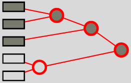
Gene Prediction with AUGUSTUS

Genome annotation: challenges in eukaryotes and consequences for evolutionary genomics, 13 February 2018

Ingo Bulla
Institut für Mathematik und Informatik
Universität Greifswald



About the speaker

- PhD in mathematics about a non-applied topic, switched to bioinformatics in 2006
- Main research topic: Sequence analysis, phylogeny, evolution, epidemiology and public health of HIV
- Now working with Mario Stanke (developer of AUGUSTUS) on improving the algorithm used by AUGUSTUS
- Limited experience in genomics, has only applied AUGUSTUS once in a research project
  → Speaker will have a Skype with
  - Mario Stanke or
  - Katharina Hoff (long-time user of AUGUSTUS, implementer of BRAKER)

  during the lunch talk if questions come up he cannot answer
- Ingénieur de recherche in Perpignan from 1st of April on, in a wetlab group (Christoph Grunau, Guillaume Mitta)

1 Overview on Gene Prediction

2 with RNA-Seq

RGASP Assessment

BRAKER1

3 homology-based

Structural Genome Annotation Problem

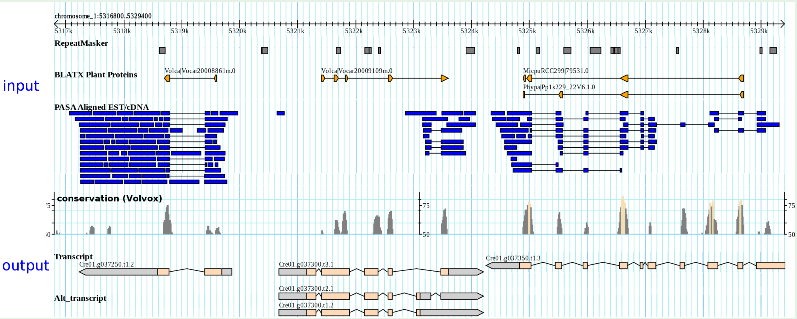
Input

- genome assemblie(s)
- extrinsic evidence, e.g. from RNA-Seq, MS/MS, protein database

Output

- start- and end positions of genes, CDS, exons and introns (`.gff`)

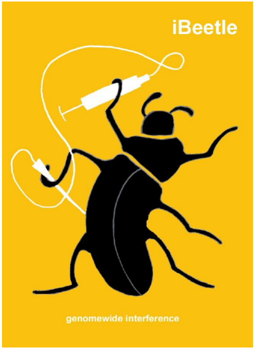
Example (12 600 bp from algae *Chlamydomonas reinhardtii*, with JGI)

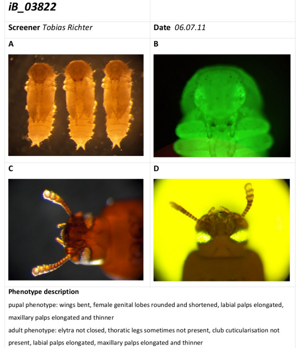


Example Application

iBeetle: RNAi screen for the beetle *Tribolium castaneum*

1. predict genes
2. design primers based on prediction
3. produce dsRNA for each gene
4. knock down each gene in larval and pupal stage
5. observe phenotype
6. study function for select genes





iB_03822

Screener *Tobias Richter*	**Date** *06.07.11*

Phenotype description

pupal phenotype: wings bent, female genital lobes rounded and shortened, labial palps elongated, maxillary palps elongated and thinner

adult phenotype: elytra not closed, thoracic legs sometimes not present, club cuticularisation not present, labial palps elongated, maxillary palps elongated and thinner

Major Approaches to Protein-Coding Gene Prediction

approach	extrinsic evidence used	programs
ab initio	-	GeneMark, Augustus, Snap, Fgenesh
transcript-based	transcript seqs, e.g. RNA-Seq	BRAKER, Exonerate Augustus, mGene
protein homology	protein sequences	Augustus-Ppx, GeneWise, Exonerate
comparative (*de novo*)	additional (unannotated) genomes	Augustus, Contrast, N-Scan
proteogenomics	peptides from mass spectrometry	Augustus
combiners/ selectors	other gene predictions + transcript seqs + proteins + ?	Jigsaw, Glean, Maker2, Pasa

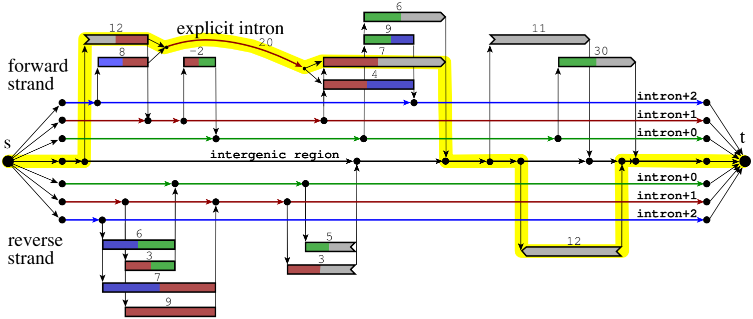
State of the art usually requires a combination of approaches:

Use for every part of a gene all evidence available for that gene or region.

Single species gene-finding: 1-species graph

Assumptions: no alternative splicing, no gene overlap

- graph represents all candidate gene structures
  - nodes: exon candidates (EC)
  - edges: introns and intergenic regions
- each path from *s* to *t* is one gene structure
- single species gene-finding in linear time: longest path algorithm



Gene finder AUGUSTUS

- developed since 2002 (PI: Mario Stanke)
- based on conditional random field (generalization of HMM)
- probabilistic model of gene structures given signals, CDS, evidence
- get most likely genes structure or a sample of likely ones

Some genome annotation collobarations using AUGUSTUS

Aedes aegypti	yellow fewer mosquito: dengue fever	*Science*, 2007
Brugia malayi	parasitic worm, causes elephantiasis	*Science*, 2007
Tribolium castaneum	red flour beetle, pest and model organism	*Nature*, 2008
Schistosoma mansoni	parasite causing bilharziosis	*Nature*, 2009
Coprinus cinereus	fungus	*PNAS*, 2010
Nasonia vitripennis	wasp	*Science*, 2010
Amphimedon queenslandica	sponge	*Nature*, 2010
Culex pipiens	common mosquito	*Science*, 2010
Ricinus communis	castor bean	*Nature Biotechnology*, 2010
Chlamydomonas reinhardtii	green algae	*Proteomics*, 2011
Galdieria sulphuraria	red algae	*Science*, 2013
Arabidopsis thaliana	plant model organism	*PNAS*, 2008
Heliconius melpomene	butterfly	*Nature*, 2012
Apis mellifera	honey bee	*BMC Genomics*, 2014

1 Overview on Gene Prediction

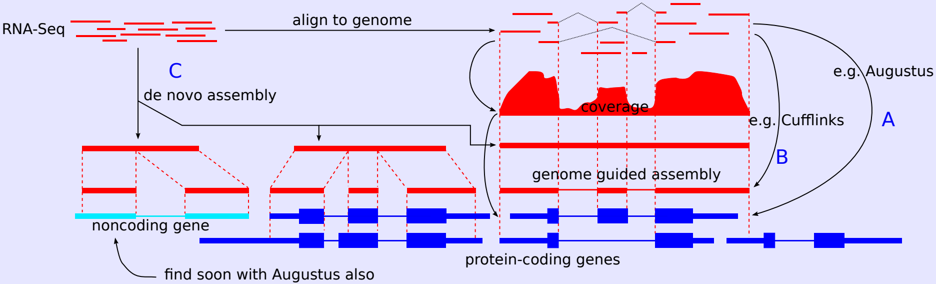
2 with RNA-Seq

RGASP Assessment

BRAKER1

3 homology-based

Three Major Approaches to Gene-Finding with RNA-Seq



A evidence integration into gene finder

(e.g. AUGUSTUS, FGENESH, MGENE, GENEID)

1. align reads to genome first
2. integrate evidence from coverage and spliced alignments into gene finder

B purely alignment-based (e.g. Cufflinks)

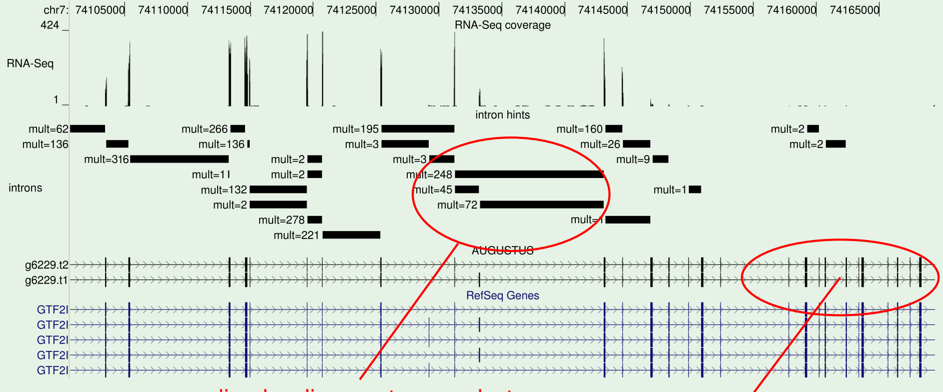
1. align reads to genome first
2. construct transcripts from spliced alignments (no gene finding)

C de novo assembly of reads (e.g. Trinitry, TransDecoder, Velvet + AUGUSTUS)

1. assemble transcriptome reads into transcript contigs
2. use contigs for gene finding or just align them

AUGUSTUS using RNA-Seq

Using RNA-Seq only (on human)



spliced alignments used to predict alternative splicing

ab initio model dominates where little or no evidence

RGASP: RNA-Seq Genome Annotation Assessment Project

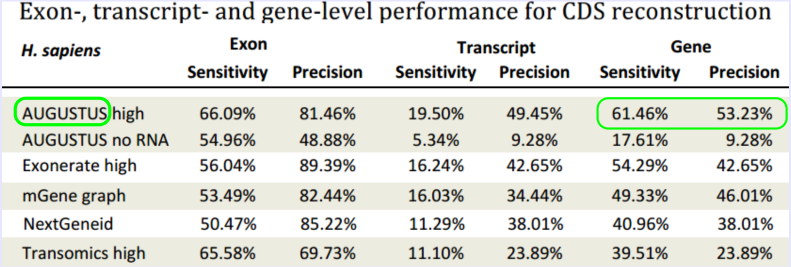
Assessment of transcript reconstruction methods for RNA-seq

Steijger et al., *Nature Methods*, Nov. 2013

- assessed the progress of automatic gene building using RNAseq
- part of ENCODE project
- 17 participating groups submitted, all on same data

Excerpt of RGASP assessment results on human

Calling transcripts and proteins:



Exon-, transcript- and gene-level performance for CDS reconstruction

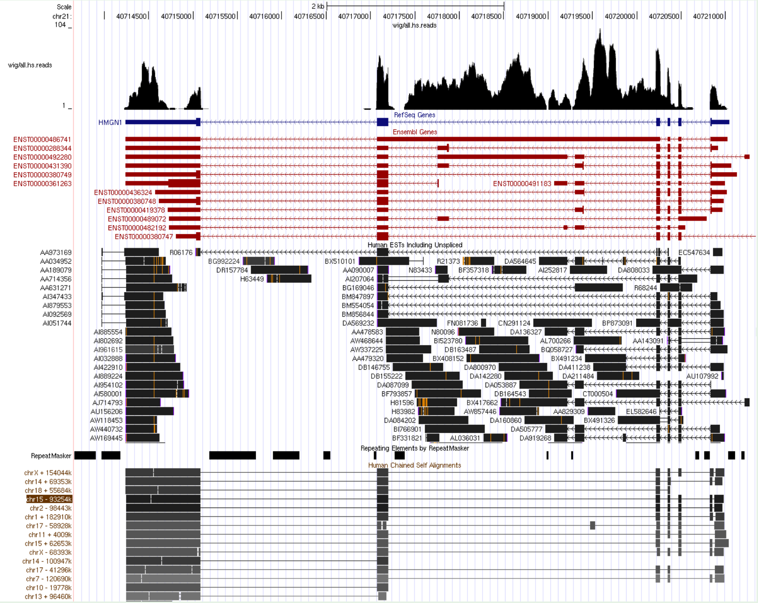
H. sapiens	Exon		Transcript		Gene	
	Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision
AUGUSTUS high	66.09%	81.46%	19.50%	49.45%	61.46%	53.23%
AUGUSTUS no RNA	54.96%	48.88%	5.34%	9.28%	17.61%	9.28%
Exonerate high	56.04%	89.39%	16.24%	42.65%	54.29%	42.65%
mGene graph	53.49%	82.44%	16.03%	34.44%	49.33%	46.01%
NextGeneid	50.47%	85.22%	11.29%	38.01%	40.96%	38.01%
Transomics high	65.58%	69.73%	11.10%	23.89%	39.51%	23.89%

Best results on

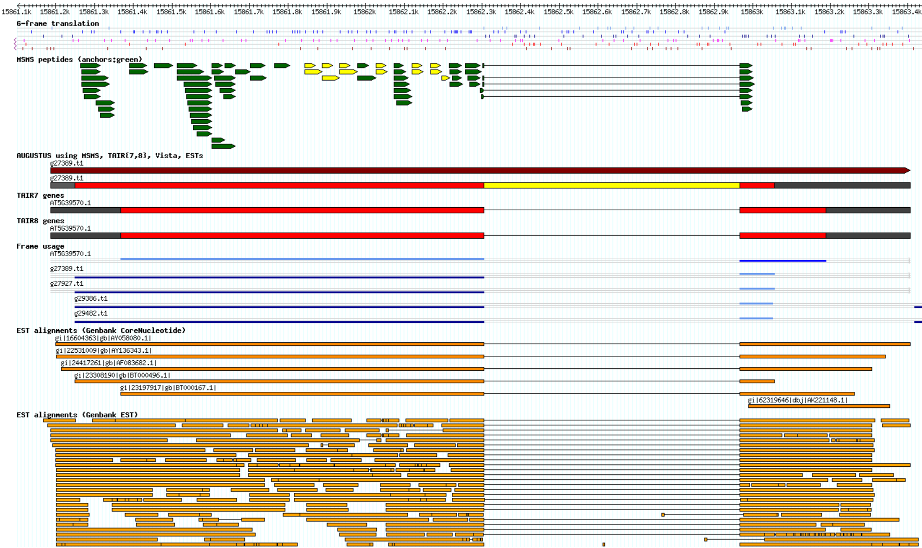
	transcript sensitivity	gene sensitivity	
fly	24%	49%	(AUGUSTUS)
worm	48%	61%	(TRANSOMICS)

Why was the accuracy not better?

Problems: intronic transcription, self-similarity of genome



Reminder: RNA-Seq does not give you the protein sequence



BRAKER1

Collaboration with former competitor

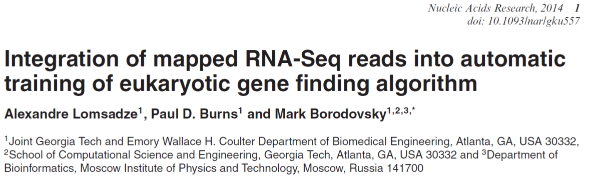
- MAKER2 pipeline uses GENEMARK and AUGUSTUS
- Why not throw together
  - GENEMARK-ET that self-trains on RNA-Seq and
  - AUGUSTUS that predicts with RNA-Seq

  ourselves?
- easy to use:

  ```
  braker.pl [OPTIONS]
  -genome=genome.fa -bam=rnaseq.bam
  ```
- fast (1 day for fly on 1 CPU)

Mark Borodovsky (GENEMARK)

GeneMark-ET (2014): unsupervised training of parameters



Nucleic Acids Research, 2014 1
doi: 10.1093/nar/gku557

Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm

Alexandre Lomsadze[1], Paul D. Burns[1] and Mark Borodovsky[1,2,3,*]

[1]Joint Georgia Tech and Emory Wallace H. Coulter Department of Biomedical Engineering, Atlanta, GA, USA 30332, [2]School of Computational Science and Engineering, Georgia Tech, Atlanta, GA, USA 30332 and [3]Department of Bioinformatics, Moscow Institute of Physics and Technology, Moscow, Russia 141700

GeneMark does not use RNA-Seq for prediction.

Anchors from RNA-Seq for training

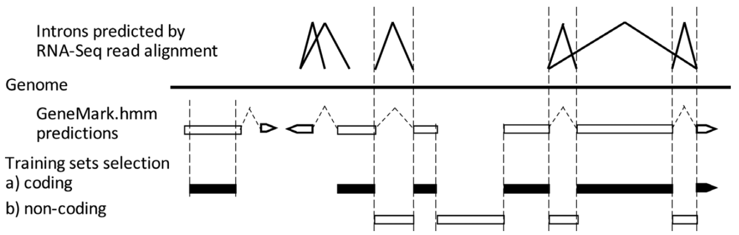
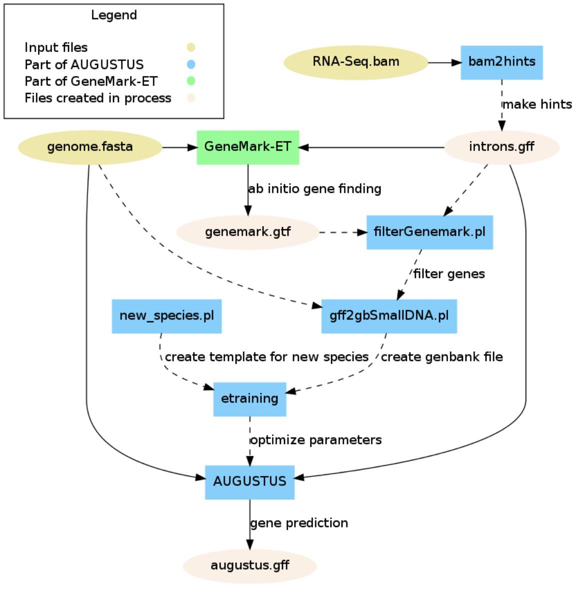


Figure 3. Selection of elements of training set in GeneMark-ET for the next iteration. The new training set of protein-coding regions is comprised from exons with at least one 'anchored splice site' as well as long exons predicted *ab initio* (>800 nt).

BRAKER1 Pipeline



Legend

Input files ●
Part of AUGUSTUS ●
Part of GeneMark-ET ●
Files created in process ●

RNA-Seq.bam

bam2hints

make hints

genome.fasta

GeneMark-ET

introns.gff

ab initio gene finding

genemark.gtf

filterGenemark.pl

filter genes

new_species.pl

gff2gbSmallDNA.pl

create template for new species

create genbank file

etraining

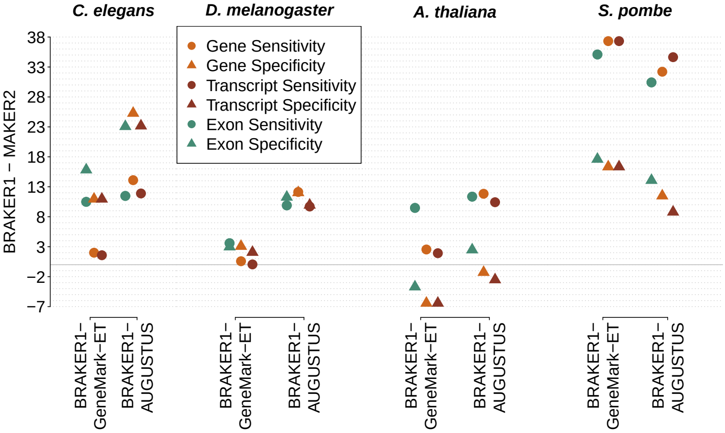
optimize parameters

AUGUSTUS

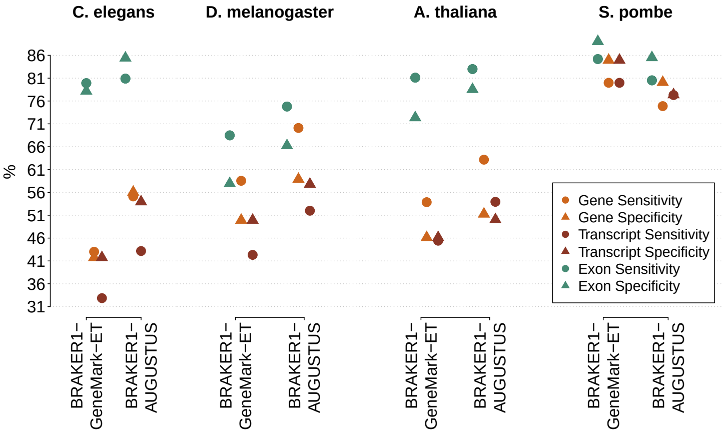
gene prediction

augustus.gff

Comparing BRAKER1 to MAKER2 (using RNA-Seq only)



Accuracy of BRAKER1



1 Overview on Gene Prediction

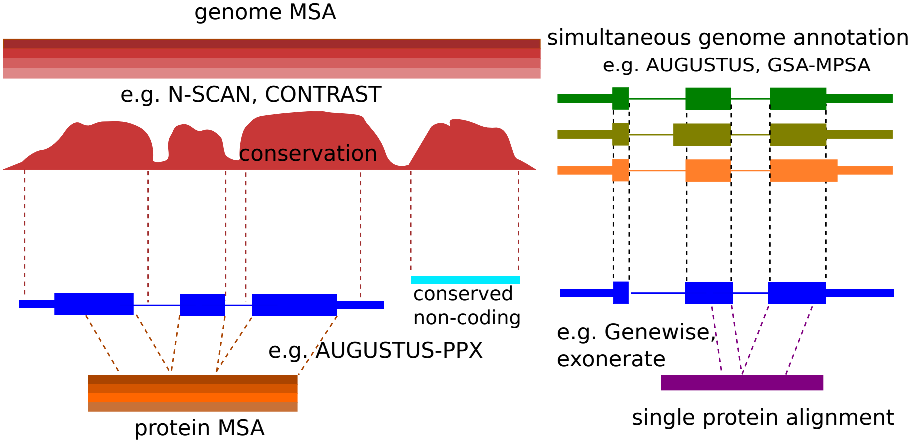
2 with RNA-Seq

RGASP Assessment

BRAKER1

3 homology-based

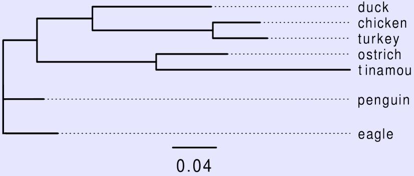
Homology-Based Gene-Finding Approaches



Example application for comparative gene prediction

$k = 47$ bird species

MSA of genomes (genome sizes ≈1Gb each)



```
scaffold702     954964 51 -  1264172 AGCAATTATCCGAGCAAATCCTTGGCTT
chr9           1515518 51 + 25554352 AGCAATTATCTGAGAAATTTCTTGGCTT
11            21279039 51 - 24221871 AGCAATTATCTGAGAAAATTCTTGGCTT
scaffold182    2077047 52 -  2532513 AGCAATTATCTGAGTAAGTTCTTGGCTT
scaffold362     124565 30 -   180957 AGCAATGACCCGAGCAGGCTCTTGAGCA
...
Scaffold679     885067 51 -  2350160 AGCAATTATCTGAGCAAGTTCGTGGCTA
...
scaffold17530    12417 51 +    51700 AGCAATTATCTGAGCAAGTTCTTGGCTA
```

Comparative gene prediction problem

Find all genes in all genomes,

optionally using existing annotations or evidence for some genomes.

Other potential target clades

- i5k insect clades (e.g. beetles, spiders, bees)
- vertebrate clades from the genome 10K project
- bacterial pan-genomes
- a polyploid genome (e.g. wheat, *Verticillium longisporum*)

Homology

Conservation of gene structure

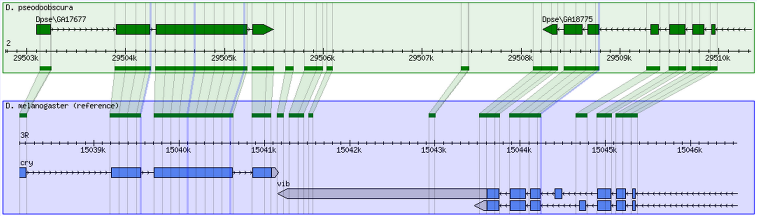
some Lamin gene structures from fish, mosquito, sponge, flea, beetle

```
T. rubripes  -----|--------|-----|--------|--|-----|--|--|-----|-----|-----------|--
T. rubripes  -----|--------|-----|--------|--|-----|--|--|-----|--------|-----------
T. rubripes  --|--|--------|-----|--|-----|--|-----|--|--|-----|--------------|-----
T. rubripes  -----|--------|-----|--------|--|-----|--|--|-----|-----------|--------
A. aegypty   -----|--------------------|--------------|--------|--------|-----------
A. queensl.  -----|--|-----------|--------|--|-----|--------------------------------
D. pulex     -----|-----|-----|-----------|-----|--|--|-----|--|--------|-----------
T. castaneum -----------------|-----------------------------------|-----------------

             -- exon (any length)
             |  intron (aligned)
```

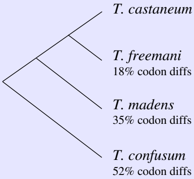
*(*example by Martin Kollmar)

Complementary to RNA-Seq: Genome Comparisons



Gbrowse_syn display of syntenic regions from *D. mel.* and *D. pseudoobscura* (50% codon diffs)

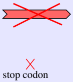
How can synteny help annotation?


T. castaneum

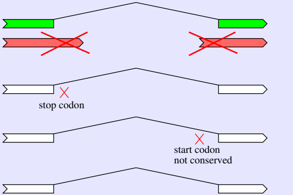
T. freemani
18% codon diffs

T. madens
35% codon diffs

T. confusum
52% codon diffs



reading frame disruption
in close relative helps
remove false positive genes/exons



two red genes not conserved
but all splice sites of intron conserved
correct split gene