Partially reproducing *The Cerebral Microvasculature in Schizophrenia*
Statistics for Genomics 140.688

Description

Genome wide association studies have found that a pool of around 100 genes have low associations with schizophrenia and overlap bipolar disorder [6]. However, finding expression differences between schizophrenia patients and controls is a complicated task, in part given the population diversity and weak signal. A study using microarrays was performed comparing two cell types as well as control, schizophrenia and bipolar disorder patients [5]. They validated their biological findings using 12 known cell differentiators between endothelia and neuronal cells. But failed to find any genes differentially expressed between control and schizophrenia groups.

Methods

Data was downloaded from NCBI GEO (accession: `GSE12679`) using `GEOquery` [3] for all 32 samples. The raw data was analyzed in `R` [10] using `affy` [4] using the appropriate array information [9, 1]. Exploration of the raw data was performed using `allyQCReport` [7]. After RMA normalization, differential expression was performed using `limma` [12] with a design matrix using the cell type, age, gender, group membership (control, bipolar, schizophrenia), and the post mortem interval (PMI).

Pre-processing

The raw data was explored with `affy` [4]. Some heterogeneity was observed as expected, with the clear outlier being sample 2 as a finger print can be seen as shown in figure 1.

Intensity values were explored with `affyQCReport` [7], revealing distributional differences as observed in figure 2. The QC report revealed array differences in positive and negative border elements (see `qcReport.pdf` page 3).

Further exploration of array 2 with an MA plot versus the pseudo-median reference reveals a large difference as shown in 3 as the loess curve spans negative M values. Thus, it is crucial to normalize the data. Further exploration (data not shown, but code included) revealed that normalizing the data improved the loess curve in the MA plot as it was symmetric around 0.

RMA was used to perform quantile normalization and background correction.

Differential expression

Differential expression was performed with `limma` [12] using all 32 samples and a model as follows:

$$Y = \beta_0 + \beta_1 I\,(\text{Cell} = \text{neuronal}) + \beta_2 \text{Age} + \beta_3 I\,(\text{Gender} = \text{male}) + \beta_4 I\,(\text{Group} = \text{bipolar}) + \beta_5 I\,(\text{Group} = \text{schizophrenia}) + \beta_6 \text{PMI}$$

Empirical Bayes was then used to compute moderated statistics and p-values were adjusted for FDR control using the Benjamini-Hochberg method. Top genes were retrieved for coefficients β_1 (cell type), β_3 (gender), β_4 and β_6 (group).

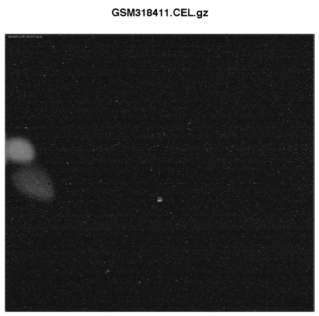

Figure 1: Reconstructed image for sample 2. A fingerprint is clearly noticeable.

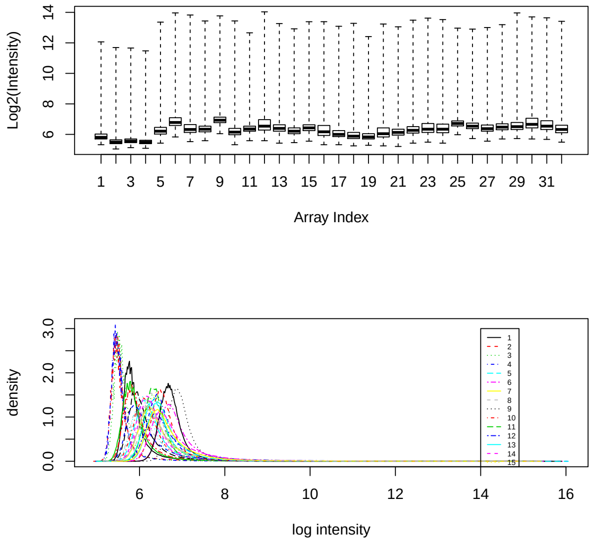

Figure 2: (Top) Log2 intensity distributions. There are clear differences, in particular between the first 4 arrays and the rest. (Bottom) Log intensity densities. Arrays group together into approximately 4 groups (3 clear, one diffused).

Biological conclusions

Table 1 shows the that 6 out of the 12 genes known to differentiate cell types were included in the 525 DE genes between cell types with FDR < 0.05. The observed fold changes match in 5 out of 6 cases the expected direction.

Similarly, table 1 shows the only 7 probes to be differentially expressed between genders with FDR < 0.10. These probes correspond to genes XIST, DDX3Y, and LINC01120 with the first two having known gender-related functions [11]. Thus both sub-analyses biologically validate the results.

When comparing the control vs bipolar groups, 0 genes were found to be differentially expressed at FDR < 0.10. A single gene, TEX10, was found to be DE for control vs schizophrenia in contrast to the previous analysis [5]. While TEX10 stands for *testis expressed 10*, more recent work [2] has found that when depleted it leads to p53-dependent G1 arrest and in general is involved in

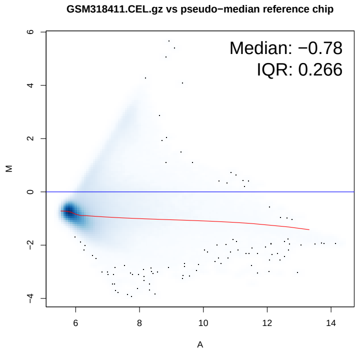

Figure 3: MA plot for raw data from array 2 versus the pseudo-median reference chip. Lowess line shown in red is considerably away from M = 0.

	probe	logFC	AveExpr	t	adj.P.Val	status	SYMBOL	question
1	216356_x_at	0.94	4.90	5.12	8.23E-03	endothelial	BAIAP3	cell type
2	204584_at	1.83	4.43	7.09	5.78E-04	endothelial	L1CAM	cell type
3	225665_at	-1.53	2.93	-5.27	7.26E-03	neuronal	ZAK	cell type
4	203571_s_at	-0.52	2.43	-4.77	1.50E-02	neuronal	ADIRF	cell type
5	200092_s_at	-1.24	8.28	-4.69	1.64E-02	neuronal	RPL37	cell type
6	38710_at	1.32	7.23	5.11	8.23E-03	neuronal	OTUB1	cell type
7	224588_at	-4.76	4.49	-8.70	1.34E-04		XIST	gender
8	221728_x_at	-3.35	3.49	-7.50	1.19E-03		XIST	gender
9	214218_s_at	-2.79	2.88	-6.90	3.62E-03		XIST	gender
10	224590_at	-2.83	2.93	-6.71	4.37E-03		XIST	gender
11	227671_at	-1.31	2.91	-6.44	7.16E-03		XIST	gender
12	240546_at	-0.40	2.24	-5.65	4.72E-02		LINC01120	gender
13	205000_at	2.16	3.61	5.34	9.26E-02		DDX3Y	gender
14	218104_at	-1.30	2.74	-6.14	7.72E-02		TEX10	schizophrenia

Table 1: Top genes for cell type, gender and schizophrenia vs control coefficients. Status column shows which cell type is expected to be up-regulated based on previous literature, with only OTUB1 mismatching the expected logFC direction. Gender results identify known gender-related genes. Schizophrenia vs control only identified one gene. P-values are FDR controlled by BH.

the regulation of the ribosome biogenesis. Further validation is needed to evaluate the relation of TEX10 in schizophrenia, or to discard it.

In particular, a recent study [8] found 1331 genes as differentially expressed between control and schizophrenia samples in neuron cells (9 samples per group). Their DE gene list is currently not available (it is a April 10 2014 pre-print) so we cannot check if TEX10 is present in their list. Nevertheless, their results could in theory be reproduced using NCBI GEO `GSE37981` in future work.

Reproducibility

The code, main results, and main plots are available at `https://github.com/lcolladotor/hw2_140.688`.

References

[1] M. Carlson. *hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2)*. R package version 2.14.0.

[2] C. D. Castle, E. K. Cassimere, and C. Denicourt. “LAS1L interacts with the mammalian Rix1 complex to regulate ribosome biogenesis”. eng. In: *Molecular biology of the cell* 23.4 (Feb. 2012). PMID: 22190735 PMCID: PMC3279398, pp. 716–728. ISSN: 1939-4586. DOI: `10.1091/mbc.E11-06-0530`.

[3] S. Davis and P. Meltzer. “GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor”. In: *Bioinformatics* 14 (2007), pp. 1846–1847.

[4] L. Gautier et al. “affy—analysis of Affymetrix GeneChip data at the probe level”. In: *Bioinformatics* 20.3 (2004), pp. 307–315. ISSN: 1367-4803. DOI: `http://dx.doi.org/10.1093/bioinformatics/btg405`.

[5] L. W. Harris et al. “The cerebral microvasculature in schizophrenia: a laser capture microdissection study”. eng. In: *PloS one* 3.12 (2008). PMID: 19088852 PMCID: PMC2597747, e3964. ISSN: 1932-6203. DOI: `10.1371/journal.pone.0003964`.

[6] International Schizophrenia Consortium et al. “Common polygenic variation contributes to risk of schizophrenia and bipolar disorder”. eng. In: *Nature* 460.7256 (Aug. 2009). PMID: 19571811 PMCID: PMC3912837, pp. 748–752. ISSN: 1476-4687. DOI: `10.1038/nature08185`.

[7] C. Parman, C. Halling, and R. Gentleman. *affyQCReport: QC Report Generation for affyBatch objects*. R package version 1.42.0.

[8] C. Y. Pietersen et al. “Molecular Profiles of Pyramidal Neurons in the Superior Temporal Cortex in Schizophrenia”. ENG. In: *Journal of neurogenetics* (Apr. 2014). PMID: 24702465. ISSN: 1563-5260. DOI: `10.3109/01677063.2014.882918`.

[9] T. B. Project. *hgu133plus2cdf: hgu133plus2cdf*. R package version 2.14.0.

[10] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2014. URL: `http://www.R-project.org/`.

[11] M Rebhan et al. “GeneCards: integrating information about genes, proteins and diseases”. eng. In: *Trends in genetics: TIG* 13.4 (Apr. 1997). PMID: 9097728, p. 163. ISSN: 0168-9525.

[12] G. K. Smyth. “Limma: linear models for microarray data”. In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Ed. by R. Gentleman et al. New York: Springer, 2005, pp. 397–420.

Some useful links:

- `http://www.genecards.org/cgi-bin/carddisp.pl?gene=XIST`
- `http://www.genecards.org/cgi-bin/carddisp.pl?gene=DDX3Y`
- `http://www.genecards.org/cgi-bin/carddisp.pl?gene=TEX10`

A Analysis code

```r
##### Analysis code ######

## Attempted data sets:
# GSE11800
# GSE30272
# GSE25673

## Get data from GEO
library("GEOquery")
geo <- getGEO("GSE12679")
geo.files <- getGEOSuppFiles("GSE12679")
save(geo, file="geo.Rdata")

## Manually un-tar the RAW data

## Load into R using the download pheno data
files <- list.files("GSE12679/GSE12679_RAW", full.names = TRUE)
library("affy")
Data <- ReadAffy(filenames = files, phenoData = phenoData(geo[[1]]))
save(Data, file="Data.Rdata")

## Briefly explore the data summary
library("hgu133plus2cdf")
Data

## Create a QC report
library("affyQCReport")
QCReport(Data, file = "qcReport.pdf")

pdf("signalDist.pdf")
signalDist(Data)
dev.off()

## Visually explore the data using functions from the affy package
image(Data)
## Results
# ok: 1, 3:11, 13:15, 17, 19, 21:28, 31:32
# fingerprint: 2
# maybe a fingerprint: 18, 29
# slightly brighter: 12, 16, 20
png("image-2.png")
image(Data[, 2])
dev.off()

pdf("maPlots-%03d.pdf", onefile=FALSE)
MAplot(Data, plot.method = "smoothScatter")
dev.off()
# number 2 is one of the most different ones vs pseudo median chip

## Not useful since there are too many reps
if(FALSE) {
    pdf("maPlot-pairs.pdf", width=14, height=14)
    MAplot(Data, pairs = TRUE, plot.method = "smoothScatter")
    dev.off()
}


## PM intensities
raw <- pm(Data)
log.raw <- log2(raw)

## Explore QN
qn <- preprocessCore::normalize.quantiles(log.raw)

## MA plot for sample 1 vs 2, then again after QN

pdf("maPlots-pair1-%03d.pdf", onefile=FALSE)
lapply(list(log.raw, qn), function(dat) {
    avg  <- (dat[, 1] + dat[, 2])/2
    diff <-  dat[, 1] - dat[, 2]

    lowess.curve <- lowess(x = avg, y = diff, f = 0.05,)
    smoothScatter(avg, diff, ylim = c(-6, 6), xlab="Average expression (log2)", ylab="Expression ratio (log2)", main="Sample 1 vs 2")
    abline(h=0,col="deepskyblue3", lwd=4)
    lines(lowess.curve, col = "deeppink3", lwd = 4)
})
dev.off()

## Perform RMA
eset <- rma(Data)
e <- exprs(eset)
save(e, eset, file="rma.Rdata")

## Obtain characteristics of interest
pd <- pData(Data)
pd.char <- as.character(pd$characteristics_ch1)
chars <- data.frame(
        "CellType" = factor(gsub("(Cell type: )|(,.*)", "", pd.char)),
        "Age" = as.numeric(gsub("(.*age:)|(, gender.*)", "", pd.char)),
        "Gender" = factor(gsub("(.*gender:)|(,.*)", "", pd.char)),
        "Group" = relevel(factor(gsub("(.*group: )|(,.*)", "", pd.char)), "control"),
        "PMI" = as.numeric(gsub("(.*PMI:)|(h)", "", pd.char))
    )


## Differential expression
design.matrix <- with(chars, model.matrix(~CellType + Age + Gender + Group + PMI))
library("limma")
fit <- lmFit(object = e, design = design.matrix)
eFit <- eBayes(fit)
save(fit, eFit, design.matrix, file="fits.Rdata")

## Explore resulting p-values
pdf("pvalues-dist-%03d.pdf", onefile=FALSE)
for(i in 2:7) {
    hist(eFit$p.value[, i], main=paste("P-value distribution for", colnames(fit$coefficients)[i]), xlab="p-values")
}
dev.off()

pdf("volcanoPlots-%03d.pdf", onefile=FALSE)
for(i in 2:7) {
    volcanoplot(eFit, coef=i, main=paste("Volcano plot for", colnames(fit$coefficients)[i]), highlight = 5)
}
dev.off()


## Get top genes
library("hgu133plus2.db")
columns(hgu133plus2.db)

table <- topTable(eFit, coef = 2, adjust.method="BH", p.value = 0.05, number=1000)

## Explore top genes
head(table)

probeNames <- rownames(table)
info <- select(hgu133plus2.db, probeNames, c("SYMBOL", "ALIAS"), "PROBEID")

previous <- data.frame(alias = c("FN1", "SPARC", "ITGAV", "CDH5", "EPAS1", "GJA4", "L1CAM", "SNAP25", "SYP", "SCN3B", "SLC17A7", "THY1"),
    status = rep(c("endothelial", "neuronal"), each=6))
compLit <- cbind(table[match(previous$alias, info$ALIAS), ], "status" = previous$status)
compLit <- compLit[!is.na(compLit$t),]
## Not completely matching published results
# Using only the control samples and a model matrix testing cell type leads to only two results matching.
# The difference must lie in the normalizations used.


tableB <- topTable(eFit, coef = 5, adjust.method="BH", p.value = 0.10, number=1000)
dim(tableB)
# 0

tableG <- topTable(eFit, coef = 4, adjust.method="BH", p.value = 0.10, number=1000)
dim(tableG)
tableG$SYMBOL <- select(hgu133plus2.db, rownames(tableG), c("SYMBOL"), "PROBEID")$SYMBOL
tableG
## 3 genes, 2 of which make sense. The other one needs more bio info.
# http://www.genecards.org/cgi-bin/carddisp.pl?gene=XIST
# http://www.genecards.org/cgi-bin/carddisp.pl?gene=DDX3Y
# http://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=389043

tableS <- topTable(eFit, coef = 6, adjust.method="BH", p.value = 0.10, number=1000)
dim(tableS)
tableS$SYMBOL <- select(hgu133plus2.db, rownames(tableS), c("SYMBOL"), "PROBEID")$SYMBOL
tableS
## One hit. But it's unclear whether it's bio related.
# http://www.genecards.org/cgi-bin/carddisp.pl?gene=TEX10

save(info, probeNames, previous, compLit, table, tableB, tableG, tableS, file="table-results.Rdata")
```

B Report code

```r
## Libs used in analysis, for session info
library("GEOquery")
library("hgu133plus2cdf")
library("affy")
library("affyQCReport")
library("limma")
library("hgu133plus2.db")

## Load data
load("table-results.Rdata")

## For printing
library("xtable")
```

```r
## Combine results
tableS$status <- NA
tableG$status <- NA
compLit$SYMBOL <- select(hgu133plus2.db, rownames(compLit), c("SYMBOL"), "PROBEID")$SYMBOL

res <- rbind(compLit, tableG, tableS)
res$question <- rep(c("cell type", "gender", "schizophrenia"), c(nrow(compLit), nrow(tableG), nrow(tableS)))

## Print out
res2 <- cbind("probe" = rownames(res), res[,-c(4, 6)])
rownames(res2) <- seq_len(nrow(res2))
print.xtable(xtable(res2, label = "tab1", display = c("d", "s", "f", "f", "f", "E", "s", "s", "s"),
    caption = "Top genes for cell type, gender and schizophrenia vs control coefficients. Status column shows which cell type is expected to be up-regulated based on prev
    size="tiny", table.placement="ht!")
```

C Session information

- R version 3.1.0 (2014-04-10), x86_64-apple-darwin10.8.0
- Locale: en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
- Base packages: base, datasets, graphics, grDevices, methods, parallel, stats, utils
- Other packages: affy 1.42.0, affyQCReport 1.42.0, AnnotationDbi 1.26.0, Biobase 2.24.0, BiocGenerics 0.10.0, DBI 0.2-7, GenomeInfoDb 1.0.2, GEOquery 2.30.0, hgu133plus2.db 2.14.0, hgu133plus2cdf 2.14.0, knitr 1.5.27, lattice 0.20-29, limma 3.20.1, org.Hs.eg.db 2.14.0, RSQLite 0.11.4, xtable 1.7-3
- Loaded via a namespace (and not attached): affyio 1.32.0, affyPLM 1.40.0, annotate 1.42.0, BiocInstaller 1.14.1, Biostrings 2.32.0, evaluate 0.5.3, formatR 0.10, gcrma 2.36.0, genefilter 1.46.0, grid 3.1.0, highr 0.3, IRanges 1.22.3, preprocessCore 1.26.0, RColorBrewer 1.0-5, RCurl 1.95-4.1, simpleaffy 2.40.0, splines 3.1.0, stats4 3.1.0, stringr 0.6.2, survival 2.37-7, tools 3.1.0, XML 3.98-1.1, XVector 0.4.0, zlibbioc 1.10.0