Affymetrix analysis pipeline shows mislabelling in experimental design.

FGT ICA, B217754 Microarray processing

Brief background on dataset

The original dataset was part of an experiment investigating the role of Calcium (Ca) in the development and pathogenesis of diet-induced obesity and insulin resistance (1). In additional to faecal sample assessment, the experimental included a genomic assessment using an Affymetrix Mouse Genome 430 2.0 array (Mouse430_2). The experimental setup was characterised by a treatment with one control and one treatment level of Calcium in diet (seen in Table 1). In the course of its analysis, this report manages systematic variation by normalising the data, controls error rates in the findings by applying multiple testing correction applies an accepted FDR rate and logFC threshold to assess the presence of biological variation in the experiment.

Table 1 Overall experimental design of the orignial dataset.

Control (3 samples)	Treatment (3 samples)
Low/LCA= 50mmol/kg	High/HCa= 150mmol/kg

Although (for the purpose of this ICA) the experiment was required to have a single level of investigation with 3 replicates (treatment vs. control), this was not the case. Instead the samples were paired by the region of the small intestine from which the sample was excised. This means that samples were not in replicates of 3, and all of my further analysis must assume (incorrectly) that the differential gene expression signatures are differential across the whole small intestine. Therefore, whilst this experiment treats the data as replicates of 3, the true biological signature of each region of the small intestine requires another repeat each for statistical validity. In a real analysis, I would have asked for an additional sample from each intestinal region and performed analysis between two regions. With further domain knowledge, I would check for specific internal control genes to ensure that the data quality was in the same realm of those expected and that experimental design had captured control genes.

Part 1

EDA and QC

Table 2 Data from the targetfile_B217754.txt file used to cohesively format experimental data in this pipeline. Curated from the GEO accession viewer (2), information such as an abbreviated Name and a Colour scheme were added.

Name	Filename	Sample	Group	Colour	Description
HCa.1	GSM213045.CEL	GSM213045	High_calcium	blue	SI1_wk2_HF (high fat diet high calcium)
HCa.2	GSM213051.CEL	GSM213051	High_calcium	blue	SI2_wk2_HF (high fat diet high calcium)
HCa.3	GSM213057.CEL	GSM213057	High_calcium	blue	SI3_wk2_HF (high fat diet high calcium)
LCa.1	GSM462207.CEL	GSM462207	Low_calcium	red	SI1_wk2_HF (high fat diet low calcium)
LCa.2	GSM462208.CEL	GSM462208	Low_calcium	red	SI2_wk2_HF (high fat diet low calcium)
LCa.3	GSM462209.CEL	GSM462209	Low_calcium	red	SI3_wk2_HF (high fat diet low calcium)

Using the annotations from the targets file previewed in Table 2), present/absent counts were used to show how well the wet-lab portion of the investigation succeeded in capturing any genes present. Typical target ranges for a chip such as the Mouse430_2 which captures a range of gene families is 40%-60%. As seen in Table 3, the data from this experiment was satisfactory with over 45% genes present in the experiment, allowing investigations to continue with all samples included.

Table 3 Manually curated Present/absent calls, with an overall statement on the % of genes present in the data.

Sample	Absent	Present	Present (%)
LCa.1	22731	20971	47.99
LCa.2	23629	20197	46.08
LCa.3	23314	20578	46.88
HCa.1	22518	21334	48.65
HCa.2	23399	20407	46.58
HCa.3	23280	20528	46.86

To further this assessment of the presence of gene signals, Figure 1 shows the log intensities of the samples. We expect to see an equivalent peak and distribution across all samples, which would again indicate that wet-lab probe-base capture had succeeded and there was some of the desired biology occurring in the samples.

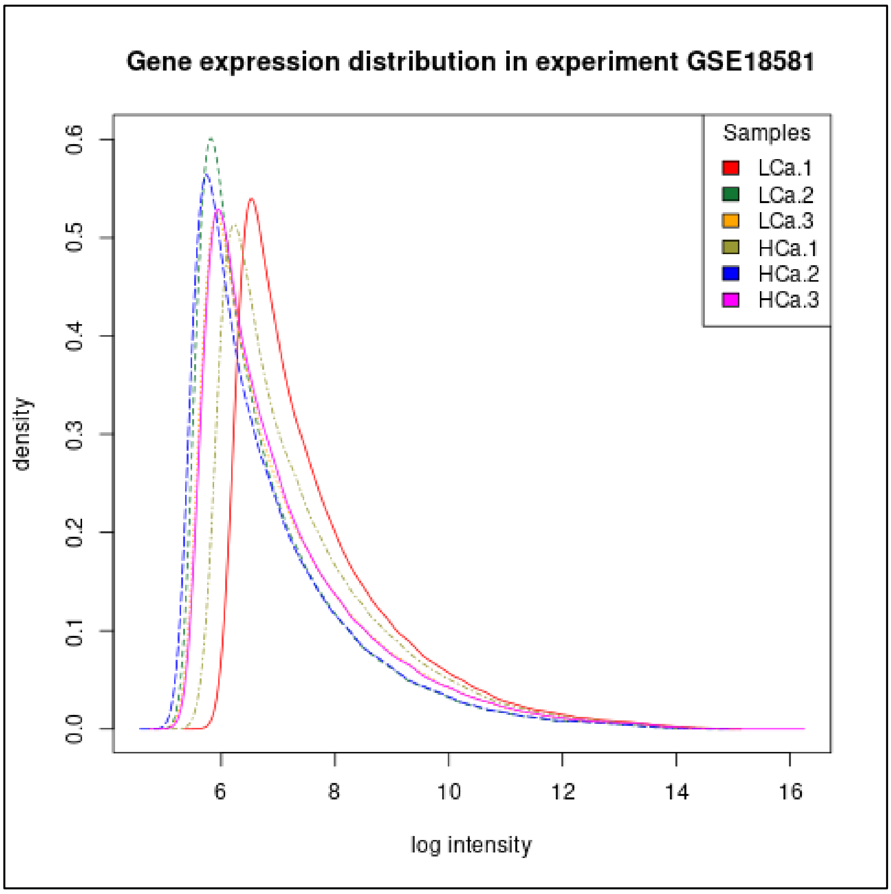


Figure 1 A Histogram of Log intensity scores for the data in each sample. Peaks appear broadly similar, with two distinct peaks separating from the rest at log intensity ~5.8.

Figure 1's peak are all similar, sitting within 0.5-0.6 density and between log (6)-log (7) intensity. There seems to be no discernible trend between samples of the same level of control (Lo or High Ca), but there are two peaks at ~5.8 log intensity, which correspond to LCa.2 and HCa.2, both of which were sampled from the same region of intestine. Since more signals were detected by the gene probesets in the mouse 430_2 chip in this region of the small intestine, it may be possible that this region has a more diverse gene expression profile or has higher expression of the same genes overall compared to the other samples. To statistically show this, we would need another sample from the '.2' section of the small intestine, which would give the ability to calculate an average peak.

Data normalisation with RMA

The shape of the log intensities had indicated a skewed distribution in expression levels, and boxplot assessments of the samples showed this, as seen in Figure 2.

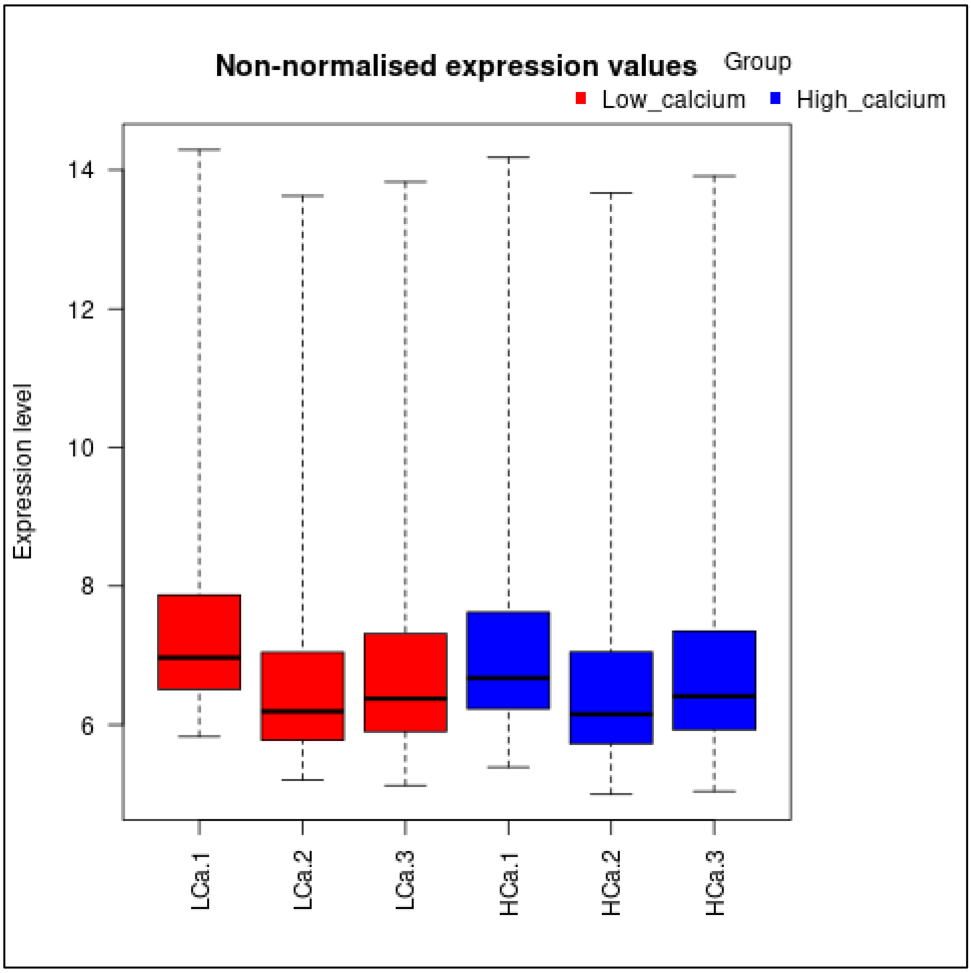


Figure 2 A boxplot of the sample sets before any data treatment has occurred. There is variety in the distributions of each sample set which requires normalisation.

The non-normalised expression values showed a noticeable variation across the samples in the median, IQC and outer boundaries of their expression levels. However, despite significant overall in their distributions (internal IQCs overlap) there appeared to be some correspondence between samples from the same region of the intestine, showing that region '.1' had higher overall expression before normalisation than '.2' and '.3'. This again highlighted the need for more samples in this facet to determine possible average differences between intestinal regions.

RMA normalisation was performed to reduce systematic biases in the data, eliminating the possible influences of intensity bias and unique chip variation. The default RMA method of quantile normalisation was used, and the resulting distributions are visible in Figure 3. Fitting the sample data to an average reference distribution displayed some gene expression values as outliers, however, the ranges of each sample were now far more congruent.

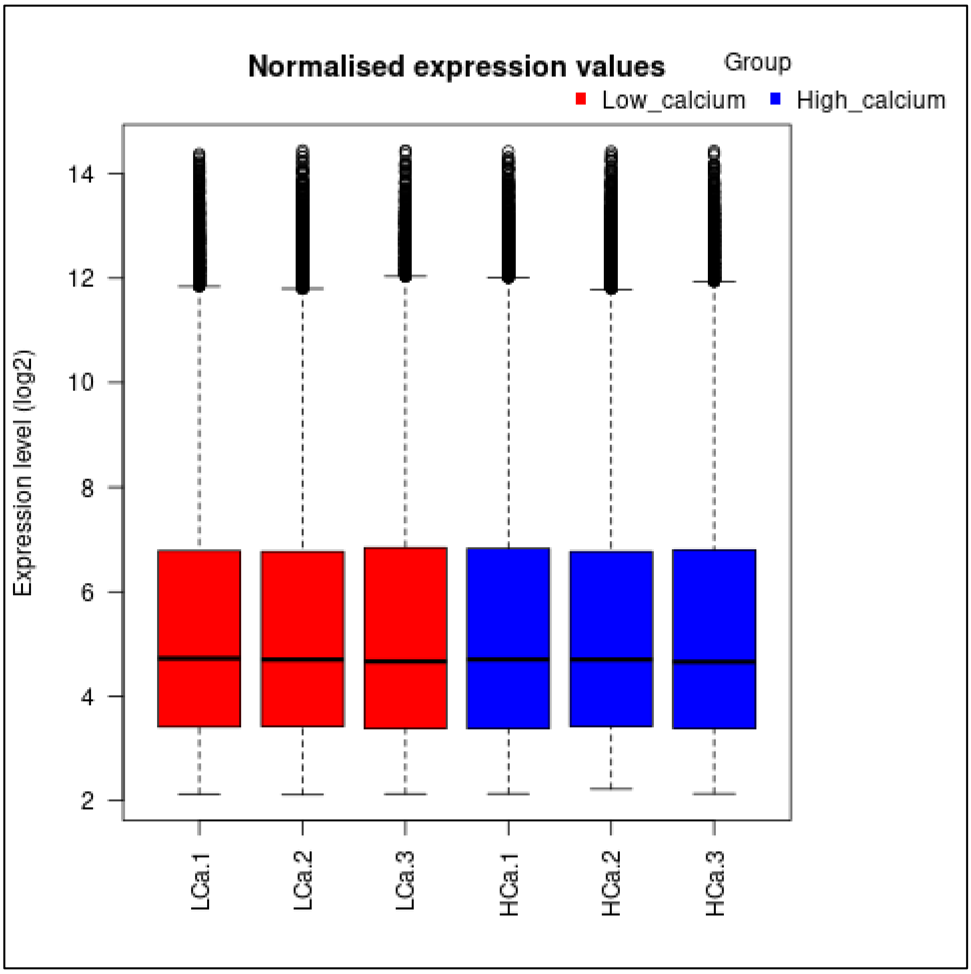


Figure 3 A boxplot of normalised data, showing that the sample sets have been normalised to a reference distribution (and therefore appear to have many outliers).

To further assess the efficacy of the data normalisation and to gain an indication of gene expression changes between the samples, M vs A plots of the log2-transformed fold-change (M) versus the log2-transformed average expression (A) were visualised in Fig.4 . In appropriateley normalised data, the majority of would likely centre along the horizontal 0 line.

Several points lay outside the central region of each plot, where the log2-fold change is close to zero, which mirrored the outlier identification fo the boxplots. This was least pronounced in the comparison of LCa.1 and LCa.2, where outliers only reached absolute fold changes of ~0.3. Although there was some dispersion in these data, there was little evidence ofsystematic bias such as normalization issues or remenant sample quality issues from before normalisation.

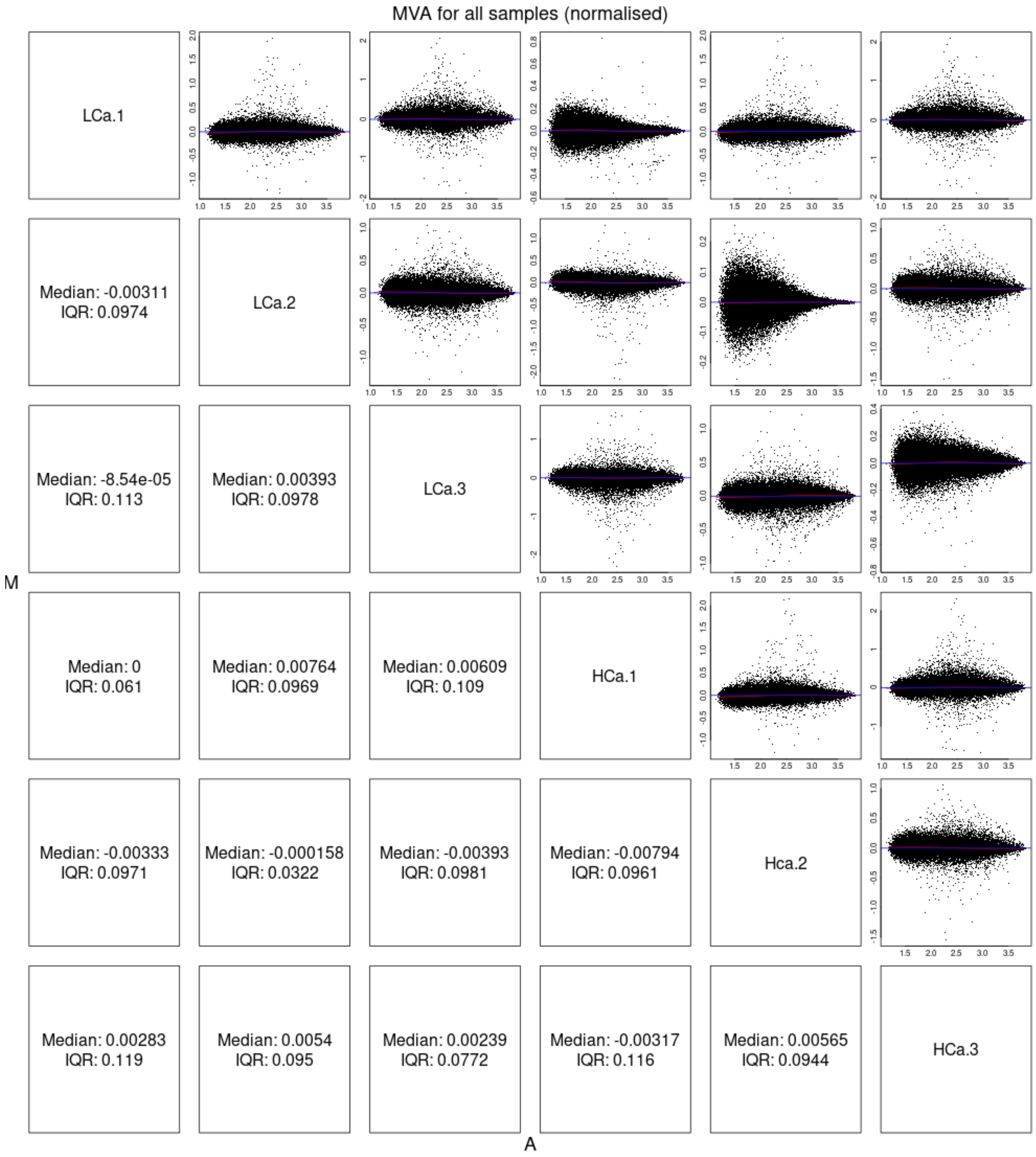


Figure 4 MvsA (MA) plot showing the average expression of each sample on a pair-wise comparison.

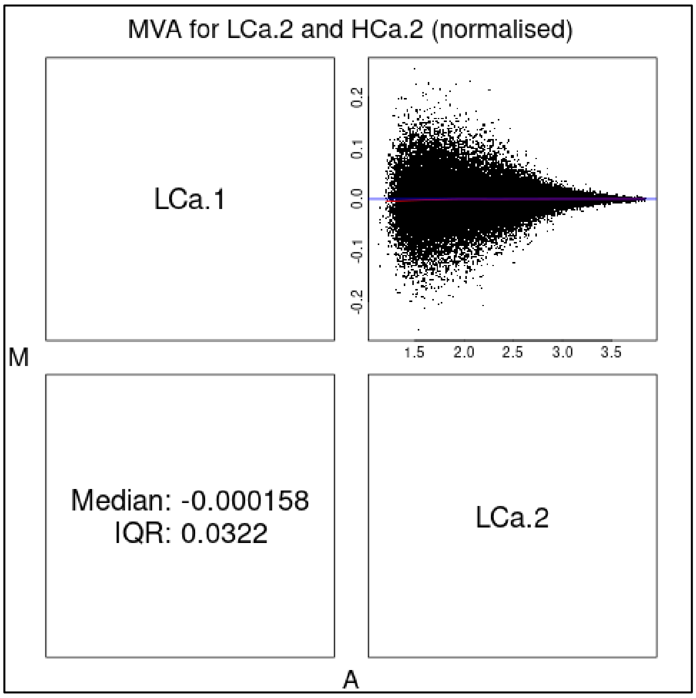


Figure 5 MvsA plot comparing the samples LCa.2 and HCa.2. The spread of the data near A 1.75 suggests that the average expression was lower overall.

Shown in more detail in Fig.5, the HCa.2/LCa.2 plot showed a prominent leftward shift towards A=1.75, resulting a lack of the typical ‘cigar’ shape distribution. This suggests that a set of certain genes were expressed less than others overall, which in this context highlights the discrepency between two biological signatures of the same treatment level in this experiment. Once again, this indicates un-verifiable differences in the genomic profiel of small intestinal regions.

Given that shortcomings in the design of the experiement were becoming evident in the distribution of the data, it was important to determine the relation of each sample to the other. A cluster dedrogram (Fig.6) by global average of similarity across all genes and Pearson correlation tests showed that samples of each treatment level (Low/High Ca) did pair up depending on the region of intestine which they originated from, however the grouping based on gene expression values was distinctly into 3 pairs and this was satisfactory for this report.

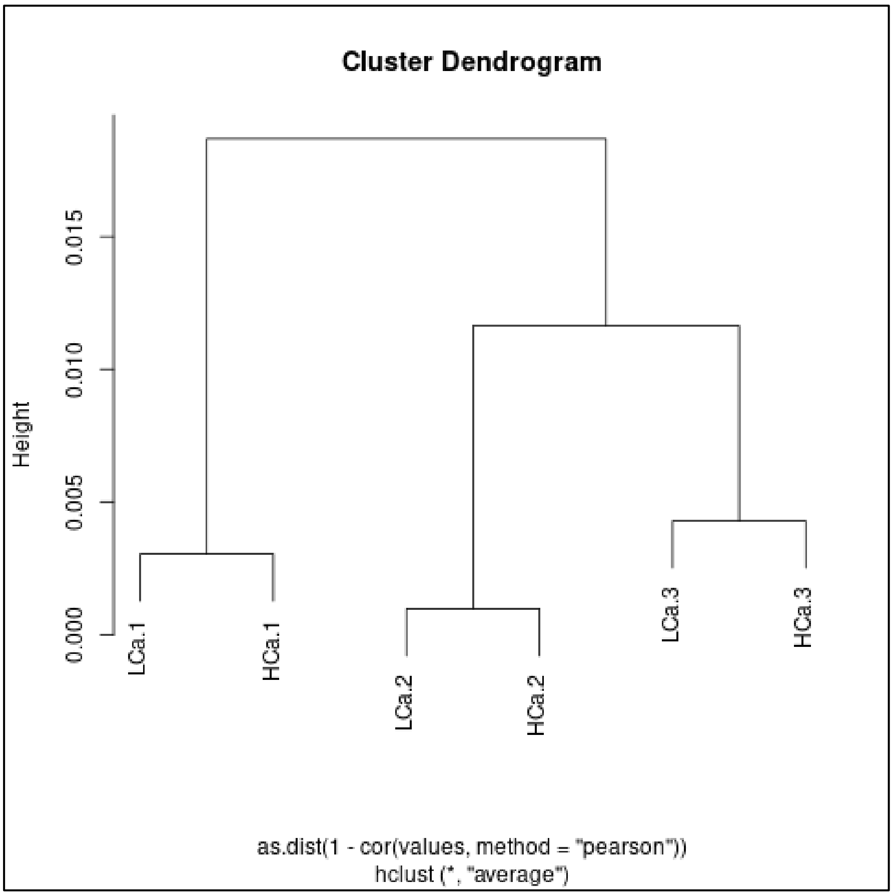


Figure 6 Cluster dendrogram showing the overall relationship between the samples based on expression values.

To gain a more obvious perspective on the differences in the data's relation, a PCA plot was used. Initial plotting showed that samples differ largely in PC3, so the plot presented in Figure 7 is focussed on this PC. Interestingly, the samples from regions '.2' and '.1' cluster closely in the three-dimensional space, suggesting a close overall similarity of gene expression levels. However, the samples from region '.3' do not cluster closely to each other, nor to the other samples of their same treatment level (HCa/LCa). Whilst a large separation of a treatment and control sample such as that between LCa.3 and HCa.3 would usually be encouraging of a significant difference in expression levels, the confounding factor here of region keeps the nature of the differences ambiguous.

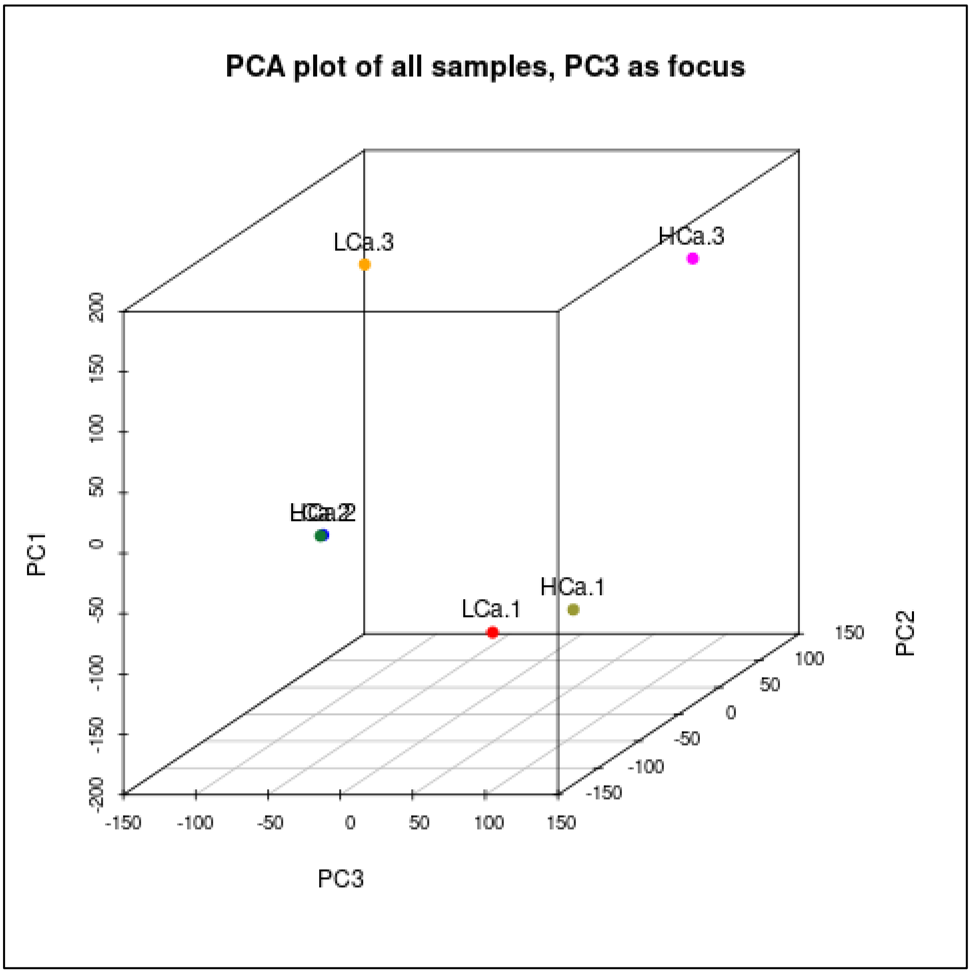


Figure 7 PCA plot of the samples in the dataset, reframed to focus on PC3 for interpretability.

A final check of the presence of genes in the treatment levels of this experiment was visualised with a Venn Diagram in Figure 8. As can be seen, only 2 genes were not expressed in the LCa level of treatment which were expressed in HCa samples. This could indicate the activation of new genes under treatment conditions, and indicates the next direction for candidate gene selection in further investigation.

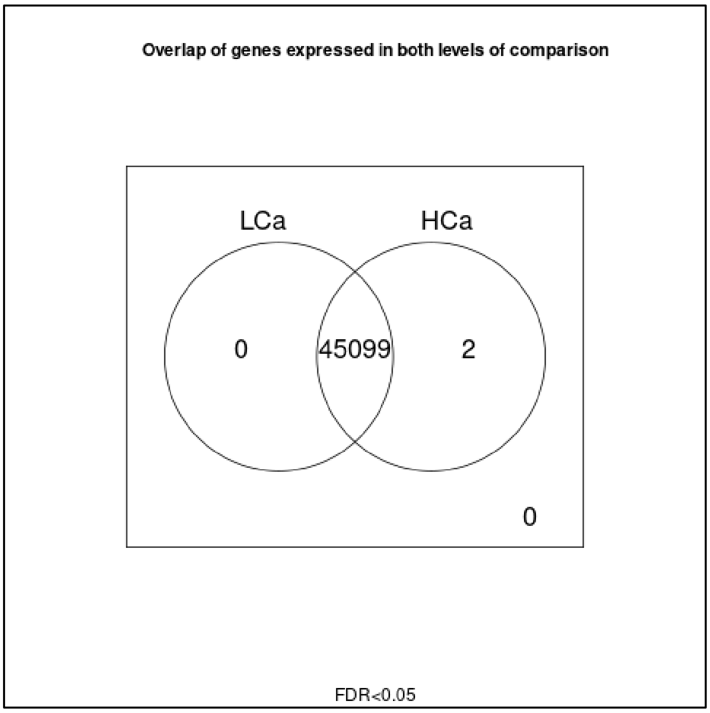


Figure 8 Venn Diagram showing how many genes were expressed in both the LCa and Hca treatment levels. FDR of 0.05.

Signal analysis (Part 2)

Differentially expressed genes

Manual calculation of the top 10 log2FC with mean log expression values was performed to give a crude estimation of the differentially expressed genes in the treatment samples. The genes 9seen in Table 5) were primarily related to gastrointestinal homeostasis as expected from the sample type.

Table 4 Manually curated log2FC values, reported in log2 for interpretability.

Fold change (Log2)	Probeset ID	Gene Symbol	Gene Name
2.68	1453132_a_at	Gkn2	gastrokine 2
1.74	1429286_at	Gkn3	gastrokine 3
1.67	1438648_x_at	Gkn3	gastrokine 3
1.65	1423404_at	Gkn1	gastrokine 1
1.55	1422448_at	Tff2	trefoil factor 2 (spasmolytic protein 1)
1.44	1437340_x_at	Gkn1	gastrokine 1
1.09	1428942_at	Mt2	metallothionein 2
0.86	1422352_at	Mcpt1	mast cell protease 1
0.79	1448156_at	Tff1	trefoil factor 1
0.72	1422557_s_at	Mt1	metallothionein 1

Functional Enrichment Analysis

To determine if a differentially expressed gene in the dataset was related to an enriched or depleted functional process, functional enrichment analysis was performed with `Mroast`. This functional enrichment tool was chosen as it handles self-contained data from a single experiment. As can be seen in Table 5, Mroast asseses the bulk differential expression in the dataset under treatment and identifies the biological processes which is likely to be affected. Downregulation of apoptosis and upregulation of mTOR signalling were among the most confident predictions, which suggests the overall profile of a possible oncotic progression. It is worth noting that the results in Table 5 all have very poor FDR values (>0.5), meaning that after adjustment from regular p-values with the BH method (discussed in Methods), it is likely that over half of the genes believed to be differentially expressed were false discoveries (false positives) and they are therefore unreliable findings. Only the top 4 pathways are shown in Table 5, as anything beyond was at an FDR of 1, meaning all genes associated and differentially expressed were likely to be false discoveries.

Table 5 Summary of the top four entries from mroast analysis

Process name	NGenes	PropDown	PropUp	Direction	PValue	FDR	PValue.Mixed	FDR.Mixed
HALLMARK_APOPTOSIS	476	0.025210084	0.010504202	Down	0.014	0.51875	0.9495	0.97375
HALLMARK_ESTROGEN_RESPONSE_EARLY	654	0.024464832	0.016819572	Down	0.021	0.51875	0.9385	0.97375
HALLMARK_ESTROGEN_RESPONSE_LATE	561	0.023172906	0.012477718	Down	0.0435	0.720833333	0.9305	0.97375
HALLMARK_MTORC1_SIGNALING	633	0.001579779	0.023696682	Up	0.181	0.97975	0.915	0.97375

Part 2

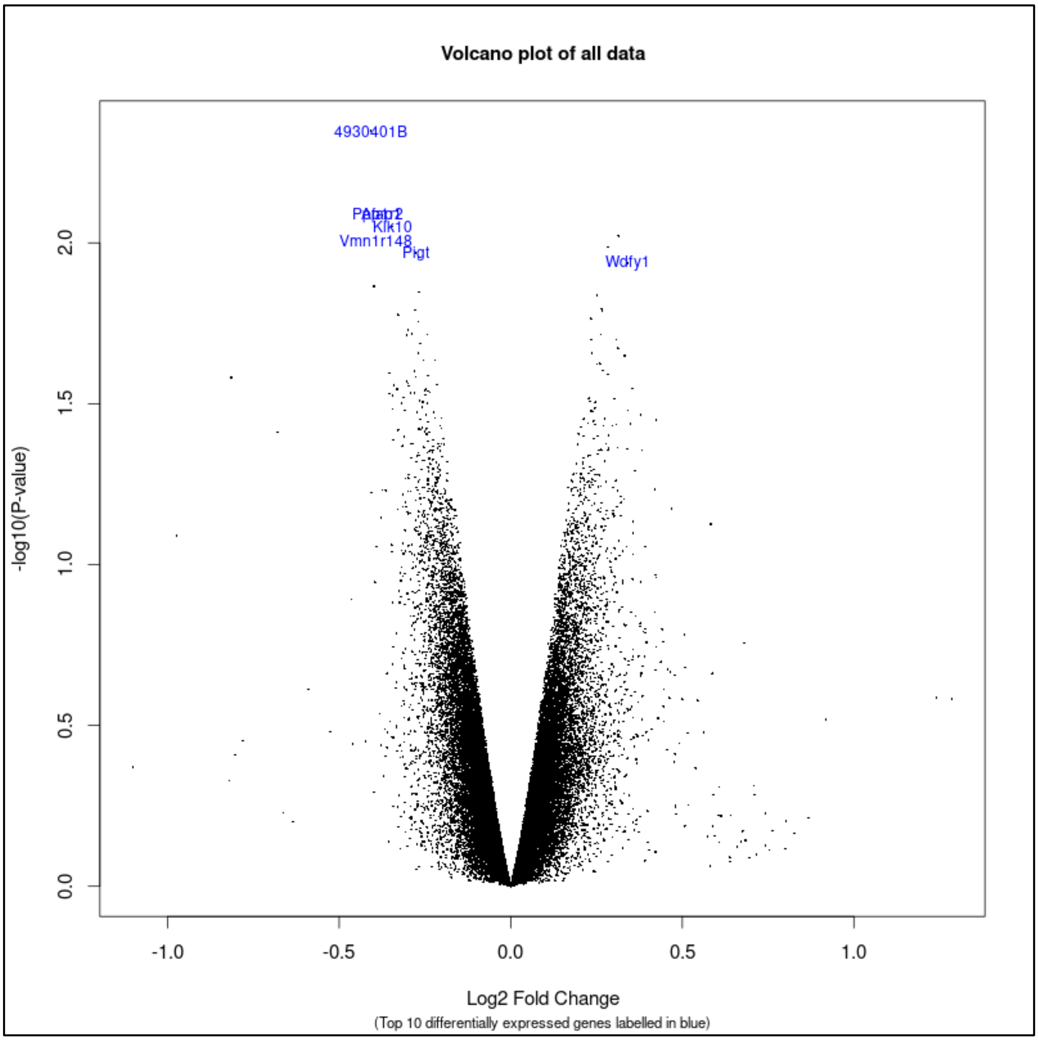
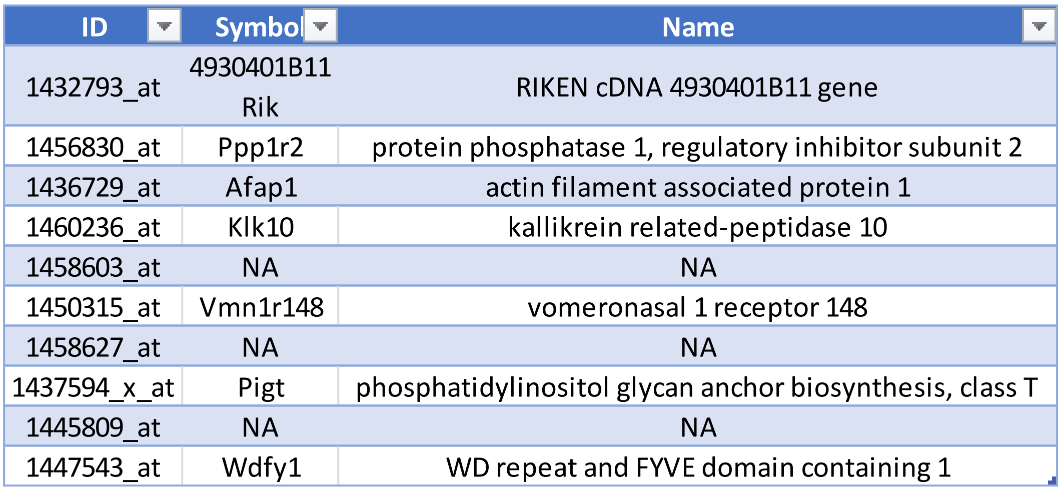


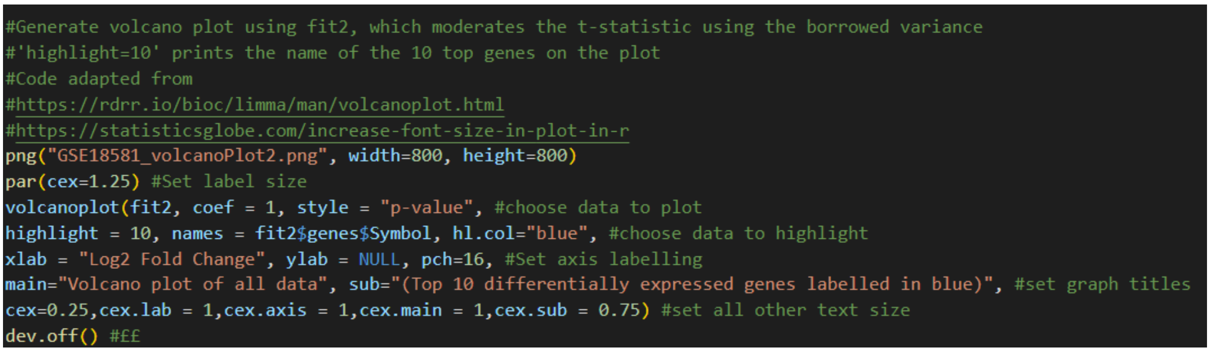
Figure 9 Volcano plot highlighting the top 10 differentially expressed genes according to FDR.

To visualise the nature of differential expression throughout the dataset, a volcano plot was generated and the top 10 differentially expressed genes (identified by the plotting command as those with the lowest adjusted p-value) were highlighted in Figure 15. The names of these genes are shown in Table 6 for legibility. In addition to these FDR-determined differentially expressed genes, there appears to be several points to the horizontal boundaries of the plot which indicate highly different log2 Fold Change signatures. These are also likely to be of interest in follow-up investigation of the affected biological processes, though none of these correspond to those found in Table 4's manually curated list.

Table 6 showing the names of the top 10 differentially expressed genes according to the volcano plot, for clarity.



ID	Symbol	Name
1432793_at	4930401B11 Rik	RIKEN cDNA 4930401B11 gene
1456830_at	Ppp1r2	protein phosphatase 1, regulatory inhibitor subunit 2
1436729_at	Afap1	actin filament associated protein 1
1460236_at	Klk10	kallikrein related-peptidase 10
1458603_at	NA	NA
1450315_at	Vmn1r148	vomeronasal 1 receptor 148
1458627_at	NA	NA
1437594_x_at	Pigt	phosphatidylinositol glycan anchor biosynthesis, class T
1445809_at	NA	NA
1447543_at	Wdfy1	WD repeat and FYVE domain containing 1



```
#Generate volcano plot using fit2, which moderates the t-statistic using the borrowed variance
#'highlight=10' prints the name of the 10 top genes on the plot
#Code adapted from
#https://rdrr.io/bioc/limma/man/volcanoplot.html
#https://statisticsglobe.com/increase-font-size-in-plot-in-r
png("GSE18581_volcanoPlot2.png", width=800, height=800)
par(cex=1.25) #Set label size
volcanoplot(fit2, coef = 1, style = "p-value", #choose data to plot
highlight = 10, names = fit2$genes$Symbol, hl.col="blue", #choose data to highlight
xlab = "Log2 Fold Change", ylab = NULL, pch=16, #Set axis labelling
main="Volcano plot of all data", sub="(Top 10 differentially expressed genes labelled in blue)", #set graph titles
cex=0.25,cex.lab = 1,cex.axis = 1,cex.main = 1,cex.sub = 0.75) #set all other text size
dev.off() #££
```

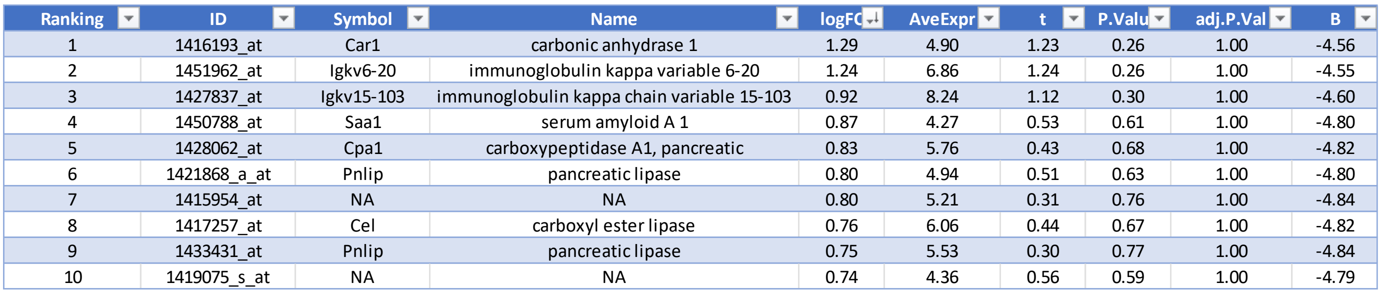
Figure 10 Code snippet showing the command used to generate the volcano plot for the highlighting of differentially expressed genes.

Similarly to the message of the functional enrichment results, it is important to employ both statistical values such as adjusted p-values and effect size measures such as Fold change to properly identify differentially expressed genes. These concepts are two halves of the overall image of a difference arising from treatment because fold change would give the magnitude and direction of the change in expression, whilst FDR filtering gives the confidence that said finding is true and can be relied on. To this end selecting differentially expressed genes by thresholding with FDR and Fold change will reduce findings and probably produce lists which differ to using FDR alone, because the FDR selects on the chance that the finding is true whilst Fold change ranks findings by their effect size.

Generating two lists of differentially expressed genes using topTable and topTreat.

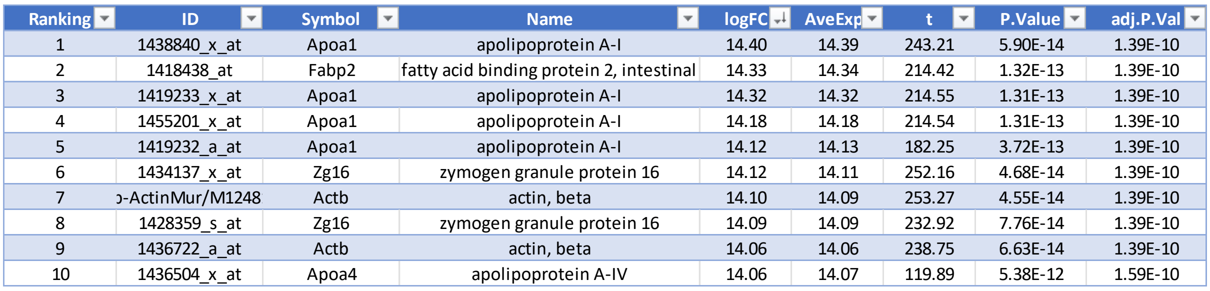
Each table was filtered by a fold change value (0.5, meaning a 1.5 times increase in expression) and a chosen value of FDR (0.05). topTreat showed 45101 differentially expressed genes whilst topTable showed 86 values after filtering for both metrics. There is liekly an error with the FDR value of topTable which led to this problem, as the log FC values (seen in Table 7) seem appropriate whilst the adj.P.value does not. Even so, there were no genes with a logFC larger than 1.29 according to topTable. Neither of the genes in either top 10 list appeared in the other, leading me to believe that there was an underlying issue with the R code used to generate them.

Table 7 topTable top10 differentially expressed genes.



Ranking	ID	Symbol	Name	logFC	AveExpr	t	P.Valu	adj.P.Val	B
1	1416193_at	Car1	carbonic anhydrase 1	1.29	4.90	1.23	0.26	1.00	-4.56
2	1451962_at	Igkv6-20	immunoglobulin kappa variable 6-20	1.24	6.86	1.24	0.26	1.00	-4.55
3	1427837_at	Igkv15-103	immunoglobulin kappa chain variable 15-103	0.92	8.24	1.12	0.30	1.00	-4.60
4	1450788_at	Saa1	serum amyloid A 1	0.87	4.27	0.53	0.61	1.00	-4.80
5	1428062_at	Cpa1	carboxypeptidase A1, pancreatic	0.83	5.76	0.43	0.68	1.00	-4.82
6	1421868_a_at	Pnlip	pancreatic lipase	0.80	4.94	0.51	0.63	1.00	-4.80
7	1415954_at	NA	NA	0.80	5.21	0.31	0.76	1.00	-4.84
8	1417257_at	Cel	carboxyl ester lipase	0.76	6.06	0.44	0.67	1.00	-4.82
9	1433431_at	Pnlip	pancreatic lipase	0.75	5.53	0.30	0.77	1.00	-4.84
10	1419075_s_at	NA	NA	0.74	4.36	0.56	0.59	1.00	-4.79

Table 8 topTreat top 10 differentially expressed genes



Ranking	ID	Symbol	Name	logFC	AveExp	t	P.Value	adj.P.Val
1	1438840_x_at	Apoa1	apolipoprotein A-I	14.40	14.39	243.21	5.90E-14	1.39E-10
2	1418438_at	Fabp2	fatty acid binding protein 2, intestinal	14.33	14.34	214.42	1.32E-13	1.39E-10
3	1419233_x_at	Apoa1	apolipoprotein A-I	14.32	14.32	214.55	1.31E-13	1.39E-10
4	1455201_x_at	Apoa1	apolipoprotein A-I	14.18	14.18	214.54	1.31E-13	1.39E-10
5	1419232_a_at	Apoa1	apolipoprotein A-I	14.12	14.13	182.25	3.72E-13	1.39E-10
6	1434137_x_at	Zg16	zymogen granule protein 16	14.12	14.11	252.16	4.68E-14	1.39E-10
7	ɔ-ActinMur/M1248	Actb	actin, beta	14.10	14.09	253.27	4.55E-14	1.39E-10
8	1428359_s_at	Zg16	zymogen granule protein 16	14.09	14.09	232.92	7.76E-14	1.39E-10
9	1436722_a_at	Actb	actin, beta	14.06	14.06	238.75	6.63E-14	1.39E-10
10	1436504_x_at	Apoa4	apolipoprotein A-IV	14.06	14.07	119.89	5.38E-12	1.59E-10

Reporting a ranked list (non-parametric test) instead of thresholding, could be a reasonable alternative way to report data, but required caution from the reader to understand that the FDR values show the confidence of a true finding, and the difference between the p-value metrics. It is possible to rank all genes by intensity. When there are little to no replicate levels, there is no data to create a confident and verifiable conclusion of the more intense gene., so this could lead to more erroneous results, as at least 5 replicates are required to create statistical confidence.

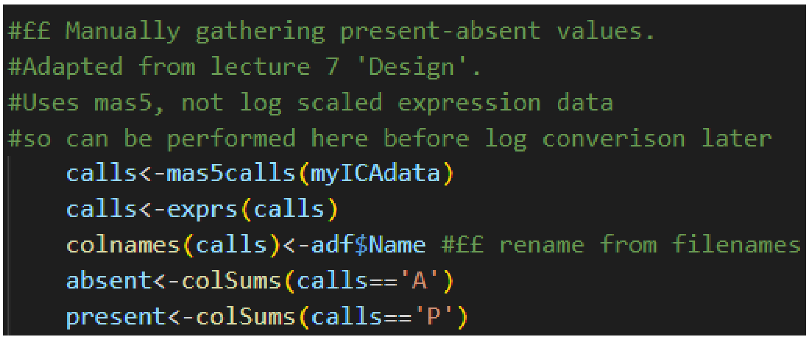
Methods

The R script *FGT_ICA_B217754.R* was written and run on R version 4.2.2 to execute the following pipeline (for further links to library sources, please see the script). The script is submitted in coordination with this report and uses the tutorial materials from the Function Genomic Technologies course except where '££' symbols denote major changes or insertions of code.

Experimental data was downloaded from the GEO accession viewer (accession GSE18581) (2) and CEL files were extracted from the tar file. A targets file was manually curated by considering details of the experimental design in the webpage's *Samples* section. After reading in the data, the target file was used to rename the samples and the ReadAffy command from the library affy converted CEL files to expression data in R.

Quality Assessment and Control

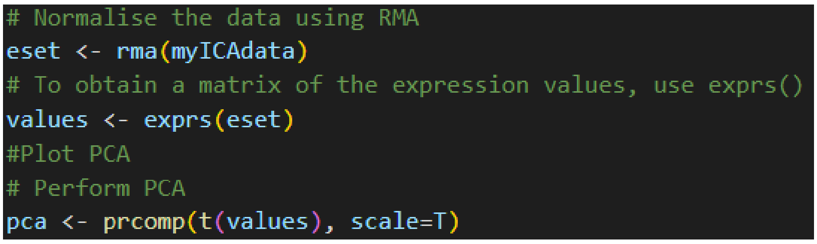
Present/absent counts were gathered by using `mas5calls` in place of an `affyQC` report (results held in 'Presence_metrics.csv'), as this version of R was incompatible with the Bioconductor library `affyQC`. A histogram of log intensities was also generated to verify the activity of each sample.



```
#££ Manually gathering present-absent values.
#Adapted from lecture 7 'Design'.
#Uses mas5, not log scaled expression data
#so can be performed here before log converison later
    calls<-mas5calls(myICAdata)
    calls<-exprs(calls)
    colnames(calls)<-adf$Name #££ rename from filenames
    absent<-colSums(calls=='A')
    present<-colSums(calls=='P')
```

Figure 11 Code snippet showing the manual curation of Present/Absent calls for QC

Data normalisation was performed using RMA and the normalised data was plotted as a boxplot, then through an MvsA plot. Hierarchical clustering of the normalised data was performed, as well as PCA to assess the relationship between each sample in terms of their expression values.

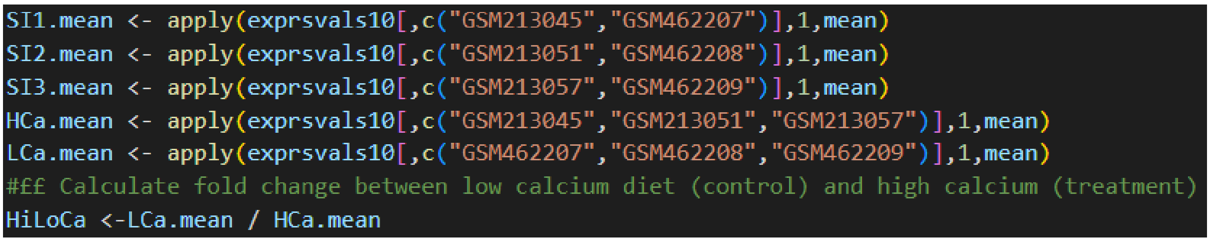


```
# Normalise the data using RMA
eset <- rma(myICAdata)
# To obtain a matrix of the expression values, use exprs()
values <- exprs(eset)
#Plot PCA
# Perform PCA
pca <- prcomp(t(values), scale=T)
```

Figure 12 Code snippet of data normalisation, expression value extraction and PCA analysis.

Analysis

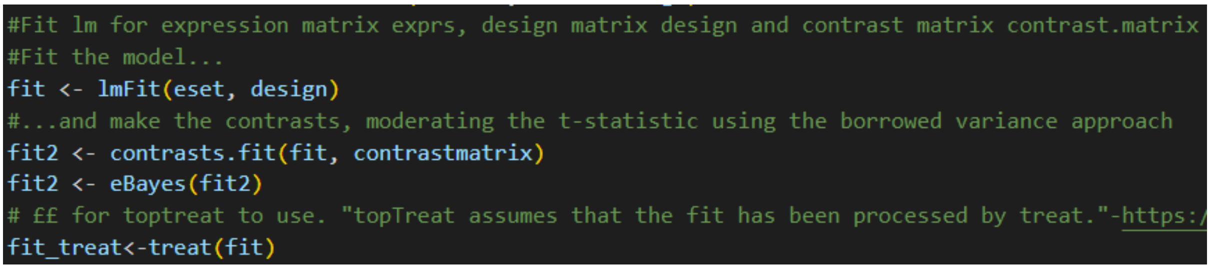
Fold filtering was performed in the first instance manually by calculating fold change means of each level in the experiment. Given that this experiment was mistakenly labelled, there were means for two factors with a total of five levels to be calculated. Theoretically, the concept of a fold change was not applicable to the factor determining the region of intestine as there is no reference between the three. A summary of the calculations is held in 'Group_means.csv'. Further table editing was performed on the top 10 probesets which differed in fold change to match the Probeset IDs to the MOE430v2 chip-specific annotation and present this as Table 5.



```
SI1.mean <- apply(exprsvals10[,c("GSM213045","GSM462207")],1,mean)
SI2.mean <- apply(exprsvals10[,c("GSM213051","GSM462208")],1,mean)
SI3.mean <- apply(exprsvals10[,c("GSM213057","GSM462209")],1,mean)
HCa.mean <- apply(exprsvals10[,c("GSM213045","GSM213051","GSM213057")],1,mean)
LCa.mean <- apply(exprsvals10[,c("GSM462207","GSM462208","GSM462209")],1,mean)
#££ Calculate fold change between low calcium diet (control) and high calcium (treatment)
HiLoCa <-LCa.mean / HCa.mean
```

Figure 13 Code snippet showing fold calculations.

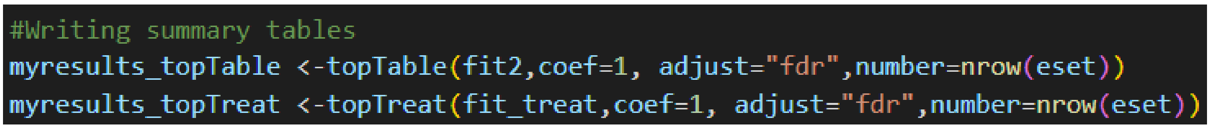
Limma was used to statistically analyse the expression values, creating a list of differentially expressed genes which was subsequently annotated with chip-specific annotations of the MOE430v2 chip. A linear model was generated on the contrasts of a single factor (low vs. high calcium), calculating the moderated t-stat. for each gene across the corresponding levels. eBayes was employed to moderate the variance (by borrowing variance information from genes at similar intensities to estimate variance based on the overall average expression level). This helps to address the lack of technical replicates in the experiment.



```
#Fit lm for expression matrix exprs, design matrix design and contrast matrix contrast.matrix
#Fit the model...
fit <- lmFit(eset, design)
#...and make the contrasts, moderating the t-statistic using the borrowed variance approach
fit2 <- contrasts.fit(fit, contrastmatrix)
fit2 <- eBayes(fit2)
# ££ for toptreat to use. "topTreat assumes that the fit has been processed by treat."-https:/
fit_treat<-treat(fit)
```

Figure 14 Code snippet showing linear model fitting and adjustment with eBayes or treat.

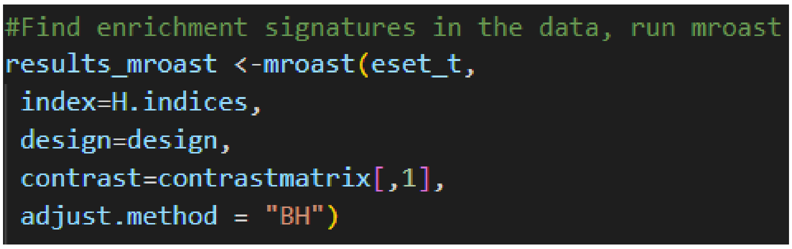
With this complete linear model, topTable was used to identify the top differentially expressed genes. The `adjust` parameter was set to ‘`fdr`’ to generate adjusted p-values using the Benjamini-Hochberg method (BH), addressing the issue of high multiple testing bias in the original p-values. A similar model was generated using `treat` instead of eBayes and was passed to `topTreat` to filter entries by logFC instead of FDR. A Venn Diagram was generated to show the overlap in differentially expressed genes.



```
#Writing summary tables
myresults_topTable <-topTable(fit2,coef=1, adjust="fdr",number=nrow(eset))
myresults_topTreat <-topTreat(fit_treat,coef=1, adjust="fdr",number=nrow(eset))
```

Figure 15 Code snippet showing the main commands used to generate the differential expression tables. topTable results should be filtered by the resulting adjusted p values (FDR), whilst topTreat is meant to be ranked by log Fold Change.

The list of differentially expressed genes was then assessed for functional enrichment. The `mouse4302.db` annotations were employed to describe which biological processes were enriched. `Mroast` was selected to calculate functional enrichment, as this suited the experimental setup which was self-contained. Again, BH was specified to counter the multiple testing issues. Finally, a volcano plot was generated using the Limma `volcanoplot` command. This highlighted the top ten differentially expressed genes identified by topTable in blue.



```
#Find enrichment signatures in the data, run mroast
results_mroast <-mroast(eset_t,
 index=H.indices,
 design=design,
 contrast=contrastmatrix[,1],
 adjust.method = "BH")
```

Figure 16 Code snippet of the main command for determining functional enrichment using mroast.

Note: if repeating this analysis by running the pipeline, please copy the script *FGT_ICA_B217754.R* and *targetfile_B217754.txt* to the present working directory in bash and execute with the command: `Rscript FGT_ICA_B217754.R`

References

(1) de Wit NJW, Bosch-Vermeulen H, Oosterink E, Müller M, van der Meer R. Supplementary dietary calcium stimulates faecal fat and bile acid excretion, but does not protect against obesity and insulin resistance in C57BL/6J mice. Br J Nutr 2011 -04;105(7):1005-1011.

(2) de Wit N, Oosterink E, Bosch-Vermeulen H, Muller M, van de Meer R. Supplementary dietary calcium stimulates faecal fat and bile acid excretion, but does not protect against obesity and insulin resistance in C57BL/6J mice. Wageningen University 2009 Oct 15,.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.