

Applying Statistical Correction for Brain Tissue RNA Degradation to Gene Expression Differences in Schizophrenia

Amy Peterson¹, Andrew E Jaffe^{*2}, and Leonardo Collado-Torres^{†2}

¹Johns Hopkins School of Public Health MPH

²Lieber Institute for Brain Development

*andrew.jaffe@libd.org †leo.collado@libd.org

18 May 2018

Abstract

MLR for 140.624

Contents

1	Introduction	2
2	Load the data	2
3	Full qSVA analysis	3
4	Multiple linear regressions for one gene	5
4.1	ELP6: case where it's not significant after qSVA	5
4.2	AHCYL1: case where it's significant after qSVA	8
4.3	Adjusted covariates interpretation	11
5	Conclusions	11
	Session info	11

1 Introduction

This document highlights some of the multiple linear regressions that were performed for the analysis of BrainSeq data using the qSVA method to adjust for RNA-quality degradation, which confounds the signal of differential gene expression between schizophrenia cases and non-psychiatric controls in the DLPFC and HIPPO brain regions.

2 Load the data

First, we need to load the BrainSeq data. This includes loading the `qsvBonf`, `qSVs`, `mod`, and `modQsva` objects which were saved from a previous script but commented below. The top qSVs were obtained using Bonferroni correction on the PC's obtained from PCA on the degradation matrix. The top k PC's were selected using the `sva::num.SV` function (k = 16 in this case). The model saved as `mod` included diagnosis, RNA-quality, and demographic covariates. The `modQsva` object contains `mod` and the qSVs.

```
## Load R packages required for this analysis
library('devtools')
library('SummarizedExperiment')
library('limma')
library('edgeR')
library('broom')

#load qsvBonf, qSVs, mod, modQsva object from brainseq data
## get qSVs for top bonferroni
# qsvBonf = prcomp(t(log2(assays(cov_rse)$counts+1)))
## qsva
# k = num.SV(log2(assays(cov_rse)$counts+1), mod)
# qSVs = qsvBonf$x[,1:k]
# getPcaVars(qsvBonf)[1:k]
# # [1] 57.800 15.700  3.360  2.700  2.360  1.160  1.040  0.848  0.750  0.661
# # [11] 0.587  0.511  0.461  0.419  0.356  0.312  0.291  0.284  0.271  0.259
## modQsva = cbind(mod, qSVs)
## mod = model.matrix(~Dx + Age + Sex + mitoRate + Region + rRNA_rate +
## totalAssignedGene + RIN + snpPC1 + snpPC2 +snpPC3 + snpPC4 + snpPC5,
## data = colData(rse_gene))
#qsvBonf, qSVs, mod, modQsva were then saved in a previous script to rdas/brainseq_phase2_qsvs.Rdata
load("/dcl01/lieber/ajaffe/lab/brainseq_phase2/expr_cutoff/rse_gene.Rdata", verbose = TRUE)
## Loading objects:
##   rse_gene
load('/dcl01/ajaffe/data/lab/qsva_brain/brainseq_phase2_qsv/rdas/brainseq_phase2_qsvs_age17_noHGold_HIPO.Rda')
## Loading objects:
##   qsvBonf
##   qSVs
##   mod
##   modQsva
##   keepIndex

## Drop prenatal
```

```
rse_gene <- rse_gene[, keepIndex]

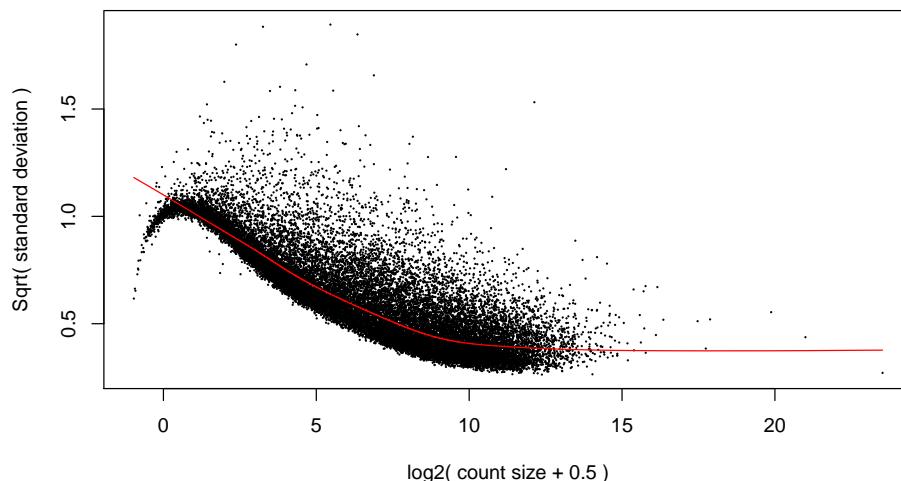
## Keep the correct samples
keepIndex = which(rse_gene$Age>17 &
                  rse_gene$Region == "HIPPO")
rse_gene <- rse_gene[, keepIndex]
mod <- mod[keepIndex, ]
modQsva <- modQsva[keepIndex, ]
```

3 Full qSVA analysis

As seen the overall workflow on slide 7, and slides 12 through 14 of the original presentation, below is code for the qSVA analysis that was performed using the three different models. Model 1 was a naive model that included diagnosis only. Model 2 was a model including diagnosis, and adding in RNA-quality and demographic covariates. Model 3 is a model including the qSVs, in addition to diagnosis, RNA-quality, and demographic covariates.

```
##### GENE #####
dge = DGEList(counts = assays(rse_gene)$counts,
              genes = rowData(rse_gene))
## Warning in as.data.frame(x, row.names = NULL, optional = optional, ...):
## Arguments in '....' ignored
#calculate library-size adjustment
dge = calcNormFactors(dge)
vGene = voom(dge,modQsva, plot=TRUE)
```

voom: Mean-variance trend

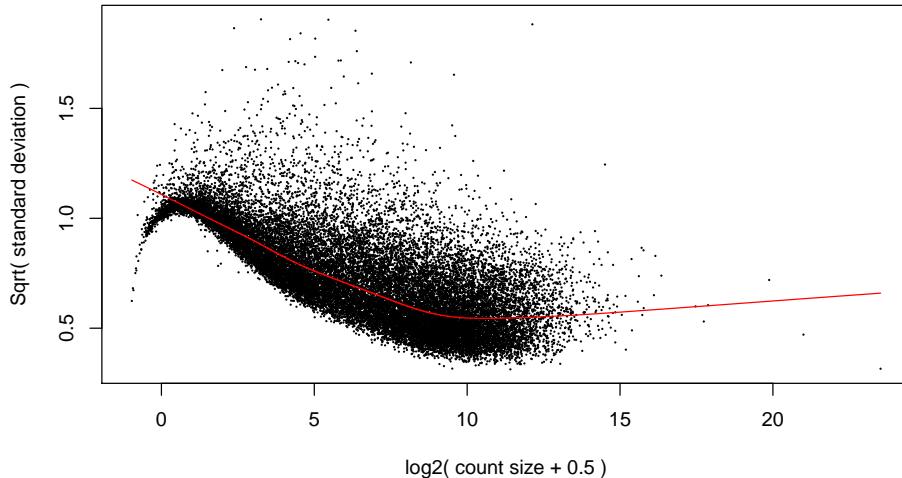


```
fitGene = lmFit(vGene)
eBGene = eBayes(fitGene)
sigGene = topTable(eBGene, coef=2,
                   p.value = 1, number=nrow(rse_gene))
outGene = sigGene[rownames(rse_gene), ]
```

Applying Statistical Correction for Brain Tissue RNA Degradation to Gene Expression Differences in Schizophrenia

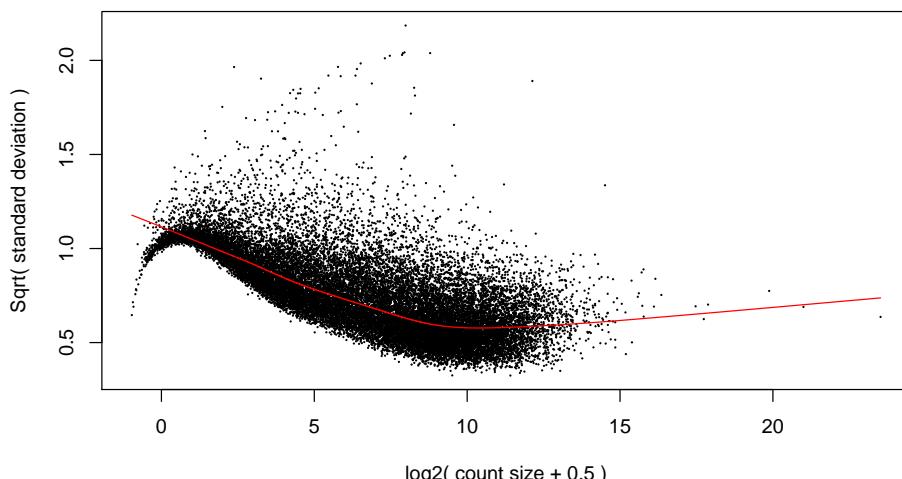
```
## no qSVA  
vGene0 = voom(dge,mod, plot=TRUE)
```

voom: Mean-variance trend



```
fitGene0 = lmFit(vGene0)  
eBGene0 = eBayes(fitGene0)  
sigGene0 = topTable(eBGene0,coef=2,  
    p.value = 1,number=nrow(rse_gene))  
outGene0 = sigGene0[rownames(rse_gene),]  
  
## no adjustment vars  
vGeneNoAdj = voom(dge, with(colData(rse_gene), model.matrix( ~ Dx)), plot=TRUE)
```

voom: Mean-variance trend



```
fitGeneNoAdj = lmFit(vGeneNoAdj)  
eBGeneNoAdj = eBayes(fitGeneNoAdj)  
sigGeneNoAdj = topTable(eBGeneNoAdj,coef=2,
```

```
p.value = 1, number=nrow(rse_gene))  
outGeneNoAdj = sigGeneNoAdj[rownames(rse_gene), ]
```

4 Multiple linear regressions for one gene

4.1 ELP6: case where it's not significant after qSVA

In the example gene *ELP6* seen below, the gene shown is initially seen as significant in the first two models in the HIPPO brain region. After adding in the qSVs, the gene is not seen as significant anymore. This is an example of the importance of including all of the covariates in the final model to reduce the number of false positives, or genes that are incorrectly identified as differentially expressed when in fact the perceived differential expression is a result of inadequate control of the confounding effect of RNA quality.

```
## Find a gene that is not DE after qSVA but is in previous models  
i <- which(outGene$adj.P.Val > 0.05 & outGene0$adj.P.Val < 0.05)[14]  
outGene[i, ]  
## Length gencodeID ensemblID gene_type  
## ENSG00000163832.15 3313 ENSG00000163832.15 ENSG00000163832 protein_coding  
## Symbol EntrezID Class meanExprs NumTx gencodeTx  
## ENSG00000163832.15 ELP6 54859 InGen 4.082769 15 ENST0000....  
## passExprsCut logFC AveExpr t P.Value  
## ENSG00000163832.15 TRUE -0.02909726 3.142082 -1.73303 0.08408999  
## adj.P.Val B  
## ENSG00000163832.15 0.5296745 -4.859558  
  
## Explore that gene in more detail  
explore_gene <- function(i, expr, model = 1) {  
    y <- expr$E[i, ]  
    df <- as.data.frame(cbind(y, modQsva))  
    if(model == 1) {  
        f <- lm(y ~ DxSchizo, data = df)  
    } else if (model == 2) {  
        f <- lm(y ~ DxSchizo + Age + SexM + mitoRate + rRNA_rate +  
                 totalAssignedGene + RIN +.snpPC1 + .snpPC2 +  
                 .snpPC3 + .snpPC4 + .snpPC5, data = df)  
    } else if (model == 3) {  
        f <- lm(y ~ DxSchizo + Age + SexM + mitoRate + rRNA_rate +  
                 totalAssignedGene + RIN +.snpPC1 + .snpPC2 +  
                 .snpPC3 + .snpPC4 + .snpPC5 + PC1 + PC2 + PC3 +  
                 PC4 + PC5 + PC6 + PC7 + PC8 + PC9 + PC10 + PC11 +  
                 PC12 + PC13 + PC14 + PC15 + PC16, data = df)  
    }  
    summary(f)  
}  
  
## Model 1: Dx only  
explore_gene(i, vGene, model = 1)
```

Applying Statistical Correction for Brain Tissue RNA Degradation to Gene Expression Differences in Schizophrenia

```
##  
## Call:  
## lm(formula = y ~ DxSchizo, data = df)  
##  
## Residuals:  
##      Min      1Q  Median      3Q     Max  
## -0.73648 -0.14387  0.00291  0.14168  0.57722  
##  
## Coefficients:  
##                 Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  3.20595   0.01549 206.966 < 2e-16 ***  
## DxSchizo     -0.15990   0.02451  -6.524 2.57e-10 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.2191 on 331 degrees of freedom  
## Multiple R-squared:  0.1139, Adjusted R-squared:  0.1113  
## F-statistic: 42.56 on 1 and 331 DF,  p-value: 2.57e-10  
  
## Model 2: with all quality and demographic covariates  
explore_gene(i, vGene, model = 2)  
##  
## Call:  
## lm(formula = y ~ DxSchizo + Age + SexM + mitoRate + rRNA_rate +  
##       totalAssignedGene + RIN +.snpPC1 + .snpPC2 + .snpPC3 + .snpPC4 +  
##       .snpPC5, data = df)  
##  
## Residuals:  
##      Min      1Q  Median      3Q     Max  
## -0.58882 -0.11890  0.01071  0.12565  0.46696  
##  
## Coefficients:  
##                 Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  1.551e+00  1.698e-01  9.130 < 2e-16 ***  
## DxSchizo    -9.670e-02  2.279e-02 -4.243 2.89e-05 ***  
## Age        -8.480e-05  7.511e-04 -0.113 0.910184  
## SexM        3.670e-03  2.212e-02  0.166 0.868345  
## mitoRate    -5.459e-01  3.298e-01 -1.655 0.098893 .  
## rRNA_rate   1.059e+03  4.589e+02  2.308 0.021631 *  
## totalAssignedGene 2.298e+00  4.476e-01  5.135 4.92e-07 ***  
## RIN         7.236e-02  1.335e-02  5.419 1.18e-07 ***  
## .snpPC1     -1.737e-01  2.182e-01 -0.796 0.426660  
## .snpPC2     -2.225e+00  2.154e+00 -1.033 0.302449  
## .snpPC3     3.970e+00  1.135e+00  3.497 0.000537 ***  
## .snpPC4     2.449e+00  6.516e+00  0.376 0.707223  
## .snpPC5     -2.612e+00  3.442e+00 -0.759 0.448426  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.1806 on 320 degrees of freedom  
## Multiple R-squared:  0.4177, Adjusted R-squared:  0.3958
```

Applying Statistical Correction for Brain Tissue RNA Degradation to Gene Expression Differences in Schizophrenia

```
## F-statistic: 19.13 on 12 and 320 DF, p-value: < 2.2e-16

## Model 3: with all of the above and the qSVs (1 to 16)
explore_gene(i, vGene, model = 3)
##
## Call:
## lm(formula = y ~ DxSchizo + Age + SexM + mitoRate + rRNA_rate +
##     totalAssignedGene + RIN + snpPC1 + snpPC2 + snpPC3 + snpPC4 +
##     snpPC5 + PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 +
##     PC9 + PC10 + PC11 + PC12 + PC13 + PC14 + PC15 + PC16, data = df)
##
## Residuals:
##      Min      1Q   Median      3Q      Max 
## -0.41808 -0.07620  0.00074  0.08385  0.29247 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 3.422e+00 2.775e-01 12.332 < 2e-16 ***
## DxSchizo    -3.053e-02 1.650e-02 -1.850 0.06526 .  
## Age        -3.560e-04 5.667e-04 -0.628 0.53035  
## SexM       -1.723e-02 1.554e-02 -1.109 0.26843  
## mitoRate    1.156e+00 4.183e-01  2.765 0.00605 ** 
## rRNA_rate   1.333e+02 3.797e+02  0.351 0.72567  
## totalAssignedGene -1.137e+00 6.305e-01 -1.803 0.07244 .  
## RIN         8.450e-03 1.266e-02  0.667 0.50511  
## snpPC1     8.715e-02 1.644e-01  0.530 0.59650  
## snpPC2    -5.921e-01 1.524e+00 -0.389 0.69782  
## snpPC3     1.875e+00 9.368e-01  2.001 0.04625 *  
## snpPC4     7.557e+00 4.561e+00  1.657 0.09854 .  
## snpPC5     4.608e-01 2.399e+00  0.192 0.84778  
## PC1        1.138e-02 1.398e-03  8.137 1.05e-14 ***
## PC2       -6.114e-03 1.463e-03 -4.179 3.84e-05 *** 
## PC3       -2.011e-02 1.812e-03 -11.095 < 2e-16 ***
## PC4       -2.705e-03 2.389e-03 -1.132 0.25847  
## PC5       -7.619e-03 3.355e-03 -2.271 0.02387 *  
## PC6        9.541e-04 3.486e-03  0.274 0.78448  
## PC7       -5.148e-03 3.262e-03 -1.578 0.11558  
## PC8       -1.520e-03 4.136e-03 -0.368 0.71348  
## PC9        6.011e-03 3.815e-03  1.576 0.11615  
## PC10      1.633e-02 5.378e-03  3.036 0.00261 ** 
## PC11      -6.256e-03 5.390e-03 -1.161 0.24662  
## PC12      4.907e-03 5.599e-03  0.876 0.38156  
## PC13      -3.009e-02 5.213e-03 -5.772 1.94e-08 *** 
## PC14      -1.672e-02 5.473e-03 -3.056 0.00244 ** 
## PC15      -2.117e-02 5.344e-03 -3.961 9.29e-05 *** 
## PC16      -1.052e-02 6.123e-03 -1.718 0.08673 .  
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1234 on 304 degrees of freedom
## Multiple R-squared:  0.7418, Adjusted R-squared:  0.718
```

Applying Statistical Correction for Brain Tissue RNA Degradation to Gene Expression Differences in Schizophrenia

```
## F-statistic: 31.19 on 28 and 304 DF, p-value: < 2.2e-16

## Export for editing with Excel
write.csv(tidy(explore_gene(i, vGene, model = 3)), file = 'gene1.csv')
```

The interpretation of the p-value in the model for each gene determines whether or not the gene is considered differentially expressed between schizophrenic cases and non-psychiatric controls. For *ELP6*, using the diagnosis only model, the gene is considered differentially expressed ($p = 2.57 \times 10^{-10}$). In the second model, once the RNA-quality and demographic covariates are added in, *ELP6* is still considered differentially expression ($p = 2.89 \times 10^{-5}$). In the final model, with qSVs added in addition to diagnosis, RNA-quality, and demographic covariates, *ELP6* is no longer considered differentially expressed ($p = 0.06526$) between cases and controls with a log2 fold change of -3.053×10^{-2} (higher expression in schizophrenia cases).

4.2 AHCYL1: case where it's significant after qSVA

In the gene *AHCYL1* below, the gene remains significant throughout the 3 different models between schizophrenia cases and non-psychiatric controls for the HIPPO brain region. This demonstrates the ability to correctly identify genes that are differentially expressed between schizophrenic cases and non-psychiatric controls. The ability to identify a gene as differentially expressed throughout all 3 models increases the confidence that the genes that are identified as differentially expressed using the final model are truly differentially expressed.

```
## Example gene that is DE after qSVA
i <- which(outGene$adj.P.Val < 0.05)[3]
outGene[i, ]
##             Length      gencodeID      ensemblID      gene_type
## ENSG00000168710.17    5113 ENSG00000168710.17 ENSG00000168710 protein_coding
##             Symbol EntrezID Class meanExprs NumTx      gencodeTx
## ENSG00000168710.17 AHCYL1     10768 InGen   45.23379      6 ENST0000.....
##                  passExprsCut      logFC AveExpr          t      P.Value
## ENSG00000168710.17           TRUE 0.1196015 7.390143 4.563977 7.257743e-06
##                  adj.P.Val      B
## ENSG00000168710.17 0.01626526 3.257424

## Model 1: Dx only
explore_gene(i, vGene, model = 1)
##
## Call:
## lm(formula = y ~ DxSchizo, data = df)
##
## Residuals:
##       Min     1Q     Median      3Q     Max 
## -1.21091 -0.27162  0.00802  0.25068  1.61573 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 7.34063   0.02864 256.279 < 2e-16 ***
## DxSchizo    0.12398   0.04532   2.736  0.00656 ** 
## ---
```

Applying Statistical Correction for Brain Tissue RNA Degradation to Gene Expression Differences in Schizophrenia

```
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4051 on 331 degrees of freedom
## Multiple R-squared:  0.02211,   Adjusted R-squared:  0.01915
## F-statistic: 7.483 on 1 and 331 DF,  p-value: 0.006564

## Model 2: with all quality and demographic covariates
explore_gene(i, vGene, model = 2)
##
## Call:
## lm(formula = y ~ DxSchizo + Age + SexM + mitoRate + rRNA_rate +
##     totalAssignedGene + RIN + snpPC1 + snpPC2 + snpPC3 + snpPC4 +
##     snpPC5, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.36669 -0.26087  0.00539  0.24864  1.32903
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.898e+00  3.659e-01 16.121 < 2e-16 ***
## DxSchizo              1.401e-01  4.909e-02  2.855  0.00459 **
## Age                   4.142e-03  1.618e-03  2.560  0.01093 *
## SexM                 -4.491e-02  4.765e-02 -0.942  0.34671
## mitoRate              1.735e-01  7.105e-01  0.244  0.80722
## rRNA_rate              1.543e+03  9.885e+02  1.561  0.11956
## totalAssignedGene    2.185e+00  9.642e-01  2.266  0.02410 *
## RIN                  4.557e-03  2.876e-02  0.158  0.87420
## snpPC1                -7.153e-02  4.701e-01 -0.152  0.87914
## snpPC2                -1.041e+00  4.641e+00 -0.224  0.82271
## snpPC3                -3.998e+00  2.445e+00 -1.635  0.10305
## snpPC4                -1.666e+01  1.404e+01 -1.187  0.23607
## snpPC5                1.567e+00  7.413e+00  0.211  0.83275
## ...
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3891 on 320 degrees of freedom
## Multiple R-squared:  0.1279, Adjusted R-squared:  0.09515
## F-statistic: 3.909 on 12 and 320 DF,  p-value: 1.279e-05

## Model 3: with all of the above and the qSVs (1 to 16)
explore_gene(i, vGene, model = 3)
##
## Call:
## lm(formula = y ~ DxSchizo + Age + SexM + mitoRate + rRNA_rate +
##     totalAssignedGene + RIN + snpPC1 + snpPC2 + snpPC3 + snpPC4 +
##     snpPC5 + PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 +
##     PC9 + PC10 + PC11 + PC12 + PC13 + PC14 + PC15 + PC16, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.36669 -0.26087  0.00539  0.24864  1.32903
```

Applying Statistical Correction for Brain Tissue RNA Degradation to Gene Expression Differences in Schizophrenia

```
## -0.90810 -0.11455  0.00455  0.11754  0.66928
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)            7.726e+00  4.408e-01 17.528 < 2e-16 ***
## DxSchizo              1.188e-01  2.621e-02  4.533 8.38e-06 ***
## Age                   2.674e-03  9.001e-04  2.971 0.003207 **
## SexM                  3.160e-02  2.468e-02  1.280 0.201419
## mitoRate              2.874e-01  6.643e-01  0.433 0.665551
## rRNA_rate             -7.429e+02  6.030e+02 -1.232 0.218925
## totalAssignedGene    -3.527e-01  1.001e+00 -0.352 0.724922
## RIN                  -5.098e-02  2.011e-02 -2.534 0.011764 *
## snpPC1               -4.712e-01  2.612e-01 -1.804 0.072202 .
## snpPC2               1.508e+00  2.420e+00  0.623 0.533693
## snpPC3               1.165e+00  1.488e+00  0.783 0.434364
## snpPC4               5.053e+00  7.243e+00  0.698 0.486007
## snpPC5               6.464e+00  3.810e+00  1.697 0.090799 .
## PC1                  3.711e-03  2.220e-03  1.672 0.095649 .
## PC2                  -3.707e-02  2.324e-03 -15.955 < 2e-16 ***
## PC3                  1.161e-02  2.878e-03  4.034 6.95e-05 ***
## PC4                  1.613e-02  3.794e-03  4.251 2.84e-05 ***
## PC5                  3.307e-02  5.329e-03  6.206 1.78e-09 ***
## PC6                  1.885e-02  5.536e-03  3.404 0.000752 ***
## PC7                  -3.677e-02  5.181e-03 -7.097 9.04e-12 ***
## PC8                  5.024e-02  6.568e-03  7.649 2.69e-13 ***
## PC9                  -1.114e-01  6.059e-03 -18.382 < 2e-16 ***
## PC10                 3.449e-02  8.541e-03  4.038 6.82e-05 ***
## PC11                 -8.359e-04  8.560e-03 -0.098 0.922271
## PC12                 3.693e-02  8.893e-03  4.152 4.28e-05 ***
## PC13                 4.191e-02  8.280e-03  5.062 7.21e-07 ***
## PC14                 -1.993e-02  8.692e-03 -2.293 0.022546 *
## PC15                 7.445e-03  8.487e-03  0.877 0.381053
## PC16                 1.925e-02  9.725e-03  1.979 0.048678 *
## ...
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.196 on 304 degrees of freedom
## Multiple R-squared:  0.7897, Adjusted R-squared:  0.7704
## F-statistic: 40.78 on 28 and 304 DF,  p-value: < 2.2e-16

## Export for editing with Excel
write.csv(tidy(explore_gene(i, vGene, model = 3)), file = 'gene2.csv')
```

As previously explained, the interpretation of the p-value in the model for each gene determines whether or not the gene is considered differentially expressed between schizophrenic cases and non-psychiatric controls. For *AHCLY1*, using the diagnosis only model, the gene is considered differentially expressed ($p = 0.00656$). In the second model, once the RNA-quality and demographic covariates are added in, *AHCLY1* remains differentially expression ($p = 0.00459$). In the final model, with qSVs added in addition to diagnosis, RNA-quality, and

Applying Statistical Correction for Brain Tissue RNA Degradation to Gene Expression Differences in Schizophrenia

demographic covariates, *AHCLY1* remains differentially expressed ($p = 8.38 \times 10^{-6}$) between cases and controls with a log₂ fold change of 1.188×10^{-1} (higher expression in non-psychiatric controls).

4.3 Adjusted covariates interpretation

Further, in terms of interpretation of other coefficients relevant to the model, RNA integrity number (RIN) is a covariate that changes significance once the qSVs are added in to the model in model 3. This increases the confidence that RNA quality confounding is adequately controlled for, since RIN is statistically significant in Model 2, but no longer appears significant in Model 3.

For *ELP6*, the false positive example that appeared to be differentially expressed until the qSVs were added in for Model 3. In Model 2, where *ELP6* still appears to be differentially expressed, RIN is statistically significant ($p=1.18 \times 10^{-7}$). In Model 3, where *ELP6* is no longer differentially expressed, RIN is not statistically significant ($p=0.50511$).

For *AHCYL1*, the gene that remains differentially expressed throughout the three models, RIN remains non-statistically significant in both Model 2 and 3. In Model 2 RIN is not statistically significant ($p=0.87420$). In Model 3, where *AHCYL1* remains differentially expressed, RIN is not statistically significant ($p=0.072202$).

5 Conclusions

The MLR analysis shown for the two example genes was completed for all of the approximately 25,000 genes contained in the BrainSeq dataset, a case-control dataset comparing schizophrenic cases to non-psychiatric controls. In the examples shown, *ELP6* was a false positive gene incorrectly identified as differentially expressed between the two groups until the qSVs were added, and *AHCLY1* was a gene that remained differentially expressed throughout all 3 models. The ability to correctly identify the genes in the dataset as differentially expressed between the two groups was dependent on adequate control of RNA quality confounding. The effectiveness of the statistical correction for RNA quality confounding was shown by the DEqual plots displayed on slides 12 - 14.

Overall, the qSVA method greatly reduced the number of differentially expressed genes between schizophrenia cases and non-psychiatric controls, thus increasing the confidence in the results.

Session info

This section shows the R session information for reproducing the analysis. This document was generated using the JHPCE computing cluster due to the size of the data analyzed.

```
## Session info -----
## setting  value
## version  R version 3.5.0 Patched (2018-04-30 r74679)
## system   x86_64, linux-gnu
## ui        X11
## language (EN)
```

Applying Statistical Correction for Brain Tissue RNA Degradation to Gene Expression Differences in Schizophrenia

```
## collate en_US.UTF-8
## tz      US/Eastern
## date   2018-05-18
## Packages -----
## package      * version  date     source
## assertthat    0.2.0    2017-04-11 CRAN (R 3.5.0)
## backports     1.1.2    2017-12-13 CRAN (R 3.5.0)
## base         * 3.5.0    2018-05-02 local
## bindr          0.1.1    2018-03-13 CRAN (R 3.5.0)
## bindrcpp      0.2.2    2018-03-29 CRAN (R 3.5.0)
## Biobase        * 2.40.0   2018-05-02 Bioconductor
## BiocGenerics   * 0.26.0   2018-05-03 Bioconductor
## BiocParallel   * 1.14.1   2018-05-17 Bioconductor
## BiocStyle       * 2.8.1    2018-05-17 Bioconductor
## bitops          1.0-6    2013-08-17 CRAN (R 3.5.0)
## bookdown        0.7      2018-02-18 CRAN (R 3.5.0)
## broom          * 0.4.4    2018-03-29 CRAN (R 3.5.0)
## compiler        3.5.0    2018-05-02 local
## datasets        * 3.5.0    2018-05-02 local
## DelayedArray    * 0.6.0    2018-05-02 Bioconductor
## devtools        * 1.13.5   2018-02-18 CRAN (R 3.5.0)
## digest          0.6.15   2018-01-28 CRAN (R 3.5.0)
## dplyr           0.7.4    2017-09-28 CRAN (R 3.5.0)
## edgeR            * 3.22.1   2018-05-17 Bioconductor
## evaluate        0.10.1   2017-06-24 CRAN (R 3.5.0)
## foreign          0.8-70   2017-11-28 CRAN (R 3.5.0)
## GenomeInfoDb    * 1.16.0   2018-05-03 Bioconductor
## GenomeInfoDbData 1.1.0    2018-04-17 Bioconductor
## GenomicRanges   * 1.32.3   2018-05-17 Bioconductor
## glue             1.2.0    2017-10-29 CRAN (R 3.5.0)
## graphics         * 3.5.0    2018-05-02 local
## grDevices        * 3.5.0    2018-05-02 local
## grid              3.5.0    2018-05-02 local
## htmltools         0.3.6    2017-04-28 CRAN (R 3.5.0)
## IRanges          * 2.14.10  2018-05-17 Bioconductor
## knitr            1.20     2018-02-20 CRAN (R 3.5.0)
## lattice           0.20-35  2017-03-25 CRAN (R 3.5.0)
## limma            * 3.36.1   2018-05-17 Bioconductor
## locfit            1.5-9.1  2013-04-20 CRAN (R 3.5.0)
## magrittr          1.5      2014-11-22 CRAN (R 3.5.0)
## Matrix            1.2-14   2018-04-13 CRAN (R 3.5.0)
## matrixStats       * 0.53.1   2018-02-11 CRAN (R 3.5.0)
## memoise           1.1.0    2017-04-21 CRAN (R 3.5.0)
## methods          * 3.5.0    2018-05-02 local
## mnormt            1.5-5    2016-10-15 CRAN (R 3.5.0)
## nlme              3.1-137  2018-04-07 CRAN (R 3.5.0)
## parallel          * 3.5.0    2018-05-02 local
## pillar            1.2.2    2018-04-26 CRAN (R 3.5.0)
## pkgconfig         2.0.1    2017-03-21 CRAN (R 3.5.0)
## plyr              1.8.4    2016-06-08 CRAN (R 3.5.0)
## psych             1.8.4    2018-05-06 CRAN (R 3.5.0)
```

Applying Statistical Correction for Brain Tissue RNA Degradation to Gene Expression Differences in Schizophrenia

```
## purrr          0.2.4    2017-10-18 CRAN (R 3.5.0)
## R6              2.2.2    2017-06-17 CRAN (R 3.5.0)
## Rcpp            0.12.16   2018-03-13 CRAN (R 3.5.0)
## RCurl           1.95-4.10 2018-01-04 CRAN (R 3.5.0)
## reshape2        1.4.3    2017-12-11 CRAN (R 3.5.0)
## rlang            0.2.0    2018-02-20 CRAN (R 3.5.0)
## rmarkdown        1.9      2018-03-01 CRAN (R 3.5.0)
## rprojroot       1.3-2    2018-01-03 CRAN (R 3.5.0)
## S4Vectors       * 0.18.2   2018-05-17 Bioconductor
## stats            * 3.5.0   2018-05-02 local
## stats4           * 3.5.0   2018-05-02 local
## stringi          1.2.2    2018-05-02 CRAN (R 3.5.0)
## stringr          1.3.1    2018-05-10 CRAN (R 3.5.0)
## SummarizedExperiment * 1.10.1  2018-05-17 Bioconductor
## tibble           1.4.2    2018-01-22 CRAN (R 3.5.0)
## tidyverse         0.8.0    2018-01-29 CRAN (R 3.5.0)
## tools             3.5.0    2018-05-02 local
## utils            * 3.5.0   2018-05-02 local
## withr            2.1.2    2018-03-15 CRAN (R 3.5.0)
## xfun              0.1      2018-01-22 CRAN (R 3.5.0)
## XVector          0.20.0   2018-05-03 Bioconductor
## yaml              2.1.19   2018-05-01 CRAN (R 3.5.0)
## zlibbioc         1.26.0   2018-05-02 Bioconductor
```