Bias Correction in RNAseq

Walter L. (Larry) Ruzzo

Computer Science and Engineering
Genome Sciences (Adjunct)
University of Washington
Fred Hutchinson Cancer Research Center (Joint Member)
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

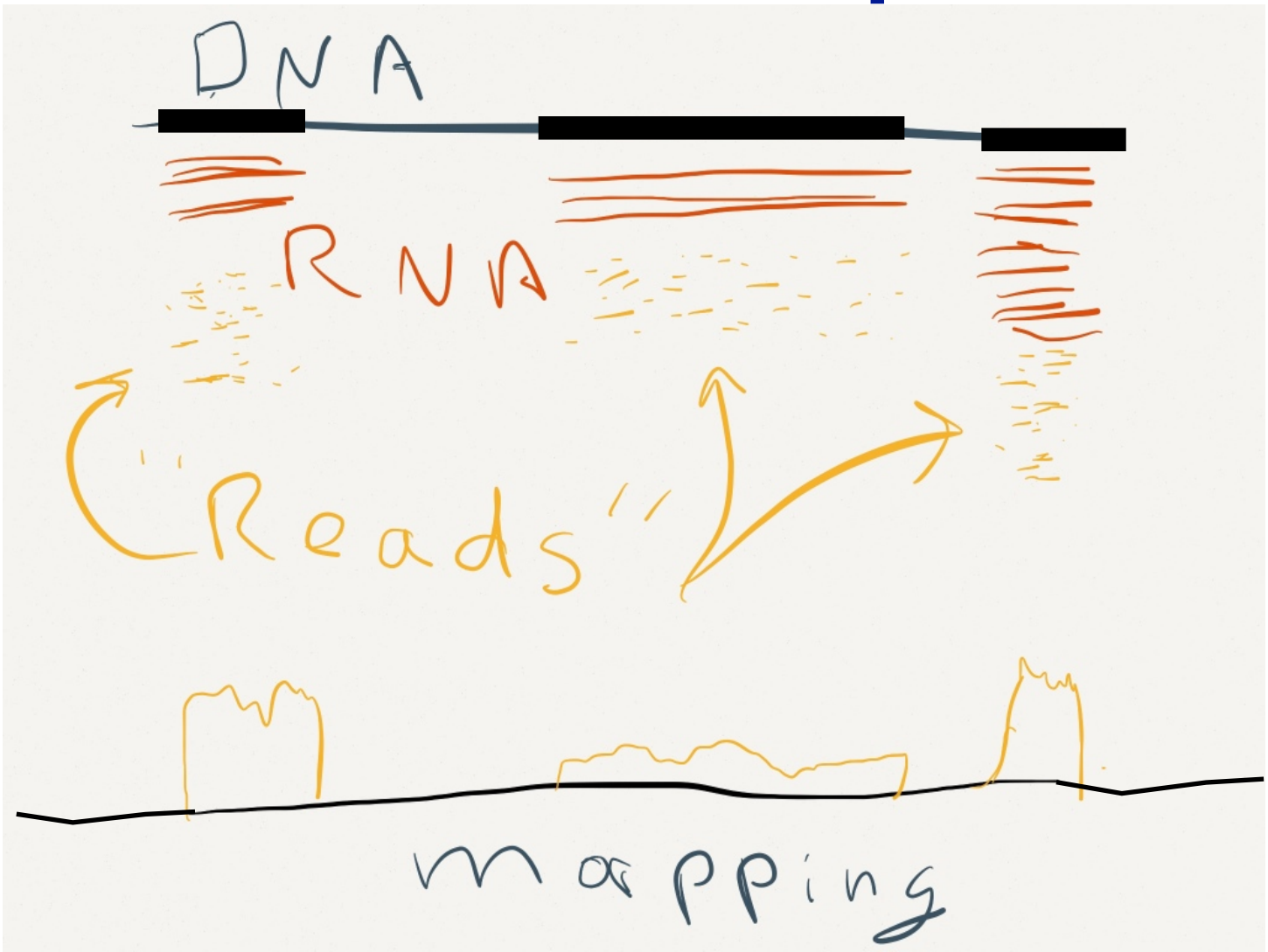
Associate Editor: Alex Bateman

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

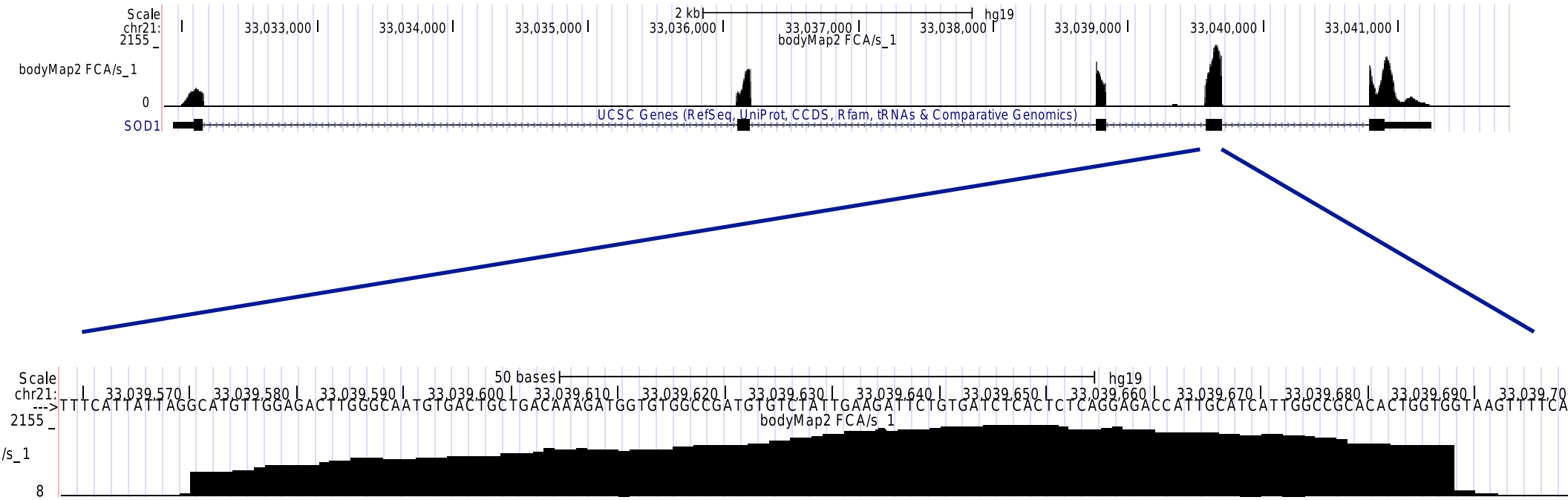
These biases may adversely [illegible] low level [illegible]

RNAseq

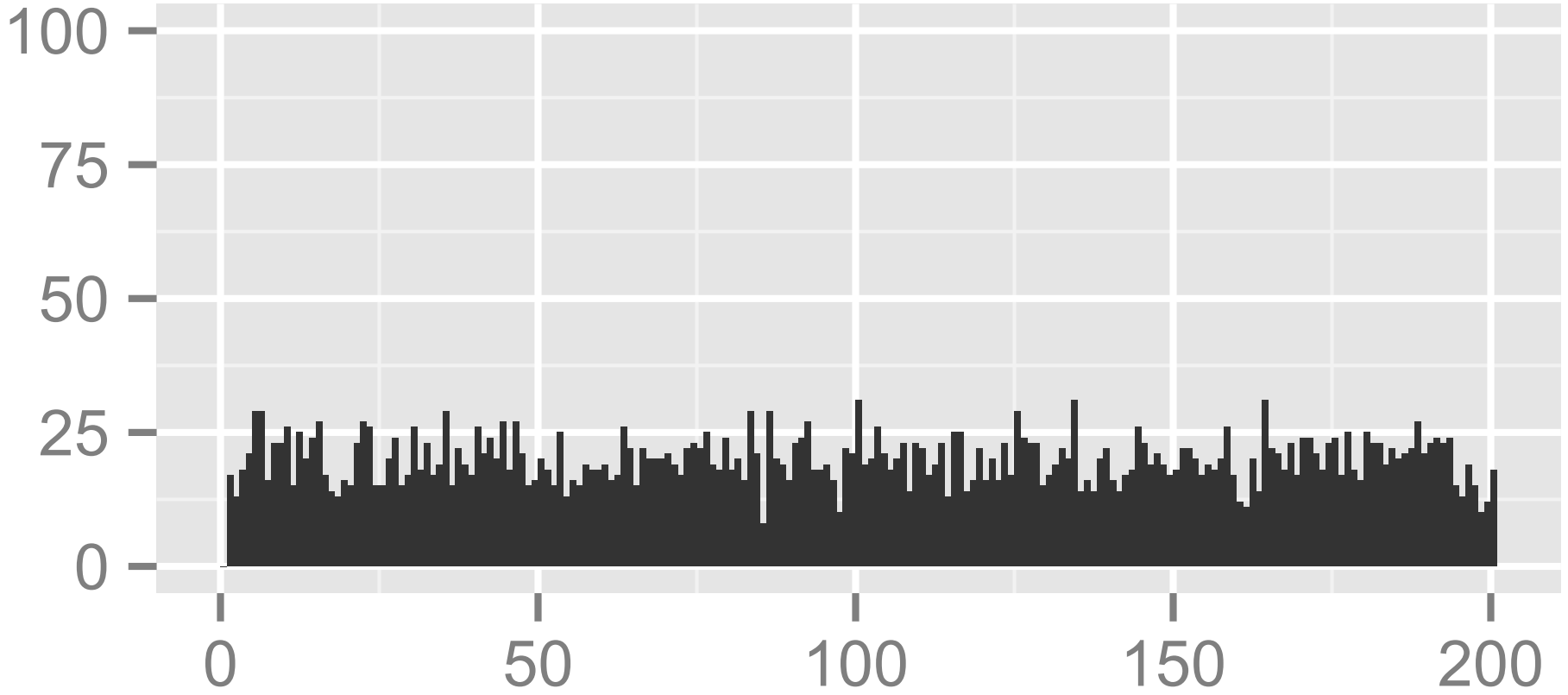


Extract RNA. Fragment it. Sequence it. Map it. Count it.
A random sampling process.

Example


Scale
chr21:
2155
33,033,000
33,034,000
33,035,000
2 kb
33,036,000
33,037,000
33,038,000
hg19
bodyMap2 FCA/s_1
33,039,000
33,040,000
33,041,000
bodyMap2 FCA/s_1
0
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
SOD1
Scale
chr21:
50 bases
hg19
33,039,570
33,039,580
33,039,590
33,039,600
33,039,610
33,039,620
33,039,630
33,039,640
33,039,650
33,039,660
33,039,670
33,039,680
33,039,690
33,039,70
--->TTTCATTATTAGGCATGTTGGAGACTTGGGCAATGTGACTGCTGACAAAGATGGTGTGGCCGATGTGTCTATTGAAGATTCTGTGATCTCACTCTCAGGAGACCATTGCATCATTGGCCGCACACTGGTGGTAAGTTTTCA
2155
bodyMap2 FCA/s_1
/s_1
8

What we expect: Uniform Sampling



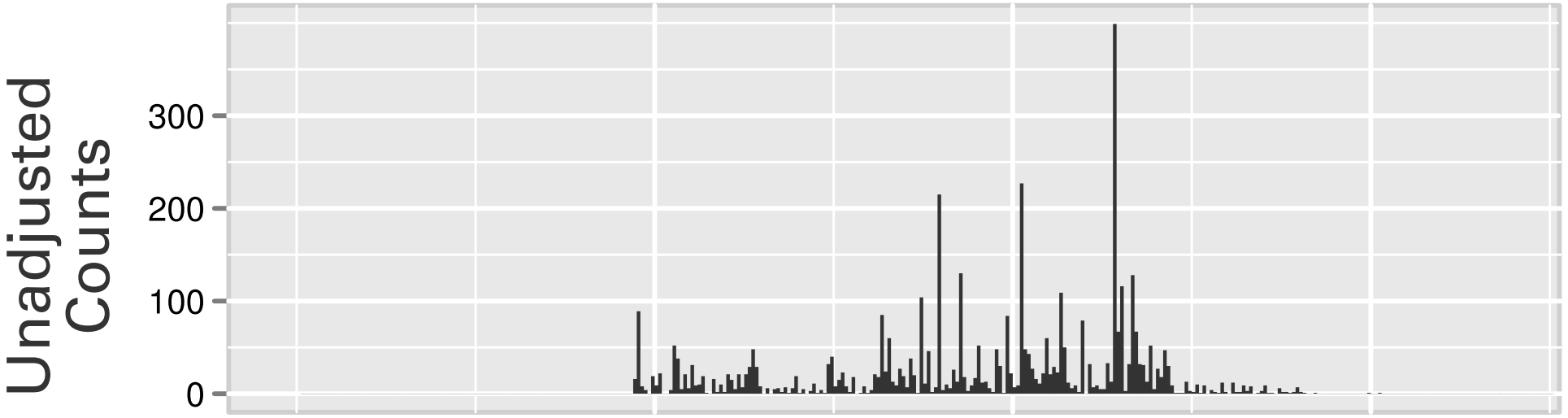
Uniform sampling of 4000 "reads" across a 200 bp "exon."

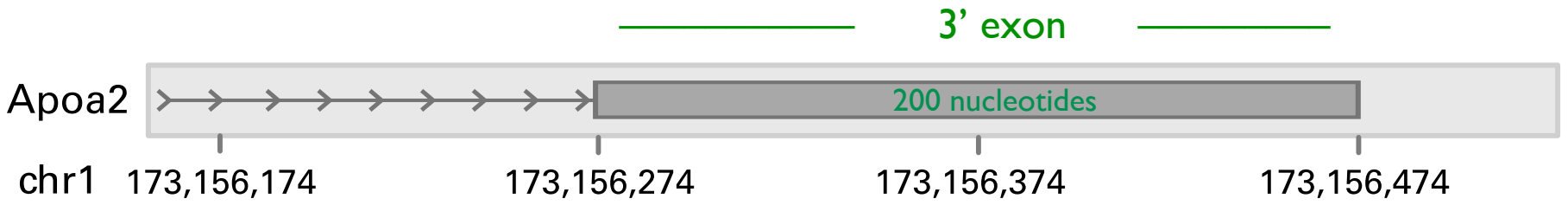
Average 20 ± 4.7 per position, min ≈ 9, max ≈33

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

What we get: *highly* non-uniform coverage

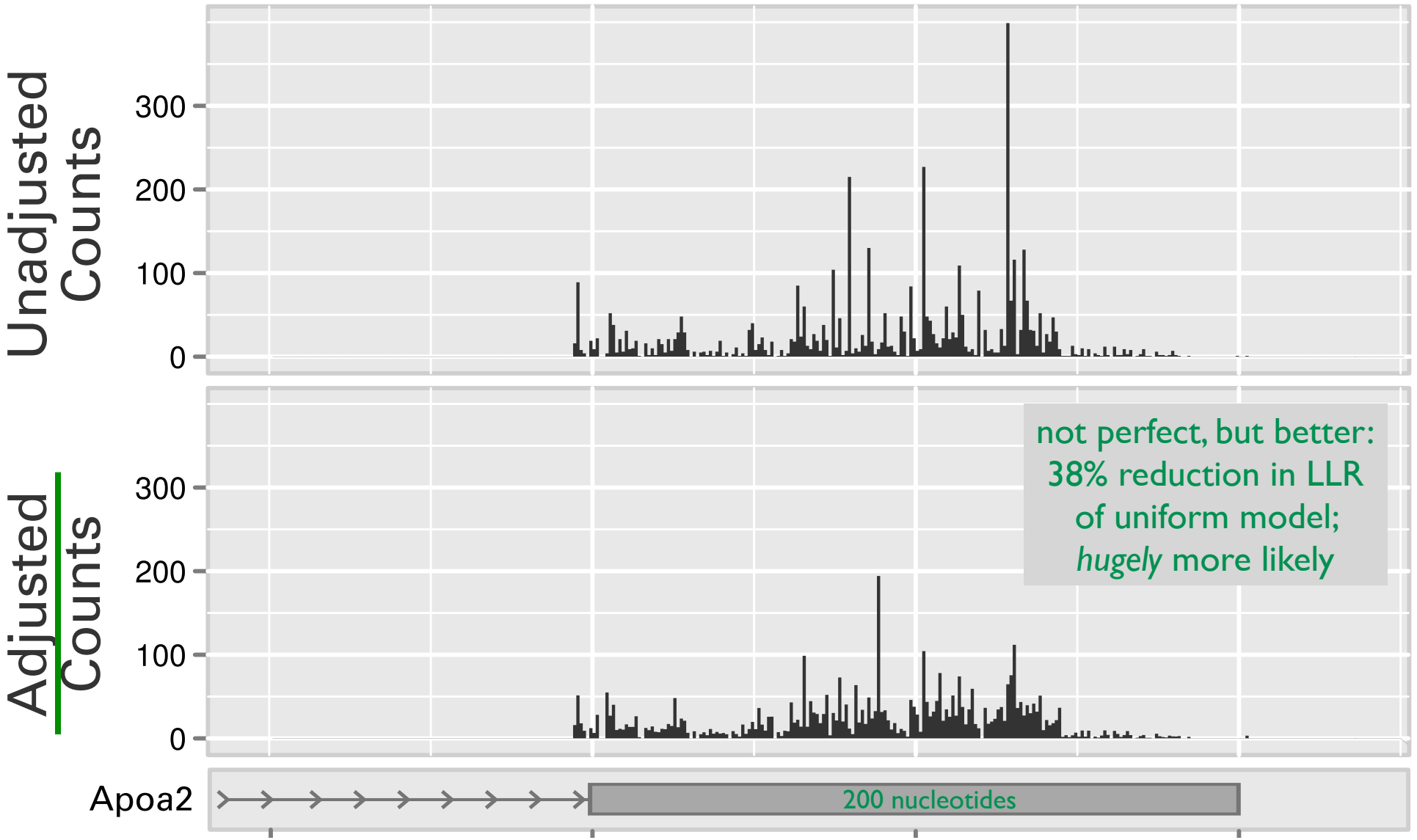
E.g., assuming uniform, the 8 peaks above 100 are ≳ +10σ above mean





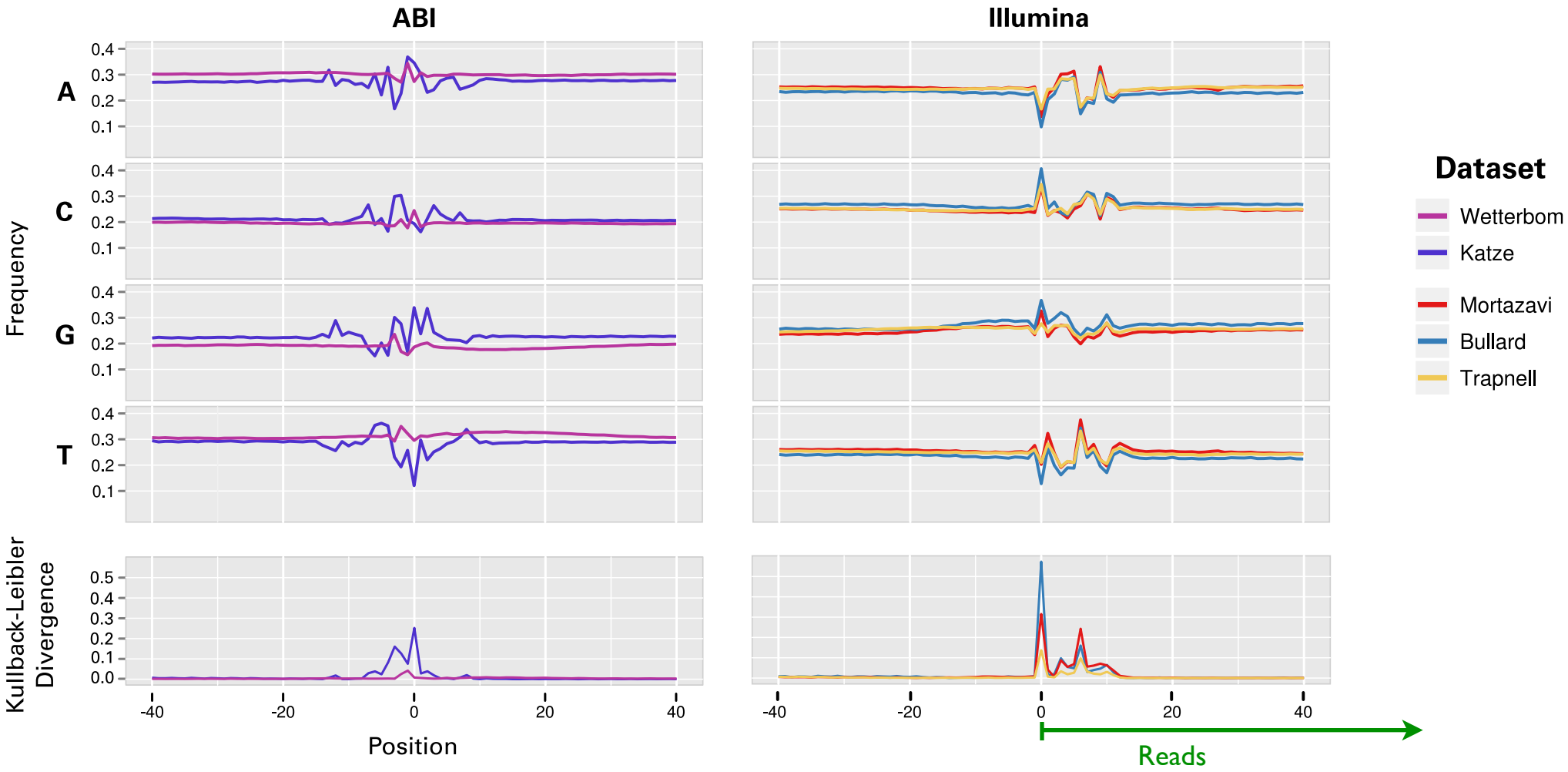
Mortazavi data

What we get: *highly* non-uniform coverage



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

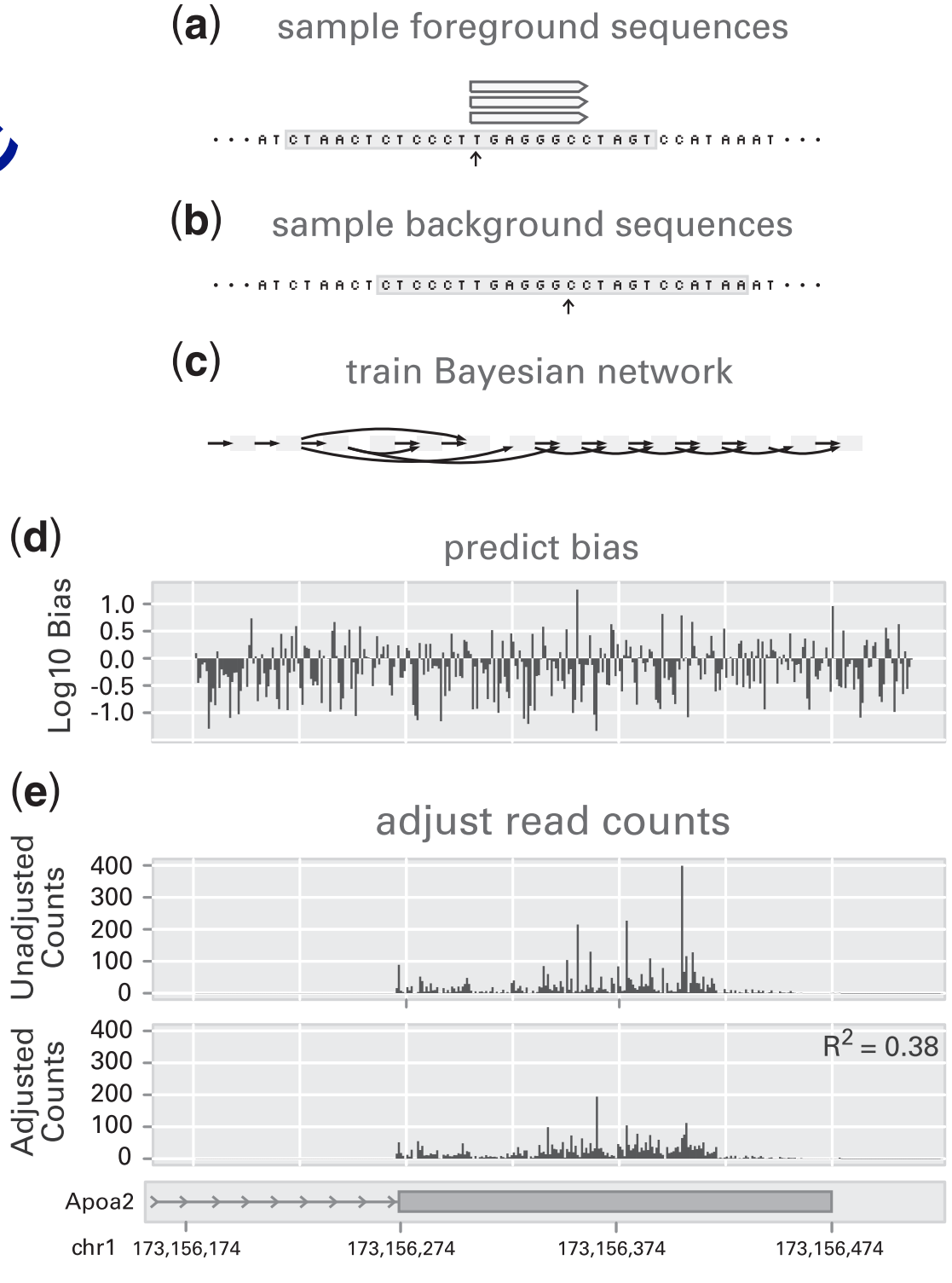
Bias is sequence-dependent



and platform/sample-dependent

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

Method Outline



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

Some obvious choices

Full joint distribution: 4^k-1 parameters

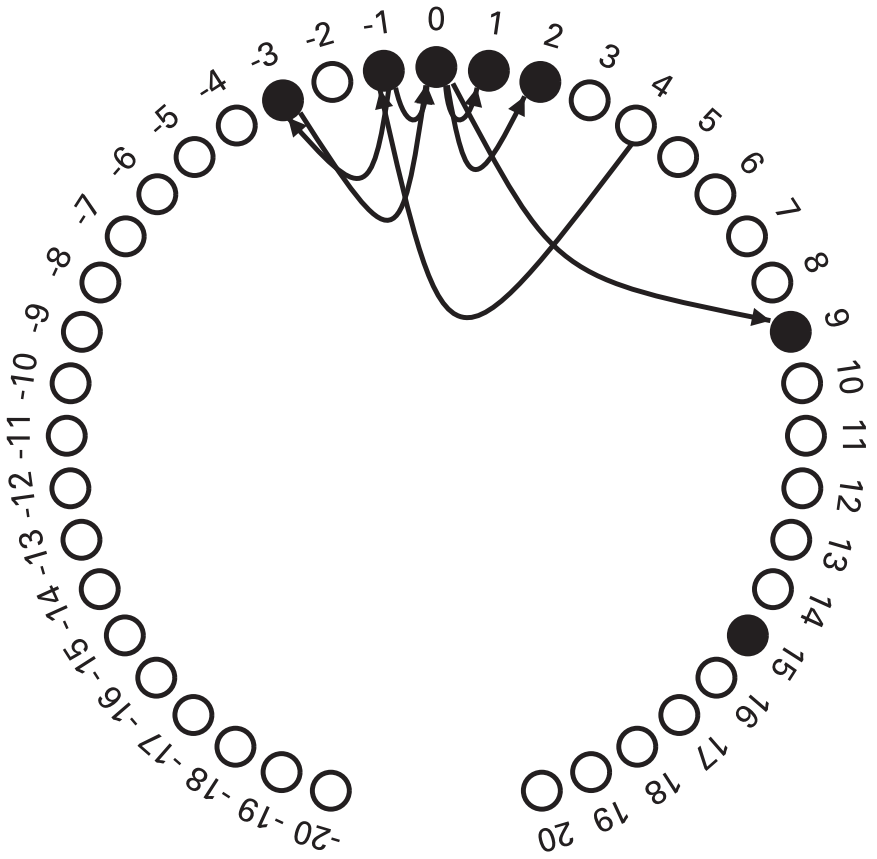
PWM (0-th order Markov): (4-1)•k parameters

Something intermediate

Directed Bayes network

Form of the models:

Directed Bayes nets



Wetterbom (282 parameters)

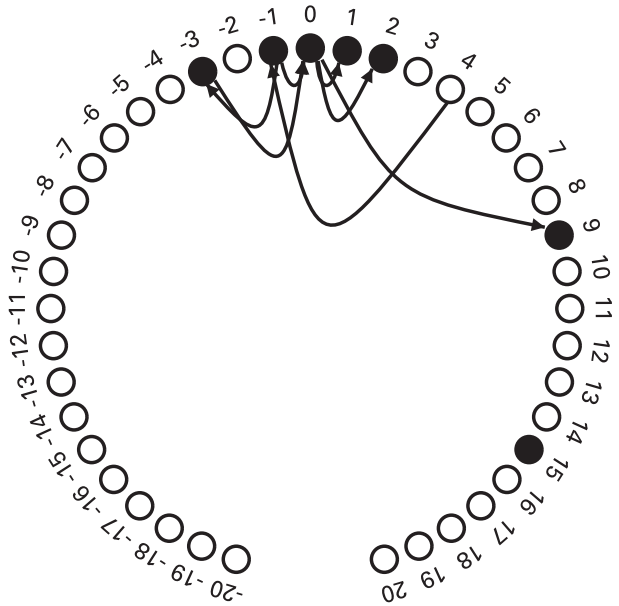
One "node" per nucleotide, ±20 bp of read start

- 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

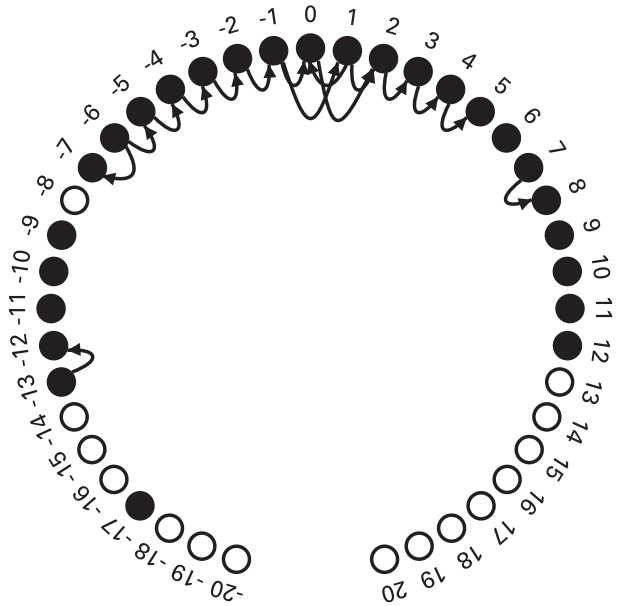
How–optimize:

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

ABI



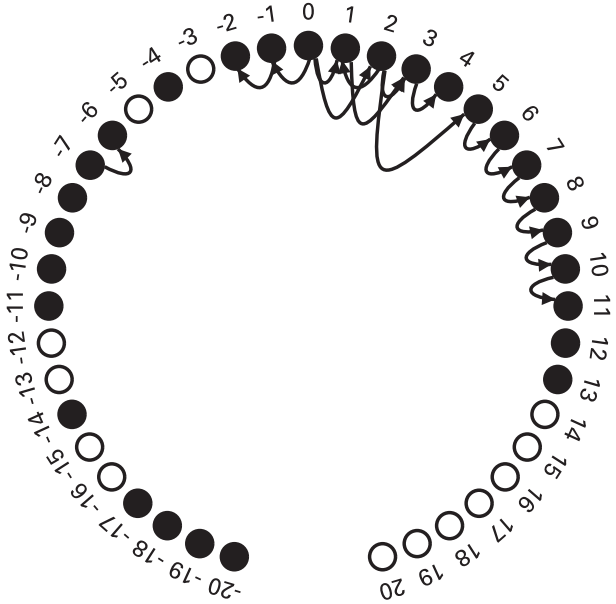
Wetterbom
(282 parameters)



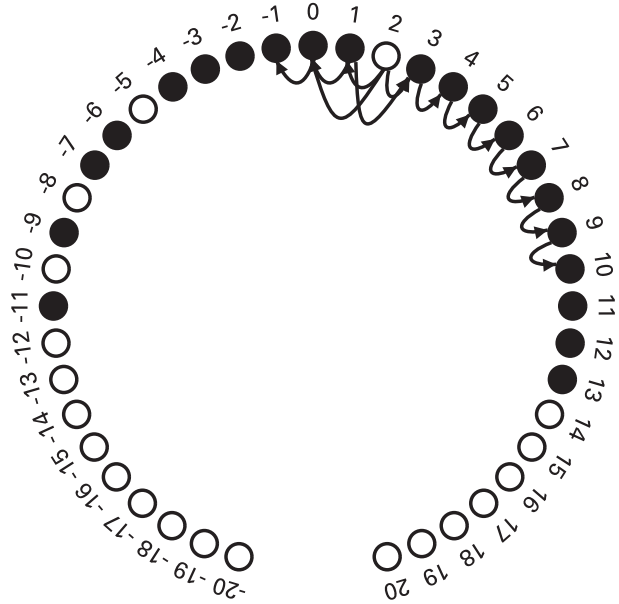
Katze
(684 parameters)

NB:
- Not just initial hexamer
- Span ≥ 19
- All include negative positions

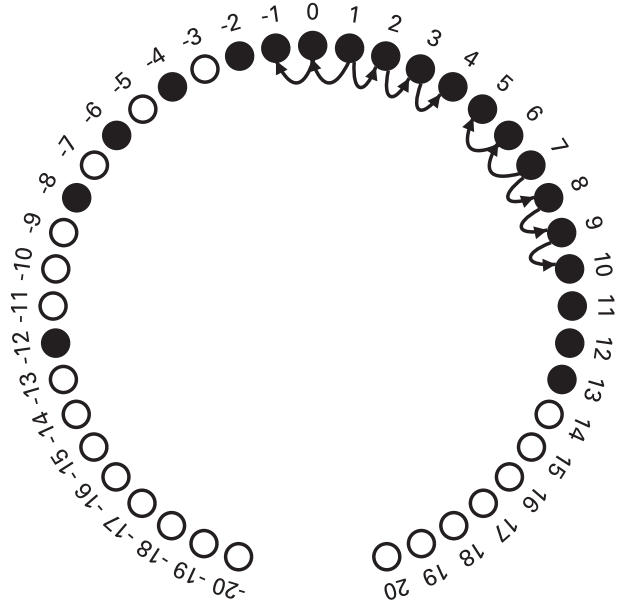
Illumina



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

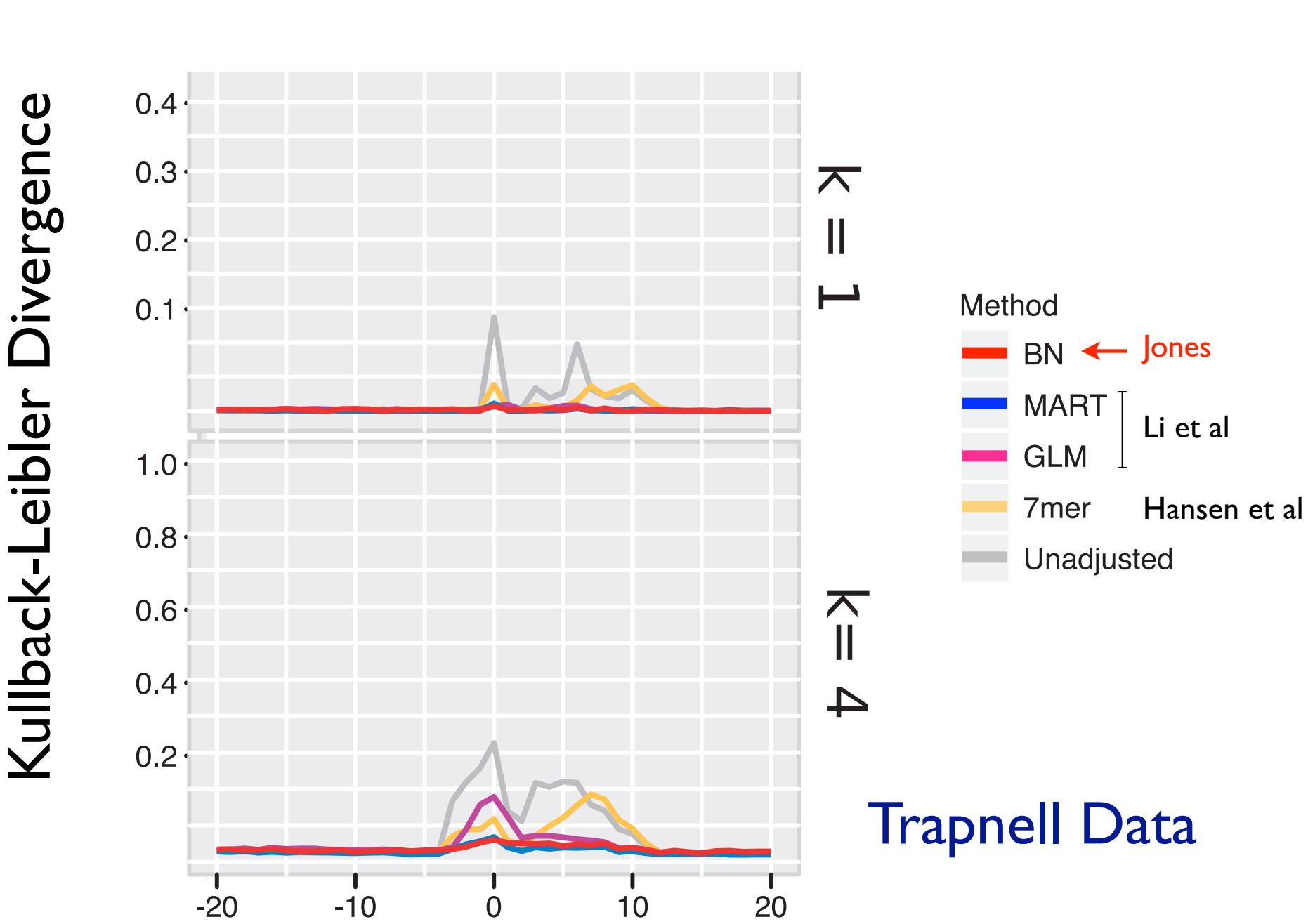
Find a graph that maximizes conditional log-likelihood.

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

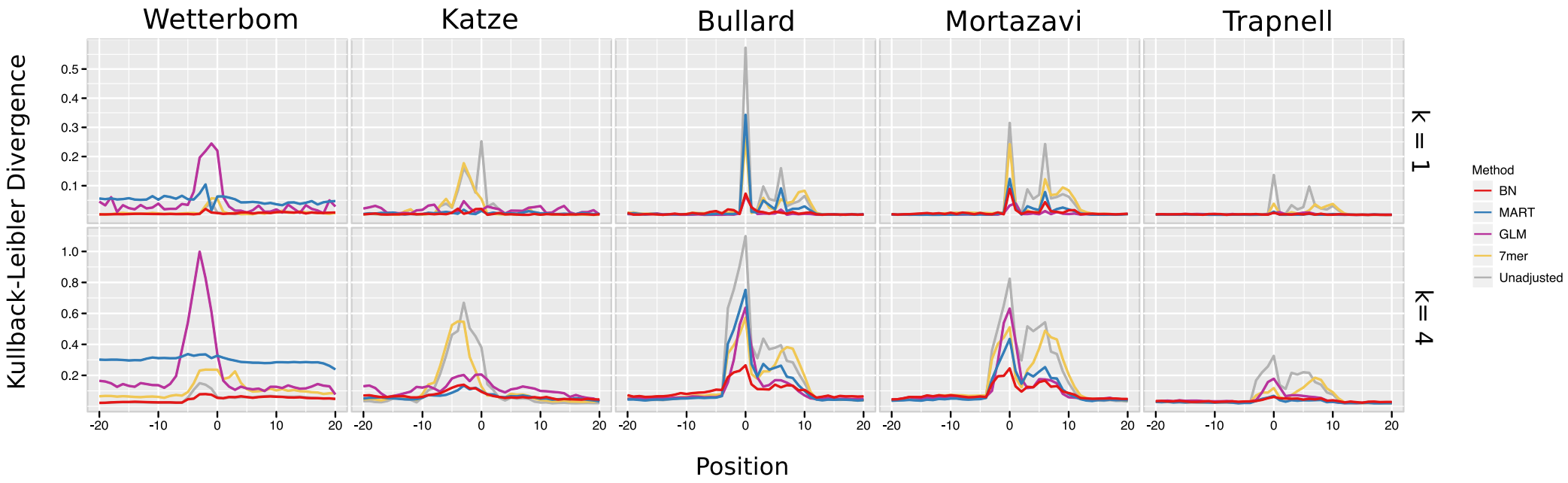
We need to penalize for model complexity as well.

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

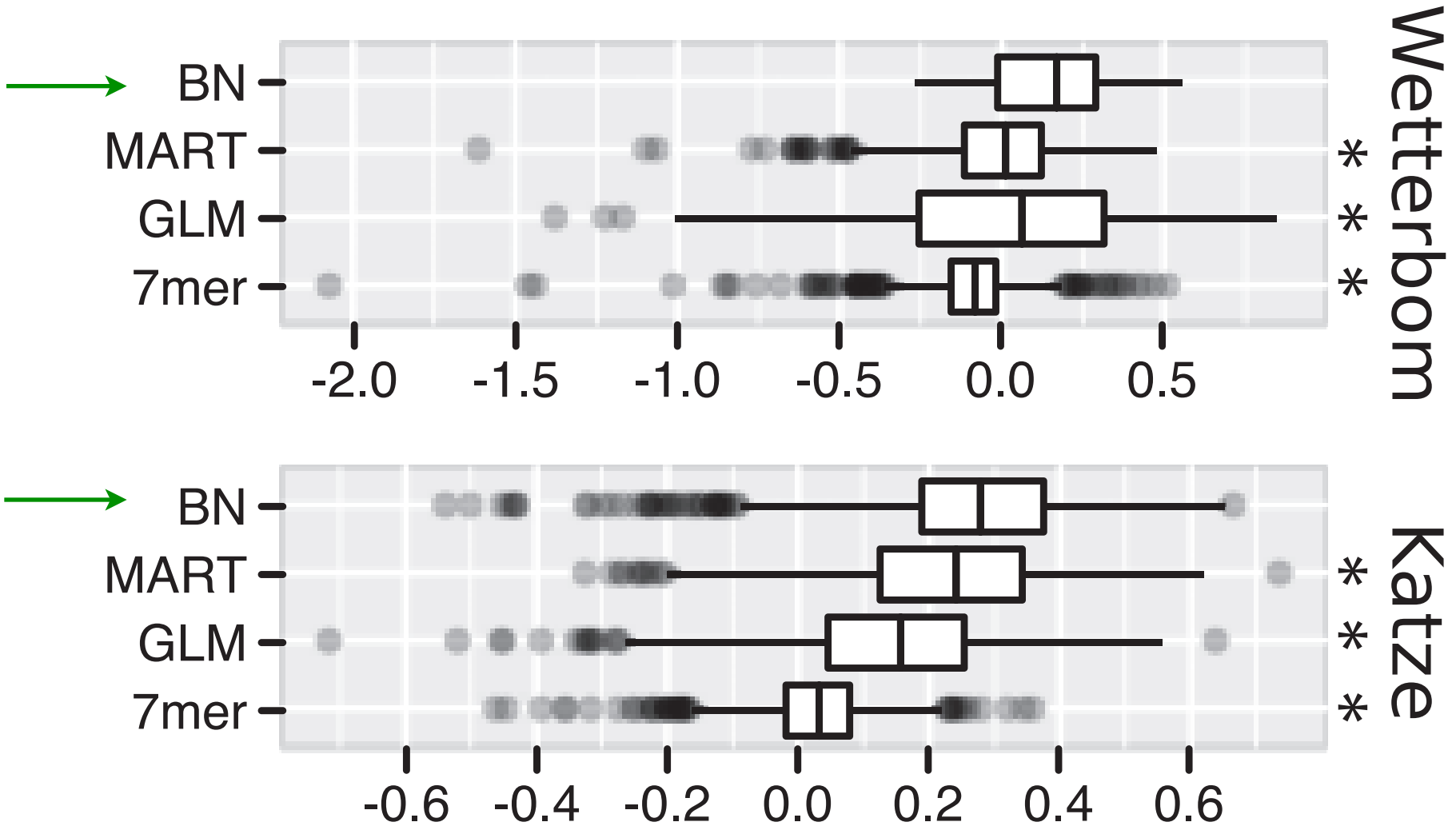
Result – Increased Uniformity



Trapnell Data


Wetterbom
Katze
Bullard
Mortazavi
Trapnell
Kullback-Leibler Divergence
k = 1
k = 4
Position
Method
BN
MART
GLM
7mer
Unadjusted

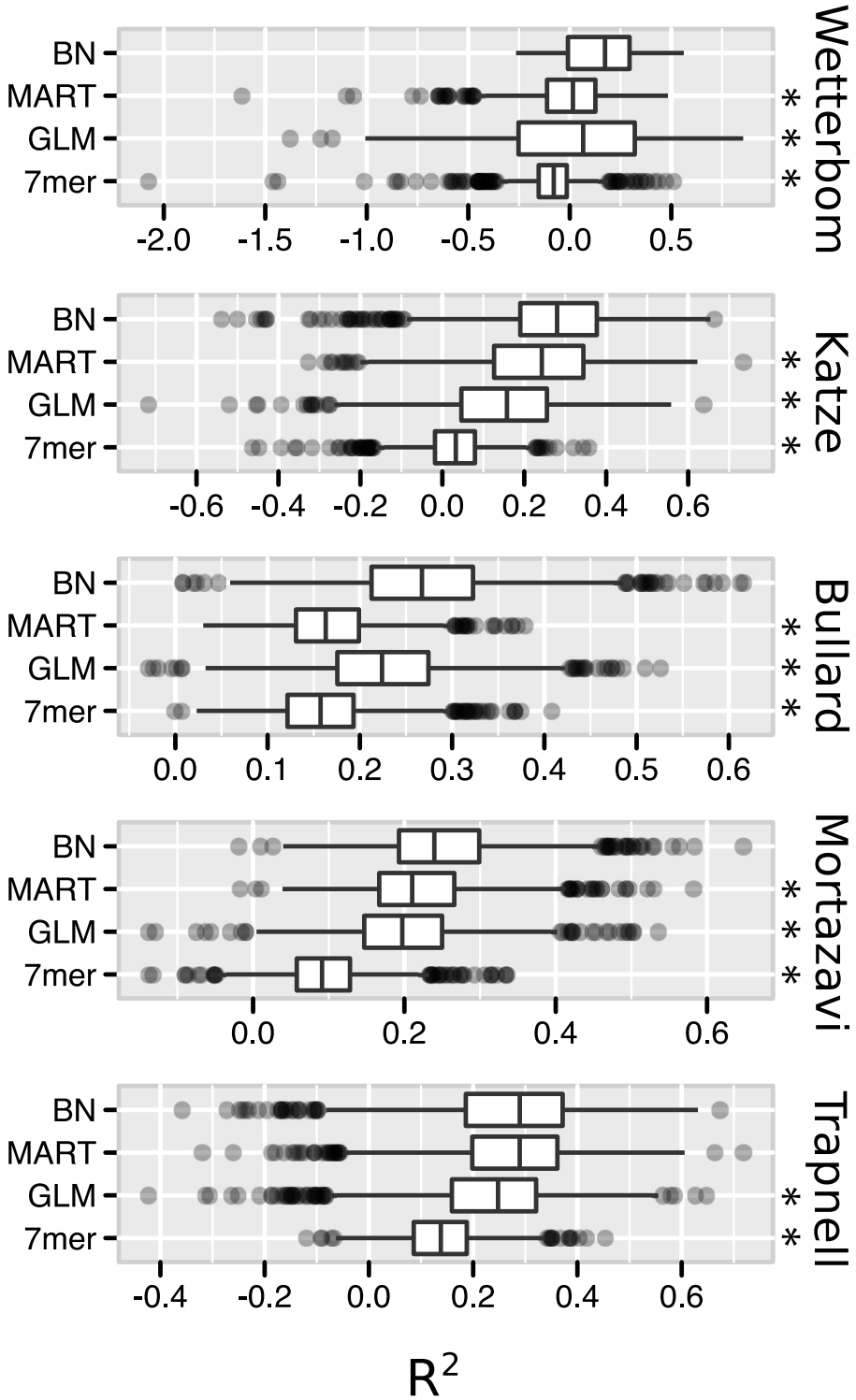
Result – Increased Uniformity



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
Katze
Bullard
Mortazavi
Trapnell
BN
MART
GLM
7mer
*
-2.0
-1.5
-1.0
-0.5
0.0
0.5
-0.6
-0.4
-0.2
0.2
0.4
0.6
0.1
0.3
R2

"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. ($E[X] = 3$)

how different are two distributions?

Given: r-sided die, with probs $p_1 \ldots p_r$ 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

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

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

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

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

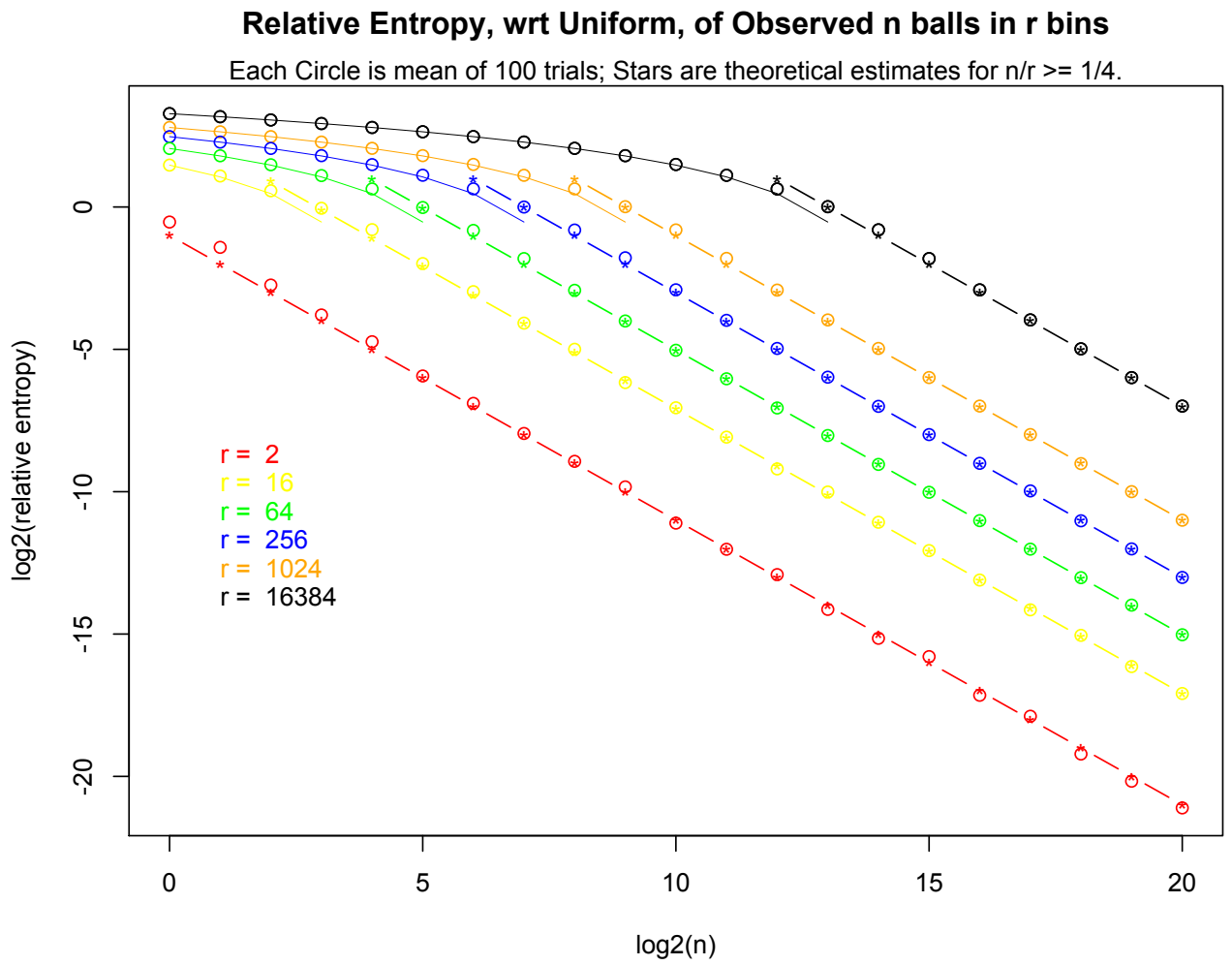
$$H(Q||P) \approx \frac{1}{2n} \sum_{i=1}^{r} \frac{(nq_i - np_i)^2}{np_i}$$
$$= \frac{1}{2n} \sum_{i=1}^{r} \frac{(X_i - E[X_i])^2}{E[X_i]}$$

… and after a modicum of algebra:

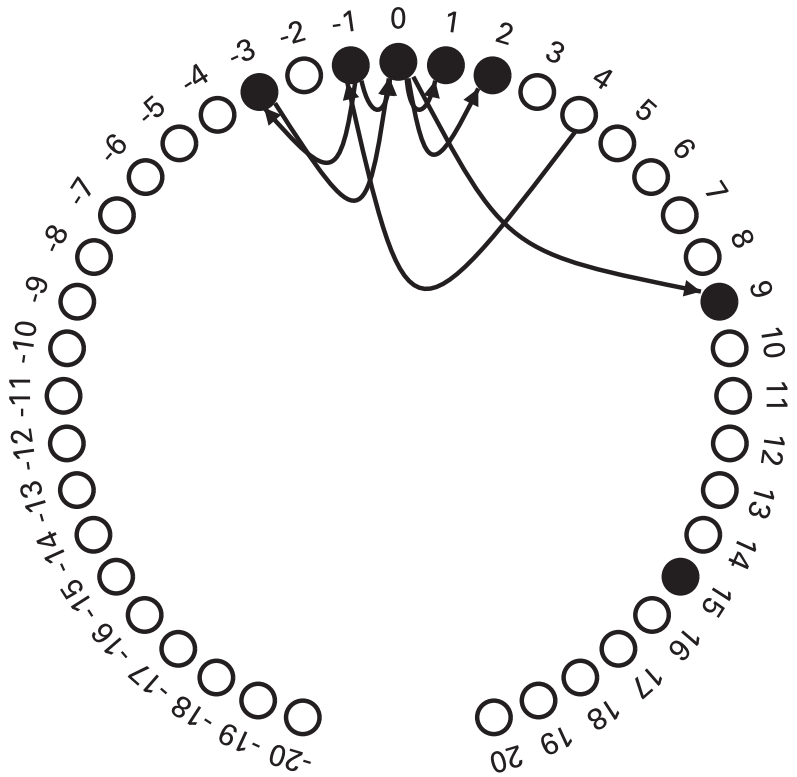
$$E[H(Q||P)] \approx \frac{r-1}{2n}$$

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

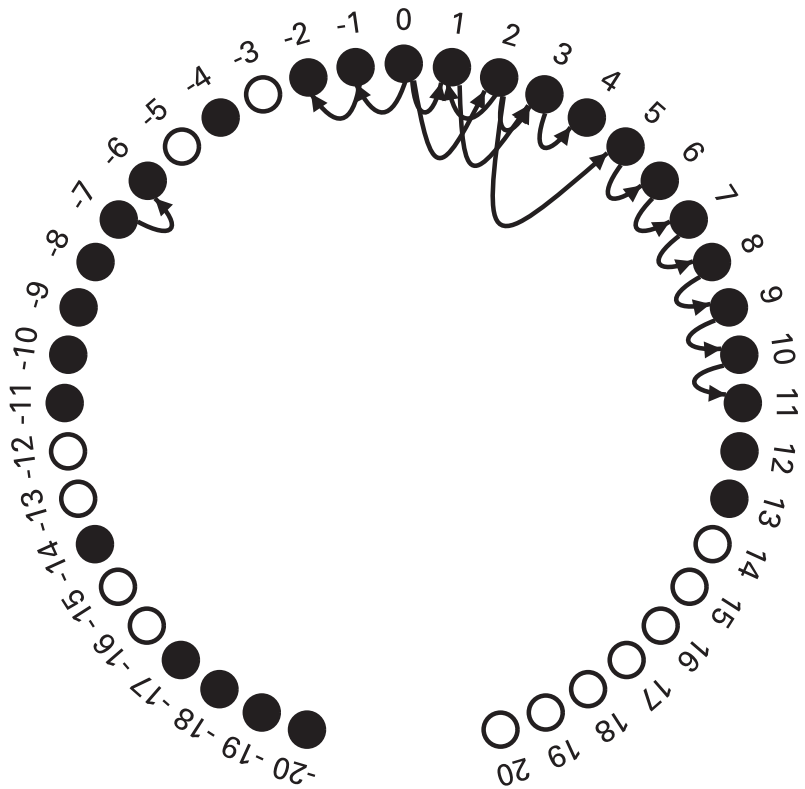
… which empirically is a good approximation:



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) ...

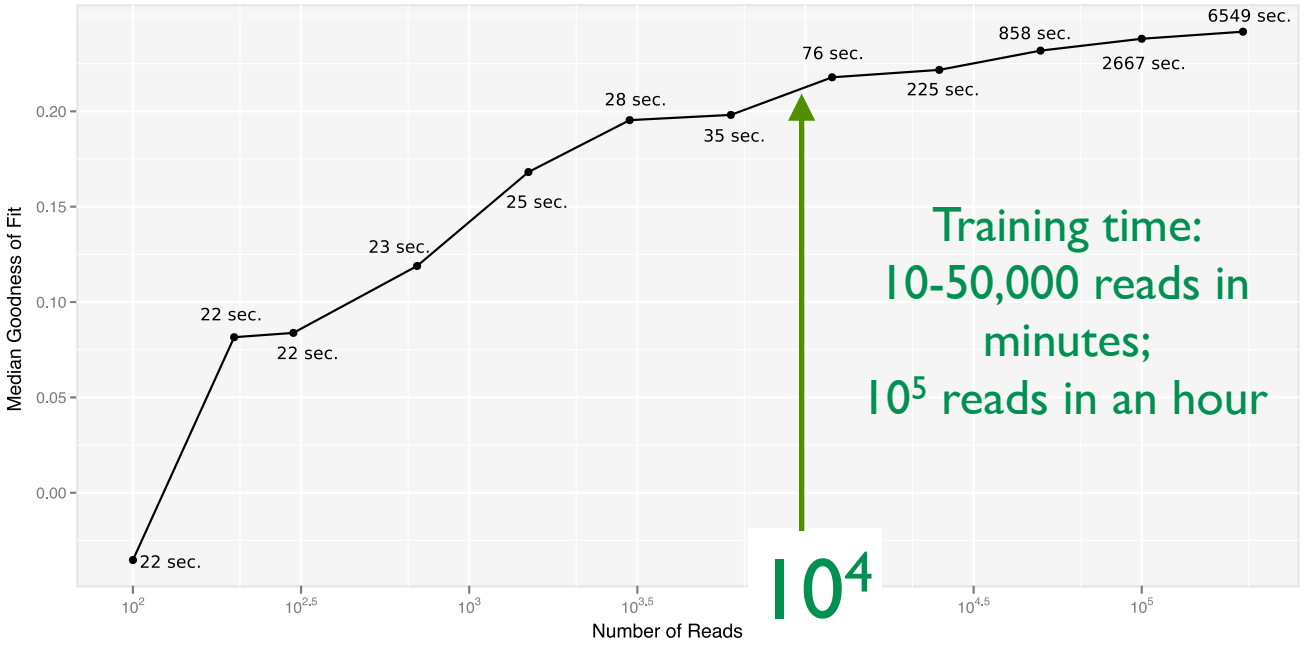
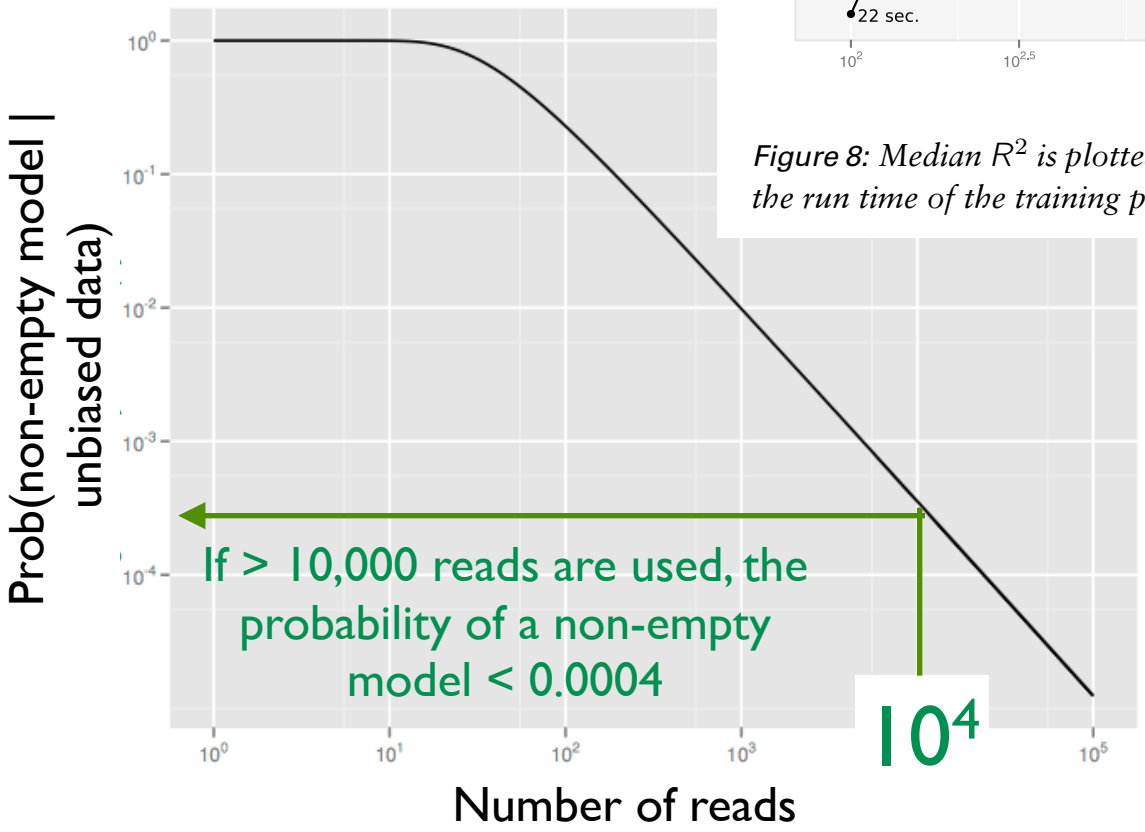


Figure 8: Median R^2 is plotted against training set size. Each point is additionally labeled with the run time of the training procedure.



… 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:

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

SNPs and/or alternative splicing might have a big effect, if samples 1 & 2 are from different individuals and/or engender changes in isoform usage

Some experiments are *not "typical,"* e.g., imprinting, allele specific expression, xenograft studies

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

In Progress: Isolator

Soon to be the world's best isoform quantitation tool

Availability

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home | Install | Help

Home » Bioconductor 2.12 » Software Packages » seqbias

seqbias

Estimation of per-position bias in high-throughput sequencing data

Bioconductor version: Release (2.12)

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

Author: Daniel Jones <dcjones at cs.washington.e

Maintainer: Daniel Jones <dcjones at cs.washingt

To install this package, start R and enter:

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

To cite this package in a public

```
citation("
```

Docum

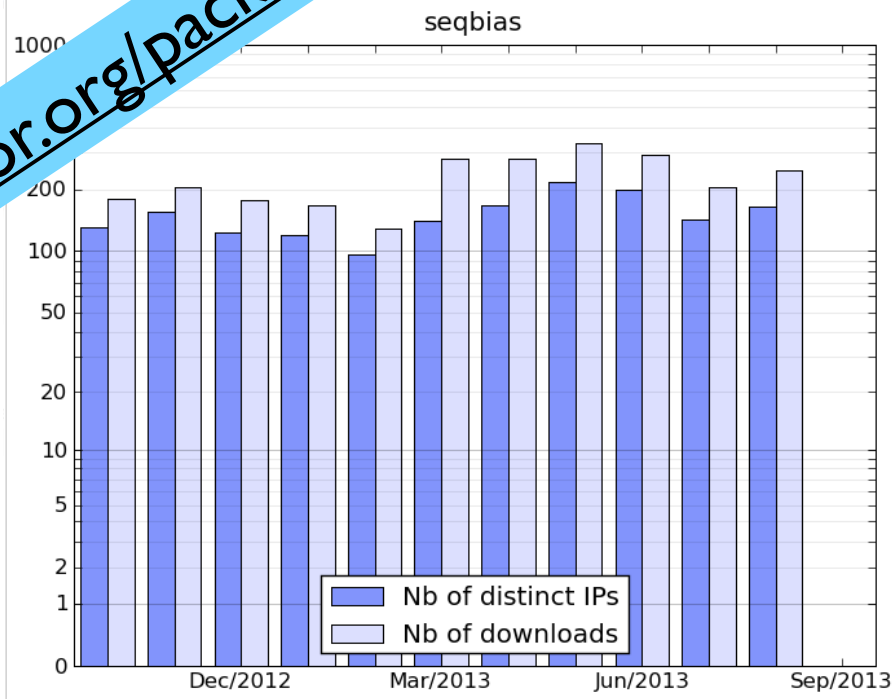
Assessing and Adjusting for
Reference Manual

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

Download stats for Software package seqbias

This page was generated on 2013-09-02 07:28:58 -0700 (Mon, 02 Sep 2013).

seqbias home page: release version, devel version.



Month	Nb of distinct IPs	Nb of downloads
Oct/2012	131	180
Nov/2012	156	204
Dec/2012	123	176
Jan/2013	119	168
Feb/2013	96	129
Mar/2013	140	282
Apr/2013	167	280
May/2013	217	333
Jun/2013	200	293
Jul/2013	142	205
Aug/2013	164	248
Sep/2013	0	0
All months	**1321**	**2498**