1 Supplemental Figures

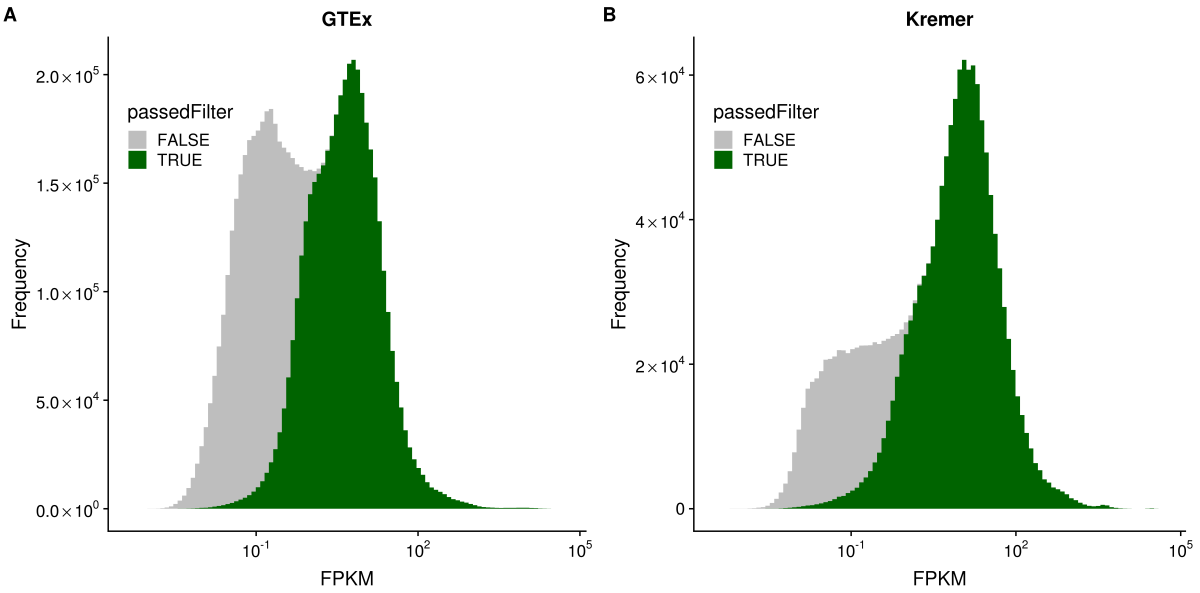

Figure S1: **Filtering of genes.** **(A)** Histogram of the FPKM values for the GTEx data set grouped according to the filter status. Green indicates the genes that passed the filter and gray those that were filtered out. **(B)** Same as A, but for the Kremer data set.

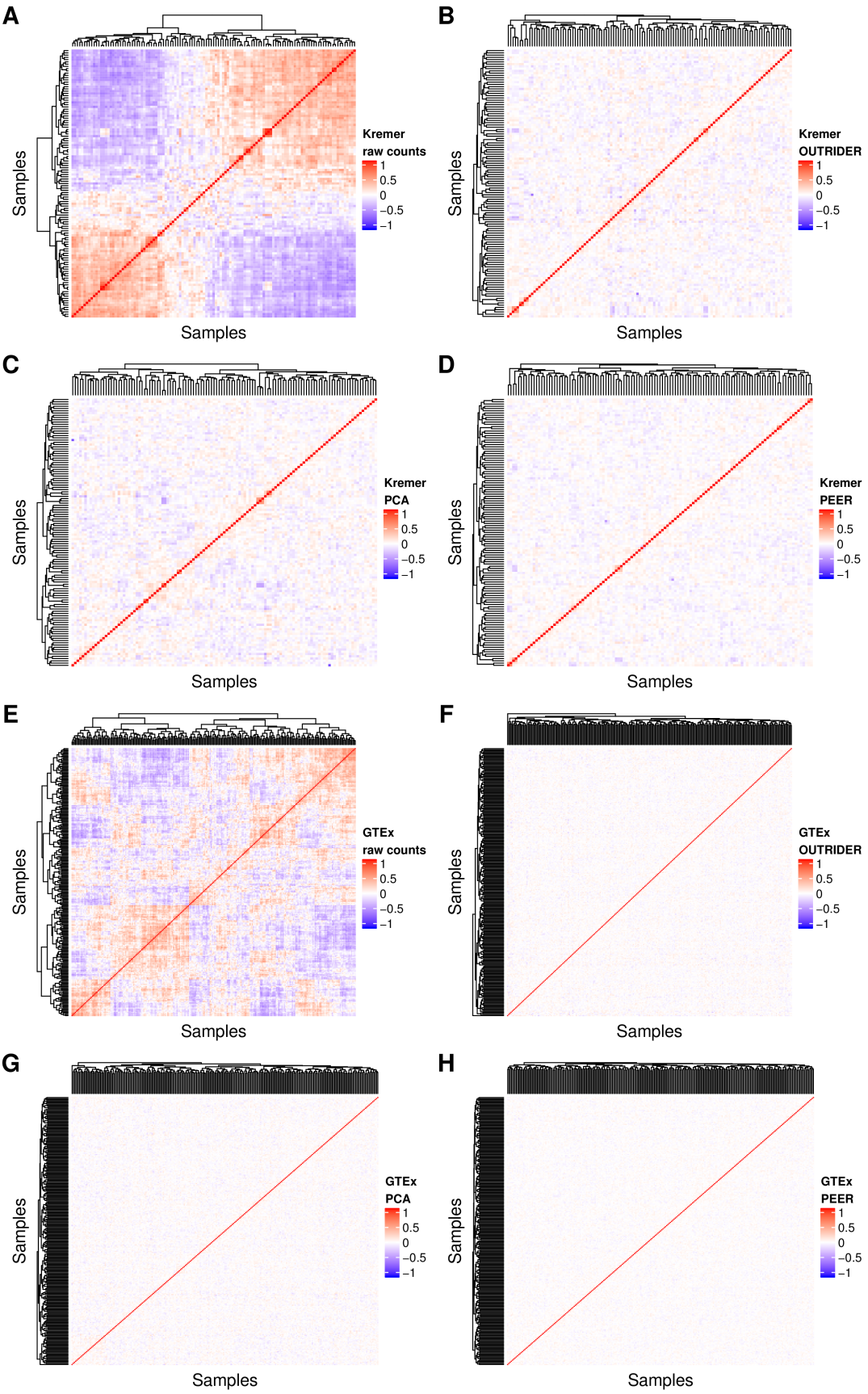

Figure S2: **Controlling count data for covariation.** **(A)** Correlation matrix of row-centered log-transformed read counts for the Kremer data set (119 samples and 10,556 genes). Red indicates a positive correlation and blue a negative correlation. The dendrogram represents the sample-wise hierarchical 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).

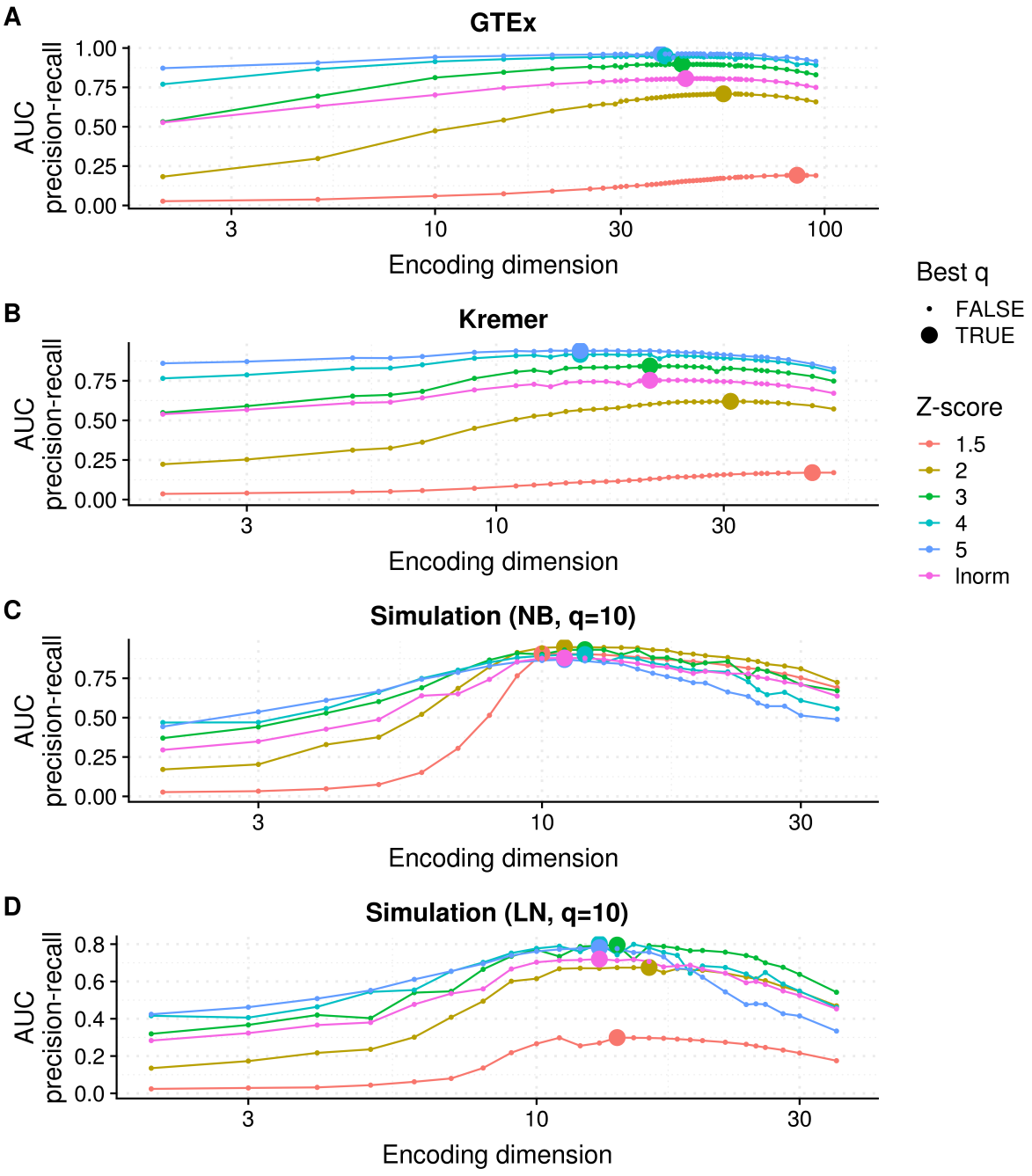

Figure S3: **Fitting the encoding dimension. (A, B, C, D)** Area under the precision–recall curve plotted against the autoencoder encoding dimension q for different Z-score amplitudes of corrupted read counts (colors). The pink curve corresponds to Z-score amplitudes sampled from a log-normal distribution (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).

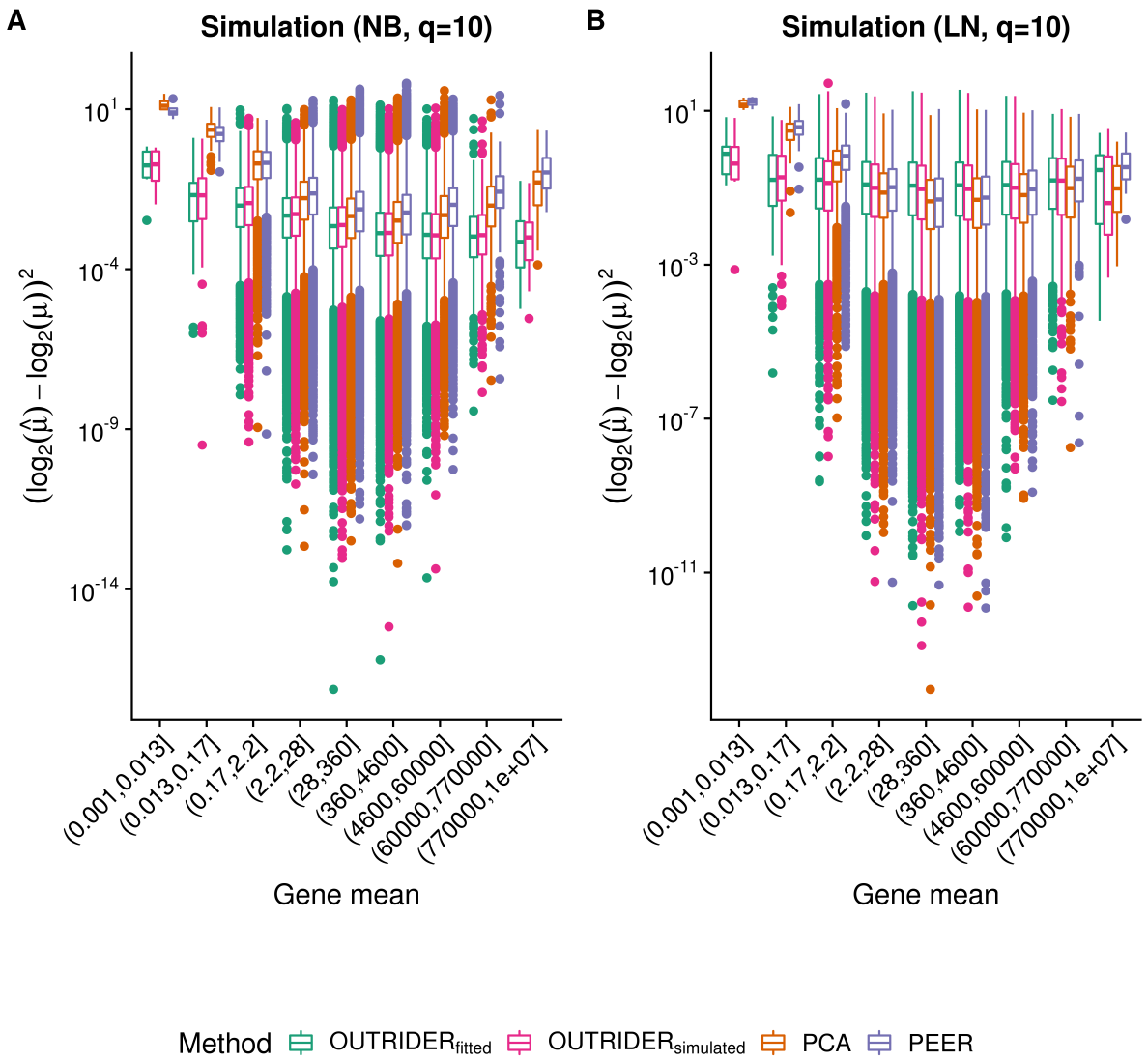

Figure S4: **OUTRIDER recovers expected counts on simulated data.** Boxplots of squared differences between $\log_2$ of fitted means and $\log_2$ of simulated means binned into 9 logarithmically spaced mean gene expression bins for OUTRIDER, PCA and PEER on simulated data. **(A)** Corresponds to the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

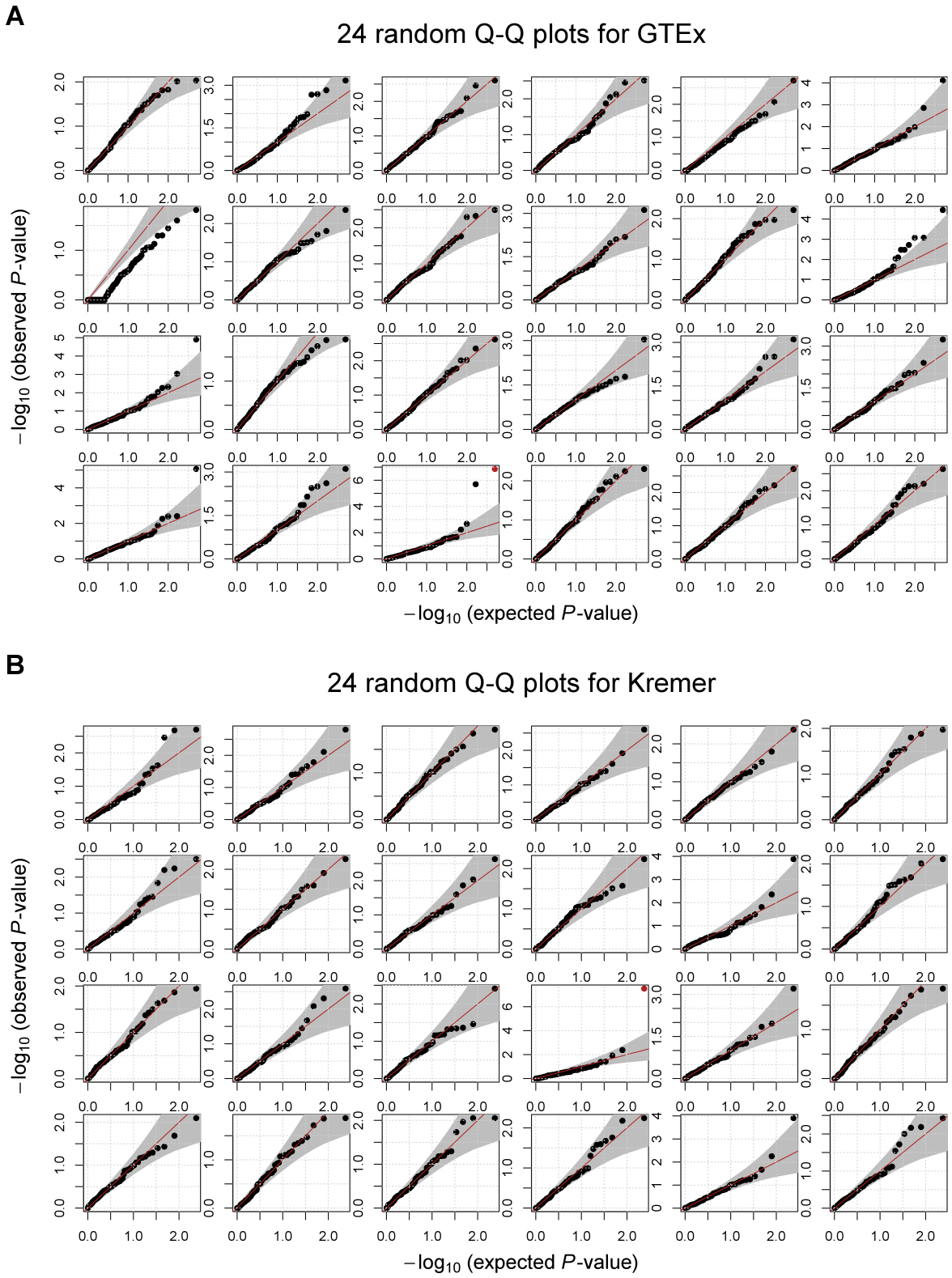

Figure S5: **Negative binomial distribution fits for individual genes.**
(A) Quantile–quantile plots for 24 randomly selected genes from the GTEx data set.
(B) Same as in A but for the Kremer data set.

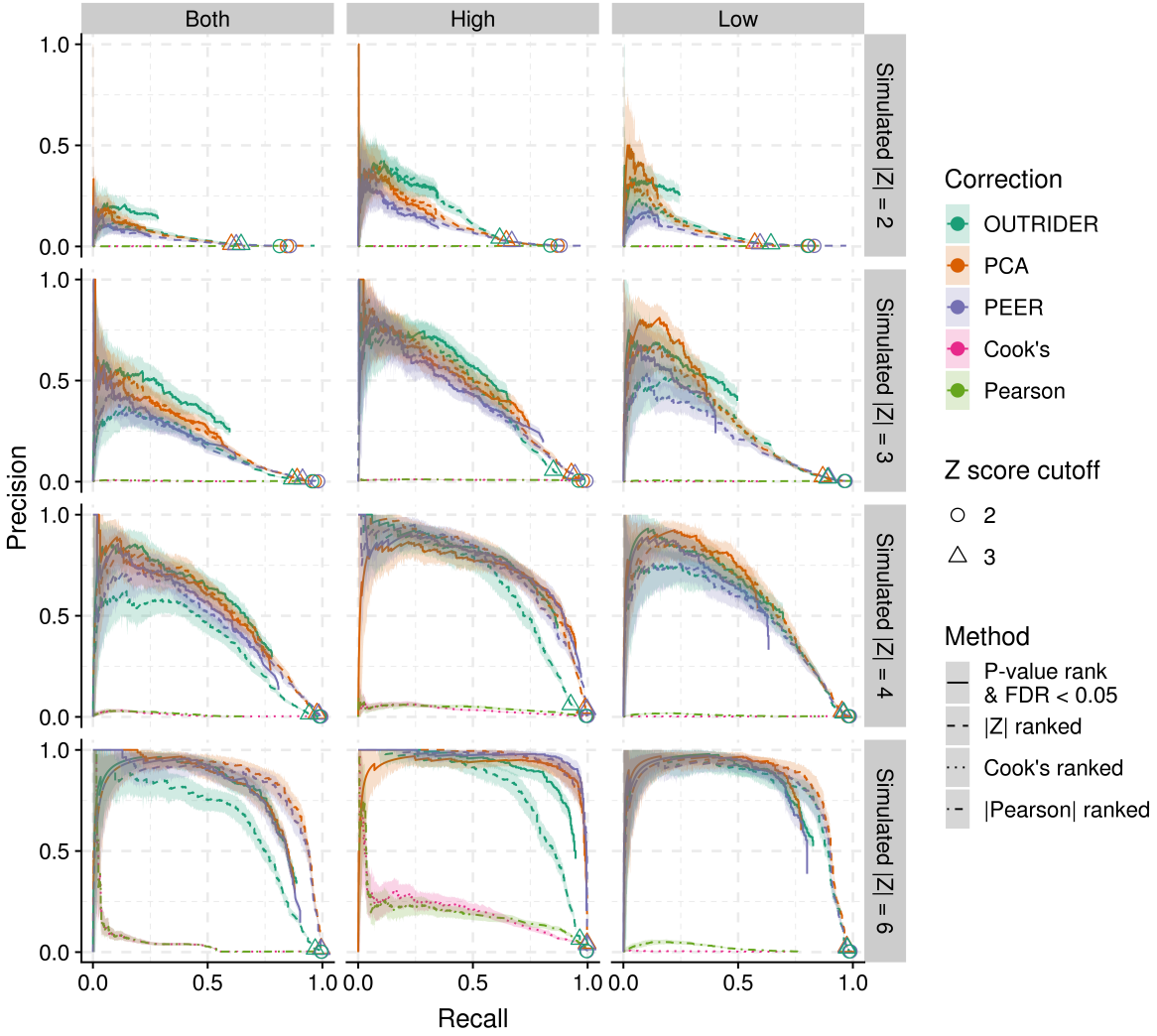

Figure S6: **Outlier detection benchmark in GTEx.** The proportion of simulated outliers among reported outliers (precision) plotted against the proportion of reported simulated outliers among all simulated outliers (recall) for 8 different ranking methods. The 8 ranking methods are OUTRIDER (green solid), 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.

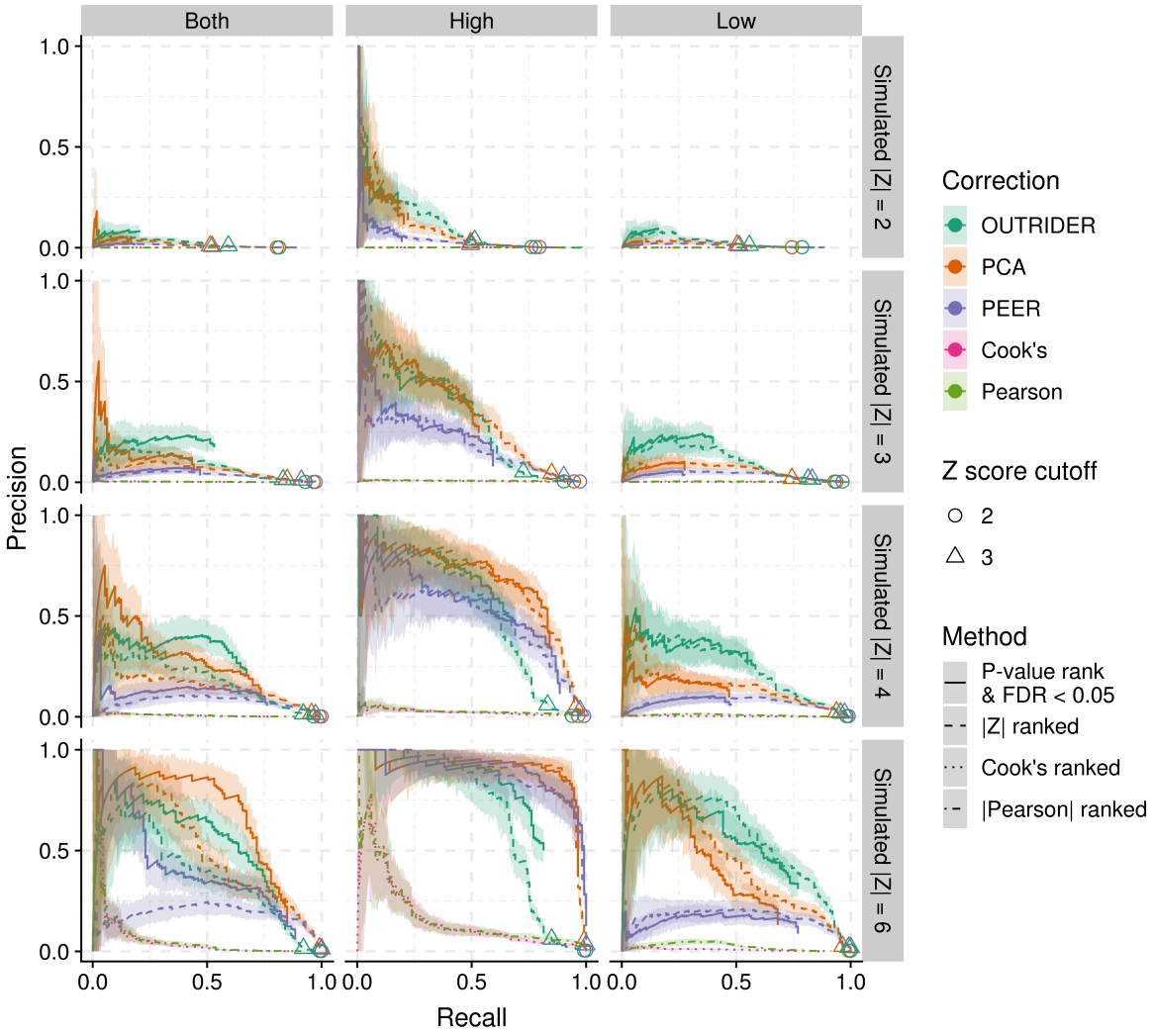

Figure S7: **Outlier detection benchmark in Kremer.** Same as S6 but for the Kremer data set.

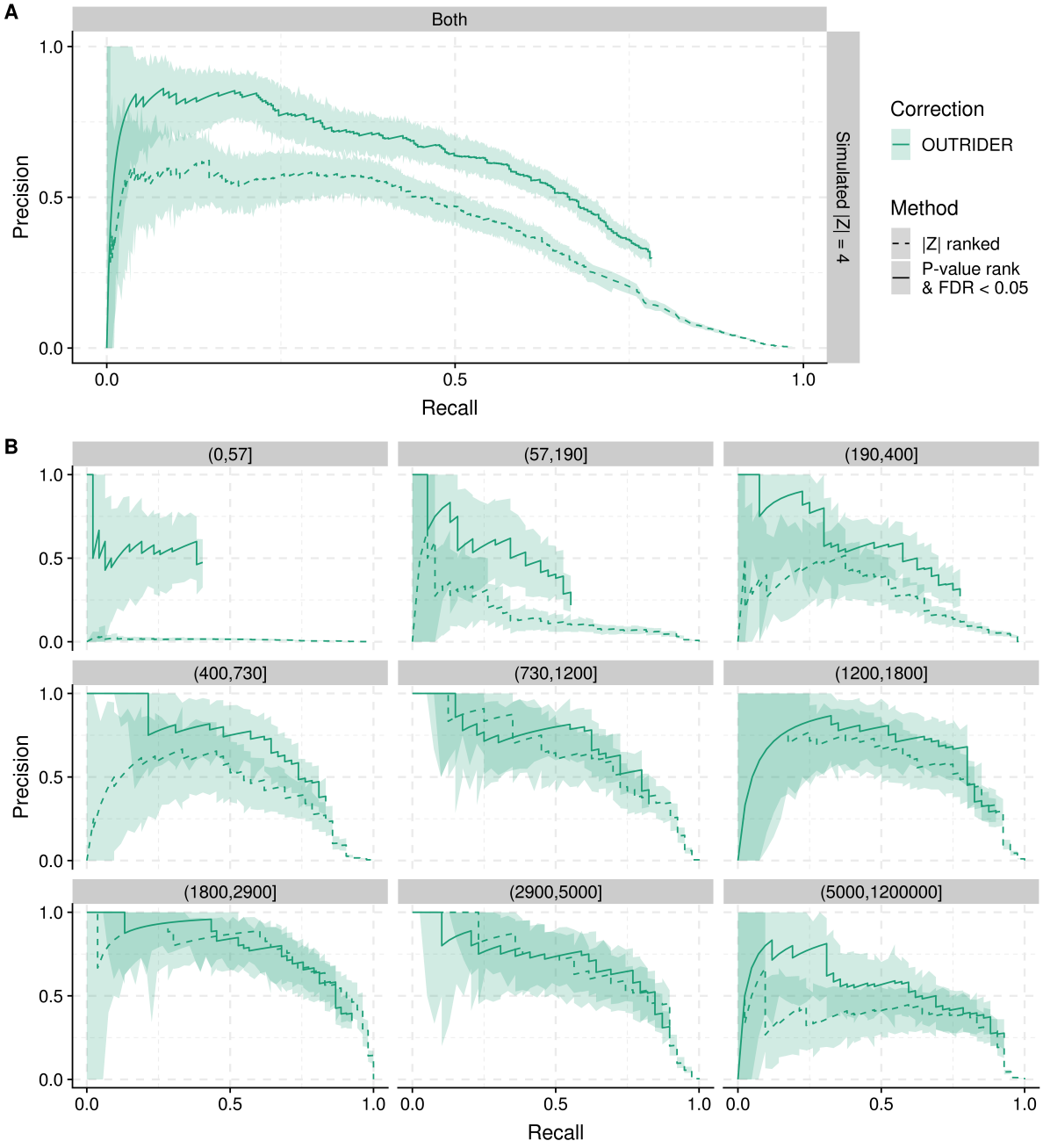

Figure S8: **Expression level dependent recall.** Precision versus recall for artificially injected high and low expression outliers with a Z-score of 4 for OUTRIDER ranked by *P*-values with FDR < 0.05 (solid) and ranked by Z-score (dashed). **(A)** For all the injected outliers. **(B)** Splitted into 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.

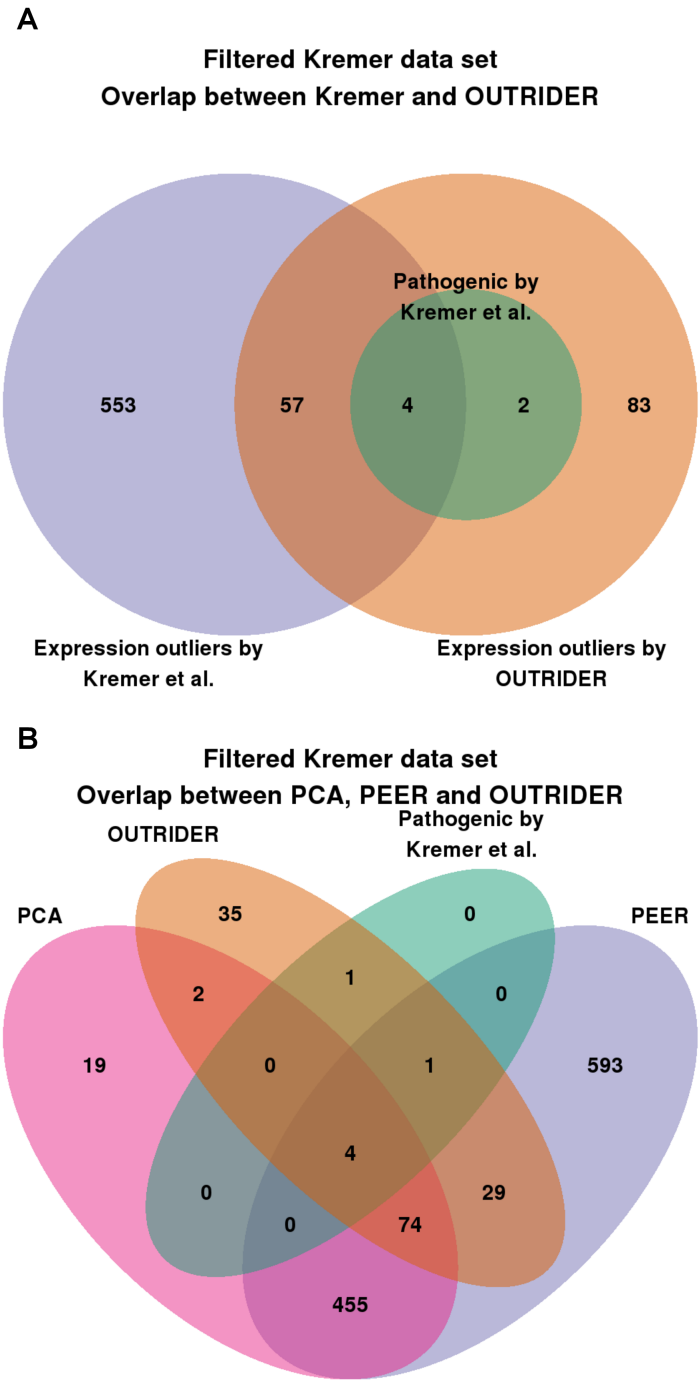

Figure S9: **Benchmark of OUTRIDER using validated genes.** **(A)** Venn diagram of the expression outliers detected by OUTRIDER (orange), expression outliers detected by Kremer et al. (violet), and pathogenic outliers validated by Kremer et al. (green) within the 48 samples of individuals undiagnosed 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.

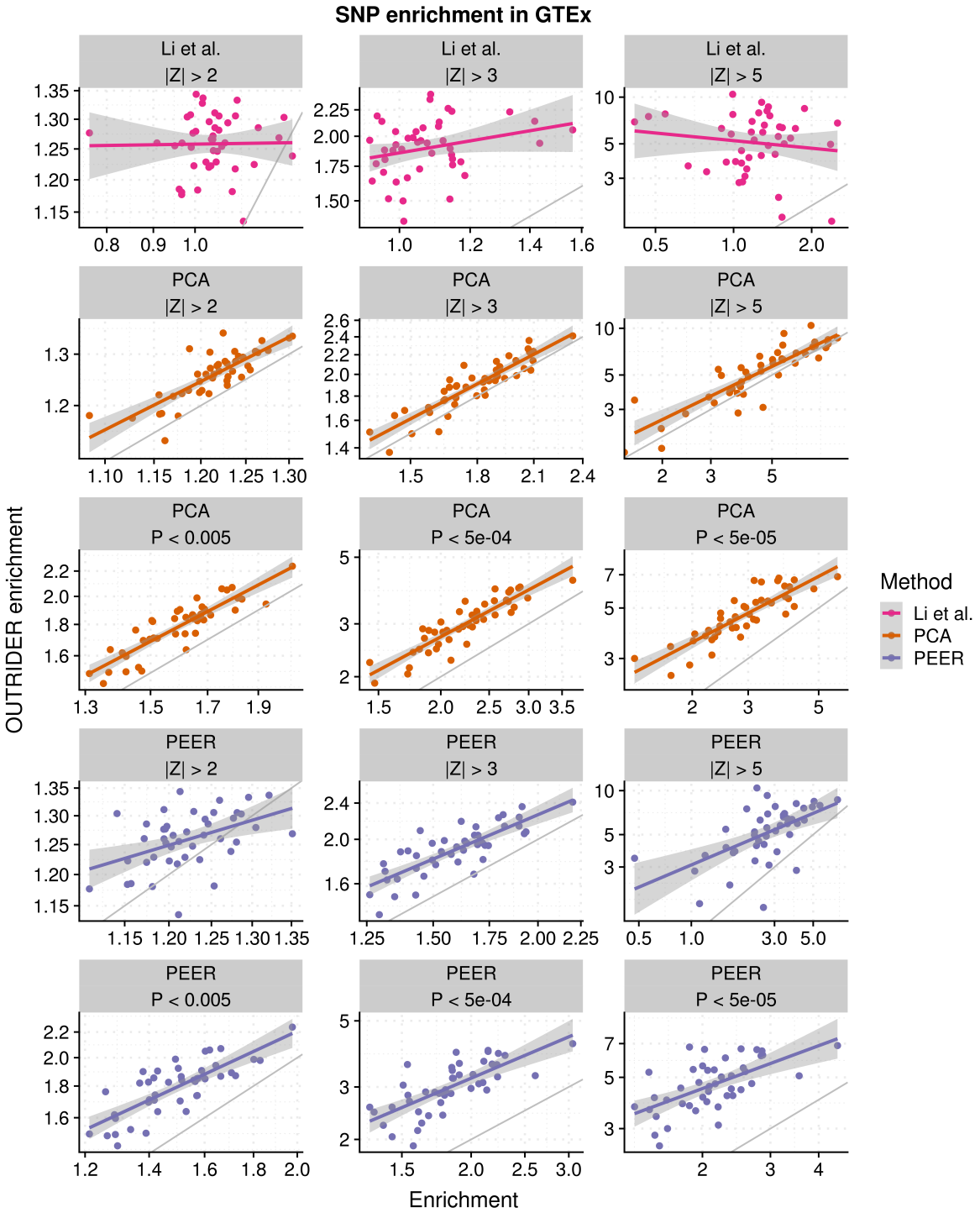

Figure S10: **OUTRIDER SNP enrichment.** Enrichment of rare (MAF < 0.05) moderate and high impact variants (according to VEP) computed on genes found to be aberrantly expressed using OUTRIDER plotted against enrichments computed on genes found to be aberrantly expressed using Z-scores published by Li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

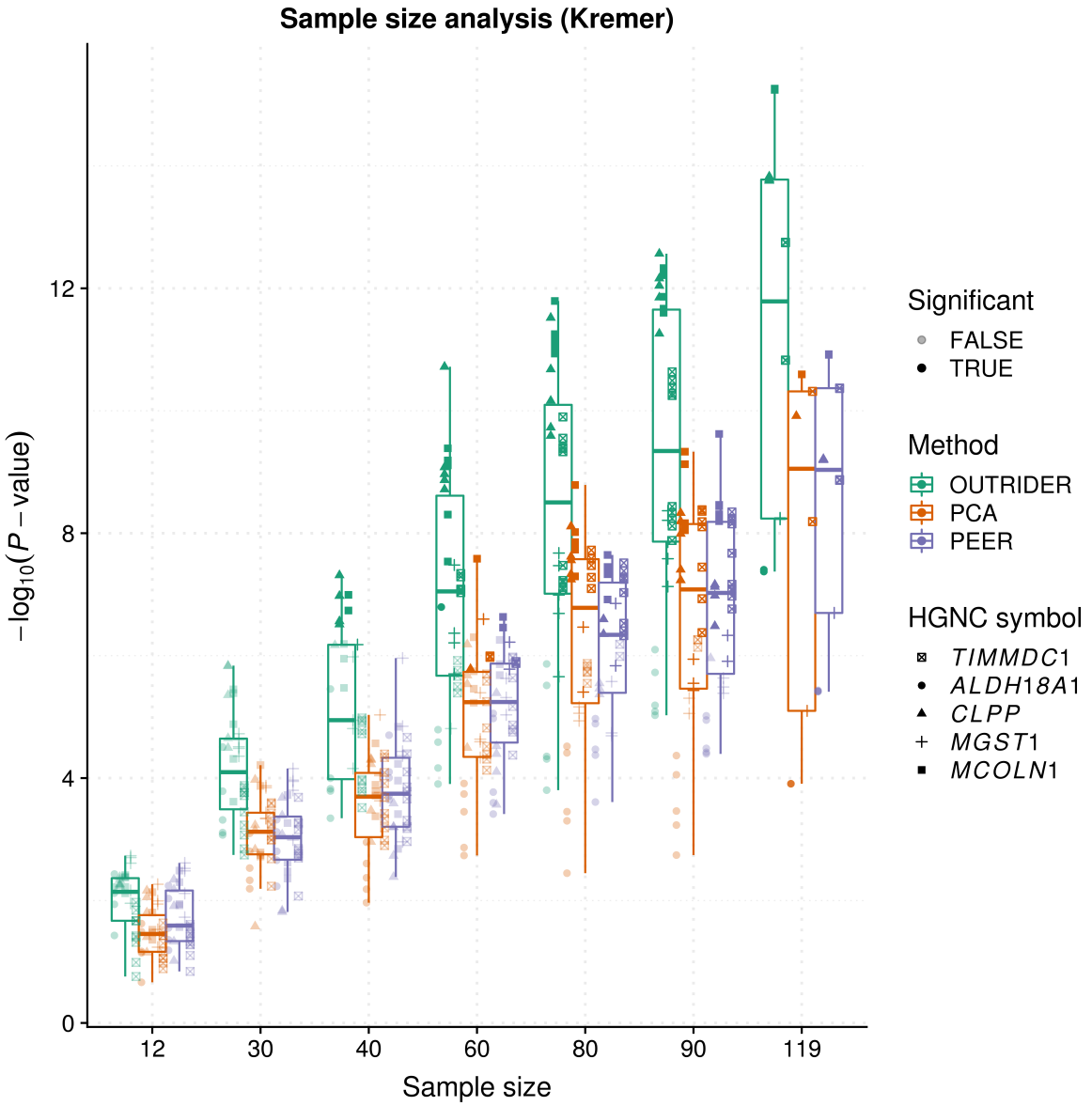

Figure S11: **Sample size analysis.** Negative $\log_{10}$ P-values are plotted against the number of samples in the data set, for 6 pathogenic genes (validated in Kremer et al. [1]). For each data set size, five random sets of samples containing the samples with the known outliers were drawn. Genes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