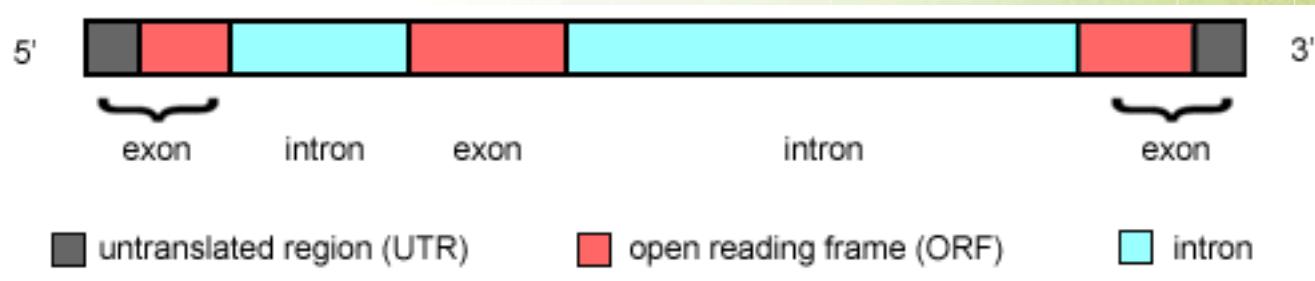


# Annotation of eukaryote genomes and transcriptomes



Henrik Lantz, BILS/SciLifeLab

Enabler for Life Sciences

# Lecture synopsis

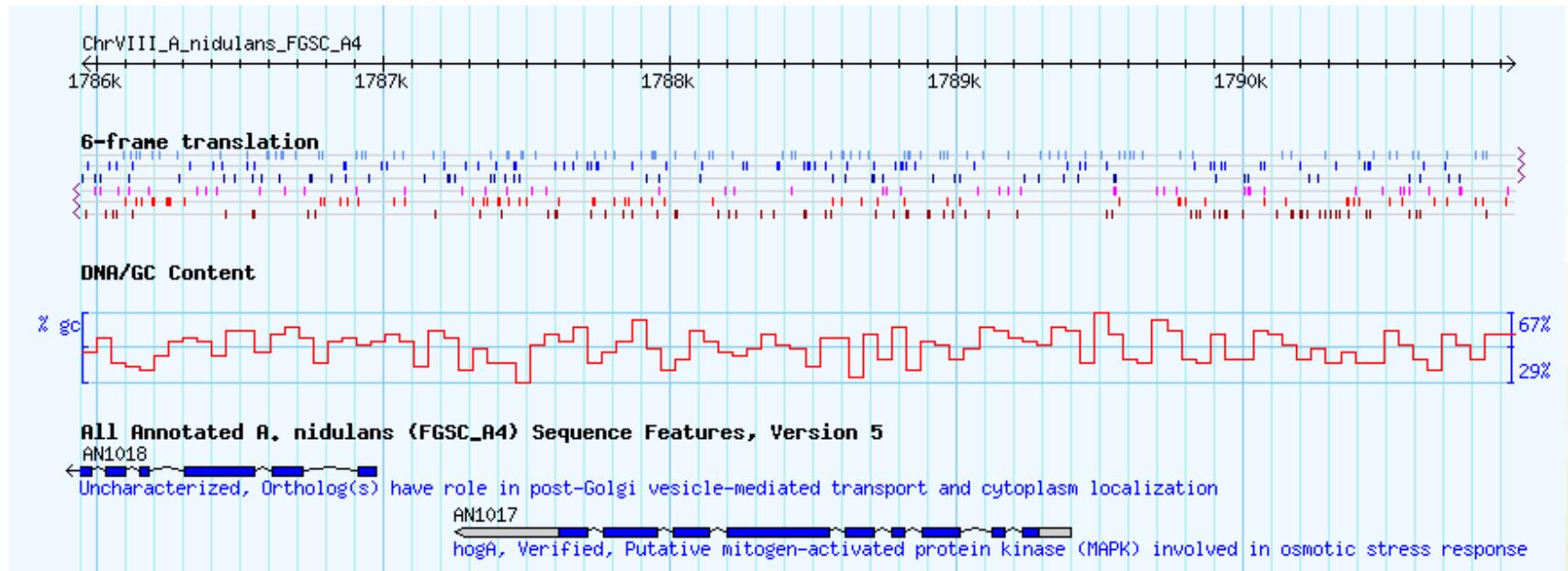
- What is annotation?
- Structural genome annotation
- Types of data used
- Transcriptome annotation
- Functional annotation

# What is annotation?

- Identification of regions of interest in sequence data

# From a genome...

# ...to an annotated gene



# GFF file format

```

###ff-version 3

scaffold_7    maker  gene   133848 144662 . - . ID=C55462A8A3E2878A71E2ABDFB34661;Name=maker-scaffold_7-augustus-gene-0_11
scaffold_7    maker  mRNA   133848 144662 . - . ID=A649E92346ADE2184E579FA9124ABD;Parent=C55462A8A3E2878A71E2ABDFB34661;Name=1:cornix-all_reads.72406.1;_AED=1.00;_eAED=1.00;_OI=71|0|0|0|0|0|8|414|347
scaffold_7    maker  exon   138974 139077 . - . ID=A649E92346ADE2184E579FA9124ABD;exon:7;Parent=A649E92346ADE2184E579FA9124ABD
scaffold_7    maker  exon   135098 135281 . - . ID=A649E92346ADE2184E579FA9124ABD;exon:6;Parent=A649E92346ADE2184E579FA9124ABD
scaffold_7    maker  exon   139616 139834 . - . ID=A649E92346ADE2184E579FA9124ABD;exon:5;Parent=A649E92346ADE2184E579FA9124ABD
scaffold_7    maker  exon   144511 144662 . - . ID=A649E92346ADE2184E579FA9124ABD;exon:4;Parent=A649E92346ADE2184E579FA9124ABD
scaffold_7    maker  exon   136342 136437 . - . ID=A649E92346ADE2184E579FA9124ABD;exon:3;Parent=A649E92346ADE2184E579FA9124ABD
scaffold_7    maker  exon   133848 134001 . - . ID=A649E92346ADE2184E579FA9124ABD;exon:2;Parent=A649E92346ADE2184E579FA9124ABD
scaffold_7    maker  exon   134262 134303 . - . ID=A649E92346ADE2184E579FA9124ABD;exon:1;Parent=A649E92346ADE2184E579FA9124ABD
scaffold_7    maker  exon   144248 144296 . - . ID=A649E92346ADE2184E579FA9124ABD;exon:0;Parent=A649E92346ADE2184E579FA9124ABD
scaffold_7    maker  five_prime_UTR 144592 144662 . - . ID=A649E92346ADE2184E579FA9124ABD;five_prime_utr;Parent=A649E92346ADE2184E579FA9124ABD
scaffold_7    maker  CDS   144511 144591 . - . ID=A649E92346ADE2184E579FA9124ABD;cds;Parent=A649E92346ADE2184E579FA9124ABD
scaffold_7    maker  CDS   144138 144291 . - . ID=A649E92346ADE2184E579FA9124ABD;cds;Parent=A649E92346ADE2184E579FA9124ABD
scaffold_7    maker  CDS   141262 141383 . - . ID=A649E92346ADE2184E579FA9124ABD;cds;Parent=A649E92346ADE2184E579FA9124ABD
scaffold_7    maker  CDS   139616 139834 . - . ID=A649E92346ADE2184E579FA9124ABD;cds;Parent=A649E92346ADE2184E579FA9124ABD
scaffold_7    maker  CDS   138974 139077 . - . ID=A649E92346ADE2184E579FA9124ABD;cds;Parent=A649E92346ADE2184E579FA9124ABD
scaffold_7    maker  CDS   136342 136437 . - . ID=A649E92346ADE2184E579FA9124ABD;cds;Parent=A649E92346ADE2184E579FA9124ABD
scaffold_7    maker  CDS   135098 135281 . - . ID=A649E92346ADE2184E579FA9124ABD;cds;Parent=A649E92346ADE2184E579FA9124ABD
scaffold_7    maker  CDS   134262 134303 . - . ID=A649E92346ADE2184E579FA9124ABD;cds;Parent=A649E92346ADE2184E579FA9124ABD
scaffold_7    maker  CDS   133848 134001 . - . ID=A649E92346ADE2184E579FA9124ABD;cds;Parent=A649E92346ADE2184E579FA9124ABD
scaffold_7    maker  three_prime_UTR 134261 . - . ID=A649E92346ADE2184E579FA9124ABD;three_prime_utr;Parent=A649E92346ADE2184E579FA9124ABD
scaffold_7    maker  gene   83101 117593 . + . ID=D30BD9A5F277975648A42E890F6B999;Name=maker-scaffold_7-augustus-gene-0_6-mRNA-1;_AED=1.00;_eAED=1.00;_OI=0|0|0|0|0|0|21|4|706
scaffold_7    maker  mRNA  83101 117593 . + . ID=CF5FD0A190832937C45A0D02E674C926;Parent=D30BD9A5F277975648A42E890F6B999
scaffold_7    maker  exon   95748 95871 . + . ID=CF5FD0A190832937C45A0D02E674C926;exon:8;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  exon   99113 9917 . + . ID=CF5FD0A190832937C45A0D02E674C926;exon:7;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  exon   99113 9948 . + . ID=CF5FD0A190832937C45A0D02E674C926;exon:6;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  exon   119231 110356 . + . ID=CF5FD0A190832937C45A0D02E674C926;exon:5;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  exon   112609 113679 . + . ID=CF5FD0A190832937C45A0D02E674C926;exon:4;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  exon   94057 94117 . + . ID=CF5FD0A190832937C45A0D02E674C926;exon:3;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  exon   84578 84670 . + . ID=CF5FD0A190832937C45A0D02E674C926;exon:2;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  exon   115452 115536 . + . ID=CF5FD0A190832937C45A0D02E674C926;exon:15;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  exon   111579 111669 . + . ID=CF5FD0A190832937C45A0D02E674C926;exon:16;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  exon   102917 103016 . + . ID=CF5FD0A190832937C45A0D02E674C926;exon:17;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  exon   96766 96849 . + . ID=CF5FD0A190832937C45A0D02E674C926;exon:18;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  exon   86666 86750 . + . ID=CF5FD0A190832937C45A0D02E674C926;exon:19;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  exon   99944 100109 . + . ID=CF5FD0A190832937C45A0D02E674C926;exon:20;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  exon   109766 109860 . + . ID=CF5FD0A190832937C45A0D02E674C926;exon:21;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  exon   93154 93282 . + . ID=CF5FD0A190832937C45A0D02E674C926;exon:22;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  exon   114737 114825 . + . ID=CF5FD0A190832937C45A0D02E674C926;exon:23;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  exon   83101 83155 . + . ID=CF5FD0A190832937C45A0D02E674C926;exon:24;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  exon   903533 904005 . + . ID=CF5FD0A190832937C45A0D02E674C926;exon:25;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  exon   10777 117593 . + . ID=CF5FD0A190832937C45A0D02E674C926;exon:26;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  exon   106779 106865 . + . ID=CF5FD0A190832937C45A0D02E674C926;exon:27;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  exon   105743 105835 . + . ID=CF5FD0A190832937C45A0D02E674C926;exon:28;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  CDS   83101 83155 . - . ID=CF5FD0A190832937C45A0D02E674C926;cds;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  CDS   84578 84670 . - . ID=CF5FD0A190832937C45A0D02E674C926;cds;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  CDS   86666 86750 . - . ID=CF5FD0A190832937C45A0D02E674C926;cds;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  CDS   90664 90748 . - . ID=CF5FD0A190832937C45A0D02E674C926;cds;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  CDS   93154 93282 . - . ID=CF5FD0A190832937C45A0D02E674C926;cds;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  CDS   94057 94117 . - . ID=CF5FD0A190832937C45A0D02E674C926;cds;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  CDS   95748 95871 . - . ID=CF5FD0A190832937C45A0D02E674C926;cds;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  CDS   96766 96849 . - . ID=CF5FD0A190832937C45A0D02E674C926;cds;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  CDS   99944 100109 . - . ID=CF5FD0A190832937C45A0D02E674C926;cds;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  CDS   109766 109860 . - . ID=CF5FD0A190832937C45A0D02E674C926;cds;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  CDS   114737 114825 . - . ID=CF5FD0A190832937C45A0D02E674C926;cds;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  CDS   83101 83155 . - . ID=CF5FD0A190832937C45A0D02E674C926;cds;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  CDS   903533 904005 . - . ID=CF5FD0A190832937C45A0D02E674C926;cds;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  CDS   10777 117593 . - . ID=CF5FD0A190832937C45A0D02E674C926;cds;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  CDS   106779 106865 . - . ID=CF5FD0A190832937C45A0D02E674C926;cds;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  CDS   105743 105835 . - . ID=CF5FD0A190832937C45A0D02E674C926;cds;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  CDS   114737 114825 . - . ID=CF5FD0A190832937C45A0D02E674C926;cds;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  CDS   111579 111669 . - . ID=CF5FD0A190832937C45A0D02E674C926;cds;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  CDS   112609 113679 . - . ID=CF5FD0A190832937C45A0D02E674C926;cds;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  CDS   114737 114825 . - . ID=CF5FD0A190832937C45A0D02E674C926;cds;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  CDS   115452 115536 . - . ID=CF5FD0A190832937C45A0D02E674C926;cds;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  CDS   117477 117589 . - . ID=CF5FD0A190832937C45A0D02E674C926;cds;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  three_prime_UTR 117593 117593 . + . ID=CF5FD0A190832937C45A0D02E674C926;three_prime_utr;Parent=CF5FD0A190832937C45A0D02E674C926
scaffold_7    maker  gene   22451 37514 . - . ID=E1E9A456071E5077490C938E70F79E56;Name=maker-scaffold_7-augustus-gene-0_10-mRNA-1;_AED=1.00;_eAED=1.00;_OI=1|1|0|0|0|0|11|762|414
scaffold_7    maker  mRNA   22451 37514 . - . ID=2B838792886FC0887J47087E6500CCS;Parent=E1E9A456071E5077490C938E70F79E56;Name=maker-scaffold_7-augustus-gene-0_10-mRNA-1;_AED=1.00;_eAED=1.00;_OI=3008|0|0|0|0|0|10|1|0|0|400
scaffold_7    maker  exon   26810 37514 . - . ID=10d62981725f8971755858a23623A6;Parent=E1E9A456071E5077490C938E70F79E56;Name=maker-scaffold_7-augustus-gene-0_10-mRNA-1;_AED=1.00;_eAED=1.00;_OI=3008|0|0|0|0|0|10|1|0|0|400
scaffold_7    maker  exon   26810 27088 . - . ID=10d629817081725f8971755858a23623A6;exon:10;Parent=10d629817081725f8971755858a23623A6
scaffold_7    maker  exon   27950 28989 . - . ID=2B838792886FC0887J47087E6500CCS;exon:9;Parent=2B838792886FC0887J47087E6500CCS;exon:10;Parent=10d629817081725f8971755858a23623A6
scaffold_7    maker  exon   33513 33610 . - . ID=2B838792886FC0887J47087E6500CCS;exon:8;Parent=2B838792886FC0887J47087E6500CCS;exon:9;Parent=2B838792886FC0887J47087E6500CCS;exon:10;Parent=10d629817081725f8971755858a23623A6

```

# GFF3 file format

<b>Seqid</b>	<b>source</b>	<b>type</b>	<b>start</b>	<b>end</b>	<b>score</b>	<b>strand</b>	<b>phase</b>	<b>attributes</b>
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"; 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	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 "6.8105889357"; 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 "6.8105889357"; 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"; 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	exon	10522	13208	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	transcript	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 "57.7578565123"; 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.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	exon	1097	2009	1000	.	.	gene_id "CUFF.8"; transcript_id "CUFF.8.1"; FPKM "6.241510"; frac "1.000000"; conf_lo "4.624813"; conf_hi "69.749694"; cov "8277.399810";
Sb_20131119_contig_100023	Cufflinks	transcript	1	123	1000	.	.	gene_id "CUFF.8"; transcript_id "CUFF.8.1"; exon_number "1"; FPKM "6.241510"; frac "1.000000"; conf_lo "4.624813"; conf_hi "69.749694"; cov "8277.399810";
Sb_20131119_contig_100023	Cufflinks	exon	1	123	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	transcript	1	221	1000	.	.	gene_id "CUFF.18"; transcript_id "CUFF.18.1"; exon_number "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.11"; transcript_id "CUFF.11.1"; exon_number "1"; FPKM "3.216803"; conf_hi "9.184747"; cov "41.309085";
Sb_20131119_contig_100107	Cufflinks	transcript	2	255	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 "65.7363"; cov "24.668034";
Sb_20131119_contig_100107	Cufflinks	exon	2	255	1000	.	.	gene_id "CUFF.14"; transcript_id "CUFF.14.1"; FPKM "3960.0556774823"; frac "1.000000"; conf_lo "56.622026"; conf_hi "18.740088"; cov "4272.822539";
Sb_20131119_contig_100107	Cufflinks	transcript	2041	2331	1000	.	.	gene_id "CUFF.14"; transcript_id "CUFF.14.1"; exon_number "1"; FPKM "3960.0556774823"; frac "1.000000"; conf_lo "56.622026"; conf_hi "18.740088"; cov "4272.822539";
Sb_20131119_contig_100107	Cufflinks	exon	2041	2331	1000	.	.	gene_id "CUFF.17"; transcript_id "CUFF.17.1"; exon_number "1"; FPKM "23.640408"; conf_lo "23.640408"; conf_hi "34.395386"; cov "75.482038";
Sb_20131119_contig_100121	Cufflinks	transcript	1756	2236	1000	.	.	gene_id "CUFF.20"; transcript_id "CUFF.20.1"; exon_number "1"; FPKM "24.0908132000"; conf_lo "1.000000"; conf_hi "7.484121"; cov "49.916331";
Sb_20131119_contig_100121	Cufflinks	exon	1756	2236	1000	.	.	gene_id "CUFF.20"; transcript_id "CUFF.20.1"; exon_number "1"; FPKM "24.0908132000"; conf_lo "1.000000"; conf_hi "7.484121"; cov "49.916331";
Sb_20131119_contig_100192	Cufflinks	transcript	1840	2212	1000	.	.	gene_id "CUFF.23"; transcript_id "CUFF.23.1"; FPKM "25.0508552879"; frac "1.000000"; conf_lo "10.500448"; conf_hi "18.137870"; cov "45.523657";
Sb_20131119_contig_100192	Cufflinks	exon	1840	2212	1000	.	.	gene_id "CUFF.13"; transcript_id "CUFF.13.1"; exon_number "1"; FPKM "18.5005852879"; frac "1.000000"; conf_lo "13.100000"; conf_hi "23.43223"; cov "88.805088";
Sb_20131119_contig_100192	Cufflinks	transcript	1	616	1000	.	.	gene_id "CUFF.19"; transcript_id "CUFF.19.1"; exon_number "1"; FPKM "42.2951435149"; frac "1.000000"; conf_lo "23.43223"; conf_hi "33.43223"; cov "88.805088";
Sb_20131119_contig_100192	Cufflinks	exon	1	616	1000	.	.	gene_id "CUFF.15"; transcript_id "CUFF.15.1"; FPKM "4118.4602721559"; conf_lo "53.966826"; conf_hi "84.228598"; cov "4249.067452";
Sb_20131119_contig_100191	Cufflinks	transcript	219	353	1000	.	.	gene_id "CUFF.15"; transcript_id "CUFF.15.1"; exon_number "1"; FPKM "4118.4602721559"; conf_lo "53.966826"; conf_hi "84.228598"; cov "4249.067452";
Sb_20131119_contig_100040	Cufflinks	exon	219	353	1000	.	.	gene_id "CUFF.38"; transcript_id "CUFF.38.1"; FPKM "69.4702369875"; frac "1.000000"; conf_lo "60.274596"; conf_hi "71.712087"; cov "143.103653";
Sb_20131119_contig_100040	Cufflinks	transcript	945	2276	1000	.	.	gene_id "CUFF.38"; transcript_id "CUFF.38.1"; exon_number "1"; FPKM "69.4702369875"; frac "1.000000"; conf_lo "60.274596"; conf_hi "71.712087"; cov "143.103653";
Sb_20131119_contig_100040	Cufflinks	exon	945	1966	1000	.	.	gene_id "CUFF.16"; transcript_id "CUFF.16.1"; exon_number "1"; FPKM "180.540054118"; conf_lo "1.000000"; conf_hi "23.150255"; cov "44.484805";
Sb_20131119_contig_100121	Cufflinks	transcript	2079	2276	1000	-	.	gene_id "CUFF.16"; transcript_id "CUFF.16.1"; exon_number "1"; FPKM "180.540054118"; conf_lo "1.000000"; conf_hi "23.150255"; conf_lo "23.43223"; conf_hi "33.43223"; cov "88.805088";
Sb_20131119_contig_100121	Cufflinks	exon	2079	2276	1000	-	.	gene_id "CUFF.19"; transcript_id "CUFF.19.1"; exon_number "1"; FPKM "40.1295751927"; conf_lo "1.000000"; conf_hi "23.43223"; cov "88.805088";
Sb_20131119_contig_100121	Cufflinks	transcript	1510	2616	1000	.	.	gene_id "CUFF.22"; transcript_id "CUFF.22.1"; exon_number "1"; FPKM "10.2575159127"; frac "1.000000"; conf_lo "11.194395"; conf_hi "24.833916";
Sb_20131119_contig_100121	Cufflinks	exon	1510	2616	1000	.	.	gene_id "CUFF.21"; transcript_id "CUFF.21.1"; exon_number "1"; FPKM "11.6559124939"; conf_lo "1.000000"; conf_hi "11.688997"; conf_lo "11.688997"; conf_hi "14.524101"; cov "32.588772";
Sb_20131119_contig_100121	Cufflinks	transcript	933	3862	1000	.	.	gene_id "CUFF.24"; transcript_id "CUFF.24.1"; exon_number "1"; FPKM "13.7863265731"; conf_lo "1.000000"; conf_hi "14.094946"; conf_lo "14.094946"; conf_hi "14.805328"; cov "32.698360";
Sb_20131119_contig_100121	Cufflinks	exon	933	3862	1000	.	.	gene_id "CUFF.18"; transcript_id "CUFF.18.1"; exon_number "1"; FPKM "44.1667134200"; conf_lo "1.000000"; conf_hi "34.717689"; conf_lo "34.717689"; conf_hi "42.267104"; cov "77.065569";
Sb_20131119_contig_100080	Cufflinks	transcript	2857	6365	1000	.	.	gene_id "CUFF.18"; transcript_id "CUFF.18.1"; exon_number "2"; FPKM "44.1667134200"; conf_lo "1.000000"; conf_hi "34.717689"; conf_lo "34.717689"; conf_hi "42.267104"; cov "77.065569";
Sb_20131119_contig_100080	Cufflinks	exon	2857	6365	1000	.	.	gene_id "CUFF.22"; transcript_id "CUFF.22.1"; exon_number "1"; FPKM "10.2575159127"; frac "1.000000"; conf_lo "11.194395"; conf_hi "24.833916";
Sb_20131119_contig_100080	Cufflinks	transcript	39	135	1000	-	.	gene_id "CUFF.21"; transcript_id "CUFF.21.1"; exon_number "1"; FPKM "11.6559124939"; conf_lo "1.000000"; conf_hi "11.688997"; conf_lo "11.688997"; conf_hi "14.524101"; cov "32.588772";
Sb_20131119_contig_100080	Cufflinks	exon	39	135	1000	-	.	gene_id "CUFF.24"; transcript_id "CUFF.24.1"; exon_number "1"; FPKM "13.7863265731"; conf_lo "1.000000"; conf_hi "14.094946"; conf_lo "14.094946"; conf_hi "14.805328"; cov "32.698360";
Sb_20131119_contig_100080	Cufflinks	transcript	207	1041	1000	-	.	gene_id "CUFF.18"; transcript_id "CUFF.18.1"; exon_number "2"; FPKM "44.1667134200"; conf_lo "1.000000"; conf_hi "34.717689"; conf_lo "34.717689"; conf_hi "42.267104"; cov "77.065569";
Sb_20131119_contig_100080	Cufflinks	exon	207	1041	1000	-	.	gene_id "CUFF.49"; transcript_id "CUFF.49.1"; exon_number "1"; FPKM "49.5933867311"; conf_lo "1.000000"; conf_hi "35.335602"; conf_lo "43.965643"; cov "84.354408";
Sb_20131119_contig_10026	Cufflinks	transcript	1	1002	1000	.	.	gene_id "CUFF.49"; transcript_id "CUFF.49.1"; exon_number "1"; FPKM "49.5933867311"; conf_lo "1.000000"; conf_hi "35.335602"; conf_lo "43.965643"; cov "84.354408";
Sb_20131119_contig_10026	Cufflinks	exon	1	1002	1000	.	.	gene_id "CUFF.25"; transcript_id "CUFF.25.1"; exon_number "1"; FPKM "43.3120736956"; conf_lo "1.000000"; conf_hi "12.379816"; conf_lo "12.379816"; conf_hi "22.244981"; cov "78.396470";
Sb_20131119_contig_10031	Cufflinks	transcript	26	377	1000	.	.	gene_id "CUFF.25"; transcript_id "CUFF.25.1"; exon_number "1"; FPKM "43.3120736956"; conf_lo "1.000000"; conf_hi "12.379816"; conf_lo "12.379816"; conf_hi "22.244981"; cov "78.396470";
Sb_20131119_contig_10031	Cufflinks	exon	26	377	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 "52.744998"; cov "584.846045";
Sb_20131119_contig_10032	Cufflinks	transcript	5174	5386	1000	.	.	gene_id "CUFF.35"; transcript_id "CUFF.35.1"; exon_number "1"; FPKM "5.916726598"; conf_lo "1.000000"; conf_hi "6.795913"; cov "12.795227";
Sb_20131119_contig_10032	Cufflinks	exon	5174	5386	1000	.	.	gene_id "CUFF.35"; transcript_id "CUFF.35.1"; exon_number "1"; FPKM "5.916726598"; conf_lo "1.000000"; conf_hi "6.795913"; cov "12.795227";
Sb_20131119_contig_10032	Cufflinks	transcript	160	3305	1000	.	.	gene_id "CUFF.30"; transcript_id "CUFF.30.1"; FPKM "90.3397464519"; conf_lo "1.000000"; conf_hi "9.089087"; conf_lo "9.089087"; conf_hi "19.663630"; cov "118.213868";
Sb_20131119_contig_100403	Cufflinks	transcript	822	1053	1000	.	.	gene_id "CUFF.30"; transcript_id "CUFF.30.1"; exon_number "1"; FPKM "90.3397464519"; conf_lo "1.000000"; conf_hi "9.089087"; conf_lo "9.089087"; conf_hi "19.663630"; cov "118.213868";
Sb_20131119_contig_100403	Cufflinks	exon	822	1053	1000	.	.	gene_id "CUFF.32"; transcript_id "CUFF.32.1"; exon_number "1"; FPKM "4.9761112120"; conf_lo "1.000000"; conf_hi "4.259535"; conf_lo "4.259535"; conf_hi "6.262152"; cov "11.356891";
Sb_20131119_contig_100412	Cufflinks	transcript	226	2367	1000	.	.	gene_id "CUFF.36"; transcript_id "CUFF.36.1"; exon_number "1"; FPKM "55.9644866986"; conf_lo "1.000000"; conf_hi "52.744998"; cov "119.323264";
Sb_20131119_contig_100412	Cufflinks	exon	226	2367	1000	.	.	gene_id "CUFF.36"; transcript_id "CUFF.36.1"; exon_number "1"; FPKM "55.9644866986"; conf_lo "1.000000"; conf_hi "52.744998"; cov "119.323264";
Sb_20131119_contig_100831	Cufflinks	transcript	1415	3072	1000	.	.	gene_id "CUFF.36"; transcript_id "CUFF.36.1"; exon_number "1"; FPKM "55.9644866986"; conf_lo "1.000000"; conf_hi "52.744998"; cov "119.323264";
Sb_20131119_contig_100831	Cufflinks	exon	1415	3072	1000	.	.	gene_id "CUFF.31"; transcript_id "CUFF.31.1"; FPKM "5.2492336693"; frac "1.000000"; conf_lo "3.07818"; conf_hi "5.98178"; cov "12.544434";
Sb_20131119_contig_100843	Cufflinks	transcript	287	3748	1000	.	.	gene_id "CUFF.31"; transcript_id "CUFF.31.1"; exon_number "1"; FPKM "5.2492336693"; frac "1.000000"; conf_lo "3.07818"; conf_hi "5.98178"; cov "12.544434";
Sb_20131119_contig_100843	Cufflinks	exon	287	3748	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";
Sb_20131119_contig_100832	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";

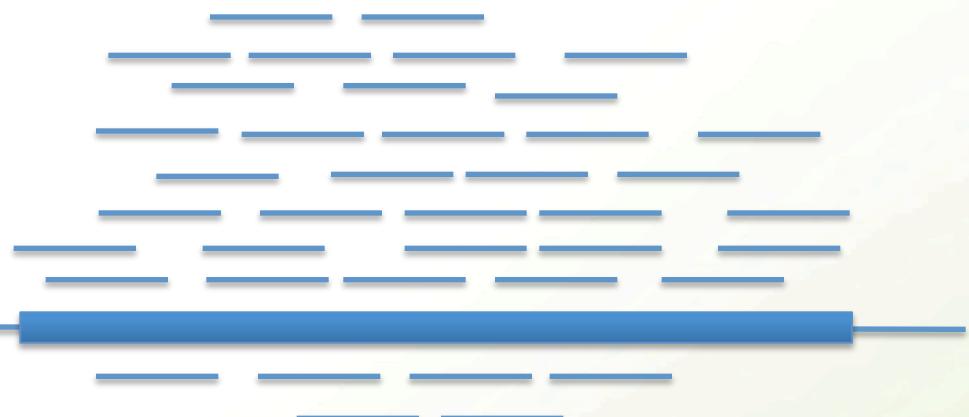
# GTF file format

<b>Seqid</b>	<b>source</b>	<b>type</b>	<b>start</b>	<b>end</b>	<b>score</b>	<b>strand</b>	<b>phase</b>	<b>attributes</b>
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

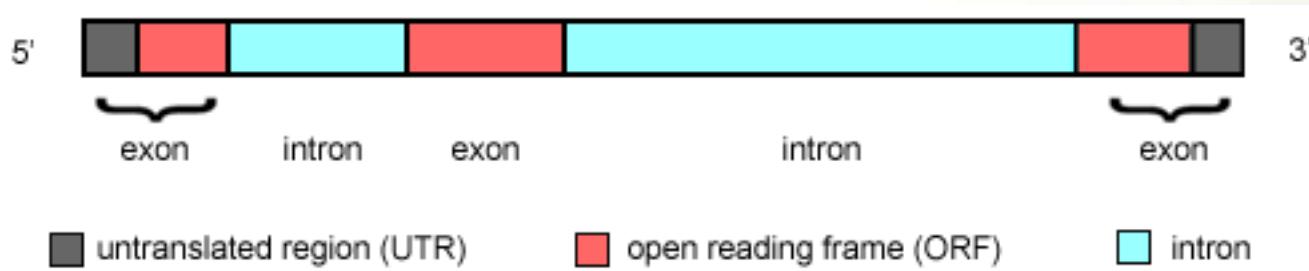
# Why is annotation important?

RNA-seq reads



# There are two major parts of annotation

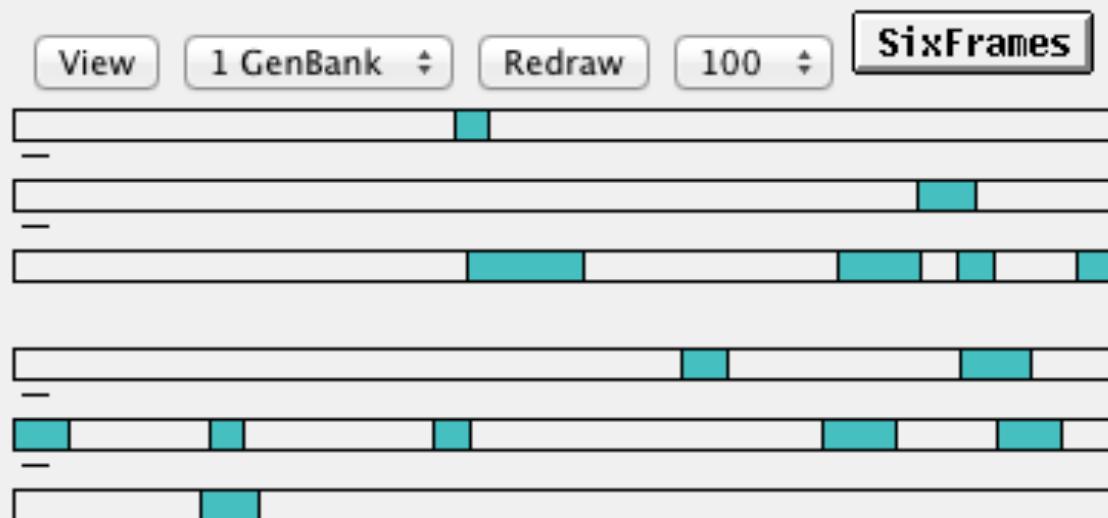
- 1) Structural: Find out where the regions of interest (usually genes) are in the genome and what they look like. How many exons/introns? UTRs? Isoforms?



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

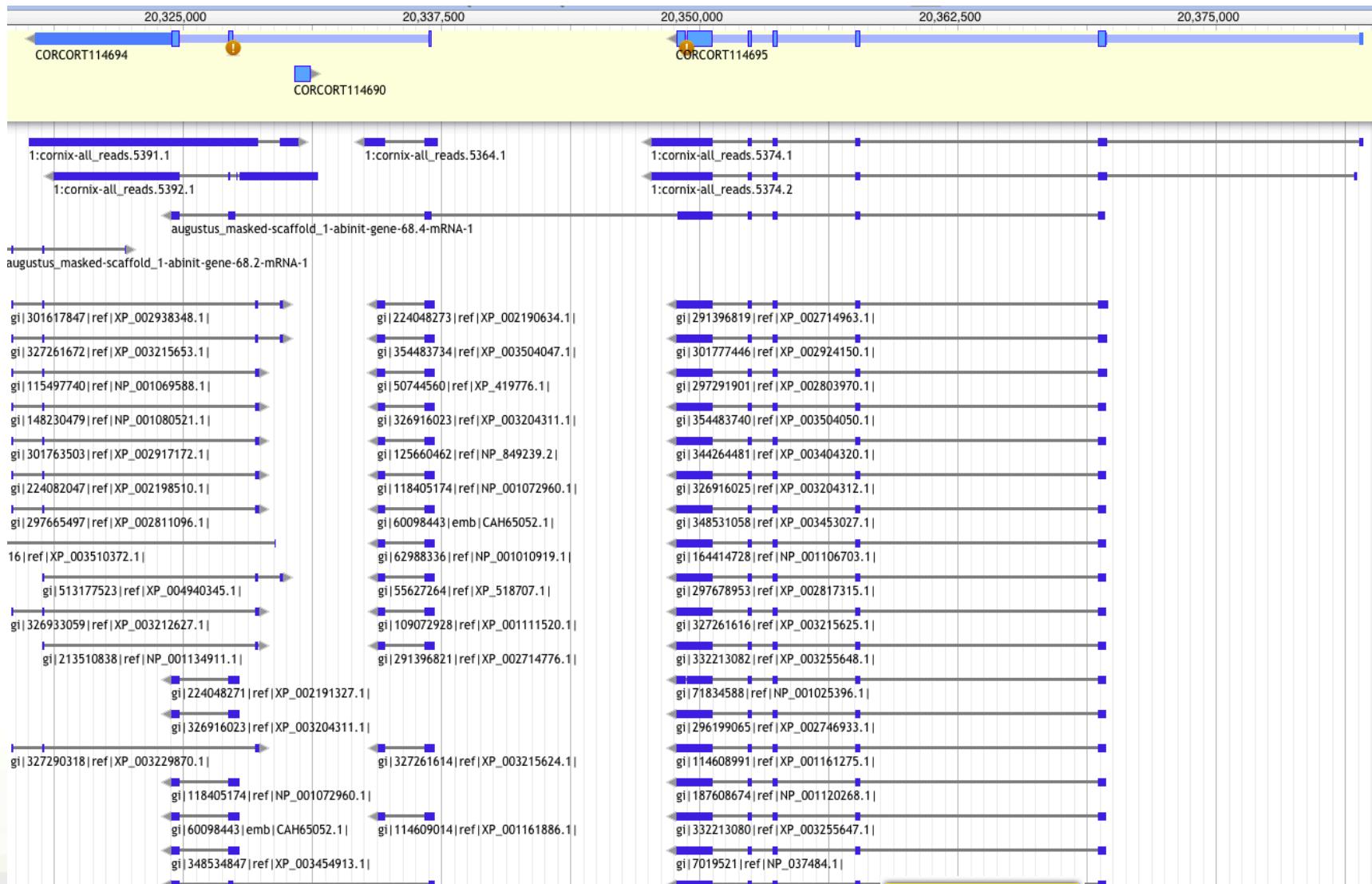
# Open reading frames

## Anonymous



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



# Combine data - use Maker!

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



MAKER  
Annotate this!

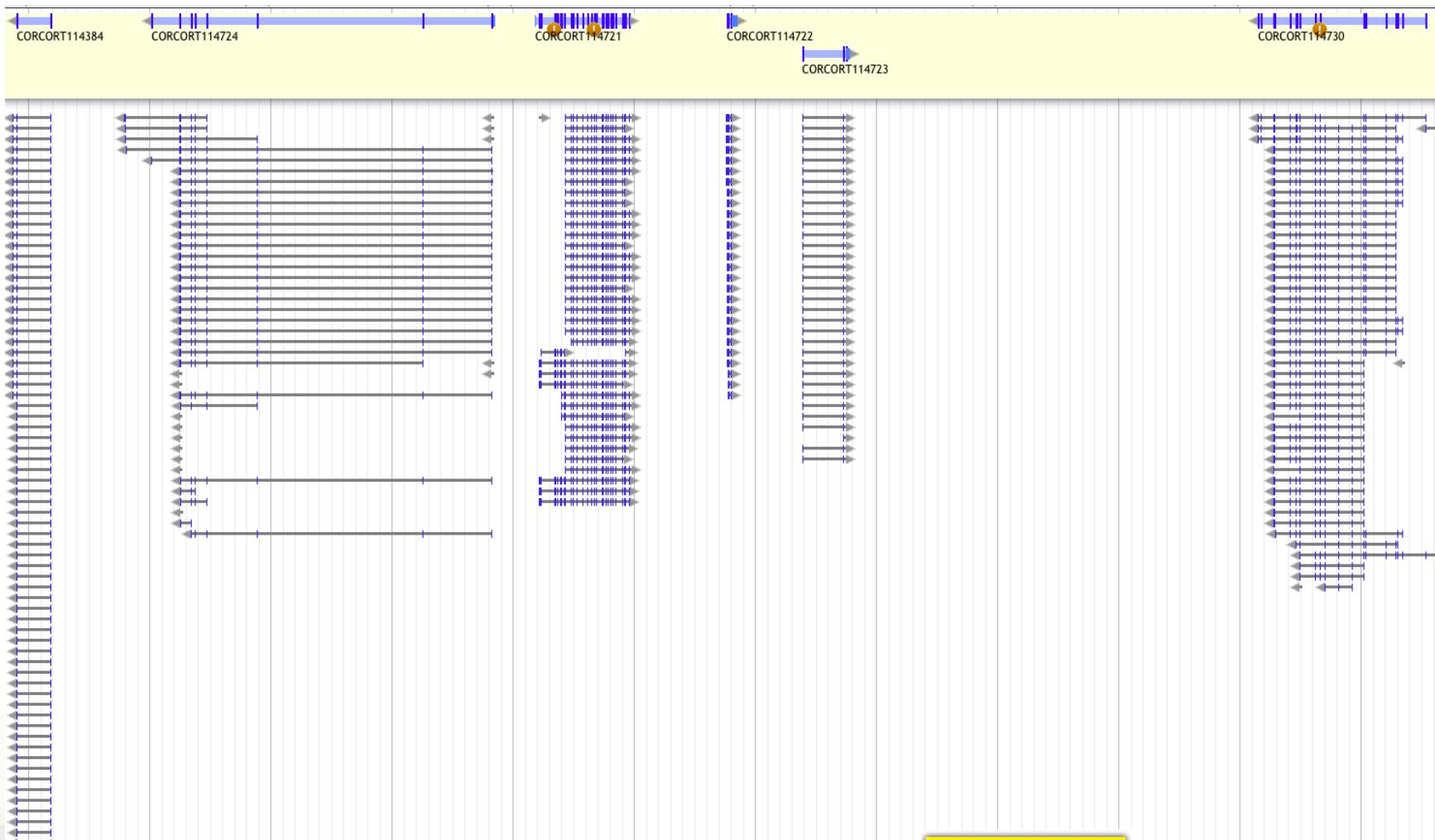
Combined annotation

# Transcriptomes are different but have their own challenges

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

```
>asmb1_2719
AGCACCTAGACGAGATGGGAGGTCTCTCTTGCTGTGCAGAGGCAGATCTCTTTCC
AACACCTAGCAGTATGAACACTAGTGAGCTCTGACTGTTTCCAGTGGTAATGAGGTGTGA
CCCGCTGCACTGCAACTGAAATTCTTCAGTCCCCGAGGCCAGCCCAGCTGTGGGC
AATGCTTTGGTGTGCTGTGACCATTC
>asmb1_2702
GTCCTGACTGGGAATGCCCTGGAGCAGAACCTTGCATGGATAAGGACACTACATT
CTGGTGTAAAGGTGAATATAACCTTCAAGGTTAAAGGATGACATTAAATTCAATTACAGCT
TGCTCTTGTAAAGCTAACAGGTTAACTAACAAAGCTATACTGTGACTCACCCCTTAGATCA
ATAGCTGGGAAACATCACCTCCCCAAATACTCCACCTTAACTGCACTCTTGAAAG
AAGTACAGGCCAGGTTAGCTGATCCATCTGTGCTAACTGCTGTGCTTACAAGCTG
CAATATTTTAAACACCAGAACATTGGTAGAGGTTAACATCAGCAAGCTTCAATT
TACAGCAGGTTAACCTCTGAACACTGTGATCACTGATATTTGGTCAAGTCAGATGT
CTTGTAGTGTCTT
>asmb1_2701
ACAAACAAAACAAAAAAACAAAGGAACAAAGCAAAAAAAACATCATACAATCCCAGT
TGTCAGAGCTTACTGTGAAATCAATCTGGATCAAACAAATGAAAGCTTCCAGA
TTCTGTATTCAGGCTGAGACAAGTTGTTAAACTCTTCCAGAAATTGCAACAGCTG
CAGGGTAACATCTTAATGCAACCTCTGTACAGAAATGCAAGACCTTAACTCTT
CAGGCCCTCCCCAGTCAACACCCAGTATAAACTCTGCCCTTCACTTGTGGAATATCTA
TCATAAGGGAAAGCATTTTTAGGCTGAGAAATACAATCCACCTTGTGAGGCCGGTCA
GCATATCATGGCTATGCTGTGATAGGTTGACCAAGCACTCTGTGAGAAATA
CTTAGAGGTGACCTAACAGGTCATTTGCAACTAACCTGGTGGCAGTATCTTTA
TTCCAACACTCCCACCTTCCCCAGAGAACAAAGCTGTATGGCAGTAGCAGTGTGTTG
AAGGTAACCTGCACTGTACTAGTAGCTGCAAGGCCAACACTTCCACACTAGCCAG
CTAGTCTAAAGTAACCTTCTGCAACAGGAAGAACATGGAACACACAGGCCAACCTTCAAG
AGGATCTGAGGCTGAGCTGCTTCCAGGAGGCCATGGGTCAAGCAGTACAGAAGG
GCACTGATAAGGTGCTTCAACCAAGCTGGCACAGAGGGCTGATCAGGAAA
ACCCACCATCAGCACAAAAGGAGCCCTGCAAACTCAGCCAGTGTAGGTTACTGGGTG
GAAATCAACTCTGCTGTGAAAGCTCTGTATACCCACATTACCTCCAGTGTACT
CAGAACACAAAGAGAGGAAATGGGAGGGACAGGAATGTGCACTGAGGAAAGCAGGG
CATCTTGTGCTCAGGCTGCTGCACTGAGCTTCACTGCCAGGGCAGTCTGAGTCC
TACCGGTGAGGACATCTGTCATGTAATGCCCCCTCTGTGACAGCAATCTCAAG
TGGTCTTAAATGGCTCTCACTTGGGAGCTCACTGGCACAGCTACTGCCA
GGAAACAAAAGGCTCAACAGGGTGGGAAACAAACTATTGTCAGTGTGAGGAAATG
TGGATAGAAACAGGGTTGGTAACTCTGACTGATAAGAAGAGGAGGTGCAAGATAGG
CTGAGAAAGACTCTAGGTTGGAAACAAGCTGTATAAAAGCTTAAAGGGTGAATGT
GAAAAAAATATGCCGAGCAGAATAAAAGGACTCTATTCCATCCCACCTGGAATCTGA
ACCCAGTTCAGAGTAATGAAGGGCTTGTGTTGTGCTGTGAGGAGATCACCAGT
AAGCAATGCTCAGGCTGCCCTCTGAGCTCAGGAAAGGACTCACCCCTGAGA
GGTTGATGTGCTGCAACAGGCCACAGCTGGCATTTGGAGGTGTTAGGTGCTT
GGGTATGCTAAACCAACGTTGAATGAAAGTGTCTGTCACTCACTGAGTGAAGGGAGA
GAACATTTCAGGAGCTGGAAACTCTGGAGGGACTGACTTAATTGTTGAGTCAAAAC
GGTCTGTGCTGGAAATACTCTGCAATGCTCTGAAAGTGTGTTGGGGGGAGGGCGTT
GCACTGAAGCTTGTGAGGAGGTATCACAAAGATATGCAATGCTAGCACAAGC
CAGAACCTTGCTGAGGCCAGGGTCTACTCTGCTGCTGGGACAGTCTCCCTGTGAG
GTACAGTTGCACTCCCTGTGAGCTGCCAGCTGGAGGTGCTCCCTCTGT
CTGGAGGAGACTGCGCACACCTCTGGAGATTGGAGCTGTAATGTTGAGGCTCTG
TGAATGGAGACTGGCTCTGTGTTGGGGACCTTCTCTGGCCAGCAGACTGTTAG
CTGAGGCCAGGCCACTCAACTCAGCATAGTGTCCAGAAAGAAGTAAAGTCCCAT
CTGCTTCCCTGTGATCTAACCTGTCAAAGGCCATTAGGTTGGCATTGTTGCTTGAAT
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

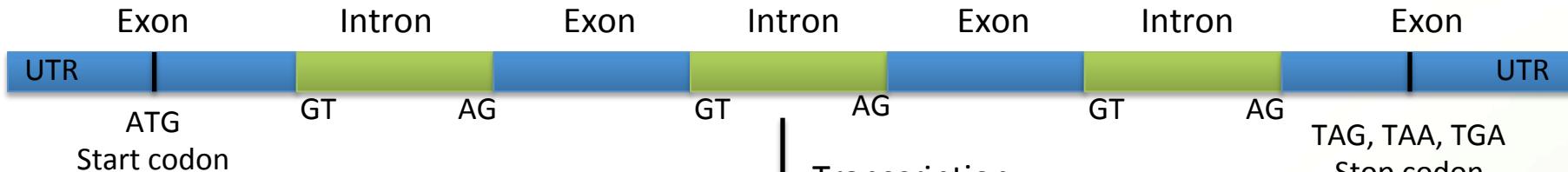
```
>ENSTGUP00000017616 pep:novel chromosome:taeGut3.2.4:8_random:2849599:2959678:-1 gene:ENSTGUG00000017338 transcript:ENSTGUT0000001801  
RSPNATEYNWHHLRYPKIPERLNPPAAAGPALSTAEGWMLPWNGQHPLLARAPGKGRER  
DGKELIKPKTFKFTLKKKKKKKKTFK  
>ENSTGUP00000017615 pep:novel chromosome:taeGut3.2.4:23_random:205321:209117:1 gene:ENSTGUG00000017337 transcript:ENSTGUT00000018017  
PDLRELVLMFEHLHRVRNGGFRNSEVKWPDRSPPPYHSFTPQAQKSFSLAGCSGESTKMG  
IKERMRLSSQRQGSRGRRQQHLGPPPLHRSPSPEDVAEATSPTKVQKSWSFNDRTRFRASL  
RLKPRIPAEGDCPPEDSGEERSSPCDLTFEDIMPRAVKTIRAVRILKFVAKRKFKETLR  
PYDVKDVIQEYQYSAGHLDMLGRIKSLQTRVEQIVGRDRALPADKKVREKGEKPALEAELVD  
ELSMMGRVVVKVERQVQSIEHKDLLLGLYSRCLRKGANSVLAAVRVPPGEPDVTSDYQ  
SPVEHEDISTSAQSLISRLASTNMD
```

# Data used - Proteins

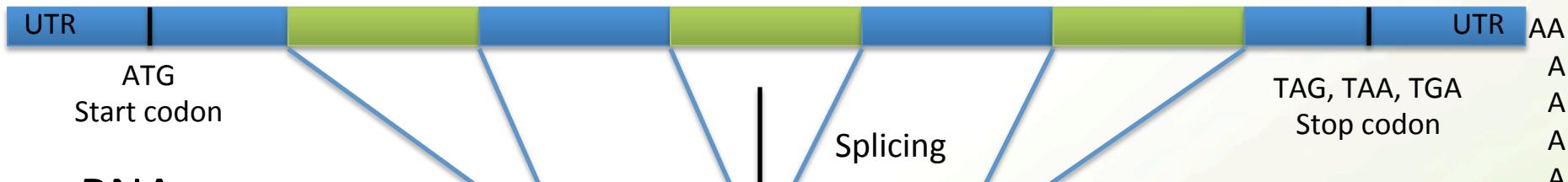
- Maker will align proteins for you: Blast -> Exonerate
- Blast is not structure aware, Exonerate is (splice sites, start/stop codons)
- Preferred file-format: fasta

# RNA-seq

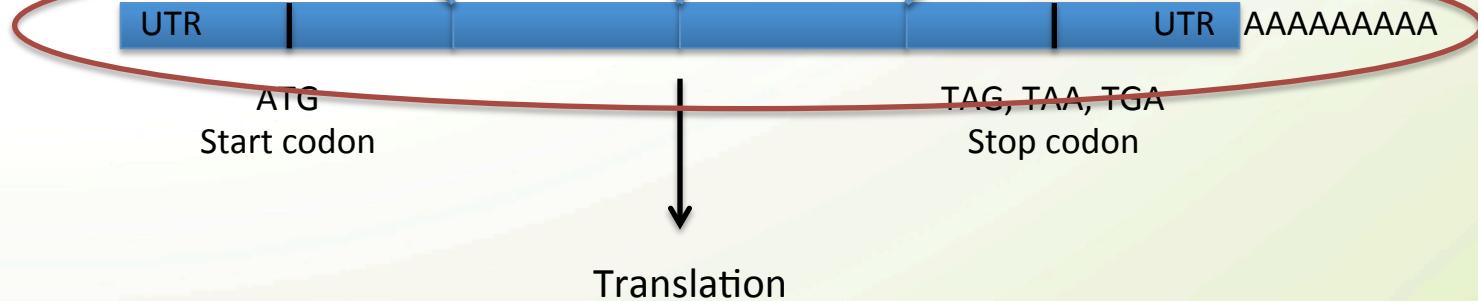
## DNA



## Pre-mRNA



## mRNA

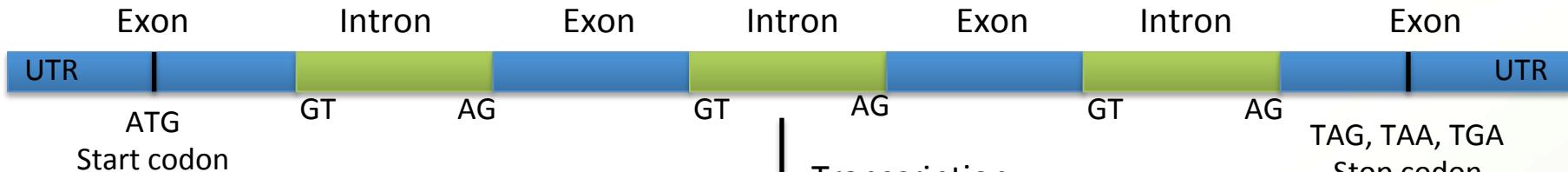


## Data used - RNA-seq

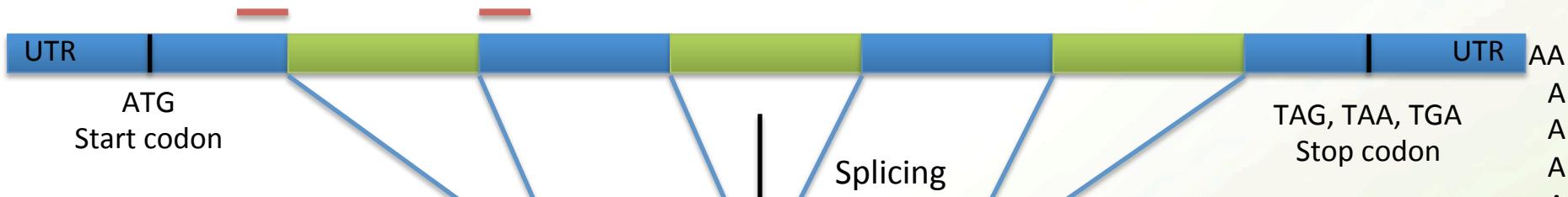
- Should always be included in an annotation project
- From the same organism as the genomic data  
=> unbiased
- Can be very noisy (tissue/species dependent), can include pre-mRNA
- PASA, or some other filtering method, often needed

# Spliced reads

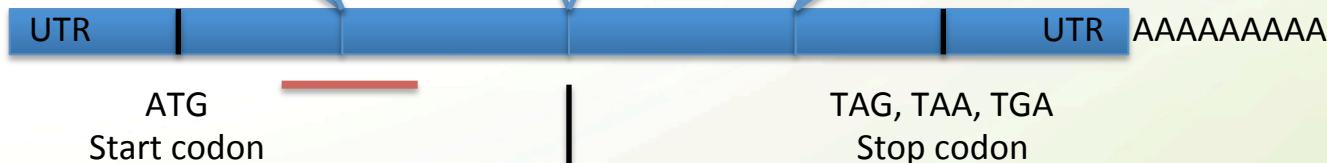
DNA



Pre-mRNA

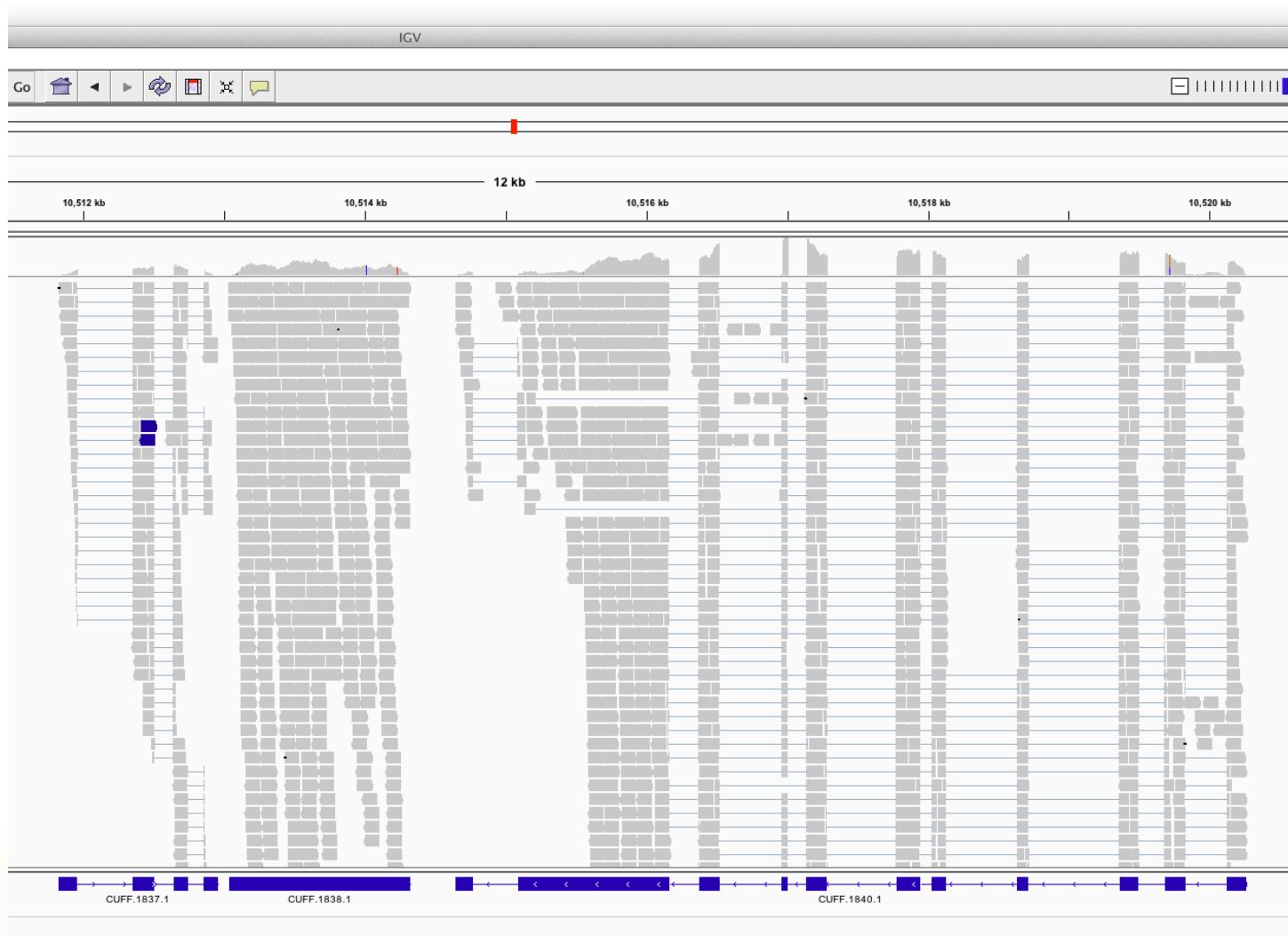


mRNA



Translation

# RNA-seq - Spliced reads

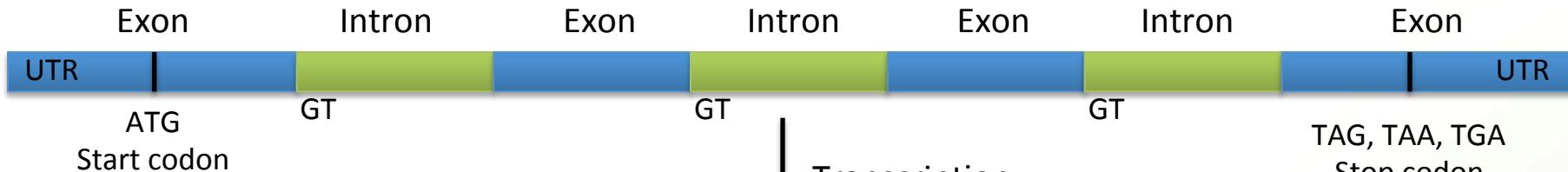


# Pre-mRNA

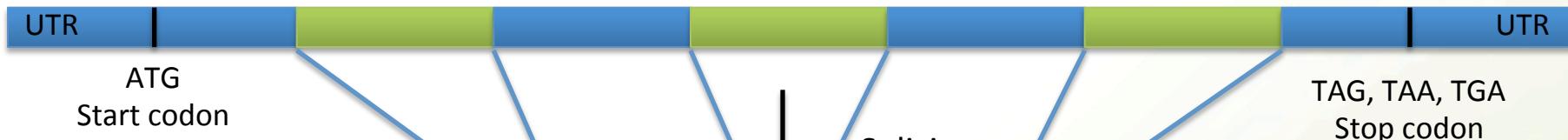


# Pre-mRNA

## DNA



## Pre-mRNA

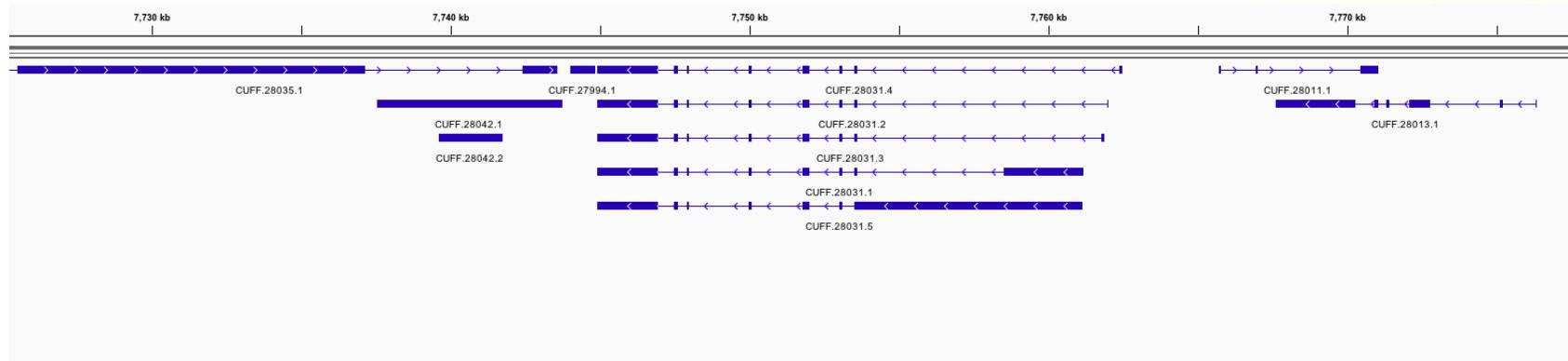


## mRNA

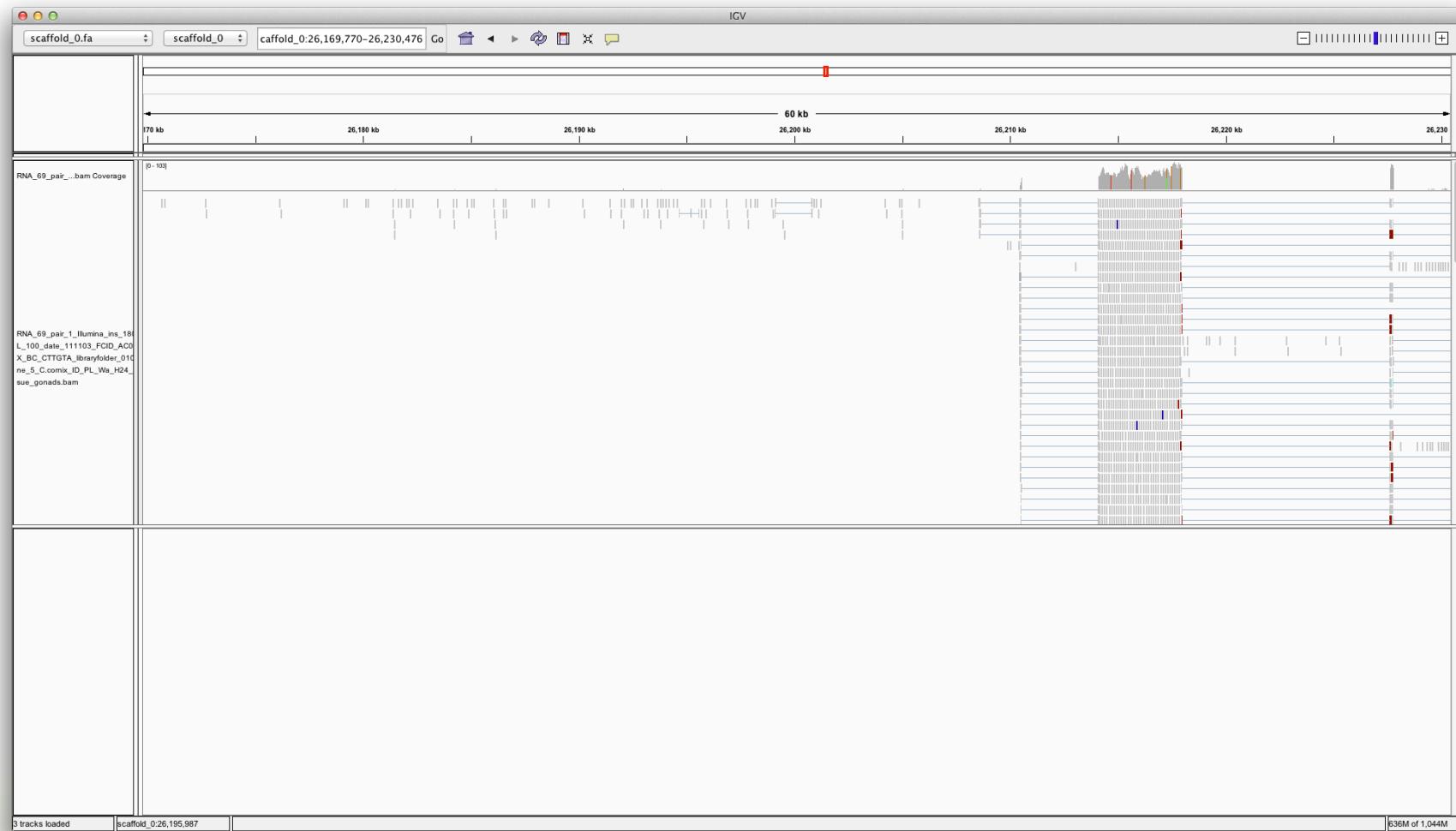


Translation

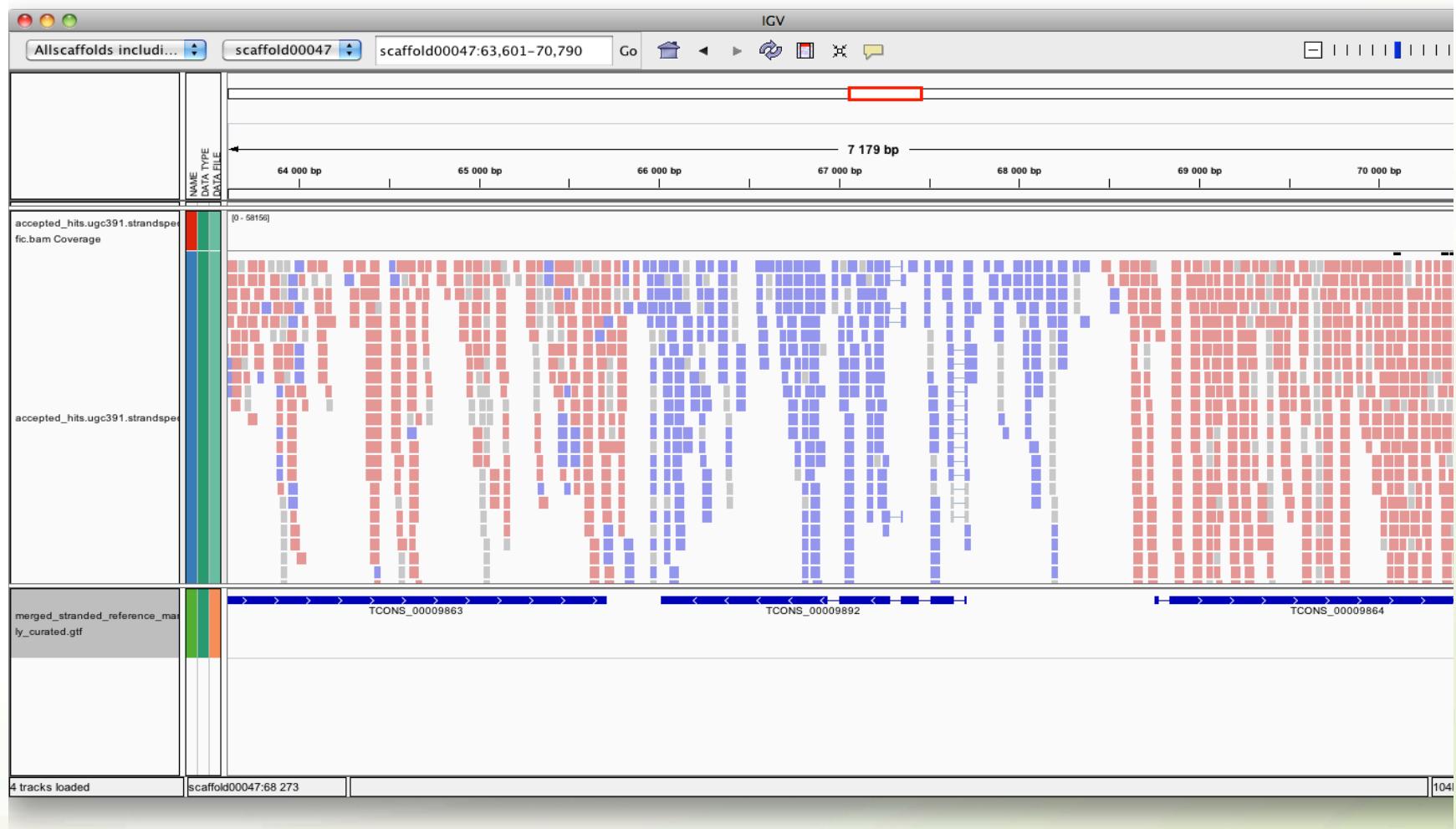
# Pre-mRNA



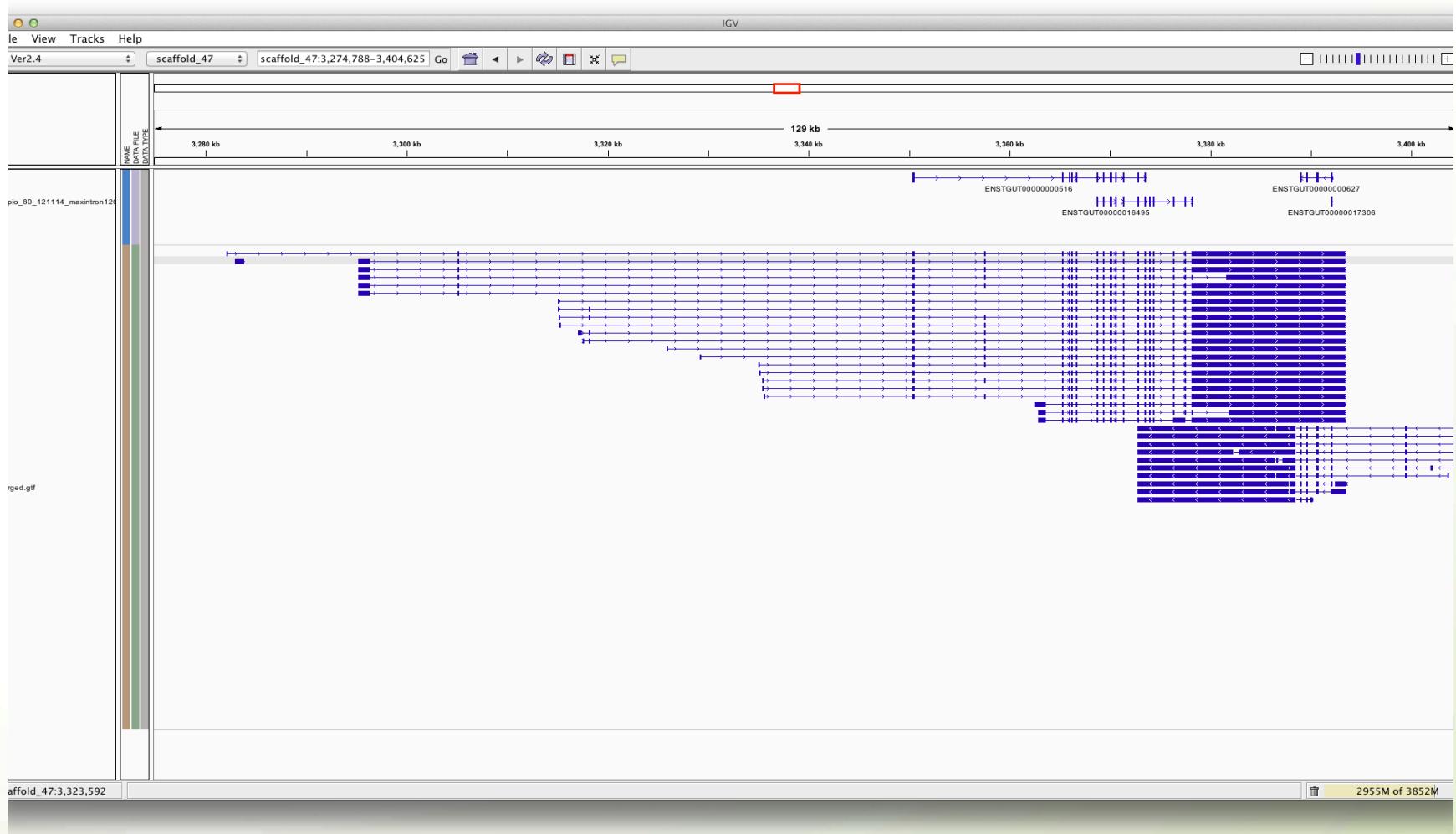
# Includes everything that is transcribed



# Stranded rna-seq

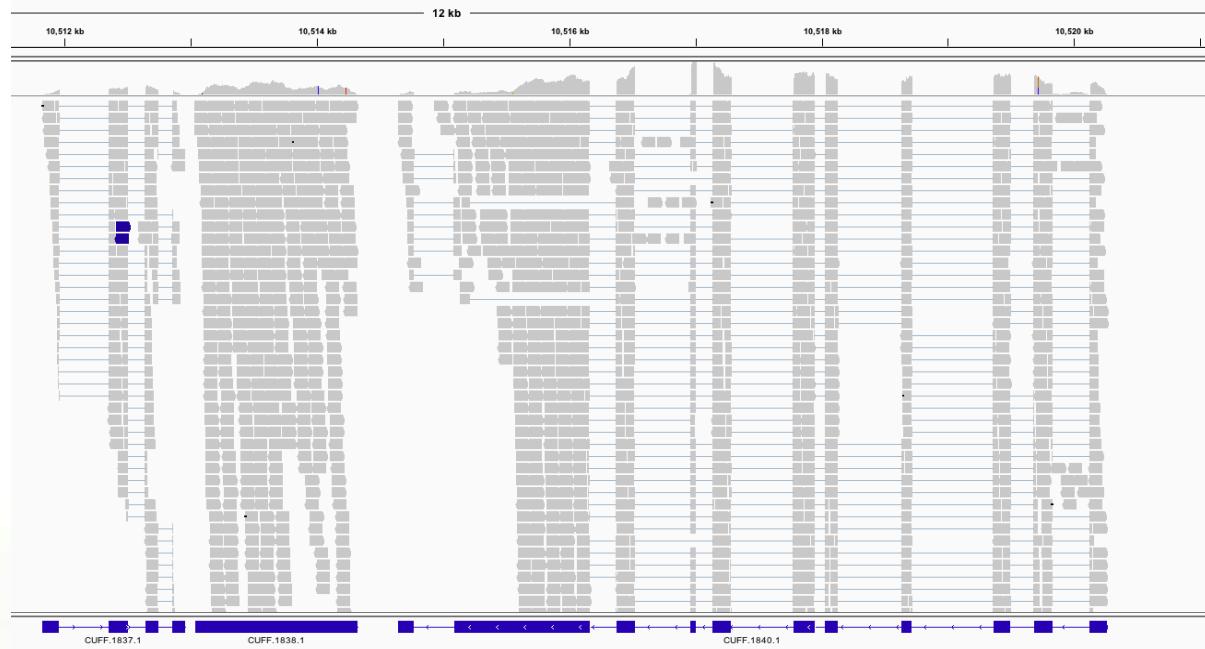


# Three-prime bias in polyA-selected rna-seq



# How to use RNA-seq

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

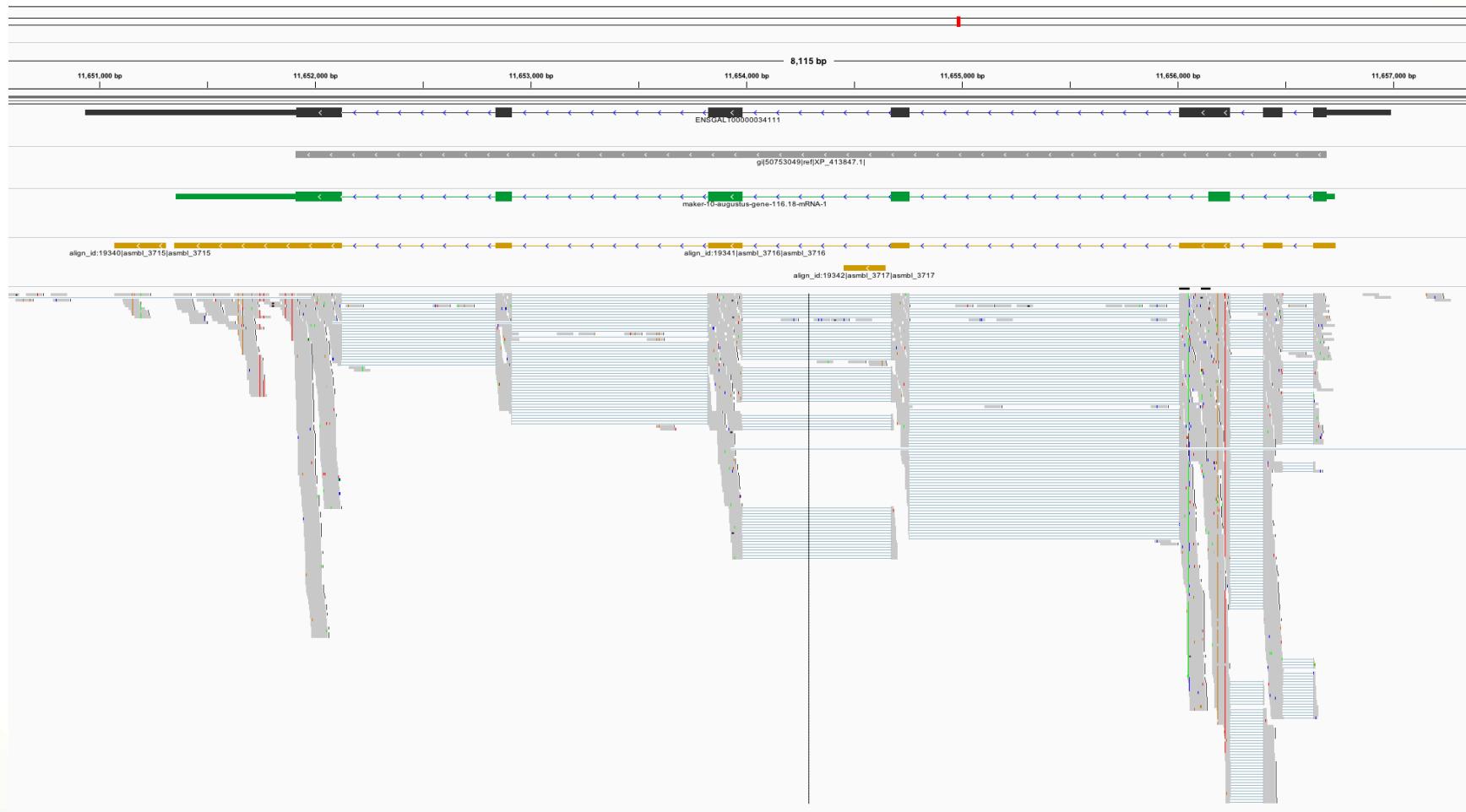


# How to use RNA-seq

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

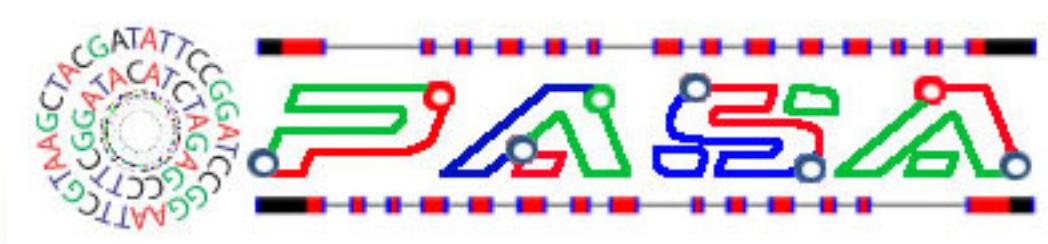


# Mapped Trinity-assembled transcripts



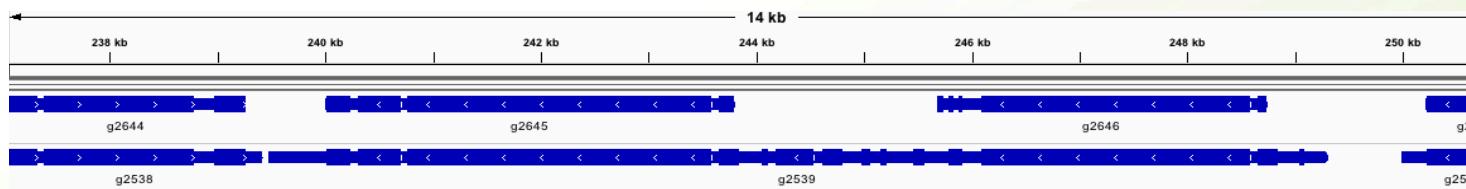
# How to use RNA-seq

- 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



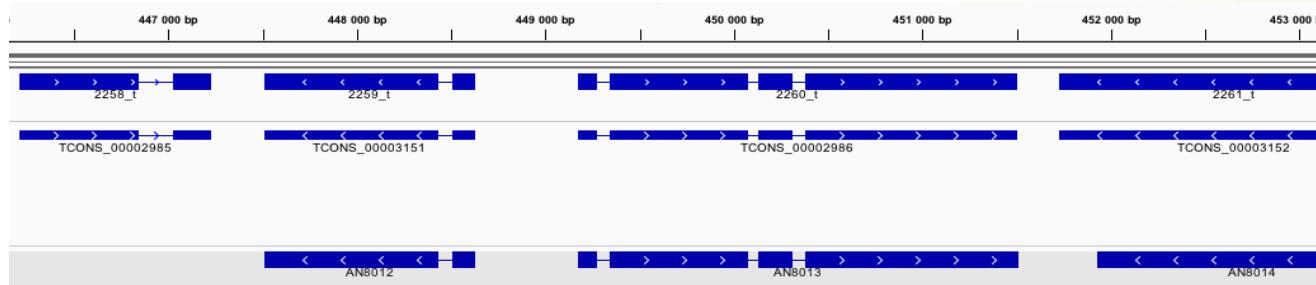
# Ab initio gene finders are used in Maker

- 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!

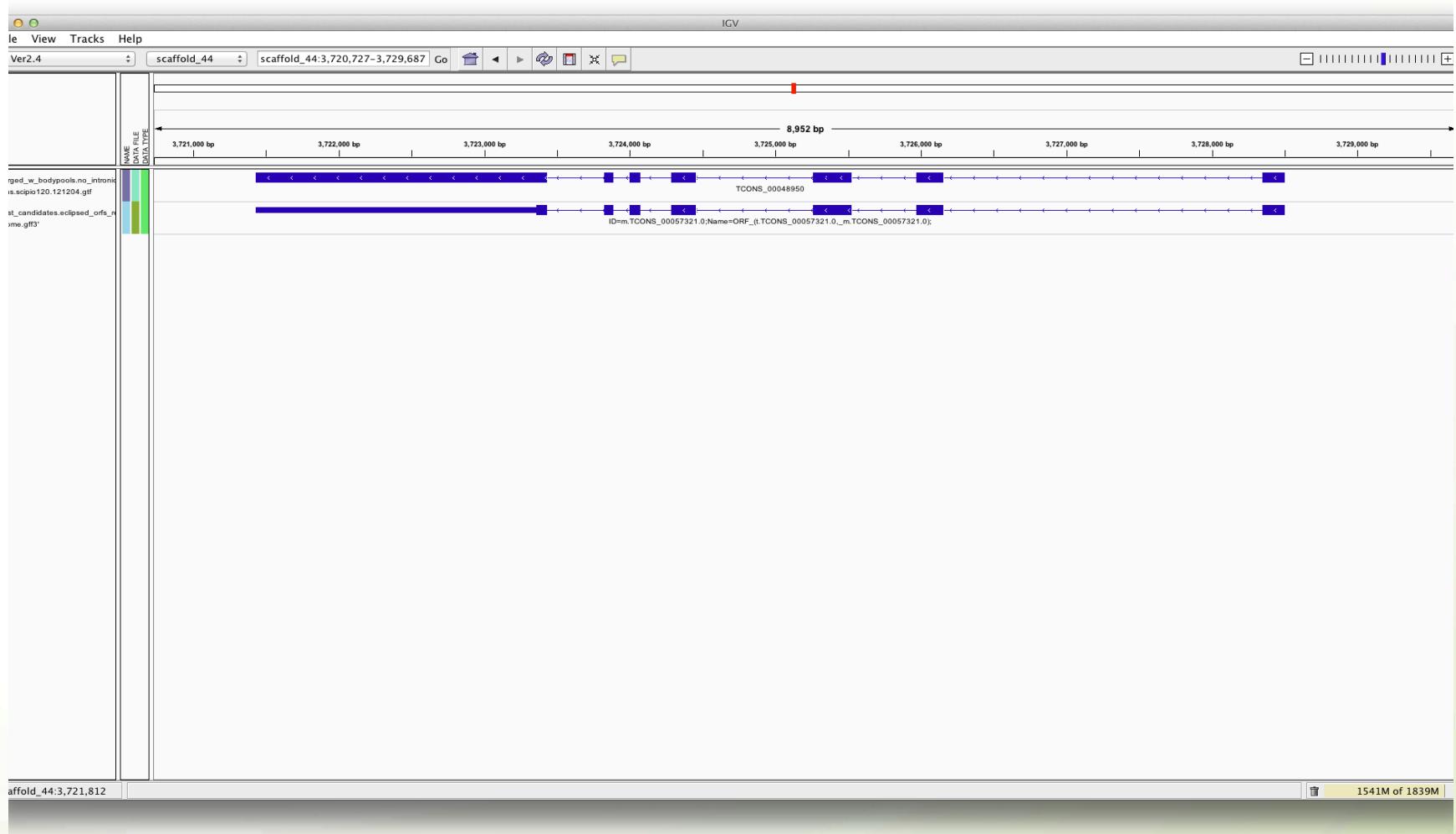


MAKER  
Annotate this!

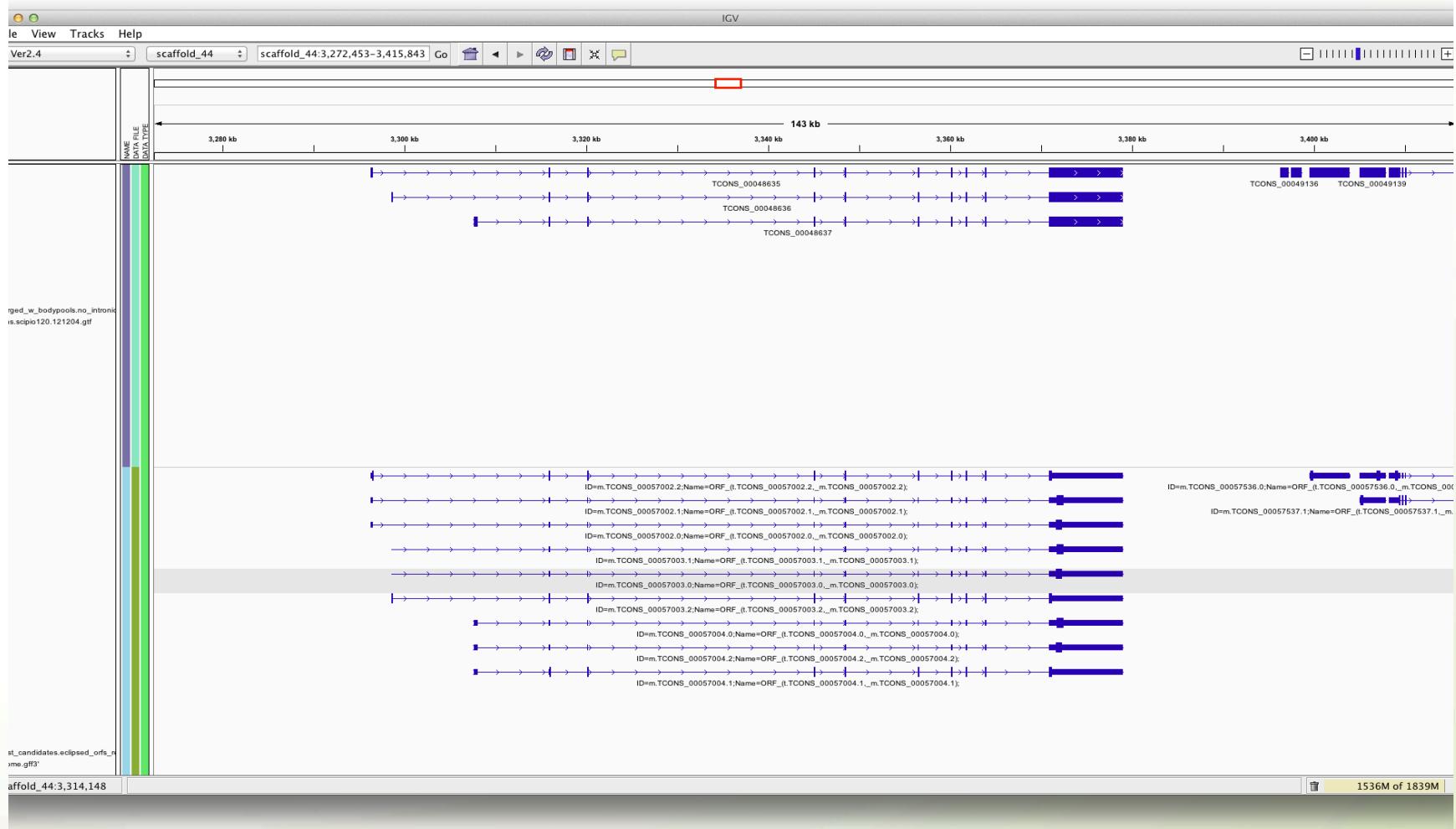
# Transcript annotation

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

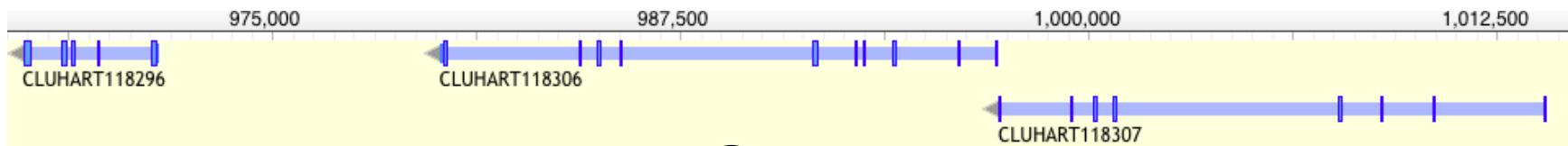
# Transdecoder



# Transdecoder



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



Insulin receptor?

Vesicle-trafficking protein?

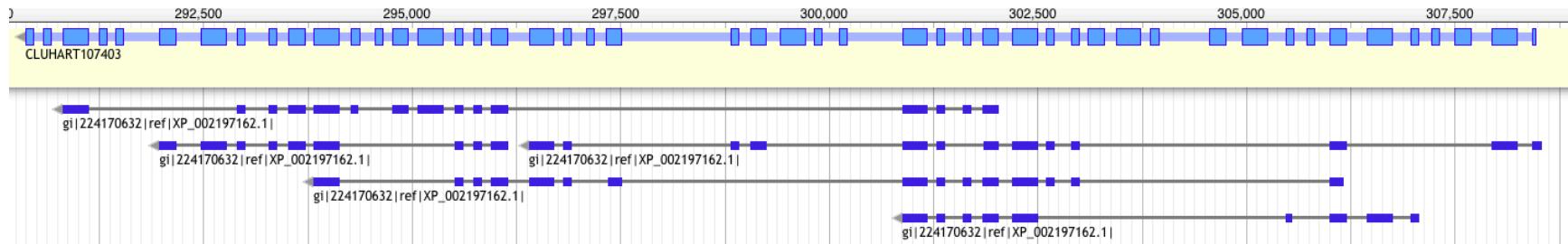
Alcohol dehydrogenase?

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



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

A screenshot of a web browser window showing the Trinotate homepage. The title bar reads "Trinotate: Trinity Transcriptome Functional Annotation". The page features the "Trinotate" logo at the top center. Below it are logos for NCBI BLAST, HMMER/Pfam, UniProt, eggNOG version 3.0, SQLite, and the Gene Ontology. A flowchart at the bottom illustrates the analysis pipeline: RNA-Seq leads to Trinity, which leads to Transcripts/Proteins, which leads to Functional Data, which finally leads to Discovery. A blue banner at the bottom states "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	#G...	GO IDs	Enzyme	InterPro
3884	gene_3884 GeneMark...	c6 transcription	977	20	1.0E-171	59.85%	7	<span style="color:blue;">F</span> transcription factor activity; <span style="color:red;">F</span> zinc ion binding; <span style="color:green;">P</span> regulation of transcription, DNA-dependent; <span style="color:blue;">C</span> transcription factor complex; <span style="color:red;">F</span> transporter activity; <span style="color:blue;">C</span> membrane; <span style="color:green;">P</span> transmembrane transport		IPR005829; IPR007219
3885	gene_3885 GeneMark...	hypothetical protein NFIA_039100 [Neosartorya fischeri NTRL 181]	312	20	1.0E-39	63.15%	1	<span style="color:blue;">C</span> viral capsid		no IPS match
3886	gene_3886 GeneMark...	sin3 complex subunit	870	20	0.0	73.2%	0			-
3887	gene_3887 GeneMark...	mitochondrial intermembrane space translocase subunit	87	20	1.0E-40	88.55%	5	<span style="color:blue;">F</span> metal ion binding; <span style="color:green;">P</span> rotein import into mitochondrial inner membrane; <span style="color:blue;">C</span> mitochondrial inner membrane; <span style="color:blue;">C</span> mitochondrial intermembrane space protein transporter complex; <span style="color:green;">P</span> transmembrane transport		IPR004217; PTHR11038 (PANTHER), PTHR11038:SF8 (PANTHER)
3888	gene_3888 GeneMark...	lysyl-tRNA synthetase	592	20	0.0	73.55%	7	<span style="color:blue;">C</span> ytosol; <span style="color:green;">P</span> auxin biosynthetic process; <span style="color:red;">F</span> nucleic acid binding; <span style="color:blue;">F</span> lysine-tRNA ligase activity; <span style="color:blue;">P</span> lysyl-tRNA aminoacylation; <span style="color:red;">F</span> ATP binding; <span style="color:green;">P</span> 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 GeneMark...	transcription factor conserved hypothetical protein [Aspergillus clavatus NTRL 1]	1569	20	0.0	70.9%	0			-
3890	gene_3890 GeneMark...	udp-gluc gal endoplasmic reticulum nucleotide	240	20	1.0E-51	56.25%	0	<span style="color:blue;">C</span> integral to membrane; <span style="color:green;">C</span> endoplasmic reticulum membrane; <span style="color:green;">P</span> transmembrane transport; <span style="color:green;">P</span> carbohydrate transport		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 Gene Ontology

Genome assembly Wiki SILL BILS projects Wikipedia Google Maps

the Gene Ontology

Search  go!

Downloads Tools Documentation Projects About 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:

GO!

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 A
- 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 activity, 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	#G...	GO IDs	Enzyme	InterPro
3884	gene_3884 GeneMark...	c6 transcription	977	20	1.0E-171	59.85%	7	<span style="color:blue;">F</span> transcription factor activity; <span style="color:red;">F</span> zinc ion binding; <span style="color:green;">P</span> regulation of transcription, DNA-dependent; <span style="color:blue;">C</span> transcription factor complex; <span style="color:red;">F</span> transporter activity; <span style="color:blue;">C</span> membrane; <span style="color:green;">P</span> transmembrane transport		IPR005829; IPR007219
3885	gene_3885 GeneMark...	hypothetical protein NFIA_039100 [Neosartorya fischeri NTRL 181]	312	20	1.0E-39	63.15%	1	<span style="color:blue;">C</span> viral capsid		no IPS match
3886	gene_3886 GeneMark...	sin3 complex subunit	870	20	0.0	73.2%	0			-
3887	gene_3887 GeneMark...	mitochondrial intermembrane space translocase subunit	87	20	1.0E-40	88.55%	5	<span style="color:blue;">F</span> metal ion binding; <span style="color:green;">P</span> rotein import into mitochondrial inner membrane; <span style="color:blue;">C</span> mitochondrial inner membrane; <span style="color:blue;">C</span> mitochondrial intermembrane space protein transporter complex; <span style="color:green;">P</span> transmembrane transport		IPR004217; PTHR11038 (PANTHER), PTHR11038:SF8 (PANTHER)
3888	gene_3888 GeneMark...	lysyl-tRNA synthetase	592	20	0.0	73.55%	7	<span style="color:blue;">C</span> ytosol; <span style="color:green;">P</span> auxin biosynthetic process; <span style="color:red;">F</span> nucleic acid binding; <span style="color:blue;">F</span> lysine-tRNA ligase activity; <span style="color:blue;">P</span> lysyl-tRNA aminoacylation; <span style="color:red;">F</span> ATP binding; <span style="color:green;">P</span> 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 GeneMark...	transcription factor conserved hypothetical protein [Aspergillus clavatus NTRL 1]	1569	20	0.0	70.9%	0			-
3890	gene_3890 GeneMark...	udp-gluc gal endoplasmic reticulum nucleotide	240	20	1.0E-51	56.25%	0	<span style="color:blue;">C</span> integral to membrane; <span style="color:green;">C</span> endoplasmic reticulum membrane; <span style="color:green;">P</span> transmembrane transport; <span style="color:green;">P</span> carbohydrate transport		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

www.ebi.ac.uk/interpro/interproscan.html

Reader

An Introduction to the Gene Ontology

About InterProScan 5 < InterPro < EMBL-EBI

Services | Research | Training | About us

EMBL-EBI

**InterPro**  
Protein sequence analysis & classification

Search InterPro... Examples: IPR020405, kinase, P51587, PF02932, GO:0007165

Home | Release notes | Training & tutorials | FAQs | Download | About InterPro | 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, Sébastien Pesquet, 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 \(324Kb\)](#)

---

**EMBL-EBI**

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

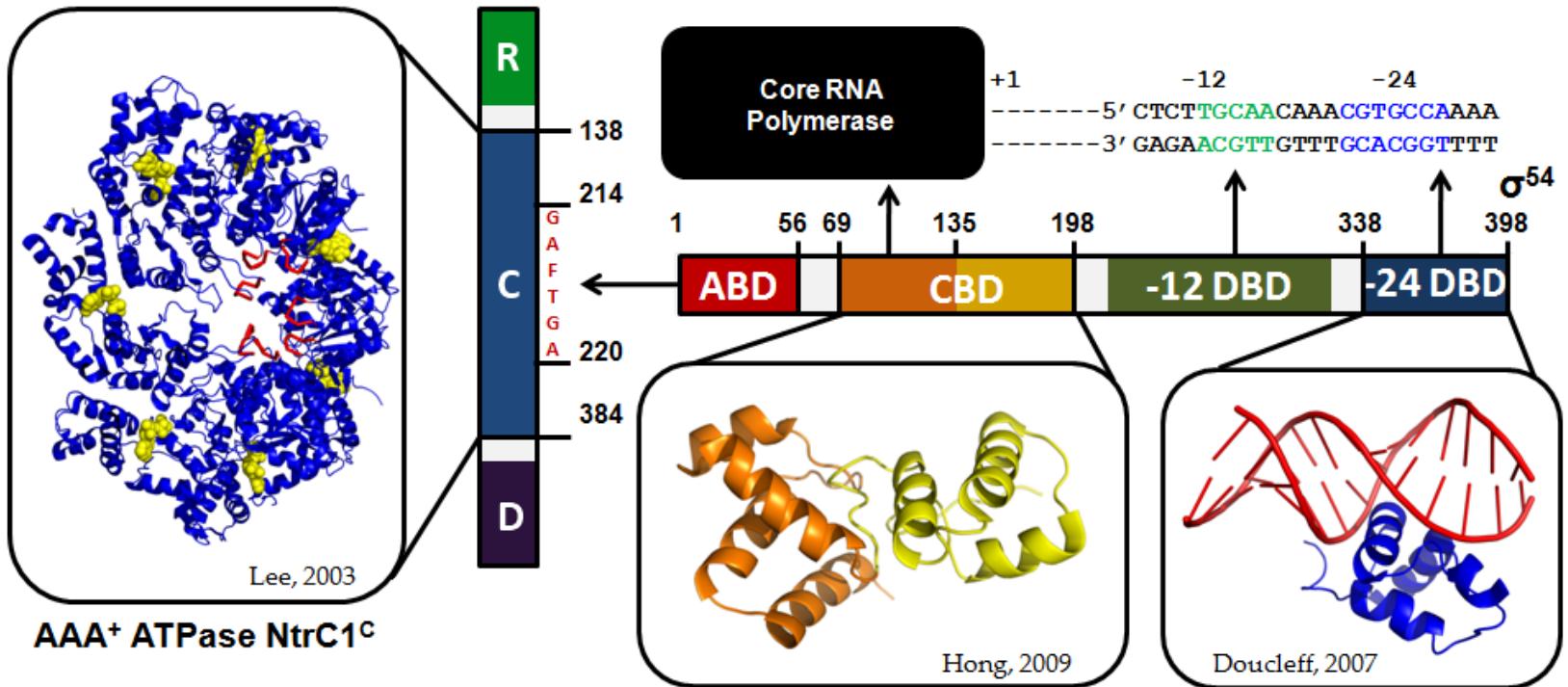
**Industry**

Overview  
Members Area  
Workshops  
SME Forum  
Contact Industry programme

**About us**

Overview  
Leadership  
Funding  
Background  
Collaboration  
Jobs  
People & groups  
News

# Sequence domains

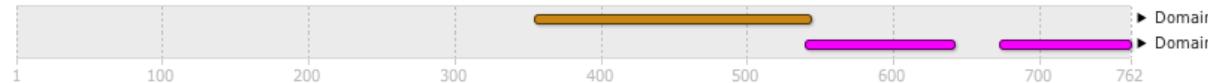


# Interproscan results

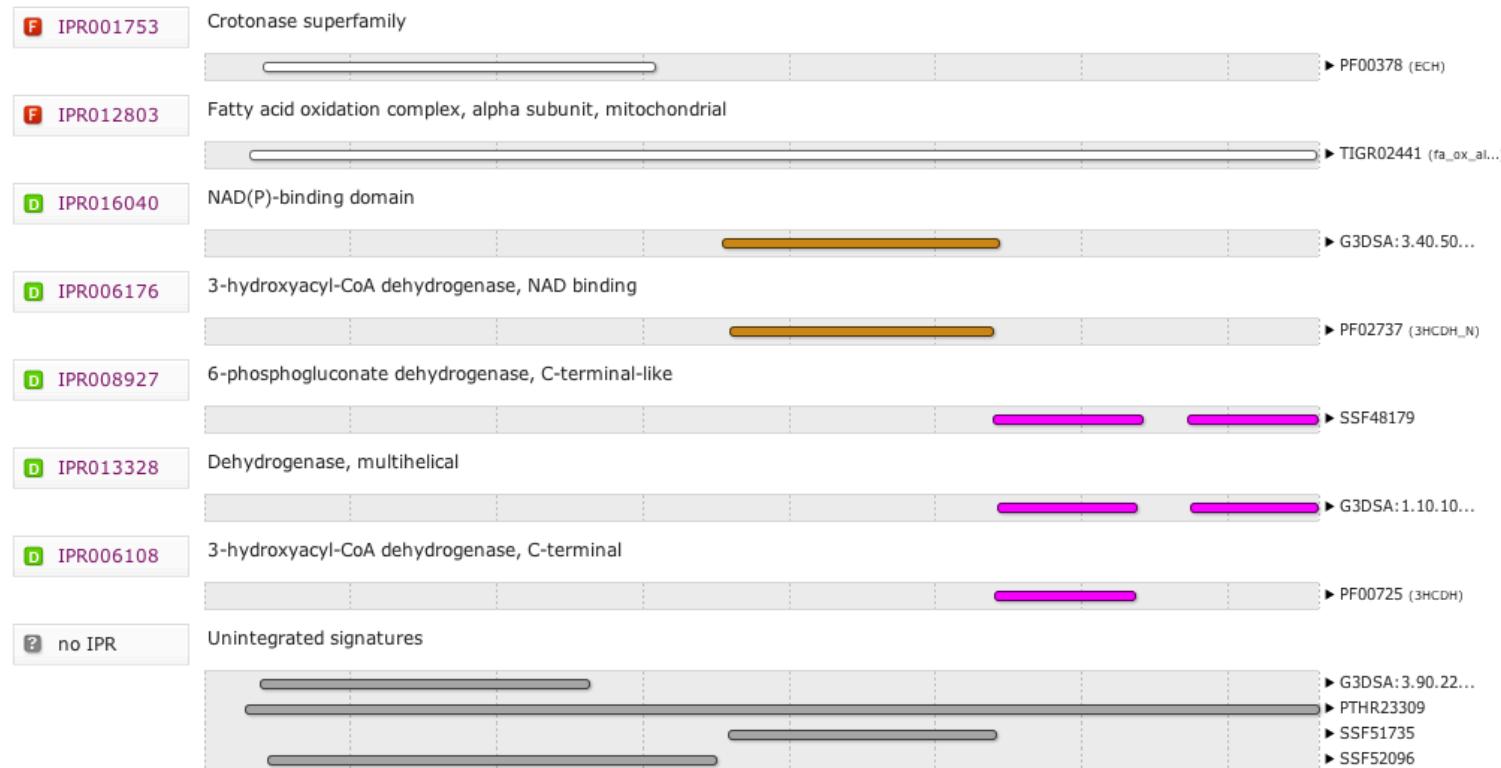
## Protein family membership

- ↳ 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	#G...	GO IDs	Enzyme	InterPro
3884	gene_3884 GeneMark...	c6 transcription	977	20	1.0E-171	59.85%	7	<span style="color:blue;">F</span> transcription factor activity; <span style="color:red;">F</span> zinc ion binding; <span style="color:green;">P</span> regulation of transcription, DNA-dependent; <span style="color:blue;">C</span> transcription factor complex; <span style="color:red;">F</span> transporter activity; <span style="color:blue;">C</span> membrane; <span style="color:green;">P</span> transmembrane transport		IPR005829; IPR007219
3885	gene_3885 GeneMark...	hypothetical protein NFIA_039100 [Neosartorya fischeri NTRL 181]	312	20	1.0E-39	63.15%	1	<span style="color:blue;">C</span> viral capsid		no IPS match
3886	gene_3886 GeneMark...	sin3 complex subunit	870	20	0.0	73.2%	0			-
3887	gene_3887 GeneMark...	mitochondrial intermembrane space translocase subunit	87	20	1.0E-40	88.55%	5	<span style="color:blue;">F</span> metal ion binding; <span style="color:green;">P</span> rotein import into mitochondrial inner membrane; <span style="color:blue;">C</span> mitochondrial inner membrane; <span style="color:blue;">C</span> mitochondrial intermembrane space protein transporter complex; <span style="color:green;">P</span> transmembrane transport		IPR004217; PTHR11038 (PANTHER), PTHR11038:SF8 (PANTHER)
3888	gene_3888 GeneMark...	lysyl-tRNA synthetase	592	20	0.0	73.55%	7	<span style="color:blue;">C</span> ytosol; <span style="color:green;">P</span> auxin biosynthetic process; <span style="color:red;">F</span> nucleic acid binding; <span style="color:blue;">F</span> lysine-tRNA ligase activity; <span style="color:blue;">P</span> lysyl-tRNA aminoacylation; <span style="color:red;">F</span> ATP binding; <span style="color:green;">P</span> 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 GeneMark...	transcription factor conserved hypothetical protein [Aspergillus clavatus NTRL 1]	1569	20	0.0	70.9%	0			-
3890	gene_3890 GeneMark...	udp-gluc gal endoplasmic reticulum nucleotide	240	20	1.0E-51	56.25%	0	<span style="color:blue;">C</span> integral to membrane; <span style="color:green;">C</span> endoplasmic reticulum membrane; <span style="color:green;">P</span> transmembrane transport; <span style="color:green;">P</span> carbohydrate transport		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

```

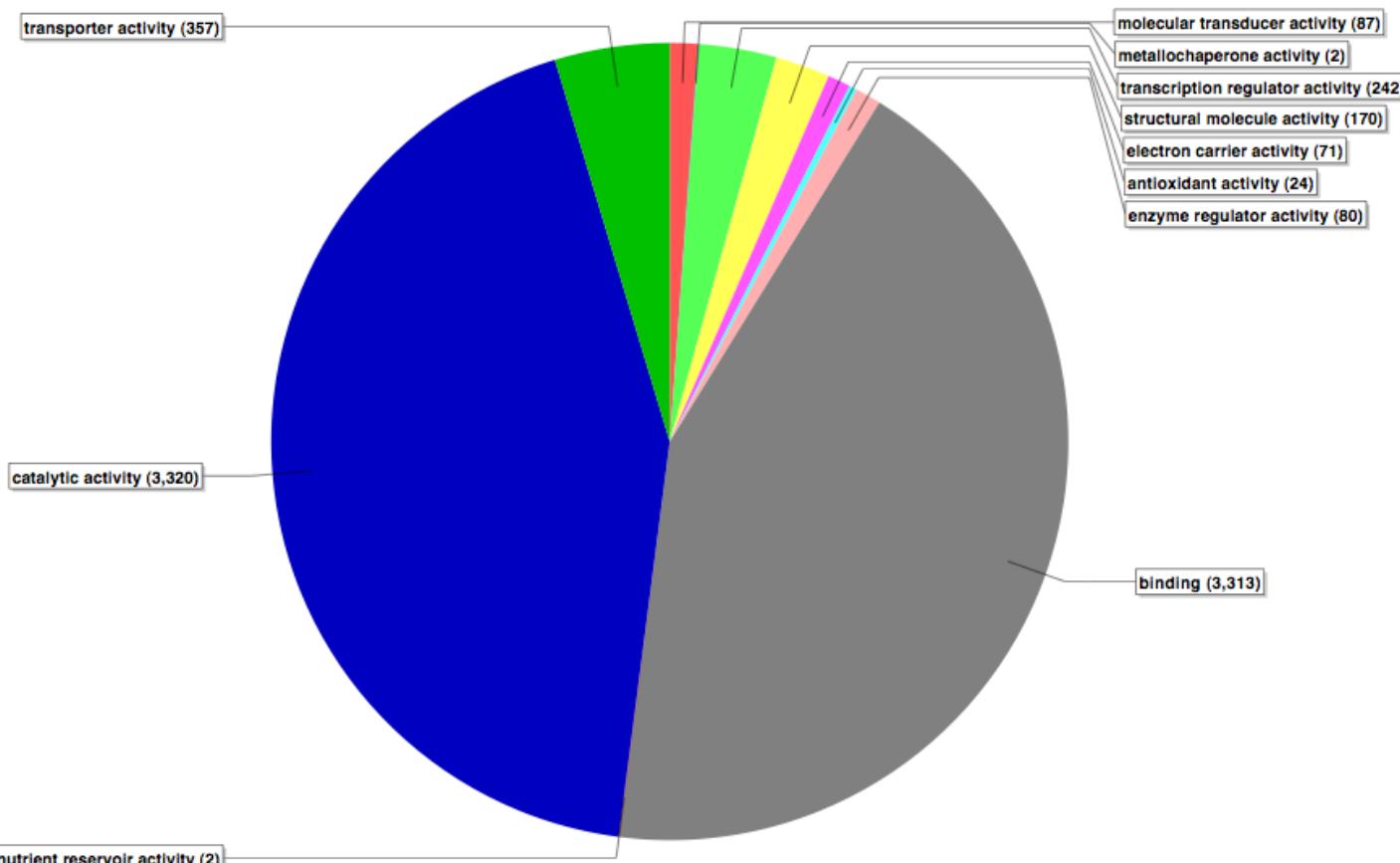
Annotation already running



## 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

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 F:ATP binding; F:succinate-CoA ligase (GDP-forming) activity: D: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
- Benzene degradation via CoA ligation
- Valine, leucine and isoleucine biosynthesis
- Reductive carboxylate cycle (CO<sub>2</sub> 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
- Cysteine and methionine metabolism
- Terpenoid backbone biosynthesis
- Histidine metabolism
- T cell receptor signaling pathway
- Tropamine, piperidine and pyridine alkaloid biosynthesis
- One carbon pool by folate
- Pentose and glucuronate interconversions
- Phosphatidylinositol signaling system

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](mailto:support@bils.se)

# Biosupport.se

Swedish Bioinformatics Support  
www.biosupport.se

Reader login about faq

SBS Swedish Bioinformatics Support Questions Tags Users Badges Unanswered Ask a Question

All Questions [RSS](#)

active newest hottest most voted

oligo file for mothur  
0 votes 1 answer 90 views 454amplicon oligosfile mothur 03 Jun, 15:13 dahlö ♦♦ 252

Tophat results more than one accepted\_hits file  
0 votes 1 answer 518 views cufflinks tophat 27 May, 11:47 eeeee(suspended)

Cloud services for bioinformatics?  
0 votes 2 answers 85 views cloud-computing service 27 May, 11:22 zashah 1

Memory overflow error performing (/mpileup/) in Uppmax  
0 votes 2 answers 108 views mpileup uppmax 22 May, 15:18 Mikael Borg ♦♦ 96

detect small RNA from normal RNA-seq data?  
0 votes 3 answers 132 views rna-seq 08 May, 14:12 walker 10

Common resources for bowtie / tophat for hg19  
0 votes 1 answer 175 views uppmax rna-seq annotation transcriptome 06 May, 10:24 S Dilorenzo ♦♦ 71

AddOrReplaceReadGroups picard problem: Caused by:  
java.io.IOException: No space left on device  
1 vote 1 answer 197 views picard 29 Apr, 12:11 S Dilorenzo ♦♦ 71

List of NGS services in Sweden?  
1 2 975

63 Questions  
126 answers  
questions

Welcome  
to the bioinformatics support forum!  
Feel free to ask any questions  
regarding bioinformatics.  
Please refer to the [FAQ](#) or the [about section](#) if you want more information  
about this community.

Privacy  
If you would feel better keeping your  
correspondence private, or if your  
question is too project specific,  
please send the question as a mail to  
[support@biosupport.se](mailto:support@biosupport.se) instead.

Recent tags  
cloud-computing mothur oligosfile  
454amplicon mpileup haplotype  
snps muscle radseq python