UBC Bioinformatics Class

Topic 5: RNAseq and analysis of differential gene expression

OUTLINE: RNAseq

1. Introduction and background
2. Overview of the methods and workflow
3. Quantifying expression levels
4. Analyzing patterns in expression
5. Technical considerations

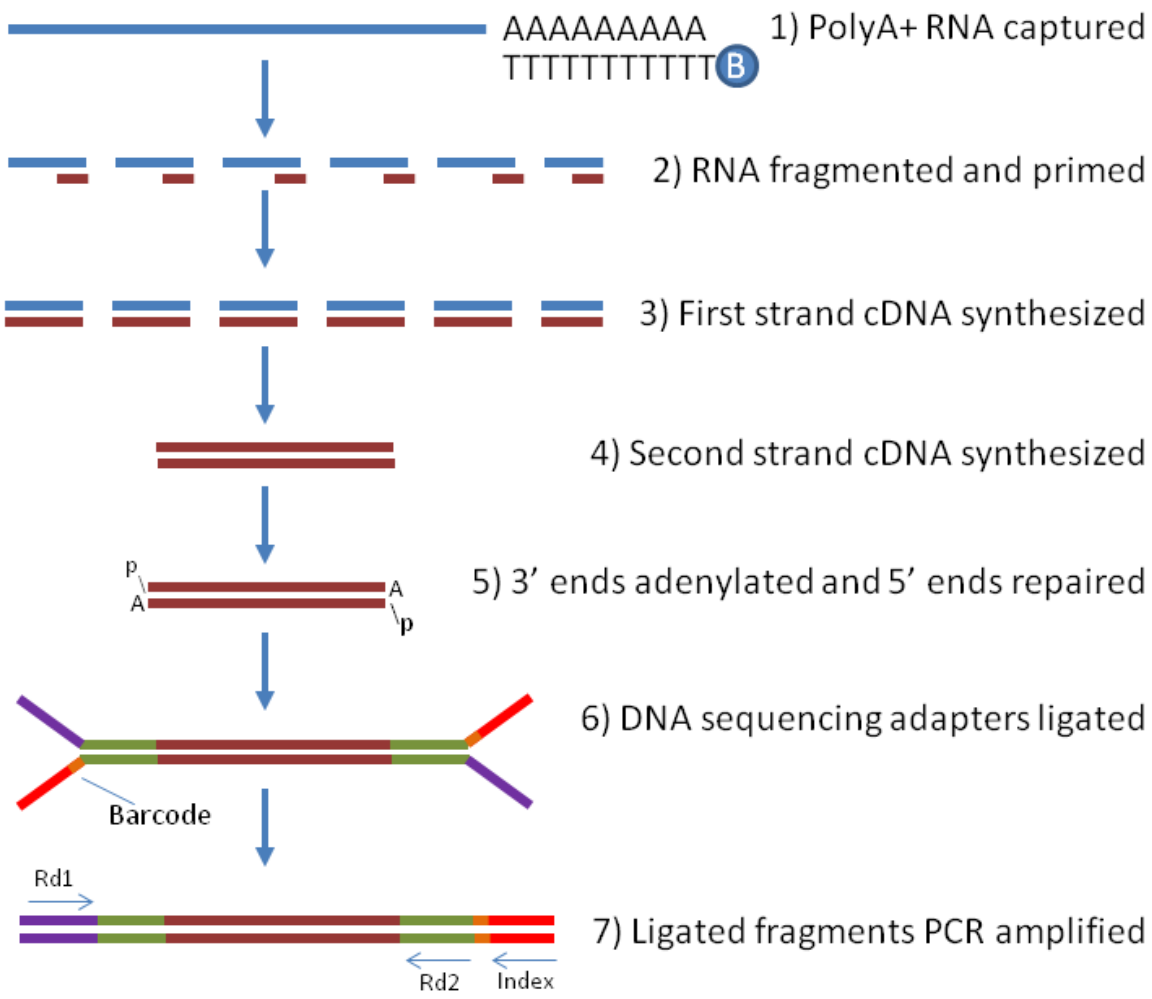
Introduction and background

Why use RNAseq?

- **Assembling** the gene space of a genome
- **Genotyping** individuals for variants that occur within the transcribed region of their genome
- **Quantify patterns of gene expression** across:
  - Organ, tissue, or cell types
  - Timepoints and development
  - Experimental treatments or observational categories

Introduction and background

How is RNAseq data generated?



SUMMARY: mRNA is isolated, fragmented, and cDNA is synthesized and sequenced

Standard Illumina paired-end data will thus represent a snapshot of the mRNA present in your sample

All bioinformatics should keep in mind the biological origins of the data

Overview of methods

Quantifying patterns of gene expression:

1. RNAseq extraction protocol & sequencing
2. Clean and filter reads
3. Map reads to a reference
4. Count number of reads per gene in each individual
5. Statistical analysis of differences in read counts

Cleaning and filtering reads

- Cleaning and filtering should be done aggressively prior to running any transcriptome assemblies (**SnoWhite** pipeline is good)
- Mapping/aligning reads to a reference is more forgiving (bad quality reads won't align), but cleaning may give greater confidence in the results and will run faster

A note on terminology:

Mapping: placement of a read in the correct region of the reference

Alignment: detailed placement of each base in a read

Overview of methods

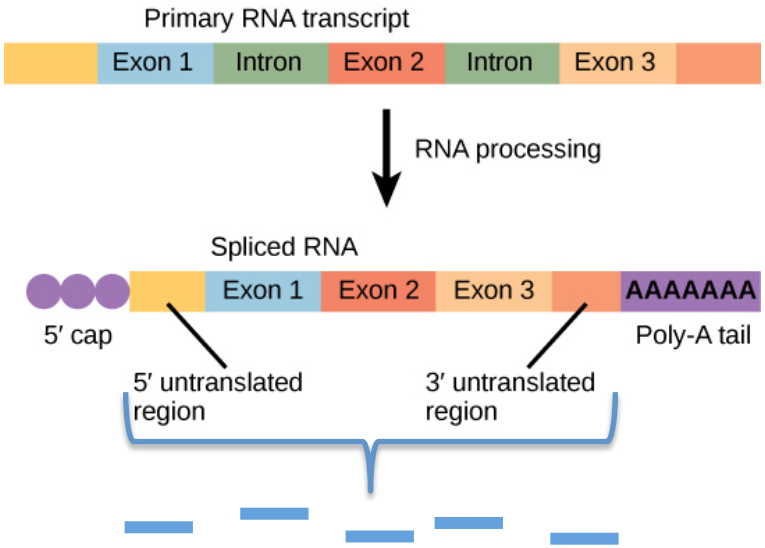
Quantifying patterns of gene expression:

1. RNAseq extraction protocol & sequencing
2. Clean and filter reads
3. Map reads to a reference
4. Count number of reads per gene in each individual
5. Statistical analysis of differences in read counts

Quantifying expression levels

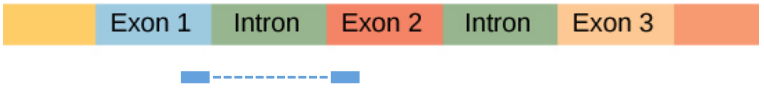
Map filtered reads to a reference:

```
@HWI-ST521:81:C0HKCACXX:1:2106:4602:62241 1:N:0:CGATGT
GAGTCGTCGGCCTCTGCCTTCAGGCGGGACTCTGGGAAGCGGAATTATGGACGCGATGAGAGCCAGTCTGCAAGTGCCAAGCGAGACAAAGAAACAGACA
+
CCCDFFFFHHHHHJIJJJJJJJJJJJJJ<EEHFHHHFFEDDDBB@CDEDDDDDDDBDDDDDDDCDCDDDDDDC>ACDDDDCDDDBDDDDCDDDDDCAAB
@HWI-ST521:81:C0HKCACXX:1:1207:4542:126997 1:N:0:CGATGT
TCCACGACTATCAATGAAAGTAATGCCTTCCAAATGTCCAACTAGAGTTCCAGCAGGCTGTGCCCTAGATGTAAGGCATCGCCTATCCCAAACCTTGCAA
+
CCCFFFFFHHHHHJJJJJJIFHHIJJJJJJJJJJJIIJIJGIGGIJJGJJJJJIIIHGGGFIJJJJIJIIHHHHHGFDEDFDDDDCDDDDDDDDDDDDC
@HWI-ST521:81:C0HKCACXX:1:1101:14997:93997 1:N:0:CGATGT
TGAAATCCCAAACTGAGCCGAATATTCCGATAACCAGCAAAAGCTTGGCTGAAACTCCAGTAATCGATCTCTTCAATGAAGCAAATGGCATCAGAAAGAC
```



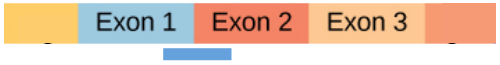
Our filtered RNAseq reads come from the mature transcript

The genome sequence looks like this:



Read fails to map across intron

The transcriptome sequence looks like this:



Read maps correctly

Some reads span two exons, and would not map to the genome using conventional approaches

Quantifying expression levels

Challenge #1: Mapping reads across intron-exon boundaries

Solutions:

- Map reads to a transcriptome (e.g. RSEM)
- Two-stage mapping to the genome (e.g. TopHat)
  - Use an "unspliced read aligner" to map the reads within a single exon
  - Split unmapped reads into shorter segments and attempt to re-map
- "Seed extension" methods map small chunks to the genome and extend to junctions (e.g. GSNAP)

Quantifying expression levels

Differential expression of alternatively spliced transcripts?

Splice variant #1

Splice variant #2

If there are two known splice variants, a read spanning exon **1** & **2** or **1** & **3** will identify which variant is present

Splice variant #1

Splice variant #2

Alternatively, if a read aligns to exon **2** then differential expression of isoforms can be inferred, relative to the expression levels of other isoforms

Quantifying expression levels

Challenge #2: Identifying abundance of alternatively spliced transcripts

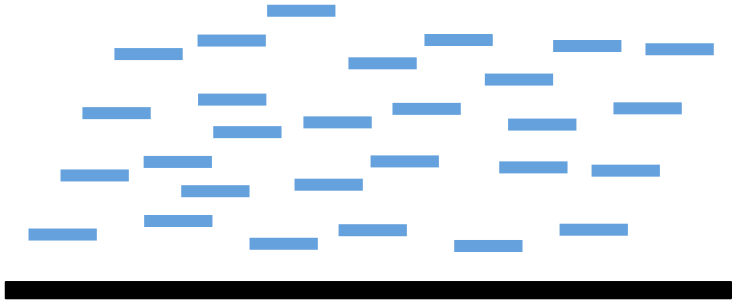
Solutions:

- Identify expression levels for reads spanning diagnostic splice sites, relative to expression levels in non-diagnostic exons
- Multiple complex algorithms for sorting reads based on compatibility with different isoform models (e.g. Cufflinks, etc.)

Quantifying expression levels

Map filtered reads back to the transcriptome and count the number of reads that align to each contig

28 reads align to the contig



Transcriptome contig
825 base pairs long

What if there is a paralog?

Partially paralogous gene

Both paralogs and alternatively spliced transcripts (isoforms) can give the problem of "multireads": a read that maps with high score to several places

Li et al. (2010) found that 17% (mouse) or 52% (maize) of reads were multireads

Quantifying expression levels

Challenge #3: Dealing with multireads at the gene- and isoform-level

Solutions:

- Discard them and estimate expression from only uniquely mapping reads
- "rescue" multireads by allocating fractions of them to each contig, in proportion to the number of uniquely mapping reads mapping to each contig
- Maximum-likelihood algorithms to assign multireads and sum across all isoforms to get a gene-level estimate (e.g. RSEM)

Quantifying expression levels

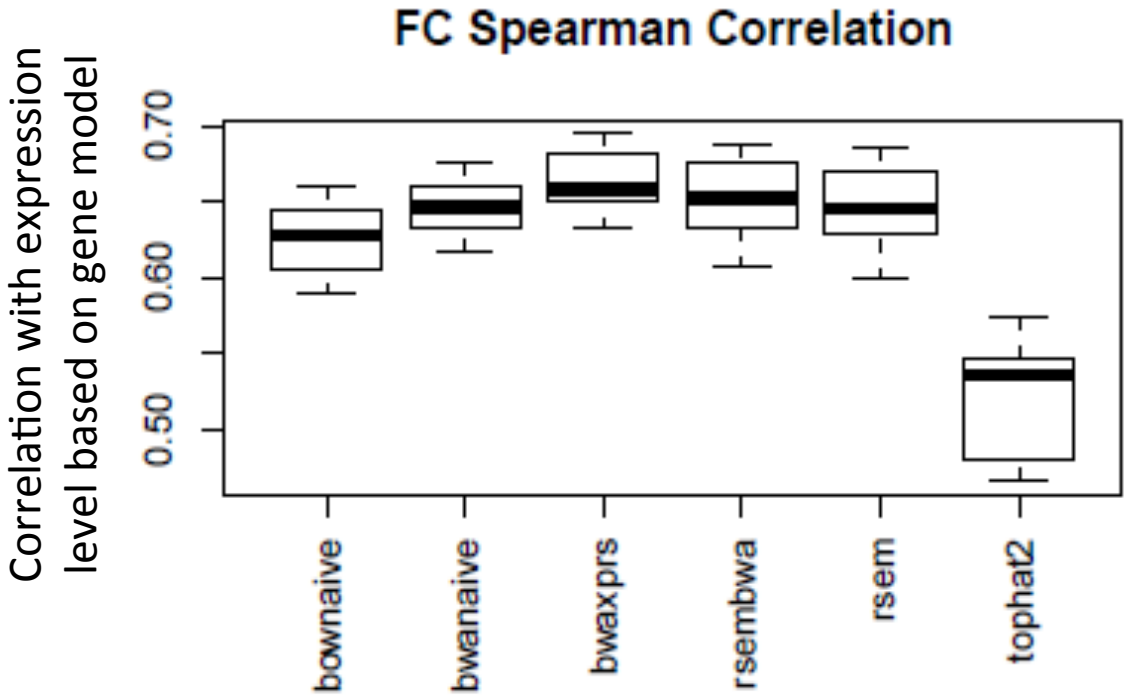
Practical approaches: RSEM

- RSEM provides a single pipeline to align sequence reads and estimate expression counts for each contig
- When used in conjunction with a Trinity-built transcriptome assembly, it will estimate isoform-level expression counts
- In contrast to TopHat and other approaches, it does not require a sequenced genome, and reasonable reference transcriptomes can be built *de novo* using Trinity non-model organisms
- In the exercise following lecture, we will work through a simple example dataset with RSEM

Quantifying expression levels

Practical approaches: TopHat + Cufflinks

- TopHat + Cufflinks provide a joint approach to mapping reads to the genome and require a good reference genome
- Tophat may be less accurate than RSEM:



Courtesy of Eric Aronesty

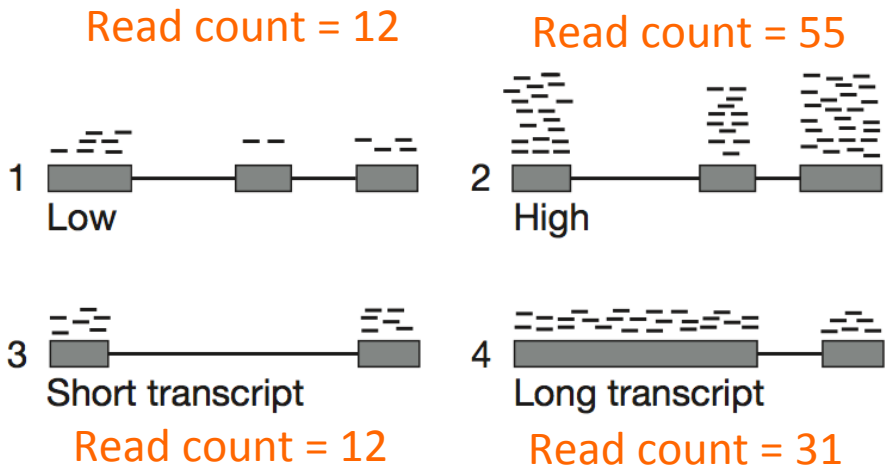
Overview of methods

Quantifying patterns of gene expression:

1. RNAseq extraction protocol & sequencing
2. Clean and filter reads
3. Map reads to a reference
4. Count number of reads per gene in each individual
5. Statistical analysis of differences in read counts

Analyzing patterns of expression

RNAseq data is highly fragmented, so there are more reads from a long transcript than from a short transcript:

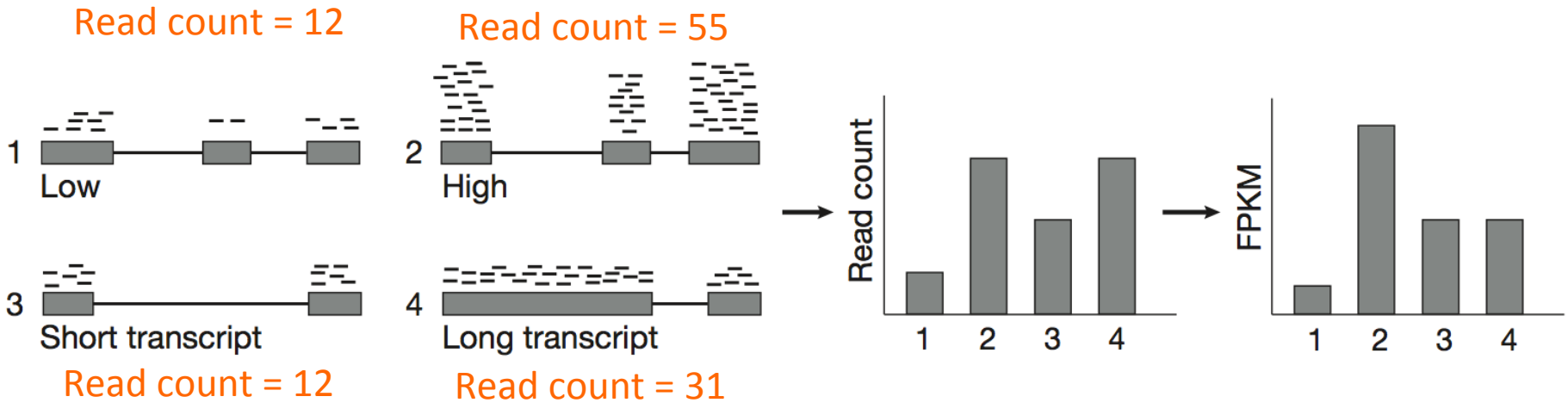


Also, some individuals have more data sequenced than others, giving higher total expression counts, even for the same amount of expression

Garber et al. 2011

Analyzing patterns of expression

RNAseq data is highly fragmented, so there are more reads from a long transcript than from a short transcript:



FPKM: Fragments **P**er **K**ilobase of transcript per **M**illion reads mapped, corrects both of these issues by normalizing by both transcript length and the total size of the mapped library

Garber et al. 2011

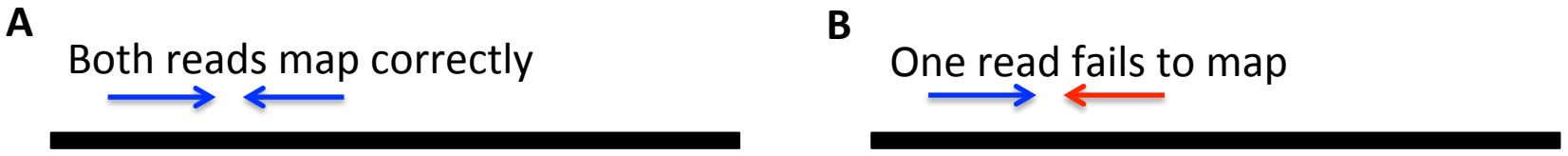
Analyzing patterns of expression

RPKM vs. FPKM

FPKM: Fragments **P**er **K**ilobase of transcript per **M**illion reads mapped

RPKM: Reads **P**er **K**ilobase of transcript per **M**illion reads mapped

FPKM corrects for the non-independence of two reads when you have paired-end data:



RPKM would count that A had 2x more expression than B, giving an underestimate for B. FPKM adjusts this count for paired end data

Analyzing patterns of expression

Practical implementation

Simple: most programs will estimate FPKM or RPKM for you

Sample output from RSEM

gene_id	transcript_id(s)	length	effective_length	expected_count	TPM	FPKM	
comp10000_c0	comp10000_c0_seq1	1502	1299.85	3	60.1	34.36	
comp100017_c0	comp100017_c0_seq1	735	532.87	1	48.87	27.94	
comp10002_c0	comp10002_c0_seq1	4182	3979.85	7	45.8	26.19	
comp100037_c0	comp100037_c0_seq1	1921	1718.85	0	0	0	
comp100052_c0	comp100052_c0_seq1	679	476.89	0	0	0	
comp10005_c0	comp10005_c0_seq1	1764	1561.85	0	0	0	
comp100064_c0	comp100064_c0_seq1	631	428.92	0	0	0	
comp10006_c0	comp10006_c0_seq1	2680	2477.85	4	42.04	24.04	

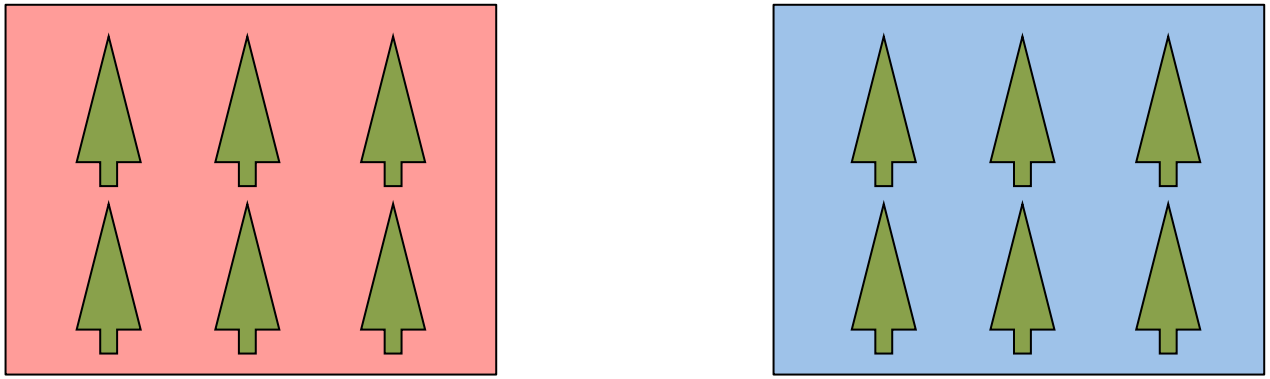
Overview of methods

Quantifying patterns of gene expression:

1. RNAseq extraction protocol & sequencing
2. Clean and filter reads
3. Map reads to a reference
4. Count number of reads per gene in each individual
5. Statistical analysis of differences in read counts

Analyzing patterns of expression

Fitting models to expression data



We wish to study differences in gene expression between seedlings that were exposed to hot vs. cool conditions

We have 2 treatments (hot, cold) and 6 individuals per treatment, and we sequence one library from each individual

Analyzing patterns of expression

How to go from raw expression counts

comp10109_c2	0.00	0.00	0.00	0.00
comp10109_c20	0.00	0.00	0.00	0.00
comp10109_c22	176.00	13.00	5.00	9.00
comp10109_c23	0.00	0.00	0.00	0.00
comp10109_c25	0.00	0.00	2.00	2.00
comp10109_c31	0.00	0.00	0.00	0.00
comp10109_c32	0.00	0.00	0.00	0.00
comp10109_c33	1.00	0.00	0.00	0.00
comp10109_c35	148.00	403.87	327.20	117.14
comp10109_c36	0.00	0.00	0.00	0.00
comp10109_c37	0.00	0.00	0.00	0.00
comp10109_c38	1.00	1.00	0.00	0.00
comp10109_c40	0.00	0.00	0.00	0.00
comp10109_c41	96.00	51.00	61.00	24.00
comp10109_c42	15.00	0.00	0.00	1.00
comp10109_c7	0.00	0.00	0.00	0.00
comp1010_c0	483.00	2125.91	2397.11	526.00

To biologically meaningful results?

Analyzing patterns of expression

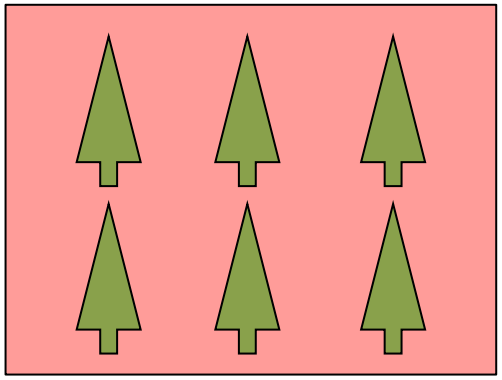
Two broad approaches to analysis:

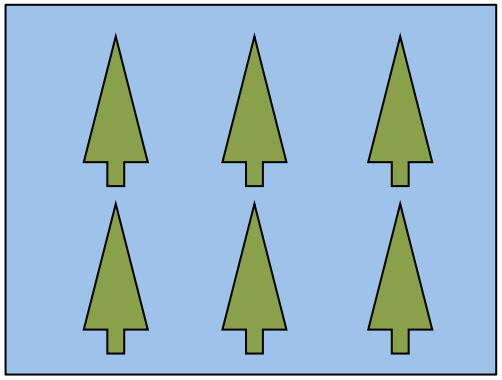
1. Analysis of differential gene expression on gene-by-gene basis (e.g. DESeq, EdgeR, limma)
   - Examine how each gene is affected by a given treatment
   - Use ANOVA or other similar statistical tools to identify genes where significantly more expression variance is partitioned among than within treatments.

2. Analysis of patterns of gene co-expression and identification of clusters of genes that have similar patterns of expression

   - Identify clusters of genes that are upregulated in treatment X and downregulated in treatment Y

Identifying differential gene expression

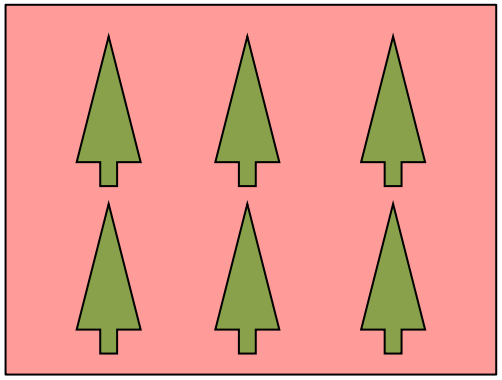


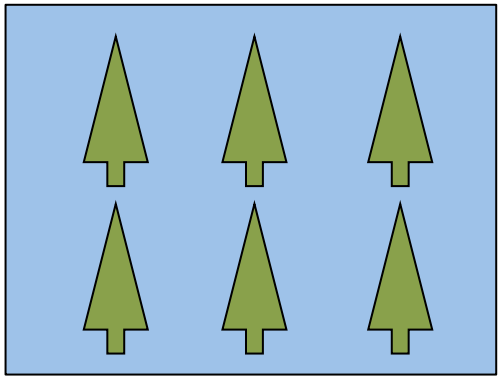


General approach:

- Data has **biological replication** (hot vs. cold) **and technical replication** (here, there is no technical replication)

Identifying differential gene expression

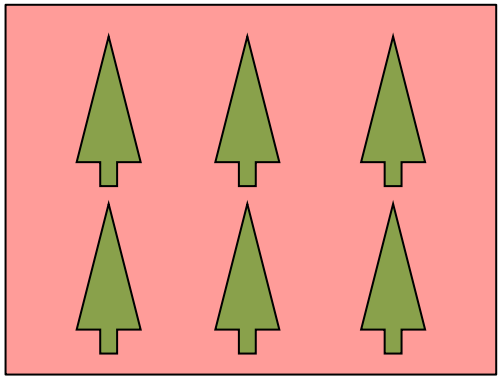


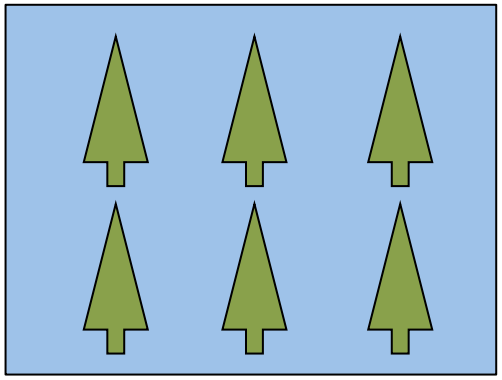


A side-note on sources of variation in RNAseq

- *Biological variation* in expression arises from real differences between samples, due to either uncontrolled sources that should be homogenous across all individuals or controlled sources that arise from experimental treatment/design.
- *Technical variation* in expression arises from measurement error inherent in the sequencing process. If we could sequence an infinite number of reads, technical error would disappear

Identifying differential gene expression





General approach:

- Data has **biological replication** (hot vs. cold) and **technical replication** (here, there is no technical replication)
- Regression of normalized counts on variable(s) of interest
- Can incorporate models representing the variance in expression counts due to both technical and biological replication
- Yields estimates of the **fold-change in expression** among factor levels and an estimate of significance

Identifying differential gene expression

Using the approach from edgeR as an example

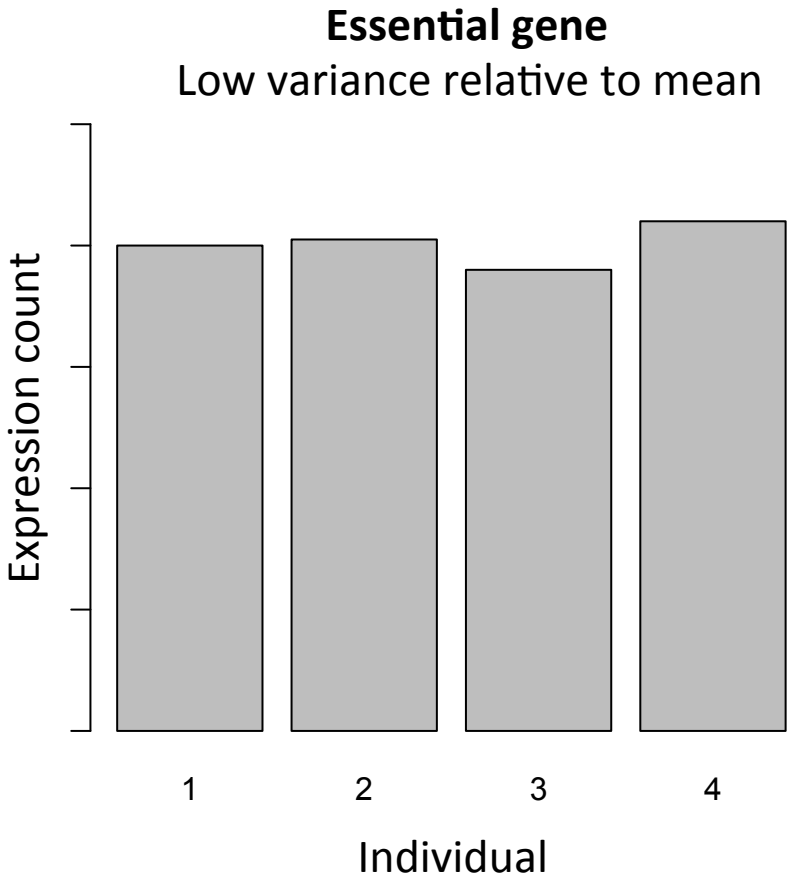
- Assumes that the expression count at a given locus can be modeled as a Poisson process where:

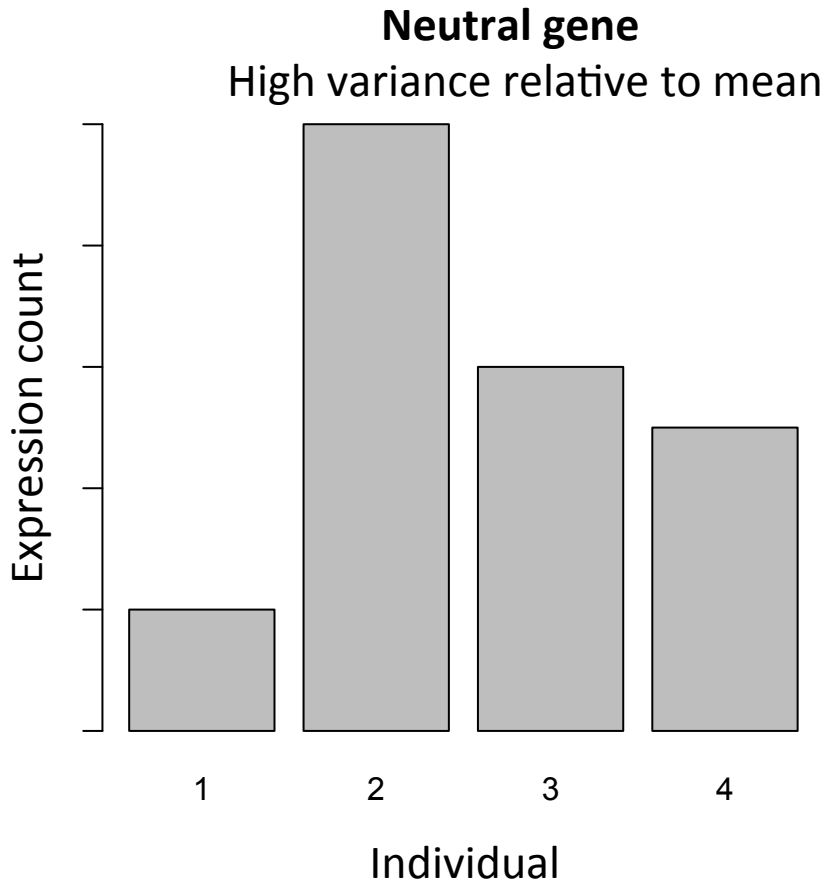
$$\text{Total } CV^2 \text{ in expression} = \text{Technical } CV^2 + \text{Biological } CV^2$$

- If a Poisson process has the **same Biologial CV** among genes, then the proportional relationship between gene-wise standard deviations and gene-wise means should be **the same for all genes**
- This simplifying assumption means that means and variances in expression can be estimated with great power across all genes and used in model fitting with General Linear Models (GLMs)

Identifying differential gene expression

However, real genes do not all behave identically. Some are much more constrained in their expression than others, because they are more critical to functioning





Identifying differential gene expression

Using the approach from edgeR as an example

- Therefore, the relationship between mean and variance in expression may vary among genes and accurate model fitting should reflect this
- EdgeR allows for two broad approaches to fitting GLM to data using a negative binomial model:
  - **common dispersion:** the relationship between mean and variance is estimated across all genes
  - **tagwise dispersion:** the common dispersion estimate is modified for each gene based on a bayesian estimate of the per-gene relationship between mean and variance
- Best practice is to fit a tagwise dispersion model after first estimating a common dispersion model.

Identifying differential gene expression

Using the approach from edgeR as an example

Model fitting results in estimation of log fold change (logFC) in expression, p-value, and estimation of False Discovery Rate (FDR)

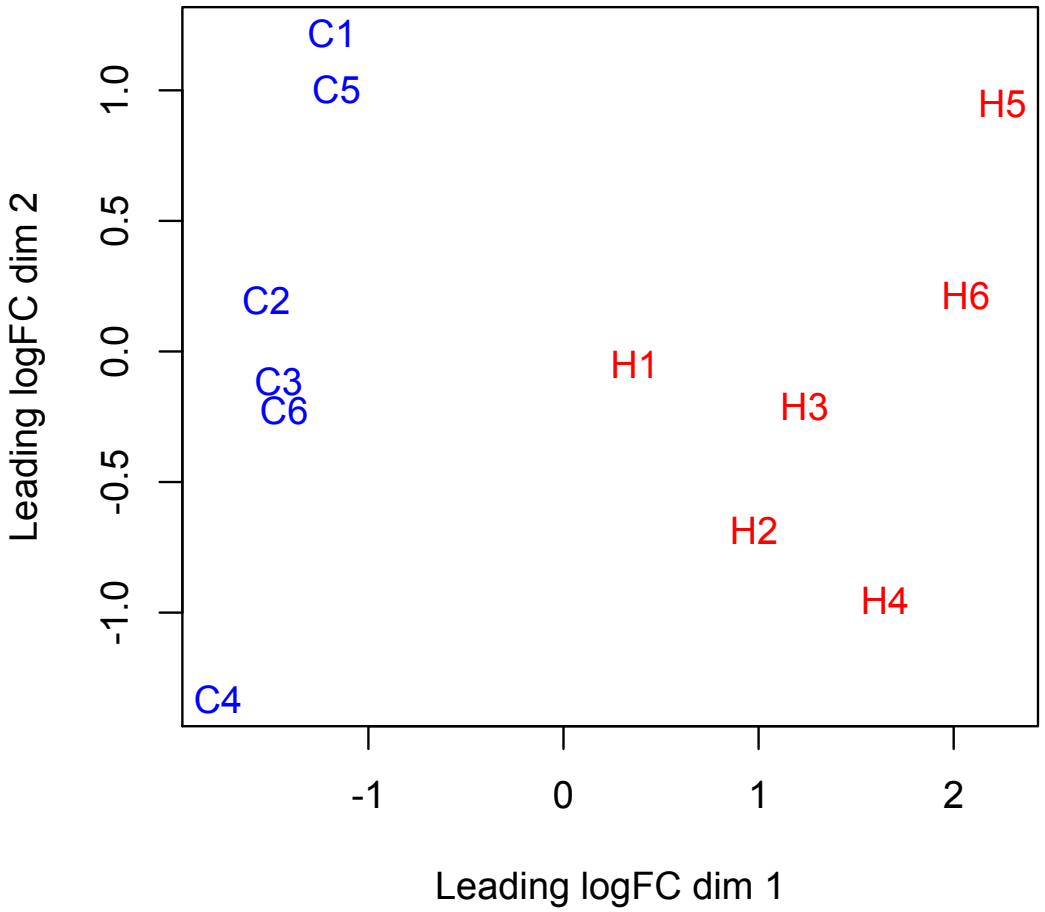
	logFC	logCPM	LR	PValue	FDR
comp520_c0	8.997022	10.663572	175.7591	4.087401e-40	7.584581e-36
comp626_c0	8.489396	8.474038	166.4056	4.510882e-38	4.185197e-34
comp29033_c0	-3.427787	2.914473	153.7321	2.650165e-35	1.639215e-31
comp3737_c0	4.121830	5.796822	134.5117	4.222342e-31	1.958744e-27
comp6840_c0	4.319808	5.063555	126.0793	2.954429e-29	1.023962e-25
comp14716_c0	-2.772885	5.115474	125.8532	3.310934e-29	1.023962e-25

Multiple approaches to fitting models, with and without intercepts

EdgeR allows multiple factors for more complex designs

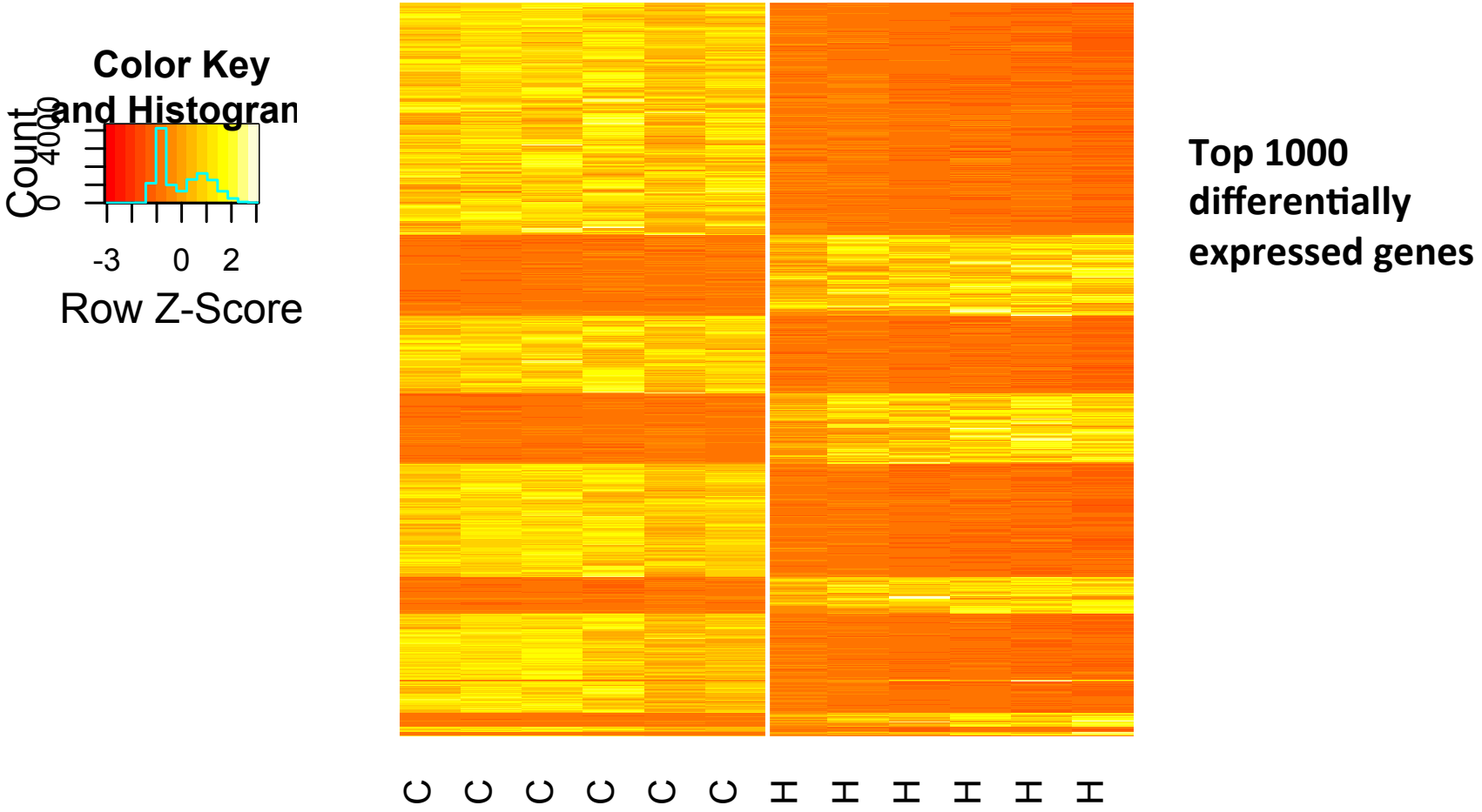
Identifying differential gene expression

Approaches to visualizing trends in data: Multi-Dimensional Scaling plot (like principle components, but allows missing data)



Identifying differential gene expression

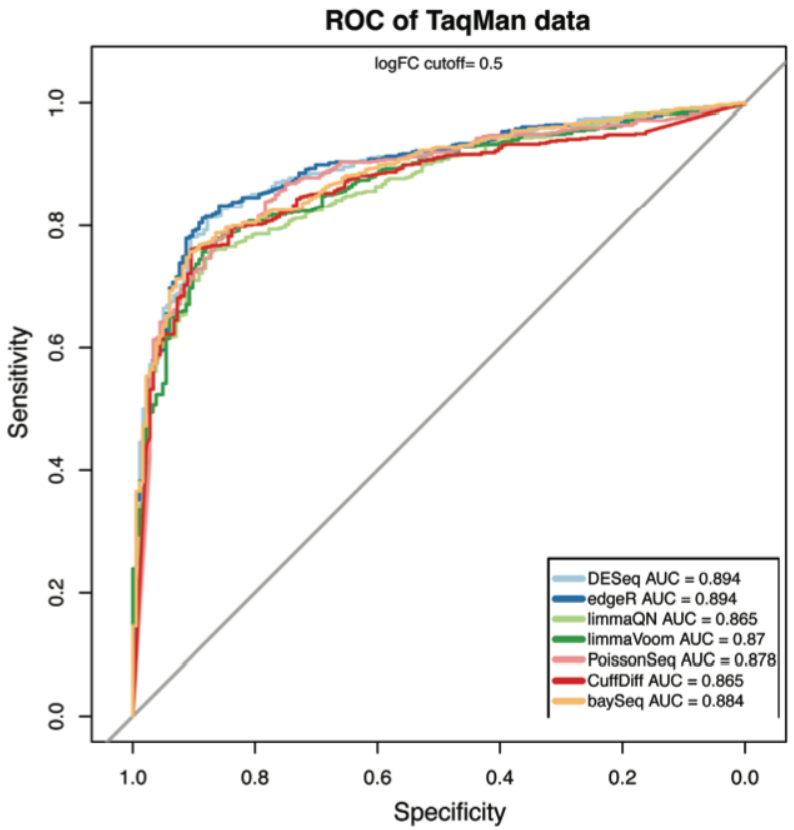
Approaches to visualizing trends in data: Heatmaps to show patterns of expression in the most differentially expressed genes



Identifying differential gene expression

Numerous programs have been developed to detect differences in gene expression:

- DESeq
- edgeR
- limmaQN
- limmaVoom
- PoissonSeq
- CuffDiff
- baySeq



Fortunately, they are relatively similar in their power and accuracy; edgeR is consistently found to slightly outperform many others

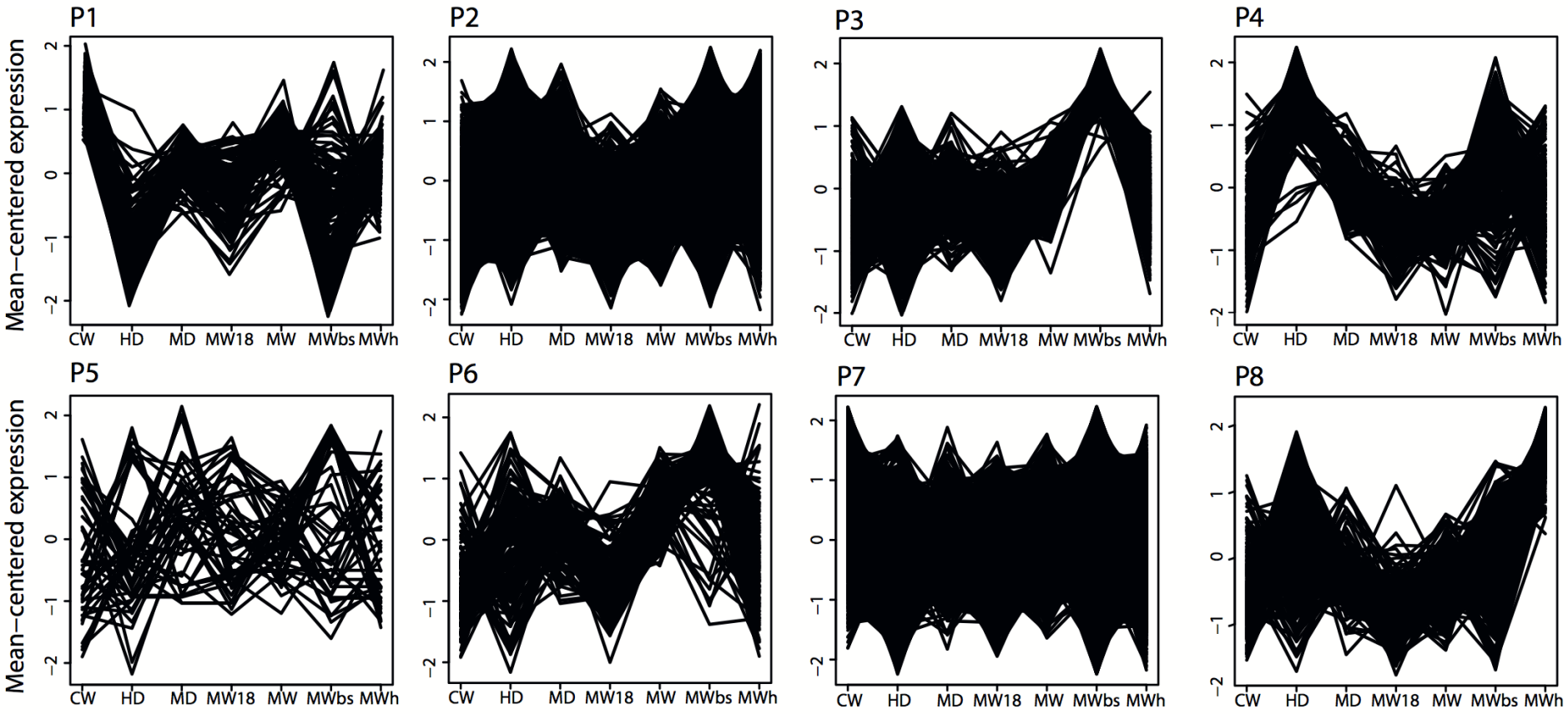
Gene co-expression networks

Genes that tend to be up-regulated and down-regulated together will have higher correlation in their expression counts across treatments:

- Calculate pairwise correlations between each gene
- Perform clustering algorithm on the correlation table, grouping like with like
- Can also group genes that have opposite patterns of expression
- Requires many treatments to get high power

Gene co-expression networks

Example (from WGCNA): 8 clusters showing gene expression in lodgepole pine over 7 treatments



Yeaman et al. (2014)

Now what?

- As with many approaches in genomics, there is a “too much data” problem
- Annotation of genes with extraordinary patterns and comparison with other species can help
- Useful for identification of genes involved in plasticity and response: are these genes also involved in adaptation? Do they have signatures of selection?
- Strong experimental design necessary to go from purely descriptive to insightful

Technical considerations

Depth of coverage

- Highly dependent upon study organism and transcriptome size
- Little power to detect changes in expression when < 50 counts per million per gene
- Too many individuals per lane can increase your technical variation

Technical considerations

- Single-cell sequencing has found considerable variation among cells of the same type sampled at the same time. Pooled cell represents an “average” snapshot
- Ablation methods and micro-dissection have also found substantial variation among cell types of the same tissue
- No substitute for biological replication
- Important that replicates be randomized or blocked by sequencing lane due to lane effects

Technical considerations

De novo assembly

- De novo assembly is quite feasible using Trinity but requires large amounts of RAM
- Lodgepole pine transcriptome assembly with 40Gbp of pooled sequence data took 200 GB of RAM
- Pooling samples from multiple tissues will yield greater increases in number of transcripts assembled than different growing conditions (in conifers, at least)
- Haploid tissue from a single individual is best
- Feasible to pool data from multiple individuals but difficult to know whether putative isoforms are “good” or just different genotypes

References & further reading

Garber et al. 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. Nature Methods. 8:469-477.

Marinov et al. 2014. From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. Genome Research. 24:496–510.

Rapaport et al. 2013. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biology. 14:R95.

Seyednasrollah et al. 2013. Comparison of software packages for detecting differential expression in RNA-seq studies. Briefings in Bioinformatics.

Tarazona et al. 2011. Differential expression in RNA-seq: A matter of depth. Genome Res. 21: 2213-2223

http://www.labome.com/method/RNA-seq-Using-Next-Generation-Sequencing.html

http://deweylab.biostat.wisc.edu/rsem/

http://www.mi.fu-berlin.de/wiki/pub/ABI/GenomicsLecture12Materials/rnaseq1.pdf

http://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf