Building an automated bioinformatician: more accurate, large-scale genomic discovery using parameter advising

Dan DeBlasio
dandeblasio.com/biodata18
@danfdeblasio

with Kwanho Kim and Carl Kingsford

Carnegie Mellon University
School of Computer Science

Tunable parameters

Biostars — BIOINFORMATICS EXPLAINED

Welcome to Biostar! Community Log In Sign Up

Question: (Closed) Questions about Salmon

Hello All! I had some questions involving alignment based v alignment free mapping done by Salmon as well as some other general questions pertaining specifically to our experiments. Please excuse any perceived ignorance as I'm more of a molecular than computation biologist and

0

1. If one will probably map the reads in order to visuali
alignment based mapping so that Salmon agrees w
2. Our experiments involve looking at ribosome profilin
read counts. We also have the corresponding RNA-
one/both data sets? My idea was to not use either c
ribosome profiling data set.
3. Could someone provide a better explanation for fldM
changing the default settings be more useful? I had

Biostars — BIOINFORMATICS EXPLAINED

Community Log

Question: Salmon libtype - Does it matter?

Hi,

I was wondering what happens if you have the libtype wrong for Salmon? Or why it requires
wasn't sure if the first read of my paired-end reads was forward (ISF) or reverse (ISR), so I r
the mapping efficiencies were identical between the runs.

1

Thanks, Matt

salmon alignment •

ADD COMMENT • link •

Biostars — BIOINFORMATICS EXPLAINED

Welcome to Biostar! Community Log In Sign Up

Question: Salmon unmapped reads

Hi All

ave some samples with low mapping in Salmon (40% and less) that have higher alignments in Tophat, and
ing to troubleshoot.

icked some of the unmapped reads (from writeunmapped salmon parameter) and Blat them to human.

me have 2 or more matches with identity 99% to 100% And some have many many matches, I need to scroll
e page down too much. Many of these matches are 100% and some range between 85% to 100% identity.

12 weeks ago by Sharon • 150

pping, but couldn't

Biostars — BIOINFORMATICS EXPLAINED

Welcome to Biostar! Community Log In Sign Up

Question: RNA-Seq analysis using STAR and Salmon

Hello! I am having some trouble figuring out how to use Salmon. I have around 30 different samples which I trimmed using bbmap then aligned them using STAR. I have all of the BAM files from this alignment. Should I

so
I am
, that

nths ag

Biostars — BIOINFORMATICS EXPLAINED

Welcome to Biostar! Community Log In Sign Up

Question: Salmon: Optimal k-mer size (for indexing) for RNA-seq data alignment using reference genome

Dear all,

I am using salmon to evaluate how the L.donovani gene expression varies in the two diffe
(promastigotes and amastigotes). For that I downloaded two RNA-seq data from SRA and
reference L.donovani coding sequence coding sequences (CDS)

1

In order to quantify gene expression with salmon I have to index the CDS using a specific
salmon manual they use a 31-mer for 75bp reads. My question is whether is reasonable t
value i.e. 0,413*reads length or if there's any other advisable value. I heard that some peo*
length whereas I also found this post where is is metioned between 0.5 and 0.66*reads le

low mapping rate ? #160

Open **atasub** opened this issue on Oct 6, 2017 · 7 comments

atasub commented on Oct 6, 2017 • edited by rob-p

I recently ran Salmon by quasi-mapping-based mode and when I checked the salmon_quant.log file, saw that mapping rate was around ~%65-68 for all of the samples. Do you have any suggestions to improve the mapping rate? I used "--libType A" to to infer the library type info and got a warning that "Greater than 5% of the fragments disagreed with the provided library typ", but I guess this is not an issue. This is an example for one of the "lib_format_counts.json" files:

Biostars — BIOINFORMATICS EXPLAINED

Community Log In Sign Up

Question: Salmon very low mapping

Hi All

Tunable parameters

```
Quant
==========
Perform dual-phase, mapping-based estimation of
transcript abundance from RNA-seq reads

salmon quant options:


basic options:
  -v [ --version ]                      print version string
  -h [ --help ]                         produce help message
  -i [ --index ] arg                    Salmon index
  -l [ --libType ] arg                  Format string describing the library  type
  -r [ --unmatedReads ] arg             List of files containing unmated reads  of (e.g. single-end reads)
  -1 [ --mates1 ] arg                   File containing the #1 mates
  -2 [ --mates2 ] arg                   File containing the #2 mates
  -o [ --output ] arg                   Output quantification file.
  --discardOrphansQuasi                 [Quasi-mapping mode only] : Discard  orphan mappings in quasi-mapping mode.  If this flag is passed then only paired mappings
                                        will be considered toward  quantification estimates.  The default  behavior is to consider orphan mappings if no valid paired
                                        mappings exist.   This flag is independent of the option  to write the orphaned mappings to file  (--writeOrphanLinks).
  --allowOrphansFMD                     [FMD-mapping mode only] : Consider  orphaned reads as valid hits when  performing lightweight-alignment.  This option will
                                        increase sensitivity (allow more reads to map and more transcripts  to be detected), but may decrease  specificity as orphaned
                                        alignments are  more likely to be spurious.
  --seqBias                             Perform sequence-specific bias  correction.
  --gcBias                              [beta for single-end reads] Perform  fragment GC bias correction
  -p [ --threads ] arg                  The number of threads to use  concurrently.
  --incompatPrior arg                   This option sets the prior probability  that an alignment that disagrees with  the specified library type (--libType)  results
                                        from the true fragment origin.  Setting this to 0 specifies that  alignments that disagree with the  library type should be
                                        "impossible",  while setting it to 1 says that  alignments that disagree with the  library type are no less likely than  those
                                        that do
  -g [ --geneMap ] arg                  File containing a mapping of  transcripts to genes.  If this file is  provided Salmon will output both  quant.sf and
                                        quant.genes.sf files,  where the latter contains aggregated  gene-level abundance estimates.  The  transcript to gene mapping
                                        should be  provided as either a GTF file, or a in  a simple tab-delimited format where  each line contains the name of a
                                        transcript and the gene to which it  belongs separated by a tab.  The  extension of the file is used to  determine how the file
                                        should be  parsed.  Files ending in '.gtf', '.gff' or '.gff3' are assumed to be in GTF  format; files with any other extension
                                        are assumed to be in the simple format. In GTF / GFF format, the  "transcript_id" is assumed to contain  the transcript
                                        identifier and the  "gene_id" is assumed to contain the  corresponding gene identifier.
  -z [ --writeMappings ] [=arg(=-)]     If this option is provided, then the  quasi-mapping results will be written  out in SAM-compatible format.  By  default, output
                                        will be directed to  stdout, but an alternative file name  can be provided instead.
  --meta                                If you're using Salmon on a metagenomic dataset  consider setting this flag to  disable parts of the abundance  estimation model
```

--reduceGCMemory	If this option is selected, a more memory efficient (but slightly slower) representation is used to compute fragment GC content. Enabling this will reduce memory usage, but can also reduce speed. However, the results themselves will remain the same.
--biasSpeedSamp arg	The value at which the fragment length PMF is down-sampled when evaluating sequence-specific & GC fragment bias. Larger values speed up effective length correction, but may decrease the fidelity of bias modeling results.
--strictIntersect	Modifies how orphans are assigned. When this flag is set, if the intersection of the quasi-mappings for the left and right is empty, then all mappings for the left and all mappings for the right read are reported as orphaned quasi-mappings
--fldMax arg	The maximum fragment length to consider when building the empirical distribution
--fldMean arg	The mean used in the fragment length distribution prior
--fldSD arg	The standard deviation used in the fragment length distribution prior
-f [--forgettingFactor] arg	The forgetting factor used in the online learning schedule. A smaller value results in quicker learning, but higher variance and may be unstable. A larger value results in slower learning but may be more stable. Value should be in the interval (0.5, 1.0].
-m [--maxOcc] arg	(S)MEMs occuring more than this many times won't be considered.
--initUniform	initialize the offline inference with uniform parameters, rather than seeding with online parameters.
-w [--maxReadOcc] arg	Reads "mapping" to more than this many places won't be considered.
--noLengthCorrection	[experimental] : Entirely disables length correction when estimating the abundance of transcripts. This option can be used with protocols where one expects that fragments derive from their underlying targets without regard to that target's length (e.g. QuantSeq)
--noEffectiveLengthCorrection	Disables effective length correction when computing the probability that a fragment was generated from a transcript. If this flag is passed in, the fragment length distribution is not taken into account when computing this probability.
--noFragLengthDist	[experimental] : Don't consider concordance with the learned fragment length distribution when trying to determine the probability that a fragment has originated from a specified location. Normally, Fragments with unlikely lengths will be assigned a smaller relative probability than those with more likely lengths. When this flag is passed in, the observed fragment length has no effect on that fragment's a priori probability.
--noBiasLengthThreshold	[experimental] : If this option is enabled, then no (lower) threshold will be set on how short bias correction can make effective lengths. This can increase the precision of bias correction, but harm robustness. The default correction applies a threshold.
--numBiasSamples arg	Number of fragment mappings to use when learning the sequence-specific bias model.
--numAuxModelSamples arg	The first <numAuxModelSamples> are used to train the auxiliary model parameters (e.g. fragment length distribution, bias, etc.). After ther first <numAuxModelSamples> observations the auxiliary model parameters will be assumed to have converged and will be fixed.
--numPreAuxModelSamples arg	The first <numPreAuxModelSamples> will have their assignment likelihoods and contributions to the transcript abundances computed without applying any auxiliary models. The purpose of ignoring the auxiliary models for the first <numPreAuxModelSamples> observations is to avoid applying these models before thier parameters have been learned sufficiently well.
--useVBOpt	Use the Variational Bayesian EM rather than the traditional EM algorithm for optimization in the batch passes.
--rangeFactorizationBins arg	Factorizes the likelihood used in quantification by adopting a new notion of equivalence classes based on the conditional probabilities with which fragments are generated from different transcripts. This is a more fine-grained factorization than the normal rich equivalence classes. The default value (0) corresponds to the standard rich equivalence classes, and larger values imply a more fine-grained factorization. If range factorization is enabled, a common value to select for this parameter is 4.
--numGibbsSamples arg	Number of Gibbs sampling rounds to perform.
--numBootstraps arg	Number of bootstrap samples to generate. Note: This is mutually exclusive with Gibbs sampling.
--thinningFactor arg	Number of steps to discard for every sample kept from the Gibbs chain. The larger this number, the less chance that subsequent samples are auto-correlated, but the slower sampling becomes.
-q [--quiet]	Be quiet while doing quantification (don't write informative output to the console unless something goes wrong).
--perTranscriptPrior	The prior (either the default or the argument provided via --vbPrior) will be interpreted as a transcript-level prior (i.e. each transcript will be given a prior read count of this value)
--vbPrior arg	The prior that will be used in the VBEM algorithm. This is interpreted as a per-nucleotide prior, unless the --perTranscriptPrior flag is also given, in which case this is used as a transcript-level prior
--writeOrphanLinks	Write the transcripts that are linked by orphaned reads.
--writeUnmappedNames	Write the names of un-mapped reads to the file unmapped_names.txt in the auxiliary directory.
-x [--quasiCoverage] arg	[Experimental]: The fraction of the read that must be covered by MMPs (of length >= 31) if this read is to be considered as "mapped". This may help to avoid "spurious" mappings. A value of 0 (the default) denotes no coverage threshold (a single 31-mer can yield a mapping). Since coverage by exact matching, large, MMPs is a rather strict condition, this value should likely be set to something low, if used.

Tunable parameters

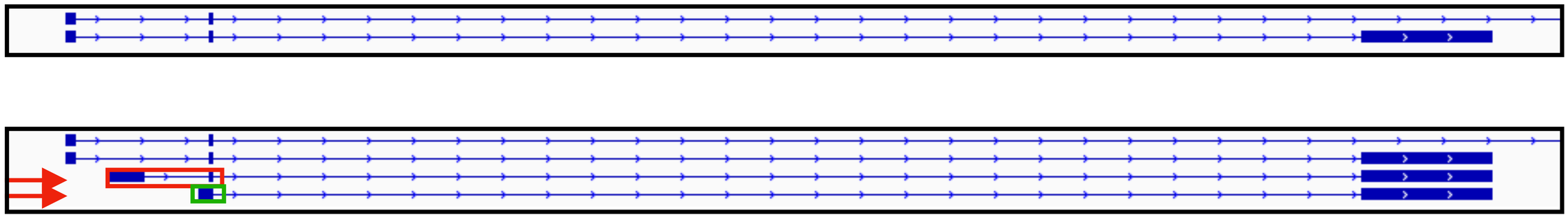
Most users rely on the default parameter settings,

- which are meant to work well on average,
- but the most interesting examples are not typically "average".

Transcript assemblies using different parameter choices

Default Parameter Vector

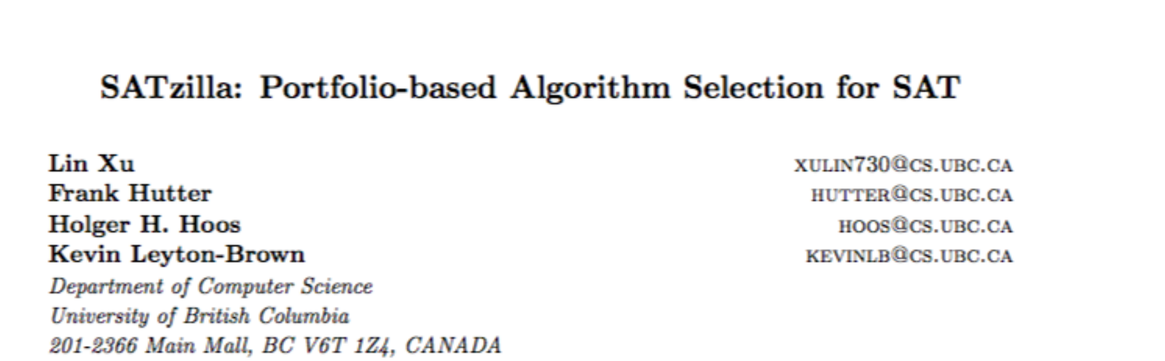
Optimized Parameter Vector



The default parameter choices miss two transcripts that are supported by the data and in the reference transcriptome.

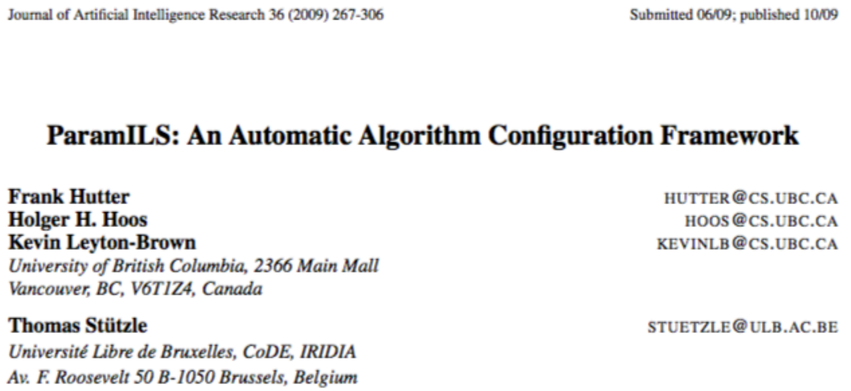
Tunable parameters

It's not just a problem in computational biology!



SATzilla: Portfolio-based Algorithm Selection for SAT

Lin Xu XULIN730@CS.UBC.CA
Frank Hutter HUTTER@CS.UBC.CA
Holger H. Hoos HOOS@CS.UBC.CA
Kevin Leyton-Brown KEVINLB@CS.UBC.CA
Department of Computer Science
University of British Columbia
201-2366 Main Mall, BC V6T 1Z4, CANADA

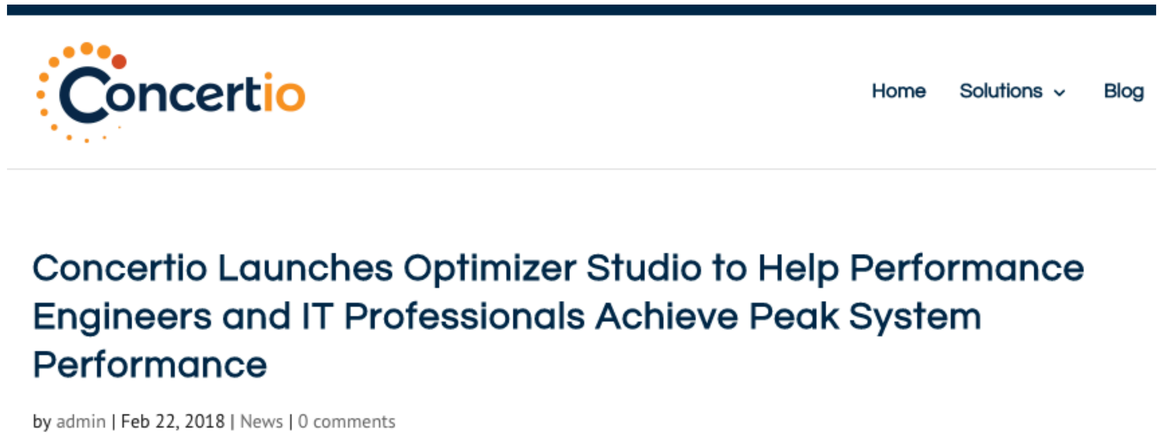


Journal of Artificial Intelligence Research 36 (2009) 267-306 Submitted 06/09; published 10/09

ParamILS: An Automatic Algorithm Configuration Framework

Frank Hutter HUTTER@CS.UBC.CA
Holger H. Hoos HOOS@CS.UBC.CA
Kevin Leyton-Brown KEVINLB@CS.UBC.CA
University of British Columbia, 2366 Main Mall
Vancouver, BC, V6T1Z4, Canada

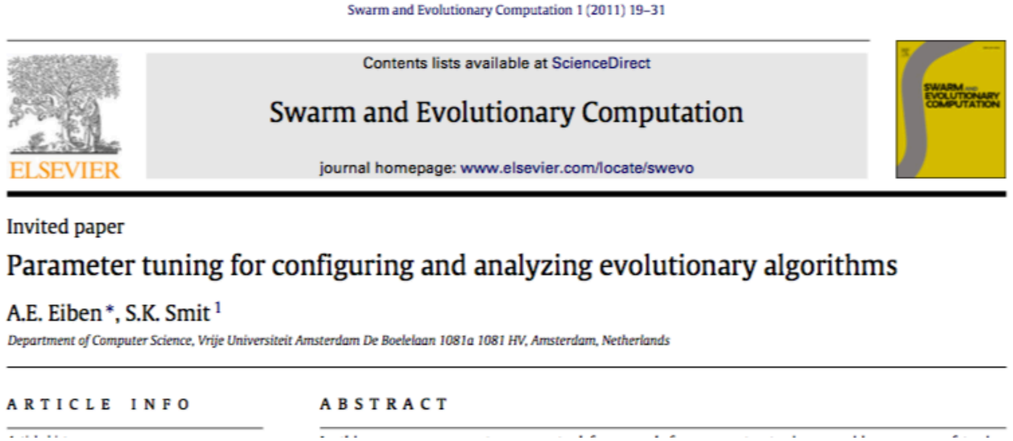
Thomas Stützle STUETZLE@ULB.AC.BE
Université Libre de Bruxelles, CoDE, IRIDIA
Av. F. Roosevelt 50 B-1050 Brussels, Belgium



Concertio Home Solutions Blog

Concertio Launches Optimizer Studio to Help Performance Engineers and IT Professionals Achieve Peak System Performance

by admin | Feb 22, 2018 | News | 0 comments



Swarm and Evolutionary Computation 1 (2011) 19–31

Contents lists available at ScienceDirect

Swarm and Evolutionary Computation

journal homepage: www.elsevier.com/locate/swevo

Invited paper

Parameter tuning for configuring and analyzing evolutionary algorithms

A.E. Eiben *, S.K. Smit [1]
Department of Computer Science, Vrije Universiteit Amsterdam De Boelelaan 1081a 1081 HV, Amsterdam, Netherlands

ARTICLE INFO ABSTRACT

Automated bioinformatician

Almost all pieces of scientific software have tunable parameters.

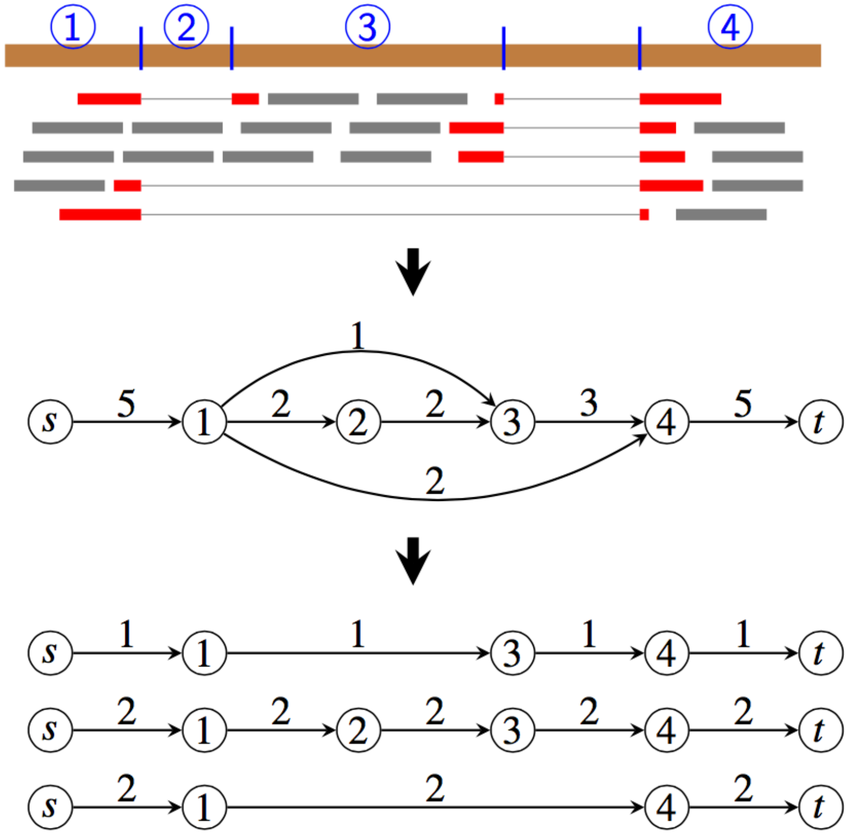
- Their settings can greatly impact the quality of output.
- Default parameters are best on average but may be bad in general.
- Mis-configuration can lead to missed or incorrect conclusions.

Can we remove parameter choice
as a source of error
in transcriptome analysis?

Automated bioinformatician

Transcript assembly

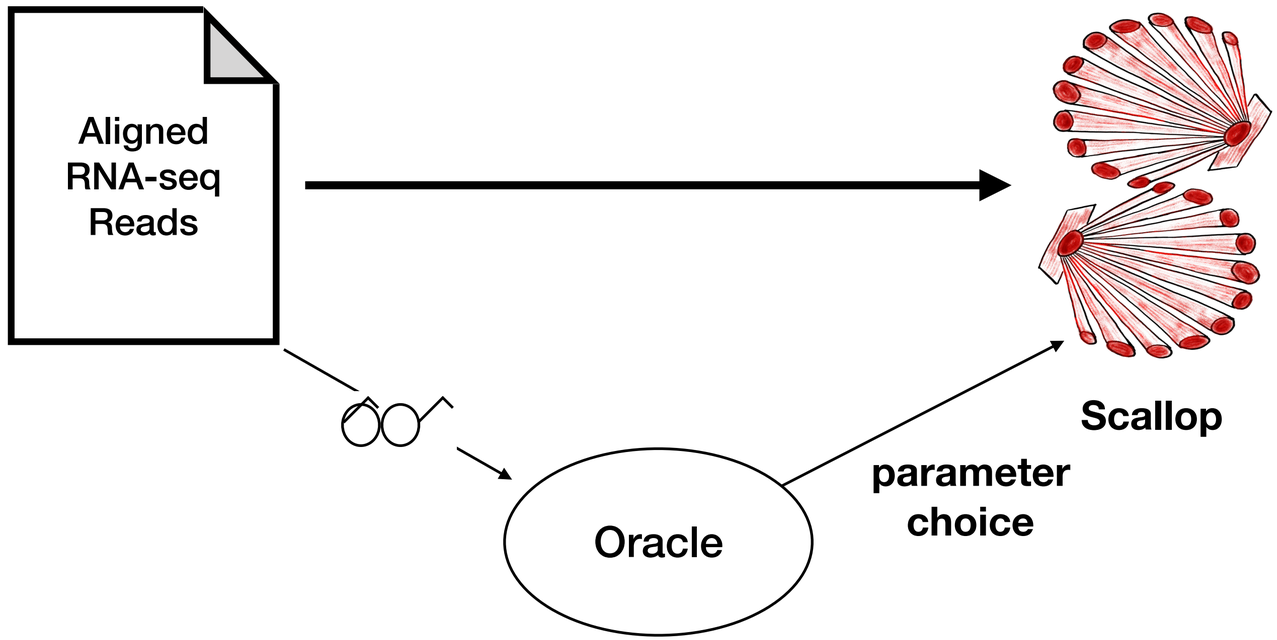
- is a fundamental problem in transcriptomics
- but is computationally difficult
- and easily impacted by choices of parameters.
- We will focus on the Scallop assembler.



[Shao and Kingsford, Nature Biotechnology 2017]

Automated bioinformatician

The goal is to find the parameter choice for a given input.



Automated bioinformatician

In machine learning, this is the hyper-parameter tuning problem

- multi-armed bandit
- simulated annealing
- bayesian frameworks
- etc.

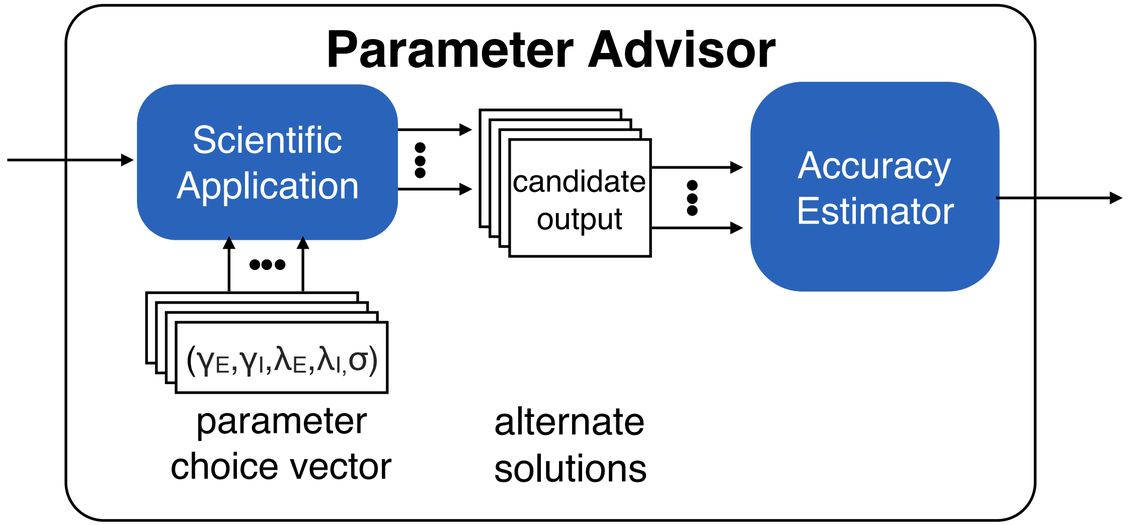
Issue is that running time is increased greatly since

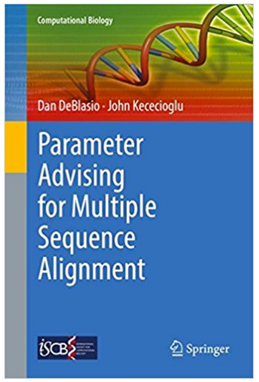
- the application needs to be run multiple times, and
- those instances need to be sequential.

Parameter advising

Given an input an advisor

- uses an advisor set of parameter choice vectors to obtain candidates, then
- returns the most accurate of these solutions based on the accuracy estimation.



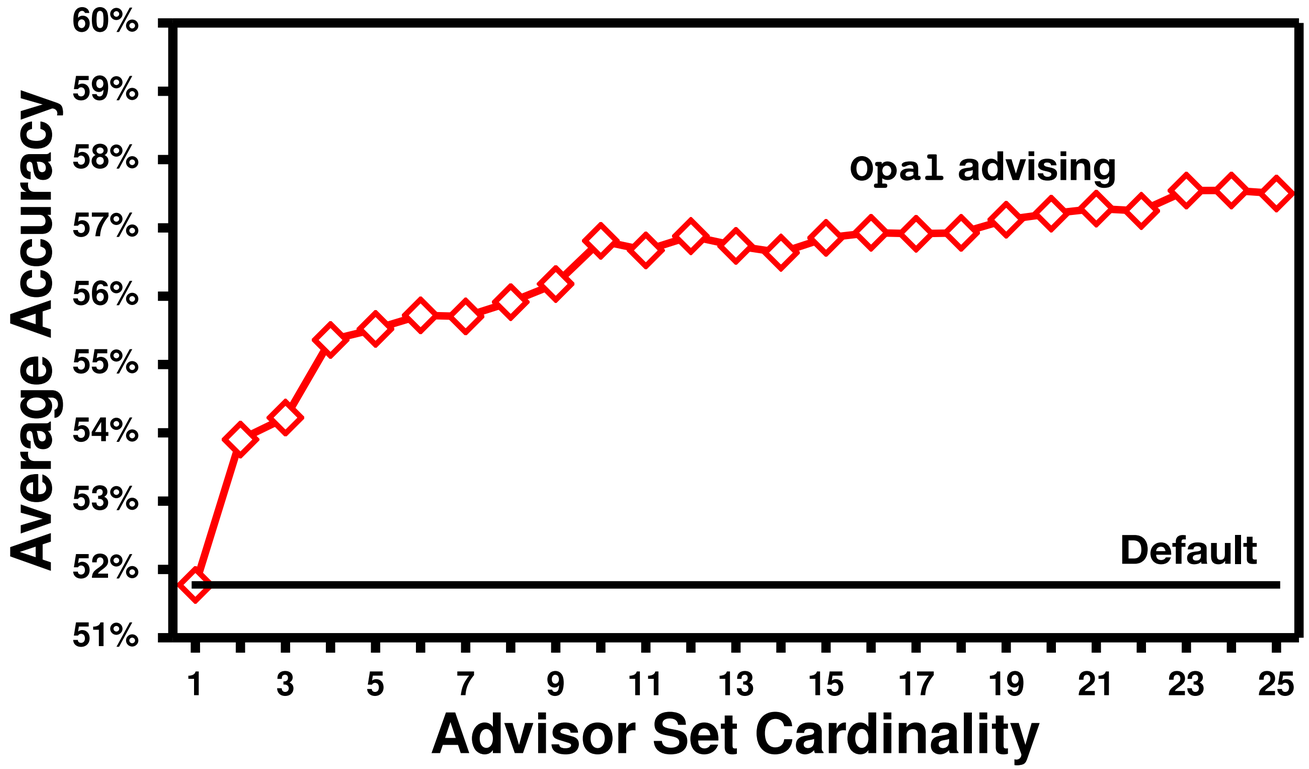


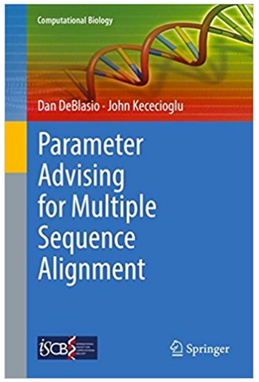
[DeBlasio and Kececioglu, Springer International, 2017]

Parameter advising

Increases accuracy for multiple sequence alignment by

- choosing a parameter choice for each input and
- accuracy increases with advisor set size, but
- so does the resource requirement.



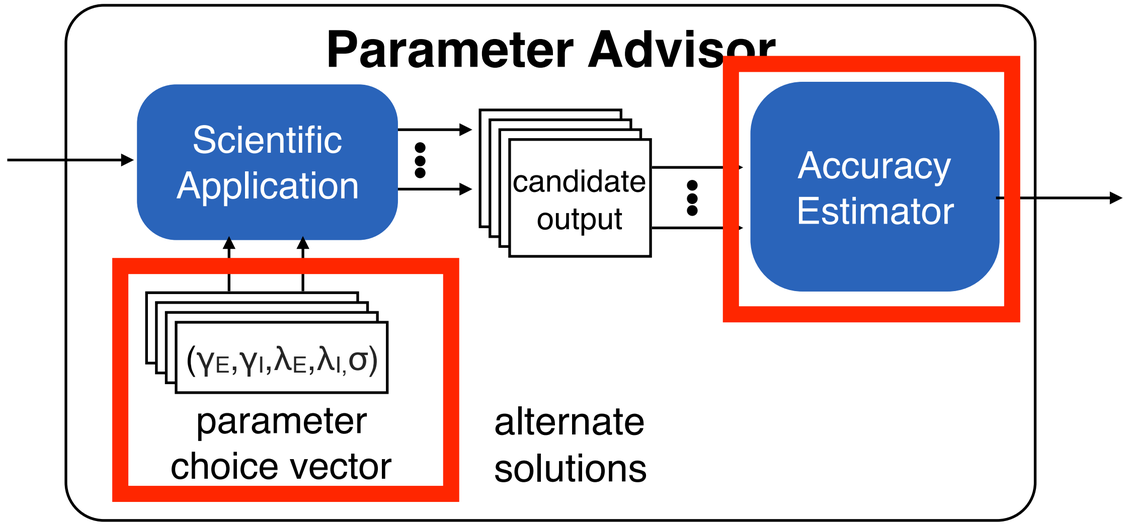


[DeBlasio and Kececioglu, Springer International, 2017]

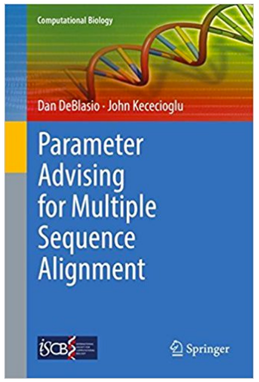
Parameter advising

Given an input an advisor

- uses an advisor set of parameter choice vectors to obtain candidates, then
- returns the most accurate of these solutions based on the accuracy estimation.



Steps to applying advising to a new domain defining: (1) the advisor set, and (2) an accuracy estimator

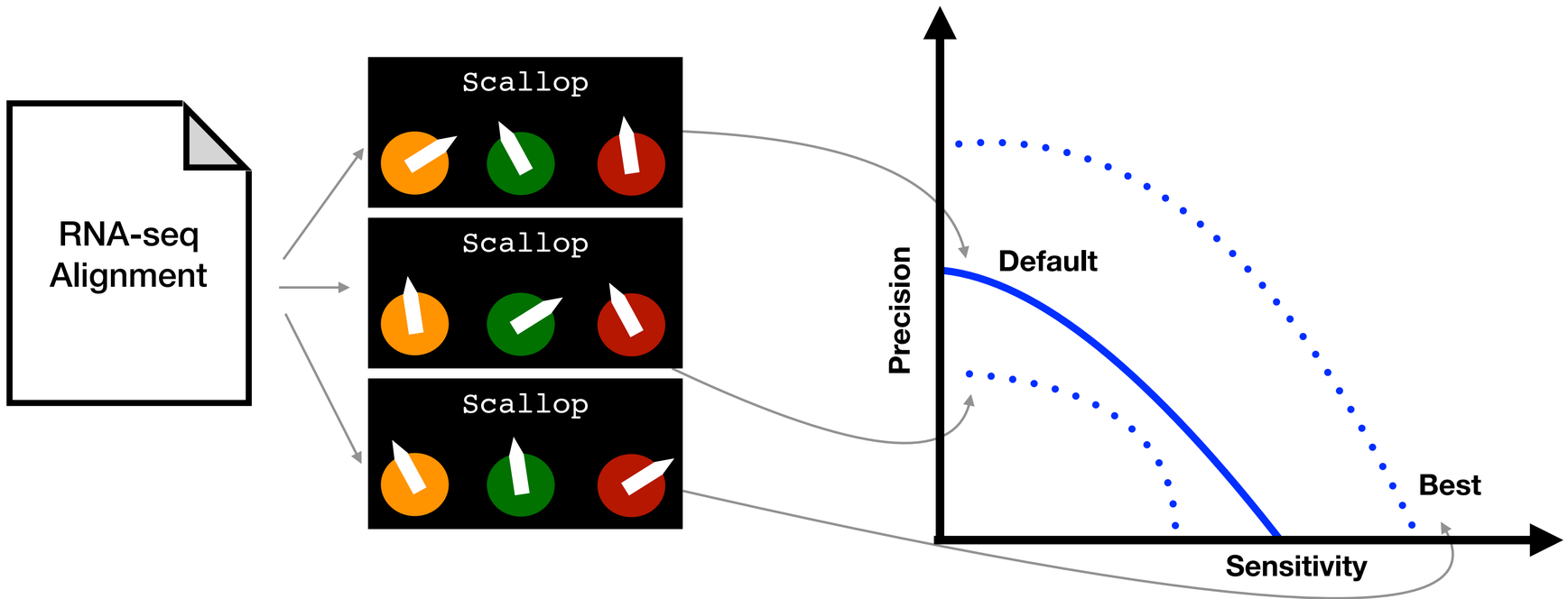


[DeBlasio and Kececioglu, Springer International, 2017]

Scallop advising

Advising for transcript assembly is different from previous implementations;

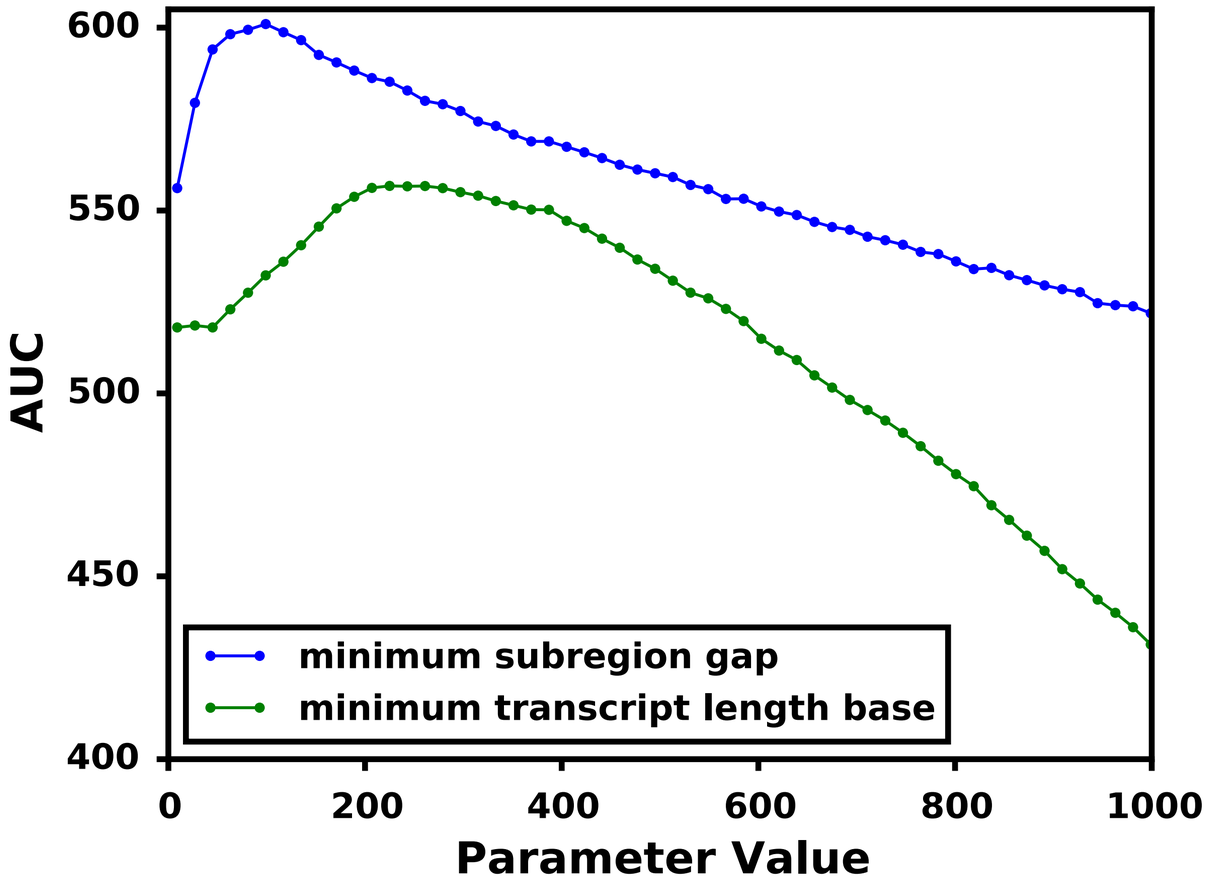
- the reference transcriptome is available so
- the area under the receiver operating characteristic curve can be used as both a measure and an objective, but
- the number of tunable parameters is very large.



Scallop advising

Cannot test all combinations of parameter values.

- Tested the behavior of each parameter in isolation.
- Each parameter had a single global maximum on the large regions tested.
- In general, we did not see non-global local maxima.



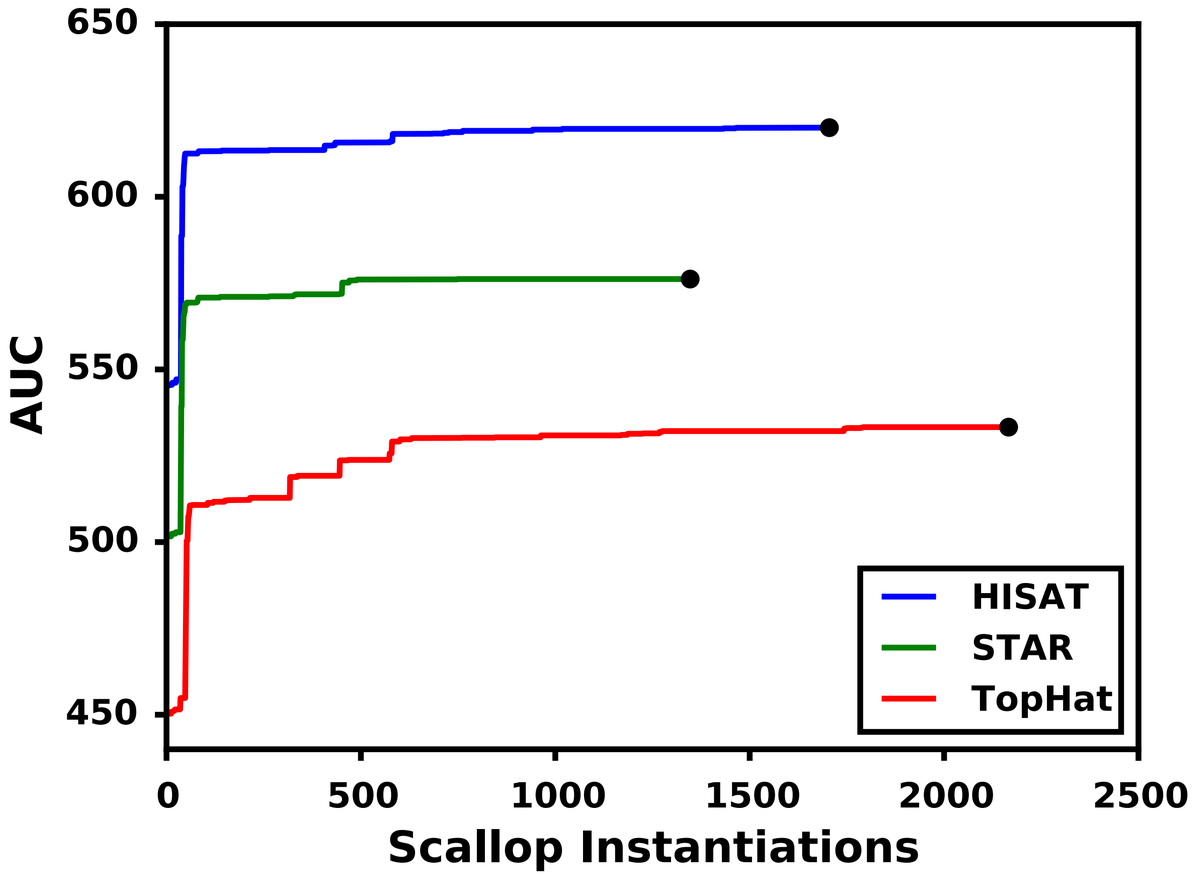
Scallop advising

Parameter curve smoothness means

- coordinate ascent will work well
- but is slow since Scallop's running time is significant.

We can use coordinate ascent

- on a set of training samples
- to find settings for all 18 tunable parameters
- which can be used as a set for parameter advising.
- That set is precomputed and doesn't impact the advising time.



Scallop advising

ID	Experiment	Aligner	ICC	NE	BG	EL	FL	MQ	NH	SBH	SG	SL	SO	TLB	TLI	UM	US
1	SRR545723	TopHat	2	1000	1359	20	3	11	20	1	39	11	1.5	454	8	false	false
2	SRR307911	TopHat	2	1000	1032	20	2	11	21	1	52	15	1.5	432	19	false	false
3	SRR534291	TopHat	0.75	51	785	25	3	1	9	1	121	16	1.51	503	8	false	false
4	SRR387661	TopHat	0.22	1000	999	20	3	11	20	1	53	15	1.5	429	19	false	false
5	SRR315334	TopHat	1.5	1000	1425	20	3	11	20	1	50	16	1.5	528	1	false	false
6	SRR307903	TopHat	1.61	250	733	2	0	11	2	1	23	26	0.75	521	2	false	false
7	SRR545695	TopHat	2	1000	852	20	2	2	104	1	5	7	1.5	432	19	false	false
8	SRR534319	TopHat	2.27	1000	306	20	0	1	20	1	7	11	1.5	346	21	false	false
9	SRR315323	TopHat	0.22	205	945	20	0	2	20	1	39	15	1.16	409	8	true	true
10	SRR534307	TopHat	2.22	1000	78	32	3	11	16	2	0	32	1.5	512	8	false	false
11	SRR545723	HISAT	1.75	11000	628	20	3	11	20	2	8	7	1.5	408	8	false	false
12	SRR307911	HISAT	99.66	1000	1828	3	5	11	3	1	43	15	0.17	392	8	false	false
13	SRR534291	HISAT	1	11000	851	22	5	11	3	1	120	3	0.17	267	30	false	false
14	SRR387661	HISAT	1.05	454	130	26	4	11	38	1	23	0	0.17	880	0	false	false
15	SRR315334	HISAT	0	282	870	20	4	11	1	1	48	8	1.28	307	21	false	false
16	SRR307903	HISAT	1.42	168	745	9	5	11	2	1	23	6	0.17	396	0	false	false
17	SRR545695	HISAT	5.56	1000	778	26	2	11	25	2	0	0	1.5	360	8	false	false
18	SRR534319	HISAT	2	1000	1669	20	0	11	81	1	4	7	1.5	150	73	false	false
19	SRR315323	HISAT	2	685	782	15	0	11	14	1	40	7	1.5	180	50	false	false
20	SRR534307	HISAT	2	146	309	41	8	11	10	3	2	0	1.17	370	4	false	false
21	SRR545723	STAR	2	11000	455	22	6	11	3	1	3	21	1.26	252	16	false	false
22	SRR307911	STAR	0.22	1000	490	20	3	11	8	1	53	26	0.5	289	24	false	false
23	SRR534291	STAR	0.22	11000	481	24	0	2	11	1	123	26	0.24	352	8	false	false
24	SRR387661	STAR	2	1000	734	20	4	11	35	2	42	29	1.5	411	1	false	false
25	SRR315334	STAR	0.5	1000	1070	8	0	11	8	1	41	29	0.75	267	33	false	false
26	SRR307903	STAR	2	1000	976	8	5	11	20	1	35	35	0.17	433	8	false	false
27	SRR545695	STAR	2	1000	388	20	8	11	89	2	20	23	1.5	307	19	false	false
28	SRR534319	STAR	2	800	1574	17	4	2	20	1	8	18	1.5	307	4	false	false
29	SRR315323	STAR	2	11000	1043	20	3	7	20	1	33	35	1.5	150	50	true	true
30	SRR534307	STAR	1.58	292	54	20	0	11	9	3	0	29	1.5	320	8	false	false
	Default		2.00	1,000	50	20	3	1	20	1	3	15	1.50	150	50	false	false
	Initial step size		10.00	10,000	100	100	10	10	100	10	10	20	10.00	500	100	n/a	n/a

- maximum dynamic programming table size (DP)
- maximum edit distance (ED)
- maximum intron contamination coverage (ICC)
- maximum number of exons (NE)
- minimum bundle gap (BG)
- minimum exon length (EL)
- minimum flank length (FL)
- minimum mapping quality (MQ)
- minimum number or hits in a bundle (NH)
- minimum router count (RC)
- minimum splice boundary hits (SBH)
- minimum subregion gap (SG)
- minimum subregion length (SL)
- minimum subregion overlap (SO)
- minimum transcript length base (TLB)
- minimum transcript length increase (TLI)
- uniquely mapped reads only (UM)
- use the secondary alignment (US)

Scallop advising

ID	Experiment	Aligner	ICC	NE	BG	EL	FL	MQ	NH	SBH	SG	SL	SO	TLB	TLI	UM	US
1	SRR545723	TopHat	2	1000	1359	20	3	11	20	1	39	11	1.5	454	8	false	false
2	SRR307911	TopHat	2	1000	1032	20	2	11	21	1	52	15	1.5	432	19	false	false
3	SRR534291	TopHat	0.75	51	785	25	3	1	9	1	121	16	1.51	503	8	false	false
4	SRR387661	TopHat	0.22	1000	999	20	3	11	20	1	53	15	1.5	429	19	false	false
5	SRR315334	TopHat	1.5	1000	1425	20	3	11	20	1	50	16	1.5	528	1	false	false
6	SRR307903	TopHat	1.61	250	733	2	0	11	2	1	23	26	0.75	521	2	false	false
7	SRR545695	TopHat	2	1000	852	20	2	2	104	1	5	7	1.5	432	19	false	false
8	SRR534319	TopHat	2.27	1000	306	20	0	1	20	1	7	11	1.5	346	21	false	false
9	SRR315323	TopHat	0.22	205	945	20	0	2	20	1	39	15	1.16	409	8	true	true
10	SRR534307	TopHat	2.22	1000	78	32	3	11	16	2	0	32	1.5	512	8	false	false
11	SRR545723	HISAT	1.75	11000	628	20	3	11	20	2	8	7	1.5	408	8	false	false
12	SRR307911	HISAT	99.66	1000	1828	3	5	11	3	1	43	15	0.17	392	8	false	false
13	SRR534291	HISAT	1	11000	851	22	5	11	3	1	120	3	0.17	267	30	false	false
14	SRR387661	HISAT	1.05	454	130	26	4	11	38	1	23	0	0.17	880	0	false	false
15	SRR315334	HISAT	0	282	870	20	4	11	1	1	48	8	1.28	307	21	false	false
16	SRR307903	HISAT	1.42	168	745	9	5	11	2	1	23	6	0.17	396	0	false	false
17	SRR545695	HISAT	5.56	1000	778	26	2	11	25	2	0	0	1.5	360	8	false	false
18	SRR534319	HISAT	2	1000	1669	20	0	11	81	1	4	7	1.5	150	73	false	false
19	SRR315323	HISAT	2	685	782	15	0	11	14	1	40	7	1.5	180	50	false	false
20	SRR534307	HISAT	2	146	309	41	8	11	10	3	2	0	1.17	370	4	false	false
21	SRR545723	STAR	2	11000	455	22	6	11	3	1	3	21	1.26	252	16	false	false
22	SRR307911	STAR	0.22	1000	490	20	3	11	8	1	53	26	0.5	289	24	false	false
23	SRR534291	STAR	0.22	11000	481	24	0	2	11	1	123	26	0.24	352	8	false	false
24	SRR387661	STAR	2	1000	734	20	4	11	35	2	42	29	1.5	411	1	false	false
25	SRR315334	STAR	0.5	1000	1070	8	0	11	8	1	41	29	0.75	267	33	false	false
26	SRR307903	STAR	2	1000	976	8	5	11	20	1	35	35	0.17	433	8	false	false
27	SRR545695	STAR	2	1000	388	20	8	11	89	2	20	23	1.5	307	19	false	false
28	SRR534319	STAR	2	800	1574	17	4	2	20	1	8	18	1.5	307	4	false	false
29	SRR315323	STAR	2	11000	1043	20	3	7	20	1	33	35	1.5	150	50	true	true
30	SRR534307	STAR	1.58	292	54	20	0	11	9	3	0	29	1.5	320	8	false	false
	Default		2.00	1,000	50	20	3	1	20	1	3	15	1.50	150	50	false	false
	Initial step size		10.00	10,000	100	100	10	10	100	10	10	20	10.00	500	100	n/a	n/a

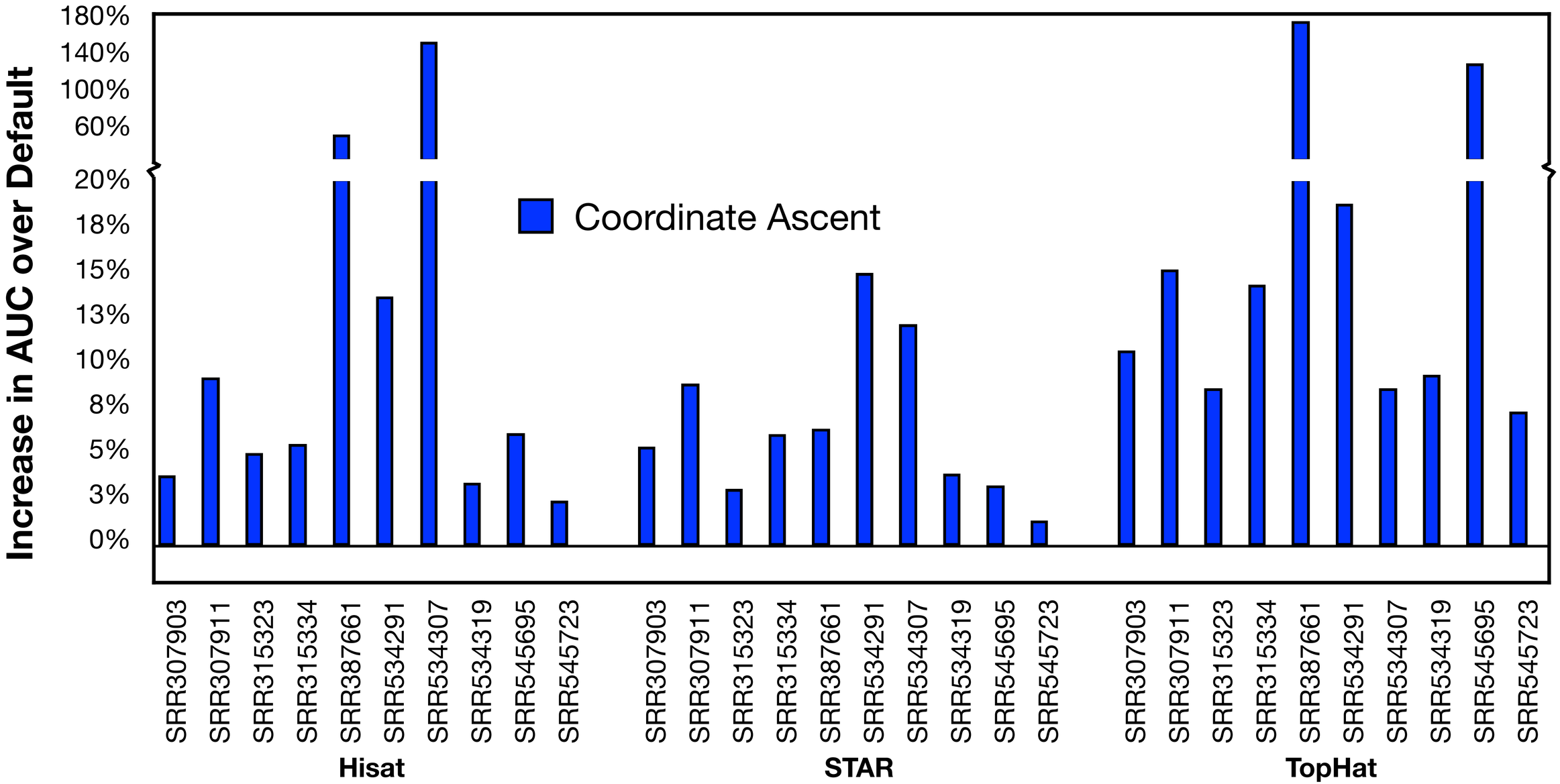
- maximum dynamic programming table size (DP)
- maximum edit distance (ED)
- maximum intron contamination coverage (ICC)
- maximum number of exons (NE)
- minimum bundle gap (BG)
- minimum exon length (EL)
- minimum flank length (FL)
- minimum mapping quality (MQ)
- minimum number or hits in a bundle (NH)
- minimum router count (RC)
- minimum splice boundary hits (SBH)
- minimum subregion gap (SG)
- minimum subregion length (SL)
- minimum subregion overlap (SO)
- minimum transcript length base (TLB)
- minimum transcript length increase (TLI)
- uniquely mapped reads only (UM)
- use the secondary alignment (US)

Scallop advising

ID	Experiment	Aligner	ICC	NE	BG	EL	FL	MQ	NH	SBH	SG	SL	SO	TLB	TLI	UM	US
1	SRR545723	TopHat	2	1000	1359	20	3	11	20	1	39	11	1.5	454	8	false	false
2	SRR307911	TopHat	2	1000	1032	20	2	11	21	1	52	15	1.5	432	19	false	false
3	SRR534291	TopHat	0.75	51	785	25	3	1	9	1	121	16	1.51	503	8	false	false
4	SRR387661	TopHat	0.22	1000	999	20	3	11	20	1	53	15	1.5	429	19	false	false
5	SRR315334	TopHat	1.5	1000	1425	20	3	11	20	1	50	16	1.5	528	1	false	false
6	SRR307903	TopHat	1.61	250	733	2	0	11	2	1	23	26	0.75	521	2	false	false
7	SRR545695	TopHat	2	1000	852	20	2	2	104	1	5	7	1.5	432	19	false	false
8	SRR534319	TopHat	2.27	1000	306	20	0	1	20	1	7	11	1.5	346	21	false	false
9	SRR315323	TopHat	0.22	205	945	20	0	2	20	1	39	15	1.16	409	8	true	true
10	SRR534307	TopHat	2.22	1000	78	32	3	11	16	2	0	32	1.5	512	8	false	false
11	SRR545723	HISAT	1.75	11000	628	20	3	11	20	2	8	7	1.5	408	8	false	false
12	SRR307911	HISAT	99.66	1000	1828	3	5	11	3	1	43	15	0.17	392	8	false	false
13	SRR534291	HISAT	1	11000	851	22	5	11	3	1	120	3	0.17	267	30	false	false
14	SRR387661	HISAT	1.05	454	130	26	4	11	38	1	23	0	0.17	880	0	false	false
15	SRR315334	HISAT	0	282	870	20	4	11	1	1	48	8	1.28	307	21	false	false
16	SRR307903	HISAT	1.42	168	745	9	5	11	2	1	23	6	0.17	396	0	false	false
17	SRR545695	HISAT	5.56	1000	778	26	2	11	25	2	0	0	1.5	360	8	false	false
18	SRR534319	HISAT	2	1000	1669	20	0	11	81	1	4	7	1.5	150	73	false	false
19	SRR315323	HISAT	2	685	782	15	0	11	14	1	40	7	1.5	180	50	false	false
20	SRR534307	HISAT	2	146	309	41	8	11	10	3	2	0	1.17	370	4	false	false
21	SRR545723	STAR	2	11000	455	22	6	11	3	1	3	21	1.26	252	16	false	false
22	SRR307911	STAR	0.22	1000	490	20	3	11	8	1	53	26	0.5	289	24	false	false
23	SRR534291	STAR	0.22	11000	481	24	0	2	11	1	123	26	0.24	352	8	false	false
24	SRR387661	STAR	2	1000	734	20	4	11	35	2	42	29	1.5	411	1	false	false
25	SRR315334	STAR	0.5	1000	1070	8	0	11	8	1	41	29	0.75	267	33	false	false
26	SRR307903	STAR	2	1000	976	8	5	11	20	1	35	35	0.17	433	8	false	false
27	SRR545695	STAR	2	1000	388	20	8	11	89	2	20	23	1.5	307	19	false	false
28	SRR534319	STAR	2	800	1574	17	4	2	20	1	8	18	1.5	307	4	false	false
29	SRR315323	STAR	2	11000	1043	20	3	7	20	1	33	35	1.5	150	50	true	true
30	SRR534307	STAR	1.58	292	54	20	0	11	9	3	0	29	1.5	320	8	false	false
	Default		2.00	1,000	50	20	3	1	20	1	3	15	1.50	150	50	false	false
	Initial step size		10.00	10,000	100	100	10	10	100	10	10	20	10.00	500	100	n/a	n/a

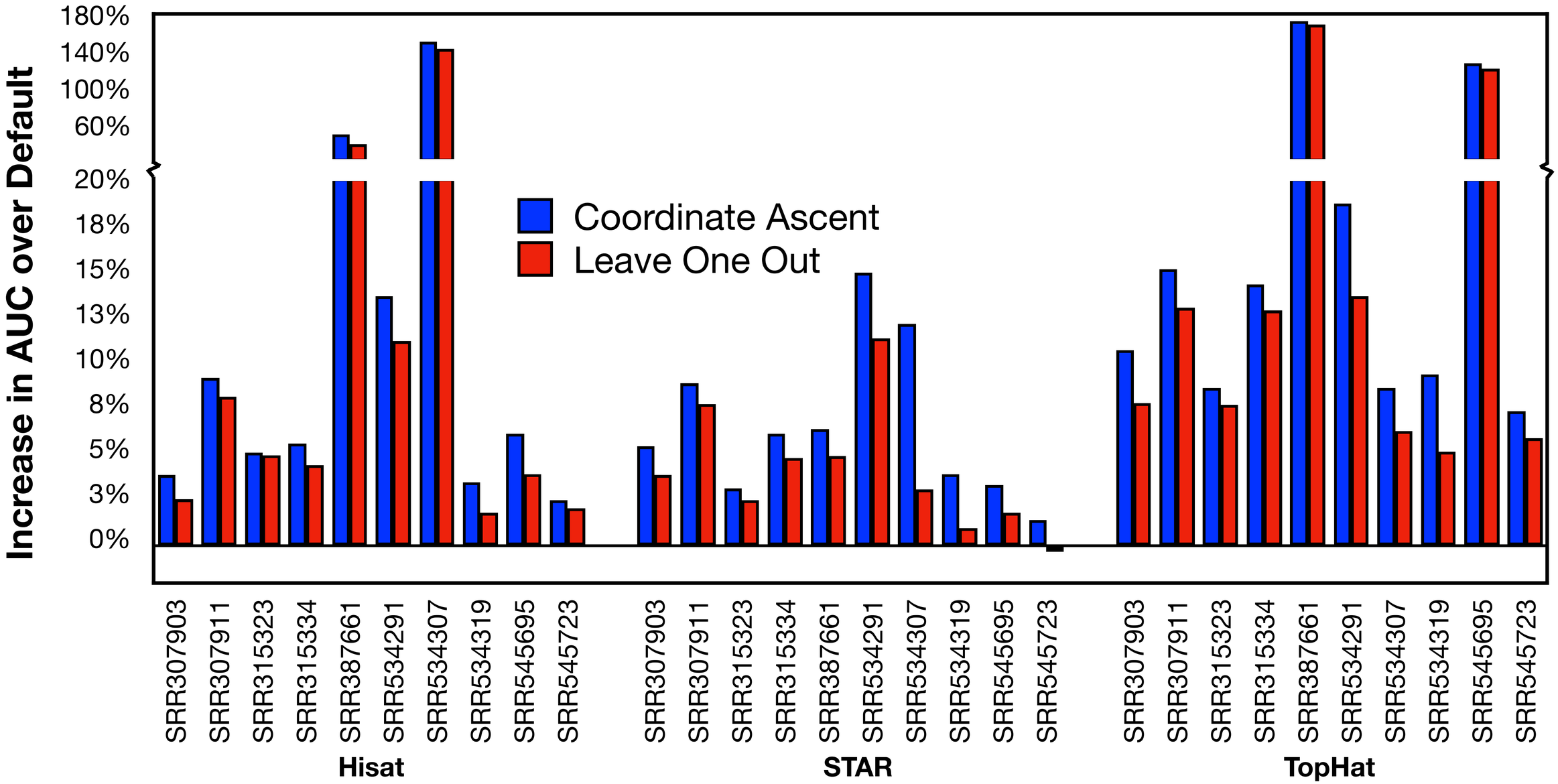
- maximum dynamic programming table size (DP)
- maximum edit distance (ED)
- maximum intron contamination coverage (ICC)
- maximum number of exons (NE)
- minimum bundle gap (BG)
- minimum exon length (EL)
- minimum flank length (FL)
- minimum mapping quality (MQ)
- minimum number or hits in a bundle (NH)
- minimum router count (RC)
- minimum splice boundary hits (SBH)
- minimum subregion gap (SG)
- minimum subregion length (SL)
- minimum subregion overlap (SO)
- minimum transcript length base (TLB)
- minimum transcript length increase (TLI)
- uniquely mapped reads only (UM)
- use the secondary alignment (US)

Scallop advising



Average of 18.1% increase in AUC using Coordinate Ascent

Scallop advising

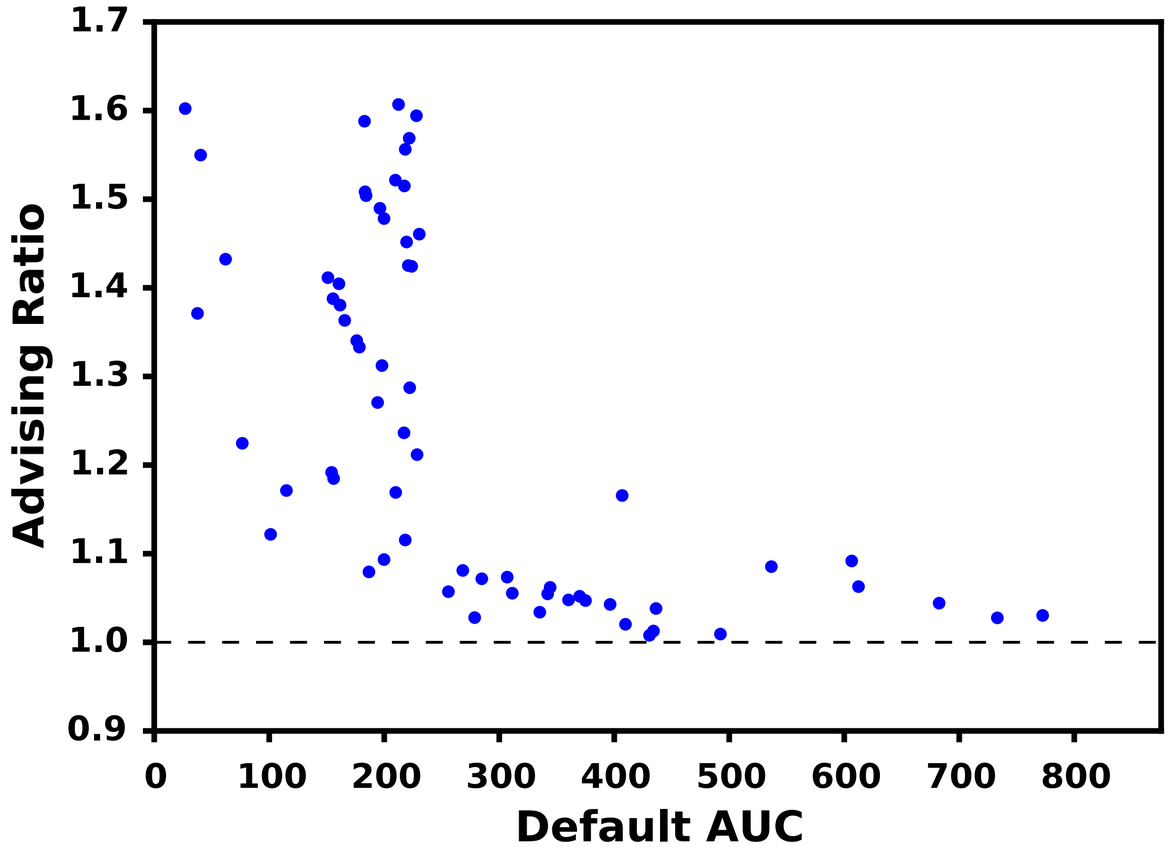


Average of 15.5% increase in AUC for leave-one-out

Scallop advising

65 ENCODE dataset

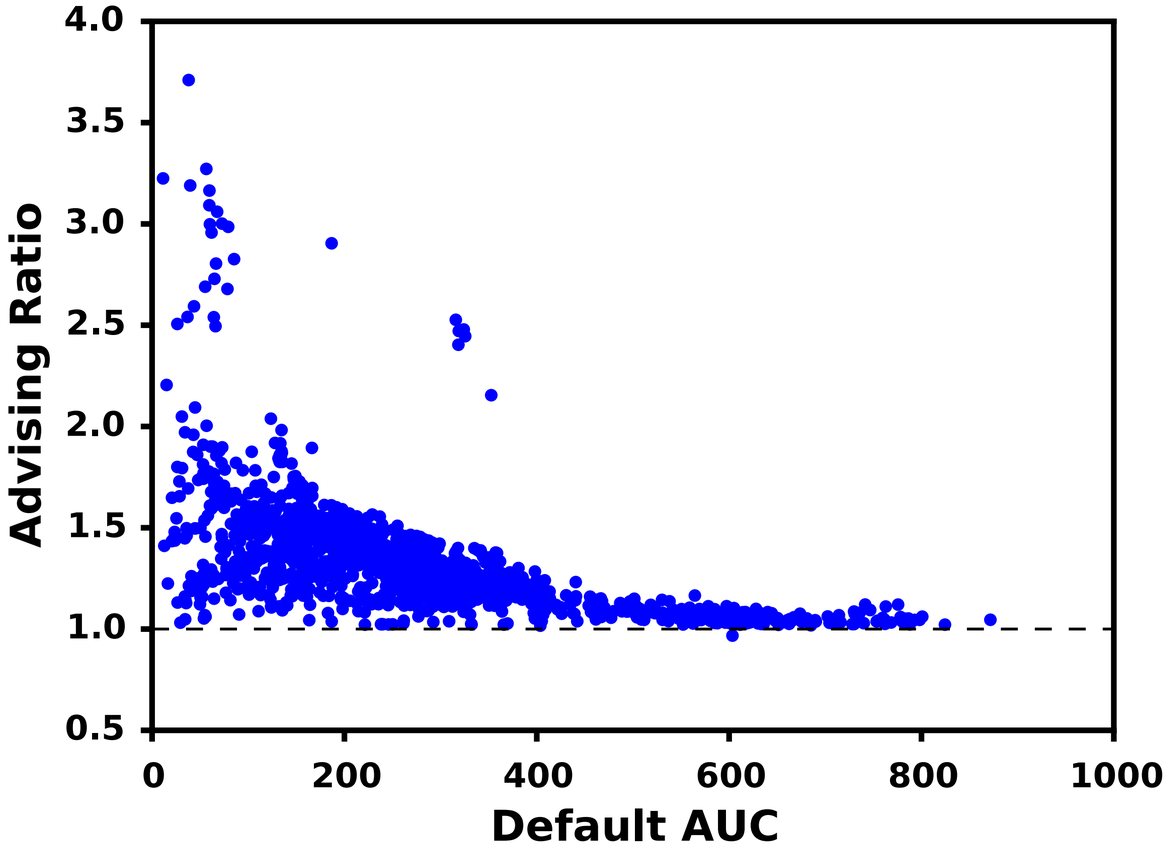
- all of the aligned RNA-seq experiments from ENCODE
- aligned using a variety of aligners
- using either the current or legacy reference genome
- stands in for the performance of advising on generic input
- **average 25.7% increase in AUC**



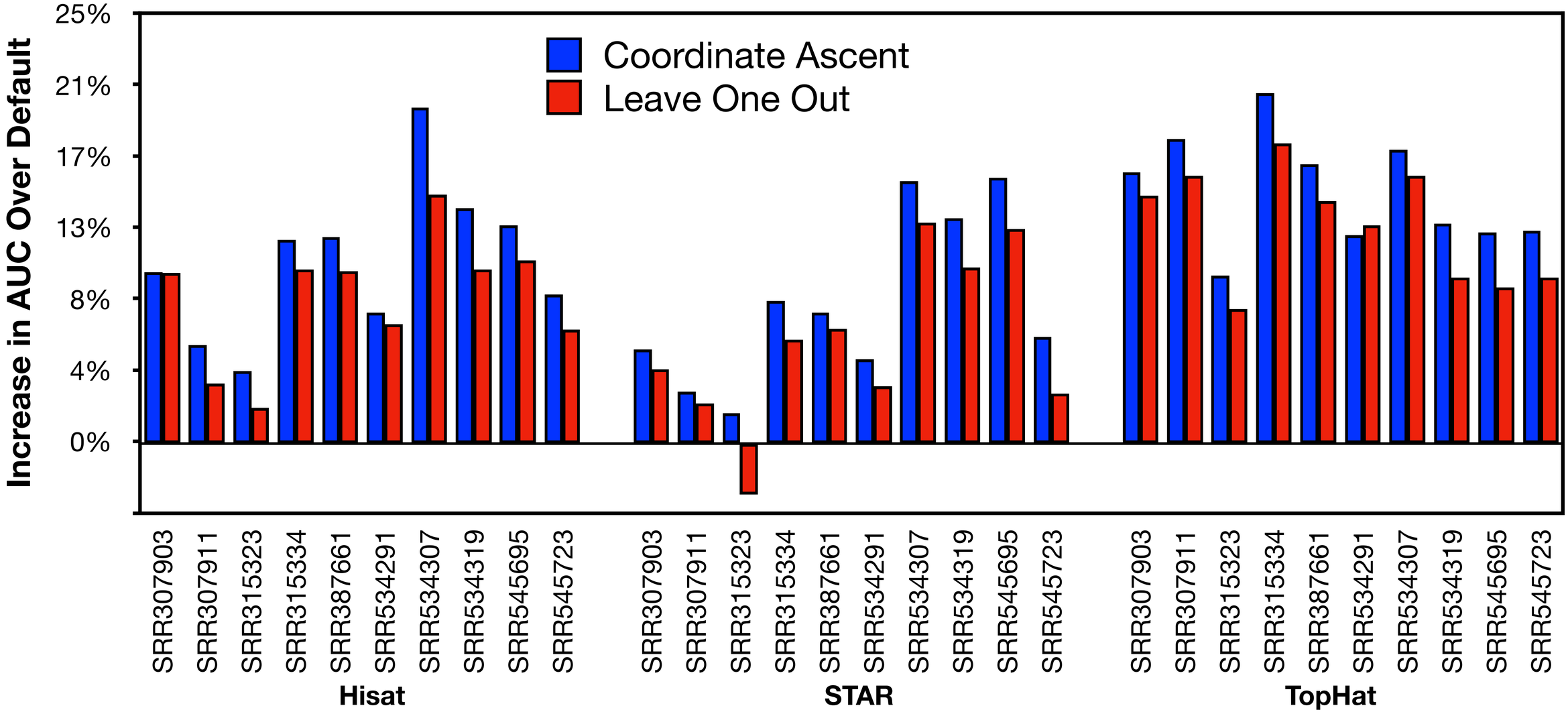
Scallop advising

SRA dataset

- all 1595 RNA-Seq experiments from the SRA
- aligned using STAR to the same reference genome
- represents performance of advising in a high-throughput experiment
- **average of 38.2% increase in AUC**



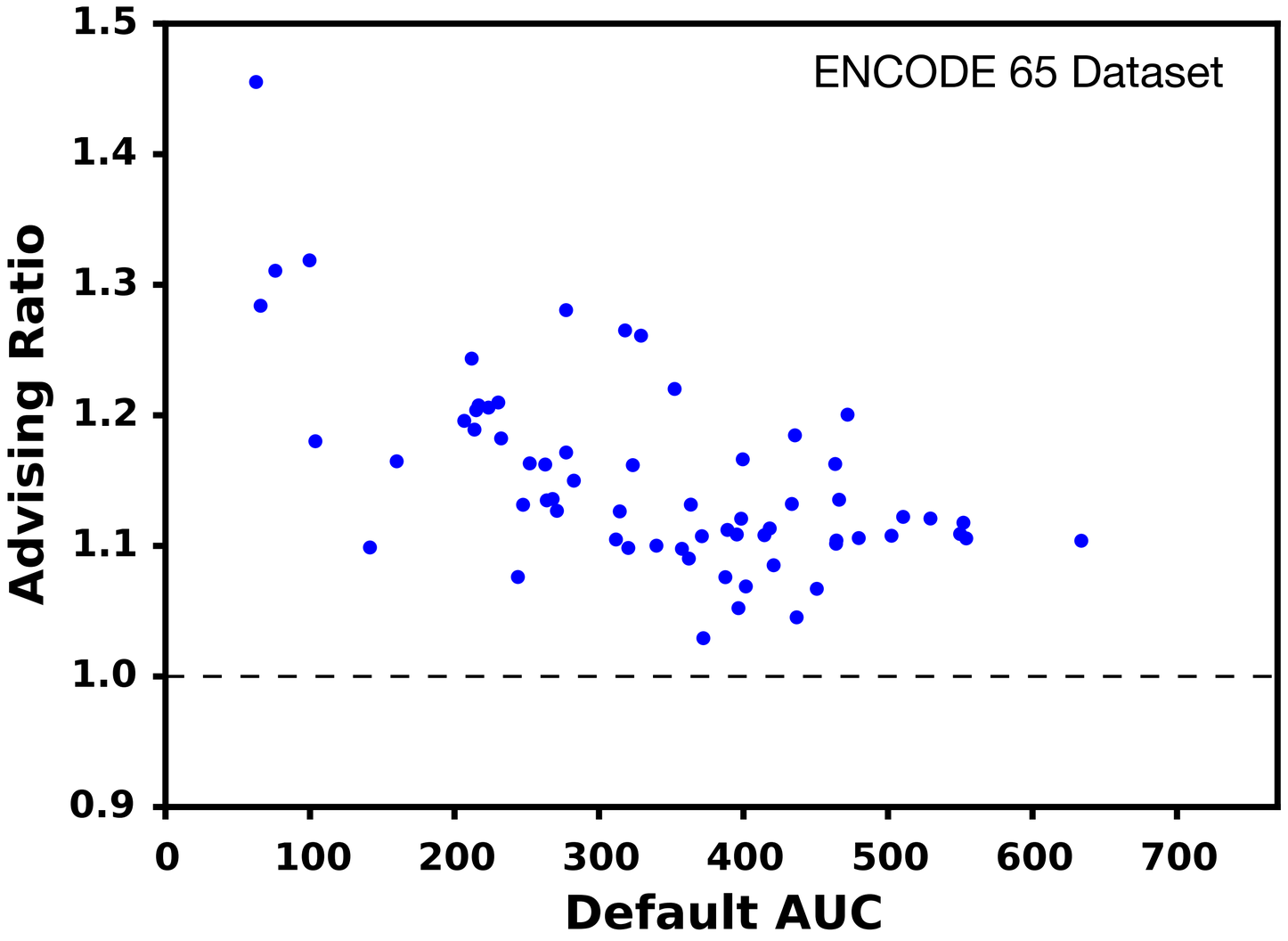
StringTie advising



11.1% and 9.0% average increase in accuracy for Coordinate Ascent and Leave One Out

[Pertea, *et al.*, Nature Biotechnology 2015]

StringTie advising

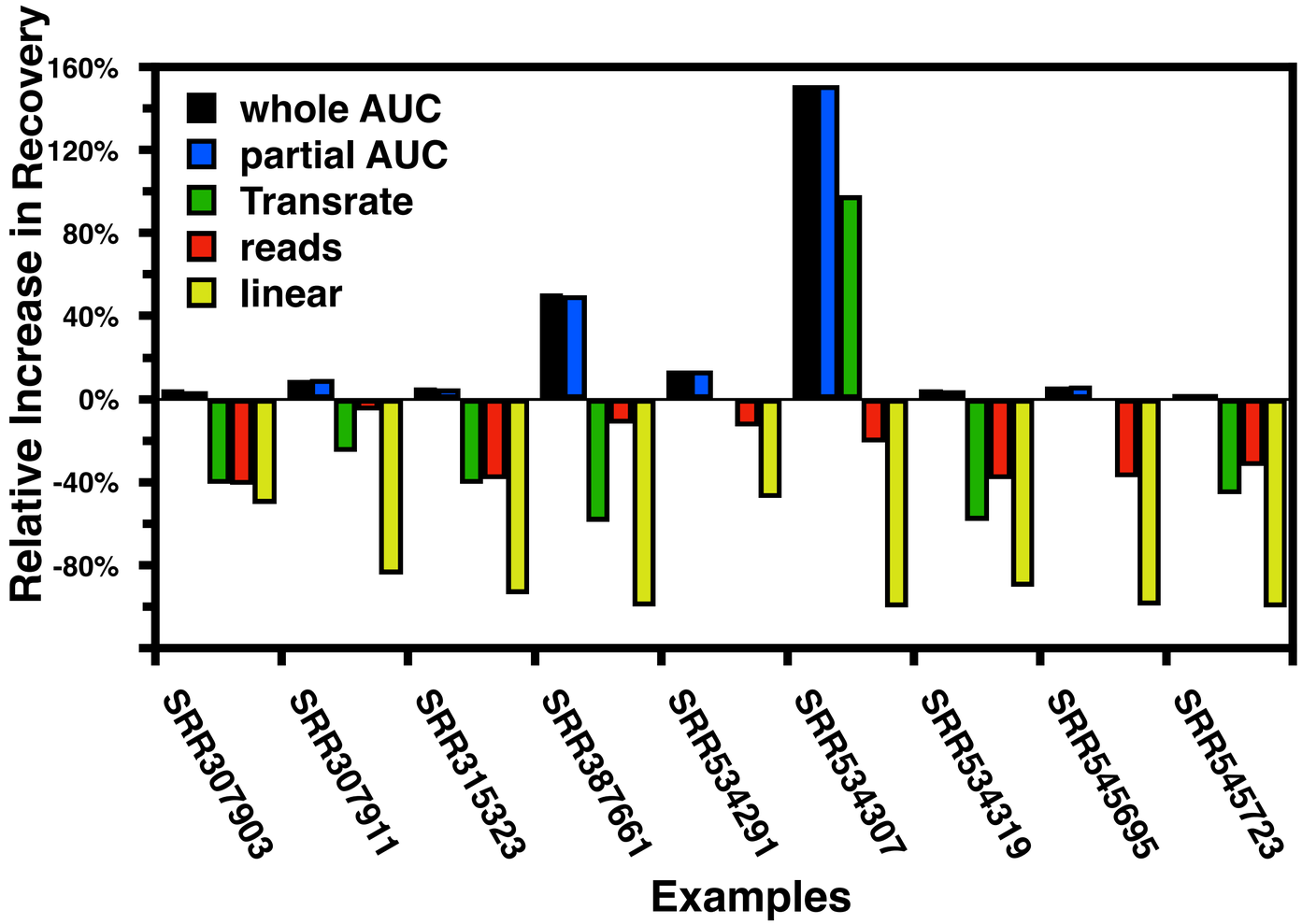


average 15.1% increase in AUC

AUC vs other metrics

AUC penalizes all transcripts that don't map to the reference

- simulated data where we know the "novel" transcripts
- optimized using coordinate ascent and various metrics
- recovery rate of the reference & novelty compared to default



AUC is the only tested method to increase recovery when optimized

Summary

Parameter advising increases AUC for transcript assembly.

- Coordinate ascent is a novel method for advisor set construction.
- Advisor subsets can be used to reduce the resource requirements.
- Improvements are seen for both Scallop and StringTie.
- AUC is currently the best optimization metric.

Preprint available on bioRxiv
doi:10.1101/342865

Acknowledgments

Kingsford Group

Mingfu Shao
Guillaume Marçais
Heewook Lee
Brad Solomon
Natalie Sauerwald
Cong Ma
Hongyu Zheng
Laura Tung

Funding

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

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

US National Institutes of Health (R01HG007104 and R01GM122935)

The Shurl and Kay Curci Foundation

Slides (and links): **dandeblasio.com/biodata18**

Preprint available on bioRxiv
doi:10.1101/342865

Scallop Advising: https://github.com/Kingsford-Group/scallopadvising

Carnegie Mellon University
School of Computer Science

GORDON AND BETTY
MOORE
FOUNDATION