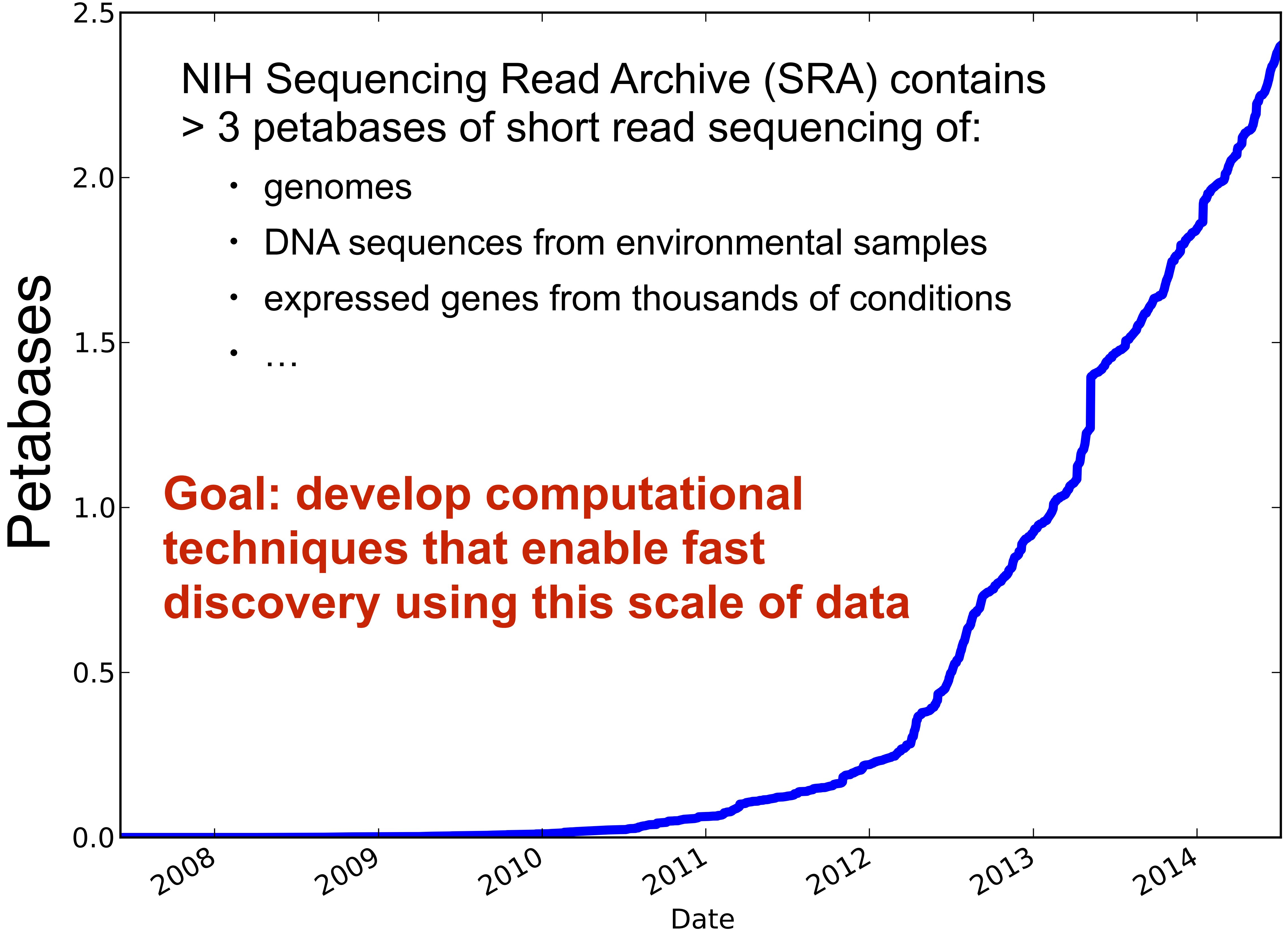
Accurate, Fast, and Model-Aware Transcript Expression Quantification

Carl Kingsford
Associate Professor, Computational Biology Department

Carnegie Mellon University

Joint work with Rob Patro & Geet Duggal

Challenge of Large-Scale Genomics



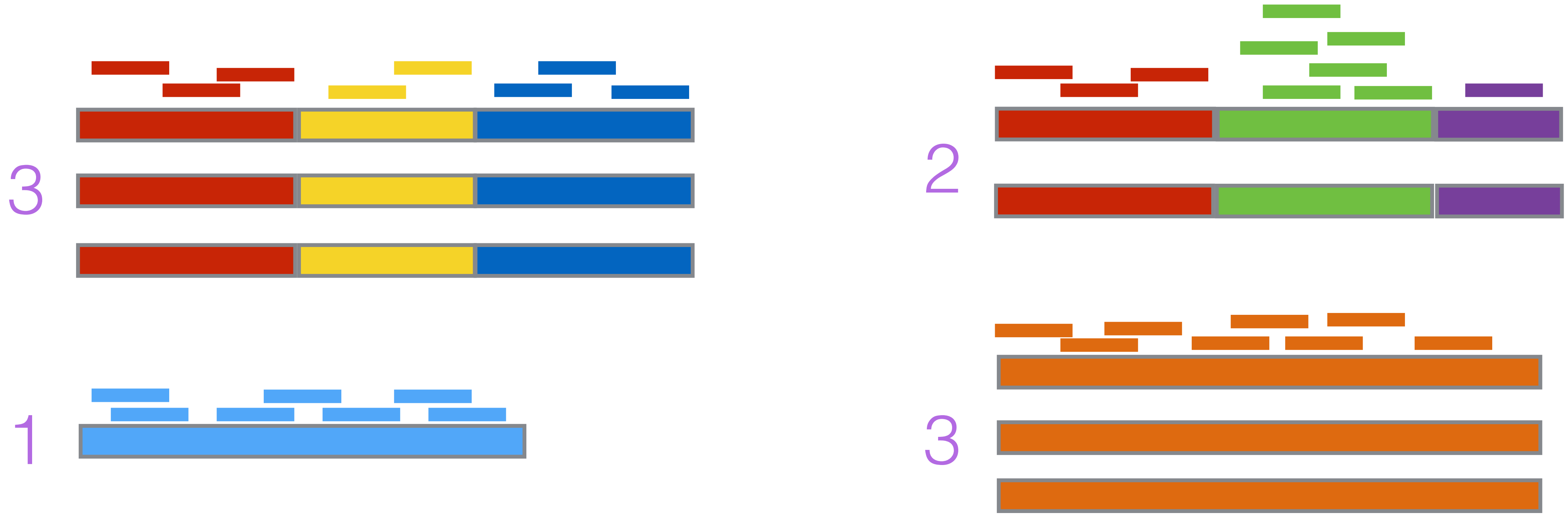
NIH Sequencing Read Archive (SRA) contains
> 3 petabases of short read sequencing of:

- genomes
- DNA sequences from environmental samples
- expressed genes from thousands of conditions
- …

Goal: develop computational techniques that enable fast discovery using this scale of data

Problem: Fast gene expression estimation from RNA-seq

Goal: estimate the abundance of each kind of transcript given short reads sampled from the expressed transcripts.



Challenges:

- hundreds of millions of short reads per experiment
- finding locations of reads (mapping) is traditionally slow
- alternative splicing creates ambiguity about where reads came from
- sampling of reads is not uniform

Why is simple counting not sufficient?

Bad approaches:

Union: treat a gene as the union of its exons
Intersection: treat a gene as the intersection of its exons

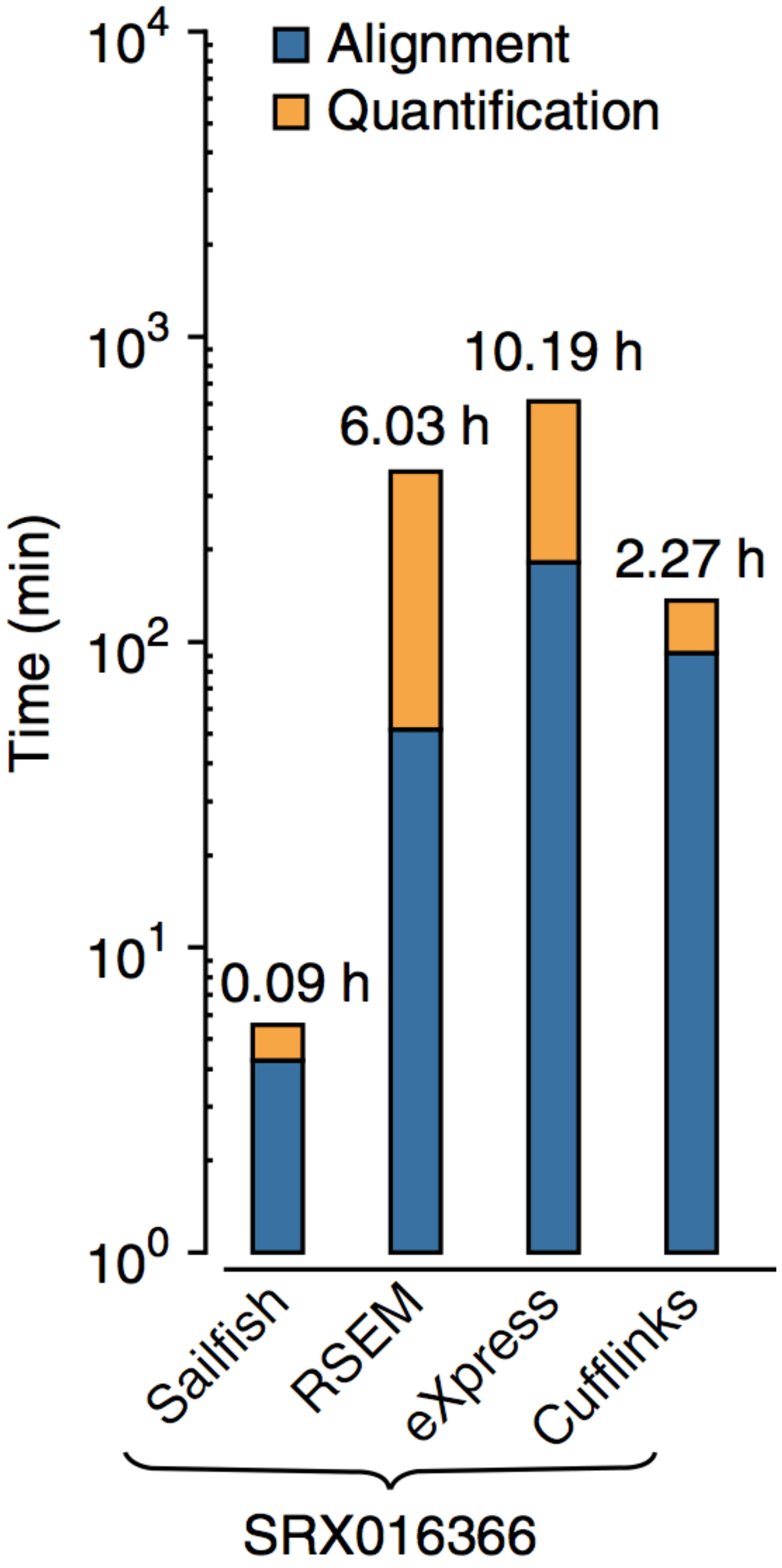
- Can't correct for positional biases / insert length distributions since they don't model which transcript reads come from
- Intersection may throw away many reads

Trapnell et al. "Differential analysis of gene regulation at transcript resolution with RNA-seq." Nature Biotechnology 31 (2013): 46-53.

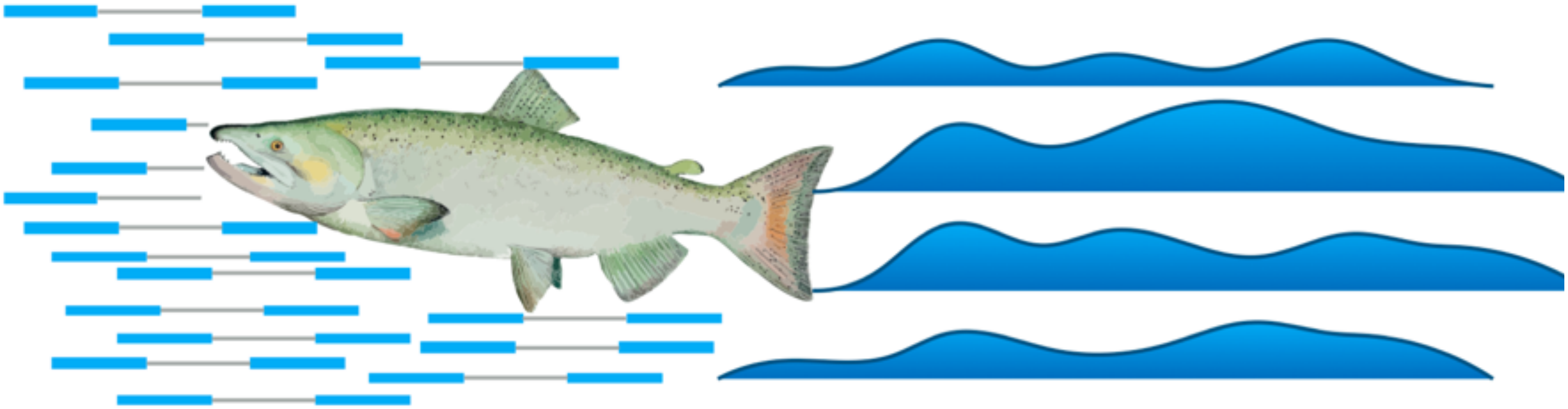
→ Many more sophisticated approaches: Cufflinks (Trapnell, 2010), RSEM (Li, 2010), TIGAR (Nariai, 2014), eXpress (Roberts, 2013), Sailfish (Patro, 2014), Kallisto (Bray, 2015), …

Sailfish: Ultrafast Gene Expression Quantification

- Fast expectation maximization algorithm
- Extremely parallelized
- Uses small data atoms rather than long sequences
- More tolerant of genetic variation between individuals



Patro, Mount, Kingsford, *Nature Biotech*, 2014

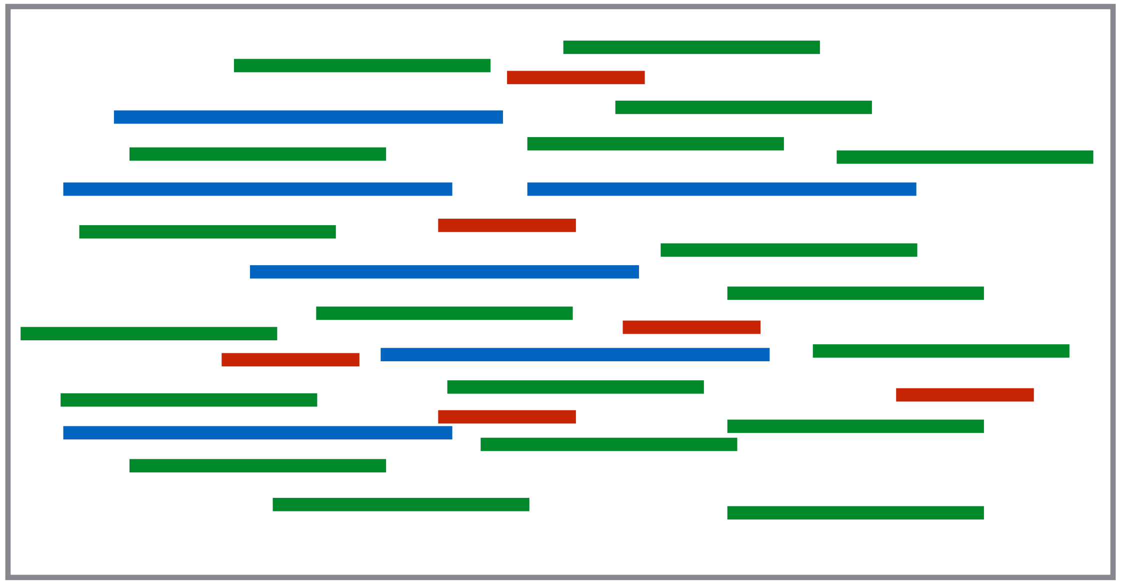


Salmon: fast & accurate method for RNA-seq-based quantification

http://biorxiv.org/content/early/2015/10/03/021592

Inference Problem

Experimental mixture:



length(━━━━━━━━) = 100	x 6 copies	= 600 nt	~ 30% blue
length(━━━━━) = 66	x 19 copies	= 1254 nt	~ 60% green
length(━━━) = 33	x 6 copies	= 198 nt	~ 10% red

These values η = [0.3, 0.6, 0.1] are the *nucleotide fractions*;
they are the quantities we want to infer

Maximum Likelihood Model

nucleotide fractions

true read origins

assumes independence of fragments

$$\Pr\{\mathcal{F} \mid \boldsymbol{\eta}, \boldsymbol{Z}, \mathcal{T}\} = \prod_{j=1}^{N} \Pr\{f_j \mid \boldsymbol{\eta}, \boldsymbol{Z}, \mathcal{T}\}$$

observed fragments (reads)

$$= \prod_{j=1}^{N} \sum_{i=1}^{M} \Pr\{t_i \mid \boldsymbol{\eta}\} \cdot \boxed{\Pr\{f_j \mid t_i, \boldsymbol{z}_{ji} = 1\}}$$

Prob. of selecting t_i *given* **$\boldsymbol{\eta}$**

Prob. of generating fragment f_j *given* t_i

Depends on abundance estimate

Independent of abundance estimate

“Bias” Model

a fragment
starting at given position

$$\Pr\{f_j \mid t_i\} = \Pr\{\ell \mid t_i\} \cdot \Pr\{p \mid t_i, \ell\} \cdot \Pr\{o \mid t_i\} \cdot \Pr\{a \mid f_j, t_i, p, o, \ell\}$$

a fragment
of the given length

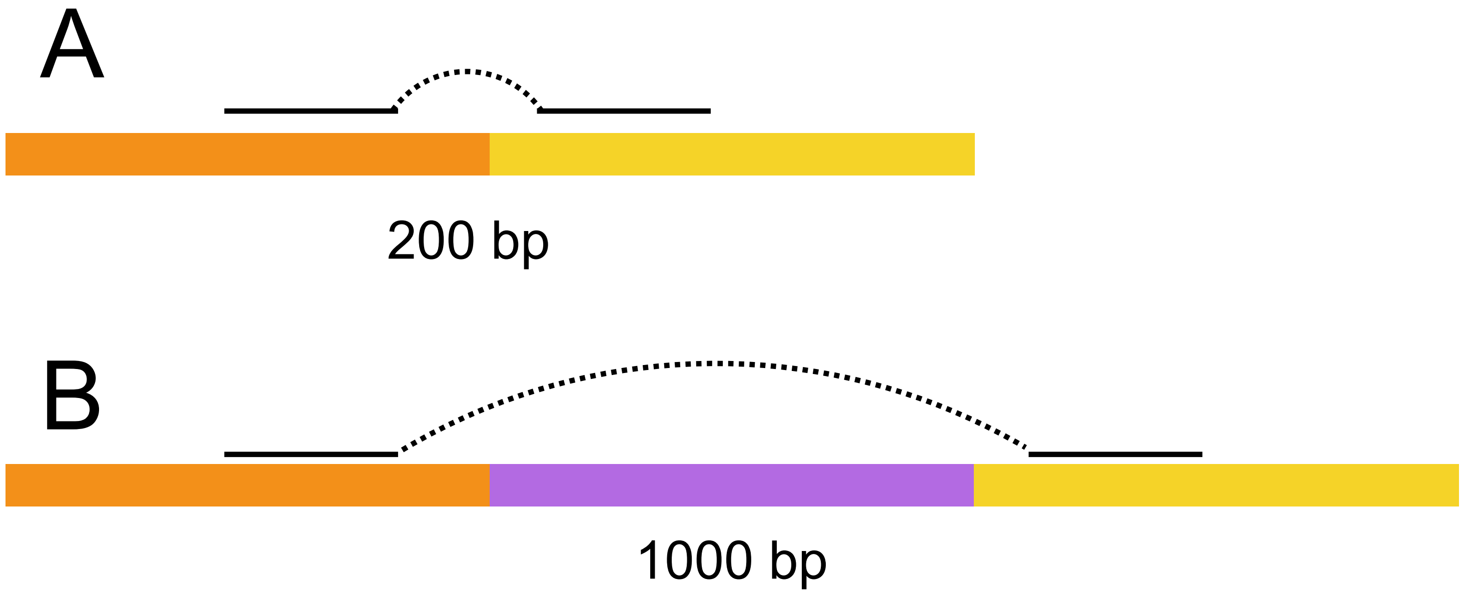
a fragment
of given orientation

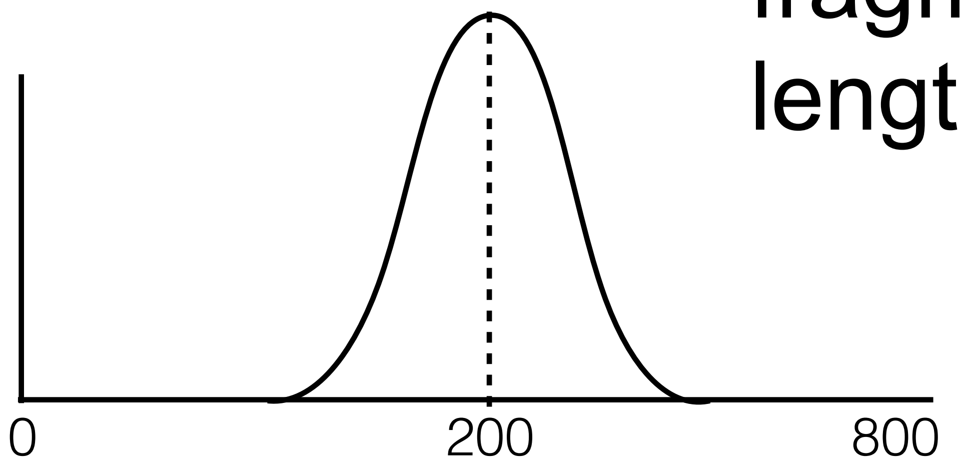
generating the given
alignment

- Salmon estimates an auxiliary model *from the data* for each term (e.g. fragment length, fragment start position, etc.)
- Accounts for sample-specific parameters and biases.

Why does this matter?

“Bias” model can provide strong information about origin of a fragment. For example:



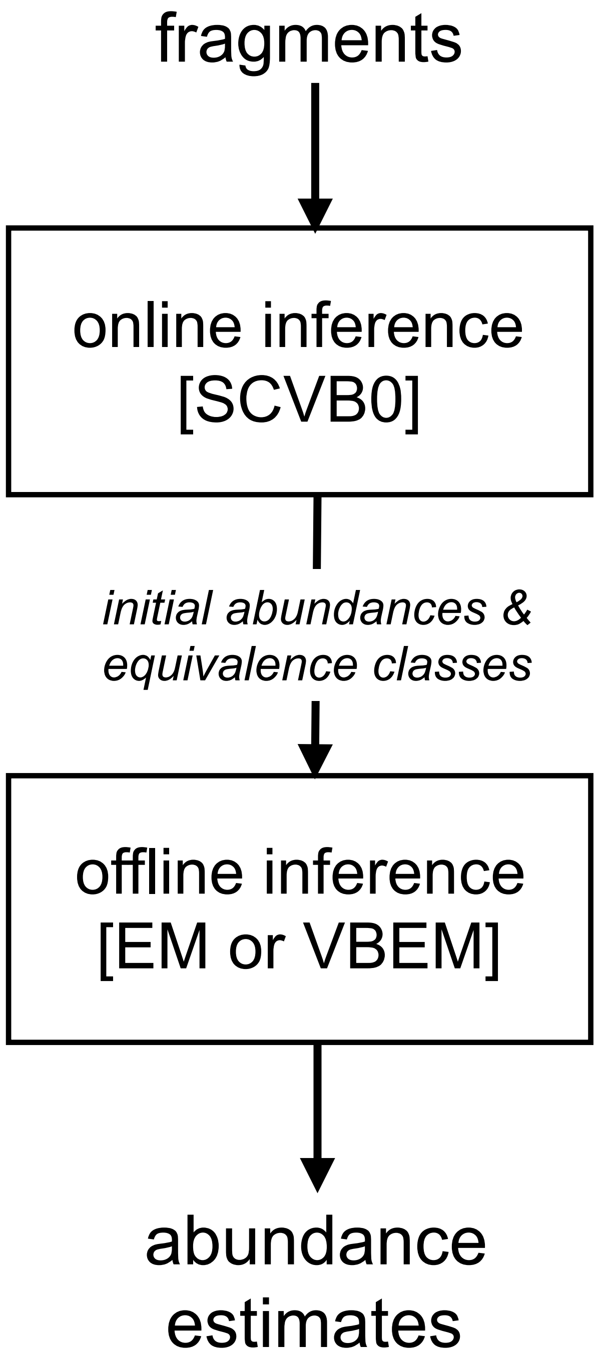


fragment
length distribution

Salmon's two phase inference procedure

Optimizes the full model using a streaming algorithm & trains the "bias" model parameters

Refines the abundance estimates using a reduced representation.


fragments

online inference [SCVB0]

initial abundances & equivalence classes

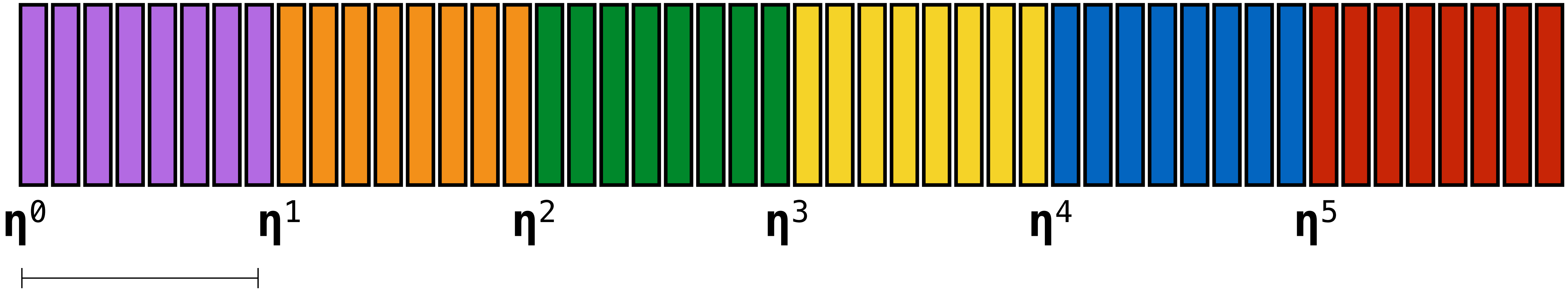
offline inference [EM or VBEM]

abundance estimates

Phase 1: Online Inference

Based on: Foulds et al. Stochastic collapsed variational Bayesian inference for latent Dirichlet allocation. ACM SIGKDD, 2013.

Process fragments in batches:



Compute local $\boldsymbol{\eta}'$ using $\boldsymbol{\eta}^{t-1}$ & current “bias” model to allocate fragments

Update global nucleotide fractions: $\boldsymbol{\eta}^{t} = \boldsymbol{\eta}^{t-1} + a^t \boldsymbol{\eta}'$

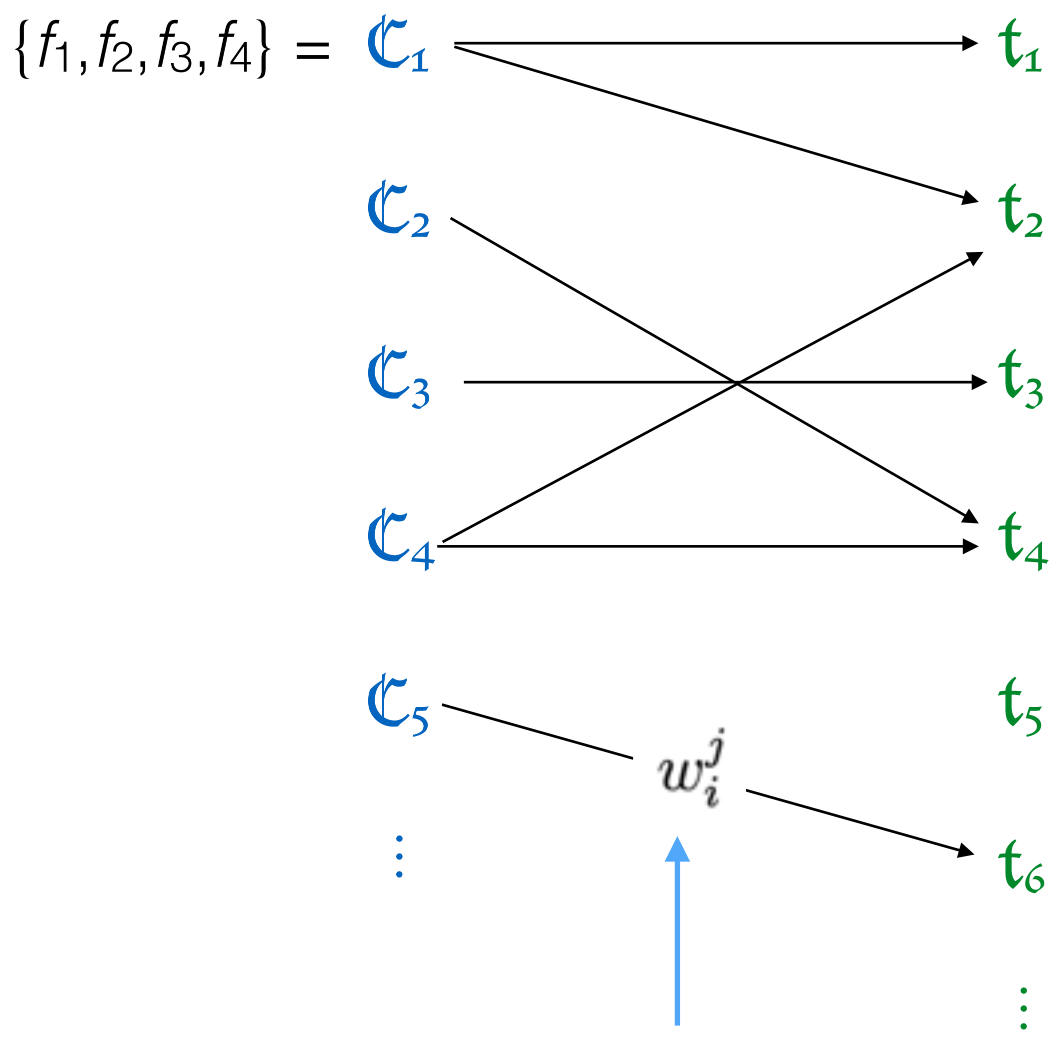
Update “bias” model

Weighting factor that decays over time

Often converges very quickly.
Compare-And-Swap (CAS) for synchronizing updates of different batches

Equivalence Classes & Affinities

Equivalence classes & affinities are computed during the online inference phase.

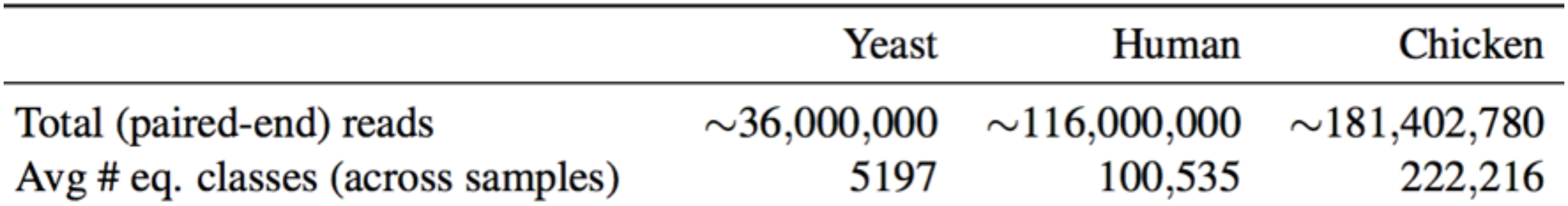


“Affinity” of class *j* to transcript *i* according to the “bias” model.

Two fragments are put into the same equivalence class if they can map to the same set of transcripts.

Affinities encode $\Pr\{f_j \mid t_i\}$ aggregated for all fragments in a class.

Benefit of Equivalence Classes



	Yeast	Human	Chicken
Total (paired-end) reads	∼36,000,000	∼116,000,000	∼181,402,780
Avg # eq. classes (across samples)	5197	100,535	222,216

The # of equivalence classes grows with the complexity of the transcriptome — independent of the # of sequence fragments.

Typically, many fewer equivalence classes than sequenced fragments.

The time for the offline inference algorithm scales in # of equivalence classes.

Phase 2: Offline Inference

Repeatedly reallocate fragments according to current abundance estimates & “bias” model until convergence:

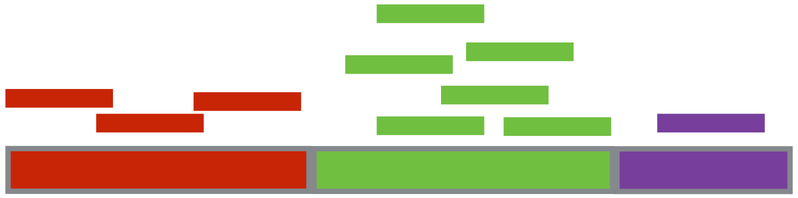
size of equivalence class *j*

reads are allocated ∝ current estimate weighted by affinity

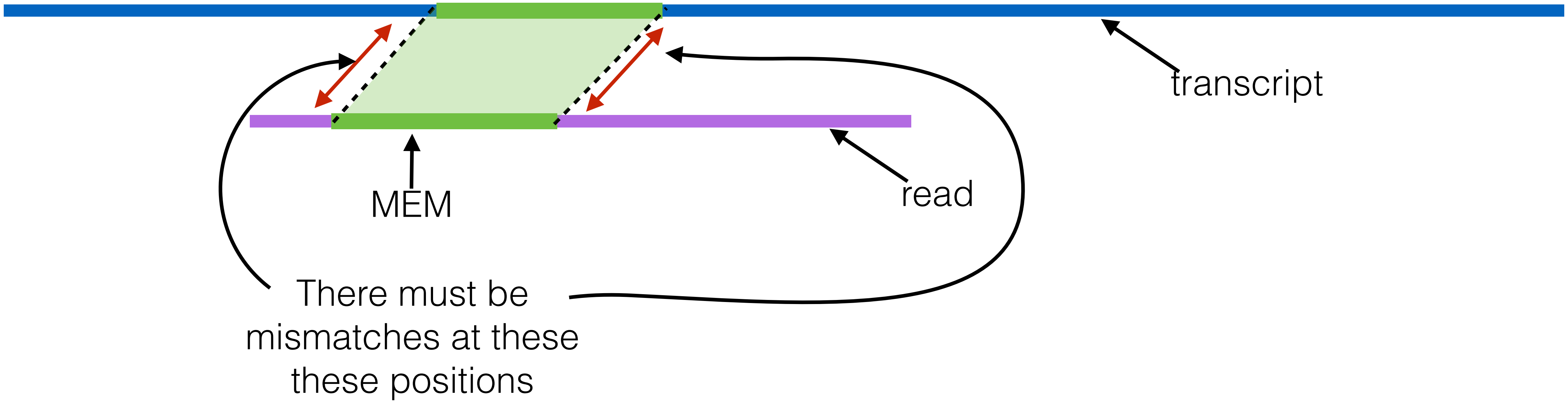
$$\alpha_i^{u+1} = \sum_{c^j \in \boldsymbol{c}} \boldsymbol{d}^j \left(\frac{\alpha_i^u w_i^j}{\sum_{t_k \in \boldsymbol{t}^j} \alpha_k^u w_k^j} \right)$$

# of reads assigned to transcript *i*

Lightweight alignment



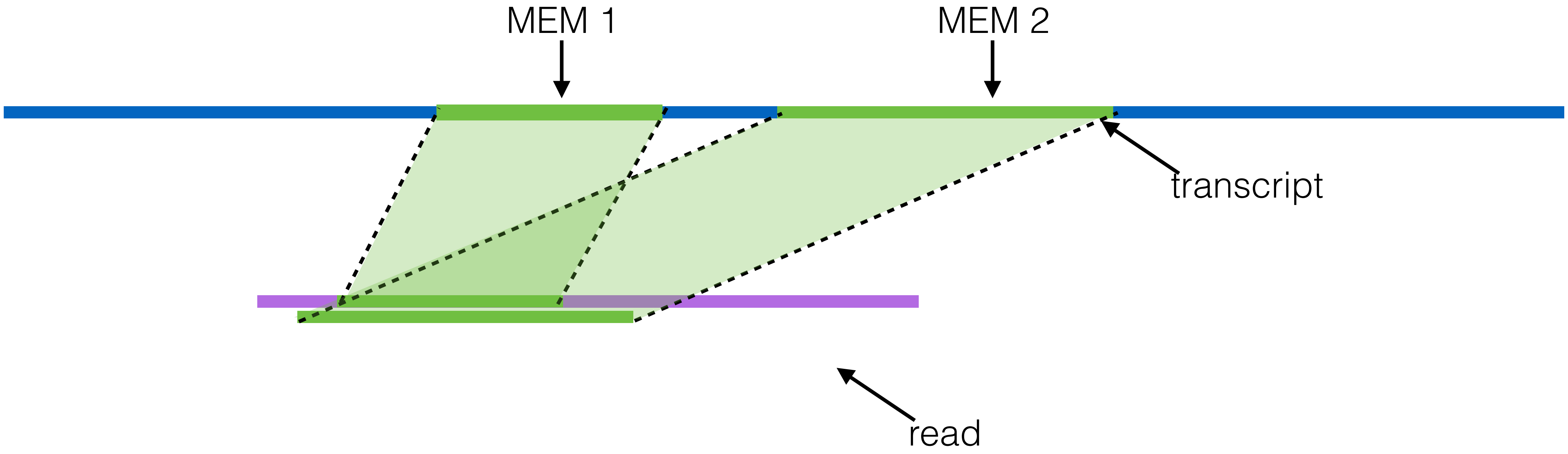
- Salmon replaces the time-consuming read alignment step with a new approach that quickly finds chains of “maximal exact matches”:



A maximal exact match is an exact match between the read and a transcript that can’t be extended in either direction.

SMEMs

A **super** maximal exact match (Li, 2013) is a MEM that is not contained in any other MEM in either the query or the reference:

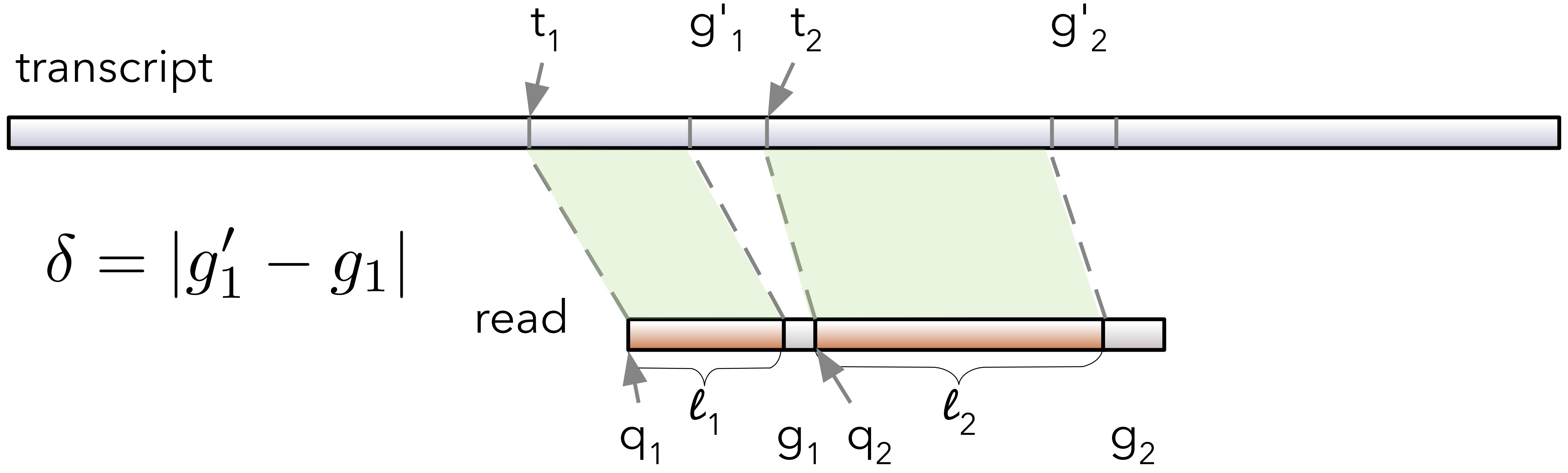


MEM 1 is not an SMEM, while MEM 2 is.

Lightweight alignment

Lightweight alignment looks for δ-consistent chains of SMEMs.

A chain of SMEMs is δ-consistent if the total difference in gap sizes between the SMEMs is ≤ δ



Salmon requires the SMEMs to cover at least 65% of the read.

Revising the Challenges

- finding locations of reads (mapping) is traditionally slow

→ Use lightweight alignment

- alternative splicing creates ambiguity about where reads came from

→ Use 2-phase EM inference algorithm

- sampling of reads is not uniform

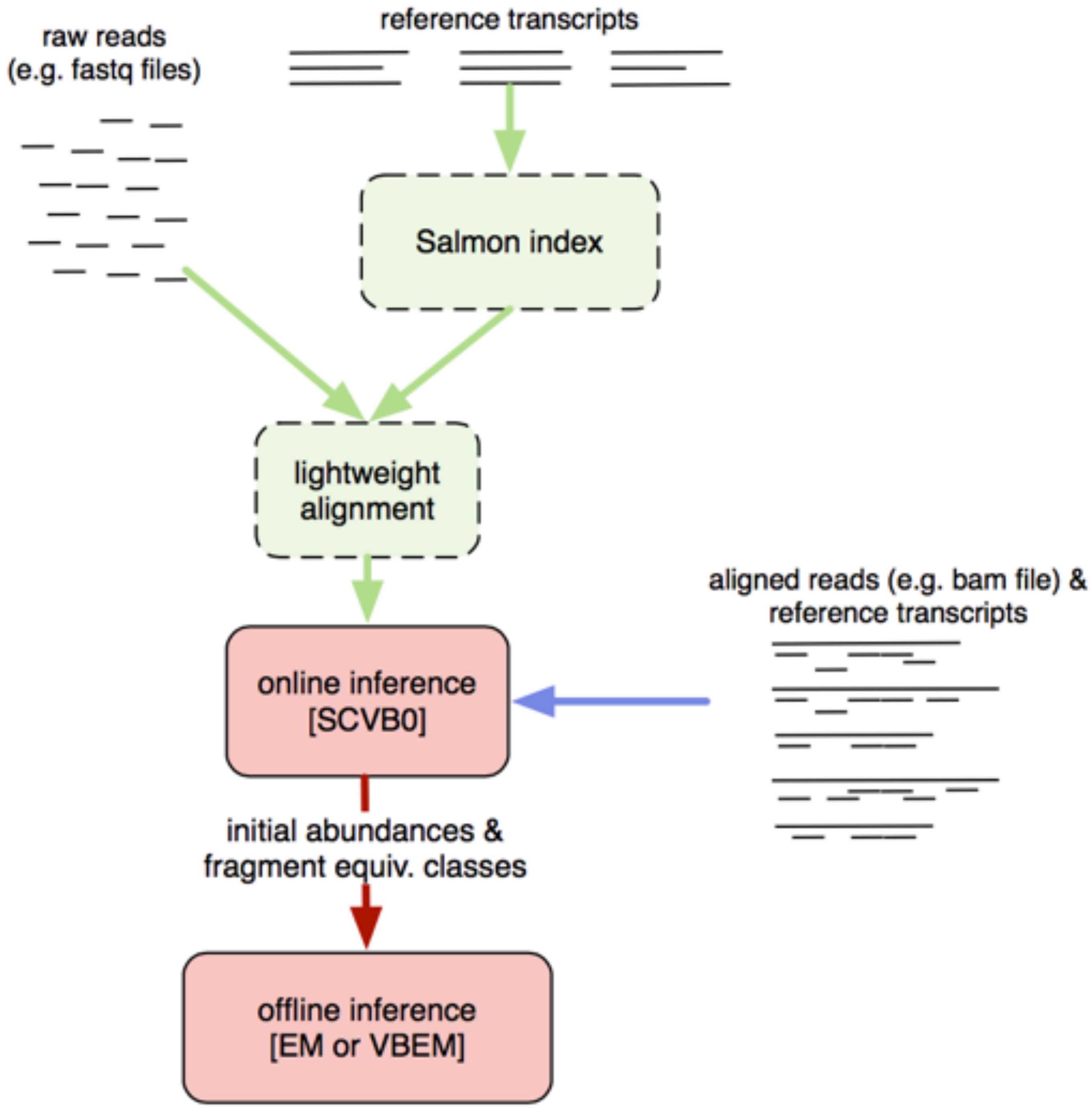
→ Use bias model learned from data

Other Salmon Features

- Variational Bayes inference procedure as an option

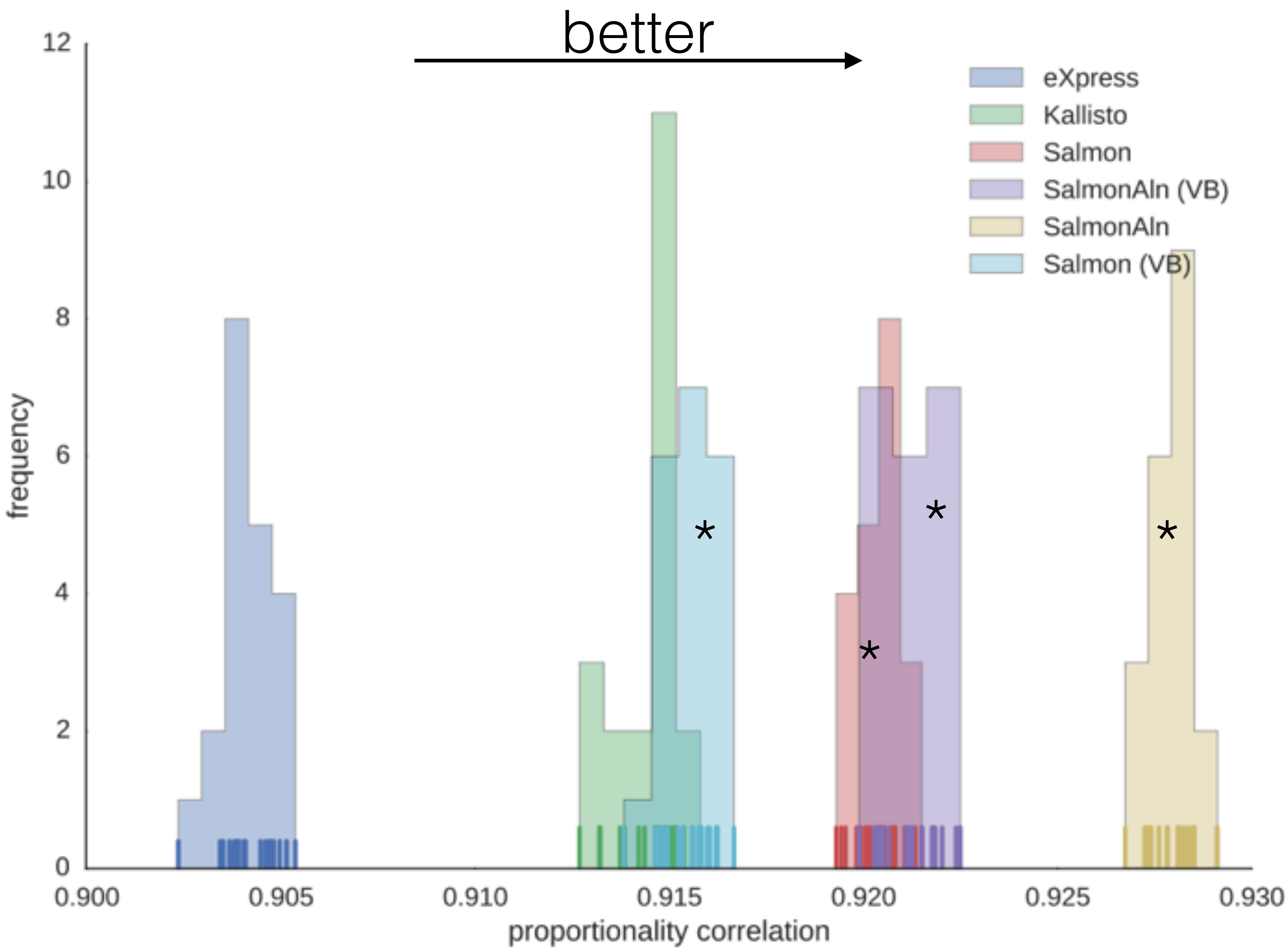
- Can provide your own alignments if you want (SalmonAln)

- Several “fast” alignment modes (not just the one based on SMEMs)


raw reads
(e.g. fastq files)
reference transcripts
Salmon index
lightweight alignment
aligned reads (e.g. bam file) & reference transcripts
online inference [SCVB0]
initial abundances & fragment equiv. classes
offline inference [EM or VBEM]

Salmon is Accurate

Human reads simulated with RSEM-sim:



Salmon is Accurate

Reads simulated with FluxSim (Griebel et al., 2012):

H. *sapiens*	Salmon	SalmonAln	eXpress	Kallisto
Proportionality corr.	**0.79**	0.76	0.75	0.76
Spearman corr.	0.73	0.7	0.63	**0.79**
MARD	**0.14**	0.19	0.25	0.2

Z. *mays*	Salmon	SalmonAln	eXpress	Kallisto
Proportionality corr.	**0.92**	0.91	0.89	0.91
Spearman corr.	**0.91**	0.90	0.85	0.89
MARD	**0.17**	0.19	0.34	0.20

Proportionality Correlation

$$\rho_p = \frac{2\mathrm{Cov}\{\log \boldsymbol{x}, \log \boldsymbol{y}\}}{\mathrm{Var}\{\log \boldsymbol{x}\} + \mathrm{Var}\{\log \boldsymbol{y}\}}$$

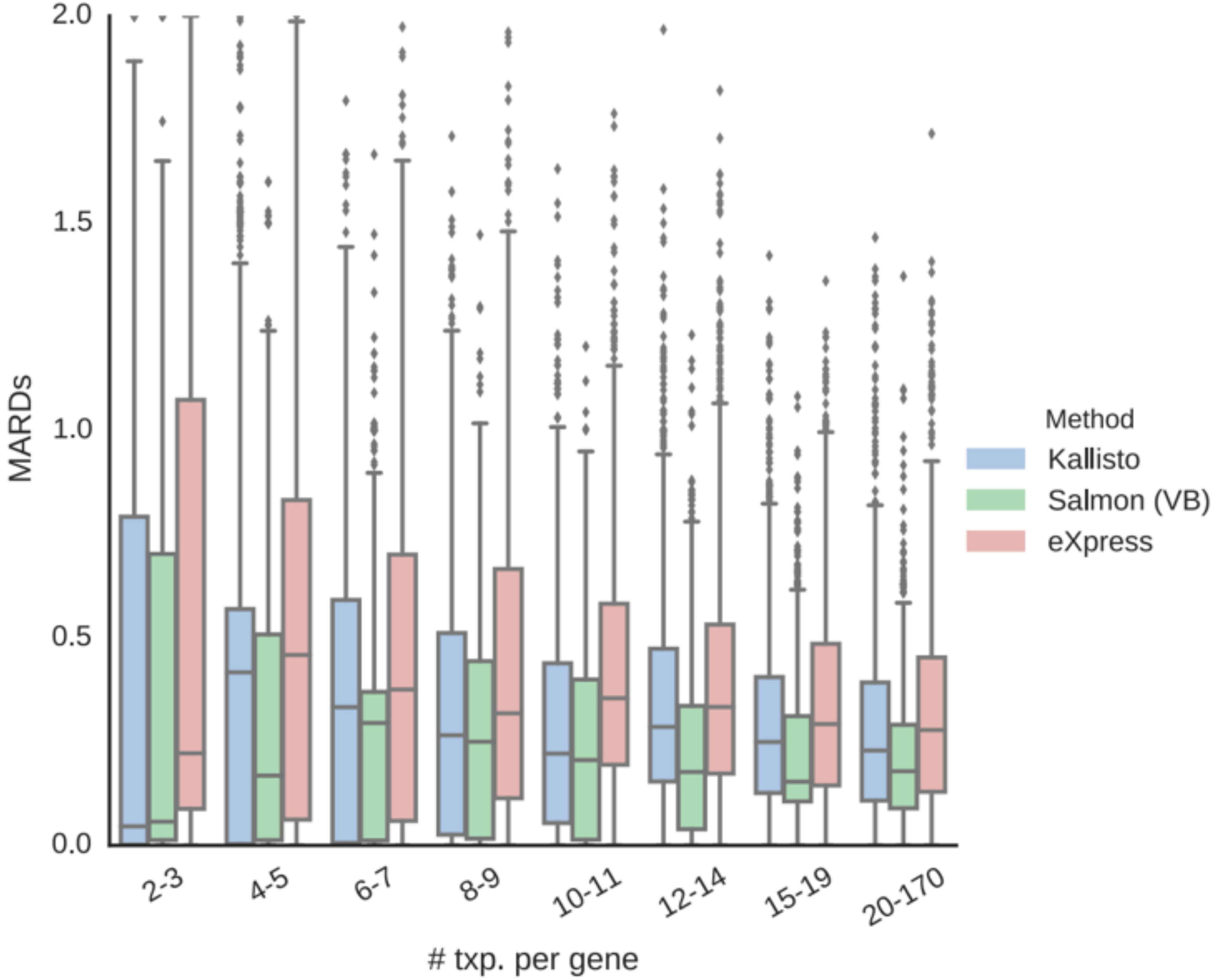
Lovell et al. argue this is
good for relative quantities

MARD

$$\mathrm{MARD} = \frac{1}{M}\sum_{i=1}^{M} \mathrm{ARD}_i.$$

$$\mathrm{ARD}_i = \begin{cases} 0 & \text{if } x_i = y_i = 0 \\ \frac{|x_i - y_i|}{0.5|x_i + y_i|} & \text{otherwise} \end{cases},$$

Salmon is accurate when there are many isoforms



GC “Bias” model → more accurate differential expression

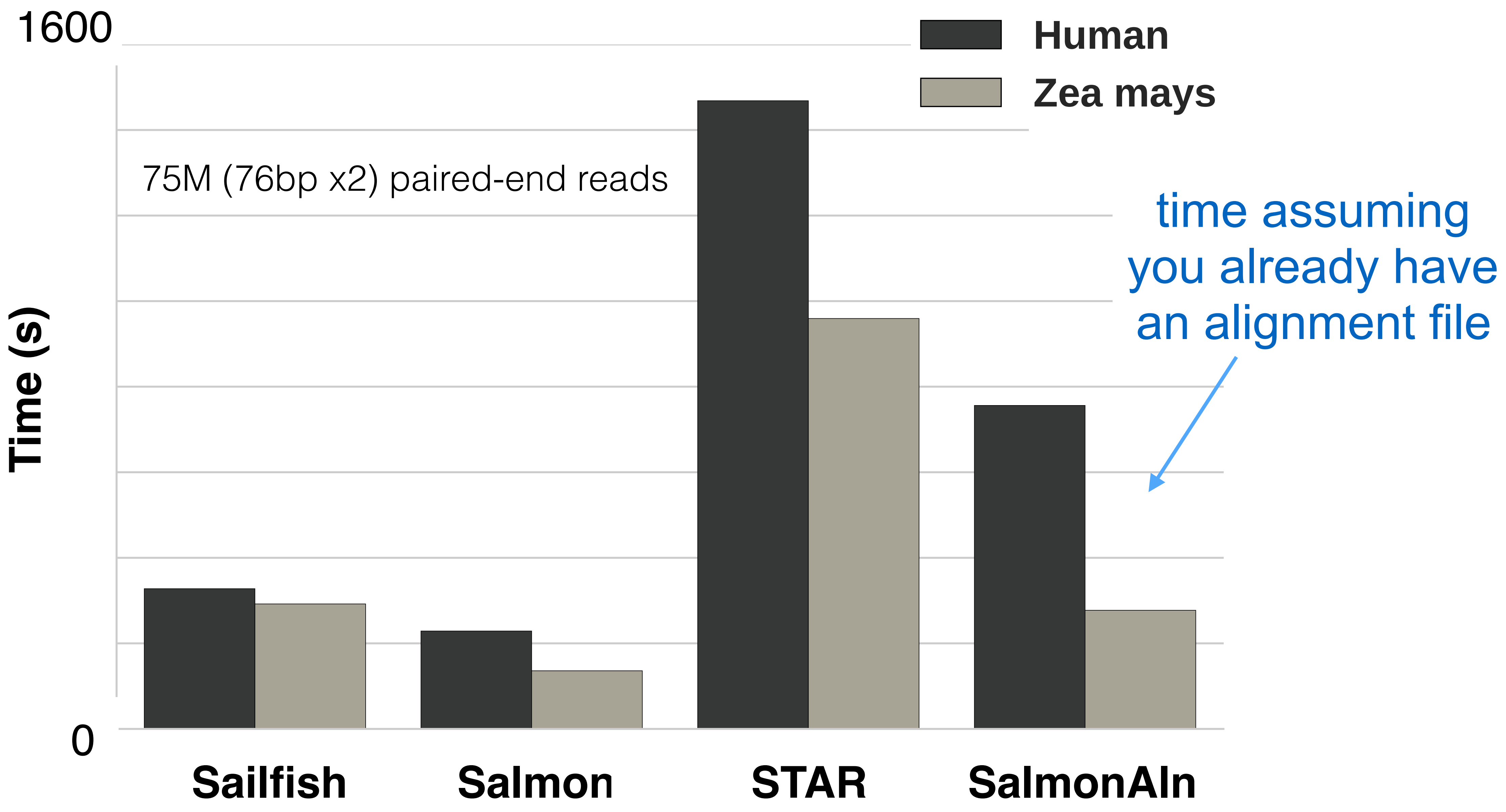
30 samples from Lappalainen et al. (2013):
15 samples from UNIGE sequencing center
15 samples from CNAG_CRG sequencing center
All same population (TSI) and cell type (lymphoblastoid)

DE of data between centers (FDR < 1%) (TPM > 0.1)

	Salmon	RSEM	Kallisto	Cufflinks
All genes	**1,325**	2,829	2,826	2,510
2-isoform genes	**225**	577	548	562

Courtesy Michael Love.
http://biorxiv.org/content/early/2015/08/28/025767

Salmon is Fast



time assuming you already have an alignment file

Both datasets take ~**5 min** using 16 threads on a 2.6GHz Xeon; **including lightweight alignment**.

Conclusion

- Salmon is a fast, accurate, flexible way to quantify expression from RNA-seq data.
- Expressive model means new types of bias can be learned and accounted for.
- Open source:

  Code: https://github.com/COMBINE-lab/salmon

  News: http://combine-lab.github.io/salmon/

  User group: https://groups.google.com/forum/#!forum/sailfish-users

Acknowledgements & Thanks!

This work:

Rob Patro, Ph.D. ← Salmon
Geet Duggal, Ph.D. ← Salmon
Brad Solomon ← SBT

Gordon and Betty Moore Foundation's Data-Driven Discovery Initiative through grant GBMF4554

Current group (who all help with feedback):

Guillaume Marçais, Ph.D.
Mingfu Shao, Ph.D.
Heewook Lee, Ph.D.
Hao Wang
Tim Wall
Natalie Sauerwald
Cong Ma

NIH R01 HG007104 (3D Genome Structure)
NIH R21 HG006913 (Jellyfish & Sailfish)

NSF CCF-1256087 (Metagenomics)
NSF CCF-1053918 (CAREER)

2012 Alfred P. Sloan Research Fellowship