
WELCOME!
Summer School will begin at
8:30 a.m. PDT

EMSL

Summer School Day 4: Transcriptomics

Rachel Richardson & Lisa Bramer
Biostatistics & Data Science

07.27.2023

8:30-8:35 a.m.	Introduction	Lisa Bramer
8:35-9:25	Express Yourself!: Transcriptomics from sample to counts	Bill Nelson
9:25-10:05	Pre-processing of Transcriptomics Data	Rachel Richardson
10:05-10:15	Networking Break	
10:15-11:00	Statistical Models for Transcriptomics Data	Lisa Bramer
11:00-11:45	Post-differential Expression Analyses (or how to interpret that big ole pile of data you've just generated)	Jason McDermott

8:30-8:35 a.m.	Introduction	Lisa Bramer
8:35-9:25	**Express Yourself!: Transcriptomics from sample to counts**	**Bill Nelson**
9:25-10:05	Pre-processing of Transcriptomics Data	Rachel Richardson
10:05-10:15	Networking Break	
10:15-11:00	Statistical Models for Transcriptomics Data	Lisa Bramer
11:00-11:45	Post-differential Expression Analyses (or how to interpret that big ole pile of data you've just generated)	Jason McDermott

EMSL

Express Yourself!:
Transcriptomics from sample to counts

Bill Nelson
Computational Scientist

Overview

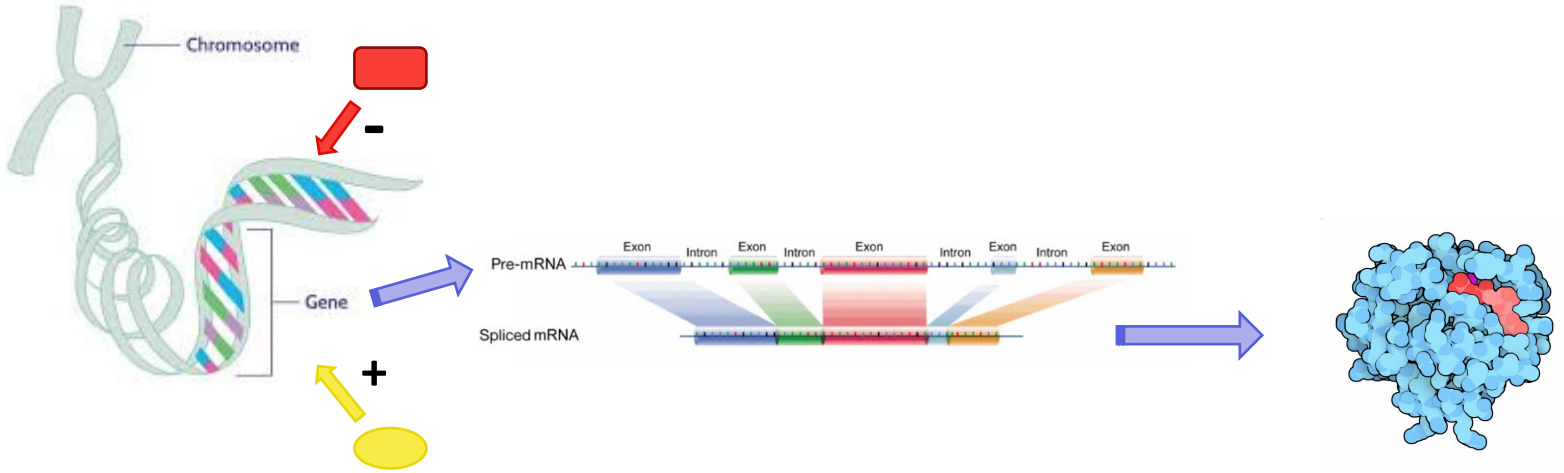
- What is transcriptomics?
- What techniques can be used to interrogate a transcriptome?
  - EST
  - SAGE
  - Microarray
  - RNA-Seq
- RNA-Seq study design considerations
- RNA-Seq process
  - Sample preparation
  - Library preparation
  - Illumina Sequencing
  - Sequence data quality checks
- Bonus round
  - Long-read sequencing
  - Single-cell transcriptomics

What is transcriptomics?

Analyzing the complement of RNA in a sample

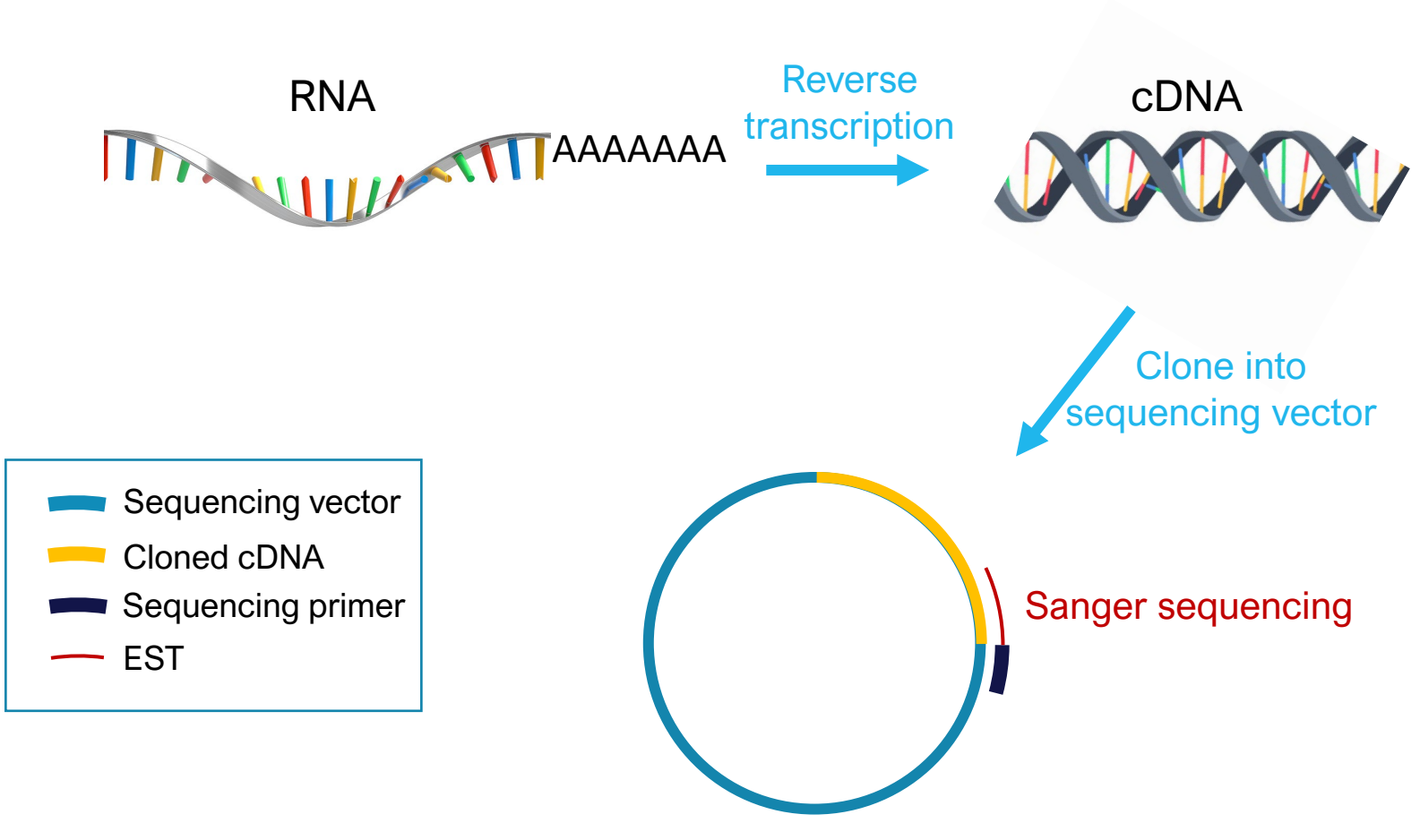
Functional genomics

- Why transcriptomics?
  - Identify expressed genes
    - Non-coding genes
  - Determine changes in gene expression under different conditions
  - Determine regulatory relationships between genes
  - Examine splice variants [Eukaryotes]
  - Transcription start sites; use of alternate promoters



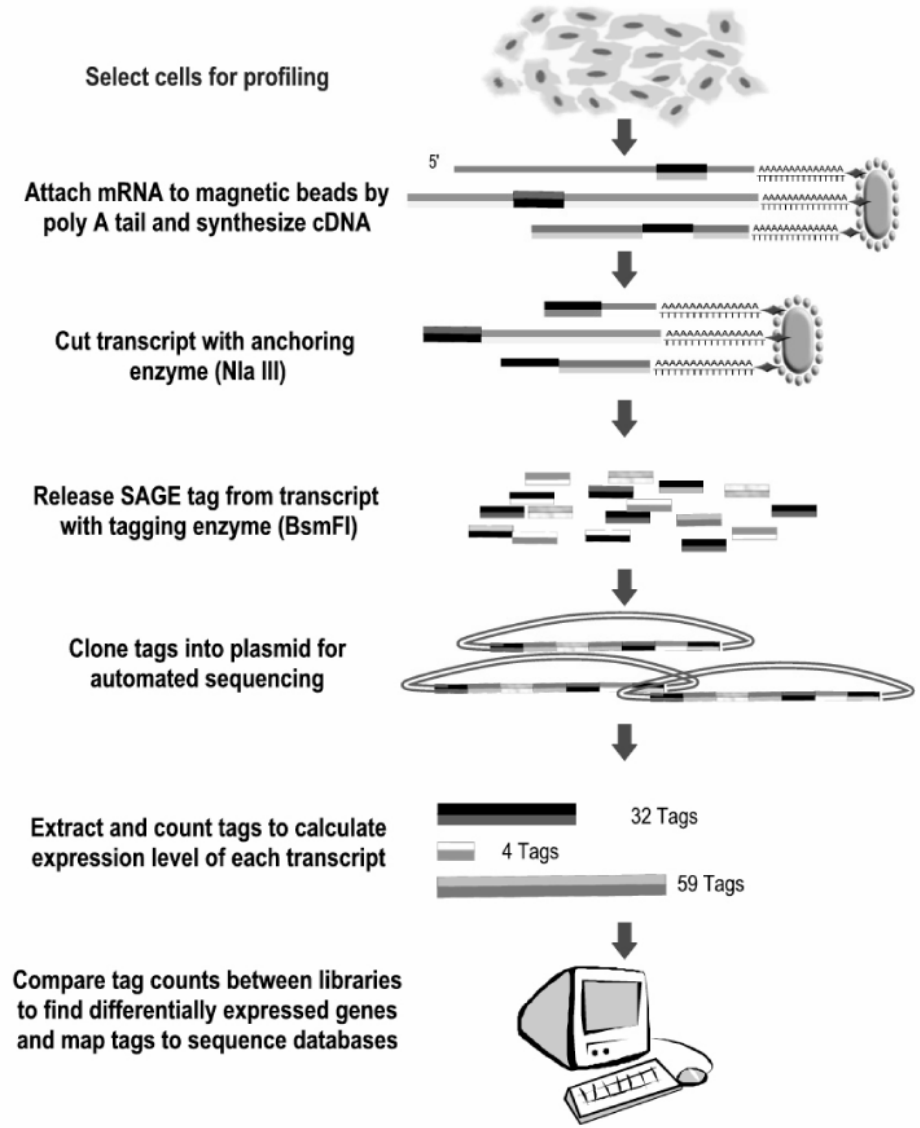
Other technologies from the past

Expressed sequence tags (EST)



Other technologies from the past

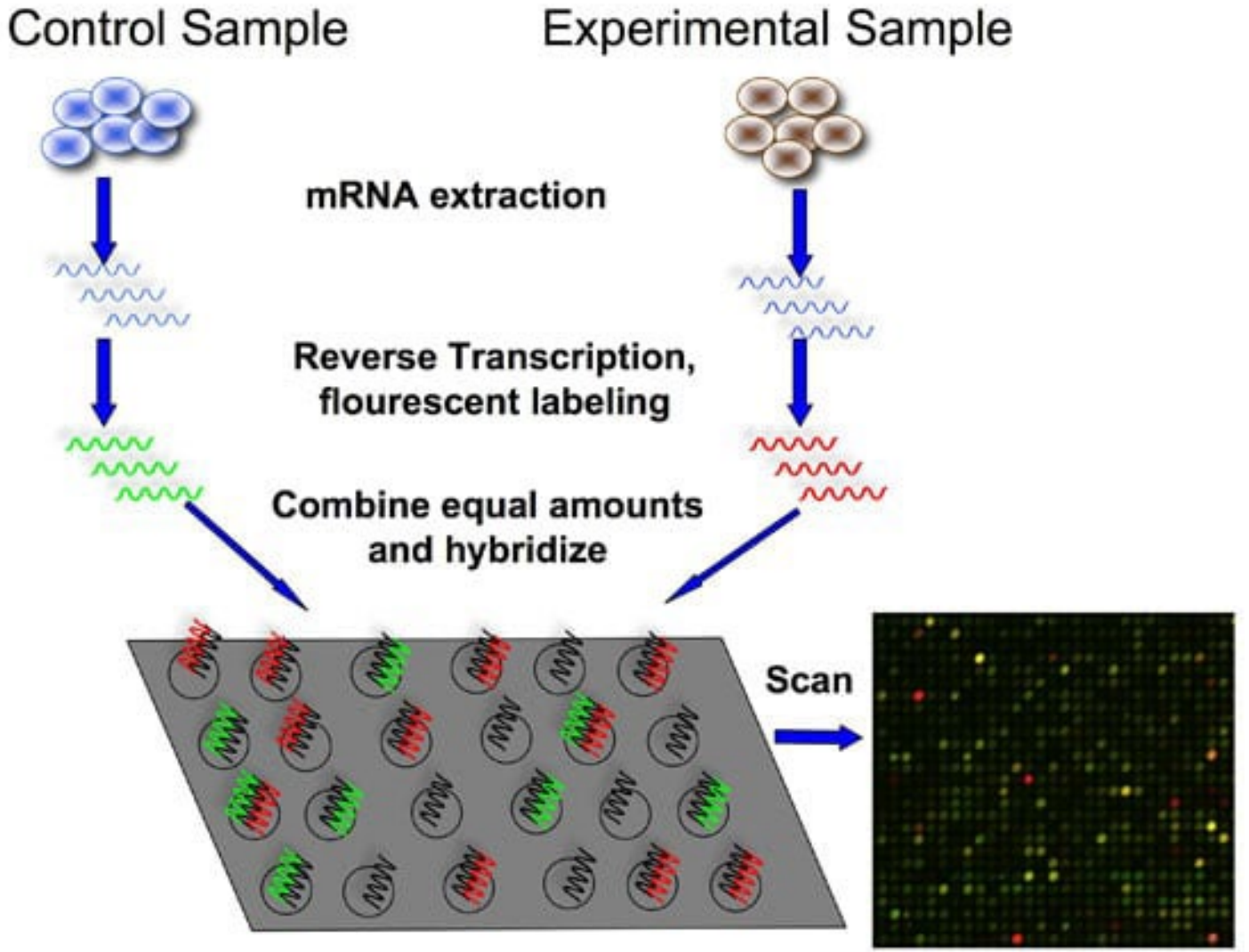
Serial Analysis of Gene Expression (SAGE)



https://www.researchgate.net/figure/Outline-of-SAGE-approach-for-transcript-profiling_fig1_11678335

Other technologies from the past

Microarray

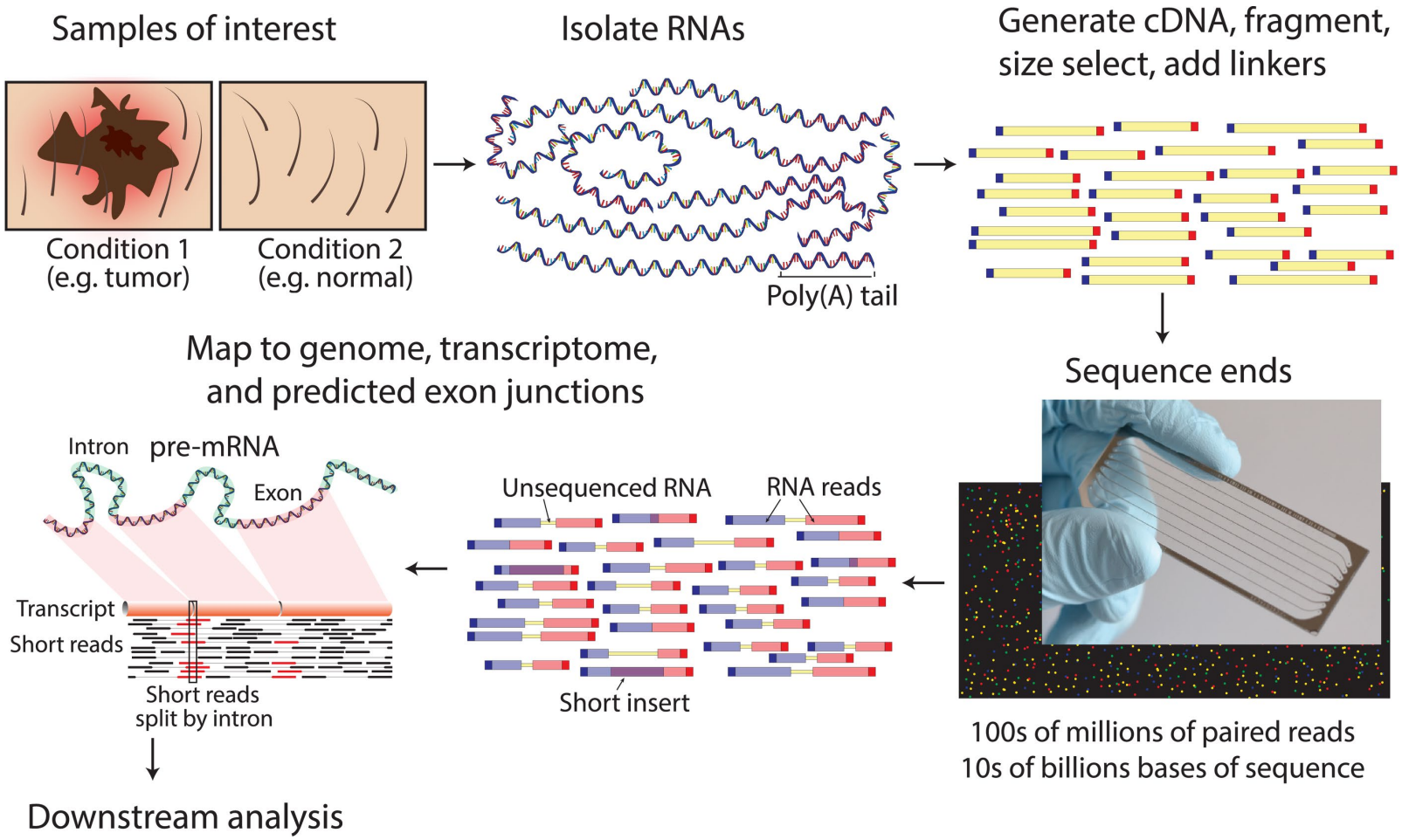


Red=up
Green=down
Yellow=same

https://bitesizebio.com/7206/introduction-to-dna-microarrays/

The reigning champion

RNA-Seq



By Malachi Griffith, Jason R. Walker, Nicholas C. Spies, Benjamin J. Ainscough, Obi L. Griffith - http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=53055894

Experimental design

Considerations

- Number of samples vs depth of sequence
- Sequence depth
  - Complexity of sample
  - Total expected length of the transcriptome
    - Number of genes
    - Length of genes
  - Expected dynamic range of expression
- Replication
  - 5 or more biological replicates is preferable
    - Insures against technical failures
  - Technical variation can be counteracted without additional expense
    - Pool parallel preparations before sequencing
    - Biological replication
- Reference data set
- Verification
  - What other techniques can you use to verify results?

Preparing for RNA-Seq

Sample prep

- Test RNA isolation technique on non-precious samples
  - Yields can be low
- Stabilization of samples
  - Snap-freeze samples
  - RNA*later*®
- Contamination
  - Are there steps you can take to remove contaminants prior to isolation?
- Poly-A enrichment (for eukaryotes)
- rRNA depletion
  - Can comprise up to 98% of RNA content

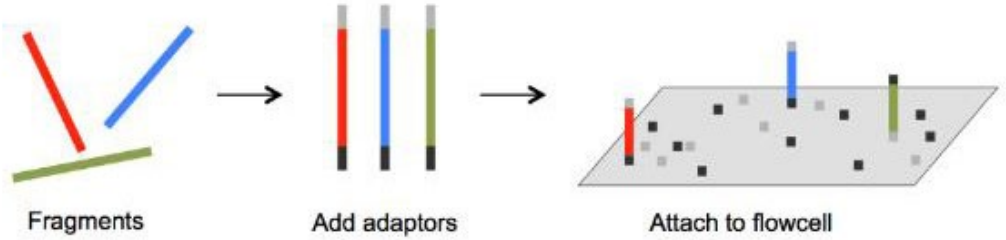
Preparing for RNA-Seq

Library prep

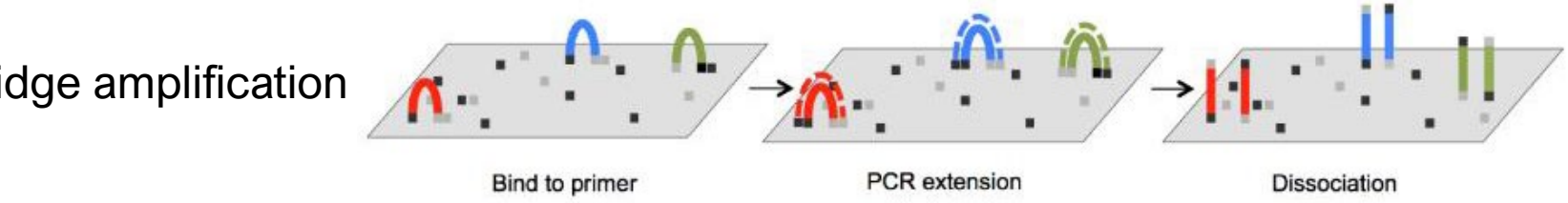
- Method will depend on choice of sequencing technology
- cDNA synthesis
  - DNA more stable
  - Allows amplification
- Fragmentation and size selection
  - Produces uniform range of fragments suitable for sequencing
- Ligation of primers/adapters
  - Sequencing primer site
  - Barcode for multiplexing
- Amplification
  - Enrich for proper construction
  - Compensate for low input

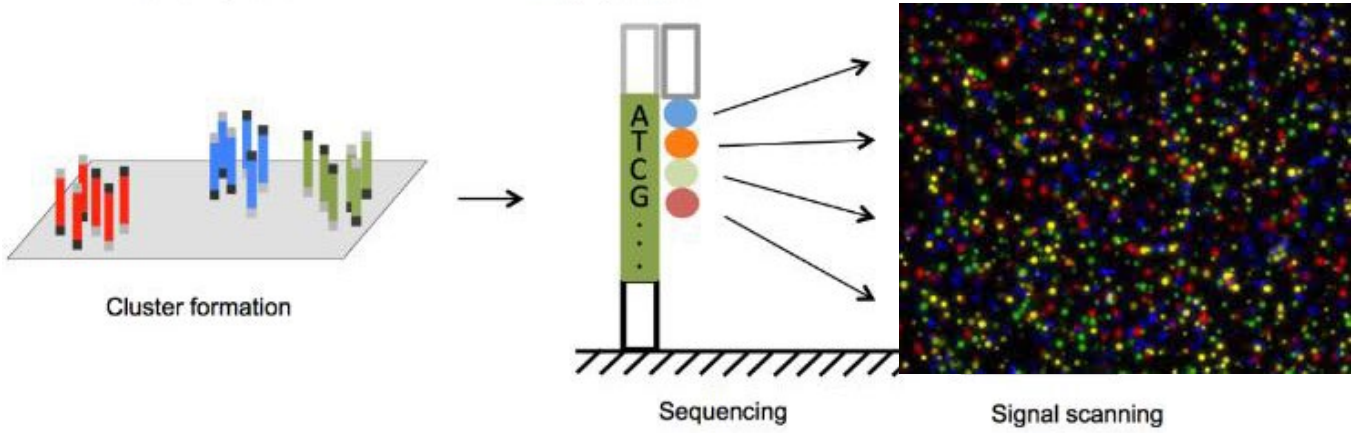
Performing RNA-Seq

Sequencing



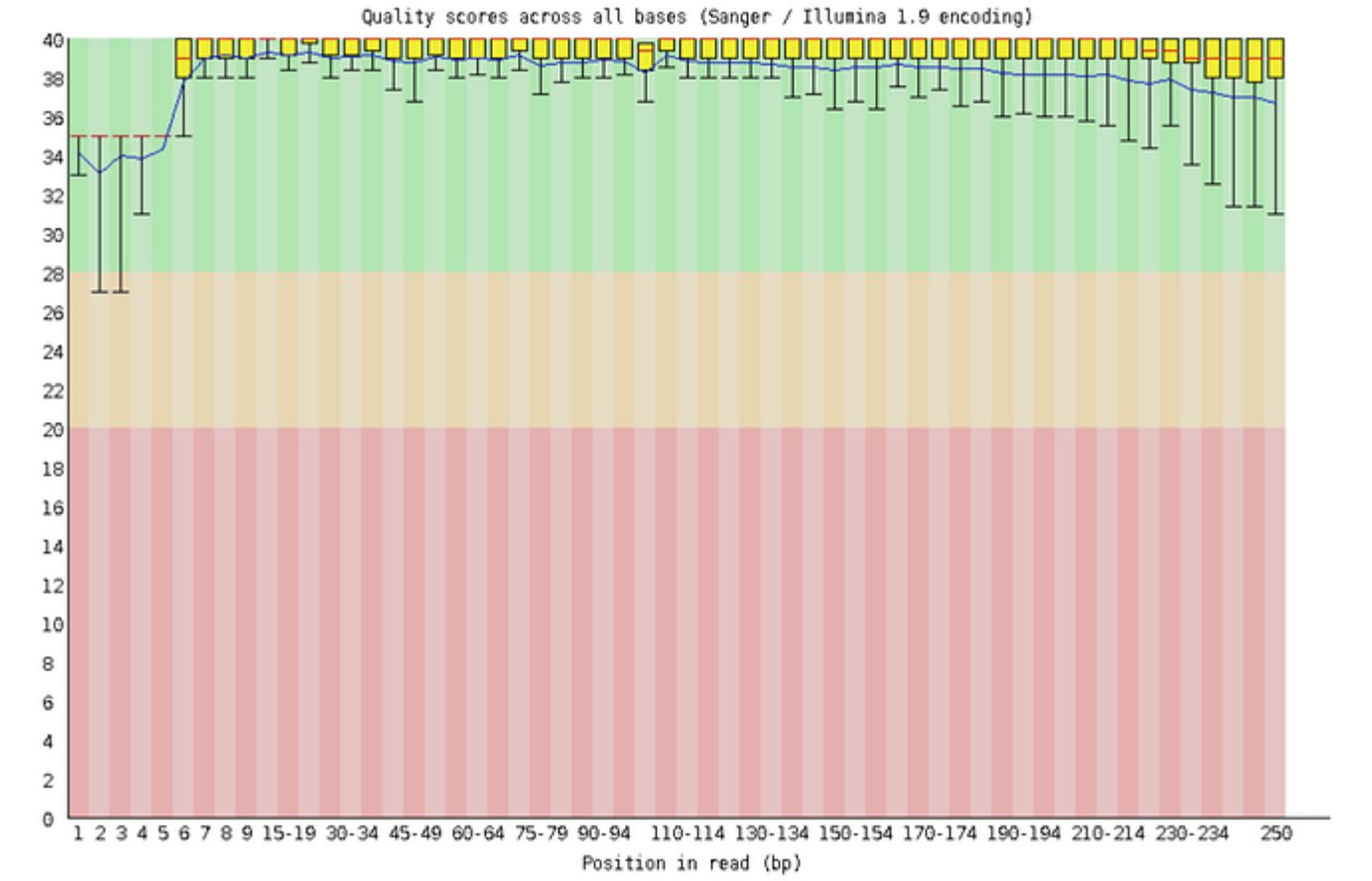
Bridge amplification





Sequence data assessment

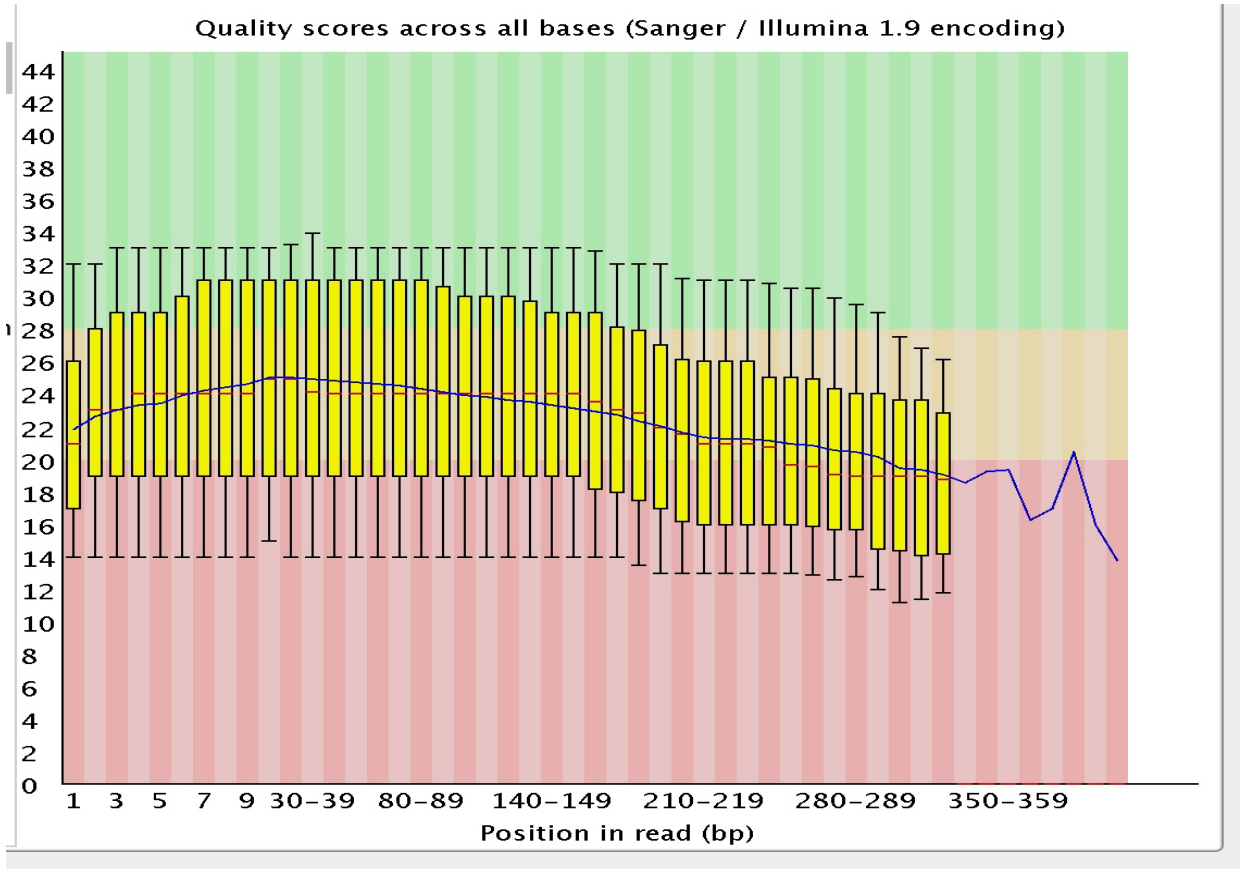
Quality assessment



Sequence quality scores: higher scores are better
Image from FastQC

Sequence data assessment

Quality assessment



Sequence quality scores: higher scores are better

Sequence data processing

Grooming

- Quality trimming based off quality scores
  - Eliminate short remainders
- Adapter trimming
- Low-complexity screen
- Screen for ‘contaminants’
  - rRNA
  - Host RNA

Dataset analysis

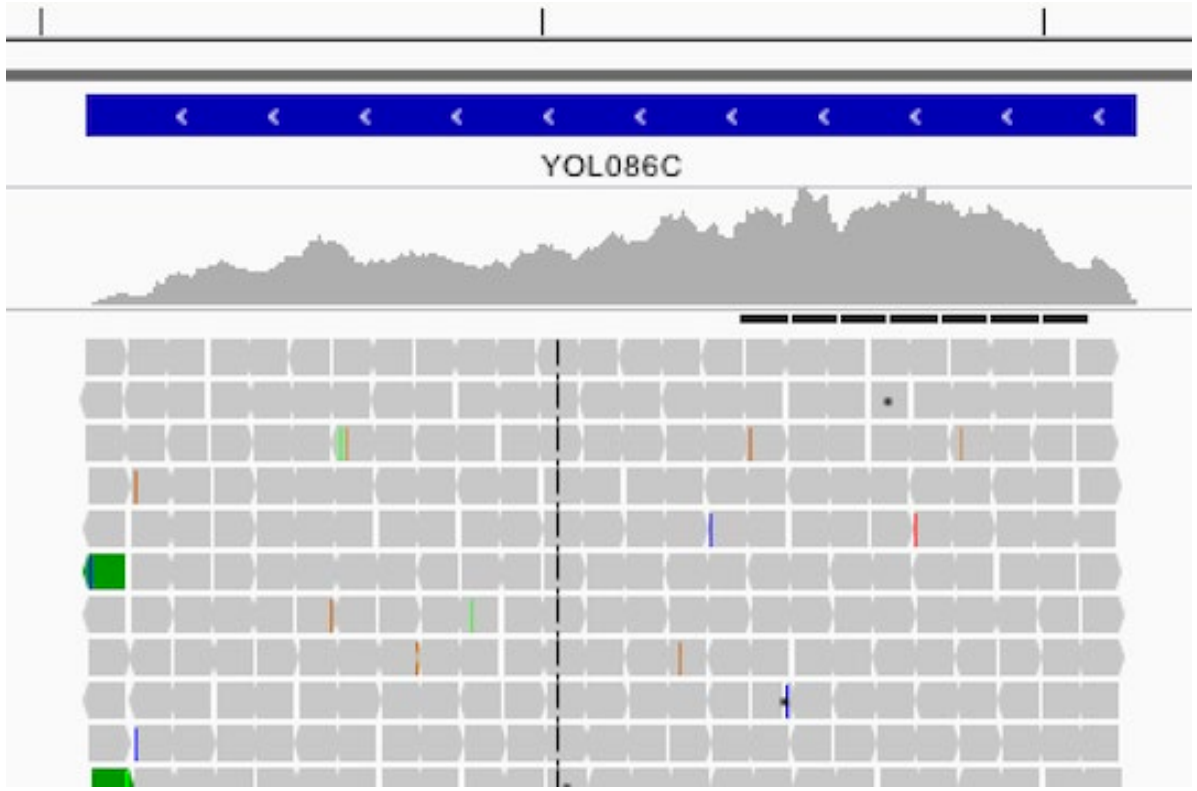
Alignment to reference

- Best if reference is derived from the same sample
  - Metatranscriptomes benefit from deeply sequenced, assembled metagenome
- Reference will typically be the genes, not the genome
  - Unless there is concern about splice variants
- Fast aligners that leverage exact matches
  - Burroughs-Wheeler aligner (BWA)
  - Bowtie2
- You can do reference-free analysis, but it is more complicated and limited.
  - Assemble transcripts
  - Align reads against transcript contigs

Dataset analysis

Count Quantitation

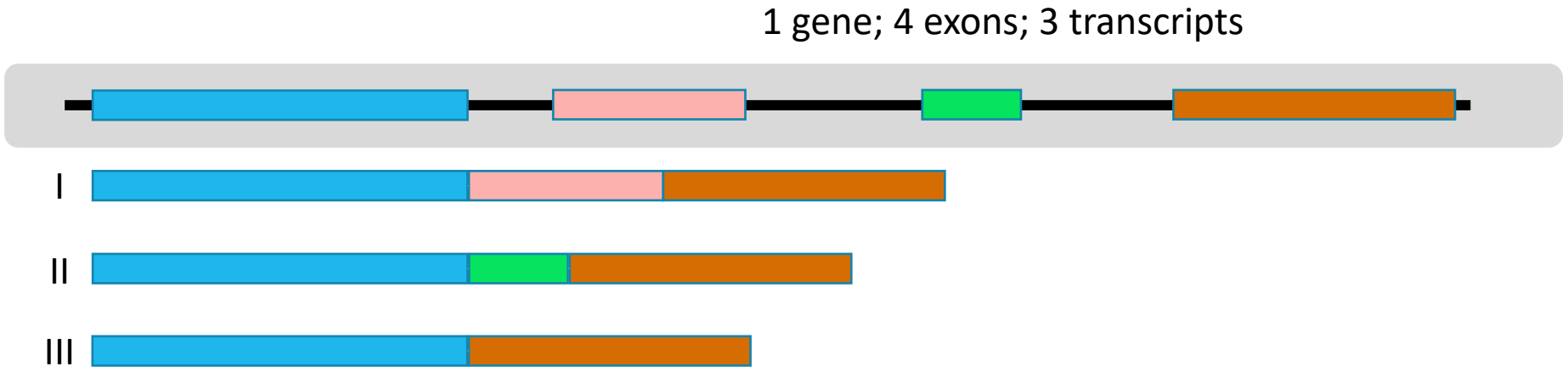
- Read coverage



Dataset analysis

Eukaryotic gene structures

- Quantitate by
  - Gene
  - Exon
  - Transcript



Dataset analysis

Normalization

- Normalize by **dataset size**
  - Per million reads analyzed
- Normalize by **target length**
  - Per kilobase of gene length
- **RPKM** – reads per kb per million
  - PM = total reads aligned/1M
  - RPM = reads mapped/PM
  - RPKM = RPM/gene length in kb
- **FPKM** – fragments per kb per million
  - Accounts for paired-end sequencing
- **TPM** – Transcripts per million
  - RPK = (reads mapped/gene length in kb)
  - PM = RPK_{Tot}/1M
  - TPM = RPK/PM

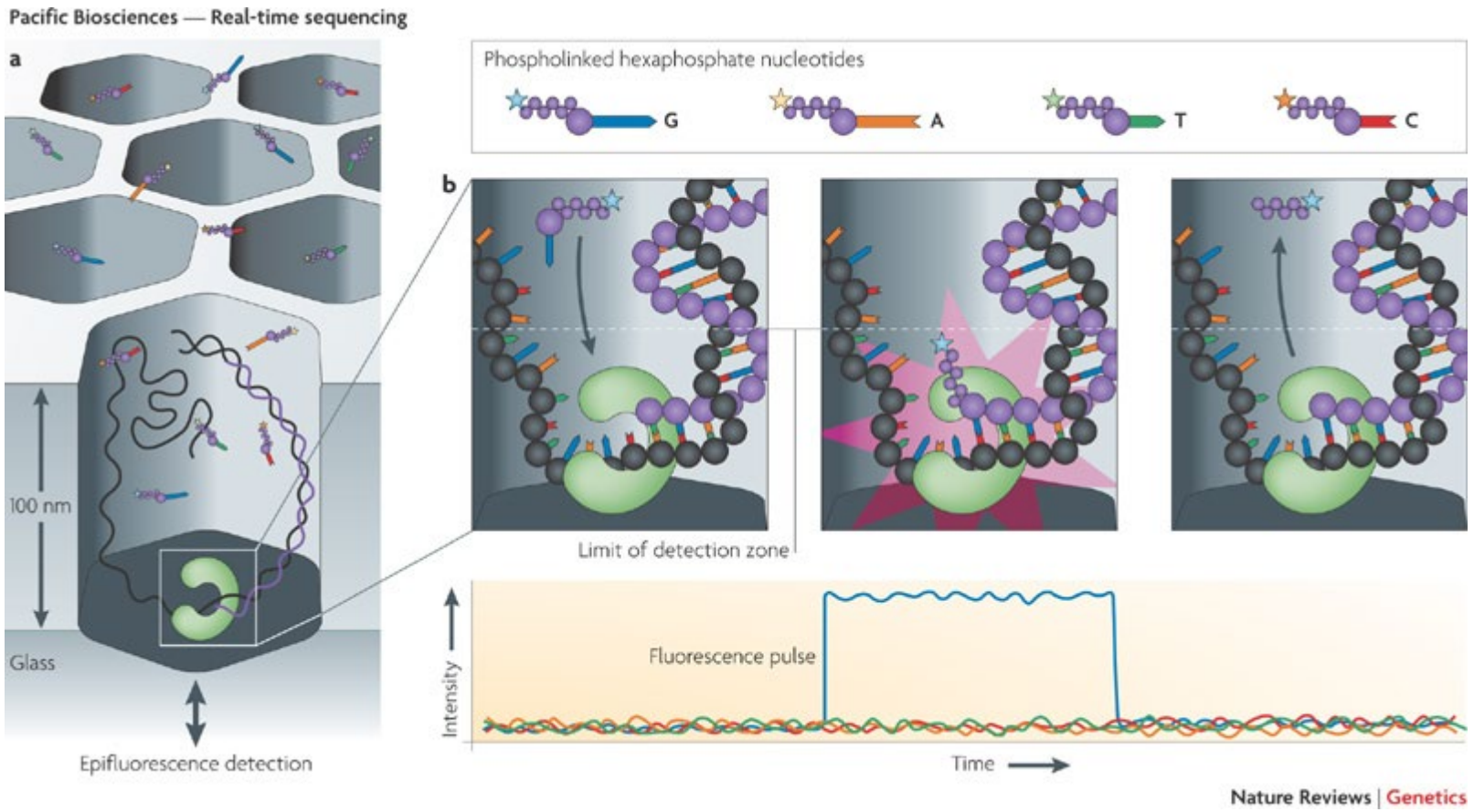
Dataset analysis

Counts table

Gene	Cond1Rep1	Cond1Rep2	Cond1Rep3	Cond2Rep1	Cond2Rep2	Cond2Rep3	Cond3Rep1	Cond3Rep2	Cond3Rep3
Gene0001	6286	8478	2579	8548	2746	6401	890	8724	2784
Gene0002	2425	7199	8735	5153	2287	9552	8776	5691	9166
Gene0003	5027	1561	9483	5721	999	1304	3927	9575	1711
Gene0004	4802	7969	2070	3603	3320	654	4970	6732	3824
Gene0005	2222	2031	7334	1779	7490	5787	5484	3168	8653
Gene0006	5744	8043	3697	6928	8897	7642	1574	4900	1742
Gene0007	9849	3258	5418	1933	2289	8225	6464	1978	7949
Gene0008	8584	6564	7867	4645	4040	6269	9850	7820	6141
Gene0009	1223	6891	5187	3159	8320	4461	5119	5622	3070
Gene0010	3177	8360	9914	682	5093	7976	6144	8234	1781
Gene0011	3291	4337	8829	128	1405	2576	601	9814	1585
Gene0012	7427	2268	4973	2128	430	5199	4994	3998	3845
Gene0013	3801	333	6462	1362	562	6715	630	928	7195
Gene0014	3318	3755	1595	5437	3642	9762	9202	3177	1967
Gene0015	8256	962	2949	9523	6939	8729	6705	4026	231
Gene0016	5674	476	312	4731	849	6150	9569	5112	1362
Gene0017	2717	3538	322	9158	9325	9270	8735	6213	6352
Gene0018	5793	5111	346	1232	4936	4320	3868	3537	9489
Gene0019	4827	6522	5993	2353	3510	7687	936	2227	5498
Gene0020	9283	5366	2738	9583	8087	3724	3118	3029	8892
Gene0021	7857	4550	8474	1900	4997	3103	5586	7392	7517
Gene0022	7079	1944	5091	8256	239	5957	7113	1800	6481
Gene0023	3880	786	7076	1994	7298	7056	1815	1296	2288
Gene0024	5104	9561	855	7164	8955	5429	5331	6118	152
Gene0025	217	8693	131	9236	12	2205	5998	6902	7103

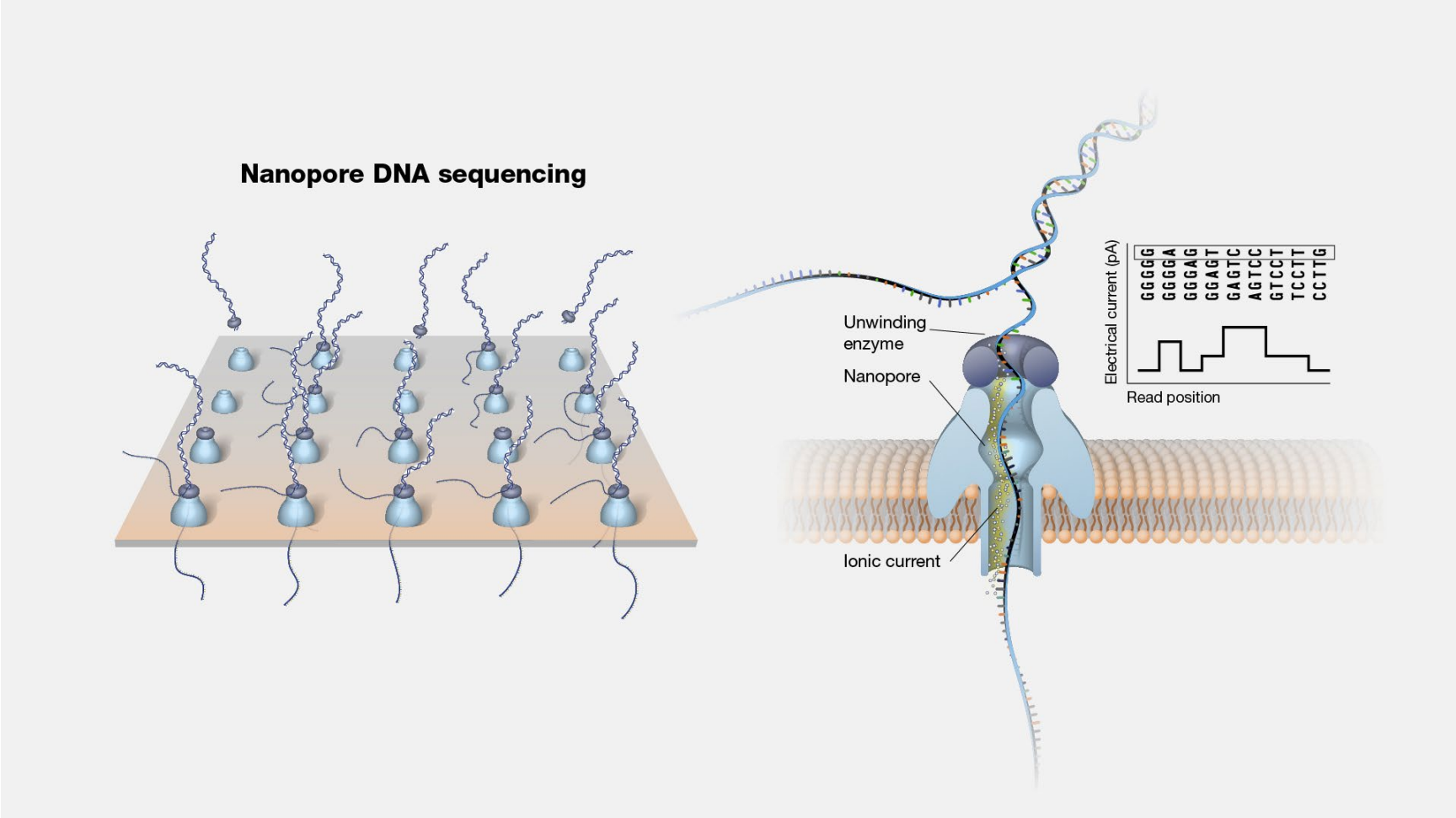
PacBio sequencing

Nanoscale reaction chambers



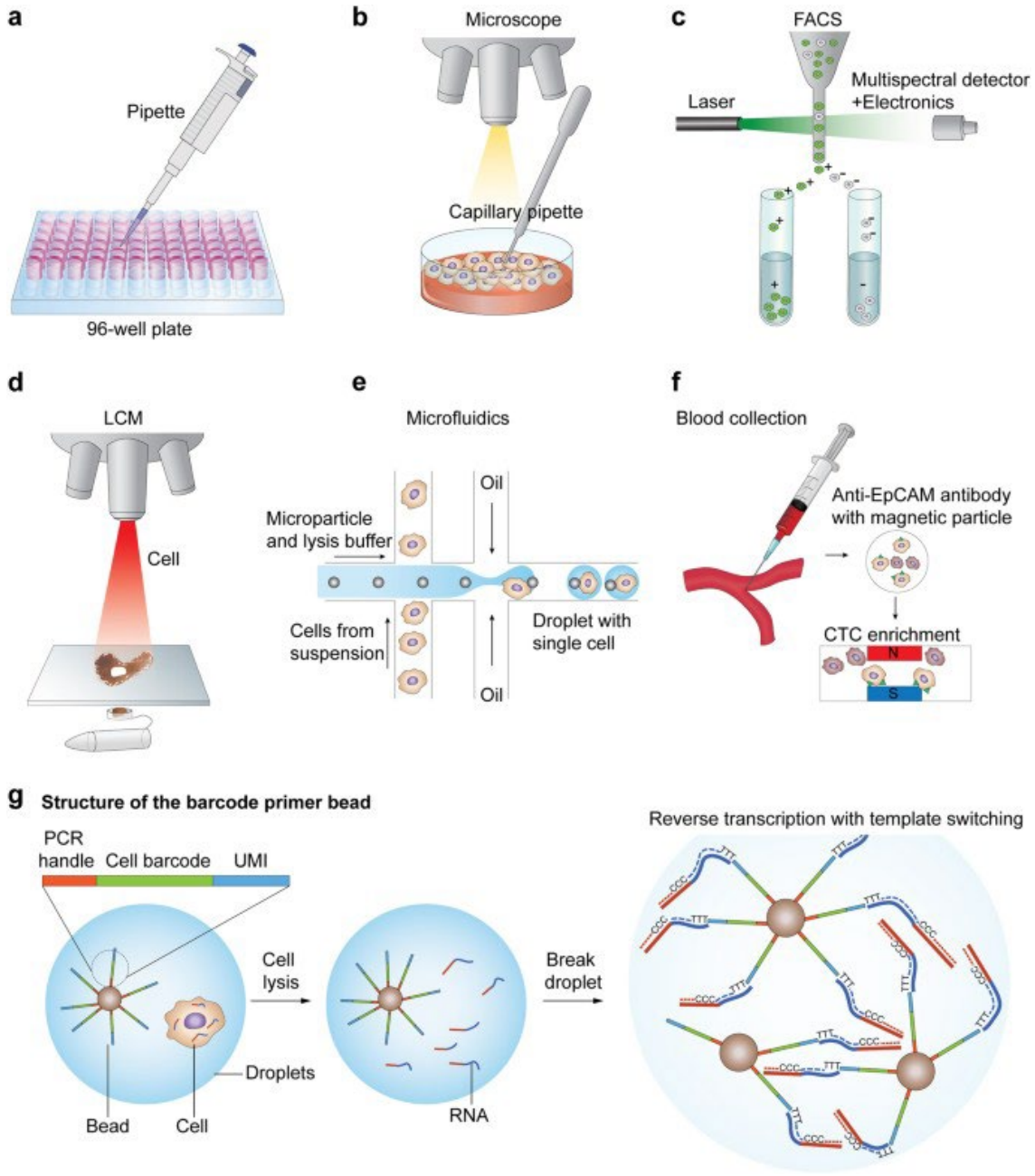
Nanopore sequencing

Sequence determined by measuring interruption of electrical current



https://www.genome.gov/genetics-glossary/Nanopore-DNA-Sequencing

Single-cell transcriptomics



https://www.nature.com/articles/s12276-018-0071-8

Questions?

8:30-8:35 a.m.	Introduction	Lisa Bramer
8:35-9:25	Express Yourself!: Transcriptomics from sample to counts	Bill Nelson
9:25-10:05	**Pre-processing of Transcriptomics Data**	**Rachel Richardson**
10:05-10:15	Networking Break	
10:15-11:00	Statistical Models for Transcriptomics Data	Lisa Bramer
11:00-11:45	Post-differential Expression Analyses (or how to interpret that big ole pile of data you've just generated)	Jason McDermott

EMSL

Day 4
Transcriptomics

Pre-processing of Transcriptomic Data

Rachel Richardson
Data Scientist I

Our definition

After
transcriptomic counts have been cross-tabulated for further analysis

Before
differential expression analysis is performed

Is there a vaguer term than preprocessing?

“Next time, don’t start the presentation by asking, ‘Can you tolerate ambiguity?’!”

Our definition

After
transcriptomic counts have been cross-tabulated for further analysis

Before
differential expression analysis is performed

Is there a vaguer term than preprocessing?

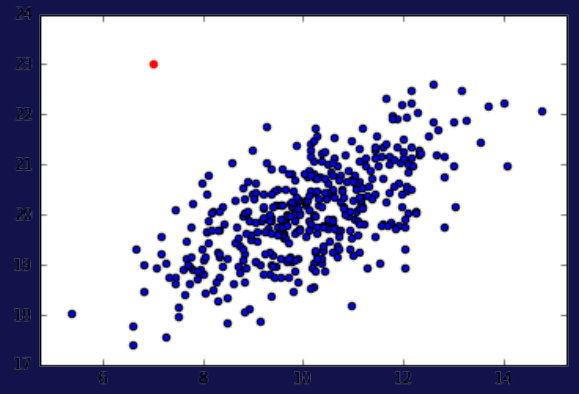
Samples

Genes/Transcripts

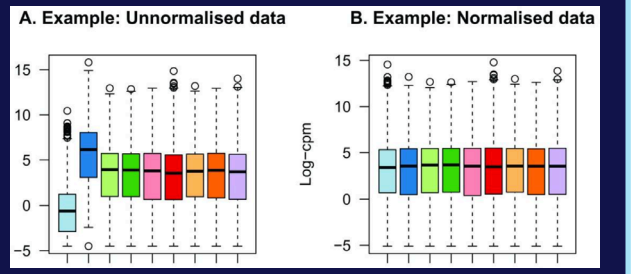
	MCL1.DG	MCL1.DH	MCL1.DI	MCL1.DJ	MCL1.DK	MCL1.DL	MCL1.LA
497097	438	300	65	237	354	287	0
100503874	1	0	1	1	0	4	0
100038431	0	0	0	0	0	0	0
19888	1	1	0	0	0	0	10
20671	106	182	82	105	43	82	16
27395	309	234	337	300	290	270	560
18777	652	515	948	935	928	791	826
100503730	0	1	0	0	0	0	0
21399	1604	1495	1721	1317	1159	1066	1334
58175	4	2	14	4	2	2	170
108664	769	752	1062	987	995	903	1381

Counts

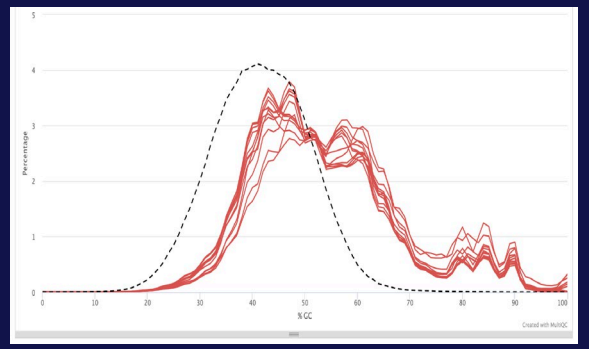
Outlier detection



Transformation/Normalization



General visualizations



Why do we need to do preprocessing?

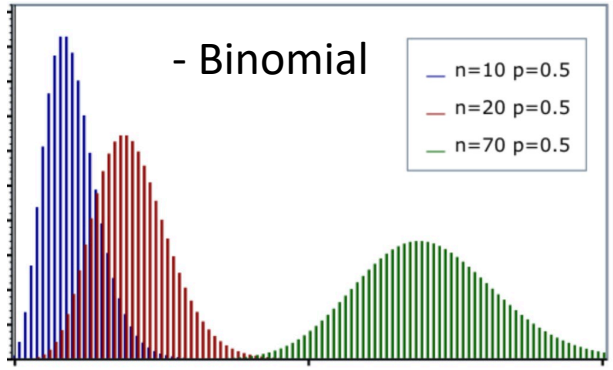
- Quality control
  - Not every sample or transcript turns out the way we hope!
  - Understanding why a sample or transcript is showing an unusual trend is important
- Removing non-biological or experimentally irrelevant variation
  - Lots of variation between individuals of a population
  - Variations in preparation introduce unwanted variation
- Better understanding of the data
  - Check assumptions before running downstream analyses
  - Understand why results might look a certain way

Big takeaway
We need to consider how RNA data is collected to apply appropriate preprocessing steps

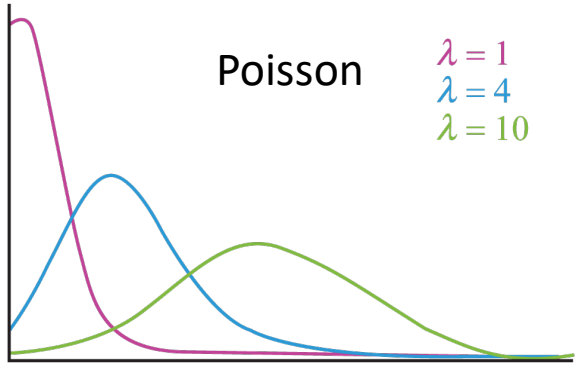
Thinking ahead
What preprocessing steps might be sensitive to zeros?

What is special about transcriptomics?

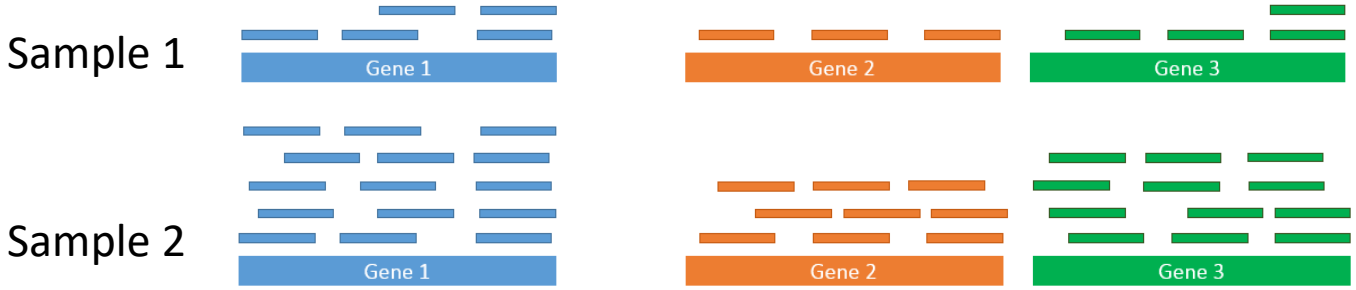
- Count data
  - Zeros matter!
  - Typically modeled by negative binomial or Poisson distributions



- Library size considerations
  - All transcript counts in a sample = library
  - Non-biological variation can change the efficiency of read transcription
  - The total number of transcripts in a library often used as a normalization factor



Are these actually different?



Experimental Highlights

-6 experimental groups
-12 total mice
~26,000 annotated gene transcripts detected

Citation

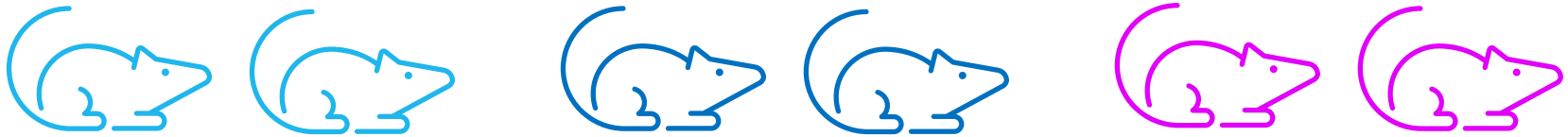
Chen Y, Lun ATL and Smyth GK. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline [version 2; peer review: 5 approved]. *F1000Research* 2016, **5**:1438 (https://doi.org/10.12688/f1000research.8987.2)

What does preprocessing look like with real data?

Basal stem cells (B)

Virgin | Pregnant | Lactating

Committed luminal cells in mammary gland (L)



Virgin | Pregnant | Lactating

*All mice ♀ and genetically identical

**RNA-seq data generated via Illumina sequencing, 100bp single end reads

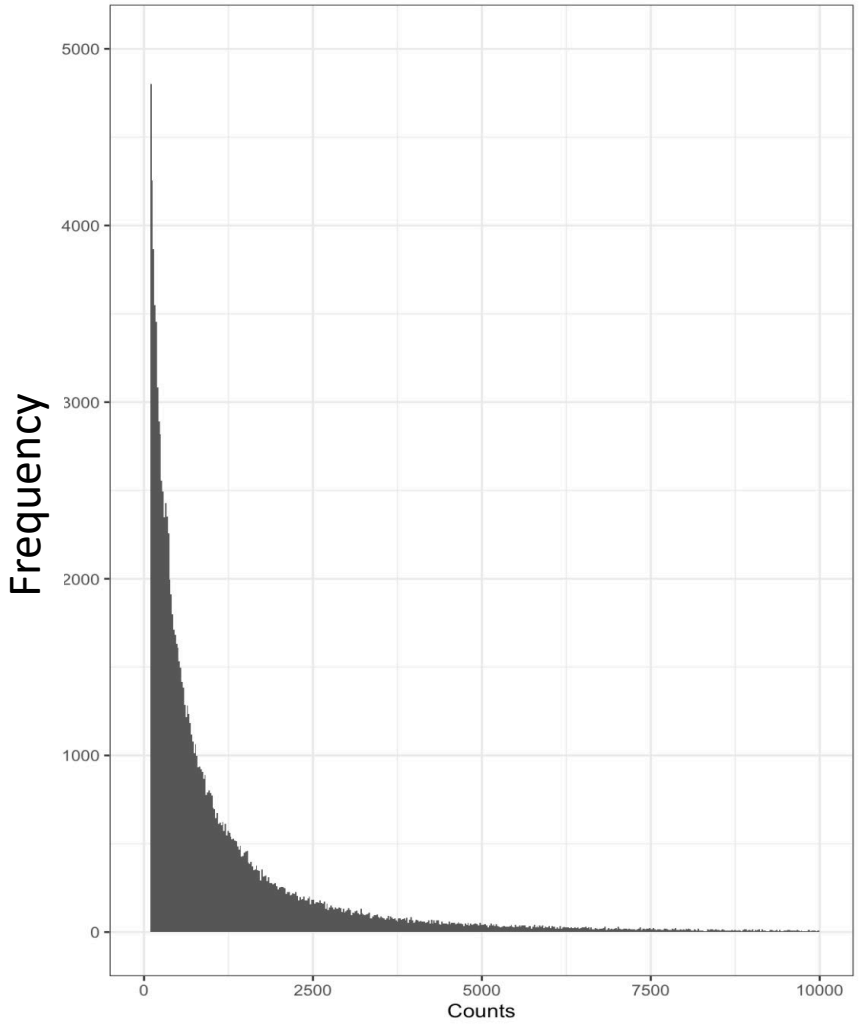
Experimental Highlights

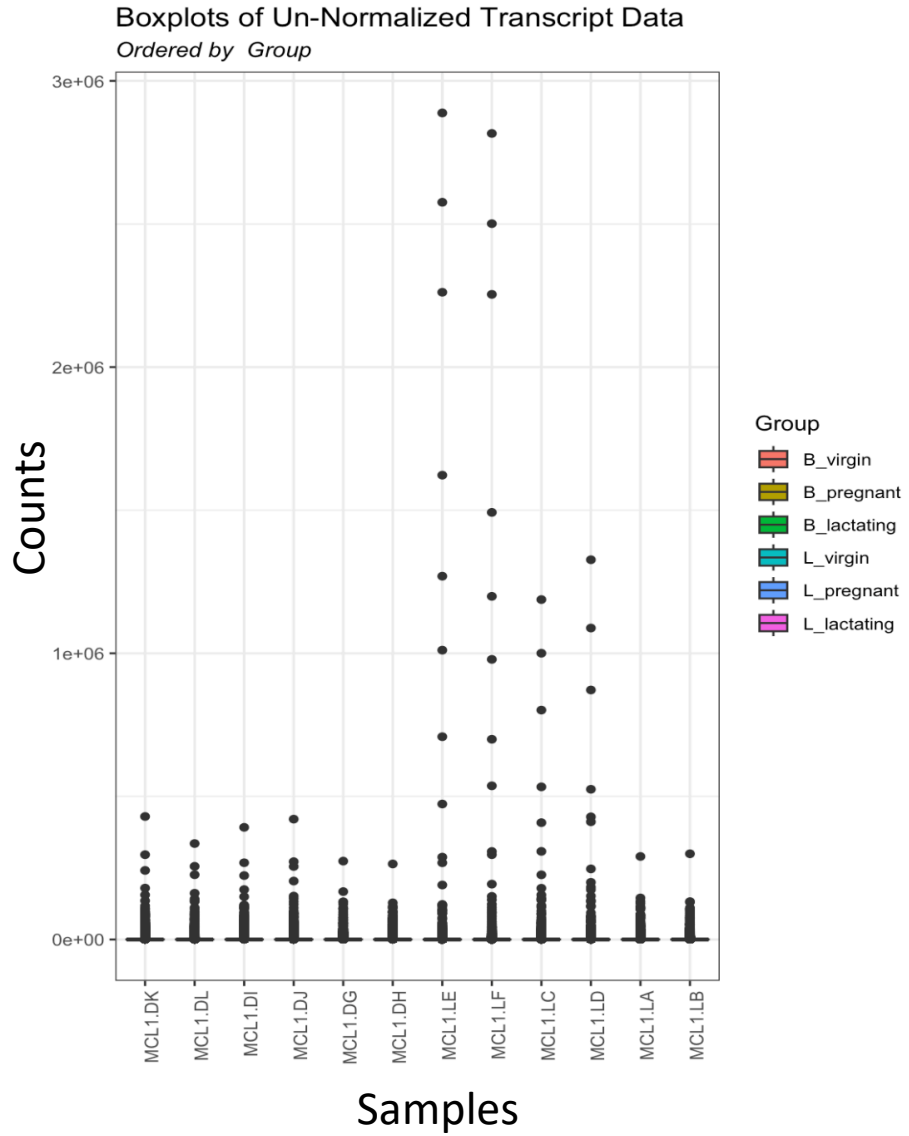
-6 experimental groups
-12 total mice
~26,000 annotated gene transcripts detected

Citation

Chen Y, Lun ATL and Smyth GK. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline [version 2; peer review: 5 approved]. *F1000Research* 2016, **5**:1438 (https://doi.org/10.12688/f1000research.8987.2)

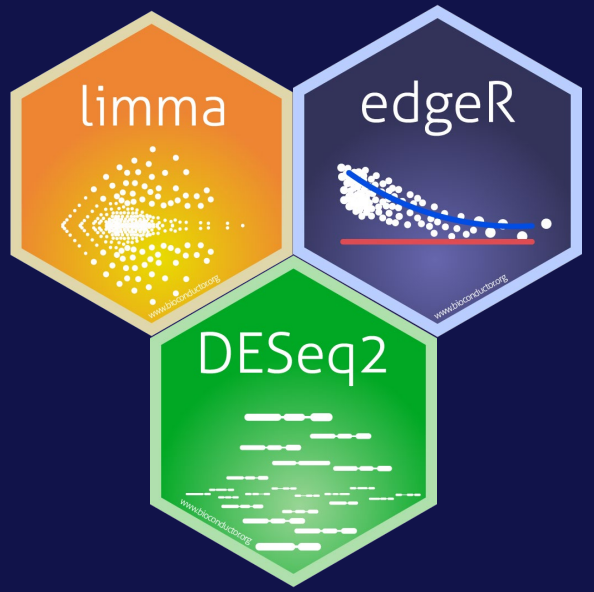
What does preprocessing look like with real data?





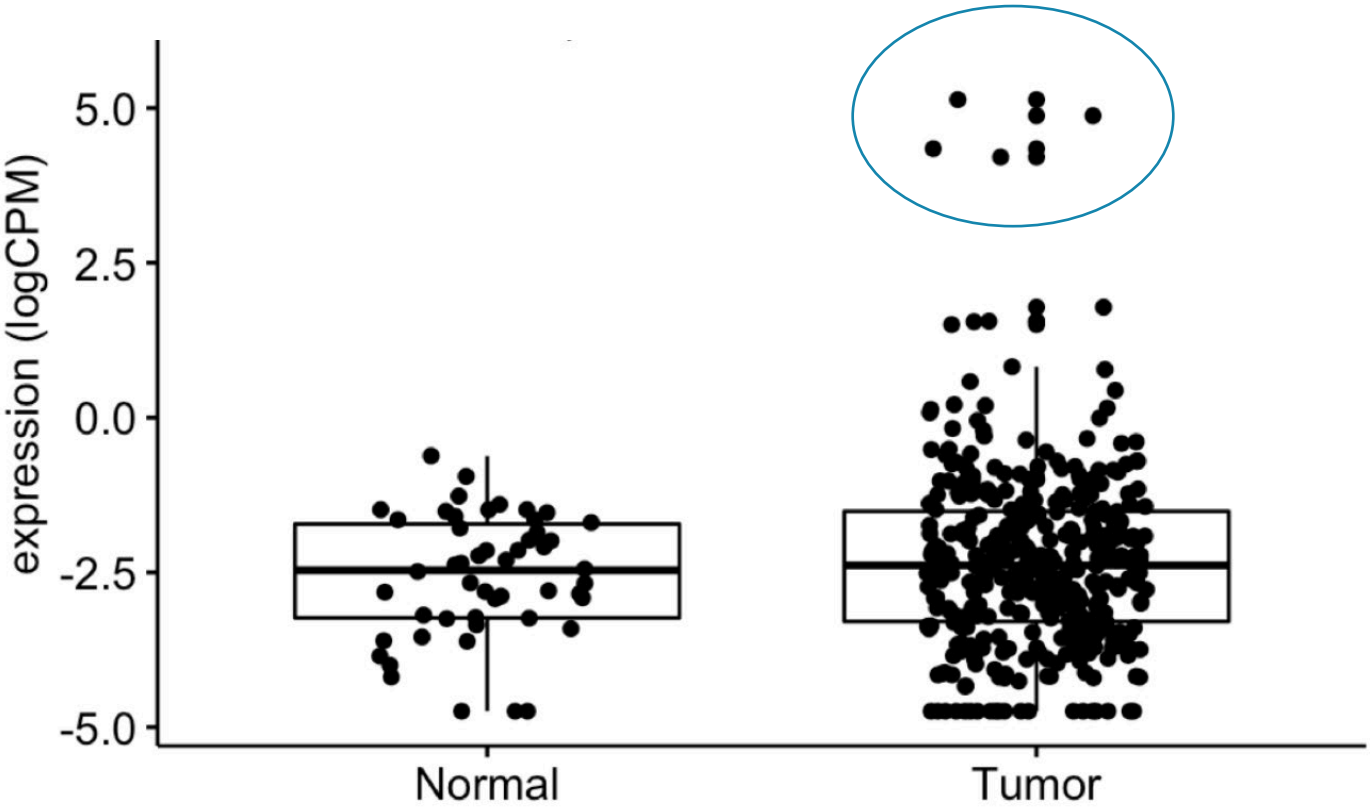
Data exploration requiring scaling (or adjustments)

-Boxplots/barplots
-Principal component analysis
-Mean-variance plots



Can and should we scale transcriptomic data?

- For visualizing and data exploration? Absolutely!
- For use in differential expression? Typically, no



Common scaling

Counts **p**er **m**illion:
CPM

Log **c**ounts **p**er **m**illion:
LCPM

Reads **p**er **k**ilobase of transcript per **m**illion:
RPKM

Fragments **p**er **k**ilobase of transcript per **m**illion:
FPKM

Regularized **log - rlog**
Variance **s**tabilizing **t**ransformation - **vst**

How do we scale transcriptomic data?

Gene = g
sample = i

Counts for a gene = r
Sum of counts = R = **Library size**

edgeR:

$$CPM = \frac{r_{gi}}{R_i} \times 10^6$$

$$LCPM = log_2\left(CPM + \frac{2}{R_i}\right)$$

Limma-voom:

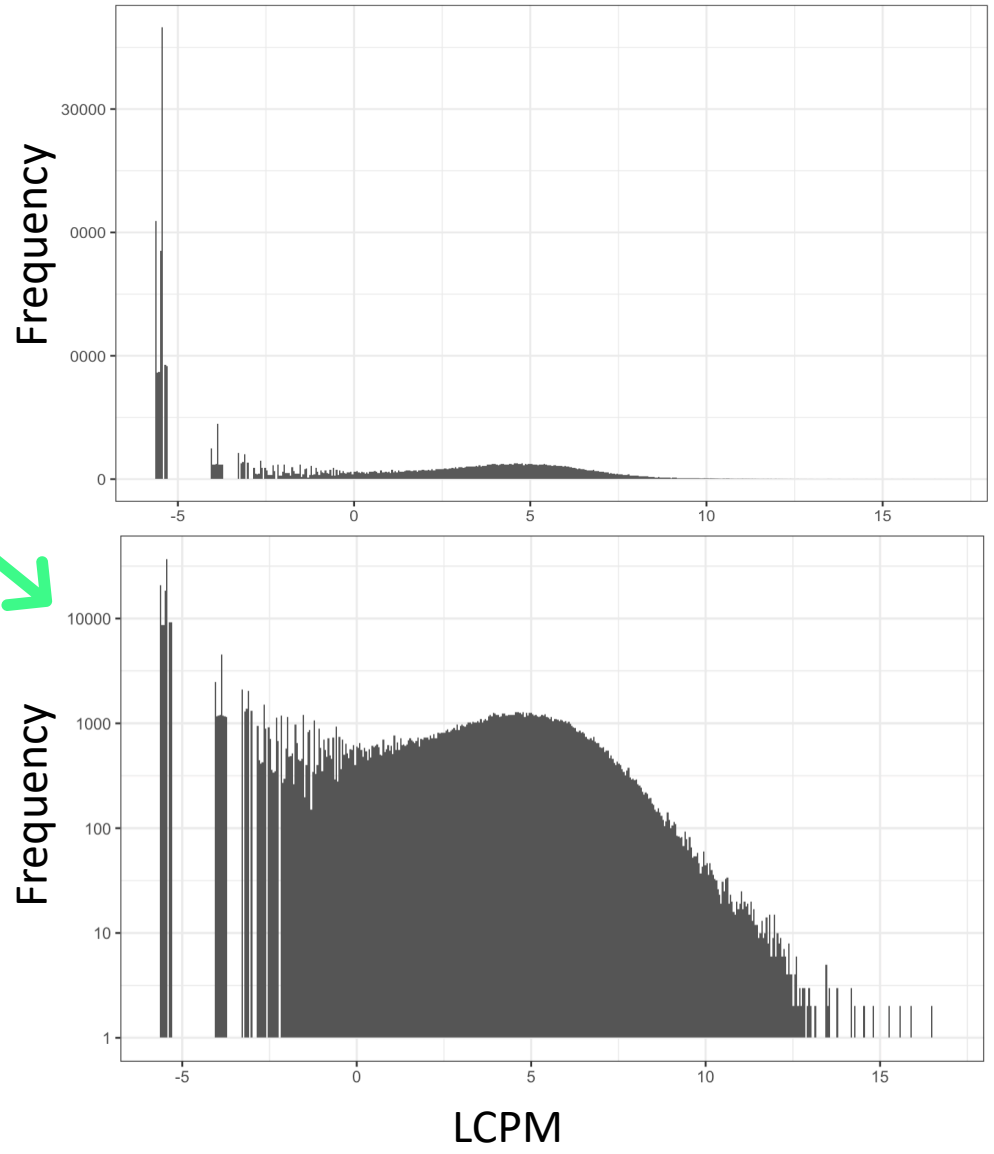
$$CPM = \frac{r_{gi} + 0.5}{R_i + 1} \times 10^6$$

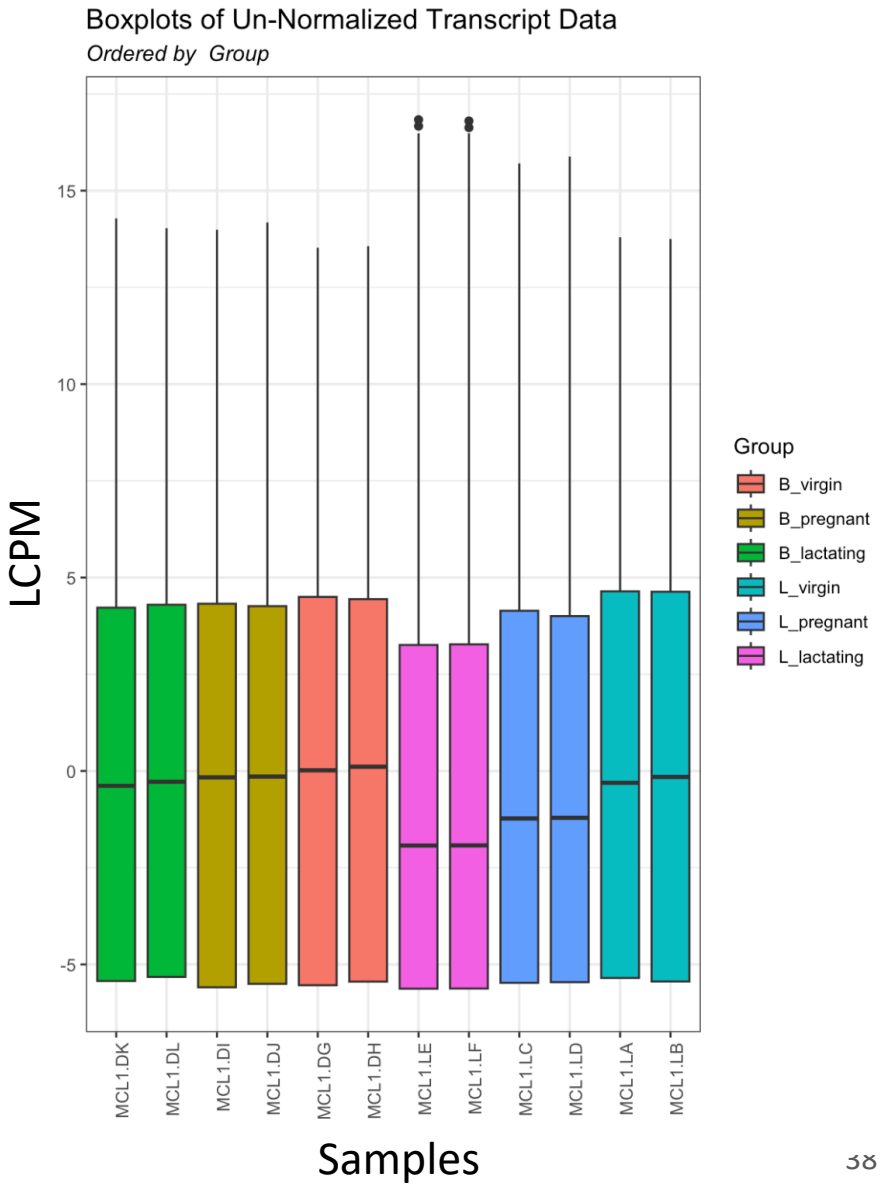
$$LCPM = log_2(CPM)$$

*The 0.5 and 2 are arbitrary! Just needs to be a small number ☺
**The +1 for voom ensures the proportion is less than 1 in all cases

What does the transformed data look like?

Scaling can explore visual representations in new ways



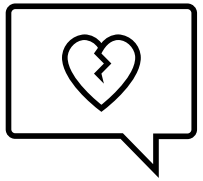


Curated data is not the same as cherry-picked data

What else should we be looking for?

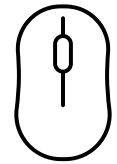
Data exploration can sniff out odd occurrences

- The enzymes denatured or did not cooperate
  - Check the number of reads across the samples



- A contamination occurred
  - Check if there are an unusual number of transcripts observed or not observed in the samples

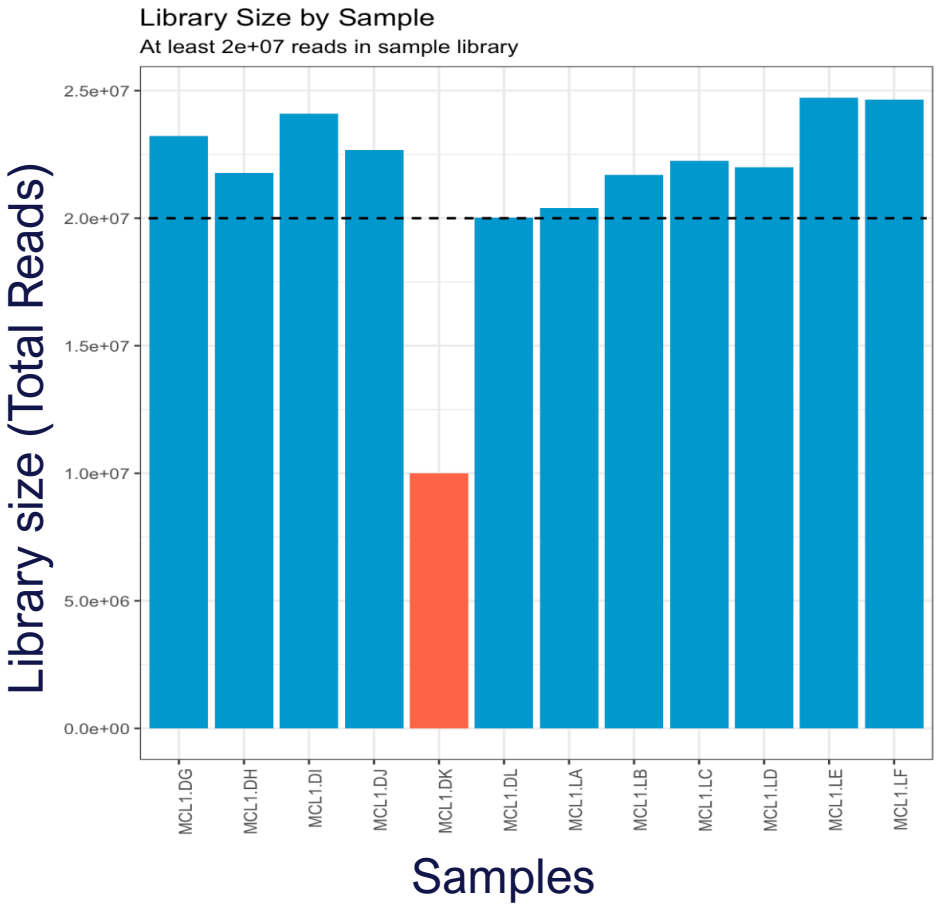
- The labels for two mice might be mixed up
  - Check if the transcript profiles are similar to expected groups



When should I be alarmed?

Grapher Beware:
Pay attention to the scale of number of reads too, not just the similarity of total reads

Note:
The total number of reads can be plotted without transformation

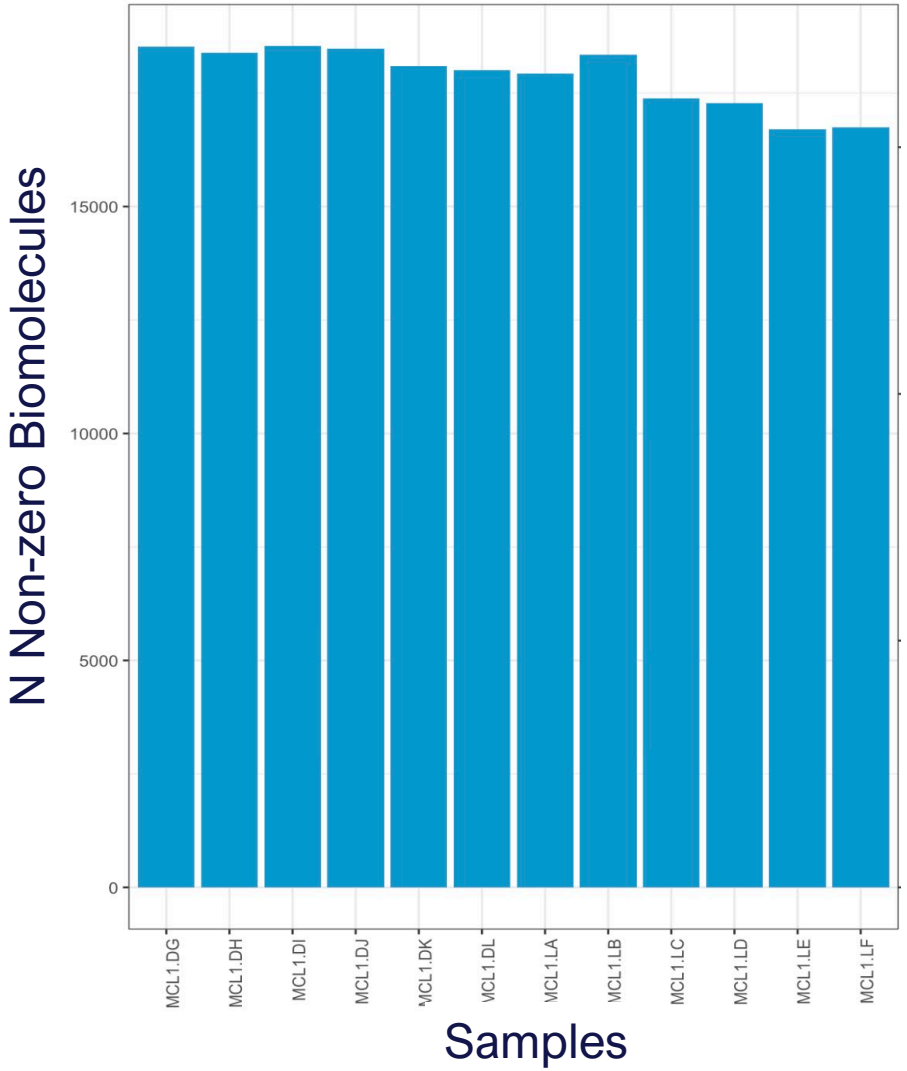


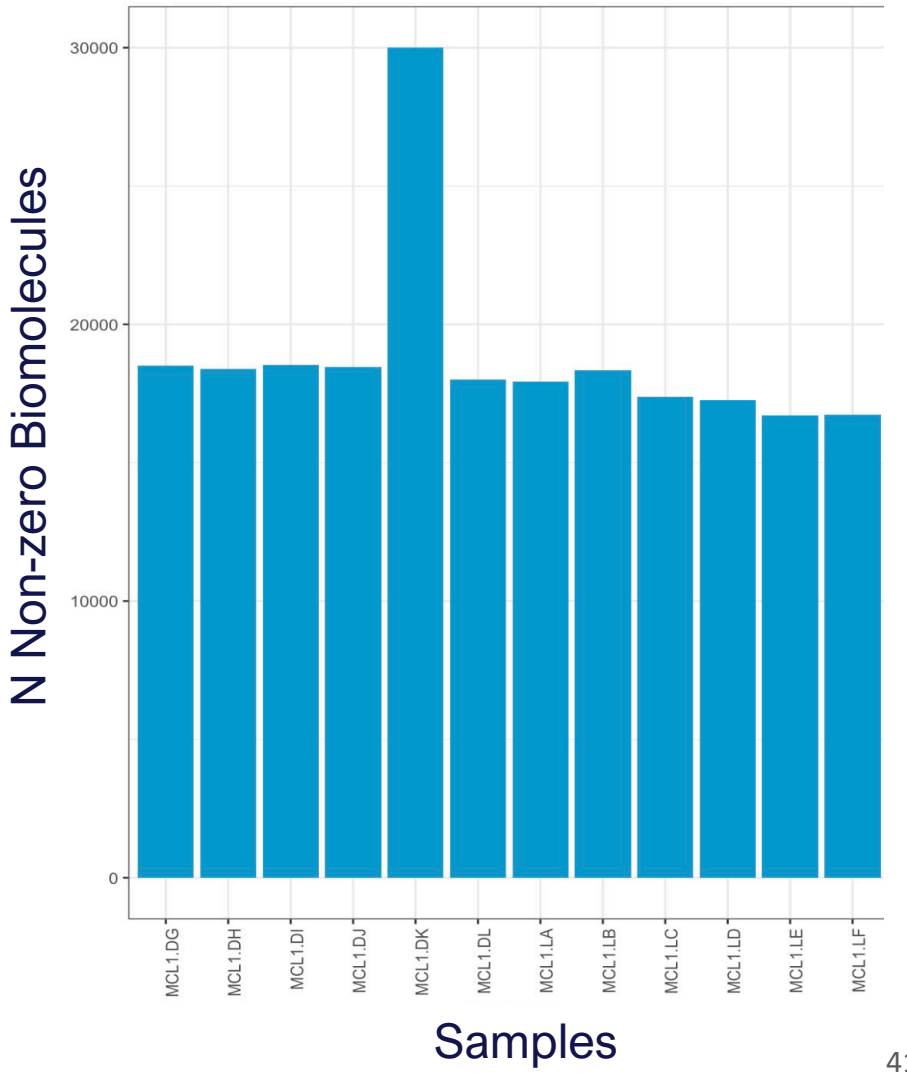
Hot tip: Some sequencing software might flag this earlier, check with your analysts

When should I be alarmed?

Grapher Beware:
Pay attention to the scale of number of reads too, not just the similarity of total reads

Note:
The total number of reads can be plotted without transformation

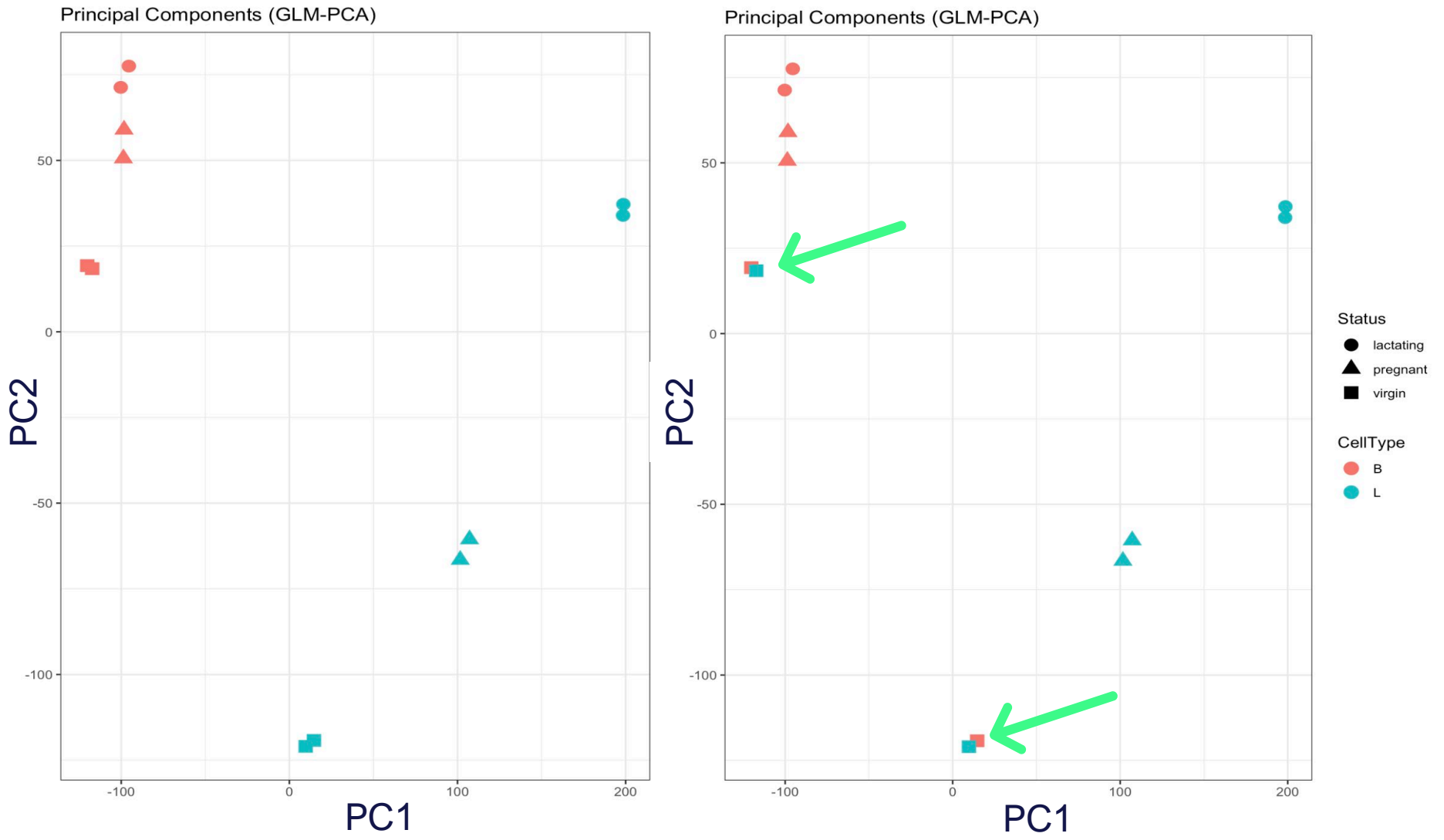




When should I be alarmed?

Principal Component Analysis (PCA)
Typically requires normal data, doesn't work well with raw counts!

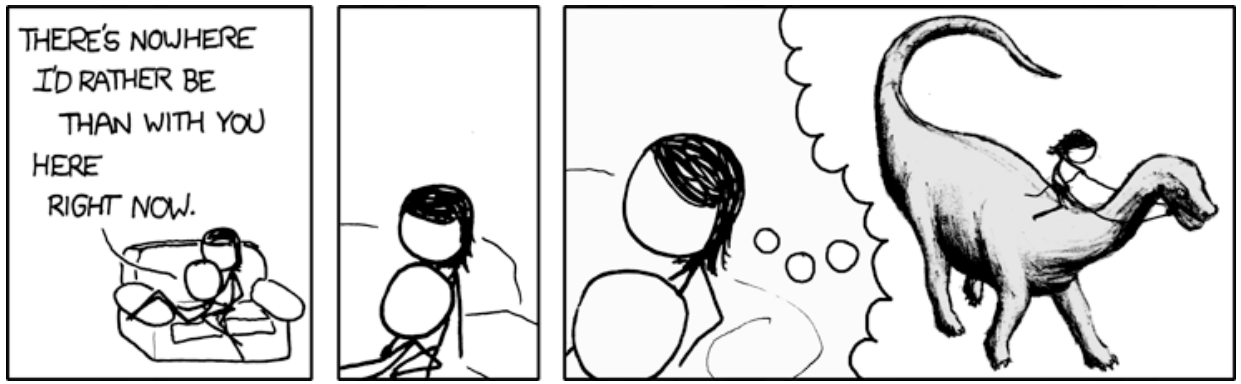
Other PCA methods
-PCAHubert
-GLM-PCA
-Sparse PCA



Now what?

- Preprocessing
  - Transformation
  - Data exploration
  - Filtering

You are **Here**



- Normalization?
- Differential expression?

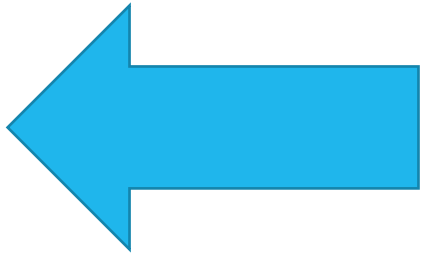
Are we ready to be **Here**?

Know thyself
Statistical analyses almost always come with assumptions about what the data looks like

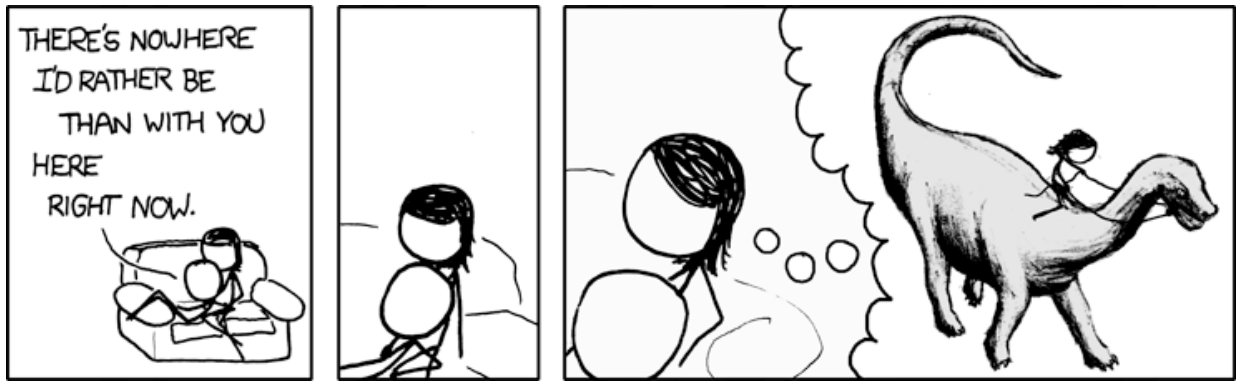
Staple differential expression methods
- DESeq2
- limma-voom
- edgeR

Now what?

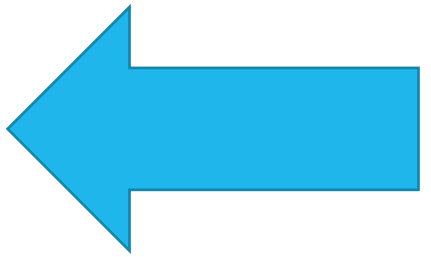
- Preprocessing
  - Transformation
  - Data exploration
  - Filtering



You are **Here**



- ~~Normalization?~~
- Differential expression?

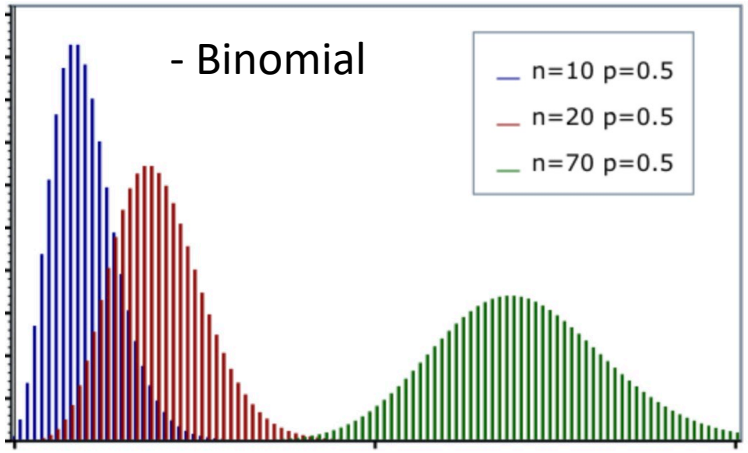


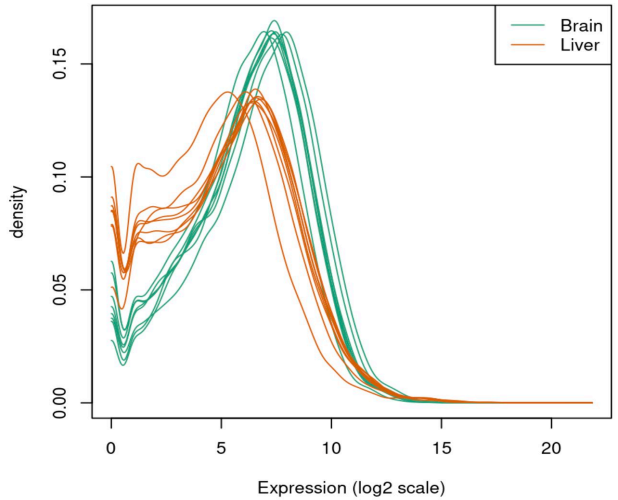
Note
Despite the same assumptions, DESeq2 and edgeR take different normalization approaches

- Transcripts need to be observed in 2+ samples
- Low count levels can mess up ratios and quantiles

What should we know about the statistical tests?

- DESeq2 and edgeR: Counts, negative binomial distribution, assumes that most genes are not DE and normalizes as such
- Limma-voom: Counts, linear weighted model, resulting test statistics are approximately normally distributed



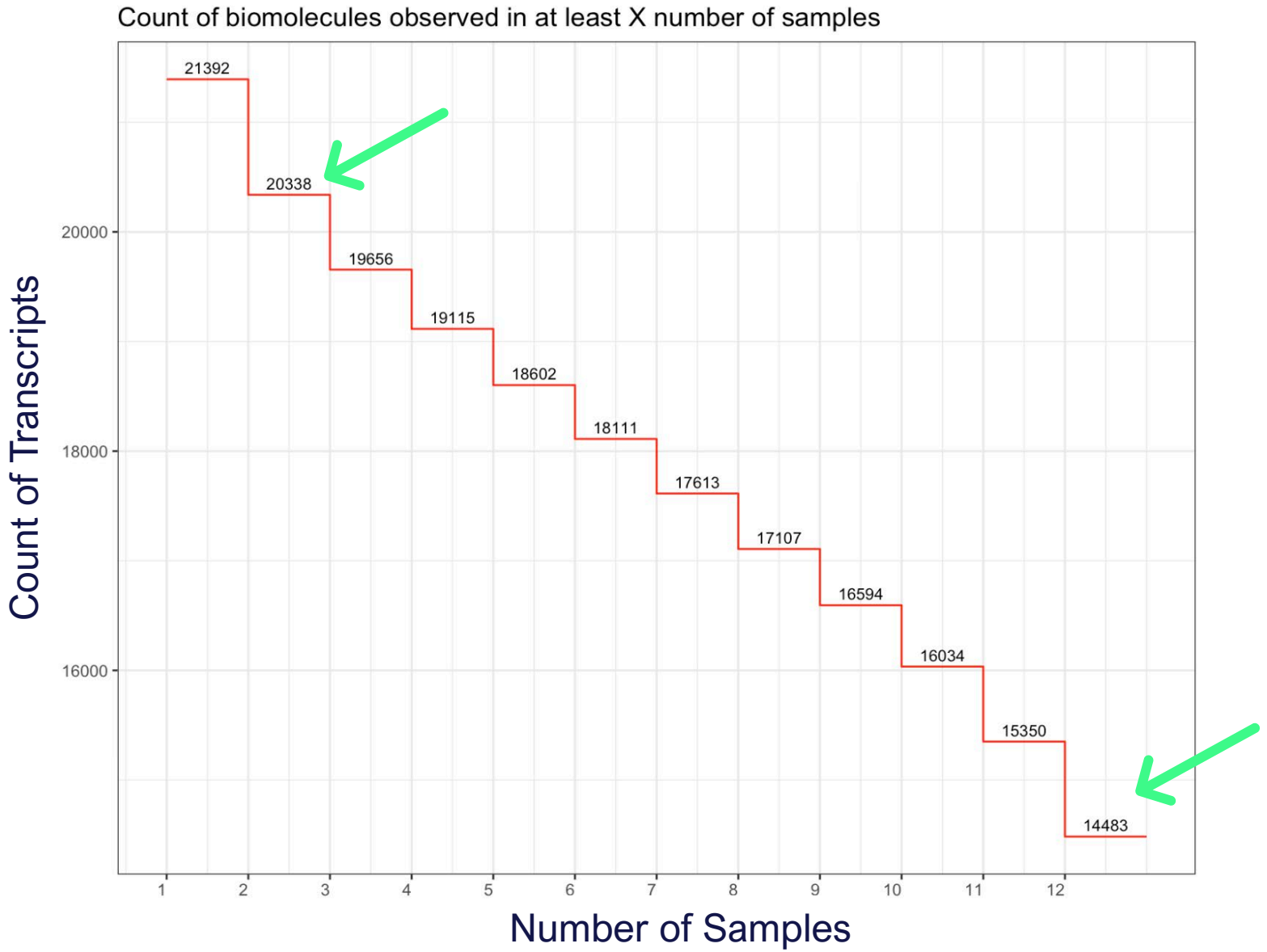


Relies heavily on ratios and distributions

(More on these in our next talk!)

How many transcripts can we actually use?

At least 2 non-zero samples is a good idea for analysis – often more required to assure the biomolecule is represented in all groups



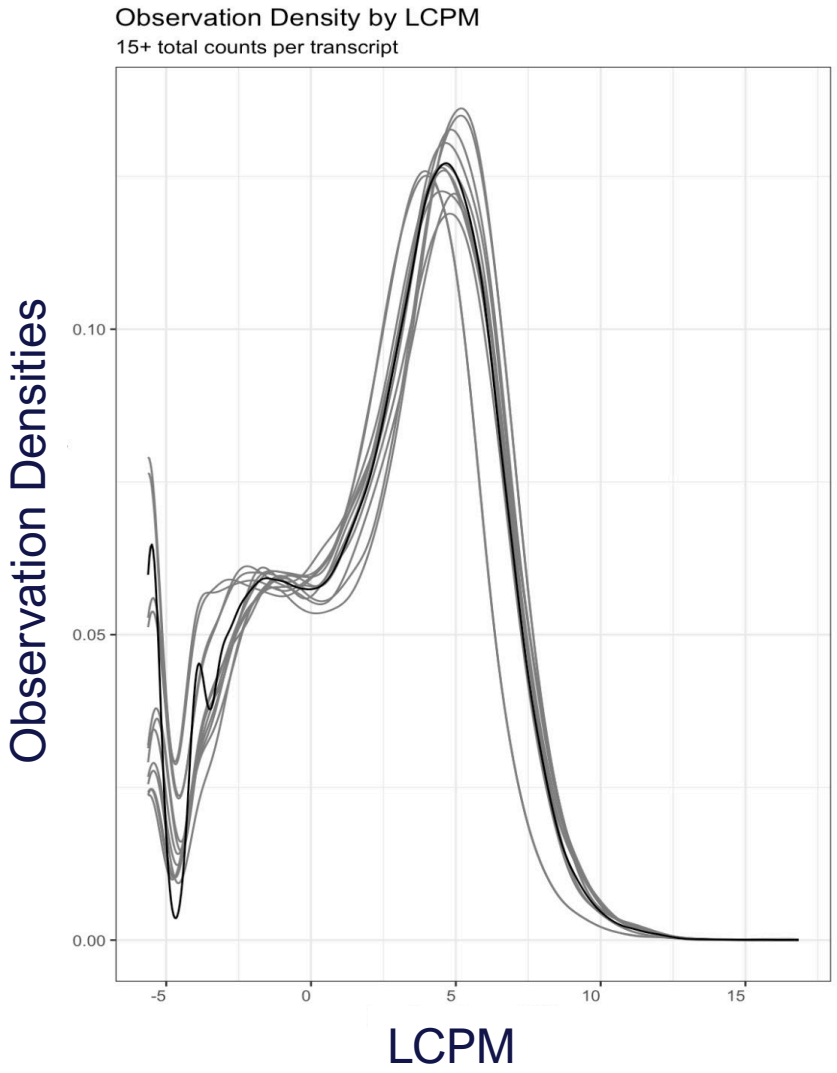
Which transcripts have enough counts?

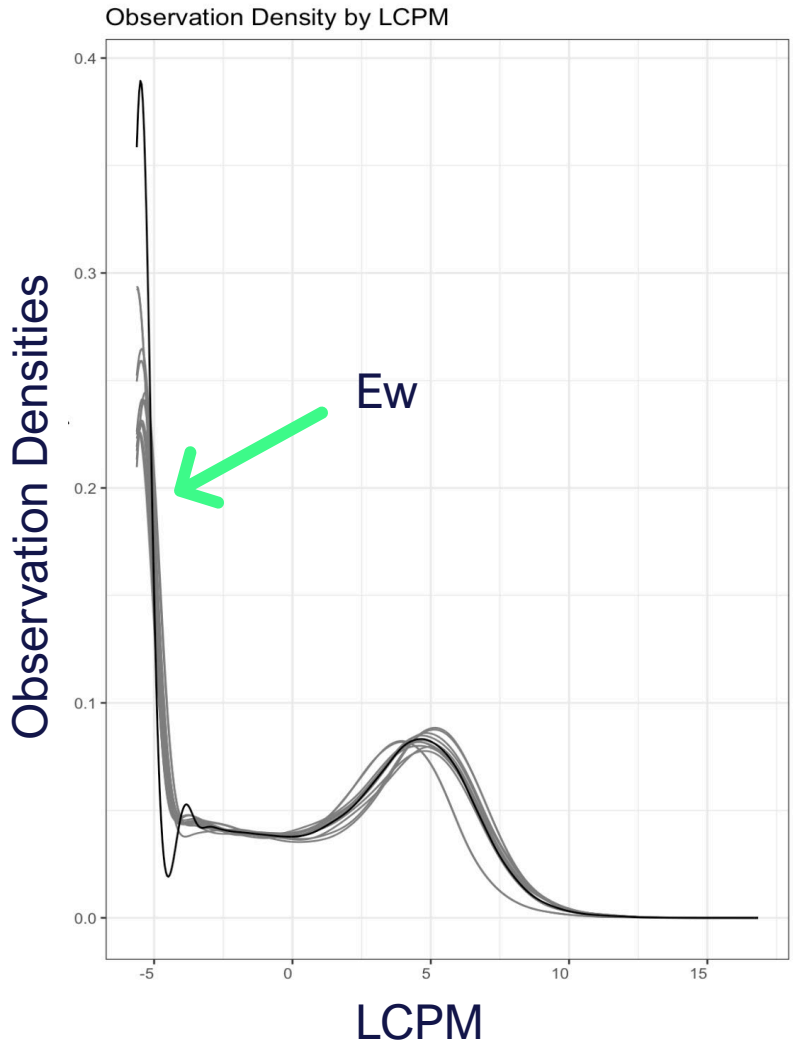
Recommendation from the top

At least **15** counts across all samples

AND/OR

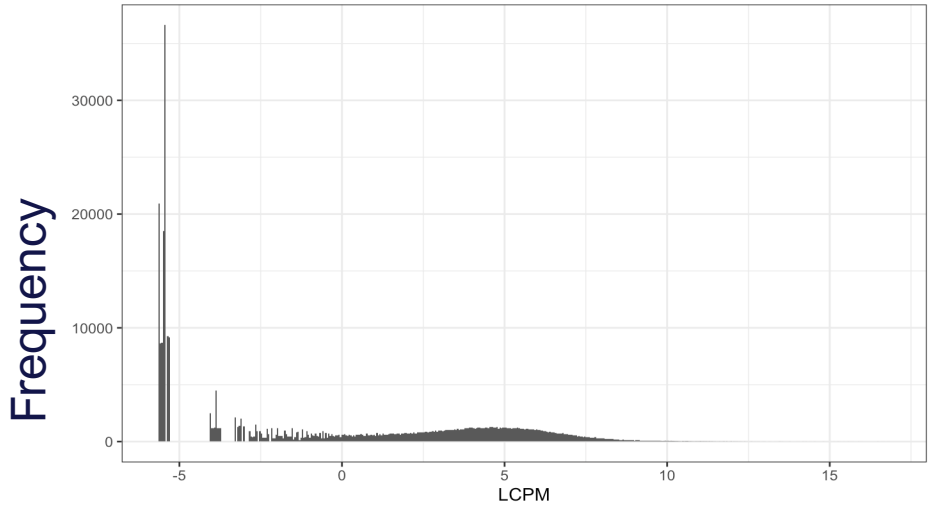
~10 counts per sample with library size taken into account

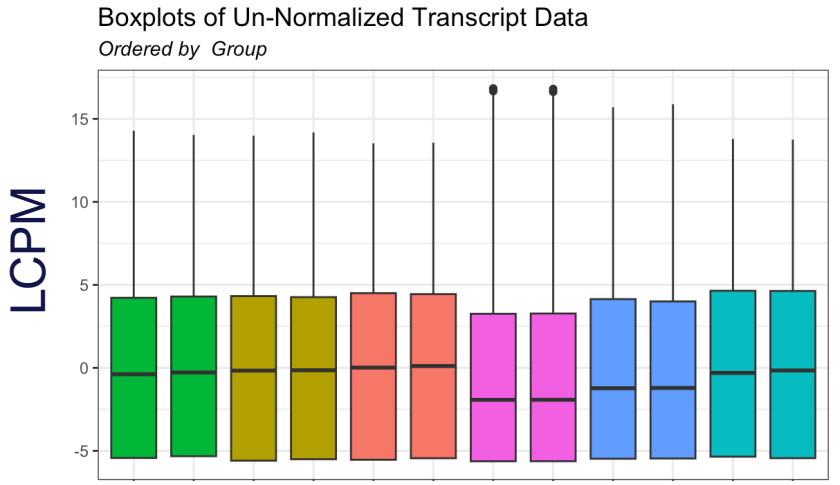




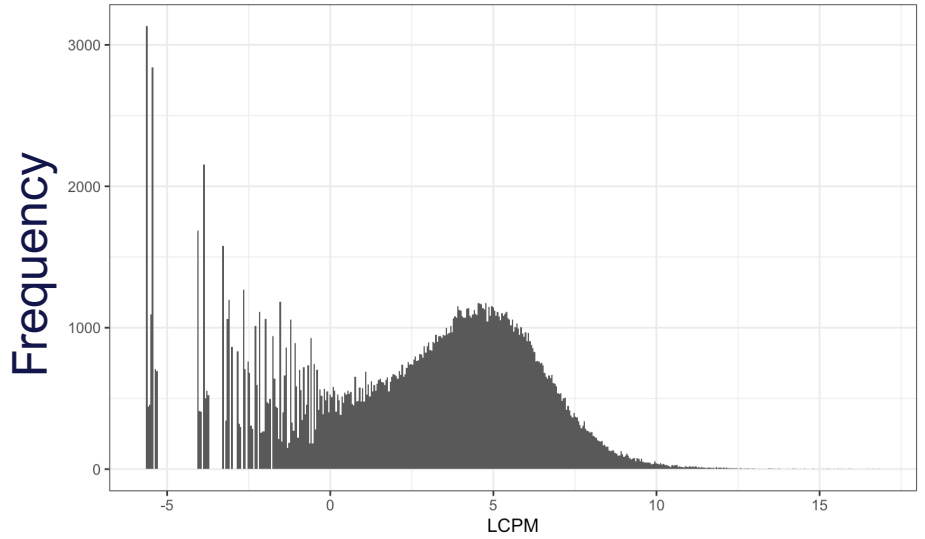
Now what does our data look like?

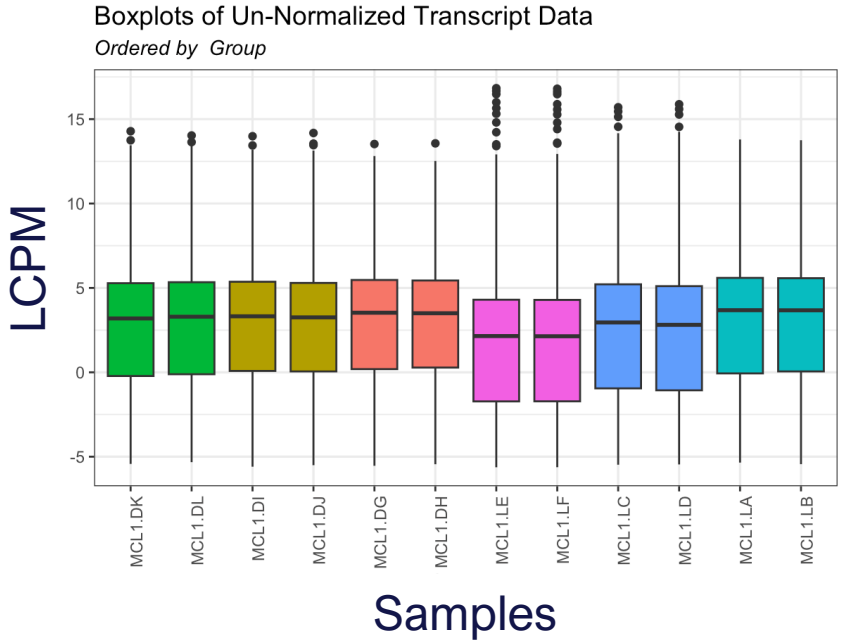
Before





After





Other caveats
-Batch effects
-Iterative data exploration

Other data exploration methods may be just as useful!

Are we done yet?

Recap

- Transformation methods of visualization and data exploration
- Data exploration via
  - Boxplots
  - Library size
  - Genes detected per sample
  - Principal component methods
  - Expression counts
- A touch of info on popular differential expression methods

Now our data is ready for differential expression algorithms!

Thank you for listening!

Questions?

Rachel.Richardson@pnnl.gov

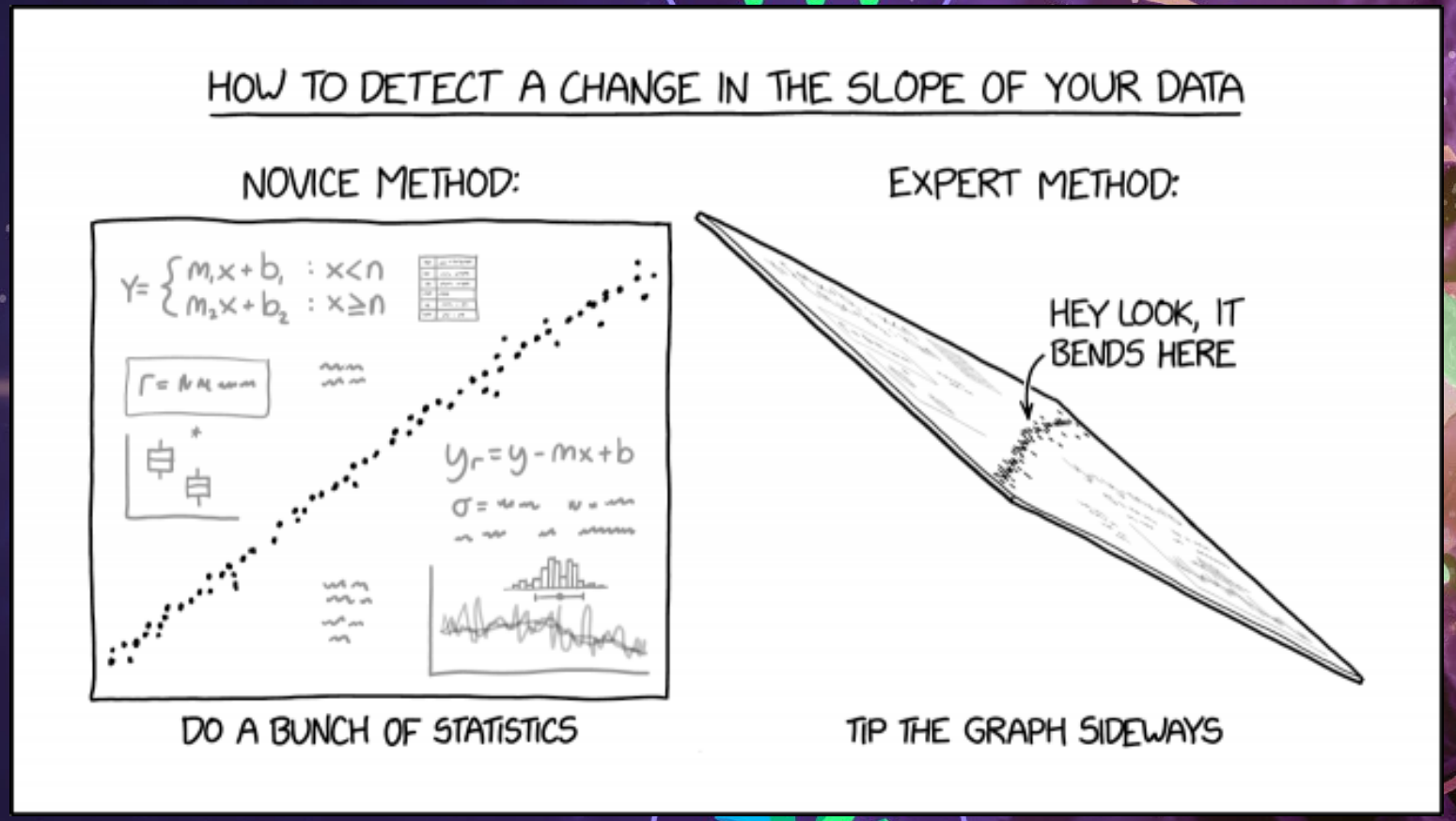

Networking Break
10:05 – 10:15 a.m.

8:30-8:35 a.m.	Introduction	Lisa Bramer
8:35-9:25	Express Yourself!: Transcriptomics from sample to counts	Bill Nelson
9:25-10:05	Pre-processing of Transcriptomics Data	Rachel Richardson
10:05-10:15	Networking Break	
10:15-11:00	**Statistical Models for Transcriptomics Data**	**Lisa Bramer**
11:00-11:45	Post-differential Expression Analyses (or how to interpret that big ole pile of data you've just generated)	Jason McDermott

Day 4
Transcriptomics

Modeling and Statistics

Lisa Bramer



Instructor Intro

Lisa Bramer

- Senior Statistician
- Computational Biology Group
  - Team Lead of Biostatistics and Data Science

lisa.bramer@pnnl.gov

Is there a vaguer term than preprocessing?

Our definition

After transcriptomic counts have been cross-tabulated for further analysis

Before differential expression analysis is performed

Samples

Genes/Transcripts

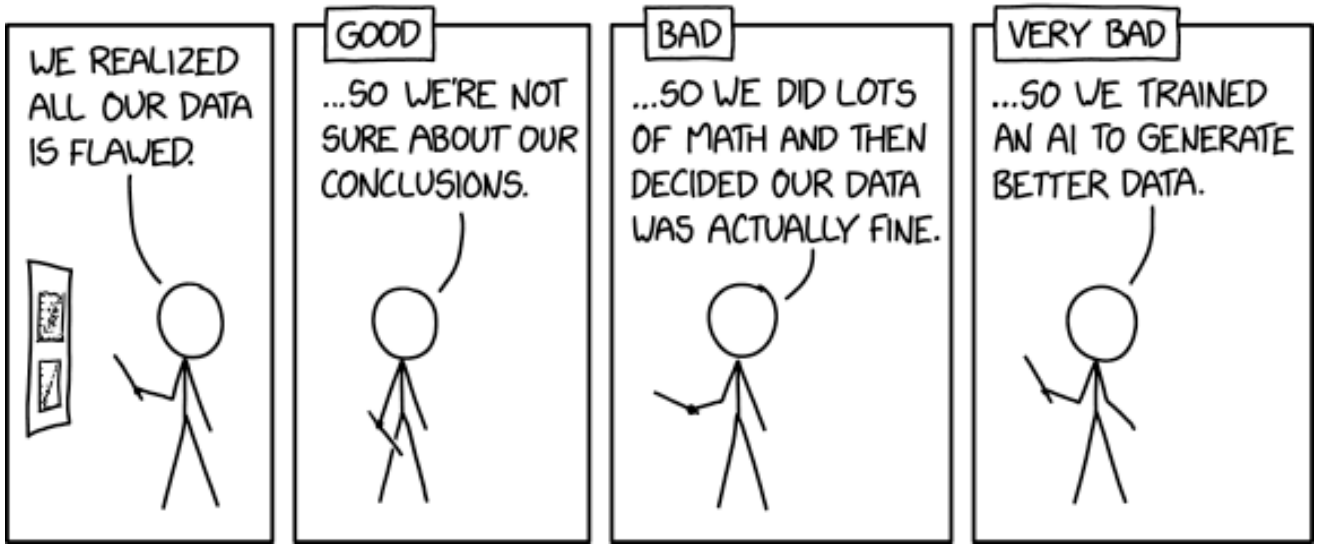
	MCL1.DG	MCL1.DH	MCL1.DI	MCL1.DJ	MCL1.DK	MCL1.DL	MCL1.LA
497097	438	300	65	237	354	287	0
100503874	1	0	1	1	0	4	0
100038431	0	0	0	0	0	0	0
19888	1	1	0	0	0	0	10
20671	106	182	82	105	43	82	16
27395	309	234	337	300	290	270	560
18777	652	515	948	935	928	791	826
100503730	0	1	0	0	0	0	0
21399	1604	1495	1721	1317	1159	1066	1334
58175	4	2	14	4	2	2	170
108664	769	752	1062	987	995	903	1381

Counts

After Preprocessing: Now Where Do we Go?

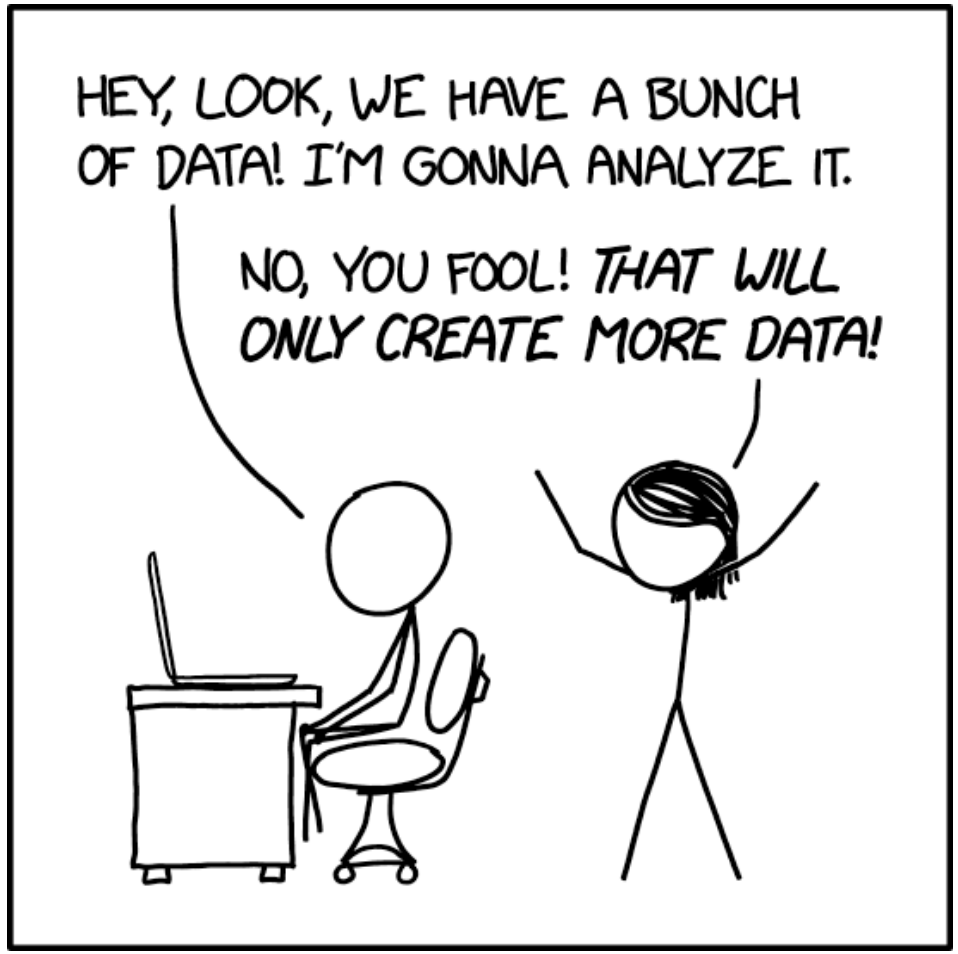
Transcriptomic counts have been cross-tabulated for further analysis

Preprocessing



Data Modeling: differential expression analysis is performed

Data Modeling and Analysis



Normalizing Count Data

- Library size considerations
  - All transcript counts in a sample = library
  - Non-biological variation can change the efficiency of read transcription
  - The total number of transcripts in a library often used as a normalization factor

Lib. 1	**Gene1: 5** **Gene2: 10** **Gene3: 200** **…** **Total: 10M reads**	**Lib. 2**	**Gene1: 2** **Gene2: 5** **Gene3: 100** **…** **Total: 5M reads**

*2-fold changes in expression would be expected across all genes for identical samples

Common scaling

Counts **p**er **m**illion:
CPM

Log **c**ounts **p**er **m**illion:
LCPM

Reads **p**er **k**ilobase of transcript per **m**illion:
RPKM

Fragments **p**er **k**ilobase of transcript per **m**illion:
FPKM

Regularized **log - rlog**
Variance **s**tabilizing **t**ransformation - **vst**

How do we scale transcriptomic data?

Gene = g
sample = i

Counts for a gene = r
Sum of counts = R = **Library size**

edgeR:

$$CPM = \frac{r_{gi}}{R_i} \times 10^6$$

$$LCPM = log_2\left(CPM + \frac{2}{R_i}\right)$$

Limma-voom:

$$CPM = \frac{r_{gi} + 0.5}{R_i + 1} \times 10^6$$

$$LCPM = log_2(CPM)$$

*The 0.5 and 2 are arbitrary! Just needs to be a small number ☺
**The +1 for voom ensures the proportion is less than 1 in all cases

How do we scale transcriptomic data?

Major among-individual differences for some transcripts can affect perceived differences for other transcripts.

A Different Solution?

TMM (Trimmed mean of M-value normalization)

$$M_g = \log_2 \frac{Y_{gk} / N_k}{Y_{gk'} / N_{k'}}$$

Y_{gk} / N_k: count for gene *g* in library *k* divided by total counts in library *k*

$Y_{gk'} / N_{k'}$: count for gene *g* in ref. lib. divided by total counts in ref. lib.

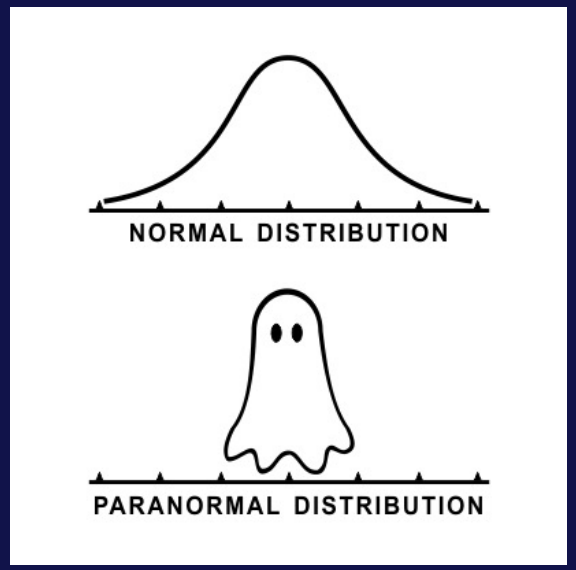
⋮

Calculate a trimmed across-gene mean of *M*

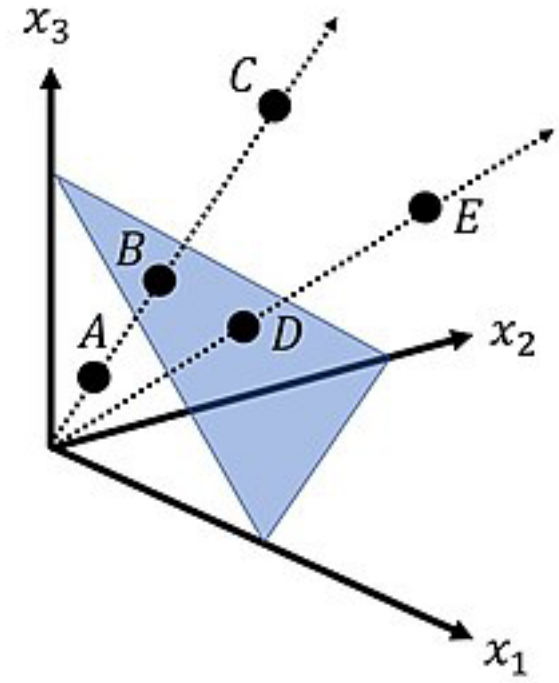
Identifies scaling factors (one for each library), which minimize log-fold changes between samples for **most of the genes.**

Assumption: Most of the genes are not differentially expressed in the biological sense, which is why basing the scaling factors on a trimmed mean of log fold changes is justified.

Attempting to be Normal



- Normalize ➔ Model assuming data follow Normal/Gaussian distribution
  - Limma-voom
  - TMM
  - ALDEx2 – Bayesian compositional approach

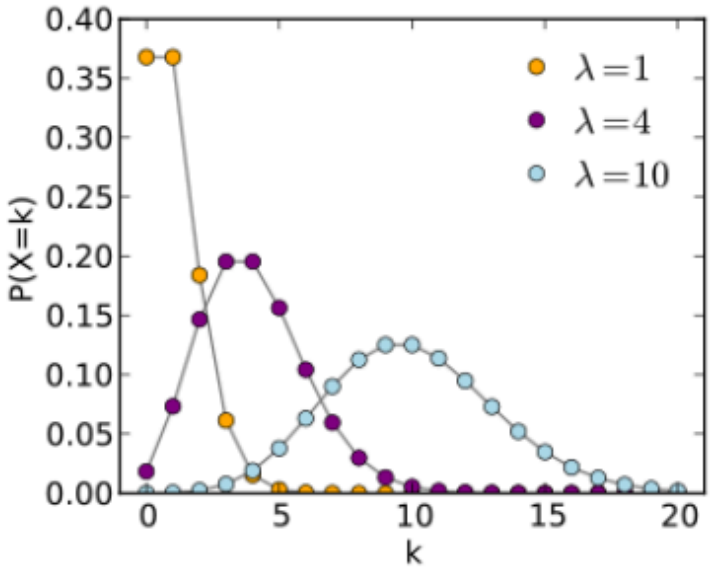


Differential Expression Analysis
(Back to thinking about raw counts)

***Want to test whether the mean number of reads for gene *X* drawn from sample type A differs from the number drawn from type B**

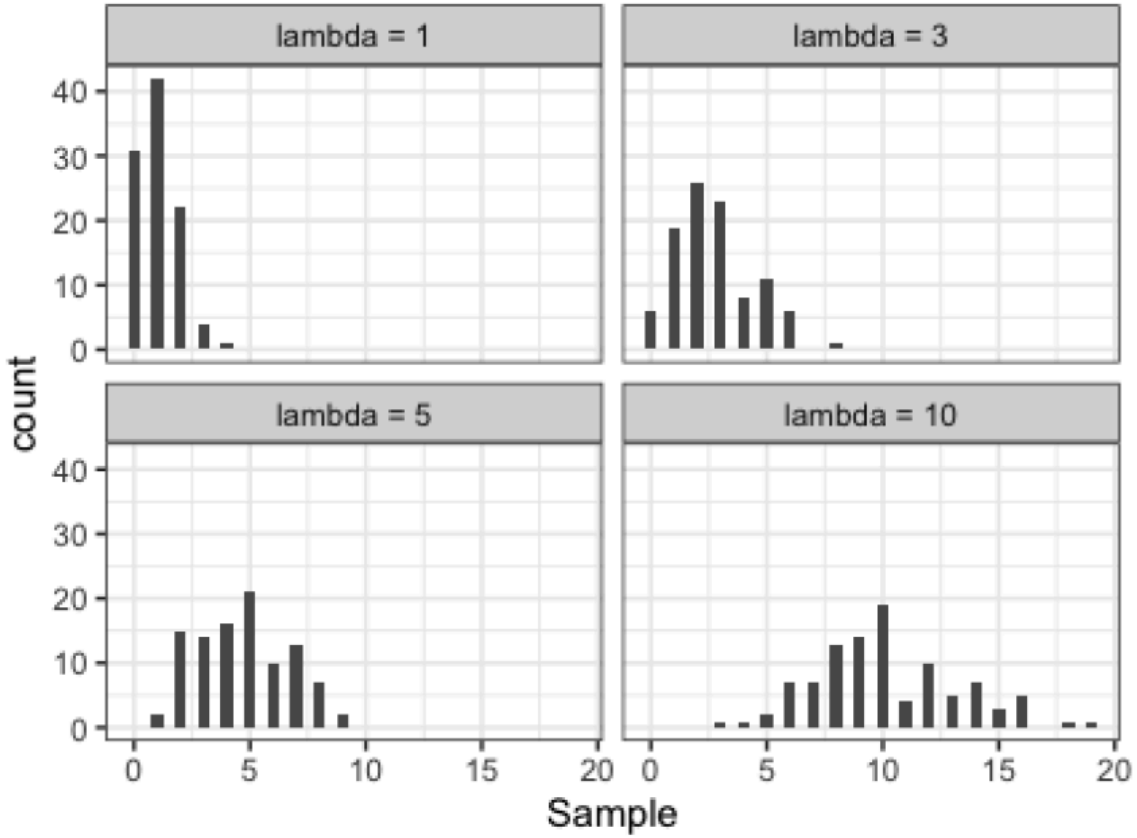
Poisson Distribution

Discrete value distribution



- General Interpretation: number of events occurring in a fixed interval of time or space (events occur with a constant mean rate and independently of the time since the last event)
- Parameters: $\lambda > 0$
- Support: $k \in \{0, 1, 2, 3, ...\}$
- PMF: $p(x = k) = \frac{\lambda^k e^{-\lambda}}{k!}$
- Mean: λ
- Variance: λ

Differential Expression Analysis
(Back to thinking about raw counts)

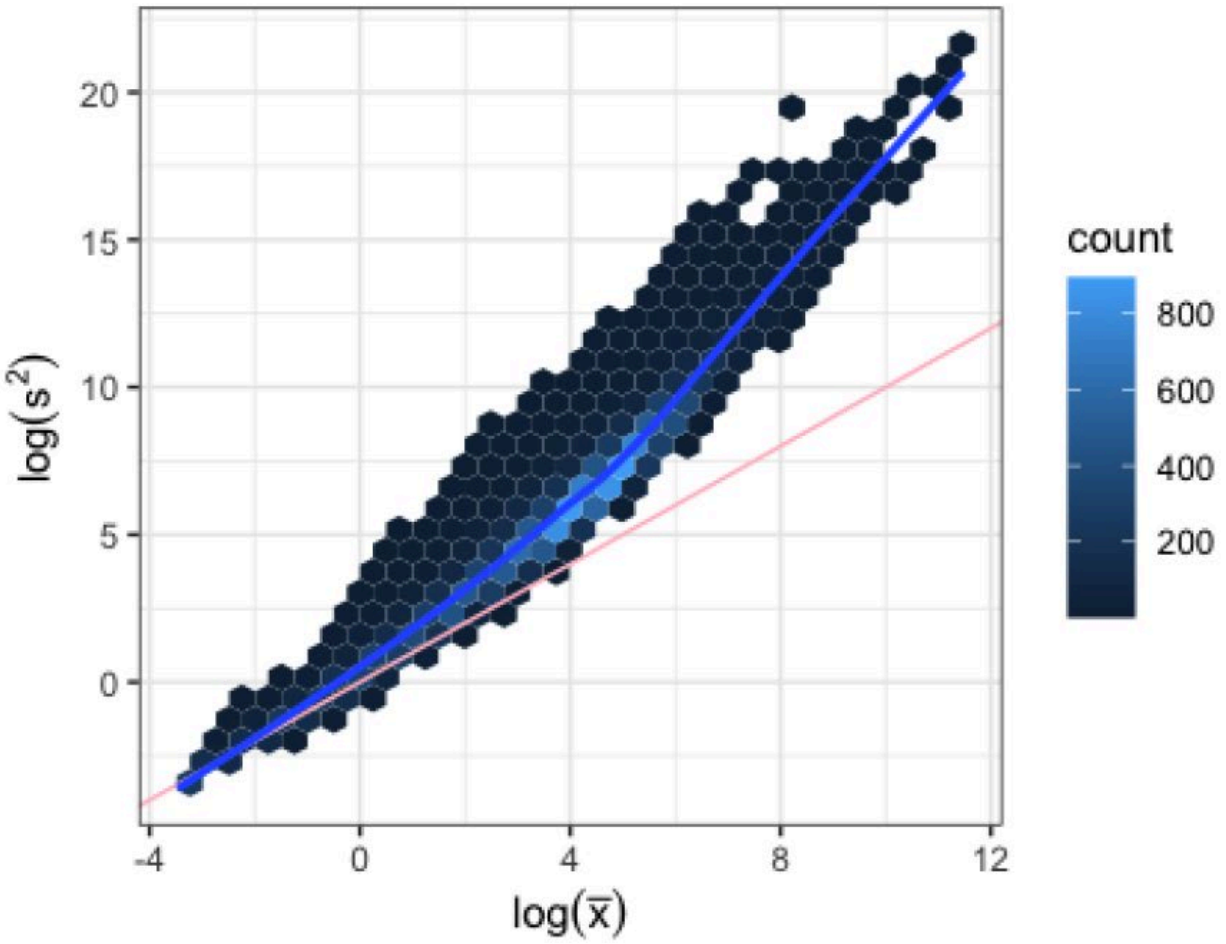


λ = mean = variance

***Early RNA-seq studies:**

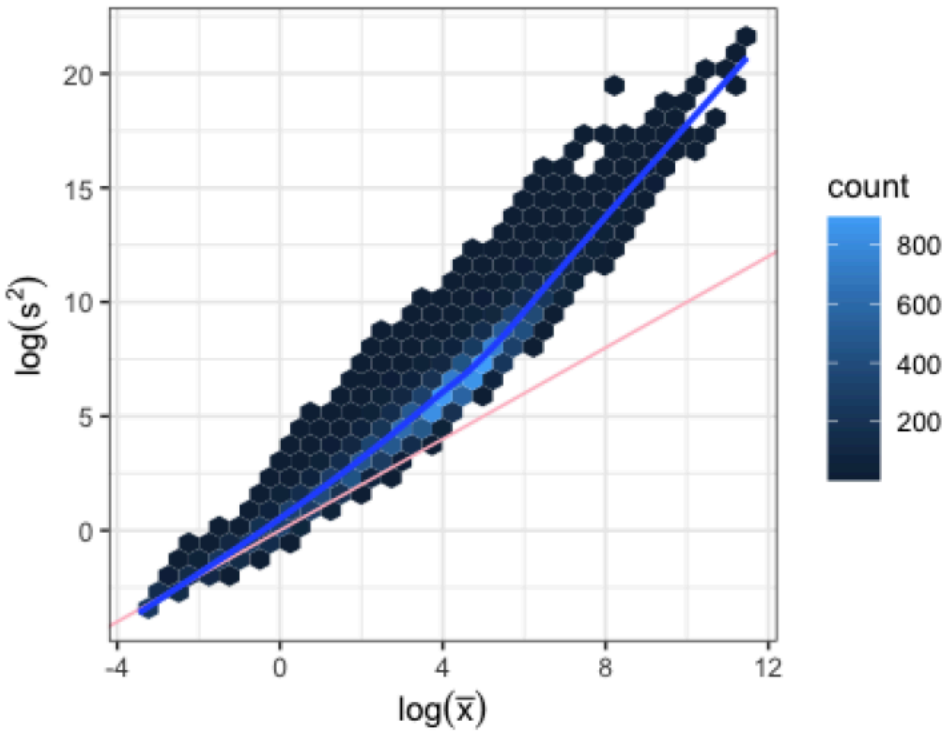
Based on technical replicates, Poisson is a good approximation

Investigate Poisson Assumption



Are the data Poisson distributed?
If not, how do they deviate from the expectation?

Differential Expression Analysis (Back to thinking about raw counts)



RNA-Seq data are overdispersed…

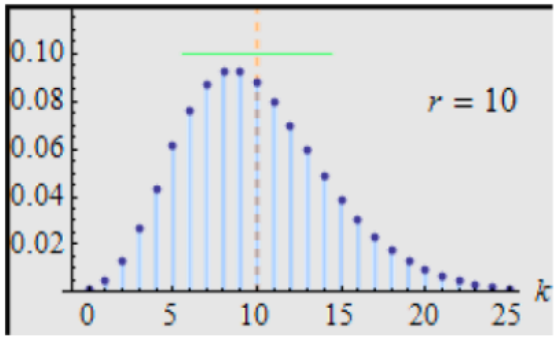
Need a different distribution

***For biological replicates, variance > mean (especially for high counts)**

Differential Expression Analysis

The Negative Binomial Distribution

*Discrete distribution with additional variance parameter



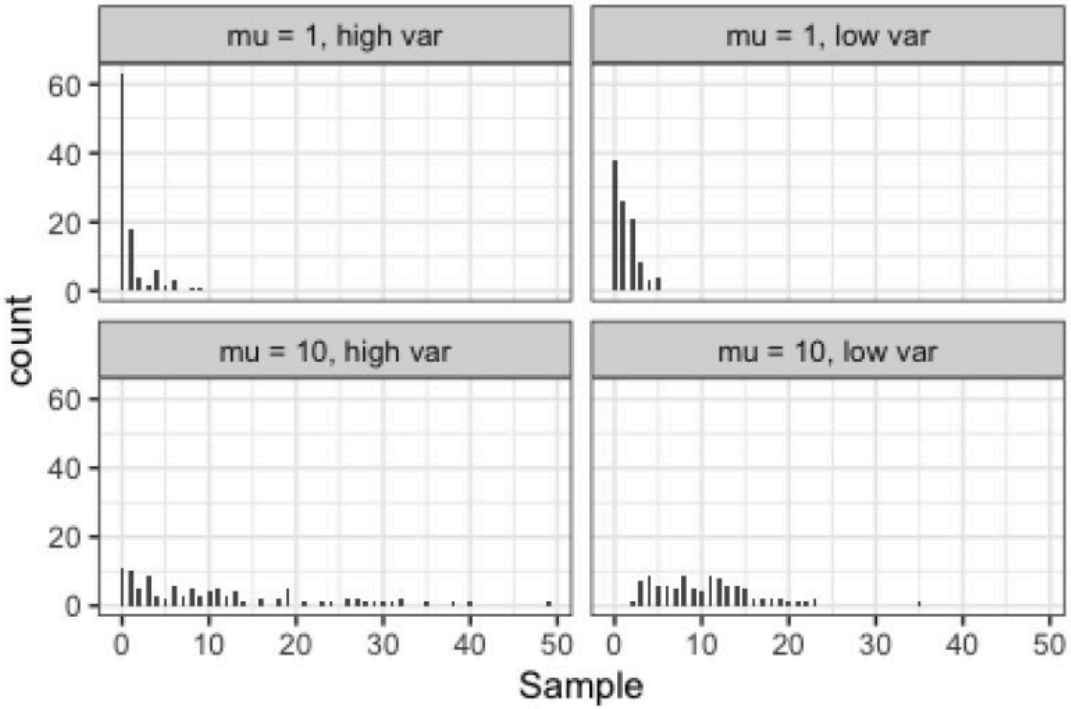
- General Interpretation: number of successes in a sequence of independent and identically distributed Bernoulli trials before a specified number of failures, r, occur
- Parameters: $0 \leq p \leq 1$, $r > 0$
- Support: $k \in \{0, 1, 2, 3, ...\}$
- PMF: $p(x = k) = \binom{k + r - 1}{k} p^k (1 - p)^r$
- Mean: $\frac{pr}{1-p}$
- Variance: $\frac{pr}{(1-p)^2}$

Differential Expression Analysis

*Discrete distribution with additional variance parameter

Alternative Parameterization

$$\Pr(X = k) = \binom{k + \frac{\mu^2}{\sigma^2 - \mu} - 1}{k} \left(\frac{\mu}{\sigma^2}\right)^{\left(\frac{\mu^2}{\sigma^2 - \mu}\right)} \left(\frac{\sigma^2 - \mu}{\sigma^2}\right)^k$$



Differential Expression Analysis

The Negative Binomial Distribution

Problem: Variance must be estimated for every gene, and RNA-Seq studies tend to have few replicates.

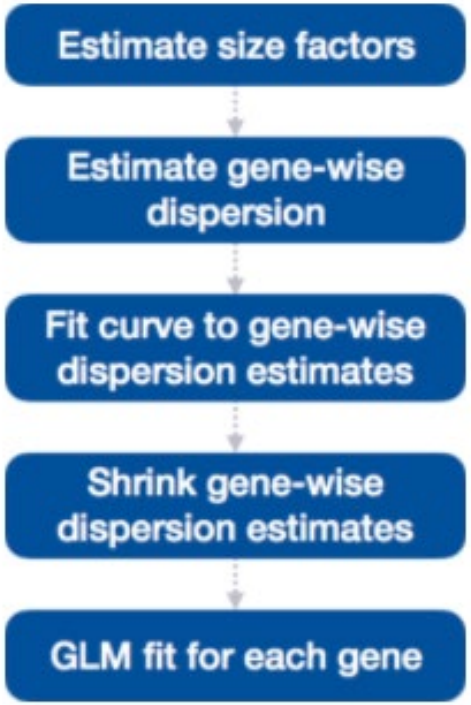
Solution: "Borrow" information about variance from the other genes to inform the genewise estimates.

Commonly used modeling approach

R packages: edgeR, DESeq (DESeq2), etc.

DESeq2

- Assumes most genes are not differentially expressed



Count Normalization:

- For each gene, compute:

$$CF_{ig} = \frac{y_{ig}}{\sqrt[n]{\prod_{i=1}^{n} y_{ig}}}$$

- Then for each sample, compute:

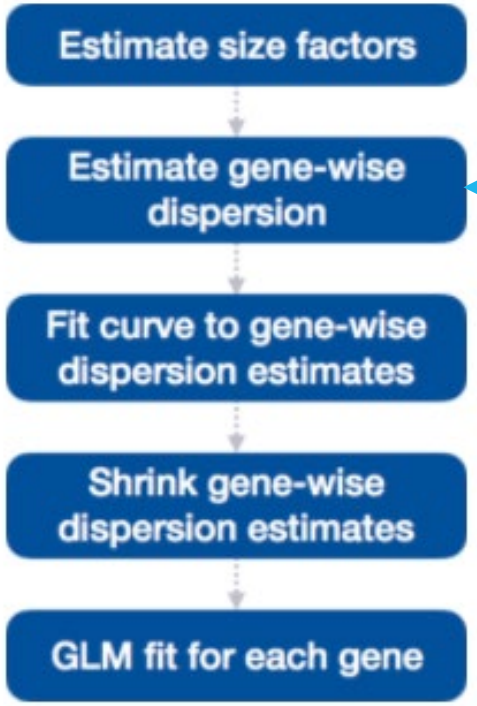
$$CF_i = median_g\{CF_{ig}\}$$

- Finally, normalize the count data for each sample:

$$y_{ig}^{*} = y_{ig} / CF_i$$

DESeq2

- Assumes most genes are not differentially expressed

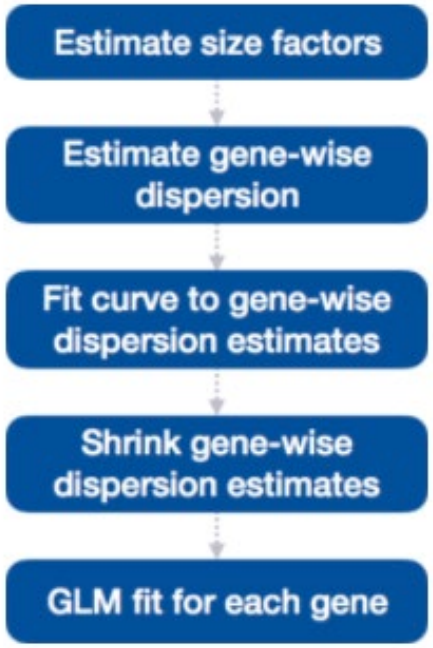


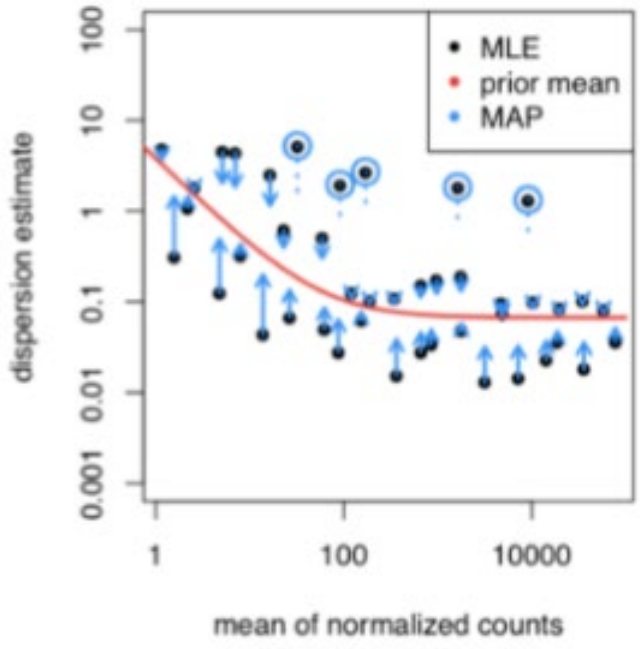
For each gene, estimate α in:

$$Var(y_g^*) = \overline{y_g^*} + \alpha(y_g^*)^2$$

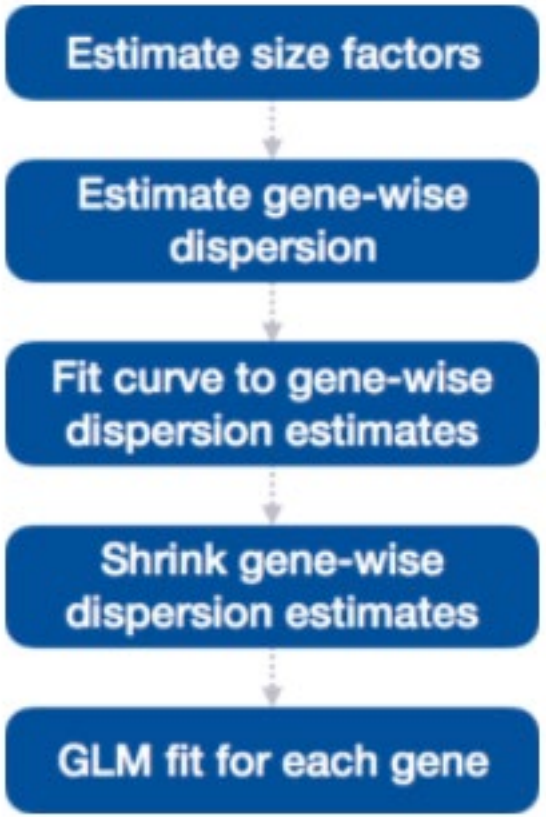
DESeq2

- Genes with low dispersion estimates are shrunken towards the curve ,and shrunken values are output for fitting of the model and differential expression testing.
- Dispersion estimates that are slightly above the curve are also shrunk toward the curve
- Genes with extremely high dispersion values are not. This is due to the likelihood that the gene does not follow the modeling assumptions and has higher variability than others for biological or technical reasons





DESeq2



$$y_{ig}^{*+} = factor1 + factor2 \ \ldots$$

$$y_{ig}^{*+} \sim NB$$

Likelihood Ratio Test
Multiple Test/FDR Correction

edgeR

Gene = g
sample = i

Counts for a gene = r
Sum of counts = R = **Library size**

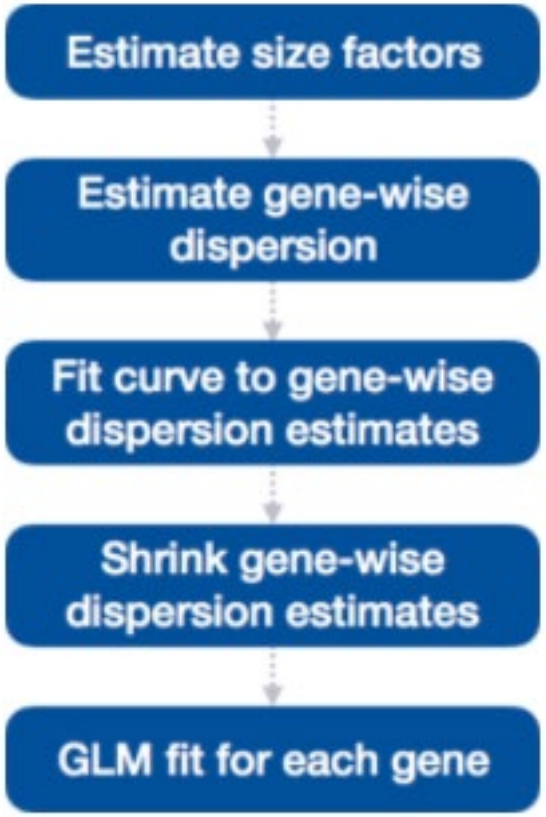
Initial Scaling

edgeR:

$$CPM = \frac{r_{gi}}{R_i} \times 10^6$$

$$LCPM = log_2\left(CPM + \frac{2}{R_i}\right)$$

edgeR



Dispersion Estimation and Adjustment:

- Similar to DESeq2
- But there is a robust dispersion estimation function which reduces the effect of individual outlier counts, and a robust arguments to estimation so that hyperparameters are not overly affected by genes with very high within-group variance

Model and DE Approach

- Also generalized linear model (GLM) assuming Negative Binomial distribution

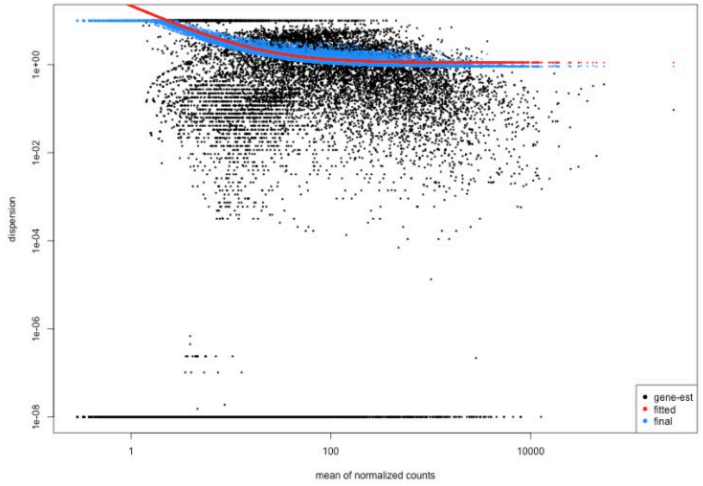
What are the Differences?

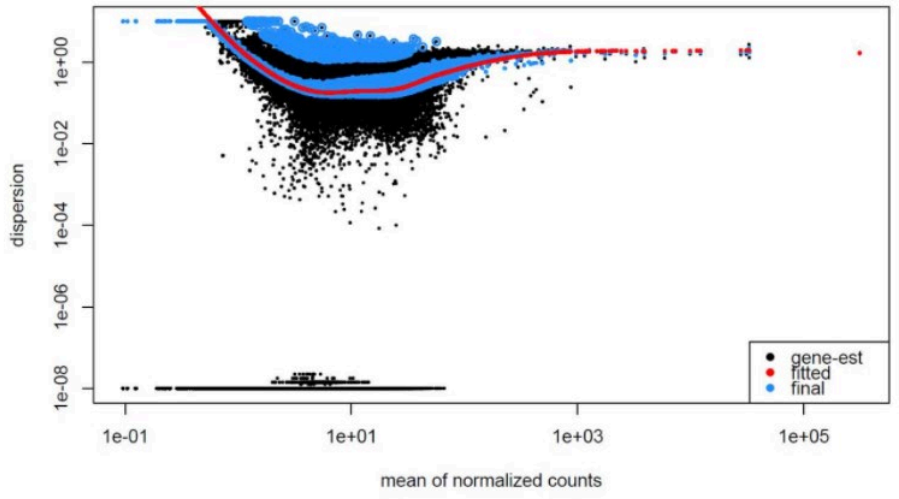
Primary Differences

- Methods for “normalization” of counts
- Estimation method for dispersion parameters
  - DESeq (which tends to overestimate dispersion) ➔ higher FDR
- DESeq2 by default (which can all optionally be turned off):
  - it finds an optimal value at which to filter low count genes
  - flags genes with large outlier counts or removes these outlier values when there are sufficient samples per group (n>6)
- edgeR is more flexible, but requires better understanding of what you’re doing and why

CAUTION AREAS

- Assumes most genes are not differentially expressed
  - Approaches are less suitable for comparing libraries that are very different, with many truly differentially expressed genes (e.g. very different tissues)
- Rely on Negative Binomial distribution and dispersion estimation





Transformation Based Methods: Limma w/voom

- Assumes a transformation can get us to an approximate Normal distribution
  - limma w/voom (log CPM)

$$y_{ga} = \log_2 \left(\frac{r_{gi} + 0.5}{R_i + 1.0} \times 10^6 \right)$$

to avoid zero in numerator

To ensure that this quotient is > 0 and < 1

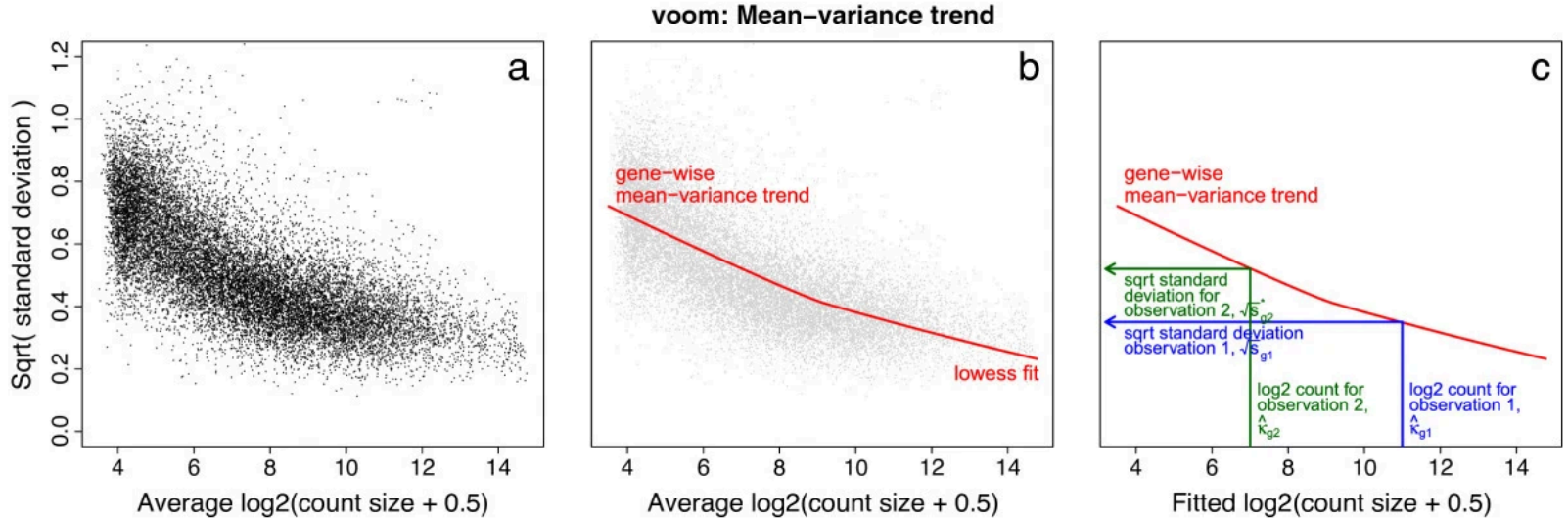
r_{gi} = **count for gene *g* in library *i***

R_i = **total count in library *i***

Transformation-Based Methods

limma w/voom (log CPM)

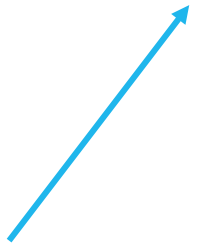
- “Different samples may be sequenced to different depths, so different count sizes may be quite different even if the cpm values are the same. For this reason, we wish to model the mean-variance trend of the log-cpm values at the individual observation level, instead of applying a gene-level variability estimate to all observations from the same gene”



Transformation Based Methods: ALDEx2

- Assumes a transformation can get us to an approximate Normal distribution
  - ALDEX2 (compositional approach)

$$clr(y_{ig}) = \ln \frac{y_{ig} + \varepsilon}{\sqrt[n]{\prod_{i=1}^{n} y_{ig} + \varepsilon}}$$



denominator can be altered based on data (e.g. sparsity differences)

Transformation-Based Methods

- ALDEX2
  - Assume a Bayesian model where counts are drawn from a Dirichlet distribution

$$p\,[n1, n2, \ldots] | \sum N = \mathrm{Dir}\,([n1, n2, \ldots] + ½)$$

$$c_{i,j} = \log_2 (p_{i,j}) - \mathrm{meanlog}_2 (p_j)$$

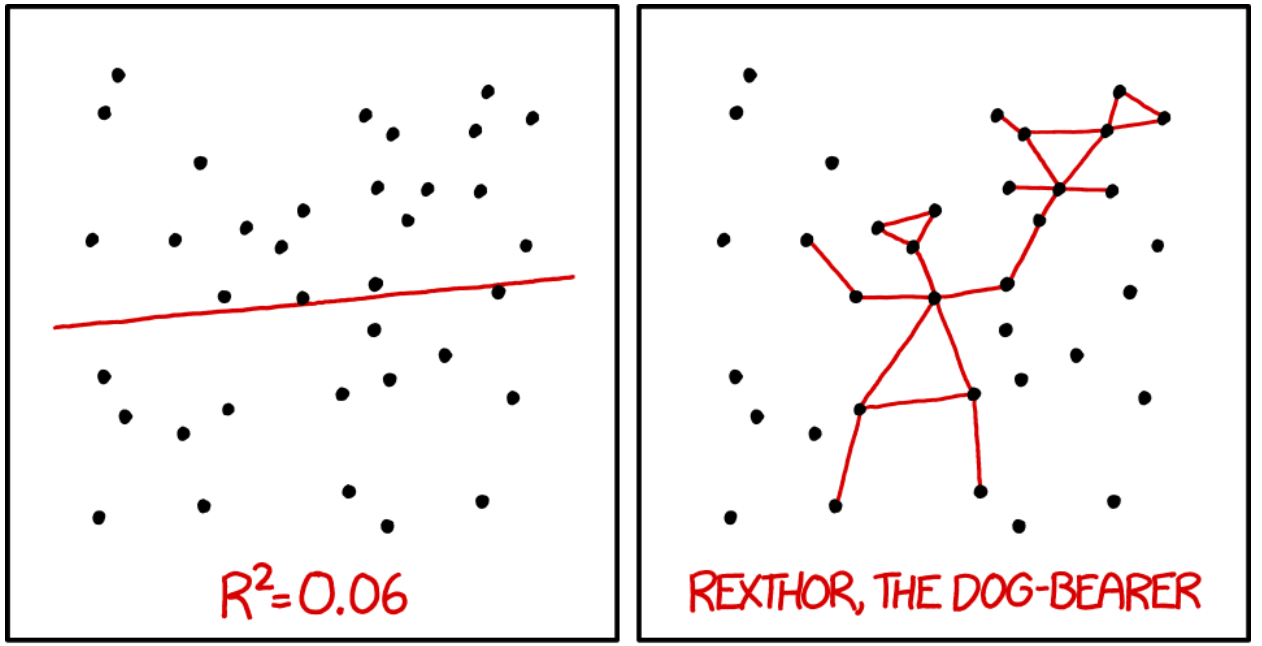
Step		C1	C2	C3	E1	E2	E3
counts + prior		69.5	185.5	70.5	511.5	659.5	462.5
Monte Carlo Dirichlet instances	1)	2.21e–5	2.94e–5	2.55e–5	1.35e–4	1.32e–4	1.23e–4
	2)	2.13e–5	2.98e–5	2.44e–5	1.25e–4	1.41e–4	1.22e–4
	3)	2.61e–5	3.06e–5	2.33e–5	1.16e–4	1.34e–4	1.20e–4
clr transform	1)	6.50	7.58	6.58	9.30	9.48	9.13
	2)	6.45	7.60	6.50	9.19	6.73	7.59
	3)	6.73	7.59	6.47	9.12	9.46	9.07

Transformation-Based Methods

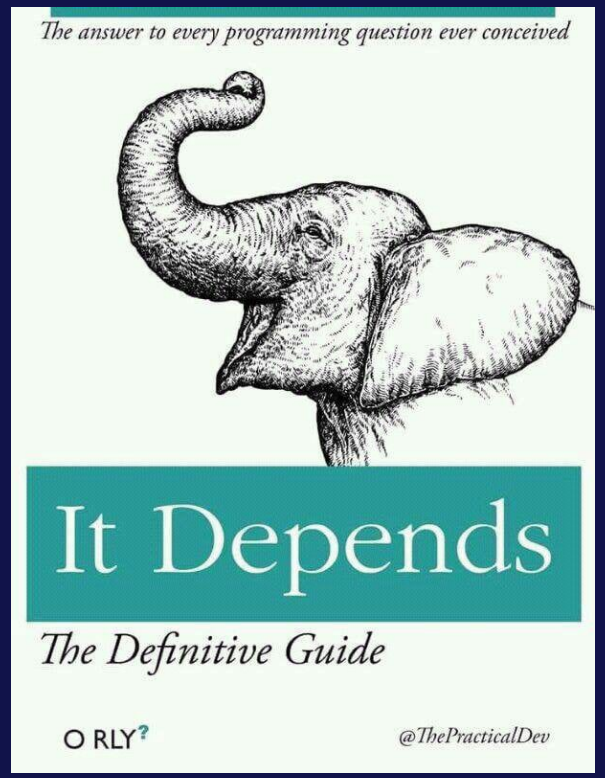
- Run traditional statistics (e.g. ANOVA, linear models, etc.)
  - limma w/voom – post-hoc FDR
  - ALDEX – Bayesian draws quantify uncertainty and are leveraged for FDR control

Step		C1	C2	C3	E1	E2	E3
counts + prior		69.5	185.5	70.5	511.5	659.5	462.5
Monte Carlo Dirichlet instances	1)	2.21e−5	2.94e−5	2.55e−5	1.35e−4	1.32e−4	1.23e−4
	2)	2.13e−5	2.98e−5	2.44e−5	1.25e−4	1.41e−4	1.22e−4
	3)	2.61e−5	3.06e−5	2.33e−5	1.16e−4	1.34e−4	1.20e−4
clr transform	1)	6.50	7.58	6.58	9.30	9.48	9.13
	2)	6.45	7.60	6.50	9.19	6.73	7.59
	3)	6.73	7.59	6.47	9.12	9.46	9.07
significance test	1)	0.01375					
	2)	0.01457	Mean:	**0.014**			
	3)	0.01349					
FDR adjustment	1)	0.0778					
	2)	0.0795	Mean:	**0.078**			
	3)	0.0761					

Which Method is “Best”?



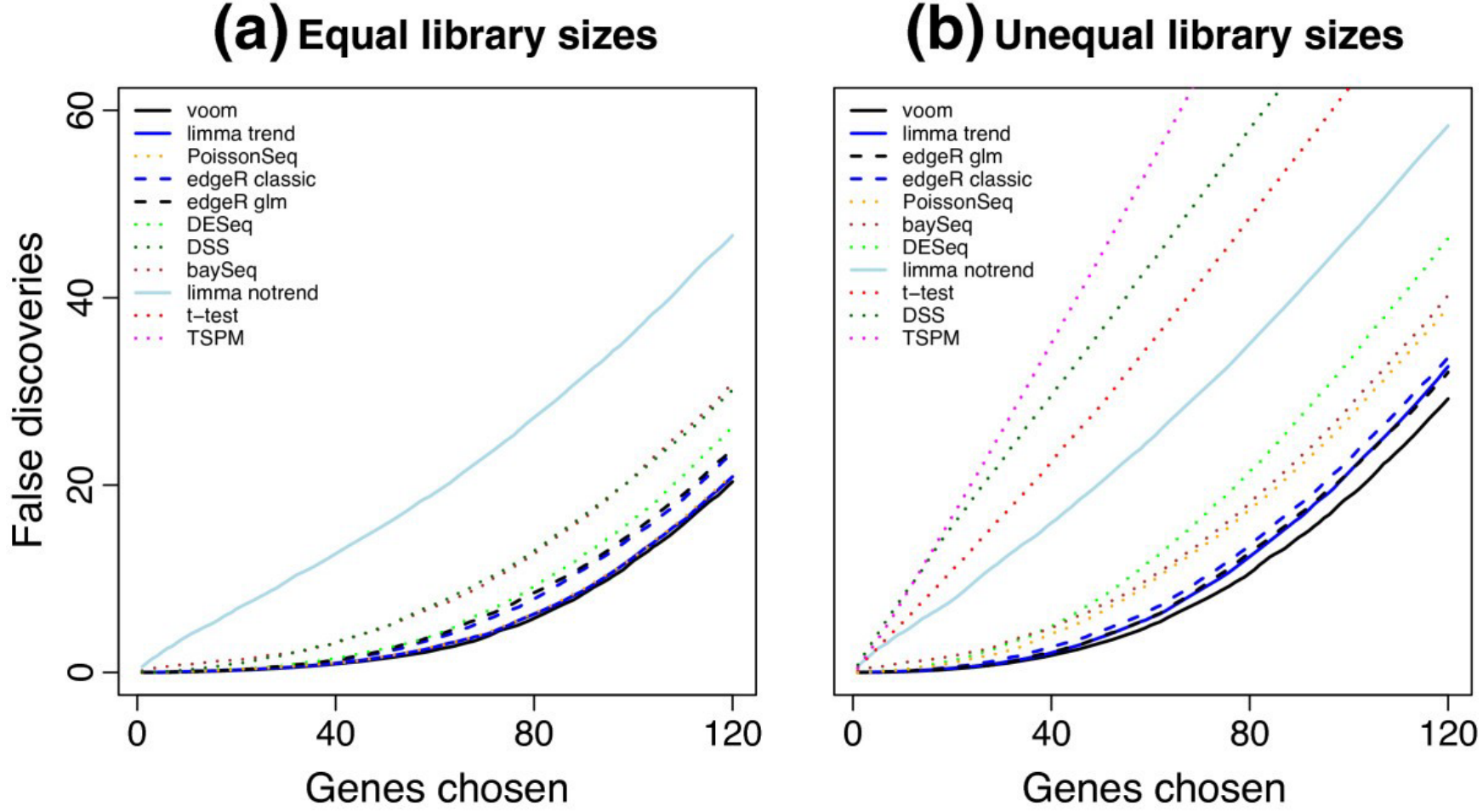
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.



Which Method is “Best”?

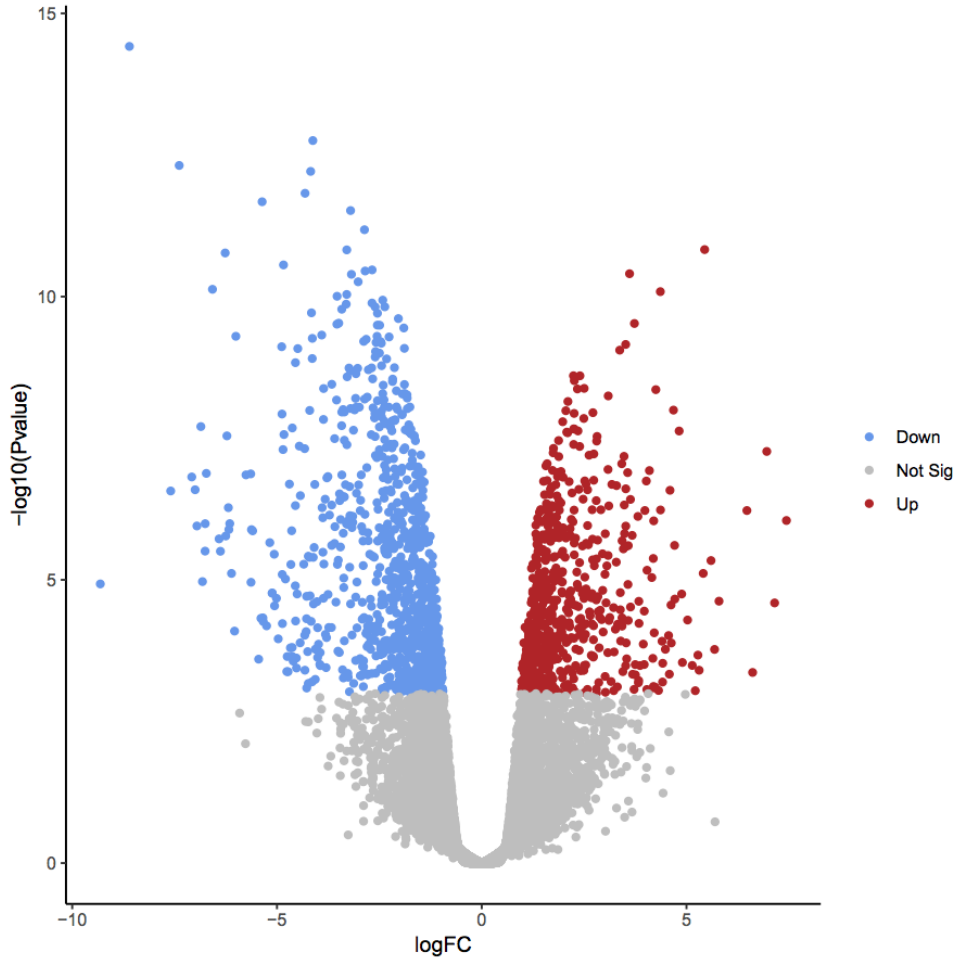
- So . . . which method is “best”?
  - It depends
    - on the data and experimental design
  - Assumptions for DESeq2 and edgeR are more stringent
    - AND SHOULD BE EVALUATED WITH EVERY DATASET
  - limma w/ voom or edgeR with quasi-likelihood are computationally MUCH more efficient
  - IMO, it does seem that limma-voom and ALDEX2 do a better job at always being under the nominal FDR (although they can have reduced sensitivity compared to DESeq2 and edgeR)
    - Especially, when the sample sizes are small (n=3 per group), when the fold changes are small, or when the counts are small

Which Method is “Best”?



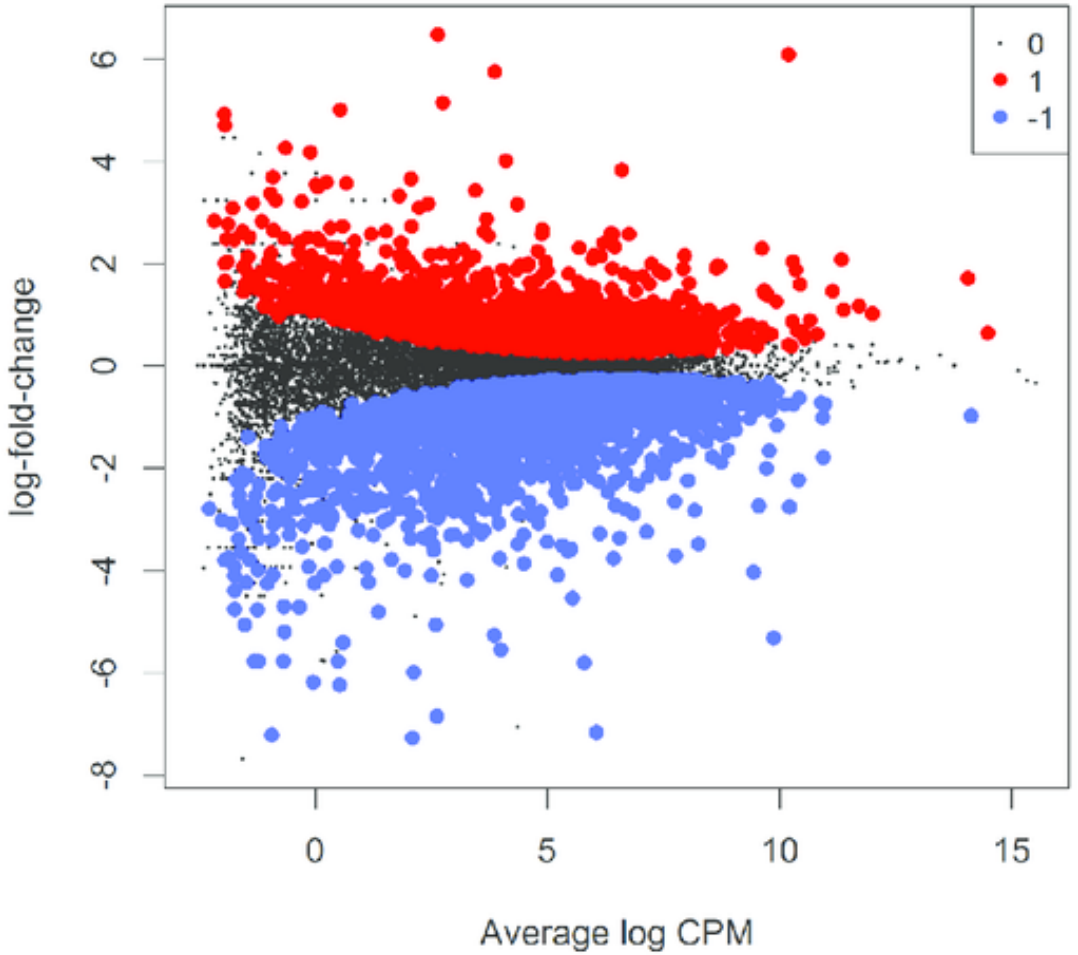
Results Output Types

- Volcano Plots – one per comparison



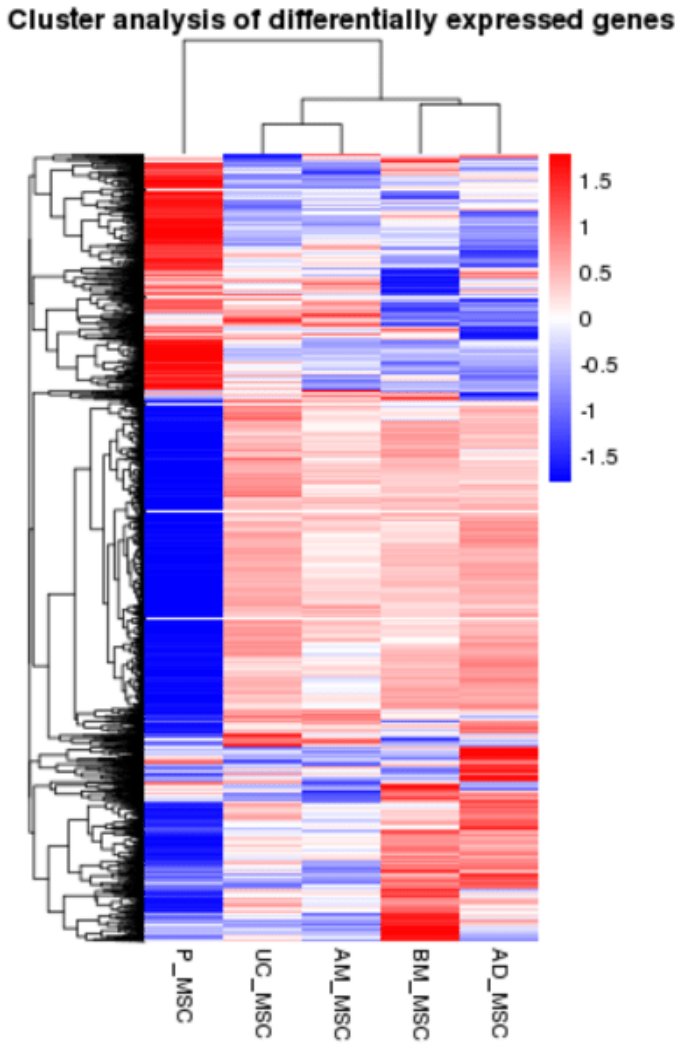
Results Output Types

- Fold-Change vs Normalized Count – one per comparison

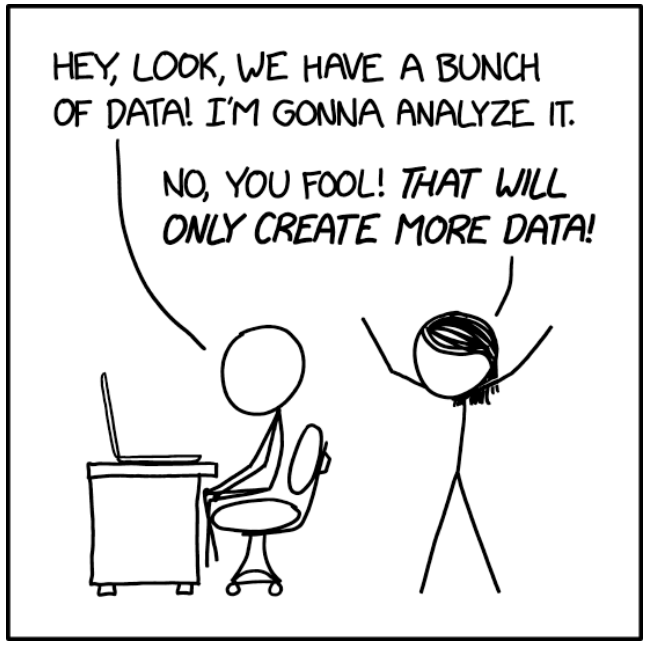


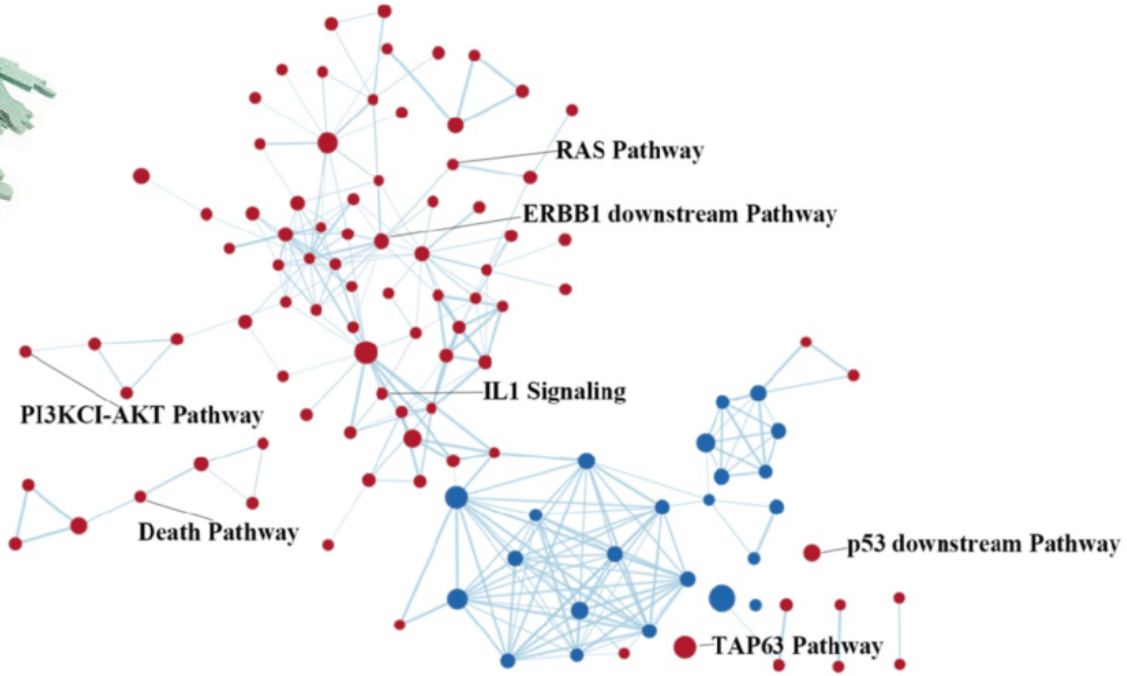
Results Output Types

- Heatmap – multiple comparisons



Results – Putting Things in Perspective





Thank you for listening!

Questions?
lisa.bramer@pnnl.gov

8:30-8:35 a.m.	Introduction	Lisa Bramer
8:35-9:25	Express Yourself!: Transcriptomics from sample to counts	Bill Nelson
9:25-10:05	Pre-processing of Transcriptomics Data	Rachel Richardson
10:05-10:15	Networking Break	
10:15-11:00	Statistical Models for Transcriptomics Data	Lisa Bramer
11:00-11:45	**Post-differential Expression Analyses (or how to interpret that big ole pile of data you've just generated)**	**Jason McDermott**

EMSL

Post-differential expression analyses

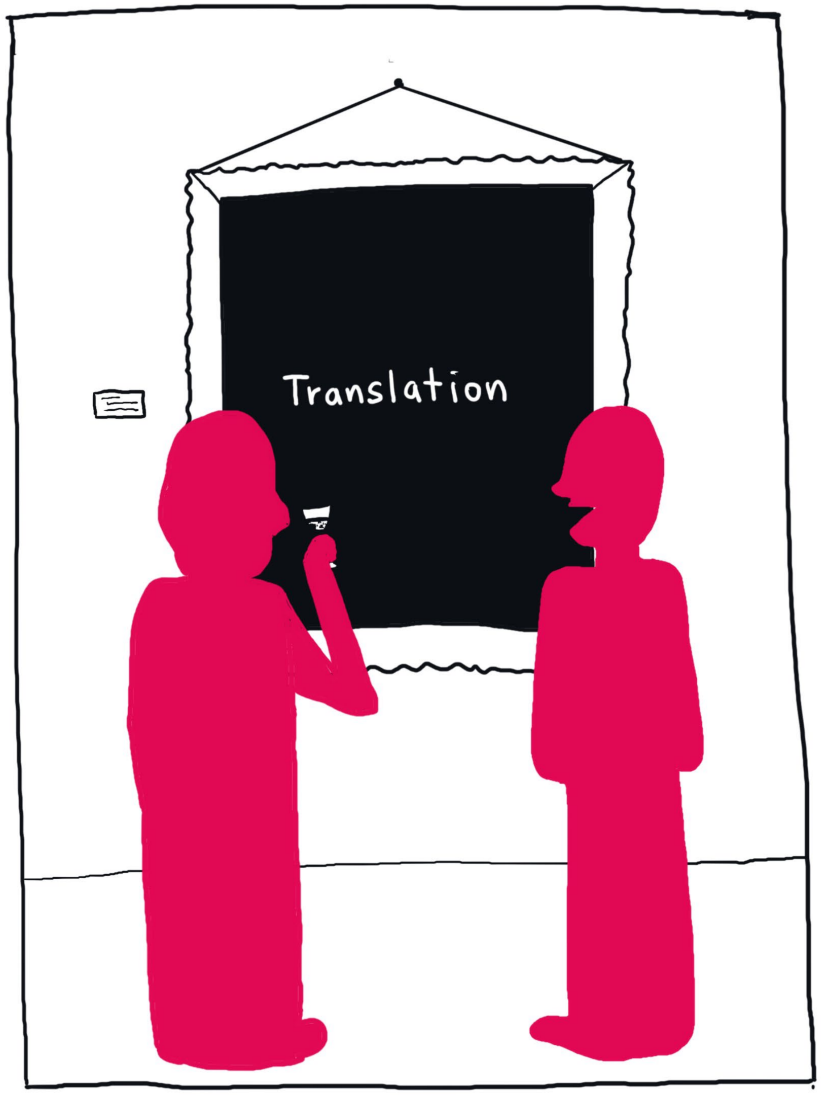
(or how to interpret that big ole pile of data you've just generated)

Jason McDermott
Team Lead, Systems Biology

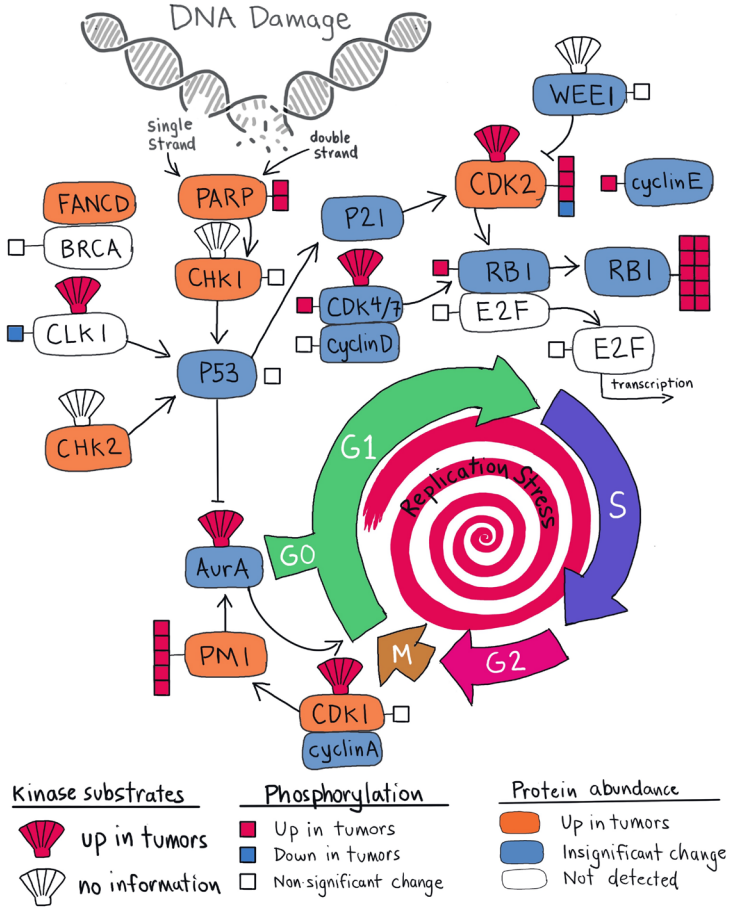
What's the overall point of post-expression analysis?

General approaches to post-expression analysis

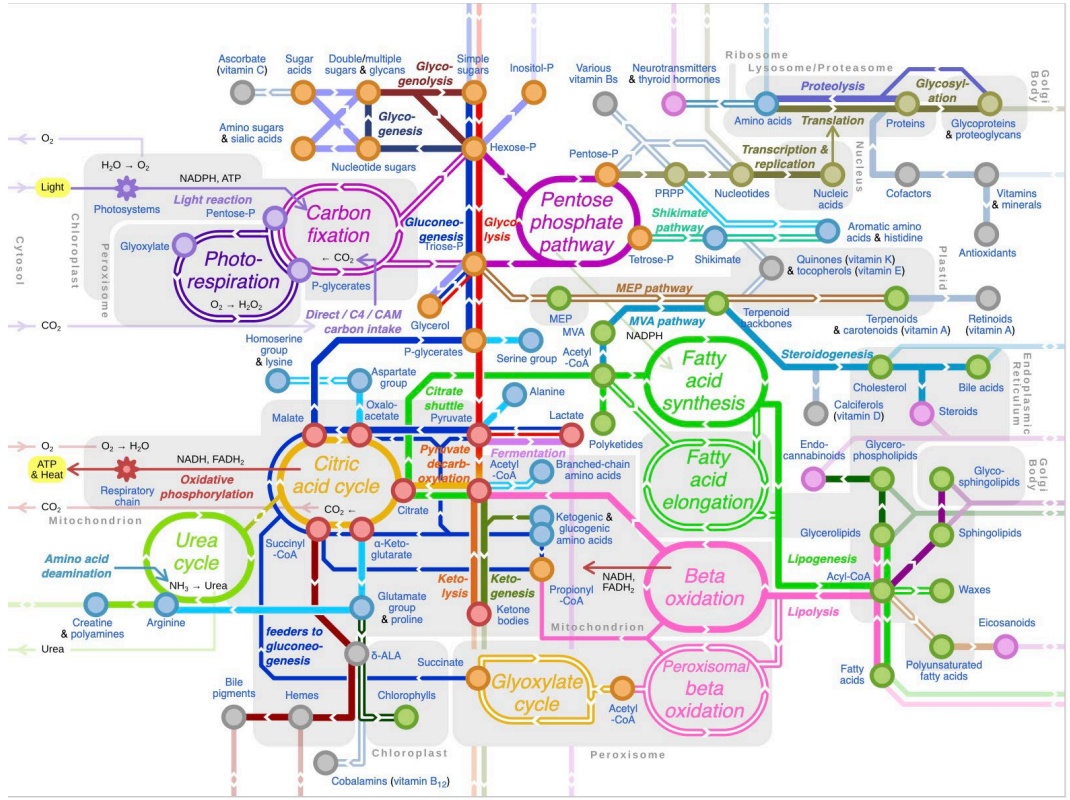
- *Visualization*
  - *Heatmaps*
  - *Ordination (PCA, etc.)*
  - *Pathway overlay*
- Enrichment approaches
  - Pathway and functional group enrichment
  - Regulon analysis
  - Interaction enrichment
- Network analysis
  - Existing known interactions
    - STRING, PPIs, regulon, metabolic, etc.
  - Inferred associations



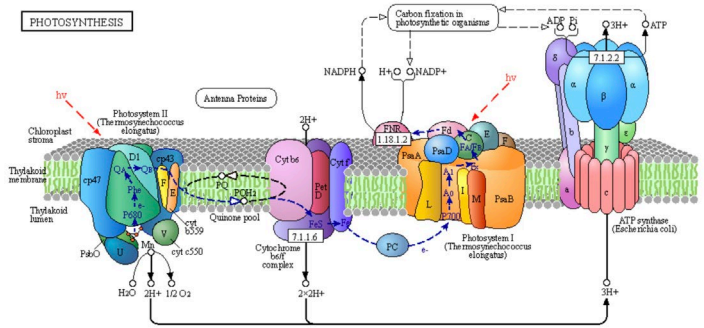
What is a ‘pathway’?



Joshi, et al. Cell Reports Medicine 2023

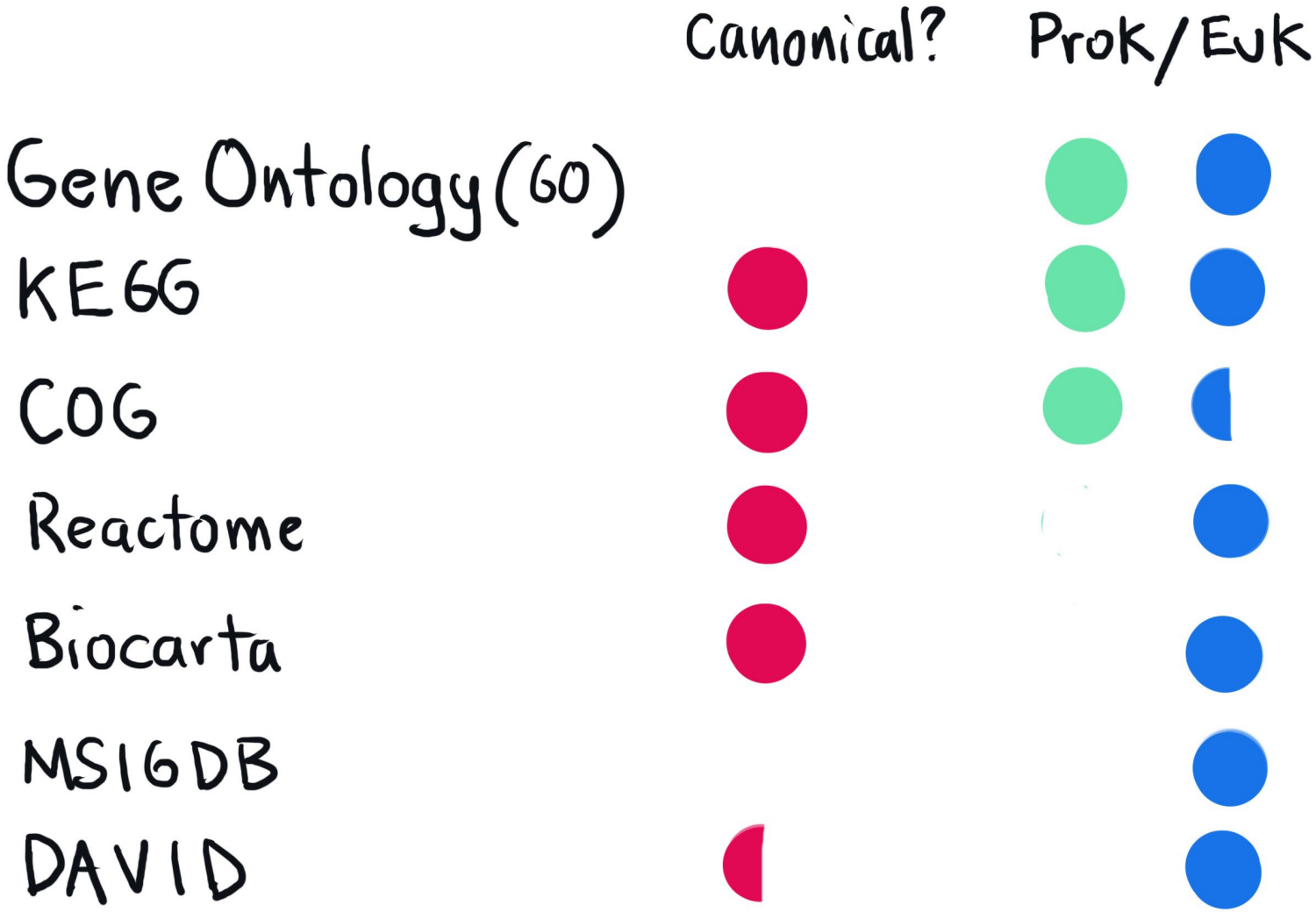


Wikipedia



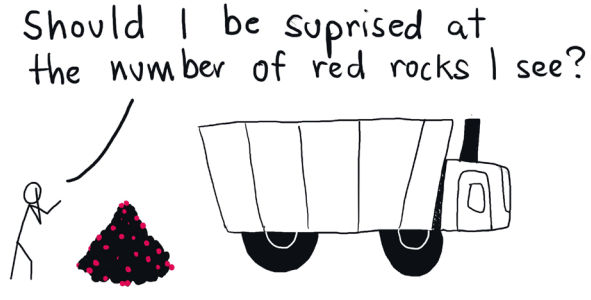
KEGG https://www.genome.jp/pathway/map00195

Pathway Databases

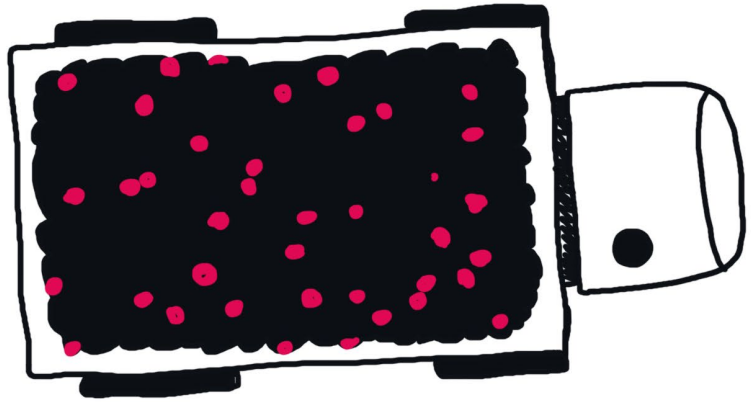


Enrichment?

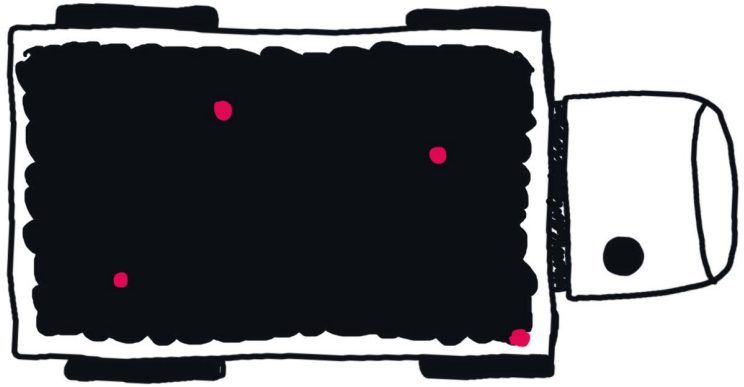
Enrichment?



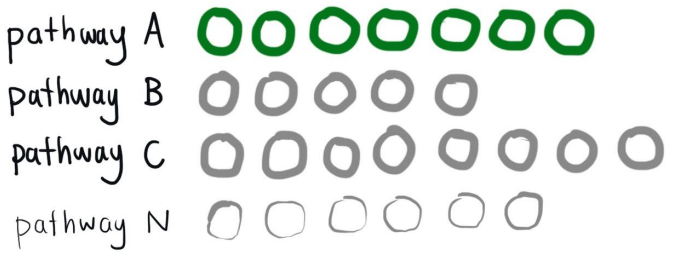
Not surprised

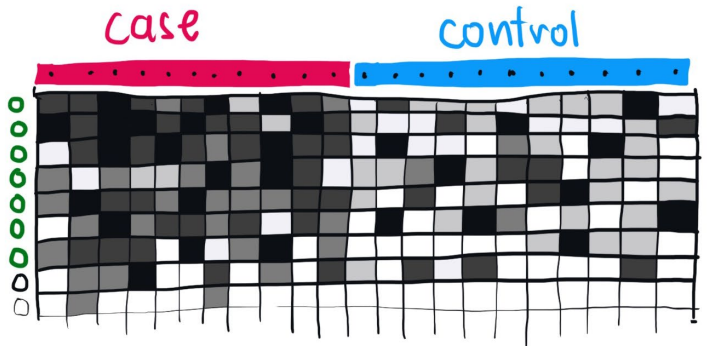


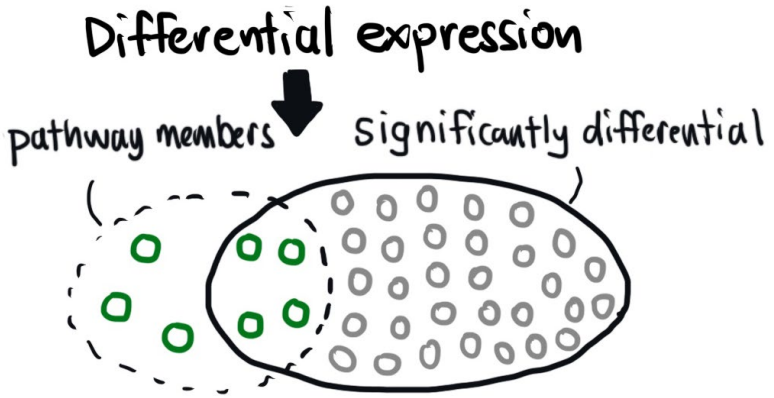
Pretty surprised



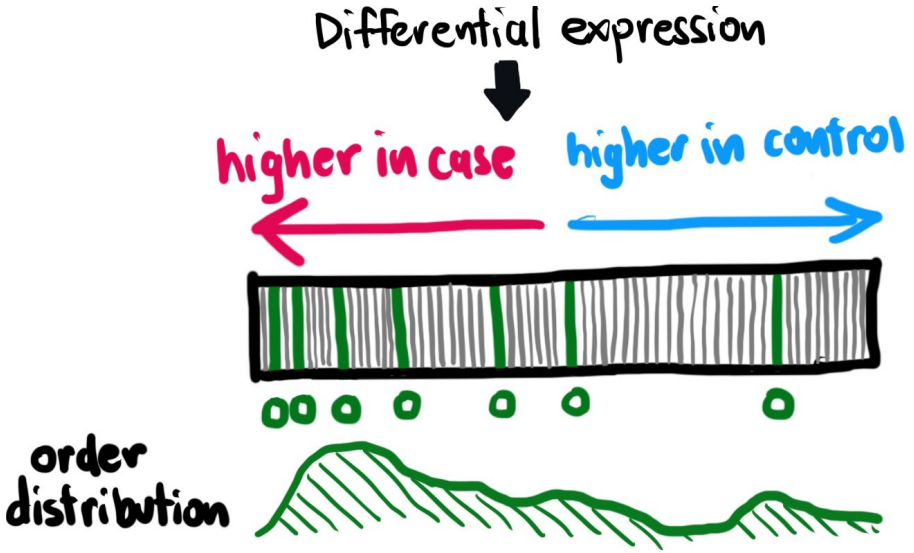
Enrichment Approaches







Hypergeometric test (DAVID, etc.)



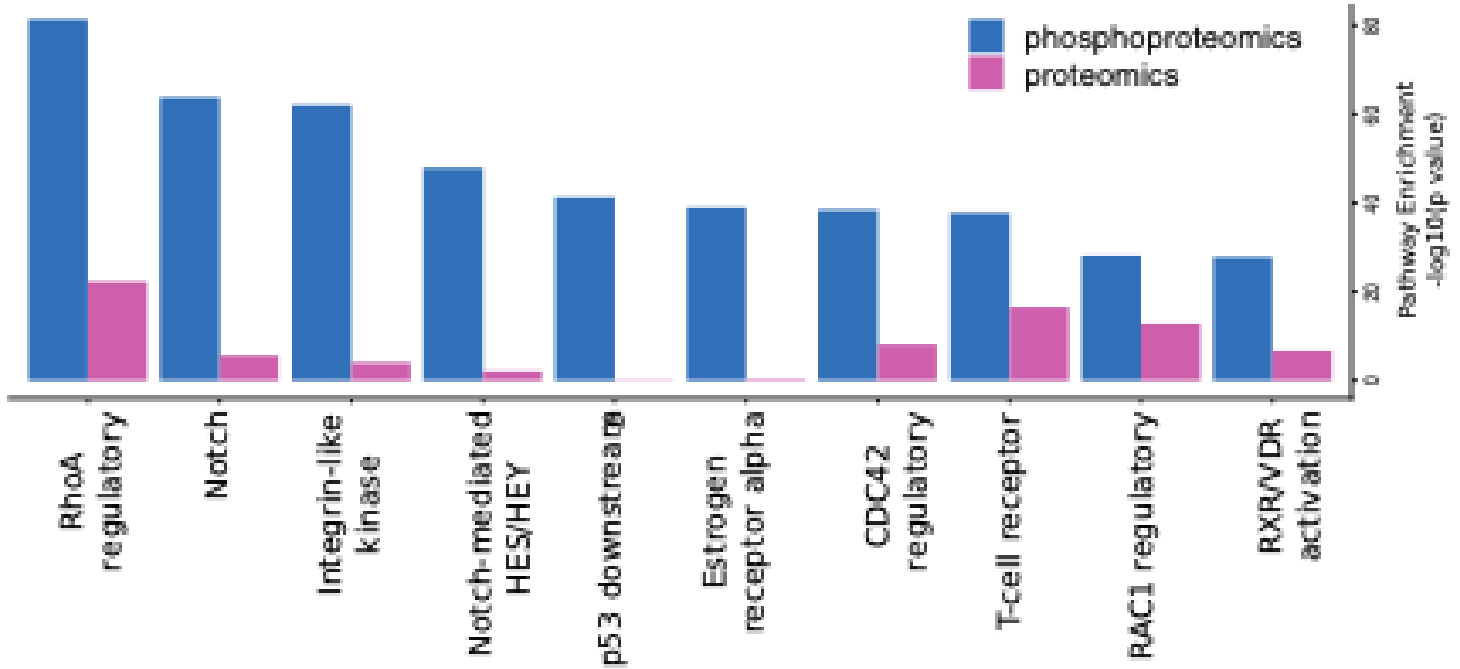
Gene Set Enrichment Analysis (GSEA)

Danna, et al. Journal of Proteome Research 2021

How to interpret enrichment results

- Look at how many genes are represented
- Do you want to consider negative enrichment (depletion?)
- Consider p-value (how surprised I am) and magnitude (how much of a difference) together
- Know the type of pathway database you're using
- Mind the background!!!
  - Use an appropriate gene set
  - For comparisons what is being compared to?

McDermott, et al. Cell Reports Medicine 2020



Example Questions That Can Be Asked with Enrichment

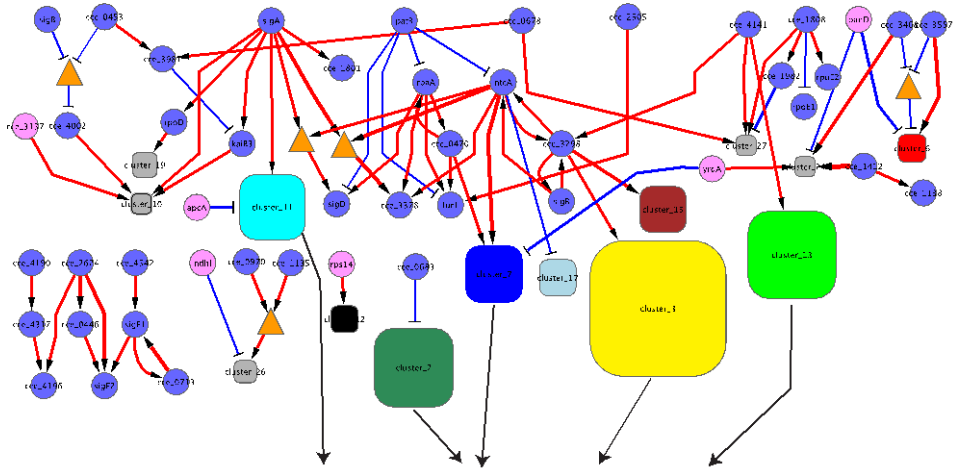
- What are the pathways active in condition A versus a control?
- What are the pathways active in condition A versus condition B?
- What are the pathways active in individual samples relative to other samples?
- What are the pathways active in a specific subset of genes (e.g. a particular cluster?)
- What pathways correlate over a range of samples with a measured variable (e.g. CO_2 production?)

Pathway Enrichment Tools

- DAVID (https://david.ncifcrf.gov/)
- GSEA (https://www.gsea-msigdb.org/gsea/index.jsp)
- Enrichr (https://maayanlab.cloud/Enrichr/)
- WebGESTALT (https://www.webgestalt.org/)
- leapR (https://github.com/PNNL-CompBio/leapR)

Network Analysis

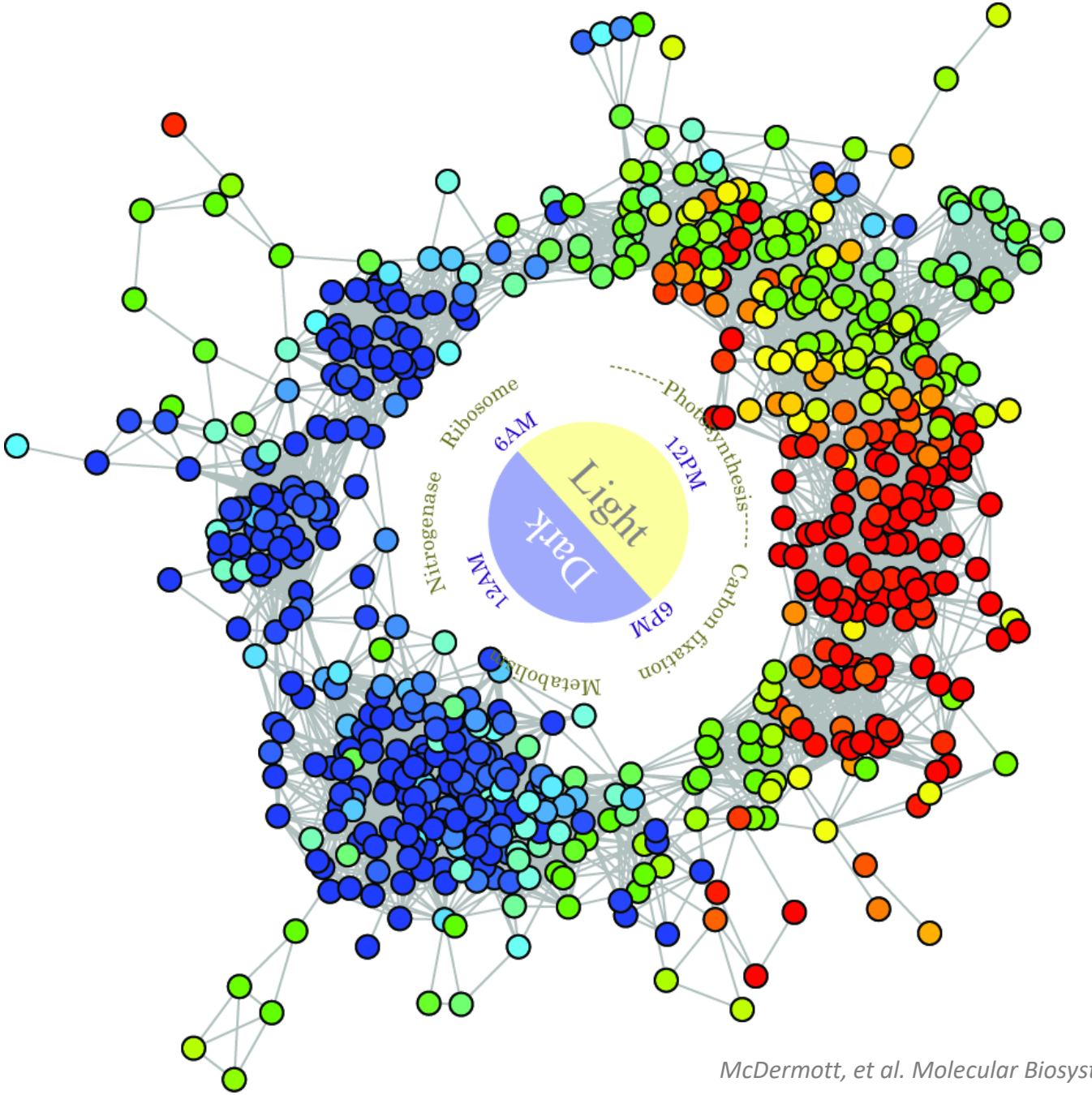
- Network analysis
  - Existing knowledge frameworks
  - Network inference
- How are proteins grouped?
  - Complexes
  - Pathways
  - Regulons
  - Functional response
- Topology
  - Critical nodes
  - Hubs



Label	Description	Pattern
cluster_13 (103):	apc, cpc, rbc, ccm, atp, glgA1, glgB1, glgB2	L1
cluster_6 (15):	psbA4	D5-D9
cluster_24 (4):	hupS	D1-D9
cluster_27 (4):	petF3, petF4, psb27	L1
cluster_3 (173):	PSII, ntcA, sigB, ftsH1, ftsH2, ndhEHI	L5
cluster_15 (42):	psbC, hypE, rpoE1, ftsW	L1-L5
cluster_17 (28):	feoA2, feoB2, hupL,	D1
cluster_7 (57):	PSI, rpaA, cmp, sigD	L5-L9
cluster_2 (90):	kaiB4, mazE, sigF1, sigF2	L1
cluster_12 (14):	ribosome, groES, groEL, dnaK, ho2, hemN	D5
cluster_26 (7):	unknown	D5-D9
cluster_11 (70):	kaiB3, rpoD, irtA, glnB, coxABC, hypA	L9-D1
cluster_10 (10):	glgP1	L9-D5
cluster_19 (11):	pnt, gnd	L9-D1

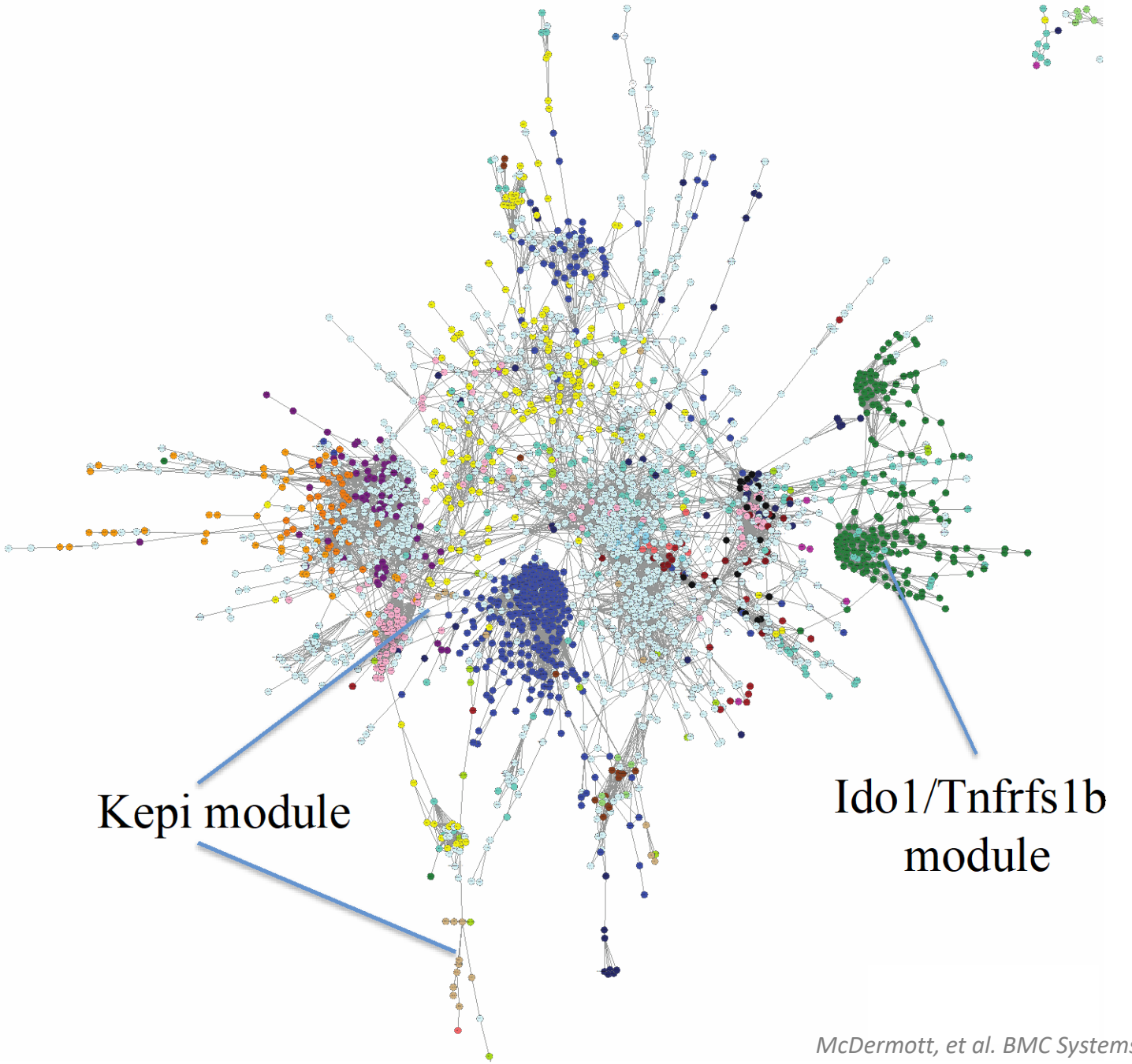
McDermott, et al. Molecular Biosystems 2011

Examples of Applications of Network Analysis



McDermott, et al. Molecular Biosystems 2011

Examples of Applications of Network Analysis



McDermott, et al. BMC Systems Biology 20186

Network Analysis Tools

- STRING (https://string-db.org/)
- KEGG (https://www.genome.jp/kegg/pathway.html)
- Cytoscape (https://cytoscape.org/)
- QIAGEN Ingenuity Pathway Analysis ($$$)

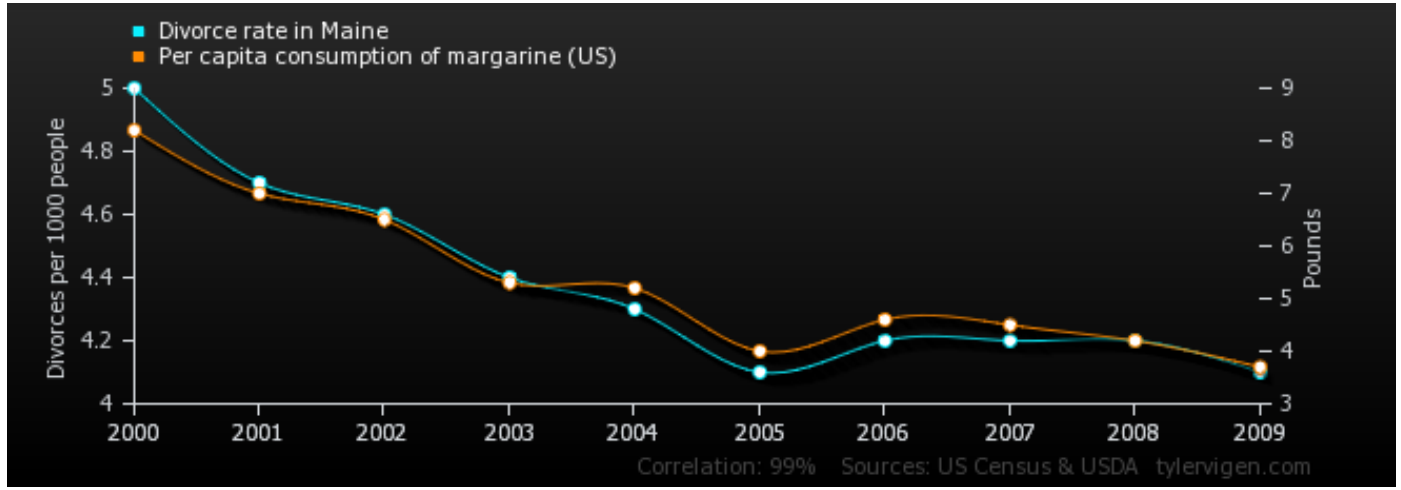
- ARACNE/CLR (http://califano.c2b2.columbia.edu/aracne)
- GENIE3 (https://bioconductor.org/packages/release/bioc/html/GENIE3.html)

Data Processing and Formats

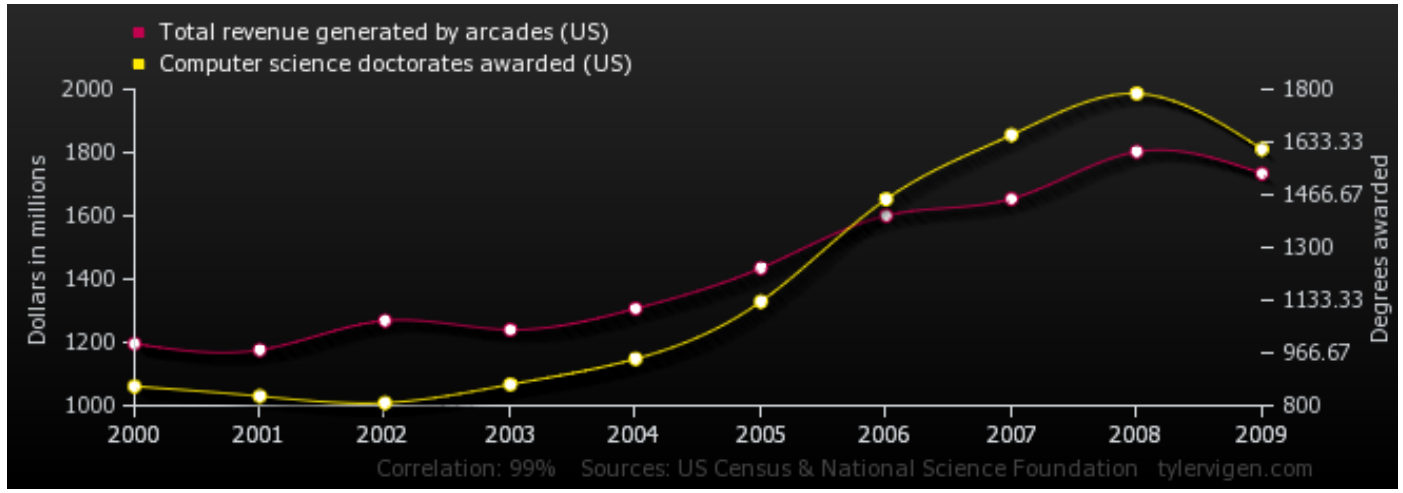
- **Identifiers**
  - **Identifiers**
  - **Identifiers!**
- Data matrix format
- Data normalization

Correlation is not always causation

Divorce rate in Maine correlates with US per capita consumption of margarine (R = 0.99)

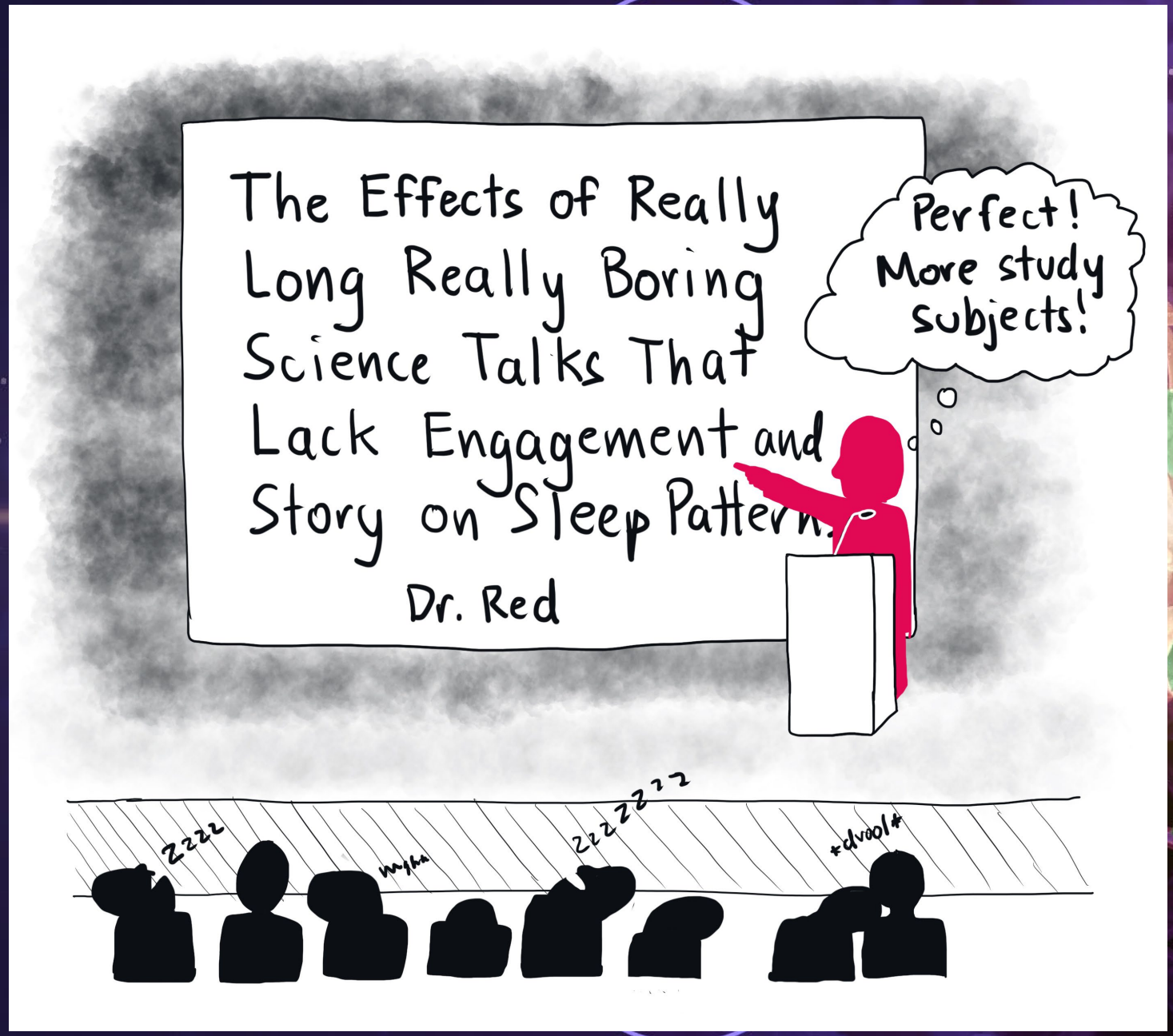


Revenue generated by arcades correlates with number of CS doctorates awarded (R = 0.99)



http://www.tylervigen.com/

Questions?


The Effects of Really Long Really Boring Science Talks That Lack Engagement and Story on Sleep Pattern.
Dr. Red
Perfect! More study subjects!
zzzz
zzzzzzz
drool

Afternoon Session

1:15-2:15 p.m.	Workshop: Formatting and Quality Control	
2:15-2:25	Networking Break	
2:25-3:15	Workshop: Statistics and Diagnostics in pmartR	
3:15-4:00	Workshop: Statistics Edge Cases	