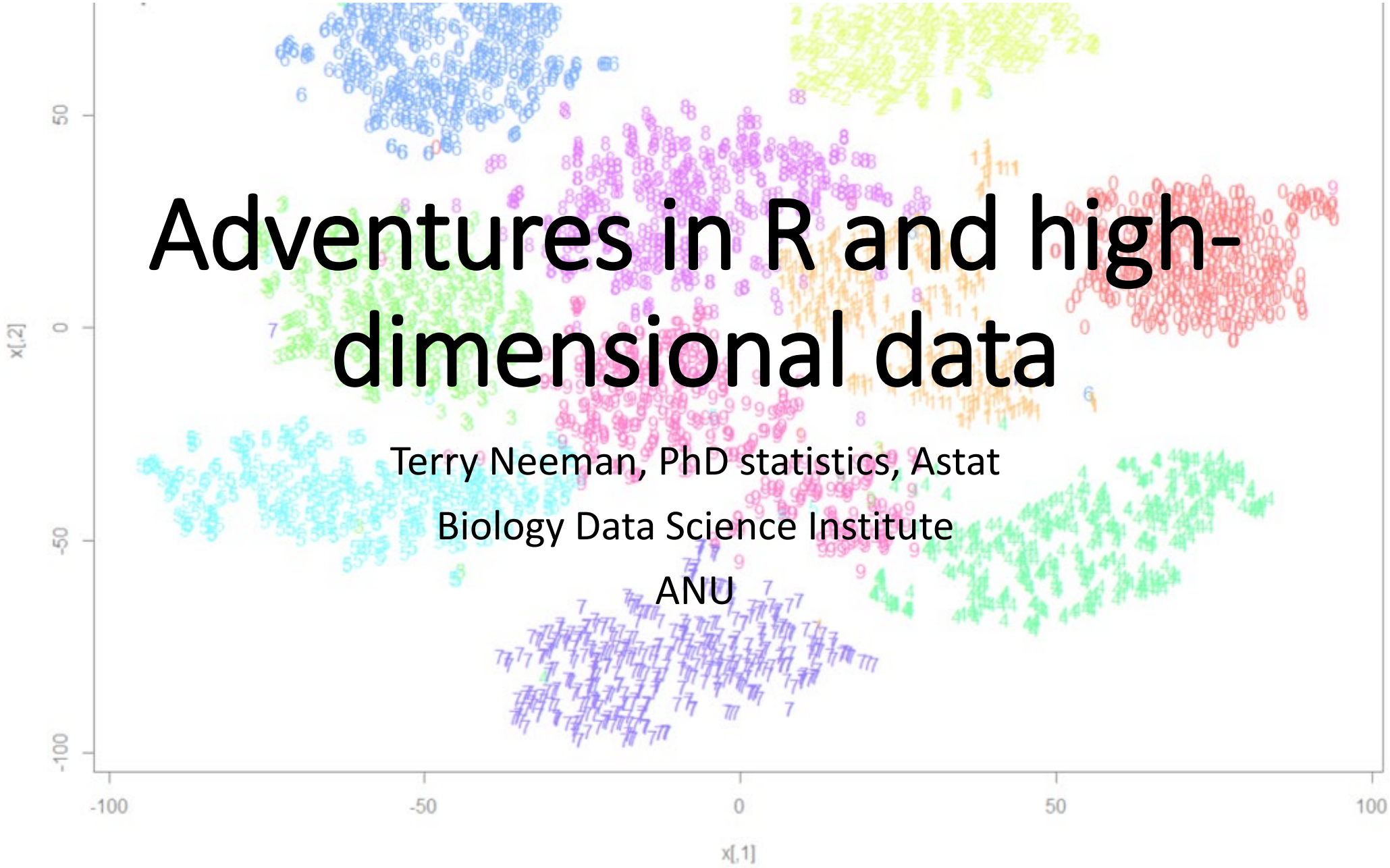


Adventures in R and high-dimensional data

Terry Neeman, PhD statistics, Astat

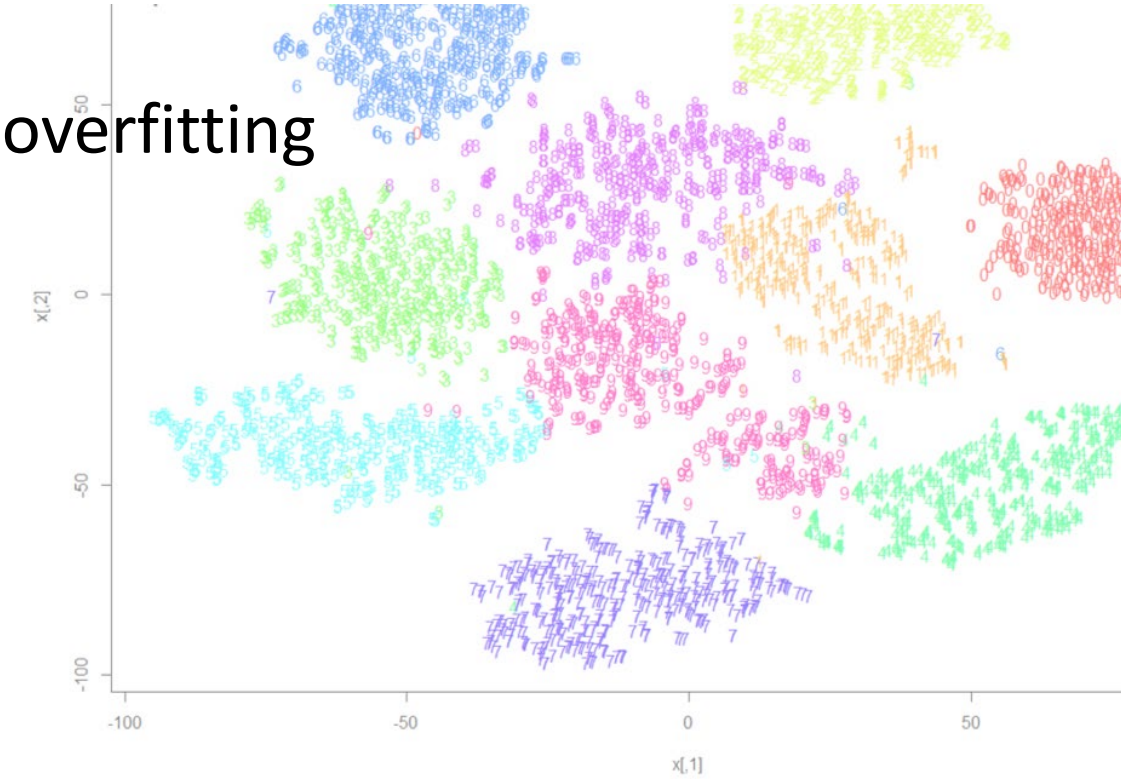
Biology Data Science Institute

ANU

Challenges of high dimensional data:

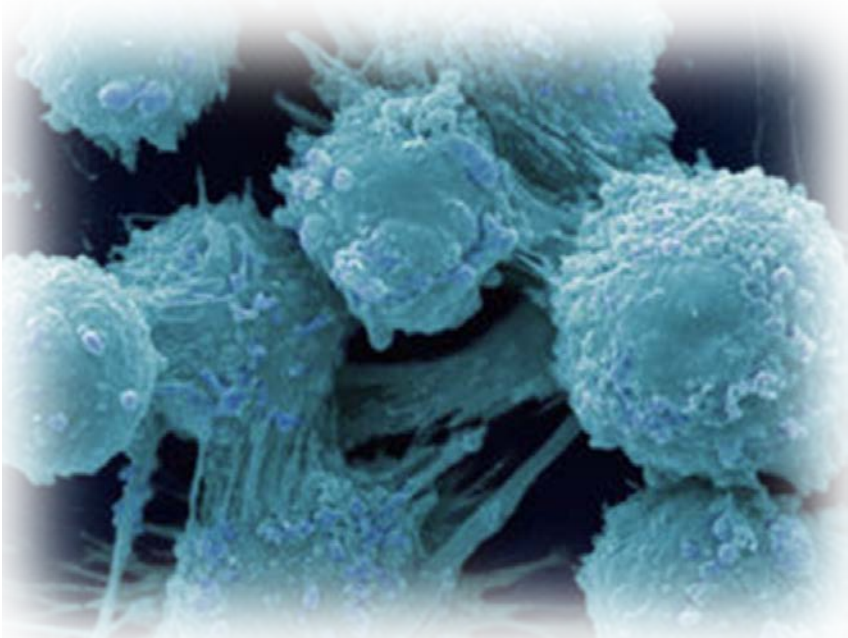
when p (features) >> n (samples)

- Visualisation
- Normalisation
- Multiple hypothesis testing
- So many features – potential for model overfitting



High-dimensional data examples

- Differential gene expression (RNA-seq)
- Comparative metabolomics (mass spectrometry)
- Differential transcriptome methylation (bisulphite sequencing)
- Comparative metagenomics
- High throughput drug screening for cancer drugs



Example: Differential Gene Expression Study

- 3 cell types in mammary gland of virgin female mice
  - Basal cells
  - Luminal progenitor cells
  - Mature luminal cells
- Run in three batches (lanes)
- 3 biological samples for each cell type
- Expression data: number of reads per “gene” (~27,000 genes)

Count data (27000+ rows, 9 columns)

sample Tags	10_6	9_6	p53	S8-2	S8-3	S8-4	S8-5	S9-P7c	S9-P8c
16178	14	37	21	37	26	132	49	25	24
16177	1180	1922	538	580	3381	2004	854	1263	895
107527	214	251	853	689	531	420	724	180	155
17082	19	18	99	100	31	79	236	20	16
16182	575	1439	107	115	2893	1389	204	1223	570
16174	97	123	12	11	79	104	23	19	16

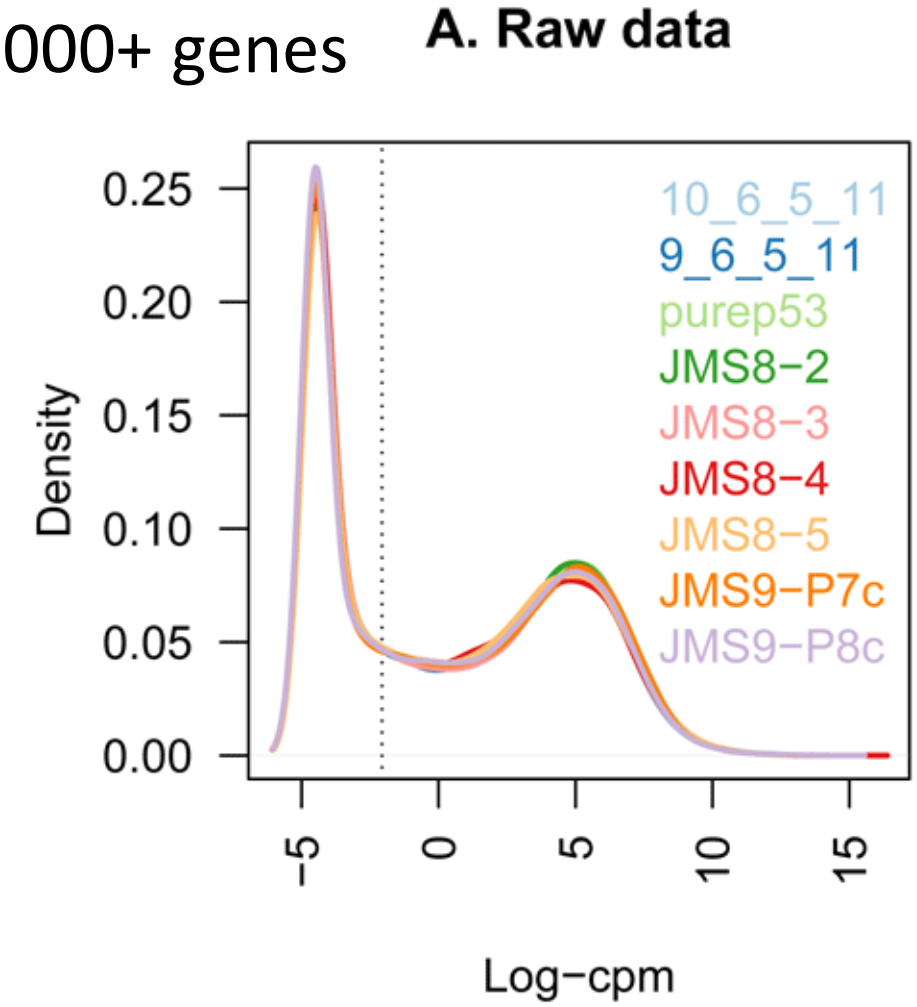
Meta-data (9 rows)

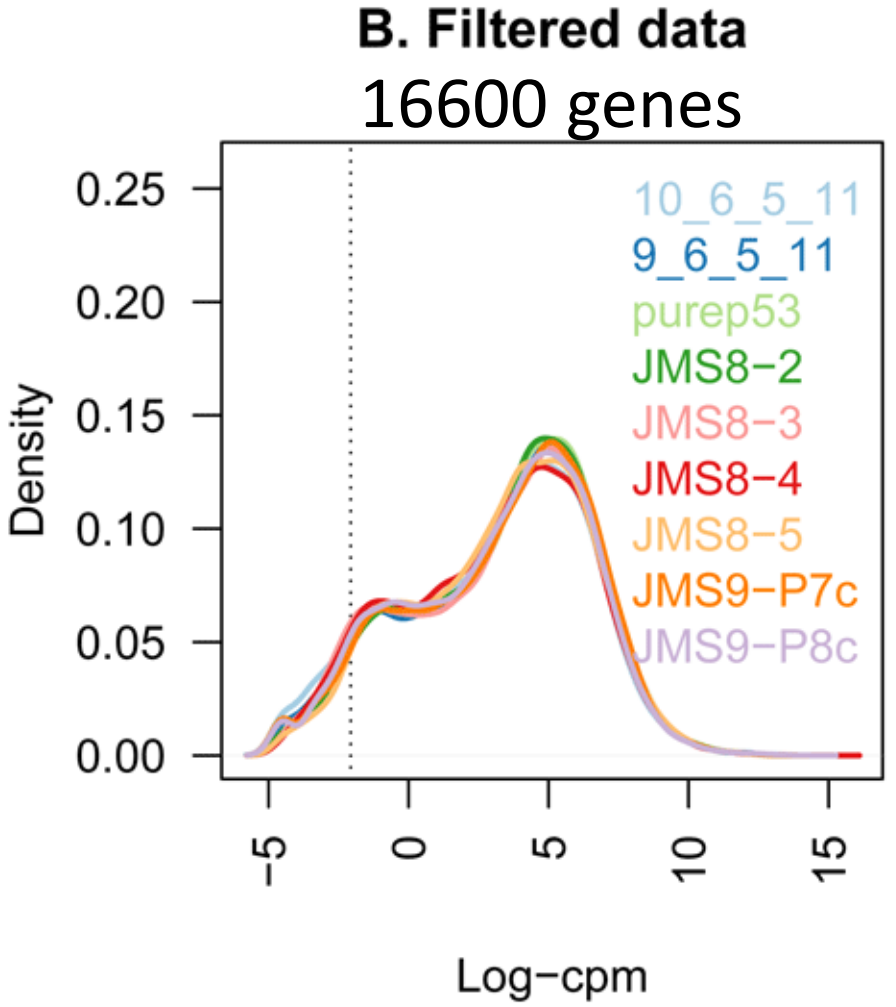
	files	group	lib.size	norm.factors	lane
10_6	GSM1545535_10_6_5_11.txt	LP	32863052	1	L004
9_6	GSM1545536_9_6_5_11.txt	ML	35335491	1	L004
p53	GSM1545538_purep53.txt	Basal	57160817	1	L004
S8-2	GSM1545539_JMS8-2.txt	Basal	51368625	1	L006
S8-3	GSM1545540_JMS8-3.txt	ML	75795034	1	L006
S8-4	GSM1545541_JMS8-4.txt	LP	60517657	1	L006
S8-5	GSM1545542_JMS8-5.txt	Basal	55086324	1	L006
S9-P7c	GSM1545544_JMS9-P7c.txt	ML	21311068	1	L008
S9-P8c	GSM1545545_JMS9-P8c.txt	LP	19958838	1	L008

Figure 1

Quality control:
Filter out genes with low counts

27000+ genes

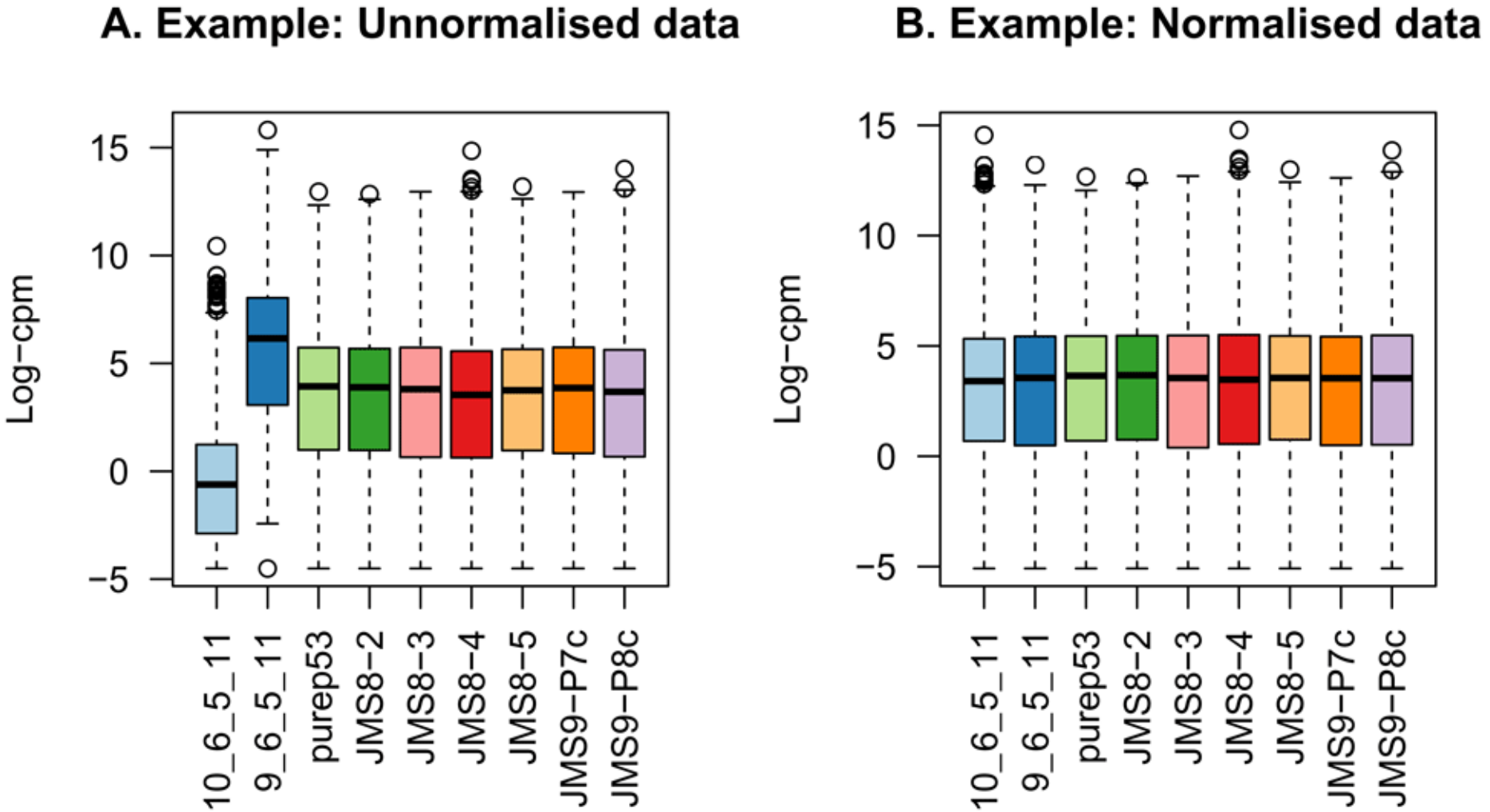




Law CW, Alhamdoosh M, Su S et al. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR [version 3]. F1000Research 2018, 5:1408 (doi: 10.12688/f1000research.9005.3)

Figure 2.

Quality control: Normalise data



Law CW, Alhamdoosh M, Su S et al. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR [version 3]. F1000Research 2018, 5:1408 (doi: 10.12688/f1000research.9005.3)

```
library(Glimma)
glMDSPlot(lcpm, labels=paste(group, lane, sep="_"),  groups=x$samples[,c(2,5)])
```

Look at global difference using MDS plots:
http://bioinf.wehi.edu.au/folders/limmaWorkflow/glimma-plots/MDS-Plot.html

Fit a linear model to each (16000+ genes)

```
##          BasalvsLP  BasalvsML  LPvsML
## Down       4648       4927      3135
## NotSig     7113       7026      10972
## Up         4863       4671      2517
```


```
design <- model.matrix(~0+group+lane)
contr.matrix <- makeContrasts( BasalvsLP = Basal-LP, BasalvsML = Basal - ML, LPvsML = LP - ML)
vfit <- lmFit(v, design)
vfit <- contrasts.fit(vfit, contrasts=contr.matrix)
efit <- eBayes(vfit)
summary(decideTests(efit))
```

Fit a linear model to each (16000+ genes) Arrange from most differentially expressed…

```
basal.vs.lp <- topTable(efit, coef=1)
head(basal.vs.lp)
```

ENTREZID	SYMBOL	CHROM	logFC	AveExpr	t	P.Value	adj.P.Val
12759	Clu	chr14	-5.46	8.86	-33.6	1.72e-10	1.71e-06
53624	Cldn7	chr11	-5.53	6.30	-32.0	2.58e-10	1.71e-06
242505	Rasef	chr4	-5.94	5.12	-31.3	3.08e-10	1.71e-06
67451	Pkp2	chr16	-5.74	4.42	-29.9	4.58e-10	1.74e-06
228543	Rhov	chr2	-6.26	5.49	-29.1	5.78e-10	1.74e-06
70350	Basp1	chr15	-6.08	5.25	-28.3	7.27e-10	1.74e-06

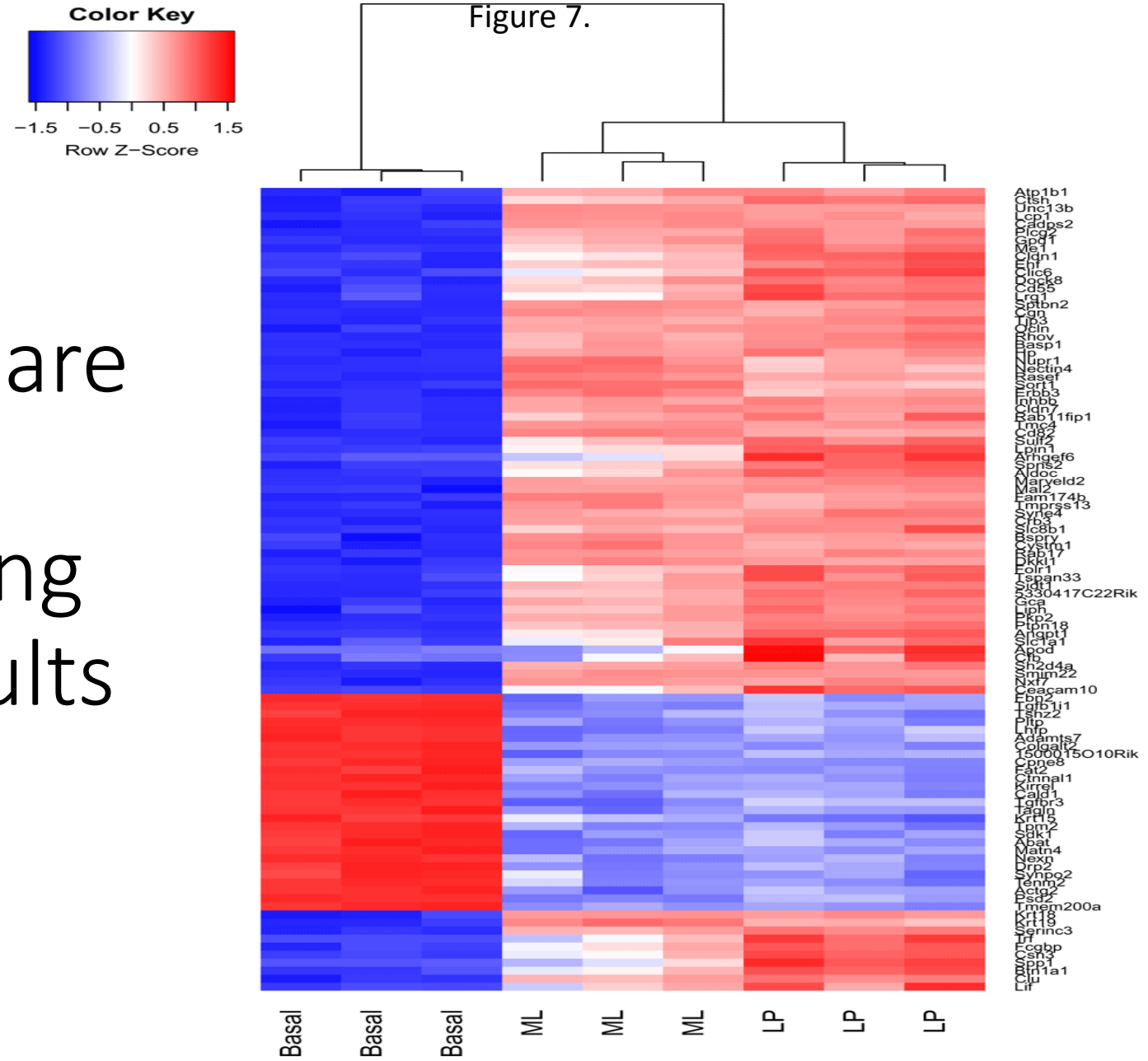
An interactive plot for examining differentially expressed genes

```
glMDPlot(efit, coef=1, status=dt, main=colnames(efit)[1], side.main="ENTREZID", counts=lcpm, groups=group)
```

An interactive plot for differential expression:
http://bioinf.wehi.edu.au/folders/limmaWorkflow/glimma-plots/MD-Plot.html

Heatmaps are useful for summarising model results

Figure 7.



Law CW, Alhamdoosh M, Su S et al. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR [version 3]. F1000Research 2018, 5:1408 (doi: 10.12688/f1000research.9005.3)

F1000Research

Reference:

Law CW, Alhamdoosh M, Su S *et al.* RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR [version 3; peer review: 3 approved]. *F1000Research*2018, **5**:1408 (https://doi.org/10.12688/f1000research.9005.3)