CSE 428

Spring 2018

Overview

Course Web Pages:

https://courses.cs.washington.edu/courses/cse428/18sp/

TAs:

Daniel Jones

Yue Zhang

Group-Project-oriented:

Typically teams of ~4 students

I will offer some projects ideas

I am open to student-generated ideas

"computers" + "biology"

(+ reasonable scope + something I can facilitate)

Project Challenges

Organization & Scheduling

Bio Jargon

Tools from elsewhere

Did I mention Organization & Scheduling?

What I hope you will learn

See previous slide!

You'll see real DNA/RNA seq data in all of them, plus

Some mixture of:

- data structures,
- algorithms,
- data analytics,
- statistics,
- biology,
- HCI,
- ML, …

Project Evaluation

Weekly Goals + Progress reports

Final written reports + oral presentations

  Including evaluation of code, test results, etc.

Peer comments

Project Ideas

3 of my 4 suggestions grow out of "bias" in RNA sequencing, outlined in the following ~2 dozen slides. For today, at least, the details are not critical; key points I hope you get are that
a) we can sequence RNA from cells
b) it's informative
c) it's quantitative
d) technical artifacts bias that quantitative information
e) we have software that ameliorates this bias, and

f) *there are unexplored issues surrounding this,* hence, project ideas: visualizing and understanding the sources and extent of the biases and their impact on various downstream analyses.

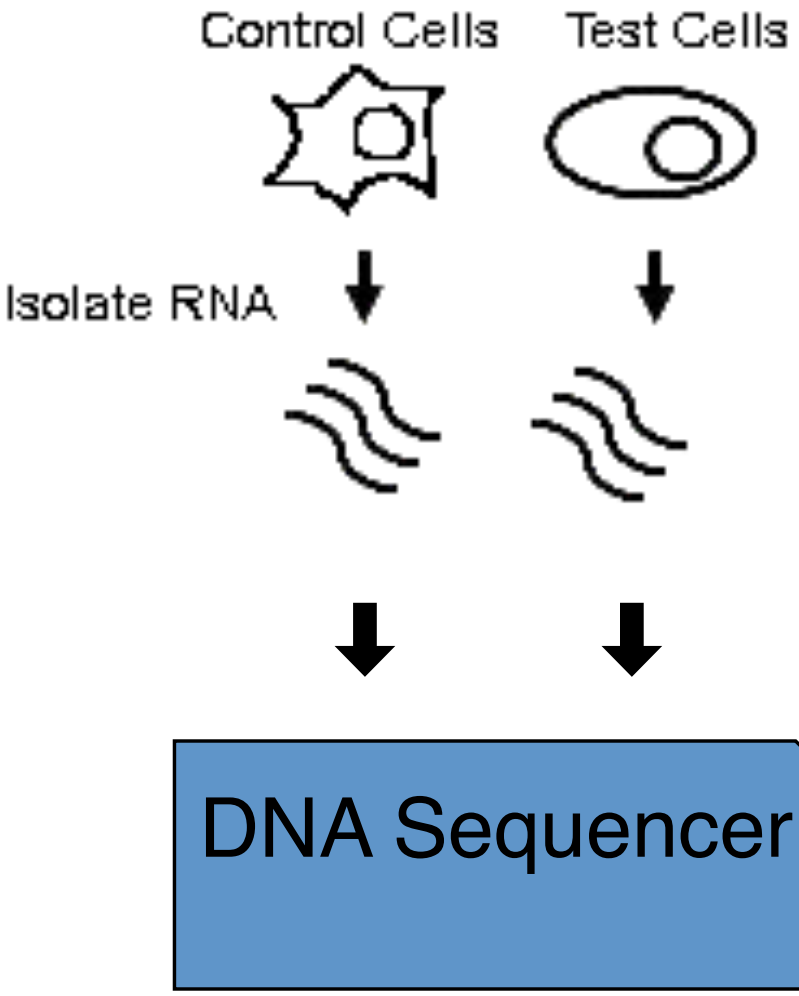
Bias in RNA sequencing and what to do about it

Walter L. (Larry) Ruzzo

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

ruzzo@uw.edu

RNAseq



DNA Sequencer

Millions of reads,
say, 100 bp each

map to genome,
compare & analyze

Goals of RNAseq

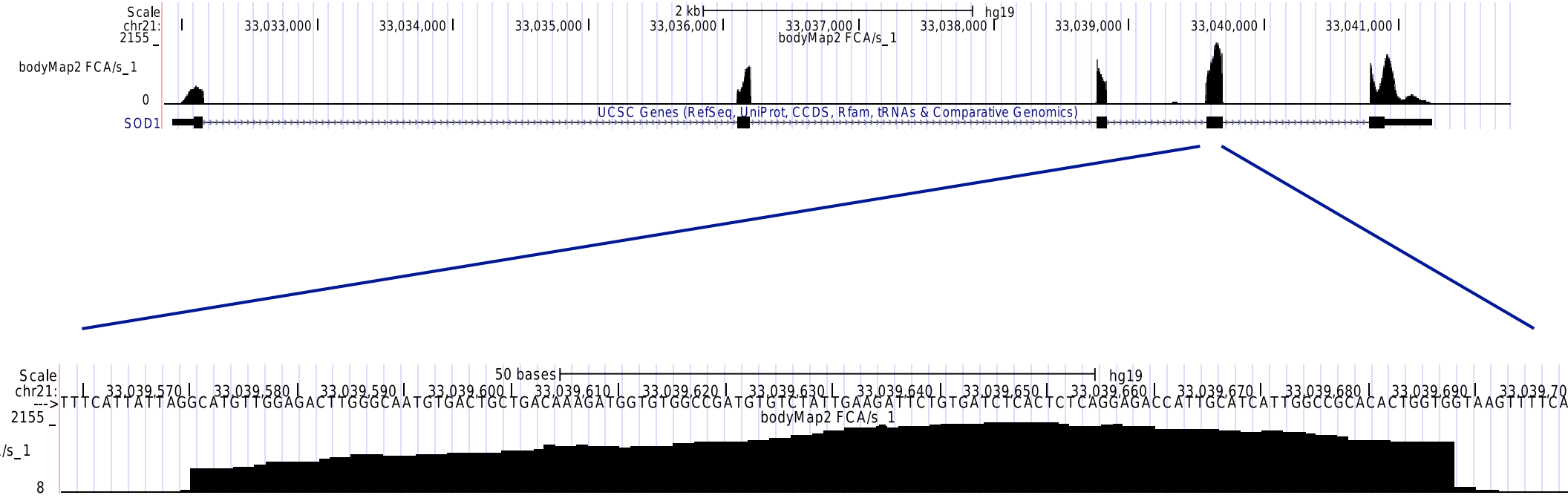
1. Which genes are being expressed?
   How? *assemble* reads (fragments of mRNAs) into (nearly) full-length mRNAs and/or *map* them to a reference genome

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. ...

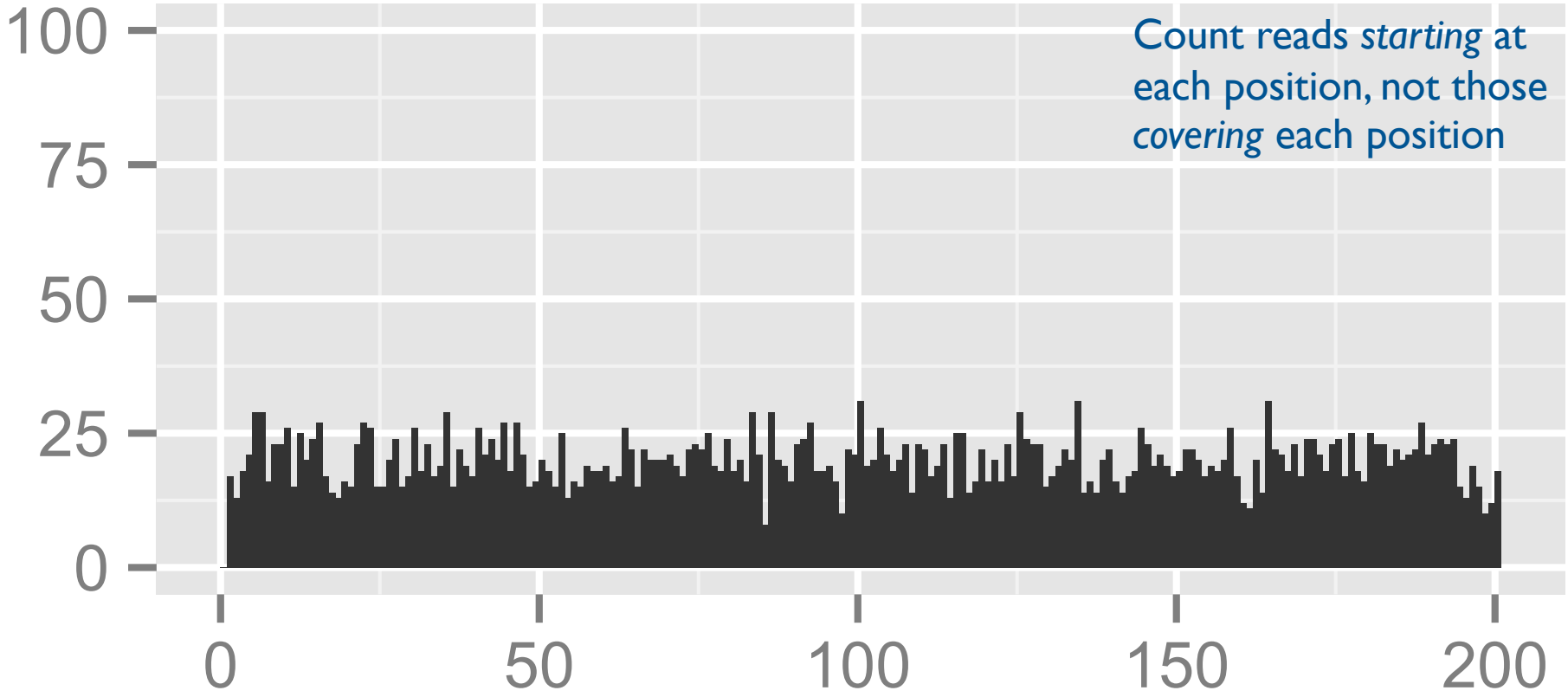
RNA seq



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



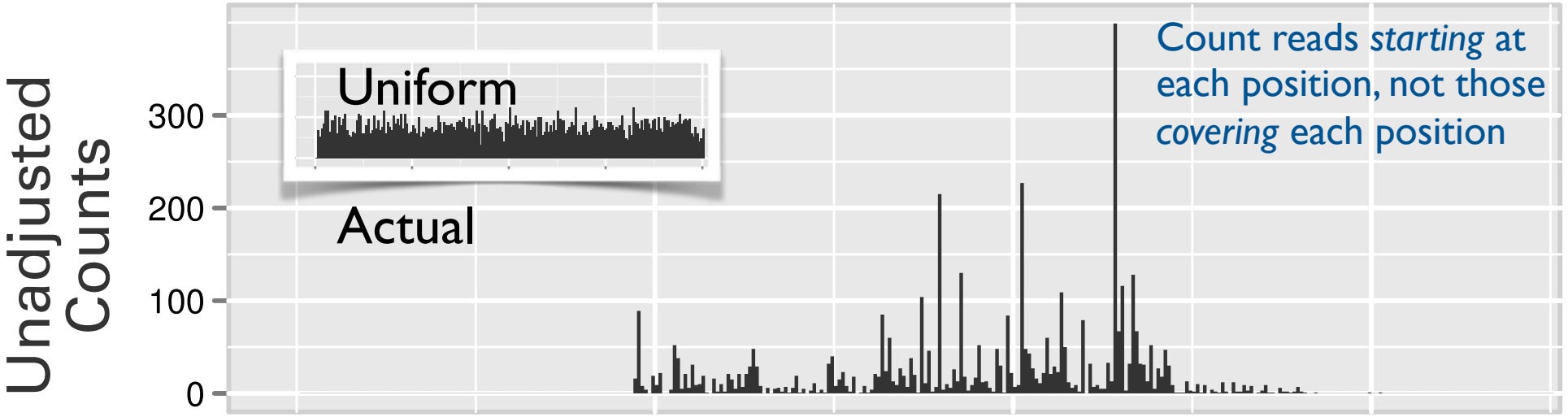
Uniform sampling of 4000 "reads" across a 200 bp "exon."

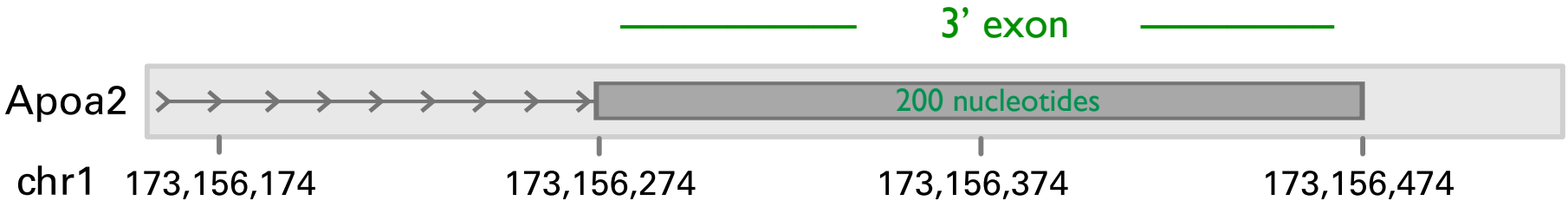
Average 20 ± 4.7 per position, min ≈ 9, max ≈33

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 ≳ +10σ above mean

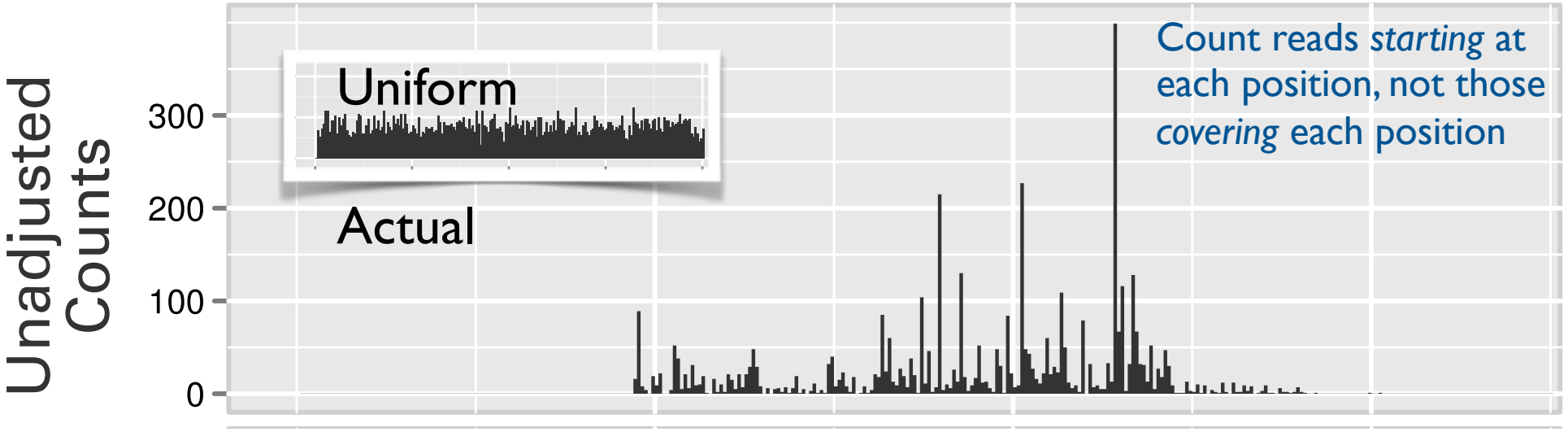




Mortazavi data

What we get: *highly* non-uniform coverage

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



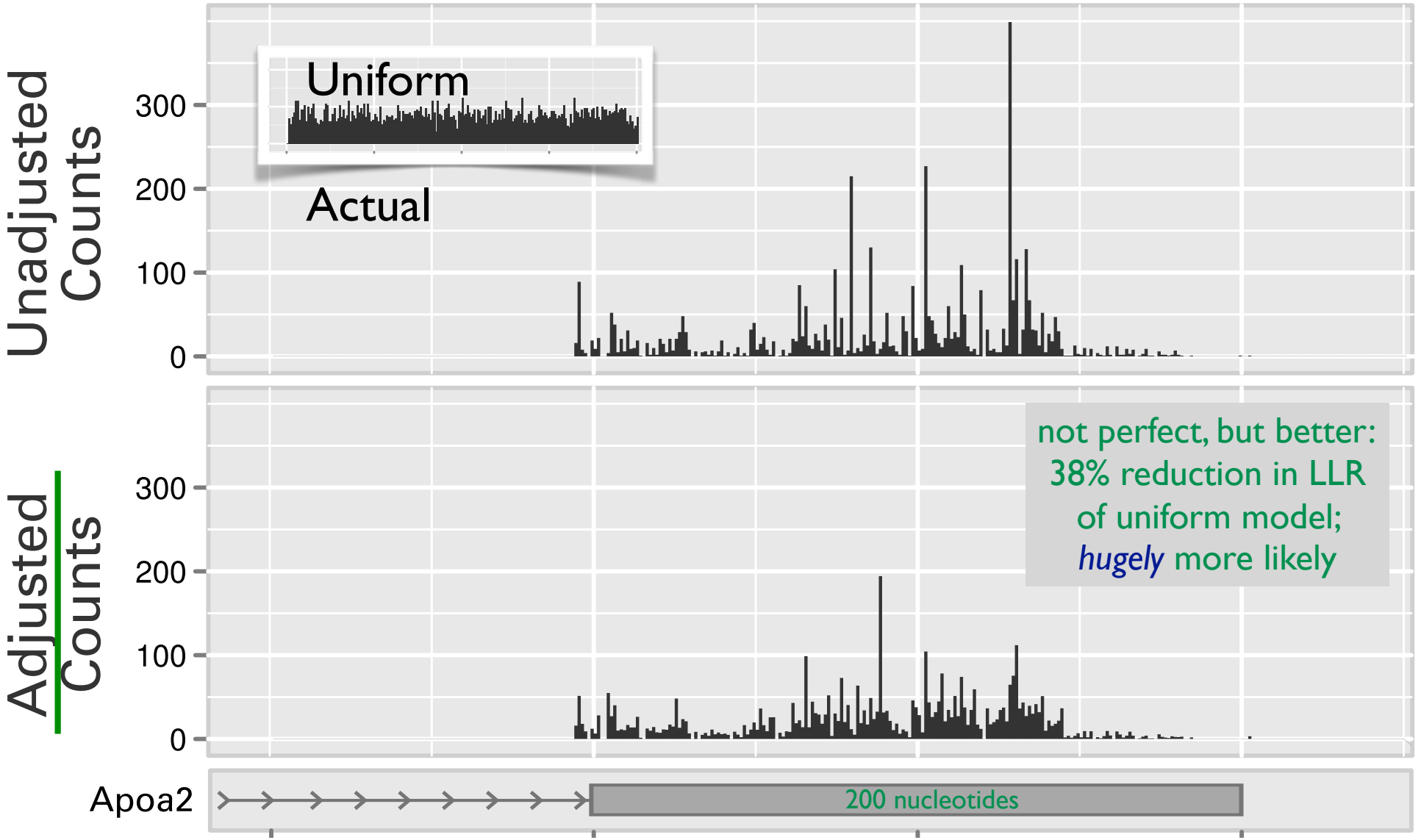
How to make it more uniform?

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

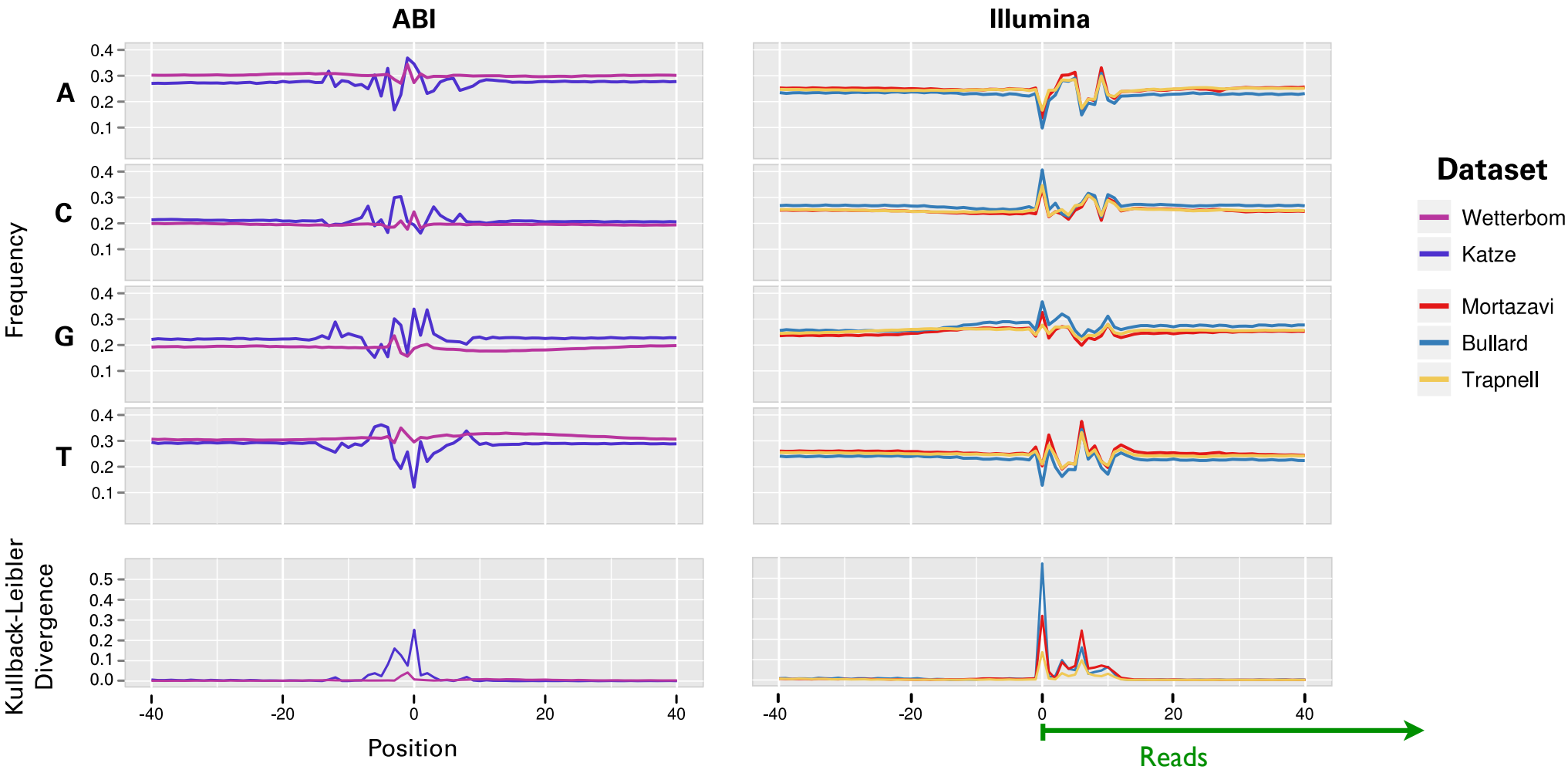
B: Try to model (aspects of) causation ← *WE DO THIS*

(& 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

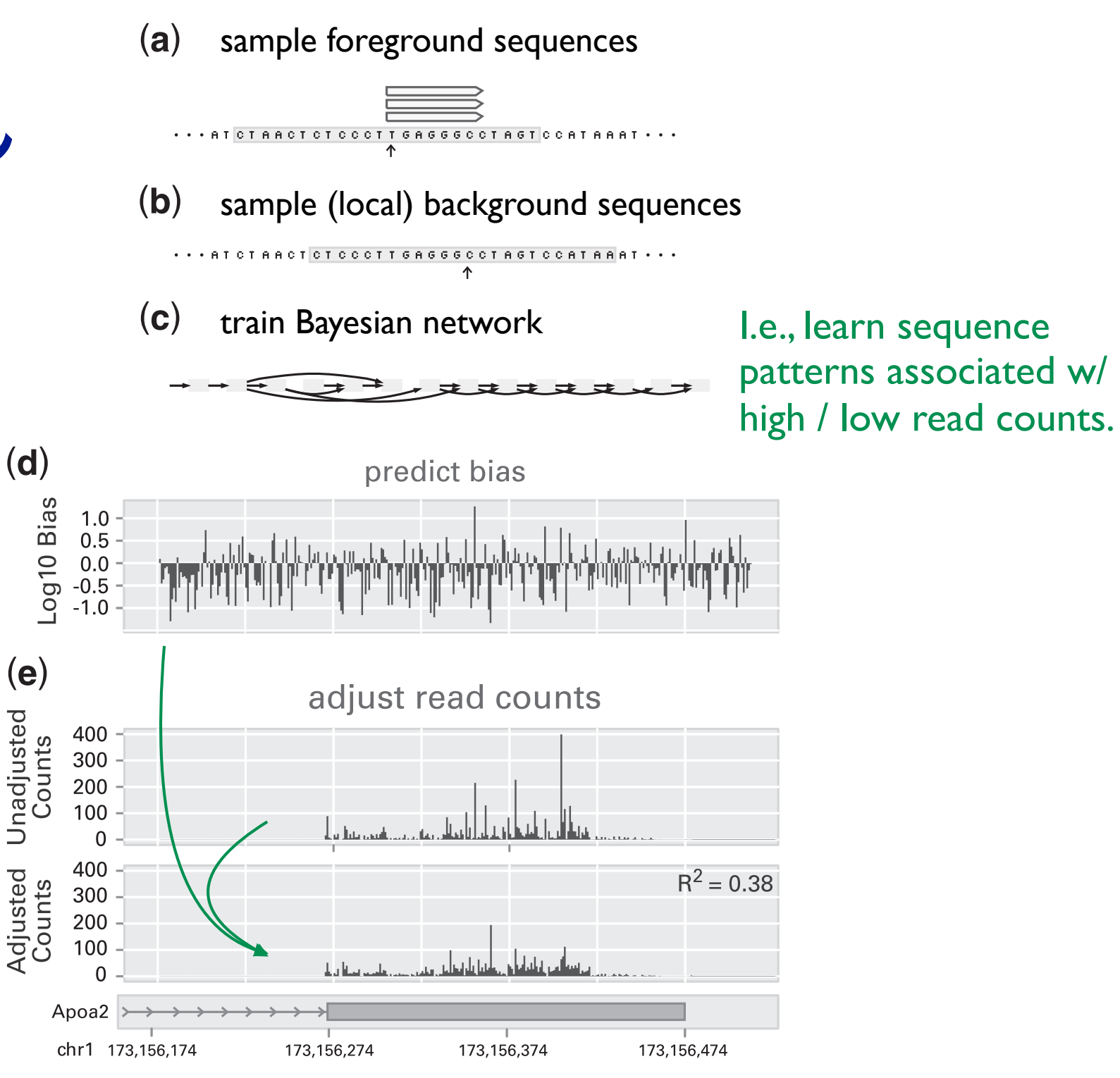
Li, et al. 2010

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:

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

From Bayes rule:

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

We define "bias" to be this factor

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

Some obvious choices:

Full joint distribution: 4^k-1 parameters

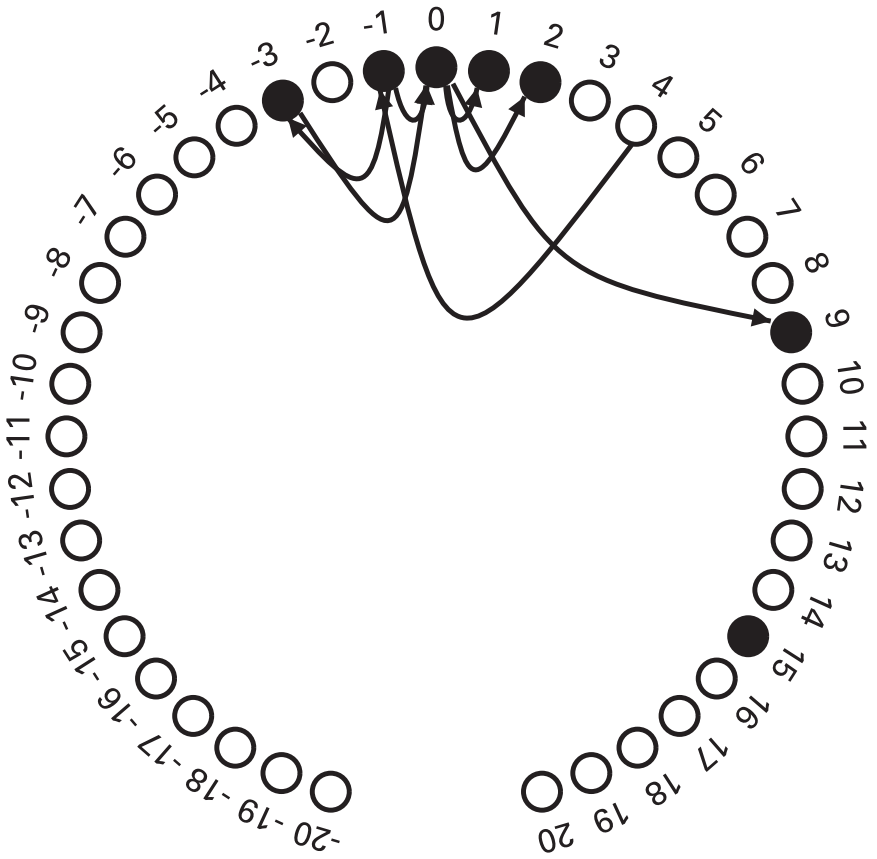
PWM (0-th order Markov): (4-1)•k parameters

Something intermediate:

Directed Bayes network

Form of the models:

Directed Bayes nets



Wetterbom (282 parameters)

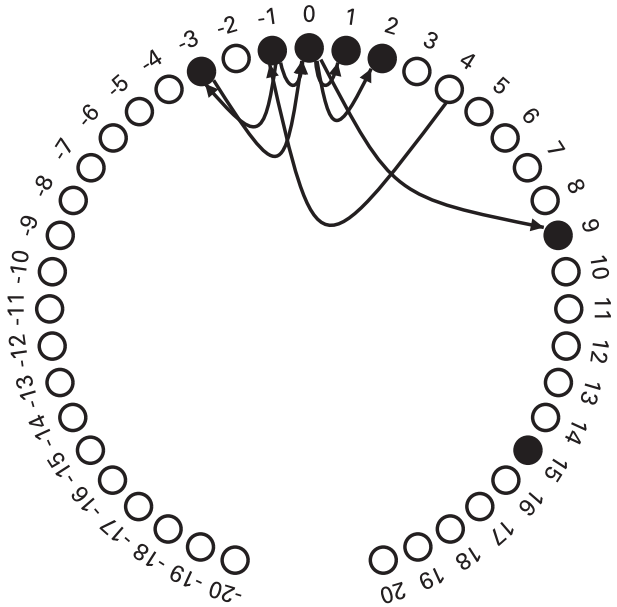
One "node" per nucleotide, ±20 bp of read start

- 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

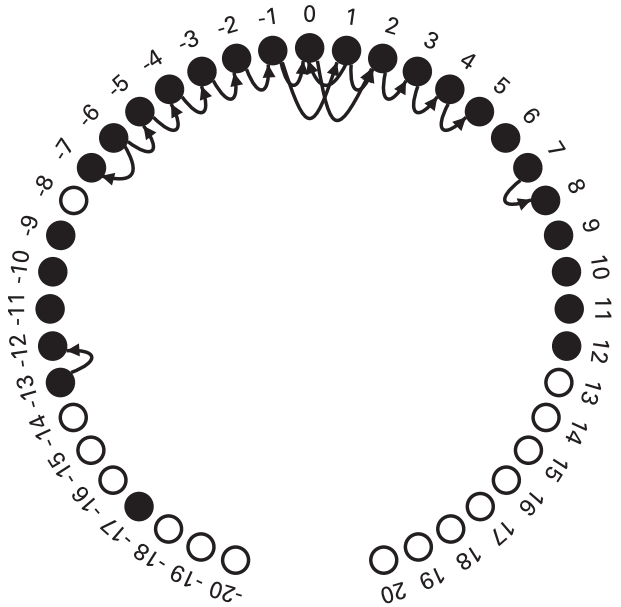
How–optimize:

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

ABI



Wetterbom
(282 parameters)

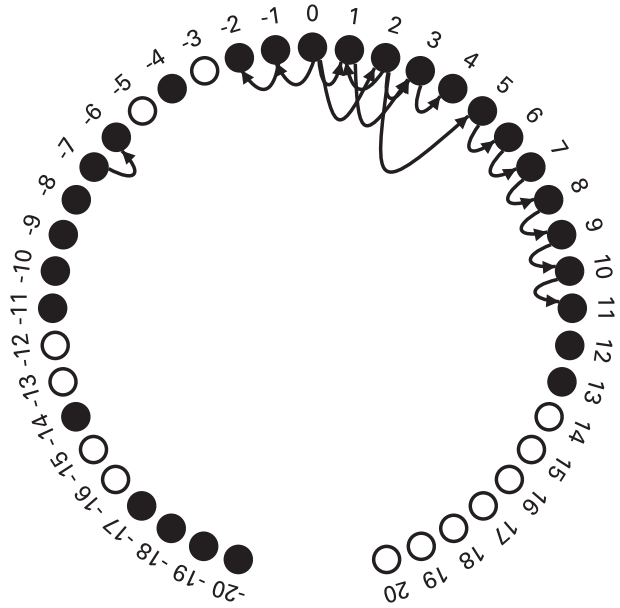


Katze
(684 parameters)

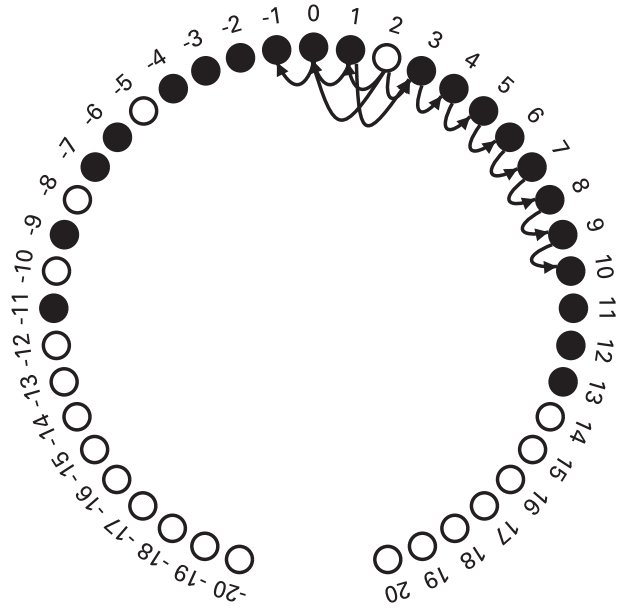
NB:

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

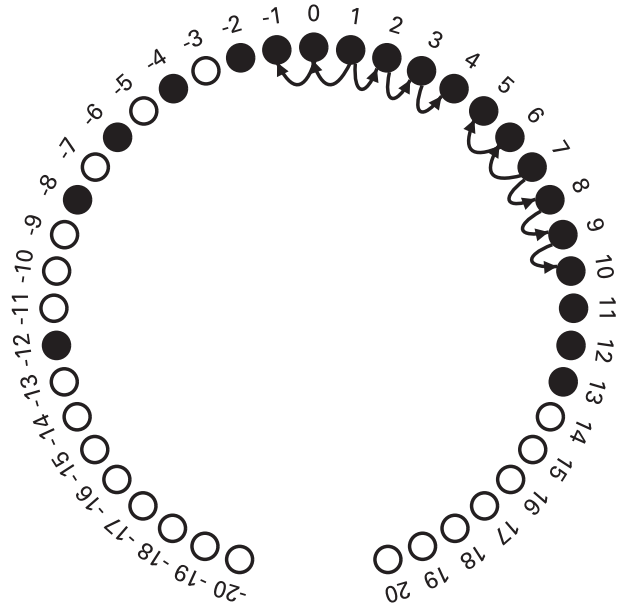
Illumina



Bullard
(696 parameters)

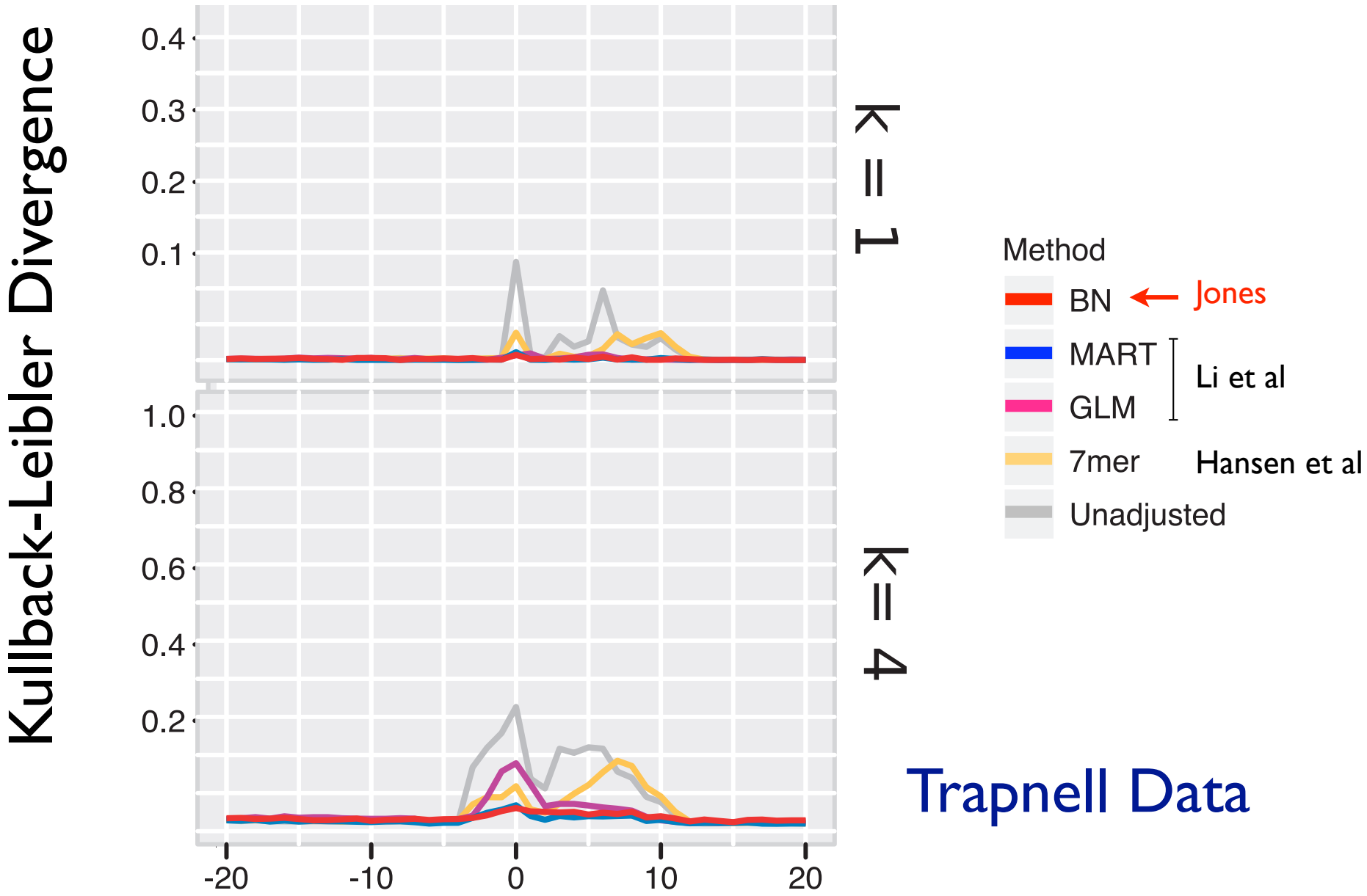


Mortazavi
(582 parameters)



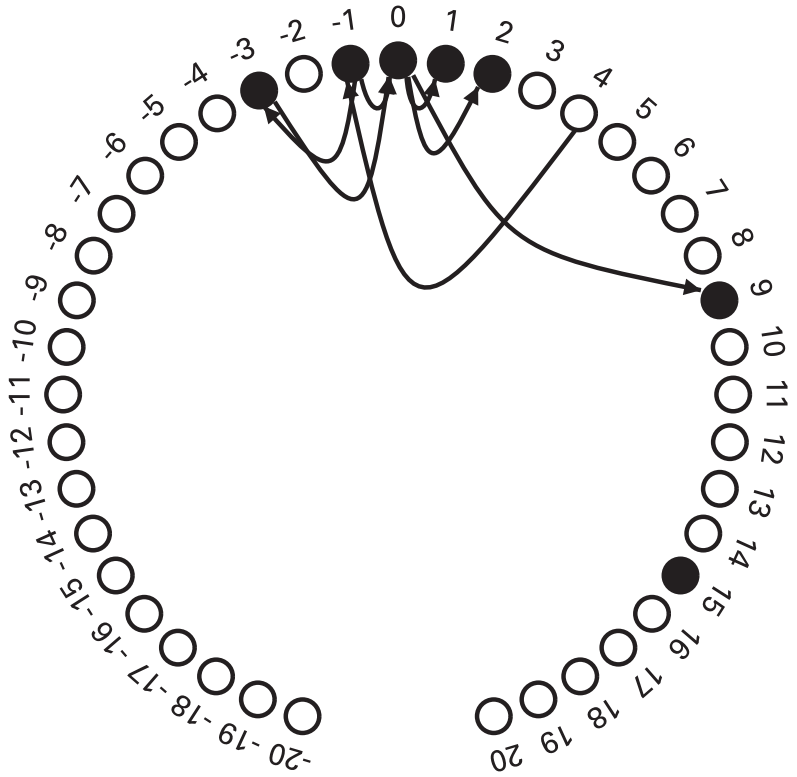
Trapnell
(360 parameters)

Result – Increased Uniformity

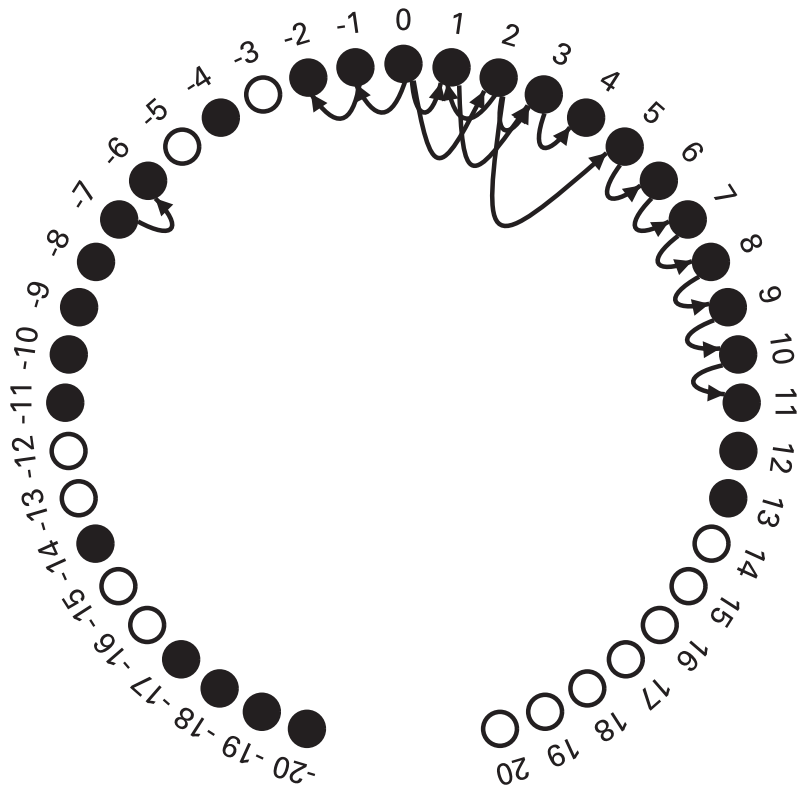


Trapnell Data

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) ...

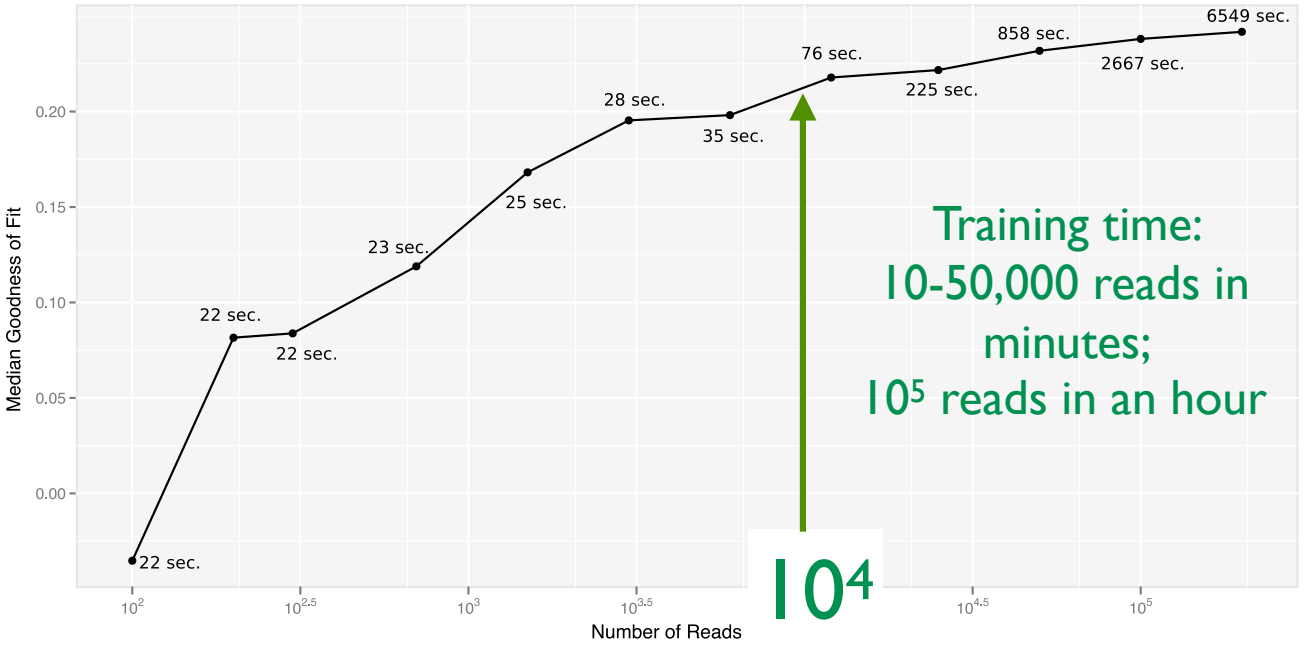
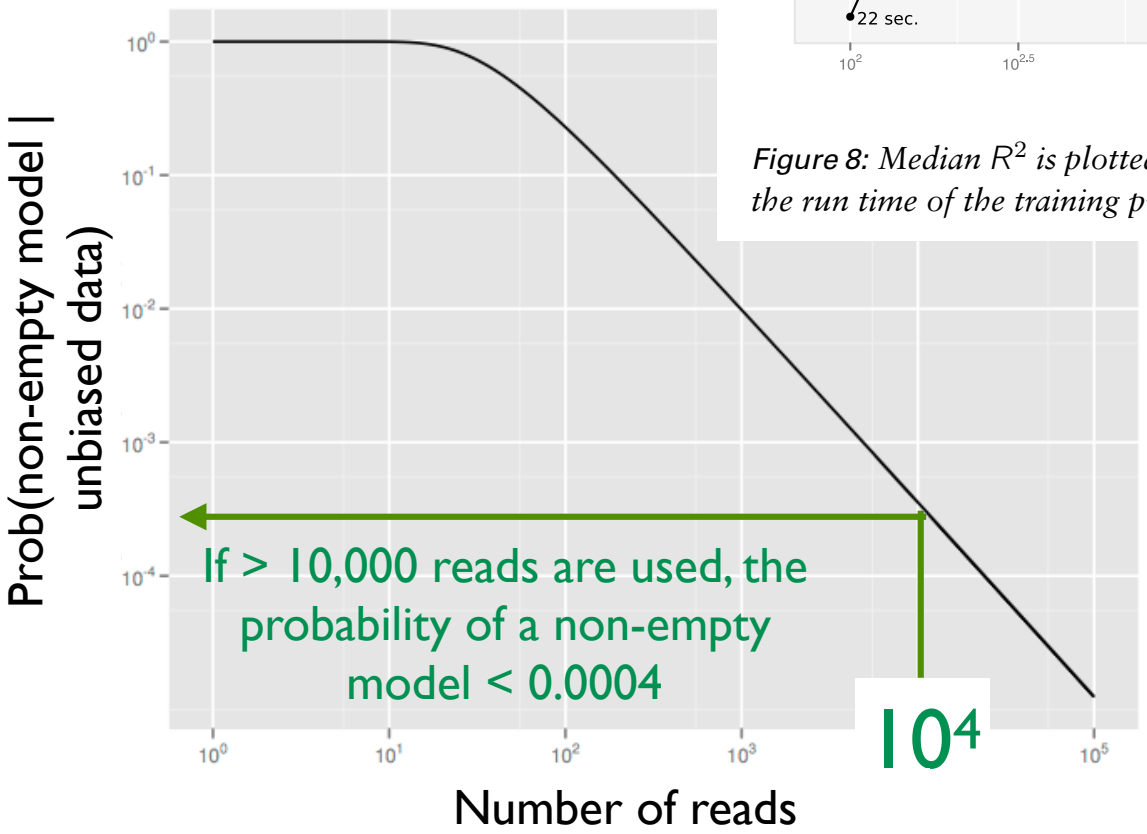


Figure 8: Median R^2 is plotted against training set size. Each point is additionally labeled with the run time of the training procedure.



… 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

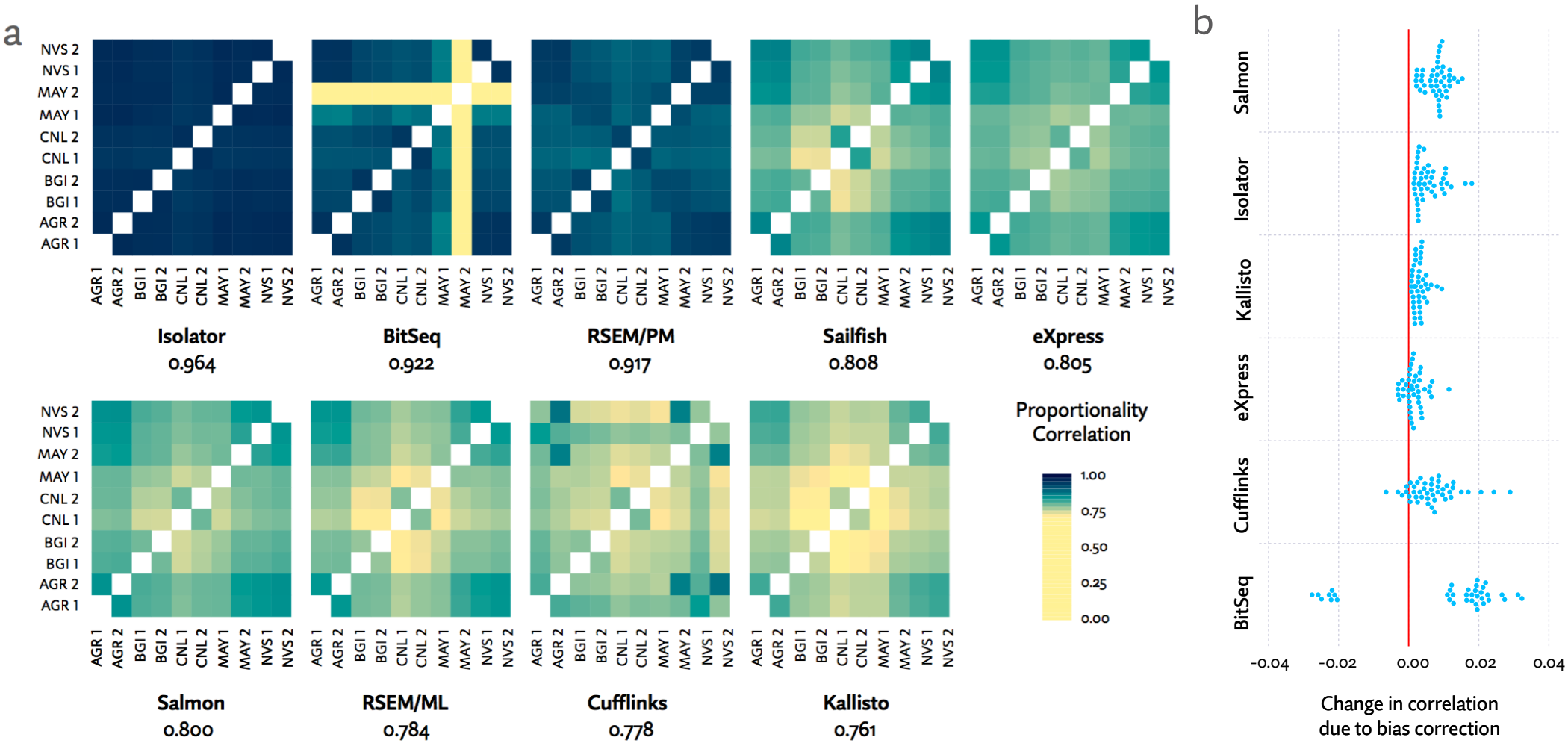
SNPs and/or alternative splicing might have a big effect, if samples are genetically different and/or engender changes in isoform usage

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.

summary

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

Announcements

Possible room change next week - watch email

Project Ideas: Next Few Slides

All are open-ended, underspecified; as you think about them, both let your imagination run free, *and* think carefully about how to scale and stage your project so you can collect low-hanging fruit before potentially getting lost in the open-ended weeds. (Fortunately, I don't think mixing metaphors is a crime in this state–at least not yet.)

Idea #1: Visualizing and Exploring SeqBias

It's hard to think about it if you can't visualize it.

Goal: Develop a tool to automatically measure, quantify, and display summaries of bias in specific RNAseq data sets, and apply this too a variety of them.

Motivating Questions: How does bias vary from one data set to another? Is more modern data less biased? How does it impact down-stream analyses?

Some Suggested Steps:

- Learn state-of-the-art in RNAseq Quality Control
- Add SeqBias, starting with figures like those in Daniel's paper
- Other metrics?
- Apply to a variety of data?
- HCI issues in presenting such data to potential users?
- Very Speculative: can we implicate *cause*s of bias?

Idea #2: Bias Distorts Allele Specific Expression Analysis?

Background: An *allele* is one variant of a gene, e.g., the A/B/O alleles that determine "Blood Type." You have 2 alleles of every gene (partially excluding those on X,Y chromosomes). E.g., if you got A from mom & B from dad, you have AB blood-type; if you have O from both, you have O blood-type.

Usually, both alleles are "expressed", i.e., made into proteins, as in the case above, but there are exceptions where only one of the two alleles is expressed ("allele specific expression" or ASE, with dozens of examples known in humans), and potentially severe consequences for disrupting this (e.g., see "Prader-Willi/Angelman syndromes").

How do you detect ASE? One way: compare DNAseq to RNAseq in an individual; if DNA shows 2 alleles, but RNA only sees one of them (or much more of one than the other), then you call it ASE.

Idea #2: Bias Distorts Allele Specific Expression Analysis?

Alleles differ in a small number of positions; bias is sensitive to sequence; so a change in bias at a few changed positions might falsely appear to be ASE, or falsely mask true ASE.

Goal: Explore the effect of SeqBias on ASE prediction. If deemed significant, develop a tool to automatically "correct" for it and apply this too a variety of data sets.

Motivating Questions: Does bias compromise our ability to detect ASE from RNAseq data? What can we do about it?

Some Suggested Steps:

- Learn state-of-the-art in ASE discovery
- Add SeqBias correction to that pipeline
- Assess whether it makes a difference
- Apply to a variety of data?

Idea #3: Impact of bias in other RNAseq use cases

Other RNAseq applications may be even more susceptible to distortion due to seqbias, e.g. ribosome foot-printing and RNA structure prediction (SHAPE).

Goal: Explore the effect of SeqBias on these tasks. If deemed significant, develop a tool to automatically "correct" for it and apply this too a variety of data sets.

Motivating Questions: Does bias compromise accuracy of our predictions from RNAseq data? What can we do about it?

Some Suggested Steps:

- Learn state-of-the-art in these applications
- Add SeqBias correction to that pipeline; a key is defining an appropriate "background"
- Assess whether it makes a difference
- Apply to a variety of data?

Idea #4: Improved crossover detection–Background

Jargon: A position in your genome where your mom's nucleotide agrees with your dad's is called *homozygous* (~99.9%); places where they disagree are *heterozygous* (the other .1%).

How might you find heterozygous sites? Perhaps DNAseq will give you "coverage" ~100 at a site, with, say 60 A's and 40 G's:

...GATAGCGATATGGAGTAGAATACCAGGAGCATTTGACCATACTAC...

Idea #4: Improved crossover detection–Background

The *phasing* problem: Given a pair of nearby heterozygous sites, say A/G at position i and G/T at position j > i, does the G at pos j appear on the same chromosome as the A at i or the G at i? I.e., do we have this:

```
          i           j
- - - A - - - - G - - -
- - - G - - - - T - - -
```

or this:

```
- - - A - - - - T - - -
- - - G - - - - G - - -
```

?

How could we tell? Again, maybe DNAseq: If there are single reads covering both pos i and pos j, do they show a mixture of `A--G` with `G--T` or a mixture of `A--T` with `G--G`?

Potential confusion to avoid: each cartoon shows one strand on each of the 2 chromosomes, not "base pairs" on one chromosome (A:T and G:C base pairs.)

Idea #4: Improved crossover detection–Background

The *crossover* problem: Given the same setup, but looking at *two individuals*, perhaps siblings, if we see this in one:

```
          i           j
- - - A - - - - G - - -
- - - G - - - - T - - -
```

and this in the other:

```
- - - A - - - - T - - -
- - - G - - - - G - - -
```

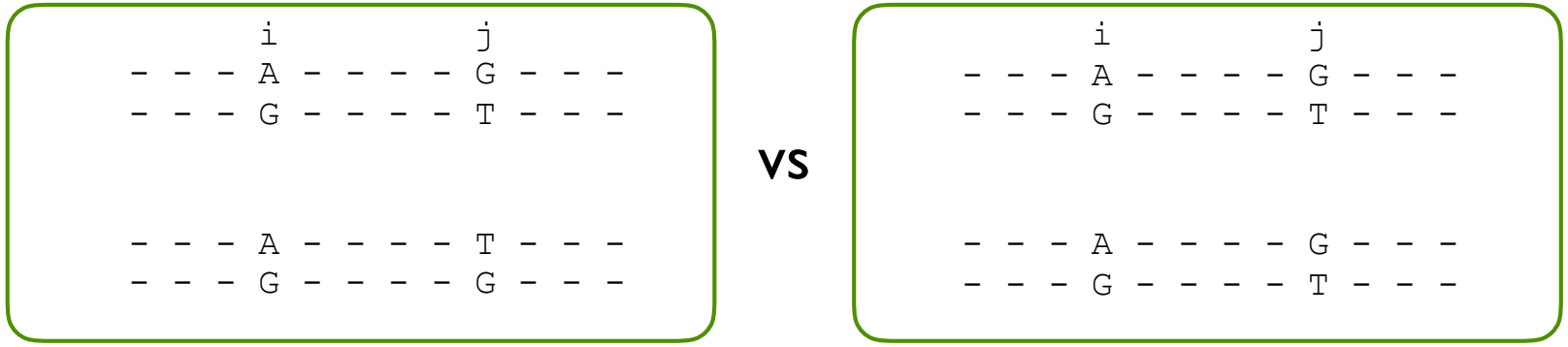
how could that be?

One likely answer: crossover/recombination (in meiosis)

Another possibility: a phasing error!

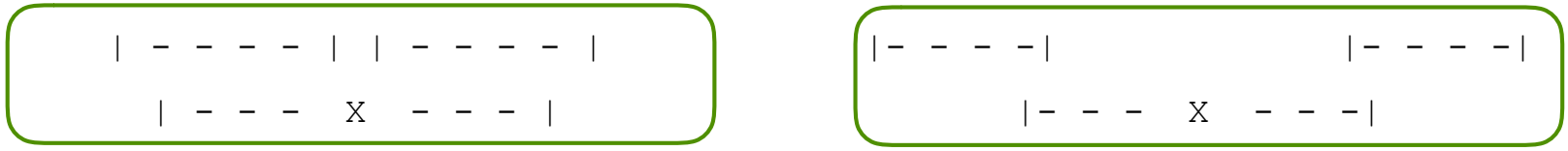
Idea #4: Improved crossover detection–Background

Is crossover distinguishable from a phasing error? Probably not in isolation, but what if we have several *overlapping* i-j pairs that are phased in both individuals? Then we can try for a probabilistic assessment. E.g., abstracting:



as :

What does this suggest?:



(If top gap is short vs long, error in "X" is more/less likely)

Idea #4: Improved crossover detection

Data from a pair of closely related individuals, after being (separately) phased, may/will show crossovers. Are they real/ how many of them are real?

Goal: build a tool to find maximum likelihood estimate of # crossovers, based simple models of xover/error.

Motivating Questions: Can we do better than blindly trusting the phasing results.

Some Suggested Steps:

Learn state-of-the-art in these applications

Model as max likelihood solution to system of linear eqns.

$x_1+x_2+e_1 \equiv 0 \pmod 2$
$x_4+x_5+e_2 \equiv 0 \pmod 2$
$x_2+x_3+x_4+e_3 \equiv 1 \pmod 2$

Good Alg? NP-hard? Good heuristics? Decomposes?

Apply to a variety of data (especially mine; phasing on up)?

Next steps

review slides
which appeals?
form groups
skim references on web
talk to/email me/TAs
we have fragments of code for parts of this
(may or may not be useful…)