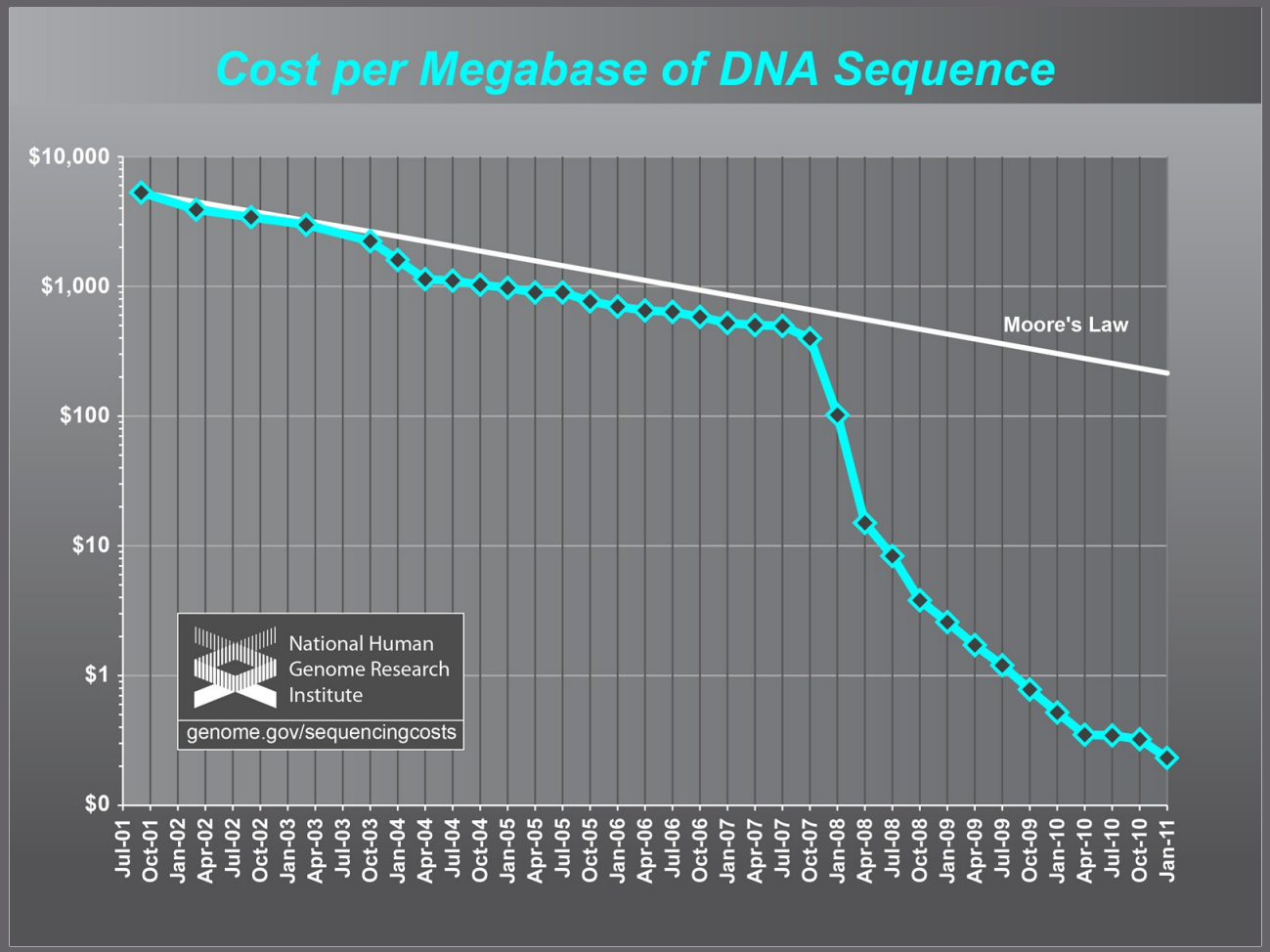
MAKER: an easy-to-use genome annotation pipeline

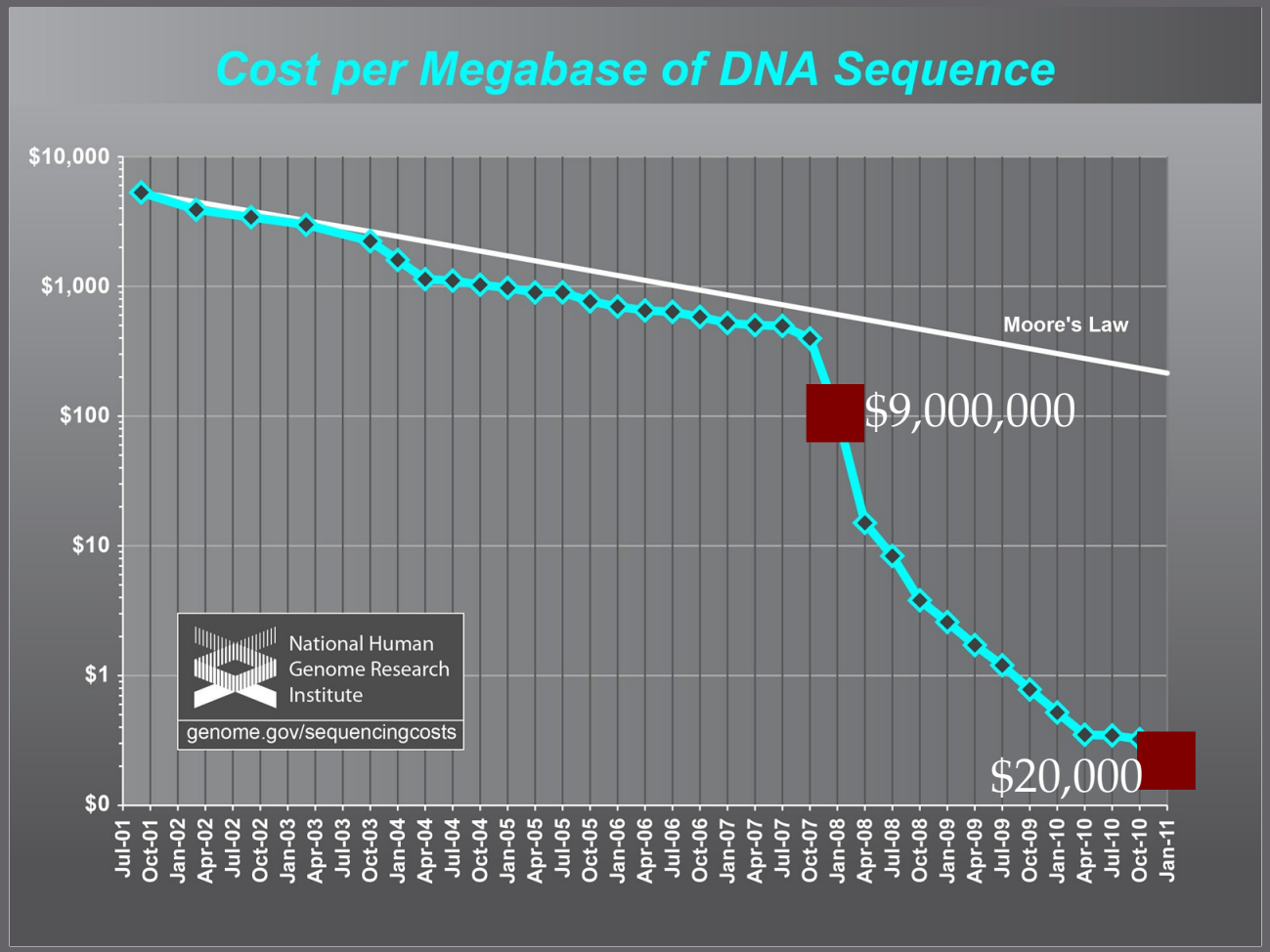
Barry Moore and Daniel Ence
Yandell Lab
University of Utah

Carson Holt
Ontario Institute for Cancer Research

Advances in Second-Generation Technology is Starting to Make Whole Genome and Transcriptome Sequencing "Routine" Even for Small Labs

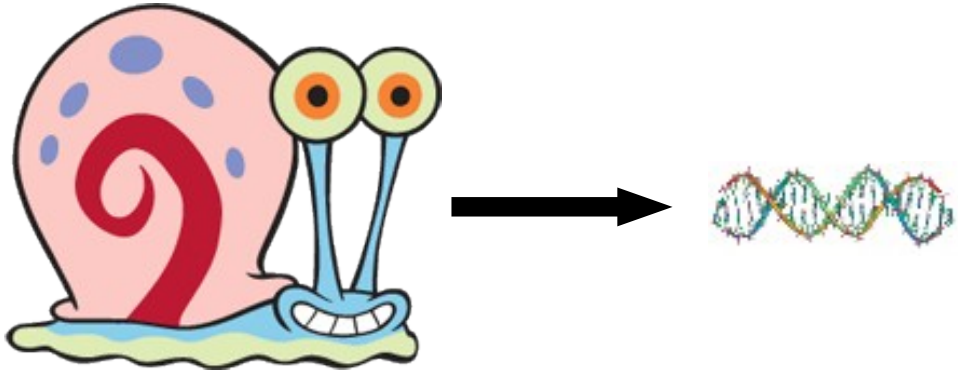


Advances in Second-Generation Technology is Starting to Make Whole Genome and Transcriptome Sequencing "Routine" Even for Small Labs

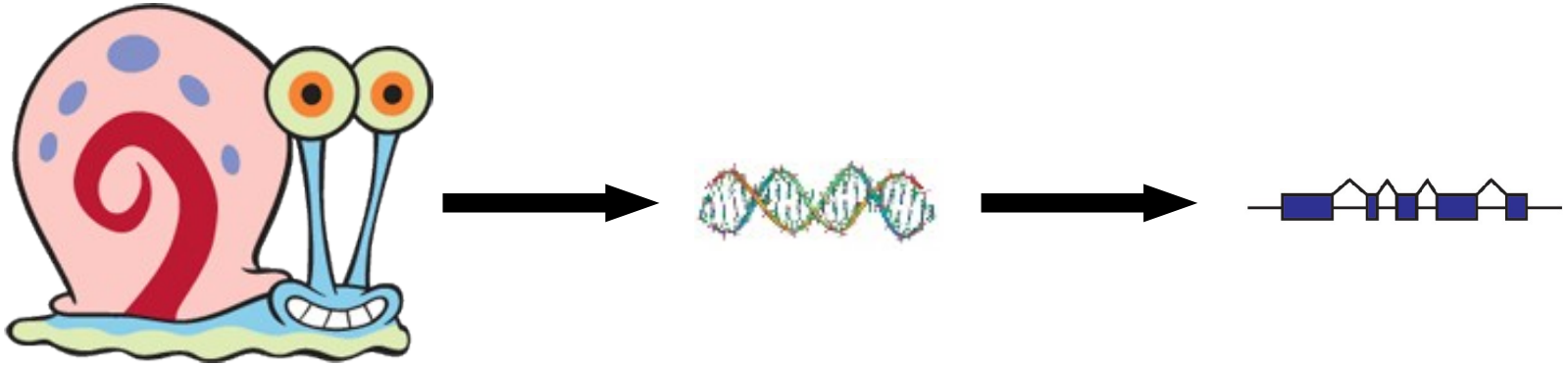


Genome Project Overview

Genome Project Overview

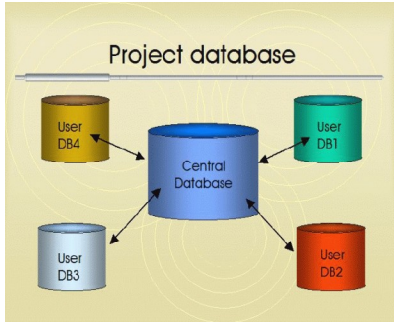


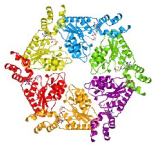
Genome Project Overview

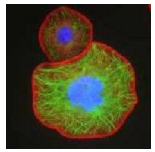


Genome Project Overview

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

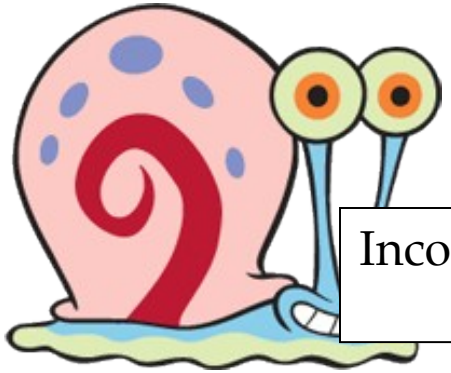






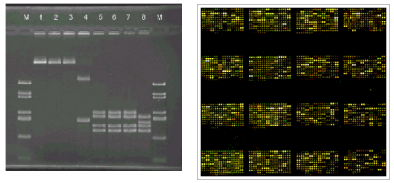
SUCCESS

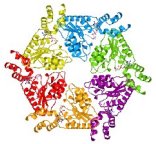
Genome Project Overview

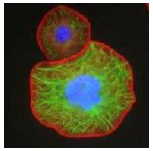


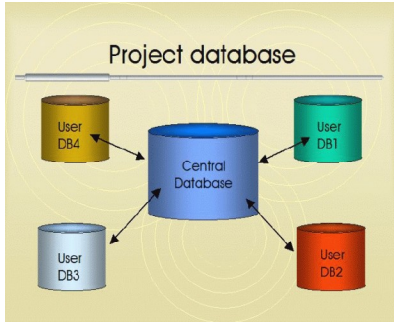
Incorrect annotations poison every experiment that uses them!!

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

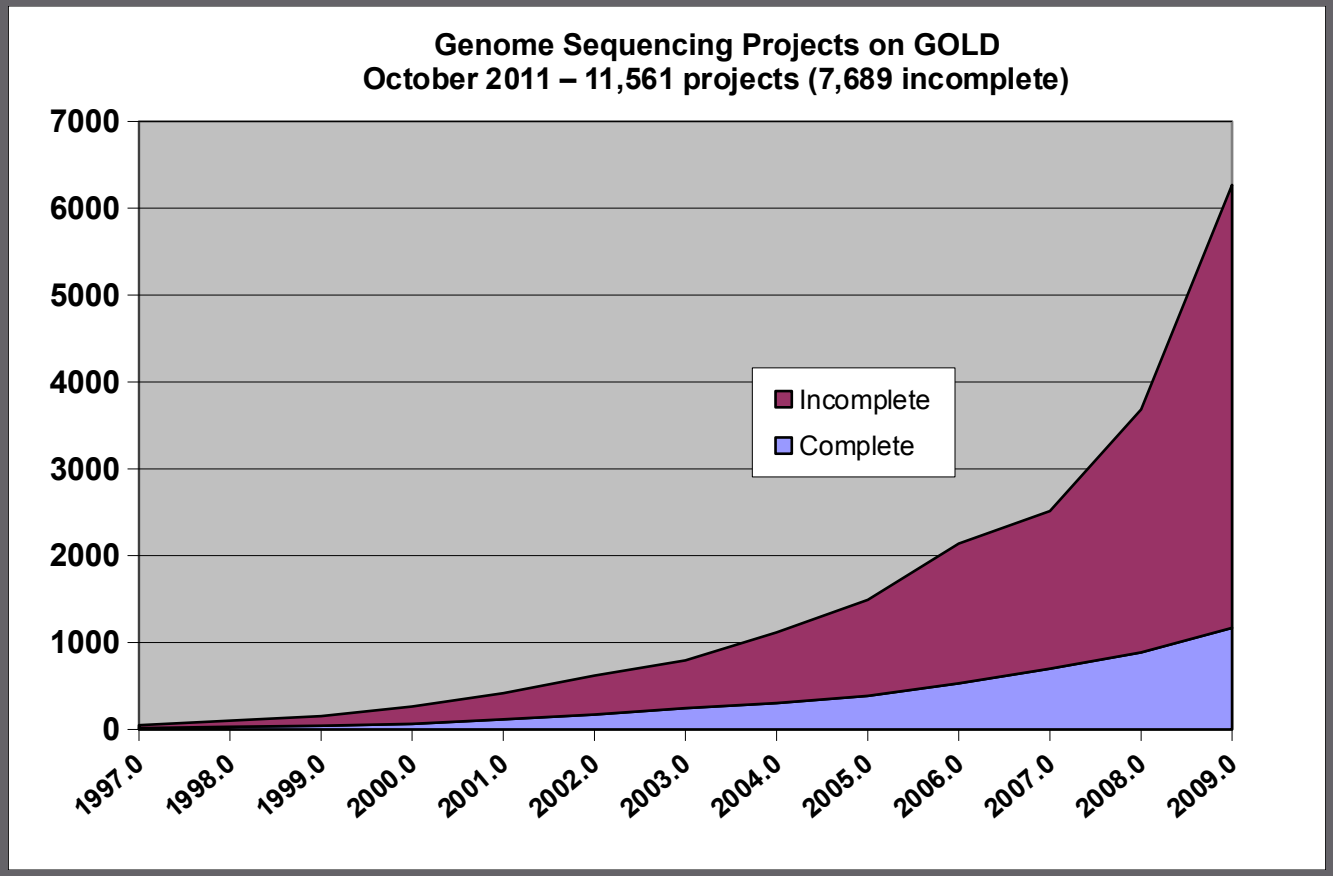








Advances in annotation technology have not kept pace with genome sequencing, and annotation is now the major bottleneck affecting modern genomics research.



- GOLD: Genomes OnLine Database. 2009.

MAKER

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

Easy-to-use by design

Easy-to-use by design

User Requirements:	Can be run by a single individual with little bioinformatics experience

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)

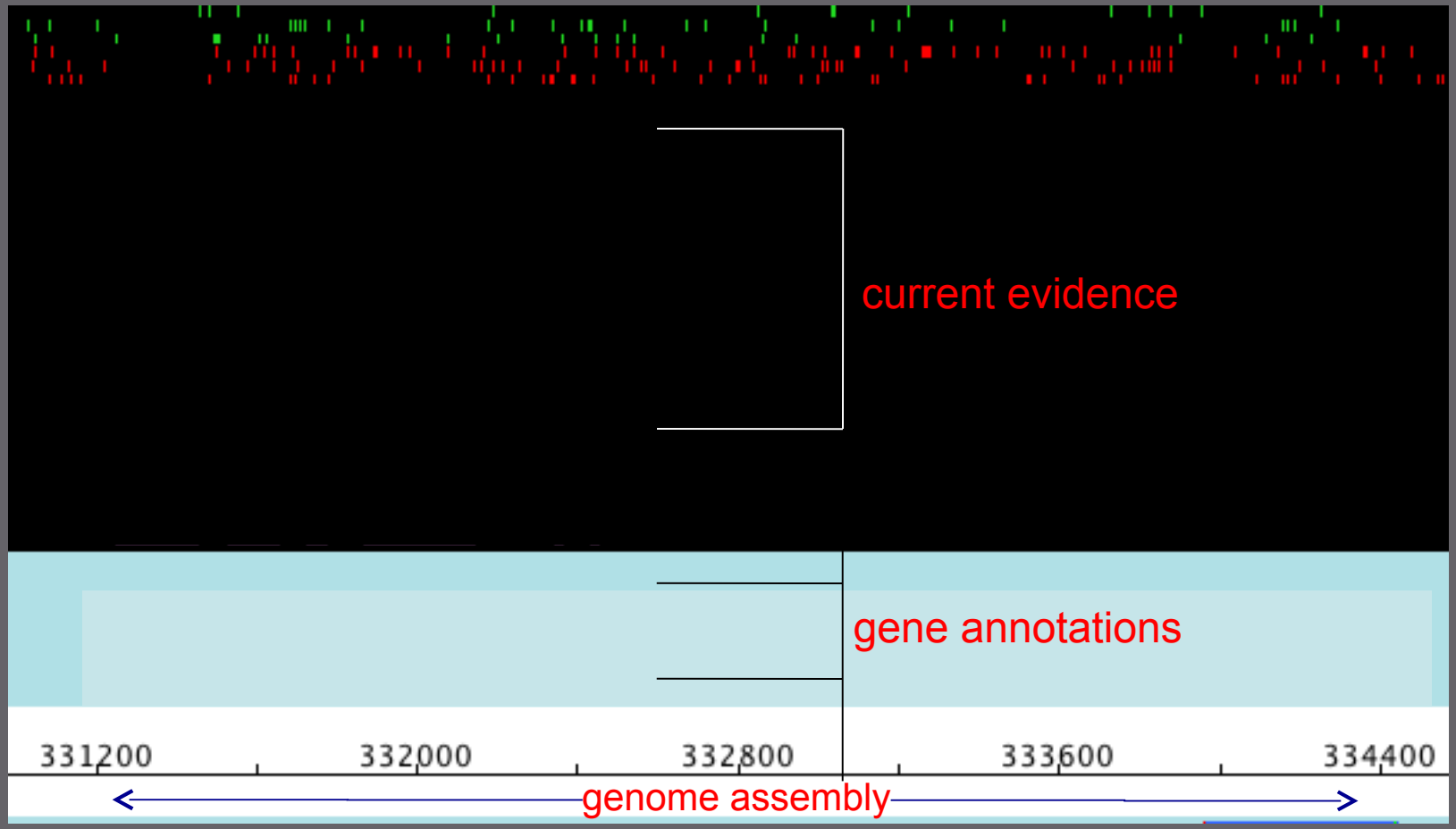
Easy-to-use by design

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 and GBrowse

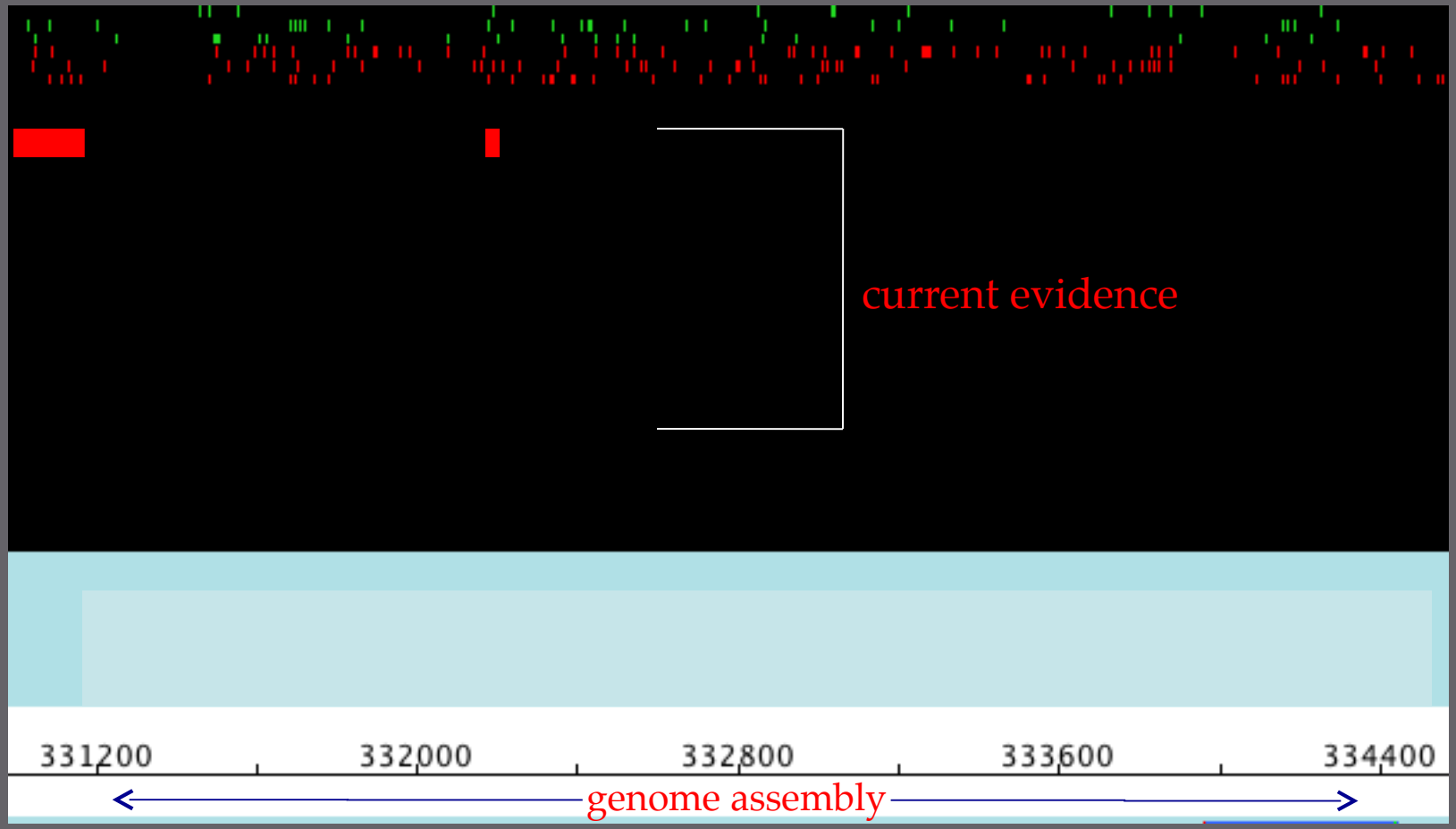
Easy-to-use by design

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 and GBrowse
Availability:	Free open source application (for the academic community)

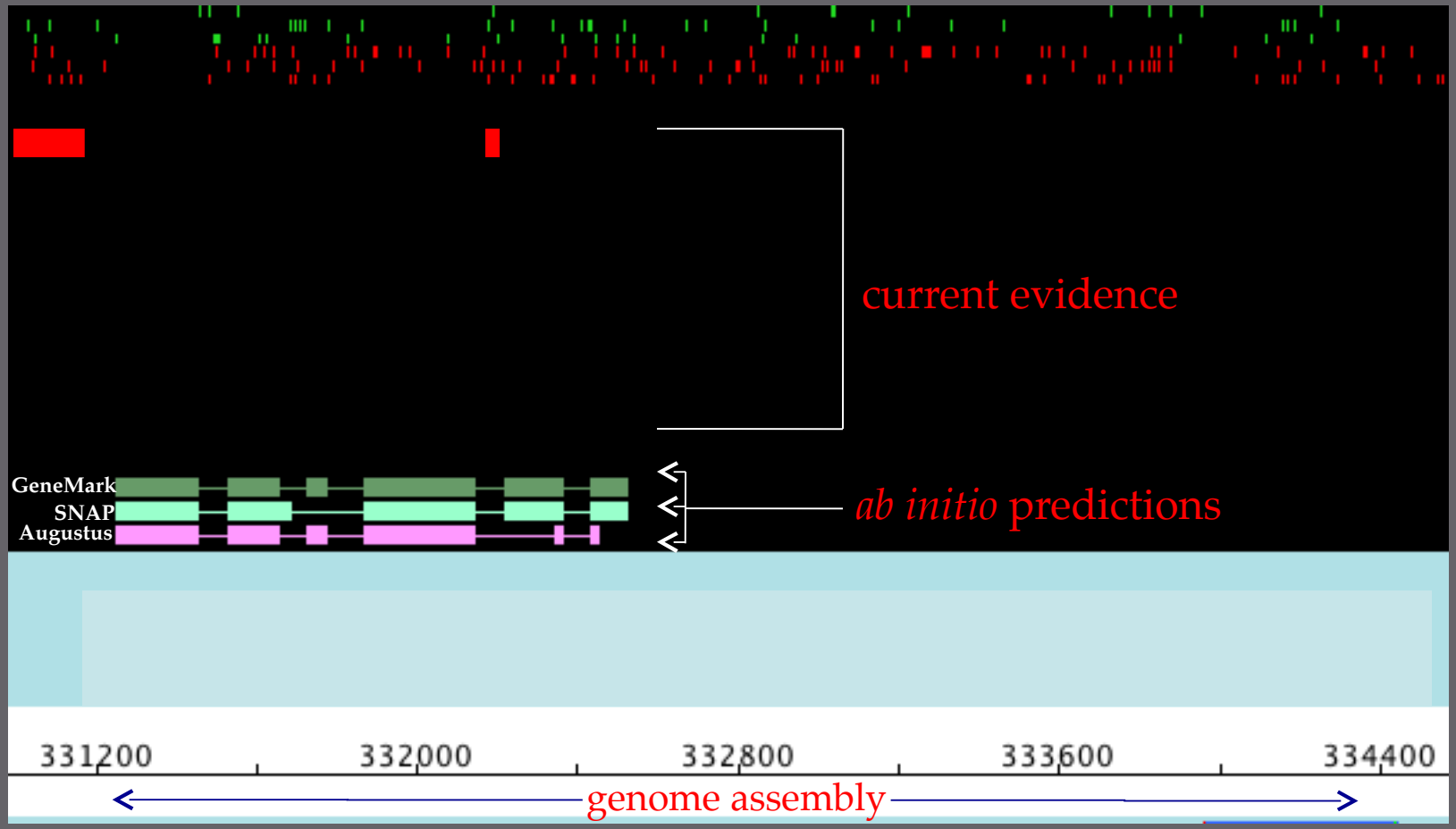
Annotating the Genome – Apollo View



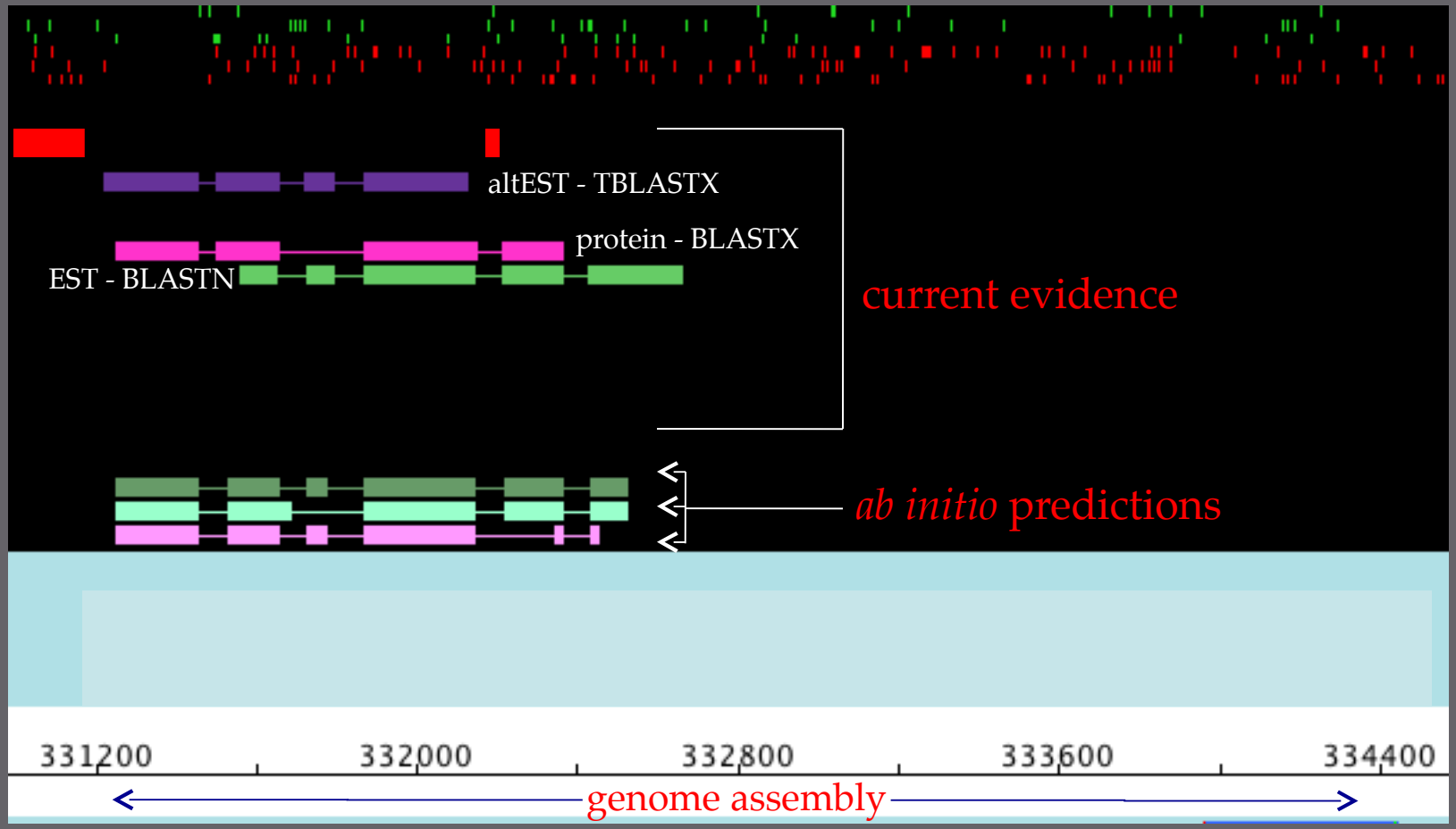
Identify and mask repetitive elements



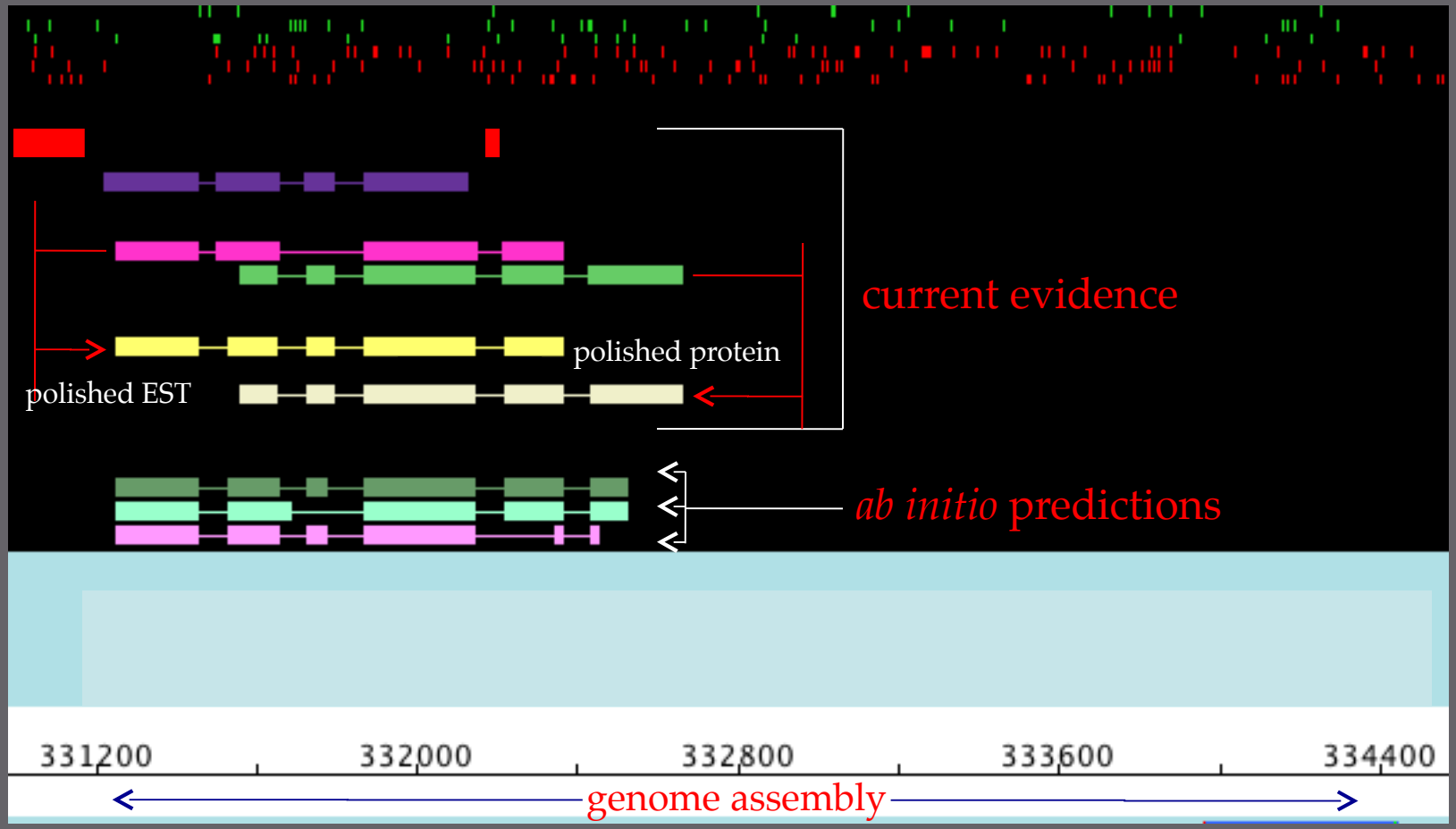
Generate *ab initio* gene predictions



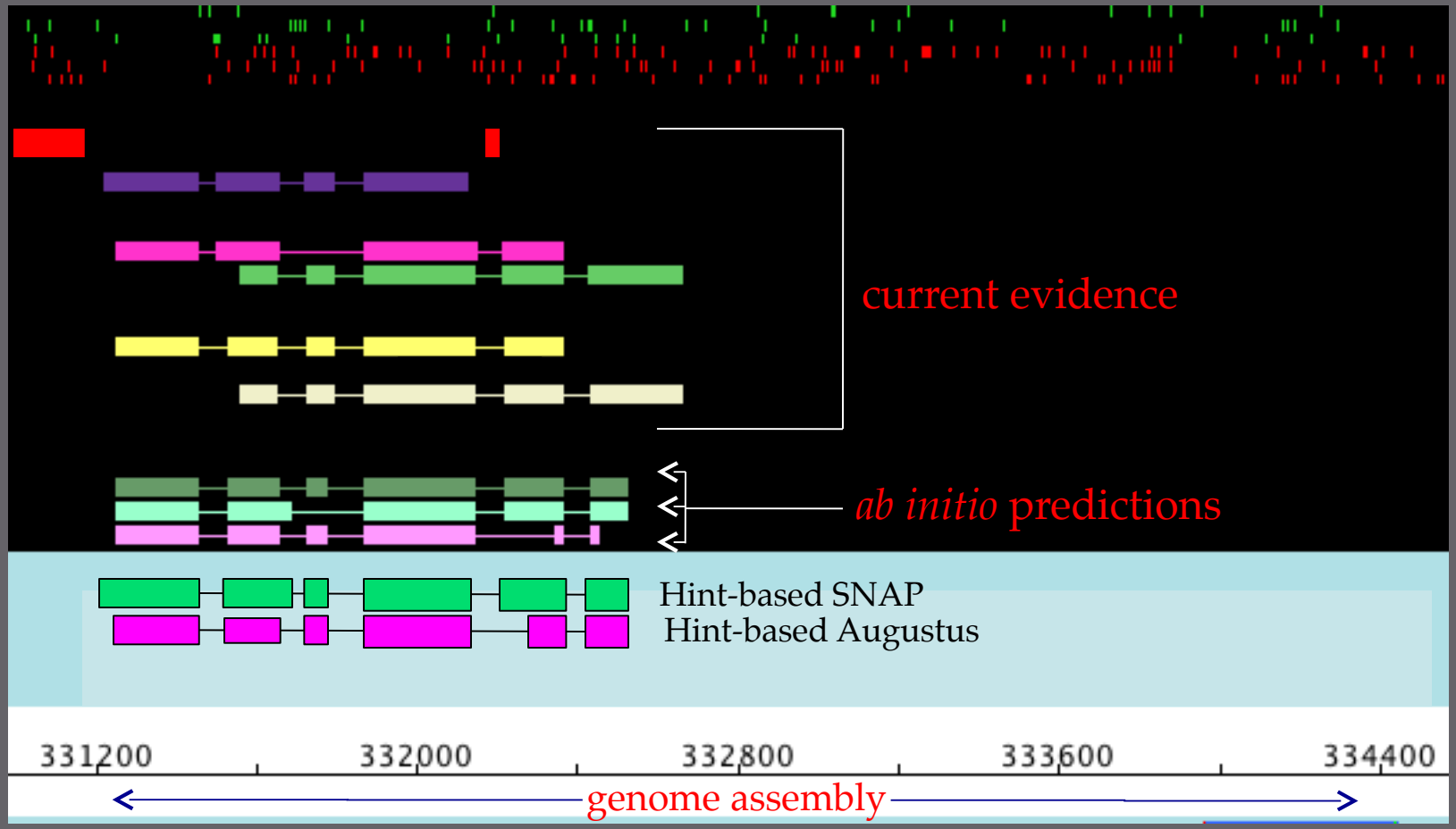
Align EST and protein evidence



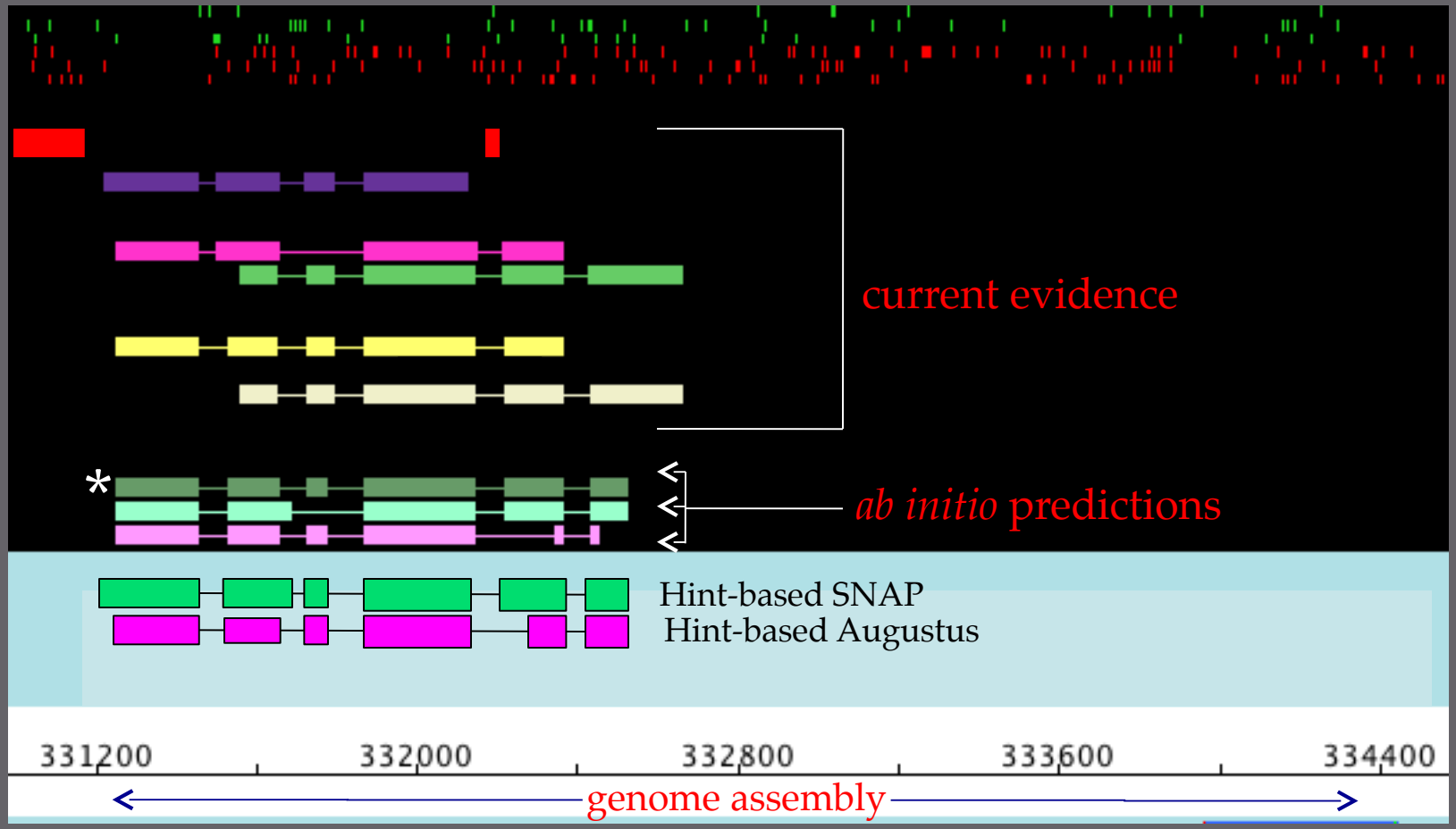
Polish BLAST alignments with Exonerate



Pass gene-finders evidence-based ‘hints’

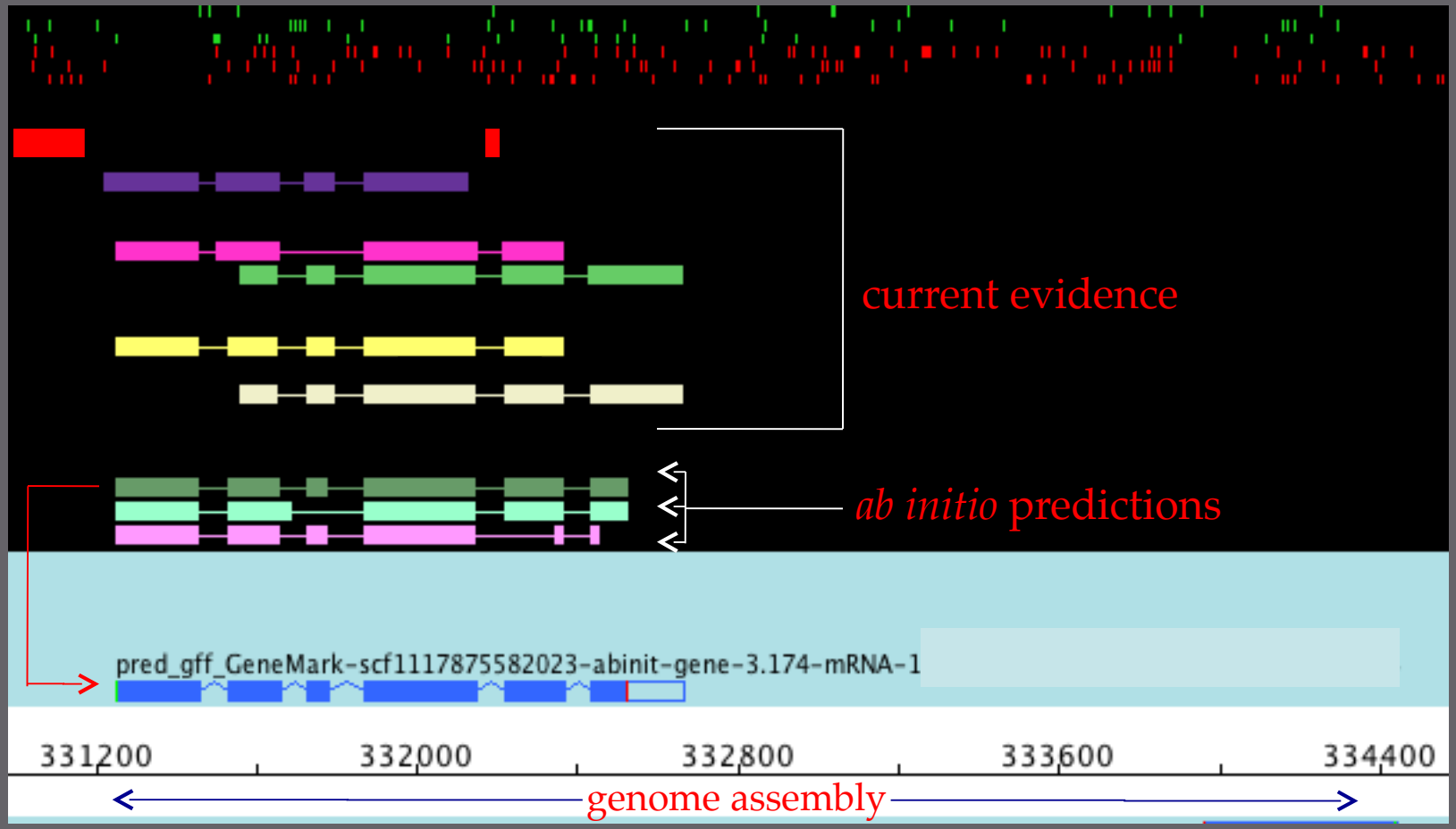


Identify gene model most consistent with evidence

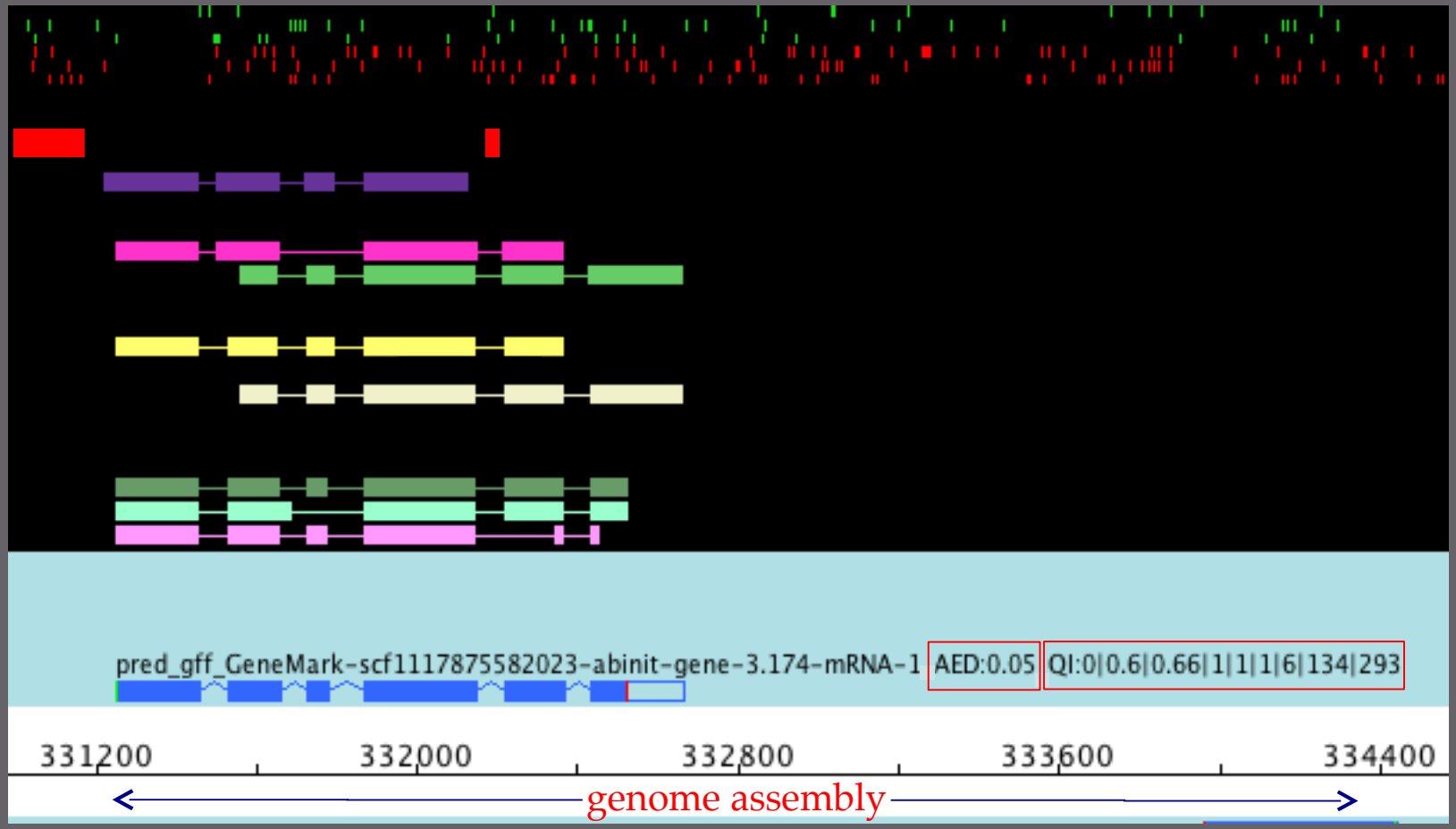


*Quantitative Measures for the Management and Comparison of Annotated Genomes
Karen Eilbeck , Barry Moore , Carson Holt and Mark Yandell BMC Bioinformatics 2009
10:67doi:10.1186/1471-2105-10-67

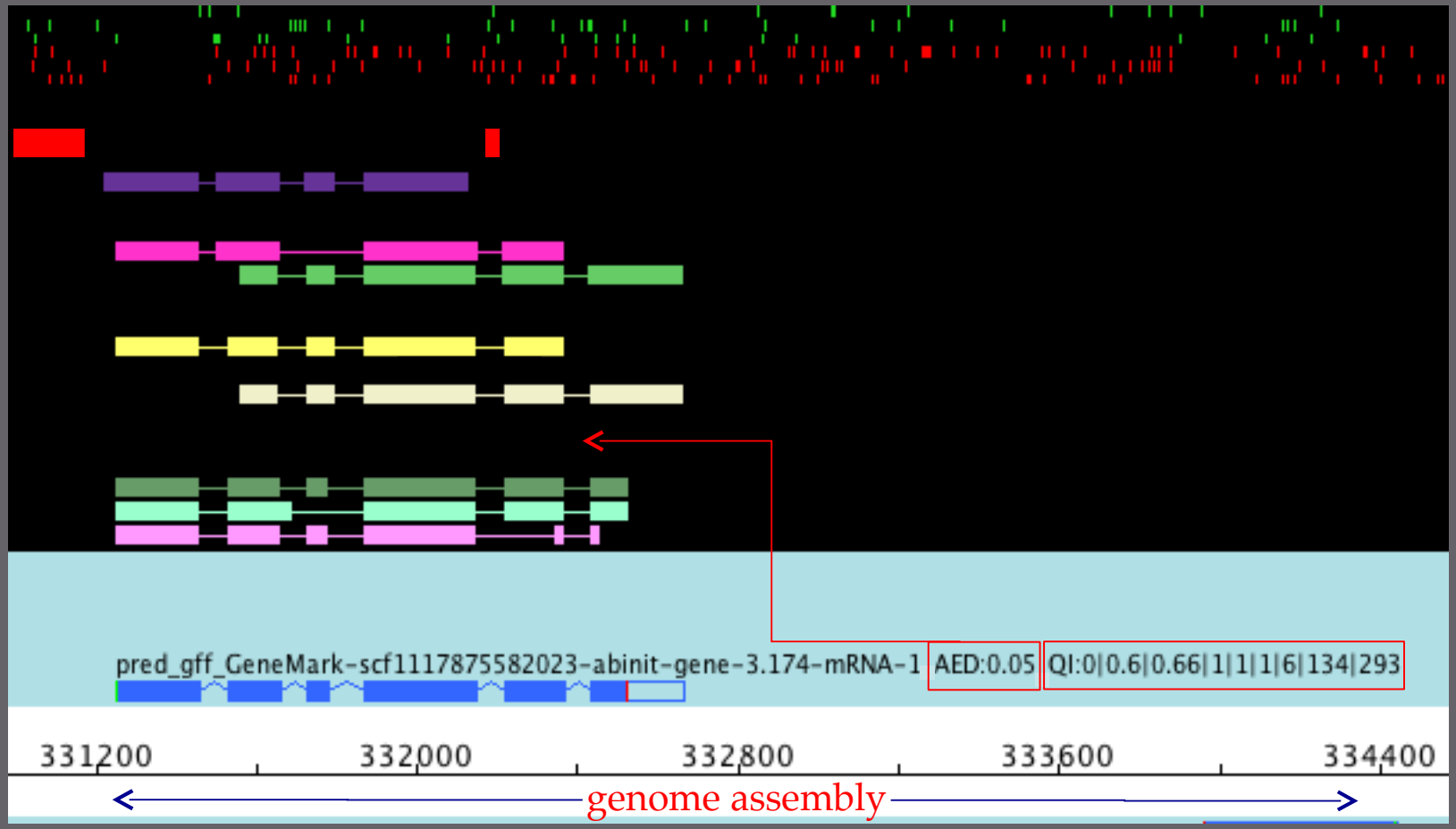
Revise it further if necessary; create new annotation



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.*

Beyond *de novo* annotation

- Quality control and data prioritization

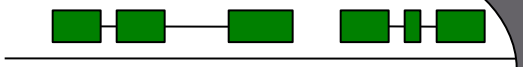
Beyond *de novo* annotation

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

Beyond *de novo* annotation

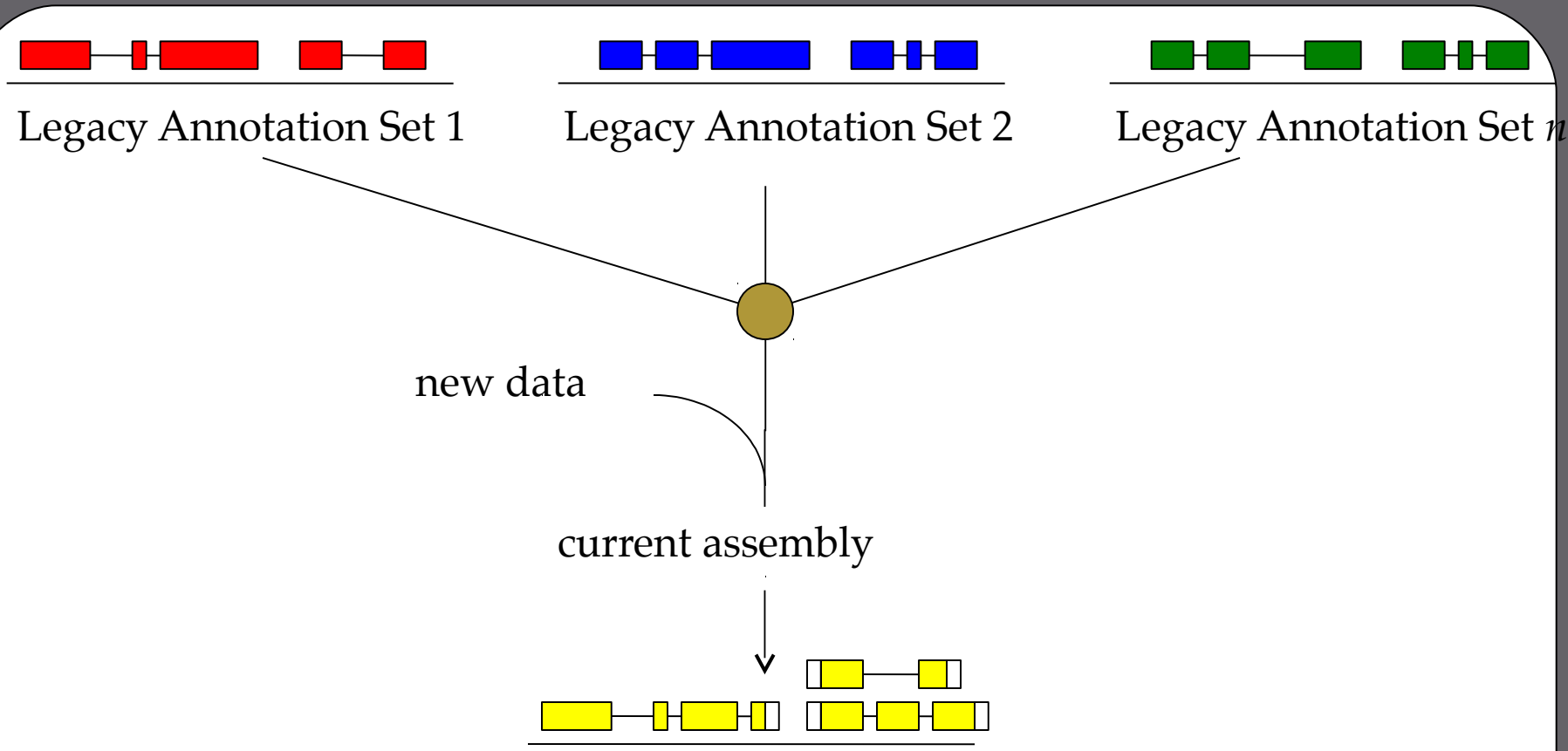
Legacy Annotation Set 1

Legacy Annotation Set 2



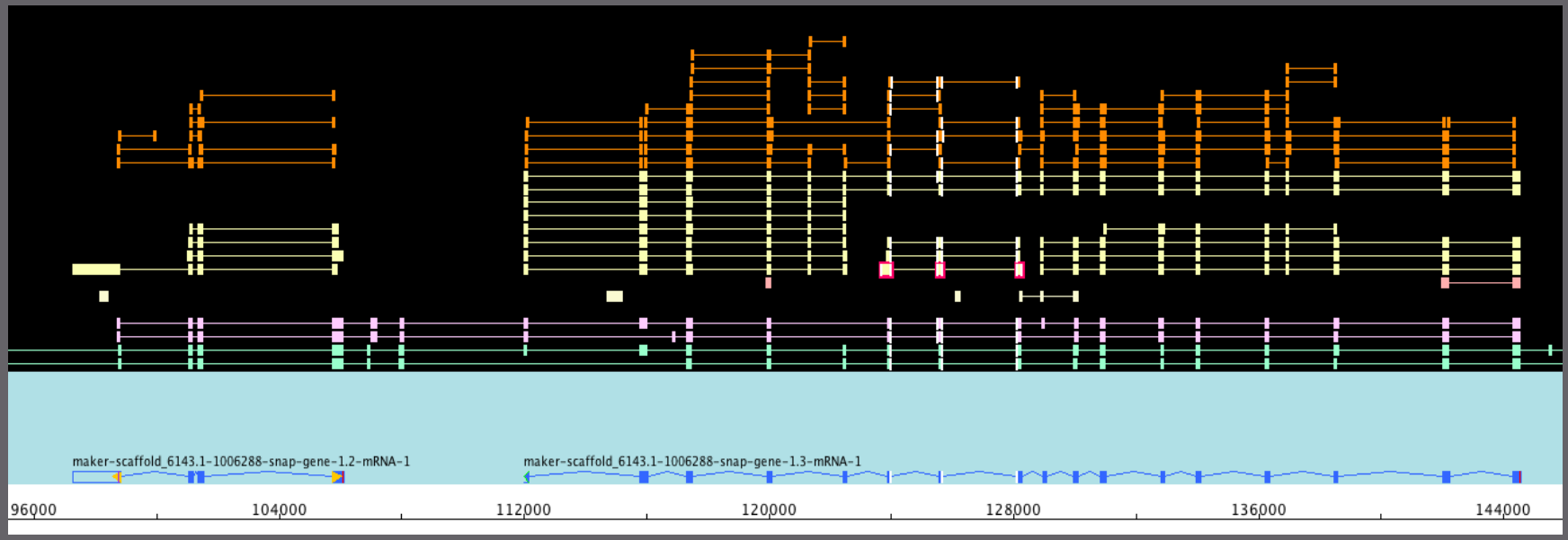
Legacy Annotation Set *n*

Beyond *de novo* annotation



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

mRNA-seq in MAKER



MAKER Web Annotation Service:

an online portal for genome annotation and analysis

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 12:35 PM
MAKER Web Annotation Service
http://derringer.genetics.utah.edu/cgi-bin/MWAS/maker.cgi?rm=job_create
Most Visited Yandell Lab Reading Group - Ya... BioPerl BioPerl-Run CHPC at the Universi... Buses jbrowse
MAKER Web Annotation Service

ab-initio gene predictions. **Optional**	No file selected... View File Contents	Upload File
Upload a GFF3 file of gene models. **Optional**	Select a file below No file selected... Select Multiple Files View File Contents	Upload File
		save

Annotation Properties

These are other options that can have an effect on the annotation process. For the most part, you can leave these set to their default values.

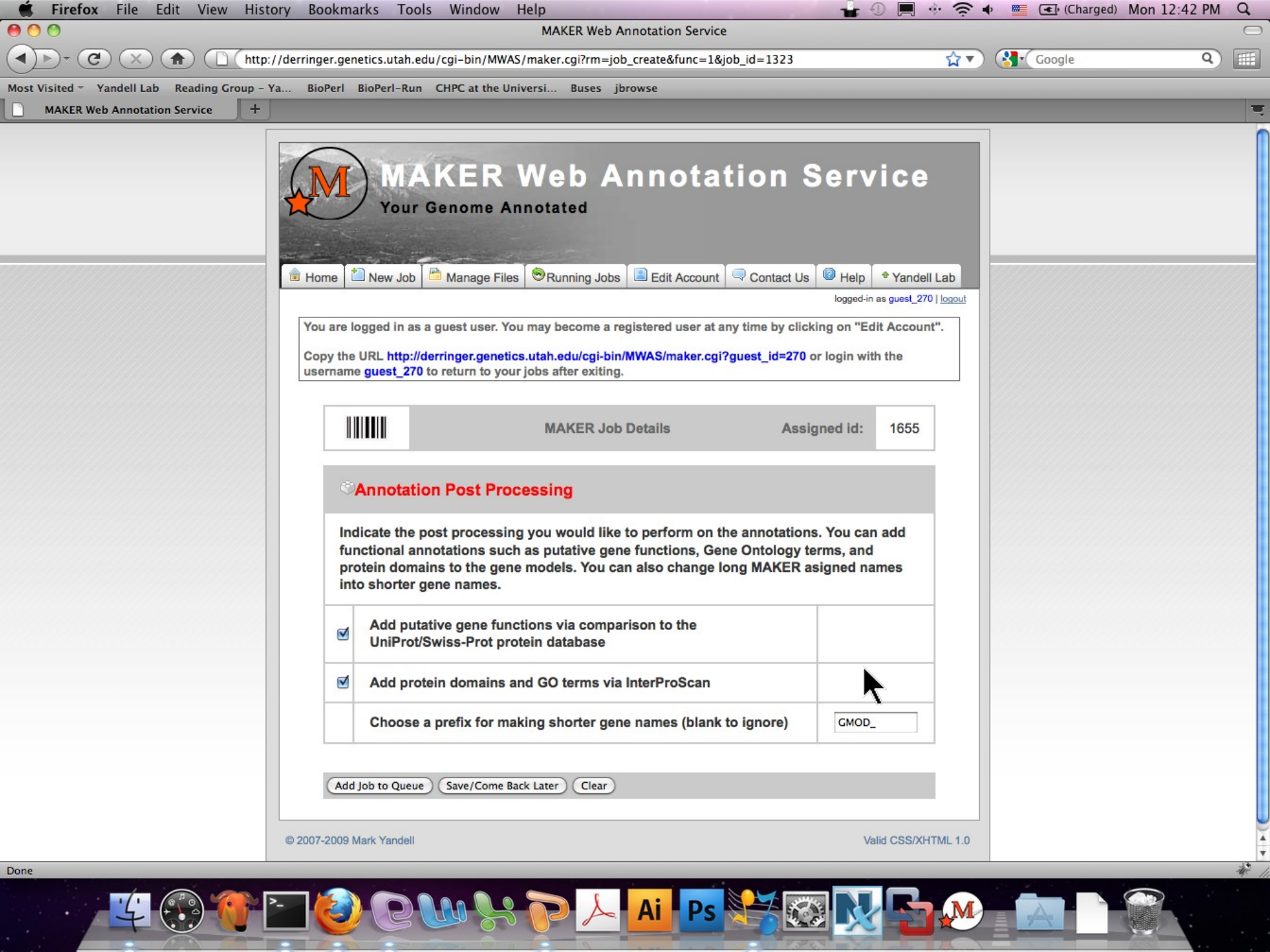
Non standard amino acids generally cause problems for alignment algorithms. Here you can select a standard amino acids to replace non standard amino acids.	C
All contigs from the input genome file below this size will be skipped.	1
All final gene annotations must produce a protein of at least this many amino acids in length.	0
Expected maximum intron size for evidence alignments (in base pairs).	10000
Consider single exon EST evidence when generating annotations. Single exon ESTs usually result from genomic contamination of the cDNA, so ignoring them is recommended.	No
Minumum length required for single exon ESTs if single exon EST alignments are used (see above option).	250

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 12:42 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 12:45 PM
GBrowse Details: id:1187
http://derringer.genetics.utah.edu/cgi-bin/gb2/gbrowse_details/MWAS_270_1334?ref=NT_010783.15;start=30454;end=382
Google
Most Visited Yandell Lab Reading Group - Ya... BioPerl BioPerl-Run CHPC at the Universi... Buses jbrowse
MAKER Web Annotation Service
GBrowse Details: id:1187

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

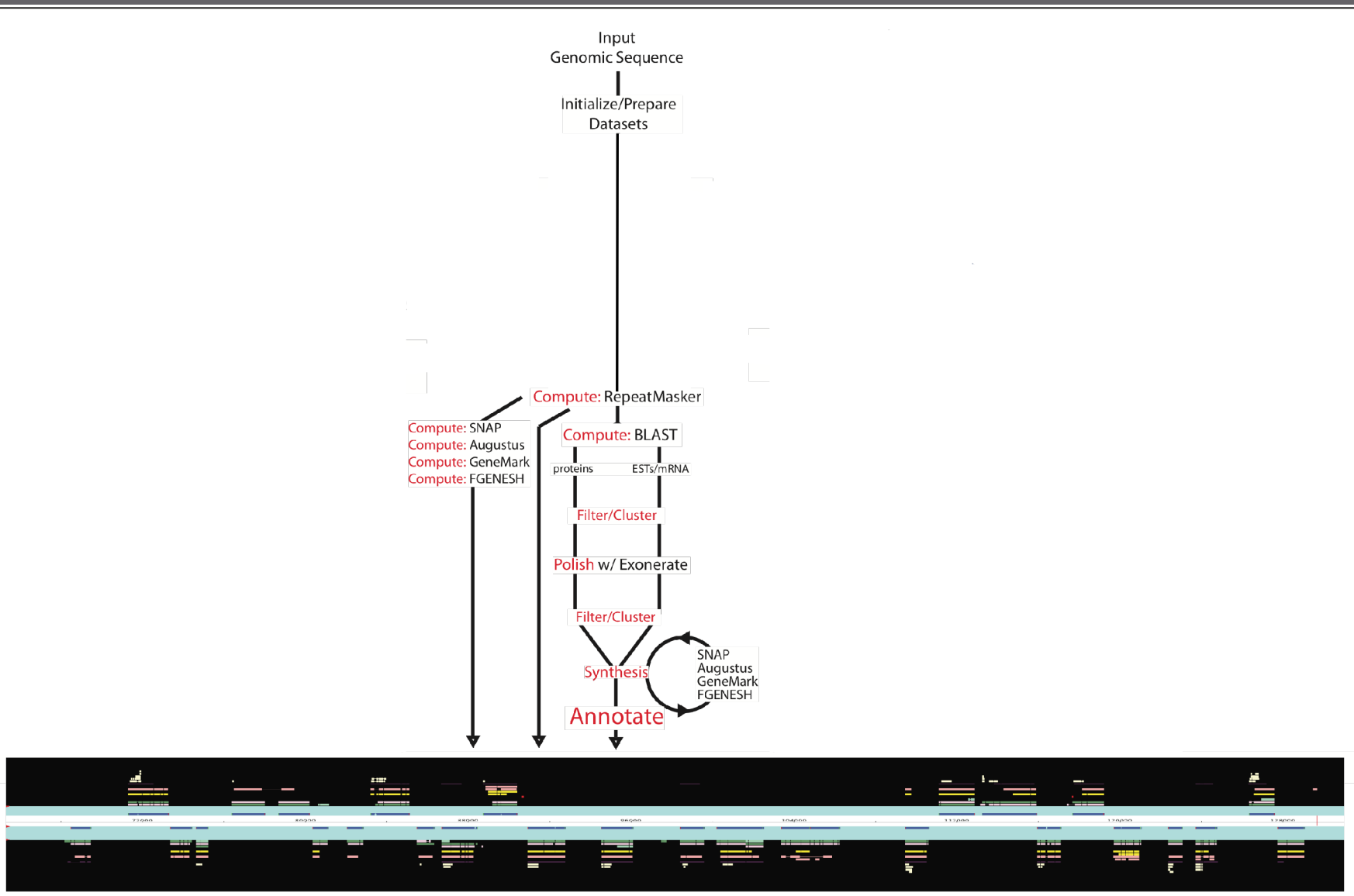
Done

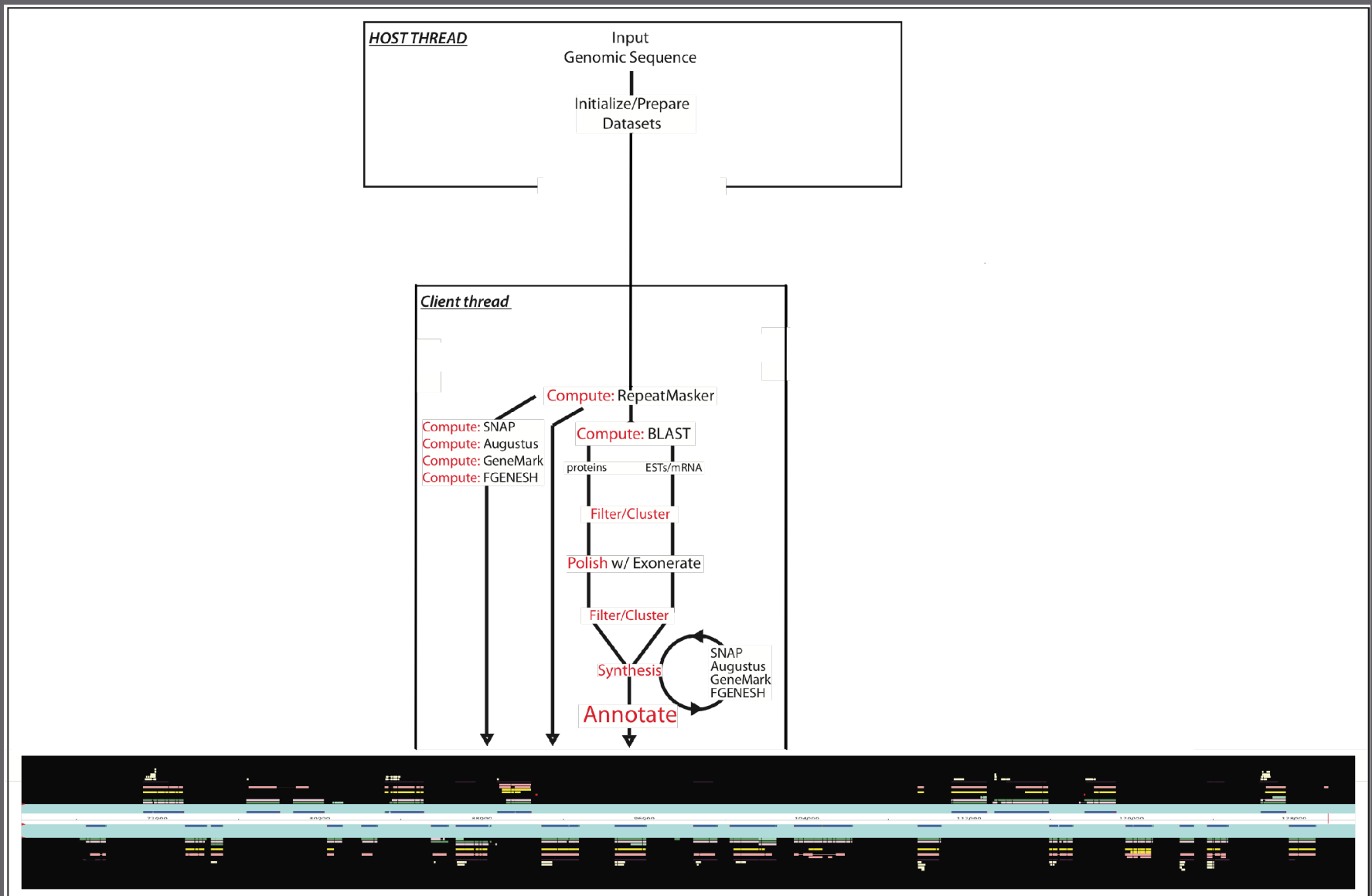
More MAKER Features

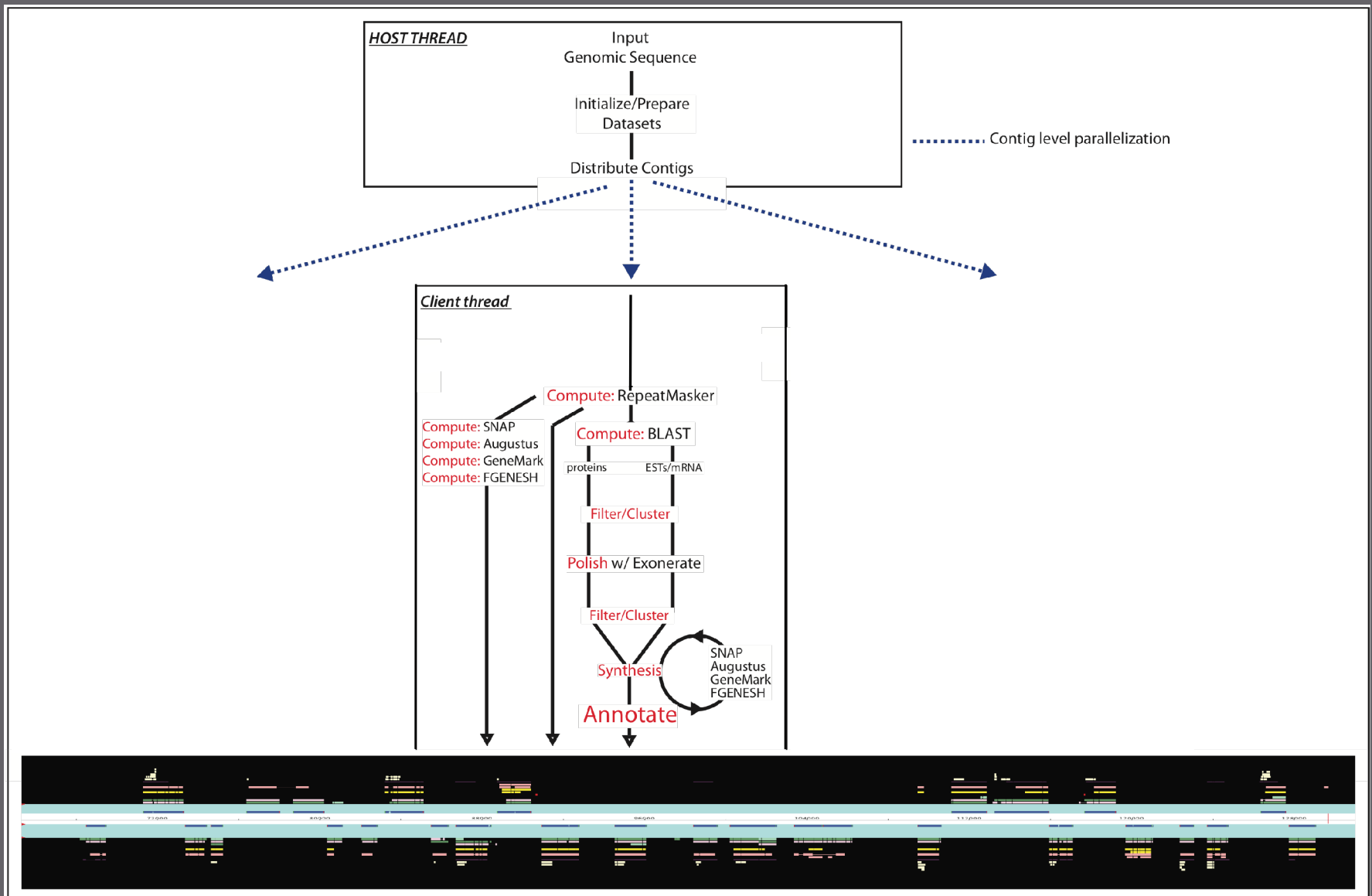
- Auto-training of SNAP
- Improved filtering of gene models
- Comma separated file lists and labels ("blastx:uniprot" or "blastx:hymenoptera")
- Local install of MWAS
- Interactive install and configuration
- Easy GMOD integration (pre-loaded with configuration files and scripts - maker2chado and maker2jbrowse)

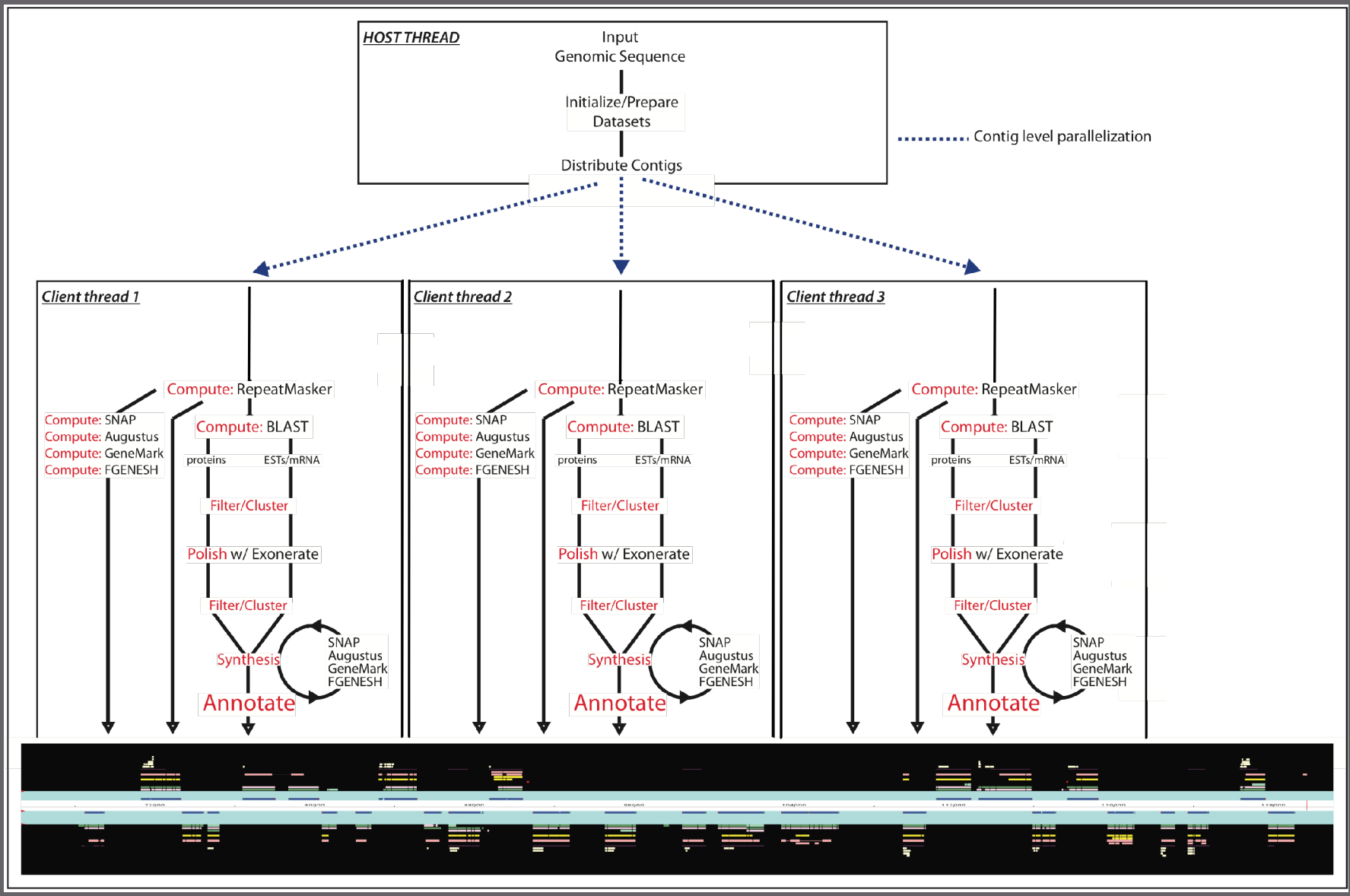
Distributed Parallelization

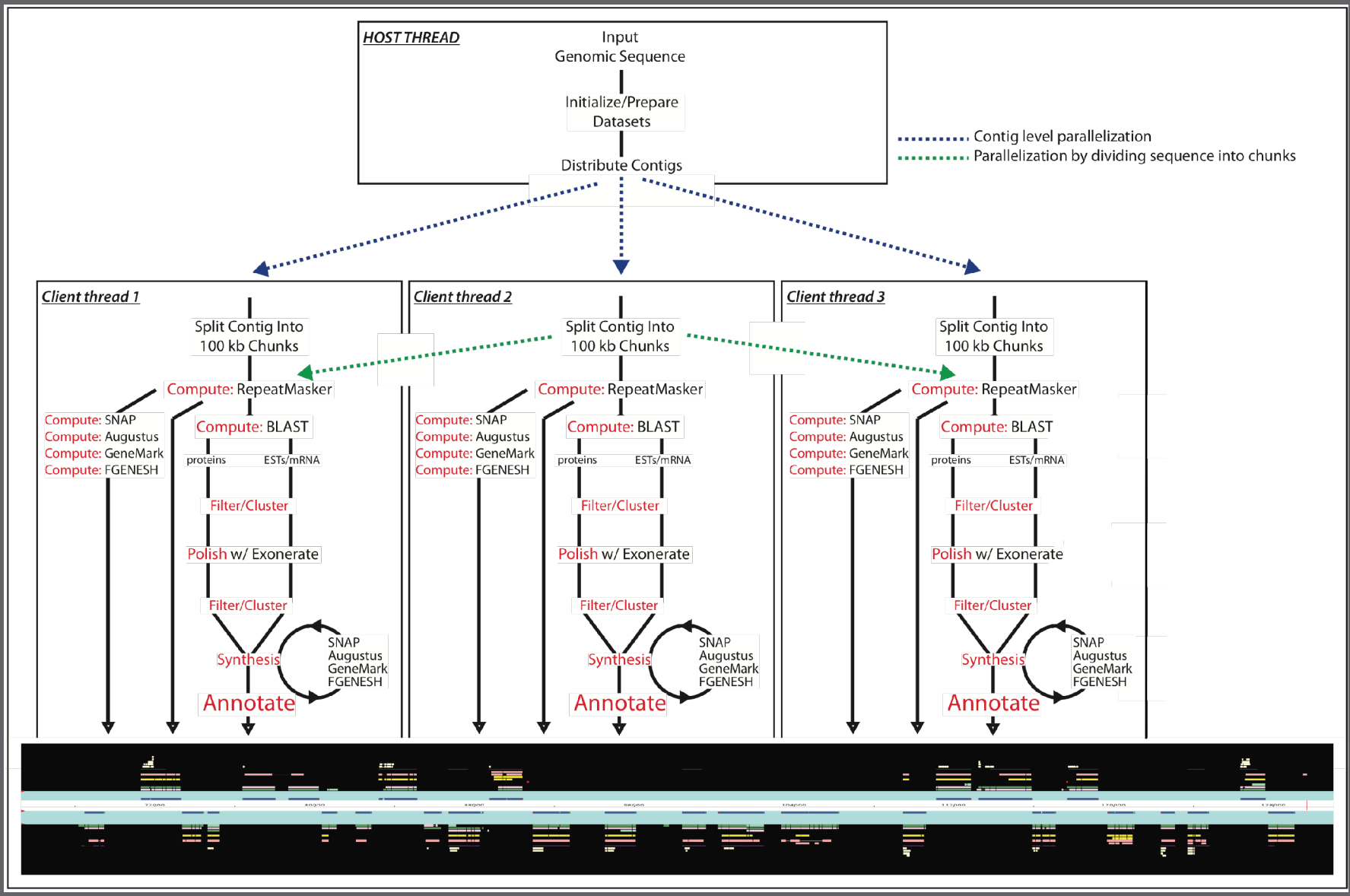
- Supports Message Passing Interface (MPI), a communication protocol for computer clusters which essentially allows multiple computers to act like a single powerful machine.

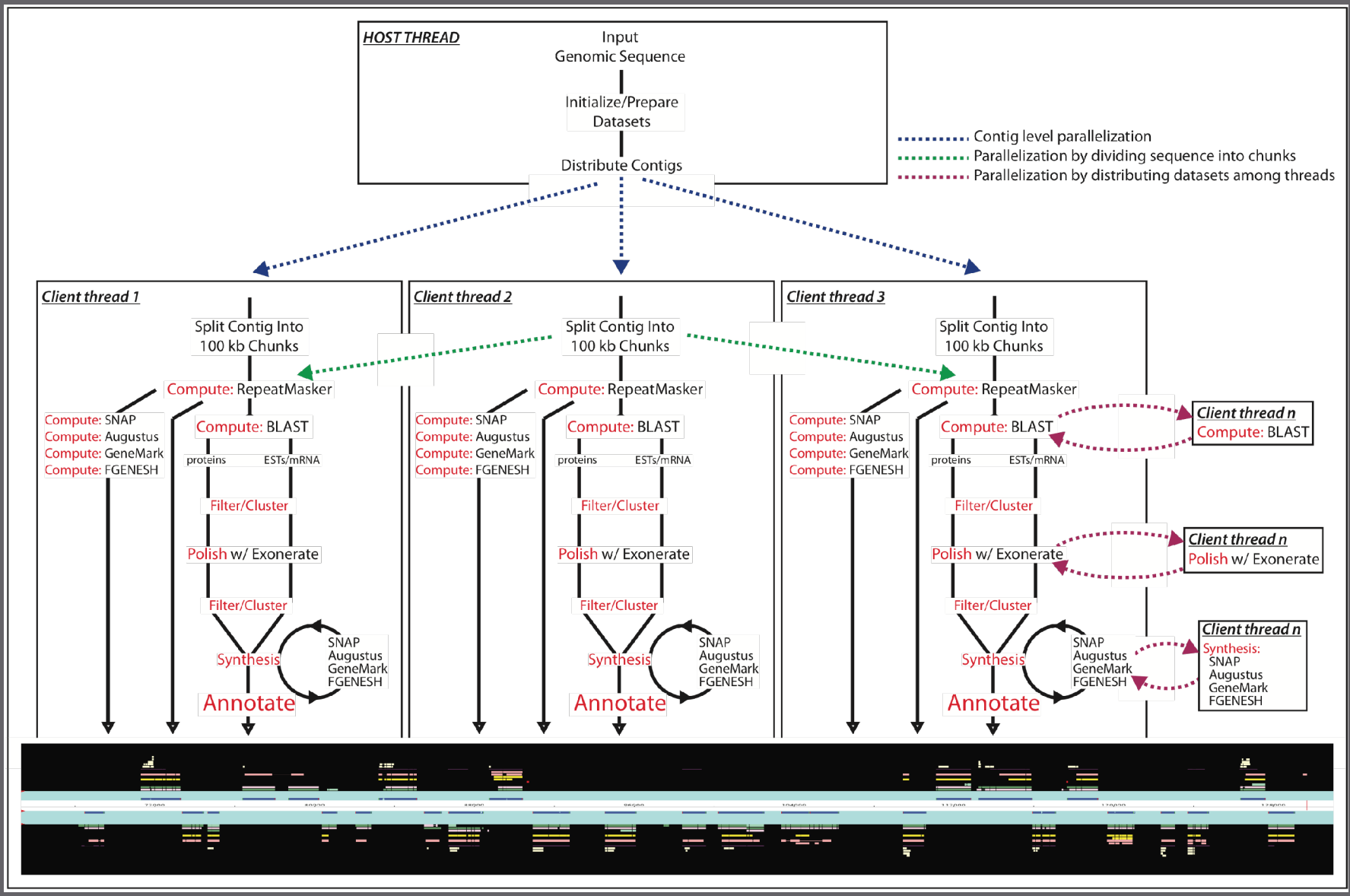

Input
Genomic Sequence
Initialize/Prepare
Datasets
Compute: RepeatMasker
Compute: SNAP
Compute: Augustus
Compute: GeneMark
Compute: FGENESH
Compute: BLAST
proteins
ESTs/mRNA
Filter/Cluster
Polish w/ Exonerate
Filter/Cluster
Synthesis
SNAP
Augustus
GeneMark
FGENESH
Annotate


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

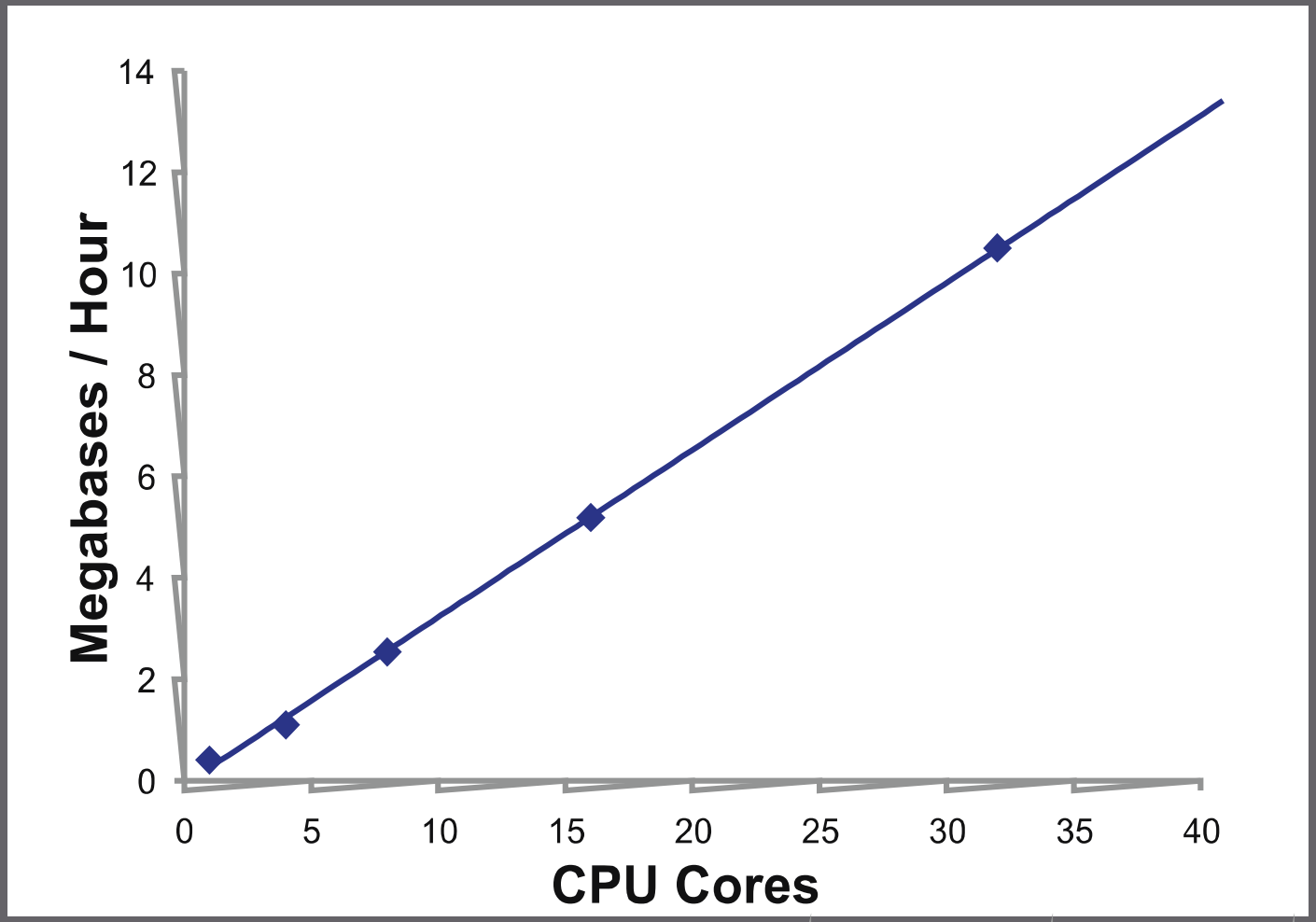

HOST THREAD
Input
Genomic Sequence
Initialize/Prepare
Datasets
Distribute Contigs
Contig level parallelization
Client thread
Compute: RepeatMasker
Compute: SNAP
Compute: Augustus
Compute: GeneMark
Compute: FGENESH
Compute: BLAST
proteins
ESTs/mRNA
Filter/Cluster
Polish w/ Exonerate
Filter/Cluster
Synthesis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

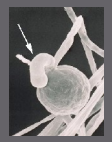

HOST THREAD
Input Genomic Sequence
Initialize/Prepare Datasets
Distribute Contigs
Contig level parallelization
Parallelization by dividing sequence into chunks
Parallelization by distributing datasets among threads
Client thread 1
Client thread 2
Client thread 3
Split Contig Into 100 kb Chunks
Compute: 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



Proof-of-principle:

MAKER: Proof of Principle

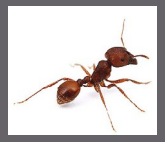


Pythium ultimum var ultimum (potato rot)

Pinus taeda (loblolly pine) *BACs Only

Fusarium circinatum (pitch canker fungus)

Atta cephalotes (leaf-cutter ant)



Pogonomyrmex barbatus (red harvester ant)

Linepithema humile (Argentine ant)

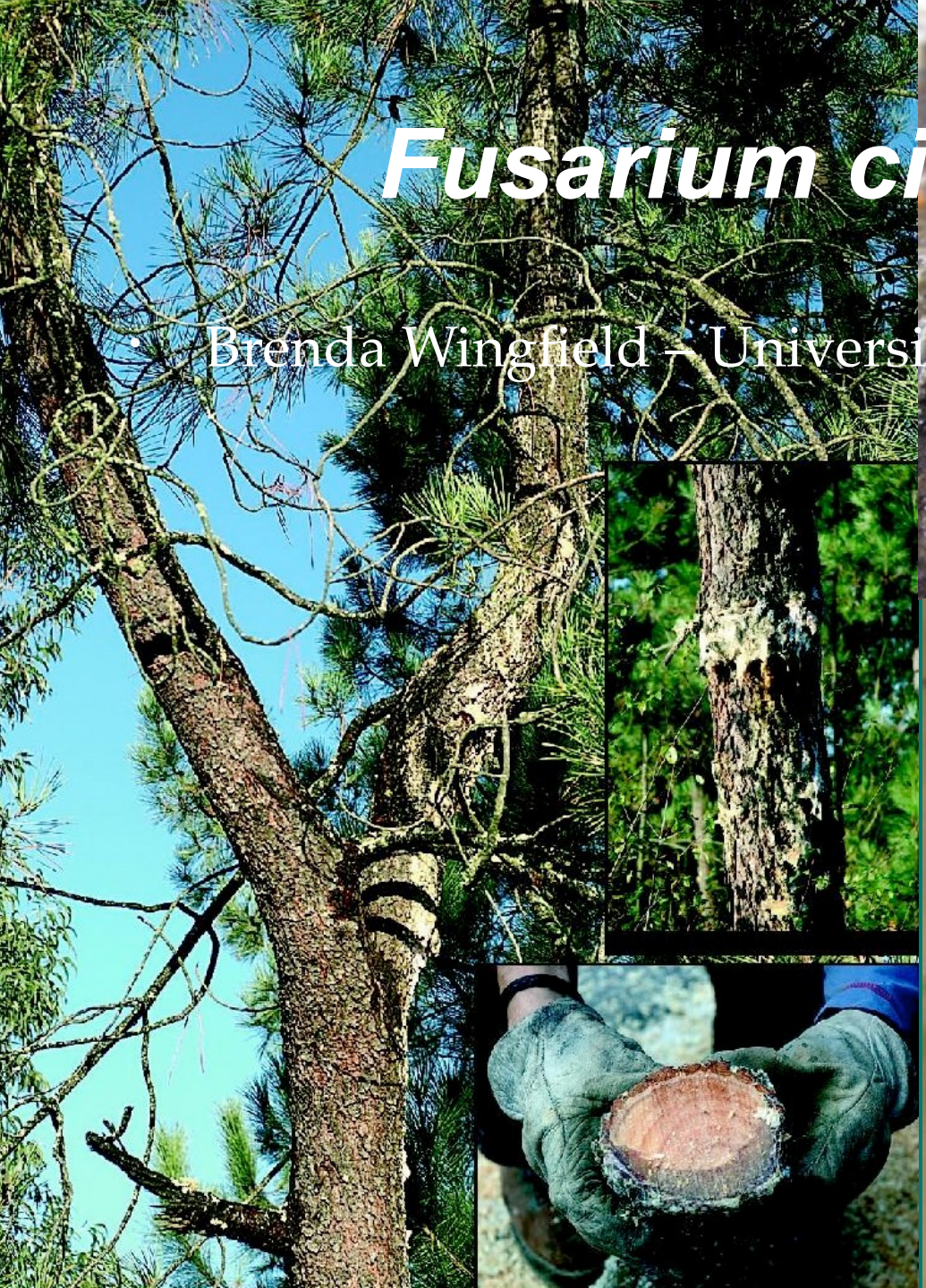
Petromyzon marinus (lamprey)

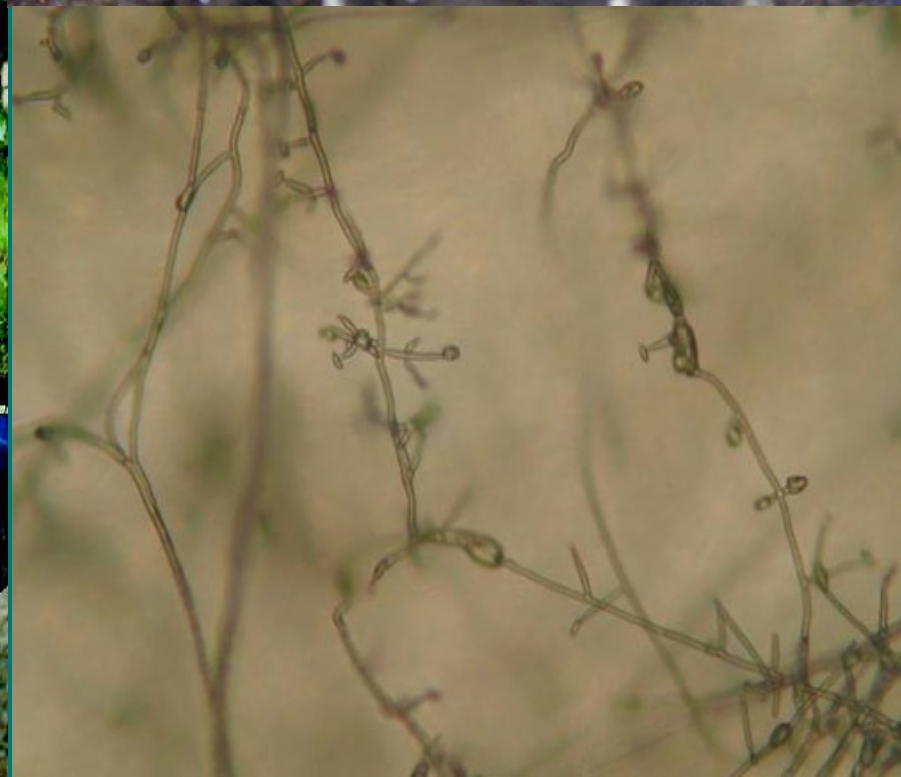
Schmidtea mediterranea (flatworm)

MAKER is being used by over 700 other projects world wide

Fusarium circinatum

Brenda Wingfield – University of Pretoria, South Africa





Fusarium circinatum

Atta Cephalotes

- Cameron Currie – University of Wisconsin-Madison

Pogonomyrmex barbatus

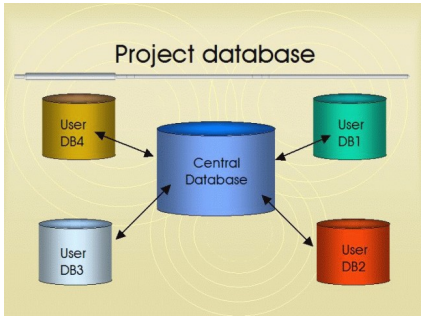
- Christopher D. Smith – San Francisco State University
- Christopher R. Smith – Earlham College
- Jürgen Gadau – Arizona State University
- Neil Tsutsui – University of California Berkeley

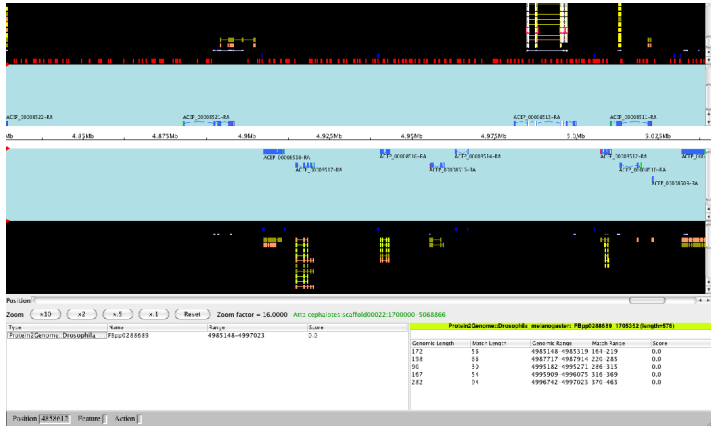
Linepithema humile

- Christopher D. Smith – San Francisco State University
- Christopher R. Smith – Earlham College
- Jürgen Gadau – Arizona State University
- Neil Tsutsui – University of California Berkeley

Pogonomyrmex barbatus, Linepithema humile, and Atta cephalotes

Community Annotation System (CAS)





Pogonomyrmex barbatus, *Linepithema humile*, and *Atta cephalotes*

- Hymenoptera Genome Database
  - http://hymenopteragenome.org/

Conclusions

Towards a Turn-key Solution to Genome Annotation

- Easy-to-use
- Fully automated
- Structural and functional annotation
- *De novo* and re-annotation
- Data prioritization and database management
- Integrates with GMOD tools (Chado, Apollo, GBrowse, etc.)

Acknowledgements

- I'd like to thank and recognize all contributions from my former advisor Mark Yandell at the University of Utah, his lab members Barry Moore and Hao Hu. And I would like to thank his former lab members Brandi Cantarel and Hadi Islam for involvment in developing MAKER.
- I would also like to recognize collaborators Ian Korf at UC Davis and his lab member Genis Parra for their contributions.
- Special thank you to Dave Clements and Scott Cain for their contributions to MAKER via GMOD.
- Genome Project Collaborators:
  - Alejandro Sánchez Alvarado - University of Utah, HHMI
  - Robin Buell - Michigan State University
  - Weiming Li - Michigan State University
  - Makadonka Mitreva - Washington University
  - Allen Kovach - UC Davis
  - Brenda Wingfield – University of Pretoria, South Africa
  - Christopher D. Smith – San Francisco State University
  - Christopher R. Smith – Earlham College
  - Jürgen Gadau – Arizona State University
  - Neil Tsutsui – University of California Berkeley
  - Cameron Currie – University of Wisconsin-Madison
- Contributions in Testing and Development
  - Sánchez Alvarado lab, University of Utah - Sofia Robb and Eric Ross
  - Washington University Genome Sequencing Center - John Martin, Kymberlie Hallsworth, Asif Chinwalla, Christopher Schuster, Mark Johnson, and William Nash
- This project is supported by funding from the NHGRI through an RO1 grant entitled *Software for the creation and quality control of genome annotations*.
- This project is supported by an NIH Genetics Training Grant