COMBINE

For Australian students and early career researchers in Bioinformatics and Computational Biology.

RNAseq analysis in R

24-25 November 2016

Walter+Eliza Hall
Institute of Medical Research

DISCOVERIES FOR HUMANITY

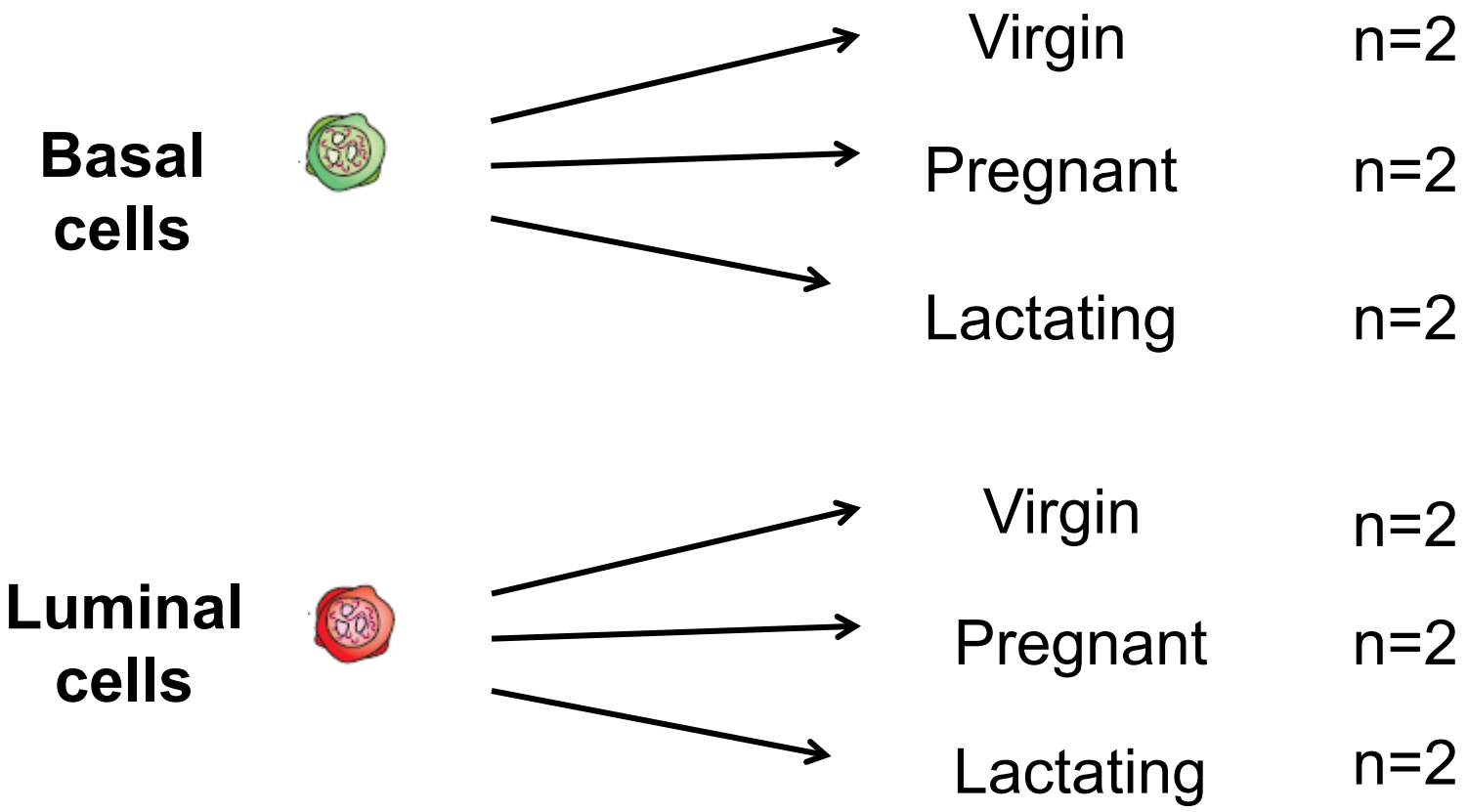
Charity Law law@wehi.edu.au

Matt Ritchie mritchie@wehi.edu.au

QC and visualisation (part 1)

Slide taken from COMBINE RNAseq workshop on 23/09/2016

RNA-seq of Mouse mammary gland



Fu *et al.* (2015) 'EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival' Nat Cell Biol

Slide taken from COMBINE RNAseq workshop on 23/09/2016

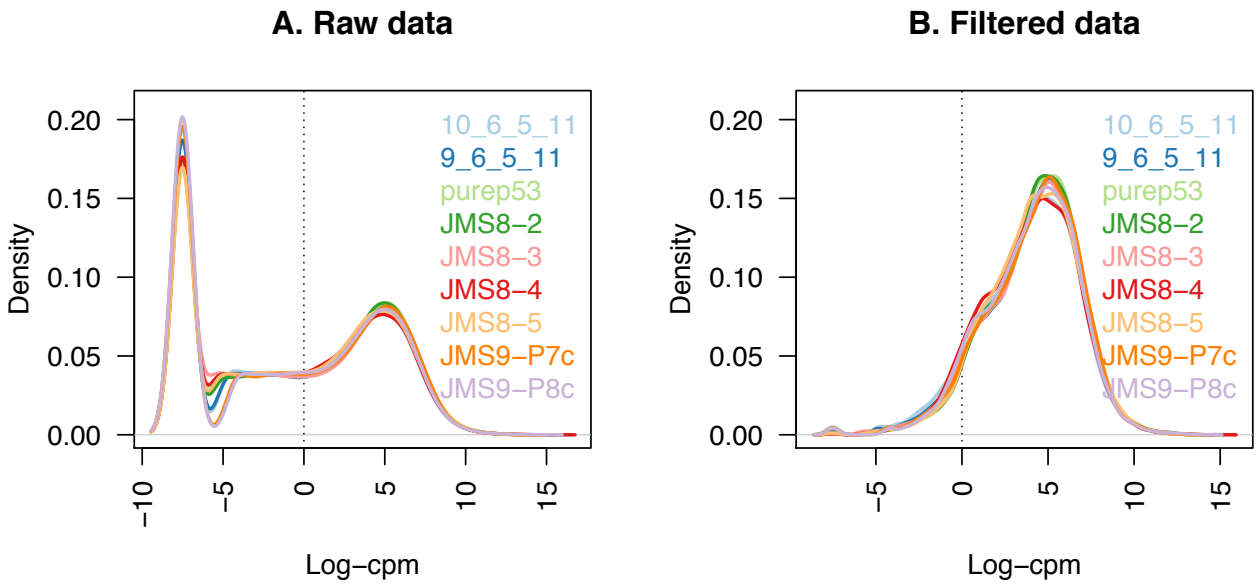
(some) questions we can ask

- Which genes are differentially expressed between basal and luminal cells?
- … between basal and luminal in virgin mice?
- … between pregnant and lactating mice?
- … between pregnant and lactating mice in basal cells?

- Reading in the data
  - counts data and sample information
- Formatting the data
  - clean it up so we can look at it easily

Filtering out lowly expressed genes

- Genes with very low counts in all samples provide little evidence for differential expression
- Often samples have many genes with zero or very low counts



Filtering out lowly expressed genes

- Testing for differential expression for many genes simultaneously adds to the multiple testing burden, reducing the power to detect DE genes.
- IT IS VERY IMPORTANT to filter out genes that have all zero counts or very low counts.
- We filter using CPM values rather than counts because they account for differences in sequencing depth between samples.

Filtering out lowly expressed genes

- CPM = counts per million, or how many counts would I get for a gene if the sample had a library size of 1M.

For a given gene:

Library size	Count	CPM
1M	1	1
10M	10	1
20M	10	0.5

Filtering out lowly expressed genes

- Use a CPM threshold to define “expressed” and “unexpressed”
- As a general rule, a good threshold can be chosen for a CPM value that corresponds to a count of 10.
- In our dataset, the samples have library sizes of 20 to 20 something million.

Library size	Count	CPM
1M	1	1
10M	10	1
20M	10	0.5

Filtering out lowly expressed genes

- Use a CPM threshold to define “expressed” and “unexpressed”
- As a general rule, a good threshold can be chosen for a CPM value that corresponds to a count of 10.
- In our dataset, the samples have library sizes of 20 to 20 something million.

Library size	Count	
1M	1	
10M	10	
20M	10	0.5

We use a CPM threshold of 0.5!

Filtering out lowly expressed genes

- Use a CPM threshold to define "expr
  "unexpressed"
- As a general rule, a good threshold
  CPM value that corresponds to a co
- In our dataset, the samples have library s to 20 something million.

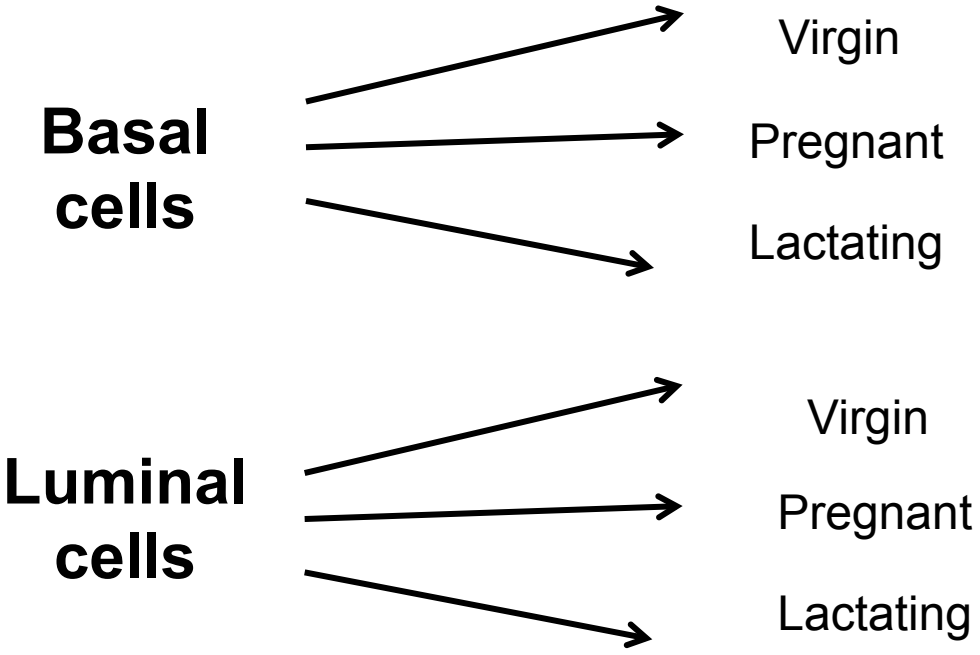
But if this is too hard to work out, a CPM threshold of 1 works well in most cases.

Library size	Count	
1M	1	
10M	10	
20M	10	0.5

We use a CPM threshold of 0.5!

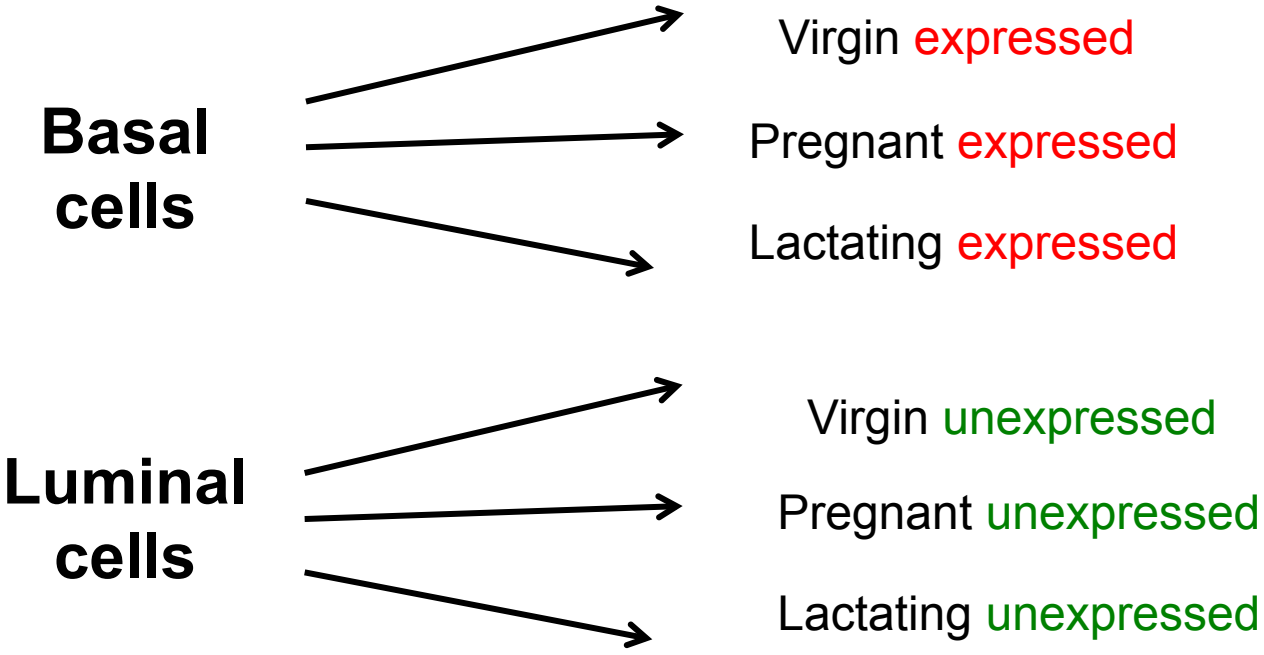
Filtering out lowly expressed genes

- We keep any gene that is (roughly) expressed in at least one group.
- 12 samples, 6 groups, 2 replicates in each group.



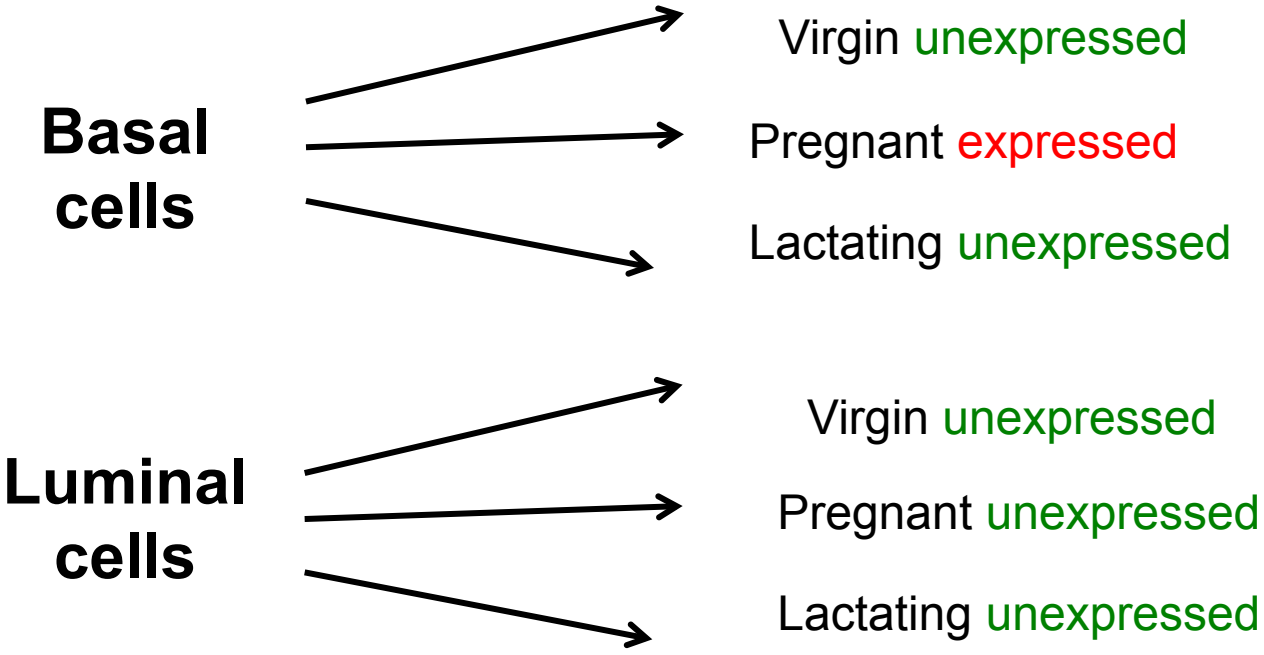
Filtering out lowly expressed genes

- We keep any gene that is (roughly) expressed in at least one group.
- 12 samples, 6 groups, 2 replicates in each group.



Filtering out lowly expressed genes

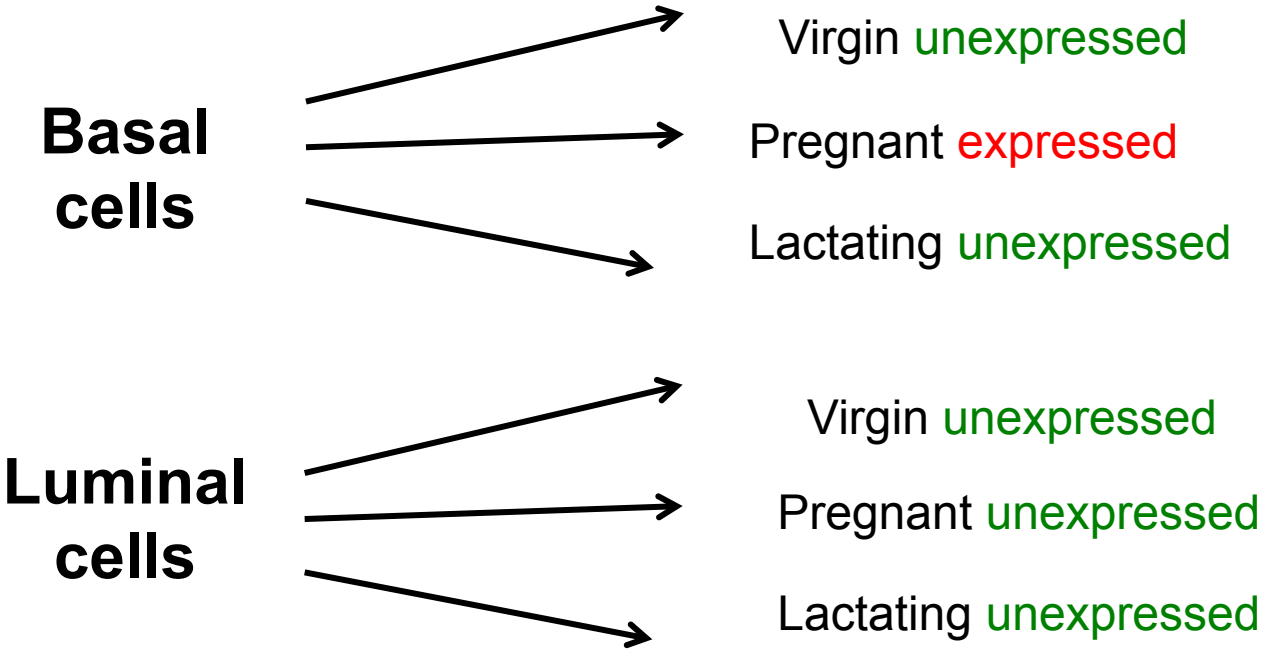
- We keep any gene that is (roughly) expressed in at least one group.
- 12 samples, 6 groups, 2 replicates in each group.



Filtering out lowly expressed genes

- We keep any gene that is (roughly) expressed in at least one group.
- 12 samples, 6 groups, 2 replicates in each group.

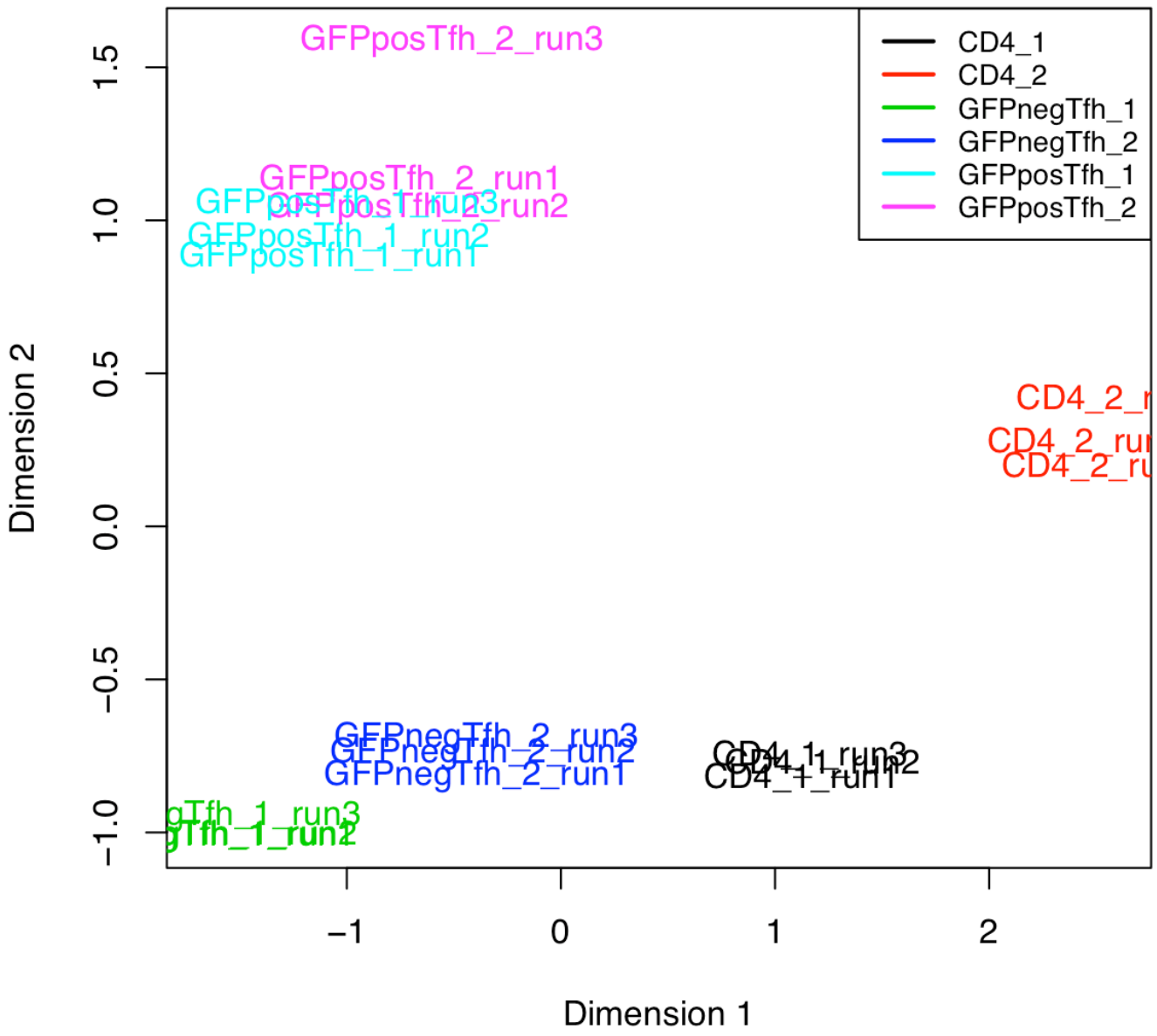
Keep gene if CPM > 0.5 in at least 2 or more samples

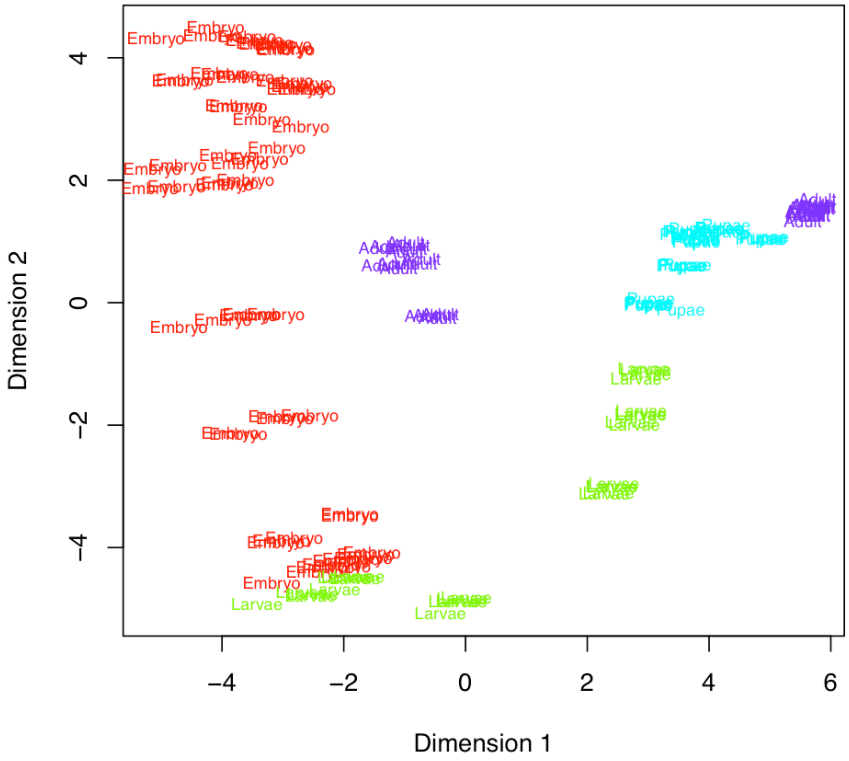


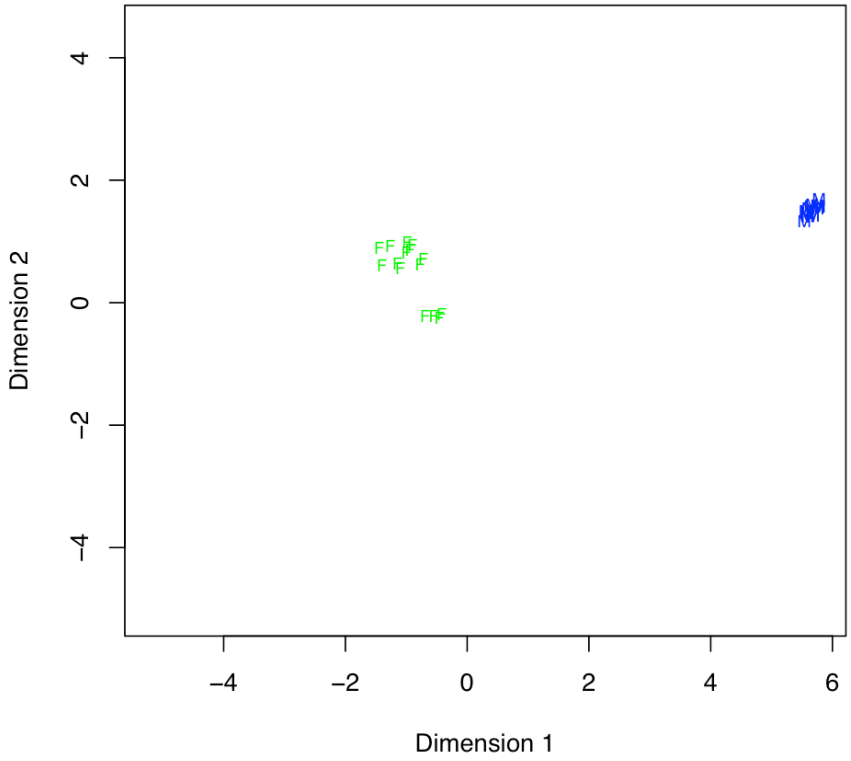
QC and visualisation (part 2)

MDS Plots

- A visualisation of a principle components analysis which looks at where the greatest sources of variation in the data come from.
- Distances represents the typical log2-FC observed between each pair of samples
  - e.g. 6 units apart = 2^6 = 64-fold difference
- Unsupervised – separation based on data, no prior knowledge of experimental design.
  - Useful for an overview of the data. Do samples separate by experimental groups?
  - Quality control
  - Outliers?

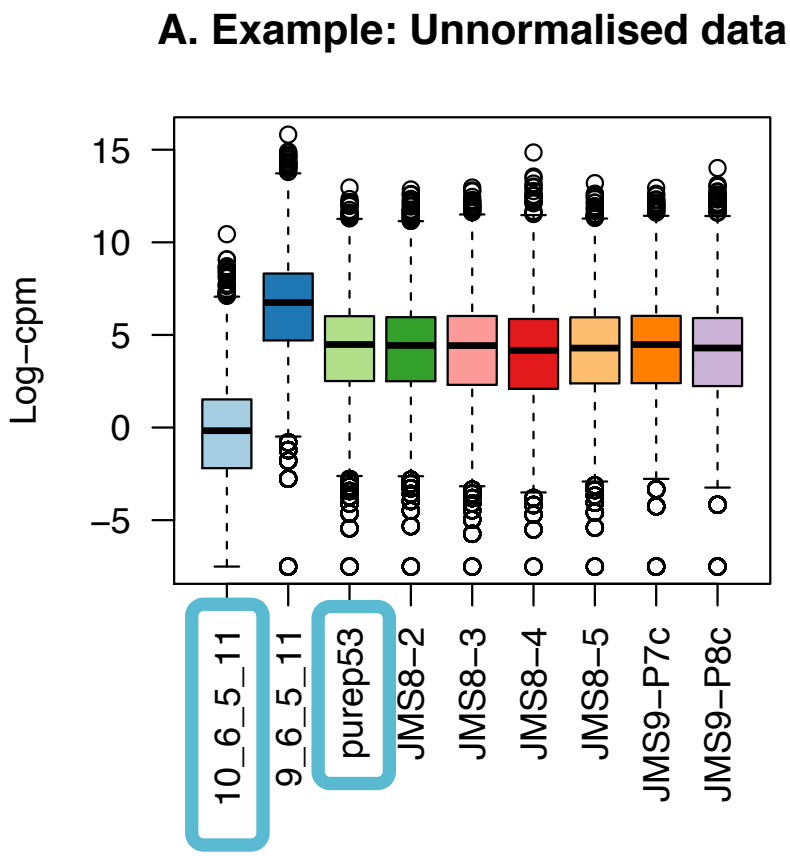

CD4_1
CD4_2
GFPnegTfh_1
GFPnegTfh_2
GFPposTfh_1
GFPposTfh_2
GFPposTfh_2_run3
GFPposTfh_2_run1
GFPposTfh_1_run2
GFPposTfh_1_run1
GFPnegTfh_2_run3
GFPnegTfh_2_run2
GFPnegTfh_2_run1
CD4_1_run3
CD4_1_run2
CD4_1_run1
Dimension 1
Dimension 2
−1
0
1
2
−1.0
−0.5
0.0
0.5
1.0
1.5


Dimension 2
Dimension 1
Embryo
Larvae
Pupae
Adult


Dimension 2
Dimension 1
F

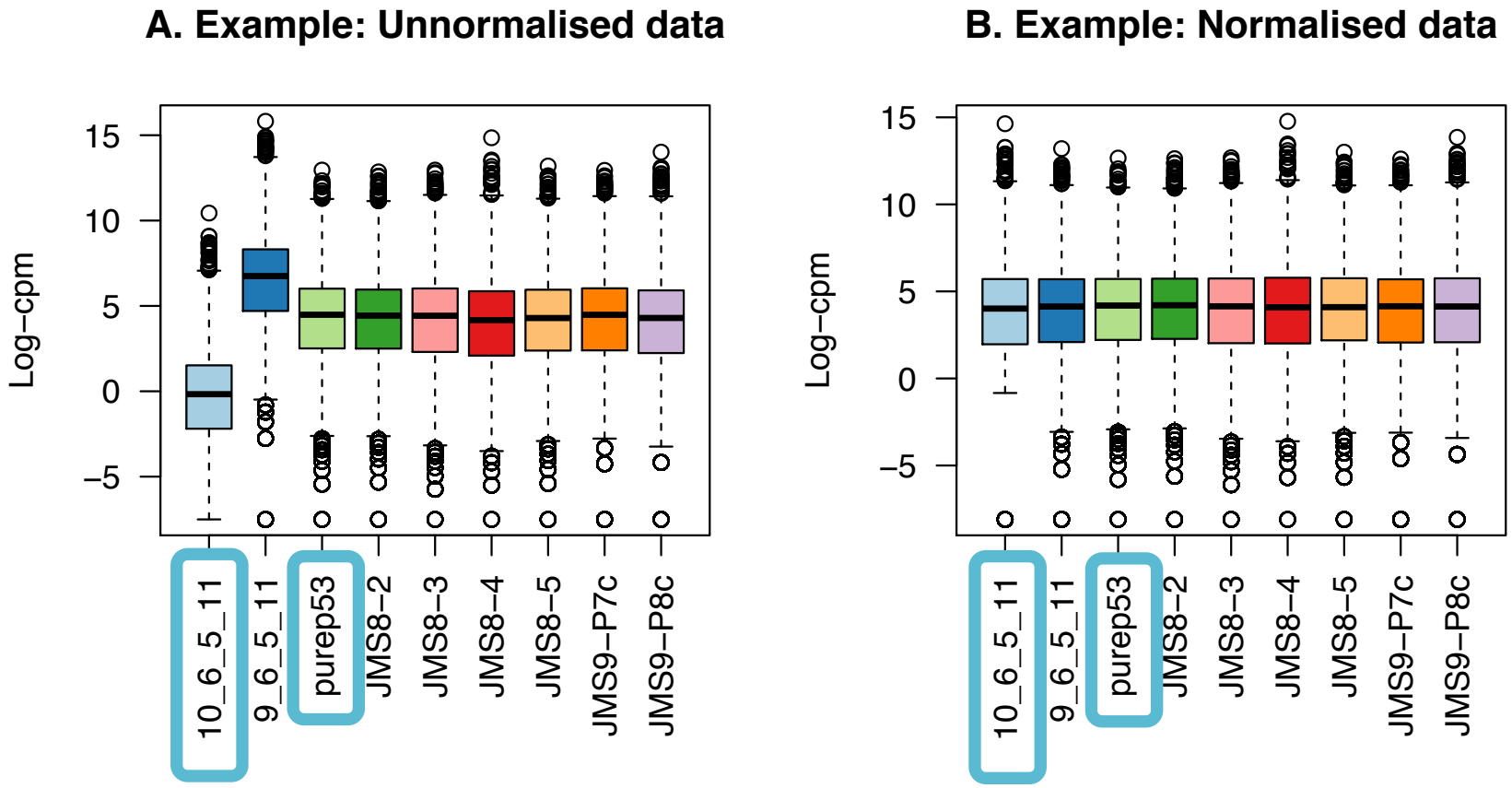
QC and visualisation (part 3)

Normalisation for composition bias



If we ran a DE analysis on Sample 1 and Sample 3, almost all genes will be down-regulated in Sample 1!!

Normalisation for composition bias



Normalisation for composition bias

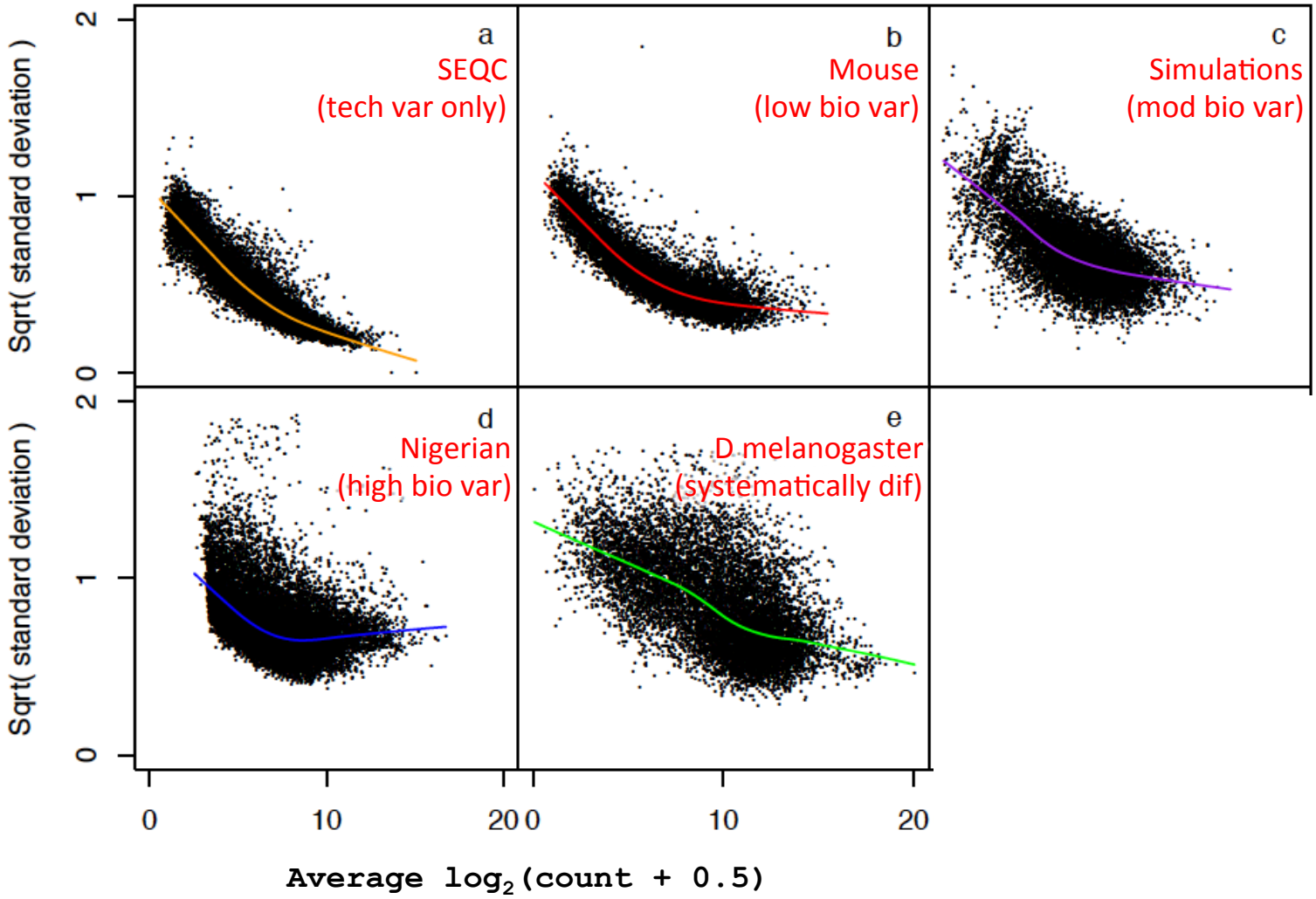
- TMM normalisation (Robinson and Oshlack, 2010)
- How do we make the expression of all the genes go UP in the one sample?
  - Scaling factors
- E.g. scale library size by 0.1 so effective library size is 1M.

Library size	Count	CPM
10M	10	1
1M	10	10

- Scaling factor <1 makes the CPM larger.

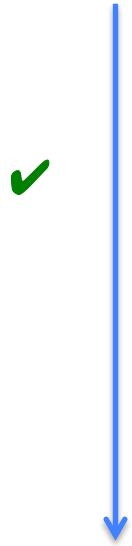
Voom

Variance of log-cpm depends on mean of log-cpm



RNA-seq data is

- discrete
- has non-constant mean-variance trend



Voom

- Transform to log-counts per million
- Remove mean-var dependence through the use of precision weights

Normal dist. assumes that the data is

- continuous
- has constant variance

Variance weights

- Obtain variance estimates for each observation using mean-var trend.
- Assign inverse variance weights to each observation.
- Weights remove mean-variance trend from the data.