

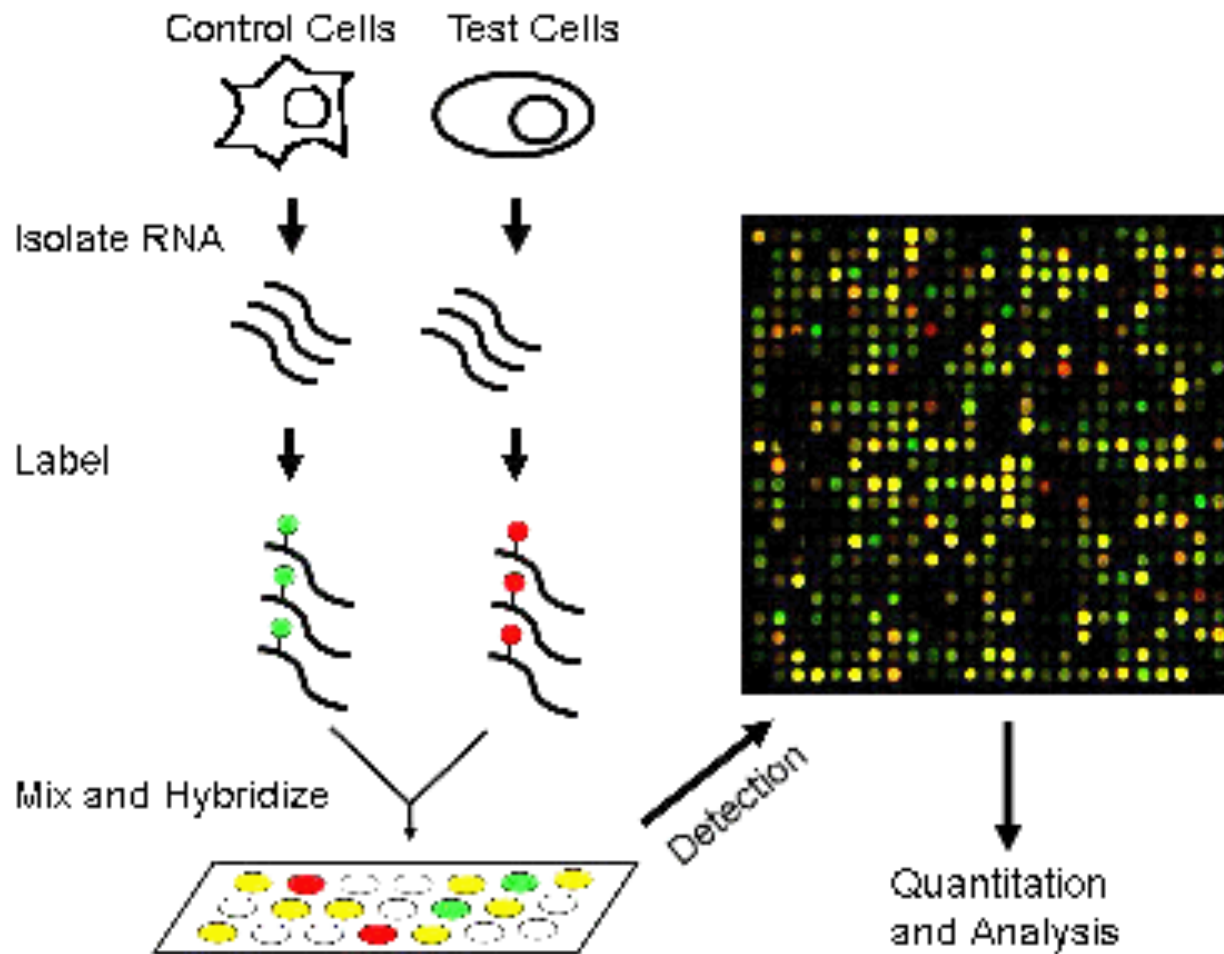
CSEP 527

Computational Biology

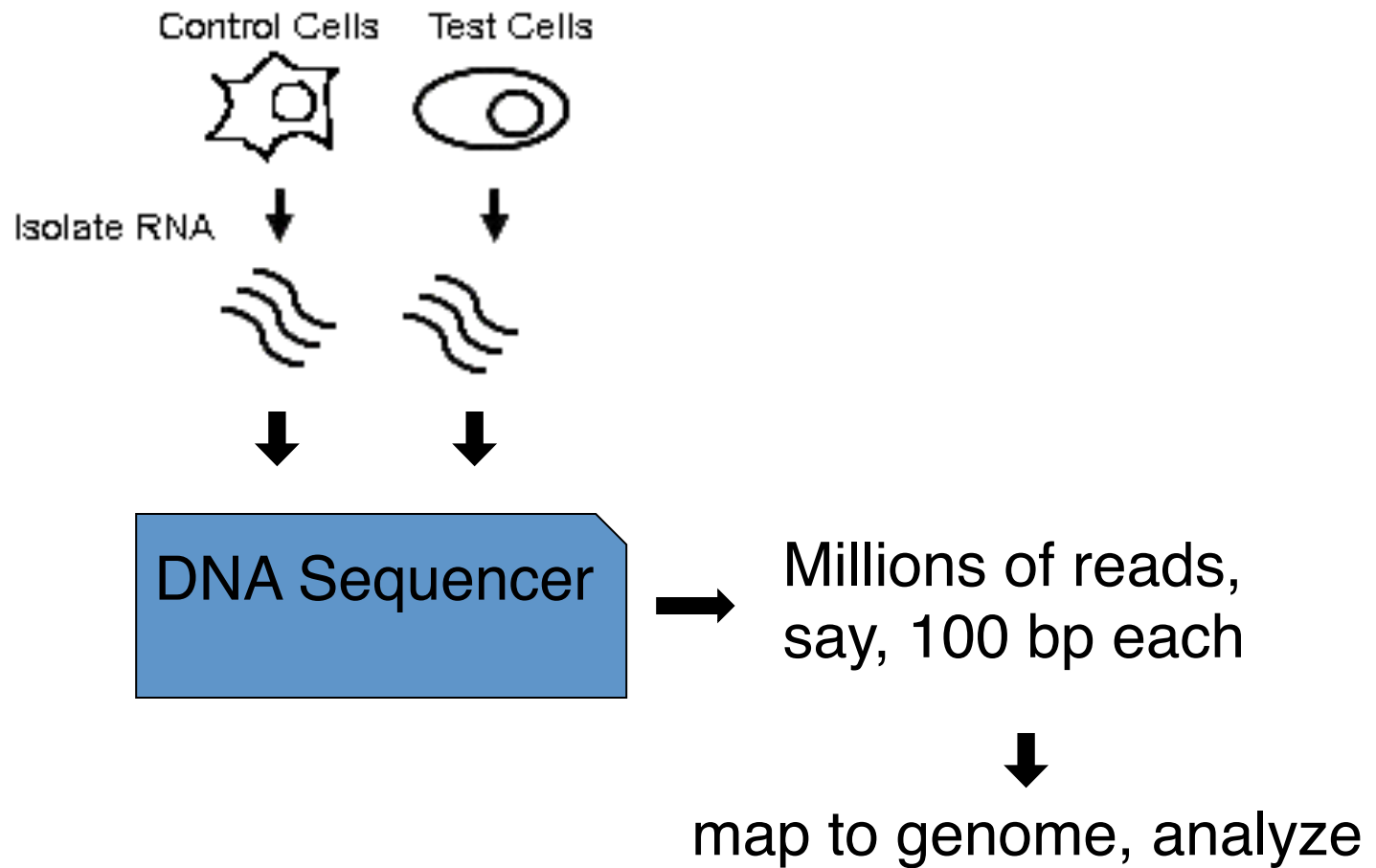
Gene Expression Analysis

Assaying Gene Expression

Microarrays



RNAseq



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?

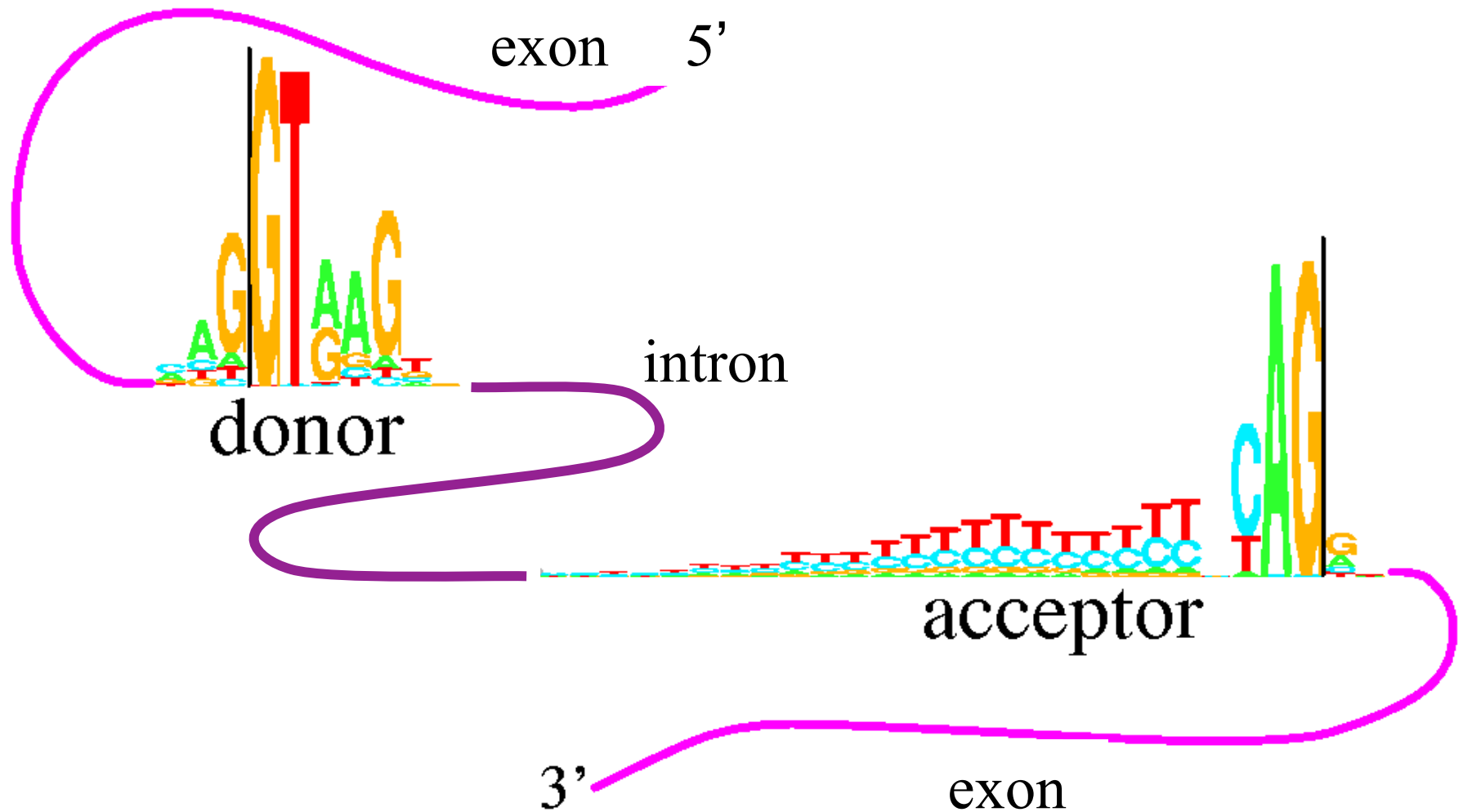
How? *count* how many fragments come from each gene—expect more highly expressed genes to yield more reads, after correcting for biases like mRNA length

#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 = 25_{mers} anchored 2 sides
- stitch original reads across these
- remap reads with minimal overlaps
- *Roughly*: 10m reads/hr, 4Gbytes
(typical data set 100m–1b reads)

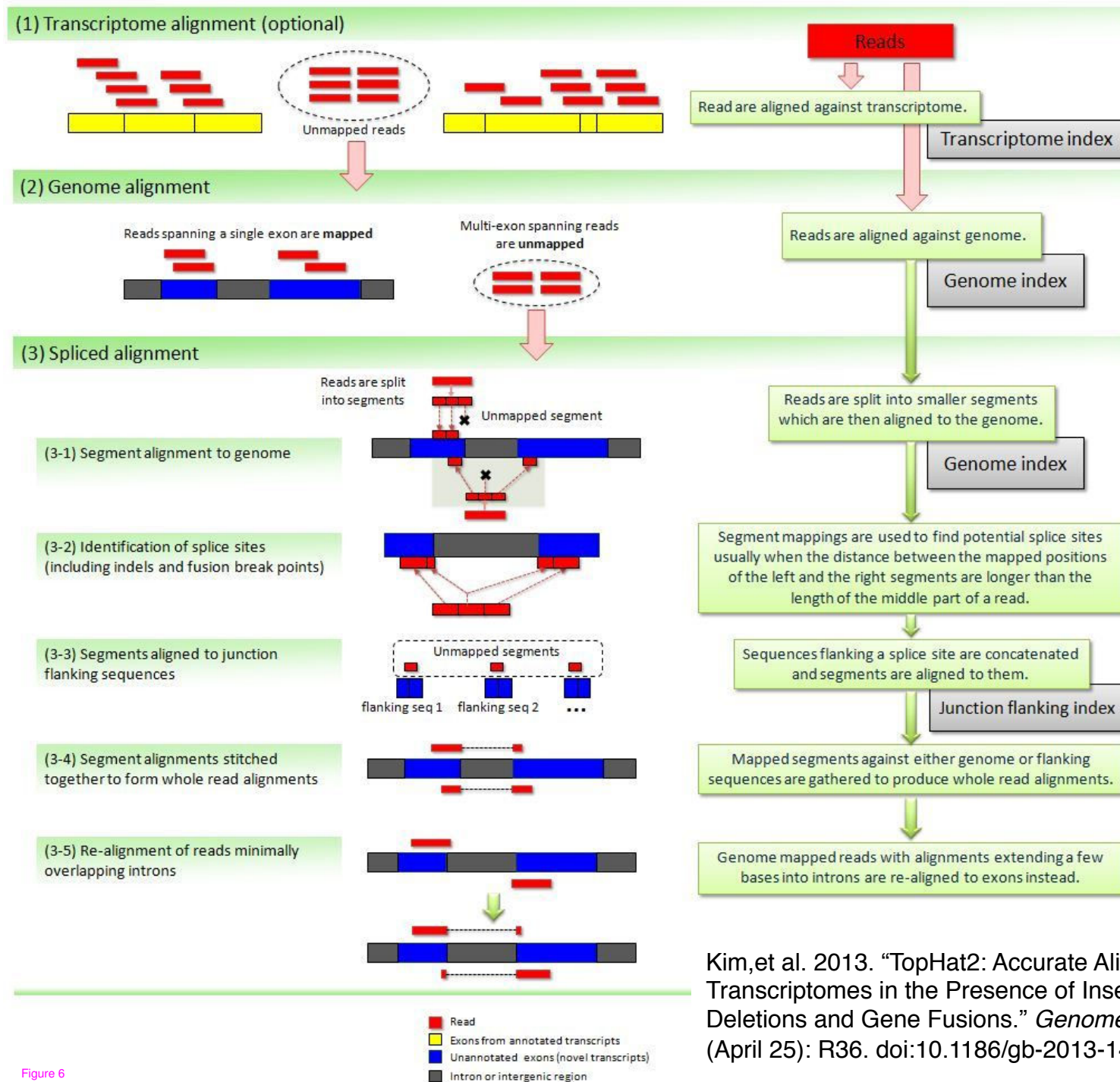


Figure 6

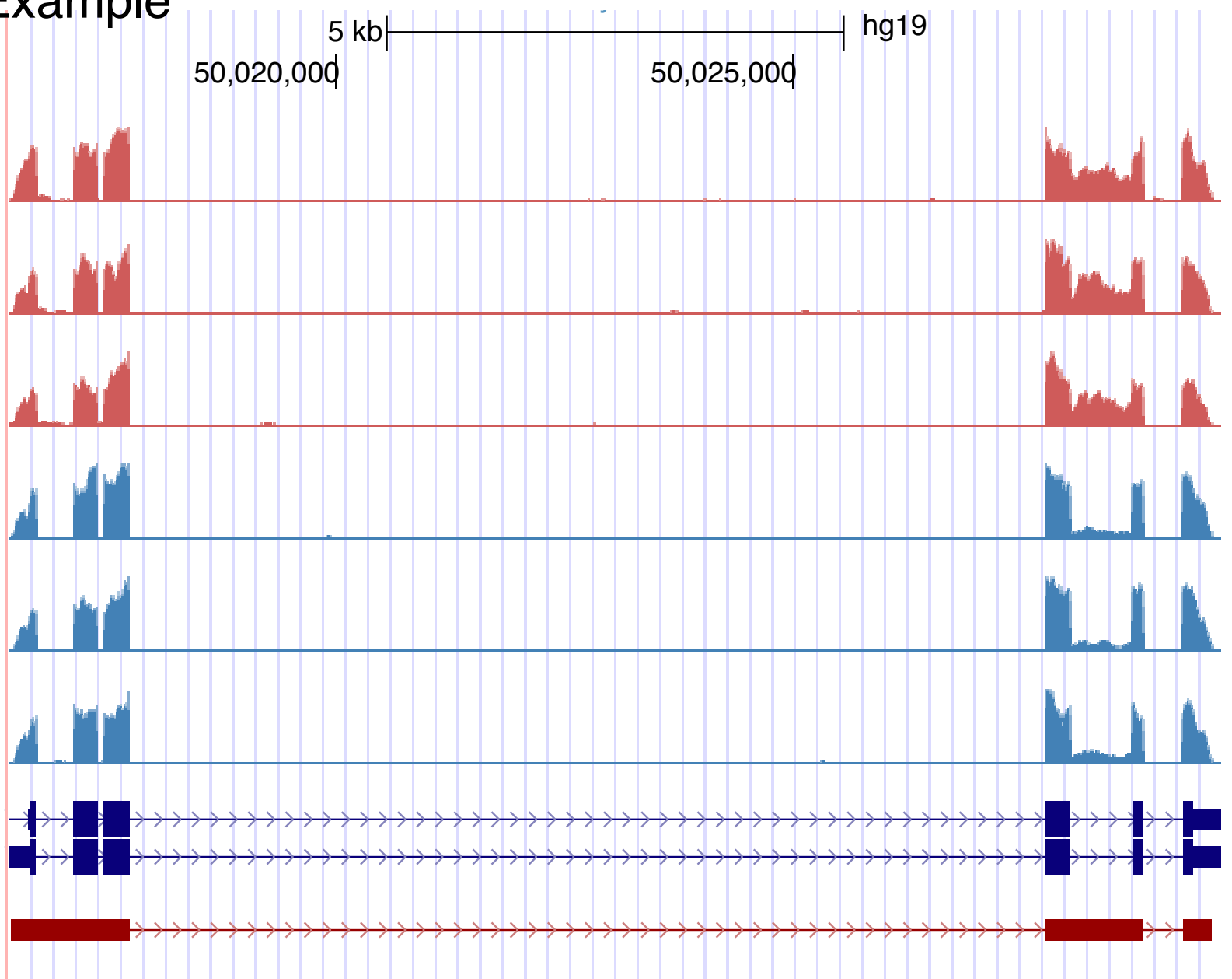
Kim, et al. 2013. "TopHat2: Accurate Alignment of Transcriptomes in the Presence of Insertions, Deletions and Gene Fusions." *Genome Biology* 14 (4) (April 25): R36. doi:10.1186/gb-2013-14-4-r36.

RNAseq Example

5 kb | hg19
50,020,000 | 50,025,000

Day 20

1 Year



RNAseq protocol (approx)

Extract RNA (either polyA ↔ polyT or tot – rRNA)

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

Two Stories:

- RNAseq Bias Correction & Isoform Quantification
- Let-7 & Cardiomyocyte Maturation

Walter L. (Larry) Ruzzo

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

Story I

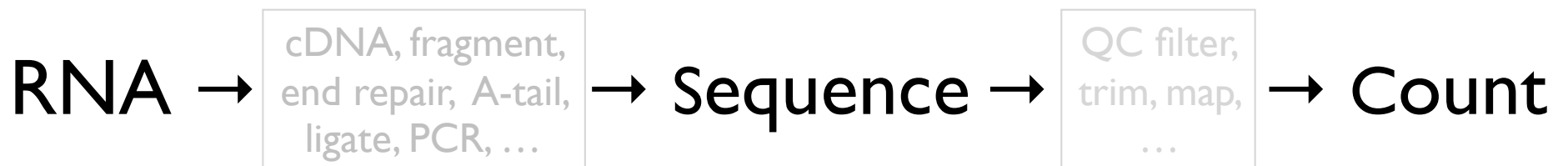
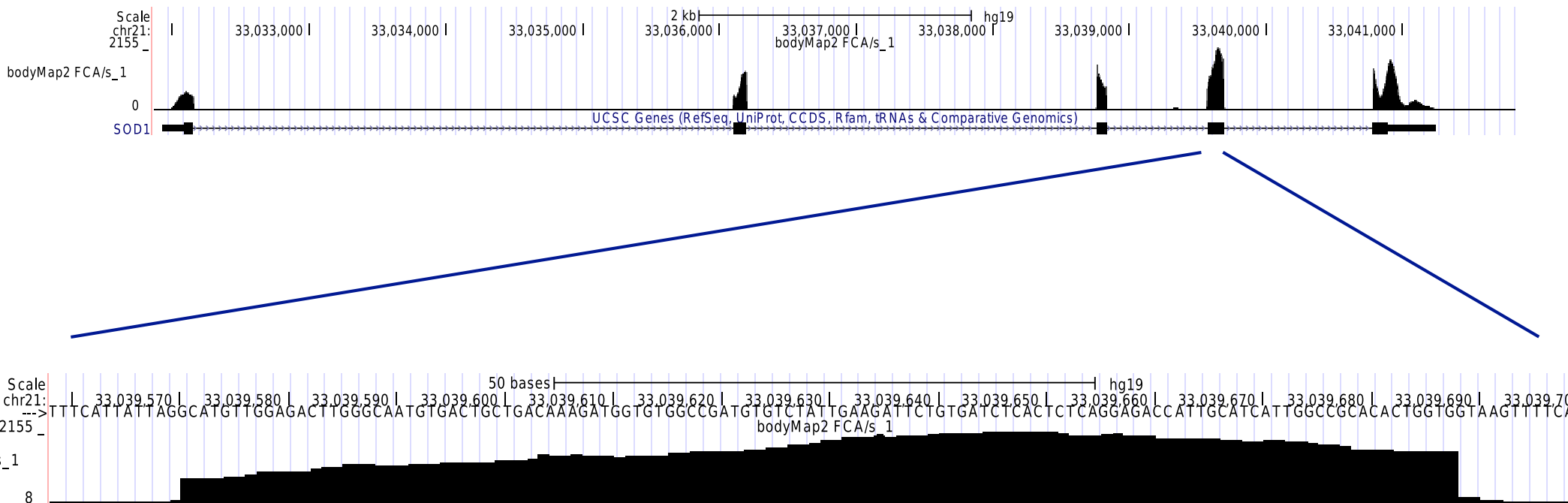
RNAseq:

Bias Correction & Alt Splicing

**“All High-Throughput
Technologies are Crap
– Initially”**

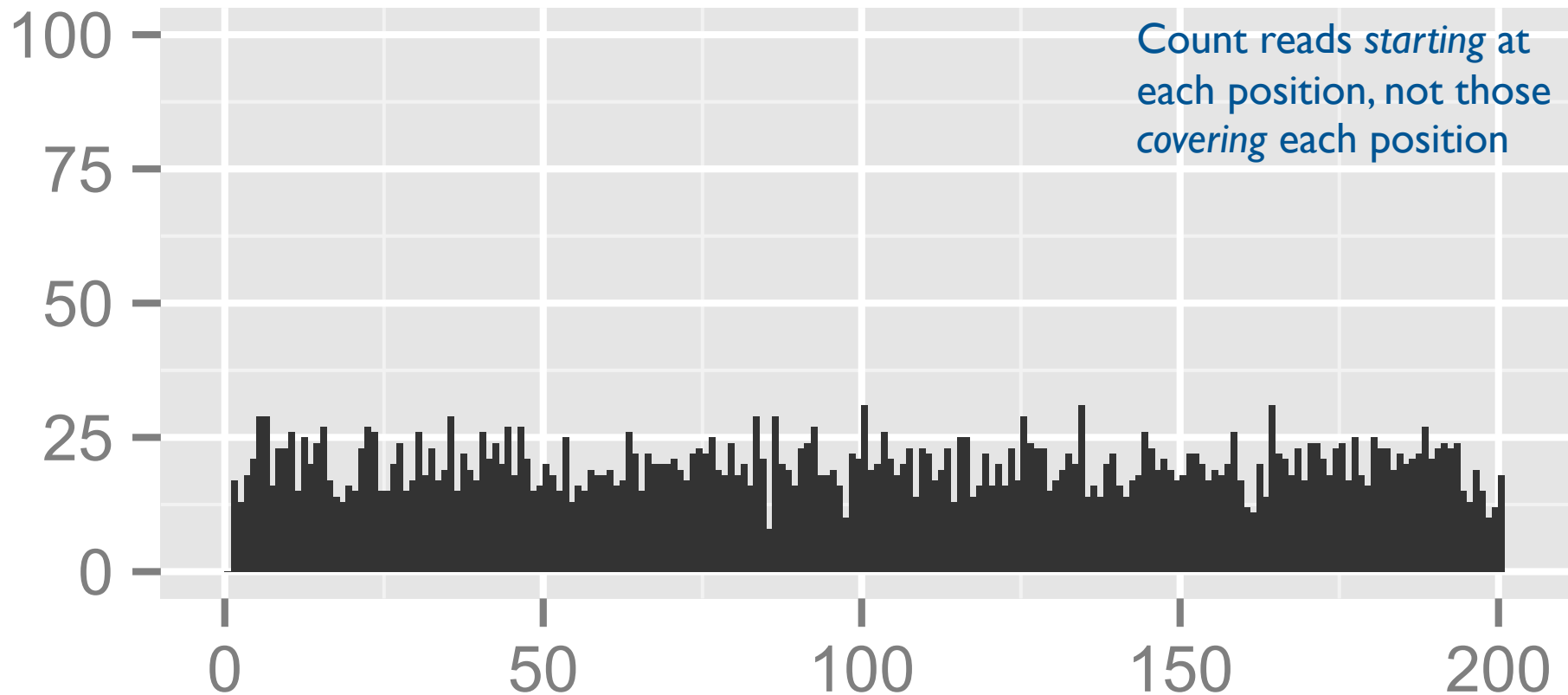
**Q. Morris
7-20-2015**

RNA seq



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

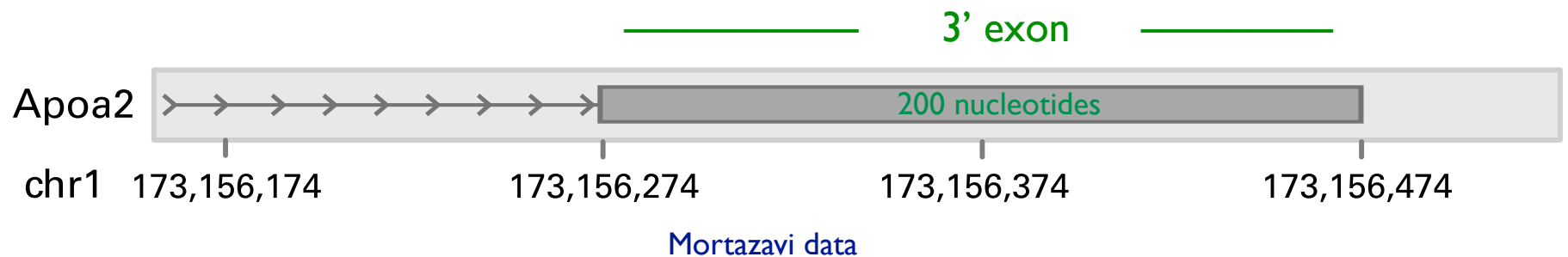
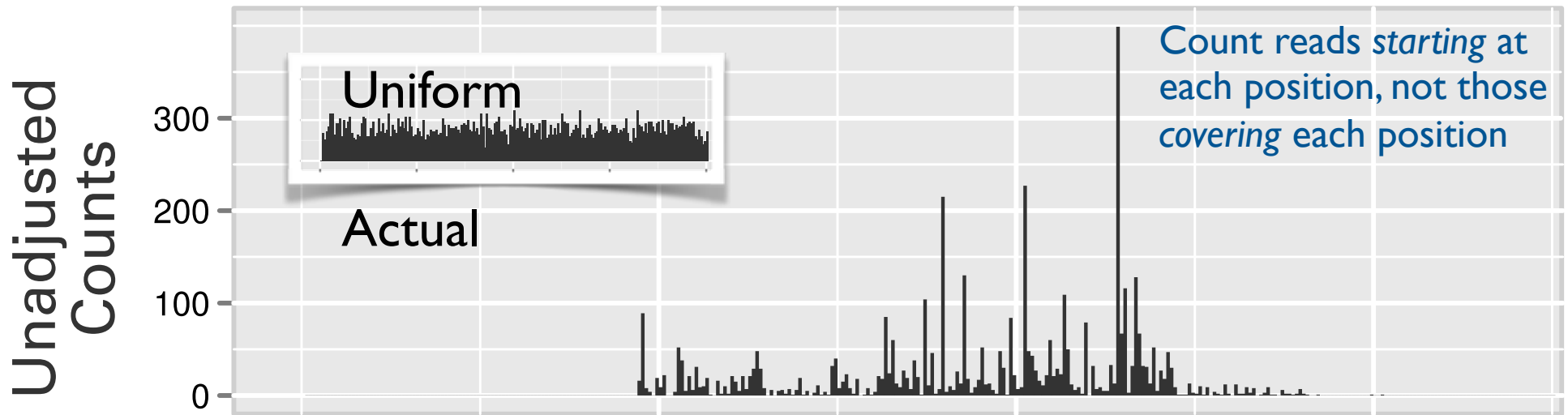
What we expect: Uniform Sampling



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 $\approx \mu \pm 3\sigma$ in 200 samples

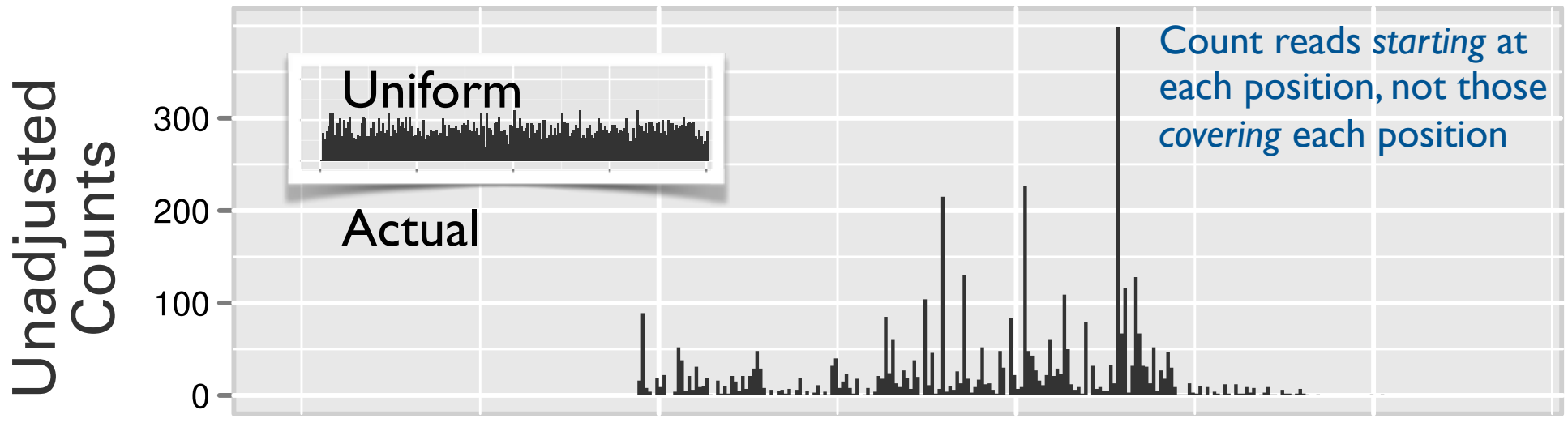
What we get: *highly non-uniform coverage*

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



What we get: *highly non-uniform coverage*

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



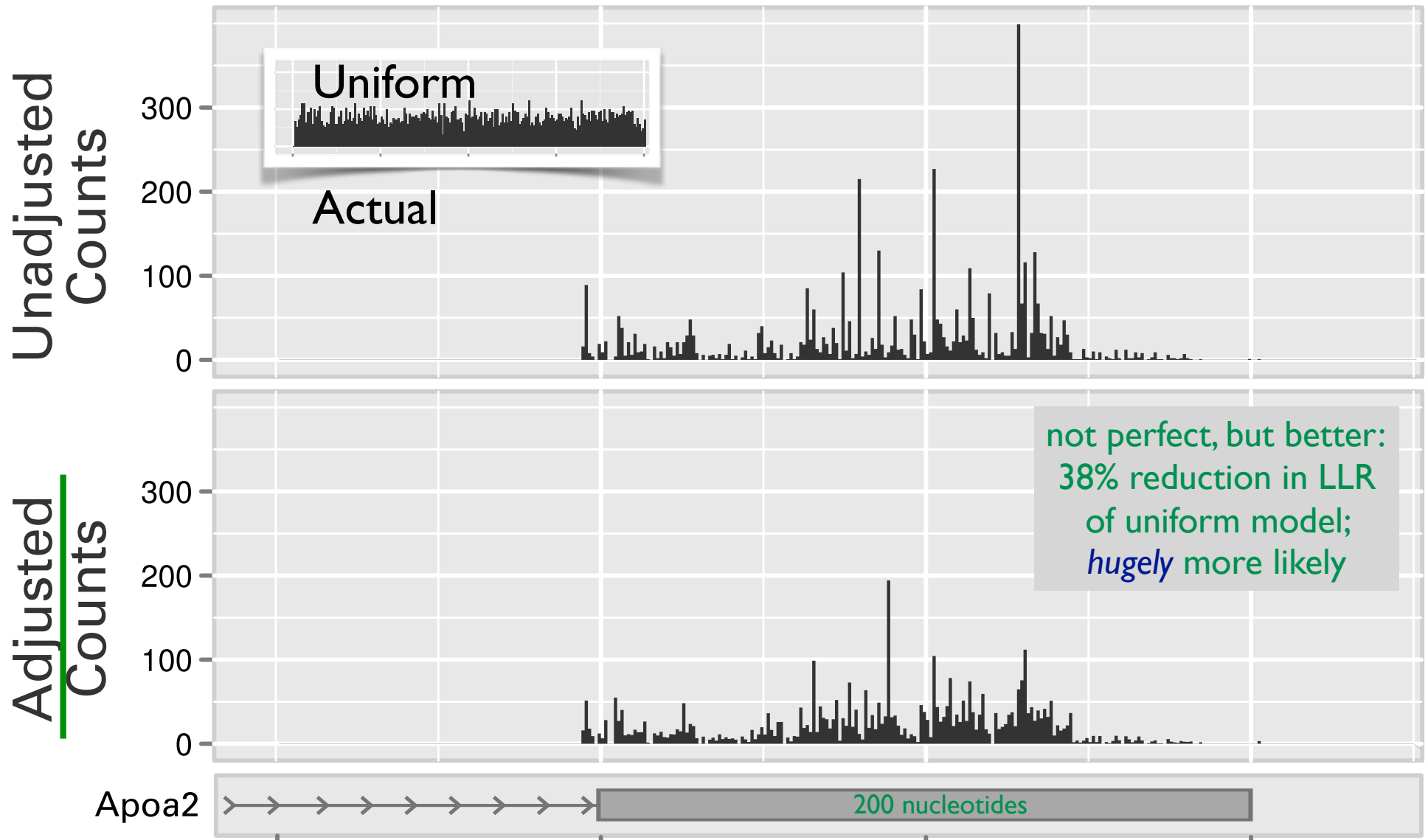
How to make it more uniform?

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

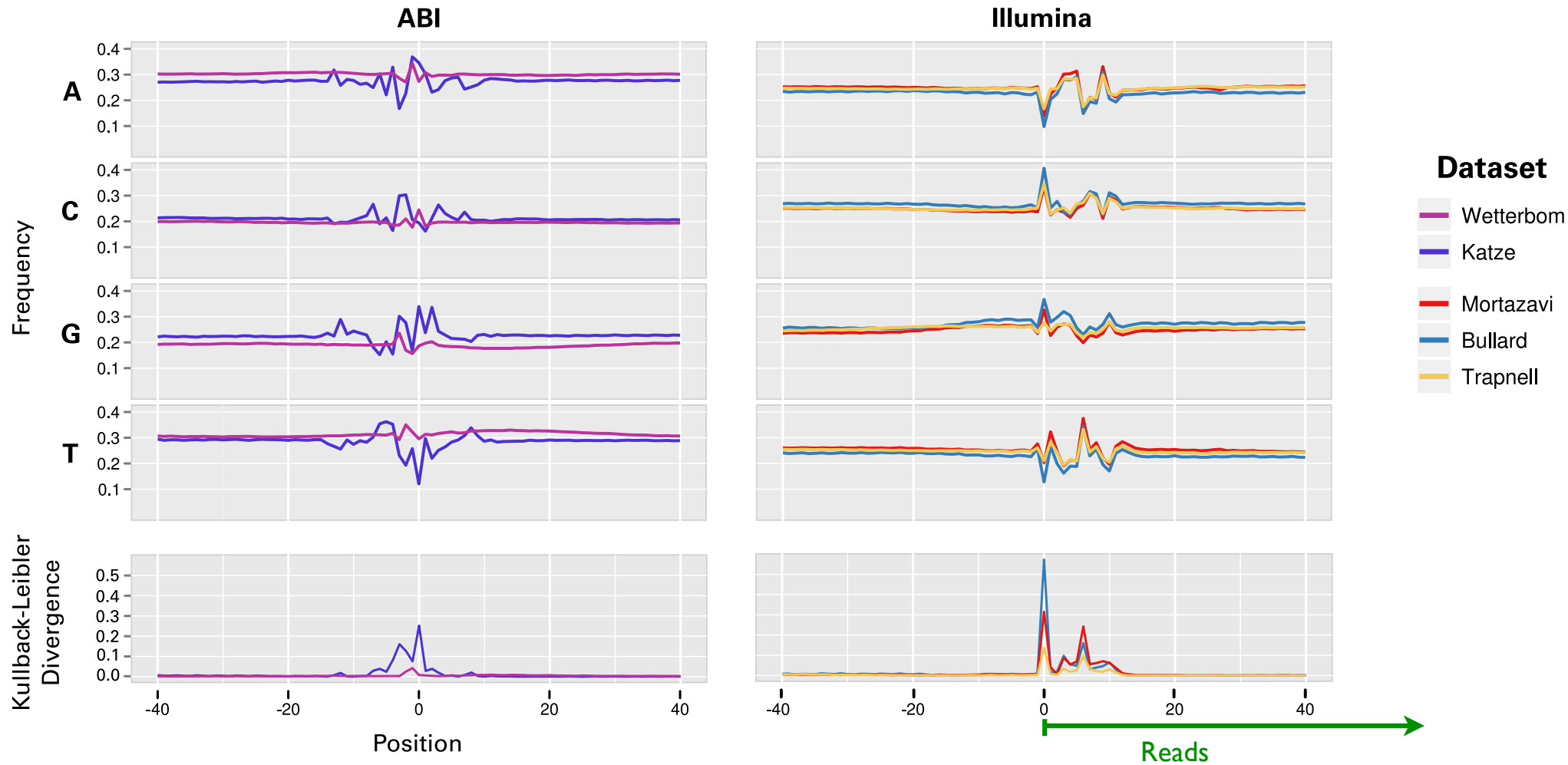
B: Try to model (aspects of) causation

← WE DO THIS

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



(in part) Bias is \wedge sequence-dependent

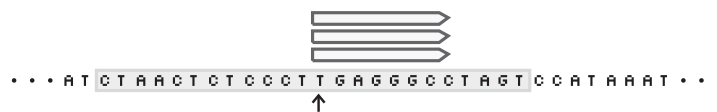


and platform/sample-dependent

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 (local) background sequences



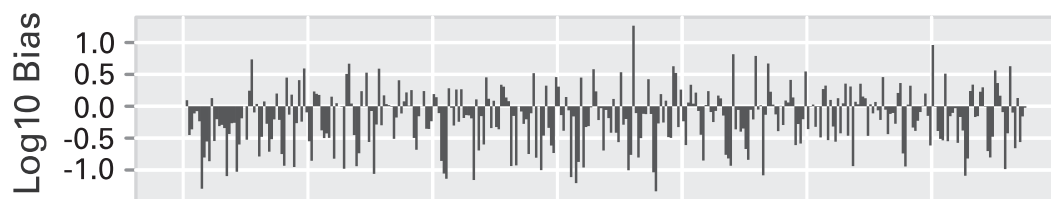
(c) train Bayesian network



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

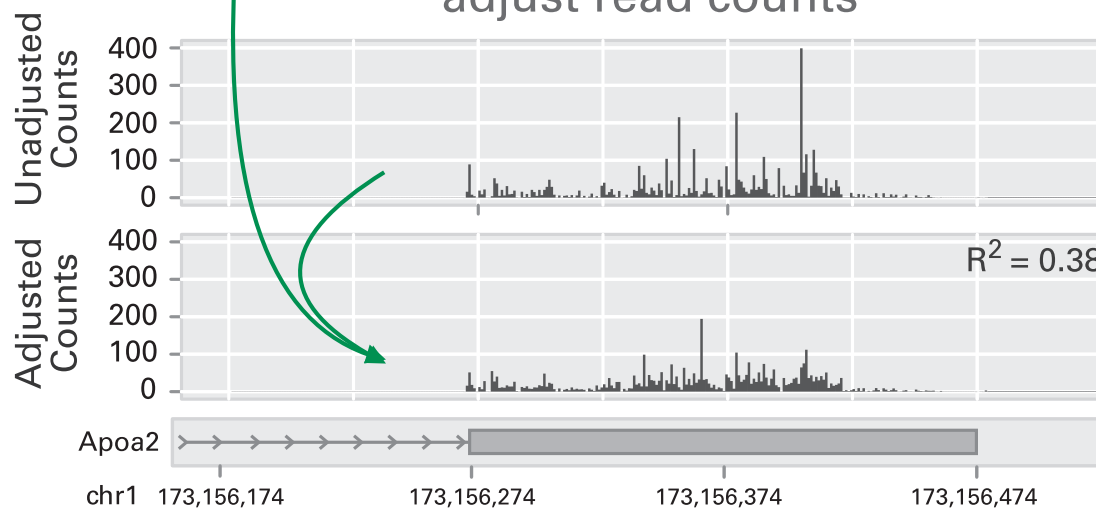
(d)

predict bias



(e)

adjust read counts



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

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

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

Some obvious choices:

Full joint distribution: $4^k - 1$ parameters

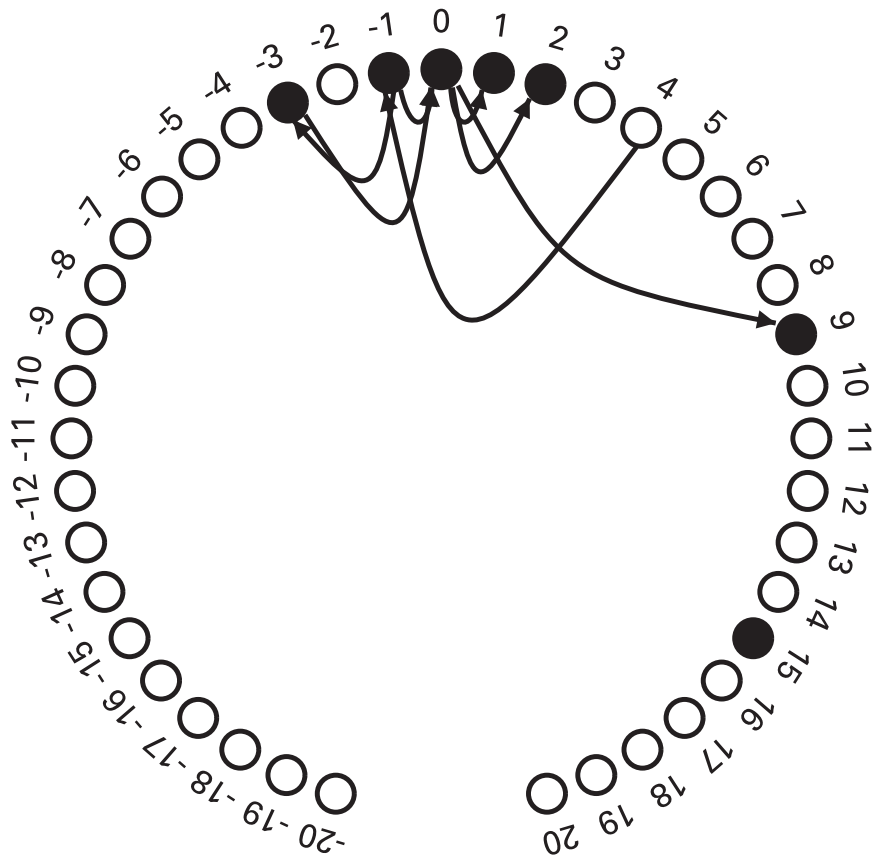
PWM (0-th order Markov): $(4 - 1) \cdot k$ parameters

Something intermediate:

Directed Bayes network

Form of the models:

Directed Bayes nets



Wetterbom
(282 parameters)

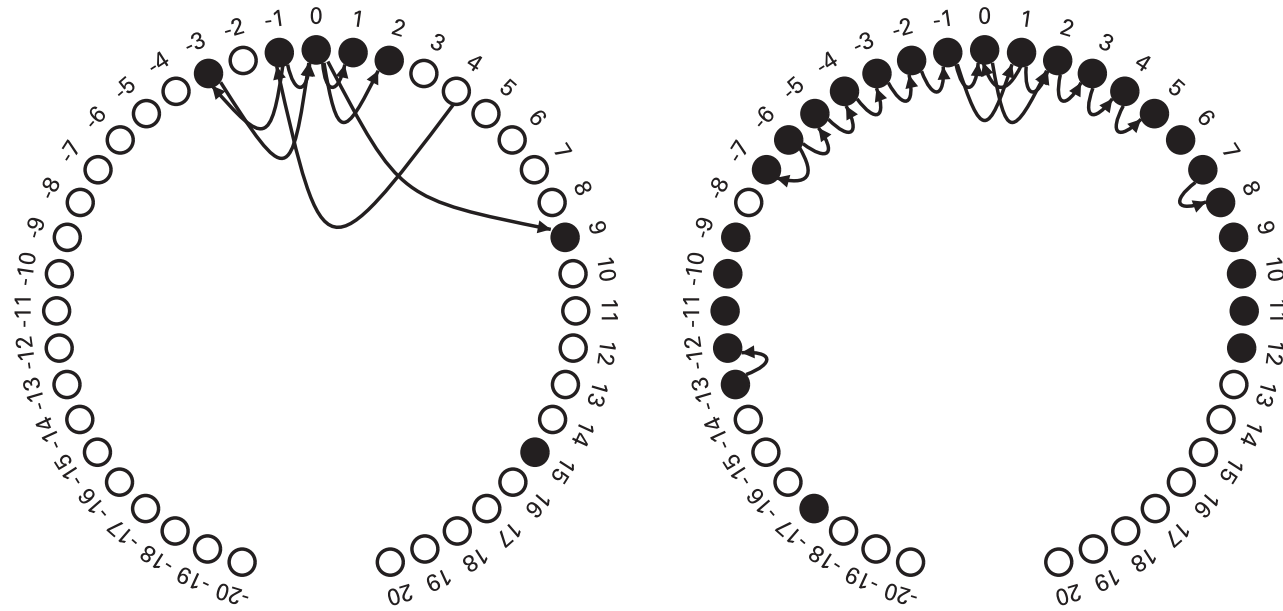
One “node” per nucleotide,
 ± 20 bp of read start

- Filled node means that position is biased
- Arrow $i \rightarrow j$ 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

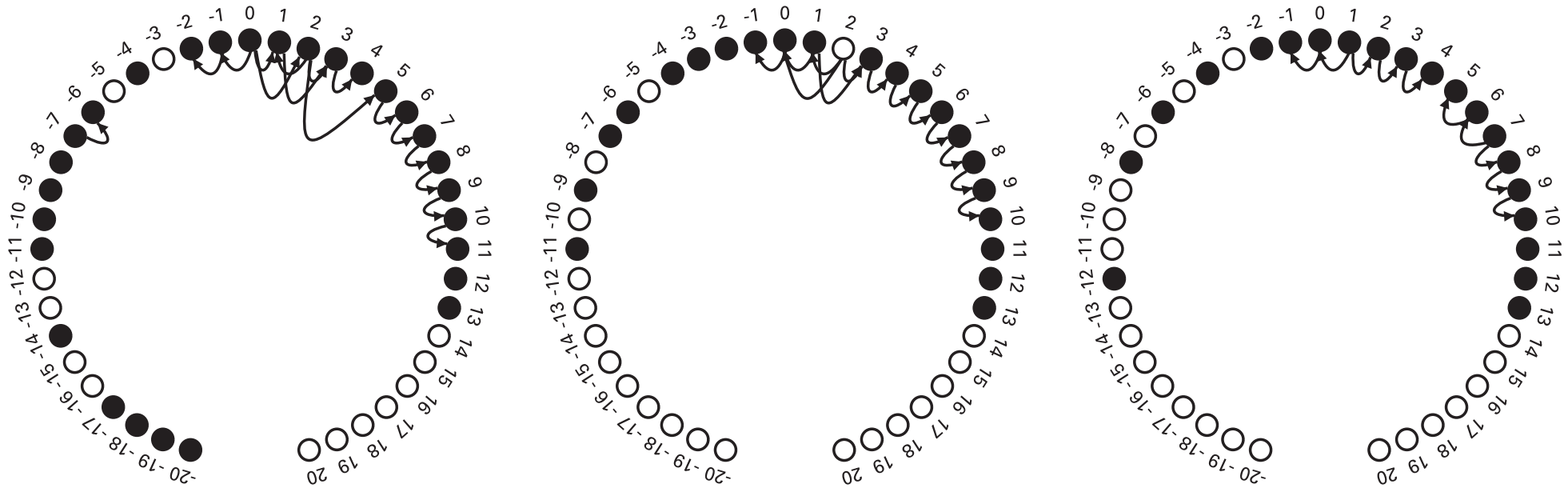


Wetterbom
(282 parameters)

Katze
(684 parameters)

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

Illumina

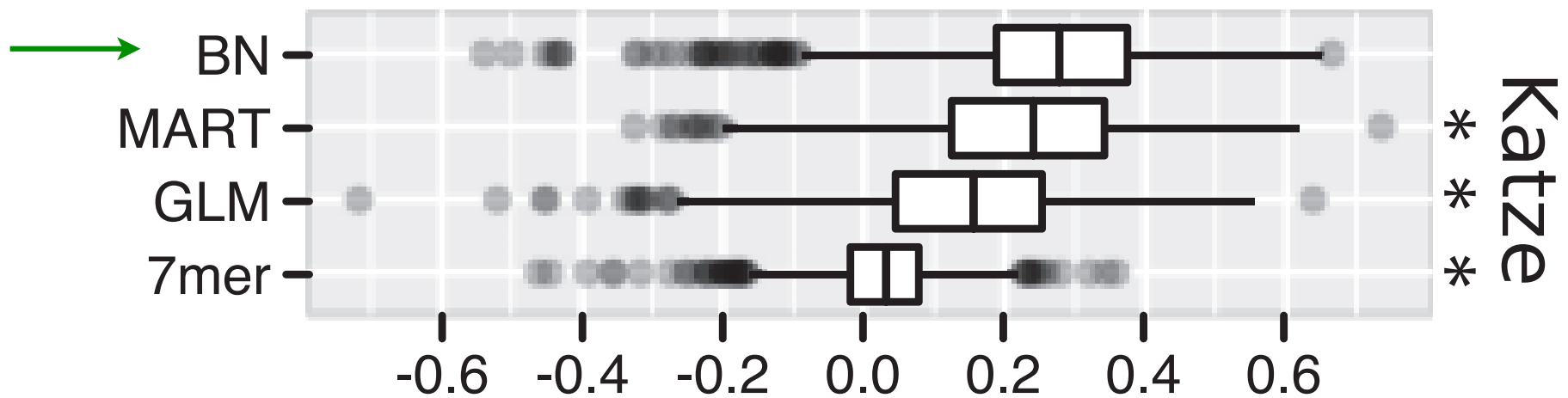
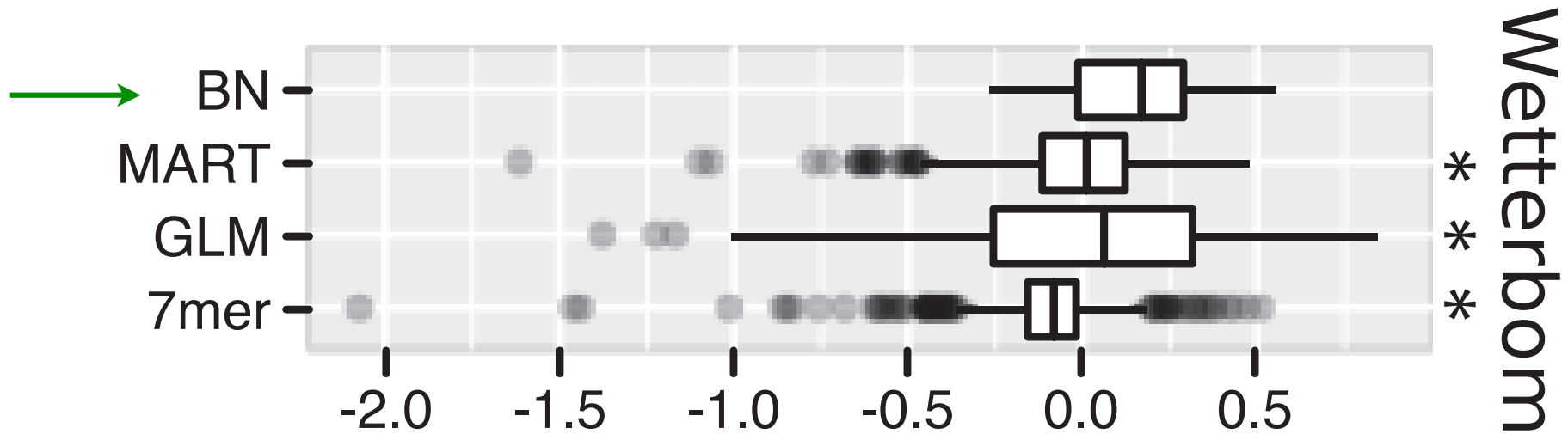


Bullard
(696 parameters)

Mortazavi
(582 parameters)

Trapnell
(360 parameters)

Result – Increased Uniformity



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

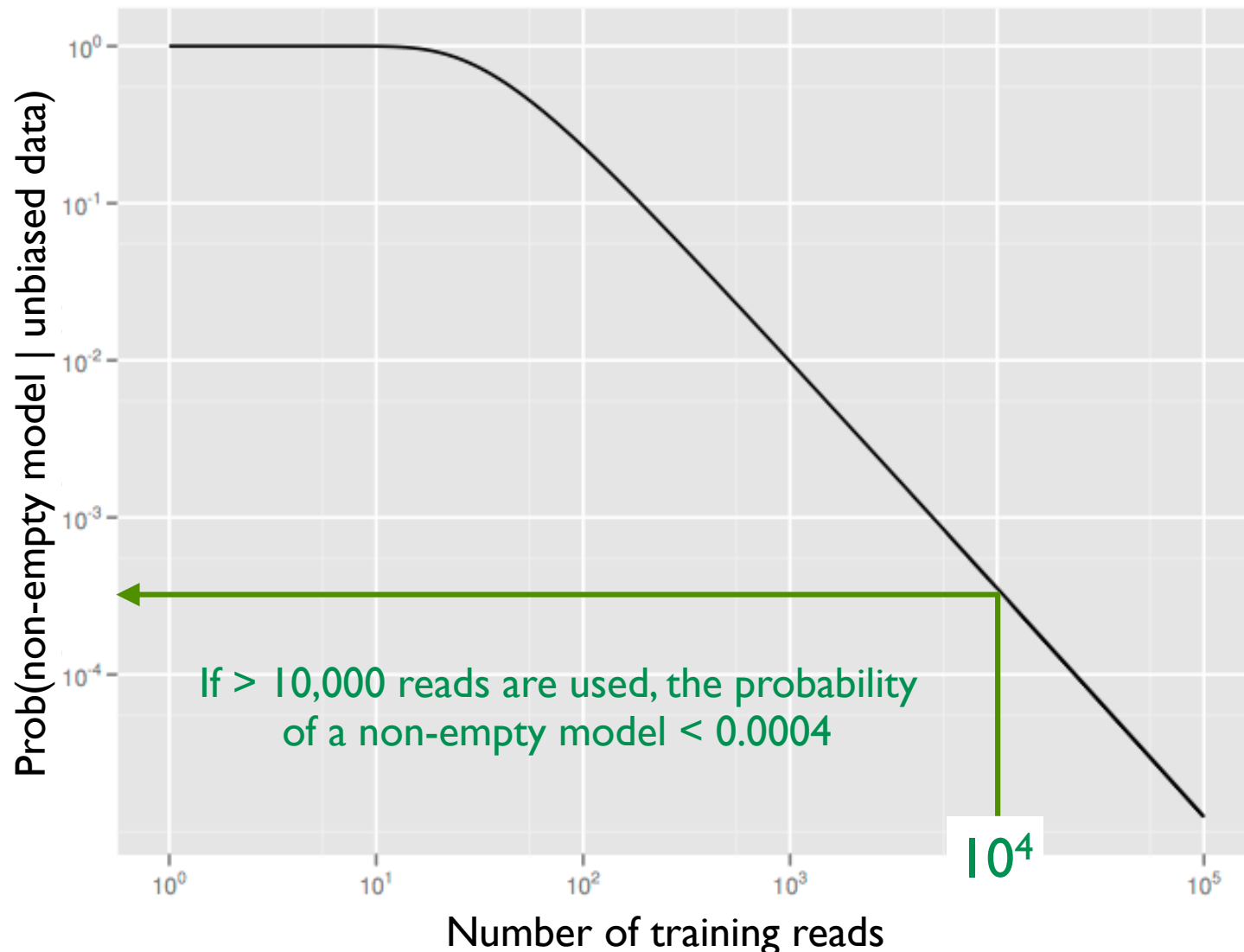
→ R^2

* = 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 .



how different are two distributions?

Given: r -sided die, with probs $p_1 \dots p_r$ of each face. Roll it $n=10,000$ times; observed frequencies = q_1, \dots, 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, \dots, 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, \dots, 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

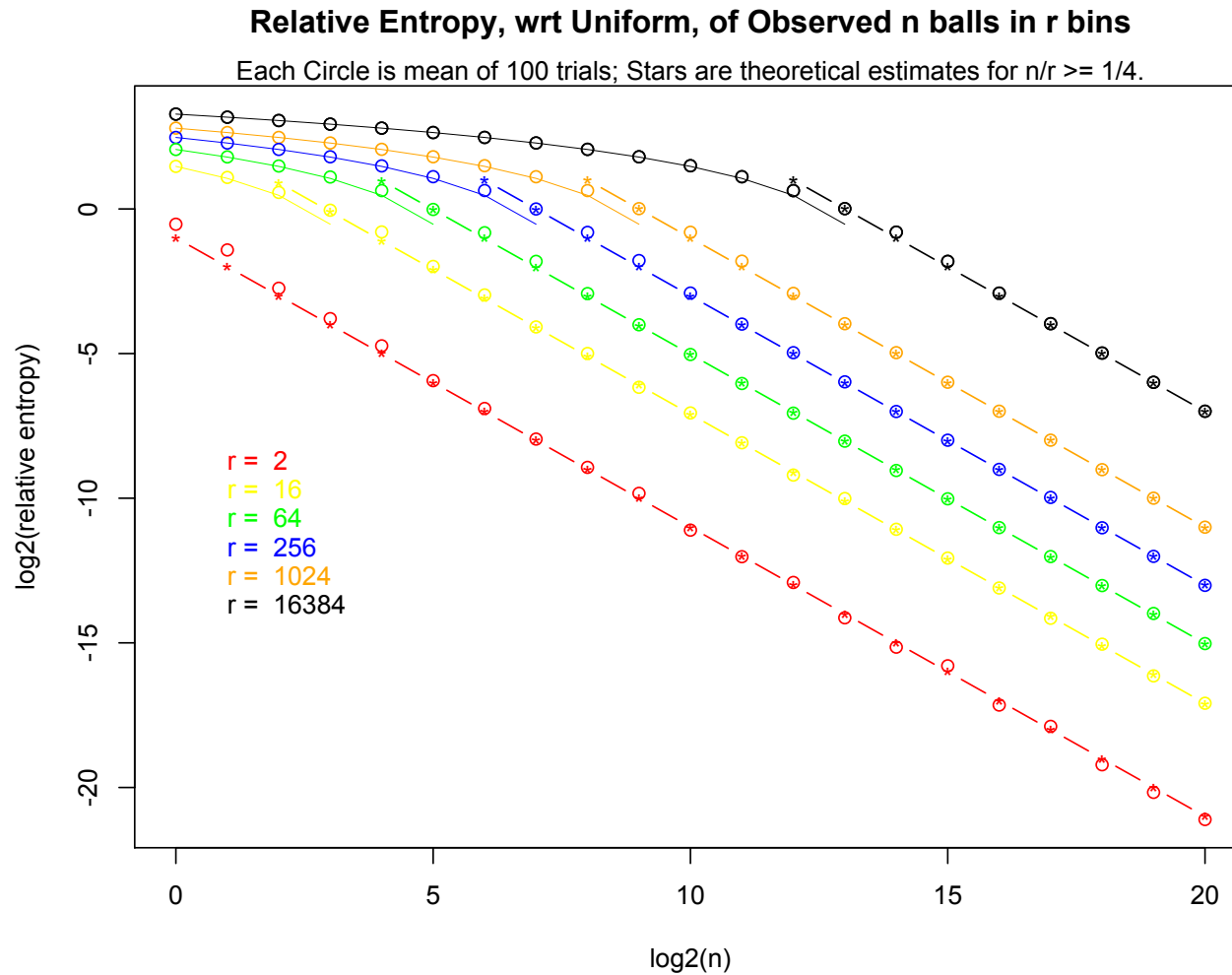
$$\begin{aligned} H(Q||P) &\approx \sum_{i=1}^r q_i \frac{q_i - p_i}{p_i} - p_i \frac{q_i - p_i}{p_i} - \frac{q_i}{2p_i} \frac{(q_i - p_i)^2}{p_i} \\ &= \sum_{i=1}^r \frac{(q_i - p_i)^2}{p_i} \left(1 - \frac{q_i}{2p_i} \right) \\ &\approx \frac{1}{2} \sum_{i=1}^r \frac{(q_i - p_i)^2}{p_i} \end{aligned}$$

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 $n \rightarrow \infty$ 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

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

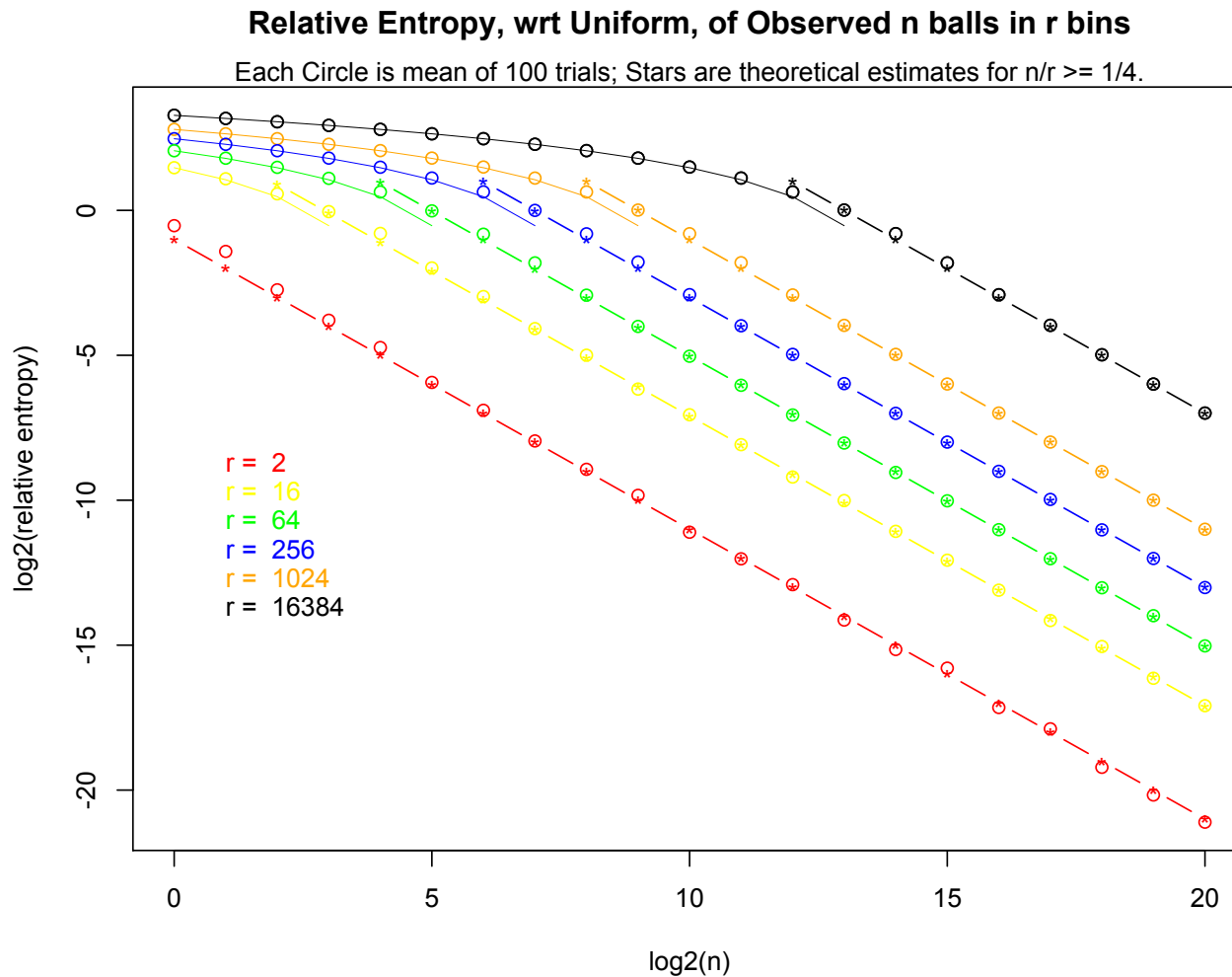


... and after a modicum of algebra:

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

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)

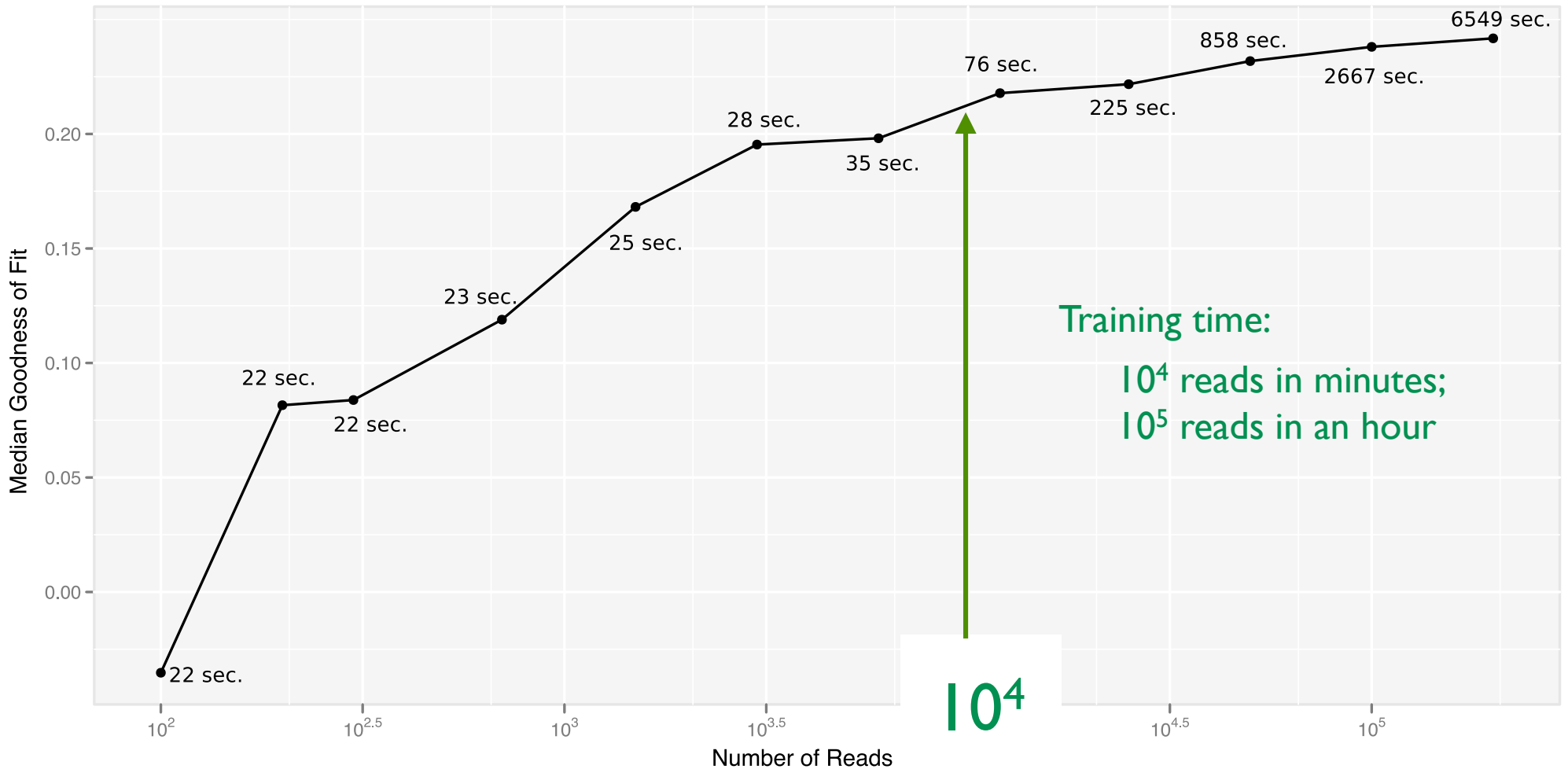


Figure 8: Median R^2 is plotted against training set size. Each point is additionally labeled with the run time of the training procedure.

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:

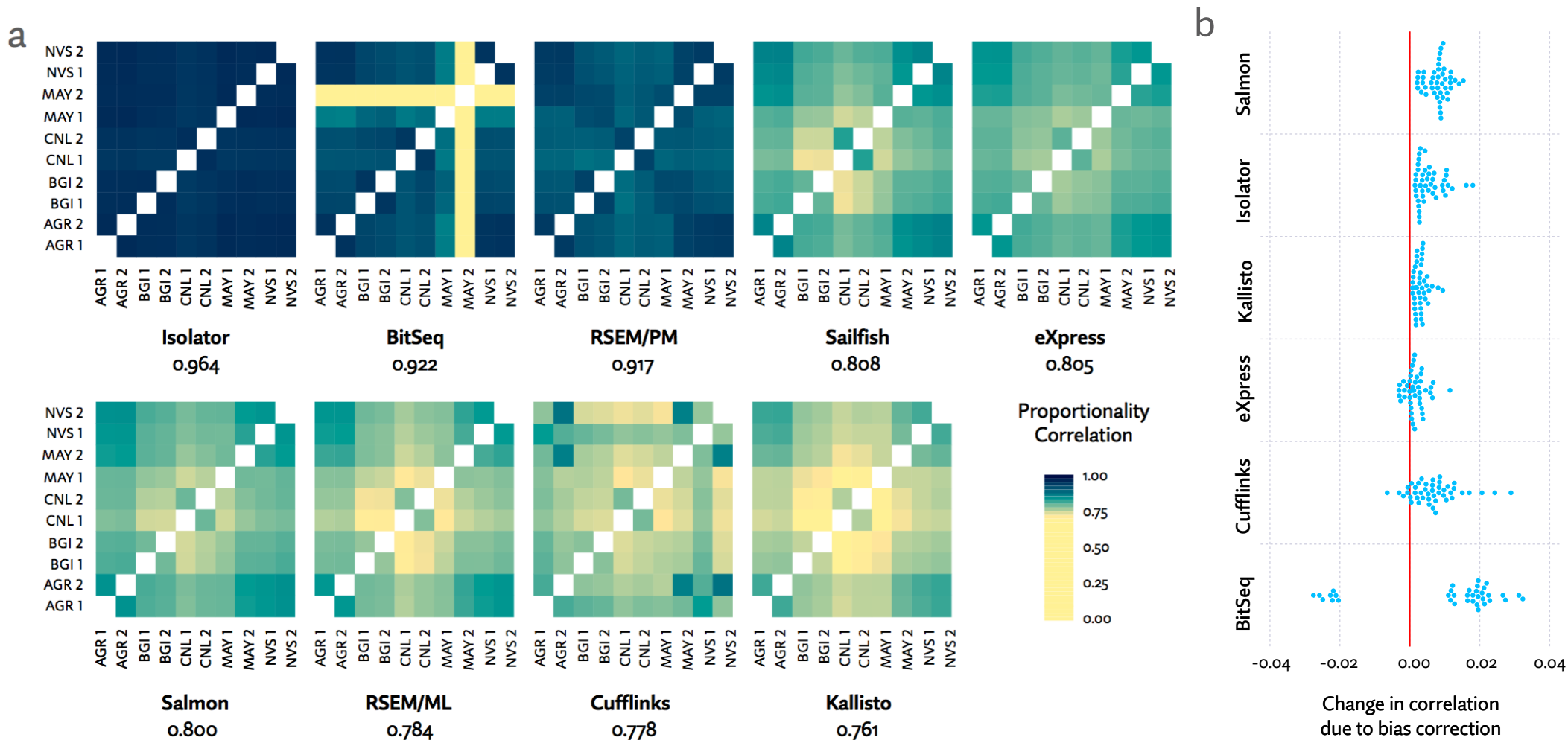
SNPs and/or alternative splicing might have a big effect, if samples are genetically different and/or engender changes in isoform usage

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.

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 Peng⁴ and Michael G. Katze⁴

¹Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195-2350,

²Department of Genome Sciences, University of Washington, Seattle, WA 98195-5065, ³Fred Hutchinson Cancer Research Center, Seattle, WA 98109 and ⁴Department 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 quantification of low level genes.





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-pos using a simple Bayesian network, the structu reads and a reference genome sequence.

Author: Daniel Jones <dcjones at cs.washing

Maintainer: Daniel Jones <dcjones at cs.wasl

To install this package, start R and enter:

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

To cite this package in a public

```
citation("
```

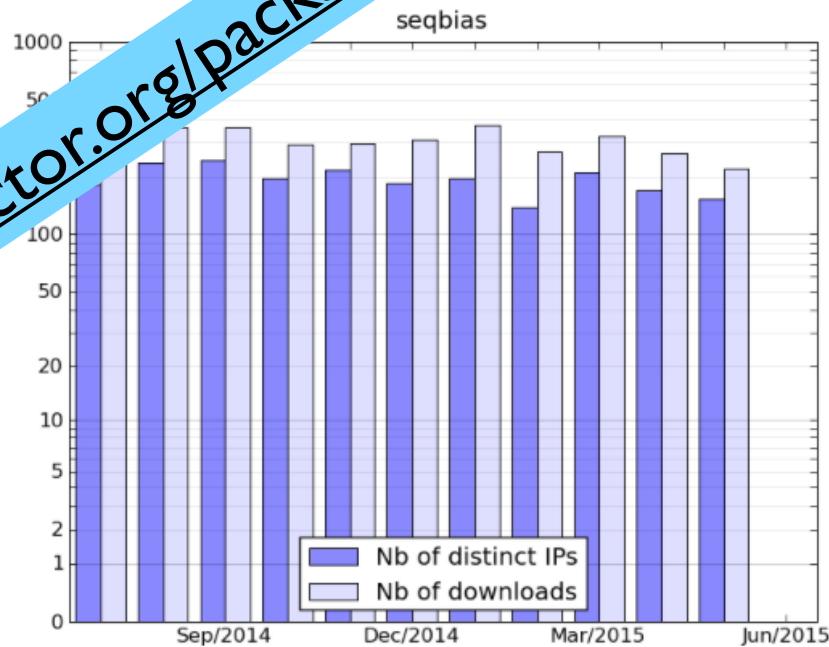
Docum

Assessing and Adjusting
Reference Manual

Download stats for Software package seqbias

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

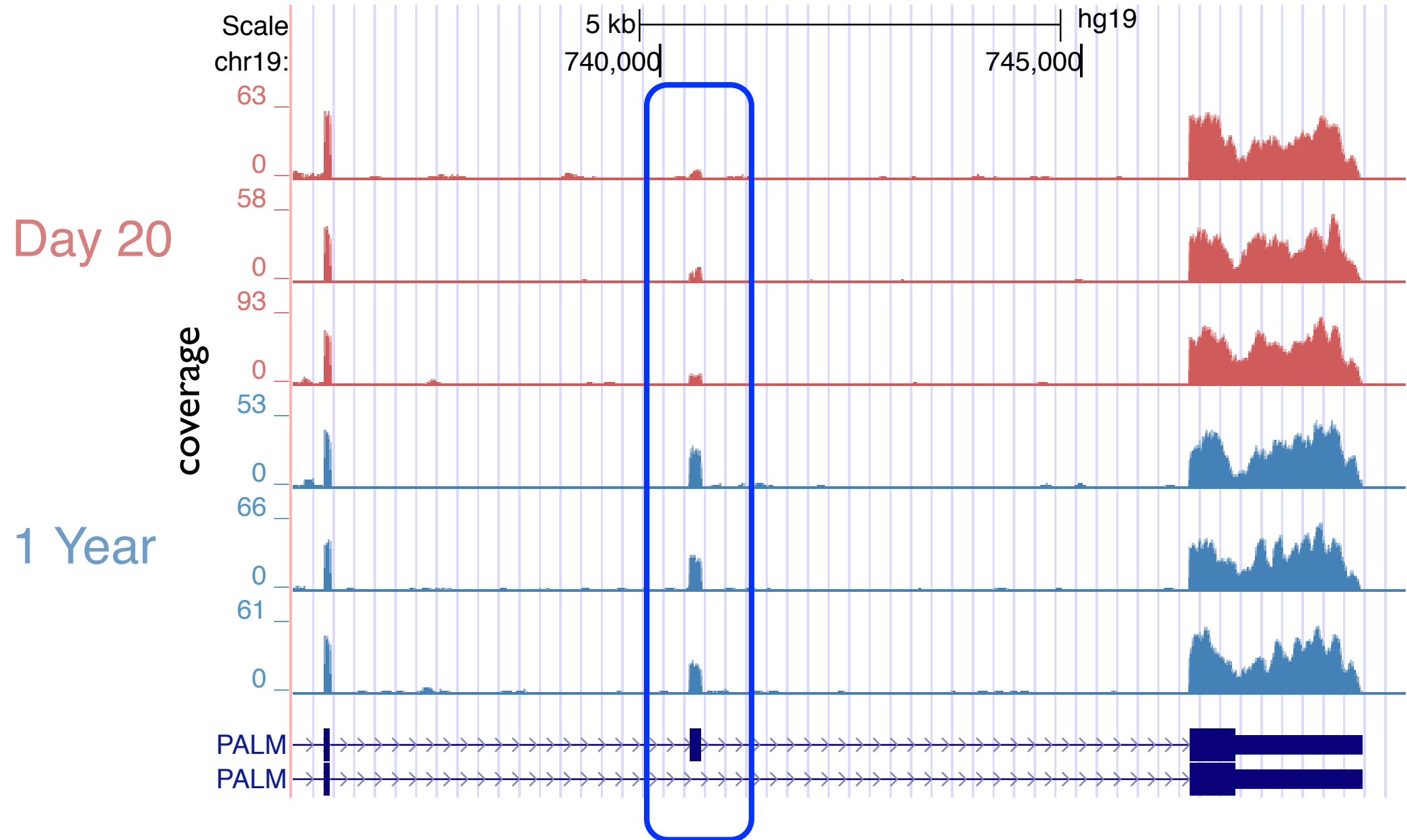
seqbias home page: [release version](#), [devel version](#).



Month	Nb of distinct IPs	Nb of downloads
Jul/2014	181	252
Aug/2014	236	360
Sep/2014	242	360
Oct/2014	197	292
Nov/2014	217	299
Dec/2014	186	311
Jan/2015	195	371
Feb/2015	138	270
Mar/2015	211	327
Apr/2015	170	264
May/2015	153	220
Jun/2015	0	0
All months	1648	3326

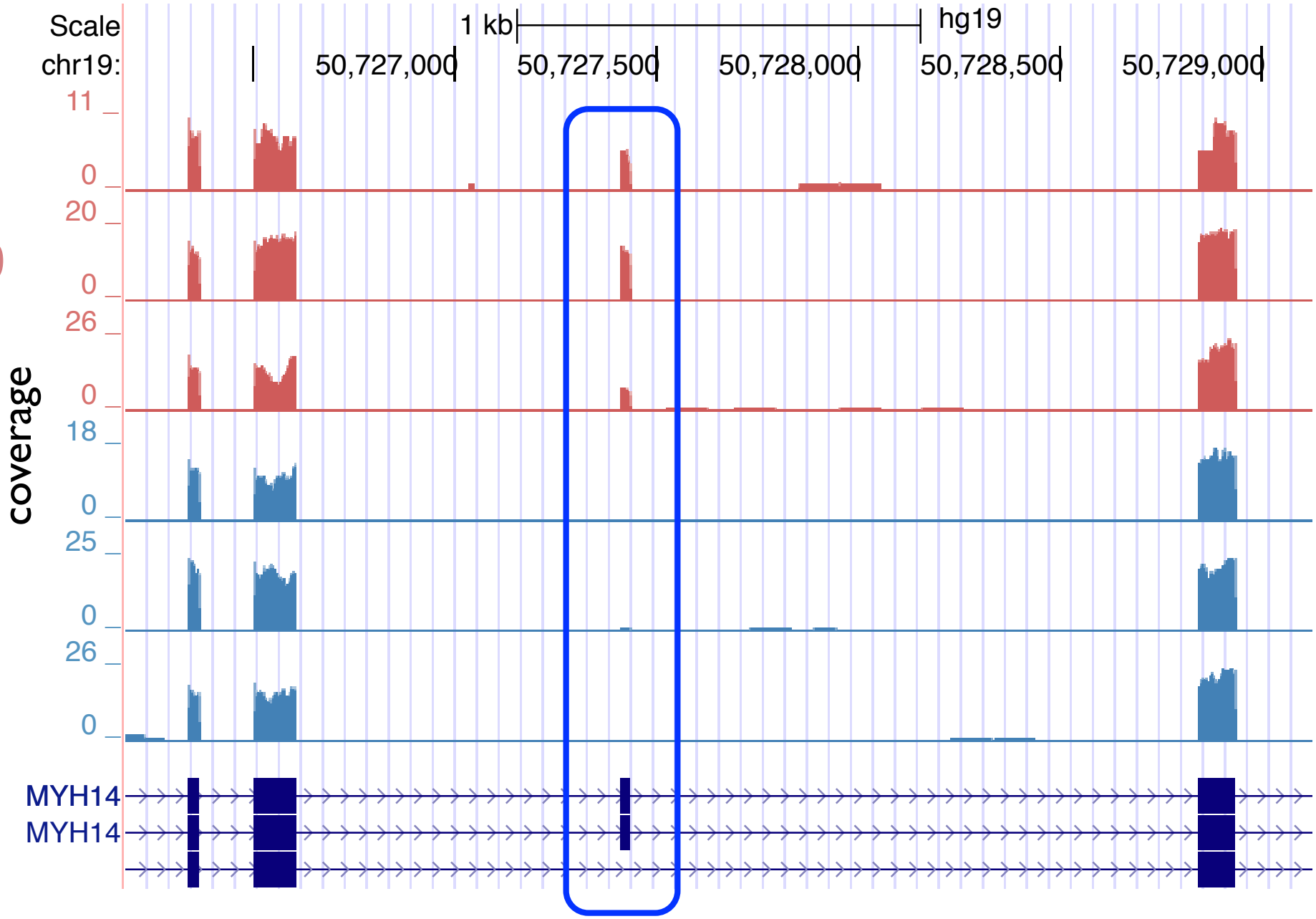
<http://bioconductor.org/packages/release/bioc/html/seqbias.html>

Alternate Splicing



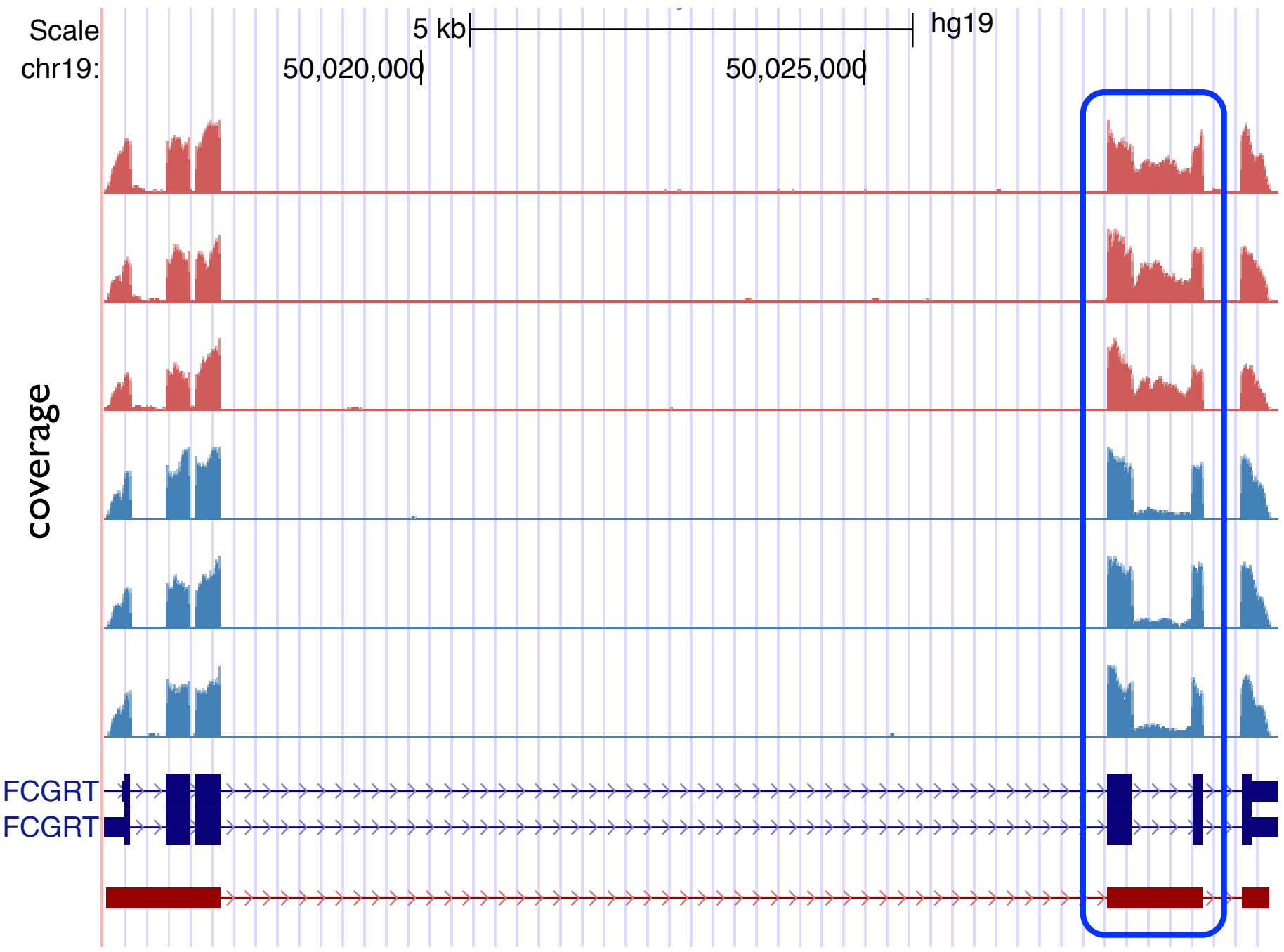
Day 20

1 Year



Day 20

1 Year



Is Isoform Quantification Hard?

Sequencing depth *per-isoform* is lower

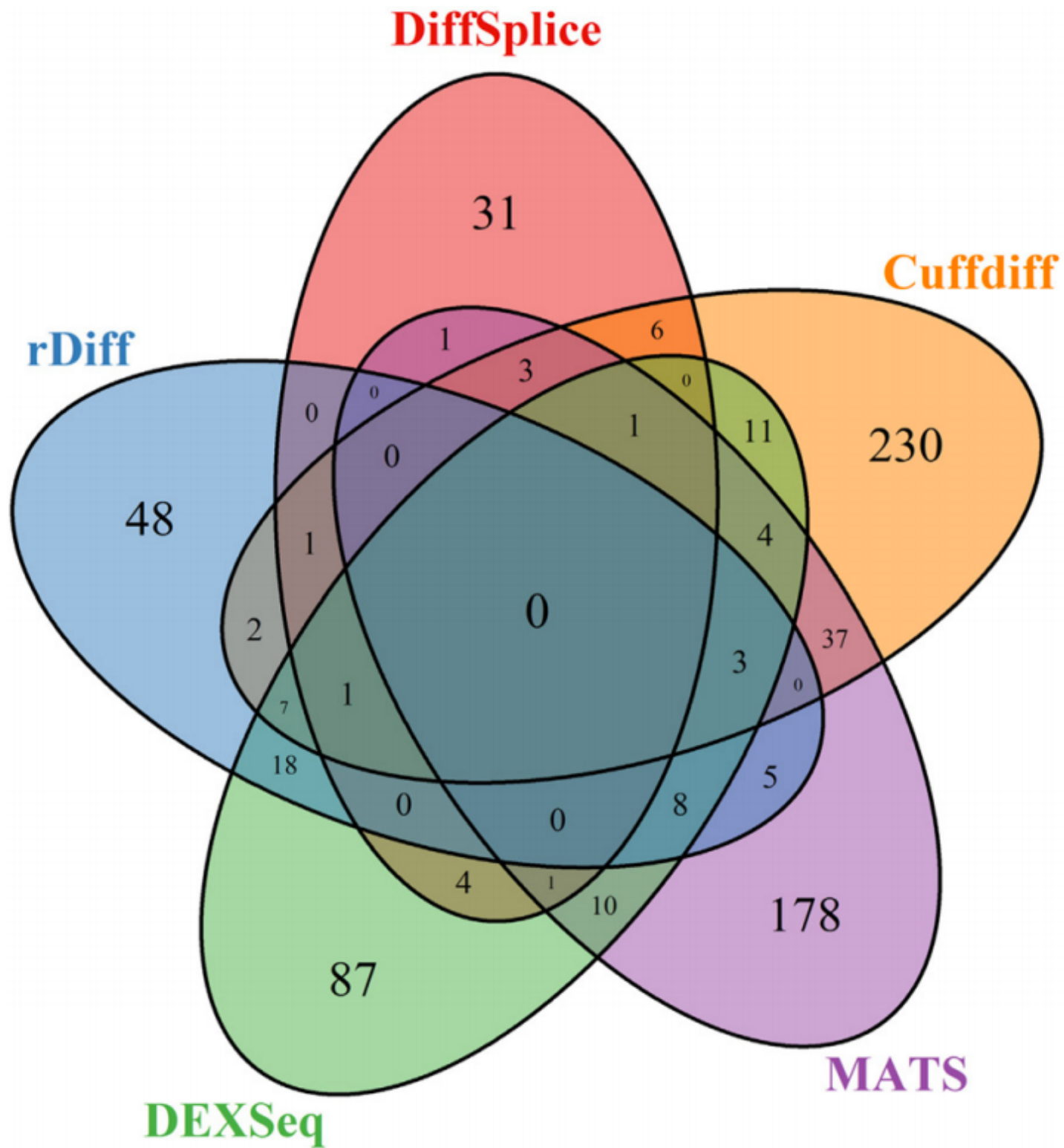
Many reads *ambiguously* mapped to multiple isoforms

Isoform *proportions* and *total expression* may both vary

All the previously-mentioned *bias issues*, including batch effects, affect all measurements

Differences among isoforms may be only a *small fraction of nucleotides* in transcript, potentially exacerbating bias

Isoform annotation is incomplete/poor



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

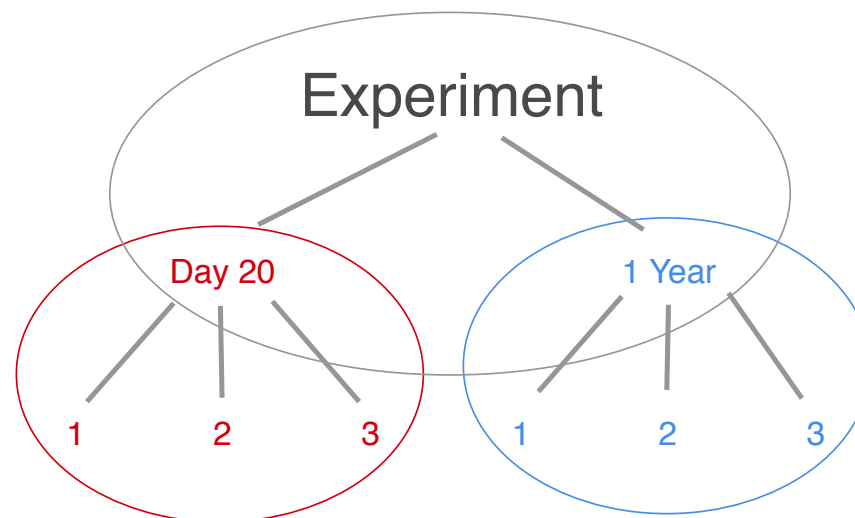
Isolator

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,
suppressing spurious changes.



Experiment

Conditions

Replicates

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

Why a Hierarchical Bayesian Model?

In a nutshell:

A natural assumption is that “nothing has changed,” unless refuted by data. (Most genes don’t change.)

Hierarchical model allows estimation of baseline expression/isoform usage/variability across *all* samples

This helps compensate for lower per-isoform coverage

Ex: Given 4 isoforms with

1, 1, 2, 2 reads in condition A vs

2, 2, 1, 1 reads in condition B


do you think all 4 are 2x different?

In a nutshell: posterior means are *more stable* than MLEs

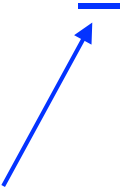
Likelihood surface max often a broad plateau, not a sharp peak

Toy example:

Isoform 1, length 1k: 

Isoform 2, length 2k: 

For simple likelihood model, one read here yields MLE expression of Iso1 twice that of Iso2



But one read here gives zero as MLE for Iso1!



OTOH, posterior mean is not zero in either case

Some Benchmarks

“Sequencing Quality Consortium” (SEQC)

4 RNA samples with spike-ins

They ran RNAseq

They did extensive PCR for “gold standard”

We ran multiple tools (on common alignment)

Evaluated “Proportionality correlation”

($2 \cdot \text{covariance} / \text{sum-of-variances}$, log-scale; usual $-1 \dots 1$ range)

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	A	B	C	D
Isolator	0.979	0.978	0.981	0.982
Salmon	0.976	0.975	0.978	0.979
Kallisto	0.972	0.972	0.973	0.976
Sailfish	0.970	0.969	0.969	0.972
Cufflinks	0.967	0.969	0.972	0.974
RSEM/PM	0.943	0.949	0.944	0.949
RSEM/ML	0.941	0.948	0.945	0.951
BitSeq	0.940	0.949	0.943	0.946
eXpress	0.931	0.939	0.935	0.942

Table 3: Proportionality correlation between known proportions of 92 ERCC spike-in controls 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 4: Proportionality correlation between transcript-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.

Method	Correlation
Isolator	0.919
Kallisto	0.887
Salmon	0.886
RSEM/ML	0.881
Cufflinks	0.881
eXpress	0.825
Sailfish	0.816
RSEM/PM	0.806
BitSeq	0.796

Table 5: Proportionality correlation between ground truth and estimates produced by each method on simulated RNA-Seq.

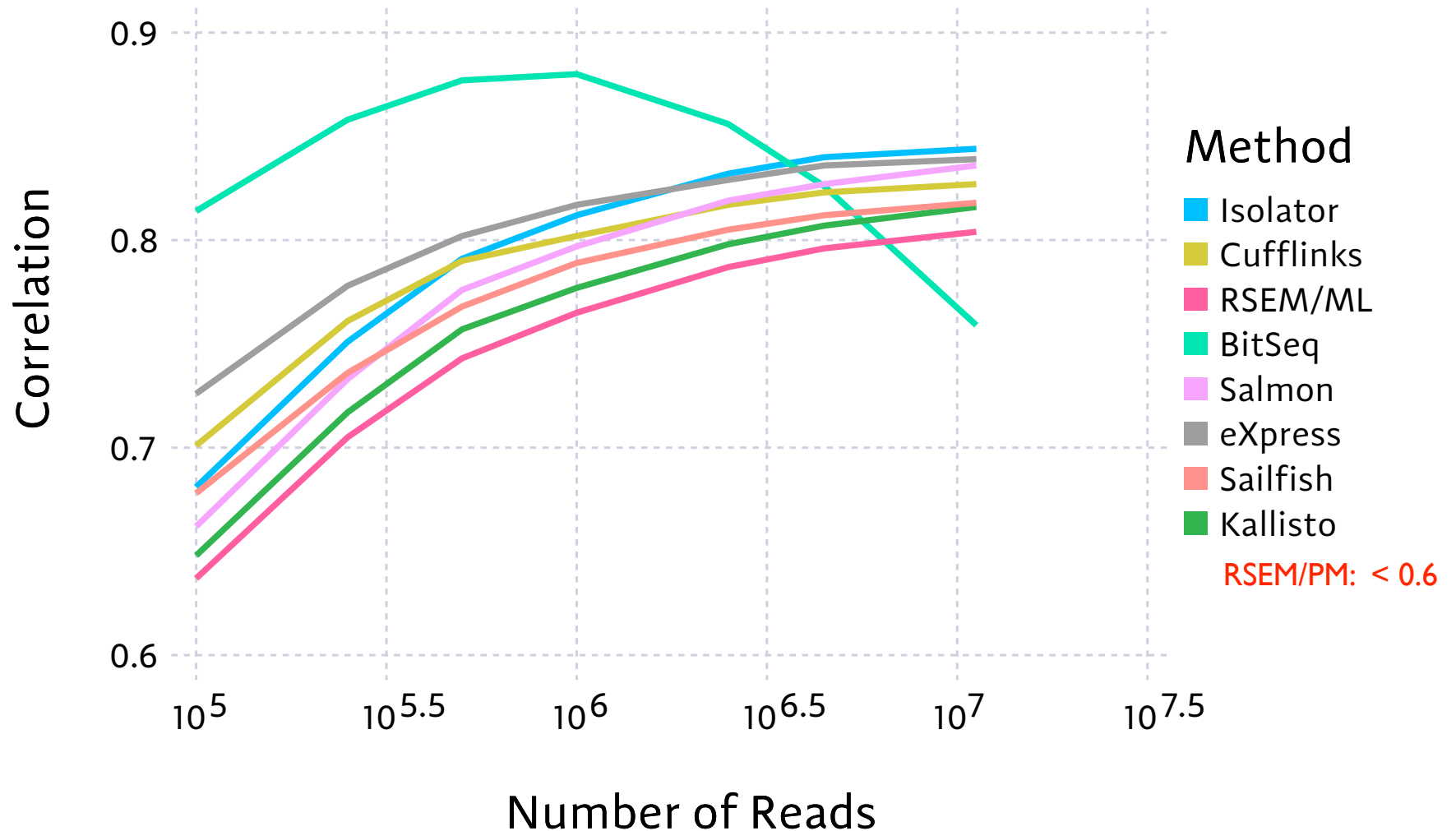
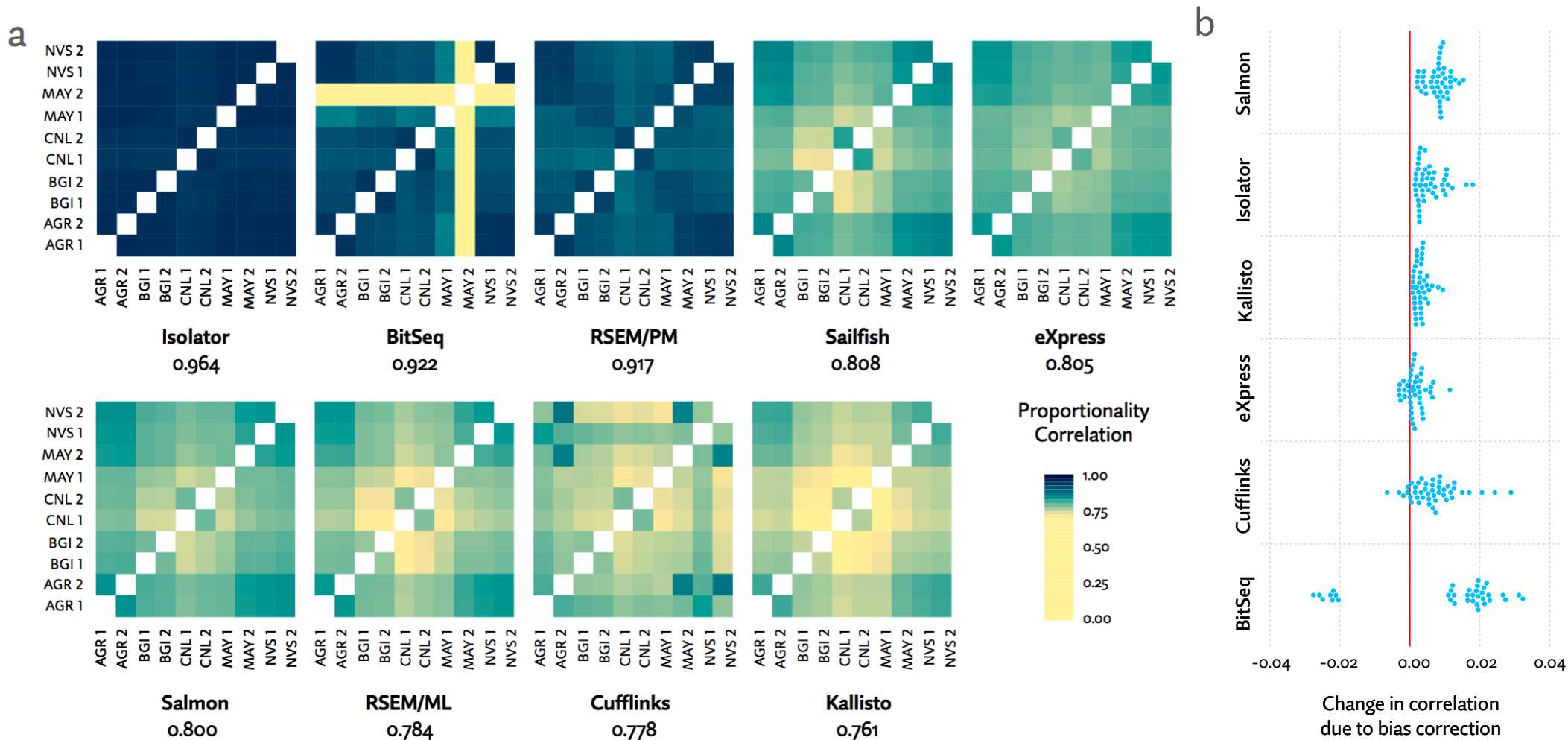


Figure 2: Proportionality correlation between estimates from 4.5 million 300nt MiSeq reads and progressively larger numbers of HiSeq 2000 reads. (100x2)

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.

Time

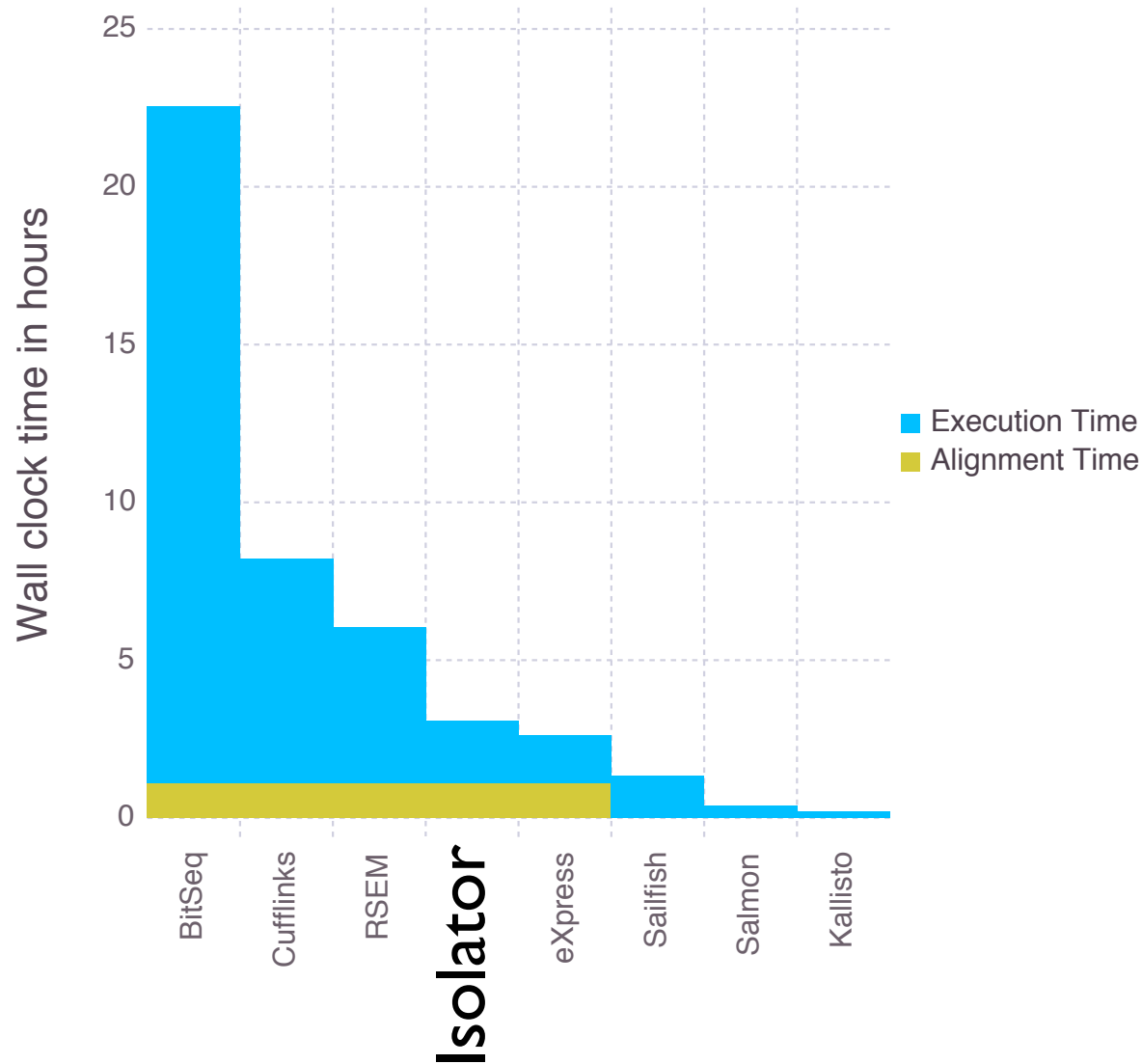


Figure 1: Run time needed to process the SEQC data presented in the Results sections. All methods were run a Google Compute Engine instance backed by four Intel Xeon cores and 52GB of a memory.

Story 2

Let-7 family of microRNA is required for maturation and adult-like metabolism in stem cell-derived cardiomyocytes

Kavitha T. Kuppusamy^{a,b}, Daniel C. Jones^c, Henrik Sperber^{a,b,d}, Anup Madan^e, Karin A. Fischer^{a,b}, Marita L. Rodriguez^f, Lil Pabon^{a,g,h}, Wei-Zhong Zhu^{a,g,h}, Nathaniel L. Tulloch^{a,g,h}, Xiulan Yang^{a,g,h}, Nathan J. Sniadecki^{f,i}, Michael A. Laflamme^{a,g,h}, Walter L. Ruzzo^{c,j,k}, Charles E. Murry^{a,g,h,i,l}, and Hannele Ruohola-Baker^{a,b,i,j,m,1}

^aInstitute for Stem Cell and Regenerative Medicine, Seattle, WA 98109; Departments of ^bBiochemistry, ^cComputer Science and Engineering, and ^dChemistry, University of Washington, Seattle, WA 98195; ^eLabCorp Genomic Services, Seattle, WA 98109; ^fDepartment of Mechanical Engineering, ^gDepartment of Pathology, ^hCenter for Cardiovascular Biology, ⁱDepartment of Bioengineering, and ^jDepartment of Genome Sciences, University of Washington, Seattle, WA 98195; ^kFred Hutchinson Cancer Research Center, Seattle, WA 98109; and ^lDepartment of Medicine/Cardiology and ^mDepartment of Biology, University of Washington, Seattle, WA 98195

Edited by Eric N. Olson, University of Texas Southwestern Medical Center, Dallas, TX, and approved April 14, 2015 (received for review December 18, 2014)

Published: 2015-05-11

It is possible to grow cardiomyocytes (heart muscle cells) from human embryonic stem cells (hESC-CMs)

Can grow billions of them

Can transplant them into animals after heart attack

Cells integrate/heart function improves (after a few weeks)

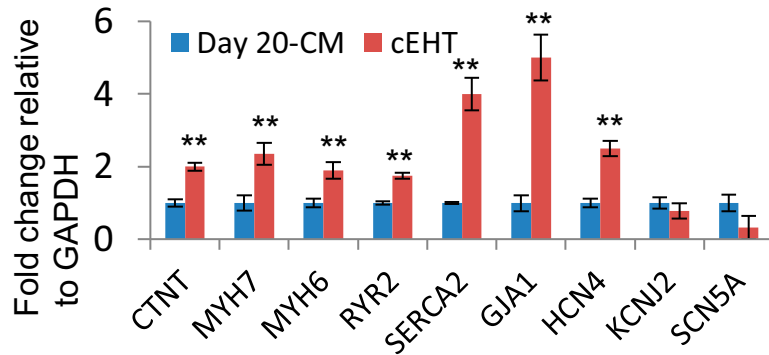
BUT – arrhythmias, at least in the early stages

Why? Probably because hESC-CMs were *immature*.

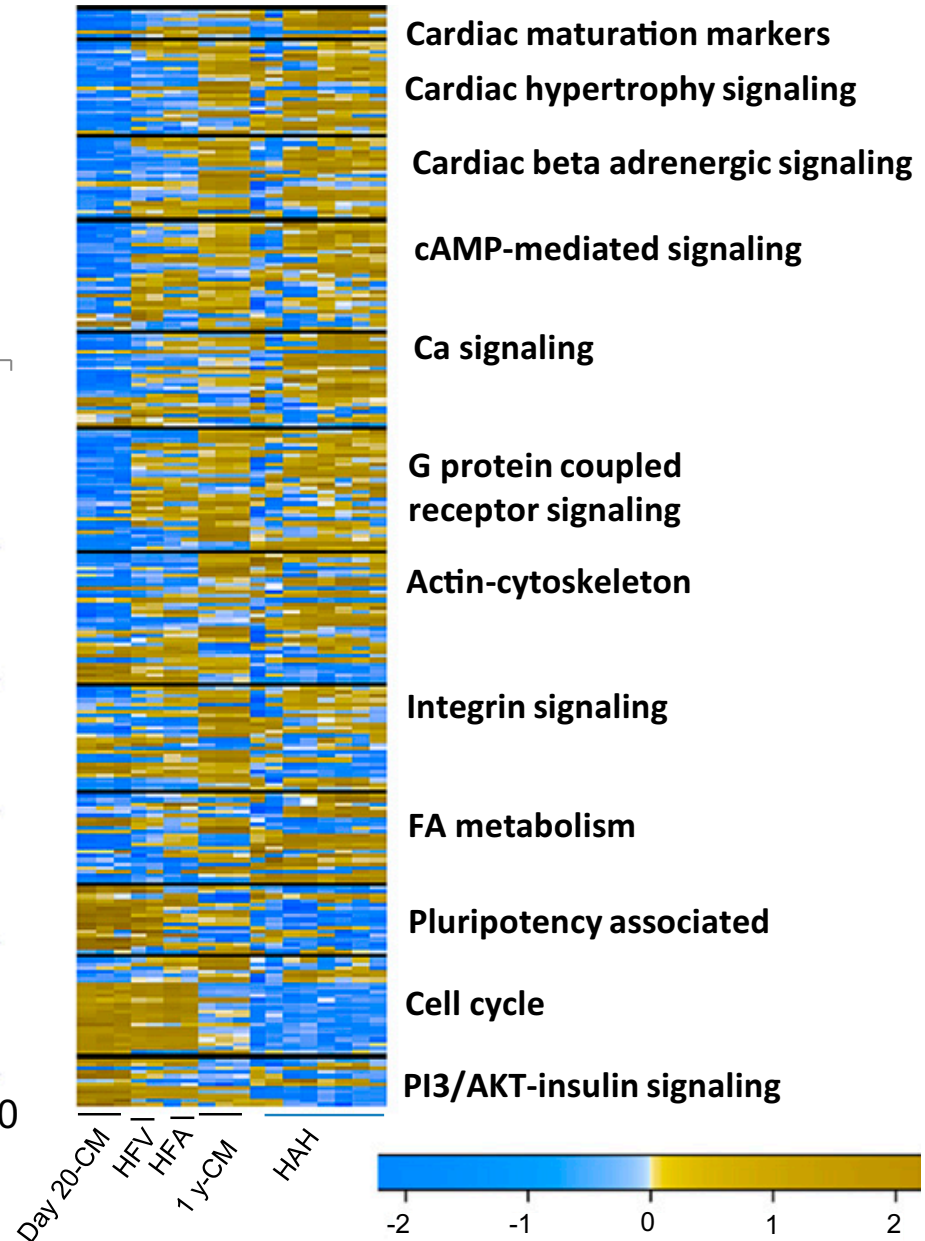
This *will* be tried in humans within a few years; ability to lab-culture *mature* hESC-CMs will greatly improve chances for success. Growing them *quickly* will greatly improve the economics. How can we do that?

step 1: find molecular biomarkers for maturity

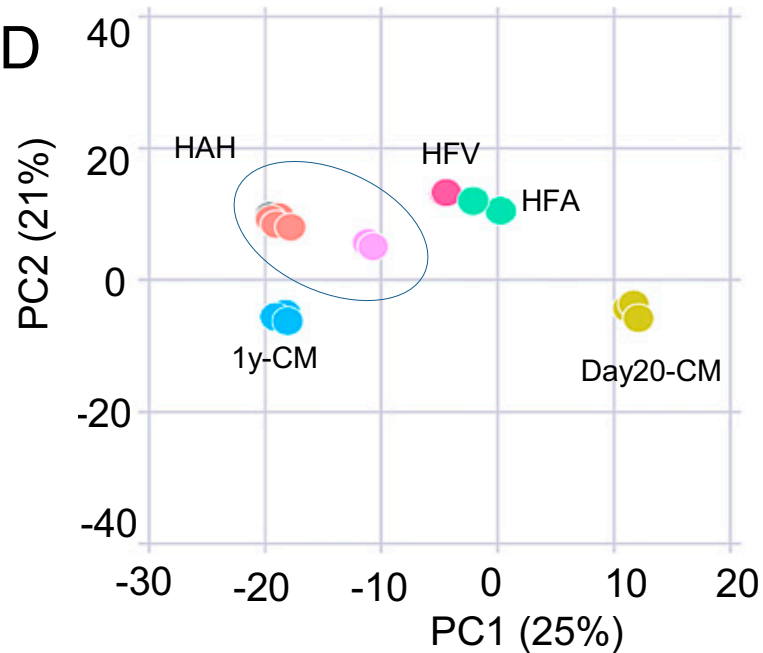
C



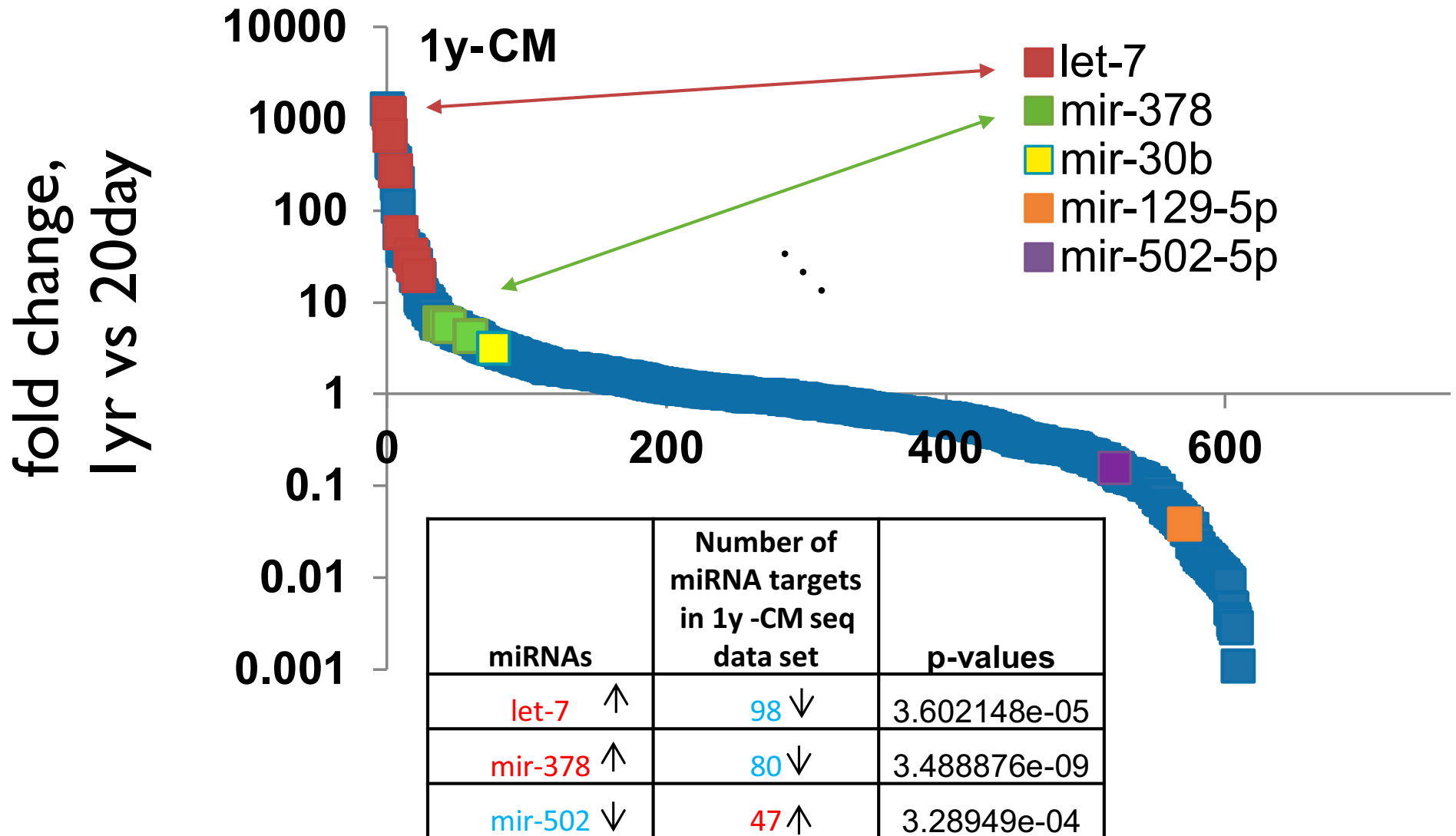
E



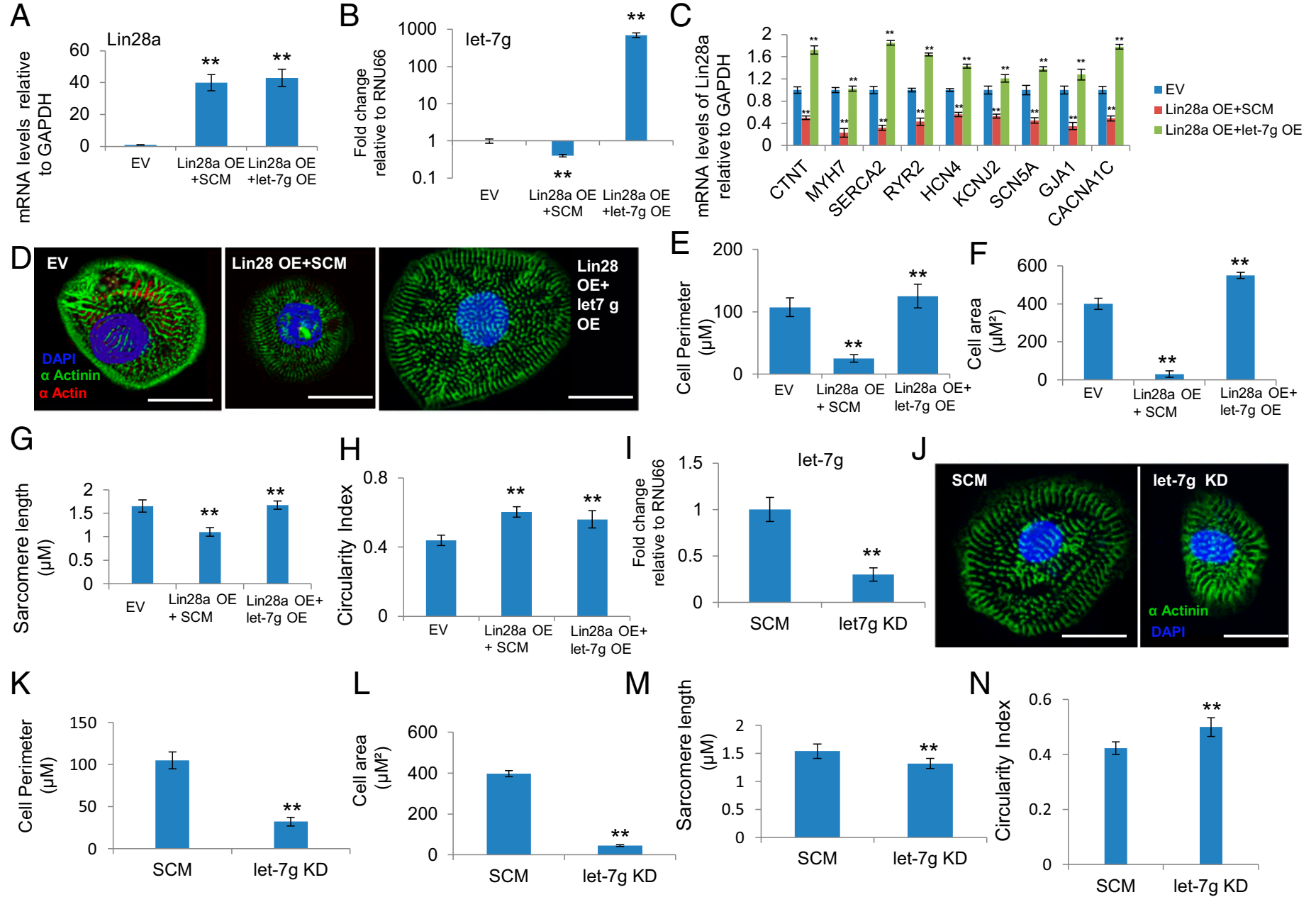
D



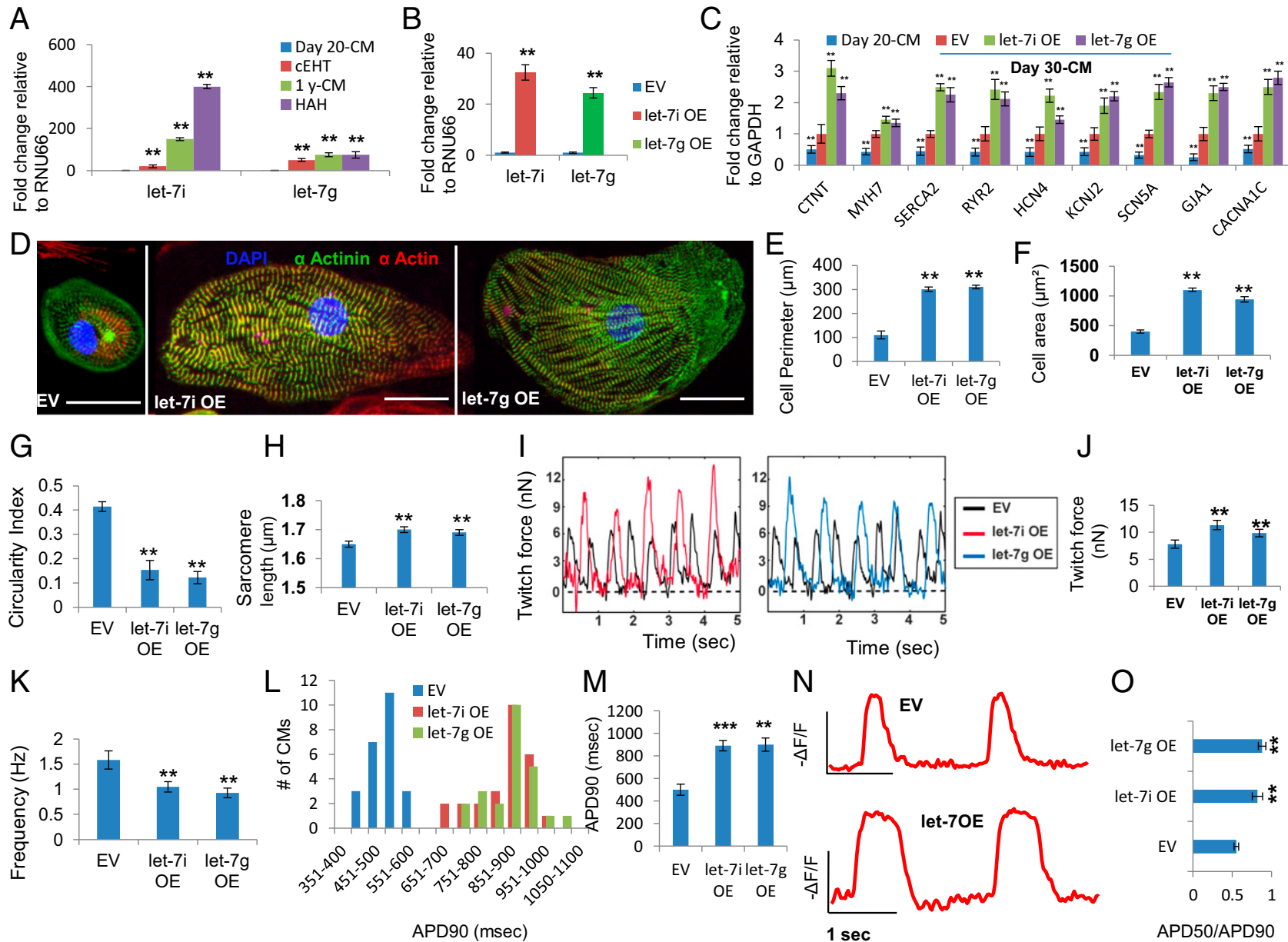
step 1 (cont.): find miRNA biomarkers for maturity, too



step2a: let-7 is driver, not passenger – it's necessary



step2b: let-7 is driver, not passenger – it's sufficient

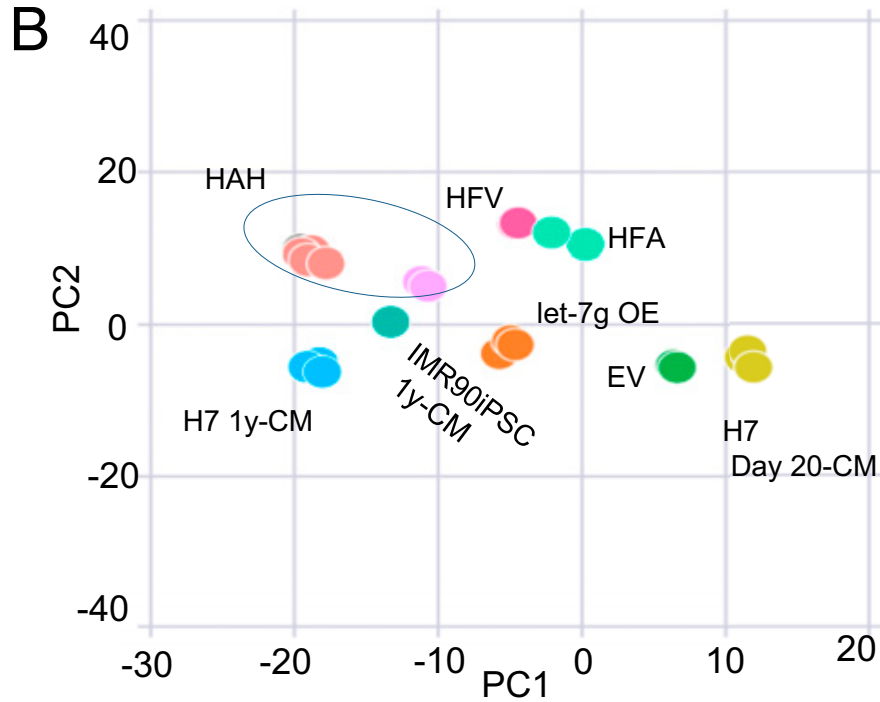


Pathways

Physiology

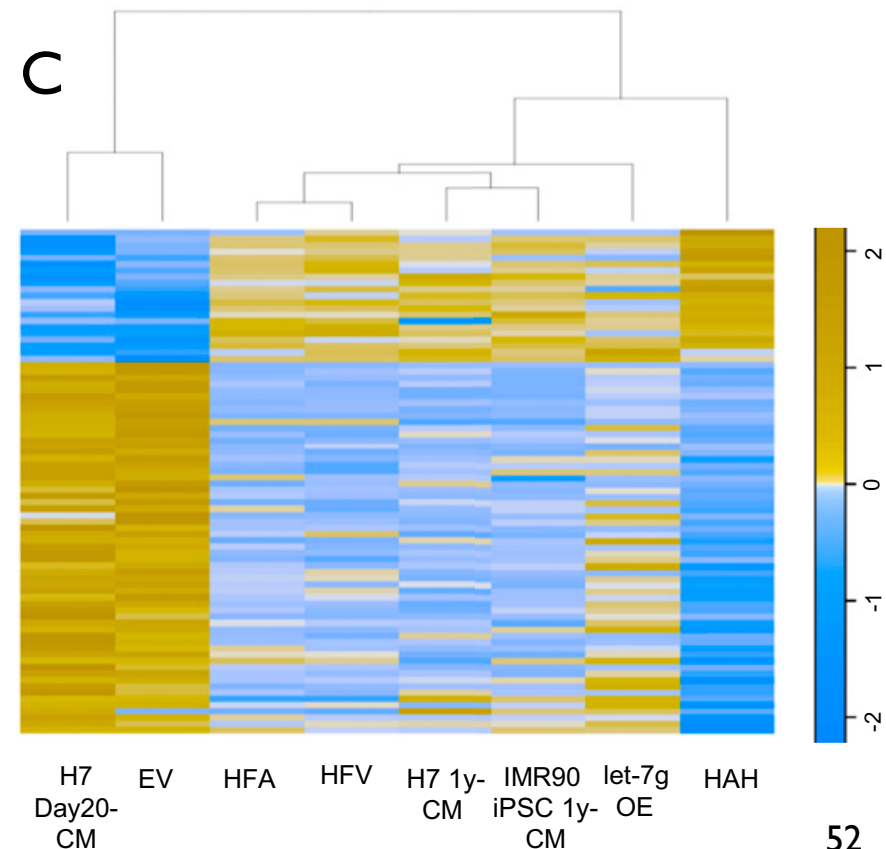
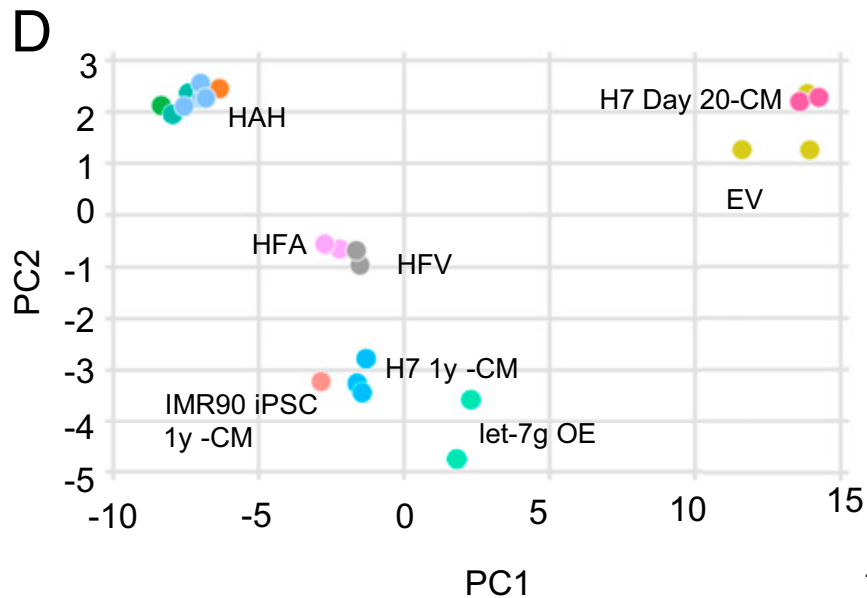
Etc.

Back to Story 1: differential splicing speaks, too



B: gene expression in cardiac- pathways (E) tracks maturation (unsurprisingly)

C/D: so does splicing (indp of level) via Isolator-detected probable monotonic changes. (Not easily assessed by MLE-based methods...)



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.

Alternative splicing changes are very hard to quantify:

lower coverage, ambiguous mapping, bias, ...

BUT careful modeling can help:

Bayesian hierarchical model borrows power across all samples

Sampling/posterior mean estimation is more robust than MLE

Sampling allows novel questions to be addressed, e.g., “is isoform shift probably monotonic in time”

It doesn't have to be slow

AND 90% of genes undergo alt splicing for a reason; you can't see what it is if you don't look

Amazing progress in stem cell technology

Ability to study and control cellular developmental pathways is one of the frontiers of modern biology

Multi-faceted, multi-disciplinary problems with rich data

In this study, microRNA let-7 identified as a key driver of cardiomyocyte maturation

Differential splicing of many transcripts clearly implicated; their exact roles remain to be determined.

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