RNAseq
Bias Correction
(& Isoform Quantification)

Walter L. (Larry) Ruzzo

Computer Science and Engineering
Genome Sciences
University of Washington
Fred Hutchinson Cancer Research Center
Seattle, WA, USA

Benasque, 2015-Jul-21

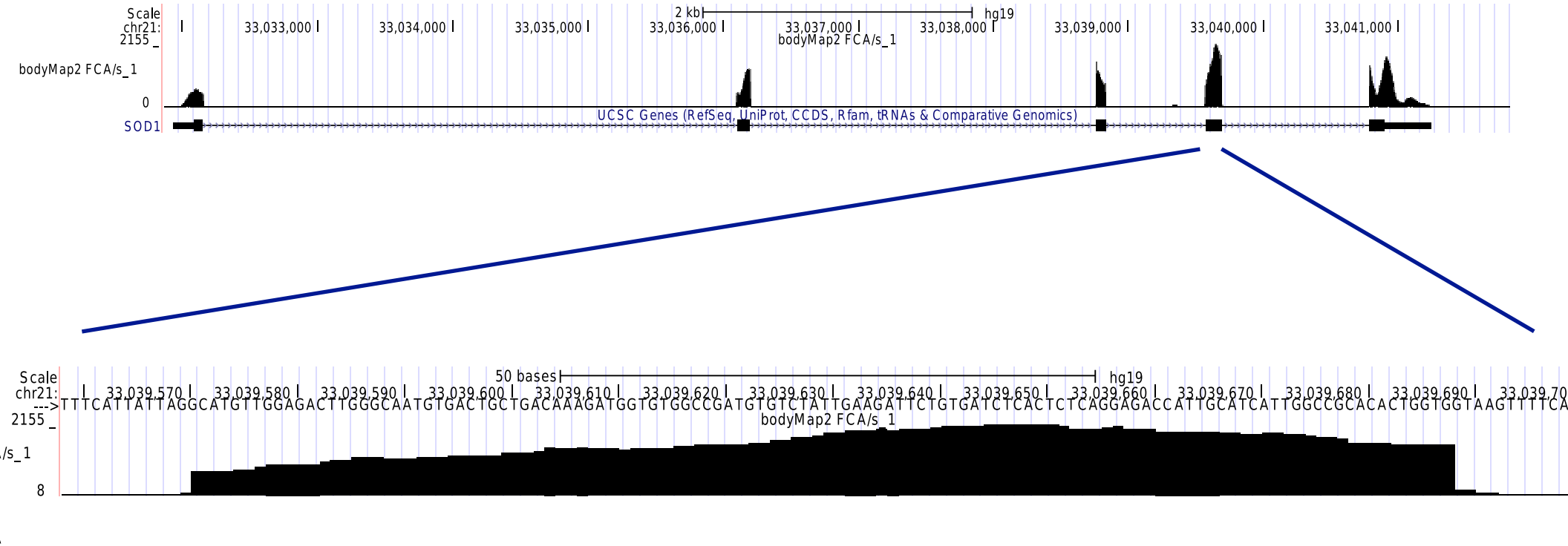
ruzzo@uw.edu

“All High-Throughput Technologies are Crap – Initially”

Q. Morris

7-20-2015

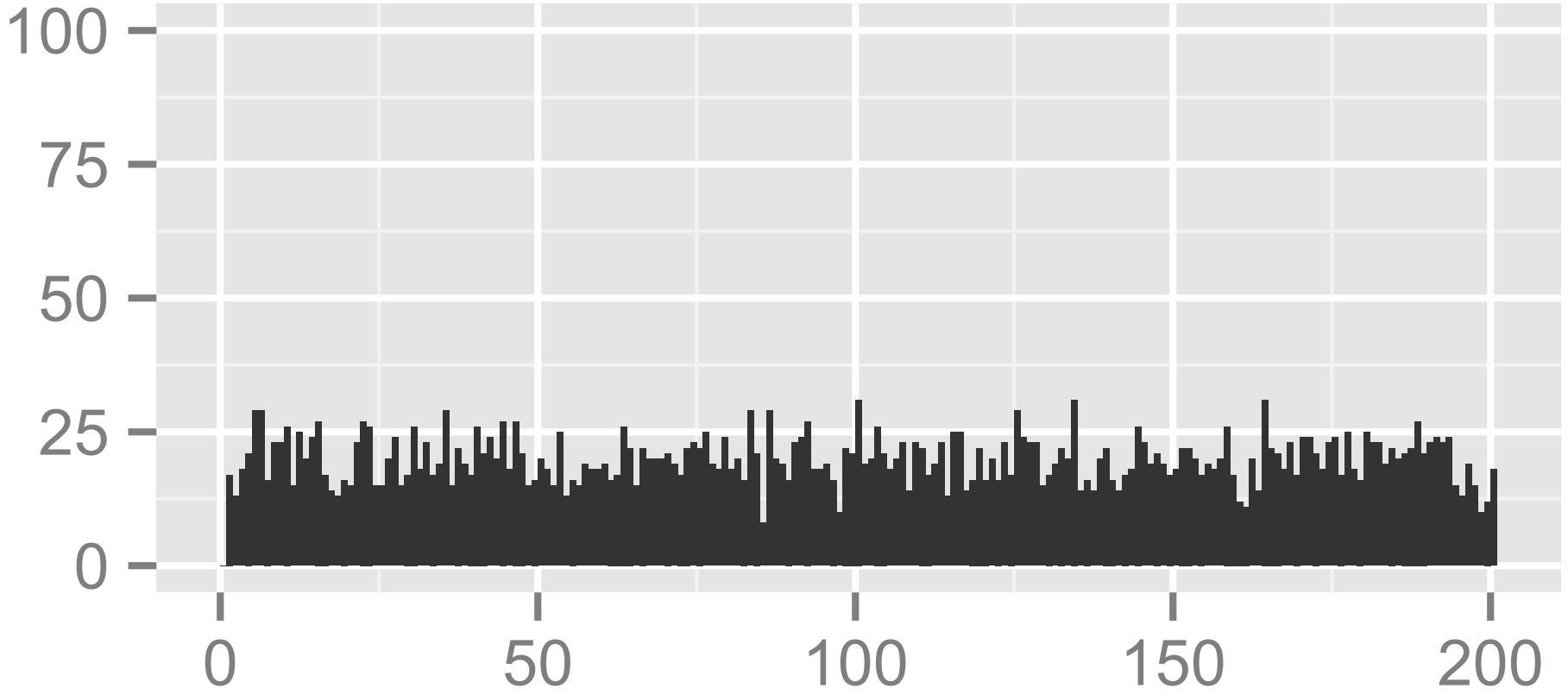
RNA seq



RNA → cDNA, fragment, end repair, A-tail, ligate, PCR, … → Sequence → QC filter, trim, map, … → Count

It's so easy, what could possibly go wrong?

What we expect: Uniform Sampling



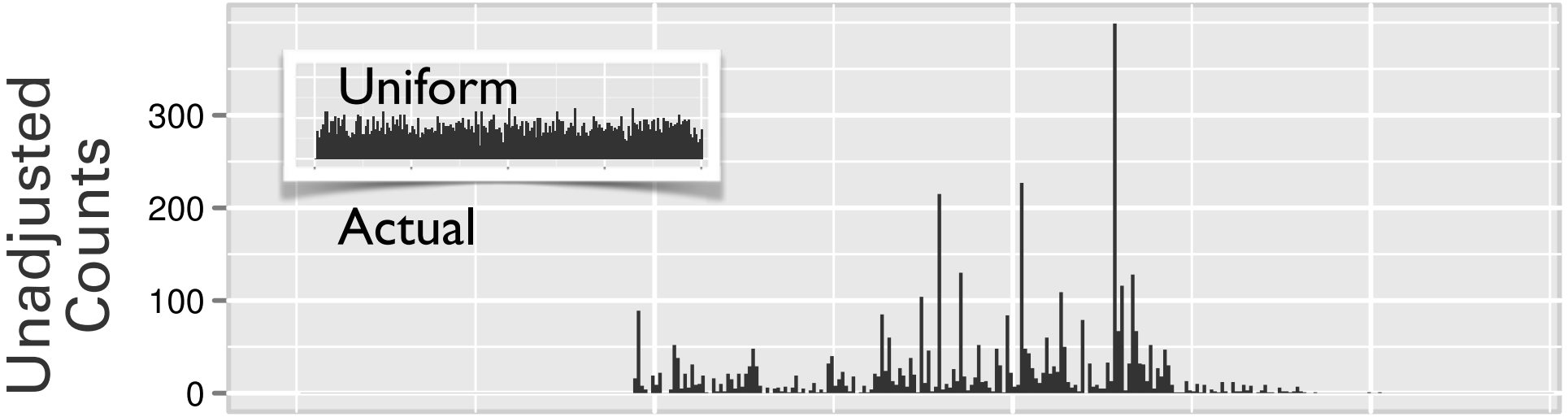
Uniform sampling of 4000 "reads" across a 200 bp "exon."

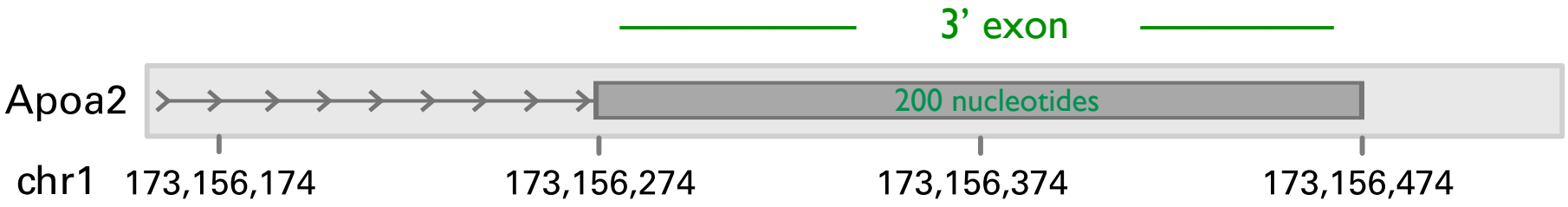
Average 20 ± 4.7 per position, min ≈ 9, max ≈33

I.e., as expected, we see ≈ μ ± 3σ in 200 samples

What we get: *highly* non-uniform coverage

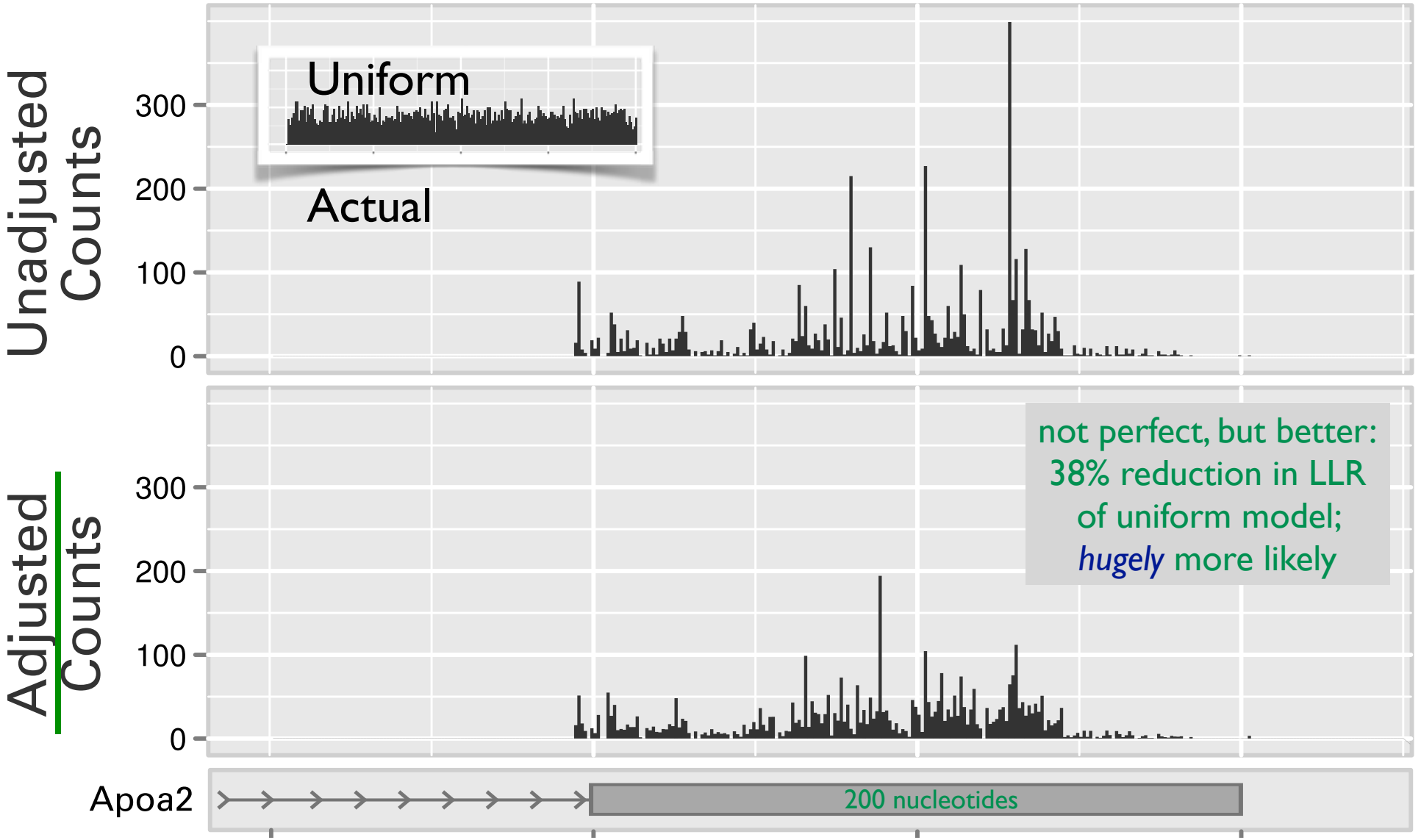
E.g., assuming uniform, the 8 peaks above 100 are ≳ +10σ above mean





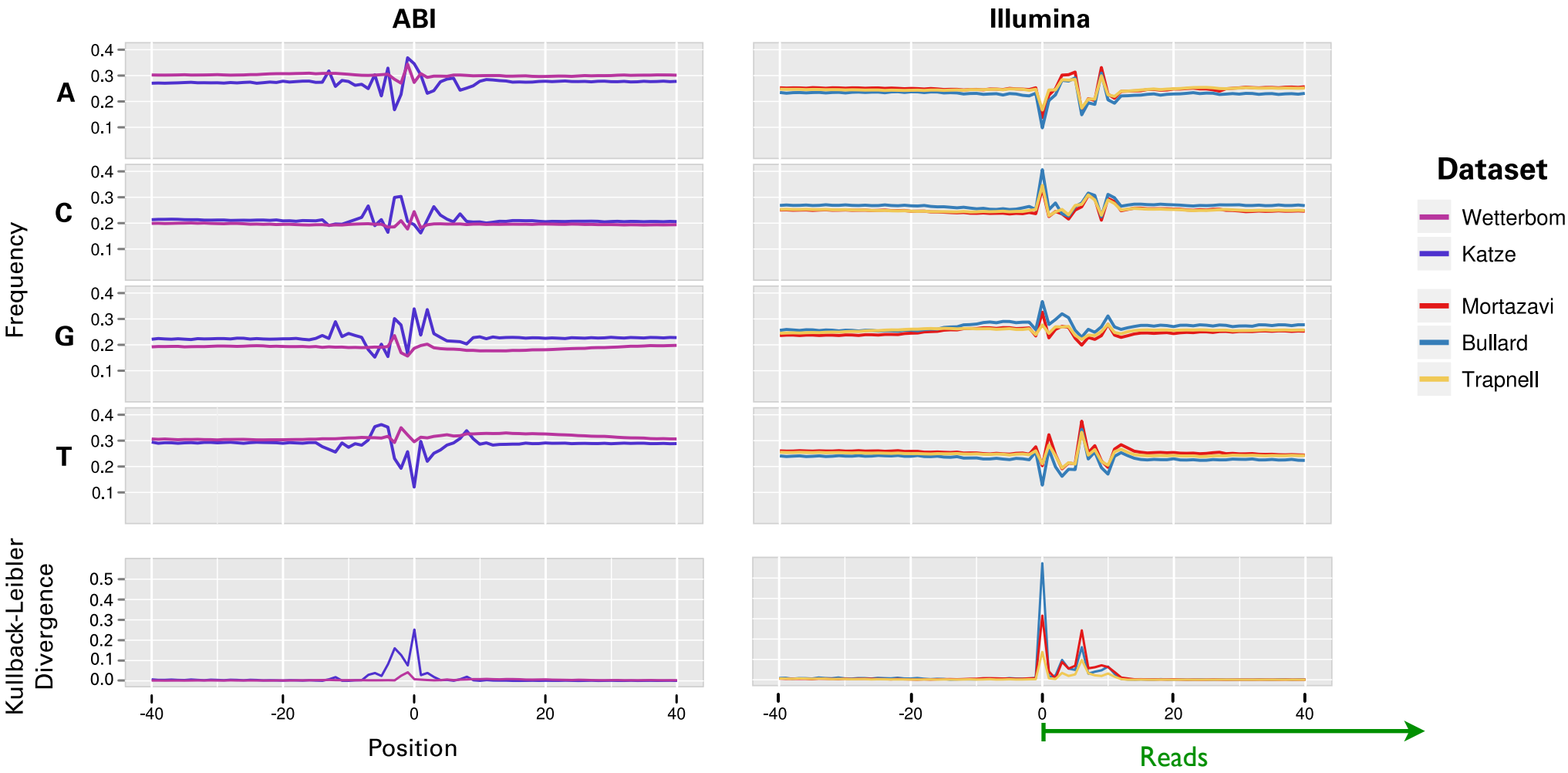
Mortazavi data

What we get: *highly* non-uniform coverage



The Good News: we can (partially) correct the bias

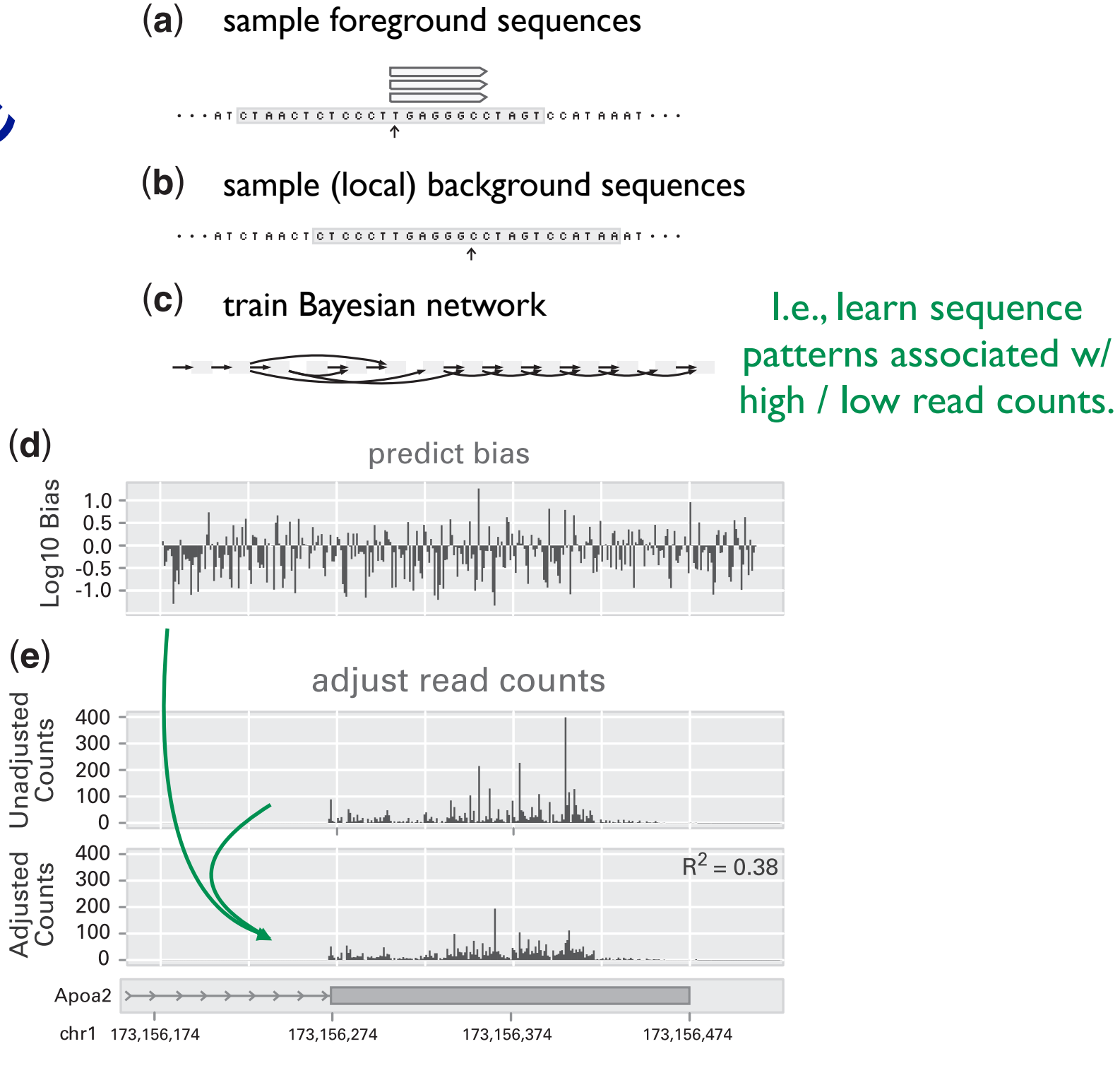
Bias is (in part) sequence-dependent



and platform/sample-dependent

Fitting a model of the sequence surrounding read starts lets us predict which positions have more reads.

Method Outline



I.e., learn sequence patterns associated w/ high / low read counts.

Want a probability distribution over k-mers, k ≈ 40?

Some obvious choices:

Full joint distribution: 4^k-1 parameters

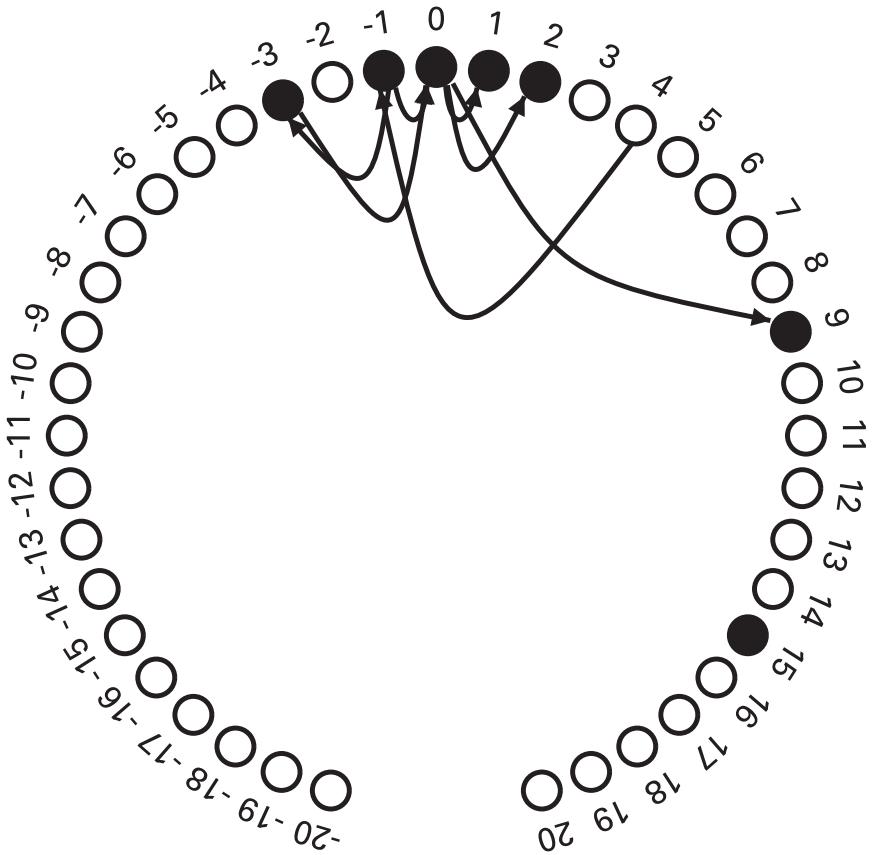
PWM (0-th order Markov): (4-1)•k parameters

Something intermediate:

Directed Bayes network

Form of the models:

Directed Bayes nets



Wetterbom (282 parameters)

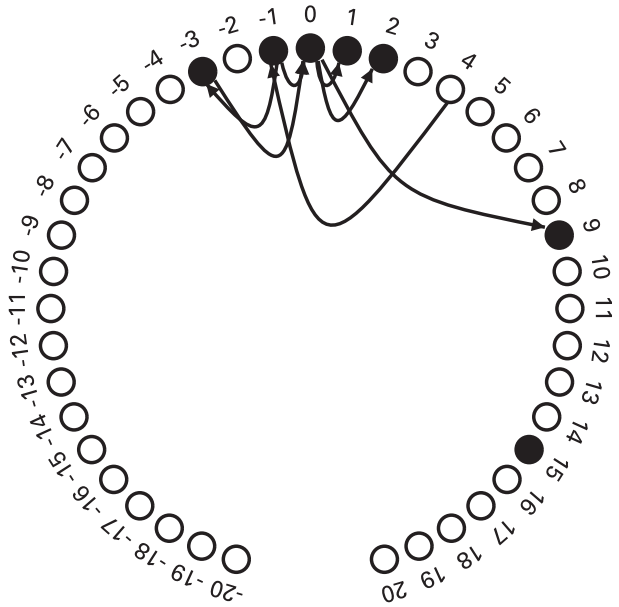
One "node" per nucleotide, ±20 bp of read start

- Filled node means that position is biased
- Arrow i → j means letter at position i modifies bias at j
- For both, numeric parameters say how much

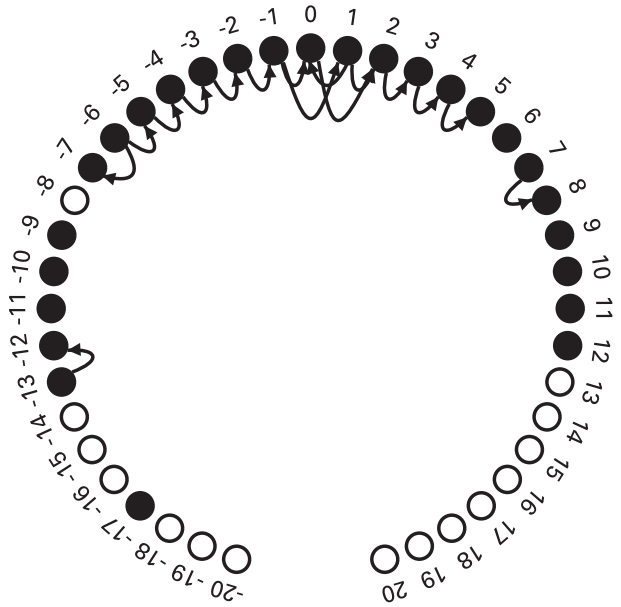
How–optimize:

$$\ell = \sum_{i=1}^{n} \log \Pr[x_i|s_i] = \sum_{i=1}^{n} \log \frac{\Pr[s_i|x_i]\Pr[x_i]}{\sum_{x\in\{0,1\}} \Pr[s_i|x]\Pr[x]}$$

ABI



Wetterbom
(282 parameters)

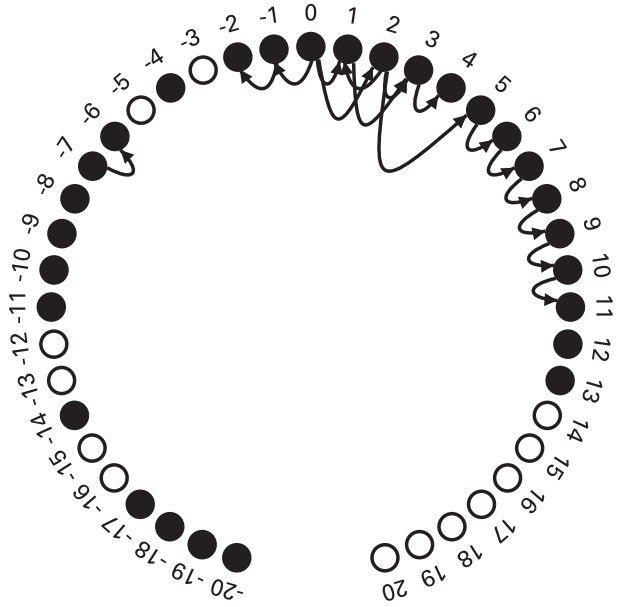


Katze
(684 parameters)

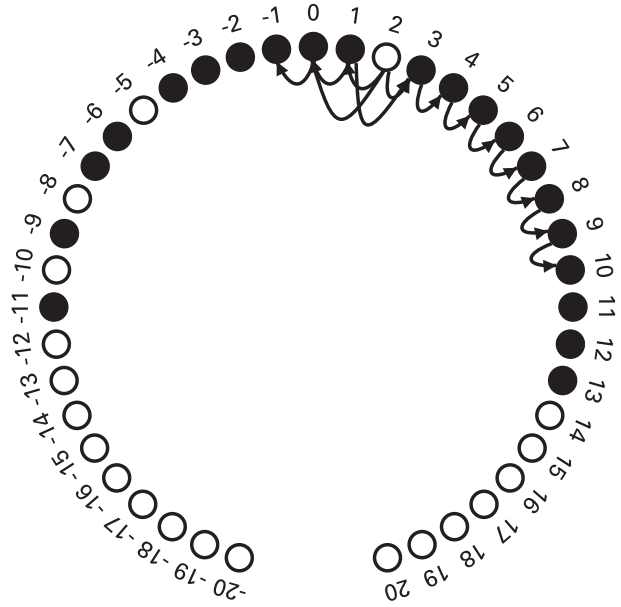
NB:

- Not just initial hexamer
- Span ≥ 19
- All include negative positions
- *All different, even on same platform*

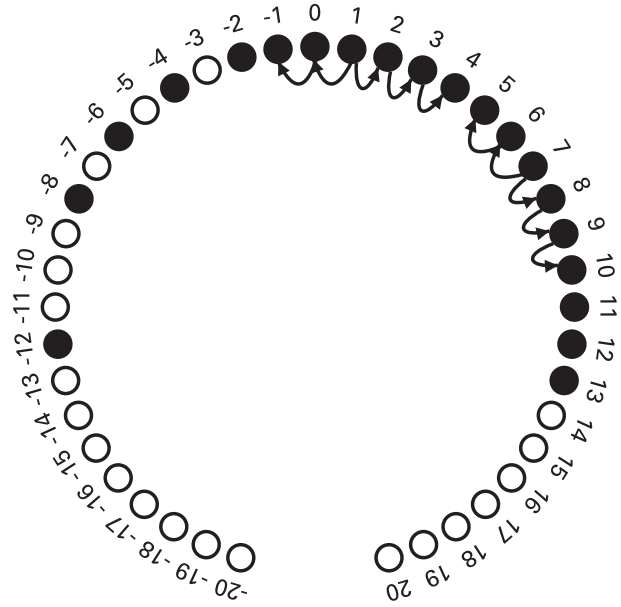
Illumina



Bullard
(696 parameters)

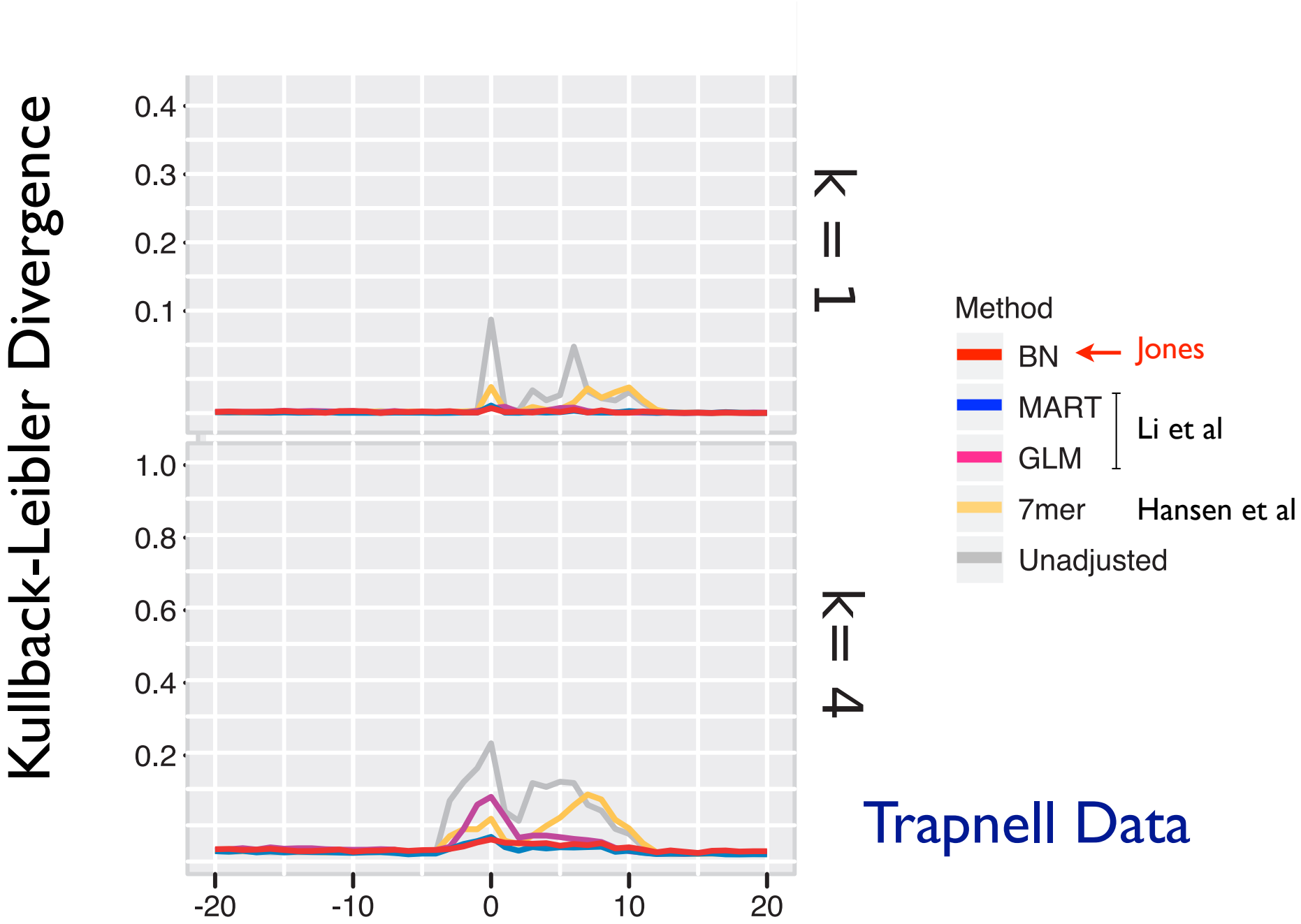


Mortazavi
(582 parameters)



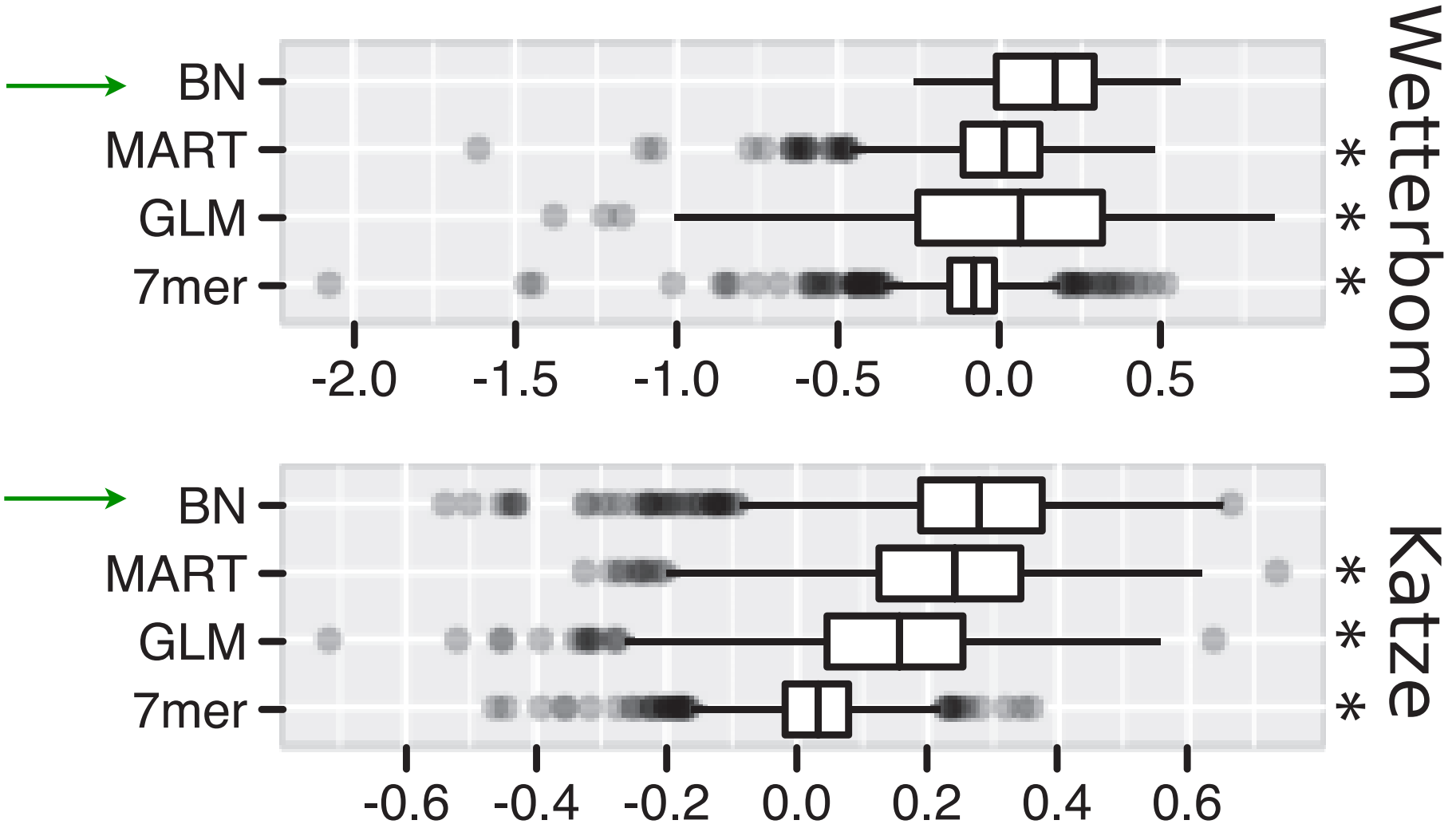
Trapnell
(360 parameters)

Result – Increased Uniformity



Trapnell Data

Result – Increased Uniformity



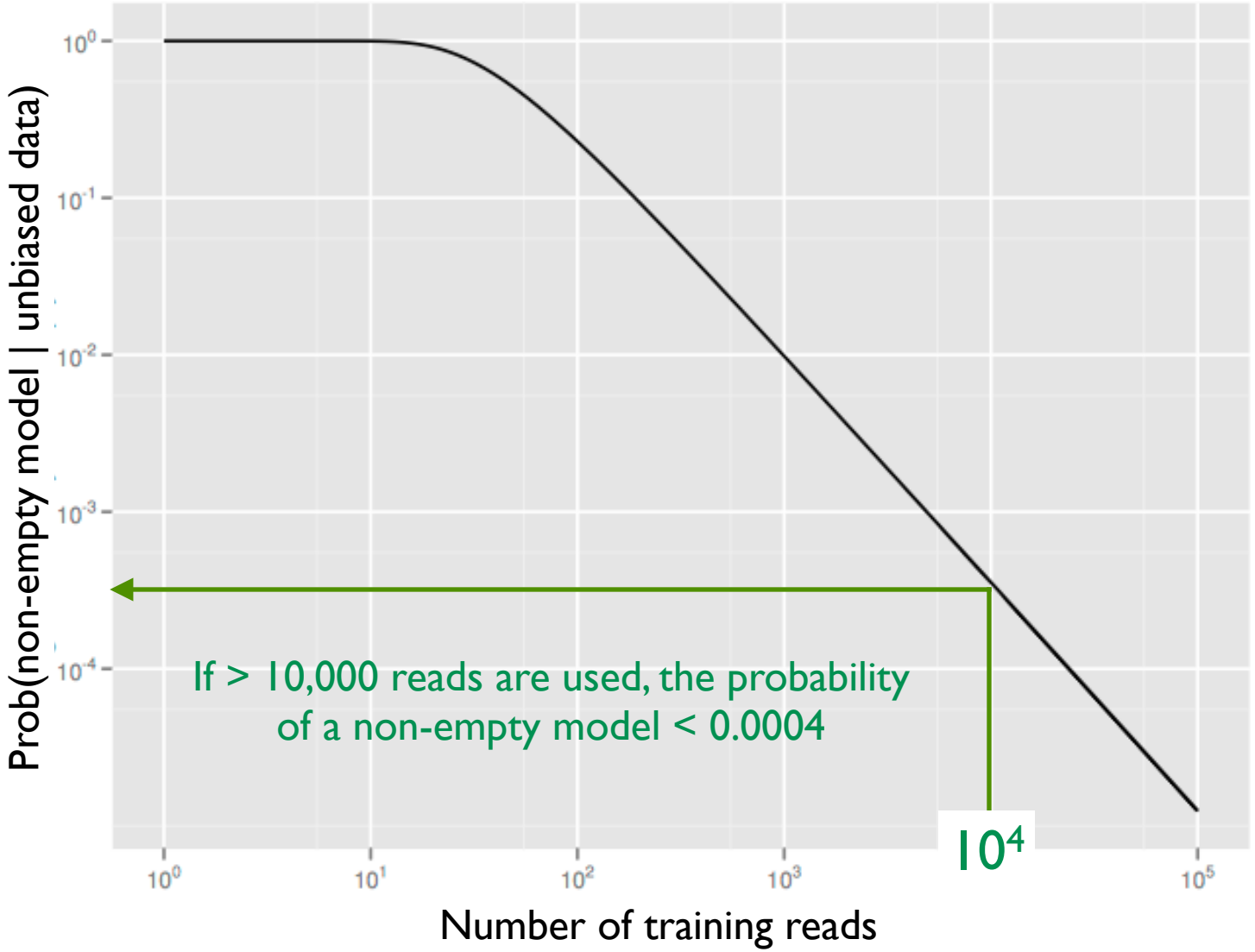
Fractional improvement in log-likelihood under uniform model across 1000 exons (R^2=1-L'/L)

* = p-value < 10^{-23}

hypothesis test:
"Is BN better than X?"
(1-sided Wilcoxon signed-rank test)

"First, do no harm"

Theorem: The probability of "false bias discovery," i.e., of learning a non-empty model from *n* reads sampled from unbiased data, declines *exponentially* with *n*.



… while accuracy and runtime rise with *n* (empirically)

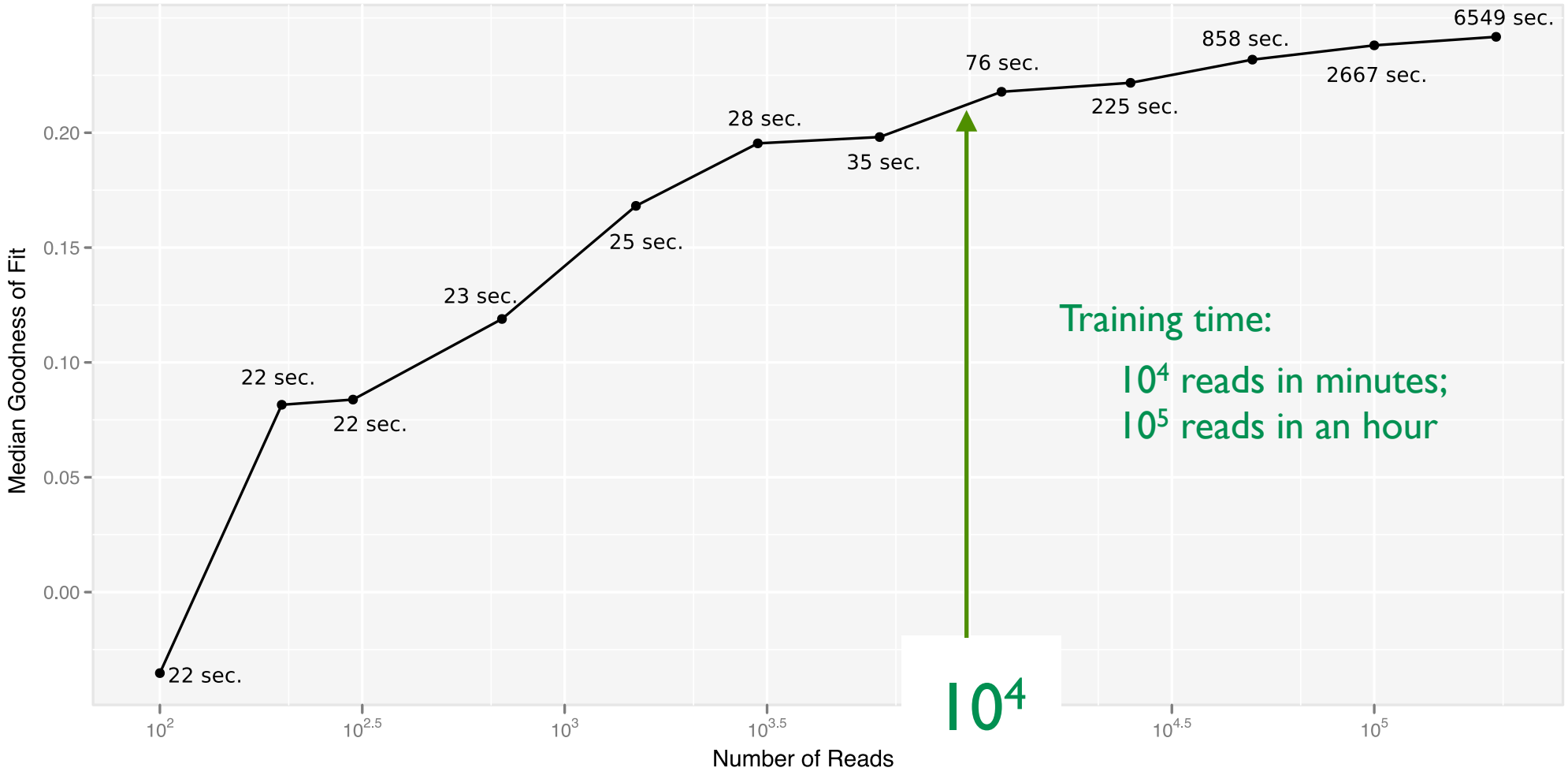


Figure 8: Median R^2 is plotted against training set size. Each point is additionally labeled with the run time of the training procedure.

Possible objection to the approach:

Typical expts compare gene A in sample 1 to *itself* in sample 2. Gene A's sequence is unchanged, "so the bias is the same" & correction is useless/dangerous

Responses:

Bias is *sample-dependent*, to an unknown degree

SNPs and/or alternative splicing might have a big effect, if samples are genetically different and/or engender changes in isoform usage

Atypical experiments, e.g., imprinting, allele specific expression, xenografts, ribosome profiling, ChIPseq, RAPseq, …

Strong control of "false bias discovery" ⇒ *little risk*

BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 7 2012, pages 921–928
doi:10.1093/bioinformatics/bts055

Gene expression

Advance Access publication January 28, 2012

A new approach to bias correction in RNA-Seq

Daniel C. Jones[1,*], Walter L. Ruzzo[1,2,3], Xinxia Peng[4] and Michael G. Katze[4]

[1]Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195-2350, [2]Department of Genome Sciences, University of Washington, Seattle, WA 98195-5065, [3]Fred Hutchinson Cancer Research Center, Seattle, WA 98109 and [4]Department of Microbiology, University of Washington, Seattle, WA

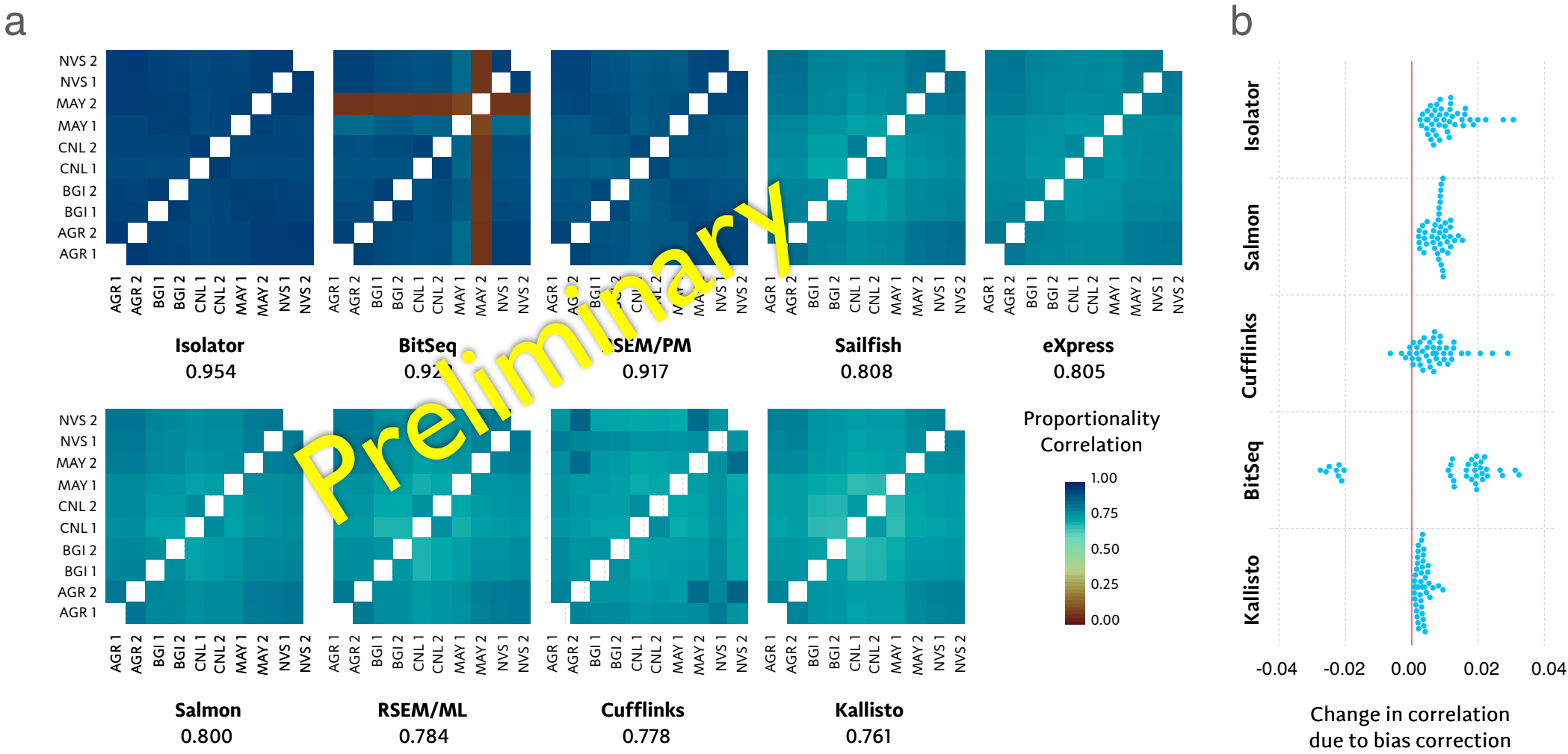
Associate Editor: Alex Bateman

ABSTRACT

Motivation: Quantification of sequence abundance in RNA-Seq experiments is often conflated by protocol-specific sequence bias. The exact sources of the bias are unknown, but may be influenced by

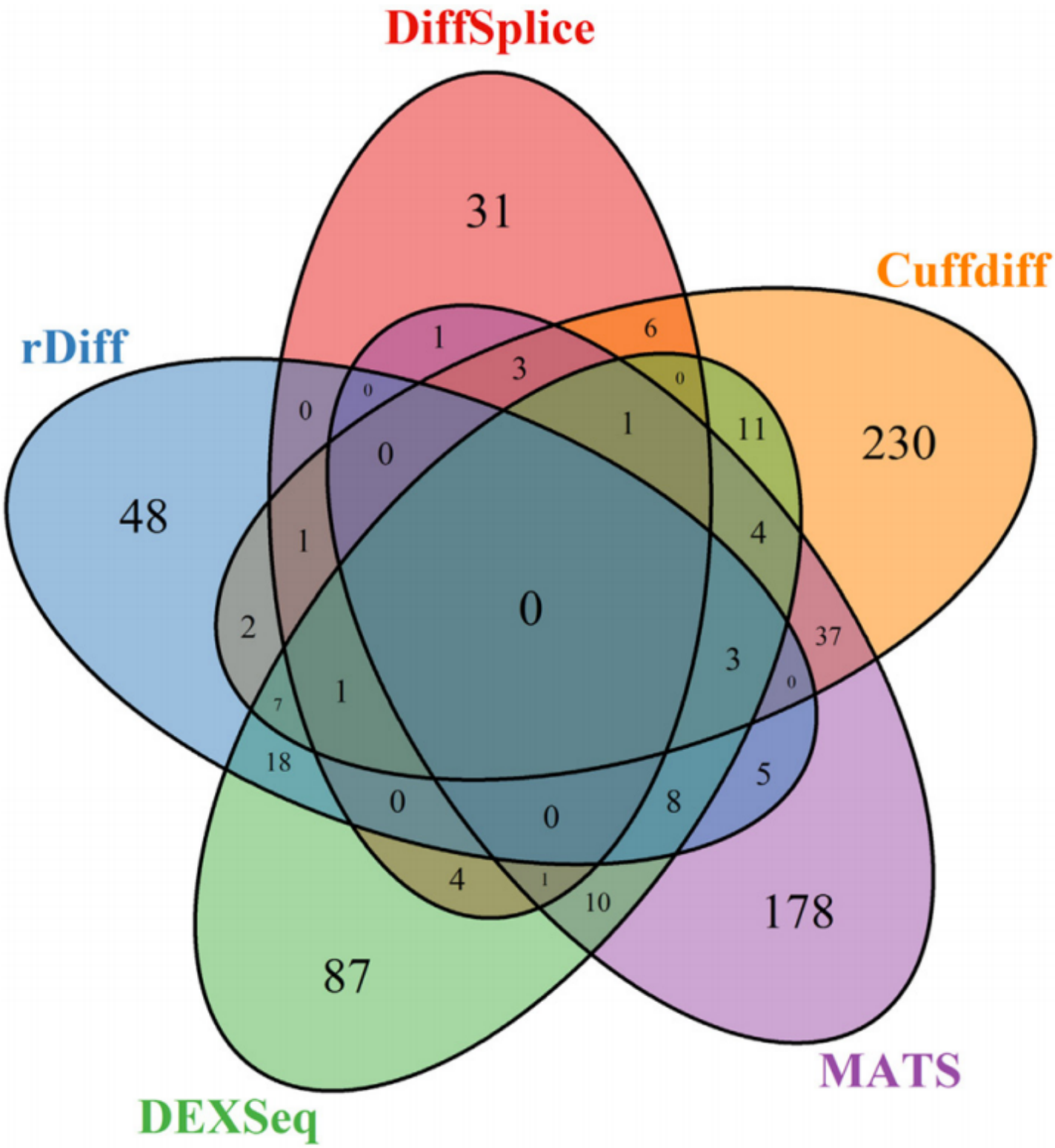
These biases may adversely affect [illegible] low level [illegible]

Batch Effects? YES!



A: Pairwise proportionality correlation between samples sequenced on 2 flowcells each at 5 sites. B: The absolute change in correlation induced by enabling bias correction (where available). For clarity, BitSeq estimates of "MAY 2", excluded; bias correction was extremely detrimental there.

Alternate Splicing



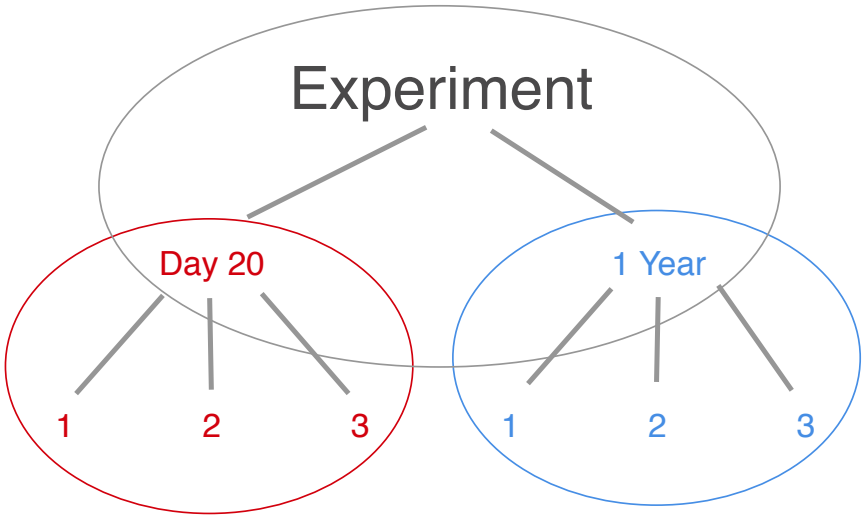
Liu, et al. BMC Bioinformatics 15.1 (2014): 364

Isolator

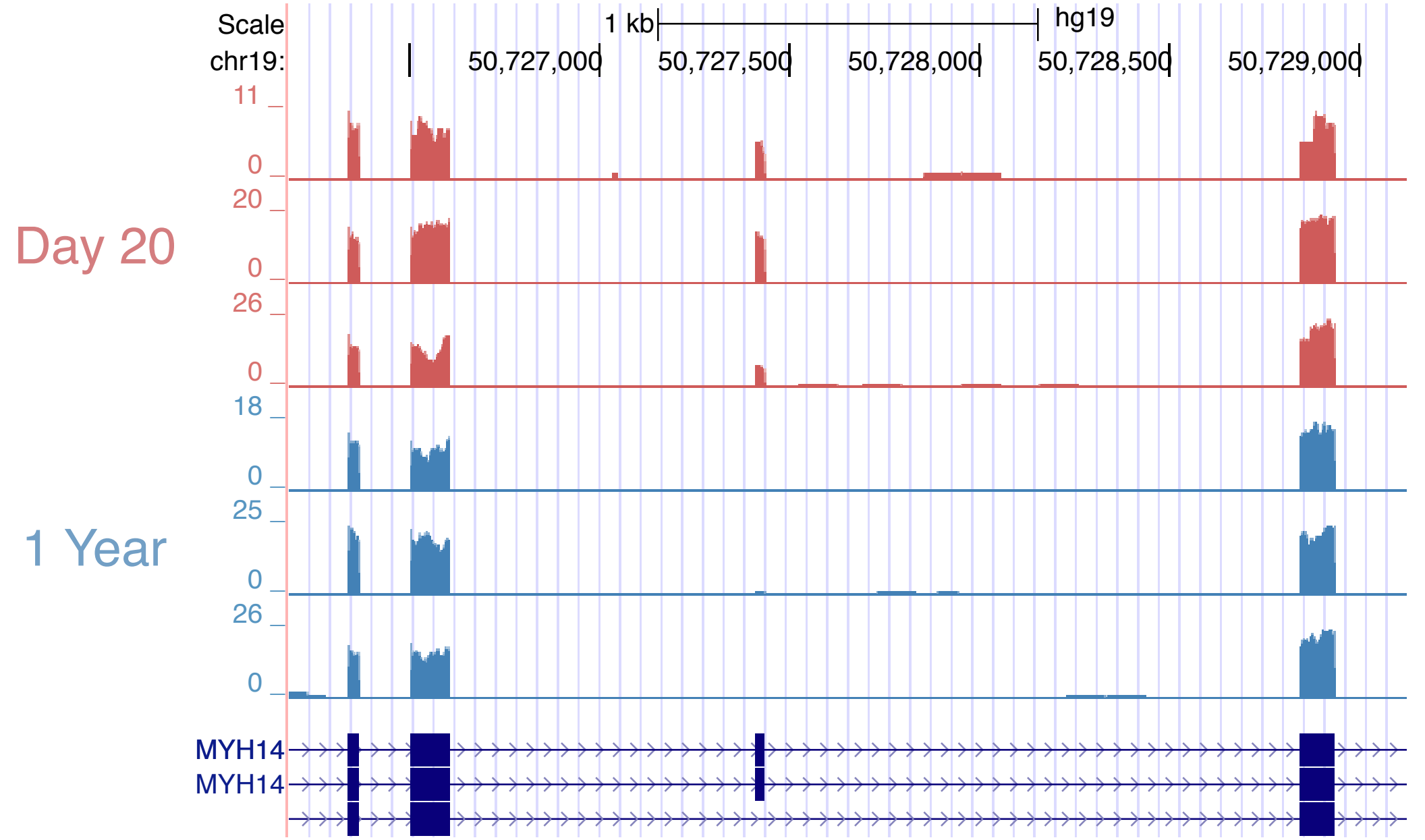
Soon to be the world's best isoform quantitation tool

Bayesian hierarchical model + fast MCMC sampler
give mean and *uncertainty* in estimates
Can handle dozens of RNAseq samples per hour

When data is lacking, estimates are shrunk towards each other, supressing suprious changes.



1 read vs. 2 reads is probably not a 2-fold change in transcription!


Day 20
1 Year
Scale
1 kb
hg19
chr19:
50,727,000
50,727,500
50,728,000
50,728,500
50,729,000
11
0
20
0
26
0
18
0
25
0
26
0
MYH14
MYH14

Method	A	B	C	D
Isolator	**0.878**	**0.866**	**0.839**	**0.852**
Cufflinks	0.870	0.856	0.799	0.841
eXpress	0.870	0.855	0.829	0.840
Salmon	0.866	0.852	0.826	0.836
RSEM/ML	0.865	0.851	0.825	0.835
BitSeq	0.840	0.821	0.802	0.813
Kallisto	0.858	0.840	0.817	0.826
Sailfish	0.844	0.814	0.797	0.802
RSEM/PM	0.840	0.822	0.803	0.811

Table 2: Proportionality correlation between gene level quantification of 18353 genes using PrimePCR qPCR and RNA-Seq quantification.

Method	c vs $0.75a + 0.25b$	d vs $0.25a + 0.75b$
Isolator	**0.975**	**0.975**
BitSeq	0.967	0.967
RSEM/PM	0.968	0.967
Sailfish	0.932	0.925
RSEM/ML	0.922	0.919
Salmon	0.916	0.914
Kallisto	0.907	0.902
eXpress	0.903	0.899
Cufflinks	0.870	0.916

Table 5: Proportionality correlation between gene-level estimates for the mixed samples C and D and weighted averages of estimates for A and B, corresponding to the mixture proportions for C and D.

RNAseq data shows strong technical biases

Of course, compare to appropriate control samples

But that's not enough, due to:

- batch effects
- SNPs/genetic heterogeneity
- alt splicing
- …

all of which tend to differently bias sample/control

"All hight-throughput technologies are crap, initially,"

BUT careful modeling can help.

Acknowledgements

Daniel Jones

Katze Lab

Michael Katze
Xinxia Peng

P01 Labs

Tony Blau, Chuck Murry, Hannele Ruohola-Baker, Nathan Palpant, Kavitha Kuppusamy, …

Funding

NIGMS, NHGR, NIAID