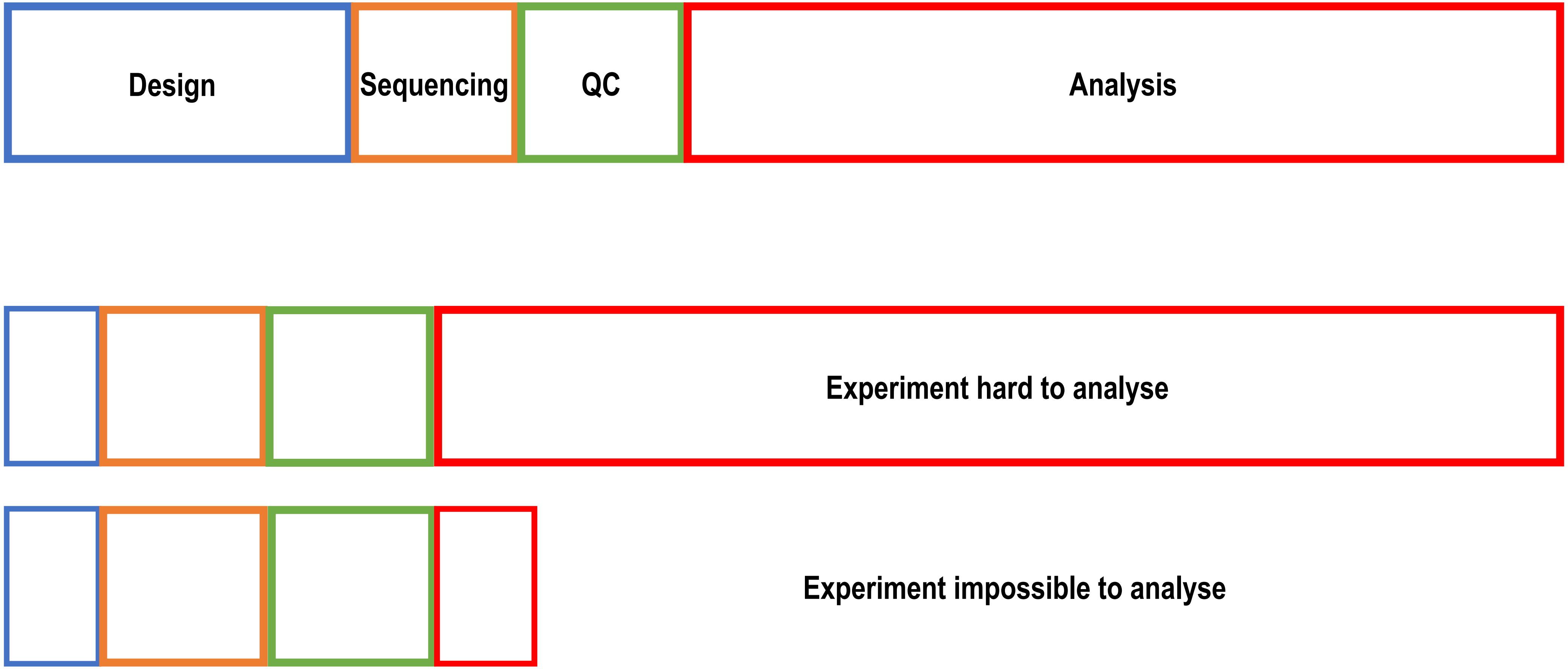
Guidelines for Experimental Design

General Considerations & Best Practices

Importance of experimental design



Design

Sequencing

QC

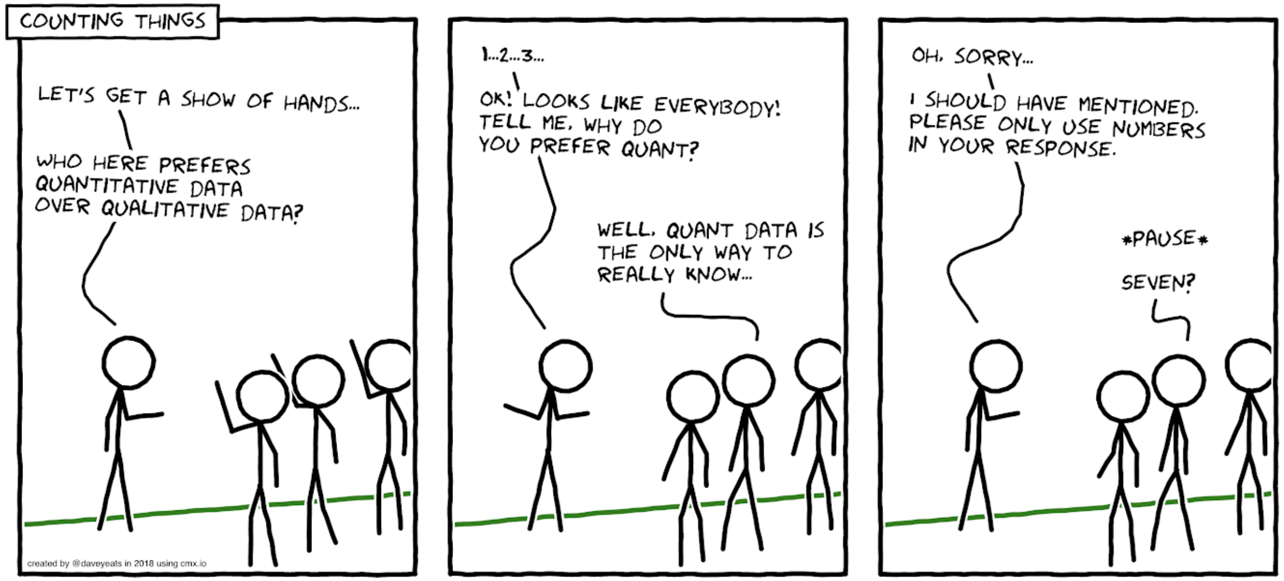
Analysis

Experiment hard to analyse

Experiment impossible to analyse

RNA-seq

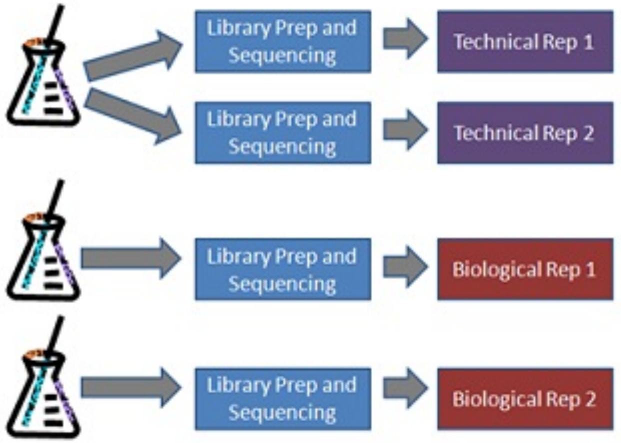
- Many types of RNA-seq experiment
- Experimental design depends on type
  - Quantitative: e.g. differential gene expression, alternative splicing
  - Qualitative: e.g. transcript discovery, identification of poly(A) sites
- Mostly focus on differential gene expression (DGE)


COUNTING THINGS
LET'S GET A SHOW OF HANDS...
WHO HERE PREFERS QUANTITATIVE DATA OVER QUALITATIVE DATA?
created by @daveyeats in 2018 using cmx.io
1...2...3...
OK! LOOKS LIKE EVERYBODY! TELL ME, WHY DO YOU PREFER QUANT?
WELL, QUANT DATA IS THE ONLY WAY TO REALLY KNOW...
OH, SORRY...
I SHOULD HAVE MENTIONED. PLEASE ONLY USE NUMBERS IN YOUR RESPONSE.
PAUSE
SEVEN?

By Dave Yeats using cmx.io

Replication

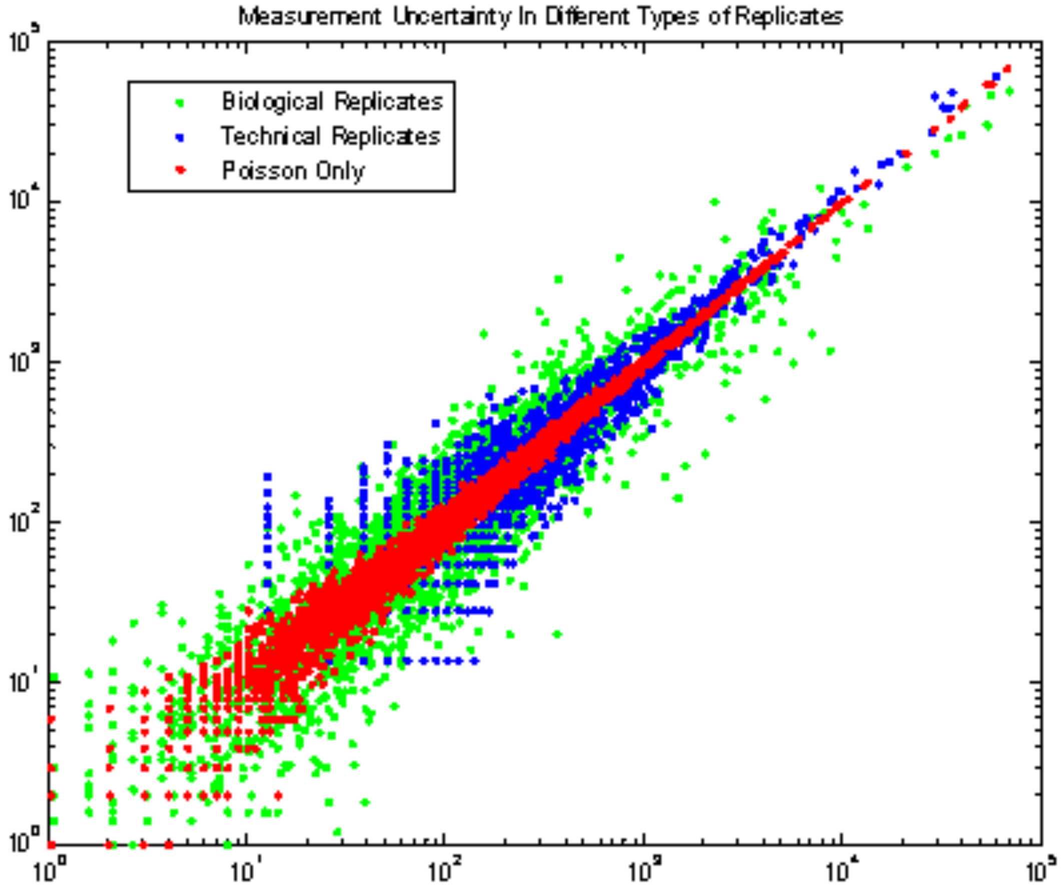
- Technical replicates:
  - Rarely needed (except during method development, where want to differentiate technical and biological variability)
  - Main source of technical variability is RNA prep and library prep, not sequencing
- Biological replicates:
  - Minimise or control for biological variability (so focus on conditions)
  - For example:
    - choose embryos from same clutch
    - or control for clutch in analysis


Library Prep and Sequencing
Technical Rep 1
Library Prep and Sequencing
Technical Rep 2
Library Prep and Sequencing
Biological Rep 1
Library Prep and Sequencing
Biological Rep 2

From http://scotty.genetics.utah.edu/help.html

Sources of variance

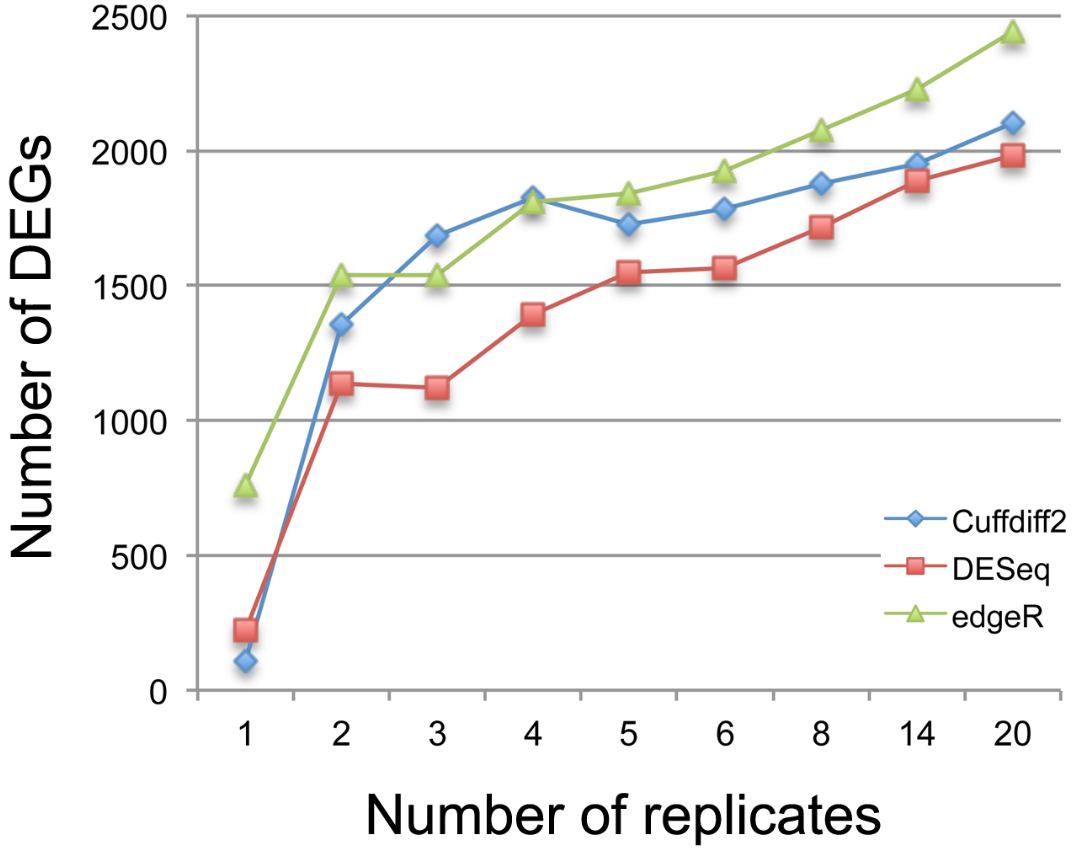
- Biological variance - natural variance
    - Zebrafish – lots of replicates, but if pool then variance reduced and can lose signal
- Technical variance - from RNA & library prep
- Poisson variance - counting noise; high variance at low counts



From Michele Busby

How many replicates? (1/2)

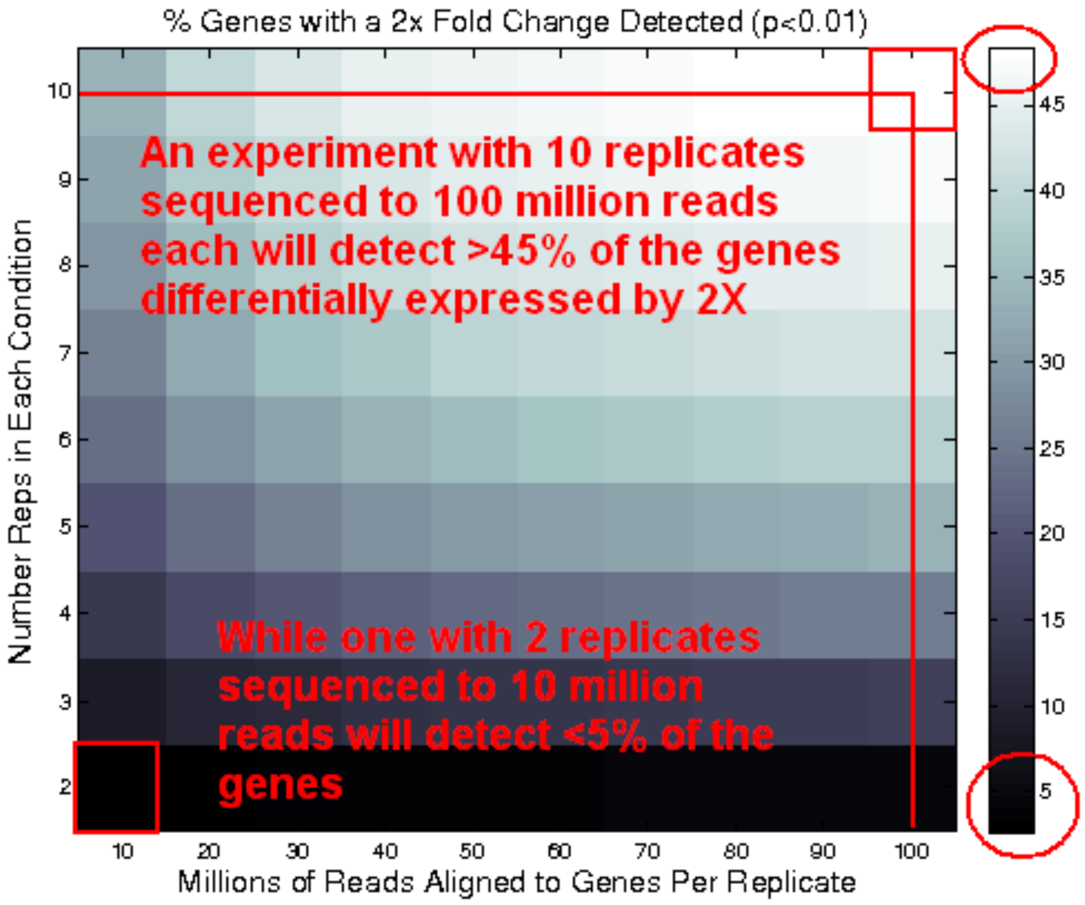
- Often trade off between number of samples and sequencing depth
- DGE:
  - More samples best (if cost allows), because reduces effect of biological variability
  - Can always sequence more deeply, but hard to add samples (batch effect)
  - Generally never < 4 samples per condition, but more better
  - We never do < 6 samples and often 12+
  - 10 million reads usually enough



From Zhang et al., 2014 – "A Comparative Study of Techniques for Differential Expression Analysis on RNA-Seq Data"

How many replicates? (2/2)

- Transcript discovery:
  - Sequencing depth important (want overlapping reads over whole transcript)
  - Enrichment for desired transcripts, e.g. by size selection
  - Range of tissues, developmental stages or treatments
- http://scotty.genetics.utah.edu/ - helps design experiment
  (requires similar or pilot data, plus costs)

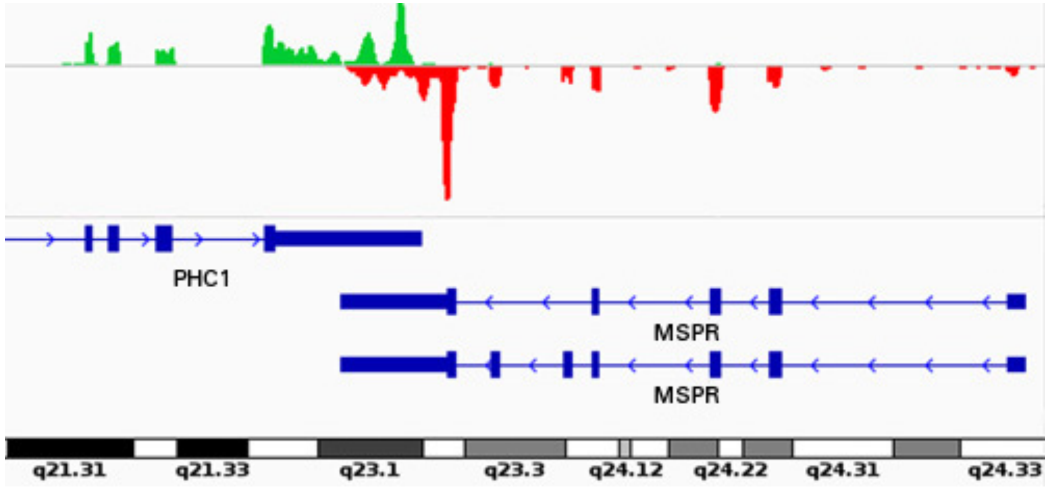


An experiment with 10 replicates sequenced to 100 million reads each will detect >45% of the genes differentially expressed by 2X

While one with 2 replicates sequenced to 10 million reads will detect <5% of the genes

What type of reads?

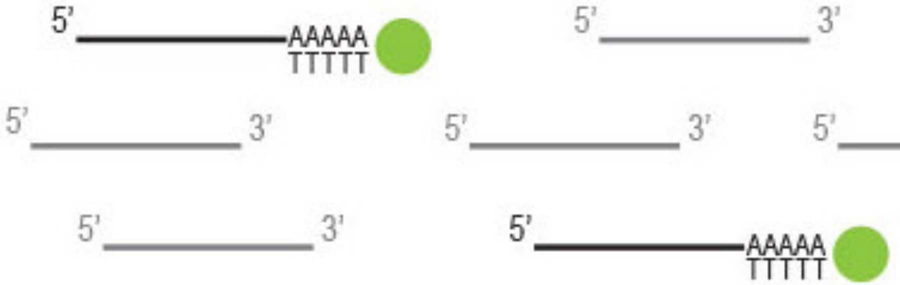
- For qualitative experiments, want:
  - Stranded library
  - Long reads (100 bp +)
  - Paired end reads
- Not important for quantitative experiments
  - 75 bp probably optimal for DGE (Chhangawala et al., 2015 – “The impact of read length on quantification of differentially expressed genes and splice junction detection”)



From Clontech

Ribosomal RNA

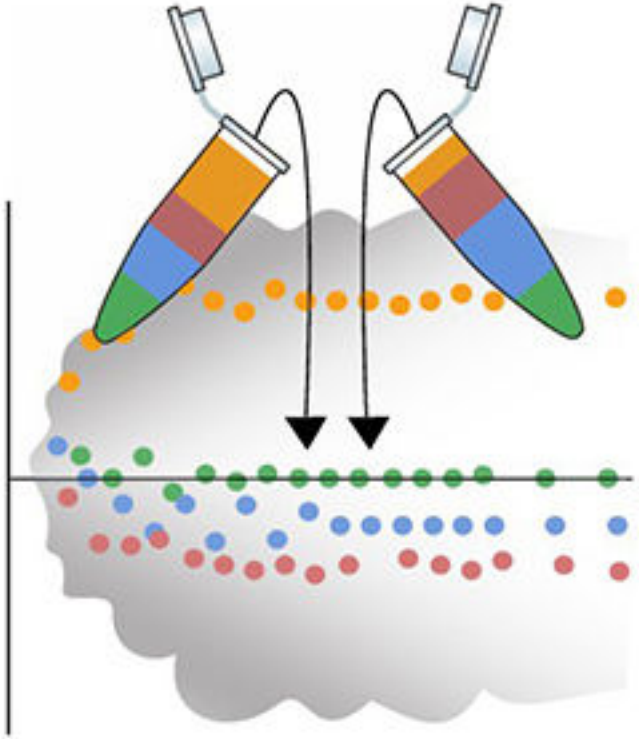
- Usually want to sequence mRNA, but total RNA is mostly rRNA
- Either enrich for mRNA or deplete rRNA
- mRNA enrichment by oligo (dT):
  - Cheaper and less noisy, but leads to 3’ bias and ignores some ncRNAs
- rRNA depletion by Ribo-Zero:
  - Expensive and doesn’t work as well with zebrafish as for other model organisms (designed for human, mouse and rat)



From Lexogen

RNA spike-ins

- ERCC spike-ins – set of transcripts of various lengths and concentrations
- Suggested to aid normalisation
- But are expensive and don't actually improve normalisation



From https://www.nist.gov/programs-projects/external-rna-controls-consortium

Batch effects

- Batch effects are technical variation between groups of samples
- RNA prep and library prep are very sensitive to batch effects
- Make sure all samples are prepared in the same way as far as possible
    - e.g. all samples prepared by same person at same time using same reagents
- Otherwise control for batch in analysis
    - But requires more samples to maintain power

Controlling for batch

Sample	Genotype
sample_1	wild_type
sample_2	wild_type
sample_3	wild_type
sample_4	wild_type
sample_5	knockout
sample_6	knockout
sample_7	knockout
sample_8	knockout

Sample	Genotype	Batch
sample_1	wild_type	Friday
sample_2	wild_type	Friday
sample_3	wild_type	Monday
sample_4	wild_type	Monday
sample_5	knockout	Friday
sample_6	knockout	Friday
sample_7	knockout	Monday
sample_8	knockout	Monday

Sample	Genotype	Batch
sample_1	wild_type	Friday
sample_2	wild_type	Friday
sample_3	wild_type	Friday
sample_4	wild_type	Monday
sample_5	wild_type	Monday
sample_6	wild_type	Monday
sample_7	knockout	Friday
sample_8	knockout	Friday
sample_9	knockout	Friday
sample_10	knockout	Monday
sample_11	knockout	Monday
sample_12	knockout	Monday

Confounding

- Don’t confound batch with conditions – otherwise analysis impossible
  - Best to randomise samples, so batches evenly distributed across conditions

Sample	Genotype	Clutch
sample_1	wild_type	clutch_1
sample_2	wild_type	clutch_1
sample_3	wild_type	clutch_1
sample_4	wild_type	clutch_1
sample_5	knockout	clutch_2
sample_6	knockout	clutch_2
sample_7	knockout	clutch_2
sample_8	knockout	clutch_2

Confounded

Sample	Genotype	Clutch
sample_1	wild_type	clutch_1
sample_2	wild_type	clutch_2
sample_3	wild_type	clutch_2
sample_4	wild_type	clutch_1
sample_5	knockout	clutch_1
sample_6	knockout	clutch_1
sample_7	knockout	clutch_2
sample_8	knockout	clutch_2

Not confounded

Clutch batch effect

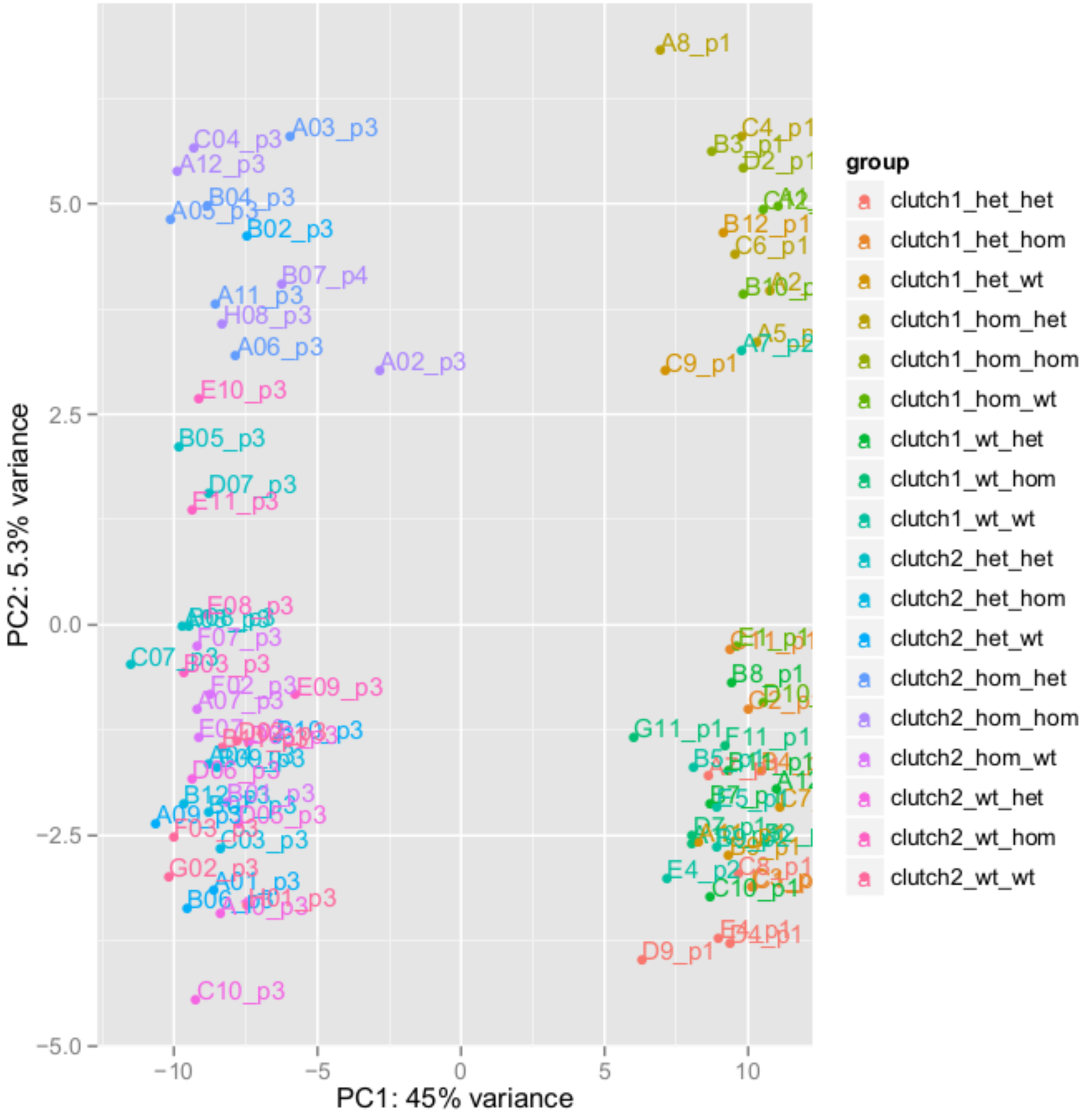
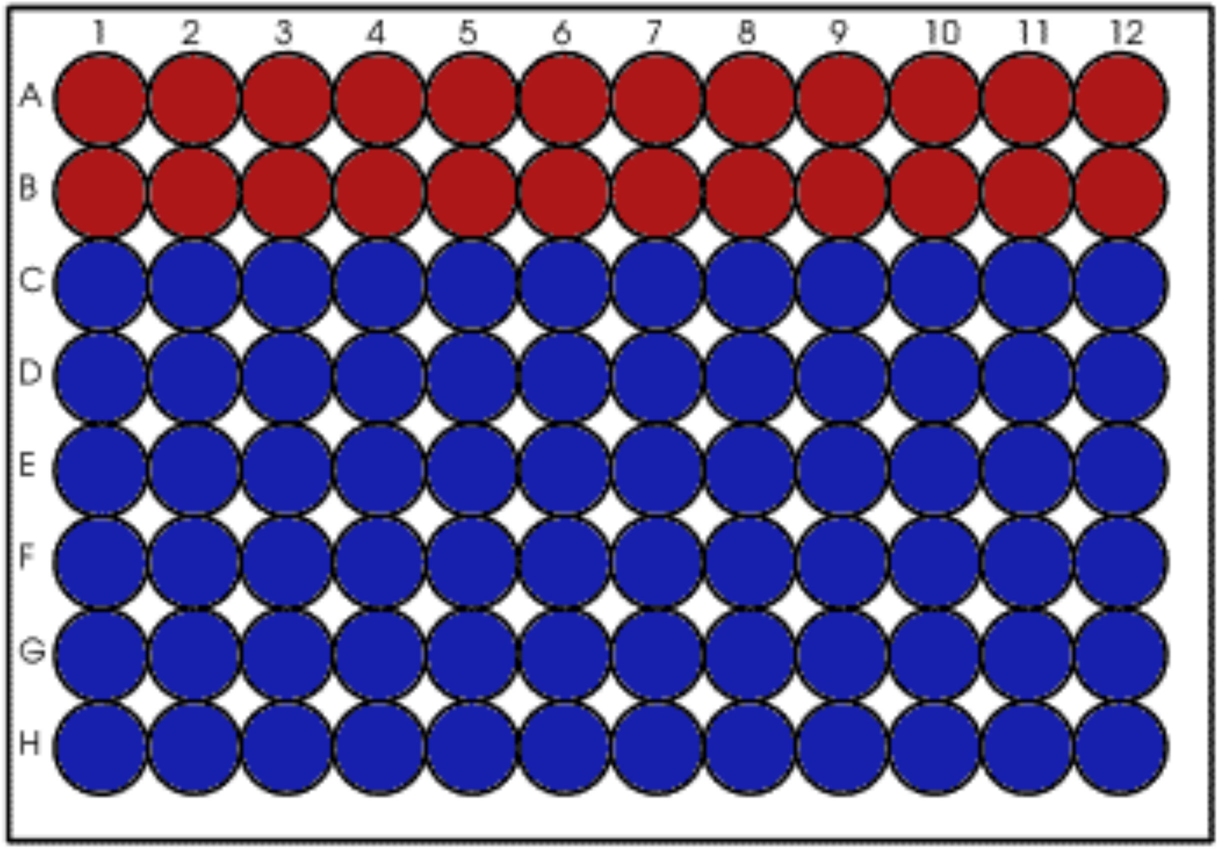


Plate effect

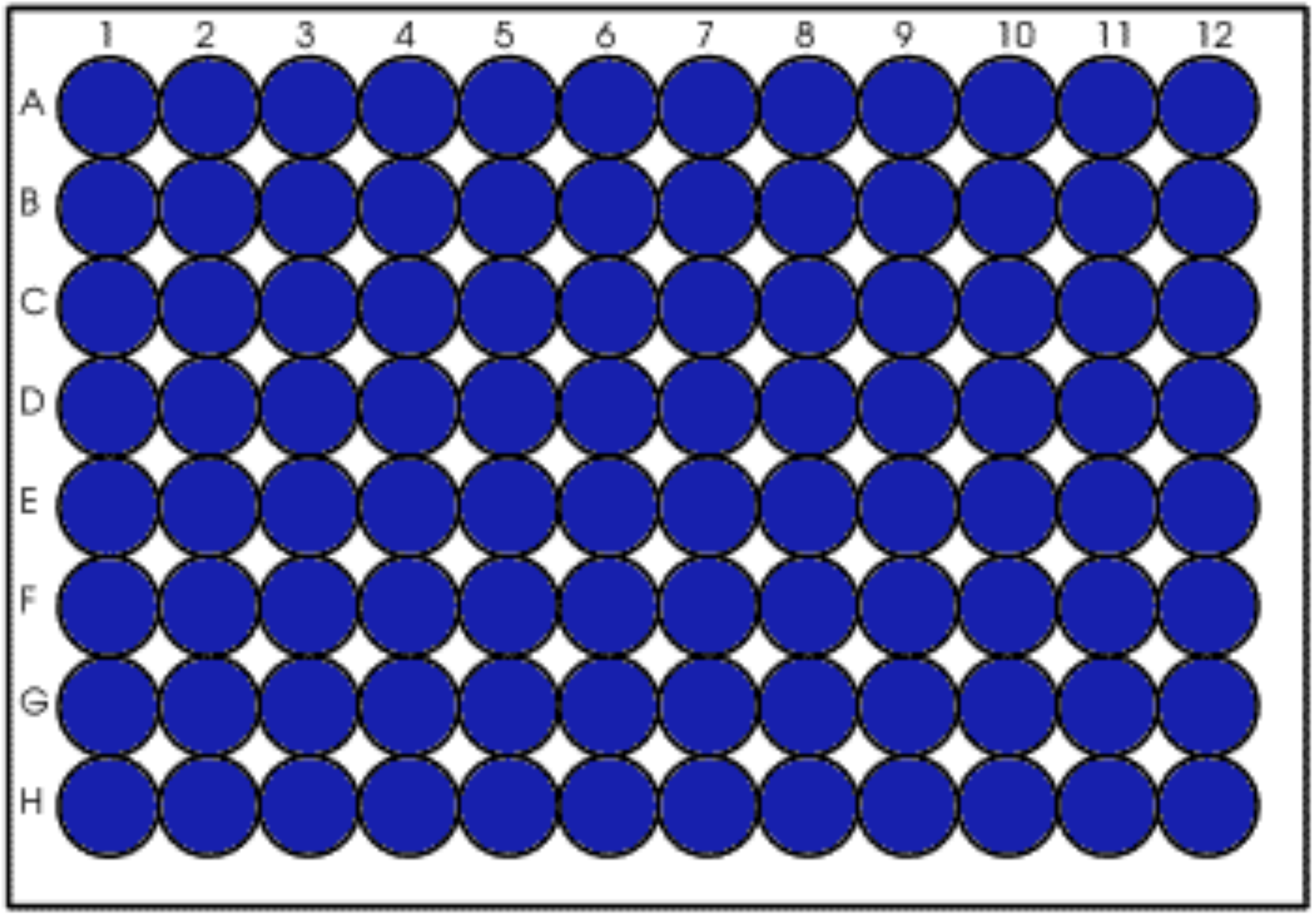


Phenotypic

Non-phenotypic

Rows	DE regions
A vs B	8
C vs D	15
D vs E	268
E vs F	692
G vs H	374
A vs H	3372
Random 12 vs 12	0
Col 1 vs Col 12	0

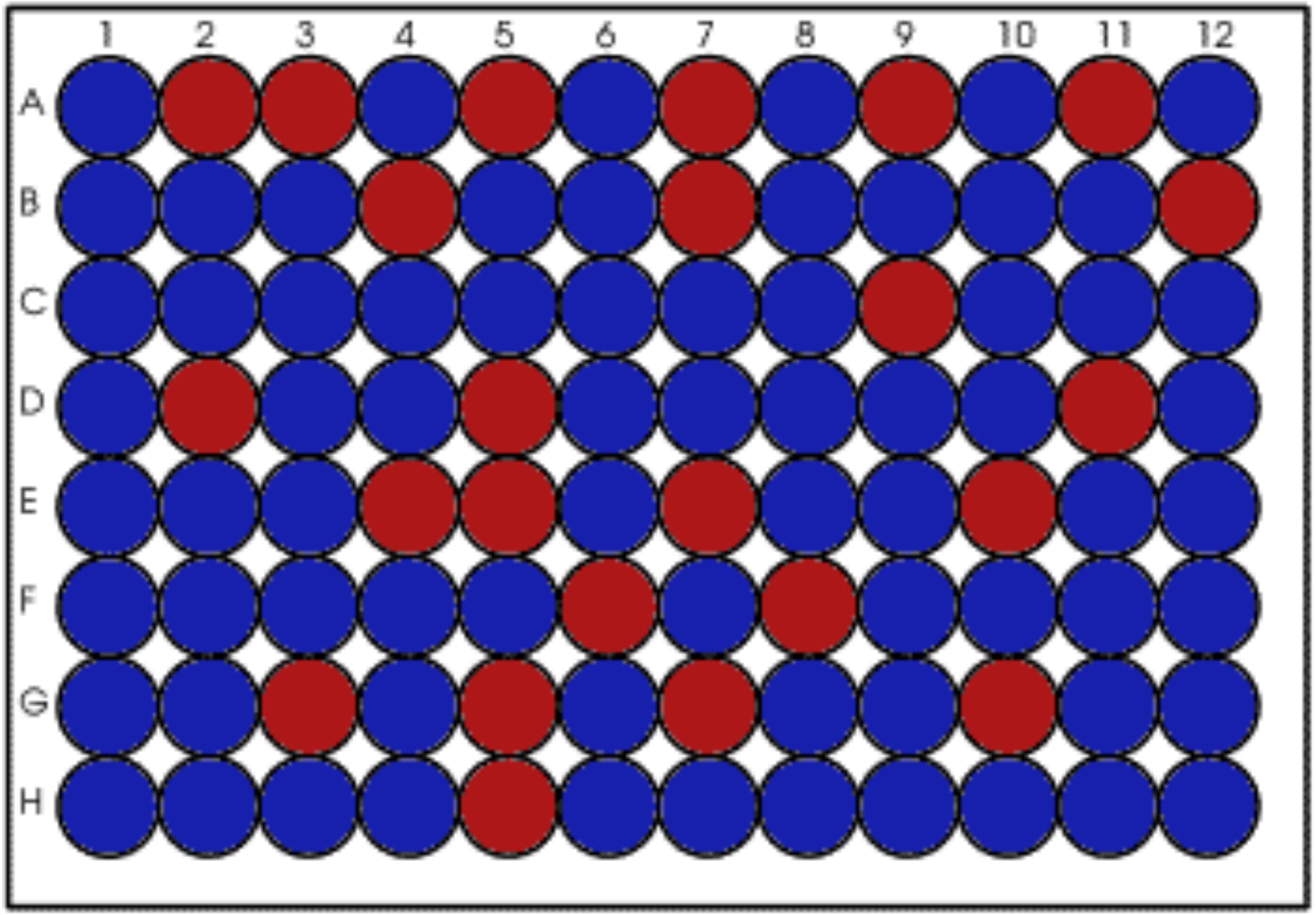
Plate effect confirmation



- 96 wild-type embryos
- RNA extracted in rows, but libraries made in columns

Columns	DE regions
1 vs 2	78
1 vs 3	749
2 vs 3	225

Better plate design



	1	2	3	4	5	6	7	8	9	10	11	12
A	🔵	🔴	🔴	🔵	🔴	🔵	🔴	🔵	🔴	🔵	🔴	🔵
B	🔵	🔵	🔵	🔴	🔵	🔵	🔴	🔵	🔵	🔵	🔵	🔴
C	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔴	🔵	🔵	🔵
D	🔵	🔴	🔵	🔵	🔴	🔵	🔵	🔵	🔵	🔵	🔴	🔵
E	🔵	🔵	🔵	🔴	🔴	🔵	🔴	🔵	🔵	🔴	🔵	🔵
F	🔵	🔵	🔵	🔵	🔵	🔴	🔵	🔴	🔵	🔵	🔵	🔵
G	🔵	🔵	🔴	🔵	🔴	🔵	🔴	🔵	🔵	🔴	🔵	🔵
H	🔵	🔵	🔵	🔵	🔴	🔵	🔵	🔵	🔵	🔵	🔵	🔵

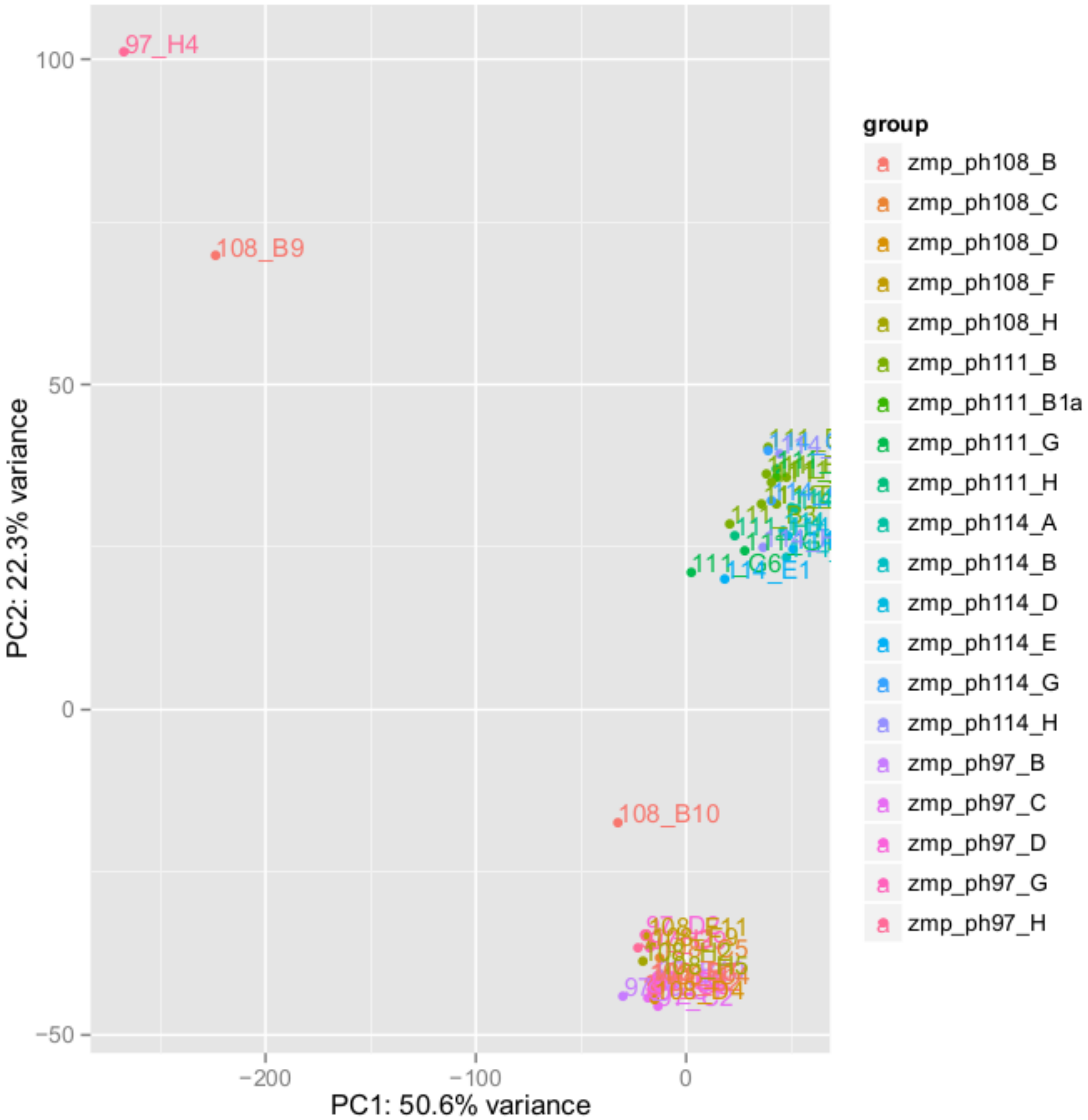
Multiplexing

- Sequencing is quite consistent, but still best to pool samples and sequence across multiple lanes
  - Reason why difficult to add more samples to an experiment
- Multiplexed libraries need to be balanced to ensure even read depth
- Can check with MiSeq run
- We prefer to exclude outliers (low read depth)
  - Another reason to have lots of samples

From Illumina

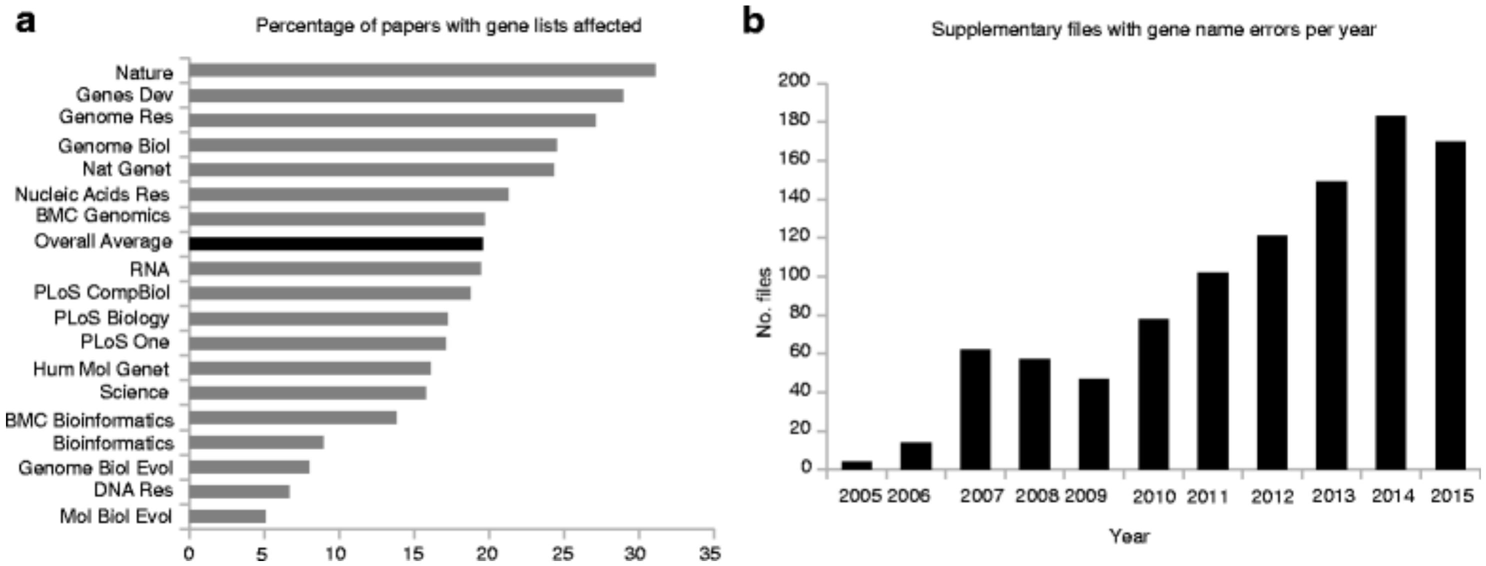
Visualisation

- Important to visualise your data at each stage of analysis
- e.g. PCA to identify outliers



Best practices (1/2)

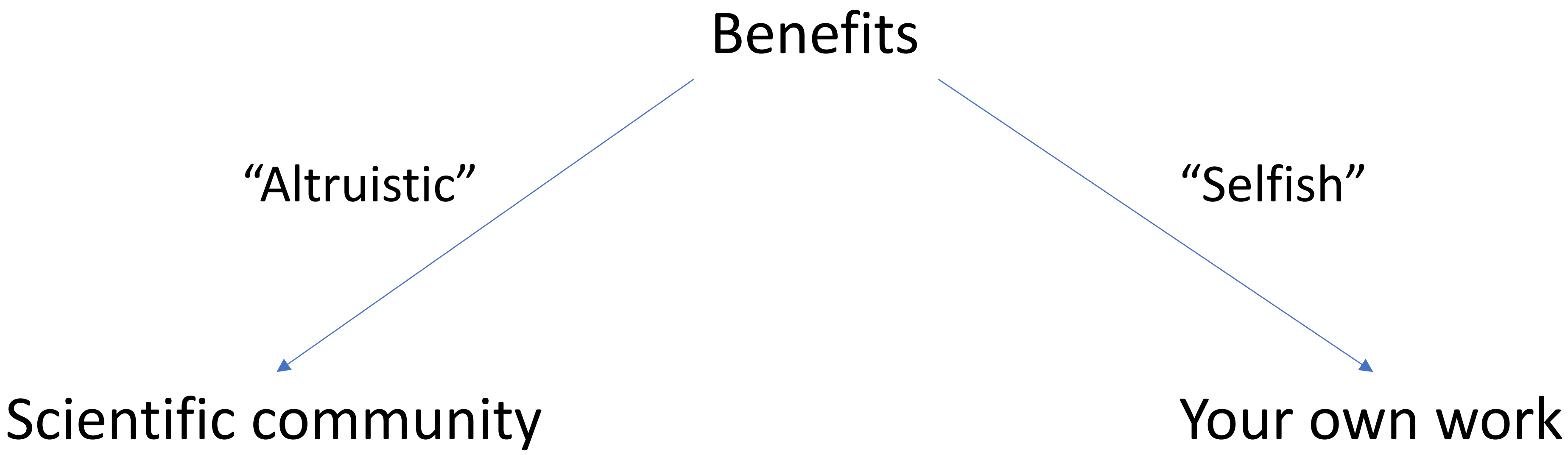
- Avoid Excel for analysis
  - Fine for exploring data, but don’t export data from Excel
  - Ziemann et al., 2016 – “Gene name errors are widespread in the scientific literature”
  - e.g. sept2 converted to 2-Sep (human gene now renamed to SEPTIN2)



Best practices (2/2)

- Don’t (subconsciously) cherrypick data
  - Conclusions should be robust and not rely on filtering data in an arbitrary way
  - e.g. can’t take a list of lipid genes and just assess those for differential expression
- Write down everything you do
  - Future you will thank you when you analyse your data and try to discover the reason for an unexpected batch effect
  - Sequence deposition requires good metadata

Data sharing



Benefits

“Altruistic”

“Selfish”

Scientific community

Your own work

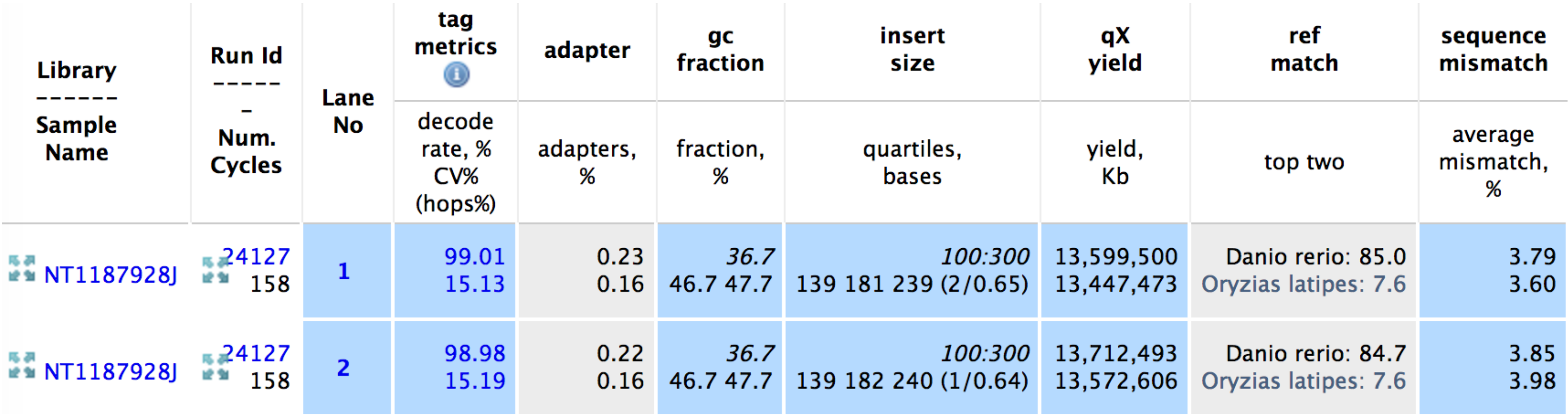
Altruistic reasons for data sharing

- Contribute to databases we use on a daily basis (e.g. Ensembl, ZFIN, GO, etc…)
- Reduce duplication of effort (Reviewer 2: “Comparison to ChIP-seq data is necessary to...”)
- Enable more discovery (other people have completely different questions; data reuse statement)
- Gives non-bioinformaticians access to NGS data

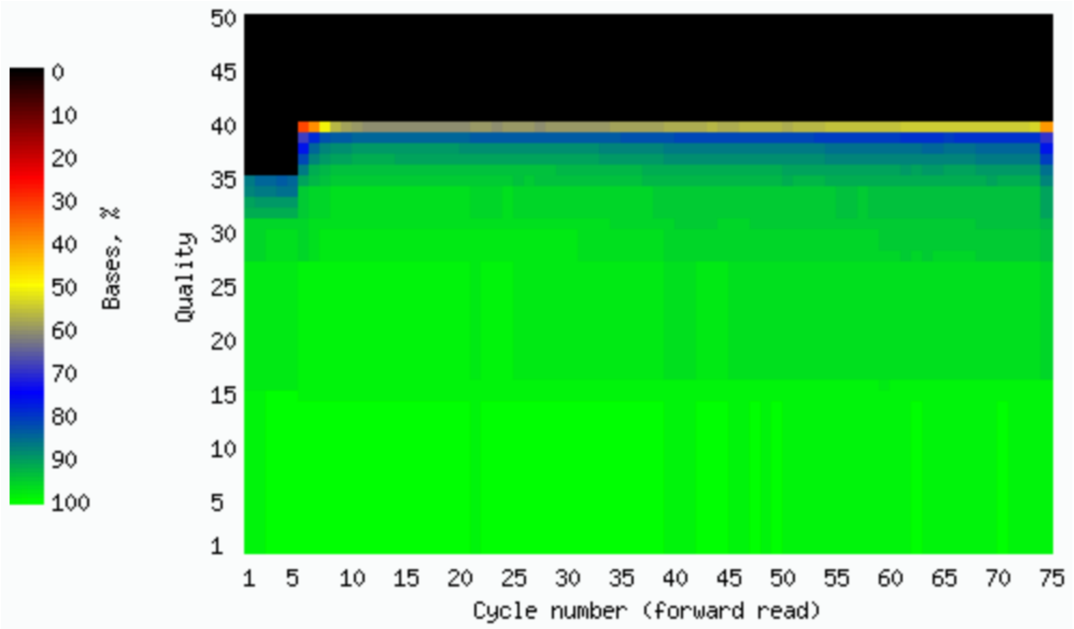
Selfish reasons for data sharing

- Encourages comprehensive metadata documentation
- Easy data access for you and for others (=> citations)
- Data access mandatory for most funders and journals
- Appreciated by reviewers (“there is tremendous utility for researchers for fully processed, discrete, clear and unambiguous annotated DE gene lists”)
- Raises awareness of your work outside your own field
- Good for your reputation – “they know what they are doing”

Analysis – In-house sequencing QC

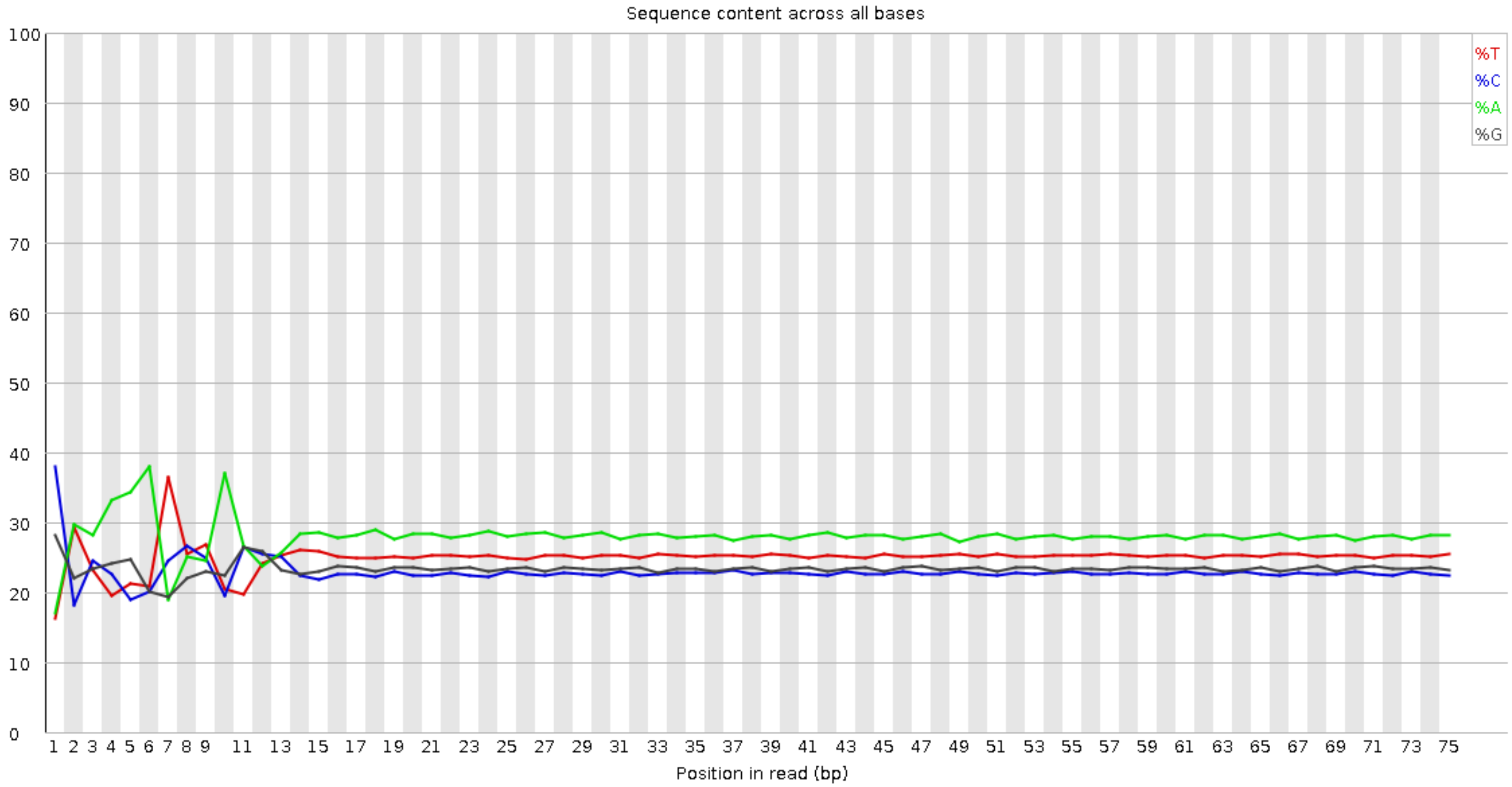


Library ------ Sample Name	Run Id ----- - Num. Cycles	Lane No	tag metrics	adapter	gc fraction	insert size	qX yield	ref match	sequence mismatch
			decode rate, % CV% (hops%)	adapters, %	fraction, %	quartiles, bases	yield, Kb	top two	average mismatch, %
NT1187928J	24127 158	1	99.01 15.13	0.23 0.16	*36.7* 46.7 47.7	*100:300* 139 181 239 (2/0.65)	13,599,500 13,447,473	Danio rerio: 85.0 Oryzias latipes: 7.6	3.79 3.60
NT1187928J	24127 158	2	98.98 15.19	0.22 0.16	*36.7* 46.7 47.7	*100:300* 139 182 240 (1/0.64)	13,712,493 13,572,606	Danio rerio: 84.7 Oryzias latipes: 7.6	3.85 3.98



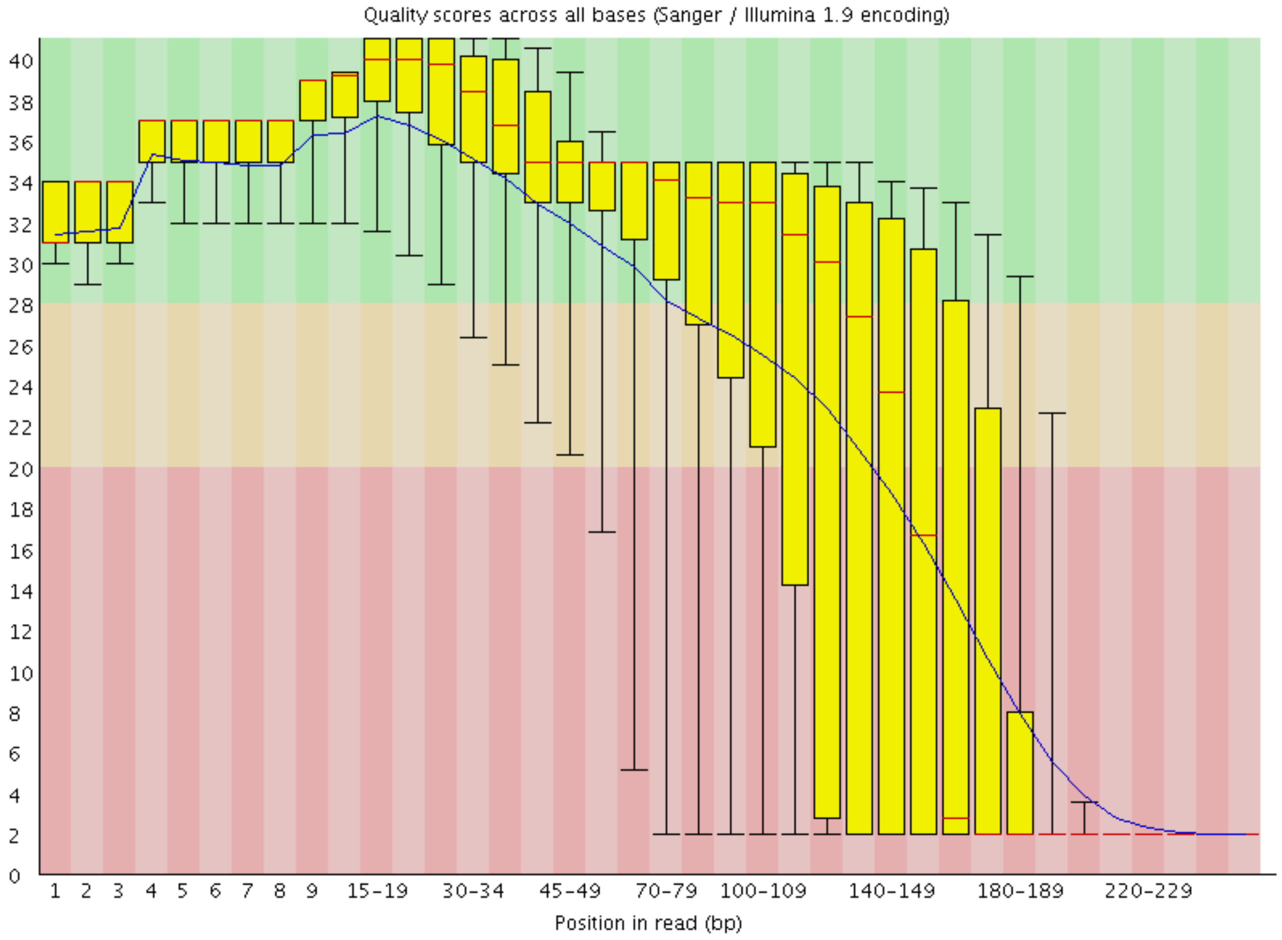
Analysis – FastQC (+ multiqc)

- Sequence quality
- Sequence content
- GC content
- N content
- Duplication
- Overrepresentation
- Adapter content



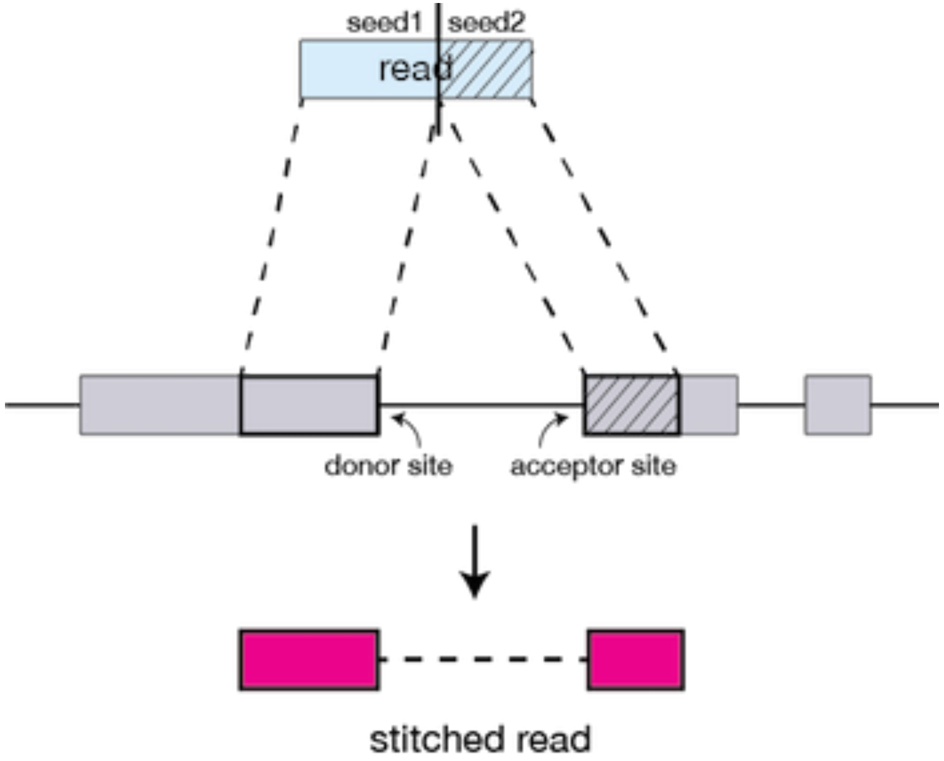
Analysis – Improving read quality

- Trim low quality bases
- Remove adapters
- Error correction
- e.g. Trim Galore! (cutadapt wrapper)



Analysis – Alignment

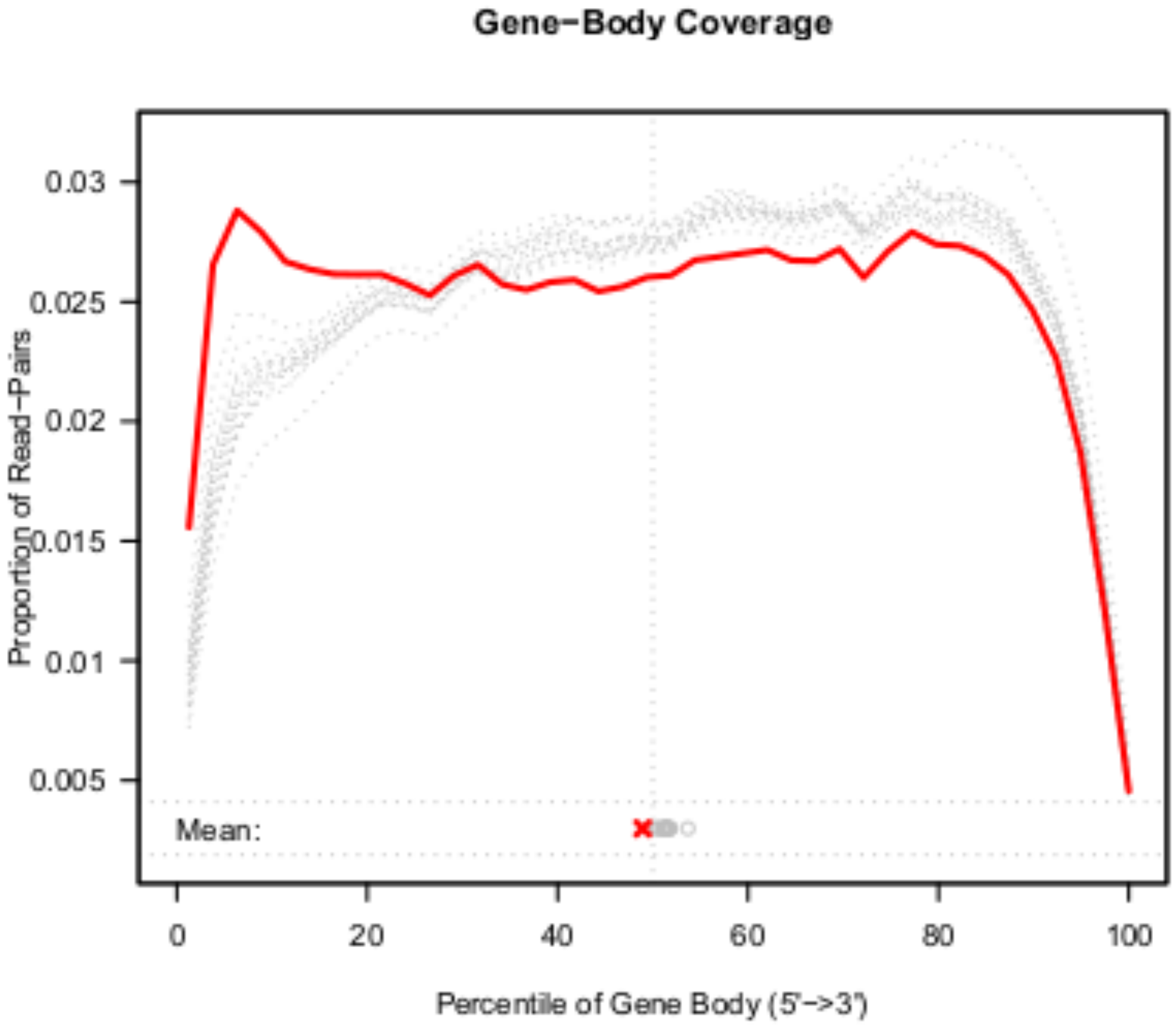
- Good zebrafish reference genome
  - Splice-aware aligner
  - Annotation optional
  - e.g. TopHat2, HISAT2, STAR
- Good zebrafish transcriptome
  - Pseudoalignment
  - Rapid
  - e.g. Salmon, kallisto



From https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

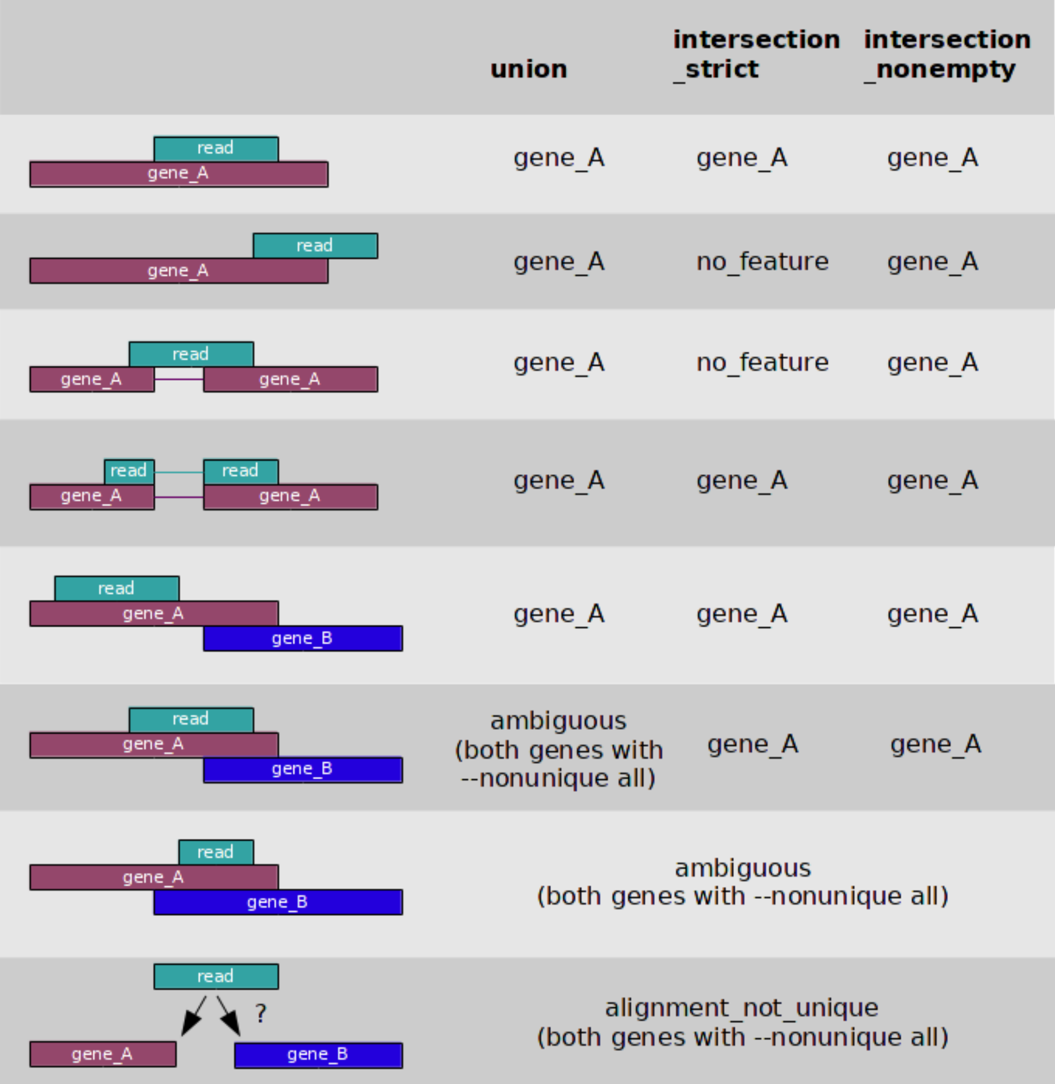
Analysis – Alignment QC

- QoRTs (Quality of RNA-seq Tool-Set)



Analysis – Quantification

- e.g. htseq-count, STAR



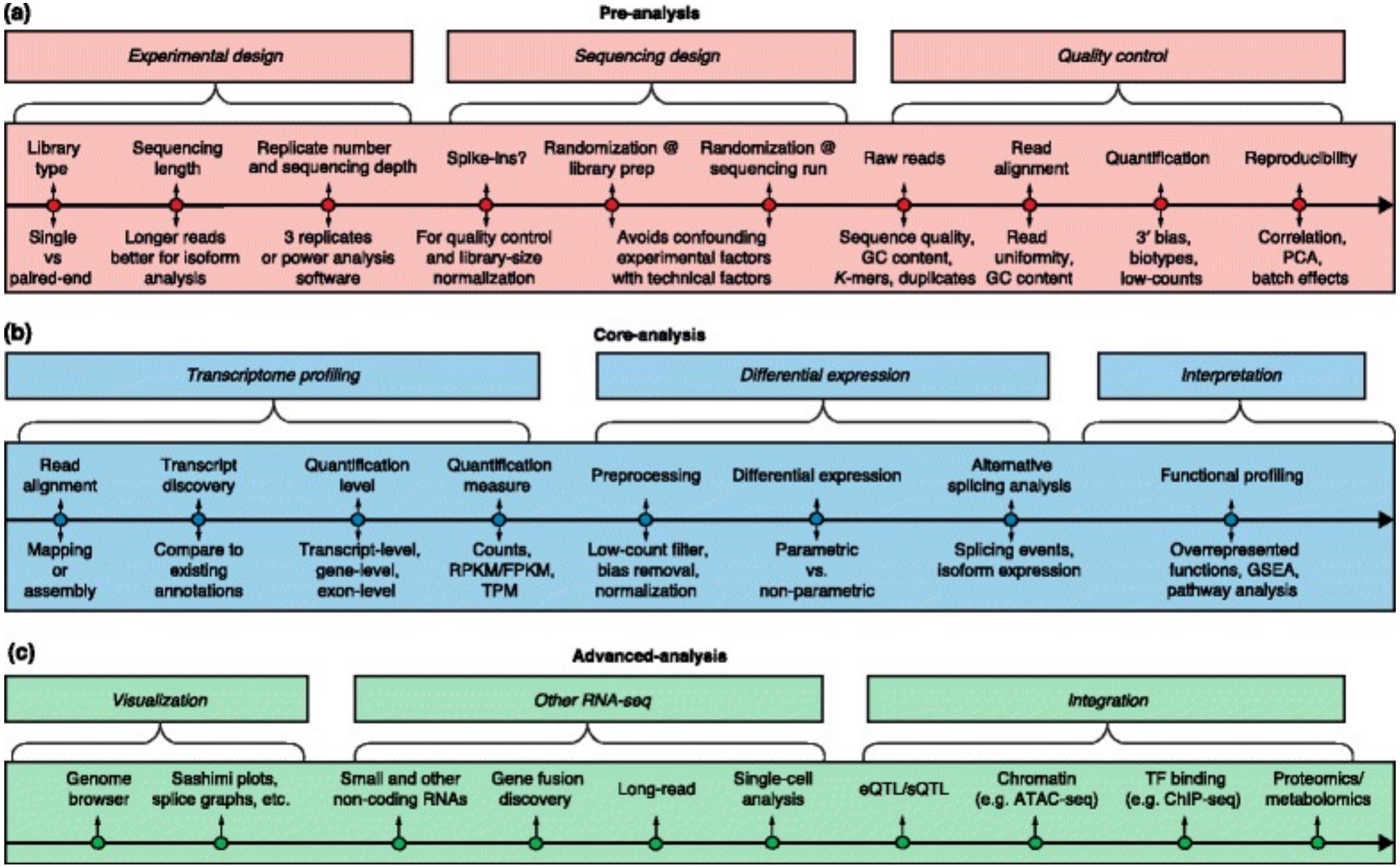
	union	intersection_strict	intersection_nonempty
read gene_A	gene_A	gene_A	gene_A
read gene_A	gene_A	no_feature	gene_A
read gene_A gene_A	gene_A	no_feature	gene_A
read read gene_A gene_A	gene_A	gene_A	gene_A
read gene_A gene_B	gene_A	gene_A	gene_A
read gene_A gene_B	ambiguous (both genes with --nonunique all)	gene_A	gene_A
read gene_A gene_B	ambiguous (both genes with --nonunique all)		
read ? gene_A gene_B	alignment_not_unique (both genes with --nonunique all)		

Analysis – Differential Expression

- e.g. DESeq2, edgeR

Gene	p-value	Adjusted p-value	Log$_2$ fold change
ENSDARG00000068969	5.13E-16	9.95E-13	4.296634713
ENSDARG00000071662	2.31E-25	8.20E-22	5.367426329
ENSDARG00000031885	2.60E-23	7.93E-20	5.248888274
ENSDARG00000043196	7.32E-08	7.80E-05	-3.715117121
ENSDARG00000075524	3.91E-15	6.94E-12	4.639355983
ENSDARG00000036787	1.22E-26	6.51E-23	4.384183256
ENSDARG00000079347	5.05E-08	5.67E-05	-2.564399561
ENSDARG00000041381	4.07E-09	5.11E-06	3.220579557
ENSDARG00000070062	3.49E-14	4.97E-11	4.454100519

Conclusion


(a)
Pre-analysis
Experimental design
Sequencing design
Quality control
Library type
Single vs paired-end
Sequencing length
Longer reads better for isoform analysis
Replicate number and sequencing depth
3 replicates or power analysis software
Spike-ins?
For quality control and library-size normalization
Randomization @ library prep
Randomization @ sequencing run
Avoids confounding experimental factors with technical factors
Raw reads
Sequence quality, GC content, K-mers, duplicates
Read alignment
Read uniformity, GC content
Quantification
3′ bias, biotypes, low-counts
Reproducibility
Correlation, PCA, batch effects
(b)
Core-analysis
Transcriptome profiling
Differential expression
Interpretation
Read alignment
Mapping or assembly
Transcript discovery
Compare to existing annotations
Quantification level
Transcript-level, gene-level, exon-level
Quantification measure
Counts, RPKM/FPKM, TPM
Preprocessing
Low-count filter, bias removal, normalization
Differential expression
Parametric vs. non-parametric
Alternative splicing analysis
Splicing events, isoform expression
Functional profiling
Overrepresented functions, GSEA, pathway analysis
(c)
Advanced-analysis
Visualization
Other RNA-seq
Integration
Genome browser
Sashimi plots, splice graphs, etc.
Small and other non-coding RNAs
Gene fusion discovery
Long-read
Single-cell analysis
eQTL/sQTL
Chromatin (e.g. ATAC-seq)
TF binding (e.g. ChIP-seq)
Proteomics/ metabolomics

From Conesa et al., 2016 – “A survey of best practices for RNA-seq data analysis”

Thank You

Any Questions?