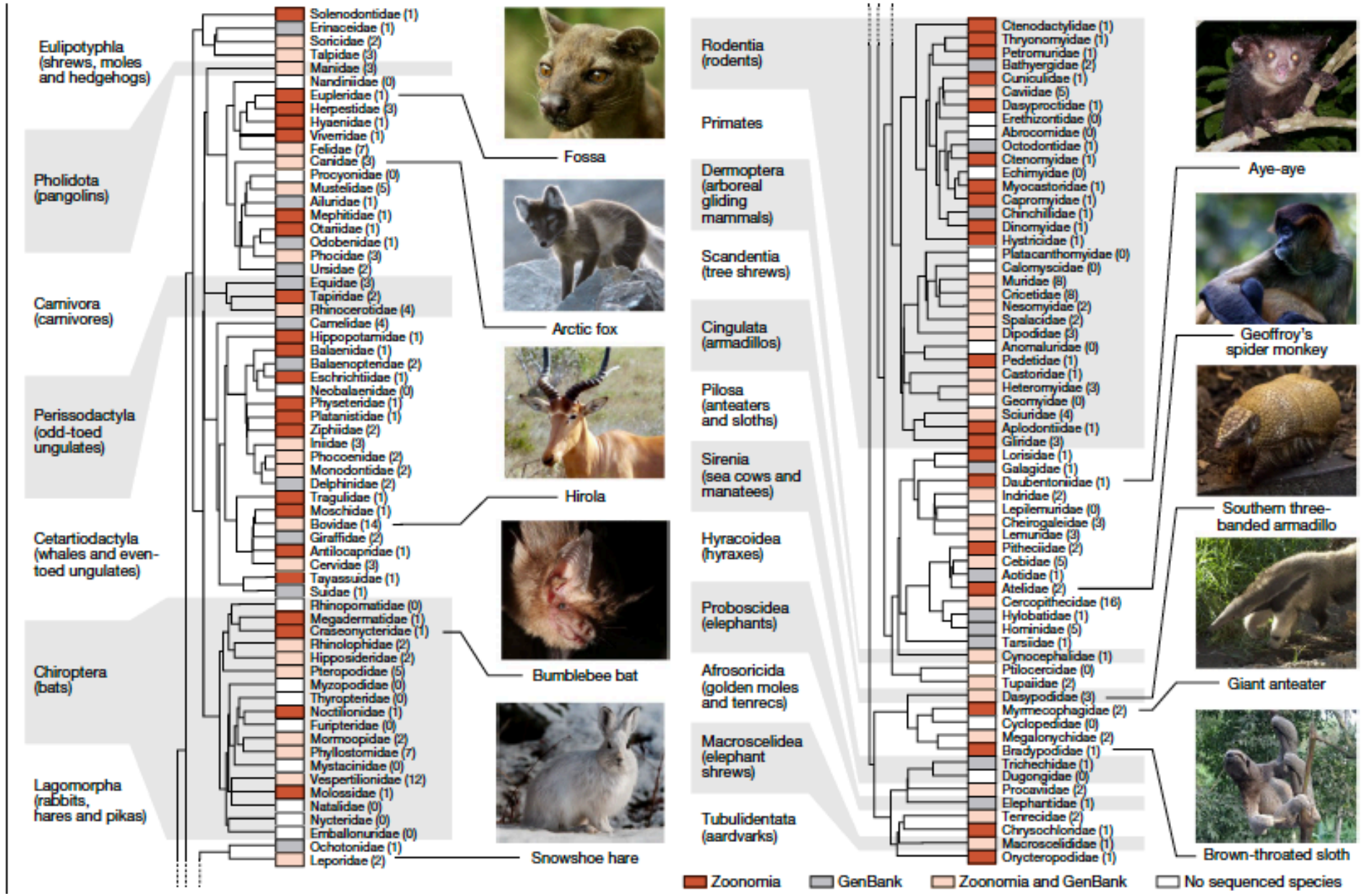
Genome Annotation II
Gene Prediction

“MODEL ORGANISMS”



Zoonomia Consortium. 2020. *Nature.*

Model Organisms versus New Genomes

Gene models

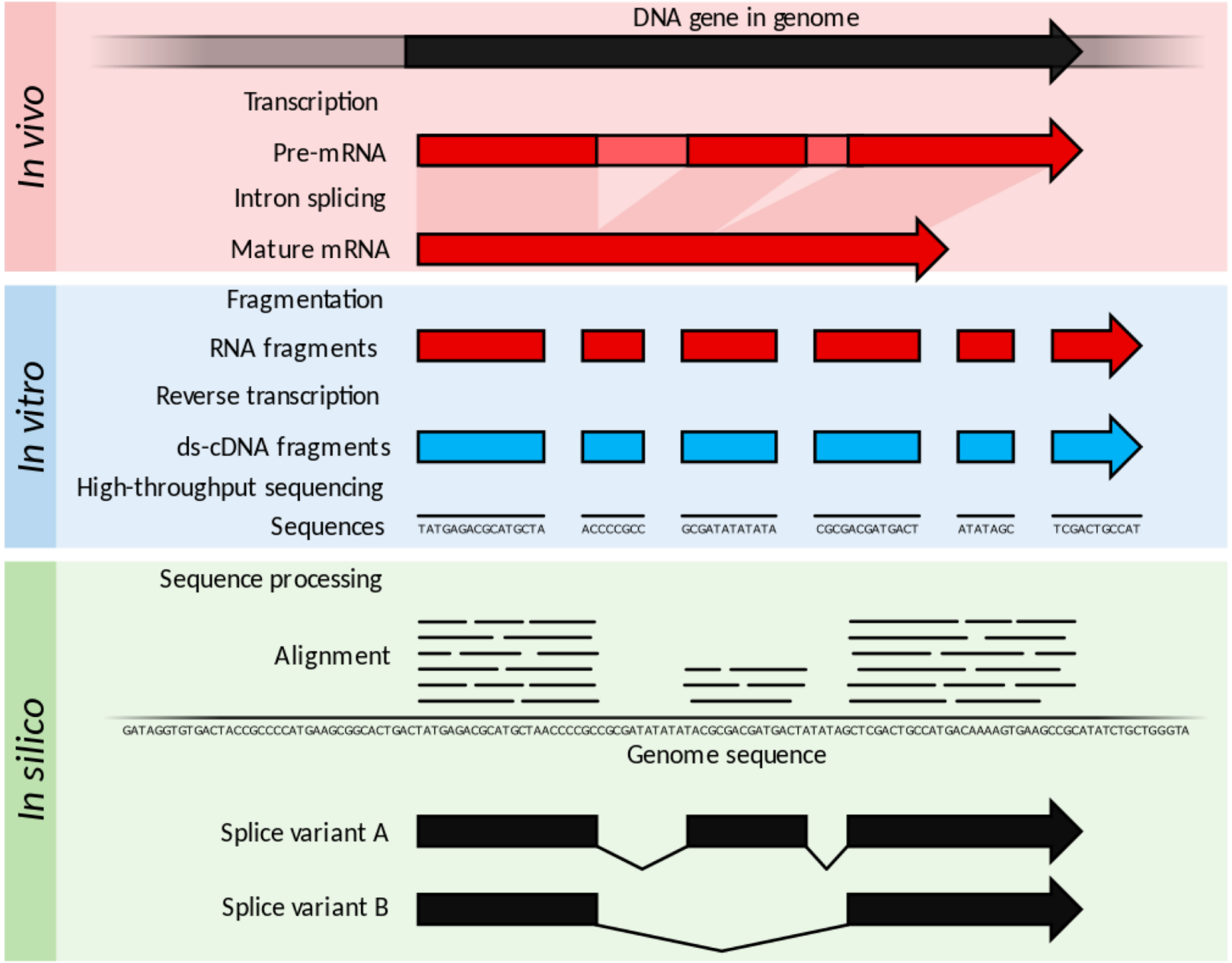
- First generation projects had large bodies of pre-existing knowledge
- *Drosophila, C. elegans, Homo sapiens*
- Allowed optimized gene prediction for these species

What about species without existing genomic resources or for whom no knowledge of protein sequences exist?

Gene Annotation Raw Materials

- *Ab-initio* gene prediction software
  - SNAP
  - Augustus
  - Requires training to your specific species
- Protein homology
  - NCBI, Ensembl, UniProt/SwissProt
- Expression data
  - RNA-Seq
  - This is the only direct evidence used

Very brief overview of RNA-Seq….



Shafee. 2017. Eukaryotic and prokaryotic gene structure. WikiJournal of Medicine.

Basic Properties of Gene Prediction Algorithms

Model must satisfy biological constraints

- Coding region must begin with a start codon: ATG
- Initial exon must occur before splice sites and introns
- Coding region must end with a stop codon: TAG, TGA, TAA

Use species-specific characteristics to improve accuracy

- Distribution of exon and intron lengths
- Base frequencies (GC content, codon bias)
- Protein sequences from the same or closely related species

Hidden Markov Models

HMM

- Type of machine learning algorithm
- Uncovers hidden labels from observed data
- During genome annotation, HMMs label individual nucleotides with a type.
  - introns
  - exons
  - splice sites

Transition probability

- The probability of switching from one nucleotide type to another

Emission probability

- The probability of observing a nucleotide od a certain type
- i.e., observing an adenine in an exon
- i.e., observing an adenine in a splice site

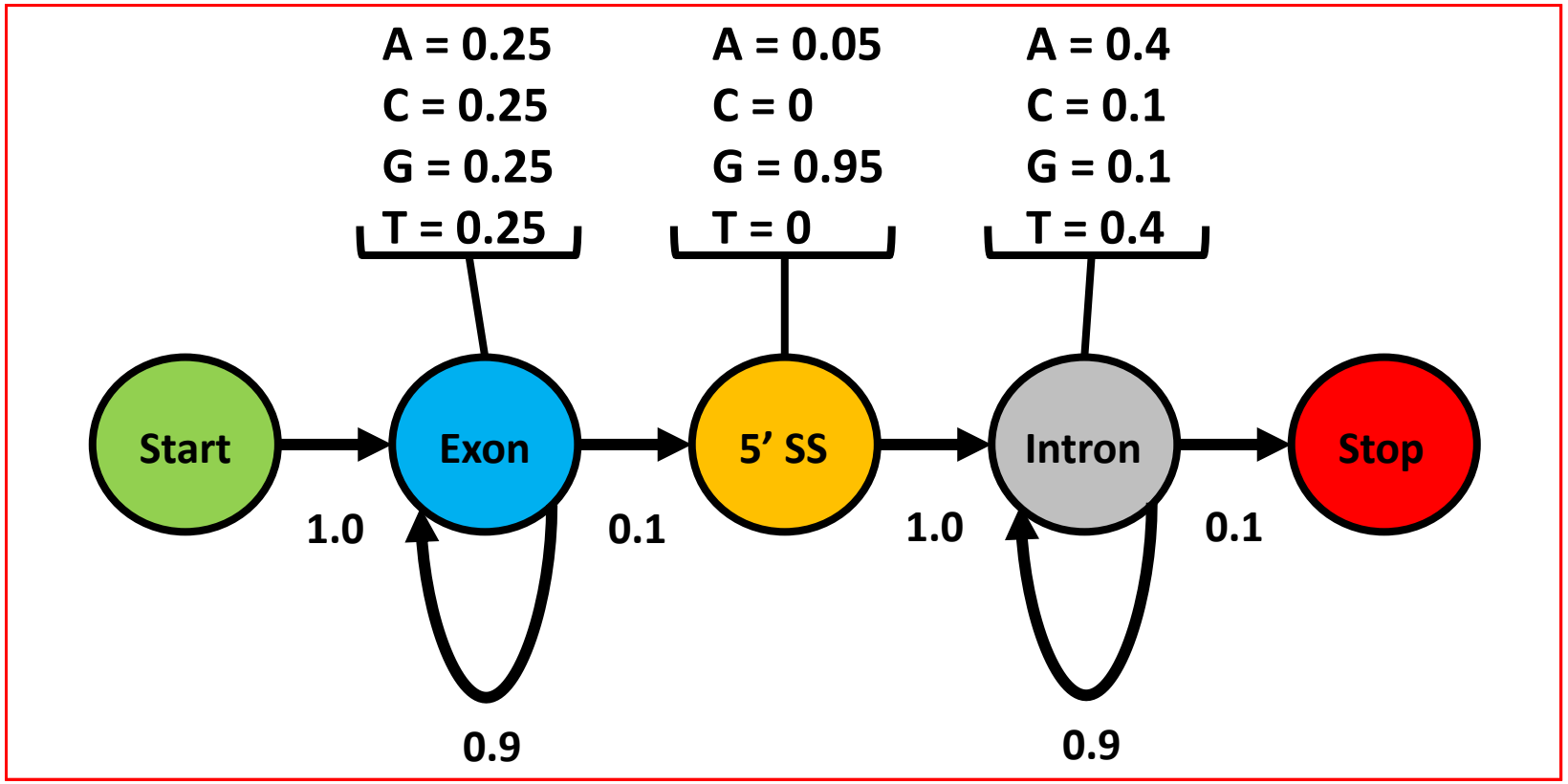
HMM Features

State Diagram

- indicates the different parts of the HMM.

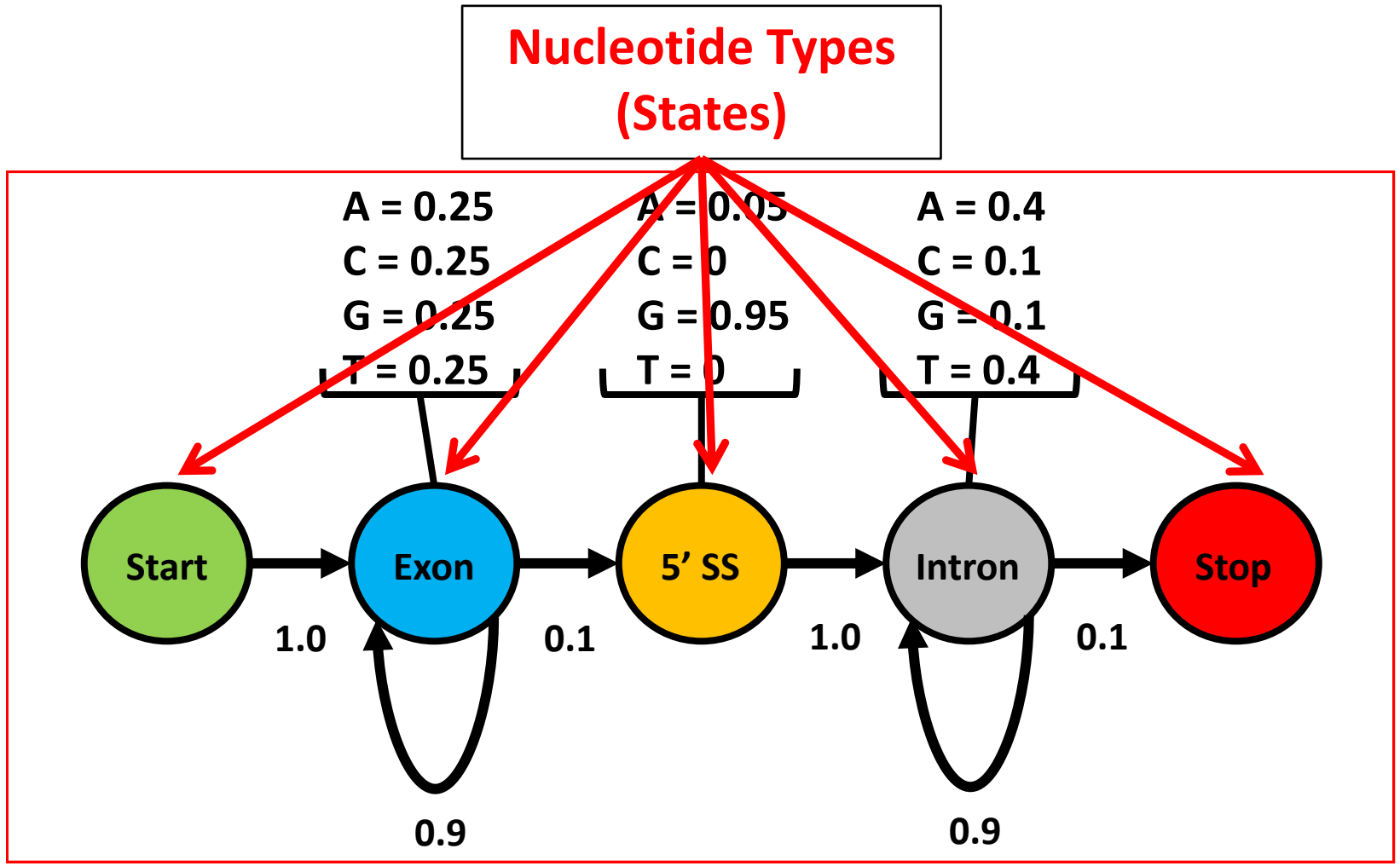
Single exon gene

State: label applied to each base



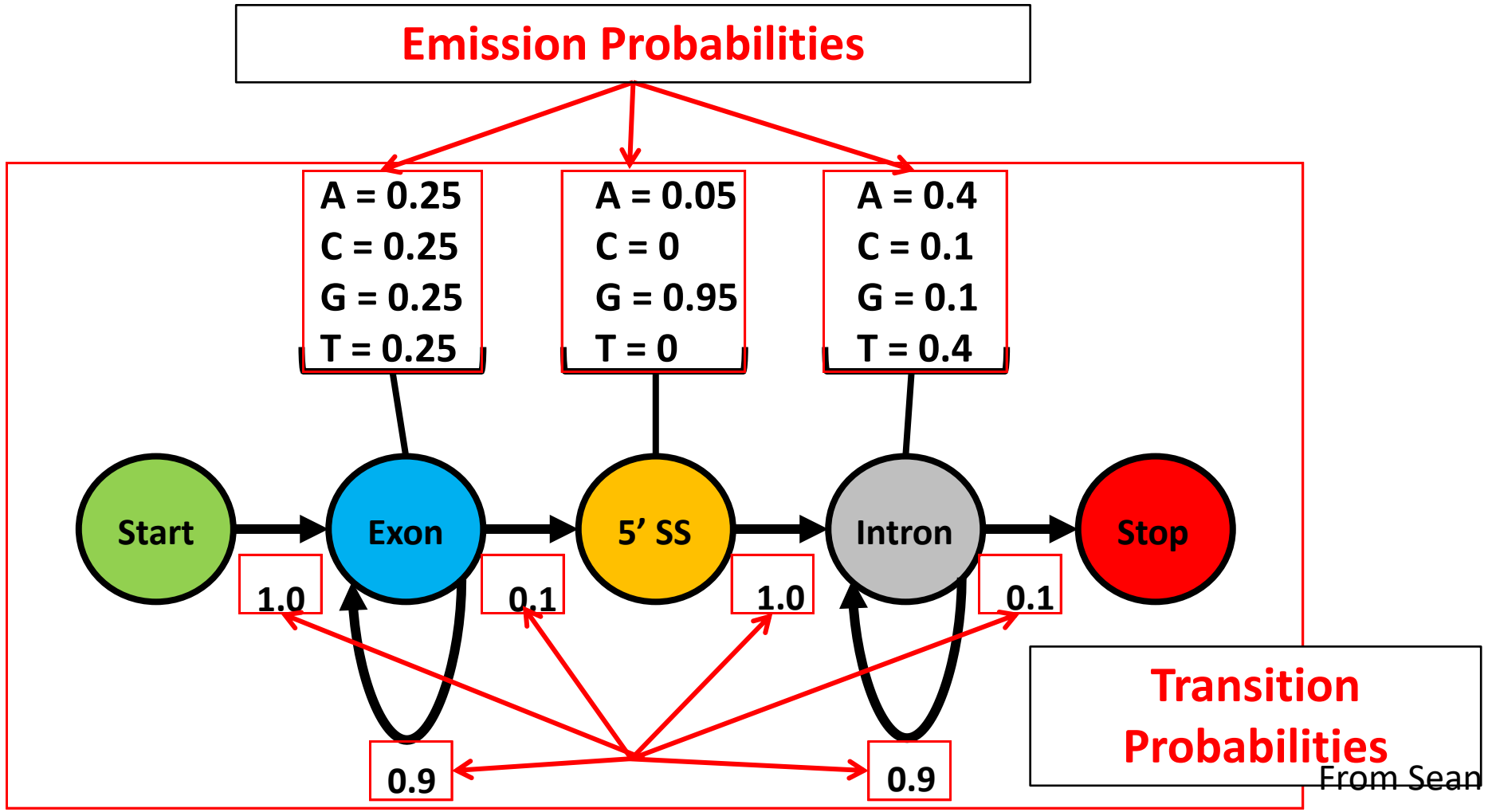
From Sean Eddy, HHMI

HMM Features



From Sean Eddy, HHMI

HMM Features



From Sean Eddy, HHMI

HMMs and Gene Prediction

Hidden Markov Models are the core of a number of gene prediction algorithms.

- GENSCAN
- Augustus
- SNAP
- GeneId
- Genemark
- GRAIL
- Twinscan

Weisstein et al. A Hands-on Introduction to Hidden Markov Models

MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects

Carson Holt and Mark Yandell. 2011. *BMC Bioinformatics*

What Does MAKER Do?

Identifies and masks out repeat elements

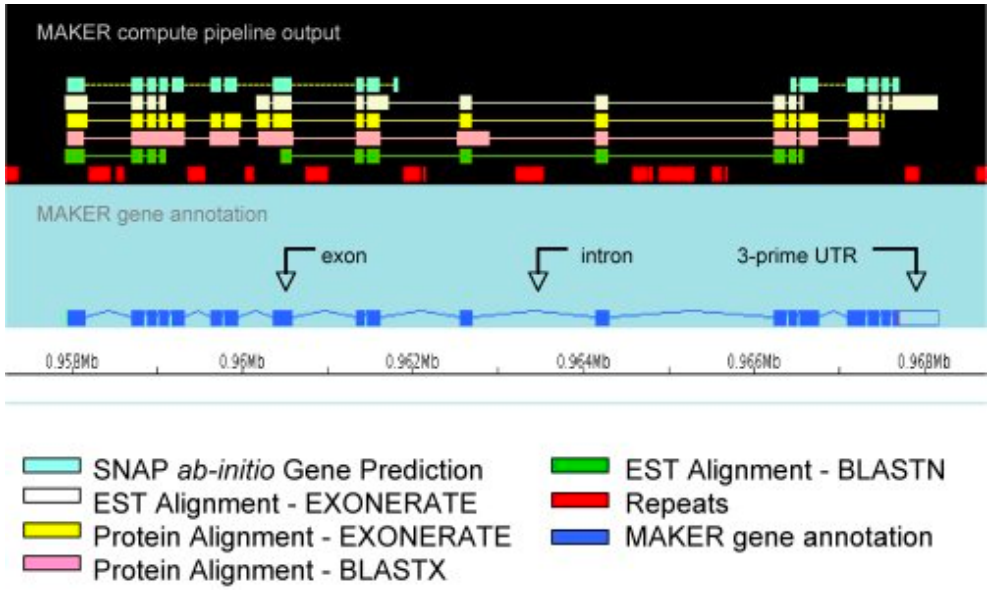
Aligns EST (RNA-Seq) to the genome

Aligns proteins to the genome

Produces *ab initio* gene predictions

Synthesizes these data into final annotations

Produces evidence-based quality values for downstream annotation management



http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/MAKER_Tutorial_for_WGS_Assembly_and_Annotation_Winter_School_2018

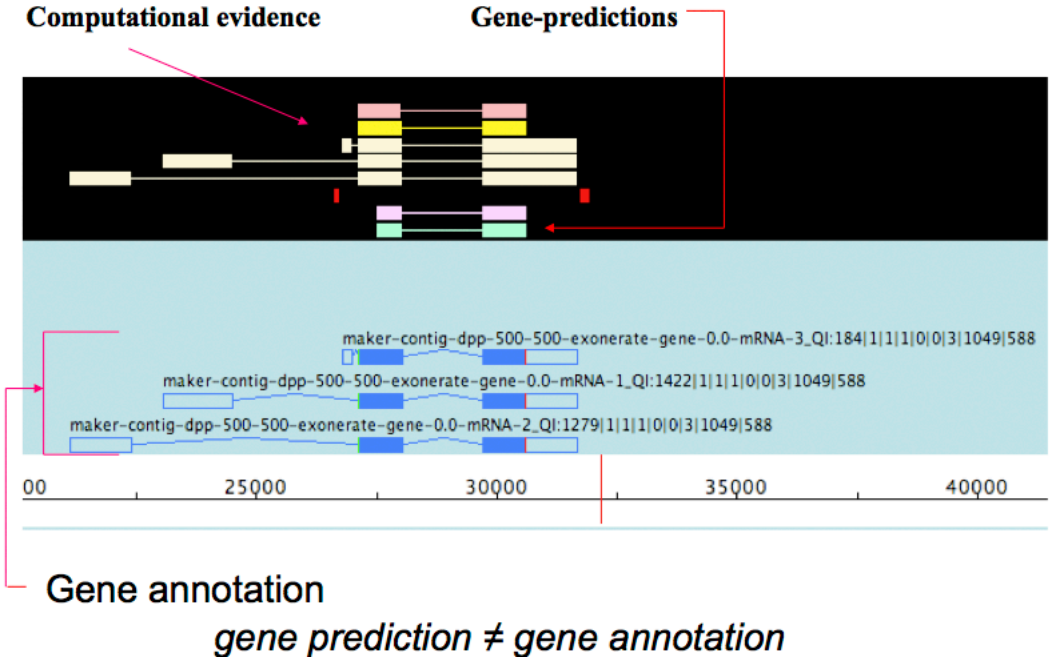
What Sets MAKER Apart from other Tools

MAKER is an annotation pipeline

- not a gene predictor.

MAKER leverages existing software

- including gene predictors
- integrates their output to produce MAKER gene models
- these are the 'best' possible models for a given location based on the evidence alignments.



gene prediction ≠ gene annotation

Gene predictions ***are*** gene models.

Gene annotations are gene models but include documented evidence supporting the model.

http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/MAKER_Tutorial_for_WGS_Assembly_and_Annotation_Winter_School_2018

Annotation Editing Distance

Sensitivity

- the number of overlapping nucleotides between the prediction and the experimental evidence divided by the total base pair count of experimental evidence

Specificity

- the number of overlapping nucleotides between the prediction and the experimental evidence divided by the number of nucleotides in the prediction

Accuracy (congruency)

- the average of sensitivity and specificity: C=(Sens + Spec)/2
- ***incongruency: D = 1-C***
- 0 means complete agreement
- 1 means no support

Next

I am looking forward to your proposals

- but I’ve received few questions
- please reach out to me before you hand in your proposal! :-)

MAKER helps to "train" *ab initio* gene models

Table 1 Gene model accuracy for gene prediction/annotation programs

Reference Organism	Performance Category	*Ab Initio* Predictions			MAKER Annotations		
		Augustus	GeneMark	SNAP	Augustus	GeneMark	SNAP
A. thaliana	Nucleotide Accuracy	77.04%	74.68%	69.78%	80.53%	79.39%	80.27%
	Exon Accuracy	67.03%	61.31%	56.40%	67.81%	69.60%	68.78%
D. melanogaster	Nucleotide Accuracy	76.08%	66.54%	69.29%	76.42%	73.66%	74.33%
	Exon Accuracy	61.37%	47.31%	47.01%	58.56%	58.03%	58.49%
C. elegans	Nucleotide Accuracy	88.29%	88.09%	85.10%	87.14%	86.29%	88.48%
	Exon Accuracy	74.62%	68.88%	61.38%	68.60%	65.03%	66.19%

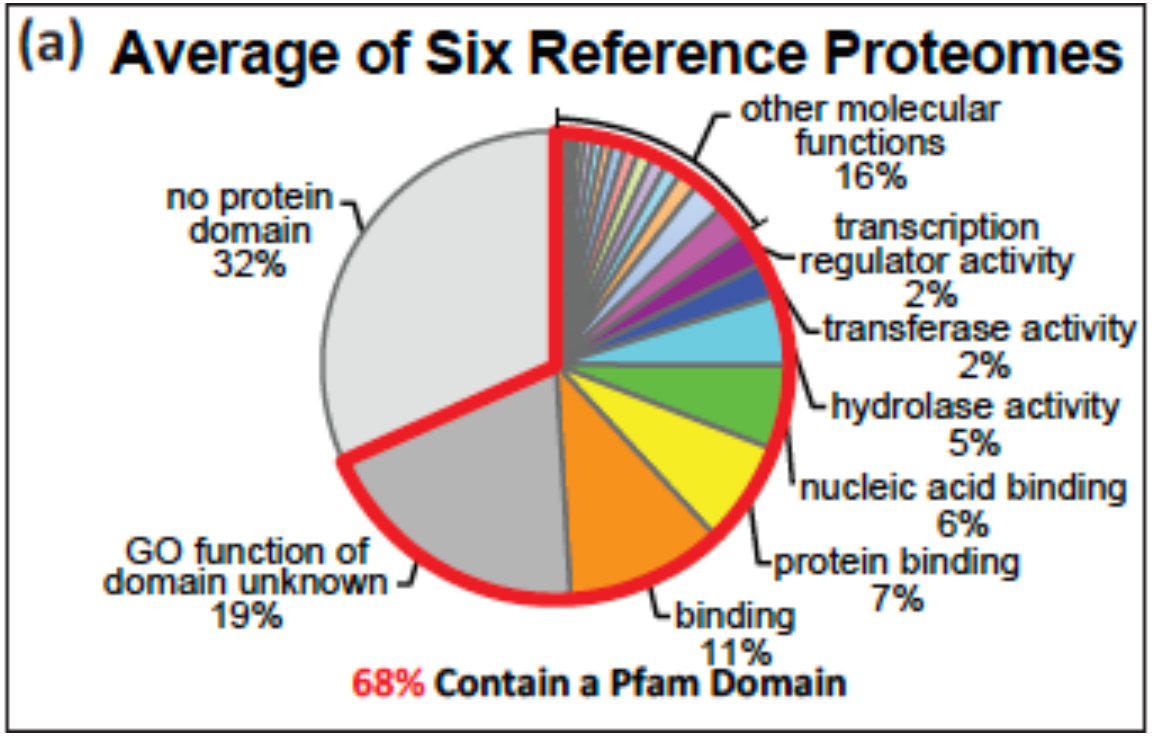
improved accuracy for gene models when ran as part of the MAKER pipeline.

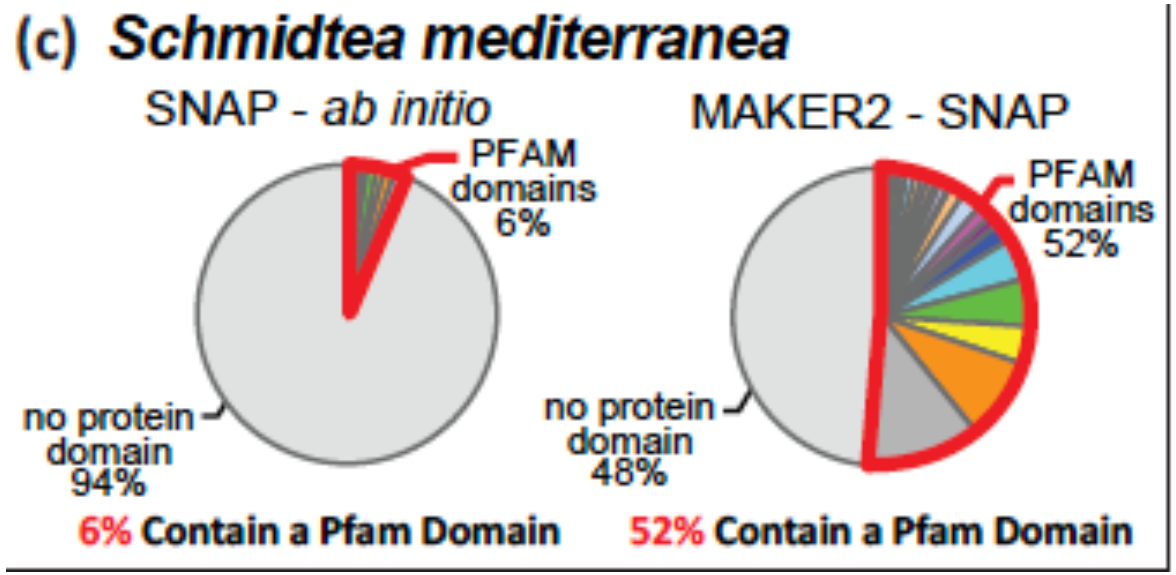
MAKER is robust – even to bad *ab initio* models

Table 2 Gene model accuracy using unmatched species parameters

Reference Organism	Performance Category	*Ab Initio* Predictions			MAKER Annotations		
		Augustus	GeneMark	SNAP	Augustus	GeneMark	SNAP
A. thaliana	Nucleotide Accuracy	57.85%	48.62%	43.84%	68.56%	57.96%	73.77%
	Exon Accuracy	30.71%	16.51%	18.58%	53.31%	28.87%	60.11%
D. melanogaster	Nucleotide Accuracy	67.47%	66.51%	48.92%	73.78%	72.83%	74.44%
	Exon Accuracy	30.62%	26.25%	19.94%	43.10%	39.74%	53.69%
C. elegans	Nucleotide Accuracy	66.18%	67.26%	68.24%	74.32%	71.92%	85.02%
	Exon Accuracy	28.33%	30.01%	35.44%	38.52%	39.42%	63.14%

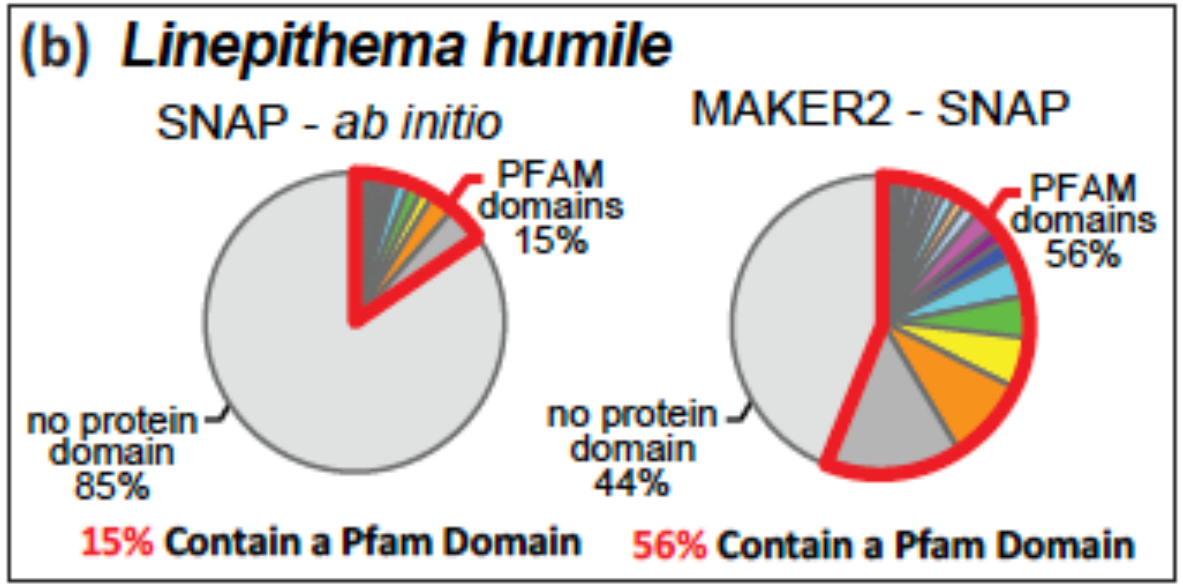
improved accuracy even when the gene models for the wrong species are used.





MAKER outperforms ab initio gene predictors.

- Gene models with evidence trails
- Ability to retrain gene predictors for improved performance.
- Far more Pfam domains found in MAKER2-predicted genes than those just predicted using SNAP



Annotation Editing Distance

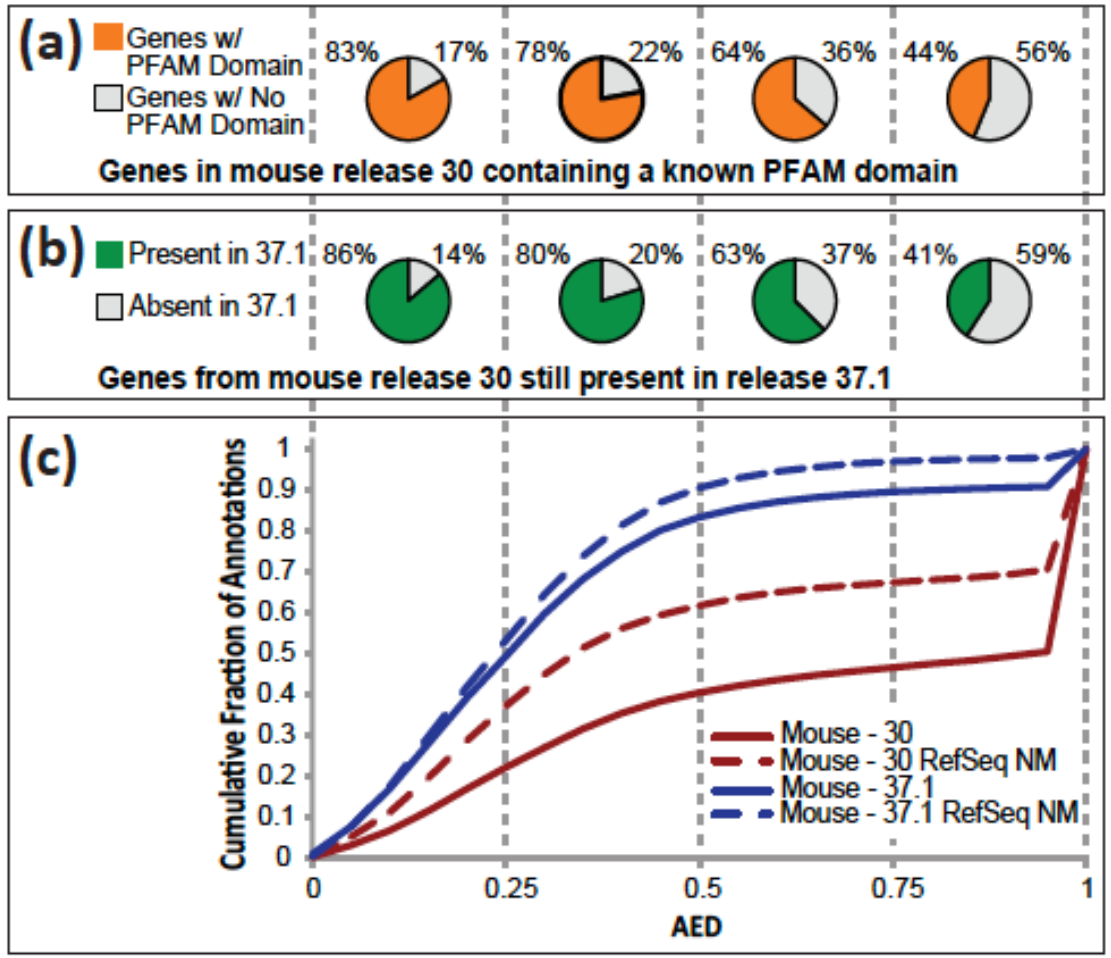
Genes with Pfam Domains

- lower AED -> more Pfam domains

AED predicts better annotations

- Genes in the updated mouse assembly have lower AED

AED quantifies improvements made between these mouse genome releases.



Annotation Editing Distance

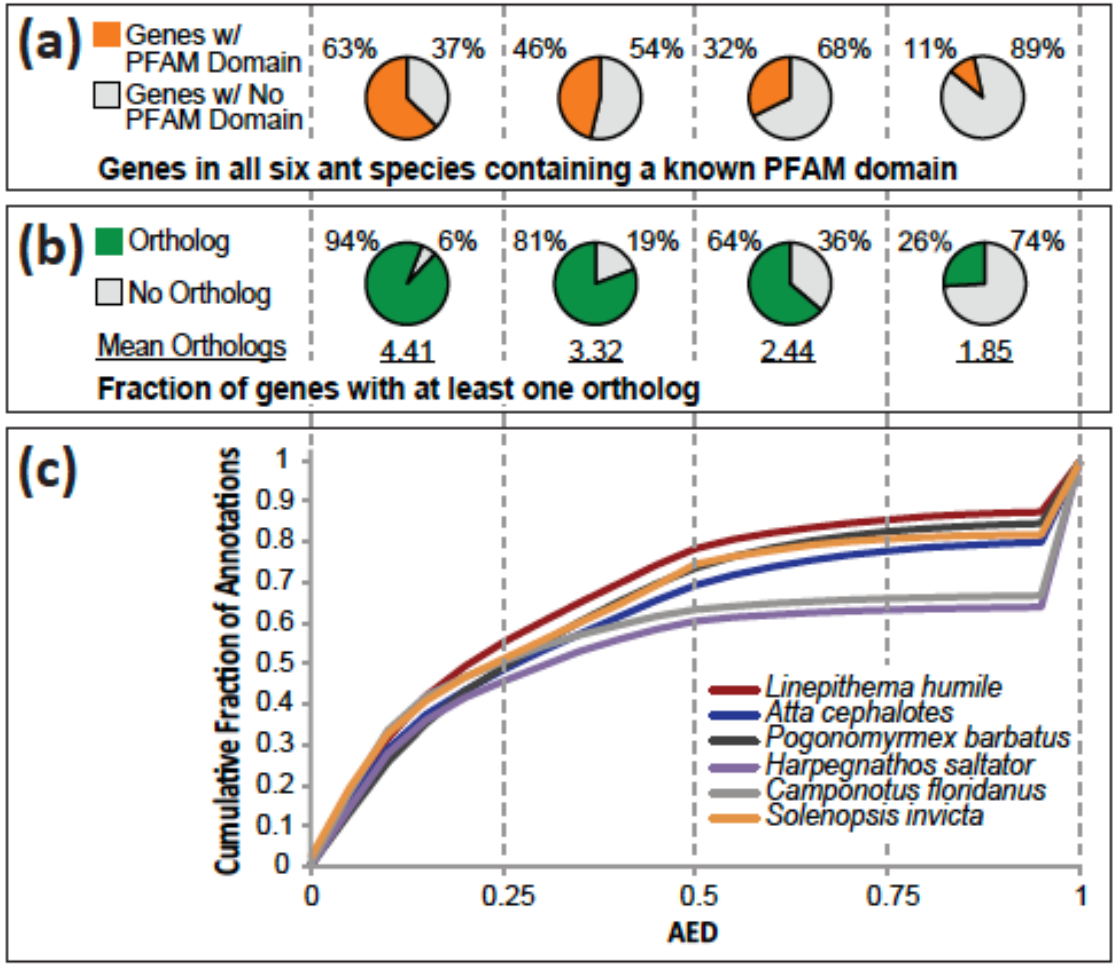
Genes with Pfam Domains

- lower AED -> more Pfam domains

AED predicts better orthology

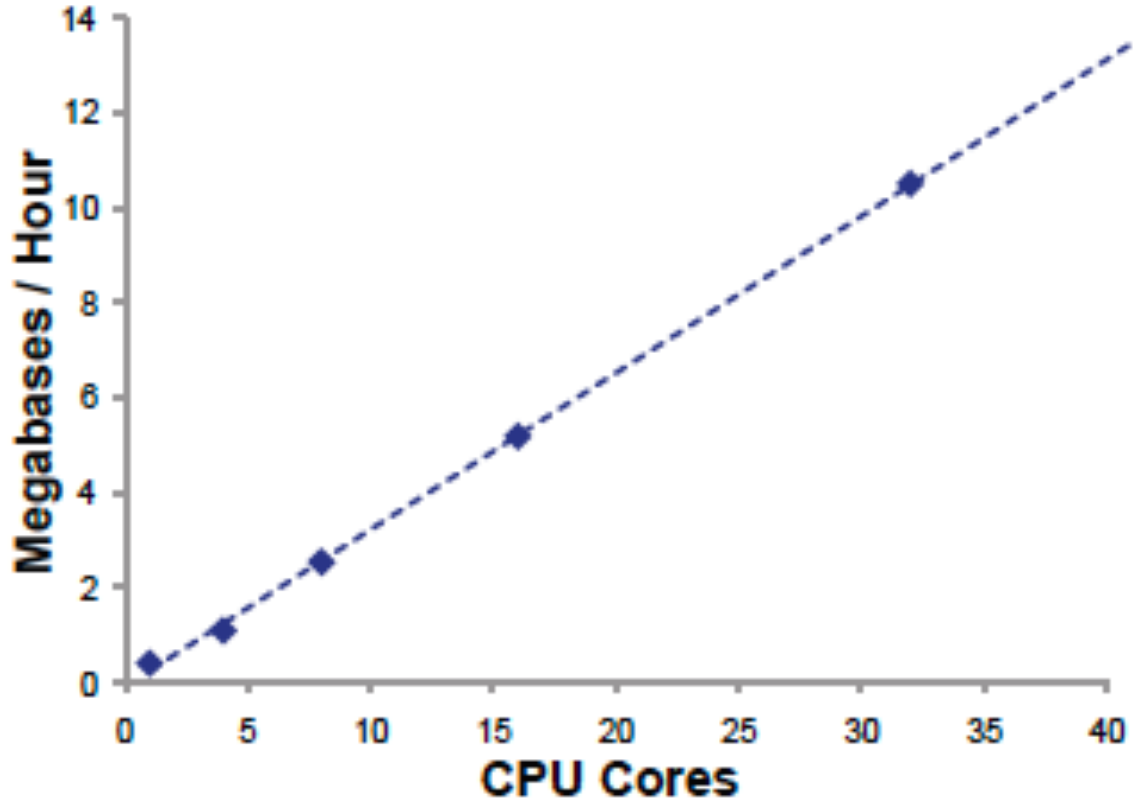
- Genes with lower AED were more likely to have an ortholog in another ant assembly.

AED allows the detection of potential false positives in non-model organism gene annotations.



Conclusion

If you have a computing cluster, use MAKER2 to annotate your genome.



But keep in mind the computational requirements.