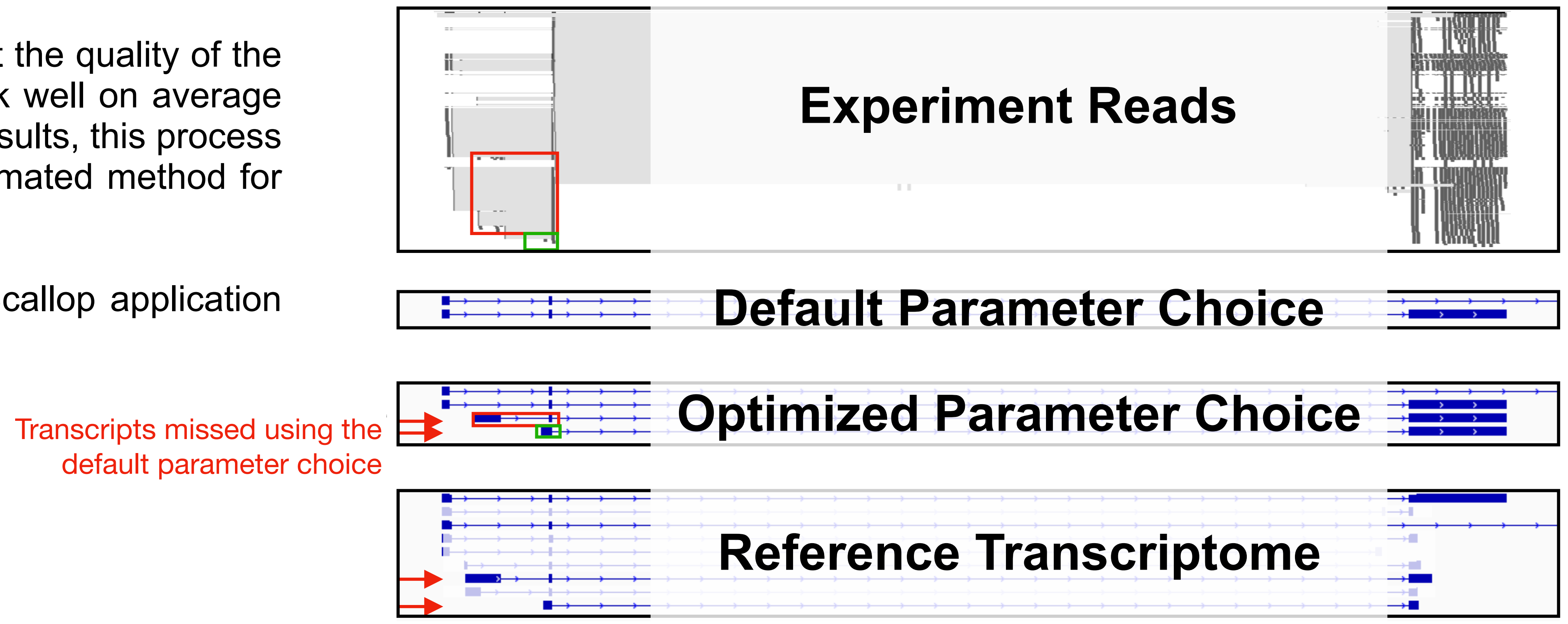
Automatically eliminating errors induced by suboptimal parameter choices in transcript assembly

Dan DeBlasio and Carl Kingsford
Computational Biology Department, Carnegie Mellon University

As the tools used for genomic analysis become more sophisticated they become reliant on user tunable parameters that can greatly impact the quality of the produced results. Most users rely on the default parameter choice defined by an application's designer. This choice was intended to work well on average across all inputs, but the most interesting cases are often not "average". While manually adjusting the parameter settings can improve the results, this process is not trivial and is not guaranteed to generalize from one input to another. Therefore, for any high throughput task we must find an automated method for parameter adjustment in order to eliminate parameter choice as a source of error.

The figure on the right shows the impact of choosing a non-ideal parameter setting for reference based transcript assembly using the Scallop application (Shao and Kingsford, 2017).

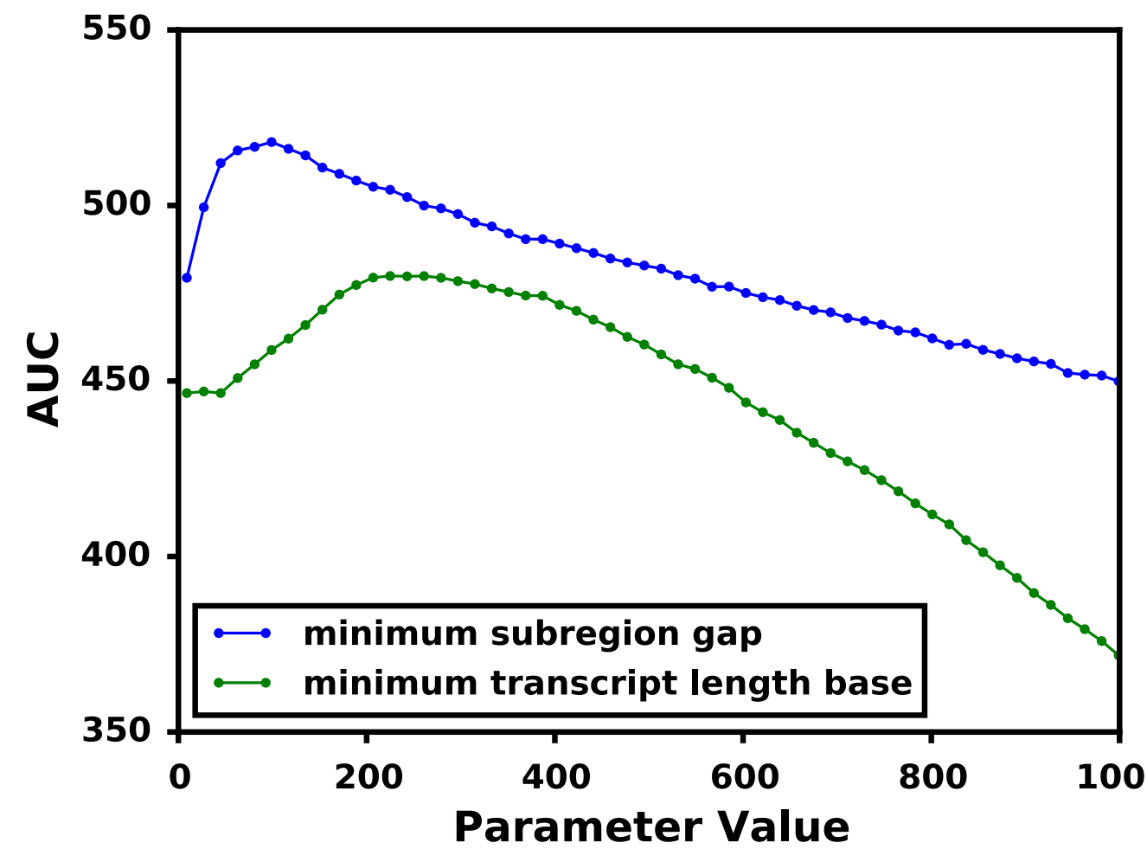

Experiment Reads
Default Parameter Choice
Optimized Parameter Choice
Reference Transcriptome
Transcripts missed using the default parameter choice

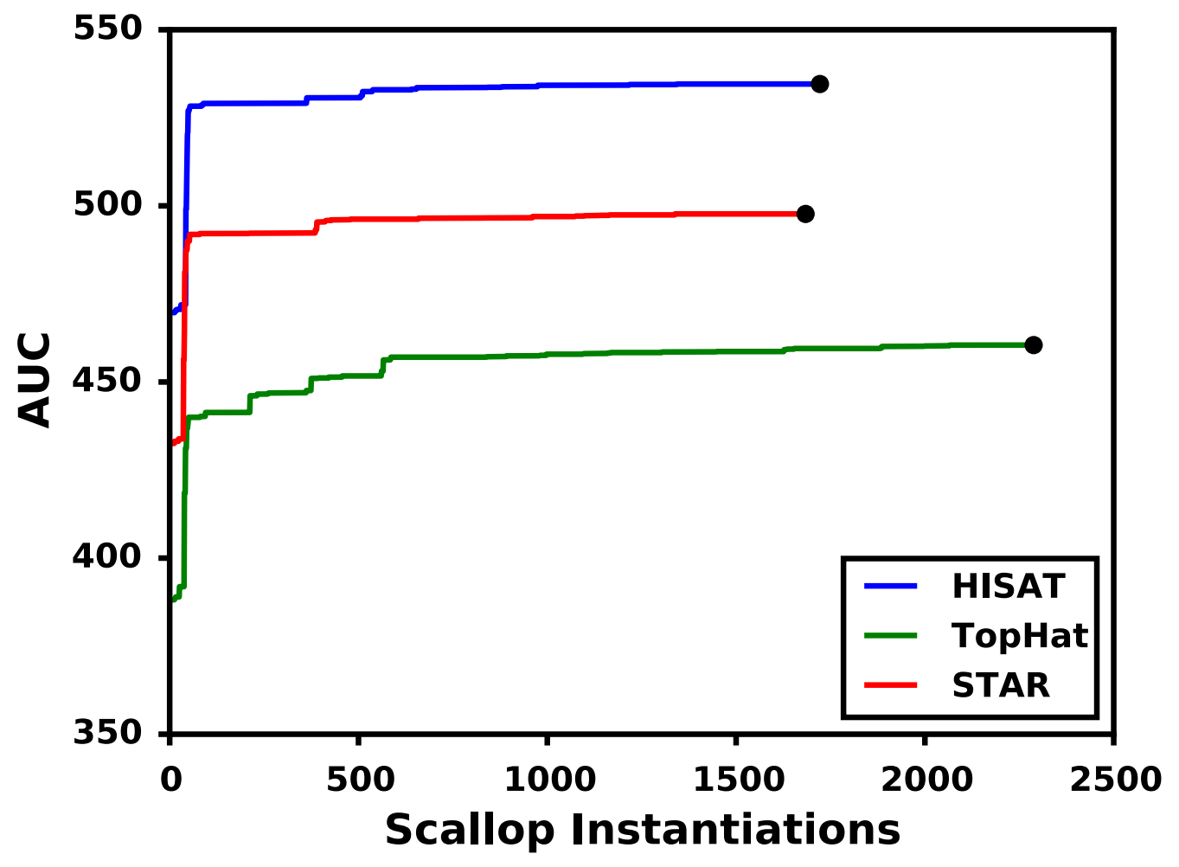
Coordinate Ascent

Parameters contain one maxima

Iterative optimization can be used to find optimal parameters by starting at the default and moving towards better AUC.

The figure show the impact on AUC when changing two parameters independently, leaving all others at the default.





Coordinate ascent increases AUC, but it's slow

Can take over >22 days to converge for a single input in some cases. The figure shows the trajectory for three instances.

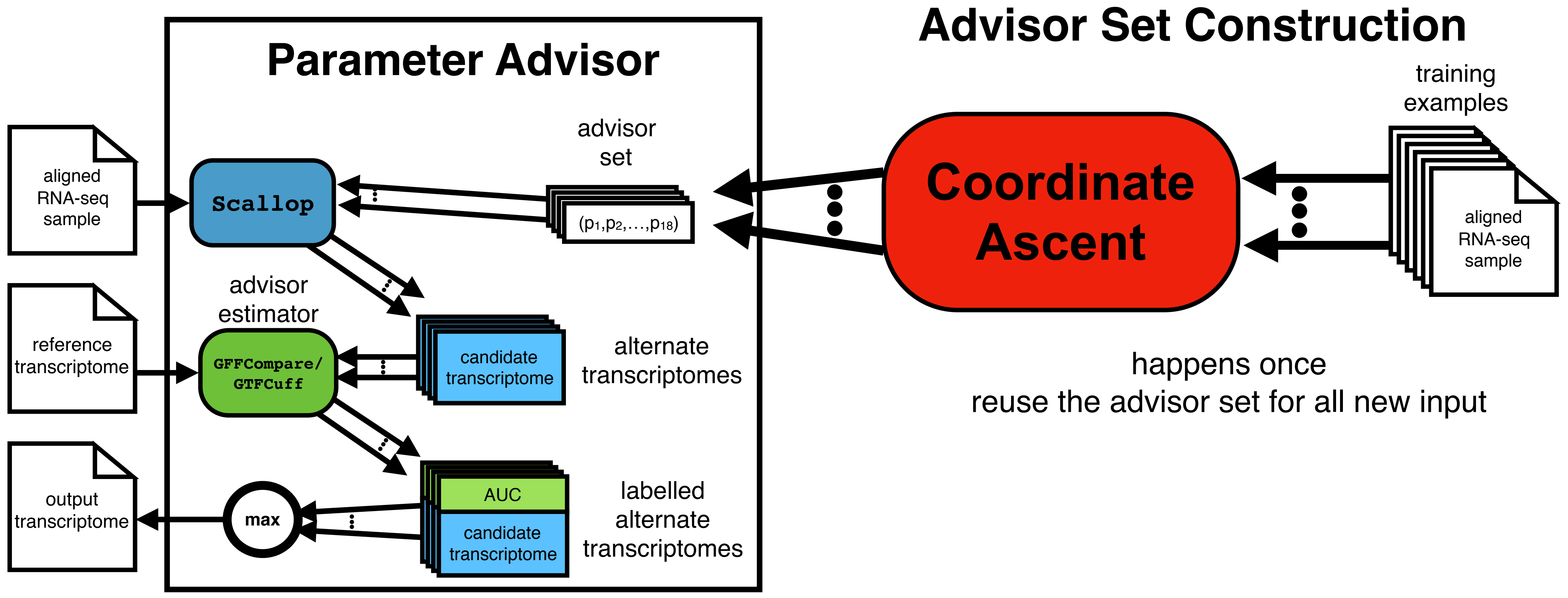
Parameter Advising

Advising is fast, but its quality relies on an advisor set

Parameter advising (DeBlasio and Kececioglu, 2017) chooses parameter values by selecting from a set of alternate assemblies constructed using parameter choices from an advisor set.

Coordinate Ascent finds Advisor Sets with a one time cost

For transcript assembly the advisor set can't be found by exhaustive search because there are too many tunable parameters. But, coordinate ascent can be used to pre-compute these sets.



happens once
reuse the advisor set for all new input

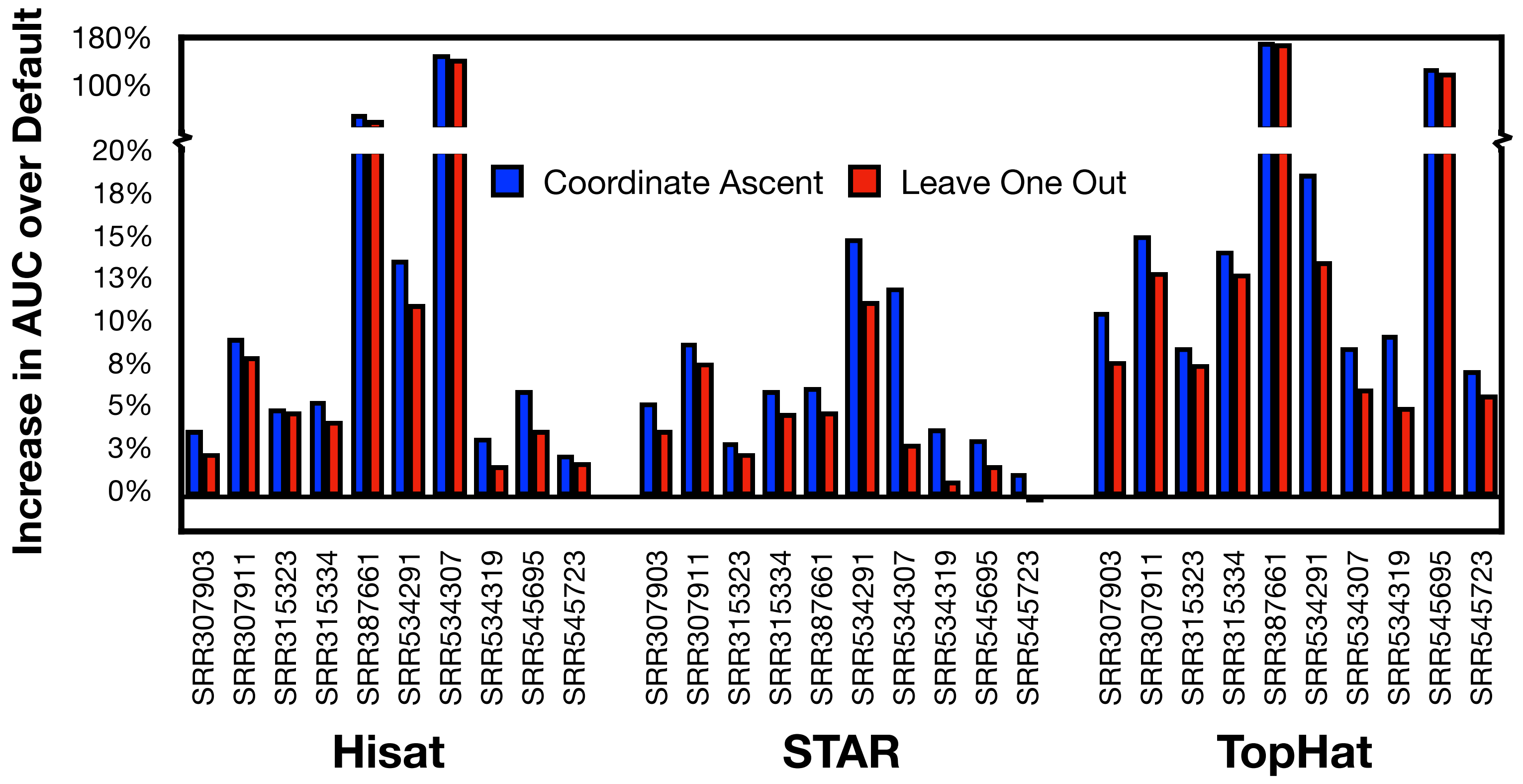
Parameter Advising Performance

Coordinate Ascent reliably improves AUC and generalizes

The advisor set was constructed using 10 experiments from ENCODE mapped to the human genome using 3 aligners.

Coordinate ascent increases AUC by 18.1% on average over using the default parameter choice.

Leave One Out tests advising limited to the 18 choices not trained on the same experiment or aligner.

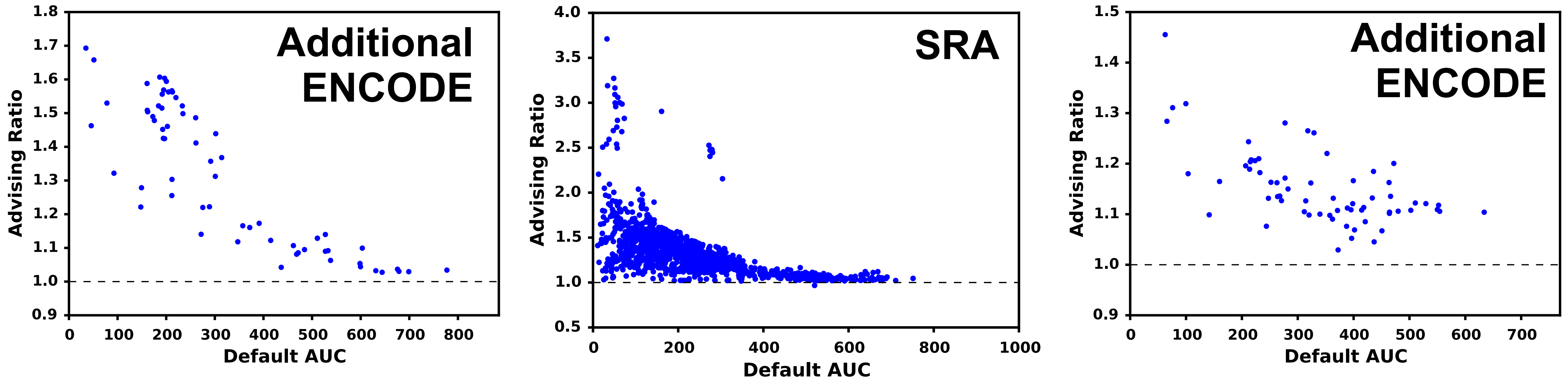


Advising increases AUC for Scallop

Parameter advising was performed on 65 other ENCODE experiments and over 1,200 experiments from the Short Read Archive (SRA). Increases AUC of Scallop by 25.7% and 38.2% on average.

Advising StringTie

Increases AUC for StringTie (Pertea, et al., 2015) by 15.1% on average for the 65 ENCODE experiments not used for finding advisor sets.



advising ratio = (AUC of advising - AUC of default) / AUC of default

References

Shao, M and Kingsford, C. 2017. **Accurate assembly of transcripts through phase-preserving graph decomposition.** Nature Biotechnology, 25, pp 1167-1169.

DeBlasio, D & Kececioglu, K. 2017. **Parameter Advising for Multiple Sequence Alignment.** Volume 26 of the Computational Biology Series. Springer Publishing.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT & Salzberg SL. 2015. **StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.** Nature Biotechnology, 33, pp 290–295

Preprint available on bioRxiv
doi:10.1101/342865

Funding

Gordon and Betty Moore Foundation's Data-Driven Discovery Initiative (GBMF4554)

US National Science Foundation (CCF-1256087, CCF-1319998)

US National Institutes of Health (R01HG007104, R01GM122935)

The Shurl and Kay Curci Foundation

Travel funding provided by the ISCB