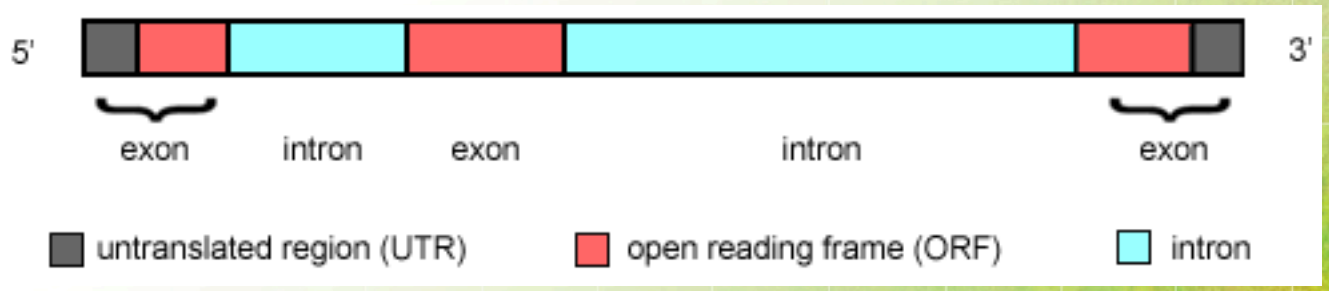
Annotation of eukaryote genomes and transciptomes



Henrik Lantz, BILS/SciLifeLab

Enabler for Life Sciences

Lecture synopsis

- What is annotation?
- Structural genome annotation
- Types of data used
- Transcriptome annotation
- Functional annotation

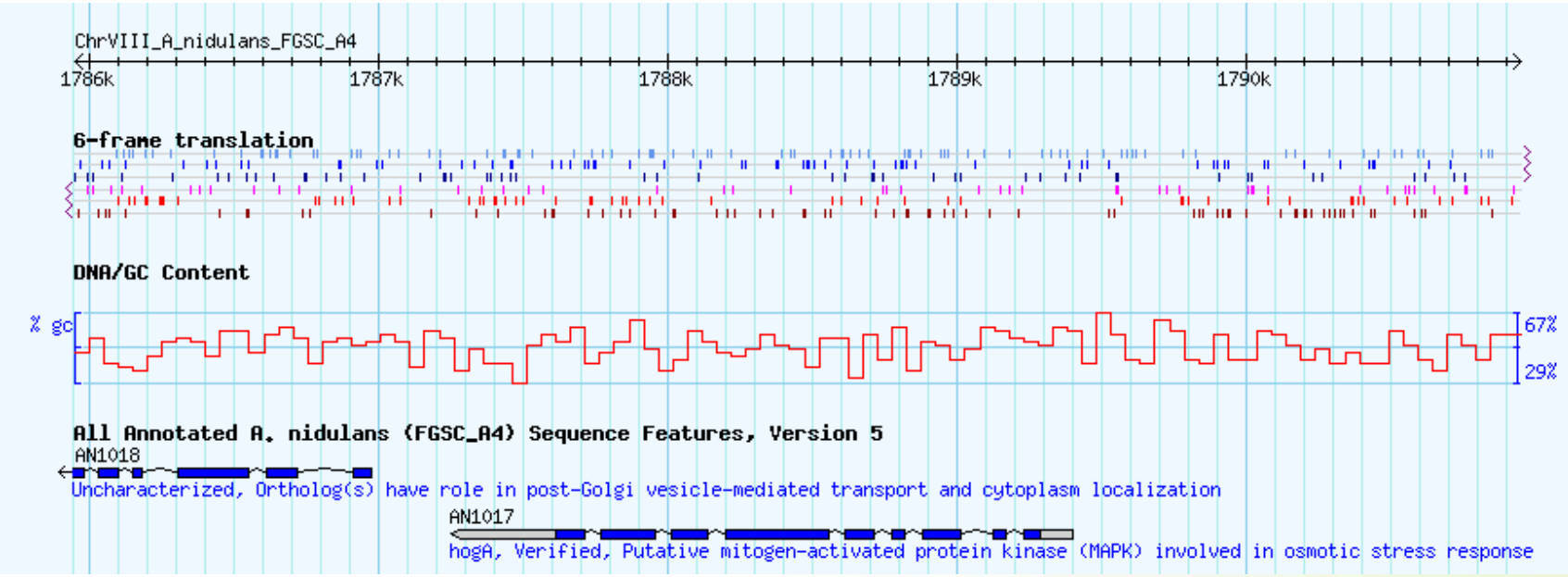
What is annotation?

- Identification of regions of interest in sequence data

From a genome…

```
>scaffold_26
AGTCACACACCCTTCAGCTTACACCCTGACTGCAGCCCTTACTCAAAACA
TTCCAGCCAGGAAGATGCTCCGACACAGCTTCTGGATGCCGCTCCTCGAC
GTCGAACGGCCCGCGCCGGGAAAATCGGCAGCGTCGGTGACCGCGGAGAT
CCGAAGCCGCCTCGGGGACCTGCGAGACAACGGGAGGCGGTCAACGAGAC
GCCGAGGGCTGGGAGTTATTCCCACACCGGGCCCGTAAGTTTTCTACCCA
AAAACCCATAGAAAAGAGATGAACCACTAAGTTTGATAACTCTTCTACTT
AACCGTGACCCTACGTGCCGGGGCAGGGCAGCTCTGACCCTAAGCGGCAC
ACGAACAAGGTGGTGCGCCCAATATAAACAAAGATGATGCAAGGGCTTGA
AATAAATCTCCGGAAGATTAATTCTCGAGCCCGACACGCTTTGAGGCAGC
GGAACCTACAGAACCACCGCAGTCACGTGAGAAGAGTCTAATACTCTCCA
AAGAGAAGTCCAAGGGAATGGAACGTGAAAAGAAGGTGCTTATCAAAAGC
GAGAAGGAAGATGGATGAGAACATCTTGTGTACTTCTTCTGGTCTCAAAA
AGCAAAAATGTAAAGATGCCAGACTAAGCCCGATCTGAGAAAGTACGCGA
GCAGAGACCCCCGCTGCCGATGTGGCCCAGAACGATGCCGATAAAGCACC
GAGACATAACAAAGCCCTGTGACACACAAGACGATGGACACAAACTACAT
AACACAGACACAAACTAAATGACACAGAGAGAAGTTGAAACTTCTGGGGA
AGTAAACATTTCTGAAACATCTACCAACAATCCGTCCATATATATTTCCA
TTCCACGGGACTCTTGGTTTGATATATGCGTGTTAACAGTAATCCCCGCT
GTAGCAATCACCACTATGCATAATTCATTAATTCTTTGGAGTTGCTGAGT
ATCATCTTATCAGTCTTATTTTTTTCCTTGGCTCTGGTTTCGGGCTTTTT
TTTTTTCTTCTGATAAGATTTTCCAGGAATGTGAAGACCCCCTGCATCCT
TCCCAAACTGACCACCCAAACTACAGACATTCTATAGCATTACATTACAC
AACCTAGGCAAAGTTTTTCTAACATTAAGGAACATGAAAAAAGCCAACAT
CACAATATATTCATAACAATTATGGAACATGCGAAAAGCCAATACCACAG
TACATTTATAACAATACCTCCCTTTTCCTTTCTTTAGAGATCATATGGCT
TGACCGCCGCCTCCTCGCCCGCCACCGCTGAGTACTGCCGTGCCGGAGTC
ACGGAGCCAGTCCCCCGCGGCCCCACCGCCTCCTCGCCCGCCGCCACGGA
GATCGGCTGCGCCACTCCCGAGCTCGGCCGTGCCATCGCCGCCCCCGCCG
GGGTCCCCCGGCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
Ntggcacagcccaggtgcccttggctttgcccacctgggcacacccagct
cgtgtccagctgctgtccccagcacccccagggcctttcccagttttcca
gcccctcttcccccagcctggagtgttccacggtgagccaagggaaggac
caggcacttggccttgttgaccctcacactgctggtctcaccccatggat
ccagcctgtcccgatccctctgcagaacctccctgccctccagcagatcc
acatcccacctgactcgatgccatctgtgaatttaccgagggtgccctca
atcccctaatccagatcatcaataaagatattgaacaggactgggcccag
cgctgatccctgggggacaccaggctgctcccaactgggtacggcaccat
tcccacccctccctgggctggccatcagccagtttccaaccctgcccaag
tgaccctgtccacgggctgcagcttttccagCACAGCAGCACCCTGGGAG
TGCTGGACTGAGGCACGGAGGCAAAATTCCATCCCAGAGAATAGCCTGAA
AAAAGGAAAAGCATCGGGAAAAGCCACAGAGGATGCTCAGAAGAAACAAC
GCAGCTCCTGGGCCTGGCGCAGGGACAGGGAAATGTGTAACTCCAGCCCA
GAGAACACGTGCTGCTAGAAATACCAACATCAGAACAGAAAACTGGATCC
GCAGAGAGAATTAACTCCCTGCACATGTGAAACAGAATATTTGGGAGAGG
TAAGAAAGCTCAGGCGAATCCCTGCCGTGTCCAAATGCTCTGCTAGAGCA
GGCTGCACCTCGGCAGCATGGCTGTCAGCACAGGGATCAGCAATCAGCGC
TTGGATGCATCCGTCCATGCCGGGAAGAGCAGGAATCCGCTGAGCGGCAG
CCAGGGCTCCCCTGCTGCAGGCGCTGCTGAGTGGGACAAGGACATTTTTG
```

…to an annotated gene



GFF file format

```
##gff-version 3
scaffold_7	maker	gene	133848	144662	.	-	.	ID=C55462A8A38E2878A71E2ABDAFB34661;Name=maker-scaffold_7-augustus-gene-0.11
scaffold_7	maker	mRNA	133848	144662	.	-	.	ID=A649E923246BADE2184E579FA9124ABD;Parent=C55462A8A38E2878A71E2ABDAFB34661;Name=1:cornix-all_reads.72406.1;_AED=1.00;_eAED=1.00;_QI=71|0|0|0|0|0|8|414|347
scaffold_7	maker	exon	138974	139077	.	-	.	ID=A649E923246BADE2184E579FA9124ABD:exon:7;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7	maker	exon	135098	135281	.	-	.	ID=A649E923246BADE2184E579FA9124ABD:exon:6;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7	maker	exon	139616	139836	.	-	.	ID=A649E923246BADE2184E579FA9124ABD:exon:5;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7	maker	exon	144511	144662	.	-	.	ID=A649E923246BADE2184E579FA9124ABD:exon:4;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7	maker	exon	136342	136437	.	-	.	ID=A649E923246BADE2184E579FA9124ABD:exon:3;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7	maker	exon	133848	134338	.	-	.	ID=A649E923246BADE2184E579FA9124ABD:exon:2;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7	maker	exon	141262	141383	.	-	.	ID=A649E923246BADE2184E579FA9124ABD:exon:1;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7	maker	exon	144138	144296	.	-	.	ID=A649E923246BADE2184E579FA9124ABD:exon:0;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7	maker	five_prime_UTR	144592	144662	.	-	.	ID=A649E923246BADE2184E579FA9124ABD:five_prime_utr;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7	maker	CDS	144511	144591	.	-	0	ID=A649E923246BADE2184E579FA9124ABD:cds;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7	maker	CDS	144138	144296	.	-	0	ID=A649E923246BADE2184E579FA9124ABD:cds;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7	maker	CDS	141262	141383	.	-	0	ID=A649E923246BADE2184E579FA9124ABD:cds;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7	maker	CDS	139616	139836	.	-	1	ID=A649E923246BADE2184E579FA9124ABD:cds;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7	maker	CDS	138974	139077	.	-	2	ID=A649E923246BADE2184E579FA9124ABD:cds;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7	maker	CDS	136342	136437	.	-	0	ID=A649E923246BADE2184E579FA9124ABD:cds;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7	maker	CDS	135098	135281	.	-	0	ID=A649E923246BADE2184E579FA9124ABD:cds;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7	maker	CDS	134262	134338	.	-	2	ID=A649E923246BADE2184E579FA9124ABD:cds;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7	maker	three_prime_UTR	133848	134261	.	-	.	ID=A649E923246BADE2184E579FA9124ABD:three_prime_utr;Parent=A649E923246BADE2184E579FA9124ABD
scaffold_7	maker	gene	83101	117593	.	+	.	ID=D3BD9A5F27797F56A844A2E890FF6B99;Name=maker-scaffold_7-augustus-gene-0.6
scaffold_7	maker	mRNA	83101	117593	.	+	.	ID=CFF5DDA190832937C45A0D2E674C9C26;Parent=D3BD9A5F27797F56A844A2E890FF6B99;Name=maker-scaffold_7-augustus-gene-0.6-mRNA-1;_AED=1.00;_eAED=1.00;_QI=0|0|0|0|0|0|21|4|706
scaffold_7	maker	exon	95748	95871	.	+	.	ID=CFF5DDA190832937C45A0D2E674C9C26:exon:8;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	exon	99113	99137	.	+	.	ID=CFF5DDA190832937C45A0D2E674C9C26:exon:9;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	exon	90664	90748	.	+	.	ID=CFF5DDA190832937C45A0D2E674C9C26:exon:10;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	exon	110231	110356	.	+	.	ID=CFF5DDA190832937C45A0D2E674C9C26:exon:11;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	exon	113609	113679	.	+	.	ID=CFF5DDA190832937C45A0D2E674C9C26:exon:12;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	exon	94057	94117	.	+	.	ID=CFF5DDA190832937C45A0D2E674C9C26:exon:13;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	exon	84578	84670	.	+	.	ID=CFF5DDA190832937C45A0D2E674C9C26:exon:14;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	exon	115452	115536	.	+	.	ID=CFF5DDA190832937C45A0D2E674C9C26:exon:15;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	exon	111579	111669	.	+	.	ID=CFF5DDA190832937C45A0D2E674C9C26:exon:16;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	exon	102917	103016	.	+	.	ID=CFF5DDA190832937C45A0D2E674C9C26:exon:17;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	exon	96766	96849	.	+	.	ID=CFF5DDA190832937C45A0D2E674C9C26:exon:18;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	exon	86666	86750	.	+	.	ID=CFF5DDA190832937C45A0D2E674C9C26:exon:19;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	exon	99944	100109	.	+	.	ID=CFF5DDA190832937C45A0D2E674C9C26:exon:20;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	exon	109766	109860	.	+	.	ID=CFF5DDA190832937C45A0D2E674C9C26:exon:21;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	exon	93154	93282	.	+	.	ID=CFF5DDA190832937C45A0D2E674C9C26:exon:22;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	exon	114737	114825	.	+	.	ID=CFF5DDA190832937C45A0D2E674C9C26:exon:23;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	exon	83101	83155	.	+	.	ID=CFF5DDA190832937C45A0D2E674C9C26:exon:24;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	exon	108533	108795	.	+	.	ID=CFF5DDA190832937C45A0D2E674C9C26:exon:25;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	exon	117477	117593	.	+	.	ID=CFF5DDA190832937C45A0D2E674C9C26:exon:26;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	exon	106779	106866	.	+	.	ID=CFF5DDA190832937C45A0D2E674C9C26:exon:27;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	exon	105743	105835	.	+	.	ID=CFF5DDA190832937C45A0D2E674C9C26:exon:28;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	CDS	83101	83155	.	+	0	ID=CFF5DDA190832937C45A0D2E674C9C26:cds;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	CDS	84578	84670	.	+	2	ID=CFF5DDA190832937C45A0D2E674C9C26:cds;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	CDS	86666	86750	.	+	2	ID=CFF5DDA190832937C45A0D2E674C9C26:cds;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	CDS	90664	90748	.	+	1	ID=CFF5DDA190832937C45A0D2E674C9C26:cds;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	CDS	93154	93282	.	+	0	ID=CFF5DDA190832937C45A0D2E674C9C26:cds;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	CDS	94057	94117	.	+	0	ID=CFF5DDA190832937C45A0D2E674C9C26:cds;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	CDS	95748	95871	.	+	2	ID=CFF5DDA190832937C45A0D2E674C9C26:cds;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	CDS	96766	96849	.	+	1	ID=CFF5DDA190832937C45A0D2E674C9C26:cds;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	CDS	99113	99137	.	+	1	ID=CFF5DDA190832937C45A0D2E674C9C26:cds;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	CDS	99944	100109	.	+	0	ID=CFF5DDA190832937C45A0D2E674C9C26:cds;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	CDS	102917	103016	.	+	2	ID=CFF5DDA190832937C45A0D2E674C9C26:cds;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	CDS	105743	105835	.	+	1	ID=CFF5DDA190832937C45A0D2E674C9C26:cds;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	CDS	106779	106866	.	+	1	ID=CFF5DDA190832937C45A0D2E674C9C26:cds;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	CDS	108533	108795	.	+	0	ID=CFF5DDA190832937C45A0D2E674C9C26:cds;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	CDS	109766	109860	.	+	1	ID=CFF5DDA190832937C45A0D2E674C9C26:cds;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	CDS	110231	110356	.	+	2	ID=CFF5DDA190832937C45A0D2E674C9C26:cds;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	CDS	111579	111669	.	+	2	ID=CFF5DDA190832937C45A0D2E674C9C26:cds;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	CDS	113609	113679	.	+	1	ID=CFF5DDA190832937C45A0D2E674C9C26:cds;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	CDS	114737	114825	.	+	2	ID=CFF5DDA190832937C45A0D2E674C9C26:cds;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	CDS	115452	115536	.	+	0	ID=CFF5DDA190832937C45A0D2E674C9C26:cds;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	CDS	117477	117589	.	+	2	ID=CFF5DDA190832937C45A0D2E674C9C26:cds;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	three_prime_UTR	117590	117593	.	+	.	ID=CFF5DDA190832937C45A0D2E674C9C26:three_prime_utr;Parent=CFF5DDA190832937C45A0D2E674C9C26
scaffold_7	maker	gene	22451	37514	.	-	.	ID=E1E94A56071E5D7940C93BE70FD79E56;Name=E1E94A56071E5D7940C93BE70FD79E56
scaffold_7	maker	mRNA	22451	37514	.	-	.	ID=2BB3B792BBFCD8B1C7A4070E76500CC5;Parent=E1E94A56071E5D7940C93BE70FD79E56;Name=scaffold_7:22450-37514;_AED=1.00;_eAED=1.00;_QI=3008|0|0|0|0|0|11|762|414
scaffold_7	maker	mRNA	26810	37514	.	-	.	ID=1D629B10725FB97175585BA23623AFA6;Parent=E1E94A56071E5D7940C93BE70FD79E56;Name=maker-scaffold_7-augustus-gene-0.10-mRNA-1;_AED=1.00;_eAED=1.00;_QI=3008|0|0|0|0|0|10|0|400
scaffold_7	maker	exon	26810	27088	.	-	.	ID=1D629B10725FB97175585BA23623AFA6:exon:40;Parent=1D629B10725FB97175585BA23623AFA6
scaffold_7	maker	exon	27950	28090	.	-	.	ID=2BB3B792BBFCD8B1C7A4070E76500CC5:exon:39;Parent=2BB3B792BBFCD8B1C7A4070E76500CC5,1D629B10725FB97175585BA23623AFA6
scaffold_7	maker	exon	33513	33610	.	-	.	ID=2BB3B792BBFCD8B1C7A4070E76500CC5:exon:38;Parent=2BB3B792BBFCD8B1C7A4070E76500CC5,1D629B10725FB97175585BA23623AFA6
maker.gff
```

GFF3 file format

Seqid	source	type	start	end	score	strand	phase	attributes
Chr1	Snap	gene	234	3657	.	+	.	ID=gene1; Name=Snap1;
Chr1	Snap	mRNA	234	3657	.	+	.	ID=gene1.m1; Parent=gene1;
Chr1	Snap	exon	234	1543	.	+	.	ID=gene1.m1.exon1; Parent=gene1.m1;
Chr1	Snap	CDS	577	1543	.	+	0	ID=gene1.m1.CDS; Parent=gene1.m1;
Chr1	Snap	exon	1822	2674	.	+	.	ID=gene1.m1.exon2; Parent=gene1.m1;
Chr1	Snap	CDS	1822	2674	.	+	2	ID=gene1.m1.CDS; Parent=gene1.m1;
		start_ codon						Alias, note, ontology_term …
		stop_ codon						

GTF file format

```
Sb_20131119_contig_1    Cufflinks       transcript      1522    2095    1000    .       .       gene_id "CUFF.1"; transcript_id "CUFF.1.1"; FPKM "2.6064385494"; frac "1.000000"; conf_lo "0.948975"; conf_hi "3.440036"; cov "4.817376";
Sb_20131119_contig_1    Cufflinks       exon    1522    2095    1000    .       .       gene_id "CUFF.1"; transcript_id "CUFF.1.1"; exon_number "1"; FPKM "2.6064385494"; frac "1.000000"; conf_lo "0.948975"; conf_hi "3.440036"; cov "4.817376";
Sb_20131119_contig_1    Cufflinks       transcript      3626    4118    1000    .       .       gene_id "CUFF.3"; transcript_id "CUFF.3.1"; FPKM "3.1548106029"; frac "1.000000"; conf_lo "0.828669"; conf_hi "3.729011"; cov "8.517668";
Sb_20131119_contig_1    Cufflinks       exon    3626    4118    1000    .       .       gene_id "CUFF.3"; transcript_id "CUFF.3.1"; exon_number "1"; FPKM "3.1548106029"; frac "1.000000"; conf_lo "0.828669"; conf_hi "3.729011"; cov "8.517668";
Sb_20131119_contig_1    Cufflinks       transcript      4855    5340    1000    .       .       gene_id "CUFF.5"; transcript_id "CUFF.5.1"; FPKM "7.0235237898"; frac "1.000000"; conf_lo "2.521814"; conf_hi "6.164435"; cov "16.171080";
Sb_20131119_contig_1    Cufflinks       exon    4855    5340    1000    .       .       gene_id "CUFF.5"; transcript_id "CUFF.5.1"; exon_number "1"; FPKM "7.0235237898"; frac "1.000000"; conf_lo "2.521814"; conf_hi "6.164435"; cov "16.171080";
Sb_20131119_contig_1    Cufflinks       transcript      5398    5975    1000    .       .       gene_id "CUFF.4"; transcript_id "CUFF.4.1"; FPKM "3.1706609980"; frac "1.000000"; conf_lo "1.178010"; conf_hi "3.651831"; cov "4.764328";
Sb_20131119_contig_1    Cufflinks       exon    5398    5975    1000    .       .       gene_id "CUFF.4"; transcript_id "CUFF.4.1"; exon_number "1"; FPKM "3.1706609980"; frac "1.000000"; conf_lo "1.178010"; conf_hi "3.651831"; cov "4.764328";
Sb_20131119_contig_10   Cufflinks       transcript      954     2795    1000    .       .       gene_id "CUFF.6"; transcript_id "CUFF.6.1"; FPKM "6.8195889357"; frac "1.000000"; conf_lo "5.175059"; conf_hi "7.577765"; cov "16.605534";
Sb_20131119_contig_10   Cufflinks       exon    954     2795    1000    .       .       gene_id "CUFF.6"; transcript_id "CUFF.6.1"; exon_number "1"; FPKM "6.8195889357"; frac "1.000000"; conf_lo "5.175059"; conf_hi "7.577765"; cov "16.605534";
Sb_20131119_contig_1    Cufflinks       transcript      4502    4718    1000    .       .       gene_id "CUFF.2"; transcript_id "CUFF.2.1"; FPKM "37.5296486924"; frac "1.000000"; conf_lo "2.510193"; conf_hi "9.099450"; cov "60.418389";
Sb_20131119_contig_1    Cufflinks       exon    4502    4718    1000    .       .       gene_id "CUFF.2"; transcript_id "CUFF.2.1"; exon_number "1"; FPKM "37.5296486924"; frac "1.000000"; conf_lo "2.510193"; conf_hi "9.099450"; cov "60.418389";
Sb_20131119_contig_1    Cufflinks       transcript      10522   13208   1000    .       .       gene_id "CUFF.23"; transcript_id "CUFF.23.1"; FPKM "55.6377793473"; frac "1.000000"; conf_lo "48.931832"; conf_hi "55.241530"; cov "121.429110";
Sb_20131119_contig_1    Cufflinks       exon    10522   13208   1000    .       .       gene_id "CUFF.23"; transcript_id "CUFF.23.1"; exon_number "1"; FPKM "55.6377793473"; frac "1.000000"; conf_lo "48.931832"; conf_hi "55.241530"; cov "121.429110";
Sb_20131119_contig_1    Cufflinks       transcript      13270   14623   1000    .       .       gene_id "CUFF.7"; transcript_id "CUFF.7.1"; FPKM "41.2374406123"; frac "1.000000"; conf_lo "31.982715"; conf_hi "39.274371"; cov "89.788421";
Sb_20131119_contig_1    Cufflinks       exon    13270   14623   1000    .       .       gene_id "CUFF.7"; transcript_id "CUFF.7.1"; exon_number "1"; FPKM "41.2374406123"; frac "1.000000"; conf_lo "31.982715"; conf_hi "39.274371"; cov "89.788421";
Sb_20131119_contig_100022       Cufflinks       transcript      3991    4547    1000    .       .       gene_id "CUFF.54"; transcript_id "CUFF.54.1"; FPKM "52.7578565123"; frac "1.000000"; conf_lo "27.382285"; conf_hi "37.895127"; cov "66.397320";
Sb_20131119_contig_100022       Cufflinks       exon    3991    4547    1000    .       .       gene_id "CUFF.54"; transcript_id "CUFF.54.1"; exon_number "1"; FPKM "52.7578565123"; frac "1.000000"; conf_lo "27.382285"; conf_hi "37.895127"; cov "66.397320";
Sb_20131119_contig_100023       Cufflinks       transcript      1097    2009    1000    .       .       gene_id "CUFF.9"; transcript_id "CUFF.9.1"; FPKM "7.4263254644"; frac "1.000000"; conf_lo "4.474632"; conf_hi "7.830606"; cov "16.282075";
Sb_20131119_contig_100023       Cufflinks       exon    1097    2009    1000    .       .       gene_id "CUFF.9"; transcript_id "CUFF.9.1"; exon_number "1"; FPKM "7.4263254644"; frac "1.000000"; conf_lo "4.474632"; conf_hi "7.830606"; cov "16.282075";
Sb_20131119_contig_100023       Cufflinks       transcript      1       123     1000    .       .       gene_id "CUFF.8"; transcript_id "CUFF.8.1"; FPKM "6279.6631156093"; frac "1.000000"; conf_lo "42.624813"; conf_hi "69.749694"; cov "8277.399810";
Sb_20131119_contig_100023       Cufflinks       exon    1       123     1000    .       .       gene_id "CUFF.8"; transcript_id "CUFF.8.1"; exon_number "1"; FPKM "6279.6631156093"; frac "1.000000"; conf_lo "42.624813"; conf_hi "69.749694"; cov "8277.399810";
Sb_20131119_contig_100040       Cufflinks       transcript      1       221     1000    .       .       gene_id "CUFF.10"; transcript_id "CUFF.10.1"; FPKM "233.0138642752"; frac "1.000000"; conf_lo "22.182837"; conf_hi "38.203775"; cov "306.188212";
Sb_20131119_contig_100040       Cufflinks       exon    1       221     1000    .       .       gene_id "CUFF.10"; transcript_id "CUFF.10.1"; exon_number "1"; FPKM "233.0138642752"; frac "1.000000"; conf_lo "22.182837"; conf_hi "38.203775"; cov "306.188212";
Sb_20131119_contig_100107       Cufflinks       transcript      2       255     1000    .       .       gene_id "CUFF.11"; transcript_id "CUFF.11.1"; FPKM "29.5909083872"; frac "1.000000"; conf_lo "3.216803"; conf_hi "9.918474"; cov "41.309085";
Sb_20131119_contig_100107       Cufflinks       exon    2       255     1000    .       .       gene_id "CUFF.11"; transcript_id "CUFF.11.1"; exon_number "1"; FPKM "29.5909083872"; frac "1.000000"; conf_lo "3.216803"; conf_hi "9.918474"; cov "41.309085";
Sb_20131119_contig_100107       Cufflinks       transcript      2041    2331    1000    .       .       gene_id "CUFF.12"; transcript_id "CUFF.12.1"; FPKM "18.4360530685"; frac "1.000000"; conf_lo "2.807793"; conf_hi "8.657363"; cov "24.668034";
Sb_20131119_contig_100107       Cufflinks       exon    2041    2331    1000    .       .       gene_id "CUFF.12"; transcript_id "CUFF.12.1"; exon_number "1"; FPKM "18.4360530685"; frac "1.000000"; conf_lo "2.807793"; conf_hi "8.657363"; cov "24.668034";
Sb_20131119_contig_100111       Cufflinks       transcript      21      129     1000    .       .       gene_id "CUFF.14"; transcript_id "CUFF.14.1"; FPKM "3960.0565774823"; frac "1.000000"; conf_lo "5.622026"; conf_hi "18.740088"; cov "4272.822539";
Sb_20131119_contig_100111       Cufflinks       exon    21      129     1000    .       .       gene_id "CUFF.14"; transcript_id "CUFF.14.1"; exon_number "1"; FPKM "3960.0565774823"; frac "1.000000"; conf_lo "5.622026"; conf_hi "18.740088"; cov "4272.822539";
Sb_20131119_contig_100121       Cufflinks       transcript      1756    2236    1000    .       .       gene_id "CUFF.17"; transcript_id "CUFF.17.1"; FPKM "52.3538569152"; frac "1.000000"; conf_lo "23.640043"; conf_hi "34.398386"; cov "75.482038";
Sb_20131119_contig_100121       Cufflinks       exon    1756    2236    1000    .       .       gene_id "CUFF.17"; transcript_id "CUFF.17.1"; exon_number "1"; FPKM "52.3538569152"; frac "1.000000"; conf_lo "23.640043"; conf_hi "34.398386"; cov "75.482038";
Sb_20131119_contig_100192       Cufflinks       transcript      1840    2212    1000    .       .       gene_id "CUFF.20"; transcript_id "CUFF.20.1"; FPKM "24.9098132799"; frac "1.000000"; conf_lo "7.484312"; conf_hi "14.786080"; cov "49.916331";
Sb_20131119_contig_100192       Cufflinks       exon    1840    2212    1000    .       .       gene_id "CUFF.20"; transcript_id "CUFF.20.1"; exon_number "1"; FPKM "24.9098132799"; frac "1.000000"; conf_lo "7.484312"; conf_hi "14.786080"; cov "49.916331";
Sb_20131119_contig_100107       Cufflinks       transcript      430     902     1000    .       .       gene_id "CUFF.13"; transcript_id "CUFF.13.1"; FPKM "25.0500552879"; frac "1.000000"; conf_lo "10.508448"; conf_hi "18.137870"; cov "45.523657";
Sb_20131119_contig_100107       Cufflinks       exon    430     902     1000    .       .       gene_id "CUFF.13"; transcript_id "CUFF.13.1"; exon_number "1"; FPKM "25.0500552879"; frac "1.000000"; conf_lo "10.508448"; conf_hi "18.137870"; cov "45.523657";
Sb_20131119_contig_100192       Cufflinks       transcript      1       616     1000    .       .       gene_id "CUFF.19"; transcript_id "CUFF.19.1"; FPKM "42.2951435419"; frac "1.000000"; conf_lo "23.433223"; conf_hi "33.823425"; cov "88.806988";
Sb_20131119_contig_100192       Cufflinks       exon    1       616     1000    .       .       gene_id "CUFF.19"; transcript_id "CUFF.19.1"; exon_number "1"; FPKM "42.2951435419"; frac "1.000000"; conf_lo "23.433223"; conf_hi "33.823425"; cov "88.806988";
Sb_20131119_contig_10011        Cufflinks       transcript      219     353     1000    .       .       gene_id "CUFF.15"; transcript_id "CUFF.15.1"; FPKM "4118.4602721559"; frac "1.000000"; conf_lo "53.966826"; conf_hi "84.228598"; cov "5420.068745";
Sb_20131119_contig_10011        Cufflinks       exon    219     353     1000    .       .       gene_id "CUFF.15"; transcript_id "CUFF.15.1"; exon_number "1"; FPKM "4118.4602721559"; frac "1.000000"; conf_lo "53.966826"; conf_hi "84.228598"; cov "5420.068745";
Sb_20131119_contig_100040       Cufflinks       transcript      945     2276    1000    -       .       gene_id "CUFF.38"; transcript_id "CUFF.38.1"; FPKM "69.4702369875"; frac "1.000000"; conf_lo "60.275496"; conf_hi "70.712087"; cov "143.103653";
Sb_20131119_contig_100040       Cufflinks       exon    945     1966    1000    -       .       gene_id "CUFF.38"; transcript_id "CUFF.38.1"; exon_number "1"; FPKM "69.4702369875"; frac "1.000000"; conf_lo "60.275496"; conf_hi "70.712087"; cov "143.103653";
Sb_20131119_contig_100040       Cufflinks       exon    2079    2276    1000    -       .       gene_id "CUFF.38"; transcript_id "CUFF.38.1"; exon_number "2"; FPKM "69.4702369875"; frac "1.000000"; conf_lo "60.275496"; conf_hi "70.712087"; cov "143.103653";
Sb_20131119_contig_100121       Cufflinks       transcript      1       150     1000    .       .       gene_id "CUFF.16"; transcript_id "CUFF.16.1"; FPKM "1080.5400540118"; frac "1.000000"; conf_lo "23.150255"; conf_hi "44.484805"; cov "1391.311243";
Sb_20131119_contig_100121       Cufflinks       exon    1       150     1000    .       .       gene_id "CUFF.16"; transcript_id "CUFF.16.1"; exon_number "1"; FPKM "1080.5400540118"; frac "1.000000"; conf_lo "23.150255"; conf_hi "44.484805"; cov "1391.311243";
Sb_20131119_contig_10022        Cufflinks       transcript      1510    2616    1000    .       .       gene_id "CUFF.22"; transcript_id "CUFF.22.1"; FPKM "10.9257515912"; frac "1.000000"; conf_lo "7.442428"; conf_hi "11.194395"; cov "24.833916";
Sb_20131119_contig_10022        Cufflinks       exon    1510    2616    1000    .       .       gene_id "CUFF.22"; transcript_id "CUFF.22.1"; exon_number "1"; FPKM "10.9257515912"; frac "1.000000"; conf_lo "7.442428"; conf_hi "11.194395"; cov "24.833916";
Sb_20131119_contig_100112       Cufflinks       transcript      933     3862    1000    .       .       gene_id "CUFF.21"; transcript_id "CUFF.21.1"; FPKM "13.6659124939"; frac "1.000000"; conf_lo "11.688997"; conf_hi "14.524101"; cov "32.588772";
Sb_20131119_contig_100112       Cufflinks       exon    933     3862    1000    .       .       gene_id "CUFF.21"; transcript_id "CUFF.21.1"; exon_number "1"; FPKM "13.6659124939"; frac "1.000000"; conf_lo "11.688997"; conf_hi "14.524101"; cov "32.588772";
Sb_20131119_contig_10022        Cufflinks       transcript      2857    6365    1000    .       .       gene_id "CUFF.24"; transcript_id "CUFF.24.1"; FPKM "13.7863265713"; frac "1.000000"; conf_lo "12.049946"; conf_hi "14.805328"; cov "32.698360";
Sb_20131119_contig_10022        Cufflinks       exon    2857    6365    1000    .       .       gene_id "CUFF.24"; transcript_id "CUFF.24.1"; exon_number "1"; FPKM "13.7863265713"; frac "1.000000"; conf_lo "12.049946"; conf_hi "14.805328"; cov "32.698360";
Sb_20131119_contig_100080       Cufflinks       transcript      39      1611    1000    -       .       gene_id "CUFF.18"; transcript_id "CUFF.18.1"; FPKM "44.1667134200"; frac "1.000000"; conf_lo "34.717689"; conf_hi "42.267104"; cov "77.065569";
Sb_20131119_contig_100080       Cufflinks       exon    39      135     1000    -       .       gene_id "CUFF.18"; transcript_id "CUFF.18.1"; exon_number "1"; FPKM "44.1667134200"; frac "1.000000"; conf_lo "34.717689"; conf_hi "42.267104"; cov "77.065569";
Sb_20131119_contig_100080       Cufflinks       exon    207     1041    1000    -       .       gene_id "CUFF.18"; transcript_id "CUFF.18.1"; exon_number "2"; FPKM "44.1667134200"; frac "1.000000"; conf_lo "34.717689"; conf_hi "42.267104"; cov "77.065569";
Sb_20131119_contig_100080       Cufflinks       exon    1128    1611    1000    -       .       gene_id "CUFF.18"; transcript_id "CUFF.18.1"; exon_number "3"; FPKM "44.1667134200"; frac "1.000000"; conf_lo "34.717689"; conf_hi "42.267104"; cov "77.065569";
Sb_20131119_contig_10026        Cufflinks       transcript      1       1002    1000    .       .       gene_id "CUFF.49"; transcript_id "CUFF.49.1"; FPKM "49.5933867311"; frac "1.000000"; conf_lo "35.335602"; conf_hi "43.965643"; cov "84.354408";
Sb_20131119_contig_10026        Cufflinks       exon    1       1002    1000    .       .       gene_id "CUFF.49"; transcript_id "CUFF.49.1"; exon_number "1"; FPKM "49.5933867311"; frac "1.000000"; conf_lo "35.335602"; conf_hi "43.965643"; cov "84.354408";
Sb_20131119_contig_10031        Cufflinks       transcript      26      377     1000    .       .       gene_id "CUFF.25"; transcript_id "CUFF.25.1"; FPKM "43.3120795956"; frac "1.000000"; conf_lo "12.379816"; conf_hi "22.244981"; cov "78.396470";
Sb_20131119_contig_10031        Cufflinks       exon    26      377     1000    .       .       gene_id "CUFF.25"; transcript_id "CUFF.25.1"; exon_number "1"; FPKM "43.3120795956"; frac "1.000000"; conf_lo "12.379816"; conf_hi "22.244981"; cov "78.396470";
Sb_20131119_contig_10032        Cufflinks       transcript      5174    5386    1000    .       .       gene_id "CUFF.29"; transcript_id "CUFF.29.1"; FPKM "386.7080614787"; frac "1.000000"; conf_lo "33.884660"; conf_hi "52.744990"; cov "584.846045";
Sb_20131119_contig_10032        Cufflinks       exon    5174    5386    1000    .       .       gene_id "CUFF.29"; transcript_id "CUFF.29.1"; exon_number "1"; FPKM "386.7080614787"; frac "1.000000"; conf_lo "33.884660"; conf_hi "52.744990"; cov "584.846045";
Sb_20131119_contig_10032        Cufflinks       transcript      160     3305    1000    .       .       gene_id "CUFF.35"; transcript_id "CUFF.35.1"; FPKM "5.9167626598"; frac "1.000000"; conf_lo "4.912969"; conf_hi "6.795913"; cov "12.795227";
Sb_20131119_contig_10032        Cufflinks       exon    160     3305    1000    .       .       gene_id "CUFF.35"; transcript_id "CUFF.35.1"; exon_number "1"; FPKM "5.9167626598"; frac "1.000000"; conf_lo "4.912969"; conf_hi "6.795913"; cov "12.795227";
Sb_20131119_contig_100403       Cufflinks       transcript      822     1053    1000    .       .       gene_id "CUFF.30"; transcript_id "CUFF.30.1"; FPKM "90.3398746415"; frac "1.000000"; conf_lo "9.098097"; conf_hi "19.663630"; cov "118.213868";
Sb_20131119_contig_100403       Cufflinks       exon    822     1053    1000    .       .       gene_id "CUFF.30"; transcript_id "CUFF.30.1"; exon_number "1"; FPKM "90.3398746415"; frac "1.000000"; conf_lo "9.098097"; conf_hi "19.663630"; cov "118.213868";
Sb_20131119_contig_100412       Cufflinks       transcript      226     2367    1000    .       .       gene_id "CUFF.32"; transcript_id "CUFF.32.1"; FPKM "5.4971611120"; frac "1.000000"; conf_lo "4.259535"; conf_hi "6.262152"; cov "11.356891";
Sb_20131119_contig_100412       Cufflinks       exon    226     2367    1000    .       .       gene_id "CUFF.32"; transcript_id "CUFF.32.1"; exon_number "1"; FPKM "5.4971611120"; frac "1.000000"; conf_lo "4.259535"; conf_hi "6.262152"; cov "11.356891";
Sb_20131119_contig_10031        Cufflinks       transcript      1415    3072    1000    .       .       gene_id "CUFF.36"; transcript_id "CUFF.36.1"; FPKM "55.9644866986"; frac "1.000000"; conf_lo "45.789638"; conf_hi "53.715558"; cov "119.323264";
Sb_20131119_contig_10031        Cufflinks       exon    1415    3072    1000    .       .       gene_id "CUFF.36"; transcript_id "CUFF.36.1"; exon_number "1"; FPKM "55.9644866986"; frac "1.000000"; conf_lo "45.789638"; conf_hi "53.715558"; cov "119.323264";
Sb_20131119_contig_100403       Cufflinks       transcript      2827    3748    1000    .       .       gene_id "CUFF.31"; transcript_id "CUFF.31.1"; FPKM "5.2492236693"; frac "1.000000"; conf_lo "3.027818"; conf_hi "5.981787"; cov "12.544434";
Sb_20131119_contig_100403       Cufflinks       exon    2827    3748    1000    .       .       gene_id "CUFF.31"; transcript_id "CUFF.31.1"; exon_number "1"; FPKM "5.2492236693"; frac "1.000000"; conf_lo "3.027818"; conf_hi "5.981787"; cov "12.544434";
Sb_20131119_contig_100328       Cufflinks       transcript      1505    3535    1000    .       .       gene_id "CUFF.33"; transcript_id "CUFF.33.1"; FPKM "4.5481695975"; frac "1.000000"; conf_lo "3.453060"; conf_hi "5.296928"; cov "10.333563";
transcripts.gtf
```

Karolinska Institutet · KTH Vetenskap och Konst · Stockholms universitet · Uppsala Universitet

BILS Bioinformatics Infrastructure for Life Sciences

SciLifeLab

GTF file format

Seqid	source	type	start	end	score	strand	phase	attributes
Chr1	Snap	exon	234	1543	.	+	.	gene_id “gene1”; transcript_id “transcript1”;
Chr1	Snap	CDS	577	1543	.	+	0	gene_id “gene1”; transcript_id “transcript1”;
Chr1	Snap	exon	1822	2674	.	+	.	gene_id “gene1”; transcript_id “transcript1”;
Chr1	Snap	CDS	1822	2674	.	+	2	gene_id “gene1”; transcript_id “transcript1”;
		start_ codon						
		stop_ codon						

Why is annotation important?

Example: Differential expression

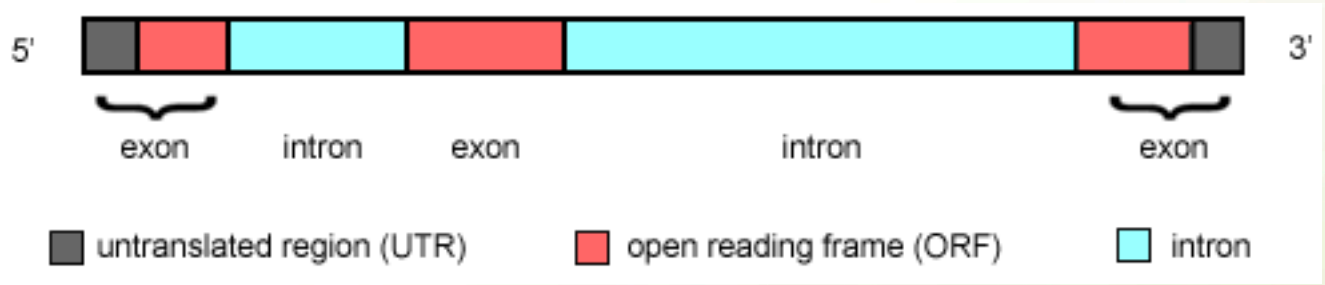
Why is annotation important?

RNA-seq reads

Genome

There are two major parts of annotation

- 1) Structural: Find out where the regions of interest (usually genes) are in the genome and what they look like. How many exons/introns? UTRs? Isoforms?

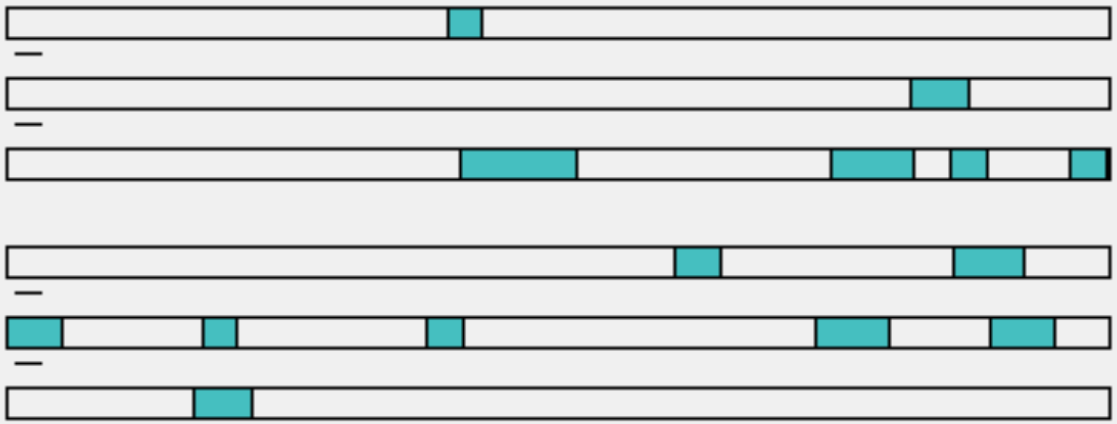


- 2) Functional: Find out what the regions do. What do they code for?

Open reading frames

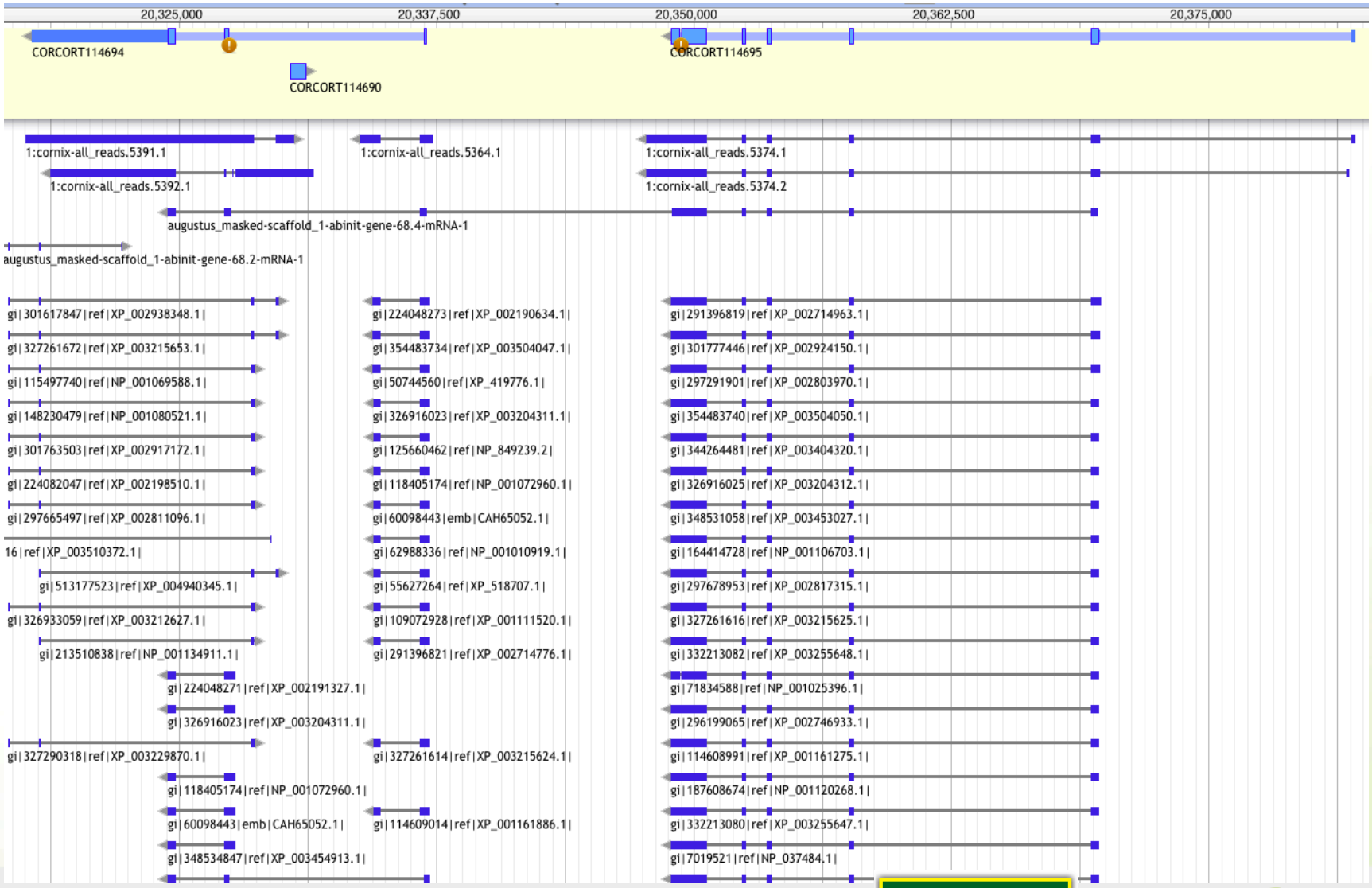
Anonymous

View | 1 GenBank | Redraw | 100 | SixFrames



Frame	from	to	Length
+3	1383	1736	354
+3	2511	2756	246
-2	2465	2686	222
-1	2880	3092	213
-2	2996	3187	192
-3	574	753	180
+2	2753	2929	177
-2	2	172	171
-1	2034	2171	138
+3	3237	3349	114
+3	2874	2984	111
-2	1283	1393	111
+1	1345	1446	102
-2	599	700	102

Difficult in practice



Karolinska Institutet · KTH · Stockholms universitet · Uppsala Universitet · BILS Bioinformatics Infrastructure for Life Sciences · SciLifeLab

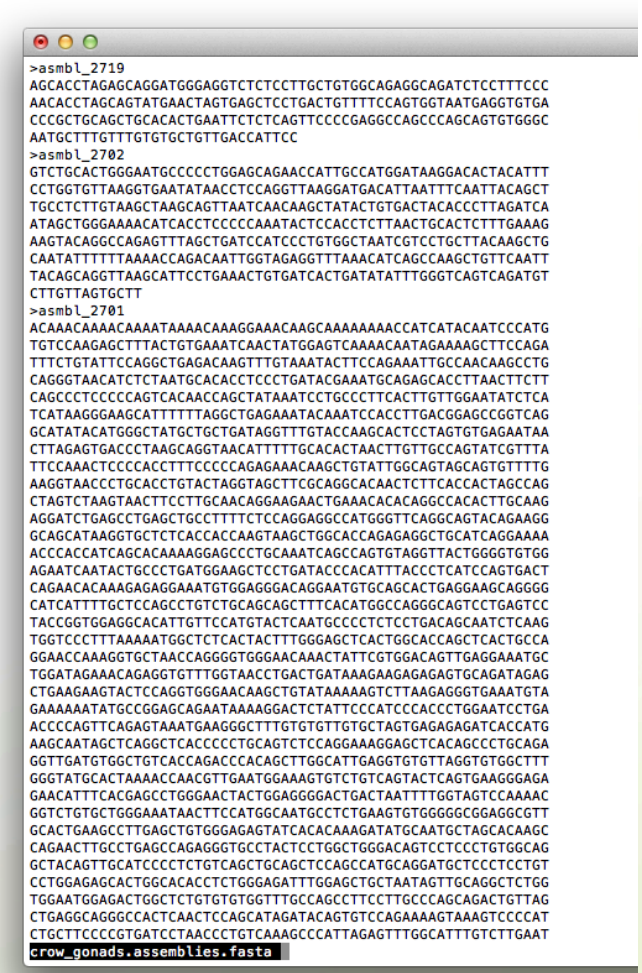
Combine data - use Maker!

- External data - proteins, rna-seq (incl. ESTs)
- Ab-initio gene finders
- (Lift-overs from closely related genomes)

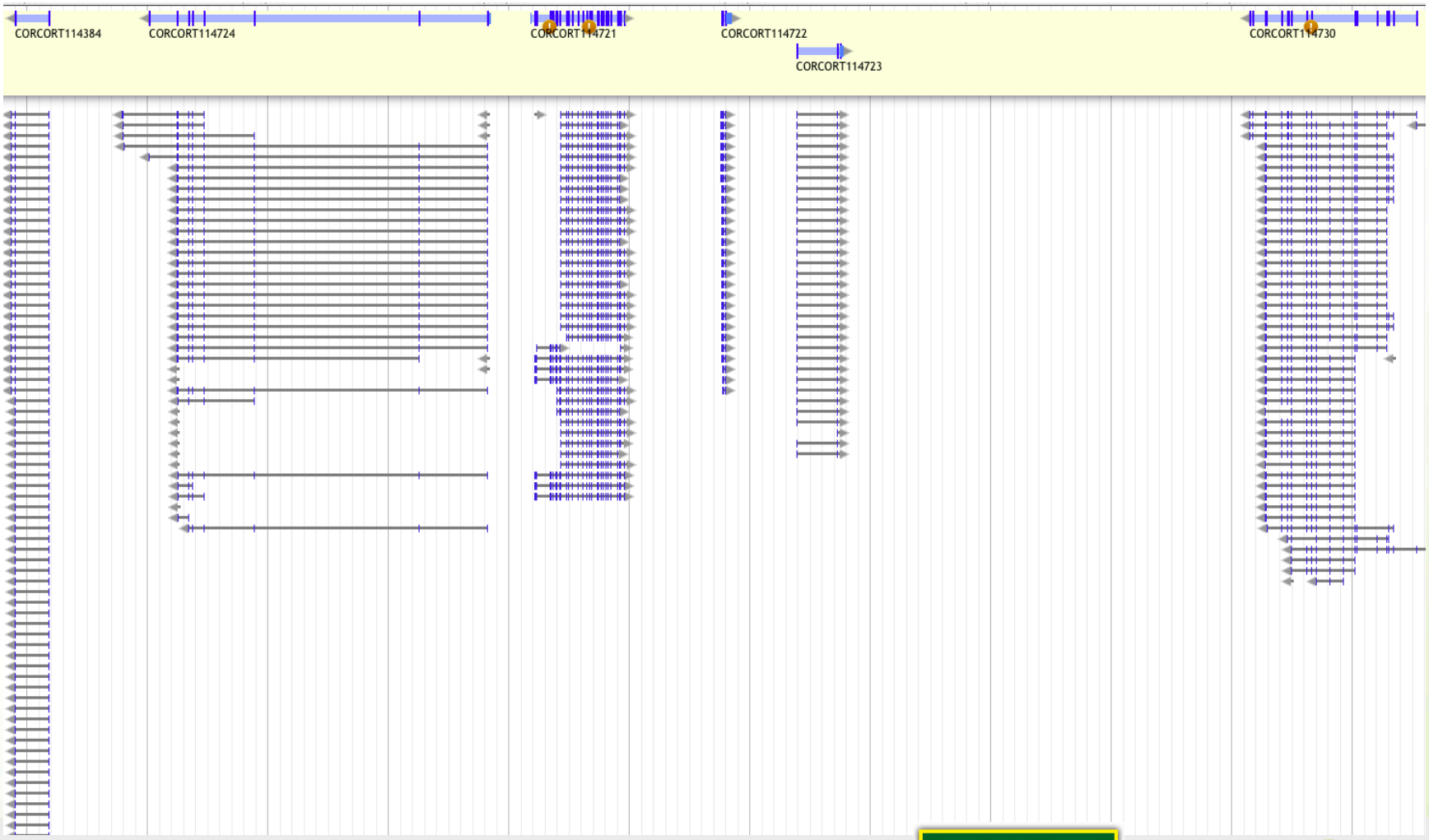
Combined annotation

Transcriptomes are different but have their own challenges

- No introns, but where are the start and stop codons?
- Still needs functional annotation



Data used - Proteins



Data used - Proteins

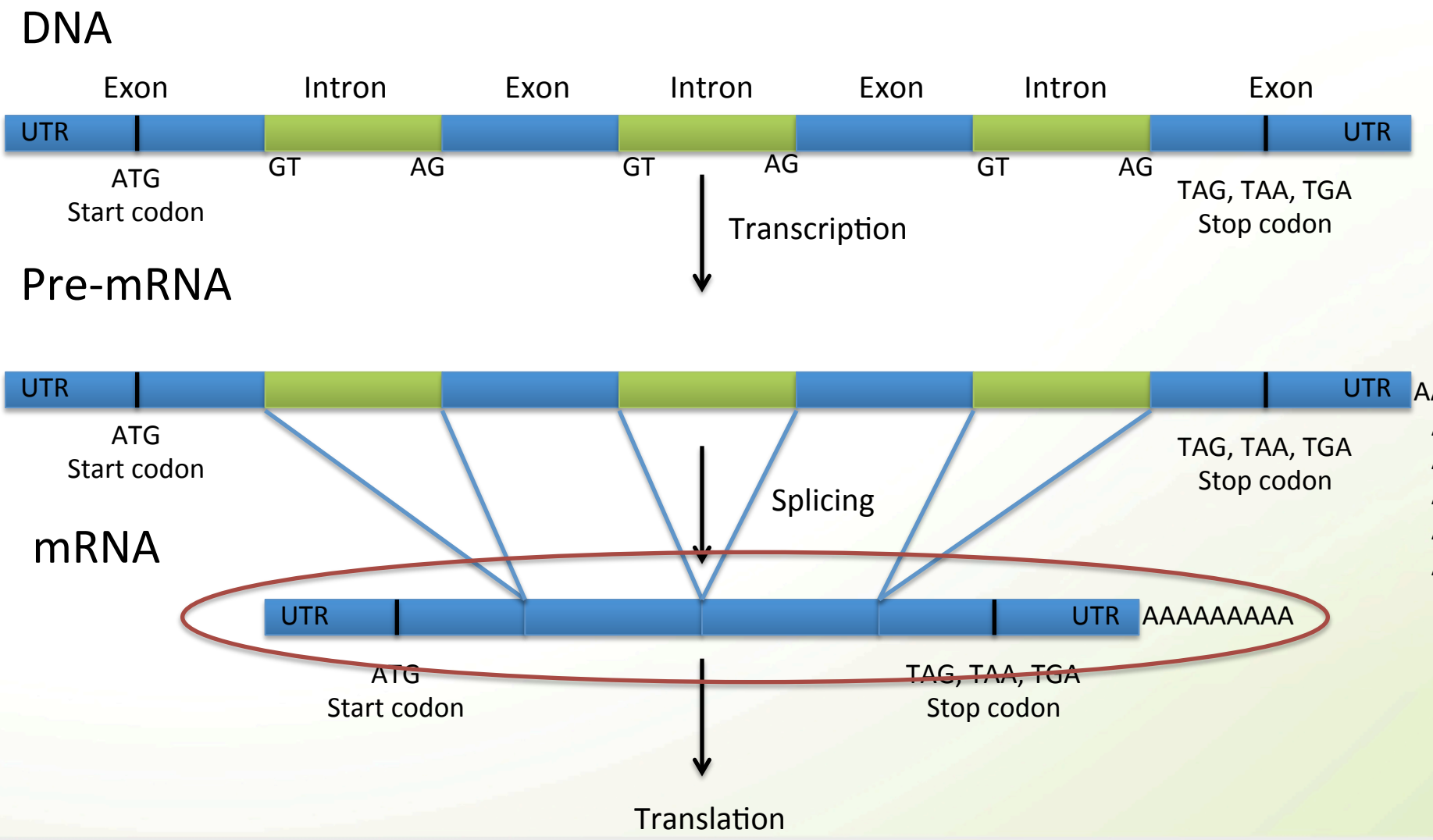
- Conserved in sequence => conserved annotation with little noise
- Proteins from model organisms often used => bias?
- Proteins can be incomplete => problems as many annotation procedures are heavily dependent on protein alignments

>ENSTGUP00000017616 pep:novel chromosome:taeGut3.2.4:8_random:2849599:2959678:-1 gene:ENSTGUG00000017338 transcript:ENSTGUT000000180
RSPNATEYNWHHLRYPKIPERLNPPAAAGPALSTAEGWMLPWGNGQHPLLARAPGKGRER
DGKELIKKPKTFKFTFLKKKKKKKKKTFK
>ENSTGUP00000017615 pep:novel chromosome:taeGut3.2.4:23_random:205321:209117:1 gene:ENSTGUG00000017337 transcript:ENSTGUT00000018017
PDLRELVLMFEHLHRVRNGGFRNSEVKKWPDRSPPPYHSFTPAQKSFSLAGCSGESTKMG
IKERMRLSSSQRQGSRGRQQHLGPPLHRSPSPEDVAEATSPTKVQKSWSFNDRTRFRASL
RLKPRIPAEGDCPPEDSGEERSSPCDLTFEDIMPAVKTLIRAVRILKFLVAKRKFKETLR
PYDVKDVIEQYSAGHLDMLGRIKSLQTRVEQIVGRDRALPADKKVREKGEKPALEAELVD
ELSMMGRVVKVERQVQSIEHKLDLLLGLYSRCLRKGSANSLVLAAVRVPPGEPDVTSDYQ
SPVEHEDISTSAQSLSISRLASTNMD

Data used - Proteins

- Maker will align proteins for you: Blast -> Exonerate
- Blast is not structure aware, Exonerate is (splice sites, start/stop codons)
- Preferred file-format: fasta

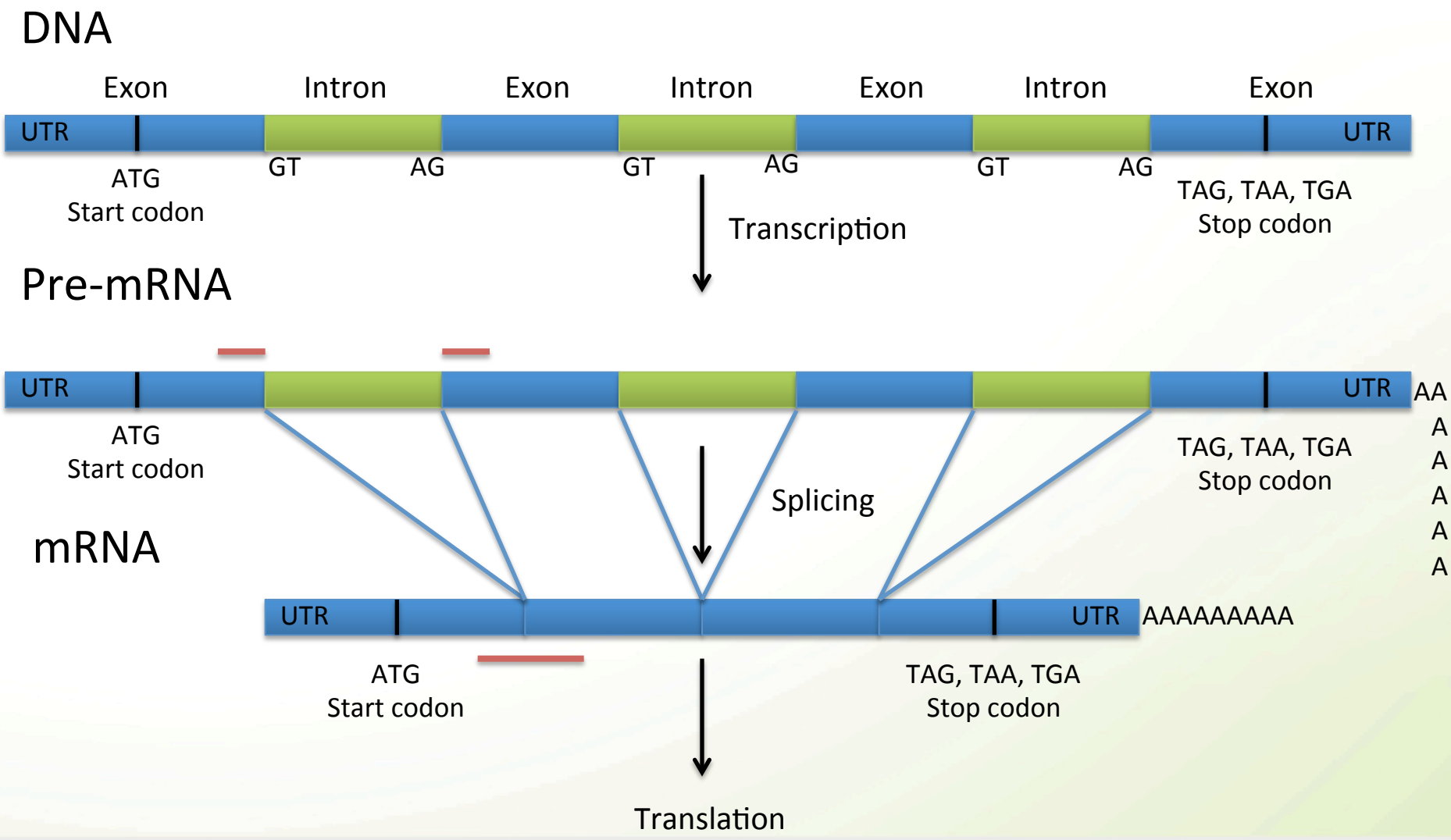
RNA-seq


DNA
Exon
Intron
Exon
Intron
Exon
Intron
Exon
UTR
UTR
GT
AG
GT
AG
GT
AG
ATG
Start codon
TAG, TAA, TGA
Stop codon
Transcription
Pre-mRNA
UTR
UTR
AA
A
A
A
A
A
ATG
Start codon
TAG, TAA, TGA
Stop codon
Splicing
mRNA
UTR
UTR
AAAAAAAAA
ATG
Start codon
TAG, TAA, TGA
Stop codon
Translation

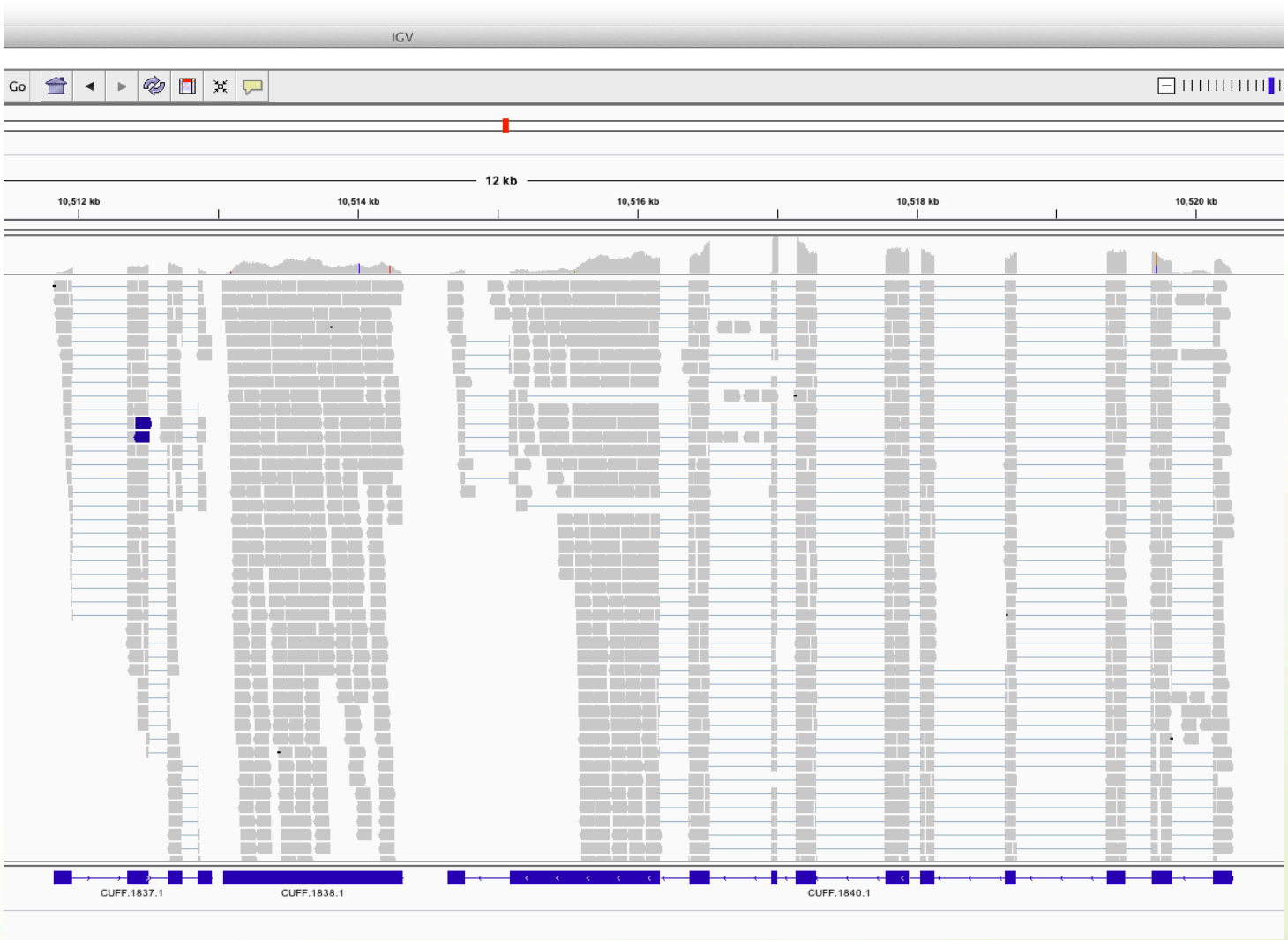
Data used - RNA-seq

- Should always be included in an annotation project
- From the same organism as the genomic data => unbiased
- Can be very noisy (tissue/species dependent), can include pre-mRNA
- PASA, or some other filtering method, often needed

Spliced reads

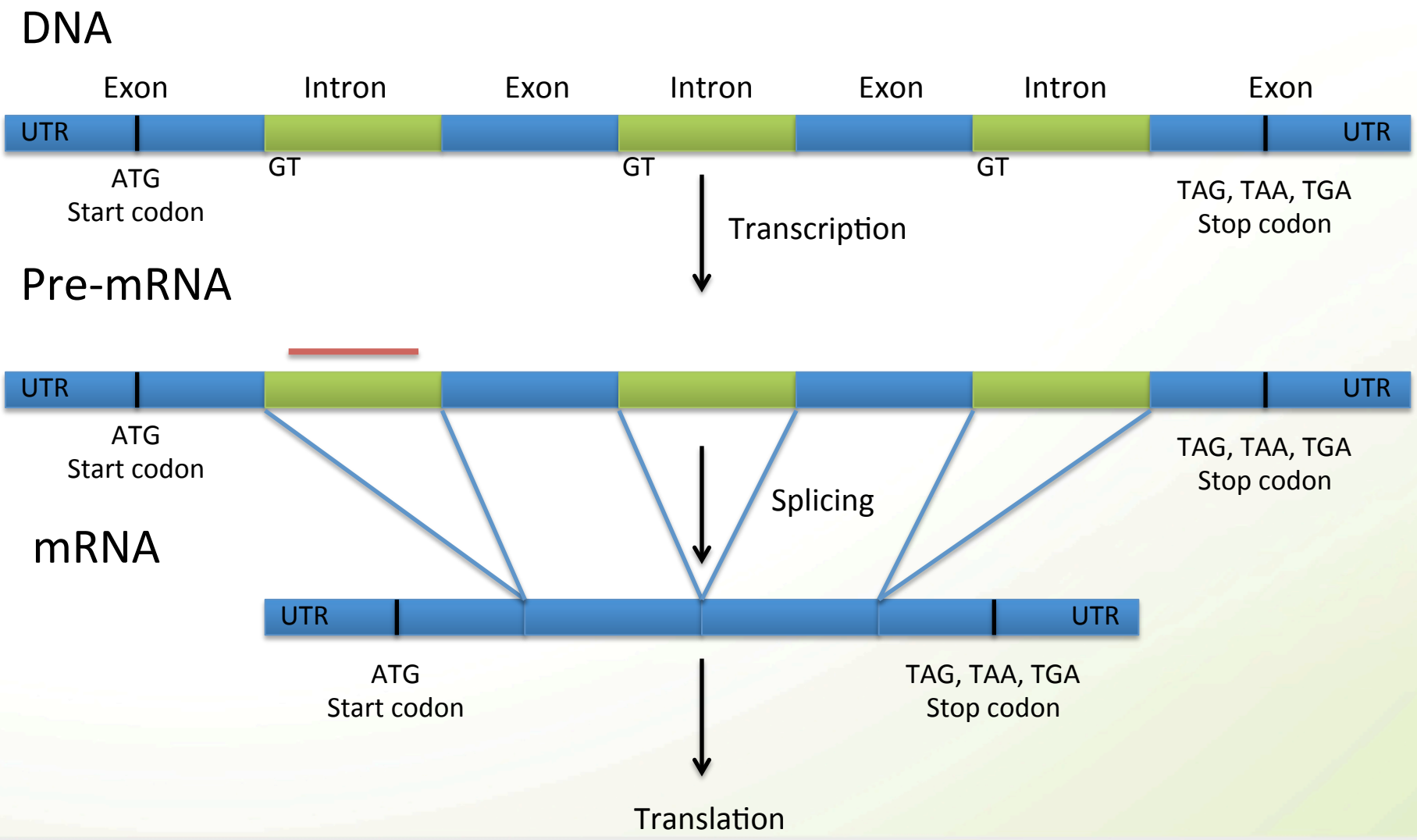


RNA-seq - Spliced reads

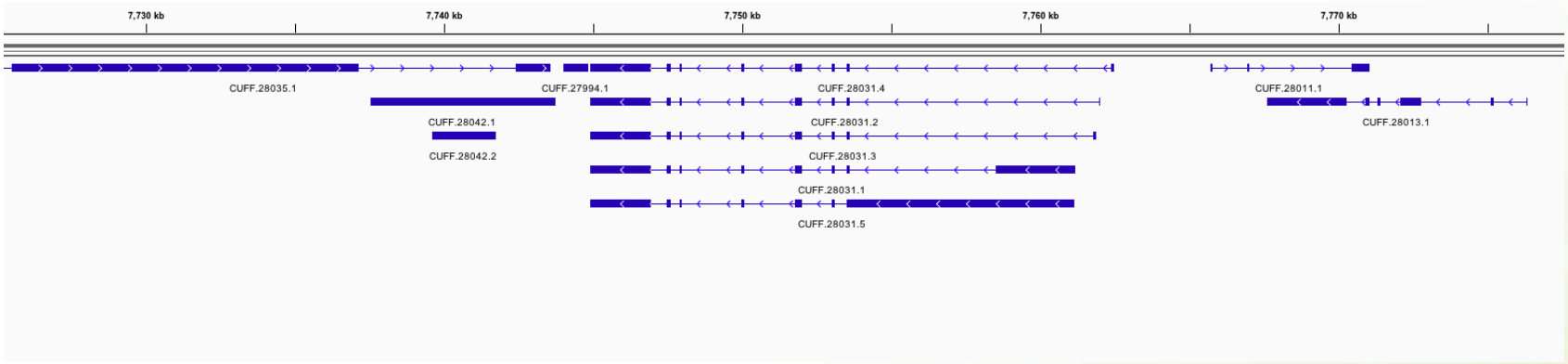


Pre-mRNA

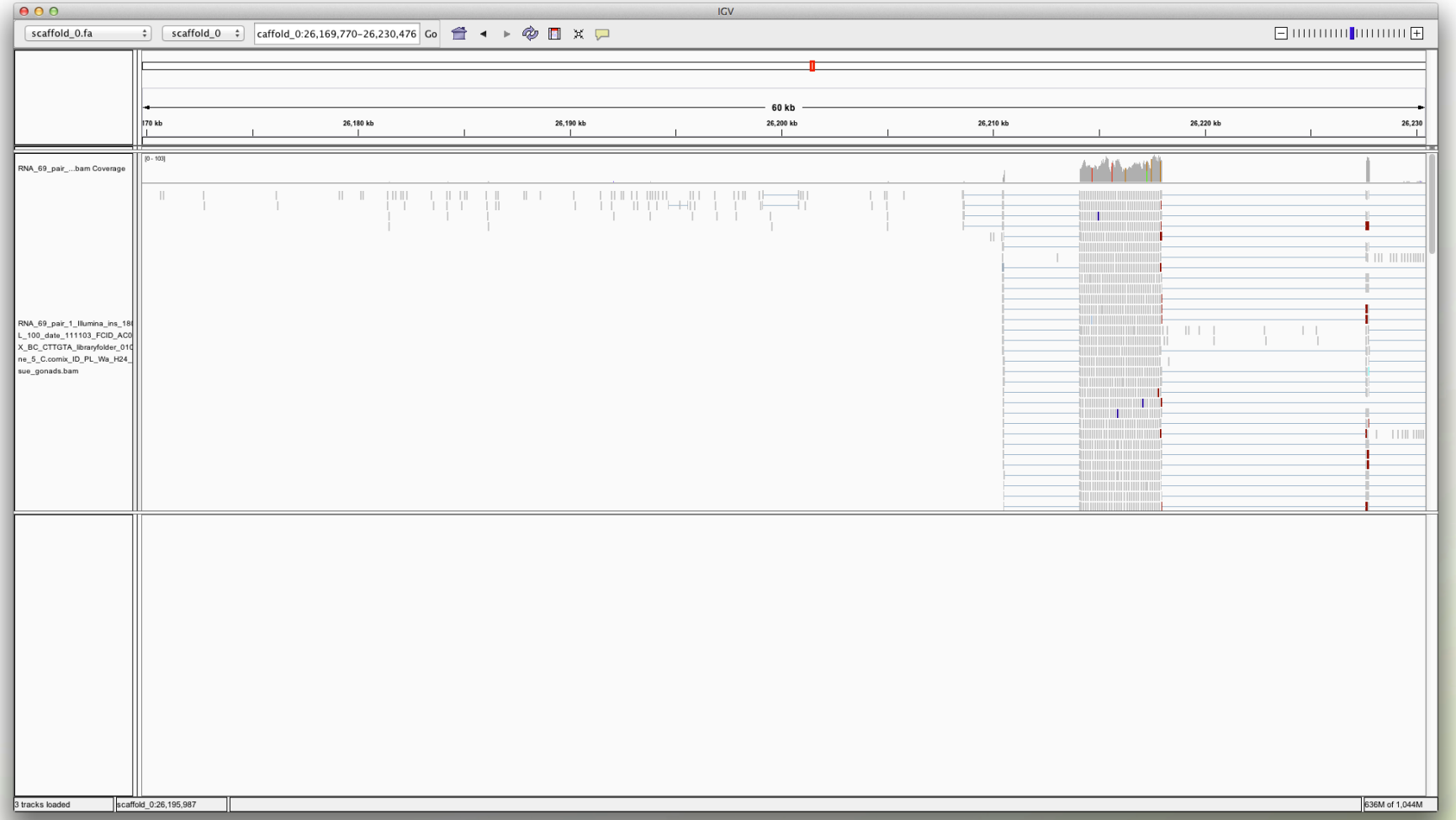
Pre-mRNA



Pre-mRNA

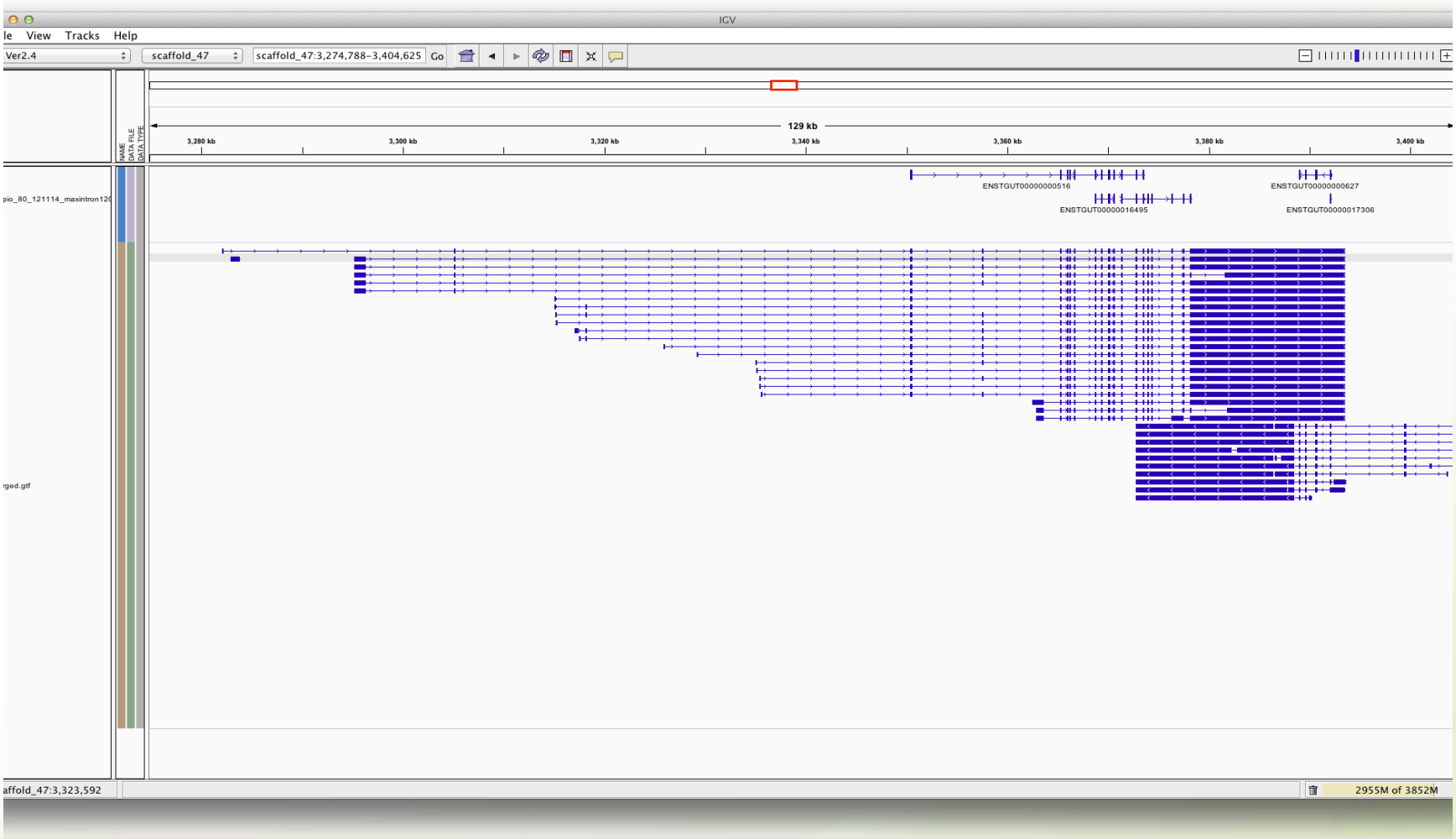

7,730 kb
7,740 kb
7,750 kb
7,760 kb
7,770 kb
CUFF.28035.1
CUFF.27994.1
CUFF.28031.4
CUFF.28011.1
CUFF.28042.1
CUFF.28031.2
CUFF.28013.1
CUFF.28042.2
CUFF.28031.3
CUFF.28031.1
CUFF.28031.5

Includes everything that is transcribed



Stranded rna-seq

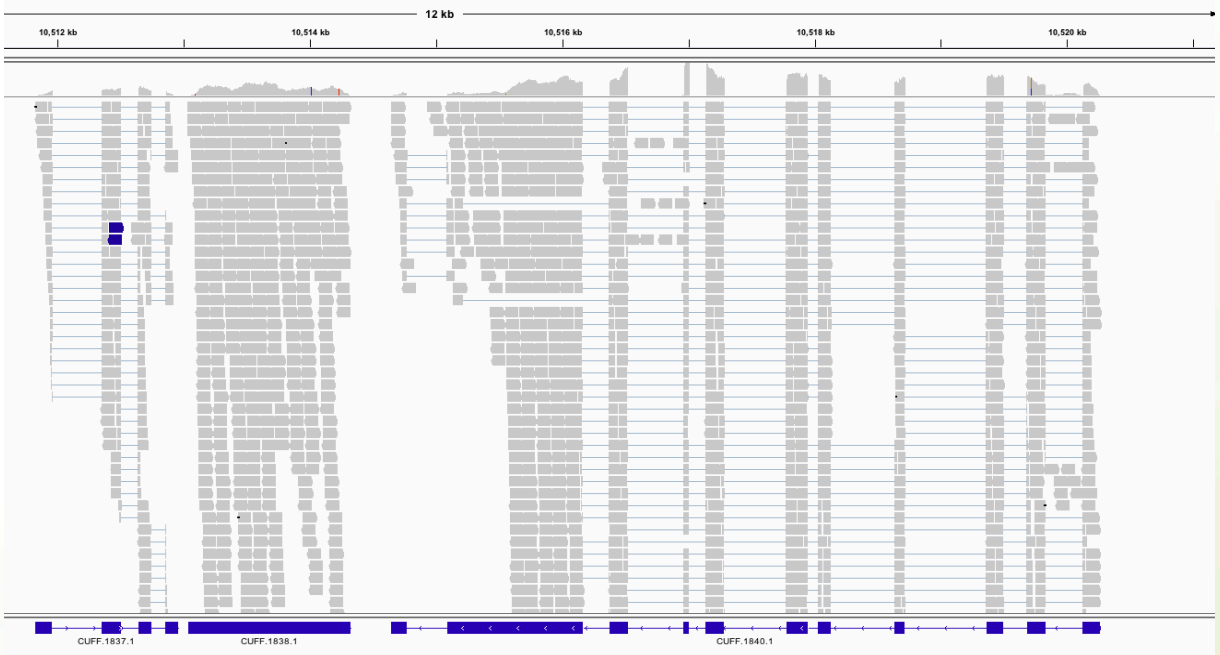
Three-prime bias in polyA-selected rna-seq



Karolinska Institutet
KTH Vetenskap och konst
Stockholms universitet
Uppsala Universitet
BILS Bioinformatics Infrastructure for Life Sciences
SciLifeLab

How to use RNA-seq

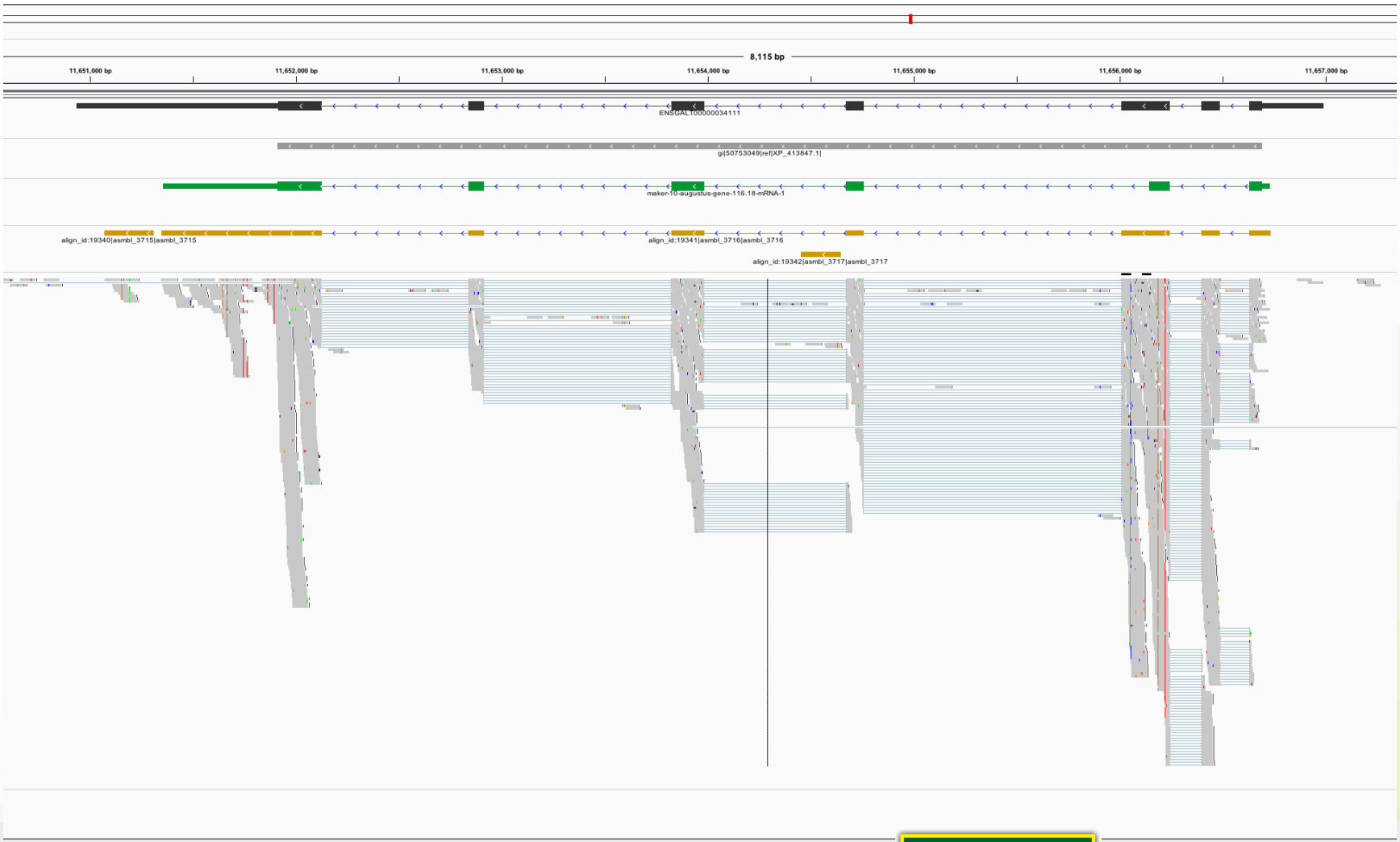
- Maker will align transcripts (ESTs), but these need to be assembled first.
- Cufflinks: mapped reads -> transcripts



How to use RNA-seq

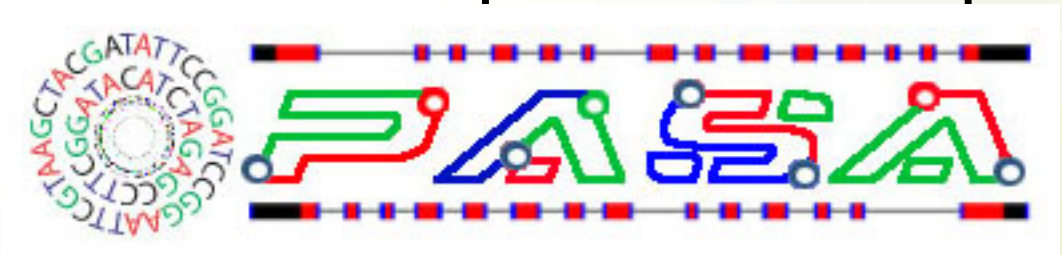
- Maker will align transcripts (ESTs), but these need to be assembled first.
- Cufflinks: mapped reads -> transcripts
- Trinity: assembles transcripts without a genome

Mapped Trinity-assembled transcripts



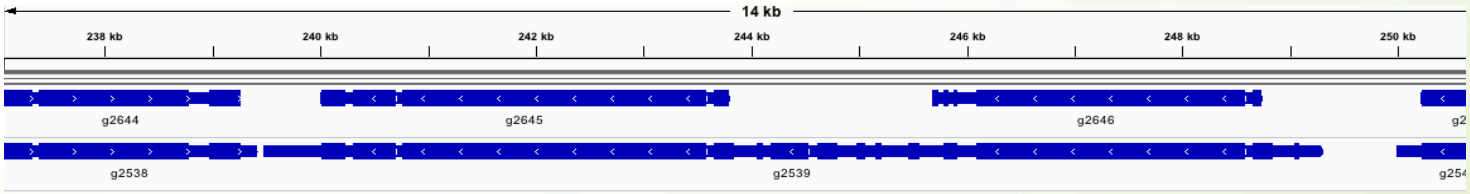
How to use RNA-seq

- Maker will align transcripts (ESTs), but these need to be assembled first.
- Cufflinks: mapped reads -> transcripts
- Trinity: assembles transcripts without a genome
- PASA can be used to improve transcript quality



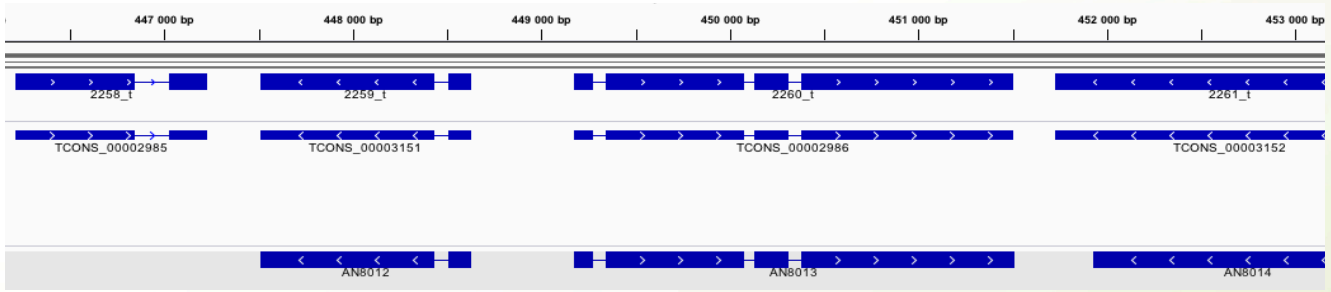
Ab initio gene finders are used in Maker

- Commonly used programs: Augustus, Snap, Genemark-ES, FGENESH, Genscan, Glimmer-HMM,…
- Uses HMM-models to figure out how introns, exons, UTRs etc. are structured
- These HMM-models need to be trained!



Liftovers are very useful for orthology determination

- Kraken
- Align the two genomes (Satsuma) and then transfer annotations between aligned regions



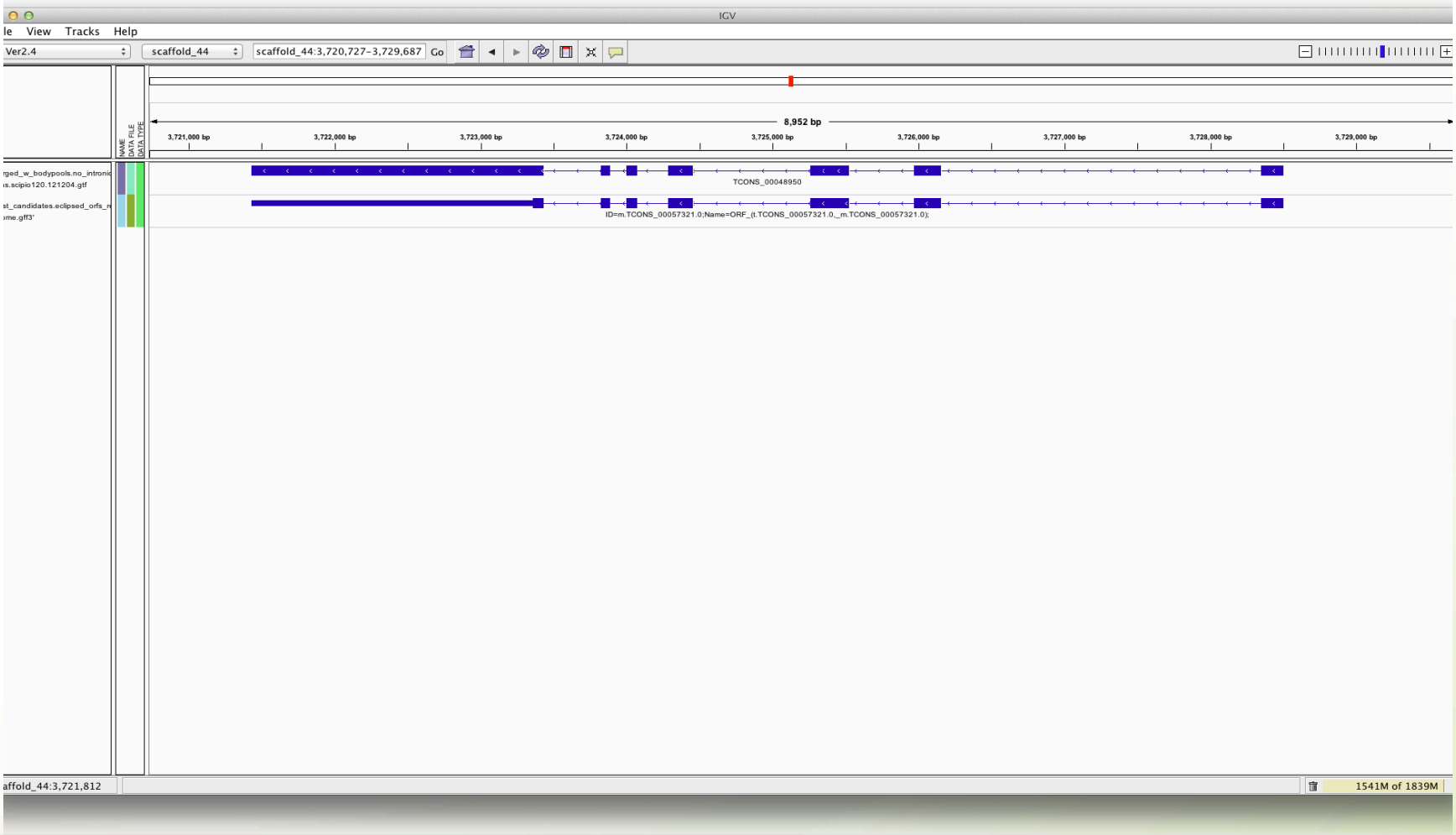
General recommendations

- Always combine different types of evidence!
- One single method is not enough!
- Use Maker!

Transcript annotation

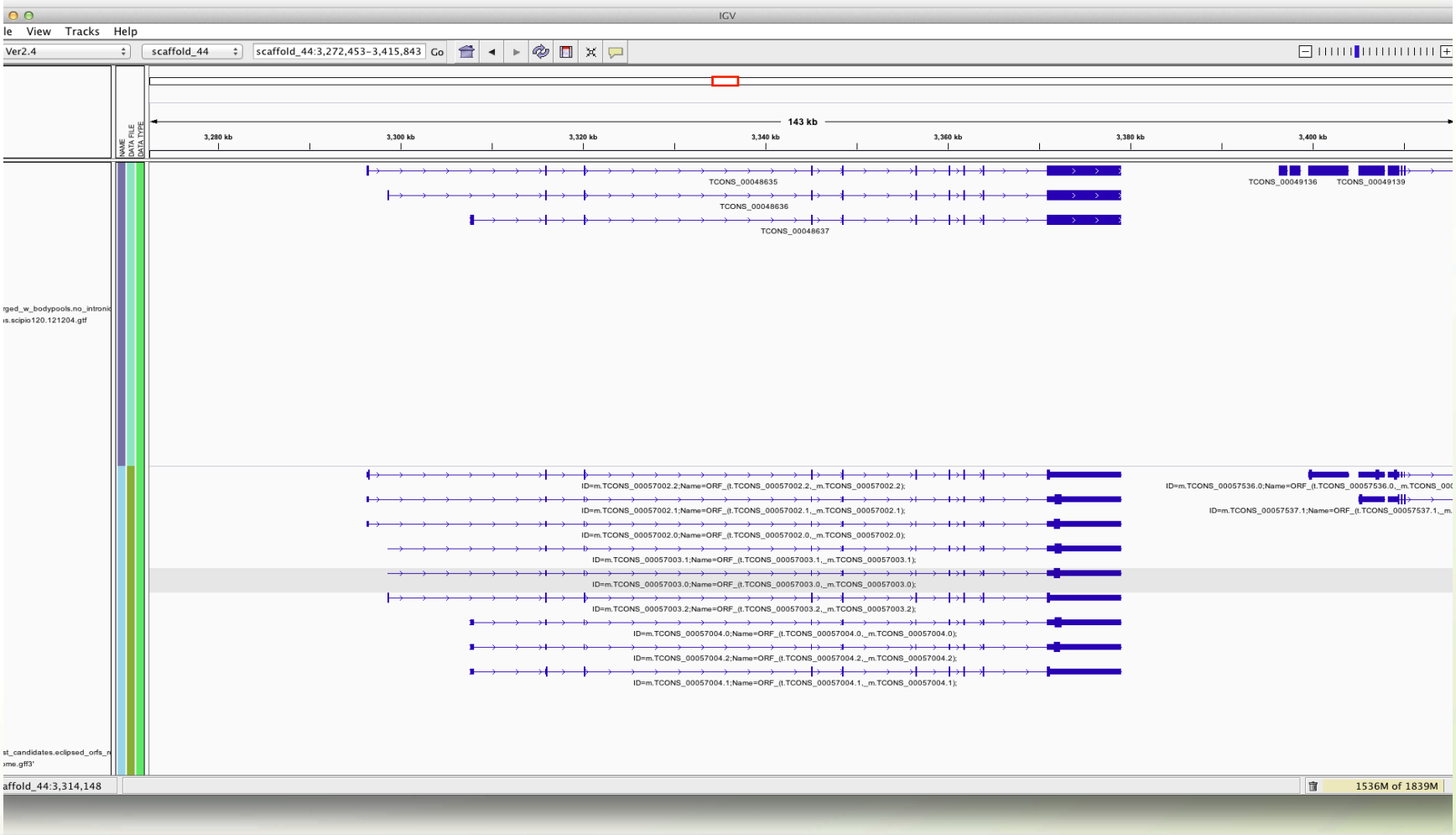
- Here the transcript is already defined. The challenge is to find where the coding regions starts and stops
- Transdecoder

Transdecoder



Karolinska Institutet
KTH Vetenskap och konst
Stockholms universitet
Uppsala Universitet
BILS Bioinformatics Infrastructure for Life Sciences
SciLifeLab

Transdecoder

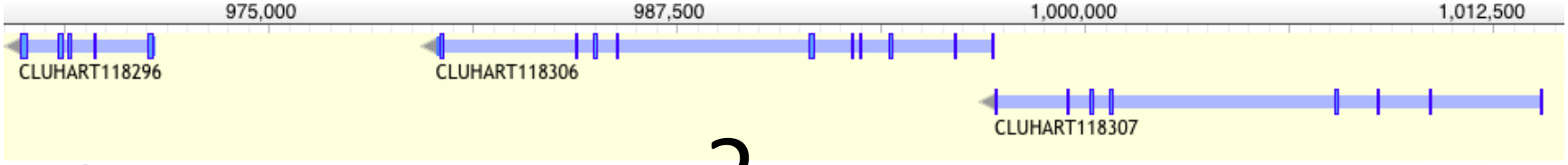


Karolinska Institutet · KTH Vetenskap och Konst · Stockholms universitet · Uppsala Universitet

BILS Bioinformatics Infrastructure for Life Sciences

SciLifeLab

Right, now we have our genes, but what do they do?



?

?

?

Insulin receptor?

Vesicle-trafficking protein?

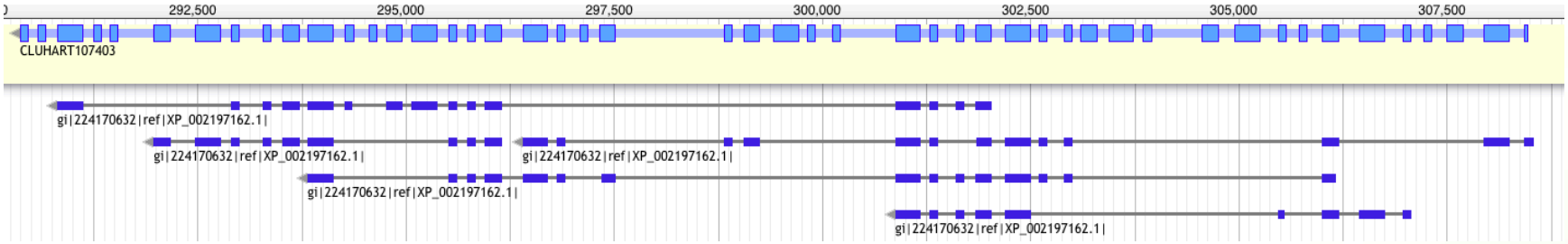
Alcohol dehydrogenase?

Aquaporin?

Transcription factor

MAP kinase kinase kinase?

But we have used proteins in our annotation!



It is actually kind of complex…

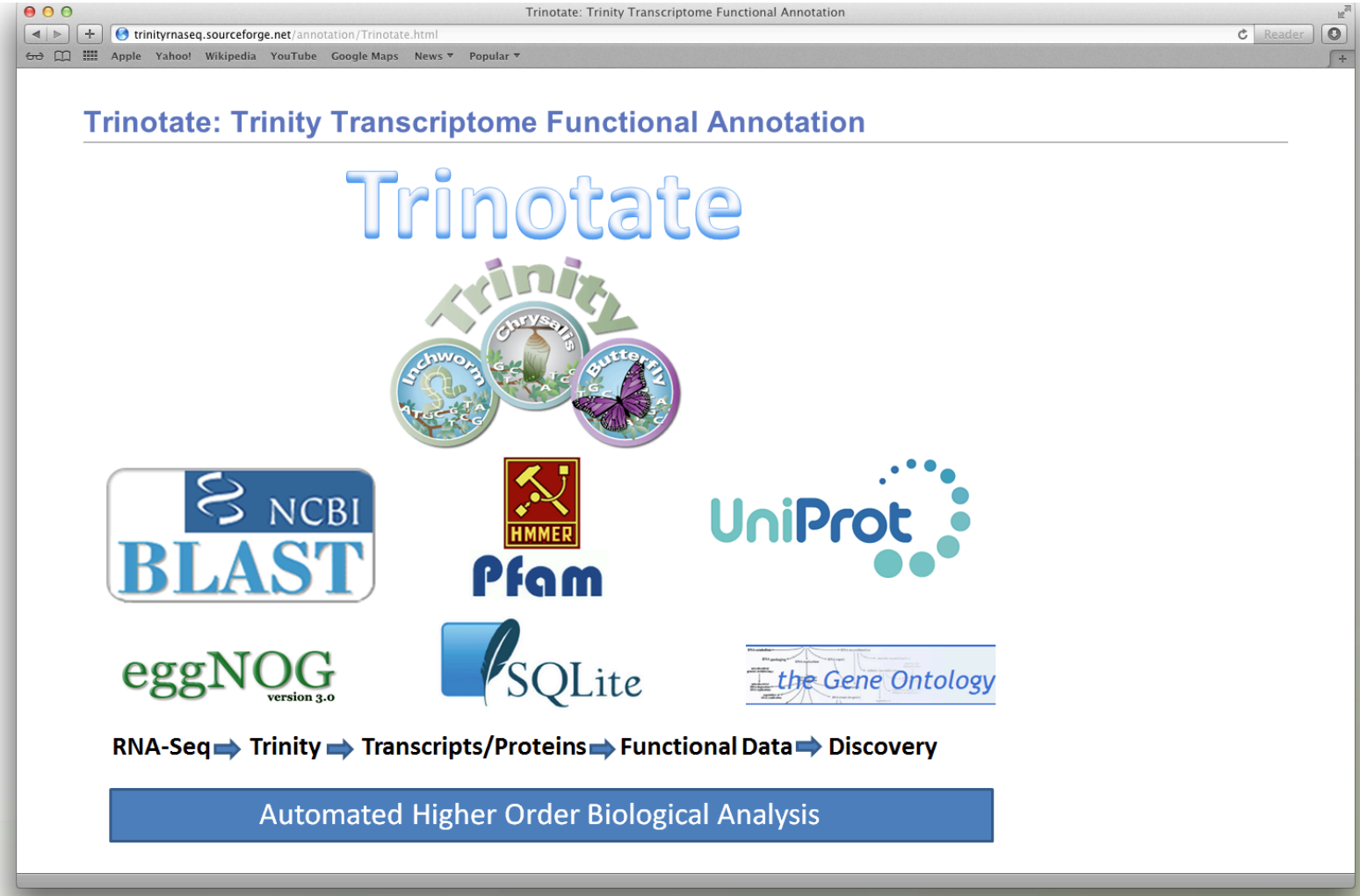
… and Maker does not do this for you.

Extract sequences -> functional annotation

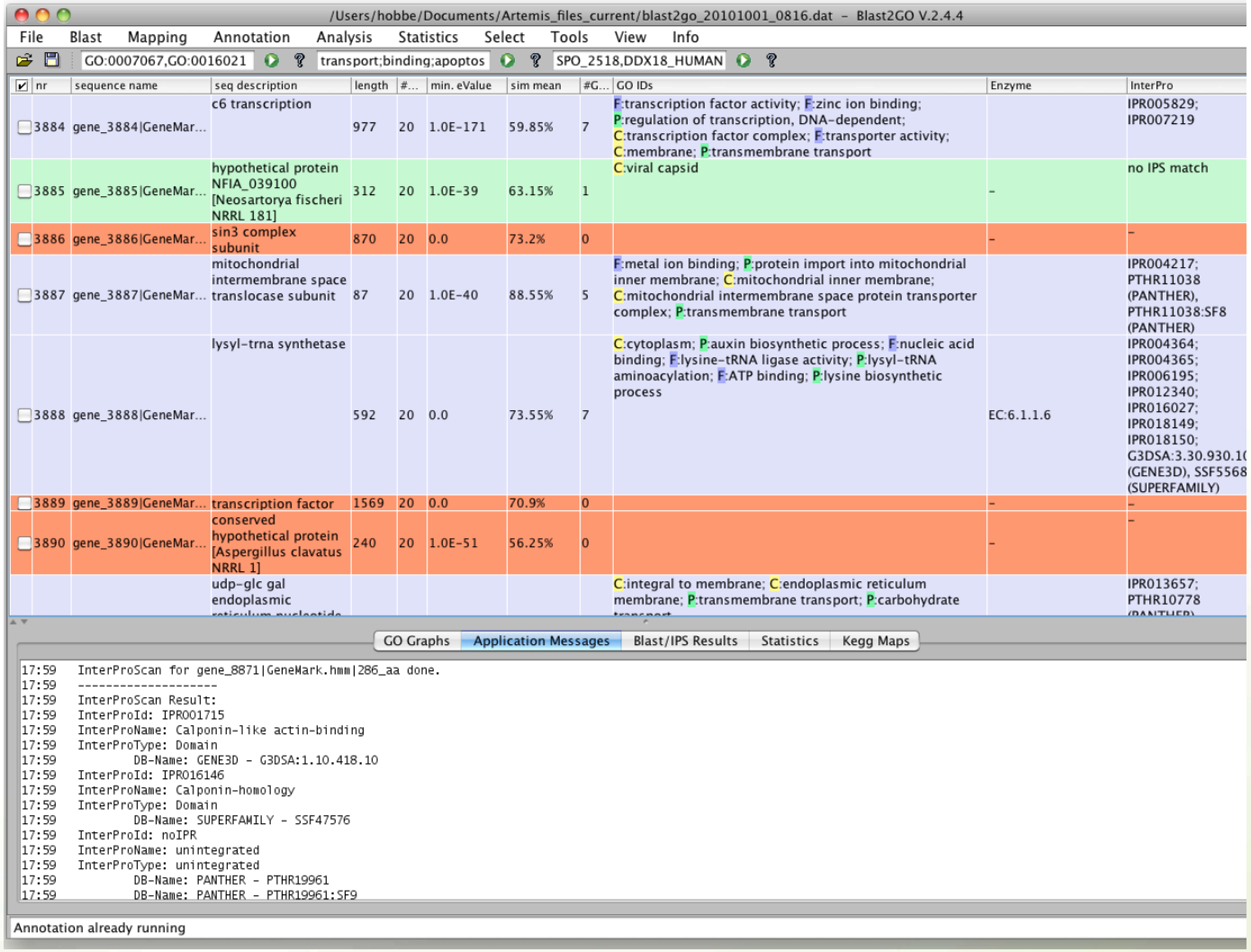
- Extract sequences from Webapollo or use gffread (in Cufflinks package)
- Annotate the sequences functionally in Blast2GO

- Full functionality now commercial…

Trinotate


Trinotate: Trinity Transcriptome Functional Annotation
trinityrnaseq.sourceforge.net/annotation/Trinotate.html
Apple Yahoo! Wikipedia YouTube Google Maps News Popular
Trinotate: Trinity Transcriptome Functional Annotation
Trinotate
Trinity
Inchworm
Chrysalis
Butterfly
NCBI
BLAST
HMMER
Pfam
UniProt
eggNOG
version 3.0
SQLite
the Gene Ontology
RNA-Seq Trinity Transcripts/Proteins Functional Data Discovery
Automated Higher Order Biological Analysis

Blast2GO



blast2go

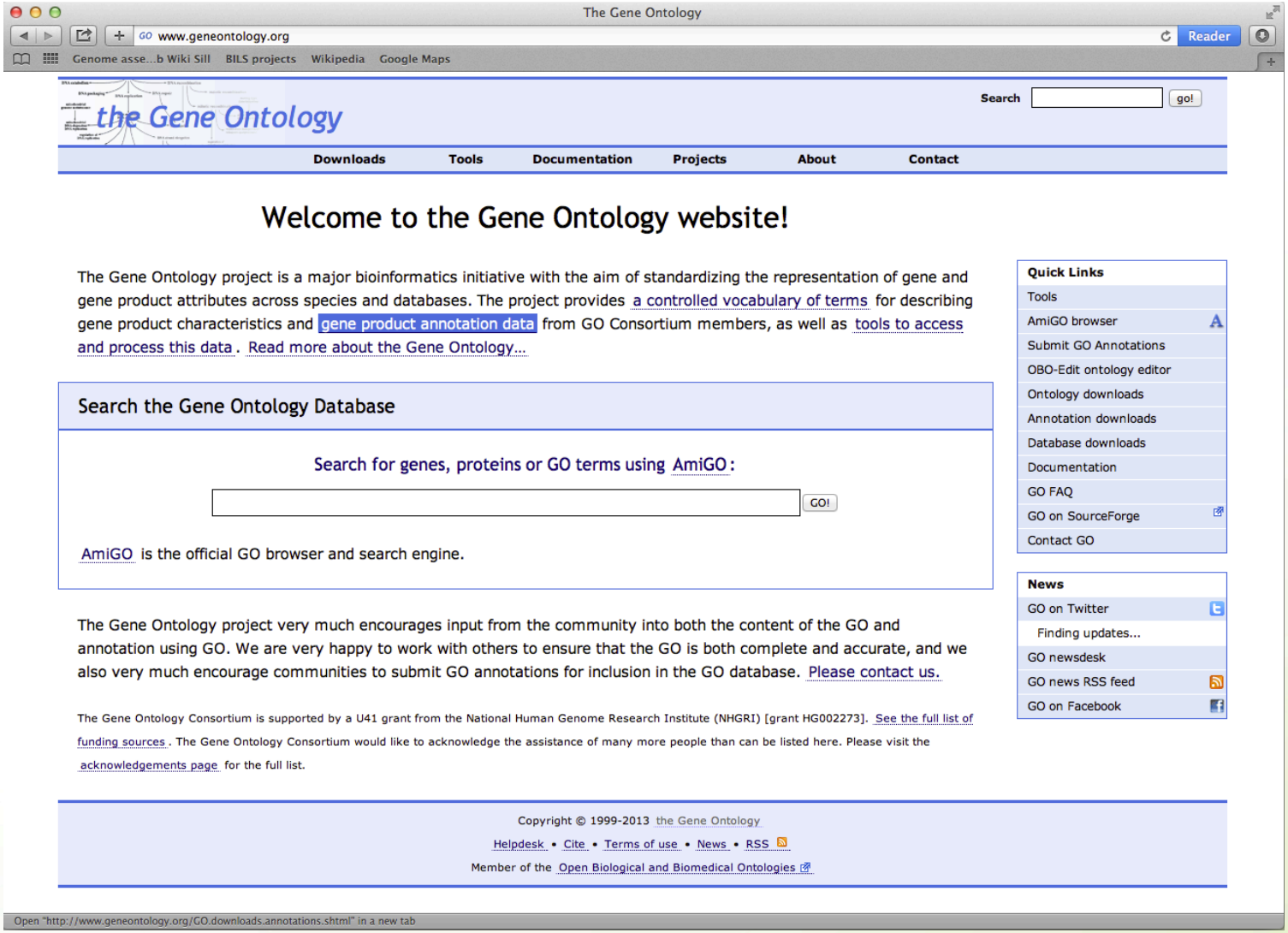
Karolinska Institutet · KTH · Stockholms universitet · Uppsala Universitet

SciLifeLab

Blast2GO

- Combines a blast-based search with a search for functional domains
- Blast at NCBI -> compares with internal database to get known GO-terms for the best blast-hits-> statistical significance test -> done!
- Interproscan

Gene Ontology

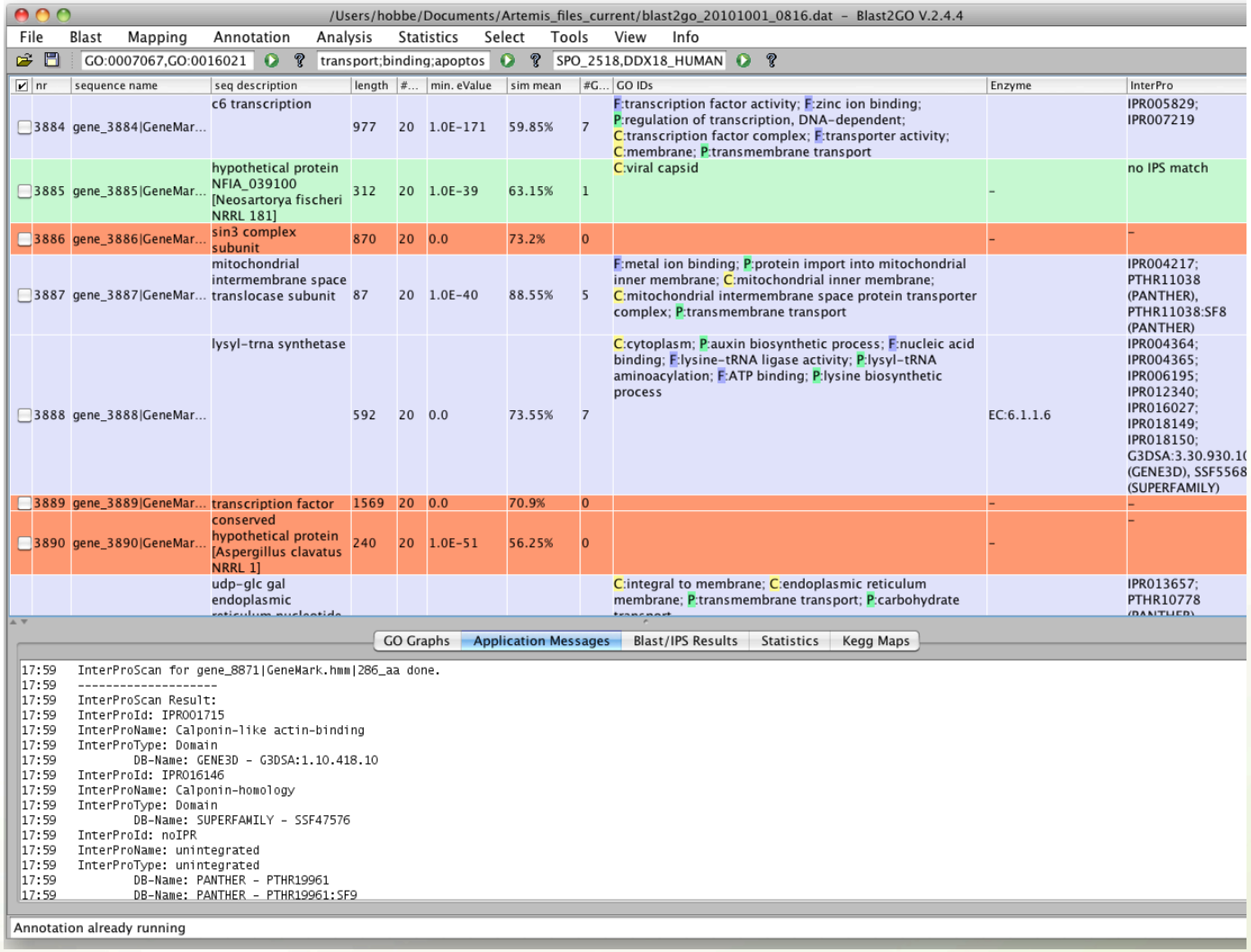


Karolinska Institutet
KTH Vetenskap och Konst
Stockholms universitet
Uppsala Universitet
SciLifeLab

Gene Ontology

- C: Cellular Component, e.g., endoplasmatic reticulum, integral to plasma membrane
- P: Biological Process, e.g., pyrimidine metabolic process
- F: Molecular Function, e.g., catalytic acticivity, transporter activity

Blast2GO

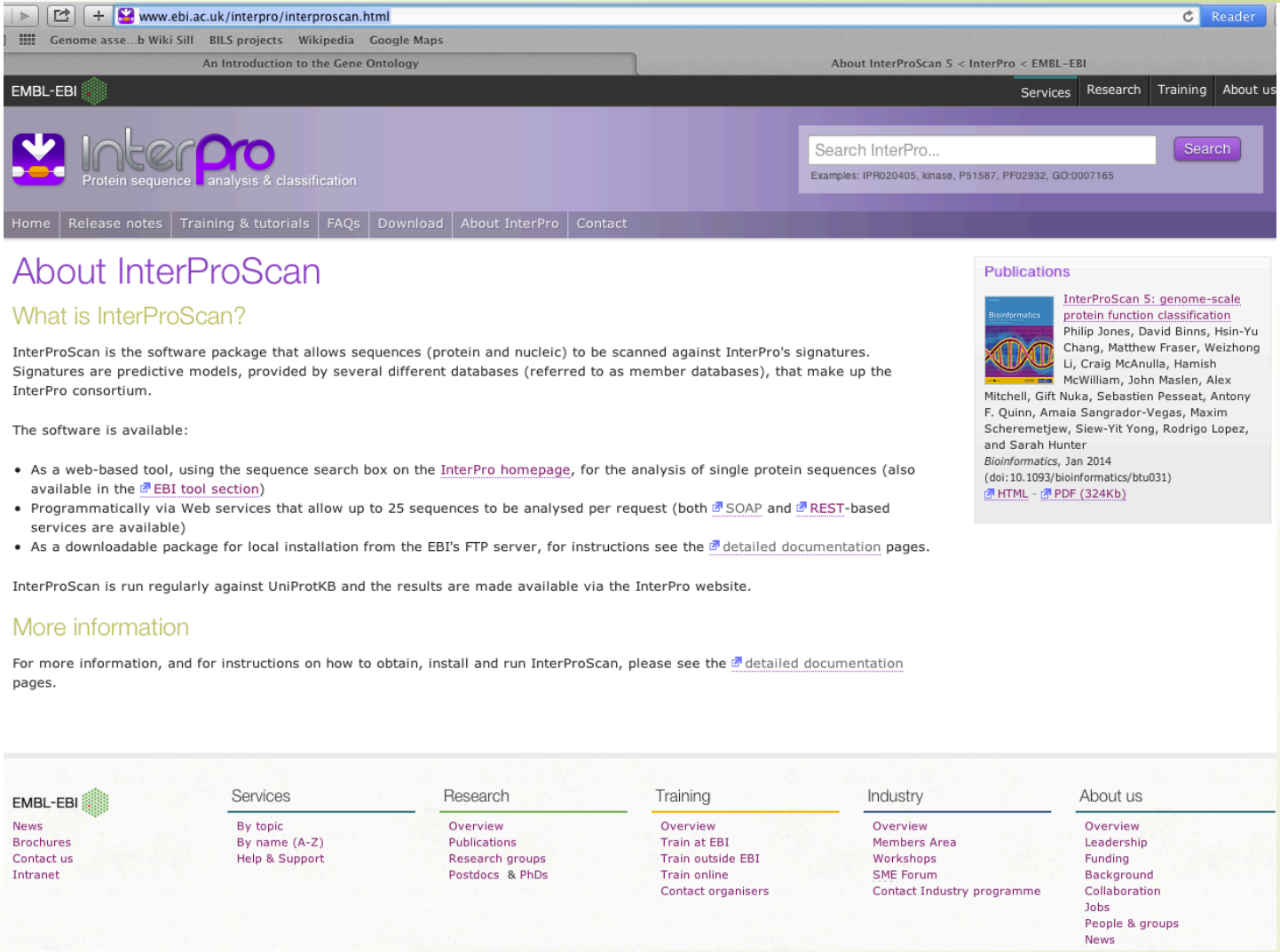


blast2go

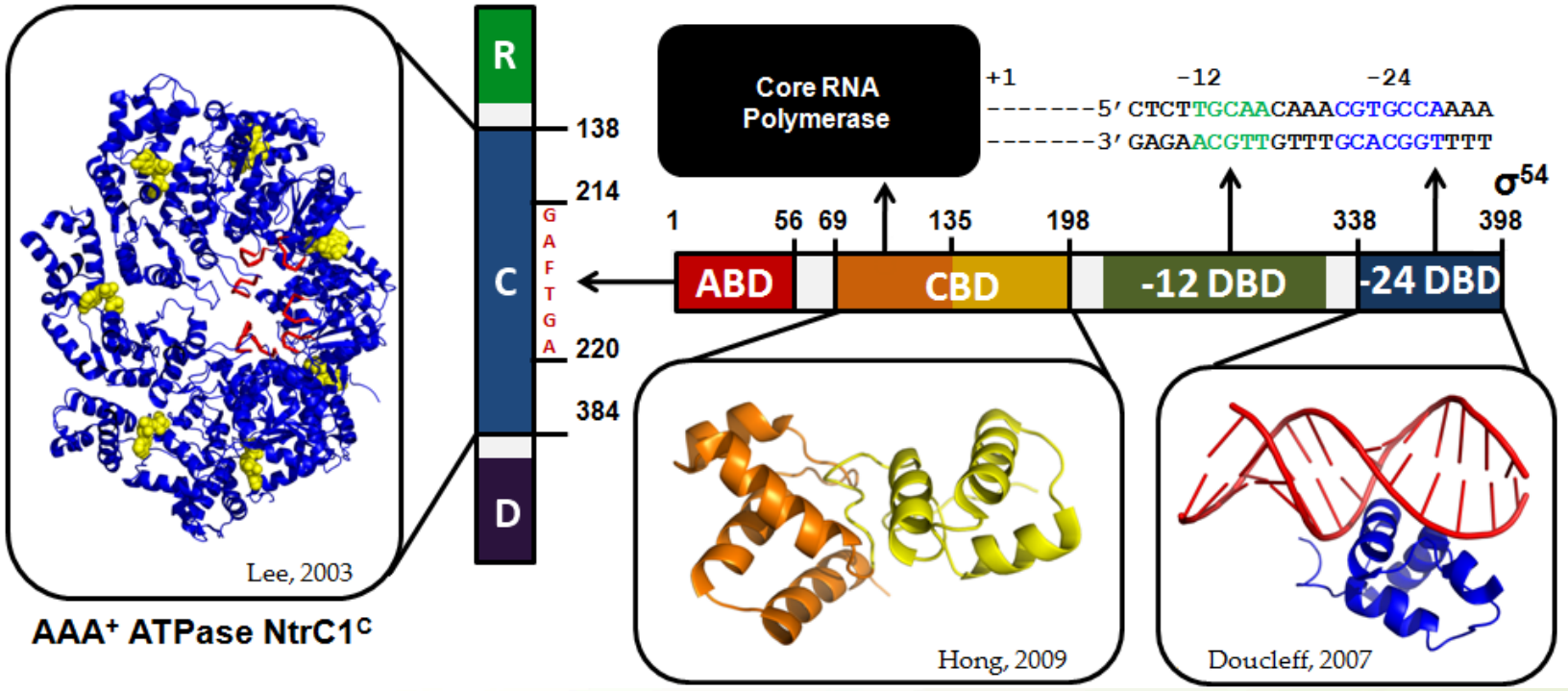
Karolinska Institutet · KTH · Stockholms universitet · Uppsala Universitet

SciLifeLab

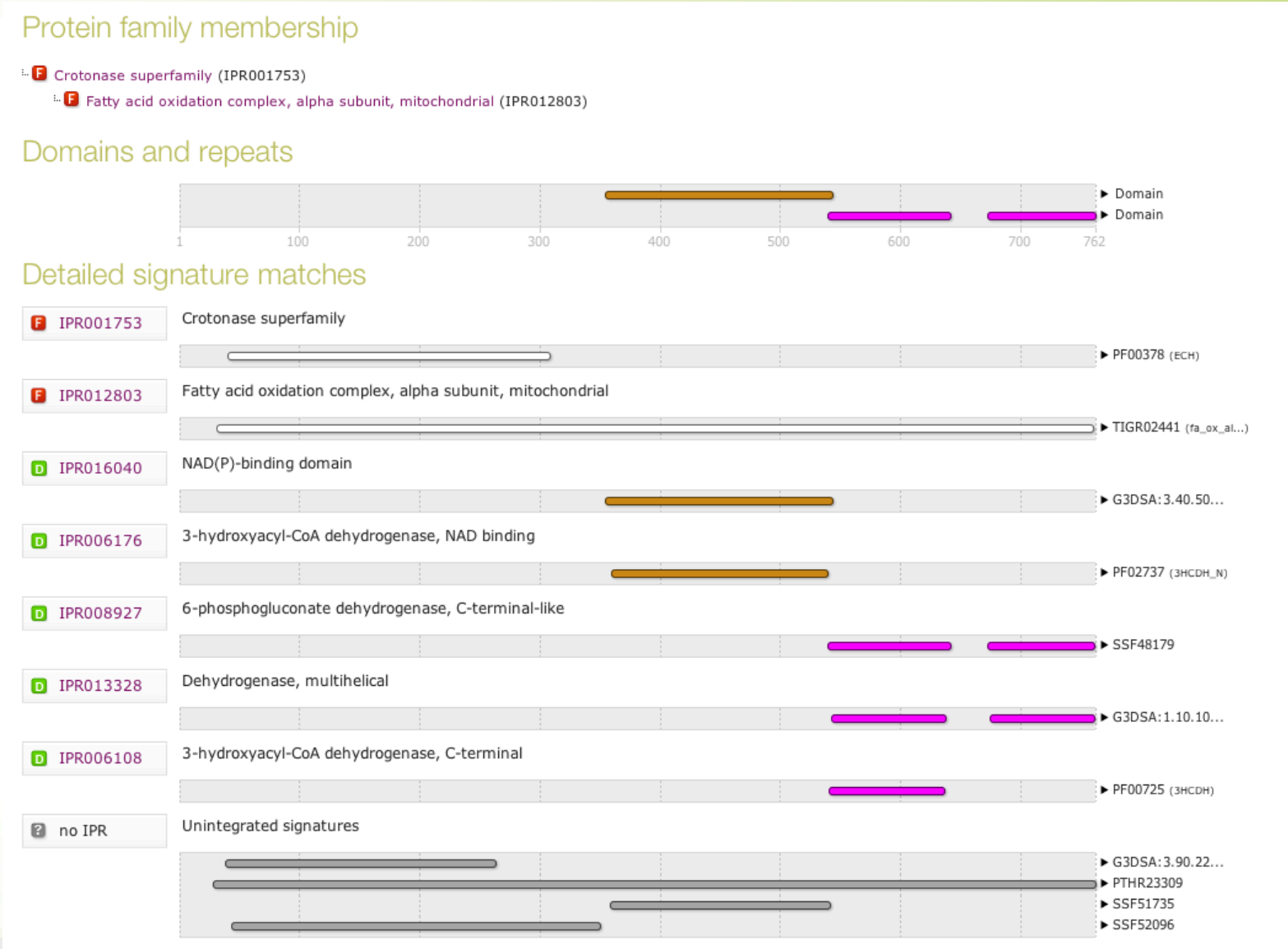
Interproscan



Sequence domains



Interproscan results



Interproscan results - GO terms

GO term prediction

Biological Process

GO:0006631 fatty acid metabolic process
GO:0006635 fatty acid beta-oxidation
GO:0008152 metabolic process
GO:0055114 oxidation-reduction process

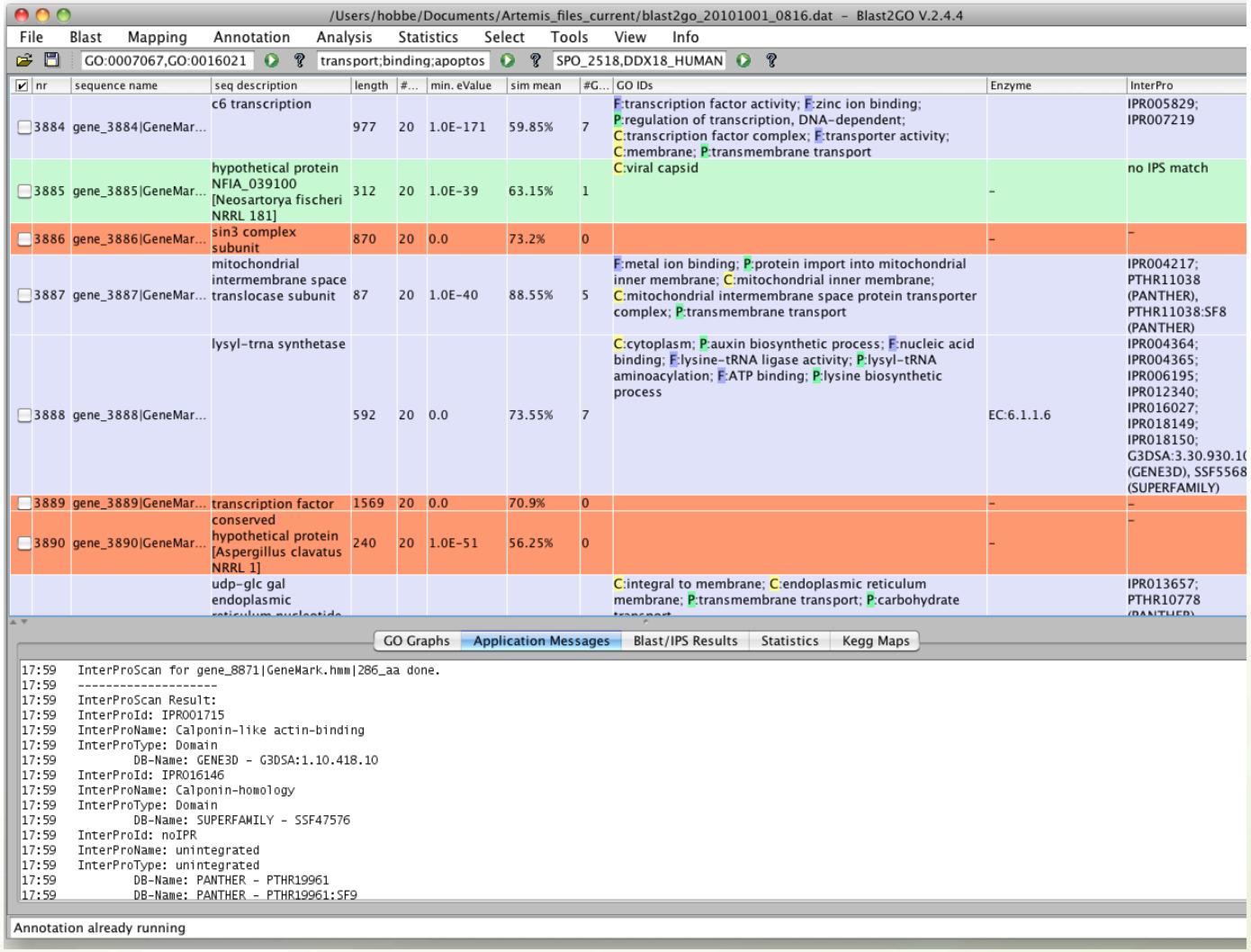
Molecular Function

GO:0003824 catalytic activity
GO:0003857 3-hydroxyacyl-CoA dehydrogenase activity
GO:0004300 enoyl-CoA hydratase activity
GO:0016491 oxidoreductase activity
GO:0016616 oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor
GO:0050662 coenzyme binding

Cellular Component

GO:0005739 mitochondrion
GO:0016507 mitochondrial fatty acid beta-oxidation multienzyme complex

Blast2GO



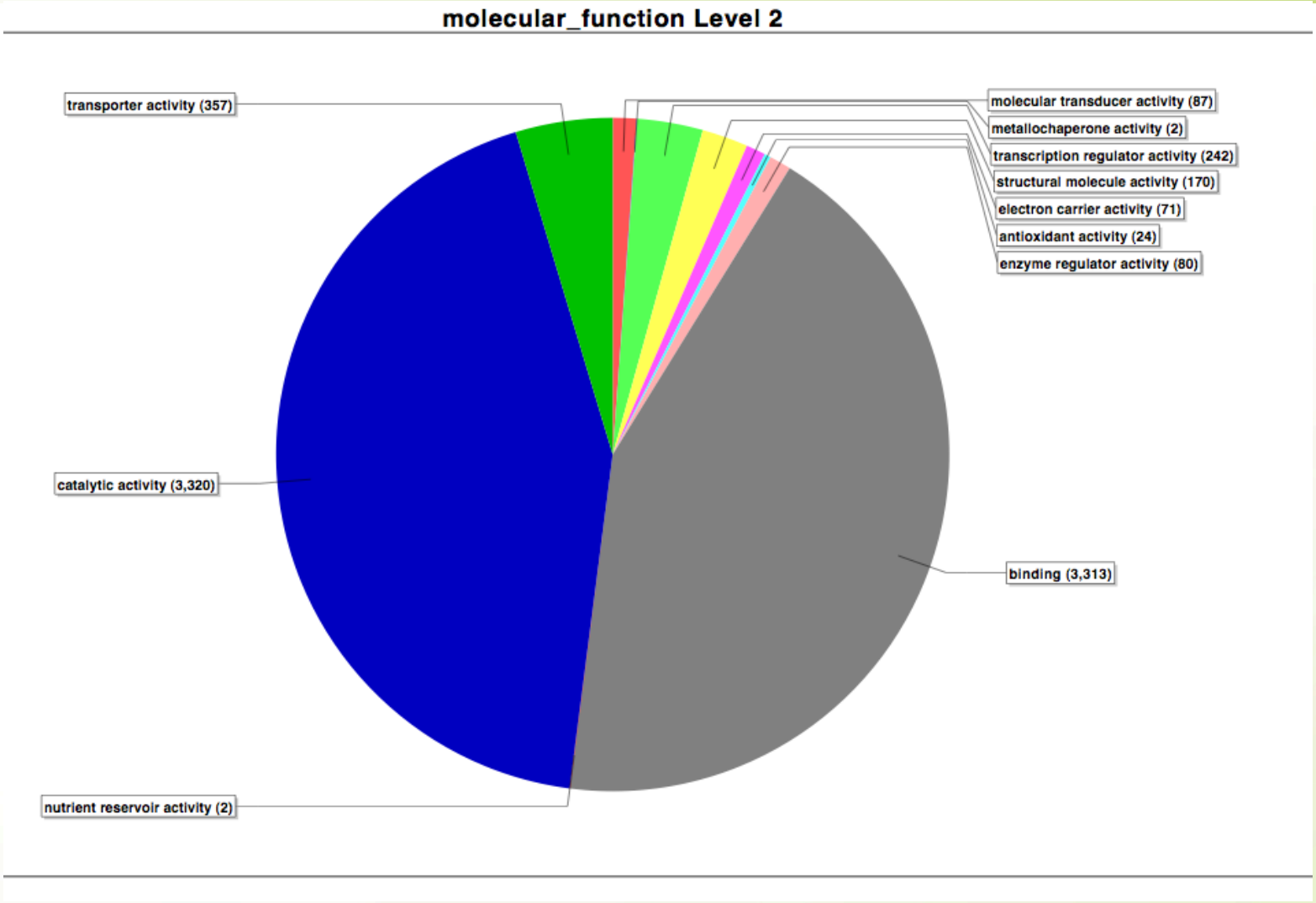
blast2go

Karolinska Institutet · KTH · Stockholms universitet · Uppsala Universitet

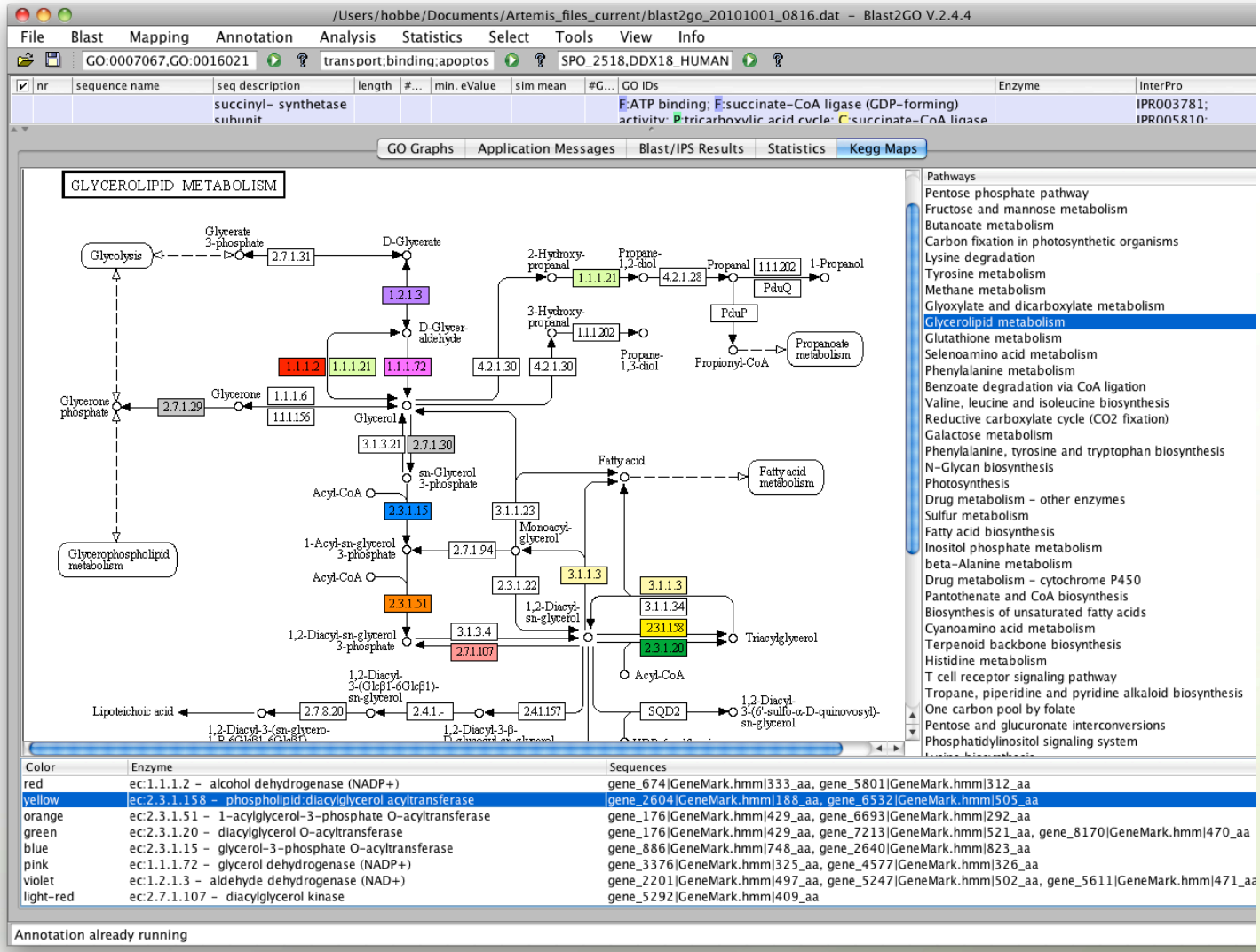
SciLifeLab

Interproscan

- Can be run on command line
- We currently combine Interproscan results with blast results using Annie -> final functional annotation


molecular_function Level 2
transporter activity (357)
molecular transducer activity (87)
metallochaperone activity (2)
transcription regulator activity (242)
structural molecule activity (170)
electron carrier activity (71)
antioxidant activity (24)
enzyme regulator activity (80)
catalytic activity (3,320)
binding (3,313)
nutrient reservoir activity (2)

KEGG-mapping

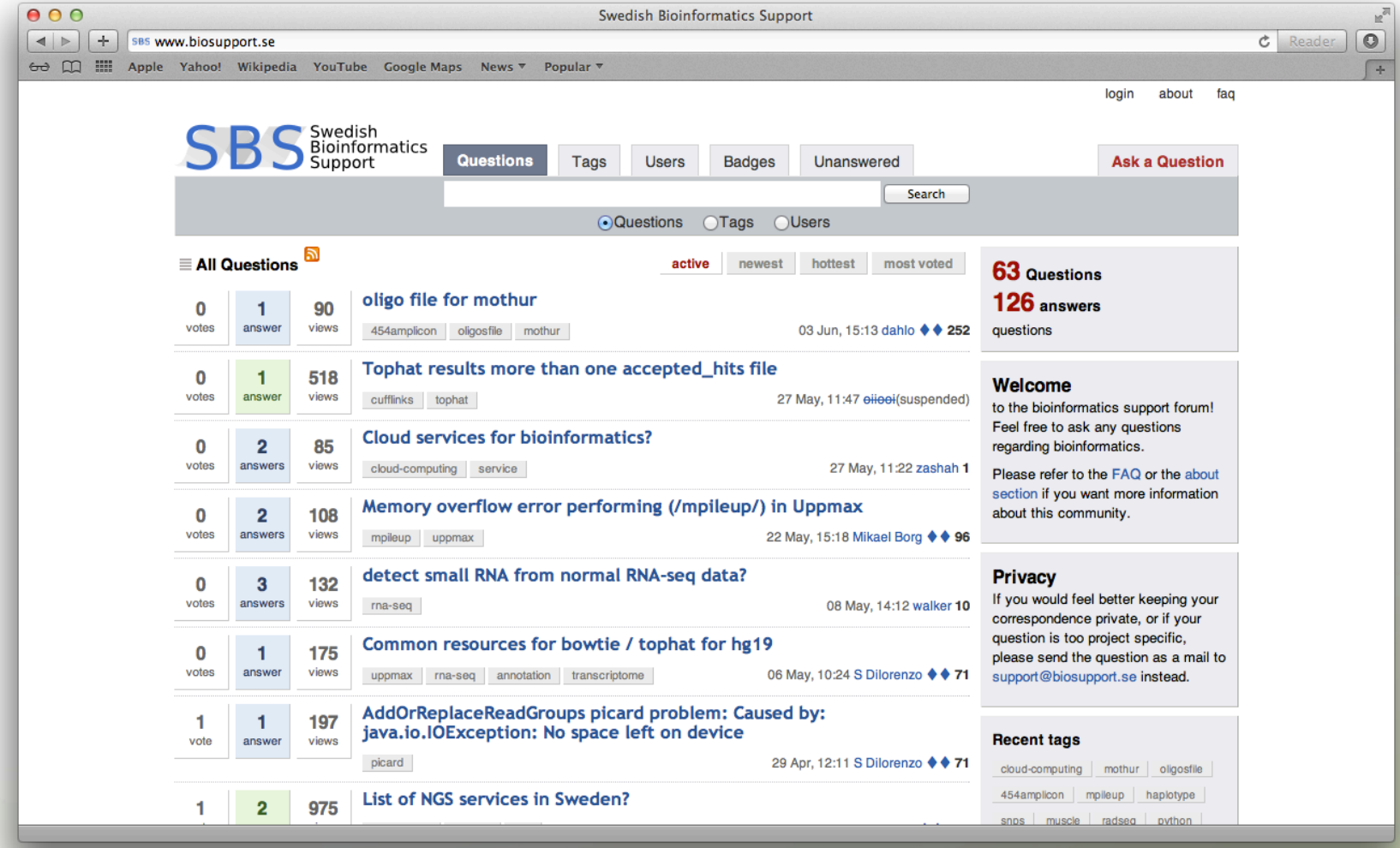


Karolinska Institutet · KTH · Stockholms universitet · Uppsala Universitet · SciLifeLab

Or get help - BILS assembly and annotation team

- Five people working with assembly and annotation
- Deliver high quality annotations
- Enable visualization and manual curation through a web interface
- Also available for consultation
- support@bils.se

Biosupport.se



Karolinska Institutet
KTH Vetenskap och konst
Stockholms universitet
Uppsala Universitet
BILS Bioinformatics Infrastructure for Life Sciences
SciLifeLab