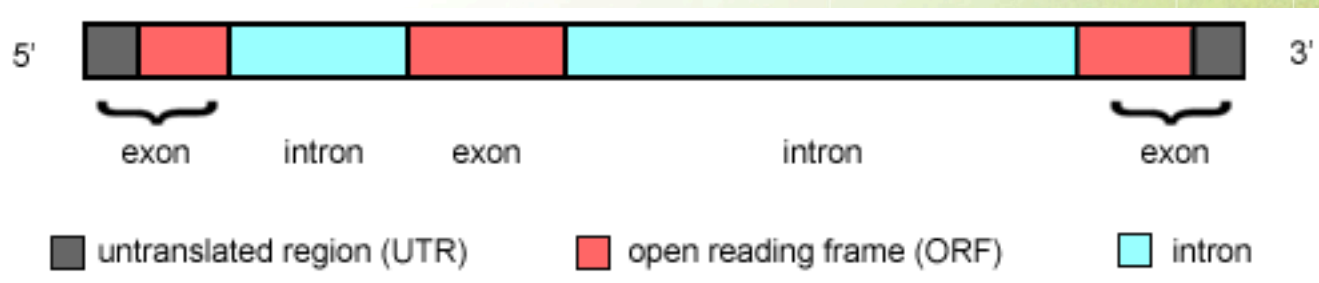


Annotation of eukaryote genomes and transcriptomes



Henrik Lantz, BILS/SciLifeLab

Enabler for Life Sciences

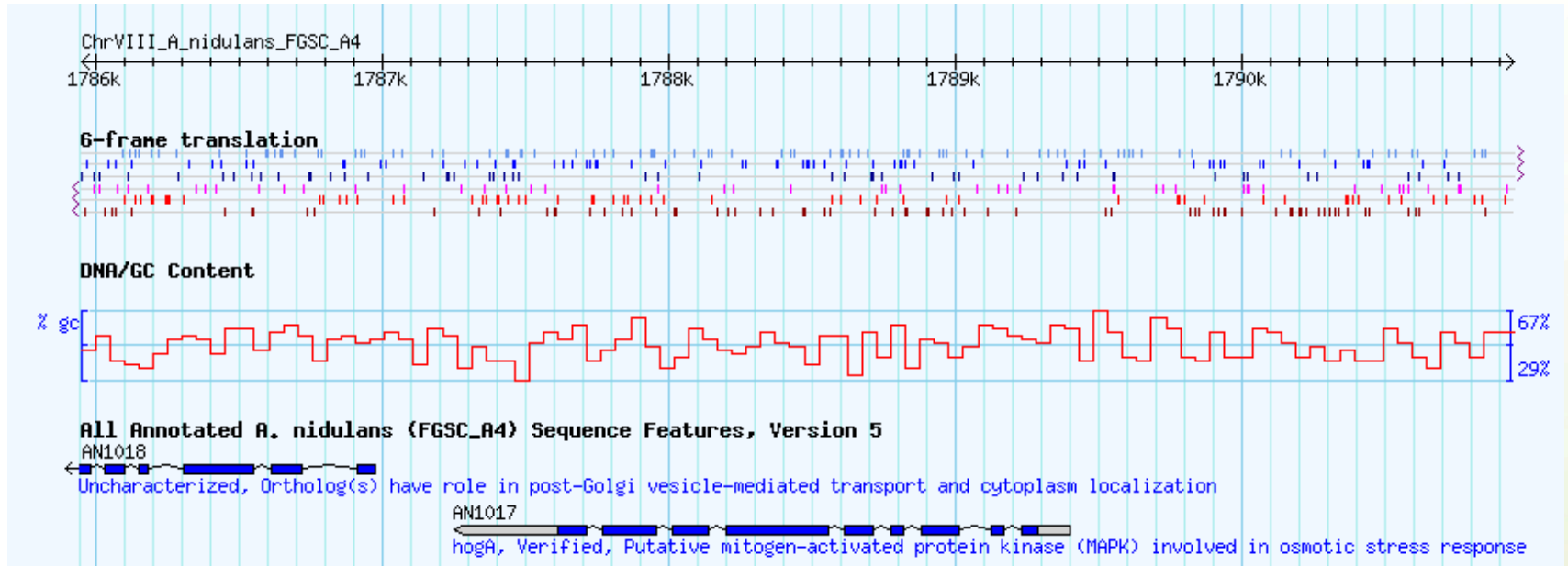
Lecture synopsis

- 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

...to an annotated gene



GFF file format

```
##gff-version 3
scaffold_7 maker gene 133848 144662 . - . ID=C554628A838E2878A71E2AD0AFB34661;Name=maker-scaffold_7-augustus-gene-0.11
scaffold_7 maker mRNA 133848 144662 . - . ID=A649E923246BADE2184E579FA9124ABD;Parent=A649E923246BADE2184E579FA9124ABD;Name=1.cornix-all_reads.72406.1_AED=1.00;_eAED=1.00;_QI=71|0|0|0|0|0|0|14|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=D38D9A5F27797F56844A2E890FF6B99;Name=maker-scaffold_7-augustus-gene-0.6
scaffold_7 maker mRNA 83101 117593 . + . ID=CF5D0A190832937C45A002E674C9C26;Parent=D38D9A5F27797F56844A2E890FF6B99;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=CF5D0A190832937C45A002E674C9C26;exon:8;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker exon 99113 99137 . + . ID=CF5D0A190832937C45A002E674C9C26;exon:9;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker exon 90664 90748 . + . ID=CF5D0A190832937C45A002E674C9C26;exon:10;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker exon 118231 118356 . + . ID=CF5D0A190832937C45A002E674C9C26;exon:11;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker exon 113609 113679 . + . ID=CF5D0A190832937C45A002E674C9C26;exon:12;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker exon 94857 94117 . + . ID=CF5D0A190832937C45A002E674C9C26;exon:13;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker exon 84578 84670 . + . ID=CF5D0A190832937C45A002E674C9C26;exon:14;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker exon 115452 115536 . + . ID=CF5D0A190832937C45A002E674C9C26;exon:15;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker exon 111579 111669 . + . ID=CF5D0A190832937C45A002E674C9C26;exon:16;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker exon 102917 103016 . + . ID=CF5D0A190832937C45A002E674C9C26;exon:17;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker exon 96766 96849 . + . ID=CF5D0A190832937C45A002E674C9C26;exon:18;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker exon 86666 86750 . + . ID=CF5D0A190832937C45A002E674C9C26;exon:19;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker exon 99944 100109 . + . ID=CF5D0A190832937C45A002E674C9C26;exon:20;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker exon 109766 109860 . + . ID=CF5D0A190832937C45A002E674C9C26;exon:21;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker exon 93154 93282 . + . ID=CF5D0A190832937C45A002E674C9C26;exon:22;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker exon 114737 114825 . + . ID=CF5D0A190832937C45A002E674C9C26;exon:23;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker exon 83101 83155 . + . ID=CF5D0A190832937C45A002E674C9C26;exon:24;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker exon 108533 108795 . + . ID=CF5D0A190832937C45A002E674C9C26;exon:25;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker exon 117477 117593 . + . ID=CF5D0A190832937C45A002E674C9C26;exon:26;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker exon 106779 106866 . + . ID=CF5D0A190832937C45A002E674C9C26;exon:27;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker exon 105743 105835 . + . ID=CF5D0A190832937C45A002E674C9C26;exon:28;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker CDS 83101 83155 . + 2 ID=CF5D0A190832937C45A002E674C9C26;cds;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker CDS 84578 84670 . + 2 ID=CF5D0A190832937C45A002E674C9C26;cds;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker CDS 86666 86750 . + 2 ID=CF5D0A190832937C45A002E674C9C26;cds;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker CDS 90664 90748 . + 1 ID=CF5D0A190832937C45A002E674C9C26;cds;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker CDS 93154 93282 . + 0 ID=CF5D0A190832937C45A002E674C9C26;cds;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker CDS 94857 94117 . + 0 ID=CF5D0A190832937C45A002E674C9C26;cds;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker CDS 95748 95871 . + 2 ID=CF5D0A190832937C45A002E674C9C26;cds;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker CDS 96766 96849 . + 1 ID=CF5D0A190832937C45A002E674C9C26;cds;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker CDS 99113 99137 . + 1 ID=CF5D0A190832937C45A002E674C9C26;cds;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker CDS 99944 100109 . + 0 ID=CF5D0A190832937C45A002E674C9C26;cds;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker CDS 102917 103016 . + 2 ID=CF5D0A190832937C45A002E674C9C26;cds;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker CDS 105743 105835 . + 0 ID=CF5D0A190832937C45A002E674C9C26;cds;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker CDS 106779 106866 . + 1 ID=CF5D0A190832937C45A002E674C9C26;cds;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker CDS 108533 108795 . + 0 ID=CF5D0A190832937C45A002E674C9C26;cds;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker CDS 109766 109860 . + 1 ID=CF5D0A190832937C45A002E674C9C26;cds;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker CDS 110231 110356 . + 2 ID=CF5D0A190832937C45A002E674C9C26;cds;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker CDS 111579 111669 . + 2 ID=CF5D0A190832937C45A002E674C9C26;cds;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker CDS 113609 113679 . + 1 ID=CF5D0A190832937C45A002E674C9C26;cds;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker CDS 114737 114825 . + 2 ID=CF5D0A190832937C45A002E674C9C26;cds;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker CDS 115452 115536 . + 0 ID=CF5D0A190832937C45A002E674C9C26;cds;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker CDS 117477 117589 . + 2 ID=CF5D0A190832937C45A002E674C9C26;cds;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker three_prime_UTR 117590 117593 . + . ID=CF5D0A190832937C45A002E674C9C26;three_prime utr;Parent=CF5D0A190832937C45A002E674C9C26
scaffold_7 maker gene 22451 37514 . - . ID=E1E94A56071E5D7940C938E70F79E56;Name=E1E94A56071E5D7940C938E70F79E56
scaffold_7 maker mRNA 22451 37514 . - . ID=D28B3792B8FCDB81C7A4070E76500C5;Parent=E1E94A56071E5D7940C938E70F79E56;Name=scaffold_7.22450-37514_AED=1.00;_eAED=1.00;_QI=3008|0|0|0|0|0|11|762|144
scaffold_7 maker mRNA 26810 37514 . - . ID=D629810725F8971755858A23623AF6A;Parent=E1E94A56071E5D7940C938E70F79E56;Name=maker-scaffold_7-augustus-gene-0.10-mRNA-1_AED=1.00;_eAED=1.00;_QI=3008|0|0|0|0|0|0|10|0|400
scaffold_7 maker exon 26810 27088 . - . ID=D629810725F8971755858A23623AF6A;exon:40;Parent=D629810725F8971755858A23623AF6A
scaffold_7 maker exon 27950 28090 . - . ID=D28B3792B8FCDB81C7A4070E76500C5;exon:39;Parent=D28B3792B8FCDB81C7A4070E76500C5.10629810725F8971755858A23623AF6A
scaffold_7 maker exon 33513 33610 . - . ID=D28B3792B8FCDB81C7A4070E76500C5;exon:38;Parent=D28B3792B8FCDB81C7A4070E76500C5.10629810725F8971755858A23623AF6A
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 "1.178011"; 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 "1.178011"; 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 "16.665534";
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 "16.665534";
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 "2.510133"; conf_hi "9.099893"; cov "160.418399";
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 "2.510133"; conf_hi "9.099893"; cov "160.418399";
Sb_20131119_contig_1 Cufflinks transcript 10522 13208 1000 . . gene_id "CUFF.23"; transcript_id "CUFF.23.1"; FPKM "55.637793473"; 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 "55.637793473"; 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 "52.757856123"; 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 "52.757856123"; frac "1.000000"; conf_lo "27.382285"; conf_hi "37.895127"; cov "66.397320";
Sb_20131119_contig_100023 Cufflinks transcript 1097 2089 1000 . . gene_id "CUFF.9"; transcript_id "CUFF.9.1"; FPKM "7.426254644"; frac "1.000000"; conf_lo "4.747632"; conf_hi "7.830684"; cov "16.282075";
Sb_20131119_contig_100023 Cufflinks exon 1097 2089 1000 . . gene_id "CUFF.9"; transcript_id "CUFF.9.1"; exon_number "1"; FPKM "7.426254644"; frac "1.000000"; conf_lo "4.747632"; conf_hi "7.830684"; 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 "233.013864257"; 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 "233.013864257"; frac "1.000000"; conf_lo "22.182837"; conf_hi "38.203775"; cov "306.188212";
Sb_20131119_contig_100040 Cufflinks transcript 2 255 1000 . . gene_id "CUFF.10"; transcript_id "CUFF.10.2"; FPKM "59.99083872"; frac "1.000000"; conf_lo "3.216803"; conf_hi "9.91847"; cov "41.309085";
Sb_20131119_contig_100040 Cufflinks exon 2 255 1000 . . gene_id "CUFF.10"; transcript_id "CUFF.10.2"; exon_number "2"; FPKM "59.99083872"; frac "1.000000"; conf_lo "3.216803"; conf_hi "9.91847"; 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 "14.786800"; 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 "14.786800"; cov "49.916331";
Sb_20131119_contig_100107 Cufflinks transcript 430 902 1000 . . gene_id "CUFF.13"; transcript_id "CUFF.13.1"; FPKM "25.050052879"; 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 "25.050052879"; 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 "23.432223"; 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 "23.432223"; 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 "53.966266"; conf_hi "84.228598"; cov "4520.868745";
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 "53.966266"; conf_hi "84.228598"; cov "4520.868745";
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 2276 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_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 "11.688977"; 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 "11.688977"; 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 1611 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 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_100080 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 "33.804660"; 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 "33.804660"; 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 "5.4971611200"; frac "1.000000"; conf_lo "4.259353"; 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 "5.4971611200"; frac "1.000000"; conf_lo "4.259353"; conf_hi "6.262152"; cov "11.356891";
Sb_20131119_contig_10041 Cufflinks transcript 345 3072 1000 . . gene_id "CUFF.36"; transcript_id "CUFF.36.1"; FPKM "55.9644866986"; frac "1.000000"; conf_lo "45.789639"; conf_hi "53.715558"; cov "119.323264";
Sb_20131119_contig_10041 Cufflinks exon 345 3072 1000 . . gene_id "CUFF.36"; transcript_id "CUFF.36.1"; exon_number "1"; FPKM "55.9644866986"; frac "1.000000"; conf_lo "45.789639"; 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 "5.2492326693"; frac "1.000000"; conf_lo "3.827818"; 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 "5.2492326693"; frac "1.000000"; conf_lo "3.827818"; 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";

```


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

Mapped reads - condition 1

Genome

Mapped reads - condition 2

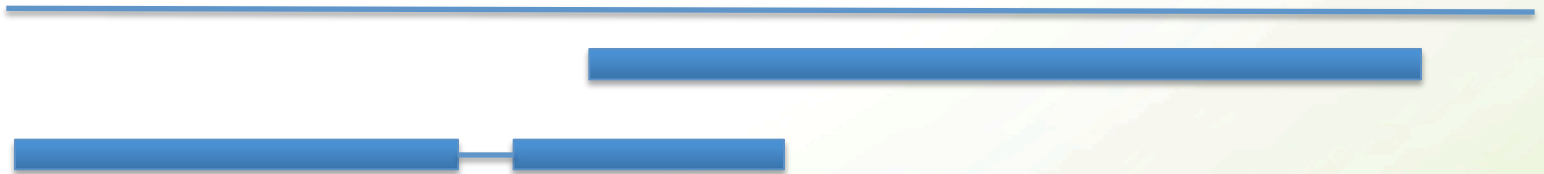


Why is annotation important?

RNA-seq reads

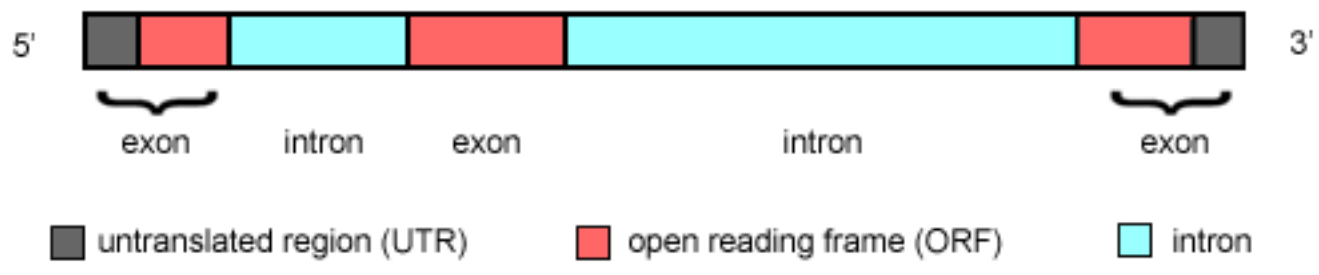


Genome



There are two major parts of annotation

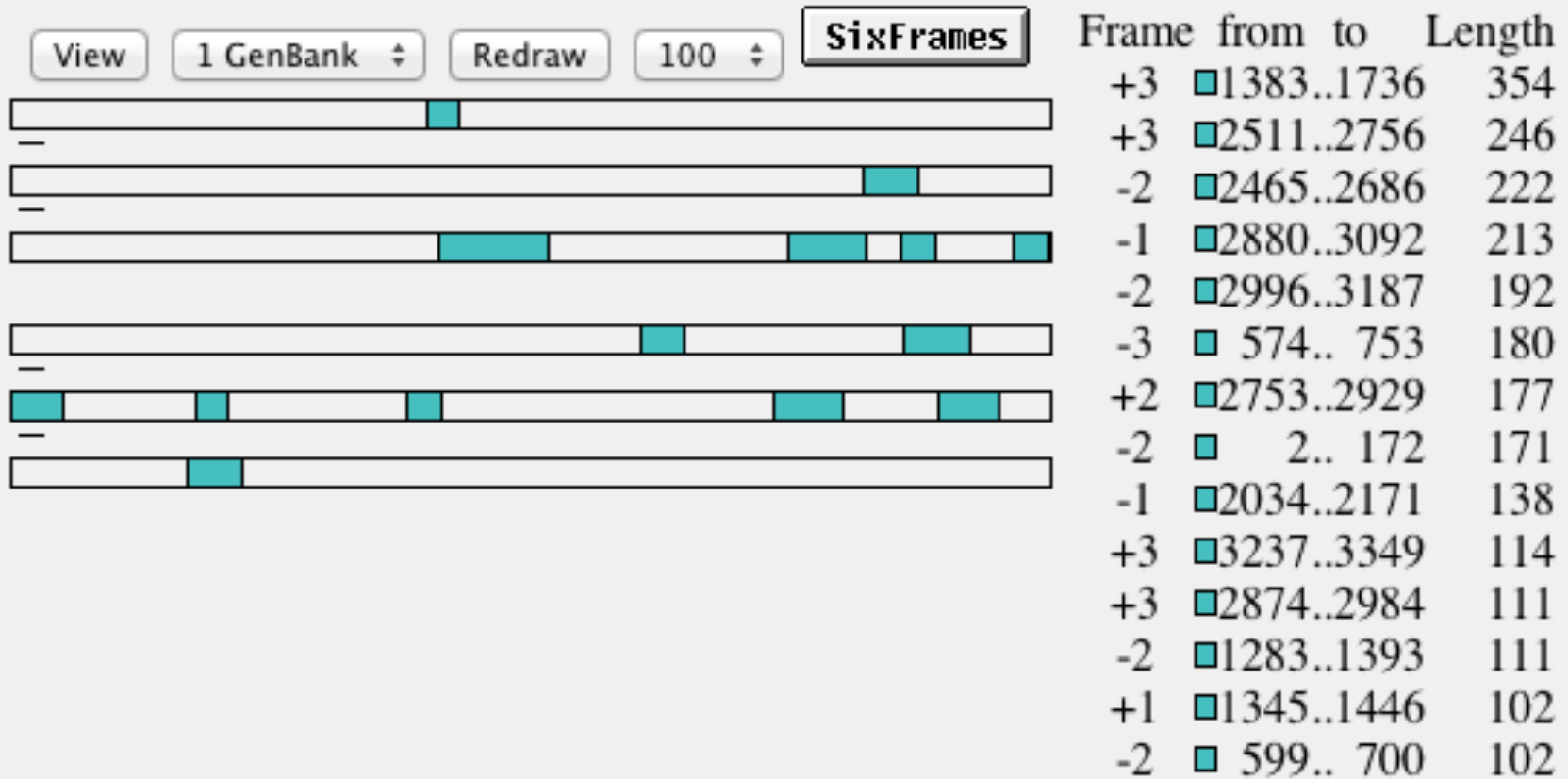
- 1) Structural: Find out where the regions of interest (usually genes) are in the genome and what they look like. How many exons/introns? UTRs? Isoforms?



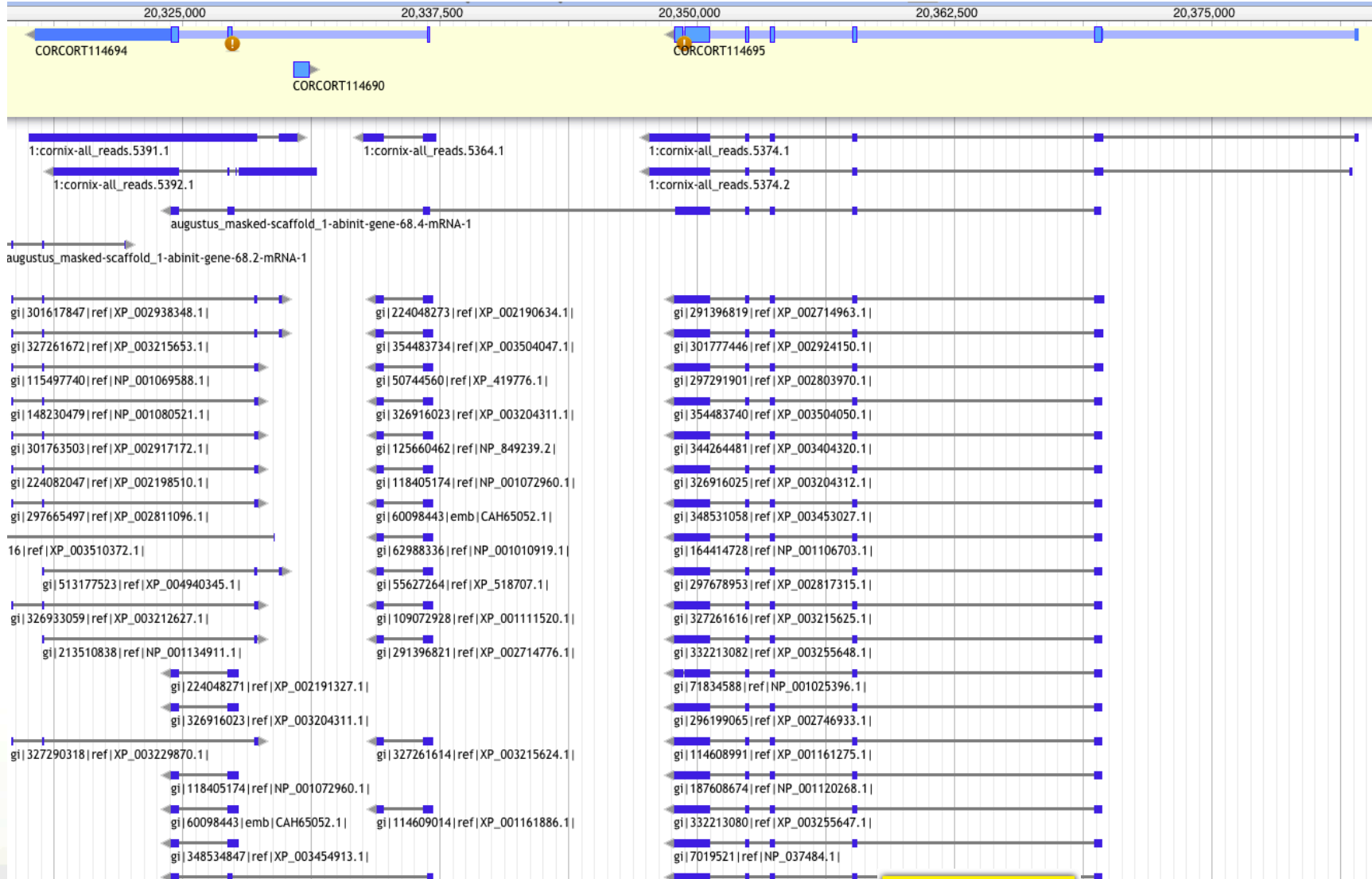
- 2) Functional: Find out what the regions do. What do they code for?

Open reading frames

Anonymous



Difficult in practice



Combine data - use Maker!

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



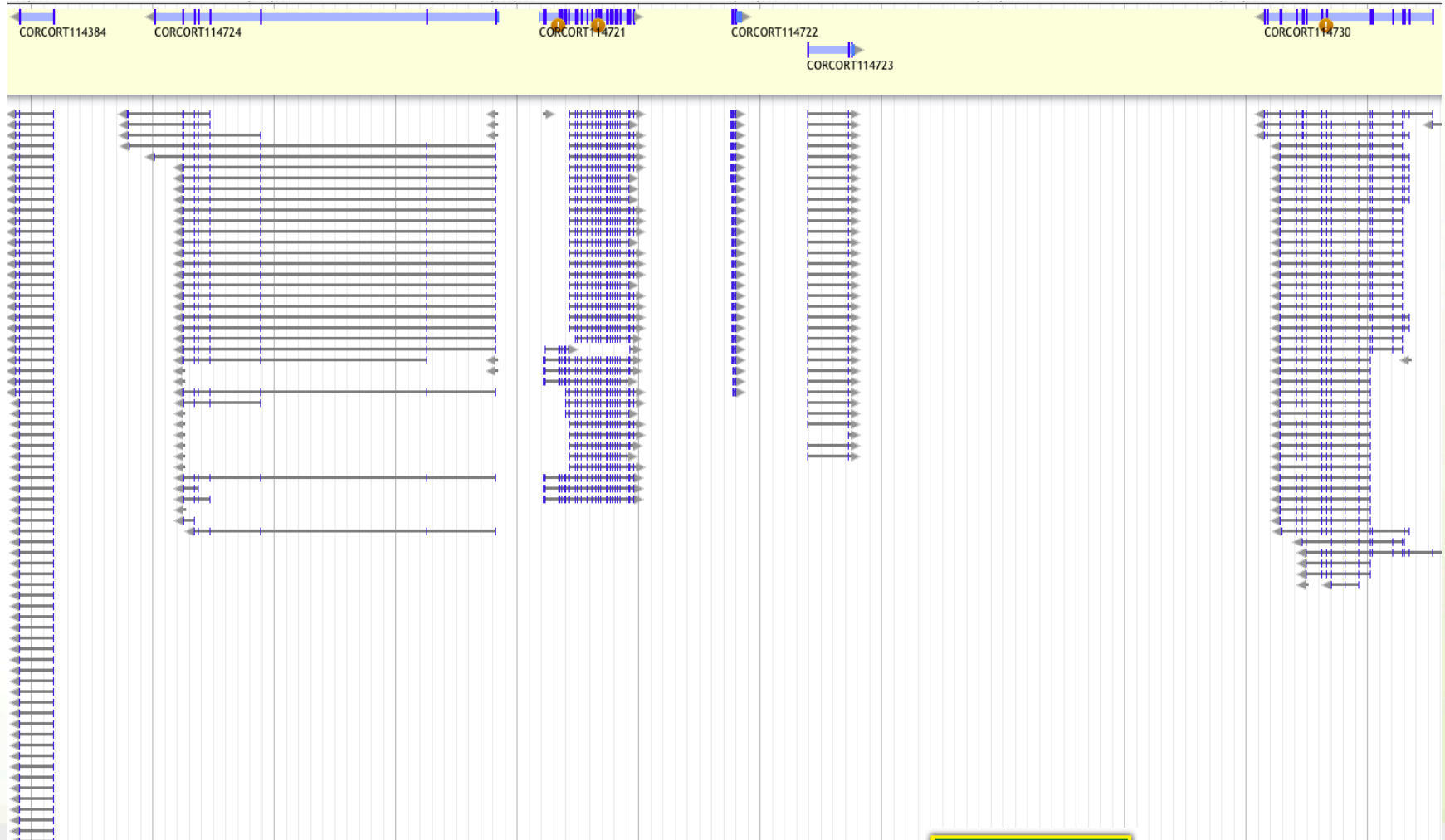
Combined annotation

Transcriptomes are different but have their own challenges

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

```
>asmbL_2719
AGCACCTAGACAGGATGGGAGGCTCTCCCTTGCTGGCAGAGGCAGATCTCCTTTCC
AACACCTAGCAGTATGAACCTAGTGAGCTCCTGACTGTTTTCCAGTGGTAAATGAGGTGTGA
CCCGCTGCAGCTGCACACTGAATTTCTCAGTTCCTCCAGGCGCAGCCAGCAGTGTGGGG
AATGCTTTGTTTGTGTGCTGTTGACCATTC
>asmbL_2702
GTCTGCACCTGGGAATGCCCTGGAGCAGAACCATTGCCATGGATAAGGACACTACATTT
CCTGGTGTAAAGGTGAATATAACCTCCAGGTTAAGGTGACATTAATTTCAATTACAGCT
TGCTCTTTGAAGCTAAGCAGTTAATCAACAAGCTATACTGTGACTACACCCCTTAGATCA
ATAGCTGGGAAAACATCACCTCCCAAACTCCACCTCTTAACTGCACCTTTTGAAG
AAGTACAGGCCAGAGTTTAGCTGATCCATCCCTGTGGCTAATCGTCTCTTACAAGCTG
CAATATTTTTAAAAACAGCAATTTGGTAGAGGTTTAAACATCAGCCAGCTGTTCAATT
TACAGCAGGTTAAGCATTCTGAAACTGTGATCACTGATATATTTGGTGCAGTCAGATGT
CTTGTAGTGCTT
>asmbL_2701
ACAAAACAAAACAAAATAAAAACAAAGGAAAACAGCAAAAAAACCATCATACAATCCCATG
TGCCAAAGAGCTTTACTGTGAAATCAACTATGGAGTCAAAAACATAGAAAAGCTTCCAGA
TTTTGTATTCCAGGCTGAGACAAGTTTGTAAATCTTCCAGAAATGGCAACAAGCTG
CAGGGTAAACATCTAATGCACACCTCCCTGATACGAAATGCAGAGCACCTTAACTTCT
CAGCCCTCCCAAGTCAACAACAGCTATAAATCTGCCCTTCACTTGTGGAATATCTCA
TCATAAGGGAAGCATTTTTTAGGTGAGAAATCAAAATCCACCTTGACGGAGCCGGTCA
GCATATACATGGGCTATGCTGCTAGGTTTGTACCAAGCACTCTAGTGTGAGAATAA
CTTAGAGTGACCTAAGCAGGTAACATTTTTGCACACTAACTTGTGCCAGTATCGTTTA
TTCCAAACTCCCACTTTTCCCAAGAGAAAACAGCTGTATTGGCAGTAGCAGTGTGTTT
AAGGTAACCTGCACCTGTACTAGTAGCTTCCAGGCACAACCTTCCACACTAGCCAG
CTAGTCTAAGTAACTTCTTGGCAACAGGAAGAACTGAAACACACAGGCCACACTTGAAG
AGGATCTGAGCTGAGCTGCCTTTTCTCCAGGAGCCATGGGTTCCAGGCAGTACAGAAG
GCAGCATAAGGTGCTCTCACCACCAAGTAAAGCTGGCACCAGAGAGGCTGCATCAGGAAA
ACCCACCATCAGCACAAAAGGAGCCCTGCAAACTCAGCCAGTGTAGGTTACTGGGGTGTGG
AGAATCAATACTGCCCTGATGGAAGCTCCTGATACCCACATTTTCCCTCATCCAGTGA
CAGAACACAAAGAGAGGAAATGTGGAGGACAGGAATGTGCAGCACTGAGGAAGCAGGG
CATCATTTTGTCTCAGCCTGTCTGCAGCAGCTTTCACATGGCCAGGGCAGTCTGAGTCC
TACCGGTGGAGGCACATTGTTCCATGACTCAATGCCCTCTCTGACAGCAATCTCAAG
TGGTCCCTTTAAAAATGGCTCTCACTACTTTGGGAGCTCACTGGCACCAGCTCACTGCCA
GGAACCAAAGGTGCTAACCGGGGTGGGAACAAATATTCTGGACAGTTGAGGAAATGC
TGGATAGAAAACAGAGGTGTTTGGTAACTGACTGATAAAGAGAGAGAGTGCAGATAGAG
CTGAAGAAGTACTCCAGGTGGGAAACAGCTGTATAAAAAGTCTTAAAGGGGTGAAATGTA
GAAAATAATGCCGGAGCAGAATAAAGGACTTATTTCCATCCCACTGGAATCCTGA
ACCCAGTTCAGAGTAAATGAAGGGCTTTGTGTGTGTGCTAGTGAGAGAGATCACCATG
AAGCAATAGCTCAGGCTCACCCCTGCACTCTCCAGGAAAGGAGCTCACAGCCCTGAGA
GGTTGATGGGCTGTCCAGCAGCCACAGCTTGGCATTAGGTTGTTAGTGTGGCTTT
GGGTATGCACATAAAACCAAGTTGAATGGAAAGTGTCTGTCAGTACTCAGTGAAGGGAGA
GAACATTTCCAGGACCTGGGAACTACTGGAGGGGACTGACTAATTTTGGTGTGTTGTTG
GGTCTGTGCTGGGAAATAAATCCATGGCAATGCCTGAAAGTGTGGGGCGGAGGCGTT
GCCTGAAAGCCTTGAGCTGTGGGAGATACACAAAAGATATGCAATGACTGACACAAGC
CAGAACTTGCCTGAGCCAGAGGGTGCCTACTCTGGCTGGGACAGTCTCCCTGTGGCAG
GCTACAGTGCATCCCTCTGTGAGCTGACGCTCAGCCATGCAGGATGCTCCCTCTGTG
CTGGAGAGCACTGGCACACCTCTGGGAGATTTGGAGCTGCTAATAGTTGCAGGCTCTGG
TGAATGGAGACTGGCTCTGTGTGGTTTGGCAGCCTTCTTCCAGCAGAGCTGTAG
CTGAGGCAGGGCCACTCAACTCCAGCATAGATACAGTGTCCAGAAAAGTAAAGTCCCAT
CTGCTCCCGGTGATCCTAACCTGTCAAAGCCATTAGAGTTGGCATTGTCTTGAAT
crow_gonads.assemblies.fasta
```

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

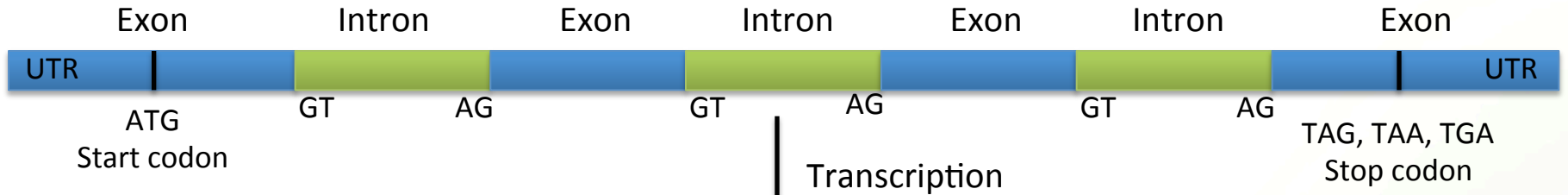
```
>ENSTGUP00000017616 pep:novel chromosome:taeGut3.2.4:8_random:2849599:2959678:-1 gene:ENSTGUG00000017338 transcript:ENSTGUT000000180
RSPNATEYNWHLRYPKIPERLNPPAAAGPALSTAEGWMLPWGNGQHPLLARAPGKGRER
DGKELIKKPKTFKFTFLKFKKFKKFKKFKKTFK
>ENSTGUP00000017615 pep:novel chromosome:taeGut3.2.4:23_random:205321:209117:1 gene:ENSTGUG00000017337 transcript:ENSTGUT00000018017
PDLRELVLVLMFEHLHRVNRNGGFRNSEVKKWPDRSPPPYHSFTPAQKSFSLAGCSGESTKMG
IKERMRLSSSRQQRGRQQLGPPPLHRSPSPEDVAEATSPTKVQKSWSFNDRTRFRASL
RLKPRIPAEGDCPPEDSGEERSSPCDLTFEDIMPAVKTLIRAVRILKFLVAKRKFKETLR
PYDVKDVEQYSAGHLDMLGRIKSLQTRVEQIVGRDRALPADKKVREKGEKPALEAELVD
ELSMMGRVVKVERQVQSIEHKLDLLGLYSRCLRKGSANSLVLA AVRVPPEPDPVTSYQ
SPVEHEDISTSQAQSLISRLASTNMD
```


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

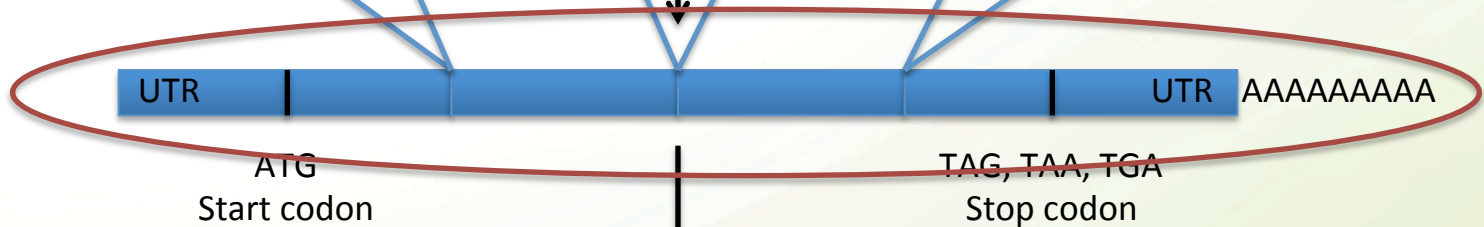
DNA



Pre-mRNA



mRNA



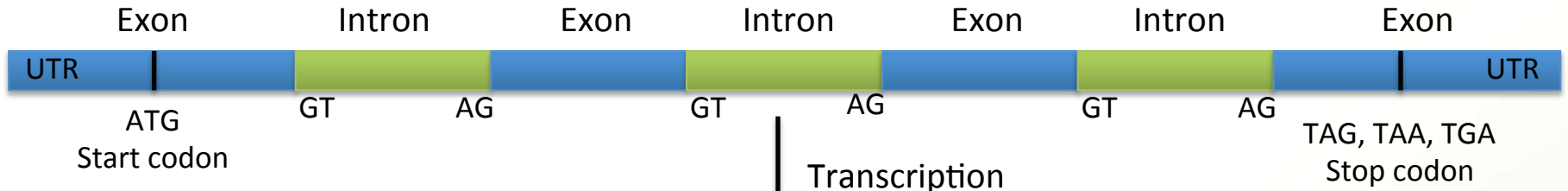
Translation

Data used - RNA-seq

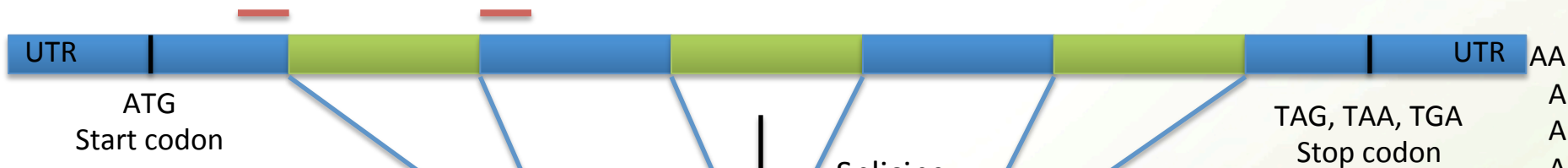
- Should always be included in an annotation project
- From the same organism as the genomic data
=> unbiased
- Can be very noisy (tissue/species dependent),
can include pre-mRNA
- PASA, or some other filtering method, often
needed

Spliced reads

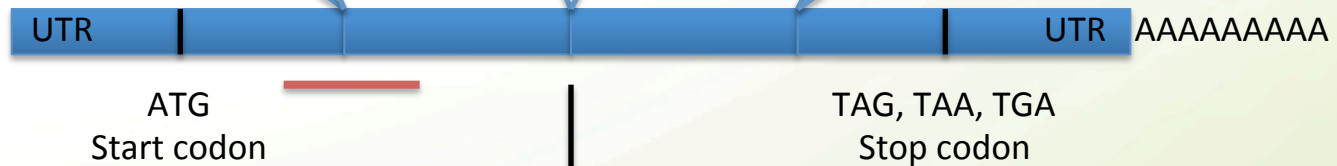
DNA



Pre-mRNA

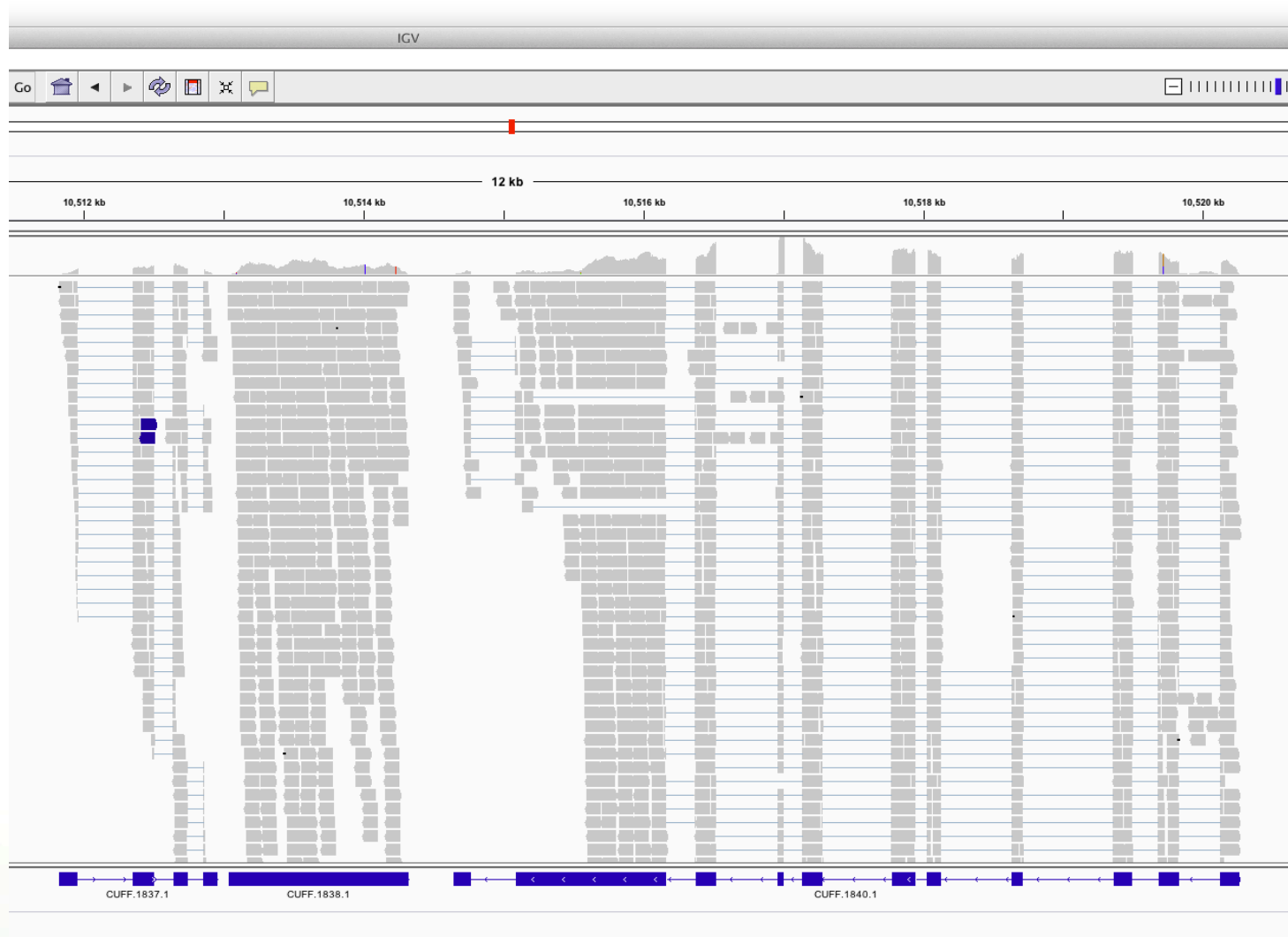


mRNA



Translation

RNA-seq - Spliced reads

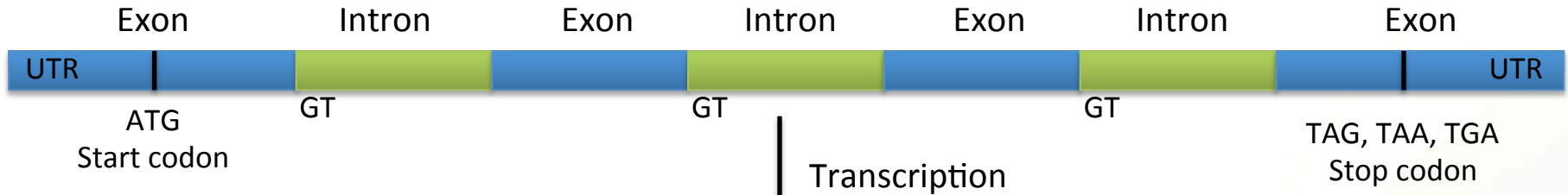


Pre-mRNA



Pre-mRNA

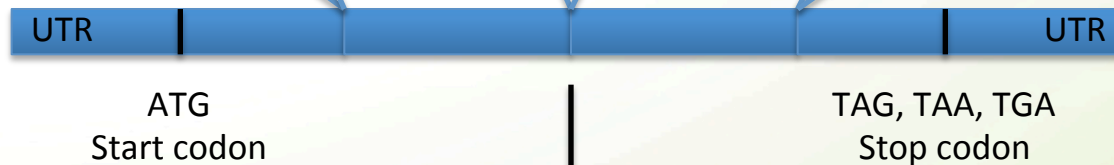
DNA



Pre-mRNA

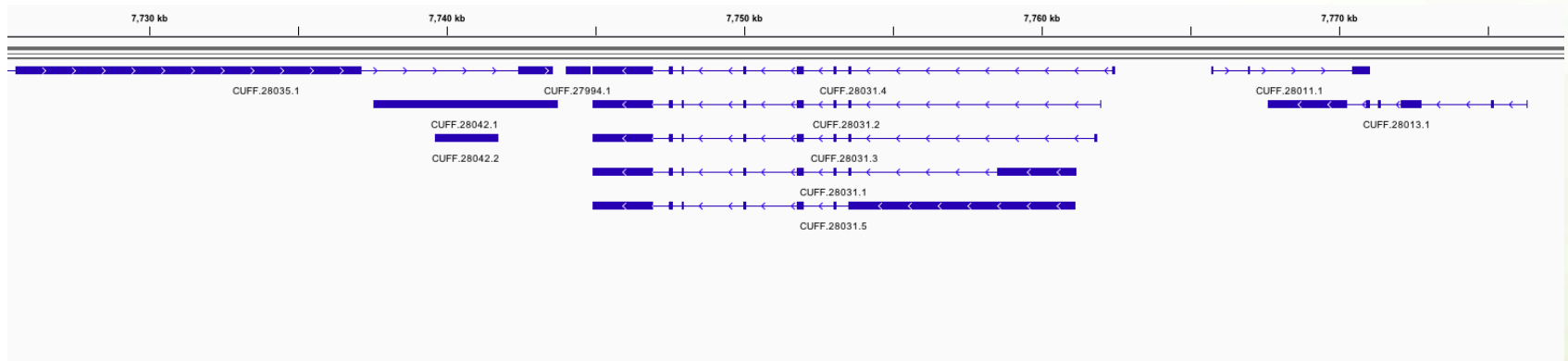


mRNA

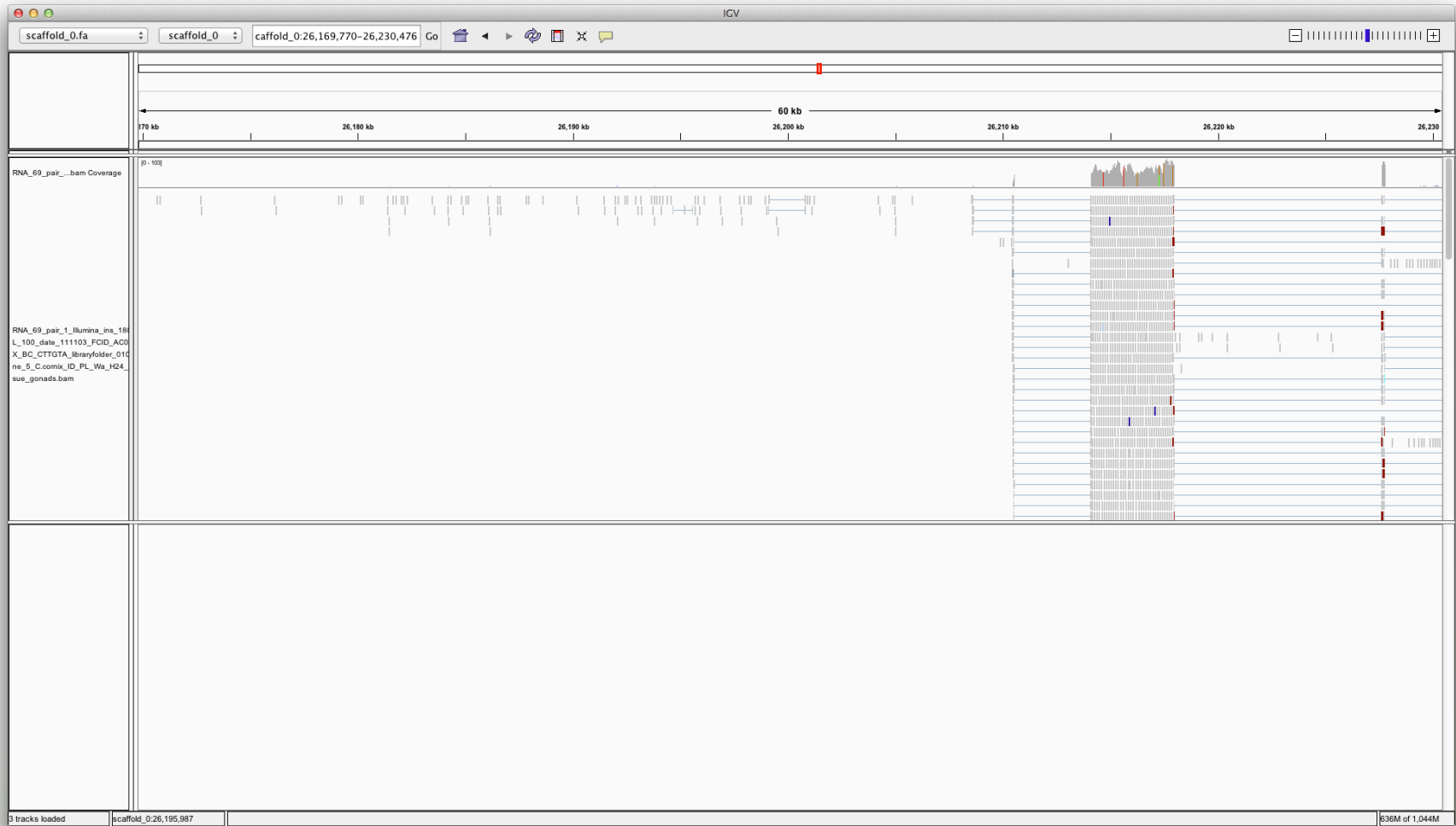


Translation

Pre-mRNA



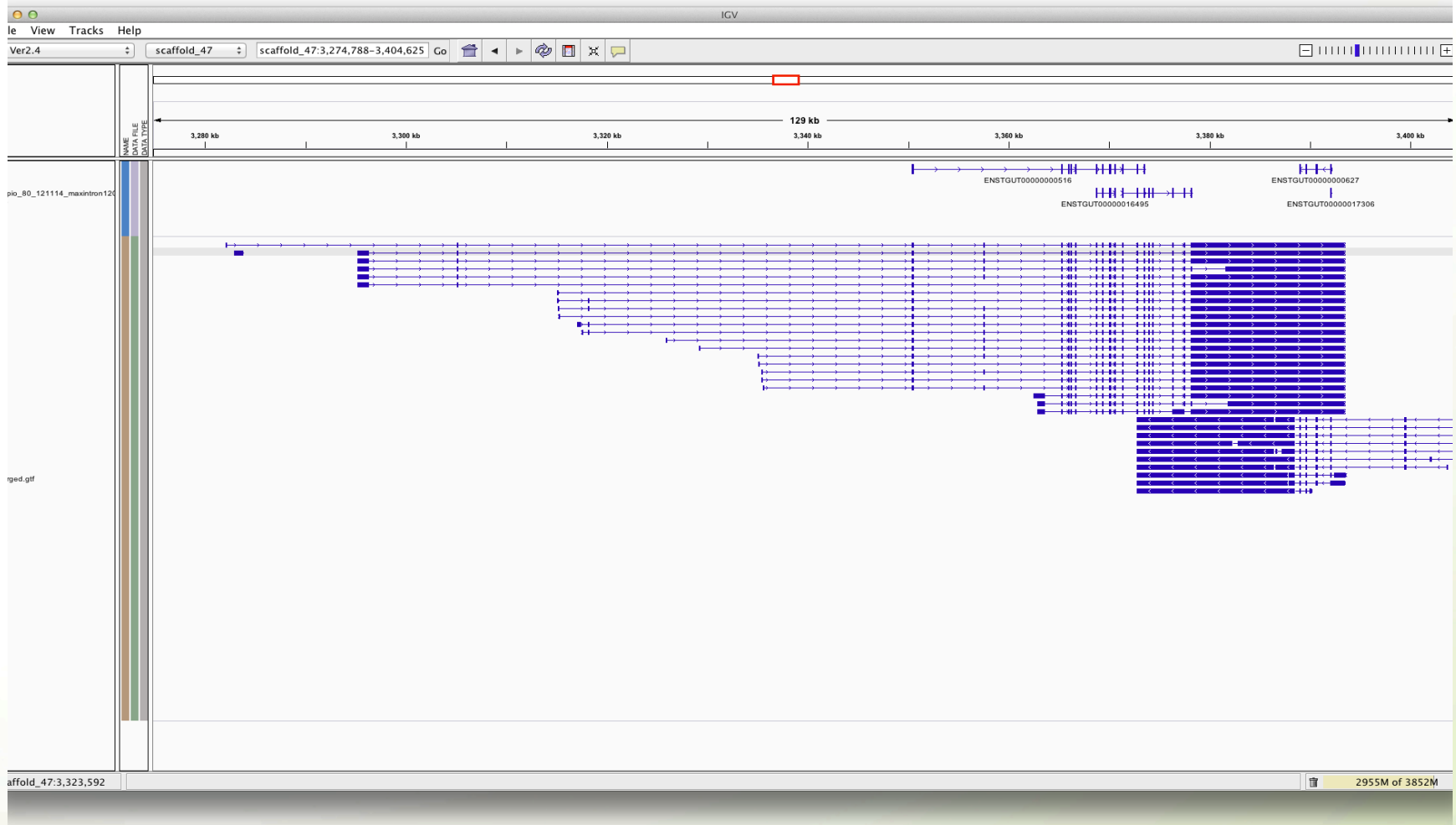
Includes everything that is transcribed



Stranded rna-seq

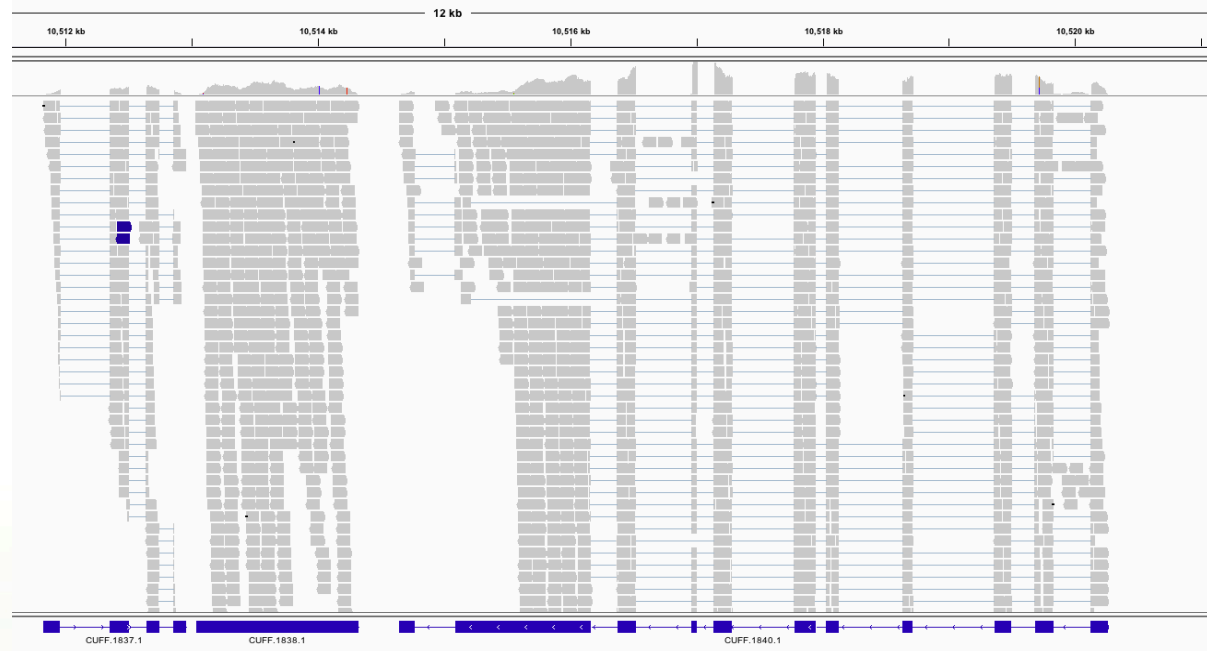


Three-prime bias in polyA-selected rna-seq



How to use RNA-seq

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

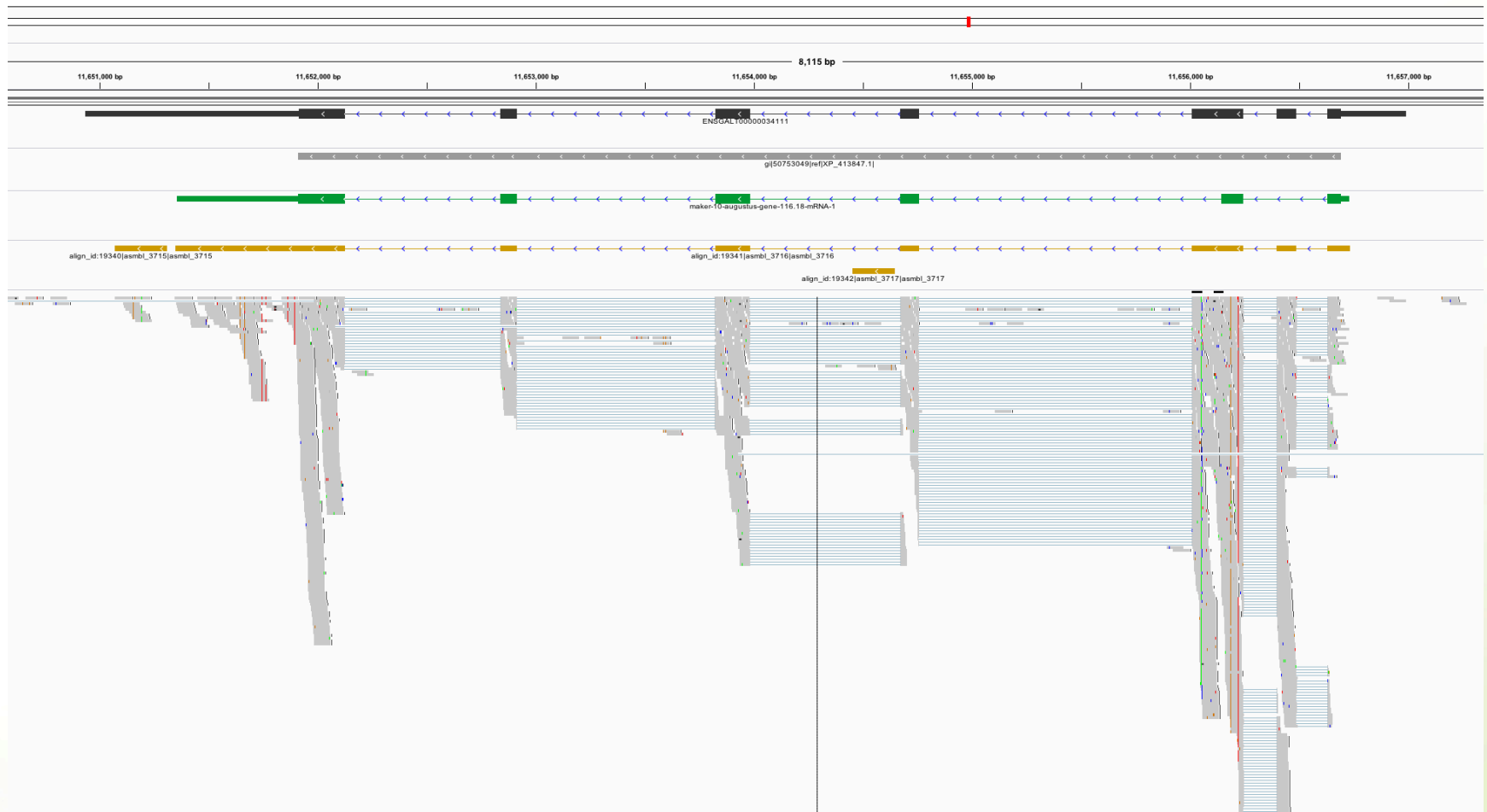


How to use RNA-seq

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

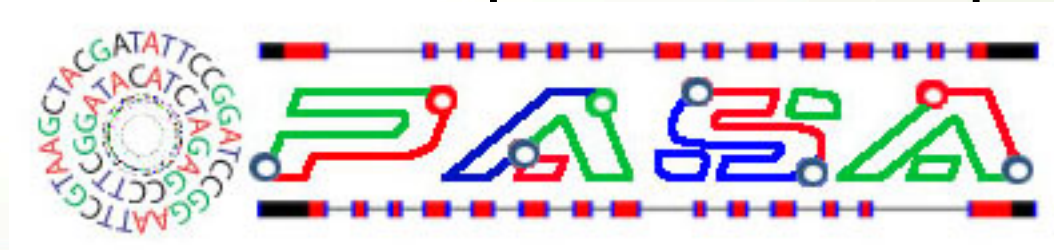


Mapped Trinity-assembled transcripts



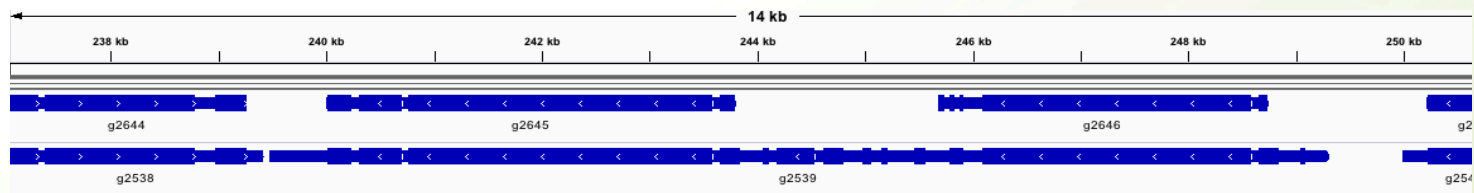
How to use RNA-seq

- 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



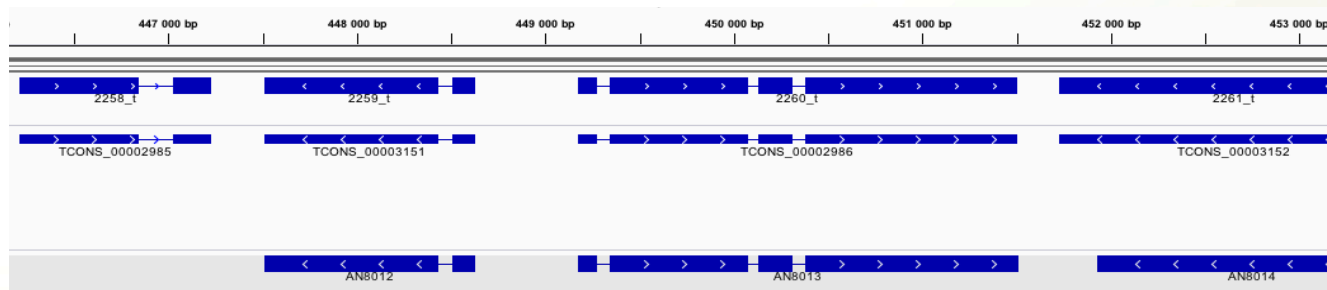
Ab initio gene finders are used in Maker

- 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

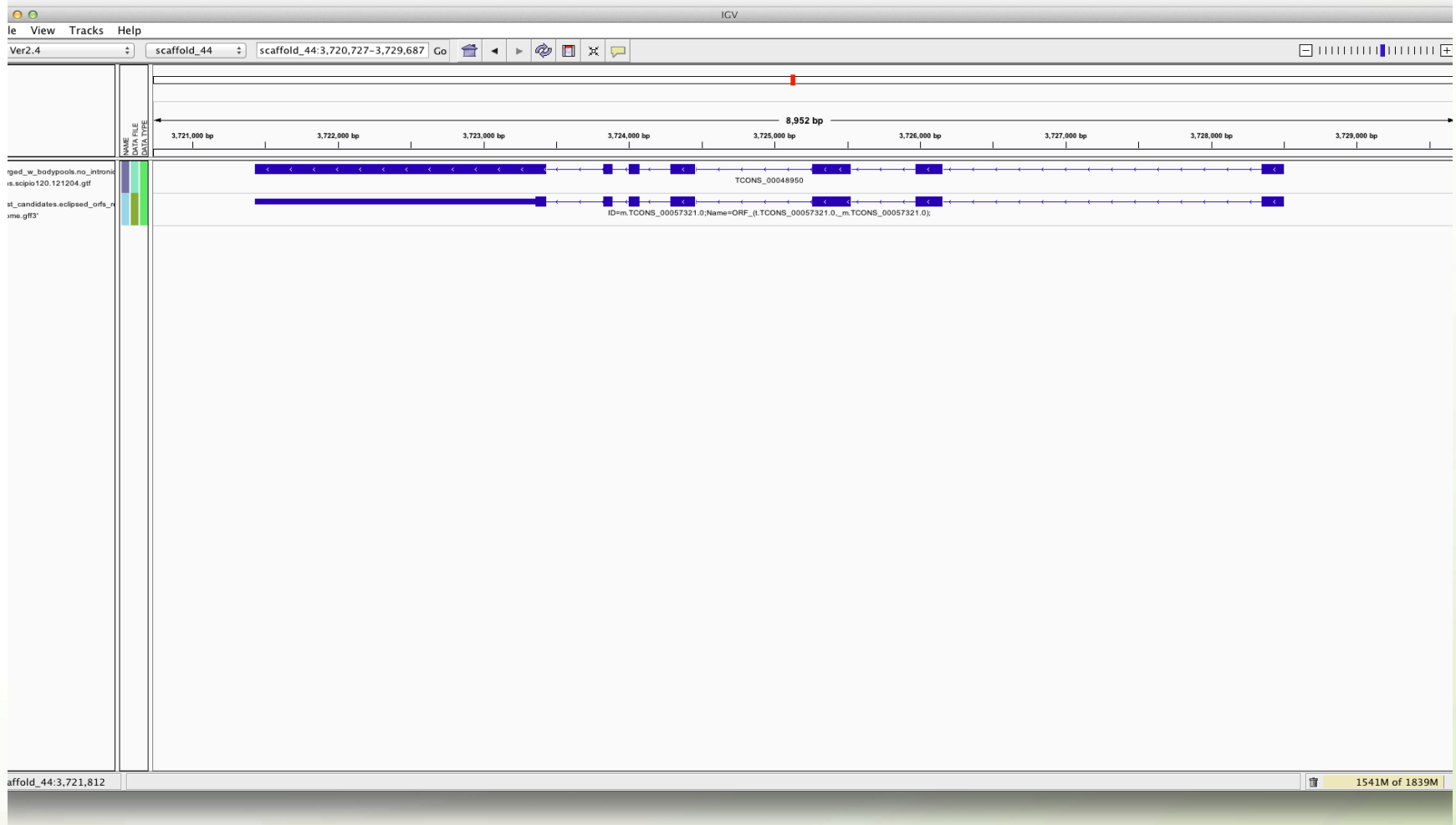
- Always combine different types of evidence!
- One single method is not enough!
- Use Maker!



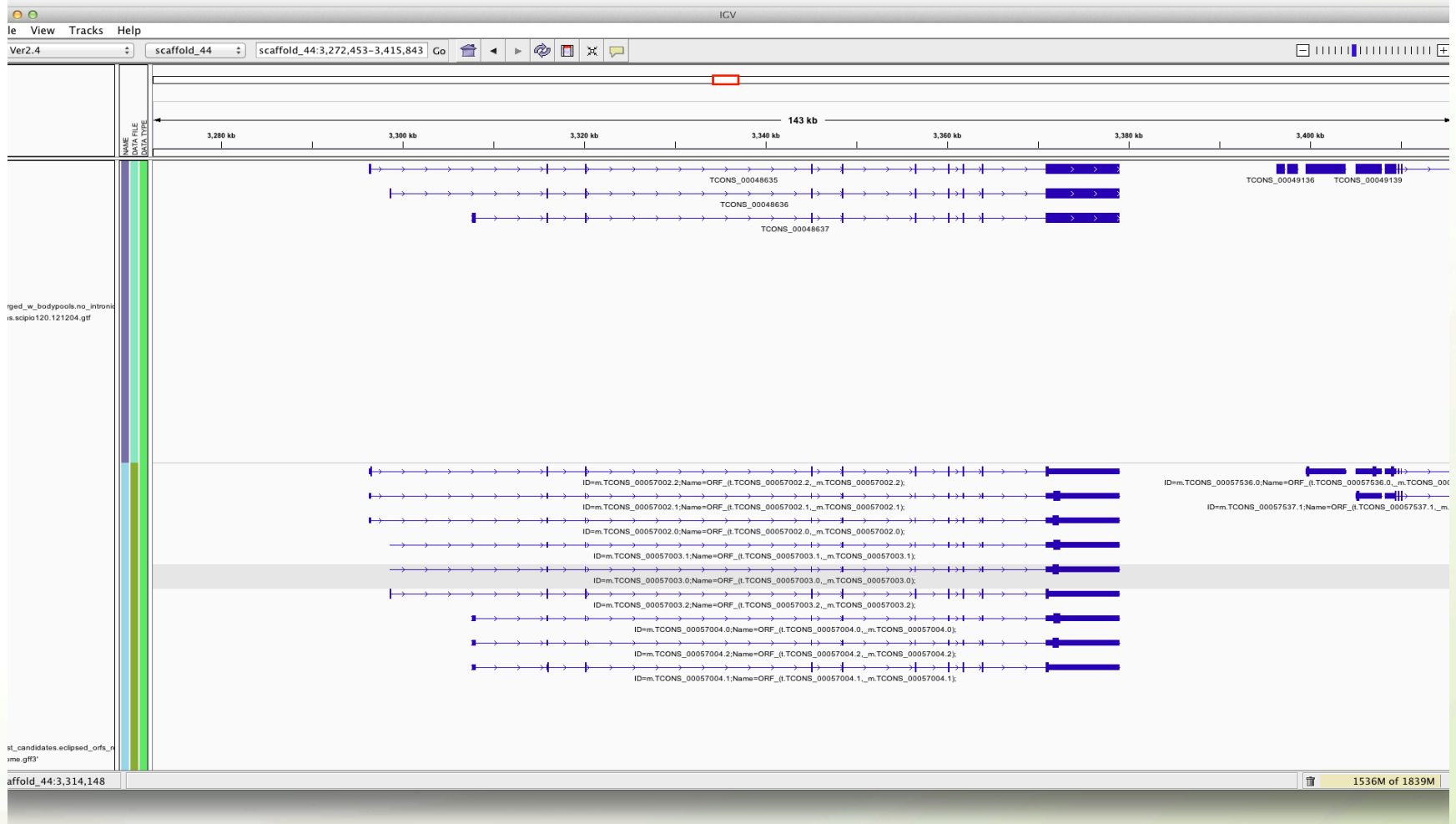
Transcript annotation

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

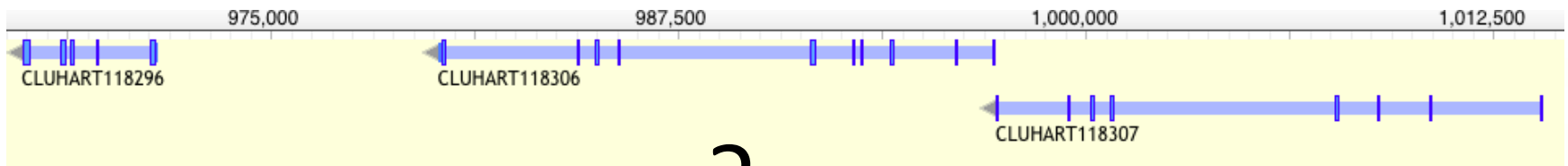
Transdecoder



Transdecoder



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



?

?

?

Insulin receptor?

Vesicle-trafficking protein?

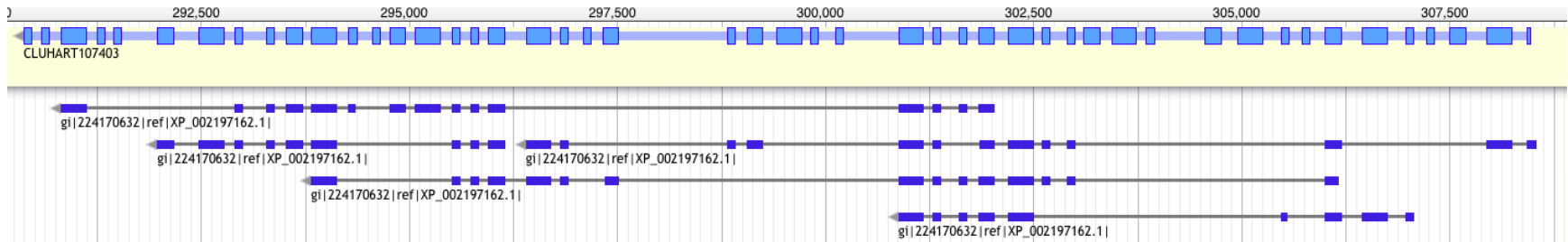
Alcohol dehydrogenase?

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

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



/Users/hobbe/Documents/Artemis_files_current/blast2go_20101001_0816.dat - Blast2GO V.2.4.4

File Blast Mapping Annotation Analysis Statistics Select Tools View Info

GO:0007067,GO:0016021 transport;binding;apoptos SPO_2518,DDX18_HUMAN

nr	sequence name	seq description	length	#...	min. eValue	sim mean	#C...	GO IDs	Enzyme	InterPro
3884	gene_3884 GeneMar...	c6 transcription	977	20	1.0E-171	59.85%	7	F:transcription factor activity; F:zinc ion binding; P:regulation of transcription, DNA-dependent; C:transcription factor complex; F:transporter activity; C:membrane; P:transmembrane transport		IPR005829; IPR007219
3885	gene_3885 GeneMar...	hypothetical protein NFIA_039100 [Neosartorya fischeri NRRL 181]	312	20	1.0E-39	63.15%	1	C:viral capsid		no IPS match
3886	gene_3886 GeneMar...	sin3 complex subunit	870	20	0.0	73.2%	0			
3887	gene_3887 GeneMar...	mitochondrial intermembrane space translocase subunit	87	20	1.0E-40	88.55%	5	F:metal ion binding; P:protein import into mitochondrial inner membrane; C:mitochondrial inner membrane; C:mitochondrial intermembrane space protein transporter complex; P:transmembrane transport		IPR004217; PTHR11038 (PANTHER); PTHR11038:SF8 (PANTHER)
3888	gene_3888 GeneMar...	lysyl-tRNA synthetase	592	20	0.0	73.55%	7	C:cytoplasm; P:auxin biosynthetic process; F:nucleic acid binding; F:lysine-tRNA ligase activity; P:lysyl-tRNA aminoacylation; F:ATP binding; P:lysine biosynthetic process	EC:6.1.1.6	IPR004364; IPR004365; IPR006195; IPR012340; IPR016027; IPR018149; IPR018150; G3DSA:3.30.930.10 (GENE3D); SSF5568 (SUPERFAMILY)
3889	gene_3889 GeneMar...	transcription factor conserved	1569	20	0.0	70.9%	0			
3890	gene_3890 GeneMar...	hypothetical protein [Aspergillus clavatus NRRL 1]	240	20	1.0E-51	56.25%	0			
		udp-glc gal endoplasmic reticulum nucleotide						C:integral to membrane; C:endoplasmic reticulum membrane; P:transmembrane transport; P:carbohydrate		IPR013657; PTHR10778 (PANTHER)

GO Graphs Application Messages Blast/IPS Results Statistics Kegg Maps

```

17:59 InterProScan for gene_8871|GeneMark.hmm|286_aa done.
17:59 -----
17:59 InterProScan Result:
17:59 InterProId: IPR001715
17:59 InterProName: Calponin-like actin-binding
17:59 InterProType: Domain
17:59 DB-Name: GENE3D - G3DSA:1.10.418.10
17:59 InterProId: IPR016146
17:59 InterProName: Calponin-homology
17:59 InterProType: Domain
17:59 DB-Name: SUPERFAMILY - SSF47576
17:59 InterProId: noIPR
17:59 InterProName: unintegrated
17:59 InterProType: unintegrated
17:59 DB-Name: PANTHER - PTHR19961
17:59 DB-Name: PANTHER - PTHR19961:SF9
    
```

Annotation already running

- 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

The screenshot shows a web browser window titled "The Gene Ontology" with the address bar displaying "www.geneontology.org". The browser's address bar also shows "GO" and "Reader" buttons. The website's header features the "the Gene Ontology" logo and a search bar with a "go!" button. Below the header is a navigation menu with links for "Downloads", "Tools", "Documentation", "Projects", "About", and "Contact".

Welcome to the Gene Ontology website!

The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The project provides [a controlled vocabulary of terms](#) for describing gene product characteristics and [gene product annotation data](#) from GO Consortium members, as well as [tools to access and process this data](#). [Read more about the Gene Ontology...](#)

Search the Gene Ontology Database

Search for genes, proteins or GO terms using [AmiGO](#):

[AmiGO](#) is the official GO browser and search engine.

The Gene Ontology project very much encourages input from the community into both the content of the GO and annotation using GO. We are very happy to work with others to ensure that the GO is both complete and accurate, and we also very much encourage communities to submit GO annotations for inclusion in the GO database. [Please contact us.](#)

The Gene Ontology Consortium is supported by a U41 grant from the National Human Genome Research Institute (NHGRI) [grant HG002273]. [See the full list of funding sources](#). The Gene Ontology Consortium would like to acknowledge the assistance of many more people than can be listed here. Please visit the [acknowledgements page](#) for the full list.

Copyright © 1999-2013 [the Gene Ontology](#)

[Helpdesk](#) • [Cite](#) • [Terms of use](#) • [News](#) • [RSS](#)

Member of the [Open Biological and Biomedical Ontologies](#)

Quick Links

- Tools
- [AmiGO browser](#)
- Submit GO Annotations
- OBO-Edit ontology editor
- Ontology downloads
- Annotation downloads
- Database downloads
- Documentation
- GO FAQ
- [GO on SourceForge](#)
- Contact GO

News

- [GO on Twitter](#)
- Finding updates...
- GO newsdesk
- [GO news RSS feed](#)
- [GO on Facebook](#)

Open "http://www.geneontology.org/GO.downloads.annotations.shtml" in a new tab

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 acticity, transporter activity

Blast2GO



/Users/hobbe/Documents/Artemis_files_current/blast2go_20101001_0816.dat - Blast2GO V.2.4.4

File Blast Mapping Annotation Analysis Statistics Select Tools View Info

GO:0007067,GO:0016021 transport;binding;apoptos SPO_2518,DDX18_HUMAN

nr	sequence name	seq description	length	#...	min. eValue	sim mean	#C...	GO IDs	Enzyme	InterPro
3884	gene_3884 GeneMar...	c6 transcription	977	20	1.0E-171	59.85%	7	F:transcription factor activity; F:zinc ion binding; P:regulation of transcription, DNA-dependent; C:transcription factor complex; F:transporter activity; C:membrane; P:transmembrane transport		IPR005829; IPR007219
3885	gene_3885 GeneMar...	hypothetical protein NFIA_039100 [Neosartorya fischeri NRRL 181]	312	20	1.0E-39	63.15%	1	C:viral capsid		no IPS match
3886	gene_3886 GeneMar...	sin3 complex subunit	870	20	0.0	73.2%	0			
3887	gene_3887 GeneMar...	mitochondrial intermembrane space translocase subunit	87	20	1.0E-40	88.55%	5	F:metal ion binding; P:protein import into mitochondrial inner membrane; C:mitochondrial inner membrane; C:mitochondrial intermembrane space protein transporter complex; P:transmembrane transport		IPR004217; PTHR11038 (PANTHER); PTHR11038:SF8 (PANTHER)
3888	gene_3888 GeneMar...	lysyl-tRNA synthetase	592	20	0.0	73.55%	7	C:cytoplasm; P:auxin biosynthetic process; F:nucleic acid binding; F:lysine-tRNA ligase activity; P:lysyl-tRNA aminoacylation; F:ATP binding; P:lysine biosynthetic process	EC:6.1.1.6	IPR004364; IPR004365; IPR006195; IPR012340; IPR016027; IPR018149; IPR018150; G3DSA:3.30.930.10 (GENE3D); SSF5568 (SUPERFAMILY)
3889	gene_3889 GeneMar...	transcription factor conserved	1569	20	0.0	70.9%	0			
3890	gene_3890 GeneMar...	hypothetical protein [Aspergillus clavatus NRRL 1]	240	20	1.0E-51	56.25%	0			
		udp-glc gal endoplasmic reticulum nucleotide						C:integral to membrane; C:endoplasmic reticulum membrane; P:transmembrane transport; P:carbohydrate		IPR013657; PTHR10778 (PANTHER)

GO Graphs Application Messages Blast/IPS Results Statistics Kegg Maps

```

17:59 InterProScan for gene_8871|GeneMark.hmm|286_aa done.
17:59 -----
17:59 InterProScan Result:
17:59 InterProId: IPR001715
17:59 InterProName: Calponin-like actin-binding
17:59 InterProType: Domain
17:59 DB-Name: GENE3D - G3DSA:1.10.418.10
17:59 InterProId: IPR016146
17:59 InterProName: Calponin-homology
17:59 InterProType: Domain
17:59 DB-Name: SUPERFAMILY - SSF47576
17:59 InterProId: noIPR
17:59 InterProName: unintegrated
17:59 InterProType: unintegrated
17:59 DB-Name: PANTHER - PTHR19961
17:59 DB-Name: PANTHER - PTHR19961:SF9
    
```

Annotation already running

Interproscan

The screenshot shows a web browser window with the URL www.ebi.ac.uk/interpro/interproscan.html. The page header includes the EMBL-EBI logo and navigation links for Services, Research, Training, and About us. The main content area features the InterPro logo with the tagline "Protein sequence analysis & classification" and a search bar labeled "Search InterPro...". Below the search bar, there are examples of search terms: IPR020405, kinase, P51587, PF02932, GO:0007165. A secondary navigation bar contains links for Home, Release notes, Training & tutorials, FAQs, Download, About InterPro, and Contact.

About InterProScan

What is InterProScan?

InterProScan is the software package that allows sequences (protein and nucleic) to be scanned against InterPro's signatures. Signatures are predictive models, provided by several different databases (referred to as member databases), that make up the InterPro consortium.

The software is available:

- As a web-based tool, using the sequence search box on the [InterPro homepage](#), for the analysis of single protein sequences (also available in the [EBI tool section](#))
- Programmatically via Web services that allow up to 25 sequences to be analysed per request (both [SOAP](#) and [REST](#)-based services are available)
- As a downloadable package for local installation from the EBI's FTP server, for instructions see the [detailed documentation pages](#).

InterProScan is run regularly against UniProtKB and the results are made available via the InterPro website.

More information

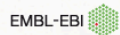
For more information, and for instructions on how to obtain, install and run InterProScan, please see the [detailed documentation pages](#).

Publications



[InterProScan 5: genome-scale protein function classification](#)
Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F. Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, and Sarah Hunter

Bioinformatics, Jan 2014
(doi:10.1093/bioinformatics/btu031)
[HTML](#) - [PDF \(324Kb\)](#)



News
Brochures
Contact us
Intranet

Services

By topic
By name (A-Z)
Help & Support

Research

Overview
Publications
Research groups
Postdocs & PhDs

Training

Overview
Train at EBI
Train outside EBI
Train online
Contact organisers

Industry

Overview
Members Area
Workshops
SME Forum
Contact Industry programme

About us

Overview
Leadership
Funding
Background
Collaboration
Jobs
People & groups
News



Karolinska
Institutet



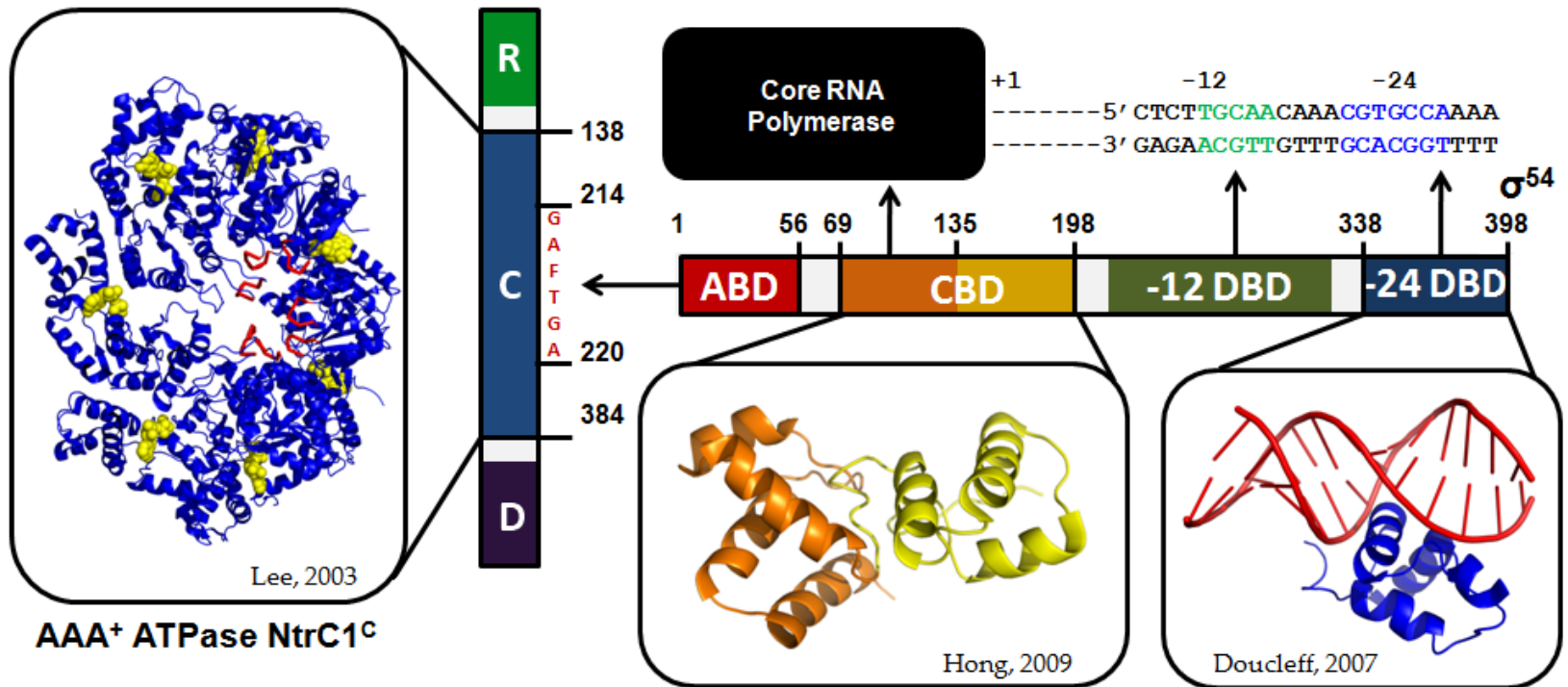
Stockholms
universitet



UPPSALA
UNIVERSITET

SciLifeLab

Sequence domains

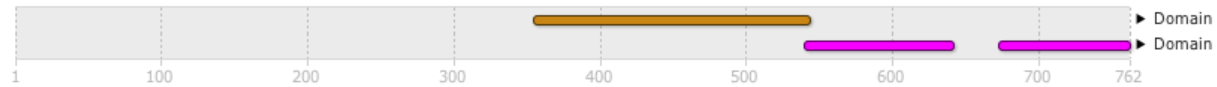


Interproscan results

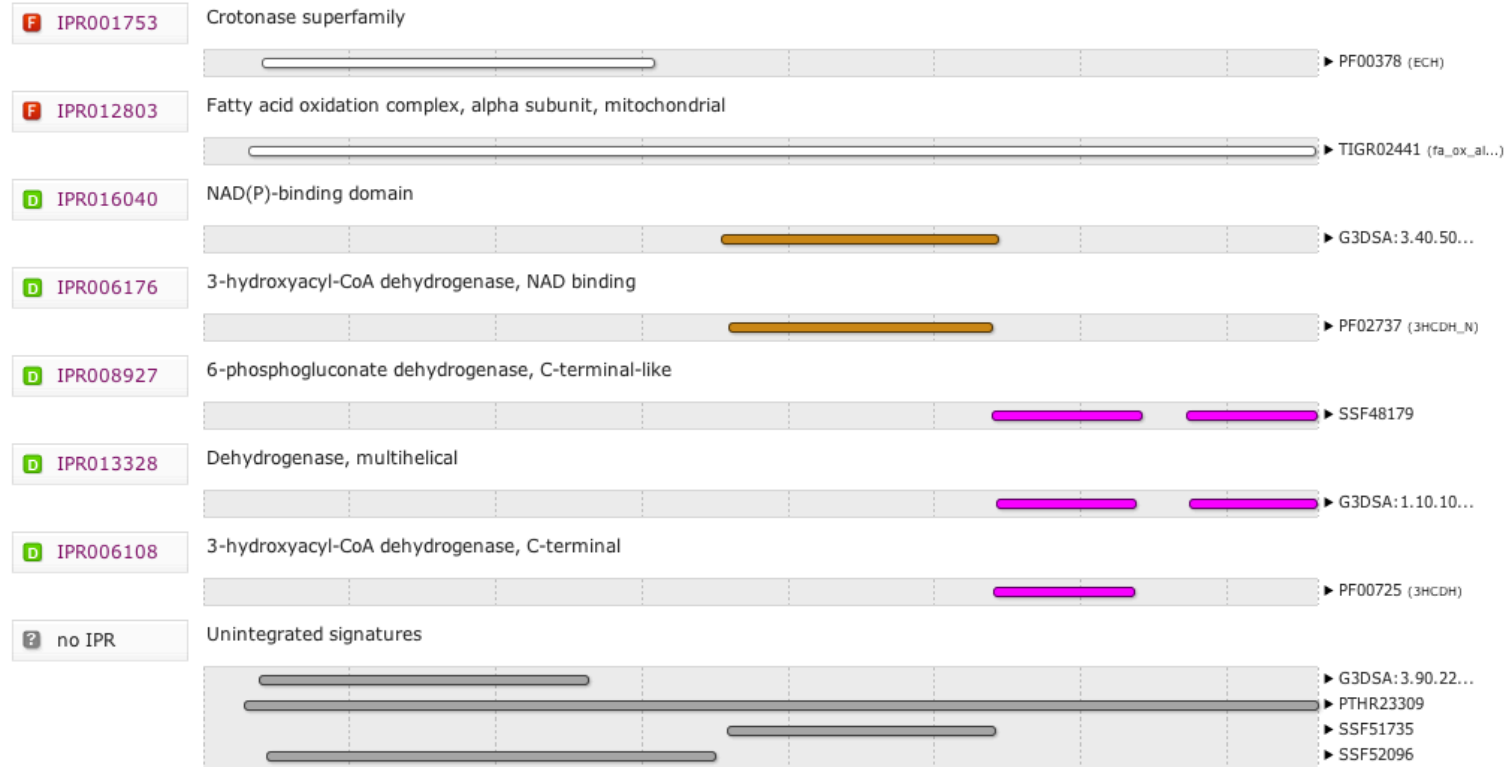
Protein family membership

- F Crotonase superfamily (IPR001753)
- F Fatty acid oxidation complex, alpha subunit, mitochondrial (IPR012803)

Domains and repeats



Detailed signature matches



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



/Users/hobbe/Documents/Artemis_files_current/blast2go_20101001_0816.dat - Blast2GO V.2.4.4

File Blast Mapping Annotation Analysis Statistics Select Tools View Info

GO:0007067,GO:0016021 transport;binding;apoptos SPO_2518,DDX18_HUMAN

nr	sequence name	seq description	length	#...	min. eValue	sim mean	#C...	GO IDs	Enzyme	InterPro
3884	gene_3884 GeneMar...	c6 transcription	977	20	1.0E-171	59.85%	7	F:transcription factor activity; F:zinc ion binding; P:regulation of transcription, DNA-dependent; C:transcription factor complex; F:transporter activity; C:membrane; P:transmembrane transport		IPR005829; IPR007219
3885	gene_3885 GeneMar...	hypothetical protein NFIA_039100 [Neosartorya fischeri NRRL 181]	312	20	1.0E-39	63.15%	1	C:viral capsid		no IPS match
3886	gene_3886 GeneMar...	sin3 complex subunit	870	20	0.0	73.2%	0			
3887	gene_3887 GeneMar...	mitochondrial intermembrane space translocase subunit	87	20	1.0E-40	88.55%	5	F:metal ion binding; P:protein import into mitochondrial inner membrane; C:mitochondrial inner membrane; C:mitochondrial intermembrane space protein transporter complex; P:transmembrane transport		IPR004217; PTHR11038 (PANTHER); PTHR11038:SF8 (PANTHER)
3888	gene_3888 GeneMar...	lysyl-tRNA synthetase	592	20	0.0	73.55%	7	C:cytoplasm; P:auxin biosynthetic process; F:nucleic acid binding; F:lysine-tRNA ligase activity; P:lysyl-tRNA aminoacylation; F:ATP binding; P:lysine biosynthetic process	EC:6.1.1.6	IPR004364; IPR004365; IPR006195; IPR012340; IPR016027; IPR018149; IPR018150; G3DSA:3.30.930.10 (GENE3D); SSF5568 (SUPERFAMILY)
3889	gene_3889 GeneMar...	transcription factor conserved	1569	20	0.0	70.9%	0			
3890	gene_3890 GeneMar...	hypothetical protein [Aspergillus clavatus NRRL 1]	240	20	1.0E-51	56.25%	0			
		udp-glc gal endoplasmic reticulum nucleotide						C:integral to membrane; C:endoplasmic reticulum membrane; P:transmembrane transport; P:carbohydrate		IPR013657; PTHR10778 (PANTHER)

GO Graphs Application Messages Blast/IPS Results Statistics Kegg Maps

```

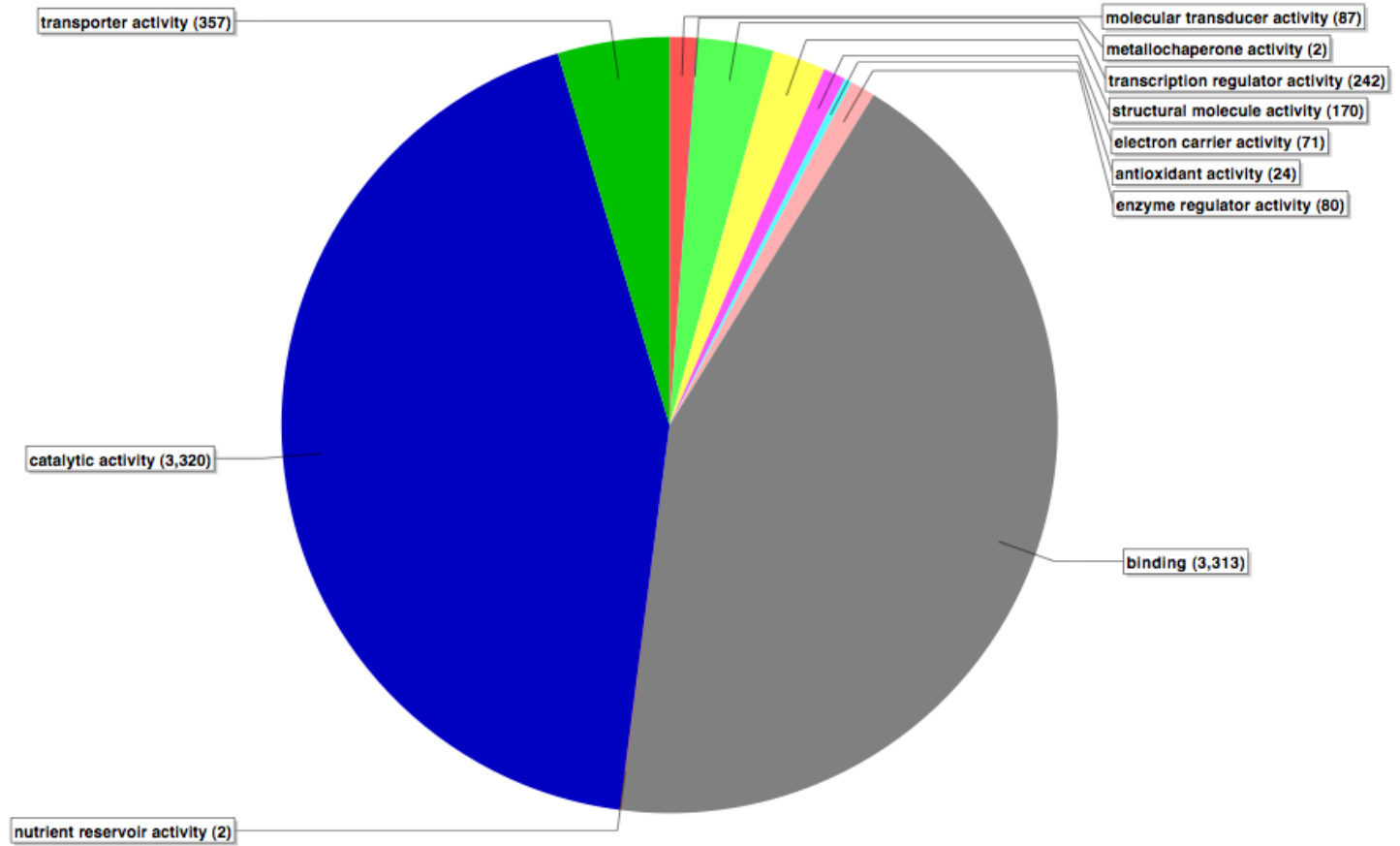
17:59 InterProScan for gene_8871|GeneMark.hmm|286_aa done.
17:59 -----
17:59 InterProScan Result:
17:59 InterProId: IPR001715
17:59 InterProName: Calponin-like actin-binding
17:59 InterProType: Domain
17:59 DB-Name: GENE3D - G3DSA:1.10.418.10
17:59 InterProId: IPR016146
17:59 InterProName: Calponin-homology
17:59 InterProType: Domain
17:59 DB-Name: SUPERFAMILY - SSF47576
17:59 InterProId: noIPR
17:59 InterProName: unintegrated
17:59 InterProType: unintegrated
17:59 DB-Name: PANTHER - PTHR19961
17:59 DB-Name: PANTHER - PTHR19961:SF9
    
```

Annotation already running

Interproscan

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

molecular_function Level 2



KEGG-mapping

/Users/hobbe/Documents/Artemis_files_current/blast2go_20101001_0816.dat - Blast2GO V.2.4.4

File Blast Mapping Annotation Analysis Statistics Select Tools View Info

GO:0007067,GO:0016021 transport:binding:apoptos SPO_2518,DDX18_HUMAN

nr	sequence name	seq description	length	#...	min. eValue	sim mean	#G...	GO IDs	Enzyme	InterPro
		succinyl- synthetase subunit						E:ATP binding; F:succinate-CoA ligase (GDP-forming) activity; P:tricarboxylic acid cycle; C:succinate-CoA ligase		IPR003781; IPR005810

GO Graphs Application Messages Blast/IPS Results Statistics **Kegg Maps**

GLYCEROLIPID METABOLISM

Pathways

- Pentose phosphate pathway
- Fructose and mannose metabolism
- Butanoate metabolism
- Carbon fixation in photosynthetic organisms
- Lysine degradation
- Tyrosine metabolism
- Methane metabolism
- Glyoxylate and dicarboxylate metabolism
- Glycerolipid metabolism**
- Glutathione metabolism
- Selenoamino acid metabolism
- Phenylalanine metabolism
- Valine, leucine and isoleucine biosynthesis
- Reductive carboxylate cycle (CO2 fixation)
- Galactose metabolism
- Phenylalanine, tyrosine and tryptophan biosynthesis
- N-Glycan biosynthesis
- Photosynthesis
- Drug metabolism - other enzymes
- Sulfur metabolism
- Fatty acid biosynthesis
- Inositol phosphate metabolism
- beta-Alanine metabolism
- Drug metabolism - cytochrome P450
- Pantothenate and CoA biosynthesis
- Biosynthesis of unsaturated fatty acids
- Cyanoamino acid metabolism
- Terpenoid backbone biosynthesis
- Histidine metabolism
- T cell receptor signaling pathway
- Tropane, piperidine and pyridine alkaloid biosynthesis
- One carbon pool by folate
- Pentose and glucuronate interconversions
- Phosphatidylinositol signaling system
- Lysine biosynthesis

Color	Enzyme	Sequences
red	ec:1.1.1.2 - alcohol dehydrogenase (NADP+)	gene_674 GeneMark.hmm 333_aa, gene_5801 GeneMark.hmm 312_aa
yellow	ec:2.3.1.158 - phospholipid:diacylglycerol acyltransferase	gene_2604 GeneMark.hmm 188_aa, gene_6532 GeneMark.hmm 505_aa
orange	ec:2.3.1.51 - 1-acylglycerol-3-phosphate O-acyltransferase	gene_176 GeneMark.hmm 429_aa, gene_6693 GeneMark.hmm 292_aa
green	ec:2.3.1.20 - diacylglycerol O-acyltransferase	gene_176 GeneMark.hmm 429_aa, gene_7213 GeneMark.hmm 521_aa, gene_8170 GeneMark.hmm 470_aa
blue	ec:2.3.1.15 - glycerol-3-phosphate O-acyltransferase	gene_886 GeneMark.hmm 748_aa, gene_2640 GeneMark.hmm 823_aa
pink	ec:1.1.1.72 - glycerol dehydrogenase (NADP+)	gene_3376 GeneMark.hmm 325_aa, gene_4577 GeneMark.hmm 326_aa
violet	ec:1.2.1.3 - aldehyde dehydrogenase (NAD+)	gene_2201 GeneMark.hmm 497_aa, gene_5247 GeneMark.hmm 502_aa, gene_5611 GeneMark.hmm 471_aa
light-red	ec:2.7.1.107 - diacylglycerol kinase	gene_5292 GeneMark.hmm 409_aa

Annotation already running

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

The screenshot shows the Biosupport.se website interface. At the top, there is a navigation bar with the site name "SBS Swedish Bioinformatics Support" and a search bar. Below the navigation bar, there are tabs for "Questions", "Tags", "Users", "Badges", and "Unanswered". A sidebar on the right contains a "login about faq" link, a "63 Questions 126 answers questions" summary, a "Welcome" message, a "Privacy" notice, and "Recent tags". The main content area displays a list of questions with their respective statistics (votes, answers, views) and titles.

Votes	Answers	Views	Question Title	Tags	Author	Date
0	1	90	oligo file for mothur	454amplicon oligofile mothur	dahlo	03 Jun, 15:13
0	1	518	Tophat results more than one accepted_hits file	cufflinks tophat	ehoei(suspended)	27 May, 11:47
0	2	85	Cloud services for bioinformatics?	cloud-computing service	zashah	27 May, 11:22
0	2	108	Memory overflow error performing (/mpileup/) in Uppmax	mpileup uppmx	Mikael Borg	22 May, 15:18
0	3	132	detect small RNA from normal RNA-seq data?	rna-seq	walker	08 May, 14:12
0	1	175	Common resources for bowtie / tophat for hg19	uppmx rna-seq annotation transcriptome	S Dilorenzo	06 May, 10:24
1	1	197	AddOrReplaceReadGroups picard problem: Caused by: java.io.IOException: No space left on device	picard	S Dilorenzo	29 Apr, 12:11
1	2	975	List of NGS services in Sweden?			