e:Med SYSTEMS MEDICINE
Federal Ministry of Education and Research

Detection of aberrant splicing events in RNA-seq data with FRASER

Christian Mertes[1*], **Ines Scheller**[1*], Vicente A. Yépez[1,2], Muhammed H. Çelik[1], Yingjiqiong Liang[1], Laura S. Kremer[3,4], Mirjana Gusic[3,4], Holger Prokisch[3,4], Julien Gagneur[1,2]

[1]Computational Genomics, Department of Informatics, Technical University of Munich, [2] Quantitative Biosciences Munich, [3]Institute for Human Genetics, Helmholtz Center Munich, [4]Institute of Human Genetics, Klinikum rechts der Isar, Technical University of Munich e-mail: mertes@in.tum.de, ines.scheller@in.tum.de

Overview

Motivation:

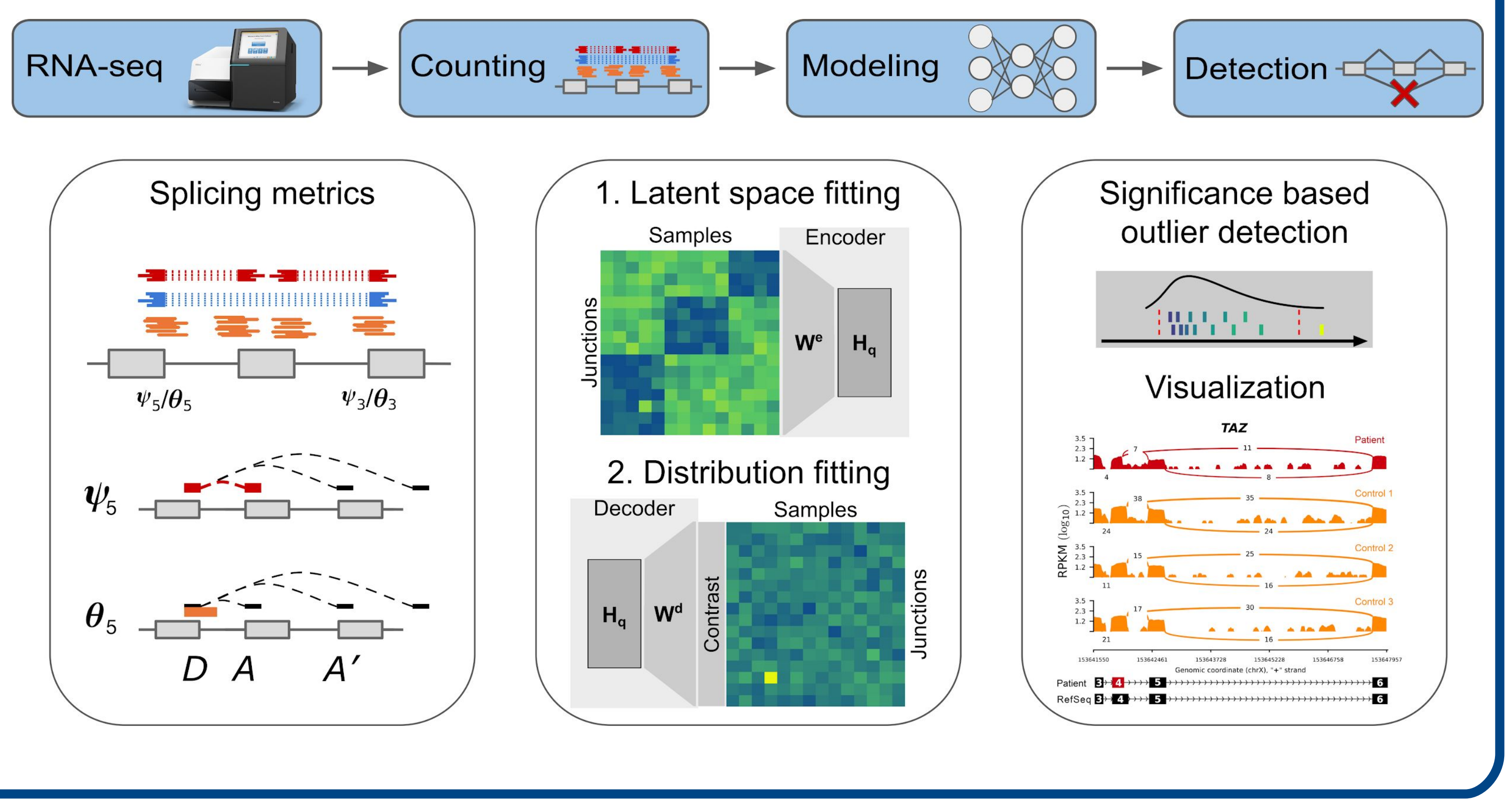
- Detection of aberrant splicing events is relevant e.g. for identifying potential causes of rare disorders, but can also be interesting in common diseases such as cancer
- Existing approaches do not assess statistical significance or do not properly control for confounders

Data:

- GTEx[1] V7
- Kremer[2] dataset: 119 samples from individuals with a suspected mitochondrial disorder

Results:

- FRASER provides automatic control for confounders and computes beta-binomial p-values which improves the detection of aberrant splicing events
- Application of FRASER to the Kremer dataset led to a new prioritization of a previously overlooked variant in *TAZ*



RNA-seq → Counting → Modeling → Detection

Splicing metrics

1. Latent space fitting

2. Distribution fitting

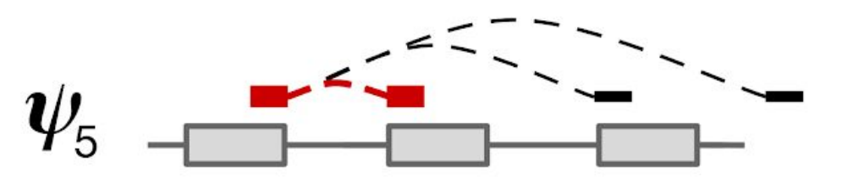
Significance based outlier detection

Visualization

[1]*GTEx Consortium, The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans, Science, 2015*
[2]*Kremer et al., Genetic diagnosis of Mendelian disorders via RNA sequencing, Nat. Commun., 2017*

Intron-centric splicing metrics $\psi_{5/3}$ and $\theta_{5/3}$

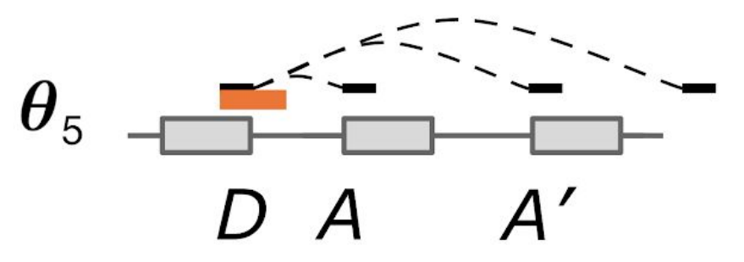
ψ (Percent Spliced-In):



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

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

θ (Splicing Efficiency):



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

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

Advantage of $\psi_{5/3}$ and $\theta_{5/3}$:
➔ quantifiable from RNA-seq data without prior exon annotations

Pervouchine et al, Bioinformatics, 2013

Statistical model of FRASER

The metrics ψ_5, ψ_3 , and θ are count proportions. We model the distribution of the numerator conditioned on the value of the denominator using the **beta-binomial distribution** with an intron-specific intra-class correlation parameter ρ_j and a sample- and intron-specific proportion expectation μ_{ij}:

$$P(k_{ij}) = BB(k_{ij} | n_{ij}, \mu_{ij}, \rho_j),$$

The **proportion expectation μ_{ij}** is jointly modeled using a latent space that captures covariations between samples:

$$\mu_{ij} = \sigma(y_{ij}) = \frac{exp(y_{ij})}{1 + exp(y_{ij})},$$

$$\mathbf{y}_i = \mathbf{h}_i \mathbf{W}_d + \mathbf{b},$$

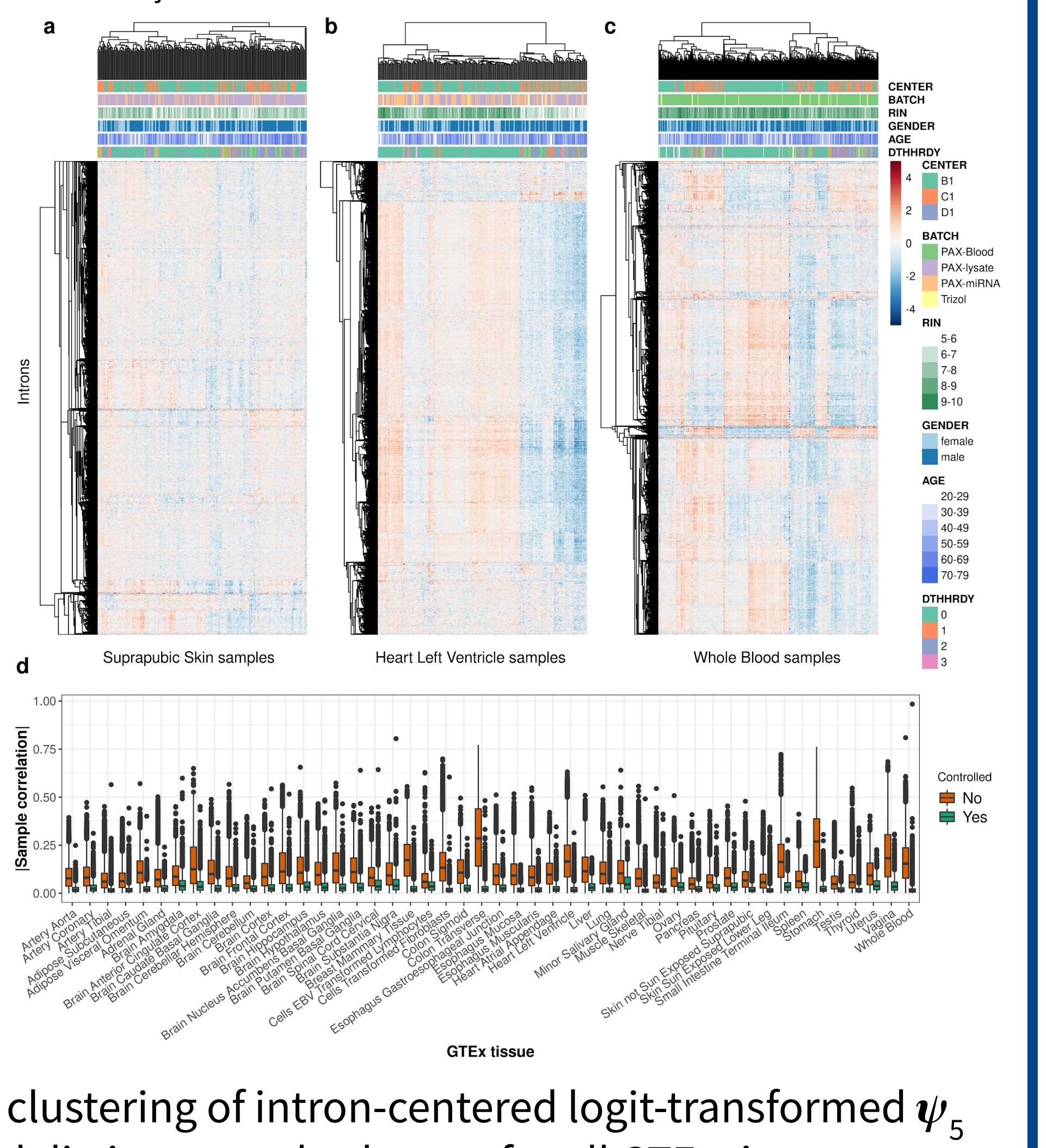
$$\mathbf{h}_i = \tilde{\mathbf{x}}_i \mathbf{W}_e,$$

The input vector $\tilde{x}_i$ is given by the centered and logit-transformed pseudo-count ratios:

$$\tilde{x}_{ij} = x_{ij} - \bar{x}_j,$$

$$x_{ij} = \text{logit}\left(\frac{k_{ij}+1}{n_{ij}+2}\right),$$

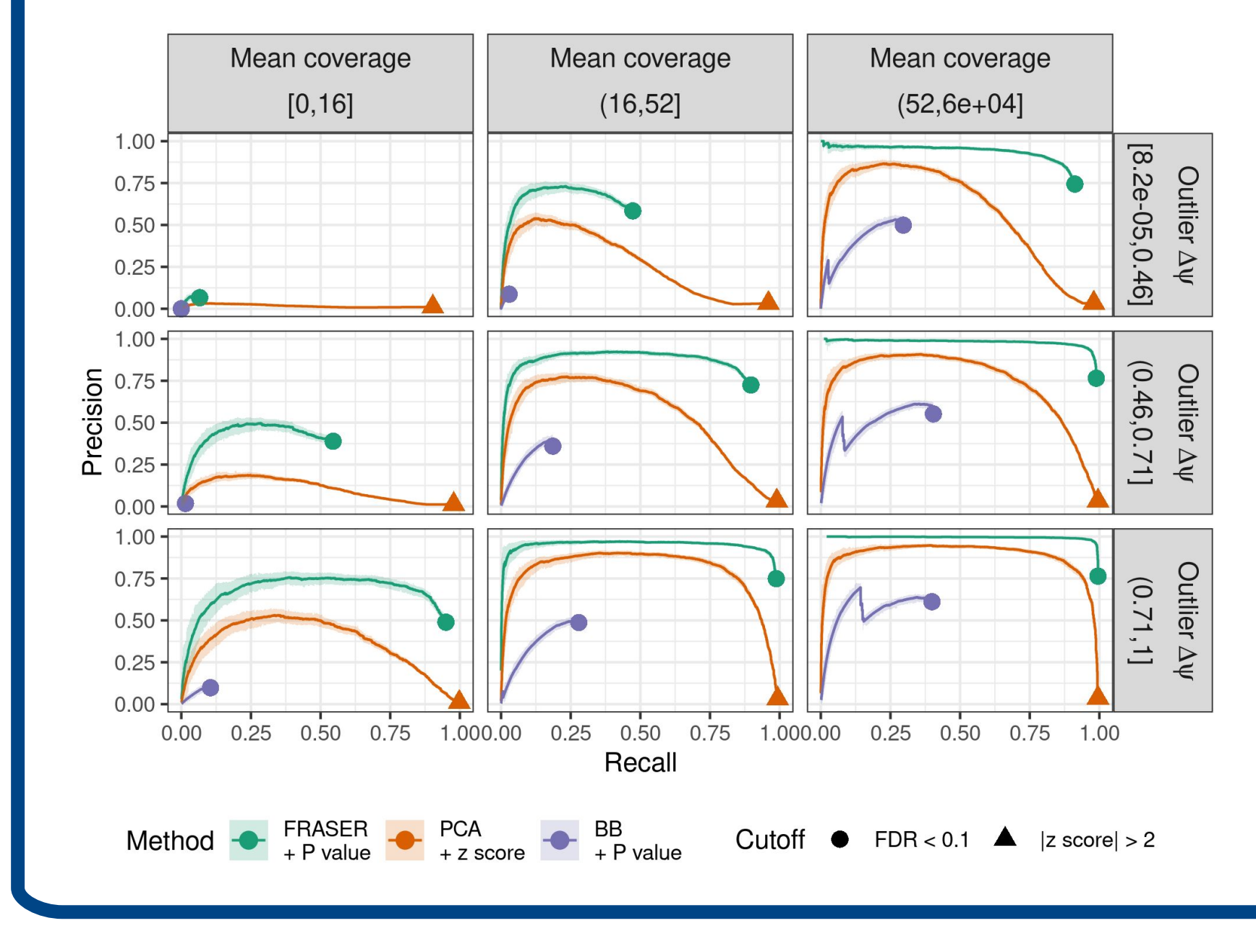
$$\text{logit}(a) = \log \frac{a}{1-a}.$$



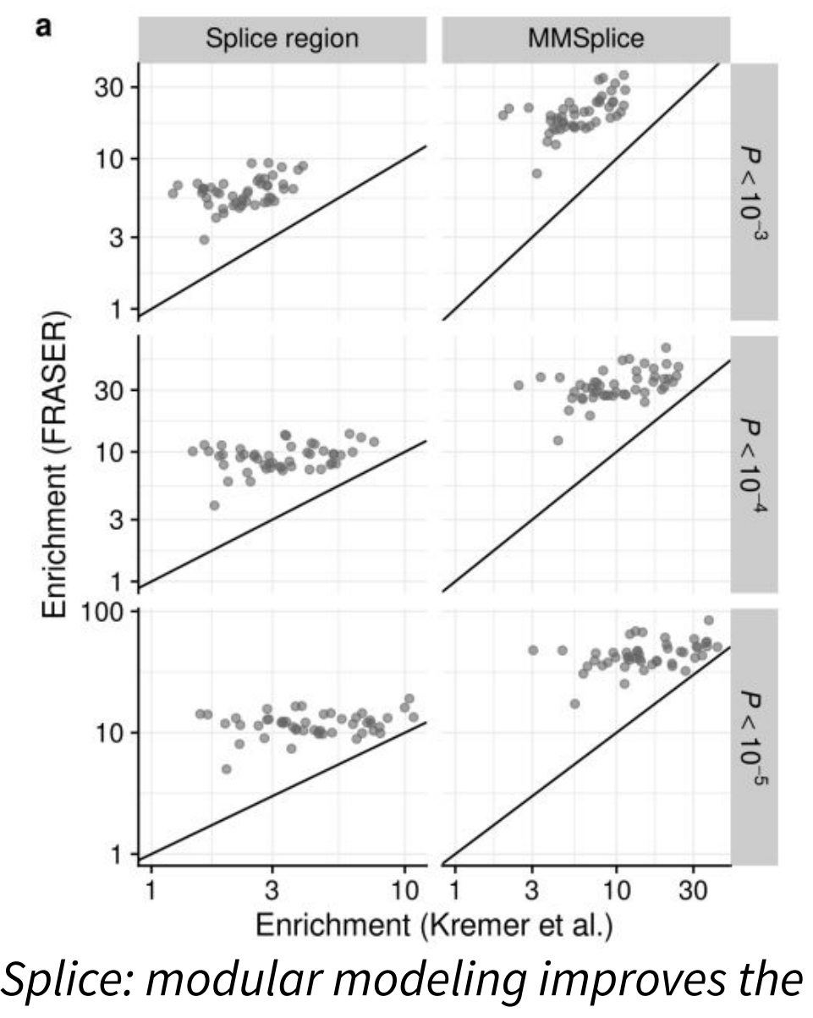
→ Hierarchical clustering of intron-centered logit-transformed ψ_5 values revealed distinct sample clusters for all GTEx tissues.
→ Controlling for the latent space reduced the between-sample correlation from 0.10 +/- 0.05 down to 0.02 +/- 0.01 (mean +/- standard deviation across tissues).

Benchmark: precision-recall and enrichment

Precision and recall of artificially injected outliers on the not sun exposed skin samples from GTEx[1] for three methods: 1) FRASER, 2) PCA + z-scores and 3) beta-binomial P values without control for confounders



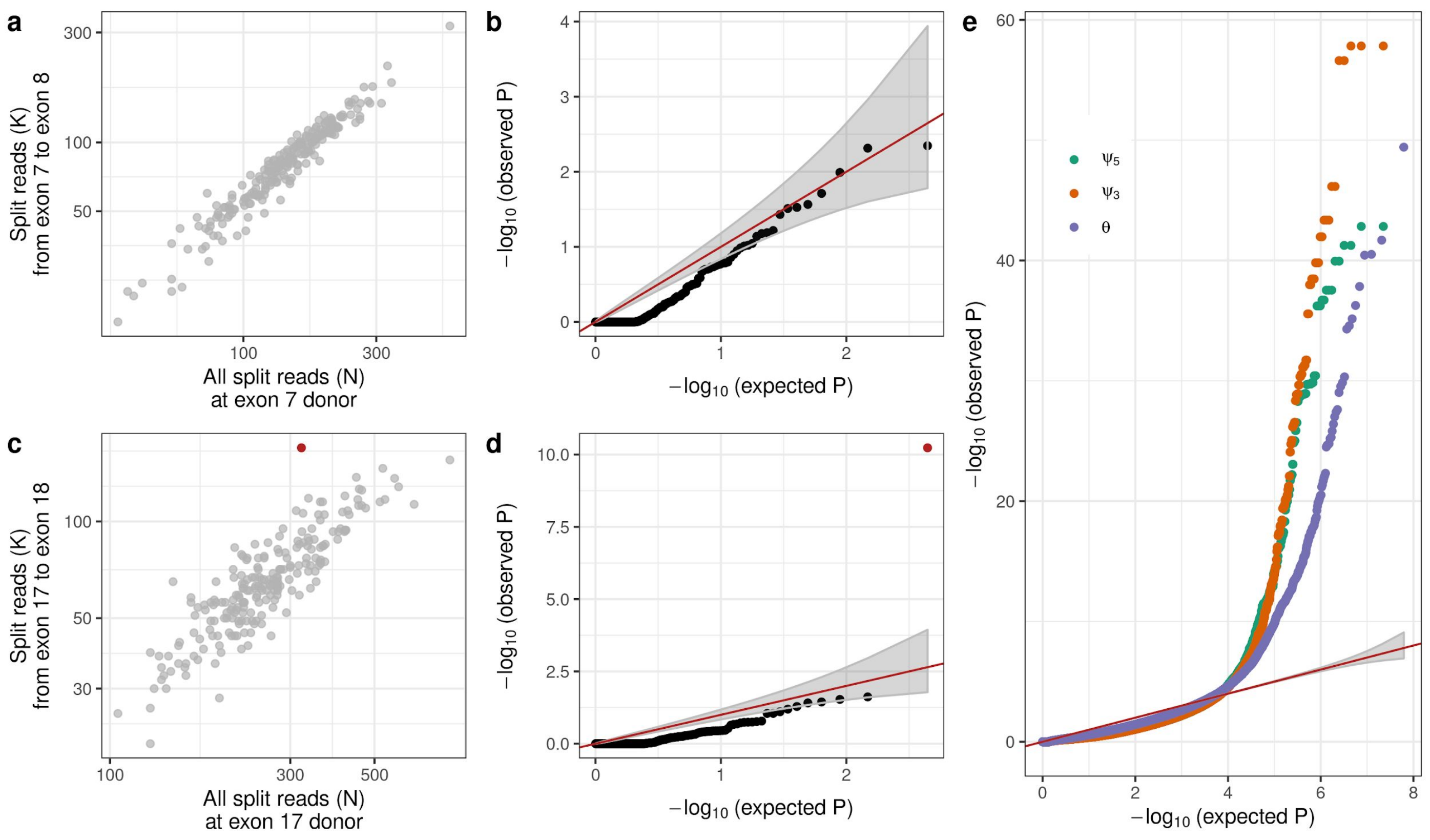
Enrichment using FRASER against enrichment using the method from Kremer et al.[2] for rare variants located in the splice region and for rare variants predicted to affect splicing by MMSplice[3] for all GTEx[1] tissues:



[3]*Cheng et al., MMSplice: modular modeling improves the predictions of genetic variant effects on splicing, Genome Biol., 2019*

Detection of aberrant splicing events

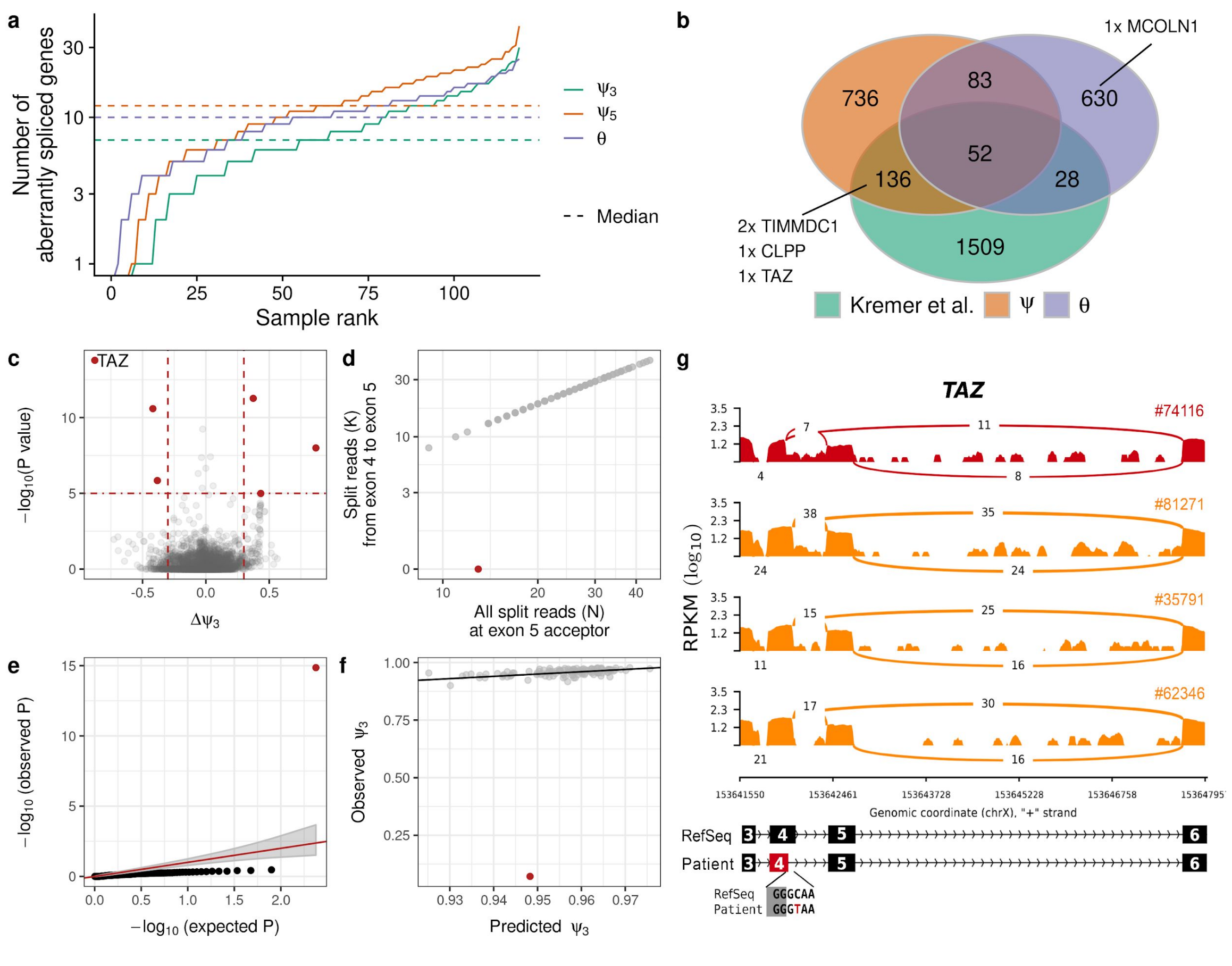
We considered the observations that significantly deviate from their expected value as outliers and computed two-sided beta-binomial p-values for each observation. P-values from introns with the same donor (or acceptor) are corrected using Holm's method. Additionally, we controlled the p-values of all introns per sample using Benjamini-Yekutieli's method.
a-b: example of a junction with no outlier
c-d: example of a junction with one outlier
e: Q-Q plot of all junctions on the Kremer dataset



Application to a rare disease cohort

Application of FRASER to the Kremer dataset:

- **a:** number of aberrant splicing events per sample for ψ_5, ψ_3, θ, at FDR < 0.1 and $\Delta\psi$ >= 0.3.
- **c-g:** aberrant alternative donor usage in the gene *TAZ* for individual 74116: $\Delta\psi_3 = -0.88$ and FDR = 1.98 x 10^{-9}:
  - → Newly created donor site located 22 bp inside the 4th exon resulted in nearly complete loss of the canonical donor site usage of the 4th exon
  - → Ablation of 8 amino acids of the protein encoded by *TAZ*, Tafazzin (catalyzes maturation of cardiolipin, a major lipid constituent of the inner mitochondrial membrane)



→ rare homozygous synonymous variant (c.348C>T) creates the new upstream donor site by introducing a GT dinucleotide.

FRASER package

FRASER is implemented as an R/Bioconductor package
Available at: https://bioconductor.org/packages/FRASER

A preprint of the paper is available on bioRxiv: https://doi.org/10.1101/2019.12.18.866830

Together with OUTRIDER[4] (a method for detecting gene expression outliers), FRASER is part of the computational workflow DROP[5] available at: https://github.com/gagneurlab/drop.git

[4]*Brechtmann et al., OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data, AJHG, 2018*
[5]*Yépez et al., Detection of aberrant events in RNA sequencing data, Protocol exchange, 2020*