A Comparative Study of SVDquartets and Other Coalescent-Based Species Tree Estimation Methods

Jed Chou, Ashu Gupta, Shashank Yaduvanshi, Ruth Davidson, Mike Nute, Siavash Mirarab, and Tandy Warnow

University of Illinois Urbana-Champaign

October 6, 2015

Support

- JC was supported by the Mathematics Department at the University of Illinois at Urbana-Champaign and NSF grant DMS-1345032.
- RD was supported by NSF grant DMS-1401591.
- AG and SY were supported by the Computer Science Department at the University of Illinois at Urbana-Champaign.
- SM was supported by a graduate fellowship from the Howard Hughes Medical Institute (HHMI).
- MN and TW were supported by the National Science Foundation grant DBI-1461364.

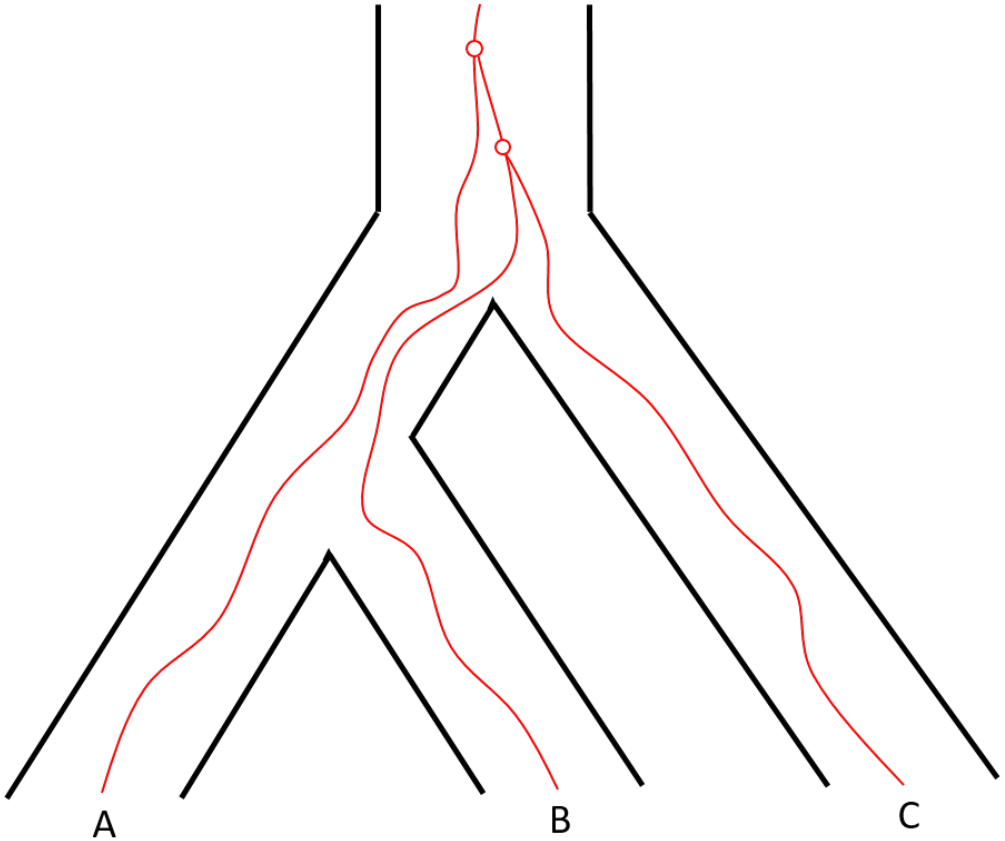
The Problem

Estimate a species tree from multi-locus sequence data.

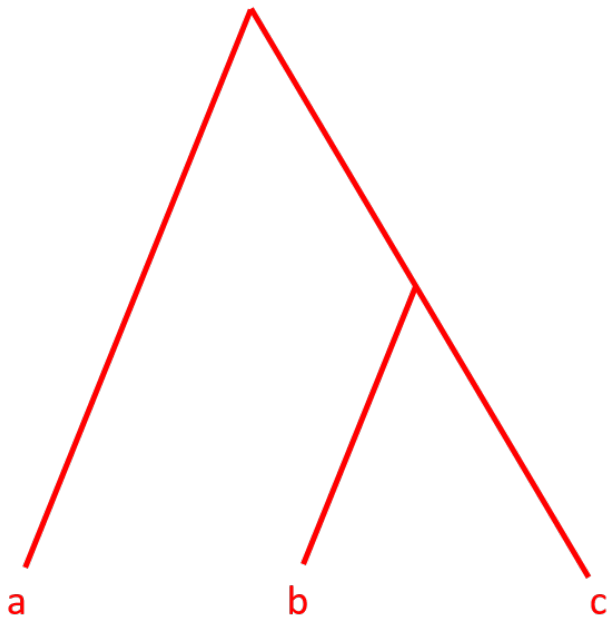
Challenges

- Large datasets
- Many sources of gene tree-species tree discord
  - **incomplete lineage sorting (ILS)**
  - horizontal gene transfer
  - gene duplication and loss
  - hybridization
  - recombination

The Multi-Species Coalescent models ILS



Gene Tree Topology



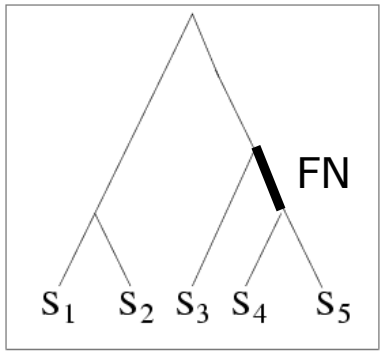
Species tree topology doesn't match gene tree topology.

Various ways to infer a species tree

- **Concatenated analysis with maximum likelihood (CA-ML)**: concatenate gene alignments and estimate species tree from supermatrix.
- **Summary methods**: infer gene trees from alignments and combine gene trees into species tree, e.g. **ASTRAL**, **MP-EST**, and **BUCKy**
- **Coestimation**: simultaneously estimate gene trees and species tree, e.g. ***BEAST** and **BEST**

Species tree estimation error

Quantifying Error



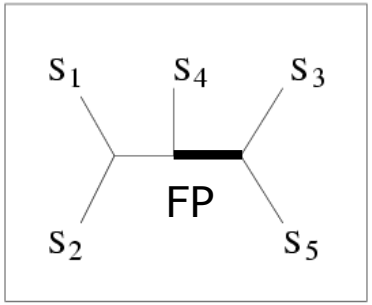
TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC
S_5	ACCAGACCGGA

DNA SEQUENCES

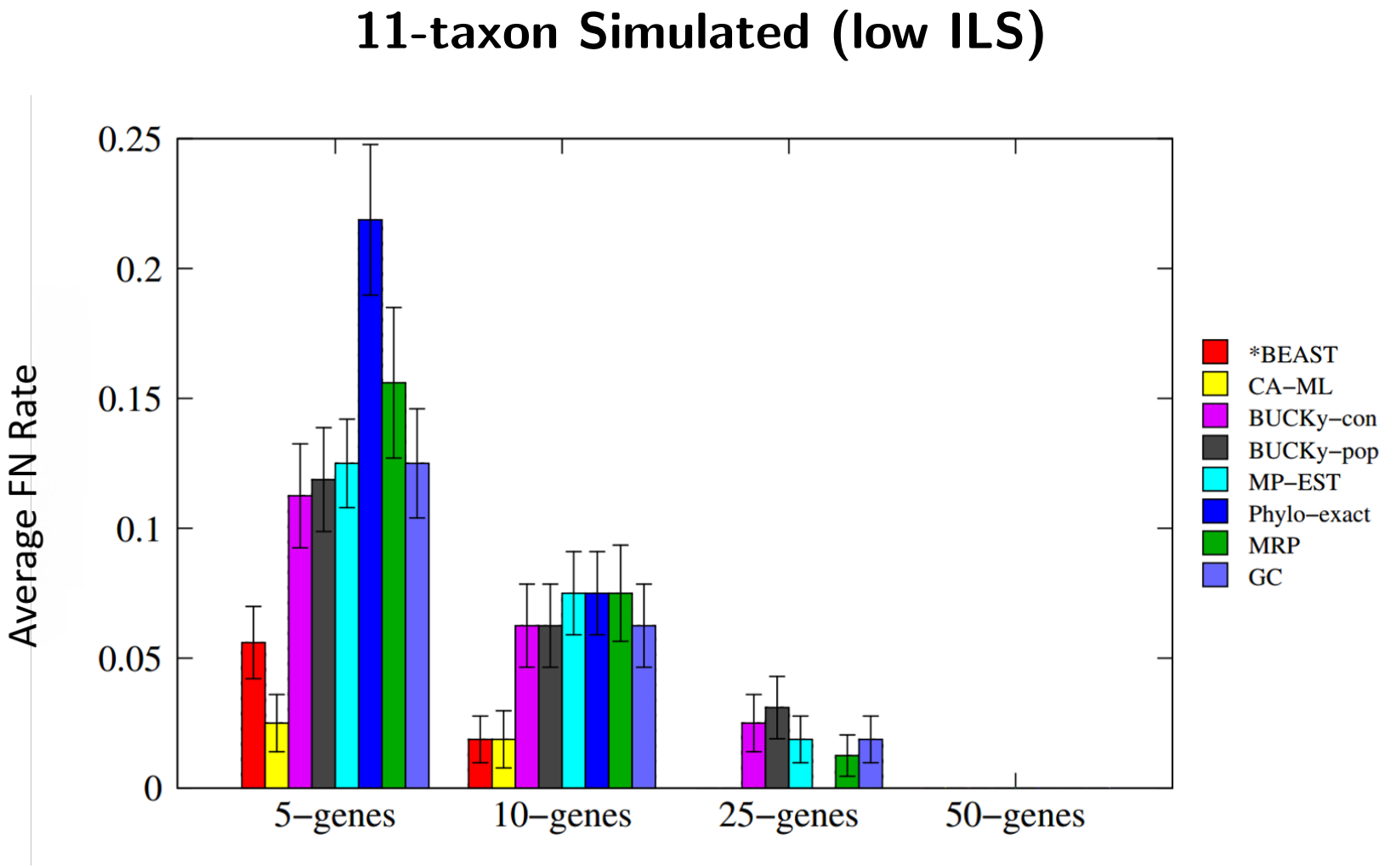
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

50% error rate



INFERRED TREE

Main advantage of CA-ML: empirical performance



Bayzid and Warnow 2013
Dataset from Chung and Ane 2011

Advantages of summary methods

- **many are statistically consistent under MSC**, whereas CA-ML can be statistically inconsistent under MSC
- summary methods more accurate than CA-ML on some datasets, less accurate on others
- summary methods faster than CA-ML

Recombination poses problems for summary methods

- recombination violates assumptions of MSC
- statistical consistency not guaranteed unless genes are recombination-free
- recombination-free genes can be very short
- short genes lead to high gene tree estimation error

Recombination a problem for summary methods

"...the **long c-genes that are required for accurate reconstruction of species trees using shortcut coalescence methods do not exist and are a delusion.** *Coalescence approaches based on SNPs that are widely spaced in the genome avoid problems with the recombination ratchet and merit further pursuit in both empirical and systematic research and simulations."*

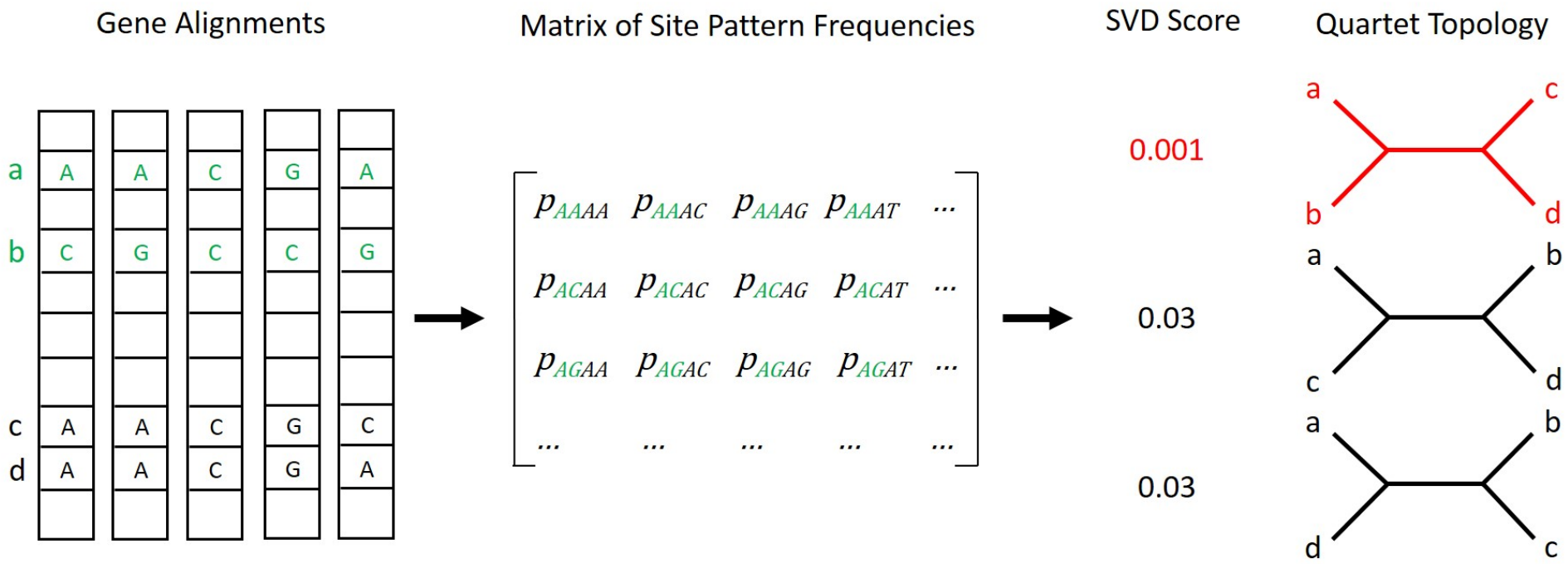
Springer and Gatesy, Molecular Phylogenetics and Evolution 2015

A new method for coalescent-based species tree estimation

Single-site methods: infer a species tree from site pattern distributions among gene alignments, e.g. **SVDquartets**

- skips gene tree estimation step
- avoids gene tree estimation error
- can be statistically consistent under the MSC

SVDquartets infers a tree topology on every quartet



Motivation for this study

- SVDquartets showed good ability to infer correct quartet topologies in initial study
- SVDquartets untested on larger simulated datasets, or against CA-ML and summary methods
- first comparison of summary methods to CA-ML on extremely short sequences

Can SVDquartets outperform summary methods and CA-ML on extremely short genes?

Species tree estimation methods used

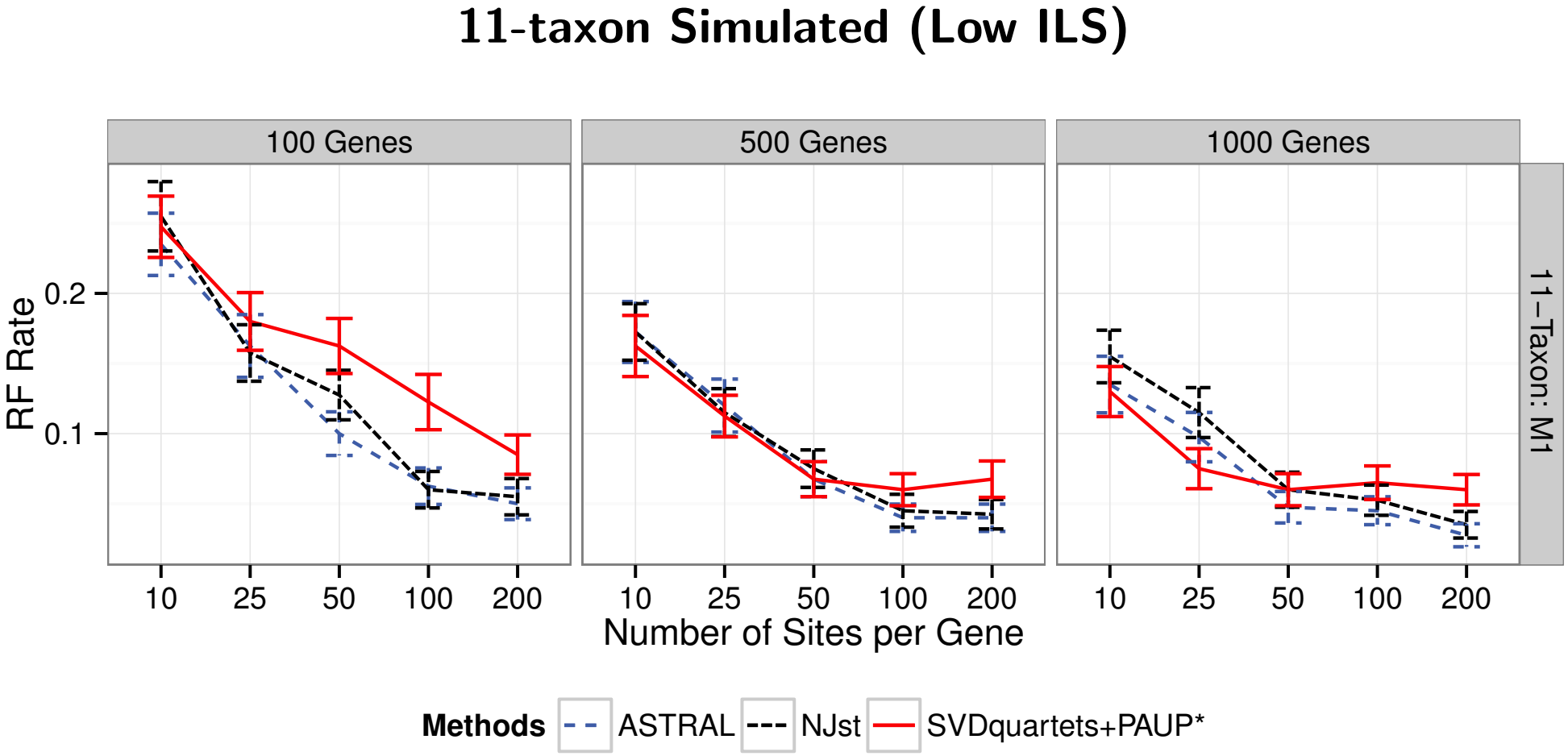
- **SVDquartets+PAUP***
  - SVDquartets to infer set of quartet trees
  - PAUP* heuristic to combine quartet trees into species tree
  - assumes number of changes proportional to time, i.e. molecular clock
- **ASTRAL-2 and NJst**
  - gene trees estimated with FastTree-2
  - statistically consistent under MSC
  - doesn't assume molecular clock
- **CA-ML**

Stats for simulated datasets

Dataset	AD	# Sites	# Genes	Clock	# Reps
11-taxon M1	16%	10-200	100-1000	No	50
11-taxon M2	38%	10-200	100-1000	No	50
11-taxon M3	66%	10-200	100-1000	No	50
11-taxon M4	85%	10-200	100-1000	No	50
15-taxon	82%	10-200	100-1000	Yes	10
37-taxon Mammalian	18%	10-200	50-200	No	20

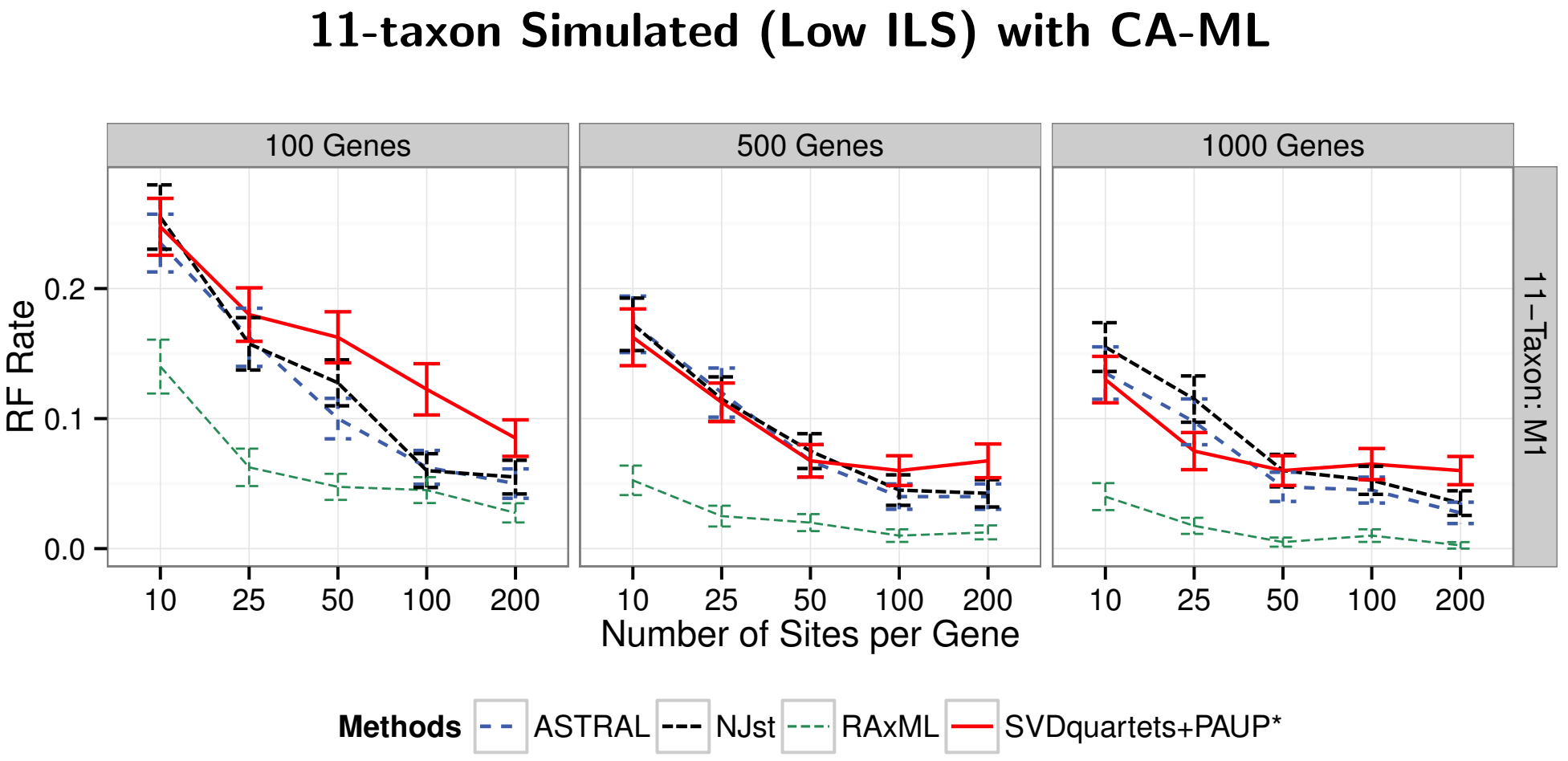
AD = average distance (Robinson-Foulds) between true gene trees and true species trees

SVDquartets competitive on shortest gene lengths



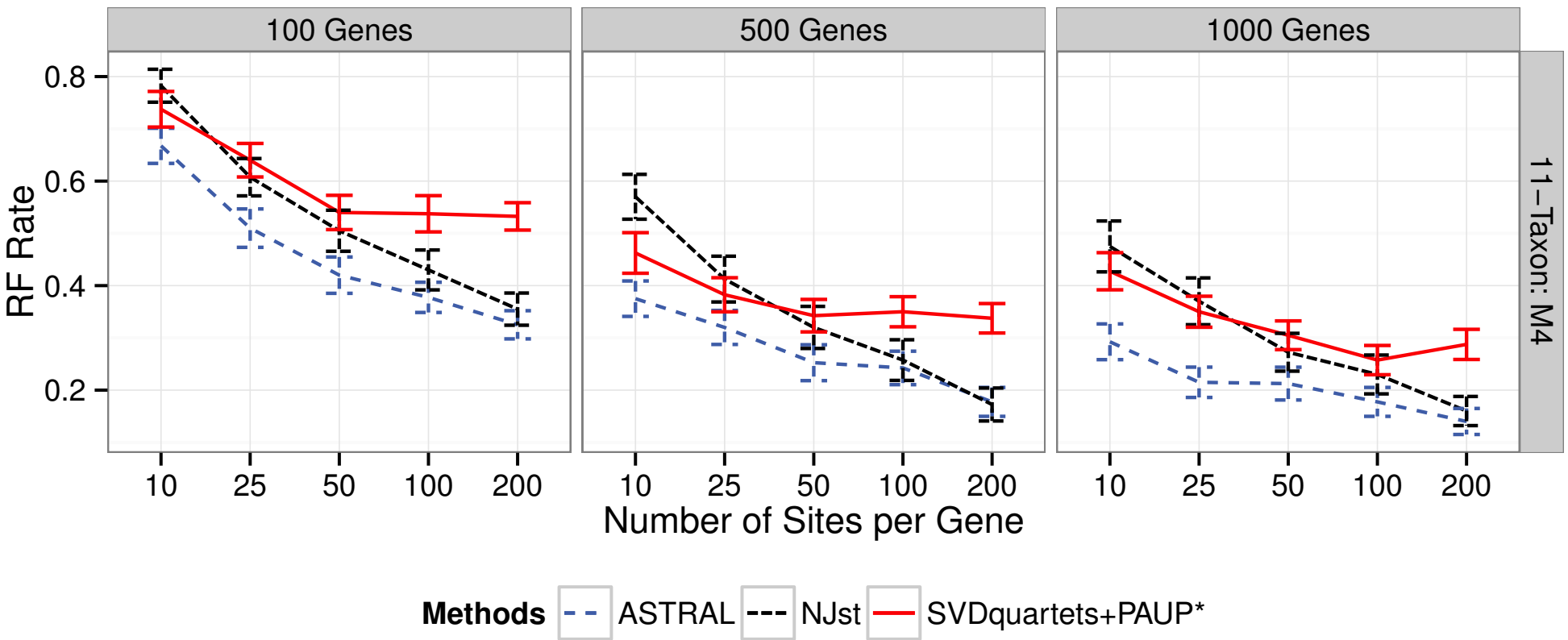
- no molecular clock, AD = 16%, 50 replicates
- *summary methods improve faster than SVDquartets as gene length increases*

CA-ML best under low ILS



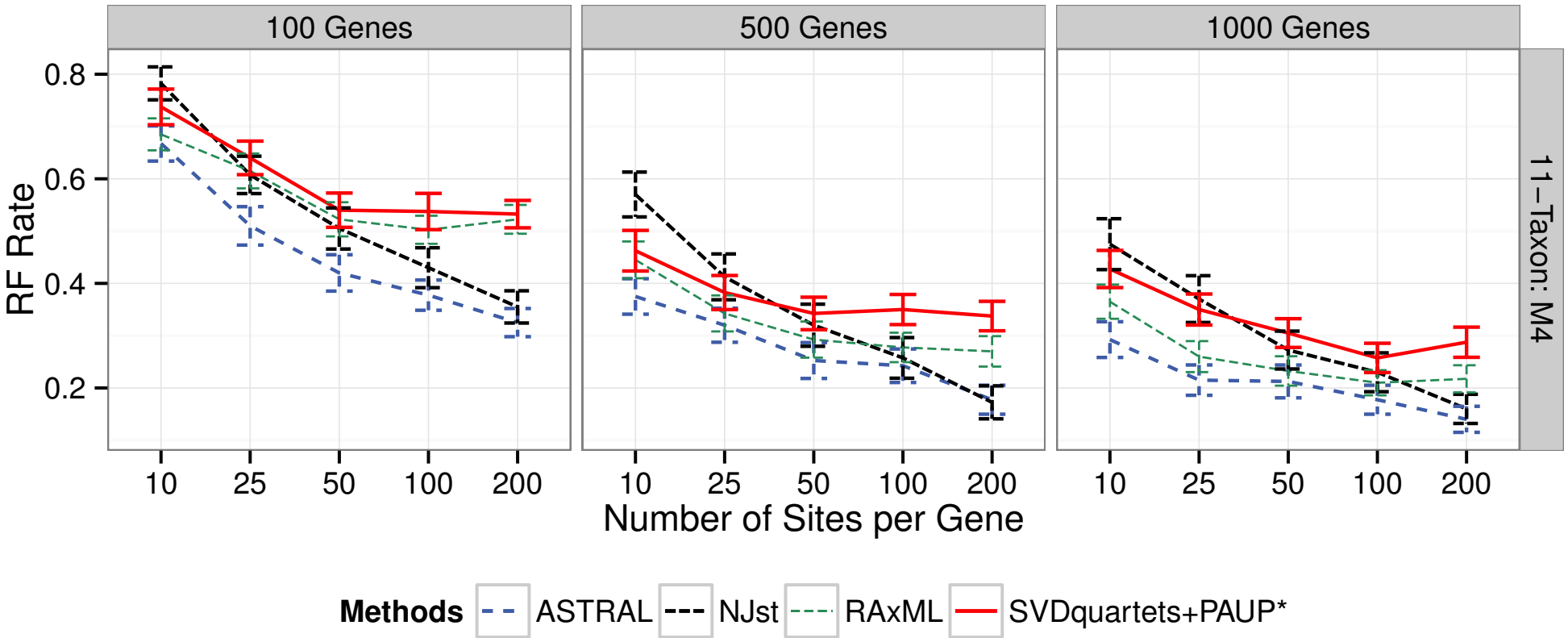
- no molecular clock, AD = 16%, 50 replicates
- *CA-ML more robust to short sequence lengths than coalescent-based methods*

ASTRAL has best performance under very high ILS



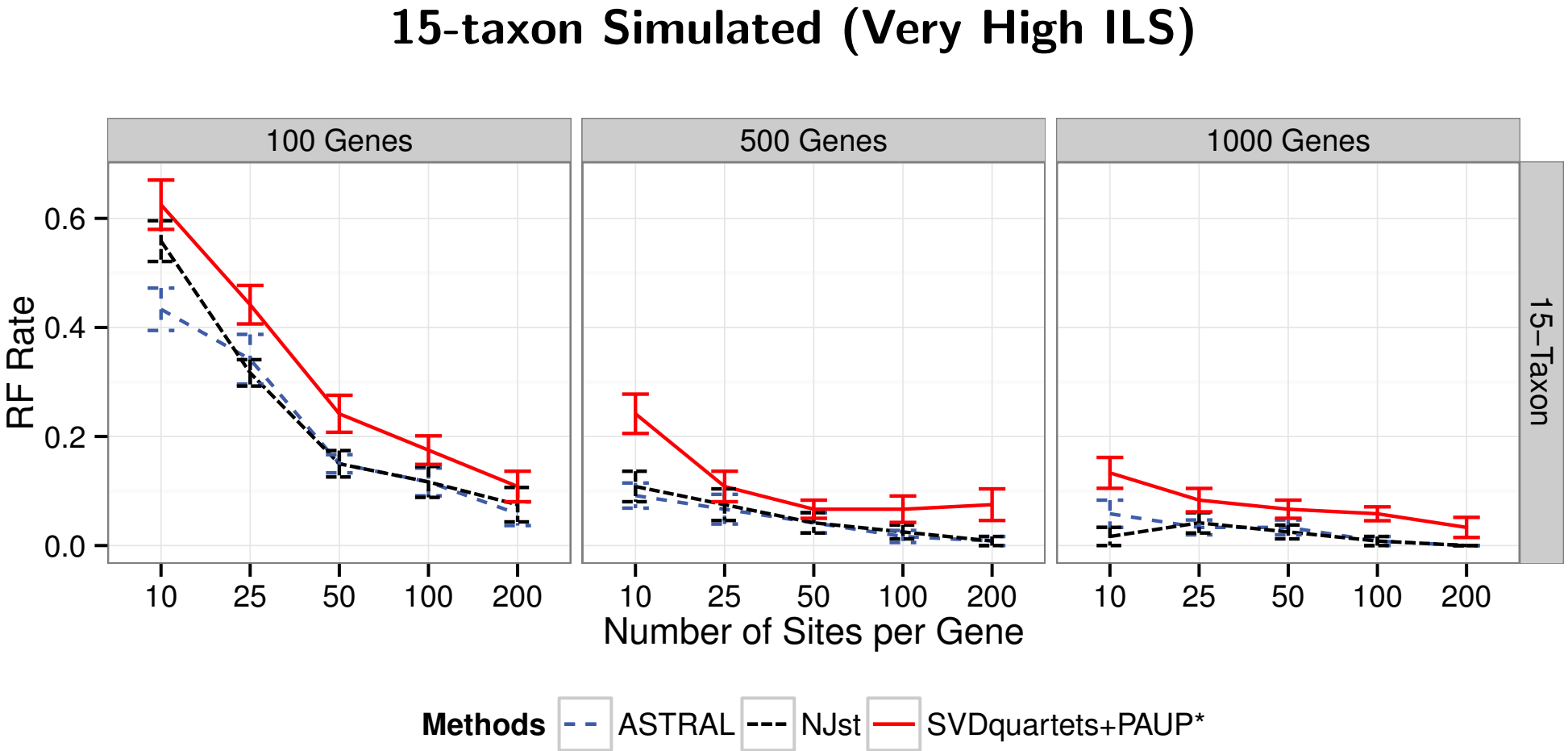
- no molecular clock, AD = 85%, 50 replicates
- *SVDquartets better than NJst, worse than ASTRAL on shortest genes*

ASTRAL better than CA-ML even on 10 bp genes



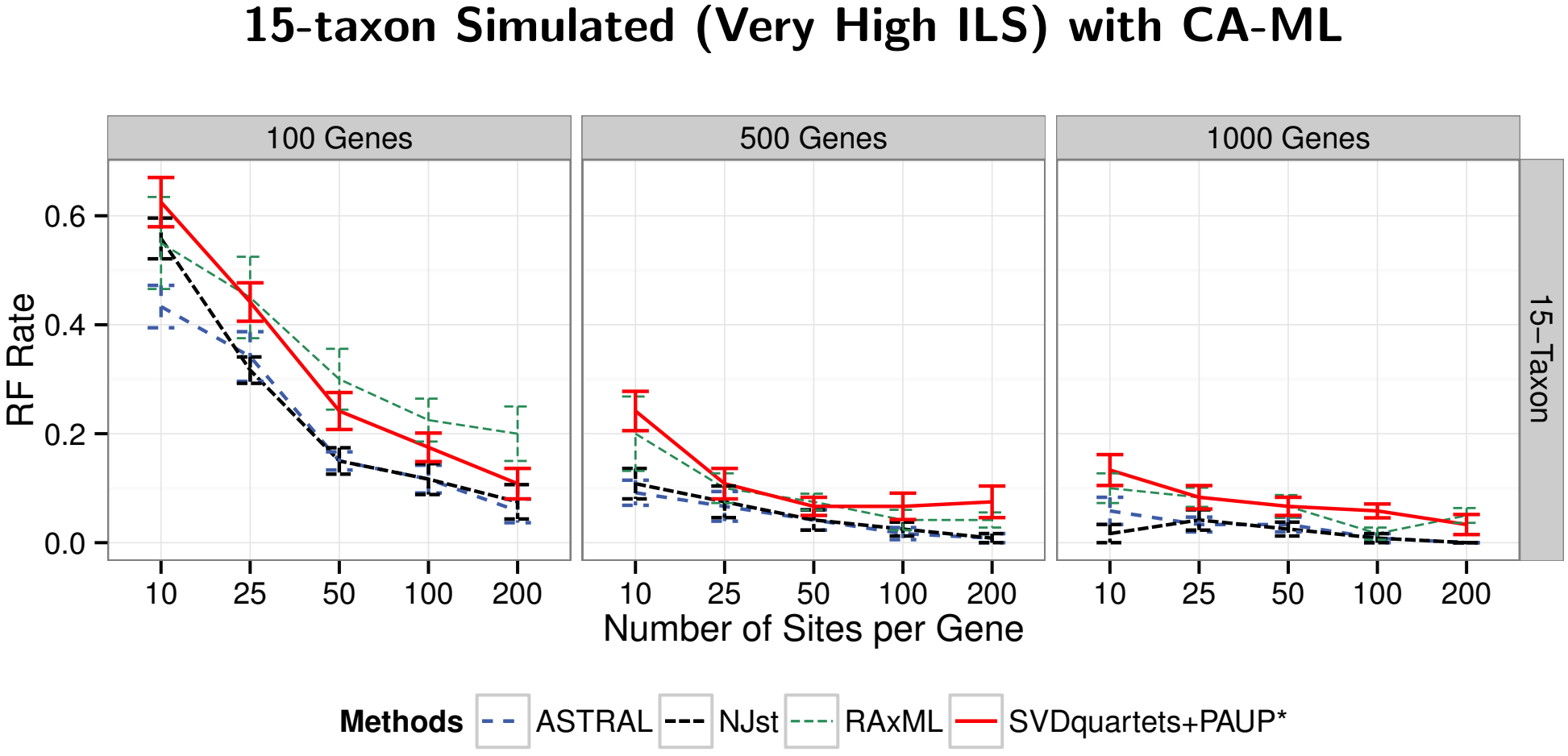
- no molecular clock, AD = 85% (very high ILS), 50 replicates
- *CA-ML better than NJst, SVDquartets on shortest genes*

All coalescent-based methods improve with clock



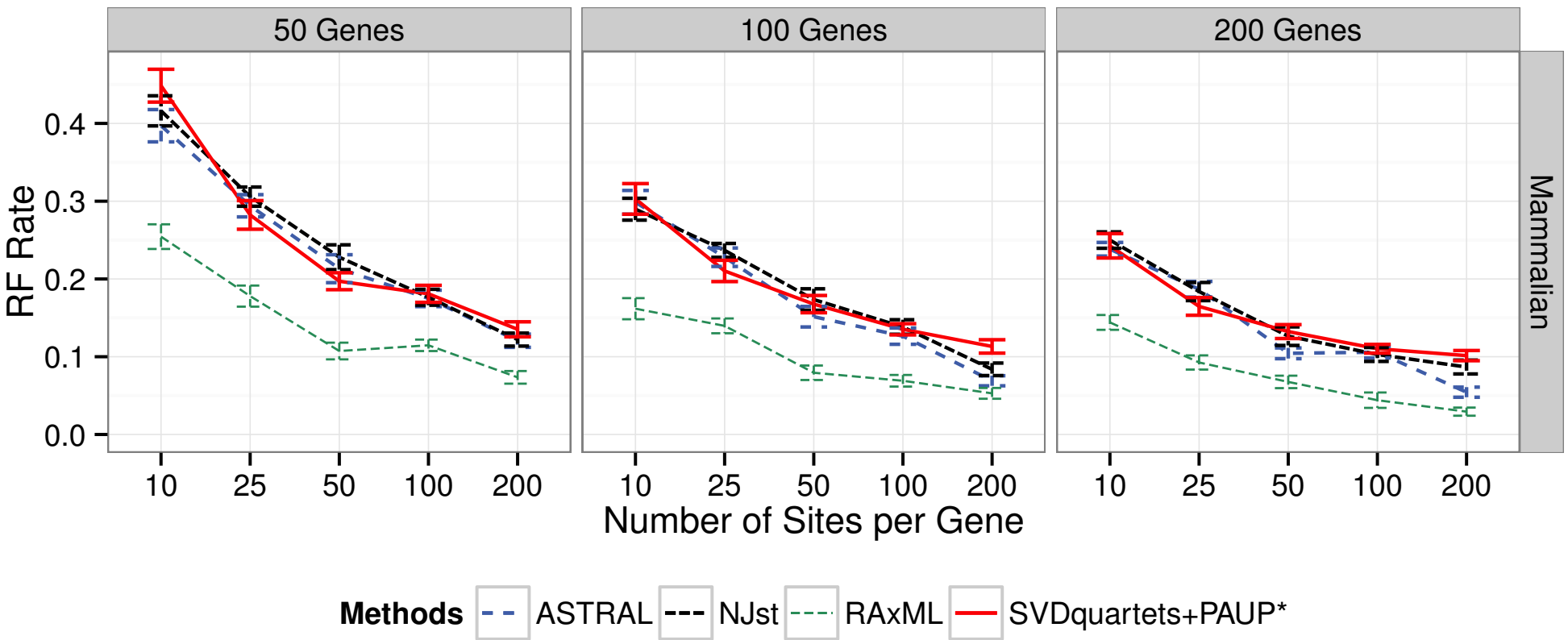
- **molecular clock**, AD = 82%, 10 replicates
- *only SVDquartets requires clock, yet summary methods more accurate at all gene lengths*

Summary methods most accurate under very high ILS



- **molecular clock**, AD = 82%, 10 replicates
- *CA-ML better than SVDquartets, worse than summary methods on shortest genes*

SVDquartets not superior under low ILS



- no molecular clock, AD = 18% (low ILS), 20 replicates
- *SVDquartets more accurate than summary methods on 25 bp genes*

Trends that hold for all methods

- increasing number of genes improves accuracy
- increasing gene lengths improves accuracy
- having molecular clock improves accuracy (though observed on only one dataset)
- increasing ILS decreases accuracy

Comparisons between methods

- under some conditions (e.g., mammalian low ILS), coalescent-based methods all have very similar accuracy
- under other conditions, the best coalescent-based method tends to be ASTRAL-2, followed closely by NJst; SVDquartets usually least accurate
- SVDquartets sometimes competitive with ASTRAL-2 under very low ILS conditions
- concatenation most accurate of all methods under low ILS

Conclusions

- leading summary methods should not be replaced by current implementation of SVDquartets
- on datasets with moderate to high ILS, use of leading summary methods recommended, even when genes are extremely short
- concatenation preferred over leading summary methods and SVDquartets on datasets with low ILS

What next?

- compare to coestimation methods (e.g. *BEAST, BEST) and fully-partitioned CA-ML, which may outperform methods in current study
- improve accuracy and theoretical guarantees of SVDquartets with different quartet amalgamation techniques

Thanks!