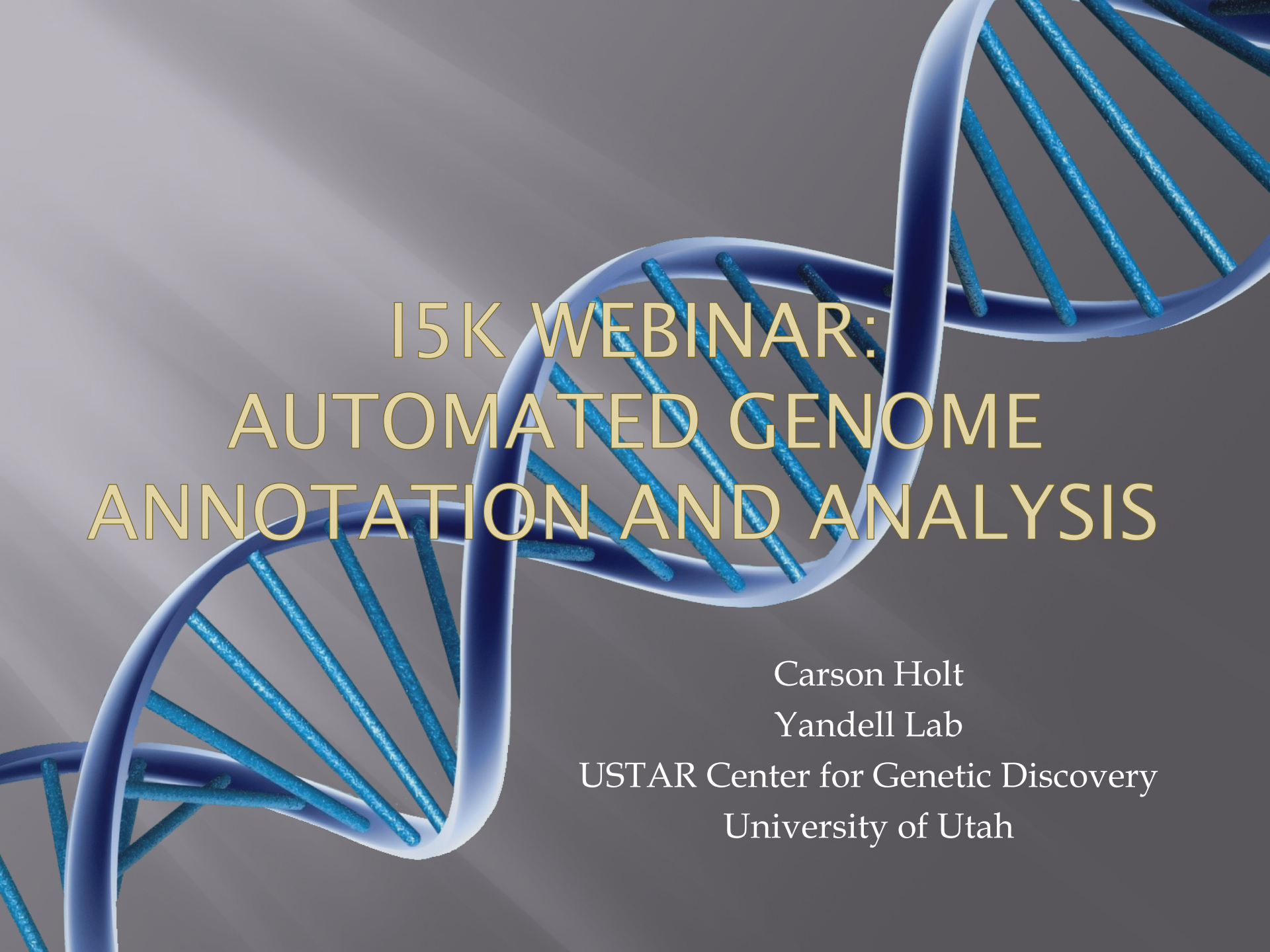
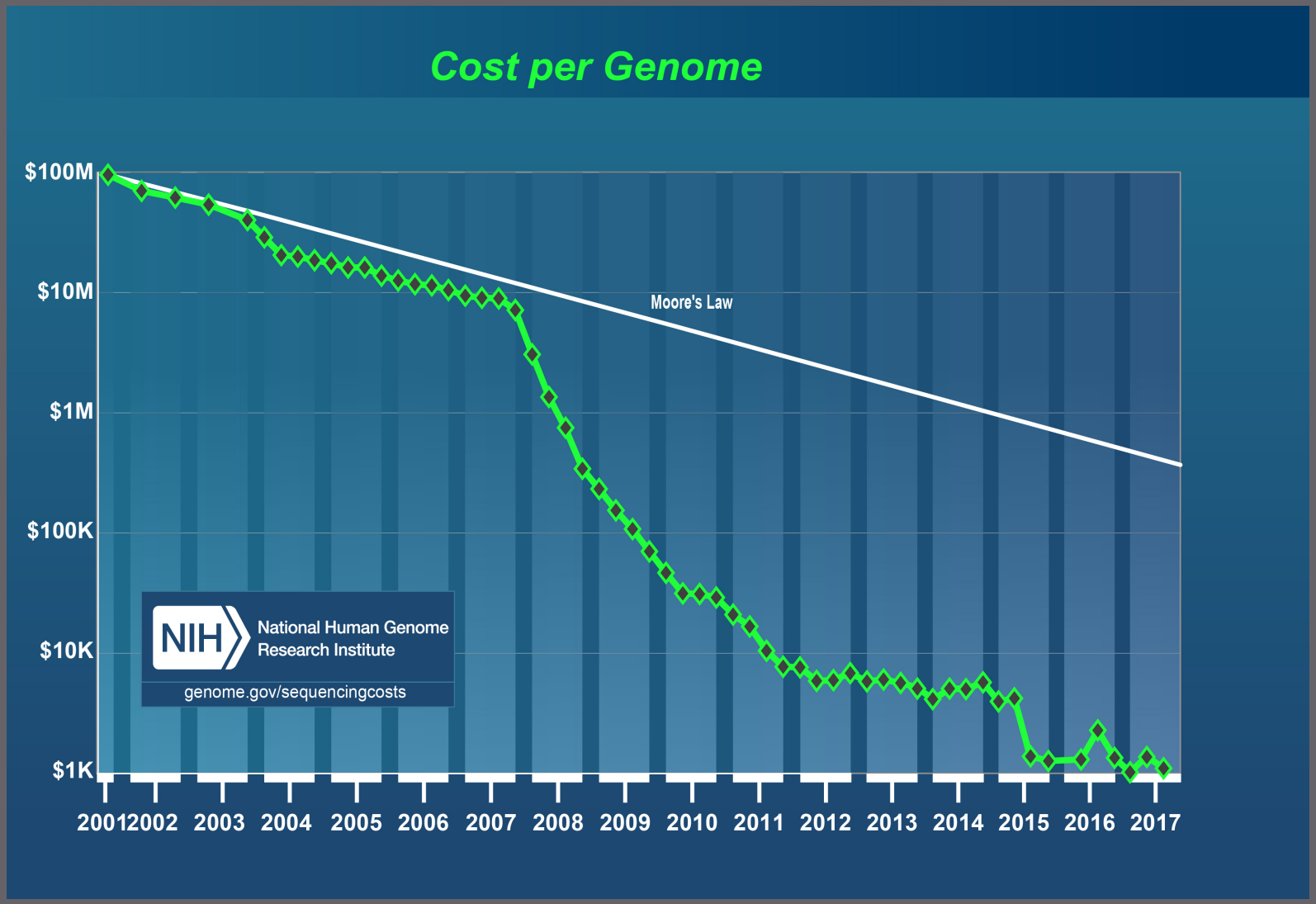


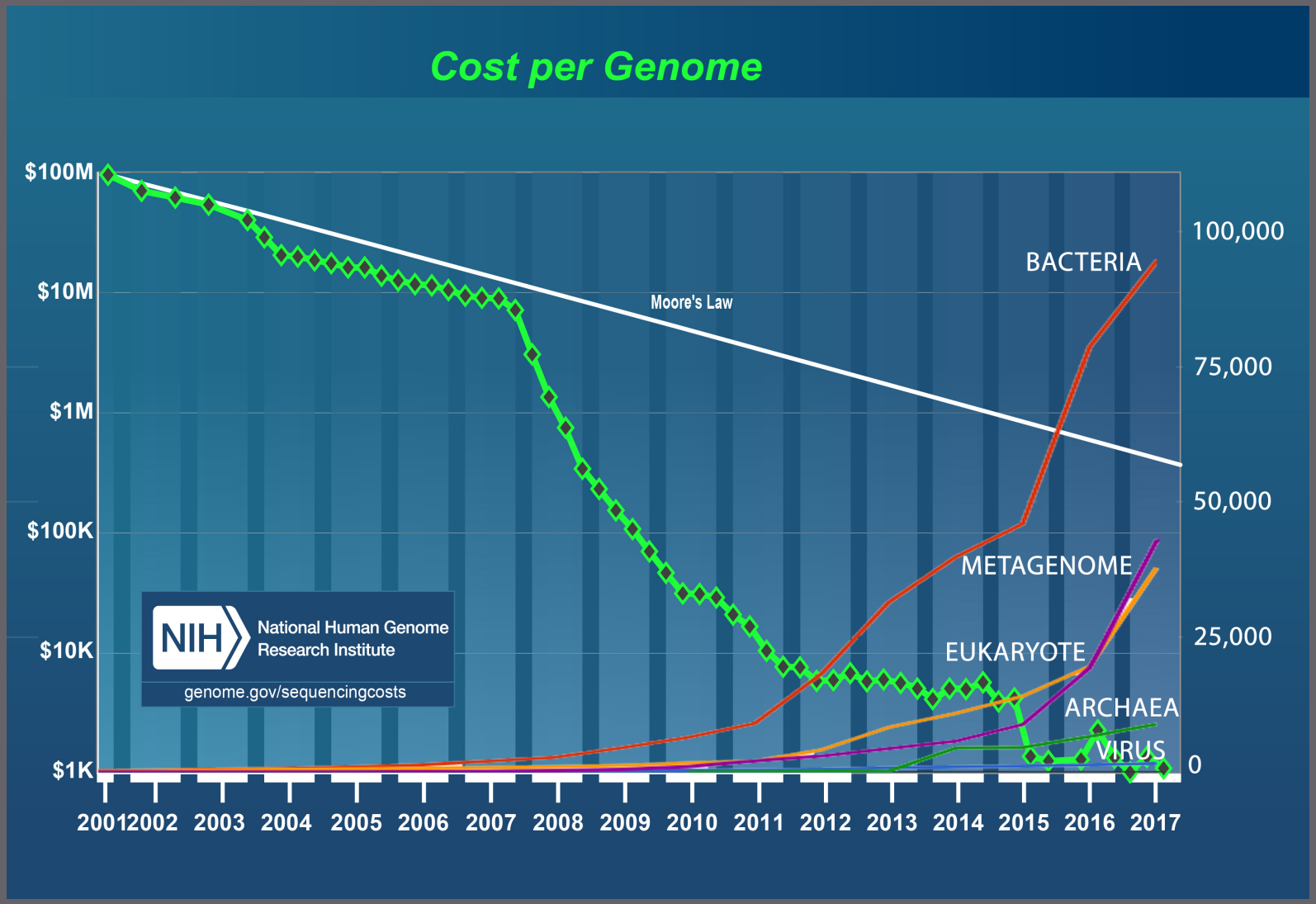
# I5K WEBINAR: AUTOMATED GENOME ANNOTATION AND ANALYSIS

Carson Holt
Yandell Lab
USTAR Center for Genetic Discovery
University of Utah

# Advances in Second-Generation Technology are Making Whole Genome and Transcriptome Sequencing "Routine" Even for Small Labs

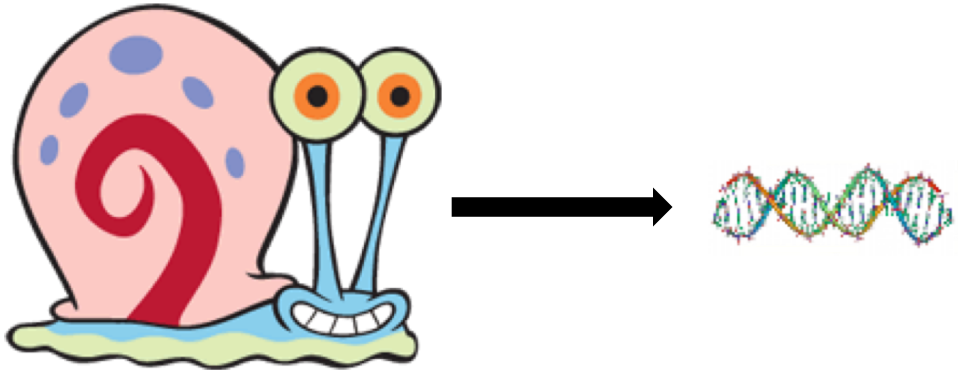


# Advances in Second-Generation Technology are Making Whole Genome and Transcriptome Sequencing "Routine" Even for Small Labs

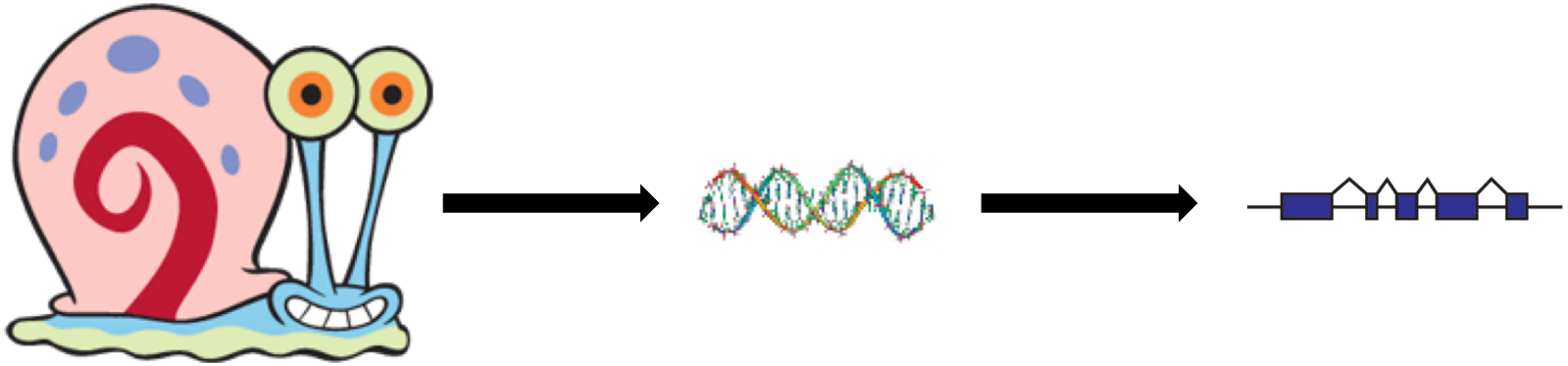


# Genome Project Overview

# Genome Project Overview



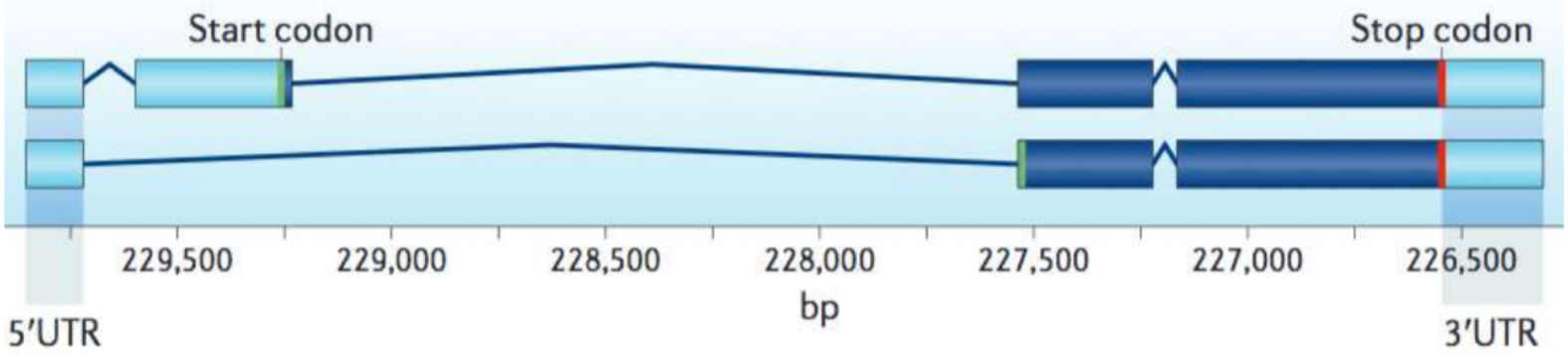
# Genome Project Overview



# Genome Project Overview

## What is an Annotation?

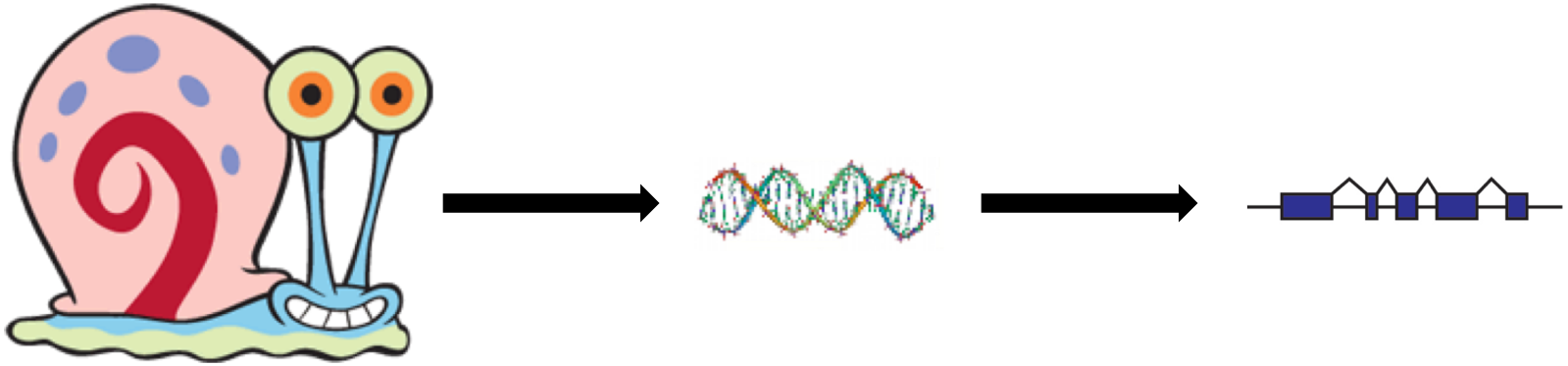
Structural



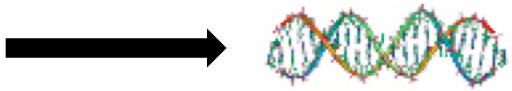

Functional

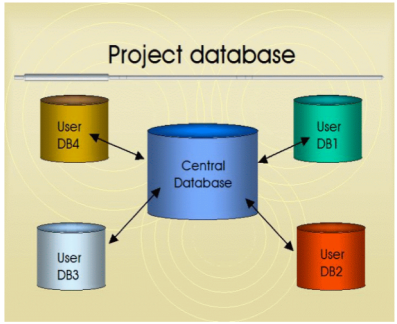
| | |
|---|---|
| **Function** | cAMP-dependent and sulfonylurea-sensitive anion transporter. Key gatekeeper influencing intracellular cholesterol transport. |
| **Subcellular location** | Membrane; Multi-pass membrane protein Ref.13 Ref.14. |
| **Domain** | Multifunctional polypeptide with two homologous halves, each containing a hydrophobic membrane-anchoring domain and an ATP binding cassette (ABC) domain. |

# Genome Project Overview

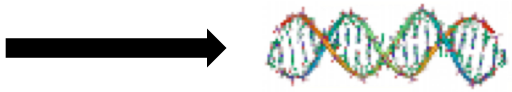


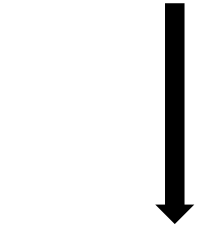
# Genome Project Overview



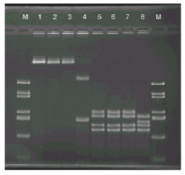


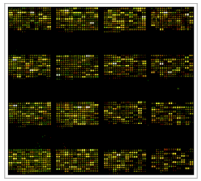
# Genome Project Overview

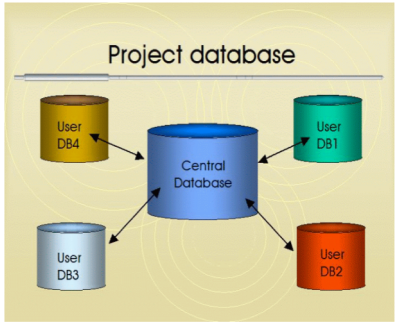


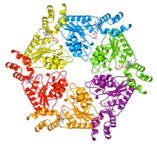


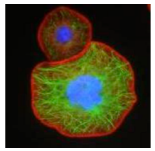
```
>Smg5
MEVTFSSGGSSNASSECAIDGGTNRCRGL
EPNNGTCILSQEVKDLYRSLYTASKQLDD
AKRNVQSVGQLFQHEIEEKRSLLVQLCKQ
IIFKDYQSVGKKVREVMWRRGYYEFIAFV
```



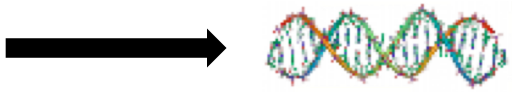




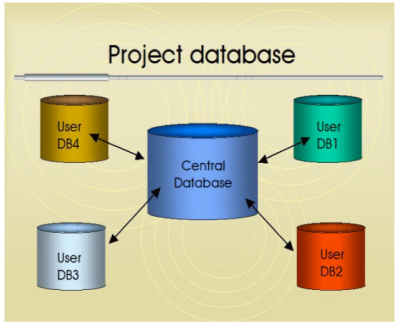



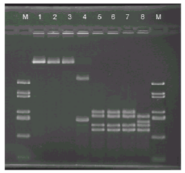


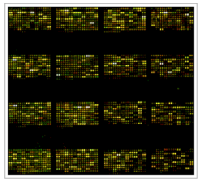
# Genome Project Overview

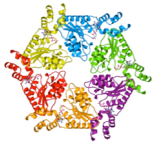


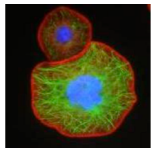
```
>Smg5
MEVTFSSGGSSNASSECAIDGGTNRCRGL
EPNNGTCILSQEVKDLYRSLYTASKQLDD
AKRNVQSVGQLFQHEIEEKRSLLVQLCKQ
IIFKDYQSVGKKVREVMWRRGYYEFIAFV
```









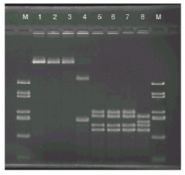
SUCCESS

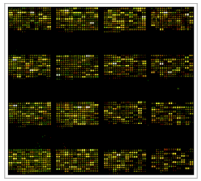
# Genome Project Overview

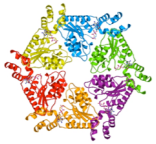
Incorrect annotations poison every experiment that uses them!!

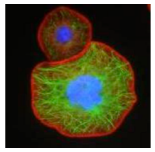
```
>Smg5
MEVTFSSGGSSNASSECAIDGGTNRCRGL
EPNNGTCILSQEVKDLYRSLYTASKQLDD
AKRNVQSVGQLFQHEIEEKRSLLVQLCKQ
IIFKDYQSVGKKVREVMWRRGYYEFIAFV
```

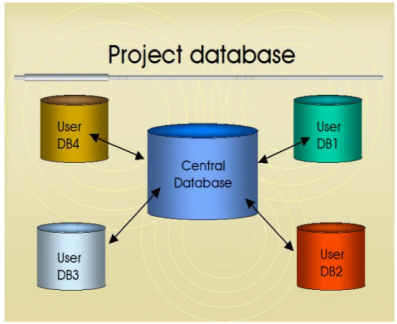
FAILURE












# MAKER

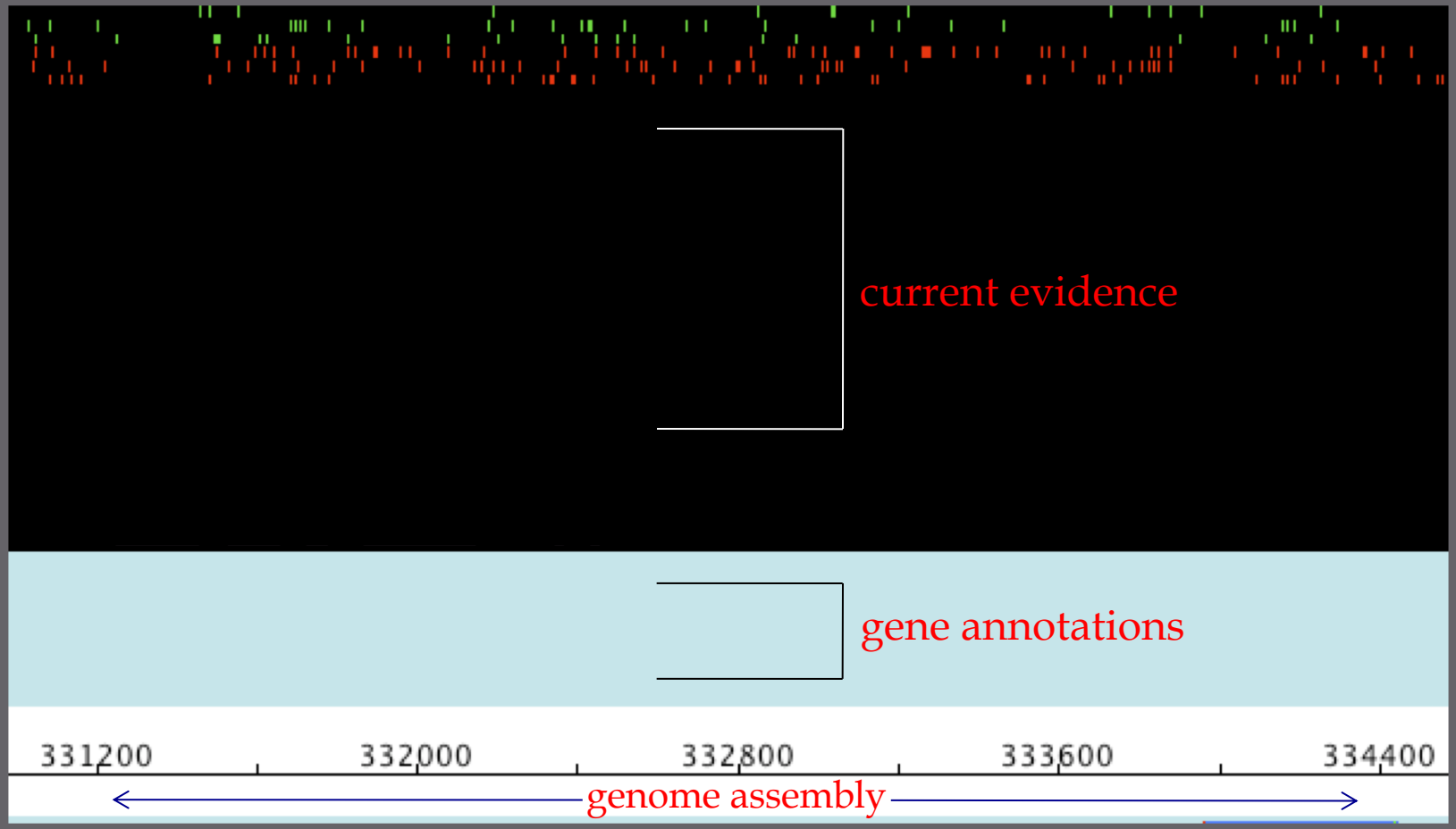
an annotation pipeline and genome-database management tool for second-generation genome projects

# Easy-to-use by design

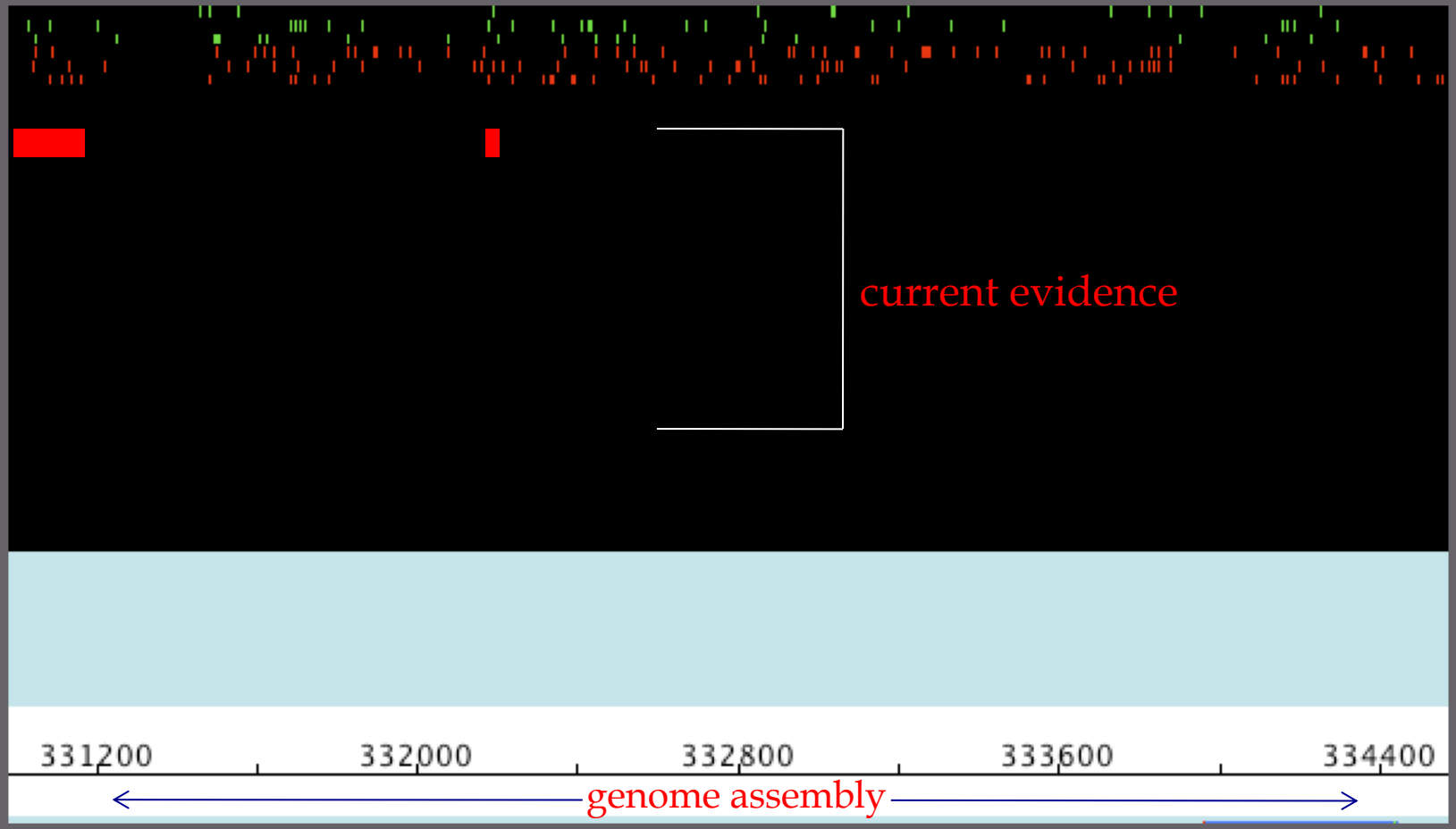
| | |
|---|---|
| **User Requirements:** | Can be run by a single individual with little bioinformatics experience |
| **System Requirements:** | Can run on laptop or desktop computers (running Linux or Mac OS X) |
| **Program Output:** | Output is compatible with popular annotation tools like Apollo,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

# How does MAKER work?

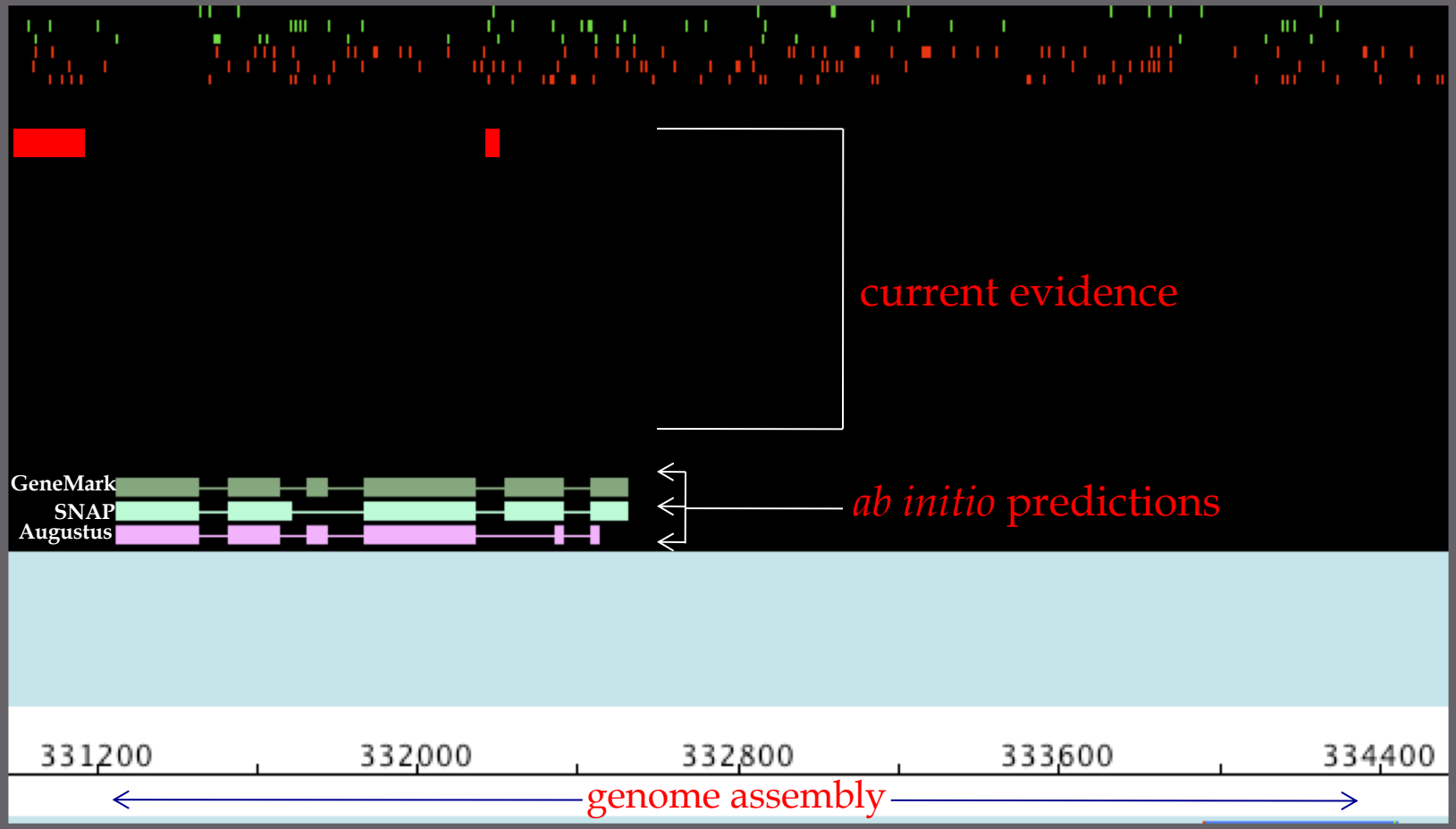
# Annotating the Genome – Apollo View



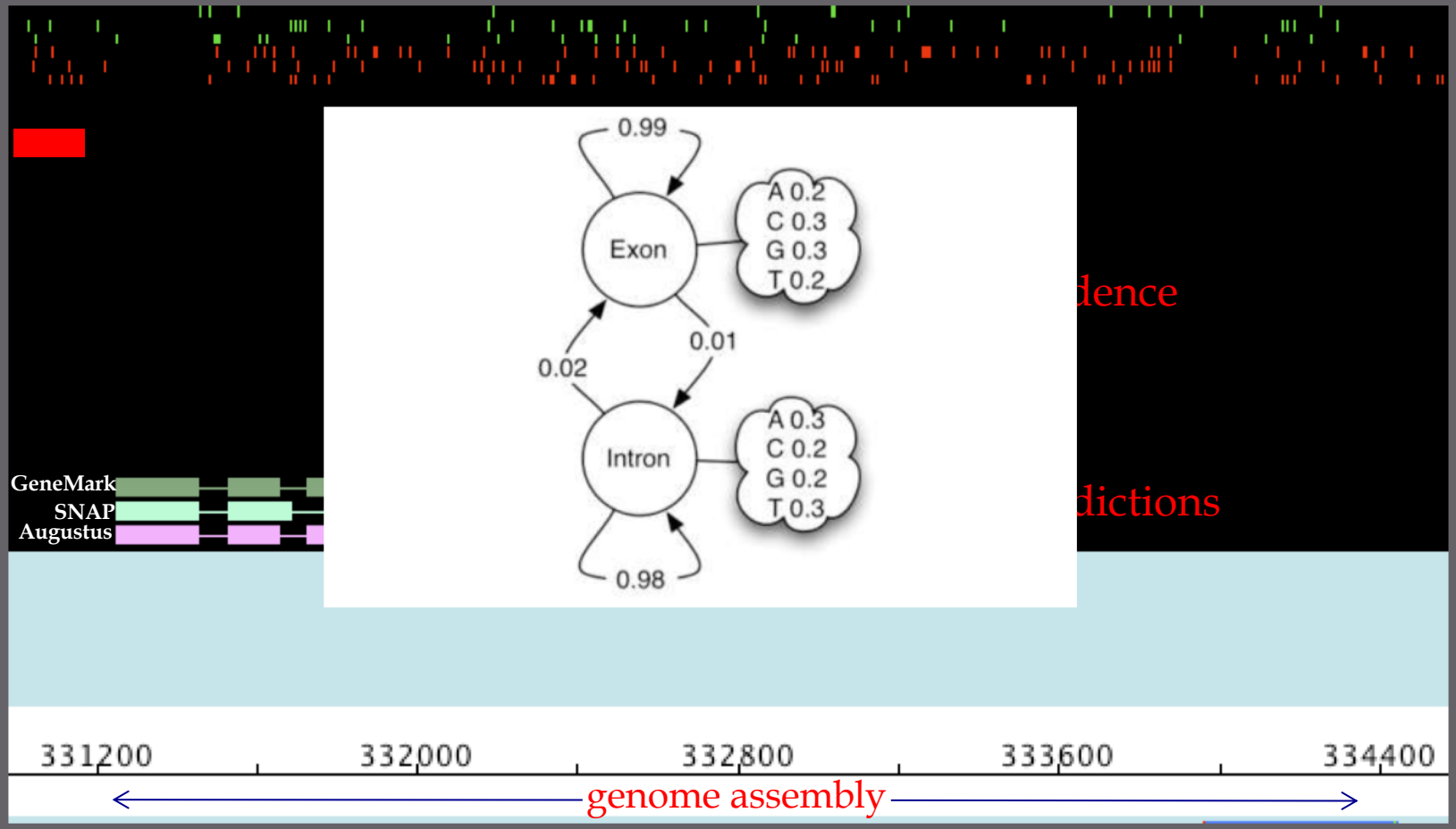
# Identify and mask repetitive elements



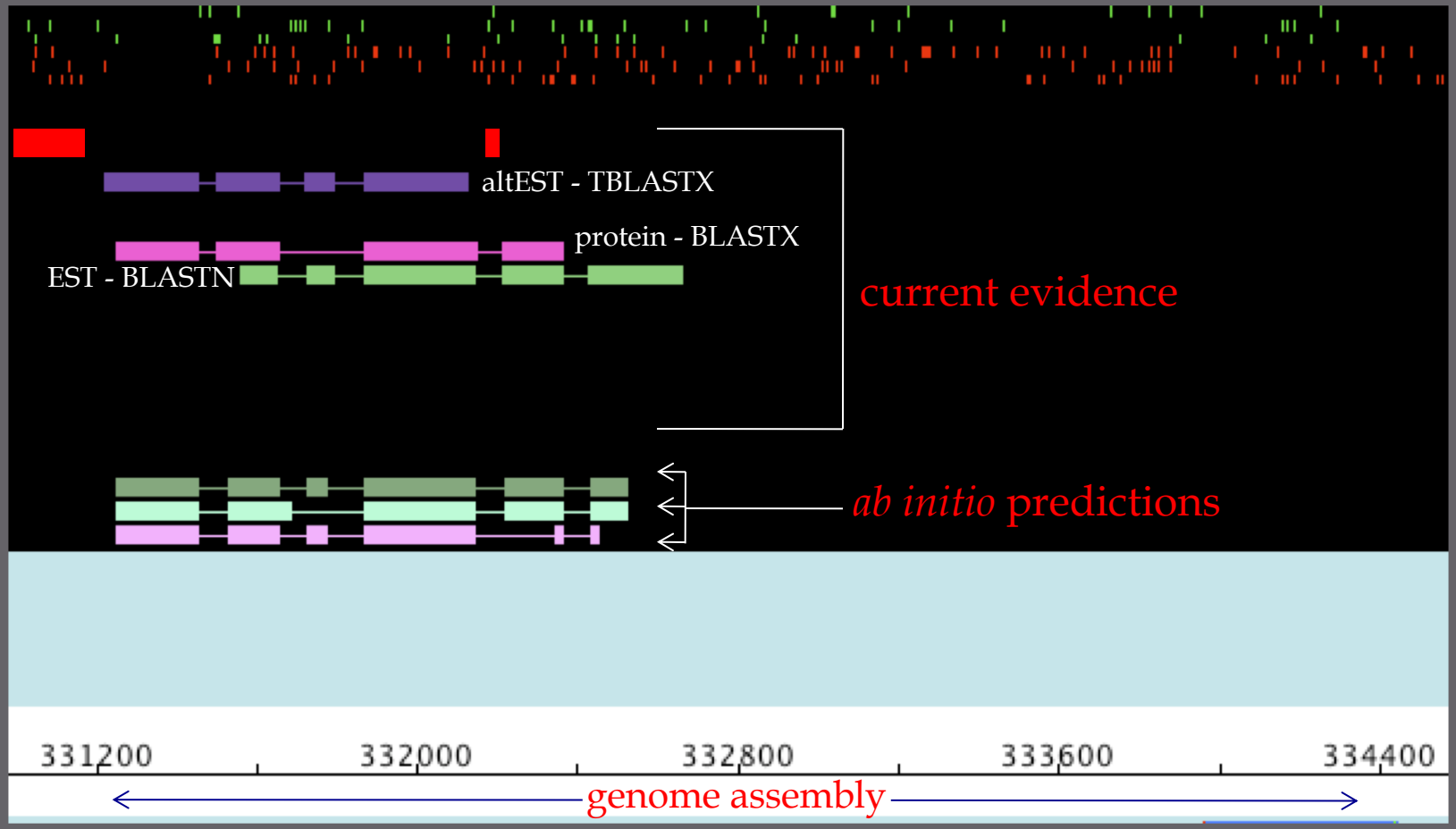
# Generate *ab initio* gene predictions



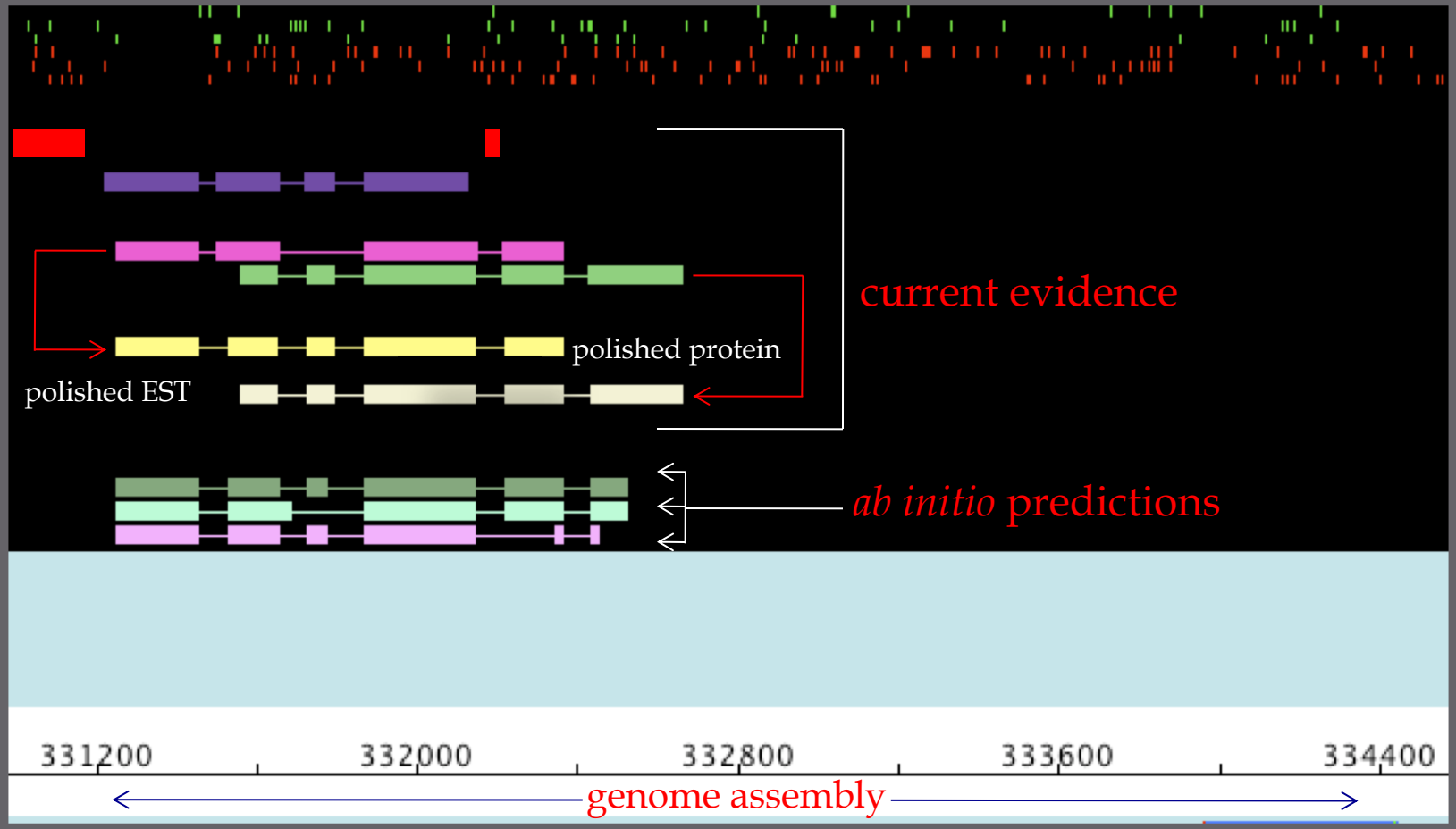
# Generate *ab initio* gene predictions



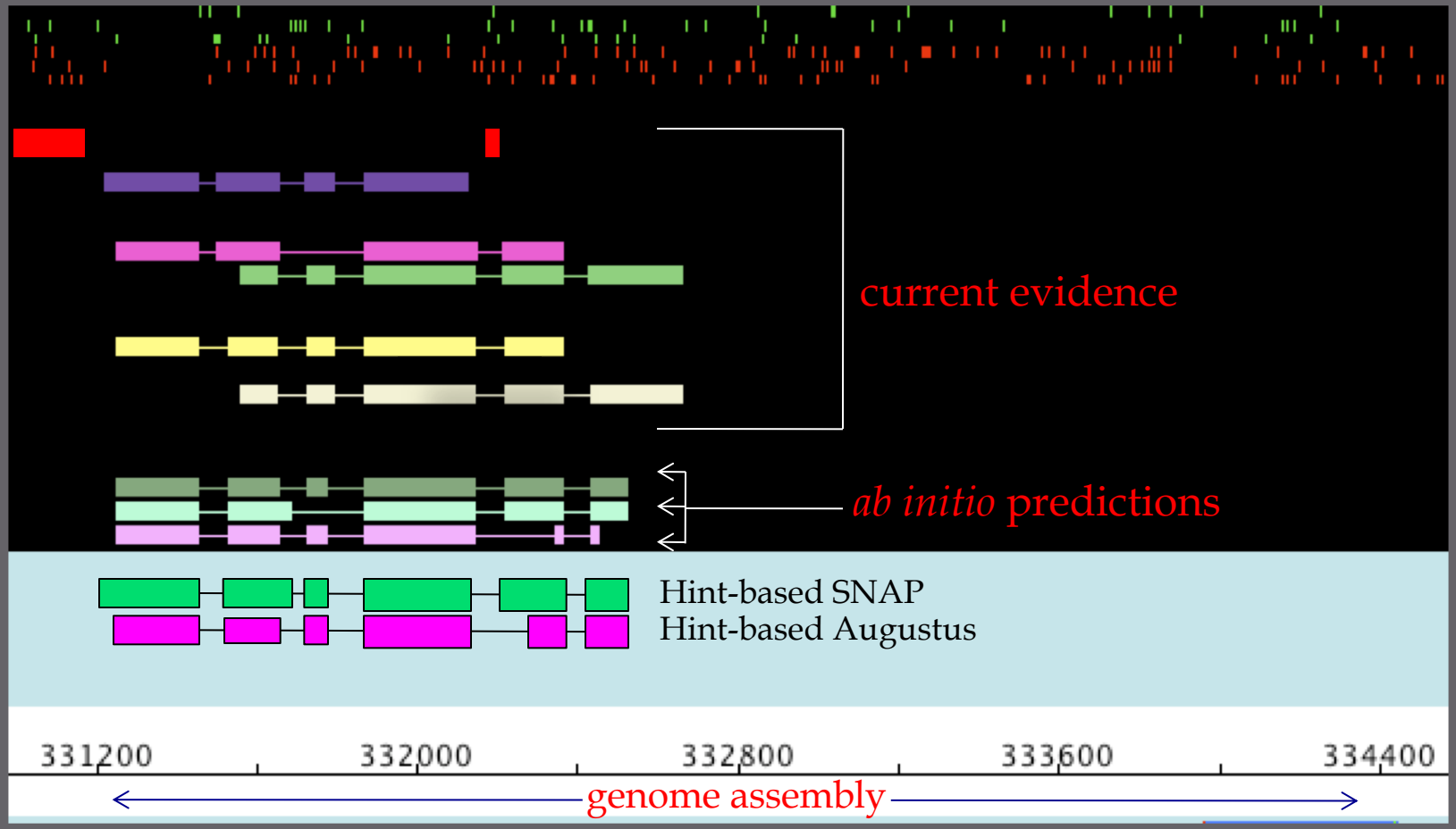
# Align EST and protein evidence



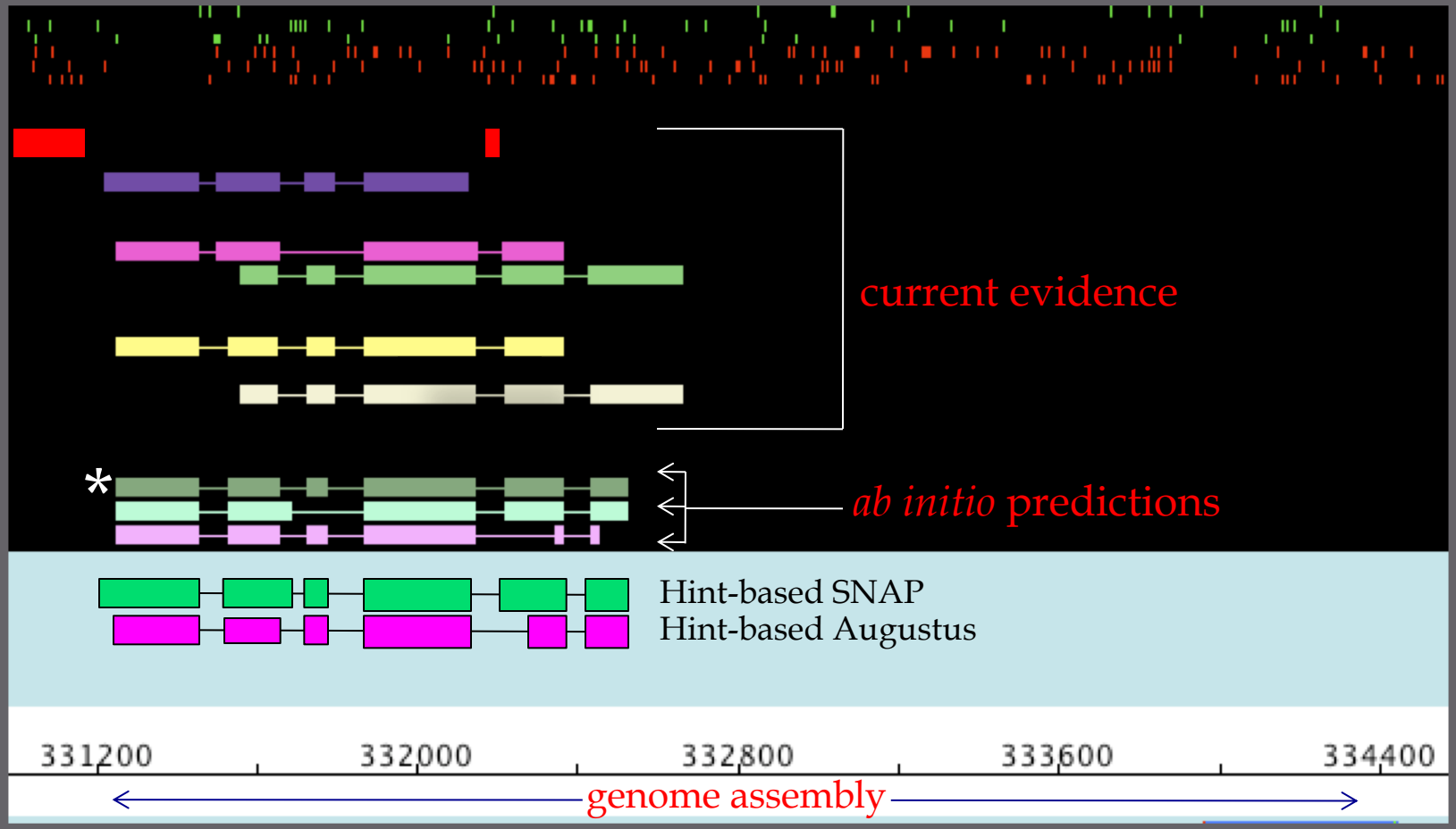
# Polish BLAST alignments with Exonerate



# Pass gene-finders evidence-based ‘hints’

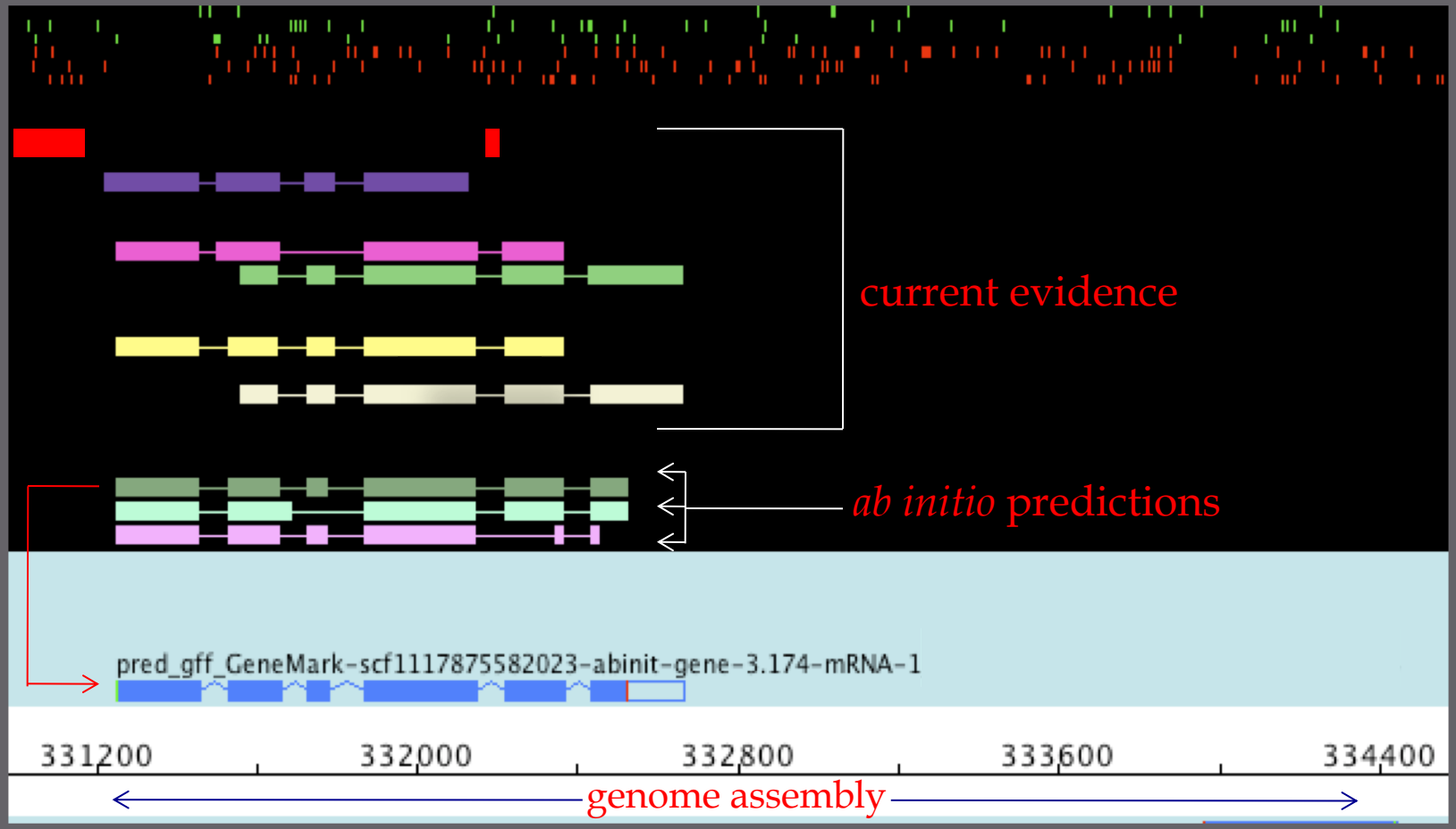


# Identify gene model most consistent with evidence

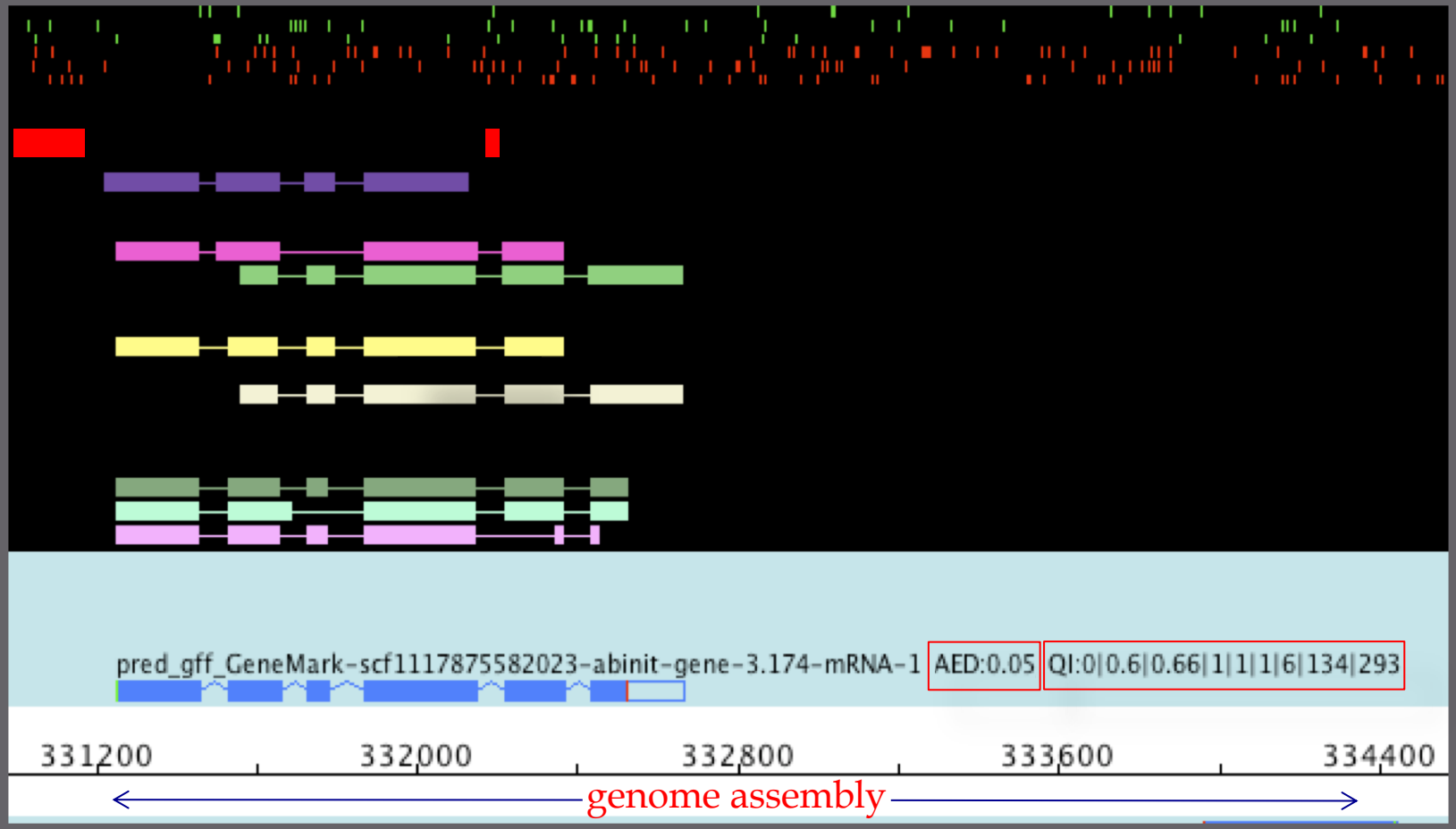



*Quantitative Measures for the Management and Comparison of Annotated Genomes
Karen Eilbeck , Barry Moore , Carson Holt and Mark Yandell BMC Bioinformatics 2009
10:67doi:10.1186/1471-2105-10-67

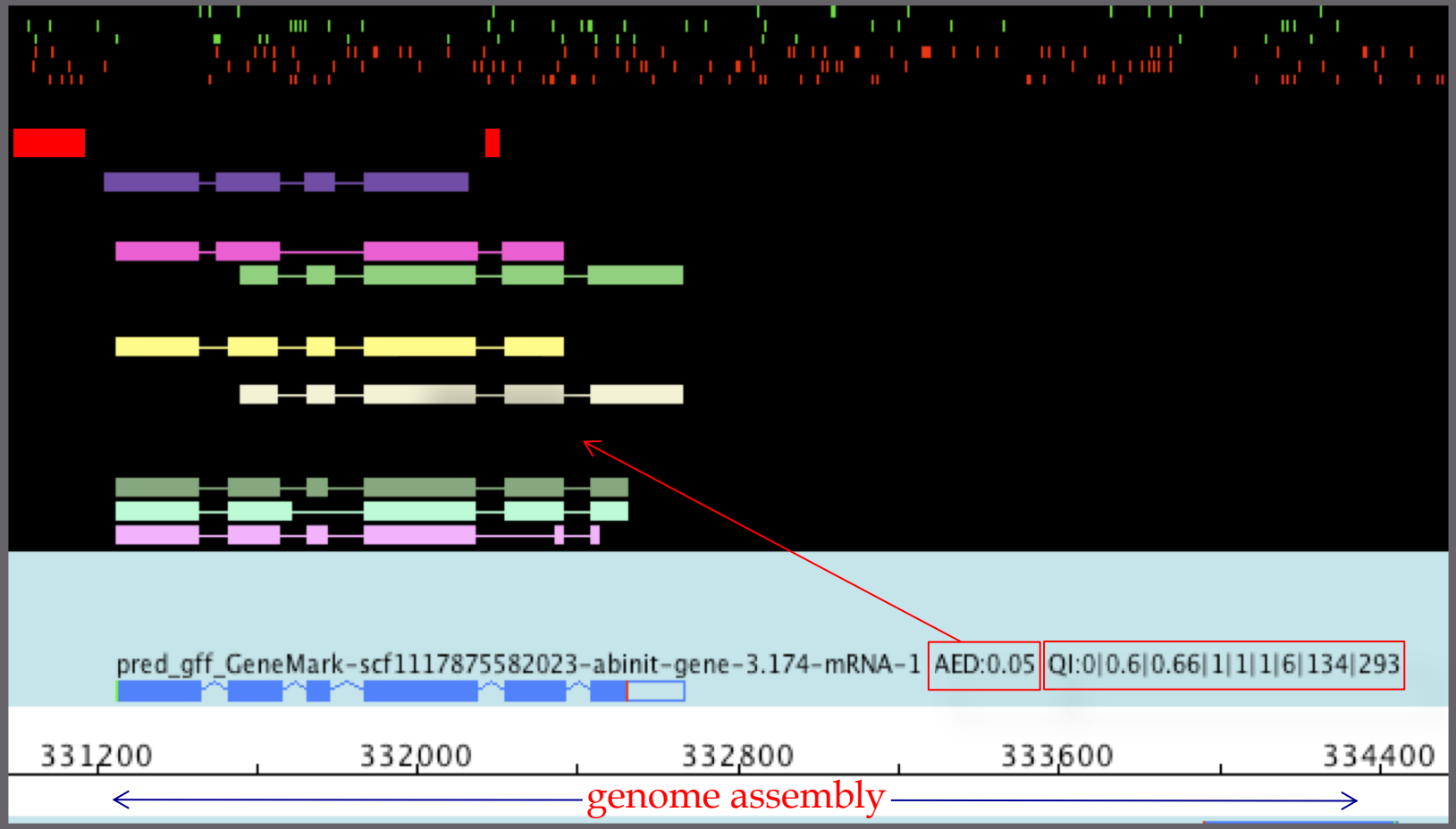
# Revise it further if necessary; create new annotation



# Compute support for each portion of gene model

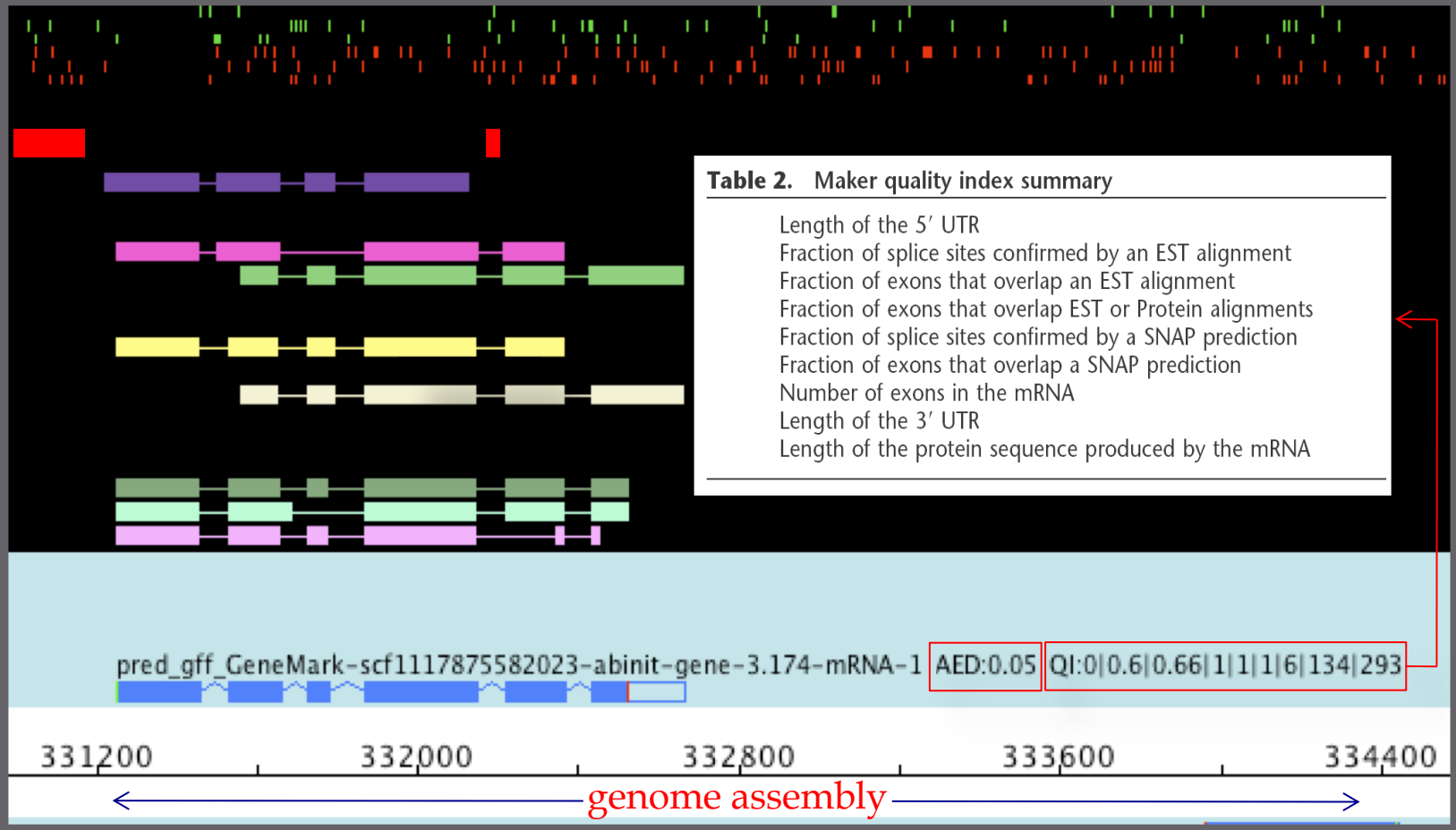


# Compute support for each portion of gene model



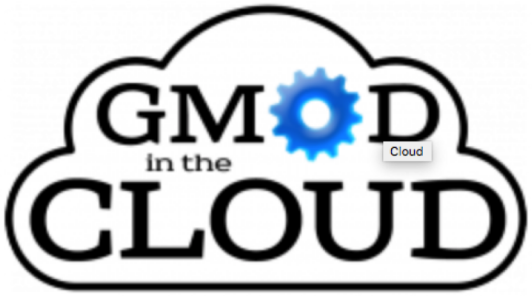

*Quantitative Measures for the Management and Comparison of Annotated Genomes
Karen Eilbeck , Barry Moore , Carson Holt and Mark Yandell BMC Bioinformatics 2009
10:67doi:10.1186/1471-2105-10-67

# Compute support for each portion of gene model




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


bin — bmoore@derringer:/data1/genomes/Homo/sapiens/36.2 — ssh
scaffold00080 repeatmasker match_part 723561 723736 914 + . ID=scaffold00080:hsp:4987;Par
scaffold00080 repeatmasker match 724950 725171 495 + . ID=scaffold00080:hit:3405;Name=specie
scaffold00080 repeatmasker match_part 724950 725171 495 + . ID=scaffold00080:hsp:4988;Par
scaffold00080 repeatmasker match 726925 727293 724 + . ID=scaffold00080:hit:3406;Name=specie
scaffold00080 repeatmasker match_part 726925 727293 724 + . ID=scaffold00080:hsp:4989;Par
scaffold00080 maker gene 56197 58302 . + . ID=ACEP_00015614;Name=ACEP_00015614;Alias=mak
scaffold00080 maker mRNA 56197 58302 . + . ID=ACEP_00015614-RA;Parent=ACEP_00015614;Name
scaffold00080 maker exon 56197 56274 . + . ID=ACEP_00015614-RA:exon:126;Parent=ACEP_0001
scaffold00080 maker exon 56569 56584 . + . ID=ACEP_00015614-RA:exon:127;Parent=ACEP_0001
scaffold00080 maker exon 56797 56906 . + . ID=ACEP_00015614-RA:exon:128;Parent=ACEP_0001
scaffold00080 maker exon 57851 57941 . + . ID=ACEP_00015614-RA:exon:129;Parent=ACEP_0001
scaffold00080 maker exon 58067 58302 18.726 + . ID=ACEP_00015614-RA:exon:130;Parent=ACEP_0001
scaffold00080 maker CDS 56197 56274 . + 0 ID=ACEP_00015614-RA:cds:124;Parent=ACEP_00015
scaffold00080 maker CDS 56569 56584 . + 0 ID=ACEP_00015614-RA:cds:125;Parent=ACEP_00015
scaffold00080 maker CDS 56797 56906 . + 2 ID=ACEP_00015614-RA:cds:126;Parent=ACEP_00015
scaffold00080 maker CDS 57851 57941 . + 1 ID=ACEP_00015614-RA:cds:127;Parent=ACEP_00015
scaffold00080 maker CDS 58067 58302 . + 2 ID=ACEP_00015614-RA:cds:128;Parent=ACEP_00015
scaffold00080 maker gene 2797 3250 . - . ID=ACEP_00015615;Name=ACEP_00015615;Alias=mak
GFF3
bin — bmoore@derringer:/data1/genomes/Homo/sapiens/36.2 — ssh
>ACEP_00015614-RA protein AED:0.401129943502825 QI:0|0|0|0.2|1|1|5|0|177
MEKSDDYQDHYVIQFLEVYMYYFVHEKQYYLKVLWKRIITLSNKYYPIGVIWAFFVPIVF
IRHPDDLKTILSNPKHIKKSFFYDNIKPWLGTSILTSEDCTLSHFIPMLSNRHFTIVHCI
VHVLGTKWQLQRKILISTFHFDILNQFVEIFEKESKNMIKSLKNAEGTVVKDLSSFI
>ACEP_00015615-RA protein AED:0.223684210526316 QI:60|0|0|0.5|0|0|2|0|93
MDTQQKHKREIPTEPGEINSVISFQICISCCDLSYCNIESPTNATNAIYSRQRRAKSKSK
RKRPGRNGVDAATRLYGSSLLWLPASLFYAYHC
>ACEP_00015613-RA protein AED:0.191135958515638 QI:9|0.4|0.33|0.83|0.2|0.5|6|0|238
MEKSDNSLTYQDHYSIFGSVYVLFGSREILWKIITLSNKFLRTFYPIGIIRAFFVPIVSI
RHPDDLKSFFYDNIKPWLGTSILINEGTKWQLQRKILIPTFHFDILNQFVEIFEKESKNM
IKSLKNAEGTVVKDLSSFISEYTLNAICESYLIPAGTVLHININGVHTDPNFWRYFQIQR
YLILIDFCPRGSEIVTLTHIYHSVLDHVIVSNLRFSKFDHERSGIVRE
>ACEP_00015612-RA protein AED:0.469714351691491 249
MYFVNFNYFYANASVTNHRNGRTSLHTELLSSKVKREQTLRNILTARR
RILSSRRPRPGDIISKDPQDVLETQETRNGIYIANARTEISIKEEIPQ
INLNATGEELESRTTKTSKNESLSPTTEKIENISSIVSTPATNQKLGDAEMRKKNTVSMP
ESLYNHFRPVESNIPVEDMSQFLYFGQKLQPDALNVTSSSNNSVETTSTNPTSSRRRYST
KRFTATIATPMEEIMNEEMNIALETLERRTVEKNQVSRIKNTFRSSGNGLYYRKPRPTAD
VVSNIIPGNESSPGIIGGSETKSEMNTVDDKFHRSDSSAHADGSRDRVPLERKNAAHVSE
VSMTTQIPRQSDVEARVIVESEIGAKLSNKTITASNGIVESLQRINDFGKDNEKKKITEA
EAGNVEKSEHKLMSVERTISTSQRESERFRESDSTETMSLVPSRAESSTLRIVVPDNSFG
FASTA


GMOD in the Cloud toolset


GBrowse: Genome annotation viewer


Galaxy: Data analysis & integration


Chado: Biological database schema


JBrowse: Super-fast genome annotation viewer


BioMart: Data mining system


WebApollo: browser-based annotation editor


MAKER: Genome annotation pipeline


GBrowse_syn: Synteny viewer


Tripal: Chado web interface


InterMine: Data warehousing

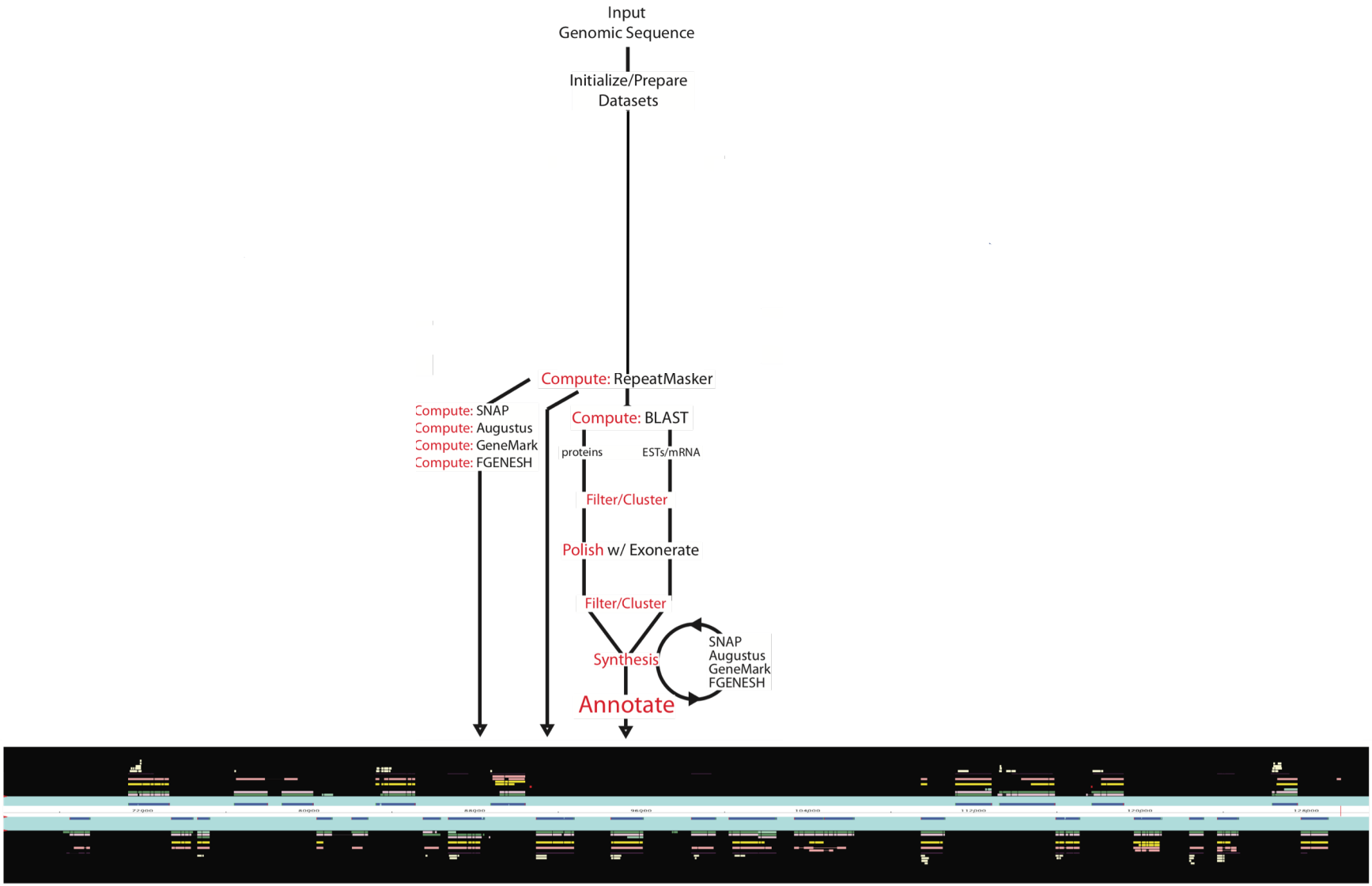

CMap: Comparative map viewer

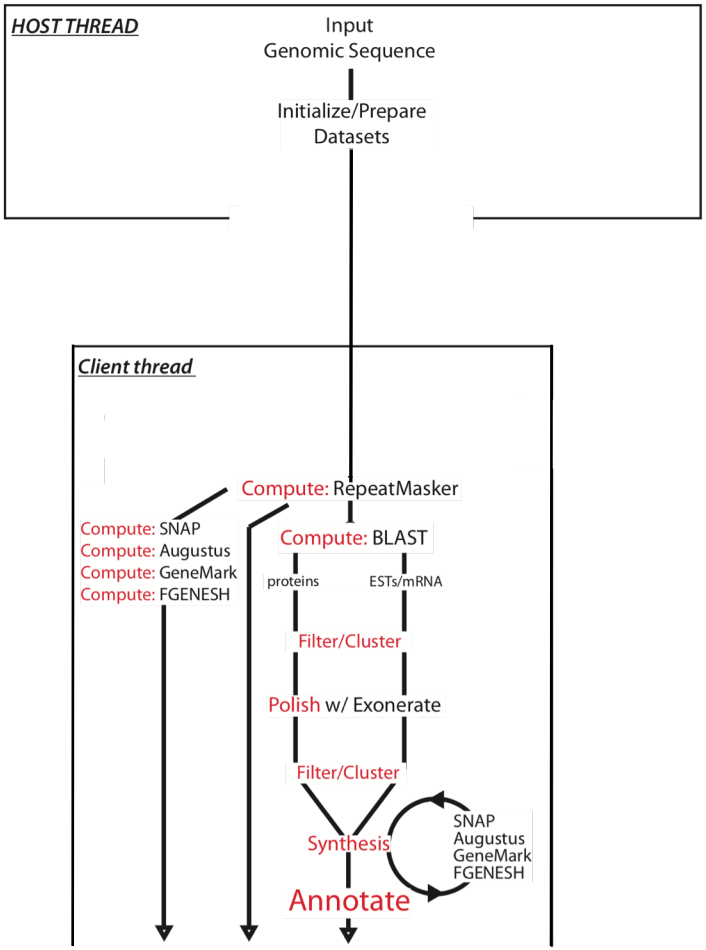

Pathway Tools: Metabolic, regulatory pathways

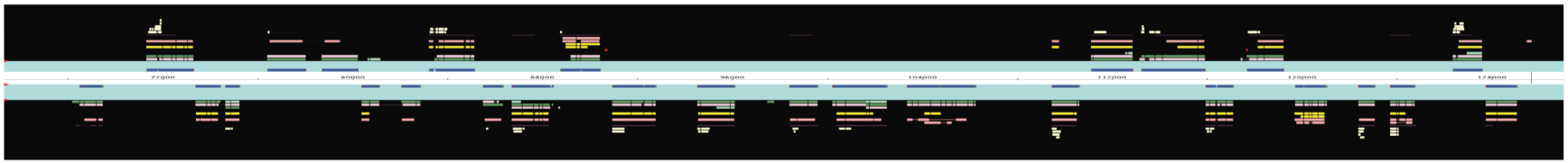

Canto: literature annotation tool

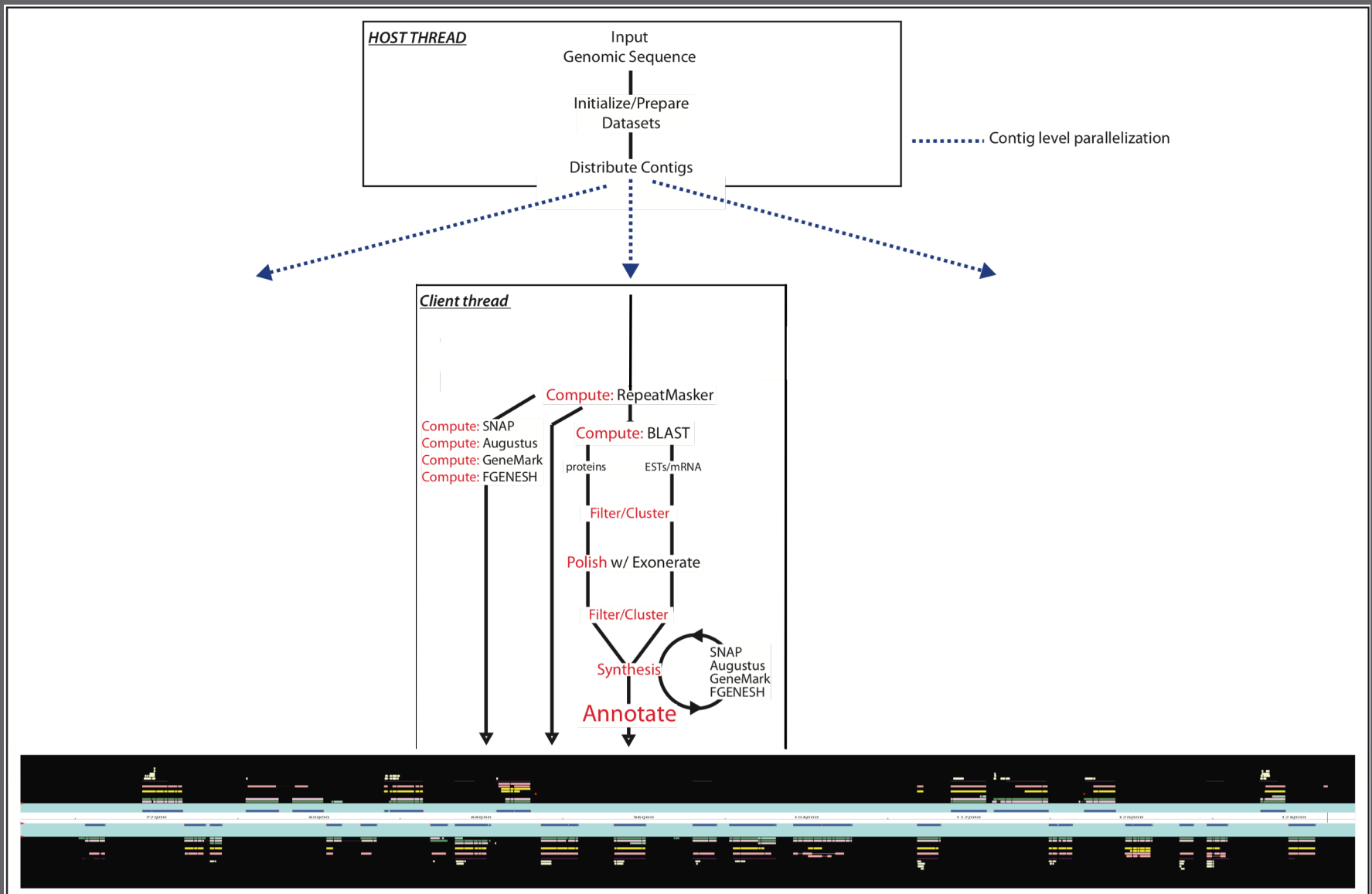
# Distributed Parallelization

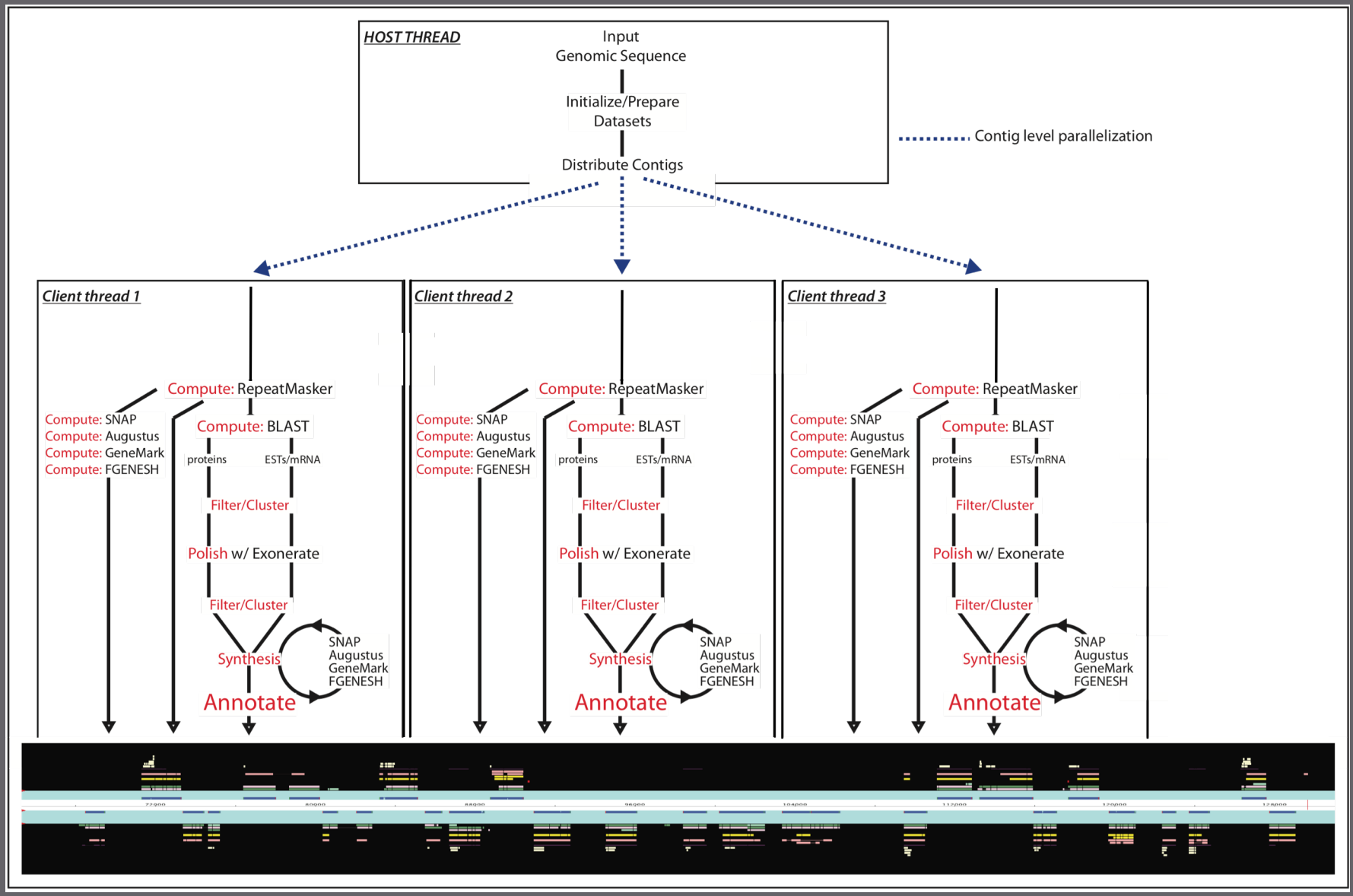
- Supports Message Passing Interface (MPI), a communication protocol for computer clusters which essentially allows multiple computers to act like a single powerful machine.

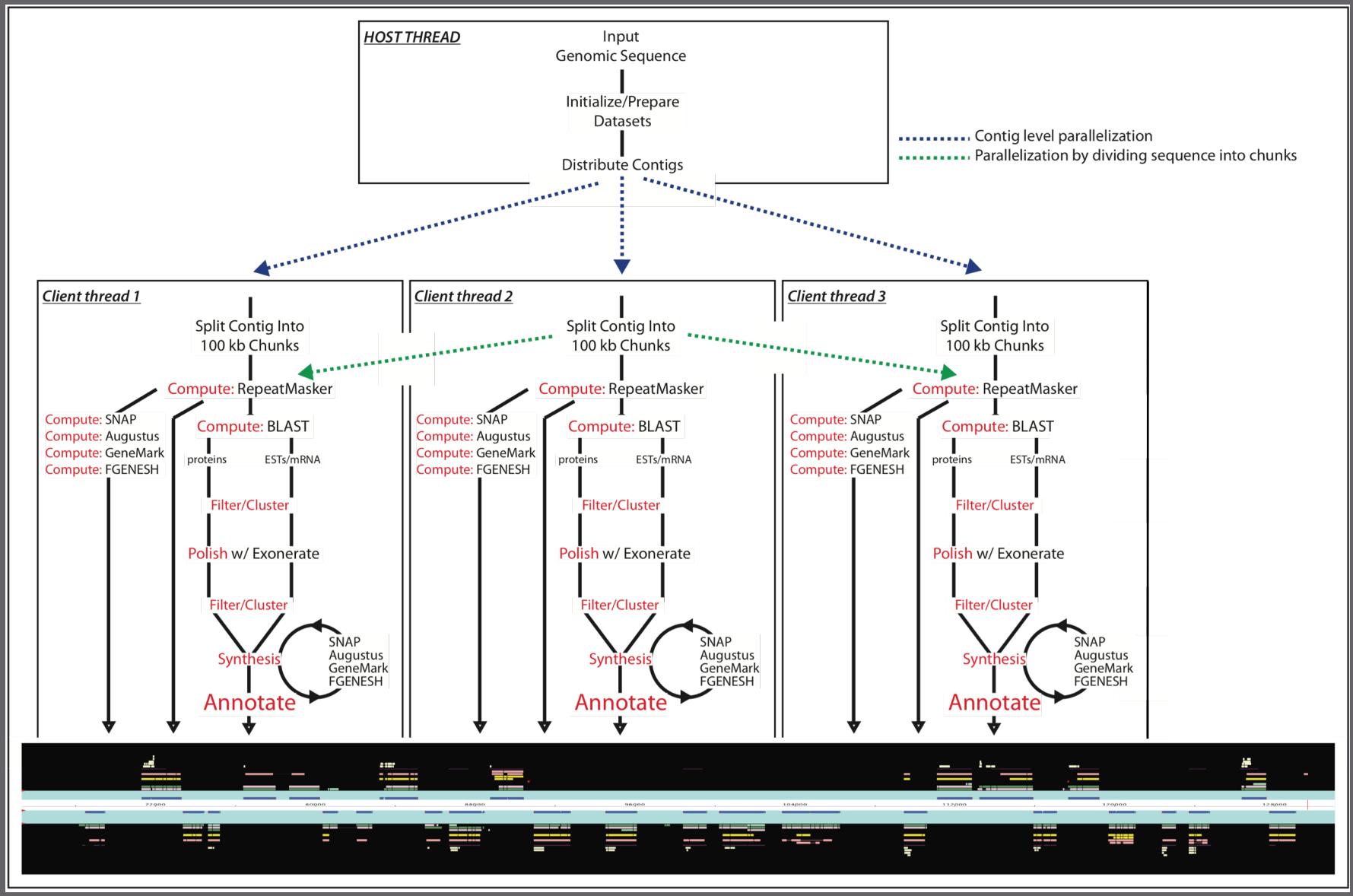

Input
Genomic Sequence
Initialize/Prepare
Datasets
Compute: RepeatMasker
Compute: SNAP
Compute: Augustus
Compute: GeneMark
Compute: FGENESH
Compute: BLAST
proteins
ESTs/mRNA
Filter/Cluster
Polish w/ Exonerate
Filter/Cluster
Synthesis
SNAP
Augustus
GeneMark
FGENESH
Annotate

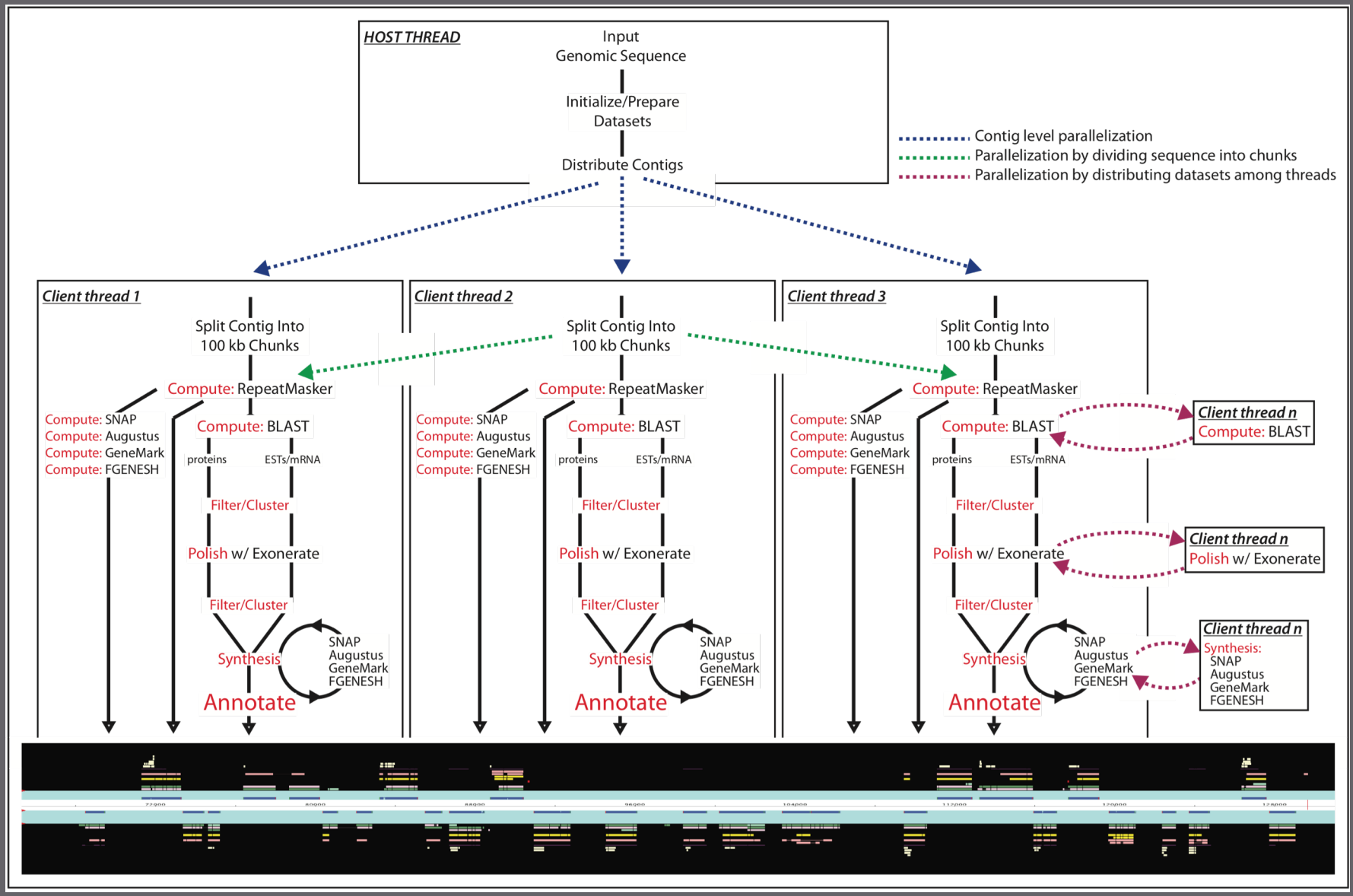

HOST THREAD
Input
Genomic Sequence
Initialize/Prepare
Datasets
Client thread
Compute: RepeatMasker
Compute: SNAP
Compute: Augustus
Compute: GeneMark
Compute: FGENESH
Compute: BLAST
proteins
ESTs/mRNA
Filter/Cluster
Polish w/ Exonerate
Filter/Cluster
Synthesis
SNAP
Augustus
GeneMark
FGENESH
Annotate



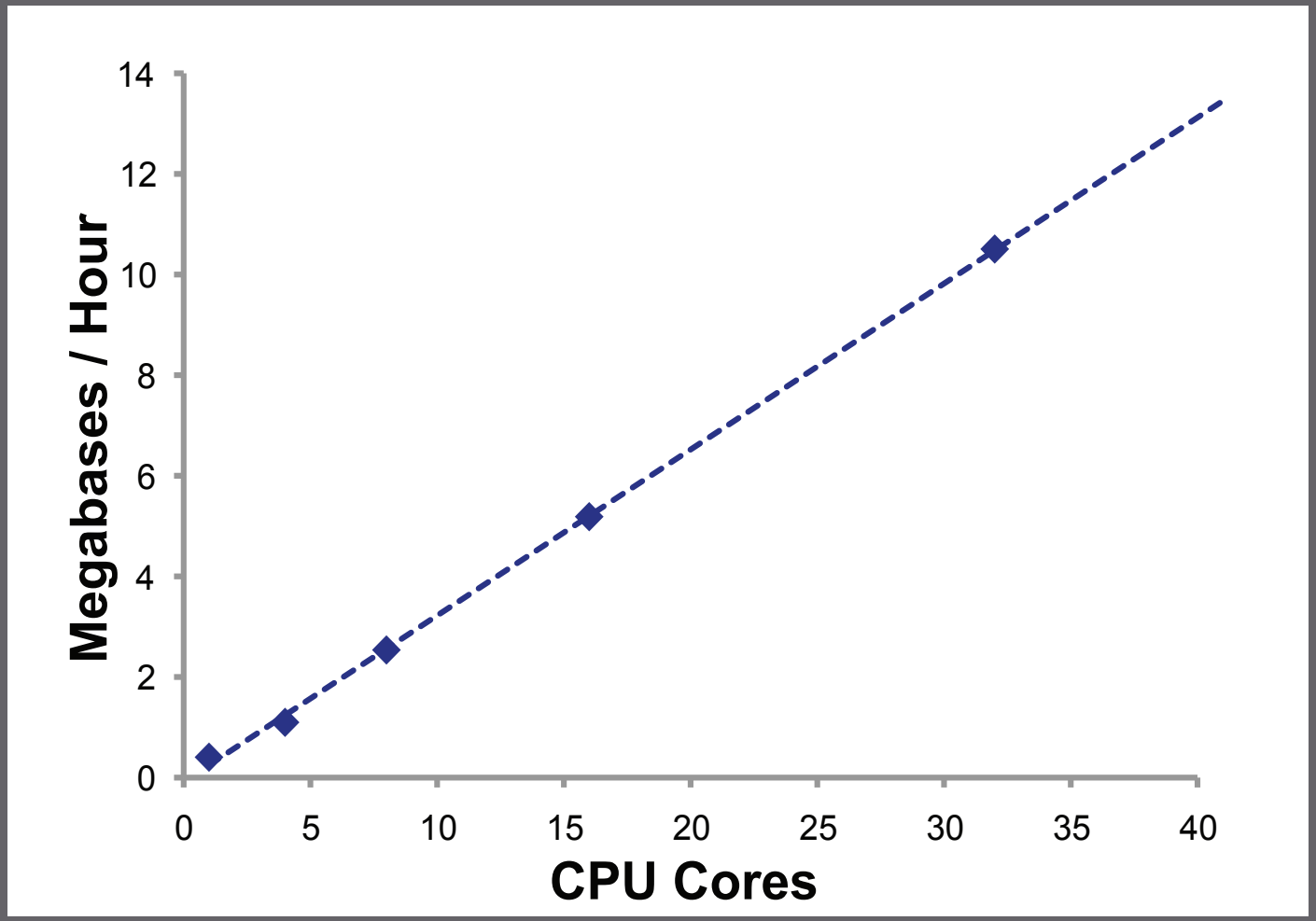

HOST THREAD
Input
Genomic Sequence
Initialize/Prepare
Datasets
Distribute Contigs
Contig level parallelization
Client thread
Compute: RepeatMasker
Compute: SNAP
Compute: Augustus
Compute: GeneMark
Compute: FGENESH
Compute: BLAST
proteins
ESTs/mRNA
Filter/Cluster
Polish w/ Exonerate
Filter/Cluster
Synthesis
SNAP
Augustus
GeneMark
FGENESH
Annotate


HOST THREAD
Input
Genomic Sequence
Initialize/Prepare
Datasets
Distribute Contigs
Contig level parallelization
Client thread 1
Client thread 2
Client thread 3
Compute: RepeatMasker
Compute: SNAP
Compute: Augustus
Compute: GeneMark
Compute: FGENESH
Compute: BLAST
proteins
ESTs/mRNA
Filter/Cluster
Polish w/ Exonerate
Filter/Cluster
Synthesis
SNAP
Augustus
GeneMark
FGENESH
Annotate

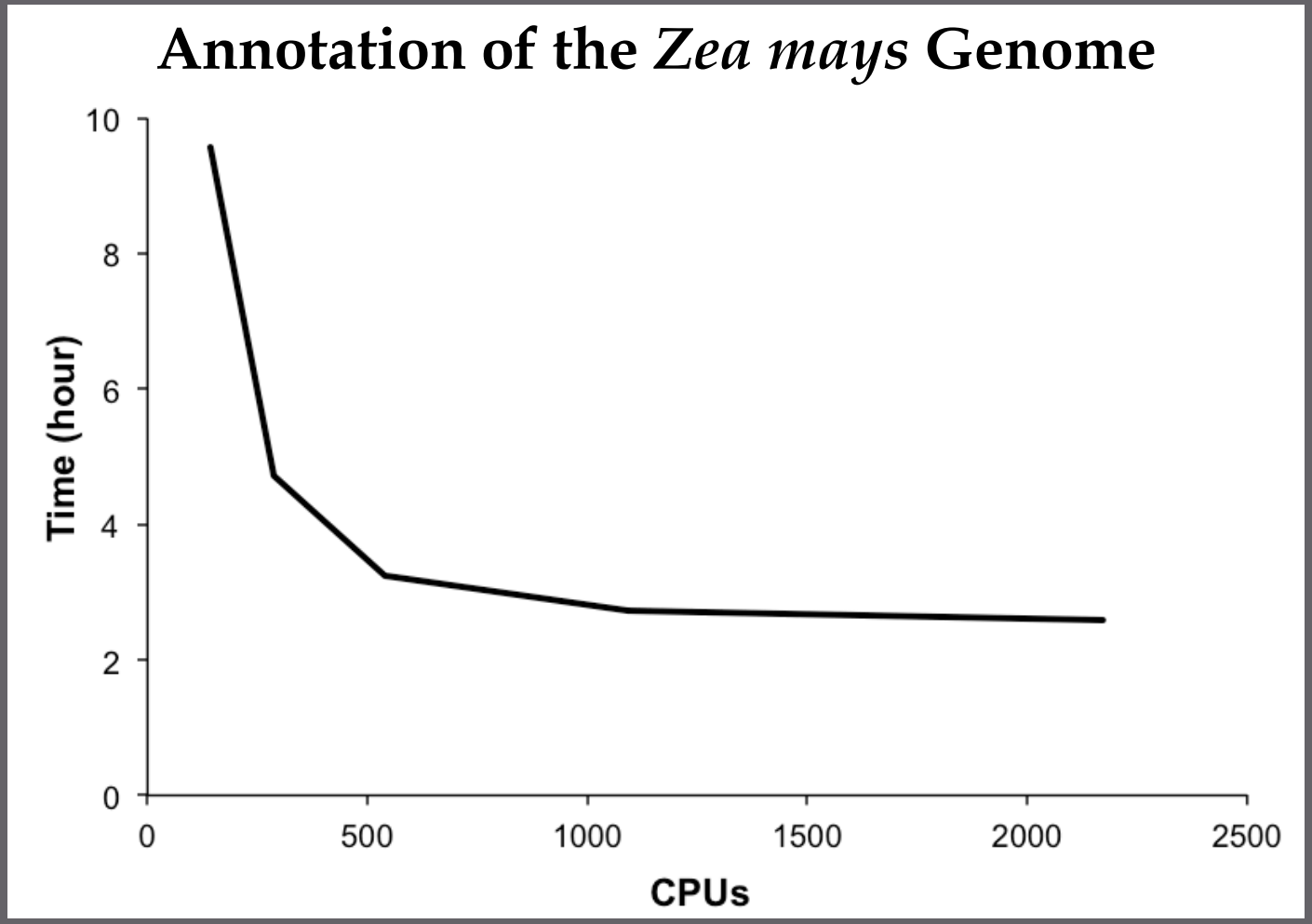

HOST THREAD
Input Genomic Sequence
Initialize/Prepare Datasets
Distribute Contigs
Contig level parallelization
Parallelization by dividing sequence into chunks
Client thread 1
Client thread 2
Client thread 3
Split Contig Into 100 kb Chunks
Compute: RepeatMasker
Compute: SNAP
Compute: Augustus
Compute: GeneMark
Compute: FGENESH
Compute: BLAST
proteins
ESTs/mRNA
Filter/Cluster
Polish w/ Exonerate
Filter/Cluster
Synthesis
SNAP
Augustus
GeneMark
FGENESH
Annotate


HOST THREAD
Input Genomic Sequence
Initialize/Prepare Datasets
Distribute Contigs
Contig level parallelization
Parallelization by dividing sequence into chunks
Parallelization by distributing datasets among threads
Client thread 1
Client thread 2
Client thread 3
Split Contig Into 100 kb Chunks
Compute: RepeatMasker
Compute: SNAP
Compute: Augustus
Compute: GeneMark
Compute: FGENESH
Compute: BLAST
proteins
ESTs/mRNA
Filter/Cluster
Polish w/ Exonerate
Filter/Cluster
Synthesis
SNAP
Augustus
GeneMark
FGENESH
Annotate
Client thread n
Compute: BLAST
Client thread n
Polish w/ Exonerate
Client thread n
Synthesis:
SNAP
Augustus
GeneMark
FGENESH

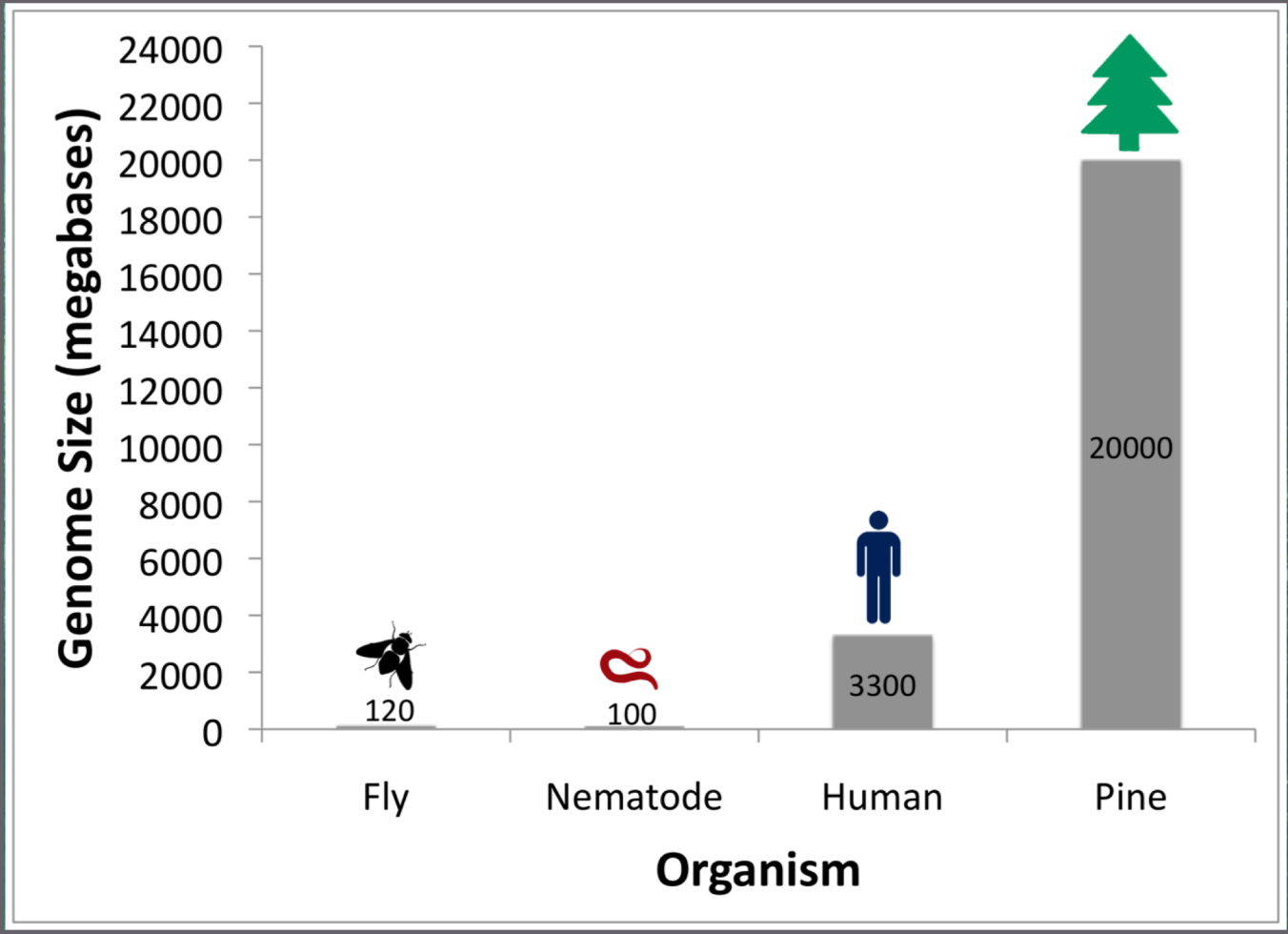
# Data throughput



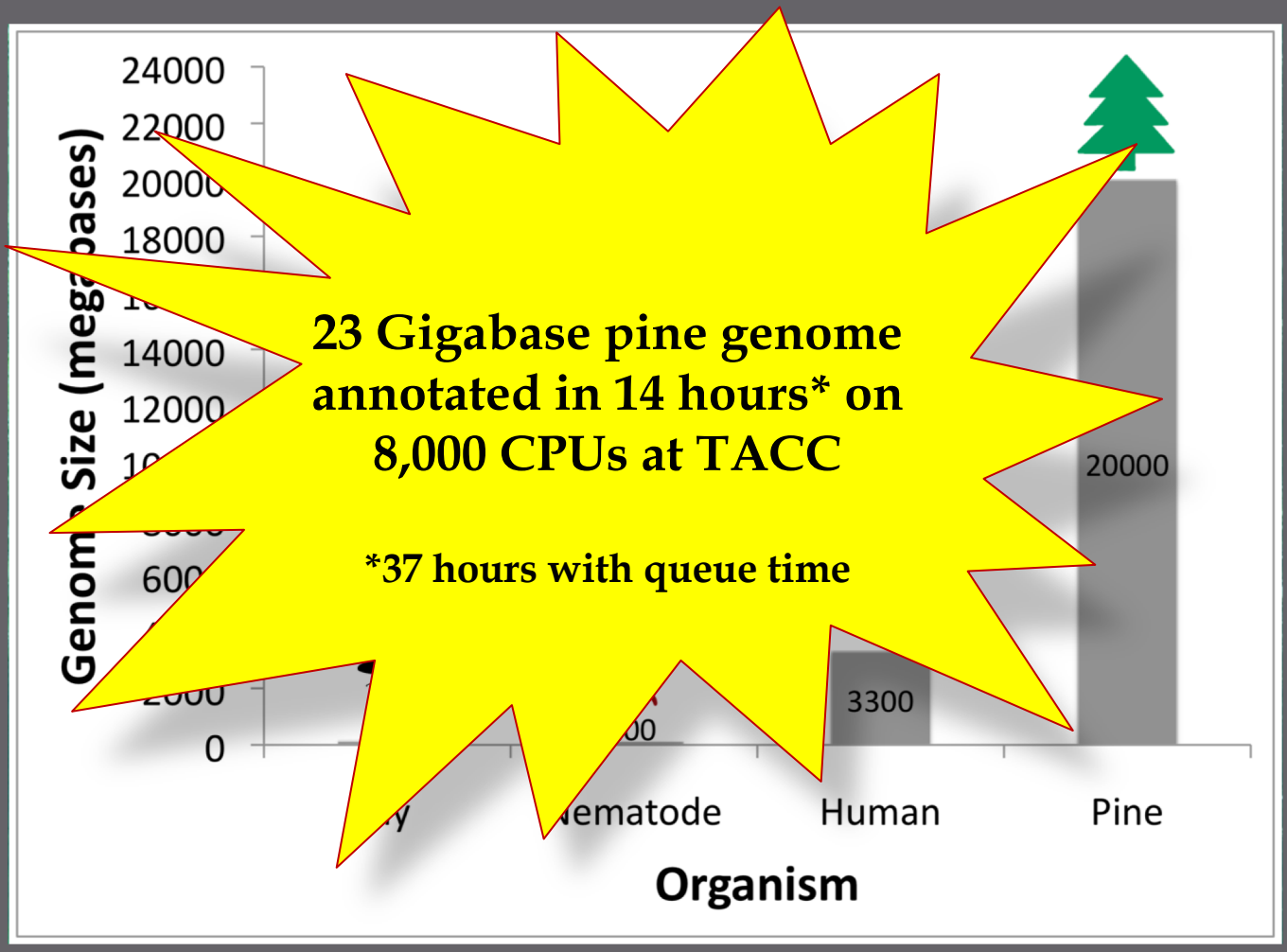
# Data throughput



# Genome Sizes



# Genome Sizes



# Next Generation of Genomics

"Second-generation " genome projects have different needs and characteristics than earlier works.

# First *versus* second-generation genomes

First-generation genomes:

- Classic experimental systems

- Large community

- Big $

- Much prior knowledge about genome

# Annotation Evaluation

**Gene model accuracies for *ab initio* prediction and genome annotation programs**

| Reference Organism | Performance Category | *Ab Initio* Prediction | | | MAKER Annotations |
|---|---|---|---|---|---|
| | | Augustus | GeneMark | SNAP | SNAP |
| *Caenorhabditis elegans* | Nucleotide Accuracy | 88.29% | 88.09% | 85.10% | 88.48% |
| *Arabidopsis thaliana* | Nucleotide Accuracy | 77.04% | 74.68% | 69.78% | 80.27% |
| *Drosophila melanogaster* | Nucleotide Accuracy | 76.08% | 66.54% | 69.29% | 74.33% |

# Annotation Evaluation

**Gene model accuracies for *ab initio* prediction and genome annotation programs**

| Reference Organism | Performance Category | *Ab Initio* Prediction | | | MAKER Annotations |
|---|---|---|---|---|---|
| | | Augustus | GeneMark | SNAP | SNAP |
| *Caenorhabditis elegans* | Nucleotide Accuracy | 88.29% | 88.09% | 85.10% | 88.48% |
| *Arabidopsis thaliana* | Nucleotide Accuracy | 77.04% | 74.68% | 69.78% | 80.27% |
| *Drosophila melanogaster* | Nucleotide Accuracy | 76.08% | 66.54% | 69.29% | 74.33% |

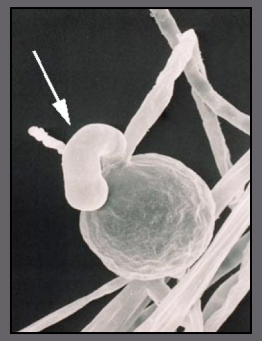
With **enough** training data, *ab initio* gene predictors can match or even out-perform annotation pipelines*

***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

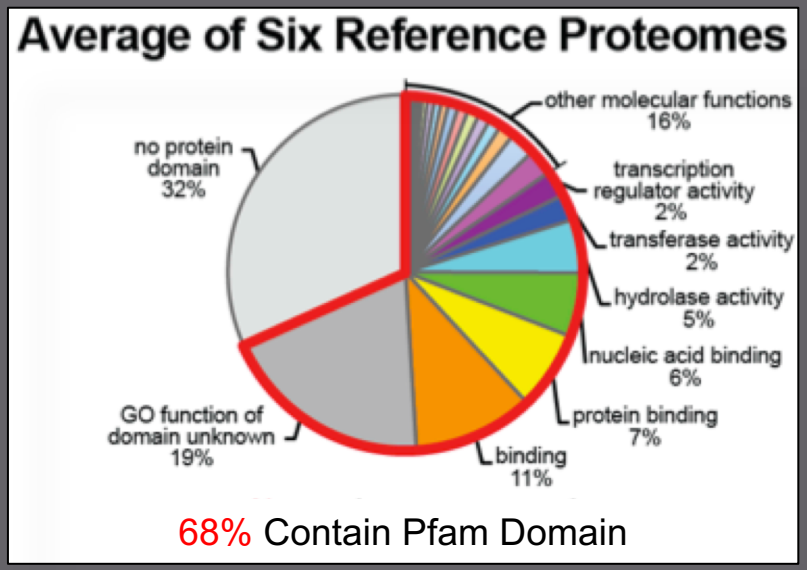
# First *versus* second-generation genomes

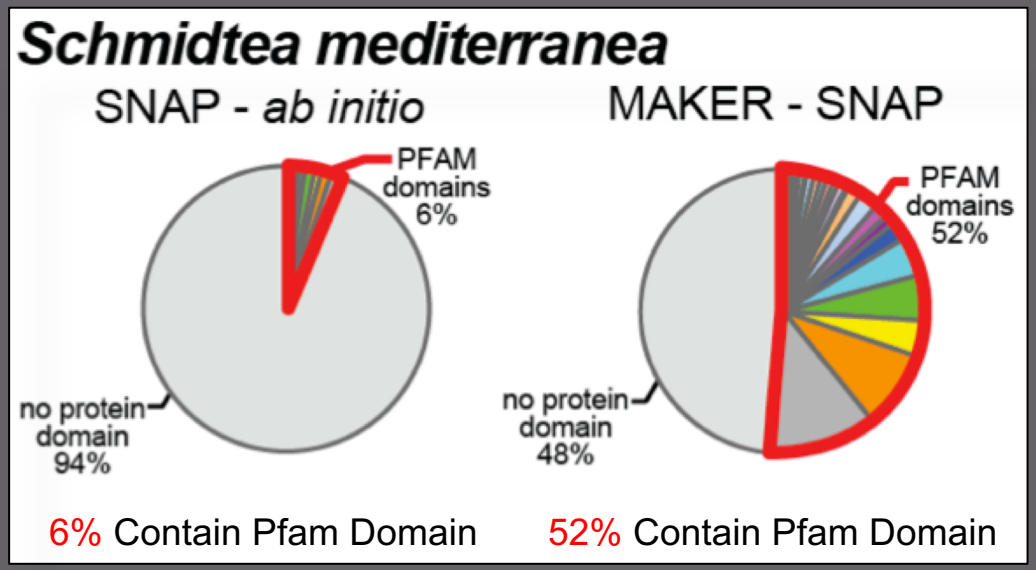
Second-generation genomes:

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



# Evaluation of *Ab initio* gene predictors on emerging model organism genomes



# Evaluation of gene models with non-matched species parameters

**Gene model accuracies for *ab initio* prediction and genome annotation programs**

| Reference Organism | Performance Category | *Ab Initio* Prediction | | |
|---|---|---|---|---|
| | | Augustus | GeneMark | SNAP |
| *Arabidopsis thaliana* | Nucleotide Accuracy | 77.04% | 74.68% | 69.78% |

# Evaluation of gene models with non-matched species parameters

**Gene model accuracies for *ab initio* prediction and genome annotation programs**

| Reference Organism | Performance Category | *Ab Initio* Prediction | | |
|---|---|---|---|---|
| | | Augustus | GeneMark | SNAP |
| *Arabidopsis thaliana* | Nucleotide Accuracy | 77.04% | 74.68% | 69.78% |

**Gene model accuracies when using unmatched species parameter files**

| Reference Organism | Performance Category | *Ab Initio* Prediction | | |
|---|---|---|---|---|
| | | Augustus | GeneMark | SNAP |
| *Arabidopsis thaliana* | Nucleotide Accuracy | 57.85% | 48.62% | 43.84% |

# Evaluation of gene models with non-matched species parameters

**Gene model accuracies for *ab initio* prediction and genome annotation programs**

| Reference Organism | Performance Category | *Ab Initio* Prediction | | |
|---|---|---|---|---|
| | | Augustus | GeneMark | SNAP |
| *Arabidopsis thaliana* | Nucleotide Accuracy | 77.04% | 74.68% | 69.78% |
| | Exon Accuracy | 67.03% | 61.31% | 56.40% |

**Gene model accuracies when using unmatched species parameter files**

| Reference Organism | Performance Category | *Ab Initio* Prediction | | |
|---|---|---|---|---|
| | | Augustus | GeneMark | SNAP |
| *Arabidopsis thaliana* | Nucleotide Accuracy | 57.85% | 48.62% | 43.84% |
| | Exon Accuracy | 30.71% | 16.51% | 18.58% |

# Evaluation of gene models with non-matched species parameters

**Gene model accuracies for *ab initio* prediction and genome annotation programs**

| Reference Organism | Performance Category | *Ab Initio* Prediction | | |
|---|---|---|---|---|
| | | Augustus | GeneMark | SNAP |
| *Arabidopsis thaliana* | Nucleotide Accuracy | 77.04% | 74.68% | 69.78% |
| | Exon Accuracy | 67.03% | 61.31% | 56.40% |

**Gene model accuracies when using unmatched species parameter files**

| Reference Organism | Performance Category | *Ab Initio* Prediction | | |
|---|---|---|---|---|
| | | Augustus | GeneMark | SNAP |
| *Arabidopsis thaliana* | Nucleotide Accuracy | 57.85% | 48.62% | 43.84% |
| | Exon Accuracy | 30.71% | 16.51% | 18.58% |

# Evaluation of gene models with non-matched species parameters

**Gene model accuracies for *ab initio* prediction and genome annotation programs**

| Reference Organism | Performance Category | *Ab Initio* Prediction | | |
|---|---|---|---|---|
| | | Augustus | GeneMark | SNAP |
| *Arabidopsis thaliana* | Nucleotide Accuracy | 77.04% | 74.68% | 69.78% |
| | Exon Accuracy | 67.03% | 61.31% | 56.40% |

**Gene model accuracies when using unmatched species parameter files**

| Reference Organism | Performance Category | *Ab Initio* Prediction | | | MAKER Annotations |
|---|---|---|---|---|---|
| | | Augustus | GeneMark | SNAP | SNAP |
| *Arabidopsis thaliana* | Nucleotide Accuracy | 57.85% | 48.62% | 43.84% | 73.77% |
| | Exon Accuracy | 30.71% | 16.51% | 18.58% | 60.11% |

# Evaluation of gene models with non-matched species parameters

**Gene model accuracies for *ab initio* prediction and genome annotation programs**

| Reference Organism | Performance Category | *Ab Initio* Prediction | | |
|---|---|---|---|---|
| | | Augustus | GeneMark | SNAP |
| *Arabidopsis thaliana* | Nucleotide Accuracy | 77.04% | 74.68% | 69.78% |
| | Exon Accuracy | 67.03% | 61.31% | 56.40% |
| *Drosophila Melanogaster* | Nucleotide Accuracy | 76.08% | 66.54% | 69.29% |
| | Exon Accuracy | 61.37% | 47.31% | 47.01% |

**Gene model accuracies when using unmatched species parameter files**

| Reference Organism | Performance Category | *Ab Initio* Prediction | | | MAKER Annotations |
|---|---|---|---|---|---|
| | | Augustus | GeneMark | SNAP | SNAP |
| *Arabidopsis thaliana* | Nucleotide Accuracy | 57.85% | 48.62% | 43.84% | 73.77% |
| | Exon Accuracy | 30.71% | 16.51% | 18.58% | 60.11% |
| *Drosophila Melanogaster* | Nucleotide Accuracy | 67.47% | 66.51% | 48.92% | 74.44% |
| | Exon Accuracy | 30.62% | 26.25% | 19.94% | 53.69% |

# Evaluation of gene models with non-matched species parameters

**Gene model accuracies for *ab initio* prediction and genome annotation programs**

| Reference Organism | Performance Category | *Ab Initio* Prediction | | |
|---|---|---|---|---|
| | | Augustus | GeneMark | SNAP |
| *Arabidopsis thaliana* | Nucleotide Accuracy | 77.04% | 74.68% | 69.78% |
| | Exon Accuracy | 67.03% | 61.31% | 56.40% |
| *Drosophila Melanogaster* | Nucleotide Accuracy | 76.08% | 66.54% | 69.29% |
| | Exon Accuracy | 61.37% | 47.31% | 47.01% |
| *Caenorhabditis elegans* | Nucleotide Accuracy | 88.29% | 88.09% | 85.10% |
| | Exon Accuracy | 74.62% | 68.88% | 61.38% |

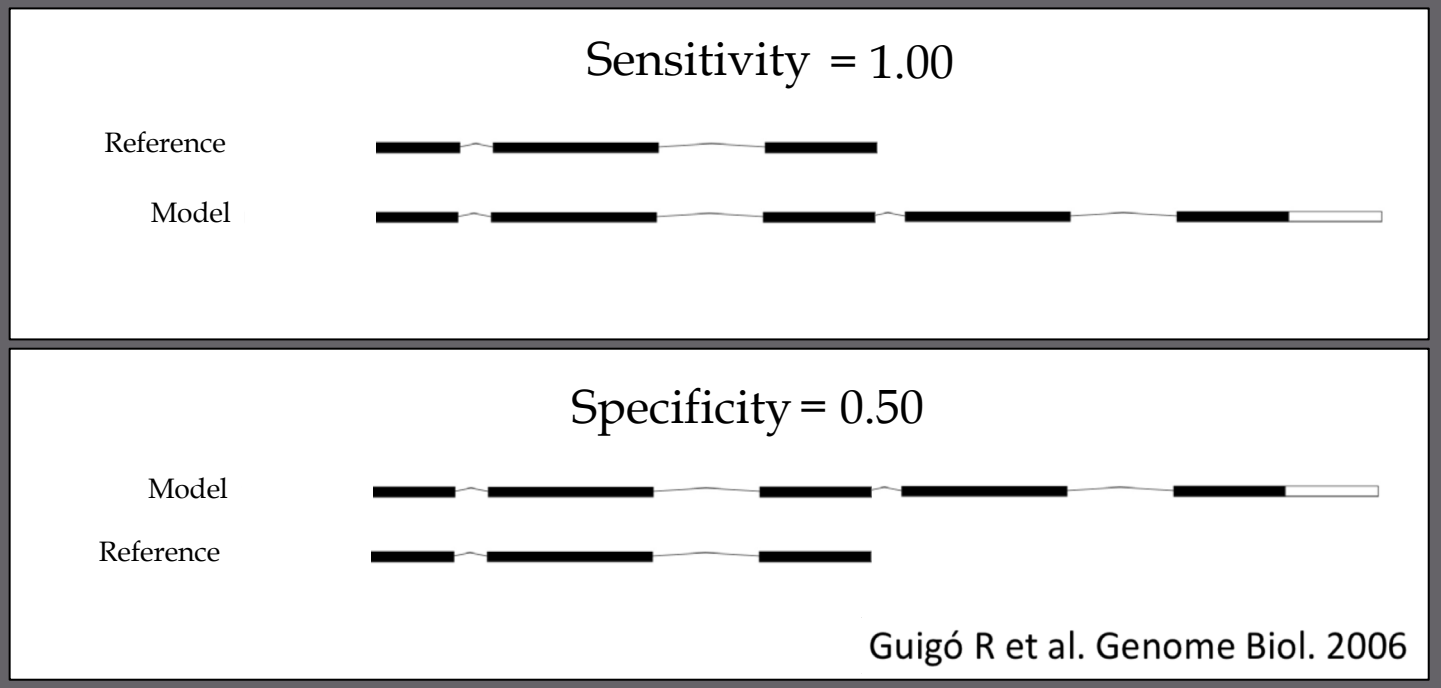
**Gene model accuracies when using unmatched species parameter files**

| Reference Organism | Performance Category | *Ab Initio* Prediction | | | MAKER Annotations |
|---|---|---|---|---|---|
| | | Augustus | GeneMark | SNAP | SNAP |
| *Arabidopsis thaliana* | Nucleotide Accuracy | 57.85% | 48.62% | 43.84% | 73.77% |
| | Exon Accuracy | 30.71% | 16.51% | 18.58% | 60.11% |
| *Drosophila Melanogaster* | Nucleotide Accuracy | 67.47% | 66.51% | 48.92% | 74.44% |
| | Exon Accuracy | 30.62% | 26.25% | 19.94% | 53.69% |
| *Caenorhabditis elegans* | Nucleotide Accuracy | 66.18% | 67.26% | 68.24% | 85.02% |
| | Exon Accuracy | 28.33% | 30.01% | 35.44% | 63.14% |

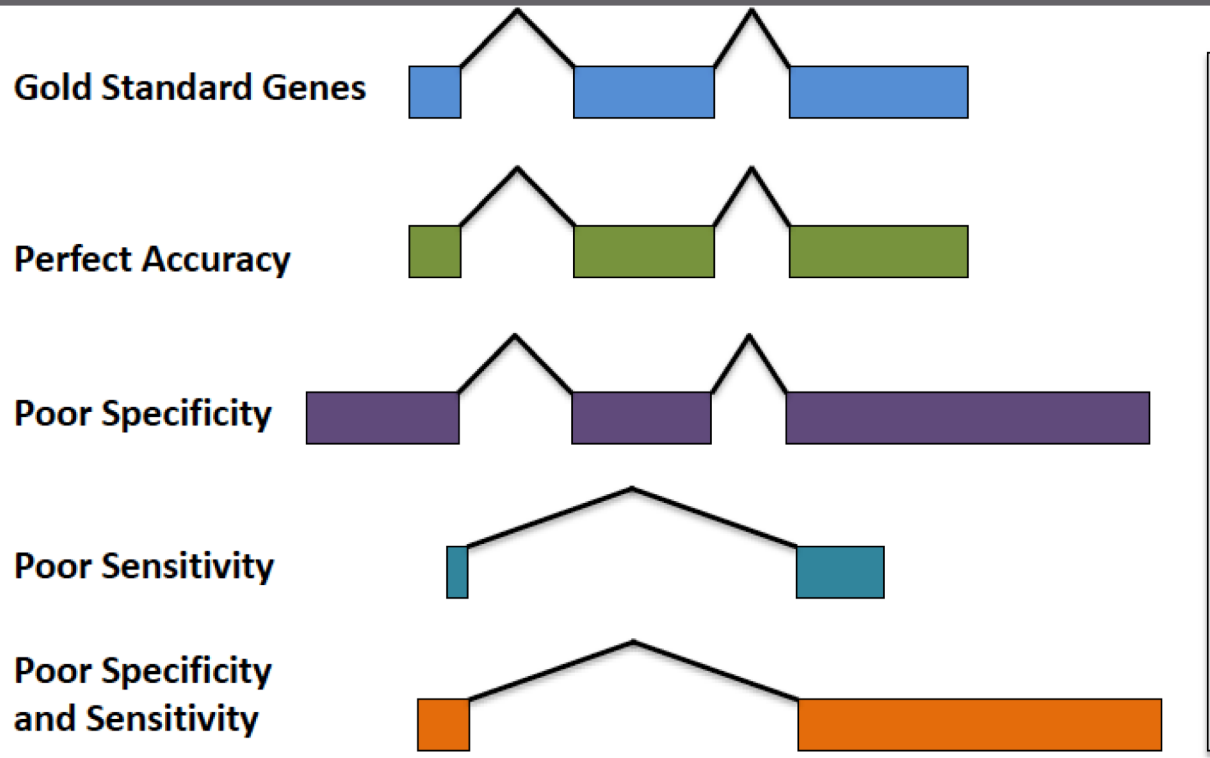
# Beyond *de novo* annotation

- Quality control and data prioritization
- Update/revise legacy annotation sets
- Integrating new evidence into existing databases

# Quality control and data prioritization



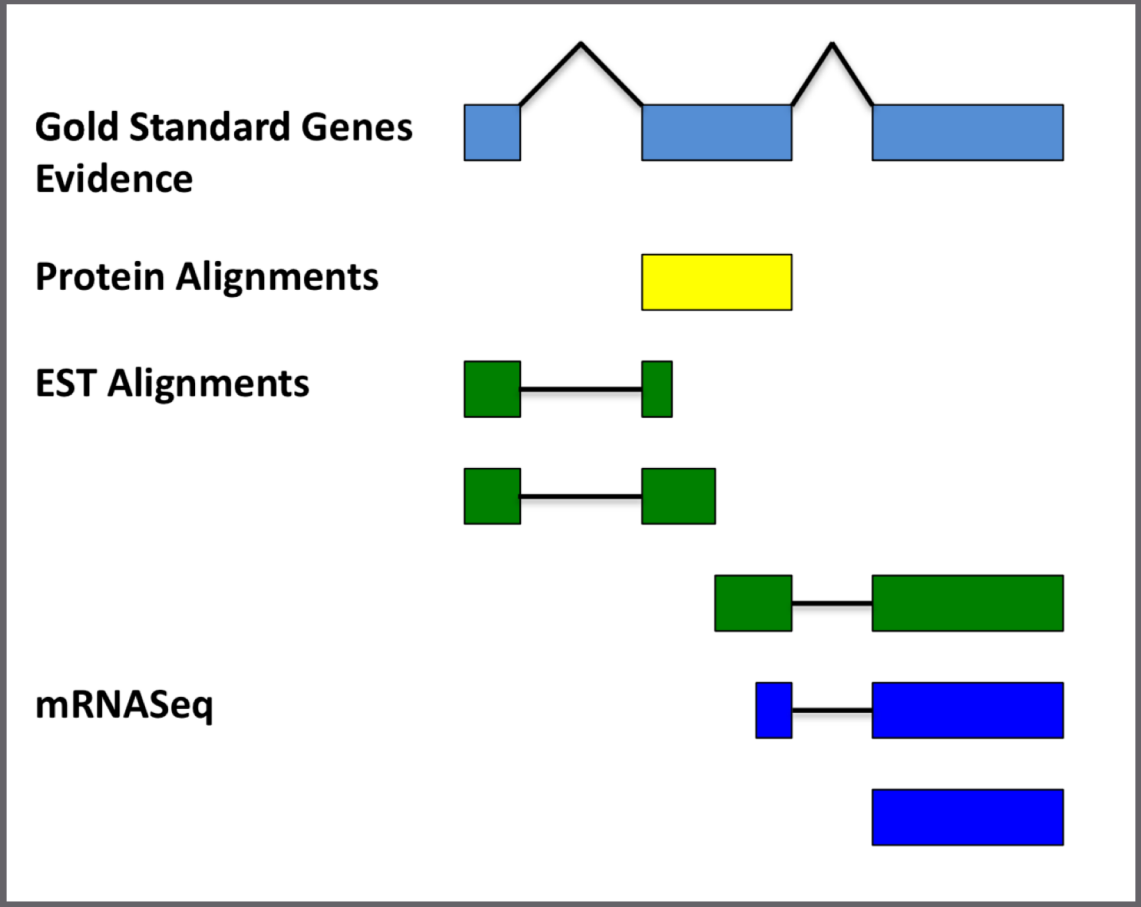
# Quality control and data prioritization




| | SN | SP | AC |
|---|---|---|---|
| Gold Standard Genes | | | |
| Perfect Accuracy | 1.0 | 1.0 | 100% |
| Poor Specificity | 1.0 | 0.5 | 75% |
| Poor Sensitivity | 0.5 | 1.0 | 75% |
| Poor Specificity and Sensitivity | 0.5 | 0.5 | 50% |

Guigó R et al. Genome Biol. 2006

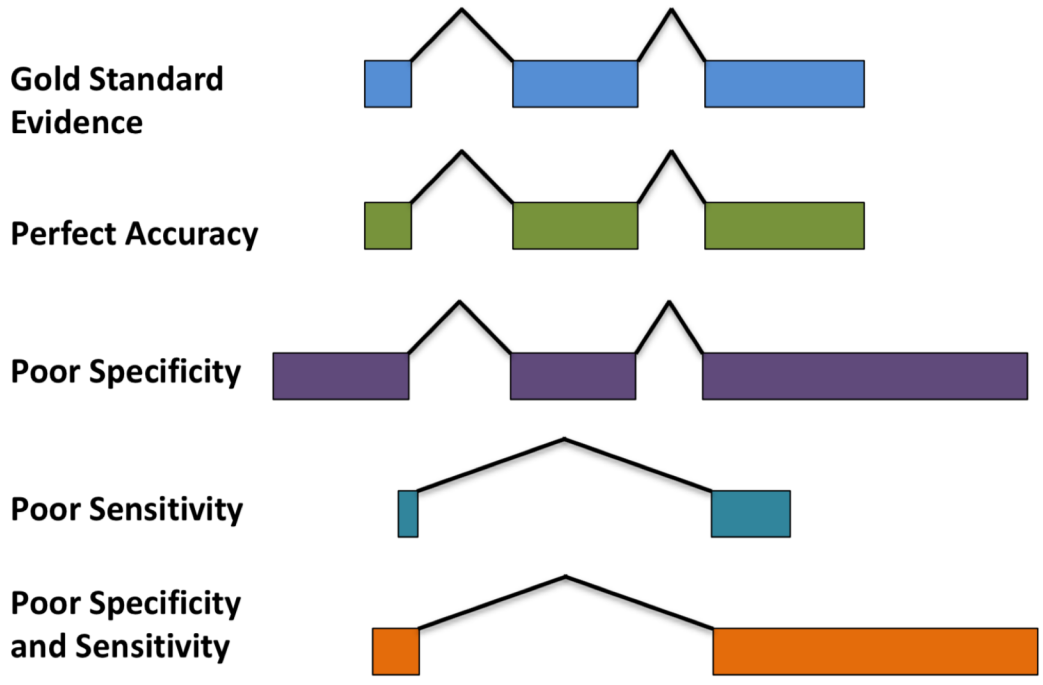
# Quality control and data prioritization




*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

# Quality control and data prioritization

$$AED = 1 - \frac{(SN + SP)}{2}$$

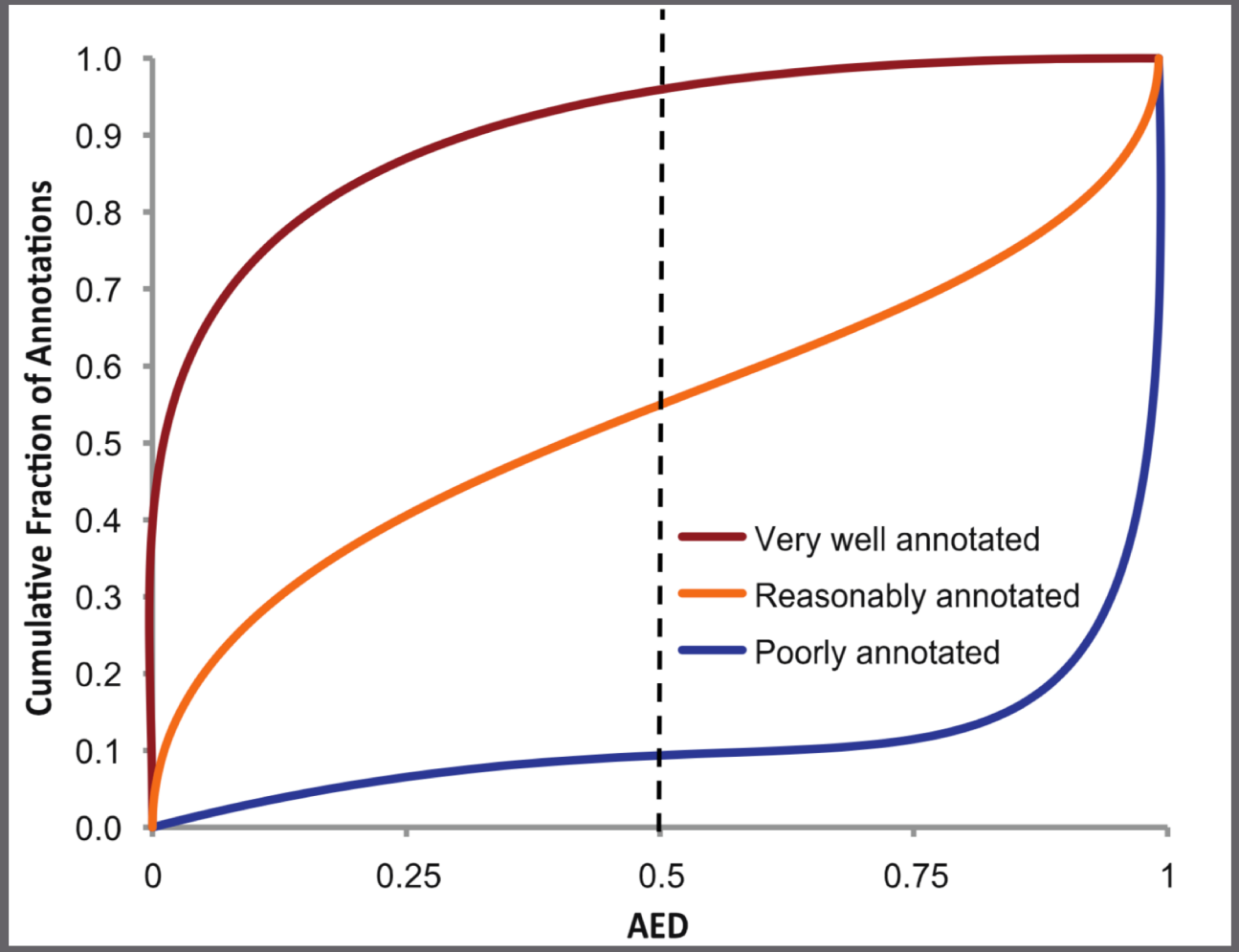



| SN | SP | AED |
|---|---|---|
| 1.0 | 1.0 | **0.0** |
| 1.0 | 0.5 | **0.25** |
| 0.5 | 1.0 | **0.25** |
| 0.5 | 0.5 | **0.5** |

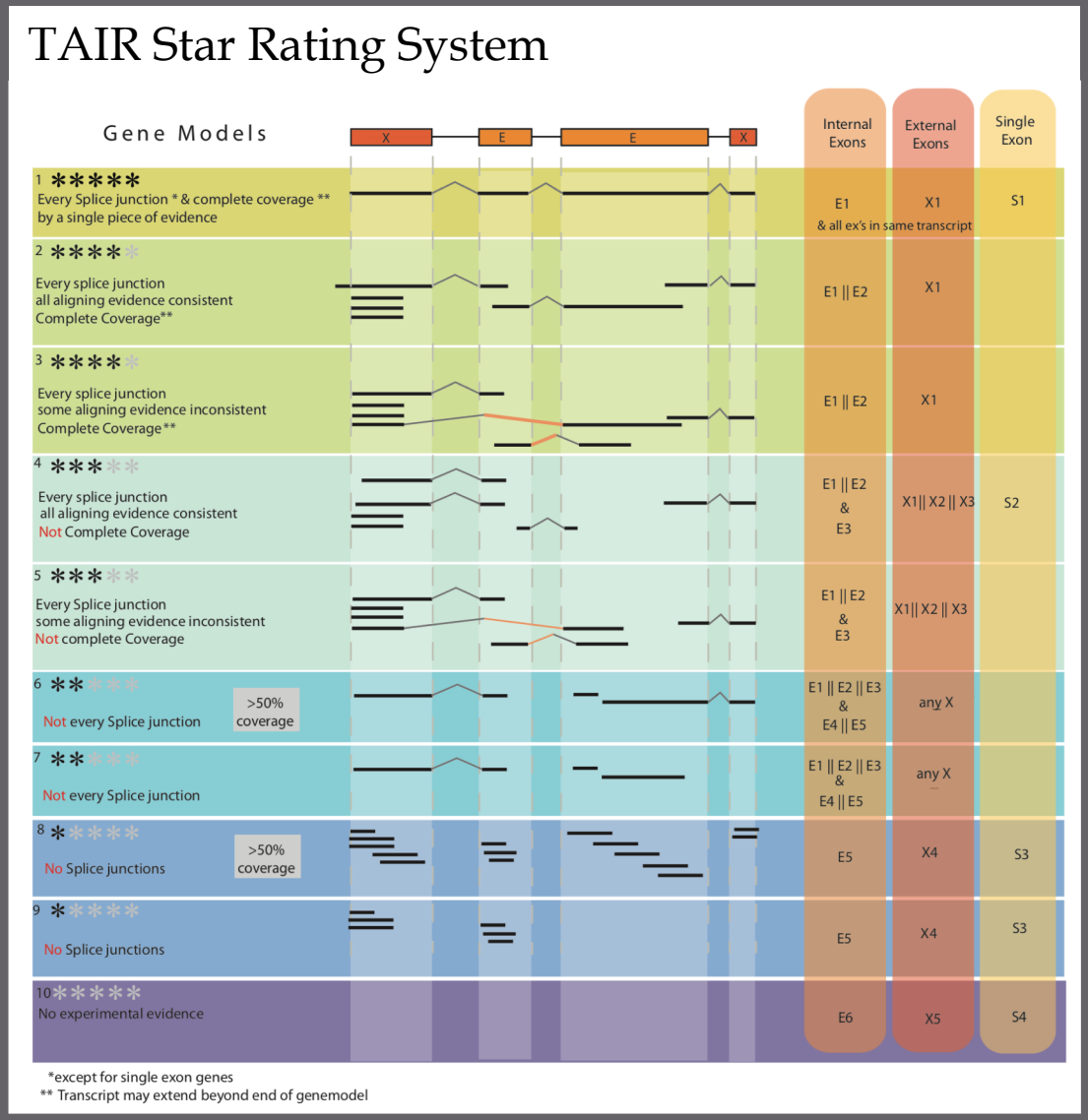
Eilbeck et al BMC Bioinformatics 2009

***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

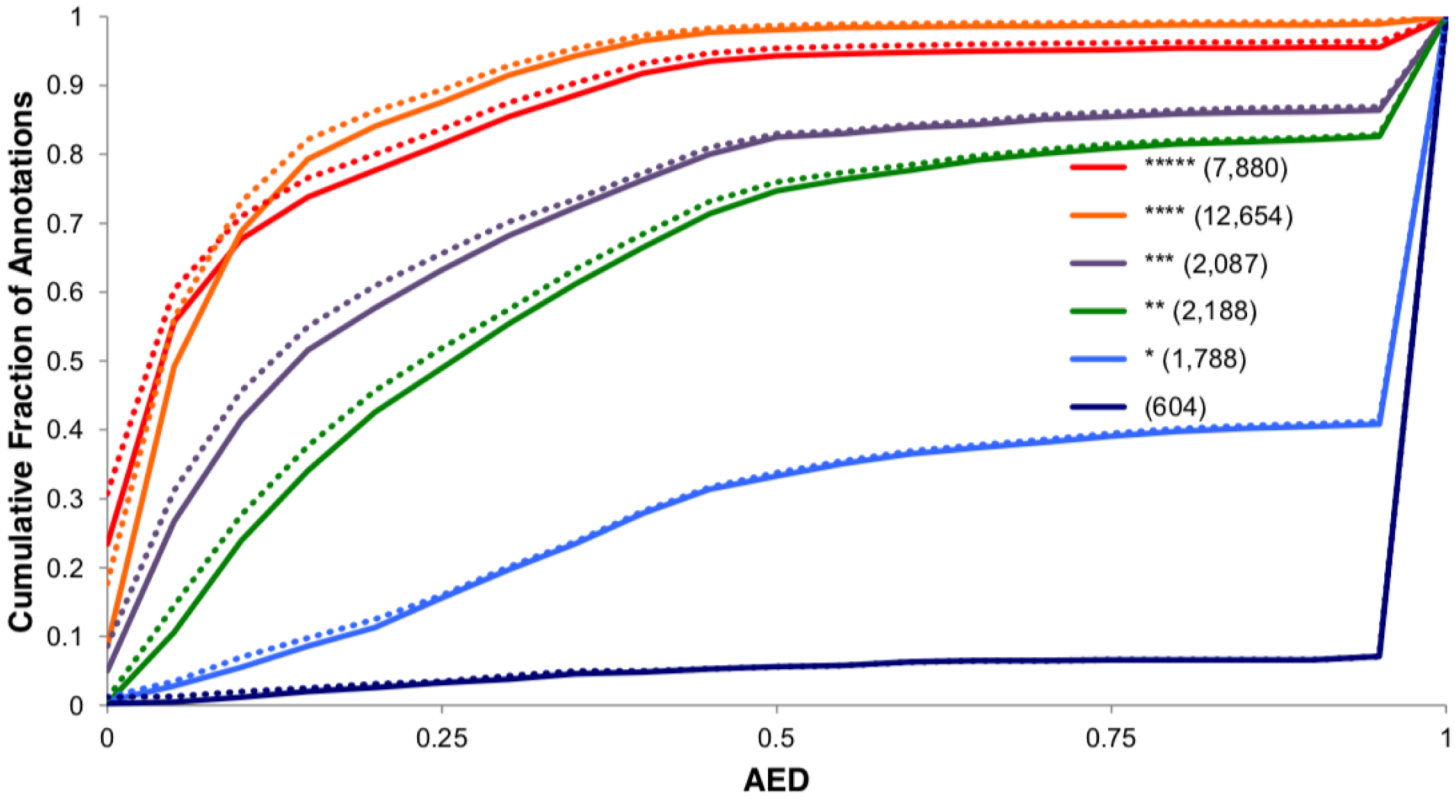
# Quality control and data prioritization



# Evaluation of AED on Arabidopsis

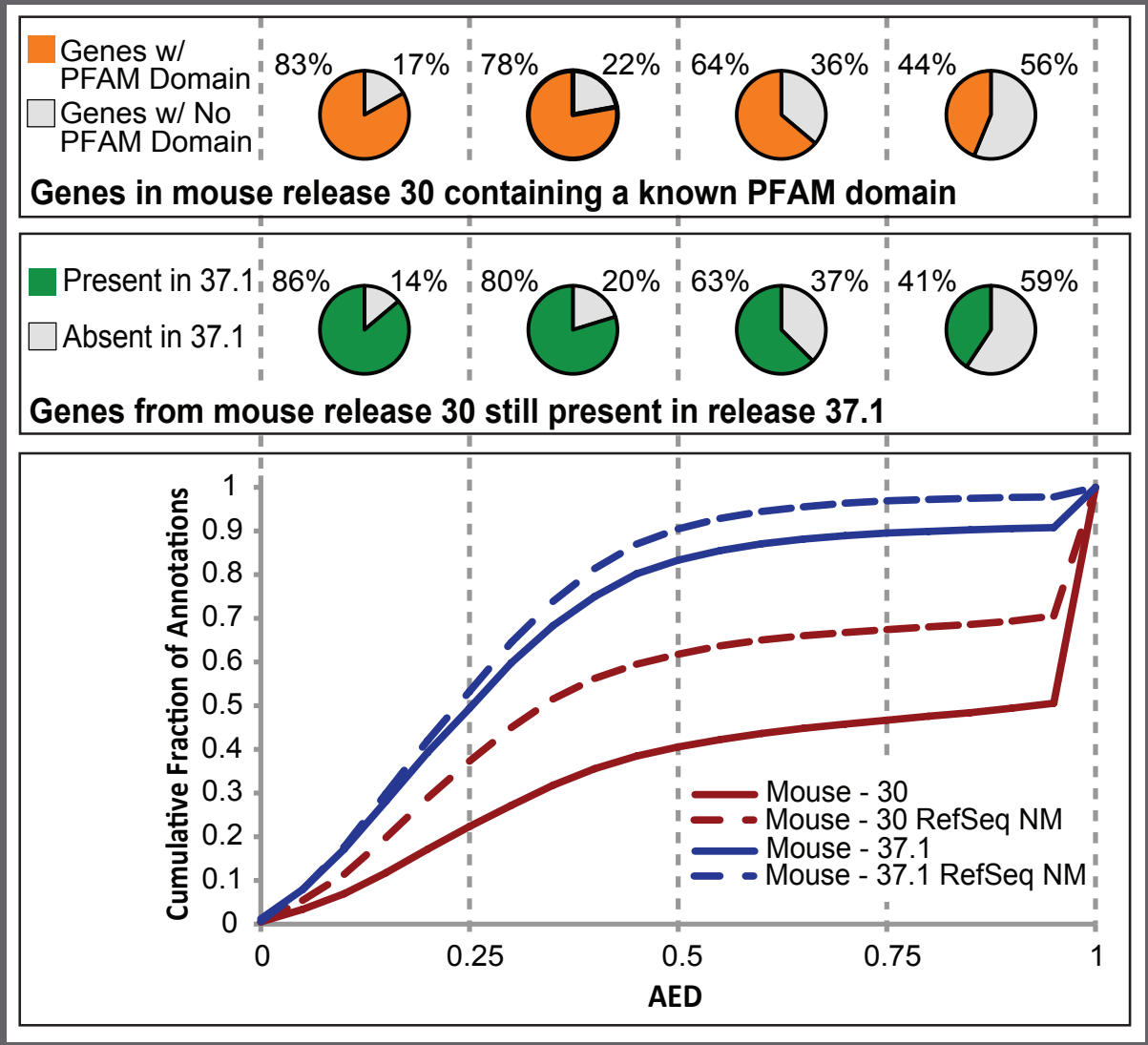


# Evaluation of AED on Arabidopsis

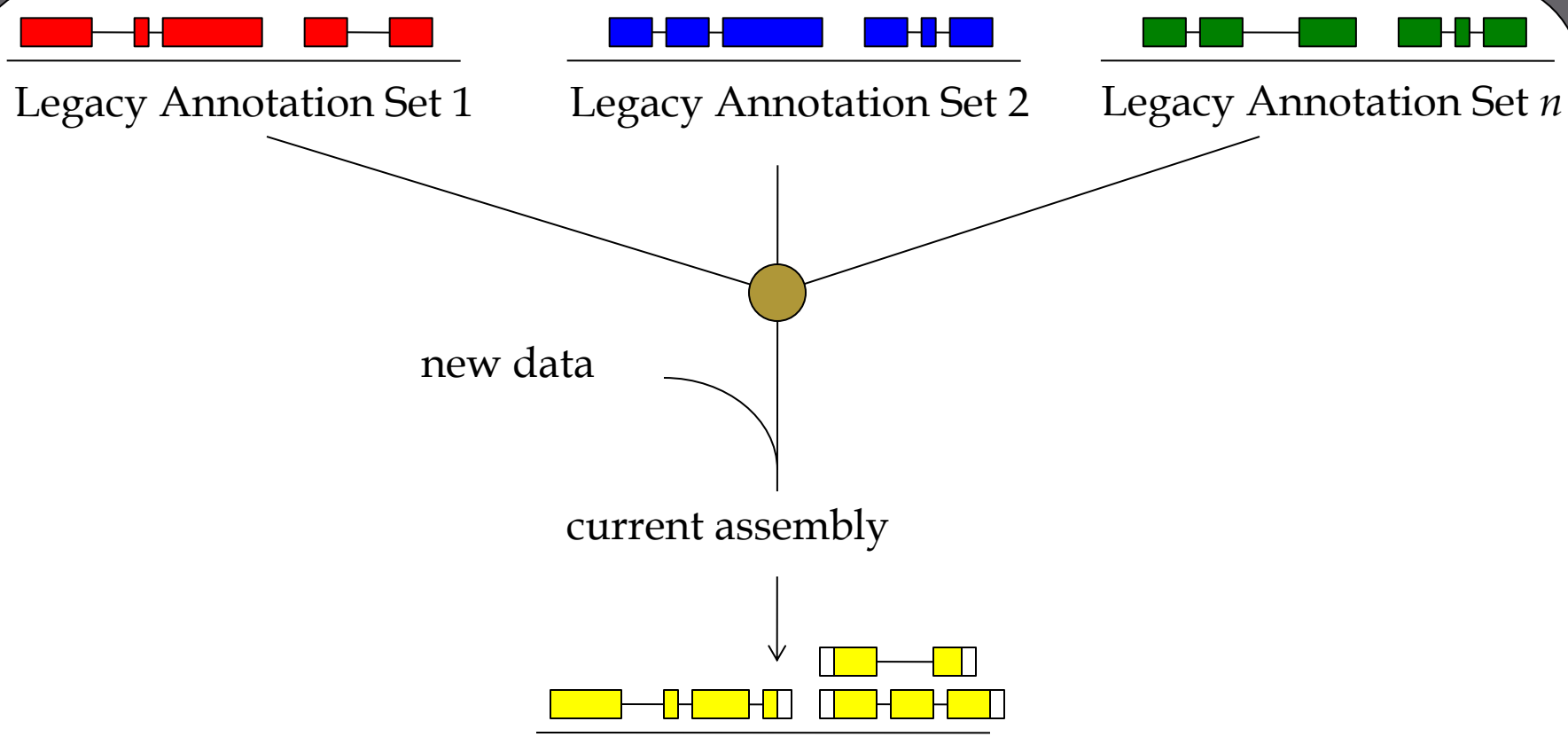



Evidence: mRNA-seq (17 experiments), ESTs, full length cDNAs, Swiss-Prot (minus Arabidopsis)

# Evaluation of AED on the mouse genome

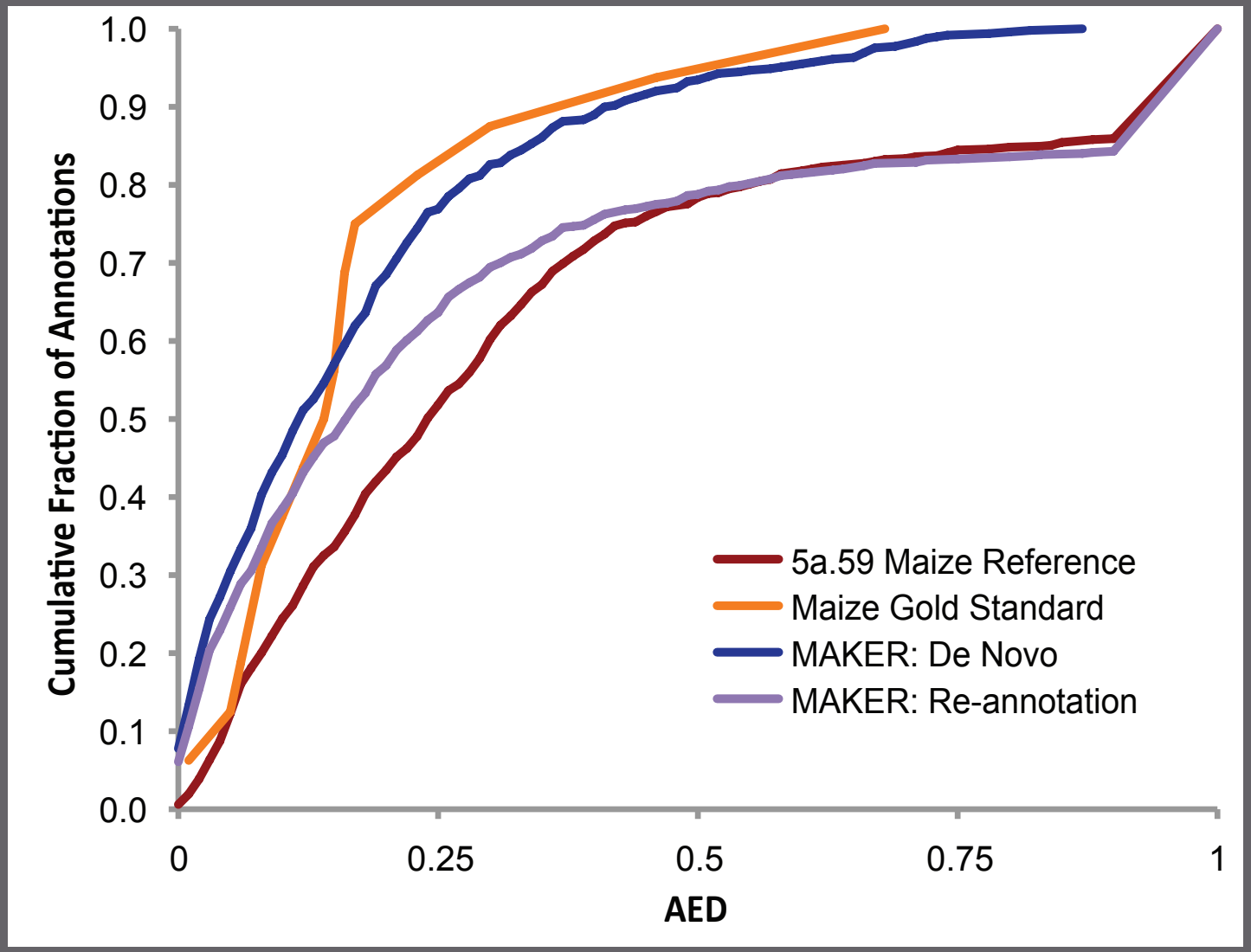


# Update/revise legacy annotation sets

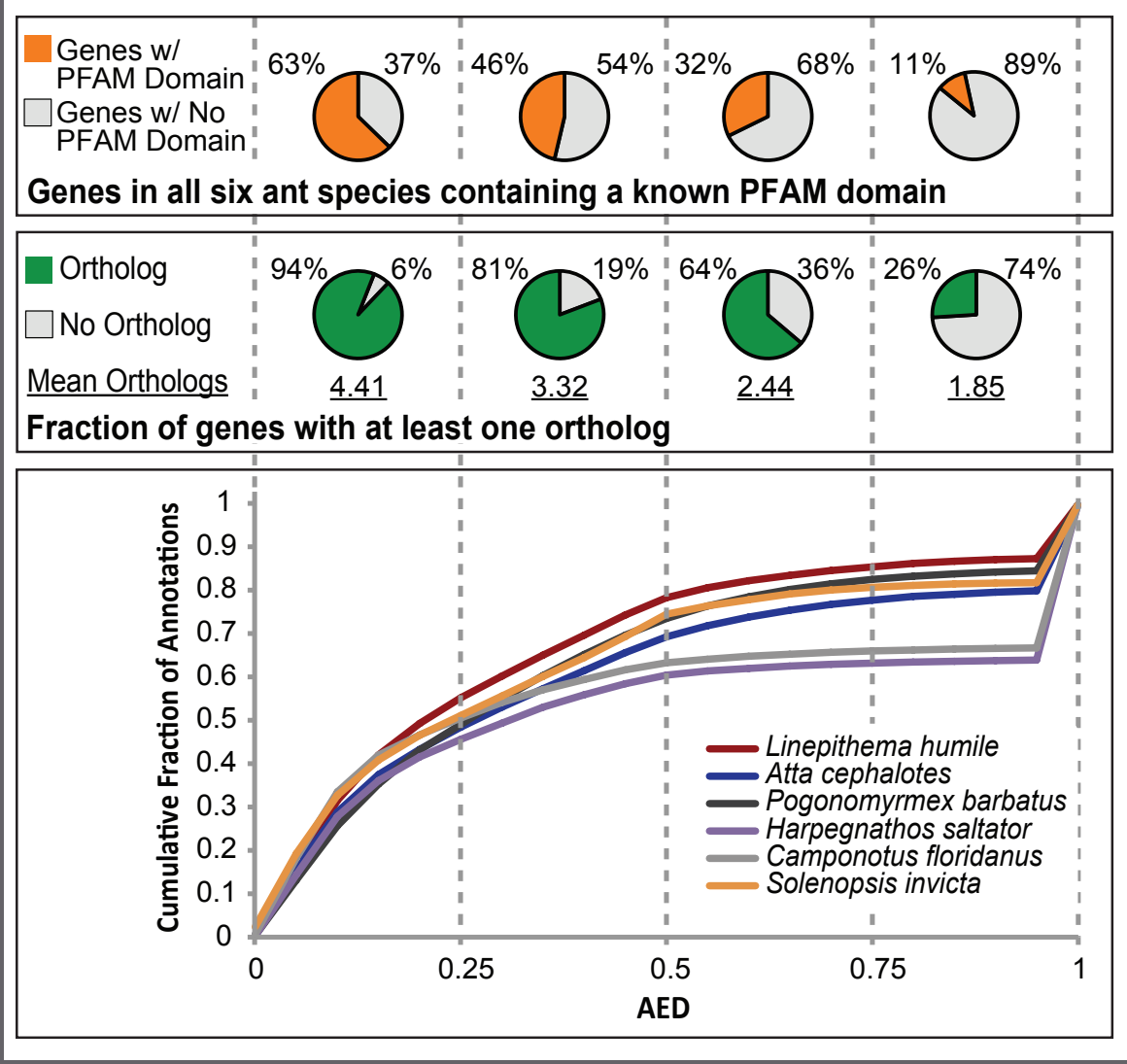



- Identify legacy annotation most consistent with new data
- Automatically revise it in light of new data
- If no existing annotation, create new one

# Re-annotation of the maize genome



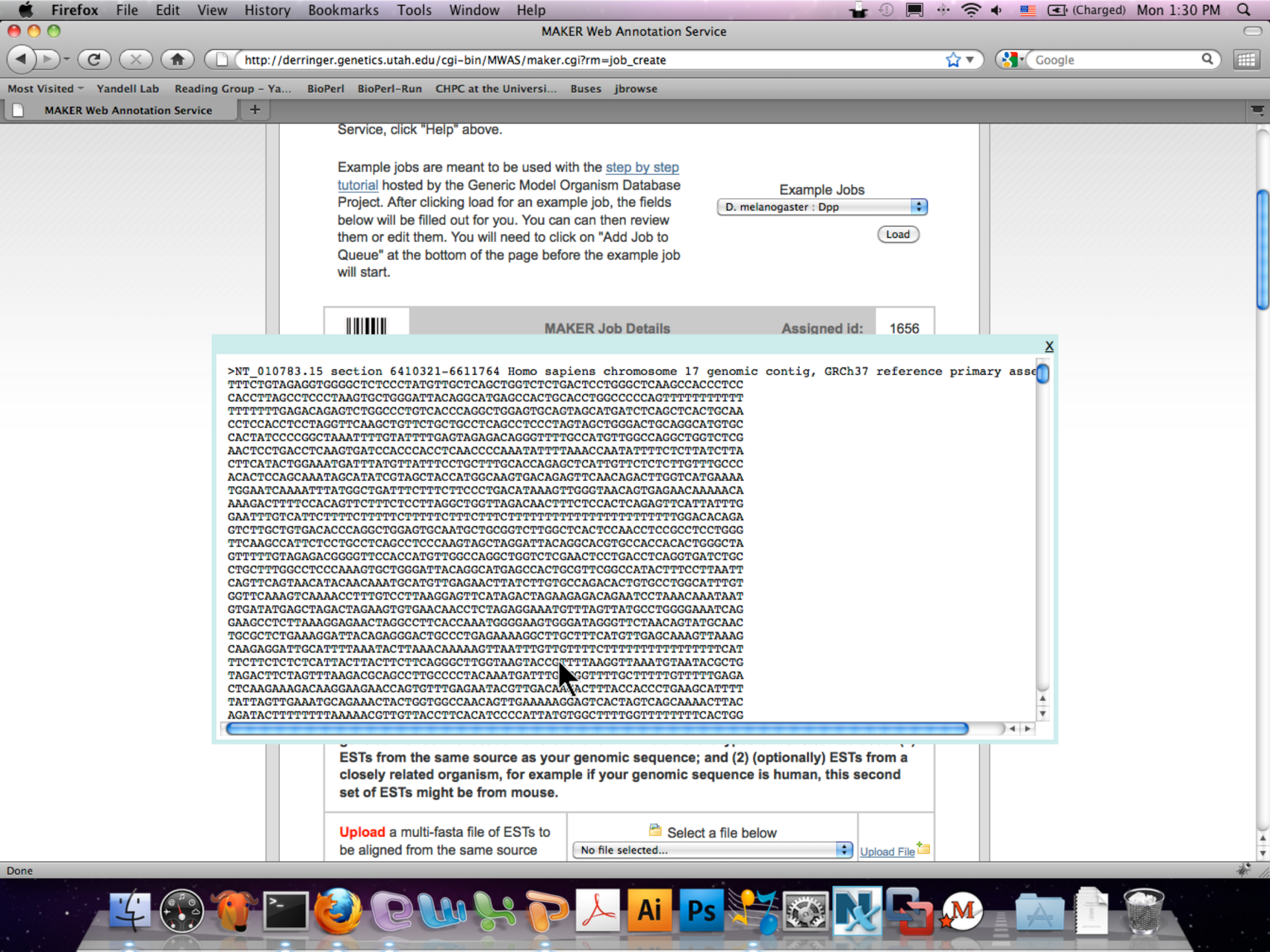
# Integrating new evidence into existing databases

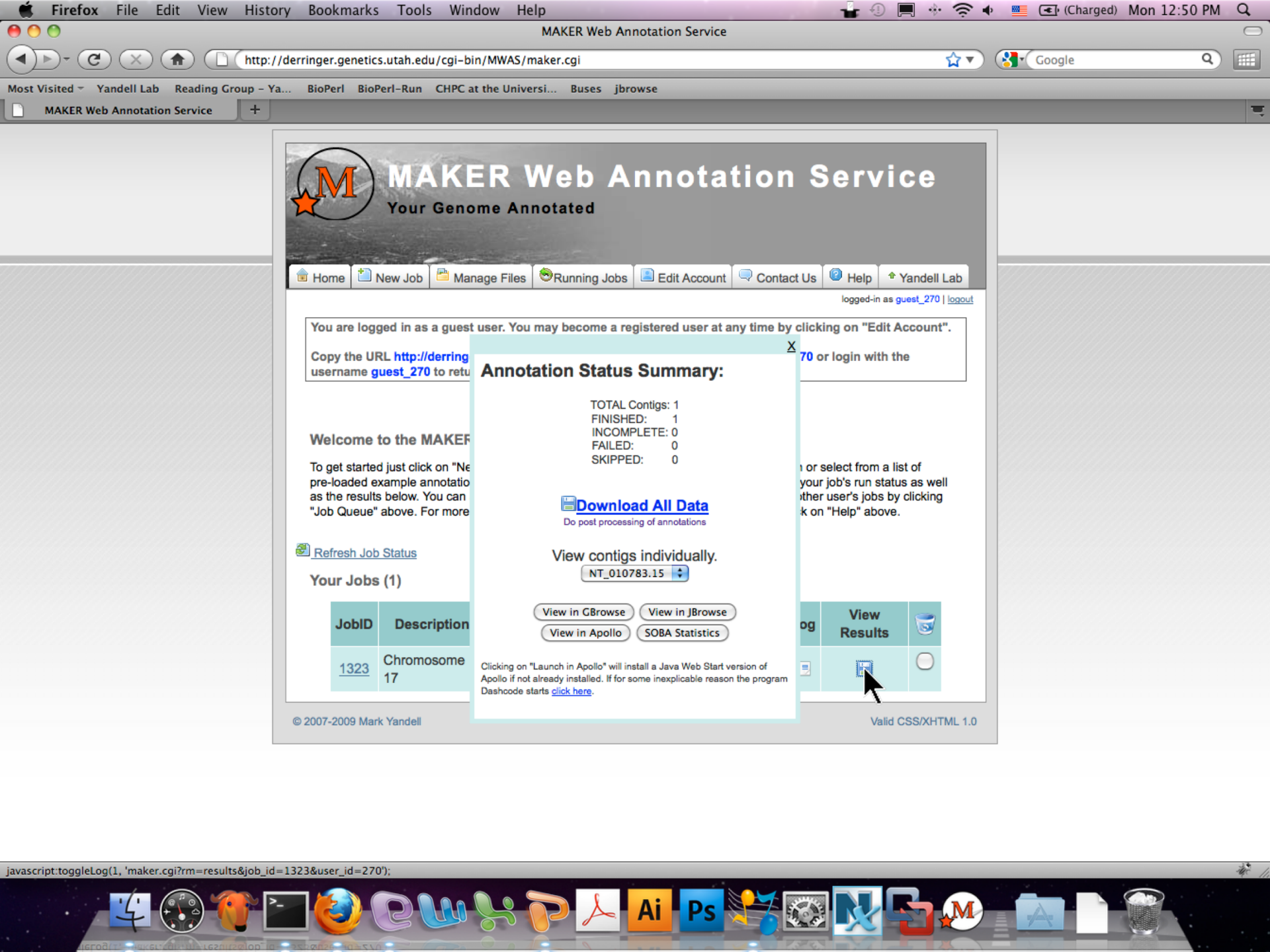


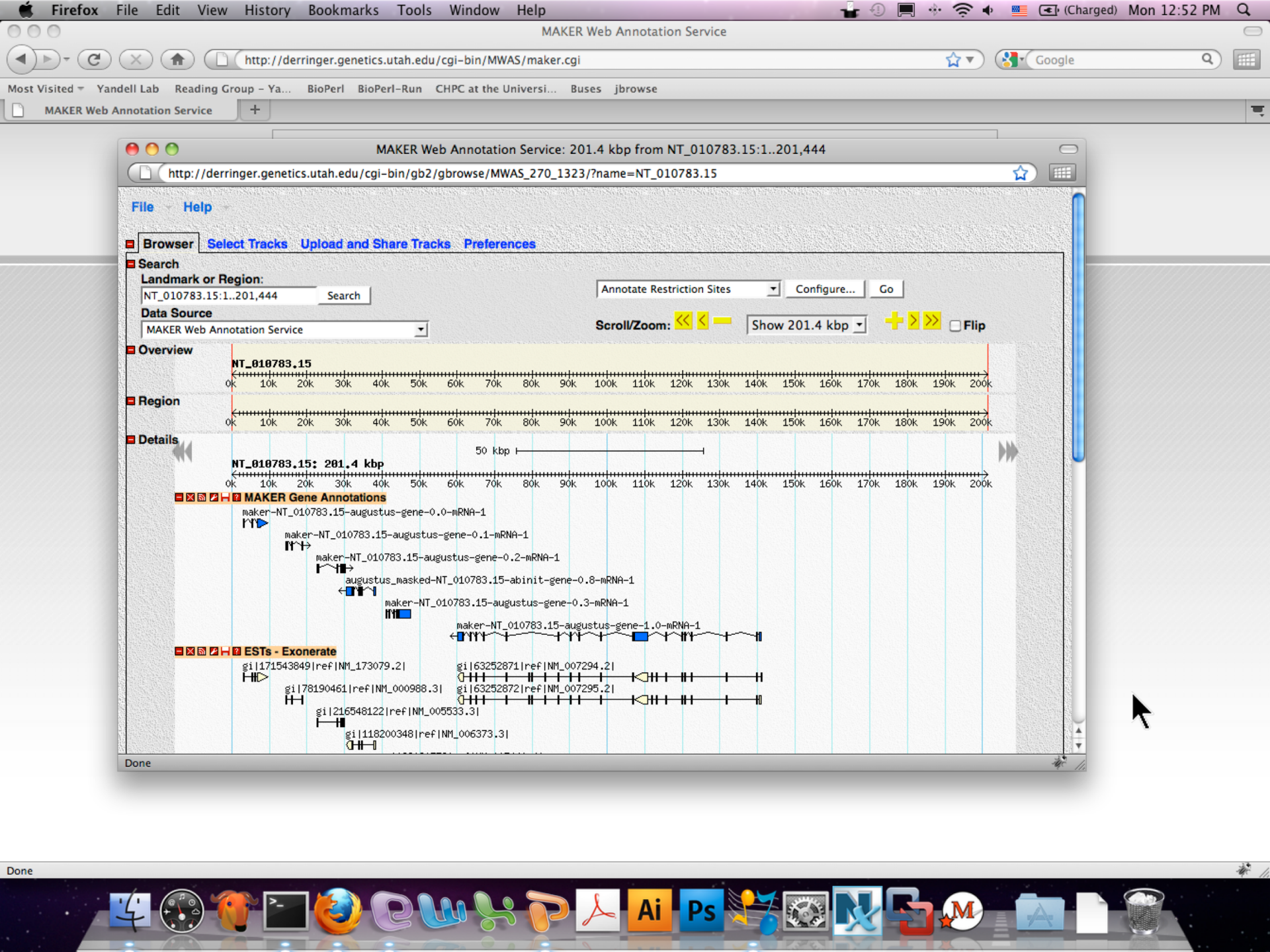
# MAKER Versions

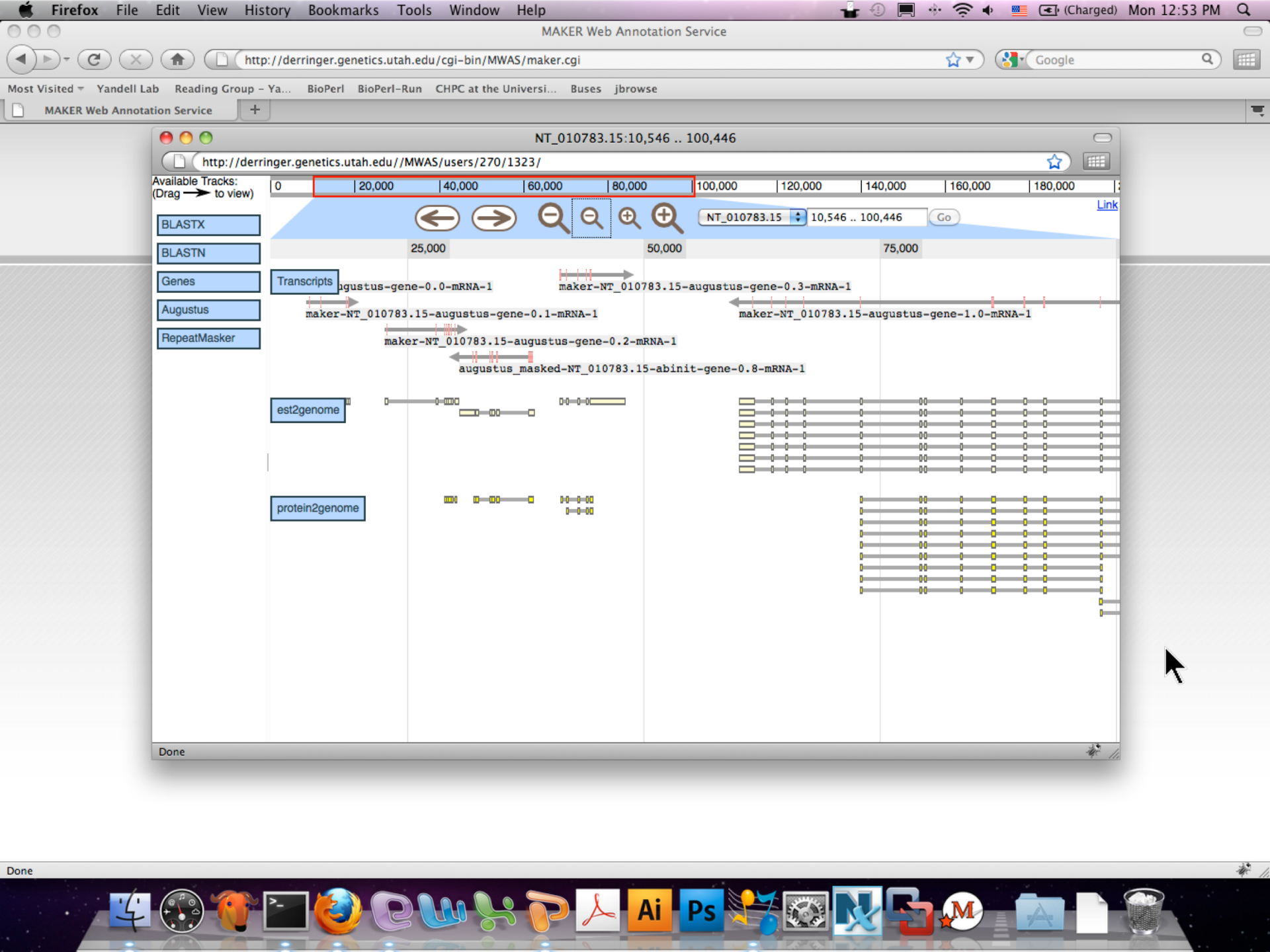
- MAKER
  - 2008. Based on earlier annotation pipelines developed by Mark Yandell
- MAKER 2 / MAKER-P
  - 2011. Introduction of MPI parallelization, support for multiple gene predictors, GFF3 pass-through, and quality metrics like AED (Annotation Edit Distance) from the Sequence Ontology consortium.
  - 2015. Support for tRNA and snoRNA annotation. Improved parallelization on large plant genomes.
- MAKER 3
  - 2016. EVM (Evidence Modeler) support for improved annotation and user defined evidence probability weighting.

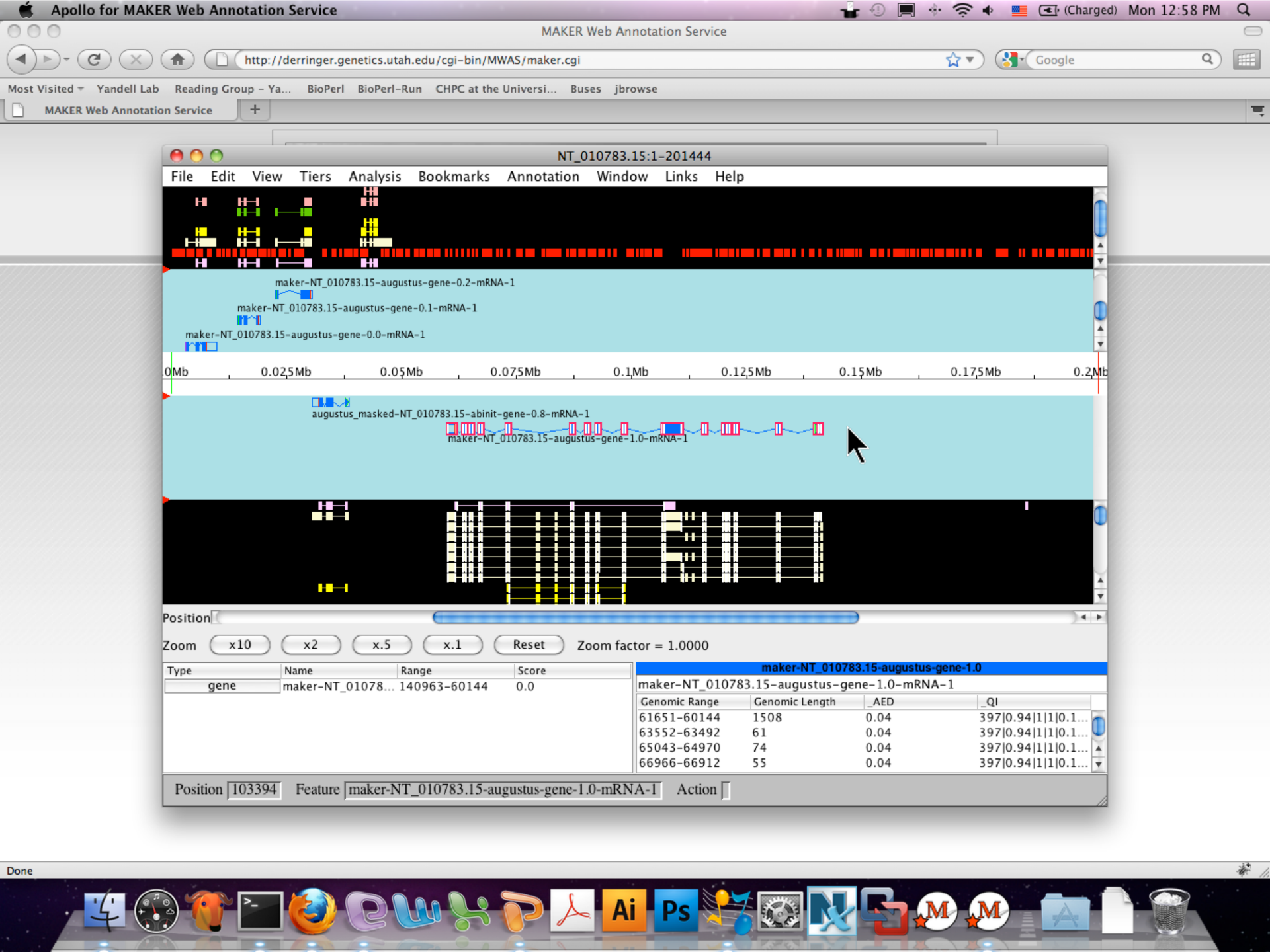

Firefox File Edit View History Bookmarks Tools Window Help
MAKER Web Annotation Service
http://derringer.genetics.utah.edu/cgi-bin/MWAS/maker.cgi
Most Visited Yandell Lab Reading Group - Ya... BioPerl BioPerl-Run CHPC at the Universi... Buses jbrowse
MAKER Web Annotation Service
MAKER Web Annotation Service
Your Genome Annotated
Home Help Yandell Lab
not logged-in | sign in
Welcome to the MAKER Web Annotation Service
Log into your account below, or you can access the server as a guest. While there is no login requirement for this site, users are highly encouraged to set up an account. Use the "New user registration" link to register a new account. Registration is free, and has several benefits. Registered users can submit up to 5,000,000 base pairs of sequence for each annotation job. Guest users are limited to 500,000 base pairs per annotation job submission. Registered users
http://www.yandell-lab.org/
Maker Web Annotation Service
User Name
Password
Remember User Name
New user registration Forgot login? Help
New Guest Account
User Sign In
© 2007-2009 Mark Yandell
Valid CSS/XHTML 1.0
Done


Firefox File Edit View History Bookmarks Tools Window Help
(Charged) Mon 1:30 PM
MAKER Web Annotation Service
http://derringer.genetics.utah.edu/cgi-bin/MWAS/maker.cgi?rm=job_create
Google
Most Visited Yandell Lab Reading Group - Ya... BioPerl BioPerl-Run CHPC at the Universi... Buses jbrowse
MAKER Web Annotation Service
Service, click "Help" above.
Example jobs are meant to be used with the step by step tutorial hosted by the Generic Model Organism Database Project. After clicking load for an example job, the fields below will be filled out for you. You can can then review them or edit them. You will need to click on "Add Job to Queue" at the bottom of the page before the example job will start.
Example Jobs
D. melanogaster : Dpp
Load
MAKER Job Details
Assigned id:
1656
X
>NT_010783.15 section 6410321-6611764 Homo sapiens chromosome 17 genomic contig, GRCh37 reference primary asse
TTTCTGTAGAGGTGGGGCTCTCCCTATGTTGCTCAGCTGGTCTCTGACTCCTGGGCTCAAGCCACCCTCC
CACCTTAGCCTCCCTAAGTGCTGGGATTACAGGCATGAGCCACTGCACCTGGCCCCCAGTTTTTTTTTTT
TTTTTTTGAGACAGAGTCTGGCCCTGTCACCCAGGCTGGAGTGCAGTAGCATGATCTCAGCTCACTGCAA
CCTCCACCTCCTAGGTTCAAGCTGTTCTGCTGCCTCAGCCTCCCTAGTAGCTGGGACTGCAGGCATGTGC
CACTATCCCCGGCTAAATTTTGTATTTTGAGTAGAGACAGGGTTTTGCCATGTTGGCCAGGCTGGTCTCG
AACTCCTGACCTCAAGTGATCCACCCACCTCAACCCCAAATATTTTAAACCAATATTTTCTCTTATCTTA
CTTCATACTGGAAATGATTTATGTTATTTCCTGCTTTGCACCAGAGCTCATTGTTCTCTCTTGTTTGCCC
ACACTCCAGCAAATAGCATATCGTAGCTACCATGGCAAGTGACAGAGTTCAACAGACTTGGTCATGAAAA
TGGAATCAAAATTTATGGCTGATTTCTTTCTTCCCTGACATAAAGTTGGGTAACAGTGAGAACAAAAACA
AAAGACTTTTCCACAGTTCTTTCTCCTTAGGCTGGTTAGACAACTTTCTCCACTCAGAGTTCATTATTTG
GAATTTGTCATTCTTTTCTTTTTCTTTTTCTTTCTTTCTTTTTTTTTTTTTTTTTTTTTTTGGACACAGA
GTCTTGCTGTGACACCCAGGCTGGAGTGCAATGCTGCGGTCTTGGCTCACTCCAACCTCCGCCTCCTGGG
TTCAAGCCATTCTCCTGCCTCAGCCTCCCAAGTAGCTAGGATTACAGGCACGTGCCACCACACTGGGCTA
GTTTTTGTAGAGACGGGGTTCCACCATGTTGGCCAGGCTGGTCTCGAACTCCTGACCTCAGGTGATCTGC
CTGCTTTGGCCTCCCAAAGTGCTGGGATTACAGGCATGAGCCACTGCGTTCGGCCATACTTTCCTTAATT
CAGTTCAGTAACATACAACAAATGCATGTTGAGAACTTATCTTGTGCCAGACACTGTGCCTGGCATTTGT
GGTTCAAAGTCAAAACCTTTGTCCTTAAGGAGTTCATAGACTAGAAGAGACAGAATCCTAAACAAATAAT
GTGATATGAGCTAGACTAGAAGTGTGAACAACCTCTAGAGGAAATGTTTAGTTATGCCTGGGGAAATCAG
GAAGCCTCTTAAAGGAGAACTAGGCCTTCACCAAATGGGGAAGTGGGATAGGGTTCTAACAGTATGCAAC
TGCGCTCTGAAAGGATTACAGAGGGACTGCCCTGAGAAAAGGCTTGCTTTCATGTTGAGCAAAGTTAAAG
CAAGAGGATTGCATTTTAAATACTTAAACAAAAAGTTAATTTGTTGTTTTCTTTTTTTTTTTTTTTTCAT
TTCTTCTCTCTCATTACTTACTTCTTCAGGGCTTGGTAAGTACCGTTTTAAGGTTAAATGTAATACGCTG
TAGACTTCTAGTTTAAGACGCAGCCTTGCCCCTACAAATGATTTGTGGTTTTGCTTTTTGTTTTTGAGA
CTCAAGAAAGACAAGGAAGAACCAGTGTTTGAGAATACGTTGACAAACTTTACCACCCTGAAGCATTTT
TATTAGTTGAAATGCAGAAACTACTGGTGGCCAACAGTTGAAAAAGGAGTCACTAGTCAGCAAAACTTAC
AGATACTTTTTTTTAAAAACGTTGTTACCTTCACATCCCCATTATGTGGCTTTTGGTTTTTTTTCACTGG
ESTs from the same source as your genomic sequence; and (2) (optionally) ESTs from a closely related organism, for example if your genomic sequence is human, this second set of ESTs might be from mouse.
Upload a multi-fasta file of ESTs to be aligned from the same source
Select a file below
No file selected...
Upload File
Done

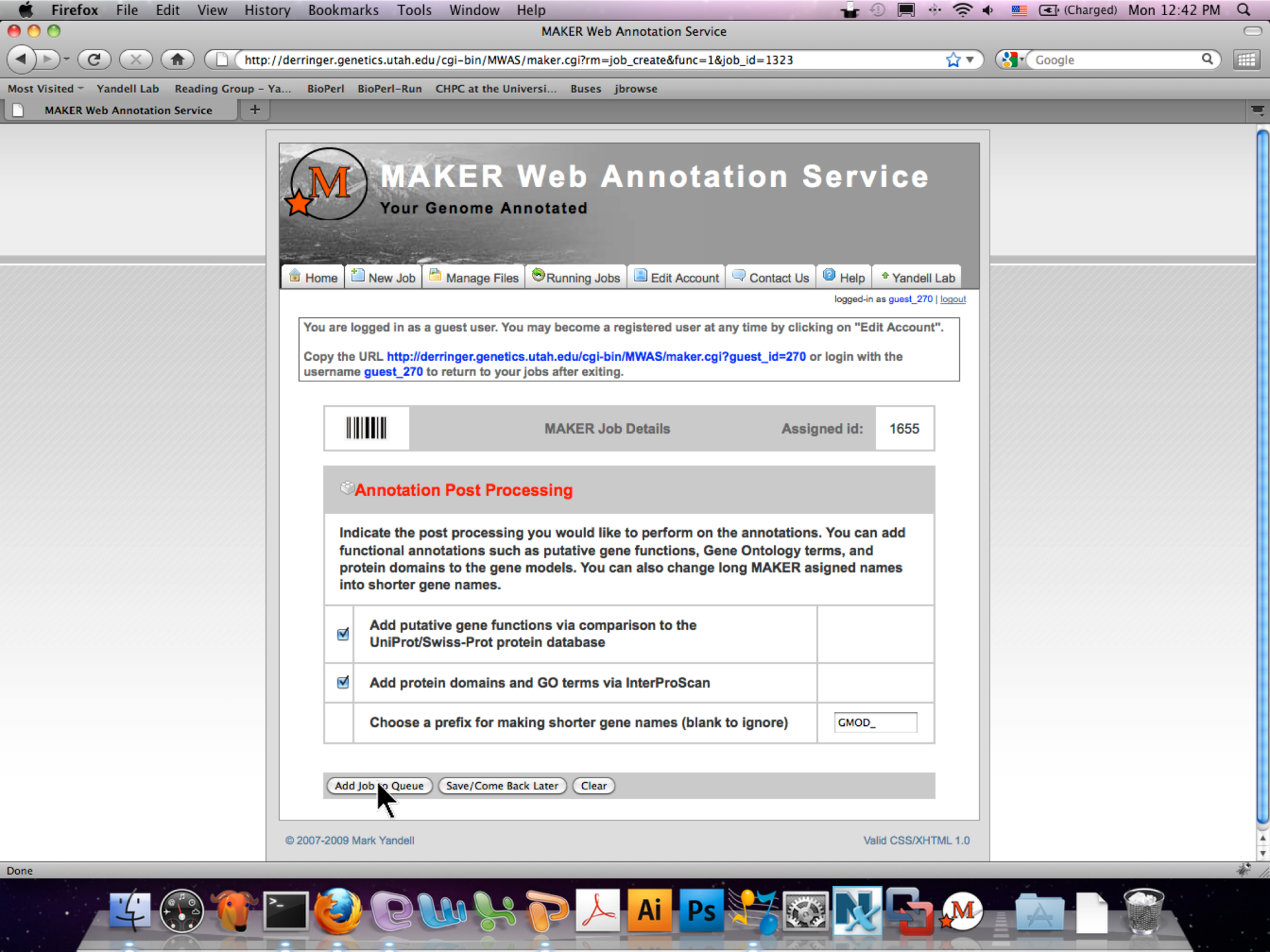

Firefox File Edit View History Bookmarks Tools Window Help
(Charged) Mon 12:50 PM
MAKER Web Annotation Service
http://derringer.genetics.utah.edu/cgi-bin/MWAS/maker.cgi
Google
Most Visited Yandell Lab Reading Group - Ya... BioPerl BioPerl-Run CHPC at the Universi... Buses jbrowse
MAKER Web Annotation Service
MAKER Web Annotation Service
Your Genome Annotated
Home New Job Manage Files Running Jobs Edit Account Contact Us Help Yandell Lab
logged-in as guest_270 | logout
You are logged in as a guest user. You may become a registered user at any time by clicking on "Edit Account".
Copy the URL http://derring 70 or login with the username guest_270 to retu
Welcome to the MAKER
To get started just click on "Ne or select from a list of pre-loaded example annotatio your job's run status as well as the results below. You can other user's jobs by clicking "Job Queue" above. For more k on "Help" above.
Refresh Job Status
Your Jobs (1)
JobID Description og View Results
1323 Chromosome 17
X
Annotation Status Summary:
TOTAL Contigs: 1
FINISHED: 1
INCOMPLETE: 0
FAILED: 0
SKIPPED: 0
Download All Data
Do post processing of annotations
View contigs individually.
NT_010783.15
View in GBrowse View in JBrowse
View in Apollo SOBA Statistics
Clicking on "Launch in Apollo" will install a Java Web Start version of Apollo if not already installed. If for some inexplicable reason the program Dashcode starts click here.
© 2007-2009 Mark Yandell
Valid CSS/XHTML 1.0
javascript:toggleLog(1, 'maker.cgi?rm=results&job_id=1323&user_id=270');

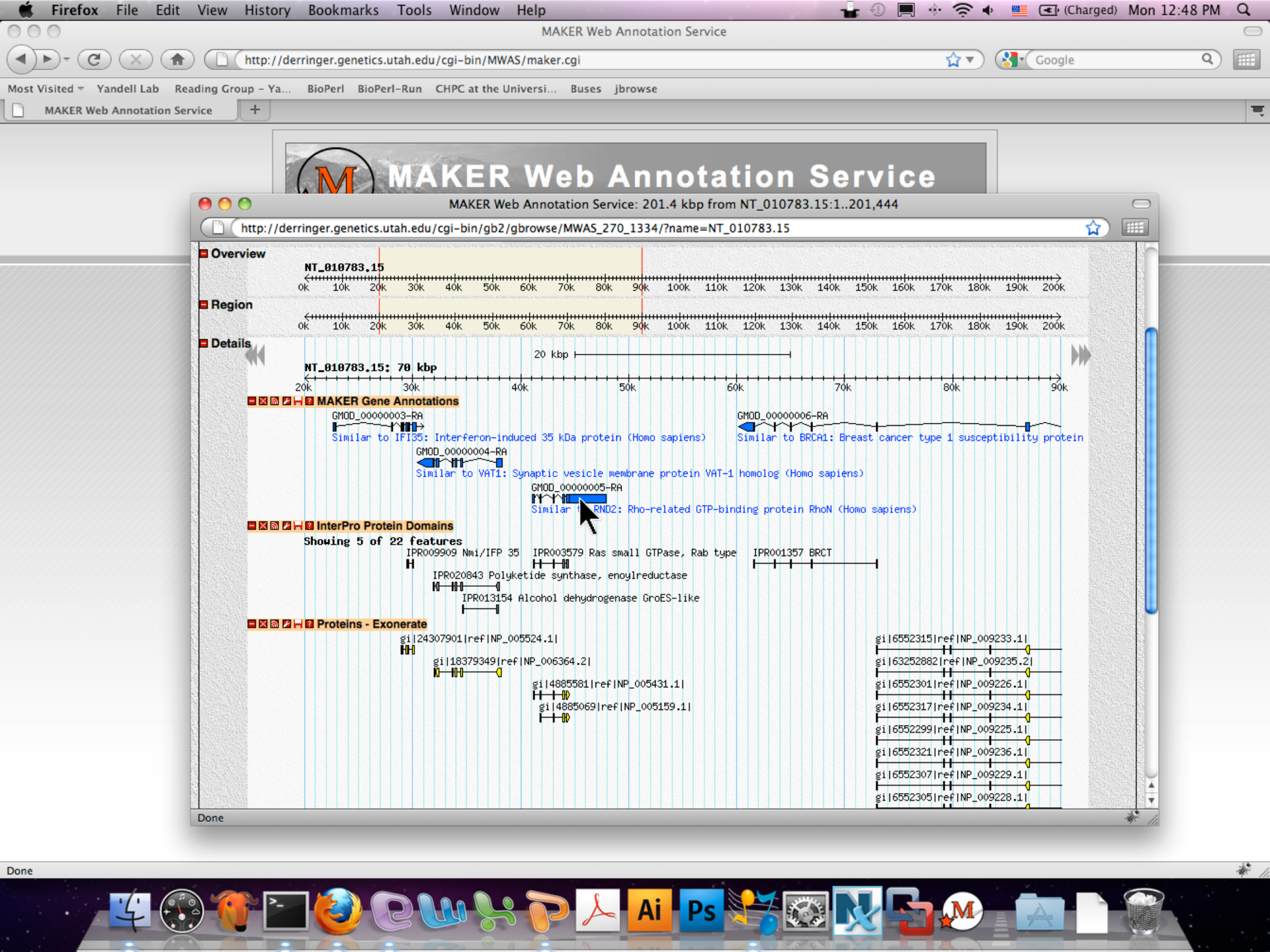

Firefox File Edit View History Bookmarks Tools Window Help
(Charged) Mon 12:52 PM
MAKER Web Annotation Service
http://derringer.genetics.utah.edu/cgi-bin/MWAS/maker.cgi
Google
Most Visited Yandell Lab Reading Group - Ya... BioPerl BioPerl-Run CHPC at the Universi... Buses jbrowse
MAKER Web Annotation Service
MAKER Web Annotation Service: 201.4 kbp from NT_010783.15:1..201,444
http://derringer.genetics.utah.edu/cgi-bin/gb2/gbrowse/MWAS_270_1323/?name=NT_010783.15
File Help
Browser Select Tracks Upload and Share Tracks Preferences
Search
Landmark or Region:
NT_010783.15:1..201,444
Search
Data Source
MAKER Web Annotation Service
Annotate Restriction Sites
Configure...
Go
Scroll/Zoom:
Show 201.4 kbp
Flip
Overview
NT_010783.15
Region
Details
50 kbp
NT_010783.15: 201.4 kbp
MAKER Gene Annotations
maker-NT_010783.15-augustus-gene-0.0-mRNA-1
maker-NT_010783.15-augustus-gene-0.1-mRNA-1
maker-NT_010783.15-augustus-gene-0.2-mRNA-1
augustus_masked-NT_010783.15-abinit-gene-0.8-mRNA-1
maker-NT_010783.15-augustus-gene-0.3-mRNA-1
maker-NT_010783.15-augustus-gene-1.0-mRNA-1
ESTs - Exonerate
gi|171543849|ref|NM_173079.2|
gi|63252871|ref|NM_007294.2|
gi|78190461|ref|NM_000988.3|
gi|63252872|ref|NM_007295.2|
gi|216548122|ref|NM_005533.3|
gi|118200348|ref|NM_006373.3|
Done
Done


Firefox File Edit View History Bookmarks Tools Window Help
(Charged) Mon 12:53 PM
MAKER Web Annotation Service
http://derringer.genetics.utah.edu/cgi-bin/MWAS/maker.cgi
Google
Most Visited Yandell Lab Reading Group - Ya... BioPerl BioPerl-Run CHPC at the Universi... Buses jbrowse
MAKER Web Annotation Service
NT_010783.15:10,546 .. 100,446
http://derringer.genetics.utah.edu//MWAS/users/270/1323/
Available Tracks:
(Drag → to view)
0
20,000
40,000
60,000
80,000
100,000
120,000
140,000
160,000
180,000
Link
NT_010783.15
10,546 .. 100,446
Go
25,000
50,000
75,000
BLASTX
BLASTN
Genes
Augustus
RepeatMasker
Transcripts
ugustus-gene-0.0-mRNA-1
maker-NT_010783.15-augustus-gene-0.3-mRNA-1
maker-NT_010783.15-augustus-gene-0.1-mRNA-1
maker-NT_010783.15-augustus-gene-1.0-mRNA-1
maker-NT_010783.15-augustus-gene-0.2-mRNA-1
augustus_masked-NT_010783.15-abinit-gene-0.8-mRNA-1
est2genome
protein2genome
Done
Done


Apollo for MAKER Web Annotation Service
Mon 12:58 PM
MAKER Web Annotation Service
http://derringer.genetics.utah.edu/cgi-bin/MWAS/maker.cgi
Most Visited
Yandell Lab
Reading Group - Ya...
BioPerl
BioPerl-Run
CHPC at the Universi...
Buses
jbrowse
MAKER Web Annotation Service
NT_010783.15:1-201444
File
Edit
View
Tiers
Analysis
Bookmarks
Annotation
Window
Links
Help
maker-NT_010783.15-augustus-gene-0.2-mRNA-1
maker-NT_010783.15-augustus-gene-0.1-mRNA-1
maker-NT_010783.15-augustus-gene-0.0-mRNA-1
0Mb
0.025Mb
0.05Mb
0.075Mb
0.1Mb
0.125Mb
0.15Mb
0.175Mb
0.2Mb
augustus_masked-NT_010783.15-abinit-gene-0.8-mRNA-1
maker-NT_010783.15-augustus-gene-1.0-mRNA-1
Position
Zoom
x10
x2
x.5
x.1
Reset
Zoom factor = 1.0000
Type
Name
Range
Score
gene
maker-NT_01078...
140963-60144
0.0
maker-NT_010783.15-augustus-gene-1.0
maker-NT_010783.15-augustus-gene-1.0-mRNA-1
Genomic Range
Genomic Length
_AED
_QI
61651-60144
1508
0.04
397|0.94|1|1|0.1...
63552-63492
61
0.04
397|0.94|1|1|0.1...
65043-64970
74
0.04
397|0.94|1|1|0.1...
66966-66912
55
0.04
397|0.94|1|1|0.1...
Position
103394
Feature
maker-NT_010783.15-augustus-gene-1.0-mRNA-1
Action
Done


Firefox File Edit View History Bookmarks Tools Window Help
(Charged) Mon 12:56 PM
MAKER Web Annotation Service
http://derringer.genetics.utah.edu/cgi-bin/MWAS/maker.cgi
Google
Most Visited Yandell Lab Reading Group - Ya... BioPerl BioPerl-Run CHPC at the Universi... Buses jbrowse
MAKER Web Annotation Service
SOBA
http://www.sequenceontology.org/cgi-bin/soba.cgi?rm=reload_files&gff_file=NT_010783.15.gff
SOBA
Sequence Ontology Bioinformatic Analysis
Home Browser Wiki GFF3 GVF Resources About Request A Term Site Map
Genome Summary
For file(s): NT_010783.15.gff
Below are the feature types and sources for the GFF3 file(s) you uploaded. You must select at least one feature type and source and then click on the headers below to view the anaylses.
Feature Types
Select All Unselect All Invert Selection
CDS
contig
exon
expressed_sequence_match
gene
match
match_part
mRNA
protein_match
Select All Unselect All Invert Selection
Term Validation:
Term usage is OK!
Sources
Select All Unselect All Invert Selection
.
augustus_masked
blastn
blastx
est2genome
maker
protein2genome
repeatmasker
Select All Unselect All Invert Selection
Done
Done


Firefox File Edit View History Bookmarks Tools Window Help
(Charged) Mon 12:42 PM
MAKER Web Annotation Service
http://derringer.genetics.utah.edu/cgi-bin/MWAS/maker.cgi?rm=job_create&func=1&job_id=1323
Google
Most Visited Yandell Lab Reading Group - Ya... BioPerl BioPerl-Run CHPC at the Universi... Buses jbrowse
MAKER Web Annotation Service
MAKER Web Annotation Service
Your Genome Annotated
Home New Job Manage Files Running Jobs Edit Account Contact Us Help Yandell Lab
logged-in as guest_270 | logout
You are logged in as a guest user. You may become a registered user at any time by clicking on "Edit Account".
Copy the URL http://derringer.genetics.utah.edu/cgi-bin/MWAS/maker.cgi?guest_id=270 or login with the username guest_270 to return to your jobs after exiting.
MAKER Job Details
Assigned id:
1655
Annotation Post Processing
Indicate the post processing you would like to perform on the annotations. You can add functional annotations such as putative gene functions, Gene Ontology terms, and protein domains to the gene models. You can also change long MAKER asigned names into shorter gene names.
Add putative gene functions via comparison to the UniProt/Swiss-Prot protein database
Add protein domains and GO terms via InterProScan
Choose a prefix for making shorter gene names (blank to ignore)
GMOD_
Add Job to Queue
Save/Come Back Later
Clear
© 2007-2009 Mark Yandell
Valid CSS/XHTML 1.0
Done


Firefox File Edit View History Bookmarks Tools Window Help
(Charged) Mon 12:48 PM
MAKER Web Annotation Service
http://derringer.genetics.utah.edu/cgi-bin/MWAS/maker.cgi
Google
Most Visited Yandell Lab Reading Group - Ya... BioPerl BioPerl-Run CHPC at the Universi... Buses jbrowse
MAKER Web Annotation Service
MAKER Web Annotation Service
MAKER Web Annotation Service: 201.4 kbp from NT_010783.15:1..201,444
http://derringer.genetics.utah.edu/cgi-bin/gb2/gbrowse/MWAS_270_1334/?name=NT_010783.15
Overview
NT_010783.15
Region
Details
20 kbp
NT_010783.15: 70 kbp
MAKER Gene Annotations
GMOD_00000003-RA
Similar to IFI35: Interferon-induced 35 kDa protein (Homo sapiens)
GMOD_00000006-RA
Similar to BRCA1: Breast cancer type 1 susceptibility protein
GMOD_00000004-RA
Similar to VAT1: Synaptic vesicle membrane protein VAT-1 homolog (Homo sapiens)
GMOD_00000005-RA
RND2: Rho-related GTP-binding protein RhoN (Homo sapiens)
InterPro Protein Domains
Showing 5 of 22 features
IPR009909 Nmi/IFP 35
IPR003579 Ras small GTPase, Rab type
IPR001357 BRCT
IPR020843 Polyketide synthase, enoylreductase
IPR013154 Alcohol dehydrogenase GroES-like
Proteins - Exonerate
gi|24307901|ref|NP_005524.1|
gi|18379349|ref|NP_006364.2|
gi|4885581|ref|NP_005431.1|
gi|4885069|ref|NP_005159.1|
gi|6552315|ref|NP_009233.1|
gi|63252882|ref|NP_009235.2|
gi|6552301|ref|NP_009226.1|
gi|6552317|ref|NP_009234.1|
gi|6552299|ref|NP_009225.1|
gi|6552321|ref|NP_009236.1|
gi|6552307|ref|NP_009229.1|
gi|6552305|ref|NP_009228.1|
Done
Done

## GMOD_00000004-RA Details

**Name:** GMOD_00000004-RA

**Type:** mRNA

**Description:** Similar to VAT1: Synaptic vesicle membrane protein VAT-1 homolog (Homo sapiens)

**Source:** maker

**Position:** NT_010783.15:30454..38291 (- strand)

**Length:** 7838

**Alias:** augustus_masked-NT_010783.15-abinit-gene-0.8-mRNA-1

**Dbxref:** Gene3D:G3DSA:3.40.50.720

InterPro:IPR002085

InterPro:IPR002364

InterPro:IPR011032

InterPro:IPR013149

InterPro:IPR013154

InterPro:IPR016040

InterPro:IPR020843

PANTHER:PTHR11695

PANTHER:PTHR11695:SF29

Pfam:PF00107

Pfam:PF08240

Prosite:PS01162

SMART:SM00829

superfamily:SSF50129

superfamily:SSF51735

**Note:** Similar to VAT1: Synaptic vesicle membrane protein VAT-1 homolog (Homo sapiens)

**Ontology_term:** GO:0003824

GO:0005488

GO:0008152

GO:0008270

GO:0016491

GO:0055114

# MAKER Wiki

http://weatherby.genetics.utah.edu/MAKER/wiki

## Post Processing of Annotations

Once you've determined where the genes are the next question is what do they do. Also it would be nice to change ugly MAKER assigned gene names to follow more standardized formats.

MAKER has a number of accessory scripts that allow you to do just that. So let's take a look at our last example.

```
cd ~/maker_tutorial/example_04_postannotation
ls -1

 hsap_contig.gff
 hsap_contig.maker.proteins.fasta
 hsap_contig.maker.transcripts.fasta
 output.blastp
 output.iprscan
 uniprot_sprot.db
```

Here we have our MAKER output GFF3 and FASTA files for proteins and transcripts (Click to see GFF3 in JBrowse). We also have output reports from the program InterProScan (a domain finder) ran on the MAKER proteins and a BLAST report of homology of the MAKER proteins to UniProt/Swiss-Prot. How to run these programs is not part of this tutorial, but how to integrate their output is.

This is an example command line for running BLASTP against UniProt/Swiss-Prot (you don't need to run it, it's just for reference):

```
blastp -query hsap_contig.maker.proteins.fasta -db uniprot_sprot.fasta -evalue 1e-6 -max_hsps 1 -max_target_seqs 1 -outfmt 6
```

This is an example command line for running InterProScan (you don't need to run it, it's just for reference):

```
interproscan.sh -appl pfam -dp -f TSV -goterms -iprlookup -pa -t p -i hsap_contig.maker.proteins.fasta -o output.iprscan
```

But first lets fix those ugly MAKER names.

MAKER comes with the script maker_map_ids to make it easy to rename genes and follow formats such as those suggested by NCBI (organism prefix and gene numbers).

```
Synopsis:

maker_map_ids --prefix PYU1_ --justify 8 genome.all.gff > genome.all.id.map

Description:
```

# Acknowledgements


- Mark Yandell
- Barry Moore
- Michael Campbell
- Daniel Ence