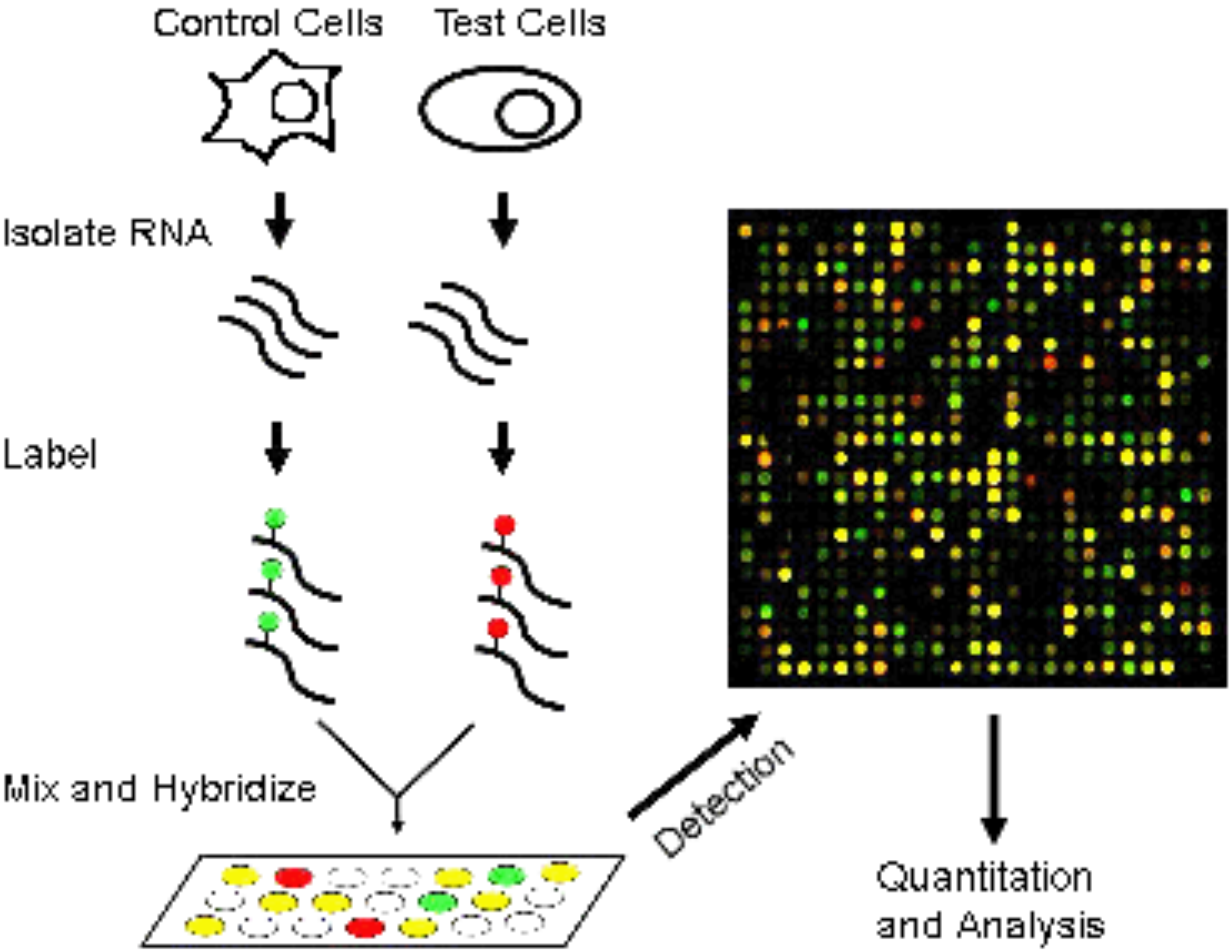
CSEP 590 B
Computational Biology

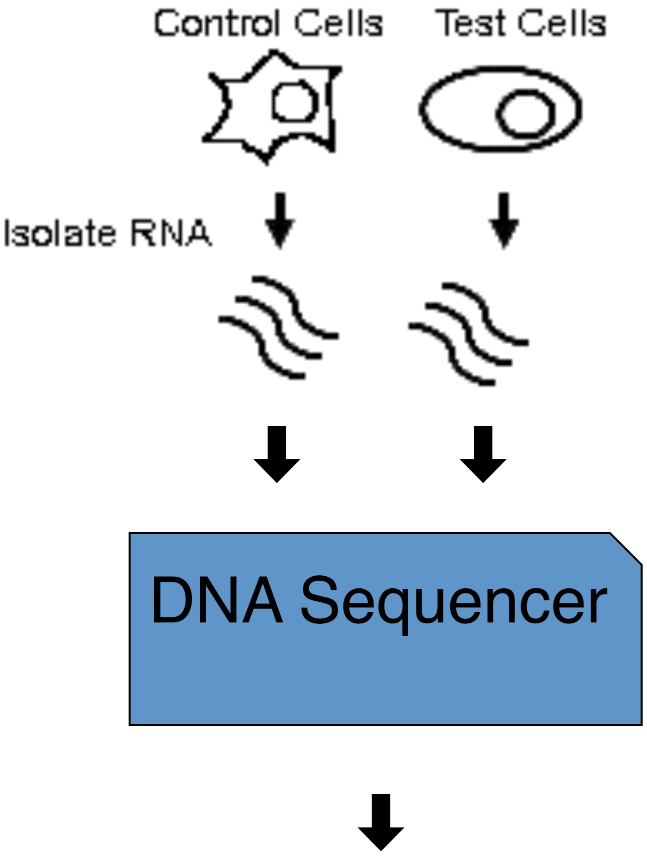
Gene Expression Analysis

Assaying Gene Expression

Microarrays



RNAseq



DNA Sequencer

map to genome, 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?

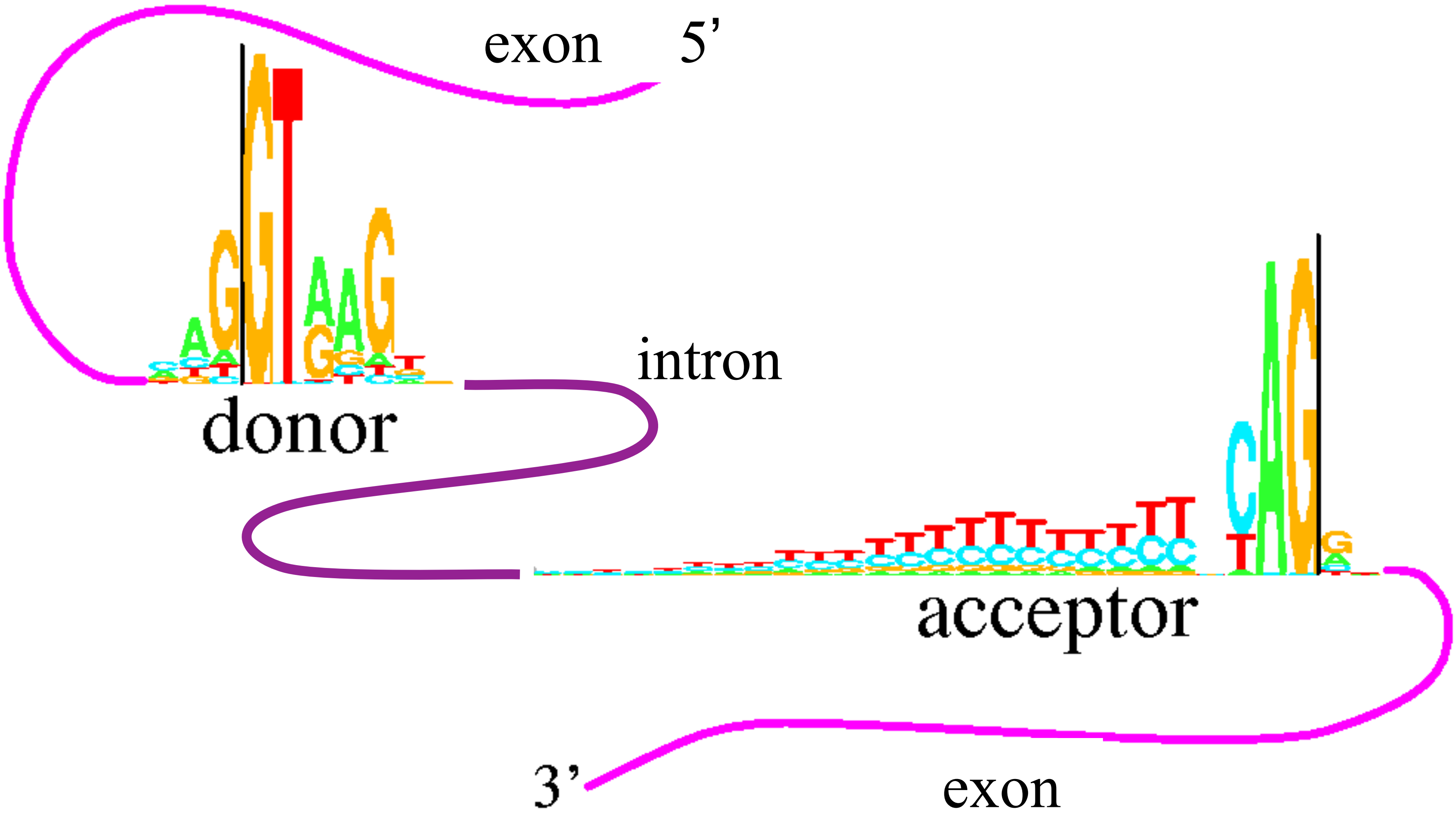
How? *count* how many fragments come from each gene–expect more highly expressed genes to yield more reads, after correcting for biases like mRNA length

#3: What’s same/diff between 2 samples

E.g., tumor/normal

#4: ...

Recall: splicing



RNAseq Data Analysis

De novo Assembly

- mostly deBruijn-based, but likely to change with longer reads
- more complex than genome assembly due to alt splicing, wide diffs in expression levels; e.g. often multiple “k’s” used
- pro: no ref needed (non-model orgs), novel discoveries possible, e.g. very short exons
- con: less sensitive to weakly-expressed genes

Reference-based (more later)

- pro/con: basically the reverse

Both: subsequent bias correction, quantitation, differential expression calls, fusion detection, etc.

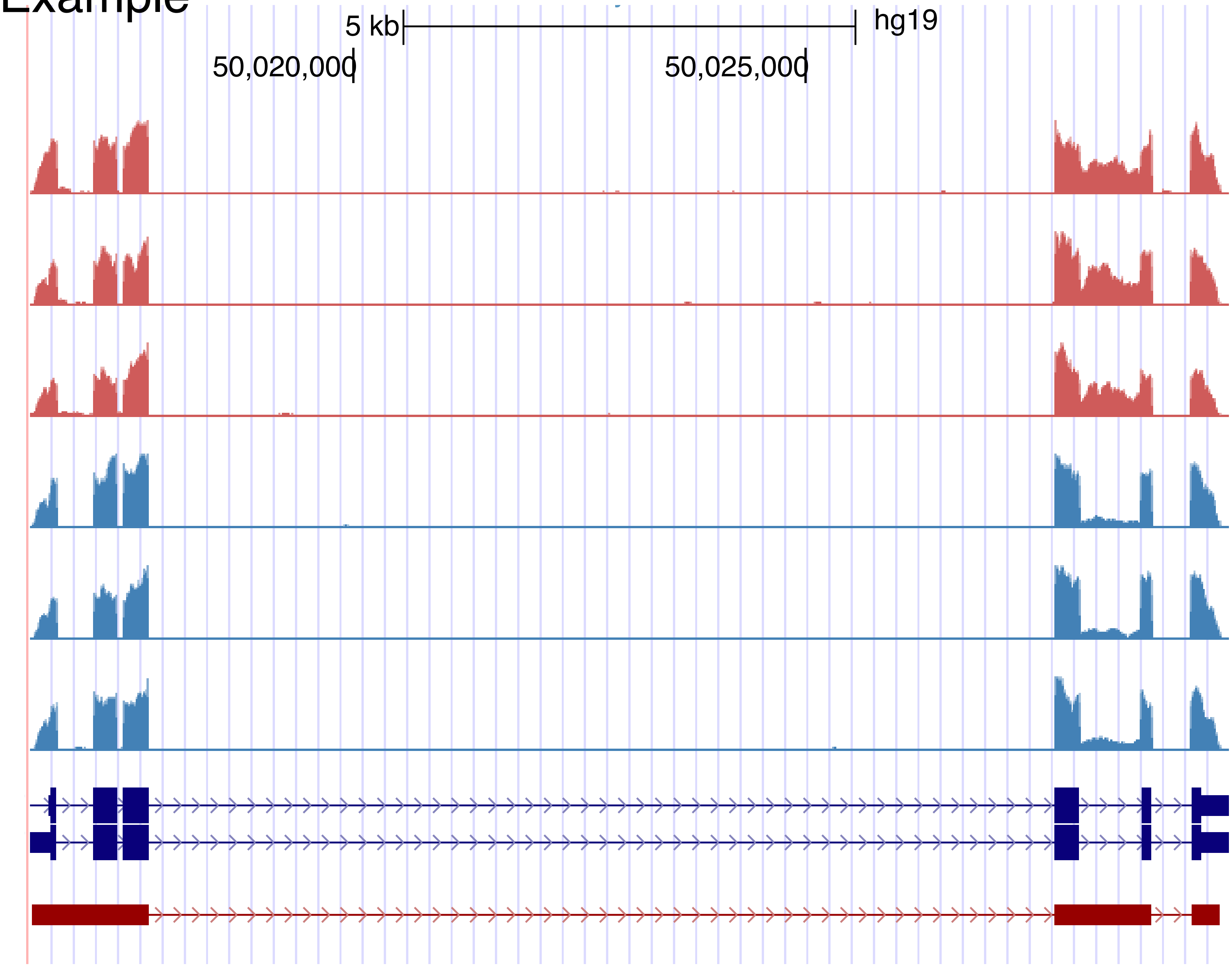
"TopHat" (Ref based example)

BWA

- map reads to ref transcriptome (optional)
- map reads to ref genome
- unmapped reads remapped as 25mers
- novel splices = 25mers anchored 2 sides
- stitch original reads across these
- remap reads with minimal overlaps

- *Roughly:* 10m reads/hr, 4Gbytes
  (typical data set 100m–1b reads)

RNAseq Example



Day 20

1 Year

RNAseq protocol (approx)

Extract RNA (maybe by polyA ↔ polyT)

Reverse-transcribe into DNA (“cDNA”)

Make double-stranded, maybe amplify

Cut into, say, ~300bp fragments

Add adaptors to each end

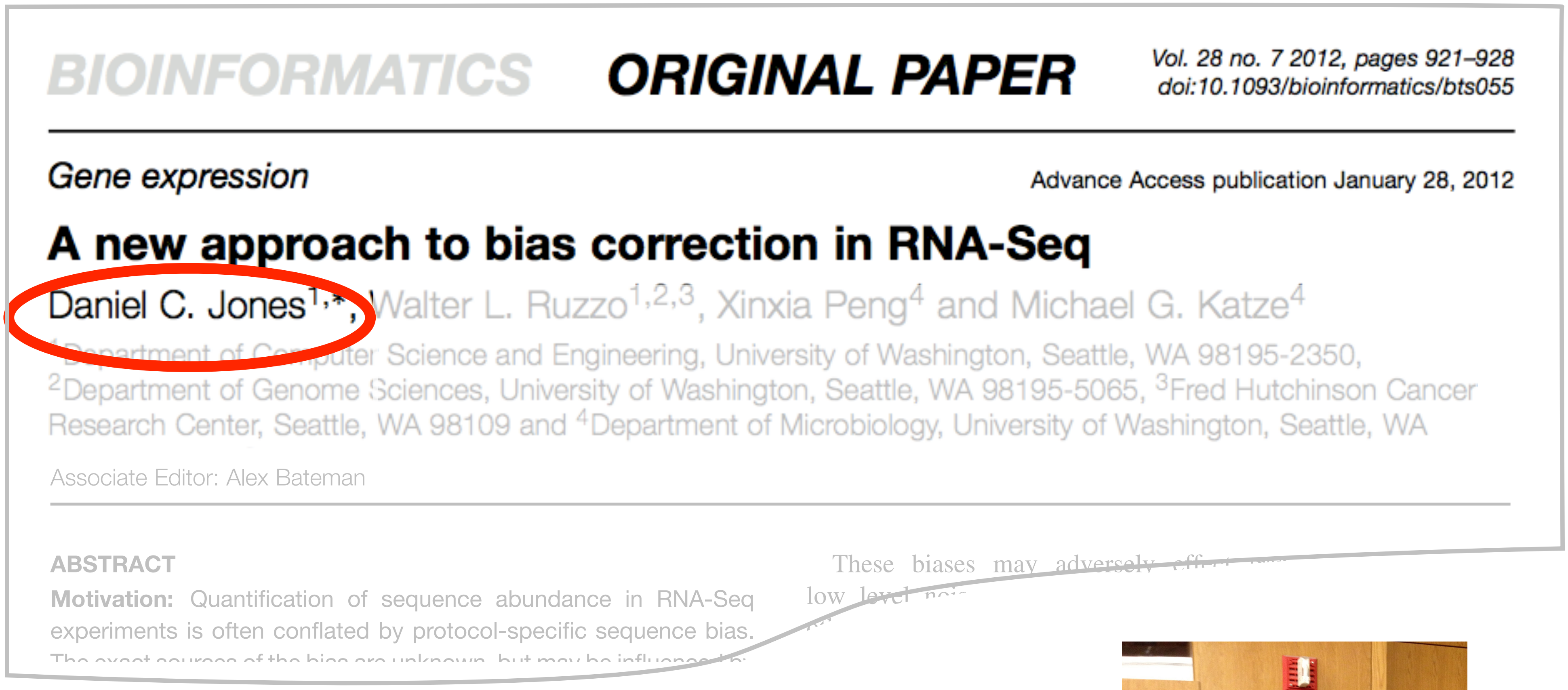
Sequence ~100-175bp from one or both ends

CAUTIONS: non-uniform sampling, sequence (e.g. G+C), 5’-3’, and length biases

Bias Correction in RNAseq

Walter L. (Larry) Ruzzo

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



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

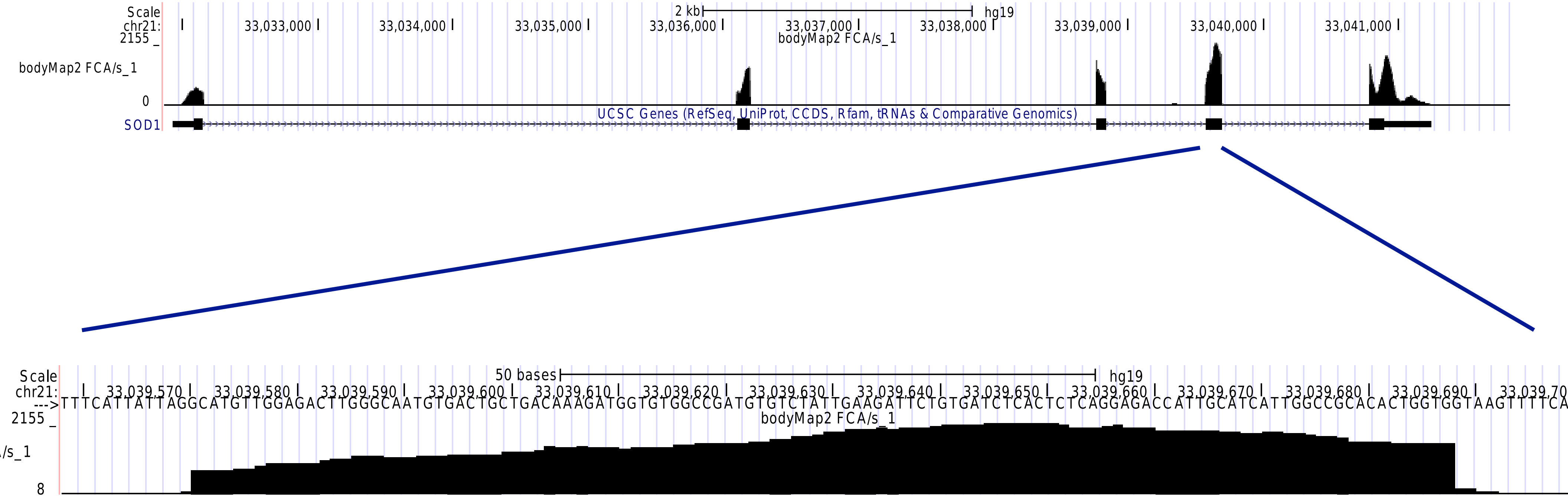
Daniel C. Jones[1,*], Walter L. Ruzzo[1,2,3], Xinxia Peng[4] and Michael G. Katze[4]
[1]Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195-2350, [2]Department of Genome Sciences, University of Washington, Seattle, WA 98195-5065, [3]Fred Hutchinson Cancer Research 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

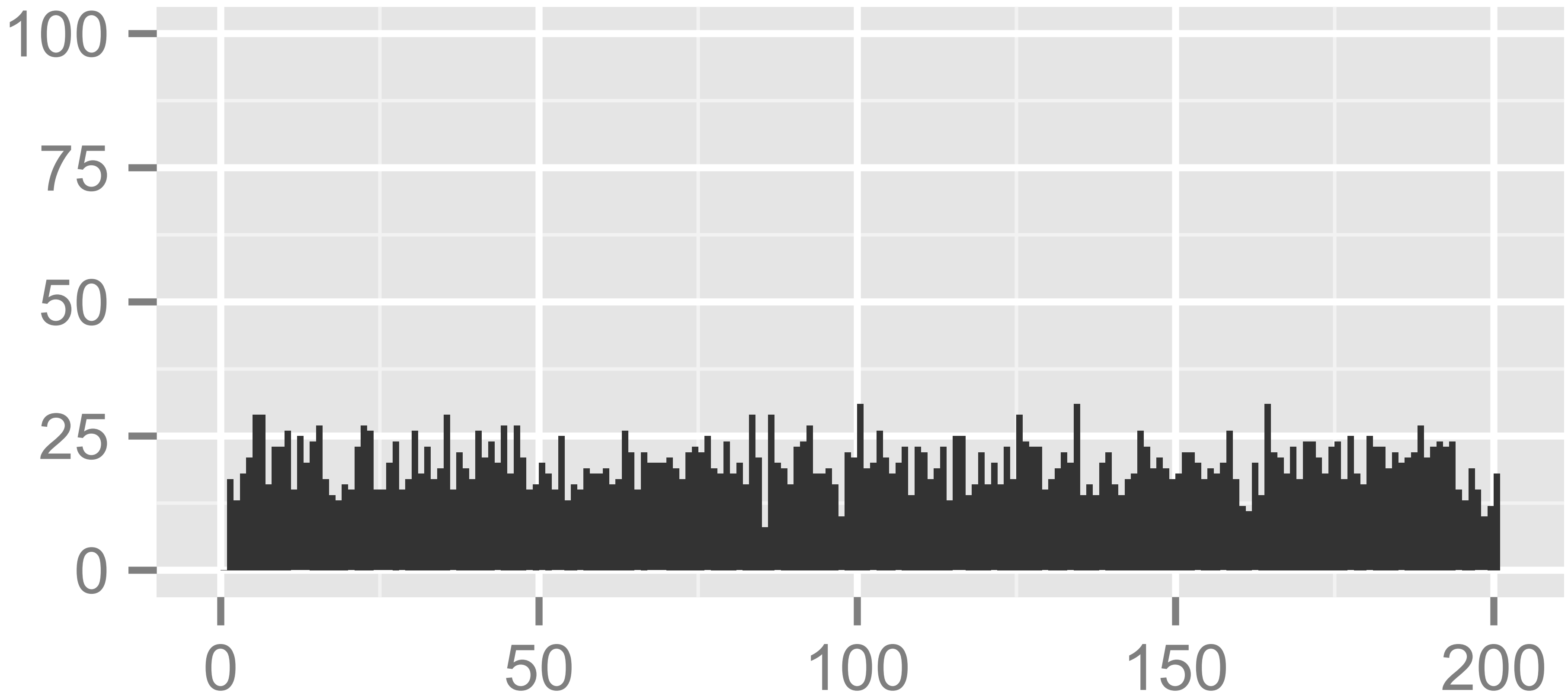
These biases may adversely
low level noi

RNA seq



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

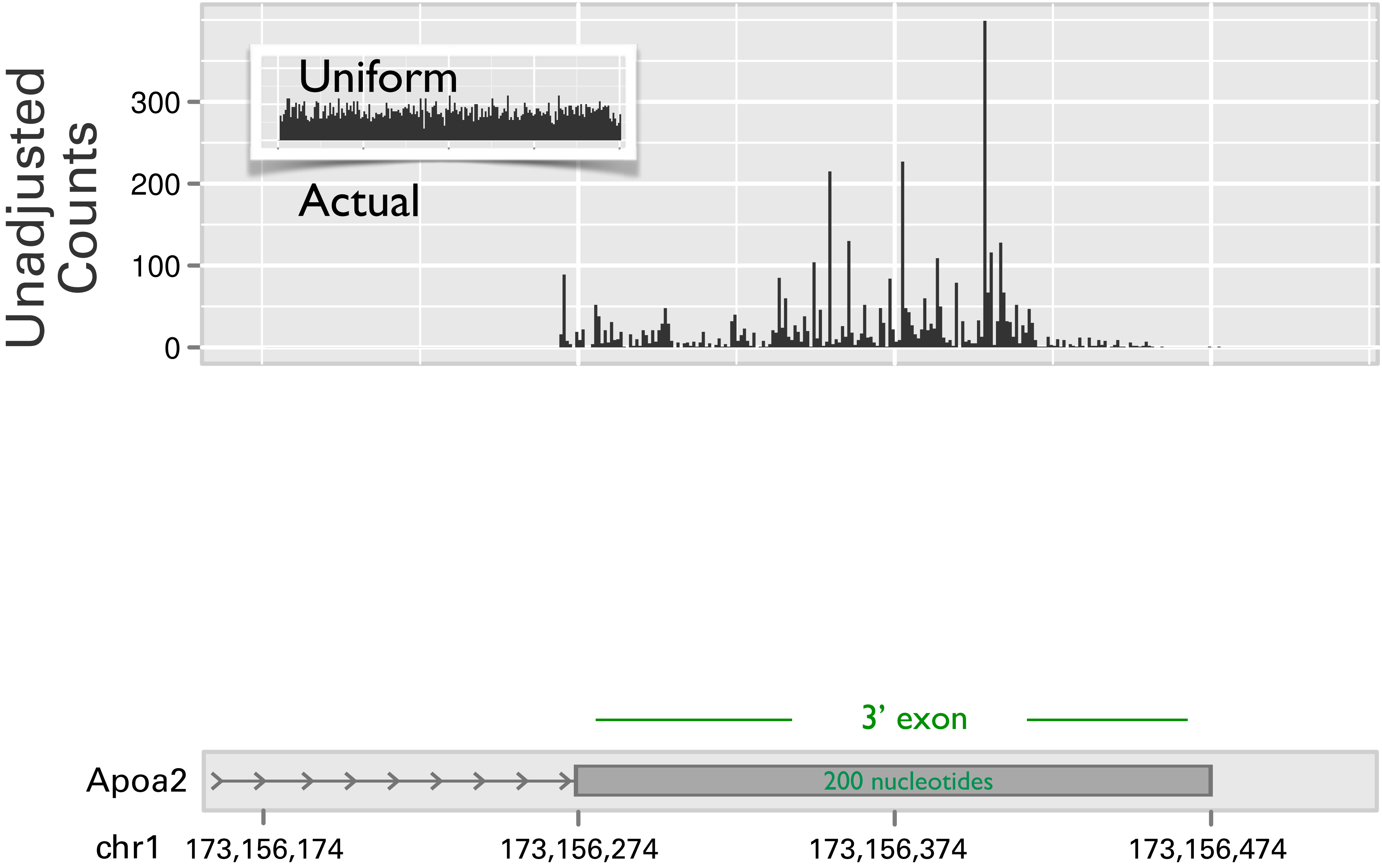
What we expect: Uniform Sampling



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

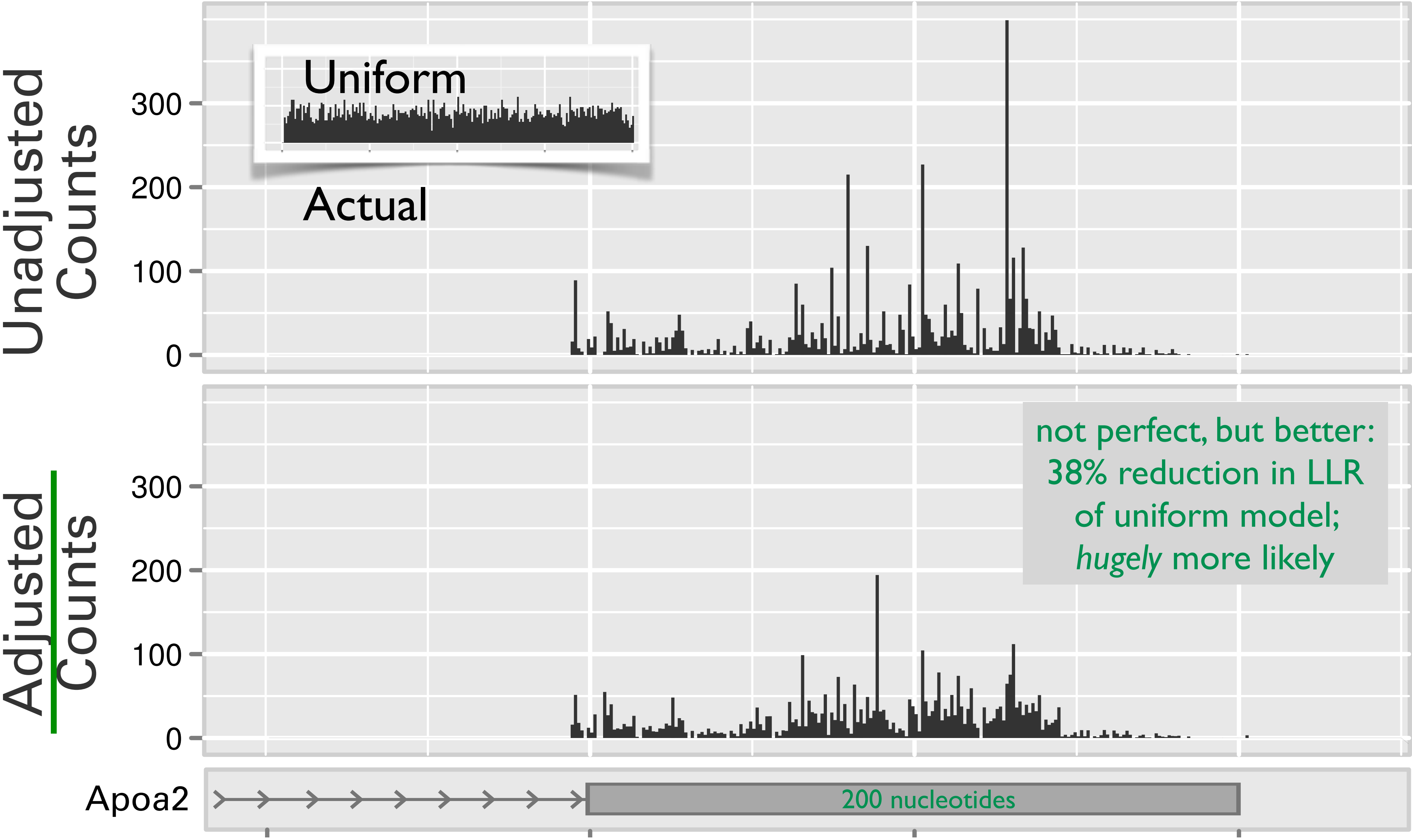
What we get: *highly* non-uniform coverage

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



Mortazavi data

What we get: *highly* non-uniform coverage

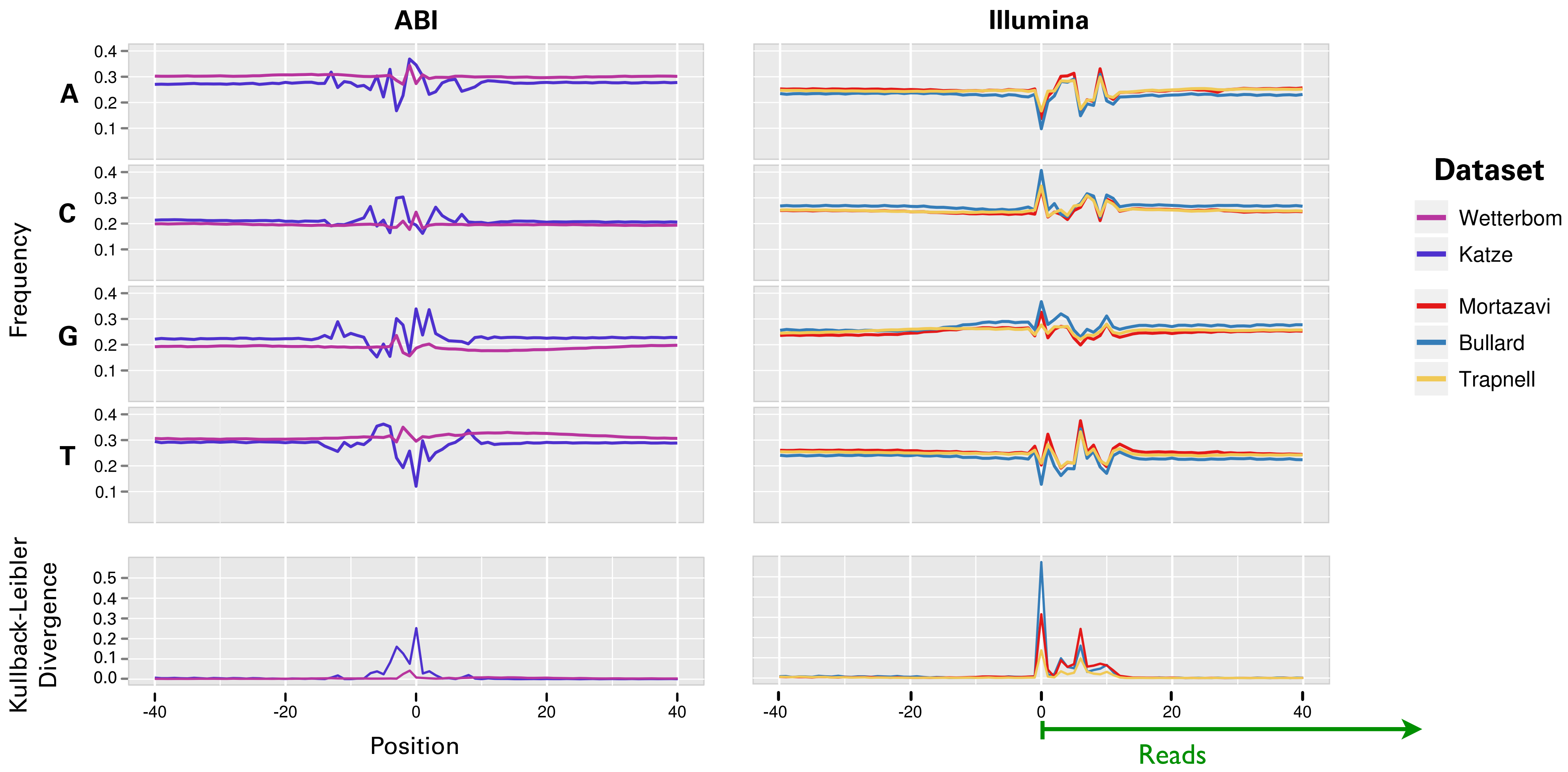


not perfect, but better: 38% reduction in LLR of uniform model; *hugely* more likely

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

(in part)

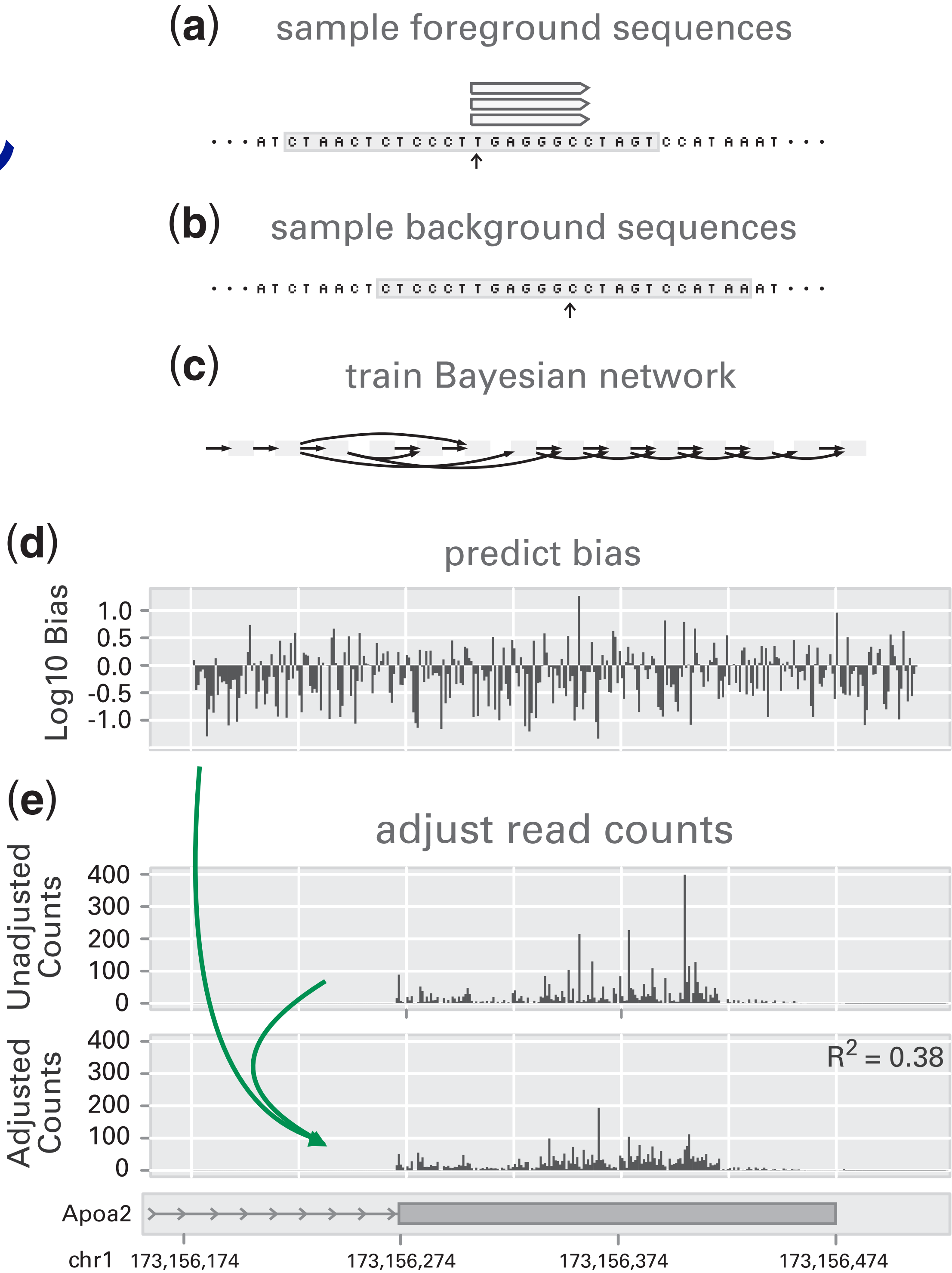
Bias is ^ sequence-dependent



and platform/sample-dependent

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

Method Outline



(a) sample foreground sequences

(b) sample background sequences

(c) train Bayesian network

(d) predict bias

(e) adjust read counts

Modeling Sequence Bias

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

Some obvious choices

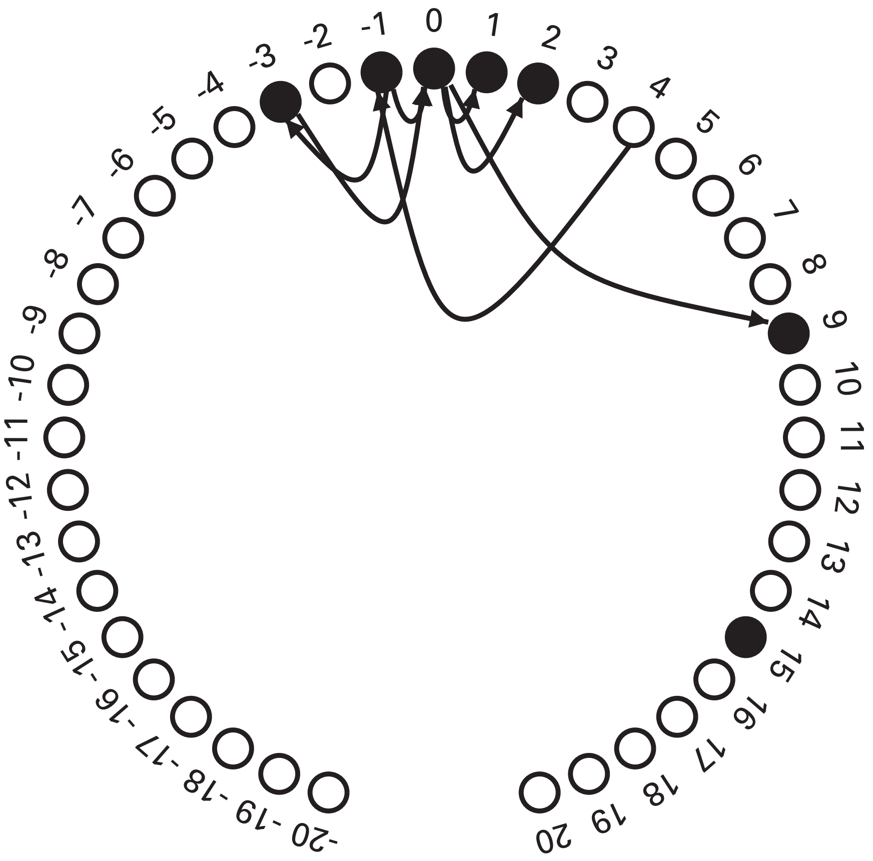
- Full joint distribution: 4^k-1 parameters
- PWM (0-th order Markov): (4-1)•k parameters

Something intermediate

- Directed Bayes network

Form of the models:

Directed Bayes nets



Wetterbom
(282 parameters)

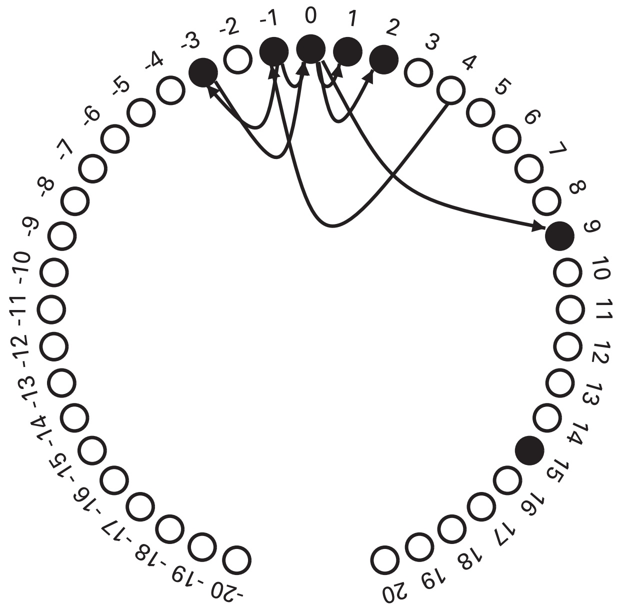
One "node" per nucleotide, ±20 bp of read start

- 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

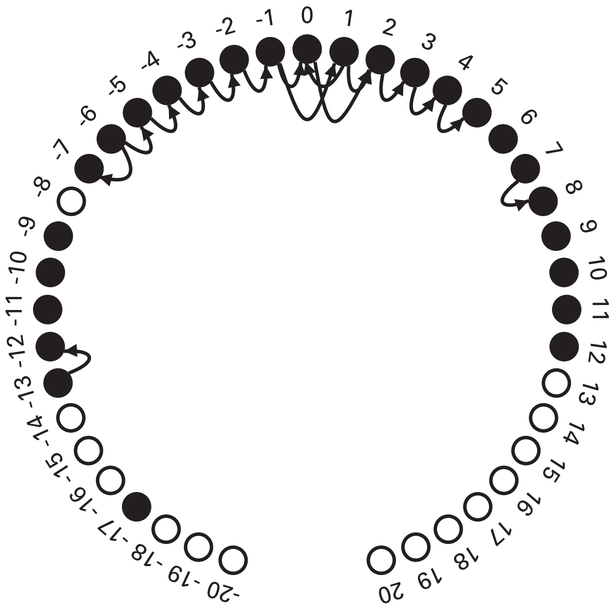
How–optimize:

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

ABI



Wetterbom
(282 parameters)

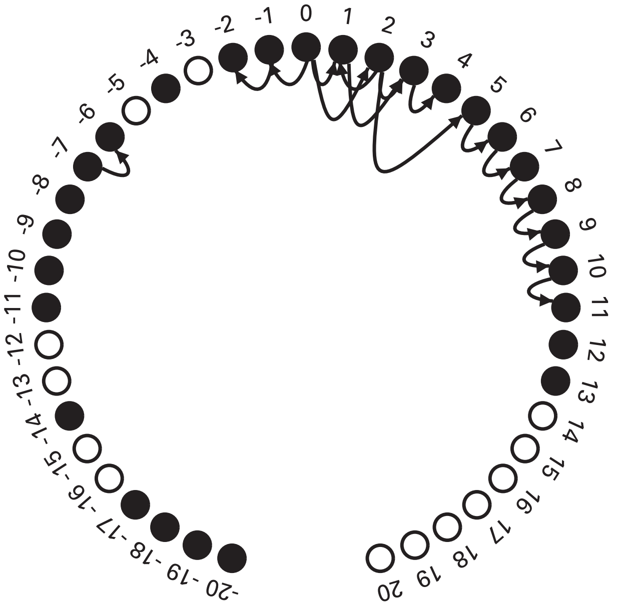


Katze
(684 parameters)

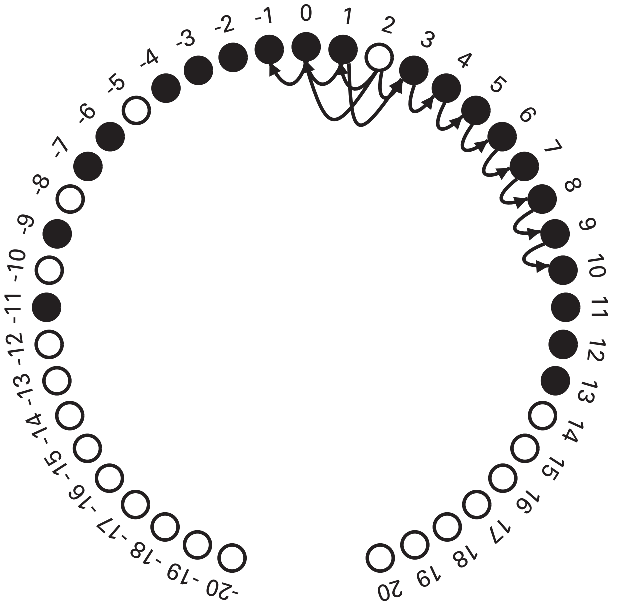
NB:

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

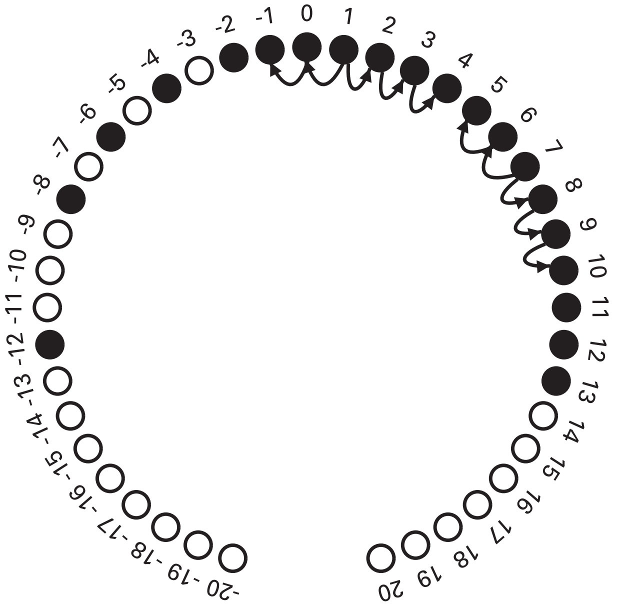
Illumina



Bullard
(696 parameters)



Mortazavi
(582 parameters)



Trapnell
(360 parameters)

Formally...

A reasonable definition of unbiasedness:

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

From Bayes...

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

So we might define **bias** as

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

Conditional Log-Likelihood

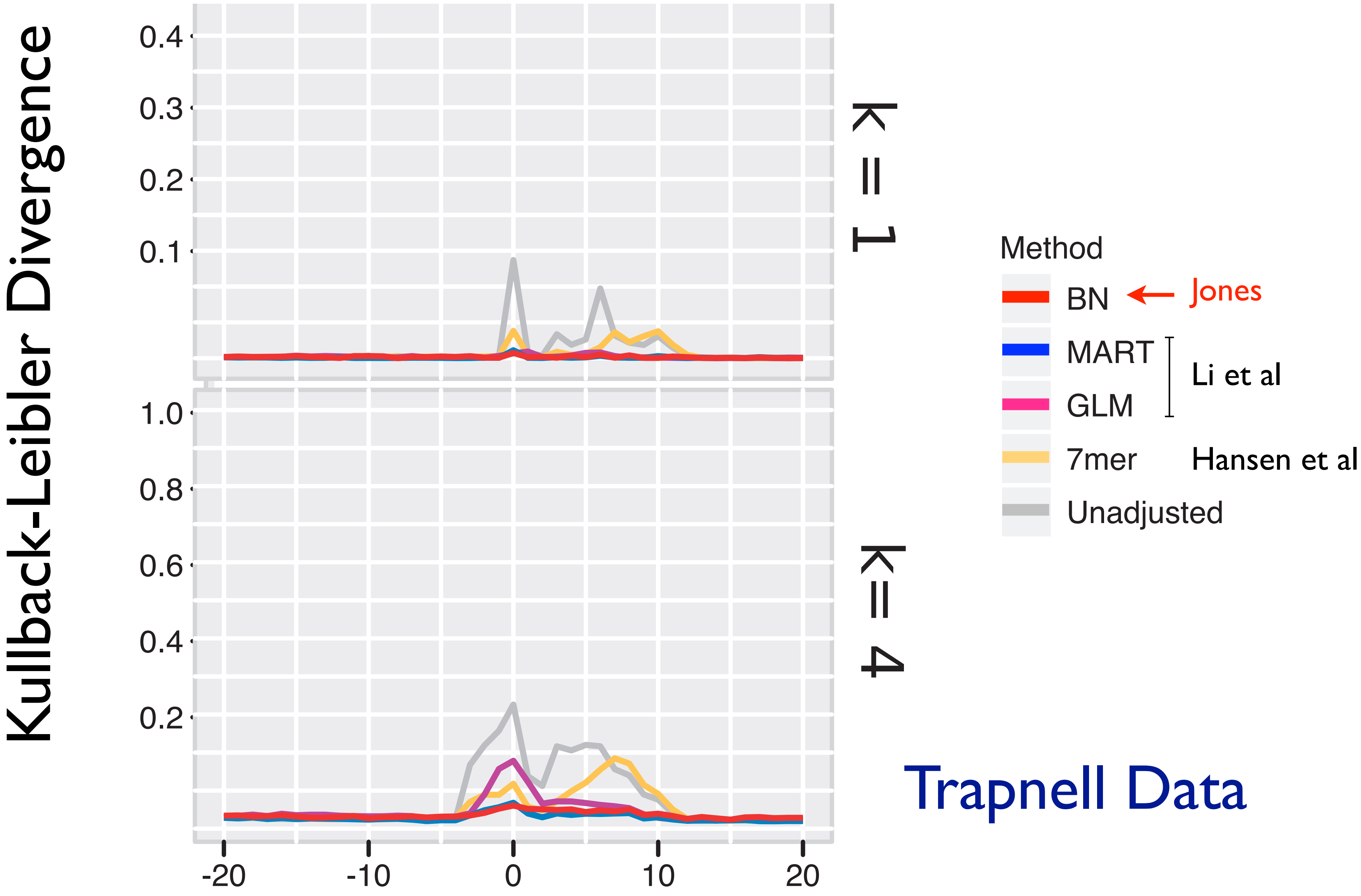
Find a graph that maximizes conditional log-likelihood.

$$\mathrm{CLL} = \sum_{i=1}^{n} \mathrm{Log}\Pr(x_i|s_i)$$

We need to penalize for model complexity as well.

$$\mathrm{CLL}' = 2\sum_{i=1}^{n} \mathrm{Log}\Pr(x_i|s_i) - m \log n$$

Result – Increased Uniformity

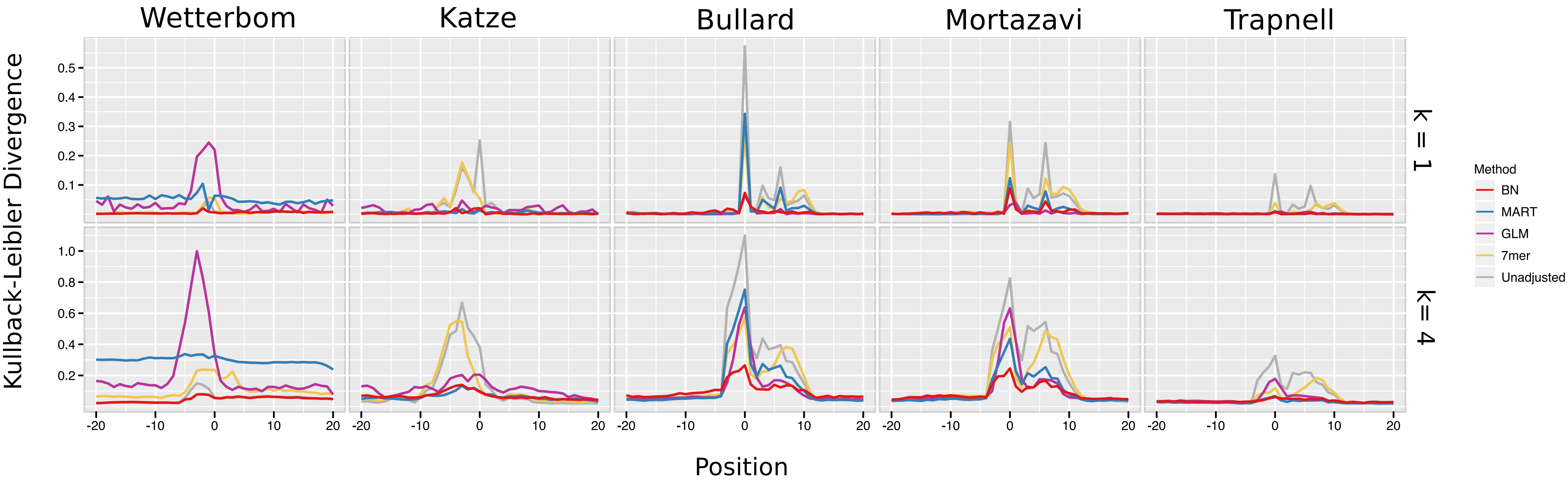


Jones

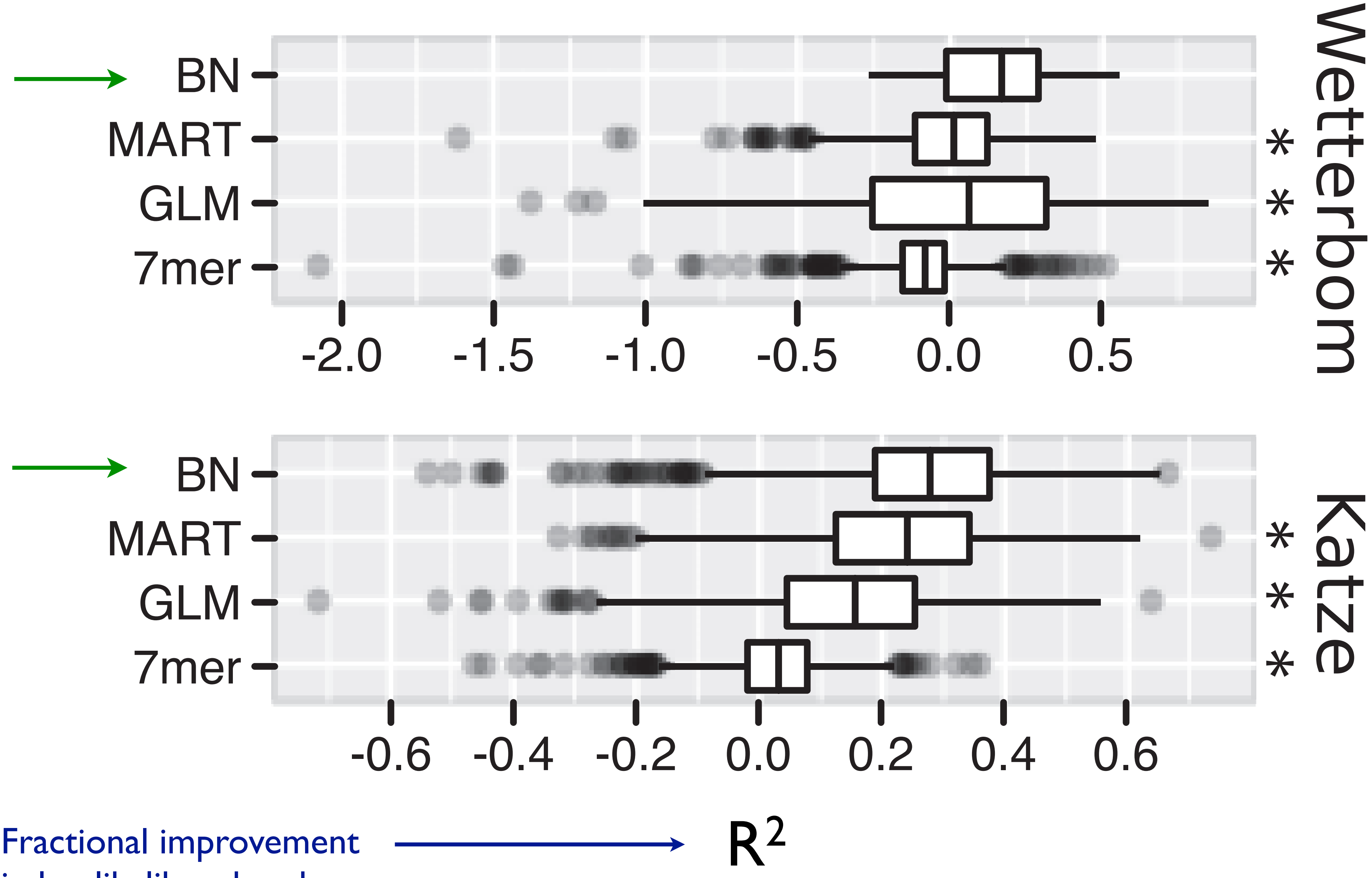
Li et al

Hansen et al

Trapnell Data


Wetterbom
Katze
Bullard
Mortazavi
Trapnell
k = 1
k= 4
Kullback-Leibler Divergence
0.5
0.4
0.3
0.2
0.1
1.0
0.8
0.6
0.4
0.2
-20
-10
0
10
20
Position
Method
BN
MART
GLM
7mer
Unadjusted

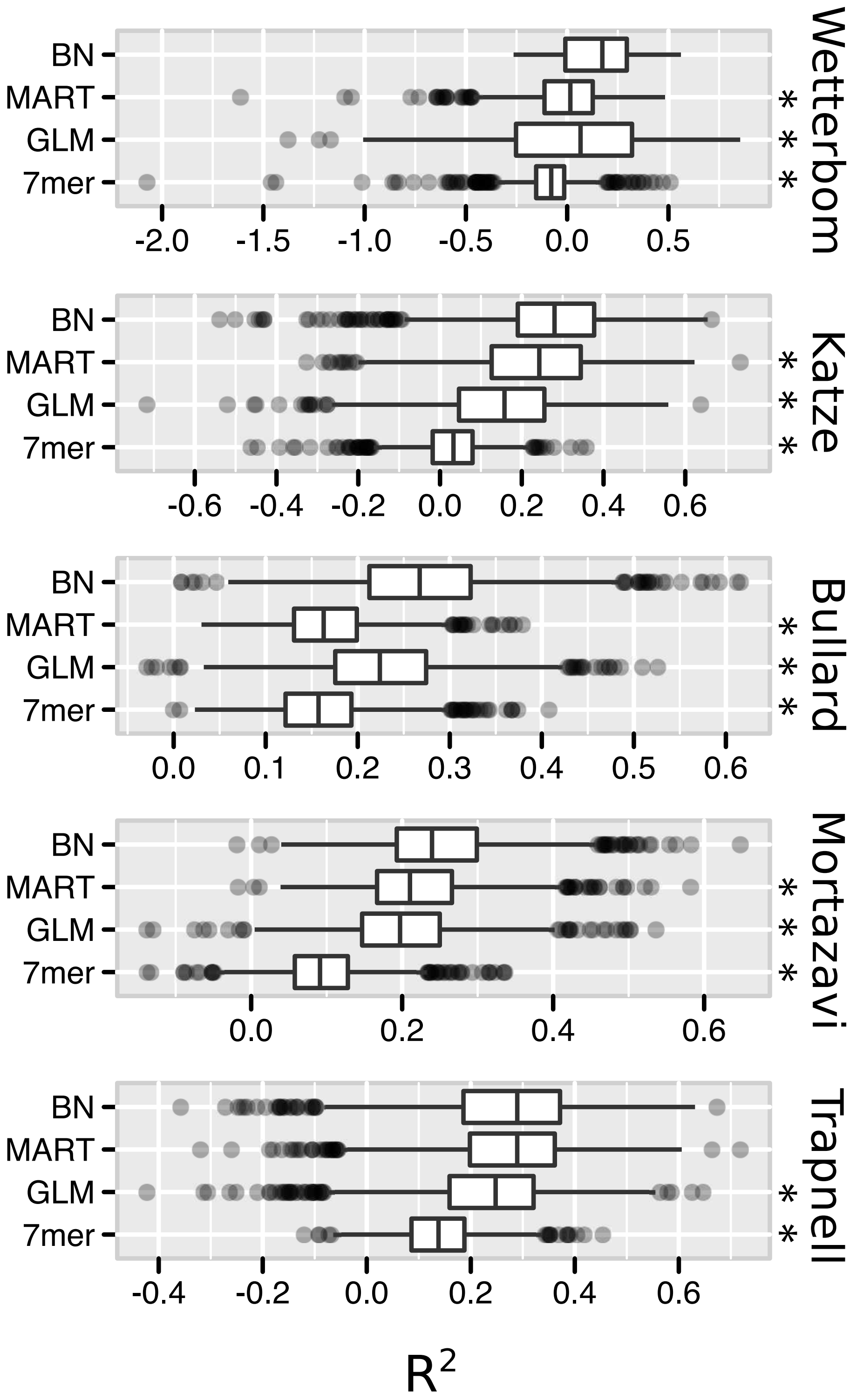
Result – Increased Uniformity



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)


Wetterbom
BN
MART
GLM
7mer
-2.0
-1.5
-1.0
-0.5
0.0
0.5
*
*
*
Katze
BN
MART
GLM
7mer
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
*
*
*
Bullard
BN
MART
GLM
7mer
0.0
0.1
0.2
0.3
0.4
0.5
0.6
*
*
*
Mortazavi
BN
MART
GLM
7mer
0.0
0.2
0.4
0.6
*
*
*
Trapnell
BN
MART
GLM
7mer
-0.4
-0.2
0.0
0.2
0.4
0.6
*
*
R^2

“First, do no harm”

Theorem:

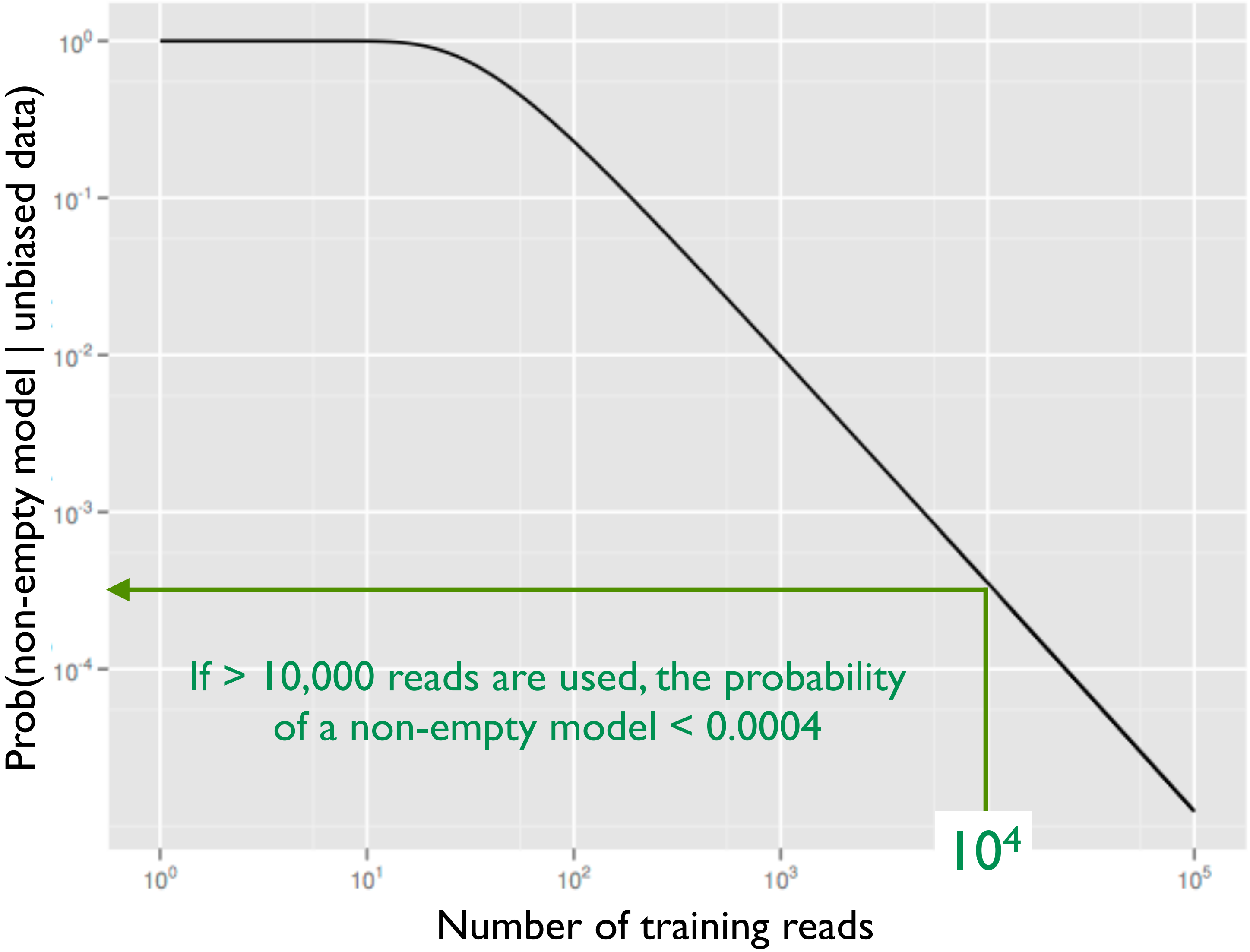
The probability of “false bias discovery,” i.e., of learning a non-empty model from *n* reads sampled from *un*biased data is less than

$$1 - (\Pr(X < 3 \log n))^{2h}$$

where *h* = number of nucleotides in the model and *X* is a random variable that (asymptotically in *n*) is χ^2 with 3 degrees of freedom. (E[*X*] = 3)

“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*.



If > 10,000 reads are used, the probability of a non-empty model < 0.0004

how different are two distributions?

Given: r-sided die, with probs $p_1...p_r$ of each face. Roll it n=10,000 times; observed frequencies = $q_1, \ldots, q_r$, (the MLEs for the unknown q_i's). How close is p_i to q_i?

Kullback-Leibler divergence, also known as *relative entropy*, of Q with respect to P is defined as

$$H(Q||P) = \sum_i q_i \ln \frac{q_i}{p_i}$$

where q_i (p_i) is the probability of observing the i^{th} event according to the distribution Q (resp., P), and the summation is taken over all events in the sample space (e.g., all k-mers). In some sense, this is a measure of the dissimilarity between the distributions: if $p_i \approx q_i$ everywhere, their log ratios will be near zero and H will be small; as q_i and p_i diverge, their log ratios will deviate from zero and H will increase.

Fancy name, simple idea: H(Q||P) is just the expected per-sample contribution to log-likelihood ratio test for "was X sampled from H_0: P vs H_1: Q?"

So, assuming the null hypothesis is false, in order for it to be rejected with say, $1000 : 1$ odds, one should choose m to be inversely proportional to $H(Q||P)$:

$$mH(Q||P) \geq \ln 1000$$

$$m \geq \frac{\ln 1000}{H(Q||P)}$$

Continuing the notation above, suppose P as an unknown distribution with parameters $p_1, \ldots, p_r$, $\sum p_i = 1$ where r is the number of points in the sample space (e.g. $r = 4^k$ in the case of k-mers). Given a random sample $X_1, X_2, \ldots, X_r$ of size $n = \sum_i X_i$ from P, it is well known that the maximum likelihood estimators for the parameters are $q_i = \frac{X_i}{n} \approx p_i$. How good an estimate for P is this distribution Q? The estimators are unbiased:

$$E[q_i] = E\left[\frac{X_i}{n}\right] = \frac{E[X_i]}{n} = \frac{np_i}{n} = p_i$$

and the standard deviation of each estimate is proportional to $1/\sqrt{n}$, so these estimates are increasingly accurate as the sample size increases. A more quantitative assessment of the accuracy of the estimator is obtained by evaluating the KL divergence:

$$H(Q||P) = \sum_{i=1}^{r} q_i \ln \frac{q_i}{p_i} = \sum_{i=1}^{r} q_i \ln \left(1 + \frac{q_i - p_i}{p_i}\right)$$

Using the first two terms of the Taylor series for $\ln(1+x)$, this is

$$\begin{aligned} H(Q||P) &\approx \sum_{i=1}^{r} q_i \left(\frac{q_i - p_i}{p_i} - \frac{1}{2}\left(\frac{q_i - p_i}{p_i}\right)^2 \right) \\ &= \sum_{i=1}^{r} q_i \frac{q_i - p_i}{p_i} - \frac{q_i}{2p_i}\frac{(q_i - p_i)^2}{p_i} \end{aligned}$$

Since $\sum_{i=1}^{r} q_i = \sum_{i=1}^{r} p_i = 1$, $\sum_{i=1}^{r} p_i \frac{q_i - p_i}{p_i} = 0$, so

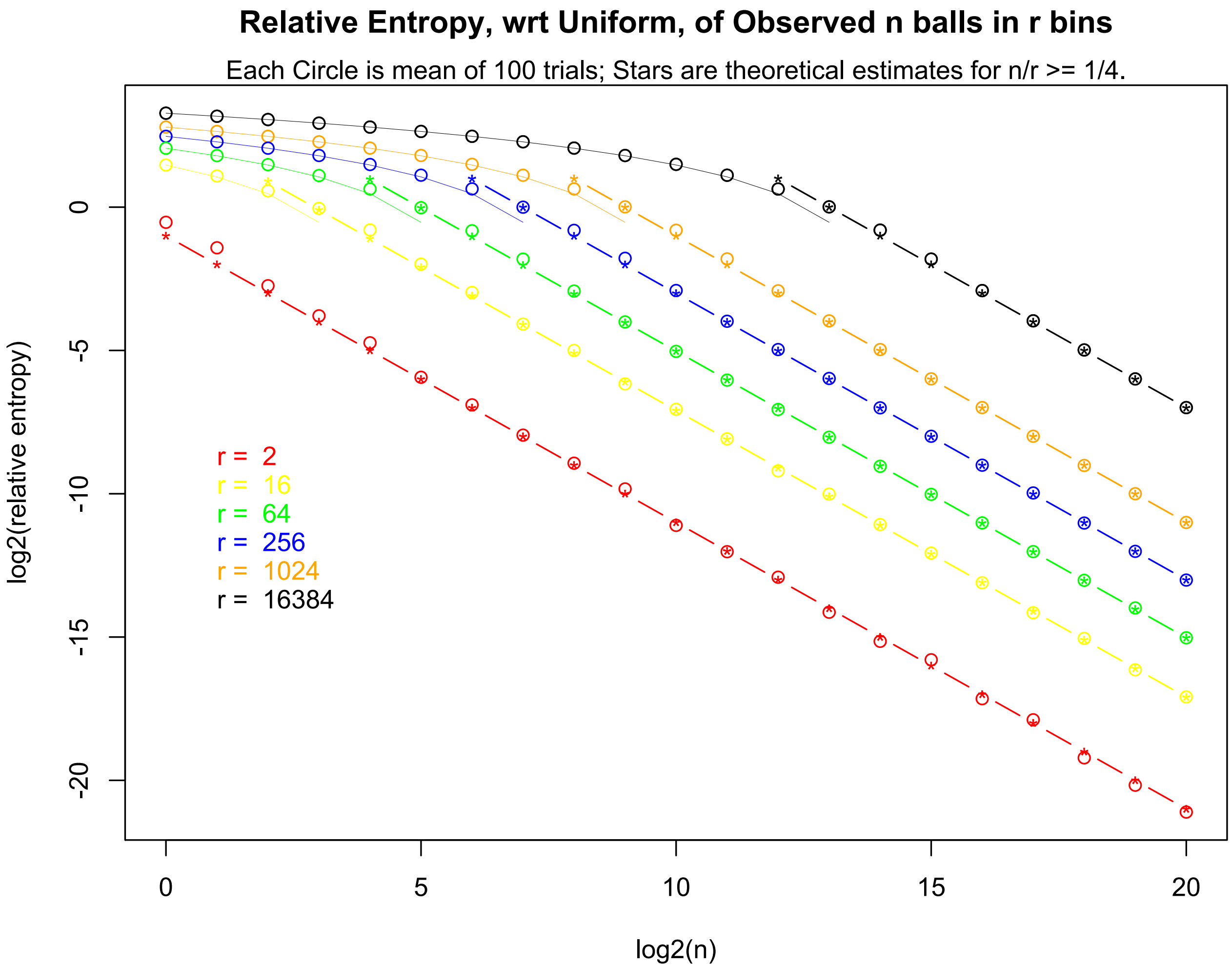
$$\begin{aligned} H(Q||P) &\approx \sum_{i=1}^{r} q_i \frac{q_i - p_i}{p_i} - p_i \frac{q_i - p_i}{p_i} - \frac{q_i}{2p_i}\frac{(q_i - p_i)^2}{p_i} \\ &= \sum_{i=1}^{r} \frac{(q_i - p_i)^2}{p_i} \left(1 - \frac{q_i}{2p_i}\right) \\ &\approx \frac{1}{2} \sum_{i=1}^{r} \frac{(q_i - p_i)^2}{p_i} \end{aligned}$$

since $q_i \approx p_i$. Multiplying by n^2/n^2 we have,

$$\begin{aligned} H(Q||P) &\approx \frac{1}{2n} \sum_{i=1}^{r} \frac{(nq_i - np_i)^2}{np_i} \\ &= \frac{1}{2n} \sum_{i=1}^{r} \frac{(X_i - E[X_i])^2}{E[X_i]} \end{aligned}$$

The summation is the test statistic for the χ^2 goodness-of-fit test for a multinomial distribution, and as $n \to \infty$ is known to follow a χ^2 distribution with $r-1$ degrees of freedom. Finally, the expected value of such a random variable is $r-1$, hence the expected KL divergence of the MLE inferred distribution Q with respect to the true distribution P is

$$E[H(Q||P)] = \frac{r-1}{2n} \tag{1}$$

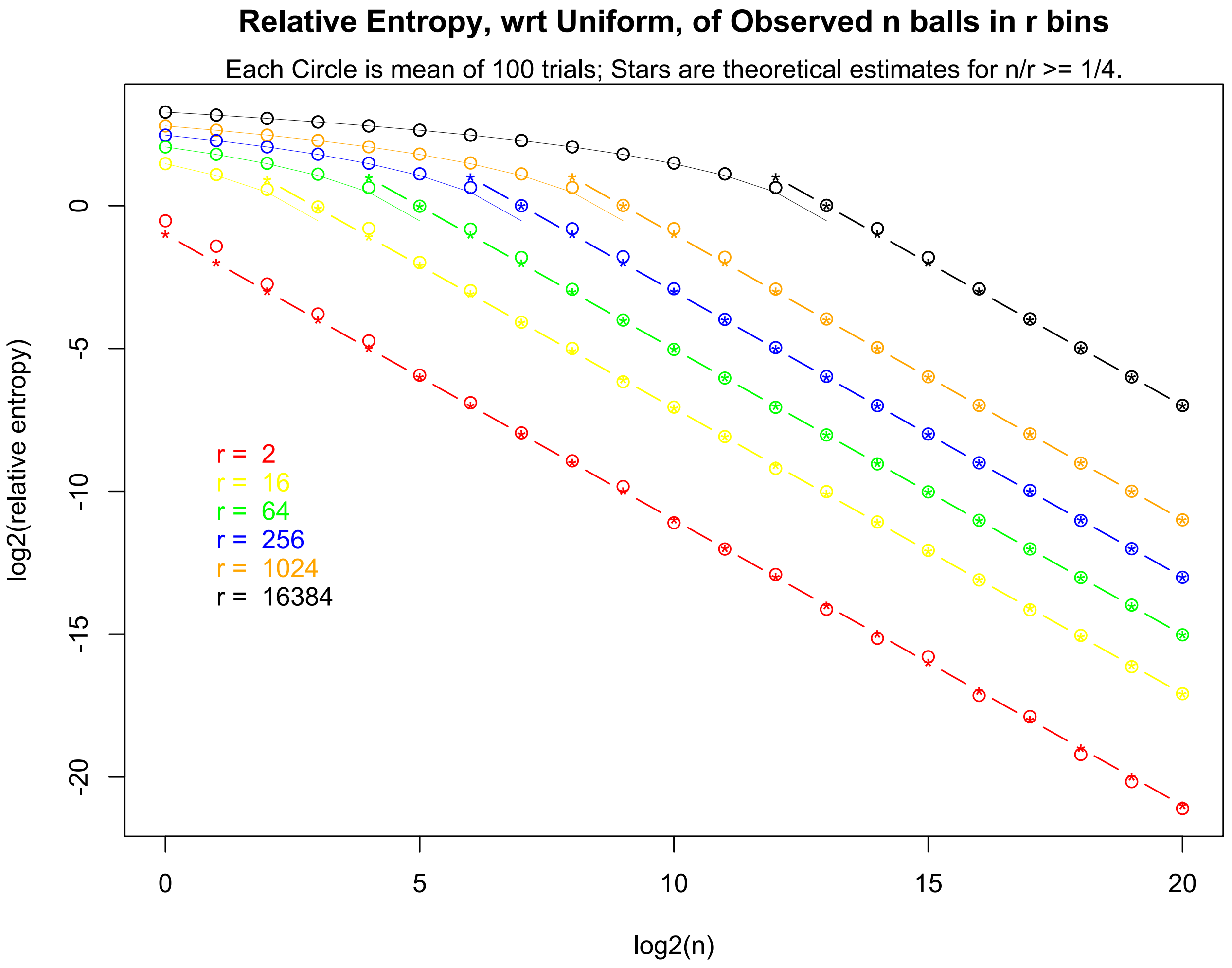


… and after a modicum of algebra:

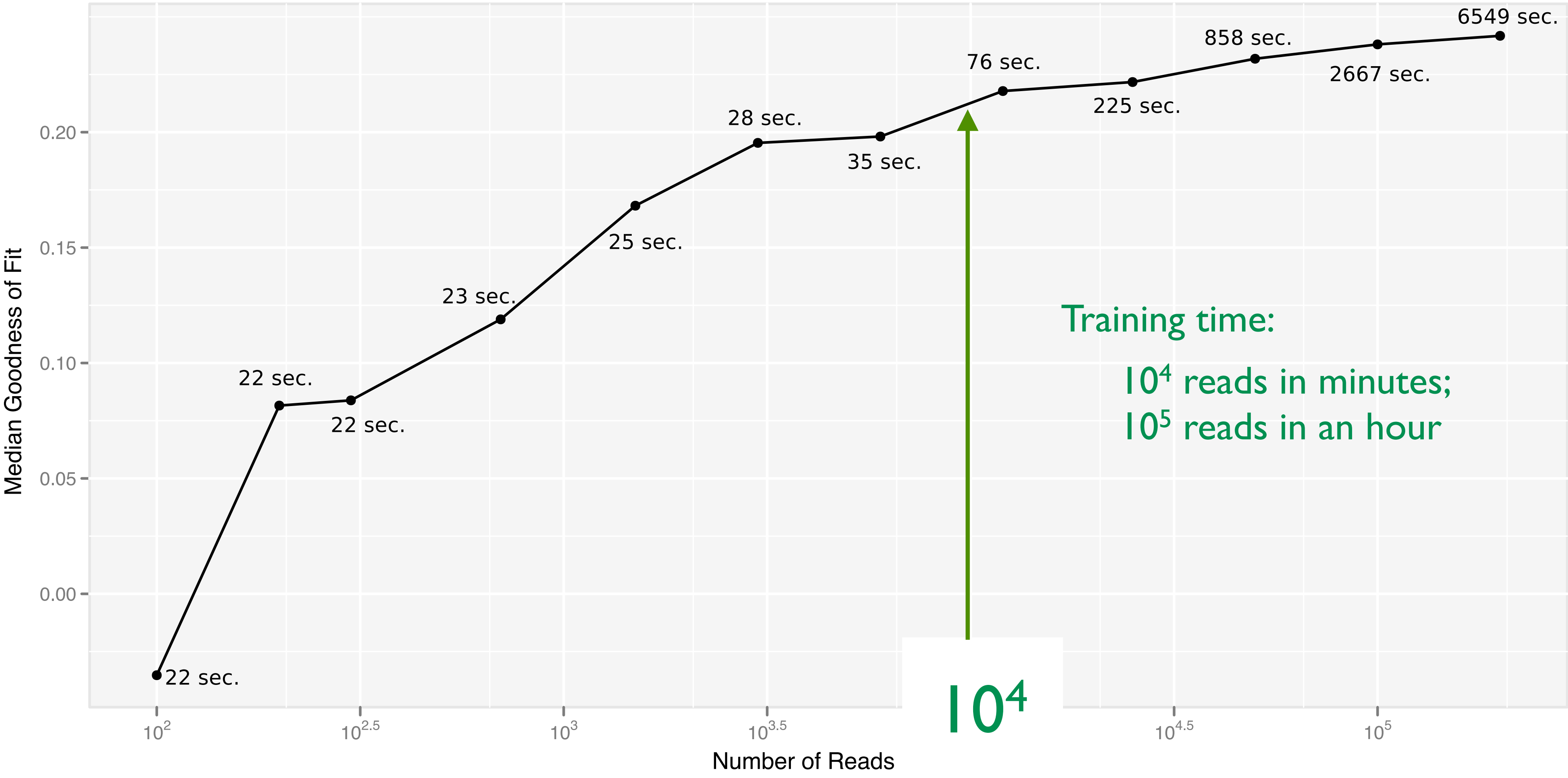
$$E[H(Q||P)] \approx \frac{r-1}{2n}$$

← LLR of error rises with number of parameters r, declines with size of training set n

… which empirically is a good approximation:



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



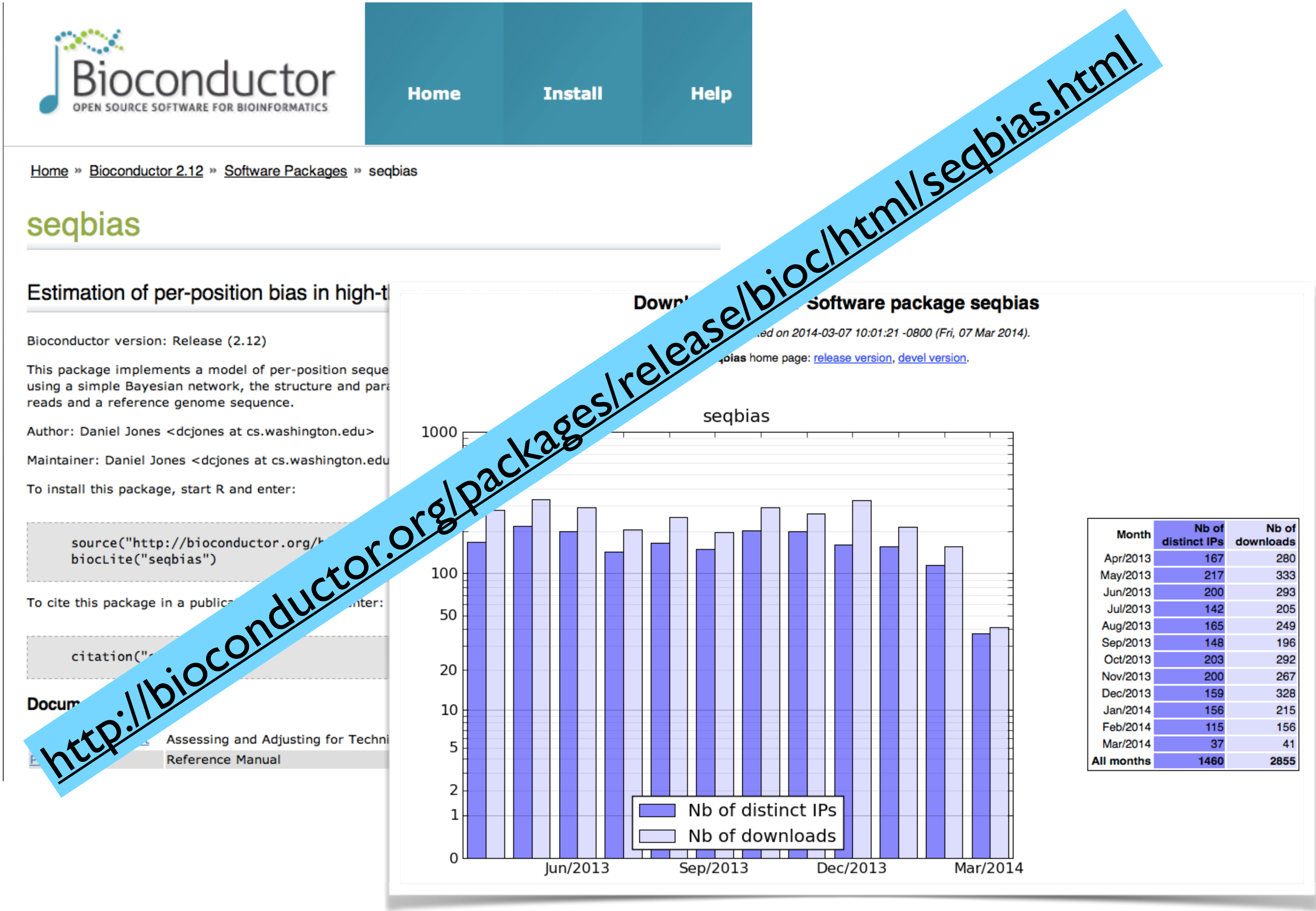
10^4

Training time:

10^4 reads in minutes;
10^5 reads in an hour

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

Availability



Home » Bioconductor 2.12 » Software Packages » seqbias

seqbias

Estimation of per-position bias in high-t

Bioconductor version: Release (2.12)

This package implements a model of per-position seque
using a simple Bayesian network, the structure and par
reads and a reference genome sequence.

Author: Daniel Jones <dcjones at cs.washington.edu>

Maintainer: Daniel Jones <dcjones at cs.washington.edu

To install this package, start R and enter:

```
source("http://bioconductor.org/b
biocLite("seqbias")
```

To cite this package in a public   nter:

```
citation("
```

Docum

Assessing and Adjusting for Techni
Reference Manual

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

Down   Software package seqbias

ed on 2014-03-07 10:01:21 -0800 (Fri, 07 Mar 2014).

bias home page: release version, devel version.

Month	Nb of distinct IPs	Nb of downloads
Apr/2013	167	280
May/2013	217	333
Jun/2013	200	293
Jul/2013	142	205
Aug/2013	165	249
Sep/2013	148	196
Oct/2013	203	292
Nov/2013	200	267
Dec/2013	159	328
Jan/2014	156	215
Feb/2014	115	156
Mar/2014	37	41
All months	1460	2855

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

CSEP 590 B
Computational Biology

Course Wrap Up

What is DNA? RNA?

How many Amino Acids are there?

Did human beings, as we know them, develop from earlier species of animals?

What are stem cells?

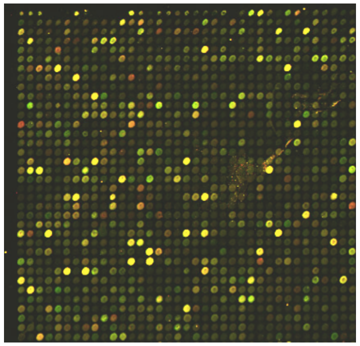
What did Viterbi invent?

What is dynamic programming?

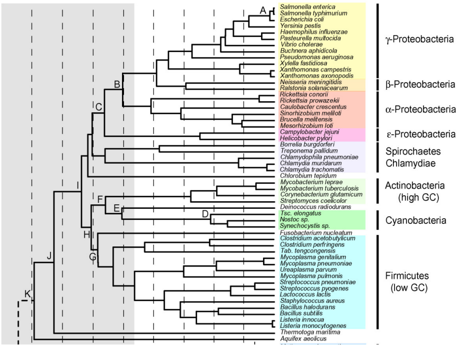
What is a likelihood ratio test?

What is the EM algorithm?

How would you find the maximum of $f(x) = ax^3 + bx^2 + cx + d$ in the interval $-10<x<25$?



“High-Throughput BioTech”

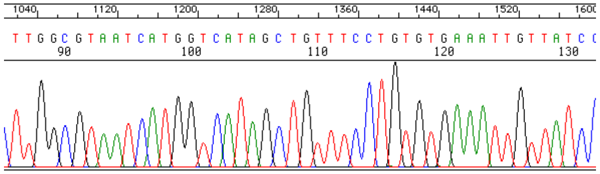


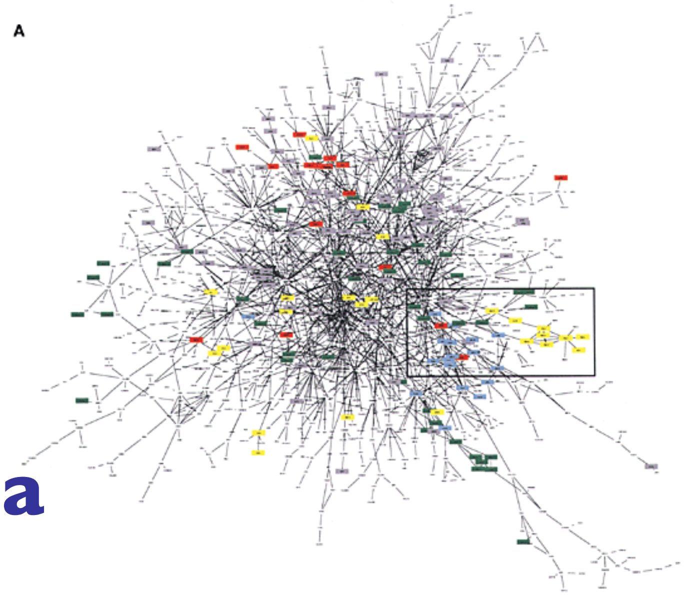
Sensors

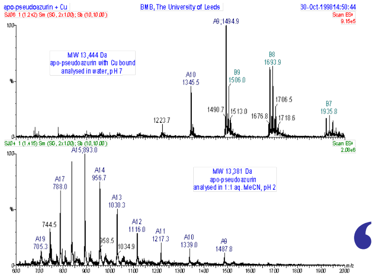
- DNA sequencing
- Microarrays/Gene expression
- Mass Spectrometry/Proteomics
- Protein/protein & DNA/protein interaction

Controls

- Cloning
- Gene editing/knock out/knock in
- RNAi







Floods of data

“Grand Challenge” problems

CS Points of Contact

Scientific visualization

- Gene expression patterns

Databases

- Integration of disparate, overlapping data sources
- Distributed genome annotation in face of shifting underlying coordinates

AI/NLP/Text Mining

- Information extraction from journal texts with inconsistent nomenclature, indirect interactions, incomplete/inaccurate models,…

Machine learning

- System level synthesis of cell behavior from low-level heterogeneous data (DNA sequence, gene expression, protein interaction, mass spec, …)

Algorithms

…

Frontiers & Opportunities

New data:

Proteomics, SNP, arrays, CGH, comparative sequence information, epigenomics, chromatin structure, ncRNA, interactome, single-cell everything

New methods:

graphical models, rigorous filtering

Data integration

many, complex, noisy sources

Systems Biology

Frontiers & Opportunities

Open Problems:

- splicing, alternative splicing
- multiple sequence alignment (genome scale, w/ RNA etc.)
- protein & RNA structure
- interaction modeling
- regulation, at all levels
- network models
- RNA trafficing
- ncRNA discovery
- …

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 before 12/7