nature genetics

TECHNICAL REPORT

https://doi.org/10.1038/s41588-017-0004-9

Annotation-free quantification of RNA splicing using LeafCutter

Yang I. Li[1,9*], David A. Knowles[1,2,3*], Jack Humphrey[4,5], Alvaro N. Barbeira[6], Scott P. Dickinson[6], Hae Kyung Im[6] and Jonathan K. Pritchard[1,7,8*]

The excision of introns from pre-mRNA is an essential step in mRNA processing. We developed LeafCutter to study sample and population variation in intron splicing. LeafCutter identifies variable splicing events from short-read RNA-seq data and finds events of high complexity. Our approach obviates the need for transcript annotations and circumvents the challenges in estimating relative isoform or exon usage in complex splicing events. LeafCutter can be used both to detect differential splicing between sample groups and to map splicing quantitative trait loci (sQTLs). Compared with contemporary methods, our approach identified 1.4–2.1 times more sQTLs, many of which helped us ascribe molecular effects to disease-associated variants. Transcriptome-wide associations between LeafCutter intron quantifications and 40 complex traits increased the number of associated disease genes at a 5% false discovery rate by an average of 2.1-fold compared with that detected through the use of gene expression levels alone. LeafCutter is fast, scalable, easy to use, and available online.

The alternative removal of introns during mRNA maturation is essential for major biological processes in eukaryotes such as cellular differentiation, response to environmental stress, and proper gene regulation[1–4]. However, researchers' ability to obtain novel insights into the regulation and function of splicing is hindered by the difficulty of estimating transcript abundances from short-read RNA-seq data.

Popular approaches for studying alternative splicing from RNA-seq estimate isoform ratios[5–8] or exon inclusion levels[9,10]. Quantification of isoforms or exons is intuitive because RNA-seq reads generally represent mature mRNA molecules from which introns have already been removed. However, the estimation of isoform abundance from conventional short-read data is statistically challenging, as each read samples only a small part of the transcript, and alternative transcripts often have substantial overlap[11]. Similarly, in estimations of exon expression levels, read depths are often overdispersed because of technical effects, and there may be ambiguity about which version of an exon is supported by a read if there are alternative 5′ or 3′ splice sites.

Further, both isoform-quantification and exon-quantification approaches rely on transcript models or predefined splicing events, both of which may be inaccurate or incomplete[12]. Predefined transcript models are particularly limiting when the splicing profiles being compared are from healthy versus disease samples, as aberrant transcripts may be disease specific, or in studies of genetic variants that generate splicing events in only a subset of individuals[13]. Even when transcript models are complete, it is difficult to estimate isoform or exon usage of complex alternative splicing events[12].

An alternative approach is to focus on what is removed in each splicing event. Excised introns may be inferred directly from reads that span exon–exon junctions; thus, there is little ambiguity about the precise intron that has been cut out, and quantification of usage ratios is very accurate[12]. The recently published MAJIQ[12] method also proposes to estimate local splicing variation from split reads and identifies complex splicing events; however, it does not scale well with more than 30 samples and has not been adapted to map sQTLs. At present, there are several software programs for sQTL mapping: GLiMMPS[14], sQTLseekeR[15], and Altrans[16]. However, all three rely on existing isoform annotations, and analyses with both GLiMMPS and sQTLseeker reported modest numbers of sQTLs.

Here we describe LeafCutter, a suite of methods that allow the identification and quantification of novel and known alternative splicing events by focusing on intron excisions. We demonstrate LeafCutter's utility by applying it to three important problems: (1) the identification of differential splicing across conditions, (2) the identification of sQTLs in multiple tissues or cell types, and (3) the assignment of molecular effects to disease-associated genome-wide association study (GWAS) loci. Using an early version of LeafCutter, we found that alternative splicing is an important mechanism through which genetic variants contribute to disease risk[17]. We now show that LeafCutter dramatically increases the number of detectable associations between genetic variation and pre-mRNA splicing, thereby enhancing understanding of disease-associated loci.

Results

Overview of LeafCutter. LeafCutter uses short-read RNA-seq data to detect intron excision events at base-pair precision by analyzing mapped split reads (Fig. 1). LeafCutter focuses on alternative splicing events, including skipped exons, 5′ and 3′ alternative splice-site usage, and additional complex events that can be summarized by differences in intron excision[12] (Supplementary Fig. 1). LeafCutter's intron-centric view of splicing is based on the observation that mRNA splicing occurs predominantly through the step-wise removal of introns from nascent pre-mRNA[18]. (Unlike isoform-quantification

[1]Department of Genetics, Stanford University, Stanford, CA, USA. [2]Department of Computer Science, Stanford University, Stanford, CA, USA. [3]Department of Radiology, Stanford University, Stanford, CA, USA. [4]UCL Genetics Institute, Gower Street, London, UK. [5]Department of Neurodegenerative Disease, UCL Institute of Neurology, London, UK. [6]Section of Genetic Medicine, Department of Medicine, The University of Chicago, Chicago, IL, USA. [7]Department of Biology, Stanford University, Stanford, CA, USA. [8]Howard Hughes Medical Institute, Stanford University, Stanford, CA, USA. Present address: [9]Section of Genetic Medicine, Department of Medicine, The University of Chicago, Chicago, IL, USA. Yang I. Li and David A. Knowles contributed equally to this work. *e-mail: yangili1@uchicago.edu; dak33@stanford.edu; pritch@stanford.edu

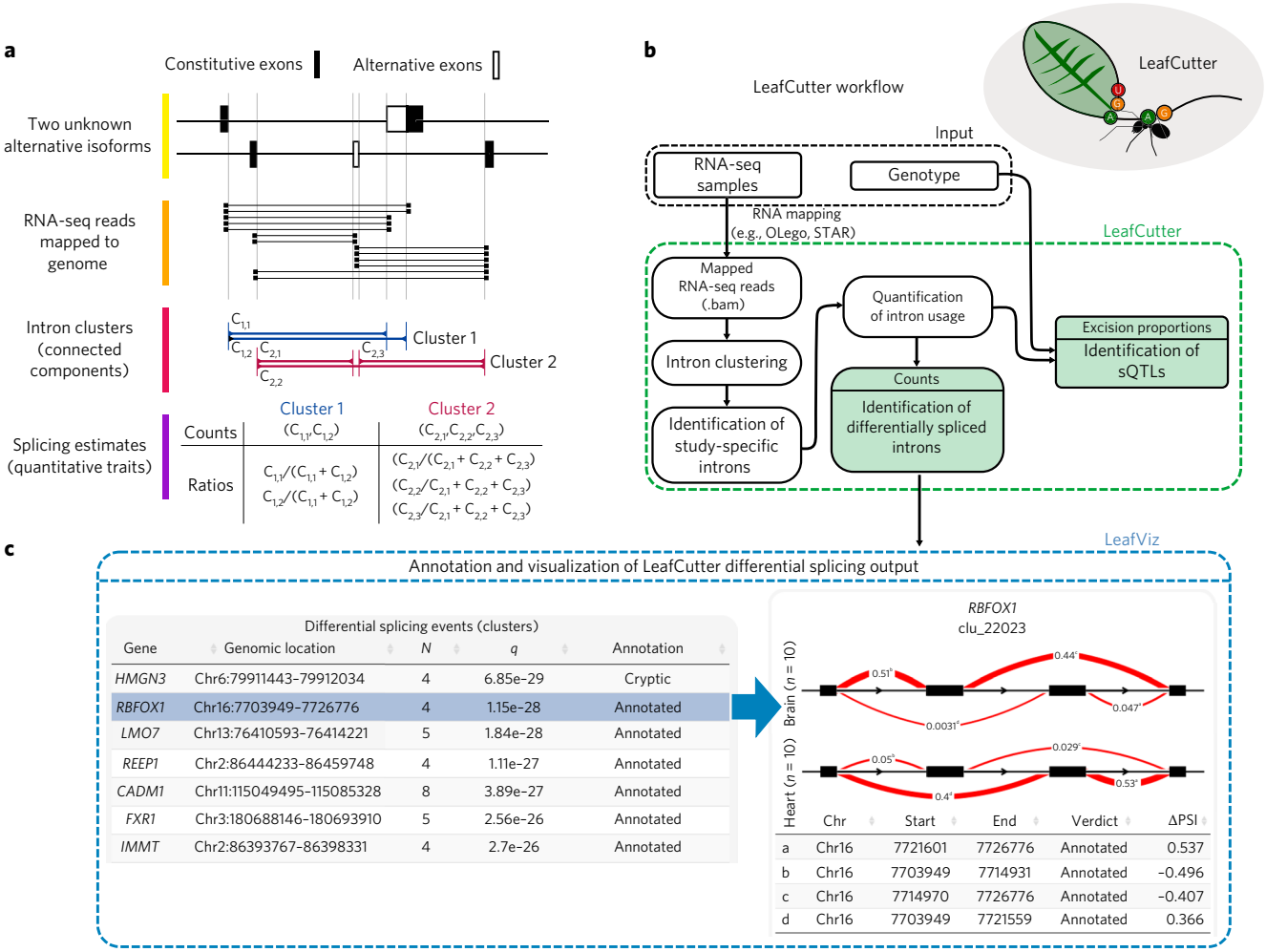

Fig. 1 | Overview of LeafCutter. a, LeafCutter uses split reads to uncover alternative intron-excision options by finding introns that share splice sites. In this example, LeafCutter identified two clusters of variably excised introns. **b**, The LeafCutter workflow. First, short reads are mapped to the genome. When SNP data are available, WASP[33] should be used to filter allele-specific reads that map with a bias. Next, LeafCutter extracts junction reads from.bam files, identifies alternatively excised intron clusters, and summarizes intron usage as counts or proportions. Finally, LeafCutter identifies intron clusters with differentially excised introns between two user-defined groups by using a Dirichlet-multinomial model, or maps genetic variants associated with intron excision levels by using a linear model. **c**, Visualization of differential splicing among ten GTEx heart and brain samples by LeafViz. LeafViz is an interactive browser-based application that allows users to visualize results from LeafCutter differential splicing analyses. In this example, we observed that *RBFOX1* showed differential usage of a mutually exclusive exon in heart compared with the usage in brain. Differential splicing is measured in terms of the change in the percent spliced in (ΔPSI). For all examples, see "URLs."

methods such as Cufflinks2[5], LeafCutter does not measure alternative transcription start sites and alternative polyadenylation directly, as they are not generally captured by intron excision events.) The major advantage of this representation is that LeafCutter does not require read assembly or inference of isoforms supported by ambiguous reads, both of which are computationally and statistically difficult. As a result, we were able to improve speed and memory requirements by an order of magnitude or more compared with those of similar methods such as MAJIQ[12].

To identify alternatively excised introns, LeafCutter pools all mapped reads from a study and finds overlapping introns demarcated by split reads. LeafCutter then constructs a graph that connects all overlapping introns that share a donor or acceptor splice site. The connected components of the graph form clusters, which represent alternative intron excision events. Finally, LeafCutter iteratively applies a filtering step to remove rarely used introns, which are defined on the basis of the proportion of reads supporting a given intron compared with other introns in the same cluster, and re-clusters leftover introns (Online Methods and Supplementary Note 1). In practice, we have found that it is important to apply this filtering step to avoid arbitrarily large clusters at read depths where noisy splicing events are supported by multiple reads.

De novo identification of RNA splicing in mammalian organs. We tested LeafCutter's novel intron-detection method by analyzing mapped RNA-seq[19] data from 2,192 samples (Supplementary Note 1) across 14 tissues from the Genotype-Tissue Expression (GTEx) Consortium[20]. We then searched for introns that were predicted to be alternatively excised by LeafCutter but were missing in three commonly used annotation databases (GENCODE v19, Ensembl, and UCSC). For this analysis, we ensured that the identified introns were indeed alternatively excised by considering only introns that were excised at least 20% of the time compared with other overlapping introns, in at least one-fourth of the samples, analyzing each tissue separately. We found that between 10.8% and 19.3% (pancreas and spleen, respectively) of alternatively spliced introns were unannotated, excluding testis, the major outlier, in which 48.5% of alternatively spliced introns were previously unidentified (Fig. 2a).

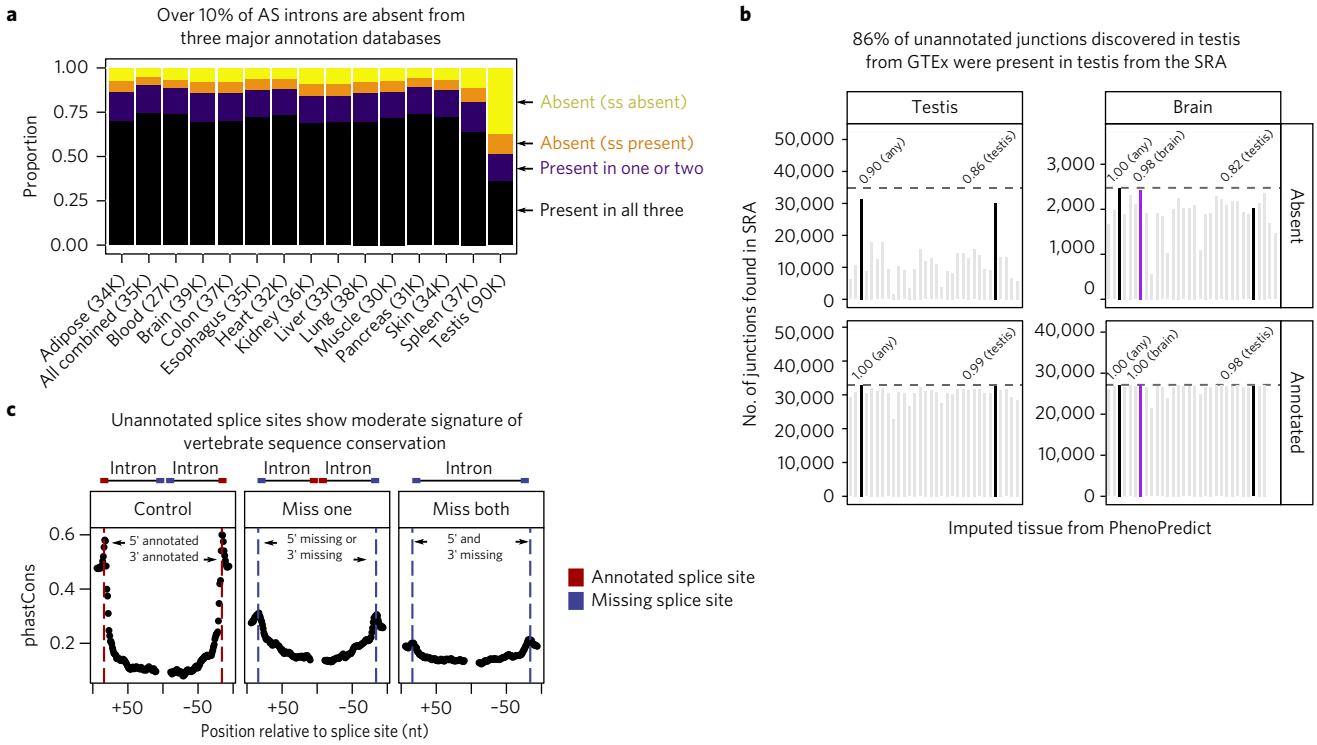

Fig. 2 | LeafCutter discovers reproducible unannotated introns. a, When we used LeafCutter to discover novel introns in 2,192 samples from the GTEx Consortium, we found that for any given tissue, >10% of alternatively spliced (AS) introns were unannotated; 48.5% of testis alternatively excised introns were unannotated. The color-coding denotes the proportion of introns for which one or more splice sites were unannotated (ss absent), both splice sites were annotated but the intron was not part of any transcript (ss present), or the intron was annotated in some but not all databases. **b**, The numbers of unannotated and annotated junctions discovered with LeafCutter that were also found in samples from the Sequence Read Archive (SRA) with Intropolis[23]. We used PhenoPredict[34] to predict the tissue type corresponding to the SRA samples analyzed in Intropolis. **c**, The unannotated splice sites of novel introns show a moderate signature of sequence conservation as determined by vertebrate phastCons scores. Miss one: conservation of the unannotated splice site of an intron for which the cognate splice site is annotated. Miss both: conservation of splice sites of introns with both splice sites unannotated.

This observation is compatible with the 'out-of-testis' hypothesis, which proposes that transcription is more permissive in testis and allows novel genes or isoforms to be selected for if they are beneficial[21,22]. Thus 31.5% of the alternatively excised introns we detected were unannotated (Supplementary Note 1), consistent with a recent study that identified a similar proportion of novel splicing events in 12 mouse tissues[12]. To further confirm that these findings were not merely mapping or GTEx-specific artifacts, we searched for junction reads in 21,504 human RNA-seq samples from the Sequence Read Archive obtained from Intropolis[23]. We found that most (86%; Fig. 2b and Supplementary Fig. 2) novel junctions identified in our study were also present in at least one RNA-seq sample from the corresponding tissue as identified in Intropolis. Furthermore, we found that, as expected, unannotated junctions tended to be tissue specific, and often involved complex splicing patterns (Supplementary Fig. 3 and Supplementary Note 1).

We next asked whether these novel introns showed evidence of functionality as determined by sequence conservation. When we averaged phastCons scores over unannotated splice sites of introns that were absent from annotation databases, we found a moderate, but significant, signature of sequence conservation (Fig. 2c). In particular, we found that a substantial number (4,616, or 15–25%) of novel splice sites are conserved across vertebrates (average phastCons ≥ 0.6; Supplementary Fig. 4), indicating that the alternative excision of thousands of introns may be functional (Supplementary Note 1).

Fast and robust identification of differential splicing. LeafCutter uses counts from the clustering step to identify introns with differential splicing between user-defined groups. Read counts in an intron cluster are jointly modeled with a Dirichlet-multinomial generalized linear model, which we found offered superior sensitivity compared with a beta-binomial generalized linear model that tests each intron independently (Supplementary Fig. 5). The implicit normalization of the multinomial likelihood avoids the estimation of library size parameters required by methods such as DEXSEQ[10].

We compared LeafCutter to other methods for differential splicing detection, including Cufflinks2[5], MAJIQ[12], and rMATS[24]. We note that comparisons between algorithms have the complication that there is typically no one-to-one mapping between the splicing events quantified by different methods. We discuss this issue and our solution in Supplementary Note 1. For comparison, we used each method to identify splicing differences between 3, 5, 10, and 15 Yoruba (YRI) and European (CEU) lymphoblastoid cell line (LCL) RNA-seq samples. In terms of runtime, we observed a large difference in scalability (Fig. 3a). In our hands, only LeafCutter completed all comparisons within 1 h, whereas Cufflinks2, rMATS, and MAJIQ took as long as 7.8, 55.7, and 66.2 h, respectively, to complete the largest comparison. In terms of memory use, we also found that LeafCutter greatly outperformed the other software, using less than 400 MB of RAM for all comparisons, whereas MAJIQ required more than 50 GB to perform the larger comparisons (Supplementary Fig. 6). Although this range of sample sizes is representative for most biological studies, the identification of differential splicing across groups in large studies such as GTEx would be impractically slow with rMATS or MAJIQ.

To compare the abilities of different methods to detect differential splicing, we reasoned that the *P* values or posterior probabilities of the tests computed by each method are not directly

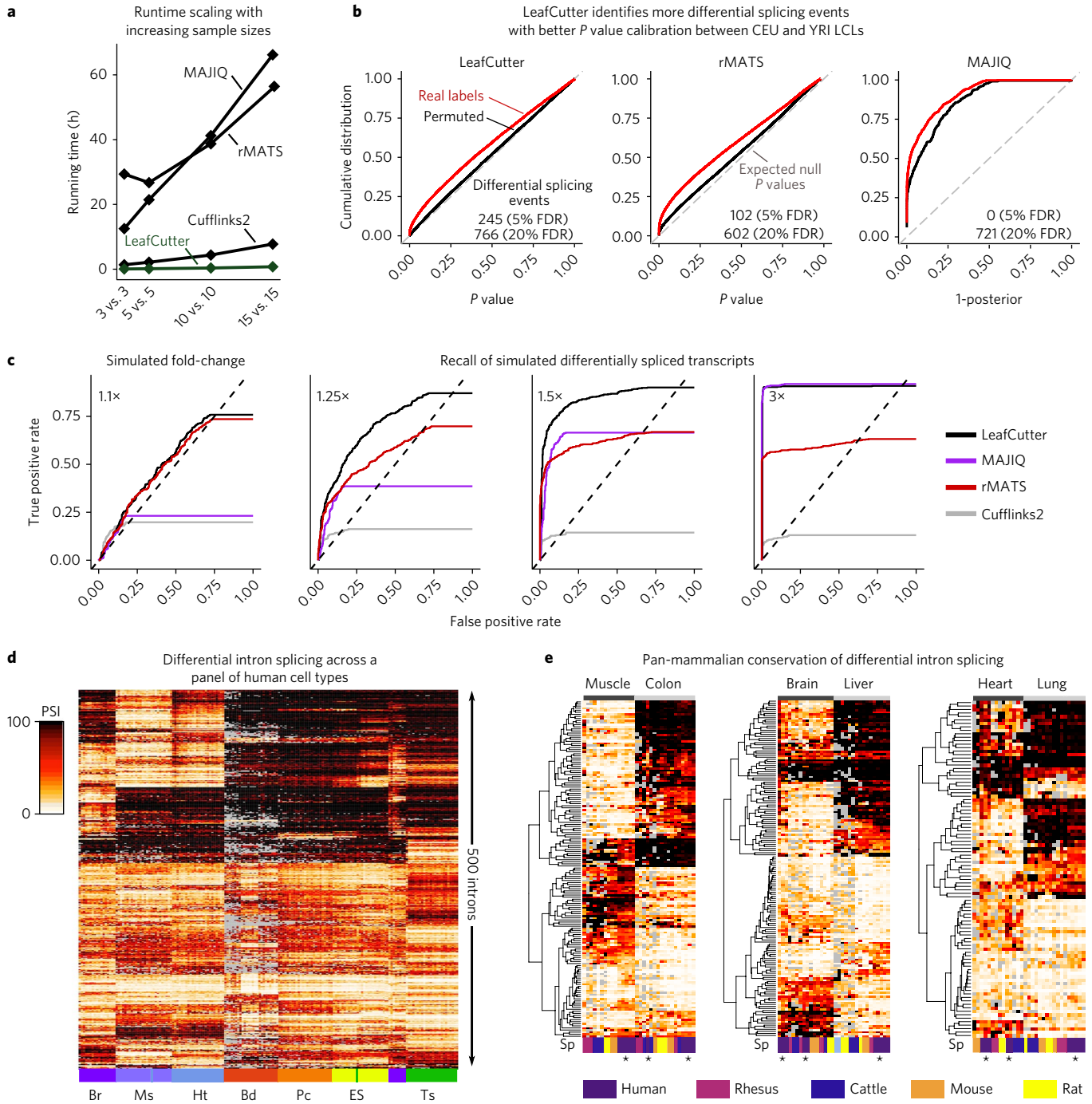

Fig. 3 | A comparison of methods for detecting differential splicing. a, The running time of differential splicing methods used to compare YRI and CEU LCL RNA-seq samples. The *x*-axis indicates the numbers of samples included. **b**, Cumulative distributions of differential-splicing-test *P* values (1-posterior for MAJIQ) for a comparison of 15 YRI and 15 CEU LCLs (red). The distribution of test *P* values for a comparison with permuted labels is also shown (black). Cufflinks2 (not shown) detected 0 significantly differentially spliced genes (Supplementary Fig. 8). **c**, Receiver operating characteristic (ROC) curves for LeafCutter, Cufflinks2, rMATS, and MAJIQ in an evaluation of differential splicing of genes with transcripts simulated to have varying levels of differential expression. ROC curves that do not reach true positive rates of 1.00 reflect genes simulated as differentially spliced that were not tested. **d**, LeafCutter identified tissue-regulated intron splicing events from GTEx organ samples. Shown is a heat map of the intron excision ratios of the top 500 introns that were found to be differentially spliced in at least one tissue pair. Tissues examined included brain (Br), muscle (Ms), heart (Ht), blood (Bd), pancreas (Pc), esophagus (Eg), and testis (Ts). PSI, percent spliced in. **e**, Heat maps showing intron exclusion ratios of introns differentially spliced between two tissues (muscle versus colon, brain versus liver, and heart versus lung). The heat maps show 100 random introns (97 for heart versus lung) that were predicted to be differentially excised in human samples with $P<10^{-10}$ (LR-test) and no more than five samples with missing data. A heat map of all introns that met our criteria can be found in Supplementary Fig. 11.

comparable. We therefore computed an empirical false discovery rate (FDR) from the *P* values of real comparisons between biologically distinct sample groups (YRI versus CEU samples here) and the *P* values of permuted comparisons between samples with permuted labels (i.e., two groups that both contained YRI and CEU samples). If the *P* values are well calibrated, the *P* value distribution of the

permuted comparisons is expected to be uniform. Indeed, we observed that the distributions of LeafCutter and rMATS *P* values for the permutations were close to uniform (Fig. 3b and Supplementary Fig. 7). However, the Cufflinks2 *P* values were overly conservative (Supplementary Note 1), and the posterior probabilities *P* reported by MAJIQ for the permuted comparisons did not track with the expected FDR of 1 – *P* (Supplementary Note 1). Overall, we found that LeafCutter *P* values showed better calibration compared with those from other methods, and that LeafCutter detected more differentially spliced events at all reasonable FDRs (≤0.2). Additionally, not only did LeafCutter detect more differentially spliced events at fixed FDRs, but it also achieved lower false negative rates when we evaluated the four methods with artificial data in which we simulated various levels of fold changes in isoform levels (Fig. 3c, Supplementary Note 1, Supplementary Fig. 8). These comparisons show that LeafCutter is a robust and highly scalable method for differential splicing analysis.

To evaluate LeafCutter's suitability to detect differential splicing in a biological setting, we searched for intron clusters that showed differential splicing between tissue pairs collected by the GTEx Consortium, using all tissues to identify intron clusters. When we combined all pairwise comparisons, we found 5,070 tissue-regulated splicing clusters at 10% FDR and with an estimated absolute effect size greater than 1.5 (Methods). As expected, GTEx samples grouped mostly by organ/tissue when hierarchically clustered according to the excision ratios of the 500 most differentially spliced introns among all tissue pairs (Fig. 3d and Supplementary Note 1).

To assess LeafCutter's applicability to studies with smaller sample sizes, we used a small subset of GTEx samples and then evaluated replication with a larger subset. When we used 220 samples (110 brain versus 110 muscle), we identified 1,906 differentially spliced clusters with estimated effect sizes greater than 1.5 at 10% FDR, compared with 885 when we used only 8 samples (4 brain versus 4 muscle). The strengths of association ($-\log_{10}$ *P* values) showed high correlation between our two analyses (Pearson $R^2 = 0.72$; Supplementary Note 1 and Supplementary Fig. 9), and 98% of alternatively spliced clusters identified at 10% FDR in the analysis with 8 total samples were replicated in the analysis with 220 samples, also at 10% FDR. These observations indicate that LeafCutter can detect differentially spliced introns even when the number of biological replicates is small.

We investigated whether the differentially spliced clusters identified with LeafCutter were likely to be functional by assessing the pan-mammalian conservation of their splicing patterns across multiple organs. Two previous studies analyzed the evolution of alternative splicing in mammals. When the researchers used gene expression levels, they saw clustering by organ as expected, but when they used exon-skipping levels, they instead saw clustering by species[25,26]. These observations indicate that a large number of alternative skipping events may lack function or undergo rapid turnover.

We initially clustered using all splicing events and found that the samples clustered mostly by species, thus confirming the previous findings[25,26] (Supplementary Fig. 10). We then focused on a subset of introns that LeafCutter identified as differentially excised across tissue pairs in human samples and found that this subset showed splicing patterns that were broadly conserved across mammalian organs (Fig. 3e and Supplementary Fig. 11). To do this, we hierarchically clustered samples from eight human organs and four organs from other mammals[25] according to the orthologous intron-excision proportions of differentially excised introns ($P < 10^{-10}$ and $\beta > 1.5$) from our pairwise analyses of human GTEx samples (Supplementary Note 1). Unlike the previous analyses, this showed striking clustering of the samples by organ, thus implying that hundreds of tissue-biased intron excisions events are conserved across mammals and are likely to have organ-specific functional roles[27]. Thus, although the majority of alternative splicing events probably undergo rapid turnover, events that show organ specificity are much more often conserved across mammals and therefore are more likely to be functionally important.

Mapping splicing QTLs with LeafCutter. To evaluate LeafCutter's ability to map sQTLs, we applied LeafCutter to 372 CEU LCL RNA-seq samples from GEUVADIS, and identified 42,716 clusters of alternatively excised introns. We used the proportion of reads supporting each alternatively excised intron identified by LeafCutter and a linear model[28] to map sQTLs (Supplementary Note 1). We found 5,774 sQTLs at 5% FDR (compared with 620 transcript ratio (tr)QTLs in the original study at 5% FDR, i.e., one-ninth as many) and 4,543 at 1% FDR. For a controlled comparison, we also processed 85 YRI LCL GEUVADIS RNA-seq samples and quantified RNA splicing events with LeafCutter, Altrans[16], and Cufflinks2[5]. We then uniformly standardized and normalized the estimates and used them as input to fastQTL[28] to identify sQTLs (Supplementary Note 1). At a similar FDR, LeafCutter identified 1.36–1.46 times and 1.83–2.06 times more sQTLs than Cufflinks2 and Altrans, respectively (Table 1). The rate of sQTL discoveries shared between methods was generally high (Storey's π_1 ranging from 0.53 to 0.72 for sQTLs identified at 10% FDR; Supplementary Note 1 and Supplementary Fig. 12), with LeafCutter sQTLs showing higher estimates of sharing ($\pi_1 = 0.70$ and 0.72 with Cufflinks2 and Altrans, respectively) than Cufflinks2 sQTLs (0.52 with Altrans) or Altrans sQTLs (0.66 with Cufflinks2).

To further ensure that our sQTLs were not simply false positives, we verified that LeafCutter found stronger associations between intronic splicing levels and single-nucleotide polymorphisms (SNPs) previously identified as exon expression quantitative trait loci (eQTLs) and trQTLs in GEUVADIS[29] compared with genome-wide SNPs (Fig. 4a). Importantly, 399 (81.3%) of the 491 top trQTLs tested were significantly associated ($P < 0.05$) with intron splicing variation as identified by LeafCutter (compared with 4.7% when our samples were permuted; Supplementary Note 1). Furthermore, we confirmed that the sQTLs we identified are located near splice sites, are close to the introns they affect (Fig. 4b and Supplementary Fig. 13), and are enriched in expected functional annotations such as "splice regions" and DNase I hypersensitivity regions (Supplementary Fig. 14).

We used LeafCutter to identify sQTLs in four tissues from the GTEx Consortium. Overall, we found 442, 1,058, 1,047, and 692 sQTLs at 1% FDR in heart, lung, thyroid gland, and whole blood, respectively (Supplementary Note 1). Using these, we estimated that 75–93% of sQTLs are replicated across tissue pairs (Fig. 4c, Supplementary Fig. 15, and Supplementary Note 1). This is in agreement with the idea of a high proportion of sharing of sQTLs across tissues[30], and contrasts with much lower pairwise sharing reported for these data previously (9–48%)[20]. The high level of replication is likely due to LeafCutter's increased power in detecting genetic associations with specific splicing events. Nevertheless, this leaves 7–25% of sQTLs that showed tissue specificity in our analysis. As expected, we found that a large proportion of tissue-specific

Table 1 | Summary of sQTLs identified in GEUVADIS samples by LeafCutter, Altrans[16], and Cufflinks2[5]

Method	YRI sQTLs (1% FDR)	YRI sQTLs (5% FDR)	CEU sQTLs (5% FDR)
LeafCutter	1,294	1,982	5,775
Altrans	624	1,083	N/A
Cufflinks2	888	1,459	N/A
GEUVADIS study	N/A	83	620

The numbers of trQTLs identified in the original GEUVADIS study[29] are also listed. N/A, not available.

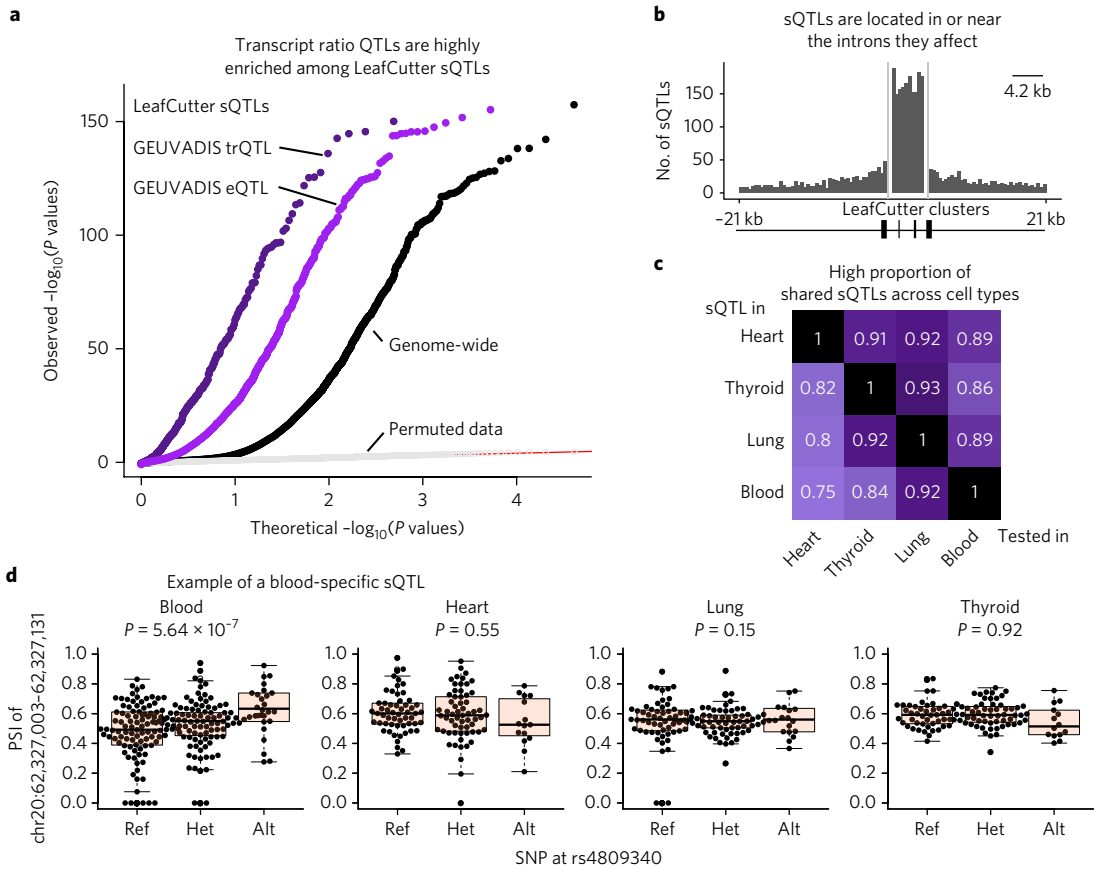

Fig. 4 | LeafCutter sQTLs augment the interpretation of GWAS hits. a, A quantile-quantile (QQ) plot showing genome-wide sQTL signal in LCLs (black), sQTL signal conditioned on exon eQTLs (purple), and sQTL signal conditioned on trQTLs (dark purple) from a previous study[29] ($n = 372$). Signal from permuted data (light gray) shows that the test was well calibrated. **b**, The positional distribution of sQTLs across LeafCutter-defined intron clusters; 1,421 of 4,543 sQTLs lay outside the boundaries (Supplementary Fig. 13 for all sQTLs). **c**, The high proportion of shared sQTLs across four tissues from a previous study[20]. **d**, An example of an SNP associated with the excision level of an intron in blood but not in other tissues. Center line, median; box edges, interquartile range (IQR); whiskers, range of data, excluding outliers beyond 1.5 × IQR; black circles, individual data points. PSI, percent spliced in; Ref, homozygous reference allele; Het, heterozygous allele; Alt, homozygous alternate allele.

sQTLs arose from trivial cases where the intron was only alternatively excised, and therefore variable, in one tissue (Supplementary Fig. 16). However, we also found cases in which the introns were alternatively excised in all tissues yet showed tissue-specific association with genotype (Fig. 4d).

LeafCutter sQTLs link disease variants to mechanism. Finally, we asked whether sQTLs identified by LeafCutter could be used to ascribe molecular effects to disease-associated variants as determined by GWASs. eQTLs are enriched for disease-associated variants, and it is likely that disease-associated variants that are eQTLs function by modulating gene expression[20,29]. We recently showed that sQTLs identified in LCLs are also enriched among autoimmune-disease-associated variants[17]. LeafCutter sQTLs can therefore be used to characterize the functional effects of variants associated with complex diseases. Indeed, when we looked at the association signals of the top eQTLs and LeafCutter sQTLs from GEUVADIS for multiple sclerosis and rheumatoid arthritis (Supplementary Note 1), we found that both QTL types were enriched for stronger associations (Fig. 5a) compared with genome-wide variants. Consistent with recent findings[17], SNPs associated with multiple sclerosis were more highly enriched among sQTLs than among eQTLs, whereas both eQTLs and sQTLs were similarly enriched among SNPs associated with rheumatoid arthritis (Fig. 5a).

To further explore the utility of LeafCutter sQTLs for understanding GWAS signals, we applied S-PrediXcan[31] to compute the association between predicted splicing quantification and 40 complex trait GWASs using models trained on GEUVADIS data (Methods and Supplementary Note 1). In an analysis with a rheumatoid arthritis GWAS, we found that considering intronic splicing allowed us to identify 18 putative disease genes (excluding genes in the extended MHC region), of which 13 were not associated on the basis of gene-expression-level measurements (Fig. 5b). Novel putative disease genes associated through intronic splicing included *CD40*, a gene previously found to affect susceptibility to rheumatoid arthritis[32]. However, we found no overall enrichment of functional categories among the 18 or 13 putative disease genes. Overall, using LeafCutter splicing quantifications allowed us to increase the number of putative disease genes by an average of 2.1-fold as compared to that obtained with the use of gene expression alone (Supplementary Dataset 1). These results demonstrate that by increasing the number of detected sQTLs, LeafCutter considerably enhanced our ability to predict the molecular effects of disease-associated variants.

Discussion

Although we applied LeafCutter to short-read RNA-seq data, the principles of LeafCutter could also be applied to long-read technologies. Long-read technologies may be particularly helpful with gene

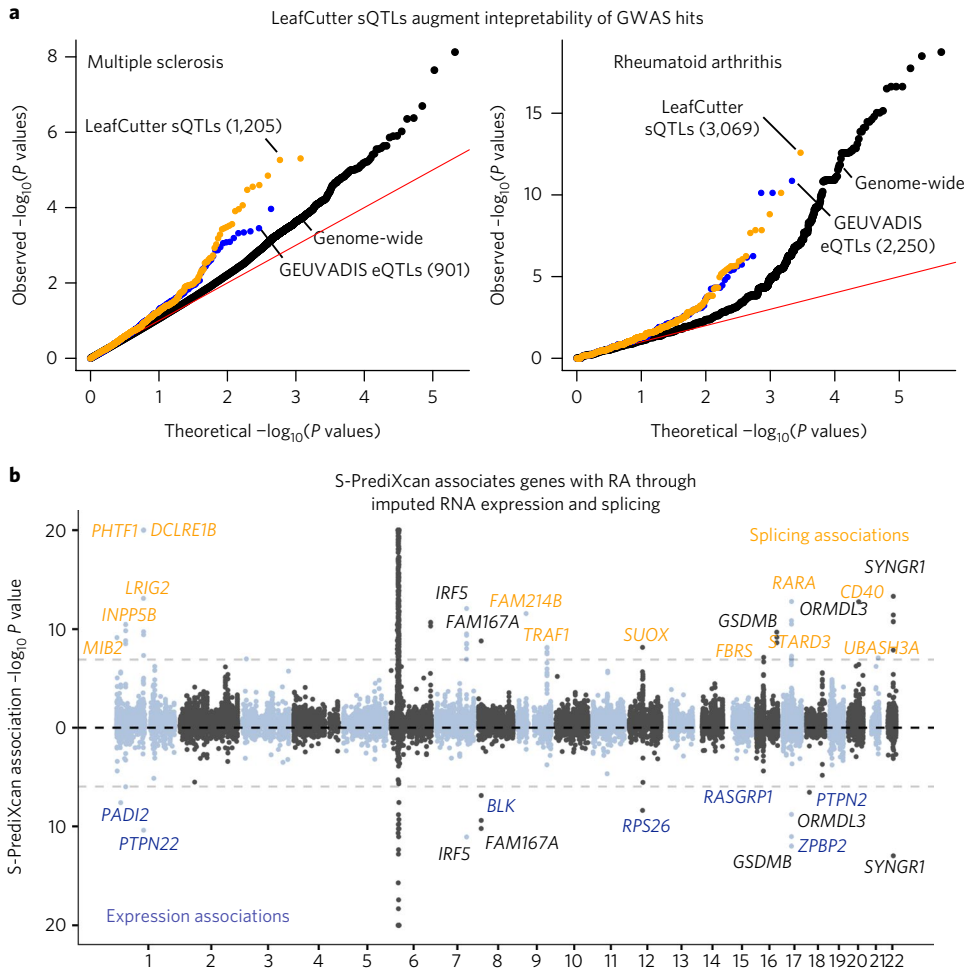

Fig. 5 | LeafCutter sQTLs enable interpretation of disease variants. a, Enrichment of low *P* value associations with multiple sclerosis and rheumatoid arthritis among LeafCutter sQTL and GEUVADIS eQTL SNPs. The numbers of top sQTLs and eQTLs tested in each GWAS are shown in parentheses. **b**, A Manhattan plot of S-PrediXcan association *P* values from prediction models for intron quantification (LeafCutter) and gene expression (GEUVADIS). Genes that were found to be associated through RNA splicing are highlighted in orange, those associated through gene expression are in purple, and those associated through both are in black. The names of associated genes from the extended MHC region are not shown. RA, rheumatoid arthritis.

families for which it is currently difficult to resolve splicing clusters with short reads because of multiple mapping.

In conclusion, our analyses show that LeafCutter is a powerful approach for studying variation in alternative splicing. By focusing on intron removal rather than exon inclusion rates, we were able to accurately measure the step-wise intron-excision process orchestrated by the splicing machinery. Our count-based statistical modeling, which accounts for overdispersion, allows the identification of robust variation in intron excision across conditions. Most important, LeafCutter allows the discovery of far more sQTLs than other contemporary methods, which improves the interpretation of disease-associated variants.

URLs. LeafCutter software, https://github.com/davidaknowles/leafcutter; LeafViz visualizations, https://leafcutter.shinyapps.io/leafviz/; rheumatoid arthritis summary statistics, http://plaza.umin.ac.jp/yokada/datasource/software.htm.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-017-0004-9.

Received: 4 April 2017; Accepted: 8 November 2017;
Published online: 11 December 2017

References

1. Han, H. et al. MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature* **498**, 241–245 (2013).
2. Calarco, J. A. et al. Regulation of vertebrate nervous system alternative splicing and development by an SR-related protein. *Cell* **138**, 898–910 (2009).
3. Brett, D., Pospisil, H., Valcárcel, J., Reich, J. & Bork, P. Alternative splicing and genome complexity. *Nat. Genet.* **30**, 29–30 (2002).
4. Pai, A. A. et al. Widespread shortening of 3′ untranslated regions and increased exon inclusion are evolutionarily conserved features of innate immune responses to infection. *PLoS Genet.* **12**, e1006338 (2016).
5. Trapnell, C. et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31**, 46–53 (2013).
6. Leng, N. et al. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* **29**, 1035–1043 (2013).
7. Patro, R., Mount, S. M. & Kingsford, C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* **32**, 462–464 (2014).
8. Bray, N., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal RNA-Seq quantification. Preprint available at https://arxiv.org/abs/1505.02710 (2015).
9. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–1015 (2010).

10. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22**, 2008–2017 (2012).
11. Lacroix, V., Sammeth, M., Guigo, R. & Bergeron, A. Exact transcriptome reconstruction from short sequence reads. In *Algorithms in Bioinformatics* (eds. Crandall, K.A. & Lagergren, J.) 50–63 (Springer, Berlin, Heidelberg, 2008).
12. Vaquero-Garcia, J. et al. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife* **5**, e11752 (2016).
13. Stein, S., Lu, Z. X., Bahrami-Samani, E., Park, J. W. & Xing, Y. Discover hidden splicing variations by mapping personal transcriptomes to personal genomes. *Nucleic Acids Res.* **43**, 10612–10622 (2015).
14. Zhao, K., Lu, Z. X., Park, J. W., Zhou, Q. & Xing, Y. GLiMMPS: robust statistical model for regulatory variation of alternative splicing using RNA-seq data. *Genome Biol.* **14**, R74 (2013).
15. Monlong, J., Calvo, M., Ferreira, P. G. & Guigó, R. Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nat. Commun.* **5**, 4698 (2014).
16. Ongen, H. & Dermitzakis, E. T. Alternative splicing QTLs in European and African populations. *Am. J. Hum. Genet.* **97**, 567–575 (2015).
17. Li, Y. I. et al. RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).
18. Tilgner, H. et al. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* **22**, 1616–1625 (2012).
19. Wu, J., Anczuków, O., Krainer, A. R., Zhang, M. Q. & Zhang, C. OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. *Nucleic Acids Res.* **41**, 5149–5163 (2013).
20. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
21. Soumillon, M. et al. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep.* **3**, 2179–2190 (2013).
22. Kaessmann, H. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* **20**, 1313–1326 (2010).
23. Nellore, A. et al. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. *Genome Biol.* **17**, 266 (2016).
24. Shen, S. et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. USA* **111**, E5593–E5601 (2014).
25. Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* **338**, 1593–1599 (2012).
26. Barbosa-Morais, N. L. et al. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587–1593 (2012).
27. Reyes, A. et al. Drift and conservation of differential exon usage across tissues in primate species. *Proc. Natl. Acad. Sci. USA* **110**, 15377–15382 (2013).
28. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).
29. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
30. Hsiao, Y. H. et al. Alternative splicing modulated by genetic variants demonstrates accelerated evolution regulated by highly conserved proteins. *Genome Res.* **26**, 440–450 (2016).
31. Barbeira, A.N. et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Preprint available at https://www.biorxiv.org/content/early/2017/10/03/045260 (2017).
32. Orozco, G. et al. Association of CD40 with rheumatoid arthritis confirmed in a large UK case-control study. *Ann. Rheum. Dis.* **69**, 813–816 (2010).
33. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).
34. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

Acknowledgements

We thank X. Lan and other members of the Pritchard lab for helpful discussions and comments. This work was supported by a CEHG fellowship (Y.I.L.), the Howard Hughes Medical Institute (J.K.P.), and the US National Institutes of Health (NIH grants HG007036, HG008140, and HG009431 to J.K.P., and MH107666 to H.K.I.).

Author contributions

Y.I.L., D.A.K., and J.K.P. conceived of the project. Y.I.L. and D.A.K. performed the analyses and implemented the software. D.A.K. developed and performed the statistical tests and modeling. J.H. implemented the visualization application. A.N.B., S.P.D., and H.K.I. performed the S-PrediXcan analyses. Y.I.L. and J.K.P. wrote the manuscript.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-017-0004-9.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to Y.I.L. or D.A.K. or J.K.P.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Identifying alternatively excised introns. To identify clusters of alternatively excised introns, split reads that map with a minimum of 6 nt into each exon are extracted from aligned.bam files. Overlapping introns defined by split reads are then grouped together. For each of these groups, LeafCutter constructs a graph in which nodes represent introns and edges represent shared splice junctions between two introns. The connected components of this graph define intron clusters. Singleton nodes (introns) are discarded. For each intron cluster, LeafCutter iteratively (1) removes introns that are supported by fewer than a specified number (default: 30) of reads across all samples or less than a proportion (default: 0.1%) of the total number of intronic read counts for the entire cluster, and (2) re-clustered introns according to the procedure described above.

Dirichlet-multinomial generalized linear model. Intron clusters identified by LeafCutter comprise two or more introns. More specifically, each intron cluster c identified by LeafCutter consists of J possible introns, which have counts $\mathbf{y}_{ij}$ for sample i and intron j (and cluster total $\mathbf{n}_{ic} = \sum_{j'} \mathbf{y}_{ij'}$), and N covariate column vectors $\mathbf{x}_i$ of length p. LeafCutter uses a Dirichlet-multinomial ($\mathscr{DM}$) generalized linear model to test for changes in intron usage across the entire cluster, instead of testing the differential excision of each intron separately across conditions or genotypes.

$$\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iJ} \mid \mathbf{n}_i \sim \mathscr{DM}\left(\mathbf{n}_{ic}, \alpha_1 p_{i1}, \ldots, \alpha_J p_{iJ}\right)$$

$$p_{ij} = \frac{\exp\left(\mathbf{x}_i \beta_j + \mu_j\right)}{\sum_{j'} \exp\left(\mathbf{x}_i \beta_{j'} + \mu_{j'}\right)}$$

where the softmax transform is used to ensure that $\sum_j p_{ij} = 1$. We perform maximum likelihood estimation for the outputs: the J coefficient row vectors β_j of length p, the intercepts μ_j, and the concentration parameters α_j. We use the following regularization to stabilize the optimization:

$$\alpha \sim \gamma(1 + 10^{-4}, 10^{-4})$$

The Dirichlet-multinomial likelihood is derived by integration over a latent probability vector π in the hierarchy

$$\pi \mid a \sim \text{Dirichlet}(a) \Rightarrow P(\pi \mid a) = \frac{\Gamma(a_\cdot)}{\prod_j \Gamma(a_j)} \prod_j \pi_j^{a_j - 1}$$

$$y_1, \ldots, y_J \mid n, \pi \sim \text{Multinomial}(n, \pi) \Rightarrow P(y \mid n, \pi) = \prod_j \pi_j^{y_j}$$

where $a_\cdot = \sum_j a_j$, to give

$$\mathscr{DM}(y \mid n, a) = \frac{\Gamma(a_\cdot) \prod_j \Gamma\left(a_j + y_j\right)}{\Gamma(a_\cdot + y_\cdot) \prod_j \Gamma(a_j)}$$

In the limit $\pi_j = e^{a_j} / \sum_{j'} e^{a_{j'}}, a_j \to \infty$ for all j, we have $\mathscr{DM}(n, a) \to \text{Multinomial}(n, \pi)$. For the generalized linear model this means that as $\alpha_j \to \infty$ we recover a multinomial model with no overdispersion. Smaller values of α_j correspond to more overdispersion.

Differential intron excision across conditions. To test differential intron excision between two groups of samples, we encode $x_i = 0$ for one group and $x_i = 1$ for the other in the Dirichlet-multinomial generalized linear model. For each cluster we compare the null model with only the intercept term to an alternative model that includes x by using a likelihood ratio test with $K - 1$ degrees of freedom, where K is the number of introns in the cluster.

We apply two filters to ensure that we perform only reasonable tests:

Only introns that are detected (i.e., that have at least one corresponding spliced read) in at least five samples are tested.

A cluster is tested only if each group includes at least four individuals with 20 spliced reads supporting introns in the cluster.

The thresholds in these filters are easily customizable as optional parameters.

Mapping splicing quantitative trait loci. For sQTL identification, RNA-seq reads are mapped onto the genome with an RNA aligner such as STAR[35] or OLego[19]. Because LeafCutter uses only reads that map across junctions to estimate intron excision rates, it is essential to remove read-mapping biases caused by allele-specific reads. This is particularly important when a variant is covered by reads that also span intron junctions, as this can lead to spurious association between the variant and intron-excision-level estimates. Subsequent to mapping, LeafCutter finds alternatively excised intron clusters and quantifies intron excision levels in all samples. LeafCutter outputs intron excision proportions, which are used as input for standard QTL mapping tools such as MatrixEQTL and fastQTL (Supplementary Note 1).

S-PrediXcan analyses. Prediction models for intron quantification (LeafCutter) and gene expression (GEUVADIS) were trained with Elastic Net on GEUVADIS data. A value of $\alpha = 0.5$ was chosen for the mixing parameter. Prediction performance for gene expression remains stable for a wide range of mixing parameters when α does not approach 0.0 (ridge regression)[36,37]. For each gene, we used SNPs within 1 Mb upstream of the transcription start site and 1 Mb downstream of the transcription end site. Similar windows around each splicing cluster were chosen.

We downloaded genome-wide association meta-analysis summary statistics for rheumatoid arthritis ("URLs") and ran S-PrediXcan using these models. We obtained a total of 4,625 gene associations for the genetic expression model, and 41,196 intron quantification cluster associations for the splicing model, that had a model prediction FDR less than 5%.

Visualizing LeafCutter differential splicing output. Using the R Shiny framework and ggplot2, we created an interactive browser-based application, LeafViz, that allows users to visualize LeafCutter differential splicing analyses. LeafViz generates LeafCutter cluster plots with information on the significance of the detected differential splicing and the estimated differences of the splicing changes. All significant clusters are labeled as "annotated" or "cryptic" by intersecting junctions with a user-defined set of transcripts (e.g., gencode v19). Users can directly download plots from the website in PDF format, and these plots can be easily edited for publication. An example of LeafViz applied to a differential splicing analysis of ten brain and ten heart samples from GTEx is available online ("URLs").

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Data availability. The datasets analyzed during the current study are available through dbGaP under accession phs000424.v6.p1 (GTEx), GEO under accession GSE41637 (RNA-seq data from mammalian organs), and ENA under accession PRJEB3366 (Geuvadis).

References

35. Wheeler, H. E. et al. Survey of the heritability and sparse architecture of gene expression traits across human tissues. *PLoS Genet* **12**, e1006423 (2016).
36. Gamazon, E. R. et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
37. Ellis, S.E., Collado Torres, L. & Leek, J. Improving the value of public RNA-seq expression data by phenotype prediction. Preprint available at http://www.biorxiv.org/content/early/2017/06/03/145656.full.pdf (2017).

nature research

Corresponding author(s): Yang Li

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

   Describe how sample size was determined. — Not applicable.

2. Data exclusions

   Describe any data exclusions. — Either no data was excluded, or random sampling was used to downsample.

3. Replication

   Describe whether the experimental findings were reliably reproduced. — We replicated our unannotated introns in one additional dataset (that uses another mapping method, Rail-RNA).

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups. — Not applicable.

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis. — Not applicable.

   Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☒	☐	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	We used standard two standard RNA-seq mappers: OLego and STAR

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	All materials are available to download online or accessible through dbgap freely.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eukaryotic cell lines were used.
b. Describe the method of cell line authentication used.	No eukaryotic cell lines were used.
c. Report whether the cell lines were tested for mycoplasma contamination.	No eukaryotic cell lines were used.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	The study did not involve human research participants.