Accurate and **Fast** Transcript (and gene) Quantification

Rob Patro

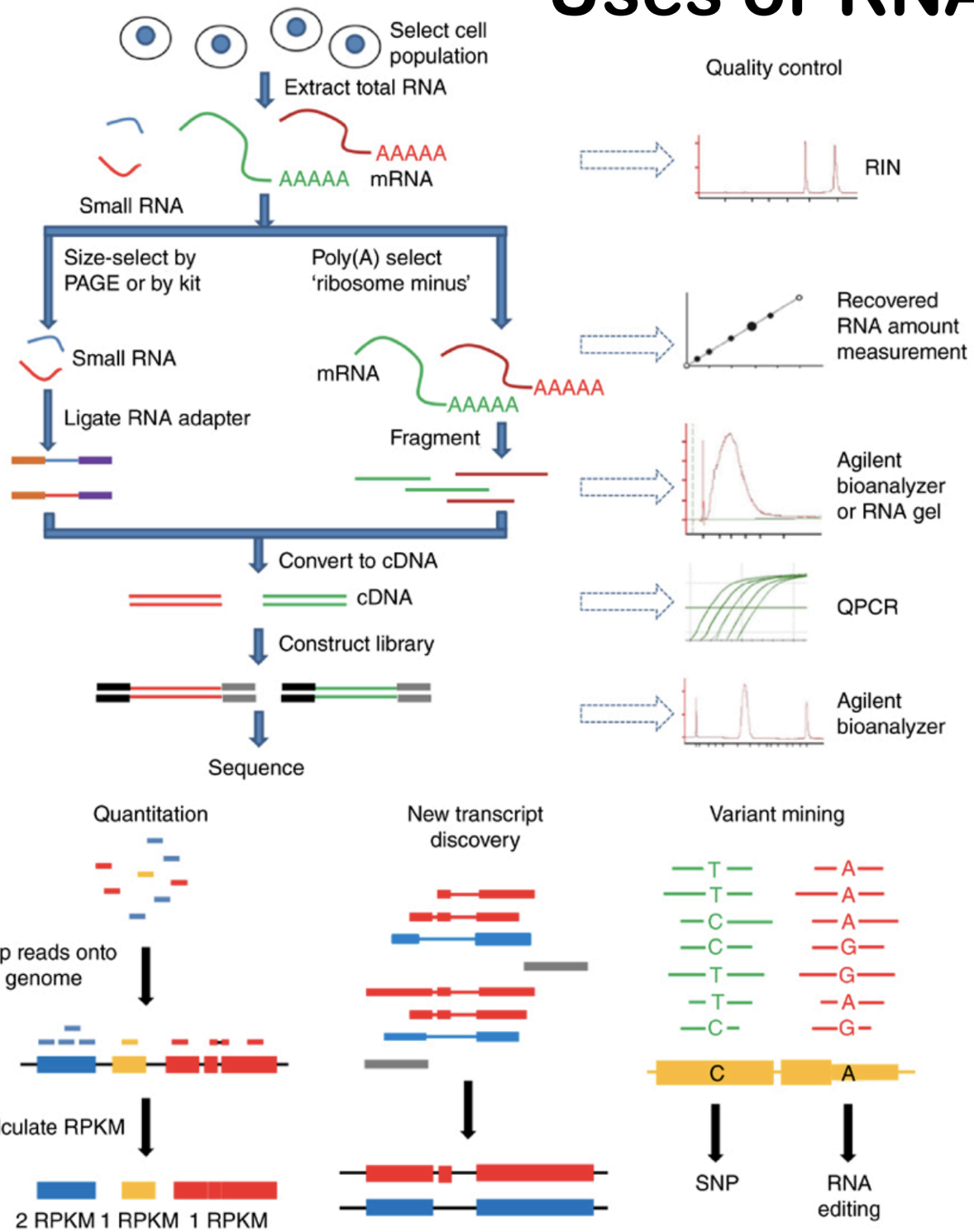
Stony Brook University

ANGUS 2016

Aug. 16 2016

Zeng & Mortazavi, Nature Immunology 2012

Uses of RNA-Seq are manifold



Whole transcriptome analysis

- Quantification & differential expression
- Novel txp discovery
  - reference-based
  - *de novo*
- Variant detection
  - Genomic SNPs
  - RNA editing

- What is dynamic & changing over time (as disease progresses)?
- What is tissue specific (in fetal development but not after)?
- What is condition specific (under stress conditions vs. not)?

Sequencing Reads

pre-proc.

align to ref.

txp. identification

reference-based

pre-proc.

de novo assembly

align to assembly

de novo

quantification

Today

DE, Alternative Splicing, etc.

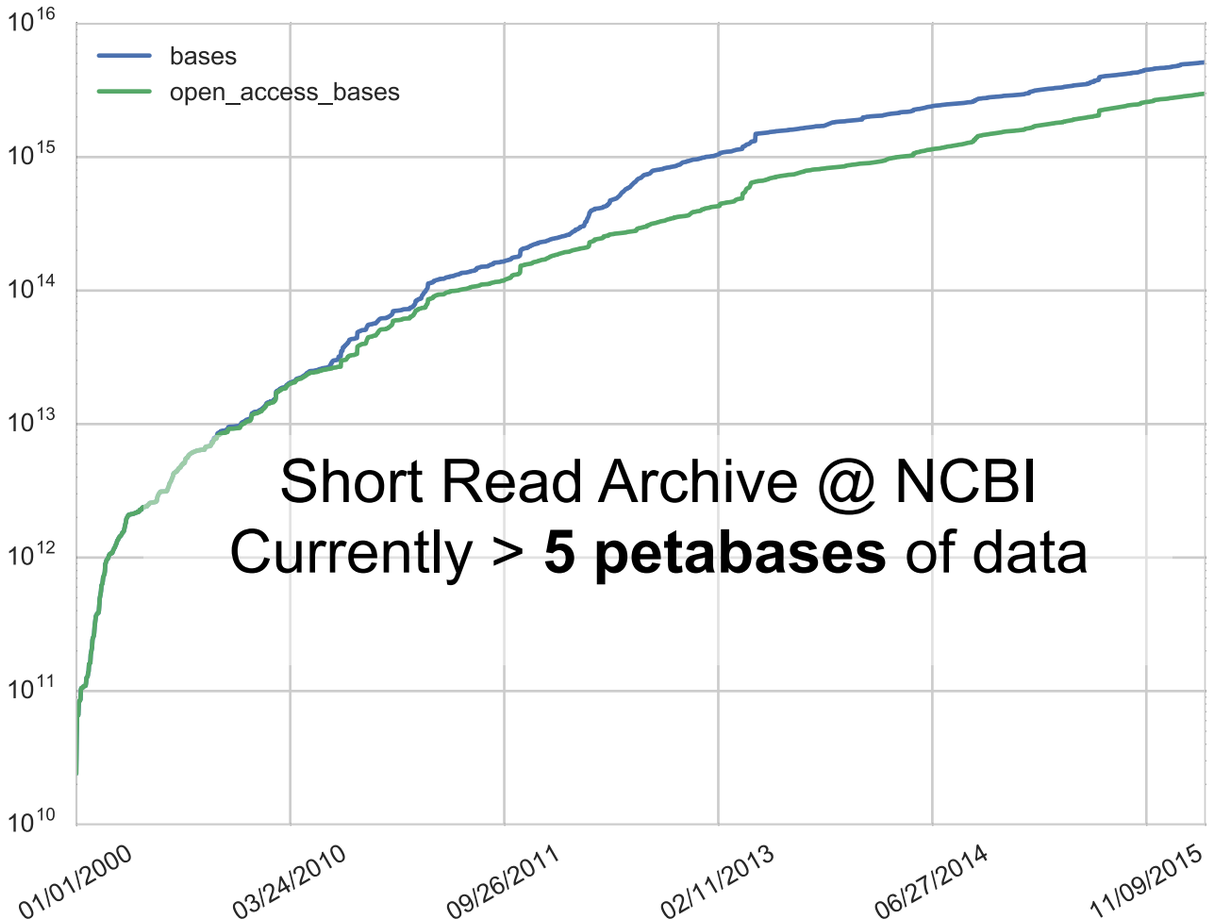
“Higher-level” analysis

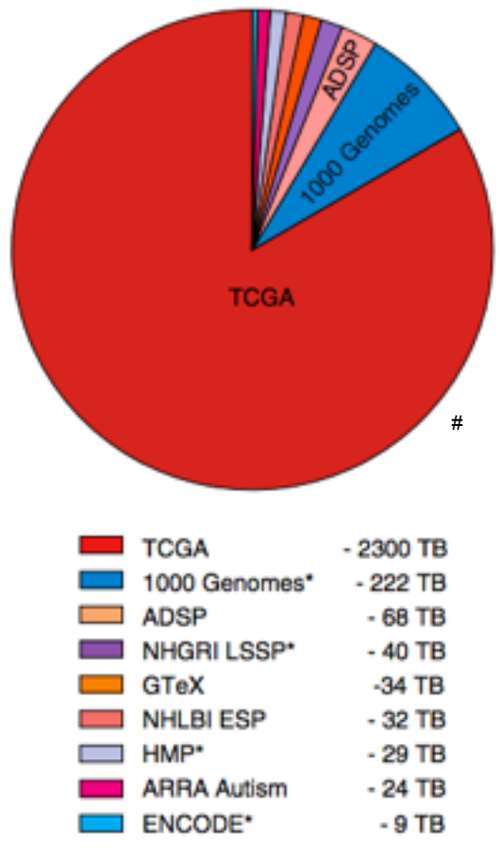
Why do we still need faster analysis?

OPINION Open Access

The real cost of sequencing: scaling computation to keep pace with data generation

Paul Muir[1,2,3], Shantao Li[4], Shaoke Lou[4,5], Daifeng Wang[4,5], Daniel J Spakowicz[4,5], Leonidas Salichos[4,5], Jing Zhang[4,5], George M. Weinstock[6], Farren Isaacs[1,2], Joel Rozowsky[4,5] and Mark Gerstein[4,5,7*]





In addition to new data, re-analysis of existing experiments often desired: In light of new annotations, discoveries, and methodological advancements.

From Fig 1 of Muir et al.

Advocating for analysis-efficient computing

- Compute *only* the information required for your analysis; ask what information you *need* to solve your problem, not what output current tools are generating
- Often the efficiency of the analysis is related to the *size* of the (processed) data's representation
- Not all analyses require such efficient solutions, should concentrate on problems where this is actually needed.

I'll provide some (hopefully) compelling examples:

- **RapMap**: Read alignment → quasi-mapping (get "core" info much faster)
- **Salmon**: Fast, state-of-the-art quantification using quasi-mapping, dual-phase inference & fragment eq. classes
- **RapClust**: Fast, accurate *de novo* assembly clustering using quasi-mapping & fragment eq. classes

We believe these ideas are **general**, and can be applied to many problems

Advocating for analysis-efficient computing

- Compute *only* the information required for your analysis; ask what information you *need* to solve your problem, not what output current tools are generating
- Often the efficiency of the analysis is related to the *size* of the (processed) data's representation
- Not all analyses require such efficient solutions, should concentrate on problems where this is actually needed.

I'll provide a (hopefully) compelling example:

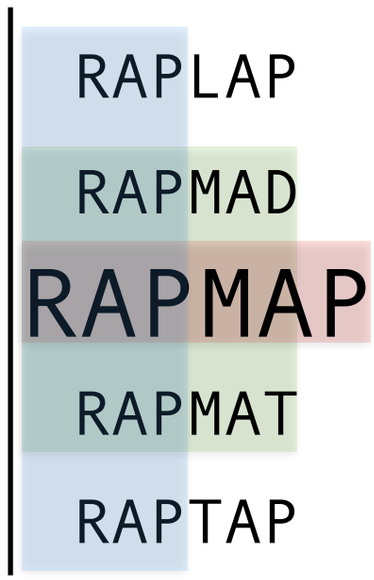
Boiler is also a beautiful example of this idea.

When we have a particular analysis in mind — transcript identification & quantification — we can compress data much more aggressively & effectively.

We believe these ideas are **general**, and can be applied to many problems

I promised to show how we can use this yesterday …

RapMap: A Rapid, Sensitive and Accurate Tool for Mapping RNA-seq Reads to Transcriptomes



GitHub repository: https://github.com/COMBINE-lab/RapMap

Paper: http://bioinformatics.oxfordjournals.org/content/32/12/i192.full.pdf

(appeared at ISMB 16)

co-authors (students): Avi Srivastava, Hirak Sarkar, Nitish Gupta

Where might we use quasi-mapping?

We believe there are ***many*** places where this replacement can be made. I'll discuss one in some depth (and mention a second):

1)Transcript-level quantification

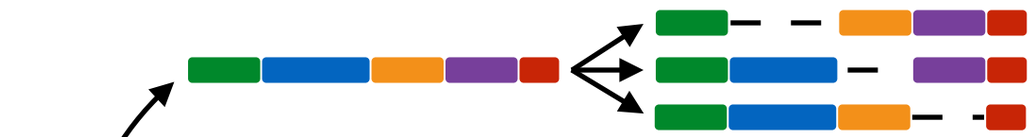
- Determine abundance of transcripts from a collection of RNA-seq reads.
- The quasi-mapping information is sufficient to yield estimates *as accurate as full alignment*.

2)*de novo* transcript clustering

- Find groups of related contigs likely from the same transcript / gene
- Such groups help improve downstream analysis (e.g. differential expression testing)

Obviously, alignments are *necessary* for certain types of analysis (e.g. variant detection).

Transcript Quantification: An Overview



1 gene ⇒ many variants (isoforms)

Gene 1

Gene M

Sample

10s-100s of millions of
short (35-300 character) “fragments”

Measurement
(RNA-seq)

Inference
(e.g. Salmon)

% Gene 1

% Gene M

isoform A

isoform B

isoform C

Abundance Estimates

Given: (1) Collection of RNA-Seq fragments
(2) A set of known (or assembled) transcript sequences

Estimate: The relative abundance of each transcript

Given: (1) Collection of RNA-Seq fragments
(2) A set of known (or assembled) transcript sequences

Estimate: The relative abundance of each transcript

Question: If we only care about “gene" abundance, can’t we just count the number of reads mapping / aligning to each gene?

Given: (1) Collection of RNA-Seq fragments
(2) A set of known (or assembled) transcript sequences

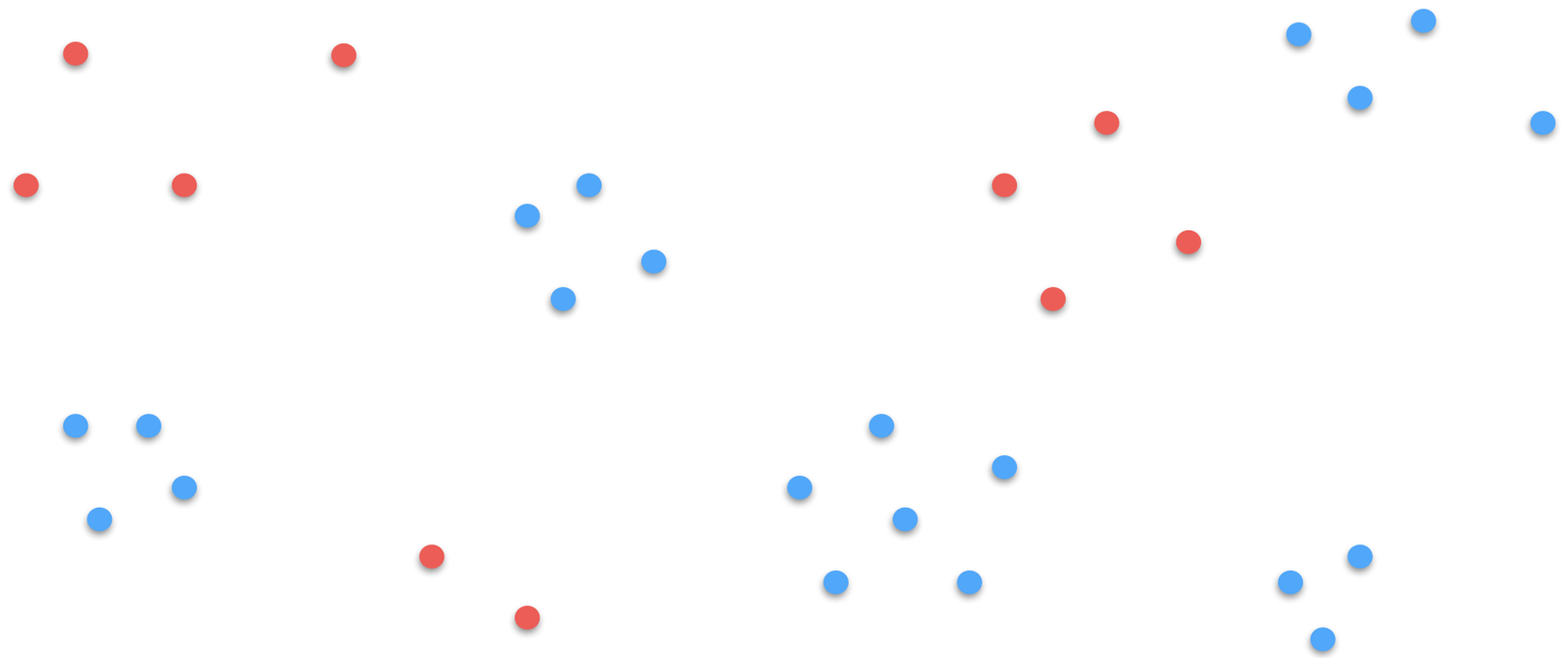
Estimate: The relative abundance of each transcript

Question: If we only care about “gene" abundance, can’t we just count the number of reads mapping / aligning to each gene?

Answer: **No.** I’ll show a general argument (and a few examples) why!

First, consider this non-Biological example

Imagine I have two colors of circle, red and blue. I want to estimate the **fraction of circles** that are red and blue. I'll *sample* from them by tossing down darts.



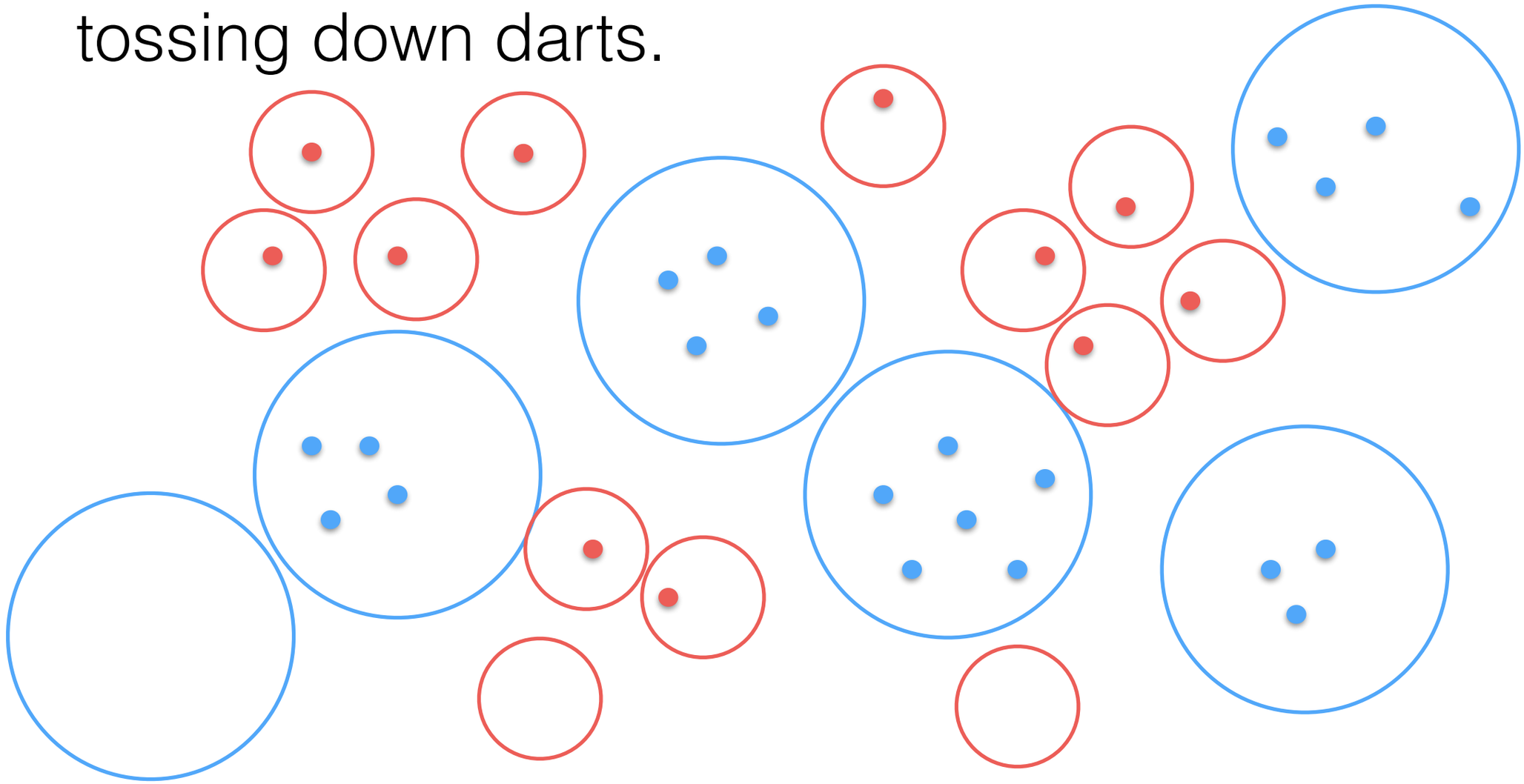
Here, a dot of a color means I hit a circle of that color.

What type of circle is more prevalent?

What is the fraction of red / blue circles?

First, consider this non-Biological example

Imagine I have two colors of circle, red and blue. I want to estimate the **fraction of circles** that are red and blue. I'll *sample* from them by tossing down darts.



You're missing a **crucial piece of information!**

The areas!

First, consider this non-Biological example

Imagine I have two colors of circle, red and blue. I want to estimate the **fraction of circles** that are red and blue. I'll *sample* from them by tossing down darts.

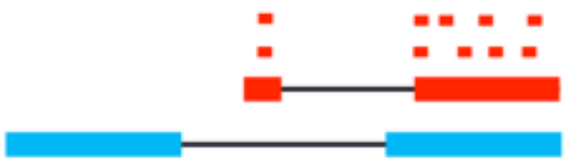
You're missing a **crucial piece of information!**

The areas!

There is an analog in RNA-seq, one needs to know the **length** of the target from which one is drawing to meaningfully assess abundance!

Resolving multi-mapping is fundamental to quantification

Isoform A is half
as long as isoform B

Condition 1	Condition 2	union-model fold-change		true fold-change
		$\log_2\left(\frac{10}{10}\right) = 0$	0 < 0.32	$\log_2\left(\frac{\frac{10}{L}}{\frac{6}{L}+\frac{4}{2L}}\right) = 0.32$
		$\log_2\left(\frac{6}{8}\right) = -0.41$	-0.41 < 0.58	$\log_2\left(\frac{6/L}{8/2L}\right) = 0.58$
		$\log_2\left(\frac{5}{10}\right) = -1$	-1 < 0	$\log_2\left(\frac{\frac{5}{L}}{\frac{10}{2L}}\right) = 0$

Key point : The length of the *actual molecule* from which the fragments derive is crucially important to obtaining accurate abundance estimates.

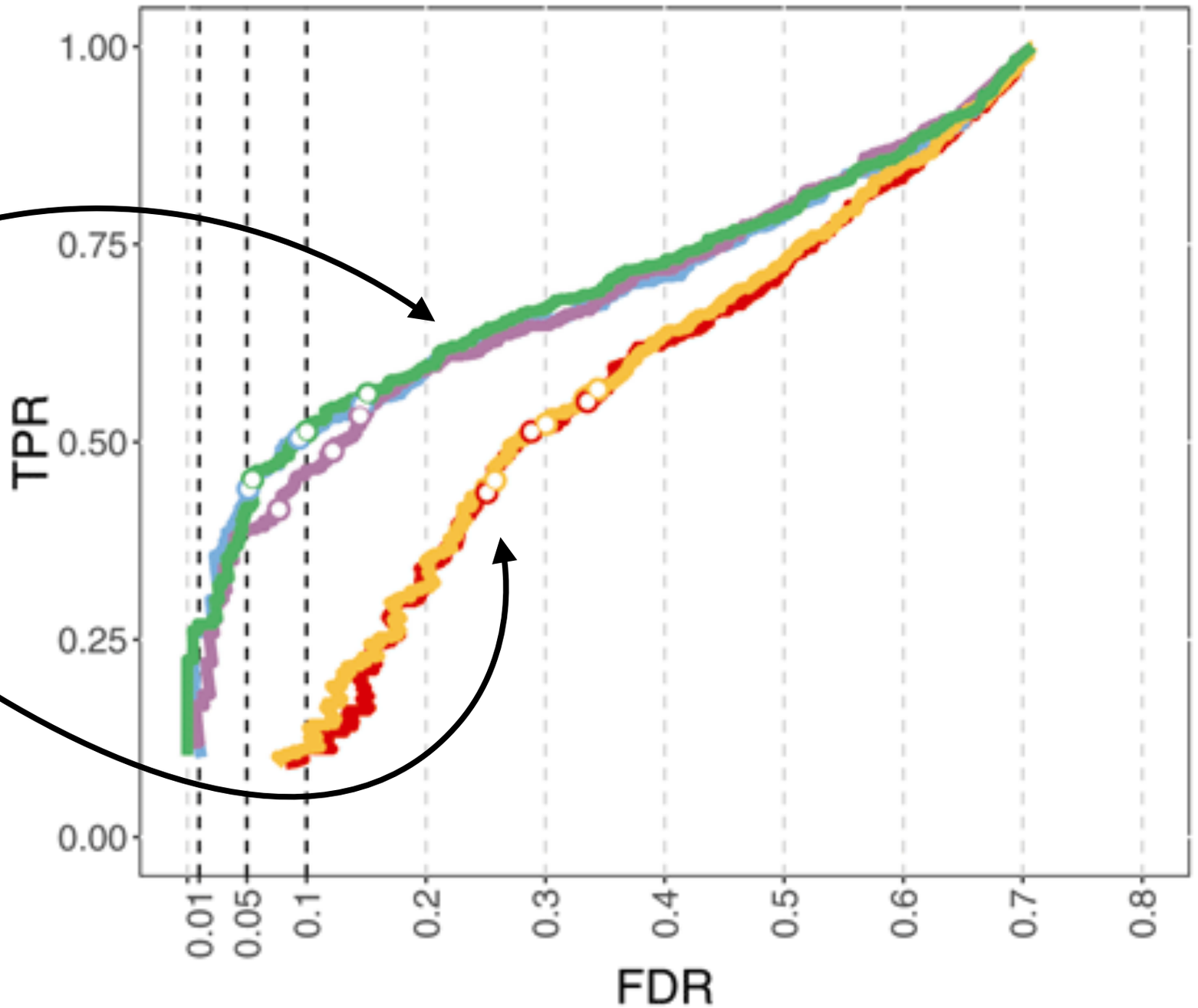
Adapted from: Trapnell, Cole, et al. "Differential analysis of gene regulation at transcript resolution with RNA-seq." Nature biotechnology 31.1 (2013): 46-53.

Resolving multi-mapping is fundamental to quantification

These errors can affect DGE calls

Variants of Salmon

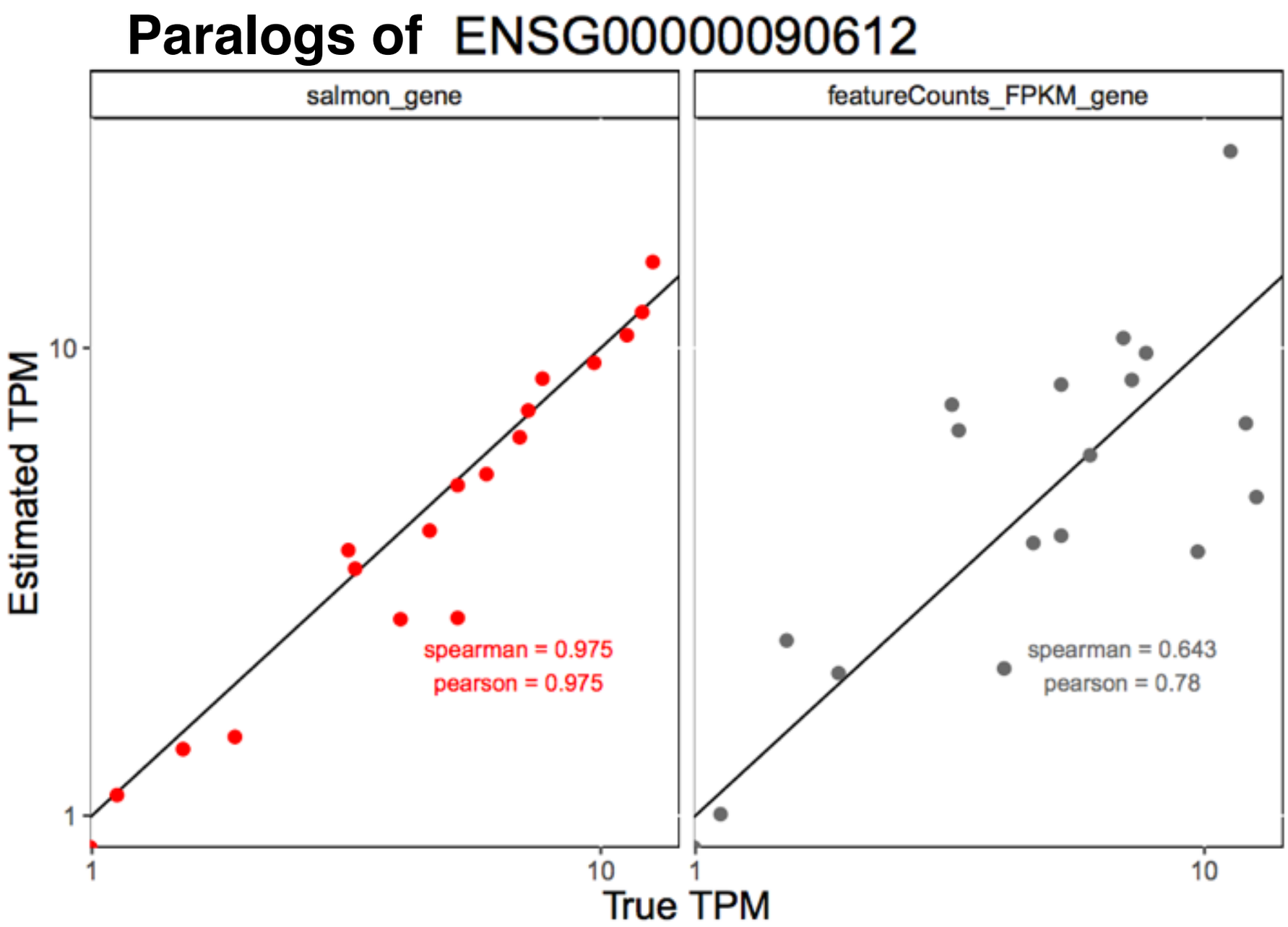
Variants of “counting”



From: Soneson C, Love MI and Robinson MD 2016 [version 2; referees: 2 approved] F1000Research 2016, 4:1521 (doi: 10.12688/f1000research.7563.2)

Resolving multi-mapping is fundamental to quantification

Can even affect abundance estimation in **absence** of alternative-splicing
(e.g. paralogous genes)

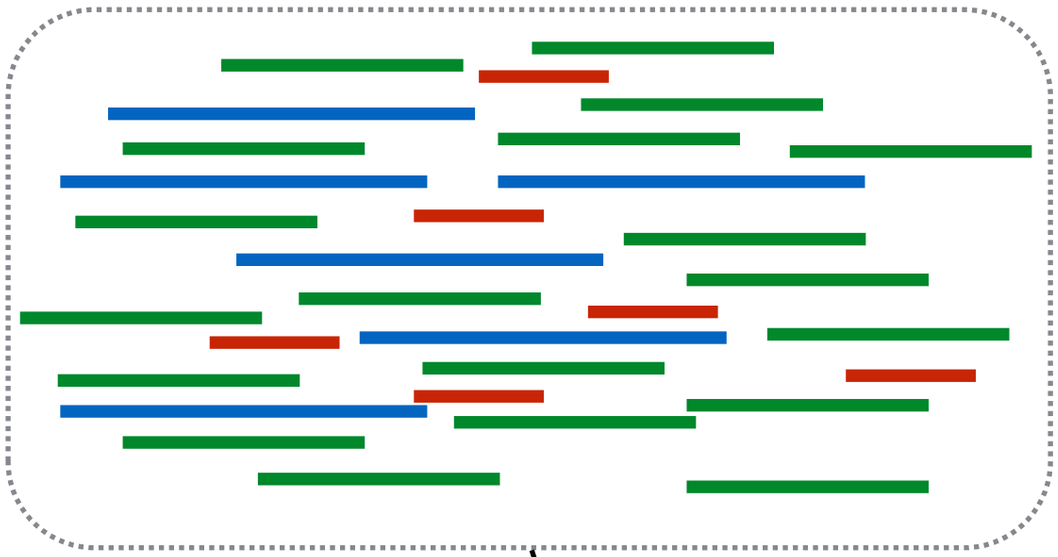


From: Soneson C, Love MI and Robinson MD 2016 [version 2; referees: 2 approved] F1000Research 2016, 4:1521 (doi: 10.12688/f1000research.7563.2)

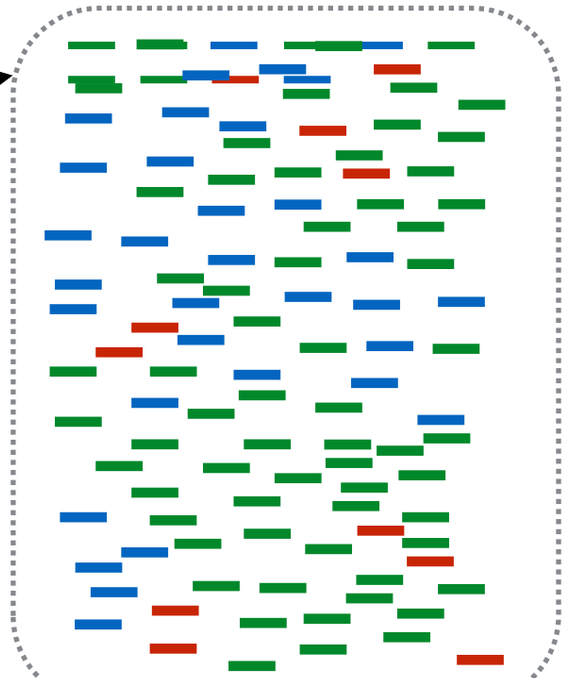
How do we do something better than “counting”?

Think about the “ideal” RNA-seq experiment . . .

Experimental Mixture



Read set



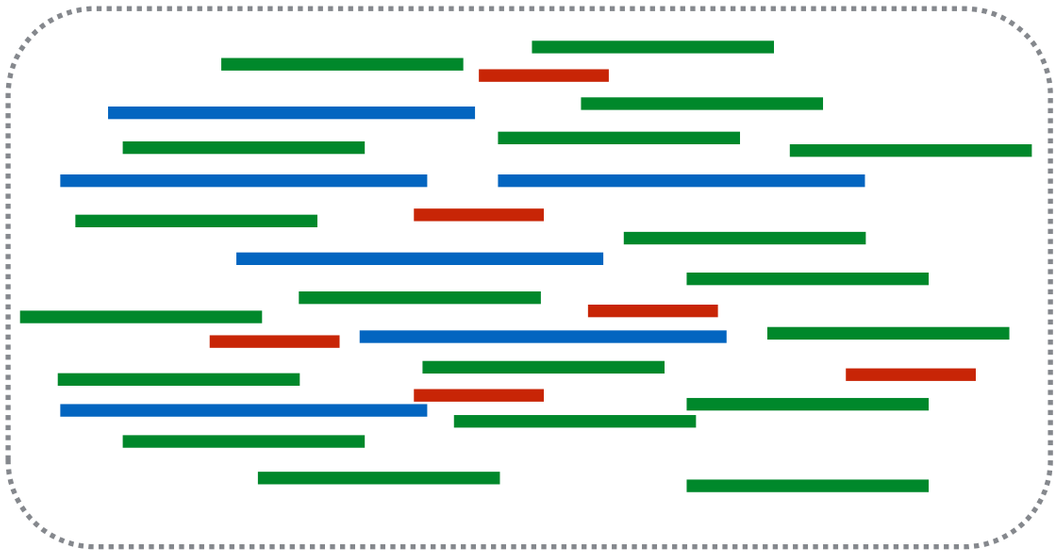
sequencing *oracle*

Pick a transcript **t** ∝ count * length

Pick a position **p** on **t** uniformly “at random”

How do we do something better than “counting”?

Experimental Mixture



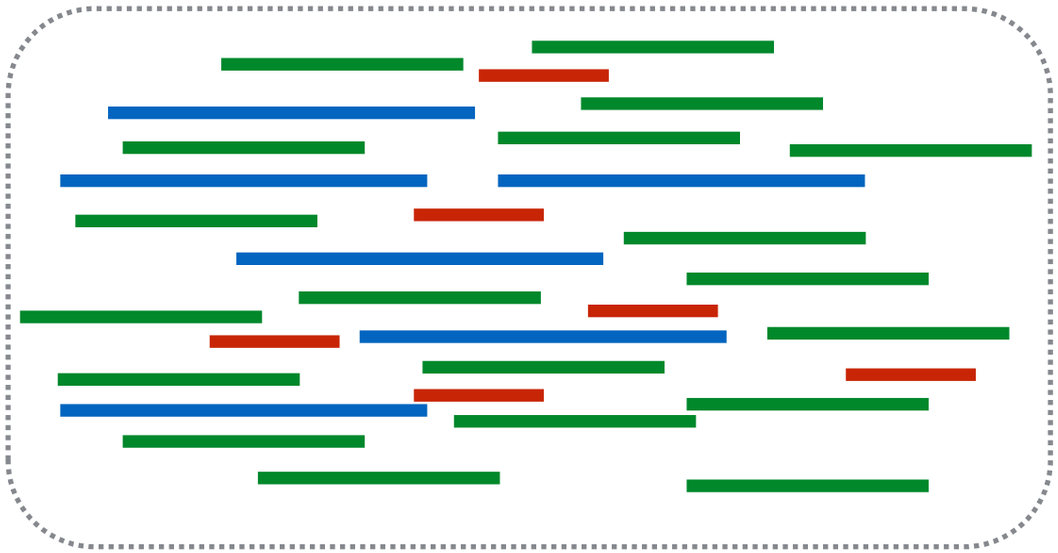
length( ) = 100 x 6 copies = 600 nt ~ 30% blue

length( ) = 66 x 19 copies = 1254 nt ~ 60% green

length( ) = 33 x 6 copies = 198 nt ~ 10% red

How do we do something better than “counting”?

Experimental Mixture



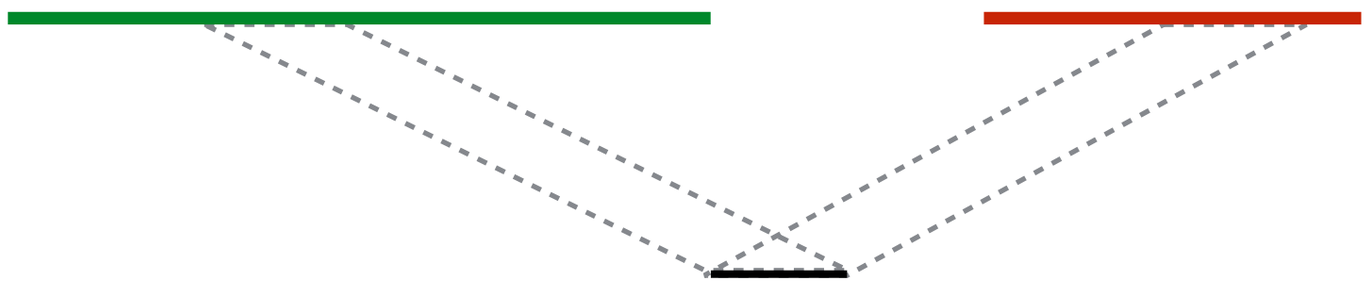
length( ) = 100 x 6 copies = 600 nt ~ 30% blue

length( ) = 66 x 19 copies = 1254 nt ~ 60% green

length( ) = 33 x 6 copies = 198 nt ~ 10% red

We call these values η = [0.3, 0.6, 0.1] the nucleotide fractions, they become the primary quantity of interest

Resolving a single multi-mapping read



Say we *knew* the η, and observed a read that mapped ambiguously, as shown above. What is the probability that it truly originated from G or R?

$$\Pr\{r \text{ from } G\} = \frac{\frac{\eta_G}{\text{length}(G)}}{\frac{\eta_G}{\text{length}(G)} + \frac{\eta_R}{\text{length}(R)}} = \frac{\frac{0.6}{66}}{\frac{0.6}{66} + \frac{0.1}{33}} = 0.75$$

$$\Pr\{r \text{ from } R\} = \frac{\frac{\eta_R}{\text{length}(R)}}{\frac{\eta_G}{\text{length}(G)} + \frac{\eta_R}{\text{length}(R)}} = \frac{\frac{0.1}{33}}{\frac{0.6}{66} + \frac{0.1}{33}} = 0.25$$

normalization factor

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

How to assess “abundance”

RPKM — Reads per kilobase per million mapped reads

FPKM — Fragments per kilobase per million mapped reads

Don’t use these measures, TPM measures the “same thing”, but in a better way.

TPM — Transcripts per million

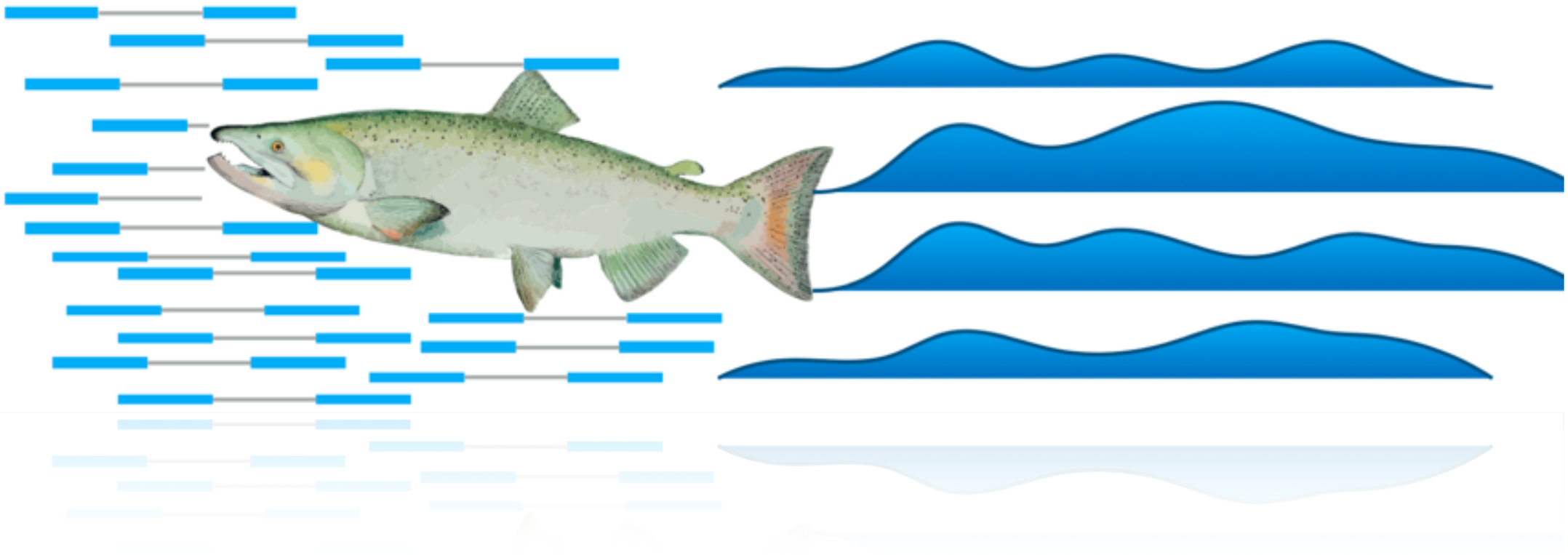
Useful for visualization / assessment etc.

(Estimated) Number of Reads

These are what are used (after normalization) for differential expression. Why can’t we use TPM?

Transcript Quantification

Salmon provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference



Official website: http://combine-lab.github.io/salmon/

GitHub repository: https://github.com/COMBINE-lab/salmon

A probabilistic view of RNA-Seq quantification

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)

We can safely truncate Pr{ti | **η**} to 0 for transcripts where a fragment doesn't map.

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

Prob. of selecting t_i *given* **η**

Depends on abundance estimate

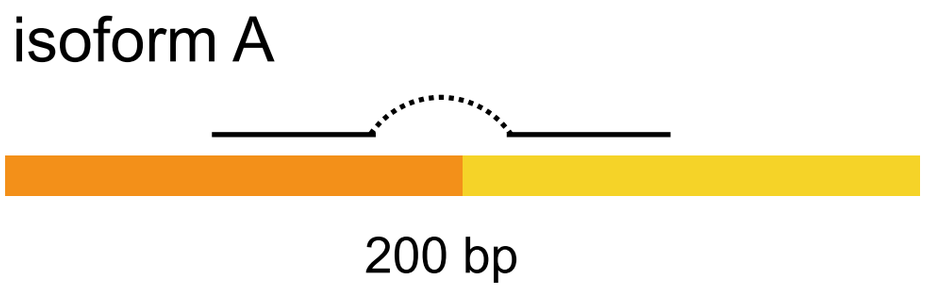
Prob. of generating fragment f_j *given* t_i

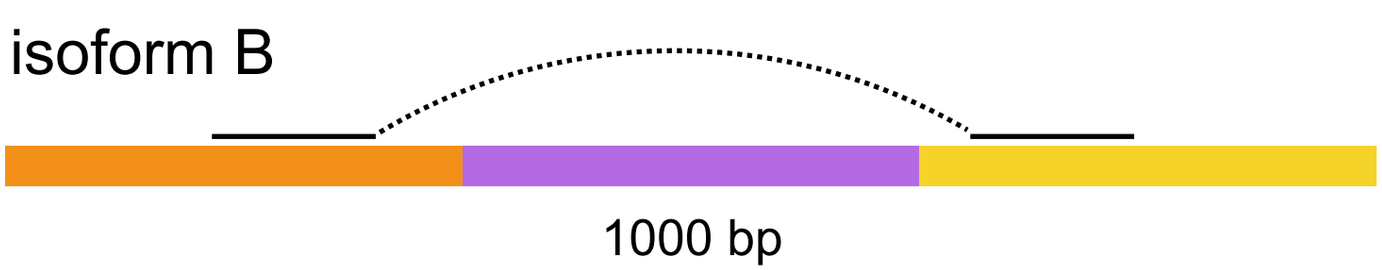
Independent of abundance estimate

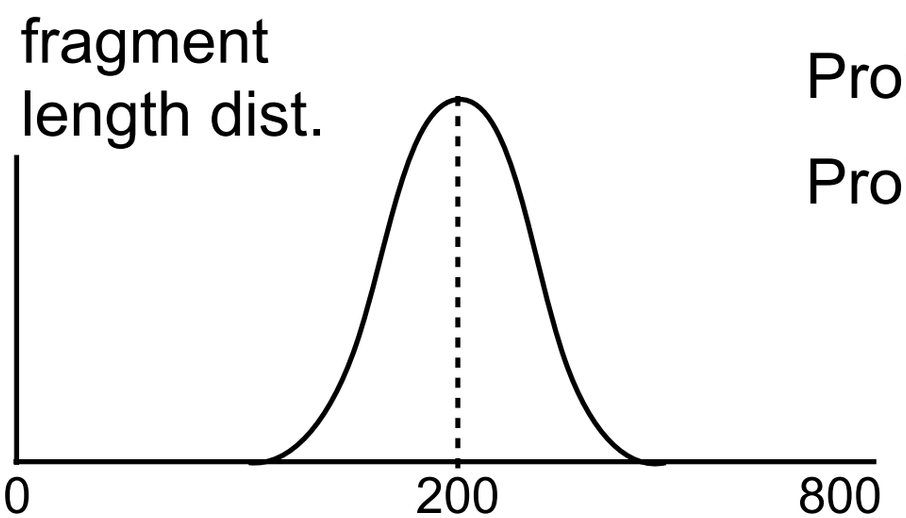
We want to find the values of **η** that ***maximize*** this probability.
We can do this (at least locally) using the EM algorithm.

Why does $\Pr\{f_j \mid t_i\}$ matter?

Consider the following scenario:





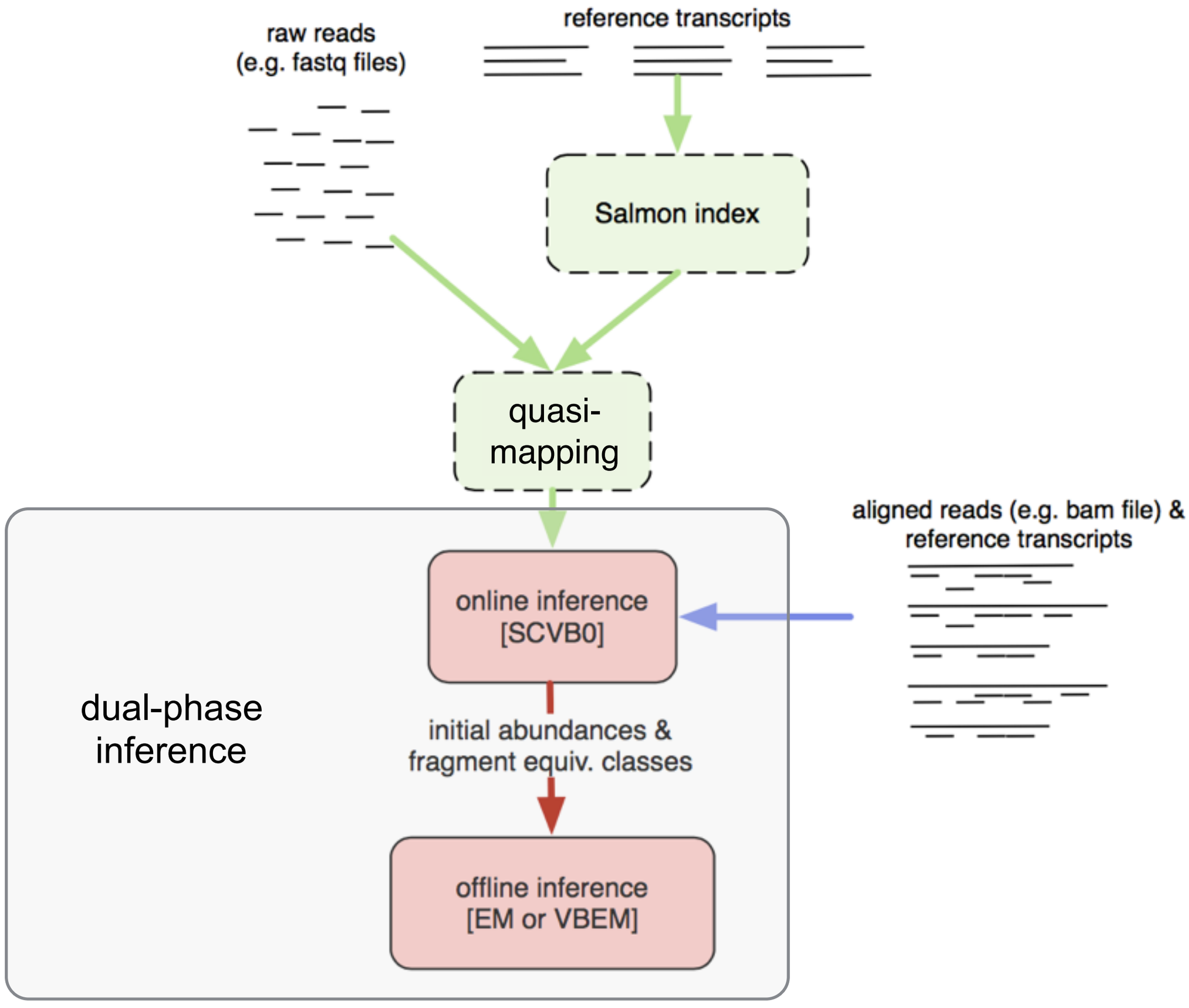


Aux. model provides *strong* information about origin of a fragment!

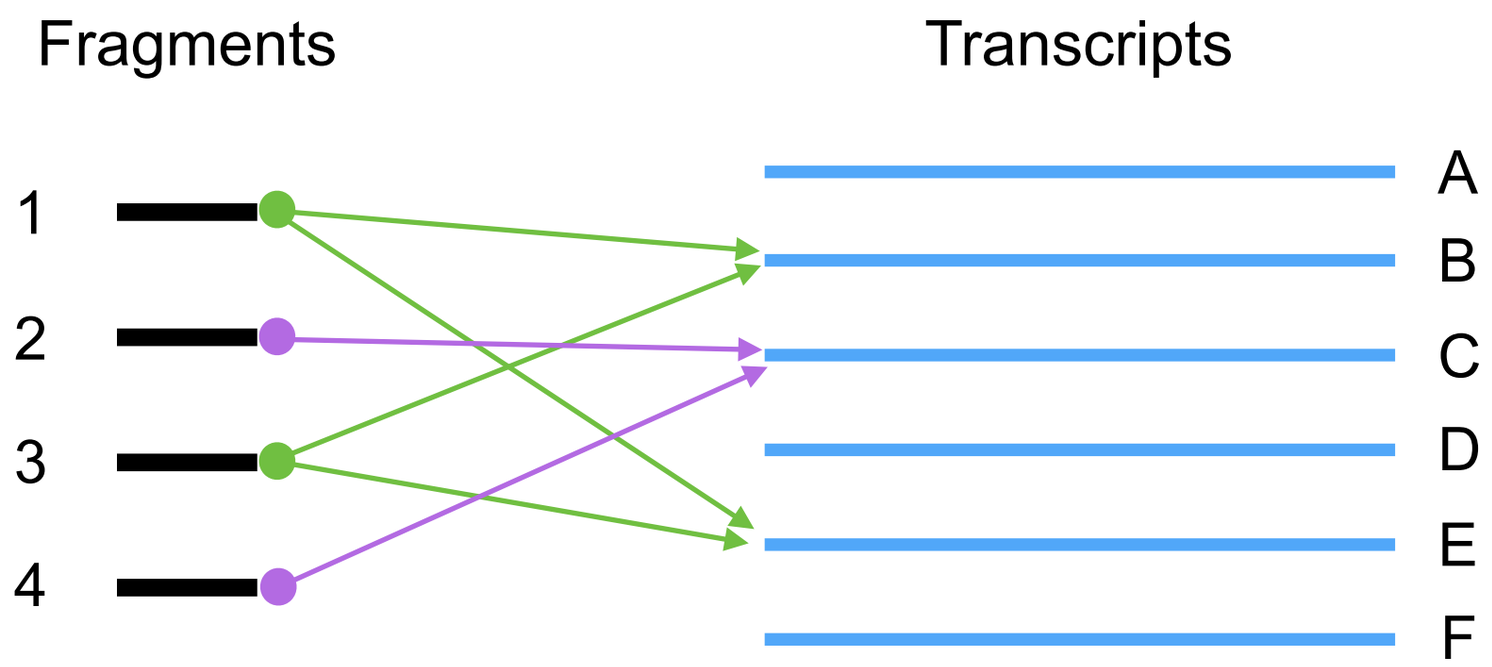
Prob of observing a fragment of size ~200 is **large**

Prob of observing a fragment of size ~1000 is **very small**

Salmon's "pipeline"



Fragment Equivalence Classes

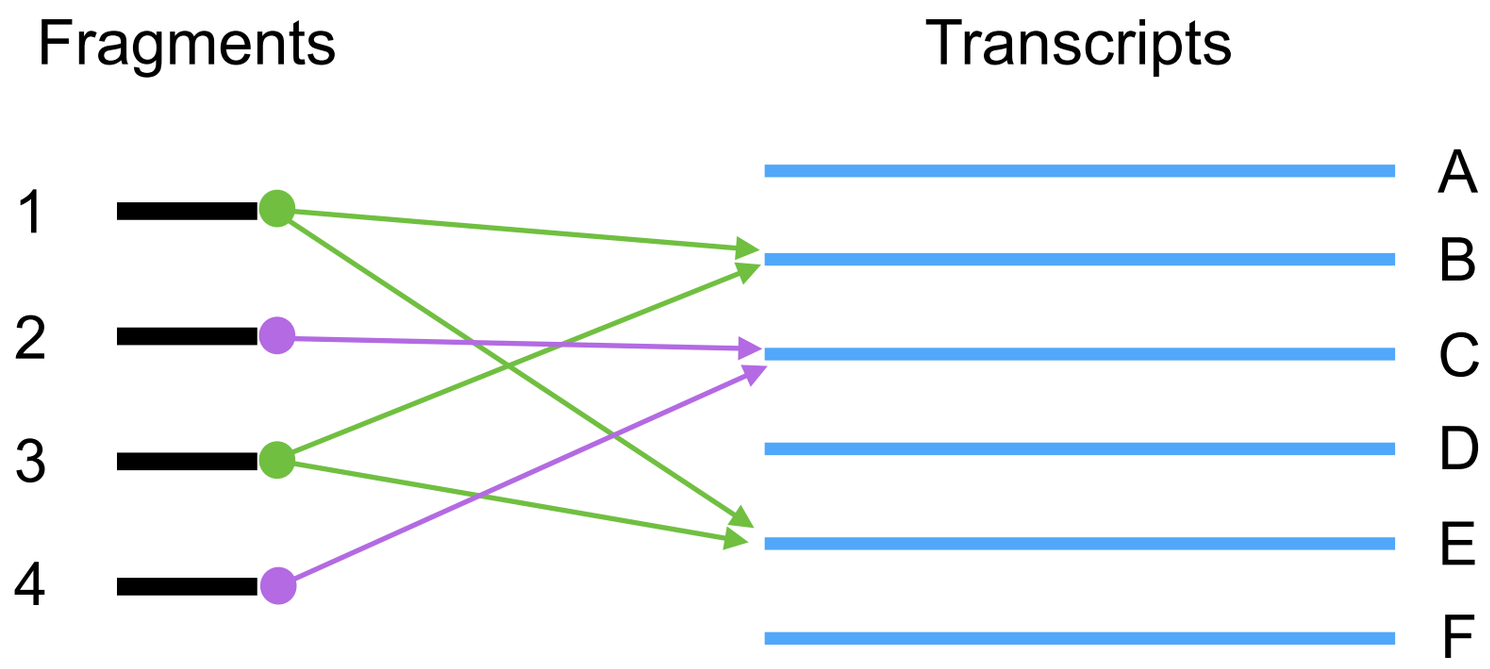


Reads 1 & 3 both map to transcripts B & E
Reads 2 & 4 both map to transcript C

We have 4 reads, but only 2 eq. classes of reads

eq. Label	Count	Aux weights
{B,E}	2	$w^{\{B,E\}}_B, w^{\{B,E\}}_E$
{C}	2	$w^{\{C\}}_C$

Fragment Equivalence Classes



w^j_i encodes the "affinity" of class *j* to transcript *i* according to the "bias" model. This is $P\{f_j \mid t_i\}$, aggregated for all fragments in a class.

Reads 1 & 3 both map to transcripts B & E
Reads 2 & 4 both map to transcript C

We have 4 reads, but only 2 eq. classes of reads

eq. Label	Count	Aux weights
{B,E}	2	$w^{\{B,E\}}_B, w^{\{B,E\}}_E$
{C}	2	$w^{\{C\}}_C$

The number of equivalence classes is small

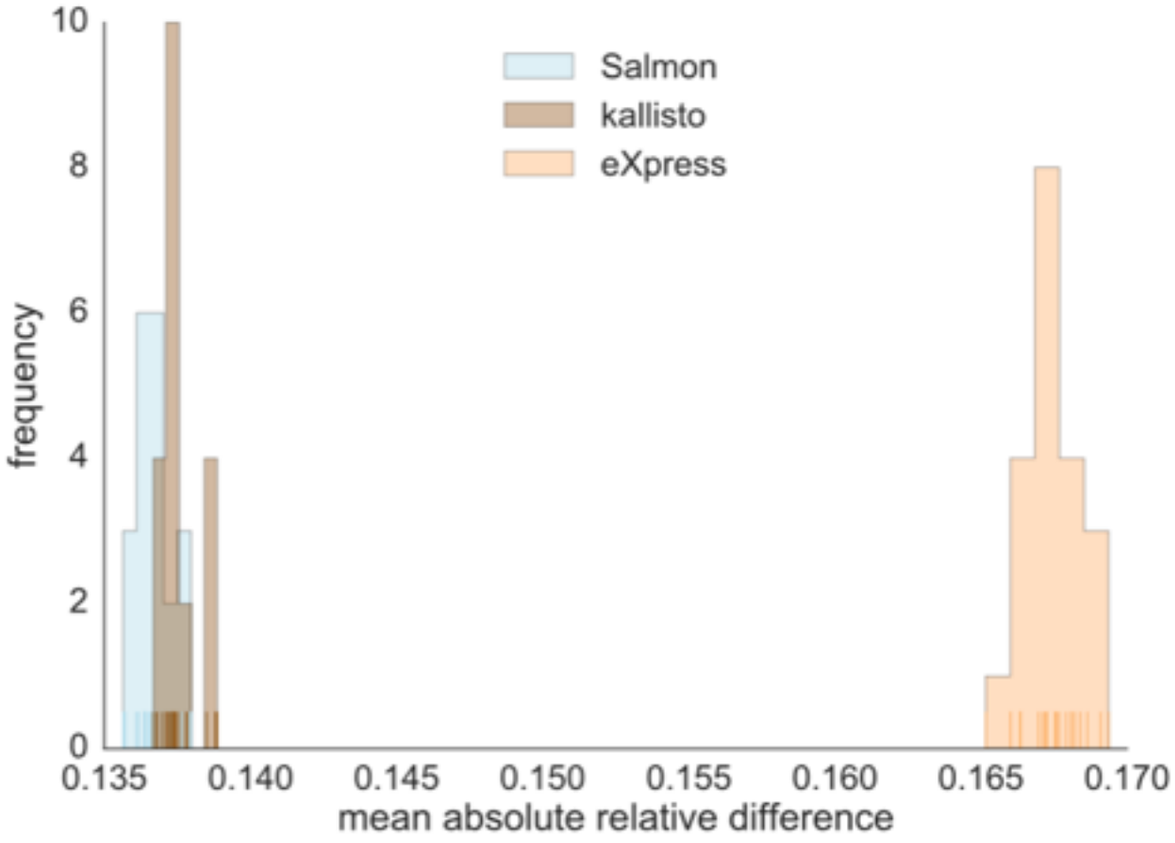
	Yeast	Human	Chicken
# contigs	7353	107,389	335,377
# samples	6	6	8
Total (paired-end) reads	~36,000,000	~116,000,000	~181,402,780
Avg # eq. classes (across samples)	5197	100,535	222,216

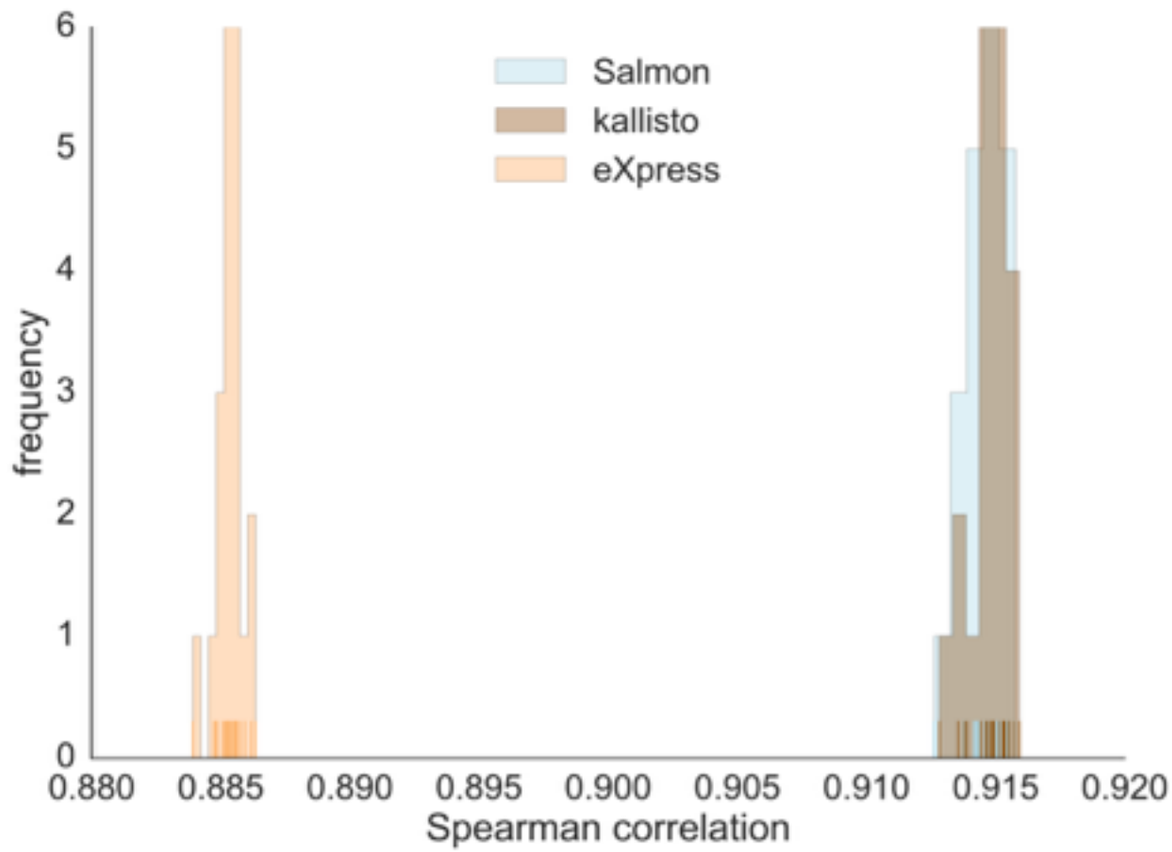
The **# of equivalence classes grows with the complexity of the transcriptome** — independent of the # of sequence fragments.

Typically, ***two or more orders of magnitude*** fewer equivalence classes than sequenced fragments.

The offline **inference** algorithm **scales in # of fragment equivalence classes**.

Transcript inference methods can be very accurate





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

Results on 20 replicates simulated (RSEM-sim) from parameters learned from NA12716_7 from GEUVADIS. Showing result distributions for kallisto[1], eXpress[2] & salmon[3]

1: Bray, Nicolas L., et al. "Near-optimal probabilistic RNA-seq quantification." Nature biotechnology 34.5 (2016): 525-527. (v0.43.0)

2: Roberts, Adam, and Lior Pachter. "Streaming fragment assignment for real-time analysis of sequencing experiments." Nature methods 10.1 (2013): 71-73. (v.1.5.1)

3: Patro, Rob, et al. "Accurate, fast, and model-aware transcript expression quantification with Salmon." bioRxiv (2015): 021592. (v0.7.0)

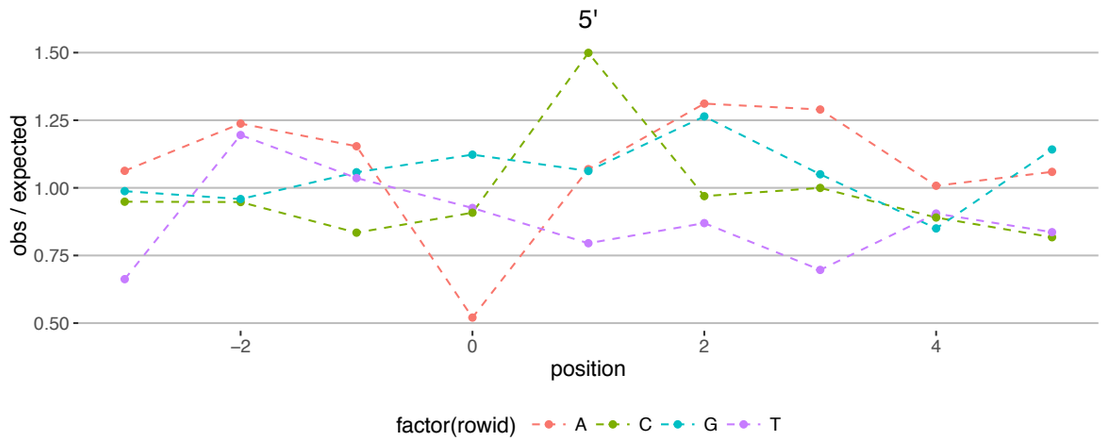
Biases abound in RNA-seq data

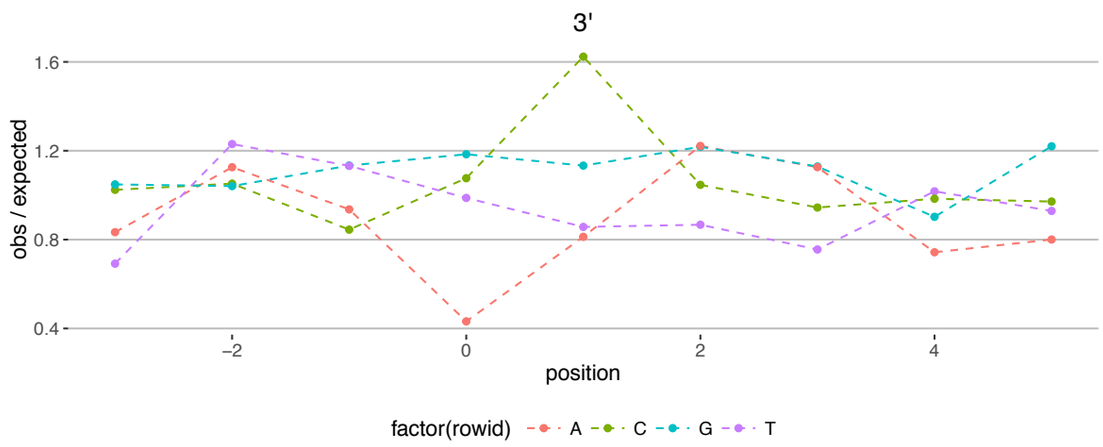
Biases in prep & sequencing can have a significant effect on the fragments we see.

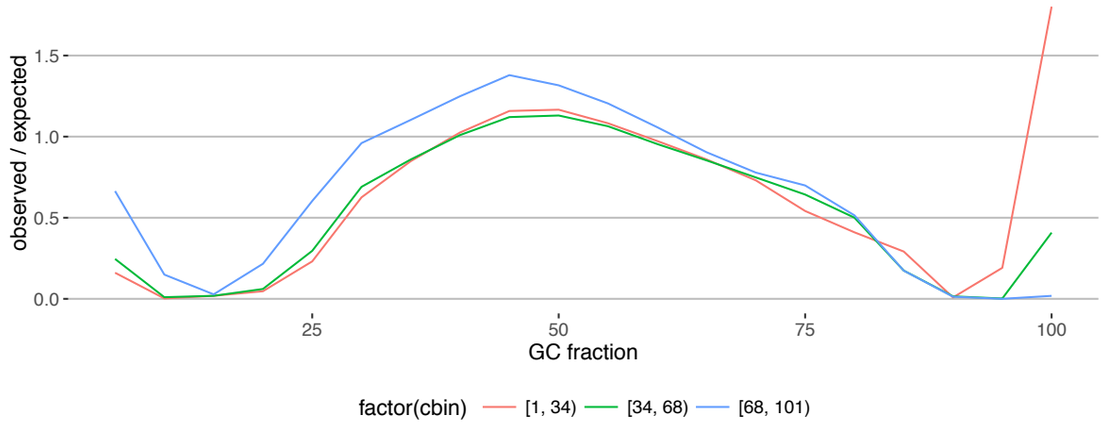
Fragment gc-bias[1]—
The GC-content of the fragment affects the likelihood of sequencing

Sequence-specific bias[2]—
sequences surrounding fragment affect the likelihood of sequencing

Positional bias[2]—
fragments sequenced non-uniformly across the body of a transcript







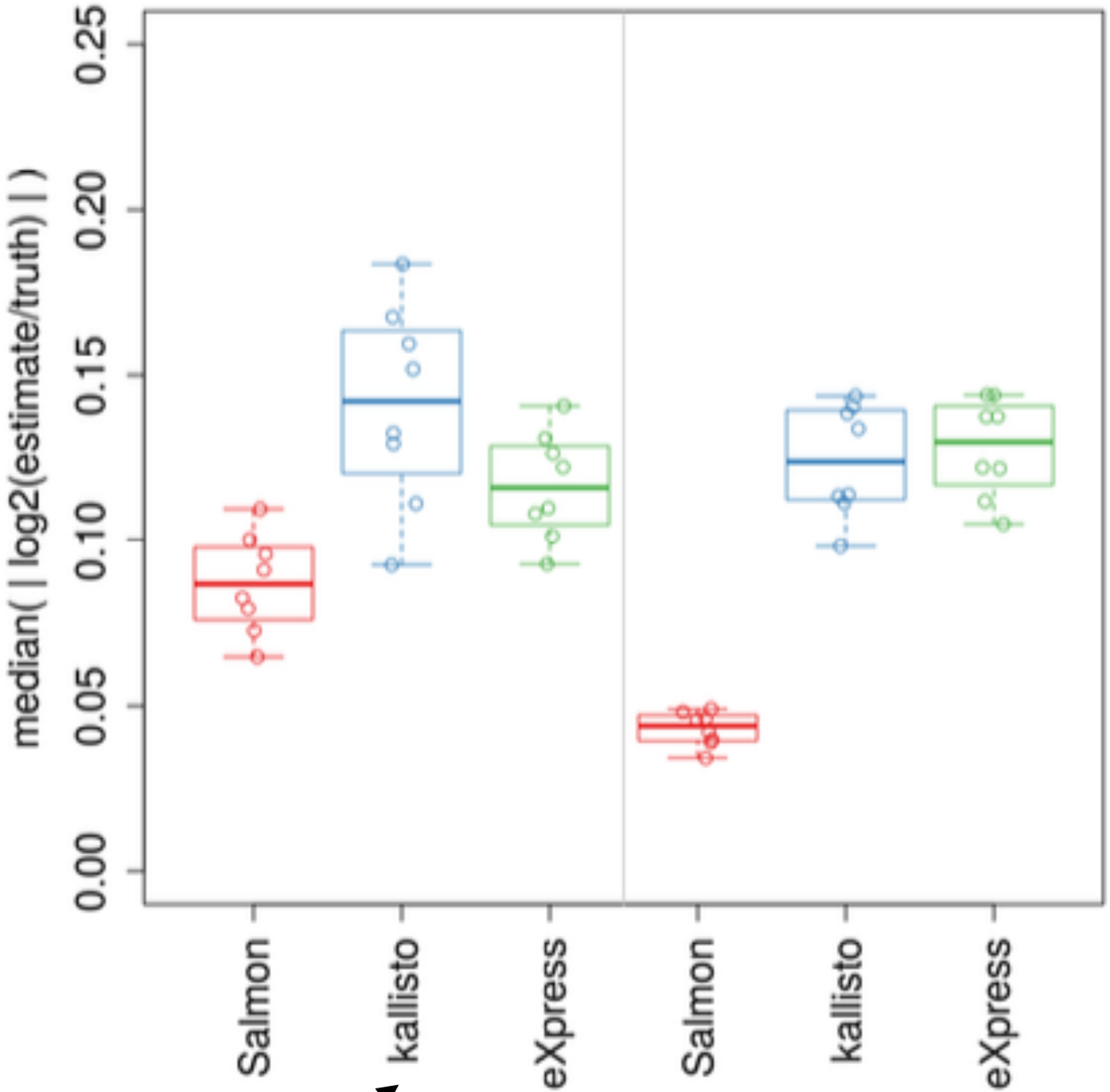
1:Love, Michael I., John B. Hogenesch, and Rafael A. Irizarry. "Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation." bioRxiv (2015): 025767.

2:Roberts, Adam, et al. "Improving RNA-Seq expression estimates by correcting for fragment bias." Genome biology 12.3 (2011): 1.

Accuracy difference can be larger with biased data

Simulated data:
2 conditions; 8 samples each

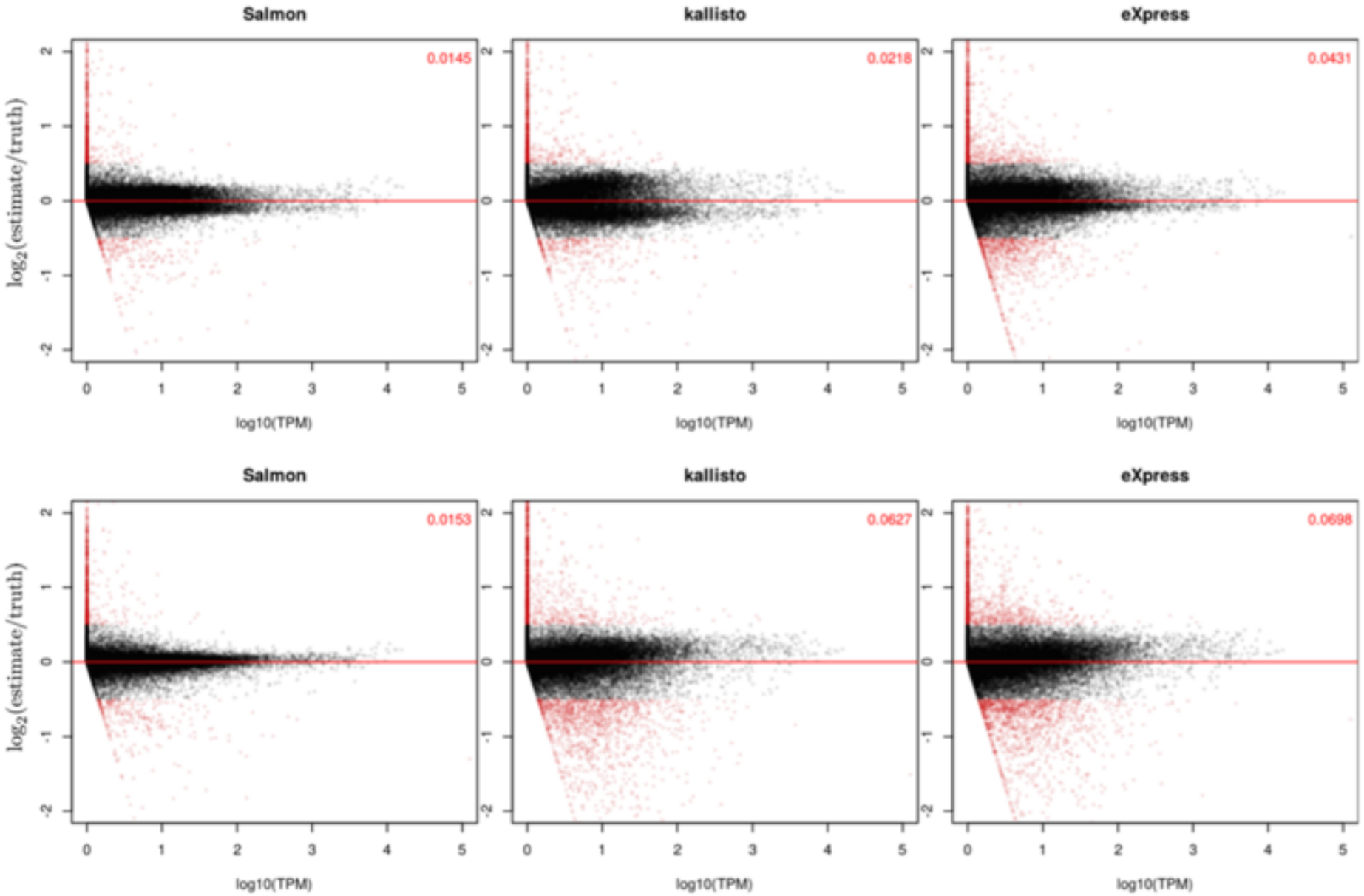
- Simulated transcripts across entire genome with known abundance using Polyester (modified to account for GC bias)
- How well do we recover the underlying relative abundances?
- How does accuracy vary with level of bias?



Sequence-bias models don’t account for fragment-level GC bias

joint work with Geet Duggal, Mike Love, Rafael Irizarry & Carl Kingsford

Accuracy difference can be larger with biased data



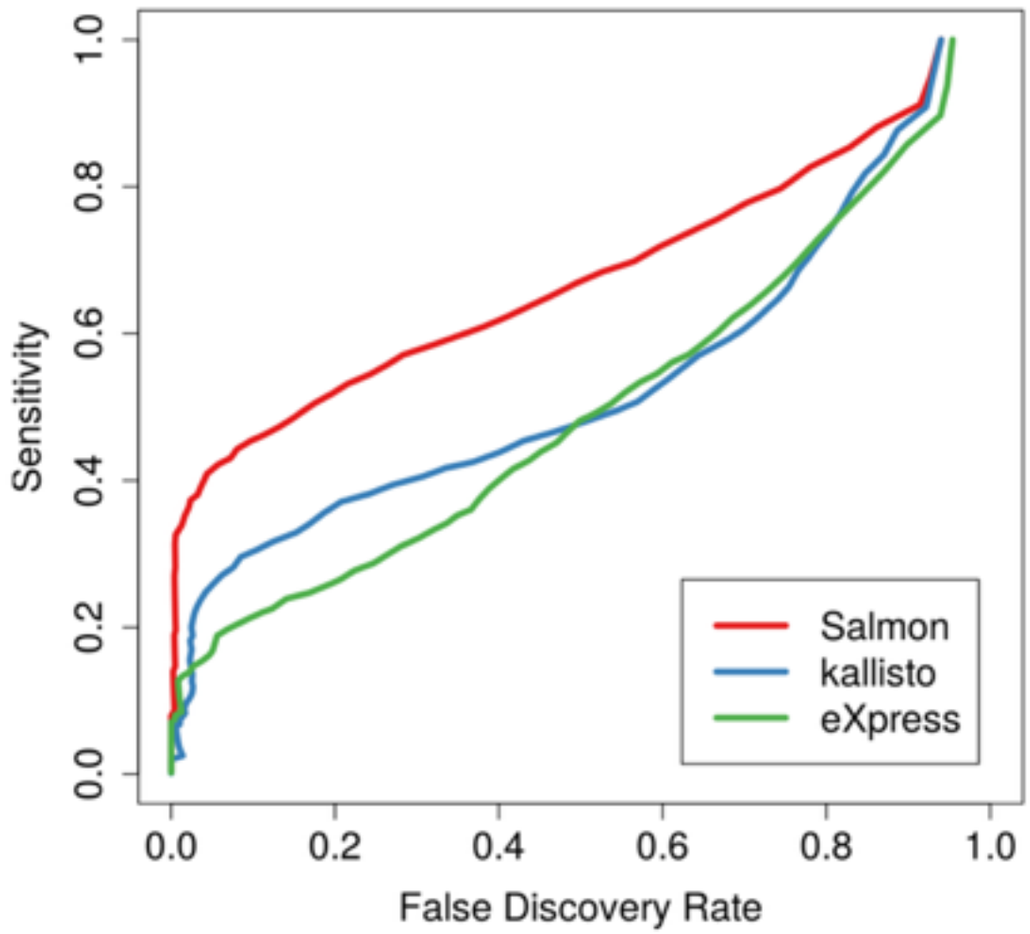
joint work with Geet Duggal, Mike Love, Rafael Irizarry & Carl Kingsford

Mis-estimates confound downstream analysis

Simulated data:
2 conditions; 8 replicates each

- set 10% of txps to have fold change of 1/2 or 2 — rest unchanged.
- How well do we recover true DE?
- Since bias is systematic, effect may be even worse than accuracy difference suggests.

Recovery of DE transcripts



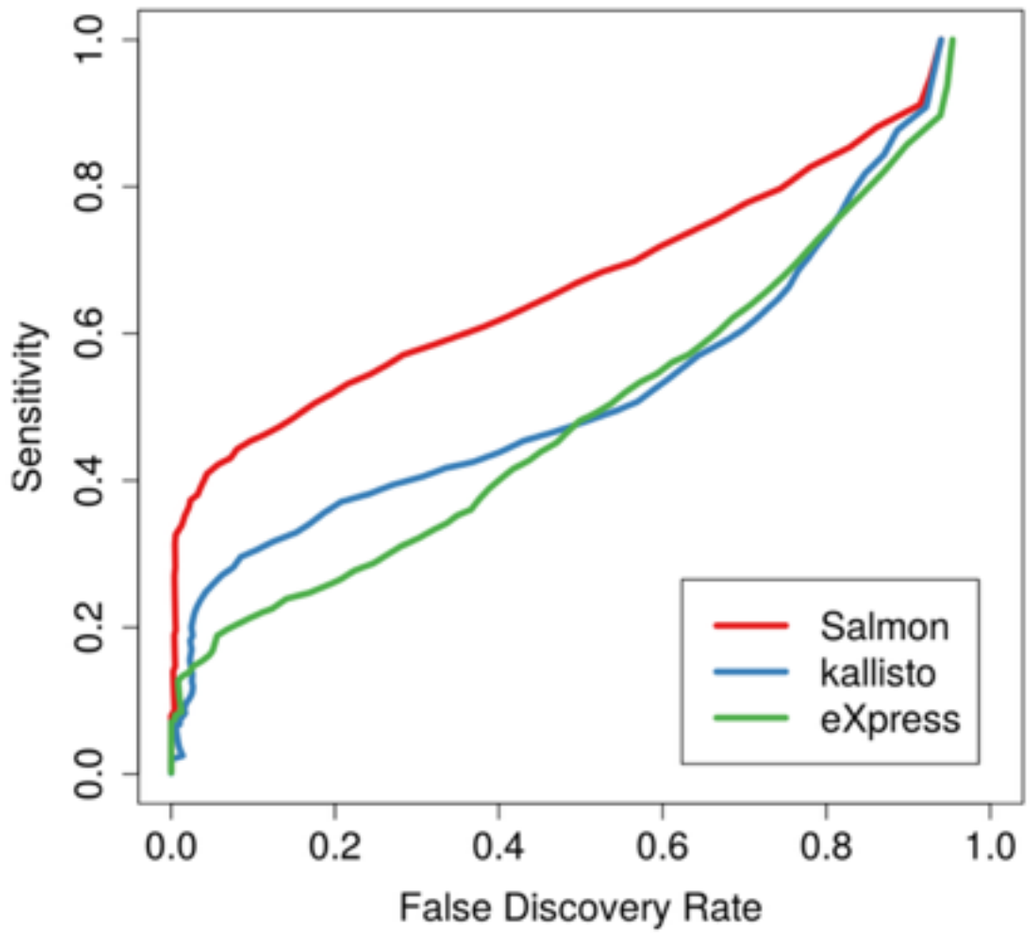
joint work with Geet Duggal, Mike Love, Rafael Irizarry & Carl Kingsford

Accuracy difference can be large with biased data!

FDR	Sensitivity at given FDR		
	Salmon	*kallisto*	*eXpress*
0.01	0.326	0.072	0.128
0.05	0.409	0.248	0.162
0.1	0.454	0.296	0.211

At the same FDR,
accuracy differences of
53 - 450%

Recovery of DE transcripts



joint work with Geet Duggal, Mike Love, Rafael Irizarry & Carl Kingsford

Importance with experimental data

30 samples from the GEUVADIS study:
- 15 samples from UNIGE sequencing center
- 15 samples from CNAG_CRG sequencing center

Same human population, expect few-to-no *real* DE (primary differences in sample prep)

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

	Salmon	Kallisto	eXpress
All transcripts	**1,171**	2,620	2,472
Transcripts of 2 isoform genes	**224**	545	531

Bias and **batch effects** are ***substantial***, and must be accounted for.

Importance with experimental data

30 samples from the GEUVADIS study:
  15 samples from UNIGE sequencing center
  15 samples from CNAG_CRG sequencing center

Same human population, expect few-to-no *real* DE (primary differences in sample prep)

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

But this is txp-level DE, and I care only about **genes!**

	Salmon	Kallisto	eXpress
All transcripts	1,171	2,620	2,472
Transcripts of 2 isoform genes	224	545	531

Bias and **batch effects** are ***substantial***, and must be accounted for.

Importance with experimental data

30 samples from the GEUVADIS study:
15 samples from UNIGE sequencing center
15 samples from CNAG_CRG sequencing center

Effects seem **at least as extreme** at the gene level

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

	Salmon	Kallisto	eXpress
All genes	**455**	1,200	1582
Transcripts of 2 isoform genes	**224**	545	531

Bias and **batch effects** are ***substantial***, and must be accounted for.

Salmon and kallisto are FAST

http://www.sbnation.com/lookit/2016/8/12/12463026/katie-ledectky-800m-gold-video-highlights-rio-olympics

Salmon and kallisto are FAST

Consider the following test:

Take all 20 replicates of the RSEM-sim simulated data above, treat them as one, giant sample. This is 20 samples x 30M paired-end reads = 600 million read pairs or 1.2 billion individual reads.

Using 30 threads[1]:

kallisto can process this sample in 20 minutes
Salmon can process this sample in 23 minutes

Just *aligning* the reads to use e.g. eXpress, Cufflinks, RSEM etc. would take dozens of hours.

1: Intel Xeon E5-4600 (2.6GHz)

One “issue” with maximum likelihood (ML)

The generative statistical model is a principled and elegant way to represent the RNA-seq process.

It can be optimized efficiently using e.g. the EM / VBEM algorithm.

but, these efficient optimization algorithms return “point estimates” of the abundances. That is, there is no notion of how *certain* we are in the computed abundance of transcript.

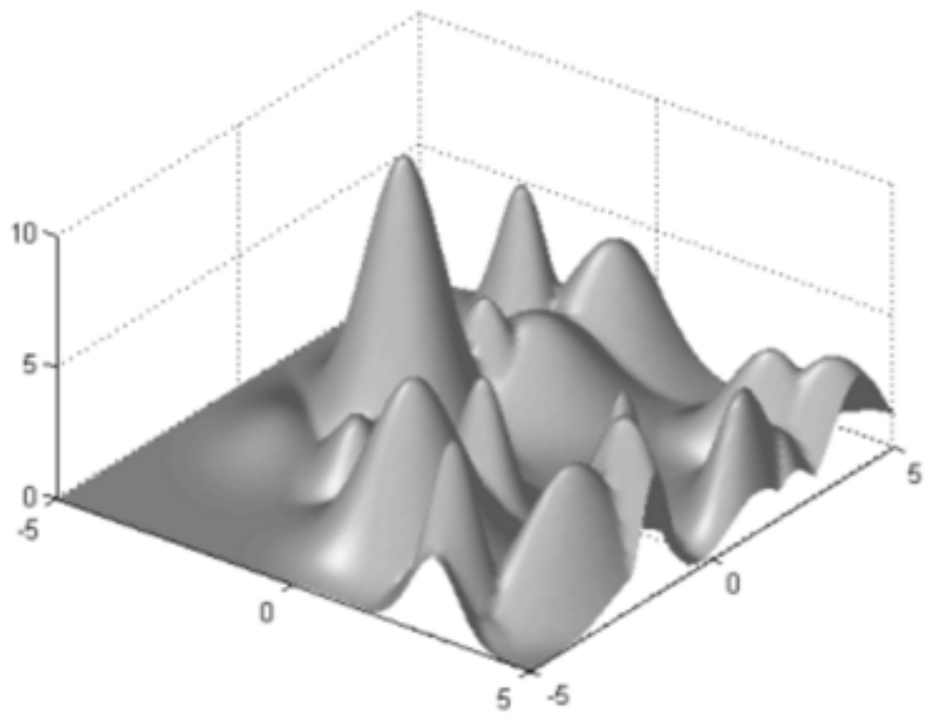
One “issue” with maximum likelihood (ML)

There are multiple sources of uncertainty e.g.

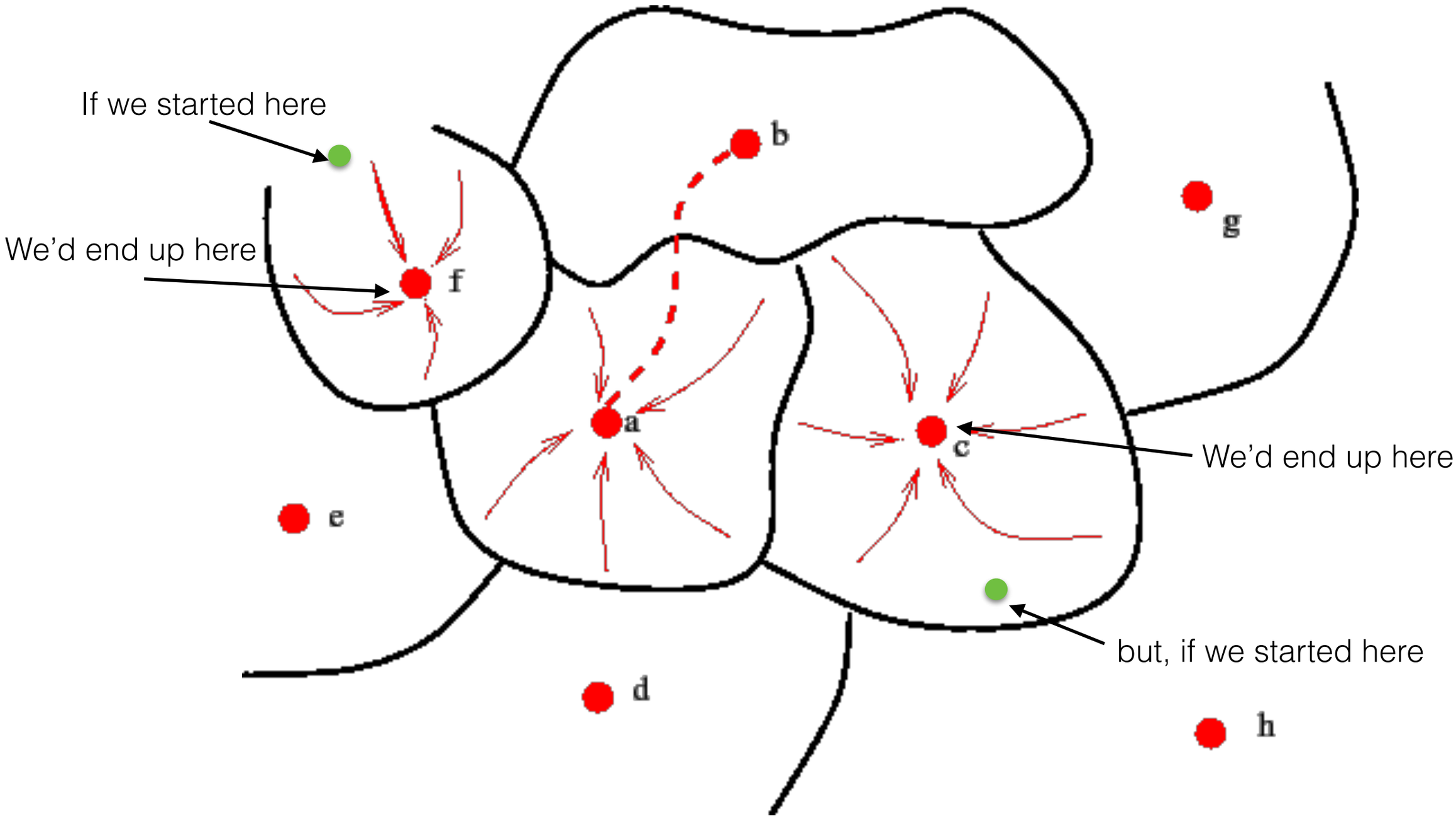
- Technical variance : If we sequenced the *exact* same sample again, we’d get a different set of fragments, and, potentially a different solution.

- Uncertainty in inference: We are almost never guaranteed to find a unique, globally optimal result. If we started our algorithm with different initialization parameters, we might get a different result.

We’re trying to find the *best* parameters in a space with 10s to 100s of thousands of dimensions!



One “issue” with maximum likelihood (ML)



https://commons.wikimedia.org/wiki/File:Local_search_attraction_basins.png (CC BY-SA 3.0)

Assessing Uncertainty

There are a few ways to address this “issue”

Do a fully Bayesian inference[1]:

Infer the entire posterior distribution of parameters, not just a ML estimate (e.g. using MCMC) — too slow!

✓ Posterior Gibbs Sampling:

Starting from our ML estimate, do MCMC sampling to explore how parameters vary — if our ML estimate is good, and taking advantage of equivalence classes, this can be made *very fast.*

✓ Bootstrap Sampling[2]:

Resample (from equivalence class counts) with replacement, and re-run the ML estimate for each sample. This can be made reasonably fast.

Happy to discuss details / implications of this further.

1: BitSeq (with MCMC) actually does this. It’s very accurate, but very slow. [Glaus, Peter, Antti Honkela, and Magnus Rattray. "Identifying differentially expressed transcripts from RNA-seq data with biological variation." Bioinformatics 28.13 (2012): 1721-1728.]

2: IsoDE introduced the idea of bootstrapping counts to assess quantification uncertainty. [Al Seesi, Sahar, et al. "Bootstrap-based differential gene expression analysis for RNA-Seq data with and without replicates." BMC genomics 15.8 (2014): 1.], but it was first made practical / fast in kallisto by doing the bootstrapping over equivalence classes.

Salmon addresses the main challenges of quantification

- finding locations of reads (mapping) is slow than necessary

→ Use quasi-mapping

- alternative splicing and related sequences creates ambiguity about where reads came from

→ Use dual-phase inference algorithm

- sampling of reads is not uniform or idealized

→ Use bias models learned from data

- uncertainty in ML estimate of abundances

→ Use posterior Gibbs sampling or bootstraps to assess uncertainty

Salmon has many other benefits

- Speed of inference makes it possible to use **bootstraps** or posterior **Gibbs sampling** to estimate variance (e.g. how certain are we in quantification estimates?).

- Quasi-mapping means no large, intermediate BAM files sitting on disk, or wasting computation time with slow disk I/O.

- Expressive model means new types of bias can be learned and accounted for.

- Separation of mapping / alignment and inference means Salmon can be used with or without existing alignments*. Here I talked only about quasi-mapping, but Salmon can use take BAM input from an aligner (if you really want!).

Many of these improvements (except dual-phase inference) have been back-ported to Sailfish, which is still *actively developed!*

https://github.com/kingsfordgroup/sailfish

Thanks!

Collaborators on Salmon

Geet Duggal (CMU / DNAnexus)

Carl Kingsford (CMU)

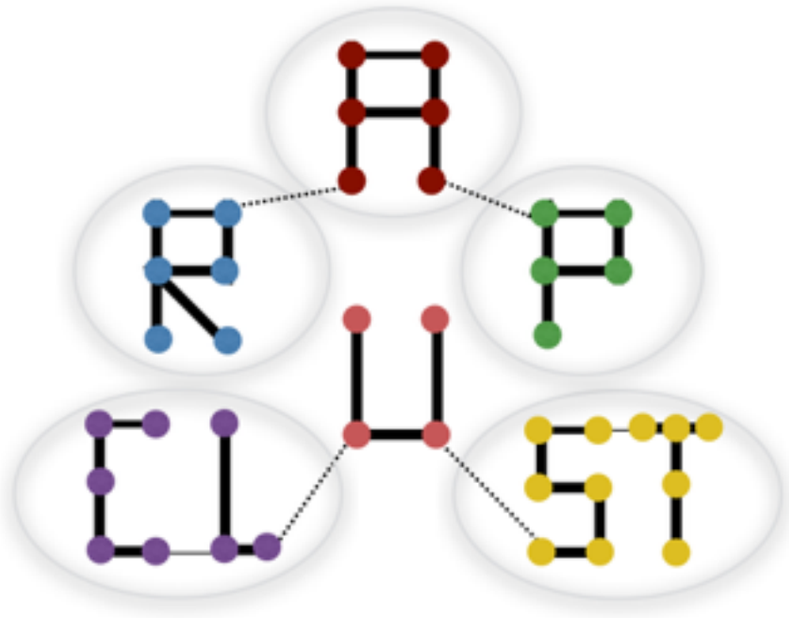
Mike Love (Harvard / UNC)

Rafael Irizarry(Harvard)

Bonus Slides

De novo transcriptome clustering

RapClust: Fast, Lightweight Clustering of de novo Transcriptomes using Fragment Equivalence Classes



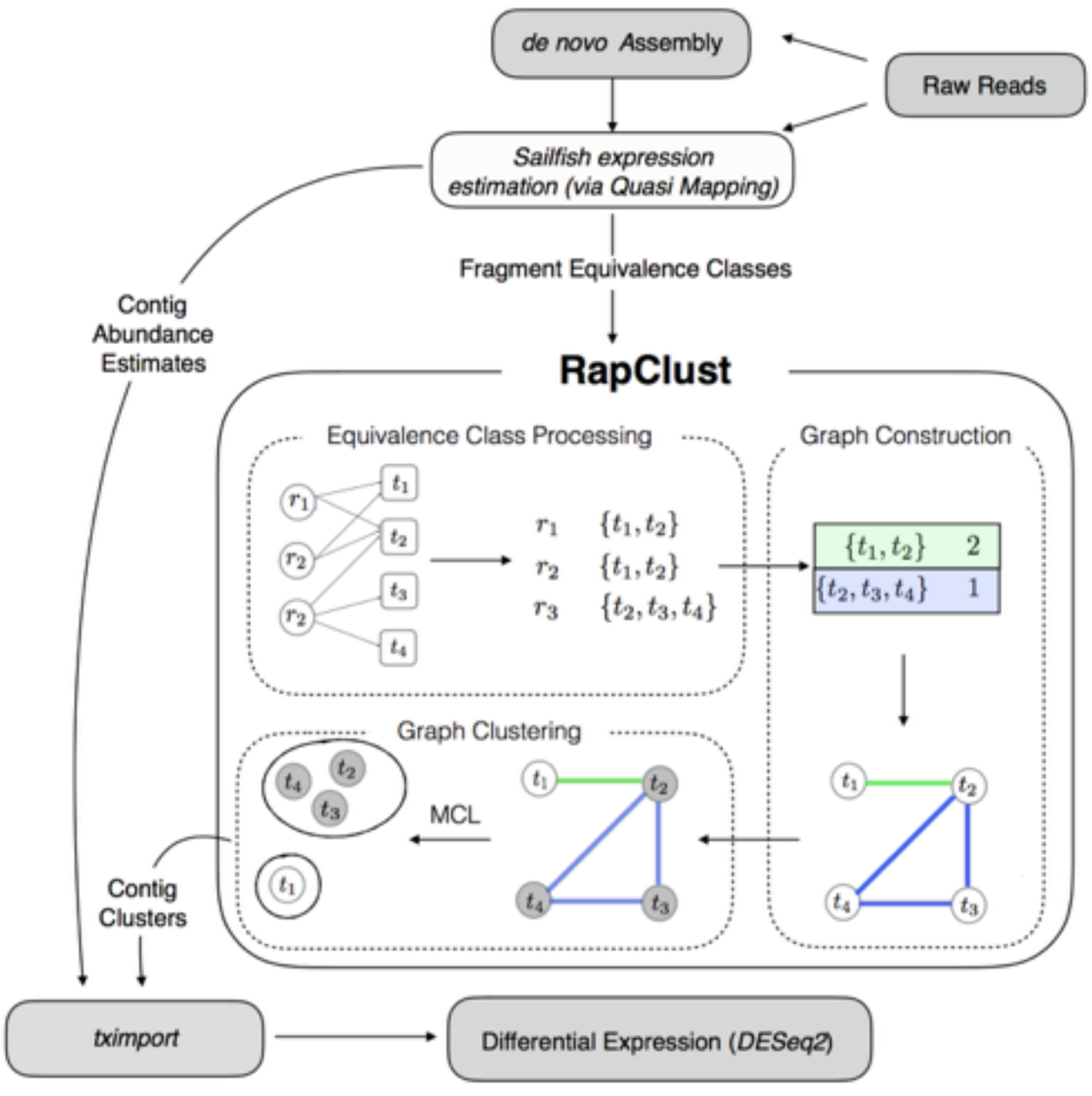
GitHub repository: https://github.com/COMBINE-lab/rapclust

Paper: https://arxiv.org/abs/1604.03250

RapClust: clustering contigs in de novo assemblies

Uses the fragment equivalence classes discussed above to cluster contigs in *de novo* assemblies.

This leads to improved downstream analysis (e.g. DE calls)



RapClust is fast

Time *including* quantification (4 threads)

	Yeast		Human		Chicken	
	`RapClust`	`Corset`	`RapClust`	`Corset`	`RapClust`	`Corset`
Time(min)	5.12	37.25	22.67	211.67	64.18	453
Space(Gb)	0.005	5.7	0.092	22	0.49	145
% of reads	88.17	62.32	93.04	77.94	88.80	60.99

Time *excluding* quantification

	Yeast			Human			Chicken		
	RC	CD	CT	RC	CD	CT	RC	CD	CT
Time(min)	0.04	0.2	2.8	0.82	4.02	16.25	5.29	36.5	87

RapClust is Fast & Lightweight

Time & Space comparison of RapClust with Corset, for *all* phases (raw reads through quantified clusters — using 4 threads).

	Yeast		Human		Chicken	
	`RapClust`	`Corset`	`RapClust`	`Corset`	`RapClust`	`Corset`
Time(min)	5.12	37.25	22.67	211.67	64.18	453
Space(Gb)	0.005	5.7	0.092	22	0.49	145
% of reads	88.17	62.32	93.04	77.94	88.80	60.99

Not having to output / rely on BAM files means the space footprint of RapClust is *orders of magnitude* smaller than that of Corset

Time comparison of RapClust (RC), Corset (CT), and CD-HIT EST (CD) for *just clustering* (using 1 thread).

	Yeast			Human			Chicken		
	RC	CD	CT	RC	CD	CT	RC	CD	CT
Time(min)	0.04	0.2	2.8	0.82	4.02	16.25	5.29	36.5	87

RapClust is accurate

Variation of Information[#] distance between the *true* clustering and the clustering computed by each method (lower is better).

#: Meila, M. (2007). "Comparing clusterings—an information based distance". Journal of Multivariate Analysis 98 (5): 873–895.

VI Distance	RapClust	CORSET	CD-HIT EST
Chicken	0.127	0.191	2.01
Human	0.712	0.735	1.24
Yeast	0.176	0.178	0.216

F1-Score of correct classification (i.e. co-clustering) of contigs from the same gene (higher is better).

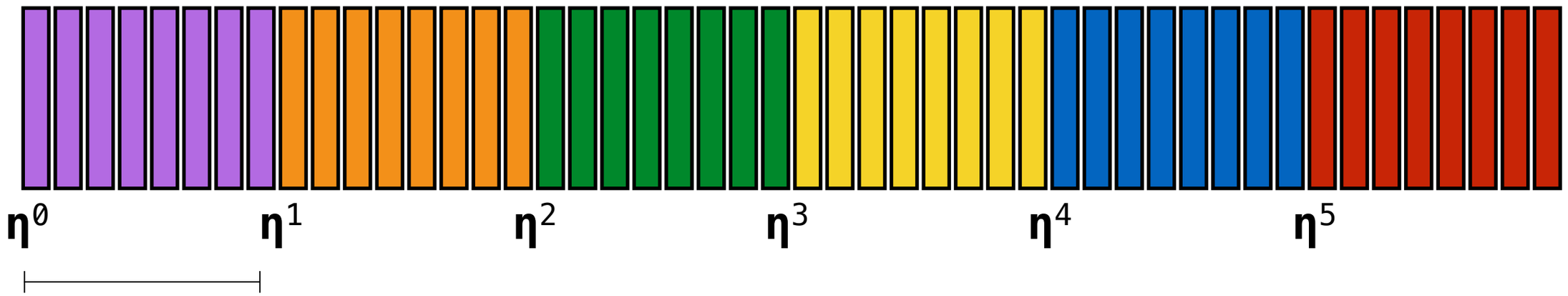
F1-Score	RapClust	CORSET	CD-HIT EST*
Chicken	97.17	95.02	13.27
Human	72.23	70.58	23.97
Yeast	46.24	45.40	21.48

*Note: RapClust & CORSET only predict clusters on an expressed subset of the data; CD-HIT EST is not directly comparable.

Phase 1: Online Inference

online inference [SCVB0] *

Process fragments in batches:



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

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

Weighting factor that decays over time

Update "bias" model

Place mappings in **equivalence classes**

- Have access to *all fragment-level information* when making these updates
- Often converges very quickly.
- Compare-And-Swap (CAS) for synchronizing updates of different batches

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

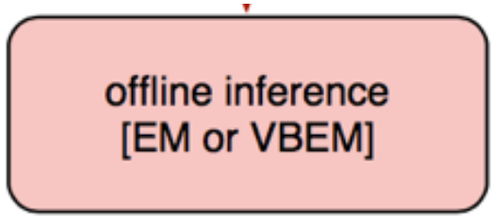
Give each transcript *appropriate* prior mass η^0 (init.)

```
For each mini-batch B^t of reads {

 For each read r in B^t {

  For each alignment a of r {
   compute (un-normalized) prob of a using η^(t-1), and aux params
  }
  normalize alignment probs & update local transcript weights η'
  add / update the equivalence class for read r
  sample a ∈ r to update auxiliary models
 }
 update global transcript weights given local transcript
 weights according to "update rule" ⟹ η^t = η^(t-1) + w^t η'
}
```

mini-batches processed in parallel by different threads

additive nature of updates mitigates effects of no synchronization between mini-batches

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*

small # of eq. classes means EM rounds are *fast!*

In practice, we re-estimate the bias terms that depend on the transcript abundances (e.g. seq-specific & fragment-GC) intermittently during optimization.