INTRODUCTION TO TRANSCRIPTOMICS

PART I

John P. McCrow, Ph.D.
Computational Staff Scientist
J. Craig Venter Institute, La Jolla, CA

Bioinformatics User Group (BUG) seminar – May 4, 2017

Transcriptomics Workshop (part II in the Fall)

Part I (Today) will be an informal discussion of transcriptomics and what would be the most helpful and interesting things to cover in more depth in a later 2-day workshop.

Thank you for organizing:

- Jessica Blanton - SIO
- Sara Rivera - SIO
- Tessa Pierce - SIO
- Lisa Komoroske - NOAA

- Timmy at the Science March SD, Apr. 22, 2017

J. Craig Venter®
INSTITUTE

Outline

- Transcriptomics what/why/how
- Considerations when designing an experiment
- Quality control of data
- Read mapping
- Differential expression
- Multi-factor experimental design
- Meta-transcriptomics
- Fall workshop ideas

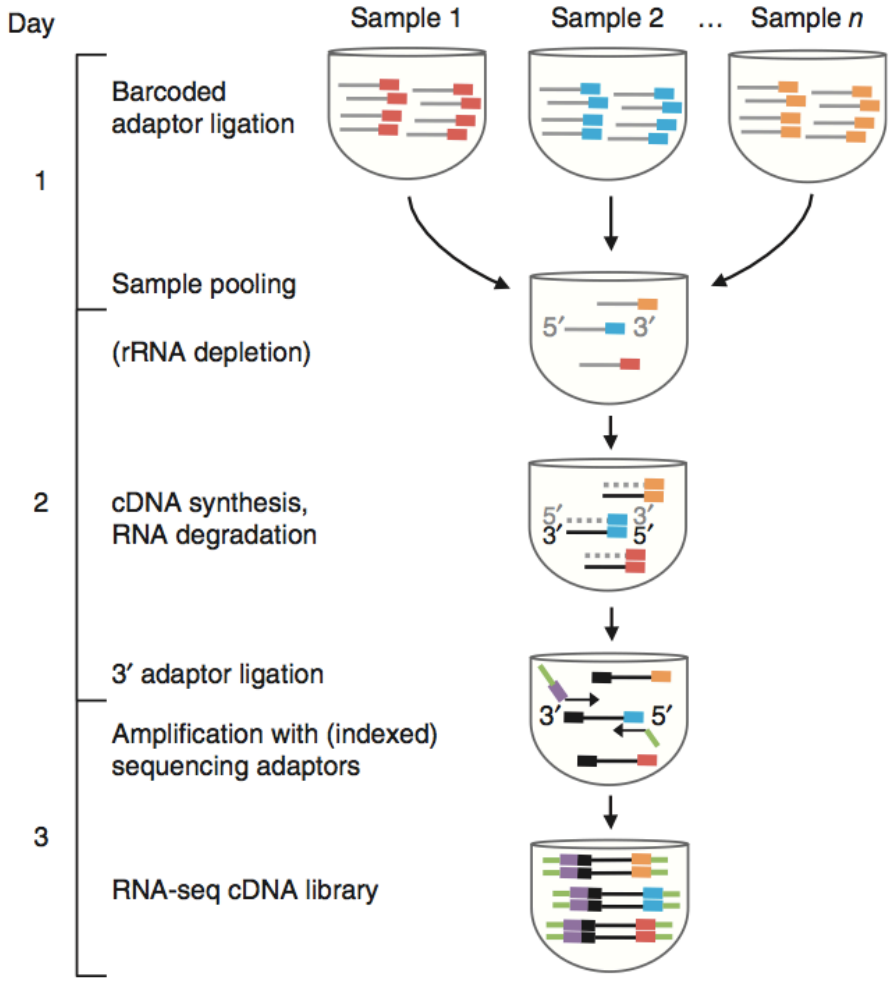
Transcriptomics

- Sequencing of cellular mRNA transcripts (cDNA)
- Transcriptomics is the most direct measure of functional response to changing environmental conditions, or cell types, or populations, etc.
- How does this compare to other measures: proteomics, genomics, even inference purely from ssu-rRNA?
- Non-omics approaches
  - RT-qPCR, RNase protection assay, Northern blot
- Why use an -omics approach?
  - If you want to know about many genes, or don't know which genes are present, or are interesting
  - Genomes are not comprised of just single-copy single-function proteins, especially in more complex euks
  - Even when measurement of a short list of genes is desired, there is value in knowing the context of other genes.
  - Good approach: Exploration and observation using transcriptomics, and validation using PCR

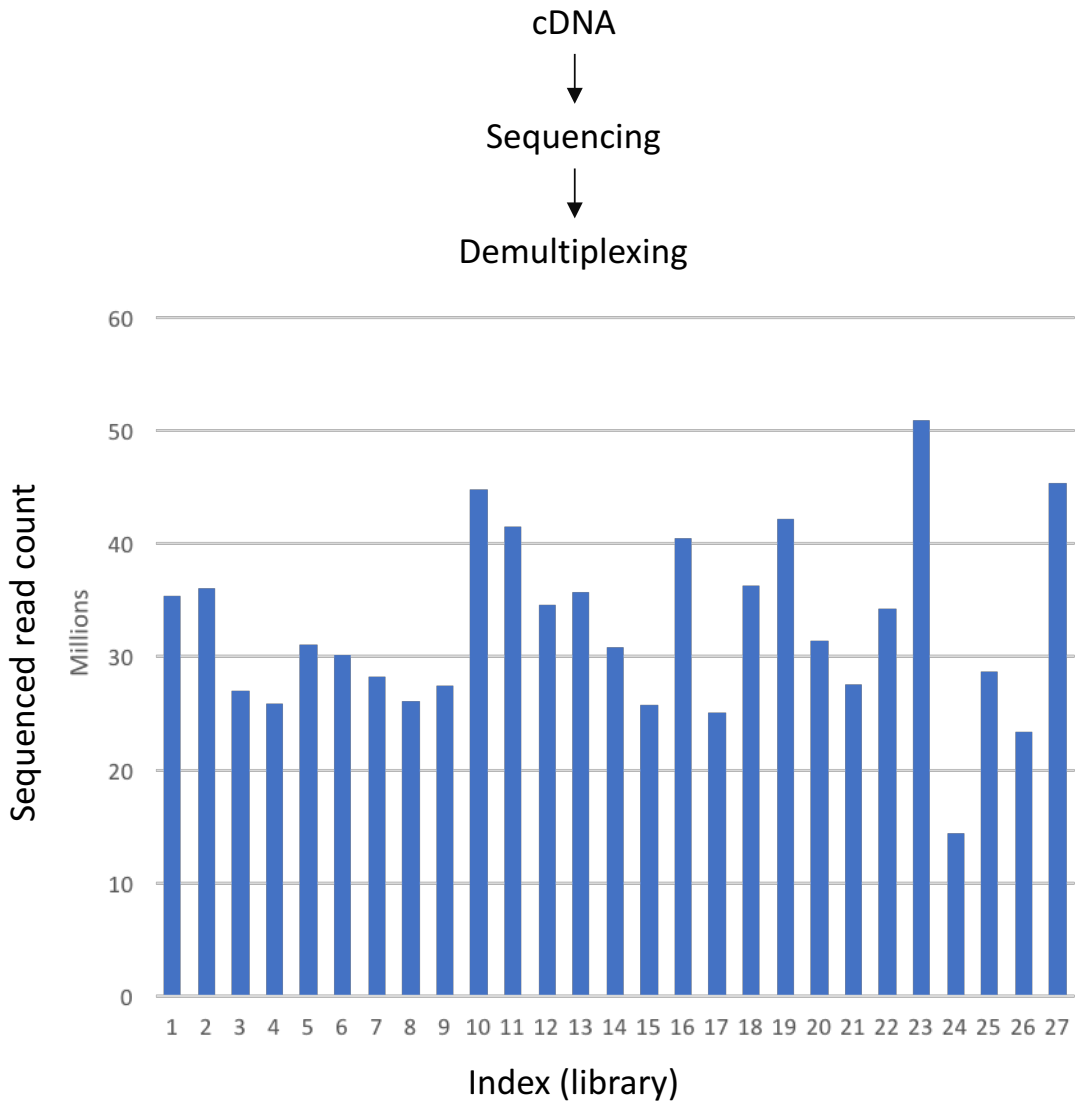
A few considerations…

- Single gene vs. -omics
- Single organism vs. metatranscriptomics
- Library prep. methods
- Single vs. paired-end sequencing
- Sequencing read length
- Depth of sequencing: number of multiplex barcodes per run
- Experimental design: Number of conditions tested and number of Biological / technical replicates
- rRNA depletion or mRNA enrichment, or other enrichments
- Standards spiked in for normalization
- Sequence QC and filtering
- Read mapping
- Differential expression
- Annotation and downstream analysis
- Comparison/validation with other datasets or data types (eg. proteomics, ssu-rRNA, RT-qPCR, metabolomics)

RNA-seq sample prep

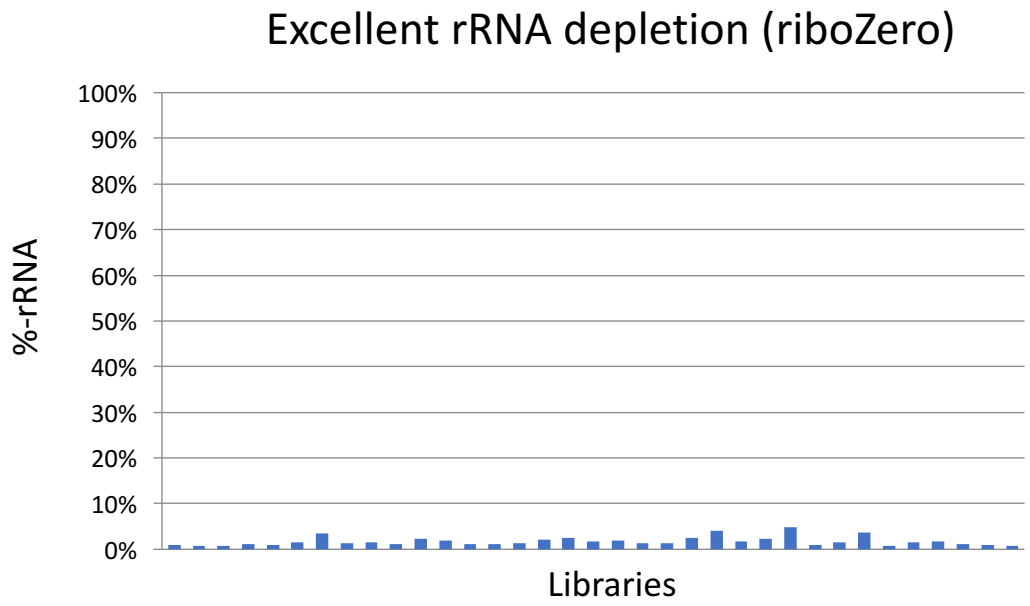


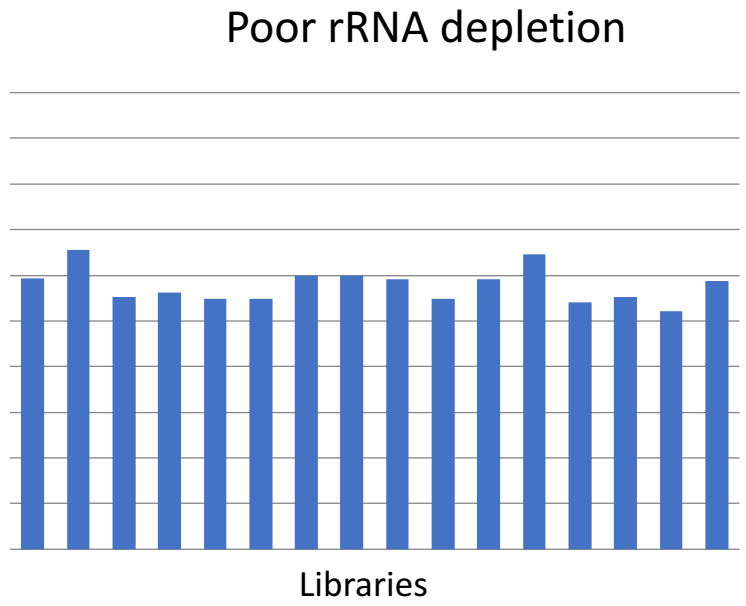
http://www.nature.com/nmeth/journal/v12/n4/full/nmeth.3313.html



rRNA Depletion & mRNA enrichment

- rRNA and tRNA can make up as much as 95% of total RNA in a typical sample
- How do we get rid of rRNA after sequencing: Ribopicker (http://ribopicker.sourceforge.net/)
- Can filtered rRNA still be used for analysis? Yes, but not going to be accurate for quantitative comparisons
- Comparison of two real examples of metatranscriptomic datasets:



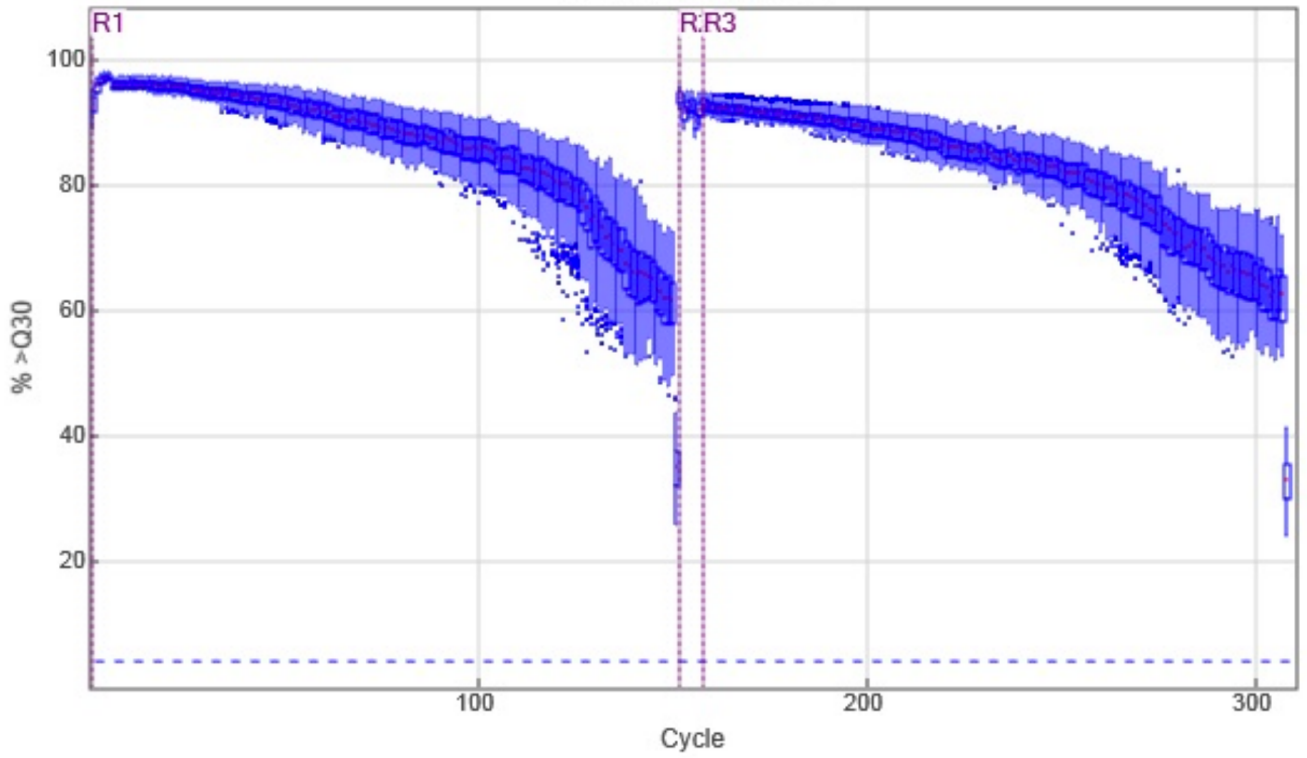


Data QC

- Always check the quality of your data at each step to confirm things are working as expected.
- Check FastQC reports
- Visual inspection
- Simple heatmap of data to confirm replicates and conditions group as expected
  - Evidence of batch effects, mislabeled samples, or other biases?

FASTQC

Does quality drop off too fast?
Are there sequence position specific biases?



Example, read 1 of NextSeq 500 paired-end

FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

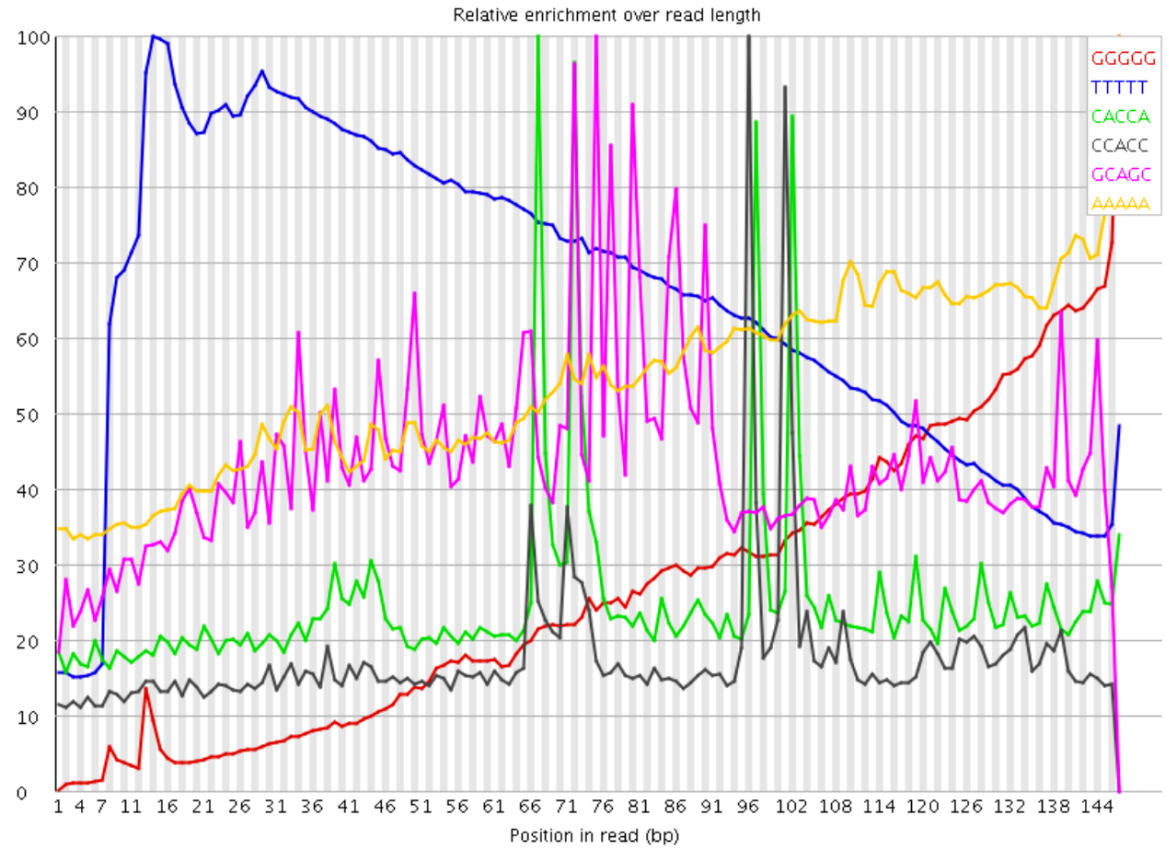
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Position bias, or just low-complexity repetitive sequence?

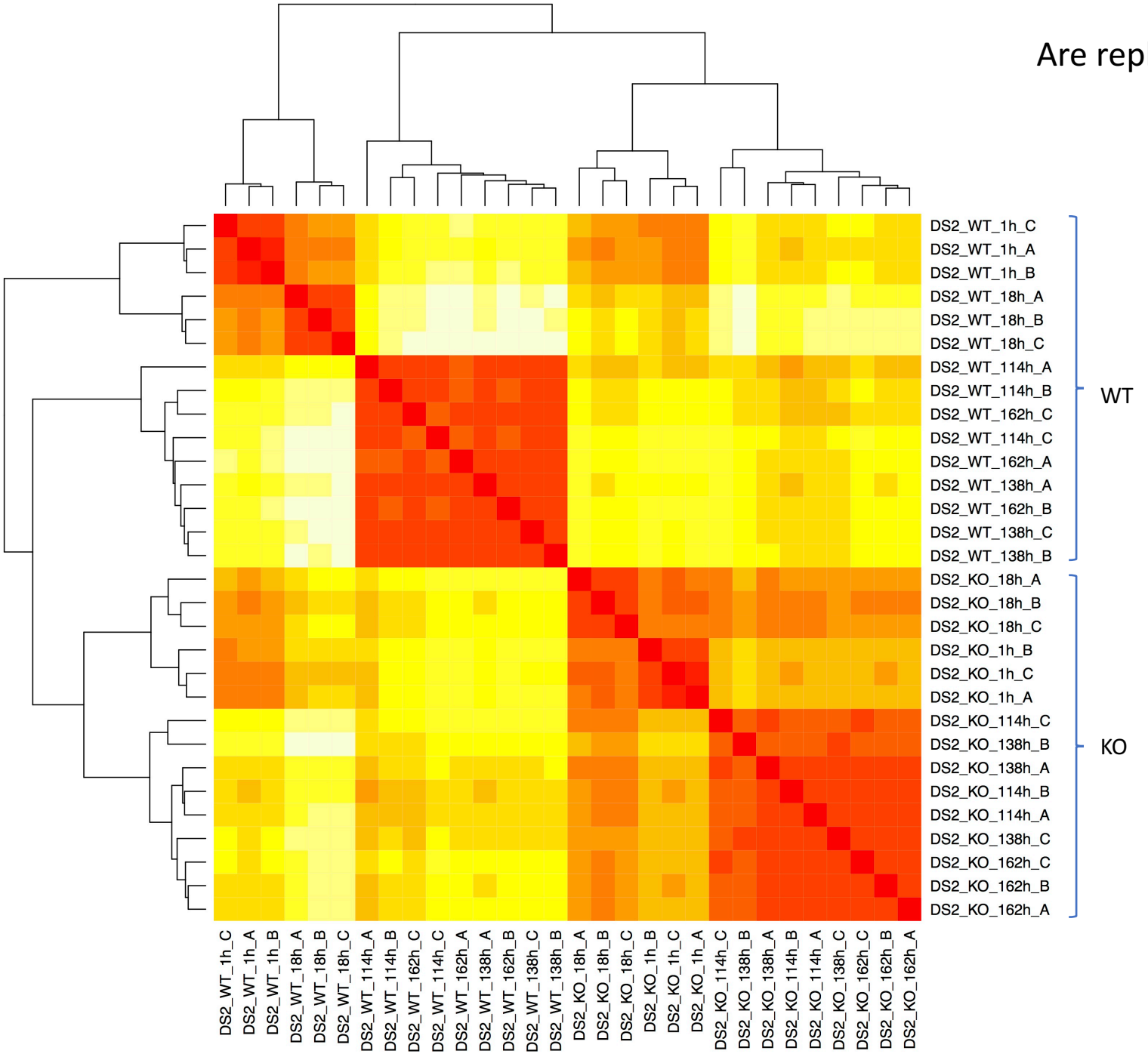
Overrepresented sequences

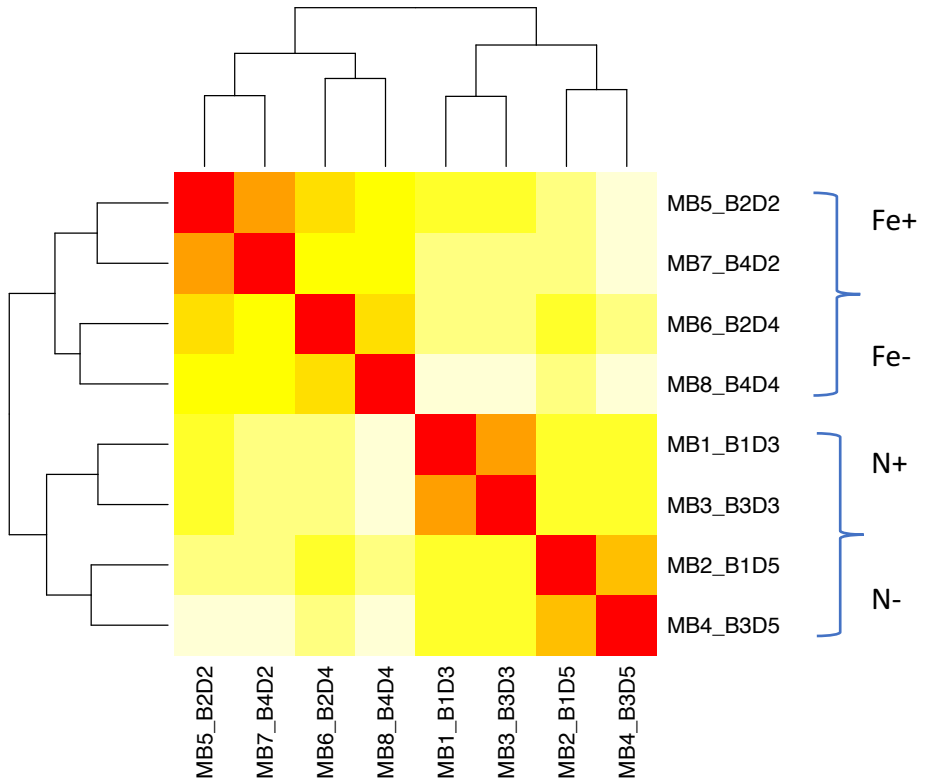
Sequence	Count	Percentage
TGTTTGATTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	173377	1.3407219438562004
CTTTGTGTTTGATTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	80299	0.6209510567705581
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	59067	0.4567642943282799
TGTTTGACTCTCTTTTCAAAGTTCTTTGCATCTTTCCCTCACGGTACTTG	42552	0.32905402766785125
CTTTGTGTTTGACTCTCTTTTCAAAGTTCTTTGCATCTTTCCCTCACGGT	37332	0.2886878398405767
GTGTTTGATTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	26656	0.20613047944900925
TGTTTGAGTGGGAATTAAGCTACCGTTTAGGTATGCGTCCATGGATCTTC	19594	0.15152013108958162
ACTTTGTGTTTGATTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	16852	0.13031628300100181
CTTTGTGTTTGAGTGGGAATTAAGCTACCGTTTAGGTATGCGTCCATGGA	15421	0.11925037978628346

Kmer Content

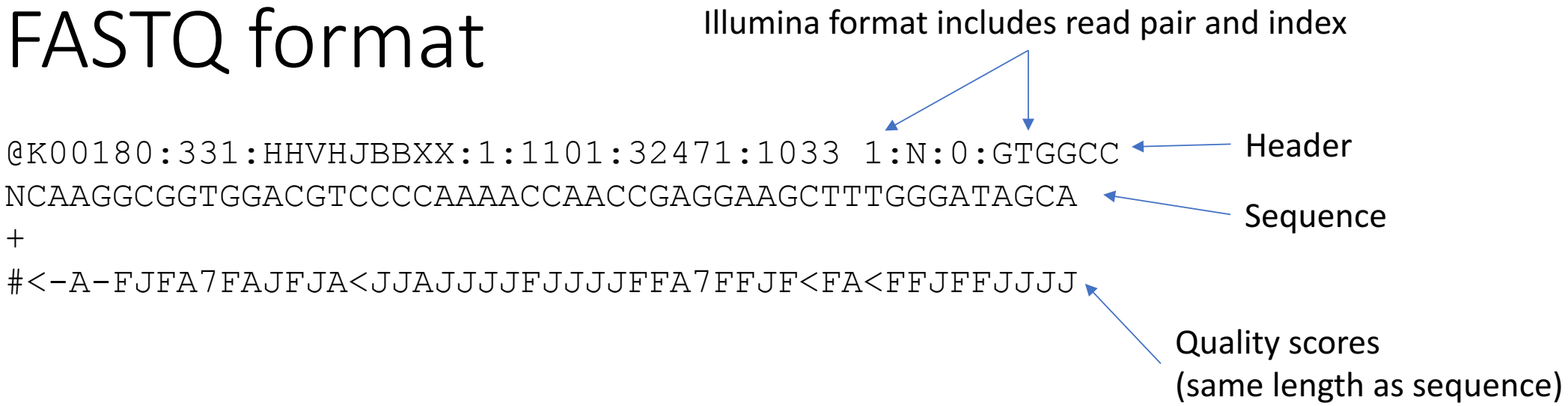


Are replicates and conditions separating as expected?

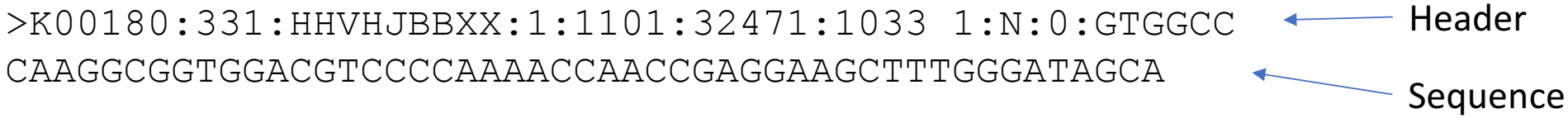




FASTQ format



FASTA format



No quality scores

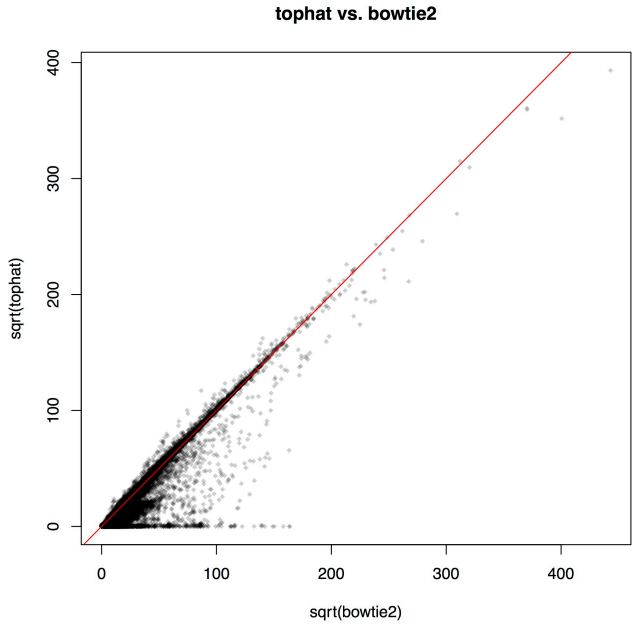
Trimmed for high-quality bases only (Q30)

Outline

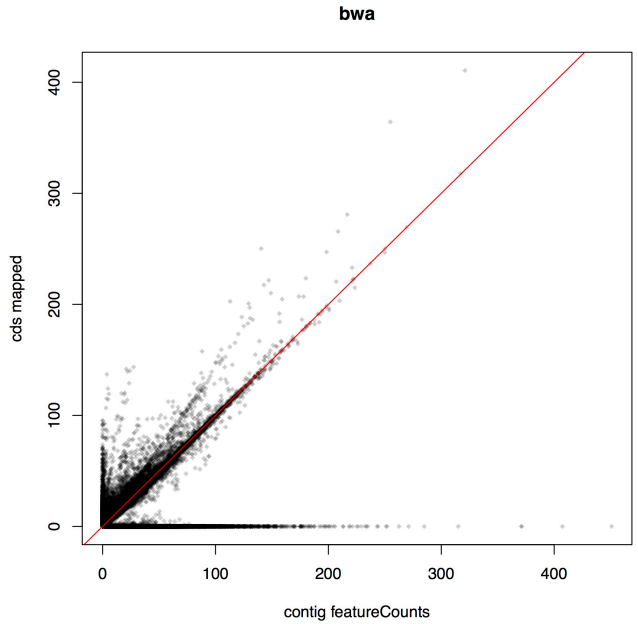
- Transcriptomics what/why/how
- Considerations when designing an experiment
- Quality control of data
- Read mapping
- Differential expression
- Multi-factor experimental design
- Meta-transcriptomics
- Fall workshop ideas

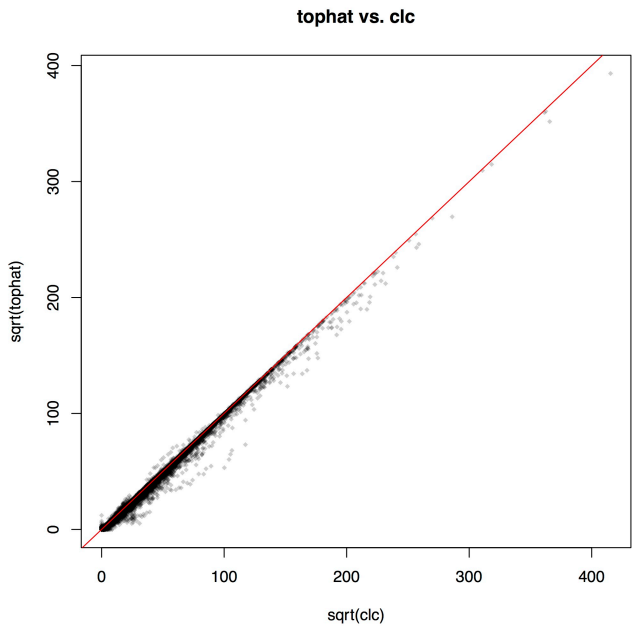
Read mapping

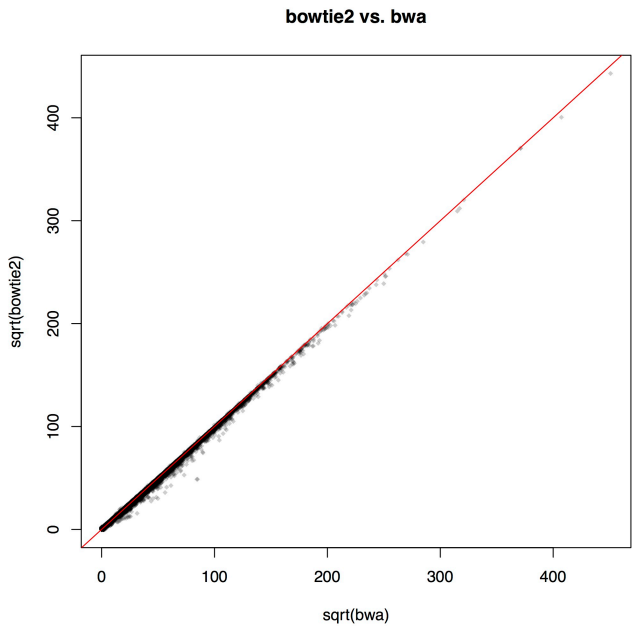
- What tools to use to map reads, and what to map to?
  - BWA, CLC, Bowtie2
- What about Eukaryotic splice junctions?
  - Tophat is slower, but will allow reads to map across introns, even without predefined exon locations. Uses Bowtie2 on split reads as a first step, then finds potential junctions where parts of reads tend to map at a distance.
  - Simply running BWA works almost as well, and is much faster
  - When reference is a transcriptome, this is not an issue



Gene (exons) read count comparisons by various read mapping tools. Reads are mapped to contigs and featureCounts is used to extract exon level counts per gene.







BWA comparison of mapping to CDS directly, versus mapping to contigs followed by featureCounts

FeatureCounts: http://bioinf.wehi.edu.au/featureCounts/
Pt genome: http://protists.ensembl.org/Phaeodactylum_tricornutum/Info/Index

Normalization

Raw read counts

	Samp 1	Samp 2	Samp 3
Gene 1	10	12	5
Gene 2	62	60	56
Gene 3	128	139	98
Gene 4	1785	1792	1104
total	1985	2003	1263

Library normalization

	Samp 1	Samp 2	Samp 3
Gene 1	0.005037783	0.005991013	0.003958828
Gene 2	0.031234257	0.029955067	0.044338876
Gene 3	0.064483627	0.069395906	0.077593032
Gene 4	0.899244332	0.894658013	0.874109264
total	1	1	1

- Total sample read counts reflect sequencing depth rather than absolute transcript abundance.
- But we can measure relative abundance. Divide counts by the total for each sample.
- Gene length should also be taken into account, as we expect longer targets to be sequenced more frequently.

Relative Abundance

- RPKM - reads per kilobase transcript (gene length) per million mapped (library size)

  $RPKM = (C * 10^9) / (L * N)$

  C = number of reads mapped for this gene
  L = length of this gene
  N = total number of mapped reads for all genes

- FPKM - count fragments the same for one read, as both in one pair
- Good measure to compare gene expression between groups of genes and across samples
- Other adaptations: Group normalized RPKM, or mixed TMM normalized RPKM.
- Why different than DE? Isn't there just one measure of transcript abundance that is real?
  - Problems with low-end abundances
  - Alternatively, you could spike in standards and have accurate cell counts
  - Different genes have different variances, fluctuating biologically and measurement error

PLOS | GENETICS

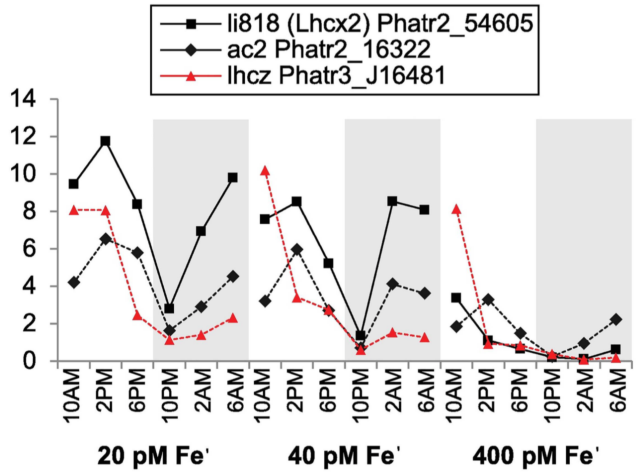
RESEARCH ARTICLE

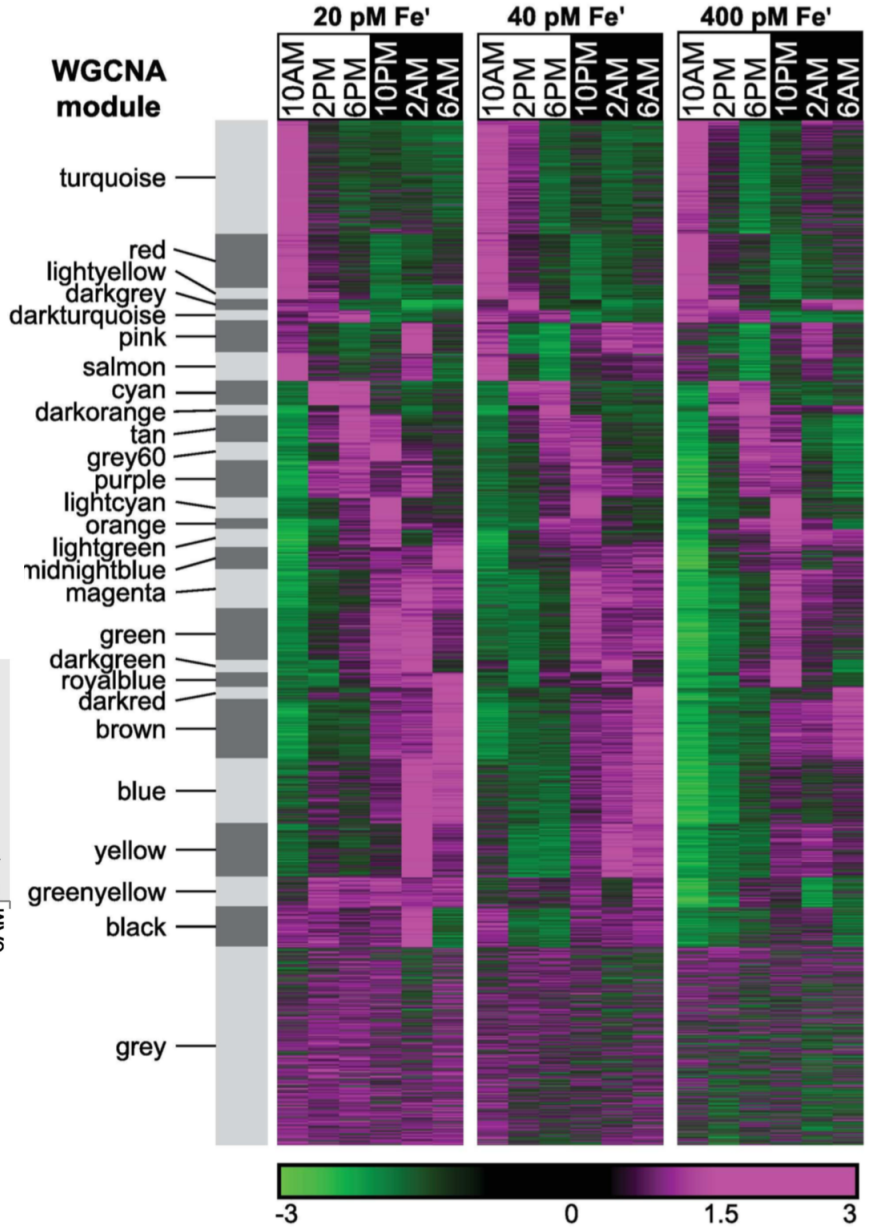
Transcriptional Orchestration of the Global Cellular Response of a Model Pennate Diatom to Diel Light Cycling under Iron Limitation

Sarah R. Smith[1,2☯], Jeroen T. F. Gillard[2,3☯], Adam B. Kustka[4], John P. McCrow[2], Jonathan H. Badger[2¤], Hong Zheng[2], Ashley M. New[4], Chris L. Dupont[2], Toshihiro Obata[5], Alisdair R. Fernie[5], Andrew E. Allen[1,2*]

http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1006490

- Vizualizations of relative abundance genes measured by RPKM
- WGCNA clustering of time-series expression patterns
- Multi-factor design: Light/dark, time, Fe concentration





Outline

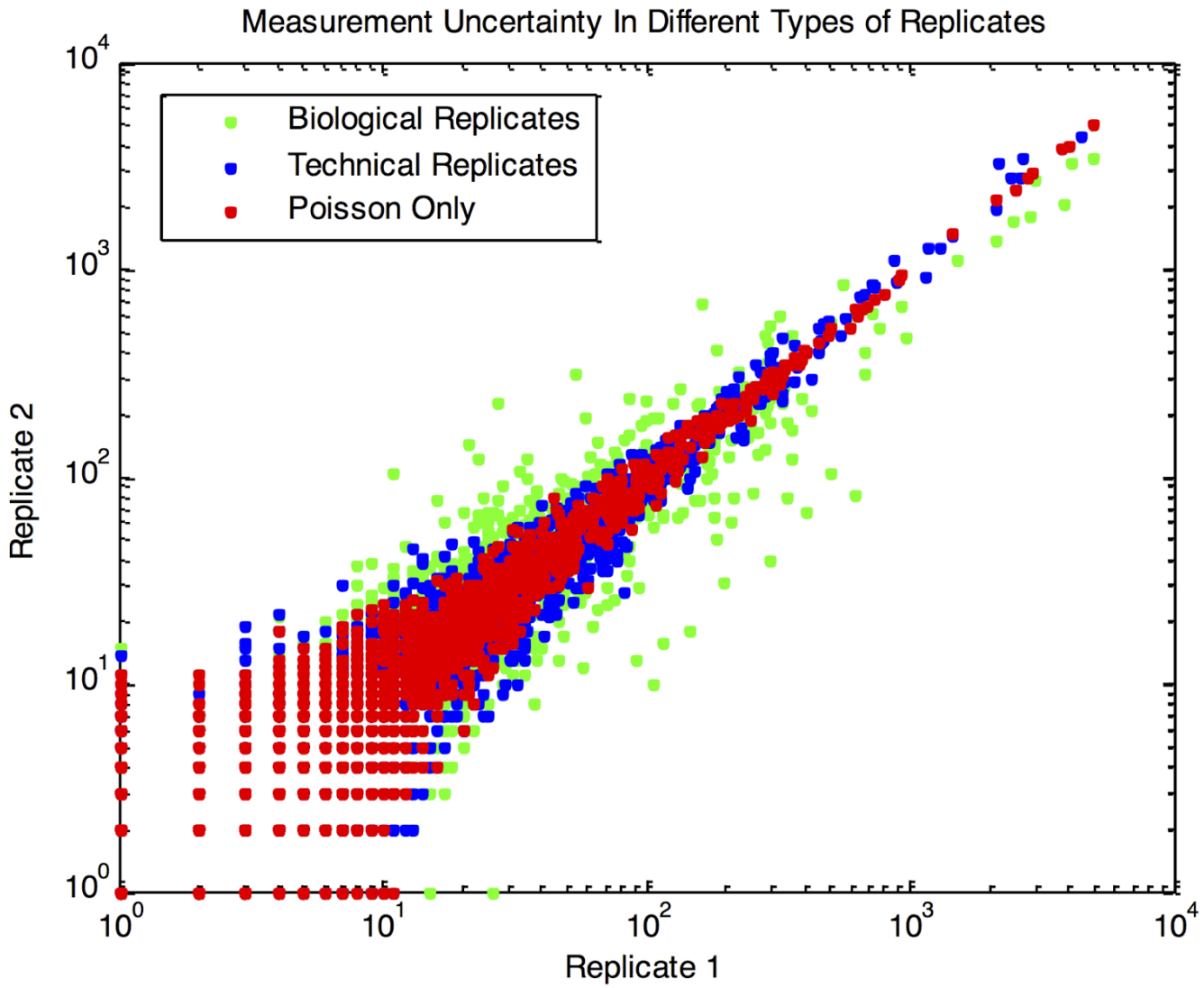
- Transcriptomics what/why/how
- Considerations when designing an experiment
- Quality control of data
- Read mapping
- Differential expression
- Multi-factor experimental design
- Meta-transcriptomics
- Fall workshop ideas

Differential Expression

Is there a difference in expression between treatment and control?

	Control A	Control B	Control C	Treatment A	Treatment B	Treatment C
Gene 10	10	14	8	145	256	189

- Best methods are edgeR and DESeq
- I prefer edgeR:
  - comparable results and accuracy to DESeq, but faster in general
  - 2 ways to make comparisons, exact test (comparable to Fisher exact test), using GLM methods LR or QL, which allow contrasts
- Both use raw read counts, and do different types of normalizations internally to take into account differences in sequencing depth, variances in gene abundances overall, as well as per gene.
- Require replicates, a minimum of 3 (2 in the control, 1 in the treatment) to even use this. But recommended minimum, is at least biological triplicates in all conditions tested.



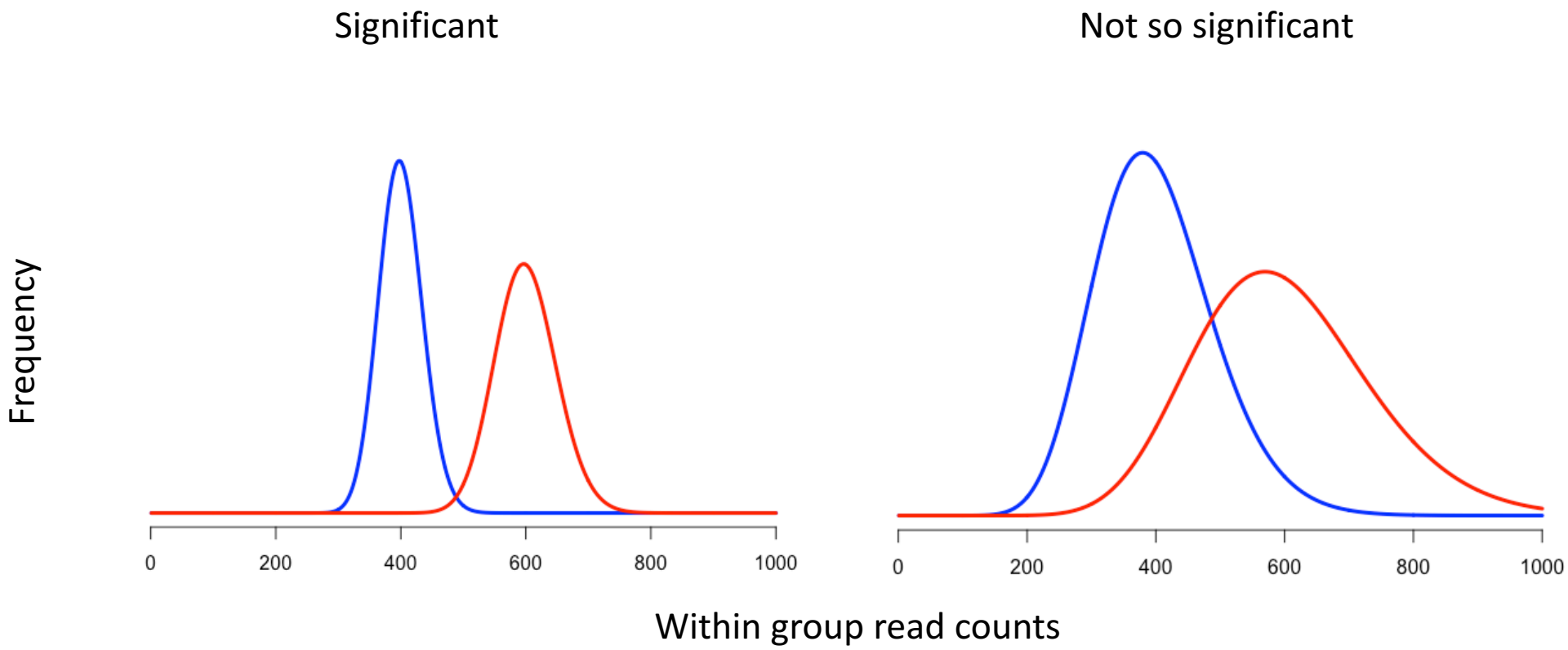
Data from 2 replicates of S. cerevisiae transcriptome:

Sources of variation, and thus measurement uncertainty, can come from changes in cellular expression (biological), anywhere from lab. prep. through sequencing and even analysis (technical). Theoretical noise due only to counting reads is based on simulations and approximates a Poisson distribution.

http://michelebusby.tumblr.com/post/26913184737/thinking-about-designing-rna-seq-experiments-to

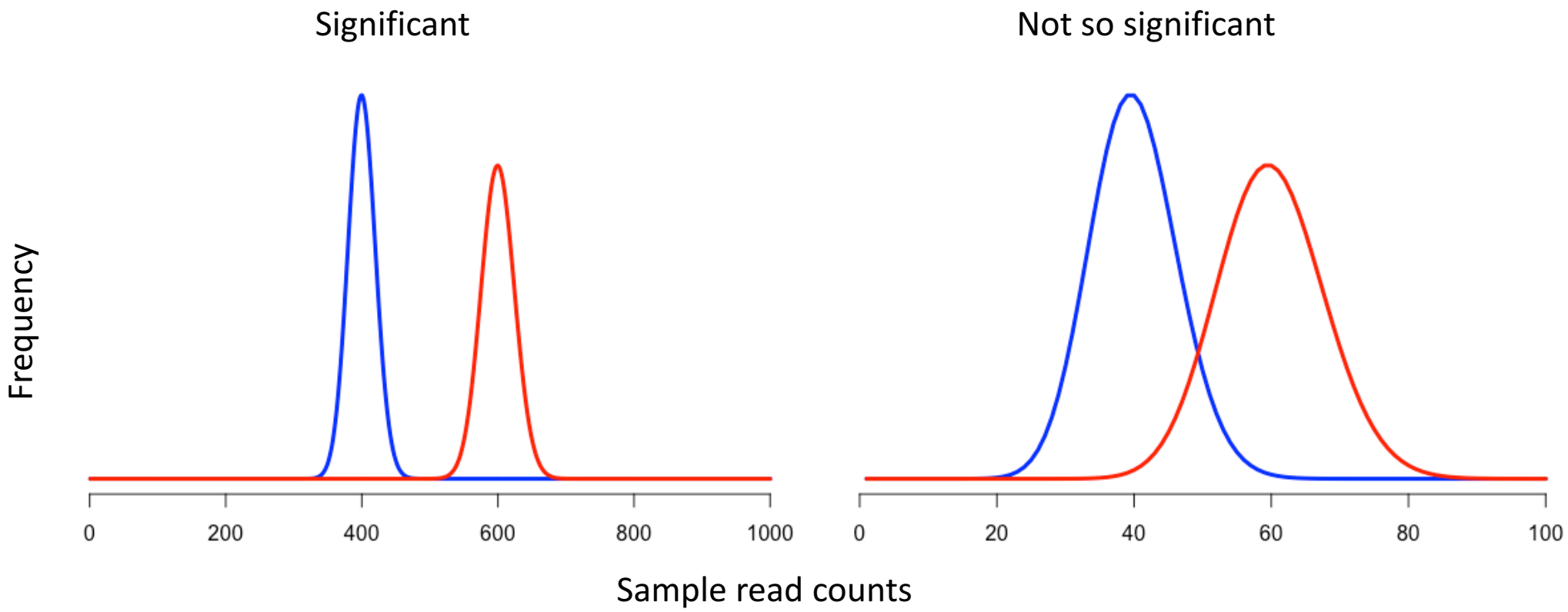
Detecting significant differential expression

Variance in expression between replicates determine our power to detect differential expression.
Given the same mean differences, a higher variance makes it more difficult to detect a difference.
Better experimental control, more read depth, and more replicates can help.



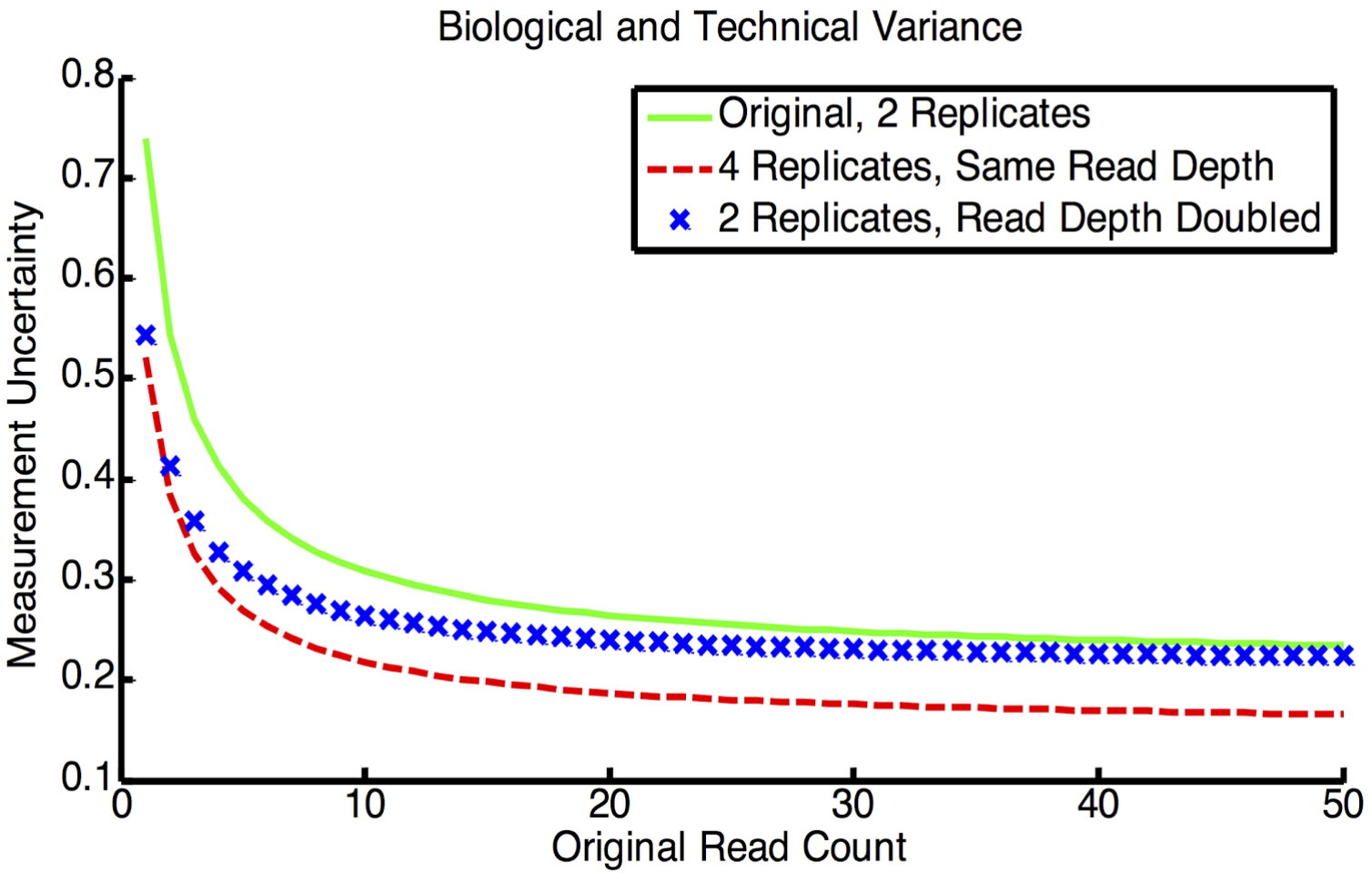
Detecting significant differential expression

Even disregarding technical and biological variance, and modelling read counts as a simple Poisson distribution, the low end is harder to detect DE. Even though both examples have a 1.5X difference, read counts at the low end have more overlap so differences are less significant.



Reducing uncertainty: Greater read depth or more replicates?

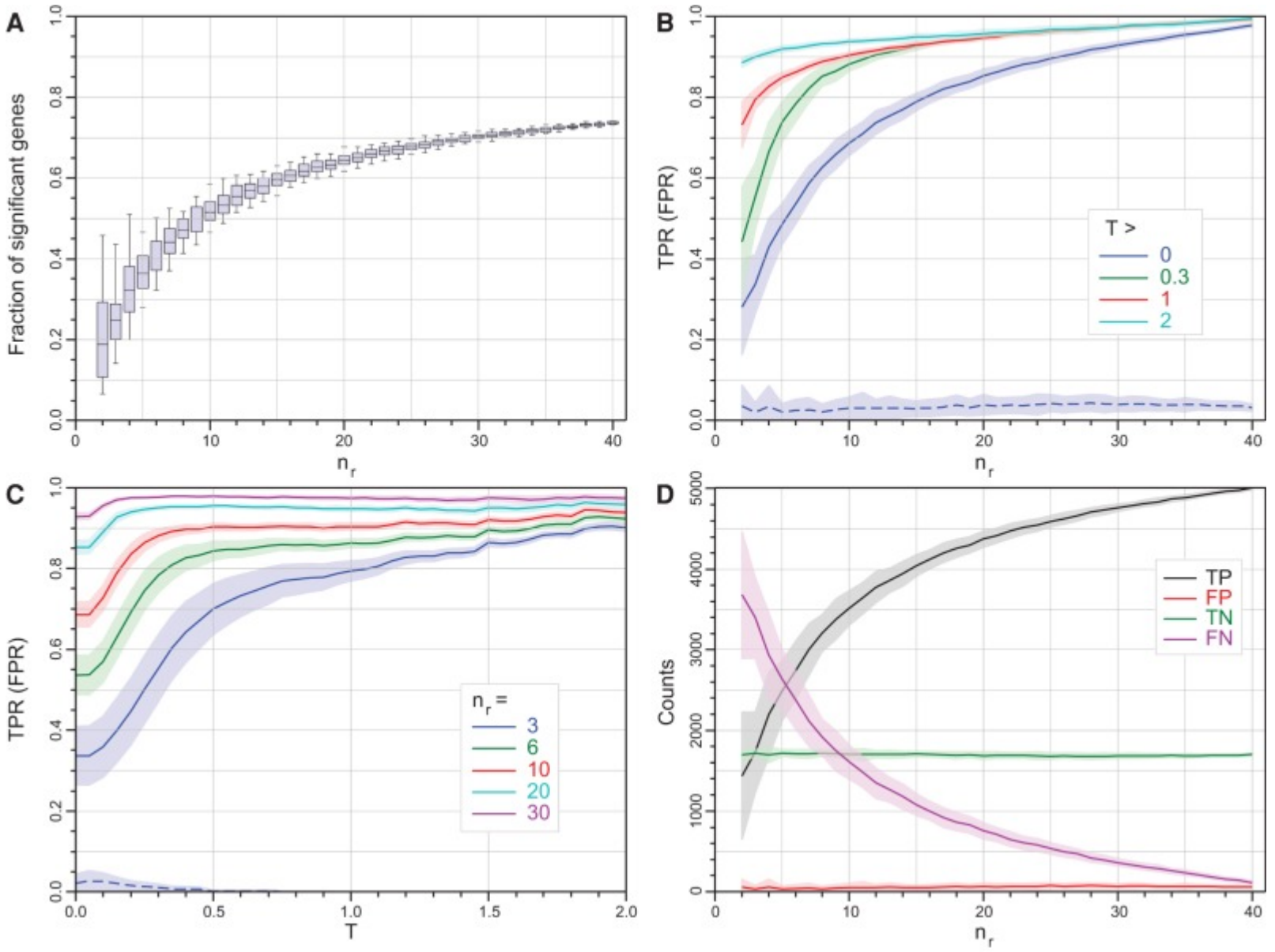
In theory only (Poisson), it doesn't matter, but in practice with real technical and biological variance it does.



https://github.com/quinlan-lab/applied-computational-genomics

I see you don’t have enough replicates!

- Timmy at Coronado Brewery (Nov. 2, 2016)



From Schurch et al. 2016 paper (University of Dundee)

Statistical properties of edgeR (exact)
T = log2 fold-change threshold
nr = number of replicates
TPR = true positive rate
FPR = false positve rate (dashed lines)

Take-aways:

1. edgeR is very good at controlling FDR
2. More power to detect significant DE for larger T, larger nr, or greater read depth

Schurch NJ, Schofield P, Gierliński M, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*. 2016;22(6):839-851. doi:10.1261/rna.053959.115.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4878611/

TABLE 2. A summary of the recommendations of this paper

	Agreement with other tools[a]	WT vs. WT FPR[b]	Fold-change threshold (T)[c]	Tool recommended for: (# good replicates per condition)[d] ≤3	≤12	>12
DESeq	Consistent	Pass	0	-	-	Yes
			0.5	-	Yes	Yes
			2.0	Yes	Yes	Yes
DESeq2	Consistent	Pass	0	-	-	Yes
			0.5	Yes	Yes	Yes
			2.0	Yes	Yes	Yes
EBSeq	Consistent	Pass	0	-	-	Yes
			0.5	-	Yes	Yes
			2.0	Yes	Yes	Yes
edgeR (exact)	Consistent	Pass	0	-	-	Yes
			0.5	Yes	Yes	Yes
			2.0	Yes	Yes	Yes
Limma	Consistent	Pass	0	-	-	Yes
			0.5	-	Yes	Yes
			2.0	Yes	Yes	Yes
cuffdiff	Consistent	Fail				
BaySeq	Inconsistent	Pass				
edgeR (GLM)	Inconsistent	Pass				
DEGSeq	Inconsistent	Fail				
NOISeq	Inconsistent	Fail				
PoissonSeq	Inconsistent	Fail				
SAMSeq	Inconsistent	Fail				

[a]Full clean replicate data set, see section "Tool Consistency with High Replicate Data" and Figure 3.
[b]See section "Testing Tool False Positive Rates" and Figure 4.
[c]See section "Differential Expression Tool Performance as a Function of Replicate Number."
[d]See Figure 2.

Schurch NJ, Schofield P, Gierliński M, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*. 2016;22(6):839-851. doi:10.1261/rna.053959.115. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4878611/

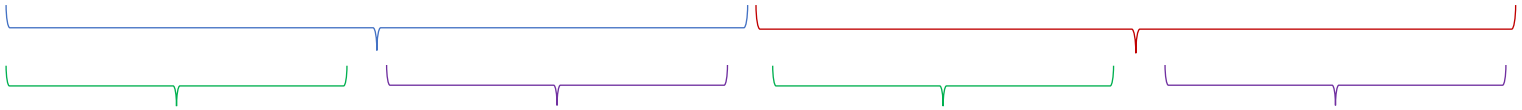
[Recommendations] when designing an RNA-seq experiment for DGE:

- 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.
- For experiments with <12 replicates per condition; use *edgeR (exact)* or *DESeq2*.
- For experiments with >12 replicates per condition; use *DESeq*.
- Apply a fold-change threshold appropriate to the number of replicates per condition between $0.1 \leq T \leq 0.5$

For additional help with experimental design, see:
Scotty - Power Analysis for RNA Seq Experiments
https://doi.org/10.1093/bioinformatics/btt015

Multi-factor experimental design

Temperature	4	4	18	18	4	4	18	18
Salinity	low	low	low	low	high	high	high	high
Gene 100	10	14	172	145	25	32	310	289



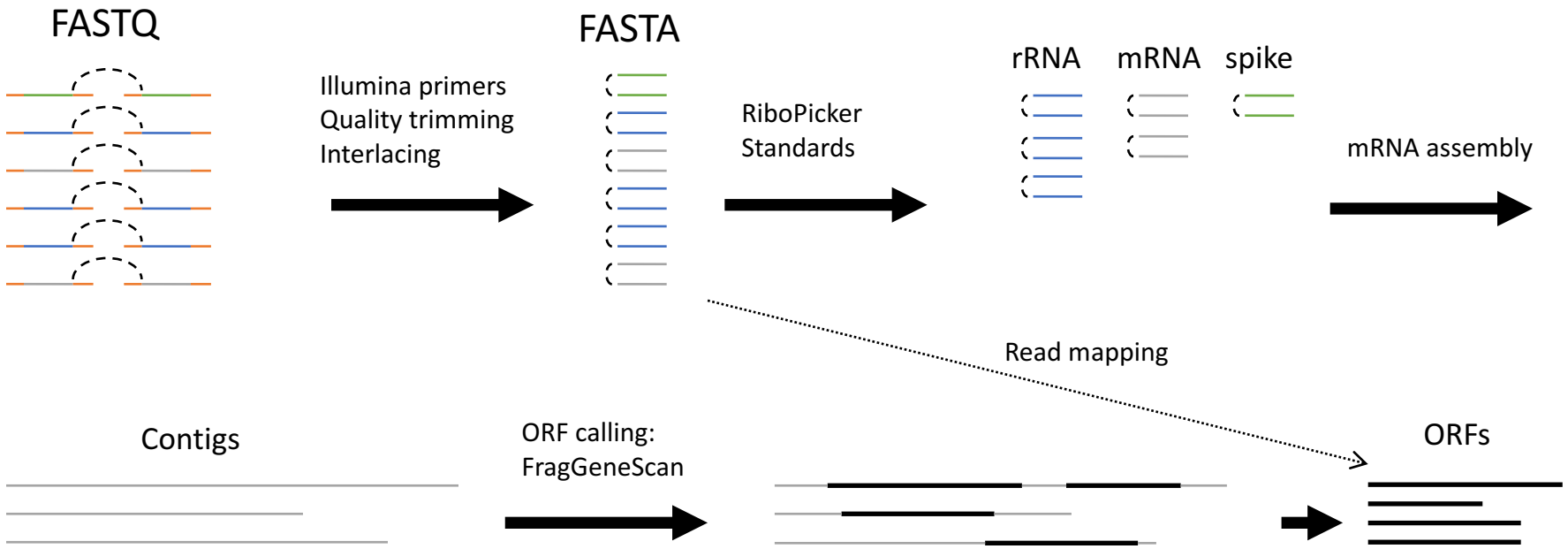
- Here there are 2 factors (temperature and salinity)
- Appears that both a raise in temp or in salinity result in higher expression
- There is also an interaction between the two factors, which is more than an additive
- Considering either one on its own makes the intra-group variance appear larger

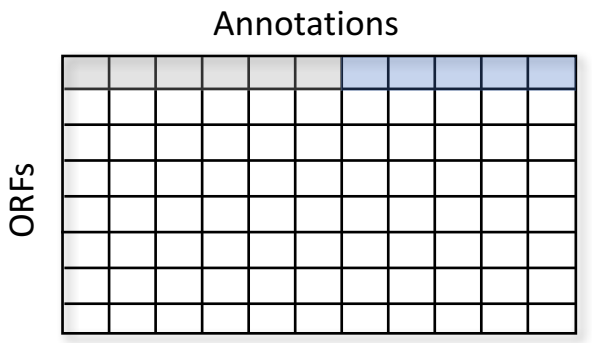
- More advanced experimental designs are available in edgeR:
  - https://www.bioconductor.org/packages/devel/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf
- For multi-factor designs set up contrasts and use generalized linear model (GLM) methods
- Factorial designs, nested interactions, blocking, all possible to set up using GLM in edgeR

Outline

- Transcriptomics what/why/how
- Considerations when designing an experiment
- Quality control of data
- Read mapping
- Differential expression
- Multi-factor experimental design
- Meta-transcriptomics
- Fall workshop ideas

RNA-Seq Annotation Pipeline (RAP)





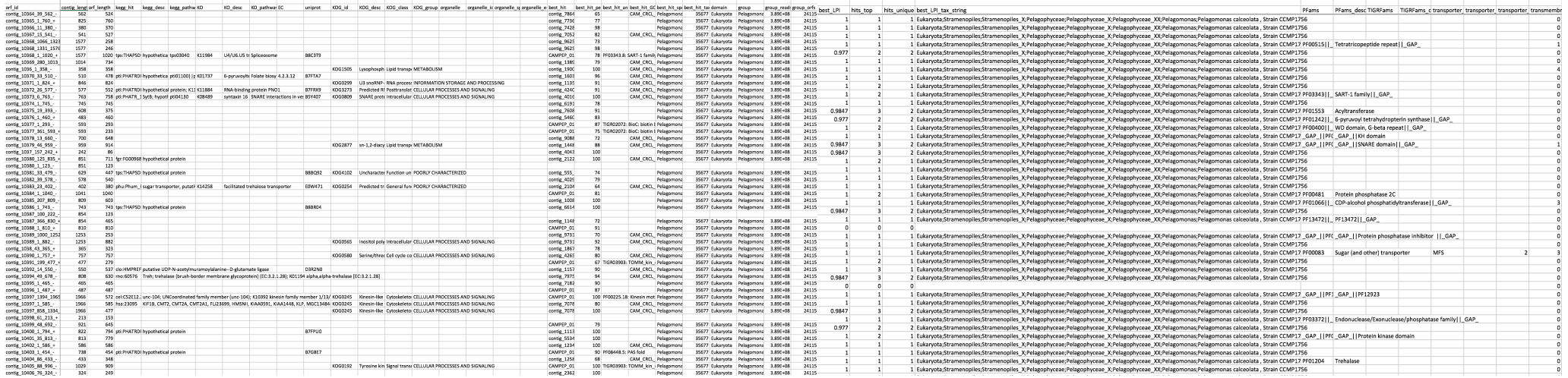
Types of annotations

Contamination: rRNA, ITS, primers
Transcript abundance: read mapping
BLASTP: PhyloDB, Kegg, organelles
HMMs: PFams, TIGRFams
Transporter: tmhmm, Pfams, BLAST
KOG, KO, Pathway, Uniprot, EC, GO, …

Secondary analysis

Differential expression
Group normalization
MCL clustering
Annotation summaries

RNA-seq *de novo* assembled ORF annotation table



ORFs

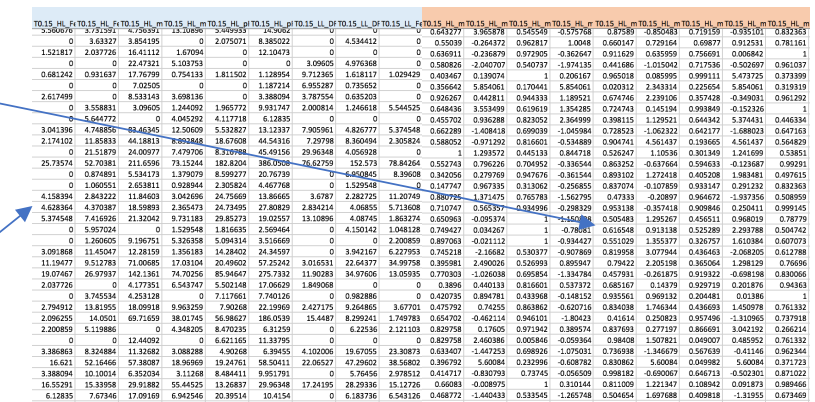
PhyloDB functional annotation

Taxonomic

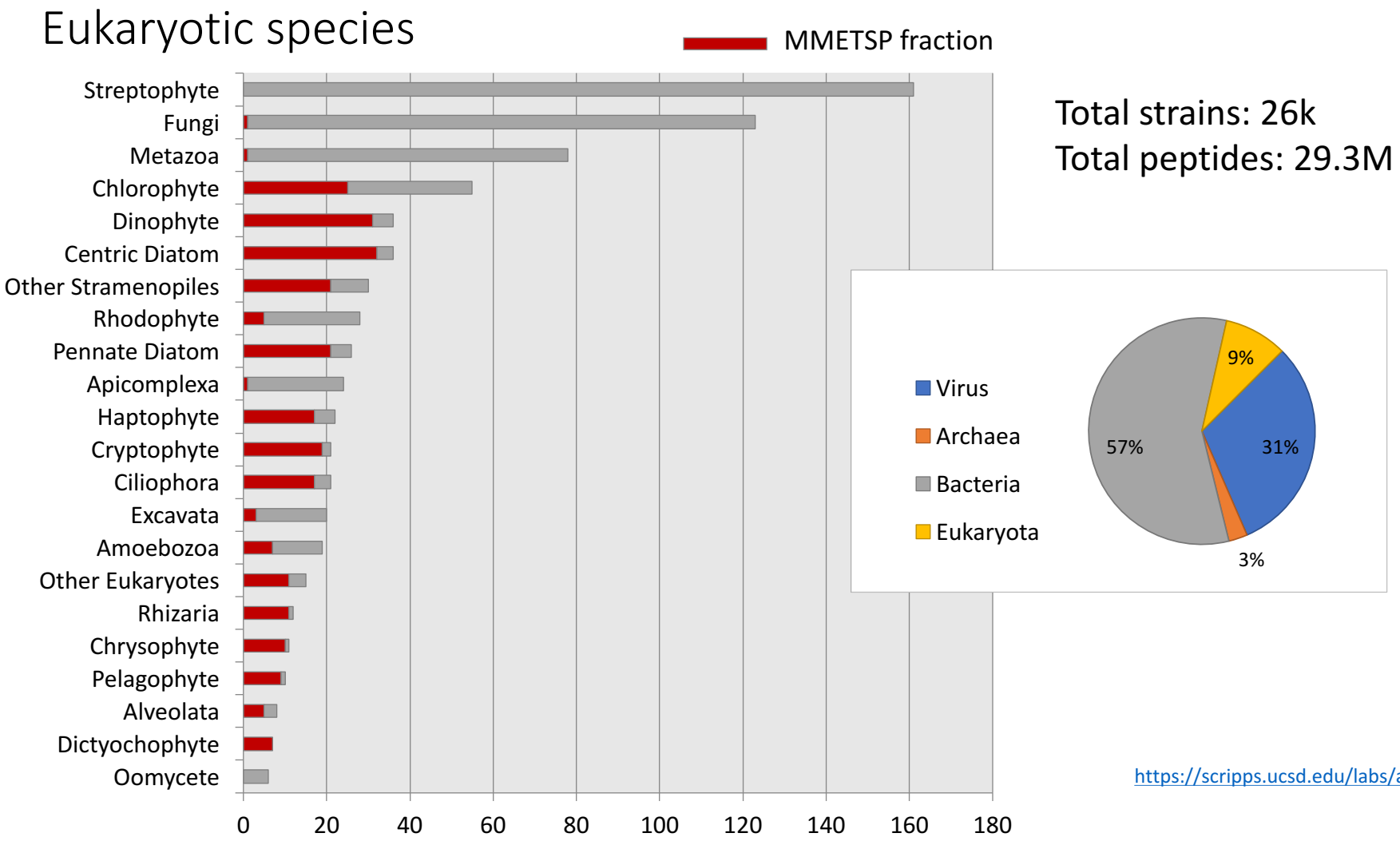
HMM

Differential expression - edgeR

RPKM abundance values



PhyloDB

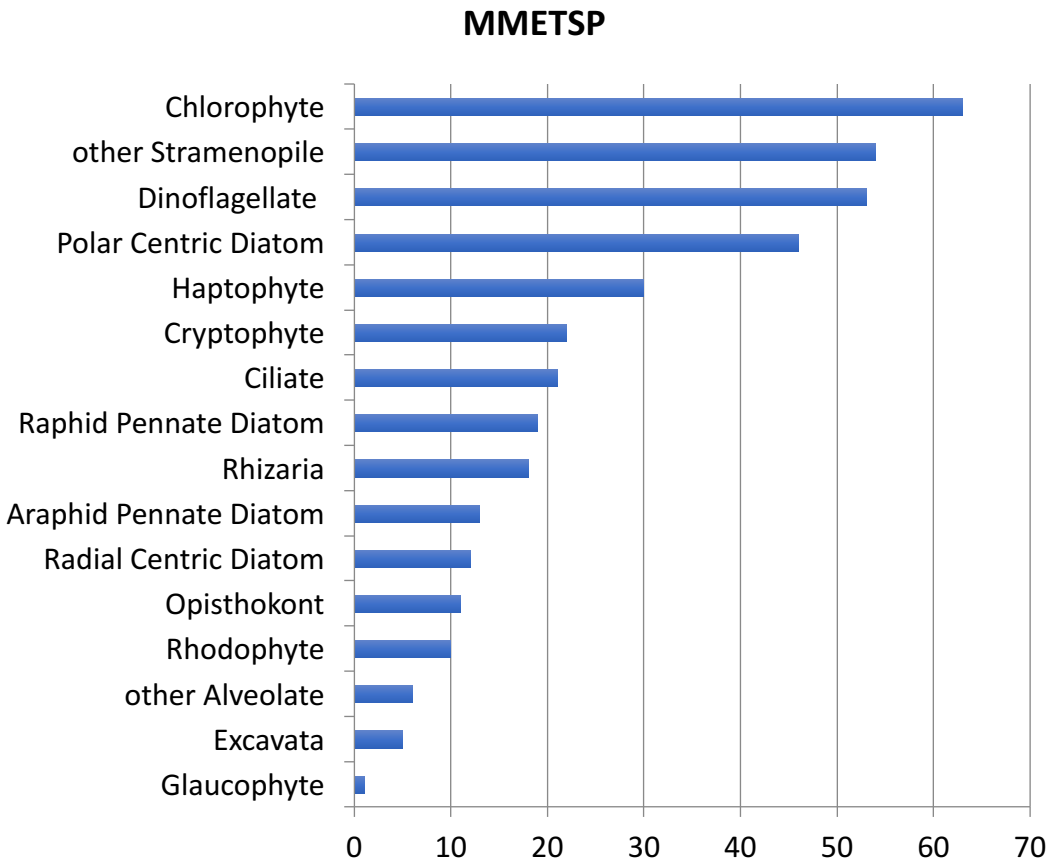


Total strains: 26k
Total peptides: 29.3M

https://scripps.ucsd.edu/labs/aallen/data/

Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP)

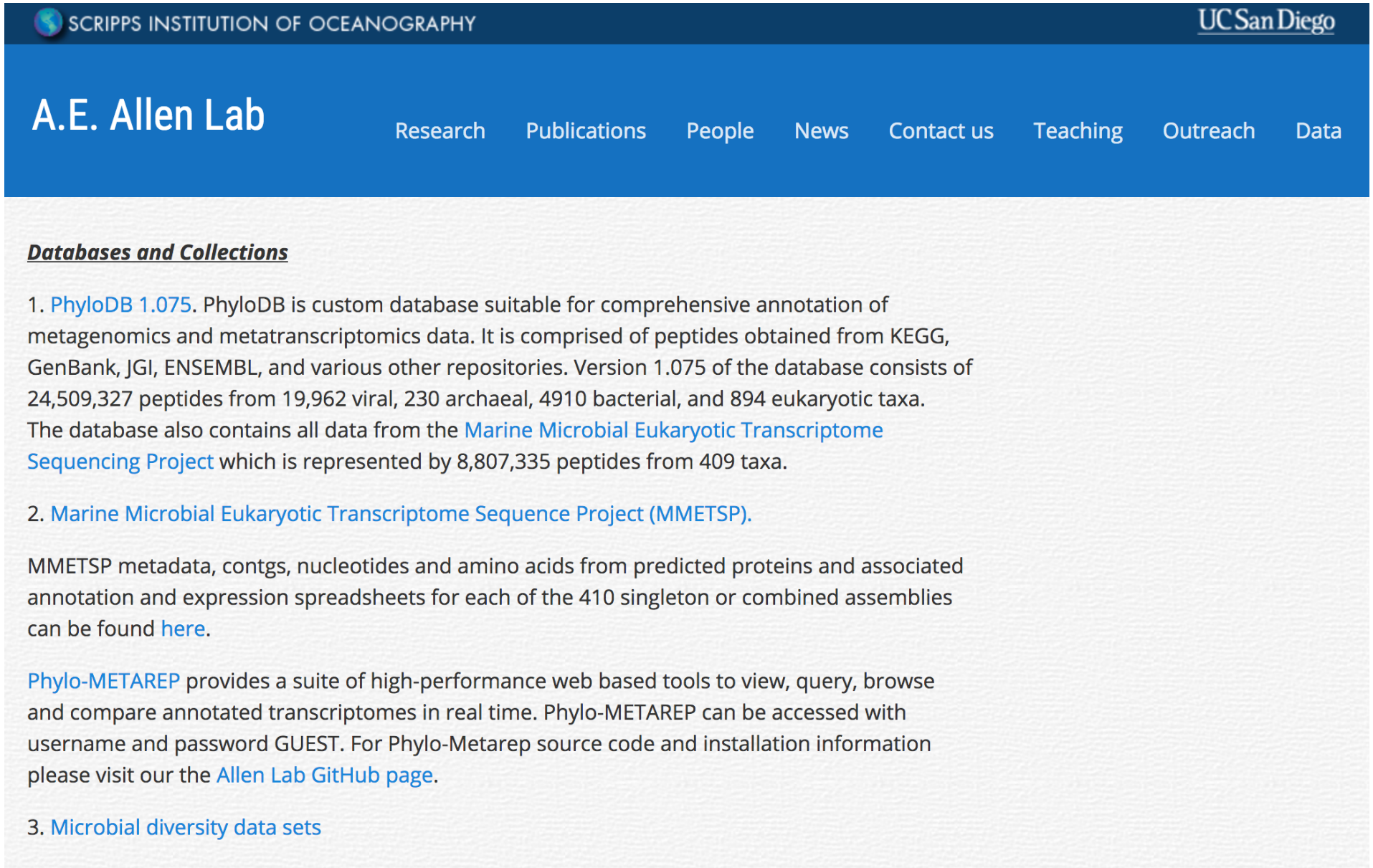
We now have the references to study marine microeukaryotes in much greater resolution.

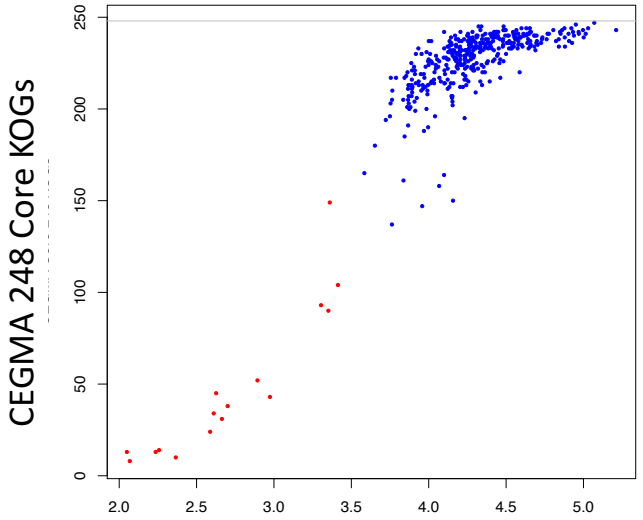


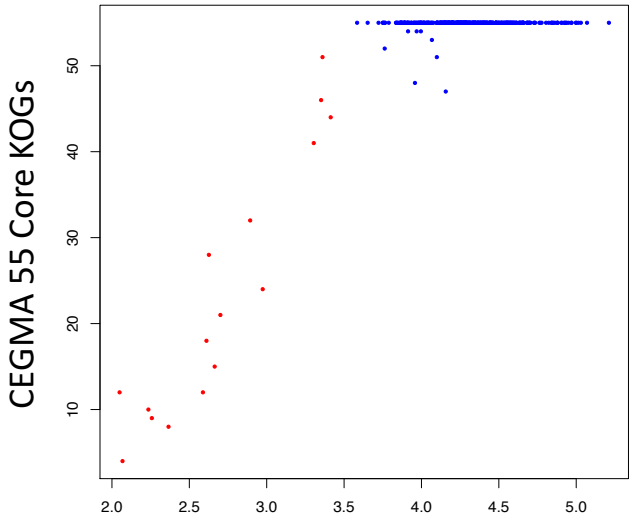
Total assemblies: 410
Samples: 678
Peptides: 9.9M

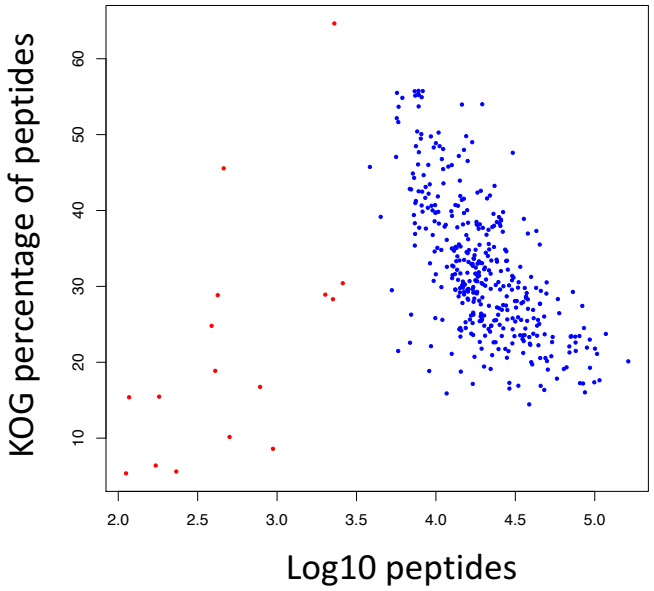
http://marinemicroeukaryotes.org
https://scripps.ucsd.edu/labs/aallen/data/

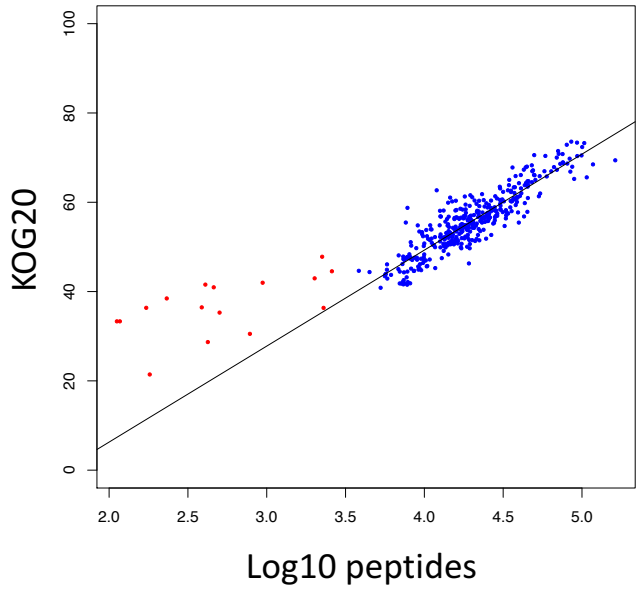
https://scripps.ucsd.edu/labs/aallen/data/







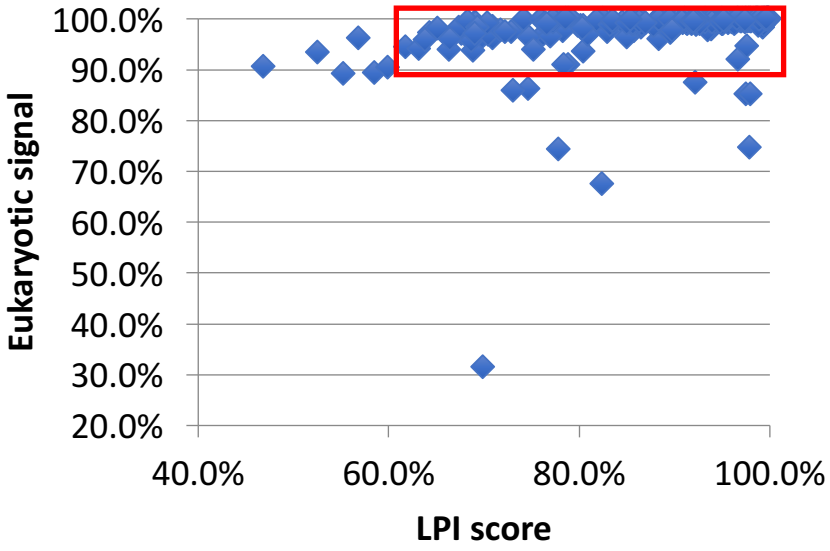




MMETSP

Filtered 410 assemblies to 384 that are high quality

KOG20 = Proportion of genome constituting largest 20% of KOGs
LPI = Lineage Probability Index (see: http://darkhorse.ucsd.edu/)



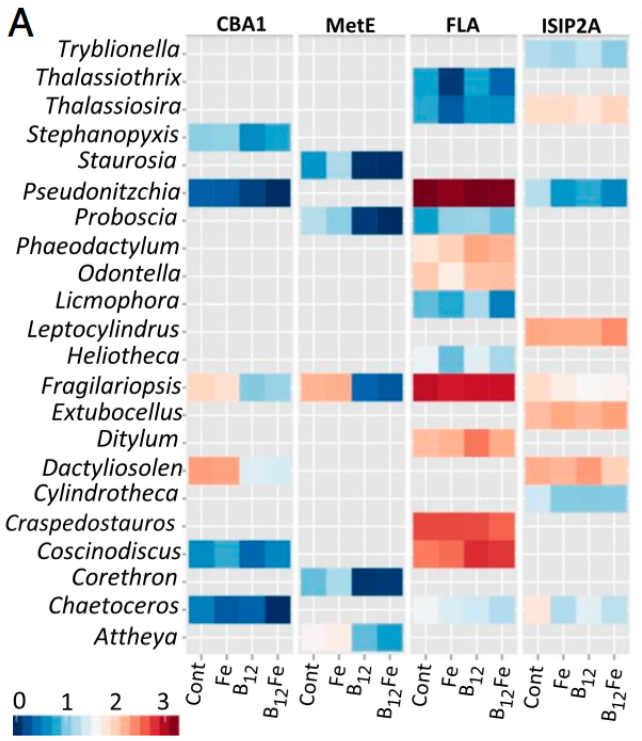
Metatranscriptomics

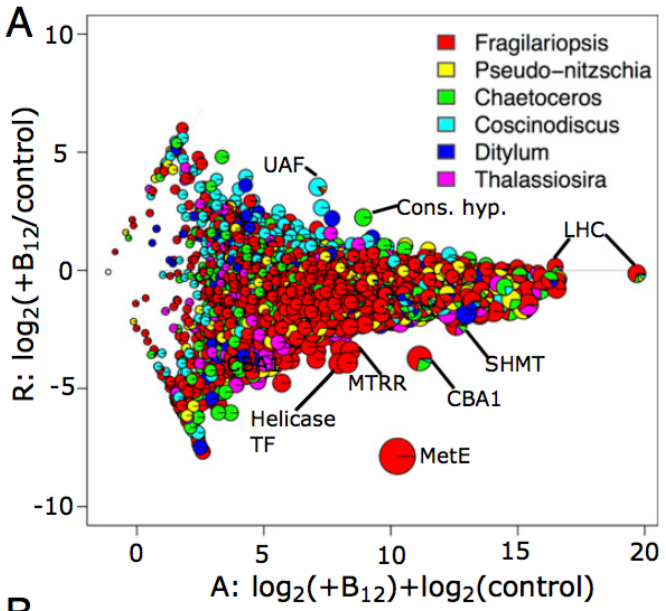
Phytoplankton–bacterial interactions mediate micronutrient colimitation at the coastal Antarctic sea ice edge

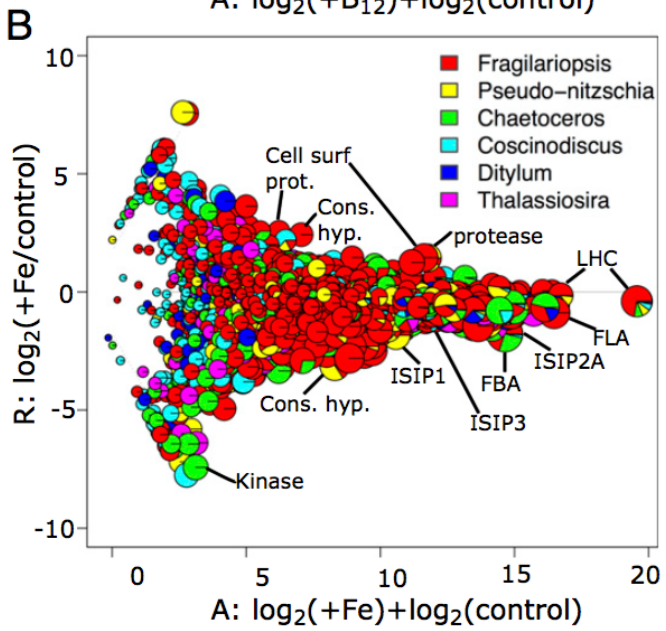
Erin M. Bertrand[a,b,1], John P. McCrow[a], Ahmed Moustafa[a,b,c], Hong Zheng[a], Jeffrey B. McQuaid[a,b], Tom O. Delmont[d], Anton F. Post[e], Rachel E. Sipler[f], Jenna L. Spackeen[f], Kai Xu[g], Deborah A. Bronk[f], David A. Hutchins[g], and Andrew E. Allen[a,b,2]

http://www.pnas.org/content/112/32/9938.long

- Vizualizations of functional and taxonomic variations
- Manta RA plots of fold-change vs. abundance, on functional clusters
- Multi-factor design: Control, Fe, B12, Fe+B12 additions







Fall Workshop – topics to cover in depth

- Data quality control
- Experimental design
  - Specific types of experiments (multi-factor, time-series, ...)
  - Power analysis
- Differential expression
- Metatranscriptomic analysis
- Tools:
  - R intro, edgeR, vegan, ...
  - BWA / samtools, ...

Last slide!