A systematic reprocessing of RNA-Seq data generated with stranded and non-stranded sequencing protocols

Executive summary

The purpose of this report of the RNA-seq investigation is to evaluate the effect of integrating RNA-seq data generated with both stranded and unstranded protocols, and investigate the best approaches for adjusting for any batch effects due to differences in data generation.

Key Findings:

1. Given the long-running nature of the DepMap project, we have generated data using two RNA-sequencing protocols as the technology has evolved data, one from a stranded RNA-seq protocol (the Illumina Tru-Seq Strand Specific Large Insert RNA Sequencing (50M pairs) v1), another from an unstranded RNA-seq protocol (the Illumina TruSeq RNA Sequencing). To date, data generated from either protocol was processed through the same bioinformatic pipeline, and the stranded information for data generated from the stranded protocol was not considered in the calculation of gene expression values.

2. This difference in protocol accounts for 6.77% of the variation in the current gene expression data.

3. After reprocessing the data generated with the stranded protocol, we have found that the global gene expression differences between the unstranded RSEM and stranded RSEM pipeline settings are largely due to the transcript overlaps with local antisense transcripts as expected, and none of them are related to cancer driver genes.

4. To account for an observed batch effect between data processed on the different protocols and pipelines, we evaluated the effect of adjusting for this batch on downstream analysis. After batch effect correction, we observed consistent clustering by cell line lineage and high overlap in important features associated with dependency in our predictive pipeline. These results suggest that this updated dataset recapitulates past results and the effects of this update should be minor.

Our conclusions are that this batch effect has relatively minor impact to the target discovery analysis for the DepMap portal, and updating the approach for integrating data across experimental protocols will improve the dataset quality.
Introduction & Motivations

DepMap’s RNAseq data contains a mixture of stranded and non-stranded sequencing protocols generated over a 10+ year long period. Our pipeline uses RSEM to estimate gene and isoform expression levels from RNAseq data [1] (Li and Dewey 2011). In particular, our pipeline was processing stranded samples in the unstranded mode of RSEM. This leads to a loss of information on the strandedness of the samples and may lead to an overestimation of the expression levels in genes with overlapping antisense transcripts. Thus, we reprocessed the data generated with a stranded protocol in the stranded mode of RSEM. In evaluating the impact of the changes, we asked the following questions:

1) Is there a batch effect between data generated with the two protocols? If so, how do we account for the batch effect present between the unstranded data and stranded data in the dataset?
2) What impact does reprocessing the stranded data in stranded mode of RSEM have on the final gene expression values?
3) What effects does a) reprocessing stranded data in stranded mode of RSEM and b) the batch effect correction of gene expression have on the gene expression levels and predictive performance of CRISPR gene effect scores?

Methods

There are multiple possible models we could look at based on changing the RSEM mode for processing the stranded samples and whether or not batch correction is applied:

Figure 1: A diagram showing the RNAseq experimental protocol and the parameter settings set for the RSEM tool during data processing.

Models including in the 24Q2 release:
1) Current gene expression dataset including stranded RNA-seq data processed in unstranded mode and unstranded RNA-seq data processed in unstranded mode. *(24Q2/CURRENT RELEASE AND PAST RELEASES)*

2) batch corrected version of gene expression 1a + 2b: stranded RNAseq samples processed in stranded mode and unstranded RNA-seq data processed in unstranded mode. This version has the correct parameter settings, and batch corrected using the Combat. *(24Q2/CURRENT RELEASE AND FUTURE RELEASES)*

Note that we are still including the original expression dataset in this release (24Q2), since our analysis suggests fairly minor changes to downstream analysis with our updated dataset. This will allow users to maintain consistency with past releases and to bridge to future releases. In the future, we anticipate only releasing the integrated data set as the batch corrected version. We will continue to release the pre-batch corrected stranded and unstranded gene expression data as separate files.

---

**Results**

**Evaluating Batch Effects**

To determine whether there exists a batch effect between the stranded and unstranded data in the current dataset, we performed principal component analysis (Figure 2). The purple dots correspond to stranded samples processed in an unstranded mode of RSEM; the red dots correspond to unstranded samples processed in an unstranded mode of RSEM; and finally, the green dots correspond to stranded samples processed in a stranded mode of RSEM. The original dataset (1b+2b) had a batch effect as shown in the distance between the clusters of purple and red dots in the leftmost panel of PCA plots. The stranded/unstranded difference was explained primarily by PC3 accounting for 6.82% of the protein coding gene expression variance in the original dataset. When the stranded data is processed with the stranded mode of RSEM, this PC accounts for 6.77% of the protein coding gene expression variance.
Correcting the Batch Effect

To do batch correction, we used ComBat [2, 3] (Behdenna et al. 2023). After applying the batch correction method to the new dataset, the distances in the PCA spaces between the cell lines from the two different batches have been sharply reduced, the explainable variance of the third
component drops to 6.13%, and we conclude that the linear batch correction method addresses the dominant batch effect from difference in protocol and pipeline.

Figure 4: PCA after batch correction of the protein coding expression for all cell line data, color coded by whether the sample is stranded or unstranded and whether the sample is processed in the stranded RSEM mode. X and y axis shows the percentage of explainable variance for each principal component. 1 means strand-specific RNA-seq data processed with strand RSEM mode, 0 means non-strand-specific RNA-seq data processed with non-stranded RSEM mode.

**Quantifying the Impact of Changes in Gene Expression**

We quantify the changes for data generated on the stranded protocol processed with the original pipeline and with the stranded-aware pipeline. We calculated log fold change as follows.

\[
log_2 FC = \exp(A) - \exp(B)
\]

where \(A\) and \(B\) are different gene expression matrices. For our purposes, we set \(A\) to the expression matrix of 1a (stranded RNA-seq samples processed in the stranded mode of RSEM) and \(B\) to the expression matrix of 1b (stranded RNA-seq samples processed in the unstranded mode of RSEM). We used the formula above to quantify the mean log2 fold change between the stranded RSEM and non-stranded RSEM settings across all the strand-specific RNA-seq data.

The results are shown in the waterfall plot below:
Figure 5: A waterfall plot of the log2 fold change between Model 3 (1a+2b) and Model 1 (1b+2b) to quantify the effect of processing the stranded RNAseq samples in the stranded mode of RSEM (a list of the genes in both tails has been provided below).

We divide the outlier genes into two groups using a log2 fold change threshold of $|\log_2(1.5)| \sim \pm 0.58$. In the first group, there are 7 genes that have $\log FC > 0.58$. In the second group, there are 23 genes that have $\log FC < -0.58$. None of these outlier genes are considered oncogenes or tumor suppressor genes when we looked for them in OncoKB [4] (Chakravarty et al. 2017). Here are the list of genes:

Genes with $\log FC > 0.58$ (higher in stranded processed data):

DHRS13 (147015), FEM1A (55527), MGAT2 (4247), NUDT3 (11165), SCO2 (9997), SFT2D3 (84826), SIX5 (147912)

Genes with $\log FC < -0.58$ (lower in stranded processed data):

ARHGAP45 (23526), CD37 (951), CEMP1 (752014), CFAP126 (257177), CRLF1 (9244), CRMP1 (1400), CYP3A5 (1577), DDTL (100037417), DENND1C (79958), ENTPD1 (953), IL17D (53342), LDLRAD2 (401944), LHFPL5 (222662), LRRC37A2 (474170), MT-ND6 (4541), MXRA8 (54587),
Necab1 (64168), Rps10-NUDT3 (100529239), Slc19A1 (6573), Slc51A (200931), Tmem140 (55281), Tssk4 (283629), Zacn (353174)

Of the 30 genes in the lists above, 27 (90%) have overlapping antisense transcripts. Thus, the list of genes with the largest differences in estimated expression when processed in a stranded-aware pipeline is relatively short and does not contain any known cancer associated gene signatures (Figure 6).

Figure 6: TCGA gene expression signature analysis across all the significantly differential expressed genes between stranded and non-stranded specific RNA-seq experiments. Each column represents one cancer type in TCGA, each row represents one gene.

**Impacts on the Predictability Pipeline**

We observe high concordance in the performance of the predictive models between the original and updated data sets (Pearson r=0.97) (Figure 5). The most variable genes in the CRISPR predictive models originated from the genes with lowly predicted CRISPR gene effect scores from both pairs of the gene expression combinations.
Figure 7: CRISPR gene effect predictive performance comparison between two sets of the models. Left-panel: x axis shows the original data set with unstranded RSEMs regardless of the sequencing protocols, y axis; right-panel: batch corrected version for the benchmarking.

If we only focus on the CRISPR genes with predictability, >0.5 Pearson correlation, then we have 72% of consensus predictive biomarkers (433 genes CRISPR screening with 3115 predictive genomic features in total) from the random forest models (Ho 1995) for the batch corrected original expression (bc(1b+2b) overlap with bc(1a+2b)) and non-batch-corrected stranded RSEM expression and unstranded RSEM (1b+2b overlap with 1a+2b) suggesting general concordance of predictability models generated with the two datasets.

**Feature Importance Thresholds**

To understand the differences in predictive models generated on the batch-corrected vs pre-batch corrected data, we compared the overlap in identified features above a feature importance threshold of 0.1. For gene expression features, we found that 42.2% overlapped between strand-specific RNA-seq and un-stranded RNA-seq with unstranded settings and increased to 85% between un-stranded RNA-seq with unstranded settings plus strand-specific RNA-seq with stranded settings, from 67.5% to 95.2% for the batch corrected version. Damaging and hotspot mutations account increases from 81.8% to 100% for both non-batch corrected and batch corrected versions. Copy number feature proportions increase from 73.8% in unstranded settings to 91.9% in stranded settings, and from 77.2% to 90.6% in the batch corrected version (Figure. 6). This shows the top predictive features for models based on different versions of gene expression data release are largely overlapped after thresholding the feature importance, and indicates that the prioritized drug target should be stable in comparison to the past release.
Figure 8: Overlapped features for each category of genomic features across the four sets of predictive models. Right panel is the 23Q4 gene expression matrix with unstranded RSEM settings with the stranded RSEM settings, the left-panel is batch corrected version model comparison. Colors represent different thresholds of feature importances. The x axis represents the core genomic features from DepMap, the y axis shows the overlapped features ratio between the two sets of predictive models, which are measured by the overlapped features for each gene in the CRISPR gene effect matrix divided by the minimum features number between the two models.

Correlation to protein abundance

We further evaluated the batch correction effect on correlation with proteomics datasets through pearson correlation analysis across samples and genes, the original and batch corrected gene expression dataset are similar in relation to protein abundance (Table 1). This suggests gene expression correlation with proteomics is stable since it is not affected by the updates in data processing or batch correction.

<table>
<thead>
<tr>
<th></th>
<th>Across samples correlation</th>
<th>Across genes correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olink Best Dilution 24Q2 with Stranded RSEM for stranded RNA-seq and unstranded RSEM for non-stranded RNA-seq</td>
<td>0.515</td>
<td>0.583</td>
</tr>
<tr>
<td>Olink Best Dilution 24Q2 with batch corrected version of above gene expression</td>
<td>0.514</td>
<td>0.577</td>
</tr>
</tbody>
</table>

Table 1. Summary of proteomics correlation with the new gene expression measurement before and after batch correction.
Sources


[4] The human genome browser at UCSC

[5] Suehnholz et al., Cancer Discovery 2023 and Chakravarty et al., JCO PO 2017