OUTRIDER: A statistical method for detecting aberrantly expressed genes in RNA sequencing data

Felix Brechtmann,[1,5] Agnė Matusevičiūtė,[1,5] Christian Mertes,[1,5] Vicente A Yépez,[1,2,3] Žiga Avsec,[1,2] Maximilian Herzog,[1] Daniel M Bader,[1,6] Holger Prokisch,[3,4] Julien Gagneur[1,2,*]

[1]Department of Informatics, Technical University Munich, Boltzmannstr. 3, 85748 Garching, Germany

[2]Quantitative Biosciences Munich, Gene Center, Department of Biochemistry, Ludwig-Maximilians Universität München, Feodor-Lynen-Strasse 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 Munich, 13 Ismaninger Str. 22, 81675 München, Germany

[5]These authors contributed equally to this work

[6]Present address: Evotec (München) GmbH, Am Klopferspitz 19a, 82152 Martinsried, Germany

*Correspondence: gagneur@in.tum.de

Abstract

RNA sequencing (RNA-seq) is gaining popularity as a complementary assay to genome sequencing for precisely identifying the molecular causes of rare disorders. An obvious and 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 they 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 co-variation among genes 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 correction of artificially corrupted data. Precision–recall analyses using simulated outlier read counts demonstrated the importance of correction for co-variation and of significance-based thresholds. OUTRIDER is open source and includes functions for filtering out genes not expressed in a data set, for identifying outlier samples with too many aberrantly expressed genes, and for the *P*-value-based detection of aberrant gene expression, with false discovery rate adjustment. Overall, OUTRIDER provides a computationally fast and scalable end-to-end solution for identifying aberrantly expressed genes, suitable for use by rare disease diagnostics platforms.

Introduction

Many patients suspected to suffer a Mendelian disorder undergo whole exome or whole genome sequencing. Nonetheless, in most cases no clear pathogenic variant can be pinpointed[1,2]. A possible reason is that the pathogenic variant is regulatory. However, precisely identifying pathogenic regulatory variants is difficult for at least two reasons. First, an individual harbors a very large number of rare non-coding variants, with about 60,000 non-coding single nucleotide variants compared with 475 protein-affecting rare variants per genome (with MAF < 0.005)[3]; and second, our understanding of regulatory sequences is much poorer than our

44 understanding of coding sequences, so the prioritization of regulatory sequences can be
45 difficult.

46 Two recent studies have shown that using RNA sequencing (RNA-seq) to directly investigate
47 gene expression defects in patients' cells provides a promising complementary method for
48 pinpointing pathogenic regulatory defects[4,5]. RNA-seq can help to reveal splicing defects, the
49 mono-allelic expression of heterozygous loss-of-function variants, and expression outliers (i.e.
50 genes aberrantly expressed outside their physiological range)[4,5]. The two studies used different
51 approaches to identify expression outliers. Cummings *et al.*[4] computed Z-scores on the
52 logarithm of gene-length normalized read counts by subtracting the mean count and dividing by
53 the standard deviation. Expression outliers were identified as read counts with an absolute Z-
54 score greater than 3. No test of statistical significance was performed. This expression outlier
55 analysis did not yield any convincing candidates. In contrast, the study by Kremer *et al.*[5]
56 identified 4 out of 6 newly diagnosed individuals as expression outliers. Read count outliers
57 were identified as those with an absolute Z-score greater than 3 and further being statistical
58 significant according to DESeq2, a statistical test originally developed for differential expression
59 analyses[6], which was applied by testing each sample in turn against the rest of the cohort.
60 DESeq2 is based on the negative binomial distribution, which can be used to model
61 overdispersed count data[7]. The application of a statistical test, appropriate for count data, may
62 be one reason for the difference in success rates between the studies. However, the reason for
63 the difference remains unclear because of the relatively small number of diagnosed cases, the
64 absence of ground truth, and the lack of a direct comparison between the two approaches
65 based on the same data.

66 The two studies did not only differ in whether or not a statistical test was applied but also
67 differed in the way the data was corrected for confounders. Cummings *et al.*[4] used RPKM

68 (reads per kilobase per million mapped reads) expression values. These control for variations in
69 sequencing depth but not for other sources of co-variation among the read counts. Controlling
70 for further sources of co-variation is important because the identification of a gene as aberrantly
71 expressed depends on the context, for example the sex of the donor. Genes encoded on the Y
72 chromosome are not present, and thus not expressed, in women. However, in men, loss of the
73 expression of a Y chromosome-encoded gene can be an aberrant expression event. Hence, not
74 taking the sex of the donor into account would not allow for the detection of aberrantly silenced
75 Y chromosome-encoded genes in males. While the sex of the donor is usually available and can
76 be controlled for, other contexts for gene expression, such as the exact tissue of origin of the
77 sample, the sample's cell type composition, the genetic background, and technical biases, may
78 not be known *a priori*, yet causing similar but less intuitive variations. Kremer *et al*.[5] corrected
79 expression levels for sex, biopsy site as inferred from the *HOX* gene set, and common technical
80 sources of variation, which were identified by visual inspection of a hierarchical clustering of the
81 samples. In a study that identified expression outliers, although not for the diagnosis of rare
82 diseases, Li *et al*.[8] corrected for sex and the top three genotype principal components, as well
83 as for hidden confounding effects estimated using the probabilistic estimation of expression
84 residuals (PEER) method[9]. However, the algorithms controlling for co-variations in RNA-seq
85 read count data used in the studies of Li *et al*.[8] and Kremer *et al*.[5] were not assessed or tuned to
86 detect aberrantly expressed genes.

87 Here, we introduce OUTRIDER (OUTlier in RNA-seq fInDER), an algorithm that provides a
88 statistical test for outlier detection in RNA-seq samples while controlling for co-variations among
89 the gene read counts. Its parameters are optimized automatically for correcting read counts
90 corrupted *in silico*. The modeling of co-variation is performed by an autoencoder that controls for
91 read count variations caused by factors not known *a priori*. Autoencoders were introduced to
92 find low-dimensional representations of high-dimensional data through a hidden layer[10–12]. They

93 have been shown to be useful for extracting meaningful biological features from RNA-seq
94 data[13]. A subclass of autoencoders, the so-called denoising autoencoders, are used to
95 reconstruct corrupted high-dimensional data through exploiting correlations in the data[14]. In
96 OUTRIDER, the autoencoder approach is used to correct for the common co-variation patterns
97 among genes. In this article we describe the OUTRIDER algorithm, its implementation, as well
98 as its performance and results on two data sets.

99 Material and Methods

100 Data sets

101 The read counts for the rare disease cohort were downloaded from Supplementary Data 1
102 published as part of the study by Kremer *et al*.[5]
103 (https://www.nature.com/articles/ncomms15824). GTEx read counts were obtained from the
104 GTEx Portal (V6p counted with RNA-SeQCv1.1.8, https://www.gtexportal.org/home/datasets)[15].
105 We considered all the expressed genes in each data set. Read counts for the Kremer *et al*. data
106 set were computed according to the UCSC annotation build hg19[16], considering the full gene
107 body. In contrast, GTEx is based on the Gencode v19 annotation[17], and the read count of a
108 gene is defined as the number of paired-end read pairs overlapping exons of that gene only.
109 FPKM (fragments per kilobase per millions of reads) values were obtained using DESeq2[6],
110 where the gene length was defined as the aggregated length of all the exons. We then filtered
111 for genes for which at least 5% of the samples had a FPKM value greater than 1 (Figure S1A
112 and B). We further filtered out samples with a sequencing depth less than the mean sequencing
113 depth minus three standard deviations (Figure S1C and D).

Statistical model

We assume that the read count k_{ij} of gene $j = 1, \dots, p$ in sample $i = 1, \dots, n$ follows a negative binomial (NB) distribution[18,19], with gene-specific dispersion parameter θ_j, and mean μ_{ij} equal to the product of a control factor c_{ij} and a gene-wise adjustment factor a_j:

$$P(k_{ij}) = \mathrm{NB}(k_{ij} | \mu_{ij} = a_j \cdot c_{ij}, \theta_j). \quad (1)$$

Both c_{ij} and a_j are required to be >0.01 during fitting because the NB distribution is not defined for $\mu \leq 0$. The correction factor c_{ij} is the product of the sample-specific size factor s_i, and the exponential of the sum of the co-variation factor y_{ij} and the mean log read count per gene $\overline{x_j} = \underset{i}{\mathrm{mean}}\left(\log\left(\frac{k_{ij}+1}{s_i}\right)\right)$:

$$c_{ij} = s_i \cdot \exp(y_{ij} + \overline{x_j}) \quad (2)$$

The size factors s_i capture variations within the sequencing depth; they are robustly estimated as the median of the ratios of the gene read counts to their geometric means[18]. The co-variation factors y_{ij} capture co-variations across genes; they are modeled using an autoencoder of encoding dimension $q < p$. Specifically,

$$\mathbf{y}_i = \mathbf{h}_i \cdot \mathbf{W}^T + \mathbf{b} \quad (3)$$

$$\mathbf{h}_\mathrm{i} = \mathbf{x}_i \cdot \mathbf{W}, \quad (4)$$

where the $p \times q$ matrix $\mathbf{W}$ is the encoding matrix and its transpose the decoding matrix, the *q*-vector $\mathbf{h}_\mathrm{i}$ is the encoded representation and the *p*-vector $\mathbf{b}$ a bias term. The input to the autoencoder, $\mathbf{x}_i$, is a vector of the log counts divided by the size factors centered on the gene mean:

$$x_{ij} = \log\left(\frac{k_{ij}+1}{s_i}\right) - \overline{x_j}, \quad (5)$$

where we add one to prevent computing the logarithm of zero. In the following, we call the combination of Equations (2 - 5) the autoencoder, or short: $c_{ij} = \mathrm{AE}(k_{ij})$.

Fitting the autoencoder parameters

To fit the autoencoder, we set the adjustment factors a_j to 1 because the goal of the autoencoder is to model as much as possible of the read count co-variation. We found that joint estimations of the dispersion parameter and the autoencoder led to poor results, possibly because one compensated for the other one. We therefore set the initial dispersion parameters for all the genes to a common value $(\theta = 25)$ that is arbitrary yet realistic. This value corresponds to a coefficient of variation for large read count expectations of 20%. The coefficient is also known as the biological coefficient of variation[19]. A biological coefficient of variation of 20% is typically seen for RNA-seq data sets of human cell lines with identical genetic background[20].

For a given encoding dimension q, adjustment factors a_j, and dispersion parameters θ_j, the parameters of the autoencoder $\mathbf{W}$ and $\mathbf{b}$ are fitted by maximizing the likelihood (Equation 1) using the L-BFGS algorithm[21] as provided by the R optimization function *optim()*. A detailed derivation of the gradient is given in the supplemental methods.

Fitting the encoding dimension

The optimal encoding dimension is obtained by assessing the autoencoder's performance in correcting corrupted data. We artificially introduce corrupted read counts randomly with a probability of 10^{-2} by shifting the true read counts 2 standard deviations on a log scale randomly up or down. The introduced values are rounded to the nearest integer.

With C denoting the set of index pairs of corrupted read counts, and $\tilde{\mathbf{k}}_i$ the vector of read counts for sample *i* after the injection of the corrupted read counts, an evaluation function was defined as:

$$\text{evaluation} = -\frac{1}{|C|}\sum_{(i,j)\in C}\log\left(\text{NB}\left(k_{ij}\middle|\mu_{ij} = a_j \cdot \text{AE}\left(\widetilde{\mathbf{k}}_i\right), \theta\right)\right) \quad (6)$$

The optimal encoding dimension for this evaluation function can be selected from the integer range [1;100] by computing the loss for each value for a given $\theta = 25$.

Estimation of the dispersion and adjustment parameters

Once the autoencoder and thus the control factors c_{ij} is fitted, the adjustment a_j and dispersion θ_j are obtained by maximizing the log-likelihood of Equation (1). To estimate the parameters,

the negative log-likelihood is minimized for each gene independently using BFGS optimization, as implemented in the *optim()* function of R[22–25]. Initial values are estimated as $\hat{a}_j = \frac{1}{p}\sum\frac{k_{ij}}{c_{ij}}$ for the adjustment and $\hat{\theta}_j = \frac{\hat{\mu}_j^2}{\hat{\sigma}_j^2 - \hat{\mu}_j}$ for the dispersion if $\hat{\sigma}_j^2 > \hat{\mu}_j$, or 1 otherwise.

P-value computation

After determining the fit parameters, the algorithm computes two-sided *P*-values using the following equation:

$$P_{ij} = 2 \cdot \min\left\{\frac{1}{2}, \sum_{k=0}^{k_{ij}} NB\left(k_{ij}\middle|\mu_{ij}, \theta_j\right), 1 - \sum_{k=0}^{k_{ij}-1} NB\left(k_{ij}\middle|\mu_{ij}, \theta_j\right)\right\}. \quad (7)$$

The term ½ is included to handle cases when both other terms exceed ½, 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 correction by the autoencoder. The computed *P*-values can therefore be correlated. Multiple testing correction was performed using the Benjamini–Yekutieli false discovery rate (FDR) method, which holds under positive dependence[26].

Z-score computation

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

181 $$Z_{ij} = \frac{l_{ij} - \mu_j^l}{\sigma_j^l} \quad (8)$$

182 where l_{ij} is the $\log_2$ corrected count calculated as $l_{ij} = \log_2\left(\frac{k_{ij}+1}{c_{ij}}\right)$, σ_j^l is the standard deviation
183 of l_{ij} for gene *j*, and μ_j^l is the mean of l_{ij} for gene *j*.

184 **Controlled counts**

185 To obtain comparable counts across samples after controlling for various effects, we compute
186 the following controlled counts: $k_{ij}^c = \frac{k_{ij}+1}{c_{ij}}\overline{c_i}$. These are all of the same magnitude, independent
187 of the control method applied. When c_{ij} are size factors, the mean of all c_{ij} is approximately 1
188 and the count k_{ij}^c is of the same magnitude as the raw count k_{ij}. When c_{ij} is the estimated
189 mean, this equation provides a similar controlled count.

190 **Benchmark by injection of outliers**

191 To assess the sensitivity and specificity of alternative outlier detection methods, artificial read
192 count outliers were injected with pre-specified amplitudes (Z-scores). This process was
193 separate from the injection of corrupted data described earlier in "Fitting the autoencoder
194 hyperparameters." The outlier injection scheme described in this section was used to
195 independently assess OUTRIDER's performance in comparison with the Z-score approach.

196 To obtain a data set with samples that contained only a few outliers, all GTEx samples from the
197 skin tissue not exposed to the sun were fitted and tested and only the non-outlier samples as
198 described above in which fewer than 0.1% of all the genes were aberrantly expressed were

199 used for the benchmark set. The read counts were then replaced with a probability of 10^{-4} by an
200 outlier read count k_{ij}^{o}, defined as:

201 $$k_{ij}^{o} = \mathrm{round}\left(s_i \cdot 2^{l_{ij}}\right),$$

202 where

203 $$l_{ij} = z^{o} \cdot \mathrm{sd}\left(l_j\right) + \overline{l_j}.$$

204 ### Implementation

205 OUTRIDER is implemented as an R package that is available through GitHub
206 (https://github.com/gagneurlab).

207 ## Results

208 We considered two RNA-seq data sets, which we refer to as the Kremer and GTEx data sets.
209 The Kremer data set contains 119 RNA-seq samples from skin fibroblasts of patients with a
210 suspected rare mitochondrial disorder[5]. This data set was analyzed in a previous study with the
211 systematic effects corrected by manual inspection of sample correlation matrices[5]. In this study
212 4 genes were identified as aberrantly expressed out of 6 pathogenic genes validated by
213 functional assays[5]. This data set therefore served as our benchmark data set for rare disease
214 applications. The GTEx data set contains 271 RNA-seq samples obtained from the not sun
215 exposed skin tissue in the Genotype-Tissue Expression project (GTEx V6p[15]). We focused on
216 these skin samples because the tissue was similar to the tissue of origin of the Kremer data set.

217 The donors of the GTEx samples did not suffer from any condition and were not under
218 treatment. Nevertheless, aberrant gene expression in these samples has been reported[8].

219 For both data sets, we filtered out genes not expressed across the whole data set, resulting in
220 10,559 genes in the Kremer set and 18,250 genes in the GTEx set (Figure S1A and B). In the
221 GTEx set, we additionally filtered out two samples because of a low sequencing depth, resulting
222 in 269 samples (Figure S1C and D). Both data sets exhibited a strong correlation structure with
223 very distinct sample clusters (Figure 2A and Figure S2A). These correlations may have arisen
224 from biological variations such as the sex of the donor, the origin of the tissue, population
225 structure, or from hidden confounders such as poorly understood systematic technical
226 variations[5,8]. Applying the autoencoder on the read counts allowed co-variations to be estimated
227 and corrected for. The dimension of the autoencoder was fitted using a scheme in which
228 artificially corrupted read counts were injected and presented as the input to the autoencoder,
229 maximizing the likelihood of the original, uncorrupted data. Based on the GTEx data set, we
230 obtained 20 as optimal encoding dimension while 13 was optimal for the Kremer data set. We
231 note that encoding dimensions close to the optimum yielded similar results. After the correction
232 was applied, the co-variation clusters disappeared from both data sets (Figure 2B and Figure
233 S2B).

234 **Testing for aberrant expression revealed aberrant samples in both data sets**

235 The OUTRIDER algorithm models RNA-seq read counts as a NB distribution with a mean that is
236 the expected value provided by the autoencoder. Expression outliers are detected as significant
237 deviations of the observed read counts from these expected values, assuming negative
238 binomial distributions with gene-specific dispersions. Quantile–quantile plots for individual genes

239 indicated that OUTRIDER reasonably modeled the count distribution (Figures 3B and D and S3)
240 and that the resulting *P*-values can be used to detect outliers (Figure 3).

241 Globally, at an FDR threshold of 0.05 (Benjamini–Yekutelli method[26]), we detected aberrantly
242 expressed genes in 168 and 63 samples for the GTEx and Kremer data sets, respectively
243 (Figure 4A and B), with 90th percentiles of 10 and 4 aberrant genes per sample. The number of
244 significant outliers exhibited a skewed distribution across samples in both data sets, consistent
245 with the results of Kremer *et al*.[5]. This suggested that some samples differed from others to a
246 considerable extent. This may have been due to strong technical artifacts in the samples or
247 because the samples were from donors that were genetically distant from the others or who had
248 diseases with a severe regulatory impact. Regardless of the cause, it was difficult to narrow
249 down the candidate pathogenic aberrantly expressed genes in these samples to a short,
250 practical list (Figure 4C and D). We therefore introduced a cutoff (number of outliers > 0.1% of
251 expressed genes) to filter out these samples. This resulted in 19 outlier samples (7.1%) in the
252 GTEx data set and 4 outlier samples (3.4%) in the Kremer data set, as well as much improved
253 global quantile–quantile plots (Figure 4E and F).

254 **Recall benchmark**

255 To benchmark OUTRIDER, we injected outlier read counts and tested the fraction that could be
256 recalled. We used the 269 skin tissue samples not exposed to the sun from the GTEx data set,
257 from which we removed 19 outlier samples. Aberrant counts were then injected into this count
258 matrix of 250 samples and 18,250 genes. Using a frequency of 10^{-4}, we injected 434 aberrant
259 counts to create three scenarios: i) all underexpressed, ii) all overexpressed, and iii) 50%
260 overexpressed and 50% underexpressed. Each scenario was repeated for three different
261 simulated amplitudes (with Z-score values of 3, 4, and 6).

262 We then ran OUTRIDER, controlling either for sequencing depth alone or for co-variations using
263 the autoencoder. Thereby, we monitored the recall of injected read count outliers and the
264 precision, i.e. the number of injected outliers among the reported outliers using the different
265 detection methods (Figure 5). Off note, the precision and recall in this setup were
266 underestimates because the original data also contained genuine outliers. We observed that the
267 proposed correction strategy outperformed controlling only for sequencing depth in almost all
268 simulations. The size factor approach performed better only for very extreme overexpressed
269 outliers (|Z| = 6). However, the recall of OUTRIDER with the autoencoder decreased by less
270 than 5% compared to control using sequencing depth alone in this case.

271 The two most commonly used Z-score cutoffs (|2| and |3|[4,8]) recalled almost all the outliers
272 (median=99%) in each simulation, yet at the cost of a high false discovery rate (precision <
273 0.02). In contrast, applying a significance cutoff corrected for multiple testing reduced the rate of
274 false discoveries and hence increased the precision, which confirms the importance of a *P*-
275 value based strategy.

276 In the scenario of overexpressed genes, the precision–recall curves with and without
277 significance overlapped. (Figure 5, middle column). This indicated that Z-scores and *P*-values
278 for high counts gave essentially the same rankings. The advantage of *P*-values is that they
279 provide a principled way to establish a cutoff that takes statistical significance and multiple
280 testing into account. With low gene expression levels, the overlap of the precision–recall curves
281 was less strong (Figure 5, right column). Only genes with a sufficiently high mean expression
282 level (mean count >80 with recall >55% and precision >60%; Figure S4B) yielded significant *P*-
283 values to recall the injected aberrantly low outliers. This is due to the non-negative nature of
284 counts, where even large Z-scores do not result in significant deviations from the mean
285 expression level (Figure S4). This limited the recall because a significant fraction of genes had a

286 mean expression level that was too low. Overall, these results demonstrated that it is important
287 to control for co-expression and to assess significance.

288 Applying OUTRIDER to the Kremer data set and filtering out the outlier samples resulted in a
289 recall of 47 events (66%) identified by Kremer *et al*. (Figure S5) based on the 48 previously
290 undiagnosed samples. Additionally, OUTRIDER detected 33 new expression outliers of which
291 19 were down-regulated. OUTRIDER was able to recall 5 out of the 6 pathogenic events (3
292 expression outliers, 1 mono-allelic expression, 2 splicing defects) validated by Kremer *et al..*
293 While Kremer at al. additionally detected the mono-allelically expressed gene *ALDH18A1* as an
294 expression outlier it was not significant using OUTRIDER. In this case, a compound
295 heterozygous defect was detected in *ALDH18A1*, where only the missense mutation was
296 expressed. Hence, there was a reduction of about 40% of the overall gene expression.
297 Nevertheless, *ALDH18A1* was the gene with the second most extreme *P*-value among the
298 down-regulated genes in this sample. Furthermore, *ALDH18A1* can also be detected with a test
299 for mono-allelic expression[5], arguably the most appropriate aberrant expression category in this
300 case. Two further genes, *CLPP* and *MCOLN1*, diagnosed by Kremer *et al*.[5] by testing for
301 splicing defects, were detected by OUTRIDER as expression outliers. Those cases exhibited
302 reduced gene expression due to severe splicing defects.

303 Discussion

304 We have introduced OUTRIDER, a new end-to-end solution for identifying aberrantly expressed
305 genes within RNA-seq data, controlling for hidden confounders in an automated fashion, and
306 providing statistical significance estimates. OUTRIDER combines an autoencoder, which allows
307 for automatic correction of technical and biological variations among genes, and a statistical test

308 based on a NB distribution. OUTRIDER has three advantages over preceding methods. First, it
309 computes *P*-values, which can be adjusted to control the false discovery rate. Z-score-based
310 approaches lack *P*-values, so the setting of cutoffs is arbitrary. Second, the parameters of the
311 model are automatically fitted evaluating the ability of the model to correct for simulated
312 corrupted read counts. A precision–recall analysis demonstrated the advantage of correcting for
313 co-variations and employing significance-based thresholds. Third, OUTRIDER is implemented
314 and made available as an open source R package (from
315 https://github.com/gagneurlab/OUTRIDER). The package provides plotting functionality and
316 allows full analyses to be made with only a few lines of code. Furthermore, the package comes
317 along with a comprehensive vignette guiding the user through a typical analysis.

318 We implemented OUTRIDER so that it is not restricted to the autoencoder, allowing the outlier
319 test to be used with any other normalization method. In particular, autoencoders with additional
320 layers could be employed to capture nonlinear relationships. However, the analysis of
321 correlations post-control did not suggest the need for a more complex autoencoder. This is
322 consistent with the study of Way and Greene, who modeled co-variations in RNA-seq samples
323 using a single layer autoencoder[13]. Independently of the way the correction is modeled,
324 OUTRIDER offers functionality for finding the optimal parameter set using a modeling scheme
325 based on corrupted count data. An advantage of this hyperparameter optimization is that no
326 manual intervention is needed to find the optimal parameters.

327 Despite controlling for co-variations, some samples exhibited a much larger number of aberrant
328 genes than others in both of the data sets investigated. This is consistent with the findings of
329 Kremer *et al.*[5]. In outlier samples where hundreds of genes are identified as aberrant, it is
330 difficult to draw any conclusion about the cause of such aberrant expression profiles. We
331 therefore label those samples as outlier samples and recommend to investigate them and their

332 global expression profile changes in a separate analysis. Some of these global changes may
333 have a technical basis, such as the two discarded samples from the GTEx data set with
334 aberrantly low coverage, whereas others may reflect the phenotype of the disease. Alternative
335 methods have been developed to detect such samples; these exploit global measures such as
336 the Mahalanobis distance to measure the dissimilarity of a sample from the population and to
337 identify aberrantly expressed gene sets[27,28]. Analysis of these genome-wide regulatory defects
338 could be interesting.

339 We have not addressed the handling of replicate samples because we do not expect them to
340 occur in diagnostic settings. The reason for this is that outliers are events that show strong
341 effects, so replicates are not essential for detecting these types of events. If a putative disease-
342 causing event is detected, such as an aberrantly expressed gene, follow-up experiments
343 involving assays complementary to RNA-seq are necessary instead of replicates to establish
344 the functional link of the event to the disease[1,29]. In contrast, if an RNA-seq sample is suspected
345 to have a technical problem, a new library can be prepared, and the new data is substituted for
346 the former. Neither of these situations results in replicate samples. When replicates are
347 available, users can combine the *P*-values of replicate samples using Fisher's method for
348 combining *P*-values[30] by assuming independence of the read counts conditioned on the
349 expected means predicted by the autoencoder.

350 In general, the autoencoder controlling scheme and the read count modeling approach benefit
351 from additional sequencing data; the more data that can be combined, the better the estimation
352 of the typical patterns within a population will be. This holds true when the overall data is equally
353 distributed across population structures or sequencing protocols, because each sample is
354 assumed to be an independent representative of the whole population. This assumption was
355 partially violated in this study because RNA-seq data sets such as GTEx comprise >85%

356 Caucasian individuals[15]. Such overrepresentation of a given population in the data set is
357 disadvantageous to our method, and additional samples from underrepresented groups would
358 be especially beneficial. More testing is needed to assess whether our strategy for controlling
359 raw read counts can control for different data sources including data from multiple sequencing
360 platforms or control data sets. The ability to control for different protocols enables count data to
361 be combined from multiple sources. This would allow studies with a few samples to merge their
362 results with sources such as the publicly available GTEx data set[15]. Currently, the best practice
363 is to use the same cell handling and library preparation protocol which reduces the analyzable
364 data set and therefore limits the statistical power.

365 The initial aim of developing OUTRIDER was to create an expression outlier detection
366 framework for RNA-seq data in a diagnostic setting. Our primary focus was to identify aberrantly
367 expressed genes in RNA-seq samples from patients with rare diseases. OUTRIDER will be
368 useful for the identification of potentially disease-causing genes in patients for whom current
369 methods, such as WES and WGS, provide variants of unknown significance rather than a
370 limited set of candidate genes. However, our approach is not restricted to such data or
371 experiments. In principle, OUTRIDER could model any count data derived from next-generation
372 sequencing. Previous studies that established links between rare variants and aberrant gene
373 expression, such as the study of Li *et al.*[8], could benefit from OUTRIDER's refined outlier
374 detection approach. Our approach could also be applied to data such as DNA accessibility from
375 ATAC-seq reads. In this case, promotor regions or enhancers would be used as features
376 instead of gene bodies. Finally, the methodology of OUTRIDER could be adopted to data from
377 proteomics or splicing experiments.

Supplemental Data

The Supplemental Data accompanying this article includes five figures and a detailed derivation of the loss function's gradient.

Declaration of Interests

The authors declare no competing interests.

Acknowledgments

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 J.G. and V.A.Y.). A Fellowship through the Graduate School of Quantitative Biosciences Munich (QBM) supports V.A.Y. and Z.A. A.M was supported by a Fellowship through the Katholischer Akademischer Ausländer-Dienst (KAAD). 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 NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 06/12/17 with dbGaP accession number phs000424.v6.p1.

Author Contributions

J.G. conceived the project and overviewed the research with the help of Z.A, H.P., and V.A.Y. F.B., A.M., C.M. analyzed the data. F.B., C.M. and A.M. developed the software. D.M.B. and M.H contributed to the software development and early stage data analysis. J.G. and Z.A. devised the statistical analysis. F.B., V.A.Y., C.M., and J.G. made the figures. F.B., C.M., A.M., V.A.Y. and J.G. wrote the manuscript. All authors performed critical revision of the manuscript.

Web Resources

OUTRIDER: https://github.com/gagneurlab/OUTRIDER

GTEx: https://www.gtexportal.org/home

Supplementary data and additional results: 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. 1000 Genomes Project Consortium, 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., et al. (2015). A global

416 reference for human genetic variation. Nature *526*, 68–74.

417 4. Cummings, B.B., Marshall, J.L., Tukiainen, T., Lek, M., Foley, A.R., Bolduc, V., Waddell, L.B.,
418 Sarah, A., Grady, G.L.O., Estrella, E., et al. (2017). Improving genetic diagnosis in Mendelian
419 disease with transcriptome sequencing. Sci. Transl. Med. *9*, 1–25.

420 5. Kremer, L.S., Bader, D.M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., Haack, T.B., Graf,
421 E., Schwarzmayr, T., Terrile, C., et al. (2017). Genetic diagnosis of Mendelian disorders via
422 RNA sequencing. Nat. Commun. *8*, 15824.

423 6. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and
424 dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

425 7. Whitaker, L. (1914). On the Poisson law of small numbers. Biometrika *10*, 36–71.

426 8. Li, X., Kim, Y., Tsang, E.K., Davis, J.R., Damani, F.N., Chiang, C., Hess, G.T., Zappala, Z.,
427 Strober, B.J., Scott, A.J., et al. (2017). The impact of rare variation on gene expression across
428 tissues. Nature.

429 9. Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic
430 estimation of expression residuals (PEER) to obtain increased power and interpretability of gene
431 expression analyses. Nat. Protoc. *7*, 500–507.

432 10. Lecun, Y. (1987). PhD thesis: Modeles connexionnistes de l'apprentissage (connectionist
433 learning models) (Paris: Universite P. et M. Curie (Paris 6)).

434 11. Bourlard, H., and Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular
435 value decomposition. Biol. Cybern. *59*, 291–294.

436 12. Hinton, G.E., and Zemel, R.S. (1994). Autoencoders, Minimum Description Length and
437 Helmholtz Free Energy. In Advances in Neural Information Processing Systems 6, J.D. Cowan,

438 G. Tesauro, and Alspector J., eds. (Morgan-Kaufmann), pp. 3--10.

439 13. Way, G.P., and Greene, C.S. (2018). Extracting a biologically relevant latent space from
440 cancer transcriptomes with variational autoencoders. Pac. Symp. Biocomput. *23*, 80–91.

441 14. Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and
442 composing robust features with denoising autoencoders. Proc. 25th Int. Conf. Mach. Learn. -
443 ICML '08 1096–1103.

444 15. The GTEx Consortium, Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins,
445 H., Klemm, A., Flicek, P., Manolio, T., et al. (2015). The Genotype-Tissue Expression (GTEx)
446 pilot analysis: multitissue gene regulation in humans. Science (80-.). *348*, 648–660.

447 16. Casper, J., Zweig, A.S., Villarreal, C., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney,
448 B.J., Lee, C.M., Lee, B.T., Karolchik, D., et al. (2018). The UCSC Genome Browser database:
449 2018 update. Nucleic Acids Res. *46*, D762–D769.

450 17. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., and Kokocinski, F.
451 (2012). GENCODE: The Reference Human Genome Annotation for The ENCODE Project.
452 Genome Res *22*, 1760–1774.

453 18. Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data.
454 Genome Biol. *11*, R106.

455 19. McCarthy, D.J., Chen, Y., and Smyth, G.K. (2012). Differential expression analysis of
456 multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. *40*,
457 4288–4297.

458 20. Yoon, S., and Nam, D. (2017). Gene dispersion is the key determinant of the read count
459 bias in differential expression analysis of RNA-seq data. BMC Genomics *18*, 408.

460 21. Byrd, R.H., Lu, P., Nocedal, J., and Zhu, C. (1995). A Limited Memory Algorithm for Bound
461 Constrained Optimization. SIAM J. Sci. Comput. *16*, 1190–1208.

462 22. Broyden, C.G. (1970). The Convergence of a Class of Double-rank Minimization Algorithms
463 1. General Considerations. IMA J. Appl. Math. *6*, 76–90.

464 23. Shanno, D.F. (1970). Conditioning of quasi-Newton methods for function minimization.
465 Math. Comput. *24*, 647–647.

466 24. Fletcher, R. (1970). A new approach to variable metric algorithms. Comput. J. *13*, 317–322.

467 25. Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. Math.
468 Comput. *24*, 23–23.

469 26. Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple
470 testing under dependency. Ann. Stat. *29*, 1165–1188.

471 27. Zeng, Y., Wang, G., Yang, E., Ji, G., Brinkmeyer-Langford, C.L., and Cai, J.J. (2015).
472 Aberrant Gene Expression in Humans. PLoS Genet. *11*, 1–20.

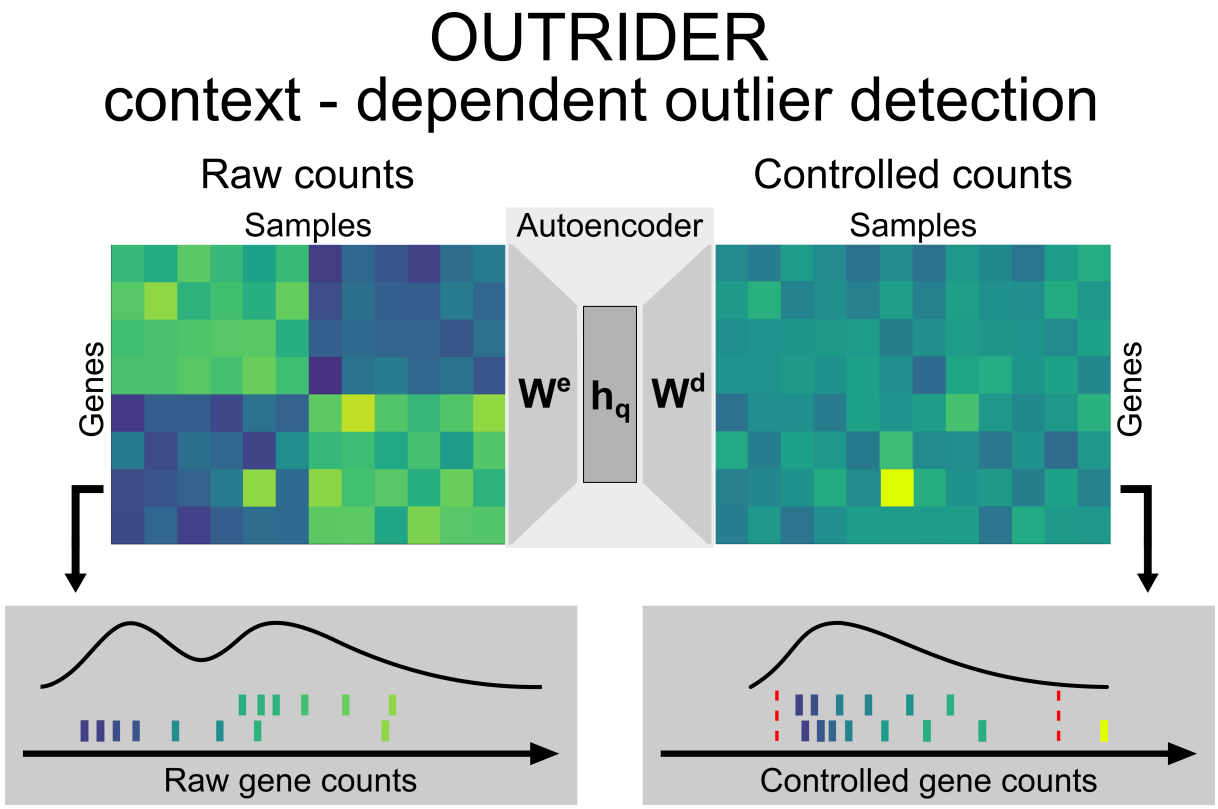
473 28. Guan, J., Yang, E., Yang, J., Zeng, Y., Ji, G., and Cai, J.J. (2016). Exploiting aberrant
474 mRNA expression in autism for gene discovery and diagnosis. Hum. Genet. *135*, 797–811.

475 29. MacArthur, D.G., Manolio, T.A., Dimmock, D.P., Rehm, H.L., Shendure, J., Abecasis, G.R.,
476 Adams, D.R., Altman, R.B., Antonarakis, S.E., Ashley, E.A., et al. (2014). Guidelines for
477 investigating causality of sequence variants in human disease. Nature *508*, 469–476.

478 30. Fisher, R.A. (1970). Statistical Methods for Research Workers, 14th edition (Oliver & Boyd).

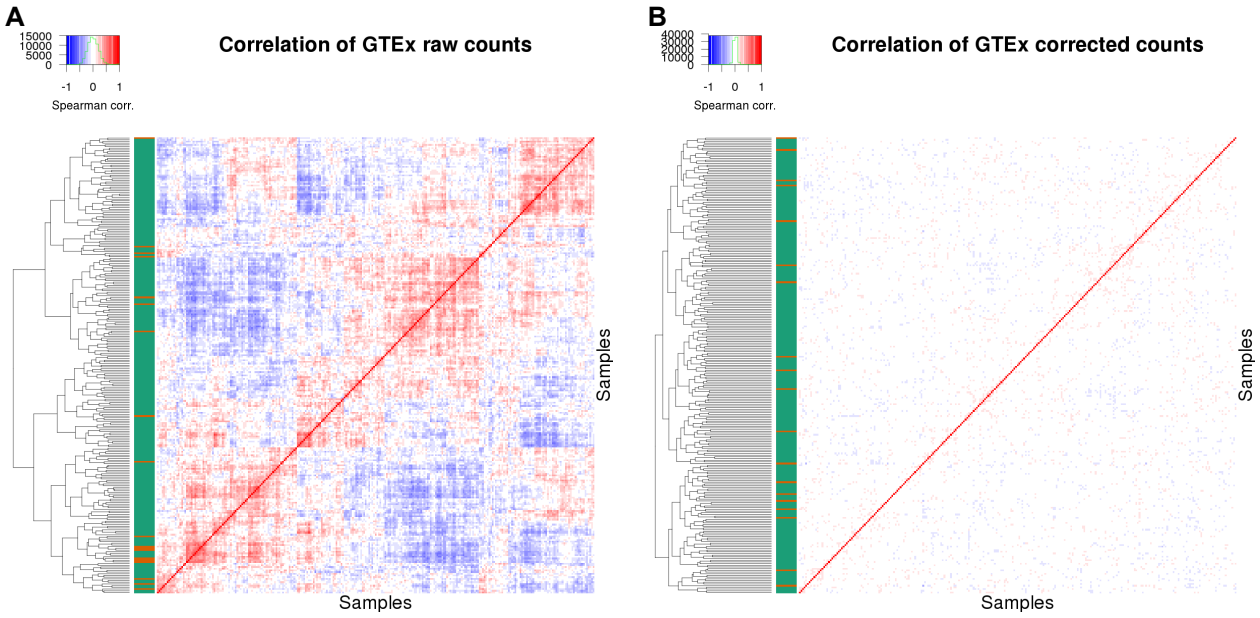
479

480 ## Figures



481

482 **Figure 1: Context-dependent outlier detection.** The algorithm identifies gene expression
483 outliers whose read counts are significantly aberrant given the co-variations typically observed
484 across genes in an RNA sequencing data set. This is illustrated by a read count (left panel, fifth
485 column, second row from the bottom) that is exceptionally high in the context of correlated
486 samples (left six samples) but not in absolute terms for this given gene. To capture commonly
487 seen biological and technical contexts, a denoising autoencoder models co-variations in an
488 unsupervised fashion and predicts read count expectations. By comparing the earlier mentioned
489 context dependent read count with these expectations it is revealed as exceptionally high (right
490 panel). The lower panels illustrate the distribution of read counts before and after applying the
491 correction for the relevant gene. The red dotted lines depict the significance cutoff.



492

493 **Figure 2: Controlling count data with autoencoders. (A)** Correlation matrix of row-centered
494 log-transformed read counts for the skin samples not exposed to the sun from the GTEx data
495 set (of 269 samples and 18,250 genes). Red indicates a positive correlation and blue a negative
496 correlation. The color code on the left indicates a normal sample as green and an outlier sample
497 as orange. The dendrogram represents the sample-wise hierarchical clustering. **(B)** Same as for
498 A, but with the autoencoder controlled read counts.

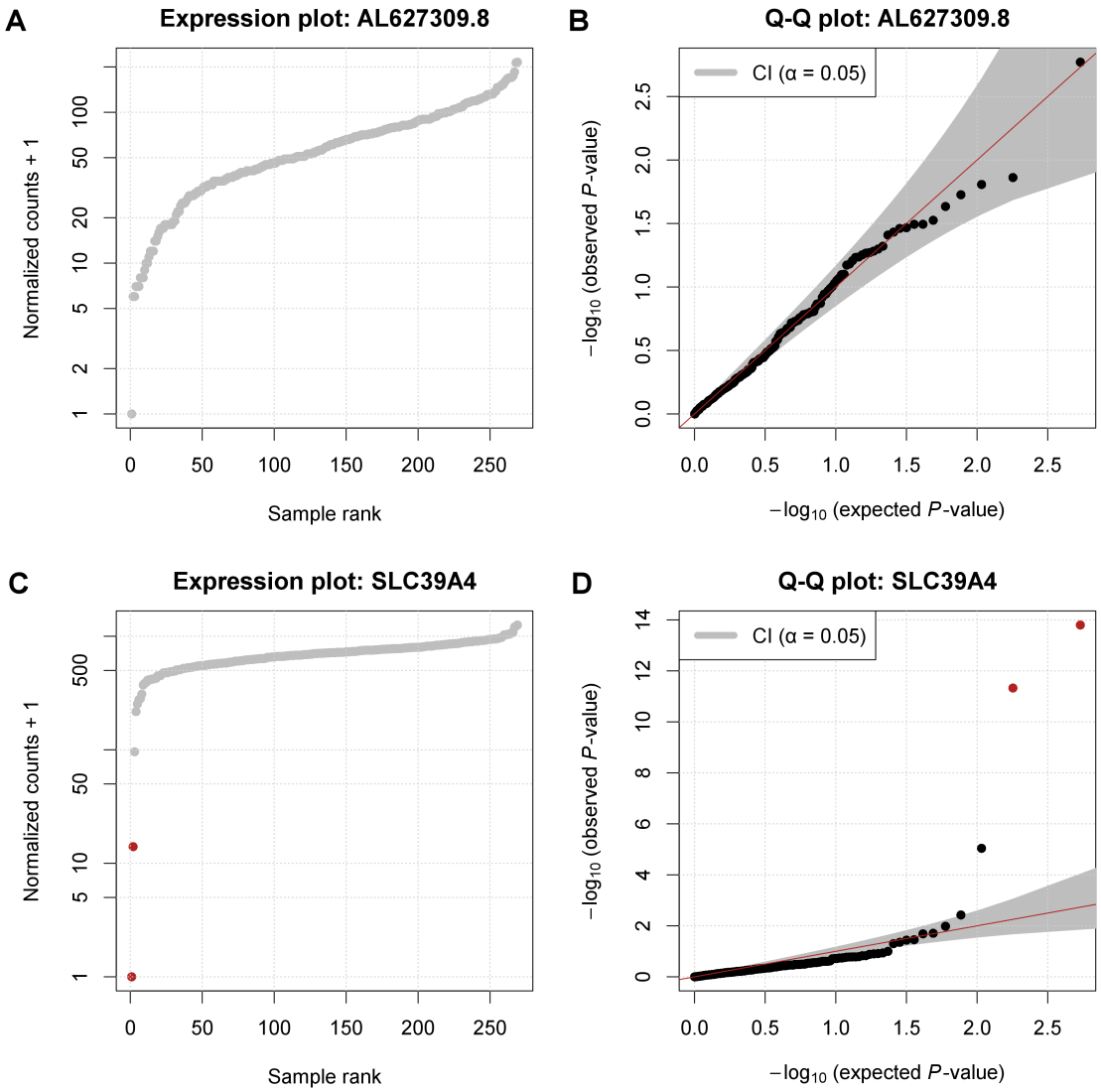
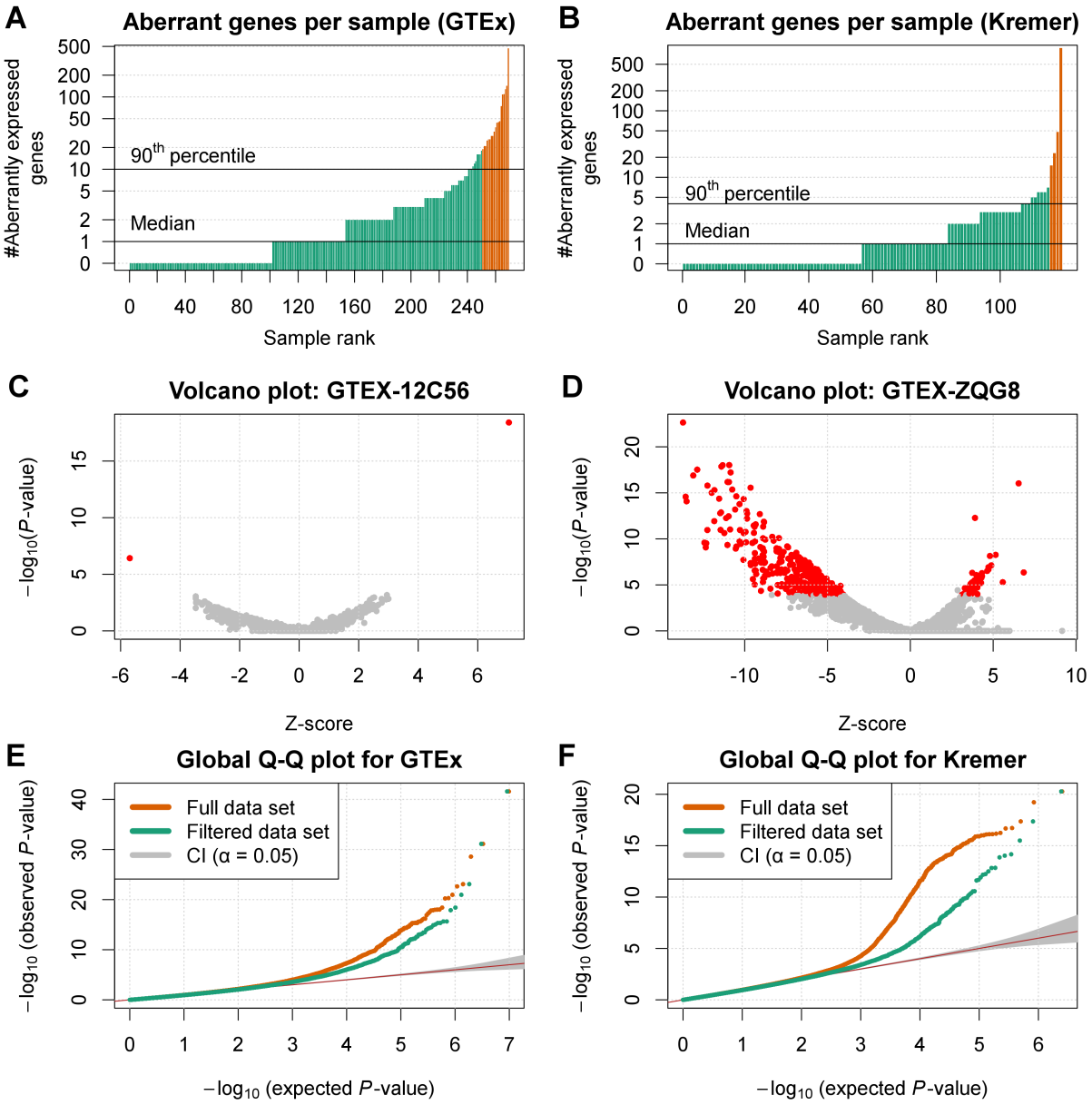
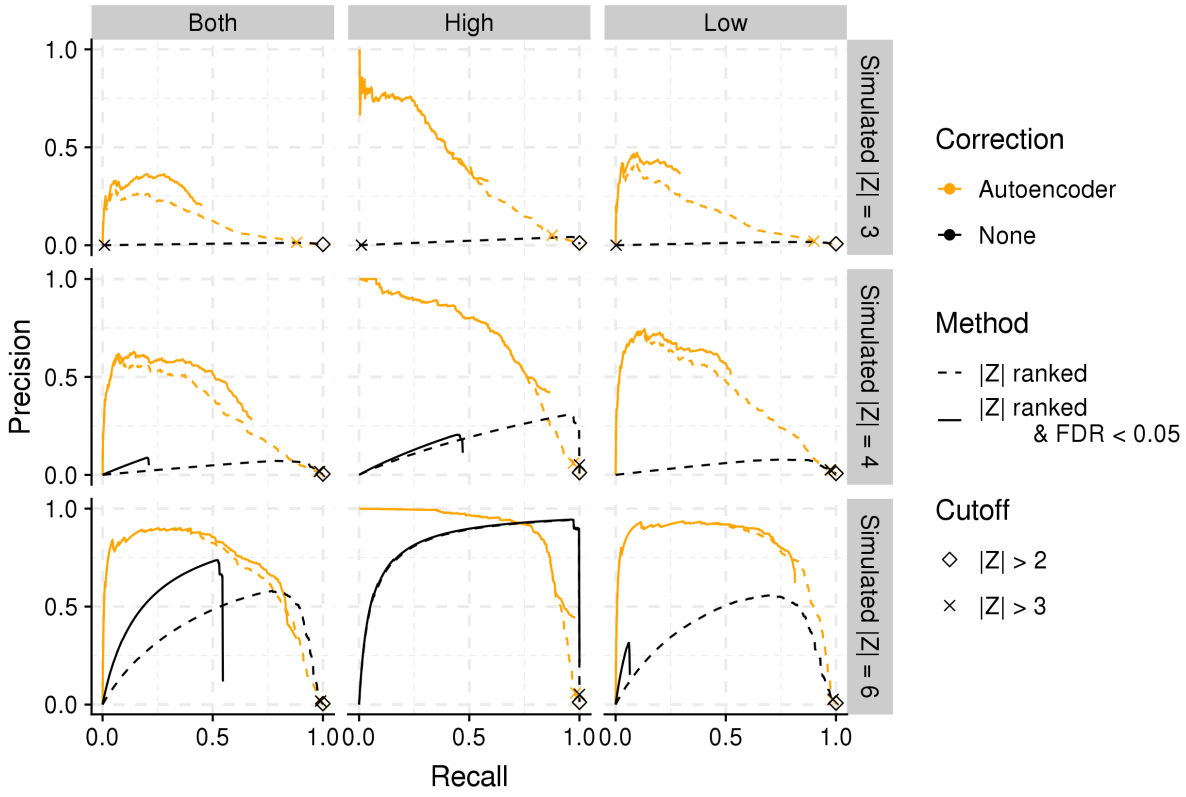


Figure 3: Using the negative binomial distribution for significance assessment. (A, C) Normalized RNA sequencing read counts plotted against their rank. **(B, D)** Quantile–quantile plots of observed *P*-values against expected *P*-values with confidence area. Outliers are shown in red (FDR < 0.05). Panels A and B show data for the gene *AL627309.8* with no detected expression outlier, and panels C and D show data for the gene *SLC39A4* with two expression outliers.



506

507 **Figure 4: RNA sequencing expression outlier detection. (A, B)** Number of aberrant genes
508 (FDR < 0.05) per sample for the GTEx (A) and Kremer data sets (B). The orange bars represent
509 abnormal samples (with >0.1% aberrantly expressed genes). **(C, D)** *P*-values versus Z-scores
510 for a representative normal sample (C) and a representative abnormal sample (D). Genes with
511 significantly aberrant read counts are marked in red. **(E, F)** Quantile–quantile plots for the data
512 shown in panel A (E), and in panel B (F). Observed *P*-values are plotted against the expected
513 *P*-values with confidence area.



514

515 **Figure 5: Outlier detection benchmark.** The proportion of simulated outliers among reported
516 outliers (precision) plotted against the proportion of reported simulated outliers among all
517 simulated outliers (recall) for decreasing Z-score and for different methods. Plots are provided
518 for three simulated amplitudes (by row, with simulated absolute Z-scores of 3, 4, and 6, top to
519 bottom, respectively) and for all combined results (left column), for the subset of aberrantly high
520 counts (middle column), and for the subset of aberrantly low counts (right column). The read
521 counts were either controlled for sequencing depth only (black) or were also controlled for gene
522 co-variation with the autoencoder (orange). The solid lines show the results restricted to
523 significant counts only (FDR < 0.05) and the dashed lines show the results for all counts. Two
524 commonly used Z-score cutoffs for outlier detection[4,8] are marked with a diamond (|Z| > 2) and a
525 cross (|Z| > 3).

Supplemental Data

Supplemental Figures

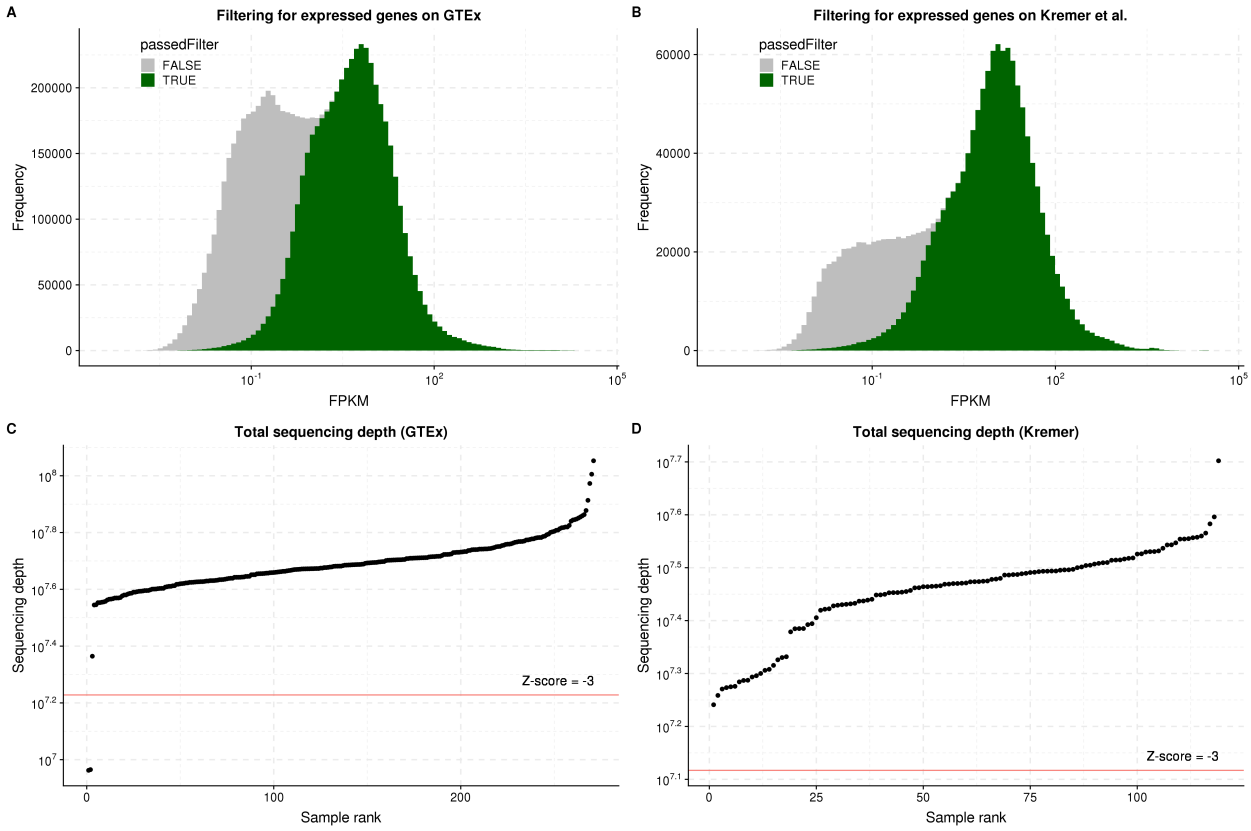
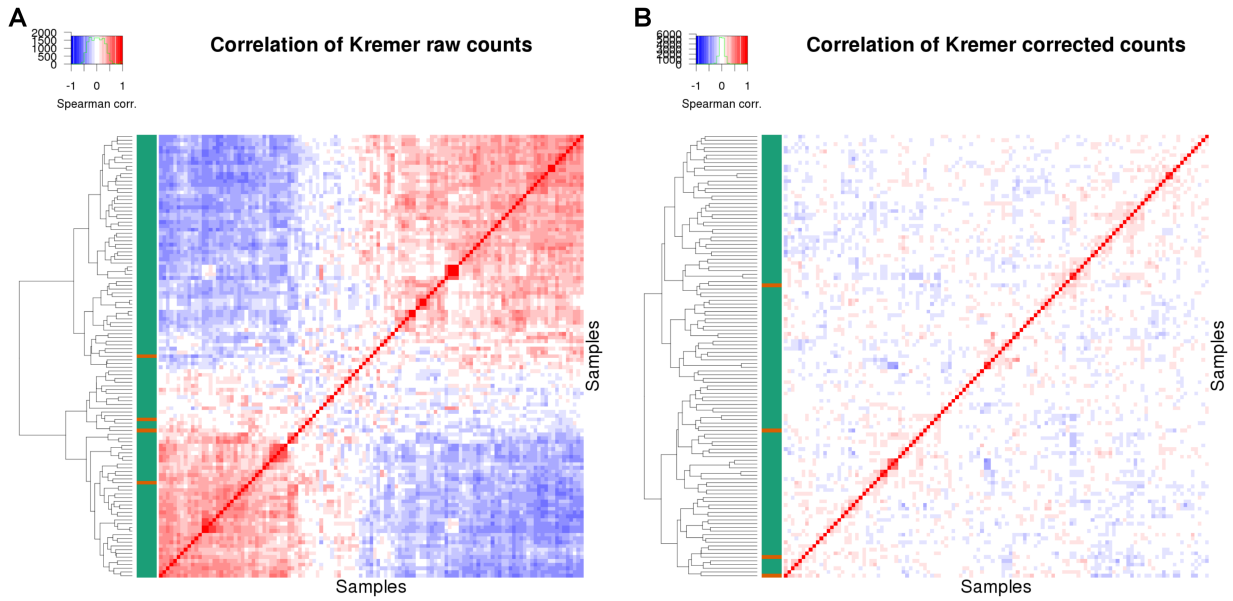
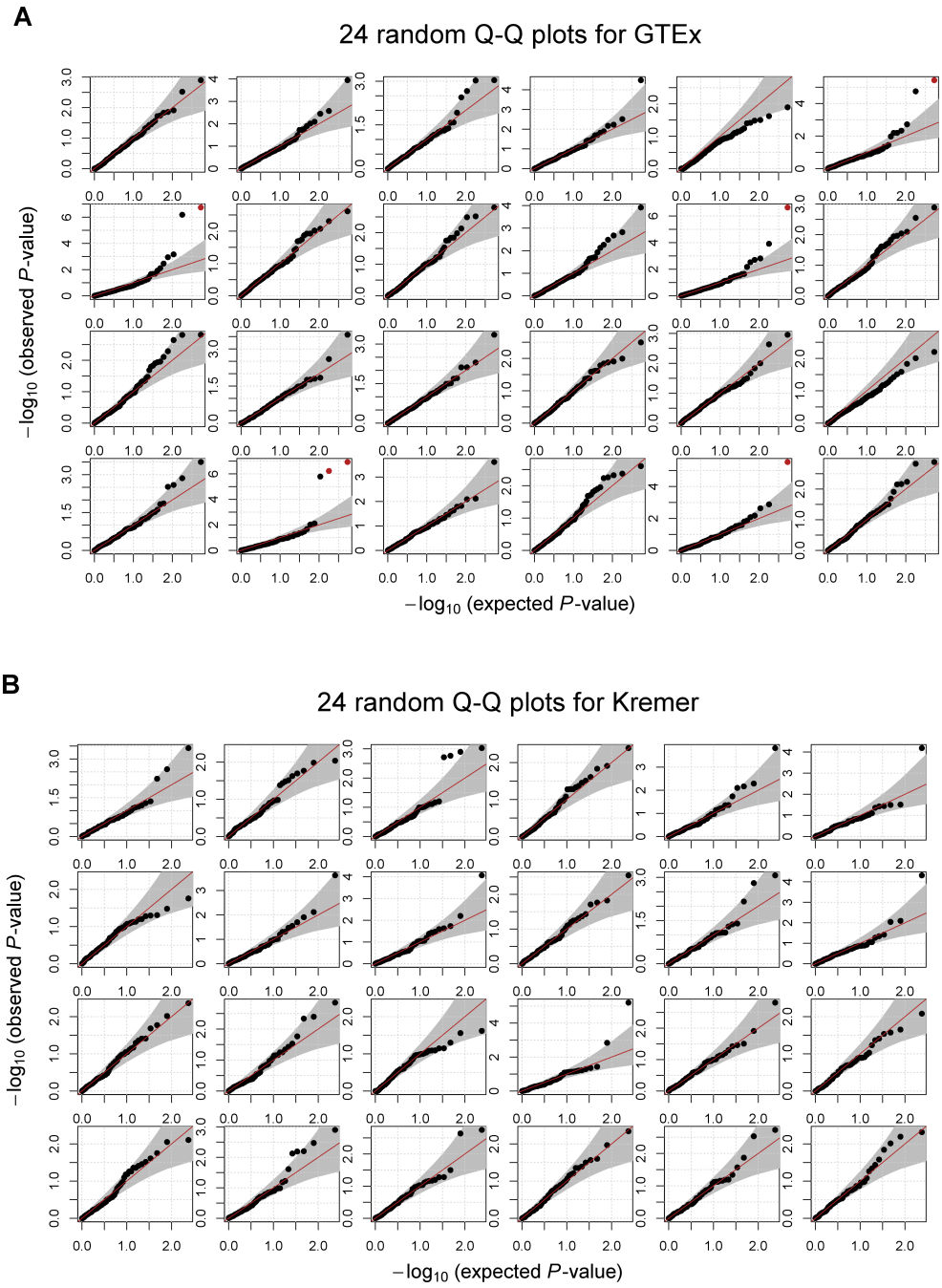


Figure S1: Filtering of genes and samples. (A) Histogram of the RPKM values for the GTEx data set grouped according to the filter status. Green indicates the genes that passed the filter and gray those that were filtered out. **(B)** The same as A, but for the Kremer data set. **(C)** The total sequencing depth sorted from low to high for all the GTEx samples. The red horizontal line represents the Z-score cutoff of -3. **(D)** The same as C, but for the Kremer data set.



534

535 **Figure S2: Controlling count data with autoencoders. (A)** Correlation matrix of row-centered
536 log-transformed read counts for the Kremer data set (with 119 samples and 10559 genes). Red
537 indicates a positive correlation and blue a negative correlation. The color code on the left
538 indicates a normal sample as green and an outlier sample as orange. The dendrogram
539 represents the sample-wise hierarchical clustering. **(B)** Same as in A, but with the autoencoder
540 controlled read counts.



541

542 **Figure S3: Using the negative binomial distribution for read counts. (A)** Quantile–

543 quantile plots for 24 randomly selected genes from the GTEx data set. **(B)** Same as in A, but

544 for the Kremer data set.

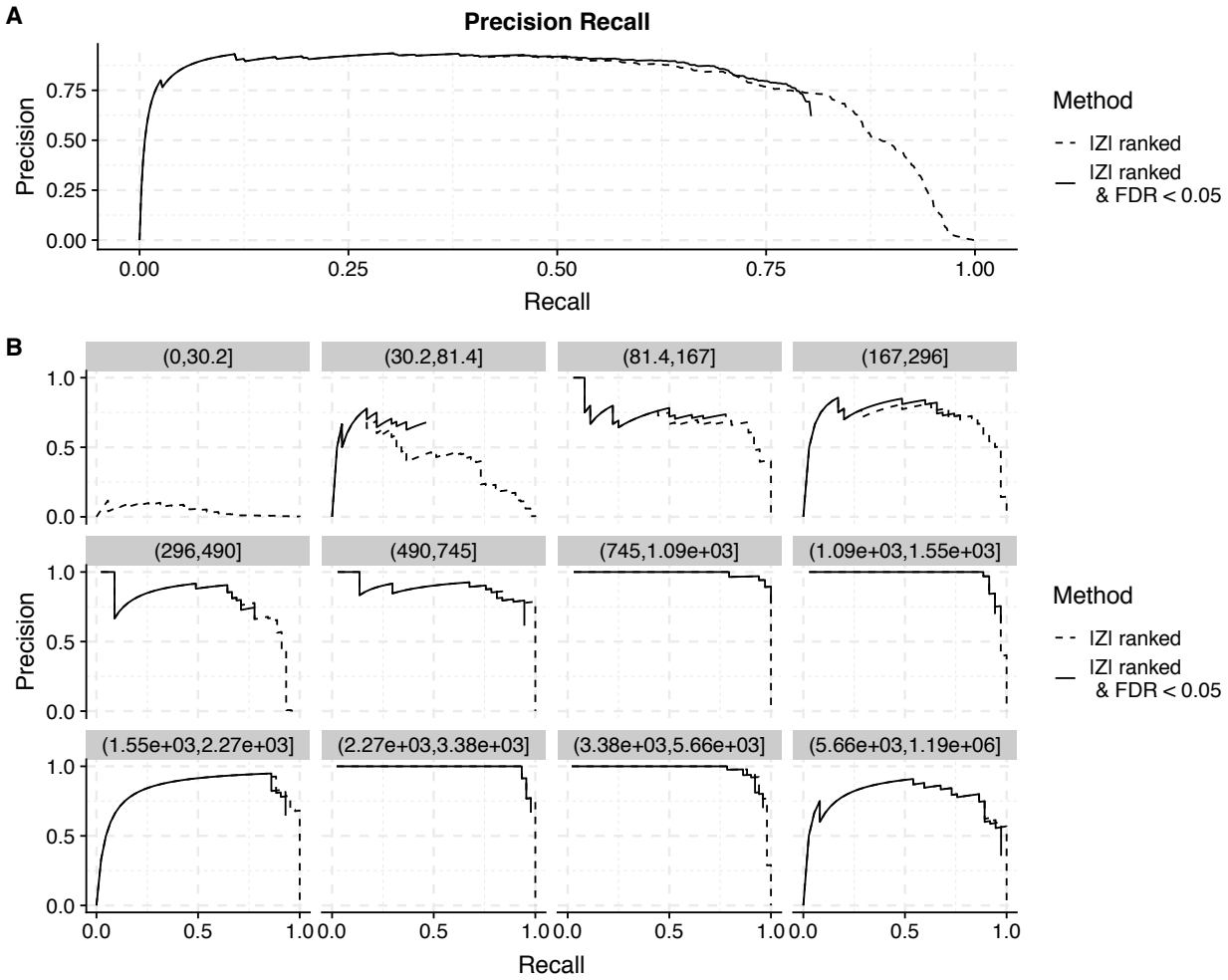
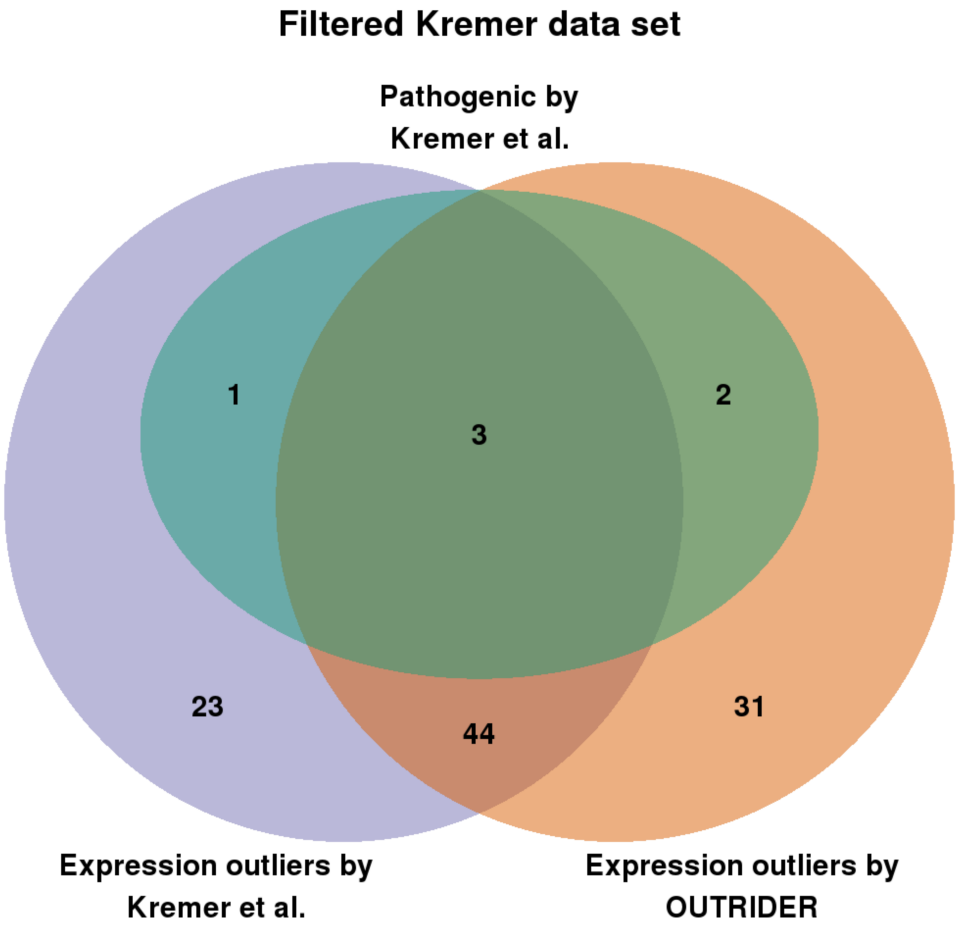


Figure S4: Expression level-dependent recall. Precision versus recall for artificially injected low expression outliers with a Z-score of 6. **(A)** For all the injected outliers. **(B)** Split into 12 equal bins according to the mean expression level of the genes. Only a small fraction of the injected outliers was significant for the two lowest bins, with a mean expression level smaller than 81 counts.



551

552 **Figure S5: Benchmark of OUTRIDER using the Kremer data set.** Venn diagram of the
553 expression outliers detected by OUTRIDER (orange), expression outliers detected by Kremer *et*
554 *al*. (violet), and pathogenic events validated by Kremer *et al*. (green) within 48 previous to
555 Kremer *et al*. undiagnosed samples[5].

556

Supplemental Methods

We use L-BFGS to fit the autoencoder model as described in the Methods. To speed up the fitting we implemented the gradient as derived below.

The expectations μ are modeled by:

$$\mu_{ij} = s_i e^{y_{ij}+\bar{x}_j}$$
$$\mathbf{Y} = \mathbf{XWW}^T + \mathbf{b}$$

where the matrix $\mathbf{X}$ is given by the matrix: $\log \frac{k_{ij}+1}{s_i} - \bar{x}_j$.

The negative binomial log likelihood is given by:

$$ll = \sum_{ij} k_{ij} \log(\mu_{ij}) + \sum_{ij} \theta \log(\theta) - \sum_{ij} (k_{ij} + \theta) \log(\mu_{ij} + \theta)$$
$$+ \sum_{ij} \log(\Gamma(\theta + k_{ij})) - \sum_{ij} \log(\Gamma(\theta) k_{ij}!)$$

For the derivation of the gradient only the first and third term need to be considered, as all other terms are independent of μ.

Computing the derivative of the first term with respect to the matrix $\mathbf{W}$ by subsituting the autoencoder model for μ. Here the operations $\log[a]$ and $\exp[a]$ are understood to be element-wise.

$$\begin{aligned}
&\frac{d}{dw_{ab}} \sum_{ij} k_{ij} \log(\mu_{ij}) \\
&= \frac{d}{dw_{ab}} \sum_{ij} k_{ij} \log[\exp[\mathbf{XWW}^T + \mathbf{b}]] \\
&= \frac{d}{dw_{ab}} \sum_{ij} k_{ij} (\mathbf{XWW}^T + \mathbf{b}) \\
&= \frac{d}{dw_{ab}} \sum_{ij} k_{ij} (\sum_{lm} x_{il} w_{lm} w_{jm} + b_j) \\
&= \sum_{ij} k_{ij} (x_{ia} w_{jb} + \delta_{aj} \sum_{l} x_{il} w_{lb}) \\
&= \sum_{ij} x_{ia} k_{ij} w_{jb} + \sum_{il} k_{ia} x_{il} w_{lb}
\end{aligned}$$

Which can be written as:

$$\mathbf{K}^T\mathbf{X}\mathbf{W} - \mathbf{X}^T\mathbf{K}\mathbf{W}$$

Equivalently the derivative of the third term is:

$$-\mathbf{L}^T\mathbf{X}\mathbf{W} - \mathbf{X}^T\mathbf{L}\mathbf{W}$$

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

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

The combined result is then:

$$\frac{dll}{d\mathbf{W}} = \mathbf{K}^T\mathbf{X}\mathbf{W} + \mathbf{X}^T\mathbf{K}\mathbf{W} - \mathbf{L}^T\mathbf{X}\mathbf{W} - \mathbf{X}^T\mathbf{L}\mathbf{W}$$

The derivative of the first term with respect to the bias $\mathbf{b}$ is computed as:

$$\begin{aligned}
&\frac{d}{db_a}\sum_{ij} k_{ij}\log(\mu_{ij}) \\
&= \frac{d}{db_a}\sum_{ij} k_{ij}(\sum_{lm} x_{il}w_{lm}w_{jm} + b_j) \\
&= \sum_i k_{ia}
\end{aligned}$$

Equivalently for the third therm the derivative is $-\sum_i l_{ia}$ and so the derivative of the loglikelihood with respect to the bias is:

$$\frac{dll}{db_a} = \sum_i k_{ia} - l_{ia}$$