# MAKER: An easy to use genome annotation pipeline

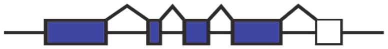
Carson Holt
Yandell Lab
Department of Human Genetics
University of Utah

# Introduction to Genome Annotation

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

# What Are Annotations?

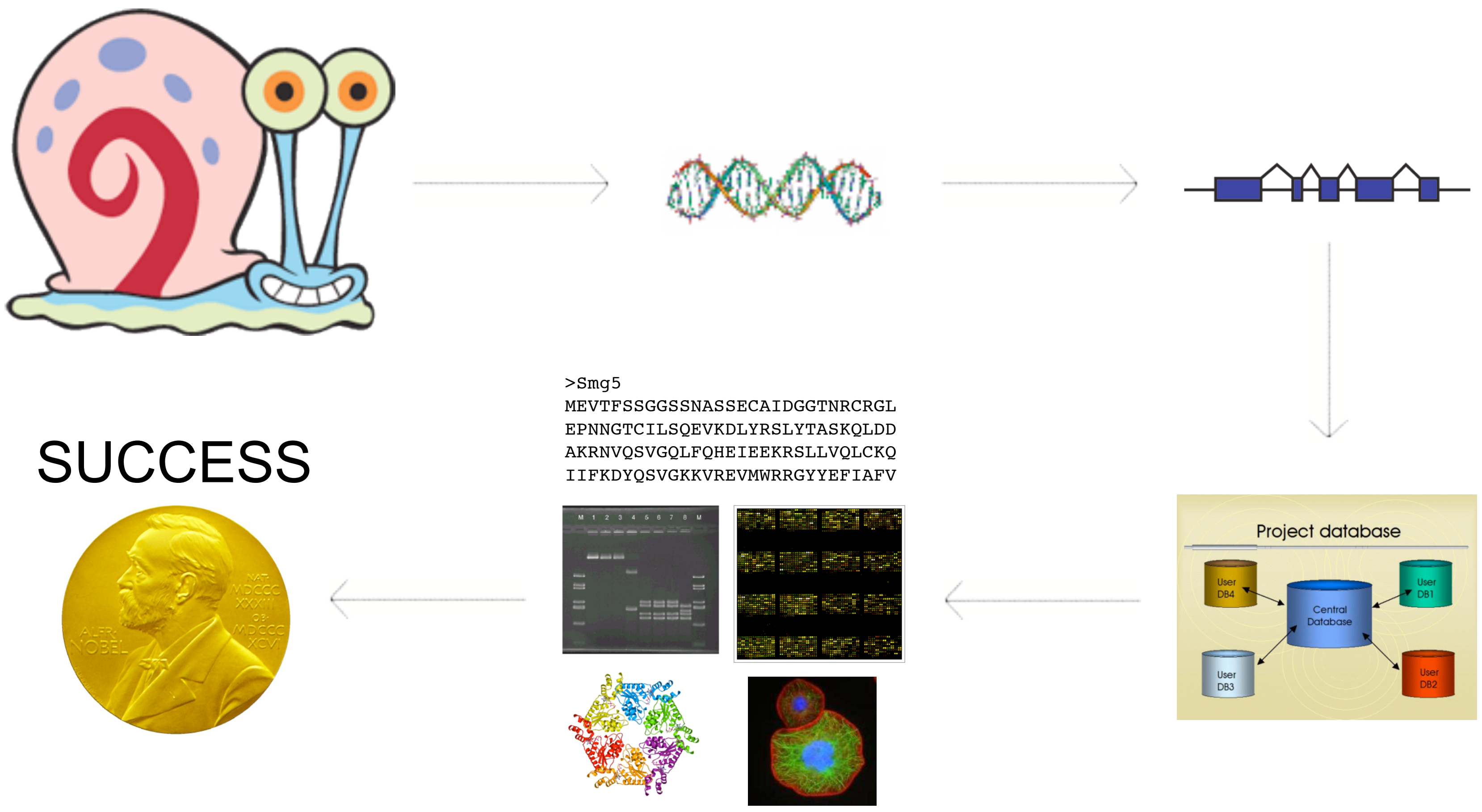
- Annotations are descriptions of features of the genome
  - Structural: exons, introns, UTRs, splice forms etc.
  - Functional: metabolism, hydrolase, expressed in the mitochondria, etc.



- 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

Background

# Why should I care about genome annotations?

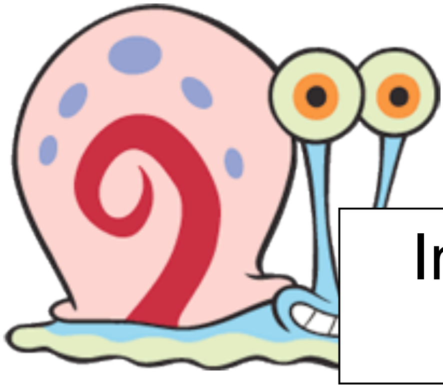


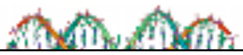
SUCCESS

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

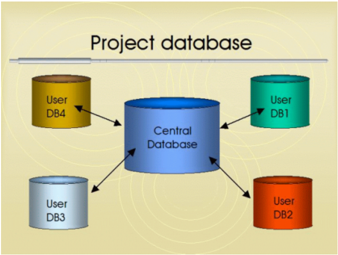
**Background**

# Why should I care about genome annotations?

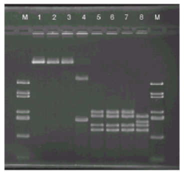


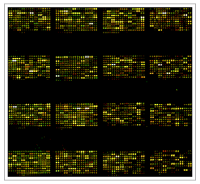


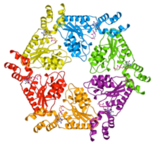
Incorrect annotations poison every experiment that uses them!!

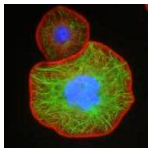


```
>Smg5
MEVTFSSGGSSNASSECAIDGGTNRCRGL
EPNNGTCILSQEVKDLYRSLYTASKQLDD
AKRNVQSVGQLFQHEIEEKRSLLVQLCKQ
IIFKDYQSVGKKVREVMWRRGYYEFIAFV
```

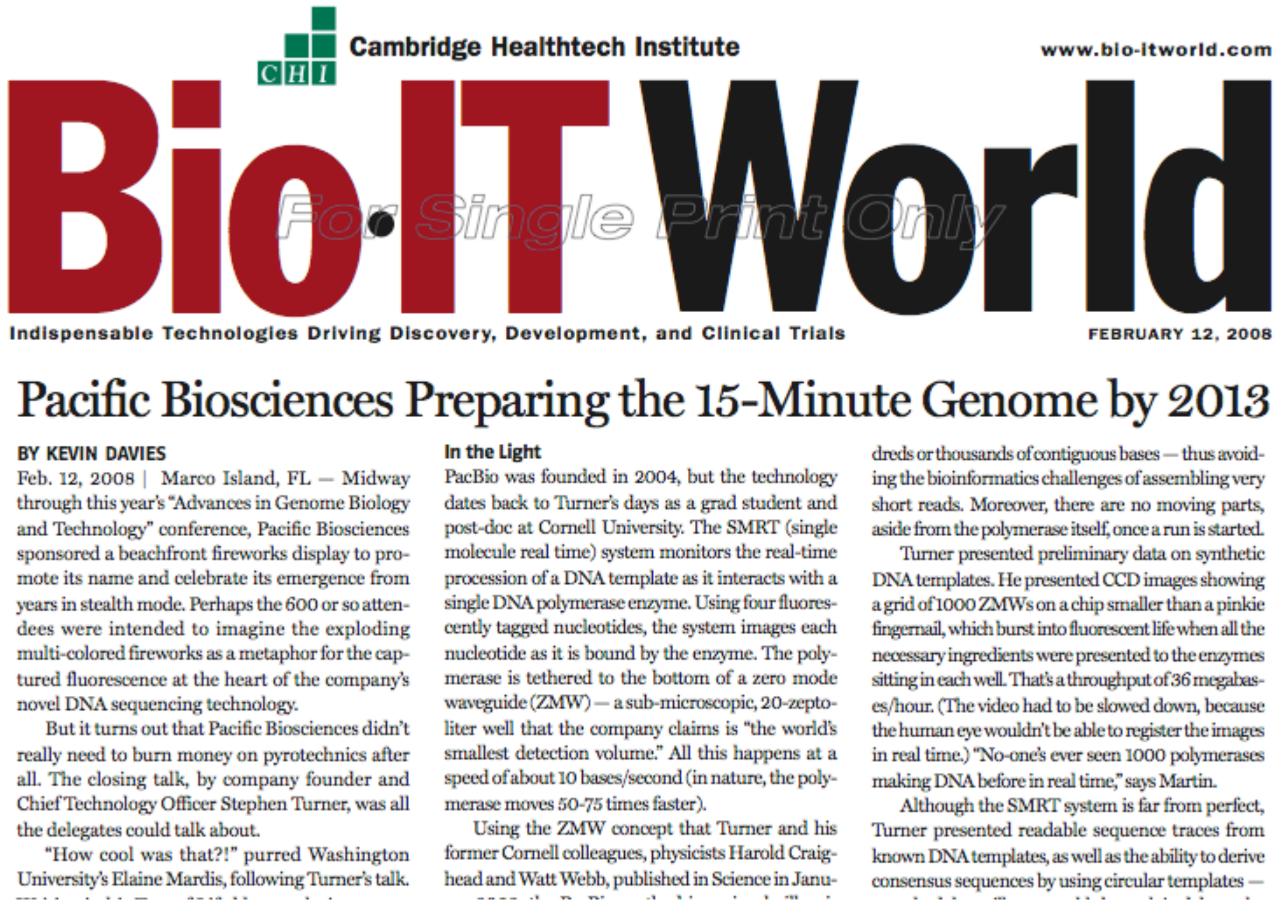








# Advances in Technology Promise to Make Whole Genome Sequencing “Routine” for Even Small Labs



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 October 2009, 222 eukaryotic genomes were fully sequenced yet unpublished.
- Currently there are over ~900 eukaryotic genome projects underway, assuming 10,000 genes per genome, that's 9,000,000 new annotations.
- There is a limit to how much data can be managed, maintained, and updated by a single organization.
- Small research groups affected disproportionately by difficulties related to genome annotation.

- .GOLD: Genomes OnLine Database. **2009**.

- 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.

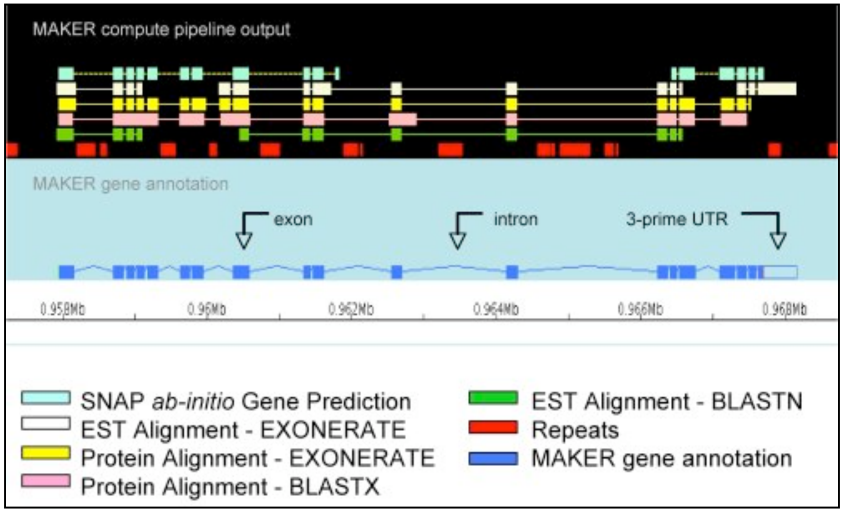
# MAKER Overview

- What does MAKER do?
- What sets MAKER apart from other tools (ab initio gene predictors, etc.)?

# 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

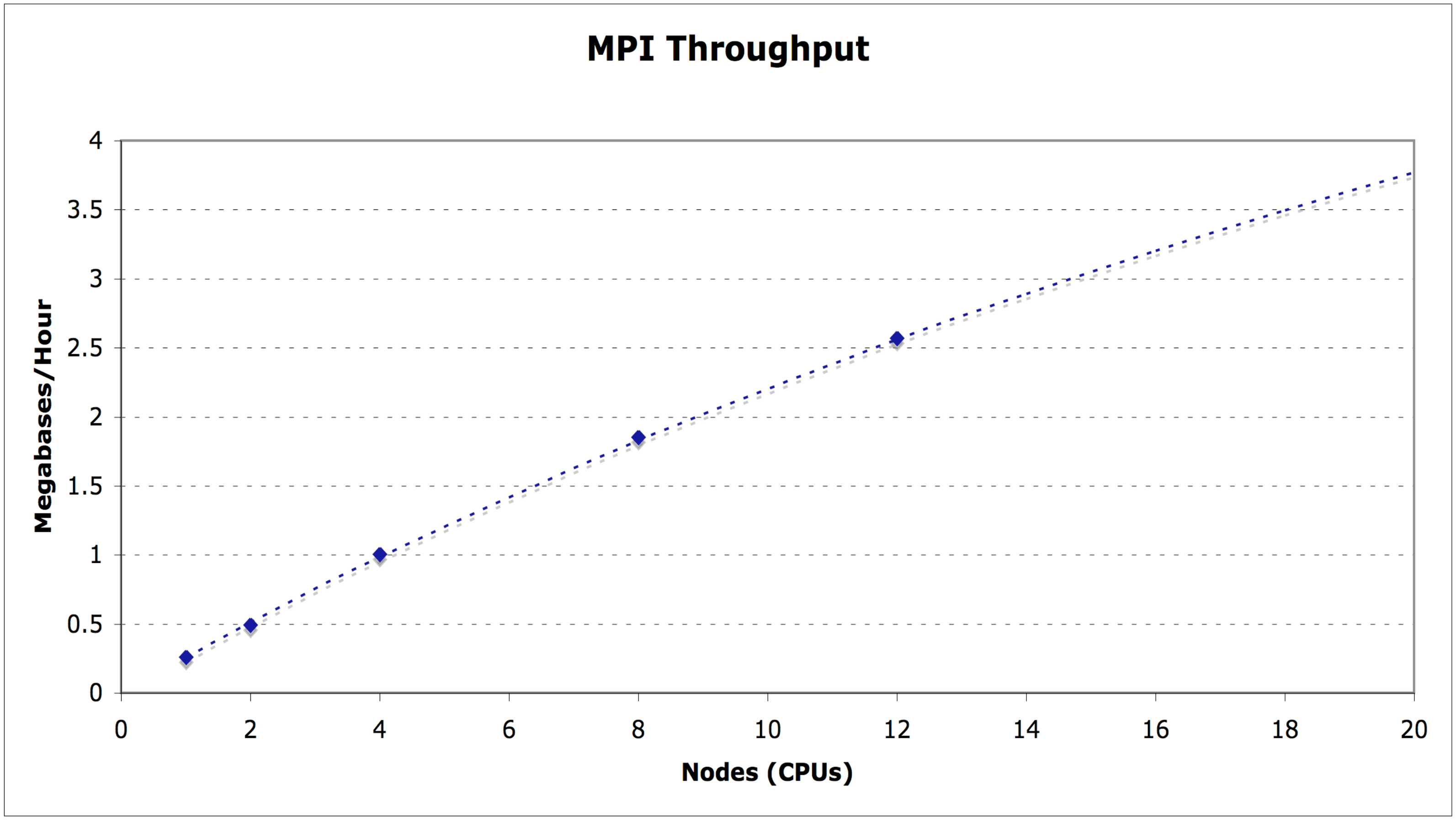
- Lewis, S.E. et al. Apollo: a sequence annotation editor. *Genome Biology* **3**, research0082.1 - 0082.14 (2002).
- Stein, L.D. et al. The Generic Genome Browser: A Building Block for a Model Organism System Database. *Genome Res.* **12**, 1599-1610 (2002).

# Other Features

# 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.

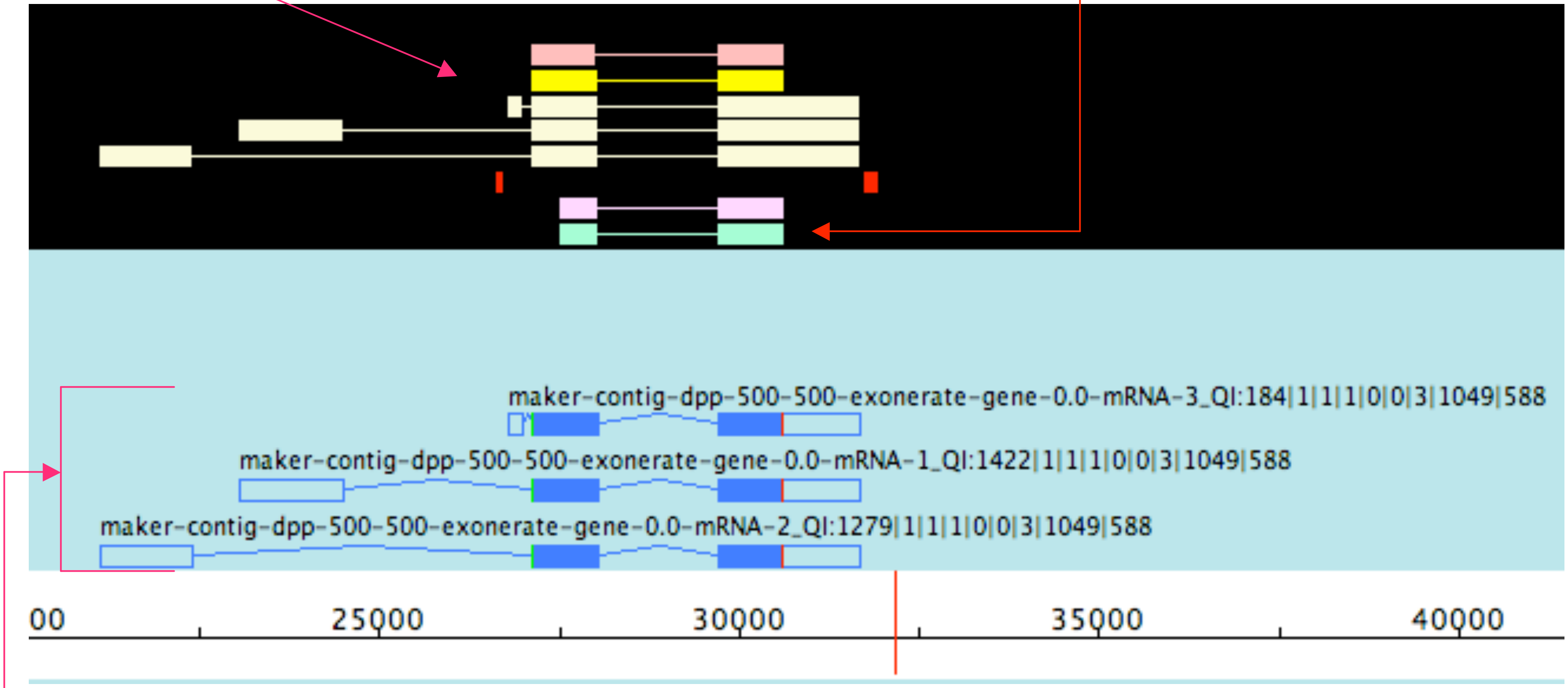
# MPI Maker


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

# 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

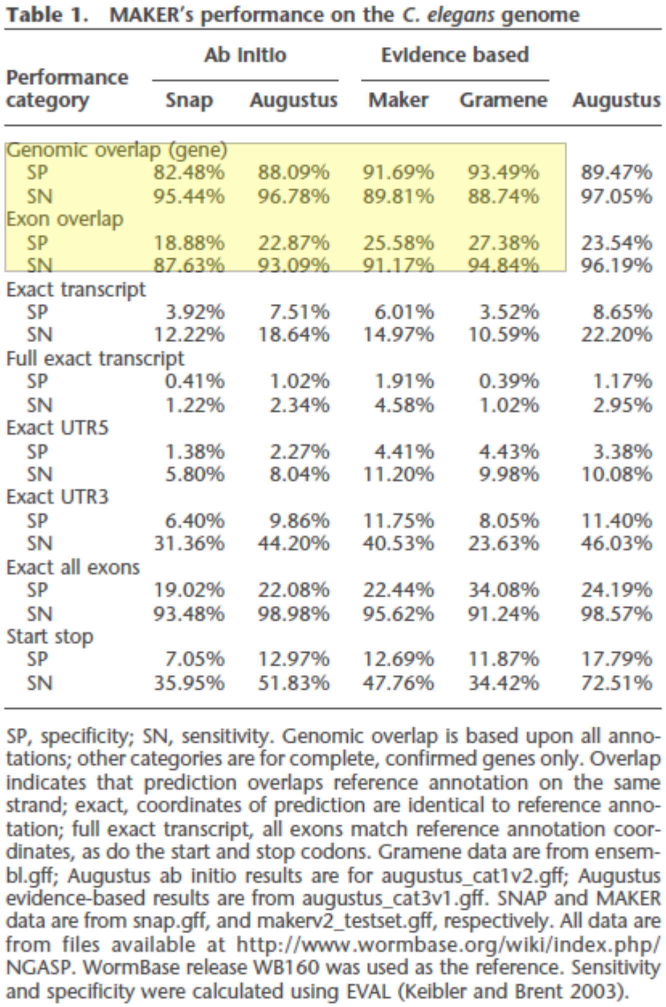


**Table 1.** MAKER's performance on the *C. elegans* genome

| 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

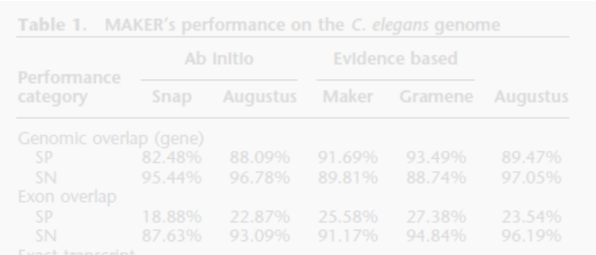
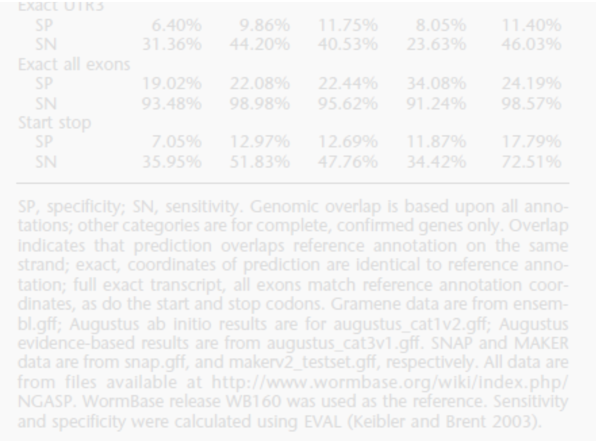


Table 1. MAKER's performance on the *C. elegans* genome

| 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 |  |  |  |  |  |

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



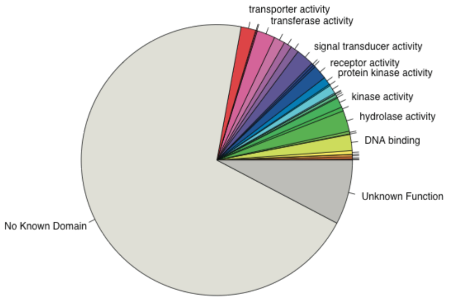
|  |  |  |  |  |  |
|---|---|---|---|---|---|
| Exact UTR3 |  |  |  |  |  |
| SP | 6.40% | 9.86% | 11.75% | 8.05% | 11.40% |
| SN | 31.36% | 44.20% | 40.53% | 23.63% | 46.03% |
| Exact all exons |  |  |  |  |  |
| SP | 19.02% | 22.08% | 22.44% | 34.08% | 24.19% |
| SN | 93.48% | 98.98% | 95.62% | 91.24% | 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).

***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*

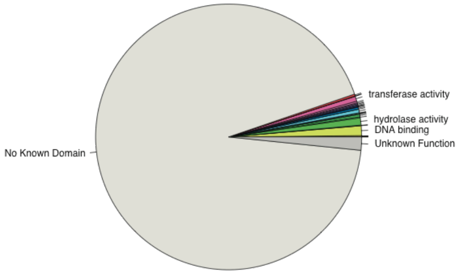
# *Ab initio* gene predictors don’t do nearly so well on emerging genomes*

**Average of seven REFERENCE proteomes**



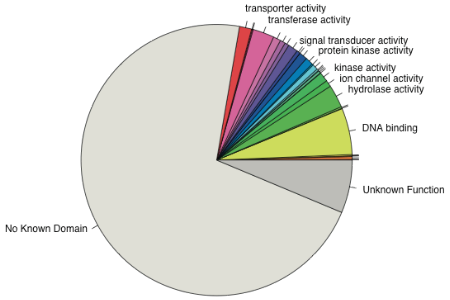
**35% contain a domain**

***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*

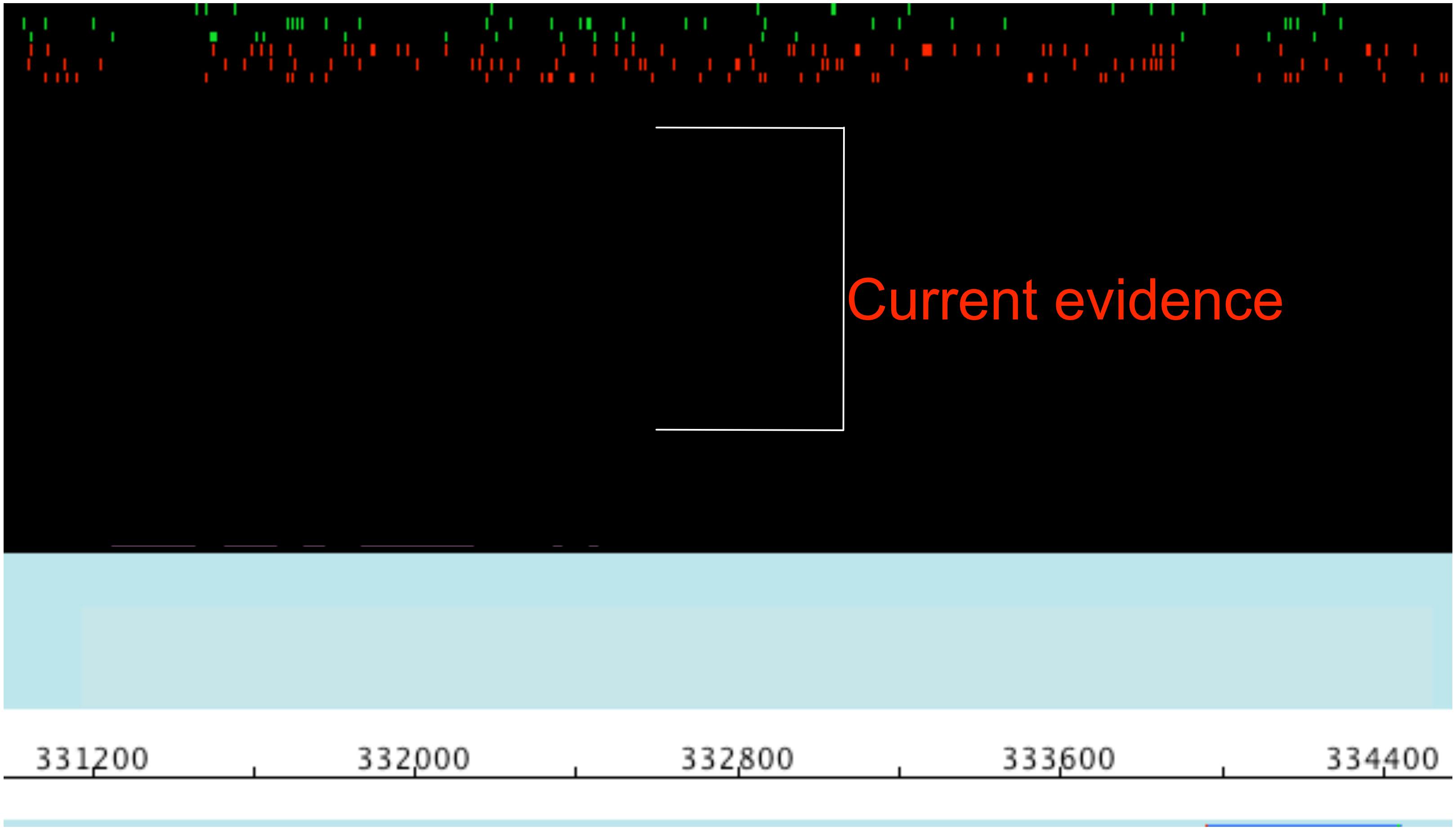
# Benefits of MAKER

- 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

# What is Happening Inside MAKER

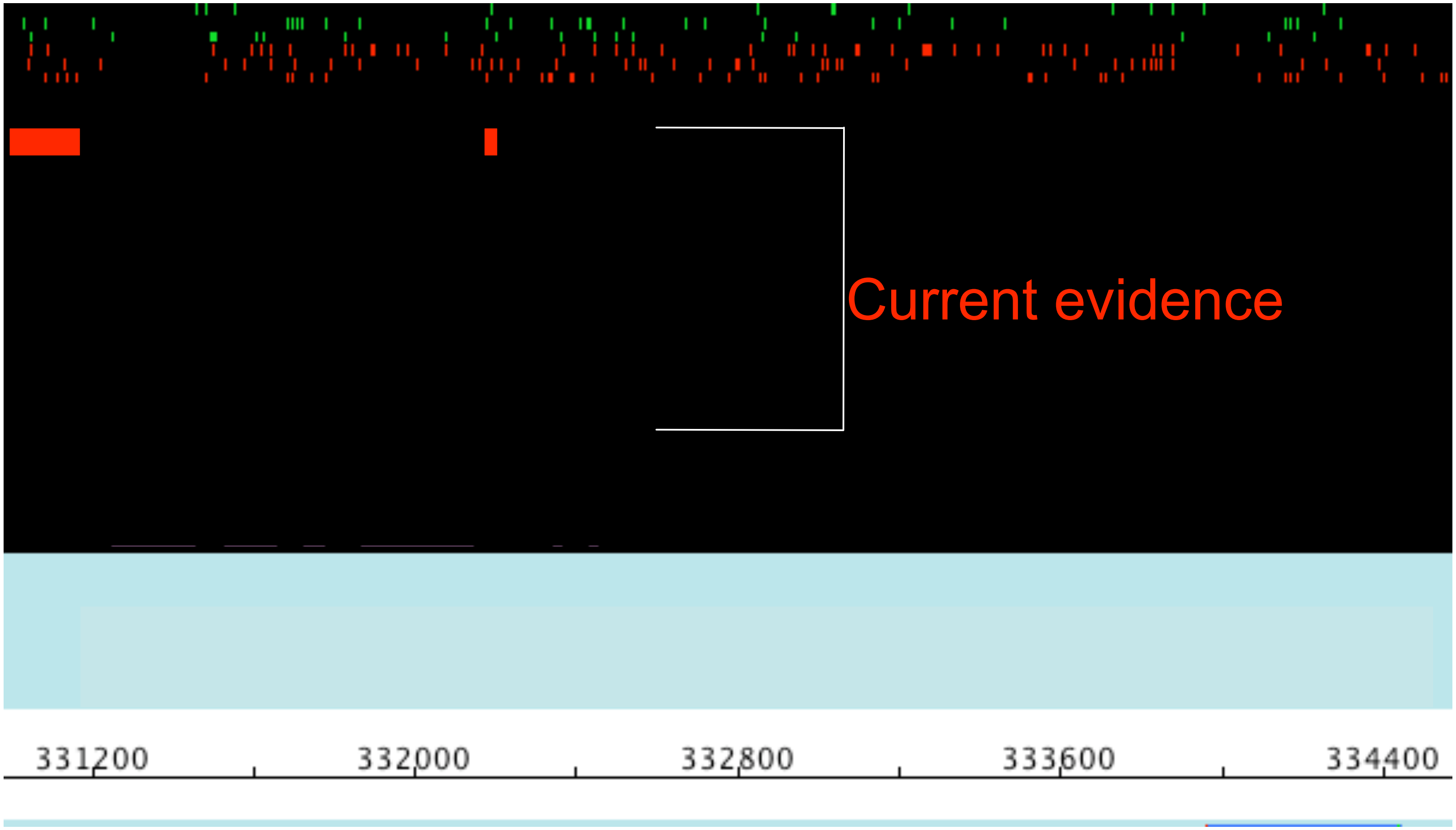
- RepeatMasking
- *Ab Initio* Gene Prediction
- EST and Protein Evidence Alignment
- Polishing Evidence Alignments
- Integrating Evidence to Synthesize Final Annotations

# Annotating the Genome – Apollo View



Current Assembly

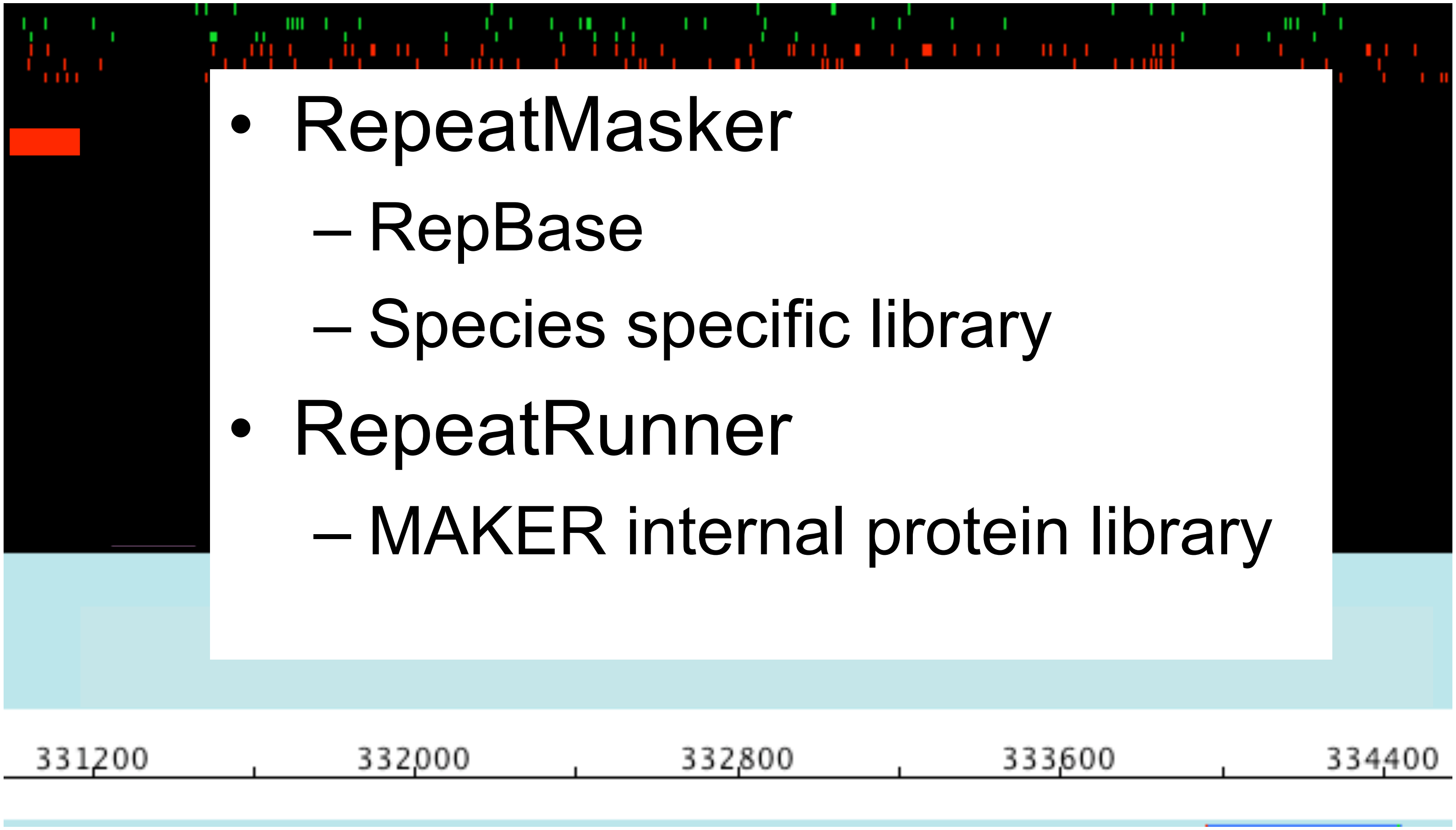
# Identify and Mask Repetitive Elements



Current evidence

Current Assembly

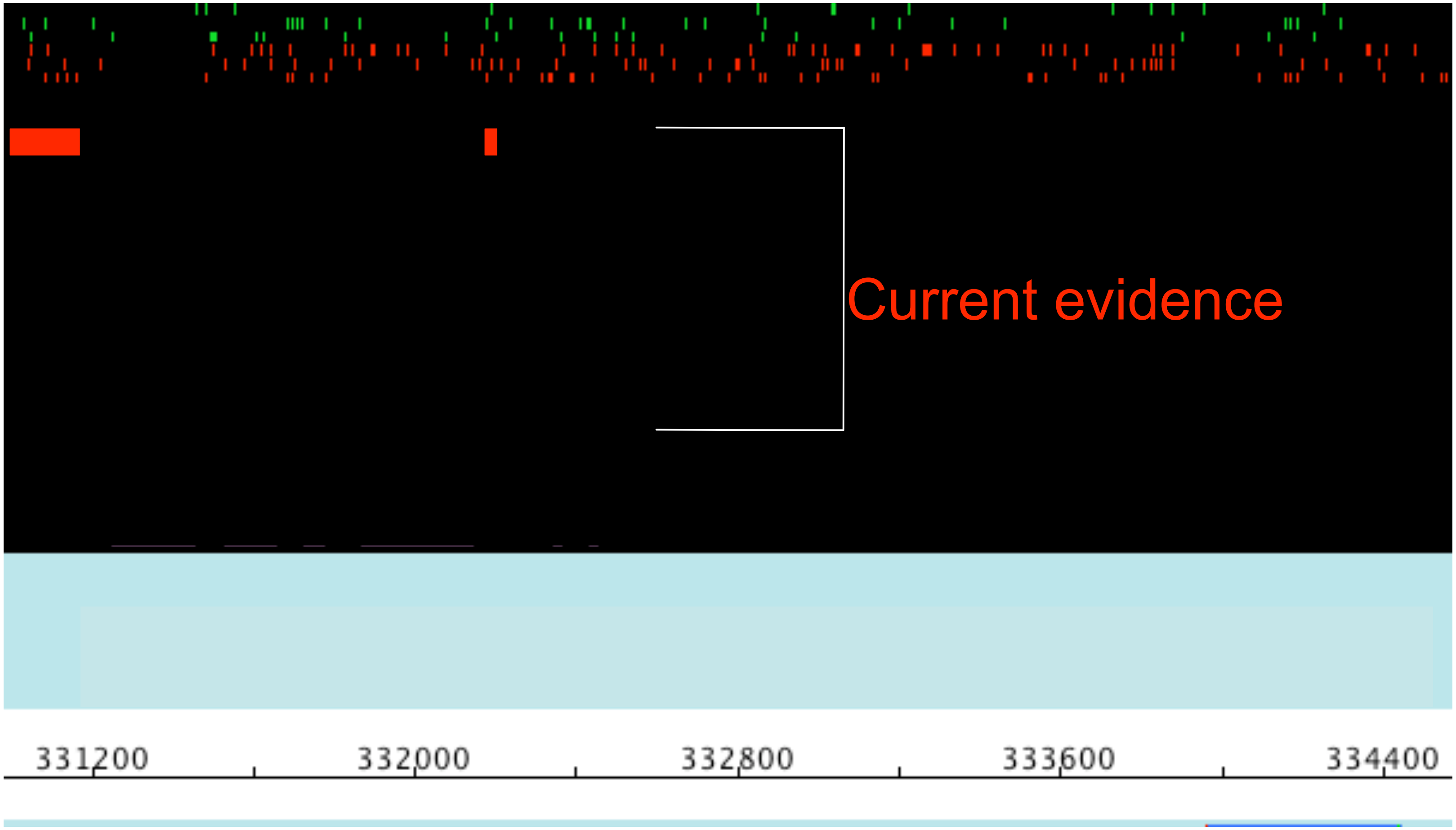
## Identify and Mask Repetitive Elements



- RepeatMasker
  - – RepBase
  - – Species specific library
- RepeatRunner
  - – MAKER internal protein library

Current Assembly

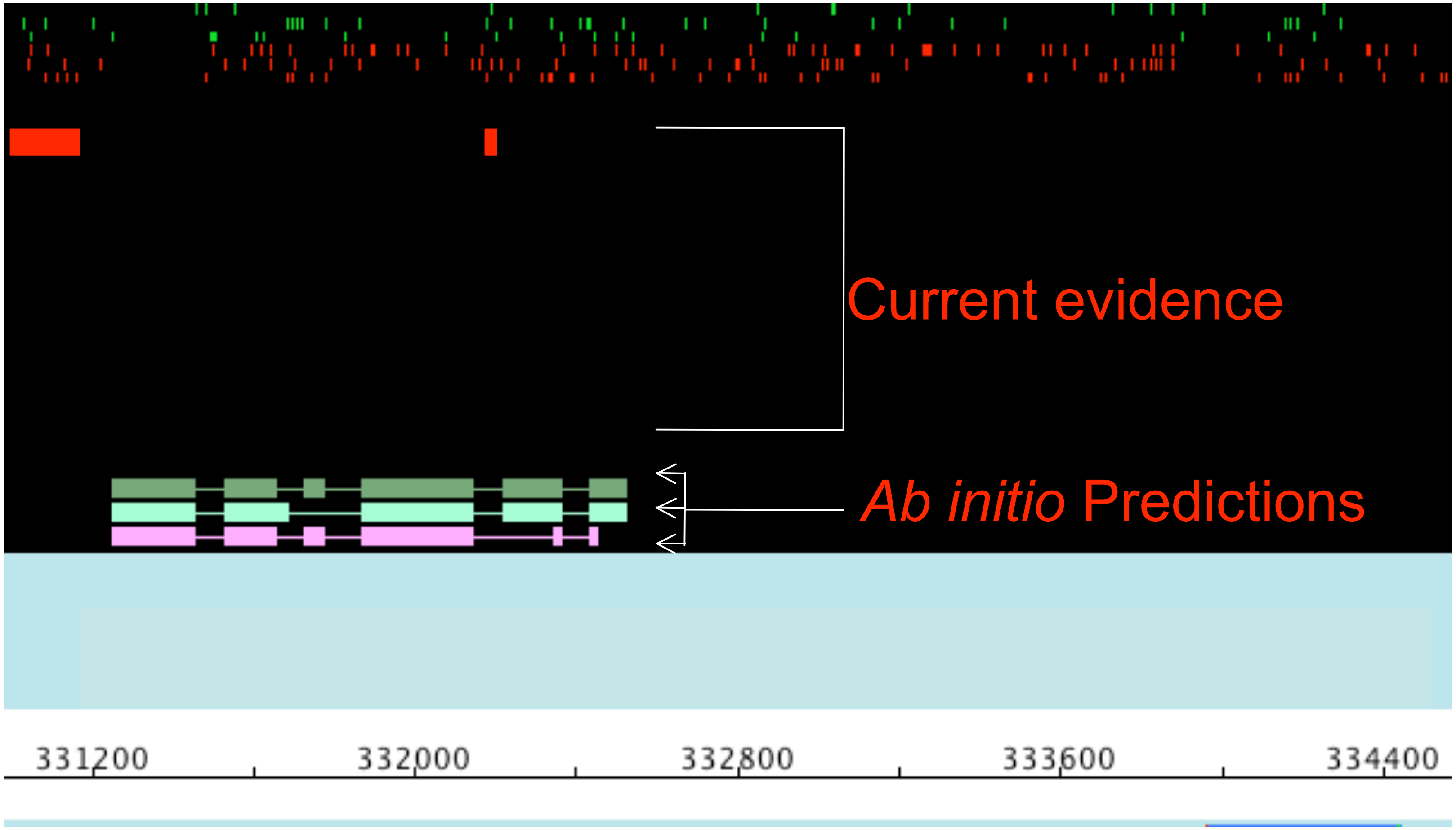
# Identify and Mask Repetitive Elements



Current evidence

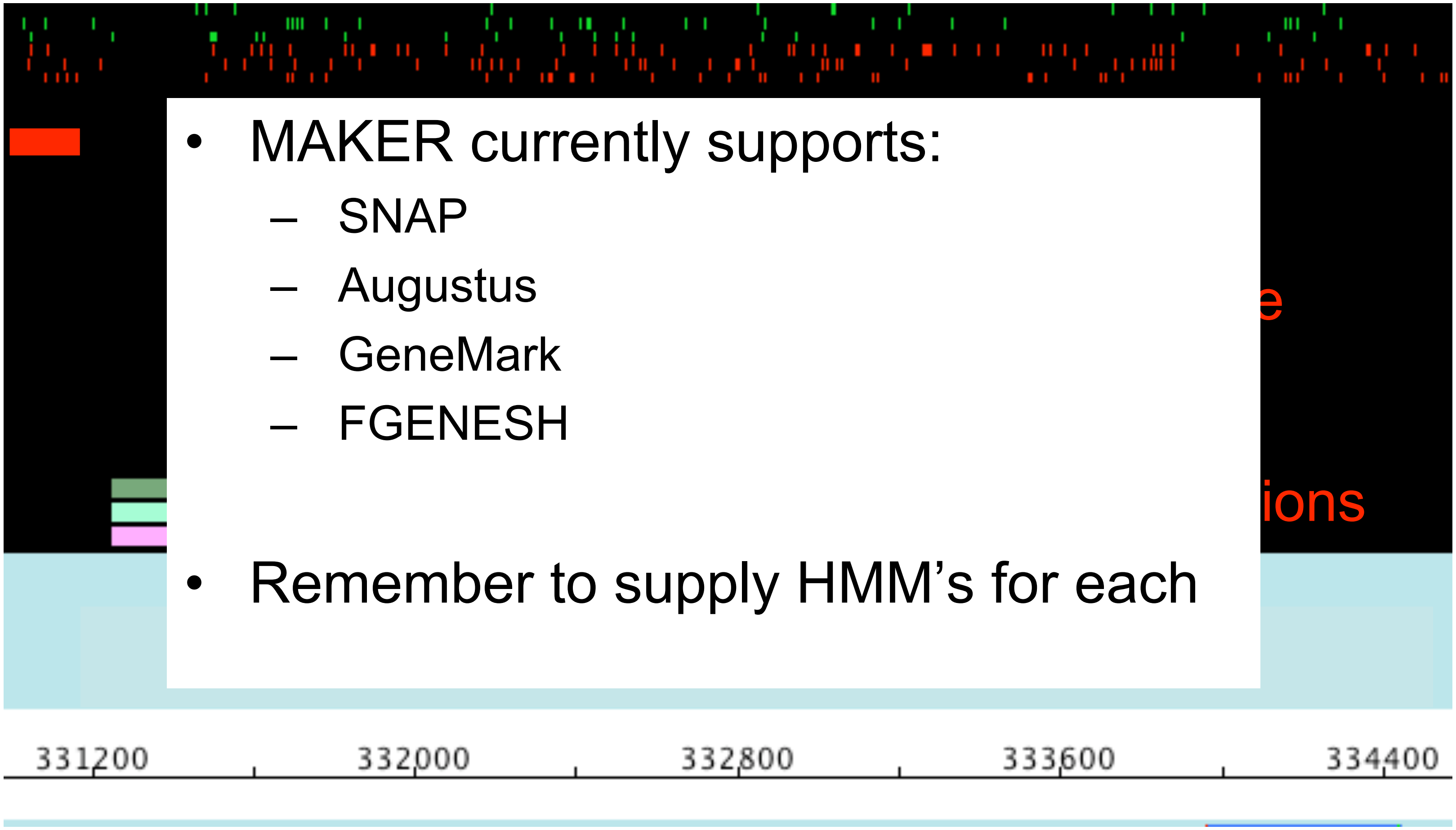
Current Assembly

# Generate *Ab Initio* Gene Predictions



Current Assembly

# Generate *Ab Initio* Gene Predictions

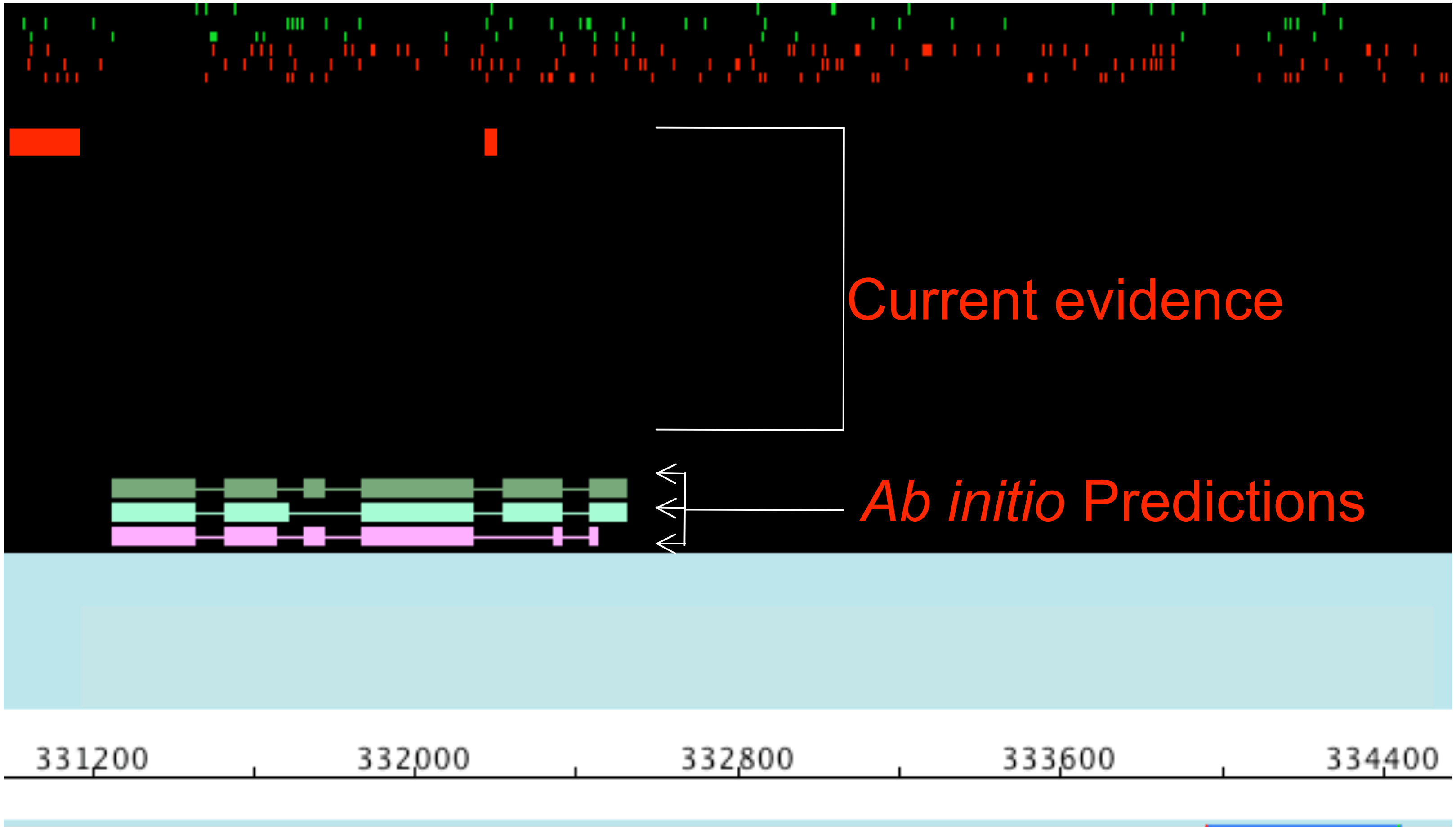


- MAKER currently supports:
  - SNAP
  - Augustus
  - GeneMark
  - FGENESH

- Remember to supply HMM’s for each

Current Assembly

# Generate *Ab Initio* Gene Predictions

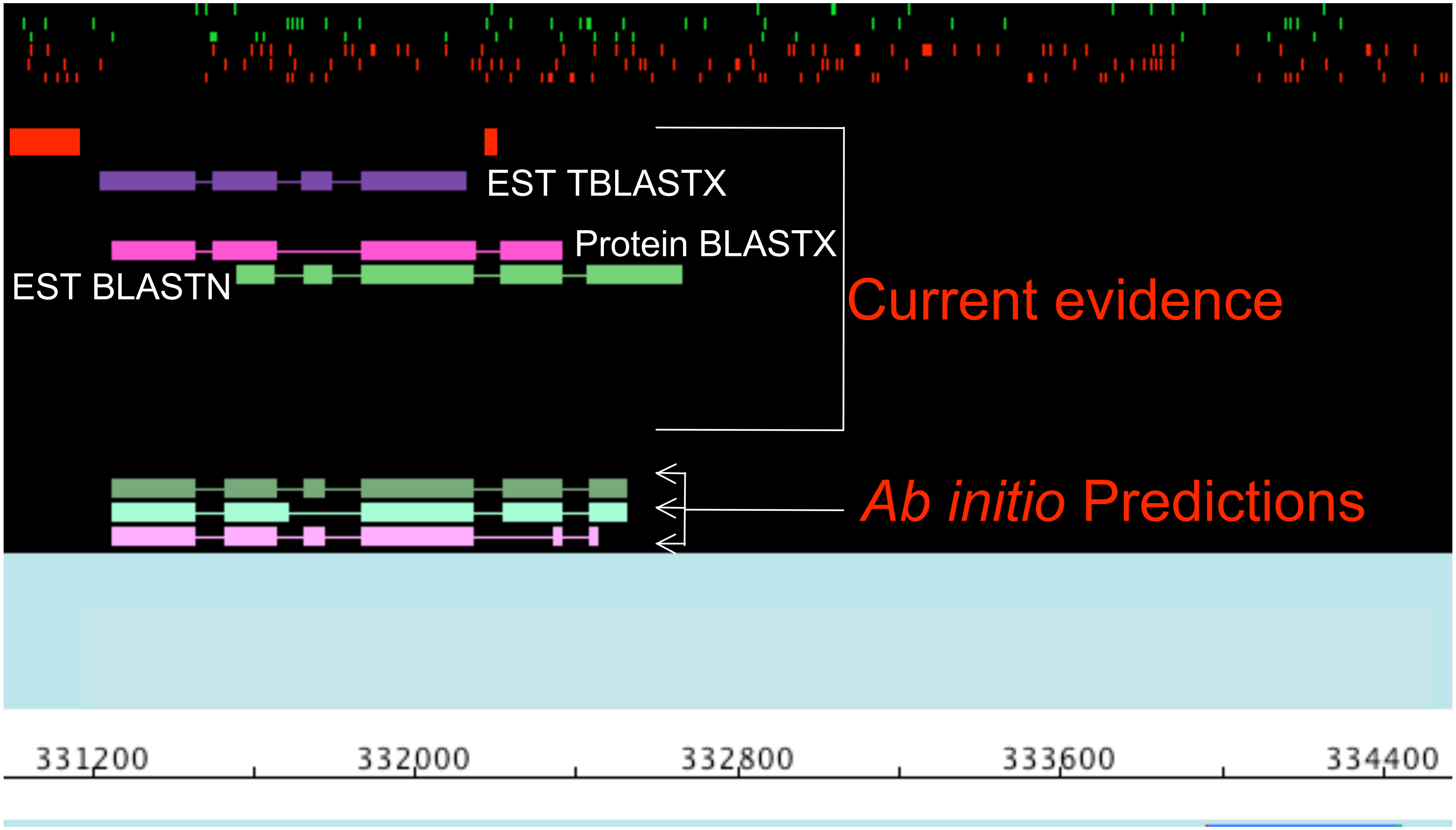


Current evidence

*Ab initio* Predictions

Current Assembly

# Align EST and Protein Evidence

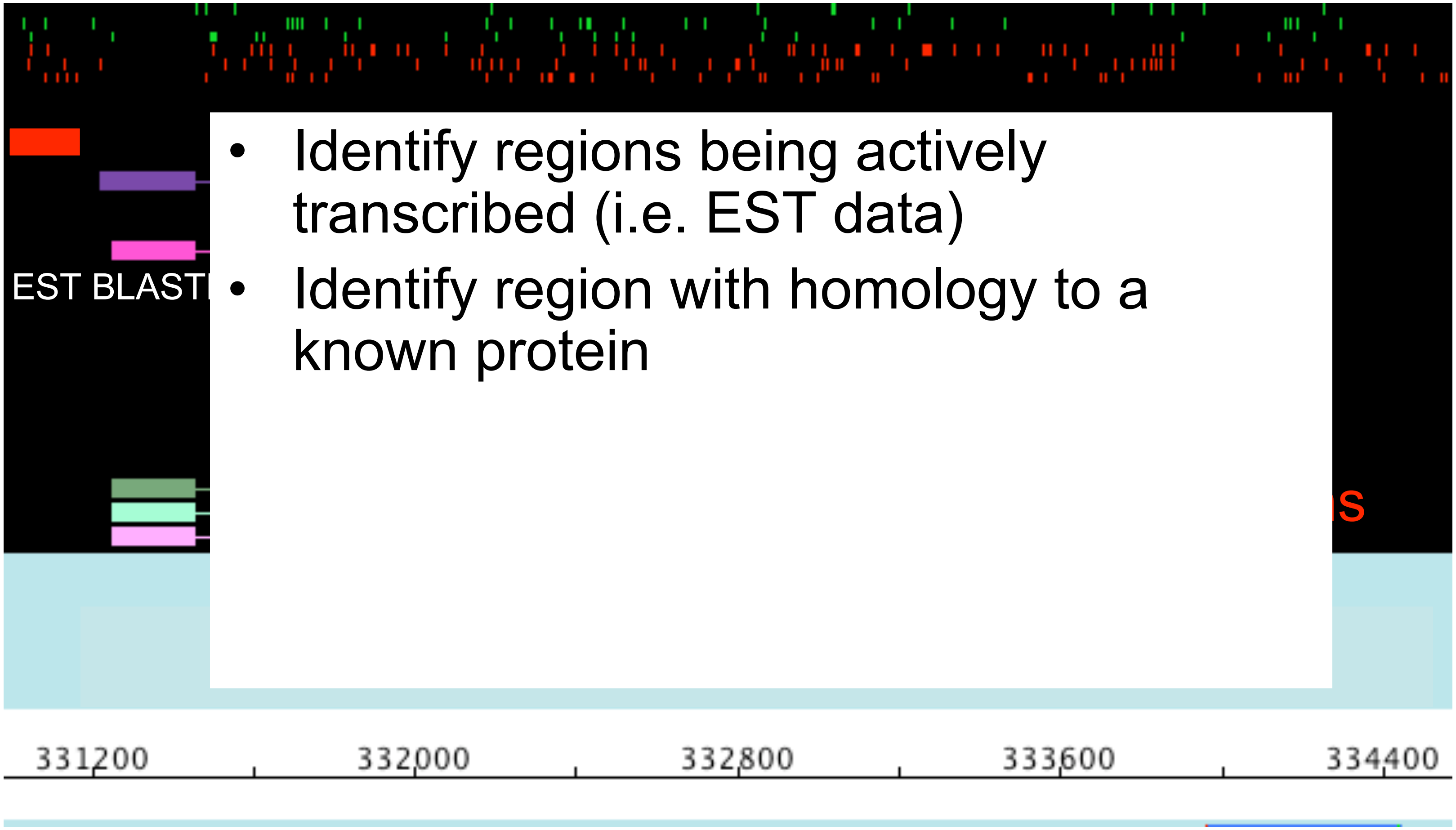


Current evidence

*Ab initio* Predictions

Current Assembly

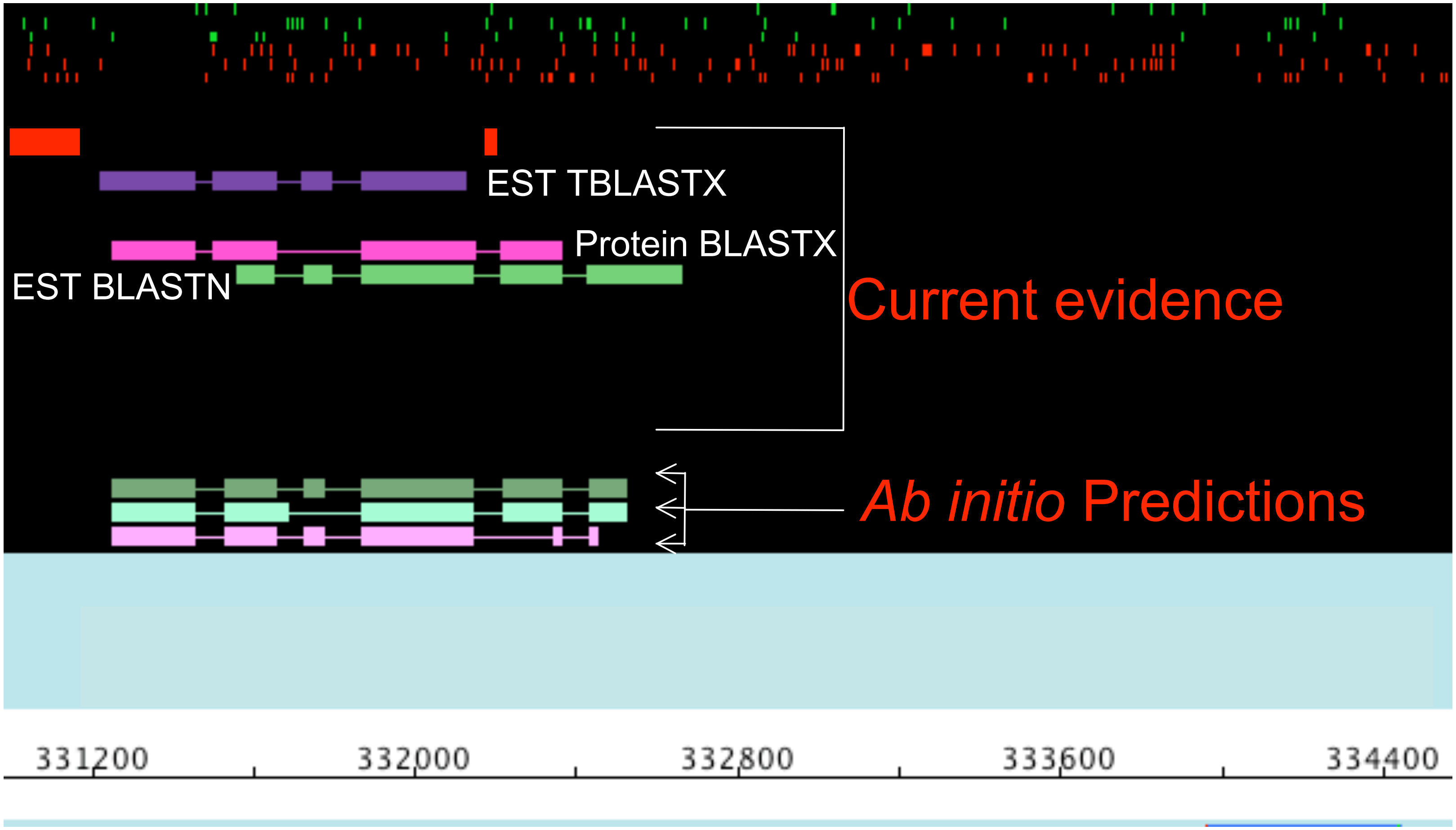
# Align EST and Protein Evidence



- Identify regions being actively transcribed (i.e. EST data)
- Identify region with homology to a known protein

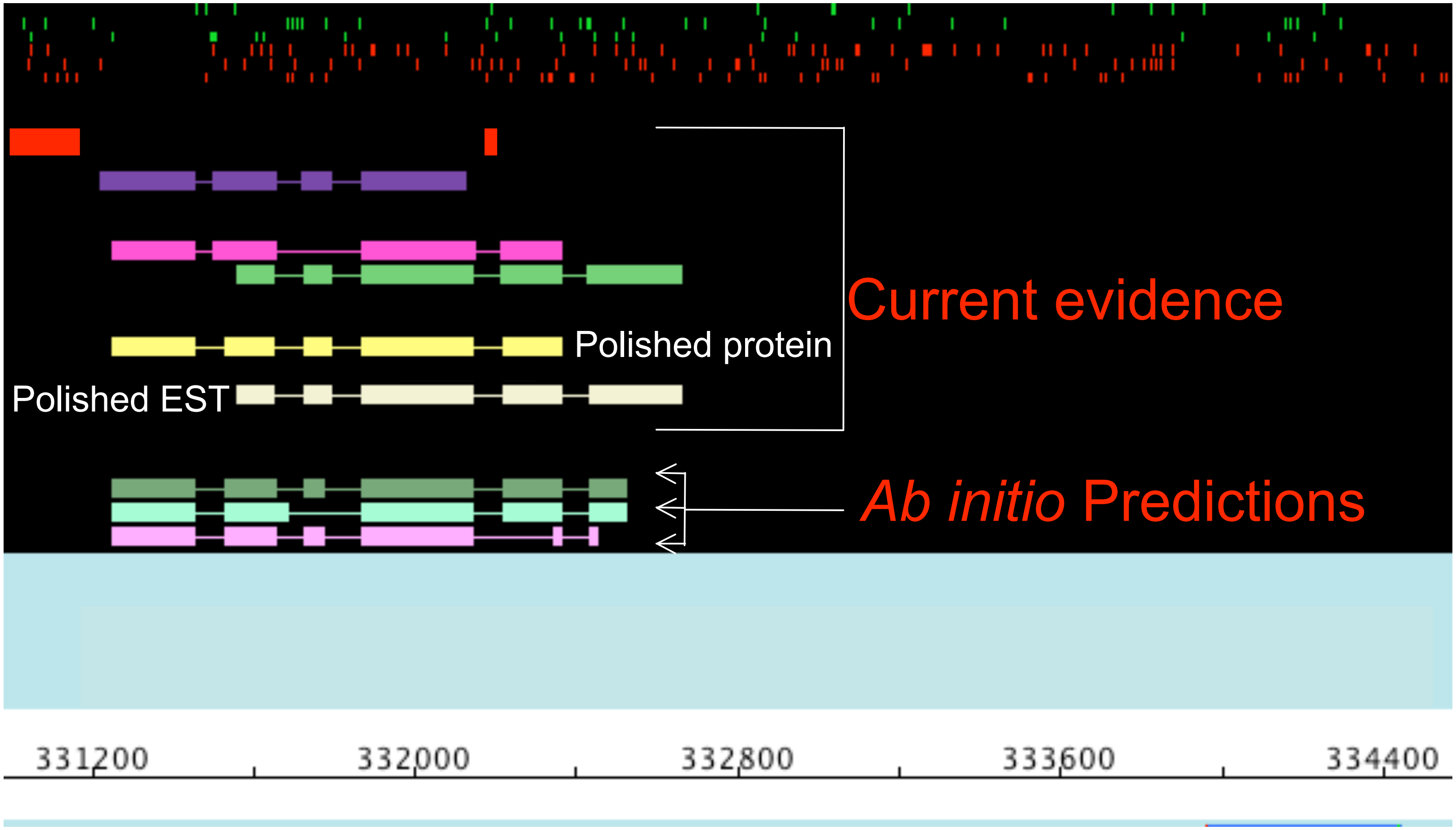
Current Assembly

# Align EST and Protein Evidence



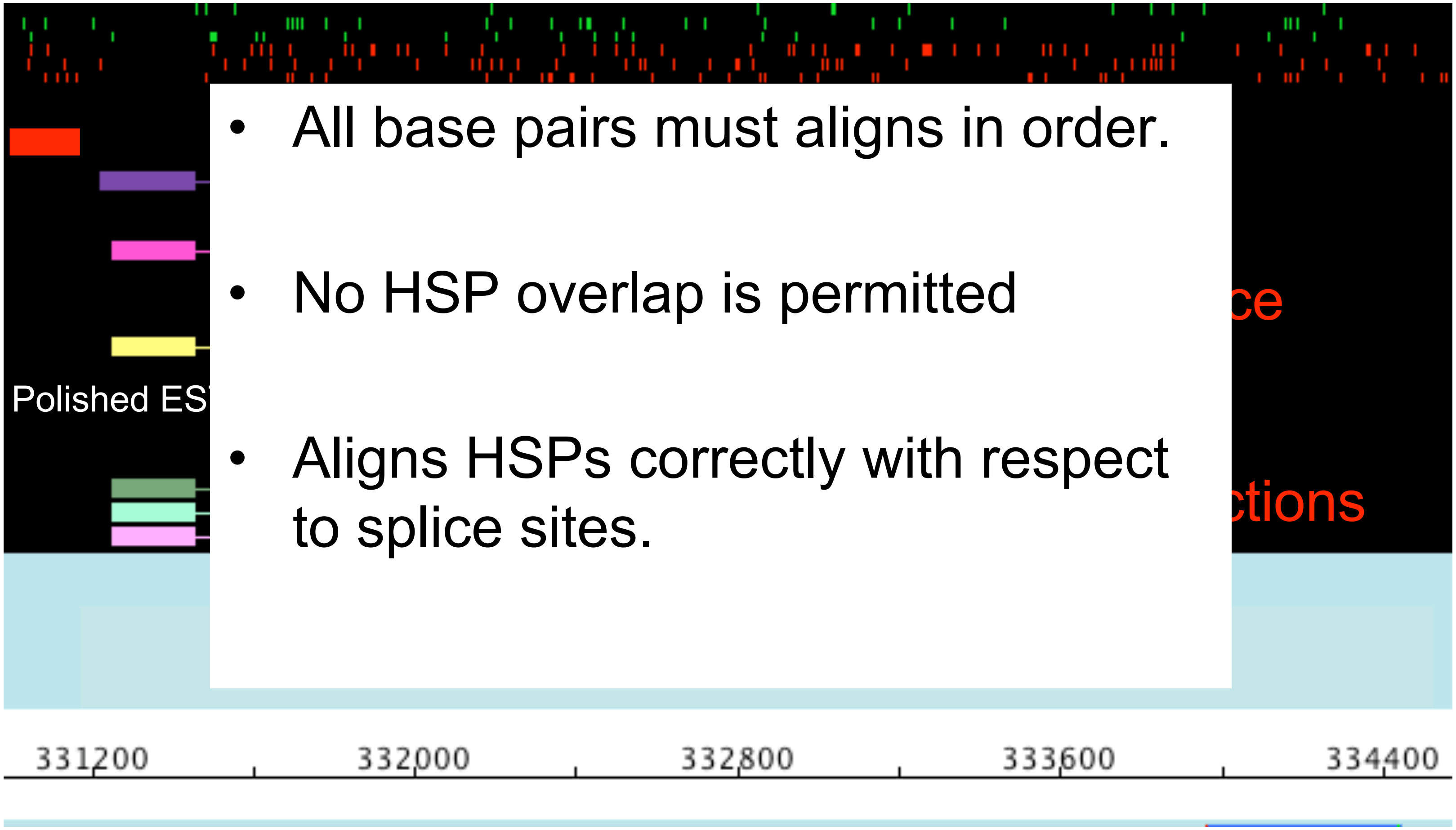
Current Assembly

# Polish BLAST Alignments with Exonerate



Current Assembly

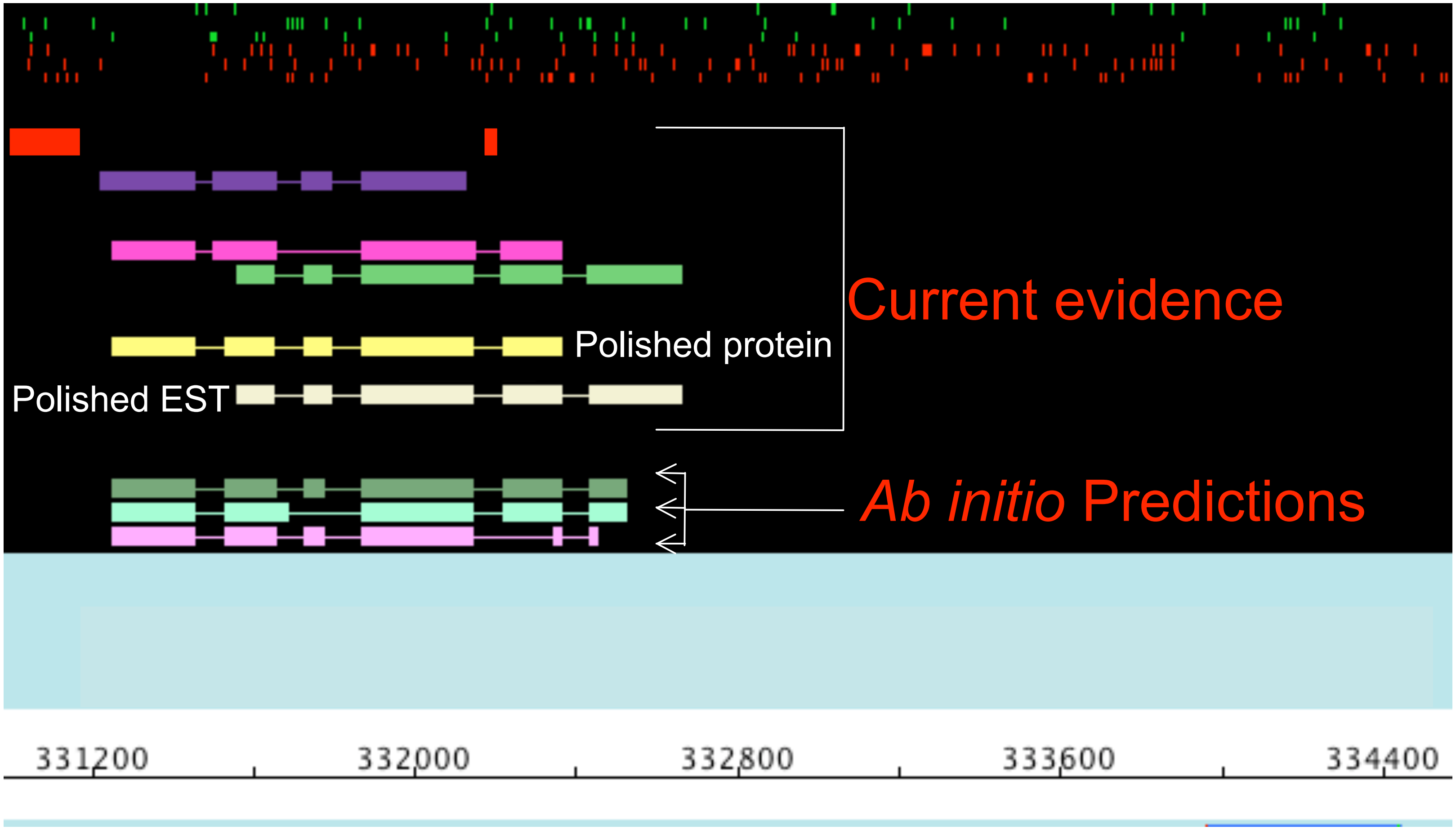
# Polish BLAST Alignments with Exonerate



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

Current Assembly

# Polish BLAST Alignments with Exonerate



Polished protein

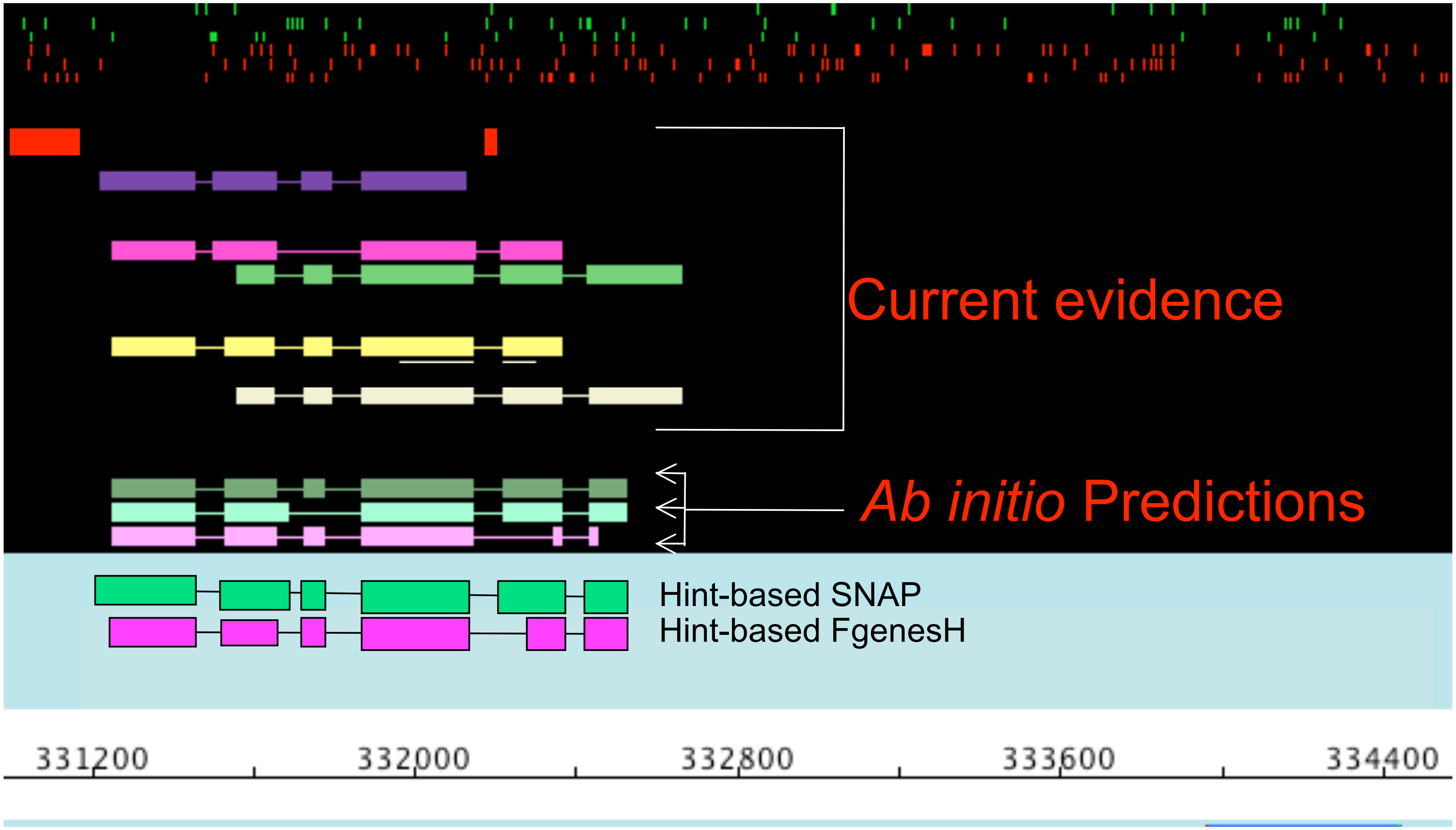
Polished EST

Current evidence

*Ab initio* Predictions

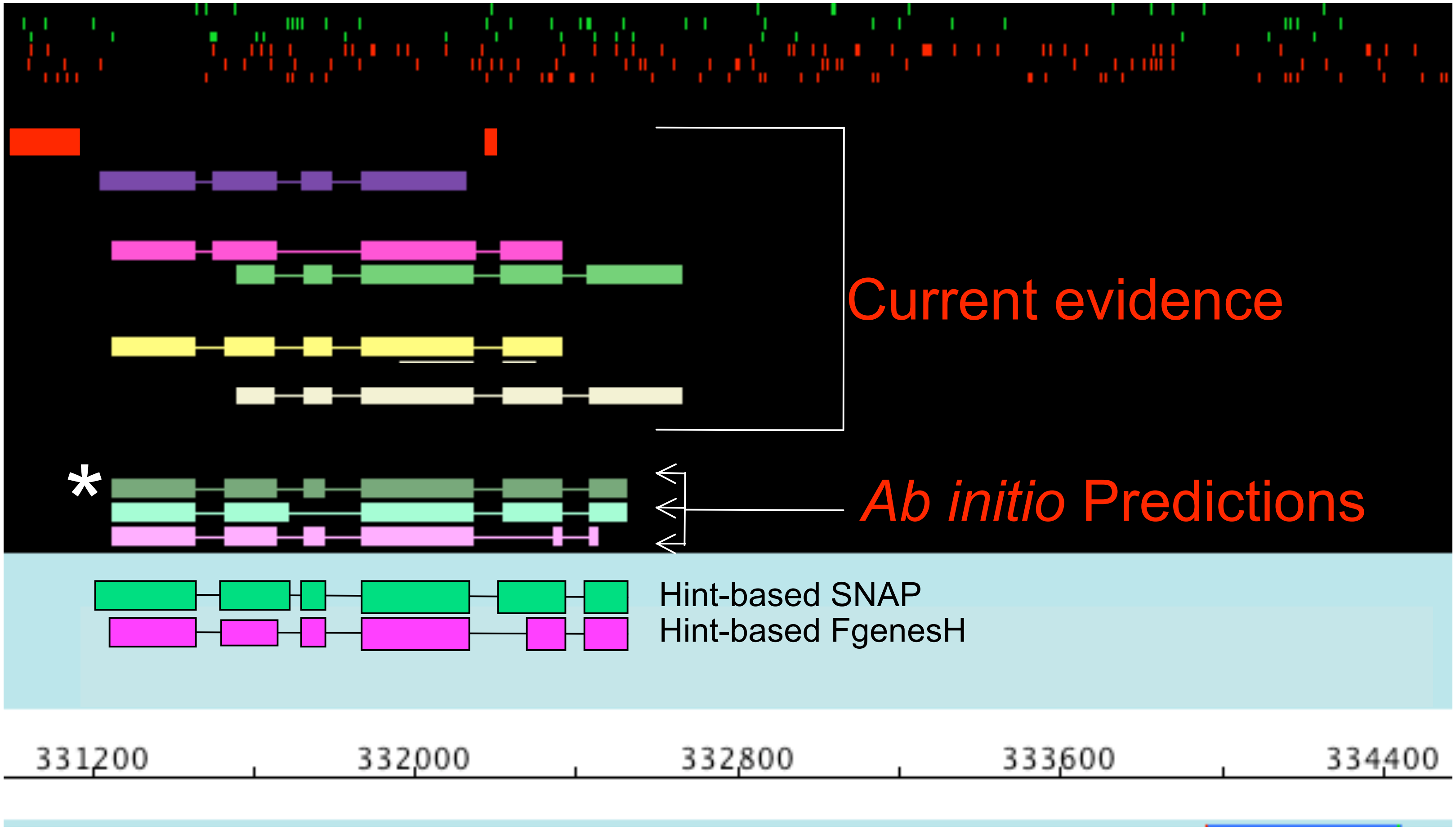
Current Assembly

# Pass Gene Finders Evidence-based ‘hints’



Current Assembly

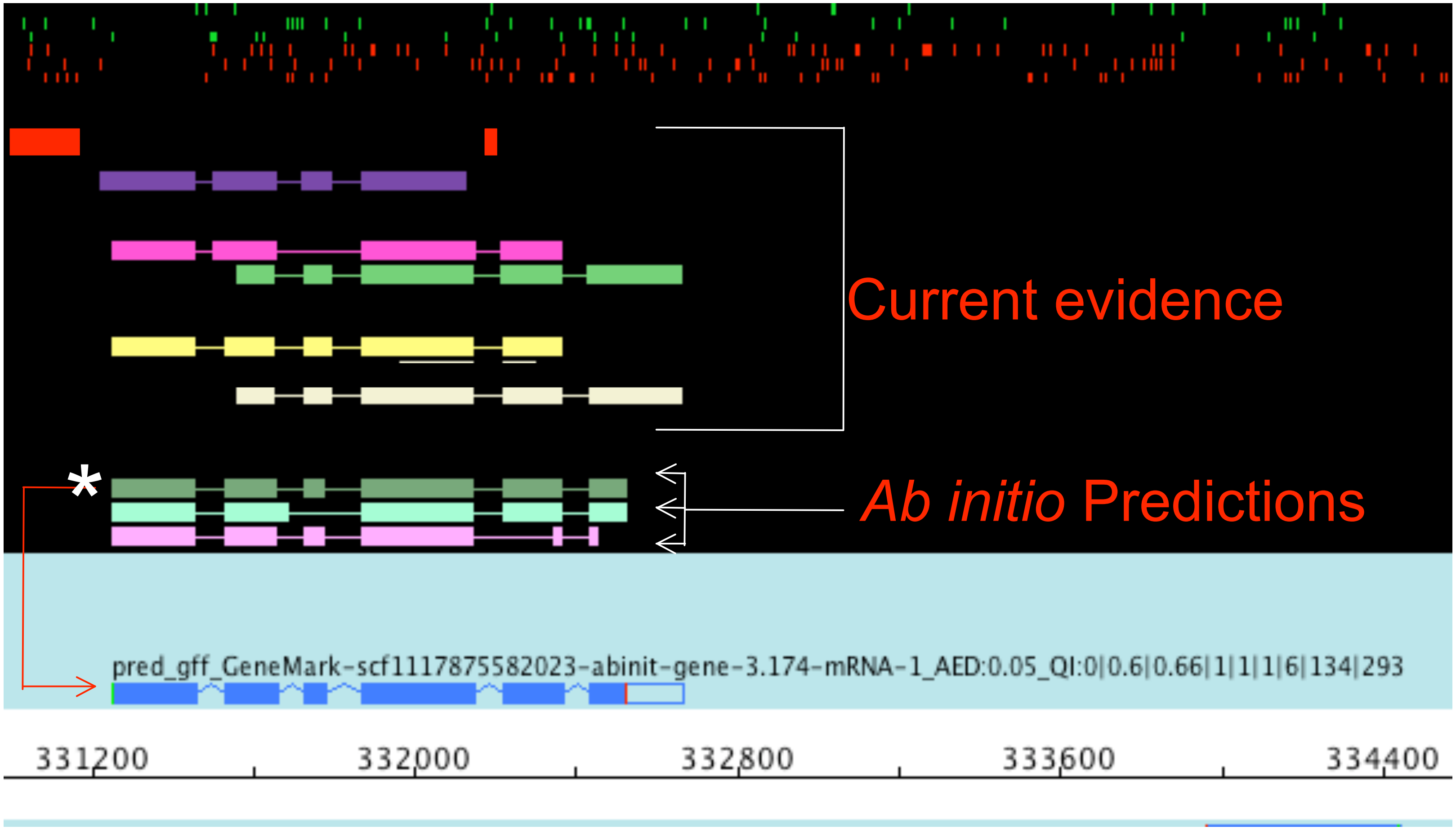
# Identify Gene Model Most Consistent with Evidence*



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


Current evidence
*
*Ab initio* Predictions
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 Assembly

# Compute Support for Each Portion of Gene Model

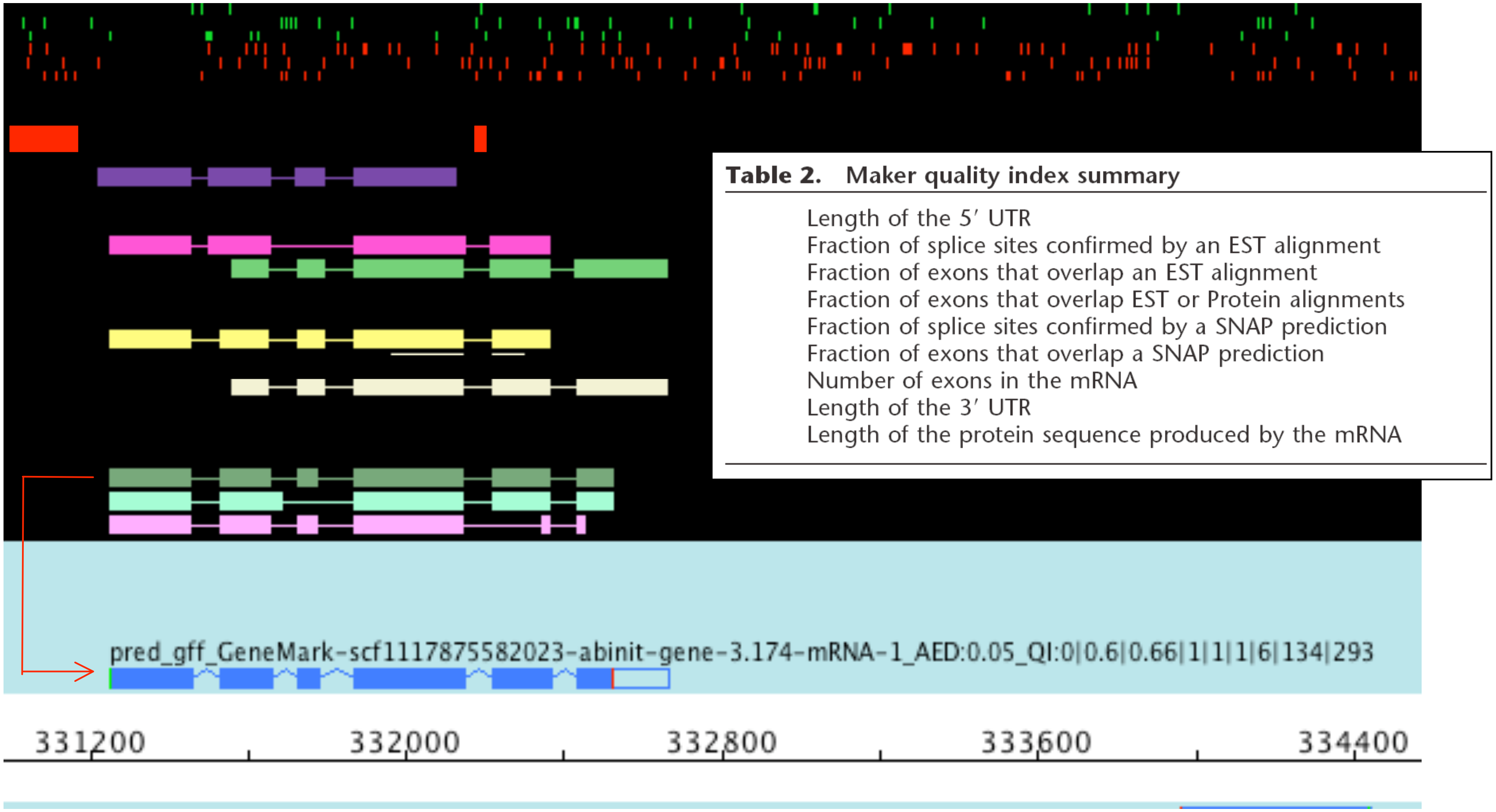
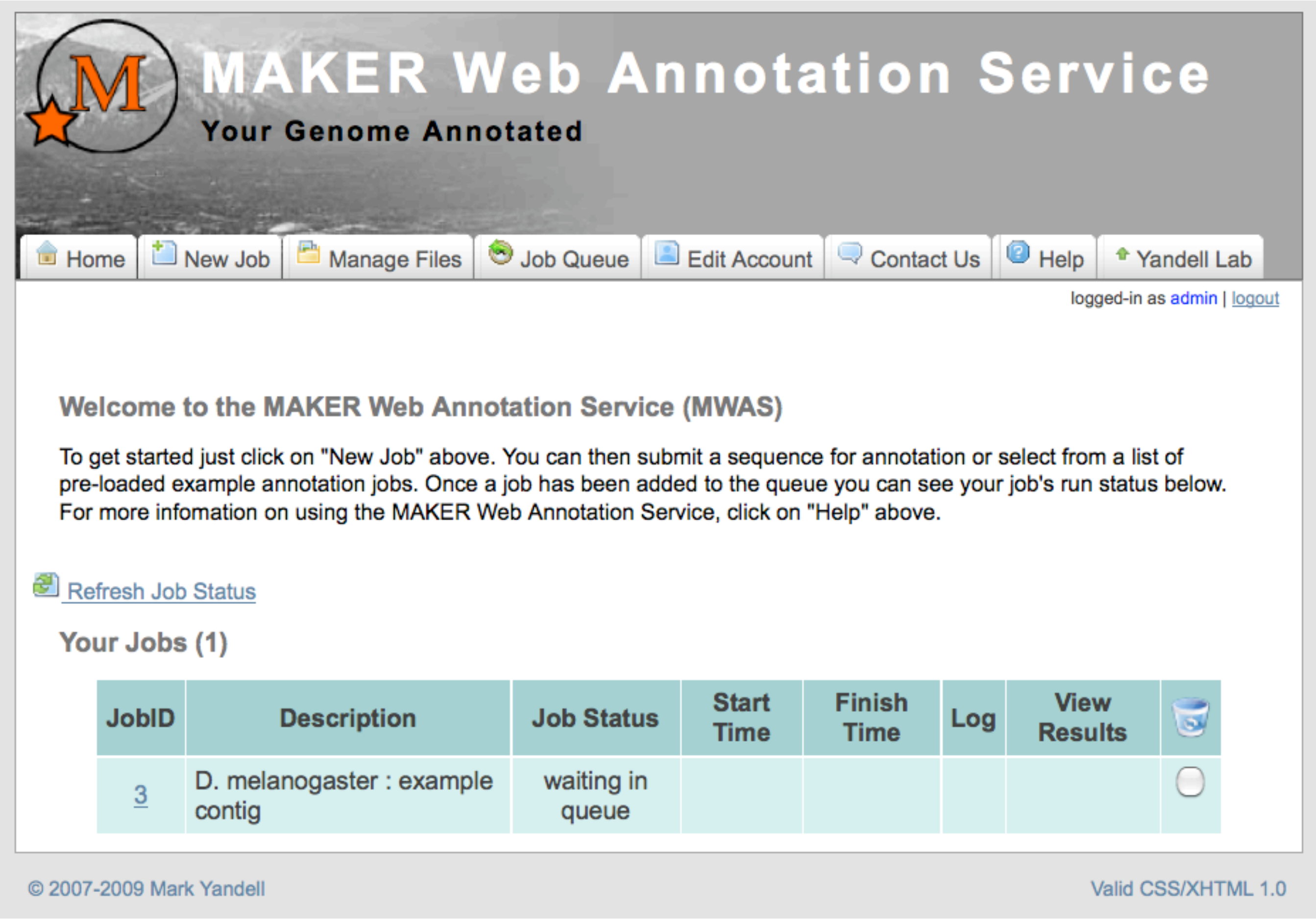

pred_gff_GeneMark-scf1117875582023-abinit-gene-3.174-mRNA-1_AED:0.05_QI:0|0.6|0.66|1|1|1|6|134|293
331200
332000
332800
333600
334400

**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

# Using MAKER

# MAKER Web Annotation Service



Welcome to the MAKER Web Annotation Service (MWAS)

To get started just click on "New Job" above. You can then submit a sequence for annotation or select from a list of pre-loaded example annotation jobs. Once a job has been added to the queue you can see your job's run status below. For more infomation on using the MAKER Web Annotation Service, click on "Help" above.

Refresh Job Status

Your Jobs (1)

| JobID | Description | Job Status | Start Time | Finish Time | Log | View Results | |
|---|---|---|---|---|---|---|---|
| 3 | D. melanogaster : example contig | waiting in queue | | | | | ○ |

© 2007-2009 Mark Yandell

Valid CSS/XHTML 1.0

# MAKER Web Annotation Service

logged-in as admin | logout

Welcome to the

To get started just c … t from a list of
pre-loaded example … s run status below.
For more infomatio

Refresh Job Status

Your Jobs (1)

| JobID | | View Results | |
|---|---|---|---|
| 3 | D. m<br>conti | | ○ |

© 2007-2009 Mark Yandell

Valid CSS/XHTML 1.0

http://www.yandell-lab.org

# *De novo* Annotation of a Newly Sequenced Genome

- You are involved in a genome project for an emerging model organism.
- You have no pre-existing gene models.
- What you do have:
  - ESTs
  - Proteins from other species available from public databases

*Go to Web*

# GFF3 pass-through: How to use external evidence

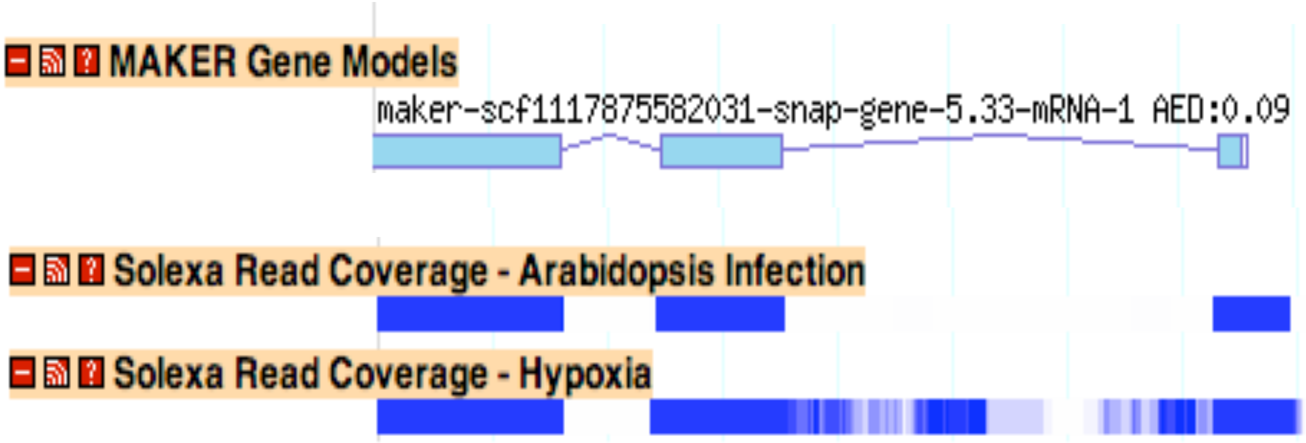
- You have an existing annotation set.
- You want to update the evidence and allow the annotation to change to reflect the new evidence.

# What if I have mRNA-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

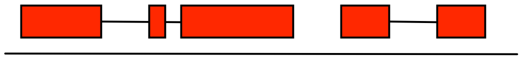
- Use BowTie and TopHat to produce, aligns reads into expression “islands” and “junctions”
- Pass data through as EST evidence via GFF3 pass-through.

*Go to Web*

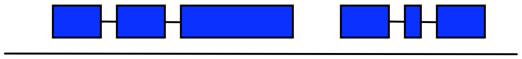
# Another issue: legacy annotations

- 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 mRNA-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

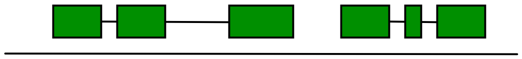
# Merging and Revising Legacy Annotation Sets



Legacy Annotation Set 1

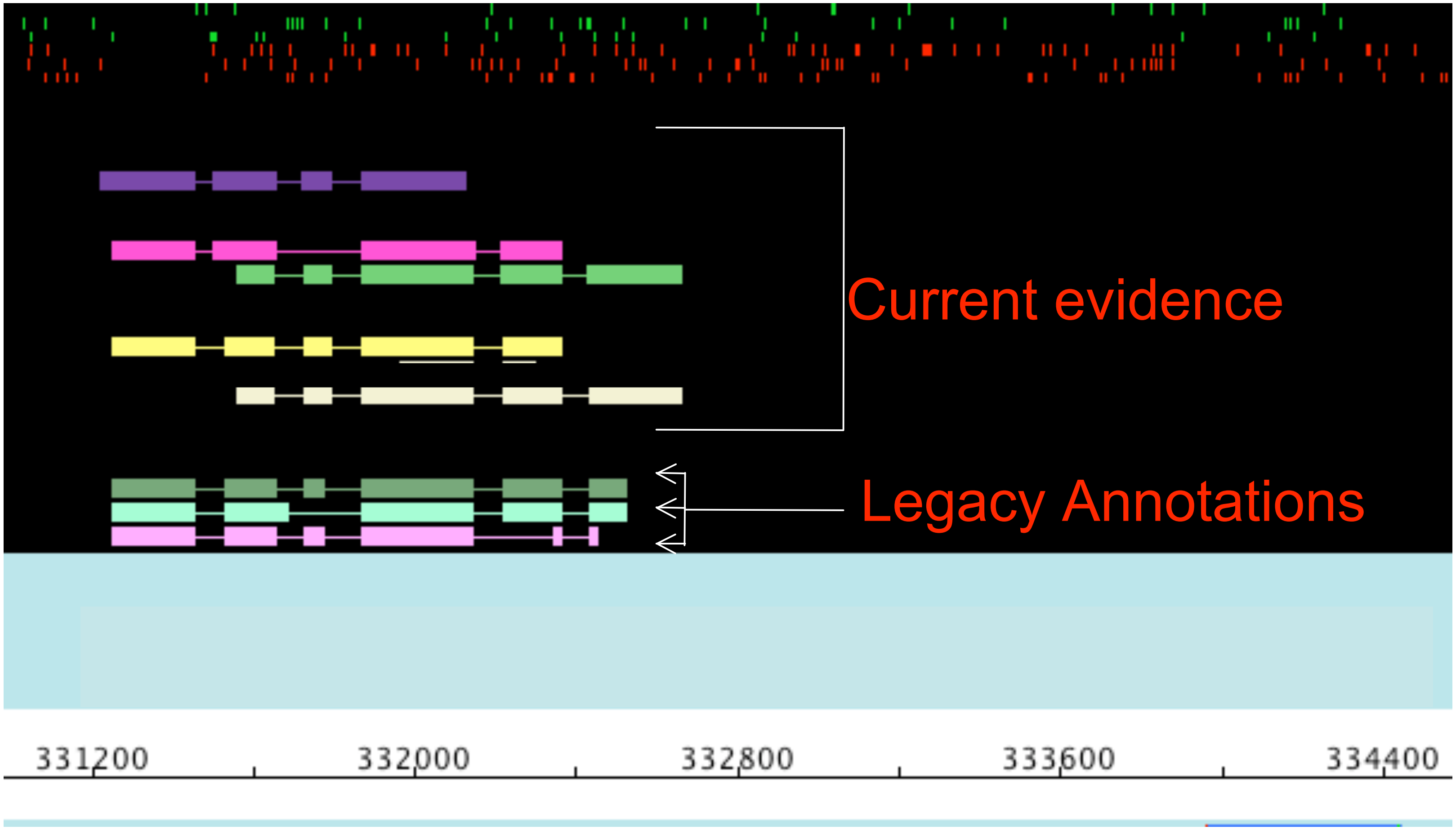


Legacy Annotation Set 2



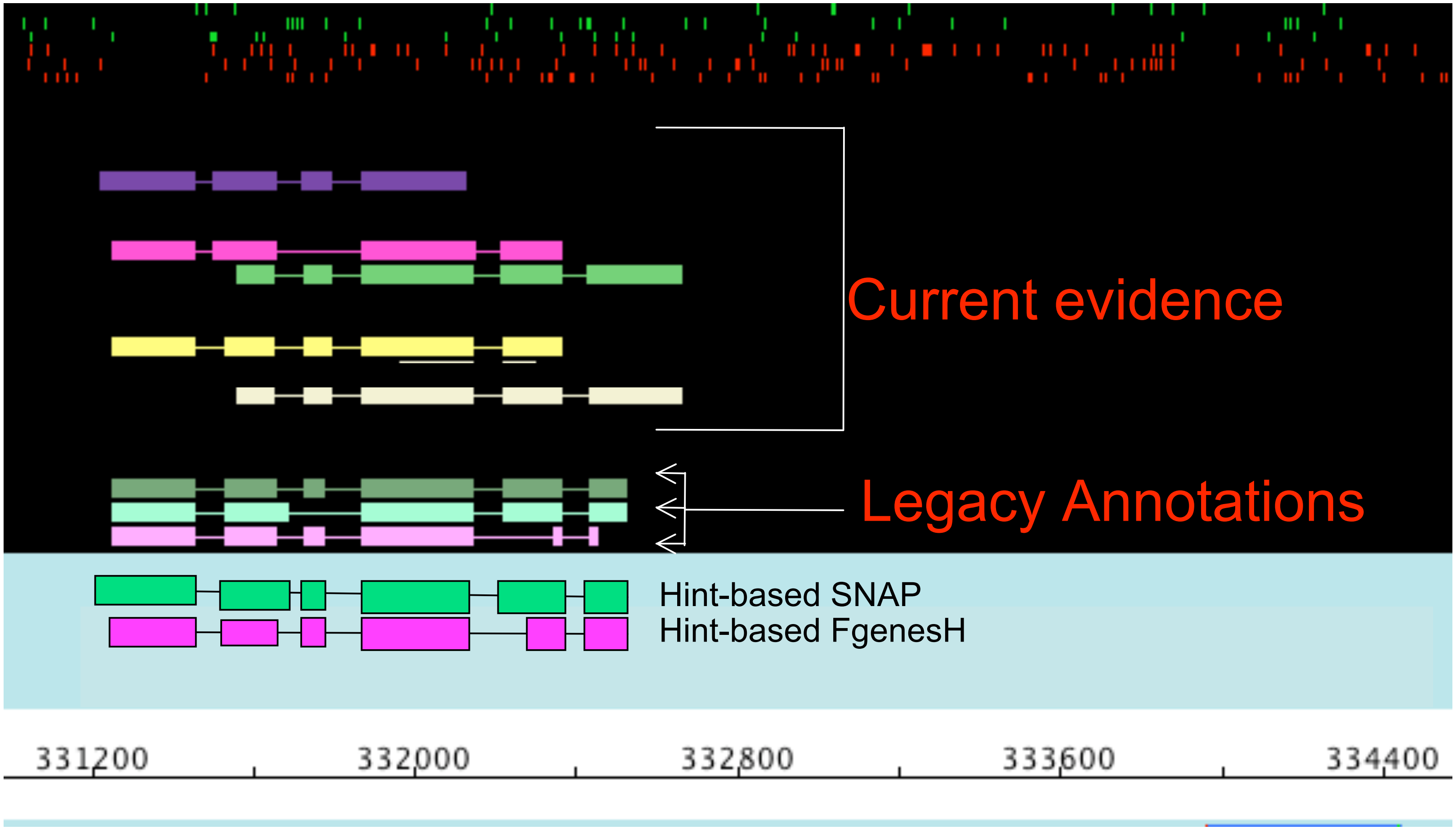
Legacy Annotation Set n

# Align Evidence and Legacy Annotations to Current Assembly



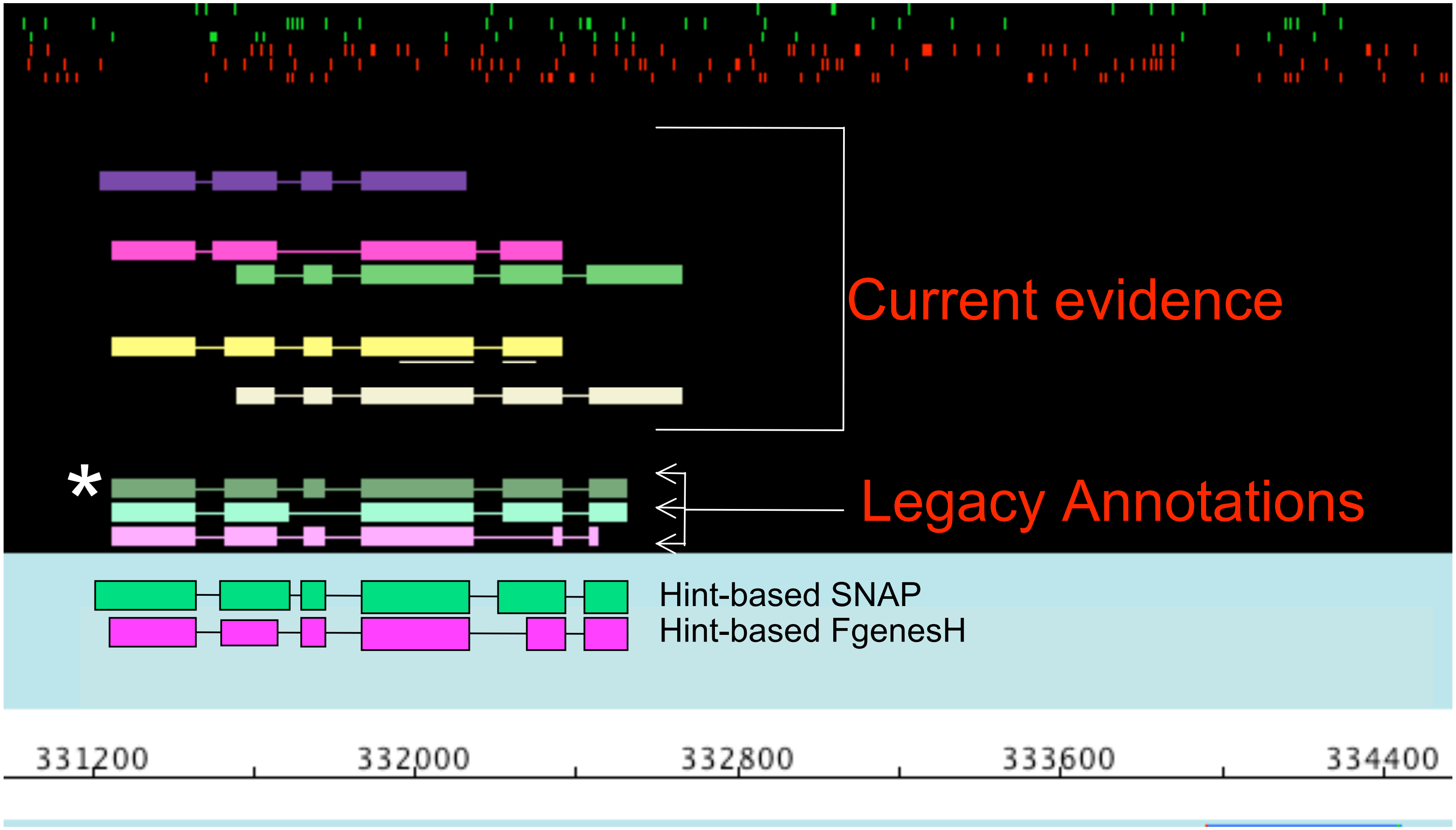
Current Assembly

# Pass Gene Finders Evidence-based ‘hints’



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

*Go to Web*

# Working with Chado

- maker2chado [OPTION] <database_name> <gff3file1> <gff3file2> ...
- maker2chado [OPTION] -d <datastore_index> <database_name>

This script takes MAKER produced GFF3 files and dumps them into a CHADO database. You must set the database up first according to CHADO installation instructions. CHADO provides its own methods for loading GFF3, but this script makes it easier for MAKER specific data. You can either provide the datastore index file produced by MAKER to the script or add the GFF3 files as command line arguments.

# Working with JBrowse

- maker2jbrowse [OPTION] <gff3file1> <gff3file2> ...
- maker2jbrowse [OPTION] -d <datastore_index>

This script takes MAKER produced GFF3 files and dumps them into JBrowse for you using pre-configured JSON tracks.