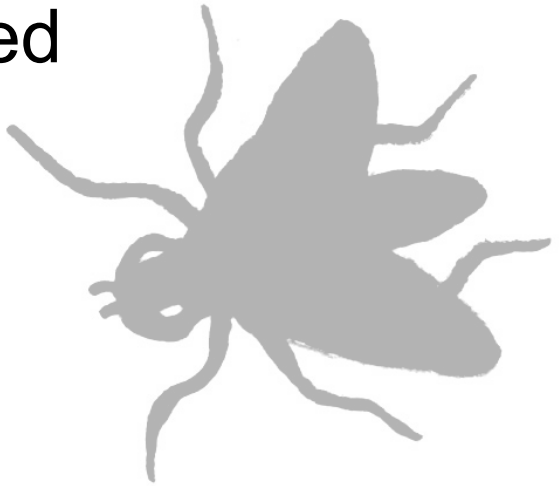
# Towards a Comprehensive Annotation of Structured RNAs in Drosophila

Rebecca Kirsch

31st TBI Winterseminar, Bled
20/02/2016



RTH
CENTER FOR NON-CODING RNA IN TECHNOLOGY AND HEALTH

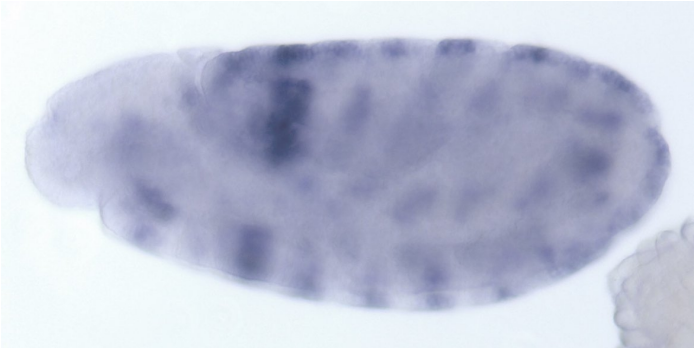
DET SUNDHEDSVIDENSKABELIGE FAKULTET
KØBENHAVNS UNIVERSITET

# Studying Non-Coding RNAs in Drosophila

**Why Drosophila?**

especially for novel molecules and processes:

- short generation time
- easy to maintain
- easy to modify and study genetically
- esp. RNAi is well-established

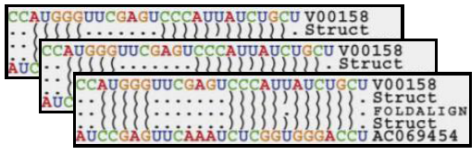


*in-situ* hybridization, embryo

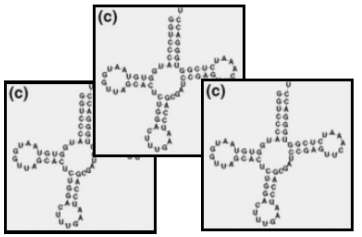
→ **ideal model organism for large-scale screens**

# Aim – Comprehensive Annotation of Structured RNAs in Drosophila

**RNA Structure Prediction**




**Annotation of Structured RNAs**




# Previous ncRNA Screens in Drosophilids

**RNAz[1] screen, relaxed**
(Rose, 2007)

**Input alignment filtering:**
> 50 nt alignment length,
< 25 % gap characters,
GC content 25 - 75 %

42,482 predictions

**RNAz[1] screen, stringent**
(Sandmann, 2007)

**More stringent input alignment filtering:**
based on most conserved regions (phastCons)

out of 2,246 predictions, 53 are **experimentally validated**

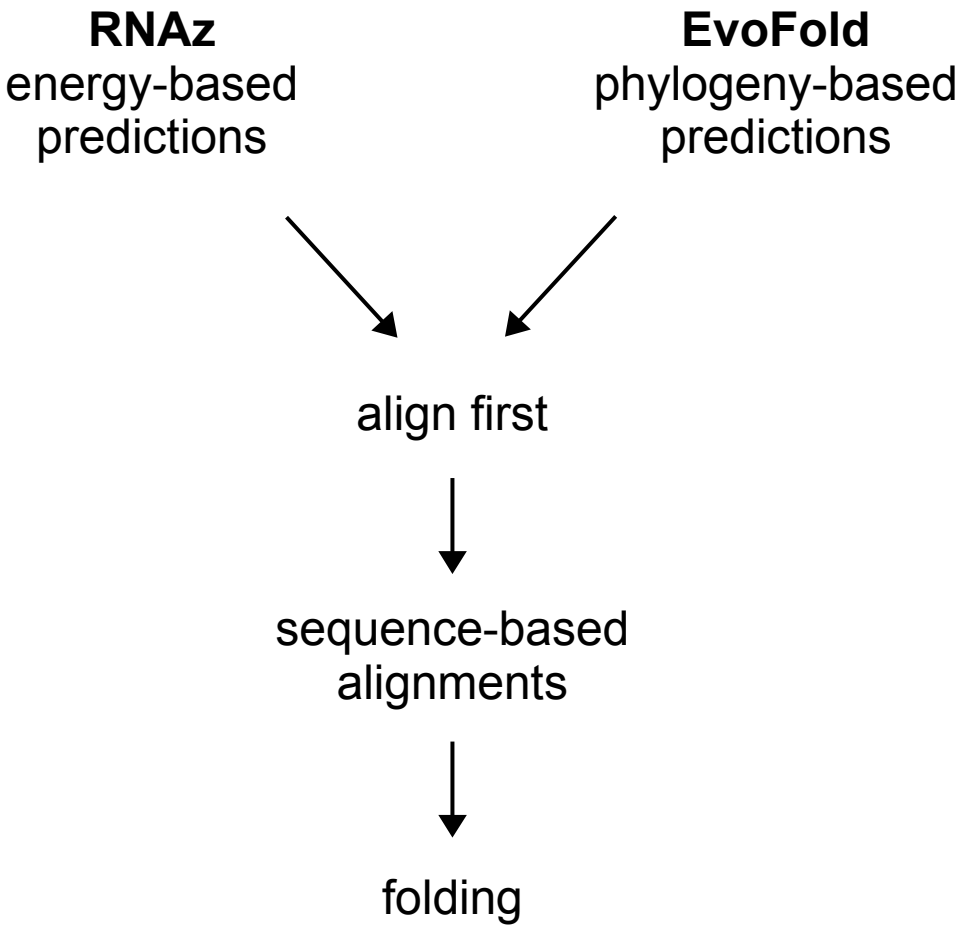
**EvoFold[2] screen**
(Stark, 2007)

**More stringent input alignment filtering:**
based on most conserved regions (phastCons)

22,682 predictions

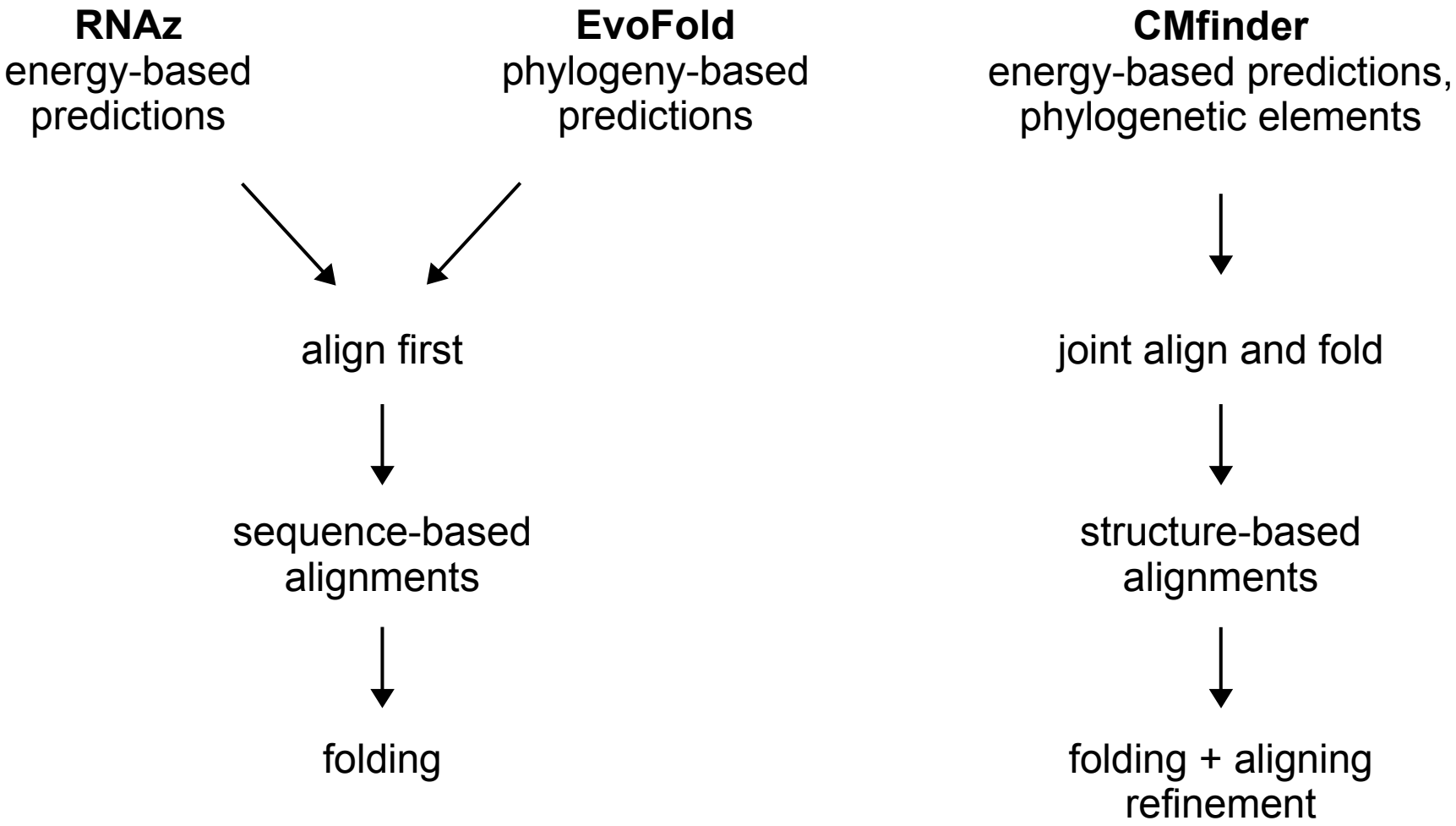
[1] Washietl, 2005 [2] Pedersen, 2006

# Structured RNA Prediction Tools



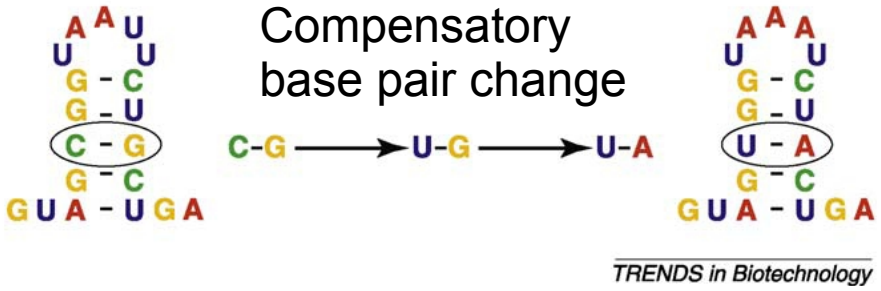

[1] Washietl, 2005 [2] Pedersen, 2006

# Structured RNA Prediction Tools



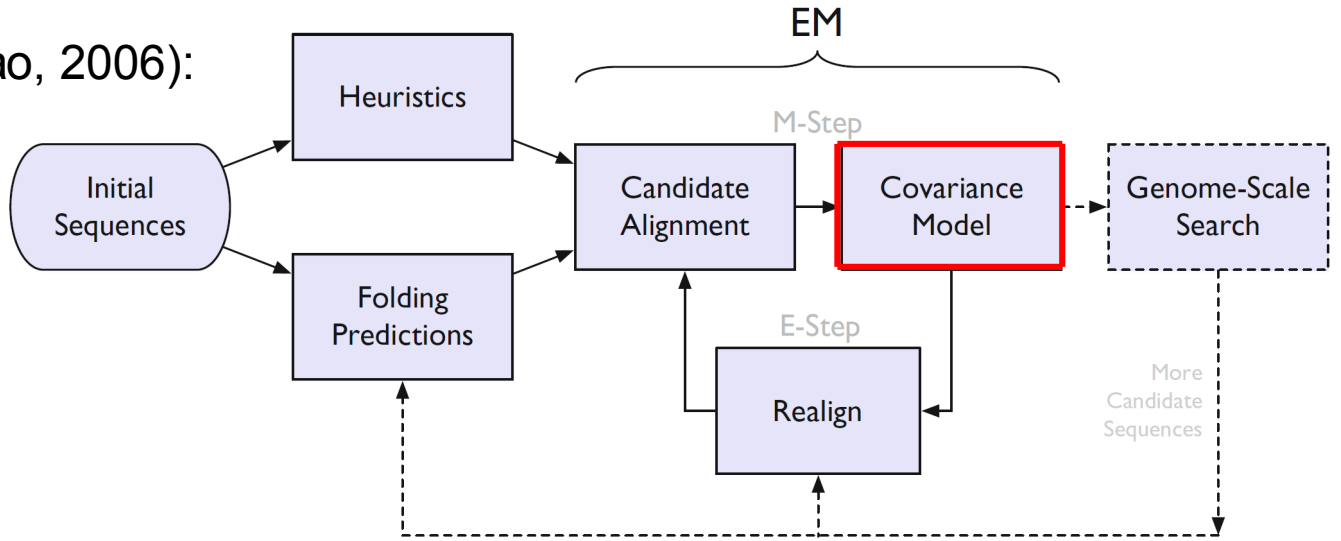

[1] Washietl, 2005 [2] Pedersen, 2006 [3] Yao, 2006

# Why Structure-Based Alignments Are Advantageous



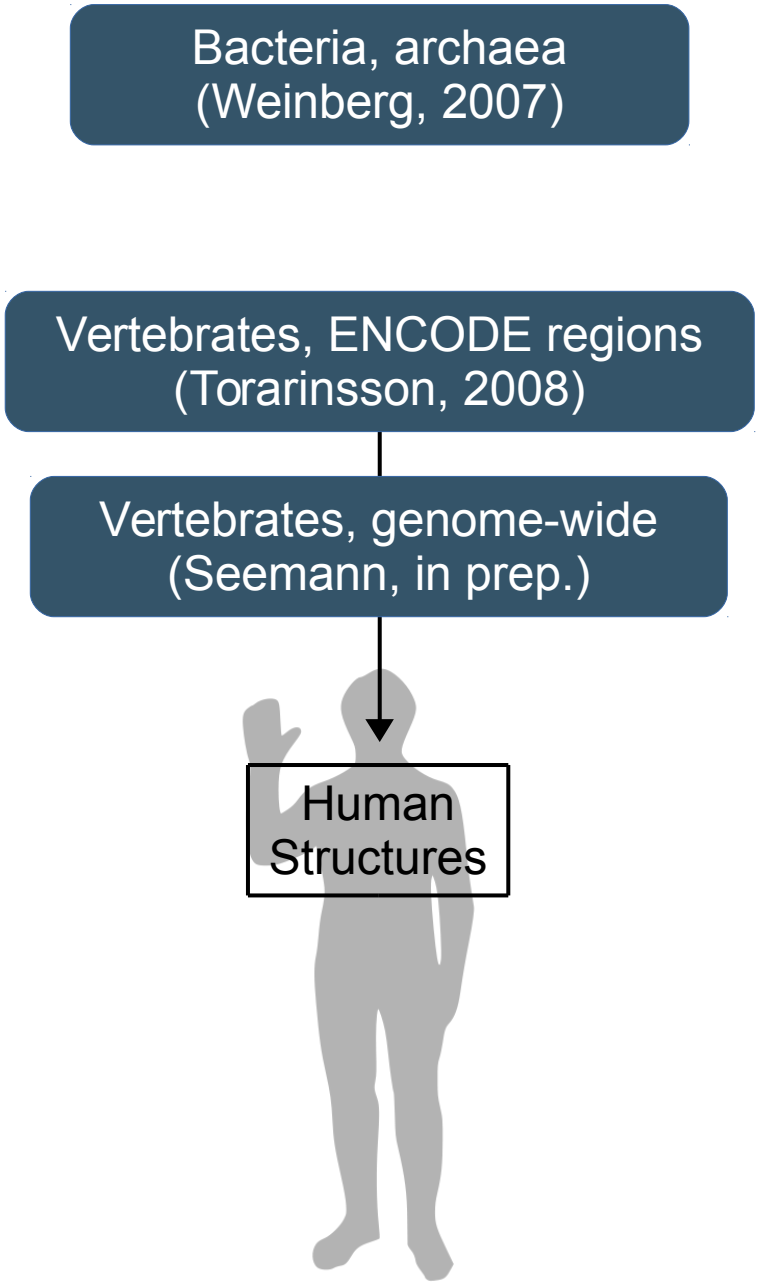

adapted from Gorodkin, 2010

**CMfinder** (Yao, 2006):




adapted from Ruzzo, 2014

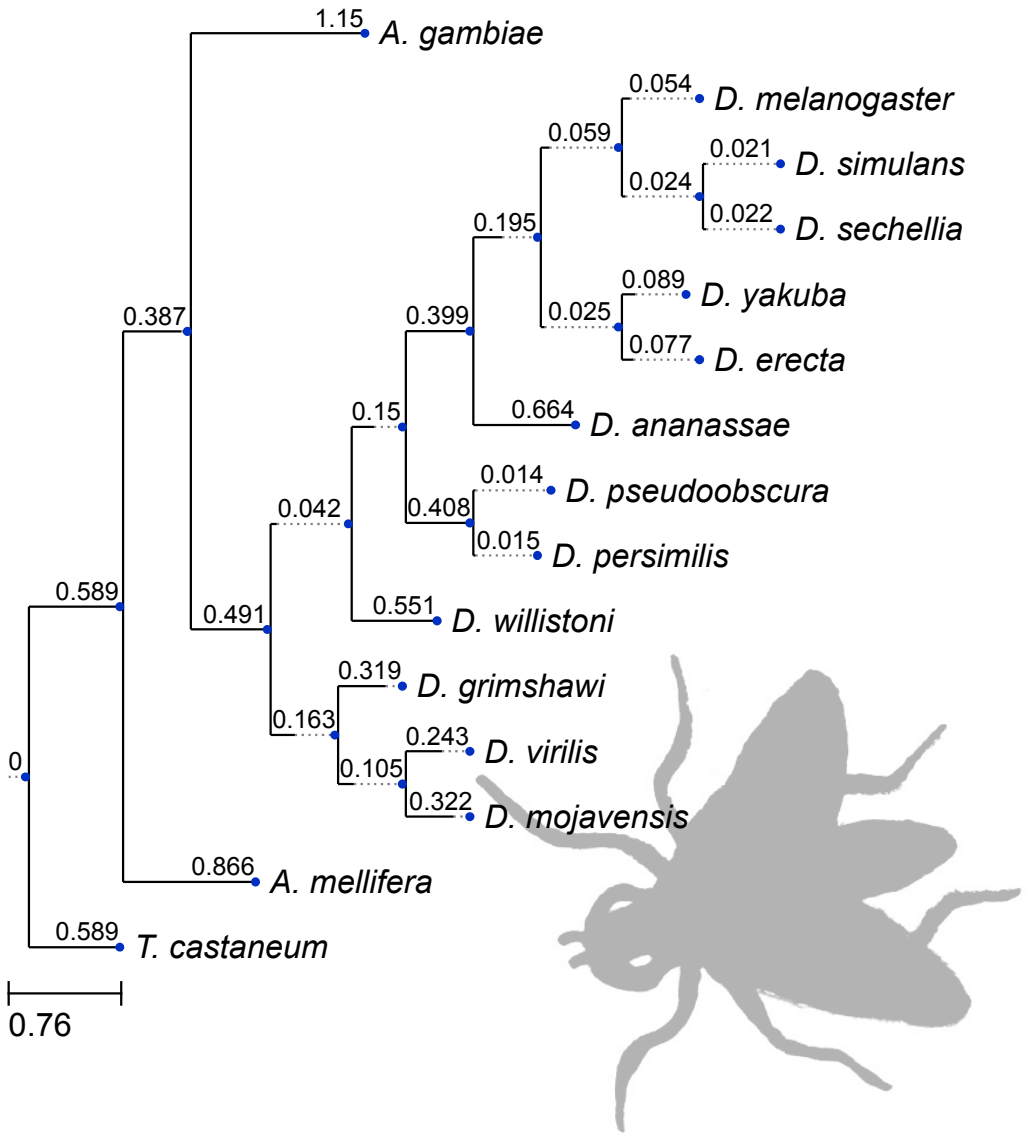
# Prediction of Structured RNAs Using CMfinder



# Prediction of Structured RNAs in Drosophilids Using CMfinder

15-way MULTIZ insect alignment

*D. melanogaster* (dm3, Apr. 2006, BDGP Release 5)
*D. simulans* (droSim1, Apr. 2005)
*D. sechellia* (droSec1, Oct. 2005)
*D. yakuba* (droYak2, Nov. 2005)
*D. erecta* (droEre2, Feb. 2006)
*D. ananassae* (droAna3, Feb. 2006)
*D. pseudoobscura* (dp4, Feb. 2006)
*D. persimilis* (droPer1, Oct. 2005)
*D. willistoni* (droWil1, Feb. 2006)
*D. virilis* (droVir3, Feb. 2006)
*D. mojavensis* (droMoj3, Feb. 2006)
*D. grimshawi* (droGri2, Feb. 2006)
*A. gambiae* (anoGam1, Feb. 2003)
*A. mellifera* (apiMel3, May 2005)
*T. castaneum* (triCas2, Sep. 2005)




tree derived from UCSC 27-way MULTIZ alignment

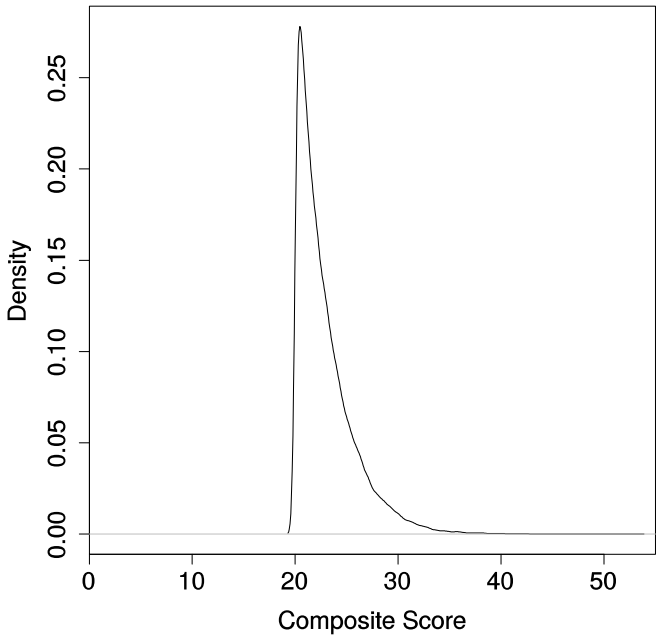
# Scoring CMfinder Predictions – Composite Score

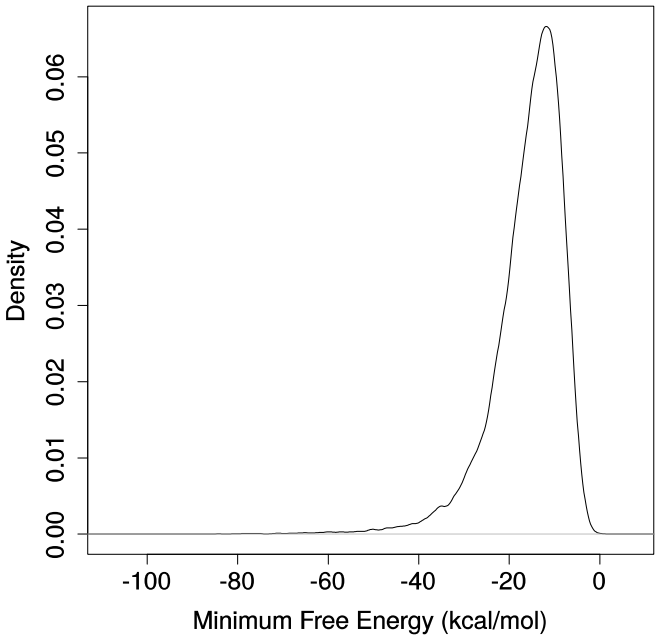
**Torarinsson et al. 2008 – CMfinder on ENCODE Regions**

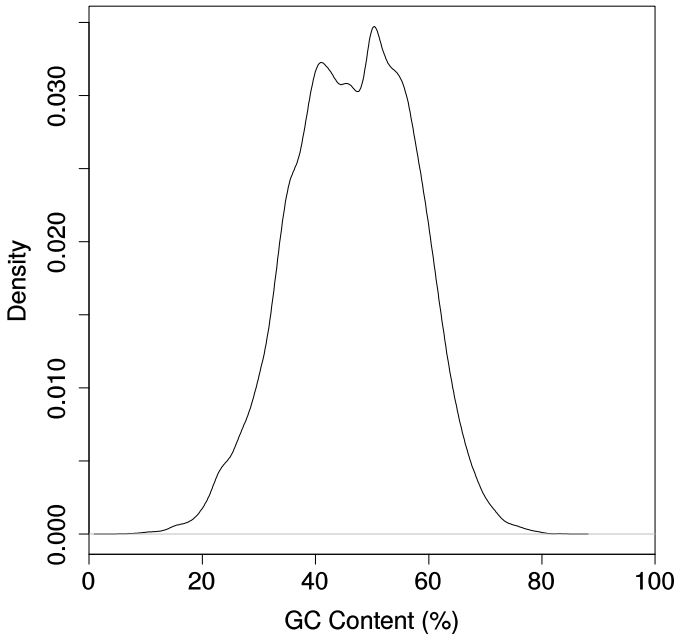
$$r = sp \cdot \frac{\sqrt{lc}}{sid} \cdot \frac{bp}{len}$$

| | |
|---|---|
| **sp** | Number of species. |
| **bp** | Number of base pairs in the consensus structure. |
| **lc** | Local sequence conservation (total size of blocks with at least 4 consecutive columns with more than 70% sequence identity). |
| **sid** | Average pairwise sequence identity. |
| **len** | Average sequence length. |

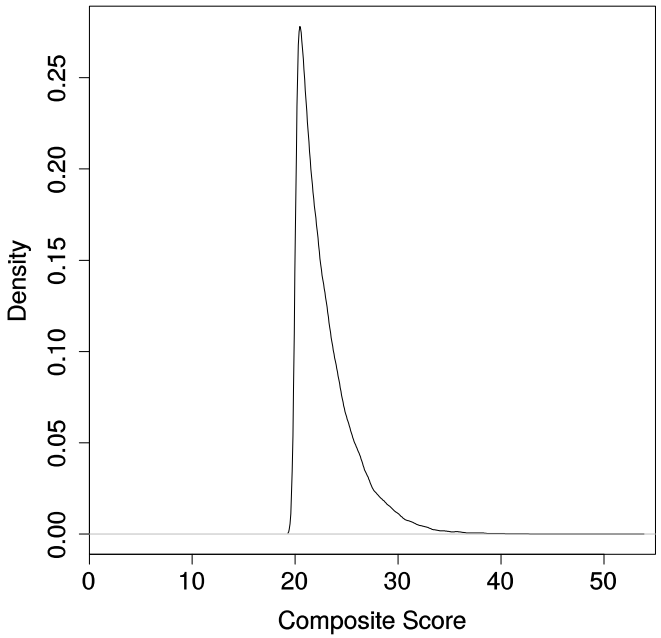
# Composite Score ≥ 20 → 67,910 Predictions

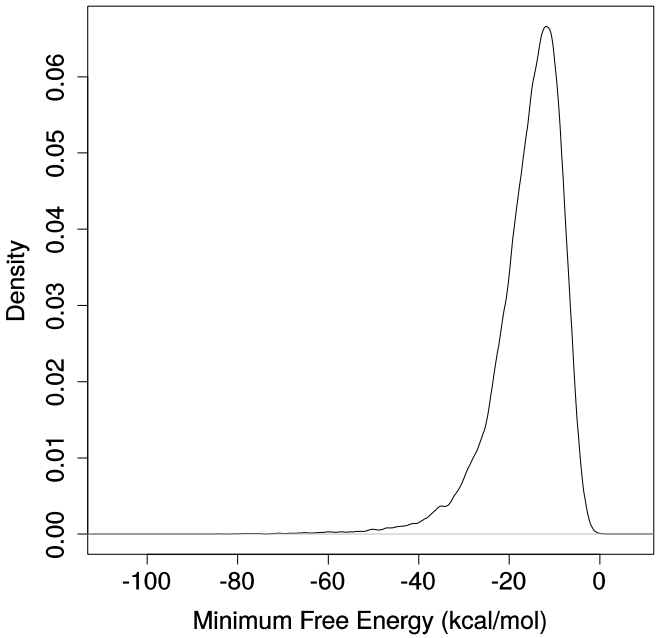


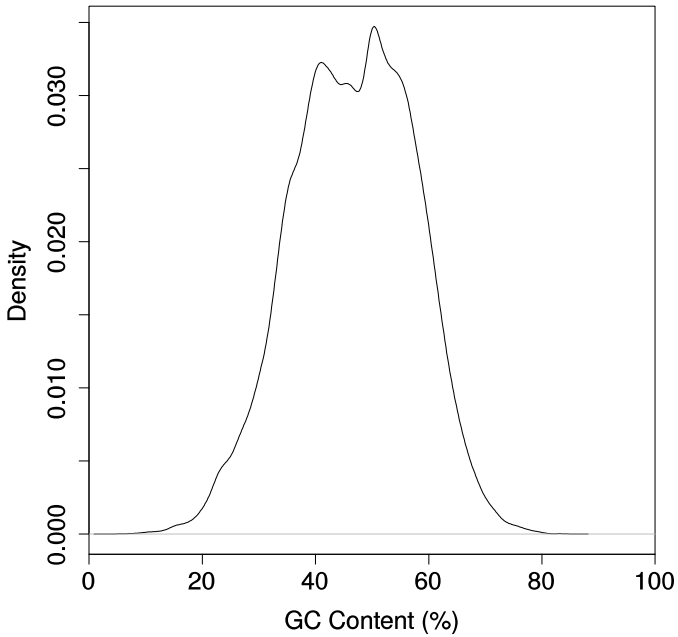



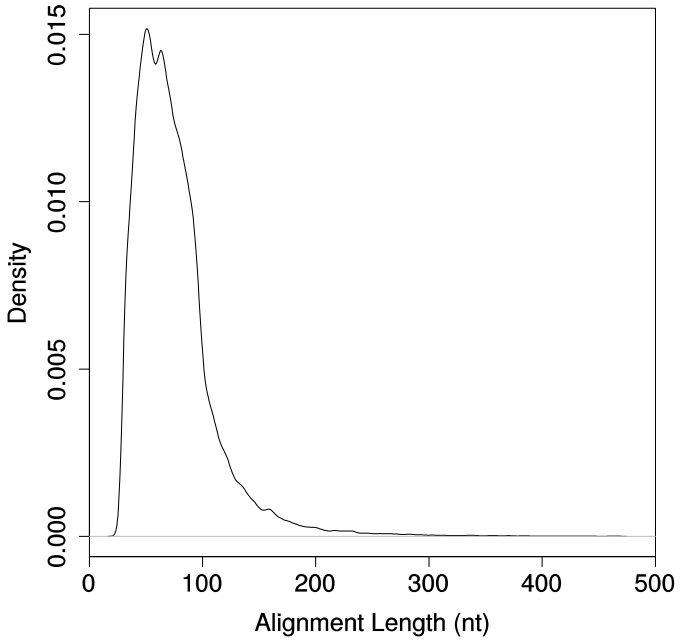



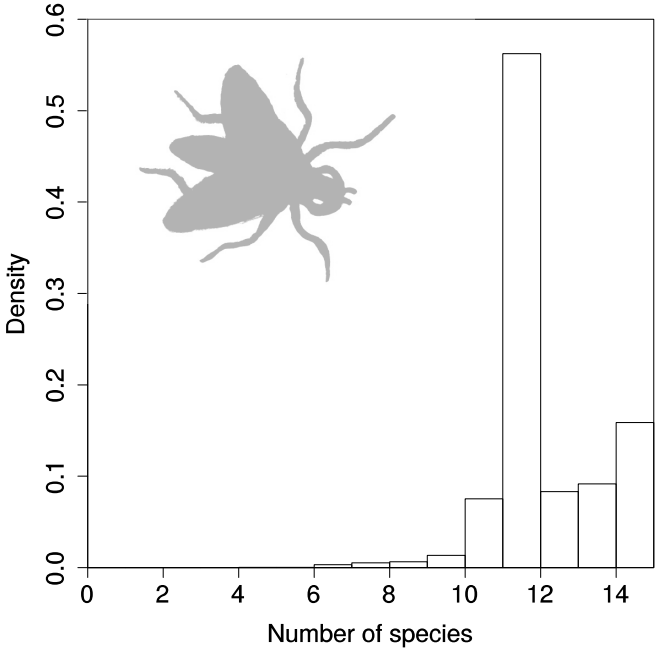

# Composite Score ≥ 20 → 67,910 Predictions

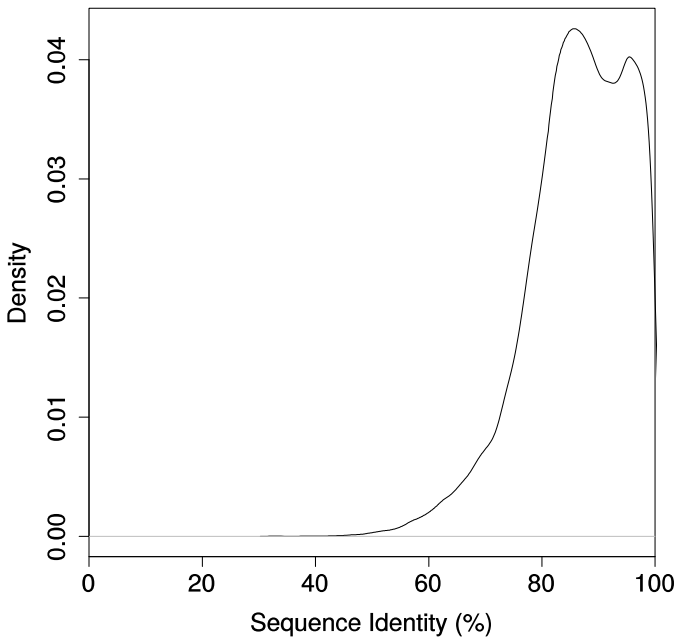


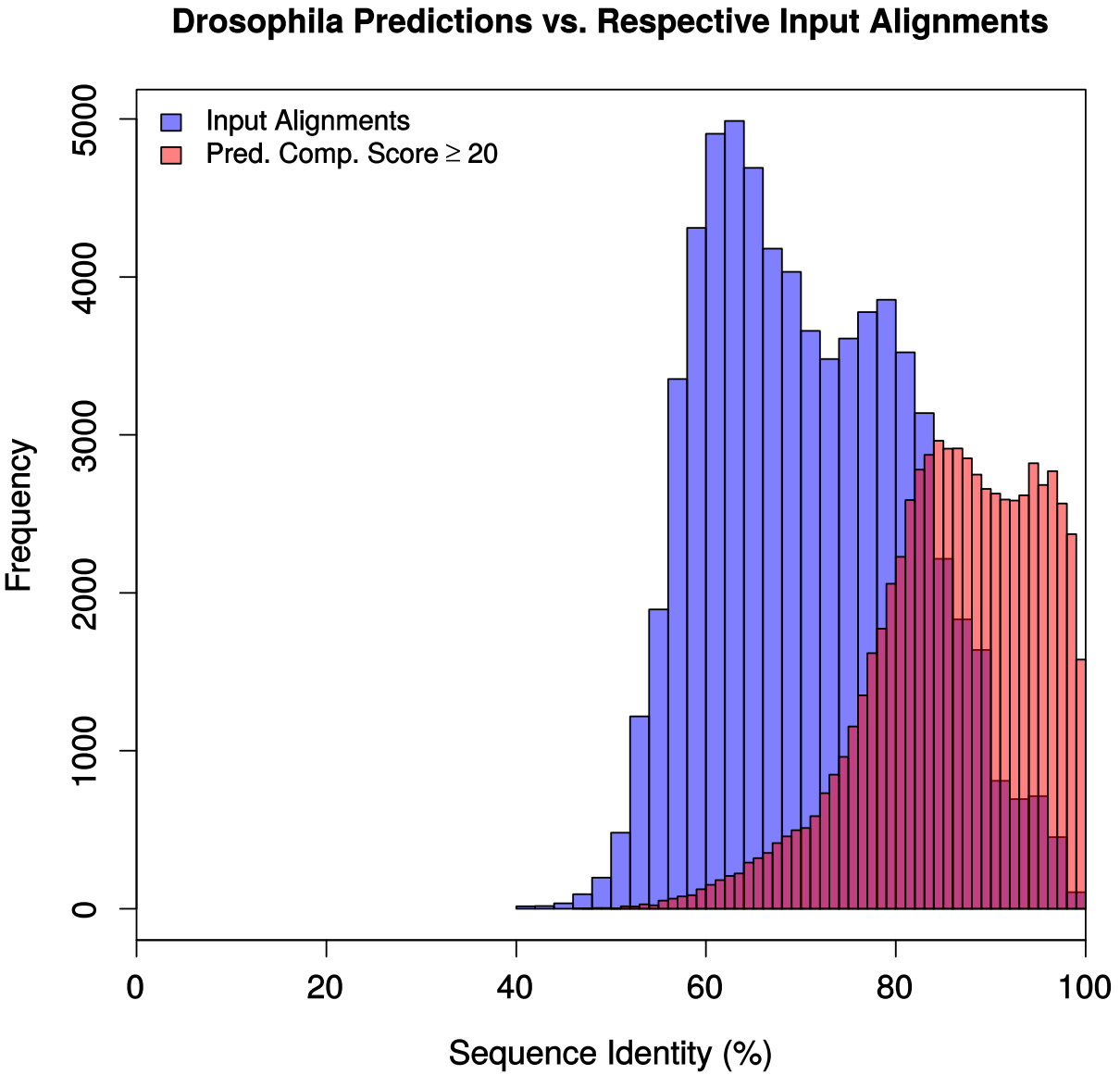



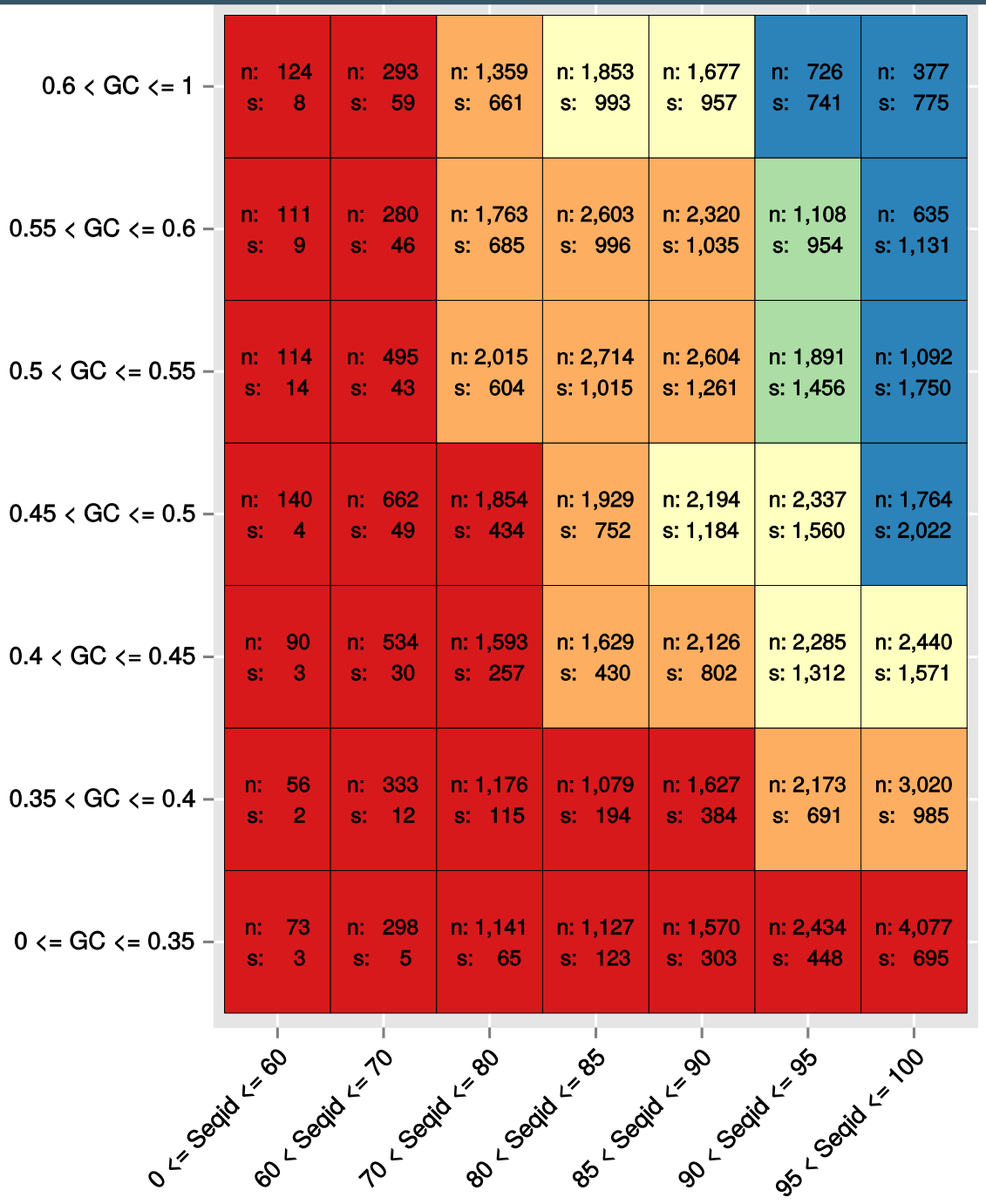



# Comparison Sequence Identity



# False Discovery Rate Depends on GC Content and Sequence Identity




$$FDR = \frac{\#shuffled\ predictions}{\#native\ predictions}$$

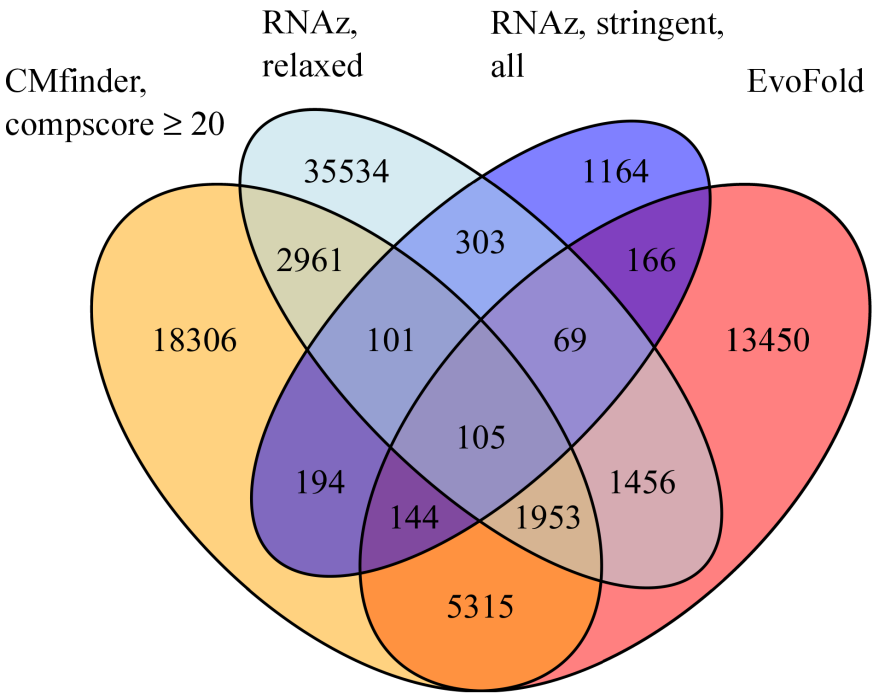
**FDR Estimate**

- (0,0.25]
- (0.25,0.5]
- (0.5,0.75]
- (0.75,1]
- (1,Inf]

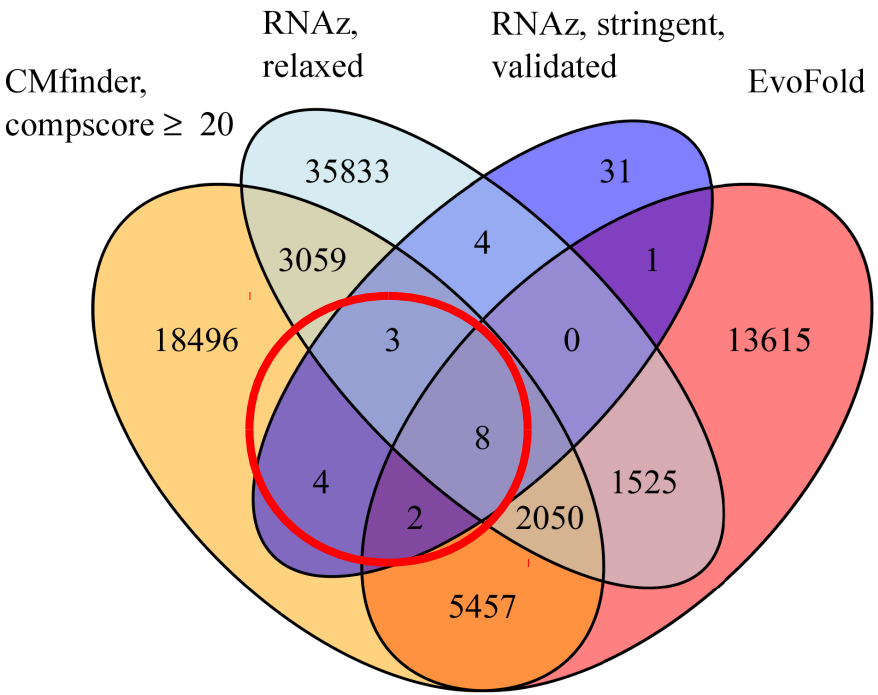
number of predictions on **n**: native and **s**: shuffled genome

# Comparison with Other Screens

Screen Overlap with Total RNAz (Stringent) Dataset



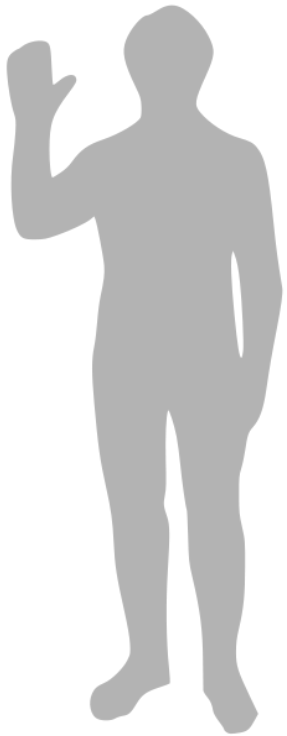

Screen Overlap with Experimentally Validated RNAz (Stringent) Dataset



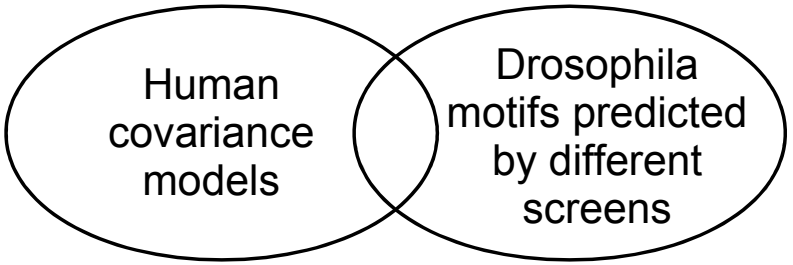

Minimum 1 bp overlap

# Outlook

Large number of predictions, high FDR
→ need to add more information to select meaningful candidates



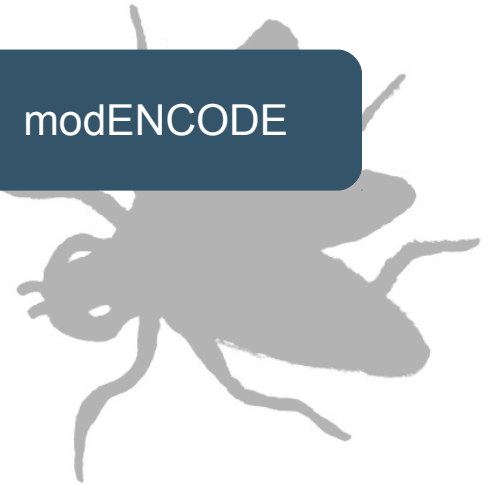
CMfinder screen on vertebrates (Seemann, in prep.)



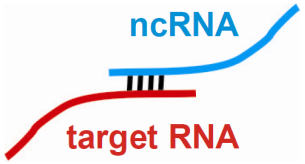

CMsearch
(Nawrocki, Eddy, 2013)

Tissue-spec. / diff. expressed transcripts conserved in Drosophilids

modENCODE



candidate selection




Functional analysis of novel ncRNAs in fly

# Thanks to...

**Leipzig**

Peter F. Stadler

**Copenhagen**

Stefan Seemann
Jan Gorodkin

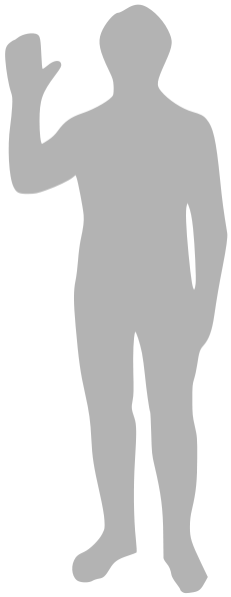
all other RTH members

Stephen M. Cohen

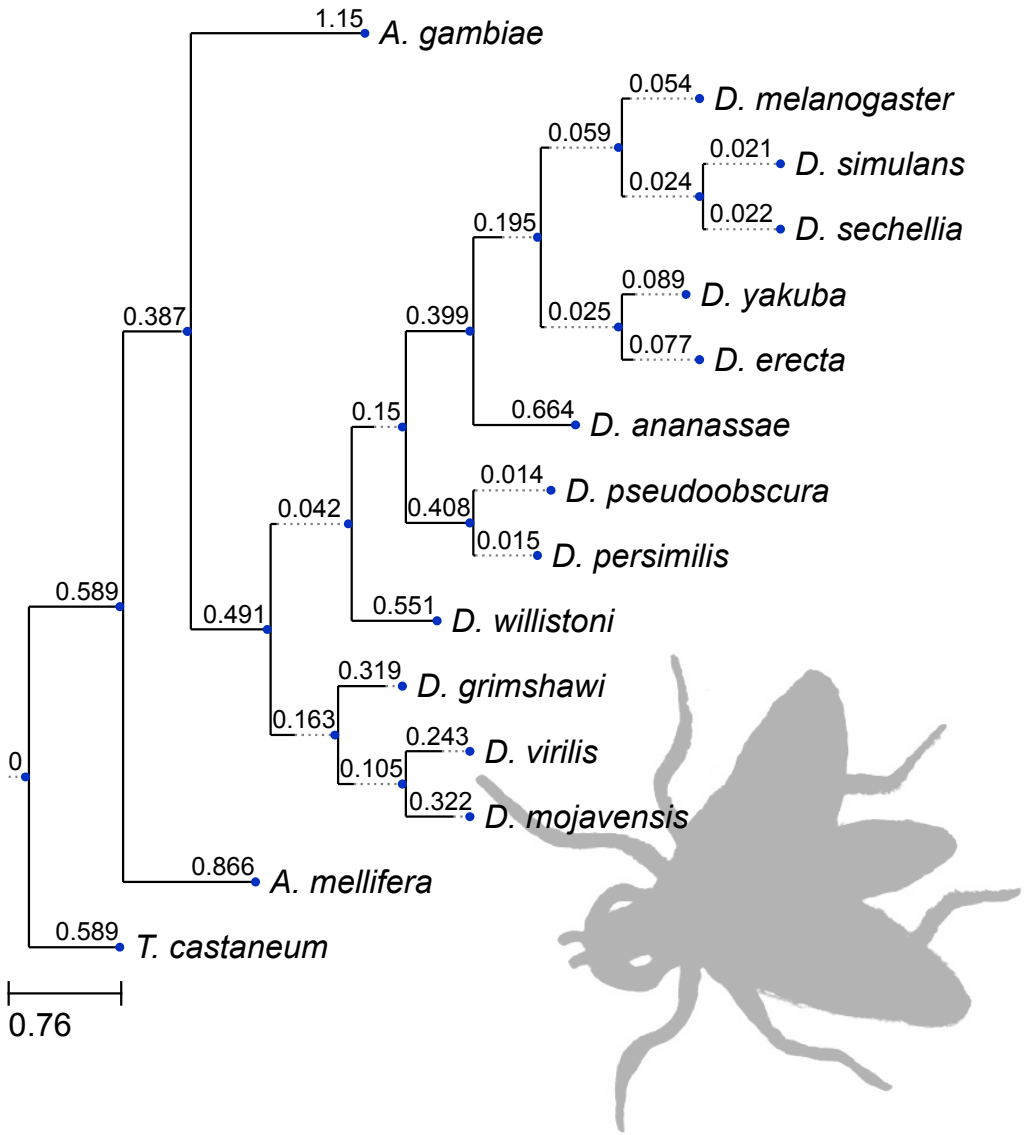
**Seattle**

Walter L. Ruzzo


# Prediction of Structured RNAs in Drosophilids






UCSC 17-way MULTIZ alignment tree

tree derived from UCSC 27-way MULTIZ alignment

# Comparison with Other Screens

| CMfinder | EvoFold | RNAz (Rose) | RNAz (Sandmann) |
|---|---|---|---|
| covariance model derived from initial local alignment, joint refinement, SCFG | compares an RNA model to an unstructured background model, takes phylogenetic information into account (tree structure, branch lengths), SCFG sliding window | measures thermodynamic stability (compares stability of native and shuffled seqences) and structural conservation, SVM sliding window | |
| joint align and fold | align first | align first | |

# Comparison with Other Screens

| CMfinder | EvoFold | RNAz (Rose) | RNAz (Sandmann) |
|---|---|---|---|
| 15-way MULTIZ alignment | 12-way MULTIZ alignment | 12-way Pecan alignment | 15-way MULTIZ alignment |
| | based on phastCons regions + flanks | > 50 nt alignment length,<br>< 25 % gap characters,<br>GC content 25 to 75 %,<br>> 3 sequences;<br>max sequence number 6;<br>overlapping hits with p > 0.5 were combined | most conserved regions (phastCons) with flanking sequences; sequences overlapping annotated genes were removed |
| | removed weak predictions (< 10 bp or > 50 % bulges in stems, predictions overlapping low-complexity repeats); in case of overlap, only highest scoring prediction is kept [22,682] | hits kept with p > 0.5 [42,482] (p > 0.9 → high-confidence set [16,377]) | hits kept with p > 0.9 |