Overview, Installation, and Basic Configuration for Annotating Genomic Sequence

MAKER I

Carson Holt
University of Utah

Introduction to Genome Annotation

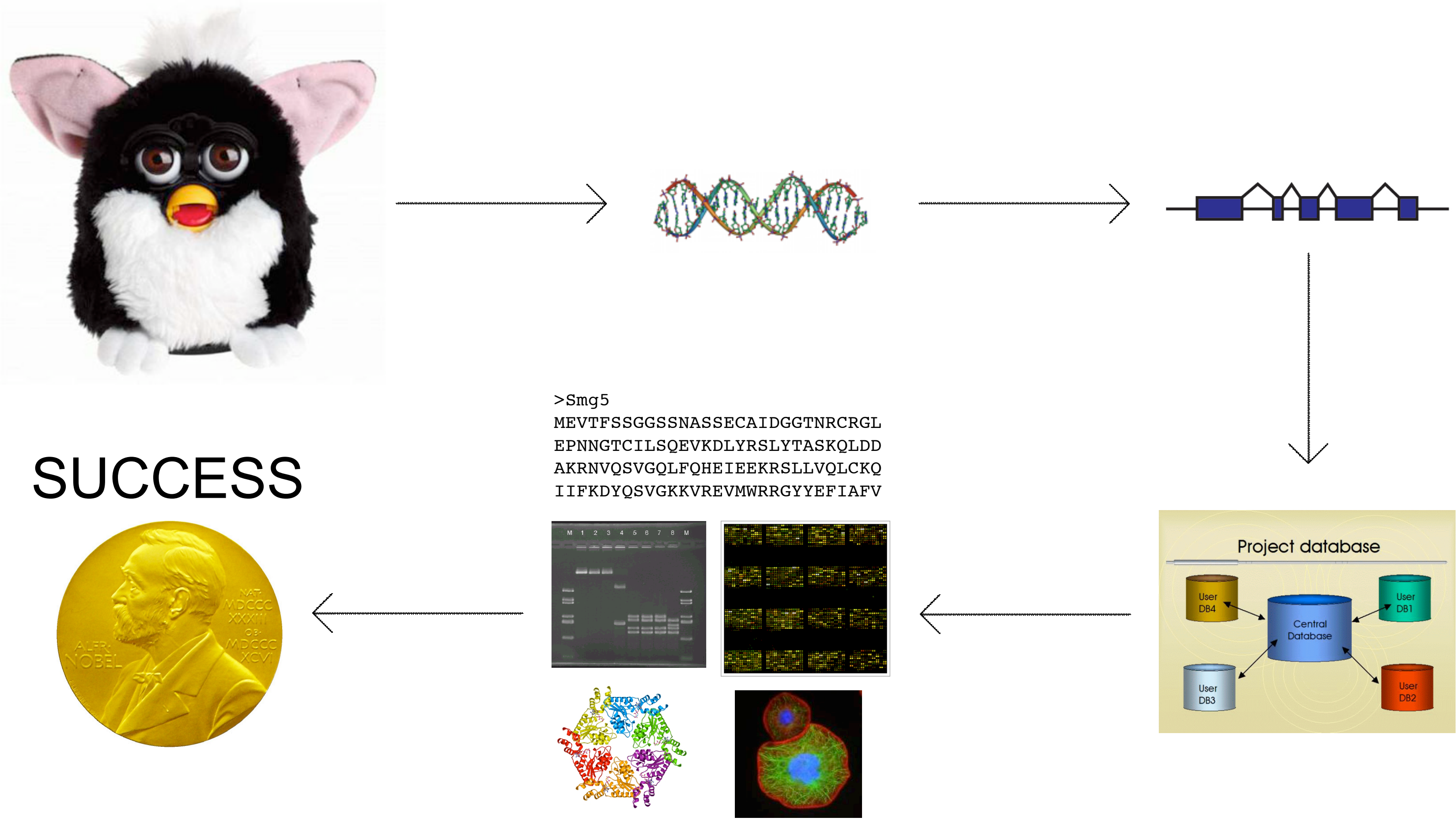
- What annotations are
- Importance of genome annotations
- Effect of next generation sequencing technologies on the annotation process

What Are Annotations?

- Annotations are descriptions of features of the genome
  - Structural: exons, introns, UTRs, splice forms etc.
  - Functional: metabolism, hydrolase, expressed in the mitochondria, etc.

- Annotations should include evidence trail
  - Assists in quality control of genome annotations
- Examples of evidence supporting a structural annotation:
  - *Ab initio* gene predictions
  - ESTs
  - Protein homology

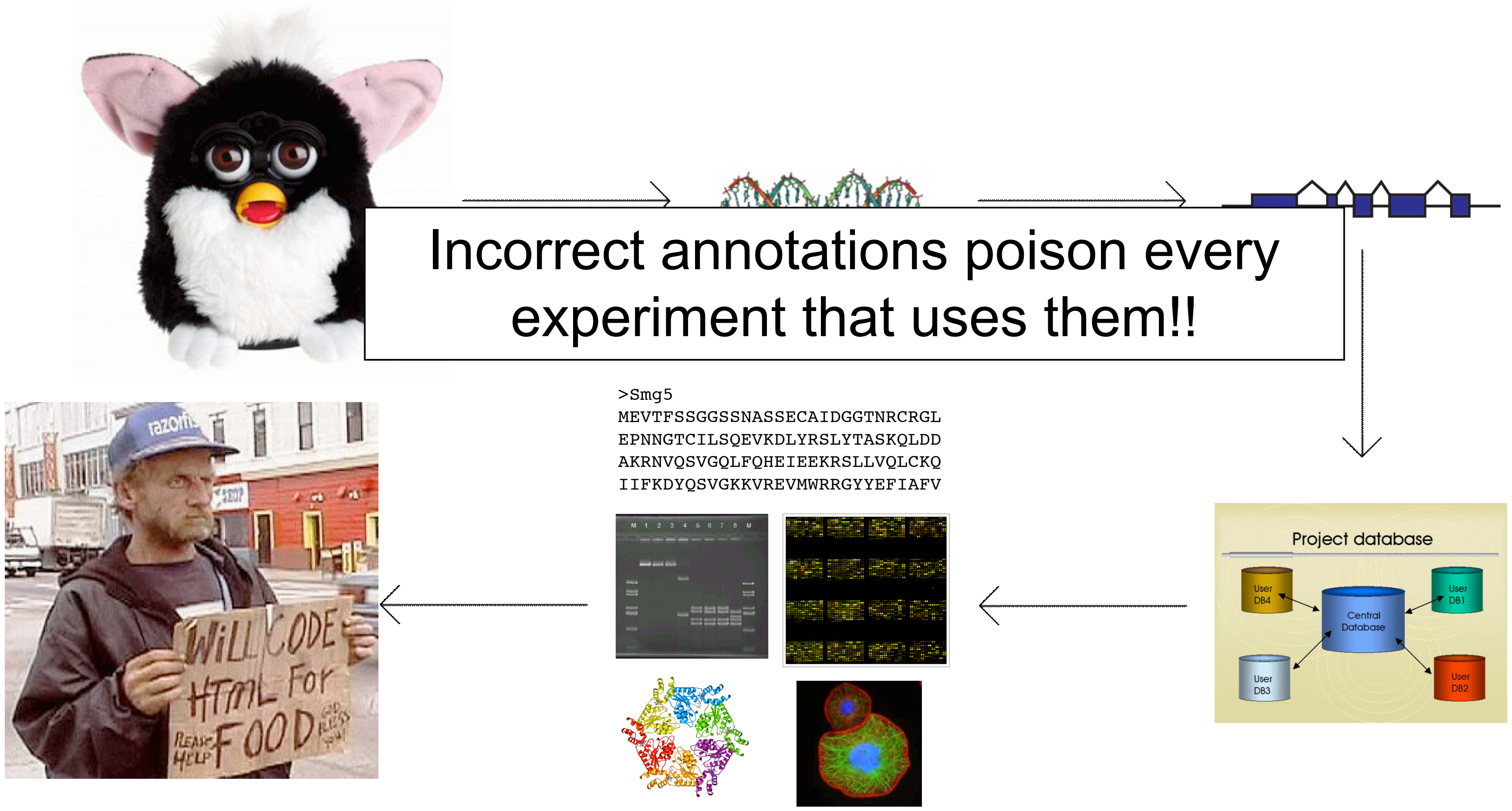
Why should I care about genome annotations?



SUCCESS

```
>Smg5
MEVTFSSGGSSNASSECAIDGGTNRCRGL
EPNNGTCILSQEVKDLYRSLYTASKQLDD
AKRNVQSVGQLFQHEIEEKRSLLVQLCKQ
IIFKDYQSVGKKVREVMWRRGYYEFIAFV
```

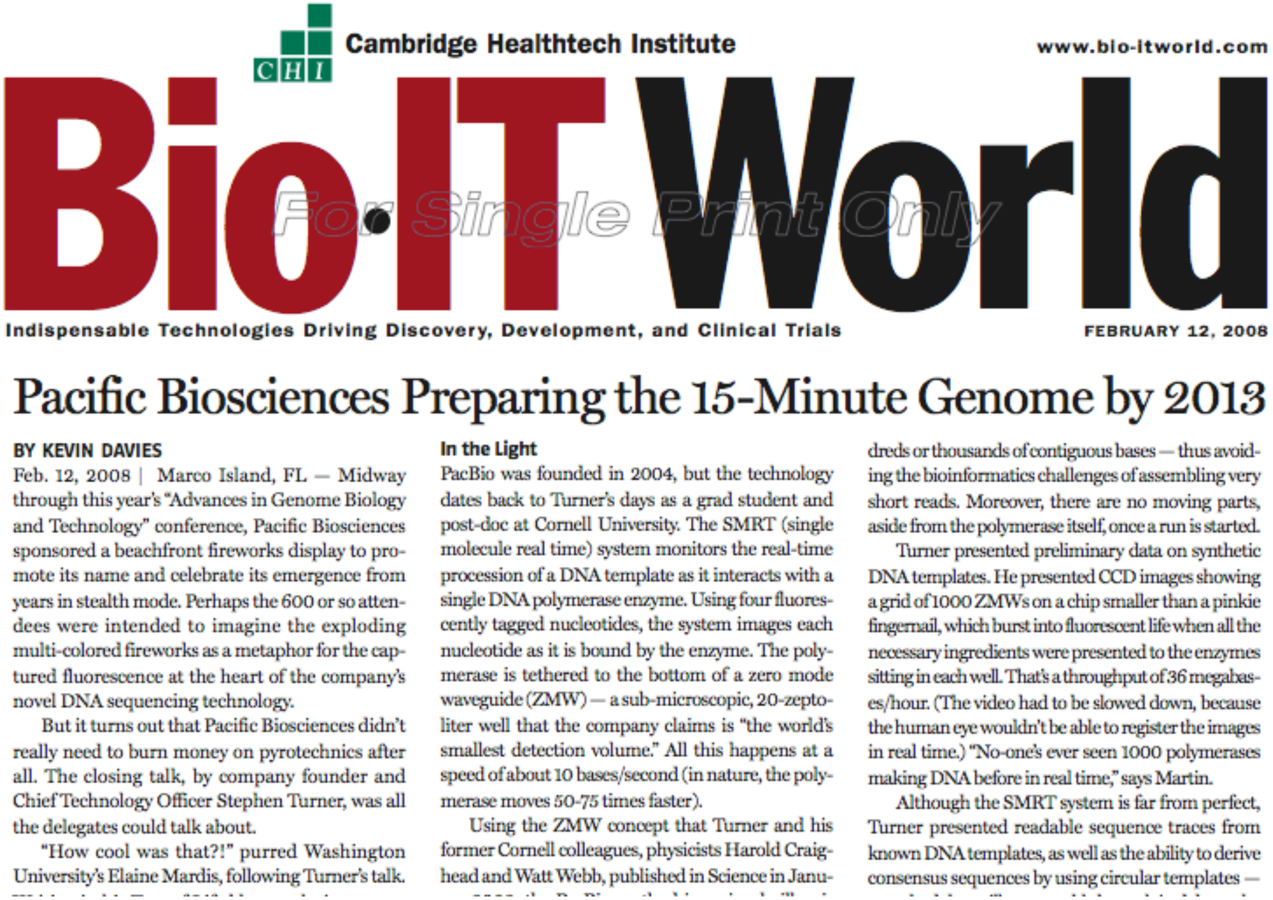
Why should I care about genome annotations?



Incorrect annotations poison every experiment that uses them!!

```
>Smg5
MEVTFSSGGSSNASSECAIDGGTNRCRGL
EPNNGTCILSQEVKDLYRSLYTASKQLDD
AKRNVQSVGQLFQHEIEEKRSLLVQLCKQ
IIFKDYQSVGKKVREVMWRRGYYEFIAFV
```

Advances in Technology Promise to Make Whole Genome Sequencing “Routine” for Even Small Labs



CHI Cambridge Healthtech Institute

www.bio-itworld.com

Bio·IT World

Indispensable Technologies Driving Discovery, Development, and Clinical Trials

FEBRUARY 12, 2008

Pacific Biosciences Preparing the 15-Minute Genome by 2013

BY KEVIN DAVIES

Feb. 12, 2008 | Marco Island, FL — Midway through this year's "Advances in Genome Biology and Technology" conference, Pacific Biosciences sponsored a beachfront fireworks display to promote its name and celebrate its emergence from years in stealth mode. Perhaps the 600 or so attendees were intended to imagine the exploding multi-colored fireworks as a metaphor for the captured fluorescence at the heart of the company's novel DNA sequencing technology.

But it turns out that Pacific Biosciences didn't really need to burn money on pyrotechnics after all. The closing talk, by company founder and Chief Technology Officer Stephen Turner, was all the delegates could talk about.

"How cool was that?!" purred Washington University's Elaine Mardis, following Turner's talk.

In the Light

PacBio was founded in 2004, but the technology dates back to Turner's days as a grad student and post-doc at Cornell University. The SMRT (single molecule real time) system monitors the real-time procession of a DNA template as it interacts with a single DNA polymerase enzyme. Using four fluorescently tagged nucleotides, the system images each nucleotide as it is bound by the enzyme. The polymerase is tethered to the bottom of a zero mode waveguide (ZMW) — a sub-microscopic, 20-zeptoliter well that the company claims is "the world's smallest detection volume." All this happens at a speed of about 10 bases/second (in nature, the polymerase moves 50-75 times faster).

Using the ZMW concept that Turner and his former Cornell colleagues, physicists Harold Craighead and Watt Webb, published in Science in Janu-

dreds or thousands of contiguous bases — thus avoiding the bioinformatics challenges of assembling very short reads. Moreover, there are no moving parts, aside from the polymerase itself, once a run is started.

Turner presented preliminary data on synthetic DNA templates. He presented CCD images showing a grid of 1000 ZMWs on a chip smaller than a pinkie fingernail, which burst into fluorescent life when all the necessary ingredients were presented to the enzymes sitting in each well. That's a throughput of 36 megabases/hour. (The video had to be slowed down, because the human eye wouldn't be able to register the images in real time.) "No-one's ever seen 1000 polymerases making DNA before in real time," says Martin.

Although the SMRT system is far from perfect, Turner presented readable sequence traces from known DNA templates, as well as the ability to derive consensus sequences by using circular templates —

Advances in annotation technology have not kept pace with genome sequencing, and annotation is rapidly becoming a major bottleneck affecting modern genomics research.

- As of February 2009, 173 eukaryotic genomes were fully sequenced yet unpublished.
- Currently there are over 1,000 eukaryotic genome projects underway, assuming 10,000 genes per genome, that's 10,000,000 new annotations.
- While there are organizations dedicated to producing and distributing genome annotations (i.e ENSEMBL and VectorBase), the shear volume of newly sequenced genomes exceeds both their capacity and stated purview.
- So unfortunately many small research groups (which often lack bioinformatics experience) must confront the difficulties associated with genome annotation on their own.

- .GOLD: Genomes OnLine Database. **2008**.

- MAKER is an easy-to-use annotation pipeline designed to help smaller research groups convert the mountain of genomic data provided by next generation sequencing technologies into a usable resource.

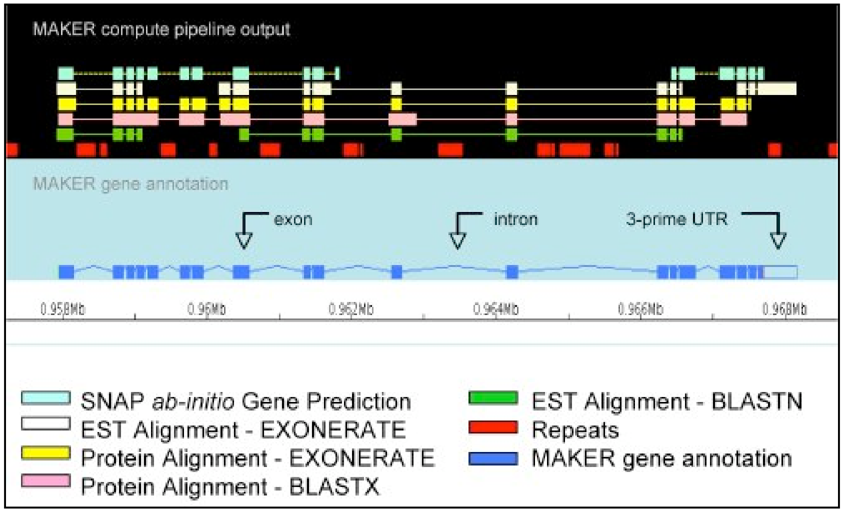
MAKER Overview

- What does MAKER do?
- What sets MAKER apart from other tools (ab initio gene predictors, etc.)?
- Emerging vs. model genomes
- Comparison of algorithm performance on model vs. emerging genomes

MAKER

- The easy-to-use annotation pipeline.

User Requirements:	Can be run by a single individual with little bioinformatics experience
System Requirements:	Can run on laptop or desktop computers running Linux or Mac OS X
Program Output:	Output is compatible with popular GMOD annotation tools like Apollo and GBrowse
Availability:	Free open source application (for academic use)


MAKER compute pipeline output
MAKER gene annotation
exon
intron
3-prime UTR
0.958Mb
0.96Mb
0.962Mb
0.964Mb
0.966Mb
0.968Mb
SNAP *ab-initio* Gene Prediction
EST Alignment - EXONERATE
Protein Alignment - EXONERATE
Protein Alignment - BLASTX
EST Alignment - BLASTN
Repeats
MAKER gene annotation

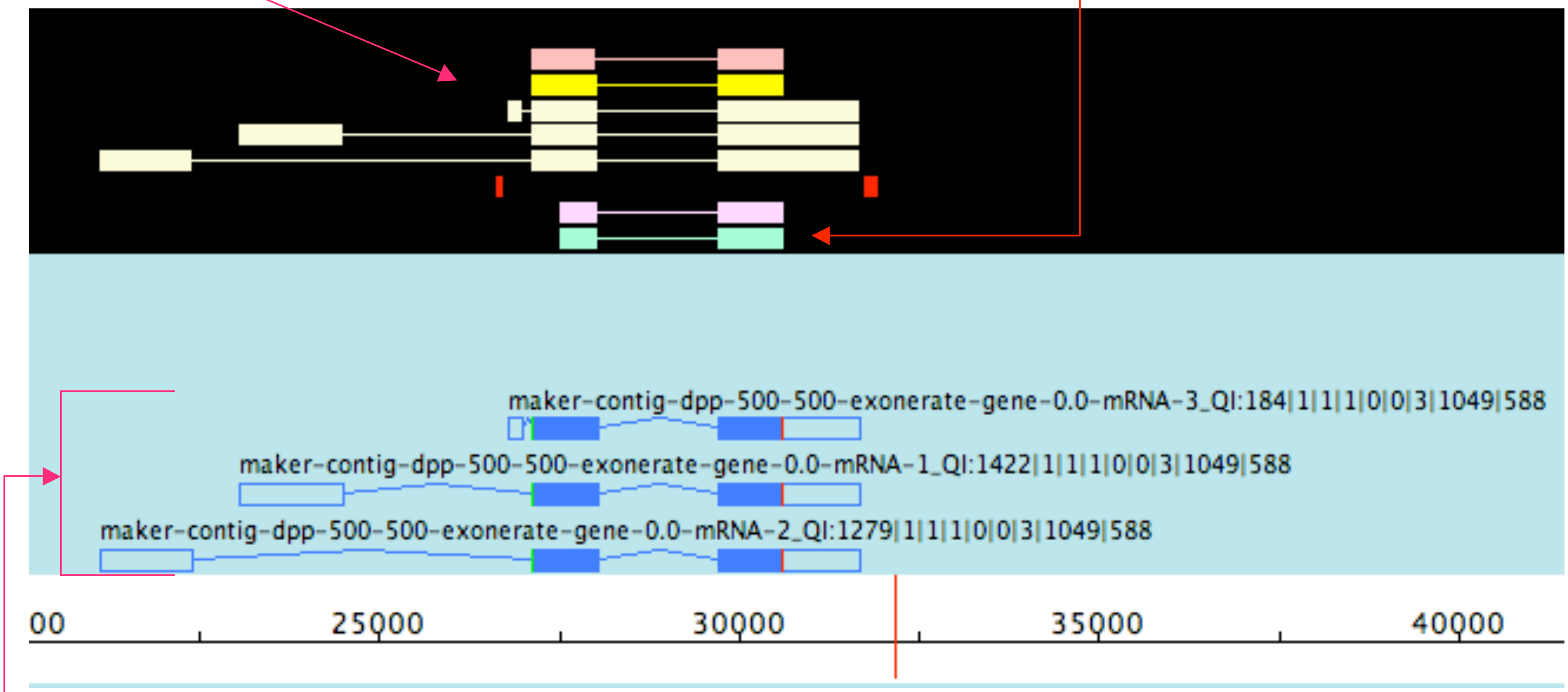
MAKER identifies repeats, aligns ESTs and proteins to a genome, produces *ab-initio* gene predictions, automatically synthesizes these data into gene annotations, and produces evidence-based quality values for downstream annotation management

- Lewis, S.E. et al. Apollo: a sequence annotation editor. *Genome Biology* **3**, research0082.1 - 0082.14 (2002).
- Stein, L.D. et al. The Generic Genome Browser: A Building Block for a Model Organism System Database. *Genome Res.* **12**, 1599-1610 (2002).

What sets MAKER apart from other tool (i.e. ab initio gene predictors)?

Computational evidence

Gene-predictions


maker-contig-dpp-500-500-exonerate-gene-0.0-mRNA-3_QI:184|1|1|1|0|0|3|1049|588
maker-contig-dpp-500-500-exonerate-gene-0.0-mRNA-1_QI:1422|1|1|1|0|0|3|1049|588
maker-contig-dpp-500-500-exonerate-gene-0.0-mRNA-2_QI:1279|1|1|1|0|0|3|1049|588
00
25000
30000
35000
40000

Gene annotation

gene prediction ≠ gene annotation

Model *versus* Emerging genomes

Model genomes:

- Classic experimental systems
- Much prior knowledge about genome
- Large community
- Big $

Examples: *D. melanogaster*, *C. elegans*, human, etc

Model *versus* Emerging genomes

Emerging genomes:

- New experimental systems
  - Genome will be the central resource for work in these systems
- Little prior knowledge about genome
  - Usually no genetics
- Small communities
- Less $

Examples: flatworms, oomycetes, the cone snail, etc.

Comparison of gene models produced by state-of-the art algorithms against a REFERENCE genome

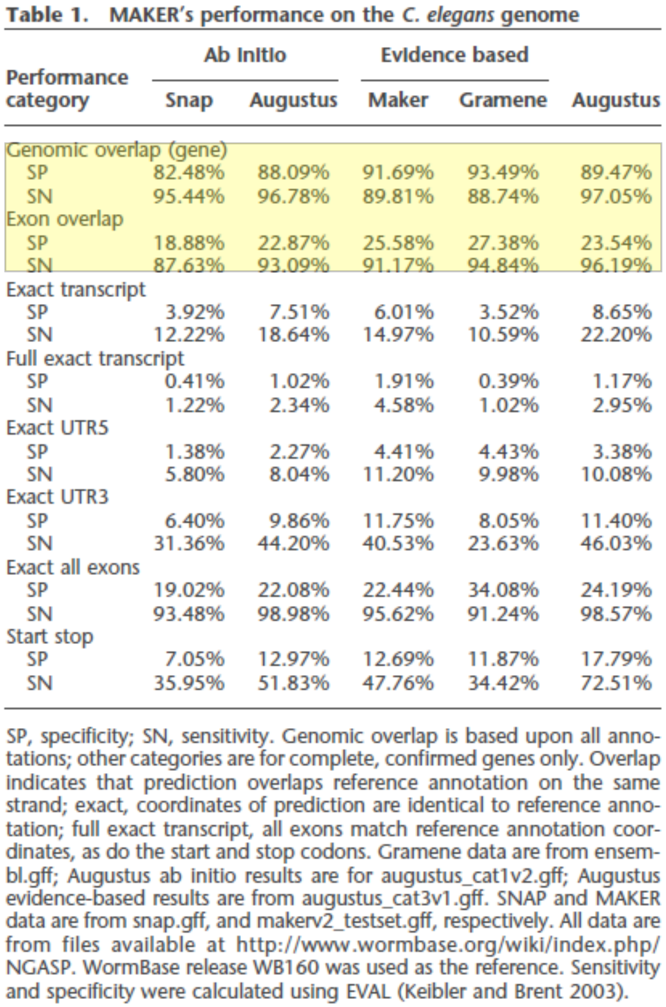


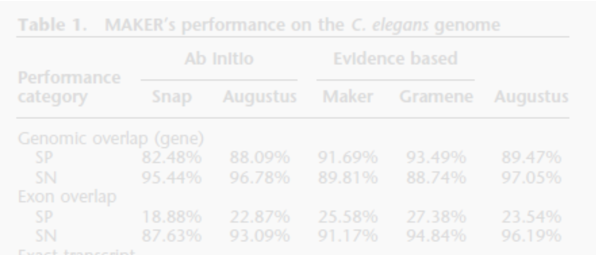
Table 1. MAKER's performance on the *C. elegans* genome

Performance category	Ab Initio		Evidence based		
	Snap	Augustus	Maker	Gramene	Augustus
Genomic overlap (gene)					
SP	82.48%	88.09%	91.69%	93.49%	89.47%
SN	95.44%	96.78%	89.81%	88.74%	97.05%
Exon overlap					
SP	18.88%	22.87%	25.58%	27.38%	23.54%
SN	87.63%	93.09%	91.17%	94.84%	96.19%
Exact transcript					
SP	3.92%	7.51%	6.01%	3.52%	8.65%
SN	12.22%	18.64%	14.97%	10.59%	22.20%
Full exact transcript					
SP	0.41%	1.02%	1.91%	0.39%	1.17%
SN	1.22%	2.34%	4.58%	1.02%	2.95%
Exact UTR5					
SP	1.38%	2.27%	4.41%	4.43%	3.38%
SN	5.80%	8.04%	11.20%	9.98%	10.08%
Exact UTR3					
SP	6.40%	9.86%	11.75%	8.05%	11.40%
SN	31.36%	44.20%	40.53%	23.63%	46.03%
Exact all exons					
SP	19.02%	22.08%	22.44%	34.08%	24.19%
SN	93.48%	98.98%	95.62%	91.24%	98.57%
Start stop					
SP	7.05%	12.97%	12.69%	11.87%	17.79%
SN	35.95%	51.83%	47.76%	34.42%	72.51%

SP, specificity; SN, sensitivity. Genomic overlap is based upon all annotations; other categories are for complete, confirmed genes only. Overlap indicates that prediction overlaps reference annotation on the same strand; exact, coordinates of prediction are identical to reference annotation; full exact transcript, all exons match reference annotation coordinates, as do the start and stop codons. Gramene data are from ensembl.gff; Augustus ab initio results are for augustus_cat1v2.gff; Augustus evidence-based results are from augustus_cat3v1.gff. SNAP and MAKER data are from snap.gff, and makerv2_testset.gff, respectively. All data are from files available at http://www.wormbase.org/wiki/index.php/NGASP. WormBase release WB160 was used as the reference. Sensitivity and specificity were calculated using EVAL (Keibler and Brent 2003).

MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes.
(2008)Cantarel B L, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M
Genome Res 18(1) 188-196

Comparison of gene models produced by state-of-the art algorithms against a REFERENCE genome



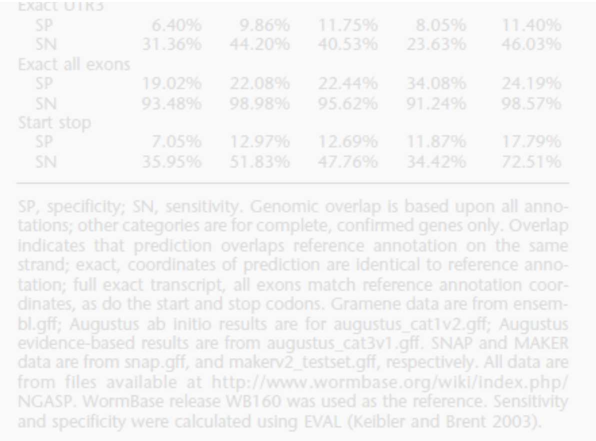


Table 1. MAKER's performance on the *C. elegans* genome

Performance category	Ab Initio		Evidence based		
	Snap	Augustus	Maker	Gramene	Augustus
Genomic overlap (gene)					
SP	82.48%	88.09%	91.69%	93.49%	89.47%
SN	95.44%	96.78%	89.81%	88.74%	97.05%
Exon overlap					
SP	18.88%	22.87%	25.58%	27.38%	23.54%
SN	87.63%	93.09%	91.17%	94.84%	96.19%
Exact transcript					
Exact UTR3					
SP	6.40%	9.86%	11.75%	8.05%	11.40%
SN	31.36%	44.20%	40.53%	23.63%	46.03%
Exact all exons					
SP	19.02%	22.08%	22.44%	34.08%	24.19%
SN	93.48%	98.98%	95.62%	91.24%	98.57%
Start stop					
SP	7.05%	12.97%	12.69%	11.87%	17.79%
SN	35.95%	51.83%	47.76%	34.42%	72.51%

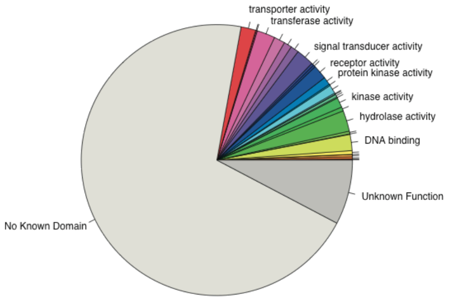
SP, specificity; SN, sensitivity. Genomic overlap is based upon all annotations; other categories are for complete, confirmed genes only. Overlap indicates that prediction overlaps reference annotation on the same strand; exact, coordinates of prediction are identical to reference annotation; full exact transcript, all exons match reference annotation coordinates, as do the start and stop codons. Gramene data are from ensembl.gff; Augustus ab initio results are for augustus_cat1v2.gff; Augustus evidence-based results are from augustus_cat3v1.gff. SNAP and MAKER data are from snap.gff, and makerv2_testset.gff, respectively. All data are from files available at http://www.wormbase.org/wiki/index.php/NGASP. WormBase release WB160 was used as the reference. Sensitivity and specificity were calculated using EVAL (Keibler and Brent 2003).

With *enough* training data, *ab-initio* gene predictors can match or even out-perform annotation pipelines*

nGASP - the nematode genome annotation assessment project *Avril Coghlan , Tristan J Fiedler , Sheldon J McKay ,Paul Flicek , Todd W Harris , Darin Blasiar , The nGASP Consortium and Lincoln D Stein BMC Bioinformatics 2008, 9:549doi:10.1186/1471-2105-9-549*

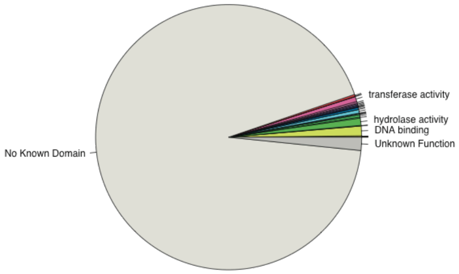
Ab initio gene predictors don’t do nearly so well on emerging genomes*

Average of seven REFERENCE proteomes



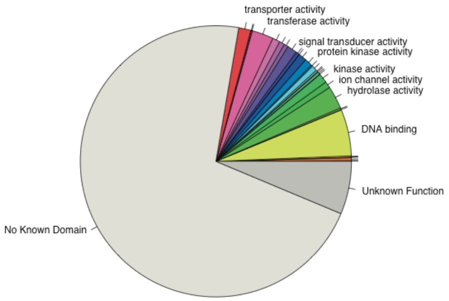
35% contain a domain

***S. mediterranea* SNAP *ab-initio* gene predictions**



7% contain a domain

MAKER *S. mediterranea* annotations



29% contain a domain

MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes.
(2008)Cantarel B L, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M
Genome Res 18(1) 188-196

Benefits of MAKER

- Provides gene models as well as an evidence trail correlations for quality control and manual curation
- Provides a mechanism to train and retrain *ab initio* gene predictors for even better performance.
- Output can be loaded into a GMOD compatible database for annotation distribution
- Annotations can be automatically updated by new evidence by simply passing existing annotation sets back into the pipeline

MAKER Installation

- Prerequisites
- The MAKER package
  - Where to get it
  - Step by step setup

Prerequisites

- Perl 5.8.0 or Higher
- BioPerl 1.6 or higher
- SNAP version 2009-02-03 or higher
- RepeatMasker 3.1.6 or higher
- Exonerate 1.4 or higher

- You must also install one of the following:
  - WU-BLAST 2.0 or higher (Now AB-BLAST)
  - NCBI BLAST 2.2.X or higher

- Optional Components:
  - Augustus 2.0 or higher
  - GeneMark-ES 2.3a or higher
  - FGENESH 2.6 or higher

- Required for optional MPI support:
  - MPICH2

Go to Wiki Here

Getting Started with MAKER

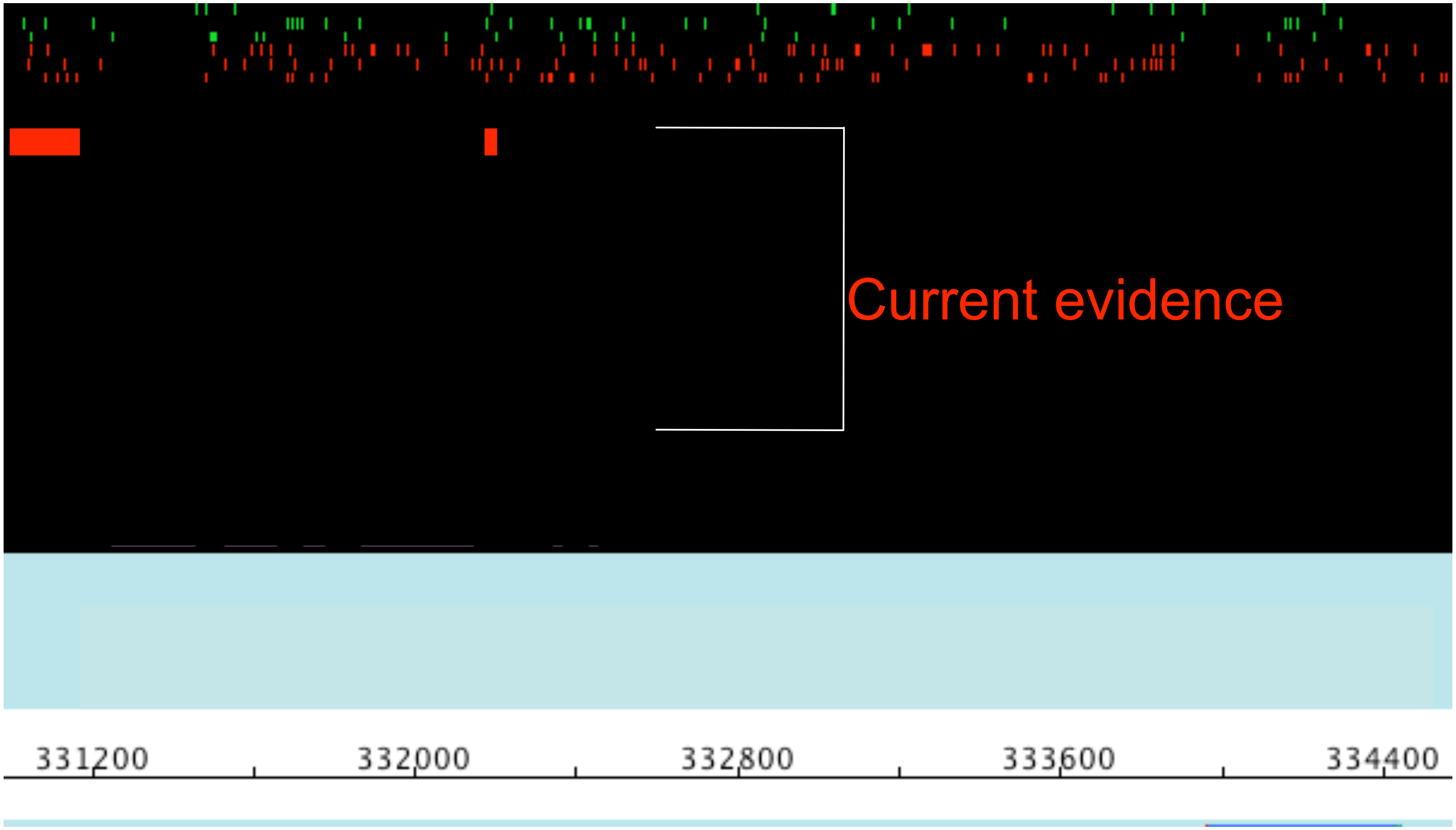
- Running example data
  - Installation sanity check
  - Familiarize yourself with very basic run configuration

Go to Wiki Here

What is Happening Inside MAKER

- RepeatMasking
- *Ab Initio* Gene Prediction
- EST and Protein Evidence Alignment
- Polishing Evidence Alignments
- Integrating Evidence to Synthesize Final Annotations

Identify and Mask Repetitive Elements



Current Assembly

Repeat Masking

- RepeatMasker - primarily nucleotide repeats
  - RepBase
  - Species specific library (built using PILER or RepeatScout).
- RepeatRunner - repeat related proteins
  - MAKER internal protein library

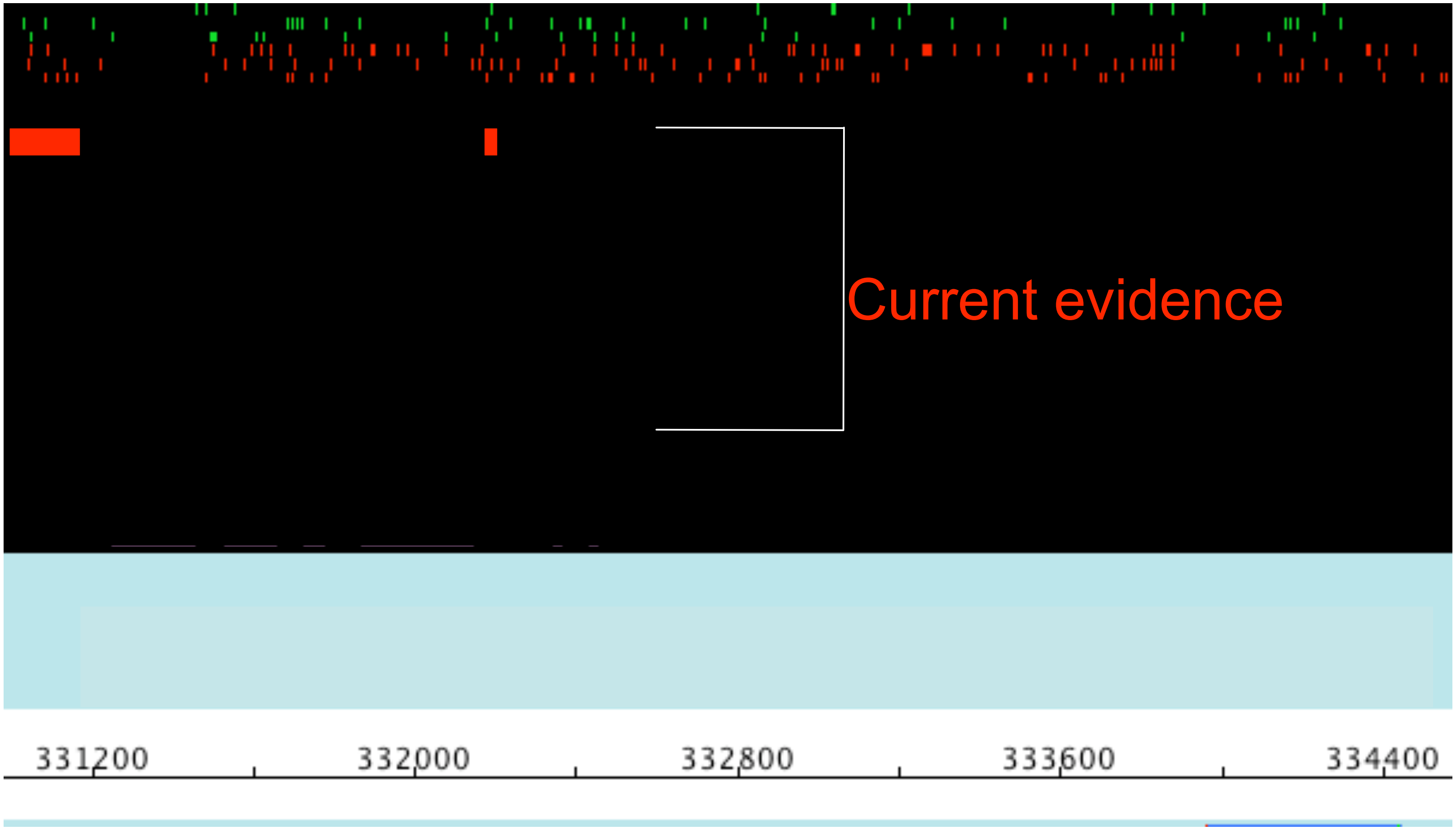
Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* **110**, 462-467 (2005).
Smith, C.D. et al. Improved repeat identification and masking in Dipterans. *Gene* **389**, 1-9 (2007).

Repeat Masking

- High-complexity repeat
  - Hard mask : AGGNNNNNNNNNACG
- Low-complexity repeat
  - Soft mask : AGGcctcctccACG

- Set unmask:1 in maker_opt.ctl to also run gene predictors on unmasked sequence

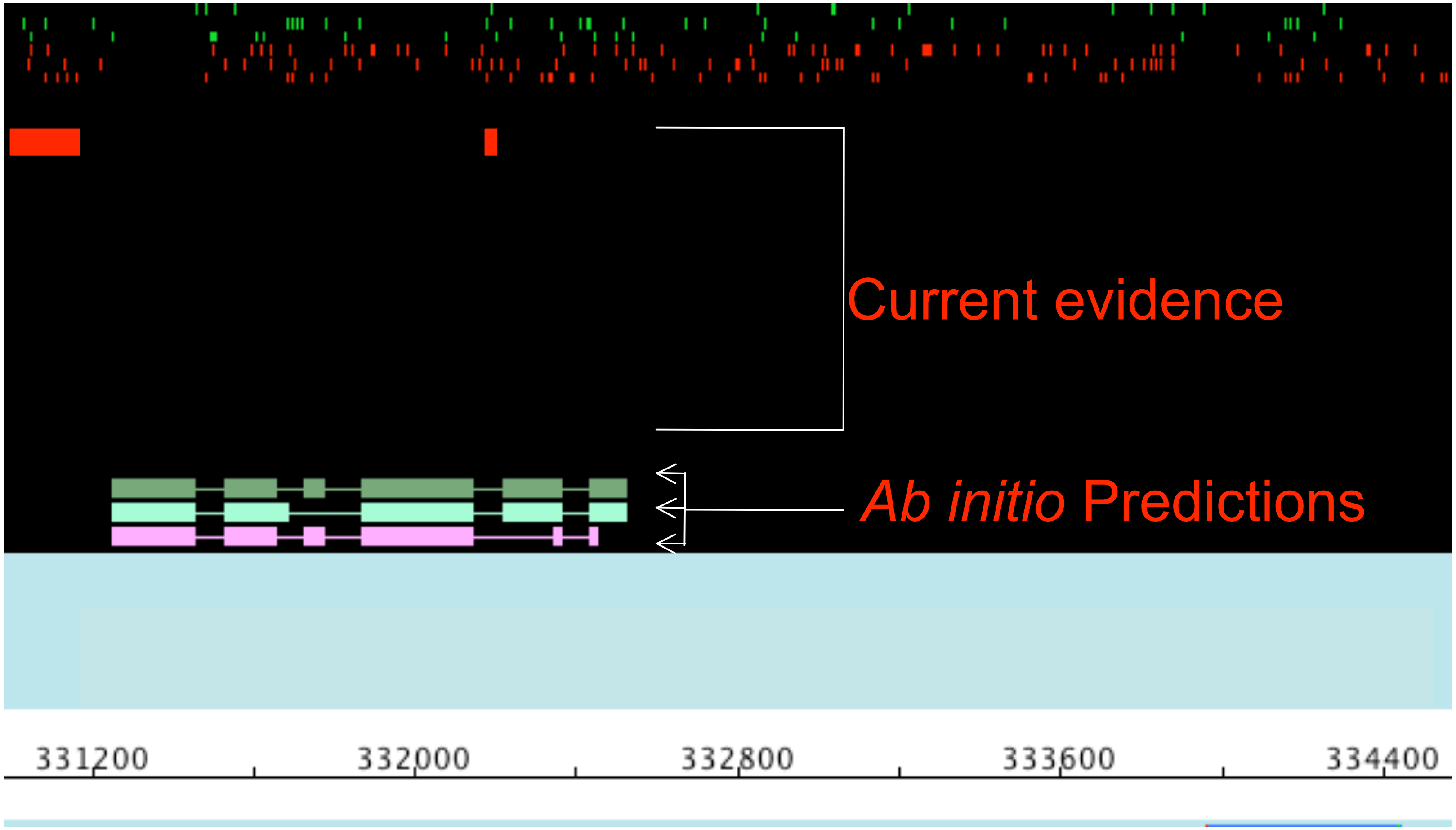
Identify and Mask Repetitive Elements



Current evidence

Current Assembly

Generate *Ab Initio* Gene Predictions



Current evidence

Ab initio Predictions

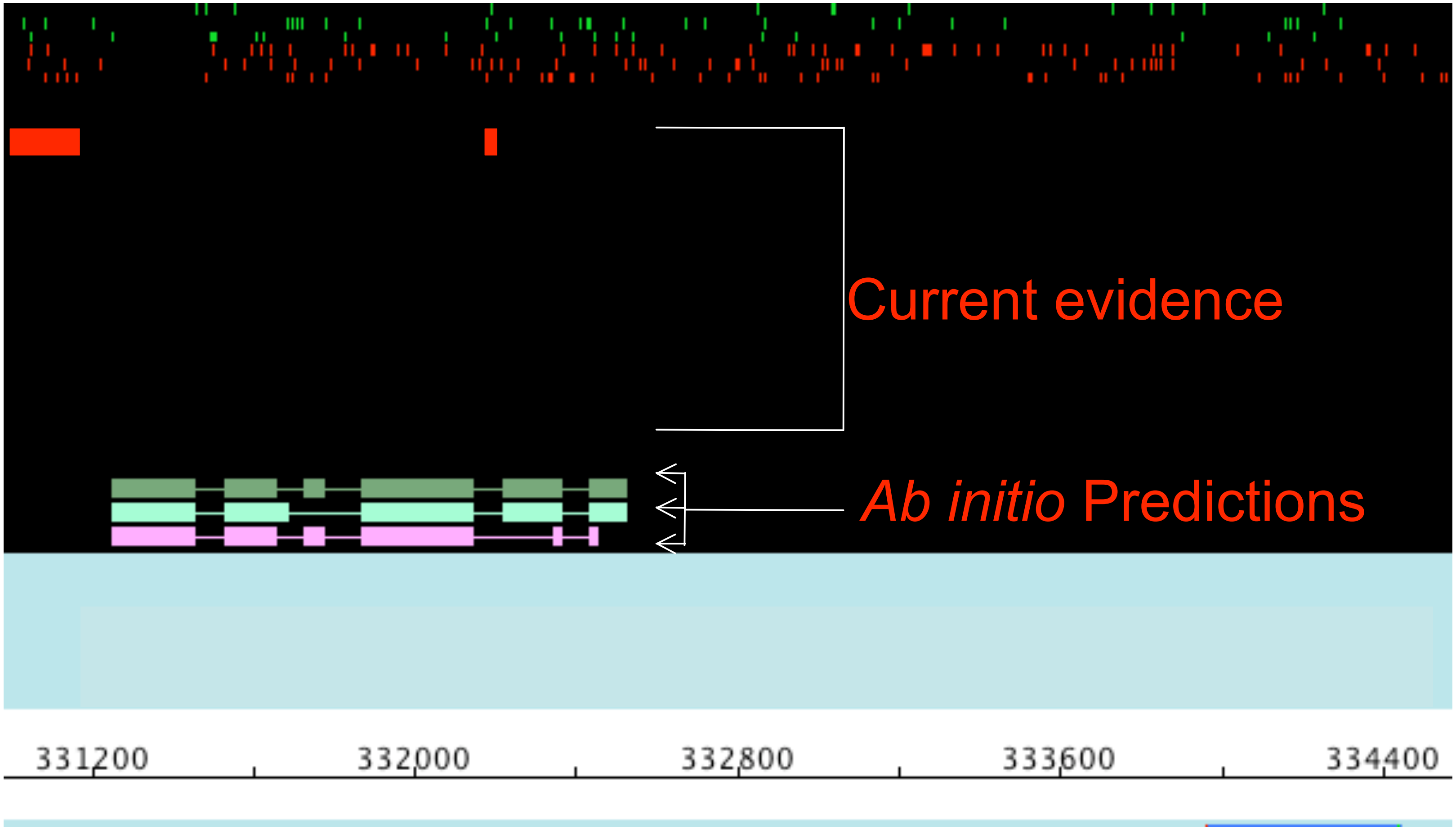
Current Assembly

Generate *Ab Initio* Gene Predictions

- MAKER currently supports:
  - SNAP
  - Augustus
  - GeneMark
  - FGENESH

- Remember to supply HMM’s for each

Generate *Ab Initio* Gene Predictions

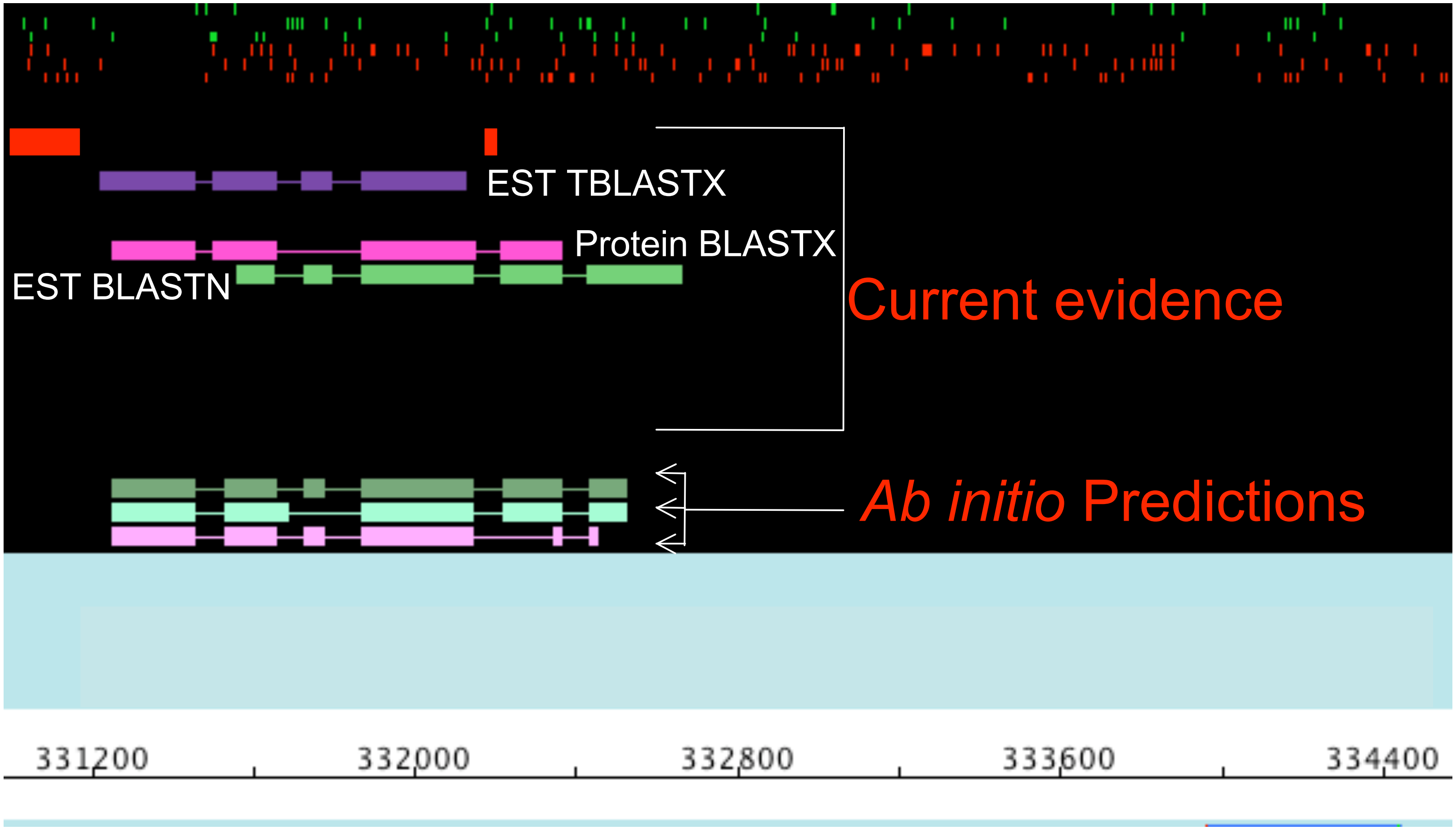


Current evidence

Ab initio Predictions

Current Assembly

Align EST and Protein Evidence



Current Assembly

EST and Protein Evidence Alignment

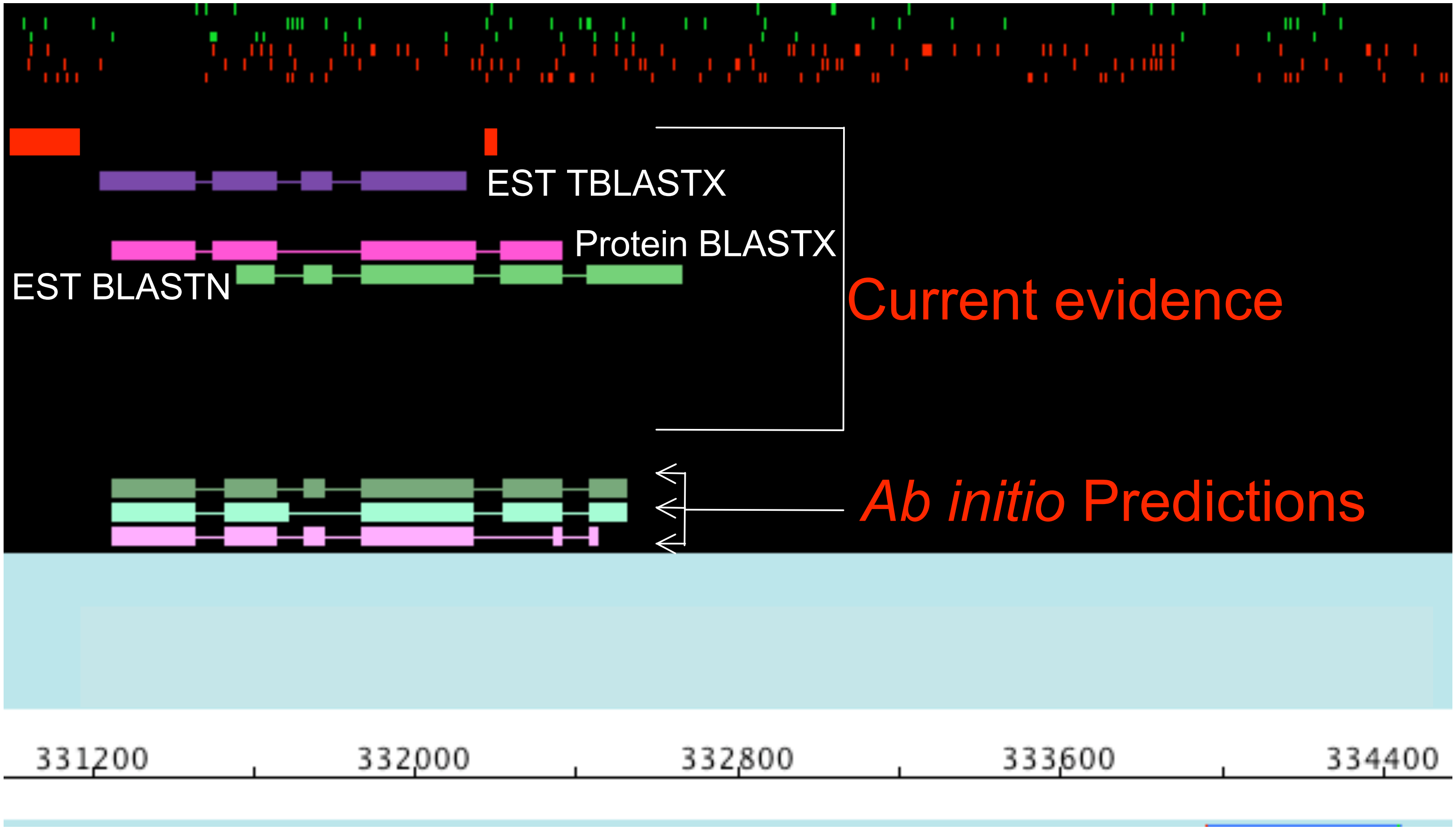
- A simple way to indicate if a sequence region is likely associated with a gene is to identify if:
  - The region is being actively transcribed (i.e. EST data)
  - The region has homology to a known protein

- ESTs from same or very closely related organism (i.e. human to chimpanzee). Use BLASTN
- ESTs from other related organisms (i.e. human to mouse). Use TBLASX
- Proteins from closely and distantly related organisms. Use BLASTX.

EST and Protein Evidence Alignment

- We are aligning against a masked genome. There are true alignments that are low complexity that are lost because of the masking.
- WU-BLAST has a feature that let’s you extend alignments through low-complexity regions, if the initial alignment starts outside of the region.
- This allows you to extend what may be real alignments while still filtering you the majority of spurious alignments. This is done via soft masking.
- Turn this feature off by setting softmask:0 in the maker_bopts.ctl file.

Align EST and Protein Evidence

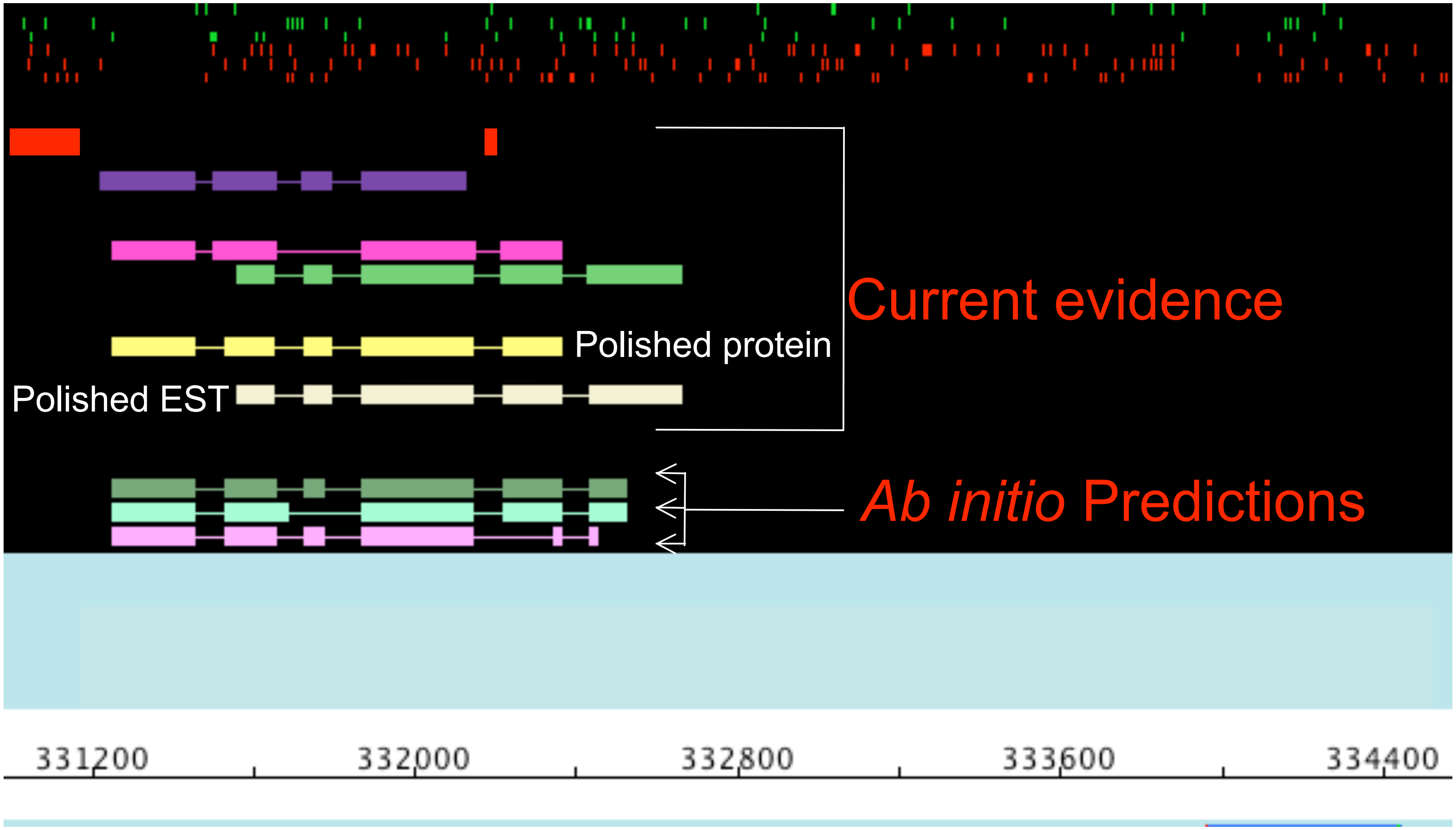


Current evidence

Ab initio Predictions

Current Assembly

Polish BLAST Alignments with Exonerate



Polished protein

Polished EST

Current evidence

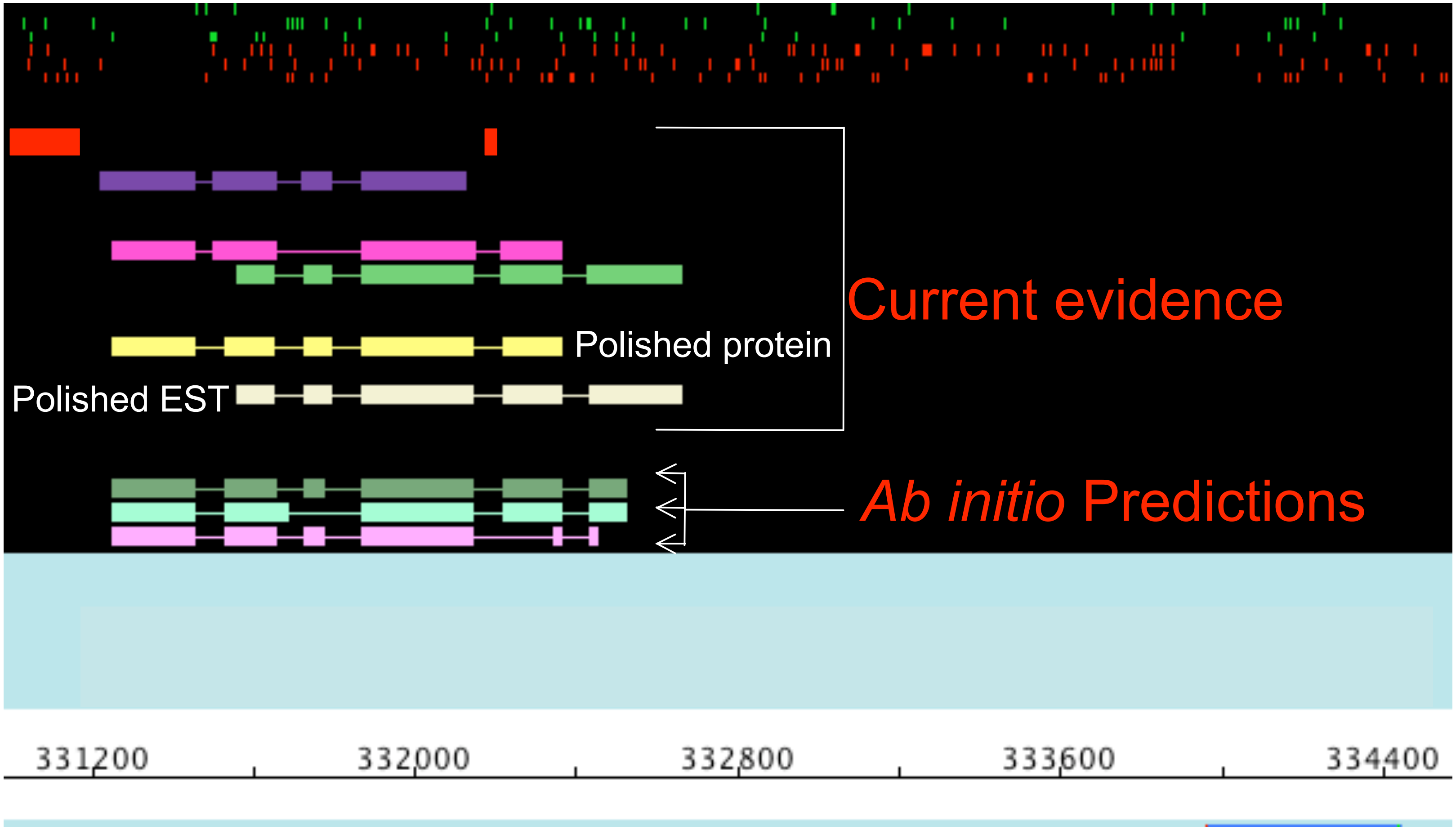
Ab initio Predictions

Current Assembly

Polish BLAST Alignments with Exonerate

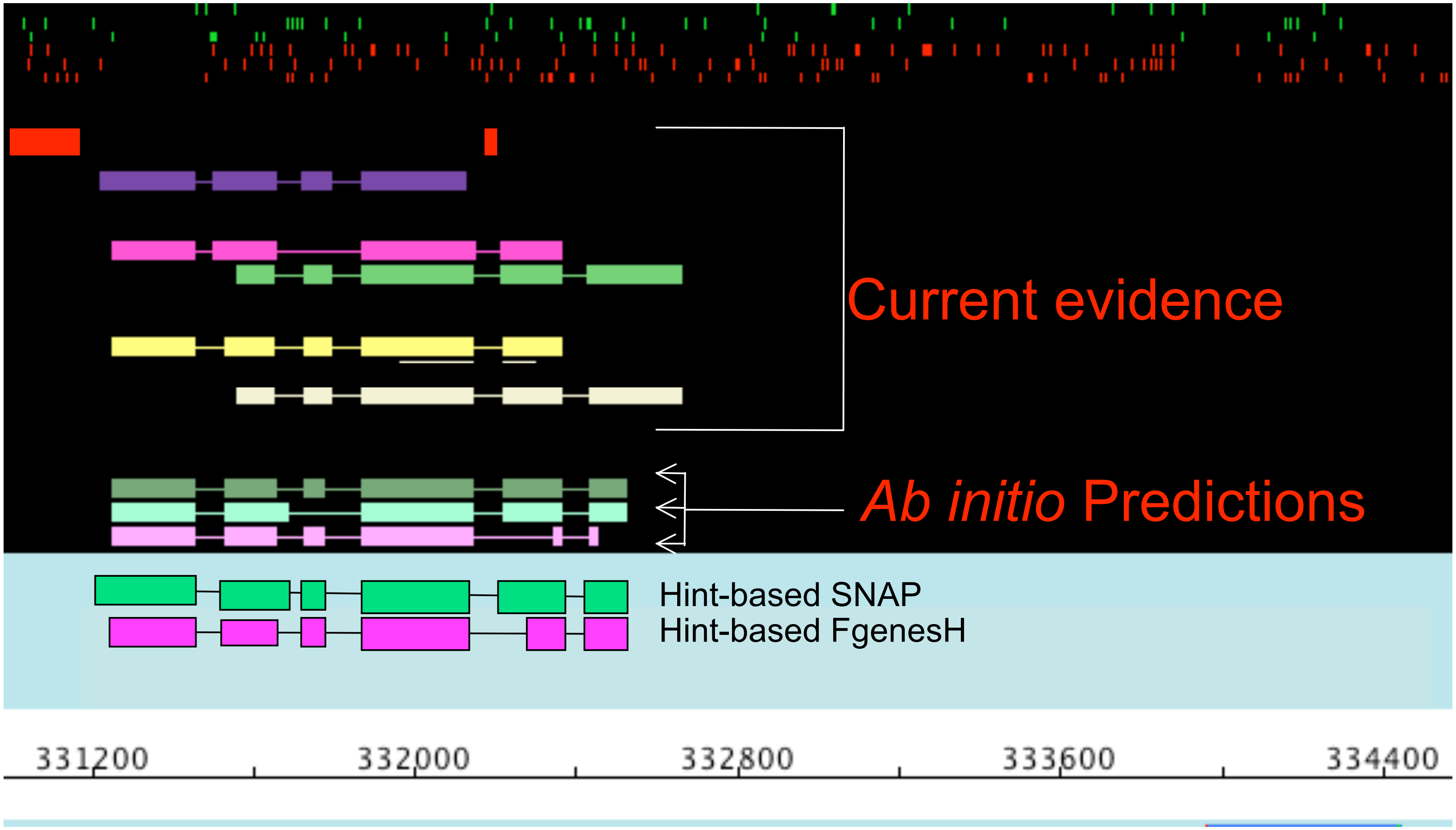
- All base pairs must aligns in order.
- No HSP overlap is permitted
- Aligns HSPs correctly with respect to splice sites.

Polish BLAST Alignments with Exonerate



Current Assembly

Pass Gene Finders Evidence-based ‘hints’



Current evidence

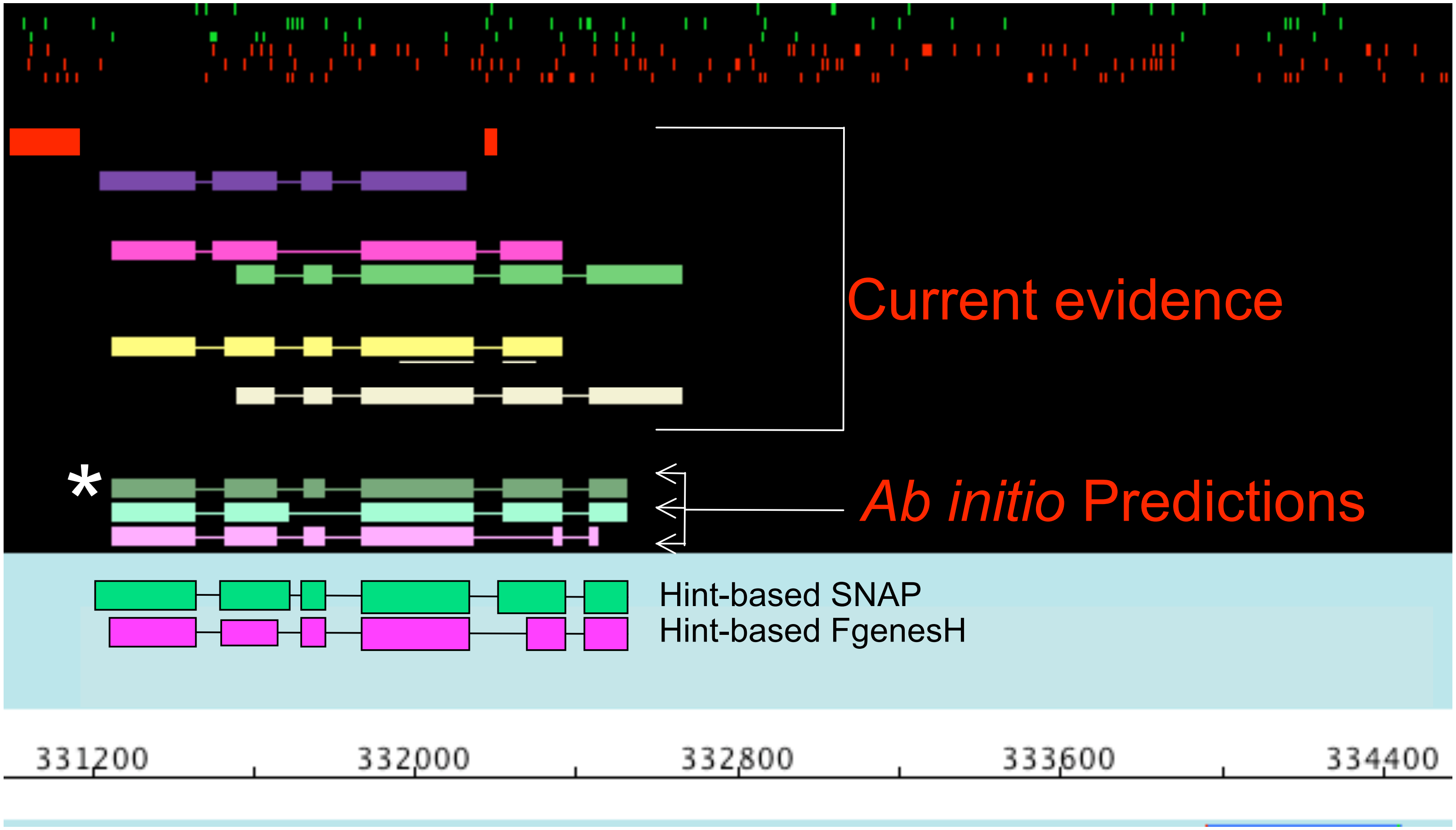
Ab initio Predictions

Hint-based SNAP

Hint-based FgenesH

Current Assembly

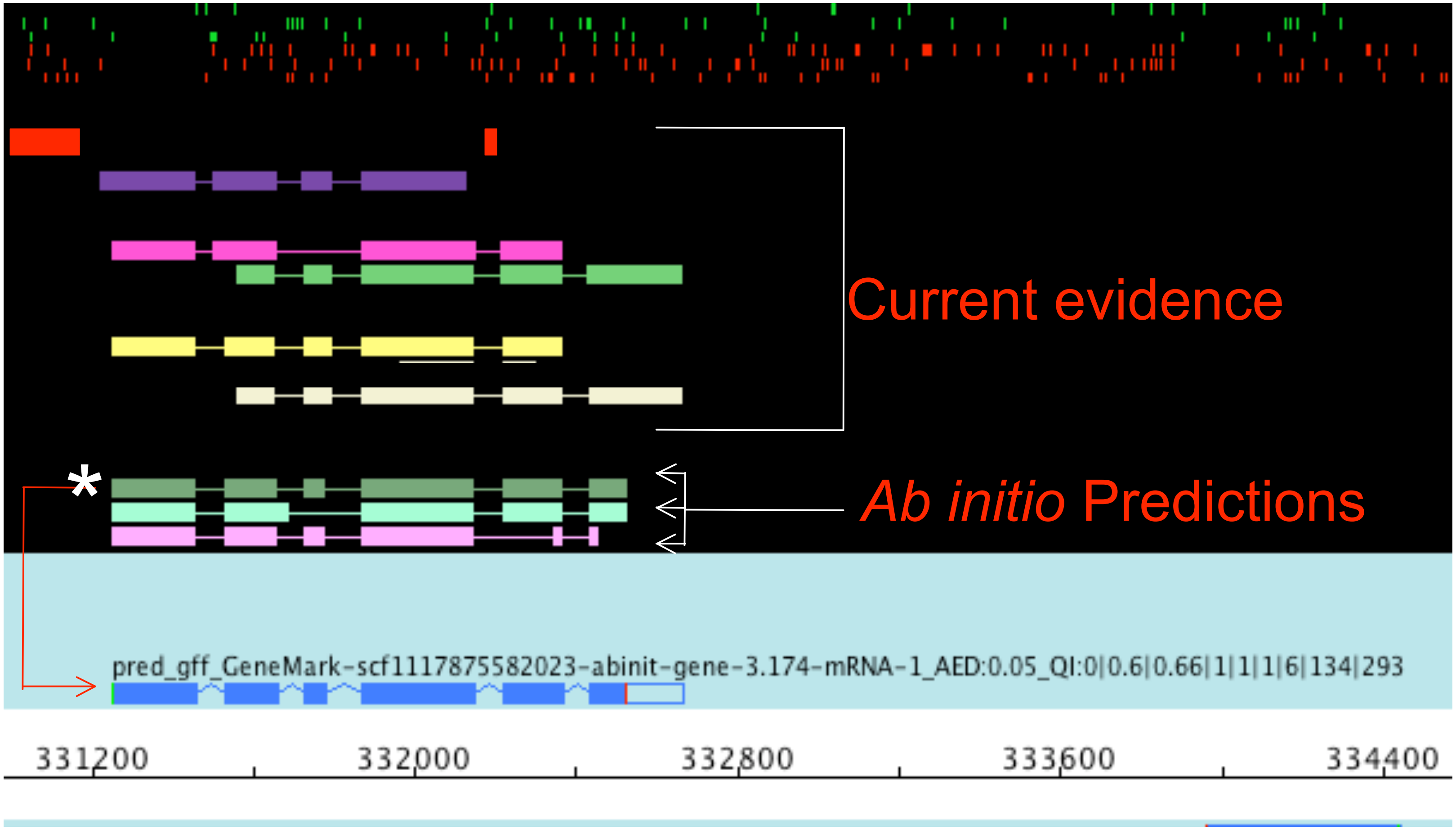
Identify Gene Model Most Consistent with Evidence*



Current Assembly

*Quantitative Measures for the Management and Comparison of Annotated Genomes
Karen Eilbeck , Barry Moore , Carson Holt and Mark Yandell BMC Bioinformatics 2009 10:67doi:10.1186/1471-2105-10-67

Revise it further if necessary; Create New Annotation


*
pred_gff_GeneMark-scf1117875582023-abinit-gene-3.174-mRNA-1_AED:0.05_QI:0|0.6|0.66|1|1|1|6|134|293
331200
332000
332800
333600
334400

Current evidence

Ab initio Predictions

Current Assembly

Compute Support for Each Portion of Gene Model

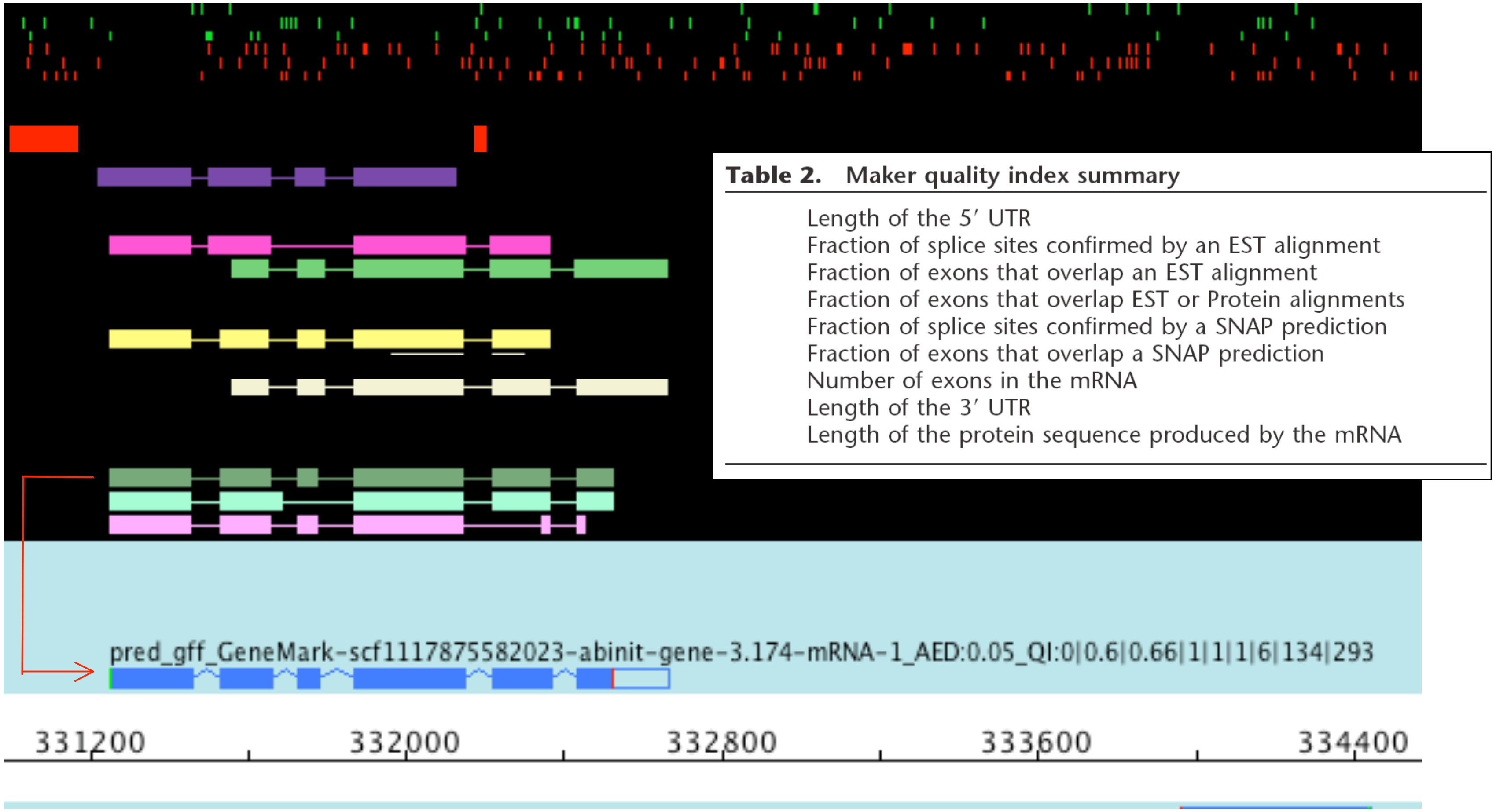


Table 2. Maker quality index summary

Length of the 5′ UTR
Fraction of splice sites confirmed by an EST alignment
Fraction of exons that overlap an EST alignment
Fraction of exons that overlap EST or Protein alignments
Fraction of splice sites confirmed by a SNAP prediction
Fraction of exons that overlap a SNAP prediction
Number of exons in the mRNA
Length of the 3′ UTR
Length of the protein sequence produced by the mRNA

pred_gff_GeneMark-scf1117875582023-abinit-gene-3.174-mRNA-1_AED:0.05_QI:0|0.6|0.66|1|1|1|6|134|293

MAKER’s Output

- Where it is
- What it is

Go to Wiki Here

Viewing MAKER Annotations

- Apollo
- GBrowse

Advanced MAKER Configuration, Re-annotation Options, and Improving Annotation Quality

MAKER II

Carson Holt
University of Utah

Configuration File Details

- Basic Input Files
- Repeat Masking Options
- Gene Prediction Options
- Other MAKER Options

Basic Input Files

- *genome* - Genomic sequence file
- *est* - ESTs from the same organism or from a very very closely related organism (i.e. chimpanzee to human).
- *altest* - These are ESTs from other closely related organisms (i.e. mouse to human).
- *protein* - proteins from the same or other organisms.

Repeat Masking Options

- *model_org* - This is a RepeatMasker option that lets you limit the repeat database to specific organisms or groups of organisms (i.e. vertebrates, Nematodes, Drosophila, primates etc).
- *repeat_protein* - This is a fasta file of transposon and virus related proteins. MAKER has an internal database it uses by default.
- *rmlib* - This is a fasta file of nucleotide repeats provided by the user.

Gene Prediction Options

- predictor - This tells MAKER what programs to run for generating annotations.
  - *est2genome* - Allows high quality spliced Exonerate EST alignments to become gene annotations.
  - *model_gff* - This allows user defined models to be used
  - *pred_gff* - Use external predictions
  - *snap*
  - *augustus*
  - *genemark*
  - *fgenes*

Gene Prediction Options

- *unmask* - Produce ab initio gene predictions for unmasked sequence as well as for masked sequence
- *snaphmm* - SNAP training file
- *gmhmm* - GeneMark training file
- *augustus_species* - Augustus species ID
- *fgenesh_par_file* - FGENESH training file

Other MAKER Options

- *evaluate* - runs an experimental annotation quality analysis program (Evaluator) on each annotation.
- *max_dna_len* - sets the length for dividing up contigs into chunks for processing.
- *min_contig* - sets the minimum length a contig must have
- *min_protein* - sets the minimum length a predicted protein must have
- *split_hit* - sets the expected max intron size for evidence alignments
- *pred_flank* - sets the length for the sequence surrounding clusters of EST and protein evidence

Other MAKER Options

- *single_exon* - tells MAKER to consider single exon EST evidence when generating annotations.
- *single_length* - sets the minimum length required for single exon ESTs if 'single_exon' is enabled
- *keep_preds* - adds non-overlapping ab-inito gene prediction to the final annotation set.
- *retry* - sets the number of times to retry a contig if there is a failure
- *clean_try* - removes all data from previous MAKER runs before retrying a contig
- *clean_up* - removes theVoid directory at the end of MAKER runs
- *TMP* - specifies a directory other than the system default temporary directory

Examples of Real Annotation Problems

- Training *ab initio* gene predictors
- Update existing annotations with new evidence
- Leverage existing data from programs not natively supported by MAKER
- Merge and resolve legacy annotations
- Using mRNAseq data

Training *ab initio* gene predictors

- You are involved in a genome project for an emerging model organism.
- You have no pre-existing gene models for training.
- What you do have:
  - ESTs

Bootstrapping SNAP

1. Train with EST based models
2. Run trained SNAP in MAKER
3. Retrain with new MAKER models
4. Run retrained SNAP in MAKER
5. Repeat steps 3 and 4 as needed.
6. Keep final annotations

Go to Wiki Here

GFF3 pass-through: How to update existing annotations

- You have an existing annotation set you don’t want to change
- You want to update the evidence associated with the annotations to include new evidence

Go to Wiki Here

GFF3 pass-through: How to use external evidence

- You have an existing annotation set.
- You want to update the evidence and allow the annotation to change to reflect the new evidence.

Go to Wiki Here

GFF3 pass-through: How to use external evidence

- You have an existing annotation set.
- You want to update the evidence and allow the annotation to change to reflect the new evidence.
- You want to rescue legacy names and apply them to updated annotations.

Go to Wiki Here

GFF3 pass-through: How to use external evidence

- You want to produce new annotations.
- You want to include gene predictions from TwinScan (not natively supported by MAKER)
- You want to include EST data from BLAT (not nativly supported by MAKER)

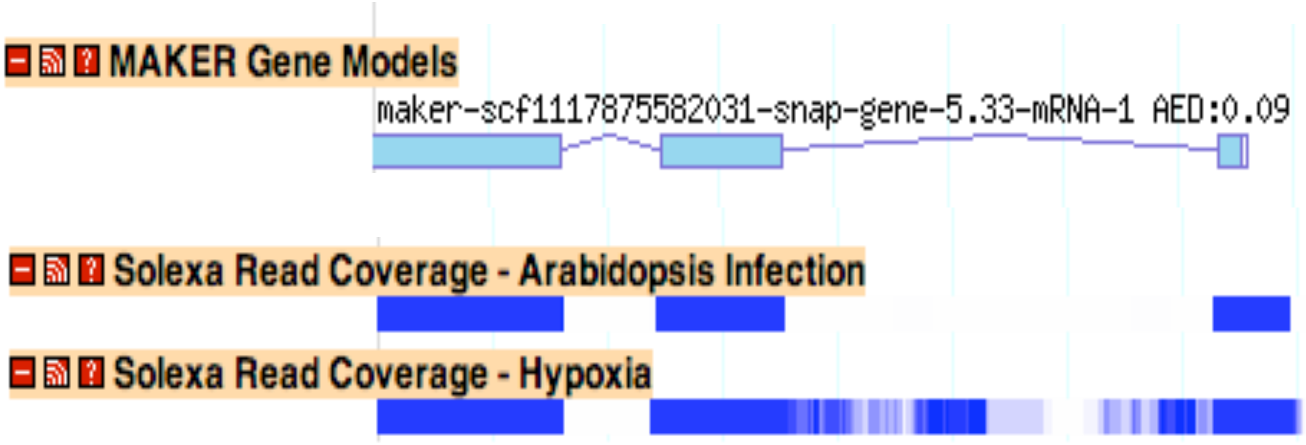
Go to Wiki Here

What if I have RNA-seq data?

RNA-seq is fundamentally changing the field of genome annotation

for both model *and* emerging genomes

RNA-seq may soon make gene prediction (mostly) a thing of the past


MAKER Gene Models
maker-scf1117875582031-snap-gene-5.33-mRNA-1 AED:0.09
Solexa Read Coverage - Arabidopsis Infection
Solexa Read Coverage - Hypoxia

- Still need to de-convolute reads & evidence (for now)
- Still need to archive and distribute annotations
- Still need to manage genome and its annotations

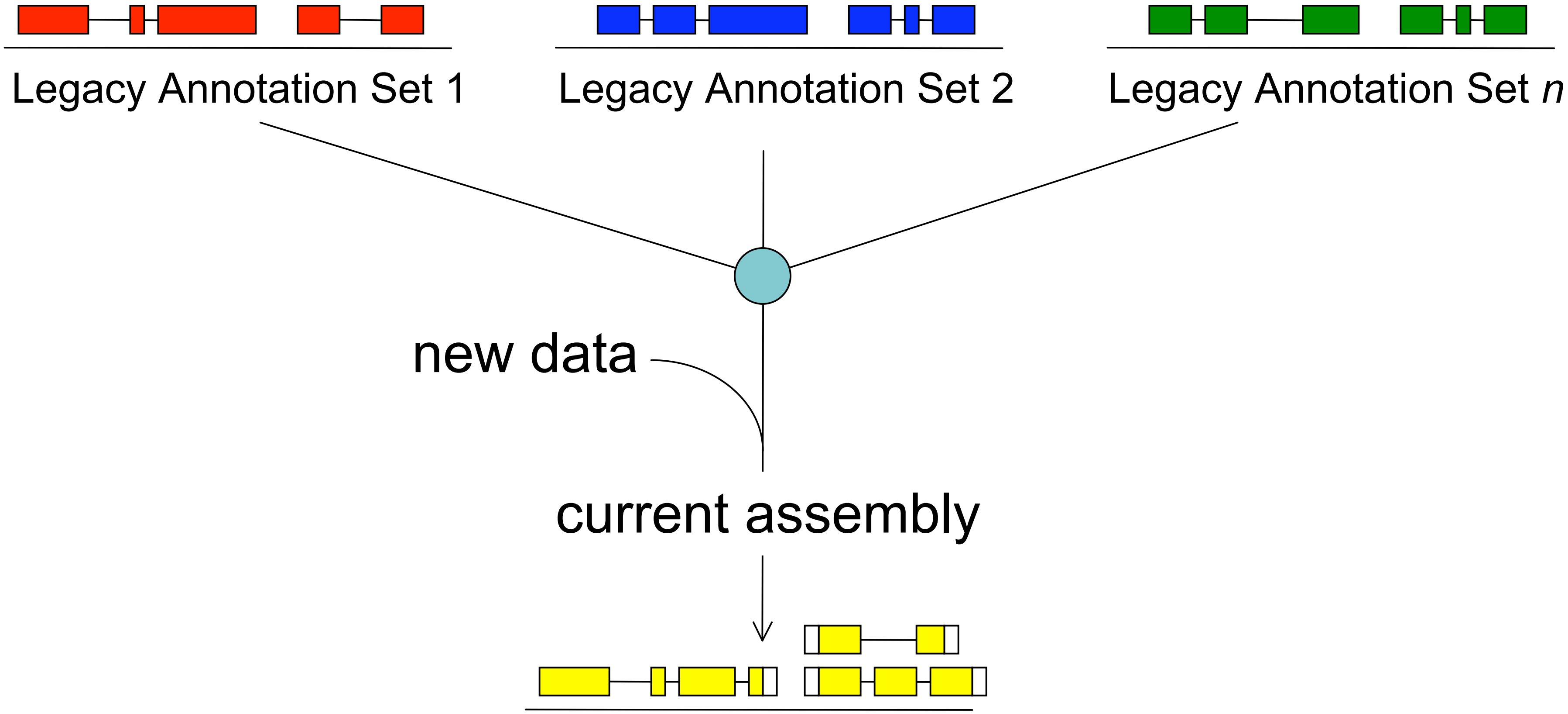
How to use RNA-seq data in MAKER

- Use BowTie and TopHat to produce, aligns reads into expression “islands” and “junctions”
- Pass data through as EST evidence via GFF3 pass-through.

Another issue: legacy annotations

- Many are no longer maintained by original creators
- In some cases more than one group has annotated the same genome, using very different procedures, even different assemblies
- The communities associated with those genomes are going to want RNA-seq data
- Many investigators have their own genome-scale data and would like a private set of annotations that reflect these data
- There will be a need to ***revise***, ***merge, evaluate***, and ***verify*** legacy annotation sets in light of RNA-seq and other data

Merging and Revising Legacy Annotation Sets



Legacy Annotation Set 1

Legacy Annotation Set 2

Legacy Annotation Set *n*

new data

current assembly

- Identify legacy annotation most consistent with new data
- Automatically revise it in light of new data
- If no existing annotation, create new one

Go to Wiki Here

MAKER Accessory Scripts

- Time permitting, let’s test them out.

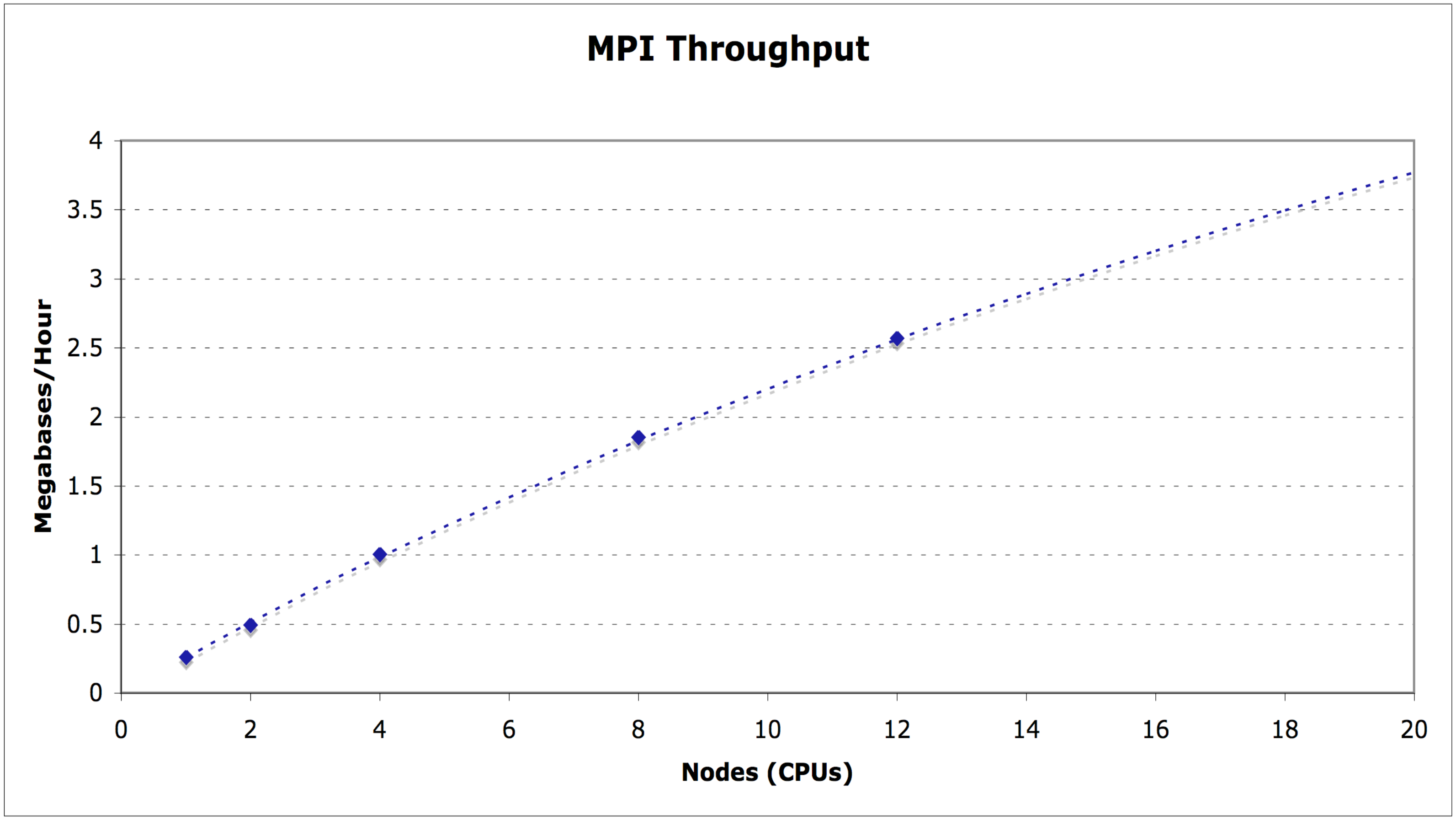
Other things MAKER

- MPI MAKER
- Web-service

MPI Support

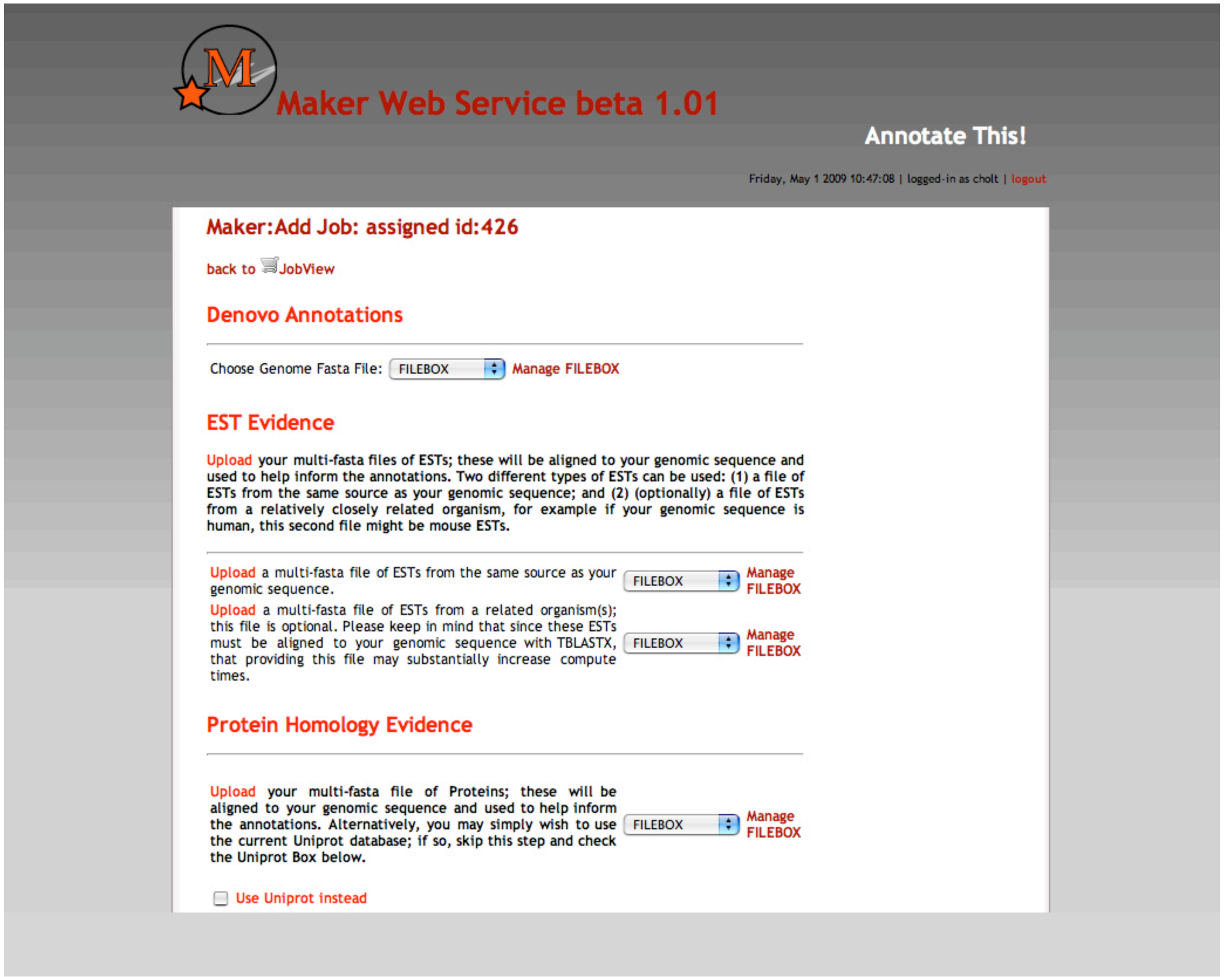
- Message Passing Interface (MPI) is a communication protocol for computer clusters which essentially allows multiple computers to act like a single powerful machine.

MPI Maker


MPI Throughput
Megabases/Hour
0
0.5
1
1.5
2
2.5
3
3.5
4
Nodes (CPUs)
0
2
4
6
8
10
12
14
16
18
20

Go to Wiki Here

MAKER will soon be available over the web



Maker Web Service beta 1.01

Annotate This!

Friday, May 1 2009 10:47:08 | logged-in as cholt | logout

Maker:Add Job: assigned id:426

back to JobView

Denovo Annotations

Choose Genome Fasta File: FILEBOX Manage FILEBOX

EST Evidence

Upload your multi-fasta files of ESTs; these will be aligned to your genomic sequence and used to help inform the annotations. Two different types of ESTs can be used: (1) a file of ESTs from the same source as your genomic sequence; and (2) (optionally) a file of ESTs from a relatively closely related organism, for example if your genomic sequence is human, this second file might be mouse ESTs.

Upload a multi-fasta file of ESTs from the same source as your genomic sequence. FILEBOX Manage FILEBOX

Upload a multi-fasta file of ESTs from a related organism(s); this file is optional. Please keep in mind that since these ESTs must be aligned to your genomic sequence with TBLASTX, that providing this file may substantially increase compute times. FILEBOX Manage FILEBOX

Protein Homology Evidence

Upload your multi-fasta file of Proteins; these will be aligned to your genomic sequence and used to help inform the annotations. Alternatively, you may simply wish to use the current Uniprot database; if so, skip this step and check the Uniprot Box below. FILEBOX Manage FILEBOX

☐ Use Uniprot instead

MAKER will soon be available over the web

Upload your multi-fasta file of Proteins; these will be aligned to your genomic sequence and used to help inform the annotations. Alternatively, you may simply wish to use the current Uniprot database; if so, skip this step and check the Uniprot Box below.

□ Use Uniprot instead