Experimental design

Charlotte Soneson
University of Zurich
Brixen 2016

"To call in the statistician
after the experiment is done
may be no more than asking him
to perform a postmortem examination:
he may be able to say what the
experiment died of."

Sir Ronald Fisher, Indian Statistical Congress, Sankhya, around 1938

Stephen John Senn
@stephensenn

Follow

Statisticians are the bad fairies of research. People forget to invite them until it's too late, at which point they send everyone to sleep.

RETWEETS 92 LIKES 93

11:22 AM - 21 Feb 2016

What is experimental design?

- The organization of an experiment, to ensure that the **right type** of data, and **enough** of it, is available to answer the **questions of interest** as clearly and efficiently as possible.

http://www.stats.gla.ac.uk/steps/glossary/anova.html#expdes

What is bad experimental design?

Treatment I

M M M M M M M M

Treatment II

F F F F F F F F

What is bad experimental design?

Treatment I

Treatment

Confounding!

What is bad experimental design?

Analysis batch I / Study center I / Processing protocol I ...

Tr	Tr	Tr	Tr	Tr	Tr	Tr	Tr

Analysis batch II / Study center II / Processing protocol II ...

Ctl	Ctl	Ctl	Ctl	Ctl	Ctl	Ctl	Ctl

What is bad experimental design?

Analysis batch I / Study center I / Pro ing protocol I ...

Tr Tr Tr Tr Tr Tr

Confounding!

Analysis batch II / Study center II / Processing protocol II ...

Ctl Ctl Ctl Ctl Ctl Ctl Ctl Ctl

What can happen with bad experimental design?

- Example: gene expression study comparing 60 CEU and 82 ASN HapMap individuals

- 26% of the genes were found to be significantly differentially expressed (78% with less restrictive multiple testing correction)

- **But**: all CEU samples were processed (sometimes years) before all the ASN samples!

Akey et al., Nature Genetics 2007; Spielman et al., Nature Genetics 2007

What can happen with bad experimental design?

- Example: gene expression study comparing 60 CEU and 82 ASN HapMap samples

- 26% of the genes were found to be significantly differentially expressed at 10% FDR (6% with less restrictive multiple testing correction)

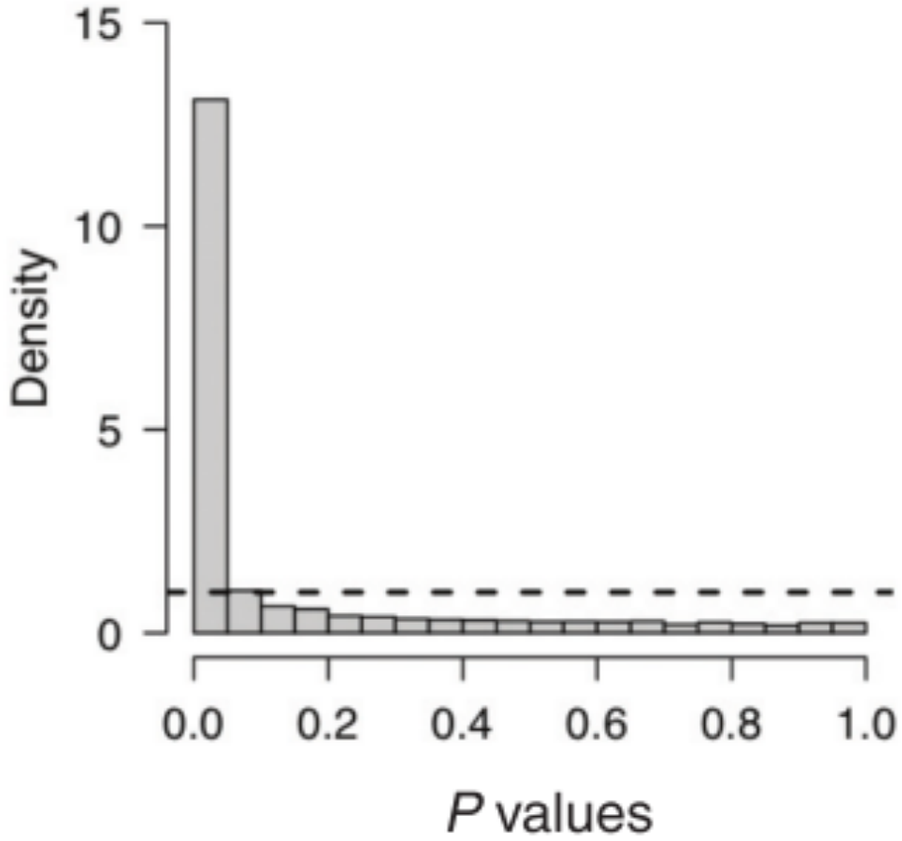
- **But**: all CEU samples were processed (sometimes years) before all the ASN samples!

Confounding!

Akey et al., Nature Genetics 2007; Spielman et al., Nature Genetics 2007

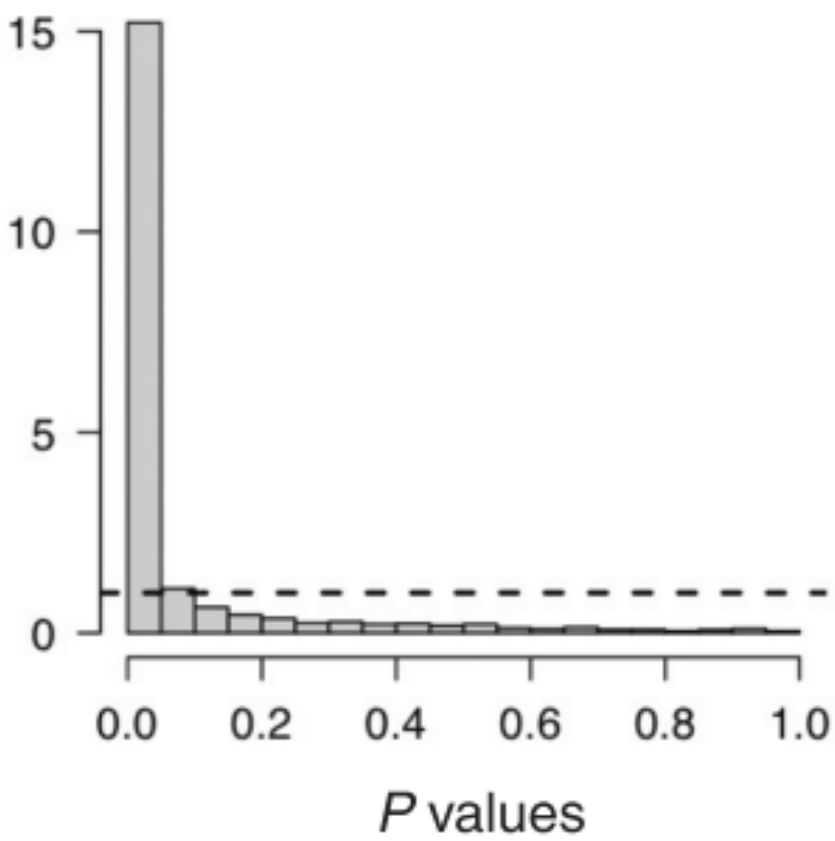
What can happen with bad experimental design?

a **Comparing CEU and ASN**



78% differentially expressed

b **Comparing processing times**

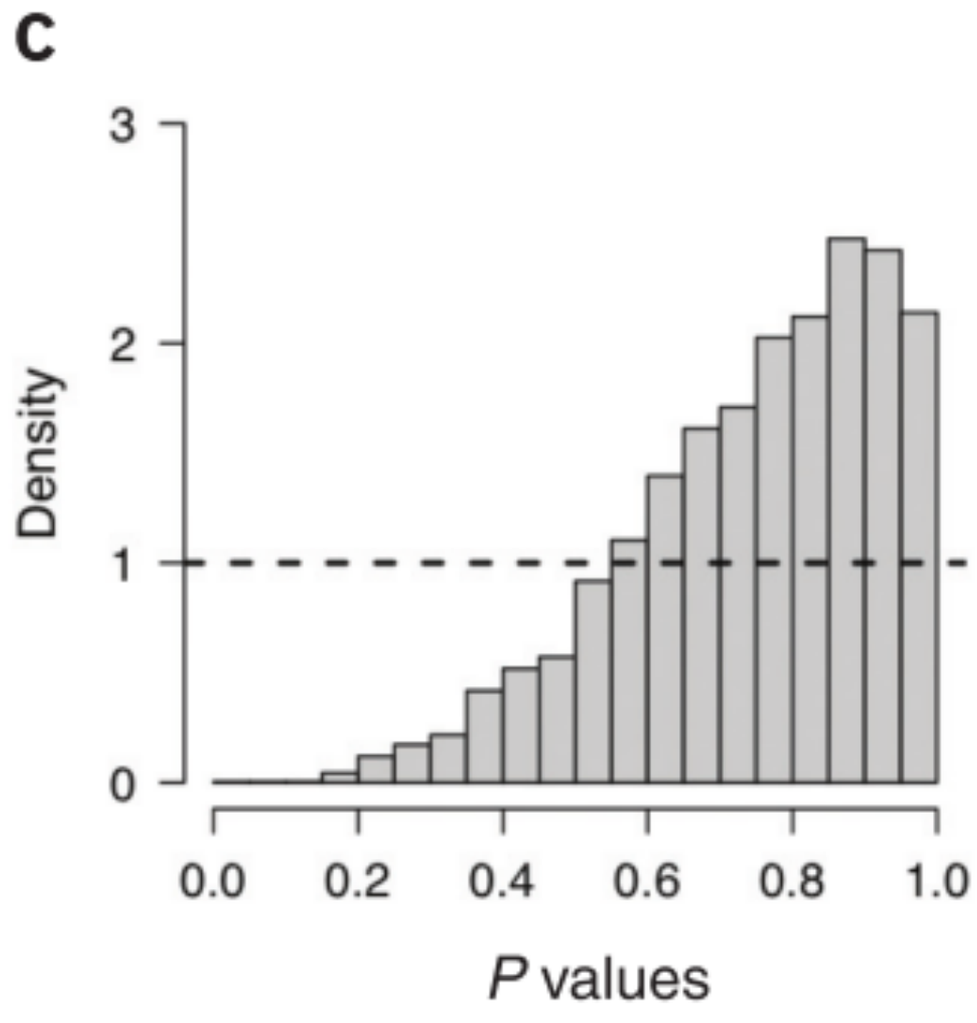


96% differentially expressed

Akey et al., Nature Genetics 2007

“Batch effect correction” won’t work here

p-values from test comparing CEU and ASN, after controlling for the processing year



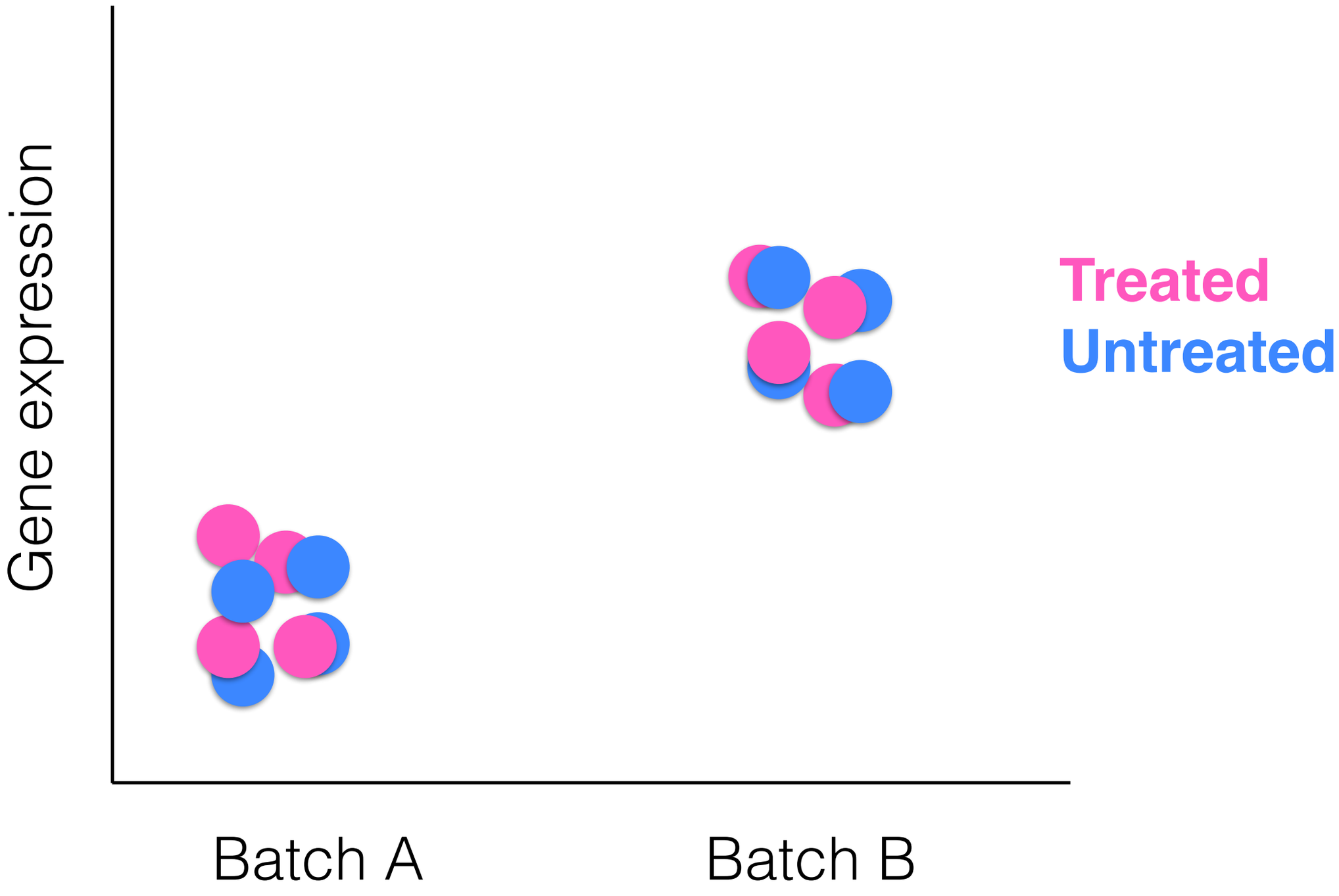
0% differentially expressed

Akey et al., Nature Genetics 2007

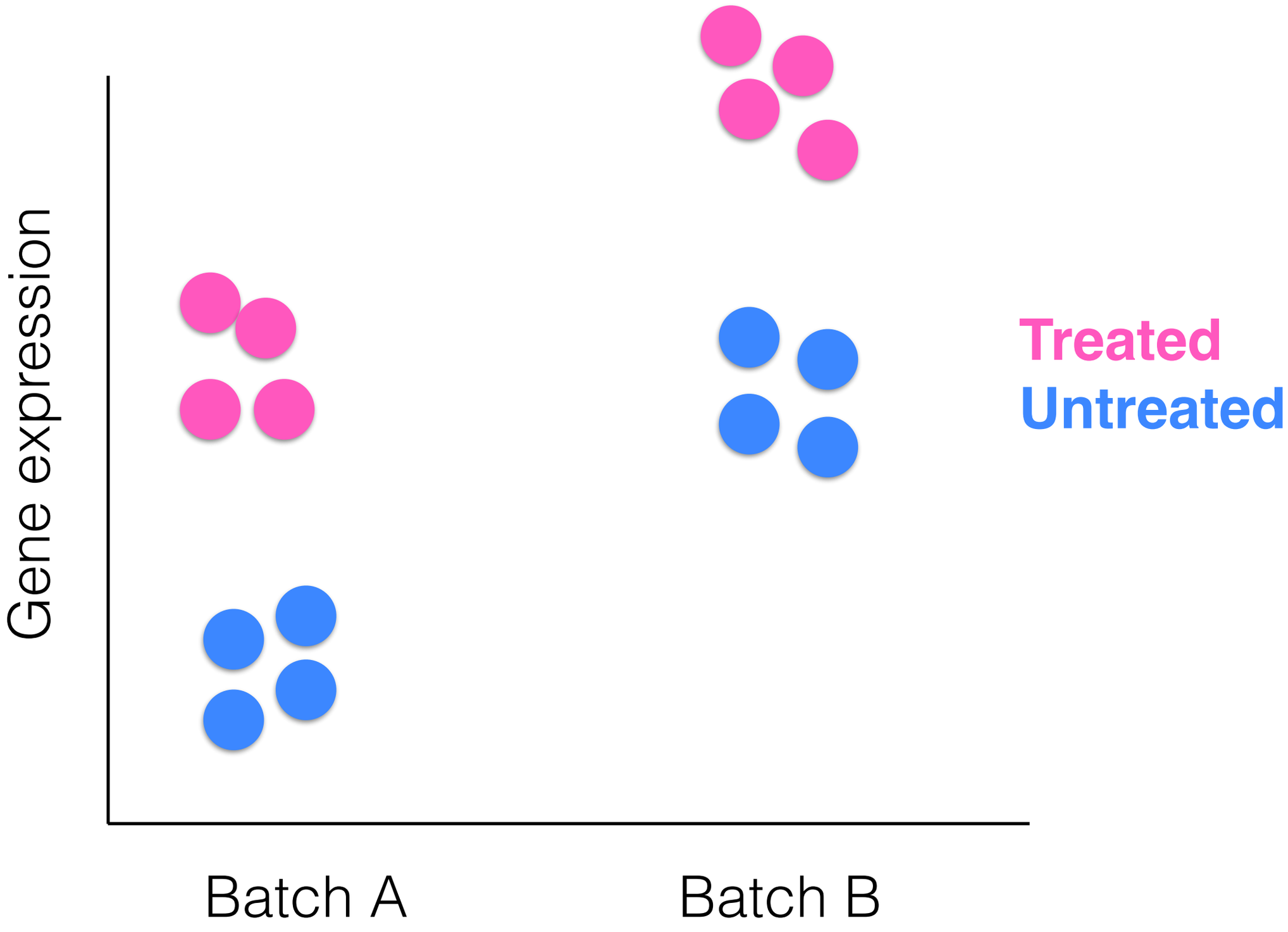
What could be a good experimental design?

- Process all samples at the same time (not always feasible)
- Minimize confounding as much as possible through
  - blocking
  - randomization
- The batch effect will still be there, but with an appropriate design we can account for it!

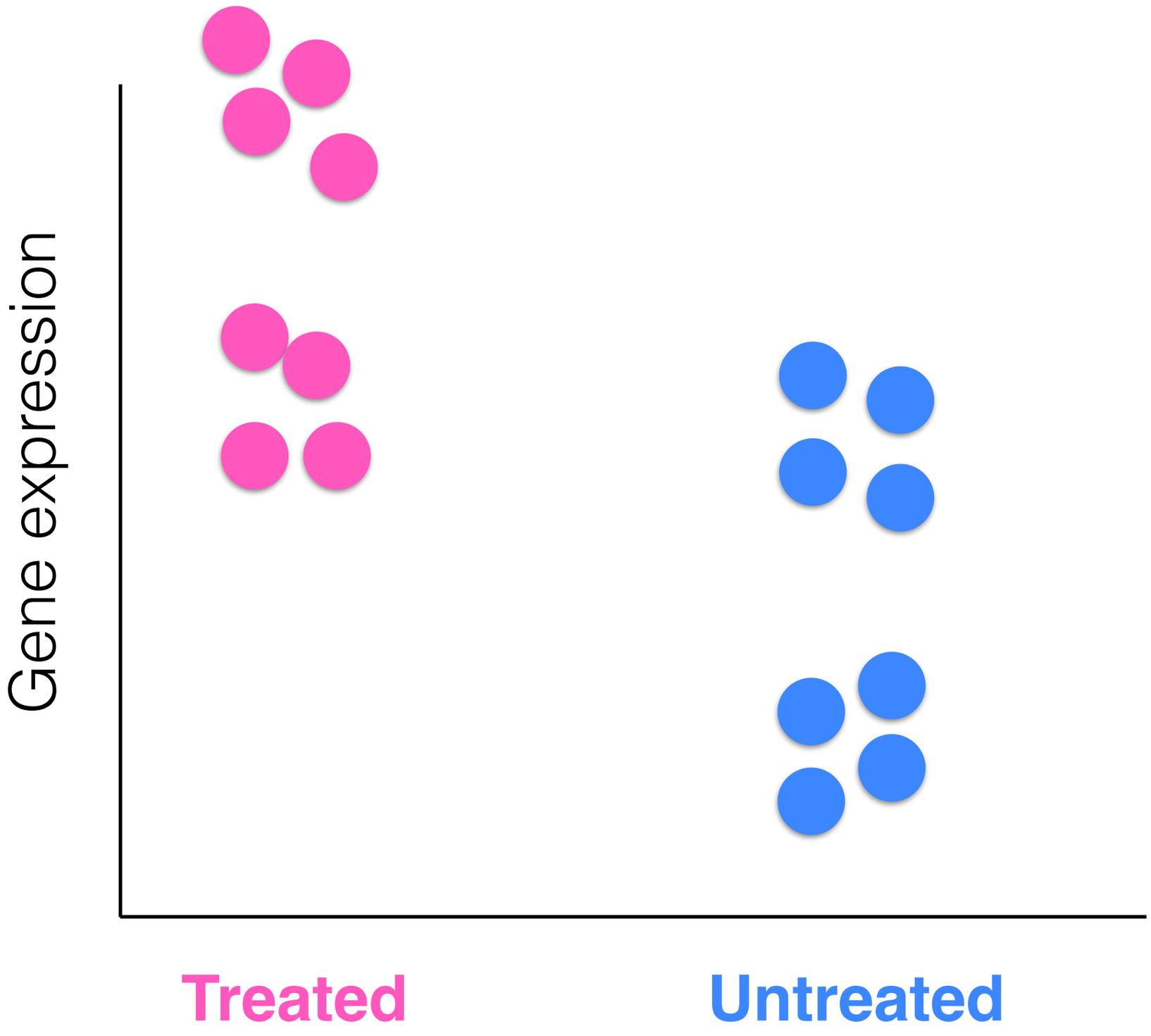
Non-confounded design



Non-confounded design



Non-confounded design



Accounting for batch effects

- In statistical modeling, batch effects can be included as **covariates** (additional predictors) in the model.

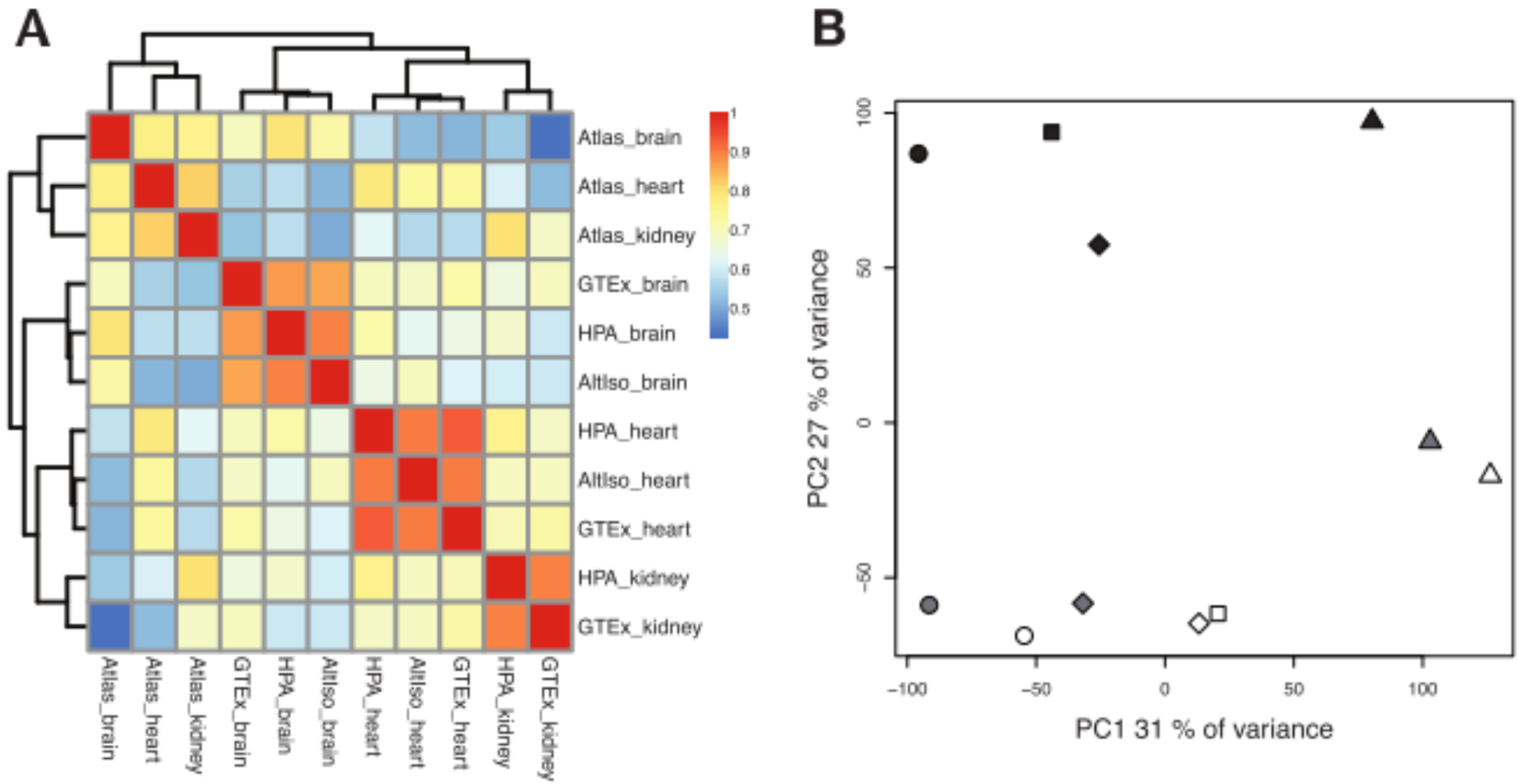
- For exploratory analysis, we often attempt to "eliminate" or "adjust for" such unwanted variation in advance, by subtracting the estimated effect from each variable.

- Even partial confounding between batch and signal of interest can lead to bias.

Nygaard et al., Biostatistics 2016

Accounting for batch effects in practice

- Public, processed RNA-seq data from 3 tissues, 4 studies show strong association with study

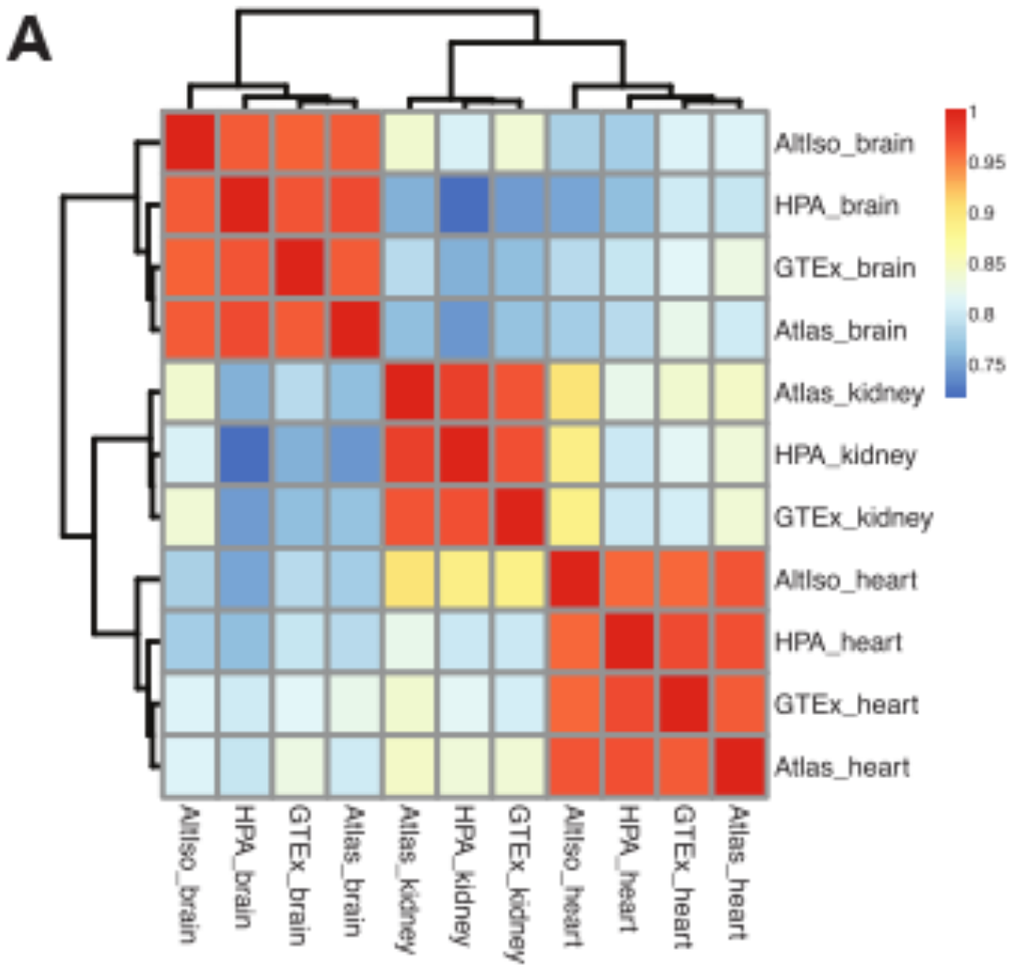


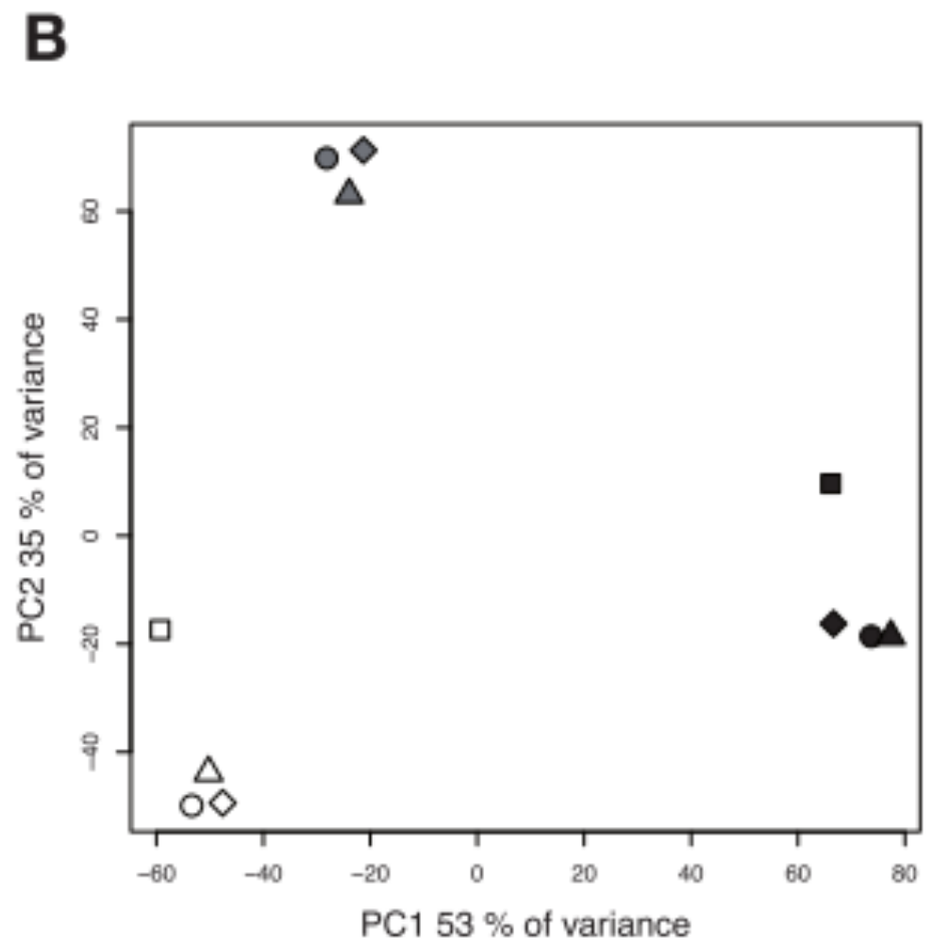
color = tissue; symbol = study (batch)

Danielsson et al., Briefings in Bioinformatics 2015

Accounting for batch effects in practice

- Accounting for the batch effect brings out signal of interest.



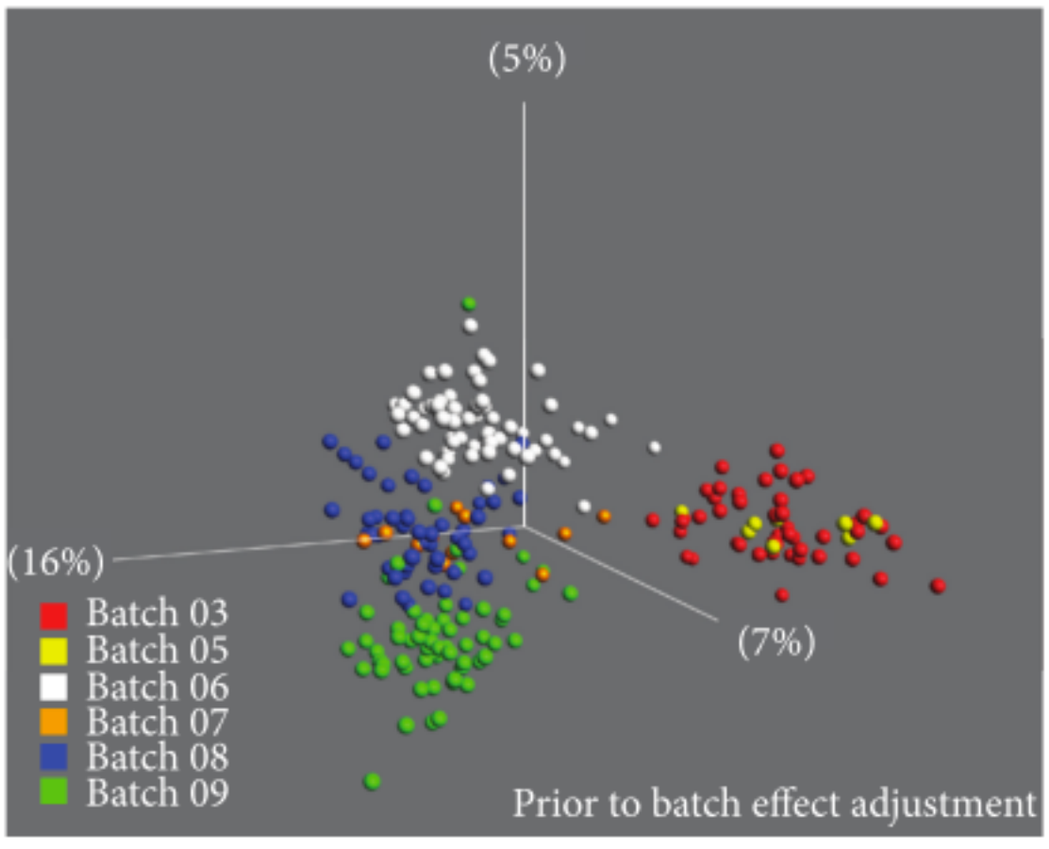


color = tissue; symbol = study (batch)

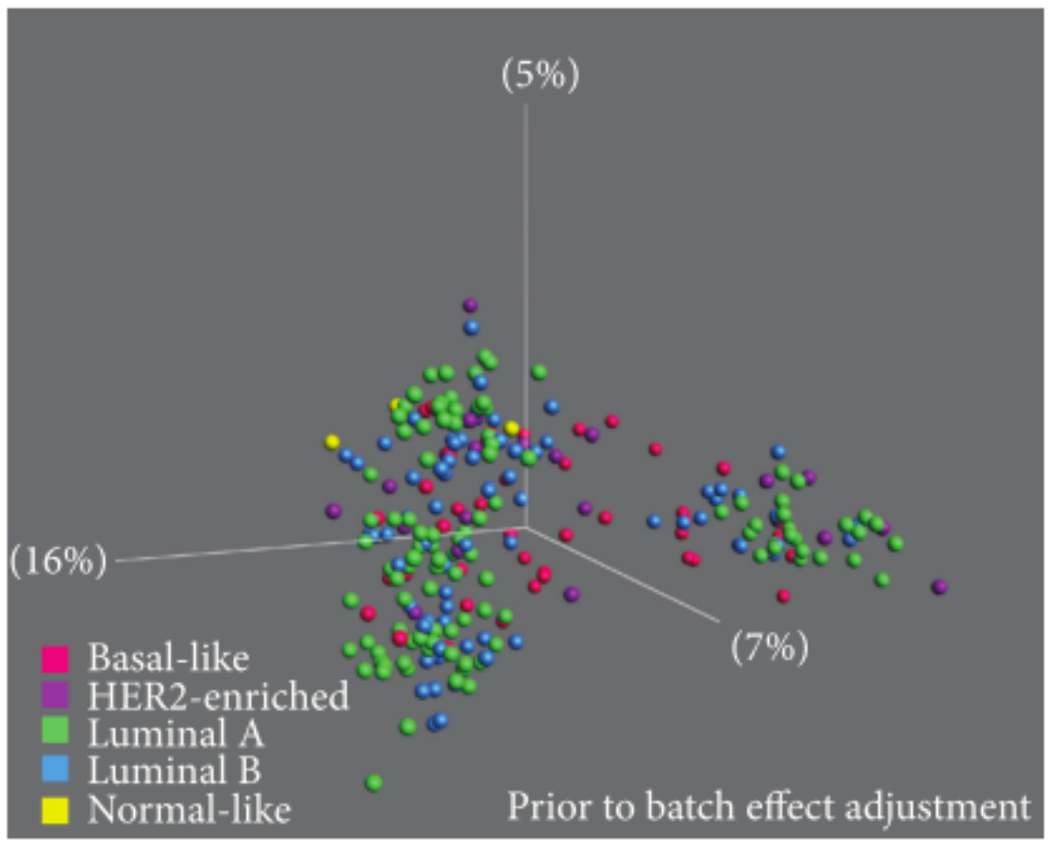
Danielsson et al., Briefings in Bioinformatics 2015

Accounting for batch effects in practice

- 5-subtype breast cancer microarray data processed in six batches.



(a)

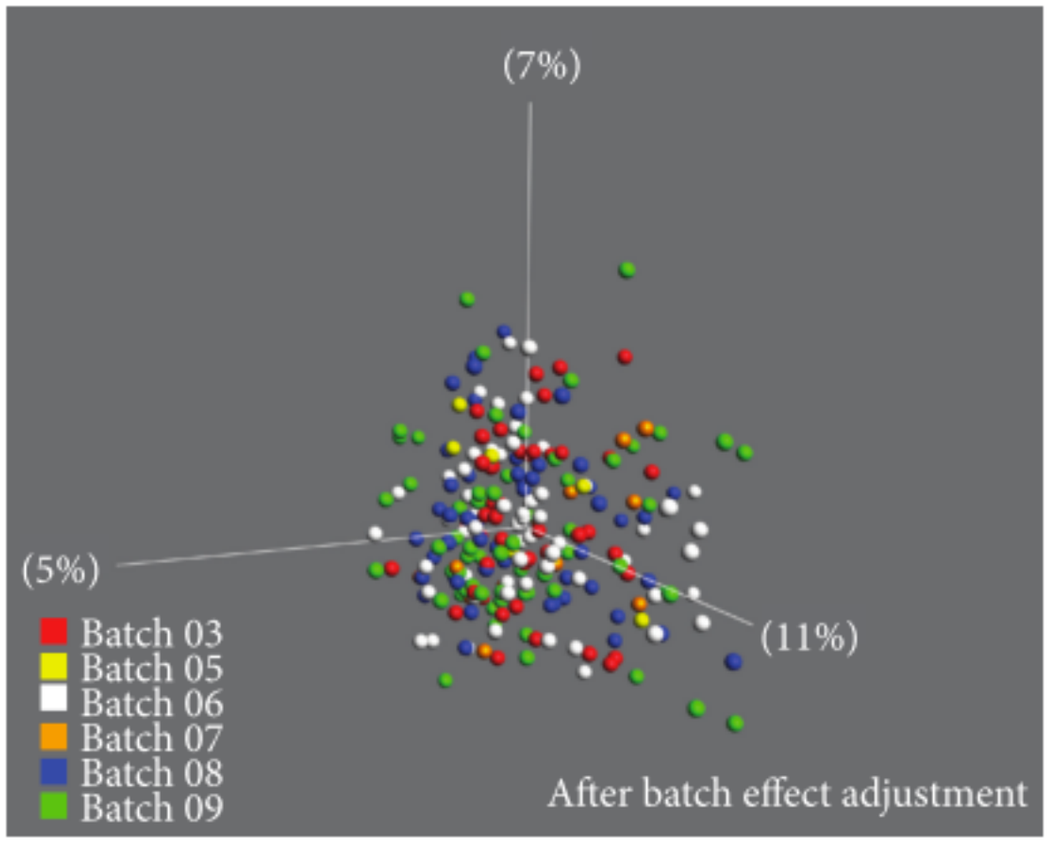


(c)

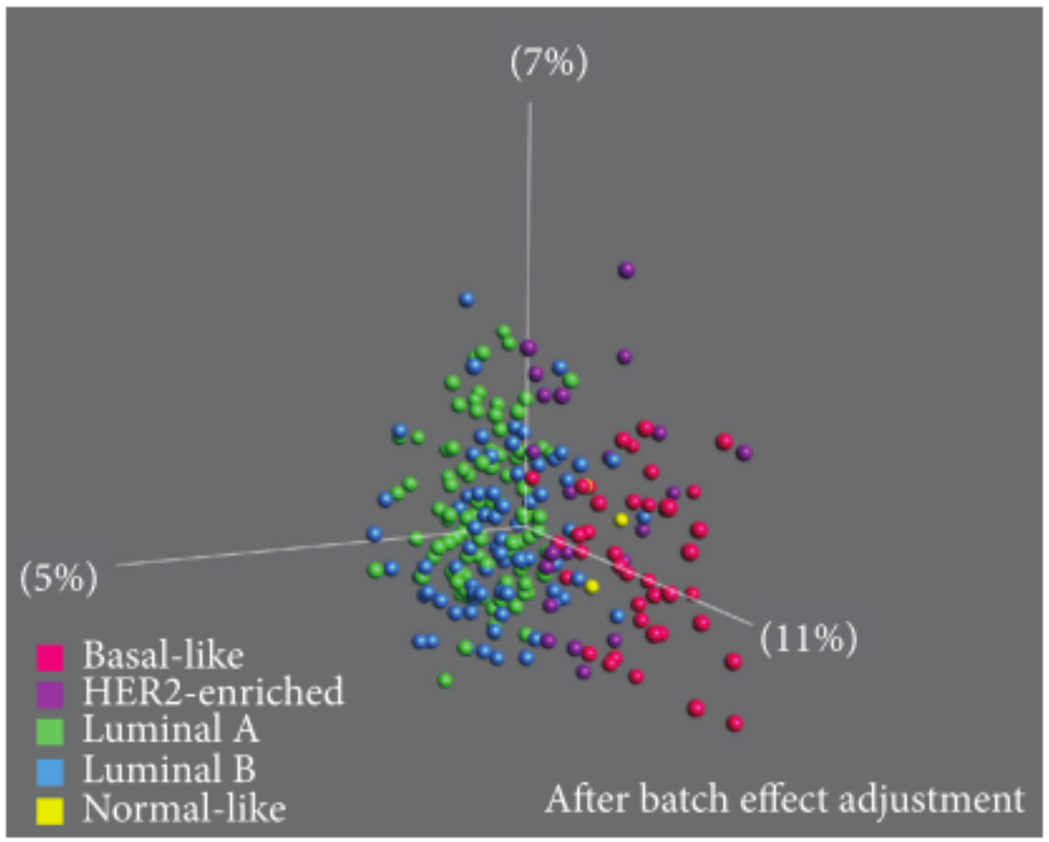
Larsen et al., BioMed Research International 2014

Accounting for batch effects in practice

- 5-subtype breast cancer microarray data processed in six batches.



(b)



(d)

Larsen et al., BioMed Research International 2014

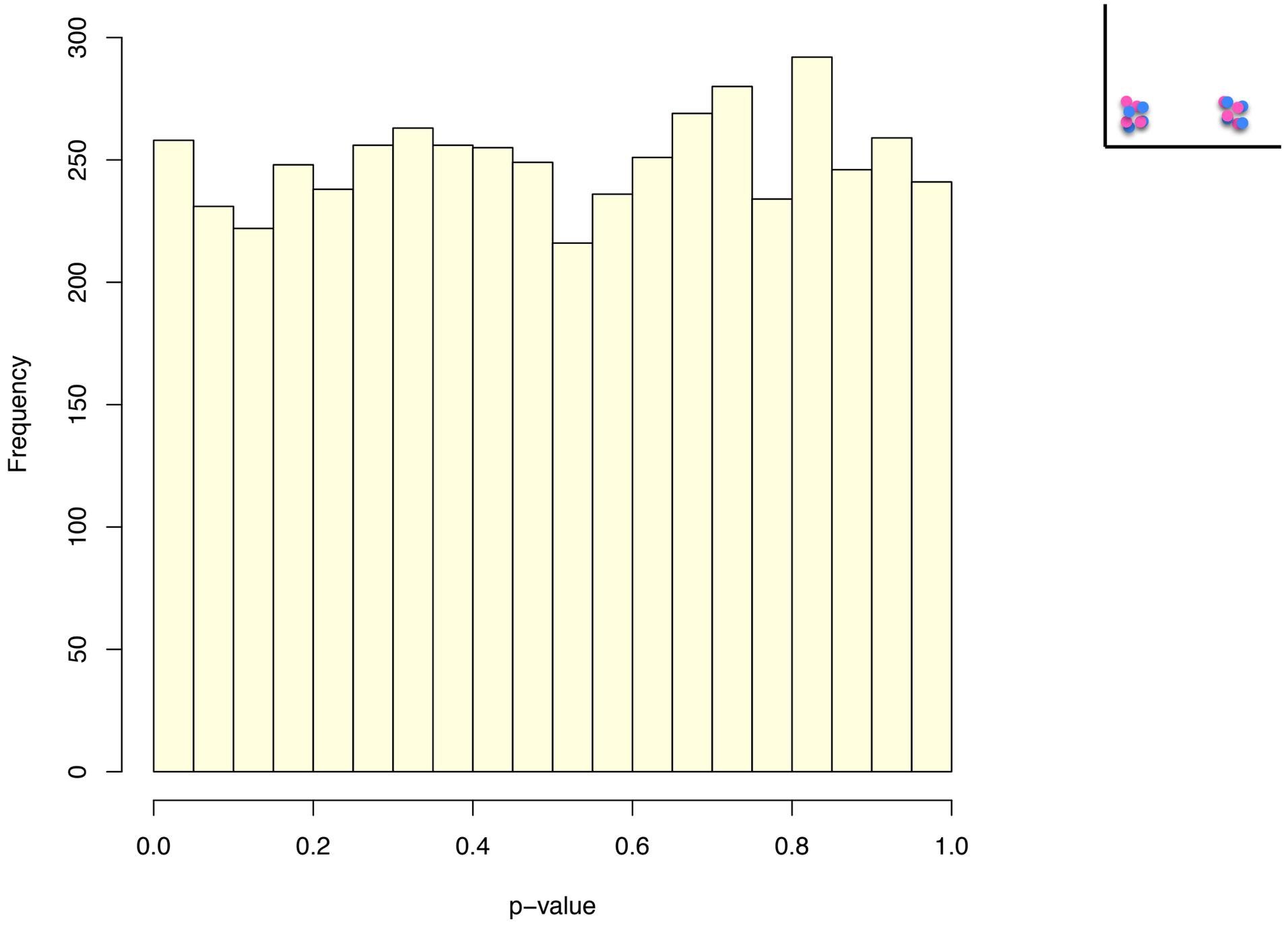
What if the batch variable is unknown?

- Manifests as systematic “unwanted variation” in data
- Identify using e.g.
  - control genes (“housekeeping” genes, spike-ins)
  - residuals after eliminating known signal
- Include estimated unwanted variation as covariate(s) in the statistical model
- **RUV**, **sva** packages commonly used in genomics

Leek & Storey (2007); Gagnon-Bartsch & Speed (2012); Leek (2014); Risso et al (2014)

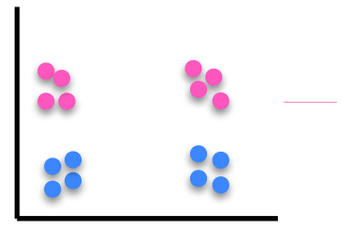
Impact of batch effect on p-value histogram

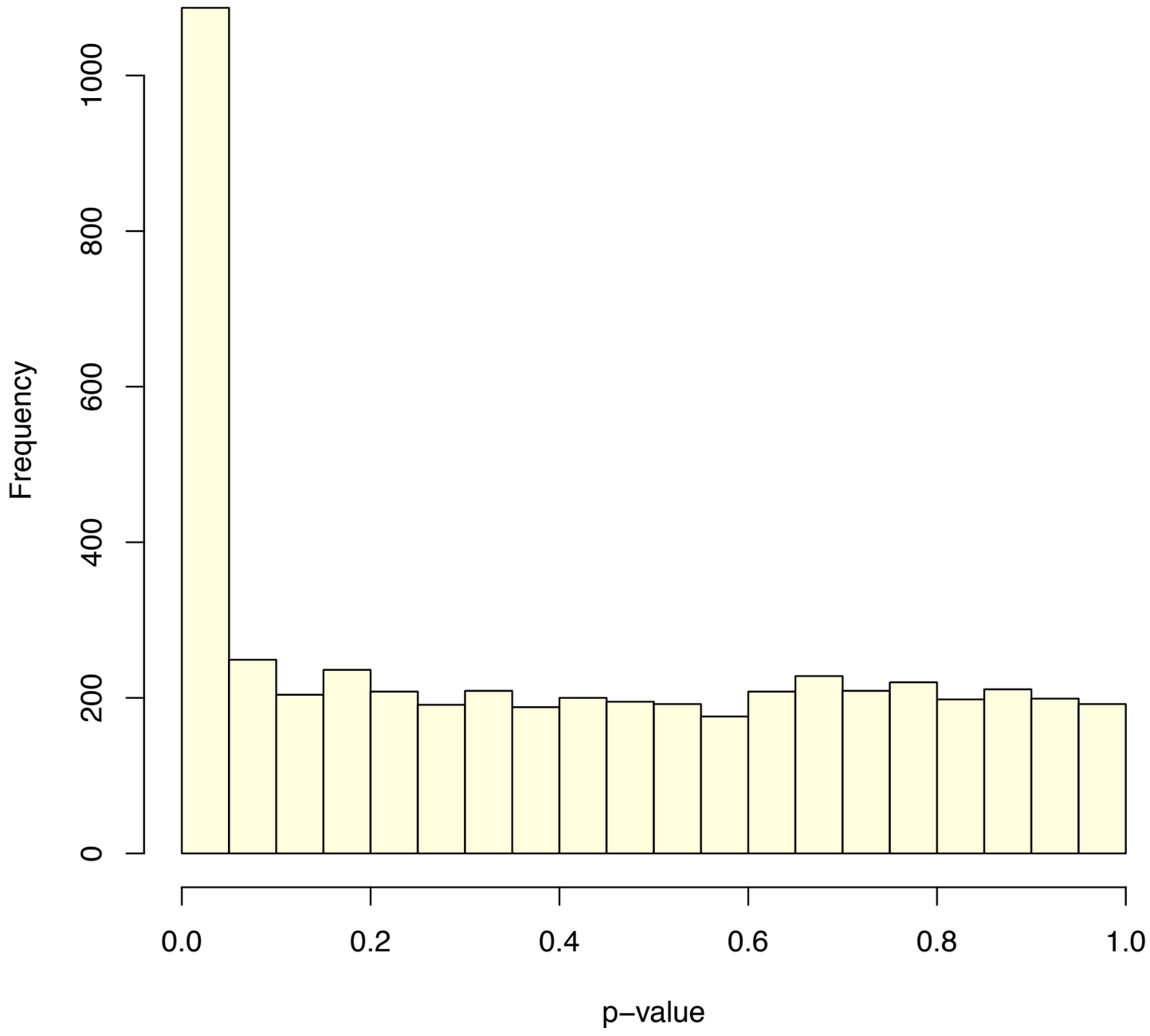
- No batch effect, no differentially expressed genes



Impact of batch effect on p-value histogram

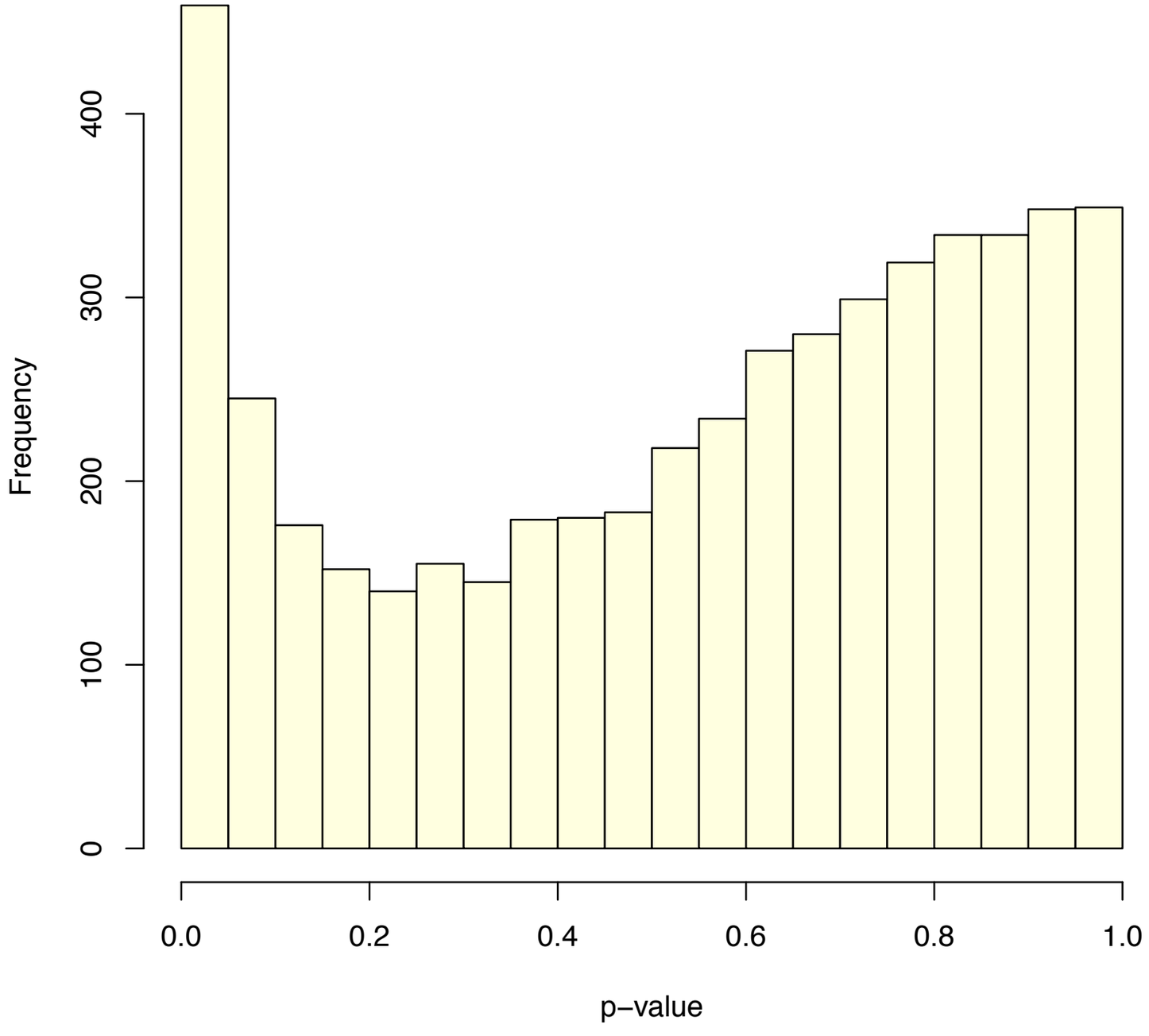
- No batch effect, some differentially expressed genes

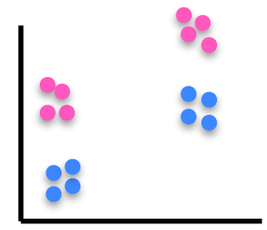


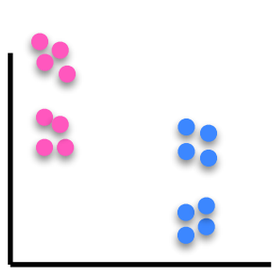


Impact of batch effect on p-value histogram

- Batch effect (no confounding), some differentially expressed genes

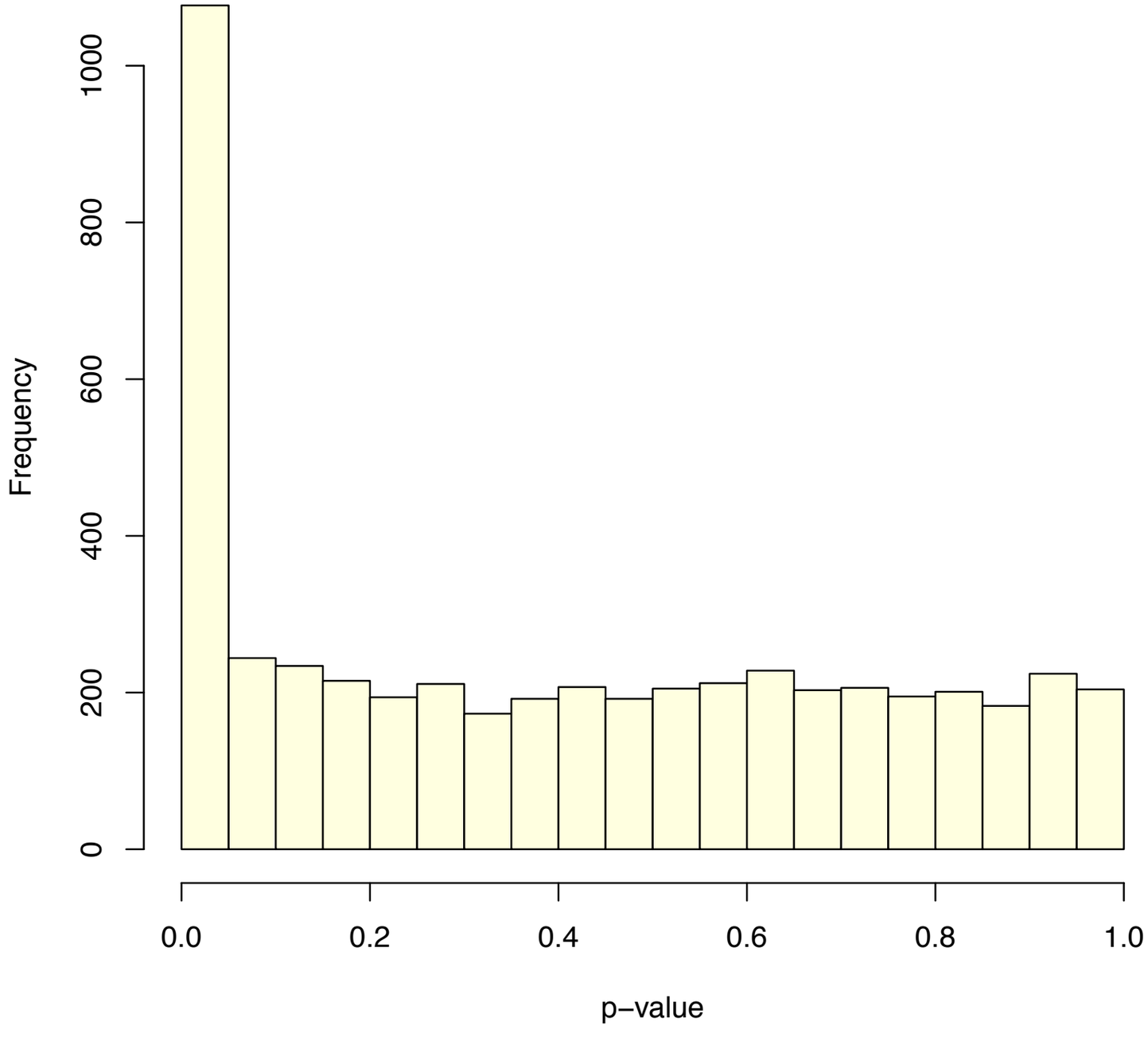


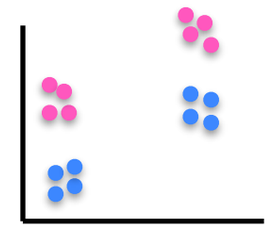


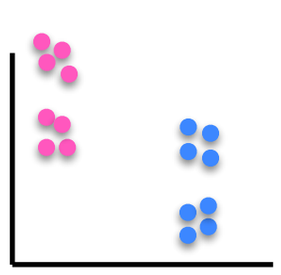


Impact of batch effect on p-value histogram

- Batch effect (no confounding), some differentially expressed genes - **after correction**

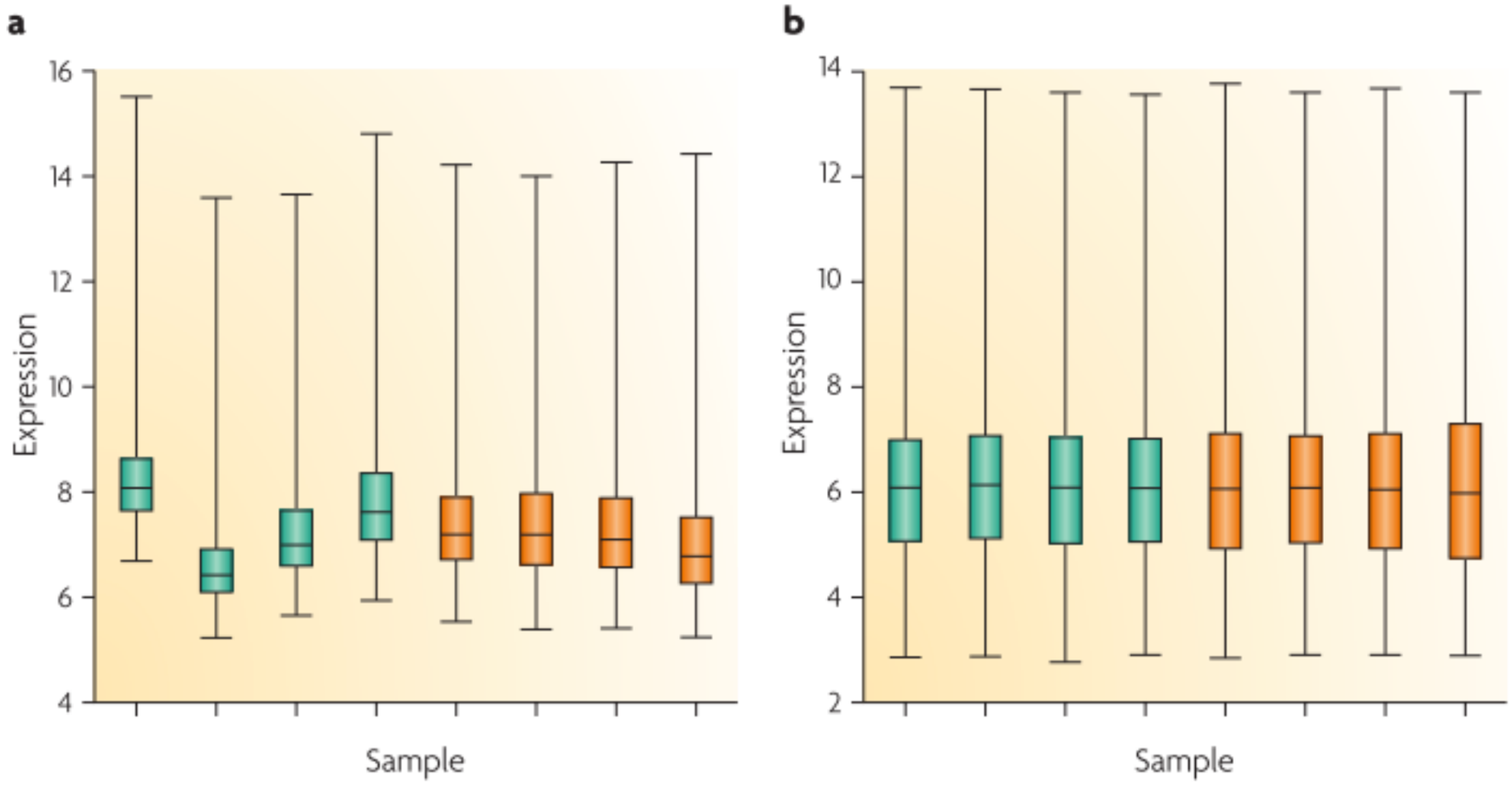






Batch effect adjustment vs normalization

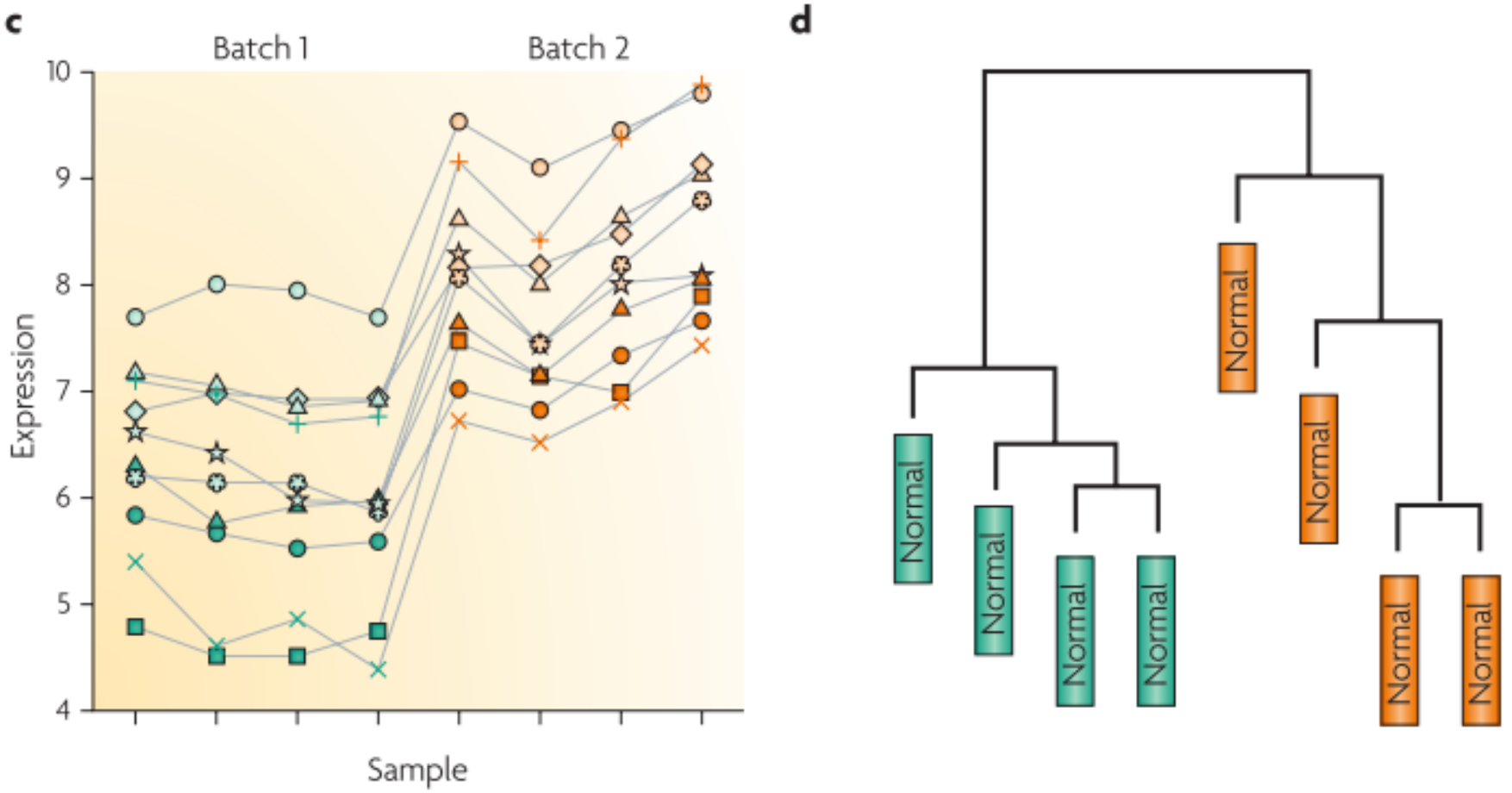
- Batch effect adjustment goes **beyond** the “global” between-sample normalization methods.



Leek et al., Nature Reviews Genetics 2010

Batch effect adjustment vs normalization

- Batch effect adjustment goes **beyond** the “global” between-sample normalization methods.

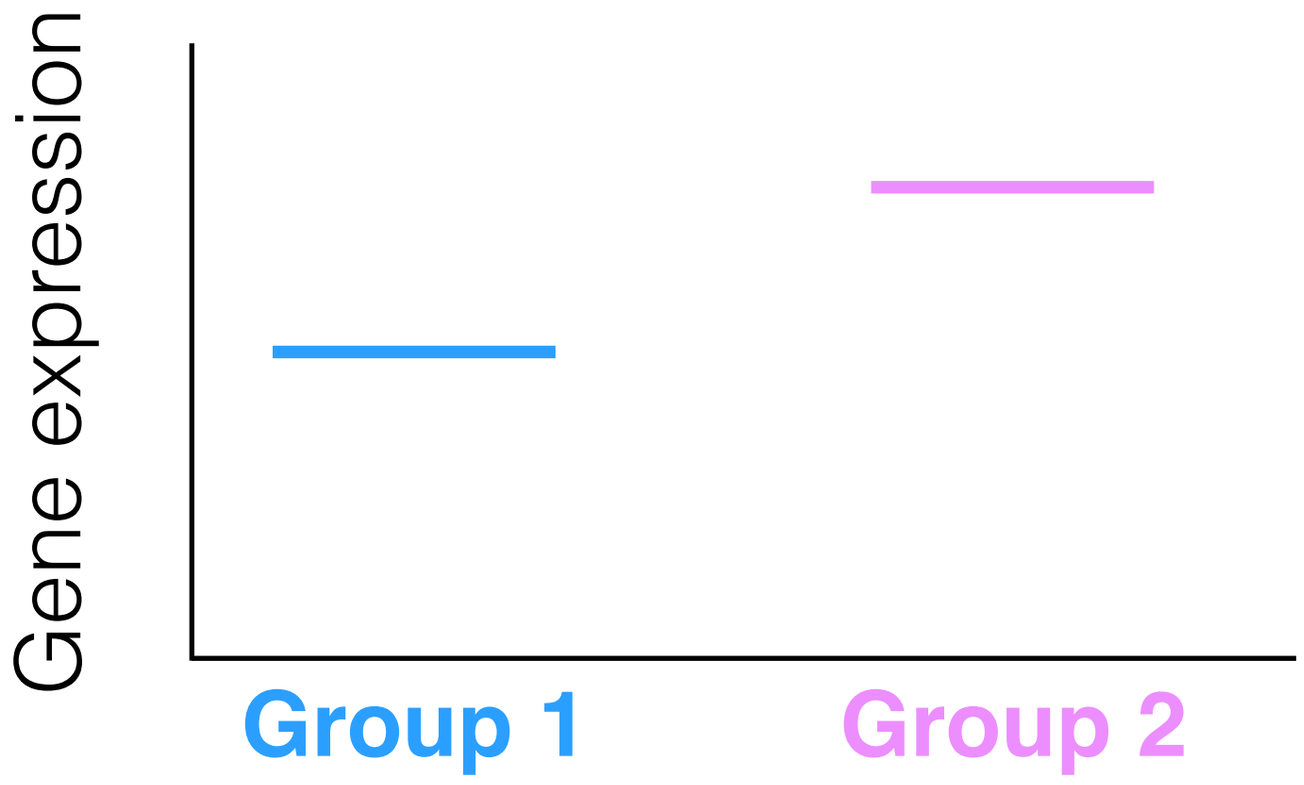


Leek et al., Nature Reviews Genetics 2010

Other design issues: replication

- Replicates are **necessary** to estimate within-condition variability.

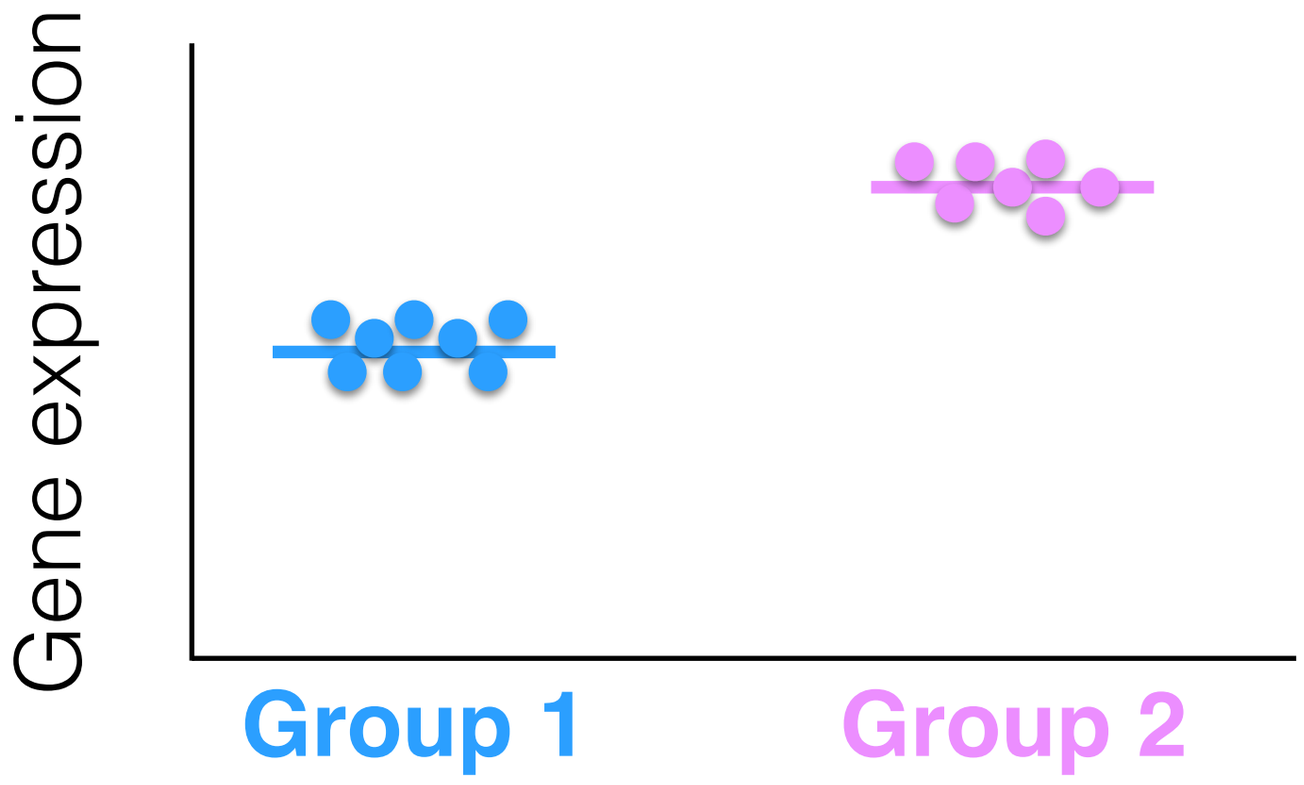
- Variability estimates are, in turn, **vital** for statistical testing.



Other design issues: replication

- Replicates are **necessary** to estimate within-condition variability.

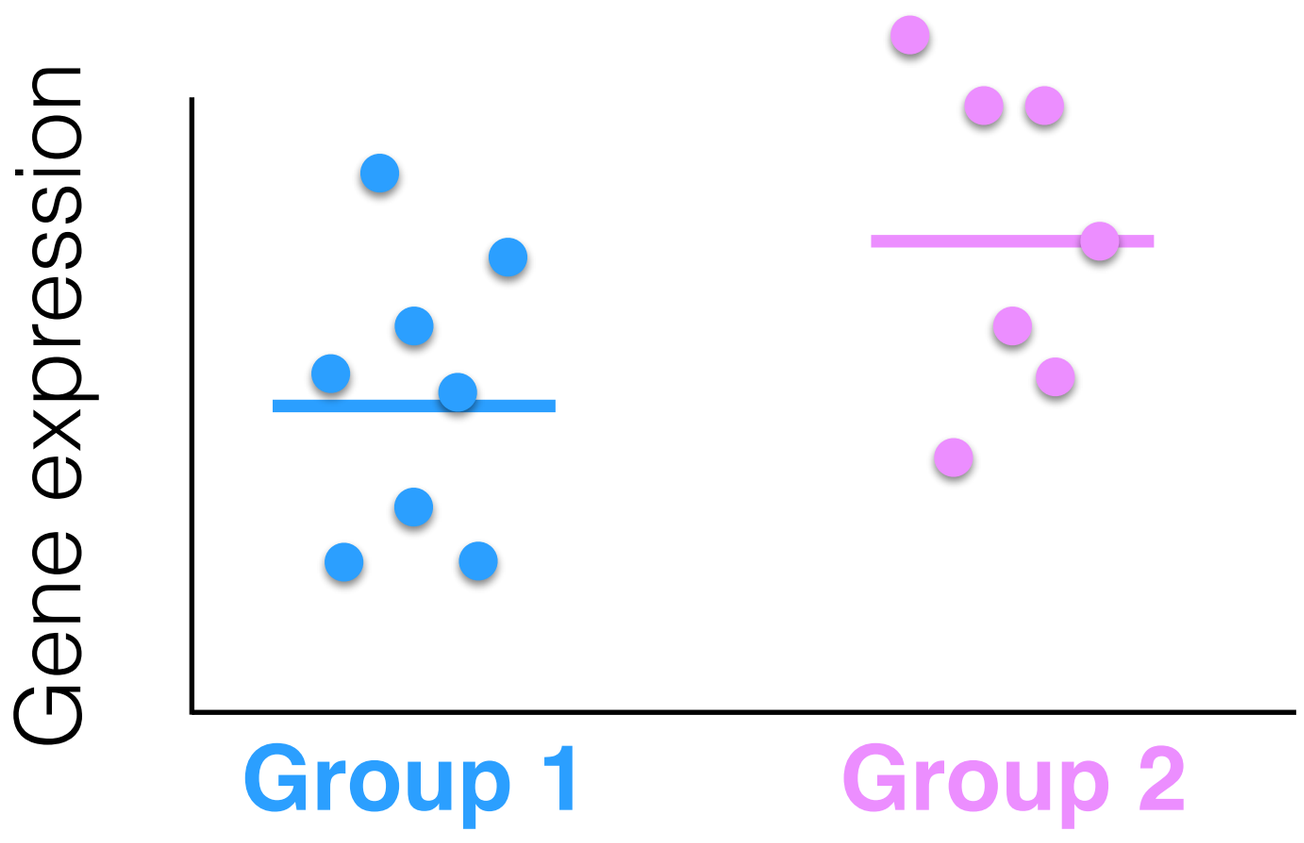
- Variability estimates are, in turn, **vital** for statistical testing.



Other design issues: replication

- Replicates are **necessary** to estimate within-condition variability.

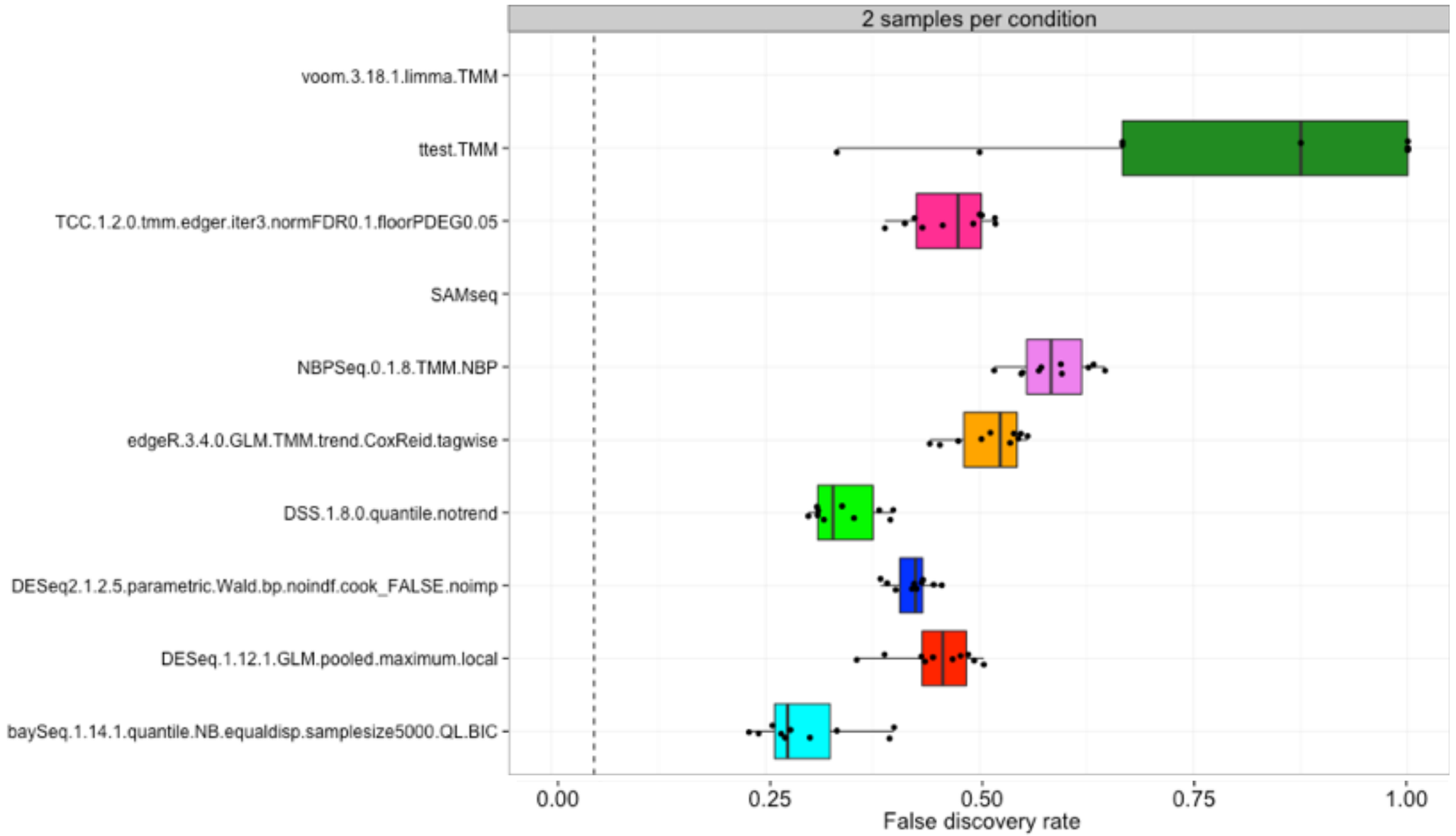
- Variability estimates are, in turn, **vital** for statistical testing.



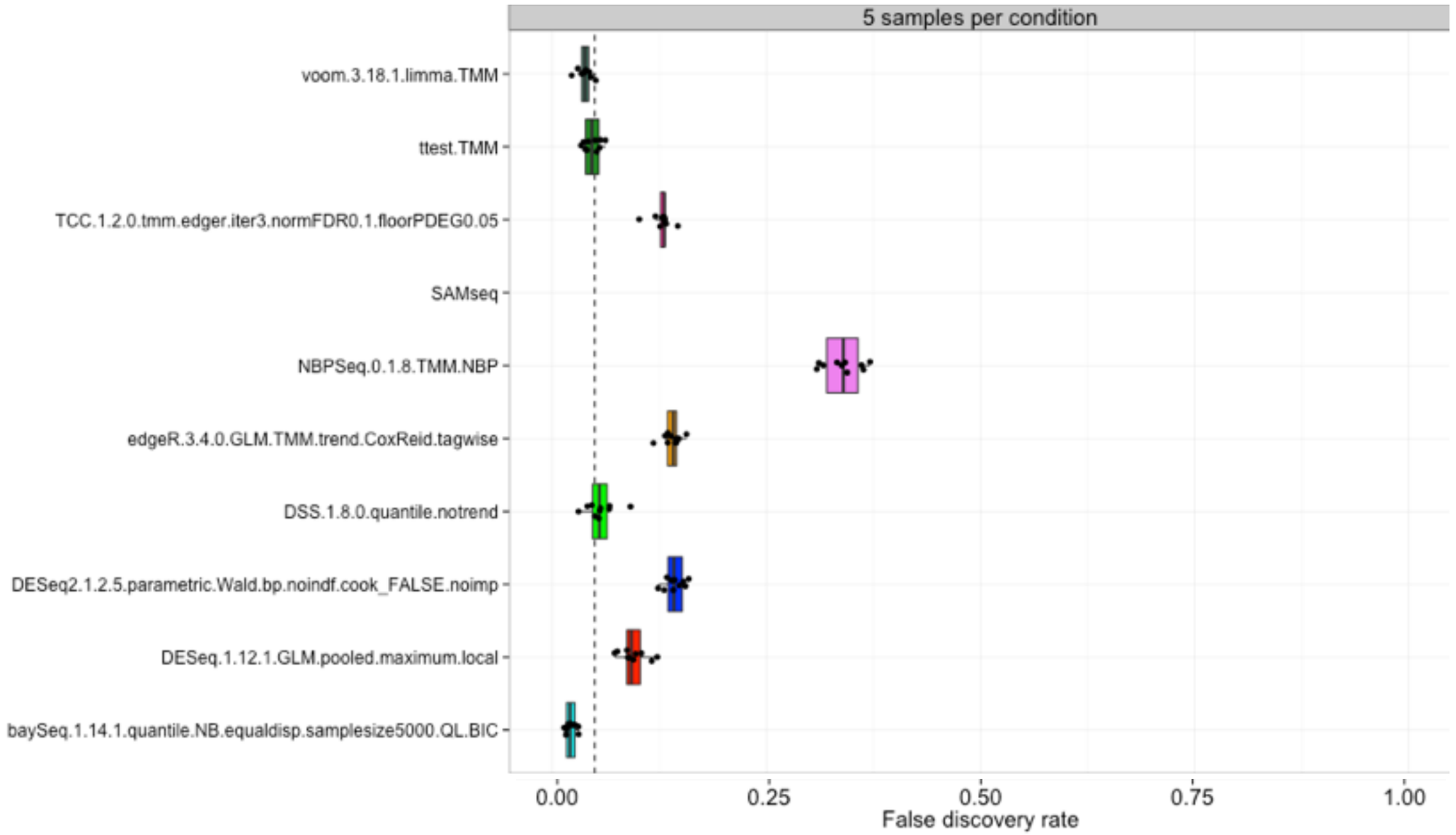
Other design issues: sample size

- As always, it depends…
    - on what we want to do (differential gene expression, variant detection, GWAS, …)
    - on the variability between samples (cell lines, inbred animals, patients, …)
    - on the magnitude of the expected effect

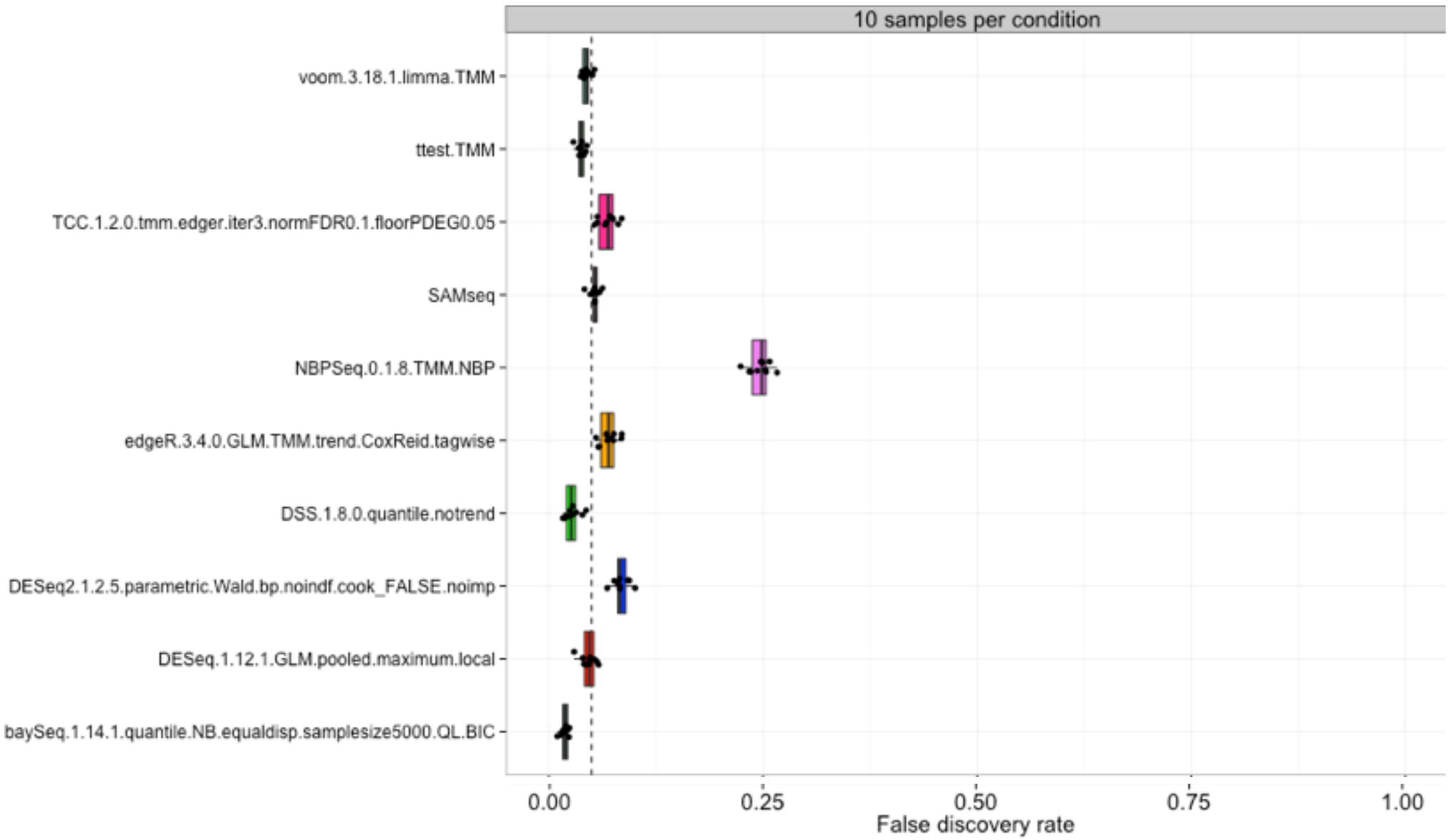
FDR, 2 replicates/condition



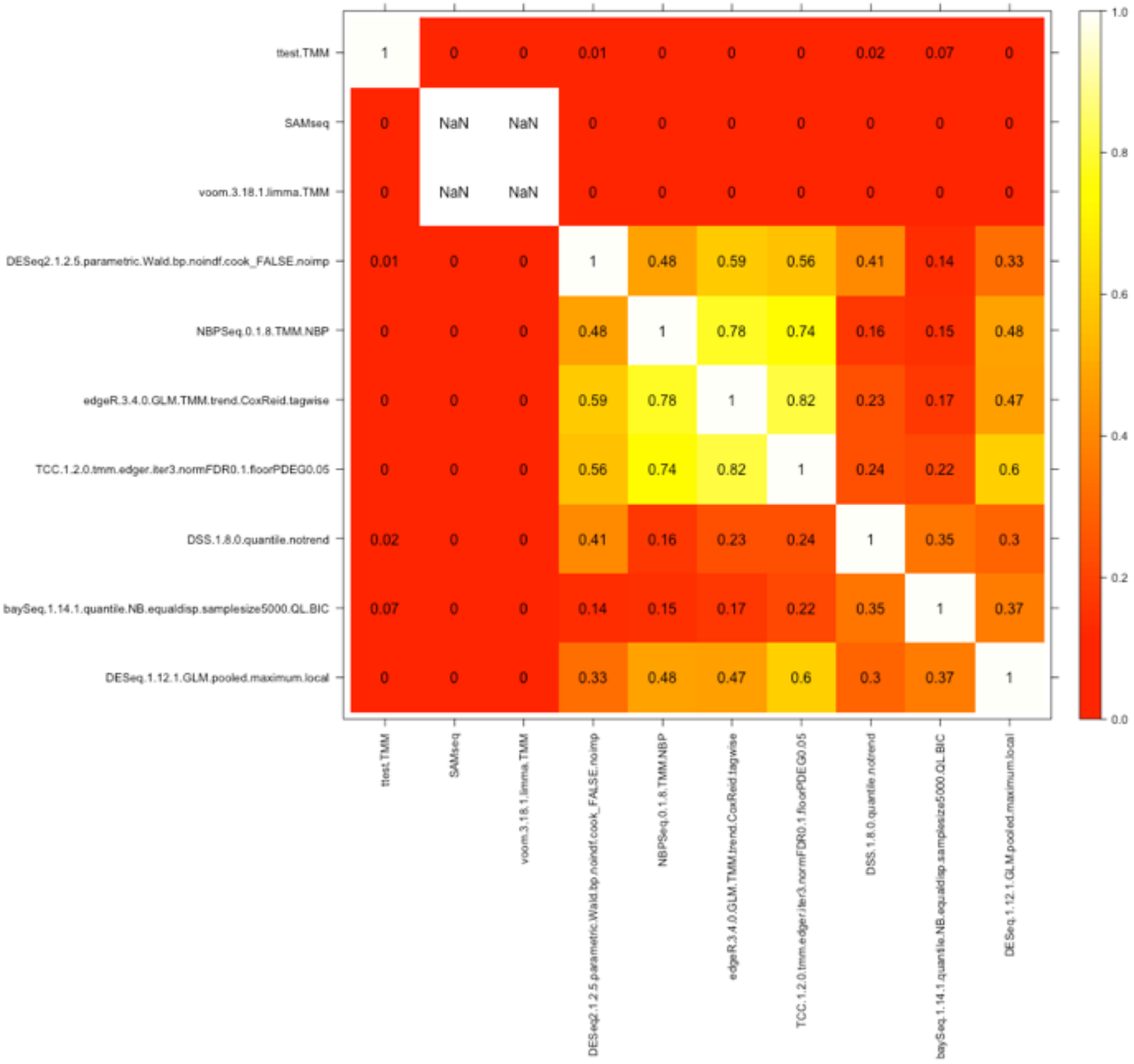
FDR, 5 replicates/condition



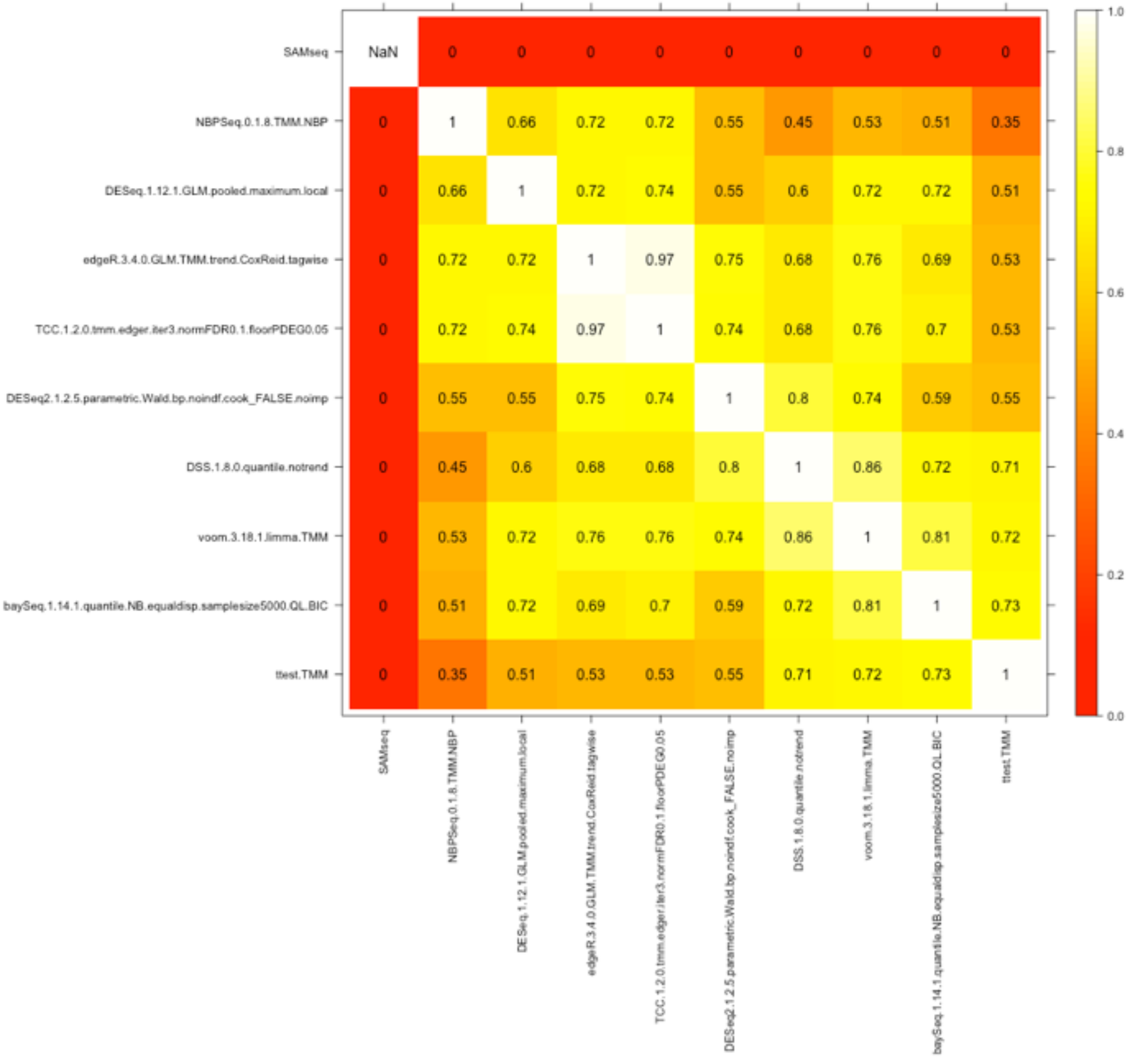
FDR, 10 replicates/condition



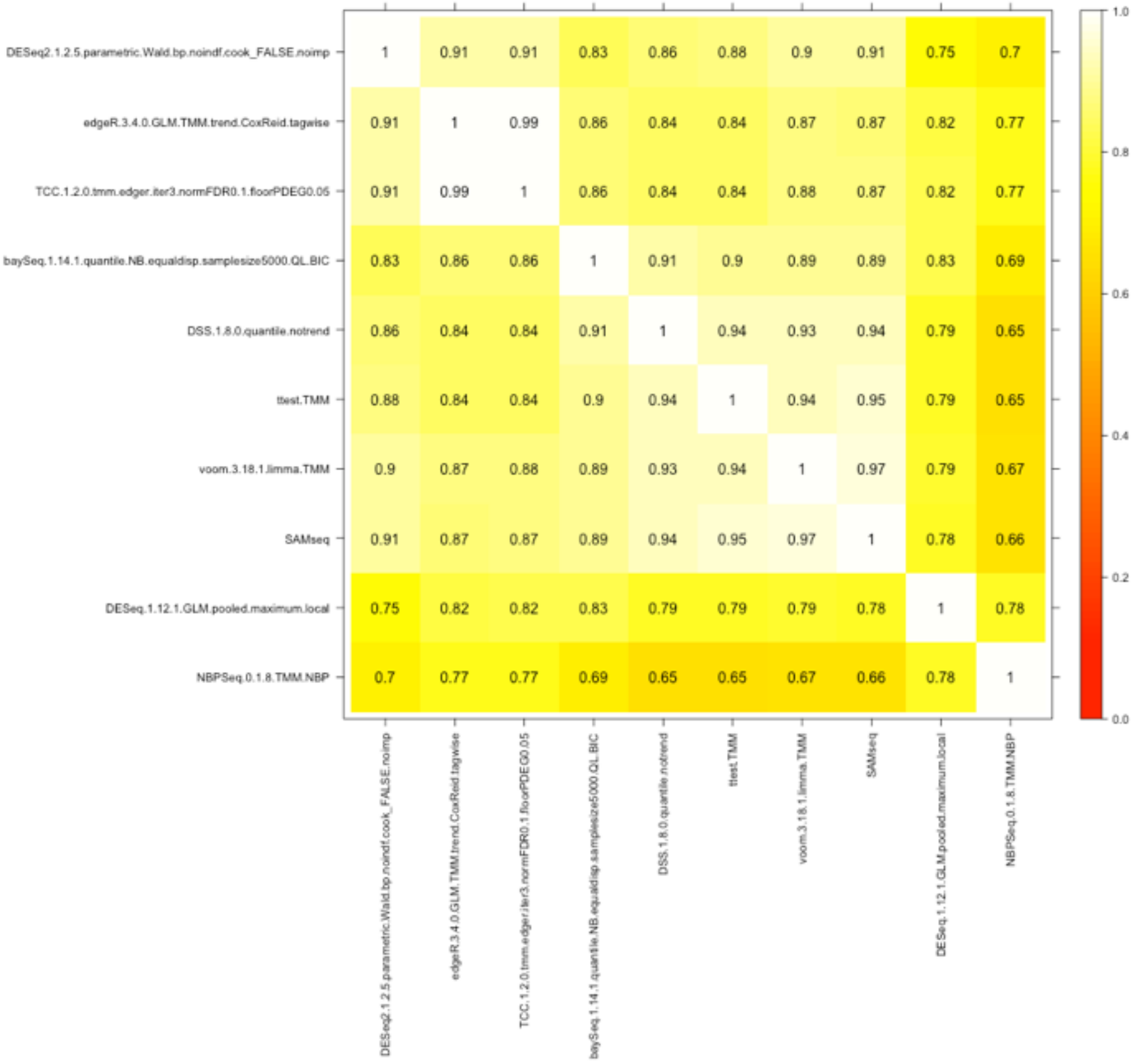
Similarity between sets of DEGs, 2 replicates/condition



Similarity between sets of DEGs, 5 replicates/condition



Similarity between sets of DEGs, 10 replicates/condition



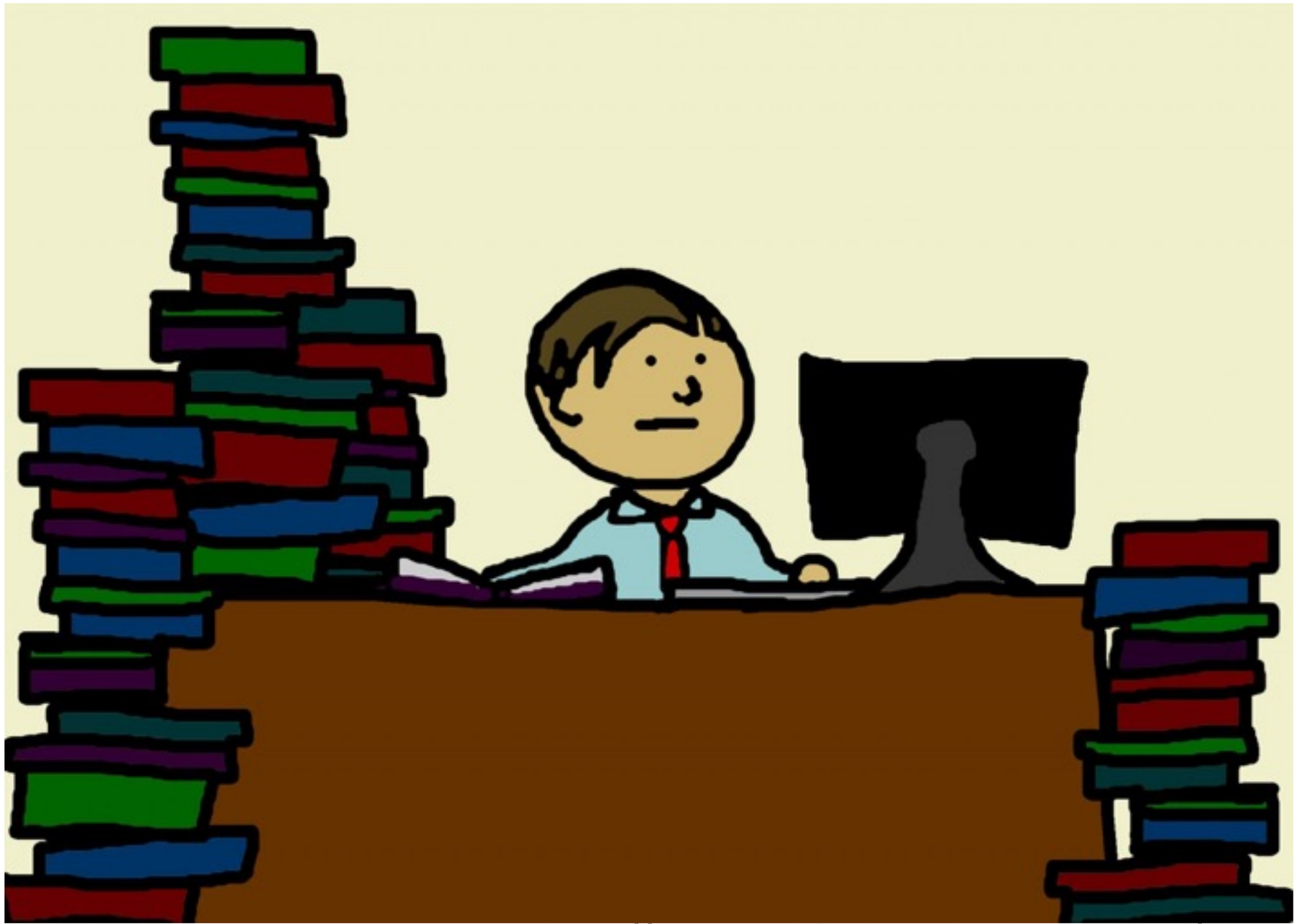
How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?

Nicholas J. Schurch[1,6], Pietá Schofield[1,2,6], Marek Gierliński[1,2,6], Christian Cole[1,6], Alexander Sherstnev[1,6], Vijender Singh[2], Nicola Wrobel[3], Karim Gharbi[3], Gordon G. Simpson[4], Tom Owen-Hughes[2], Mark Blaxter[3] and Geoffrey J. Barton[1,2,5]

At least six replicates per condition for all experiments.
At least 12 replicates per condition for experiments where identifying the majority of all DE genes is important.

Schurch et al., RNA 2016

And now for something completely different…



http://www.publicdomainpictures.net/view-image.php?imag

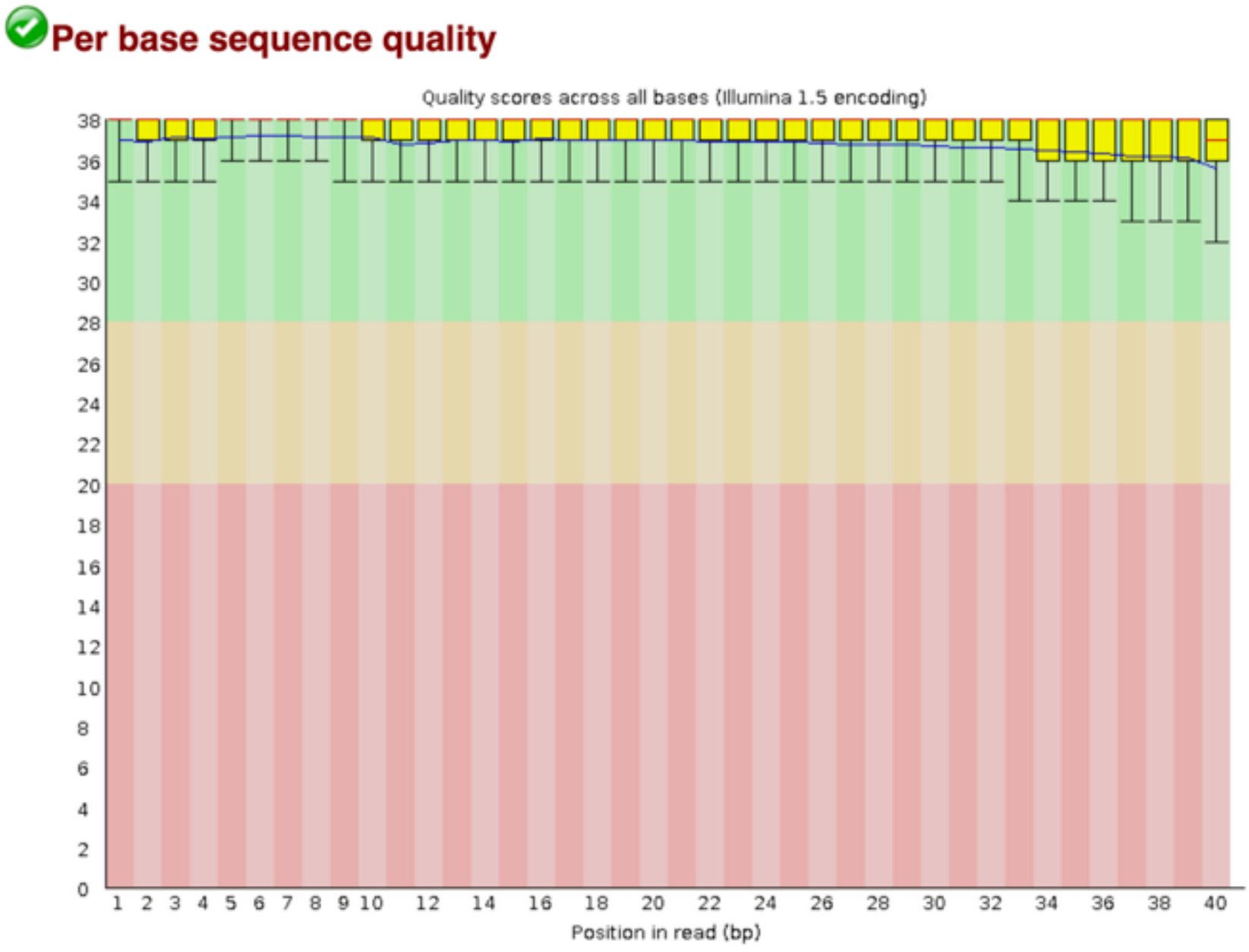
No matter how carefully you design your experiment, data can still be compromised…

- Contamination
- Sequencing failures
- Remaining adapters
- PCR duplicates
- …

QC software for NGS data - example

FastQC - raw read QC

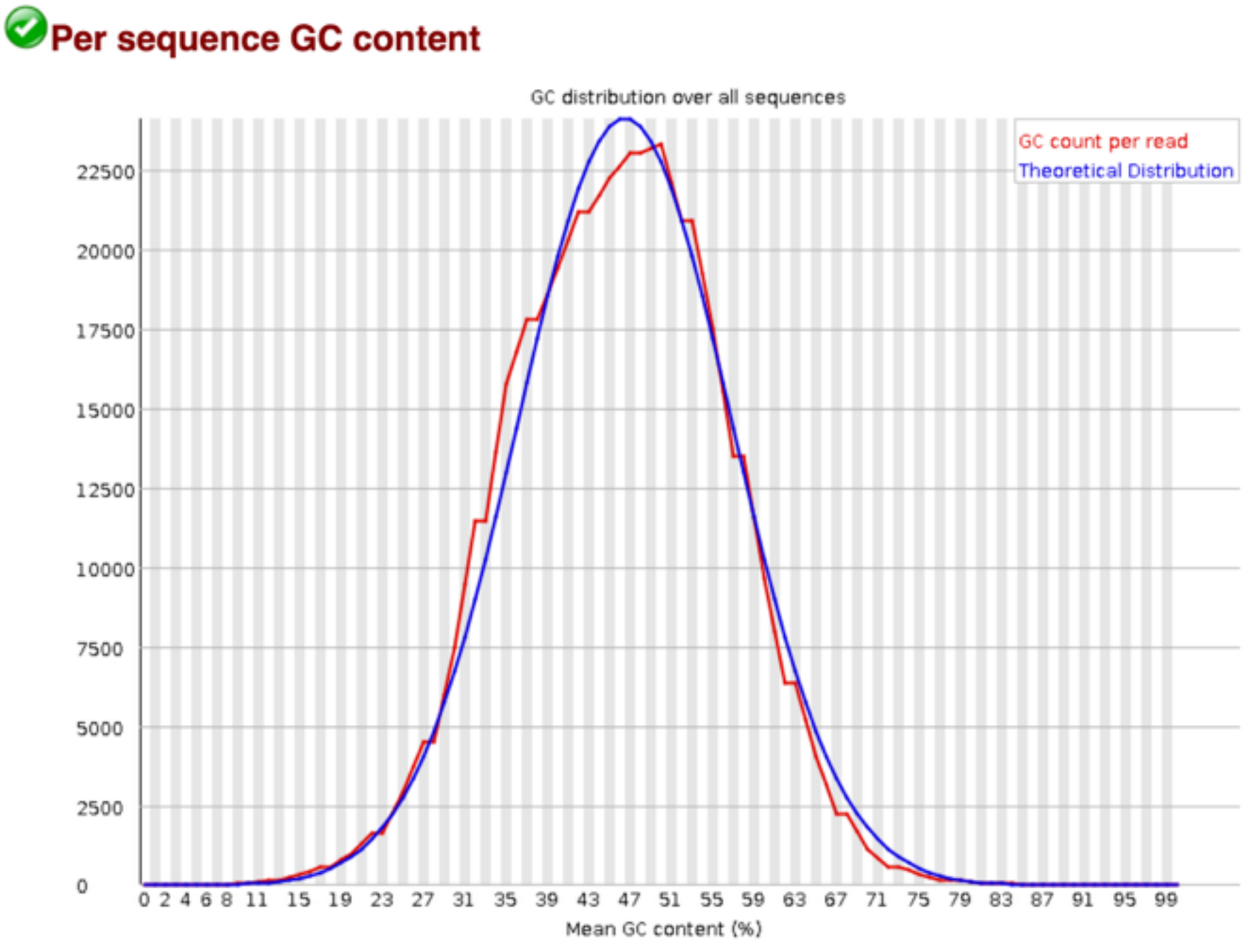
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/



QC software for NGS data - example

FastQC - raw read QC

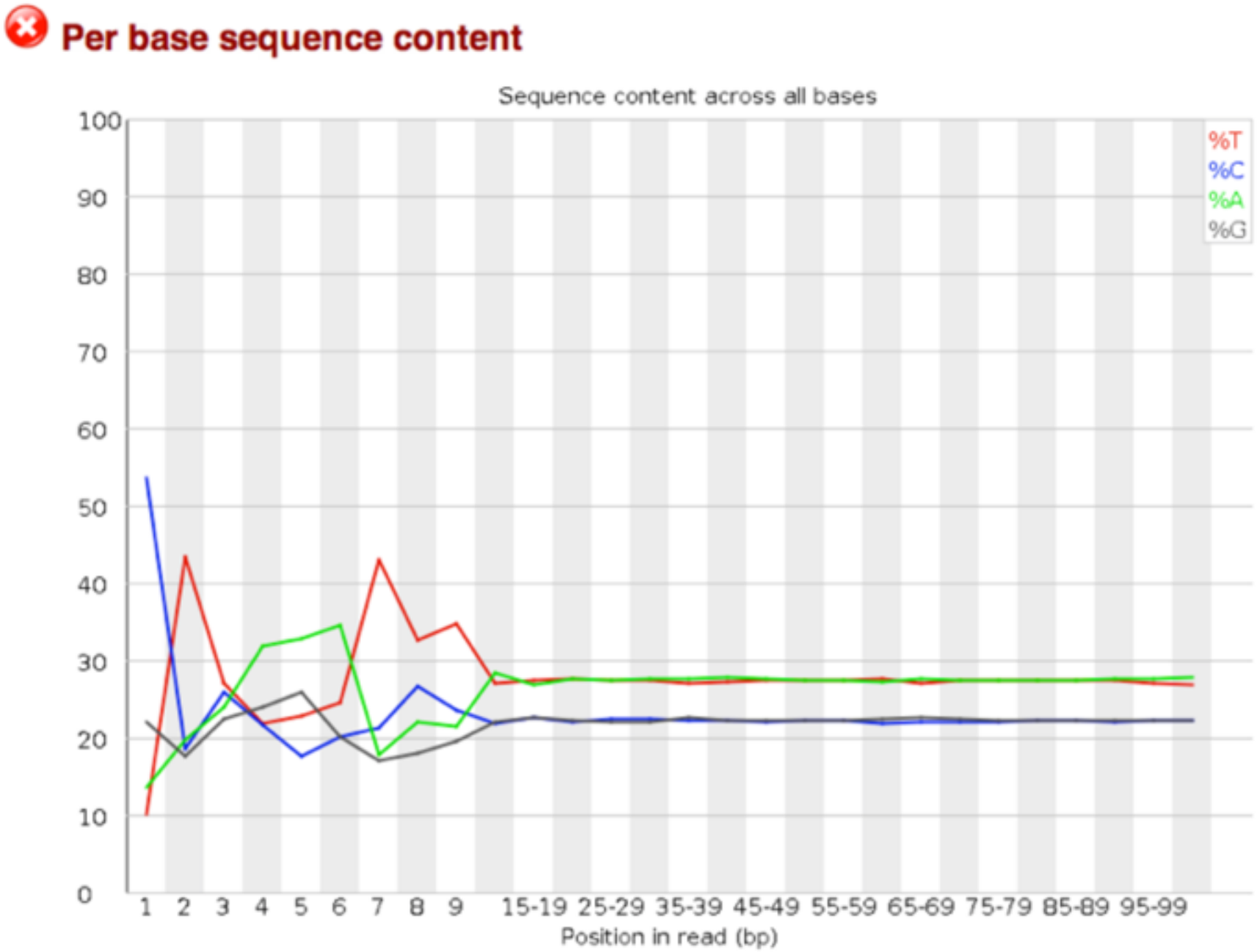
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/



QC software for NGS data - example

FastQC - raw read QC

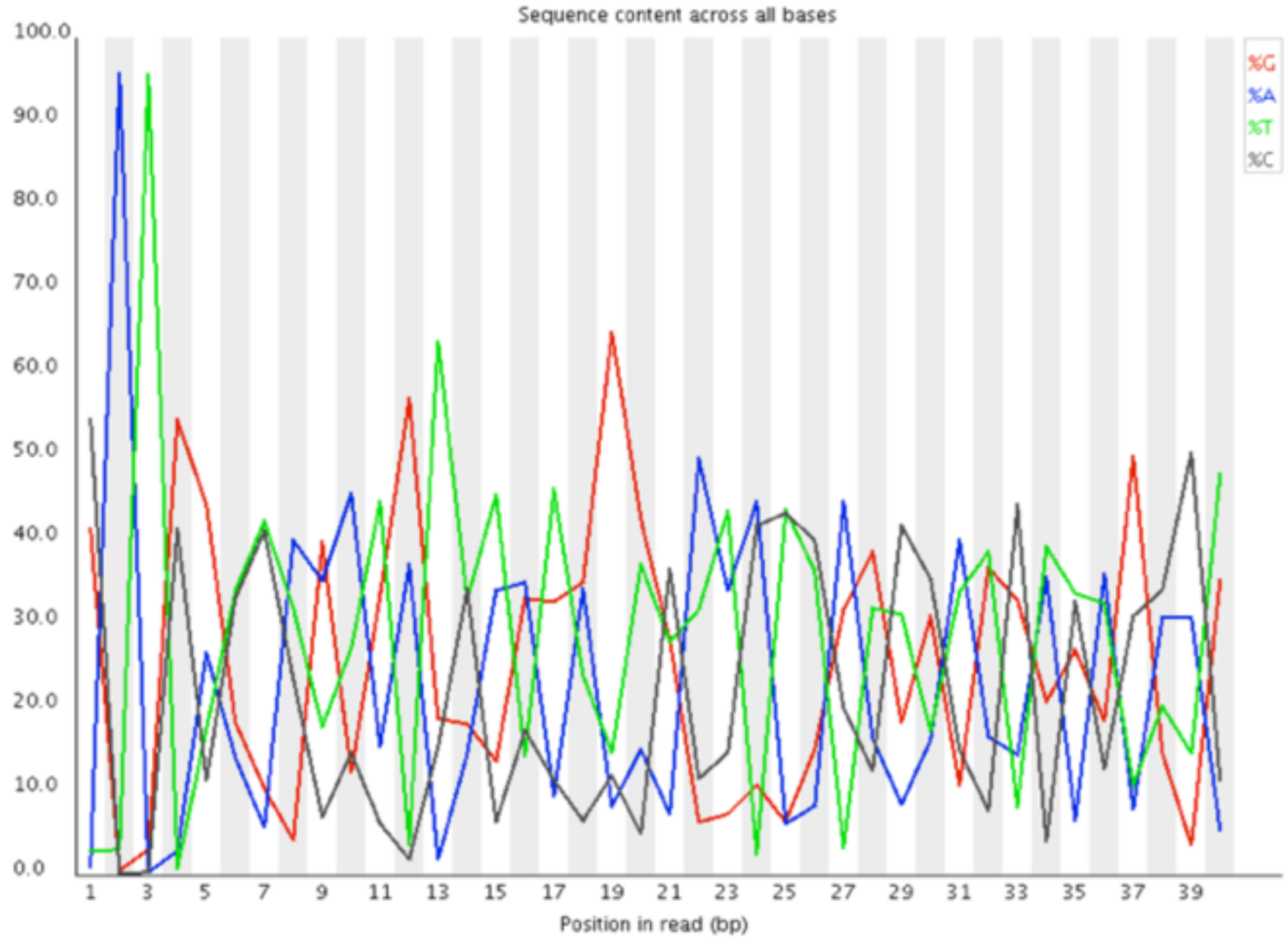
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/



QC software for NGS data - example

FastQC - raw read QC

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/



QC software for NGS data - example

FastQC - raw read QC

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT	8122	8.122	Illumina Paired End PCR Primer 2 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAG	5086	5.086	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	1085	1.085	Illumina Single End PCR Primer 1 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGAAG	508	0.508	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATTATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	242	0.242	Illumina Single End PCR Primer 1 (97% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAAGATCGGAA	235	0.23500000000000001	Illumina Paired End Adapter 2 (96% over 31bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCGAGATCGGAAGA	228	0.22799999999999998	Illumina Paired End Adapter 2 (96% over 28bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACG	205	0.20500000000000002	Illumina Paired End PCR Primer 2 (97% over 36bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGATCGGAA	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGTCGGAAG	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)

QC software for NGS data - example

multiQC - summarize results from many analyses

http://multiqc.info/docs/#

MultiQC
v0.6

- featureCounts
- STAR
- Cutadapt
- FastQC
  - Sequence Quality Histograms
  - Per Sequence Quality Scores
  - Per Base Sequence Content
  - Per Sequence GC Content
  - Per Base N Content
  - Sequence Length Distribution
  - Sequence Duplication Levels
  - Adapter Content

MultiQC

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2016-05-03, 08:05 based on data in: `/Users/philewels/Work/MultiQC_website/public_html/examples/rna-seq/data`

General Statistics

Copy table | Configure Columns | Showing 8 rows.

Sample Name	% Assigned	M Assigned	% Aligned	M Aligned	% Trimmed	% Dups	% GC	Length	M Seqs
SRR3192396	67.5%	71.9	93.7%	97.8	4.0%	78.9%	51%	97	104.4
SRR3192397	66.6%	63.0	94.7%	87.1	3.5%	77.2%	49%	97	92.0
SRR3192398	50.9%	36.5	88.2%	58.7	5.0%	55.3%	47%	97	66.6
SRR3192399	52.3%	42.3	88.2%	65.6	5.0%	57.4%	47%	97	74.3
SRR3192400	70.3%	63.4	77.3%	73.4	7.2%	74.1%	45%	93	94.9
SRR3192401	71.2%	63.8	76.4%	72.8	6.3%	76.3%	45%	94	95.2
SRR3192657	73.1%	67.1	91.2%	85.0	3.1%	82.2%	51%	98	93.1
SRR3192658	71.2%	66.9	89.7%	87.1	3.4%	82.3%	52%	97	97.1

Toolbox

References

- Ewels et al.: MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics (2016) - **multiQC**
- Akay et al.: On the design and analysis of gene expression studies in human populations. Nature Genetics 39(7): 807-808 (2007)
- Nygaard et al.: Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics 17(1):29-39 (2016)
- Danielsson et al.: Assessing the consistency of public human tissue RNA-seq data sets. Briefings in Bioinformatics 16(6):941-949 (2015)
- Larsen et al.: Microarray-based RNA profiling of breast cancer: batch effect removal improves cross-platform consistency. BioMed Research International vol. 2014, article ID 651751 (2014)
- Leek et al.: Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics 11(10:733-739 (2010)
- Leek & Storey: Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genetics 3(9):e161 (2007)
- Leek: svaseq: removing batch effects and other unwanted noise from sequencing data. Nucleic Acids Research 42(42):e161 (2014)
- Risso et al.: Normalization of RNA-seq data using factor analysis of control genes or samples. Nature Biotechnology 32(9):896-902 (2014)
- Gagnon-Bartsch & Speed: Using control genes to correct for unwanted variation in microarray data. Biostatistics 13(3):539-552 (2012)
- Schurch et al.: How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA (2016)