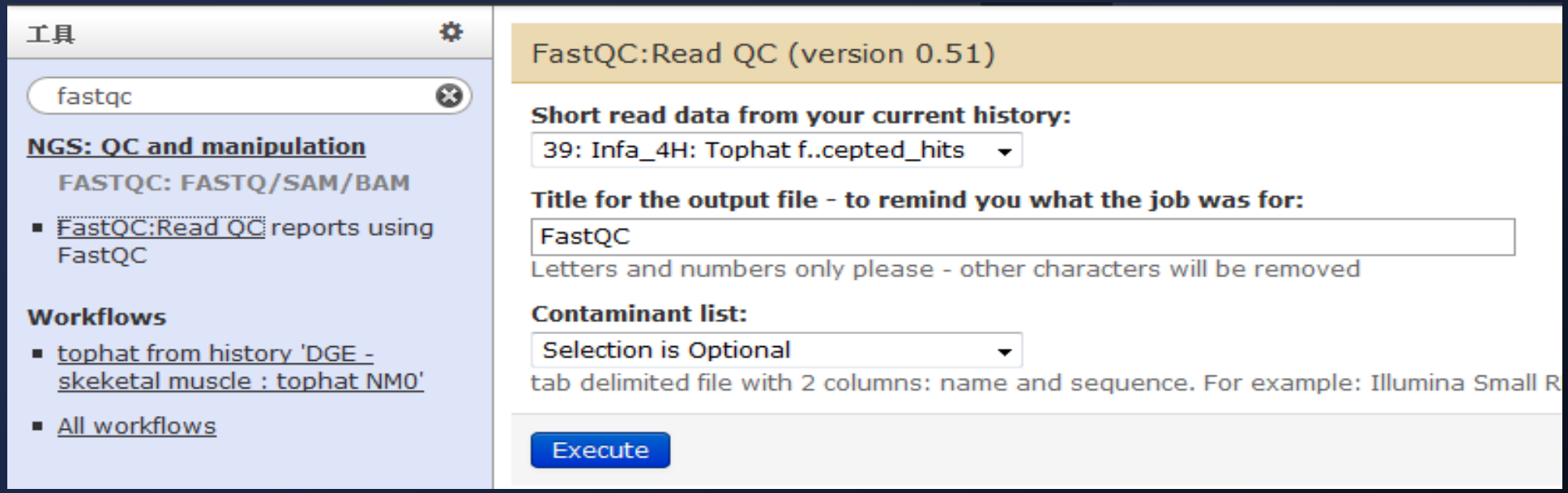
NGS Data Processing Overview



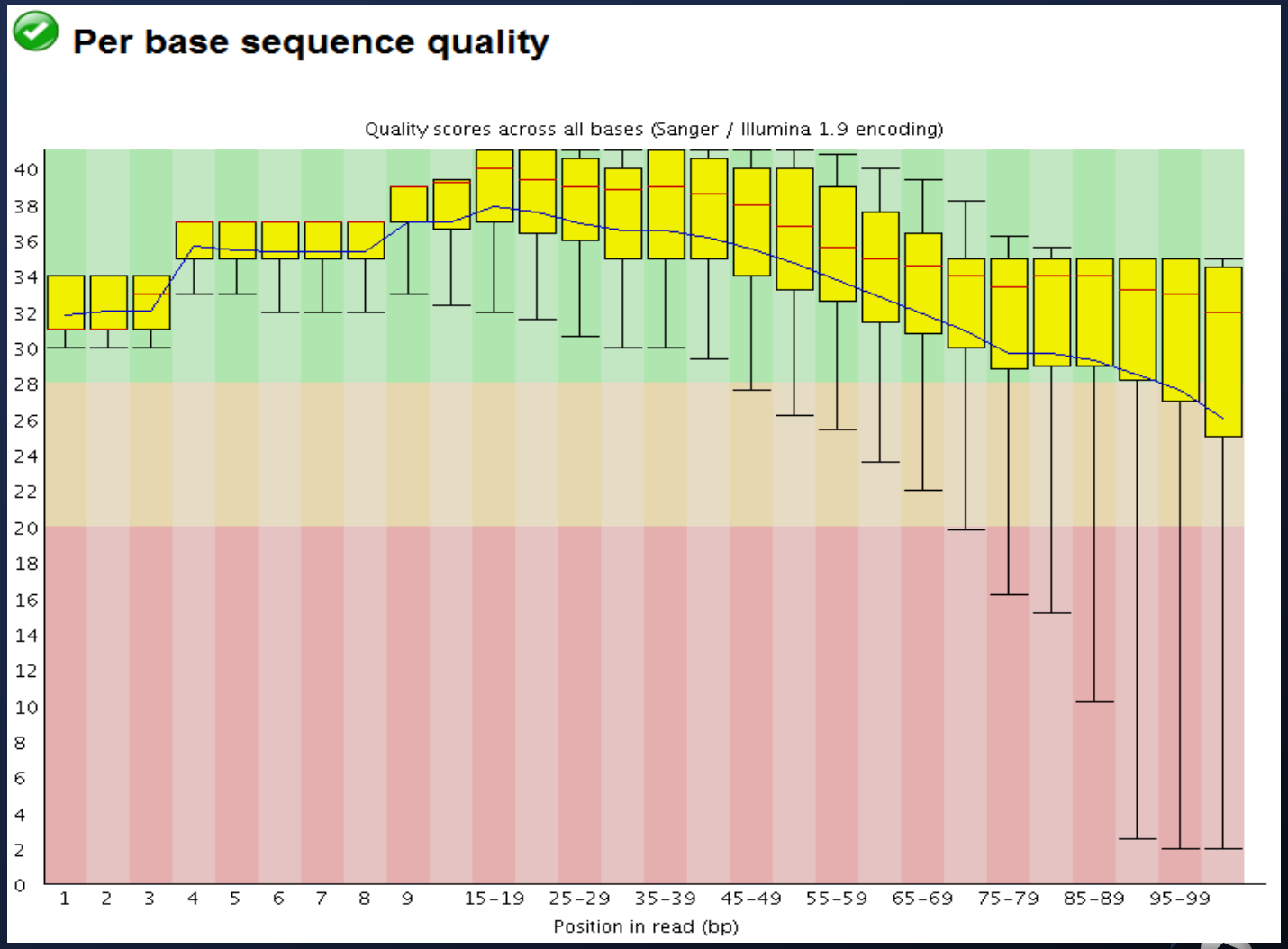
Overview of RNA-seq Data Analysis



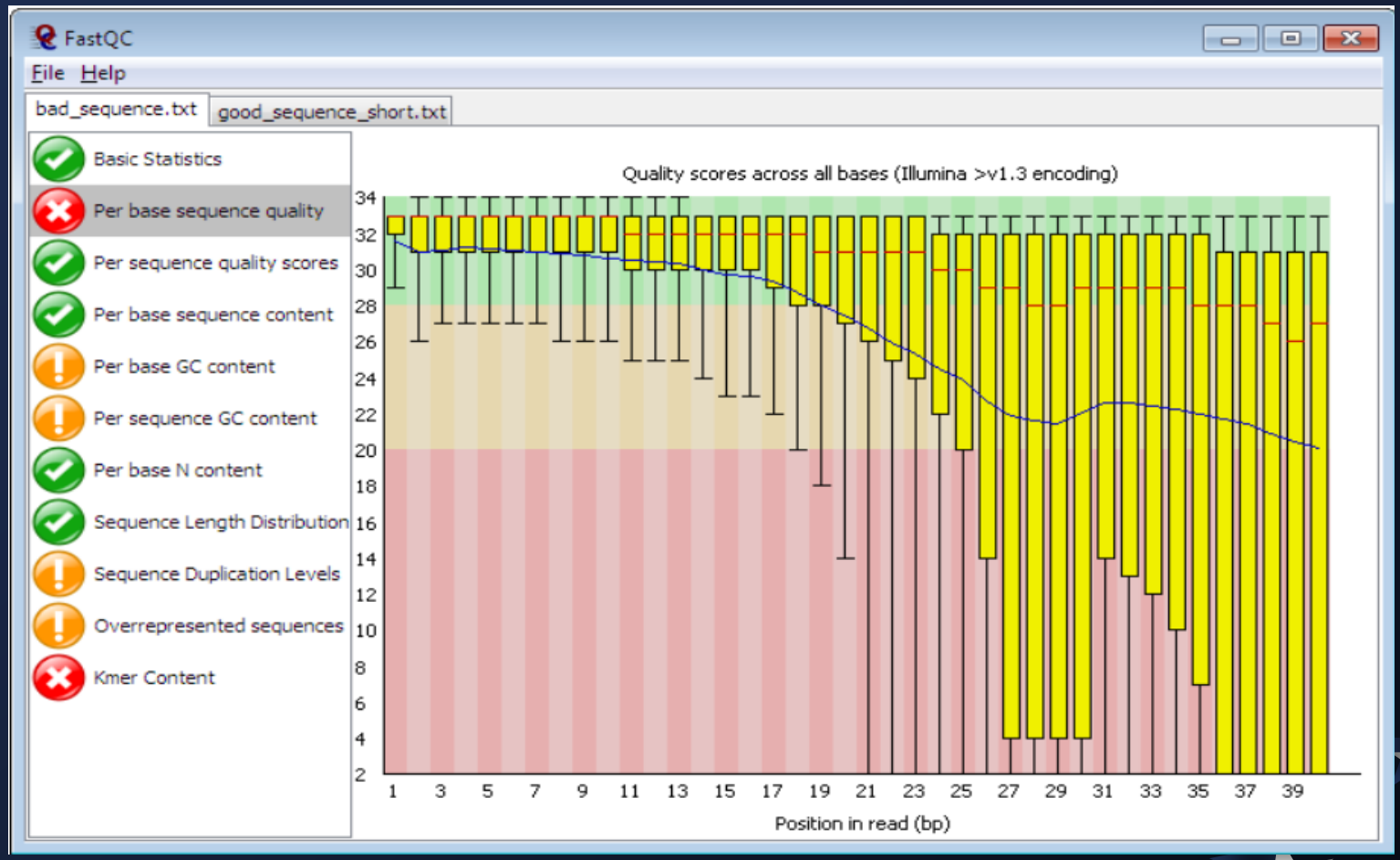
QC: FastQC


工具
fastqc
NGS: QC and manipulation
FASTQC: FASTQ/SAM/BAM
FastQC:Read QC reports using FastQC
Workflows
tophat from history 'DGE - skeketal muscle : tophat NM0'
All workflows
FastQC:Read QC (version 0.51)
Short read data from your current history:
39: Infa_4H: Tophat f..cepted_hits
Title for the output file - to remind you what the job was for:
FastQC
Letters and numbers only please - other characters will be removed
Contaminant list:
Selection is Optional
tab delimited file with 2 columns: name and sequence. For example: Illumina Small R
Execute

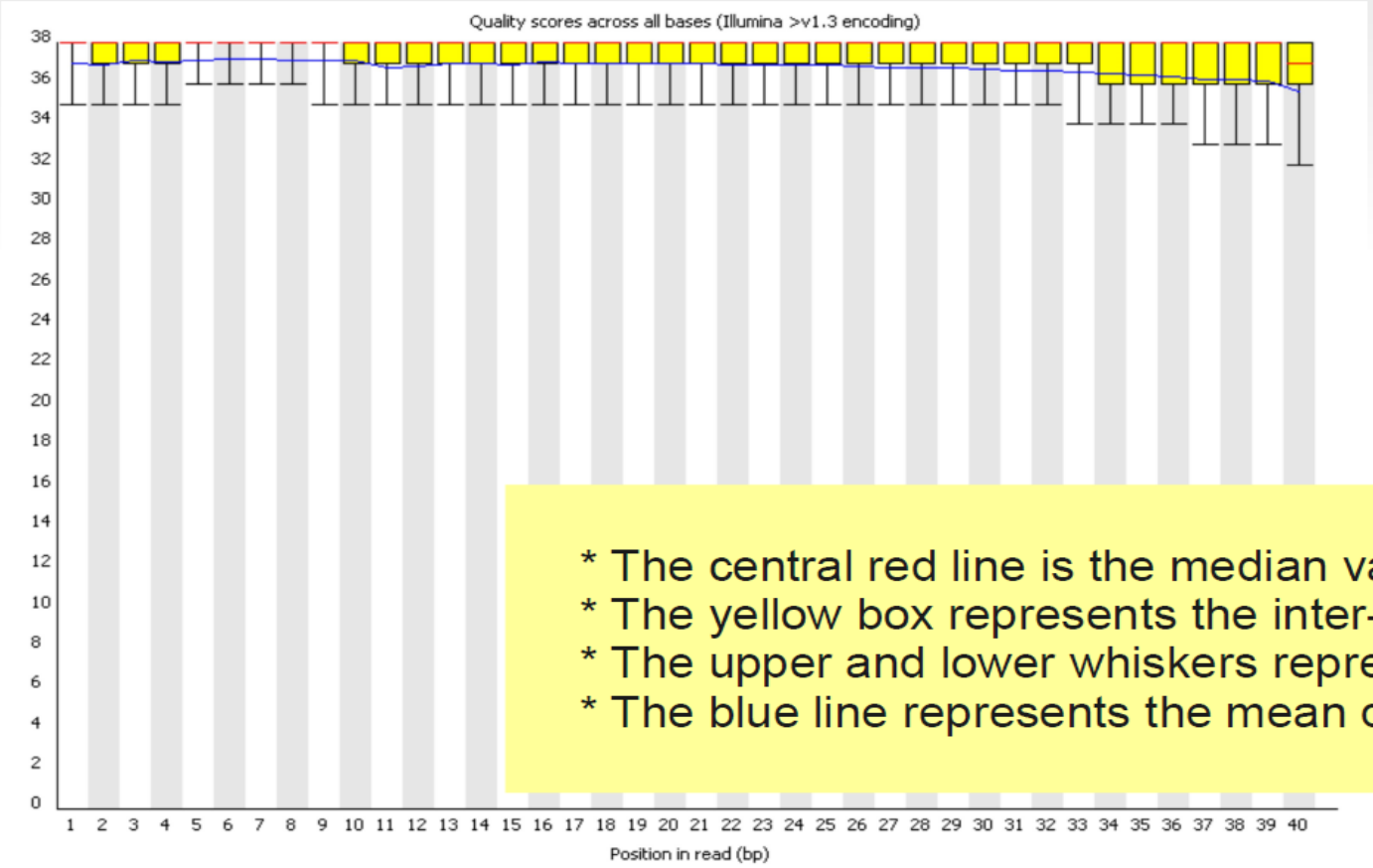
QC: FastQC


Per base sequence quality
Quality scores across all bases (Sanger / Illumina 1.9 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
15-19
25-29
35-39
45-49
55-59
65-69
75-79
85-89
95-99
Position in read (bp)

Tool for NGS Data QC: FASTQC



Sequence Quality Per base Position

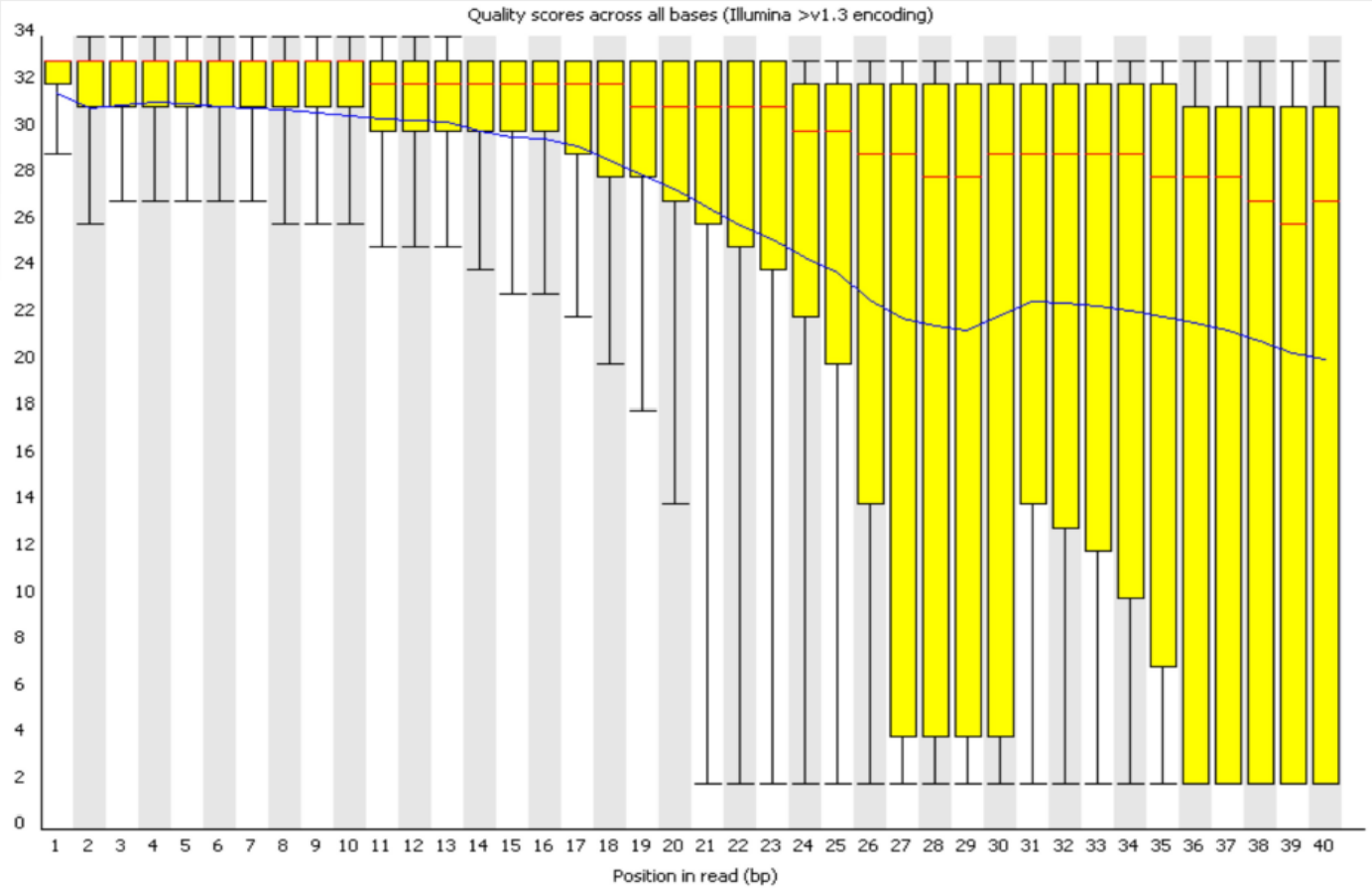


Good data

- Consistent
- High quality along the read

* The central red line is the median value
* The yellow box represents the inter-quartile range (25-75%)
* The upper and lower whiskers represent the 10% and 90% points
* The blue line represents the mean quality

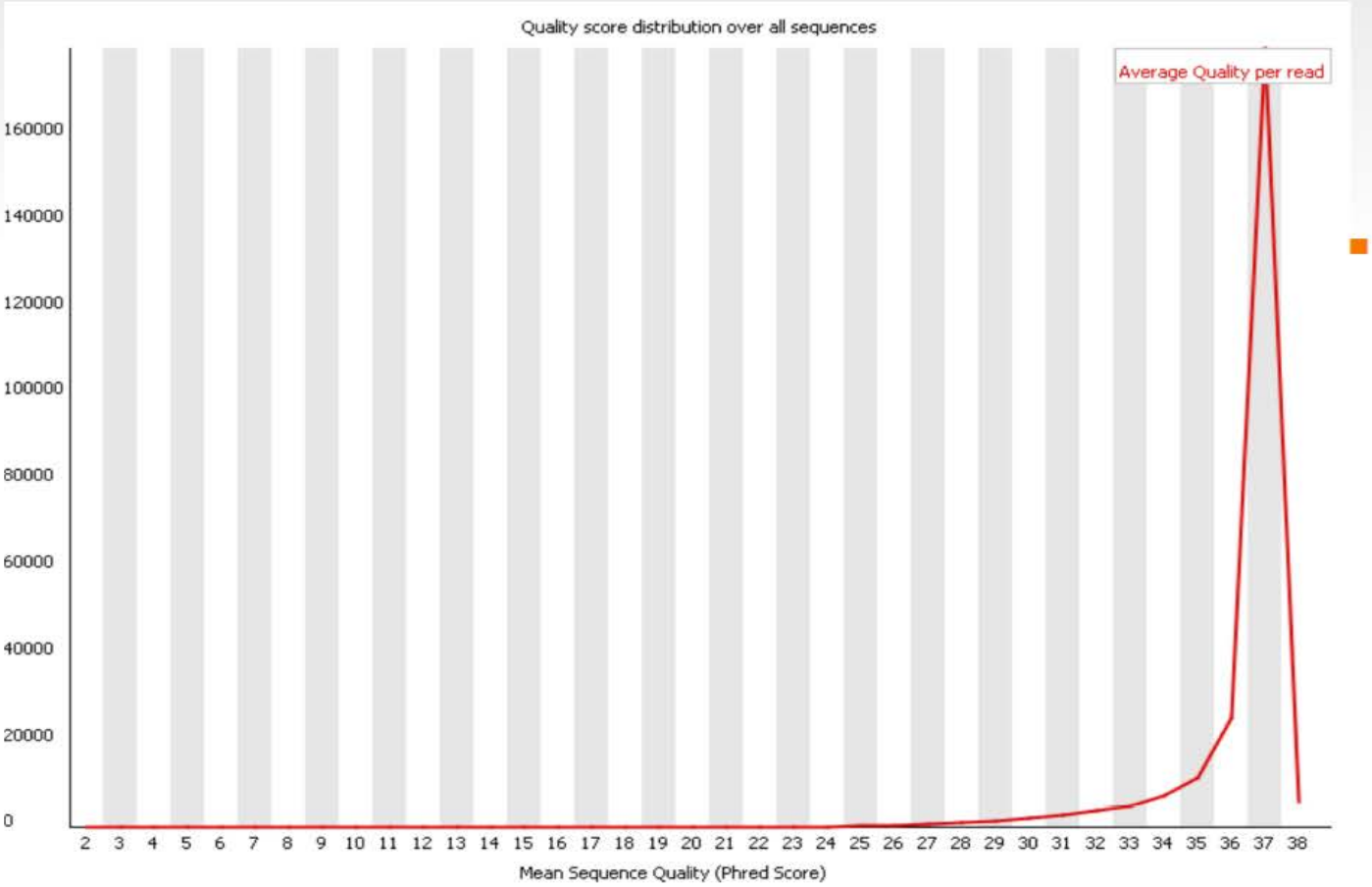
Sequence Quality Per base Position



Bad data

- High variance
- Quality decrease with length

Per Sequence Quality Distribution



Good data

- Most are high-quality sequences

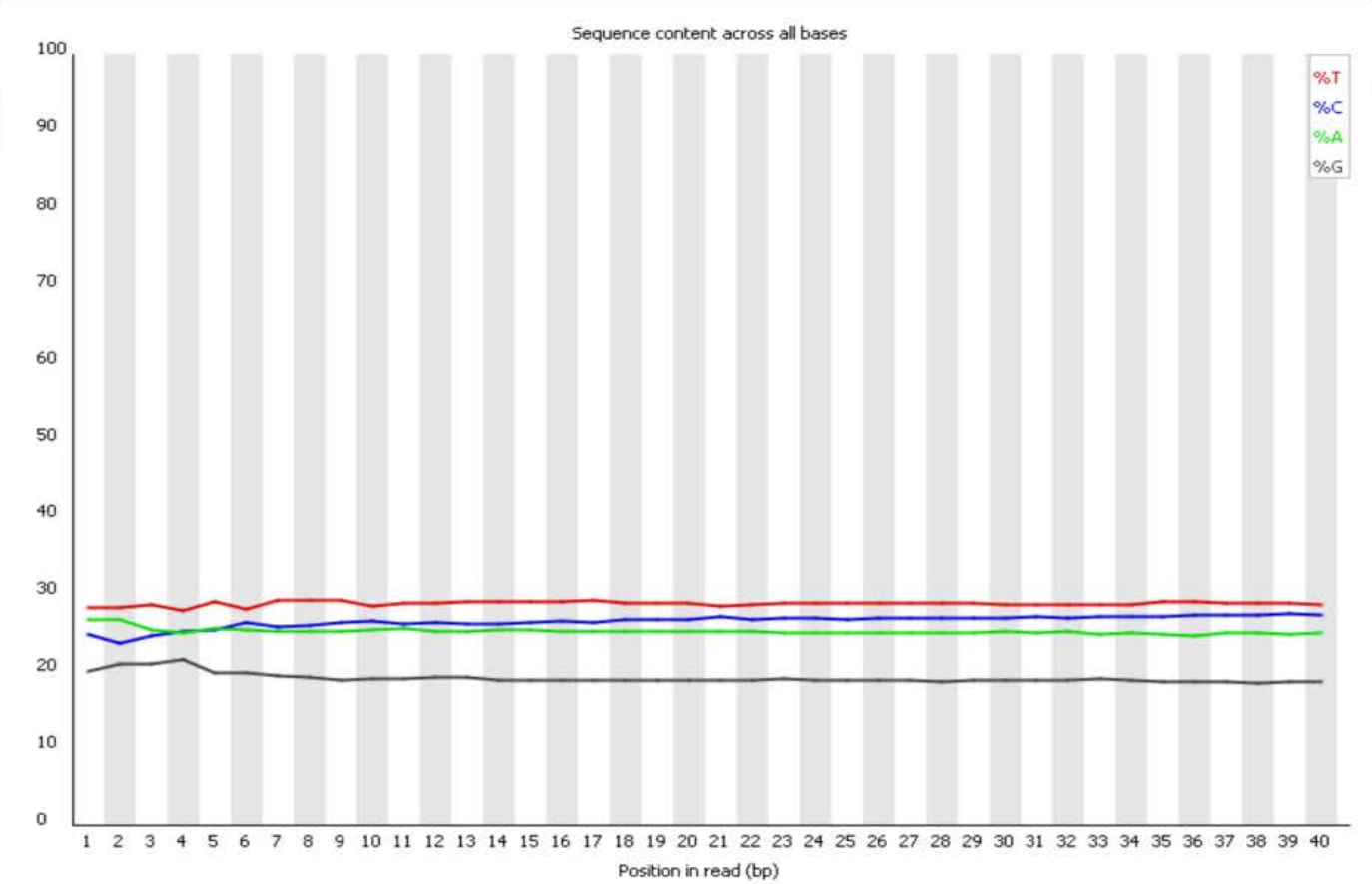
Per Sequence Quality Distribution



Bad data

- Not uniform distribution

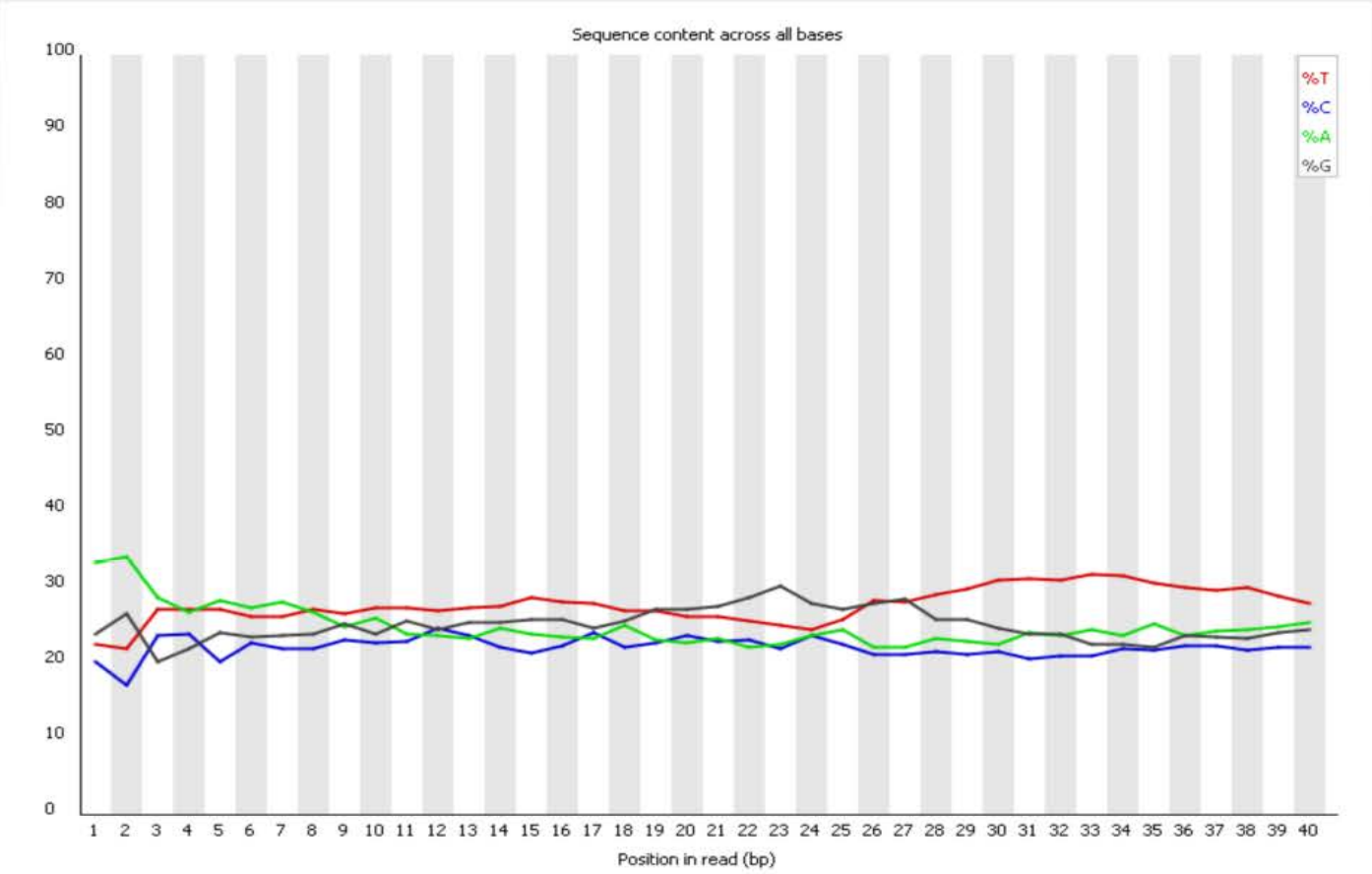
Nucleotide Content per position



Good data

- Smooth over length
- Organism dependent (GC)

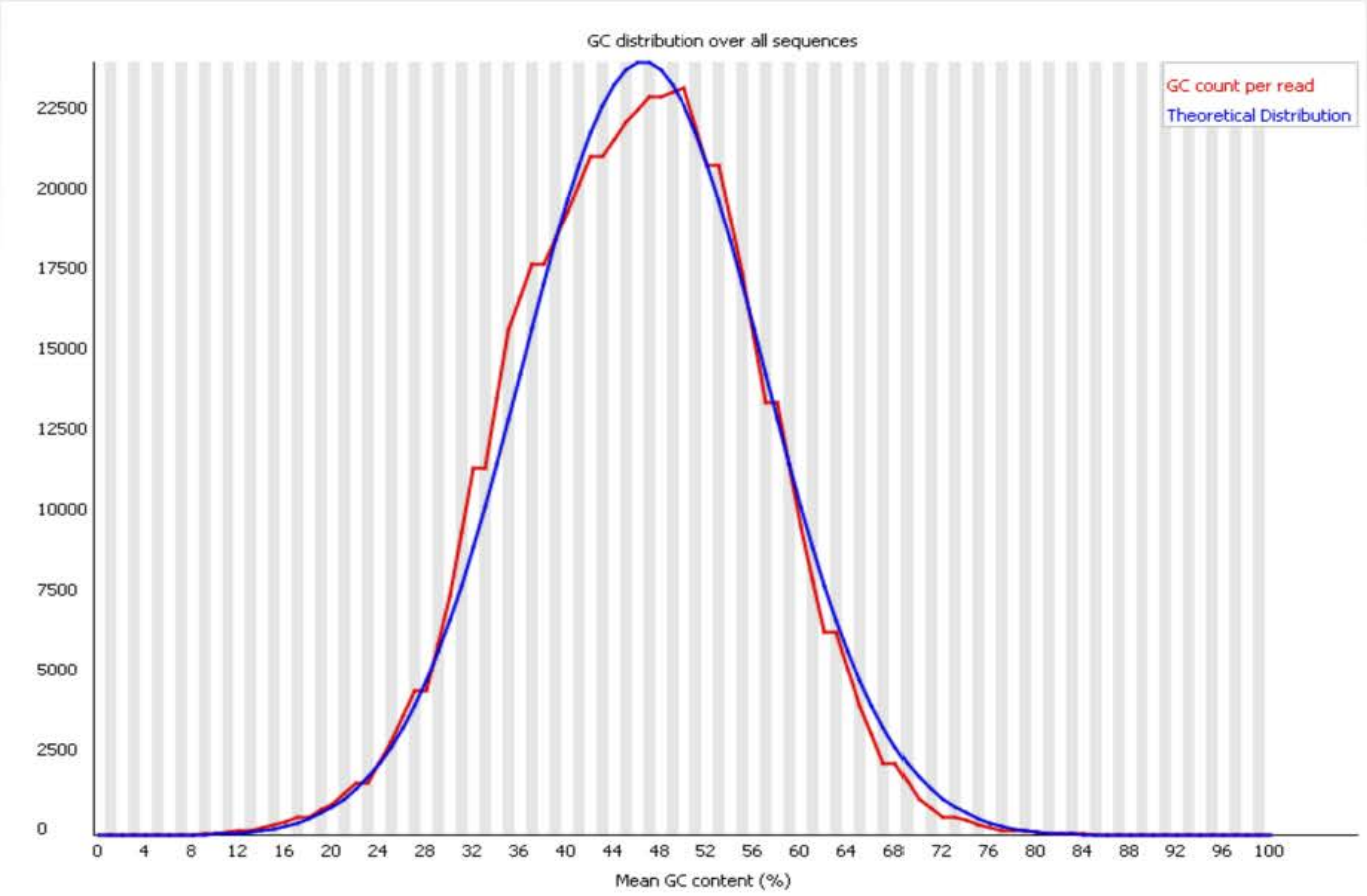
Nucleotide Content per position



Bad data

- Sequence position bias

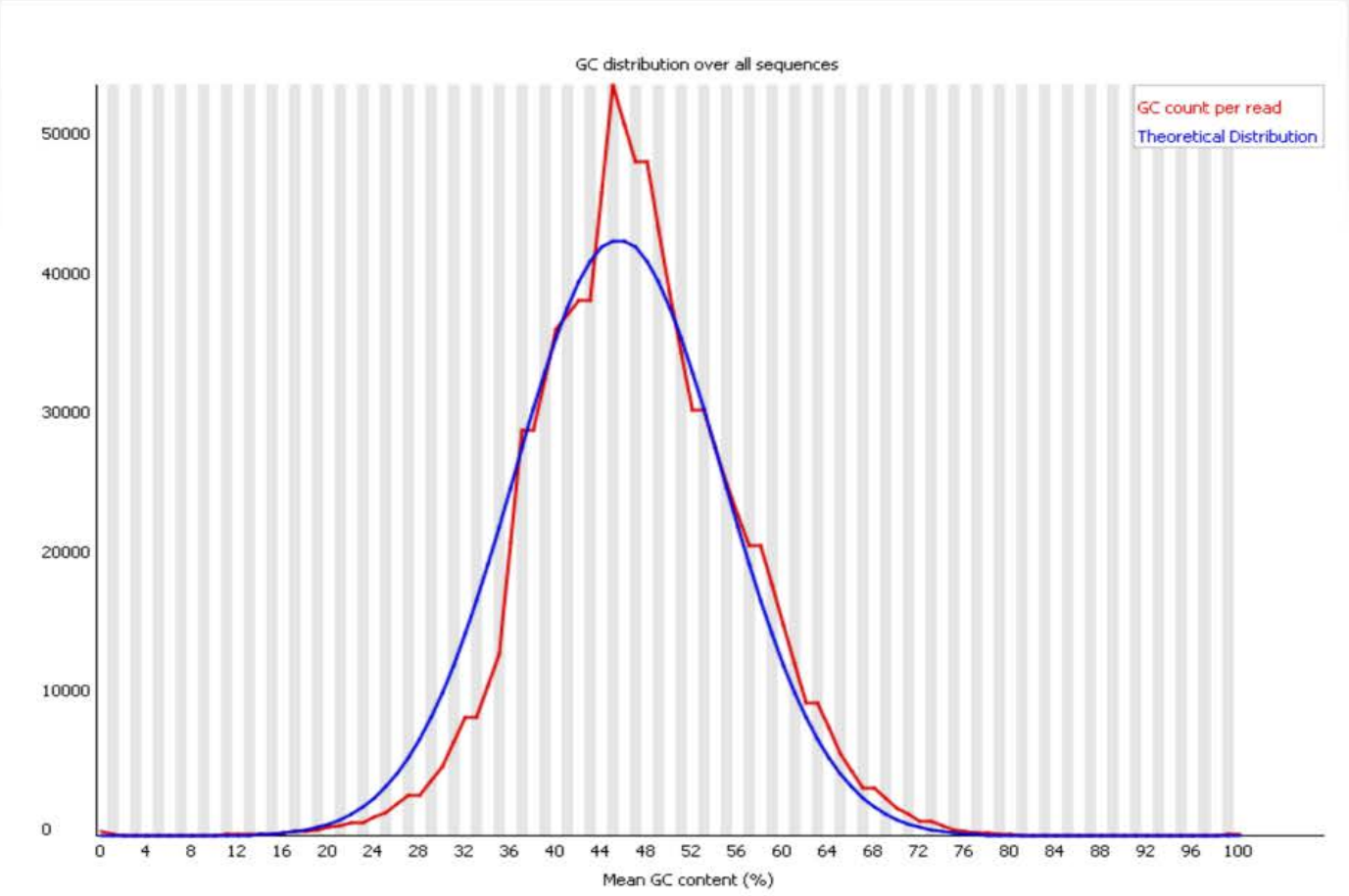
GC Distribution



Good data

- Fits with the expected
- Organism dependent

Per sequence GC Distribution

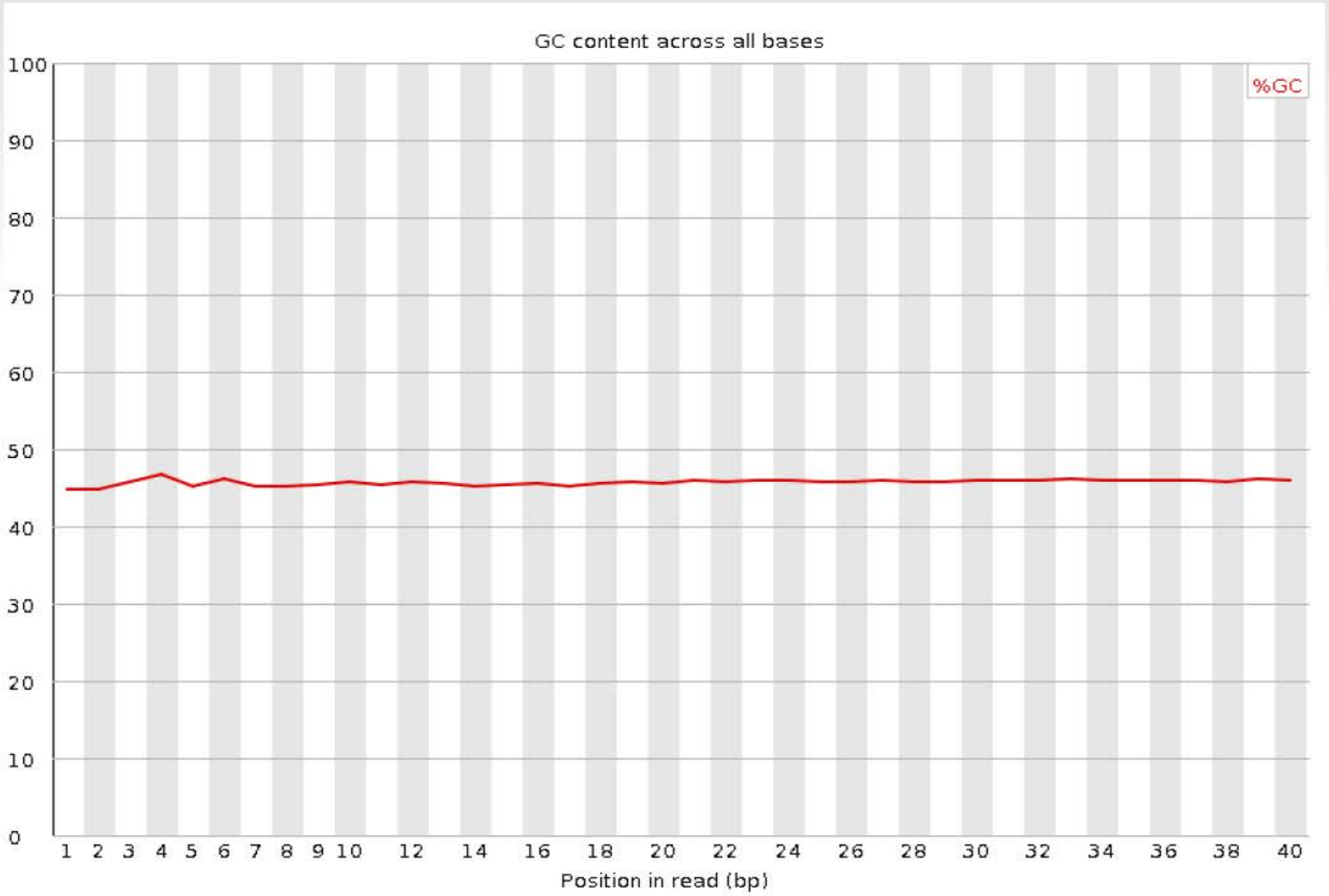


Bad data

- It does not fit with expected
- Organism dependent

Library contamination?

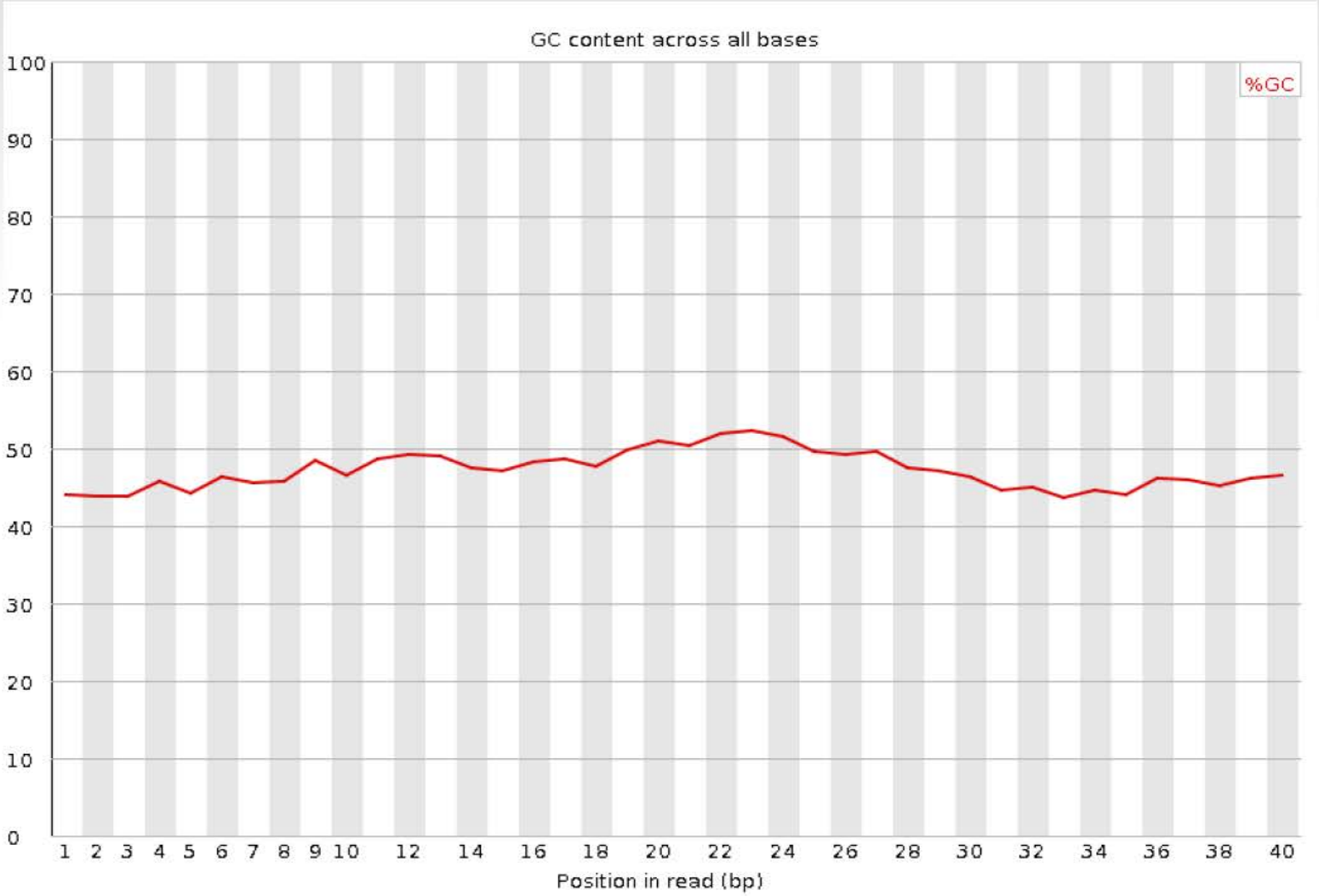
Per base GC Distribution



Good data

- No variation across read sequence

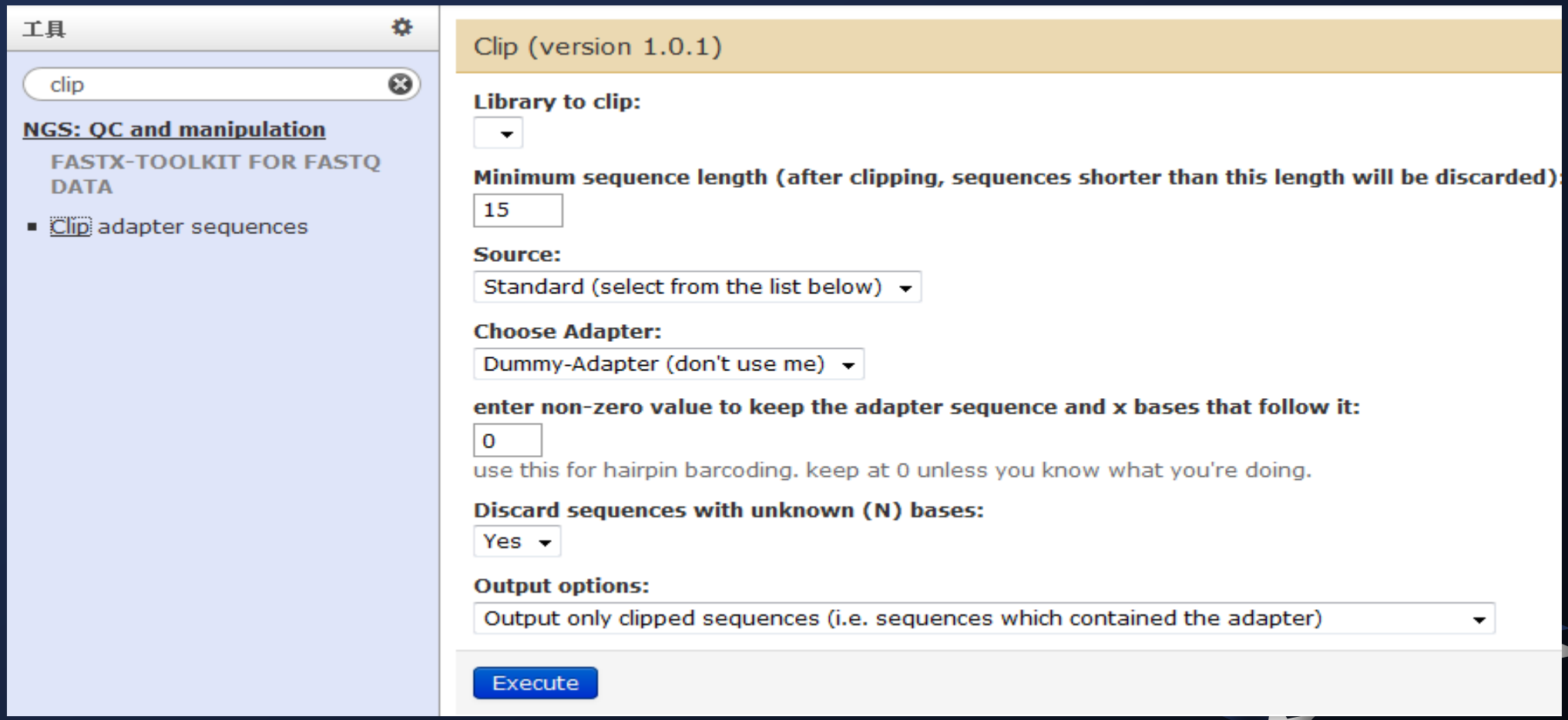
Per base GC Distribution



Bad data

- Variation across read sequence

Trim Adaptor (Fastx-tools)



Mapping Algorithms

BWA

- Short reads up to 200bp with error $< 5\,\%$
- Do not account for read quality
- Gapped alignment

Bowtie

- Very fast for short reads
- Does not align with gaps
- Use the read quality to evaluate the alignment

Tophat

- Improved Bowtie with gap alignment

Other

- ELAND (Illumina software), SOAP, MAQ, etc.

Map to Genome (Bowtie)

Map with Bowtie for Illumina (version 1.1.2)

Will you select a reference genome from your history or use a built-in index?:

Use a built-in index

Built-ins were indexed using default options

Select a reference genome:

Internal Control (YourGene Company, Klebsiella pneumoniae subsp. pneumoniae NTUH-K2044 DNA)

if your genome of interest is not listed - contact Galaxy team

Is this library mate-paired?:

Single-end

FASTQ file:

4: Select first on data 3

Must have ASCII encoded quality scores

Bowtie settings to use:

Commonly used

For most mapping needs use Commonly used settings. If you want full control use Full parameter list

Suppress the header in the output SAM file:

Bowtie produces SAM with several lines of header information by default

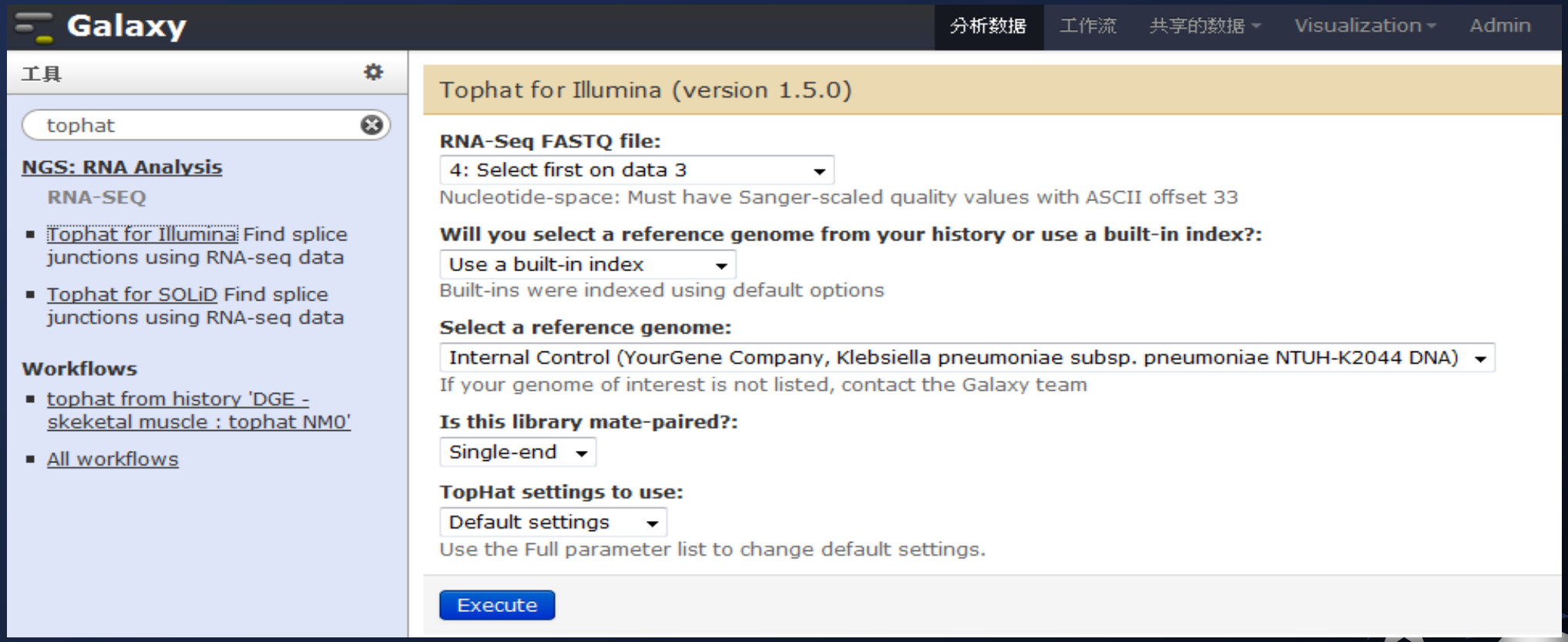
Execute

Map to Genome (Bowtie)

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	MRNM	MPOS	ISIZE	SEQ
@HD VN:1.0 SO:unsorted									
@SQ SN:chr1 LN:249250621									
@SQ SN:chr2 LN:243199373									
@SQ SN:chr3 LN:198022430									
@SQ SN:chr4 LN:191154276									
@SQ SN:chr5 LN:180915260									
@SQ SN:chr6 LN:171115067									
@SQ SN:chr7 LN:159138663									
@SQ SN:chr8 LN:146364022									
@SQ SN:chr9 LN:141213431									
@SQ SN:chr10 LN:135534747									
@SQ SN:chr11 LN:135006516									
@SQ SN:chr12 LN:133851895									
@SQ SN:chr13 LN:115169878									
@SQ SN:chr14 LN:107349540									
@SQ SN:chr15 LN:102531392									
@SQ SN:chr16 LN:90354753									
@SQ SN:chr17 LN:81195210									
@SQ SN:chr18 LN:78077248									
@SQ SN:chr19 LN:59128983									
@SQ SN:chr20 LN:63025520									
@SQ SN:chr21 LN:48129895									
@SQ SN:chr22 LN:51304566									
@SQ SN:chrX LN:155270560									
@SQ SN:chrY LN:59373566									
@SQ SN:chrM LN:16571									
@PG ID:Bowtie VN:0.12.7 CL:"bowtie -q -p 4 -S --phred33-quals /mnt/das4/lunchbox/galaxy_data/bowtie_index/hg19/hg19 /mnt/galaxy_dist/database/files/005/dataset_5001.d									
ERR030888.1	16	chr12	56668626	255	40M	*	0	0	CTTCCTTATGCAGCTGTGTTAGCCAGACAAGC
ERR030888.3	16	chr19	39369666	255	40M	*	0	0	GGGAGCTGTANGAGATTGGGGGATGTTCTGAG
ERR030888.6	16	chr6	139232438	255	59M	*	0	0	GTCTTTTCTTCANNTTTACTATCTGTTTTGGCAA
ERR030888.7	16	chr2	136956675	255	40M	*	0	0	TCACTAGTCTTAGAAAAACTTTCCAGAATACTG
ERR030888.9	16	chrM	11200	255	39M	*	0	0	CATACTTCCTATTCTACACACTAGTAGGCTCC(
ERR030888.11	16	chr6	99365329	255	35M	*	0	0	TCCATNCCACAACCATCCTTTTCTGCATAGGC
ERR030888.10	0	chr5	180668580	255	40M	*	0	0	GCGAAGGCCACACTCAGCACATCCTGGGTAT(
ERR030888.15 HWI-BRUNOP16X_0001:1:1:5506:1018#0 length=75	4	*	0	0	*	*	0	0	GTCGGACGCGCGGCTGCGCCAGGTTATTCAA(
ERR030888.17	16	chr6	32412617	255	58M	*	0	0	CTGTCTATTGCNNTTTCCTGTATCTATTTTCCT(
ERR030888.19	0	chr11	62327251	255	40M	*	0	0	CCAGGATCCAGTTTCCGCCATGTGTATGACTC
ERR030888.43	0	chr11	16996278	255	35M	*	0	0	CTCCCAGTTCAAAATGCTTGCATCTTTTAATAG
ERR030888.37	0	chr22	39713537	255	75M	*	0	0	GCCTCGAGGGGTTTCCACGTAGCCCACAATG(
ERR030888.45	0	chr5	141692297	255	39M	*	0	0	TATGCATGCAACAAGTACCTACCCACTGGACG
ERR030888.47	16	chr13	96271884	255	40M	*	0	0	ACAACTTGACCAAAAATTTGTCACAGAATTTTG
ERR030888.46	16	chr10	114836019	255	60M	*	0	0	CCTGAGCGGCTNGNNCTGTNGGTGTGAATTAC
ERR030888.50	16	chr4	91759888	255	40M	*	0	0	AGGCGCGGCCNTCGTTGTCAGTAGTTCTTTGA
ERR030888.49	16	chr16	33963662	255	40M	*	0	0	GGTTCCGGGGNGAGTATGGTTGCAAAGCTGAA
ERR030888.52	16	chr12	56553800	255	45M	*	0	0	ATGCTGCAGACAGTGGCCAAGAACACGGACC
ERR030888.53	0	chr10	22615385	255	60M	*	0	0	CAAACAACGAGAATCAAGATCACTGAGCTAAA
ERR030888.61 HWI-BRUNOP16X_0001:1:1:18792:1018#0 length=75	4	*	0	0	*	*	0	0	GTACTATGGGAGGTGGAGACGGAACACACGG(
ERR030888.41	16	chr21	15246346	255	60M	*	0	0	GTGACAATGAGATNNTCCATATGGCAGGAGTA
ERR030888.54	0	chr21	28290289	255	75M	*	0	0	GCTTAATTTTACAGTTTAAAGTCACACATAGGG



Map to Genome (TopHat)



SAM/BAM Format

SAM file example

Header

```
@HD     VN:1.0
@SQ     SN:chr20 LN:62435964
@RG     ID:L1 PU:SC_1_10 LB:SC_1 SM:NA12891
@RG     ID:L2 PU:SC_2_12 LB:SC_2 SM:NA12891
```

Alignment

```
read_28833_29006_6945 99 chr20 28833 20 10M1D25M = 28993 195 \
AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG
<<<<<<<<<<<<<<<<<<<:<9/,&,22;;<<< \
NM:i:1 RG:Z:L1
read_28701_28881_323b 147 chr20 28834 30 35M = 28701 -168 \
ACCTATATCTTGGCCTTGGCCGATGCGGCCTTGCA
<<<<<;<<<<7;:<<<6;<<<<<<<<<<<7<<<< \
MF:i:18 RG:Z:L2
```

Data in SAM/BAM Format

Table 1. Mandatory fields in the SAM format

No.	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	Bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-Based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	Extended CIGAR string (operations: MIDNSHP)
7	MRNM	Mate Reference NaMe ('=' if same as RNAME)
8	MPOS	1-Based leftmost Mate POSition
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQuence on the same strand as the reference
11	QUAL	Query QUALity (ASCII-33=Phred base quality)

Normalization (RPKM)

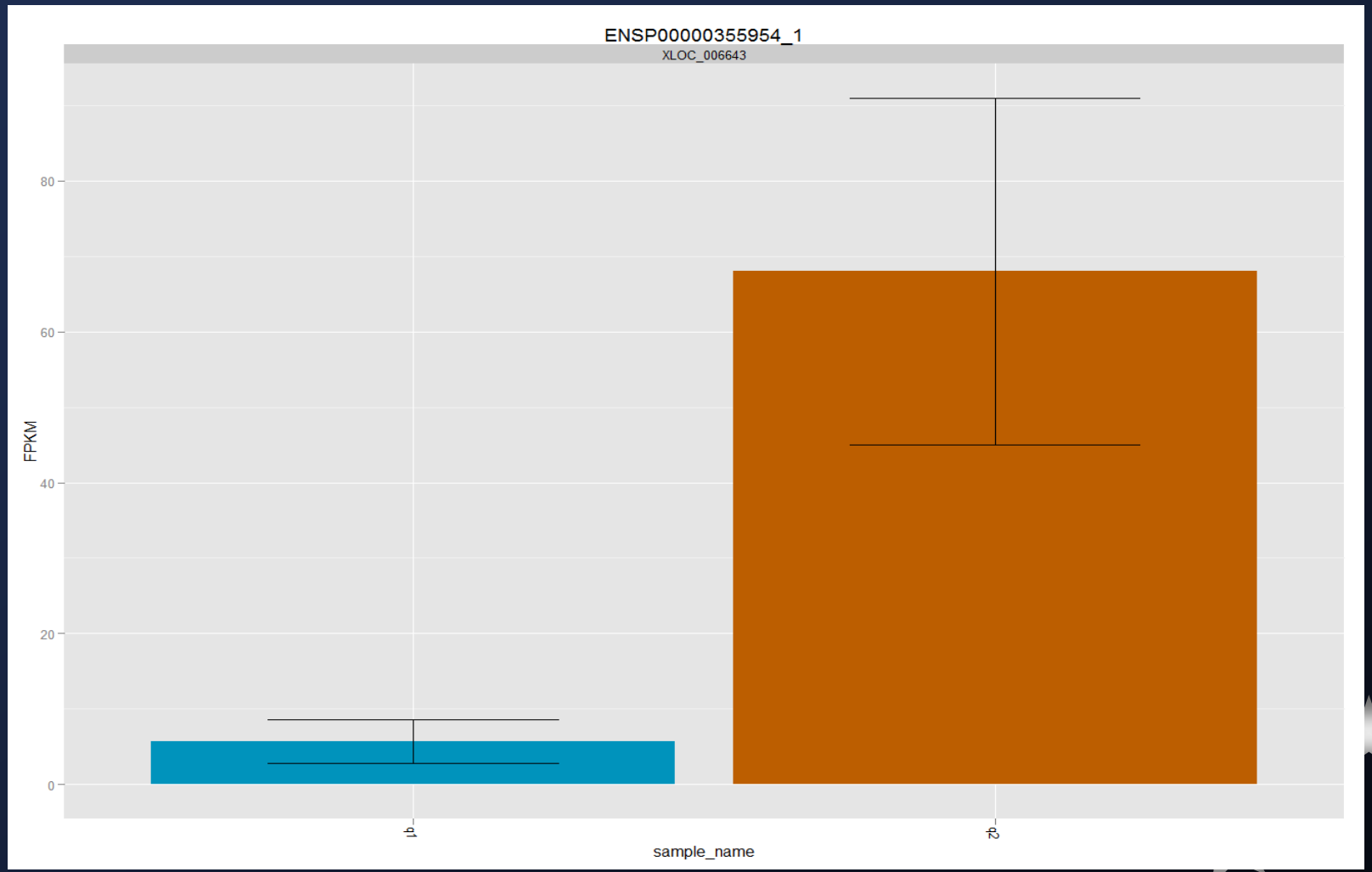
RPKM

Compute the RPKM (Reads per kilobase per million reads) value for a gene in a given experiment and compare your results with those obtained by `cufflinks`

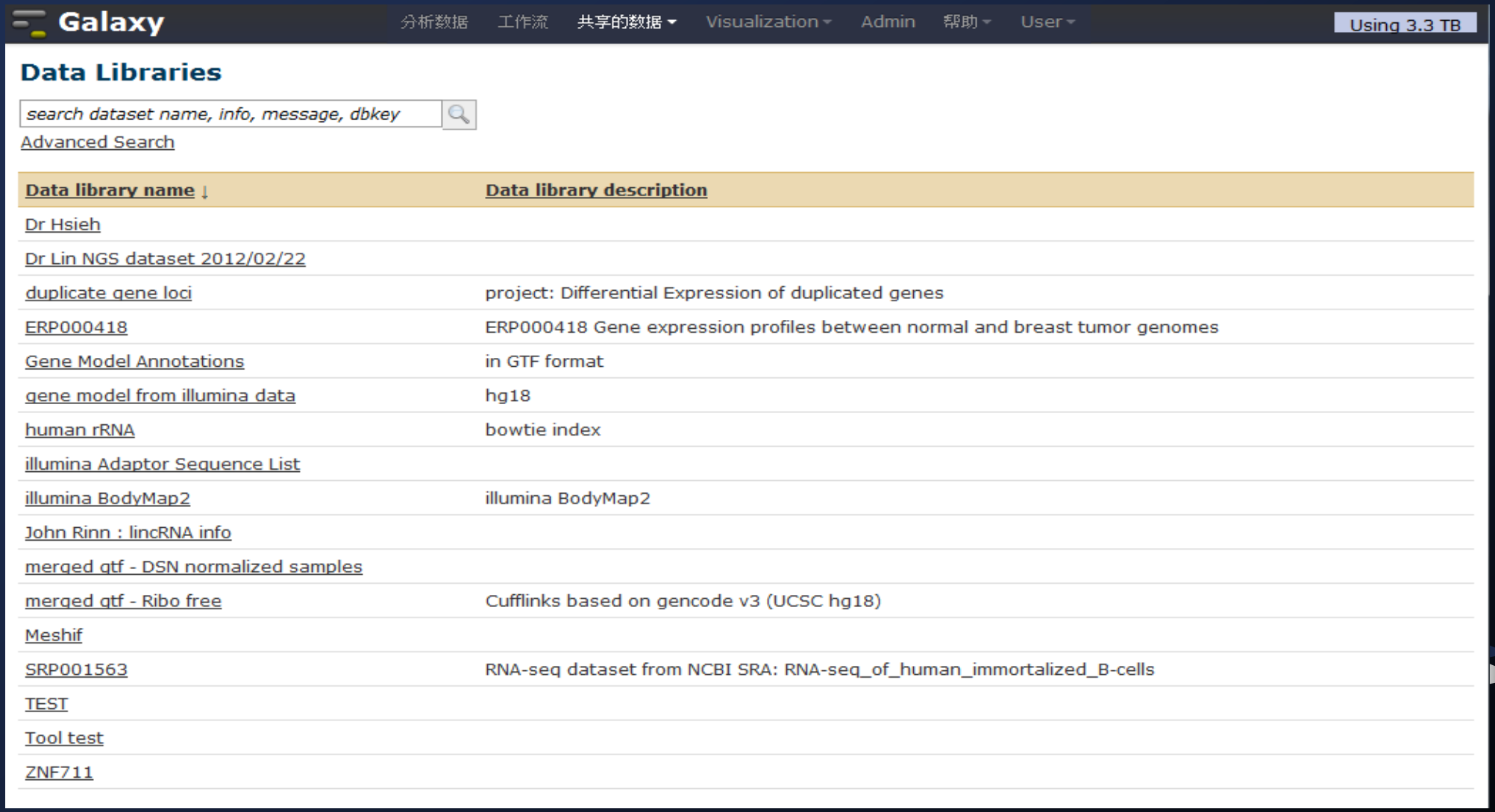
$$RPKM = \frac{\text{number of reads of the region}}{\frac{\text{total reads}}{1000000} \times \frac{\text{region length}}{1000}}$$



Examples of Differentially Expressed Genes (activating transcription factor 3)



Data Management



Data Management

