Differential Expression Analysis to Investigate SULT2B1b Knockdown in Prostate Cancer Cells

Prepared for the Ratliff Lab
Faye Zheng, Nadia Atallah, and R.W. Doerge
October 20, 2015

1. Introduction

Single-cell RNA-seq expression profiles from human LNCap prostate adenocarcinoma cells were delivered to the Doerge group in the form of a count matrix. Previously, these cells underwent sorting, sample prep and sequencing by the Ratliff lab and the Purdue Genomics Facility. Nadia Atallah of the Purdue Center for Cancer Research performed quality control, alignment, and expression quantification (see details in her report, attached here in the appendage). The treatment groups of interest consist of 1) cells that underwent siRNA knockdown of the SULT2B1b isoform, and 2) cells that were introduced to untargeted siRNA as a negative control. The goal of the analysis is to perform an unbiased analysis of the impact of SULT2B1b knockdown on gene expression. The resulting list of differentially expressed genes was delivered to Nadia Atallah for downstream pathway analyses (see details in the appendage).

2. Data Exploration

2.1 Experimental Design

Samples were prepared in separate batches, owing to restrictions on the number of plates containing cells that may be sorted per day. Each batch contains one set of control (C) and one set of knockdown (KD) cells. The schematic, as well as the number of cell replicates per group and batch, are depicted in Table 1. The ensuing differential expression analysis is only concerned with comparing gene expression between treatment groups, while accounting for potential batch effects.

Batch 1		Batch 2		Batch 3	
Control n=76	Knockdown n=64	Control n=66	Knockdown n=65	Control n=67	Knockdown n=61

Table 1. Cells were prepared in 3 batches, with each batch containing one set of control and one set of knockdown cells.

2.2 Filter Low Expression Genes

The original data comprised 36,135 sequenced genes, many of which exhibit very low expression levels. Omitting low-expression genes that contribute little to the analysis yields a more powerful statistical test overall. We chose to keep only the genes that have average counts of at least 5 across all samples; this is standard practice in the literature. We removed 25,281 genes using this criterion, comprising 70.0% of the original number; 10,854 genes remain for analysis.

2.3 Exploratory Plots

Multidimensional scaling (MDS) plots can be used to visually assess similarities and dissimilarities between samples. The distance between each pair of samples is the Euclidean distance for the genes with the highest (leading) log-fold-change between those samples. Hence, samples that are similar to each other group together. Figures 1 and 2 depict the same MDS plot, but with colors and labels highlighting the separation of samples with respect to either batch or experimental group. The plots suggest a clear degree of biological difference between knockdown and control cells (Figure 1), but no discernible difference between batches (Figure 2).

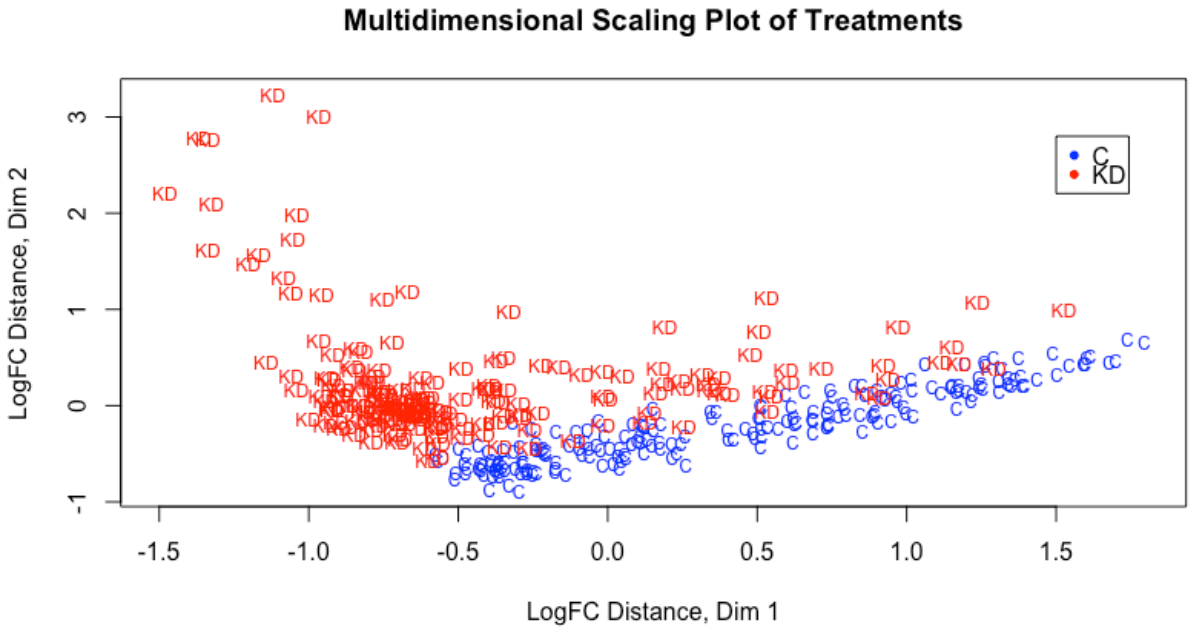

Figure 1. MDS plot of the samples, with labels and colors highlighting the experimental treatments (KD for knockdown cells, C for control). Here, the samples separate nicely by group, suggesting a clear degree of difference between groups.

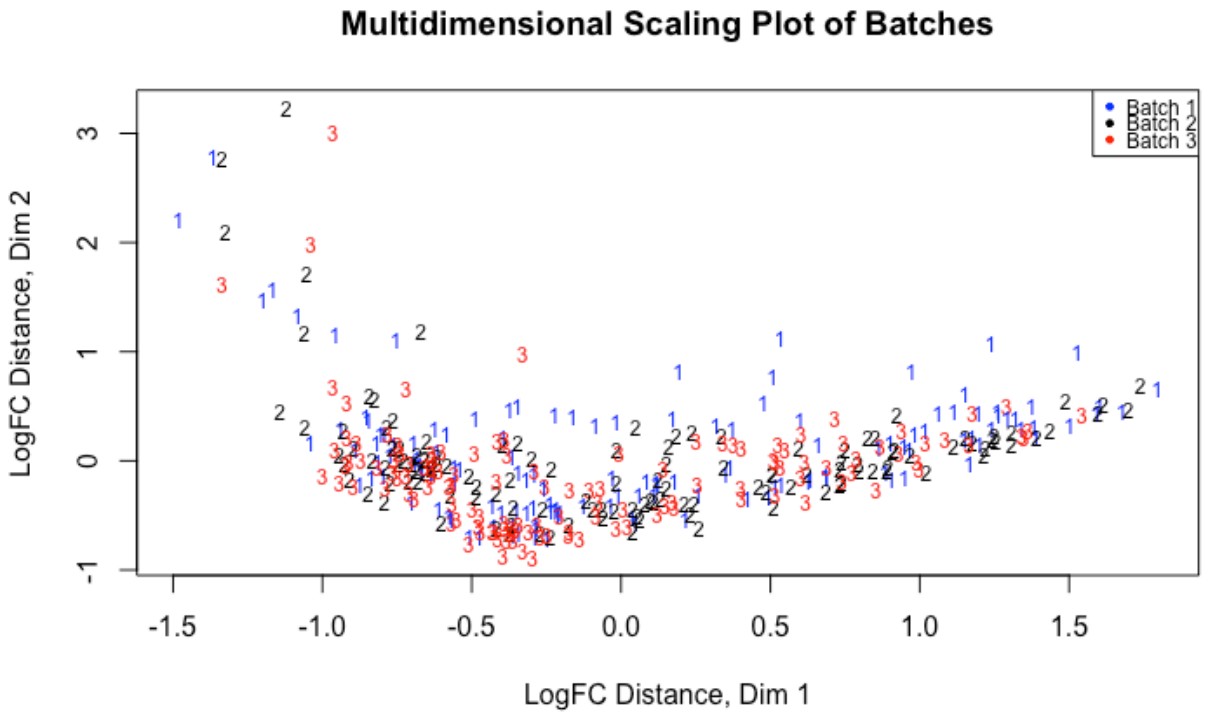

Figure 2. MDS plot of the samples, with labels and colors highlighting the batch in which each cell was collected. Here, the samples do not clearly separate by batch, suggesting that there is no discernible biological difference between batches, and by extension that batch is not a serious confounding factor.

3. Methods: Testing Effect of SULT2B1b Knockdown

3.1 Negative Binomial Generalized Linear Model

The objective of the analysis is to determine which genes are differentially expressed between control cells (C) and SULT2B1b knockdown cells (KD). Differential expression was tested using the Bioonductor package *edgeR* in R (v. 3.2.2). A negative binomial (NB) distribution, being a discrete distribution, was employed for modeling the count data generated from the RNA-seq experiments. While a Poisson distribution can also be used to model count data, the Poisson assumption that the variance is equal to the mean is generally untrue for RNA-seq data and hence too restrictive. We are allowed more flexibility with the NB distribution due to the additional use of the dispersion parameter, ϕ_g. This parameter allows for a more accurate modeling of the variability between samples. Note that when the ϕ_g is zero, the NB model reduces to the Poisson model.

Under the NB model, the data are distributed as

$$Y_{gijk} \sim NB(M_k p_{gij},\ \phi_g),$$

where Y_{gijk} is the number of reads from cell k of experimental group i (C, KD) and batch j (1, 2, 3) that are mapped to gene g; M_k is the total number of mapped reads, i.e. library size, for cell k; p_{gij} is the proportion of all reads that originate from gene g in the i^{th} group, j^{th} batch; and ϕ_g is the dispersion parameter for gene g.

The generalized linear model for this analysis is denoted by

$$\log(Y_{gk}) = \beta_{0g} + \beta_{1g}KD_k + \beta_{2g}Batch2_k + \beta_{3g}Batch3_k + logM_k + \varepsilon_{gk} \quad (1)$$

where Y_{gk} is the observed count for gene g in cell k; KD_k is an indicator variable for cell k being a knockdown as opposed to a control cell; $Batch2_k$ and $Batch3_k$ are indicator variables for cell k having been processed from batch 2 and 3, respectively, with batch 1 being used as a basis for comparison; logM_k is an offset term representing the cell k library size; and ε_{gk} is the error term.

For this model, we are primarily interested in the effect of SULT2B1b knockdown on gene expression, while the differences between batches are not of primary interest. Hence, we treat Batch as an experimental blocking factor. This will allow us to test primarily for gene expression differences between knockdown and control cells, while adjusting for any nuisance effects between batches. Specifically, we are testing for each gene g the null hypothesis that there is no effect of knockdown on expression. The hypotheses are denoted as follows for gene g.

$$H_0: \beta_{1g} = 0$$
$$H_a: \beta_{1g} \neq 0$$

The test statistic for testing the above hypotheses is as follows:

$$F_g = LRT_g\ \phi_g \sim F_{1,n-p} \text{ under } H_0 \quad (3)$$

where LRT_g is the quasi-likelihood test statistic; ϕ_g is the dispersion parameter estimation in the NB model; n is the sample size; and p is the number of parameters estimated in the model. From (3) we can see that accurately estimating the dispersion parameter is important since underestimating ϕ_g tends to cause lower p-values, resulting in a higher false discovery rate.

3.2 Adjustments for biases in RNA composition

There are situations when a set of genes is highly expressed in one sample but not in another. When this occurs, the remainder of the genes in the first sample would be "under-sampled", and thus creates a potential bias due to its particular RNA composition. To prevent this from occurring and skewing the differential expression analysis and results, we normalized the data using an empirical approach that estimates bias (Robinson and Oshlack, 2010). The scaling factors that were estimated across the 399 total samples had a middle 50% range of 1.036 to 1.141; the departure of these factors from 1 indicates the presence of compositional differences between libraries.

3.3 Estimating dispersion

As explained previously, ϕ_g is the dispersion parameter of the negative binomial model. We first estimated a common dispersion, which is the average ϕ_g across all genes, and then extended this by estimating a separate dispersion for each individual gene. This was done using an empirical Bayes method that "squeezes" the genewise dispersions towards the common dispersion, thus allowing for information-borrowing from other genes (Robinson, McCarthy, and Smyth, 2010).

3.4 Adjustments for multiple testing

Since we are performing a separate statistical test for all of the 10,854 genes, it is necessary to adjust the p-values for multiple testing (i.e., we want to control the Type I error rates across the "family" of genes rather than for "each" gene). This was accomplished using the Benjamini-Hochberg procedure for controlling the expected proportion of incorrectly rejected null hypotheses, also known as the false discovery rate (FDR) (Benjamini and Hochberg, 1995).

5. Results

Data from the 209 control and 190 knockdown cells were used to test each gene for differential expression between knockdown and control cells, using the negative binomial test (Section 3), after accounting for any batch effects. Controlling FDR within 5%, we found 2029 differentially expressed genes, representing 18.69% of all genes. Of these differentially expressed genes, 1073 had higher expression in the knockdown group, and 956 had higher expression in the control group. Table 2 lists the

top ten most significant genes. The full list of differentially expressed genes has been delivered to Nadia Atallah; she will perform pathway analyses to more deeply investigate the biological consequences of knocking out SULT2B1b.

Gene	logFC	logCPM	PValue	Adj. Pvalue
ENSG00000126709	5.5484	6.5219	4.10E-106	4.45E-102
ENSG00000187608	4.1391	7.8612	1.81E-75	9.81E-72
ENSG00000119917	10.5402	9.6884	1.04E-51	3.76E-48
ENSG00000119922	9.1443	10.4925	2.78E-50	7.53E-47
ENSG00000146677	-1.3556	6.9920	2.77E-47	6.02E-44
ENSG00000144713	-1.3548	8.1579	4.23E-47	7.66E-44
ENSG00000142089	1.6650	8.0496	2.20E-46	3.41E-43
ENSG00000185745	8.7035	9.0046	1.03E-44	1.39E-41
ENSG00000135114	11.5586	8.6546	1.02E-42	1.23E-39
ENSG00000213881	-1.0501	5.1340	5.85E-39	6.34E-36

Table 2. Top ten differentially expressed genes between knockdown and control cells; the total list of significant differentially expressed genes contains 2029 genes. Log(FC) is the log-fold change in expression between groups; positive log(FC) values indicate higher expression levels for the knockdown group. Log(CPM) is the log-counts per million (i.e. expression level) between the two groups.

6. Discussion

6.1 Methods for tissue-level vs. single-cell RNA-seq data

While we chose the Bioconductor R package *edgeR* for testing differential expression, another widely-used statistical package that is comparable to *edgeR* in performance and popularity is *DESeq2* (Love, Hubers, and Anders, 2014). *DESeq2* was attempted on these data but faced major computational shortcomings due to the large number of cell samples. This is not surprising; RNA-seq experiments on tissue-level samples rarely saw sample sizes beyond a few dozen, contrasted with the almost 400 cell samples seen in this experiment.

It is important to remark that both *edgeR* and *DESeq2* were originally developed for tissue-level RNA-seq data, and are not specific to single-cell RNA-seq (scRNA-seq) data. These methods were selected in this analysis for their well-established performance on tissue-level RNA-seq data, and can be expected to apply acceptably to single-cell data without major issue. However, scRNA-seq data is increasingly understood to exhibit properties that are distinct from tissue-level data, owing to both technical and

biological reasons. As single cells contain such low starting amounts of biological material, the resulting scRNA-seq data exhibit substantially higher technical variability than is typically observed in tissue-level data, and hence suffer from frequent non-detection of even moderately expressed genes (Brennecke, et al. 2013; Grun, Kester, and van Oudenaarden, 2014). Biologically, single-cell gene expression may be affected by latent factors such as cell cycle (Chen et al., 2013), or transcriptional bursts of individual genes or gene networks (Munsky, Neuert, and van Oudenaarden, 2012).

While there have been some ventures into method development for differential expression testing of scRNA-seq data (Kharchenko, Silberstein, and Scadden, 2014; Trapnell, et al. 2014), their performance has not yet been thoroughly demonstrated on real data, and their potential advantages over existing well-established methods such as *edgeR* and *DESeq2* has not been fully seen. However, this is becoming an increasingly active area of research, and an important goal moving forward is to incorporate new methods specific to scRNA-seq data to analyze experiments such as this one.

6.2 Treating SULT2B1b expression as continuous

The differential expression analysis presented in this report was carried out between the following experimental groups: 1) cells that underwent siRNA knockdown of the SULT2B1b isoform, and 2) cells that were introduced to untargeted siRNA as a negative control. The underlying motivation is to investigate how the presence or absence of SULT2B1b affects the expression of other genes. However, it has been observed that the siRNA knockdown of SULT2B1b isoform is only 80% efficient. While control cells do tend to have higher expression than knockdown cells, the assumption of a binary presence/absence of SULT2B1b between the two experimental groups does not hold.

Descriptive plots provided in the supplement at the end of the report treat SULT2B1b expression as continuous, rather than binary, and can be used to motivate further analysis strategies. Further work will be done to render these plots interactive via RShiny, a web application framework for R. This way, the user will be able to select various inputs of genes and other parameters, as desired.

7. References

Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society* 57, 289-300.

Brennecke, P. et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods* 10, 1096-1098.

Chen, W.C. et al. (2013). Functional interplay between the cell cycle and cell phenotypes. *Integrative Biology* 5: 523– 534.

Grun, D., Kester, L. and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nature Methods* 11, 637-640.

Kharchenko, P.V., Silberstein, L. and Scadden, D.T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods* 11, 740-742.

Love, M.I., Huber, W. and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 15:550.

Munsky, B., Neuert, G. and van Oudenaarden, A. (2012). Using gene expression noise to understand gene regulation. *Science* 336 (6078), 183-187.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139-140.

Robinson, M.D. and A. Oshlack (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* 11 (3).

Trapnell, C. et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology* 32, 381-386.

8. Supplementary Plots

8.1 SULT2B1b Expression in Each Group

Figure 1 shows SULT2B1b expression in control and knockdown groups. The percentage of cells exhibiting no SULT2B1b expression (zero counts) is 98% in the knockdown group and 79% in the controls. So, while the knockdown treatment is not perfect, and while a substantial proportion of cells in both groups exhibit no SULT2B1b expression, there is still a clear separation of the two groups with respect to SULT2B1b presence. Note that the horizontal range of the points across the x-axis is simply a graphical jitter; it is meant for visual clarity and has no further meaning.

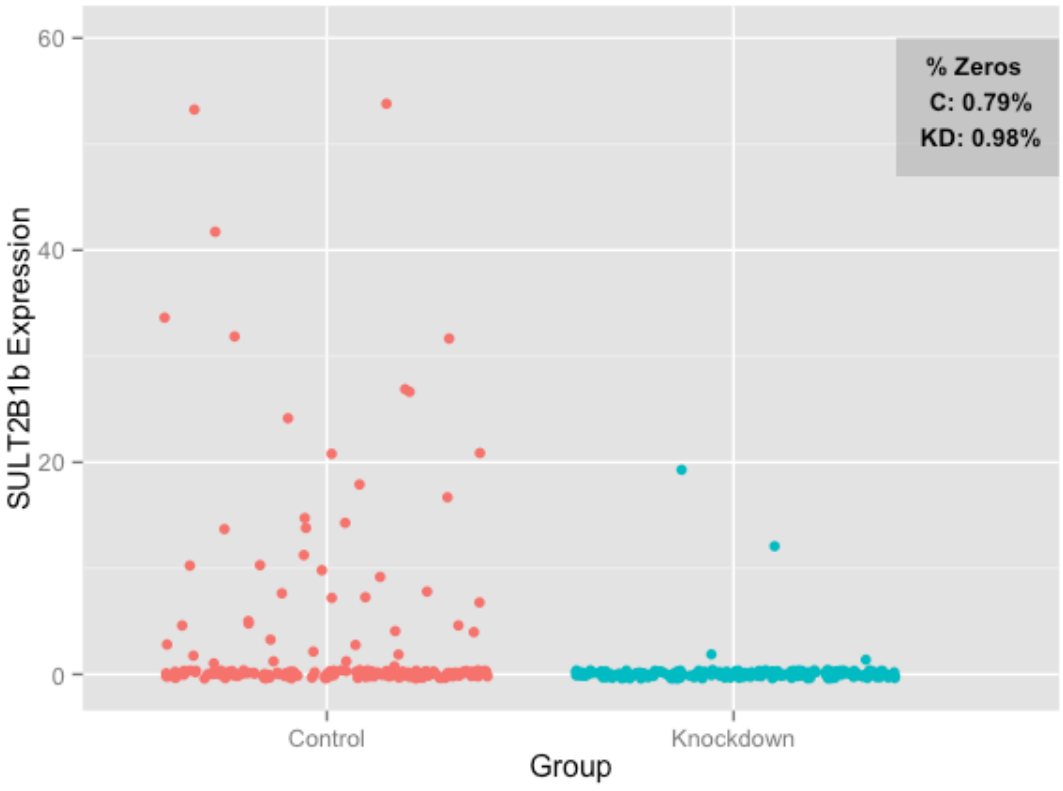

Figure 1. SULT2B1b expression in control and knockdown groups.

8.2 SULT2B1b Expression, with High and Low Thresholds

One idea is that instead of defining experimental groups as control/knockdown depending on receipt of the siRNA knockdown treatment, one could define groups as high/low depending on observed SULT2B1b expression itself. This would necessitate selecting high and low thresholds for SULT2B1b expression, in order to separate cells with high SULT2b1b expression from those with low SULT2B1b expression. Differential expression analyses with groups defined this way may be a more accurate way to distinguish biological behavior between cells with and without SULT2B1b expression. Obvious questions involve how to select the thresholds.

Figure 2 depicts SULT2B1b expression with arbitrarily chosen thresholds separating cells with high vs. low SULT2B1b expression. In this example, the high group consists of only control cells (no siRNA knockdown treatment), while the low group consists of a mix of cells from either experimental group. This highlights the added flexibility of defining groups based on SULT2B1b expression values themselves, rather than on siRNA treatment.

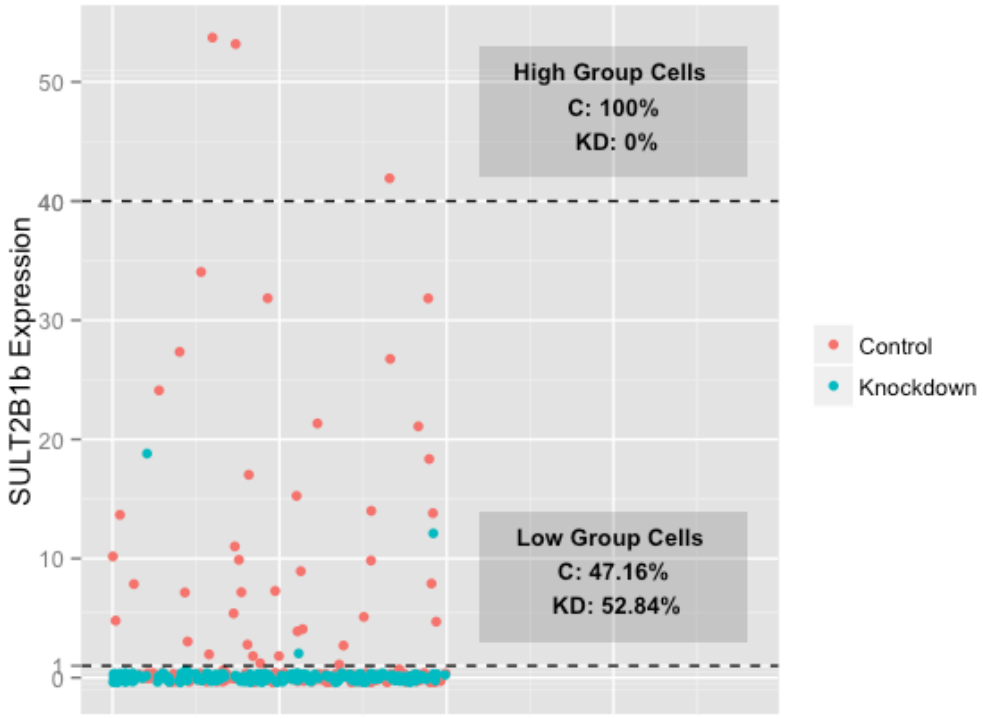

Figure 2. SULT2B1b expression, with thresholds separating cells with high vs. low SULT2B1b expression.

8.3 Correlation between SULT2B1b and other genes of interest

With SULT2B1b expression treated as continuous, one may plot correlations between SULT2B1b expression and other genes of interest. For example, Figure 4 depicts the correlation between AR expression values and SULT2B1b expression values; the points are cells, colored by siRNA treatment group. The correlation is low, in large part due to the large number of zeros in SULT2B1b expression values.

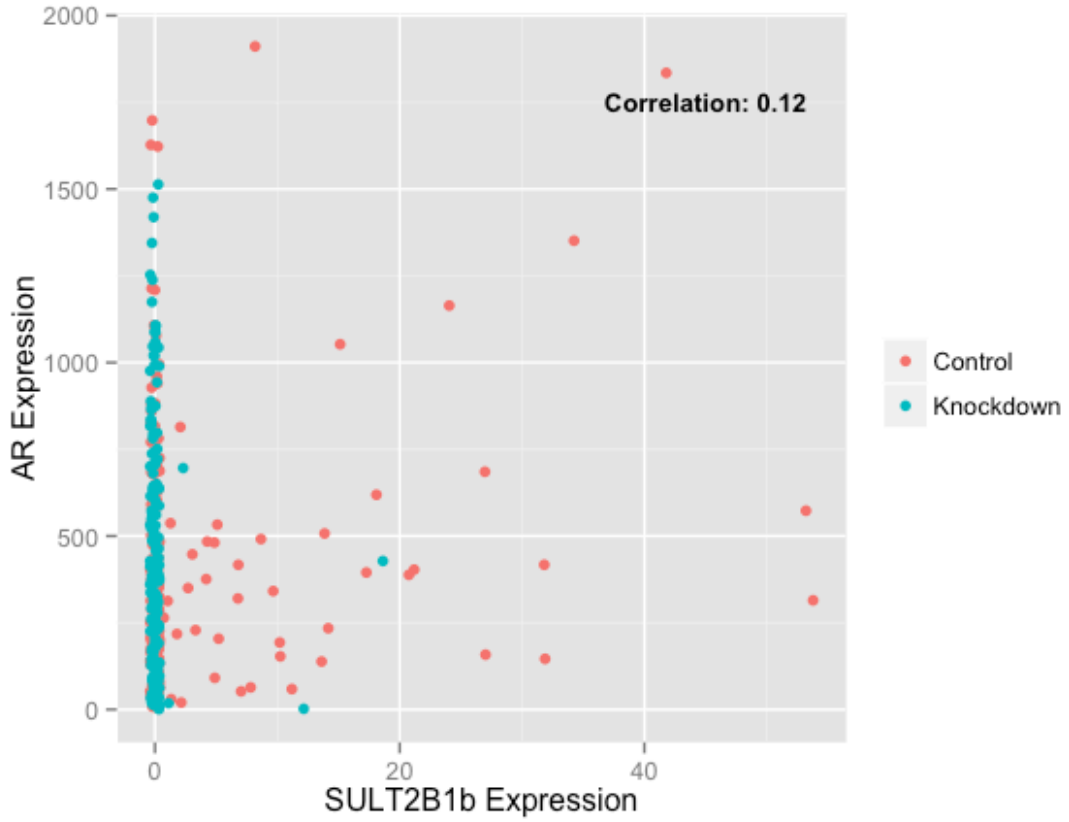

Figure 4: Correlation between SULT2B1b and AR expression values.