

Genome annotation: Independent study

Saranya Sankaranarayanan 

Advisor : Dr. R. Taylor Raborn 
Dr. Volker Brendel

Learning Goals



- What is genome annotation ?
- How to annotate a genome?
- How to compare annotations ?

Tools used



Maker 2.31.6 : Holt and Yandell (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinfo 12:491

Genometools: G. Gremme, S. Steinbiss and S. Kurtz. *GenomeTools*: a comprehensive software library for efficient processing of structured genome annotations. [IEEE/ACM Transactions on Computational Biology and Bioinformatics 2013, 10\(3\):645–656](#)

Parseval : Daniel S Standage, Volker P Brendel :ParseEval: parallel comparison and analysis of gene structure annotations. *BMC Bioinformatics* 2012, 13:187 doi:10.1186/1471-2105-13-187

Bedtools : Aaron R Quinlan,Ira M.Hall: BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* (2010) 26 (6): 841-842.doi: 10.1093/bioinformatics/btq033

xGDBVm:

CEGMA : Genis Parra, Keith Bradnam and Ian Korf. [CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes."](#) *Bioinformatics*, 23: 1061-1067 (2007)

Github

CEGMA : Genis Parra, Keith Bradnam and Ian Korf. [CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes."](#) *Bioinformatics*, 23: 1061-1067 (2007)

Introduction :





Need for genome annotation

- Genomes sequencing rate much higher than annotation
- Eukaryotic genome annotation much more complex
- Identification of genes and other features
- Many tools, many pipelines




Gene Prediction

Ab initio



- De novo
- Genome searched for protein coding genes
- Prokaryotic gene prediction simple
-  Uses HMM
-  SNAP

Empirical

-  Homology (Evidence based)
- Search genome for EST, protein alignment
- Dependent of assembly quality
- Matches: Complete or partial
- Genome Threader

Daphnia pulex



-  Crustacean found in ponds and lakes
- Standard organism for toxicity testing
- switch between clonal and sexual reproduction in response to environmental conditions
-  191 scaffolds
- Genome size ~ 200 Mb
- 30,907 protein-encoding genes







Method:

Maker



- easy-to-use genome annotation pipeline
- *de novo* annotation of newly sequenced genomes
- updating existing annotations to reflect new evidence
- combine annotations, evidence, and quality control statistics for use with other GMOD programs like Gbrowse

Maker

- Input :

Genome sequence

Transcript sequence evidence

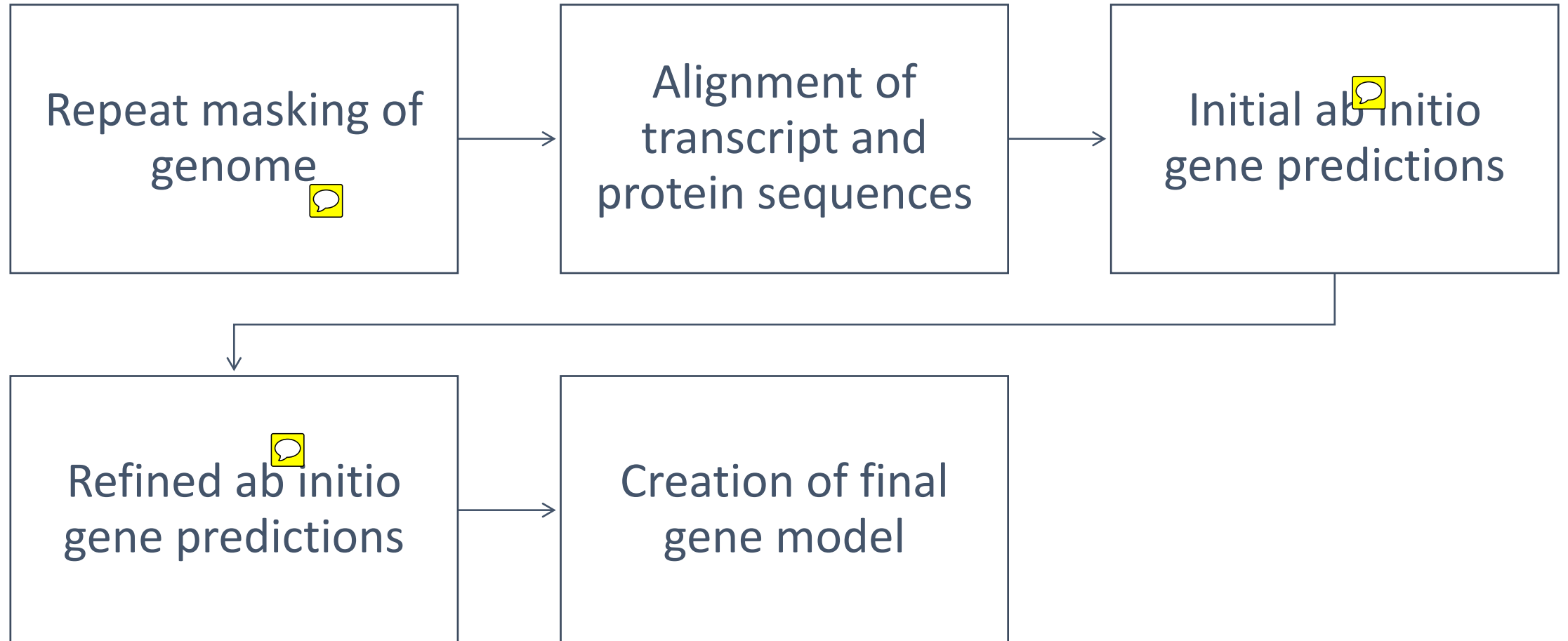
Protein sequence evidence

- Input files can be provided in several formats: FASTA, GFF
- Uses HMMs for gene prediction

Control files

maker_exectl	paths to required executables
maker_boptctl	options for BLAST and Exonerate
maker_optctl	Contains paths to input files and other options

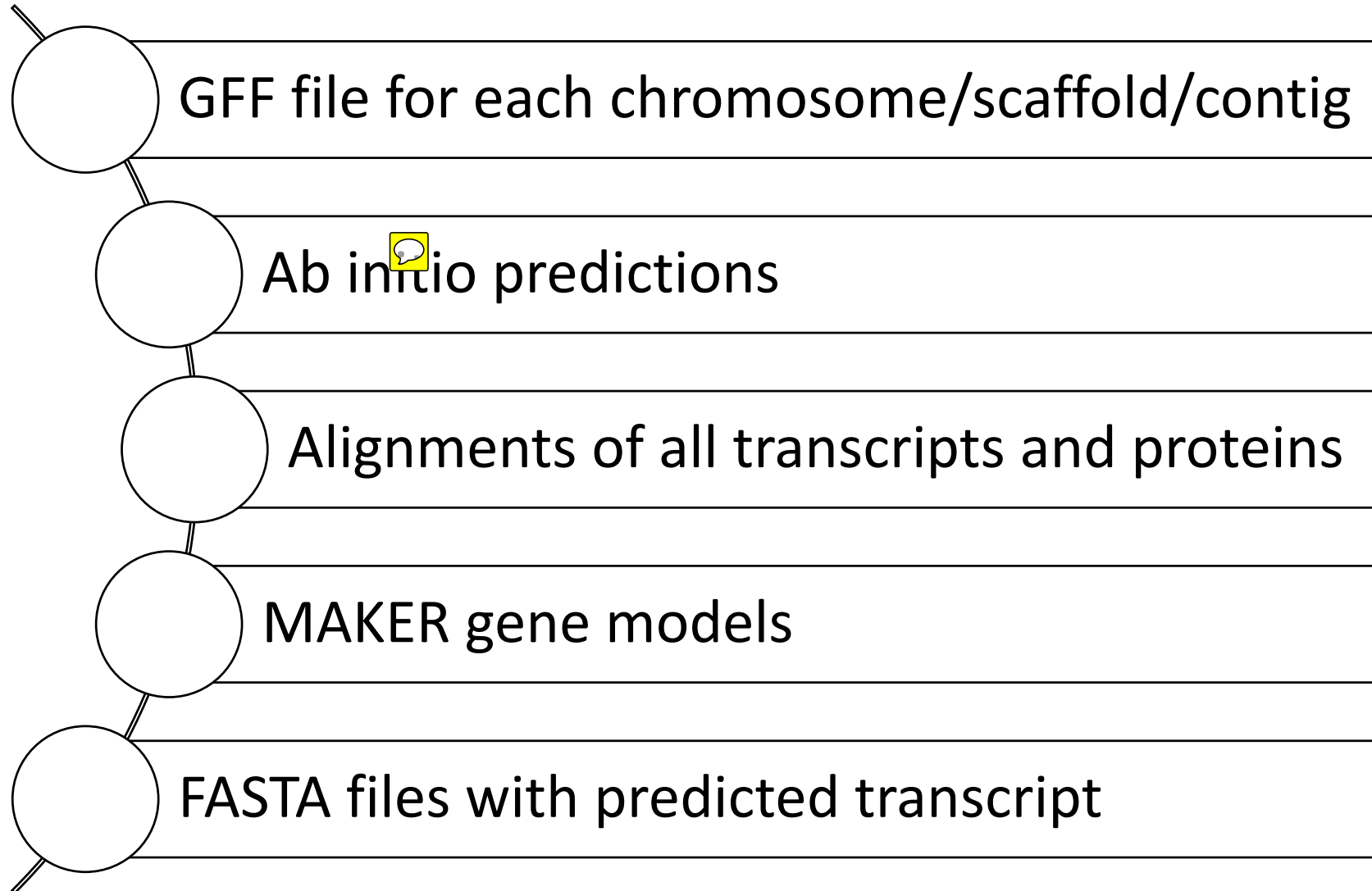
Maker pipeline



Maker^{🗨️} derived gene models

- Annotations with evidence is reported.
- BLAST alignment must have at least a corresponding gene predictor result
- Gene predictor result without evidence not reported
- AED score :amount of dissimilarity between evidence and annotation

Result from Maker



Merging the files:



Gather the predicted proteins

```
cat ./**/*/*.proteins.fasta > genome.predictedproteins.fasta
```

Gather the predicted transcripts

```
cat ./**/*/*.transcripts.fasta > genome.predictedtranscripts.fasta
```

Gather the gff

```
gff3_merge -d genome_master_datastore_index.log
```

Viewing the results:

"DaphniaGDB"

[Help with this page ?](#)

[Track Color Codes ?](#)

[Link to this Page](#)

Genome region: ?

Zoom

Annotate

BLAST

Download

Format

Add Track

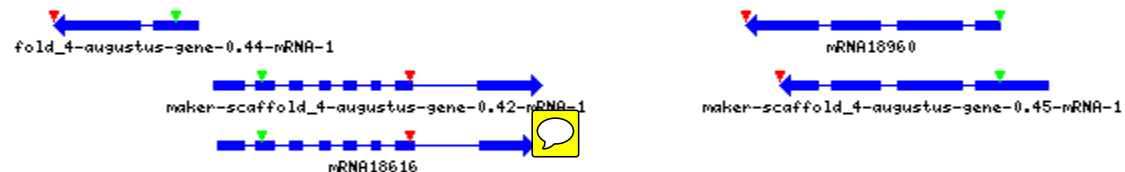
Nucleotide Level

1 501 1001 1501 2001 2501 3001 3501 4001 4501 5001 5501 6001 6501 7001 7501 8001 8501 9001 9501

▼ GDB001 - yrGATE Annotations

▼ GDB001 - Gene Models (from CpGAT)

▼ GDB001 - Gene Models (from gff3)



▼ GDB001 - Aligned Protein

▼ GDB001 - Aligned cDNA

▼ GDB001 - Aligned EST





Results:

Evaluation of annotation

Gtstats

Parseval 


Single exons BLAST

Comparison with other metazoan annotations

CEGMA



Basic statistics:

Number of 	Frozen Gene Dataset	Maker annotation
Genes	30776	11285
Exons	145385	74868
CDS	143218	72731
Introns	114575	63630
3' UTR	7961	7836
5' UTR	7219	7026



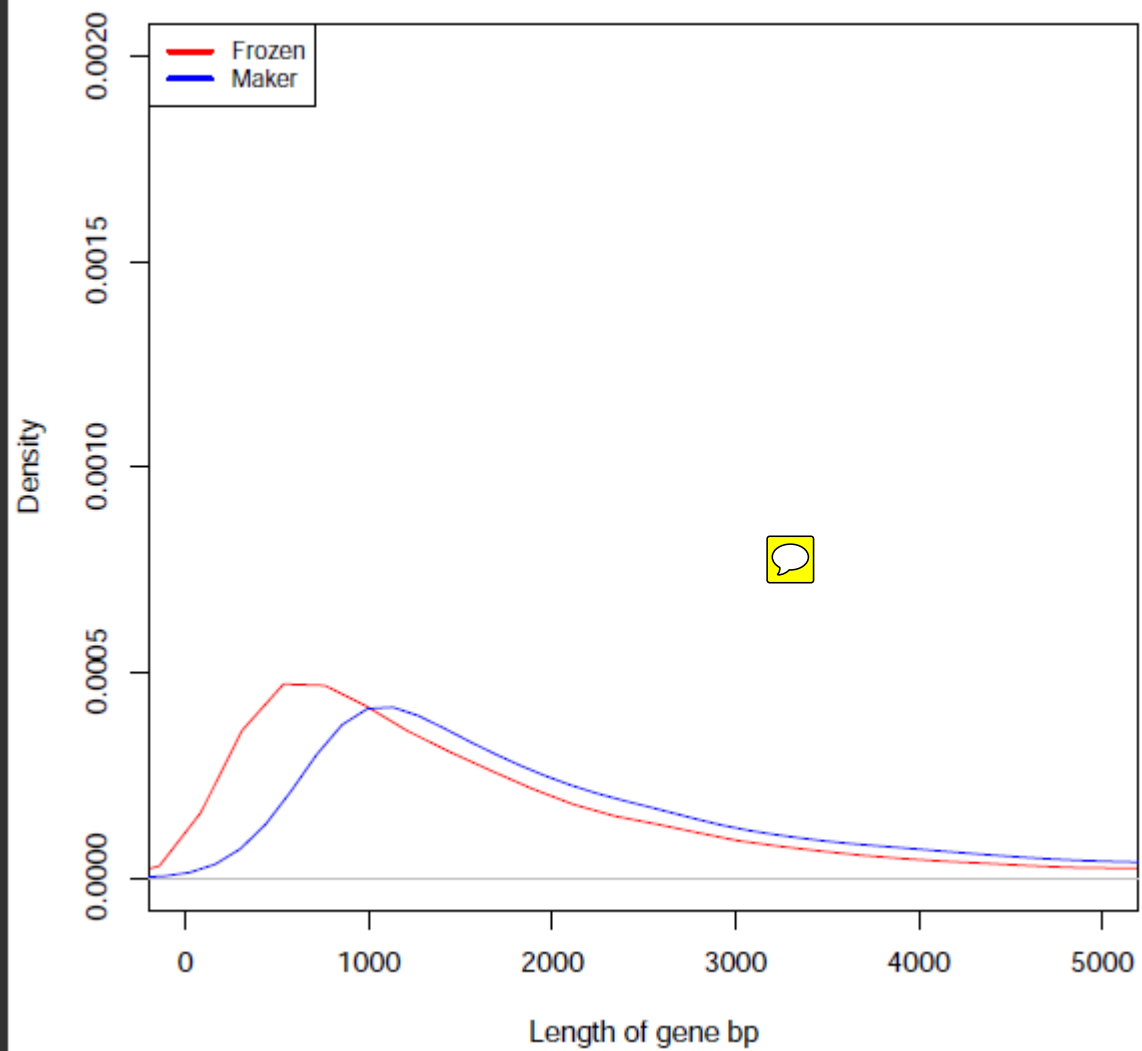
Parseval comparison

Gene Loci	
Unique to FrozenGene	Unique to Maker
17932	295

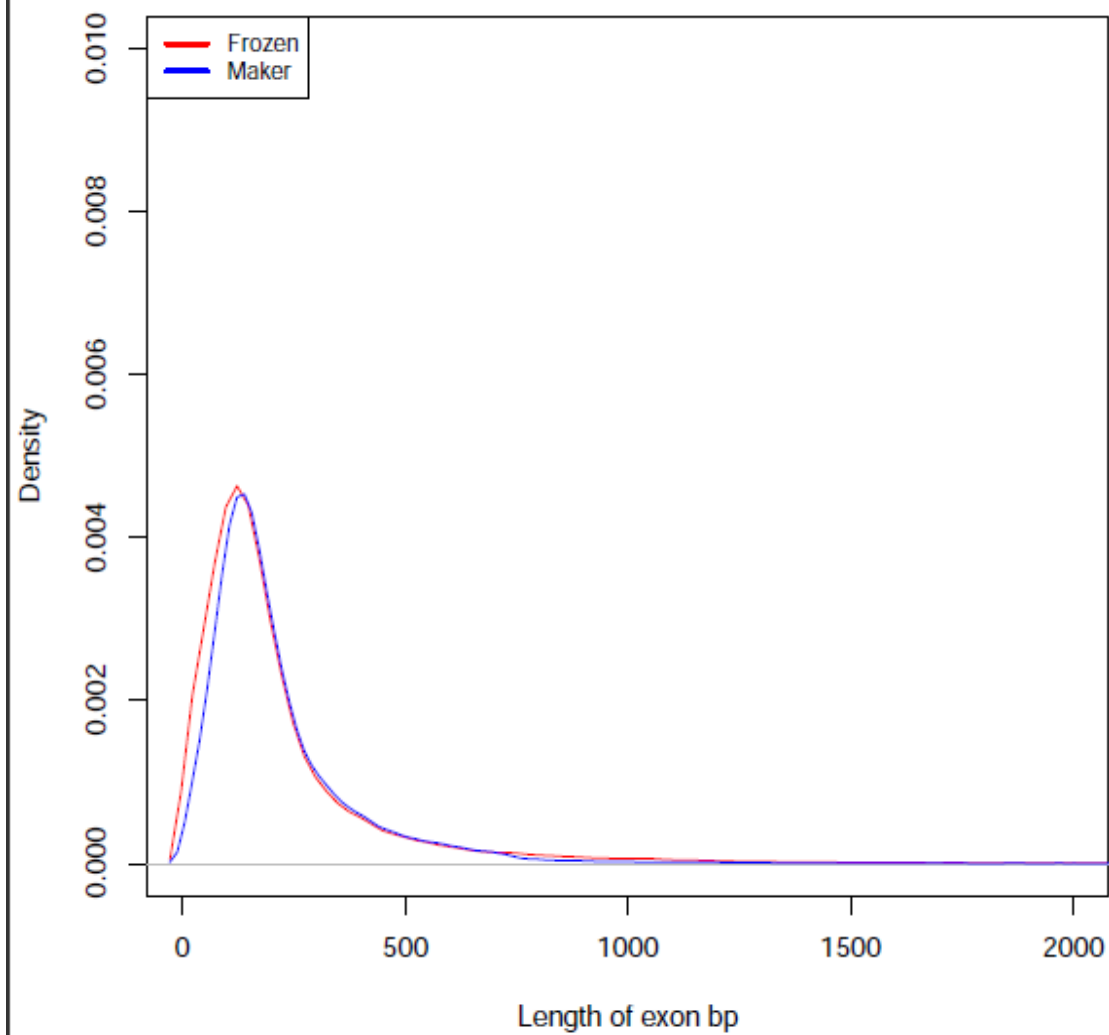
	FrozenGene	Maker
Number of genes	30776	11285
Average per locus	1.11	0.4




Frozen Vs Maker Gene length distribution



Frozen Vs Maker exon length distribution



No of exons/gene

Number of exon/gene	Frozen Gene Dataset	Maker annotation
One Exon	171	5090 
Two Exons	1055	4769
Three Exons	1584	4769
Four Exons	1604	3739
Five +	6871	11817

Number of exons/gene



Annotation

D.melanogaster

C.elegans

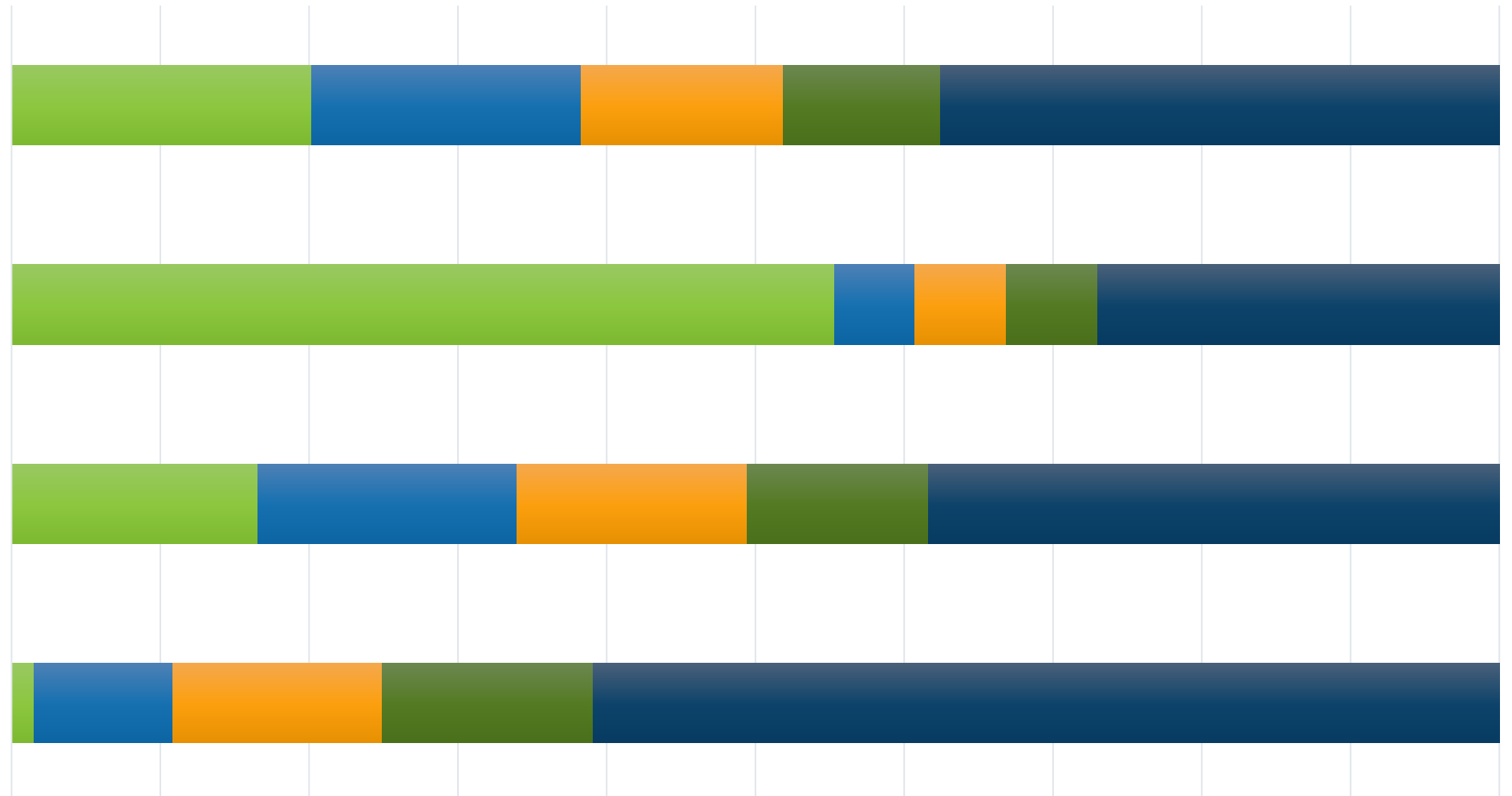
FrozenGeneDataset

Maker derived

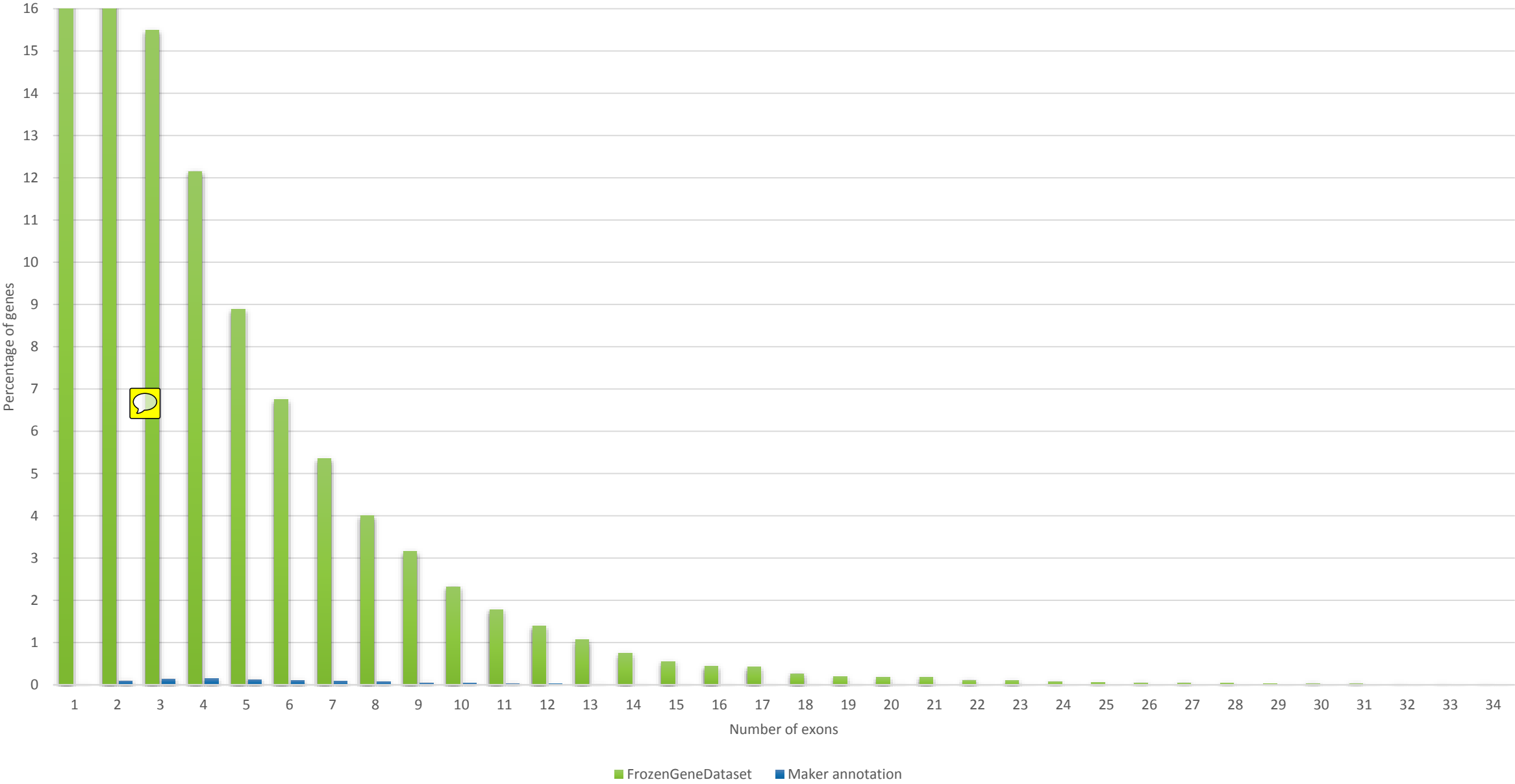
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Number of genes

1 2 3 4 5+



Number of exons/gene



Get single exonic genes from the
gff3 file : R package
GenomicFeatures (exons by 'gene')



Bedtools getfasta : fasta sequence



BLASTx



Check for the single exonic gene in
Maker gff3 : Bedtools intersect



Example:

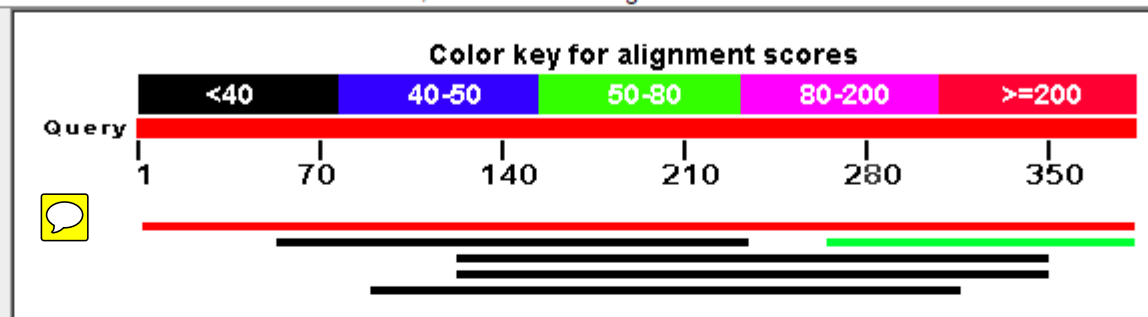


- Gene9711
 - Single exonic gene
 - On scaffold_184 from 186739 to 187119
 - SNAP derived gene
-
- One of the many single exonic genes with no evidence.

No putative conserved domains have been detected

Distribution of 6 Blast Hits on the Query Sequence



Mouse-over to show define and scores, click to show alignments



Descriptions





Sequences producing significant alignments:

Select: [All](#) [None](#) Selected: 0

 [Alignments](#)  [Download](#)  [GenPept](#) [Graphics](#) 

	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	hypothetical protein DAPPUDRAFT_263274 [Daphnia pulex]	229	229	99%	2e-74	100%	gi 321455240 EFX66378.1
<input type="checkbox"/>	hypothetical protein DAPPUDRAFT_251165 [Daphnia pulex]	57.0	57.0	30%	8e-08	64%	gi 321463868 EFX74880.1
<input type="checkbox"/>	protein of hypothetical function UPF0052 and CofD [Weissella koreensis]	37.4	37.4	59%	1.7	31%	gi 493899121 WP_006844948.1
<input type="checkbox"/>	hypothetical protein [Weissella koreensis]	36.2	36.2	59%	4.7	29%	gi 503755542 WP_013989618.1

CEGMA

- CEGMA  output.cegma.gff
- The assembly  was 96.37 % complete
- Bedtools to  convert gff3 to bed format
- KOG1078.2 issing from Frozen gene dataset
- KOG1078, Vesicle coat complex COPI, gamma subunit [Intracellular trafficking, secretion, and vesicular transport]



Questions ?

