OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data

Felix Brechtmann,[1,5] Christian Mertes,[1,5] Agnė Matusevičiūtė,[1,5] Vicente A. Yépez,[1,2] Žiga Avsec,[1,2] Maximilian Herzog,[1] Daniel M. Bader,[1] Holger Prokisch,[3,4] and Julien Gagneur[1,2,*]

RNA sequencing (RNA-seq) is gaining popularity as a complementary assay to genome sequencing for precisely identifying the molecular causes of rare disorders. A powerful approach is to identify aberrant gene expression levels as potential pathogenic events. However, existing methods for detecting aberrant read counts in RNA-seq data either lack assessments of statistical significance, so that establishing cutoffs is arbitrary, or rely on subjective manual corrections for confounders. Here, we describe OUTRIDER (Outlier in RNA-Seq Finder), an algorithm developed to address these issues. The algorithm uses an autoencoder to model read-count expectations according to the gene covariation resulting from technical, environmental, or common genetic variations. Given these expectations, the RNA-seq read counts are assumed to follow a negative binomial distribution with a gene-specific dispersion. Outliers are then identified as read counts that significantly deviate from this distribution. The model is automatically fitted to achieve the best recall of artificially corrupted data. Precision-recall analyses using simulated outlier read counts demonstrated the importance of controlling for covariation and significance-based thresholds. OUTRIDER is open source and includes functions for filtering out genes not expressed in a dataset, for identifying outlier samples with too many aberrantly expressed genes, and for detecting aberrant gene expression on the basis of false-discovery-rate-adjusted p values. Overall, OUTRIDER provides an end-to-end solution for identifying aberrantly expressed genes and is suitable for use by rare-disease diagnostic platforms.

Introduction

No clear pathogenic variant can be pinpointed for the majority of individuals suspected to suffer from a Mendelian disorder after undergoing whole-exome or whole-genome sequencing (WES or WGS, respectively).[1–3] A possible reason is that the pathogenic variant is regulatory. Accurately identifying pathogenic regulatory variants is difficult. One difficulty is that any individual harbors a very large number of rare non-coding variants, about 60,000 compared with 475 protein-affecting rare variants per genome (with minor allele frequency [MAF] < 0.005).[4] Another difficulty is that the interpretation of non-protein-coding regions of the genome remains challenging.[5]

Two recent studies have shown that using RNA sequencing (RNA-seq) to directly investigate gene expression defects in cells of affected individuals provides a complementary method to pinpoint pathogenic regulatory defects.[6,7] RNA-seq can help to reveal splicing defects, the mono-allelic expression of heterozygous loss-of-function variants, and expression outliers (i.e., genes aberrantly expressed outside their physiological range).[6,7] The two studies used different approaches to identify expression outliers. Cummings et al.[6] computed Z scores on the log-transformed gene-length-normalized read counts by subtracting the mean count and dividing by the standard deviation. Expression outliers were identified as read counts with an absolute Z score greater than 3. Cummings et al. did not apply a formal statistical test for outlier detection and also explicitly note that their outlier analysis was underpowered to draw definitive conclusions.[6] This study did not yield any convincing pathogenic expression-outlier candidates. In contrast, the study by Kremer et al.[7] identified four out of six newly diagnosed individuals as expression outliers. Read-count outliers were identified as those with an absolute Z score greater than 3 and statistical significance according to DESeq2, a statistical test originally developed for differential expression analyses,[8] which the authors applied by testing each sample against the rest of the cohort. DESeq2 is based on the negative binomial (NB) distribution, which can be used to model overdispersed count data.[9] Altogether, the reason for the difference remains unclear because of the relatively small number of diagnosed individuals, the absence of ground truth, and the lack of a direct comparison between the two approaches based on the same data.

The two studies differed not only in whether a statistical test was applied but also in the way the data were controlled for confounders. Cummings et al.[6] used RPKM (reads per kilobase per million mapped reads) expression values. These control for variations in sequencing depth but not for other sources of covariation among the read counts. Controlling for further sources of covariation is important because the identification of a gene as aberrantly expressed depends on the context, for example, the sex of the donor. Genes encoded on the Y chromosome

[1]Department of Informatics, Technical University of Munich, Boltzmannstr. 3, 85748 Garching, Germany; [2]Quantitative Biosciences Munich, Gene Center, Department of Biochemistry, Ludwig-Maximilians Universität München, Feodor-Lynen-Str. 25, 81377 München, Germany; [3]Institute of Human Genetics, Helmholtz Zentrum München, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany; [4]Institute of Human Genetics, Klinikum rechts der Isar, Technical University of Munich, 13 Ismaninger Str. 22, 81675 München, Germany
[5]These authors contributed equally to this work
*Correspondence: gagneur@in.tum.de
https://doi.org/10.1016/j.ajhg.2018.10.025.

are not present and thus not expressed in women. However, in men, loss of the expression of a Y-chromosome-encoded gene can be an aberrant expression event. Hence, not taking the sex of the donor into account would not allow for the detection of aberrantly silenced Y-chromosome-encoded genes in males. Although the sex of the donor is usually available and can be easily controlled for, other contexts for gene expression—such as the exact tissue of origin of the sample, the sample's cell-type composition, the genetic background, and technical biases—might not be known *a priori*, causing similar but less intuitive variations. Kremer et al.[7] controlled expression levels for sex, biopsy site as inferred from the *HOX* gene set, and common technical sources of variation, which were identified by visual inspection of a hierarchical clustering of the samples. In a study that identified expression outliers, although not for the diagnosis of rare diseases, Li et al.[10] controlled for sex and the top three genotype principal components, as well as for hidden confounding effects estimated by the probabilistic estimation of expression residuals (PEER) method.[11] However, the algorithms controlling for covariations in RNA-seq read-count data used in the studies of Li et al.[10] and Kremer et al.[7] were neither assessed nor tuned to detect aberrantly expressed genes.

Here, we introduce OUTRIDER (Outlier in RNA-Seq Finder), an algorithm that provides a statistical test for outlier detection in RNA-seq samples while controlling for covariations among the gene read counts (Figure 1A). The modeling of covariation is performed by an autoencoder that controls for read-count variations caused by factors not known *a priori*. Its parameters are optimized automatically for recalling read counts corrupted *in silico*. Autoencoders were introduced to find low-dimensional representations of high-dimensional data.[12–14] They have been shown to be useful for extracting meaningful biological features from RNA-seq data[15] and imputing missing values in single-cell RNA-seq data.[16] A subclass of autoencoders, the so-called denoising autoencoders, are used for reconstructing corrupted high-dimensional data by exploiting correlations in the data.[17] In OUTRIDER, the autoencoder approach is used to control for the common covariation patterns among genes.

Differential-expression algorithms, such as DESeq2[8] and edgeR,[18] have been conceived to compare small predefined groups of samples, typically treatment versus control, with a handful of replicates. To manage such small sample sizes, these approaches borrow information across genes to have robust estimates of the within-group variability. The setup for rare-disease diagnostics is different. In rare-disease diagnostics, replicates are typically not available for most individuals, and there is no typical experiment design of treatment versus control; rather, there are several tens of samples, where one sample is tested against all others. Also, one is not interested in detecting a subtle fold change between two controlled populations but rather in identifying an outlier within a large population (Figure 1B). We note that DESeq2 and edgeR also have procedures based on Cook's distance and Pearson residuals[8,19] to mark or downweight outliers but with a different purpose than for rare-disease diagnostics. These methods aim to increase the robustness of the differential-expression analysis rather than assess the significance of the outlier data points. Here, we adopt a typical approach for outlier detection for the univariate case by modeling the distribution of the population and testing each data point to determine whether it significantly deviates from this distribution.[20]

In this article, we describe the OUTRIDER algorithm, which combines an autoencoder and a statistical test for outlier detection, and delineate the added value of these two components against state-of-the-art alternatives by utilizing simulated data and two experimental datasets.

Material and Methods

Datasets

The RNA-seq read counts, in the following called counts, were downloaded from Data S1 published as part of the study by Kremer et al.[7] for the rare-disease cohort. GTEx counts were obtained from the Genotype-Tissue Expression (GTEx) Portal (V6P counted with RNA-SeQC v.1.1.8).[21] Counts for the Kremer et al. dataset were computed according to the UCSC Genome Browser build hg19[22] with consideration of the full gene body. In contrast, GTEx is based on the Gencode v.19 annotation,[23] and the count of a gene is defined as the number of paired-end read pairs overlapping exons of that gene only. Samples with a low RNA integrity number (RIN $<$ 5.7) were filtered out from the GTEx dataset. FPKM (fragments per kilobase per millions of reads) values were obtained with DESeq2,[8] where the gene length was defined as the aggregated length of all the exons. We then filtered for expressed genes, defined as genes for which at least 5% of the samples had a FPKM value greater than 1 (Figure S1). Additionally, we discarded genes that had zero counts in more than 75% of the samples.

Statistical Model

We assume that the count k_{ij} of gene $j = 1, \ldots, p$ in sample $i = 1, \ldots, n$ follows a NB distribution with gene-specific dispersion parameter θ_j and expected value c_{ij}:

$$P(k_{ij}) = \mathrm{NB}(k_{ij} \mid \mu_{ij} = c_{ij}, \theta_j). \qquad \text{(Equation 1)}$$

The used parameterization of the NB distribution can be found in the Supplemental Material and Methods. We limited the parameter range for θ_j to the interval [0.01, 1000]. The lower limit prevents convergence issues for genes with unusual high dispersion (θ_j close to zero), and the upper limit is used to avoid overfitting. The expected count c_{ij} is the product of the sample-specific size factor s_i and the exponential of the factor y_{ij}:

$$c_{ij} = s_i \cdot \exp\left(y_{ij}\right) \qquad \text{(Equation 2)}$$

The size factors s_i capture variations in sequencing depth; they are robustly estimated as the median of the ratios of the gene read counts to their geometric means as implemented in DESeq.[24] The factors y_{ij} capture covariations across genes. They

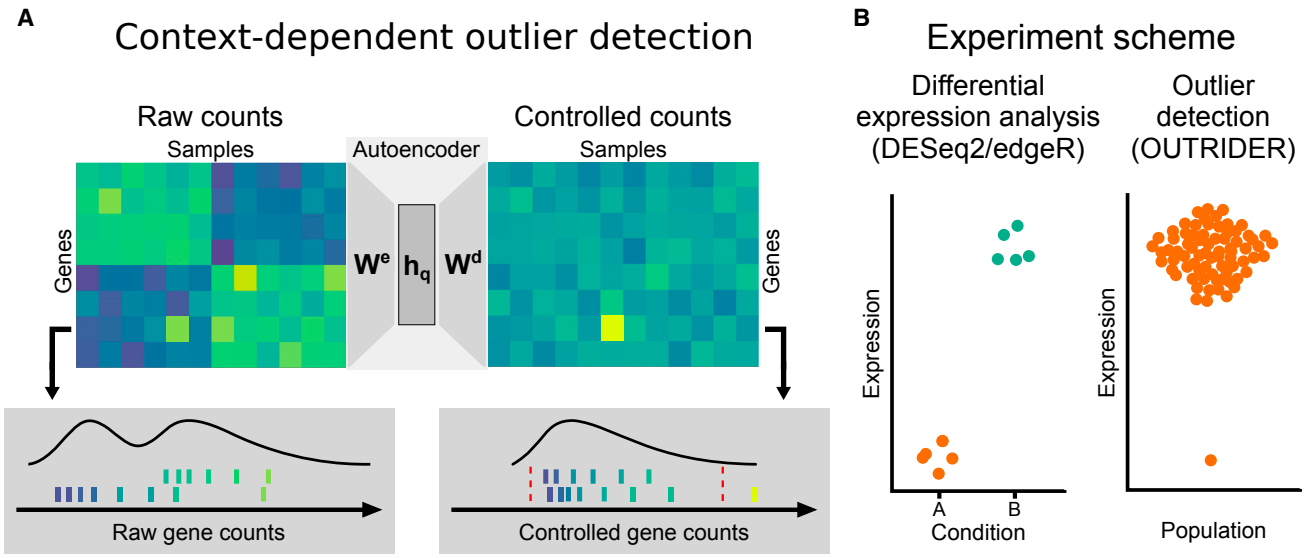

Figure 1. OUTRIDER Overview

(A) Context-dependent outlier detection. The algorithm identifies gene expression outliers whose read counts are significantly aberrant given the covariations typically observed across genes in an RNA-seq dataset. This is illustrated by a read count (left panel, fifth column, second row from the bottom) that is exceptionally high in the context of correlated samples (left six samples) but not in absolute terms for this given gene. To capture commonly seen biological and technical contexts, an autoencoder models covariations in an unsupervised fashion and predicts read-count expectations. Comparing the earlier mentioned read count with these context-dependent expectations reveals that it is exceptionally high (right panel). The lower panels illustrate the distribution of read counts before and after controlling for covariations for the relevant gene. The red dotted lines depict significance cutoffs.

(B) Schema showing the differences in the experimental designs for differential expression analyses and outlier detection analyses; relevant analysis packages are mentioned.

are modeled with an autoencoder of encoding dimension $1 < q < \min(p, n)$. Specifically,

$$\mathbf{y}_i = \mathbf{h}_i \mathbf{W}_d + \mathbf{b}, \qquad \text{(Equation 3)}$$

$$\mathbf{h}_i = \tilde{\mathbf{x}}_i \mathbf{W}_e, \qquad \text{(Equation 4)}$$

where the $p \times q$ matrix $\mathbf{W}_e$ is the encoding matrix, the $q \times p$ matrix $\mathbf{W}_d$ is the decoding matrix, the q-vector $\mathbf{h}_i$ is the encoded representation, and the p-vector $\mathbf{b}$ is a bias term. Having a decoding matrix that is not the transpose of the encoding matrix, unlike for principal-component analysis (PCA), turned out to be important, most likely because the property that the matrix inverse equals the matrix transpose does not generalize to the NB loss function. The input vector to the autoencoder $\tilde{\mathbf{x}}_i$ is computed as follows:

$$\tilde{x}_{ij} = x_{ij} - \overline{x_j}, \quad \text{where} \qquad \text{(Equation 5)}$$

$$x_{ij} = \log\left(\frac{k_{ij}+1}{s_i}\right), \qquad \text{(Equation 6)}$$

where we add 1 to prevent computing the logarithm of 0, we divide by the size factor to control for sequencing depth, and we center gene-wise by subtracting the mean $\overline{x_j}$. In the following, we call the combination of Equations 2–6 the autoencoder or, in short, $c_{ij} = AE(k_{ij})$.

Fitting the Parameters

Fitting the autoencoder is implemented as an iterative three-step procedure in which the parameters $\mathbf{W}_e$ and $\mathbf{W}_d$ and the θ_j values are iteratively updated until convergence. First, the encoder and decoder matrices are initialized with PCA through the "pca()" function provided by the package pcaMethods,[25–27] the bias vector is set to the mean of the log-transformed size-factor-normalized counts $\overline{x_j} = \text{mean}(\log((k_{ij}+1)/s_i))$, and the dispersion parameters are estimated by the method of moments, yet the dispersion parameters are restricted to the interval [0.01, 1000]. Before the iterative procedure starts, the gene-specific entries of the decoder matrix $\mathbf{W}_d$ and then the gene-specific dispersion parameters are fitted by maximum likelihood. The autoencoder is then fitted through repetition of the following three update steps: (1) the encoder matrix is updated, (2) the decoder matrix is updated per gene, and (3) the dispersion parameters are refitted per gene as detailed below. In each update step, the average negative log-likelihood is minimized with respect to the current parameters by the optimization method L-BFGS as implemented in "optim()."[28] For all three steps, detailed derivations of the used loss functions and the respective gradients can be found in the Supplemental Data.

Convergence Criteria for the Iterations of the Update Steps

The autoencoder is updated in an iterative fashion. The three update steps described above are repeated until the average negative log-likelihood of each step in one iteration does not differ more than the convergence threshold of 10^{-5} from the last step of the previous iteration or at most 15 iterations.

Fitting the Encoding Dimension

The optimal autoencoder dimension is obtained through evaluation of the performance of calling corrupted counts. To this end, we artificially introduced corrupted counts k_{ij}^c randomly with a

frequency of 10^{-2} to the given count matrix. These corrupted counts are obtained as follows:

$$u_{ij} = \log_2\left(\frac{k_{ij}}{s_i} + 1\right), \quad \text{(Equation 7)}$$

$$k_{ij}^c = \text{round}\left(s_i \, 2^{u_{ij} \pm e^z \sigma_{u_j}}\right), \quad \text{(Equation 8)}$$

where z, the amplitude of the corrupted count, is drawn from a normal distribution characterized by a mean of log(3) and a standard deviation of log(1.6). The sign of the shift is randomly selected. The optimal dimension is then selected as the dimension maximizing the area under the precision-recall curve for identifying corrupted counts.

p Value Computation

For every pair of gene j and sample i, we test the null hypothesis that the count k_{ij} follows the distribution described by Equation 1. The algorithm computes two-sided p values by using the following equation:

$$P_{ij} = 2 \cdot \min\left\{\frac{1}{2}, \sum_{k=0}^{k_{ij}} NB(k_{ij} \mid \mu_{ij}, \theta_j), 1 - \sum_{k=0}^{k_{ij}-1} NB(k_{ij} \mid \mu_{ij}, \theta_j)\right\}. \quad \text{(Equation 9)}$$

The term 1/2 is included to handle cases when both other terms exceed 1/2, which is possible because of the discrete nature of the NB distribution.

Expression levels of different genes for the same sample are correlated because of biological confounding effects such as co-regulation, which cannot be entirely excluded even after controlling by the autoencoder. The computed p values can therefore be correlated. Multiple-testing correction was performed with the Benjamini-Yekutieli false-discovery rate (FDR) method, which holds under positive dependence.[29]

Z Score Computation

Z scores Z_{ij} are computed on a logarithmic scale as follows:

$$Z_{ij} = \frac{l_{ij} - \mu_j^l}{\sigma_j^l}, \quad \text{(Equation 10)}$$

where l_{ij} is the log-transformed controlled count calculated as $l_{ij} = \log_2((k_{ij}+1)/(c_{ij}+1))$, σ_j^l is the standard deviation of l_{ij} for gene j, and μ_j^l is the mean of l_{ij} for gene j.

Benchmark by Injection of Outliers

To assess the sensitivity and specificity of alternative outlier detection methods, we injected artificial outliers with pre-specified amplitudes on the logarithmic scale (Z scores). This process was separate from the injection of corrupted data described earlier in Fitting the Encoding Dimension. We used the outlier injection scheme described in this section to independently assess OUTRIDER's performance in comparison with that of other approaches.

We used this benchmark separately for both datasets: the GTEx skin tissue not exposed to the sun and the rare-disease cohort from Kremer et al.[7] The counts were replaced with a probability of 10^{-4} by an outlier count k_{ij}^o, defined as follows:

$$k_{ij}^o = \text{round}\left(s_i \, 2^{\overline{u}_j \pm e^z \sigma_{u_j}}\right), \quad \text{(Equation 11)}$$

where $\overline{u}_j$ is the mean of u_{ij} for gene j in the log space.

Alternative Control Methods

We benchmarked OUTRIDER against PEER[30] and PCA.[25–27] Both were used instead of the autoencoder to model covariations. In the case of PCA, we obtained the matrix of expected counts by using the first q loadings as the $\mathbf{W}_e$ and $\mathbf{W}_d$ matrices, where q is the encoding dimension inferred for the autoencoder. The bias term $\mathbf{b}$ was set to the gene means. In the case of PEER, we set the number of factors to one-fourth of the number of samples as suggested by Stegle et al.[11] We then subtracted the residuals from the log-transformed counts and multiplied the size factors to obtain c_{ij}. For PEER, we used the provided residuals to compute Z scores to avoid numerical inaccuracies due to conversion to counts. For both PCA and PEER, we fitted a NB model with a per-gene adjustment a_j and a dispersion parameter $\theta_j$on top of the obtained controlled counts (Equation 12) to obtain NB p values. We used the adjustment parameter to capture deviations between the estimated mean from the log-normal model and that from the NB model:

$$P(k_{ij}) = NB(k_{ij} \mid \mu_{ij} = a_j \cdot c_{ij}, \theta_j). \quad \text{(Equation 12)}$$

Enrichment Analysis

We obtained rare variants (MAF < 0.05 within all 652 GTEx samples and in gnomAD[31]) from the GTEx WGS data (V7). We further filtered this set for those with predicted moderate or high impact according to the Variant Effect Predictor (VEP).[32] To be comparable with Li et al., we only used the 441 individuals considered in their analysis for our enrichment analysis. As described in Li et. al.,[10] we computed enrichments for rare variants found within outlier genes as the proportion of outliers having a rare variant over the proportion of non-outliers having a rare variant.

Implementation

OUTRIDER is implemented as an R package that is available through Bioconductor.

Results

We considered two RNA-seq datasets, which we refer to as the Kremer and GTEx datasets. The Kremer dataset contained 119 RNA-seq samples from skin fibroblasts of individuals with a suspected rare mitochondrial disorder.[7] This dataset was analyzed in a previous study, where the systematic effects were controlled by manual inspection of sample correlation matrices.[7] In this previous study, four genes were identified as aberrantly expressed out of six pathogenic genes detected by RNA-seq analysis and validated by functional assays.[7] This dataset therefore served as our benchmark dataset for rare-disease applications. The GTEx dataset contained 250 RNA-seq samples obtained from the not-sun-exposed skin tissue in the GTEx project (V6P).[21] Unless stated otherwise, we focused on these skin samples because the tissue was similar to the tissue of origin of the Kremer dataset. The donors of the GTEx samples did not suffer from any condition and were not under treatment. Nevertheless, aberrant gene expression in these samples has been reported.[10]

For both datasets, we filtered out genes not expressed across the whole dataset, resulting in 10,556 genes in the Kremer dataset and 17,065 genes in the GTEx dataset (Figures S1A and S1B). In the GTEx dataset, we additionally filtered out one sample because of a low RIN (<5.7), resulting in 249 samples. Both datasets exhibited a strong correlation structure with very distinct sample clusters (Figures S2A and S2E). These correlations could have arisen from biological variations such as the sex of the donor, the origin of the tissue, population structure, or hidden confounders such as poorly understood systematic technical variations.[7,10] Applying the autoencoder on the counts allowed covariations to be estimated and controlled for. The dimension of the autoencoder was fitted with a scheme in which a fraction of counts were artificially corrupted and fitted with OUTRIDER. Then the optimal dimension for the autoencoder was selected as the dimension maximizing the area under the precision-recall curve for identifying corrupted counts. We estimated the optimal encoding dimension to be 45 for the GTEx dataset and 21 for the Kremer dataset. Running the same assessment by using PCA instead of the autoencoder yielded 54 and 24 principal components, respectively. Encoding dimensions close to the estimated optimum yielded similar results. Moreover, using different corruption amplitudes had little impact on the optimal encoding dimension (Figure S3). After we controlled for covariation, clusters disappeared from both datasets (Figure S2).

Improving the Detection of Outliers by Using a NB Model

The OUTRIDER algorithm assumes that counts follow a NB distribution with a mean that is the expected value provided by the autoencoder. Assuming NB distributions with gene-specific dispersions, it detects expression outliers as significant deviations of the observed counts from these expected values. In contrast, PCA and PEER assume normal distribution and therefore require some transformation of the count data, typically the log-transformation.

To understand the differences between OUTRIDER, PCA, and PEER, we performed simulations of counts by assuming either (1) a simulation scheme corresponding to the assumptions of OUTRIDER or (2) a simulation scheme close to the assumptions of PCA and PEER. For both simulation schemes, the latent space was set to have ten dimensions. For the first simulation scheme, we drew counts from the NB distribution. For the second scheme, we drew values from a log-normal distribution and rounded them to the nearest integer to obtain counts. Means fitted by OUTRIDER were closer to the simulated means than means fitted by PCA or PEER throughout the entire range of counts on the basis of the NB simulation scheme (Figure S4A). In the log-normal case, all three methods were almost equal, but for low expression levels (simulated means lower than 30), OUTRIDER performed better than PCA or PEER on the log-normal simulated data. These simulation analyses emphasize the relevance of using a count distribution for fitting the expected counts, especially in the low count range.

We then applied OUTRIDER on experimental data. Quantile-quantile plots for individual genes indicated that OUTRIDER reasonably modeled the count distribution (Figures 2 and S5) on the GTEx and Kremer datasets and that the resulting p values can be used for detecting outliers (Figures 2C and 2D). To untangle the contribution of the autoencoder from the p value computation, we substituted the autoencoder with either PCA or PEER to estimate the expected counts. Across all genes, the OUTRIDER p values deviated less from the expected uniform distribution than the NB p values computed on top of PCA or PEER (Figures 3A and 3B). These results show that, on experimental data, the distribution of the data is better captured when modeling the covariation with OUTRIDER than when using either PCA or PEER. Consistent with these observations, the distribution of the number of outlier counts per sample at a FDR less than 0.05 (Benjamini-Yekutelli method[29]) was more even for OUTRIDER than for PCA and PEER for both datasets (Figures 3C and 3D). Moreover, samples with a high number of outliers according to OUTRIDER had similar numbers of outliers according to PCA and PEER. In contrast, in samples with a high number of outliers according to PCA and PEER, OUTRIDER did not find any or only a few outliers. To showcase this, we picked the most aberrant sample by PEER and compared it with the prediction by OUTRIDER in Figures 3E and 3F. In order to flag such aberrant samples, we introduced a cutoff (number of outliers $>$ 0.5% of expressed genes). Accumulated over all GTEx tissues, we found 9, 18, and 214 out of 8,166 samples to be aberrant by using OUTRIDER, PCA, and PEER, respectively. Altogether, the smaller number of samples with an excessive number of outliers found by OUTRIDER further indicates that OUTRIDER captures the data distribution better than PCA and PEER.

Recall Benchmark

We then benchmarked OUTRIDER for recalling outliers. To this end, we injected simulated outliers into the GTEx and Kremer datasets and monitored the fraction of these simulated outliers that could be recalled. Simulated outliers were injected with a frequency of 10^{-4} into the count matrices, resulting in 381 injected outliers for the GTEx dataset and 113 injected outliers for the Kremer dataset. The injection of outliers was done according to three scenarios: (1) all underexpressed, (2) all overexpressed, and (3) 50% overexpressed and 50% underexpressed. Each scenario was repeated for four different simulated amplitudes (with Z score values of 2, 3, 4, and 6). We monitored the recall of injected read-count outliers and the precision, i.e., the number of injected outliers among the reported outliers, by using different detection methods. We note that the precision and the recall in this setup were underestimated because the original data also contained genuine outliers. The precision-recall curves

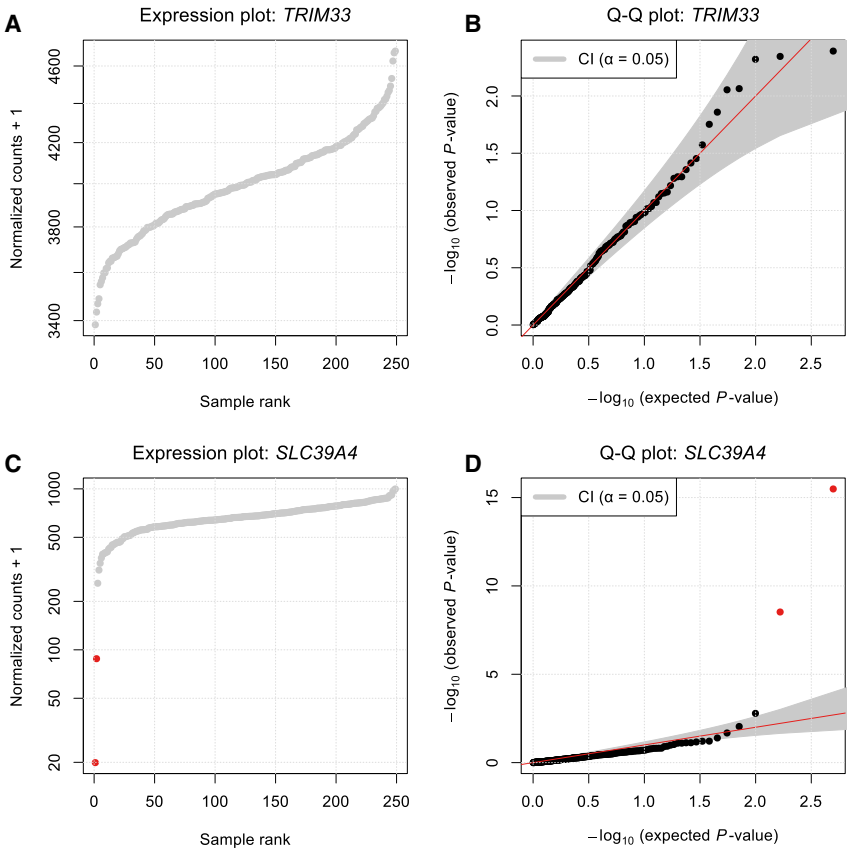

Figure 2. Using the NB Distribution for Significance Assessment
Normalized RNA-seq read counts plotted against their rank (A and C) and quantile-quantile plots of observed p values against expected p values with 95% confidence bands (B and D); outliers are shown in red (FDR < 0.05). Shown are data for *TRIM33* (MIM: 605769) with no detected expression outlier (A and B) and data for *SLC39A4* (MIM: 607059) with two expression outliers (C and D).

showed that the OUTRIDER ranking outperformed ranking by Z score with PCA or PEER, except for simulated Z scores of 6 (Figure 4). Moreover, the two most commonly used Z score cutoffs[6,10] (|2| and |3|) recalled almost all the outliers (median = 97%) for both PCA and PEER but at the cost of a high FDR (precision < 0.02). The advantage of p values is that they provide a principled way to establish a cutoff that accounts for statistical significance and multiple testing. Combining either PCA or PEER to model the expected counts with the NB model to obtain p values and FDR estimates led to improved precision-recall curves over PCA and PEER with Z score ranking, particularly for the underexpressed simulations (Figures S6 and S7). This analysis delineates the importance of using a count distribution and a p-value-based strategy.

Furthermore, we investigated the performance of an alternative strategy to detect outliers that can be easily implemented in DESeq2[8] or edgeR.[19] We controlled for known confounders as covariates and used Cook's distance, as done in DESeq2,[8] and the Pearson residuals, as done in edgeR,[19] instead of the p value to score outliers (Supplemental Material and Methods). For the Kremer dataset, these known confounders were sex and the body site inferred from the *HOX* gene set.[7] For the GTEx dataset, these were sex, age, and ischemic time. Both methods gave the same rankings because Pearson residual and Cook's distance are monotonically related. However, both showed much poorer precision-recall curves than the PCA, PEER, and OUTRIDER alternatives on the Kremer and GTEx datasets (Figures S6 and S7).

At the FDR = 0.05 cutoff, the recall was limited between 0.3 for injected outliers with $|Z| = 2$ and 0.8 for injected outliers with $|Z| = 6$ (Figure 4). To investigate which type of outliers were not recovered, we stratified the precision-recall curves by mean expression levels of the gene. Figure S8 shows the results for the $|Z| = 4$ scenario with over- and underexpressed genes, which is representative of the other $|Z|$ levels and scenarios. Here, we observed that in the lowest bin (mean count < 57; Figure S8B), the p value ranking outperformed the Z score ranking. This is due to the instability of Z scores for small counts. Nevertheless, we observed an increase in the precision for the Z score ranking with increasing mean count. Altogether, this underlines the importance of using p values instead of Z scores, particularly for genes with low expression levels.

Applying OUTRIDER to the Kremer dataset resulted in a recall of 61 outliers (9.9%) identified by Kremer et al. (Figure S9A) on the basis of the 48 previously undiagnosed samples.[7] Additionally, OUTRIDER detected 85 new expression outliers, of which 54 were downregulated. OUTRIDER was able to recall all six pathogenic events (three expression outliers, one mono-allelic expression, and two splicing defects) validated by Kremer et al. Despite the fact that *CLPP* (MIM: 601119) and *MCOLN1* (MIM: 605248) were only reported as having splice defects by Kremer et al., the resulting loss of expression was detected by OUTRIDER. On the Kremer dataset, PCA called 3.8 more outliers than OUTRIDER and missed two pathogenic events, and PEER called 7.8 times more outliers than OUTRIDER and missed one pathogenic event (Figure S9B). These results are consistent with the results from the simulations.

To further evaluate the performance of OUTRIDER on experimental data, we assessed the enrichment of rare variants among outliers, given that previous studies linked rare variants with aberrant gene expression.[10,33] We applied OUTRIDER, as well as PCA and PEER, on all

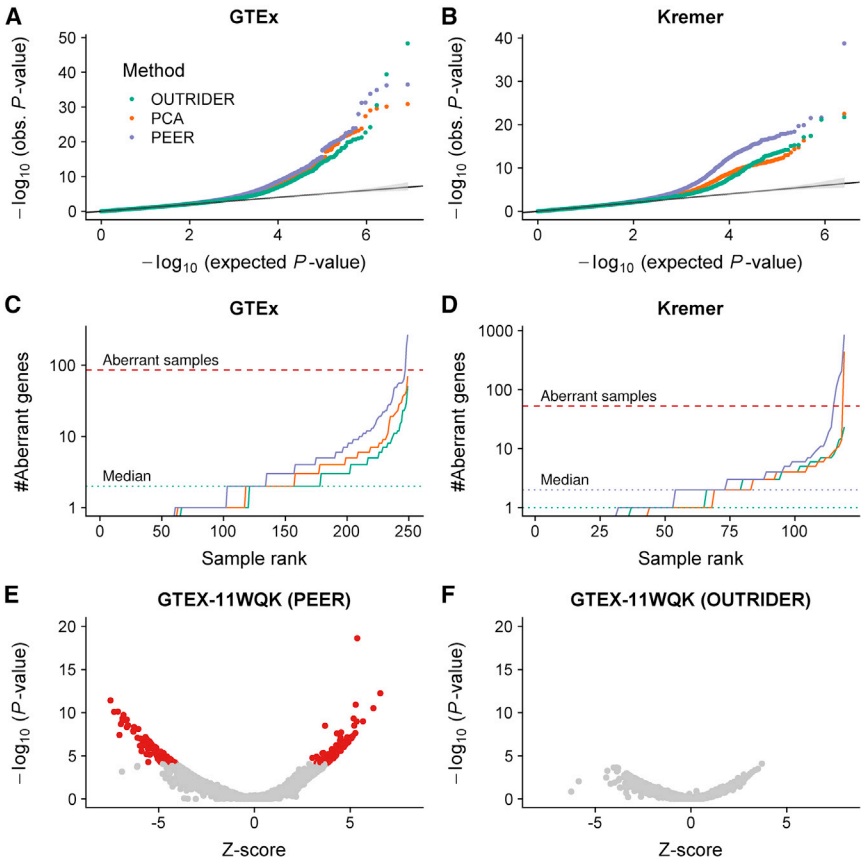

Figure 3. RNA-Seq Expression-Outlier Detection

(A and B) Quantile-quantile plots for the GTEx (A) and Kremer datasets (B). Observed p values are plotted against the expected p values for three different methods. The diagonal marks the expected distribution under the null hypothesis with 95% confidence bands (gray).

(C and D) Number of aberrant genes (FDR < 0.05) per sample for the data shown in (A) and (B) (C and D, respectively). The dashed line represents the abnormal sample cutoff (>0.5% aberrantly expressed genes).

(E and F) p values versus *Z* scores for a representative abnormal sample in PEER (E) and the same sample in OUTRIDER (F). Genes with significantly aberrant read counts are marked in red.

GTEx tissues and computed enrichments of rare variants within expression outliers on the basis of different p value cutoffs (Figure 5). We considered variants with a MAF < 0.05 and a predicted moderate or high impact according to the VEP,[32] which gave a manageable amount of variants to handle and covered the variants causing nonsense-mediated decay and transcript amplification. OUTRIDER showed higher enrichments than PCA and PEER across all tissues for three different nominal p value cutoffs (Figure 5). Furthermore, the highest enrichment was achieved with the highest cutoff regardless of the control method (Figure 5). We observed the same trend when we calculated the enrichment on the basis of *Z* score cutoffs (Figure S10). Overall, OUTRIDER had across all cutoffs the highest enrichment in comparison with the other approaches, including the original *Z* scores computed by Li et al. in a PEER-based approach.[10] Only in the case of $|Z| > 5$ did OUTRIDER and PCA achieve similar enrichment scores.

Finally, we investigated the sensitivity of sample size. To this end, we used the Kremer dataset and the six known pathogenic events to estimate the required dataset to reach significance. We monitored the nominal p values of the pathogenic events after randomly removing samples from the dataset and keeping the samples containing these six pathogenic outliers (Figure S11). As expected, the p values slowly increased toward 1 as sample size decreased. Although OUTRIDER had overall lower p values than PCA and PEER, our approach needed 60 samples to recall the majority of the pathogenic events. PCA and PEER, which had similar p values, needed at least 80 samples to recover most of those events.

Discussion

We have introduced OUTRIDER, an end-to-end solution for identifying expression outliers within RNA-seq data, controlling for hidden confounders in an automated fashion, and providing estimates of statistical significance. OUTRIDER combines an autoencoder that allows for automatically controlling technical and biological variations among genes and a statistical test based on the NB distribution. OUTRIDER outperformed preceding methods in recalling simulated outliers and pathogenic outliers from a rare-disease cohort and yielded outliers with a higher enrichment of rare variants in a cohort of healthy donors. OUTRIDER has two advantages over preceding methods. First, it computes p values that can be adjusted to control the FDR. *Z*-score-based approaches lack p values, so the setting of cutoffs is arbitrary. Second, the model's parameters are automatically fitted through optimization of the model's ability to recall corrupted counts. OUTRIDER is implemented and made available as an R Bioconductor package. The package allows for a full analysis to be made with only a few lines of code and provides plotting functionality for visualizing the results. Furthermore, the package comes along with a comprehensive vignette guiding the user through a typical analysis.

We implemented OUTRIDER so that it is not restricted to the provided autoencoder, allowing the statistical test to be used with alternative methods modeling the expected counts. In particular, PCA and PEER can be substituted for the autoencoder. Alternatively, autoencoders with

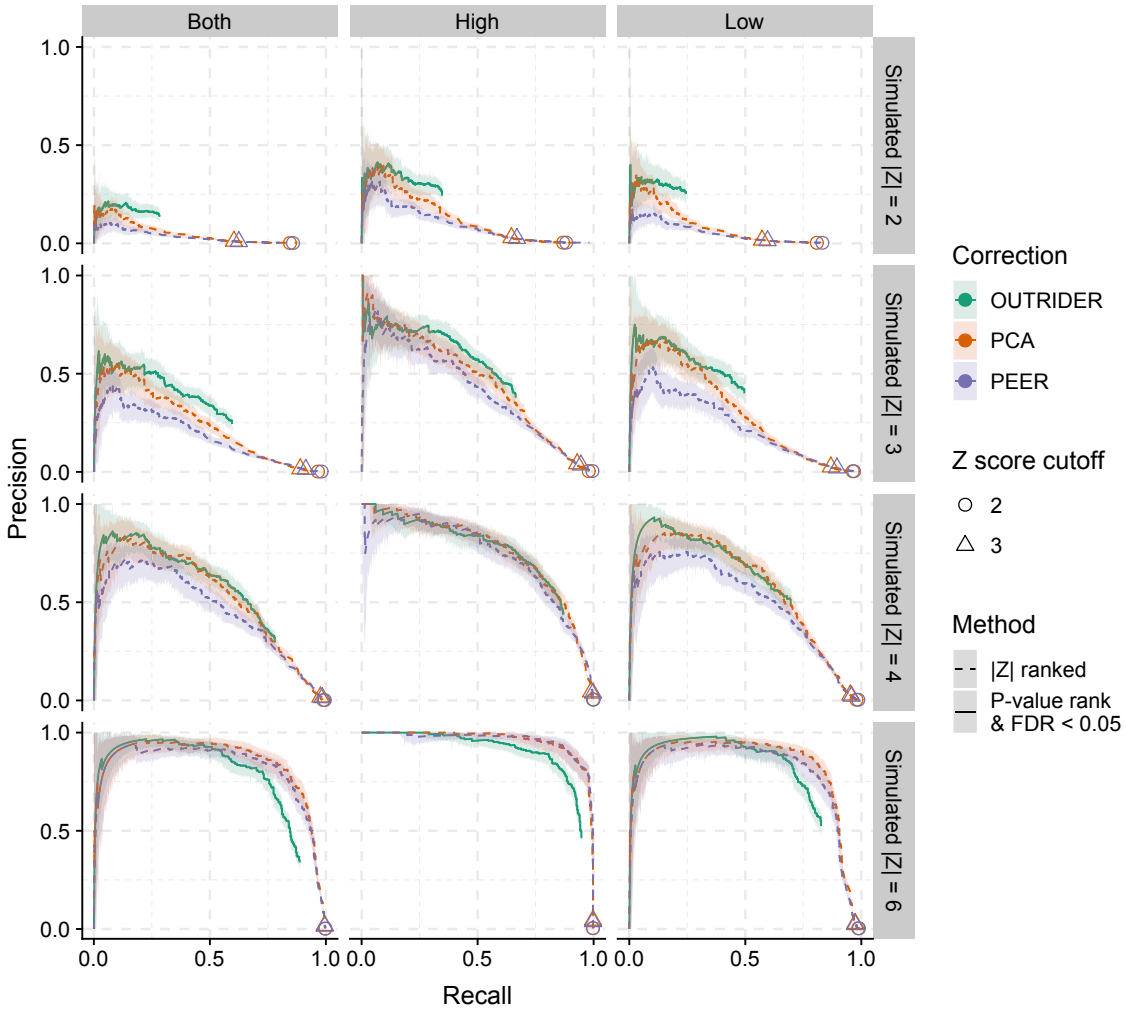

Figure 4. Outlier-Detection Benchmark
The proportion of simulated outliers among reported outliers (precision) plotted against the proportion of reported simulated outliers among all simulated outliers (recall) for increasing p values up to FDR < 0.05 (OUTRIDER) or decreasing absolute *Z* scores (PCA and PEER). Plots are provided for four simulated amplitudes (by row with simulated absolute *Z* scores of 2, 3, 4, and 6 from top to bottom) and for three simulation scenarios (by column from left to right: aberrantly high and low counts, aberrantly high counts, and aberrantly low counts). The read counts were controlled for gene covariation with OUTRIDER (green), PCA (orange), or PEER (blue). The ranking of outliers was bootstrapped to yield 95% confidence bands.

additional layers could be employed to capture nonlinear relationships. However, the analysis of correlations post-control did not suggest the need for a more complex autoencoder. This is consistent with the study of Way and Greene, who modeled covariations in RNA-seq samples by using a single-layer autoencoder.[15] Independently of the way the counts are controlled, OUTRIDER offers functionality for finding the optimal encoding dimension by using a modeling scheme based on corrupted count data. An advantage of this hyperparameter optimization is that no manual intervention is needed.

Current standard methods used to control for covariations of RNA-seq data across individuals include PEER[10,34,35] or PCA.[36] Both approaches assume the input data to be log-normal distributed, which is suboptimal for count data. To directly work on counts instead of transformed counts, we introduced OUTRIDER, which uses the NB distribution, a more suitable distribution for count data. On simulated counts, we observed better inference of the expected counts by using OUTRIDER than by using either PEER or PCA on log-transformed counts. This was especially true for genes with low expression and for underexpressed outliers. Altogether, this resulted in better rankings of outliers and a significantly improved enrichment of rare variants among expression outliers detected in RNA-seq read counts. This improved model for RNA-seq read counts could potentially boost studies that are distinct from outlier calling and rely on controlling for covariations. In particular, mapping of expression quantitative loci could be attempted but was not investigated in this study. Also, one could in principle extend OUTRIDER to include known confounding covariates, for instance, by adding them along with the latent factors before the decoder layer. However, the robustness and the practical

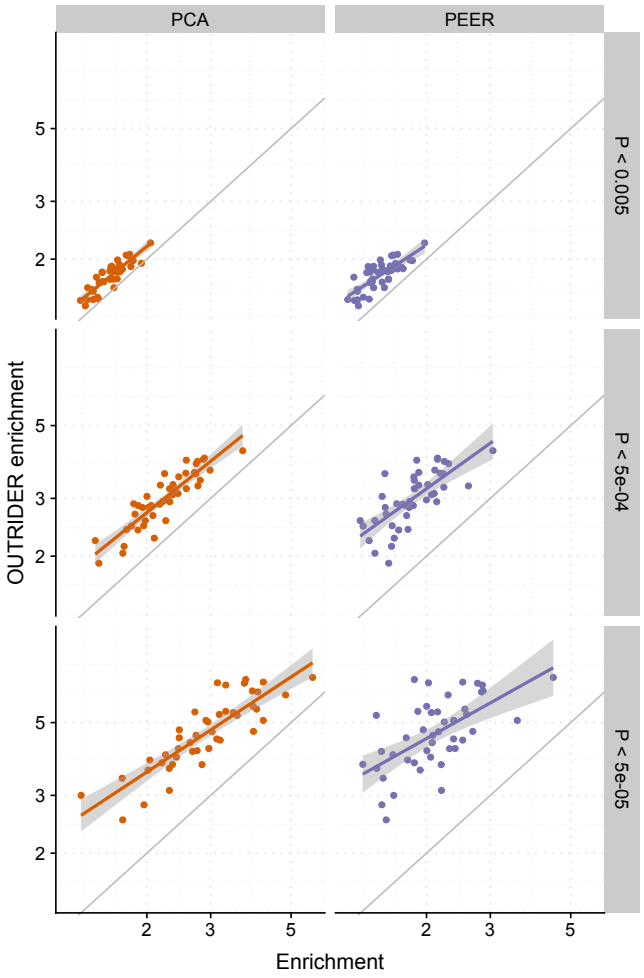

Figure 5. OUTRIDER SNP Enrichment
Enrichment of rare (MAF < 0.05) moderate- and high-impact variants (according to the VEP) computed on genes found to be aberrantly expressed by OUTRIDER is plotted against enrichments computed on genes found to be aberrantly expressed by PCA or PEER for all GTEx tissues with three different p value cutoffs.

added value of such a hybrid approach would have to be investigated.

In differential expression analyses, outlier detections are used for obtaining robust estimators of fold changes and within-group variance. Notably, DESeq2[8] uses Cook's distance to flag extreme observations, whereas edgeR[19] uses Pearson residuals to downweight the impact of extreme observations on the model. This idea is distinct from our aim of assessing the significance of outliers. Nonetheless, these or similar robust estimators[37] could be incorporated into OUTRIDER to improve the estimation of expected counts. This would, however, come with the disadvantage of adding more parameters and a certain degree of circularity.

We have not addressed the handling of replicate samples because we do not expect them to be performed by default in diagnostic settings. The reason for this is that expression outliers are events that show strong effects; therefore, replicates are not essential for detecting these types of events. If a putative disease-causing event, such as an aberrantly expressed gene, is detected, follow-up experiments involving assays complementary to RNA-seq are preferred over replicates to establish the functional link of the event to the disease.[1,5] In contrast, if an RNA-seq sample is suspected to have a technical problem, a new library can be prepared, and the new data are substituted for the former. Neither of these situations results in replicate samples. When replicates are available, users can exclude the replicated samples from the fit to not lower the specificity. Afterward, they can combine the p values of replicate samples by using Fisher's method of combining p values[38] by assuming independence of the read counts conditioned on the expected means predicted by the autoencoder. The same strategy could be applied for pseudo-replicate samples, such as affected individuals of the same family. More elaborated statistical tests have been designed to leverage family structures in normal population studies.[34,39] These methods are based on the normal distribution and log-transformed counts. Our comparison to PEER and PCA suggests that these methods could be improved by a count distribution such as the NB.

Another related issue is the modeling of multiple samples from the same individual. In analyzing GTEx samples, Li et al. have shown that some outliers are shared across multiple tissues for a given individual.[10] If tissues are fitted jointly, it can be difficult to detect outliers shared across tissues because they might be modeled as expected covariation by the autoencoder and because the large number of outlier data points could lead to a poor fit of the NB distribution. In a rare-disease diagnostic setting, we do not expect a large number of tissues to be available per individual. To study the GTEx data, we suggest following the strategy of Li et al.,[10] i.e., to fit a model per tissue and summarize results across tissues with a meta-analysis strategy. We have performed the tissue-wise p value computations with OUTRIDER and provide them on our website (Web Resources).

In general, the autoencoder controlling scheme and the count modeling approach benefit from additional sequencing data; the more data of unrelated individuals that can be combined, the better the estimation of the typical patterns within a population will be. This holds true when the overall data are equally distributed across population structures or sequencing protocols because each sample is assumed to be an independent representative of the whole population. This assumption was partially violated in this study because RNA-seq datasets such as GTEx comprise >85% individuals of European descent.[21] Such overrepresentation of a given population in the dataset is disadvantageous in general, and additional samples from underrepresented groups would be especially beneficial. More testing is needed for assessing whether our strategy for controlling counts can control for different data sources, including data from multiple sequencing platforms or control datasets. The ability to control for different protocols would enable count data to be

combined from multiple sources. This would allow studies with a few samples to merge their results with sources such as the publicly available GTEx dataset.[21] Currently, the best practice is to use the same cell-handling and library-preparation protocol that reduces the analyzable dataset and therefore limits the statistical power. According to a power analysis, 50–60 samples were enough for OUTRIDER to recall most of the known pathogenic events, which are mainly a complete loss of expression. To allow for the detection of more subtle outliers, a larger cohort size is recommended.

The initial aim of developing OUTRIDER was to create a framework for detecting expression outliers for RNA-seq data in a rare-disease diagnostic setting. OUTRIDER will be useful for the identification of potentially disease-causing genes in individuals for whom current methods, such as WES and WGS, only provide variants of unknown significance. However, our approach is not restricted to such data or experiments. Our re-analysis of the tissues of the GTEx dataset[10] indicates that OUTRIDER can provide a more accurate set of expression outliers than existing methods also for studying expression outliers in normal populations. In principle, OUTRIDER could model any count data derived from next-generation sequencing. Our approach could also be applied to data such as DNA accessibility from ATAC-seq reads. In this case, promotor regions or enhancers would be used as features instead of gene bodies. Finally, the methodology of OUTRIDER could be adapted to detect splicing outliers or outliers in proteomics or metabolomics.

Supplemental Data

Supplemental Data include 11 figures and Supplemental Material and Methods and can be found with this article online at https://doi.org/10.1016/j.ajhg.2018.10.025.

Acknowledgments

We thank Gökcen Eraslan for fruitful and inspiring discussions about the usage of autoencoders on count data and Daniel MacArthur for clarifications about the Cummings et al. study.[6] This study was supported by the German Bundesministerium für Bildung und Forschung (BMBF) through the German Network for Mitochondrial Disorders (mitoNET; 01GM1113C to H.P.), the E-Rare project GENOMIT (01GM1207 to H.P.), and the Juniorverbund in der Systemmedizin "mitOmics" (FKZ 01ZX1405A to J.G. and V.A.Y.). A fellowship through the Graduate School of Quantitative Biosciences Munich supports V.A.Y. and Z.A. A.M was supported by a fellowship through the Katholischer Akademischer Ausländer-Dienst. C.M., F.B., V.A.Y., H.P., and J.G. are supported by EU Horizon2020 Collaborative Research Project SOUND(633974). The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health and by the National Cancer Institute, National Human Genome Research Institute, National Heart, Lung, and Blood Institute, National Institute on Drug Abuse, National Institute of Mental Health, and National Institute of Neurological Disorders and Stroke. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on June 12, 2017, under accession number dbGaP: phs000424.v6.p1.

Declaration of Interests

The authors declare no competing interests.

Received: May 24, 2018
Accepted: October 25, 2018
Published: November 29, 2018

Web Resources

GTEx Portal, https://www.gtexportal.org/home
OMIM, http://www.omim.org
OUTRIDER, http://bioconductor.org/packages/OUTRIDER/
OUTRIDER analysis pipeline, https://github.com/gagneurlab/OUTRIDER-analysis/
OUTRIDER supplemental data, https://i12g-gagneurweb.in.tum.de/public/paper/OUTRIDER/

References

1. Taylor, J.C., Martin, H.C., Lise, S., Broxholme, J., Cazier, J.-B., Rimmer, A., Kanapin, A., Lunter, G., Fiddy, S., Allan, C., et al. (2015). Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nat. Genet. *47*, 717–726.
2. Wortmann, S.B., Koolen, D.A., Smeitink, J.A., van den Heuvel, L., and Rodenburg, R.J. (2015). Whole exome sequencing of suspected mitochondrial patients in clinical practice. J. Inherit. Metab. Dis. *38*, 437–443.
3. Wright, C.F., FitzPatrick, D.R., and Firth, H.V. (2018). Paediatric genomics: diagnosing rare disease in children. Nat. Rev. Genet. *19*, 253–268.
4. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
5. MacArthur, D.G., Manolio, T.A., Dimmock, D.P., Rehm, H.L., Shendure, J., Abecasis, G.R., Adams, D.R., Altman, R.B., Antonarakis, S.E., Ashley, E.A., et al. (2014). Guidelines for investigating causality of sequence variants in human disease. Nature *508*, 469–476.
6. Cummings, B.B., Marshall, J.L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A.R., Bolduc, V., Waddell, L.B., Sandaradura, S.A., O'Grady, G.L., et al.; Genotype-Tissue Expression Consortium (2017). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci. Transl. Med. *9*, 1–25.
7. Kremer, L.S., Bader, D.M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., Haack, T.B., Graf, E., Schwarzmayr, T., Terrile, C., et al. (2017). Genetic diagnosis of Mendelian disorders via RNA sequencing. Nat. Commun. *8*, 15824.
8. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.
9. Whitaker, L. (1914). On the Poisson law of small numbers. Biometrika *10*, 36–71.
10. Li, X., Kim, Y., Tsang, E.K., Davis, J.R., Damani, F.N., Chiang, C., Hess, G.T., Zappala, Z., Strober, B.J., Scott, A.J., et al.;

GTEx Consortium; Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI; Biospecimen Collection Source Site—RPCI; Biospecimen Core Resource—VARI; Brain Bank Repository—University of Miami Brain Endowment Bank; Leidos Biomedical—Project Management; ELSI Study; Genome Browser Data Integration &Visualization—EBI; and Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz (2017). The impact of rare variation on gene expression across tissues. Nature *550*, 239–243.

11. Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat. Protoc. *7*, 500–507.
12. Lecun, Y. (1987). Modeles connexionnistes de l'apprentissage (connectionist learning models). PhD thesis (Université Pierre et Marie Curie).
13. Bourlard, H., and Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. Biol. Cybern. *59*, 291–294.
14. Hinton, G.E., and Zemel, R.S. (1994). Autoencoders, minimum description length and Helmholtz free energy. In Advances in Neural Information Processing Systems 6, J.D. Cowan, G. Tesauro, and J. Alspector, eds. (Morgan-Kaufmann), pp. 3–10.
15. Way, G.P., and Greene, C.S. (2018). Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. Pac. Symp. Biocomput. *23*, 80–91.
16. Eraslan, G., Simon, L.M., Mircea, M., Mueller, N.S., and Theis, F.J. (2018). Single cell RNA-seq denoising using a deep count autoencoder. bioRxiv. https://doi.org/10.1101/300681.
17. Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th International Conference on Machine Learning (International Machine Learning Society), pp. 1096–1103.
18. 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.
19. Zhou, X., Lindsay, H., and Robinson, M.D. (2014). Robustly detecting differential expression in RNA sequencing data using observation weights. Nucleic Acids Res. *42*, e91.
20. Barnett, V., and Lewis, T. (1974). Outliers in statistical data (Wiley).
21. GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648–660.
22. Casper, J., Zweig, A.S., Villarreal, C., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Karolchik, D., et al. (2018). The UCSC Genome Browser database: 2018 update. Nucleic Acids Res. *46* (D1), D762–D769.
23. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. *22*, 1760–1774.
24. Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.
25. Wold, H. (1966). Estimation of Principal Components and Related Models by Iterative Least Squares (Academic Press).
26. Oba, S., Sato, M.A., Takemasa, I., Monden, M., Matsubara, K., and Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. Bioinformatics *19*, 2088–2096.
27. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R.B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics *17*, 520–525.
28. Byrd, R.H., Lu, P., Nocedal, J., and Zhu, C. (1995). A Limited Memory Algorithm for Bound Constrained Optimization. SIAM J. Sci. Comput. *16*, 1190–1208.
29. Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Ann. Stat. *29*, 1165–1188.
30. Stegle, O., Parts, L., Durbin, R., and Winn, J. (2010). A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput. Biol. *6*, e1000770.
31. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
32. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.
33. Zeng, Y., Wang, G., Yang, E., Ji, G., Brinkmeyer-Langford, C.L., and Cai, J.J. (2015). Aberrant gene expression in humans. PLoS Genet. *11*, e1004942.
34. Pala, M., Zappala, Z., Marongiu, M., Li, X., Davis, J.R., Cusano, R., Crobu, F., Kukurba, K.R., Gloudemans, M.J., Reinier, F., et al. (2017). Population- and individual-specific regulatory variation in Sardinia. Nat. Genet. *49*, 700–707.
35. Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al.; Geuvadis Consortium (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.
36. Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature *464*, 768–772.
37. Aeberhard, W.H., Cantoni, E., and Heritier, S. (2014). Robust inference in the negative binomial regression model with an application to falls data. Biometrics *70*, 920–931.
38. Fisher, R.A. (1970). Statistical Methods for Research Workers, Fourteenth Edition (Oliver & Boyd).
39. Li, X., Battle, A., Karczewski, K.J., Zappala, Z., Knowles, D.A., Smith, K.S., Kukurba, K.R., Wu, E., Simon, N., and Montgomery, S.B. (2014). Transcriptome sequencing of a large human family identifies the impact of rare noncoding variants. Am. J. Hum. Genet. *95*, 245–256.

The American Journal of Human Genetics, Volume *103*

Supplemental Data

OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data

Felix Brechtmann, Christian Mertes, Agnė Matusevičiūtė, Vicente A. Yépez, Žiga Avsec, Maximilian Herzog, Daniel M. Bader, Holger Prokisch, and Julien Gagneur

1 Supplemental Figures

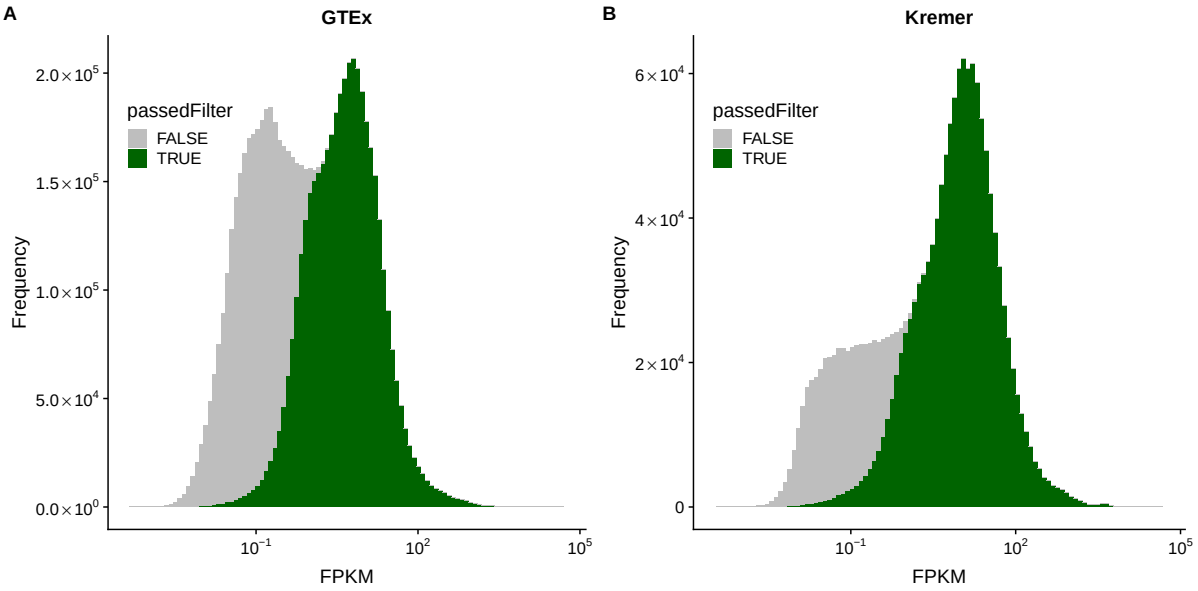

Figure S1: **Filtering of genes.** **(A)** Histogram of the FPKM values for the GTEx data set grouped according to the filter status. Green indicates the genes that passed the filter and gray those that were filtered out. **(B)** Same as A, but for the Kremer data set.

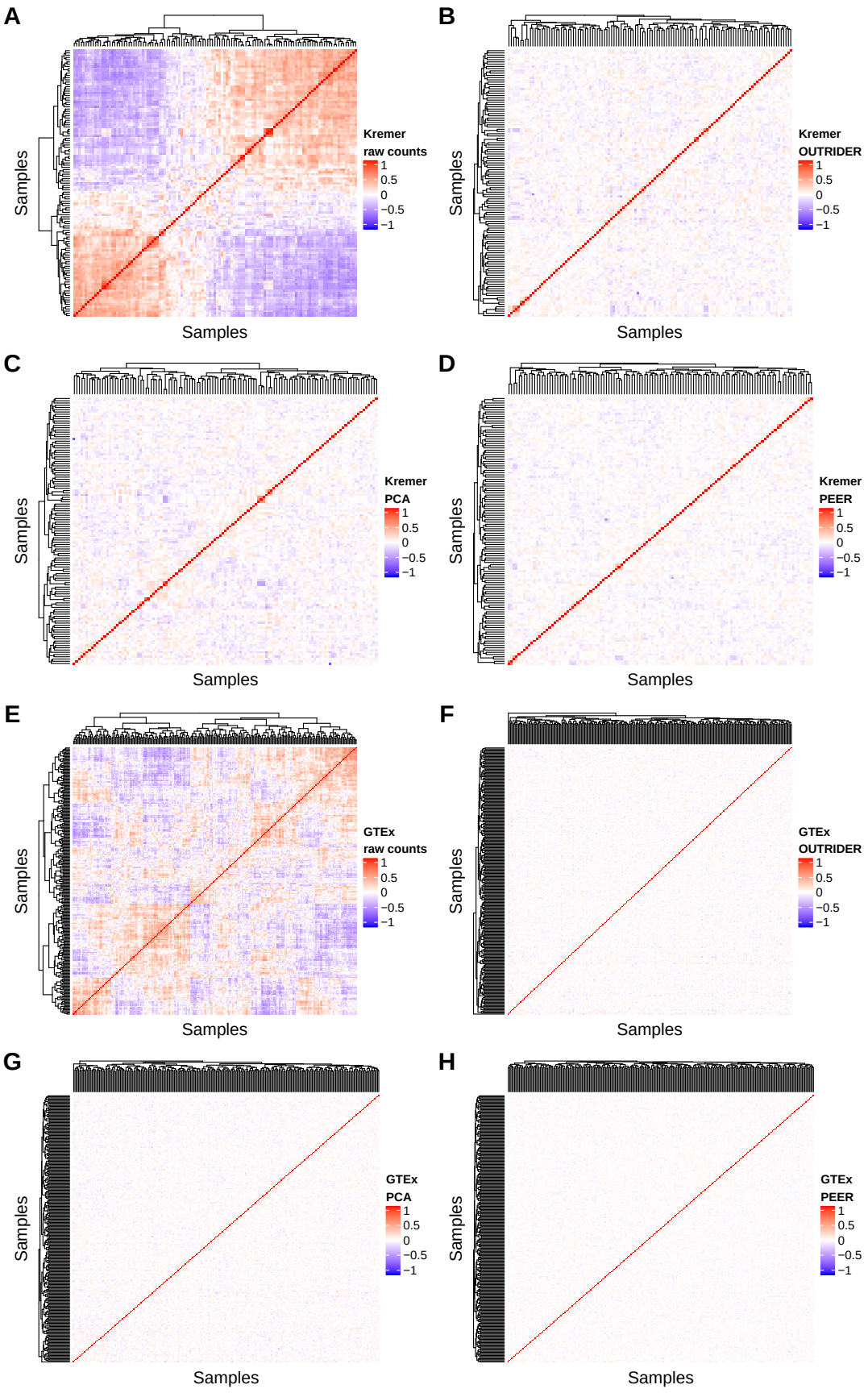

Figure S2: **Controlling count data for covariation.** **(A)** Correlation matrix of row-centered log-transformed read counts for the Kremer data set (119 samples and 10,556 genes). Red indicates a positive correlation and blue a negative correlation. The dendrogram represents the sample-wise hierarchical clustering. **(B, C, D)** Same as in A, but with OUTRIDER, PCA, or PEER controlled read counts. **(E, F, G, H)** Same as in A-D, but for the GTEx data set (249 samples and 17,065 genes).

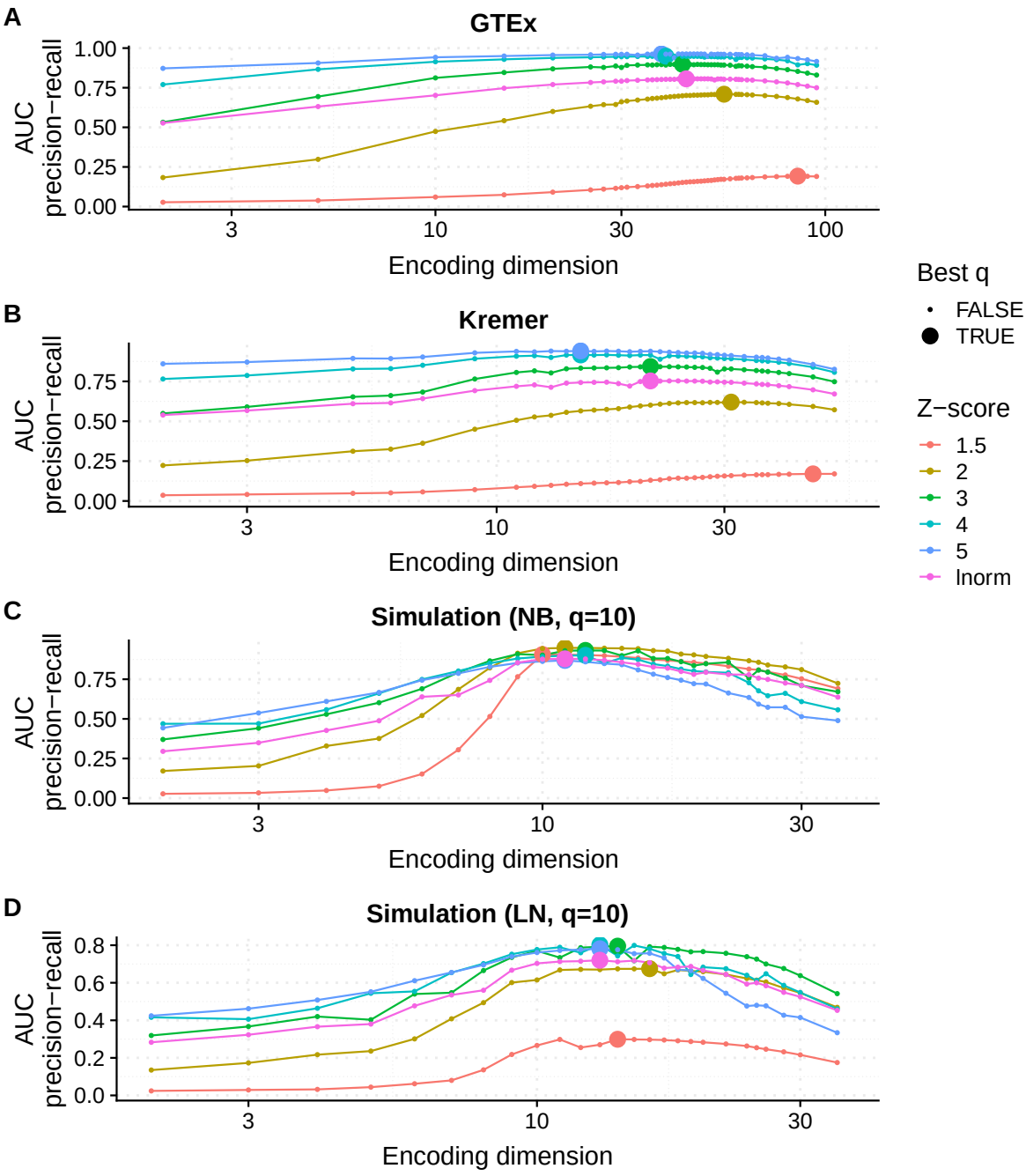

Figure S3: **Fitting the encoding dimension.** **(A, B, C, D)** Area under the precision–recall curve plotted against the autoencoder encoding dimension q for different Z-score amplitudes of corrupted read counts (colors). The pink curve corresponds to Z-score amplitudes sampled from a log-normal distribution (Materials and Methods). Large dots indicate the maximal AUC for a given Z-score amplitude of corrupted read counts. (A) is based on GTEx, (B) on Kremer, (C) on a simulation data set with a latent space of dimension 10 and counts drawn from a negative binomial distribution, and (D) on the same simulated latent space, but with counts drawn from a log-normal distribution and rounded to the nearest integer (Materials and Methods). The optimal encoding dimension was obtained using the log-normally distributed Z-score injection scheme (45, 21, 11, and 13 for GTEx, Kremer, negative binomial and log-normal simulations, respectively).

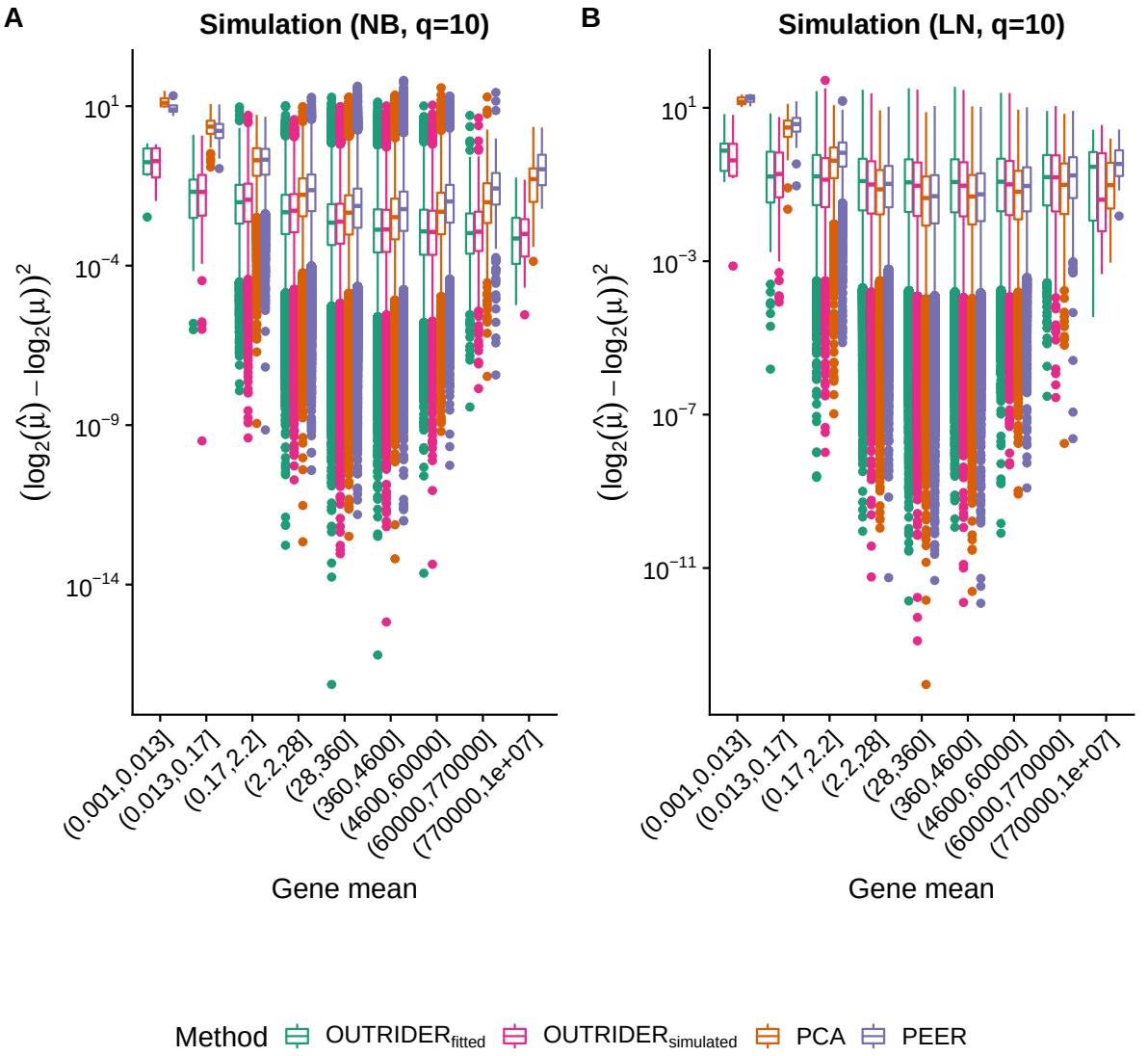

Figure S4: **OUTRIDER recovers expected counts on simulated data.** Boxplots of squared differences between $\log_2$ of fitted means and $\log_2$ of simulated means binned into 9 logarithmically spaced mean gene expression bins for OUTRIDER, PCA and PEER on simulated data. **(A)** Corresponds to the negative binomial simulation scheme as in Figure S3C and **(B)** corresponds to the log-normal simulation scheme as in Figure S3D.

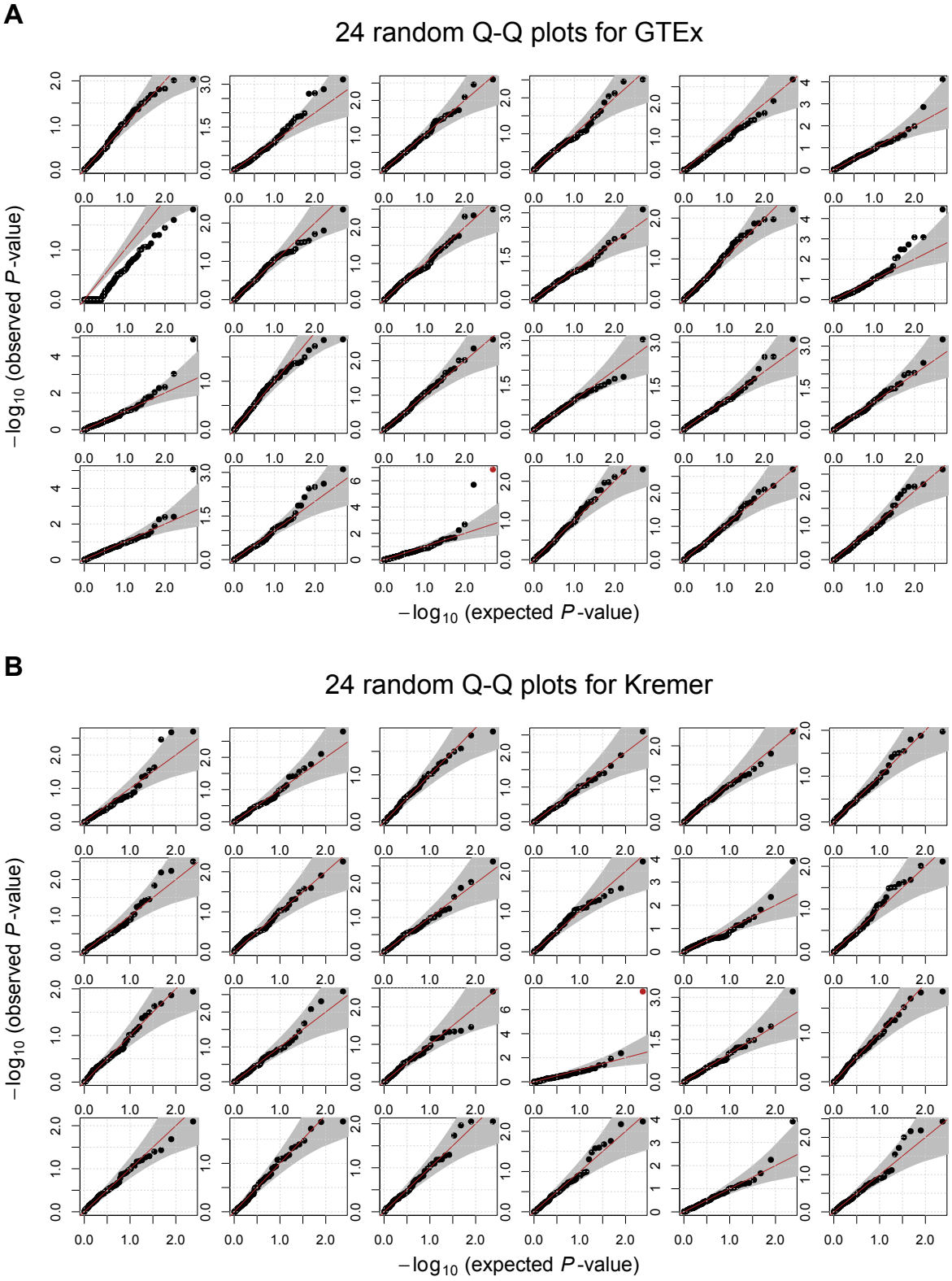

Figure S5: **Negative binomial distribution fits for individual genes.**
(A) Quantile–quantile plots for 24 randomly selected genes from the GTEx data set.
(B) Same as in A but for the Kremer data set.

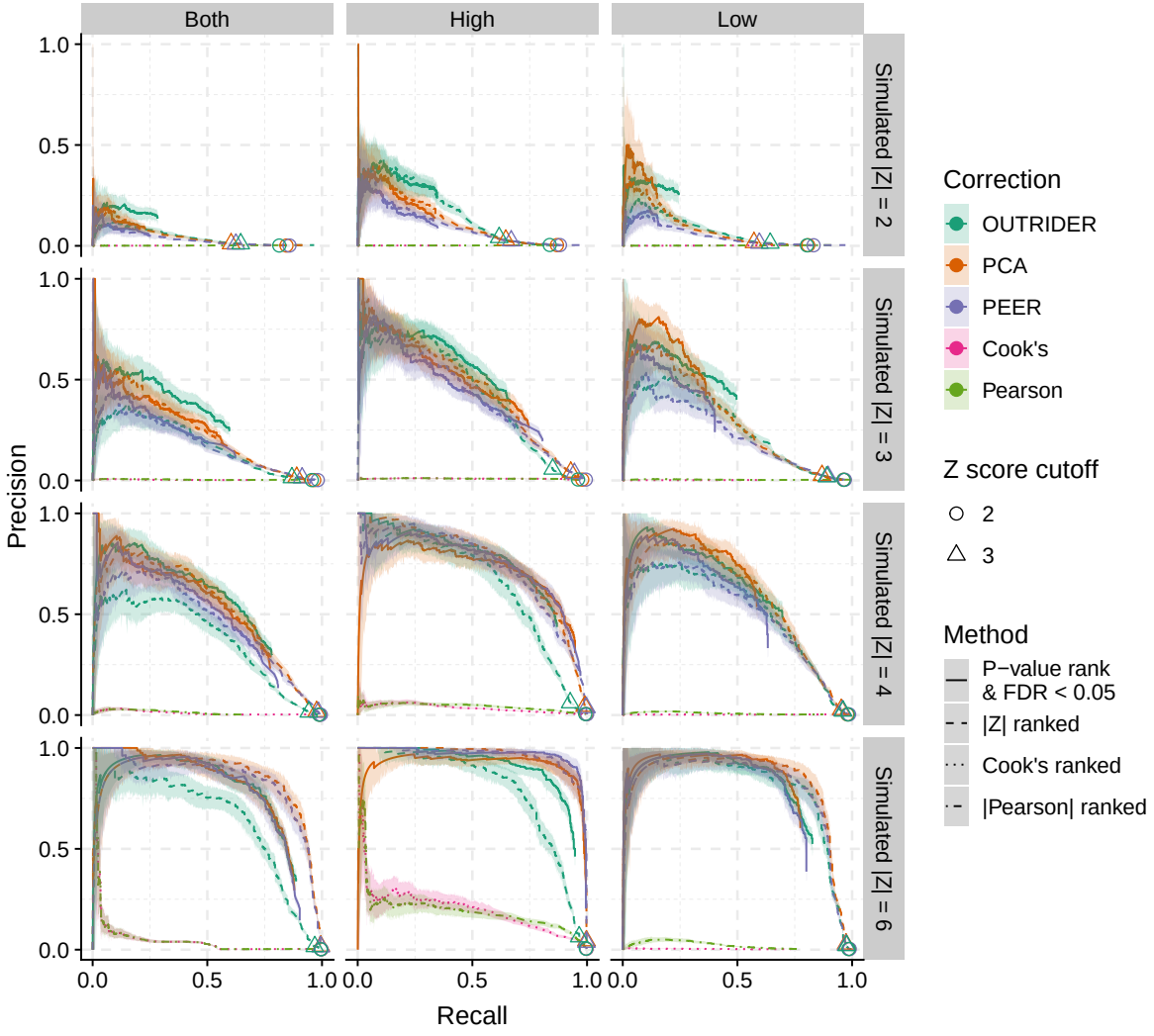

Figure S6: **Outlier detection benchmark in GTEx.** The proportion of simulated outliers among reported outliers (precision) plotted against the proportion of reported simulated outliers among all simulated outliers (recall) for 8 different ranking methods. The 8 ranking methods are OUTRIDER (green solid), PCA (orange solid), and PEER (blue solid) sorted by P-value with FDR < 0.05, OUTRIDER (green dashed), PCA (orange dashed), and PEER (blue dashed) sorted by Z-score, DESeq2 normalization with known covariates sorted by Cook's distance (pink dotted), and DESeq2 normalization with known covariates sorted by absolute value of Pearson residuals (olive green dashed and dotted). Plots are provided for four simulated amplitudes (by row, with simulated absolute Z-scores of 2, 3, 4, and 6, top to bottom, respectively) and for three simulation scenarios (by column for aberrantly high and low counts, for aberrantly high counts, and for aberrantly low counts, left to right, respectively). The ranking of outliers was bootstrapped to obtain 95% confidence areas.

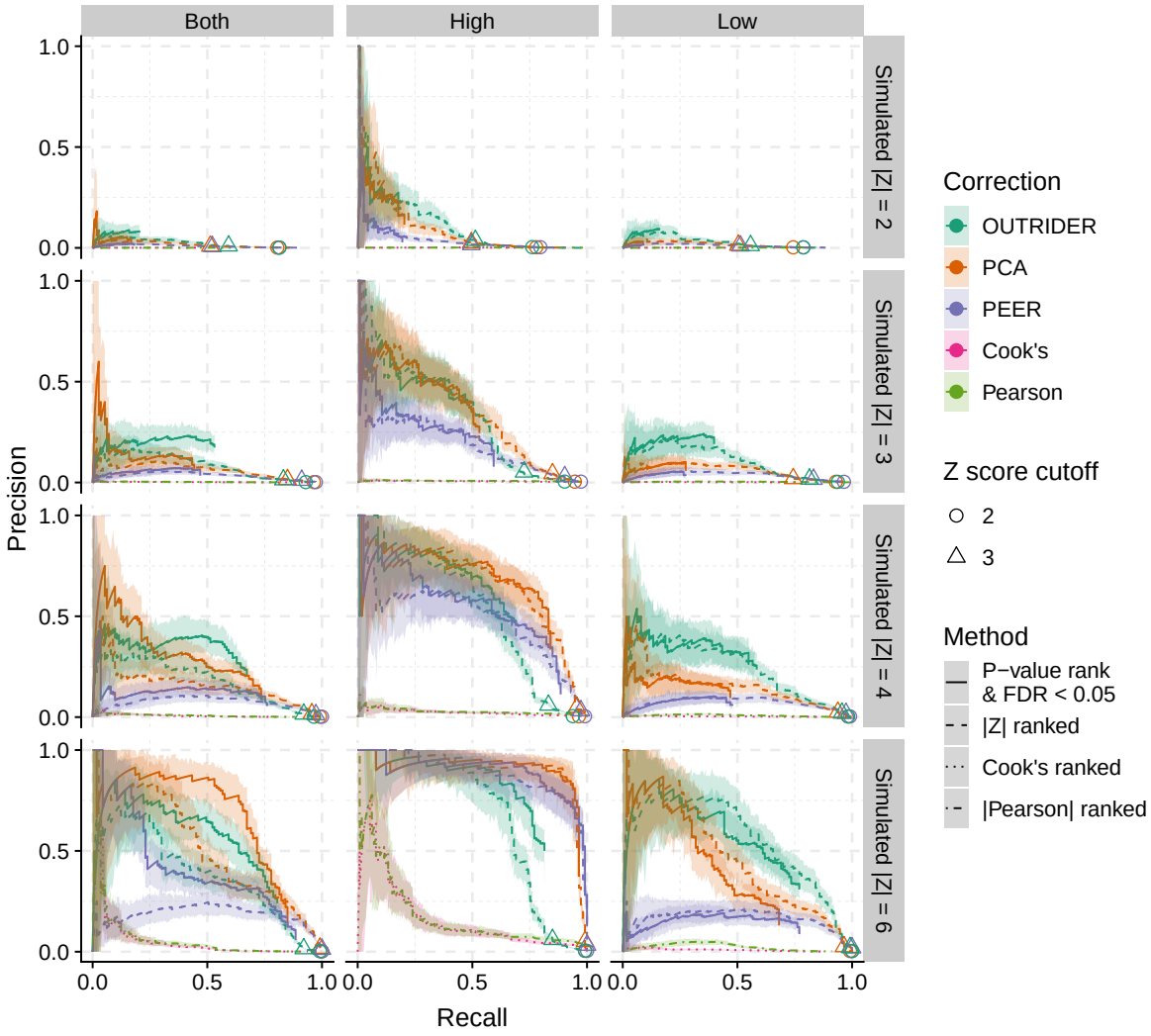

Figure S7: **Outlier detection benchmark in Kremer.** Same as S6 but for the Kremer data set.

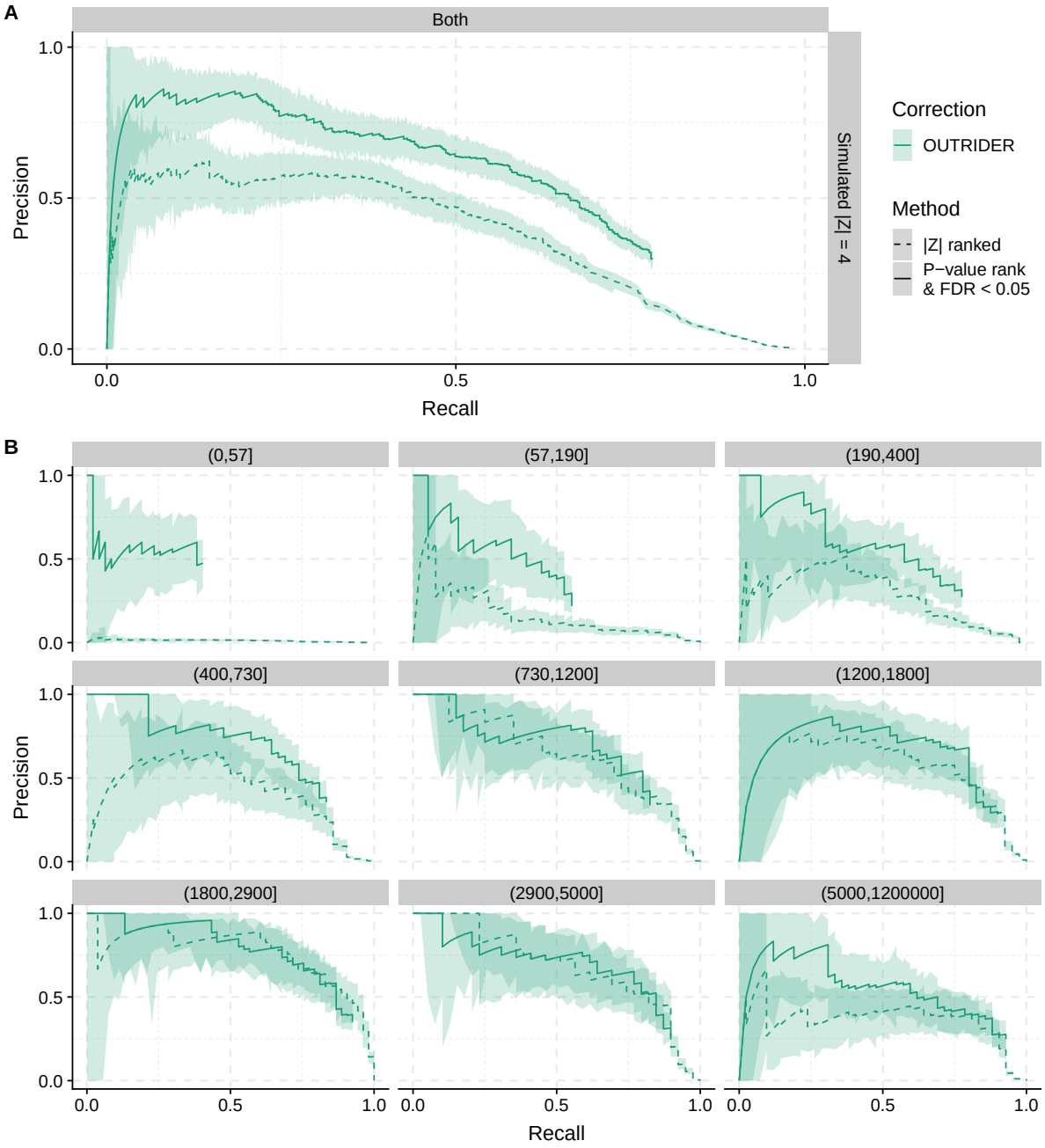

Figure S8: **Expression level dependent recall.** Precision versus recall for artificially injected high and low expression outliers with a Z-score of 4 for OUTRIDER ranked by *P*-values with FDR < 0.05 (solid) and ranked by Z-score (dashed). **(A)** For all the injected outliers. **(B)** Splitted into 9 bins, with equal number of read counts per bin, according to the mean expression level of the genes. Only a small fraction of the injected outliers was significant for the lowest bin, with a mean expression level smaller than 58.

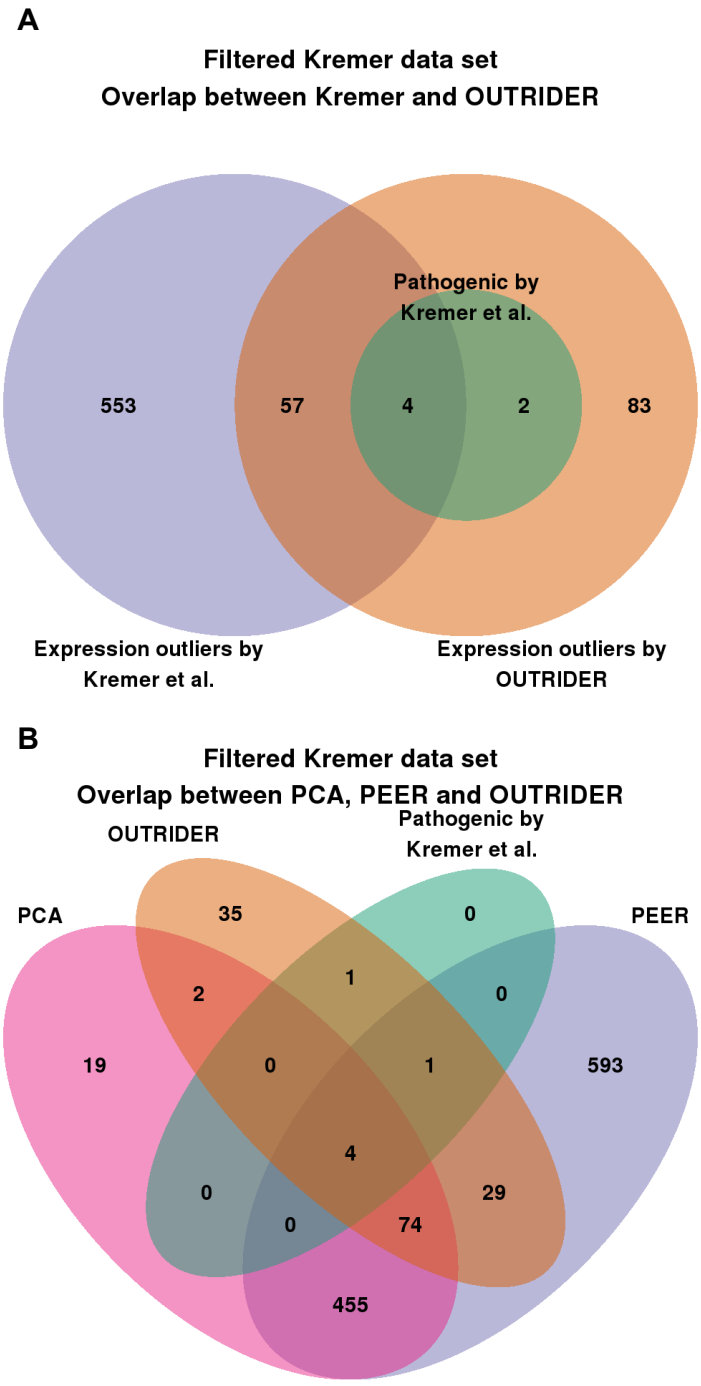

Figure S9: **Benchmark of OUTRIDER using validated genes.** **(A)** Venn diagram of the expression outliers detected by OUTRIDER (orange), expression outliers detected by Kremer et al. (violet), and pathogenic outliers validated by Kremer et al. (green) within the 48 samples of individuals undiagnosed at the start of the Kremer et al. study[1]. **(B)** Venn diagram of expression outliers detected by OUTRIDER (orange), PEER (violet), PCA (pink) and validated pathogenic events (green) within the same data as in A.

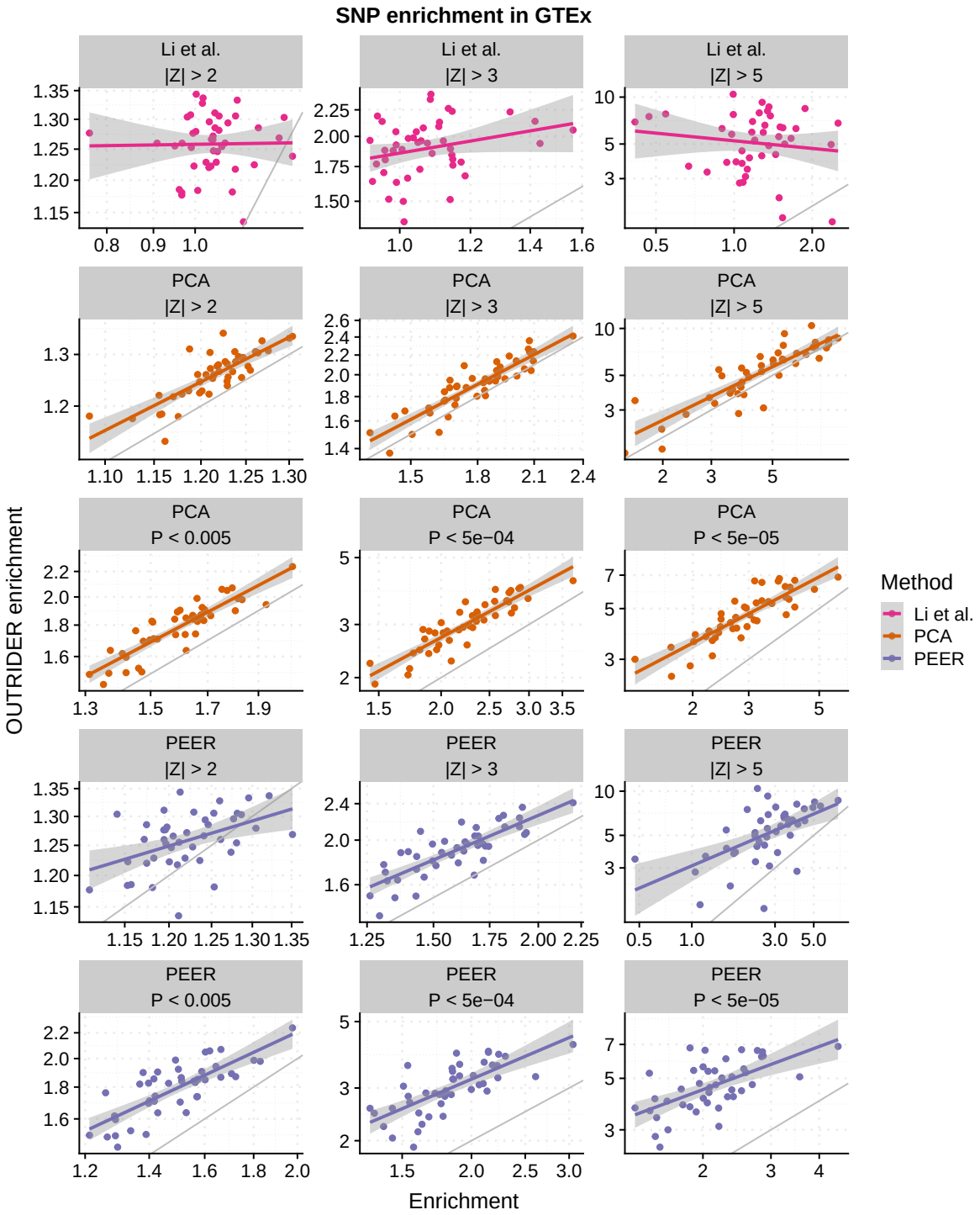

Figure S10: **OUTRIDER SNP enrichment.** Enrichment of rare (MAF < 0.05) moderate and high impact variants (according to VEP) computed on genes found to be aberrantly expressed using OUTRIDER plotted against enrichments computed on genes found to be aberrantly expressed using Z-scores published by Li et al.[2], PCA or PEER for all GTEx tissues using three different P-value or Z-score cutoffs.

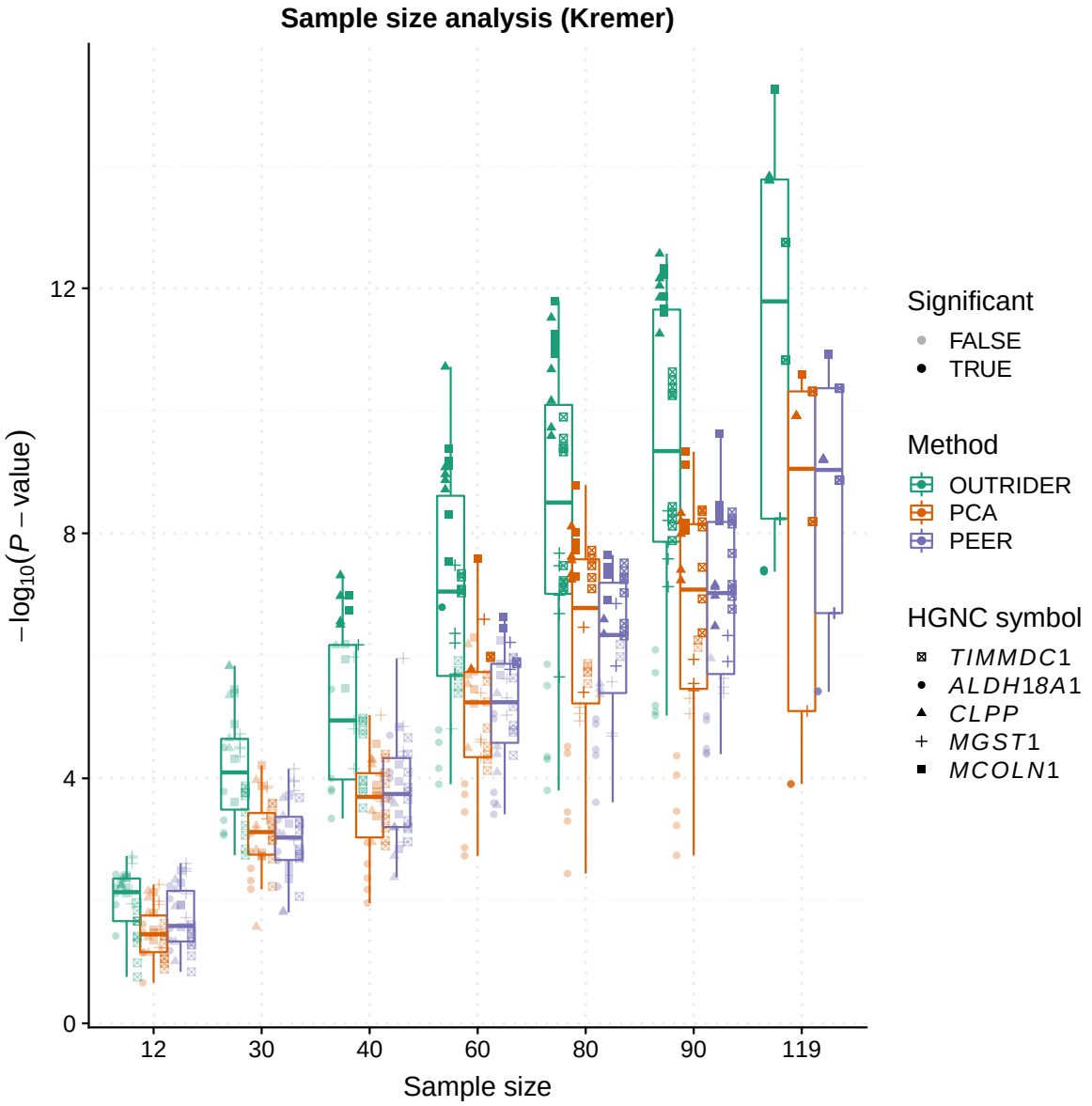

Figure S11: **Sample size analysis.** Negative $\log_{10}$ P-values are plotted against the number of samples in the data set, for 6 pathogenic genes (validated in Kremer et al. [1]). For each data set size, five random sets of samples containing the samples with the known outliers were drawn. Genes that are significant (FDR < 0.05) are marked darker.

2 Supplemental Methods

2.1 Alternative outlier detection methods

In differential expression analyses outlier detections are used to obtain robust estimators of fold changes. Notably, DESeq2[3] uses the Cook's distance to flag extreme observations, while edgeR[4] uses the Pearson residuals to downweight the impact of extreme observations on the model. To benchmark these outlier detection approaches, we calculated the Cook's distance and the Pearson residuals and included them in the benchmark.

In the case of the Cook's distance, we ran a DESeq2 model against known covariates. For GTEx, we used sex, age, and the ischemic time as covariates, while for Kremer, we used sex and the body site inferred from gene expression of the Hox family. After fitting the DESeq2 model, we extracted the Cook's distance from the object. For the Pearson residuals, we fitted the same model with DESeq2 to estimate the mean. The dispersion was estimated with the method of moments provided by DESeq2. For the count of sample i and gene j, the Pearson residuals r_{ij}^{Pearson} was then calculated as:

$$r_{ij}^{\text{Pearson}} = \frac{k_{ij} - \mu_{ij}}{\sqrt{v_{ij}}},$$
$$\text{where } v_{ij} = \mu_{ij} + \alpha_j^{\text{DESeq2}} \mu_{ij}^2,$$

where k_{ij} is the count, μ_{ij} its estimated mean, v_{ij} its estimated variance, and α_j is the gene-wise dispersion parameter (as parameterized and estimated by DESeq2).

2.2 Fitting of the parameters

All notations are introduced in the Materials and Methods section.

Negative Binomial model

We use the following parameterization of the negative binomial distribution:

$$P(k|\mu,\theta) = \frac{\Gamma(k+\theta)}{\Gamma(\theta)k!}\left(\frac{\mu}{\mu+\theta}\right)^k\left(\frac{\theta}{\mu+\theta}\right)^\theta$$

where the variance of the distribution is given by:

$$Var = \mu + \frac{\mu^2}{\theta}$$

Negative log-likelihood

The negative log-likelihood nll of the model is given by:

$$\begin{aligned}\text{nll} = &-\sum_{ij} k_{ij}\log(\mu_{ij}) - \sum_{ij}\theta_j\log(\theta_j) + \sum_{ij}(k_{ij}+\theta_j)\log(\mu_{ij}+\theta_j)\\ &-\sum_{ij}\log(\Gamma(k_{ij}+\theta_j)) + \sum_{ij}\log(\Gamma(\theta_j)k_{ij}!)\end{aligned}$$

For the optimization of the model only the first and third term of the nll need to be considered, as all other terms are independent of $\mathbf{W}_e$ and $\mathbf{W}_d$, yielding the following truncated form of the negative log likelihood:

$$\text{nll}_{\mathbf{W}} = -\sum_{ij} \left[k_{ij} \log\left(\mu_{ij}\right) - \left(k_{ij} + \theta_j\right) \log\left(\mu_{ij} + \theta_j\right)\right] \tag{1}$$

We use L-BFGS to fit the autoencoder model as described in Methods. We implemented the following gradients.

The expectations μ_{ij} are modeled by:

$$\mu_{ij} = s_i e^{y_{ij}}$$

Hence, $\text{nll}_{\mathbf{W}}$ can be rewritten as:

$$\text{nll}_{\mathbf{W}} = -\sum_{ij} \left[k_{ij} \log(s_i) + y_{ij} - \left(k_{ij} + \theta_j\right) \cdot \left(\log(s_i) + y_{ij} + \log\left(1 + \frac{\theta_j}{s_i \cdot e^{y_{ij}}}\right)\right)\right]$$

In the following the y_{ij} are the elements of the $\mathbf{Y}$ defined as:

$$\mathbf{Y} = \mathbf{X}\mathbf{W}_e\mathbf{W}_d^T + \mathbf{b}, \tag{2}$$

where the element (i, j) of the matrix $\mathbf{X}$ is given by: $\log\left(\frac{k_{ij}+1}{s_i}\right) - \bar{x}_j$.

Update of $\mathbf{W}_d$

The updating of the matrix $\mathbf{W}_d$ is performed gene-wise. For each gene, the gene-wise average negative log likelihood is minimized. To not run into convergence issues or numerical instability of the logarithm, we enforce $-700 < y_{ij}$. From Equation 1 and Equation 2, we obtain the gradients:

$$\begin{aligned}
\frac{d\text{nll}}{d\mathbf{W}_e} &= \mathbf{K}^T\mathbf{X}\mathbf{W}_d - \mathbf{L}^T\mathbf{X}\mathbf{W}_d \\
\frac{d\text{nll}}{d\mathbf{W}_d} &= \mathbf{X}^T\mathbf{K}\mathbf{W}_e - \mathbf{X}^T\mathbf{L}\mathbf{W}_e \\
\frac{d\text{nll}}{db_j} &= \sum_i k_{ij} - l_{ij}
\end{aligned}$$

where the components of the matrix $\mathbf{L}$ are computed by:

$$l_{ij} = \frac{(k_{ij} + \theta_j)\mu_{ij}}{\theta_j + \mu_{ij}}$$

References

[1] Kremer, L. S., Bader, D. M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., Haack, T. B., Graf, E., Schwarzmayr, T., Terrile, C. *et al.* (2017). Genetic diagnosis of Mendelian disorders via RNA sequencing. Nature Communications *8*, 15824.

[2] Li, X., Kim, Y., Tsang, E. K., Davis, J. R., Damani, F. N., Chiang, C., Hess, G. T., Zappala, Z., Strober, B. J., Scott, A. J. *et al.* (2017). The impact of rare variation on gene expression across tissues. Nature *550*, 239–243.

[3] 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.

[4] Zhou, X., Lindsay, H., and Robinson, M. D. (2014). Robustly detecting differential expression in RNA sequencing data using observation weights. Nucleic Acids Research *42*.