Supplementary Data

Detection of aberrant events in RNA sequencing data

Vicente A. Yépez[1,2,*], Christian Mertes[1,*], Michaela F. Müller[1], Daniela S. Andrade[1], Leonhard Wachutka[1], Laure Frésard[3], Mirjana Gusic[4,5], Ines Scheller[1], Patricia F. Goldberg[1], Holger Prokisch[4,5], Julien Gagneur[1,2,*]

[1] Department of Informatics, Technical University of Munich, Munich, Germany
[2] Quantitative Biosciences Munich, Department of Biochemistry, Ludwig-Maximilians Universität, Munich, Germany
[3] Department of Pathology, School of Medicine, Stanford University, Stanford, CA, USA
[4] Institute of Human Genetics, Helmholtz Zentrum München, Neuherberg, Germany
[5] Institute of Human Genetics, Technical University of Munich, Munich, Germany
[*] Corresponding authors

Supplementary Methods

Aberrant expression

To generate the gene counts, DROP uses the `summarizeOverlaps` function from the GenomicAlignments package[1]. The parameters mode, singleEnd, ignore.strand, and inter.feature are all specified by the user in the sample annotation (refer to main text or HT-Seq[2] documentation). Afterwards, genes in which less than 5% of the samples have a FPKM > fpkmCutoff (defined in the config file) are filtered out. Finally, the OUTRIDER fit is run for each analysis group and the results tables are created.

Aberrant splicing

The whole module is implemented as described in FRASER[3]. Briefly, split reads are counted using the `summarizeJunctions` function from the GenomicAlignments package[1], and non-split reads overlapping splice sites are counted using the `featureCounts` function from the Rsubread package[4]. Then, they are converted into the intron-centric metrics percent-spliced-in and splicing efficiency. The percent-spliced-in (ψ) index is computed as the ratio between reads mapping to the given intron and all split-reads sharing the same donor or acceptor site, respectively:

$$\psi_5(D,A) = \frac{n(D,A)}{\sum_{A'} n(D,A')} \text{ and } \psi_3(D,A) = \frac{n(D,A)}{\sum_{D'} n(D',A)},$$

where $n(D,A)$ denotes the number of split reads mapping to the intron spanning from donor *D* to acceptor *A*. To detect partial or full intron retention, the splicing efficiency (θ) metric is used. It is defined as the ratio of all split-reads and the full read coverage at a given splice site:

$$\theta_5(D) = \frac{\sum_{A'} n(D,A')}{\sum_{A'} n(D,A') + n(D)} \text{ and } \theta_3(A) = \frac{\sum_{D'} n(D',A)}{\sum_{D'} n(D',A) + n(A)},$$

where $n(D)$ denotes the number of reads spanning the exon–intron boundary at the donor splice site *D* and $n(A)$ is the number of reads spanning the exon–intron boundary at the acceptor site *A*.

Afterwards, introns with less than 20 reads in all samples and introns for which the total number of reads at the donor and acceptor splice site is zero in more than 95% of the samples are filtered out. Lastly, the FRASER fit is run and the results are extracted for each junction and aggregated by gene.

Mono-allelic expression

First, DROP subsets the VCF(s) files to obtain only SNVs using the `view` command from bcftools[5]. Then, the allelic counting is performed using the `ASEReadCounter`[6] function from GATK[7]. The negative binomial test is applied on the reads using the DESeq2 package[8] fixing the dispersion parameter to 0.05 as done in Kremer et al.[9]. Finally, allele frequencies from gnomAD[10] are added using the R packages MafDb.gnomAD.r2.1.hs37d5 and MafDb.gnomAD.r2.1.GRCh38.

VCF-BAM matching

In order to match the variants from DNA and RNA, first a set of variants that are not in linkage disequilibrium (as correlated variants can bias the results[11]) is needed. In order to obtain that set, we pooled all of the variants from the samples in the test dataset and consider only the ones in autosomal chromosomes that are not in linkage disequilibrium using the function `snpgdsLDpruning` from the R/Bioconductor package SNPRelate[12]. Applying a linkage disequilibrium threshold of 0.2, we obtained a set of *P* = 26,402 variants and their genomic positions.

For each of the *N* VCF files, we check for variants at those positions, thus generating a vector x_i =[0/0, 0/1, 1/1, …,] of size *P*, where 0/0 represents no variant, 0/1 heterozygous, 1/1 homozygous variant, and *i=1,...,N* is a counter for the VCF files. Then, we compute the allelic counts at those *P* positions using all *M* BAM files. We test if they are mono-allelically expressed and obtain a vector y_j =[NA, 0/1, 1/1, 0/0, …,] of size *P*, where 0/0 means a ratio of the alternative allele (ratioALT) < 0.2, 1/1 that ratioALT > 0.8, 0/1 that 0.2 ≤ ratioALT ≤ 0.8, and NA that the position was not expressed or had less than 10 reads, and *j=1,...,M* is a counter for the BAM files. Finally, we count the number of elements that are the same for each combination of

vectors x_i , y_j and divide it by the length of y_j after removing missing values, thus generating an $N \times M$ matrix.

Supplementary References

1. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
2. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
3. Mertes, C. *et al. Detection of aberrant splicing events in RNA-Seq data with FRASER*. https://tinyurl.com/FRASER-paper (2019).
4. Liao, Y., Smyth, G. K. & Shi, W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.* **47**, e47 (2019).
5. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
6. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* **16**, 195 (2015).
7. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
8. Anders, S. & Huber, W. Differential expression analysis for sequence count data. 12 (2010).
9. Kremer, L. S. *et al.* Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat. Commun.* **8**, 15824 (2017).
10. Karczewski, K. J. *et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes*.

http://biorxiv.org/lookup/doi/10.1101/531210 (2019) doi:10.1101/531210.

11. Slatkin, M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9**, 477–485 (2008).
12. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).

Supplementary Figures

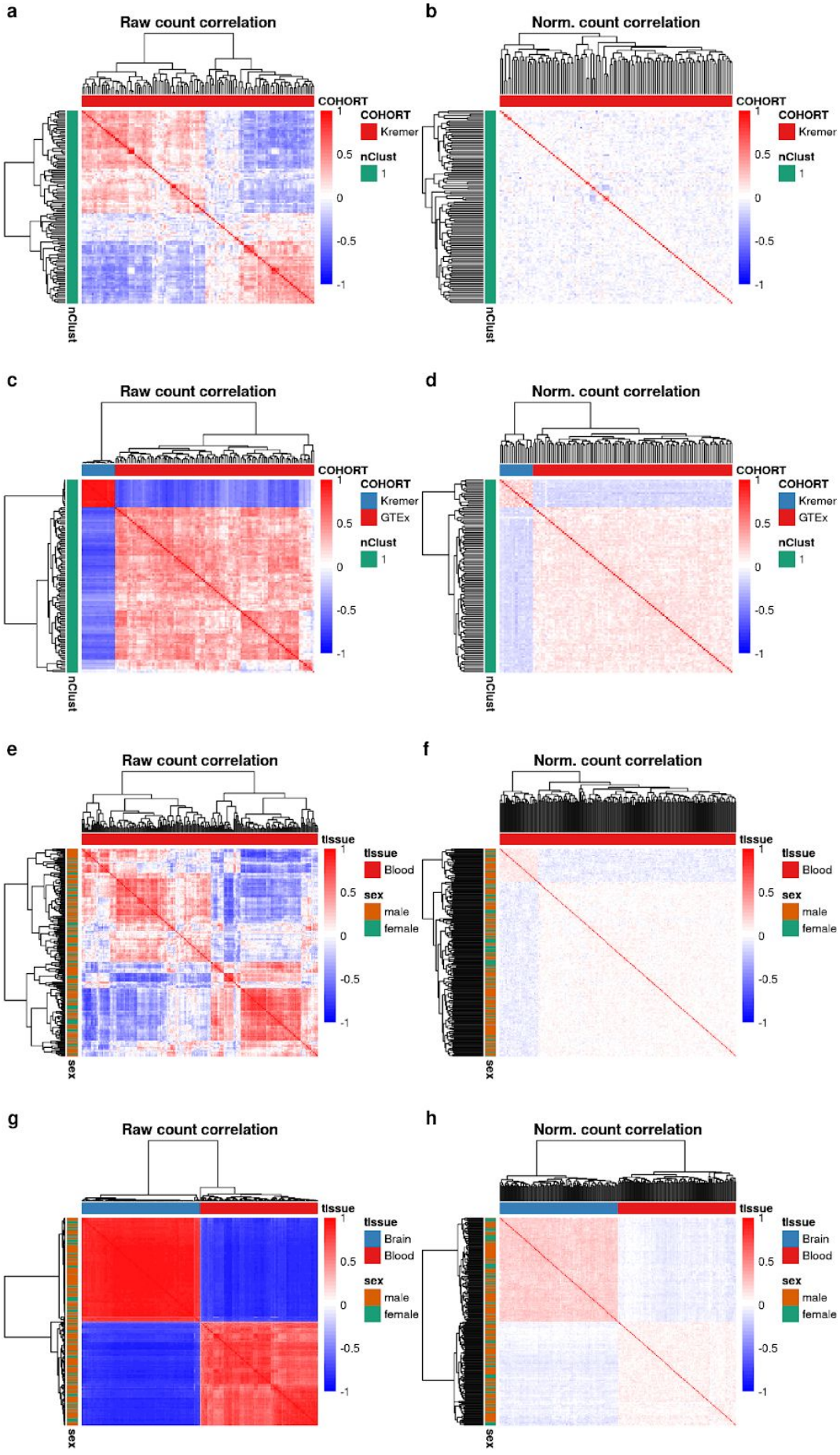
a
Raw count correlation
COHORT
COHORT
Kremer
nClust
1
1
0.5
0
-0.5
-1
nClust
b
Norm. count correlation
COHORT
COHORT
Kremer
nClust
1
1
0.5
0
-0.5
-1
nClust
c
Raw count correlation
COHORT
COHORT
Kremer
GTEx
nClust
1
1
0.5
0
-0.5
-1
nClust
d
Norm. count correlation
COHORT
COHORT
Kremer
GTEx
nClust
1
1
0.5
0
-0.5
-1
nClust
e
Raw count correlation
tissue
tissue
Blood
sex
male
female
1
0.5
0
-0.5
-1
sex
f
Norm. count correlation
tissue
tissue
Blood
sex
male
female
1
0.5
0
-0.5
-1
sex
g
Raw count correlation
tissue
tissue
Brain
Blood
sex
male
female
1
0.5
0
-0.5
-1
sex
h
Norm. count correlation
tissue
tissue
Brain
Blood
sex
male
female
1
0.5
0
-0.5
-1
sex

Fig S1 | Raw and OUTRIDER-normalized count correlation heatmaps of different datasets. **a,** Heatmap of the correlation of row-centered log-transformed read counts between samples before. The dataset consists of 119 fibroblast samples from Kremer. **b,** same as a) after autoencoder correction. **c, d,** Same as a) and b) but for a simulated heterogeneous dataset consisting of 17 samples from Kremer and 102 samples from GTEx skin not sun exposed **e, f,** Same as a) and b) but for a dataset consisting of 200 blood samples from GTEx. **g, h,** Same as c) and d) but for a dataset consisting of 100 blood and 100 brain (cerebellum) samples from GTEx.

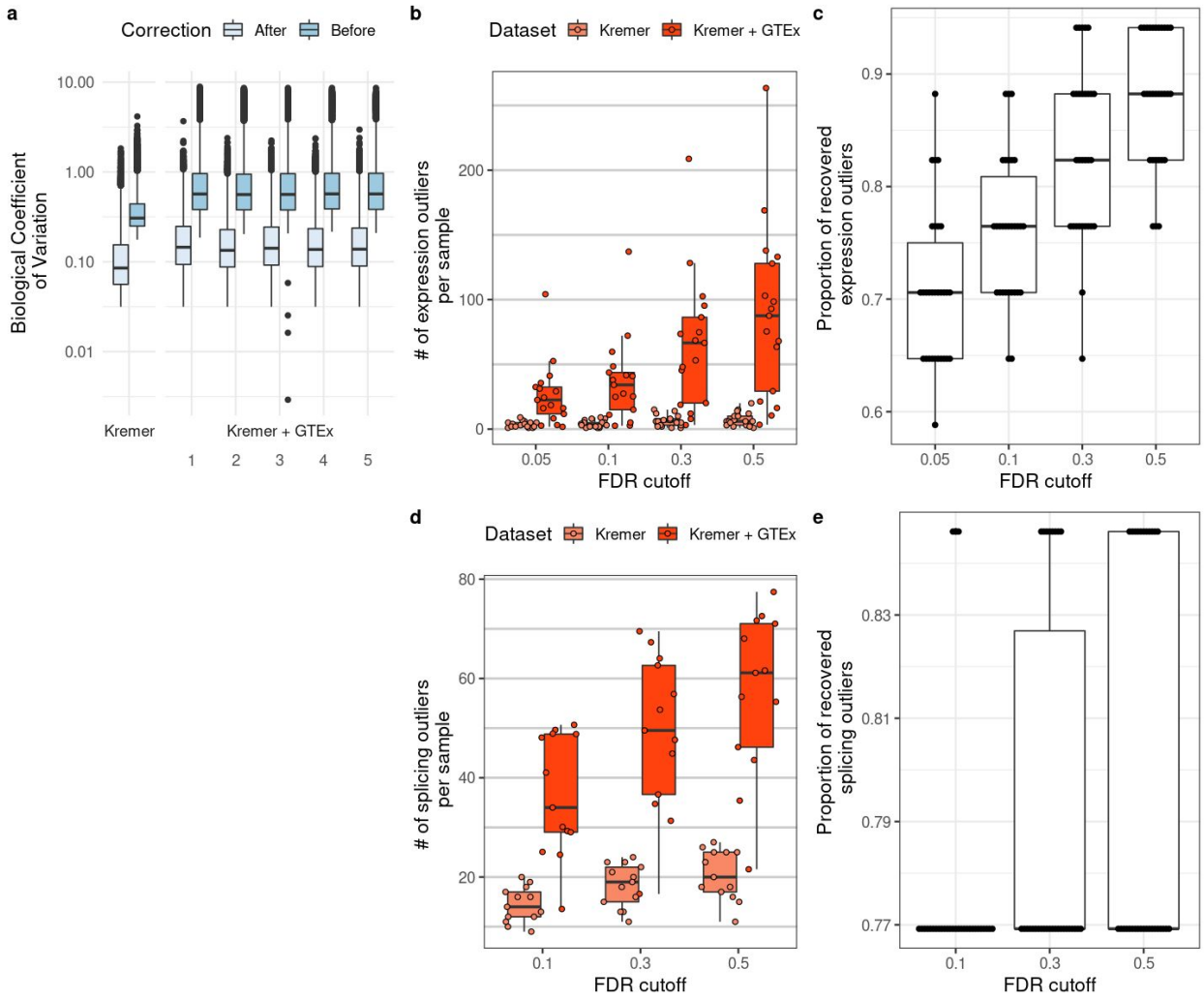

Fig S2 | Analysis of combination of patient samples with controls. **a,** Distribution of biological coefficient of variation before and after autoencoder correction shows that even though the correction yielded a lower biological coefficient of variation for all combinations, it worked better for the Kremer dataset alone. Five representative randomizations out of the 30 are shown. Each data point is a gene. **b,** Number of expression outliers per sample of the 17 true pathogenic outliers from the Kremer dataset when tested in the original dataset and in combination with random samples from GTEx skin not sun exposed. At an FDR cutoff of 0.05 the median expression outliers per sample is 3 for the Kremer dataset and 23 for the combined dataset. **c,** Proportion of the 17 true pathogenic expression outliers from Kremer recovered after combining them with GTEx. Different FDR cutoffs used. Each dot represents 1 randomization out of 30. **d,** Same as b) but using the 13 true pathogenic splicing outliers. At an FDR cutoff of 0.1 the median splicing outliers per sample is 14 for the Kremer dataset and 34 for the combined. **e,** Same as c) but for the 13 true pathogenic splicing outliers.

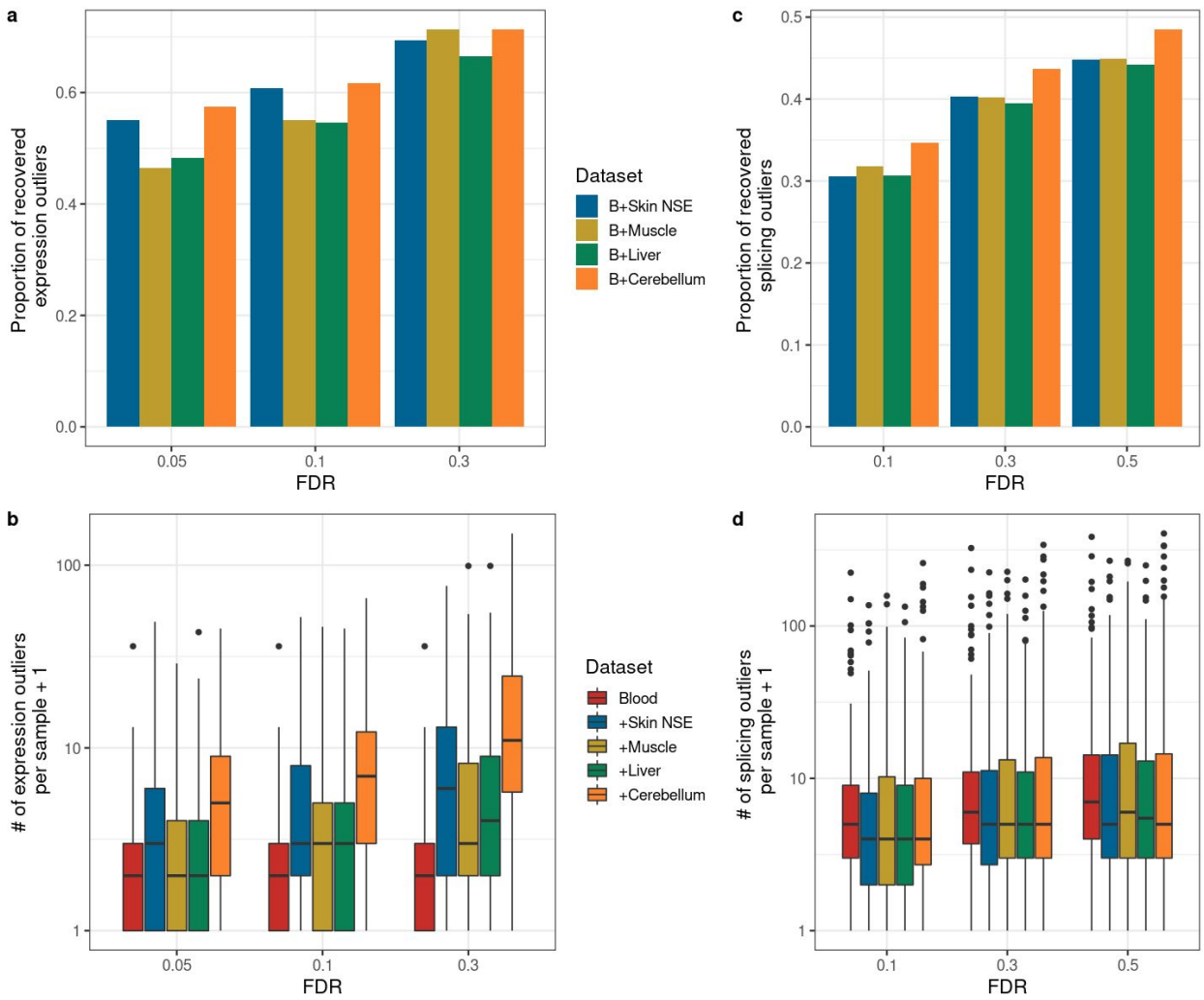

Fig S3 | Analysis of combination of different GTEx tissues. **a,** Proportion of recovered outliers after fitting samples of blood alone and after combining them with samples from skin not-sun-exposed, skeletal muscle, liver and brain cerebellum. Different FDR cutoffs used. **b,** Number of expression outliers + 1 for blood alone and after combining it with the same tissues as a). **c,** Same as a) but for splicing outliers. **d,** Same as b) but for splicing outliers.

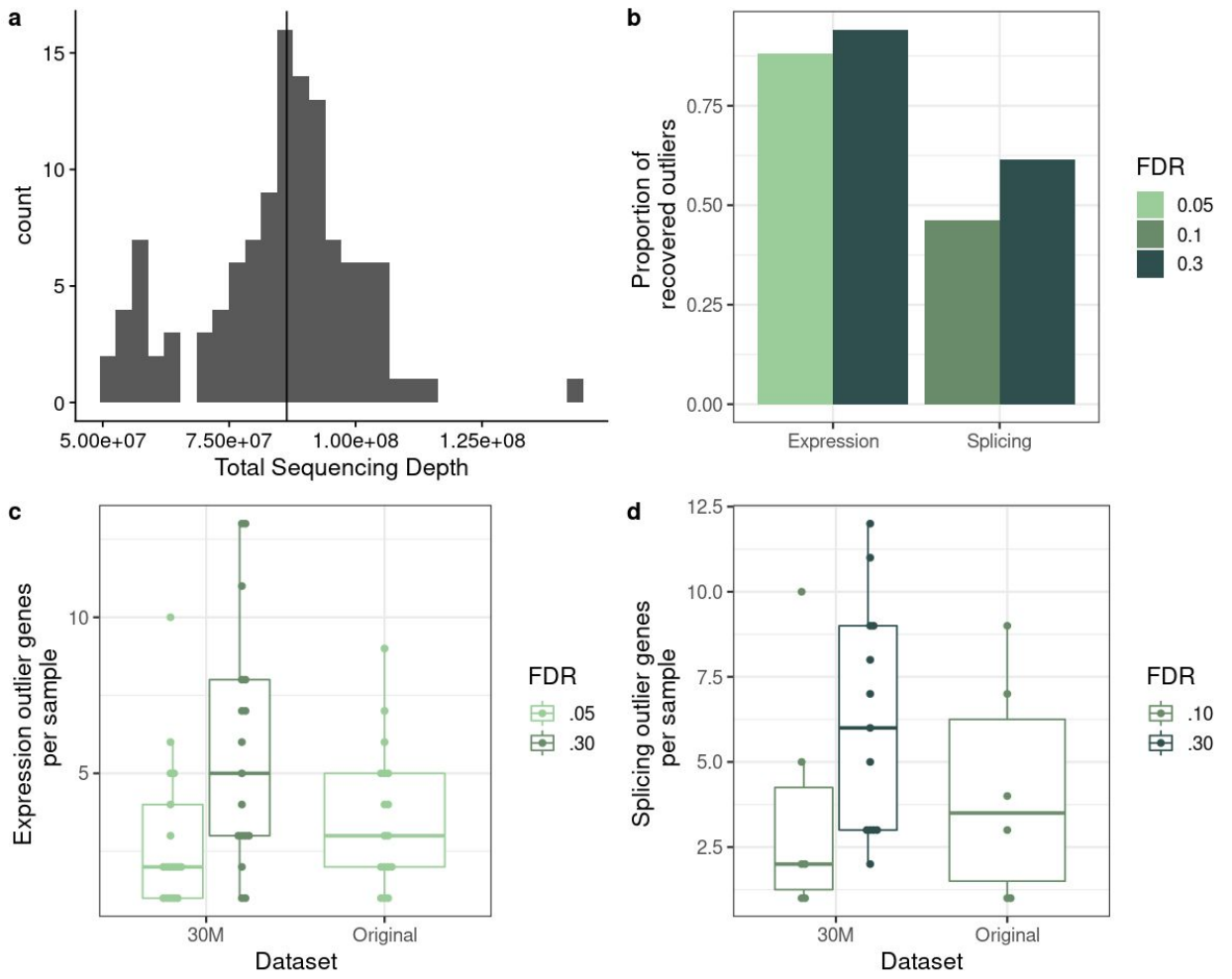

Fig S4 | Analysis of combination of different sequencing depths. **a,** Total RNA sequencing depth of the samples from the Kremer dataset (median ~86 million reads). **b,** Proportion of 17 true pathogenic expression outliers (and 13 splicing outliers) from the Kremer dataset simulated to have a sequencing depth of ~30 million reads, recovered after combining them with the rest of the dataset at its original depth depending on FDR cutoffs. **c,** Number of expression outlier genes per sample for the true positives in their original and 30 million read depth, using different cutoffs. **d,** Same as c) but for splicing outliers.

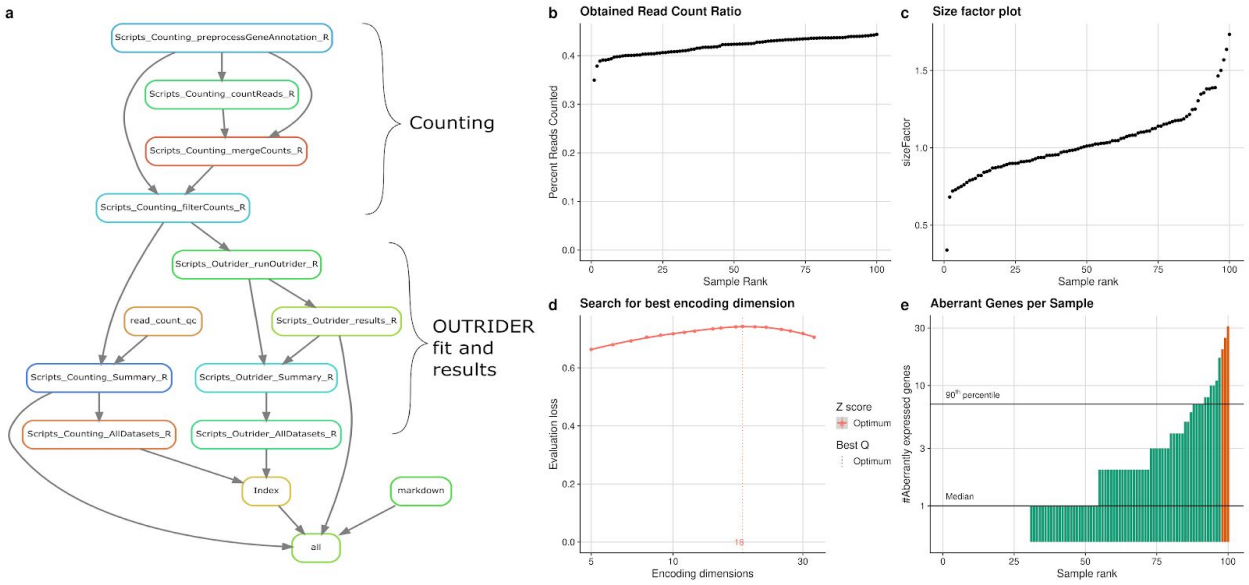

Fig S5 | Aberrant Expression Module **a,** Aberrant expression workflow. The two main steps are counting and running the OUTRIDER fit and results. **b,** Percentage of counted reads per sample. **c,** Sorted size factors. Size factors represent the relative sequencing depth of a certain sample. **d,** OUTRIDER evaluation loss for different encoding dimensions *q* and showing the optimal value (vertical dotted line). **e,** Number of aberrantly expressed genes per sample. Aberrant samples (orange) are samples with too many outlier genes.

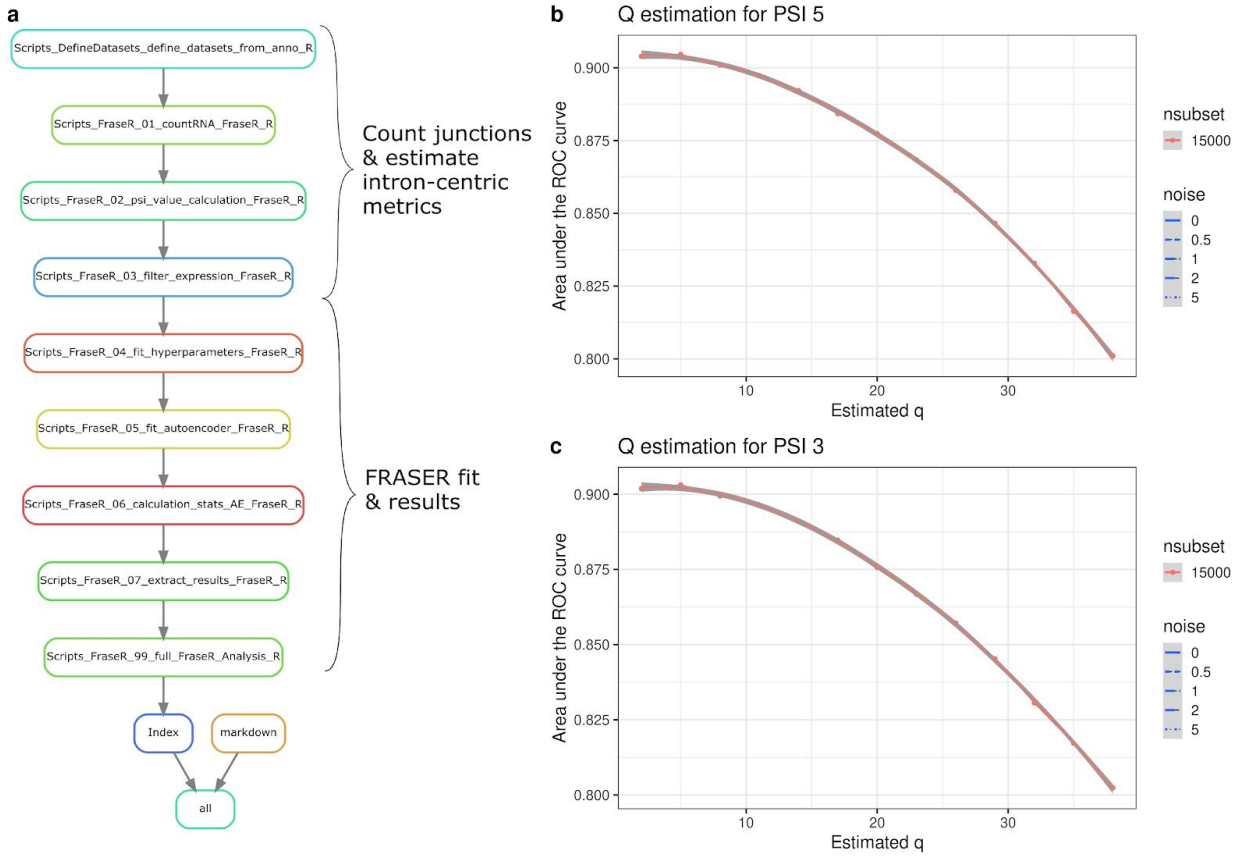

Fig S6 | Aberrant Splicing Module. **a,** Aberrant splicing workflow. The two main steps are counting the junctions and running the FraseR fit and results. **b-c,** Area under the ROC curve (y-axis) after different encoding dimensions (x-axis) and noise level injections for PSI 3 (b) and PSI 5 (c). The ranking of outliers was bootstrapped to yield 95% confidence bands.

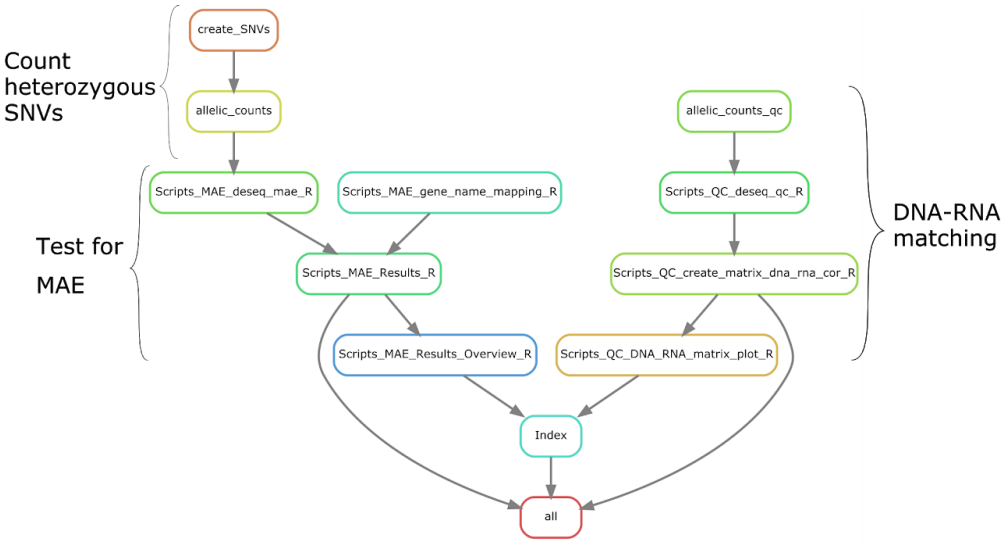

Fig S7 | Mono-allelic expression workflow. It is composed of two parts, the first one tests for heterozygous SNVs that are mono-allelically expressed and the second one matches VCF with BAM files.

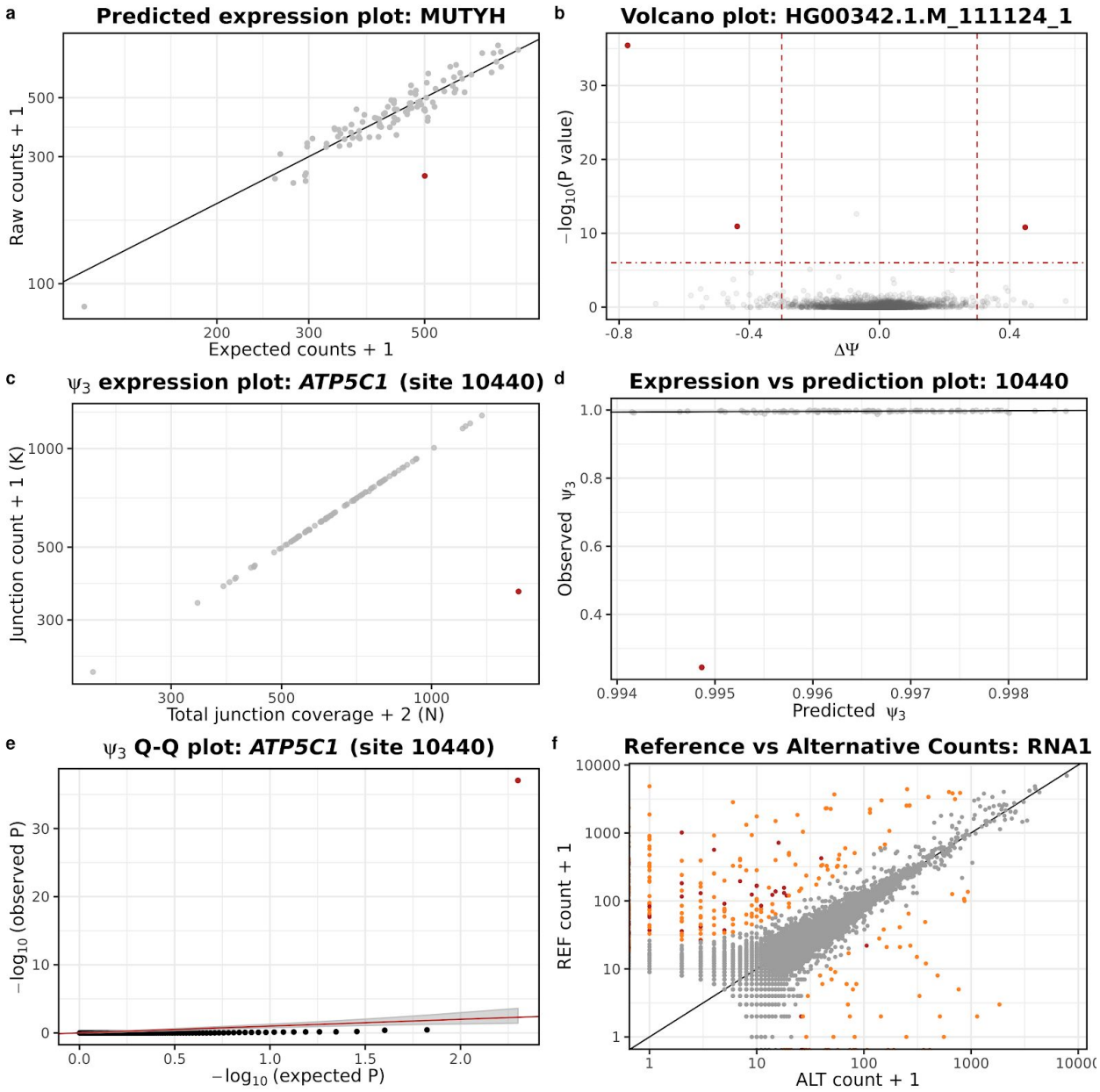

Fig. S8 | a, Raw (y-axis) versus expected (x-axis) counts of gene *MUTYH* showing one outlier (red). **b,** Negative log-transformed nominal *P* values (y-axis) versus $\Delta\psi_3$ values (x-axis) derived from all the splice sites (aggregated by gene) of sample HG00342.1.M_111124_1. Outliers are marked in red. **c,** Junction read counts (*K*, y-axis) plotted against the total split read coverage at the acceptor site (*N*, x-axis), of one junction in gene *ATP5C1*, which is the most severe outlier of panel (b). **d,** Observed (y-axis) versus expected (FRASER-predicted, x-axis) ψ_3 values of the same junction. **e,** Quantile-quantile plot of observed *P* values (-$\log_{10}$ scale, y-axis) against expected *P* values (-$\log_{10}$ scale, x-axis), with a 95% confidence band (gray) of the same junction. **f,** Counts assigned to the alternative (y-axis) versus reference allele (x-axis), highlighted by significance (orange) and significance and rarity (red), of sample HG00096.