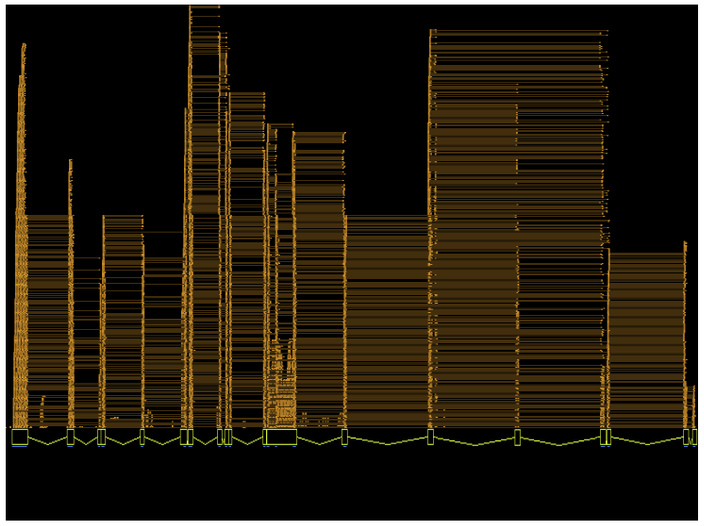
Analyzing RNA-seq gene expression data

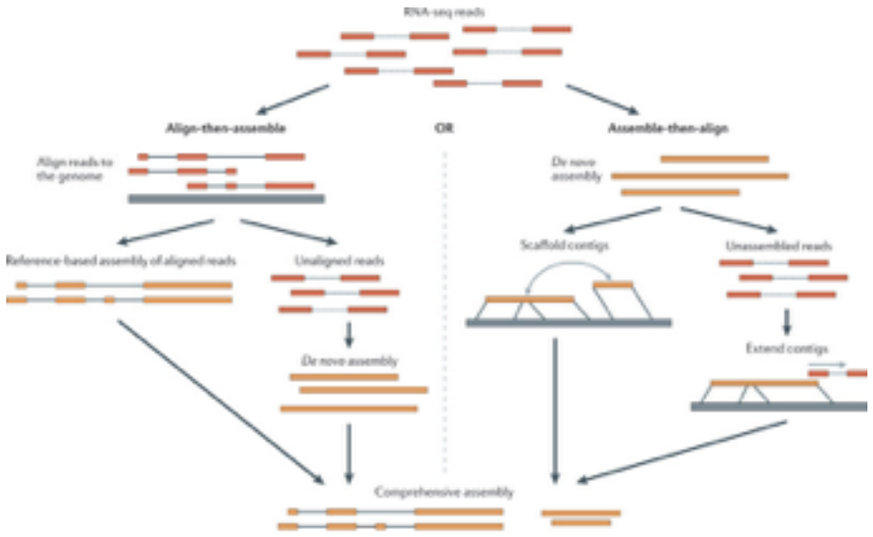
Jenny Tung and Amanda Lea
Conservation Genomics RNA-seq workshop
Asilomar, California 2016

@jtung5
@AmandaLea14
@AmboseliBaboons



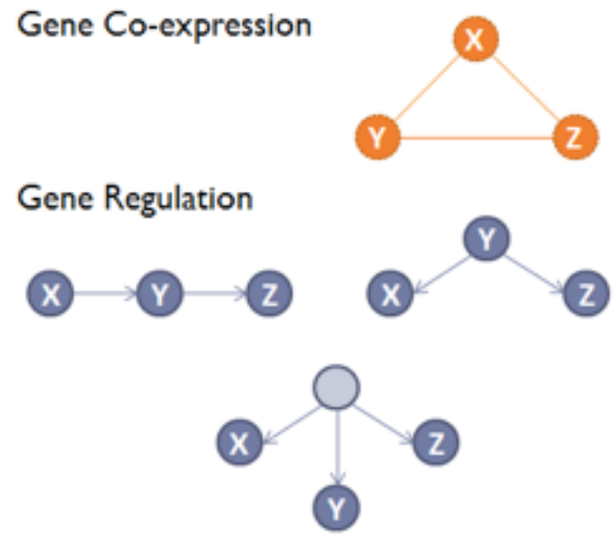
Simon White (Ensembl)

Gene annotation

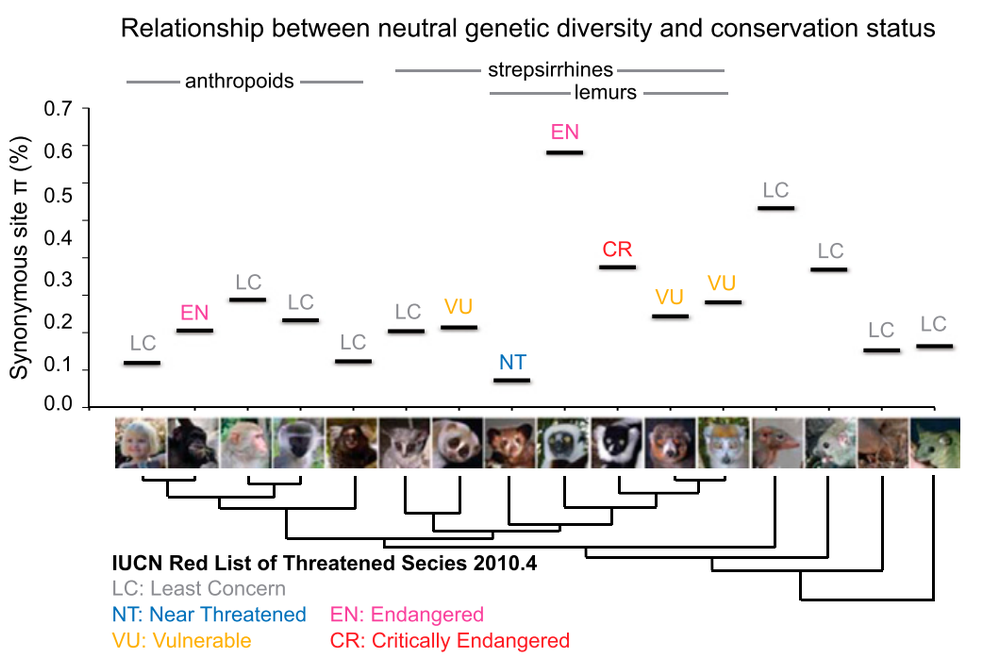


Martin and Wang 2011, NRG

Transcriptome assembly

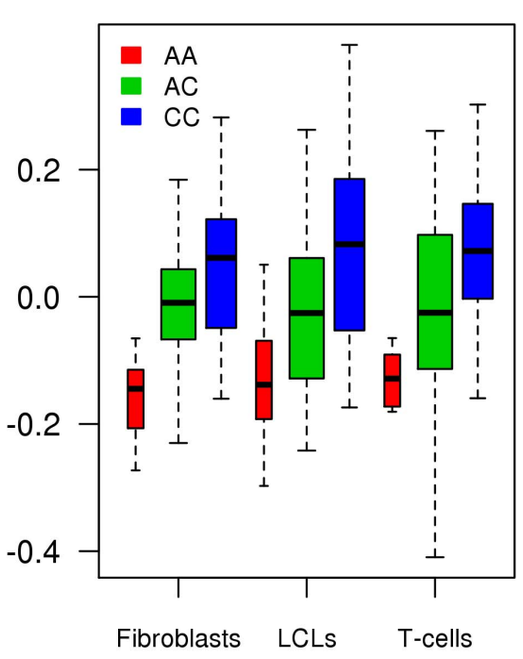


Pathway inference



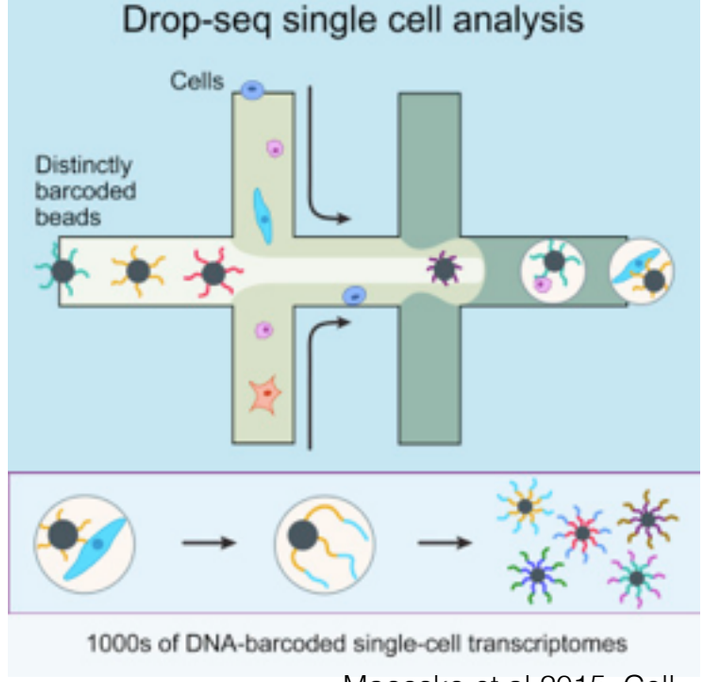
Perry et al 2011, Genome Research

Genotyping



Flutre et al 2013, PLoS Genetics

Association mapping



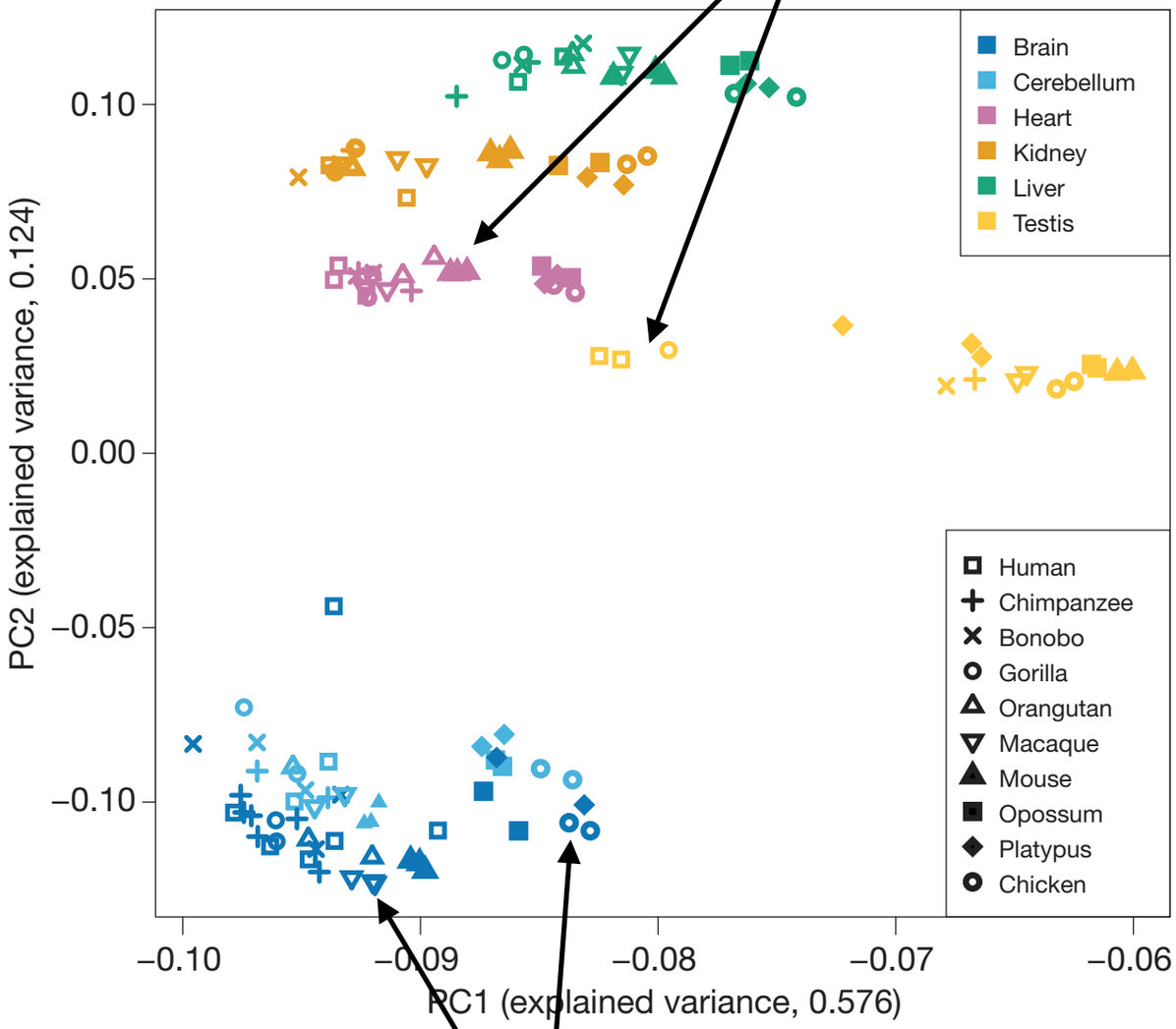
Macosko et al 2015, Cell

Single cell analysis

Our focus here is on differential gene expression analysis

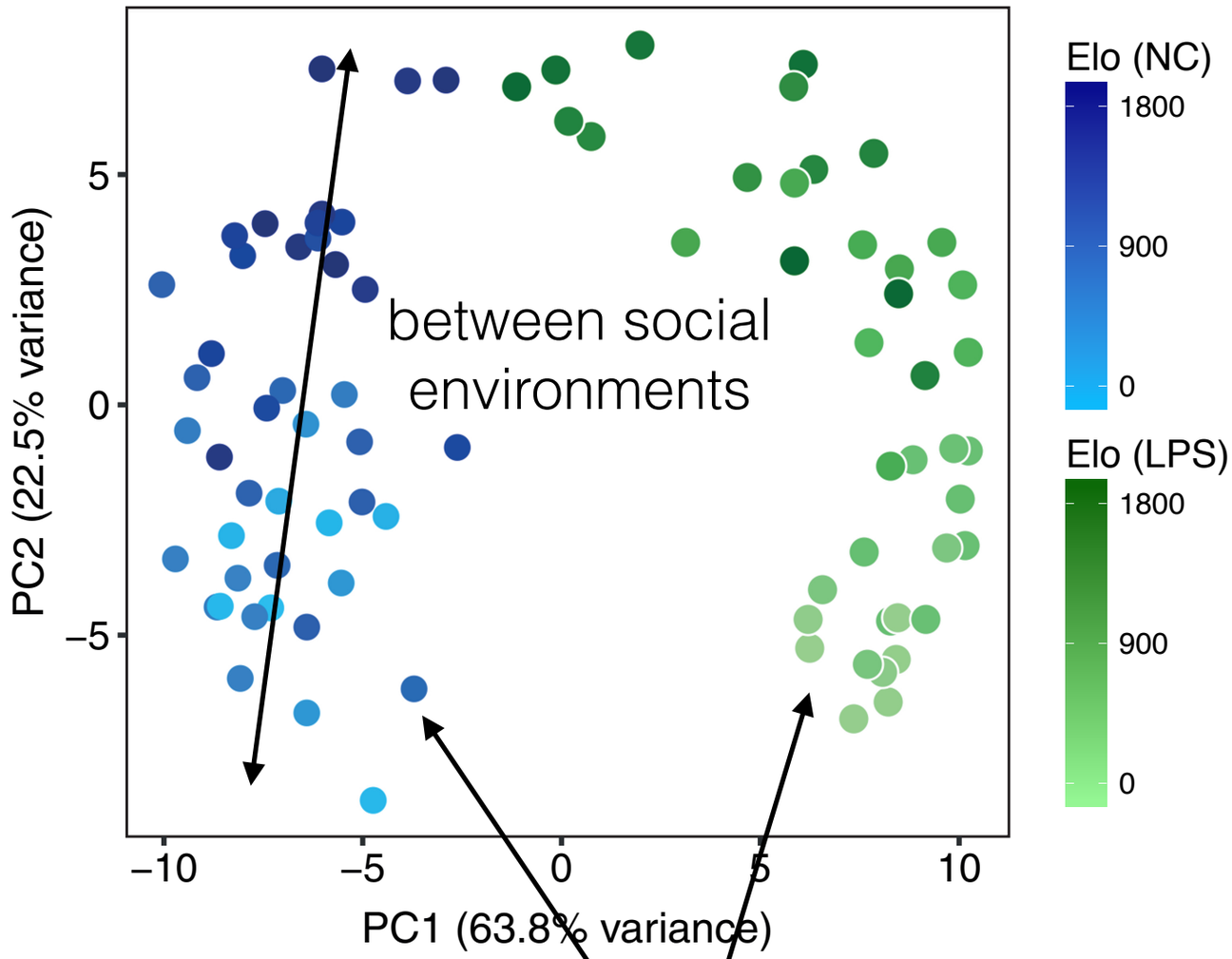
between cell types/tissues

from Brawand et al 2011 (Nature)



between species

Snyder-Mackler et al (unpublished)



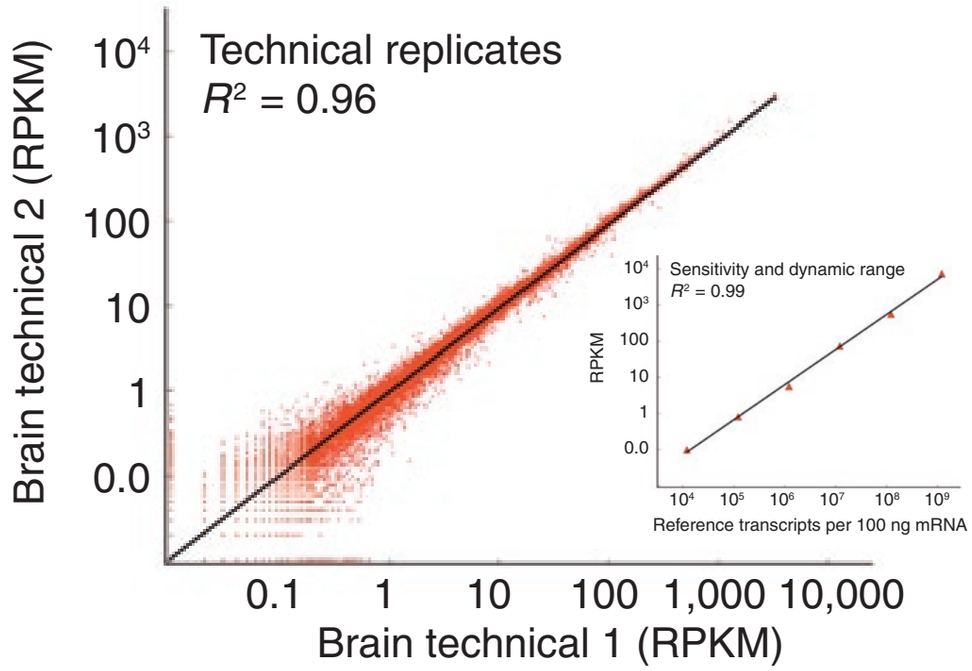
between pathogen conditions

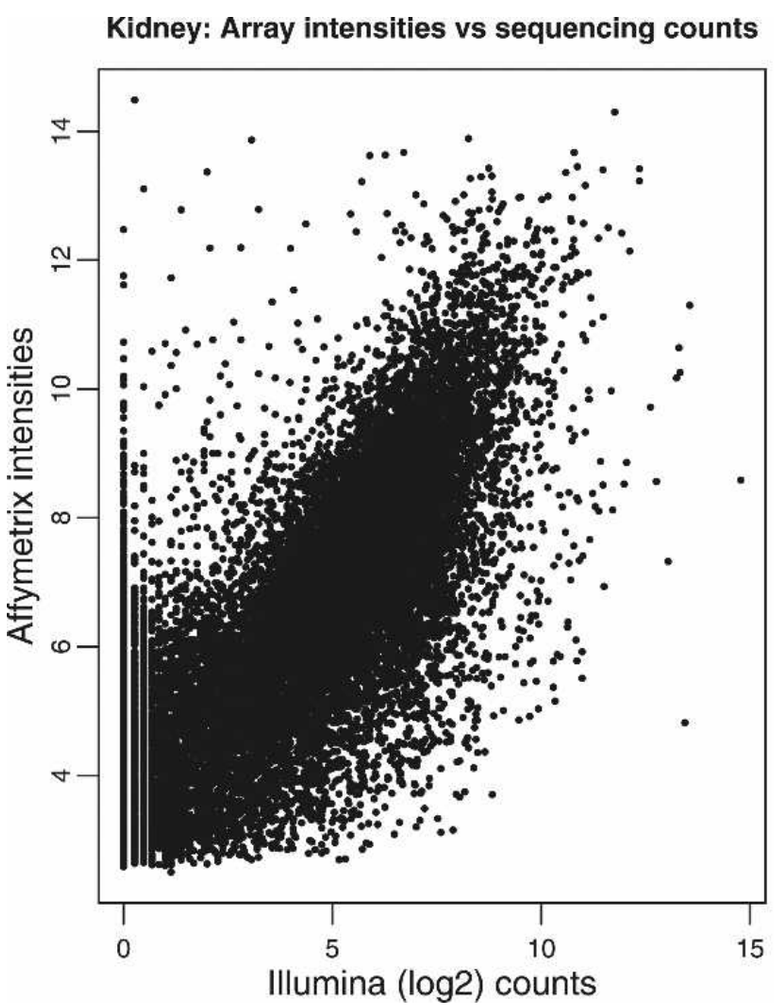
Mapping and quantifying mammalian transcriptomes by RNA-Seq

Ali Mortazavi[1,2], Brian A Williams[1,2], Kenneth McCue[1], Lorian Schaeffer[1] & Barbara Wold[1]

NATURE METHODS | VOL.5 NO.7 | JULY 2008

- suggested use of RNA-seq to revise/refine gene models
- proposed the commonly used “reads per kilobase of [exon/ gene model] per million mapped reads” (RPKM) measure of gene expression levels





Methods

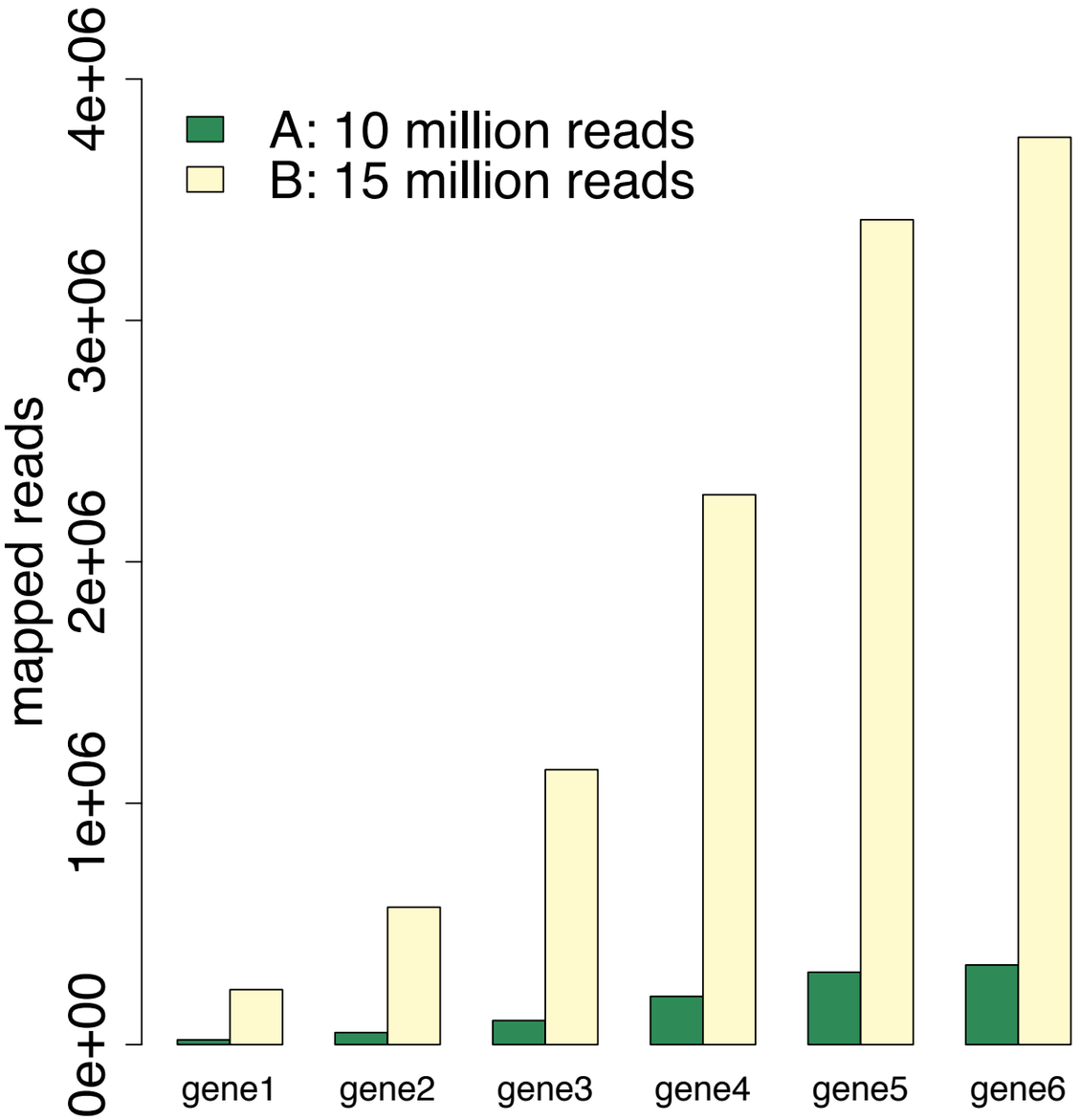
RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays

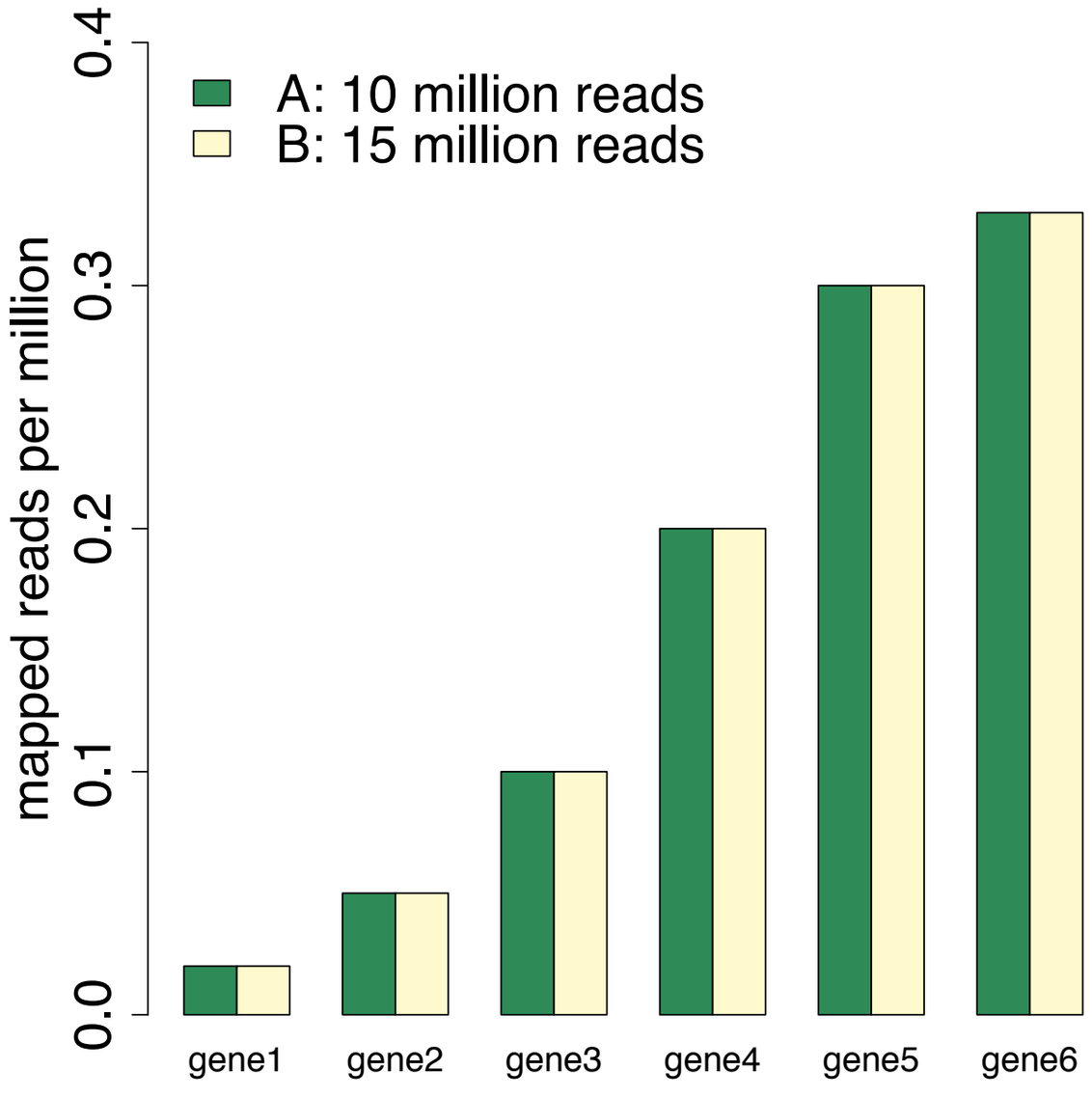
John C. Marioni,[1,6] Christopher E. Mason,[2,3,6] Shrikant M. Mane,[4] Matthew Stephens,[1,5,7] and Yoav Gilad[1,7]

CSH PRESS GENOME RESEARCH

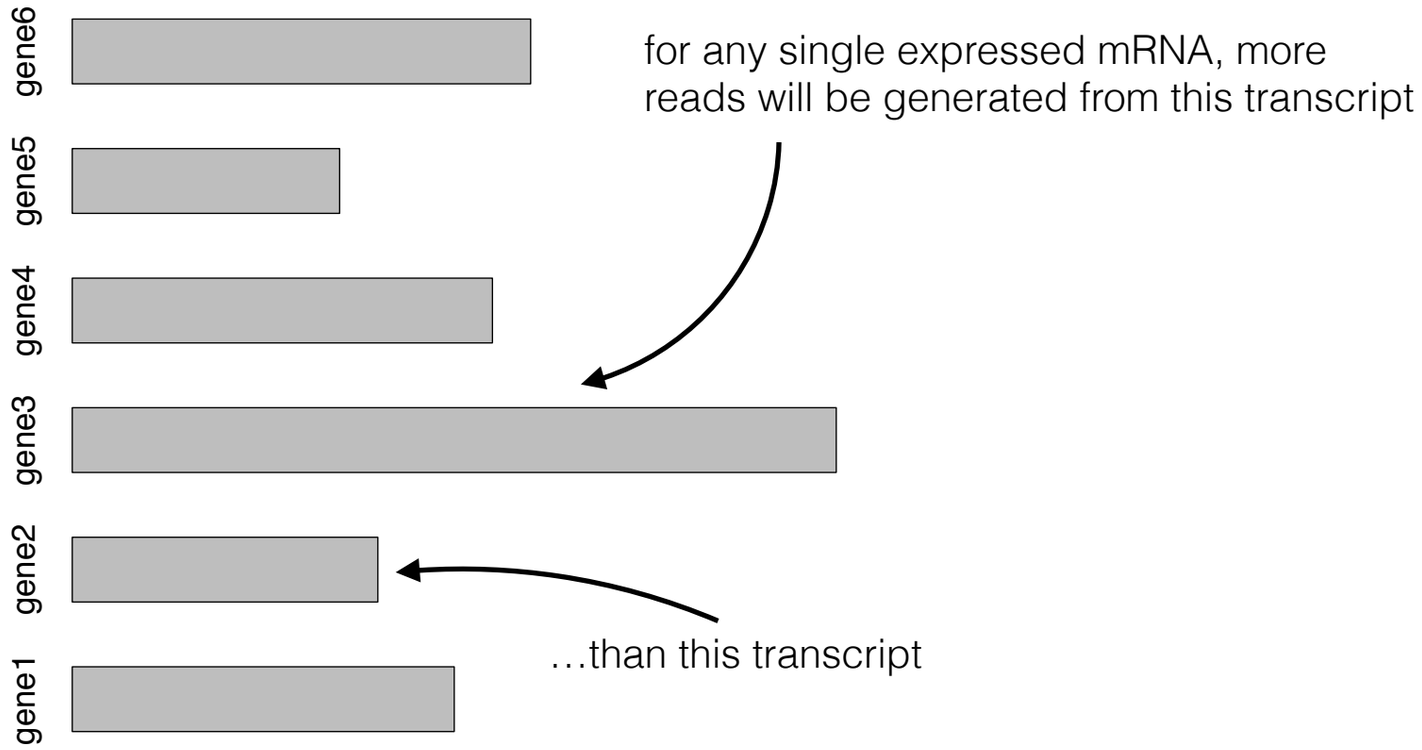
- relatively stable estimates of gene expression; few “lane” effects among technical (not biological) replicates
- good power to detect DE (kidney vs liver) relative to microarrays
- proposed a Poisson generalized linear model for analysis (but also other possibilities, including transformations to continuous data or negative binomial models)

RPKM? FPKM? TPM? CPM?

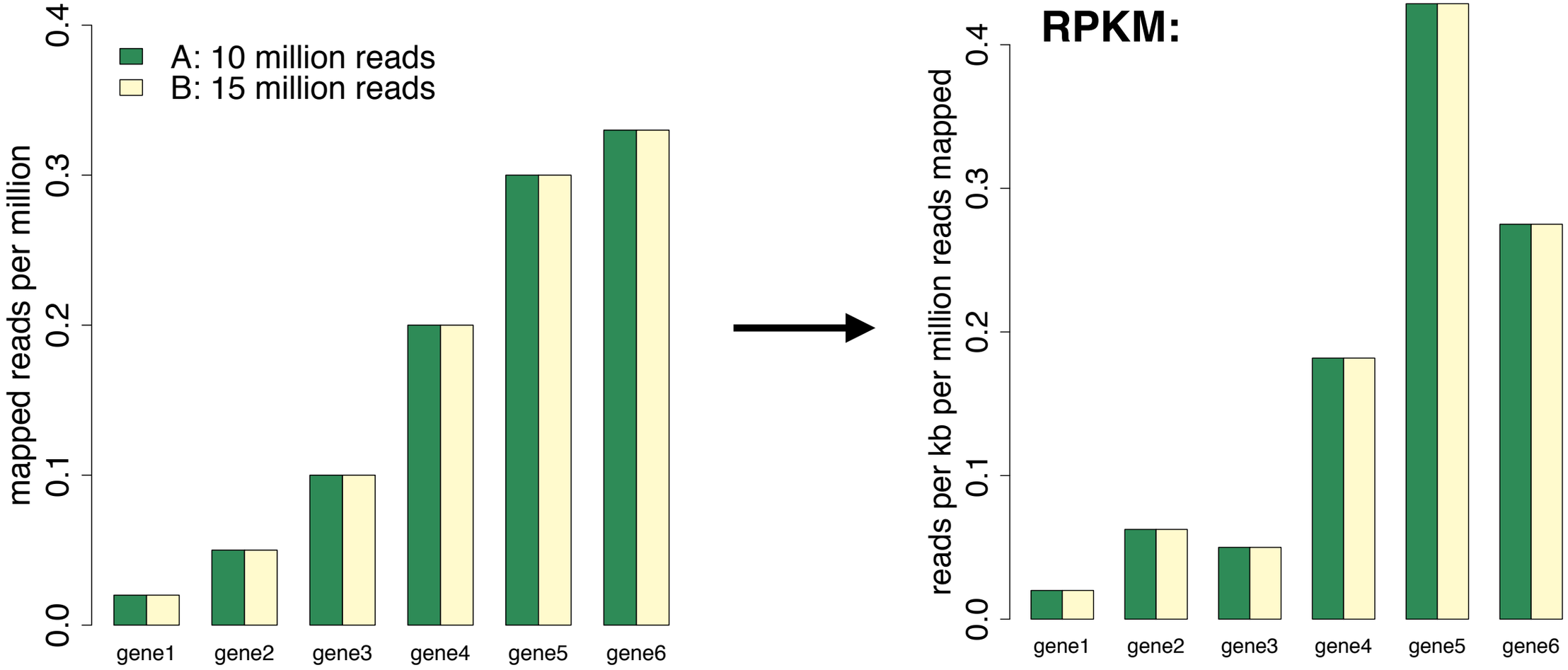




But gene lengths themselves differ:

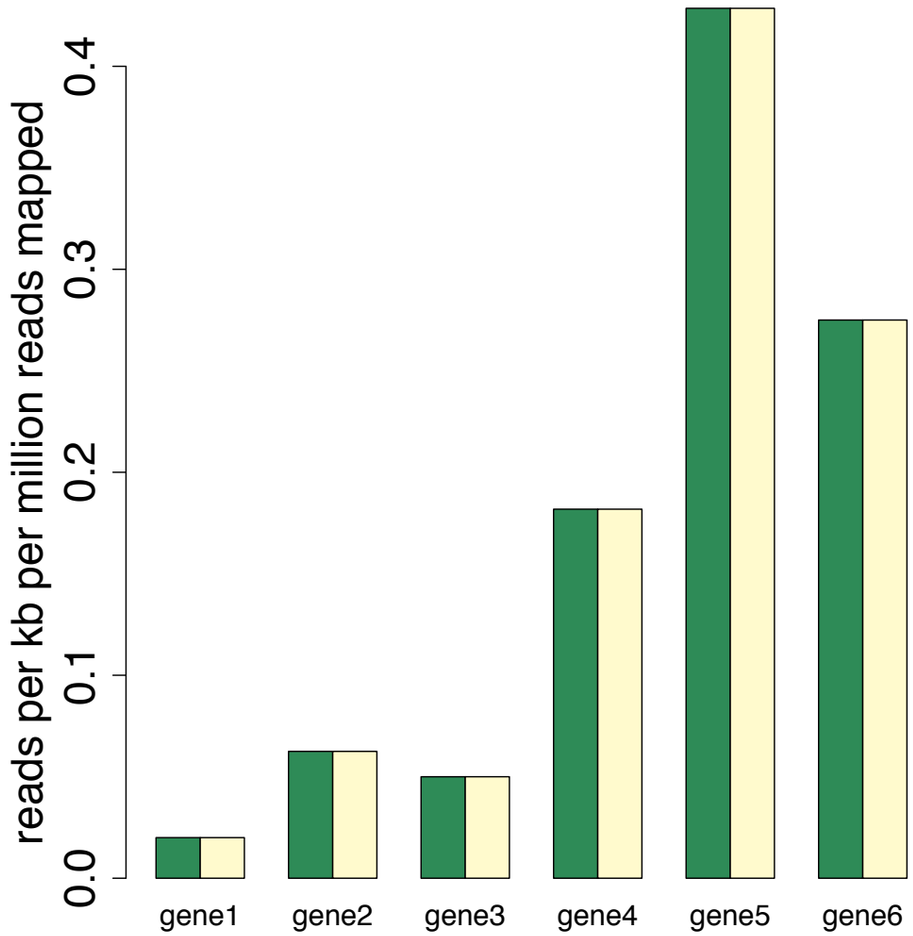


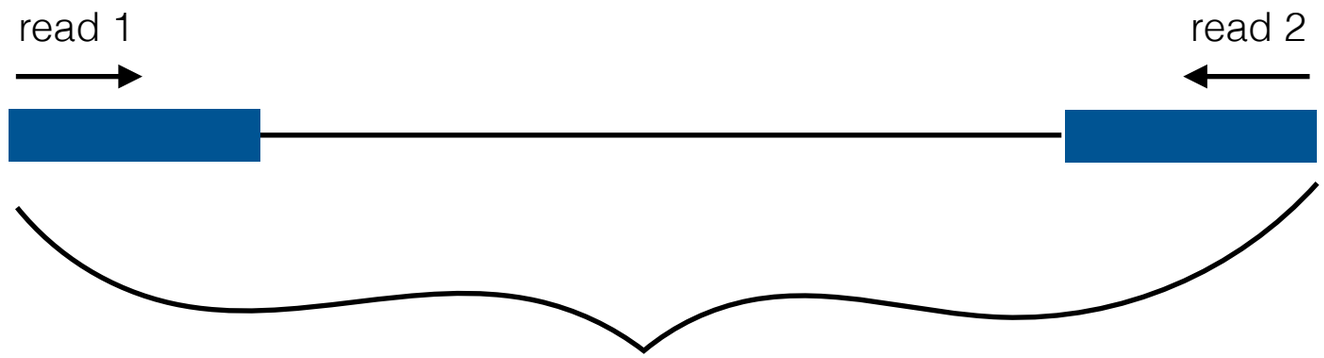
`(tag count  * 1,000,000) / (total number of tags * kilobase of transcript)`



introduced by Trapnell et al 2010, Nature Biotech

RPKM ≈ FPKM





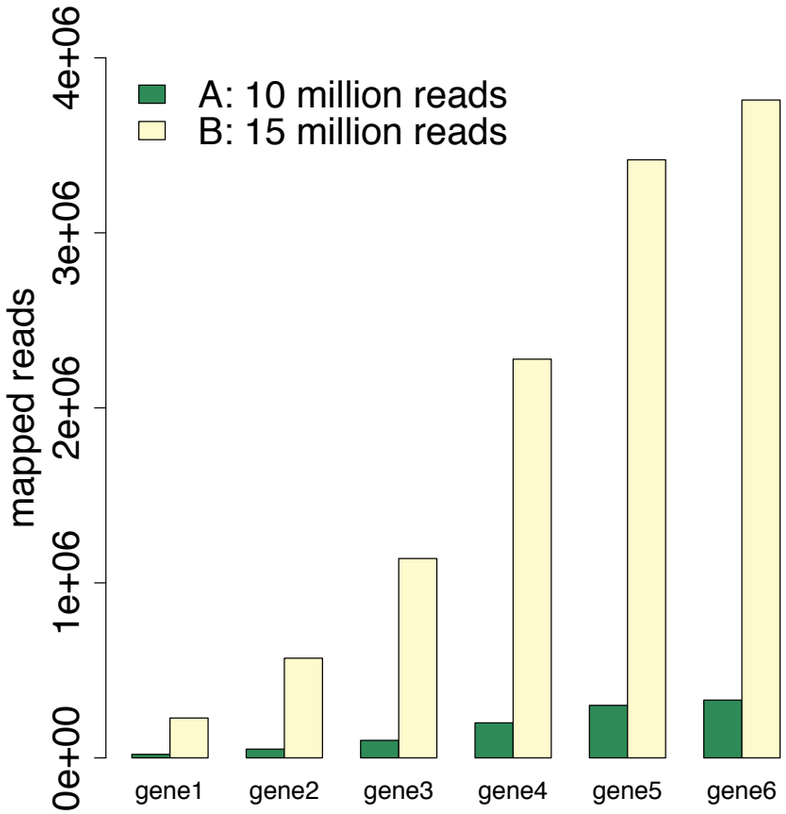
but one fragment

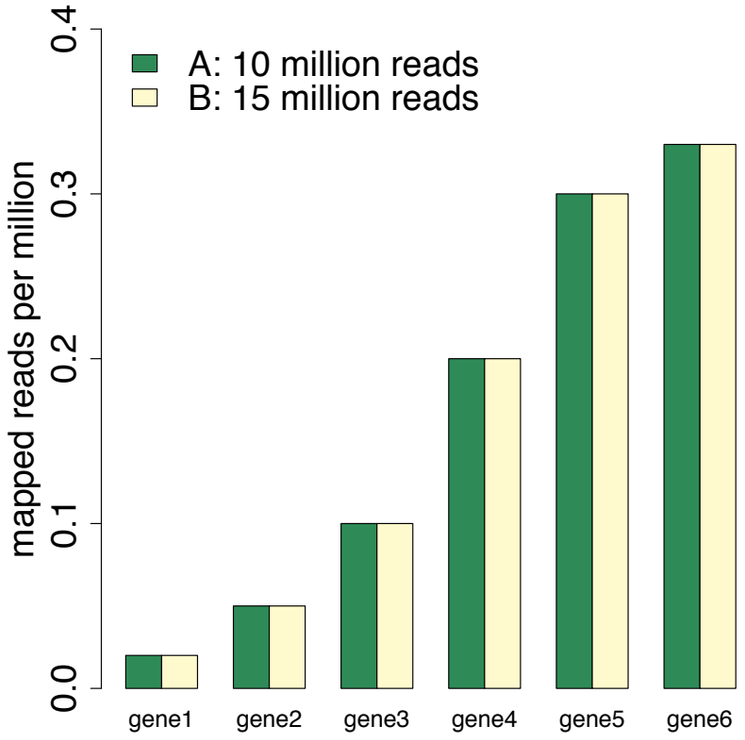
RPKM and FPKM are equivalent for single-end reads

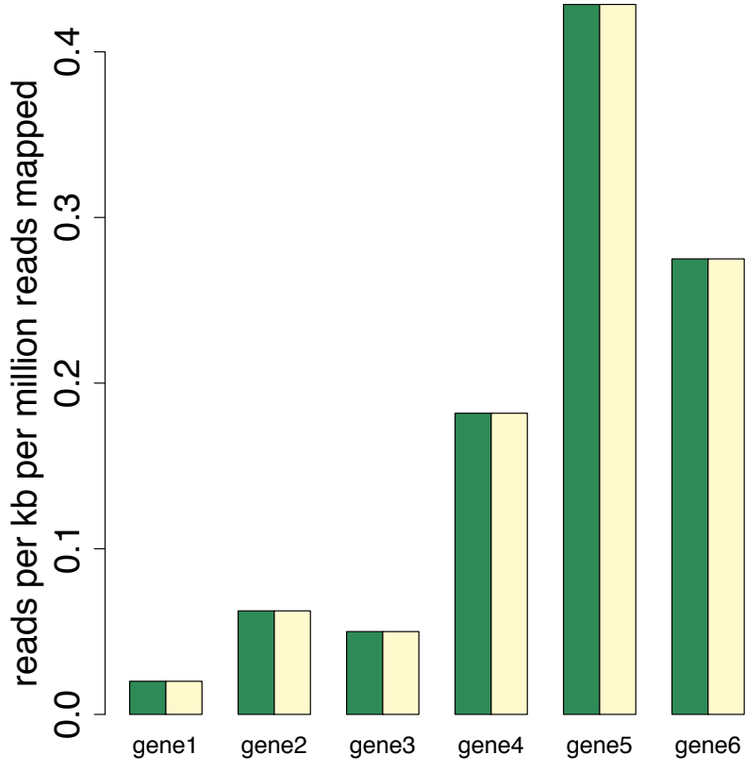
wrong relative expression w/ in sample AND wrong DE between samples

DE artifact from sequencing depth fixed; still wrong relative expression w/in sample

DE artifact from sequencing depth fixed; correct relative expression w/in sample?







RPKM is inconsistent with the expectation that mean normalized abundance across transcripts should be equal for each sample (can be fixed using **"TPM"** methods: divide by sum of *length-normalized* read counts for all transcripts instead: Li et al 2010, Wagner et al 2012) *BUT ALWAYS KEEP IN MIND WHAT YOU'RE ACTUALLY TRYING TO ANALYZE/COMPARE)!

*Resources: Lior Pachter keynote at CSHL Genome Informatics 2013 (YouTube), NextGenetics blog post: http://blog.nextgenetics.net/?e=51

2010

2011

DESeq (Anders & Huber, Genome Biology)

edgeR (Robinson et al, Bioinformatics)

BaySeq (Hardcastle & Kelley, BMC Bioinformatics)

NBPseq (Di et al, Stat Appl Gen Mol Bio)

TSPM (Auer & Doerge, Stat Appl Gen Mol Bio)

2012

2013

PoissonSeq (Li et al, Biostatistics)

cuffdiff (Trapnell et al, Nature Protocols)

NOIseq (Tarazona et al, EMBNet Journal)

ShrinkSeq (Van de Wiel, Biostatistics)

EBseq (Leng et al, Bioinformatics)

SAMseq (Li and Tibshirani, Stat Met Med Research)

2014

2015

2016

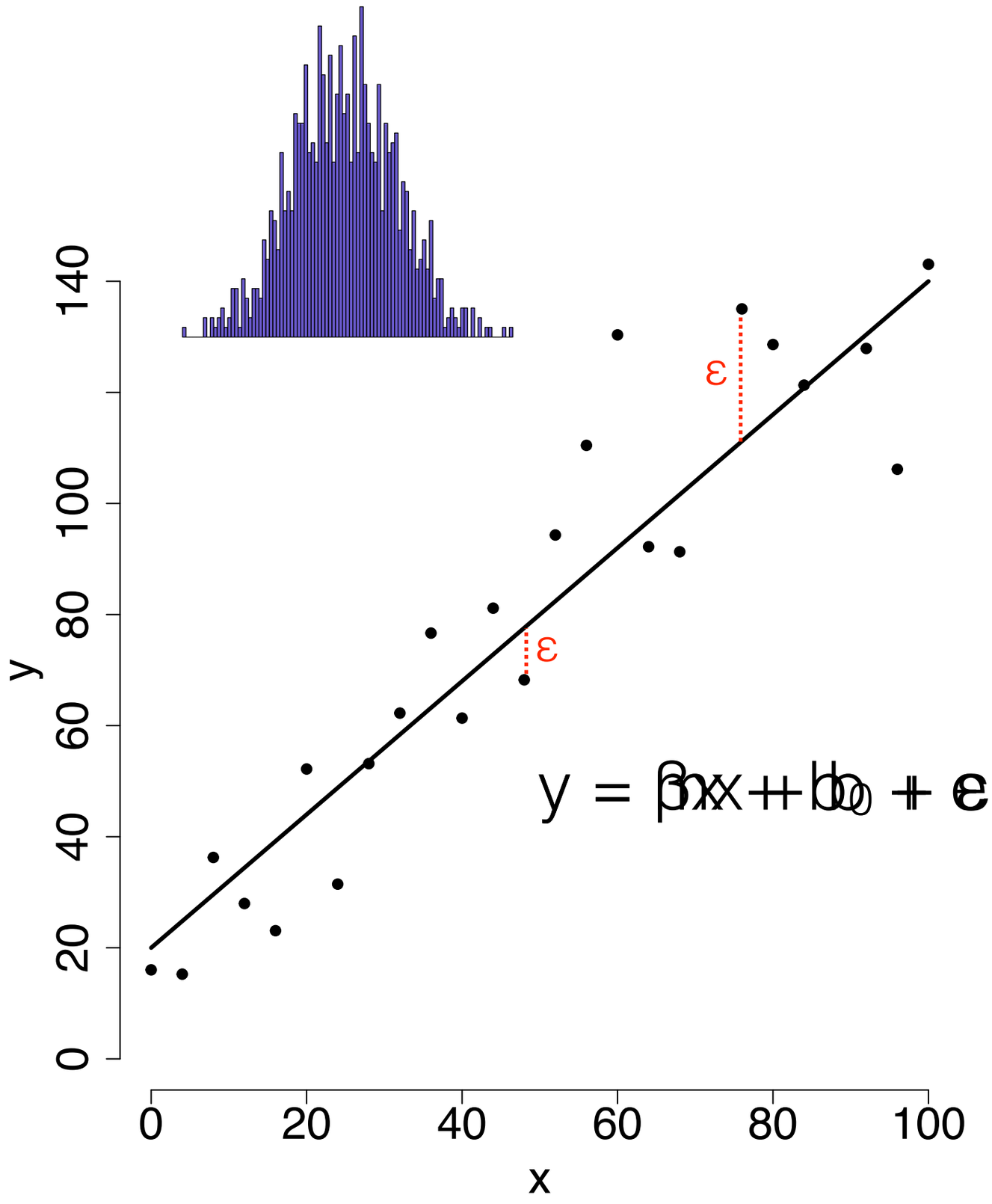
DESeq2 (Love et al, Genome Biology)

limma+voom (Law et al, Genome Biology; building on **limma**: Smyth 2005)

limma (Ritchie et al, Nuc Acids Research)

sleuth (Pimentel et al, bioRxiv)

MACAU (Sun et al, bioRxiv)


y
0
20
40
60
80
100
140
0
20
40
60
80
100
x
ε
ε
y = β1x + b0 + e

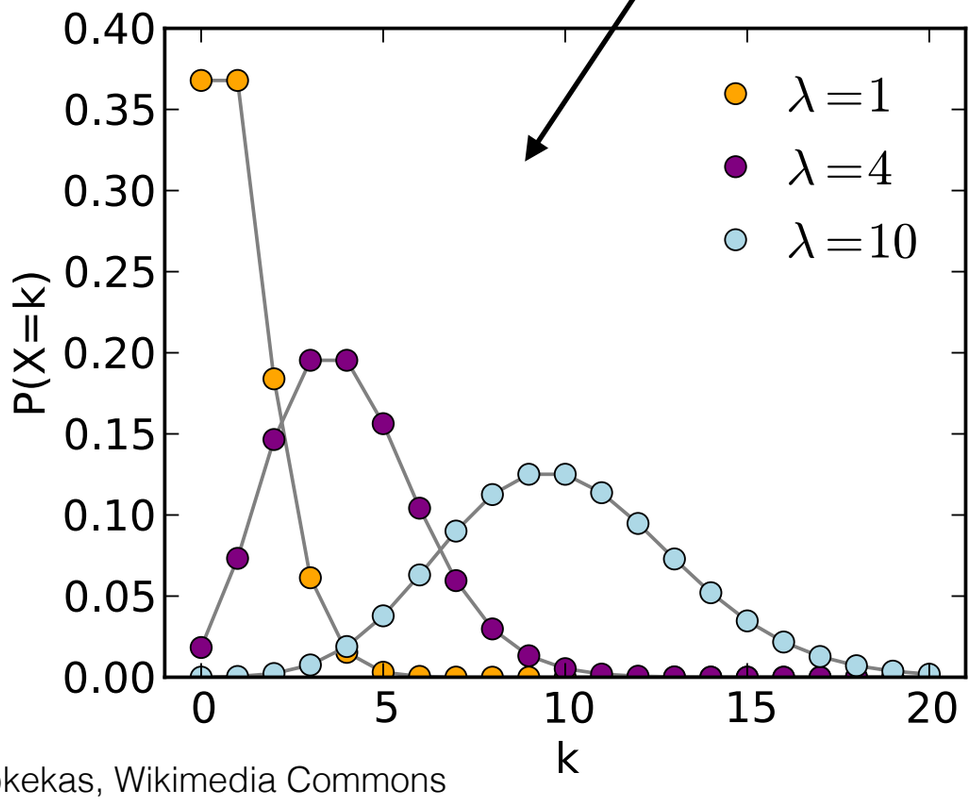
Some issues to consider

*not including library prep (stranded? PCR-free? Nextera/Tru-seq/other?), mapping (which aligner? what QC thresholds?), genome annotation, etc.

	g1	*g2*	*g3*	. . .	*gp*
s1	4	10	15	8	20
s2	8	. . .			
⋮	⋮				
⋮	⋮				
⋮	⋮				
sn					

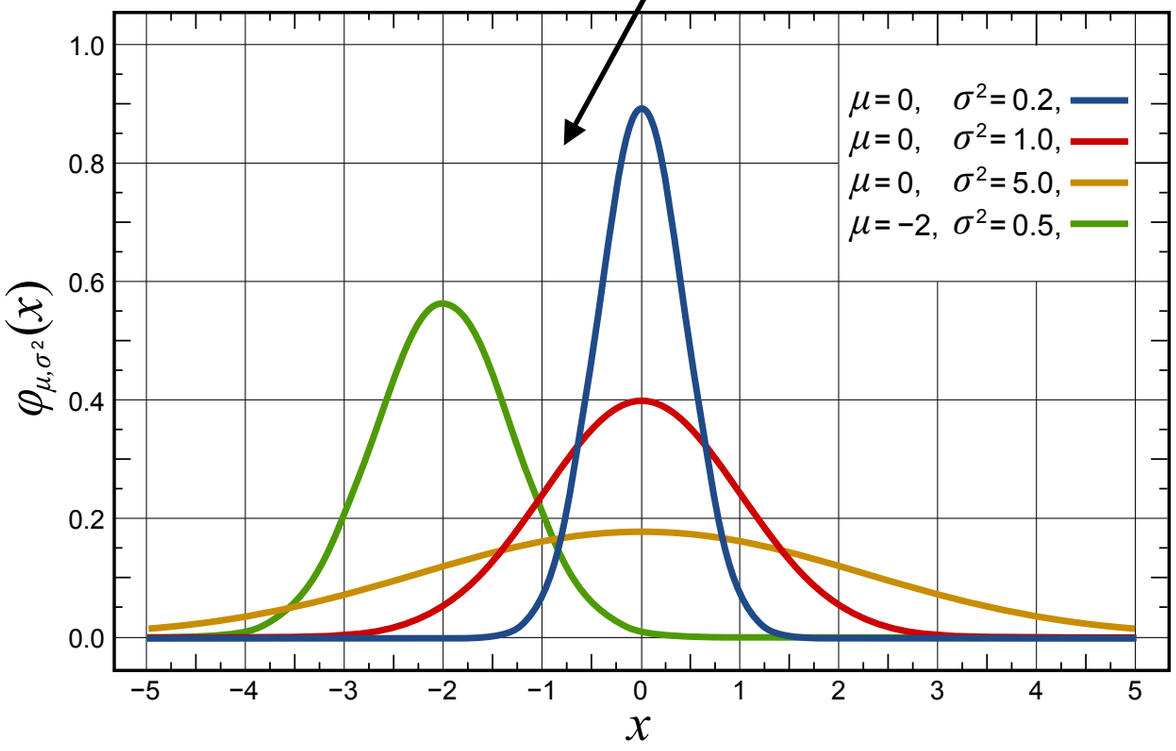
Your data are counts: keep them as counts or transform?

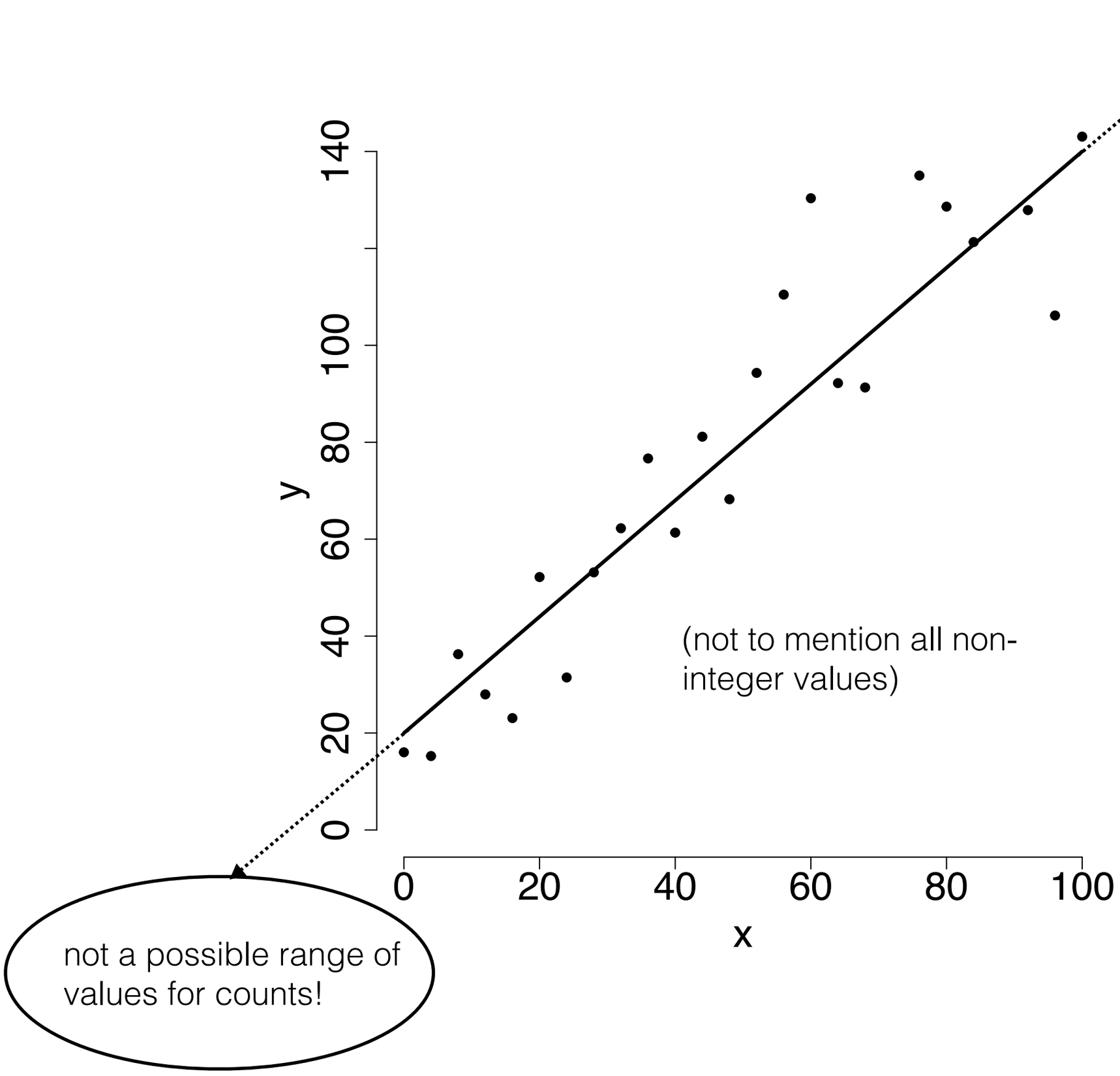
non-negative integers only; less flexible, but a "true" representation of the data structure



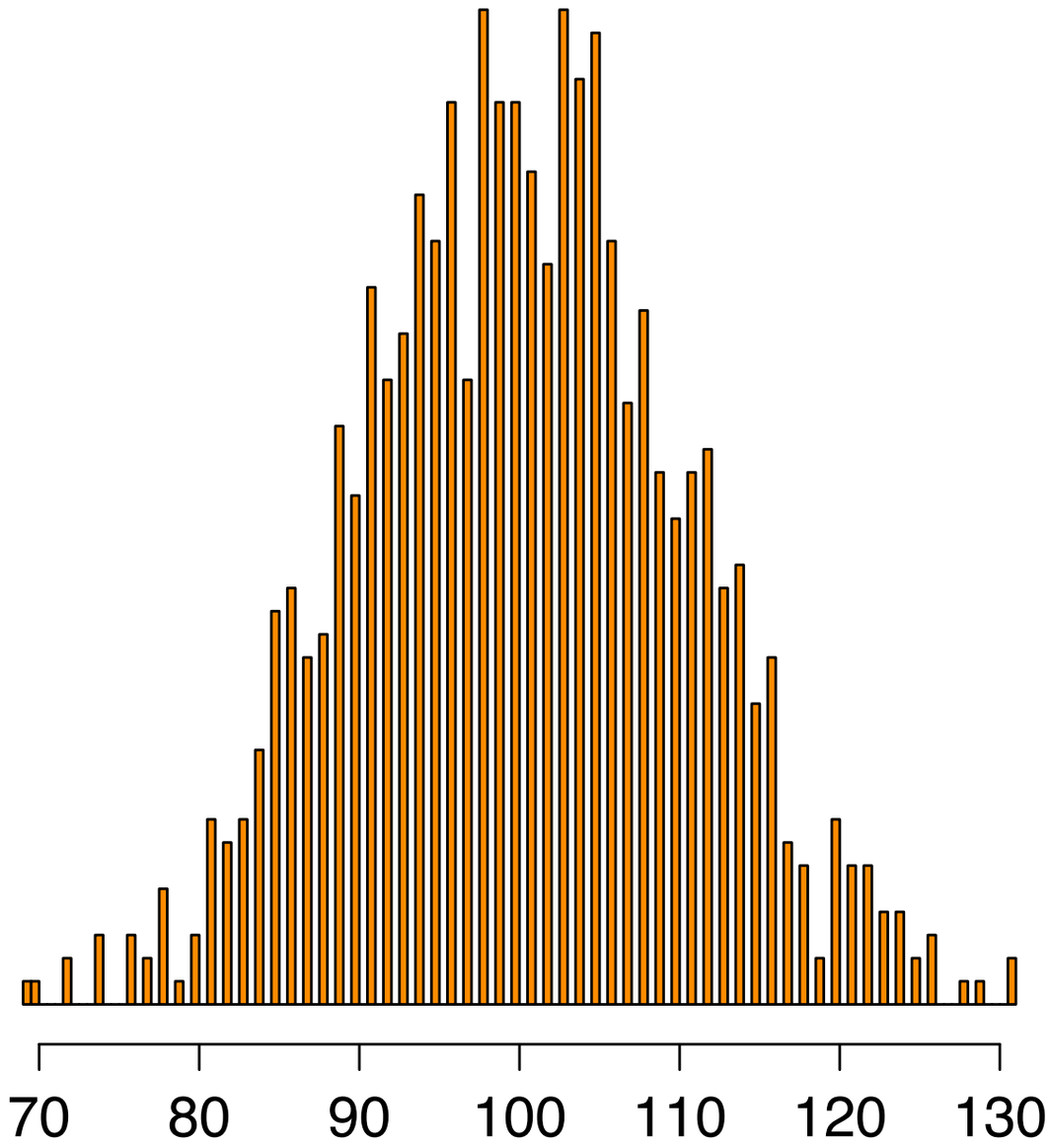
Skbkekas, Wikimedia Commons

continuous distribution; more flexible (large set of developed tools), but possible information loss



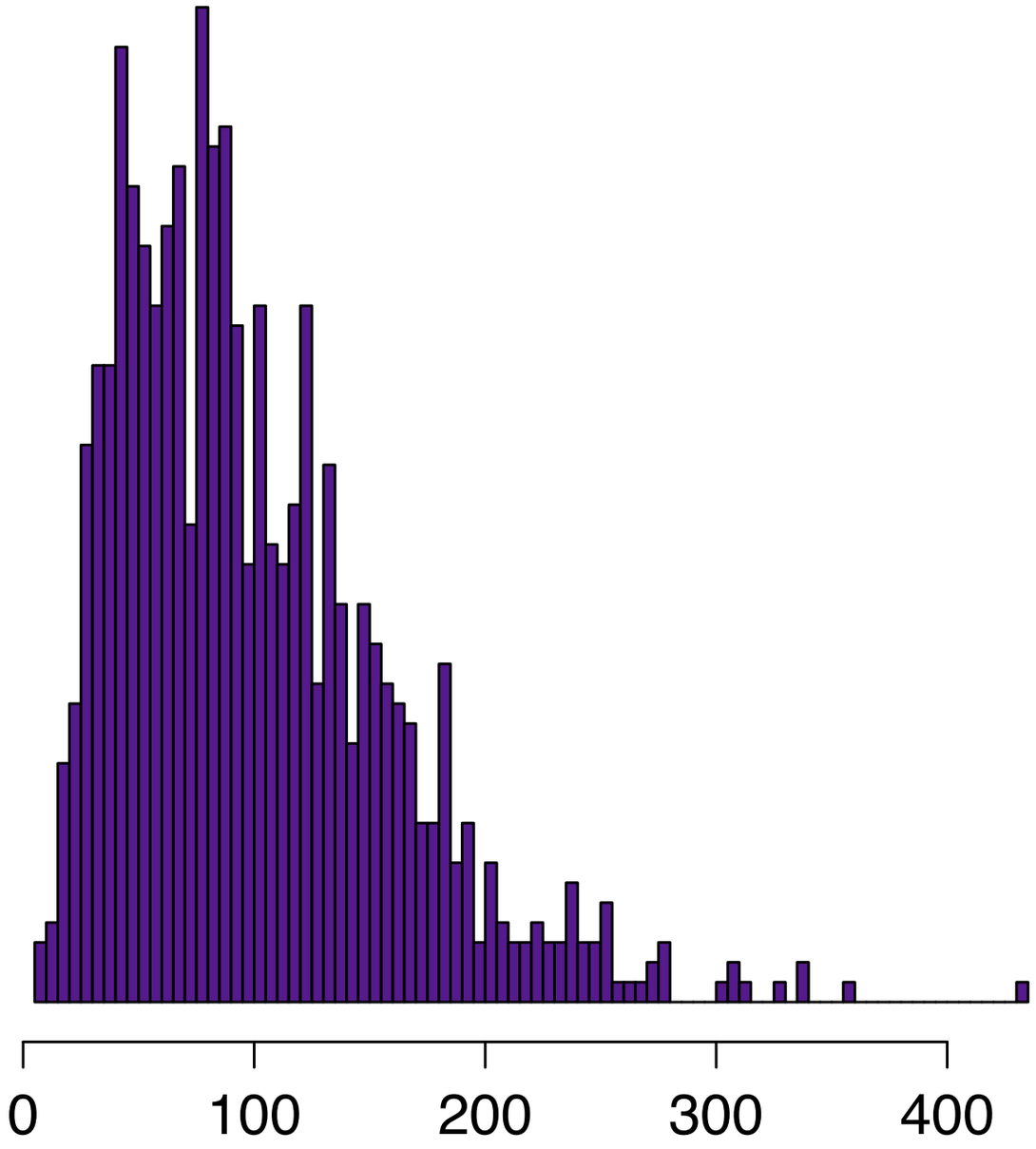

y
0
20
40
60
80
100
140
0
20
40
60
80
100
x
(not to mention all non-integer values)
not a possible range of values for counts!

Poisson distribution: all non-negative integers



(mean and variance are controlled by a single parameter)

negative binomial distribution: all non-negative integers

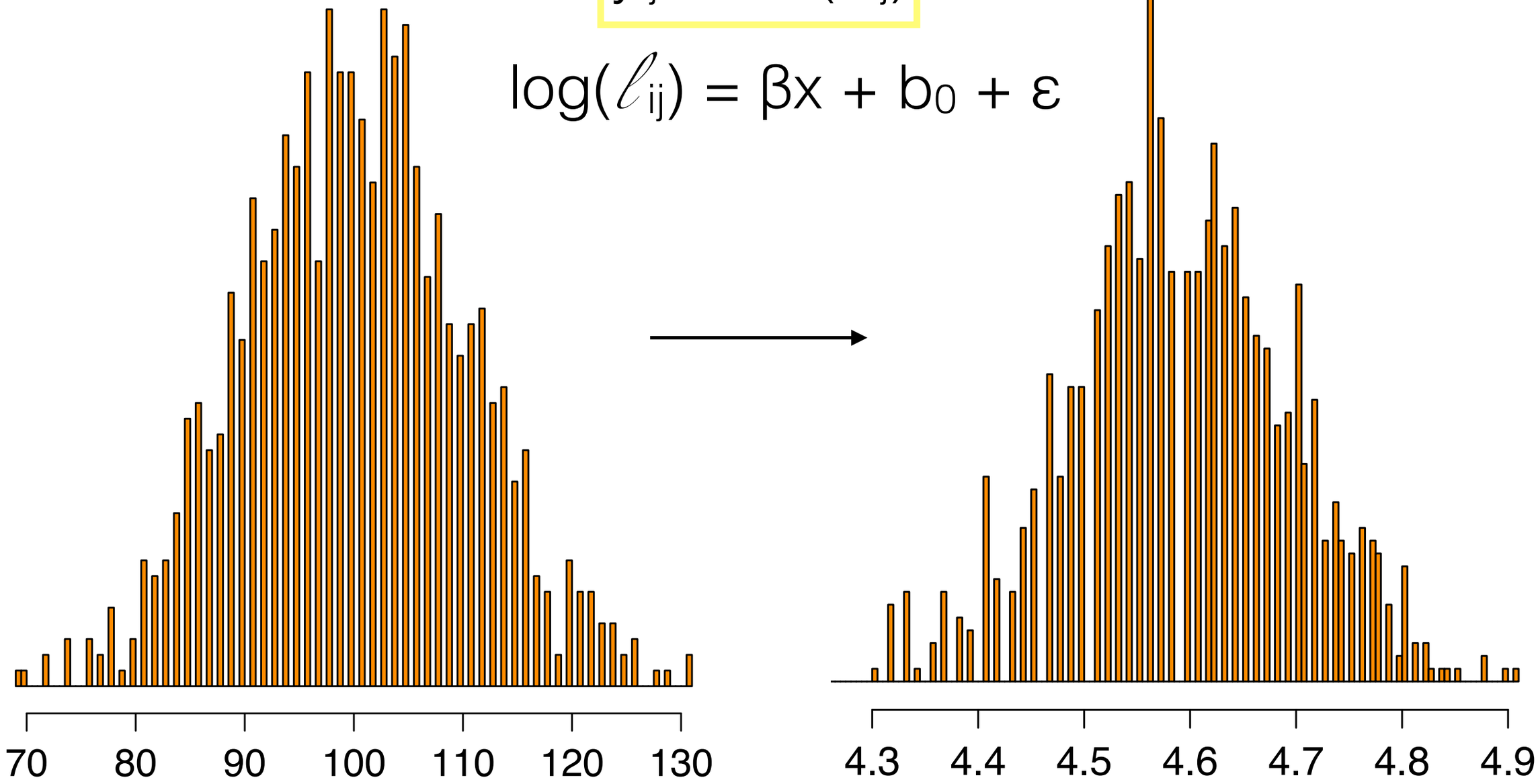


(mean and variance are controlled by separate parameters, variance scales with the mean)

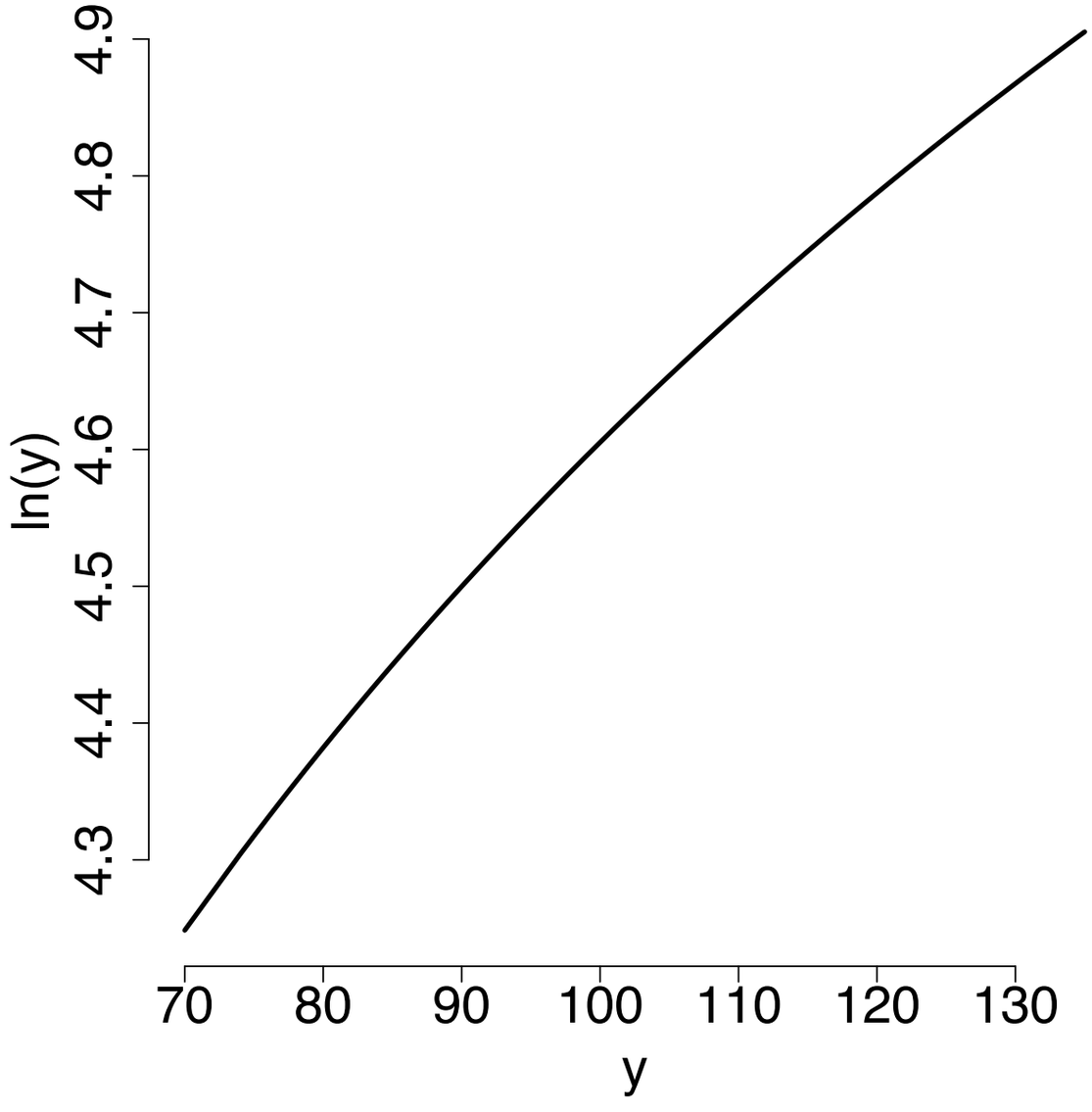
for individual *i* and gene *j*:

$$y_{ij} = \text{Poi}(\ell_{ij})$$

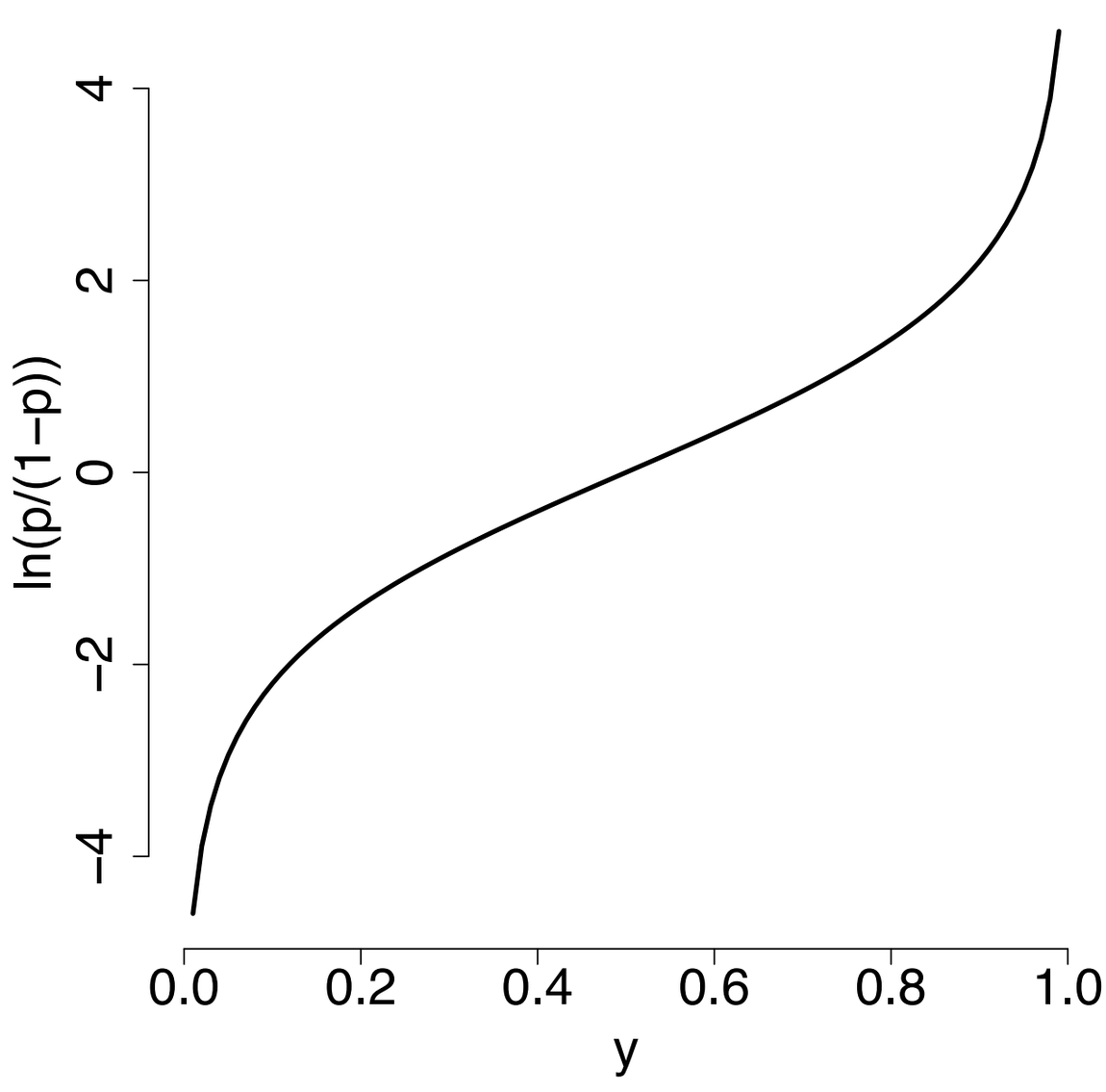
$$\log(\ell_{ij}) = \beta x + b_0 + \varepsilon$$



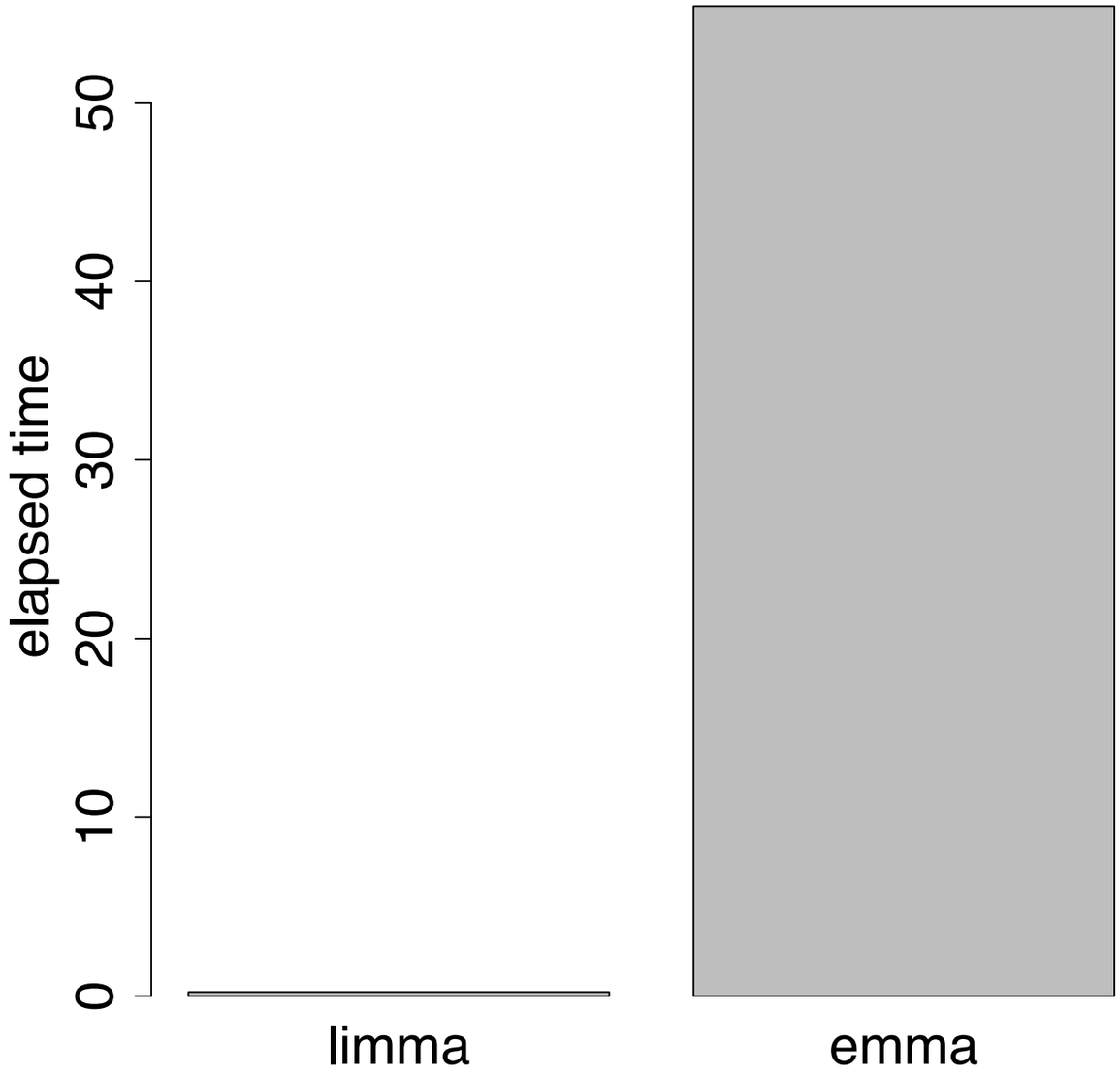
log link (typical for GLMs with Poisson distributed data)



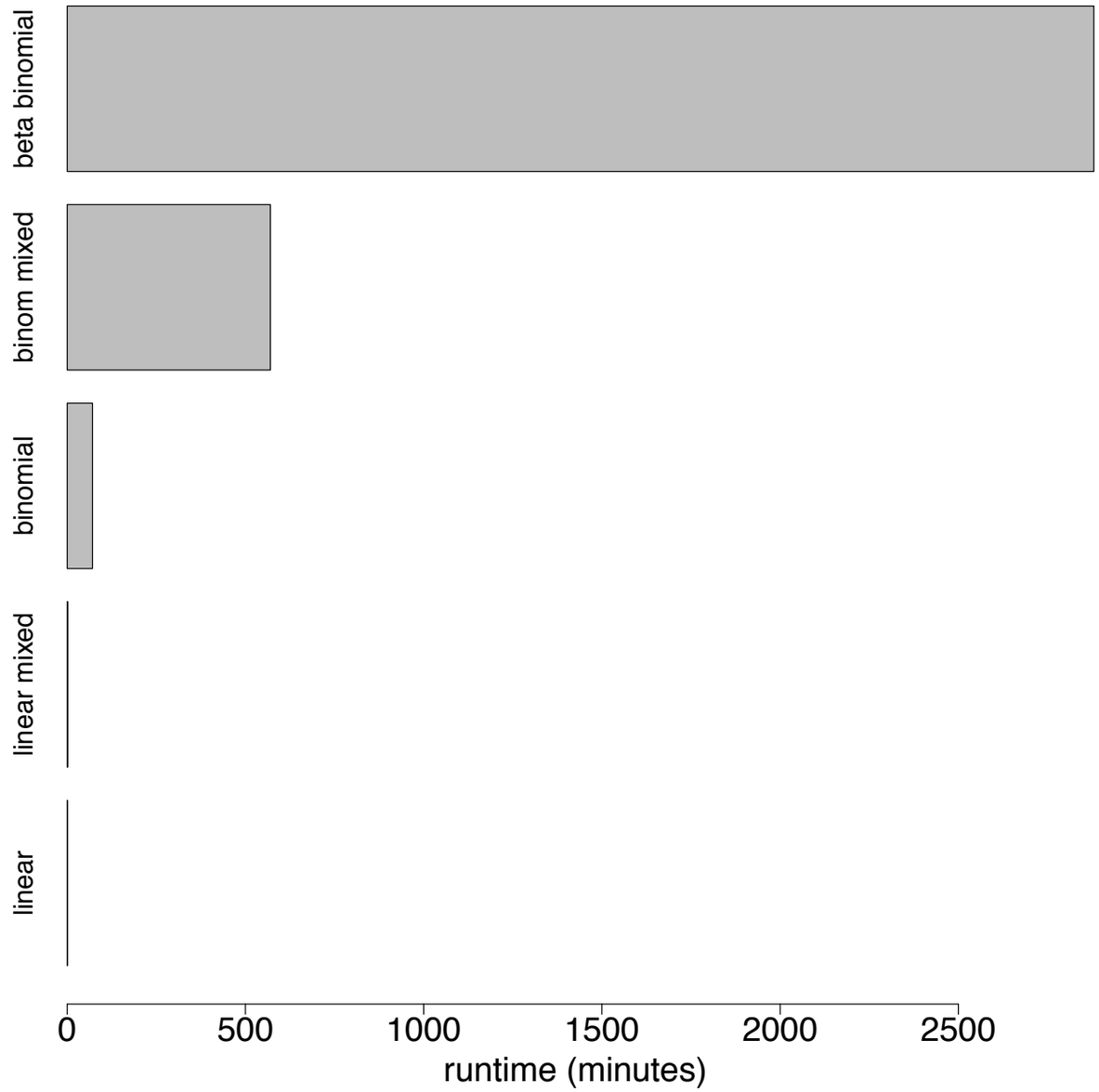
logit link (typical for GLMs with binomial distributed data)



time to run the same data set (n=20 RNA-seq samples) on my computer (1.4 GHz Intel Core i5 MacBook Air)



n=24 *Arabidopsis* WGBS samples, 830k sites, single Intel Xeon L5420 2.50 GHz core)



(Lea et al 2015, PLoS Genetics)

Some issues to consider

*not including library prep (stranded? PCR-free? Nextera/Tru-seq/other?), mapping (which aligner? what QC thresholds?), genome annotation, etc.

Your data are counts: keep them as counts or transform?

Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data

Franck Rapaport[1], Raya Khanin[1], Yupu Liang[1], Mono Pirun[1], Azra Krek[1], Paul Zumbo[2,3], Christopher E Mason[2,3], Nicholas D Socci[1] and Doron Betel[3,4*]

Genome Biology

A comparison of methods for differential expression analysis of RNA-seq data

Charlotte Soneson[1*] and Mauro Delorenzi[1,2]

BMC Bioinformatics

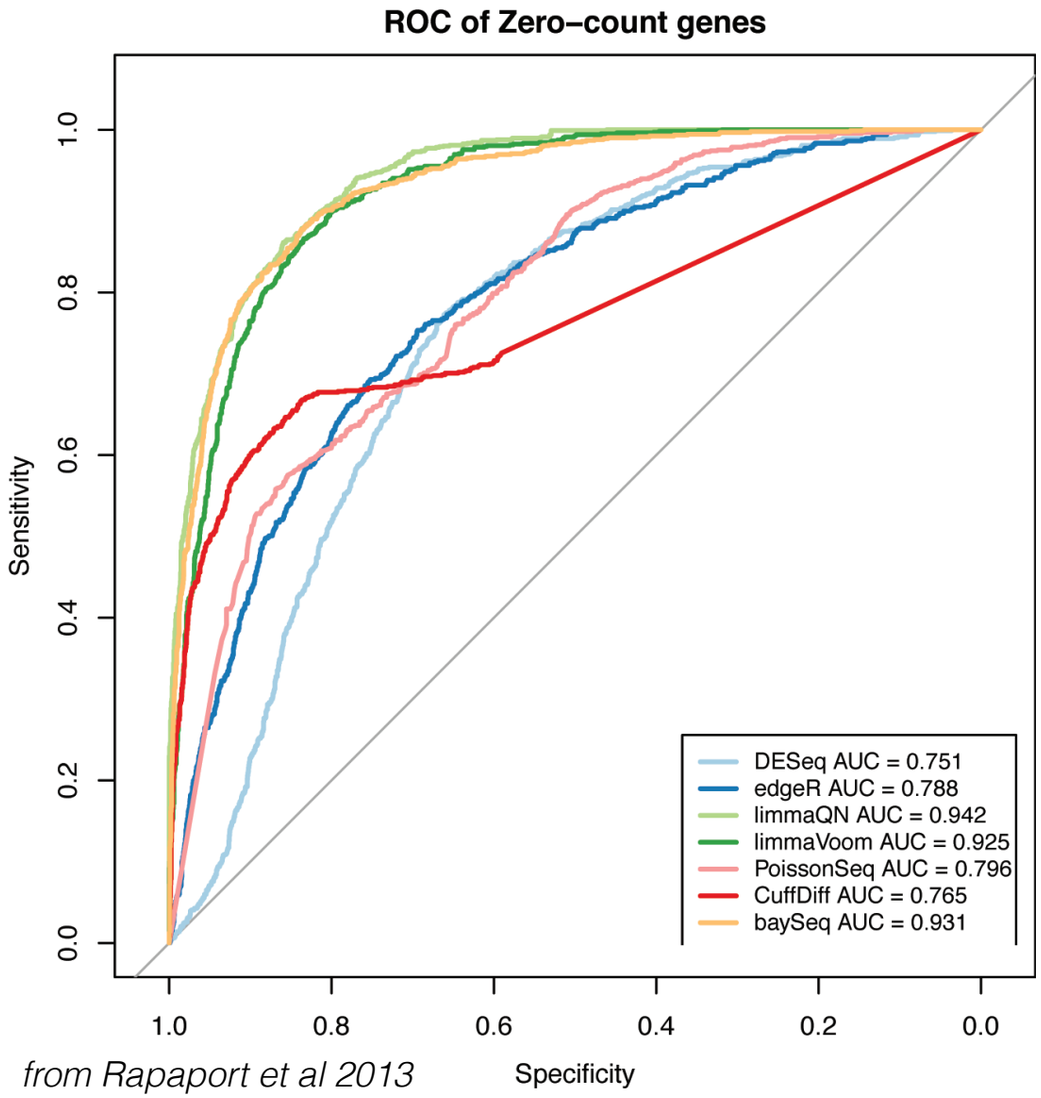
(and others . . .)

(based on technical replicates—no DE genes expected)

Table 1 Number of false differential expression genes predicted by each method at adjusted *P* values (or false discovery rate) ≤0.05 separated by gene read count quantiles.

Expression quantile	Cuffdiff	DESeq	edgeR	limmaQN	limmaVoom	PoissonSeq	baySeq
100% (high expression)	28	5	3	0	0	7	1
75%	76	6	0	0	0	0	0
50%	84	27	1	2	0	0	0
25% (low expression)	5	9	0	87	0	0	0
Total	193	47	4	89	0	7	1

from Rapaport et al 2013



from Rapaport et al 2013

"We find significant differences among the methods, but note that array-based methods adapted to RNA-seq perform comparably to methods designed for RNA-seq" (Rapaport et al 2013)

"Among the methods evaluated…those based on a variance-stabilizing transformation combined with limma (i.e., voom+limma and vst+limma) performed well under many conditions, were relatively unaffected by outliers and were computationally fast" (Soneson and Delorenzi 2013)

Lior Pachter @lpachter · Sep 9

@drchriscole @BioMickWatson In the words of the author of both: "[**voom**] wins all comparisons with other methods"

Bioinformatics Seminar – Professor Gordon Smyth ...

Gordon is well known for his development of the limma Bioconductor package for the analysis of differential gene expression using microarrays. More recently his g...

statsandgenomes.wordpress.com

6 12

In reply to Chris Cole

Lior Pachter @lpachter · Sep 9

@drchriscole @BioMickWatson E.g. in every benchmark I've done/seen over the past few years edgeR performs worse than **voom** (by same PI).

2 3

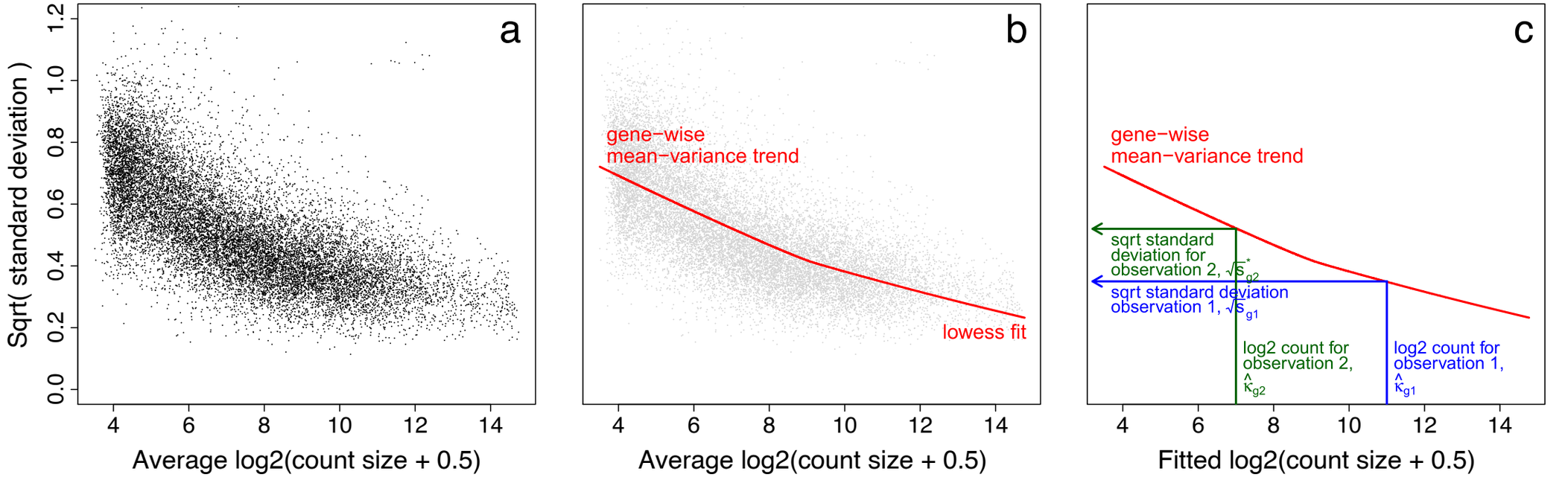
voom (+ limma)

e.g., TMM normalized

1) log transformation

$$y = \log_2(\text{count} + 0.5)/(\text{normalized library size} + 1) \times 1e6$$

2) precision weights (in limma)



from Law et al 2014

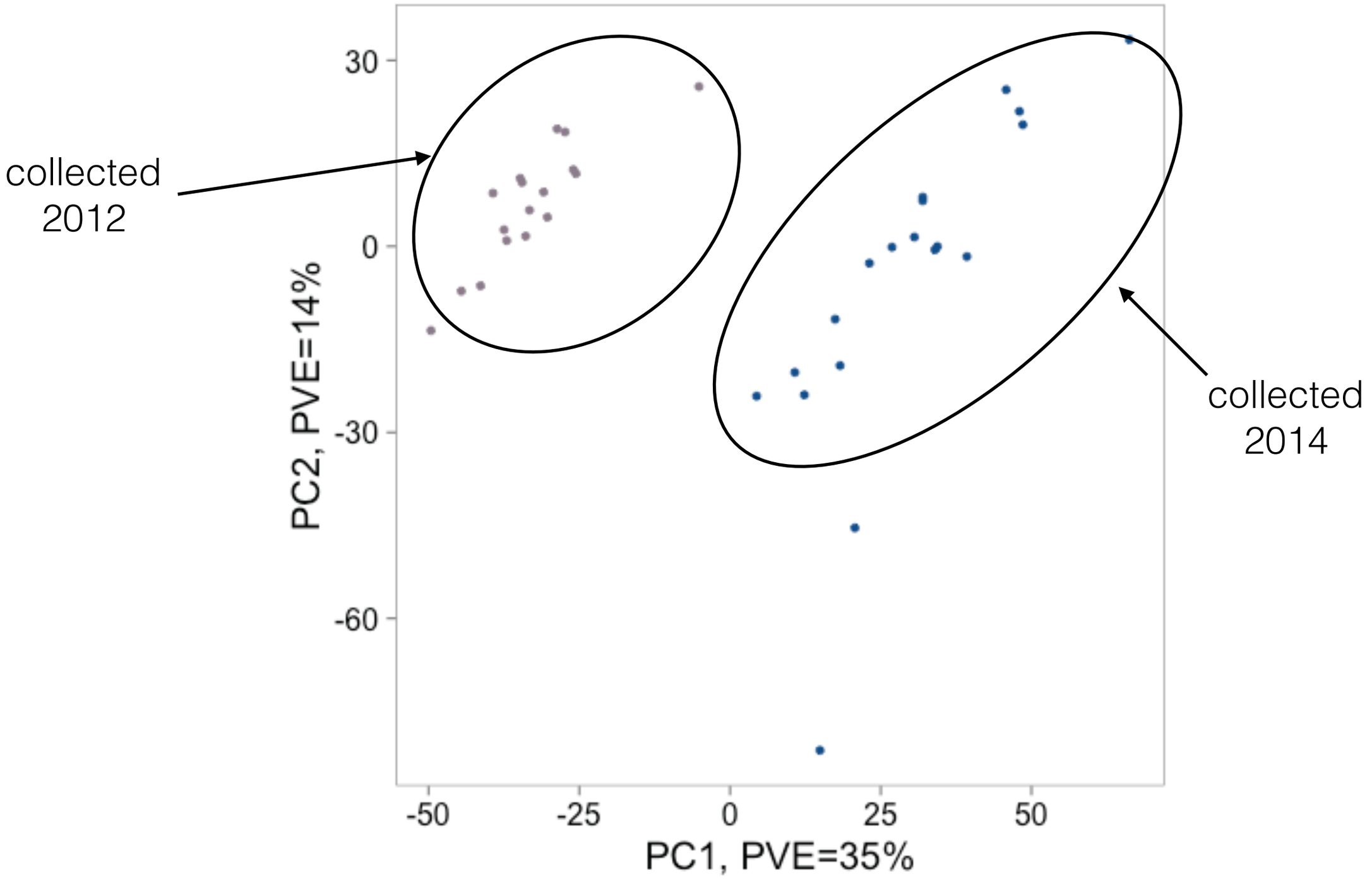
Some issues to consider

*not including library prep (stranded? PCR-free? Nextera/Tru-seq/other?), mapping (which aligner? what QC thresholds?), genome annotation, etc.

Your data contain batch effects.

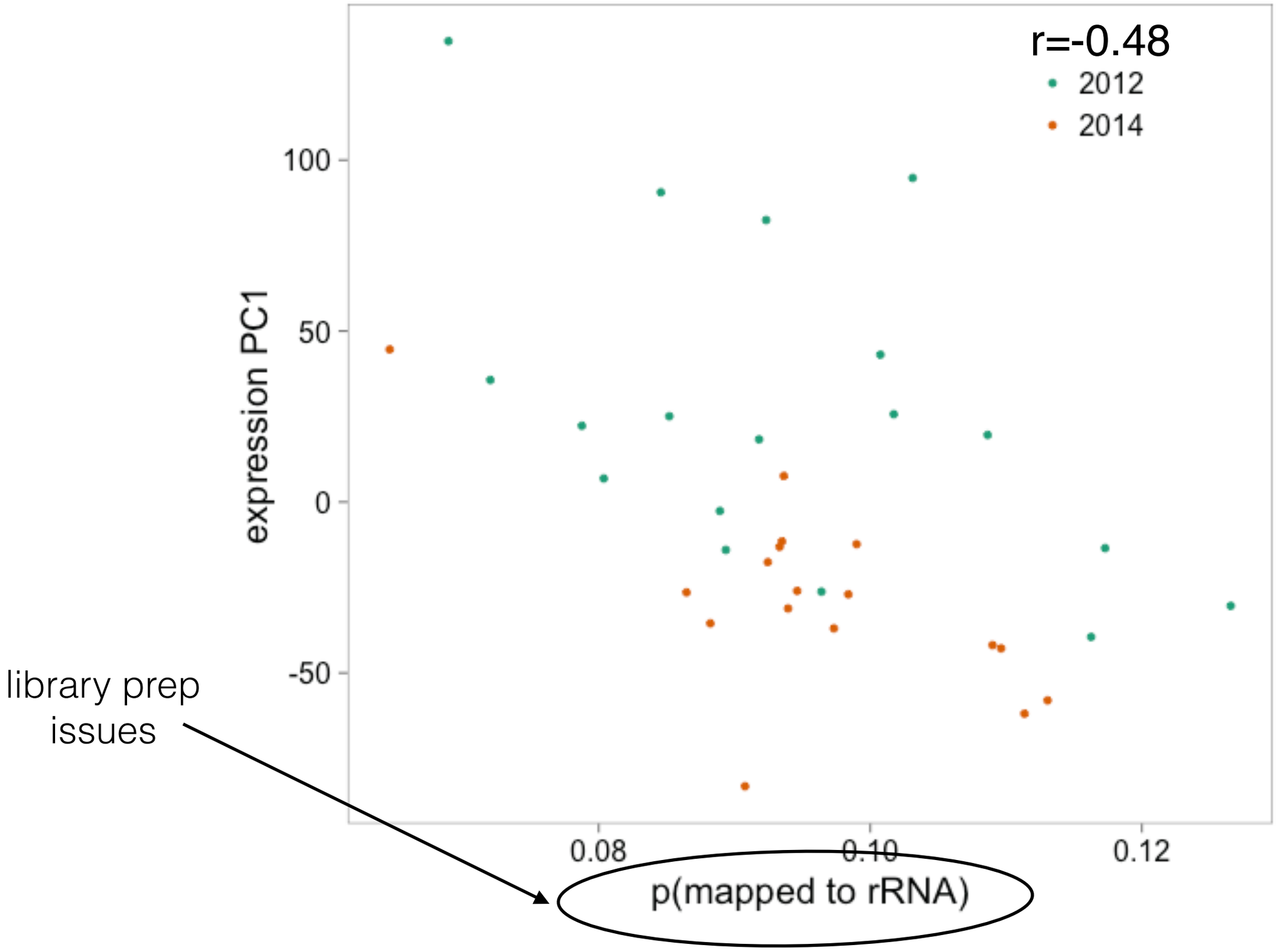
(I haven’t seem them, but I promise they do)

Pre/post-dispersal gene expression in male rhesus macaques



Snyder-Mackler and Tung, unpublished data

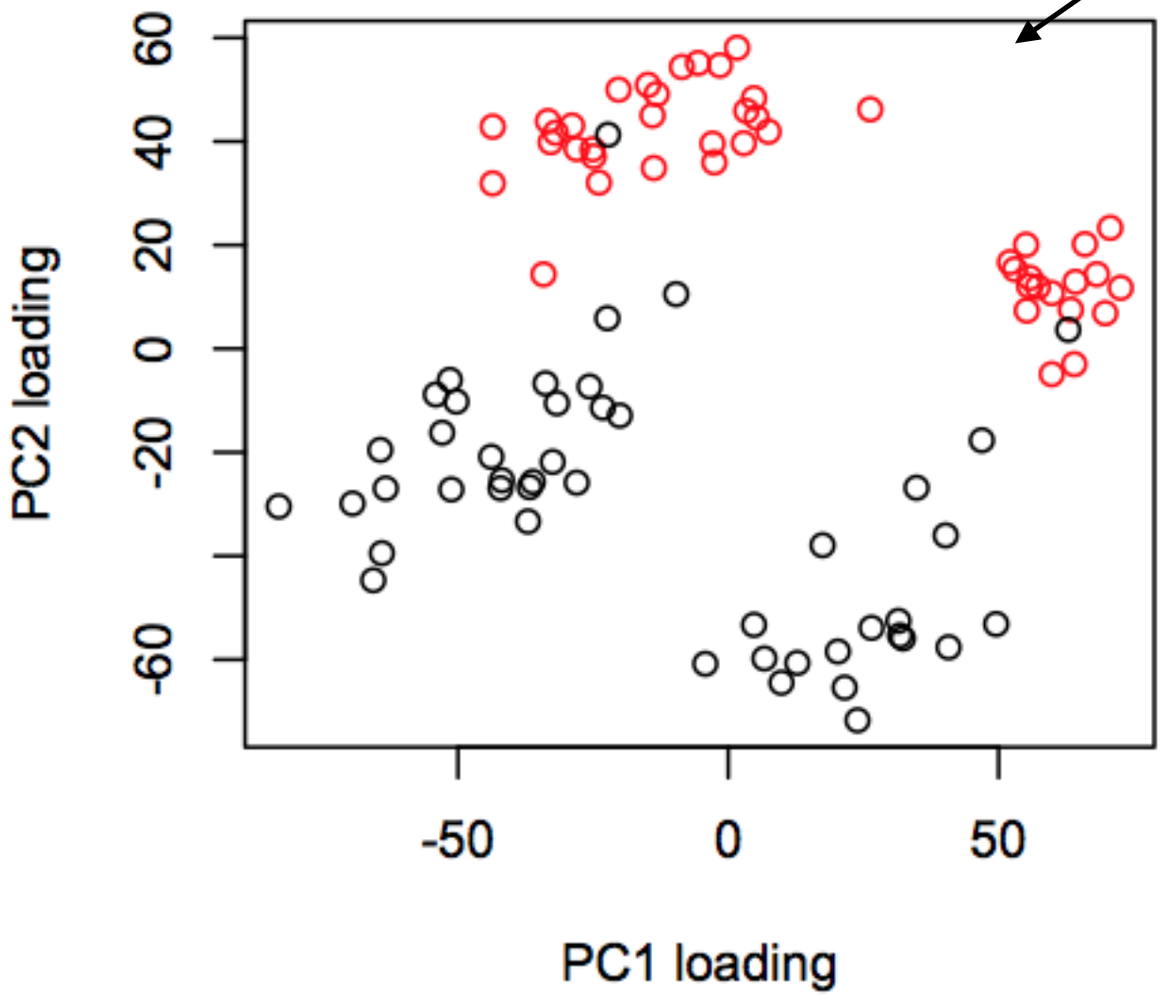
Pre/post-dispersal gene expression in male rhesus macaques



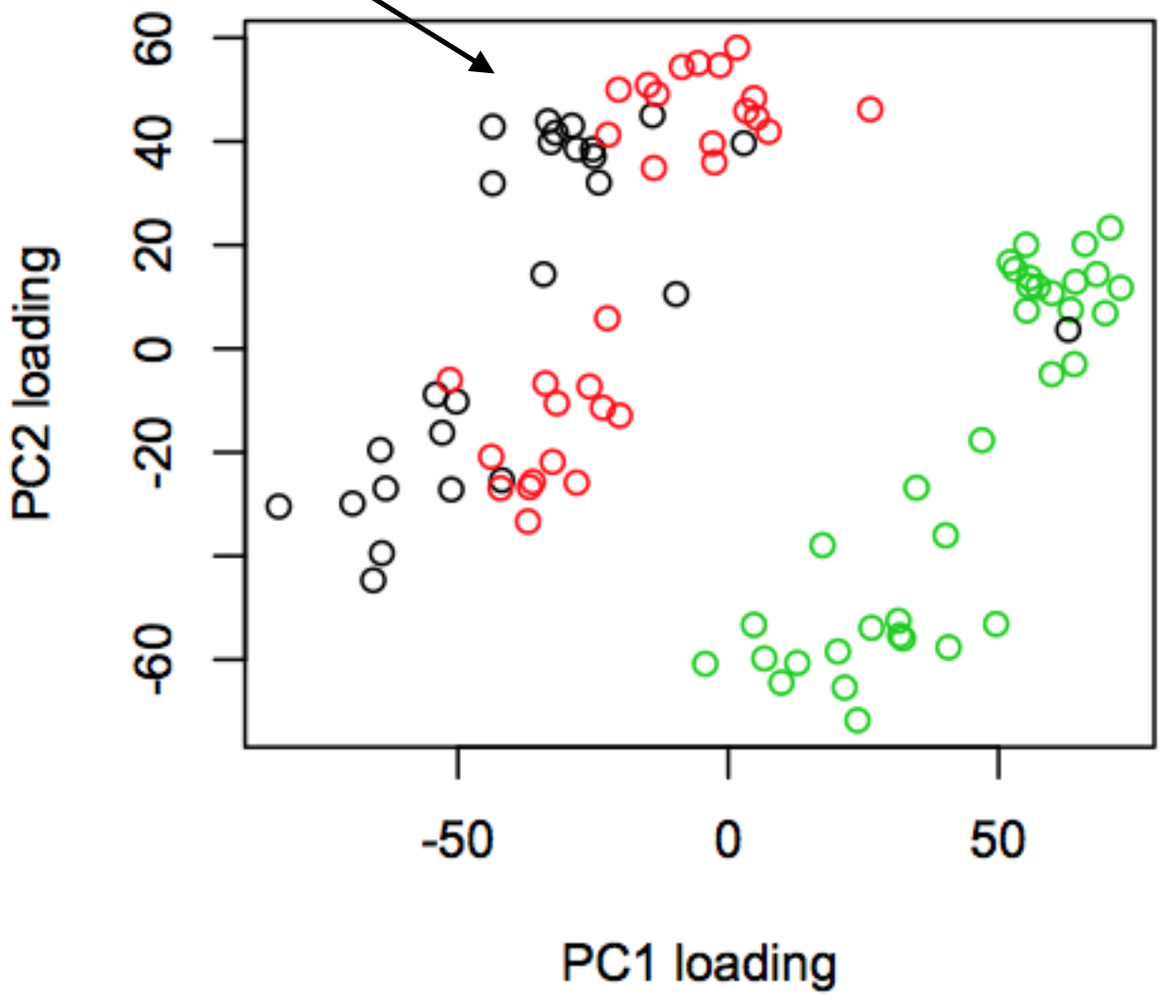
Snyder-Mackler and Tung, unpublished data

Response to immune stimulation in wild baboons

sampling effort effects layered on
top of stimulation effects



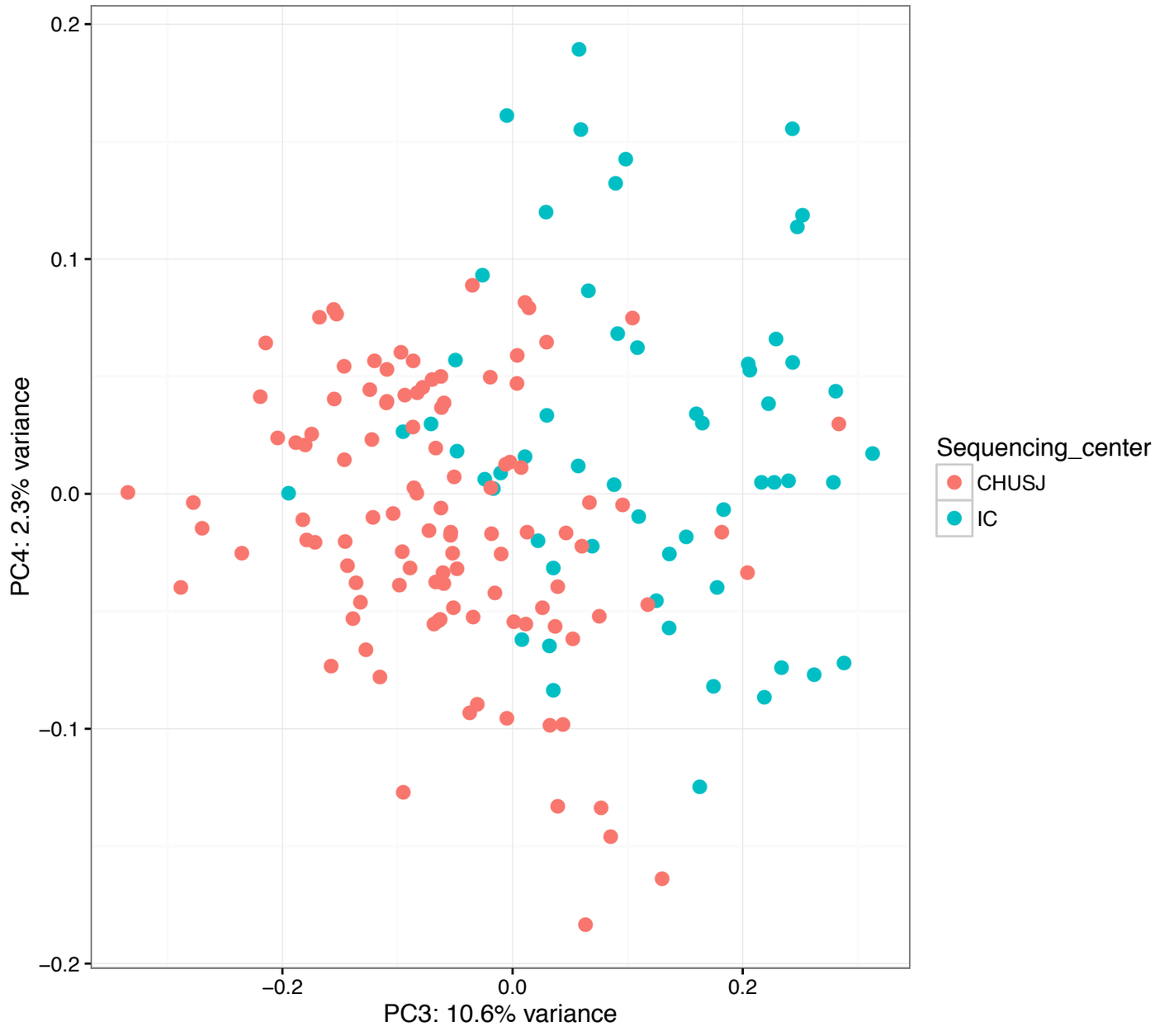
red=null; black=LPS



green=2015; red=2014; black=2013

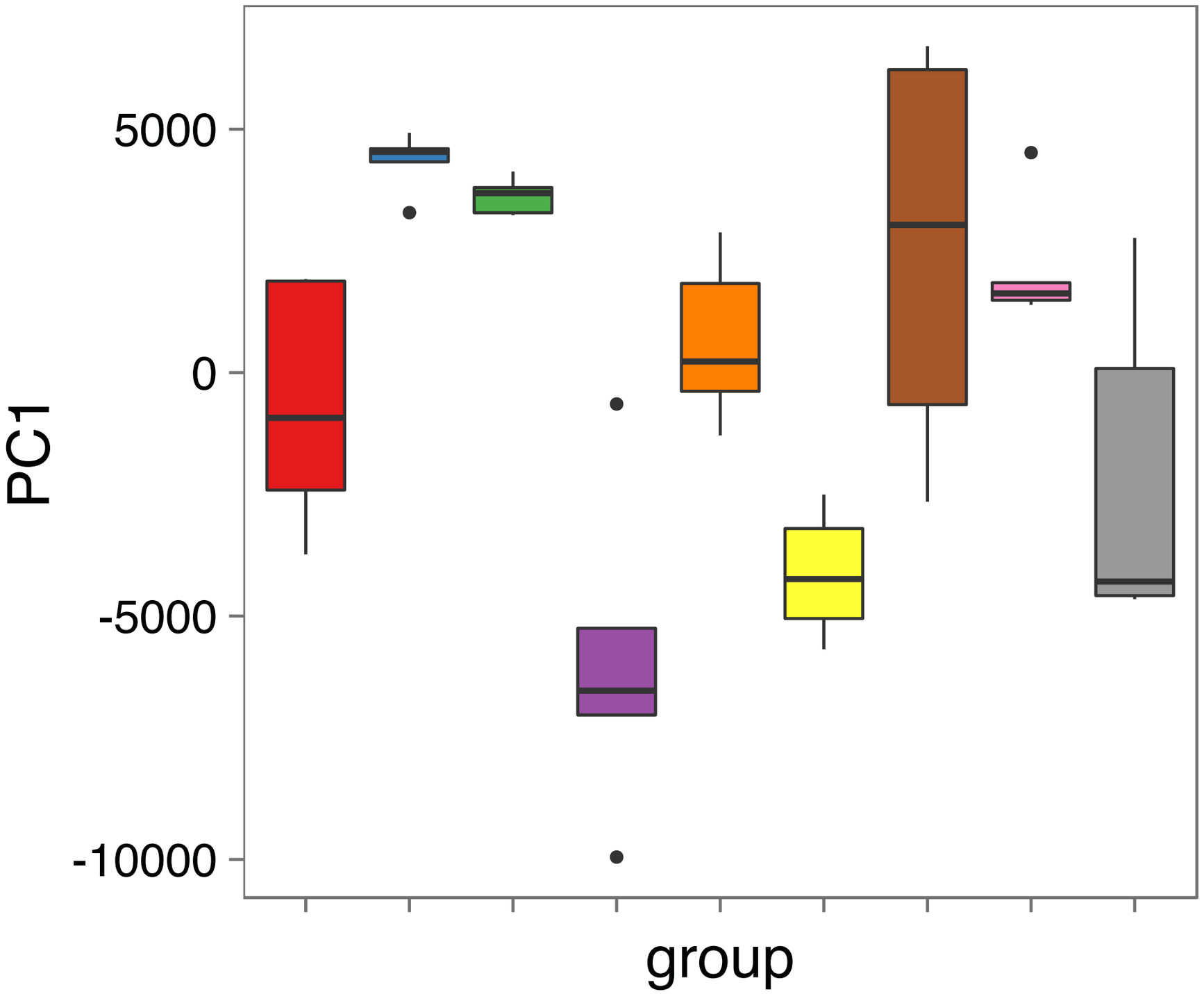
Lea and Tung, unpublished data

Listeria-infected samples from African/European ancestry humans



Nedelec et al, in press (Cell)

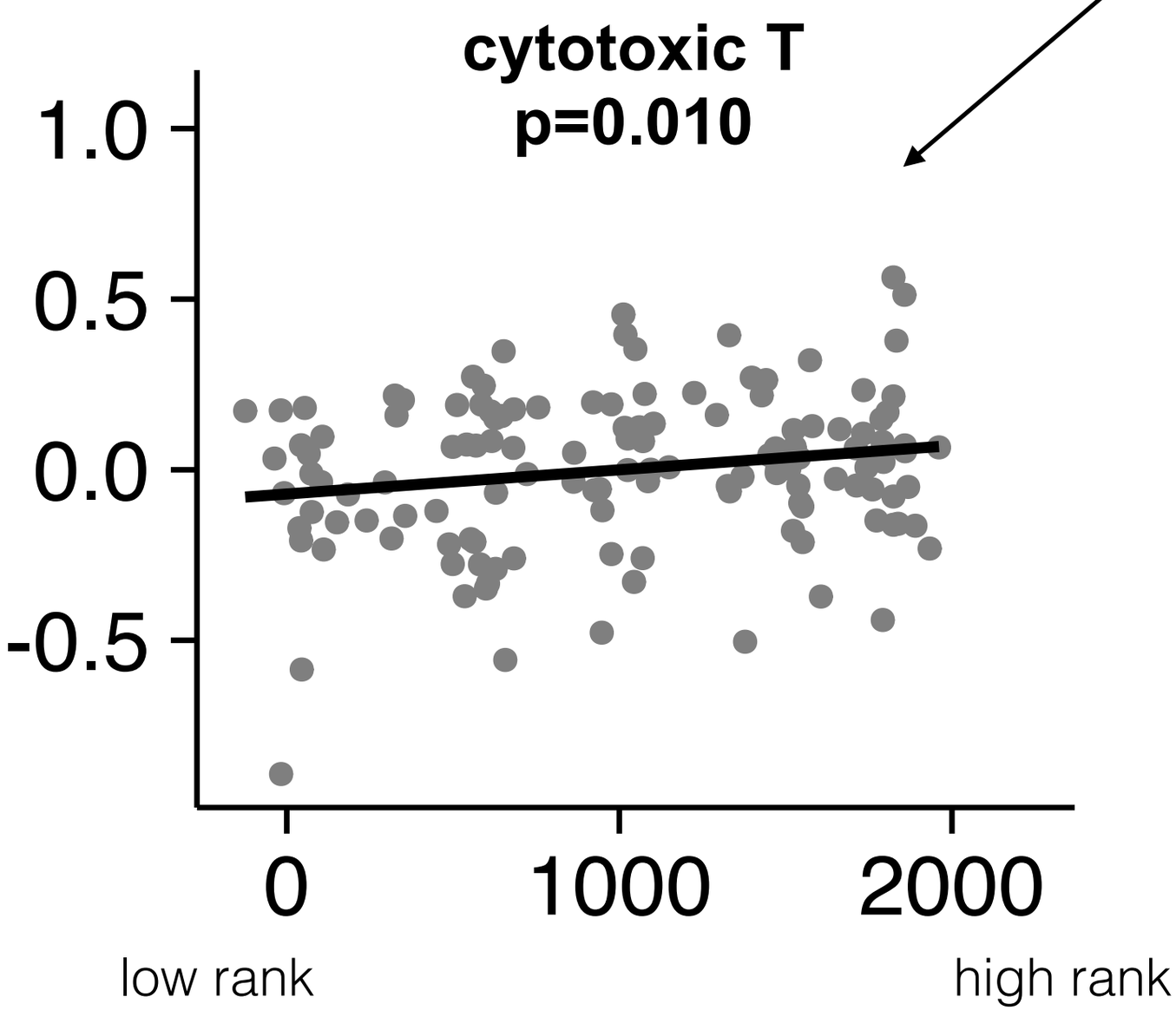
B cell gene expression in captive rhesus macaques

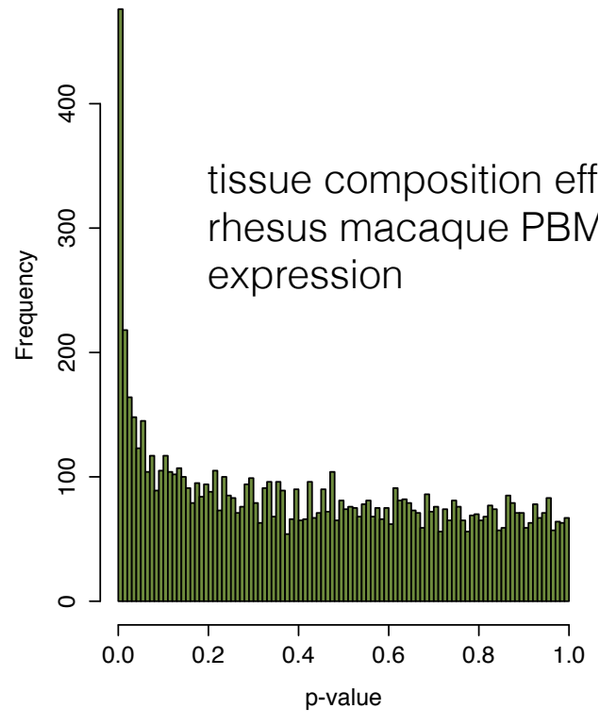


Snyder-Mackler, Sanz et al, in review

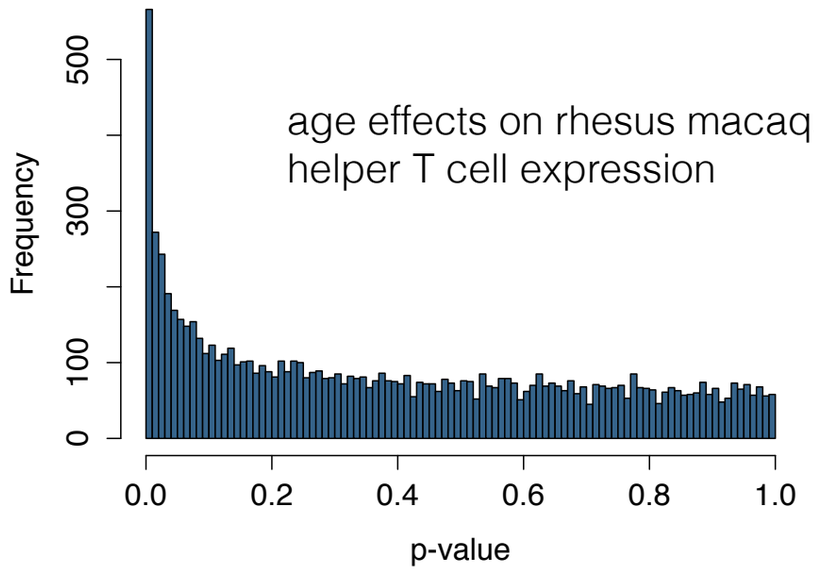
“Biological” batch effects can also occur

in heterogeneous samples, genes that are DE in cytotoxic T cells relative to other blood cells will look DE by rank *even* if there are no changes w/in cells





tissue composition effects on rhesus macaque PBMC gene expression

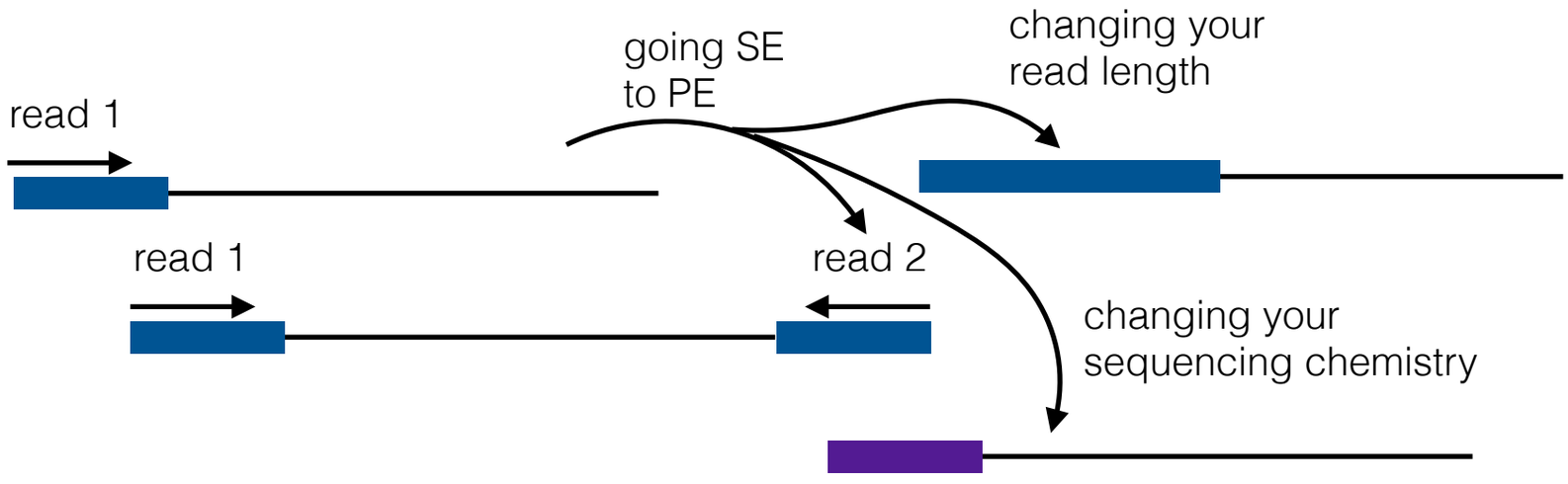


age effects on rhesus macaque helper T cell expression

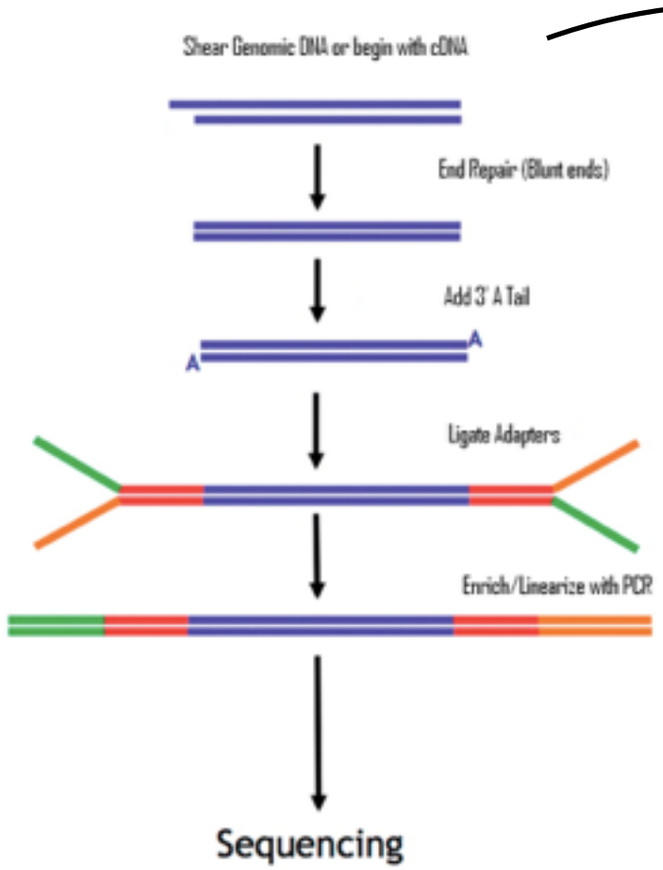
Snyder-Mackler, Sanz et al, in review; Synder-Mackler and Tung, unpublished data

…and also your past poor decisions!

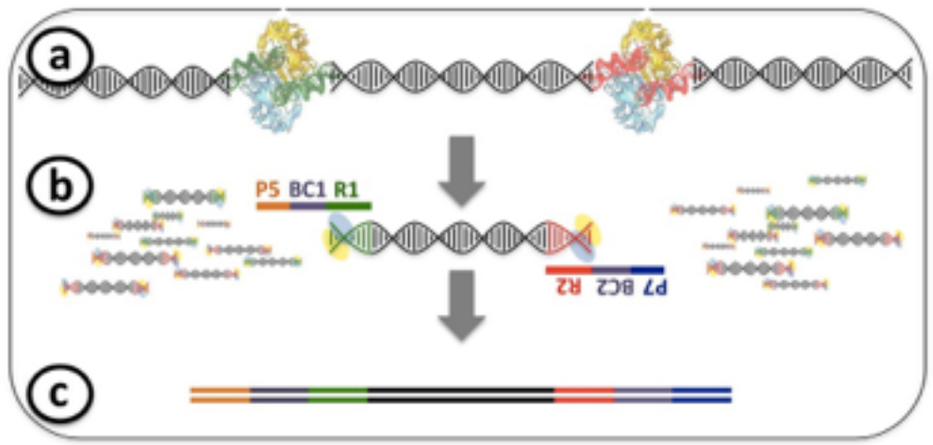
changing your sequencing strategy mid-stream



changing your library prep method

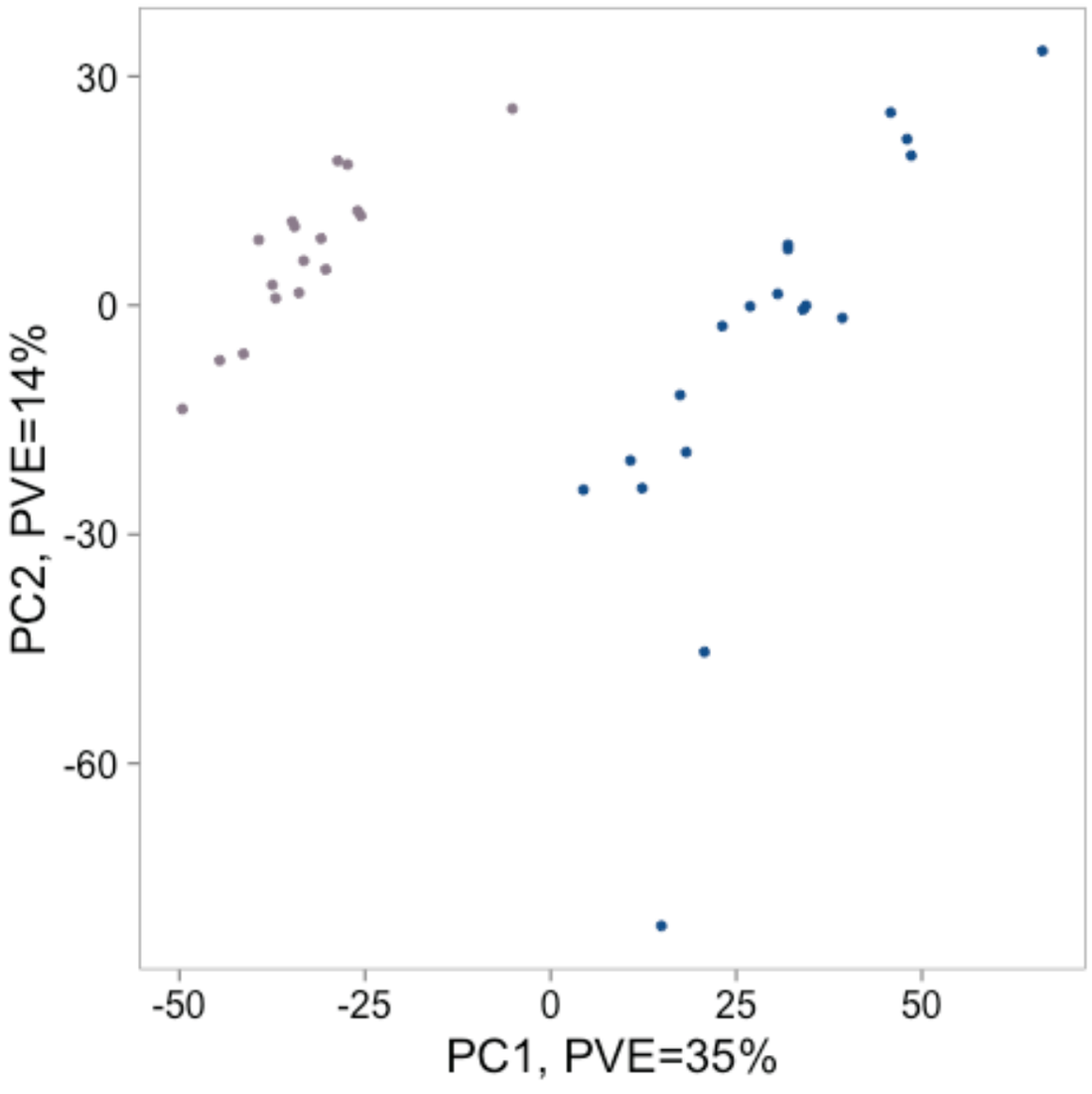


e.g., TruSeq to Nextera

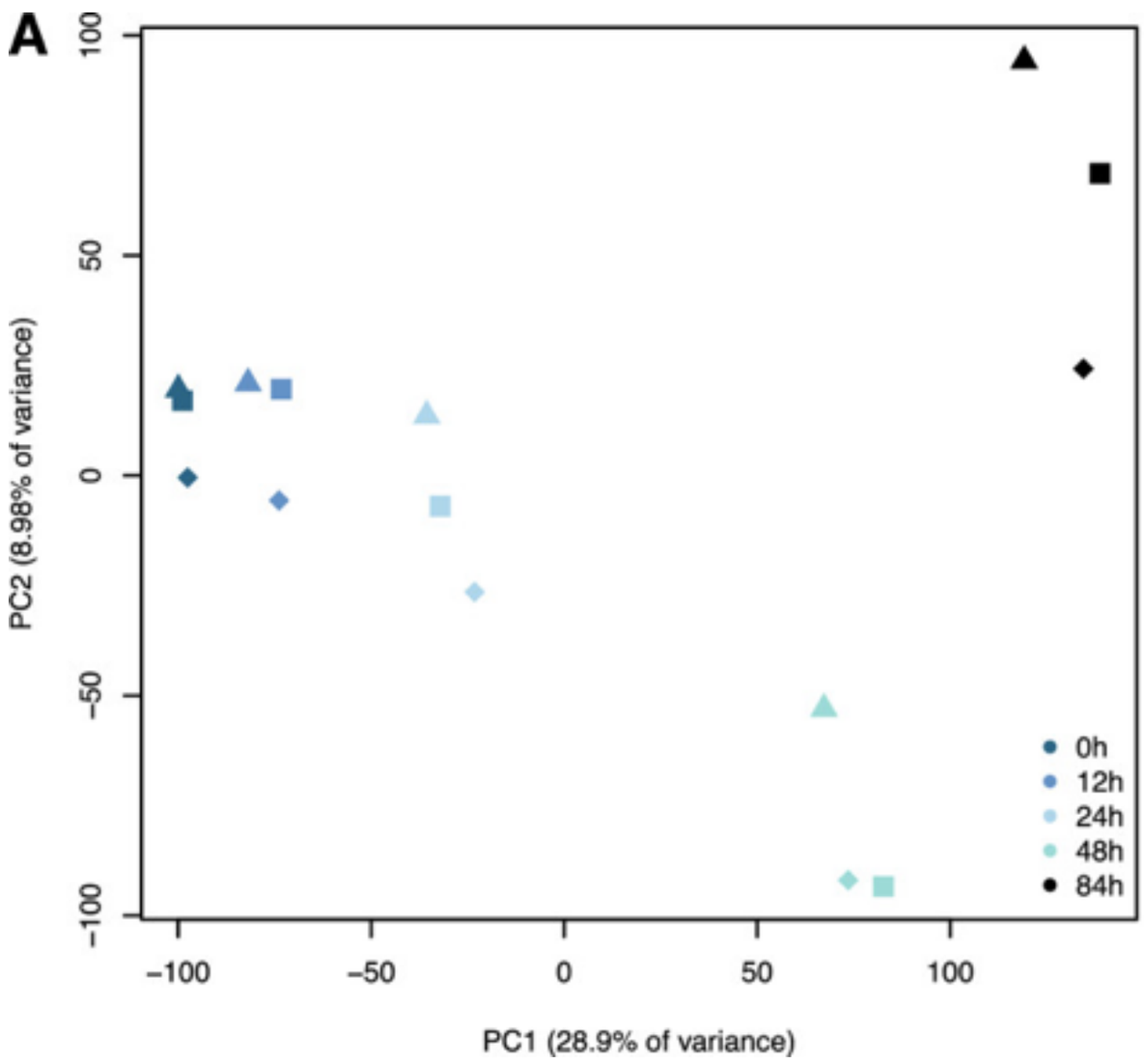


…and also your past poor decisions!

augmenting your data set later on



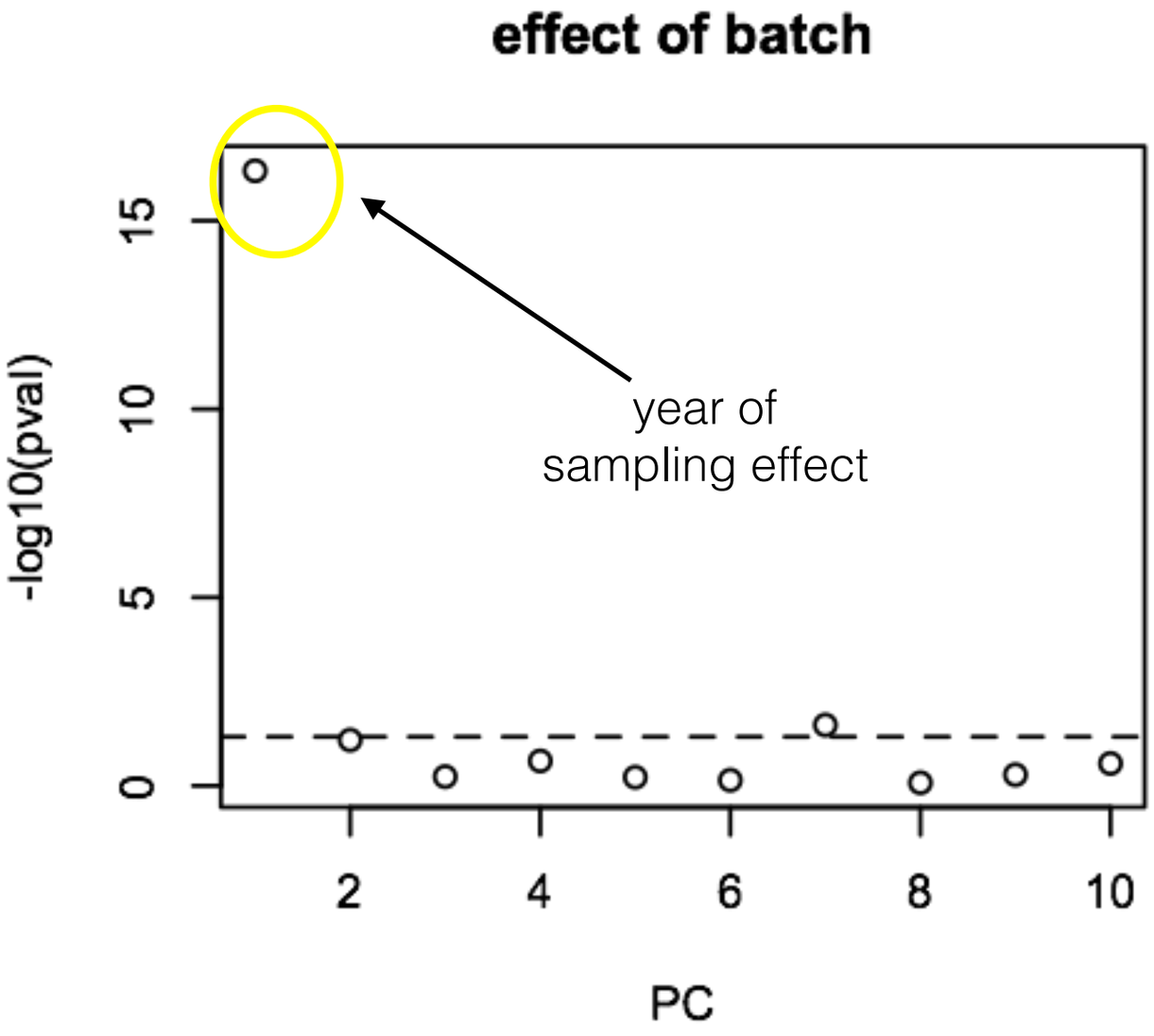
using degraded RNA



Gallego Romero et al 2014

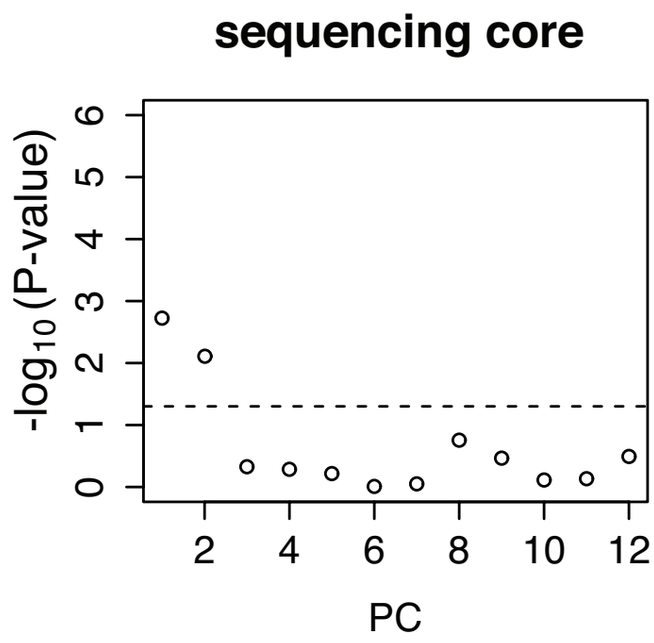
What do you do about batch effects?

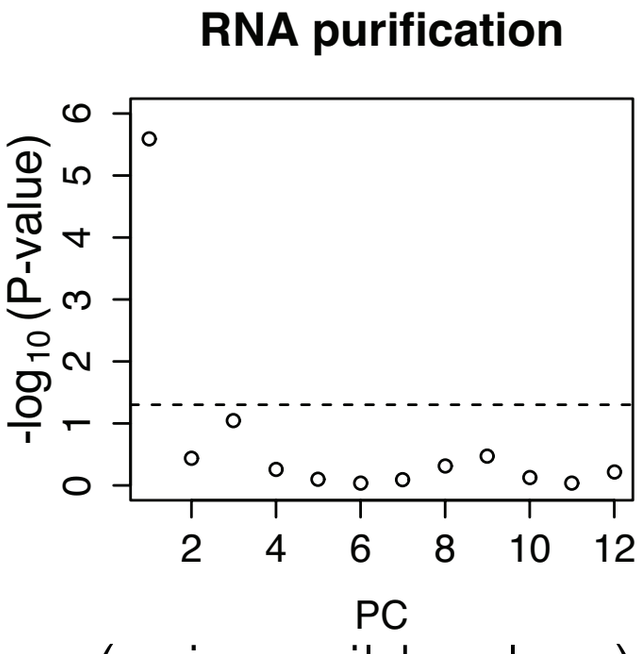
1. Try to figure out the batch effects you can model directly.
   i. Pervasive effects act in *trans* and often affect many genes



(LPS challenge data,
wild baboons)

Lea and Tung, unpublished data



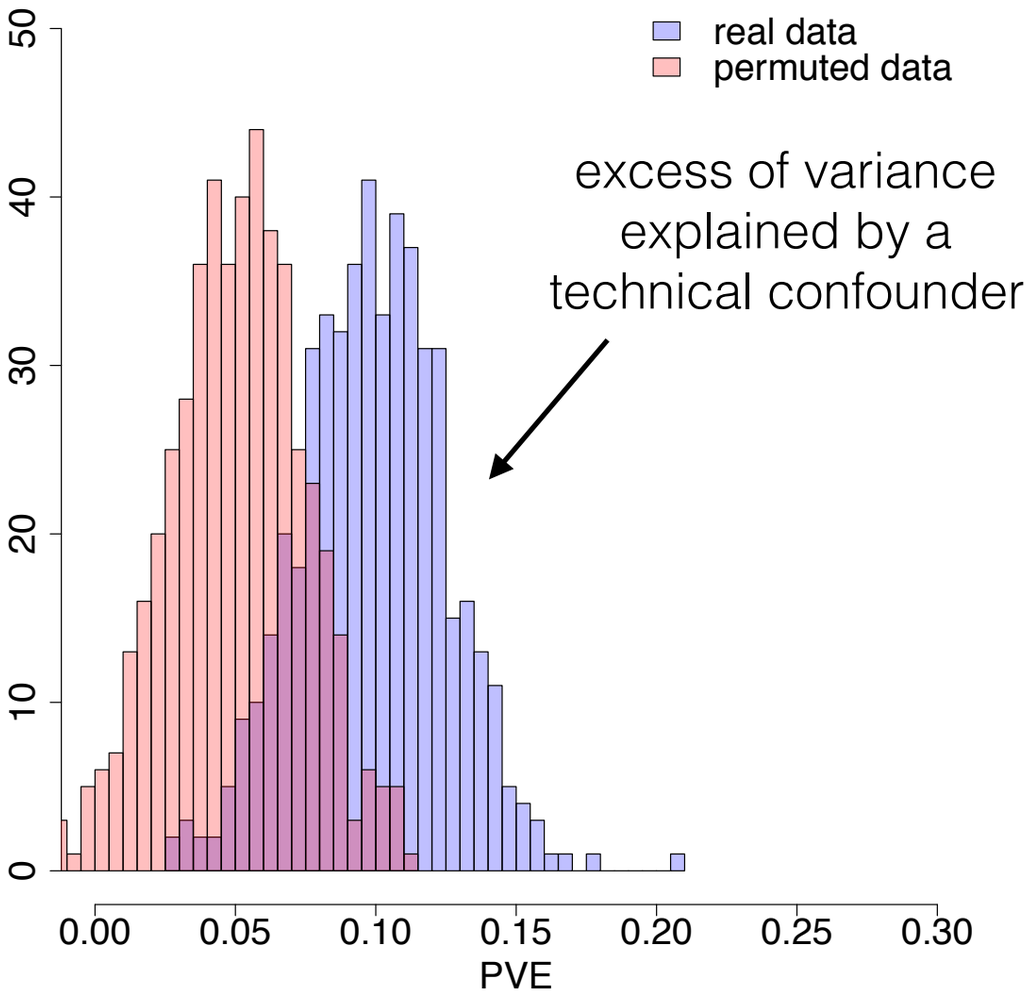


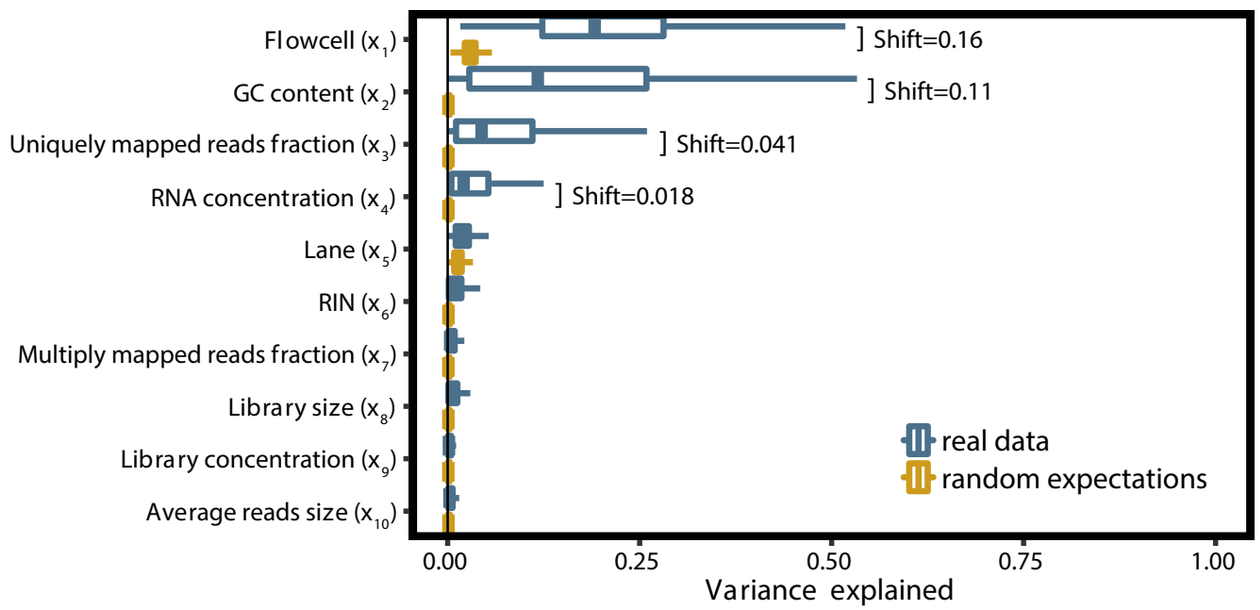
(aging, wild wolves)

Charruau, Johnston et al 2016

What do you do about batch effects?

1. Try to figure out the batch effects you can model directly.
   i. Pervasive effects act in *trans* and often affect many genes





Nedelec et al, in press (Cell)

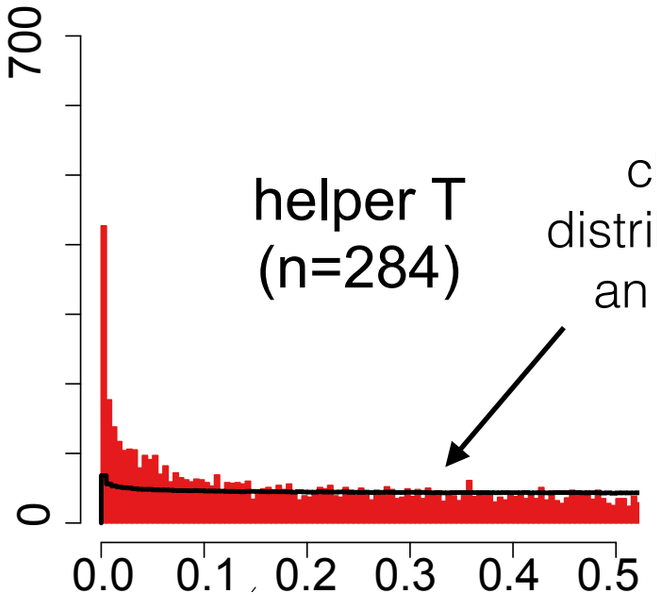
What do you do about batch effects?

1. Try to figure out the batch effects you can model directly.
   i. Pervasive effects act in *trans* and often affect many genes

2. Figure out how to take known batch effects into account.
   i. model as covariates
   ii. "remove" prior to main analysis: regress them out, use methods that "protect" the effects of known biological covariates, e.g., ComBat (Johnson et al 2007, Biostatistics), pSVA (Parker et al 2014, Bioinformatics)

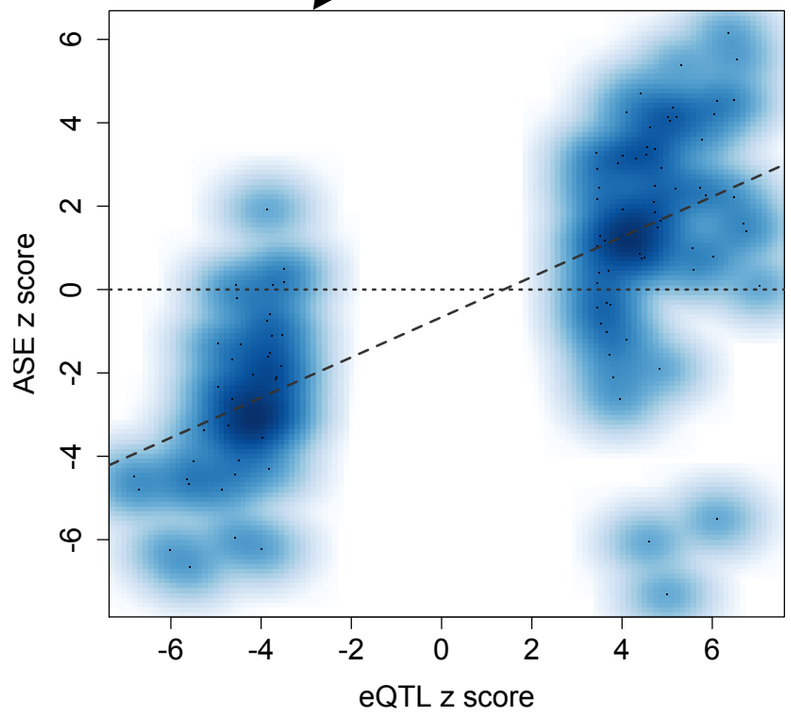
3. Try to get rid of batch effects you can't identify.
   i. PC removal (e.g., in eQTL analysis)
   ii. Surrogate variable analysis: remove hidden factors not associated with biological effects of interest (Leek & Storey 2007, PLoS Genetics)

Check for biological/statistical coherence

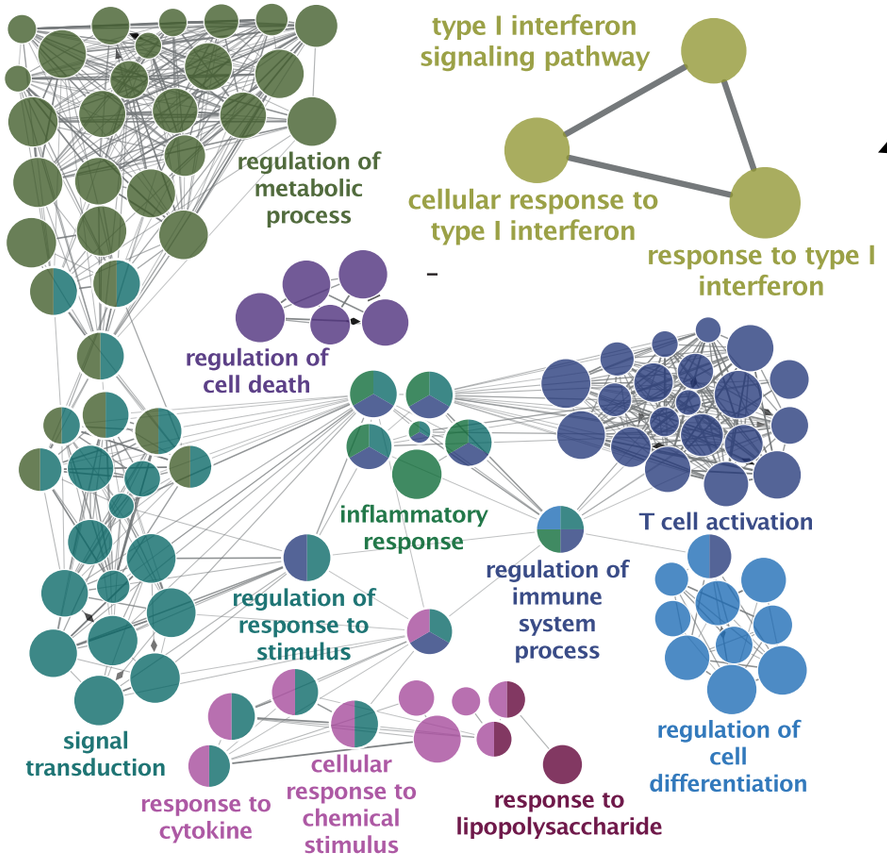


check your null distribution! (calculate an empirical FDR)

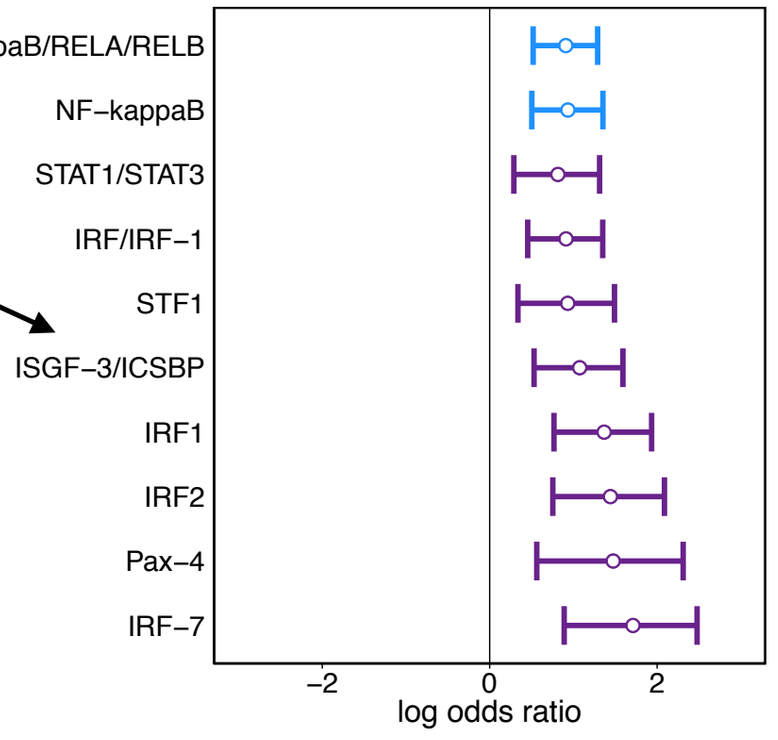
perform a complementary data analysis



look for "sensible" pathway enrichments (with caution)



generate/analyze complementary data

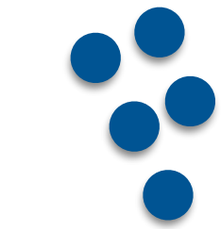


Snyder-Mackler, Sanz et al, in review; Tung et al 2015, eLife

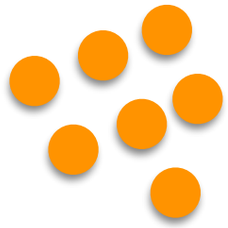
Some issues to consider

*not including library prep (stranded? PCR-free? Nextera/Tru-seq/other?), mapping (which aligner? what QC thresholds?), genome annotation, etc.

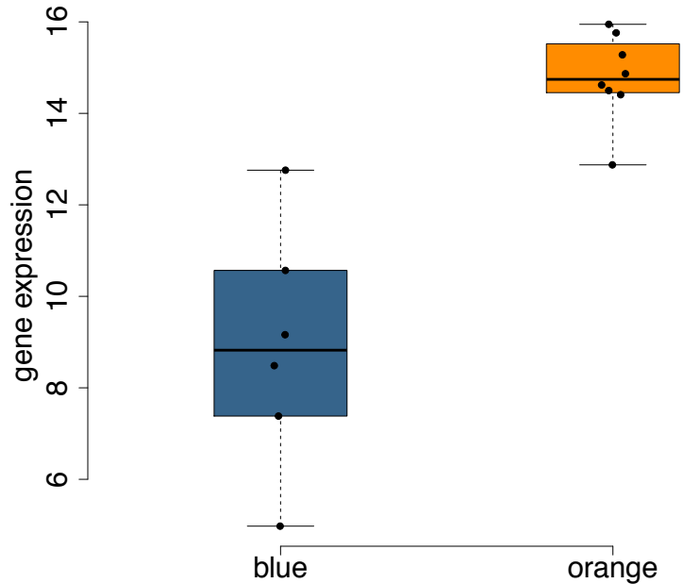
Do you have kin in your sample? Other sources of population structure?

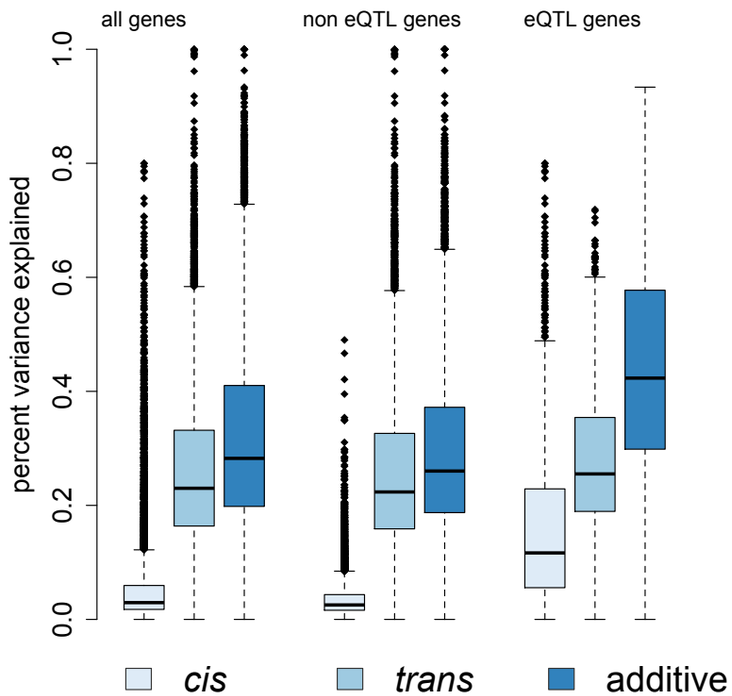


high status family

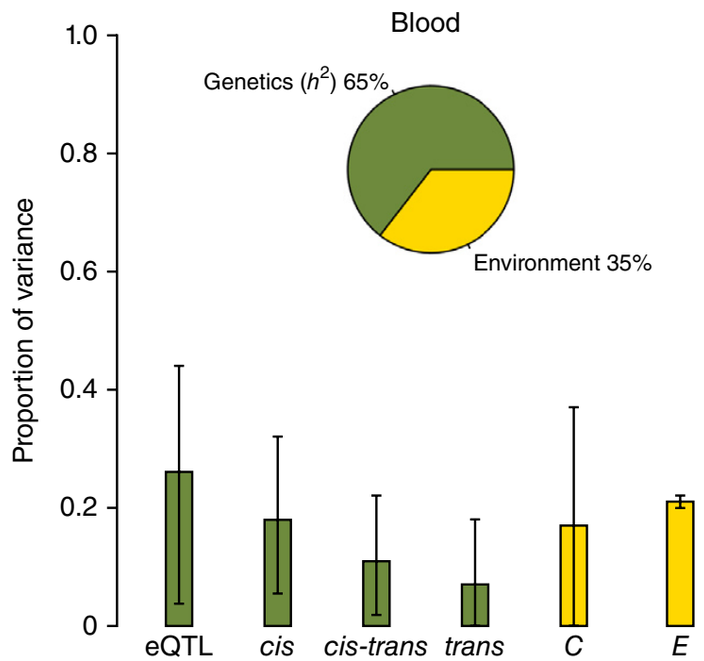


low status family





baboon h^2: Tung et al 2015, eLife



human h^2: from Buil et al 2014, Nature Genetics

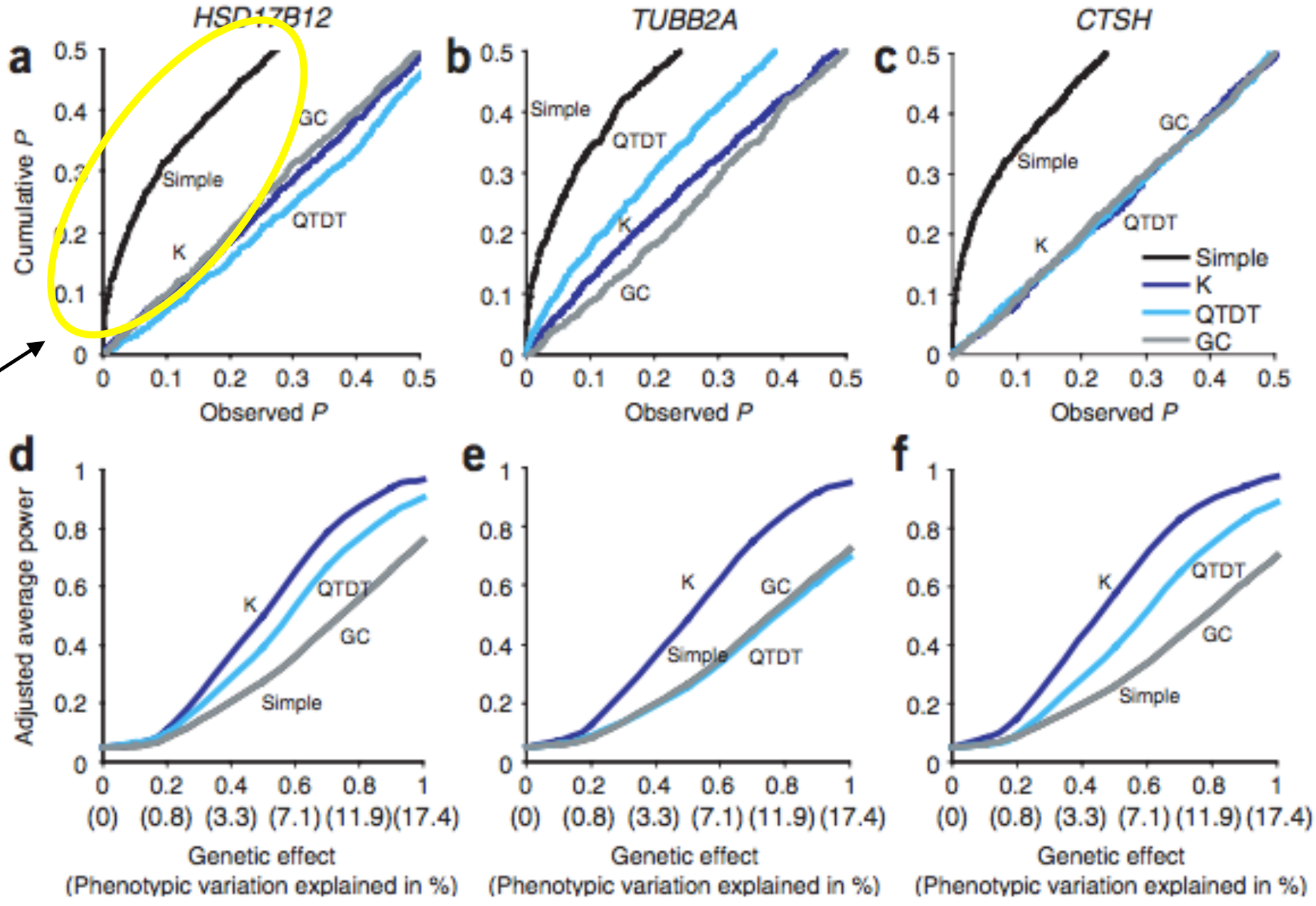
A unified mixed-model method for association mapping that accounts for multiple levels of relatedness

Jianming Yu[1,9], Gael Pressoir[1,9], William H Briggs[2], Irie Vroh Bi[1], Masanori Yamasaki[3], John F Doebley[2], Michael D McMullen[3,4], Brandon S Gaut[5], Dahlia M Nielsen[6], James B Holland[4,7], Stephen Kresovich[1,8] & Edward S Buckler[1,4,8]

nature genetics

major type I (FP) error problem

mixed model including kinship (**K**) corrects T1 error and maintains power



The basic linear mixed model

effect size of variable of interest

independent residuals

$$y_i = x_i\beta + \mathbf{w}_i^T\boldsymbol{\alpha} + g_i + \varepsilon_i$$

gene expression data

variable of interest (e.g., a SNP genotype, age, an environmental exposure)

fixed effects covariates and their effect size

the genetic component of the gene expression variance is not independent across samples:

genetic variance component

$$\mathbf{g} \sim MVN(0, \sigma^2 h^2 \boldsymbol{K})$$

n x n pairwise kinship/ relatedness matrix (from pedigree or genotype data)

the “environmental” component of the gene expression variance is independent across samples:

$$\mathbf{e} \sim MVN(0, \sigma^2(1-h^2)\boldsymbol{I})$$

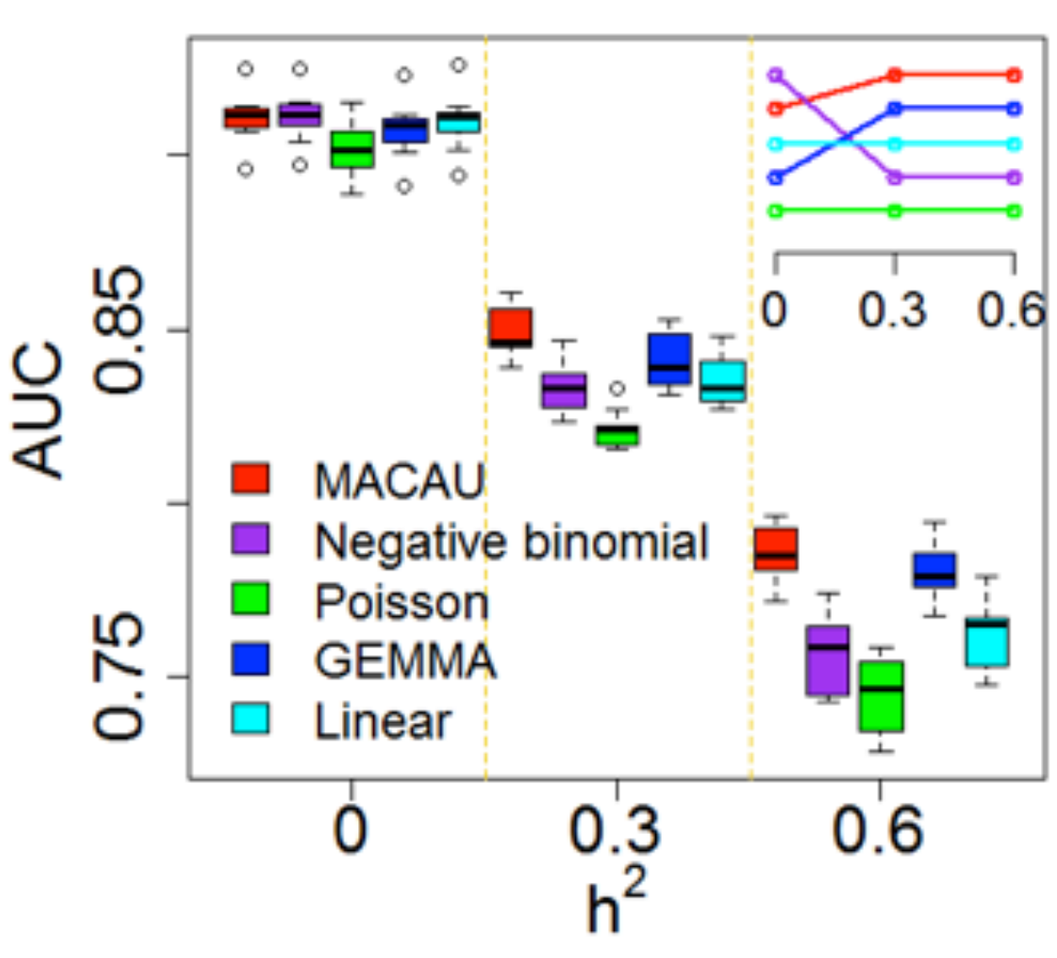
n x n pairwise identity matrix (1’s on the diagonal)

non-genetic variance component

Many efficient packages for fitting linear mixed models

- EMMA/EMMAX (Kang et al 2008, *Genetics*; Kang et al 2010, *Nature Genetics*); GEMMA (Zhou and Stephens 2012, *Nat Gen*), Fast-LMM (Lippert et al 2011, *Nat Methods*)

- Also some exploration of Poisson mixed effects models to model counts directly (MACAU: Sun et al 2016, bioRxiv)



Summary

Many options now available for RNA-seq data analysis (and also comparative literature benchmarking alternative models in different situations)

Approaches either attempt to stay faithful to the true, count-based data distribution *or* increase flexibility by transforming counts so they can be modeled using normal assumptions. *Recent benchmarks suggest that these transforms work well.*

"Batch" effects are common in RNA-seq data

Avoid introducing batch effects where possible (study design), always investigate the structure of your data, and consider what variance in your data you can model and/or exclude

Consider the structure in your data

Kinship/population structure are less likely to impact lab experimental studies, but often affects samples collected from natural populations.

"Cross-validate" your results

Against other studies, alternative analysis methods, complementary data sets

Thanks!

TUNG LAB:
Noah Snyder-Mackler
Amanda Lea
Tauras Vilgalys
Jordan Anderson
Arielle Fogel
Shauna Morrow
Amanda Shaver
Mike Yuan
Tawni Voyles
Tina Del Carpio
Reena Debray
Meghana Rao
Yingying Zhang

Xiang Zhou (Michigan)
Shiquan Sun (Michigan)
Luis Barreiro (Montreal)
Joaquin Sanz (Montreal)
Yohann Nedelec (Montreal)
Sayan Mukherjee (Duke)

Many co-authors, collaborators, and colleagues who have helped us learn about data generation, modeling, analysis, and the infinite array of things that can go wrong.