When the outlier is the signal:
Fast and sensitive expression outlier calling

Tian Tang, Mehmet Celimli, Brendan He and Andrea Raithel

Computational Modeling for System Genetics

07 Feb 2024

Abstract

The detection of aberrant genes in RNA sequencing data important in identifying the related genes of Aberrant gene events and rare diseases such as aberrant splicing, and loss of function variants in genes [5] by utilising these detection algorithms Aberrant genes can be detected and the pathogenicity of genes [8]. The first case was OUTRIDER which was able to recall six pathogenic events related to patients with Mendelian disorder encoder [2]. However, due to its reliance on matrices, it is computationally expensive and requires long processing times, as a response to the long processing times of OUTRIDER many new models have arrived to make a model for outlier detection that is quicker and possess the same effectiveness or higher then OUTRIDER. While the success of outlier detection models has been archived in many studies, few studies have implemented and compared the performance of the new and emerging outlier detection models. Our study aims to find and determine the most effective and efficient method of outlier detection from OutSingle, our developed OUTRIDERsingle, saseR, Axolotl and OUTRIDER detection models. And test the functionality of each implementation to recall and enrich the detection of loss of function genes from the detected outliers. Our results indicate that our OUTRIDERsingle implementation boasted one of the highest precision in outlier detection **(Figure 6. B).** All other implementations were more than 10x faster than OUTRIDER with OutSingle requiring the least time with a median of 0.5 hours **(Figure 3)** and all implementations required far less memory **(Figure 4).** Their performance in the recalling and enrichment of gene variants responsible for loss of function achieved a similar average recall of 0.02 0.075 0.04 from the top 10,000 outliers. We achieved our goals in this study to implement and check the performance of the different methods. There were however inconsistencies in some implementations such as issues with running jobs parallel or manually that may have caused huge variations in computational time and issues that arose from some implementations being in their preprint stage such as Axolotl [1] and saseR. Despite this, our study successfully determined the efficiency and effectiveness of the outlier models in determining aberrant genes and the most effective implementation can therefore be determined in the field of medicine.

1. Introduction

Each human harbours a large number of rare variants in its genome which range from single nucleotide substitutions to complex structural variations. These variants are frequently associated with rare Mendelian disorders and have diverse consequences. Whereas some rare variants directly affect the amino-acid composition of the gene product, most of them have an impact on gene regulation, which can be reflected in the transcriptome.[8] Let us consider an insertion into a transcription factor binding site. This event might prevent a crucial transcription factor from binding to its specific target sequence and thus cause a significant decrease in the expression of the corresponding gene. Therefore, RNA-sequencing (RNA-seq) is commonly utilized as a tool to pinpoint pathogenic variants in diagnostics through the detection of aberrant gene expression.

The utilization of outlier detection algorithms in RNA sequencing data (RNA-seq) has emerged as a potent approach to differentiate between regular and aberrant gene expression. This task is challenging due to the inevitable influence of confounders such as technical batches, environmental conditions, sample characteristics or common genetic variations. For this purpose, Brechtmann, Mertes, Matuseviciute et al. developed an advanced outlier detection method called OUTRIDER (Outlier in RNA-Seq Finder) in 2018 [2]. By determining an optimal latent space dimension followed by the employment of a denoising autoencoder, this open-source algorithm eliminates correlation noise from the RNA-seq read-counts before applying a statistical test for outlier prediction. Yet, the major limitation of the OUTRIDER model is the execution time. As the input read-count matrices are large, finding the optimal encoding dimension is computationally demanding [2].

In response to this computational challenge, more efficient outlier detection techniques began to emerge. OutSingle (Outlier Detection using Singular Value Decomposition) assumes a log-normal distribution for gene expression data and establishes an optimal hard threshold based on singular value decomposition (SVD) to control for confounders [3]. Contrary, AXOLOTL (Aberrant gene eXpression as OutLier detectiOn using deviaTion) represents an unsupervised machine learning model that controls for biological confounders by applying coexpression constraints [1]. Another method of enhanced speed is saseR (Scalable Aberrant Splicing and Expression Retrieval) which utilizes conventional bulk RNA-seq workflows for aberrant expression analysis.

All of those novel methods claim to be faster than the in-house developed OUTRIDER method while yielding comparable results. Thus, our goal for this project was to implement a faster implementation of OUTRIDER, which exploits the deterministic approach of OutSingle to calculate the encoding dimension and will be referred to as OUTRIDERSingle. Subsequently, we benchmarked this newly introduced version in regards to i) execution time, ii) memory usage, iii) optimal latent space dimension, iv) outliers per sample and v) recall of rare high-impact variants against OUTRIDER, OutSingle, saseR and AXOLOTL.

2. Methods and Datasets

To detect the outliers in the RNA-seq gene expression while addressing the hidden confounding effects not known *a priori*, Several methods have been introduced over the years, each employing various statistical tools, while also sharing common ground in their approach to analyzing RNA-seq data. The OUTRIDERSingle is the hybrid method we have combined the q value estimation of OutSingle and used that q for the OUTRIDER method. Moreover, keep in mind that, saseR and AXOLOTL remain at the preprint stage though they can be found in the Bioconductor package.

2.1 OUTRIDER

The first of our methods is OUTRIDER [2], which stands as a method to detect the aberrant gene expression in RNA-seq data to diagnose rare diseases. The previous methods for detecting aberrant gene expressions have relied on manual corrections for the confounding effects, such as applying a threshold to the Z scores. To address these, OUTRIDER has utilized an autoencoder to capture the covariations across genes establishing the confounder control.

Autoencoder is a special type of neural network composed of two parts: encoder and decoder. The encoder's purpose is to embed the raw gene counts into low dimensional representations(h_q), containing the biologically meaningful aspects of the data which is evaluated by the loss functions. The decoder reconstructs the expected gene expression to fit a negative binomial distribution. The decoder output is the controlled gene counts adjusted for covariations across genes in an RNA-seq. One arising problem is determining the optimal dimension of the latent space (q). To find it, corrupted data is given to the model by a frequency of 10^{-2} to the given data. Then the dimension that maximizes the precision-recall curve for identifying these corrupted counts k^c is selected as optimal. The two-tailed p values are calculated for the null hypothesis that the counts follow the negative binomial distribution[1]. Benjamini-Yekutieli false-discovery rate method is used for multiple testing because even after the autoencoder step not all the confounding effects are eliminated.

On top of adjusted P values, the OUTRIDER package calculates the Z scores using the following formulas[2]. l_{ij} is the log-transformed count after correction for confounders. k_{ij} is the count for gene j of sample i and c_{ij} = AE(k_{ij}).

$$P_{ij} = 2 \cdot \min\left\{\frac{1}{2}, \sum_{k=0}^{k_{ij}} NB(k_{ij} \mid \mu_{ij}, \theta_j), 1 - \sum_{k=0}^{k_{ij}-1} NB(k_{ij} \mid \mu_{ij}, \theta_j)\right\}. \tag{1}$$

$$z_{ij} = \frac{l_{ij} - \mu_j^l}{\sigma_j^l}$$

$$l_{ij} = \log_2\left(\frac{k_{ij}+1}{c_{ij}+1}\right), \tag{2}$$

After the computation of these statistical values outliers are filtered by setting different significance levels to the p-values and Z-scores. Thus, OUTRIDER offers an end-to-end function for filtering outliers, computing the statistics, and finding the aberrant genes for rare disease diagnosis.

2.2 OutSingle

OutSingle

Considering OUTRIDER being too computationally demanding despite unbiased, Salkovic et. Al. decided to model gene counts with log-normal distribution then apply an off-the-shelf matrix decomposition method singular value decomposition (SVD) to control for confounders, with optimal hard threshold (OHT) as the rank of denoised matrix. This method, claimed by Salkovic et. Al., in comparison with the denoising autoencoder (OUTRIDER), is almost instantaneous.

Gene-specific log-normal z-scores matrix $\tilde{Z}$ is generated by controlling genet counts k_{ji} with sample-specific DESeq size factors s_i, then log-transforming the controlled counts c_{ji}, finally obtaining gene-specific log-normal z-score $\tilde{z}_{ji}$ is calculated by normalizing log-controlled counts l_{ji} with gene-specific means μ_j and standard deviations τ_j.

$$s_i = \underset{i}{median} \frac{k_{ji}}{\left(\prod_{t=1}^{N} k_{jt}\right)^{\frac{1}{N}}} \tag{1}$$

$$c_{ji} = \frac{k_{ji}}{s_i} \; ; \overline{c_J} = \sum_{t=1}^{N} c_{ji} \tag{2}$$

$$l_{ji} = \log_2\left(\frac{c_{ji}+1}{\bar{c_J}+1}\right) \sim \mathcal{N}\left(\mu_j, \tau_j\right) \tag{3}$$

$$\widetilde{z_{J\iota}} = \frac{l_{ji} - \mu_j}{\tau_j} \tag{5}$$

Singular value decomposition (SVD) and optimal hard threshold (OHT)

Consider the z-score matrix $\tilde{Z}$ as perturbation of a low-rank matrix Z ("signal") with a Gaussian noise matrix E. The idea is that signal matrix Z captures all confounding factors and noise matrix E captures all outliers. With singular value decomposition, $\tilde{Z}$ could be broken down into

$$\tilde{Z} = \tilde{U}\tilde{\Sigma}\tilde{V}^T \tag{7}$$

Where $\tilde{U}$ and $\tilde{V}^T$ are the orthogonal matrices, representing rotation geometrically, with $[J, J]$ and $[N, NJ]$ dimensions. $\tilde{\Sigma}$ is a rectangular diagonal matrix, representing stretch geometrically, with dimensions $[J, N]$ and non-negative values ${\sigma^2}_N$ on the diagonal ("singular values"). The singular values are sorted in descending order: ${\sigma^2}_1 \geq {\sigma^2}_2 \geq \cdots \geq {\sigma^2}_N \geq 0$. Let's perform an artificial separation on $\tilde{\Sigma}$, assuming we know r is the rank of signal matrix Z. Then $\tilde{Z}$ could be decomposed as:

$$\tilde{Z} = \tilde{U}(\Sigma + \Sigma_E)\tilde{V}^T = \tilde{U}\Sigma\tilde{V}^T + \tilde{U}\Sigma_E\tilde{V}^T = Z + E \tag{6-10}$$

$$\text{Where } \Sigma = \begin{bmatrix} \Sigma^r_{[r,r]} & 0_{[r,N-r]} \\ 0_{[J-r,r]} & 0_{[J-r,N-r]} \end{bmatrix}, \ \Sigma_E = \begin{bmatrix} 0_{[r,r]} & 0_{[r,N-r]} \\ 0_{[J-r,r]} & \Sigma^n_{[J-r,N-r]} \end{bmatrix}$$

If we can find r, we will have the matrix E with outliers masked and unmasked by confounding factors.

Gavish and Donoho (2014) proposed to determine r (optimal hard threshold OHT) by studying the behavior of asymptotic mean square error (MSE) when dimensions of a matrix are much larger than its true rank (by keeping the rank while increasing the size of matrix). Consider matrix E as a standard normal distributed matrix E_{SND} stretched by scalar λ(level of white noise). When λ is unknown, they propose applying equation 12-15 will yield the optimal hard threshold. The equations are straightforward without complicated matrix multiplication. Despite equation 15 having no closed-form solution, $\omega(\beta)$ can be easily approximated in implementation as $0.56 \times \beta^3 - 0.95 \times \beta^2 + 1.82 \times \beta + 1.43$ (2014).

$$r = \omega(\beta)\sigma_{median}, \ \beta = \frac{N}{J} \tag{12-13}$$

$$\lambda(\beta) = \left(2(\beta+1) + \frac{8\beta}{(\beta+1) + (\beta^2 + 14\beta + 1)^{1/2}}\right)^{1/2}, \tag{14}$$

$$\int_{(1-\beta)^2}^{\mu_\beta} \frac{\left(\left((1+\sqrt{\beta})^2 - t\right)\left(t - (1-\sqrt{\beta})^2\right)\right)^{1/2}}{2\pi t} dt = 1/2 . \tag{15}$$

Given r, matrix of outliers E is then derived. The final z-scores matrix is obtained by applying gene-specific normalization. The nominal P values and false discovery rate (FDR) adjusted P values are then obtained sequentially.

OutSingle indeed proved to be much faster and less computationally demanding than OUTRIDER with comparable performance on outlier detection, which we will show in the Results section.

2.3 OUTRIDERSingle

We'd like to see whether OUTRIDER has any other performance boost other than shortening time if we apply the SVD OHT to get the optimal encoding dimension q.

2.4 saseR (Scalable Aberrant Splicing and Expression Retrieval)

Agreeing that OUTRIDER's implementation, although unbiased but time-consuming and computationally burdensome, Sergers et. al also regards the OutSingle's implementation, despite obtaining a similar performance with less computational time, fails to "account for the heteroskedasticity and the discrete nature of the count of RNA-Seq data" (2023), i.e., an inadequate distribution. Therefore, they propose that conventional bulk methods can be used to estimate the negative binomial distribution (NBD) once the optimal encoding dimension is found. They claim their implementation of saseR is both fast and scalable for parameter estimation to assess aberrant gene expression (2023).

First, saseR uses DESeq2's geometric mean to normalize counts. In our case, no known confounding factors are provided, so saseR proceeds to calculate the offset matrix and then normalize it. The normalized offset matrix is then fed into singular value decomposition to find the optimal encoding dimension.

Then saseR introduces its fast computational algorithm which replaces the negative binomial (NB) variance structure with quadratic variance structure, such that

$$Var(Y_{ij}) = \phi_j \mu_{ij}^2$$

Compounded with the log link function,

$$log(\mu_{ij}) = \eta_{ij}$$

The parameter estimation of the Newton-Raphson algorithm required in the fitting of NBD is reduced to matrix multiplication and remains unbiased, instead of the original solving by iterations.

$$\begin{aligned}\beta_j^{k+1} &= \beta_j^k + (\boldsymbol{X}^{\boldsymbol{T}}\frac{1}{\phi_j}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\frac{1}{\phi_j}\frac{1}{\boldsymbol{\mu_j}}\boldsymbol{y_j} - \boldsymbol{\mu_j} \\ &= \beta_j^k + (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\frac{(\boldsymbol{y_j} - \boldsymbol{\mu_j})}{\boldsymbol{\mu_j}},\end{aligned}$$

saseR as the authors claim is also several magnitudes faster than OUTRIDER, which we will also demonstrate in the Results section.

2.5 AXOLOTL

AXOLOTL is a new method for outlier detection in RNA-seq data. that processes normalizes and transforms raw count data into a matrix to train a Local Outlier Factor (LOF) unsupervised learning to highlight aberrant genes, it bases its determination of aberrant genes on the five customizable features;

cts_z*, *rank_z*, *cts_dev*, *rank_dev*, *ogs_pv

Outrider

The Raw gene counts were first submitted through Outrider to undergo autoencoder normalization. Optimal encoding dimension q was generated inside a specified range for noise reduction and normalization of gene expression. The z-scores were then calculated and then ranked in descending order to obtain the score matrix cts_z and rank_z respectively

Detecting Co-expression Partners; Pearson's Correlation Coefficient

Pearson's correlation coefficient (**R**) is calculated for each Gene and Sample, by extracting the gene expression values from the data table from Outrider, the covariance is calculated and able to be used to determine co-expression partners that emit similar changes in gene expression. A matrix χ was created with χ((%,%*)) where (%,%*) represents the strength of co-expression for gene % and gene %*. Determining a linear correlation between the gene coexpression strength % and %* where if a correlation for (%,%*) → 1 the likelihood of two genes emitting co-expression was greater. The genes that emitted the highest co-expression strength (top 2%) acted as constraints for the LOF model training model.

Deviation from co-expression Partner

Deviations of gene expression was determined and constrained by the co-expressing partners (cts_dev and rank_dev). The deviation was established by comparing the average z-score difference between gene % of m and co-expression partners % %* The deviation scores of gene expression were calculated based on the equations (1) and (2). Where they took the z scores matrix of ctsZ of m rows and n column then subtracted the matrix by -1/ |m∗| * the sum of the average z score of co-expression genes m' (1). |m∗| refers to the number of genes in m* where the number of genes in the coexpression partner is included (m') (m'∈m*), and m' is the z score of co-expression genes. This was repeated for the rankZ matrix (2). This way the z score of the original z score matrix was constrained by the average z score of co-expressed genes. T

$$cts\_dev(m,n) = ctsZ(m,n) - \frac{1}{|m^*|}\sum_{m'\in m^*} ctsZ(m',n) \tag{1}$$

$$rank\_dev(m,n) = rankZ(m,n) - \frac{1}{|m^*|}\sum_{m'\in m^*} rankZ(m',n) \tag{2}$$

OutSingle

The OutSingle method was implemented to determine p-values for the gene expression of each sample. The recorded p-values underwent negative logarithmic conversion creating the feature matrix ogs_pv.

Local Outlier Factor

The Local Outlier Factor (LOF) is a method for detecting outliers in the gene expression dataset by calculating the local density of a gene expression point relative to its neighbour (k) The distance and density of gene expression scores were used to determine aberrant genes by representing each gene with an anomaly score based on the deviation from aberrant gene expression, the anomaly score was further constrained by the co-expression partners %,%* values determined from Pearson's (r). The lower anomaly scores indicate a high likelihood of aberrancy and co-expression partners.

2.6 Datasets

Geuvadis: The Geuvadis dataset was obtained from the 1000 Genomes Project, the dataset consists of the raw read counts of genes of the Lymphoblastoid Cell Line after DNA sequencing. The 100 participants whose blood was obtained for the analysis were assumed to be healthy. The data was obtained by genomic sequencing.

Kremer-119: The skin fibroblast cell lines of 119 patients were extracted with the assumption that they possessed rare mitochondrial disease. The data was generated from transcriptome sequencing.

GTEx: The GTEx (Genotype Tissue Expression) is a data resource and tissue bank responsible for the Genotype Tissue Expression Project V8 (.dbGaP: phs000424.v8.p1). The data consists of Raw counts of RNA sequencing data from 17,350 RNA-seq samples 54 tissues were extracted from 948 post-mortem organ donors.

3. Results

All methods are applied via the same setup. 100GB of RAM is allocated for each job that runs the respective method. Considering the computational demand we decided that the datasets are processed in parallel three at a time. Results are logged both during and upon completion of each dataset.

3.1 Execution time

The execution time of each method for each dataset is logged from the start to the end of the method. As seen in the violin plots below, OutSingle and saseR detect the outliers in RNA-seq instantly on scales of minutes, though OutSingle is at least twice as fast. The methods OUTRIDERSingle and OUTRIDER depend on training an autoencoder to detect the outliers while considering the confounding effects. While OUTRIDERSingle finding the latent dimension size(q) significantly speeds up the process. All the iterations and matrix computations still demand heavy computational resources.

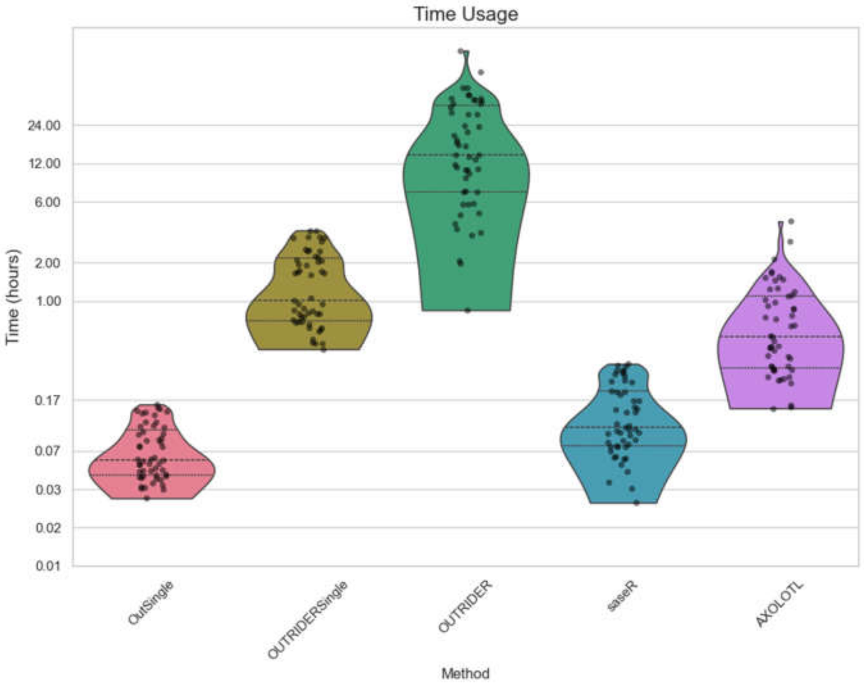


Fig.3 Violin Plot of Time required for each outlier detection implementation to process datasets in hours. The time was recorded after the implementations began processing the inputted dataset represented as points on the graph. The graph stretched represents the distribution of time required per dataset, medians are highlighted inside the violins with a dotted line.

3.2 Memory usage

To evaluate the efficiency of computational resources resident set size (RSS) is calculated for every dataset. RSS quantifies the amount of physical memory allocated for a certain process. It includes heap memory, stack, and the memory for the code segment. Similar to the total execution time comparison, OutSingle proves to be the most memory efficient in terms of RSS only followed by saseR. The OUTRIDER on the other hand on average consumes the most memory among the methods. Some of the datasets for the OUTRIDER method consume very little memory which we think might be related to the experimentation errors, but even upon further investigation, we spotted that these datasets are smaller in size compared to others. The exact cause of OUTRIDER's small memory usage remains unknown.

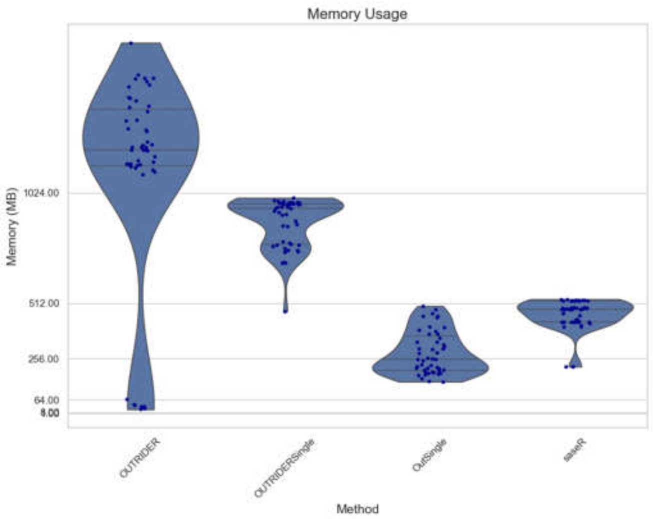


Fig.4 Violin graph of the four p-value outlier detection methods OUTRIDER, OUTRIDERsingle, OutSingle and saseR. Determined by Memory usage of Megabytes (MB) used when processing results. Jobs were done in parallel for OUTRIDER OUTRIDERSingle and OutSingle. saseR data and script ran independently. Each point represents a dataset that was read by the detection method.

3.3 Encoding dimension

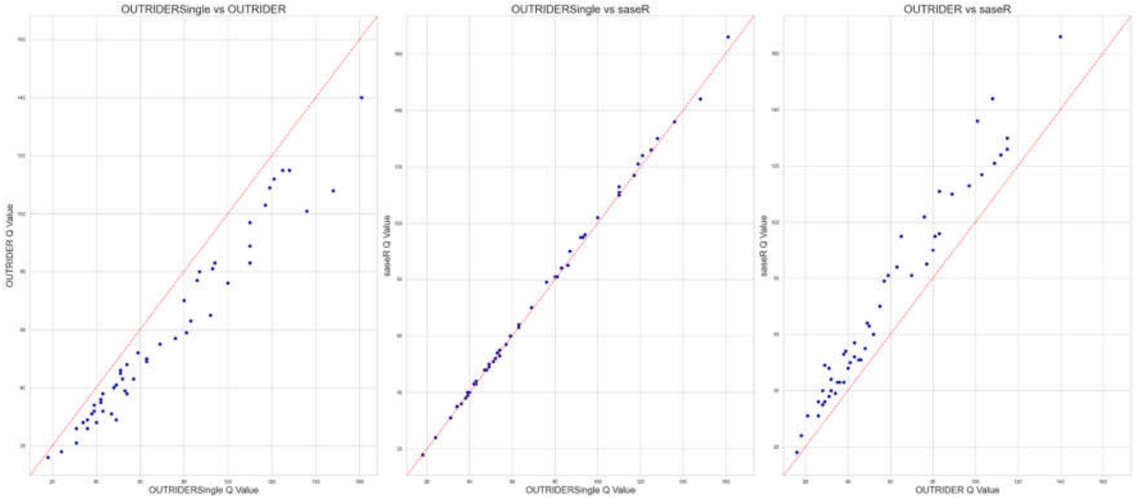


Fig.5: Latent dimension size comparison for the outlier detection methods. Each blue dot represents a dataset.

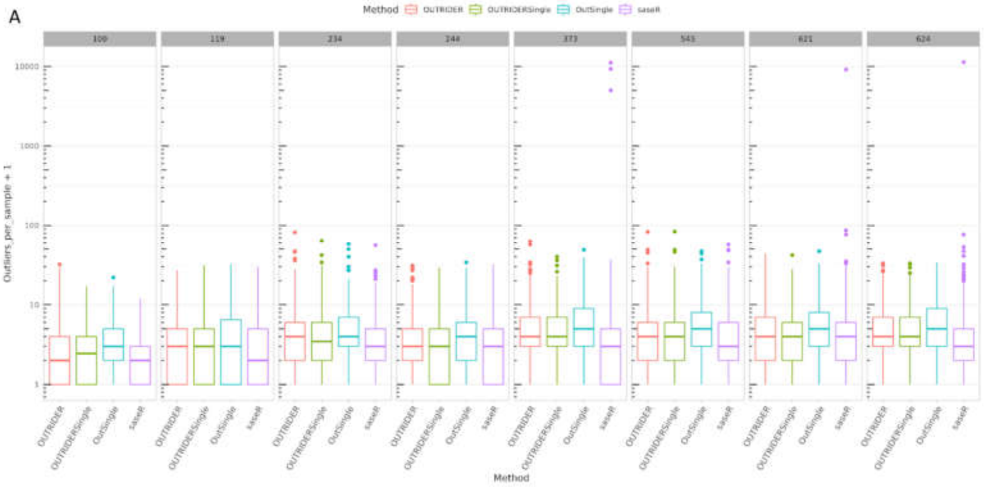
The above graph compares the q values of different methods, OutSingle was not included since q values are exactly equal to the OUTRIDERSingle's. AXOLOTL omitted the same q values were derived OUTRIDER itself. saseR and OUTRIDERSingle subplot reveal that their q value estimation is very similar. OutSingle and OUTRIDERSingle use log normalized Z scores on the other hand saseR directly uses the Z scores directly before applying an optimal hard threshold. Interestingly, OUTRIDER tends to yield the smallest q values from either of the three methods.

3.4 Number of outliers per sample

Here we'd like to focus on the number of outliers detected by each method

Figure 7. A shows that, the number of outliers per sample across datasets (here we select 8 datasets with different dimensions to be inclusive) are very stable, despite very rarely predicting the very high number of outliers on large datasets. Out of the four methods, OutSingle tends to produce a slightly higher number of outliers. Figure 7. B shows that although faint symmetric behaviour around the diagonal exists (i.eThere are samples predicted with few outliers by one method but more by another method), most data points fall on the diagonal line, meaning predictions on the number of outliers per sample are consistent across methods.

Figure 7. C is the upset plot of the outliers predicted across methods for all datasets. Almost 42k outliers are commonly predicted by all four methods, indicating some consistency among the methods; OUTRIDERSingle noticeably has the lowest number of outliers predicted only by itself and a high number of predictions shared with other methods; Conversely, saseR has the highest number of outliers predicted only by itself, and a low number of predictions shared with other three methods.



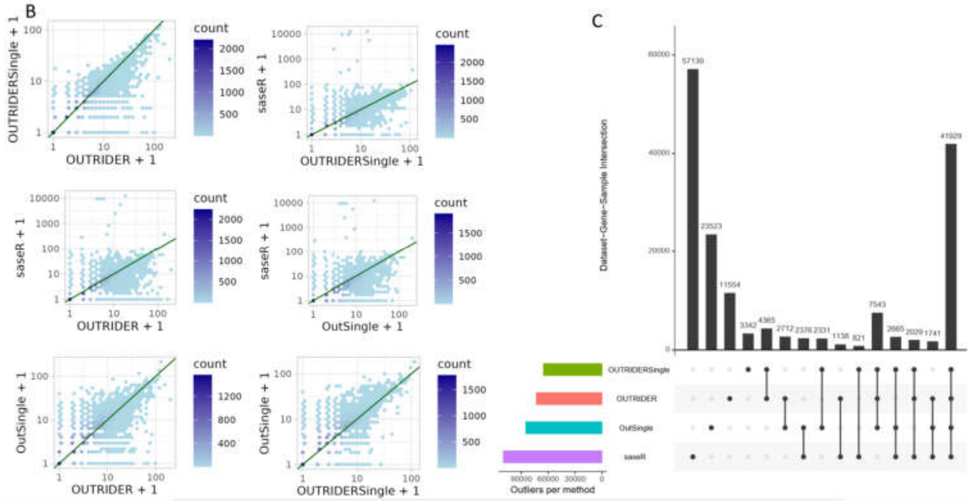


Figure 7. Compare outliers predicted by each method. A. Boxplot of number of outliers per sample on selected 8 datasets. B. Scatter plot of methods 1-1 comparison on outliers (dataset-gene-sample combination) predicted all outliers predicted across 51 datasets. C. UpSet plot of outlier predicted across methods, and all method combinations.

3.5 Precision and Recall

In addition to the RNA-seq data from the Geuvadis consortium, also a corresponding VCF file is publically available. From these variant annotations, only a specific subgroup was extracted using the vcfR package. More precisely, sample-gene-variant combinations were filtered for i) carriers of at least one alternative allele, ii) an allele frequency AF < 0.005 and iii) high-impact (frameshifts, splice-acceptor and -donor variants, stop-gained and start-lost variants). This filtering process yielded a total of 1096 variants for further downstream analysis, which in the following will be referred to as VEP-annotated rare high-impact variants.

Next, several metrics were calculated based on gene-sample pair matches between identified outliers of various algorithms and the extracted variants. As displayed in **Fig. 6 A,** the recall of VEP-annotated rare high-impact variants was obtained for outliers ranked by nominal p-value or anomaly score. Taking into account 100000 outliers, each method shows a similar curve with a steep rise for the most significantly aberrantly expressed genes that bends into a linear behaviour upon reaching a recall of approximately 7.5%. This behaviour is caused by the decreasing significance of the methods´ outputs, resulting in a linear curve with a slope determined by the probability of randomly picking an annotated rare high-impact variant amongst all possible sample-gene combinations. Thus, performance comparison should rather be derived from the first 1000 or 10000 outliers (**Fig. 6 A,** middle and right subplot). In those subplots, it becomes visible that AXOLOTL, OUTRIDER, OUTRIDERSingle and OutSingle perform equally well, whereas saseR is outperformed.

This finding is even more striking in the Venn diagram of detected VEP-annotated rare high-impact variants with an FDR-adjusted p-value <= 0.05. The saseR model fails to detect any additional outliers on top of the eleven variants identified by all other considered methods. In contrast, OUTRIDERSingle correctly classified 26 aberrant gene expression events, from which three were uniquely identified by this method. Finally, precision values were computed and visualized in **Fig. 6 C.** This metric comparison further enhances the finding that OUTRIDERSingle with a precision of 10.3% exceeds competitive methods, in particular the base model OUTRIDER.

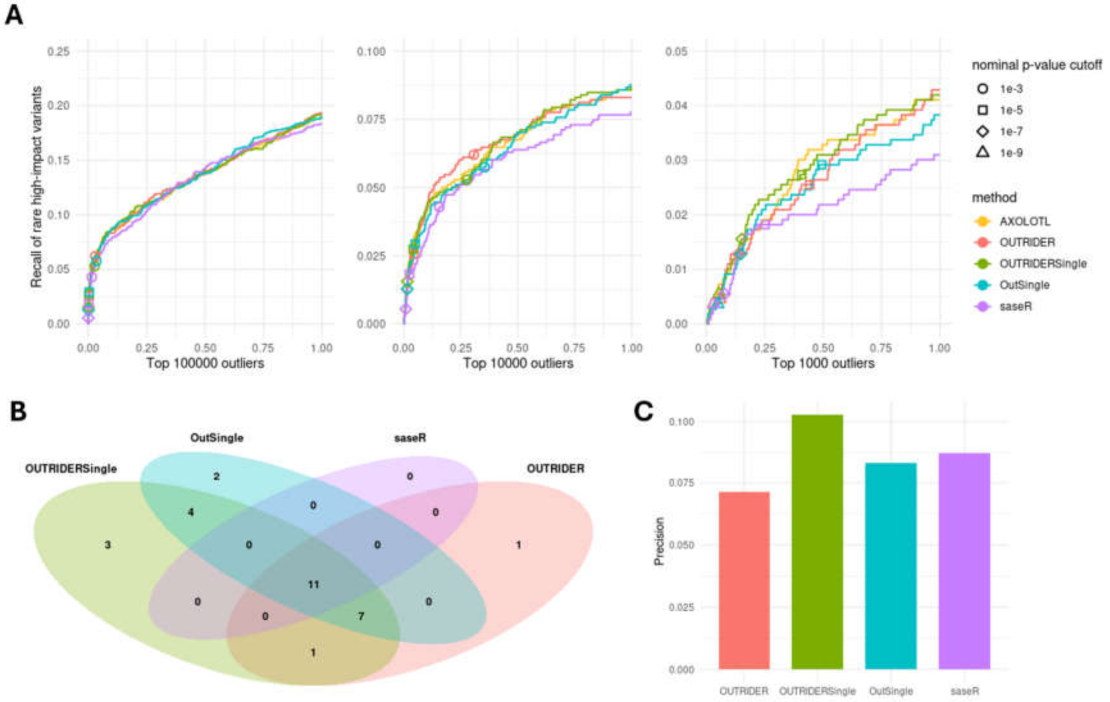


Figure 6: Benchmarking results for recall of rare high-impact variants and precision on the Geuvadis dataset across various outlier detection methods as indicated by colour. (A) Recall of VEP-annotated rare high-impact variants amongst the top 100000, 10000 and 1000 outliers for different outlier detection methods. For OUTRIDER, OUTRIDERSingle, OutSingle and saseR, detected outliers were sorted by nominal p-values with specific cutoffs indicated by different shapes. For AXOLOTL, detected outliers were sorted by anomaly scores. **(B)** Venn diagram of detected VEP-annotated rare high-impact variants with an FDR-adjusted p-value <= 0.05. **(C)** Barplot of precision values considering variants with an FDR-adjusted p-value <= 0.05.

3.6 Q-Q plot for distribution

The motivation to see the Q-Q is to see whether the generated P-values correspond to the original assumption of the designed algorithm. Due to page limitations, we only display the Q-Q plot of the Geuvadis dataset. Most observed P-values lie on the diagonal of expected P-values, suggesting there is certain soundness (at least empirically) toward the choices of distribution. For OUTRIDER, OUTRDERSingle and saseR, they choose the NBD, which is regarded as a standard choice of modelling gene count; however OutSingle employs log normal distribution, as we witness, there isn't much deviation from the diagonal of OutSingle's Q-Q plot. It's also revealed in the last section Precision and Recall that OutSingle didn't find considerably fewer true variants than other methods.

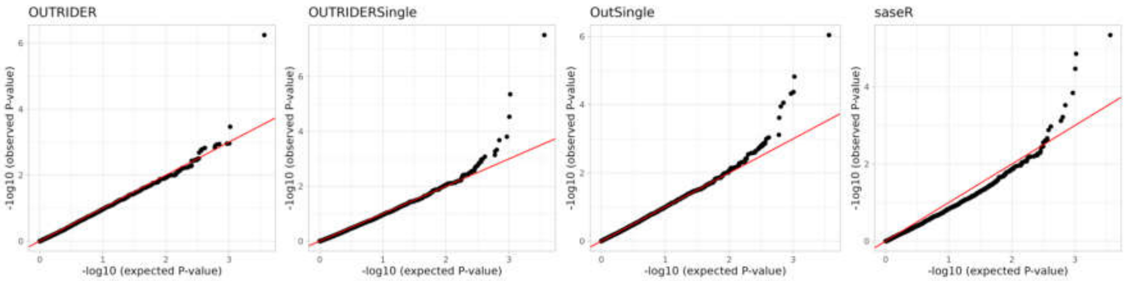


Figure 8. Q-Q plots of the Geuvadis dataset across methods. From left, OUTRIDER, OUTRIDERSingle, OutSingle, saseR.

4. Discussion

Our main aim for this study was to implement a faster implementation of the OUTRIDER model for outlier detection in RNA-seq data and to compare it to competitive methods regarding efficiency and effectiveness.

4.1 Enrichment of rare high-impact variants

To evaluate the overall behaviour of the considered methods in the recall analysis, we compared our plots **(Fig. 6 A)** to Fig. 2 B from Scheller et al. displaying an equal approach for the detection of rare splice-disrupting variants. Even though all plots show a similar flattening curve, the methods described in the paper generated distinguishable differences in the recall metric. One underlying reason is that those methods were benchmarked on the GTEx datasets, comprising a much larger consortium than the Geuvadis dataset with only 100 samples [9]. Therefore, we need to be careful about conclusions on the performance regarding the recall of VEP-annotated rare high-impact variants. Nevertheless, our findings were supported by the Venn diagram, where OUTRIDERSingle emerged as the most sensitive method. Moreover, our implementation was able to outperform competitive methods in terms of precision.

4. 2 OUTRIDERsingle was the most Effective Method

OUTRIDERsingle was able to detect outliers more precisely than the other models with a precision score of ~10% **(Figure.6.B)** and it was 10x faster than Outrider **(Figure.3),** the inclusion of OutSingle encoding dimension retrieved from OHT with OUTRIDER boosted its computational time and lowered memory requirement dramatically as the process to determine optimal threshold involves simply calculating ω(β) then multiplying the estimated noise(λ) by the obtained ω(β) to obtain the threshold. OUTRIDER required many steps. This is more hardware intensive, especially in cases when working with matrices and larger datasets this can take a lot more computational time. Many other methods do not calculate the q to the same extent which is why many of the other methods (Axolotl, saseR, OutSingle and OUTRIDERsingle) were all faster and required less memory. **(Figure.3) (Figure.4)**

4.3 OutSingle was the fastest Method

The fastest outlier detection method was OutSingle which had a median required time of ~0.5 hours to detect outliers in the sequence data (**Figure.3)**. Regarding Q encoding dimensions the Q value for each implementation varied greater, in the case of OUTRIDER single and saseR the encoding dimension Q was consist**ure.5).** However an odd case occurs when comparing saseR and Outsingle, despite using the same q estimation method they deviated, we suspect that OutSingle applies size-controlled then obtains the log-normalized z-score matrix. Then applies singular value decomposition. saseR is similar. However saseR code applies edgeR's deviance calculation, due to no known confounder in the experiment design, it’s uncertain if the deviance calculation went through to determine the same q encoding dimension.

4.4 Axolotl and its Future Applications

Axolotl is currently in its preprint stage and with the lack of p-values it is difficult to classify outliers and non-outliers. Its benefit is that unlike any method it considers the co-expression of genes whereas other samples do not it can however be used to detect aberrant genes however it requires reading publication and collecting test data such as the anomaly scores. This system is not automated and is more suited for the Prioritization of genes to investigate.

4.5 Thresholding and Comparable to P-Value Implementations

There is currently no known imitation of thresholding for the LOF model or past implementation of Leng. et al, due to no definition of anomaly score, one was considered as P-values can correlate with z-scores however with no clear definition of anomaly scores we can not come to a certain threshold.

4.6 Conclusion

Not only have models evolved to detect outliers faster, many new outlier detection methods are even being introduced for specific purposes such as saseR for detecting aberrant splicing events. These implementations can be used as complementary tools to determine causative variants of diseases that are related to aberrant gene expression events [8]. Therefore, our study findings might be a valuable contribution to medical research, if revised and evaluated on other datasets. We achieved our goal of elaborating a faster implementation of the OUTRIDER model. We not only sped up the calculation process of the optimal encoding dimension, but this technique also yielded promising results indicating an enhancement of the biological performance.

6. References

[1] Leng, F., Liu, Y., Zhang, J., Shen, Y., Liu, X., Wang, Y., & Xu, W. (2023). AXOLOTL: an accurate method for detecting aberrant gene expression in rare diseases using coexpression constraints. Retrieved from https://doi.org/10.1101/2024.01.07.574502

[2] Brechtmann, F., Mertes, C., Matusevičiūtė, A., Yépez, V. A., Avsec, Ž., Herzog, M., et al. (2018). OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. The American Journal of Human Genetics, 103(6), 907–917. doi: 10.1016/j.ajhg.2018.10.025

[3] Salkovic, E., Sadeghi, M. A., Baggag, A., Salem, A. G. R., & Bensmail, H. (2023). OutSingle: a novel method of detecting and injecting outliers in RNA-Seq count data using the optimal hard threshold for singular values. Bioinformatics, 39, btad142. doi: 10.1093/bioinformatics/btad142

[4] Segers, A., Gilis, J., Heetvelde, M. V., Baere, E. D., & Clement, L. (2023). Juggling offsets unlocks RNA-seq tools for fast scalable differential usage, aberrant splicing and expression analyses. Retrieved from https://doi.org/10.1101/2023.06.29.547014

[5] Kremer, L. S., Bader, D. M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., et al. (2017). Genetic diagnosis of Mendelian disorders via RNA sequencing. Nature Communications, 8, 15824. doi: 10.1038/ncomms15824

[6] Lappalainen, T., Sammeth, M., Friedländer, M. R., ‘t Hoen, P. A. C., Monlong, J., Rivas, M. A., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature, 501, 506–511. doi: 10.1038/nature12531

[7] Wortmann, S. B., Koolen, D. A., Smeitink, J. A., et al. (2015). Whole exome sequencing of suspected mitochondrial patients in clinical practice. Journal of Inherited Metabolic Disease, 38(3), 437–443.

[8] Yépez, V.A., Mertes, C., Müller, M.F. et al. Detection of aberrant gene expression events in RNA sequencing data. Nat Protoc 16, 1276–1296 (2021). https://doi.org/10.1038/s41596-020-00462-5

[9] Scheller IF, Lutz K, Mertes C, Yépez VA, Gagneur J. Improved detection of aberrant splicing using the Intron Jaccard Index. medRxiv [Preprint]. 2023 Apr 3:2023.03.31.23287997. doi: 10.1101/2023.03.31.23287997

[10]M. Gavish and D. L. Donoho, "The Optimal Hard Threshold for Singular Values is 4/\sqrt {3}4/\sqrt {3}," in IEEE Transactions on Information Theory, vol. 60, no. 8, pp. 5040-5053, Aug. 2014, doi: 10.1109/TIT.2014.2323359.