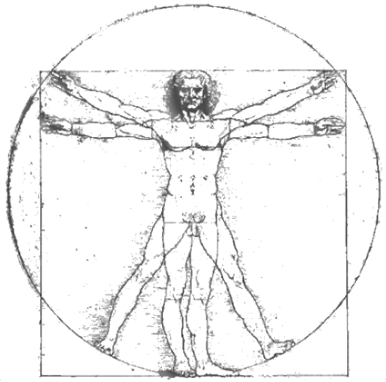
Structure-based Whole Genome Realignment Reveals Many Novel Non-coding RNAs

Sebastian Will, Michael Yu, Bonnie Berger

Previous Genome-wide Screens for Structural ncRNA

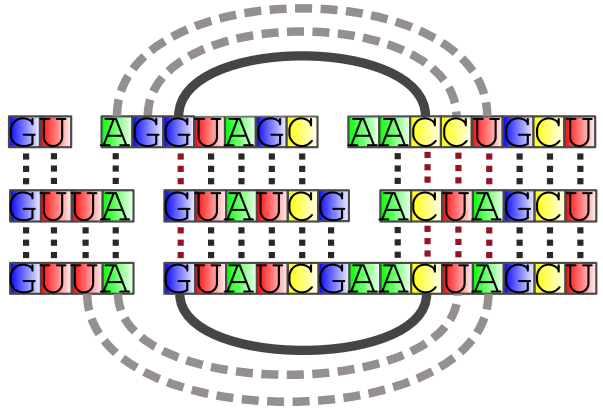


RNAz	Washietl, *Nat Biotech*, 2005	Rose, *BMC Genomics*, 2007
EvoFold	Pedersen, *PLoS CB*, 2006	Bradley, *PLoS One*, 2006

- predict thousands of putative structural ncRNAs
- energy model + comparative information:
  WGAs, *conserved stable structure*
- for efficiency: only *correctly aligned* structure

Challenges for ncRNA Screens

- WGA tools regularly misalign structural RNA



RNAz

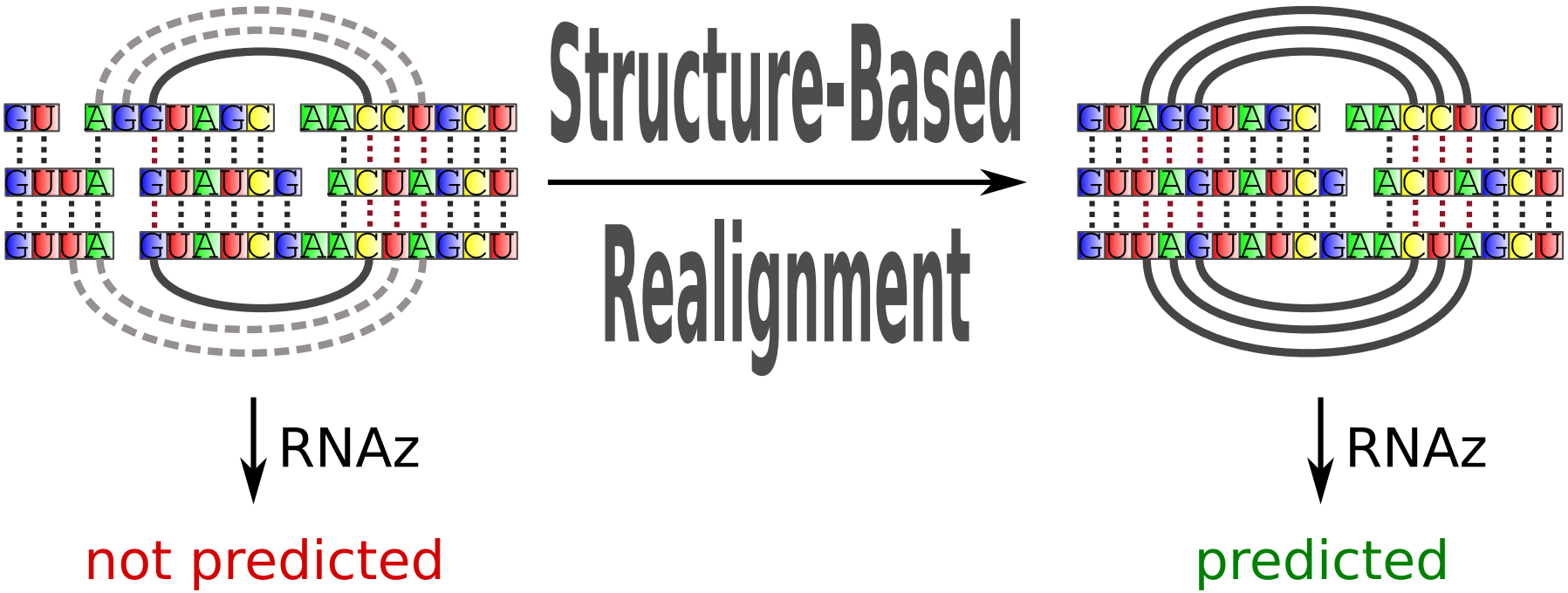
not predicted

Conjecture:

- *underestimate true number of structural ncRNAs*
- *biased for 'easy targets'*

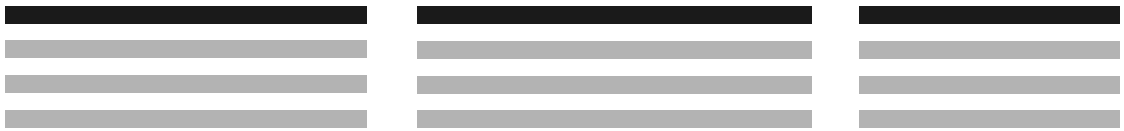
Revealing Hidden ncRNA

Idea: patch mistakes by structure-based realignment

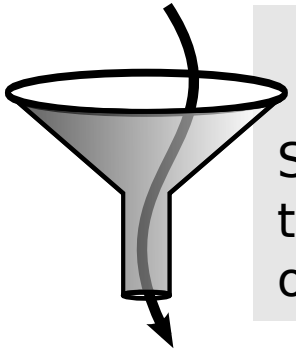


Massachusetts Institute of Technology
S.Will

The REAPR Pipeline



Whole genome alignment



1

Slice and filter by thermodynamic stability of single RNA structures

Stable loci

Unstable conserved structure of original locus alignment

Realign based on sequence and structure

2

Estimate ncRNA likelihood

3

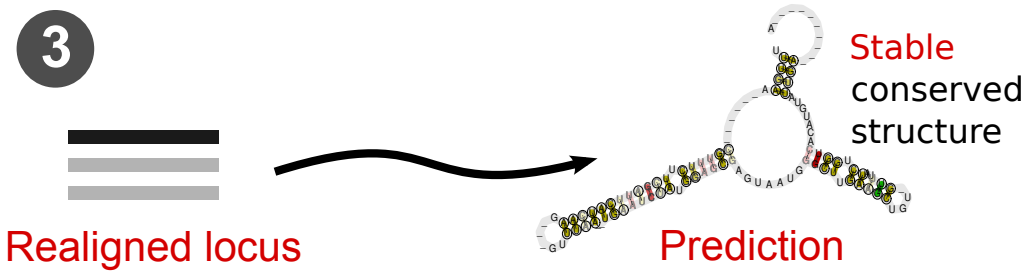
Stable conserved structure

Realigned loci

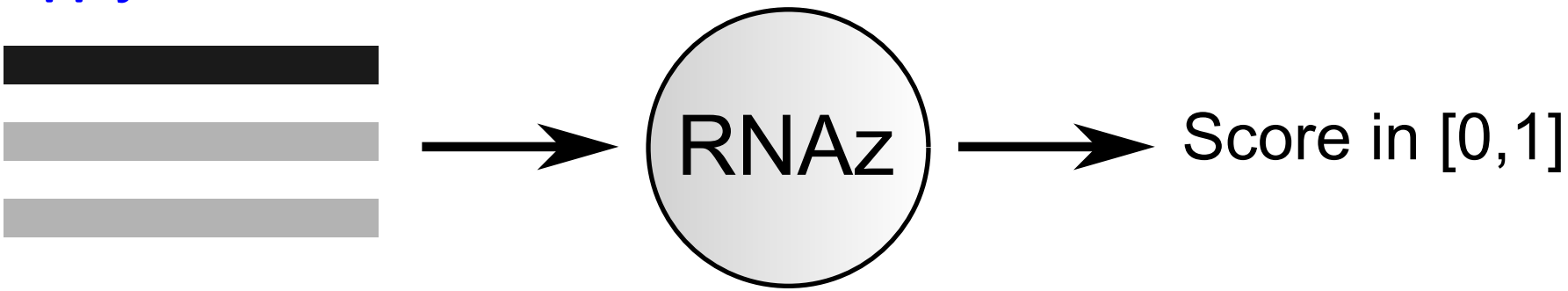
Predictions (q-values)

Massachusetts Institute of Technology
S.Will

Estimate ncRNA Likelihood



Apply RNAz

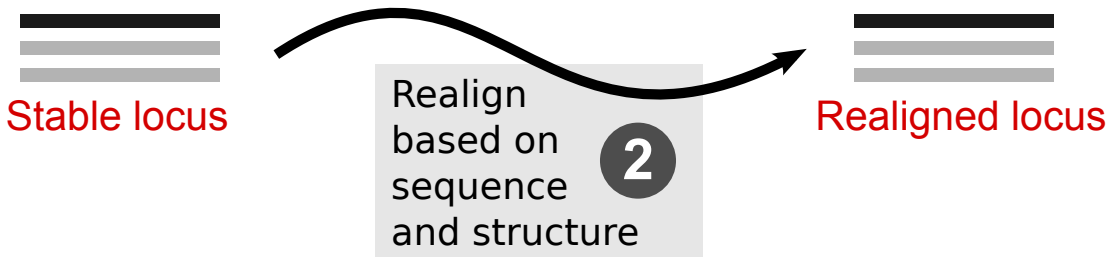


recall: RNAz sees only *correctly aligned* conserved structure

Washietl *et al.* PNAS, 2005.

Gruber *et al.* PSB 2010. *(RNAz 2.0)*

Making ncRNAs Visible for De Novo Predictors

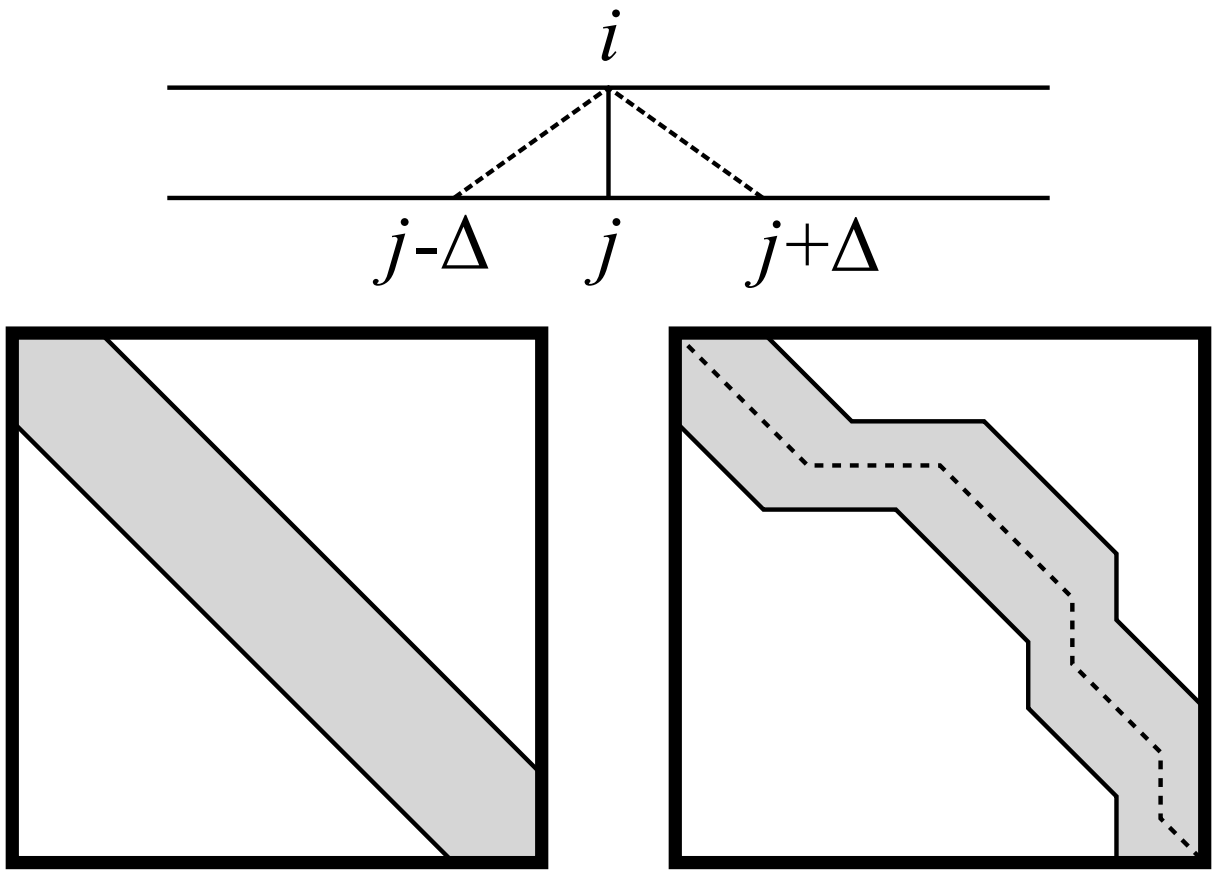


Apply "gold-standard" of structure-based RNA alignment: simultaneous alignment and folding

- use Sankoff-algorithm ? $O(n^6)$
- use fast and accurate Sankoff-variant ?
  e.g., *LocARNA* [Will et al., 2007], $O(n^4)$

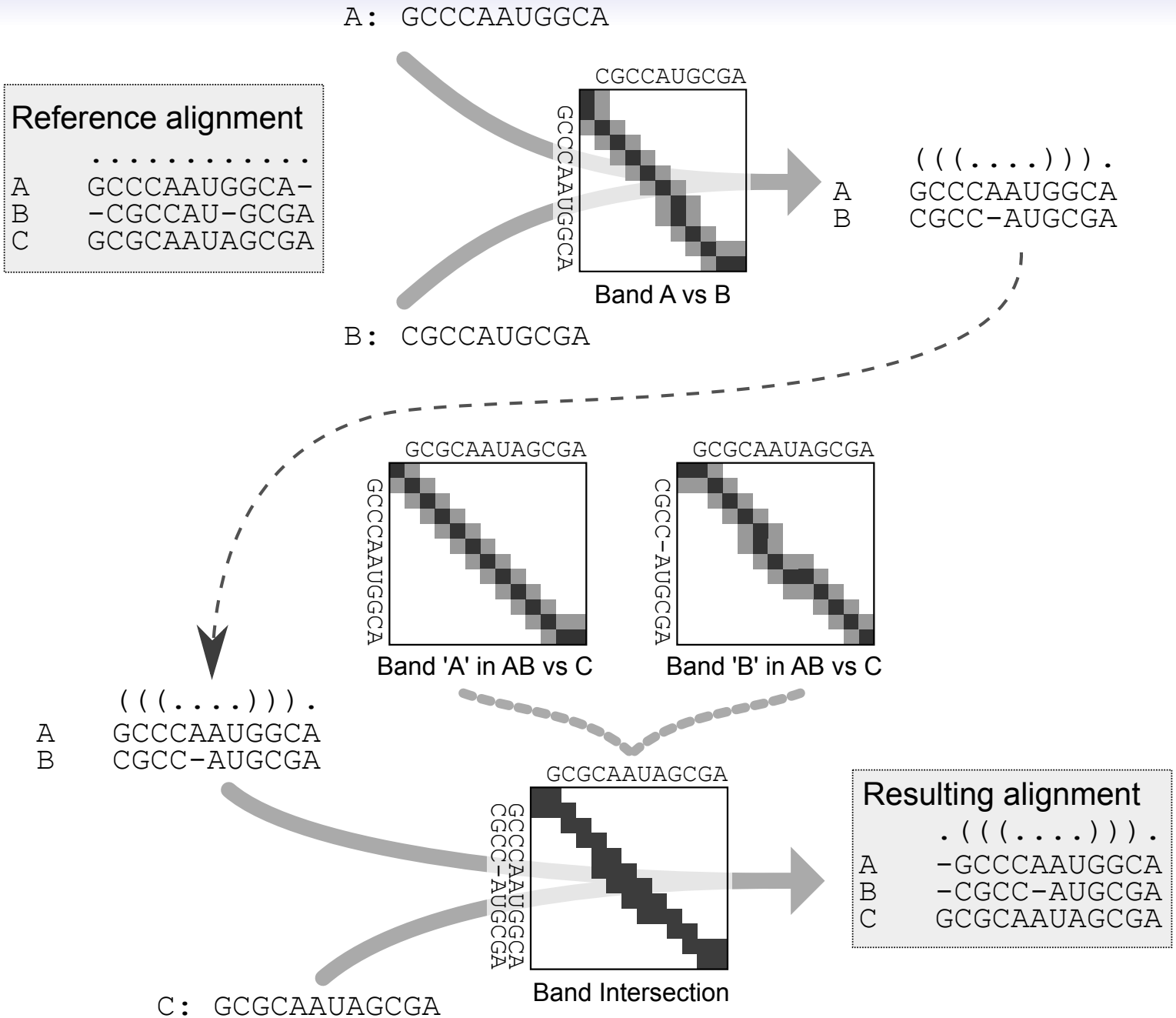
Efficient Realignment: Novel Banding Algorithm

novel **banding**: compute **realignment** *within limited deviation* Δ *from a reference alignment*, e.g., $\Delta = 20$ (± 20 nt)



main technical challenge: banding for *multiple* alignment

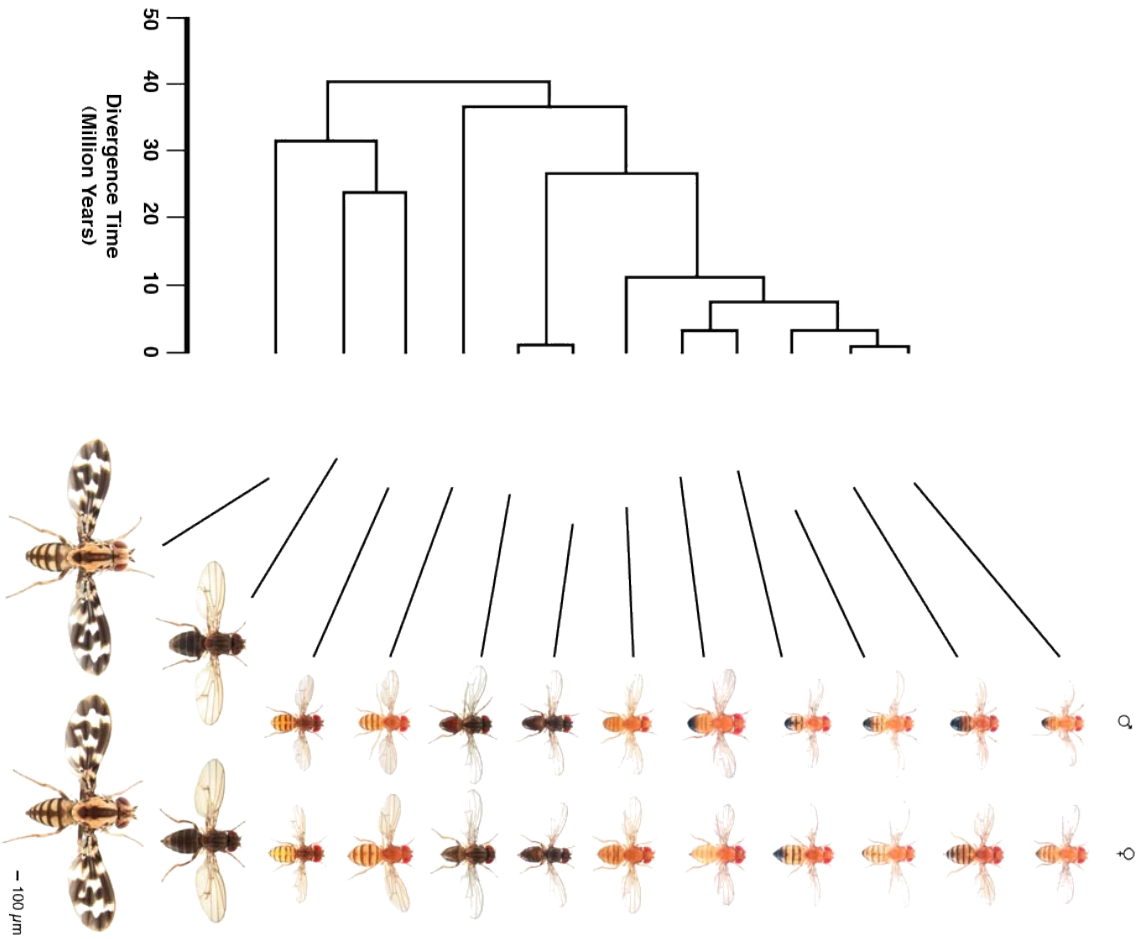
Banding for Progressive Multiple Alignment



Massachusetts Institute of Technology
S.Will

Applying REAPR: **Reaping ncRNAs in Fly**

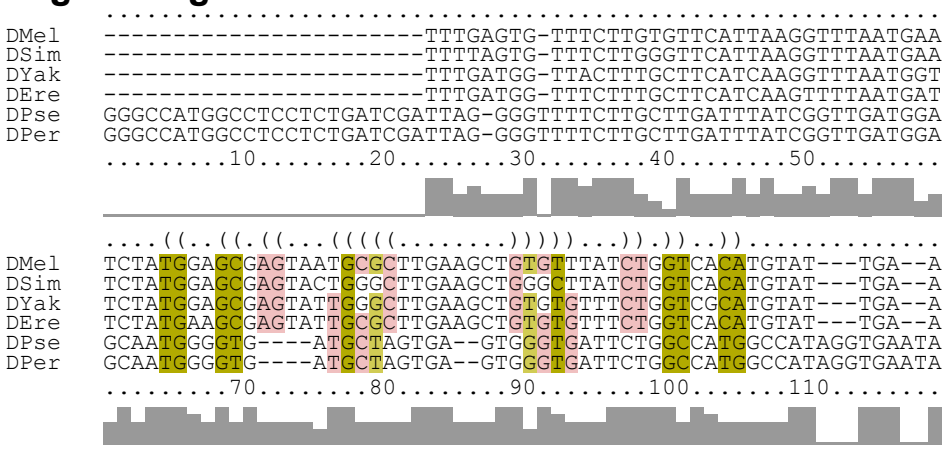
- First computational screen of a eukaryotic genome that corrects structural misalignments
- control *w/o realignment*



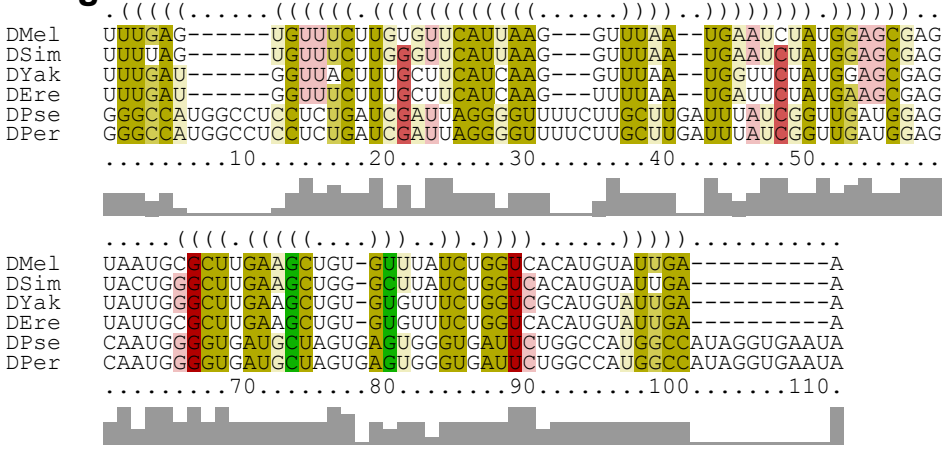
Massachusetts Institute of Technology
S.Will

A Novel Putative Motif in the Long ncRNA roX1

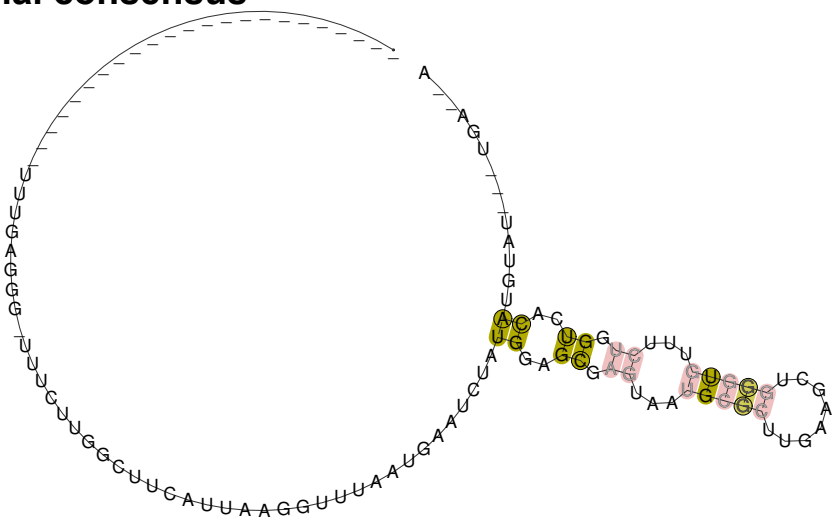
original alignment



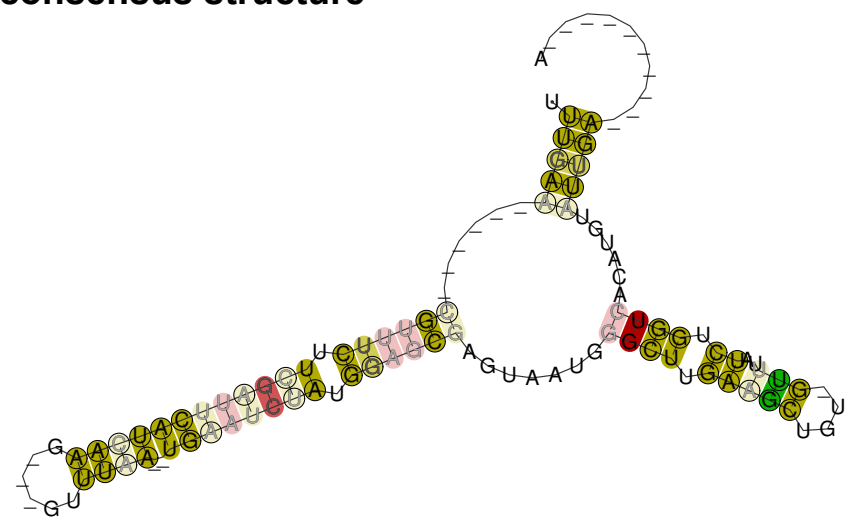
realignment



original consensus



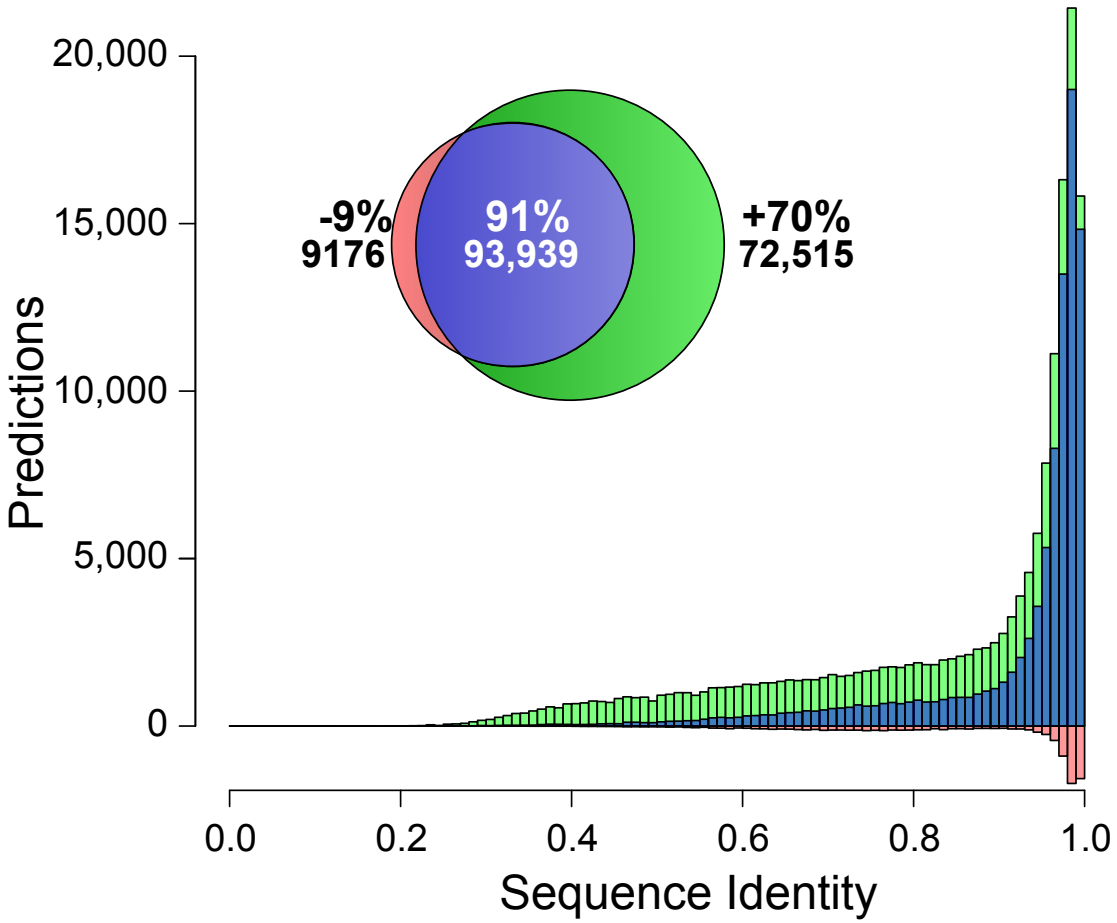
new consensus structure



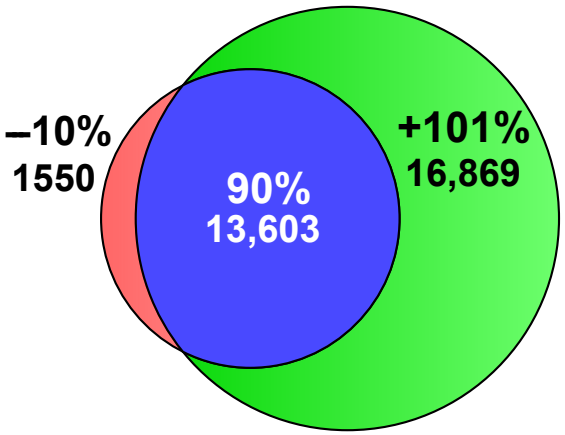
Massachusetts Institute of Technology
S.Will

Novel Predictions due to Realignment

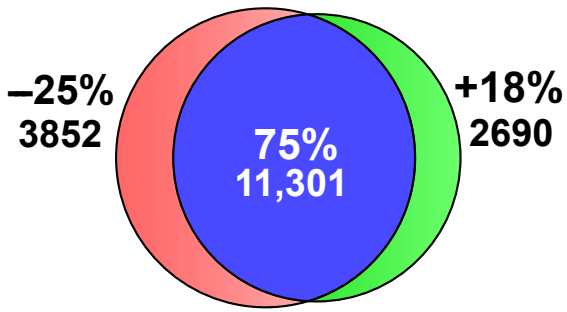
all predictions



***D. melanogaster* predictions**



REAPR with LocARNA



REAPR with Muscle

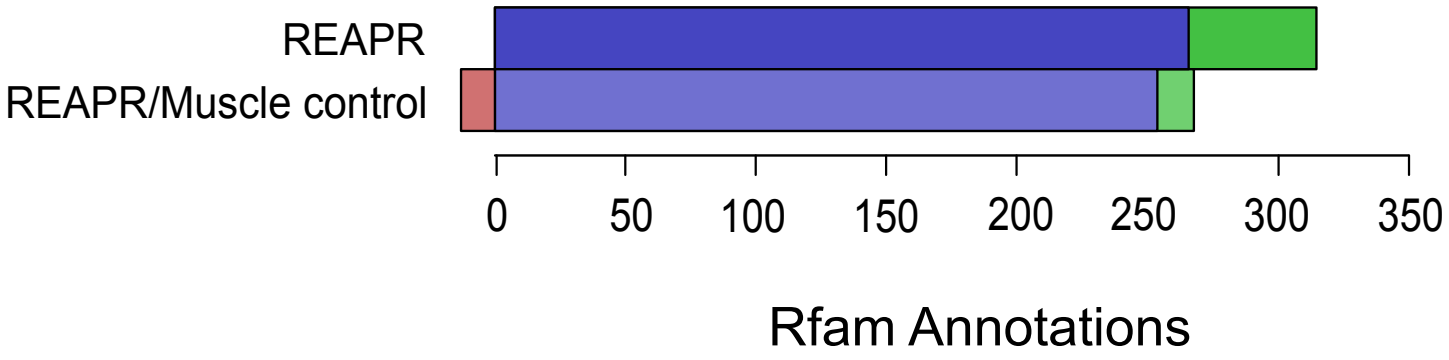
■ = common predictions by REAPR and control

■ = novel predictions by REAPR

■ = predictions only by control

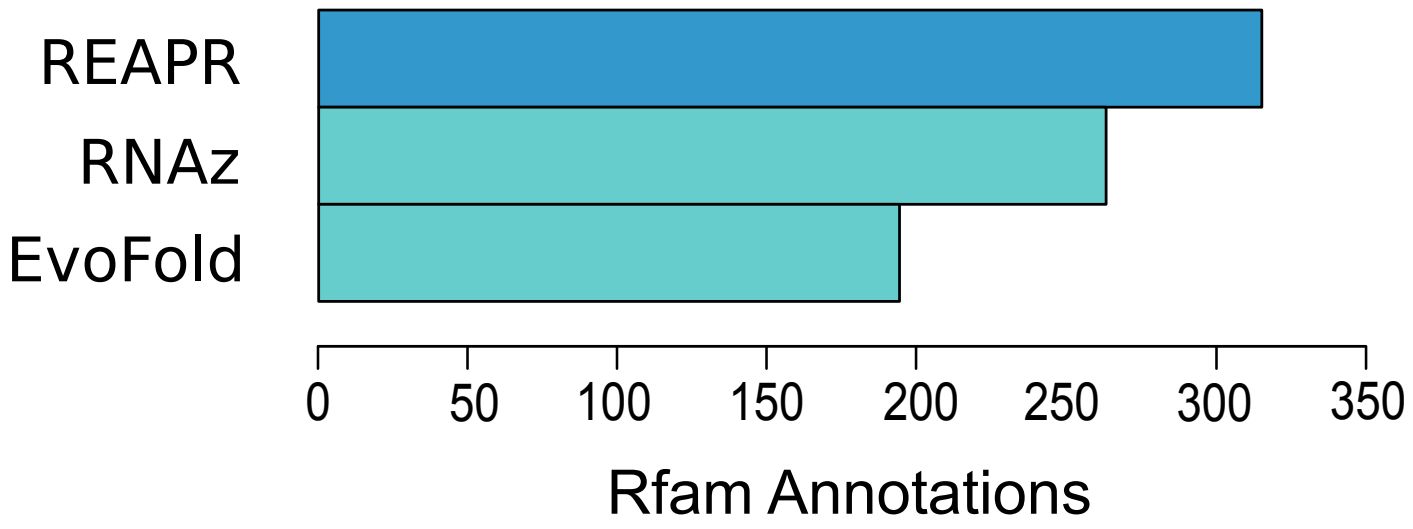
Massachusetts Institute of Technology
S.Will

Sensitivity for Rfam ncRNAs in *D. melanogaster*
REAPR vs. Controls

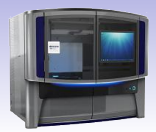


■ = common predictions of REAPR and control
■ = novel predictions by REAPR
■ = predictions only by control

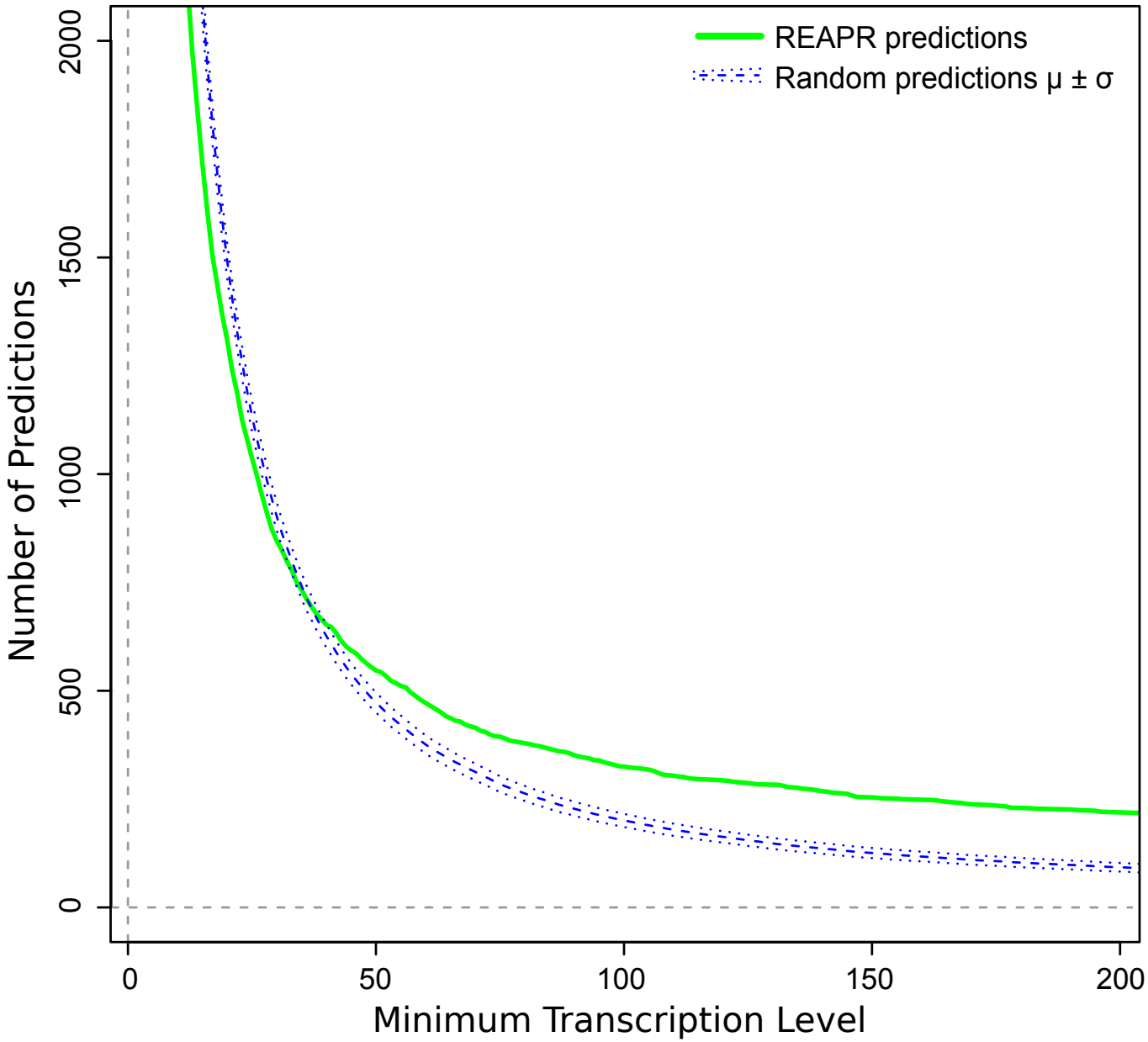
Sensitivity for Rfam ncRNAs in *D. melanogaster*
REAPR vs. Previous Screens

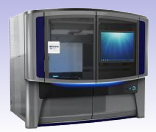


Massachusetts Institute of Technology
S.Will

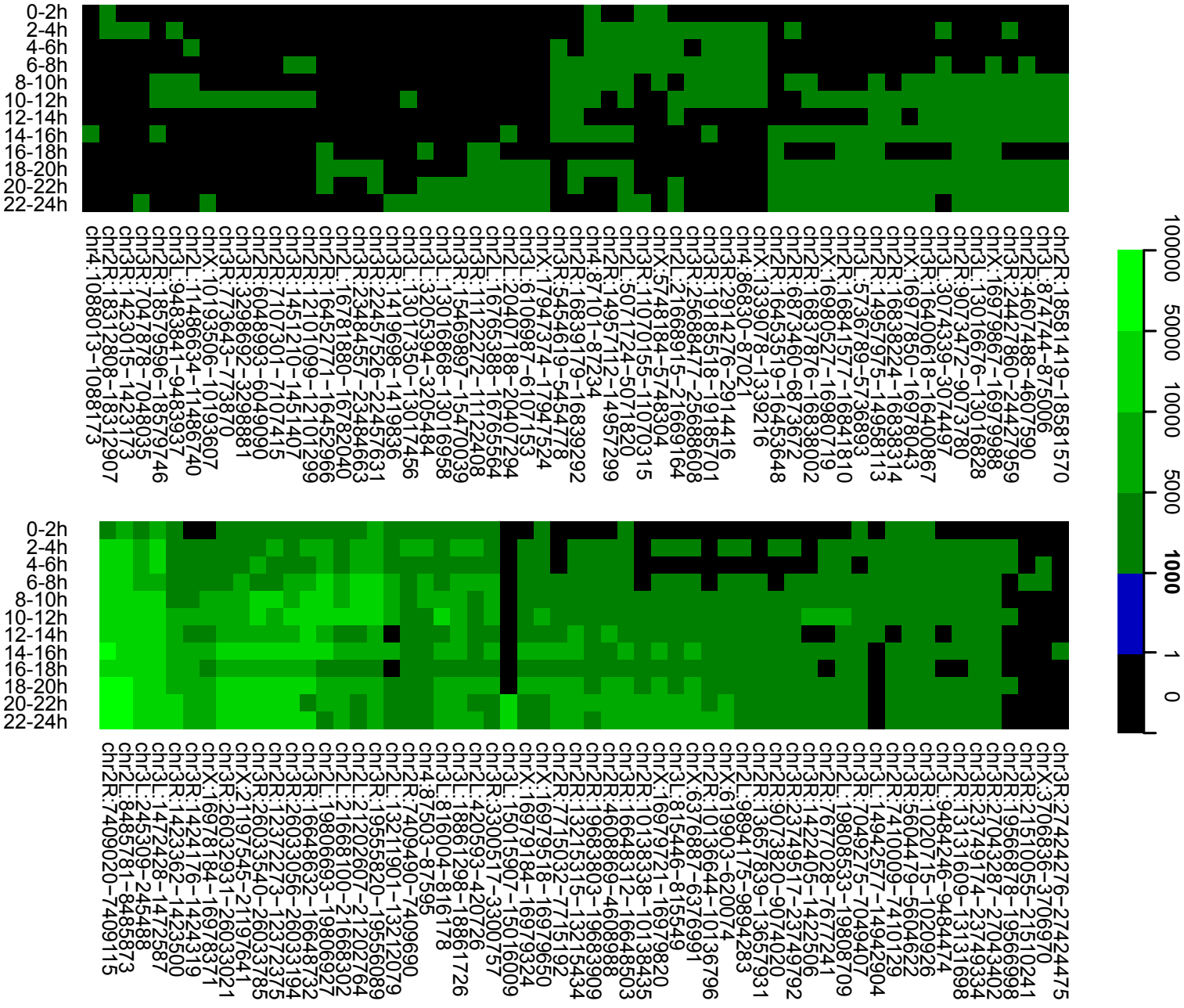


RNA-seq (modENCODE) Confirms Predictions





Combining Fly Embryo RNA-seq and REAPR



Massachusetts Institute of Technology
S. Will

Conclusions

- *REAPR*: "Realignment for Prediction of ncRNAs"
- *First* genome-wide ncRNA screen
  that *corrects structural misalignments*
- *Twice as many D. melanogaster ncRNAs*
  while *less than doubling run-time*
- *Novel banding method* for fast multiple realignment
  implemented in the RNA alignment tool LocARNA
- Verified with *annotations and modEncode RNA-seq*
- REAPR + RNA-seq combined results in interesting
  structural ncRNAs
- *Future challenge:* detect non-locally misaligned RNAs

Conclusions

- *REAPR*: "Realignment for Prediction of ncRNAs"
- *First* genome-wide ncRNA screen
  that *corrects structural misalignments*
- *Twice as many D. melanogaster ncRNAs*
  while *less than doubling run-time*
- *Novel banding method* for fast multiple realignment
  implemented in the RNA alignment tool LocARNA
- Verified with *annotations and modEncode RNA-seq*
- REAPR + RNA-seq combined results in interesting
  structural ncRNAs
- *Future challenge:* detect non-locally misaligned RNAs

Conclusions

- *REAPR*: "Realignment for Prediction of ncRNAs"
- *First* genome-wide ncRNA screen
  that *corrects structural misalignments*
- *Twice as many D. melanogaster ncRNAs*
  while *less than doubling run-time*
- *Novel banding method* for fast multiple realignment
  implemented in the RNA alignment tool LocARNA
- Verified with *annotations and modEncode RNA-seq*
- REAPR + RNA-seq combined results in interesting
  structural ncRNAs
- *Future challenge:* detect non-locally misaligned RNAs

Conclusions

- *REAPR*: "Realignment for Prediction of ncRNAs"
- *First* genome-wide ncRNA screen
  that *corrects structural misalignments*
- *Twice as many D. melanogaster ncRNAs*
  while *less than doubling run-time*
- *Novel banding method* for fast multiple realignment
  implemented in the RNA alignment tool LocARNA
- Verified with *annotations and modEncode RNA-seq*
- REAPR + RNA-seq combined results in interesting
  structural ncRNAs
- *Future challenge:* detect non-locally misaligned RNAs

Conclusions

- *REAPR*: "Realignment for Prediction of ncRNAs"
- *First* genome-wide ncRNA screen
  that *corrects structural misalignments*
- *Twice as many D. melanogaster ncRNAs*
  while *less than doubling run-time*
- *Novel banding method* for fast multiple realignment
  implemented in the RNA alignment tool LocARNA
- Verified with *annotations and modEncode RNA-seq*
- REAPR + RNA-seq combined results in interesting structural ncRNAs
- *Future challenge:* detect non-locally misaligned RNAs

Conclusions

- *REAPR*: "Realignment for Prediction of ncRNAs"
- *First* genome-wide ncRNA screen
  that *corrects structural misalignments*
- *Twice as many D. melanogaster ncRNAs*
  while *less than doubling run-time*
- *Novel banding method* for fast multiple realignment
  implemented in the RNA alignment tool LocARNA
- Verified with *annotations and modEncode RNA-seq*
- REAPR + RNA-seq combined results in interesting
  structural ncRNAs
- *Future challenge:* detect non-locally misaligned RNAs

Conclusions

- *REAPR*: "Realignment for Prediction of ncRNAs"
- *First* genome-wide ncRNA screen
  that *corrects structural misalignments*
- *Twice as many D. melanogaster ncRNAs*
  while *less than doubling run-time*
- *Novel banding method* for fast multiple realignment
  implemented in the RNA alignment tool LocARNA
- Verified with *annotations and modEncode RNA-seq*
- REAPR + RNA-seq combined results in interesting
  structural ncRNAs
- *Future challenge:* detect non-locally misaligned RNAs

Acknowledgements

Bonnie Berger

Michael Yu

Rolf Backofen

German Research Foundation DFG