UBC Bioinformatics Workshop

Topic 6: RNAseq and analysis of differential gene expression

Lecture outcomes

Explain how RNAseq is generated and used

Identify the basic steps to align and analyze RNAseq data

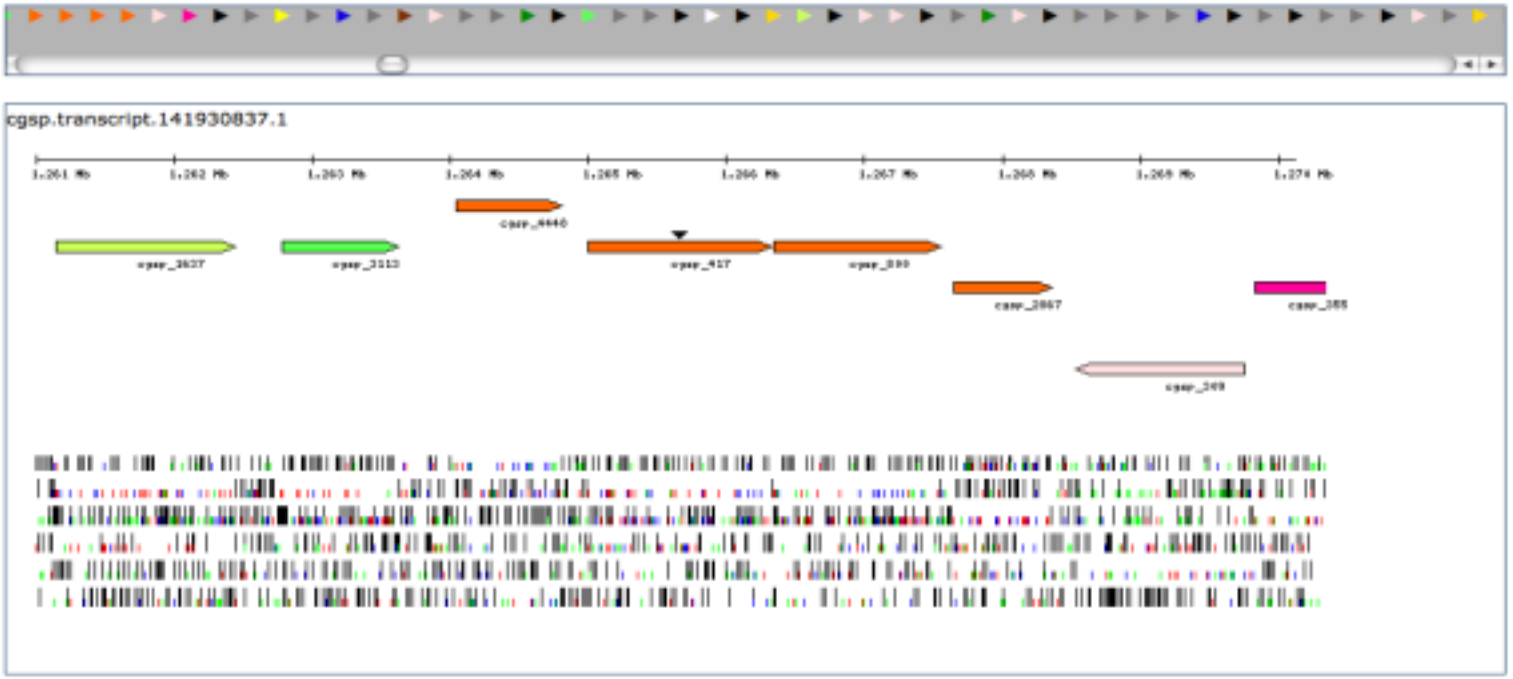
Outline

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?

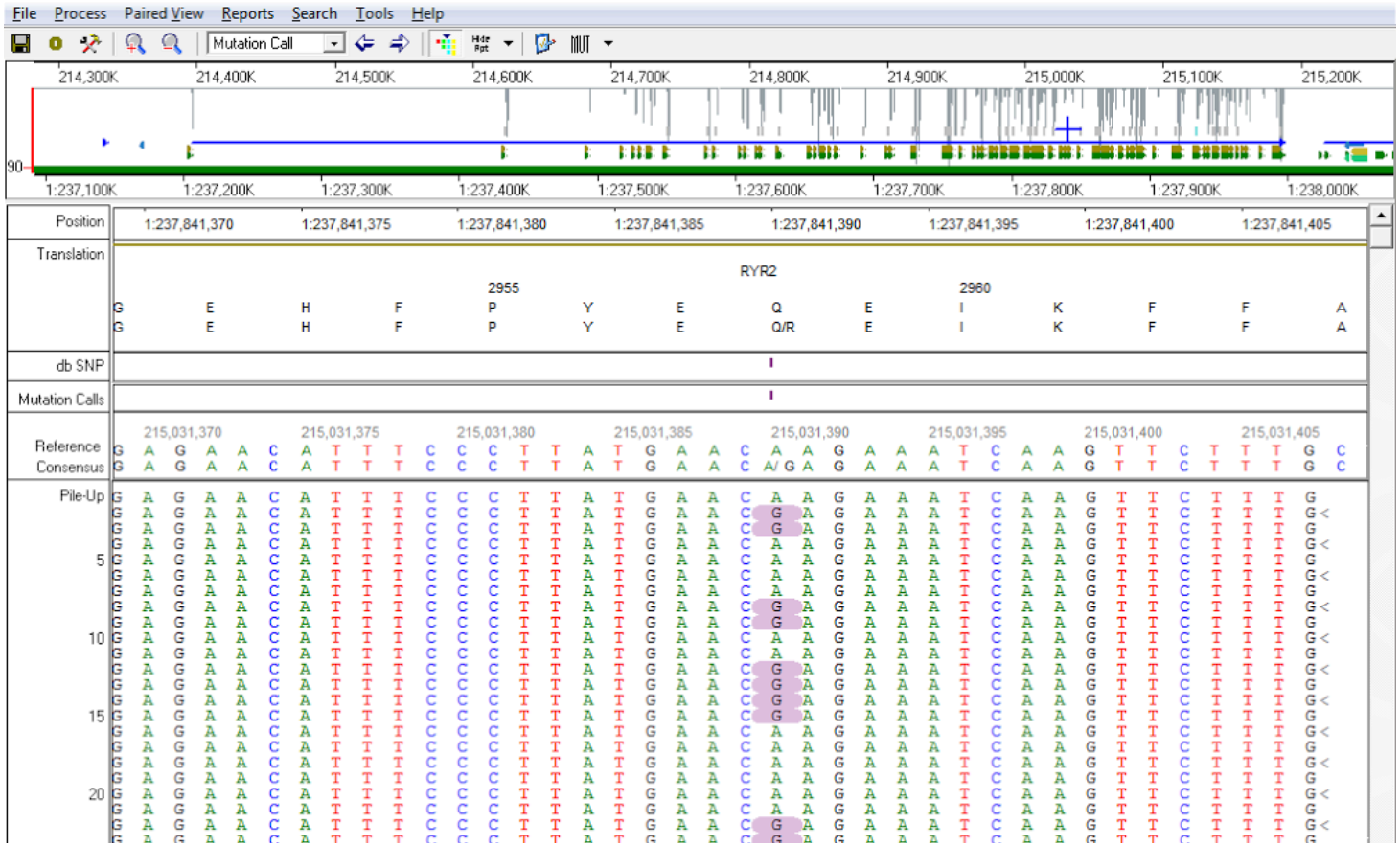
- Assembling gene space and genome annotation



Introduction and background

Why use RNAseq?

- Genotyping within the transcribed regions



Introduction and background

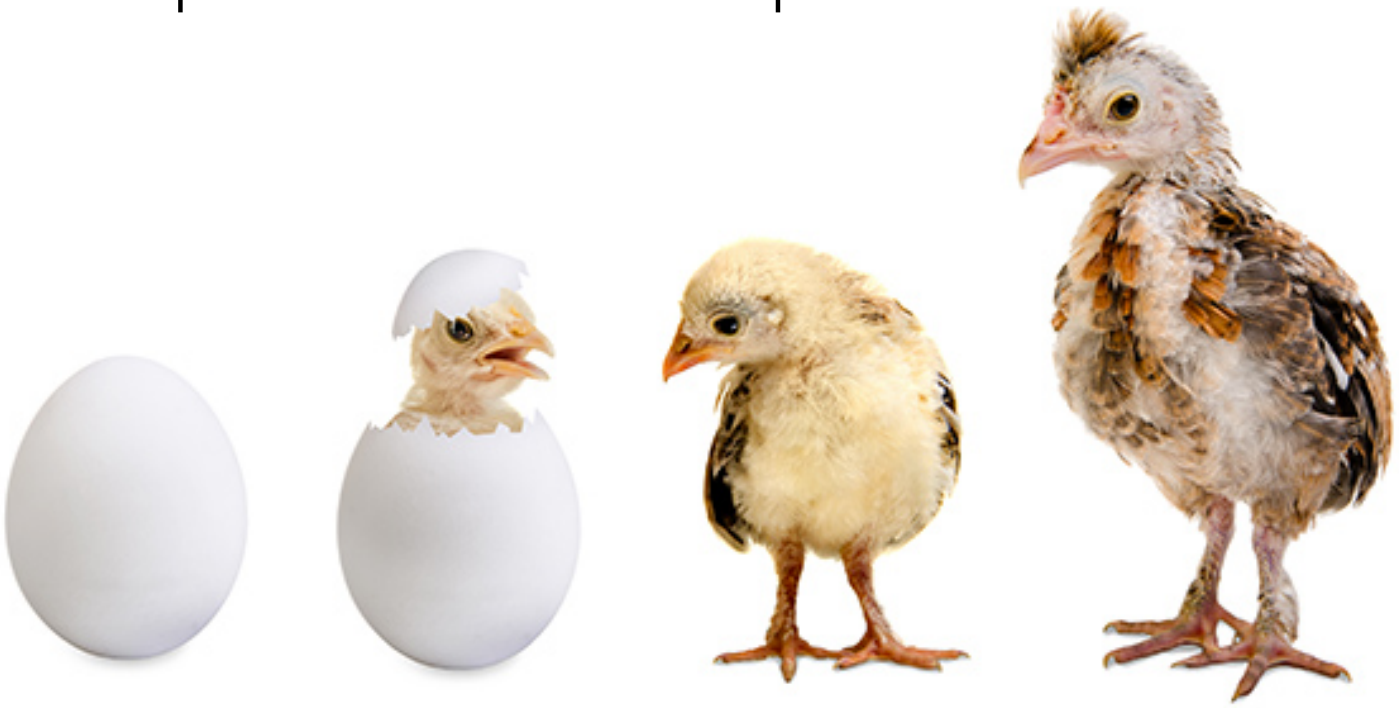
Why use RNAseq?

- Quantify patterns of gene expression
  - Organ, tissue, or cell types

Introduction and background

Why use RNAseq?

- Quantify patterns of gene expression
  - Timepoints and development



Introduction and background

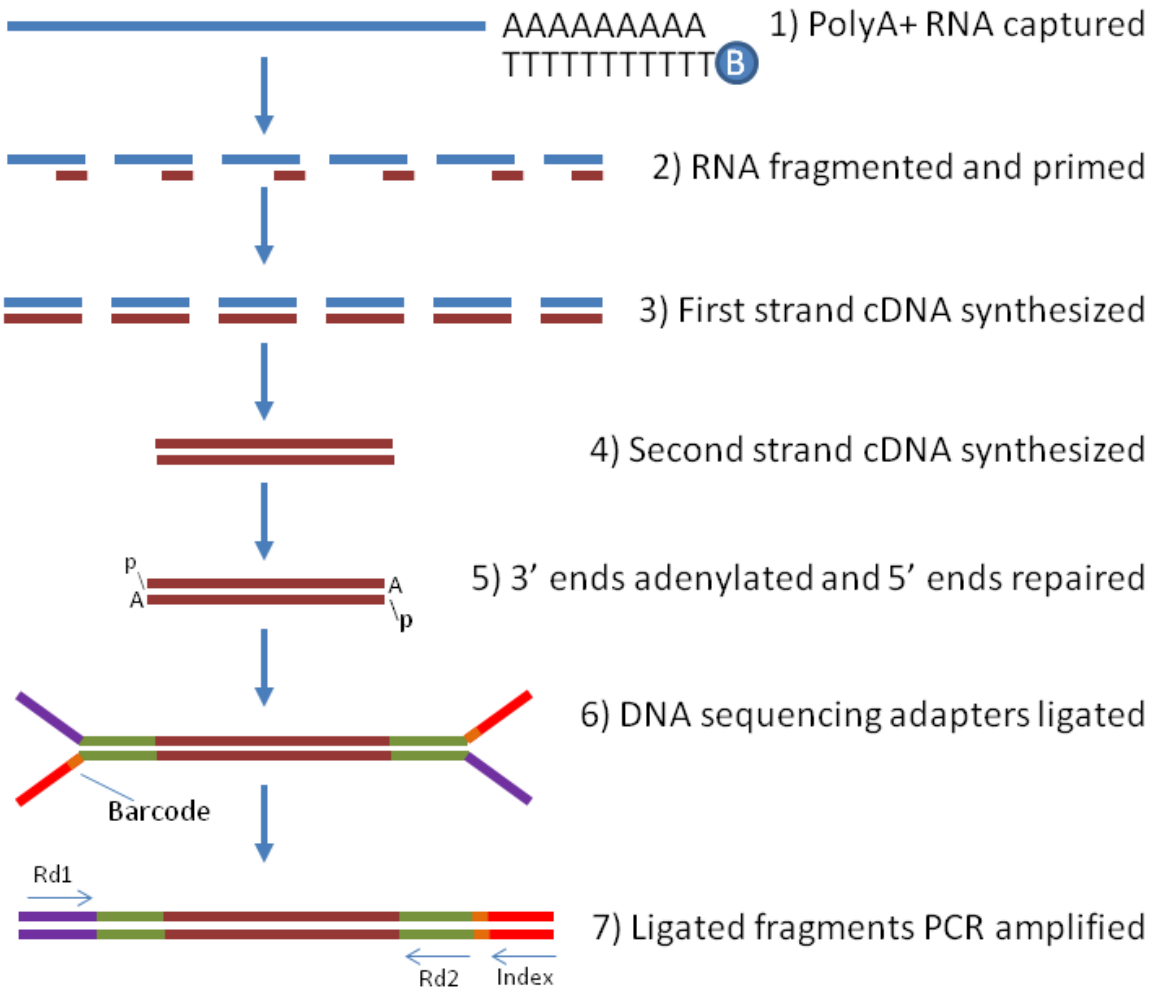
Why use RNAseq?

- Quantify patterns of gene expression
  - Experimental treatments or observational categories

Introduction and background

How is RNAseq data generated?

- 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



http://aws.labome.com/

Overview of the 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

Overview of the methods

- Cleaning and filtering is important for de novo transcriptome assembly and for SNP calling (SnoWhite pipeline is good)

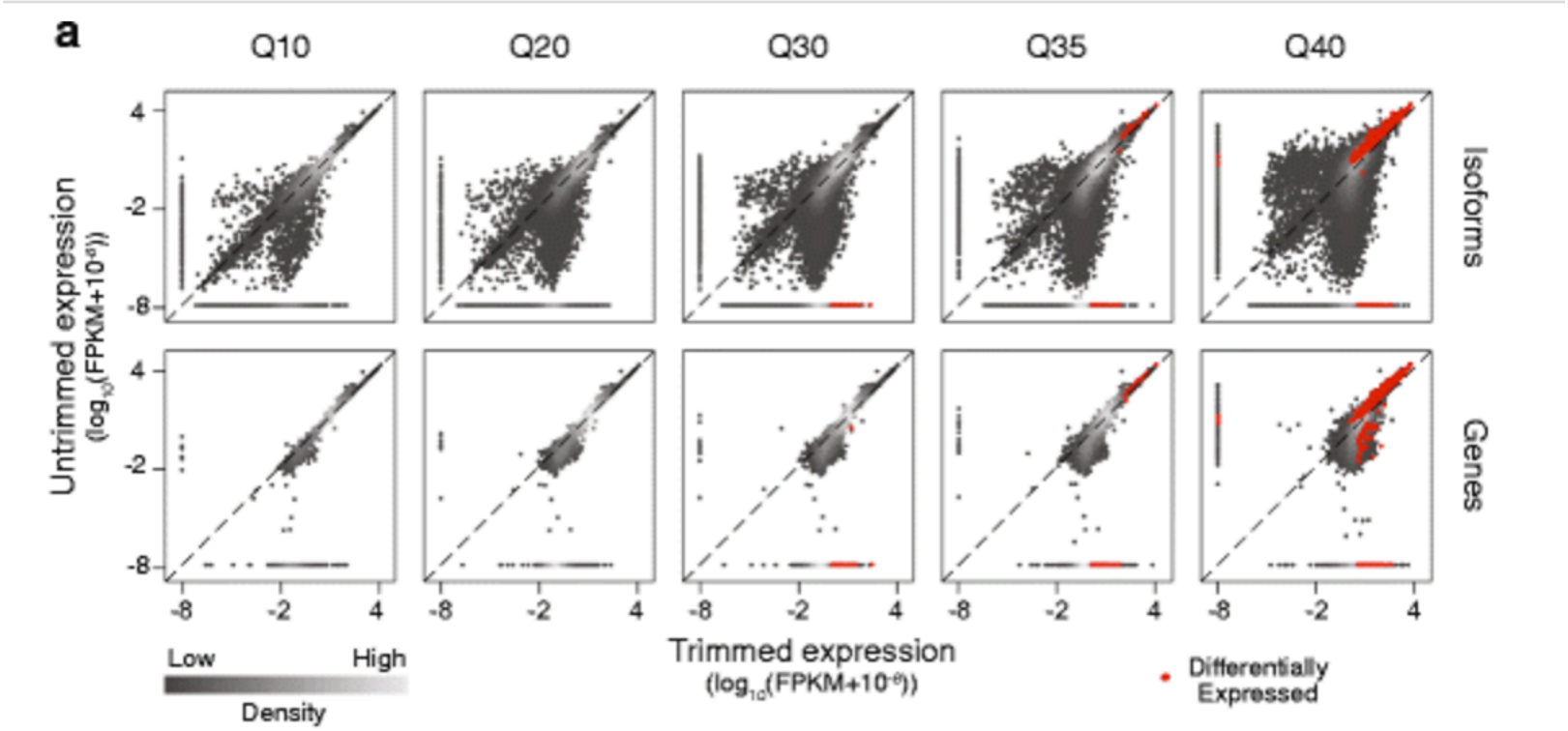
```
@HWI-ST521:81:C0HKCACXX:1:2106:4602:62241 1:N:0:CGATGT
GAGTCGTCGGCCTCTGCCTTCAGGCGGGACTCTGGGAAGCGGAATTATGGACGCGATGAGAGCCAGT
+
CCCDFFFFHHHHHJIJJJJJJJJJJJJJ<EEHFHHHFFEDDDBB@CDEDDDDDDDBDDDDDDDCDC
@HWI-ST521:81:C0HKCACXX:1:1207:4542:126997 1:N:0:CGATGT
TCCACGACTATCAATGAAAGTAATGCCTTCCAAATGTCCAACTAGAGTTCCAGCAGGCTGTGCCCTA
+
CCCFFFFFHHHHHJJJJJJIFHHIJJJJJJJJJJJIIJIJGIGGIJJGJJJJJIIIHGGGFIJJJJI
@HWI-ST521:81:C0HKCACXX:1:1101:14997:93997 1:N:0:CGATGT
TGAAATCCCAAACTGAGCCGAATATTCCGATAACCAGCAAAAGCTTGGCTGAAACTCCAGTAATCGA
```

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

Alignment: detailed placement of each base in a read

Overview of the methods

- Aggressive trimming and spurious alignments of short reads can lead to inaccurate estimates of gene expression



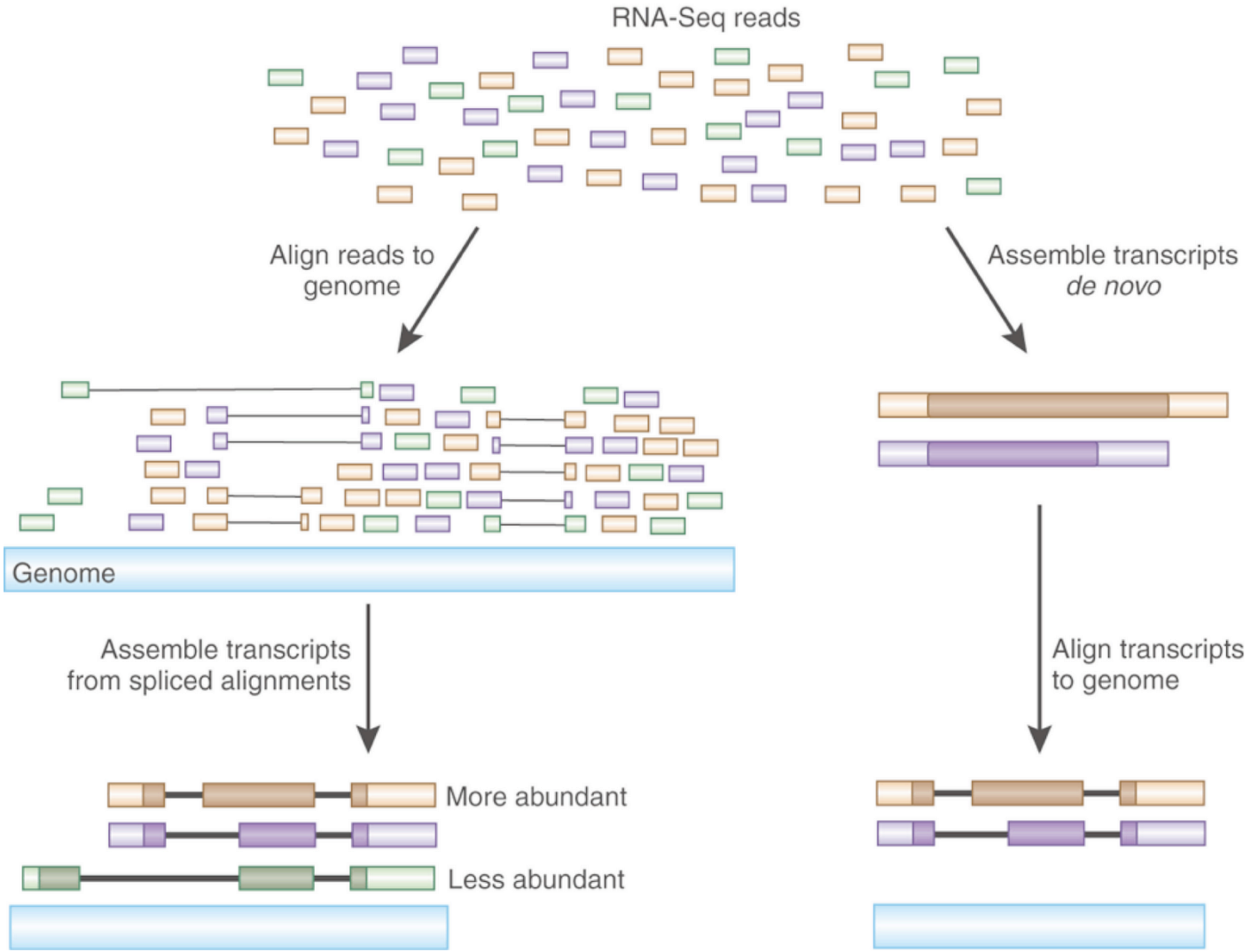
Williams et al 2016

Overview of the 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

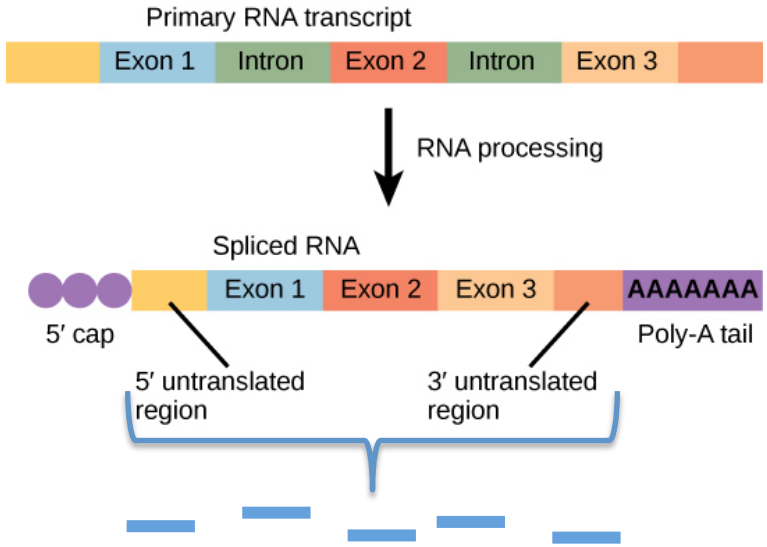


Assembling and Aligning

Haas & Zody 2010 Nature Biotechnology

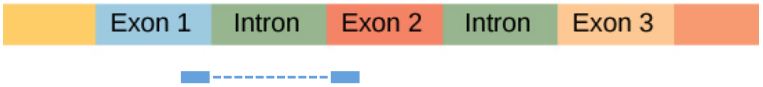
Quantifying expression levels

Challenge 1: Mapping reads across intron-exon boundaries



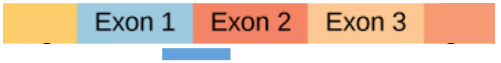
Our filtered RNAseq reads come from the mature transcript

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

http://philschatz.com/biology-concepts-book/

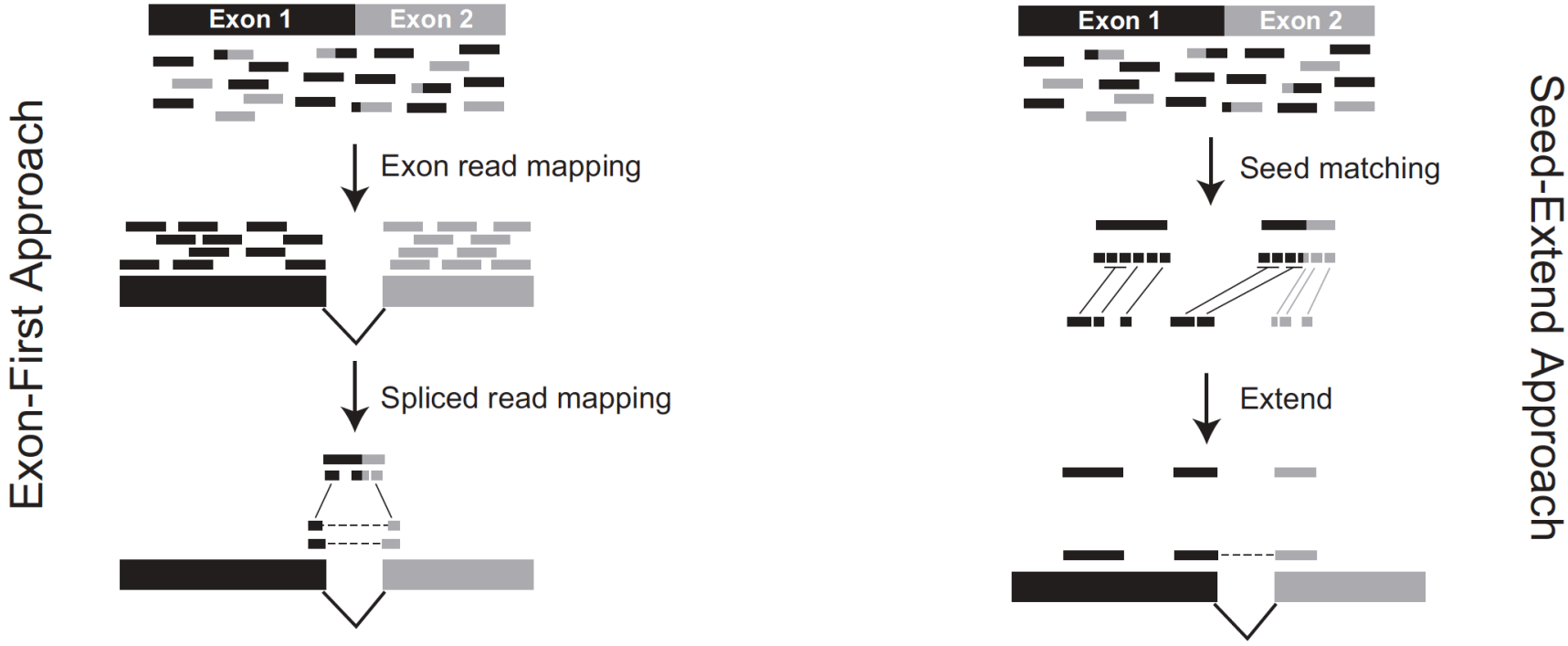
Quantifying expression levels

Challenge 1: Mapping reads across intron-exon boundaries

Solutions:

- Map reads to a transcriptome (e.g. RSEM)
- Exon first mapping to 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 and extend methods map small chunks to the genome and extend to splice sites (e.g. GSNAP)

Quantifying expression levels



Quantifying expression levels

Challenge 2: Identifying abundance 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

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

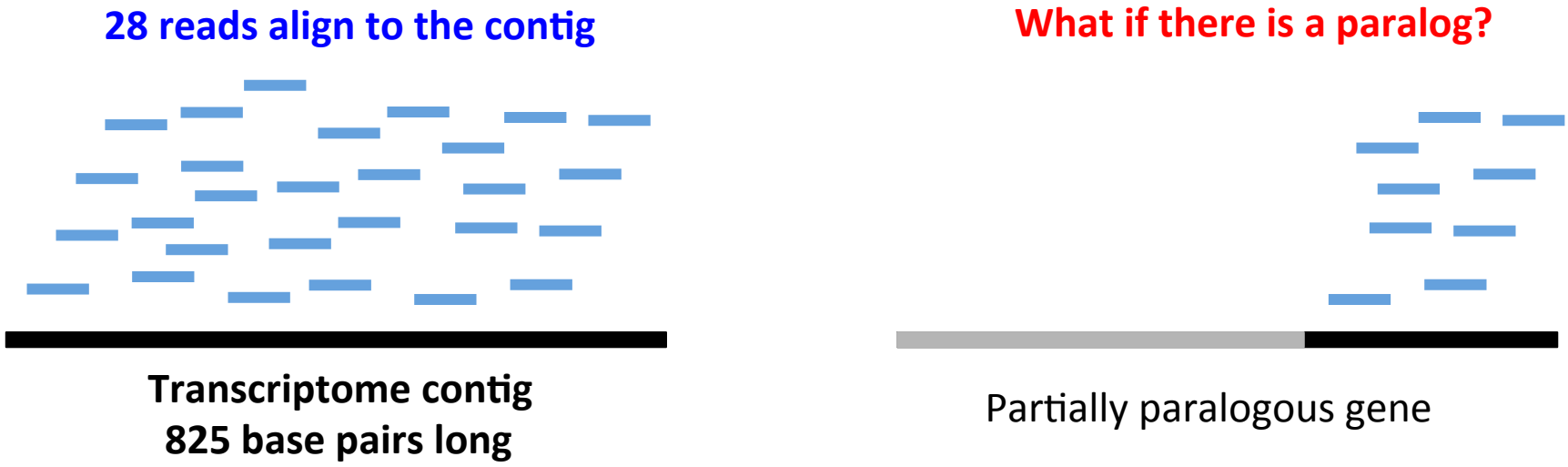
Challenge 2: Identifying abundance of alternatively spliced transcripts

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)

Quantifying expression levels

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



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 (only use uniquely mapping reads)
- “rescue” multireads by allocating fractions of them in proportion to the number of uniquely mapping reads mapping to each contig
- ML algorithms to assign multireads and sum across all isoforms for gene-level estimates (e.g. RSEM)

Quantifying expression levels

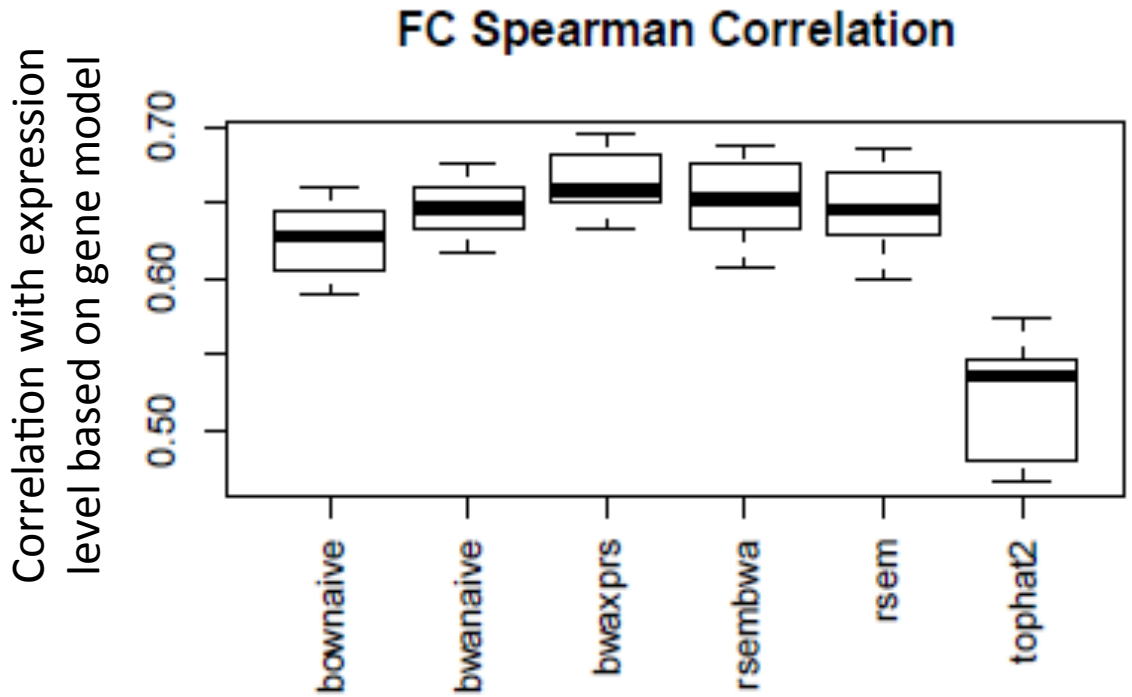
Practical approaches: RSEM

- Single pipeline to align and estimate expression
- Will estimate isoform-level expression counts (if isoforms for each gene are identified)
- No sequenced genome needed (a reasonable reference transcriptome can be built de novo using Trinity in 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 the methods

Quantifying patterns of gene expression:

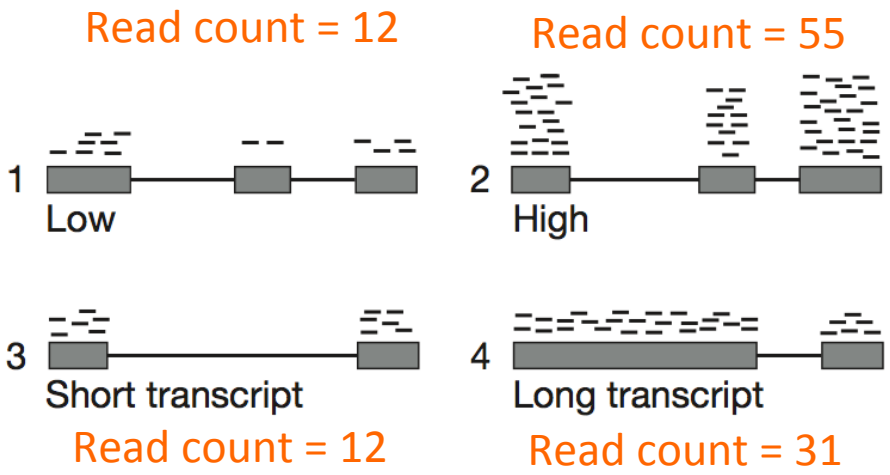
1. RNAseq extraction protocol & sequencing
2. Clean and filter reads
3. Map reads to a reference
4. Count number of reads per gene in each individual
5. Statistical analysis of differences in read counts

Analyzing patterns in expression

RNAseq normalization needed due to two systematic causes of variation:

1) Differences in the amount sequenced among individuals

2) More reads from a long transcript than from a short transcript



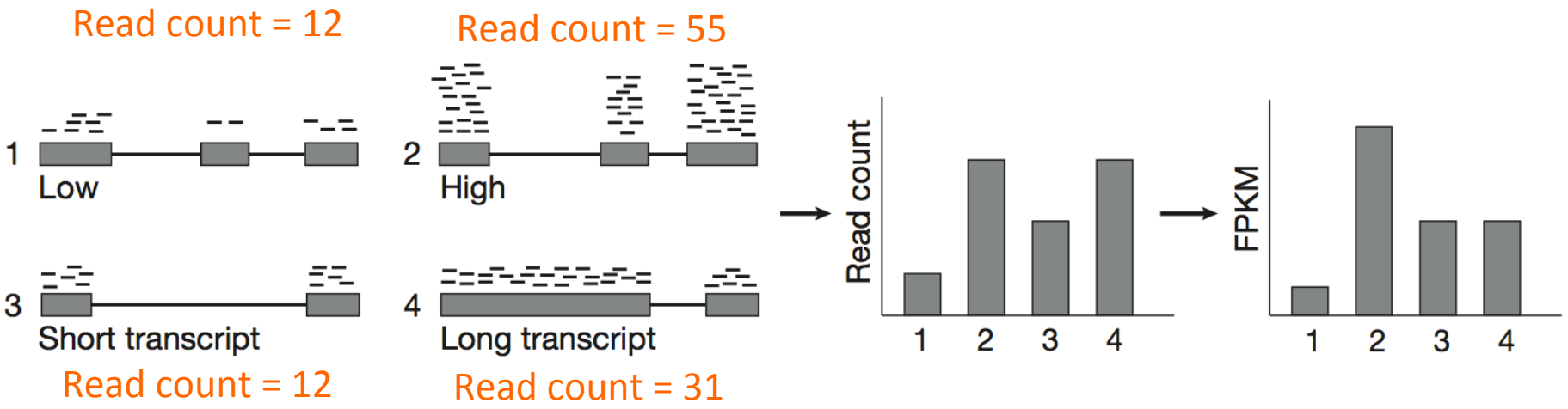
Garber et al. 2011

Analyzing patterns in expression

RNAseq normalization needed due to two systematic causes of variation:

1) Differences in the amount sequenced among individuals

2) More reads from a long transcript than from a short transcript



Garber et al. 2011

Analyzing patterns in expression

FPKM: Fragments Per Kilobase of transcript per Million reads mapped

- normalizes by transcript length and the total size of the mapped library
- correct both issues

Garber et al. 2011

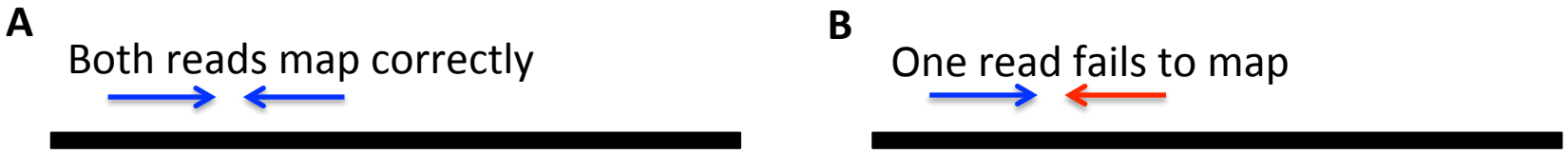
Analyzing patterns in expression

RPKM vs. FPKM

FPKM: Fragments Per Kilobase of transcript per Million reads mapped

RPKM: Reads Per Kilobase of transcript per Million 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 in expression

Practical implementation

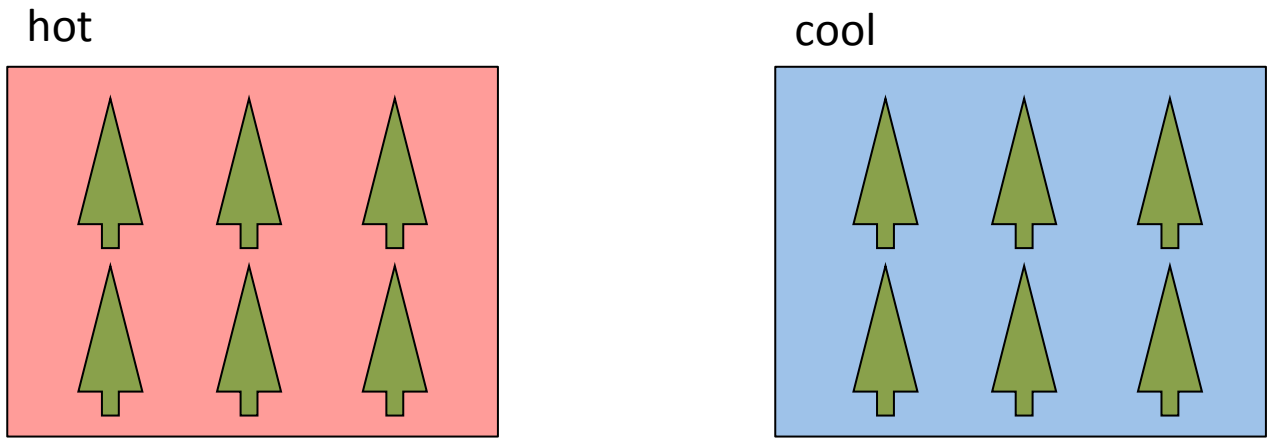
Simple: most programs will estimate FPKM or RPKM for you

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	

Tutorial part I

Align reads and estimate expression levels for three pine samples to the transcriptome reference



Tutorial part I

1. Open file README_rsem_processing_and_edgeR.txt
2. Follow the instructions to align and assess transcript abundance with RSEM (PART I)
3. Answer the following questions:

What is the expected count of comp996_c0 for each individual?

What expression measure would you use to compare gene expression between different genes and why (expected counts versus FPKM)?

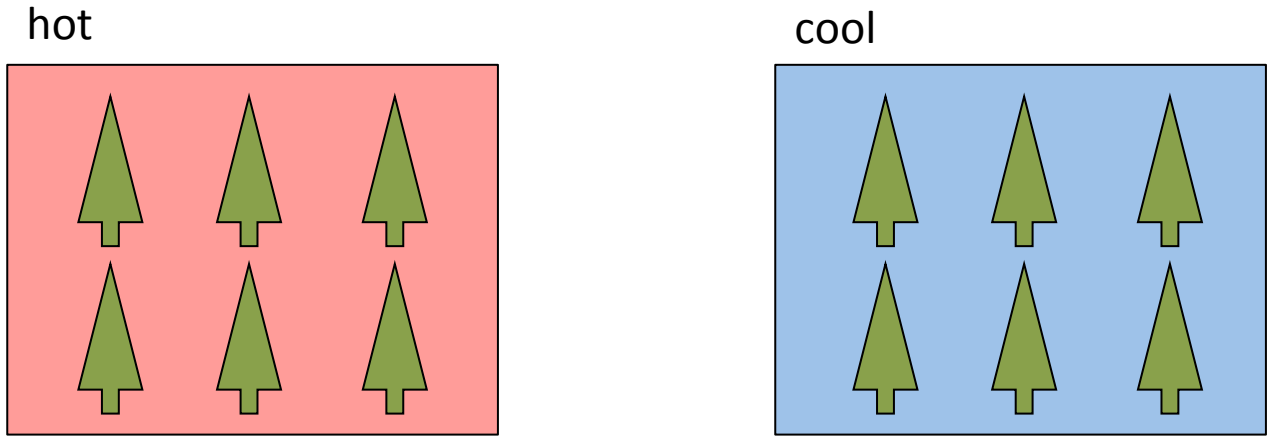
Overview of the methods

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 in expression

Fitting models to expression data



6 individuals per treatment (1 library/ind)

What genes are differentially expressed in response to temperature?

Analyzing patterns in 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 in expression

Approaches to analysis:

1. Differential gene expression on gene-by-gene basis (e.g. DESeq, EdgeR, limma)
   - Examine how each gene is affected by a factor (e.g. treatment)
   - Use glms to identify genes with significant expression differences among groups
2. Patterns of gene co-expression
   - Identify clusters of genes that are regulated together

Analyzing patterns in expression

Biological variation

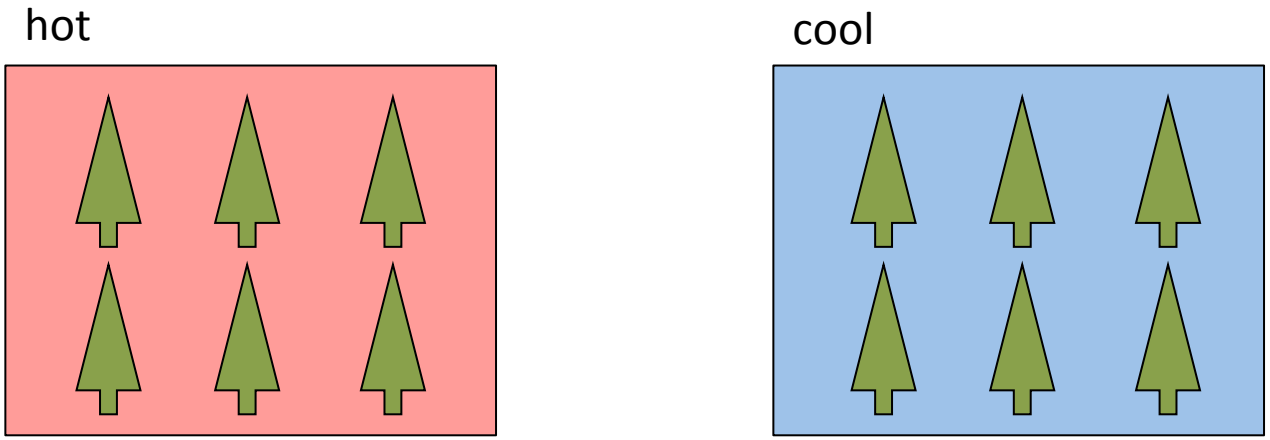
- real differences between samples due to:

1) uncontrolled sources that should be homogenous across treatments

2) controlled sources that arise from experimental treatment/design

Technical variation

- arises from measurement error inherent in the sequencing process (sequencing and library prep)

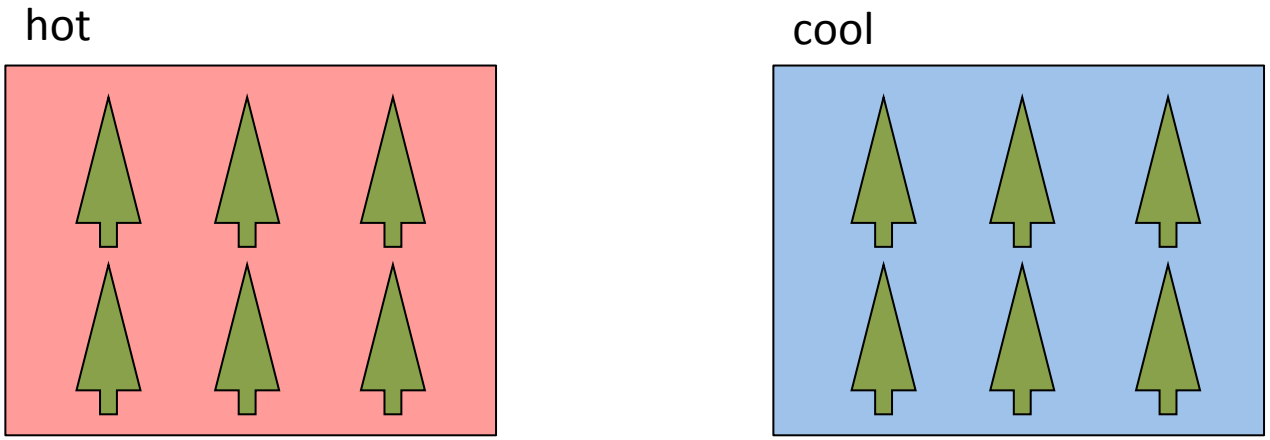


Analyzing patterns in expression

- biological replication (multiple individuals per treatment)
- technical replication (here, there is no technical replication)

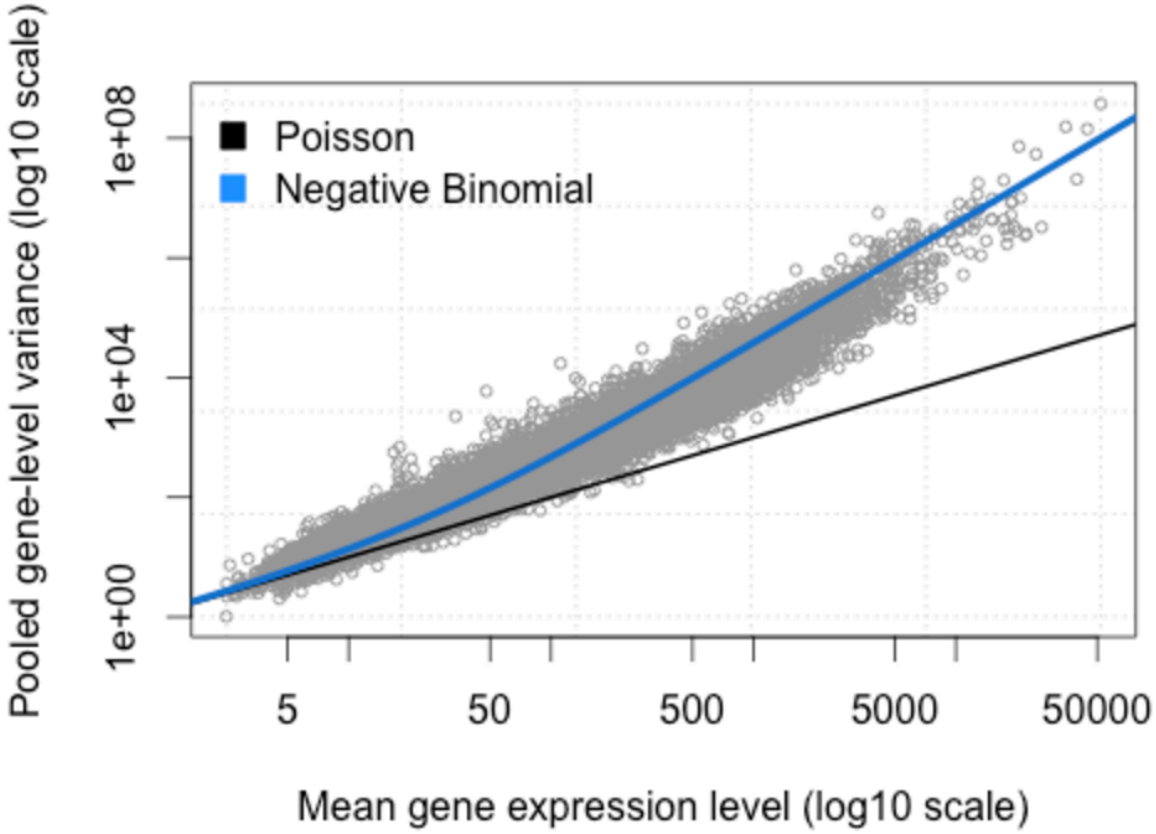
Regression of normalized counts on variable(s) of interest

- fold-change in expression among factor levels (log2(B/A))
- estimates of significance



Analyzing patterns in expression

- Count data can be modeled using the Poisson distribution (mean=variance)
- Biological variance creates over-dispersion so the mean does not equal the variance



Analyzing patterns in expression

For the negative binomial:

- $var = \mu + \phi\mu^2$
- $\sqrt{\phi}$ = CV (SD/mean)
- ϕ is called the dispersion parameter
- Total CV^2 in expression = Technical CV^2 + Biological CV^2

Biological CV (BCV) is the coefficient of variation with which the (unknown) true abundance of the gene varies between biological replicates.

Analyzing patterns in expression

Empirical Bayes for gene expression

- Many RNAseq/microarray approaches use an empirical Bayes method to “borrow” information across genes
- Prevents outliers from driving differential expression

Who were the best batters?

the worst batters?

name	H	AB	average
Frank Abercrombie	0	4	0
Horace Allen	0	7	0
Pete Allen	0	4	0
Walter Alston	0	1	0
Bill Andrus	0	9	0

the best batters?

name	H	AB	average
Jeff Banister	1	1	1
Doc Bass	1	1	1
Steve Biras	2	2	1
C. B. Burns	1	1	1
Jackie Gallagher	1	1	1

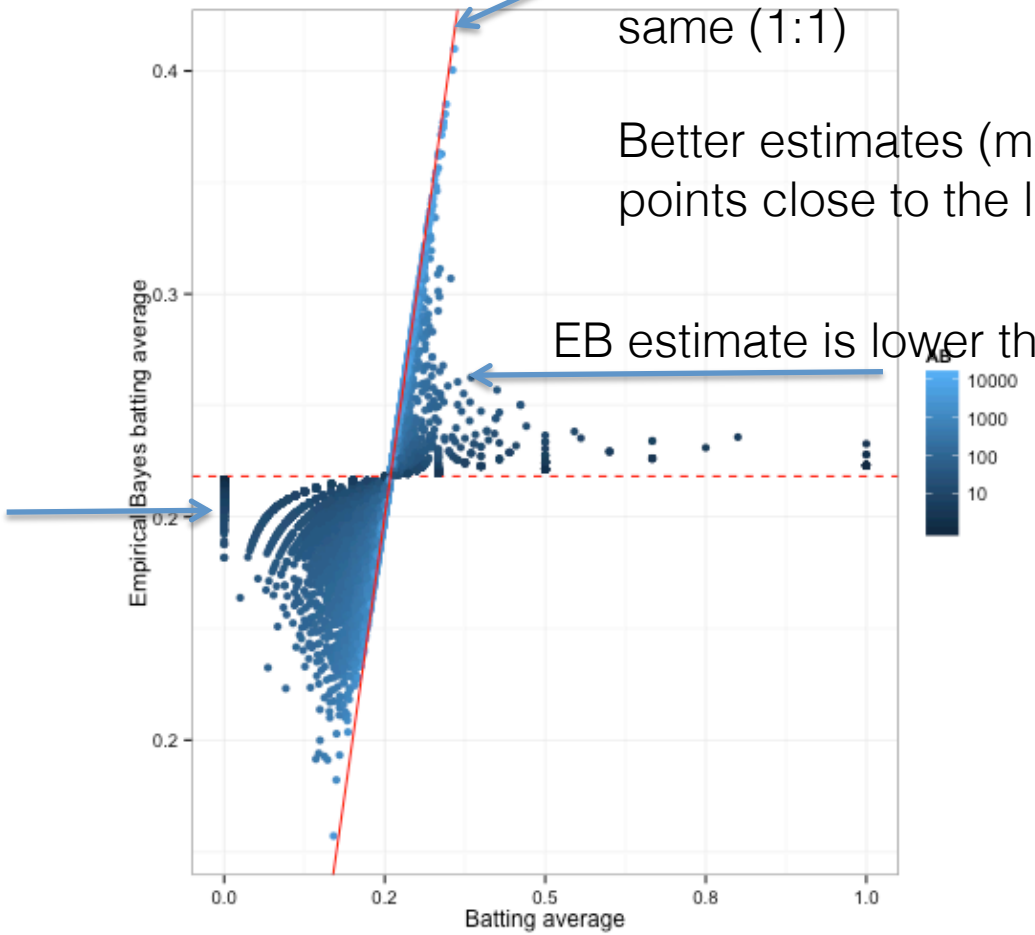
Shrinkage: EB tends to move estimates towards the mean

Estimate and measured average are the same (1:1)

Better estimates (more data) for those points close to the line

EB estimate is lower than actual estimate

EB estimate is higher than actual estimate



http://varianceexplained.org/r/empirical_bayes_baseball/

better estimates of batting averages

name	H	AB	average	eb_estimate
Rogers Hornsby	2930	8173	0.358	0.355
Shoeless Joe Jackson	1772	4981	0.356	0.350
Ed Delahanty	2596	7505	0.346	0.343
Billy Hamilton	2158	6268	0.344	0.340
Harry Heilmann	2660	7787	0.342	0.339

name	H	AB	average	eb_estimate
Bill Bergen	516	3028	0.170	0.178
Ray Oyler	221	1265	0.175	0.191
John Vukovich	90	559	0.161	0.196
John Humphries	52	364	0.143	0.196
George Baker	74	474	0.156	0.196

Analyzing patterns in expression

Estimating Dispersion

- common dispersion: same Biological CV among genes (i.e. proportional relationship between gene-wise standard deviations and gene-wise means is the same for all genes)
  - gene expression levels have non-identical and dependent distribution between genes (common dispersion too naïve)
- tagwise dispersion: the common dispersion estimate is modified for each gene based on a Empirical Bayes estimate of the per-gene relationship between mean and variance

Analyzing patterns in expression

Using the approach from edgeR as an example

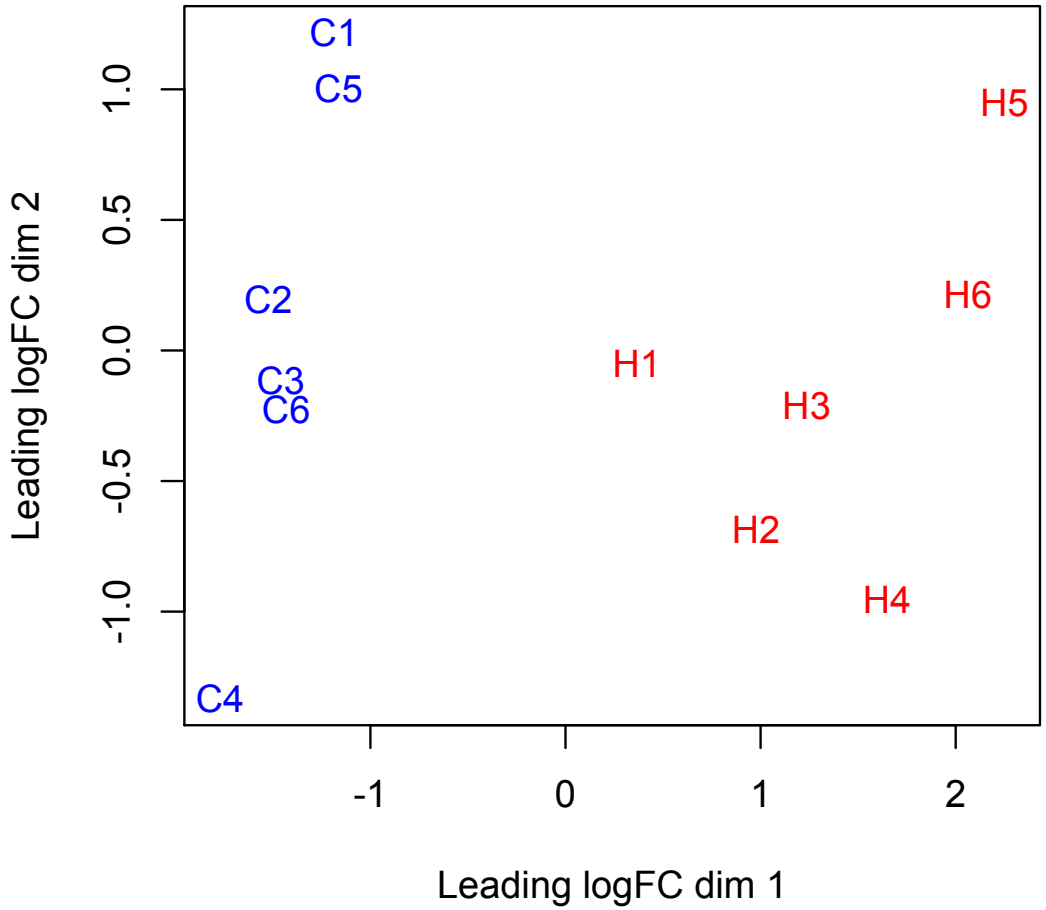
Model fitting results in estimation of log fold change (logFC) in expression, p-value, and estimation of False Discovery Rate (FDR)

	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

- EdgeR allows multiple factors for more complex designs (as does Limma)

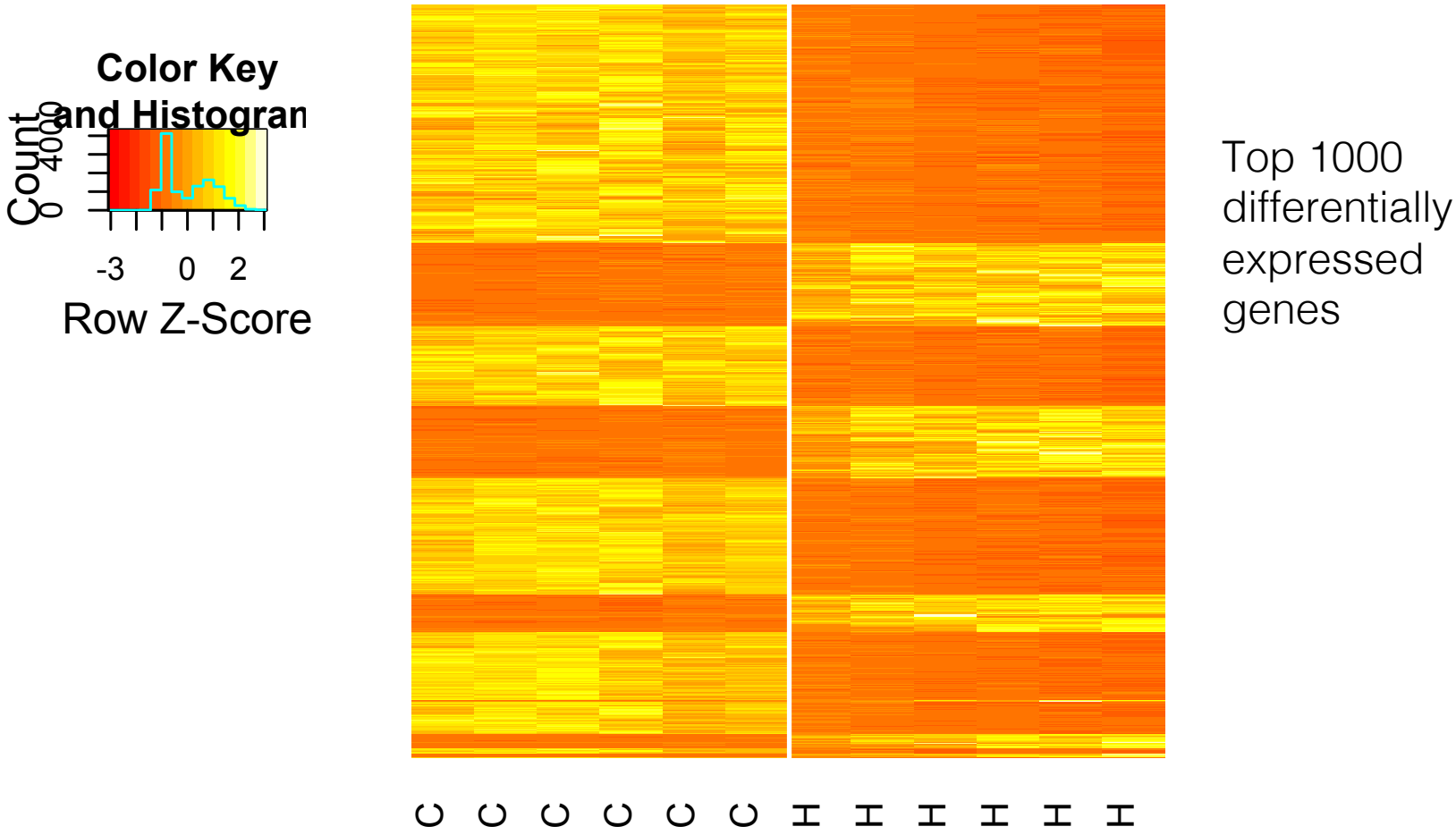
Analyzing patterns in expression

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



Analyzing patterns in expression

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



Analyzing patterns in 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

Tutorial part II

- Run EdgeR to compare expression between climate treatments (PART II)
- Answer the following questions

1. How many genes are differentially expressed by treatment in the simple contrast of C vs H (using "cold_hot_expression.txt")? How does the choice of FDR cutoff or p-value affect this number? What happens if you include genes in your analysis with low or no expression across all of the samples?

2. How many genes are differentially expressed in the three-way contrast (using "cold_hot_mwh_expression.txt")? Which treatment is driving differential expression here? How do you know?

3. How much does model fitting with common dispersion vs. tagwise dispersion affect the answers you get from the data? (think in terms of the number of DE genes, the evidence for a single gene, etc.)

Analyzing patterns in expression

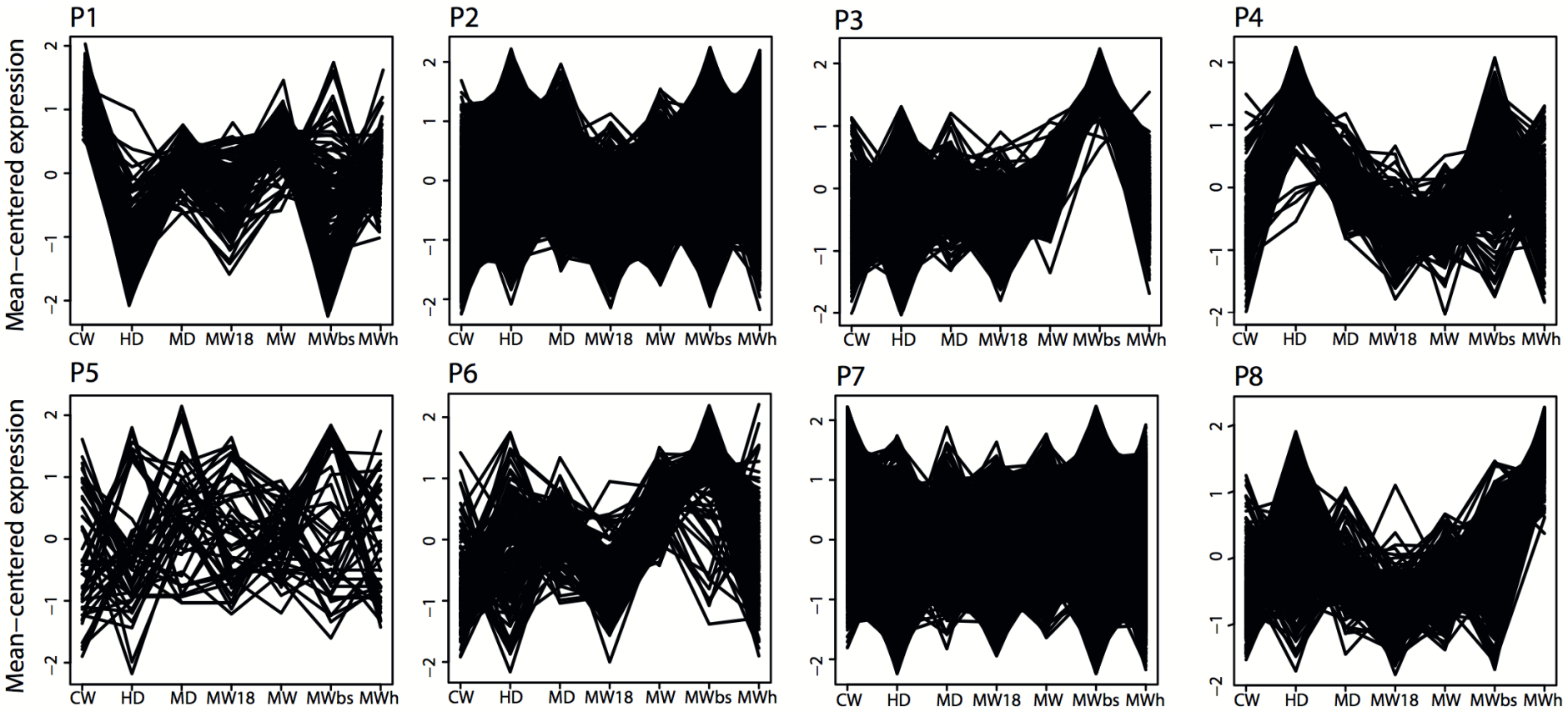
Gene co-expression networks: Finding genes that are expressed in the same way across treatments/tissues

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

Analyzing patterns in expression

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



Yeaman et al. (2014)

Now what?

- As with many approaches in genomics, there is a “too much data” problem
- Annotation of genes
- 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
  - Move from descriptive to biological insight

Technical considerations

Depth of coverage?

- Dependent on:
  1. study organism
  2. transcriptome size
  3. purpose of your study
- Low power if < 50 counts per million per gene
- Too many individuals per lane can increase your technical variation
- 10 million reads per sample is a benchmark from which to start for most eukaryotes
- Biological replication is often more valuable than higher depth of coverage per individual

Technical considerations

- Variation among cells of the same type sampled at the same time (single-cell sequencing)
- Variation among cell types of the same tissue (micro-dissection)
- 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 needs large amounts of RAM
- Lodgepole pine transcriptome assembly:

  40Gbp of sequence data = 200 GB RAM
- 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
- Pooling from multiple tissues, treatments, developmental time points

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

http://rnaseq.uoregon.edu/#analysis-trimming