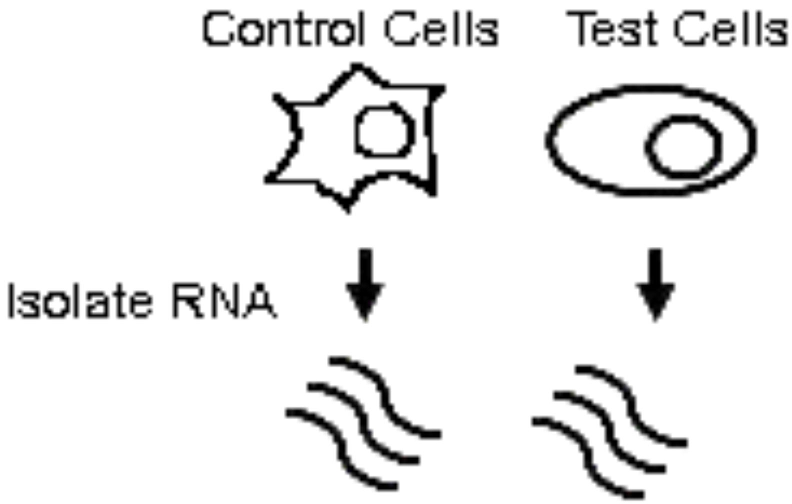
Bias in RNA sequencing and what to do about it

Walter L. (Larry) Ruzzo

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

ruzzo@uw.edu



DNA Sequencer

Millions of reads, say, 100 bp each

map to genome, compare & analyze

Goals of RNAseq

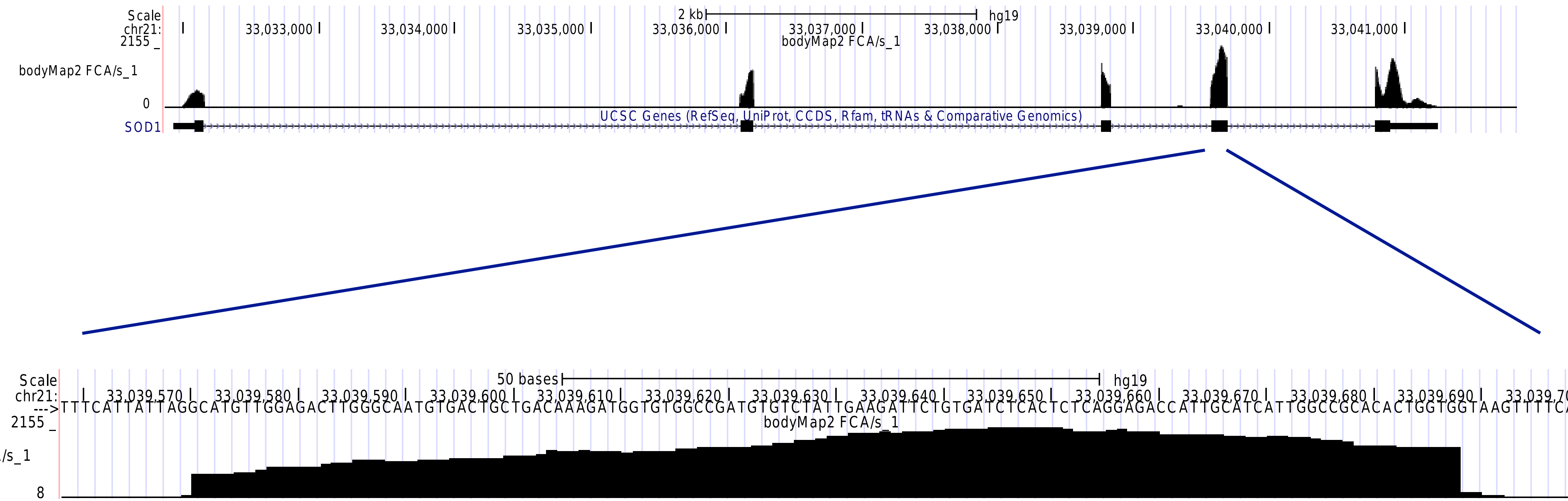
1. Which genes are being expressed?
   How? *assemble* reads (fragments of mRNAs) into (nearly) full-length mRNAs and/or *map* them to a reference genome

2. How highly expressed are they?
   How? *count* how many fragments come from each gene–expect more highly expressed genes to yield more reads per unit length

3. What's same/diff between 2 samples
   E.g., tumor/normal

4. ...

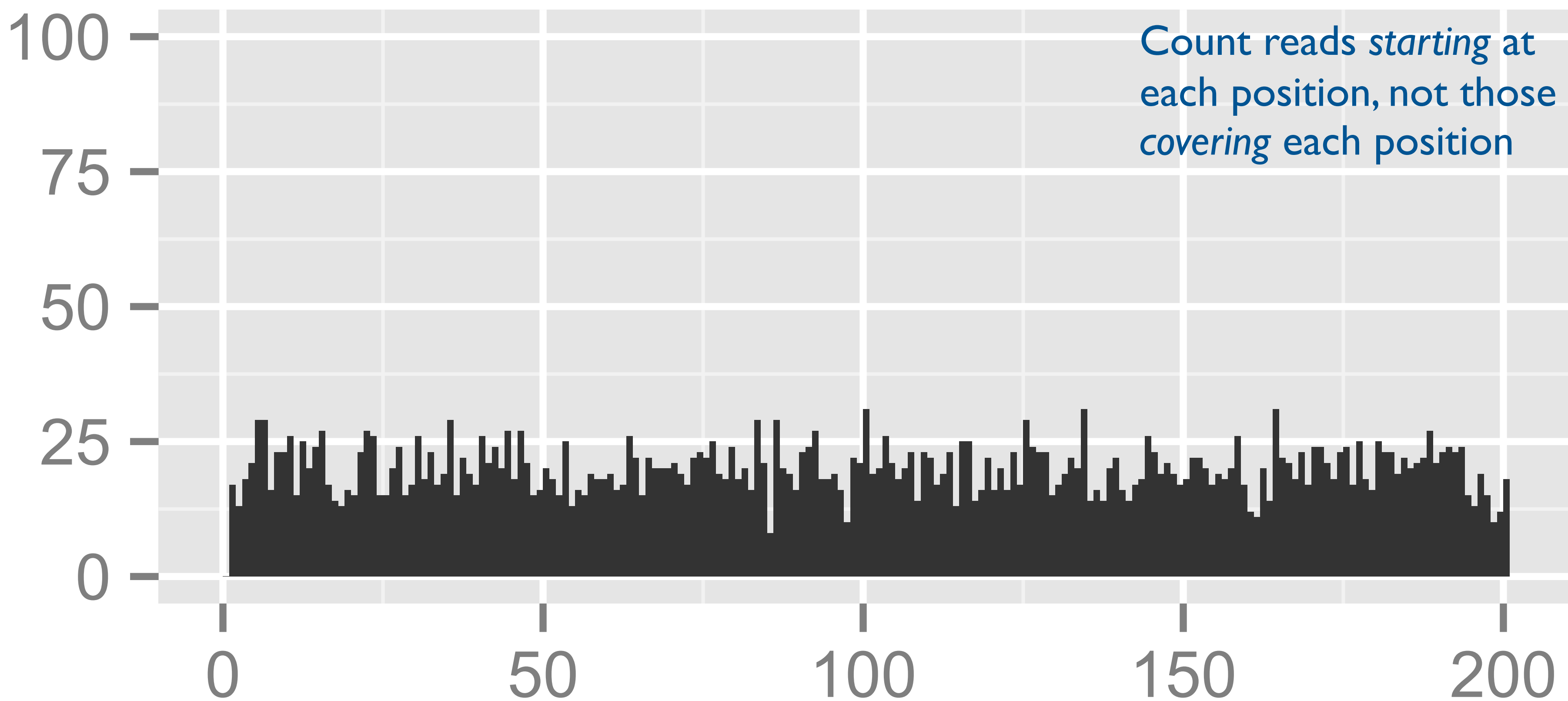
RNA seq



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

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

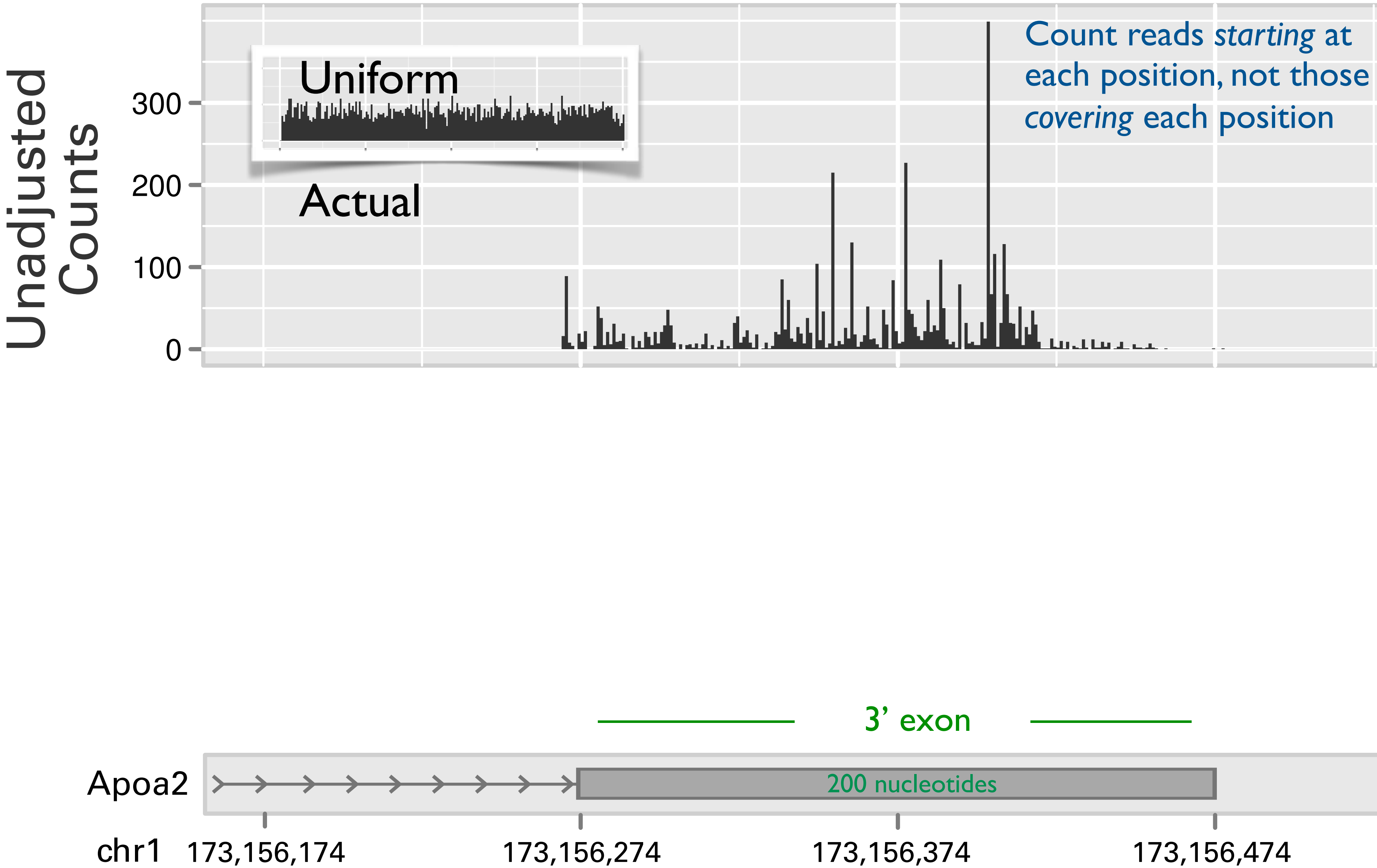
What we expect: Uniform Sampling



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

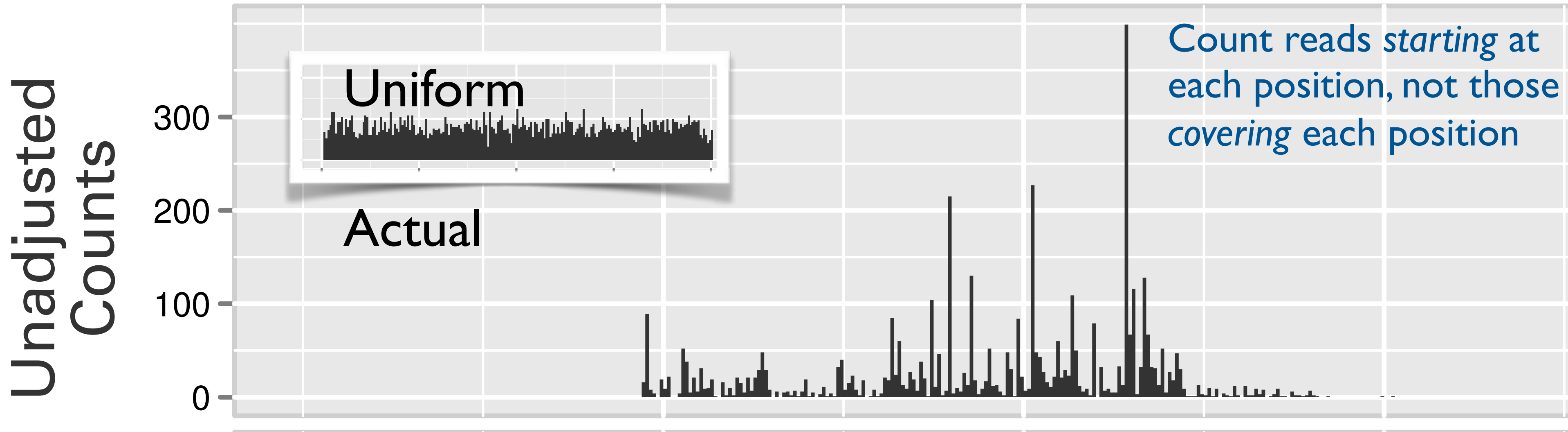
E.g., assuming uniform, the 8 peaks above 100 are ≳ $+10\sigma$ above mean



Mortazavi data

What we get: *highly* non-uniform coverage

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



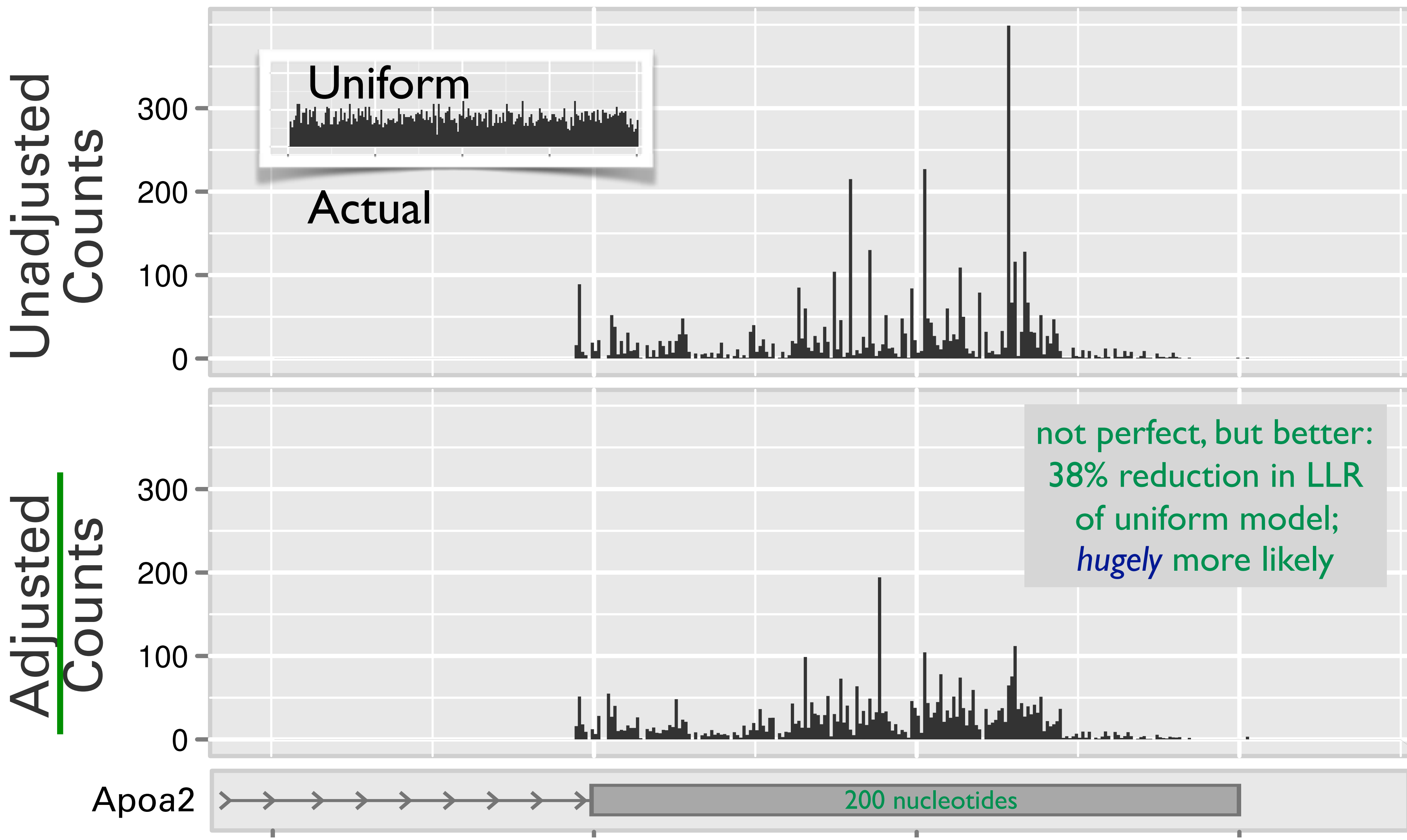
How to make it more uniform?

A: Math tricks like averaging/smoothing (e.g. "coverage") or transformations ("log"), …, or

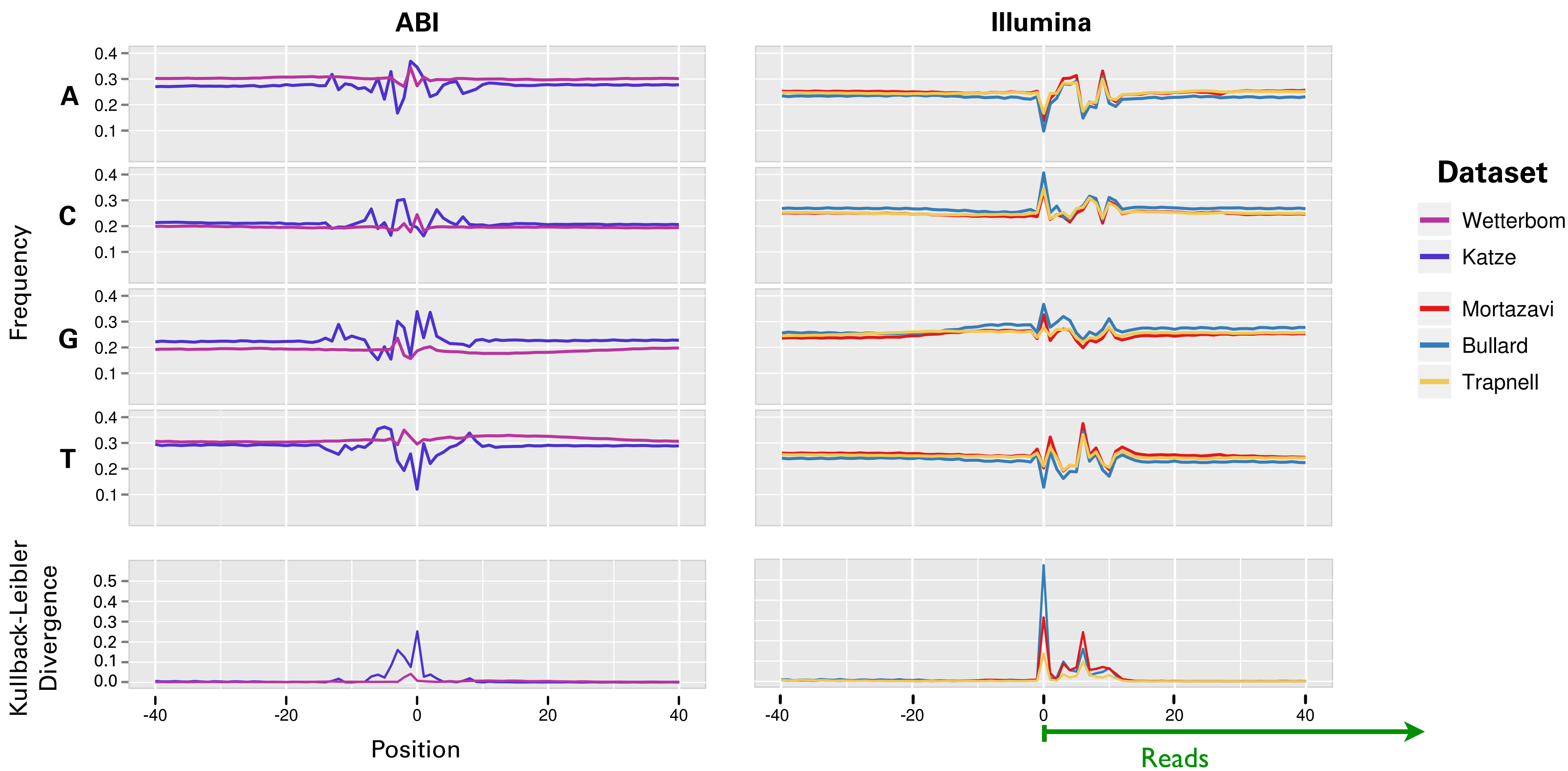
B: Try to model (aspects of) causation ← *WE DO THIS*

(& use increased uniformity of result as a measure of success)

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.

No one knows in any great detail

Speculations:

- *all* steps in the complex protocol may contribute
- E.g.,
  - primers in PCR-like amplification steps may have unequal affinities ("random hexamers", e.g.)
  - ligase enzyme sequence preferences
  - potential RNA structures
  - fragmentation biases
  - mapping biases

Hansen, et al. 2010

“7-mer” method - directly count foreground/ background 7-mers at read starts, correct by ratio

$2 * (4^7-1) = 32766$ free parameters

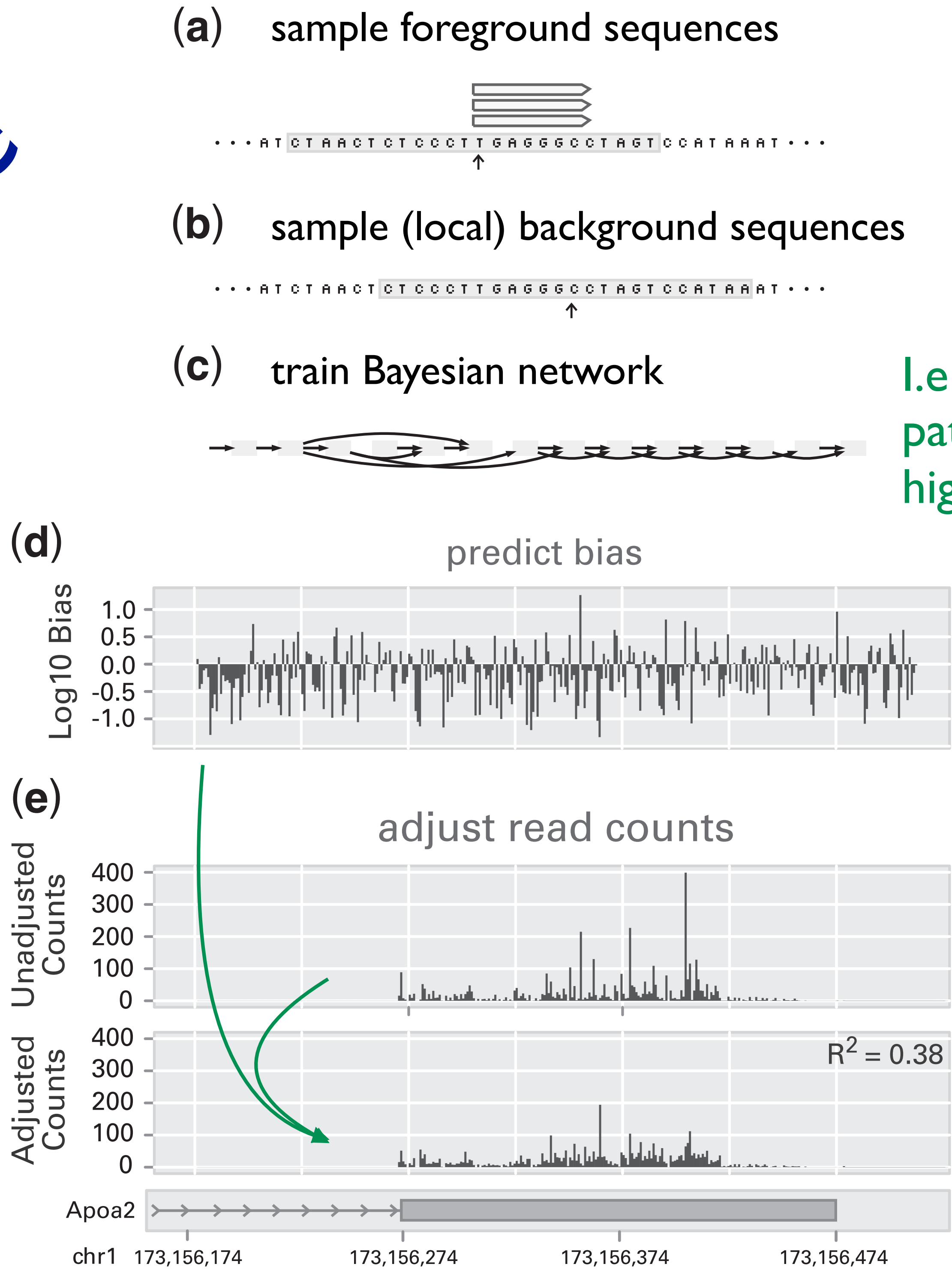
Li, et al. 2010

GLM - generalized linear model

MART - multiple additive regression trees

training requires gene annotations

Method Outline



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

Data is *Un*biased if read is independent of sequence:

$$\Pr(\text{ read at i }) = \Pr(\text{ read at i } | \text{ sequence at i })$$

From Bayes rule:

$$\Pr(\text{ read at i } | \text{ seq at i }) = \boxed{\frac{\Pr(\text{ seq at i } | \text{ read at i })}{\Pr(\text{ seq at i})}} \Pr(\text{ read at i })$$

We define "bias" to be this factor

Want a probability distribution over k-mers, $k \approx 40$?

Some obvious choices:

Full joint distribution: 4^k-1 parameters

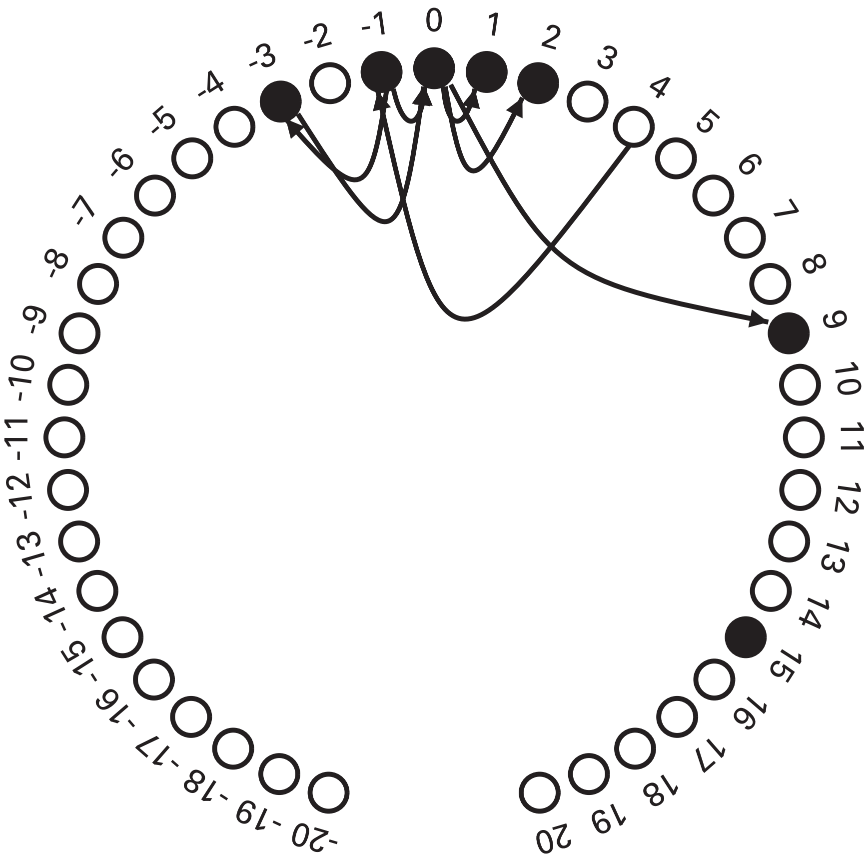
PWM (0-th order Markov): $(4-1)\cdot k$ parameters

Something intermediate:

Directed Bayes network

Form of the models:

Directed Bayes nets



Wetterbom
(282 parameters)

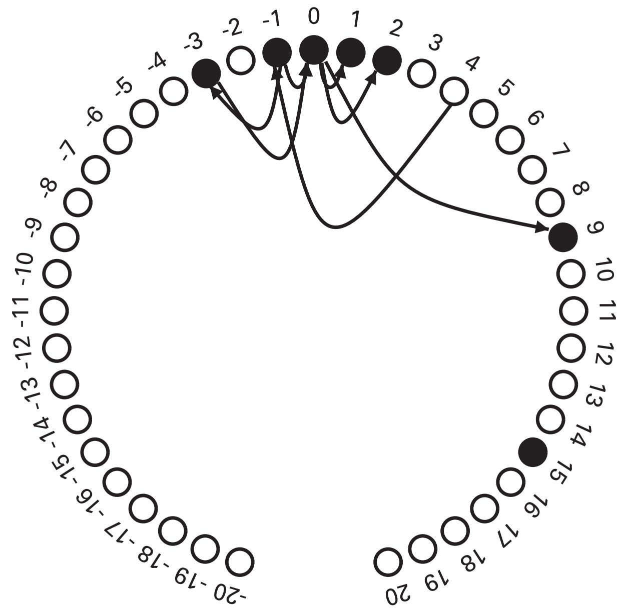
One "node" per nucleotide, ±20 bp of read start

- 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

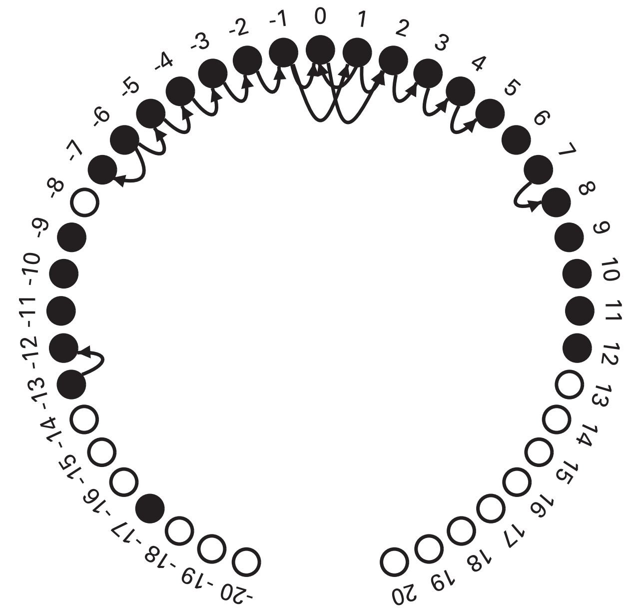
How–optimize:

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

ABI



Wetterbom
(282 parameters)

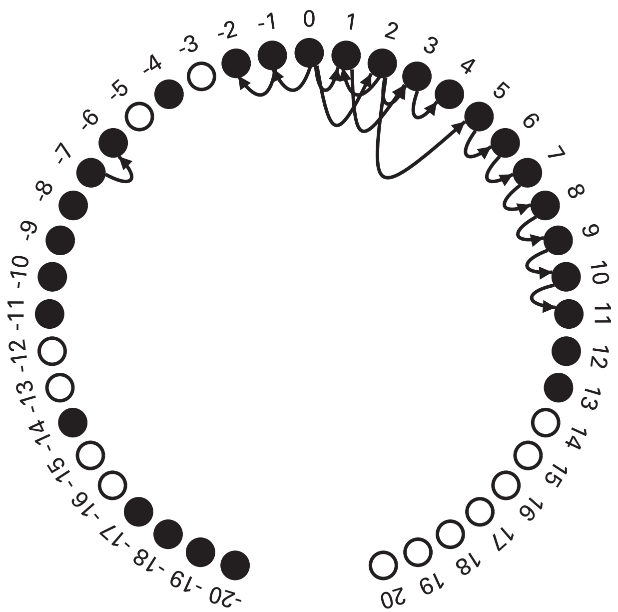


Katze
(684 parameters)

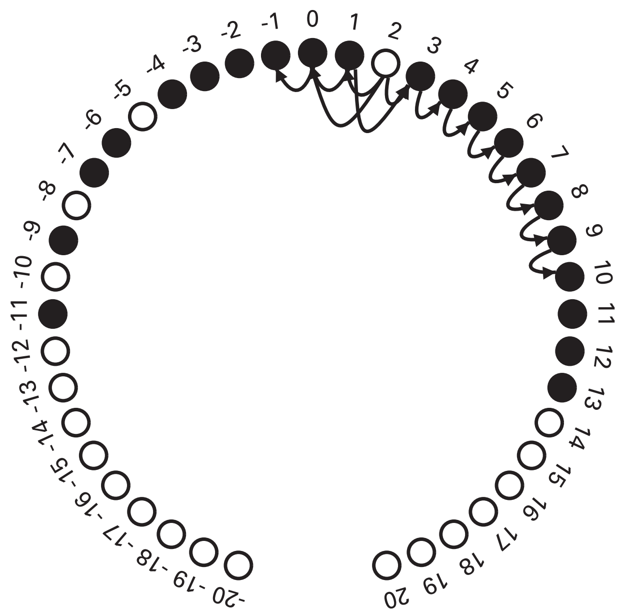
NB:

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

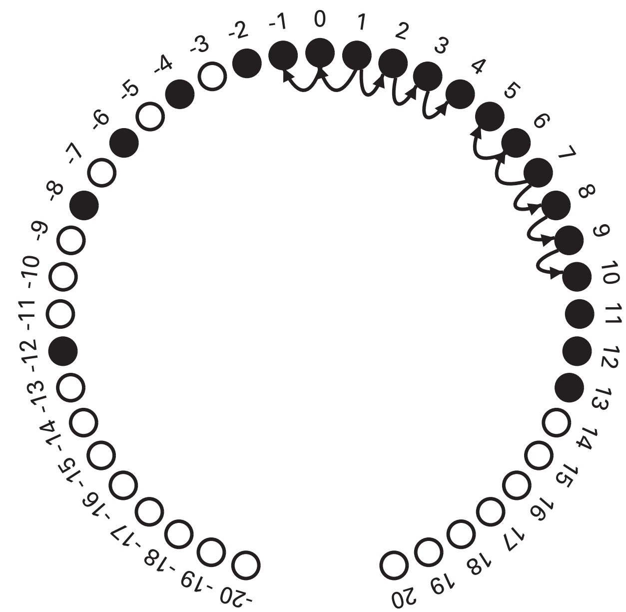
Illumina



Bullard
(696 parameters)

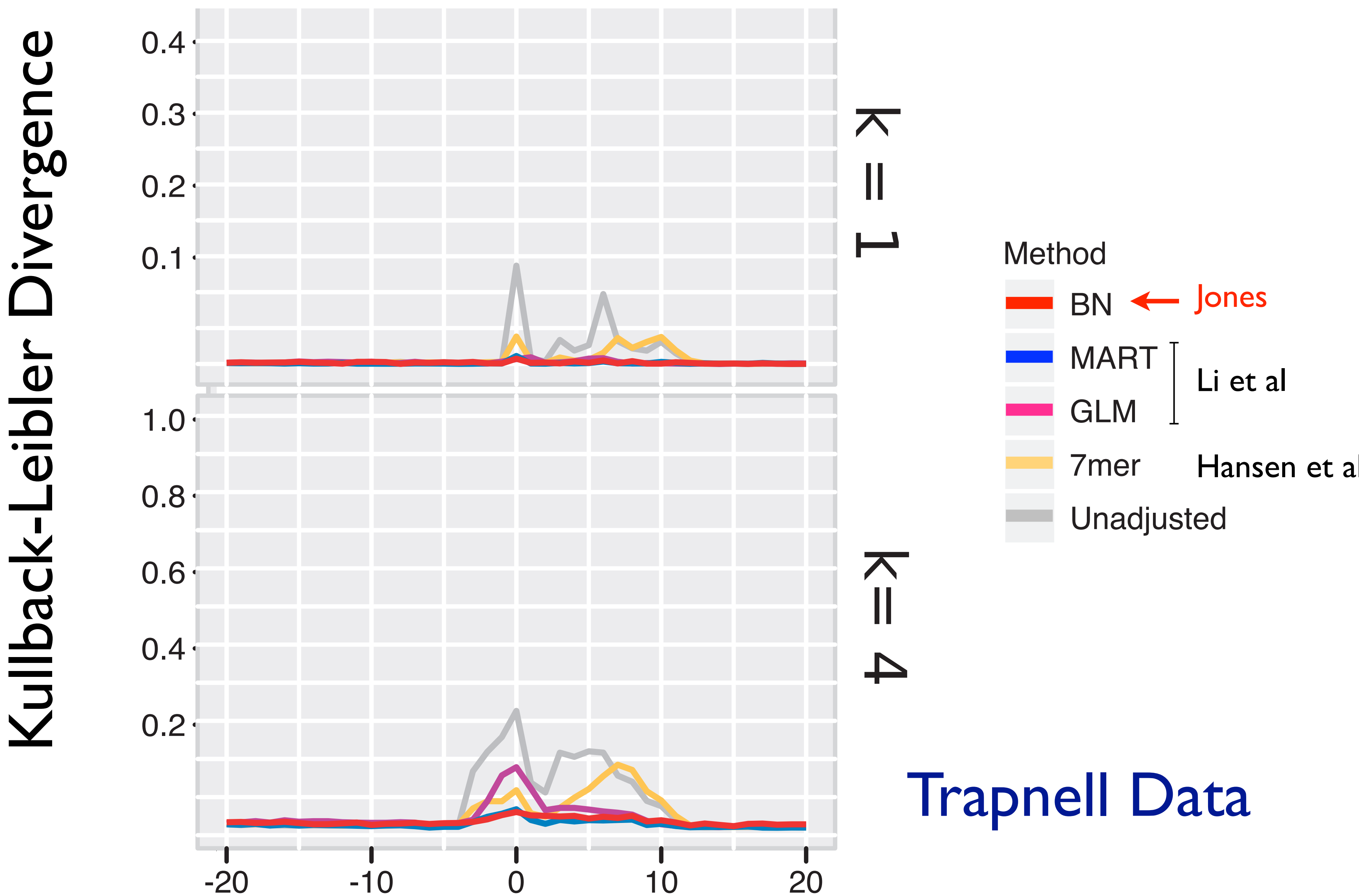


Mortazavi
(582 parameters)



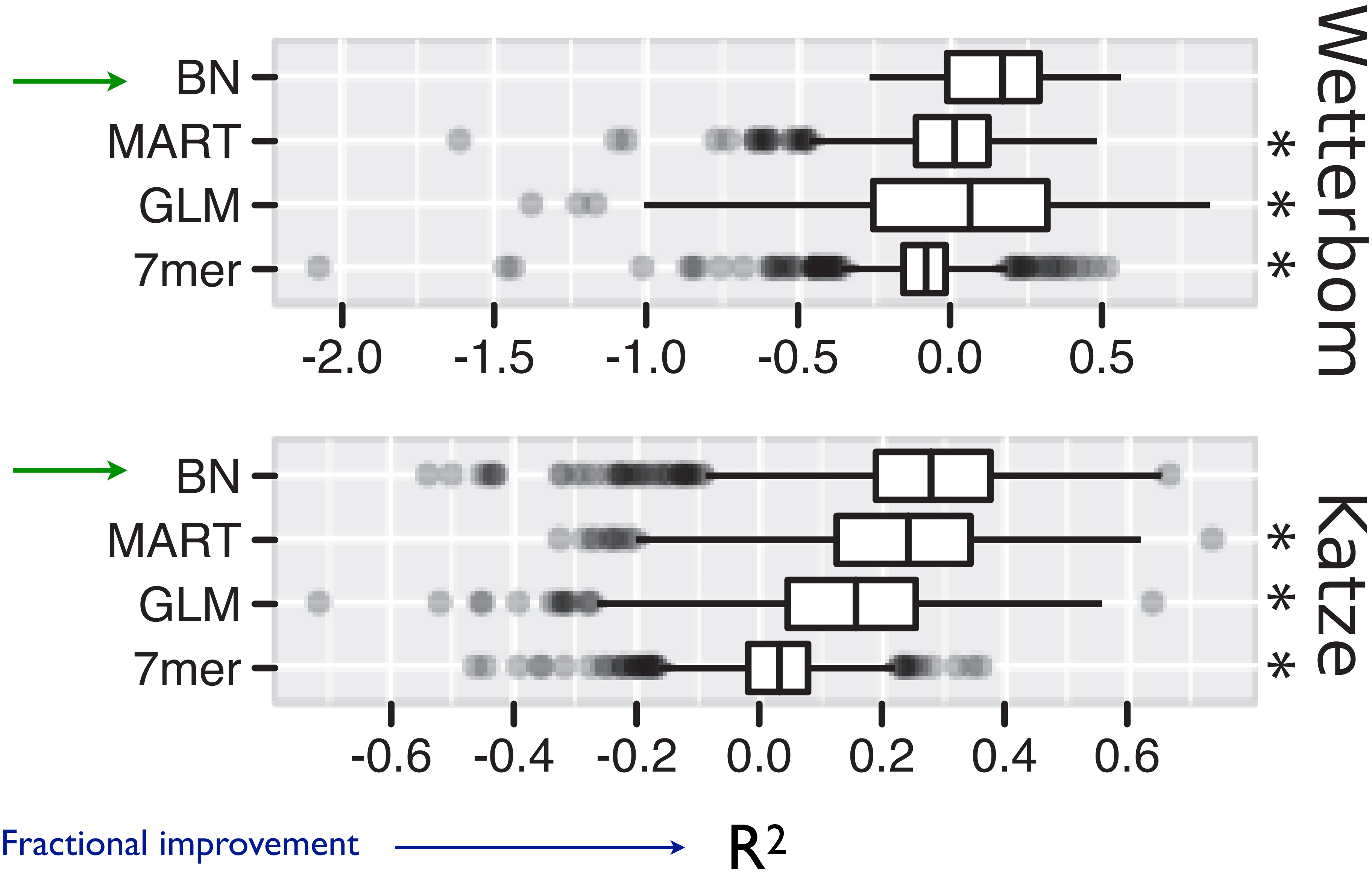
Trapnell
(360 parameters)

Result – Increased Uniformity



Trapnell Data

Result – Increased Uniformity

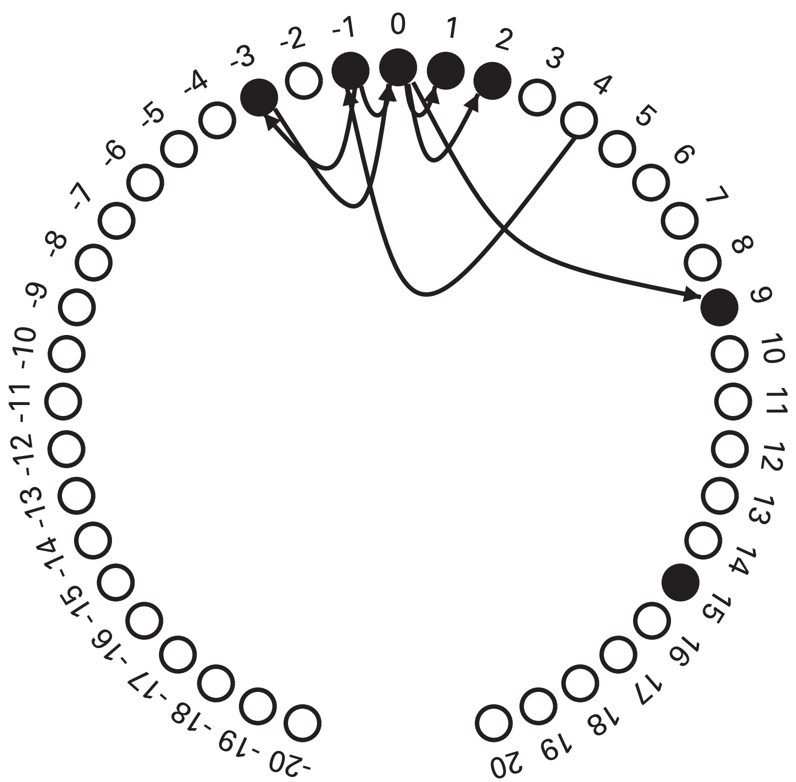


Fractional improvement in log-likelihood under uniform model across 1000 exons ($R^2=1-L'/L$)

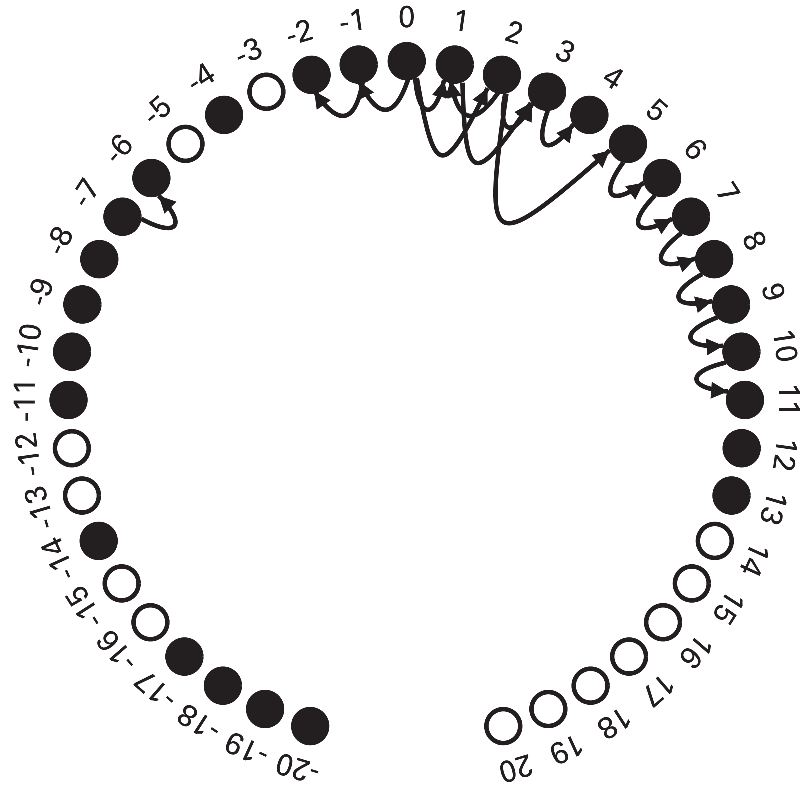
* = p-value < 10^{-23}

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

What is the chance that we will learn an incorrect model? E.g., learn a biased model from unbiased input?



Wetterbom
(282 parameters)



Bullard
(696 parameters)

How does the amount of training data effect accuracy of the resulting model?

Probability of falsely inferring "bias" from an unbiased sample declines rapidly with size of training set (provably) ...

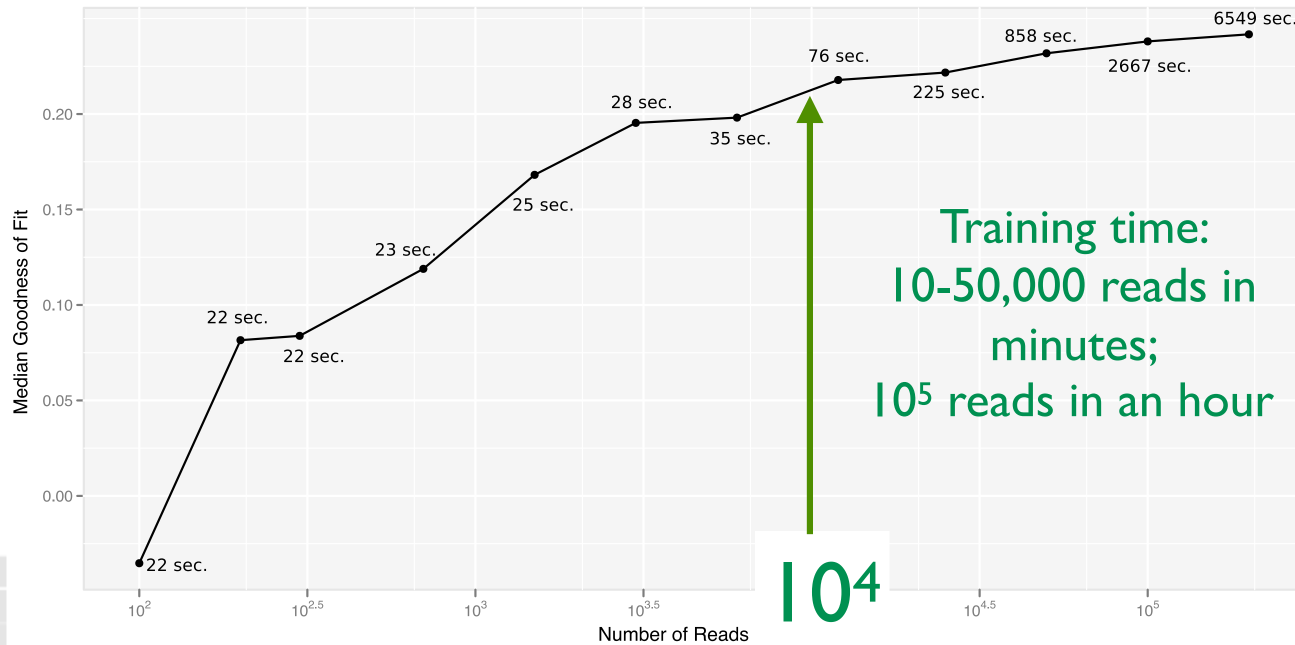
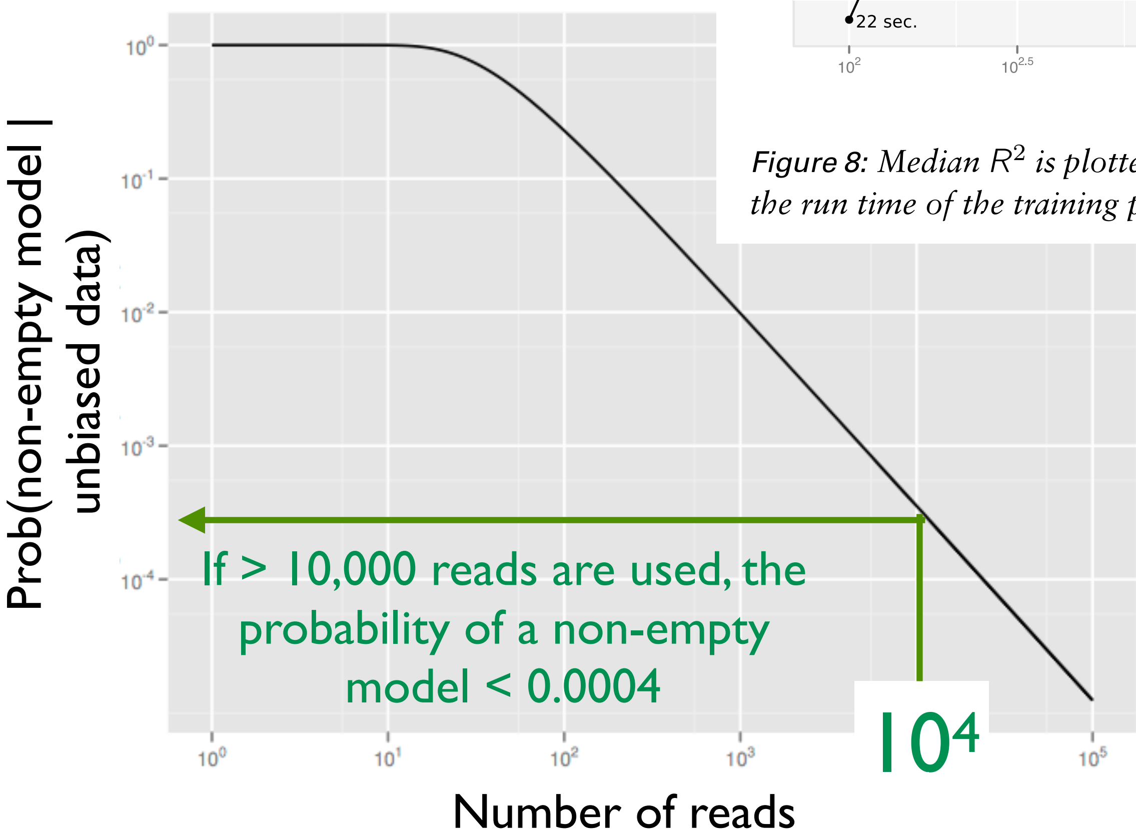


Figure 8: Median R^2 is plotted against training set size. Each point is additionally labeled with the run time of the training procedure.



… while accuracy and runtime rise (empirically)

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:

If bias changes coverage, it changes power to detect differential expression

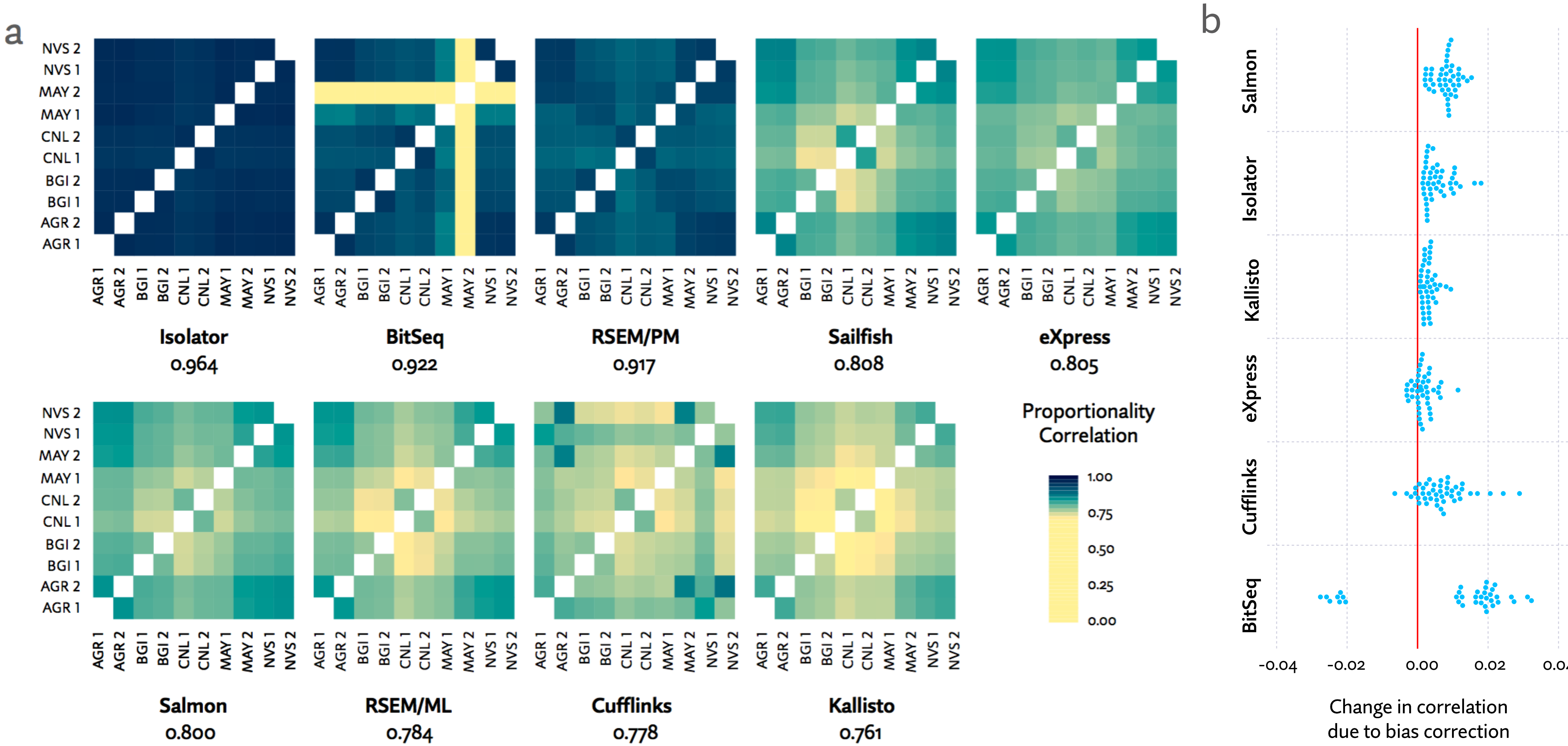
SNPs and/or alternative splicing might have a big effect, if samples are genetically different and/or engender changes in isoform usage

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

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

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

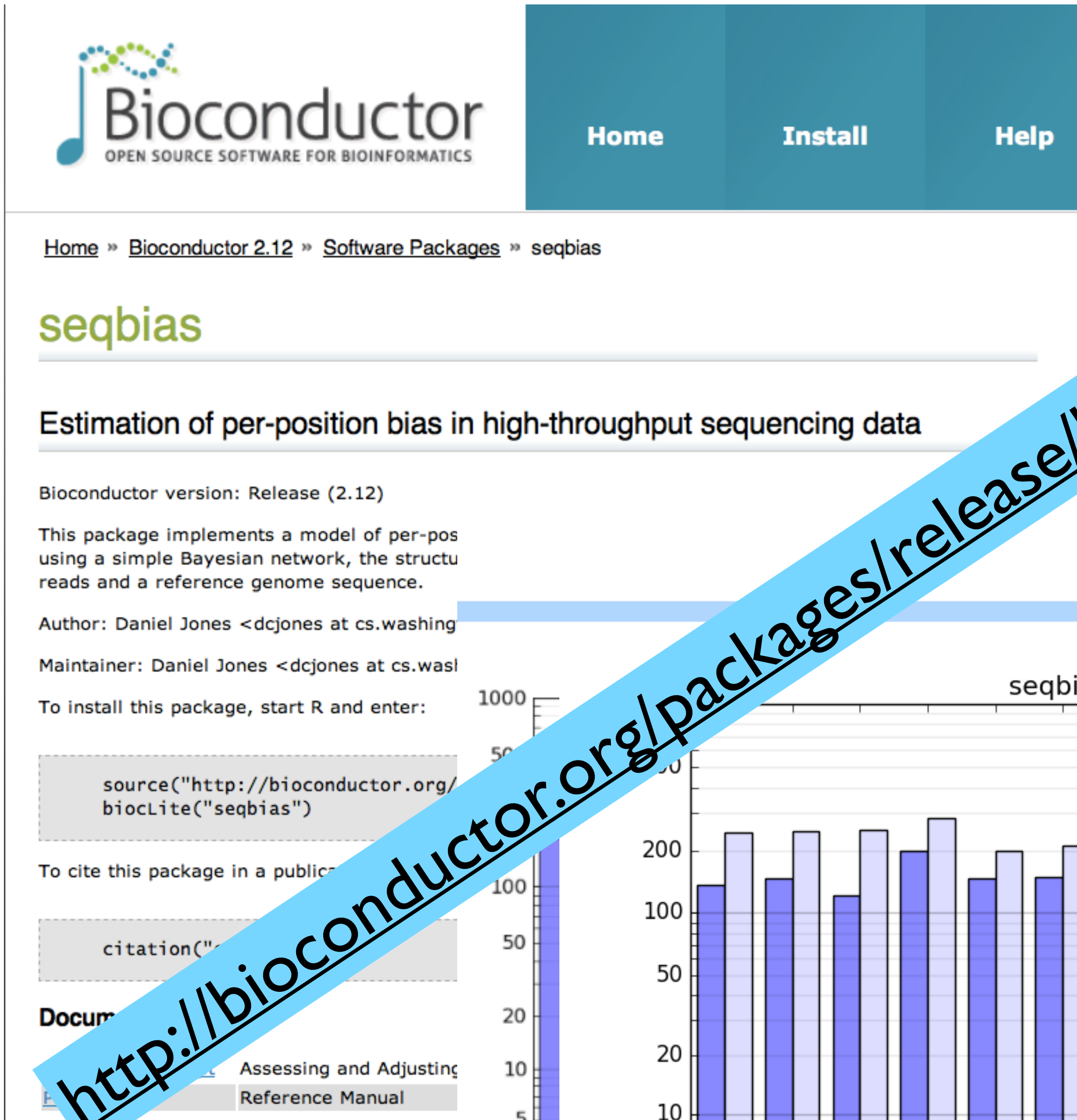
Batch Effects? YES!



A: Pairwise proportionality correlation between *technical* replicates; 1 lane of 2 flowcells each at 5 sites, all HiSeq 2000. B: The absolute change in correlation induced by enabling bias correction (where available).

For clarity, BitSeq est. of "MAY 2", excluded; bias correction was extremely detrimental there.

Availability

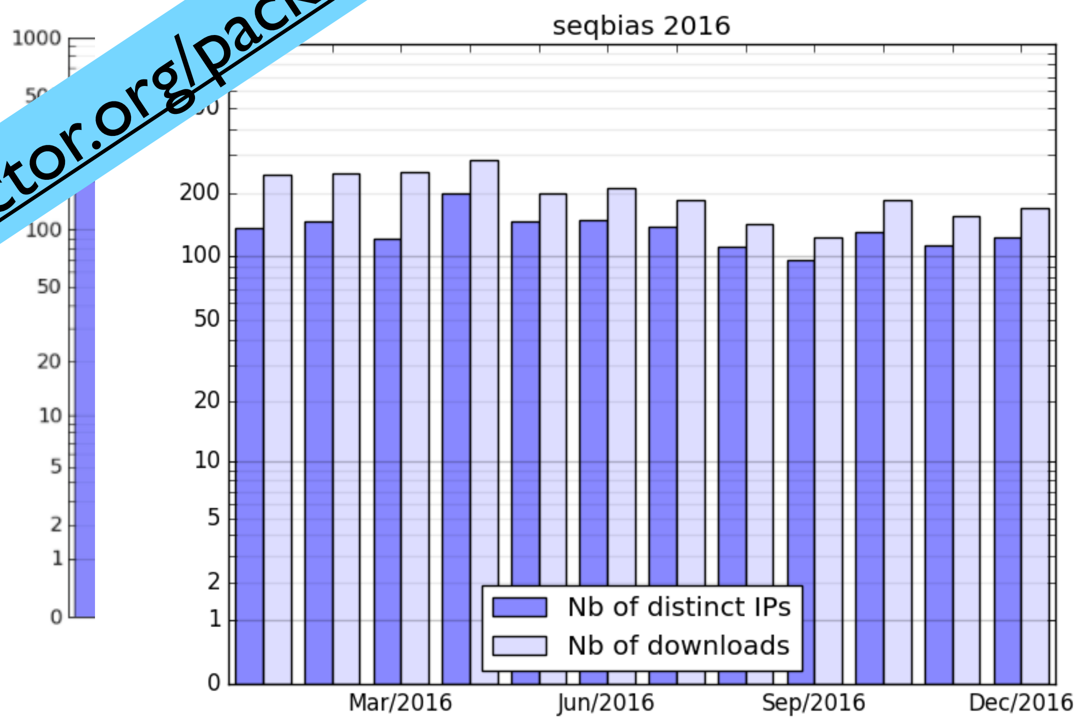


http://bioconductor.org/packages/release/bioc/html/seqbias.html

Download stats for Software package seqbias

This page was generated on 2015-06-01 06:29:02 -0700 (Mon, 01 Jun 2015).

seqbias home page: release version, devel version.



2015

...t IPs	Nb of downloads
181	252
236	360
242	360
197	292
217	299
186	311
195	371
138	270
211	327
170	264
153	220
0	0
1648	3326

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

BUT careful modeling can help.

Acknowledgements

Daniel Jones

Katze Lab

Michael Katze
Xinxia Peng

Stem Cell Labs

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

Funding

NIGMS, NHGR, NIAID

Exciting Times

"Biology is to 21st Century as Physics was to 20th"

Lots to do

Highly multidisciplinary

You'll be hearing a lot more about it

I hope I've given you a taste of it

Thanks!

PS: Please complete online course evaluation by Sunday

https://uw.iasystem.org/survey/188811