Computational Molecular Medicine
TUM School of Computation, Information and Technology
Technical University of Munich

Optimal Hard Thresholding drastically reduces the time to compute expression and splicing outliers using OUTRIDER and FRASER 2.0

Andrea Raithel, Ata Jadid Ahari, Felix Brechtmann, Christian Mertes, Ines Scheller, Vicente Yépez✉, and Julien Gagneur✉

✉ yepez(at)in.tum.de
✉ julien.gagneur(at)tum.de

April 06, 2025

1 Introduction

DROP is a computational workflow that allows for the analysis of aberrant expression and aberrant splicing from RNA-seq data [1] using OUTRIDER [2] and FRASER 2.0 [3], respectively. Both tools are based on a denoising autoencoder that controls for covariation patterns by exploiting correlations in the data. In an autoencoder's latent space, the essential features of the input data are captured in a lower-dimensional representation. Thus, employing the optimal latent space dimension is crucial to enhance the performance of the autoencoder. For this purpose, artificial outliers are injected randomly into the dataset and the encoding dimension maximizing the precision-recall AUC for identifying corrupted read counts is selected (Figure 1A). However, this approach is computationally demanding, as it requires performing a time-intensive autoencoder fit for a range of encoding dimensions that increases with sample size.

Salkovic et al. [4] proposed OutSingle, a fast outlier detection method that utilizes Optimal Hard Thresholding (OHT) for confounder control. OHT is a deterministic technique to denoise low-rank matrices based on singular value decomposition (Figure 1B). In DROP version 1.5.0 we integrate OHT to find the optimal encoding dimensions for the autoencoders. Thereby, we accelerate the OUTRIDER and FRASER 2.0 pipelines substantially, while enhancing or maintaining the enrichment of rare high-impact variants.

2 Materials and Methods

2.1 Data

Here, we used the GTEx dataset which consists of 16,213 RNA-seq samples from 49 tissues of assumed-to-be-healthy individuals in the Genotype-Tissue Expression Project V8 [5]. This extensive dataset allowed us to benchmark our methods for RNA-seq data using orthogonal DNA data.

2.2 Optimal Hard Thresholding

Optimal Hard Thresholding established by Gavish and Donoho [6] is a deterministic approach to denoise low-rank matrices relying on singular value decomposition. It assumes a model $Z = X + \gamma E$, where the noise matrix E has i.i.d. zero-mean entries and X represents the signal. For RNA-seq data we assume an unknown noise level γ, which can be determined by a robust estimator:

$$\hat{\gamma}(Z) = \frac{\sigma_{\text{med}}}{\sqrt{n \cdot \mu_\beta}}.$$

σ_{med} is the median singular value of Z and μ_β is the median of the Marchenko-Pastur distribution that can be evaluated numerically by the Adaptive Gauss-Kronrod Quadrature.

Then, the Optimal Hard Threshold can be computed as:

$$\hat{\tau}(\beta, Z) = \lambda_*(\beta) \cdot \sqrt{n}\hat{\gamma}(Z) = \frac{\lambda_*(\beta)}{\sqrt{\mu_\beta}}\sigma_{\text{med}},$$

where β is the aspect ratio of Z. The Optimal Hard Threshold coefficient $\lambda_*(\beta)$ can be derived from the formula:

$$\lambda_*(\beta) = \sqrt{2(\beta+1) + \frac{8\beta}{(\beta+1) + \sqrt{\beta^2 + 14\beta + 1}}}.$$

The optimal latent space dimension corresponds to the rank of the smallest singular value that surpasses the optimal hard threshold $\hat{\tau}(\beta, Z)$, which is visualized in Figure 1B. For OUTRIDER, the standardized log-transformed read counts serve as OHT input, while for FRASER 2.0 the computed intron Jaccard indices in the logit-scale are utilized.

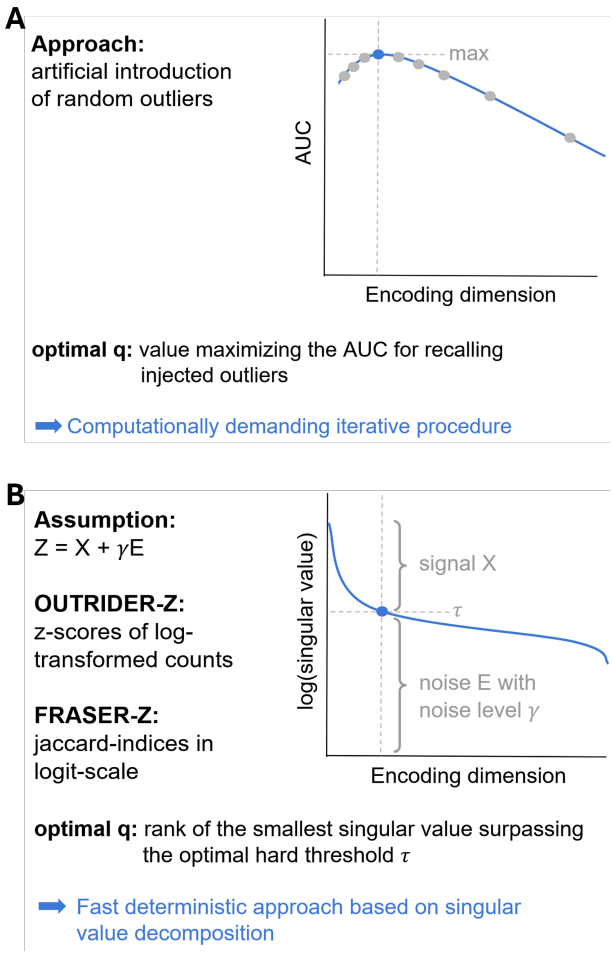

Figure 1 Contrasting overview of **A)** the original outlier injection method versus **B)** Optimal Hard Thresholding to determine the optimal encoding dimension for the autoencoder fit.

2.3 Implementation

Both for OUTRIDER (Version 1.21.0) and FRASER 2.0 (Version 2.0.0), the *estimateBestQ* function now replaces the *findEncodingDim* and *optimHyperParams* functions, respectively. The *estimateBestQ* function determines the optimal latent space dimension using OHT per default. Alternatively, the user can set `useOHT=FALSE` to estimate the optimal encoding dimension by recalling artificially introduced outliers. To visualize the OHT results, the *plotEncDimSearch* function can generate a singular value plot that resembles the plot in Figure 1B. Regarding the different splice metrices in FRASER 2.0, currently only the Intron Jaccard Index is supported for OHT. DROP was updated to incorporate the newest versions of OUTRIDER and FRASER 2.0.

2.4 Enrichment analysis

Rare variants were obtained from the GTEx WGS data (V8) and filtered for a $MAF < 0.01$. For the evaluation of the OUTRIDER results, only variants predicted to have a high impact according to the Variant Effect Predictor (VEP) [7] were considered, while the analysis of the FRASER 2.0 output was additionally restricted to splice donor and acceptor variants. A p-value threshold of 5×10^{-5} and 5×10^{-9} was applied to detect outlier genes in aberrant expression and aberrant splicing, respectively. Enrichments of rare variants among detected outliers were computed as the proportion of outlier genes associated with a rare variant over the proportion of non-outliers associated with a rare variant. To assess the statistical significance of the findings, a paired Wilcoxon signed rank test was performed.

3 Results and Discussion

We benchmarked the original tools OUTRIDER and FRASER 2.0 that search for the optimal latent space dimension among a range of values against the updated pipelines using the OHT implementation (OUTRIDER-OHT and FRASER-OHT). For aberrant expression, we additionally included two competitor methods OutSingle [4] and saseR [8] which both integrate OHT in their pipeline without fitting an autoencoder. The following sections discuss the results across various metrics, including encoding dimensions, execution time, and variant enrichment.

3.1 Moderately increased latent space dimensions

First, we compared the encoding dimensions determined by OHT with those obtained from the artificial outlier injection approach. We observed a trend in Figure 2, where the OHT-derived values are generally larger than the original optimal latent space dimensions. However, the increase is only moderate, as most points remain close to the diagonal.

A possible reason for the elevated encoding dimensions, is that the artificial outlier injection method relies on the sampling of outlier amplitudes from fixed distributions. This assumption may not fully capture the complexity of outlier behaviour and

could lead to a slight underestimation of the optimal latent space dimension.

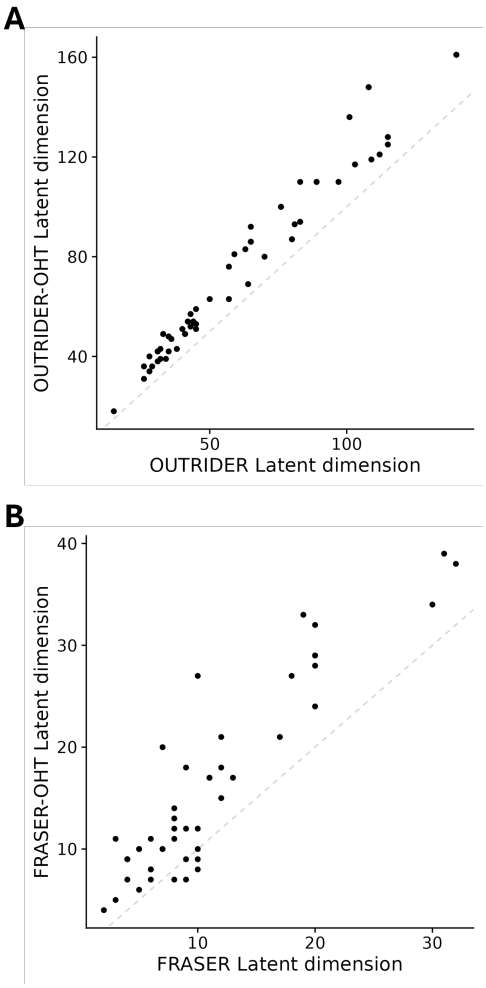

Figure 2 Comparison of encoding dimensions determined by OHT versus the original values predicted by recalling artificially injected outliers for **A)** OUTRIDER and **B)** FRASER 2.0. Each data point represents one tissue of the GTEx dataset.

3.2 Accelerated execution time

All methods were assessed for their execution time running on 30 CPU cores per tissue. While OutSingle and saseR remain the fastest tools for aberrant expression detection, replacing the artificial outlier injection method in the OUTRIDER pipeline with the deterministic OHT approach led to a reduction of the execution time by an average factor of 11 across GTEx tissues (Figure 3A). In the case of FRASER 2.0, we achieved a fourfold decrease by introducing OHT (Figure 3B). Further accelerations in both applications are limited by the time-intensive final autoencoder fit using the pre-computed encoding dimensions.

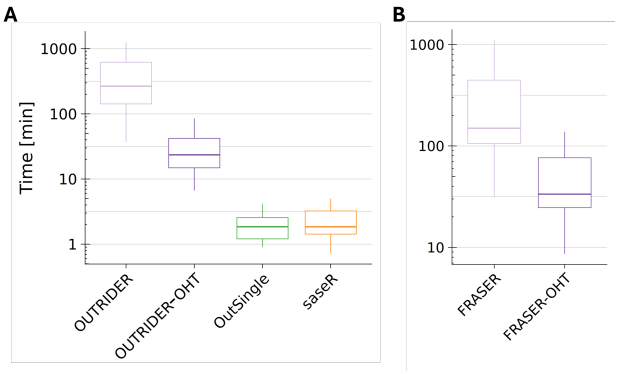

Figure 3 Median execution time of **A)** aberrant expression and **B)** aberrant splicing tools to process the GTEx data.

3.3 Enhanced variant enrichment

To ensure that the reported improvement in the execution time does not result in reduced qualitative performance, we analyzed the enrichment of rare high-impact variants among detected outlier genes. Figure 4 displays that utilizing the encoding dimensions computed by OHT for the autoencoder fit did not negatively affect the enrichment. In fact, OUTRIDER-OHT performed significantly better than the original OUTRIDER as well as OutSingle (Figure 4A). For saseR, we observed some instabilities in the outlier detection, predicting an unexpectedly high number of aberrant genes for a few samples. This impairs the interpretability of the pairwise comparison between saseR and OUTRIDER-OHT. Examining Figure 4B, we also observe a statistically significant increase in the enrichment of rare splice variants by FRASER-OHT.

These results again highlight the advantage of a deterministic approach for finding the optimal latent space dimension. The random procedure of the artificial outlier injection is based on practically derived data distributions that might not be suitable for every dataset of interest. In contrast, OHT should identify the optimal encoding dimension for any given matrix, provided that all formal requirements are met.

4 Conclusion

In conclusion, the integration of OHT substantially enhances the efficiency of both OUTRIDER and

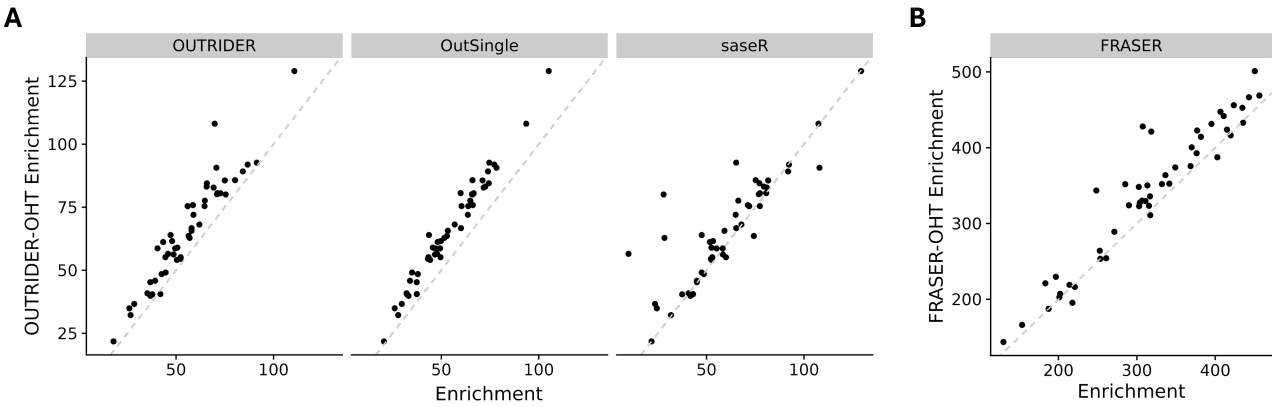

Figure 4 Enrichment of rare high-impact variants in **A)** aberrantly expressed and **B)** aberrantly spliced genes of GTEx samples. Each data point represents one tissue.

FRASER 2.0 by reducing execution times while improving or maintaining the enrichment of rare high-impact variants. The deterministic approach offers a reliable alternative to the computationally demanding iterative procedure of recalling artificially introduced outliers, enabling more scalable analyses of RNA-seq data in diagnostics and biomedical research.

References

1. Yépez VA, Mertes C, Müller MF, Klaproth-Andrade D, Wachutka L, Frésard L, Gusic M, Scheller IF, Goldberg PF, Prokisch H, and Gagneur J. Detection of aberrant gene expression events in RNA sequencing data. Nature Protocols 2021; 16:1276–96. DOI: 10.1038/s41596-020-00462-5
2. Brechtmann F, Mertes C, Matusevičiūtė A, Yépez VA, Avsec Ž, Herzog M, Bader DM, Prokisch H, and Gagneur J. OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. American journal of human genetics 2018; 103:907–17. DOI: 10.1016/j.ajhg.2018.10.025
3. Scheller IF, Lutz K, Mertes C, Yépez VA, and Gagneur J. Improved detection of aberrant splicing with FRASER 2.0 and the intron Jaccard index. American journal of human genetics 2023; 110:2056–67. DOI: 10.1016/j.ajhg.2023.10.014
4. Salkovic E, Sadeghi MA, Baggag A, Salem AGR, and Bensmail H. OutSingle: a novel method of detecting and injecting outliers in RNA-Seq count data using the optimal hard threshold for singular values. Bioinformatics 2023; 39. DOI: 10.1093/bioinformatics/btad142
5. Battle A, Brown CD, Engelhardt BE, and Montgomery SB. Genetic effects on gene expression across human tissues. Nature 2017; 550:204–13. DOI: 10.1038/nature24277
6. Gavish M and Donoho DL. The Optimal Hard Threshold for Singular Values is 4/sqrt(3). IEEE Transactions on Information Theory 2014; 60:5040–53. DOI: 10.1109/TIT.2014.2323359
7. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, and Cunningham F. The Ensembl Variant Effect Predictor. Genome biology 2016; 17:122. DOI: 10.1186/s13059-016-0974-4
8. Segers A, Gilis J, van Heetvelde M, Baere E de, and Clement L. Juggling offsets unlocks RNA-seq tools for fast scalable differential usage, aberrant splicing and expression analyses. bioRxiv 2023 :2023.06.29.547014. DOI: 10.1101/2023.06.29.547014