Reference-free analysis of mRNA isoform variation in large-scale RNA-seq datasets

Carlos F Buen Abad Najar[1], Dongyue Xie[2], Peter Carbonetto[3], Ru Feng[4], Gao Wang[4], Matthew Stephens[2,3]*, Yang I Li[1,3,5,]*

1. Section of Genetic Medicine, The University of Chicago, Chicago, IL
2. Department of Statistics, The University of Chicago, Chicago, IL
3. Department of Human Genetics, The University of Chicago, Chicago, IL
4. Center for Statistical Genetics, The Gertrude H. Sergievsky Center, Department of Neurology, Columbia University, New York, NY, USA
5. Chan Zuckerberg Chicago, Chicago, IL
* Corresponding authors

Abstract

Dysregulation of mRNA processing plays a critical role in most human diseases. Yet, identifying dysregulated mRNA isoforms remains challenging because disease-causing isoforms are often unannotated. Here, we present Torino, a pipeline based on a novel matrix factorization algorithm that simultaneously learns mRNA isoform structure and their relative expression levels from RNA-seq read coverage alone. We show that Torino accurately identifies patterns of variation in isoform structure across tissues from GTEx RNA-seq data. In addition to annotated alternative splicing events, Torino uncovers thousands of unannotated intron retention and alternative polyadenylation events that vary in usage across tissue-types, potentially corresponding to thousands of unstudied functional gene regulatory events. We also show that Torino captures the impact of inter-individual genetic heterogeneity on read coverage, helping pinpoint the effects of hundreds of disease-associated variants on mRNA isoform levels. Finally, we applied Torino on 1,193 RNA-seq samples from the Alzheimer's Disease Functional Genomics consortium. Our analysis reveals a global increase in intron retention associated with AD status and elevated neurofibrillary tangle, including in many AD genes such as *PTK2B and APBB3*. Torino is a powerful, reference-free approach for capturing biological variation in mRNA isoform levels from large RNA-seq datasets.

Introduction

RNA processing impacts transcriptome diversity and gene expression levels[1]. AS produces different messenger RNA (mRNA) isoforms from the same gene, diversifying the proteome in a tissue-specific manner. Recently, we discovered that AS of aberrant isoforms coupled with nonsense-mediated decay (NMD) affects an average of 15% of all transcripts from virtually all protein coding genes, impacting gene expression levels[2]. Intron retention also results in NMD-targeted transcripts[3], and APA leads to different 3' untranslated regions, impacting mRNA stability, localization, and other features[4]. The rapid advancement in sequencing technologies has resulted in a wealth of transcriptomic data that presents new opportunities to study these complex processes. However, one difficulty that prevents the full exploitation of this information is the lack of computational approaches that can integrate this data into interpretable models of isoform variation.

Quantification of full mRNA isoforms with short-read sequencing data (RNA-seq) relies on prior external knowledge of the transcript structures (i.e., isoform annotations), which can be detrimental when annotations are incomplete, incorrect, or when most annotated isoforms are irrelevant. On the other hand, most reference-free approaches rely on splice junctions to detect and quantify AS events[5], but they either do not work on full isoform structures, or they are unable to capture isoform variation that is not associated with splice junctions such as intron retention and APA. Additionally, differential analysis of isoform relative abundances usually require a priori stratification of the data into categorical conditions (e.g., “condition” and “controls”), which hampers our ability to find important biological variance across the samples that does not follow the preset stratification.

De-novo transcriptome assembly approaches such as Stringtie[6] address the issue of incomplete annotations by inferring the transcripts present in RNA-seq samples directly from sequence alignment files. Alternatively, long read sequencing (LRS) technologies are able to detect and quantify full isoform structures independently of annotation, and studies suggest that the majority of mRNA isoforms captured by LRS technologies are currently unknown[7]. However, LRS approaches have a higher error rate and a lower sequencing depth than the short-read counterparts, hindering its ability to capture isoform variation, and the availability of LRS datasets is much smaller than the wealth of short-read RNA-seq data on existence. While de-novo transcriptome assembly and the extensive catalogs of transcripts from LRS provides a reference-free approach to isoform quantification, recent studies showcasing the high prevalence of unproductive splicing in the transcriptome suggest that the majority of these novel transcripts are the result of the combination of random RNA processing events that result in unproductive transcripts. As a result, merely characterizing all the possible isoforms for each gene might be less informative for functional studies than learning the variable RNA processing events whose combination results in these transcripts. Furthermore, differential analysis with de-novo assembled transcriptomes and with LRS data still require *a priori* stratification of the data, since it does not perform an agnostic analysis of the isoform variance between samples.

Here we present Torino, a novel computational approach for reference-free characterization, visualization, and quantification of isoform structures from large-scale short-read RNA-seq datasets. Torino uses an approach based on topic modeling (TM) to learn isoform structures directly from short-read RNA-seq profiles[8]. TM has previously been used in multiple contexts to discover abstract “topics” from a collection of samples, e.g. to classify texts based on inferred topics and to estimate admixture from genetic data. More recently, TM was applied to model the correlated structure of sequencing read coverage across adjacent bases, revealing coordinated variation in enhancer activity from ChIP-seq coverage data. Torino leverages the ability of TM to model coordinated variation in RNA-seq coverage across thousands of samples. We show that Torino captures variable, and interpretable mRNA isoform structures of protein-coding genes and their relative usage. This allows us to capture and quantify biologically-relevant variation in isoform expression without gene annotation, or prior sample stratification.

Results

Overview of the Torino Model

Torino uses a novel Bayesian Poisson Non-negative Matrix Factorization (NMF) algorithm to learn a low dimensional representation of RNA-seq read coverages on genomic intervals. Torino infers factors (i.e. “topics”) and loadings that best explain the variability in read coverages observed across samples. Torino factors are modeled to favor similarity in coverage across adjacent bases (later referred to as spatial constraint), which is the key feature that allows it to capture patterns of exonic and intronic coverages typical of RNA-seq data, while simultaneously minimizing the effects of noise and sequencing biases.

To obtain a more intuitive understanding of Torino, consider X a $n \times m$ matrix of counts across a genomic interval (e.g., read counts from an RNA-seq experiment across a gene with length m base pairs in n biological samples). We model K factors of the data ($n > K > 1$) by approximating $X \approx LF$. F is a $k \times m$ matrix in which each row f_k, or ***factor***, is a latent structure of the data. f_k is a vector of length m, such that f_{ki} corresponds to the expected reads in base pair i contributed by the latent structure. L is a $n \times k$ matrix in which l_{jk} indicates how much the factor f_k is represented in sample j (***loading***). We model the counts as a Poisson distribution with identity link function: $x_{ij} \sim Poisson(s_i \sum_{k=1}^{K} l_{jk} f_{ki})$, where $s_i = \sum_{j=1}^{m} x_{ij}$ is the total number of counts in sample i. Notably, f_k is modeled using a smooth prior $f_k \sim g(.)$ modeled as a wavelet function to add the spatial constraint that is crucial for recovering interpretable factors that resemble mRNA isoforms (see Methods). We use the identity link to the Poisson function to model the counts as a mixture of the underlying components. This allows us to interpret the loadings as the relative abundances of the factors used to explain the coverage observed in each individual sample. In summary, Torino models RNA-seq coverage in a genomic interval as the product of latent factors (*k* possible mRNA isoforms corresponding to each factor f_k) and their relative abundances (corresponding to each loading l_k).

As input, Torino takes RNA-seq coverages (bigWig), junction files (e.g. from STAR[9] or LeafCutter[5]), and a gene annotation file (e.g. GENCODE[10] gtf). Torino compiles the RNA-seq coverages into matrices of read counts spanning genomic intervals corresponding to the body of each gene. Each row of a matrix corresponds to a single sample, while each column corresponds to a base-pair position (**Fig. 1A**). Our Bayesian Poisson NMF model simultaneously learns the main structures of variance across the gene body (factors) and relative proportion of the factors for each sample (loadings) (**Fig. 1B**). Thus, Torino infers coverage factors (**Fig. 1C**) whose linear combination predicts read coverage across samples. Finally, Torino uses read junctions to refine the inferred factors and collapses highly similar factors to increase interpretability (**Methods**).

As an illustrative example, we showcase the factors (*k = 10*) inferred by Torino when applied to read coverages at the *SRSF3* gene from 1,000 GTEx[11] samples from 10 different tissue-types. We first note that constitutively spliced exons are clearly captured by most all factors (**Fig. 1D, Supplemental Figure 1**). More interestingly, several factors capture a well-known cassette exon in *SRSF3*, while others show exon skipping. Additionally, a subset of factors capture differences in coverage at the 3’ end of *SRSF3*, consistent with a lengthening of the *SRSF3 3’* UTRs in brain samples relative to other tissue-types. Finally, two factors capture read coverage in introns, especially surrounding *SRSF3*’s cassette exon, representing a variability in the retention levels of these introns across samples.

As can be seen for *SRSF3*, Torino’s factors are easily interpretable and can be used for downstream analysis without further processing. However, to further facilitate interpretability and

to allow comparisons to annotated transcripts, we developed a pipeline to ascribe our factors to RNA isoforms by matching the spatial structures captured by each factor to junction read coordinates. Our approach also takes into consideration potential intron retention and alternative polyadenylation events (**Fig. 2A**), which are generally missed by popular annotation-free junction based approaches including LeafCutter. Applying this pipeline to the factors obtained on *SRSF3* confirmed that Torino captured all constitutive exons of *appris* principal[12] protein-coding isoforms (Gencode), as well as a poison exon in *SRSF3*'s well-known NMD-targeted isoform. Notably, the largest differences between Torino factors and GENCODE transcripts were intron retention events and alternative polyadenylation sites that were missing in GENCODE (**Fig 2A,B**). Interestingly, many factors showed variable loadings across tissues-types, with whole blood samples presenting a higher abundance of the poison exon factor, while the factor capturing the unannotated intron retention was more common in heart, lung and liver samples (**Fig 1C**).

Variation in mRNA isoform structure across human tissue-types

We applied Torino on 15,232 protein-coding genes with a minimum of 100 reads across at least 30 of the 1000 GTEx samples analyzed. To ensure that read coverages represent mRNA from a single gene, we removed regions that overlap more than one gene (Methods). In spite of this, factors learned by Torino fully match at least one GENCODE-annotated transcript for the majority of genes. Indeed, Torino factors fully match 18,813 GENCODE-annotated isoforms (**Fig. 2C**) from 11,597 genes (**Fig. 2D**), or 76% of all tested genes. When we considered only genes with no annotated overlaps, we found that 5,480 of 6274 genes (87.3%) had at least one Torino factor that fully matches with a GENCODE-annotated transcript.

We next assessed Torino's ability to capture alternative splicing events. We first searched for possible cassette exons among Torino factors, which we defined as segments that are present in one or more factors, but absent in other factors. We found 28,464 potential cassette exons across 8,774 genes. A large minority of these exons (13,058) are observed as cassette exons in Gencode or in VastDB[13] (**Fig. 2E**), while nearly all remaining exons (10,355) are annotated as constitutive (6,597) or overlap partially with annotated exons (3,758). The remaining 5,051 potential cassette exons did not overlap with any annotated exons, and their relative abundances were lower than the other cassette exons (**Supplemental Figure 2A**). Despite their lower relative abundance compared with annotated cassette exons, these structures show variability across samples (**Supplemental Figure 2B**). Given that the GTEx data is unstranded, these structures could represent overlapping cryptic transcripts that are missing from the annotation. This highlights Torino's ability to discover new variable structures, even when their characterization is not straightforward.

The factors in Torino represent structure in the read coverage that explains the most variance in coverage across the modeled samples. Indeed, we found that the Torino cassette exons that are also annotated by GENCODE (3,802) or VastDB (12,259) show high variance across the GTEx samples (**Fig. 2F**, **Supplementary Figure 3**). By contrast, the 9,567 GENCODE-annotated cassette exons and 108,328 VastDB-annotated cassette exons that appear as constitutive exons in Torino factors (i.e. present in all factors) show high exon inclusion across all samples. Furthermore, the remaining 11,449 GENCODE-annotated cassette exons and 10,812 VastDB-annotated cassette exons not detected by Torino (i.e., absent in all factors but whose flanking exons are present) show low inclusion across all samples.

These results show that Torino captures major axes of mRNA isoform variation across input samples without relying on existing annotation.

Torino reveals pervasive unannotated variation in intron retention and alternative polyadenylation

The most abundant alternative structure that Torino found across read coverages from GTEx samples was intron retention events. In total, 83,553 retained introns were identified in 12,472 genes, corresponding to 82% of all genes tested (**Fig. 2D,E**). Strikingly, the majority of retained introns (53,719) are not associated with any GENCODE transcript (**Fig. 3A**), while 17,441 overlap entirely (6,828) or partially (10,613) with retained introns from GENCODE (**Fig. 3A,B**). Only 1,992 intron retention events annotated by GENCODE were not captured by a factor with Torino. Our results therefore suggest widespread usage of intron retention events that are missed by GENCODE (**Fig. 3C**). As expected, intron retention events captured by Torino that overlap with GENCODE transcripts showed the highest coverage, followed by unannotated intron retention events captured by Torino (**Fig. 3D**). In contrast, intron retention events annotated by GENCODE but not by Torino have lower intronic coverage, suggesting that these introns are rarely retained in GTEx samples (**Fig. 3D**, **Supplementary Figure 4**). Finally, and as expected, all other introns showed very low coverage, consistent with efficient co-transcriptional splicing. Interestingly the majority of protein-coding gene introns are annotated as potentially retained in the VastDB database, with 157,721 out of 169,217 appris principal introns being annotated as retained in VastDB (Supplementary Figure 5A). This includes the majority of intron retention events found by Torino (55,077 with total overlap, 22,936 with partial overlap, and 5,540 with no overlap). However, only introns captured by Torino show appreciable coverage. Indeed, coverage at retained introns annotated in VastDB alone was indistinguishable from baseline coverage (Supplementary Figure 5B). Thus, Torino identifies and quantifies isoform structures that explain variation in read coverage across samples. The structures identified by Torino often include unannotated transcript structures that are variable across input samples, while leaving out many annotated transcripts that explain little to no variance in read coverage.

Similar to other types of splicing events, intron retention events detected using Torino capture global biological differences between GTEx tissues (Supplementary Figure 6A), indicating that they are biologically regulated and are potentially functional. As an example, Torino factors for *SYNGR1* captures two known alternative protein-coding isoforms, as well as three isoform structures with intron retention events, only one of which is annotated in GENCODE (**Fig. 3E**). Part of this region contains. As expected, isoform ENST00000328933 of *SYNGR1* is enriched in brain tissues, while both protein-coding isoforms (ENST00000328933 and ENST00000318801) are present in the other tissue-types (**Fig. 3F**). Notably, all three isoform structures with intron retention are negatively correlated with total *SYNGR1* expression level (Supplementary Figure 6B), consistent with the notion that intron retention results in rapid mRNA degradation through nonsense-mediated decay (NMD).

In addition to intron retention events, Torino factors captured 8,013 alternative polyadenylation events across 2,906 genes (**Fig. 2C,D**). We conservatively required the 3’ end of these events to differ by at least 20% of the final exon length or by more than 200 nucleotides. Still, we found that XX of these alternative polyadenylation events are unannotated in GENCODE. For example, Torino finds three different isoform structures associated with the protein-coding isoform of *SRSF3*, differing only at the 3’ UTR polyadenylation site (**Fig. 3G**). Two of these polyadenylation sites are annotated, but the third isoform corresponds to a longer 3’ UTR that is not annotated by

GENCODE. Loadings for the factor representing the unannotated 3' UTR are higher in brain tissue samples, while that for the factors representing the shortest 3' UTR was higher in skeletal muscle samples. These observations are consistent with previous findings of 3' UTR lengthening in neurons, and shortening in skeletal muscle.

In many cases, we found that some GENCODE transcripts have long 3' UTRs; however, little coverage is observed in this region across GTEx samples. Instead, we observed that novel 3' polyadenylation sites identified by Torino have higher inter-tissue variation in the GTEx tissues we analyzed compared to GENCODE-annotated alternative 3' UTRs that are not captured by Torino (**Fig. 3H**). Thus, many GENCODE-annotated transcripts representing alternative polyadenylation events have no support from GTEx samples, while Torino learns relevant mRNA isoform structure directly from RNA-seq coverage data.

Impact of common and trait-associated variants on mRNA isoform levels

We have shown so far that Torino learns structures in RNA-seq coverages that can be interpreted as different mRNA isoforms, and estimates loadings that correspond to the relative usage of each mRNA isoform. We next investigated whether the loadings estimated by Torino capture variability in mRNA isoform levels resulting from genetic variation across individuals from the GTEx samples. To do this, we turned to quantitative trait loci (QTL) mapping, using loadings as the phenotypic values of interest. Specifically, we used QTLTools to test for associations between genetic variants and the loadings of the factors obtained for each gene by Torino and each tissue separately. Since loadings for Torino factors represent the relative abundances of isoform structures, we reasoned that the impact of genetic variants could be reflected in multiple factors per gene. To account for this, we ran QTLTools[14] using the grouped permutation pass, obtaining for each gene the best QTL match. This resulted in a median of 2,829 isoform QTLs per tissue, impacting a median of 1,892 genes (**Fig. 4A**). The lead QTL variants were distributed across the gene body with a slight enrichment towards the transcription ending site (Supplementary figure), unlike eQTLs which are enriched at transcription start sites. This enrichment is consistent with previously reported distributions of QTLs affecting splicing, intron retention and alternative polyadenylation, confirming that our approach captures genetic associations with the relative abundance of isoform structures rather than total gene expression level. We observed a high degree of isoform QTL sharing across all tissues, with a median Pi1 of 0.9 for QTLs with FDR ≤ 1e-4, and a median of 0.96 across brain tissues (**Fig. 4B**). These observations confirm previous findings that variant effects on splicing, intron retention and alternative polyadenylation are highly shared across different tissues.

Variant effects on alternative splicing[15], intron retention[16], and alternative polyadenylation sites[4,17] are known to be associated with complex disease. In addition, many expression QTLs associated with common disease act through alternative splicing of unproductive isoforms that are degraded through NMD[2]. However, the molecular mechanism at most GWAS loci remains unknown. We reasoned that our reference-free approach might allow us to identify molecular effects on mRNA isoform levels that are difficult to assess without annotated isoforms, e.g. unannotated intron retention or alternative polyadenylation events. To test this, we compiled summary statistics from 65 genome-wide association studies (GWAS) and used Hyprcoloc to colocalize signals from each trait with Torino isoform QTLs. Of 6,746 lead GWAS SNPs, 815 (12.1%) colocalized with at least one isoform QTL (**Fig. 4C**). Notably, the most common event colocalized with GWAS loci was intron retention (440), followed by cassette exons (387), alternative polyadenylation (186) and other types of splicing events (126).

We found strong colocalizations between genetic variants and multiple isoform QTLs identified using Torino. Notably, these include variants that colocalize with Torino factors representing cryptic splicing events that are not annotated in GENCODE. For example, rs7740107 is an intronic variant in the gene *L3MBTL3* that has been previously identified as an expression QTL and is associated with lymphocyte counts. This variant affects the usage of a cryptic alternative 3' splicing site that is captured by Torino in whole blood samples (**Fig. 4D**). Another variant, rs11057400, is also associated with lymphocyte count and lies in the intron of *CCDC92.* rs11057400 is also strongly associated with an annotated early alternative 3' exon transcript variant that is captured by Torino's factors (Supplementary figure). Interestingly, one of the strongest associations that we found was between rs1045599, a variant associated with Parkinson's disease, and the levels of an alternative polyadenylation site of *ZSWIM7* in brain tissues (**Fig. 4E**). Changes in the relative usage of alternative polyadenylation sites has been previously identified in neurodegenerative diseases, including Parkinson's disease.

We also identified another interesting case involving rs4074793, an intronic variant in the gene *ITGA1* that is associated with coronary artery disease (CAD). rs4074793 has been previously identified as an eQTL of *ITGA1*, possibly acting through an intronic enhancer[18]. We found a strong colocalization of this variant with an increase in the relative usage of an unannotated intron retention event for *ITGA1* in heart atrial appendage and lung tissue (Supplementary Figure). Rs4074793 lies within the unannotated intron retention event, raising the possibility that it can directly impact its splicing efficiency. These results suggest an additional potential mechanism of action of rs4074793 in CAD.

Our results demonstrate that modeling isoform structures with Torino captures the genetic effects on isoform structures without reliance on transcription annotations. Importantly, many of the isoform QTL detected represent cryptic splicing, alternative polyadenylation sites, and intron retention, all of which are difficult to detect and quantify using other popular approaches.

Isoform variation in a large Alzheimer's disease cohort

Finally, we sought to demonstrate Torino's ability to identify biologically-relevant variation in mRNA isoform levels in a large-scale RNA-seq dataset with disease cases and control samples. To do that, we applied Torino on RNA-seq data compiled by the Alzheimer's disease Functional Genomics consortium[19], which comprises of a total of 1,193 post-mortem brain RNA-seq samples from AD patients (n = 390), patients with cognitive impairments other than AD (n = 388), and control individuals (n = 415). These samples were collected from three different brain regions: anterior caudate nucleus (AC), dorsolateral prefrontal cortex (DLPC), and posterior cingulate cortex (PCC). After filtering for minimum coverage per gene, we obtained Torino models for 15,232 genes. Using these, we tested for associations between Torino factor loadings and AD-relevant variables, including AD status and neurofibrillary tangle (as measured by Braak score), by running a logistic regression model for each Torino isoform, using sex and age as covariate. Finally, we computed a single association p-value for each gene by aggregating the p-value of all tested isoforms using the Aggregated Cauchy Association Test (ACAT, Methods).

This analysis revealed widespread changes in mRNA isoform levels associated with AD status. After multiple-test correction, we found a total of 4,943 genes with significant association between isoform structure changes and AD status (FDR ≤ 0.05). Interestingly, genes associated with differences in mRNA isoform levels show a strong enrichment for genes previously found to be differentially expressed in DLPC between control and AD samples (**Fig. 5A**). Additionally,

these genes are enriched in signatures associated with the nervous system such as synapses, neuron projection, axons and other signatures associated with brain activity, as well as neurological disorders such as speech impediment (**Fig. 5B**).

We next asked about the types of isoform changes that differ between healthy and AD brains. Remarkably, for 2,810 of these genes, the factors with the strongest association with AD status corresponded to an intron retention event. Furthermore, AD-associated factors that correspond to intron retention events are overwhelmingly (87.9%) positively correlated with AD-status, indicating that AD is associated with a global increase in intron retention (**Fig. 5C**). In fact, the strongest association between an isoform factor and AD status was in *PKFP*, where an unannotated intron retention increases in samples with AD across all three brain regions (**Fig. 5D**, Supplementary figure). More generally, we found that usage of the vast majority of the AD-associated intron retention isoforms increase in AD across all brain regions (supplementary figure). This includes ten causal AD genes: *APBB3*, *BCKDK*, *ICA1*, *IDUA*, *MYO15A*, *PLEKHA1*, *PTK2B*, *SLC24A4*, *TMEM106B* and *TPCN1*. Importantly, intron retention is correlated with increases in the Braak stage[20], a semi-quantitative measure of neurofibrillary tangle severity used in diagnostics for neurodegenerative diseases such as AD and Parkinson's disease (**Fig. 5E**). Our findings indicate a likely important role of splicing dysregulation in neurodegenerative disorders, mediated by a global increase in intron retention in AD brains.

Discussion

RNA processing events such as alternative splicing, intron retention, and alternative polyadenylation have critical roles in gene regulation and disease pathobiology. The vast wealth of existing transcriptomic data presents an opportunity to explore the function of RNA processing events across thousands of samples from hundreds of tissues and conditions. However, the large diversity and stochastic nature of RNA processing makes it challenging to model and study, especially when relying on existing transcript annotations. We developed Torino to overcome this challenge. Torino learns isoform structures underlying variation in mRNA isoform usage across thousands of samples in large-scale datasets. In particular, Torino identifies isoform structures that are particularly difficult to detect, including intron retention and alternative polyadenylation events which are often poorly annotated.

For the majority of genes Torino identifies at least one annotated protein-coding isoform, which corresponds to an *appris* principal isoform, confirming that our approach captures major isoforms of protein-coding genes. Additionally, Torino identifies alternative splicing events, including cassette exons and cryptic splicing events, and finds that intron retention is the most common source of variance across samples. Surprisingly, thousands of intron retention and alternative polyadenylation events captured across the transcriptome were unannotated reflecting potentially unstudied functional gene regulatory events. In addition to capturing variation across tissue-types, Torino also captures variation in mRNA isoform usage caused by genetic variants. Indeed, we discovered thousands of genetic loci impacting mRNA isoform levels, of which hundreds helped link disease-associated variants to changes in mRNA isoform levels. Notably, many of these variants impact novel cryptic splicing, intron retention, and alternative polyadenylation events, which are difficult to capture from existing approaches.

Torino uncovers variation in mRNA isoform usage in an unsupervised manner. This has several advantages. First, only the largest axis of mRNA isoform variation are captured, ensuring that potential variation identified do not reflect splicing noise or extreme outliers. This implies that

isoform factors learned from one dataset can likely be transferred and used for the analysis of another dataset, akin to projecting samples onto principal components. Second, this approach obviates the need to stratify samples before examining differences in isoform usage, which is convenient for data exploration. Last, and most importantly, Torino requires no transcript annotation, which helps find cryptic isoforms that are activated in complex diseases by global changes in splicing or in rare diseases by DNA mutations.

To showcase the power of Torino on a case-control RNA-seq dataset, we chose a dataset consisting of over one thousand brain samples from patients with Alzheimer's disease (AD) and controls. Our analysis revealed a widespread increase in intron retention events associated with AD, including novel isoforms affecting many genes previously implicated in AD. Previous studies have documented an increase in intron retention across AD brains. Here, we show here that this phenomenon is global and impacts thousands of introns across the genome. We speculate that AD is characterized by a general dysregulation of splicing efficiency which manifests itself as widespread intron retention and induction of cryptic splicing. More work is needed to determine the exact causal factors that increase intron retention in the AD brain. Possible factors include dysregulation of specific splice factors associated with AD, or specific pathobiology related to AD such as Tau pathology, which is suspected to impact splicing efficiency by impacting nuclear speckles biology.

In conclusion, Torino presents a versatile approach to studying transcriptomic variance directly from the data to uncover novel isoform structures and variable regions across genes. Torino reveals tens of thousands of novel structures, including events that are difficult to study without annotations such as intron retention and APA. Our results demonstrate that these often overlooked events are associated with complex disease, which highlights the importance of RNA processing events in human disease.

Methods

RNA-sequencing preprocessing

The input for Torino's factorization step is a matrix of read counts per base pair position of a gene across each sample. For the purpose of isoform structure characterization and quantification Torino focuses only on the coordinates of protein-coding genes. Each read is only counted once (i.e., in one base pair position only) at the 5' end mapping position. We obtain these matrices To obtain this input, for each sample first we transform the Binary Alignment Map (BAM) files into bedGraph files with read counts at the 5' base pair position with the following command:

```
bedtools genomecov -5 -bga -ibam {sample}.bam | bgzip > {sample}.bed.gz
```

Once we have a bedGraph file for all samples, we use a customized script to parse through all files and obtain the read counts in the coordinates defined for each protein coding gene in Gencode v44 plus 50 base pairs at each end. We use pytabix, a python implementation of Tabix, to parse efficiently through the data. In the end we have a matrix size $$n \times m$$,

Smooth Poisson Non-negative Matrix Factorization.

Torino applies a novel approach to factorize count data called smoothed Poisson non-negative matrix factorization (SPNMF), an especial case of Bayes Poisson NMF that uses an identity link function and assumes smoothness-inducing priors on factors.

Given $X \in n \times m$ a count matrix with n samples and m base pair positions, we model the underlying isoform structures by representing the matrix as the product of two low-dimensional matrices, referred to as loadings and factors. Namely, we model $X \approx LF$ where $L \in n \times k$ and $F \in k \times m$. The factors correspond to the spatial structures across the m base pairs, while the loadings indicate the relative contribution of the factors to each sample.

SPNMF is a special case of the empirical Bayes Poisson NMF in which we model the counts with a link identity function as:

$$x_{ij} \sim Poisson(s_j \sum_{k=1}^{K} l_{jk} f_{ki})$$

Where:

- $s_j = \sum_{i=1}^{m} x_{ij}$ is the total number of counts in sample j for the gene being modeled. This term is used to account for differences in total counts per sample, which can be attributed to differences in gene expression or in sequencing depth.
- $\boldsymbol{l}_k \sim g_{l_k}(\cdot)$ is the vector of loadings, which are modeled with a Gamma prior.
- $\boldsymbol{f}_k = \mu_k$ is the factor that represents the underlying structures of the data. $\mu_k = [\mu_{k1}, \ldots, \mu_{km}]$ is the average count numbers expected at position i contributed by factor k before adjustment with s_j. We use the identity link to ensure that the counts are modeled as the sum of the underlying components, which is the case we expect when modeling the contribution of isoforms to counts in RNA-seq data.

To ensure that μ_{jk} is spatially structured, we introduce a splitting variable b to model the over-dispersion between spatially close samples (nugget effect) as a Gaussian model:

$$\mu_k | \boldsymbol{b}_k \sim \mathcal{N}(\boldsymbol{b}_k, \sigma^2)$$

Where:

- $\boldsymbol{b}_k$ is modeled with a wavelet prior: $\boldsymbol{b} \sim g_{\boldsymbol{b},\text{smooth}}(\cdot)$

The model is fit iteratively by alternating between empirical Bayes matrix factorization (EBMF, citation Wang & Stephens 2021) for the Poisson component of the model, and Gaussian posterior approximation for the Gaussian component.

Obtaining isoforms from Torino factors

To facilitate interpretability, we developed an ad-hoc approach to characterize isoform structures from the factor matrix from Torino's factorization step. For this step, first we must define for each factor the regions that are covered by an exon versus intronic (non-covered) regions. Torino implements an ad-hoc algorithm to transform the factor shapes into isoform structures, by

classifying each position as belonging to an exon or to an intron depending on their relative coverage (see **Algorithm 1** in the Appendix).

Once a factor is binarized, Torino uses a table of coordinates for splice junctions (e.g., a .junc file containing the coordinates for the splice junctions from the BAM files) to match the exons and introns to specific splicing events. Factors with the same intron chain are then tested for differences in the 5' and 3' terminal exons. Factors with the same intron chain and no differences in terminal exons are assigned to the same isoform structure. See **Algorithm 2** in the Appendix for details.

For each gene we obtain $Z \leq K$ isoforms. For each $isoform_z$ we estimate its relative abundance in sample i as:

$$loadings(isoform_{iz}) = \sum_{k \in isoform_z} L_{ik}$$

Characterization of RNA processing events in isoform structures

For each $isoform_z$, we define its exon chain $[(a_1, a_2)_I]$ as defined by the sequence of splice junctions found with **Algorithm 2.** We sought to characterize the RNA processing events that lead to each unique isoform in a gene with the following criteria:

- Cassette exon. An $isoform_z$ is considered to have a cassette exon if its exon chain has three consecutive exons with coordinates $(a_1, a_2), (b_1, b_2), (c_1, c_2)$ and another $isoform_y$ exists such that its exon chain contains two consecutive exons $(x_1, a_2), (c_1, y_2)$.
- Retained intron. An $isoform_z$ is considered to have a retained intron if its exon chain has an exon (a_1, a_2), and there exists another $isoform_y$ whose exon chain contains the coordinates $(b_1, b_2), (c_1, c_2)$ such that $b_2 > a_2$ and $c_1 < a_2$.
- Alternative polyadenylation site. An $isoform_z$ if its last exon has coordinates (a_1, a_2) and there exists another $isoform_y$ whose last exon has coordinates (a_1, a_3) such that $|a_2 - a_3| \geq e$; and $e = \min(\frac{a_1 - a_2}{4}, 200)$.

Running Torino on GTEx data

We selected randomly 100 samples for the following GTEx tissues: Brain Anterior cingulate cortex BA24, Brain Cortex, Brain Frontal Cortex BA9, Brain Putamen basal ganglia, Heart Atrial Appendage, Liver, Lung, Muscle Skeletal, Skin Not Sun Exposed Subprapubic and Whole Blood, procuring to have 50 samples from female donors and 50 samples from male donors on all tissues where it was possible. We downloaded the BAM files directly from the GTEx server and processed them as indicated above to obtain count coverage across the bodies of all protein-coding genes.

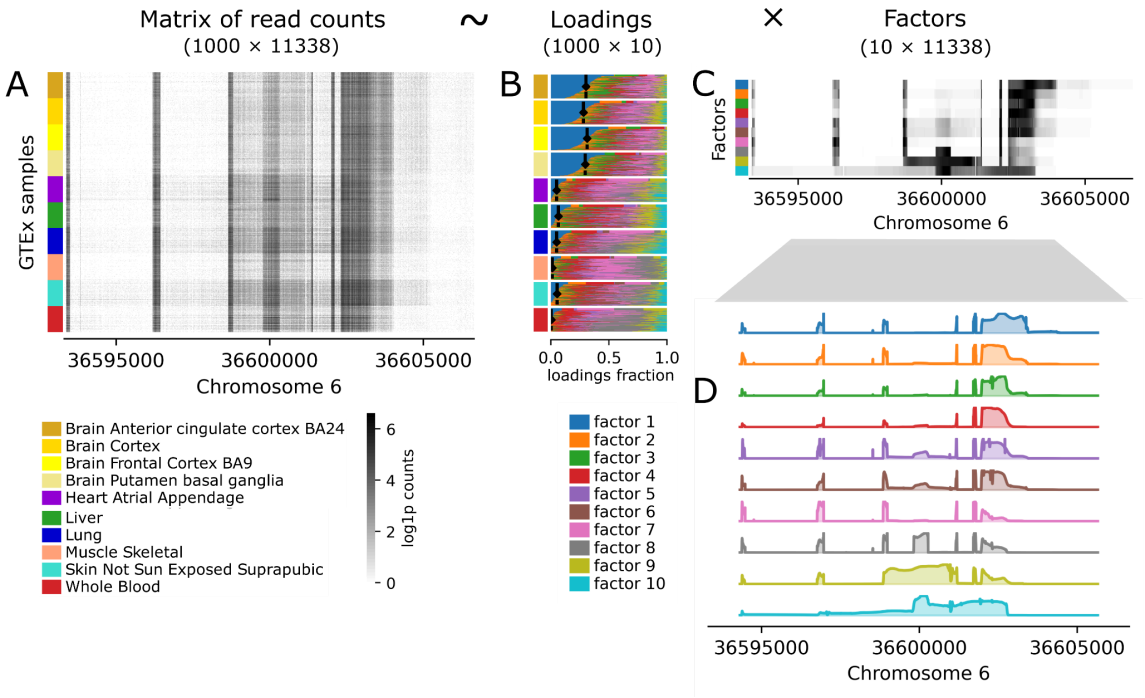


Figure 1. Torino applies matrix factorization to transcriptomic profiles to reveal underlying isoform structures. A) The input for Torino's matrix factorization step is a (N x M) matrix of read counts for each sample per base pair positions across a gene's body. Here we use the read counts of N = 1000 samples from 10 tissues from GTEx, across M = 11338 base pairs on the splicing factor *SRSF3*. The matrix factorization step results in a **B)** (N x K) loadings matrix and a **C)** (K x M) factors matrix. Here K = 10 is a preselected number of factors. **D)** is a different visualization of the factors matrix, which shows the underlying variable structures of the data. Factor 1 shows an isoform with an elongated 3' end exon compared with other factors, and it is enriched in the four brain tissues.

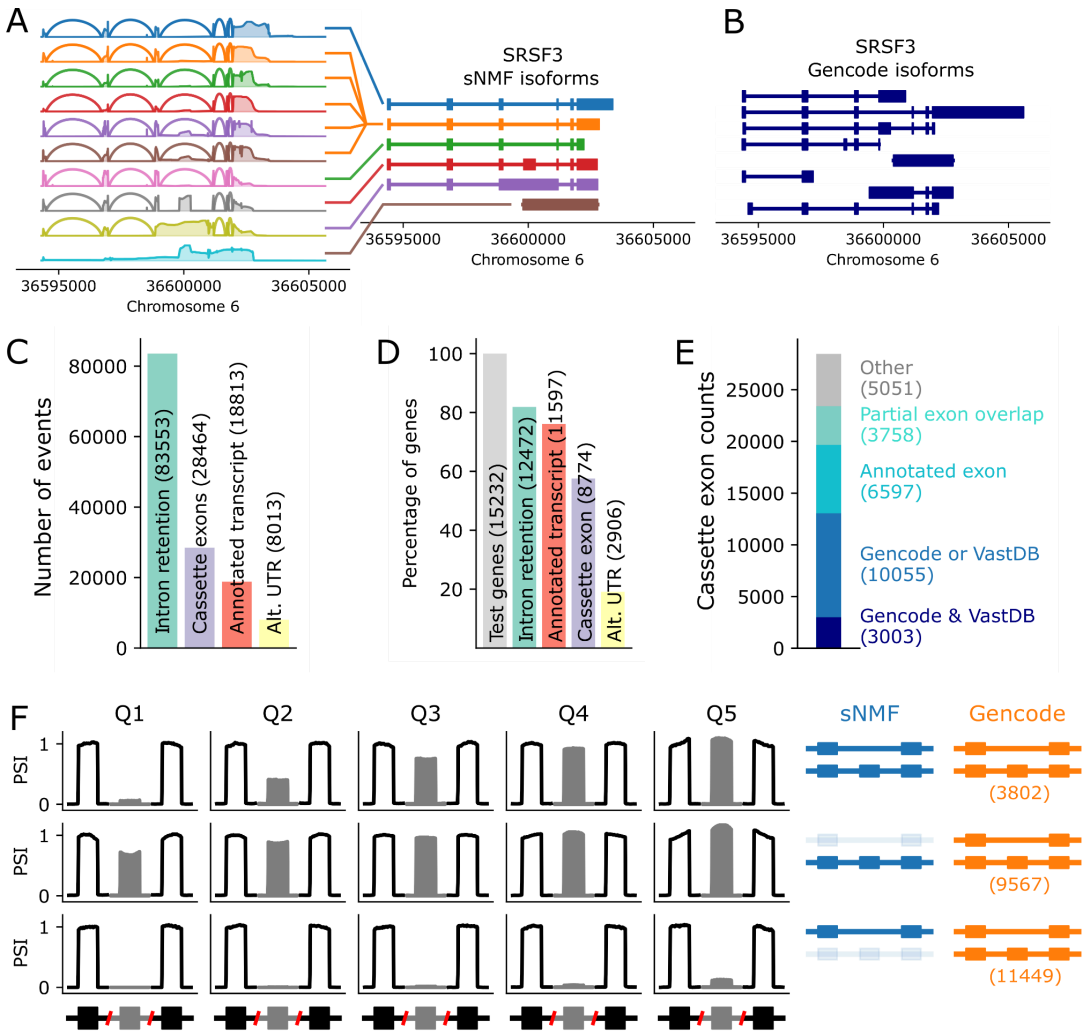


Figure 2. Torino's factors capture RNA processing events. A) Torino matches the factors to splice junctions to infer isoform structures, in this case the splicing factor *SRSF3*. **B)** Gencode annotated isoforms for *SRSF3*. **C)** Total number of events captured by Torino across some important RNA processing event categories. **D)** Percentage of genes presenting the events described in C). **E)** Annotation of the cassette exons in Torino isoforms in Gencode and VastDB. **F)** Metaplots showing the relative coverage of Gencode cassette exons across quintiles if (top) they are cassette exons in Torino's isoform structures, (middle) they are only found as constitutive exons in Torino's isoforms, and (bottom) if they are always skipped in Torino's isoform structures.

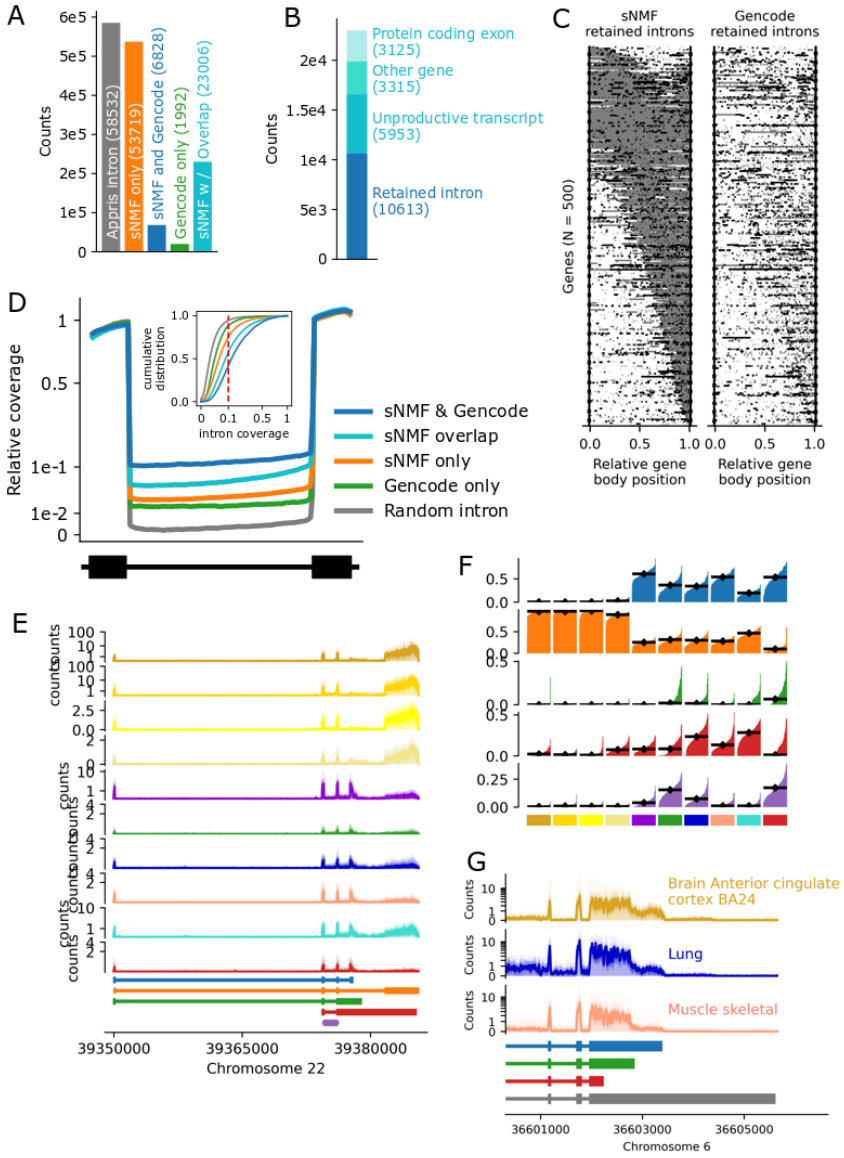


Figure 3. Torino reveals widespread variance in intron retention across the transcriptome. **A)** Number of introns that don't have intron retention or overlap with any exons, compared with intron retentions found by Torino only, by Torino and Gencode, by Gencode only, and by Torino with partial overlaps with annotated Gencode structures. **B)** Annotation of Gencode structures that overlap with retained introns found by Torino, but that do not match exactly to Gencode retained introns. **C)** Distribution of intron retention events found by Torino across 500 random protein coding genes, compared with their corresponding Gencode annotation. **D)** Metaplot showing relative coverage across intron retention events in Liver samples and (in-plot) cumulative distribution of the relative coverage. **E)** Count coverage in the gene SYNGR1 across the 10 GTEx samples used for training Torino, and the 5 isoform structures identified by Torino. Each tissue is represented by 100 samples. **F)** Loadings of the isoform structures across all 1000 GTEx samples. **G)** Count coverage on the 3' exons of SRSF3 across three different GTEx tissues, along with three protein coding Torino isoform structures and Gencode isoforms.

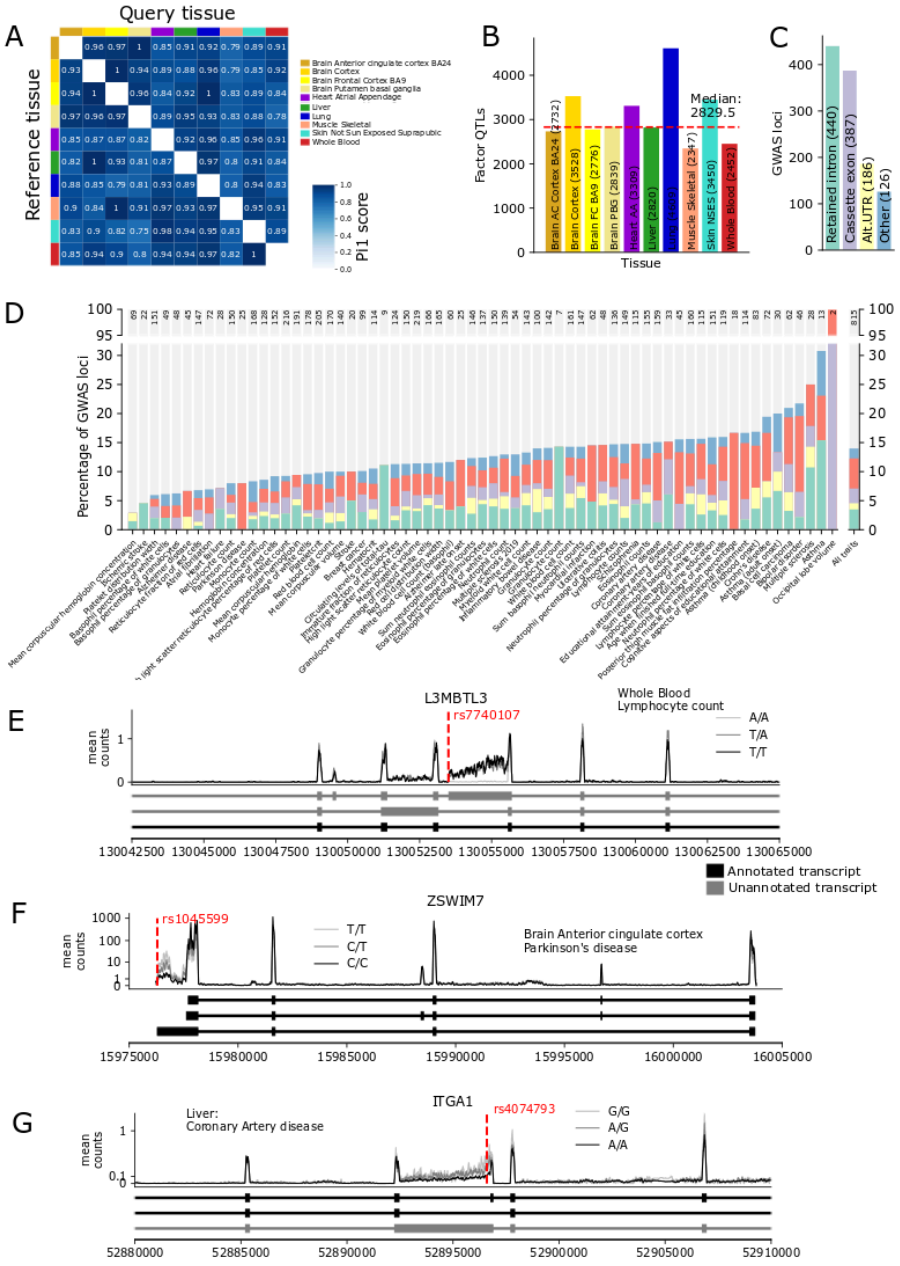


Figure 4. Novel associations between genetic variants, RNA processing events and disease revealed by Torino. A) Pi1 score on QTLs on Torino's factors with an FDR ≤ 1e-4 across 10 GTEx tissues. **B)** Number of significant (FDR ≤ 1e-1) associations between genetic variants and Torino isoform structures per tissue. **C)** Annotation of Torino isoform structures that significantly colocalize with at least one of 65 GWAS traits. **D)** Percentage of GWAS loci that colocalize with at least one significant isoform structure found by Torino. **E)** Variant rs7740107 is associated with changes in lymphocyte counts and it causes a novel cryptic alternative splicing event in *L3MBTL3*. **F)** Variant rs1045599 is associated with Parkinson's disease and it causes differential usage in an APA site in *ZSWIM7*. **G)** Pathogenic variant rs4074793 is associated with coronary artery disease and it causes increased intron retention in *ITGA1*.

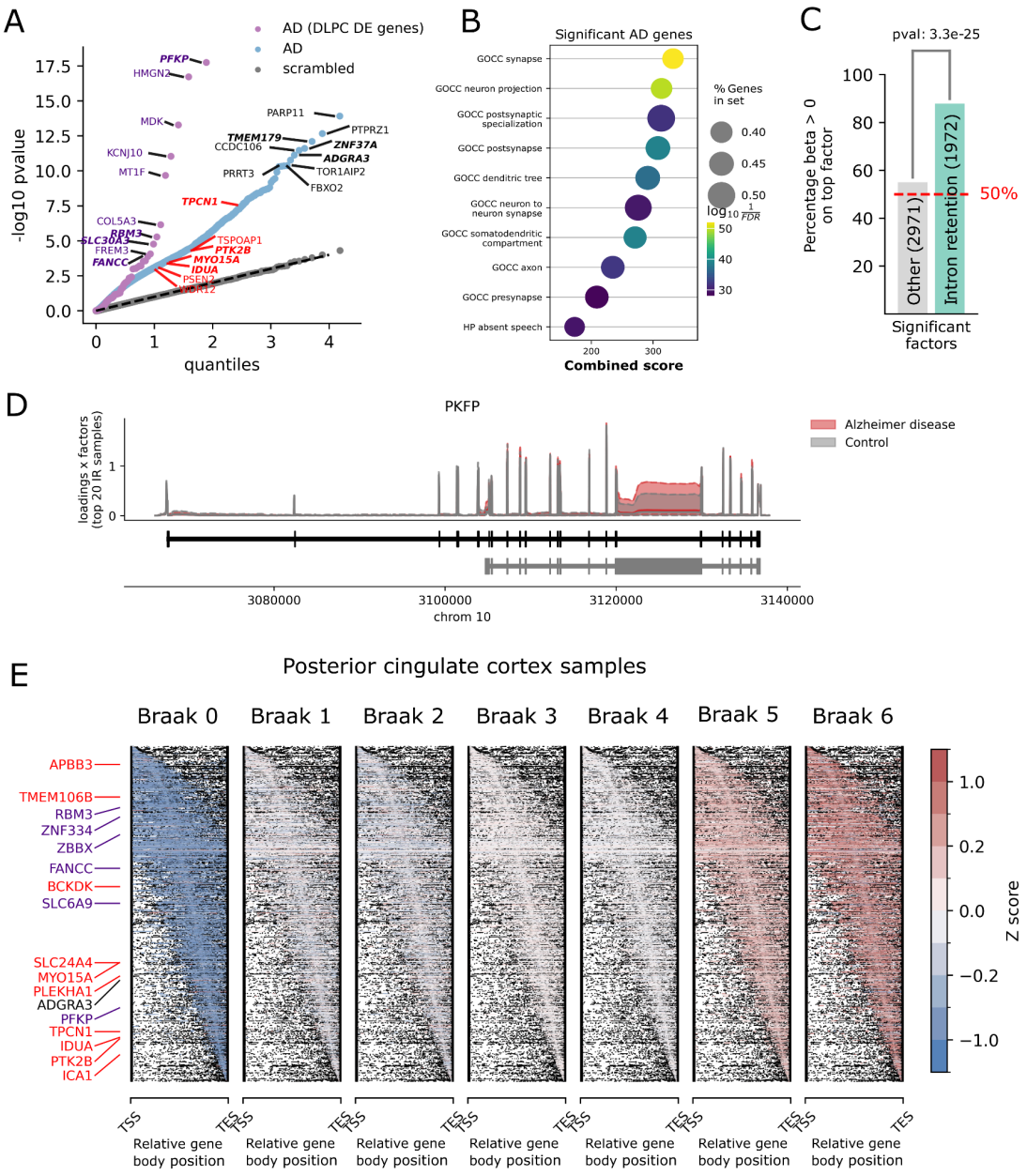
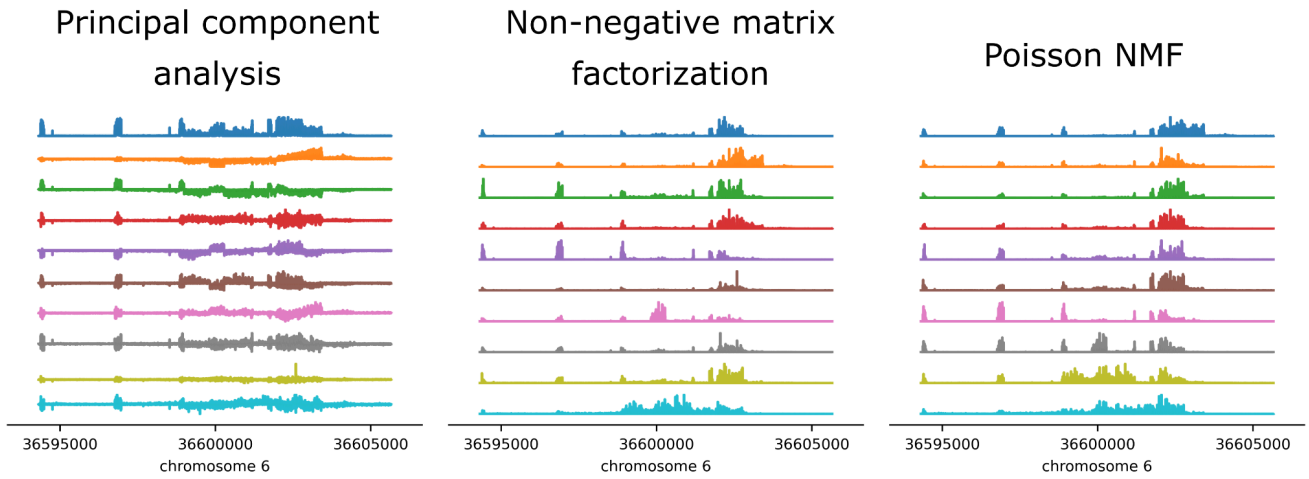


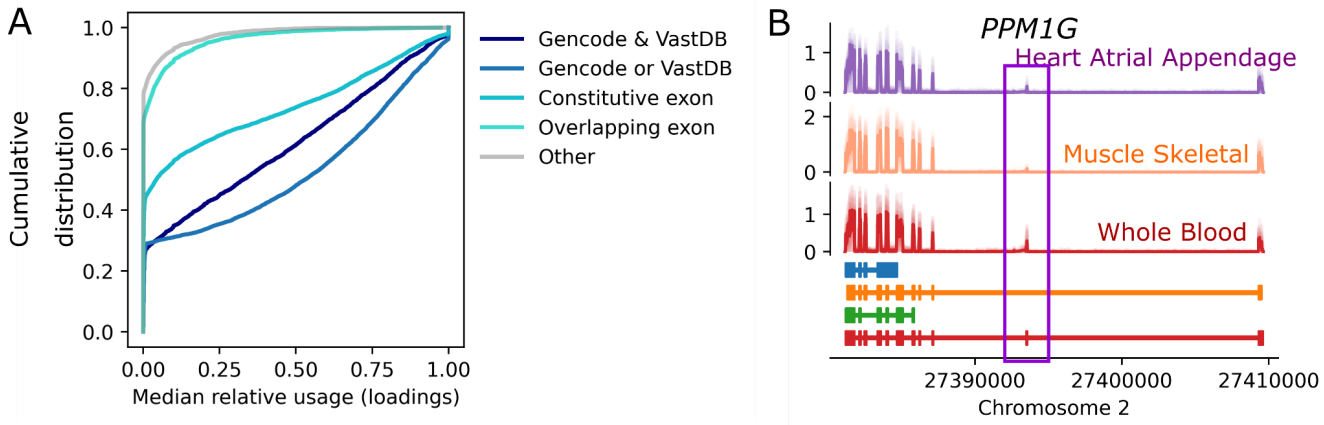
Figure 5. Torino reveals a global increase in intron retention associated with Alzheimer's disease (AD). A) QQ plot showing the association between factors found by Torino and AD through logistic regression, with some of the strongest associated genes labeled. Genes with red labels correspond to some of the top significant genes that are causal in AD. Genes marked in bold and italics have an intron retention event as the strongest or second strongest associated factor with AD. **B)** Gene ontology analysis of genes with significant (FDR ≤ 5e-2) association with AD. **C)** Percentage of positive associations between factor loadings and AD in intron retention factors vs other top significant factors. P-value corresponds to a Chi-squared test. **D)** Modeled coverage (*L*×*F*) of PKFP, showing differences at a novel intron retention event between control and AD samples. Colored area indicates the range of values per category; dashed lines indicate the values for the samples with the top and lowest intron retention values. Solid lines correspond to median values per category. **E)** Distribution of intron retention events found by Torino across genes associated significantly with AD, across 7 Braak stage categories, in PCC samples. The color of each intron corresponds to the average z-score of relative intron coverage for samples in each Braak stage category. Genes labeled in red are significant causal AD genes with significant

intron retention factors. Purple genes are differentially expressed in DLPC samples. Black genes are in the top 10 significant genes, but in neither category.

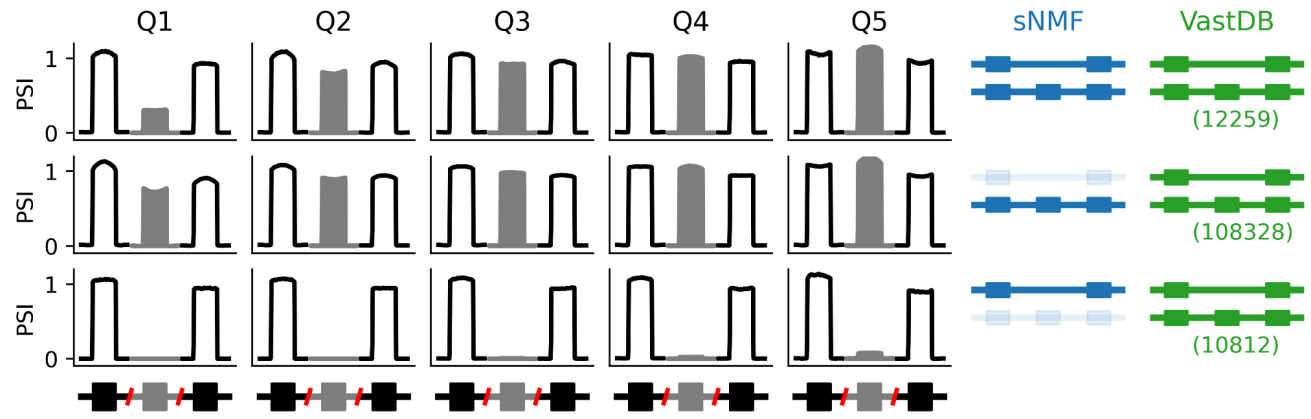
Supplementary Figures



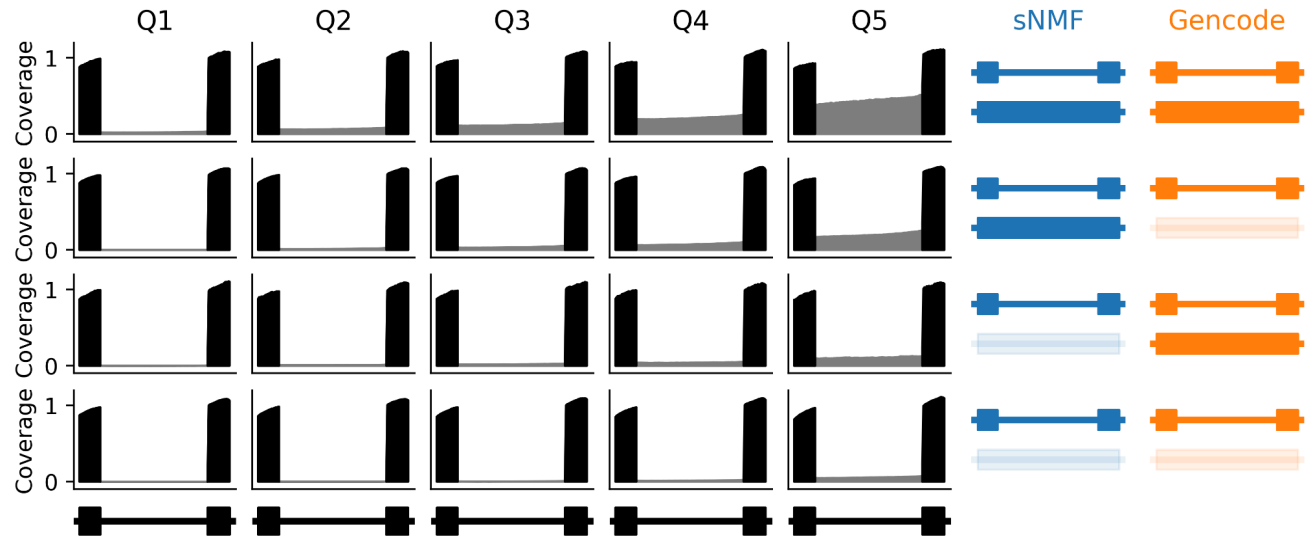
Supplementary figure 1. Decomposition of *SRSF3* counts matrix with three commonly used methods.



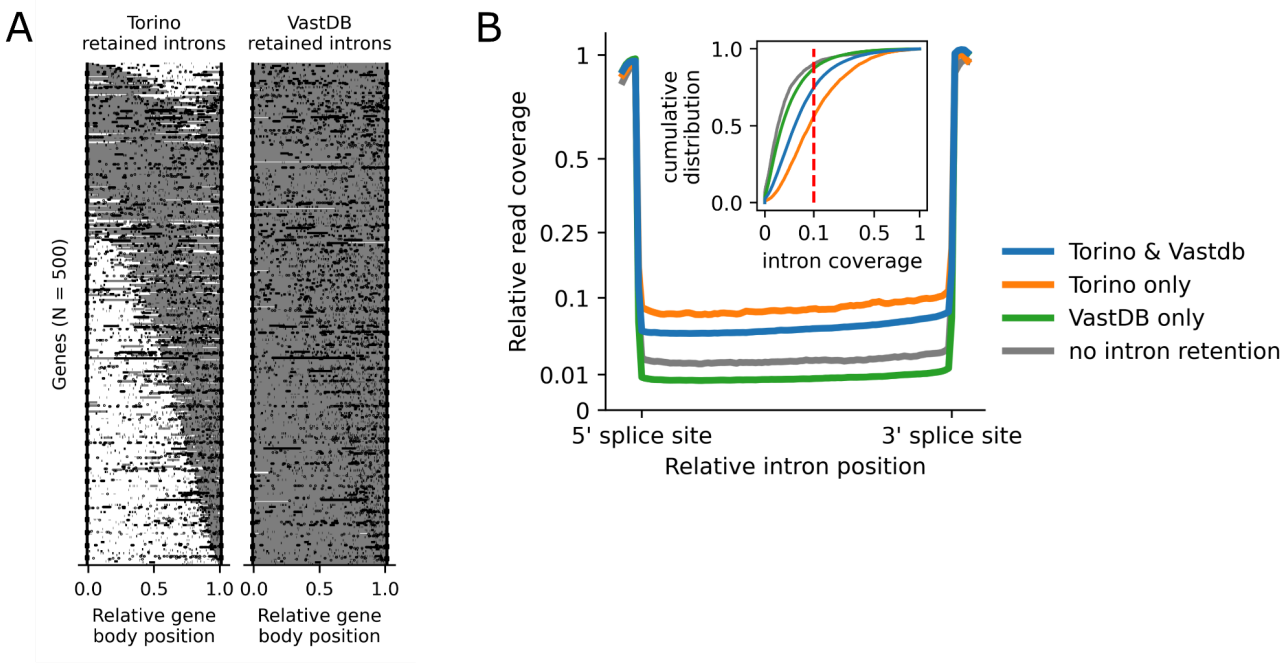
Supplementary figure 2. A) Cumulative distribution of loadings for Torino isoforms with cassette exon. Isoforms with cassette exons that do not overlap with annotated exons and do not match observed splice junctions (Other) have the lowest loadings. **B)** Region annotated as cassette exon in *PPM1G* does not match an annotated exon and is not supported by splice junctions.



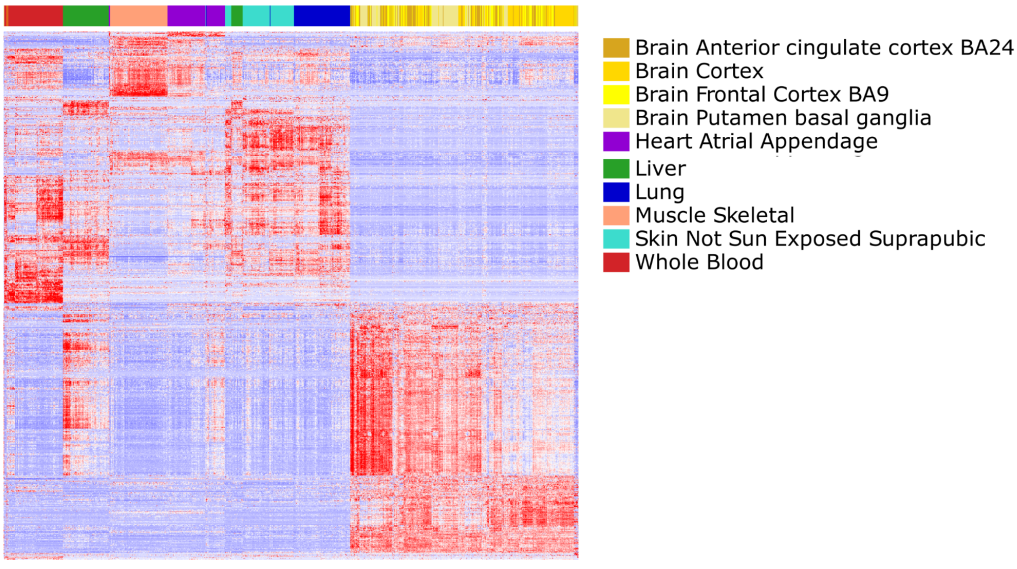
Supplementary Figure 3. Metaplots showing the relative coverage of VastDB cassette exons across quintiles if (top) they are cassette exons in Torino's isoform structures, (middle) they are only found as constitutive exons in Torino's isoforms, and (bottom) if they are always skipped in Torino's isoform structures.



Supplementary Figure 4. Metaplots of relative coverage of introns across quintiles if (top) they are annotated as retained intron by Torino and Gencode, (top-middle) they are only annotated as retained introns by Torino, (bottom-middle) they are only annotated as retained introns by Gencode, and (bottom) if they are not annotated as retained introns by either Torino or Gencode.



Supplementary Figure 5. A) Distribution of intron retention events found by Torino across 500 random protein coding genes, compared with their corresponding VastDB annotation. **B)** Metaplot showing relative coverage across intron retention events in Liver samples and (in-plot) cumulative distribution of the relative coverage.



Supplementary Figure 6. Relative loadings of the top 1000 intron retention factors.

Appendix

Algorithm 1: Binarize factor.

Input:

- $\boldsymbol{f} \leftarrow F_{k,:}$

Initialize parameters:

- $\hat{f} = \boldsymbol{f} / quant_{0.99}(\boldsymbol{f})$
- $c \leftarrow 0.25$
- Initialize $\boldsymbol{f}^b \leftarrow [0, \ldots, 0]$
- `close exon`

Forward pass:

- `for` $i \in 1 : m$:
  - `If exon is closed, then:`
    - `If` $\hat{f}_i \geq c$, `then:`
      - $\boldsymbol{f}_i^b \leftarrow 1$
      - `open exon`
  - `else:`
    - $q \leftarrow \max(\hat{f}$`in exon`$) \times 0.25$
    - `If` $\hat{f}_i \leq \max(q, 0.1)$, `then:`
      - $\boldsymbol{f}_i^b \leftarrow 0$
      - `close exon`
      - $c < -\max(\hat{f}_i, 0.25)$
    - `else:`
      - $f_i^b \leftarrow 1$
- `Reverse to 0 any` $\boldsymbol{f}_i^b$ `with less than 3 consecutive 1s`

Reverse pass:

- `close exon`
- `for` $i \in$ `reverse`$(1 : m-1)$:
  - `If` $\boldsymbol{f}_i^b = 1$, `then:`
    - `open exon`
  - `elif:`$\boldsymbol{f}_i^b = 0$ `AND` $\boldsymbol{f}_{i+1}^b = 1$, `then:`
    - $q \leftarrow \max(\hat{f}$`in exon`$) \times 0.25$
    - `If` $\hat{f}_i \geq \max(q, 0.1)$, `then:`
      - $\boldsymbol{f}_i^b \leftarrow 1$
    - `else:`
      - `close exon`

Return: f^b

References

1. Ule, J. & Blencowe, B. J. Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution. *Mol. Cell* **76**, p329-345 (2019).
2. Fair, B. *et al.* Global impact of unproductive splicing on human gene expression. *Nat. Genet.* **56**, 1851–1861 (2024).
3. Middleton, R. *et al.* IRFinder: assessing the impact of intron retention on mammalian gene expression. *Genome Biol.* **18**, 51 (2017).
4. Mittleman, B. E. *et al.* Alternative polyadenylation mediates genetic regulation of gene expression. *eLife* **9**, e57492 (2020).
5. Li, Y. I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).
6. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
7. Pardo-Palacios, F. J. *et al.* Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nat. Methods* **21**, 1349–1363 (2024).
8. Carbonetto, P., Sarkar, A., Wang, Z. & Stephens, M. Non-negative matrix factorization algorithms greatly improve topic model fits. Preprint at https://doi.org/10.48550/arXiv.2105.13440 (2022).
9. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
10. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
11. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
12. Pozo, F., Rodriguez, J. M., Martínez Gómez, L., Vázquez, J. & Tress, M. L. APPRIS principal isoforms and MANE Select transcripts define reference splice variants. *Bioinformatics* **38**, ii89–ii94 (2022).

13. Tapial, J. *et al.* An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. (2017) doi:10.1101/gr.220962.117.

14. Delaneau, O. *et al.* A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* **8**, 15452 (2017).

15. Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).

16. Li, H.-D. *et al.* Integrative functional genomic analysis of intron retention in human and mouse brain with Alzheimer's disease. *Alzheimers Dement.* **17**, 984–1004 (2021).

17. Shah, A., Mittleman, B. E., Gilad, Y. & Li, Y. I. Benchmarking sequencing methods and tools that facilitate the study of alternative polyadenylation. *Genome Biol.* **22**, 291 (2021).

18. Aragam, K. G. *et al.* Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. *Nat. Genet.* **54**, 1803–1815 (2022).

19. De Jager, P. L. *et al.* A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Sci. Data* **5**, 180142 (2018).

20. Braak, H., Alafuzoff, I., Arzberger, T., Kretzschmar, H. & Del Tredici, K. Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. *Acta Neuropathol. (Berl.)* **112**, 389–404 (2006).