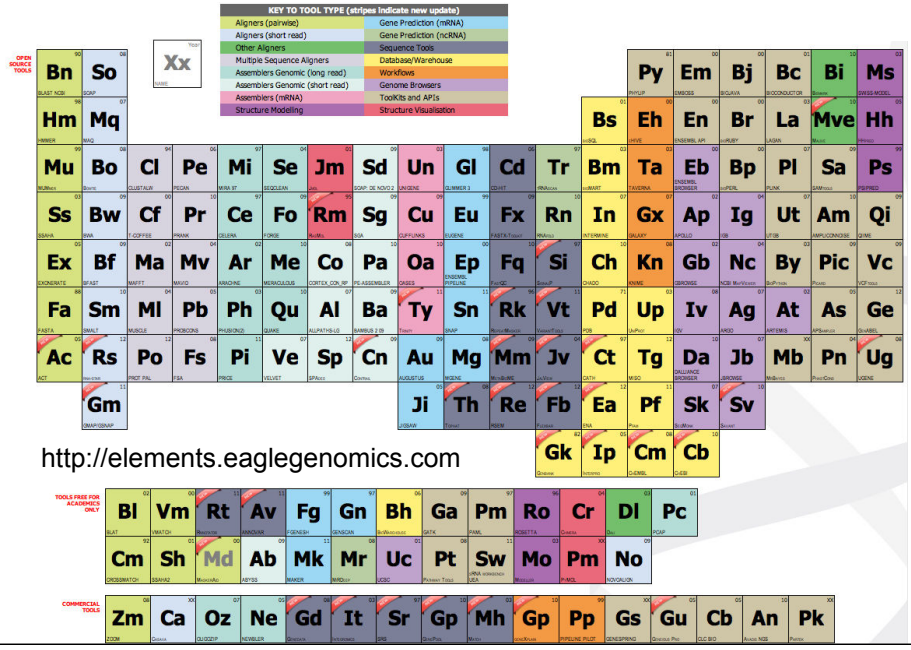
shRNA-seq analysis in Galaxy

Matthew Ritchie

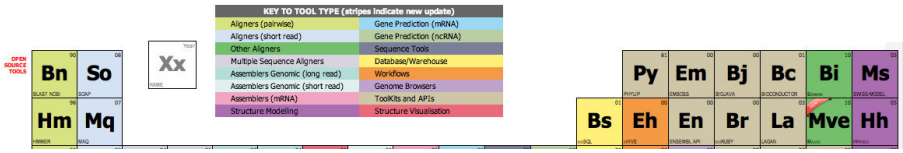
Molecular Medicine
Walter + Eliza Hall Institute

Elements of Bioinformatics



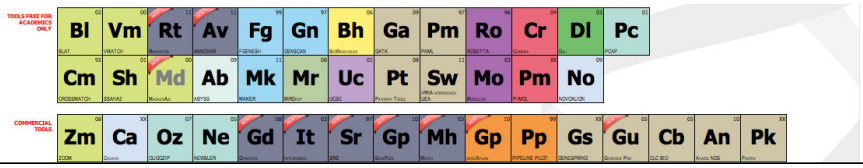
http://elements.eaglegenomics.com

Elements of Bioinformatics



We author a range of R/Bioconductor packages tailored for genomics research

limma
edgeR
crlmm
Rsubread

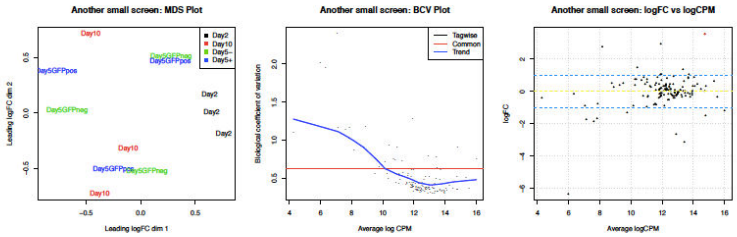


Typical R script…

```
# Read in sample & hairpin information
sampleanno = read.table("Samples2.txt", header=TRUE, sep="\t")
sampleanno

hairpinseqs = read.table("Hairpins2.txt", header=TRUE, sep="\t")
hairpinseqs[1:5,]

# Process raw sequences from fastq file
x2 = processHairpinReads("screen2.fastq", "Samples2.txt", "Hairpins2.txt", ver
# Make an MDS plot to visualise relationships between replicate samp
par(mfrow=c(1,3))
plotMDS(x2, labels=x2$samples$group, col=rep(1:4, times=3), main="An
legend("topright", legend=c("Day2", "Day10", "Day5-", "Day5+"), col=

# Begin differential representation analysis
# We will use GLMs in edgeR in this case since there are more than 2
# Set up design matrix for GLM
des = model.matrix(~x2$samples$group)
des

# Estimate dispersions
```

Compile script into a report

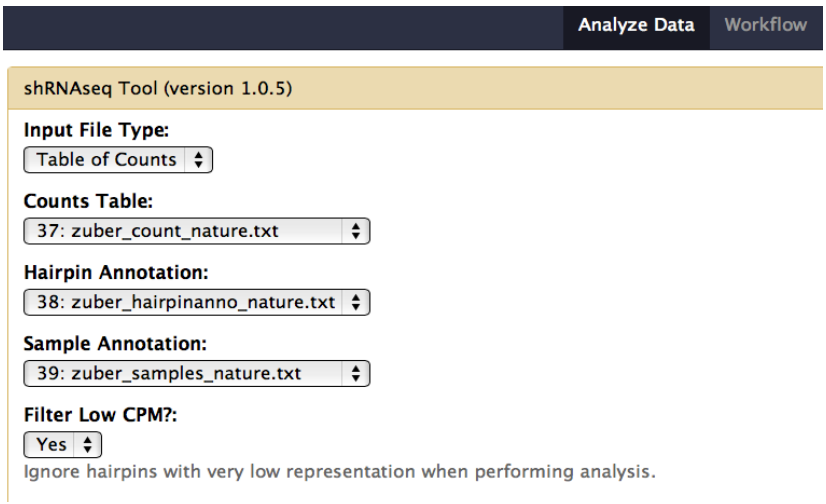
```
# Select hairpins with FDR < 0.05 to highlight on plot
thresh = 0.05
top2 = topTags(lrt, n = Inf)
top2ids = top2$table[top2$table$FDR < thresh, 1]

# Plot logFC versus logCPM
plotSmear(lrt, de.tags = top2ids, pch = 20, cex = 0.6, main = "Another small screen: logFC
abline(h = c(-1, 0, 1), col = c("dodgerblue", "yellow", "dodgerblue"), lty = 2)
```



The biological coefficent of variation (BCV) plot (middle panel) summarises the variability in the screen as a functon of hairpin abundance. These plots tend to have a characteristic shape of decreasing variability as hairpin abundance increases, which is similar to what is observed for other applications such as RNA-seq. The individual black points show hairpin-specific (referred to as *'Tagwise'* variability, while the blue line shows the trend value as hairpin abundance changes (*'Trended'*) and the red line is the common value (calculated by assuming all counts come from the same hairpin).

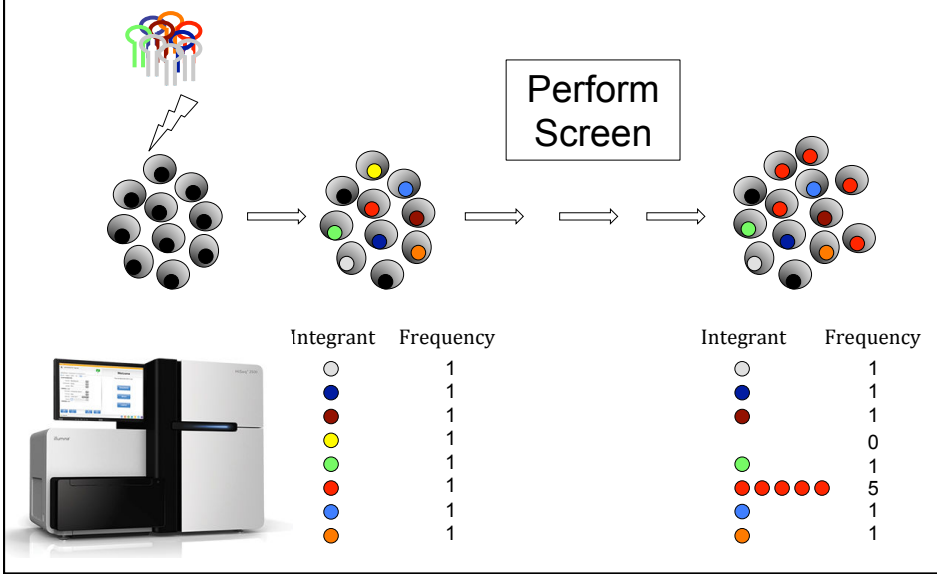
Summary: In this second small screen, the variation between replicate samples is much higher than

From scripts to Galaxy tools

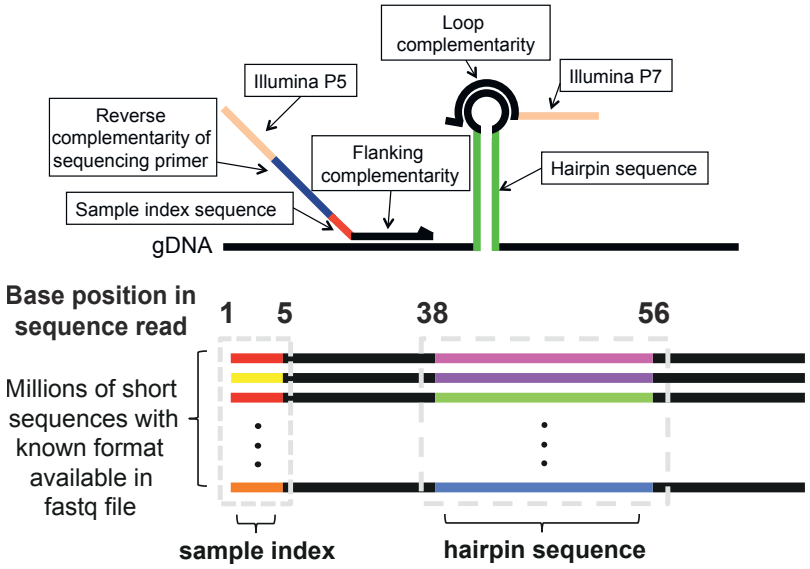


Search for 'shRNAseq' on Tool Shed

Representation of shRNAs within a pool can be established by sequencing



Amplicon sequencing



Summary of edgeR workflow

Step	Function
Data Import	`processHairpinReads`

Quality Assessment, Plotting & Normalization	`plotMDS`
	`plotBCV`
	`plotSmear`
	`calcNormFactors`

Differential Representation Analysis	`estimateDisp`
	`exactTest`
	`glmFit`
	`glmLRT`
	`topTags`

Gene-level Analysis	`camera` `roast`

Robinson, McCarthy, Smyth (2010) Bioinformatics
Dai *et al.* (2014) [in preparation]

Input files for shRNA-seq analysis

```
@B1ADDFFHHHHGIJJIJJJIJFHHGGHHHIIIJJGIJGGCHIJHIJJJJGJIGIJJHIGHHGHHF>DFD#,;A==ACCCCDDDDB####++2?BDDDDD
@HWI-ST226:335:C2T7MACXX:1:1101:1227:2083 1:N:0:GTGAGGGTTCTTTCCC
TAACGGTAGCCCCTTGAATTCCGAGGCAGTAGGCACCAGAAATTATGTGACTATATATACATCTGTGGCTNCACTAAGATCGGAAGAGCTCGTATGCCGT
+
@C@FFFFFHHHHHJJJJJJJJJJJJJJJJJJJJIJJJJJJJJJIJJJJJIHJIHIJJJJJIJJGHHHGHFF#,5?CCDDDDDDDDDDDDDDDDDDDDDDD
@HWI-ST226:335:C2T7MACXX:1:1101:1226:2110 1:N:0:GAGGGGGCTCTTTCCC
CACAAGTAGCCCCTTGAATTCCGAGGCAGTAGGCAGTTGAAAGAAGGTTATATTAAATACATCTGTGGCTNCACTAAGATCGGAAGAGCTCGTATGCCGT
+
@@?D?DDDFFFFF@GFEG,AAF;@<E?DB:CDG<0?9999B?D*/?FCFICFEEAFI;@##########################################
@HWI-ST226:335:C2T7MACXX:1:1101:1228:2125 1:N:0:GTGGGGCGTCTTTCCC
CAAACGTAGGCCCCTTGAATTCCGAGGCAGTAGGCACAAGCTTCTGGTTAGAGACAAATACATCTGTGGCNTCACTAAGATCGGAAGAGCTCGTATGCCG
+
BBBFFFFFGFHHHIJJIJ>FHHIIGIJJIJCGHIJEIIJGGIGIIJJFHHHIGGIJJJGHHHHHFFFFFC#,;?B@CDDDDDDDDDDDDDDDDDDDDDD
@HWI-ST226:335:C2T7MACXX:1:1101:1235:2194 1:N:0:GGGGTGGGTCTTTCCC
GACACGTAGCCCCTTGAATTCCGAGGCAGTAGGCAGTAGAAAGCTAACAAAGCTATACATCTGTGGCTTCNCTAAGATCGGAAGAGCTCGTATGCCGTCT
+
@C@FFFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJIIJIJGGGIIJIIJIJCHGIJJJJIHHHHGDFFF#,;=A?CDCDD@BDDDCDDDBDCDCDDBDD
@HWI-ST226:335:C2T7MACXX:1:1101:1215:2195 1:N:0:GGGTGCGCTCTTTCCC
GATACGTAGCCCCTTGAATTCCGAGGCAGTAGGCAGTAGAAAGCTAACAAAGCTATACATCTGTGGATTCNCTAAGATCGGAAGAGCTCGTATGCCGTCT
+
1114422=ADBCFGIIIGGICHF<GH?=FHGIIDGGB<FCGHI>FH>CEF@GFHEGH7@EEIH?EH)=@D#,5;5;>A>ADDB?AD<CDB<<ACC+99>B
@HWI-ST226:335:C2T7MACXX:1:1101:1241:2210 1:N:0:GGTCGCAATCTTTCCC
TAAATGTAGCCCCTTGAATTCCGAGGCAGTAGGCAGTTGGACTATCATGTTTATGTATACATCTGTGGCTNCACTAAGATCGGAAGAGCTCGTATGCCGT
+
```

Input files for shRNA-seq analysis

Samples2a.txt

ID	Sequences	group	Replicate
3	GAAAG	Day 2	1
6	GAACC	Day 10	1
9	GAAGA	Day 5 GFP neg	1
16	GAATT	Day 5 GFP pos	1
18	GACAC	Day 2	2
21	GACCA	Day 10	2
28	GACGT	Day 5 GFP neg	2
31	GACTG	Day 5 GFP pos	2
33	GAGAA	Day 2	3
40	GAGCT	Day 10	3
43	GAGGG	Day 5 GFP neg	3
46	GAGTC	Day 5 GFP pos	3

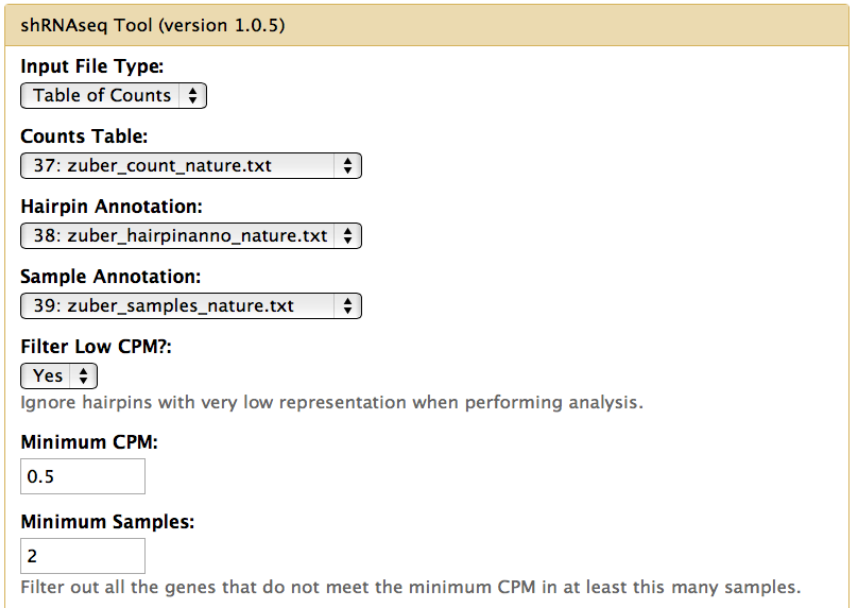
Input files for shRNA-seq analysis

Hairpins2a.txt

ID	Sequences	Gene
Control1	TCTCGCTTGGGCGAGAGTAAG	2
Control2	CCGCCTGAAGTCTCTGATTAA	2
Control3	AGGAATTATAATGCTTATCTA	2
Hairpin1	AAGGCAGAGACTGACCACCTA	4
Hairpin2	GAGCGACCTGGTGTTACTCTA	4
Hairpin3	ATGGTGTAAATAGAGCTGTTA	4
Hairpin4	CAGCTCATCTTCTGTGAAGAA	4
Hairpin5	CAGCTCTGTGGGTCAGAAGAA	4
Hairpin6	CCAGGCACAGATCTCAAGATA	4
Hairpin7	ATGACAAGAAAGACATCTCAA	7
Hairpin8	GAAGATGATGTTTATGGTTTA	7
Hairpin9	AAAGACCATAGAGAAGTTCAA	8
Hairpin10	TAGAGAAGTTCAAGACCATAA	8
Hairpin11	ACCCTTGATCTTCGAGTGCAA	8
Hairpin12	CCGAGGATAGTAGGACTTCTA	8
Hairpin13	CTGGTGTTTCTTTGCTAGGAA	8
Hairpin14	TACGACTGTGTTGTCCTCTTT	8
Hairpin15	AACGAAGAATGGGAACCTATA	8
Hairpin16	CTGCGAGACATCGACCATGAA	9
Hairpin17	AAGGTGATCCTTATGCTGTTA	9
Hairpin18	TTCATGCTGCACCAAGATCTA	9
Hairpin19	CACAGGACAGGTGGACGTCAA	9

Demo

- Analyse data from Zuber *et al.* (2011) Nature
- AML mouse model
- ~ 1000 shRNAs (3-6 per gene) targeting known chromatin regulators
- Samples taken at Day 0 and Day 14. Hits identified by comparing hairpin abundance between these two time points, and looking for shRNAs that drop out over time



Analysis Type:

Generalised Linear Model

Classic Exact Tests are useful for simple comparisons across two sampling groups. Generalised linear models allow for more complex contrasts and gene level analysis to be made.

Contrasts of interest:

Day 14 - Day 0

Specify equations defining contrasts to be made. Eg. KD-Control will result in positive fold change if KD has greater expression and negative if Control has greater expression.

Perform Gene Level Analysis?:

Yes

Analyse LogFC tendencies for hairpins belonging to the same gene.

Minimum Hairpins:

2

Only genes with at least this many hairpins will be analysed.

Gene Selection Method:

By Gene Identifier

Symbols of Genes to Plot:

Brd4

Select genes based on their identifier in the 'Gene' column of the sample information file. Please ensure exact match with the values in input file and separate selections with commas.

FDR Threshold:

0.05

All observations below this threshold will be highlighted in the smear plot.

Absolute LogFC Threshold:

6.0

In additional to meeting the FDR requirement, the absolute value of the log-fold-change of the observation must be above this threshold to be highlighted.

Execute

EdgeR Analysis Output:

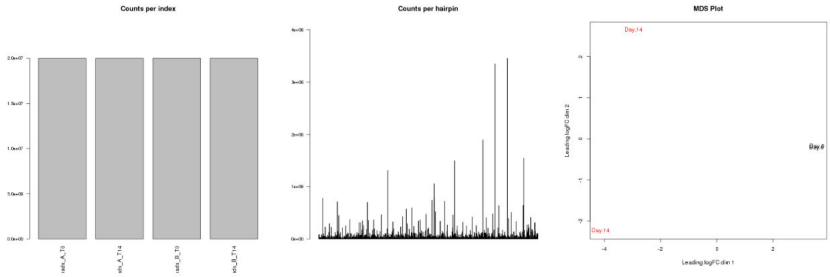
Input Summary:

- Number of Samples: 4
- Number of Hairpins: 1105
- Number of annotations provided: 1105
- Number of annotations matched to hairpin: 1105

The estimated common biological coefficient of variation (BCV) is: 0.9264115

Output:

All images displayed have PDF copy at the bottom of the page, these can exported in a pdf viewer to high resolution image format.



Plots:

Counts per Index Barplot (.pdf)
Counts per Hairpin Barplot (.pdf)
MDS Plot (.pdf)
Smear Plot(Day.14-Day.0) (.pdf)
Barcode Plot(Day.14-Day.0) (.pdf)

Tables:

Top Tags Table(Day.14-Day.0) (.tsv)
Gene Level Analysis Table(Day.14-Day.0) (.tsv)

alt-click any of the links to download the file, or click the name of this task in the galaxy history panel and click on the floppy disk icon to download all files in a zip archive.

.tsv files are tab seperated files that can be viewed using Excel or other spreadsheet programs

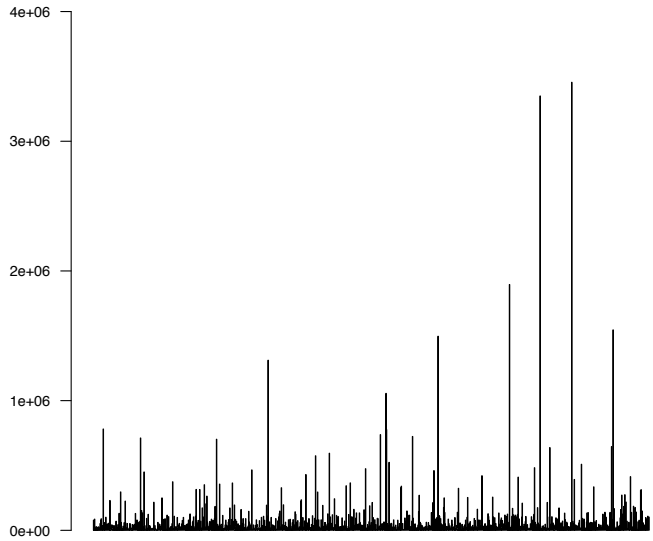
Task started at: 2014-03-23 18:30:07
Task ended at: 2014-03-23 18:30:50

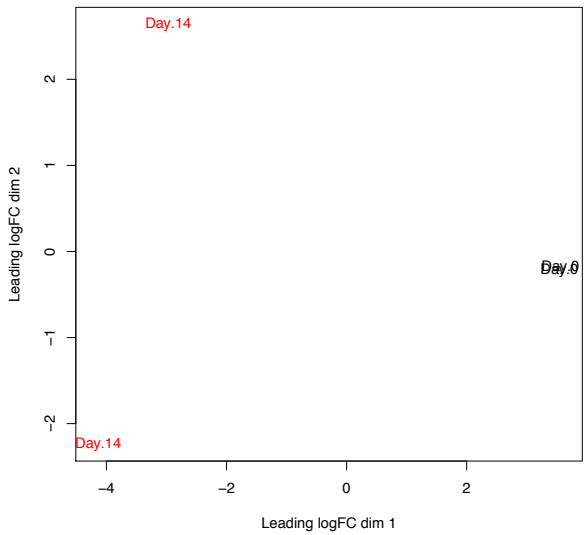
EdgeR Analysis Output:

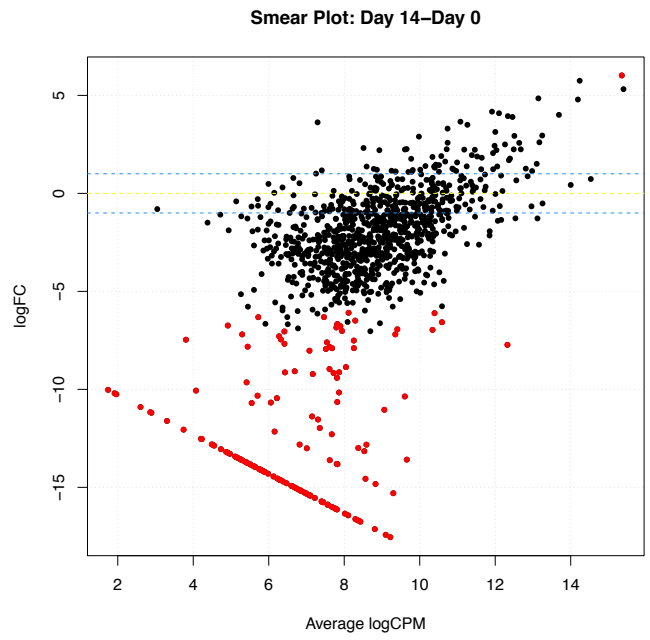
Input Summary:

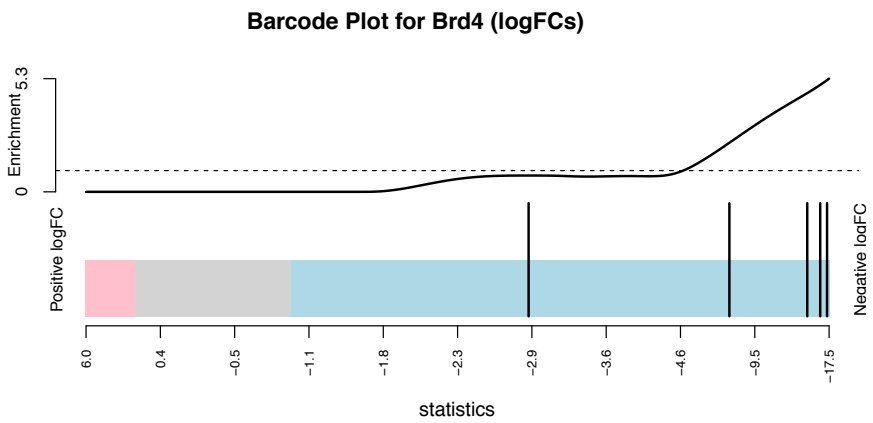
- Number of Samples: 4
- Number of Hairpins: 1105
- Number of annotations provided: 1105
- Number of annotations matched to hairpin: 1105

The estimated common biological coefficient of variation (BCV) is: 0.92641


Counts per hairpin
4e+06
3e+06
2e+06
1e+06
0e+00


MDS Plot
Day.14
Day.0
Day.0
Day.14
Leading logFC dim 2
Leading logFC dim 1
−4
−2
0
2
−2
−1
0
1
2





P-value = 0.0003

Wu *et al.* (2010) Bioinformatics

Future work: RNA-seq analysis workflow

- Goal is to put together a workflow that uses WEHI developed tools:
  - *subread* for mapping short reads to the genome
  - *featureCounts* to obtain gene-level summaries in each sample
  - *limma-voom* to perform differential expression analysis

Liao *et al.* (2013) Nucleic Acids Res
Liao *et al.* (2013) Bioinformatics
Law *et al.* (2014) Genome Biology

Acknowledgements

Molecular Medicine
Shian Su
Jenny Dai
Cynthia Liu

Toby Sargeant
Jarny Choi
Nick Seidenman

Jamie Gearing
Darcy Moore
Natasha Jansz
Kelan Chen
Andrew Keniry
Marnie Blewitt

Mark McKenzie
Ross Dickins

Doug Hilton

Bioinformatics
Gordon Smyth
Yunshen Chen
Aaron Lun
Yang Liao, Wei Shi
Matthew Wakefield
Keith Satterley

Stem Cells and Cancer
Julie Sheridan

Laura Galvis
Marie-Liesse Asselin-Labat

CSCD
Iris Tan, Grant Dewson

IMP, Vienna
Johannes Zuber

AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Australian Government
National Health and Medical Research Council