The American Journal of Human Genetics, Volume *110*

## Supplemental information

# Improved detection of aberrant splicing with FRASER 2.0 and the intron Jaccard index

**Ines F. Scheller, Karoline Lutz, Christian Mertes, Vicente A. Yépez, and Julien Gagneur**

# Supplemental Figures

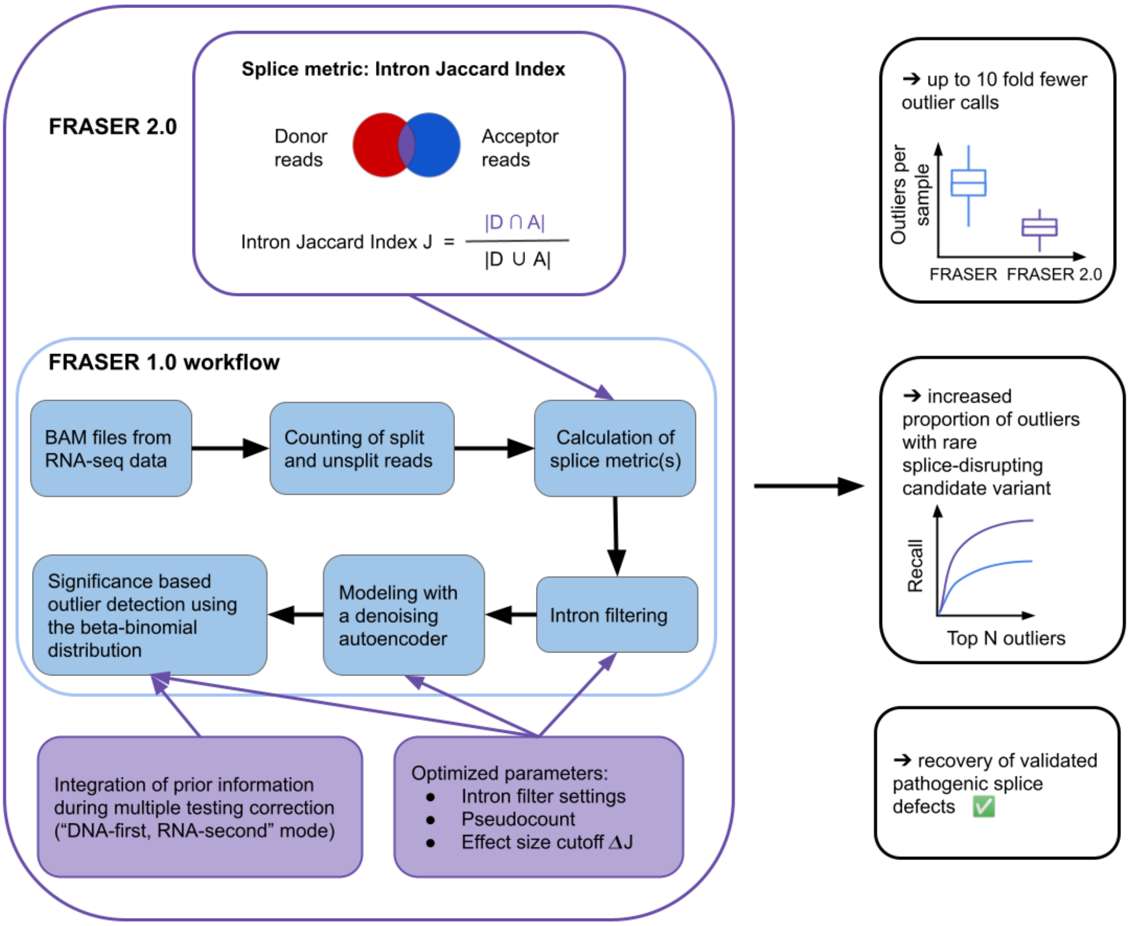
FRASER 2.0
Splice metric: Intron Jaccard Index
Donor reads
Acceptor reads
Intron Jaccard Index J = |D ∩ A| / |D ∪ A|
FRASER 1.0 workflow
BAM files from RNA-seq data
Counting of split and unsplit reads
Calculation of splice metric(s)
Intron filtering
Modeling with a denoising autoencoder
Significance based outlier detection using the beta-binomial distribution
Integration of prior information during multiple testing correction ("DNA-first, RNA-second" mode)
Optimized parameters:
● Intron filter settings
● Pseudocount
● Effect size cutoff ΔJ
➔ up to 10 fold fewer outlier calls
Outliers per sample
FRASER
FRASER 2.0
➔ increased proportion of outliers with rare splice-disrupting candidate variant
Recall
Top N outliers
➔ recovery of validated pathogenic splice defects ✅

**Figure S1. Schematic overview of the differences between FRASER 2.0 and FRASER.** Schematic overview highlighting the improvements of FRASER 2.0 (purple boxes) in comparison to the workflow of FRASER 1.0 (blue boxes). FRASER 2.0 uses the Intron Jaccard Index as its splice metric (top left), whereas FRASER 1.0 uses the metrics $\psi_3$, $\psi_5$ and $\theta$. Additionally, FRASER 2.0 introduces an option to integrate prior information during multiple testing correction and optimized parameters affecting the intron filtering, the modeling and outlier detection steps (bottom left). The benefits of these improvements are described on the right.

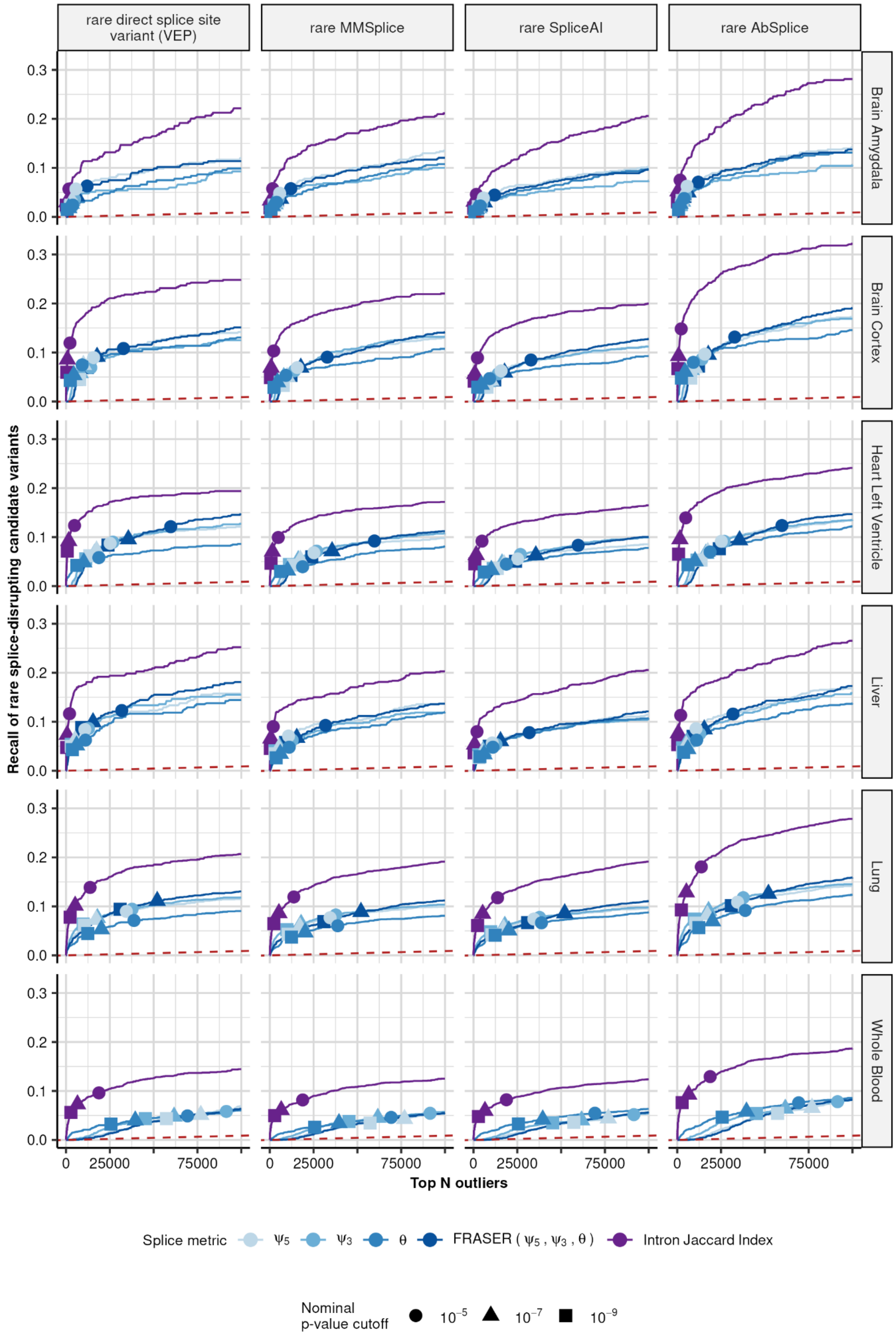
rare direct splice site variant (VEP)
rare MMSplice
rare SpliceAI
rare AbSplice
Brain Amygdala
Brain Cortex
Heart Left Ventricle
Liver
Lung
Whole Blood
Recall of rare splice-disrupting candidate variants
0.0
0.1
0.2
0.3
0
25000
75000
Top N outliers
Splice metric
ψ5
ψ3
θ
FRASER ( ψ5 , ψ3 , θ )
Intron Jaccard Index
Nominal p-value cutoff
10^-5
10^-7
10^-9

**Figure S2. Intron Jaccard Index increases recall of splice-disrupting candidate variants over FRASER's splice metrics on several GTEx tissues.**
Same as Fig. 1D), but for the FRASER adaption using Intron Jaccard Index (purple) compared to individual and combined metrics of FRASER ($\psi_3$, $\psi_5$, $\theta$) on several GTEx tissues (rows) and four different sets of rare splice-disrupting candidate variants (columns).

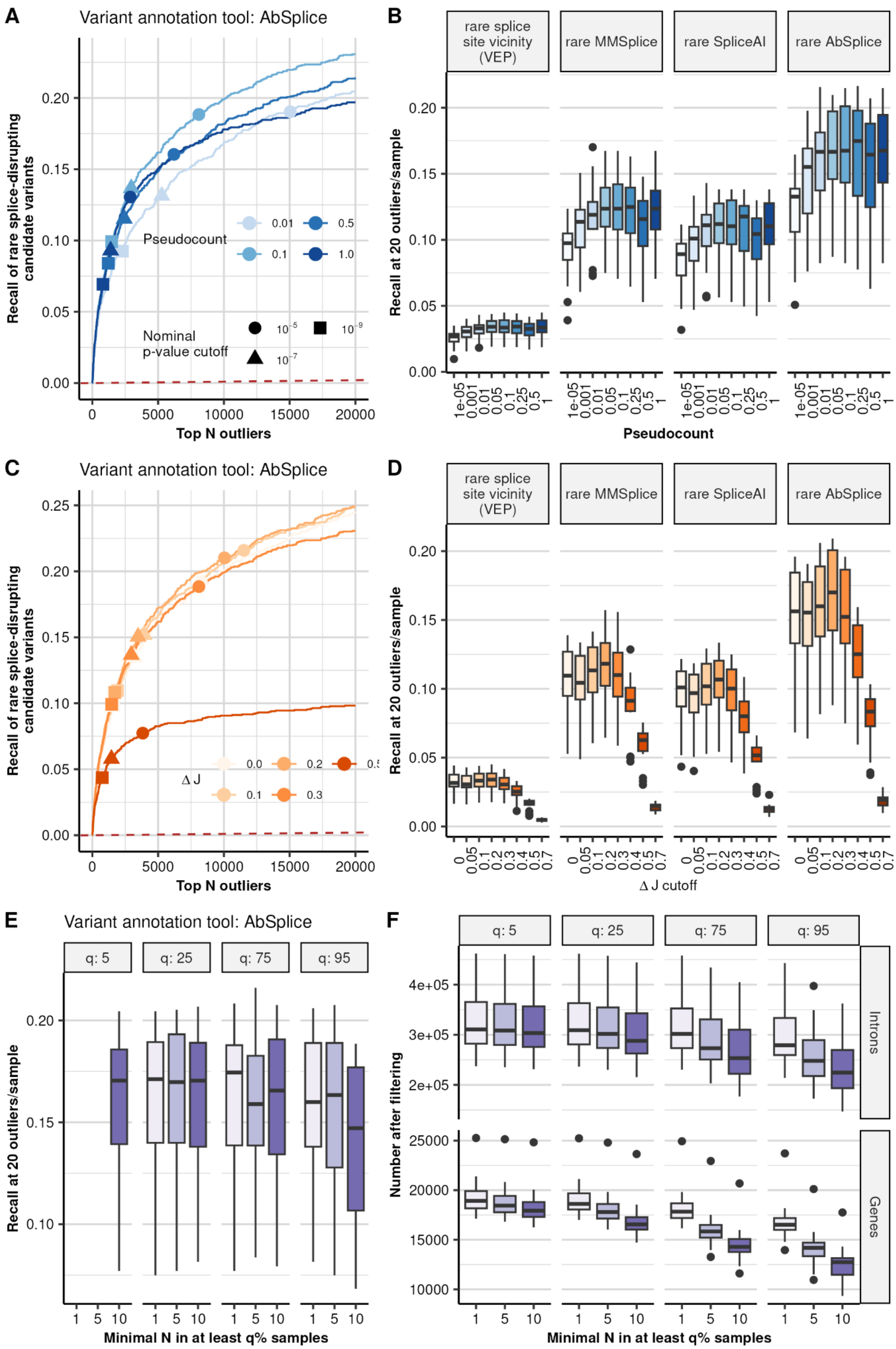
A
Variant annotation tool: AbSplice
Recall of rare splice-disrupting candidate variants
0.00
0.05
0.10
0.15
0.20
0
5000
10000
15000
20000
Top N outliers
Pseudocount
0.01
0.1
0.5
1.0
Nominal p-value cutoff
$10^{-5}$
$10^{-7}$
$10^{-9}$
B
rare splice site vicinity (VEP)
rare MMSplice
rare SpliceAI
rare AbSplice
Recall at 20 outliers/sample
0.00
0.05
0.10
0.15
0.20
1e-05
0.001
0.01
0.05
0.1
0.25
0.5
1
Pseudocount
C
Variant annotation tool: AbSplice
Recall of rare splice-disrupting candidate variants
0.00
0.05
0.10
0.15
0.20
0.25
0
5000
10000
15000
20000
Top N outliers
Δ J
0.0
0.1
0.2
0.3
0.5
D
rare splice site vicinity (VEP)
rare MMSplice
rare SpliceAI
rare AbSplice
Recall at 20 outliers/sample
0.00
0.05
0.10
0.15
0.20
0
0.05
0.1
0.2
0.3
0.4
0.5
0.7
Δ J cutoff
E
Variant annotation tool: AbSplice
q: 5
q: 25
q: 75
q: 95
Recall at 20 outliers/sample
0.10
0.15
0.20
1
5
10
Minimal N in at least q% samples
F
q: 5
q: 25
q: 75
q: 95
Introns
Genes
Number after filtering
2e+05
3e+05
4e+05
10000
15000
20000
25000
1
5
10
Minimal N in at least q% samples

**Figure S3. Identification of optimal FRASER 2.0 parameters pseudocount, Δ jaccard cutoff and filtering settings. (A)** Recall of rare splice-disrupting candidate variants as defined by AbSplice versus the rank of nominal *P*-values from FRASER with the Intron Jaccard Index metric using different values of the pseudocount on the GTEx skin not-sun-exposed dataset. Nominal *P*-value cutoffs are indicated with shapes. **(B)** Boxplots of the recall of rare splice-disrupting candidate variants at the rank corresponding to a value of 20 outliers per sample for different values of the pseudocount across 48 GTEx tissues. Facets indicate different tools to define the set of candidate splice-disrupting variants: VEP (annotated as splice donor/acceptor or splice region), MMSplice (absolute MMSplice Δlogit $\Psi \geq 2$), SpliceAI (SpliceAI score ≥ 0.5) and AbSplice (max. AbSplice score ≥ 0.05). **(C)** Same as (A) but comparing different cutoff values on the predicted Δ Intron Jaccard Index for FRASER with the Intron Jaccard Index metric with pseudocount set to 0.1. **(D)** Same as (B) but comparing different cutoff values on the predicted Δ Intron Jaccard Index for FRASER with the Intron Jaccard Index metric with pseudocount set to 0.1. **(E)** Same as (B) but comparing different filtering settings. Facets indicate the quantile and x-axis indicate the minimal value of N at this quantile to pass the filter. **(F)** Boxplots of the number of introns (top row) and genes (bottom row) after applying the respective filtering setting defined by the quantile (columns) and the minimal value of N at this quantile (x-axis). Boxplots: center line = median; box limits = first and third quartiles; whiskers span all data within 1.5 interquartile ranges of the lower and upper quartiles (applicable here and in all the following figures).

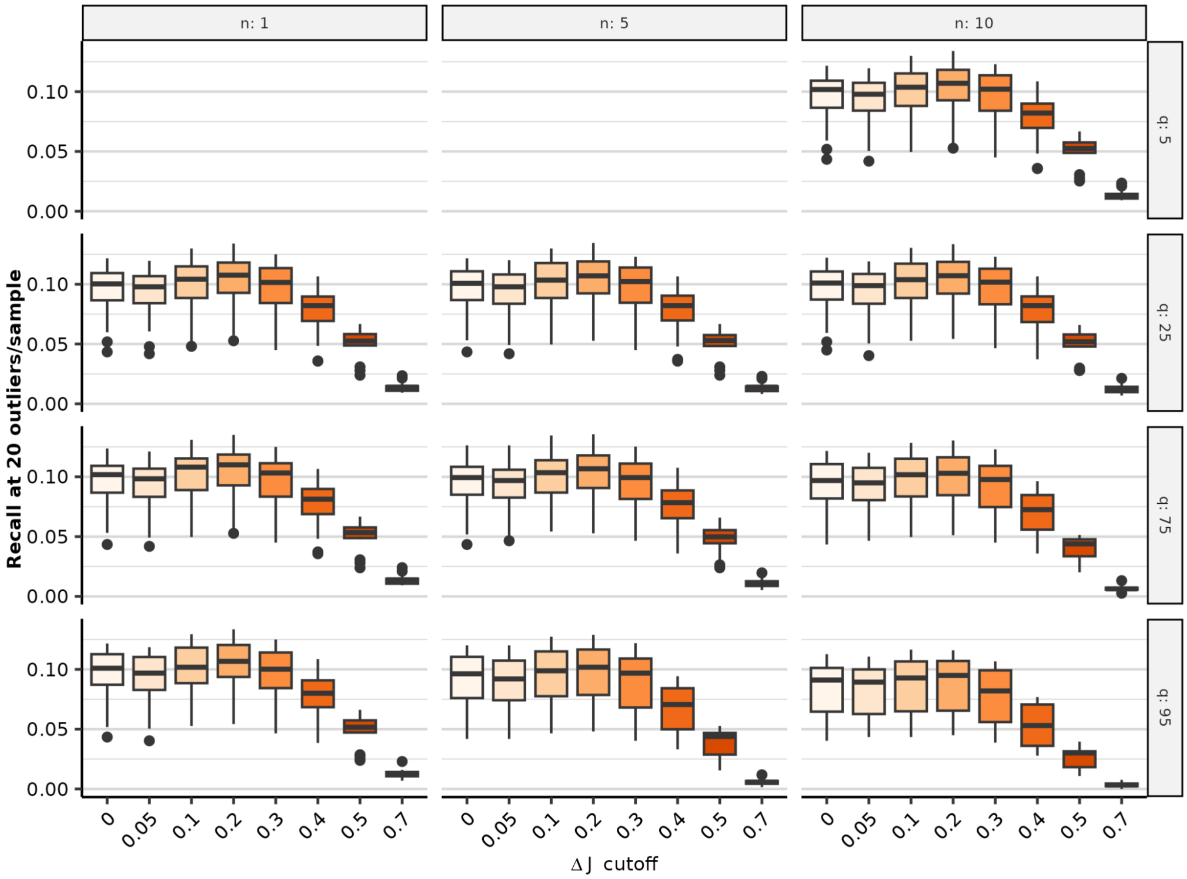

**Figure S4. Combined parameter optimization results for different combinations of ΔJ and filtering cutoffs.**

Boxplots of the recall of rare splice-disrupting candidate variants as predicted with SpliceAI at the rank corresponding to a value of 20 outliers per sample for different values of the ΔJ cutoff across 48 GTEx tissues. Facets indicate different intron filtering settings, defined by the minimal required n (columns) in at least q% of the samples (rows).

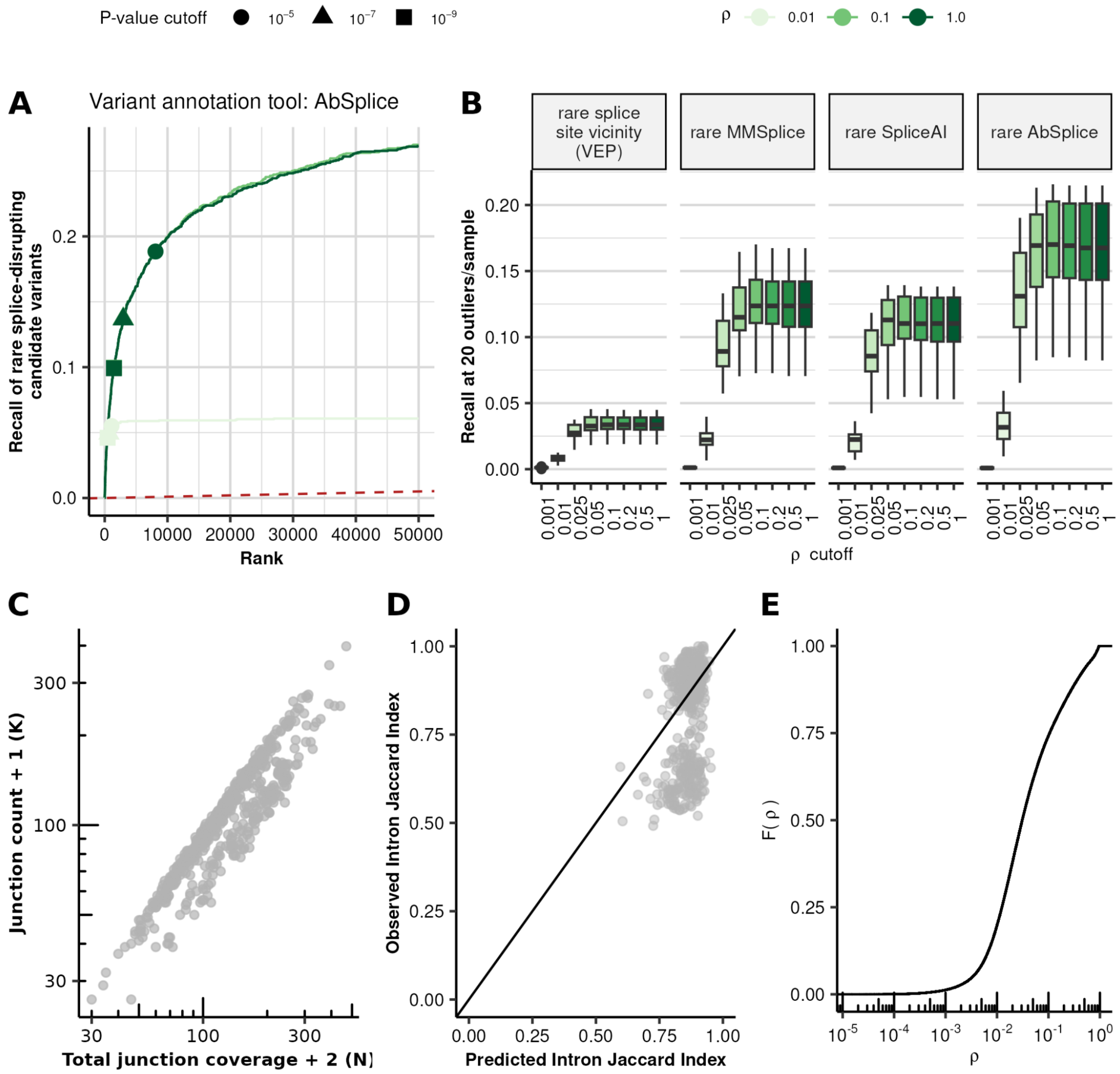

**Figure S5. Goodness-of-fit cutoff does not improve splicing outlier calls. (A)** Recall of rare (MAF < 0.001) splice-disrupting candidate variants as defined by AbSplice versus the rank of nominal *P*-values for FRASER with the Intron Jaccard Index metric and pseudocount of 0.1 for different values of the goodness-of-fit cutoff on $\varrho$ (overdispersion parameter of the beta-binomial distribution, shown in shades of green). Nominal *P*-value cutoffs are indicated with shapes. **(B)** Boxplots of the recall of rare splice-disrupting candidate variants at the rank corresponding to a value of 20 outliers per sample for different values of the goodness-of-fit cutoff across 48 GTEx tissues. Facets indicate different tools to define the set of candidate splice-disrupting variants: VEP (annotated as splice donor/acceptor or splice region), MMSplice (absolute MMSplice Δlogit $\Psi \geq 2$), SpliceAI (SpliceAI score ≥ 0.5) and AbSplice (max. AbSplice score ≥ 0.05). **(C)** Intron counts (y-axis) against the denominator of the jaccard metric (x-axis) of intron chr1:1485839-1486109:+ affected by a sQTL loci in cis across samples in GTEx skin (suprapubic). **(D)** Predicted against observed Intron Jaccard Index values for the intron shown in (C). **(E)** Empirical cumulative density function of $\varrho$ on GTEx suprapubic skin tissue with the Intron Jaccard Index metric and pseudocount of 0.1.

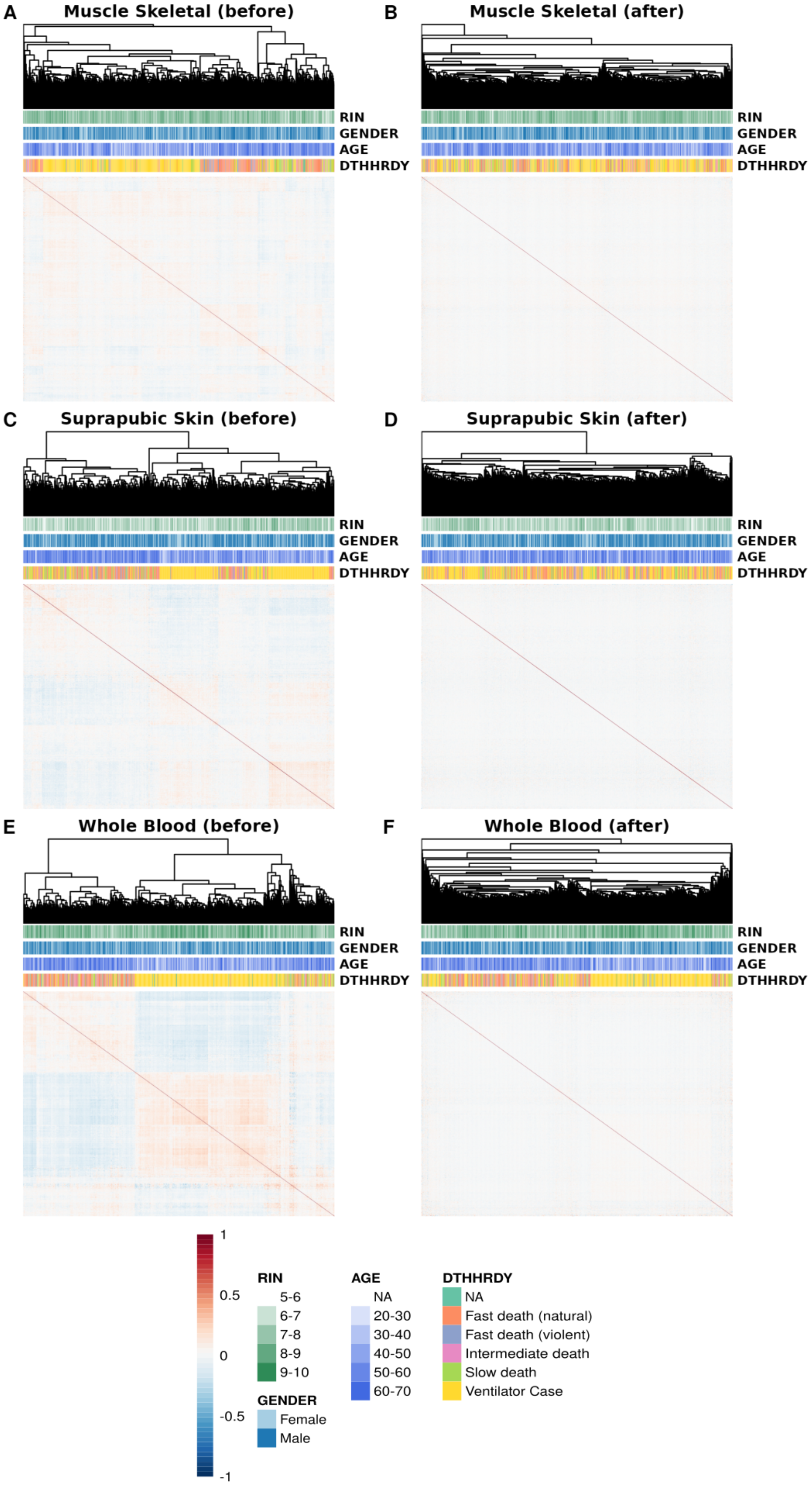
A
Muscle Skeletal (before)
B
Muscle Skeletal (after)
C
Suprapubic Skin (before)
D
Suprapubic Skin (after)
E
Whole Blood (before)
F
Whole Blood (after)
RIN
GENDER
AGE
DTHHRDY
1
0.5
0
-0.5
-1
RIN
5-6
6-7
7-8
8-9
9-10
GENDER
Female
Male
AGE
NA
20-30
30-40
40-50
50-60
60-70
DTHHRDY
NA
Fast death (natural)
Fast death (violent)
Intermediate death
Slow death
Ventilator Case

**Figure S6. Sample correlation heatmaps.** Heatmaps of sample-sample correlations of Intron Jaccard Index metric before **(A, C, E)** and after **(B, D, F)** FRASER 2.0's autoencoder correction on three GTEx tissues: muscle skeletal (**A, B**, *N*=782), not sun-exposed suprapubic skin (**C, D**, *N*=582) and whole blood (**E, F**, *N*=735). A dendrogram of the sample clustering is shown on top of each heatmap alongside sample metadata: RNA integrity number (RIN), age, gender, and cause of death (Hardy scale classification, DTHHRDY).

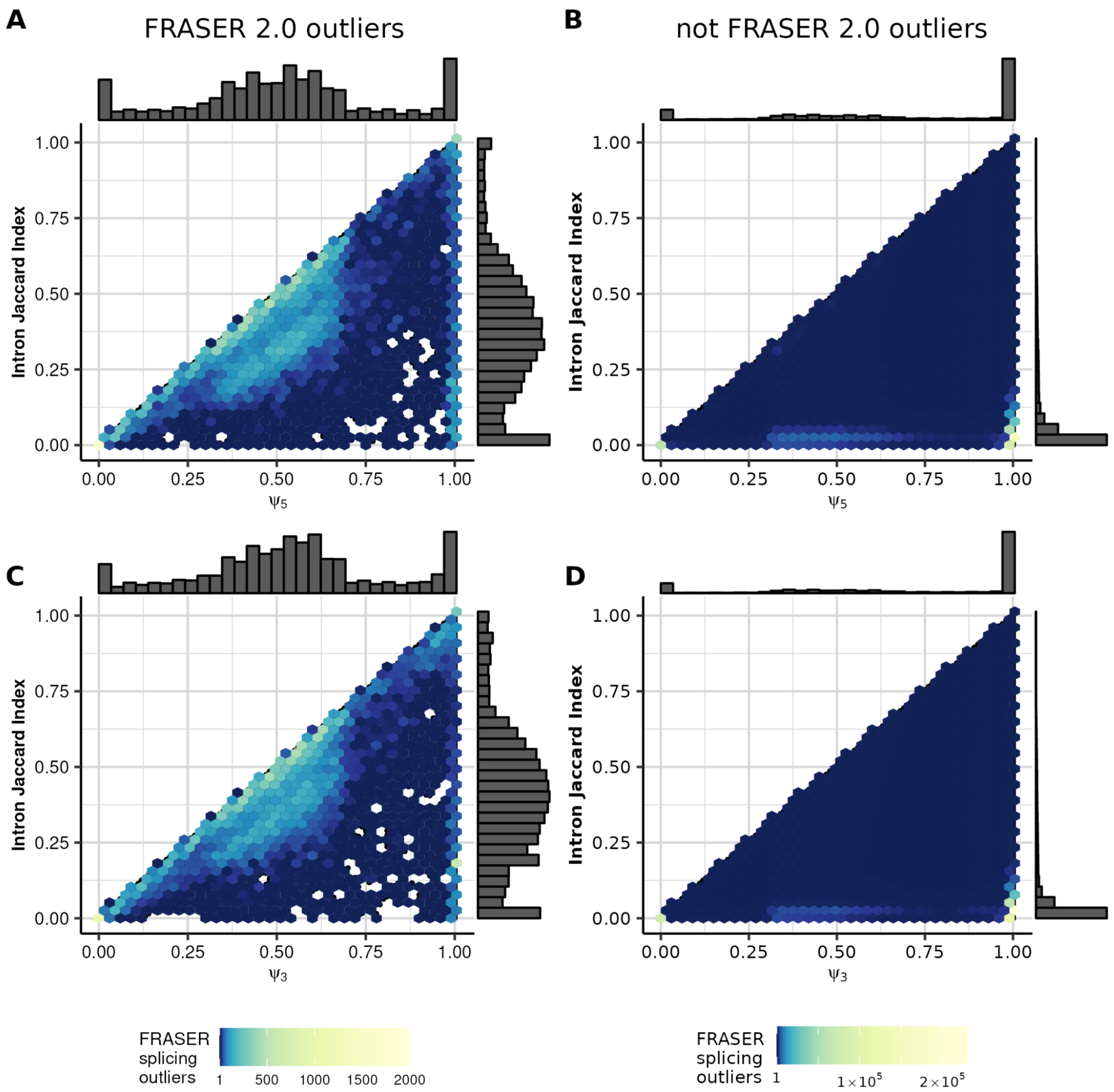

**Figure S7. Comparison of Intron Jaccard Index values to FRASER's $\psi_5$ and $\psi_3$ metrics of FRASER outliers.**

Scatterplot with density of $\psi_5$ values of FRASER $\psi_5$ outliers against Intron Jaccard Index values across GTEx tissues for introns that are also reported as outliers by FRASER 2.0 **(A)** and introns that are not outliers in FRASER 2.0 **(B). (C, D)** Same as (A, B), but for $\psi_3$ values. Outliers that are found both by FRASER and FRASER 2.0 **(A, C)** tend to lie alongside the diagonal, whereas most FRASER outliers that are not reported by FRASER 2.0 **(B, D)** have small values in the Intron Jaccard Index metric while having $\psi_5$ or $\psi_3$ close to 1.

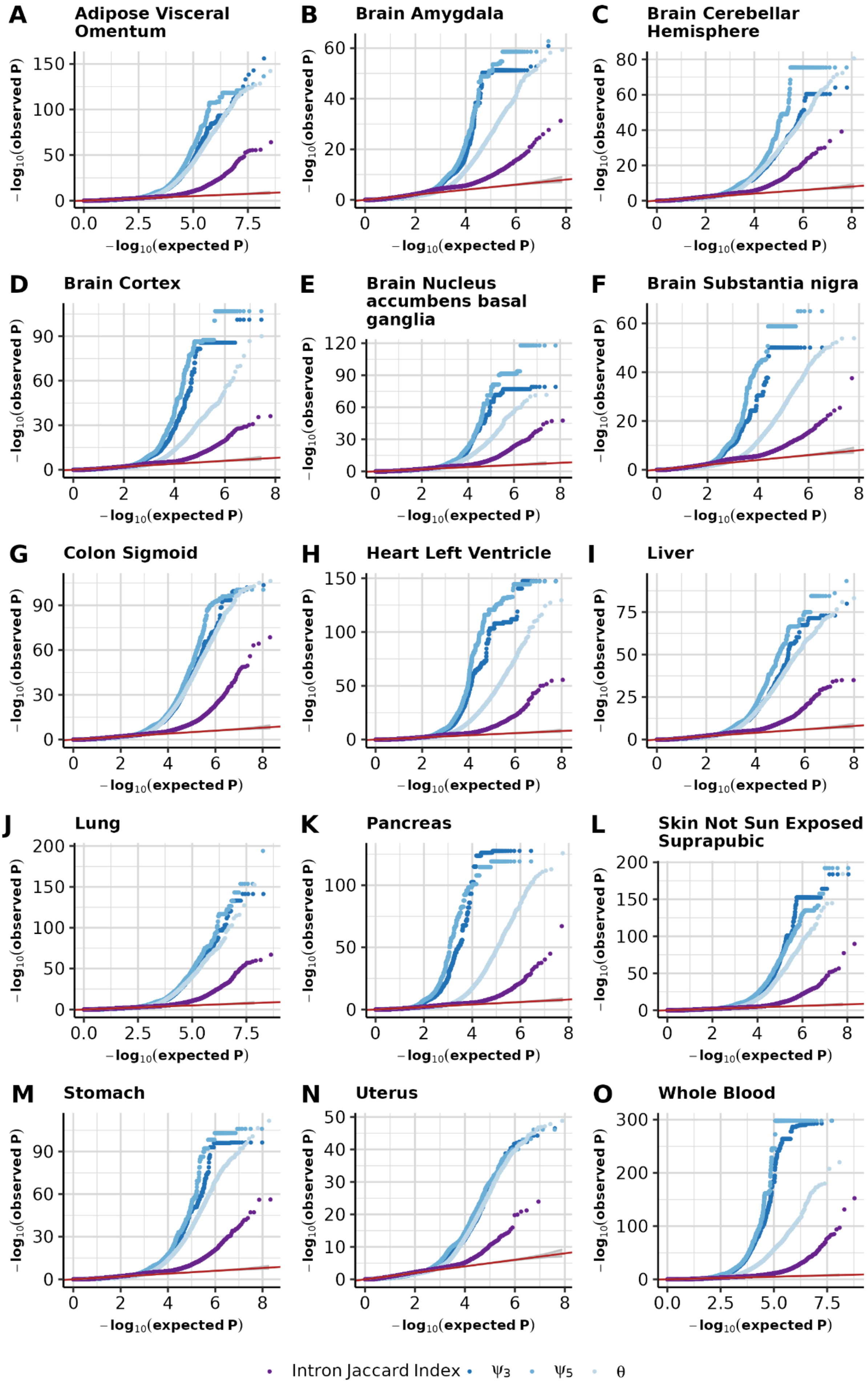
A
Adipose Visceral Omentum
B
Brain Amygdala
C
Brain Cerebellar Hemisphere
D
Brain Cortex
E
Brain Nucleus accumbens basal ganglia
F
Brain Substantia nigra
G
Colon Sigmoid
H
Heart Left Ventricle
I
Liver
J
Lung
K
Pancreas
L
Skin Not Sun Exposed Suprapubic
M
Stomach
N
Uterus
O
Whole Blood
$-\log_{10}$(observed P)
$-\log_{10}$(expected P)
0.0
2.5
5.0
7.5
0
2
4
6
8
0
50
100
150
0
20
40
60
80
0
30
60
90
120
0
25
50
75
200
300
0
10
20
30
40
50
Intron Jaccard Index
$\psi_3$
$\psi_5$
$\theta$

**Figure S8. Quantile-quantile plots of FRASER 2.0 *P*-values.** Quantile-quantile plots of expected against observed *P*-values obtained using the 3 splice metrics from FRASER (different shades of blue) and the Intron Jaccard Index of FRASER 2.0 (purple) on 15 GTEx tissues **(A-O)**. Under the null hypothesis, the data are expected to lie along the diagonal (red, 95% confidence bands in gray).

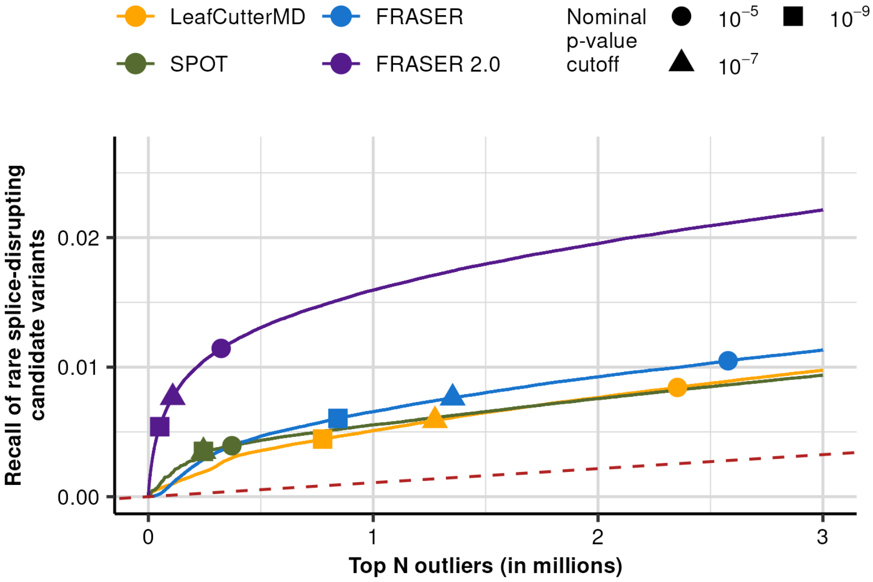

**Figure S9. Recall of rare variants in the splice site vicinity as defined by VEP.**
Recall of rare splice-disrupting candidate variants as defined by the variant annotation tool VEP versus the rank of nominal *P*-values combined across GTEx tissues for FRASER (blue), FRASER 2.0 (purple), LeafCutterMD (yellow), and SPOT (green). Nominal *P*-value cutoffs are indicated with shapes.

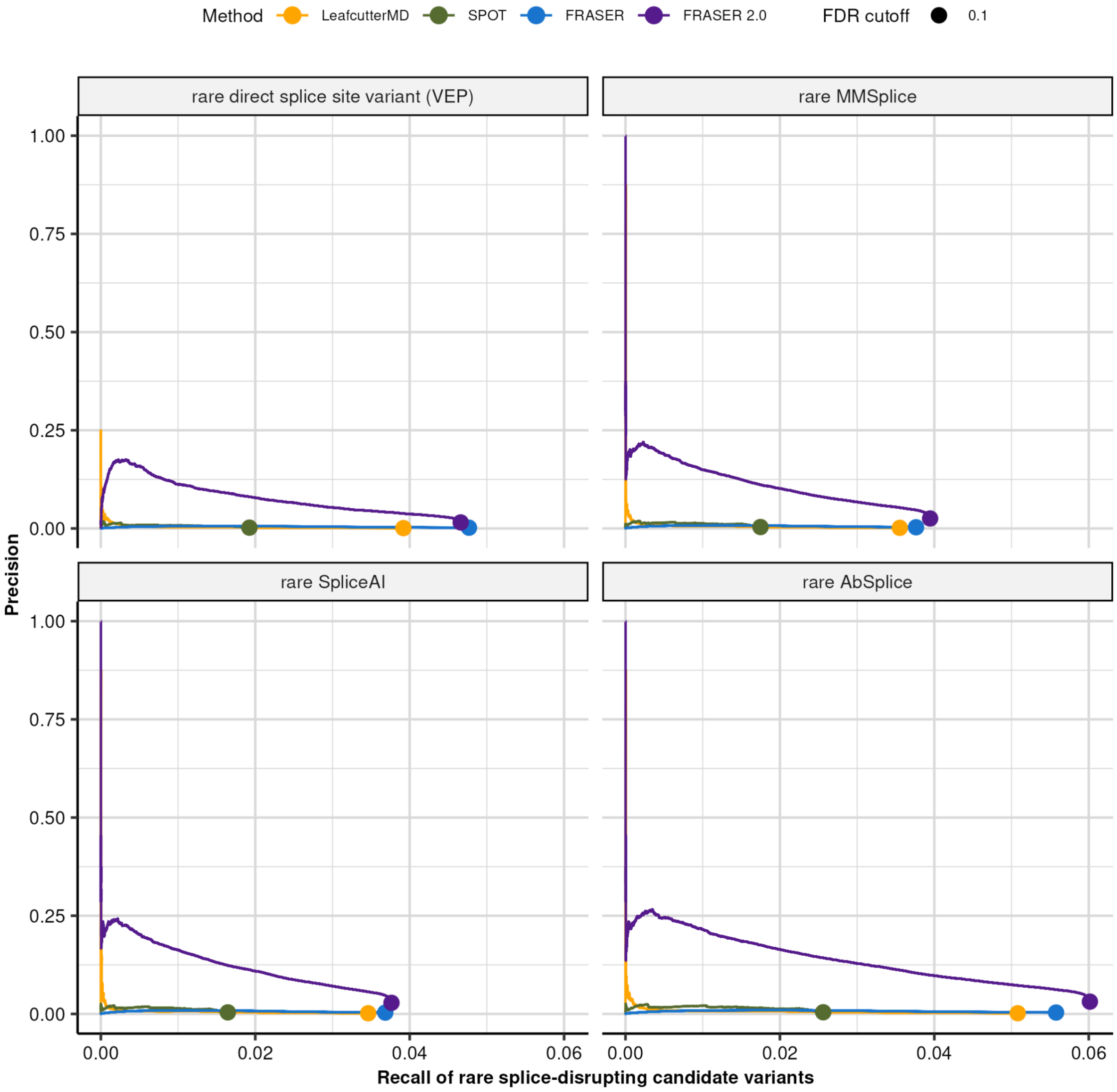

**Figure S10. Improved precision of FRASER 2.0 at FDR cutoff.** Precision-recall plot on candidate rare splice-disrupting variants as defined by the variant annotation tools VEP, MMSplice, SpliceAI, and AbSplice (facets) on nominal P-values combined across GTEx tissues for FRASER (blue), FRASER 2.0 (purple), LeafcutterMD (yellow), and SPOT (green) for significant results at FDR ≤ 0.1. The FDR cutoff is indicated with a circle.

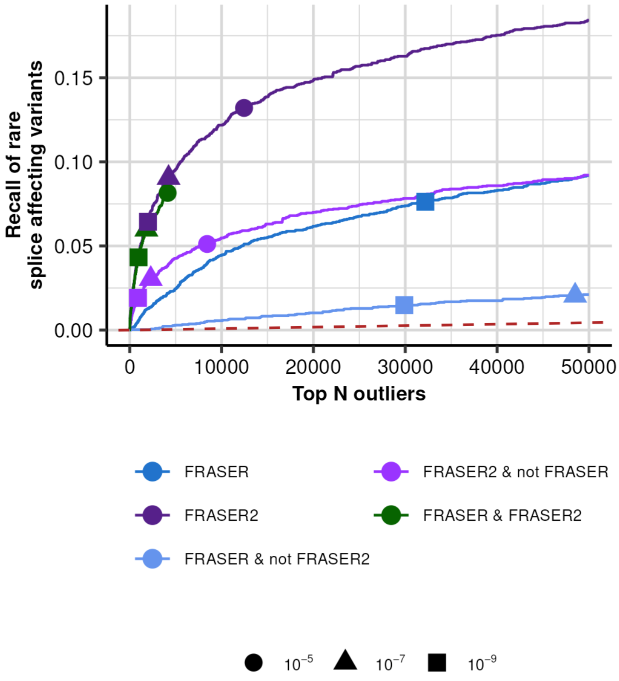

**Figure S11. Recall of splice-disrupting candidate variants by regions of the Venn diagram for FRASER and FRASER 2.0.** Recall of rare splice-disrupting candidate variants as defined by SpliceAI score 0.5 versus the rank of nominal *P*-values combined across GTEx tissues for different regions of the Venn diagram between FRASER (middle blue) and FRASER 2.0 (dark purple) with FRASER-only outliers in light blue, FRASER 2.0 only outliers in light purple and both FRASER and FRASER 2.0 outliers in green. Different nominal *P*-value cutoffs are indicated with shapes.

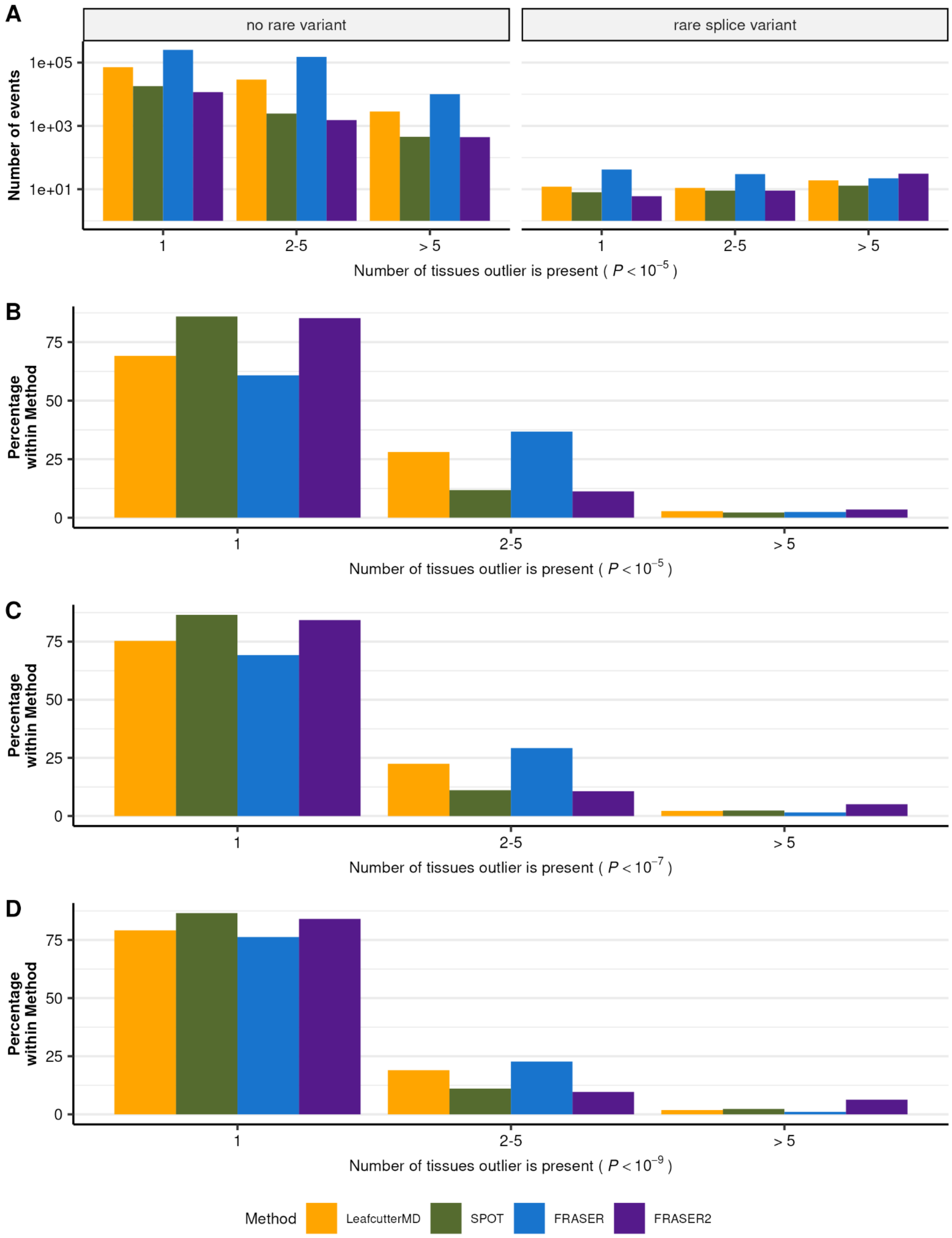

**Figure S12. Reproducibility of splicing outlier calls across GTEx tissues.**
**(A)** Barplot of the number of gene-level splicing outliers (y-axis) against their reproducibility (x-axis) across GTEx tissues. The reproducibility is defined as the number of tissues an event is observed at a nominal *P*-value < $10^{-3}$ given it was observed at least once at a nominal *P*-value < $10^{-5}$. Data is stratified by associated variant status (defined by VEP) and grouped by method. **(B)** Same as (A) but plotted as the proportion (y-axis) of reproducible gene-level splicing outlier calls in GTEx tissues. **(C-D)** Same as (B) but with at least one call at a nominal *P*-value < $10^{-7}$ (C) and nominal *P*-value < $10^{-9}$ (D).

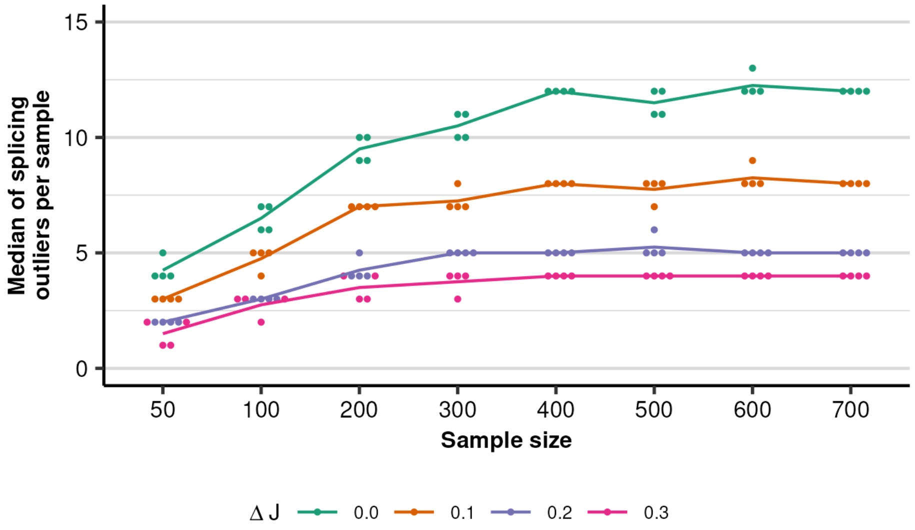

**Figure S13. Splicing outliers per sample comparison with sample size.**
For the skeletal muscle dataset in GTEx (*N*=782), the median of splicing outliers across all samples using FRASER 2.0 (y-axis) is plotted against the simulated sample size (x-axis) for different cutoffs on the effect size ΔJ. The lines connect the average per sample size and effect size ΔJ.  Each sample size was sampled 5 times.

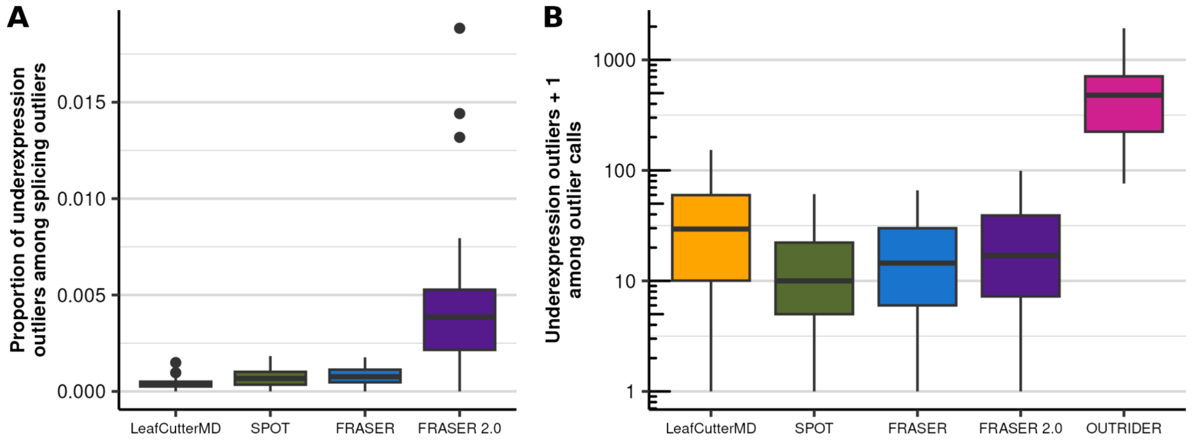

**Figure S14. Gene underexpression outliers among splicing outliers.** Proportion **(A)** and total number **(B)** of gene underexpression outliers (FDR < 0.05) per GTEx tissue detected with OUTRIDER among all splicing outliers reported by different aberrant splicing detection methods. The distribution of the number of outliers called by OUTRIDER is also shown for reference (rightmost box).

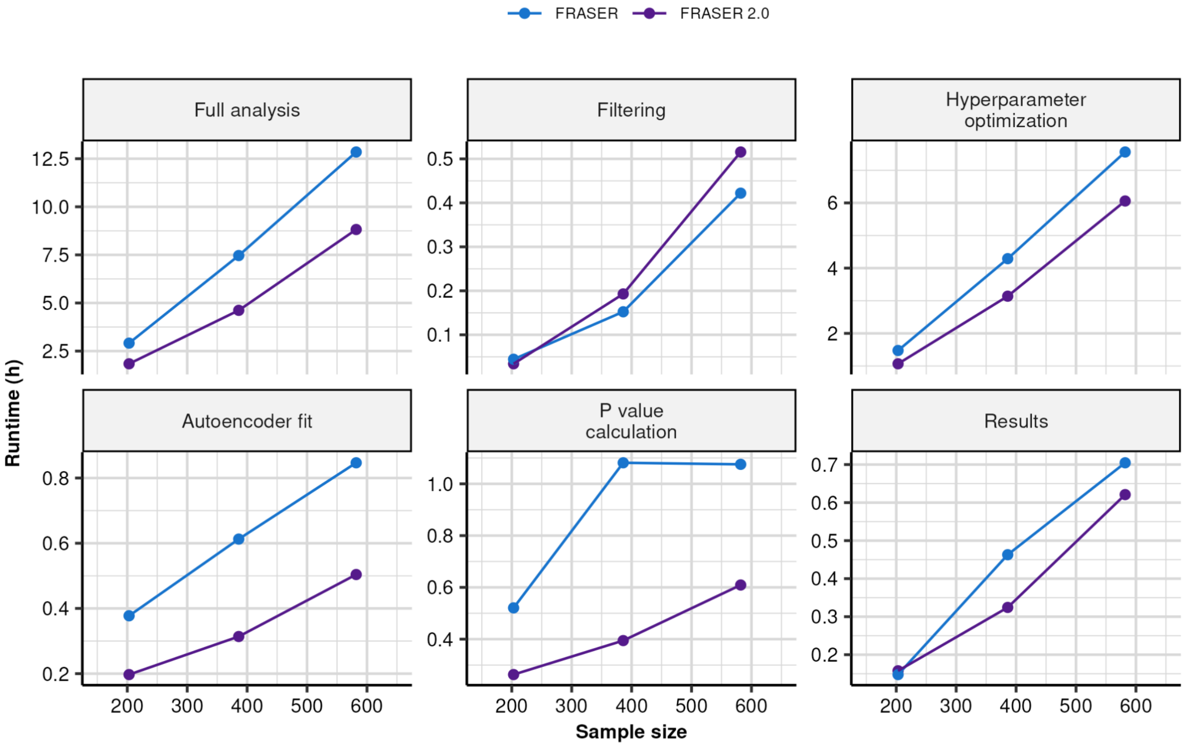

**Figure S15. Runtime comparison of FRASER 2.0 and FRASER**. Runtime in hours on the GTEx tissues Liver (N=203), Heart left ventricle (N=386) and Skin not sun exposed suprapubic (N=582) for different computational steps of FRASER (blue) and FRASER 2.0 (purple).

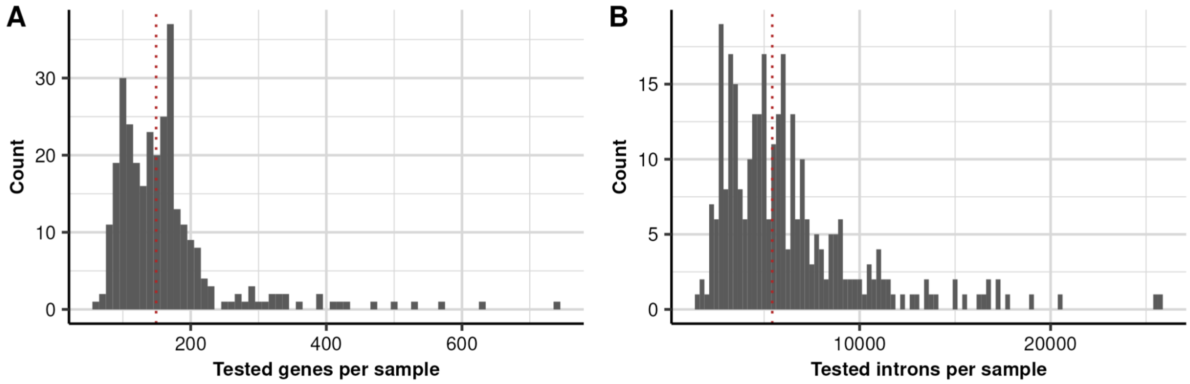

**Figure S16. Tested features per sample on the OMIM + rare variant subset.** Distribution of the total number of tested genes **(A)** and introns **(B)** per sample on the Yépez et al. dataset when including only OMIM genes harboring a rare variant. The red line denotes the median across samples: genes=149 and introns=5,427.

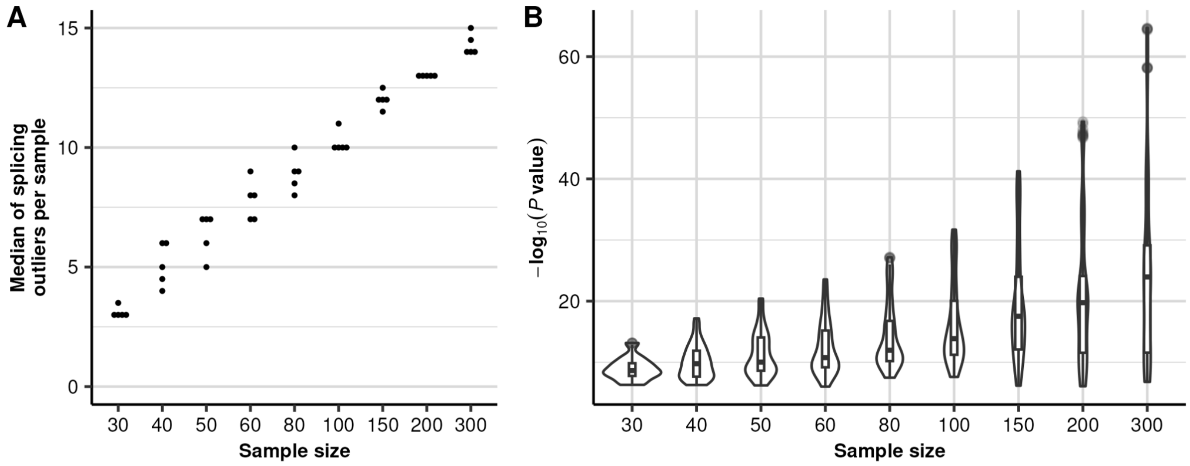

**Figure S17. Power analysis of FRASER 2.0 on the Yépez et al. dataset (*N*=303). (A)** The median of splicing outliers across all samples using default cutoffs (y-axis) is plotted against the taken sample size (x-axis). Each sample size was sampled 5 times. **(B)** The negative log$_{10}$ *P*-value for all known disease-causing splicing outliers (y-axis) is plotted against the taken sample size (x-axis). The violin depicts the density of the data points.

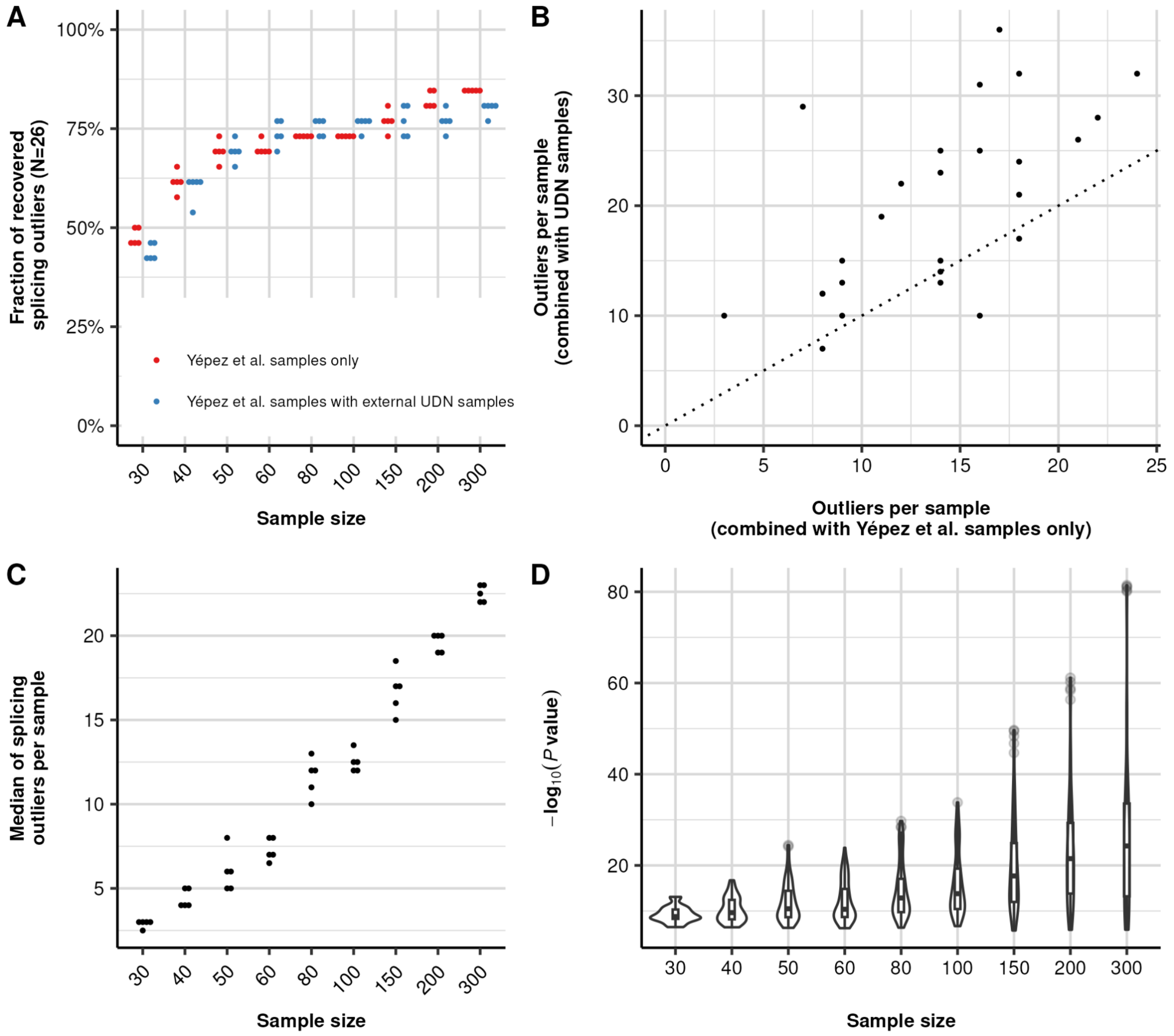

**Figure S18. Power analysis of FRASER 2.0 on the solved cases of the Yépez et al. dataset combined with external samples from UDN. (A)** Fraction of recovered pathogenic splicing outliers from the Yépez et al. dataset (y-axis, total *N*=26) when subsampling to different sample sizes (x-axis), either combining with samples from the same dataset (red) or with external samples (blue). Each sample size was randomly sampled 5 times. **(B)** FRASER 2.0 outlier calls per sample at the gene-level for the 26 samples with validated pathogenic splicing defects from Yépez et al., when combined either with 274 samples from Yépez et al. (x-axis), or with 274 samples from UDN (y-axis, median across 5 runs for each). The dotted line indicates the diagonal. **(C) (D)** Same as Fig. S17A,B, but for the dataset combining the cases with pathogenic events from the Yépez et al. dataset with the external UDN samples.

## Supplemental Tables

**Table S1: FRASER 2.0 results for the 26 validated pathogenic splice defects from Yépez et al.** This table contains information about the 26 pathogenic splice defects identified in the Yépez et al. study, annotated with the results of FRASER 2.0 for these samples.