

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).

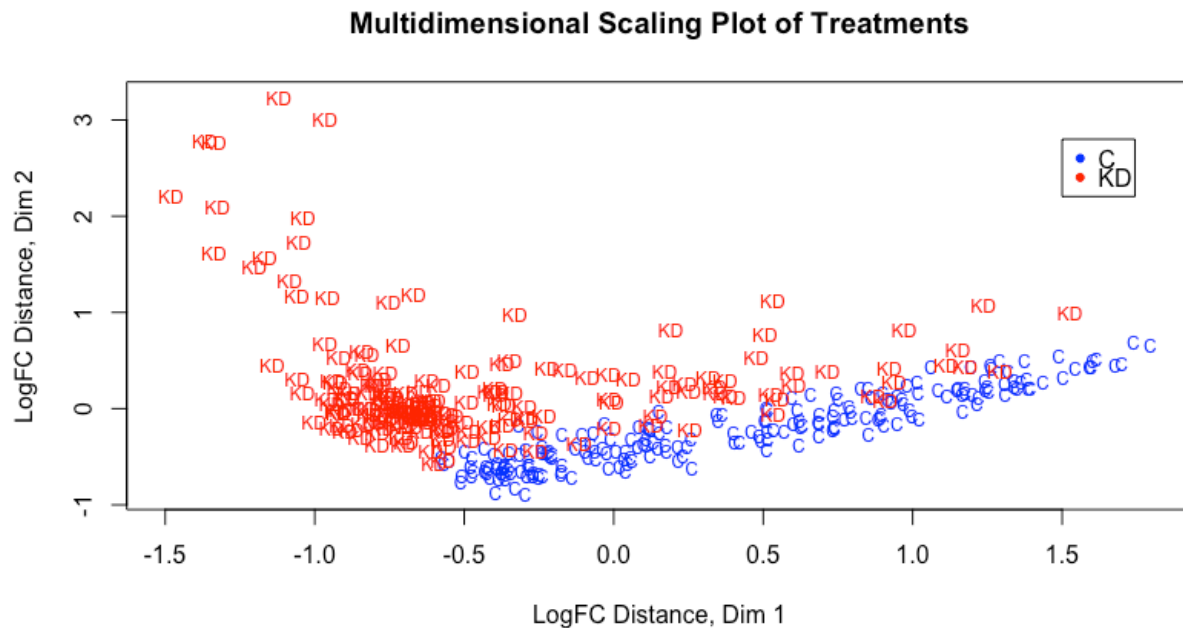


Figure 1. MDS plot of the samples, with labels and colors highlighting the experimental treatments (KD for knockdown cells, C for control). Here, the samples separate nicely by group, suggesting a clear degree of difference between groups.

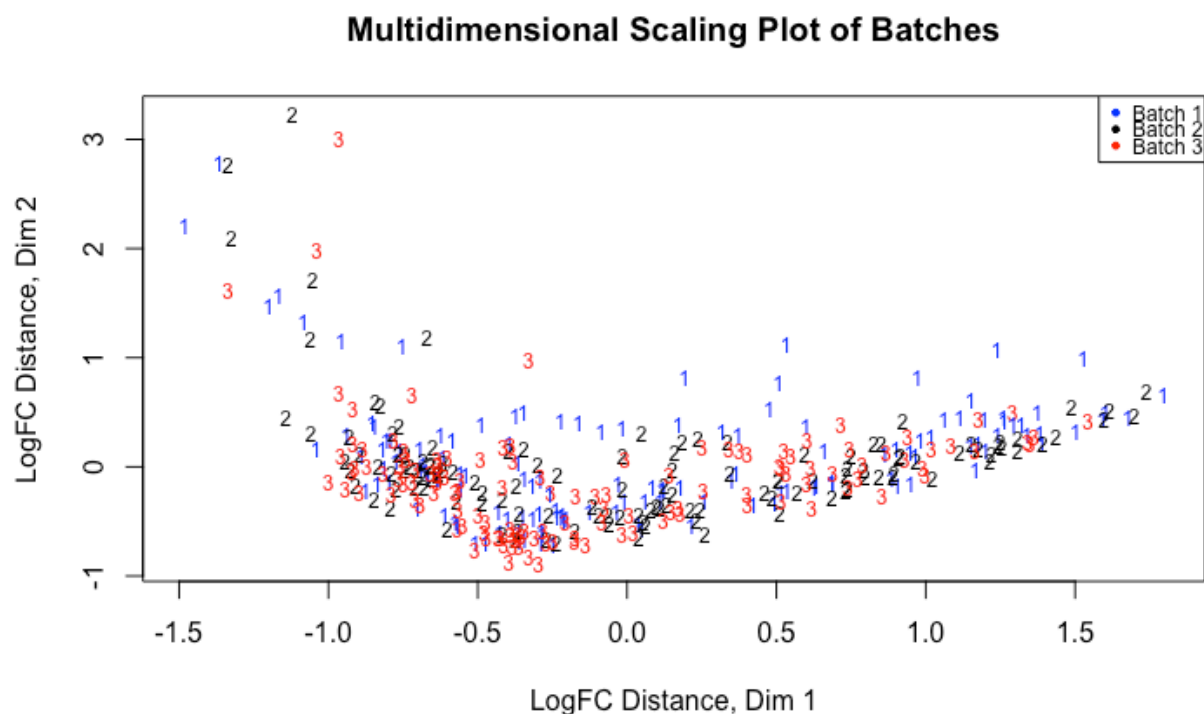


Figure 2. MDS plot of the samples, with labels and colors highlighting the batch in which each cell was collected. Here, the samples do not clearly separate by batch, suggesting that there is no discernible biological difference between batches, and by extension that batch is not a serious confounding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

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

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; $\log M_k$ 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.

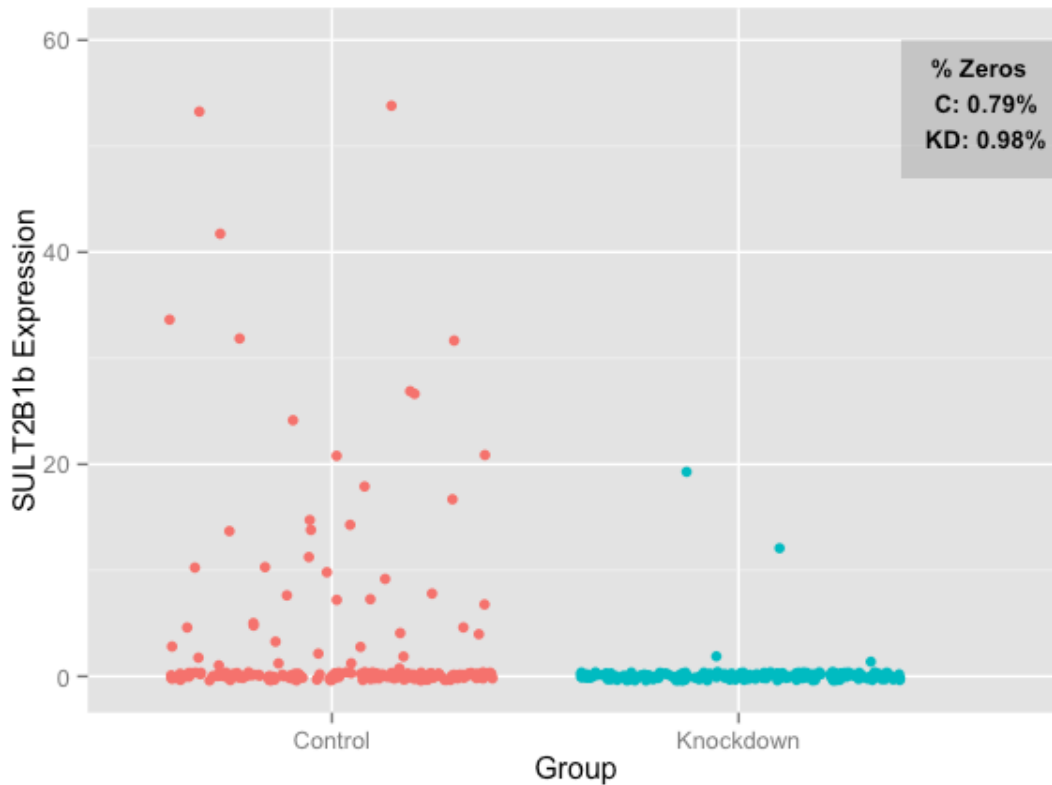


Figure 1. SULT2B1b expression in control and knockdown groups.

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.

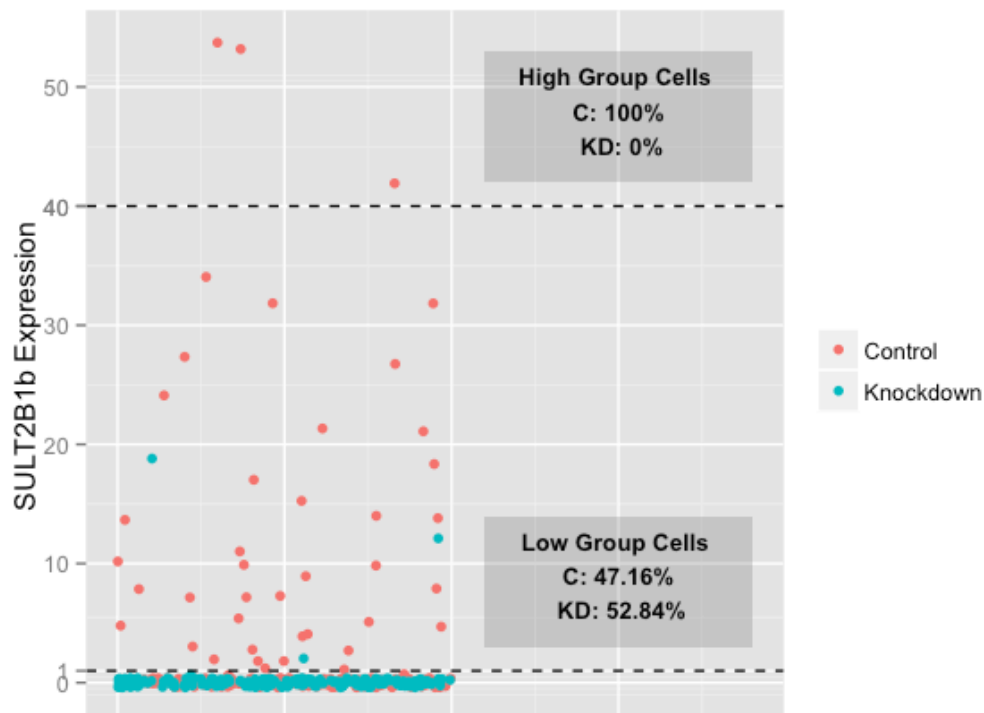


Figure 2. SULT2B1b expression, with thresholds separating cells with high vs. low SULT2B1b expression.

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.

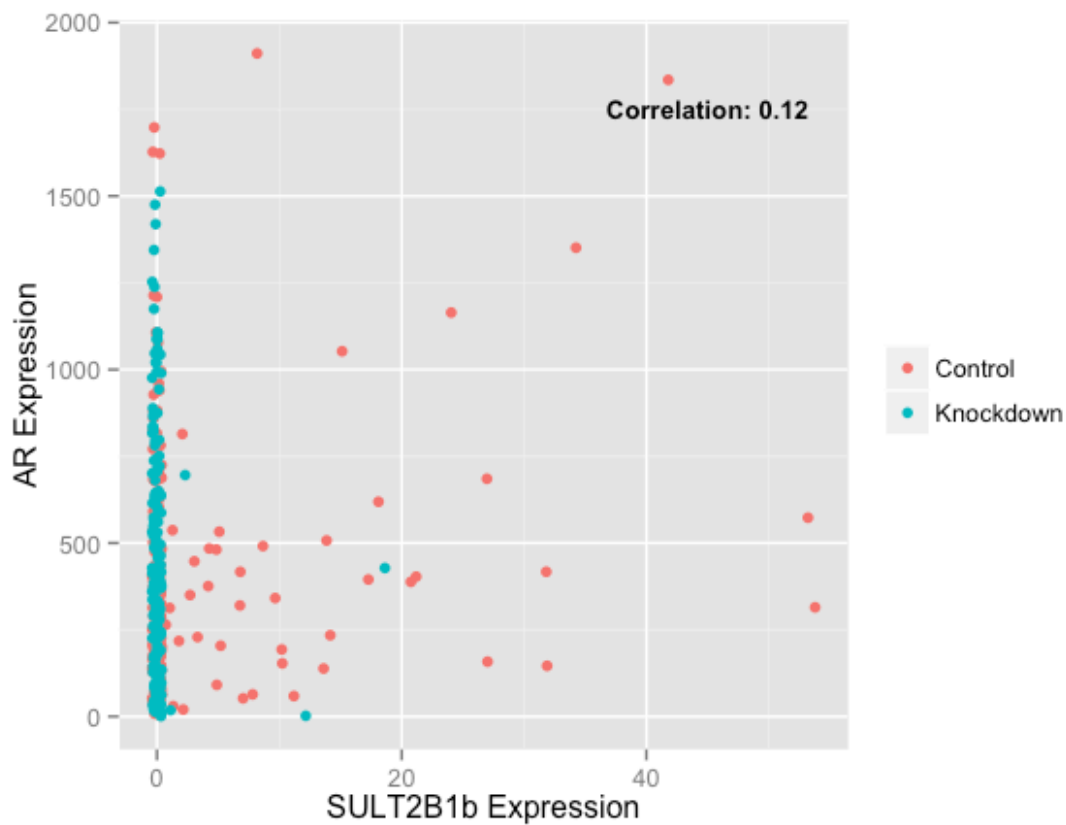


Figure 4: Correlation between SULT2B1b and AR expression values.

SULT2B1b knock-down single-cell RNA-seq

Nadia Atallah

September 29, 2015

Introduction

Cholesterol sulfate is present at high concentrations in prostate cancer cells and is a potential indicator for prostate cancer (Eberlin et al., 2010), however the implications of these high concentrations of cholesterol sulfate and likewise the mechanism leading to them are unknown. Sulfonation of cholesterol is performed by SULT2B1b, the most highly expressed SULT2B (sulfotransferase 2B) isoform (Higashi et al., 2004). The Ratliff lab recently found that SULT2B1b knock-down affects the viability of prostate cancer cells, however the reasons for this are unknown. Single-cell RNA-seq provides an opportunity to observe the effects of SULT2B1b at a transcriptional level, however challenges exist in the analysis of single-cell data. Statistical methodology that appropriately handles the specifics of single-cell data still need to be identified/developed (Ding et al., 2015; Stegle, Teichmann, & Marioni, 2015). The goals of this study are a) to perform an unbiased analysis on the impact of SULT2B1b siRNA knock-down on gene expression and b) to develop/identify a statistical methodology for performing normalization and differential expression analysis on single-cell RNA-seq data.

Human LNCaP prostate adenocarcinoma cells were targeted with either a construct that leads to the siRNA knock-down of SULT2B1b or with a negative control, targeting no gene. The SULT2B1b knock-down efficiency is ~80%. The C1 Fluidigm Single-Cell Auto Prep system was used to capture single-cells, perform live-dead screening, perform reverse transcription, and to amplify cDNA (Fluidigm, San Francisco, CA). The SMARTer Ultra Low RNA Kit for Illumina Sequencing (Clontech, Mountain View, CA) was used in conjunction with the Nextera kit (Illumina, San Diego, CA) to prepare sequencing libraries. SULT2B1b is needed for cell viability and so live/dead screening was performed prior to introduction into the C1 Fluidigm and then again prior to cDNA synthesis. It was decided that ~1 million reads would be sequenced per cell and up to 96 cells will be sequenced per lane. Paired-end reads are utilized in order to identify gene-fusions and to aid in unambiguous alignment of reads to the reference genome. Cell libraries from all cells on a plate will be barcoded and cDNA combined into one lane. At this time, it is only feasible for one 96-well plate of single-cell cDNA to be prepared per day. For this reason, plates were prepared in batches. Sequencing was performed in three batches using three flow cells with 2 lanes per flow cell. Each batch contains a 96 well plate with knock-out cells and a 96 well plate with control cells. Cells with SULT2B1b knocked-down are being statistically compared to those that had only negative control RNA delivered. Differences in gene expression patterns amongst these cells will give us insight into novel pathways to study and potential mechanisms for how this protein is functioning within LNCaP cells. It is expected that the androgen receptor (AR) gene (ENSG00000169083), PSA (ENSG00000142515) and other androgen driven products will be lower in expression in the knock-down cells compared to control cells. It is also expected that the TNF α gene (ENSG00000232810) will exhibit higher expression in knock-down cells compared to in control cells.

Data was sequenced on an Illumina Hi-Seq2500 on RapidRun mode. Illumina and Clontech adapters were removed via Trimmomatic (Bolger, Lohse, & Usadel, 2014). Quality trimming was performed using FastX-Toolkit (Gordon, 2009) and Trimmomatic. FastQC (Andrews, 2010) and FastX-

Toolkit both were used to provide quality control summaries and graphs. Tophat2 (Kim et al., 2013; Trapnell, Pachter, & Salzberg, 2009) was used to align reads to the Ensembl Genome Reference Consortium Human Build 38 (GRCh38.p3). HTSeq (Anders, Pyl, & Huber, 2015) was used to count reads mapping to features. Matrices were then generated and biomaRt (Durinck et al., 2005; Durinck, Spellman, Birney, & Huber, 2009; Smedley et al., 2015) was used in obtaining the annotations for each gene. Matrices were subsequently delivered to Faye Zheng in Dr. Doerge's group for analysis. The program FusionCatcher (Nicorici et al., 2014) was run to identify potential gene fusions in the cells. Pathway analysis was performed using QIAGEN's Ingenuity Pathway Analysis (IPA, QIAGEN Redwood City, www.qiagen.com/ingenuity).

Methods

Cells Sorting, Library Preparation, and Sequencing

Data was collected for the all three batches at the Purdue Cell Cytometry Facility using the C1 Fluidigm instrument with SMARTer chemistry (Clontech, Mountain View, CA) to generate cDNA from captured single cells. This chemistry uses a modified oligo(dT) primer to prime 1st strand synthesis. The Purdue Genomics Facility prepared libraries using a Nextera kit (Illumina, San Diego, CA). Unstranded 2x100 bp reads were sequenced using the HiSeq2500 on rapid run mode. Before library preparation the dsDNA quality was checked using an Agilent Bioanalyzer with the High Sensitivity DNA Chip. The "S" samples are the experimental (SULT2B1b knock-down) group and the "C" samples are the control group. For each batch, bulk RNA-seq samples were prepared - one for each condition; these RNA samples were not obtained from cells captured by the C1 Fluidigm.

Adapter and Quality Trimming of Reads

Files were downloaded and stored in the Purdue University Data Depot. Reads were quality trimmed and adapter sequences were removed using Trimmomatic v. 0.32 (Bolger et al., 2014). Trimmomatic is a program that removes adapter sequences and trims short Illumina reads based on quality. Adapters removed were from the SMARTer kit (Clontech, Mountain View, CA) as well as the Nextera kit (Illumina, San Diego, CA). FastQC v. 0.11.2 (Andrews, 2010) was run in order to observe data quality both before and after quality trimming/adapter removal. FastX-Toolkit v. 0.0.13.2 (Gordon, 2009) quality trimmer was used to further trim reads based on quality score and FASTX-Toolkit quality chart was used to make read per-base quality plots. A FastX trimscore of 30 (the minimum quality score for trimming reads was 30) and a trim length of 50 (reads shorter than 50 bases will be discarded) were used. Maximum length was set to 151 bases. All plots were checked to ensure the reads that would be used in the remainder of the analysis were of high quality and that there were no obvious problems.

Aligning Reads to the Reference Genome and Counting Reads

Before any statistical analysis can be performed, reads resulting from sequencing must be aligned to a reference in order to quantify relative amounts of genes/transcripts. Tophat2 (Kim et al., 2013; Trapnell et al., 2009) was used to align reads to the reference genome. Tophat2 was run with defaults except that the number of mismatches allowed was 1 and due to the non-strand-specificity of the library, the library-type was set to "fr-unstranded". Upon careful consideration, it was decided that aligning the reads to a reference transcriptome would not be ideal. Aligning reads to a reference transcriptome is

very useful when there is either no reference genome or when the reference genome is of poor quality, however this is not the case for human data. Likewise an analysis to identify differential splicing was not performed because the depth of sequencing would not permit identification of alternatively spliced isoforms.

The htseq-count script in HTSeq v.0.6.1 (Anders et al., 2015) was run to count the number of reads mapping to each gene. HTSeq used Biopython v.2.7.3 in the analysis. In order to test parameters and determine the best method for aligning reads to the genome, on Batch1 of the data, HTSeq was run three different ways. HTSeq was run once using the human Ensembl GTF file with rRNA genes removed on “intersection-nonempty” mode, HTSeq was run once with using the entire Ensembl GTF file on “union” mode, and was run once using the Ensembl GTF file with rRNA genes removed on “union” mode. Subsequently, HTSeq was run using the Ensembl GTF file with rRNA genes removed on “intersection-nonempty” mode. The HTSeq modes specify how to handle reads that overlap with more than one gene (Figure 1). The HTSeq feature was set to “exon” (this specifies which feature from the

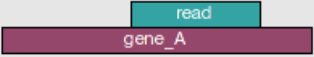
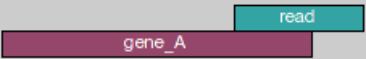

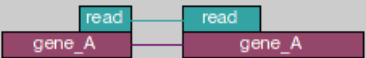
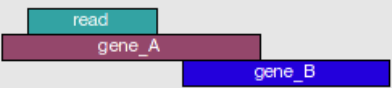

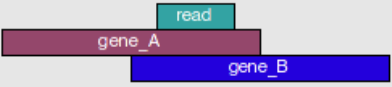
	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Figure 1. HTSeq modes for dealing with reads that overlap more than one feature. Image was taken from the HTSeq manual. This image details the way in which a read is scored depending on the feature(s) the read maps to and the mode in which HTSeq is run.

GTF file is to be used). The HTSeq attribute parameter was set to “gene_id”, which specifies that the Ensembl gene IDs are to be used as rownames in the count files. The --stranded=no option was set because the RNA-Seq libraries prepared were not strand-specific. Once the pipeline was completely done running, all error and output files were checked to ensure that everything ran as expected. The Purdue Bioinformatics Core’s RNA-Seq pipeline was used in trimming reads, running fastQC, running

Tophat, and running HTSeq. The Samtools (Li et al., 2009) flagstat command was run to generate read alignment statistics for each BAM file resulting from the alignment of reads to the reference genome.

Gene Fusions were Found using Tophat-Fusion and FusionCatcher

FusionCatcher (Nicorici et al., 2014) was run using default parameters to search for possible fusion genes. FusionCatcher has a relatively stringent filtering stage in which it runs through the potential fusions and filters out those that are likely to be false positives (such as genes that are overlapping, genes that are known false positive due to the inability to biologically validate, genes that are on mitochondrial chromosomes, or genes in which one gene is the other gene's pseudogene). Reads show support for a gene fusion if a read maps on an exon-exon junction belonging to a candidate fusion gene or if it maps to one of the genes forming a candidate and the other mate maps to the other gene of the candidate fusion (Figure 2).

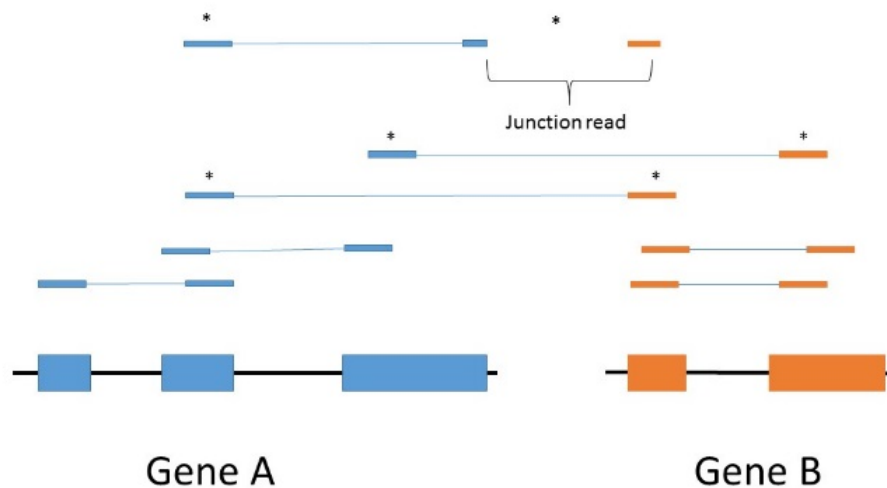


Figure 2. Starred reads are spanning read pairs, which give support to the fusion of genes A and B. The junction read supports the fusion of A and B and aligns to a splice junction. In order for a putative fusion to be reported, both reads mapping to an exon-exon junction belonging to candidate fusion genes and reads in which one mate maps to one of the genes forming a candidate and the other mate maps to the other gene of the candidate fusion must be present. At least two junction reads and at least 3 spanning read pairs must be present for a putative fusion to be identified.

Pathway analysis was performed with IPA

Differentially expressed genes identified using edgeR were uploaded, along with associated false discovery rate, log fold-changes and logCPM (log counts per million), into Ingenuity IPA software and a network analysis was performed (IPA, QIAGEN Redwood City, www.qiagen.com/ingenuity). CPM, or counts per million mapped reads are counts scaled by the number of reads you sequenced, multiplied

by one million. An upstream regulator analysis, mechanistic networks analysis, causal network analysis, and downstream effects analysis were performed using IPA. Top overrepresented canonical pathways were also identified within IPA. Default settings were used and results were filtered based on p-value (results were deemed significant if the $p\text{-value} < 0.01$). The upstream regulator analysis identifies potential upstream regulators that are connected to the dataset genes (the differentially expressed genes) either directly or indirectly. This allows for the identification of molecules that are associated with the observed genes expression changes in the data. The mechanistic networks analysis builds networks based on the putative upstream regulators by connecting regulators that are likely to be involved in the same signaling pathways or involved in the same processes. The causal effects analysis connects upstream regulators to the differentially expressed genes both directly and also indirectly by adding intermediate regulators which may be involved in the networks. Finally, the downstream effects analysis identifies biological functions and diseases that are downstream of the differentially expressed genes and, where possible, predicts whether these functions are likely to be up-regulated or down-regulated as the result of SULT2B1b knock-down. Algorithms and details of each type of network analysis are presented in (Kramer, Green, Pollard, & Tugendreich, 2014).

Results

Quality Control Filtering and Data Visualization of FASTQ Files Shows that Data is of High Quality

Unfortunately in the S1 (SULT2B1b knock-down batch 1) samples, libraries G12, H12, and ctrl-S1 (the batch control sample for S1) have very few reads. Likely a problem occurred during library preparation because all three failed samples were in the same column on the 96 well plate in which the libraries were constructed. The Genomics Center ran these cDNA samples again during the Batch 2 sequencing run; G12 and H12 were successfully sequenced, however ctrl-S1 failed again. All but one (F6) of the control C2 samples were successfully sequenced, and only four of the S2 knock-down samples failed (A6, 2b-tube-control, H12, H6)). All S3 and C3 samples were successfully sequenced. FastQC (Andrews, 2010) is a program that provides simple visual quality checks for data. FastQC was originally developed for DNA sequencing data and thus not all the plots are useful for RNA-Seq analysis. The FastQC output that is the most useful for RNA-seq data are the per base sequence quality graphs (Figure 2), the per sequence quality score graphs (Figure 3), and a list given for each FASTQ files of overrepresented sequences. All FastQC output was reviewed and the data looked to be of high quality. The average number of reads post-trimming, number of cells sequenced, and read numbers (both before and after trimming) are shown in Table 1. Due to the high quality of the sequencing data, 85.39% of reads (1,705,636,658/1,997,518,356) remain after trimming.

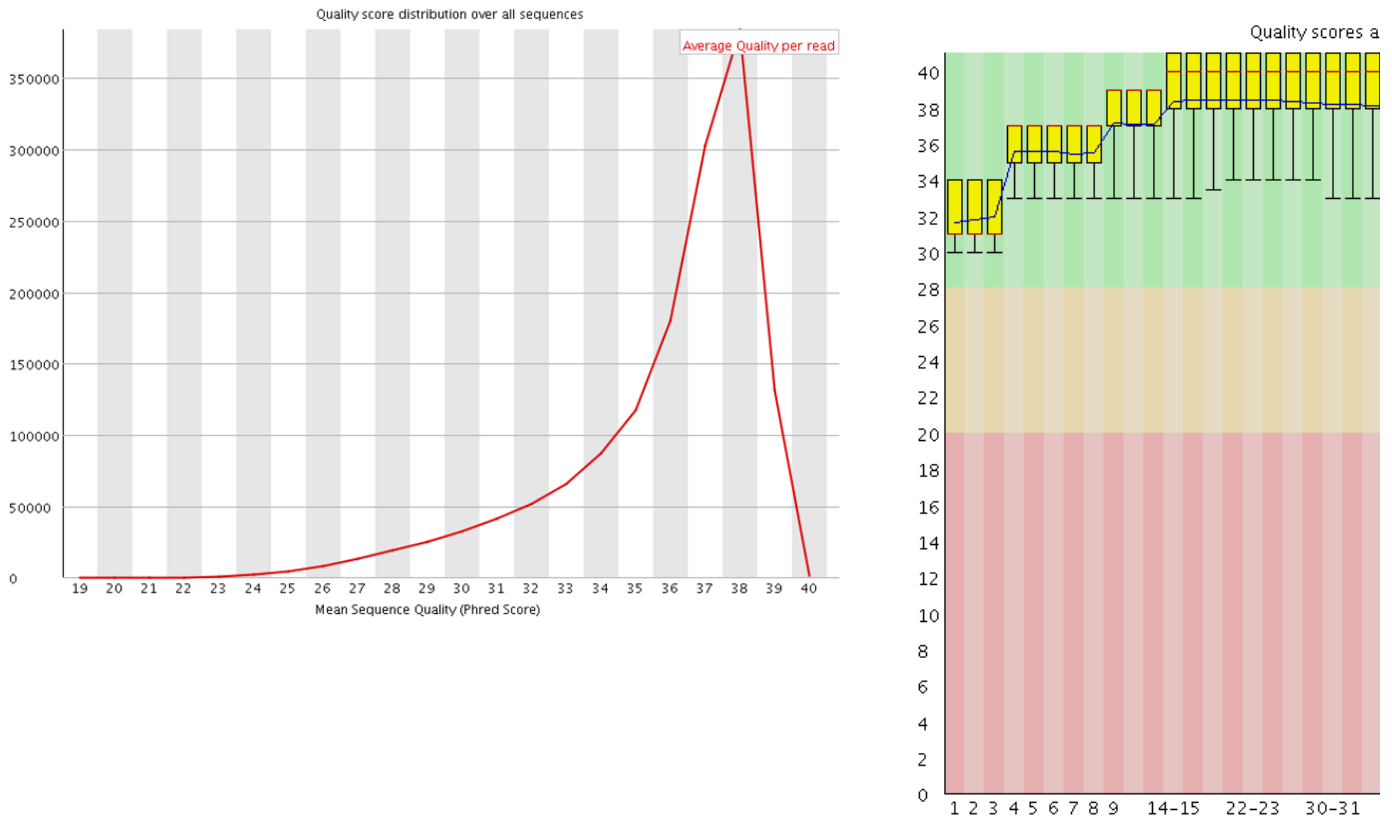


Figure 2. Per base sequence quality for C1-A10 trimmed right reads. The y-axis shows quality scores and the x-axis shows the position in the read. The red line is the median and the blue line is the mean. The yellow box represents the inter-quartile (25%-75%) range, and the upper and lower whiskers represent the 10% and 90% points. The dip in sequence quality at the end of the read is normal and is nearly always observed in Illumina sequencing data. Due in part to trimming, all positions in the reads have high quality. This graph is typical for what was observed across all read files.

Figure 3. Per sequence quality score for C1-A10 trimmed right reads. This plot shows allows identification of samples in which a subset of sequences in the FASTQ file have universally low quality scores. The Y axis shows the number of sequences and the x axis shows the mean quality score. This plot shows that the reads post-trimming have very high quality. This plot looks similar to all the other per sequence quality score plots seen across the trimmed FASTQ files.

Table 1. Read and cell numbers for each condition and batch of data. “C” designates control datasets and “S” designates SULT2B1b knock-down samples. The number designates the batch number. “C1 Redo” shows the read and cell numbers for the batch 1 control samples that failed in the first sequencing run and were resequenced during the second sequencing run. Total number of cells sequenced lists the numbers of cells for which sequencing was successful and does not include cells that were captured from the C1 but were not successfully sequenced.

	C1	S1	C1 Redo	C2	S2	C3	S3
Starting Reads	344,982,788	323,562,478	10,200,012	322,057,980	313,393,742	335,869,862	347,451,494
Total Reads After Trimmomatic	302,800,144	282,552,262	4,442,722	272,385,984	269,183,966	299,192,938	311,809,650
Total Reads After FastX	298,456,776	278,125,063	3,897,830	265,882,007	264,094,088	290,550,640	304,630,254
Ave No. Reads/Sample Post Trim	3,876,062	4,345,704	4,365,458	3,968,388	4,062,986	4,272,804	4,913,391
Total No. Cells Sequenced	77	64	2	67	65	68	62

90% of Reads Map to 36,136 Features

TopHat2 (Kim et al., 2013) was successfully run on all samples and overall 1,522,074,914 reads were mapped out of the 1,693,999,386 input reads (~90%). Subsequently, HTSeq (Anders et al., 2015) was run to count reads aligning to features (genes).

Ribosomal RNA (rRNA) must be removed from samples prior to performing RNA-Seq. If no RNA selection is done prior to library preparation, up to ~95% of the reads can be expected to map to rRNA. For the current experiment, PolyA selection was performed, and therefore polyadenylated RNA were selected. This practice excludes the majority of rRNA. Unfortunately polyA selection is not 100% efficient and so there is always some rRNA contamination in sequencing reads. In this experiment, the polyA selection seems to have worked very well, and few samples used in the analysis had greater than 10% rRNA contamination (Table 2). Since reads were being aligned to a reference genome, the presence of rRNA reads is not a huge concern. However it is plausible that the uneven distribution of rRNA reads amongst the libraries could throw off the statistical analysis. Therefore steps should be taken to remove rRNA prior to statistical analysis. Thus the rRNA genes were removed from the GTF file prior to running HTSeq (Anders et al., 2015). The count matrix generated from the GTF file without rRNA will not have rRNA genes in it and will have the rRNA reads counted in the “Map to No Feature” Category.

Table 2. Number of samples with greater than 10% rRNA contamination. For each of the three batches sequenced, the number of control samples (#C samples), the number of SULT2B1b knock-down samples (#S samples), and the total number of samples with greater than 10% of reads mapping to rRNA is listed, along with the total number of samples sequenced.

	Batch 1	Batch 2	Batch 3
#C Samples >10% rRNA	15	1	2
#S samples >10% rRNA	6	0	2
Total Samples >10% rRNA	21	1	4

Total Samples	141	132	130
----------------------	-----	-----	-----

HTSeq (Anders et al., 2015) was run in two modes, union and intersection-nonempty (see Figure 1 for details on HTSeq modes). Although the “union” default mode is the mode recommended for use, it throws away all reads mapping to two features, thus no counts will be obtained for genes that overlap. Indeed, running HTSeq on intersection-nonempty mode on the Batch1 data gains counts for an additional 453 genes. Therefore HTSeq was run on “intersection-nonempty” mode. A matrix was made of all counts and was given to Faye Zheng in Dr. Doerge’s group. Faye performed a statistical analysis to find differentially expressed genes.

FusionCatcher Identified Fusion Genes, although with Low Sensitivity

Chromosomal rearrangements resulting in gene fusions define many oncogenes and the identification of gene fusions could potentially lead to the identification of therapeutic targets. Gene fusions found in cancer sometimes lead to the translation of a fusion protein or an oncogene may even be aberrantly expressed through fusion to the regulatory elements of a different gene (Maher et al., 2009). Paired-end sequencing allows for the detection of putative fusion genes in bulk RNA-seq data. However it was unknown whether fusion genes could be detected from single-cell data due to the relatively low number of sequencing reads generated from single-cell data. It has been suggested though that increasing sequencing depth past ~500,000 reads would yield diminishing returns, thus increasing the sequencing depth in single-cell experiments is unlikely to yield better results (Pollen et al., 2014). FusionCatcher (Nicorici et al., 2014) tends to perform better than other methods for finding fusion genes. A cell line such as the LNCaP cells used in the present study is an excellent system in which to test fusion finding software on single-cell data because we expect to see the same fusions in each cell. Unfortunately, while many putative fusions were found, the overlap of fusions found between cells was small. For example, 2,196 putative gene fusions were found, however the vast majority of these fusions were found in only 1 cell. Only 42 gene fusions are present in more than one cell and only 24 are present in more than 2 cells. It is likely that many of these fusions (or most even) are false positives as fusion finding programs are known to have high false positive rates. However it does appear that some bona fide fusions were found. The most common fusion found, the MIPOL1--DGKB fusion, was found in 42 cells and is an already known fusion in prostate cancer. Additionally, a DANCER-ETV1 and a ERGIC2-COBL1 fusion were found by in the present dataset. ERG fusions are common in prostate cancer as well as ETV1 fusions (Kwok, Liu, Mangel, & Daskal, 2006; Robinson et al., 2015). Circos plots were produced to depict gene fusions (Figure 4). While several oncogenes have been identified as potential fusions, many other fusions often found in prostate cancer cells (Kumar-Sinha, Tomlins, & Chinnaiyan, 2008; Kwok et al., 2006; Robinson et al., 2015) are not present in the FusionCatcher data. It is possible that the read depth is not great enough to detect many fusions. It is also possible that these fusions may simply not be present in the cell line used in this study or that the drop-out of genes often seen in single-cell sequencing studies lead to the software being unable to detect fusions. Regardless though it is clear based on how few of the putative gene-fusions are detected in more than one cell, that with the currently available software and technology, single-cell sequencing at the current depth is not sufficient to be able to find gene fusions with confidence. Additionally, it seems unlikely that changing FusionCatcher parameters will aid in better identification of fusion genes in the case of the single-cell

data, especially seeing as both the sensitivity and specificity are likely not good when handling single-cell data. Specifically, it is highly likely that many or even most of the putative fusions identified are false positives. We could set the criteria for detecting fusions more stringently and request that more junction reads and spanning read pairs be present than in default settings. However this will almost certainly lead to a loss of sensitivity and already we are not able to detect the same fusions in multiple cells with even a moderate degree of sensitivity.

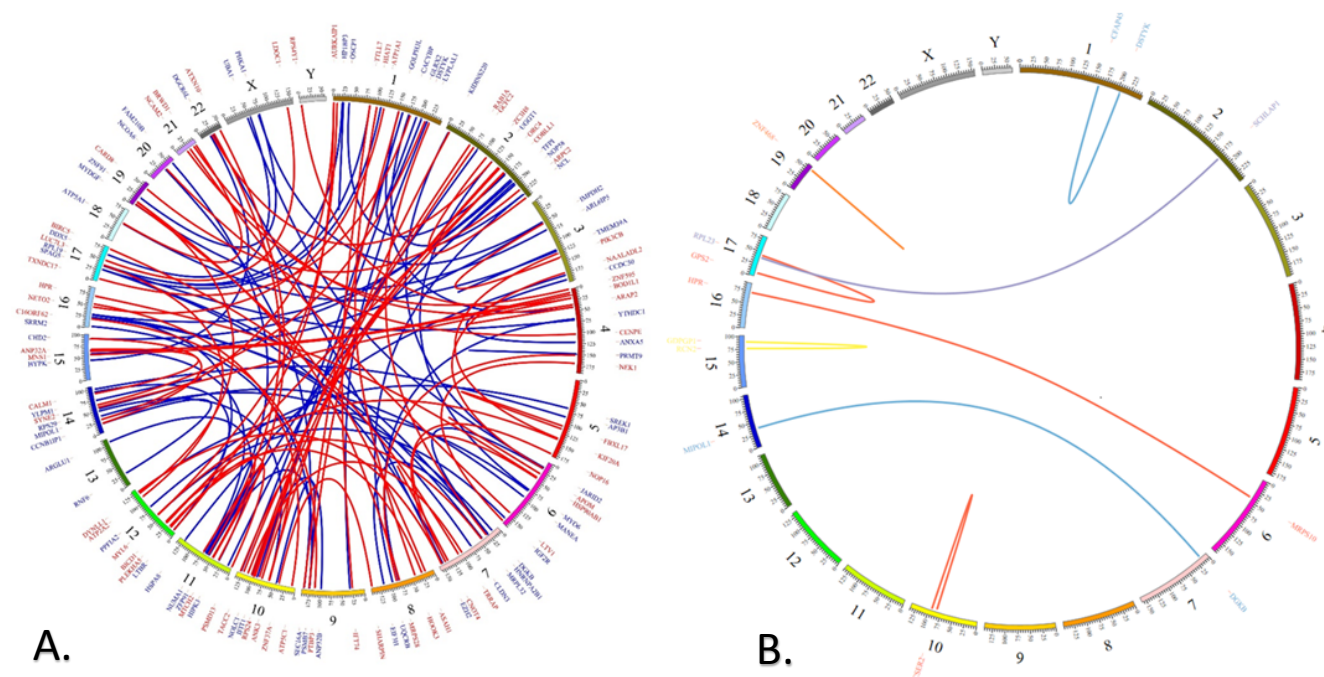


Figure 4. Circos plots for batch 1 data. To enhance readability, only the putative gene fusions for batch1 data are shown. A. All predicted in-frame (red) and out-of-frame (blue) mutations. Only in-frame and out-of-frame predicted fusions were shown to enhance readability. B. All fusions found in more than one cell. Predicted in-frame fusions are shown in red, out-of-frame in blue, CDS(truncated)/UTR in light blue, intronic/CDS(truncated) in light orange, CDS(complete)/UTR in grey, UTR/UTR in yellow, and CDS(truncated)/CDS(no-known-start-or-end) in orange.

A Pathway Analysis Finds Overrepresented Canonical Pathways and Provides Testable Hypotheses for the Future

Figure 5 shows top hits for canonical pathways overrepresented amongst the differentially expressed genes. The molecules in these pathways do not change based on data input, however molecules that are encoded by differentially expressed genes are highlighted. Significance is calculated based on the number of genes/molecules that map to a biological function, pathway, or network. The Fisher's Exact Test was used to test the null hypothesis that the proportion of differentially expressed genes mapping to the pathway are similar to the proportion of genes in the pathway that map in the entire population (the reference set). If the proportions are similar, then there is no significant biological effect. Figure 6 shows the relationship and overlap between the overrepresented canonical pathways. A number of

cancer-related pathways are overrepresented based on the differentially expressed genes, including the prostate cancer signaling pathway (Figure 7). The results of a recent large-scale, multi-institutional genomics project found a number of pathways that were found to be altered in metastatic castration resistant prostate cancer (mCRPC) (Robinson et al., 2015). The results of this pathway analysis show significant overlap with many of the pathways found to be significantly overrepresented amongst the present data. The Robinson et al. study found that the androgen receptor pathway, PI3K pathway, WNT pathway, DNA repair pathway, and cell-cycle pathway were all commonly altered. With the exception of the WNT pathway, our current results show that each of these pathways were overrepresented amongst the differentially expressed genes. As seen in Figures 5 and 6, the following pathways were amongst the overrepresented canonical pathways, thus suggesting that these pathways are affected by SULT2B1b knock down: role of BRCA1 in DNA Damage response, PI3K/AKT signaling, role of CHK proteins in cell cycle check-point control, prostate cancer signaling, and cell cycle: G2/M DNA damage checkpoint regulation.

An upstream regulator analysis was performed, which identifies molecules which could be responsible for the observed gene expression changes in the uploaded data. For all pathway analyses, examples using the prostate cancer signaling pathway or androgen receptor (AR) are shown. However, the diagrams shown in this report can be created for any of the upregulated canonical pathways or putative upstream regulators. Supplemental tables listing all the putative upstream regulators are included with this report. Attached in Table S1 are putative upstream regulators, as predicted by the IPA software. Figure 8 shows the results of the upstream regulator analysis for predicted regulator androgen receptor (AR). AR was a highly significant predicted upstream regulator (p-value 1.34×10^{-12}). The upstream regulator analysis determines likely upstream regulators that are connected to dataset genes through a set of direct or indirect relationships. This analysis also predicts whether the proposed regulators are activated or inhibited based on the gene expression patterns given by the data and then integrates findings and predictions with previous knowledge from peer-reviewed scientific literature. The upstream regulator analysis ultimately takes a gene-expression dataset and attempts to identify the upstream biological causes as well as downstream probable events. For some regulators, it also predicts whether they are activated or inhibited based on the up- and down- regulation of differentially expressed genes and determines which relationships in the literature are relevant for the given data. This can aid in generating mechanistic hypotheses as well as being used to find potential upstream regulators with a response opposite to the observed expression pattern, which can be useful for the prediction of therapeutic compound effects. P-values are from the Fisher's exact test, which uses the overlap of observed and predicted regulated gene sets in determining biological significance. The activation z-score assesses the match of observed and predicted up/down regulation patterns. In this way, the z-score is both a significance measure and a predictor for the state (activated or inhibited) of the putative regulator. The default cutoff for significance is an activation z-score that is more extreme than $|2|$. Note that the activation z-score computed by IPA is not the same as the standard score used in statistics to indicate how many standard deviations an element is from the mean. One-sided Fisher's exact tests were used in the upstream regulator analyses. The top three most significant predicted upstream regulators are the interferon IFN LAMBDA 1 (IFNL1) with a p-value of 1.14×10^{-35} , IFN ALPHA 2 (IFNA2) with a p-value of 1.42×10^{-27} , and the E2F transcription factor 4 (E2F4) with a p-value of 3.86×10^{-27} . The upstream regulatory network of androgen receptor and other regulatory proteins it is predicted to be connected to is shown in Figure 8. The mechanistic network for androgen receptor is shown in Figure 9. It is important to keep in mind that the networks may not be complete due to a lack of

statistical evidence for missing edges or due to the missing data that is common in single-cell sequencing. The mechanistic network analysis is based off the results of the Causal Analysis and thus, similarly to in the Causal Analysis, AR is colored orange, indicating predicted activation, however based on the activation z-score and the default settings in IPA, predictions about how AR is regulated by SULT2B1b are not able to be made.

The Causal Network Analysis, shown in Figure 10, is a generalization of the upstream regulator analysis that formulates hypothesis networks. The networks generated connect putative upstream regulators that are predicted to be part of the same signaling or causal mechanism. The Causal Network analysis predicts that AR is inhibited based on gene expression data. This fits with previous findings from the Ratliff lab, in which qRT-PCR was used to determine that SULT2B1b modulates AR expression, either directly or indirectly, and that knock-down of SULT2B1b leads to a decrease in AR expression levels. Recall that in the Upstream Analysis and Mechanistic Networks analysis, that AR was colored in orange, signifying upregulation by SULT2B1b knock down. However this is somewhat the activation z-score predicting up or down regulation was not significant. Additionally IPA did not predict an activation state (upregulated or downregulated) for AR. Thus, in both of these analyses, while AR is colored as if it is upregulated, there is not sufficient evidence to determine whether it is likely up or downregulated. In the Causal Networks analysis however, the activation z-score is significant and thus AR is predicted to be downregulated by SULT2B1b knock-down. The reason for the ability to predict the activation state of AR in the causal networks analysis but not in the upstream analysis or mechanistic networks analysis has to do with the topology of the network. In the Upstream Analysis, AR has 46 downstream targets (from the uploaded data file) and the regulation observed for these molecules (in conjunction with the regulation observed in published literature) is used to make the prediction inference for AR. Please note that the path between AR and the downstream targets is "one-step" interaction (i.e., there are no intermediate nodes or regulators between AR and the downstream data set molecules). In the Causal Networks, on the other hand, AR (which is identified as the Master Regulator) is connected to 292 downstream targets (from the uploaded data file) via 46 "intermediate molecules" (also called the Participating Molecules) through 1 or 2 or 3 steps. These additional nodes in the causal path now cause a change in the prediction pattern for AR.

A Downstream Effect Analysis was also performed in order to identify the likely impacts of SULT2B1b knock-down on biological functions and diseases that are downstream of the differentially expressed genes. Many of the identified possible downstream effects fit with what we know about the effects of SULT2B1b knock-down on LNCaP cells. The three most significant functions predicted to be increased by knock-down of SULT2B1b are cell death of tumor cell lines ($p\text{-value}=2.68 \times 10^{-24}$), cell death ($p\text{-value}=6.94 \times 10^{-21}$), and necrosis ($p\text{-value}=9.03 \times 10^{-19}$). The three most significant functions predicted to be decreased by the knock-down of SULT2B1b are proliferation of cells ($p\text{-value}=4.32 \times 10^{-17}$), proliferation of tumor cell lines ($p\text{-value}=1.62 \times 10^{-13}$), and replication of viral replicon ($p\text{-value}=6.02 \times 10^{-10}$).

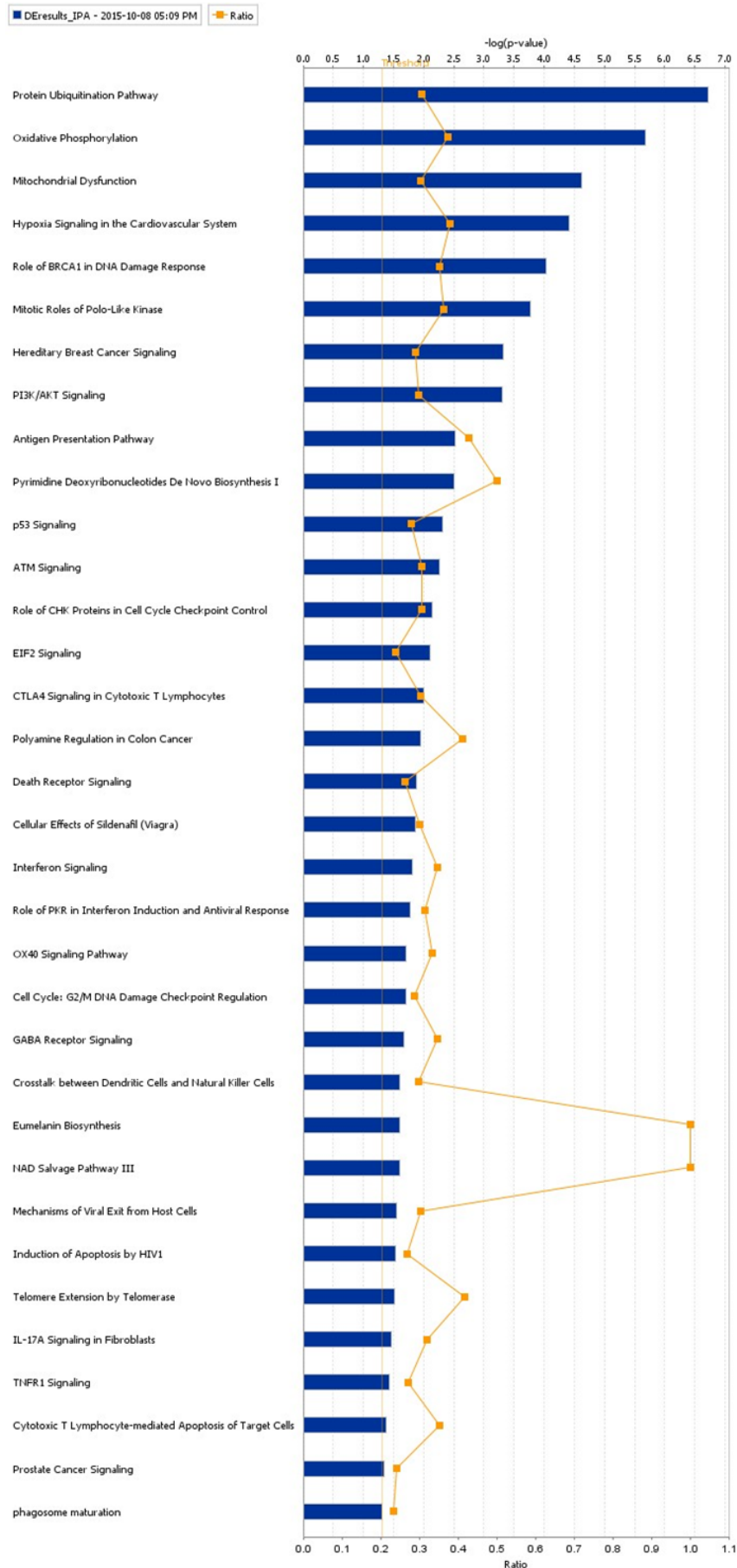


Figure 5. The top hits for canonical pathways amongst the differentially expressed genes is shown. The columns show the $-\log(p\text{-values})$ while the orange points show the ratio of the number of genes which meet cutoff criteria/the number of genes in the pathway. A number of cancer-related pathways are on the list, including the “Prostate Cancer Signaling”, “p53 Signaling”, and “Hereditary Breast Cancer Signaling” pathways.

intensity of the color indicates the degree of down or upregulation. The expression value used in coloring the molecules was the log fold-change.



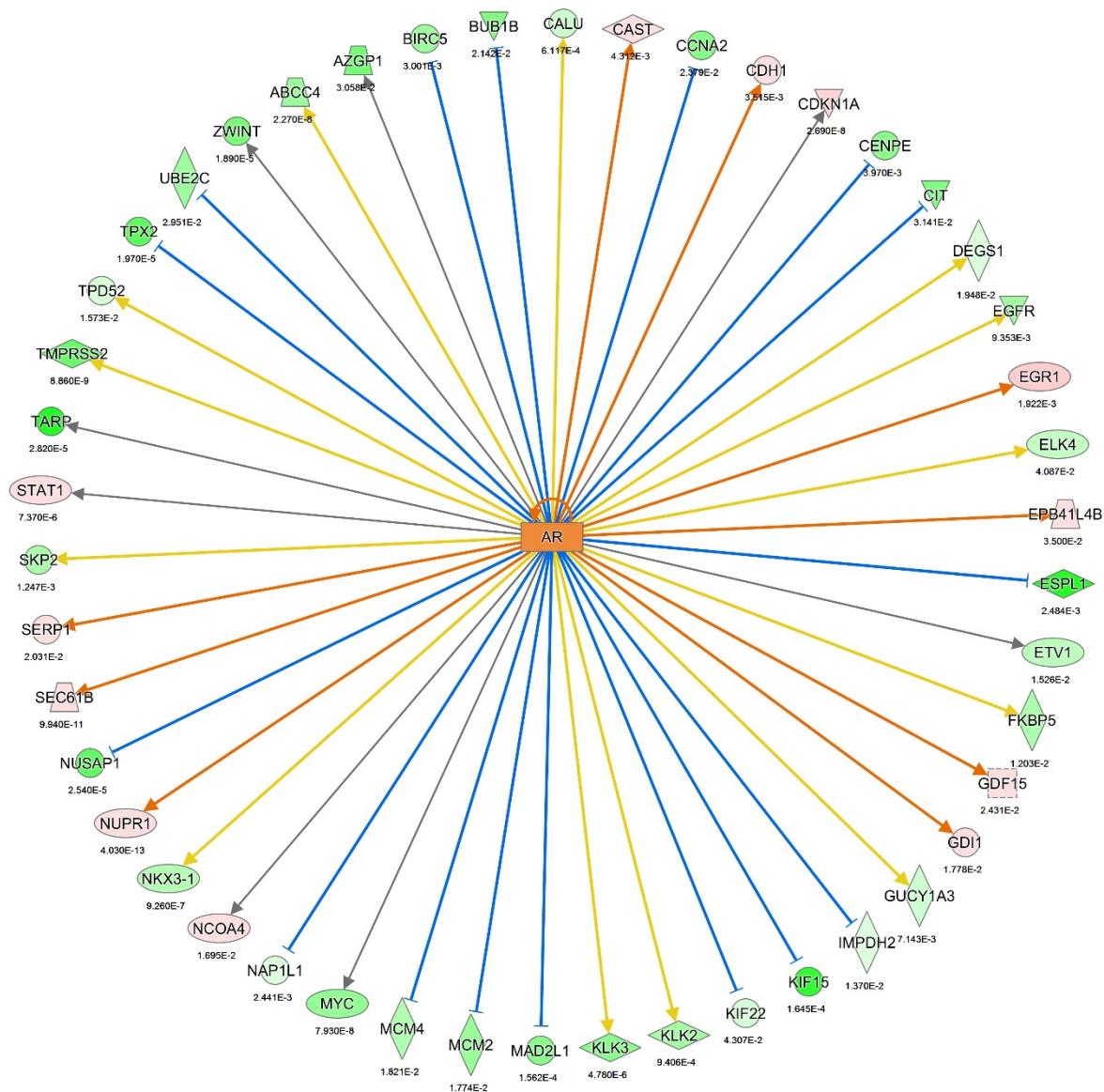


Figure 8. Upstream regulator androgen receptor (AR) and predicted target molecules in the dataset. The list of differentially expressed genes, along with log fold-change and false discovery rates were used as input for the analysis. AR is one of the top predicted upstream regulators, connecting directly to 46 additional putative regulators, shown here. Molecules in red are up-regulated and molecules in green are down-regulated in SULT2B1b knock-down cells. Androgen receptor is shown in orange, which signifies that it is predicted to be upregulated in response to SULT2B1b knock down based on the 46 genes regulated directly by it, however its activation z-score (1.225) does not show significant up or downregulation based on the threshold used of $z=|2|$. Orange edges denote predicted activating

relationships, blue edges predict repressing relationships, grey edges do not have an effect predicted, and yellow edges denote interactions that are inconsistent with the proposed interactions based on the Ingenuity Knowledge Base (peer-reviewed scientific literature) and hence, a prediction of the nature of the relationship cannot be made. Arrowheads on edges indicated activating relationships in the gene expression data and blunt lines at the end of edges indicate inactivating relationships in the gene expression data. All edges are solid lines and are thus predicted to be direct interactions between AR and the target molecules.

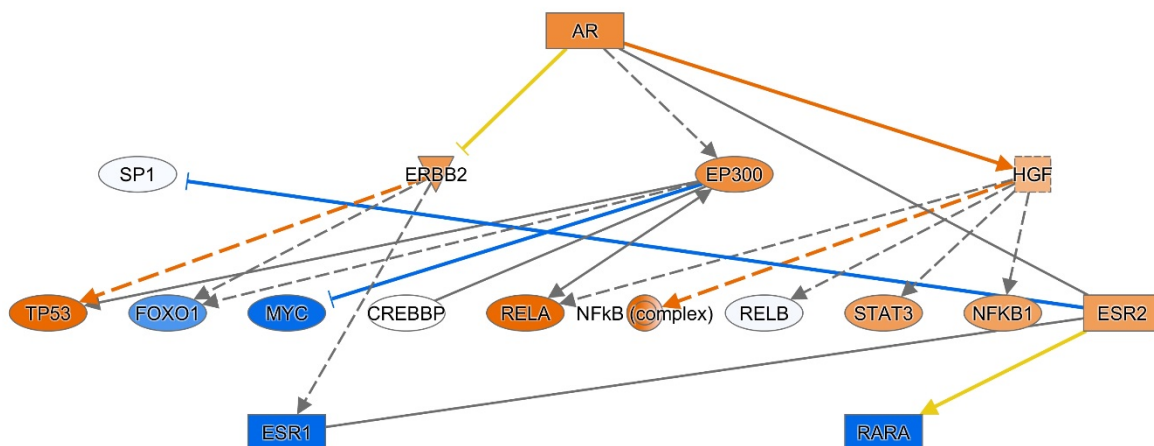


Figure 9. Mechanistic network for androgen receptor (AR). In this network AR is proposed to directly activate EP300 (the transcriptional regulator E1A binding protein p300) and HGF (hepatocyte growth factor). The list of differentially expressed genes, along with log fold-change and false discovery rates were used as input for the analysis. A similar figure and analysis may be performed for any of the putative upstream regulators listed in Supplementary Table S1. Nodes shown in orange are also predicted to be upstream activators and those that are predicted to be inhibitors are shown in blue. As in the Upstream Regulator Analysis, the AR z-activation score is not significant and thus it should be considered to neither be activated nor inactivated. Orange edges show activating relationships and blue show inhibiting relationships. Black edges do not have an effect predicted and yellow edges show literature findings which are inconsistent with the state of the downstream molecule. This mechanistic network determines which network edges between pre-determined upstream regulators are likely relevant for the causal mechanism behind the dataset. The mechanistic network algorithm is run upon the regulators from the upstream regulator analysis results. Solid edges show direct interactions and dashed lines show indirect interactions between molecules.

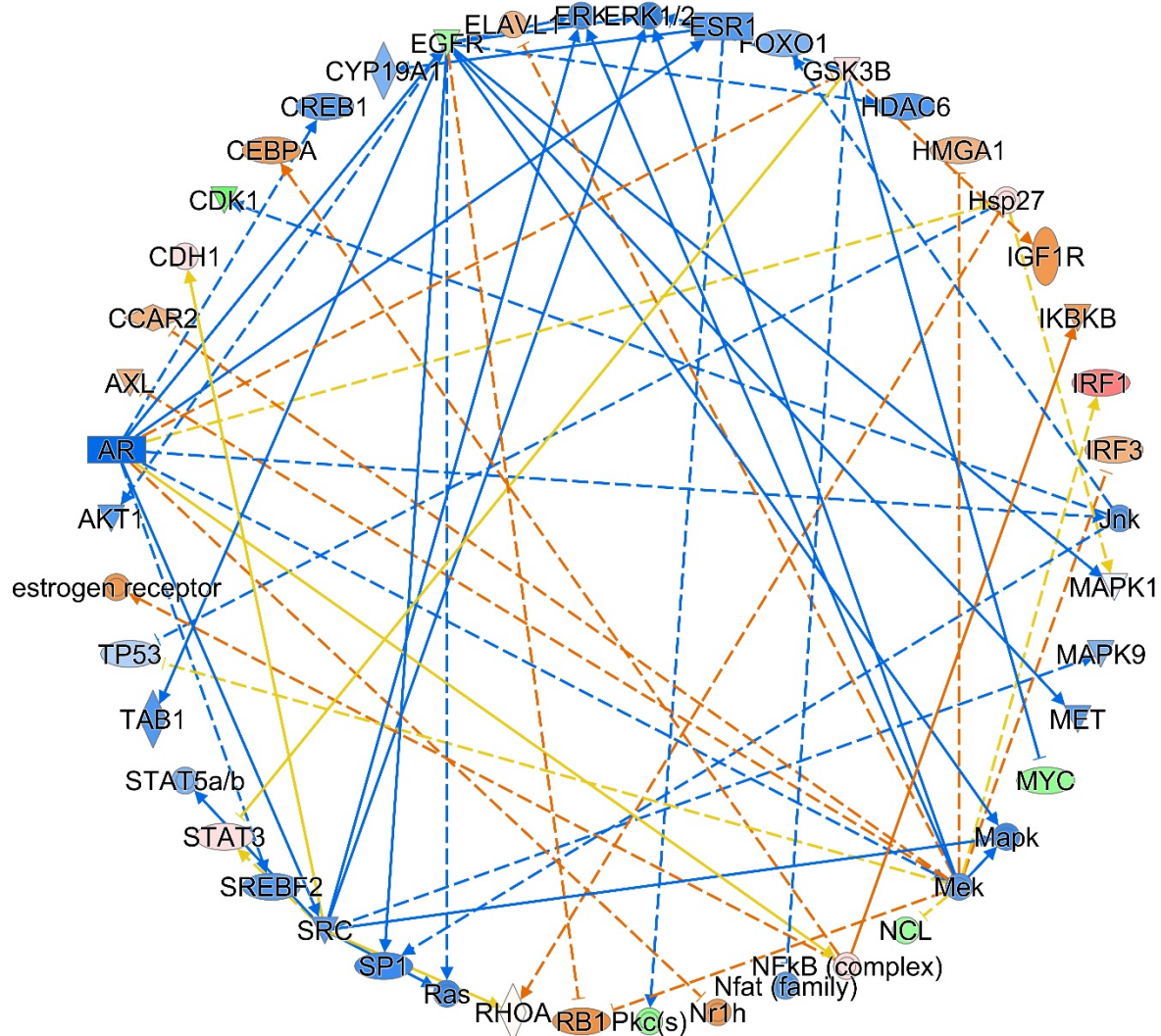


Figure 10. Causal network analysis for androgen receptor (AR). Causal network analysis seeks to generate a more complete visualization of the network by building on the Upstream Regulatory Analysis and connecting regulators to dataset molecules, while taking advantage of paths that can involve intermediate regulators. The list of differentially expressed genes, along with log fold-change and false discovery rates were used as input for the analysis. Blue molecules are significantly inactivated and orange are significantly activated. Molecules in red are up-regulated and molecules in green are down-regulated in SULT2B1b knock-down cells, as observed from the differential expression analysis results.

Summary and Discussion

Overall, the data appears to overall be of high quality. The analysis of these datasets showed that high quality single-cell data can be obtained with the use of the C1 FLuidigm instrument. Overall, the mapability of the reads was excellent and few reads were lost during quality control.

While this data shows that it is possible to identify some fusion genes in single-cell data, with the current sequencing technology and fusion-finding software available, based on this study it is not recommended to use single-cell data to identify fusion genes. Over 2,000 putative gene fusions were found, which is not surprising given that these fusion-finding software packages are known to have high false positive rates. However, even though the transcriptomes of 403 cells were sequenced, few fusions were found in more than 1 cell. Since LNCaP cells are a single line of cells, we expect to be able to identify the same fusions in each cell, however this was not the case. It is likely that the depth of sequencing in single-cell data along with the lack of advanced fusion-finding software packages lead to the lack of sensitivity in identifying fusion genes.

The pathway analysis performed fits with what is known about prostate cancer and also provides testable hypotheses to explore further, both computationally and biologically. A list of putative upstream regulators was identified based on the input list of differentially expressed genes was identified. Several types of network analyses may be done using each of these upstream regulators, as demonstrated using the putative upstream regulator AR as an example. Many of the canonical pathways found to be overrepresented fit with recently published literature by Robinson et al., which identified pathways that are often modified in prostate cancer (Robinson et al., 2015). Overrepresented pathways which fit with the findings of this study include role of BRCA1 in DNA Damage response, PI3K/AKT signaling, role of CHK proteins in cell cycle check-point control, prostate cancer signaling, and cell cycle: G2/M DNA damage checkpoint regulation. The WNT pathway was the only pathway identified by Robinson et al. which was not represented amongst the overrepresented canonical pathways identified by IPA. The Mechanistic Networks analysis and the more complete networks generated by the Causal Network Analysis allow for predictions to be made about how SULT2B1b may be modulating downstream genes through the predicted upstream regulators, such as AR. Finally, the Downstream Effect Analysis identified a number of biological processes and even diseases which based on the differentially expressed genes, may be affected by SULT2B1b knock-down.

Overall, this project shows that successful single-cell isolation and sequencing can be performed using the C1 Fluidigm in combination with Next Generation Sequencing. This data has allowed the identification of genes which are differentially expressed in SULT2B1b knock-down cells. These genes, along with the predicted targets and mechanisms obtained from the network analysis provide a starting place for future experiments and a set of testable hypotheses. Additionally the network analysis has provided clear evidence linking SULT2B1b with prostate cancer.

Supplemental Legends

Table S1 lists the results of the Upstream Regulator analysis performed using Ingenuity IPA software to find putative upstream regulators that could potentially explain the differential expression results. Each row shows information about a predicted regulator and the molecules in the dataset that it targets. Putative upstream regulators are listed, along with the logFC found by Faye using edgeR, the Molecule type, predicted activation state (is the molecule predicted to be activated, inhibited, or in some cases no

state can be predicted), the activation z-score (both a measure of significance and also takes into account predicted activation state), p-value of overlap (from a one-sided Fisher's exact test), a list of the target molecules in the list of differentially expressed genes, and the number of genes in the mechanistic network. The number of genes in the mechanistic network provides the number of genes that each putative regulator connects to, with the number of genes that are directly connected to the regulator listed in parentheses. Note that not all of the genes predicted to be upstream regulators are found amongst the differentially expressed genes. Genes that have no log fold-change are predicted to be upstream regulators based on differentially expressed genes, but are not themselves differentially expressed.

Table S2 lists the results of the Downstream Effects Analysis. Each row in the table provides the functional category, the associated disease or function, the Fisher's exact test p-value, the predicted effect on the function (increasing or decreasing), the activation z-score, the molecules associated with the functional category, the number of molecules associated with this category, and whether the function relates to general biological functions/disease or toxicity.

References

- Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2), 166-169. doi: 10.1093/bioinformatics/btu638
- Andrews, S. (2010). FastQC.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114-2120. doi: 10.1093/bioinformatics/btu170
- Ding, B., Zheng, L., Zhu, Y., Li, N., Jia, H., Ai, R., . . . Wang, W. (2015). Normalization and noise reduction for single cell RNA-seq experiments. *Bioinformatics*, 31(13), 2225-2227. doi: 10.1093/bioinformatics/btv122
- Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16), 3439-3440. doi: 10.1093/bioinformatics/bti525
- Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc*, 4(8), 1184-1191. doi: 10.1038/nprot.2009.97
- Eberlin, L. S., Dill, A. L., Costa, A. B., Iza, D. R., Cheng, L., Masterson, T., . . . Cooks, R. G. (2010). Cholesterol sulfate imaging in human prostate cancer tissue by desorption electrospray ionization mass spectrometry. *Anal Chem*, 82(9), 3430-3434. doi: 10.1021/ac9029482
- Gordon, A. (2009). FastX-Toolkit. Retrieved from http://hannonlab.cshl.edu/fastx_toolkit/download.html
- Higashi, Y., Fuda, H., Yanai, H., Lee, Y., Fukushige, T., Kanzaki, T., & Strott, C. A. (2004). Expression of cholesterol sulfotransferase (SULT2B1b) in human skin and primary cultures of human epidermal keratinocytes. *J Invest Dermatol*, 122(5), 1207-1213. doi: 10.1111/j.0022-202X.2004.22416.x
- Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, 14(4), R36. doi: 10.1186/gb-2013-14-4-r36
- Kramer, A., Green, J., Pollard, J., Jr., & Tugendreich, S. (2014). Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics*, 30(4), 523-530. doi: 10.1093/bioinformatics/btt703

- Kumar-Sinha, C., Tomlins, S. A., & Chinnaiyan, A. M. (2008). Recurrent gene fusions in prostate cancer. *Nat Rev Cancer*, 8(7), 497-511. doi: 10.1038/nrc2402
- Kwok, S. C., Liu, X., Mangel, P., & Daskal, I. (2006). PTX1(ERGIC2)-VP22 fusion protein upregulates interferon-beta in prostate cancer cell line PC-3. *DNA Cell Biol*, 25(9), 523-529. doi: 10.1089/dna.2006.25.523
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078-2079. doi: 10.1093/bioinformatics/btp352
- Maher, C. A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., Jing, X., . . . Chinnaiyan, A. M. (2009). Transcriptome sequencing to detect gene fusions in cancer. *Nature*, 458(7234), 97-101. doi: 10.1038/nature07638
- Nicorici, D., Satalan, M., Edgren, H., Kangaspeska, S., Murumagi, A., Kallioniemi, O., . . . Kilkku, O. (2014). FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data. *bioRxiv*. doi: <http://dx.doi.org/10.1101/011650>
- Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., Leyrat, A. A., Lui, J. H., . . . West, J. A. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat Biotechnol*, 32(10), 1053-1058. doi: 10.1038/nbt.2967
- Robinson, D., Van Allen, E. M., Wu, Y. M., Schultz, N., Lonigro, R. J., Mosquera, J. M., . . . Chinnaiyan, A. M. (2015). Integrative clinical genomics of advanced prostate cancer. *Cell*, 161(5), 1215-1228. doi: 10.1016/j.cell.2015.05.001
- Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., . . . Kasprzyk, A. (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res*, 43(W1), W589-598. doi: 10.1093/nar/gkv350
- Stegle, O., Teichmann, S. A., & Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet*, 16(3), 133-145. doi: 10.1038/nrg3833
- Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9), 1105-1111. doi: 10.1093/bioinformatics/btp120