

MAKER—the easy-to-use genome annotation pipeline

Mark Yandell
Eccles Institute of Human Genetics
University of Utah & School of Medicine

Overview

- Issues in Genome Annotation
- MAKER
- Annotating the *S. mediterranea* genome
- Some Comparative Genomics
- Some Functional Genomics: genome-scale image-based RNAi screen

Overview

- Issues in Genome Annotation
- MAKER
- Annotating the *S. mediterranea* genome
- Some Comparative Genomics
- Some Functional Genomics: genome-scale image-based RNAi screen

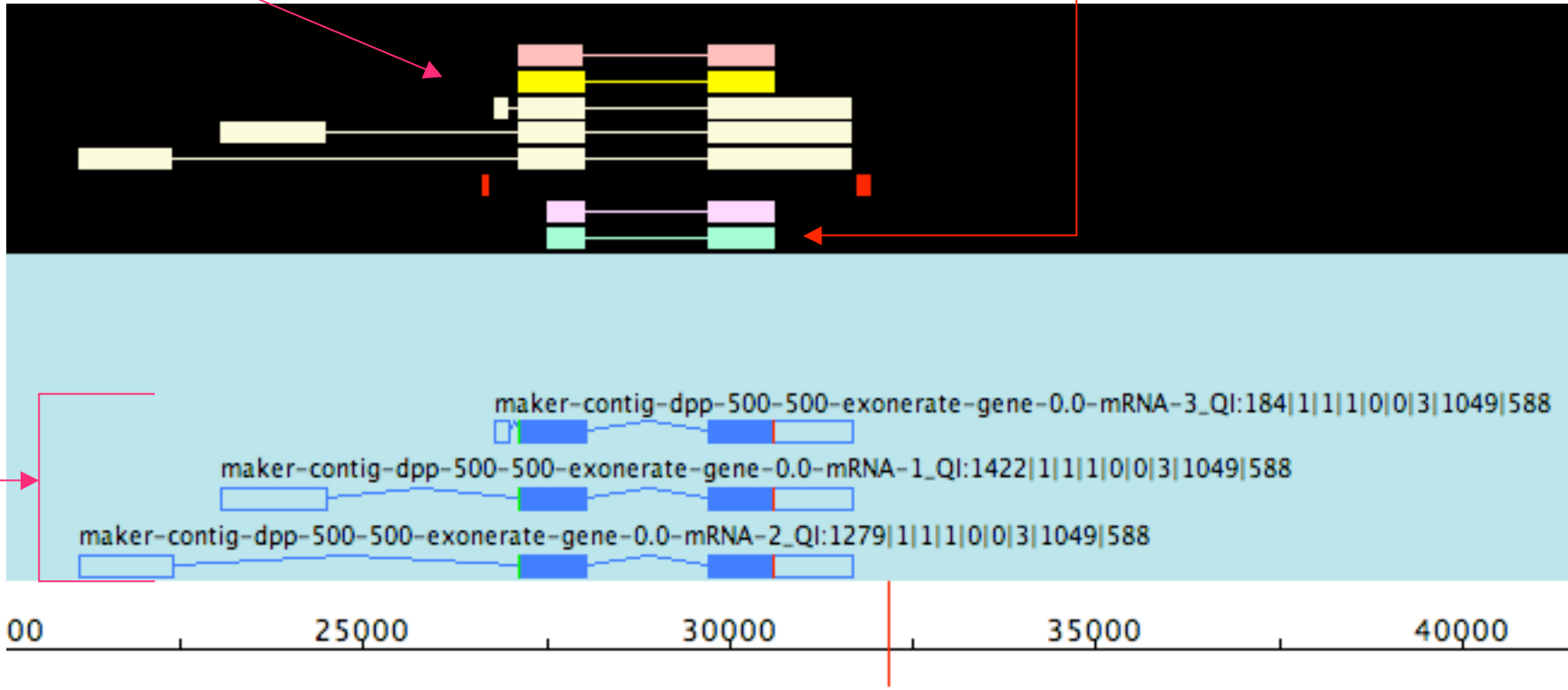
Why annotations are important

- Genome sequences not too useful in themselves
- Genes are the reason genomes are sequenced
- Gene annotations are crucial for downstream applications
- Incorrect gene annotations poison every experiment that makes use of them

What is a gene annotation?

Computational evidence

Gene-predictions



Gene annotation

gene prediction \neq gene annotation

Within 5 years thousands of plant and animal genomes will be sequenced

- Fall in costs will enable individual investigators to sequence genomes
- Most will fall into the 'emerging model organism' category
- We are developing software to meet the needs of these projects

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

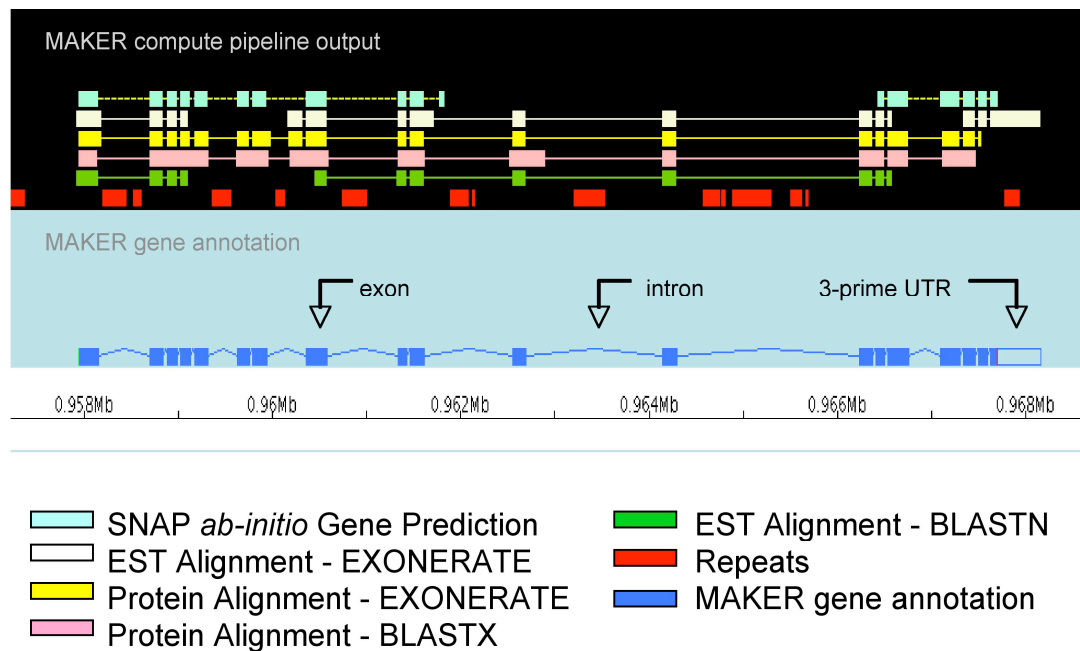
Overview

- Issues in Genome Annotation
- **MAKER**
- Annotating the *S. mediterranea* genome
- Some Comparative Genomics
- Some Functional Genomics: genome-scale image-based RNAi screen

MAKER genome annotation tool

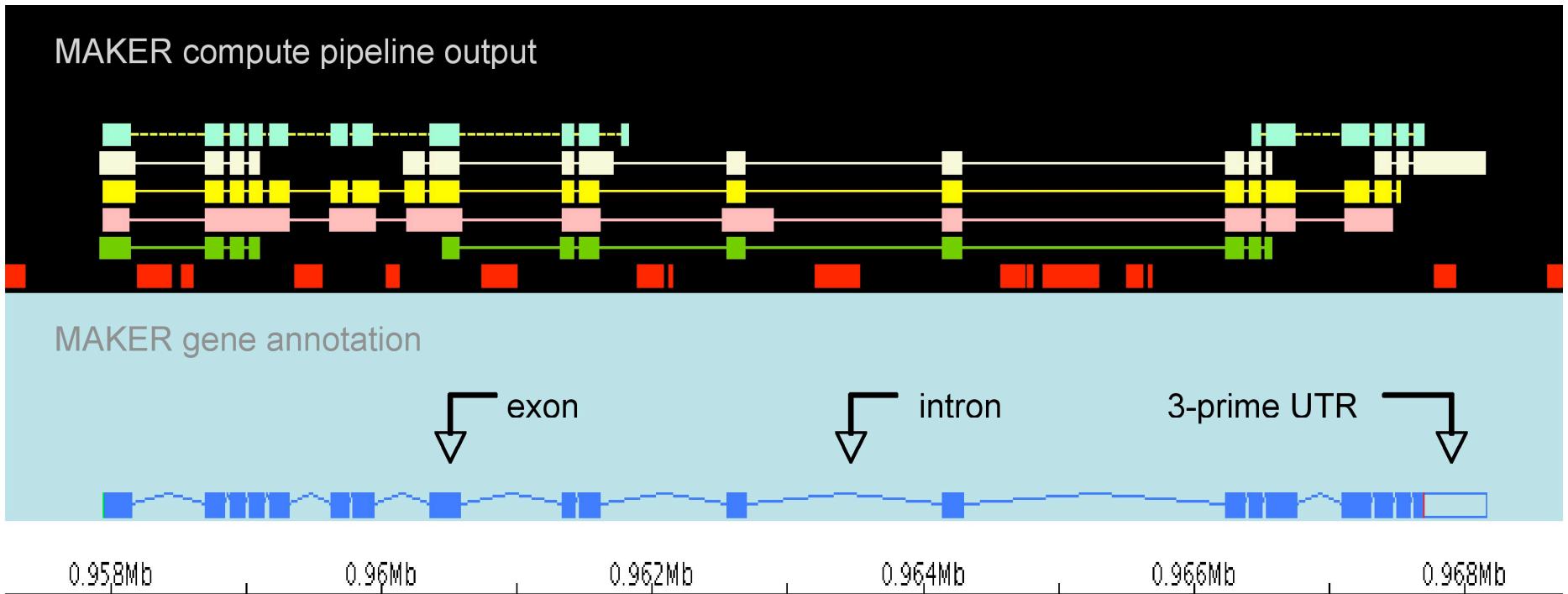




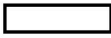




- ✓ Easy to use; excellent user support
- ✓ Annotate a single BAC or an entire genome
- ✓ View outputs directly in Apollo and GBrowse
- ✓ Automatically generate a GMOD database from MAKER output



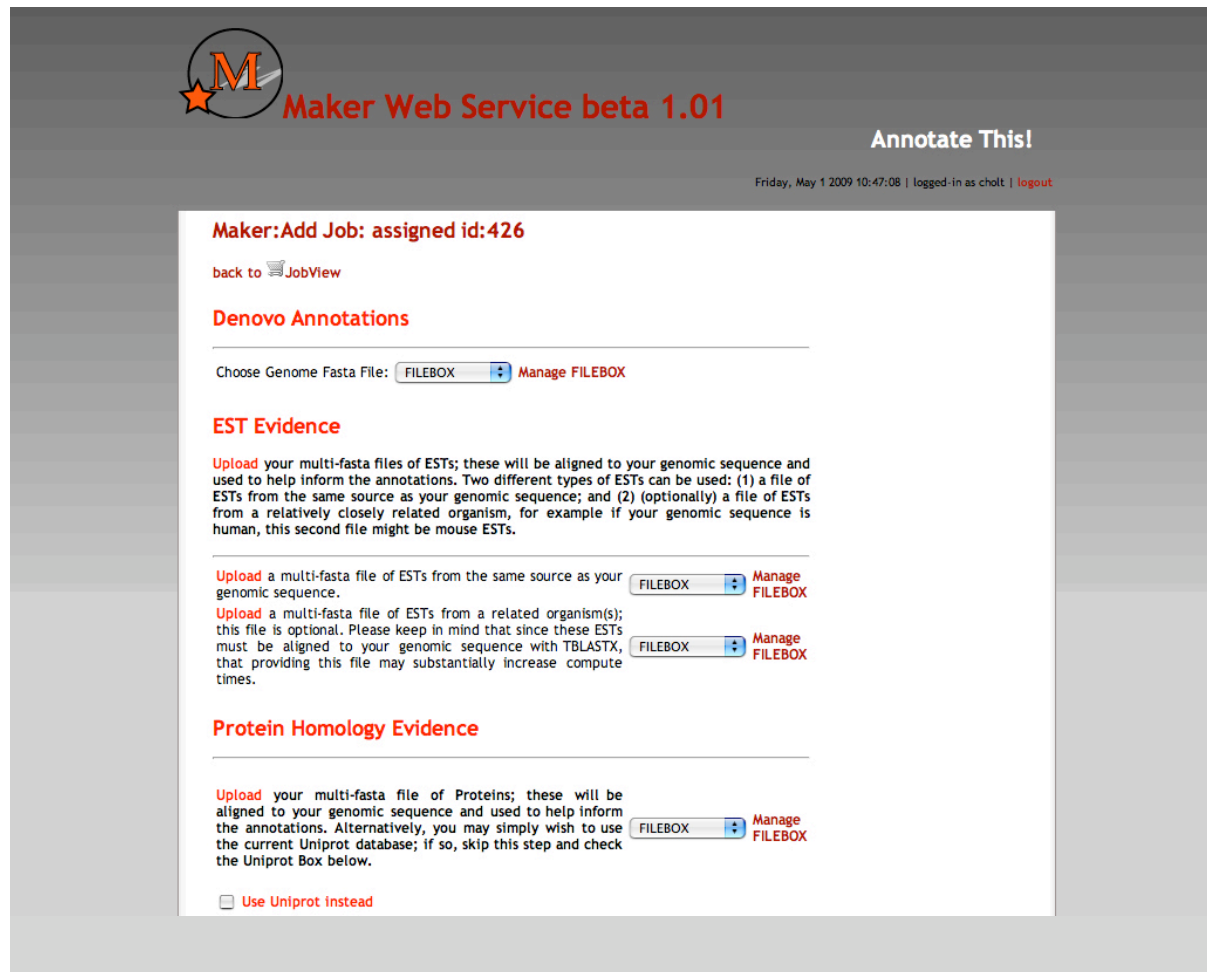
MAKER identifies repeats, aligns ESTs and proteins to a genome, produces *ab-initio* gene predictions, and automatically synthesizes these data into gene annotations www.yandell-lab.org


What MAKER does



- | | |
|---|--|
|  SNAP <i>ab-initio</i> Gene Prediction |  EST Alignment - BLASTN |
|  EST Alignment - EXONERATE |  Repeats |
|  Protein Alignment - EXONERATE |  MAKER gene annotation |
|  Protein Alignment - BLASTX | |

MAKER will soon be available over the web




 **Maker Web Service beta 1.01**

Annotate This!

Friday, May 1 2009 10:47:08 | logged-in as chott | [logout](#)

Maker:Add Job: assigned id:426

[back to](#)  JobView

Denovo Annotations

Choose Genome Fasta File: [Manage FILEBOX](#)

EST Evidence

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

Upload a multi-fasta file of ESTs from the same source as your genomic sequence. [Manage FILEBOX](#)

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

Protein Homology Evidence

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

[Use Uniprot Instead](#)

Who is Using MAKER?

Current Genome Project Collaborations



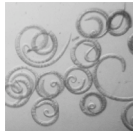
Schmidtea mediterranea (flatworm)



Pythium ultimum var ultimum (potato rot)



Petromyzon marinus (lamprey)



Trichinella Spiralis (parasitic nematode)



Trichuris suis (parasitic nematode)



Pinus taeda (pine tree)



Conus bullatus (cone snail)

MAKER is being used by over 200 projects worldwide.

Issues peculiar to annotating emerging genomes

- Needs to be easy
- Repeat identification/masking
- Gene finder training

MAKER's performance on REFERENCE genomes is comparable to other-state-of-the-art pipelines

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

MAKER's performance on REFERENCE genomes is comparable to other-state-of-the-art pipelines

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%

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

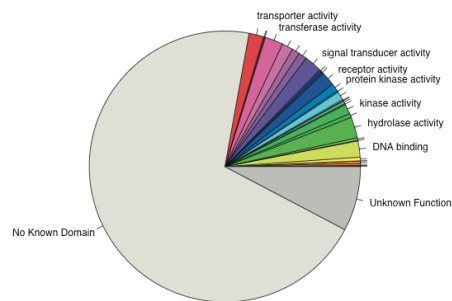
EXACT UTTRS					
SP	6.40%	9.86%	11.75%	8.05%	11.40%
SN	31.36%	44.20%	40.53%	23.63%	46.03%
Exact all exons					
SP	19.02%	22.08%	22.44%	34.08%	24.19%
SN	93.48%	98.98%	95.62%	91.24%	98.57%
Start stop					
SP	7.05%	12.97%	12.69%	11.87%	17.79%
SN	35.95%	51.83%	47.76%	34.42%	72.51%

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 ensemble.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

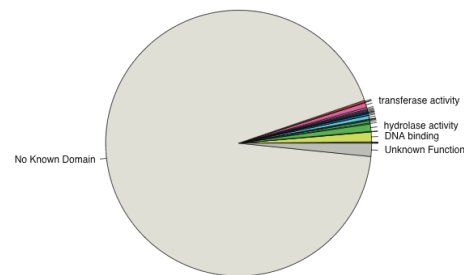
ab-initio gene predictors don't do nearly so well on emerging genomes*

**Average of seven
REFERENCE proteomes**



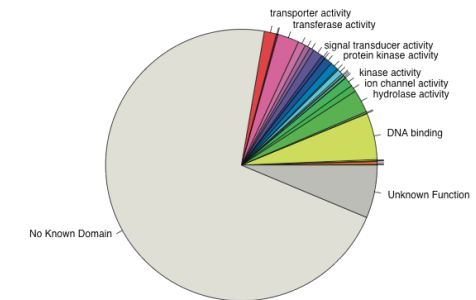
35% contain a domain

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



7% contain a domain

**MAKER *S. mediterranea*
annotations**

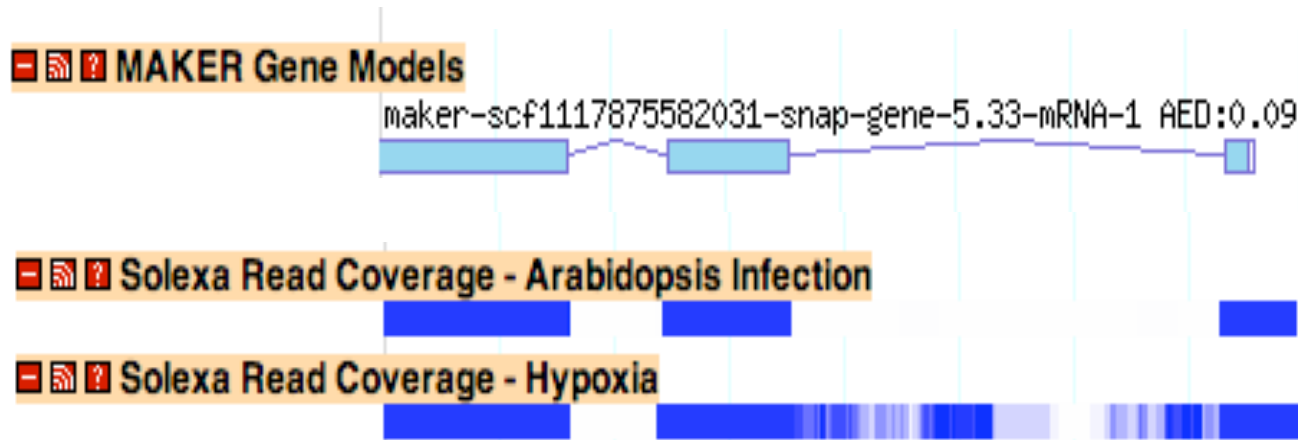


29% contain a domain

***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

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

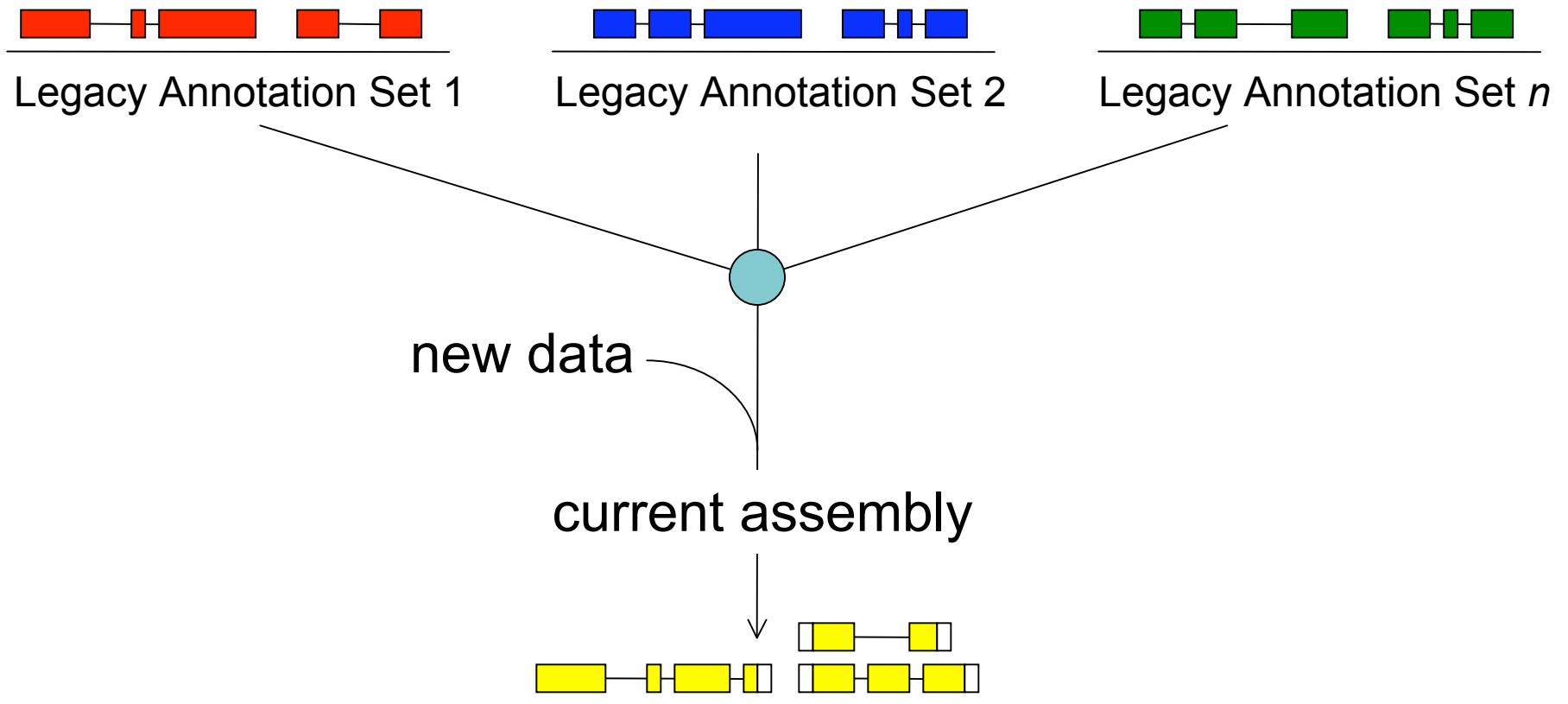


- Still need to de-convolute reads & evidence (for now)
- Still need to archive and distribute annotations
- Still need to manage genome and its annotations

Another issue: legacy annotations

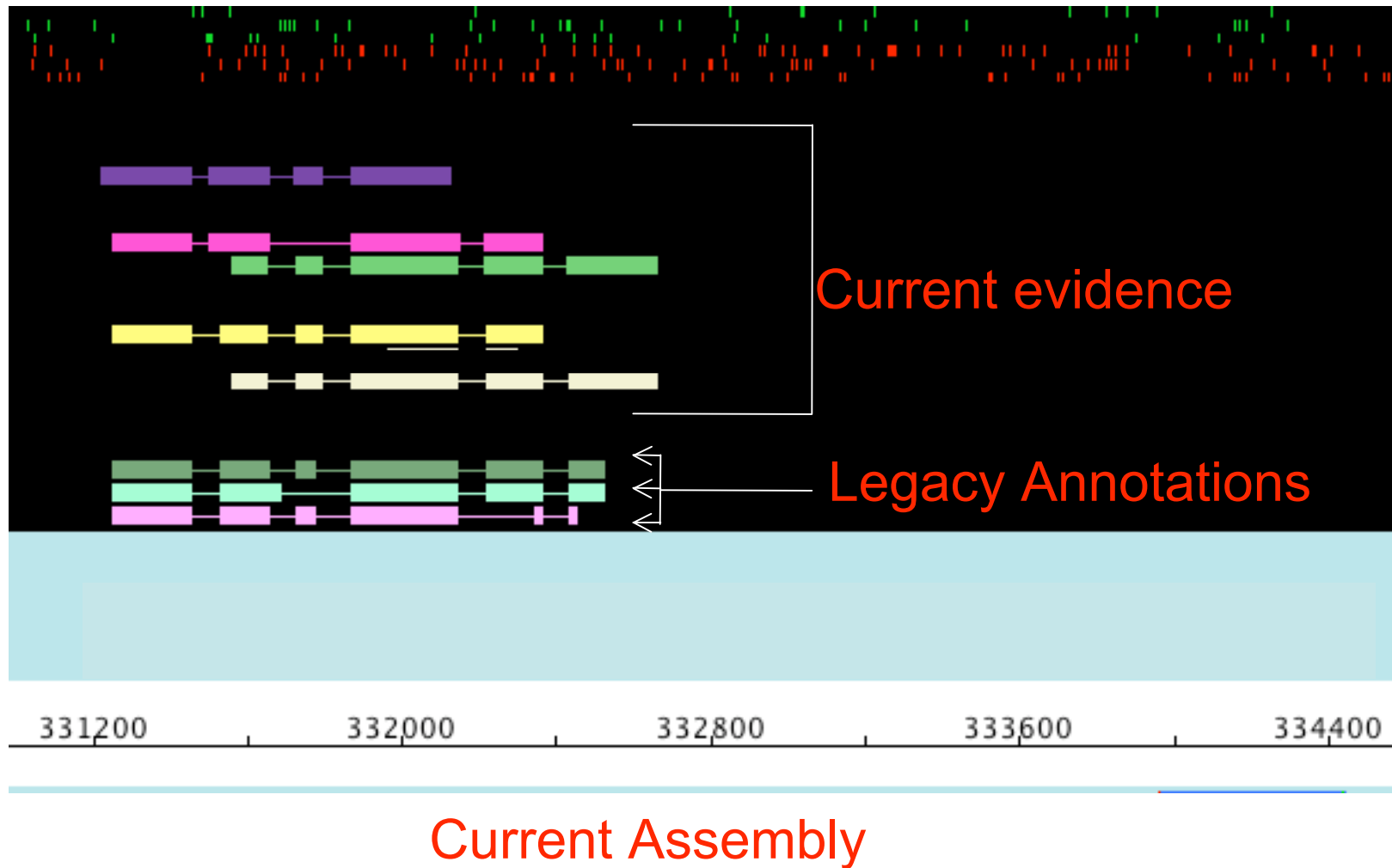
- Many are no longer maintained by original creators
- In some cases more than one group has annotated the same genome, using very different procedures, even different assemblies
- The communities associated with those genomes are going to want RNA-seq data
- Many investigators have their own genome-scale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

Merging and Revising Legacy Annotation Sets

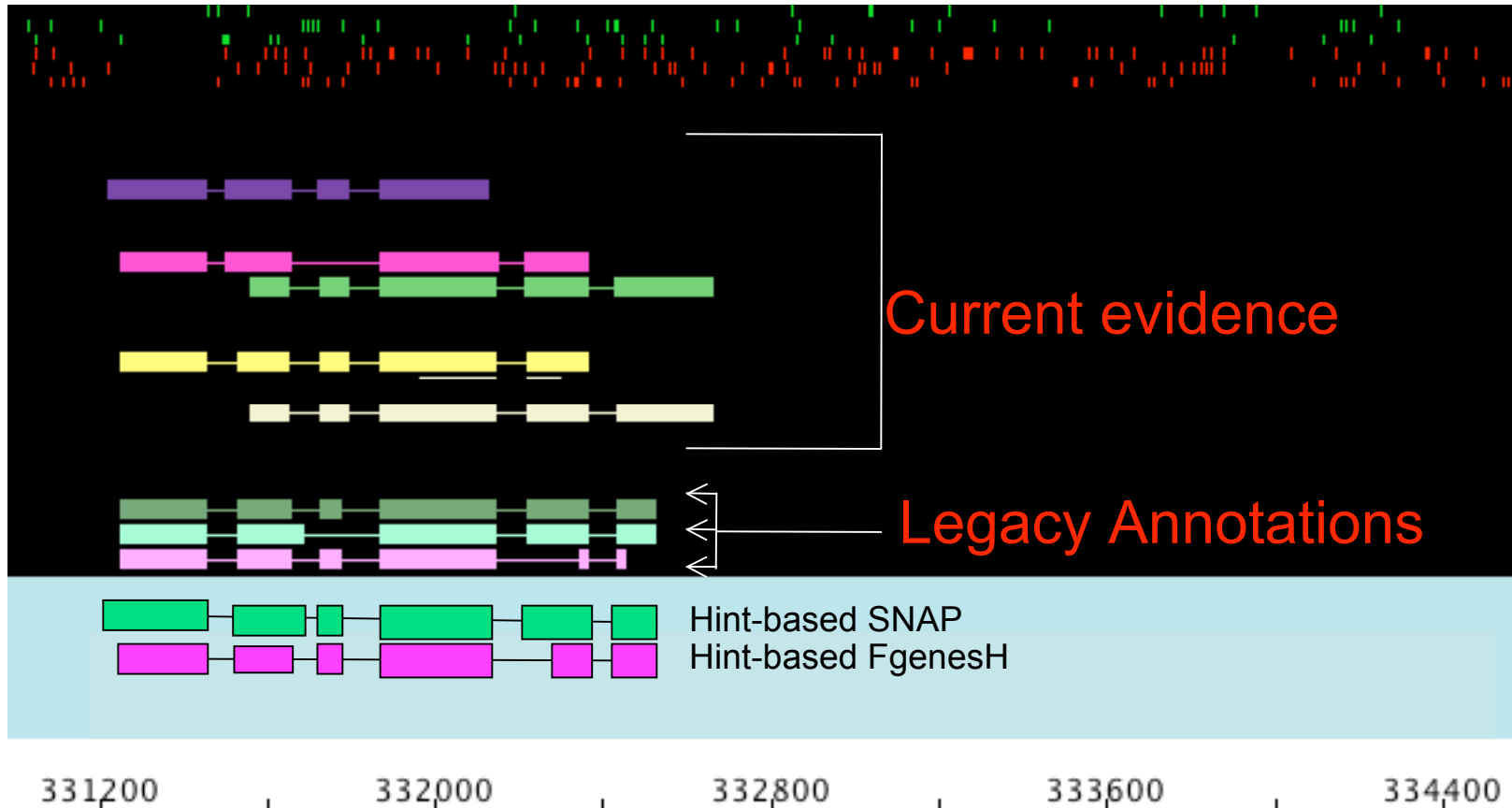


- Identify legacy annotation most consistent with new data
- Automatically revise it in light of new data
- If no existing annotation, create new one

Align Evidence and Legacy Annotations to Current Assembly

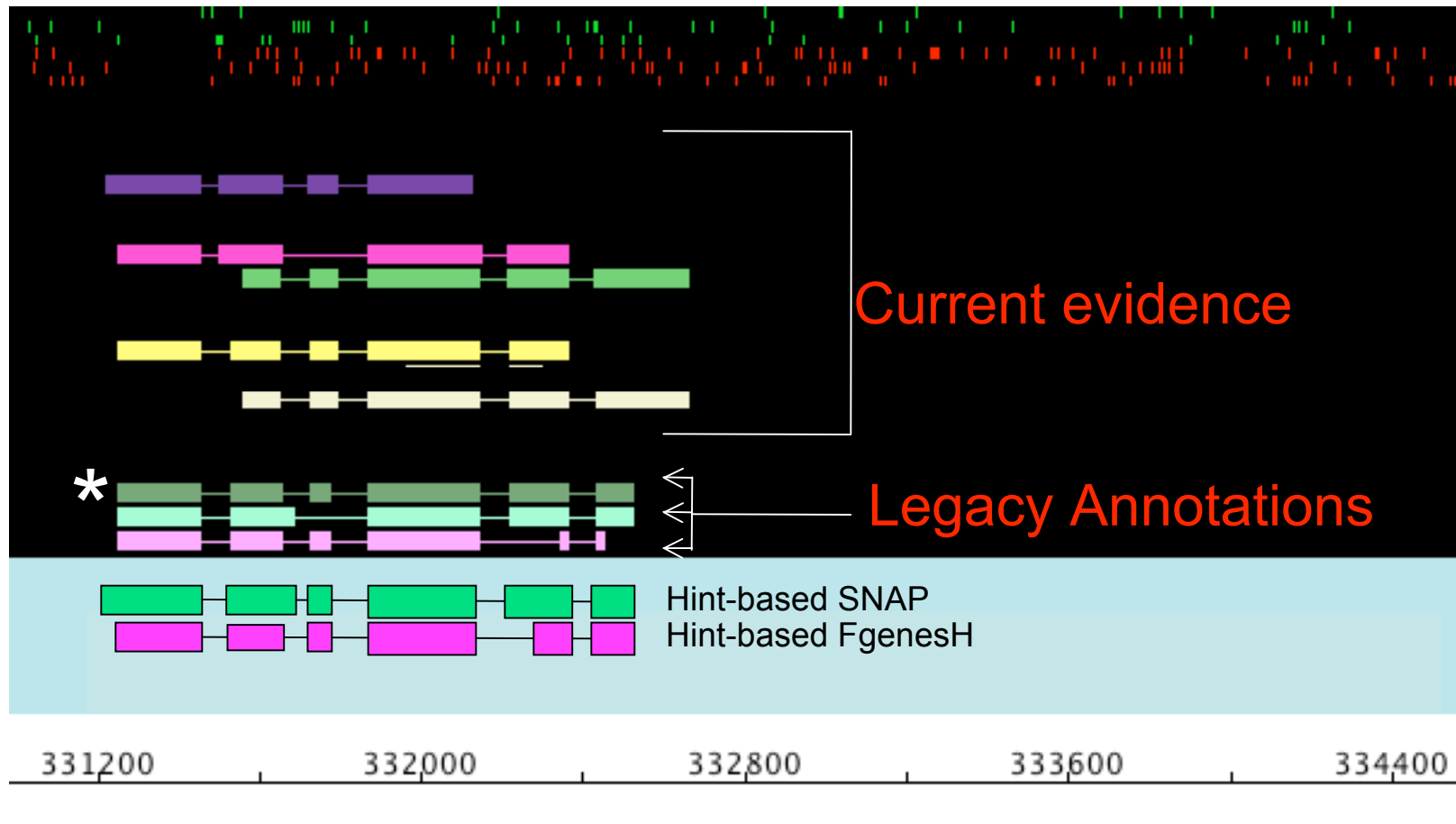


Pass Gene Finders Evidence-based 'hints'



Current Assembly

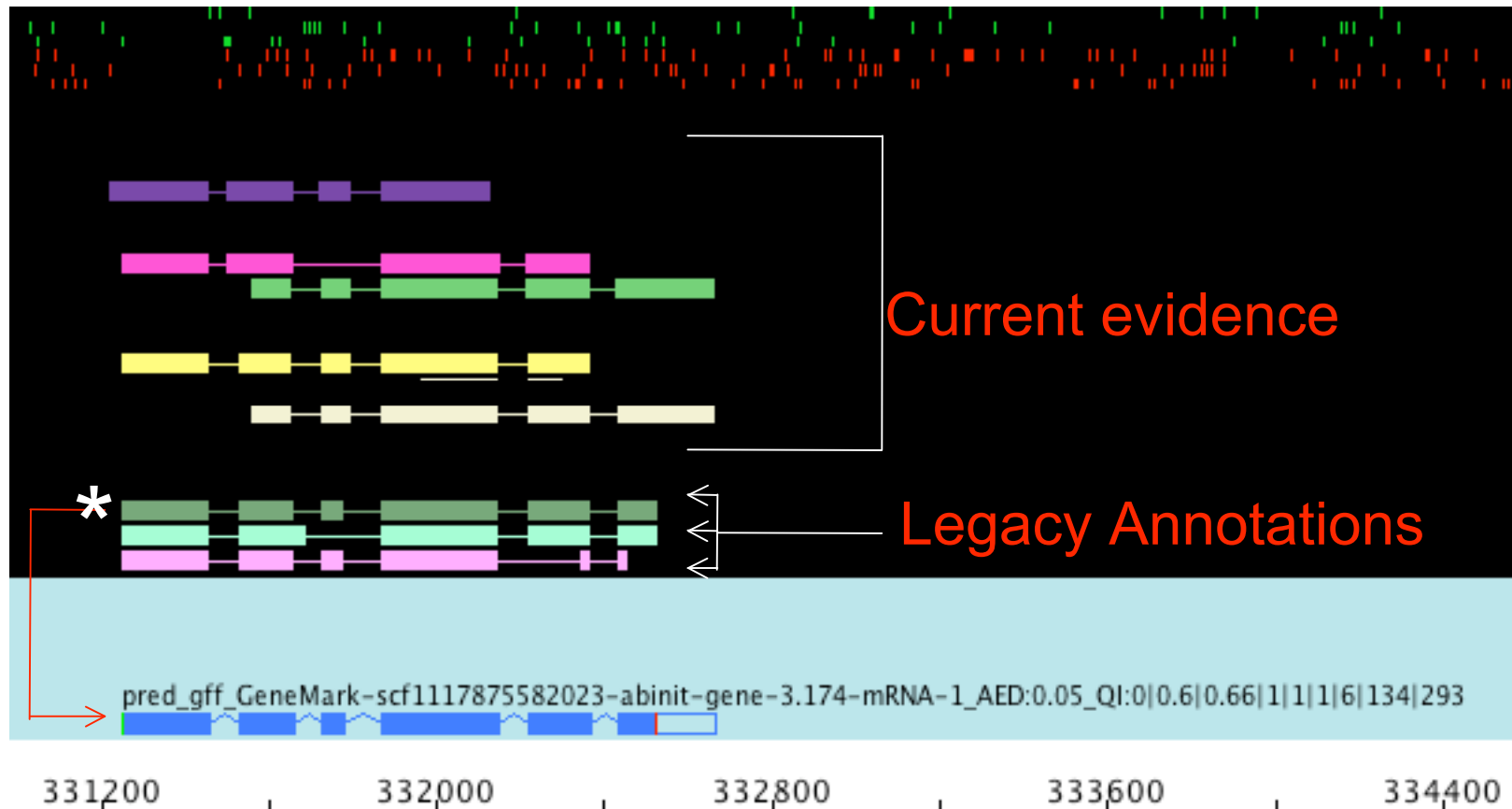
Identify Gene Model Most Consistent with Evidence*



Current Assembly

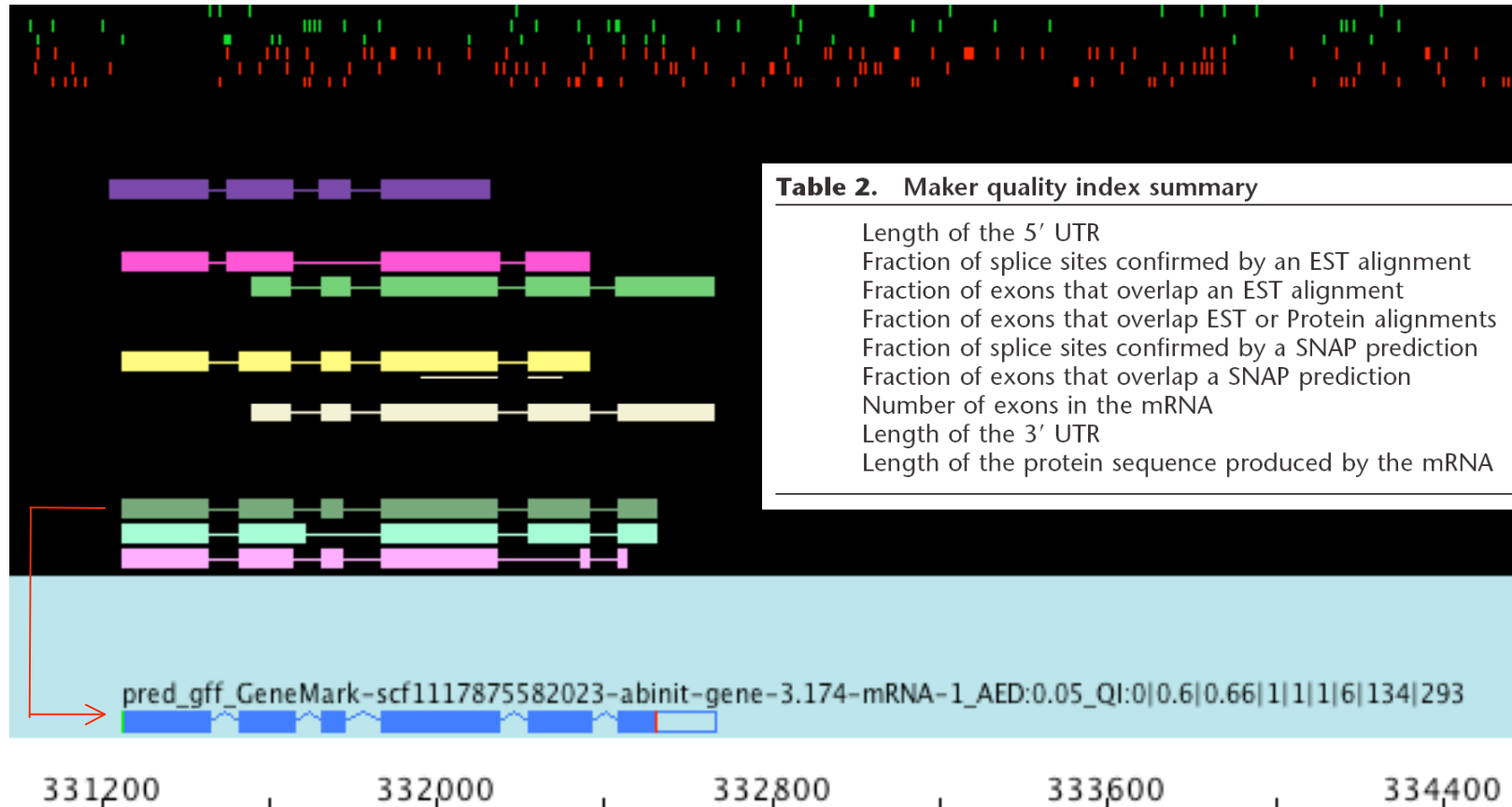
***Quantitative Measures for the Management and Comparison of Annotated Genomes**
Karen Eilbeck , Barry Moore , Carson Holt and Mark Yandell BMC Bioinformatics 2009
10:67doi:10.1186/1471-2105-10-67

Revise it further if necessary; Create New Annotation



Current Assembly

Compute Support for each portion of Gene Model



Overview

- Issues in Genome Annotation
- MAKER
- Annotating the *S. mediterranea* genome
 - ✓ a typical emerging genome
- Some Comparative Genomics
- Some Functional Genomics: genome-scale image-based RNAi screen

Proof of principle on an emerging genome:

The *S. mediterranea* genome



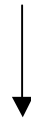
Collaboration with the
Sánchez-Alvarado lab

Why work on *S. mediterranea* ?

- Easy to grow and maintain.
- Invertebrate model of choice for wound healing and regeneration, e.g. *they have stem cells*
- No genetics, but good mol. bio *including RNAi*
- Simple morphology makes them ideal for image based screens– *they are flat!*

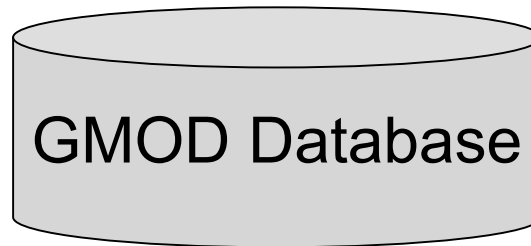
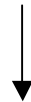
Automatic creation of the *S. mediterranea* genome database

Genome sequence

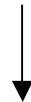


MAKER

GFF3



GMOD Database



Community access

MAKER's automatic annotations provide a starting point for further work

THE UNIVERSITY OF UTAH
Schmidtea mediterranea Genome Database

Home | Genome Browser | Blast | Blat | Search | Help

AAATCAATCT GTGGTGGCAT
CAGCC CCAAATAAGA TCCAC
GATTTAGCAA AGTTGAACCT
TCAAGGCCTGTT ATAG
AGAT CCGAATGT ATGTTG
GGTAAATTCC CTTAGTA
ACAATCTC GTAT
CACGGAAATA CACGAGA
ACTGAA CGT AA AACCA
TGCCAATG TTTTTTGA
TATC GGGAAATATG
AAGT ATTCAAT CCAA
ATCATTATCC AATCGAAGAA

Chromosome Start End

no_name

20000 25000 30000

Planarians are free-living (non-p of all three germ layers (i.e., ec renowned for their development generations of biologists (<http://>

Among all flatworm species stu rapidly emerging as a key mod homeostasis and stem cell biology. The Schmidtea mediterranea Genome Database (SmedGD) is a GMOD compliant database that integrates in a single web-accessible portal all available data associated with the planarian genome, including predicted and annotated genes, ESTs, protein homologies, gene expression patterns and RNAi phenotypes.

If you use SmedGD in your research please reference the following citation:
Sofia M.C. Robb, Eric Ross and Alejandro Sánchez Alvarado (2007) SmedGD: the Schmidtea mediterranea Genome Database Nucleic Acids Research, doi:10.1093/nar/gkm684 (In Press)

To take full advantage of what SmedGD has to offer, we recommend reading the [Help](#)

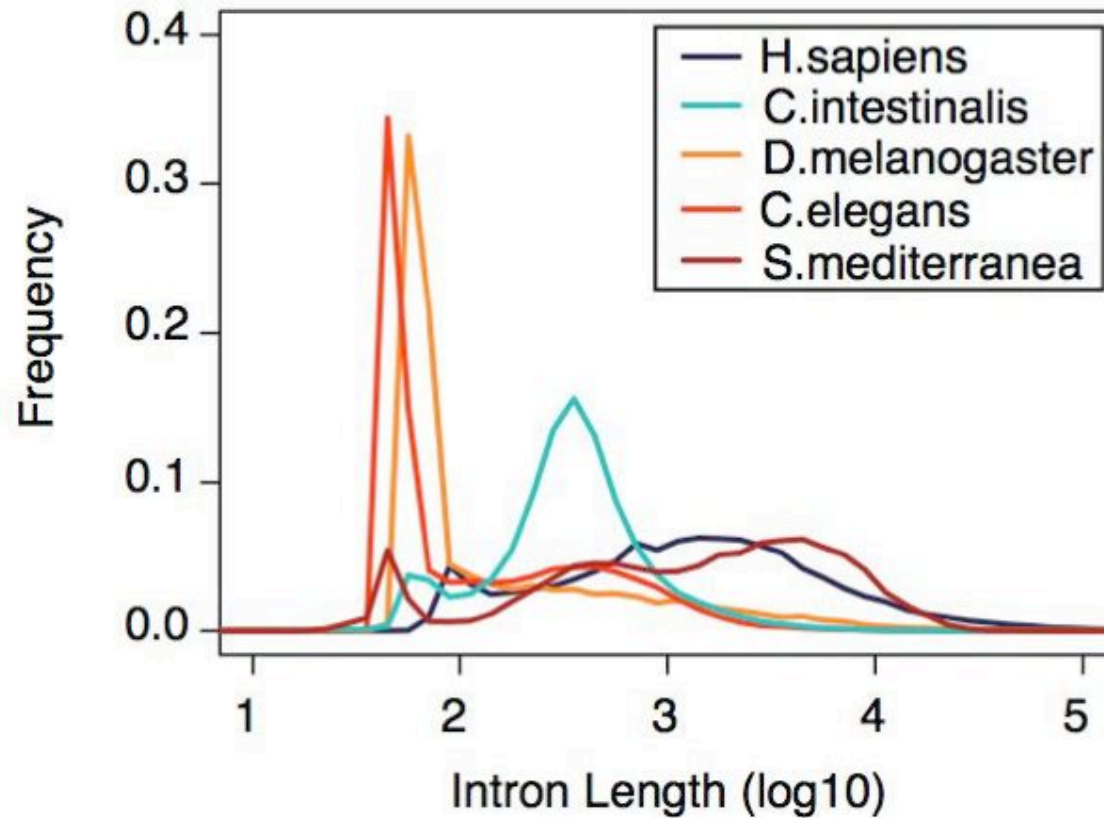
smedgd.neuro.utah.edu

What we have found out to date.

Comparative genomics

The *S. mediterranea* genome from a comparative and evolutionary perspective

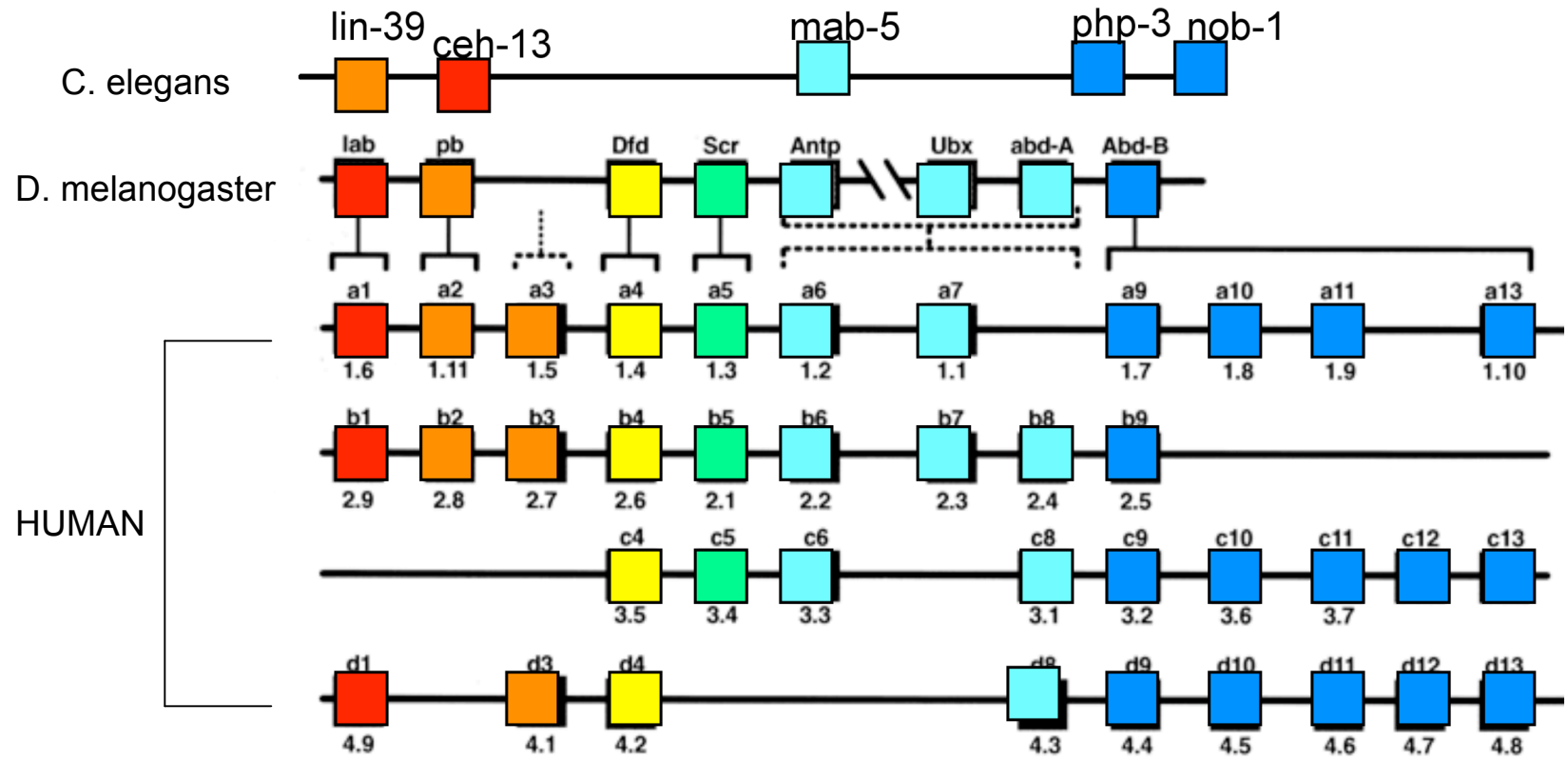
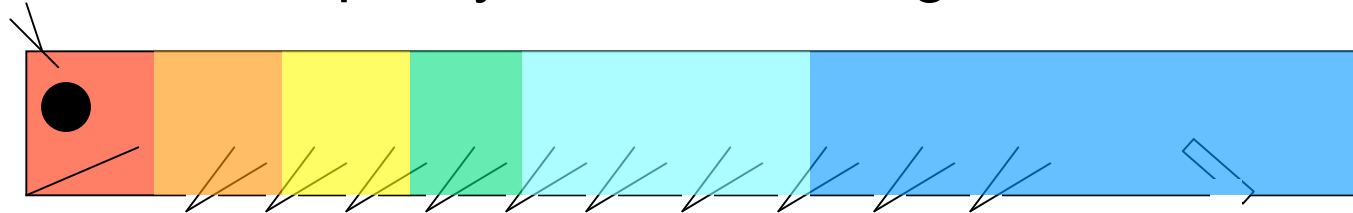
S. mediterranea introns are large and abundant



In these respects its genes are surprisingly deuterostome-like

Many of *S. mediterranea*'s gene families are surprisingly deuterostome-like as well, c.f. HOX genes.

HOX genes provide a conserved system used to specify cell fates along the A-P axis

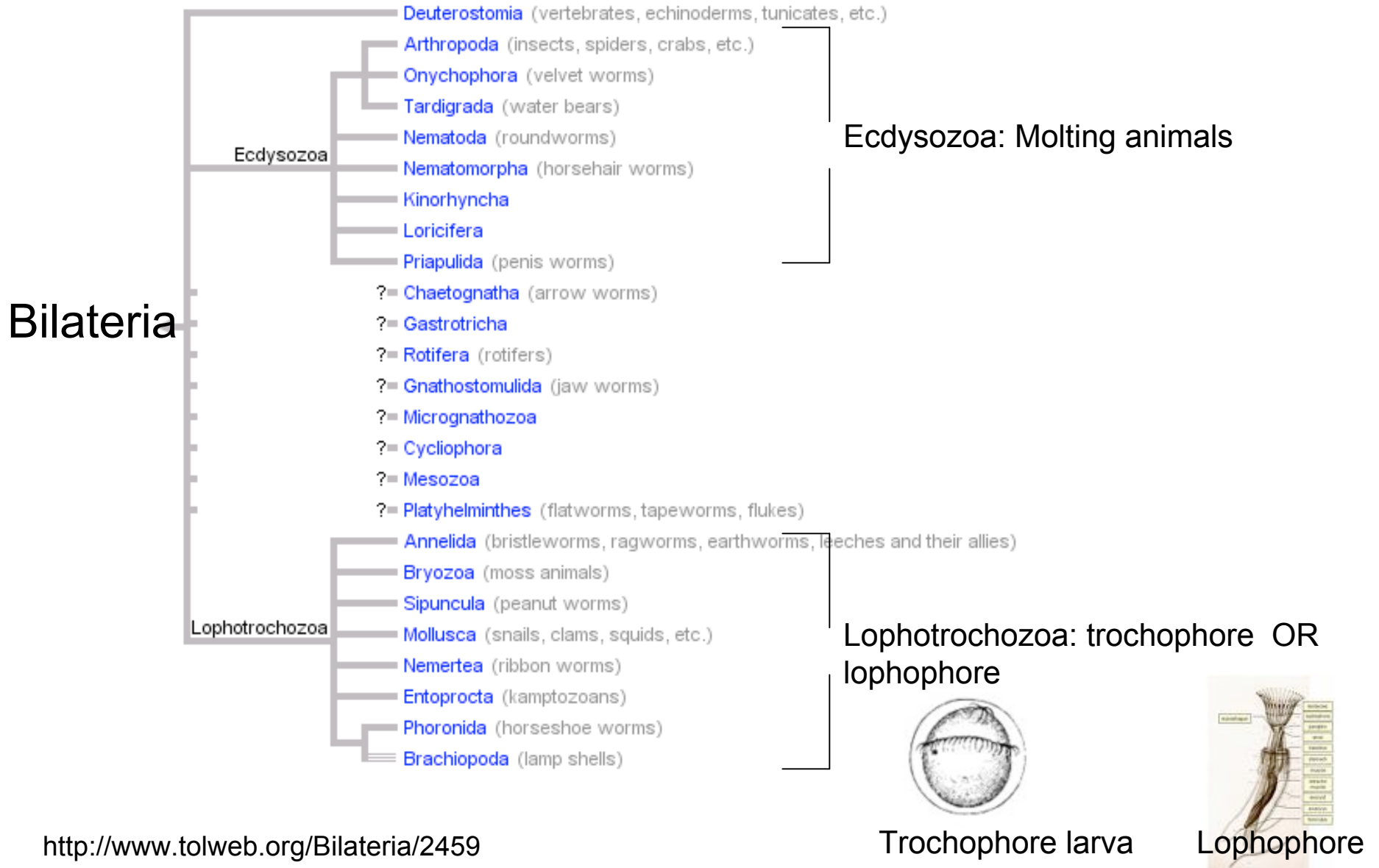


The *S. mediterranea* HOX family is surprisingly complex, and *deuterostome-like*

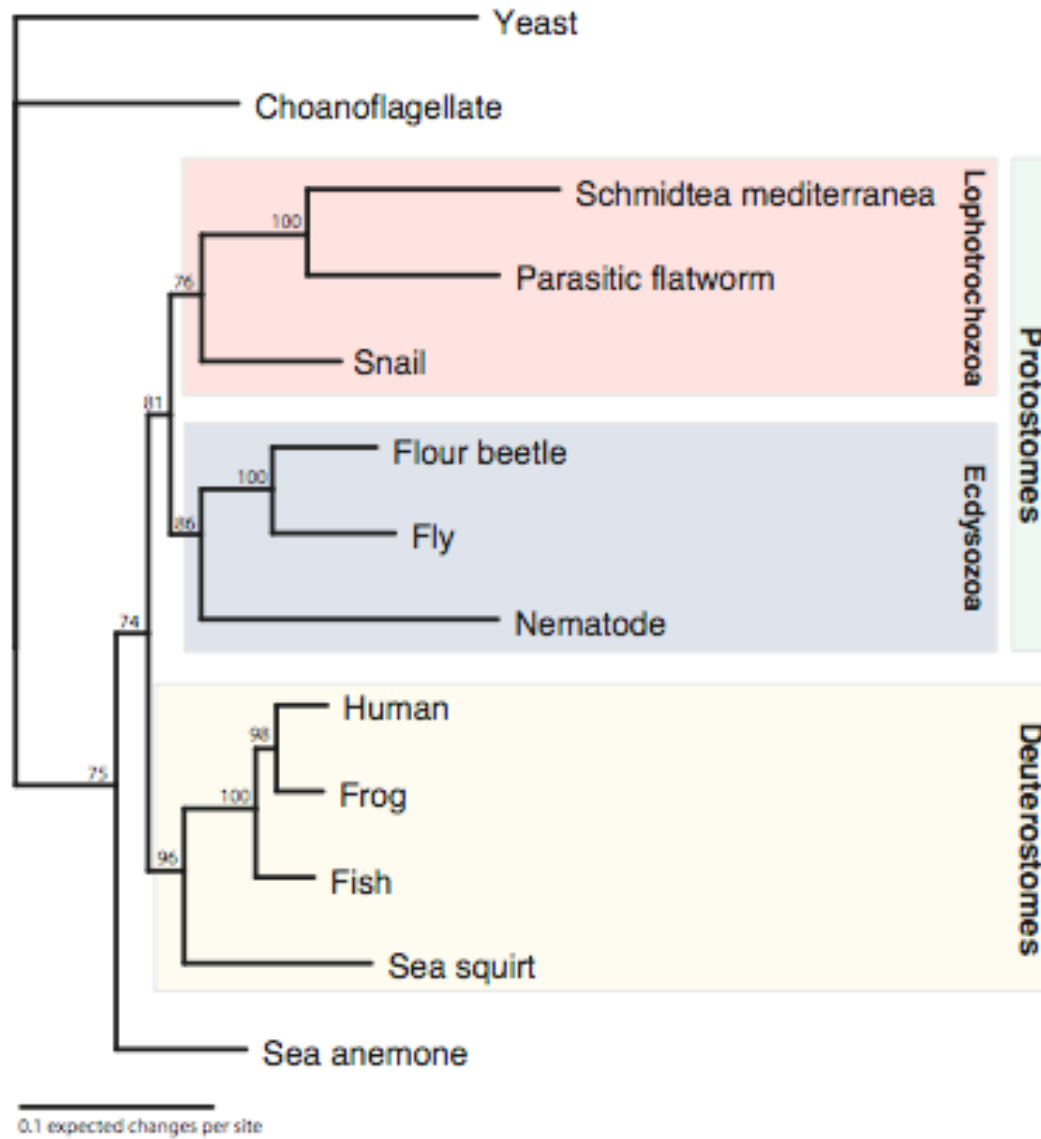


True for many gene families, including those known to be involved with RNAi and Stem cells & regeneration

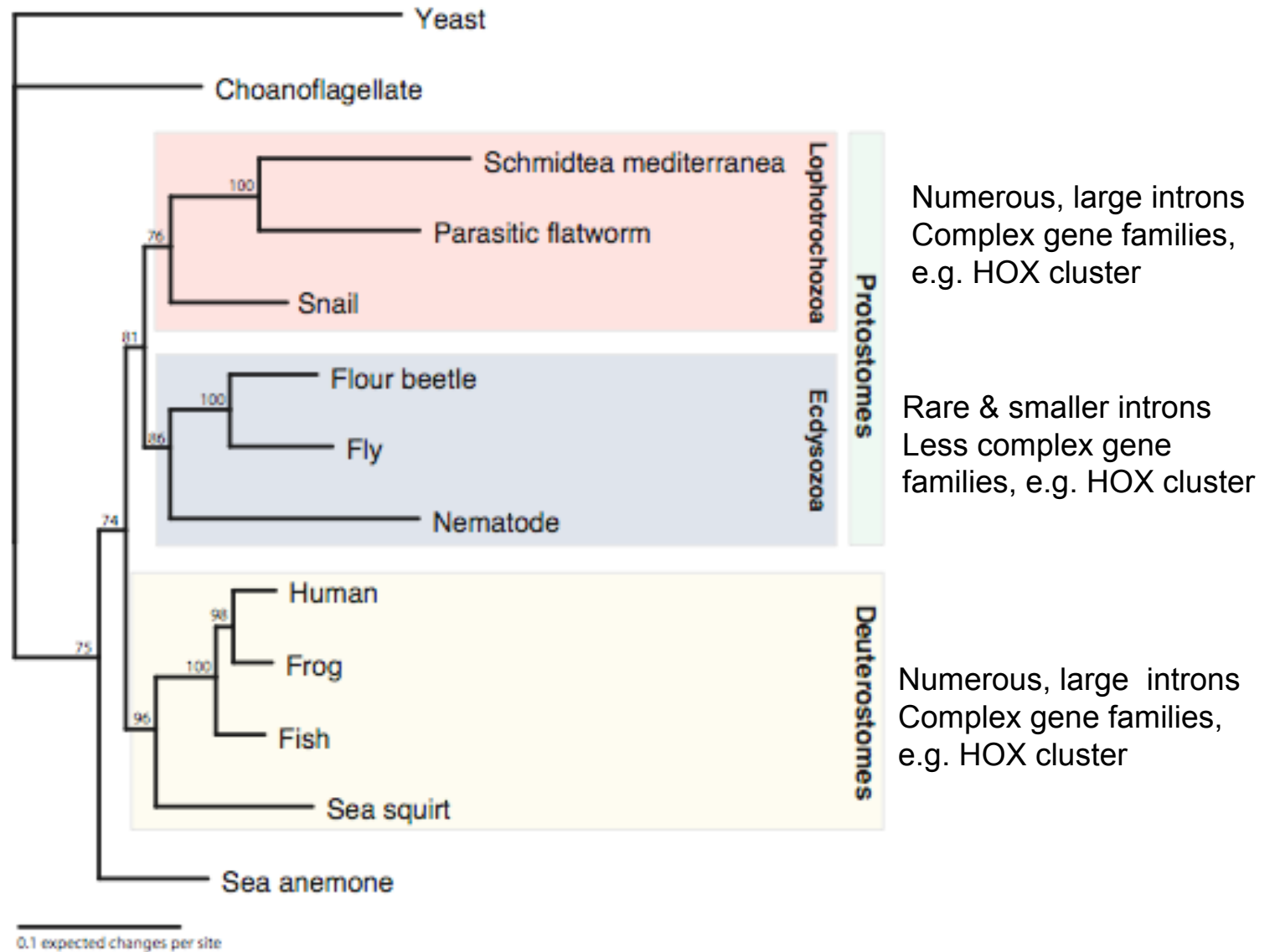
Whither Flatworms?



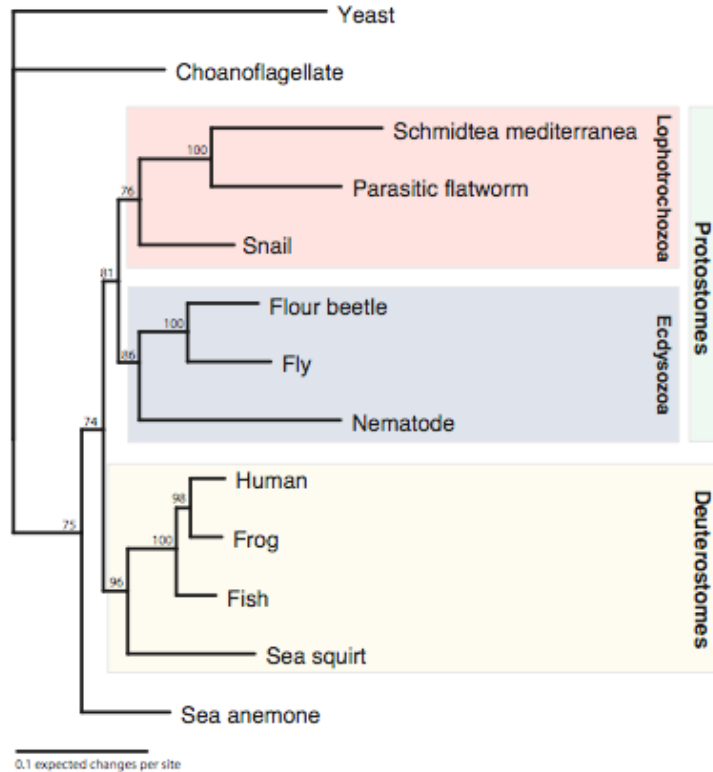
Where flatworms sit in the tree of life



Where flatworms sit in the tree of life



Where flatworms sit in the tree of life



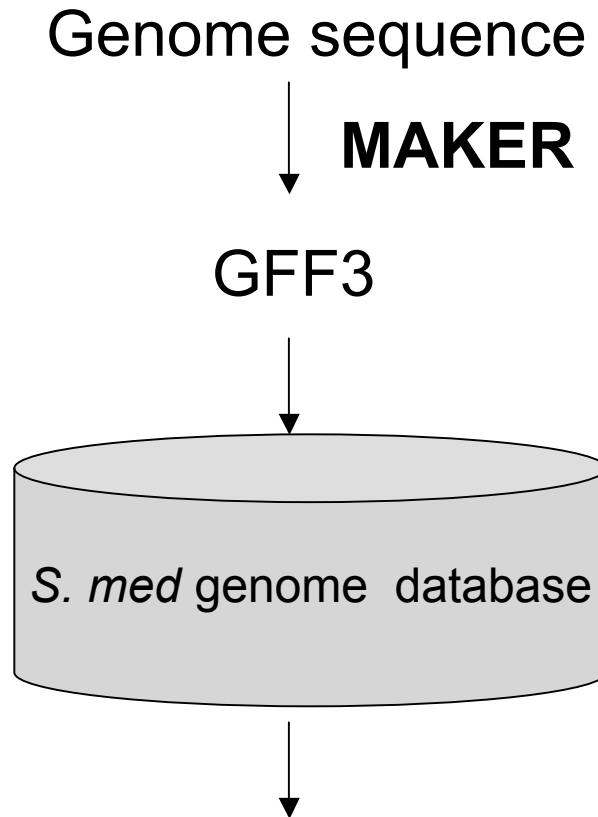
conclusions:

- Ancestral bilaterian genome likely had numerous, large introns.
- Elaborate HOX gene repertoire.
- *D. melanogaster* & *C. elegans* genomes highly streamlined.

Overview

- Issues in Genome Annotation
- MAKER
- Annotating the *S. mediterranea* genome
- Some Comparative Genomics
- **Some Functional Genomics:** genome-scale image-based RNAi screen

What we want to do with the *S. mediterranea* annotations:



Genome-wide, image-based RNAi screens for genes involved in tissue homeostasis & regeneration

‘Functional genomics application’

Surfing the data deluge

High-throughput
Sequencing



ABI, Solexa, 454, etc.

High-throughput
Imaging



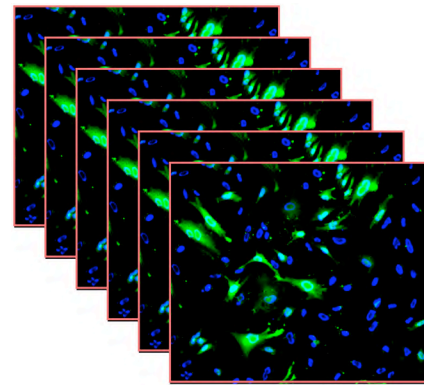
BD Pathway

The data deluge

High-throughput
Imaging

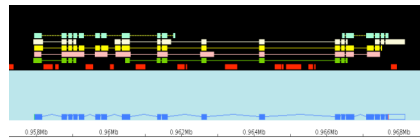


BD Pathway



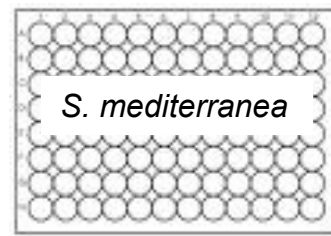
Reams of
digital images

A whole genome screen for novel regulators of tissue homeostasis and regeneration



Genome Annotations

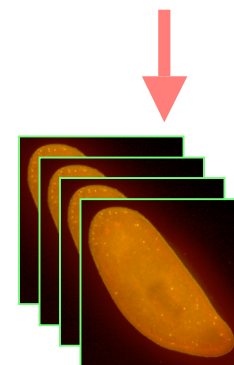
↓
RNAi
constructs



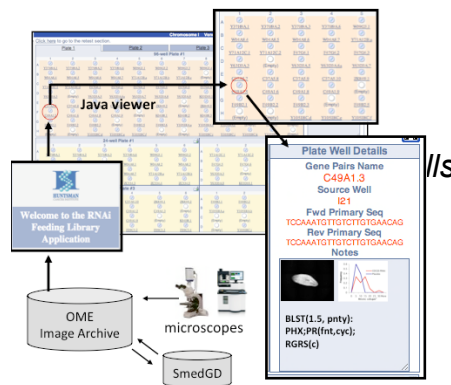
96-well plates



Automated confocal
microscope



Large numbers of images

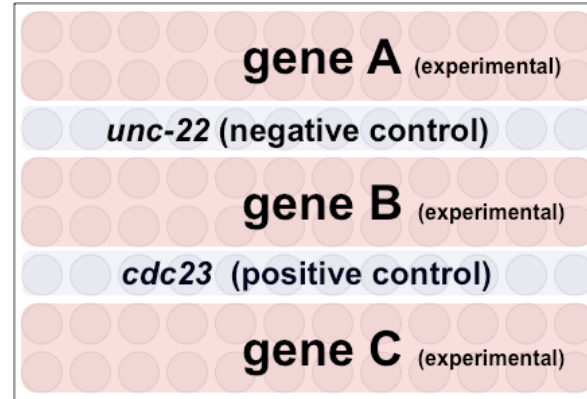
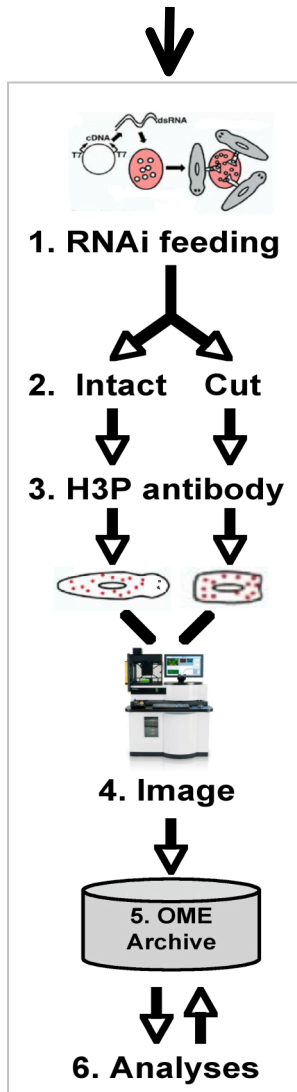


OME Archive & Automated image
processing

Details of the screen

5000 Planarian genes

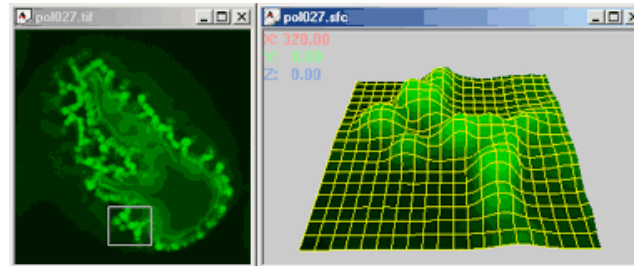
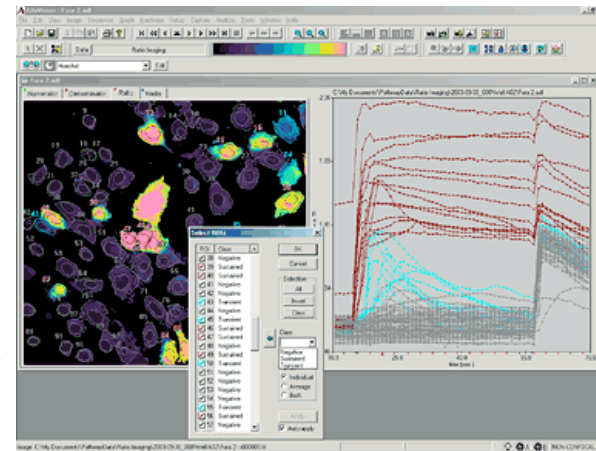
homologous to human genes, not yet implicated in homeostasis or regeneration



		Density of H3P ⁺ nuclei in whole animals		
		Up	Down	Unchanged
Density of H3P ⁺ nuclei in cut animals	Up	W+, C+ General regulators of cell division	W-, C+ Reciprocal Regulators (Class 3)	W=, C+ Regeneration Specific (Class 2)
	Down	W+, C- Reciprocal Regulators (Class 3)	W-, C- General regulators of cell division	W=, C- Regeneration Specific (Class 2)
	Unchanged	W+, C= Homeostasis Specific (Class 1)	W-, C= Homeostasis Specific (Class 1)	W=, C= Uninvolved

Most image analysis software is still GUI-based (point & click)

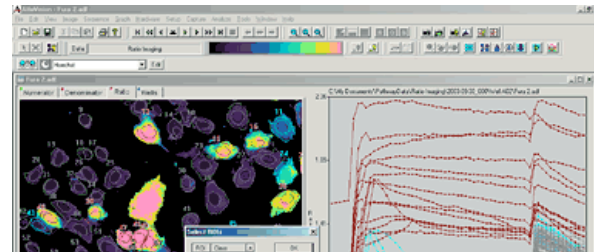
High-throughput Imaging



BD Pathway

Most image analysis software is still GUI-based (point & click)

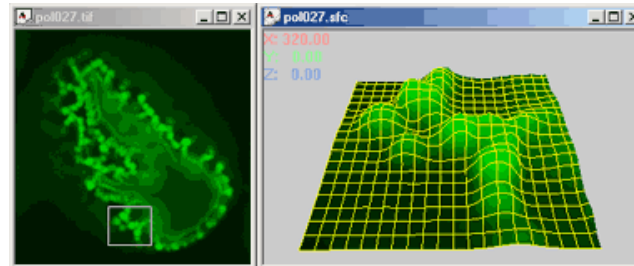
High-throughput Imaging



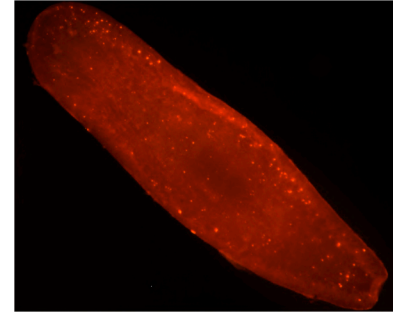
GUI-based software doesn't scale to large screens
RNAi experiment assaying 25,000 genes X 10 images each
= 250,000 images. That a lot of pointing and clicking.



BD Pathway

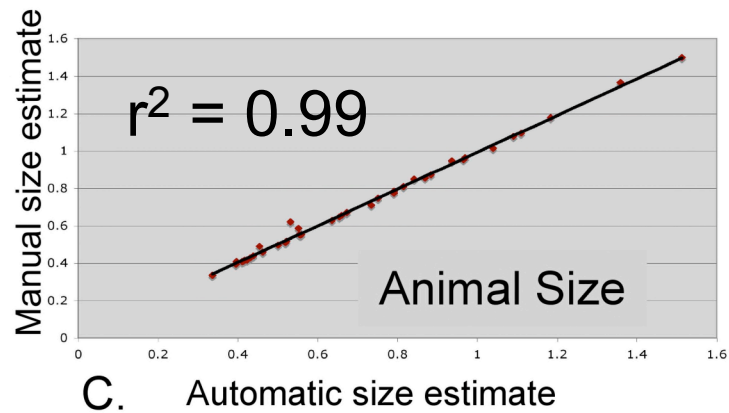
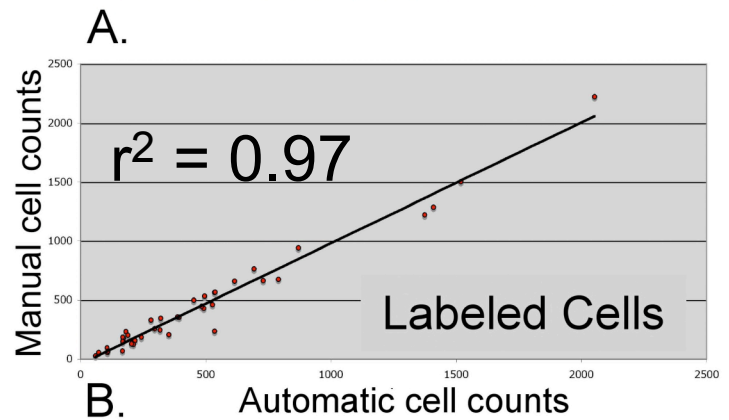
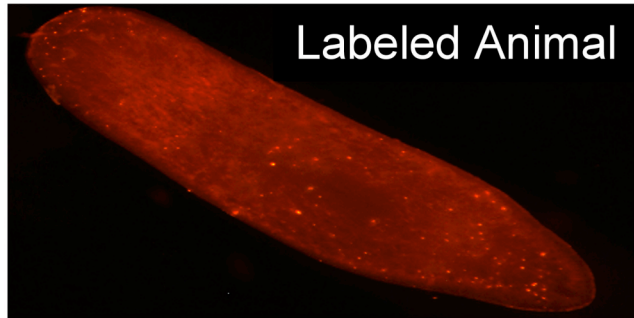


The screen will require:

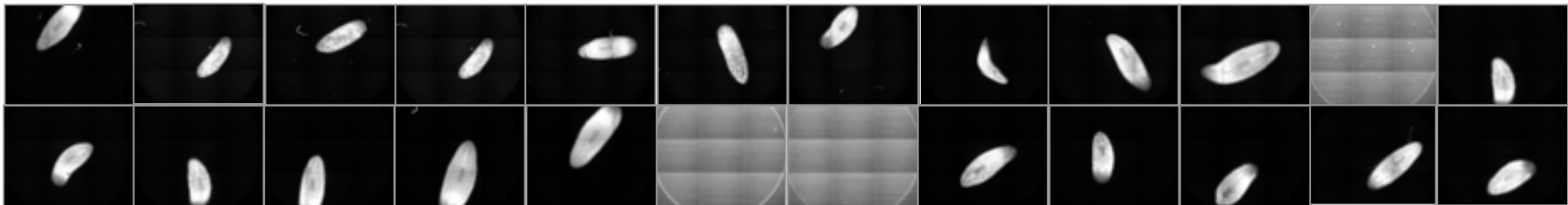
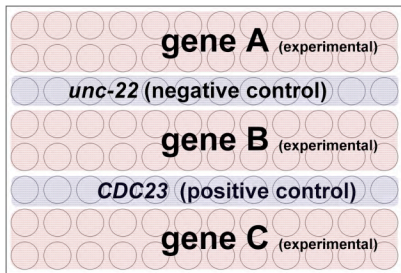


- Ability to identify neoblasts
- Ability to automatically determine the density of neoblasts
 - count labeled cells
 - determine animal size
- Ability to automatically identify RNAi-induced changes in the locations of neoblasts

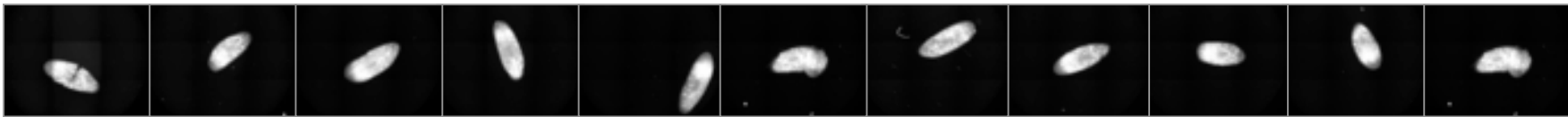
Task One: Accurate & automated detection of labeled cells and determination of animal size



Reality check: running the pipeline on real data

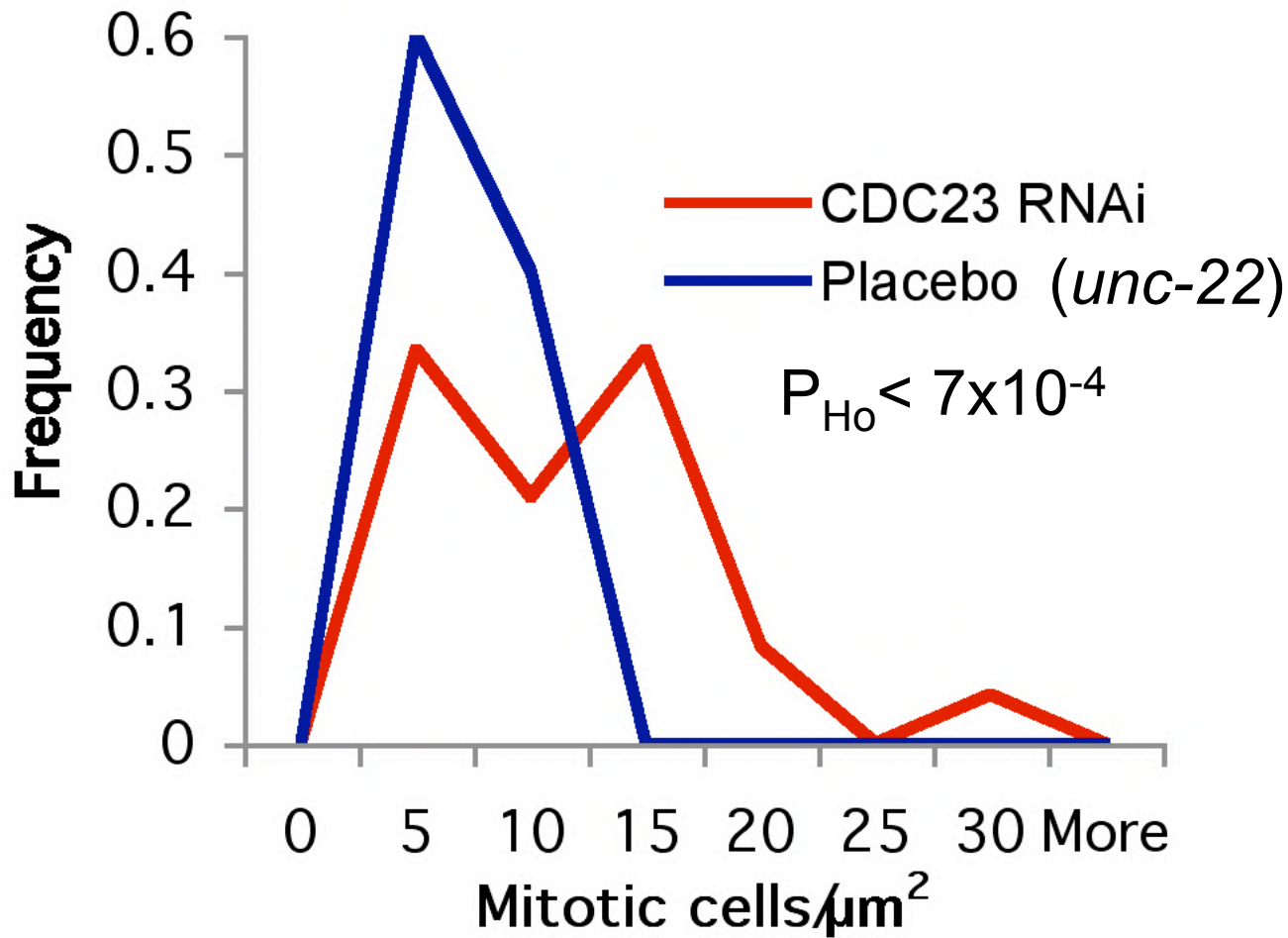


cdc23



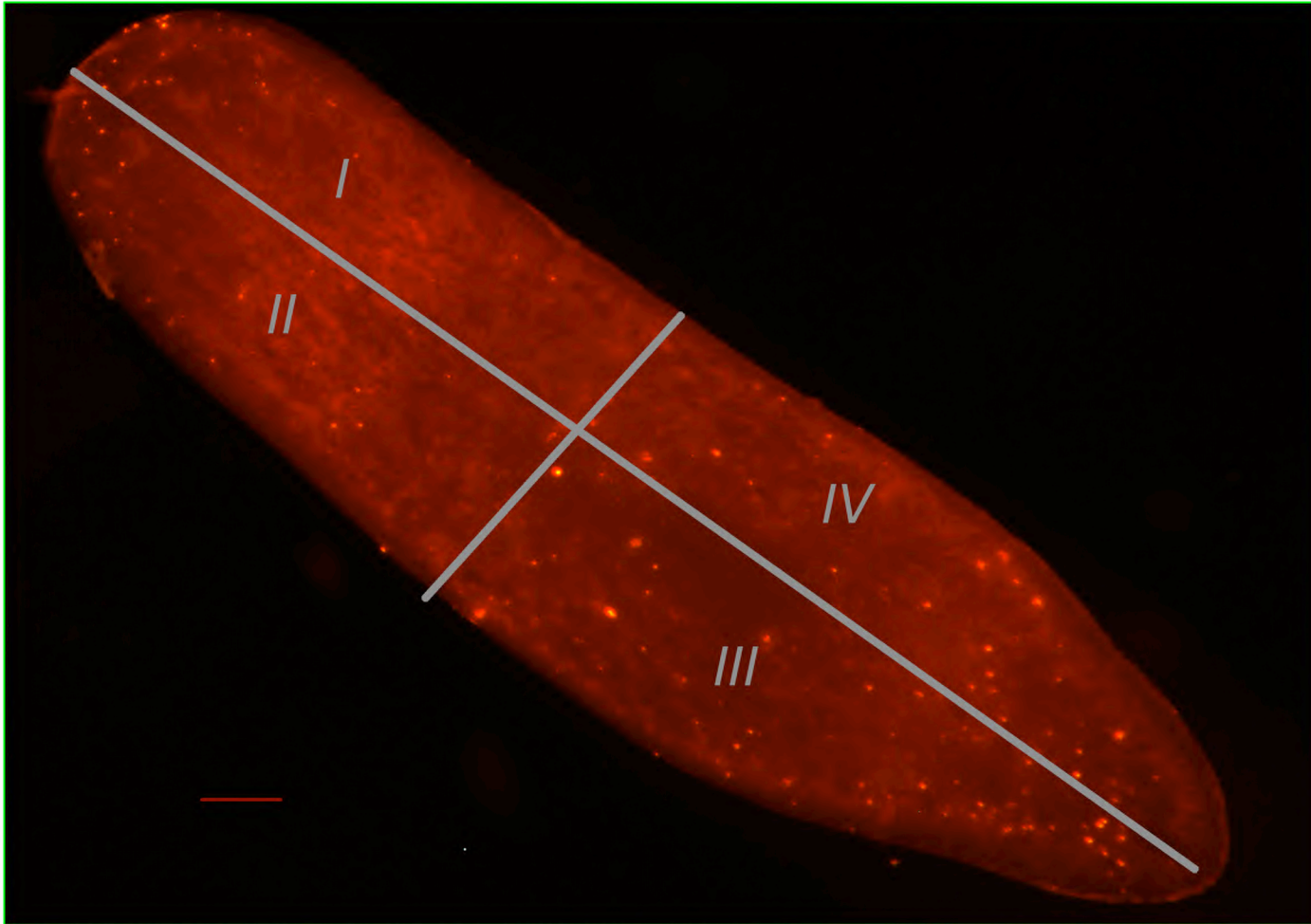
unc-22 (negative control)

Automatic detection of a 2-fold increase in the number of dividing cells induced by the **cdc23*** RNAi construct

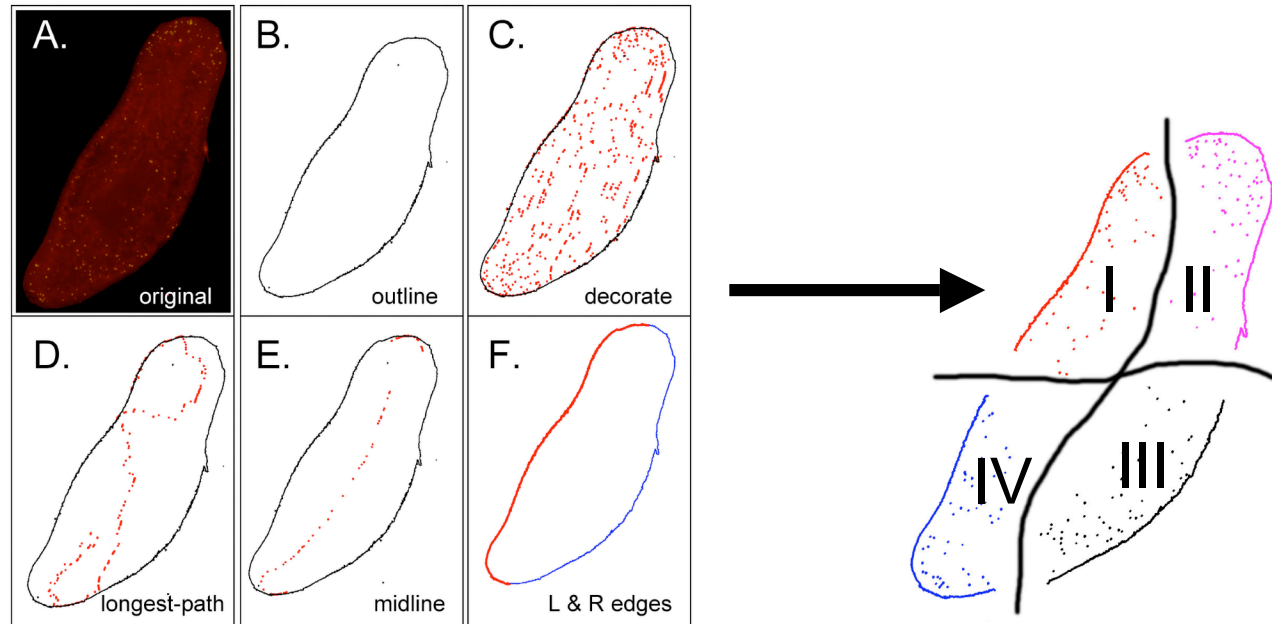


*Reddien, et al. Dev. Cell, 2005 May;8(5):635-49.

Task two: Automatically identify RNAi-induced changes in the locations of neoblasts



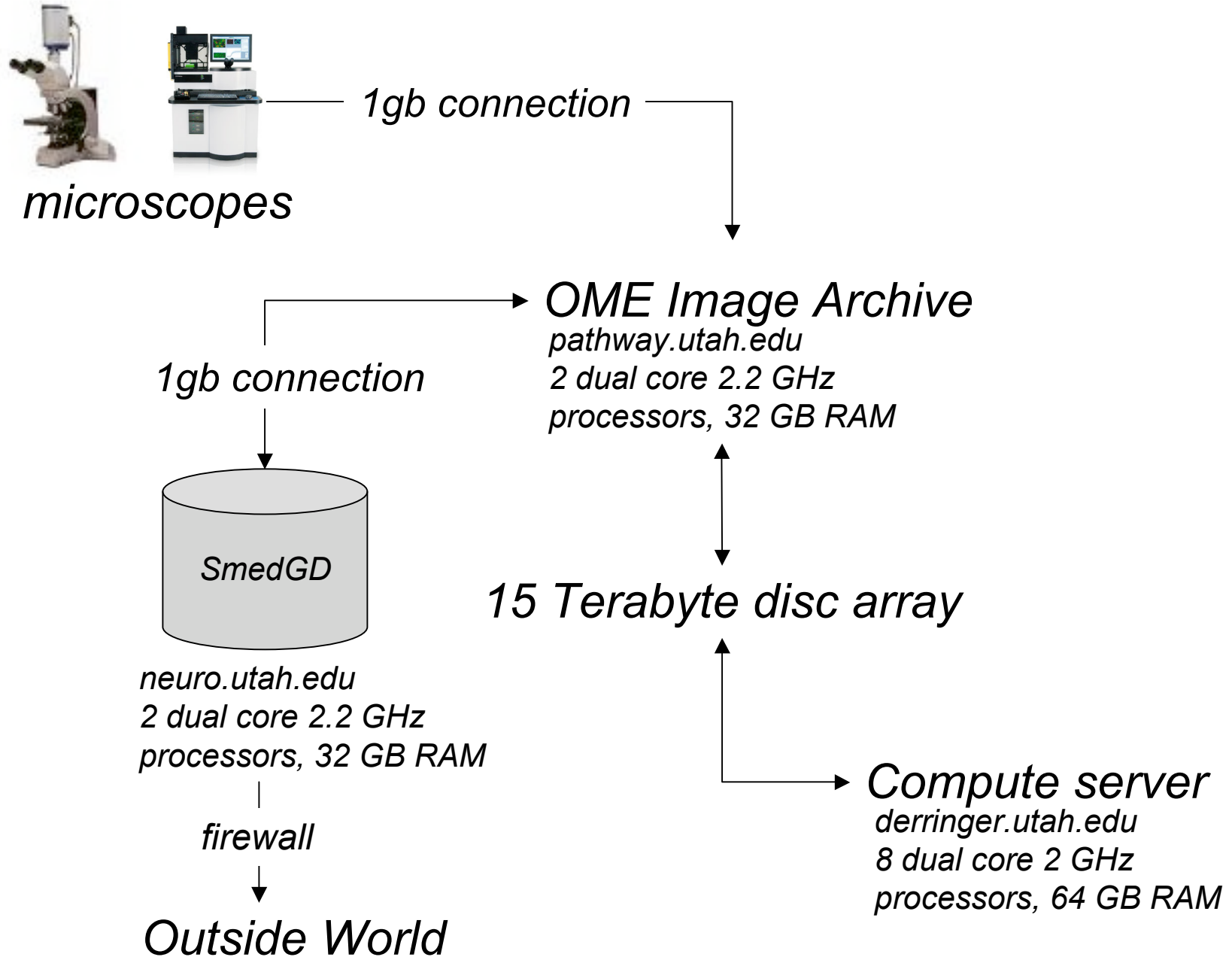
An algorithm[†] for identifying image midlines & L/R axes enables automated screens for asymmetric distributions of labeled* cells



* *mitotic, apoptotic, in situ, etc*

[†] Peng, et al. *Bioinformatics*, 2008 Jan 15;24(2):234-42.

..and of course you need a cyberinfrastructure



Conclusions

- **MAKER** greatly simplifies the genome annotation process with little tradeoff in accuracy
- **MAKER** is the first annotation tools designed to automatically *revise*, *merge*, *verify* and *re-evaluate* legacy annotation sets.
- Our collaboration with **the *S. mediterranea* genome project** provides a case study for the annotation and analysis of an emerging model organism genome.
- **Bioimage informatics** is a new area of bioinformatics facing problems of scale similar to those of genome annotation

Acknowledgements

Sánchez Alvarado Lab (UofU)

- Sofia Robb
- Joya Robb
- Eric Ross
- Jason Pellettieri
- Bret Pearson

Buell Lab (MSU)

- Robin Buell
- John Hamilton
- *P. ultimum* community

Suzanna Lewis (LBL)

Korf Lab (UCDGC)

- Genis Parra
- Keith Bradnam

Yandell Lab (UofU)

- Brandi Cantarel
- Carson Holt
- Barry Moore
- Hao Hu
- Deepak Anthony
- Hadi Islam

NHGRI, HHMI, NSF and USDA, and U of U School of Medicine