SplicER: A novel analytic scheme for the analysis of Splicing Efficiency in RNA-seq data

L. Simon[1], T. Hsu[2], N. Neill[2], J. Wang[2], E. Chen[2], T. Westbrook[2] & C. Shaw[2]

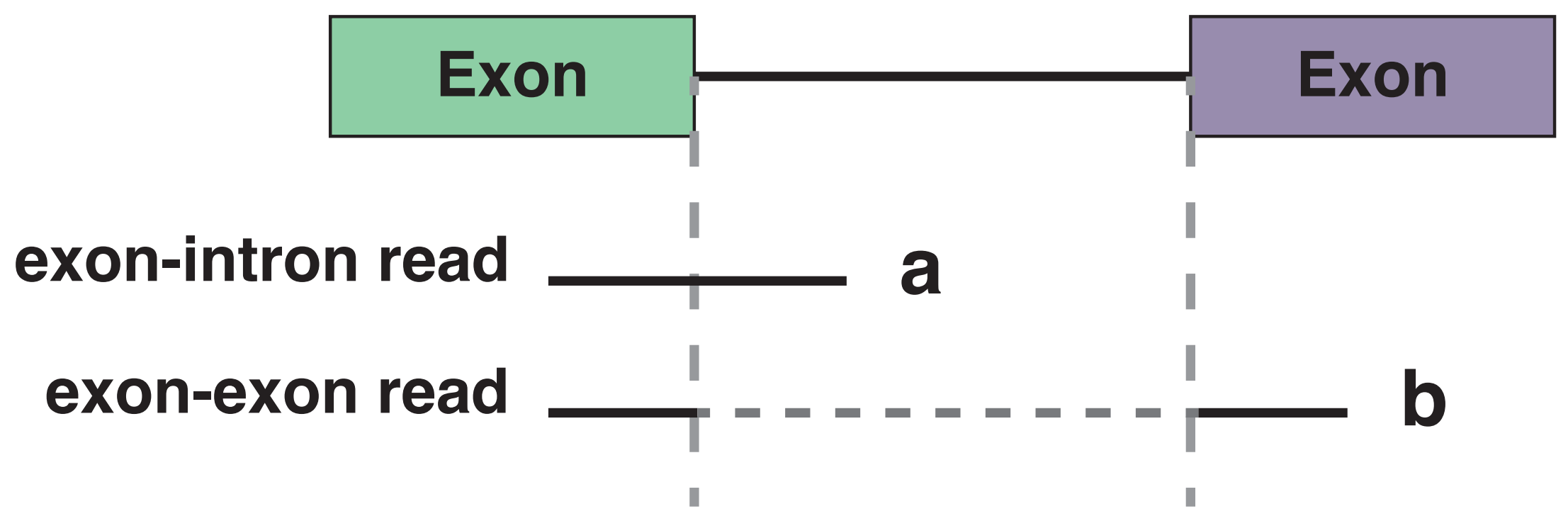
1 Structural and Computational Biology and Molecular Biophysics | 2 Molecular and Human Genetics | Baylor College ofMedicine | Houston, TX

Introduction

The advent of RNA-seq technology has created the opportunity to study transcriptomics at an unprecedented level. The nature of this technology allows researchers to make inquiry into diverse properties of genome-wide transcription. Recent studies have suggested that global splicing efficiency is a crucial transcriptional property that impacts cellular phenotypes and disease such as cancer. We define as a measure of splicing efficiency the proportion of purely exonic to mixed exon-intron reads among those observed at exon-intron boundaries. Spliced alignments of RNA-seq read-level data allow direct calculation of this measure of splicing efficiency. To implement and explore this measure of splicing efficiency, we have developed a new computational Python package we call SplicER to analyze this phenomena. To validate the performance of the package and our measure of splicing efficiency, we used data from the Gene Expression Omnibus on known splicing mutants, and we were able to recapitulate known differences. To extend the analyses with our package, we applied it to data from the Geuvadis project. We observed that there is great inter-individual variation in our measure of splicing efficiency, and that this variation correlates with the expression of spliceosomal machinery.

Methods

For each junction, reads spanning the corresponding genomic position are extracted using the pysam module. Next, each read is classified as exon-exon (**a**) or exon-intron (**b**). Classification is based on the CIGAR string of each read. If the read maps into or skips the first base of the intron it is classified as **a** or **b**, respectively. Total counts of **a** and **b** are recorded for each junction for determination of the splicing efficiency measure. Aggregation of this junction-level measure at the level of genes and individual samples permits higher scale summary of splicing efficiency across genes within a sample and among samples considering the ensemble of all junctions.

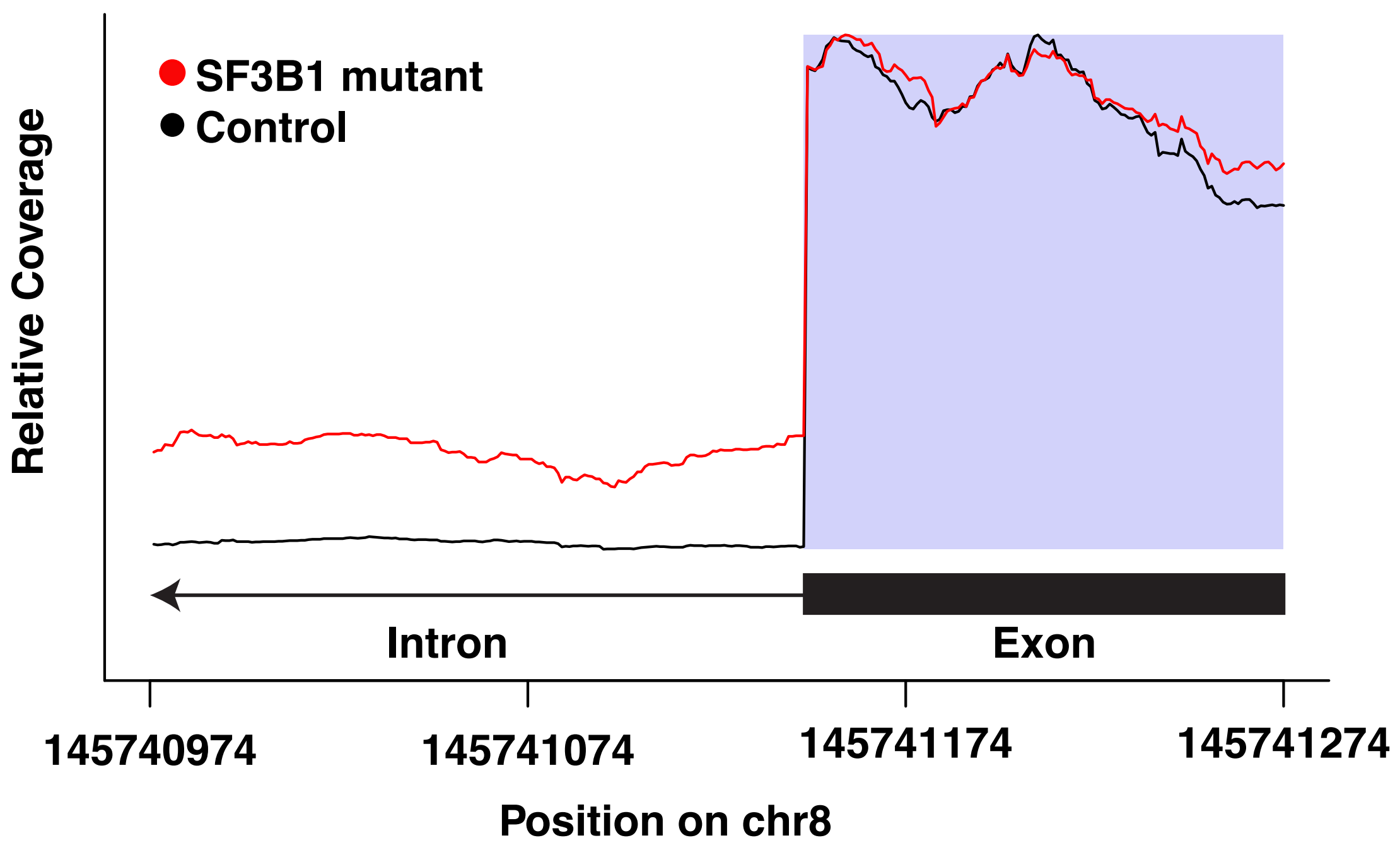


Results

We searched publicly available RNA-seq data sets for "splicing efficiency" and identified the following experiments:

Knockdown Gene	First Author	Year	ArrayExpress ID
SF3B1	Kfir et al.	2015	E-GEOD-65644
SART1	Oka et al.	2014	E-GEOD-59376
UBL5	Oka et al.	2014	E-GEOD-59376

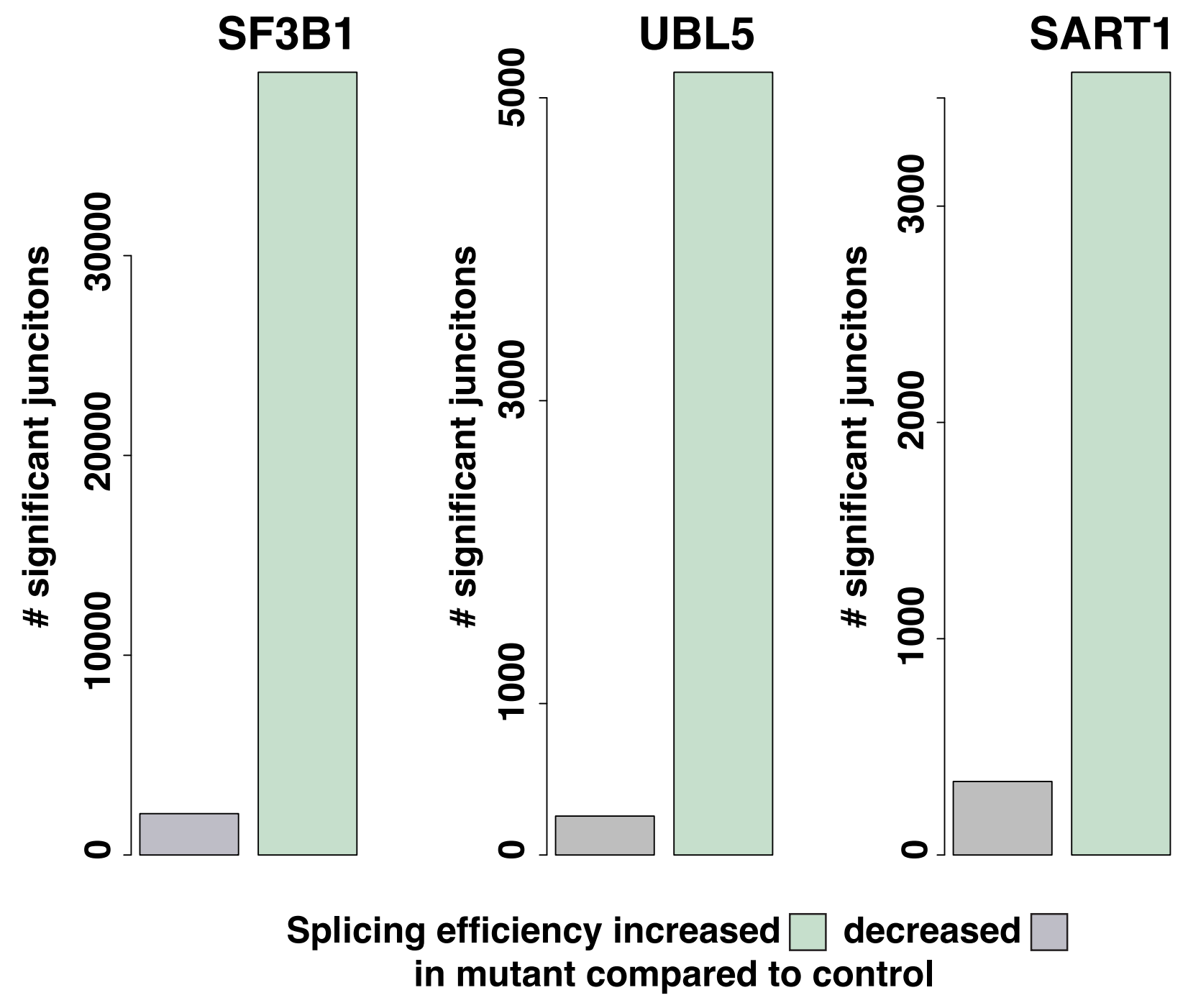
We applied our SplicER method to each of these datasets. To account for alternative splicing, we used a set of ~170k constitutive junctions (present in each isoform of a given gene). To statistically evaluate differences between mutants and controls at a junction-level we perfomed proportion tests. The plot on the left shows an examplary junction with a significantly higher proportion of exon-intron reads in the SF3B1 mutant compared to the control.



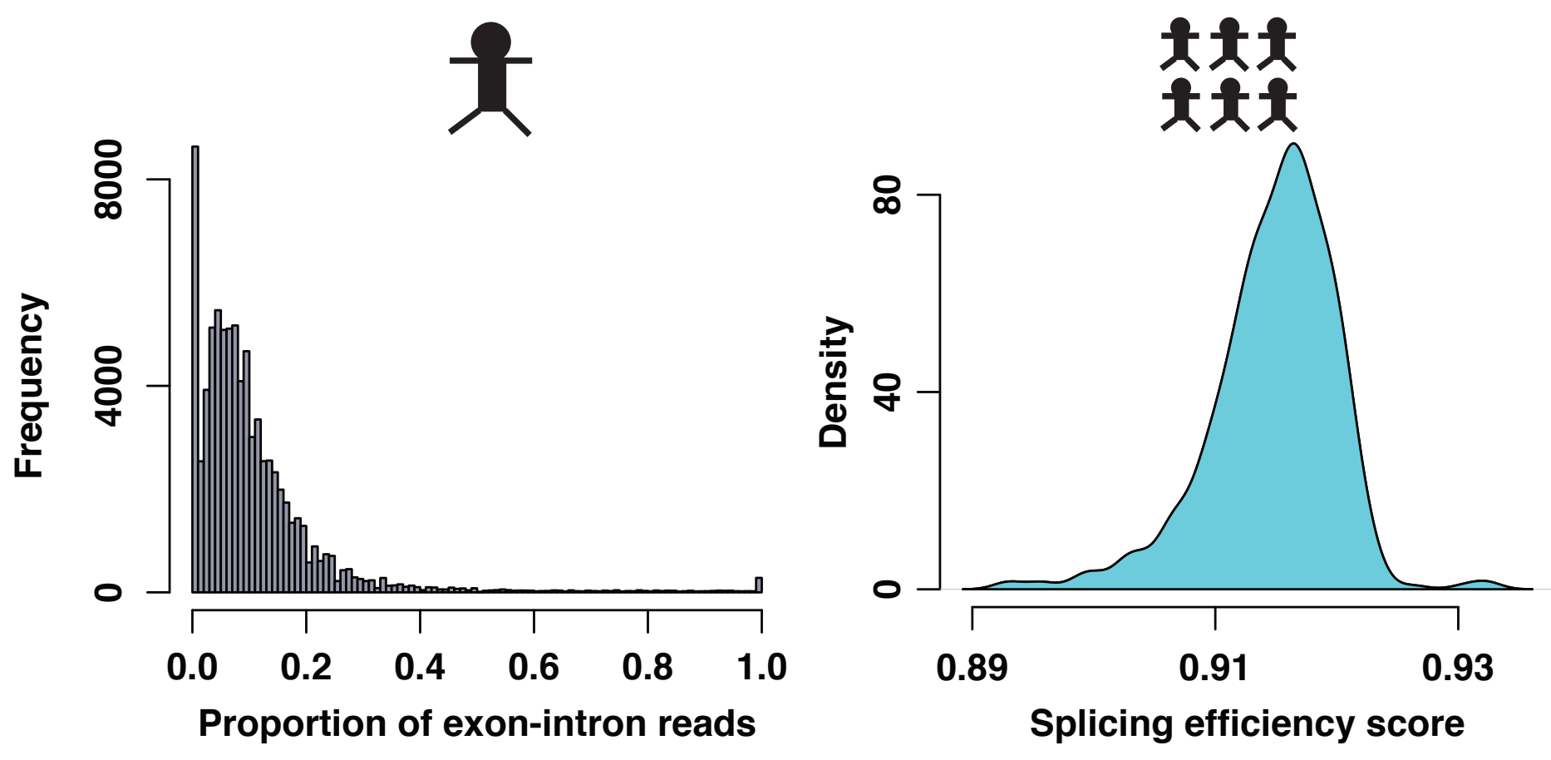
RNA-seq coverage across a junction in gene *RECGL4*. Y-axis shows read coverage relative to treatment. X-axis shows genomic coordinates. The SF3B1 mutant and control are colored in red and black, respectively.

Results

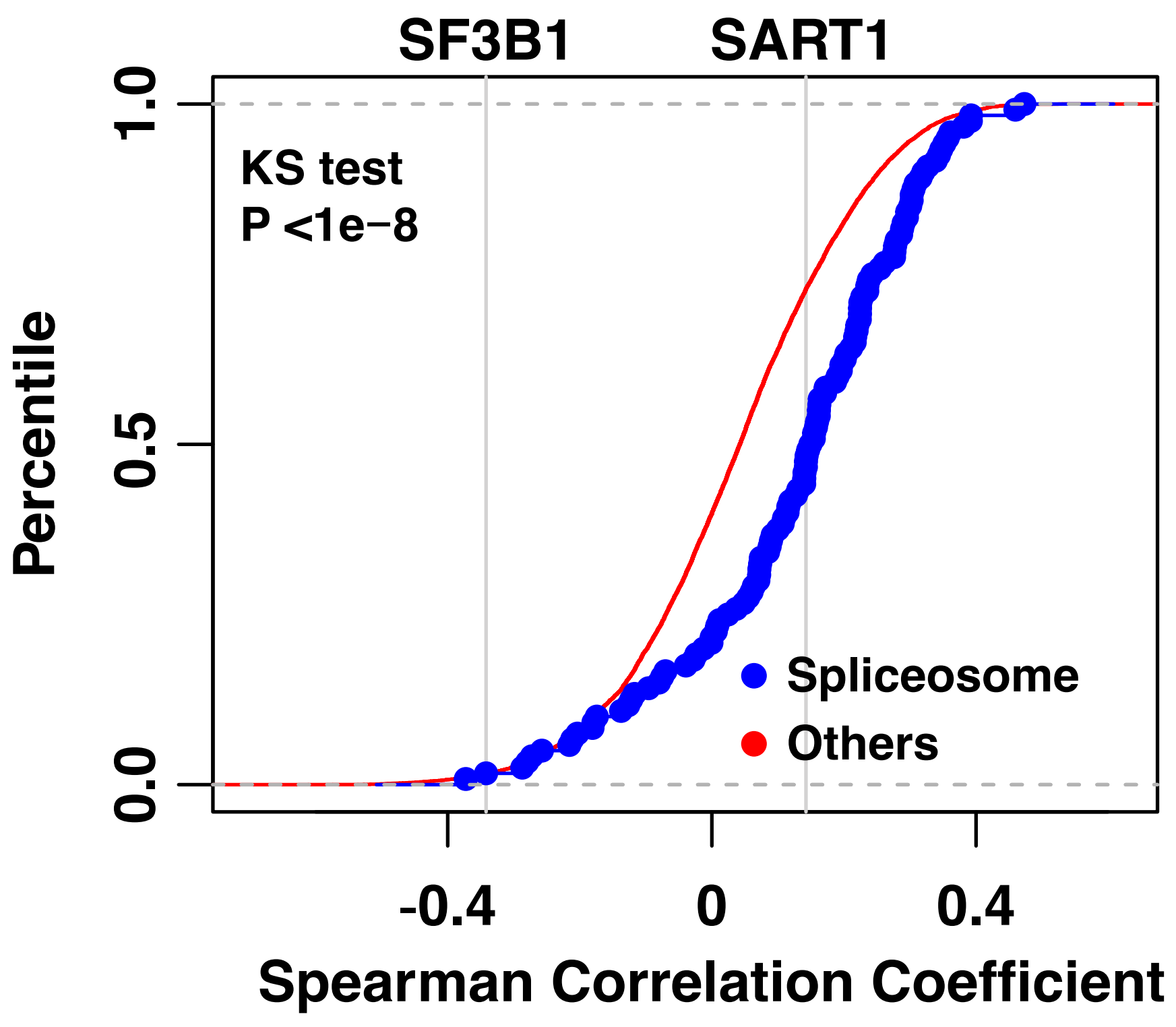
The majority of junctions with a marginal effect (P<0.05) contained more exon-intron reads in the mutant compared to controls. This demonstrates a decrease of splicing efficiency in the mutant relative to control, validating our splicing metric is able to detect deficiency in splicing efficiency when we expect that they will exist. Next, we applied SpliceER to the Geuvadis dataset. This project contains RNA-seq data from the 1000 Genomes samples. To summarize splicing efficiency across all junctions within a single sample, we define splicing efficiency as the trimmed mean of the proportion of exon-intron reads across all junctions containing at least 10 reads. The set of genes annotated to "Spliceosome" was significantly enriched among genes highly correlated with splicing efficiency. EFTUD2 was the spliceosomal component with the strongest association with splicing efficiency. To test for association between splicing efficiency and genetic variation we performed a genome-wide quantitative trait analysis. No association passing genome-wide significance (P<1e-6) was found.



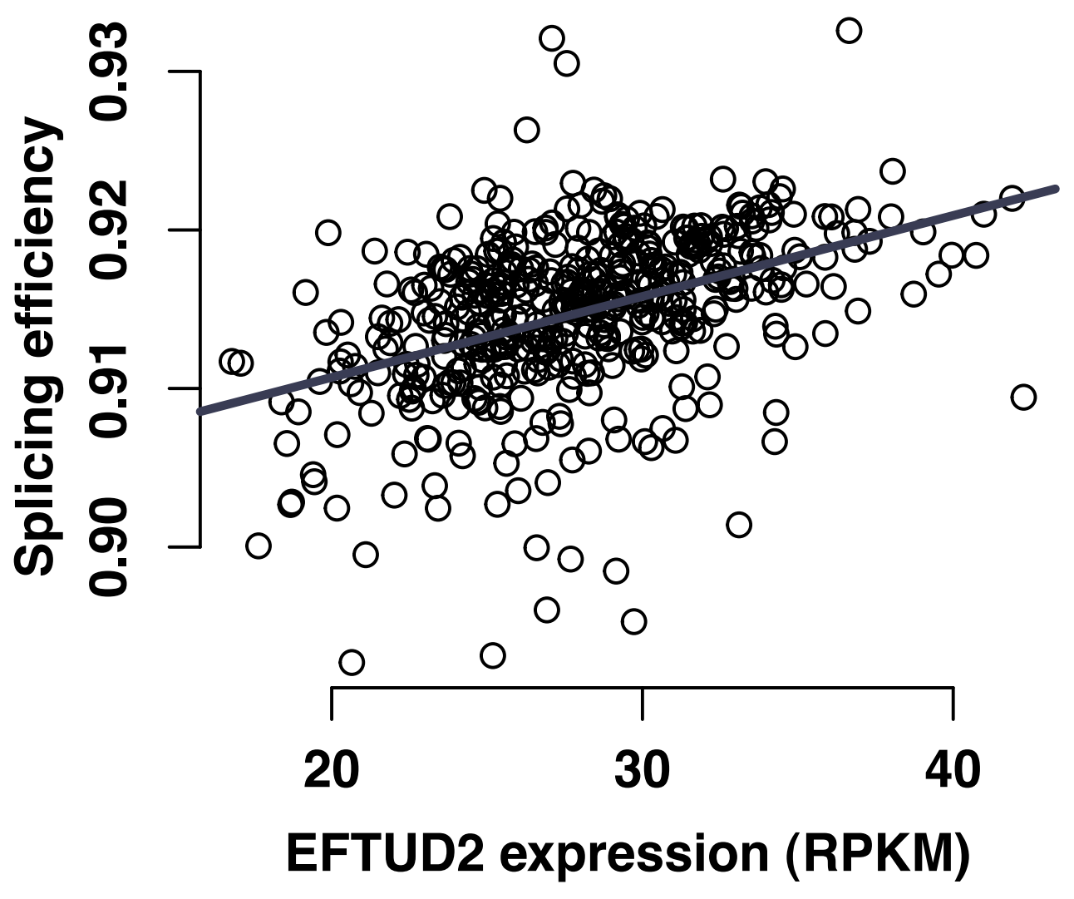
Barplots show the number of junctions with a marginally significant effect (P<0.05). Most junctions contained a larger proportion of exon-intron reads in the mutant compared to control.



Left) Distribution of proportion of exon-intron reads across junctions with more than 10 reads within a single sample. Right) Distribution of splicing efficiency scores across all Geuvadis samples.



Empirical cumulative distribution of Spearman correlation coefficients. Genes annotated to the "Spliceosome" geneset are colored in blue. Shift of the blue points indicates enrichment of spliceosomal genes.



Scatterplot of EFTUD2 expression levels (X-axis) and splicing efficiency (Y-axis).

Performance

Our SplicER algorithm is fast and extremely memory efficient. Analysis of ~15Mio junction reads in a 5GB bamfile takes about 20min on a single thread using <100Mb of memory.

Discussion

Our novel method allows researchers to quantify splicing efficiency from RNA-seq alignments and enables large scale investigation of splicing quality at the junction and sample levels as a transcriptional phenotype. Code is freely available from: **github.com/lkmklsmn/splicer.**

Acknowledgements

I would like to thank Han Liang, Joel Neilson, John Belmont, Neil Hanchard and Chad Shaw for their helpful mentorship. I would like to thank all the members of the Shaw and Westbrook laboratories for their support.