Benchmarking Statistical Multiple Sequence Alignment

Michael Nute

joint work with Ehsan Saleh and Tandy Warnow

August 17, 2018

Montpelier, France

Statistical Multiple Sequence Alignment (MSA)

- What:
  - MSA method that adheres to a statistical model
- One particular method of interest:

BAli-Phy (Redelings & Suchard, 2005)

  - Models sequence evolution along a tree *including* insertions & deletions (indels).
  - Bayesian method:
    - Uses MCMC sampling (like MrBayes, for example)
    - Generates many alignment-tree pairs

Strongside/Weakside: Bayesian MCMC Methods

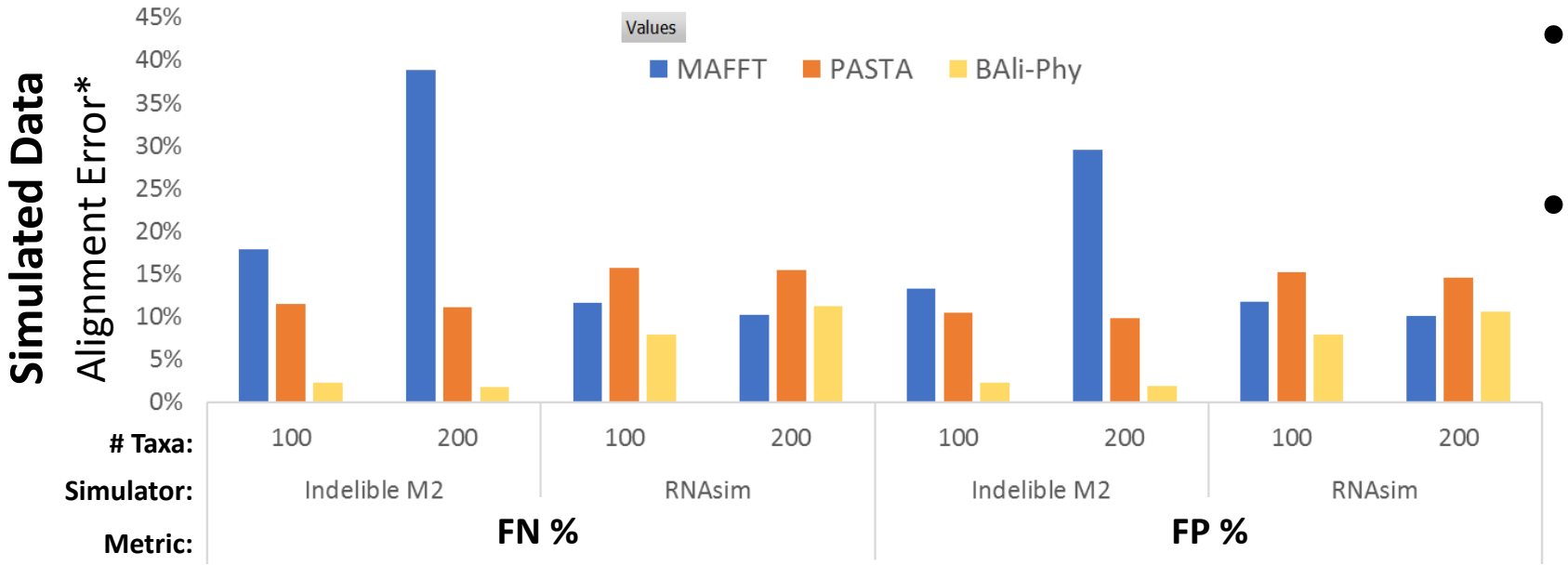
Pro:

- Statistically robust.
- Responds to the data flexibly, according to credibility (i.e. sample size).
- Outputs a *distribution* of parameters.

Con:

- Statistically robust ≠ effective in practice
- Requires a large number of MCMC samples to be useful.
  - More samples = more compute cycles
- No way to definitively say how many samples is “large enough”
  - Potential falsifiability problem

Why the Interest? Preliminary Analysis Shows Promise.

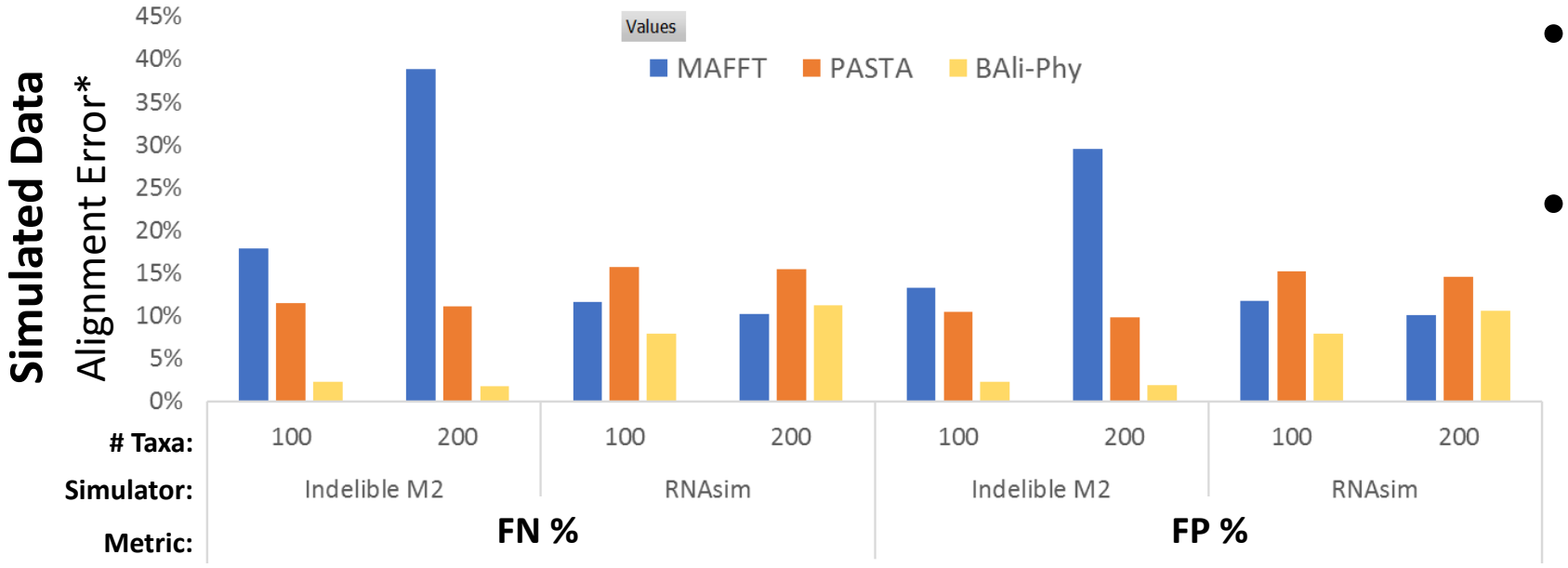


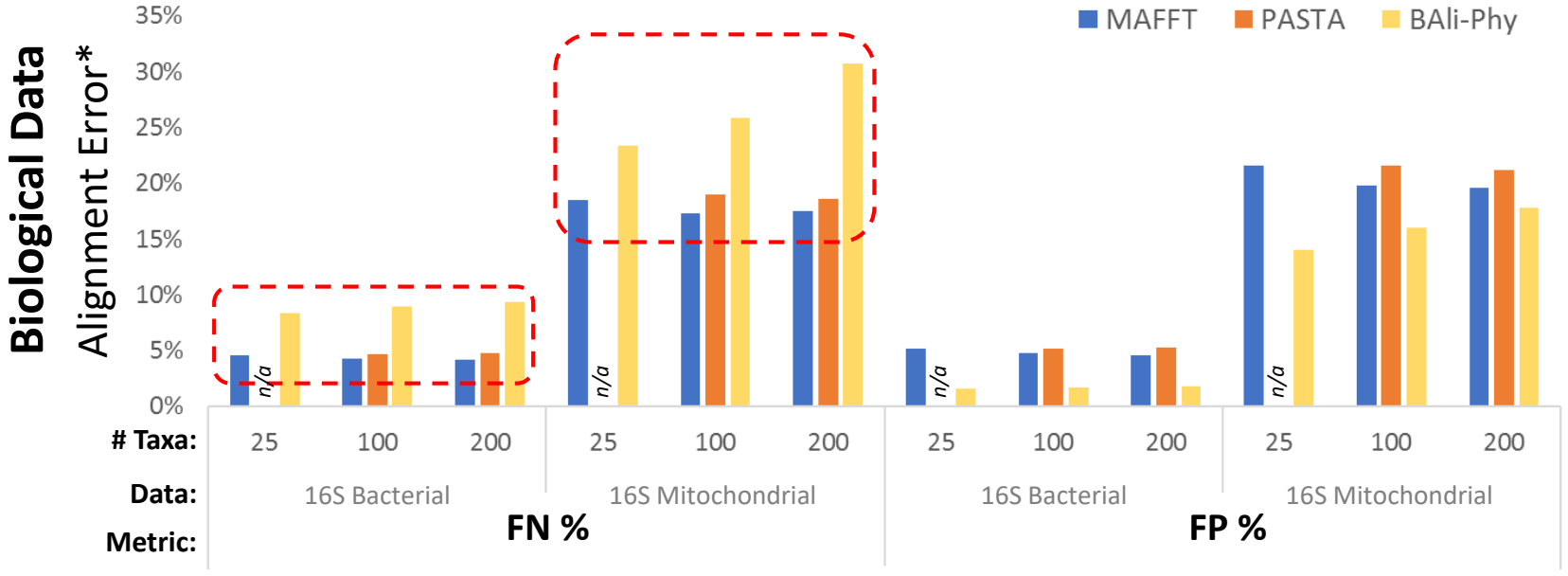
- FN %: False Negative Rate
  - "Missing Homologies"
- FP %: False Positive Rate
  - "Incorrect Homologies"

Simulated Data:
- Indelible (Fletcher & Yang, 2009)
- RNAsim (Guo, et al. 2009)

**Averages over 10 replicates*

Bio vs. Simulated: Note the Differences





- FN %: False Negative Rate
  - "Missing Homologies"
- FP %: False Positive Rate
  - "Incorrect Homologies"

Simulated Data:
- Indelible (Fletcher & Yang, 2009)
- RNAsim (Guo, et al. 2009)

Biological Data:
- 16S alignments from Comparative RNA Website (Cannone, et al. 2002)
- Random Subsets of Bacterial and Mitochondrial seqs (10 replicates)

Limitations of Nucleotide Study

- Biological data not very representative
  - Samples from *two curated alignments* of *closely related genes*
  - Curated by one lab (Gutell)
- Convergence questions linger
  - Maybe BAli-Phy just needed more time?
  - Issue does seem to scale with alignment size.

Better study needed!

Performance Study (Protein Data Only, on bioRxiv)

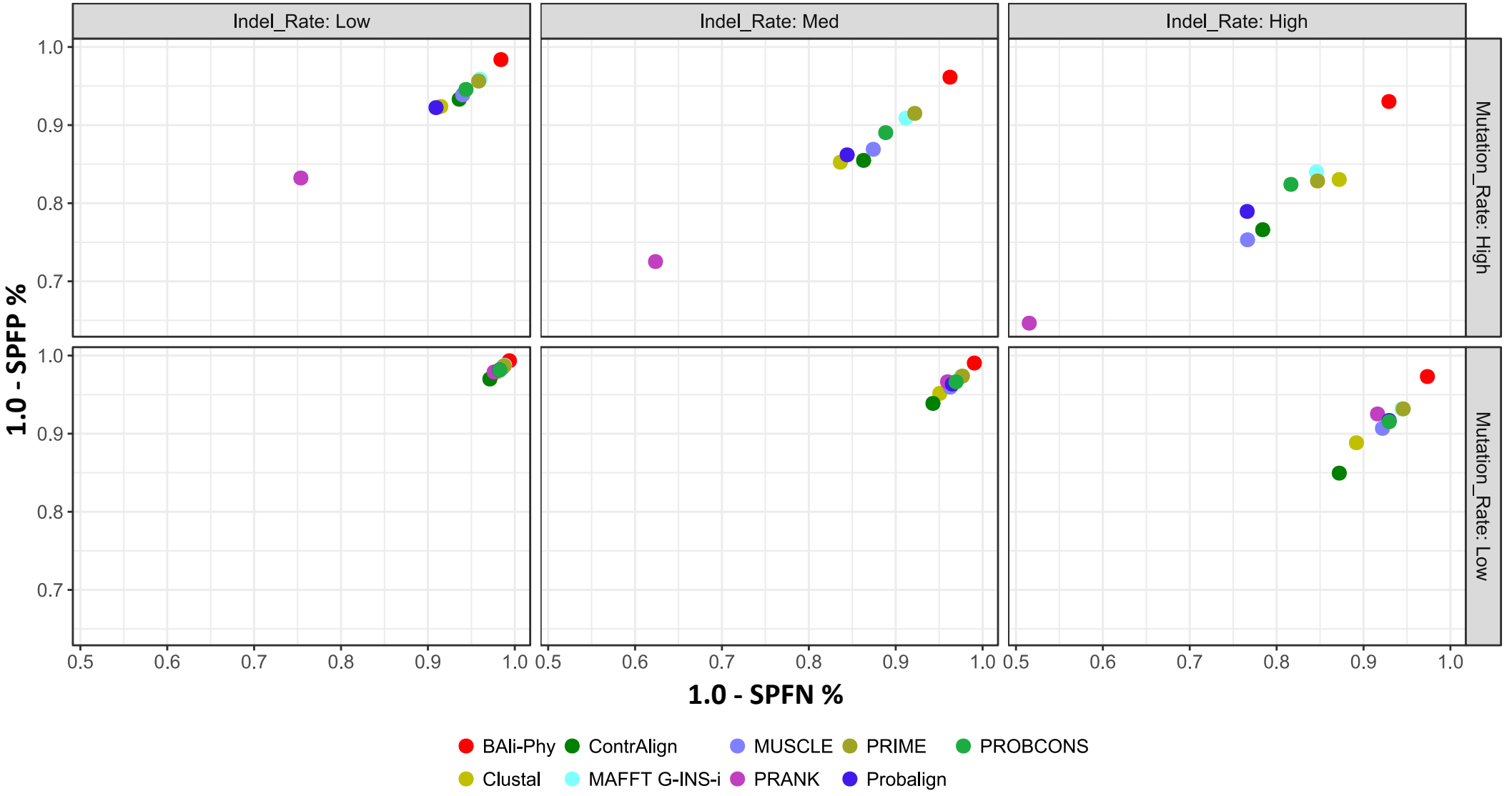
- **Goal**:
  - First extensive study of BAli-Phy accuracy on biological data
  - Eliminate convergence questions with overwhelming force
    - 2 months of CPU time per data set (4-27 sequences each)
  - Compare to leading protein alignment methods.
- **Data**:
  - Simulated: 6 Model Conditions (20 replicates each)
  - Biological: curated structural alignments from 4 separate benchmarks*:
    - BAliBase (Bahr, et al., 2001): 658 alignments
    - Homstrad (Mizuguchi et al., 1998): 231
    - MattBench (Daniels, et al., 2012): 202
    - Sisyphus (Andreeva, et al., 2007): 101

**Excludes some alignments where not all methods completed.*

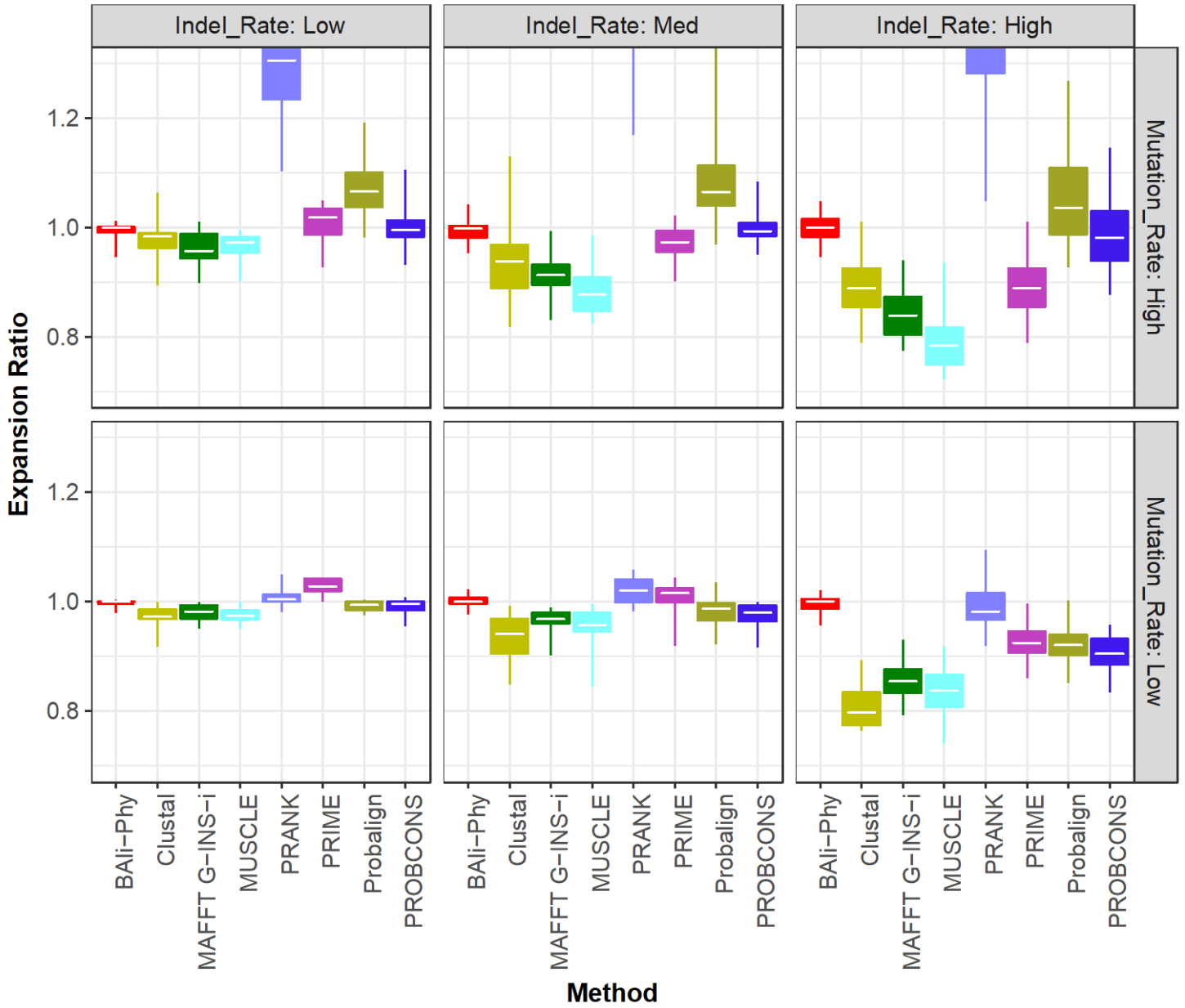
Performance Study (Protein Data Only , on bioRxiv)

- **Evaluation Metrics**:
  - (1 – FP %) (a.k.a. Modeller Score, Precision)
  - (1 – FN %) (a.k.a. SP-Score, Recall)
  - Expansion ratio (estimated alignment length / reference alignment length)

Results on Simulated Data

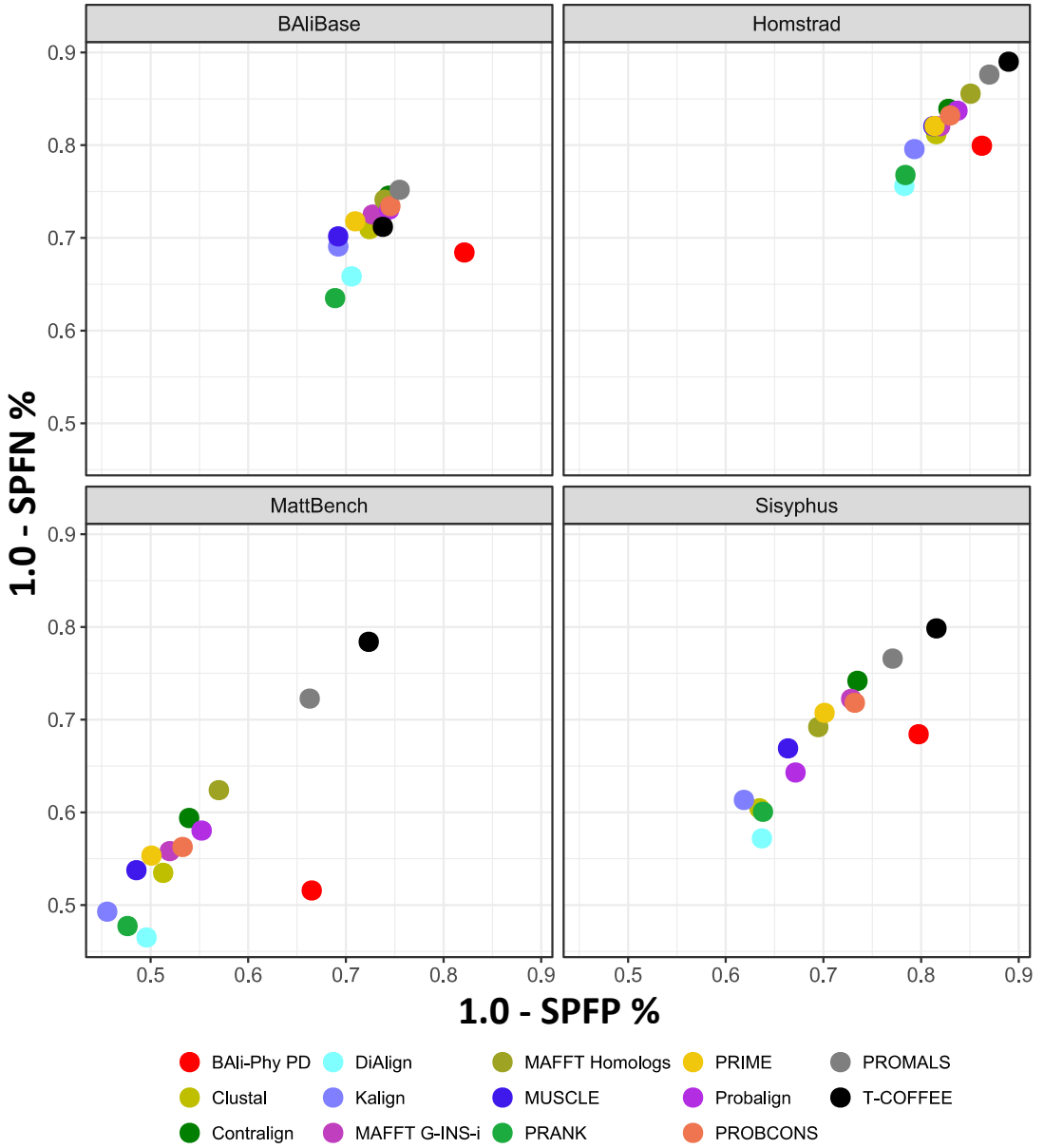


Simulated Data: Expansion Ratios



Results on Biological Data

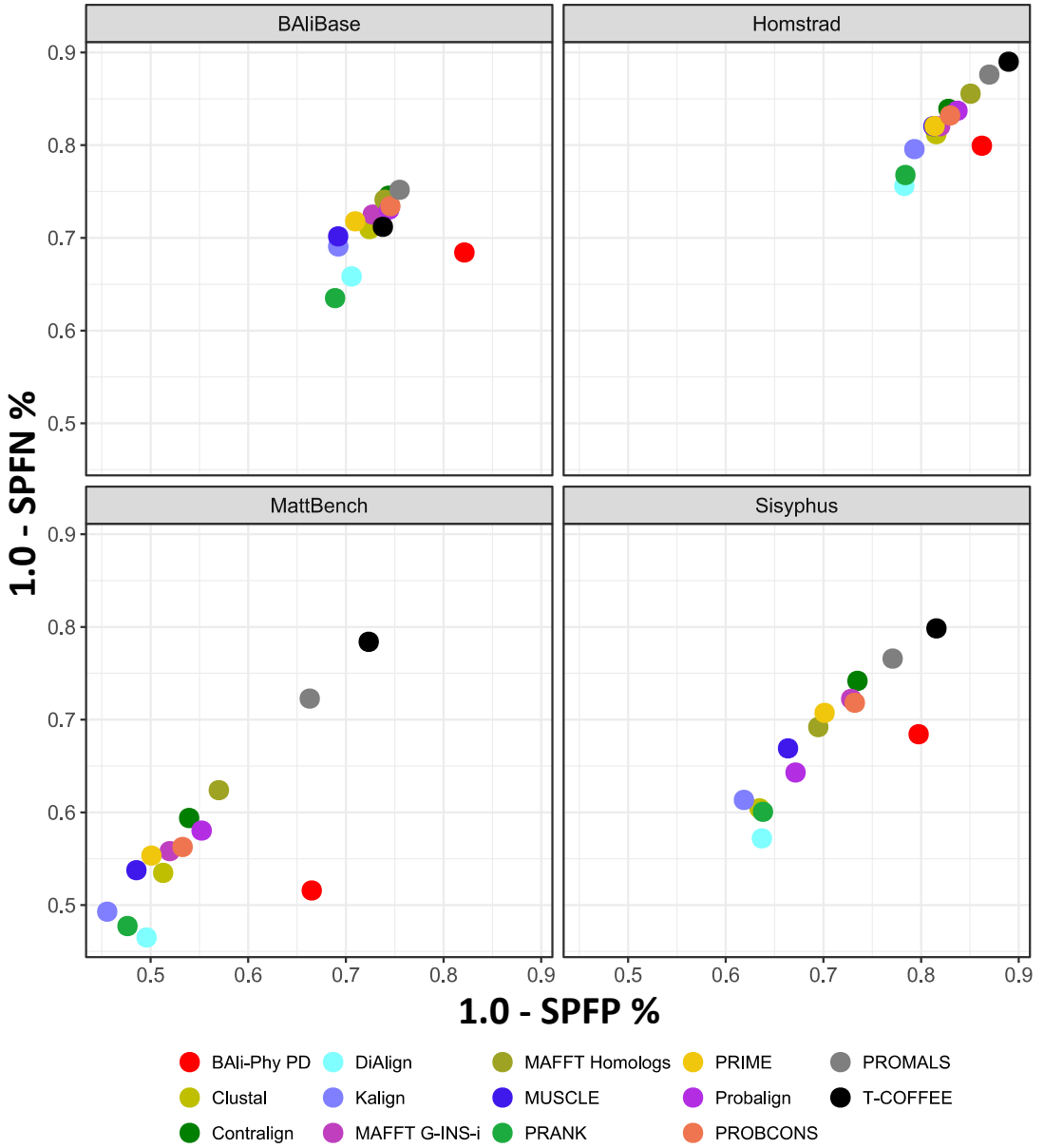
- BAli-Phy still has low SPFP and high SPFN on biological data.
  - Across all 4 benchmarks.
- Methods that recruit homologs do best. I.e.:
  - T-COFFEE (Notredame et al., 2000)
  - PROMALS (Pei & Griffin, 2007)
  - MAFFT Homologs (Katoh et al., 2002)



Results on Biological Data

- BAli-Phy still has low SPFP and high SPFN on biological data.
  - Across all 4 benchmarks.
- Methods that recruit homologs do best. I.e.:
  - T-COFFEE (Notredame et al., 2000)
  - PROMALS (Pei & Griffin, 2007)
  - MAFFT Homologs (Katoh et al., 2002)

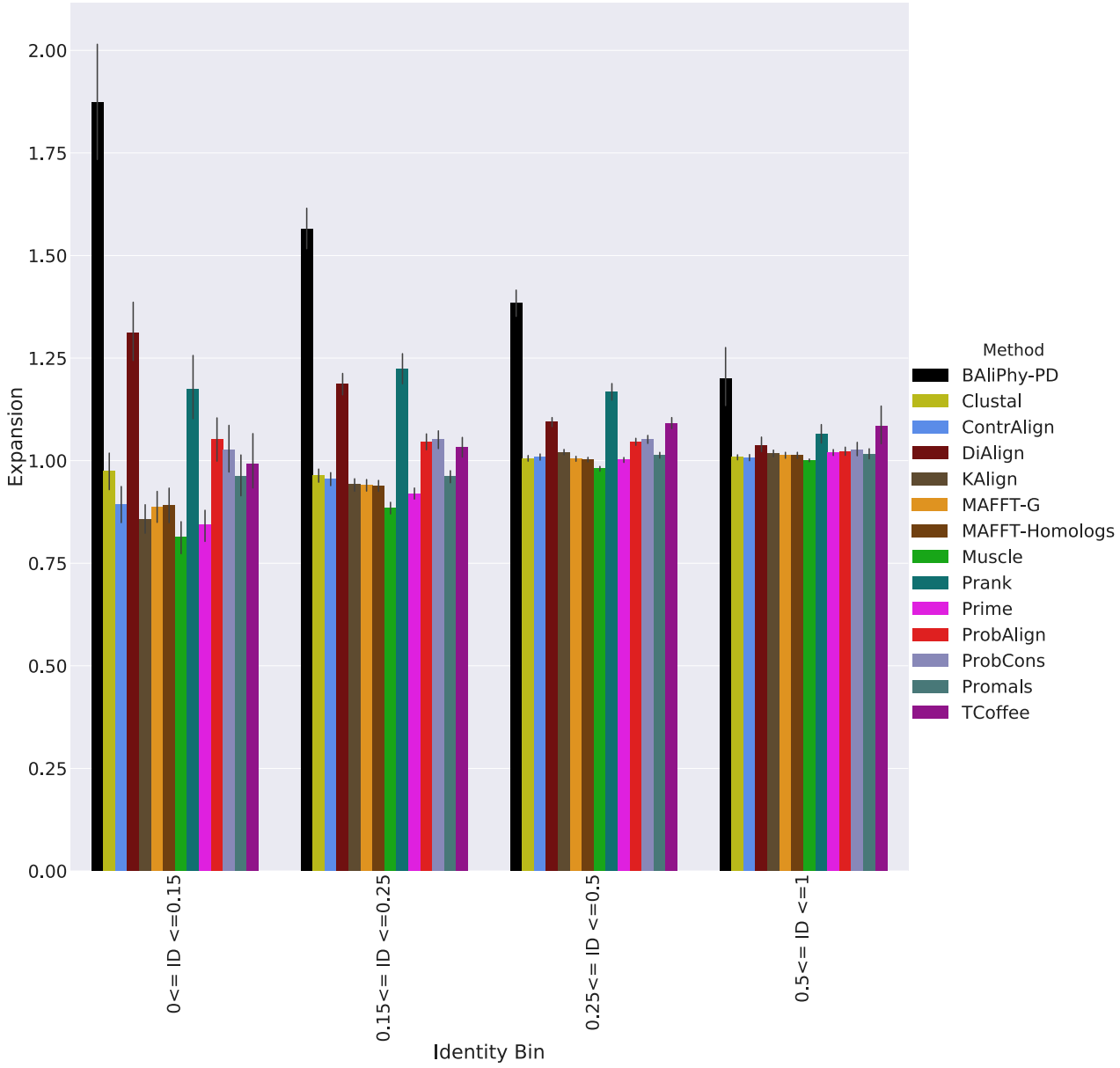
Note the difference from simulated data!



Biological Data: Expansion Ratios

- BAli-Phy *under-aligns* on biological data.
- Problem seems to get worse with low sequence identity.

No indication of this problem on simulated data.



?????

Conclusion:

- BAli-Phy is highly accurate on simulated data
- On biological data, BAli-Phy systematically under-aligns

?????

Conclusion:

- BAli-Phy is highly accurate on simulated data
- On biological data, BAli-Phy systematically under-aligns

Possible Explanations:

- BAli-Phy model is wrong somehow.
- Structural alignment ≠ Evolutionary alignment
- Systematic over-alignment in manual curation process

Conflicting Opinions

> ***…and when you get a Ph.D., you realize that nobody else knows anything either.***
>
> —old joke (as told by Ping Ma)

Conflicting Opinions

> ***…and when you get a Ph.D., you realize that nobody else knows anything either.***
>
> —old joke (as told by Ping Ma)

Reviewer #1 Comment:

Is the proposed explanation "that many of the reference sequence alignments in these benchmark data sets have substantial error" reasonable, given that four different established benchmark databases were used? These benchmarks have been evaluated in (multiple) different studies…

Conflicting Opinions

…and when you get a Ph.D., you realize that nobody else knows anything either.

—old joke (as told by Ping Ma)

Reviewer #1 Comment:

Is the proposed explanation "that many of the reference sequence alignments in these benchmark data sets have substantial error" reasonable, given that four different established benchmark databases were used? These benchmarks have been evaluated in (multiple) different studies…

Reviewer #2 Comment:

The conclusions of this paper do not align with data presented. In the analyses of simulated data…BaliPhy consistently correctly aligns across indels (expansion ratios close to 1). When they find similar results in biological data - that most alignment software generates more compressed alignments than does BaliPhy - they draw the opposite conclusion, (that) BaliPhy has underaligned these data sets and is therefore 'less accurate'….

Despite common human tendencies to see patterns where none exist, the possibility that these reference data sets have been 'overaligned'…is discarded as 'not seeming very likely'.

Acknowledgements

Collaborators

Tandy Warnow
(UIUC)

Ehsan Saleh
(UIUC)

Funding

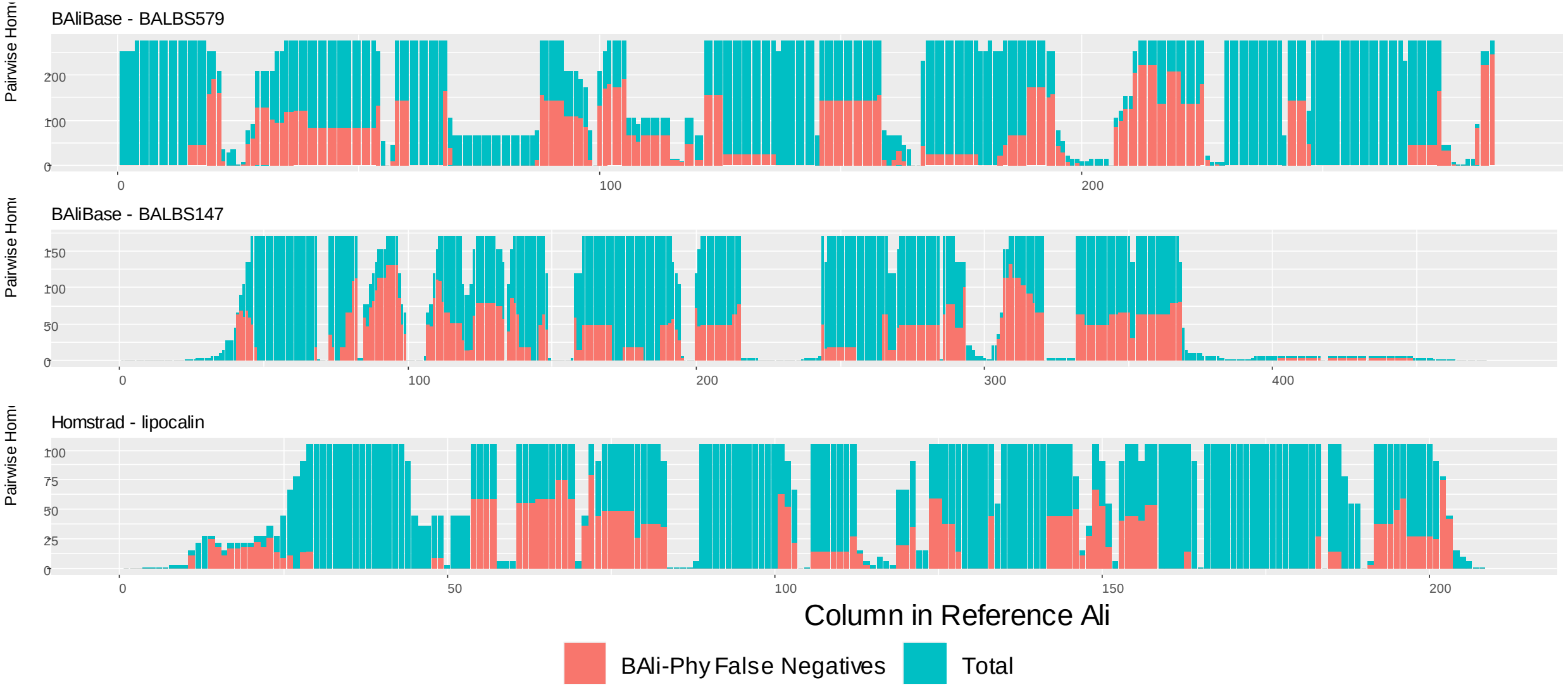
- This work was funded by NSF grant III:AF:1513629 (PI: Tandy Warnow) and by a fellowship from the CompGen Initiative at UIUC to M.N.

Blue Waters

- This research is part of the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (awards OCI-0725070 and ACI-1238993) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications.

Appendix: Patterns of BAli-Phy False Negatives

Three examples showing where in the reference alignment BAli-Phy is getting it wrong:



Biological Data: Grouped by Degree of Difficulty

Results grouped by Percent Identity (i.e., Difficulty)