Article Title:

ASTK: a machine learning-based integrative software for alternative splicing analysis

Author names and affiliation:

Shenghui Huang[1 2], Jiangshuang He[1 2], Lei Yu[1 2], Jun Guo[1 2], Zhaoxia Sun[1 2], Linghui Cheng[3], Xing Chen[4], Xiang Ji[5], Yi Zhang[1 2 6 7 8]*

1 Zhejiang Provincial Key Laboratory of Medical Genetics, Key Laboratory of Laboratory Medicine, Ministry of Education, China, School of Laboratory Medicine and Life Science, Wenzhou Medical University, Wenzhou, Zhejiang Province 325035, China

2 Institute of Genomic Medicine, Wenzhou Medical University, Wenzhou, Zhejiang Province 325035, China

3 Scientific Research Center, Wenzhou Medical University, Wenzhou, Zhejiang Province 325035, China

4 School of Informatics, University of Edinburgh, Edinburgh, UK

5 Department of Mathematics, School of Science & Engineering, Tulane University, New Orleans, LA, USA

6 Key Laboratory of Alzheimer's Disease of Zhejiang Province, Institute of Aging, Wenzhou Medical University, Wenzhou, Zhejiang Province 325035, China

7 The Eye-Brain Research Center, State Key Laboratory of Ophthalmology, Optometry and Visual Science, Wenzhou, Zhejiang Province 325027, China

8 Oujiang Laboratory, Zhejiang Lab for Regenerative Medicine, Vision and Brain Health, Wenzhou, Zhejiang Province 325101, China

*Corresponding author: Yi Zhang, Institute of Genomic Medicine, Wenzhou Medical University, Wenzhou, Zhejiang Province 325035, China, Email: yizhang@wmu.edu.cn

Abstract

Alternative splicing (AS) analysis of massive amounts of transcriptomic data is challenging due to limited functions of current software packages. To fill this gap, we present ASTK, a software package covering upstream and downstream analysis of AS. First, we provide a module to perform enrichment analysis at both the gene- and exon-level to incorporate various impacts by different spliced events on a single gene. We further cluster AS genes and spliced exons into three groups based on spliced exon sizes (micro-, mid-, and macro-), which are preferentially associated with distinct biological activities. Second, decoding the regulatory splicing codes has been a daunting task. We demonstrate that ASTK can efficiently extract sequence features and epigenetic marks related to AS events. More significantly, by utilizing machine learning algorithms, we discovered key features that affect the inclusion levels of spliced exons for most AS types. Specifically, the splice site strength is a defining factor to decide exon inclusion levels of alternative 3'/5' splice site. Multiple sequence and epigenetic features act together to control exon inclusion/exclusion for alternative first exons and skipping exon. Our analysis proves that ASTK can facilitate the functional characterization of AS events and improve understanding of splicing regulation.

Key words: Alternative splicing, machine learning, splicing code, functional enrichment, sequence features, epigenetic marks

Introduction

Alternative splicing (AS) is an essential mechanism to increase transcriptomic and proteomic diversity. This process occurs for more than 95% of multi-exonic genes in human genome to generate diverse splice isoforms from a single mRNA [1]. Alternative splicing is critical for normal physiological functions, thus aberrant alternative splicing events give rise to a variety of diseases [2]. Over the past few decades, AS studies have been focusing on two main aspects: functionality and regulation. Until now, researchers usually studied the two aspects separately, due to the lack of proper tools and resources. However, recent advances call for using a more thorough approach to analyze AS events. For this purpose, by employing the most up-to-date algorithms and techniques, we developed a comprehensive software package to address the needs: i) characterizing the distribution patterns of AS events, ii) systematically analyzing functional effects of AS events at the gene- and exon-level respectively, iii) dissecting out splicing-associated sequence features and epigenetic marks, iv) inferring the splicing regulatory code for each type of AS.

Regarding functionality, numerous studies have shown that AS significantly affects biological functions in cellular processes, tissue specificity, developmental states, and disease conditions [1, 3]. Traditionally, AS analysis is centered on genes. It is noteworthy that this gene-central approach may bring many potential pitfalls since not all spliced exons have a functional impact. In addition, a single gene can have multiple AS events simultaneously, which may affect the gene's functions differently. Therefore, we also need to investigate the AS events at the exon level. In ASTK, we provide a module to perform the functional enrichment analysis of AS at both the gene- and exon-level.

Under evolutionary constraints, splicing is tightly regulated to ensure correct forms of the pre-mRNAs. Exon conservation in length and sequence has been implicated in splicing regulation [4]. Statistically, the majority of spliced exons are less than 300

nucleotides (nt) in the human and mouse genomes [5, 6]. Comparative genomics of 76 vertebrate species, including human, demonstrated that exons of less than 250 nt are highly conserved in length. And more importantly, there exists a strong correlation between exon length and exon sequence conservation for these exons [4], suggesting that exon size is an evolutionarily conserved feature. Meanwhile, exons longer than 250 nt are low in length-conservation [4]. In another report, over 13,000 small sized exons (≤ 51 nt), named micro-exons, are identified in the human genome and are highly conserved across species [7]. Recent studies have shown that micro-exons exhibit tissue-specific expression patterns and play important roles in cell differentiation, cell migration, and neuronal and glial functions [8-10]. Mis-regulated micro-exons appear to contribute to the development of neurological disorders and cancer [11]. In this work, we demonstrated that AS events can be clustered into three groups based on the length of spliced exons: micro-exons (≤ 51 nt), mid-sized exons (52-250 nt), and macro-exons (≥ 251 nt). At the gene level, after clustering, we found that the three groups are likely enriched in distinct biological pathways in our case studies. Similar observations were made at the exon level too.

Splicing of a pre-mRNA is catalyzed by the spliceosome, a complex macromolecular machine consisting of five core snRNPs and more than 300 different RNA binding proteins (RBPs) [12]. As a highly regulated process, the decision as to which exon is included or removed depends on complex interactions between RNA sequence elements, splicing regulators, and epigenetic components. Several sequence features, such as the splice site strength, the exon/intron architecture, and GC contents, influence the exon recognition and selection [13, 14]. Additionally, it has been shown that histone modifications, chromatin conformation and accessibility also impact the splicing outcomes [1, 15-20]. Although an increasing number of studies have established the link between sequence and epigenetic features and AS, these studies are typically restricted to test the association of skipping exon (SE) with one type of the features at a time. The regulatory mechanisms in other types of AS are largely

unexplored. Just as important, how sequence and epigenetic features work together to control divergent AS remains elusive. Taking advantage of our ASTK software, we systematically examined the sequence features and epigenetic marks associated with each AS class in our test cases. Intriguingly, we uncovered key features that strongly influence the exon inclusion levels in six classes except for mutually exclusive exons (MX): alternative 3' splice site (A3), alternative 5' splice site (A5), alternative first exons (AF), alternative last exons (AL), retained intron (RI), and SE. Moreover, the contribution of sequence features and epigenetic marks to different AS classes varies dramatically, suggesting the existence of distinct regulatory mechanisms for different types of AS events.

With the development of high-throughput sequencing technologies and the reduction of cost, huge amounts of transcriptomic data accumulated up to now have provided unprecedented resources and challenges to fully characterize AS events. Despite multiple software packages available for AS analysis, most of them focus on only one or few aspects [21-26]. Here, we proposed a new software package, ASTK, with a command-line interface for comprehensive analysis of AS, including clustering AS events based on the size of spliced exons, differential AS analysis, functional enrichment analysis at the gene- and exon-level separately, extracting sequence and epigenetic features, identifying RBP binding motifs, and machine learning-based modeling to rank feature importance and decrypt the splicing codes. We demonstrated the use of ASTK with published multi-omics data. Our results indicate that gene- and exon-level function enrichment of AS are different and can be complementary. In terms of AS regulation, we found that sequence features are the primary determinant to influence exon inclusion levels for most AS classes during embryonic mouse forebrain development. More specifically, the splice site strength alone is a potent factor to determine exon inclusion levels of A3 and A5. Meanwhile, besides sequence features, several histone modifications and open chromatin states also play important roles in exon selection and usage for AF and SE. Therefore, given the dynamic nature

of transcription regulation, ASTK will aid in characterizing the splicing code in a spatial-temporal, condition-specific manner.

Materials and methods

The workflow and components of ASTK are illustrated in Figure 1A. ASTK consists of several modules and the usage of each module is provided in the online manual. Each function in the modules is scripted to provide the command line interface, where the scripts can be conveniently launched and customized by users.

AS events detection, quantification, and differential splicing analysis

As shown in Figure 1A, we developed built-in functions based on SUPPA2 [24] for AS events detection and quantification, with flexibility for samples under different conditions. Also, ASTK can utilize output generated by other programs, including SUPPA2, rMATS [23], and EventPointer [27, 28], for downstream analysis. Differential AS events can be selected using cutoffs of ΔPSI (Percent Spliced-In) and p-values. Furthermore, users have an option to cluster AS genes and spliced exons according to the alternative exon lengths.

AS events distribution pattern analysis

ASTK implements functions to compute statistics and analyze expression patterns of AS events. In the program, barplots are used to depict the occurrence of different types of AS events across experimental conditions, and UpSet plots are used to visualize the common and distinct AS events between conditions. Meantime, principal component analysis (PCA) using percent spliced in (PSI) values is used to visualize variation between samples, useful for quality control. Heatmaps of PSI values can help us quickly identify meaningful changes in AS events between samples. Volcano plots based on ΔPSI values (x-axis) and $-\log_{10}$(p-values) (y-axis) show the AS events with statistical significance.

Functional enrichment analysis for AS events

In ASTK, we implemented three functional enrichment methods, including over-representation analysis (ORA), gene set enrichment analysis (GSEA), and the network-based enrichment method for AS events (NEASE) [22]. ORA and GSEA provide gene-level enrichment analysis for differential AS genes by running clusterProfiler R package [29]. As the ORA method often produces redundant GO terms, we employ simplifyEnrichment R package [30] to simplify enrichment results. Exon-level enrichment analysis is performed by the NEASE method, which allows us to specifically identify the structural impacts of spliced exons and how the protein-protein interactions remodeled [22].

RBPs motif analysis

ASTK performs RBP binding motif enrichment analysis for flanking regions surrounding 5' and 3' splice sites by applying CentriMo algorithm [31] with CISBP-RNA and ATtRACT motif databases [32, 33]. Identified significant binding motifs are displayed as sequence logos. The density curve is the probability of the named motif at each position.

Sequence and epigenetic feature extraction

To crack the splicing codes, we developed multiple functions to extract sequence and epigenetic features of AS events in ASTK. Currently, ASTK supports the computation of exon/intron lengths, GC contents, and splice site strength for all AS types. The GC content can be calculated for user-defined flanking regions of each splice site. Based on MaxEntScan [13] Perl script, we further developed a function to compute splice site strength simultaneously for all splice sites of a given AS event. Epigenetic features including histone modifications and chromatin accessibility can be efficiently extracted by ASTK. To associate a specific epigenetic feature with AS events, the read coverage along the flanking regions surrounding 5' and 3' splice sites is calculated

using mapped reads for each ChIPseq or ATACseq sample. The read alignment BAM files are normalized and converted into bigwig file format with customized bin size.

Feature visualization and statistical comparison

ASTK provides a function for the statistical comparison of individual sequence features between conditions using multiple statistical test methods. The distribution of epigenetic signals in the flanking region is displayed with heatmaps centered on each splice site.

Machine learning modeling and feature importance

ASTK developed a machine learning module based on Scikit-learn API [34], which supports three major machine learning algorithms: random forest, support vector machine, and k-nearest neighbors. To maximize the model performance, ASTK has an option to conduct a grid search for hyperparameter tuning. Cross-validation is used to test the effectiveness of models. Multiple classical metrics, such as accuracy, AUC, recall, precision, and F1 scores, were calculated to evaluate the model performance in ASTK. To identify feature effects for better model interpretability, the feature importance is quantified by Gini importance, an index summing the decrease of impurity caused by each feature across all trees: the higher, the more important the feature.

Availability of software and data

ASTK is developed in Python and is available in the Python Package Index (PyPi), the official third-party software repository for Python packages. ASTK can be installed with a few commands (e.g., using pip, https://pypi.org/project/astk/). ASTK is also packaged in a Docker image to minimize sophisticated configuration of the computing environment and dependencies. The Docker image is freely available on DockerHub at https://hub.docker.com/r/huangshing/astk. We provide a detailed online

instruction at https://huang-sh.github.io/astk-doc/. The description and accession numbers of datasets used in this study are listed in Supplemental Table 1.

Results

Testing ASTK with published datasets

In this section, we used public data to evaluate the performance of ASTK, including multi-omics datasets of C57BL/6 mouse forebrain development (E11.5-P0) downloaded from ENCODE and RNAseq datasets of human embryonic stem cells at seven early neural differentiation time points (0-72 hours) obtained from the NCBI Gene Expression Omnibus (GEO).

AS events distribution patterns

Using E11.5 of C57BL/6 mouse forebrain as the control, the rest time points E12.5, E13.5, E14.5, E15.5, E16.5, and P0 were selected as the treatment groups. Each time point has two biological replicates. AS events were initially identified between the control and each treatment group. Subsequently, differential alternative splicing events were defined using cutoff values of $|\Delta PSI| > 0.1$ and p-value < 0.05. Like PCA on gene expression data (Figure 2A), the PCA plot based on PSI values shows a similar separation pattern of RNAseq samples at different time points (Figure 2B). The heatmap generated from PSI values displays significant changes in AS events across different developmental stages (Figure 2C). Meantime, the volcano plot derived from ΔPSI values allows us to quickly identify statistically significant AS events (Figure 2D). Using the UpSet plot, we can infer the complex association between AS events sets, for example, the overlapping of differential AS events across developmental stages (Figure 2E). Similarly, human embryonic stem cells at 0 h (undifferentiated H1-ESCs) as the control, six different time points of early neural differentiation (3, 6, 12, 24, 48, and 72 h) were used as the treatment groups. AS events were then detected accordingly.

Functional enrichment analysis

Differential AS genes of the SE class during C57BL/6 mouse forebrain development (E11.5 to P0) were chosen as examples for functional enrichment analysis at the gene level. First, current enrichment analysis often produces highly redundant biological terms, which are difficult to interpret and can mask the underlying key biological events. We calculated semantic similarities among enriched GO terms and cluster highly similar GO terms. For example, 538 significant GO terms are clustered into a few major clusters for 761 differential AS genes identified between E11.5 and E16.5 (Figure 3A). However, GO terms sometimes overstate the significance of individual differentially expressed genes that are simply selected using arbitrary cutoffs. Both ORA and GSEA methods can complement each other. In Figure 3B, the GSEA result shows that AS genes with large ΔPSI values between E11.5 and E16.5 are significantly enriched in positive regulation of cell junction assembly and synapse assembly.

Furthermore, before functional enrichment analysis, we attempted to cluster differential AS genes into three groups according to the size of spliced exons. In total, 413 differential AS genes were identified between E11.5 and E12.5, 75 genes with micro-exons spliced, 26 genes with macro-exons spliced, and 336 genes with mid-sized exons spliced. Interestingly, GO enrichment analysis shows that the three groups are largely associated with distinct functions (Figure 3C). Likewise, after clustering, differential AS genes identified between human stem cells 0h and 6h are enriched into different GO terms as well (Figure 3D).

In fact, not all AS events have functional relevance, therefore, identifying the meaningful changes in the protein structure and protein-protein interactions caused by spliced exons will provide important biological insights into individual AS events or isoform switches. Using RNAseq data of human embryonic stem cells at seven early neural differentiation time points (0-72 h) [35], we demonstrated the functional

impact of spliced exons, which are complementary to the gene-level enrichment analysis. We used differential AS events between 0 and 6 hours as an example, Figure 3E shows the results of KEGG enrichment analysis at the exon-level, which displays clear differences for micro-, mid-sized, and macro-exons.

RBP motif enrichment analysis

RBPs broadly regulate RNA transcripts at both transcriptional and post-transcriptional levels, such as splicing, modification, and translation [36, 37]. Regulation of AS by RBPs depends on their binding site positions and preferences along the splice sites. Supplemental Figure 1A shows the sequence logo of a RBP, SRSF3. The binding affinity difference of SRSF3 in the A3 region between the high PSI and low PSI events for human embryonic stem cells at 72 hours is observed (Supplemental Figure 1B).

Specific sequence features and epigenetic marks associated with AS

In eukaryotic cells, the majority of AS events occur co-transcriptionally, which are highly sensitive to sequence features, RNA polymerase II elongation rate, and chromatin modifications and structure [1, 14, 20, 38-40]. Thus, an optimal splicing code requires the assembling of various features that act together coherently at multiple levels.

We analyzed ENCODE RNAseq, ChIPseq, and ATACseq datasets during mouse forebrain development (E12.5-16.5). ChIPseq data contains 8 major histone modifications: H3K4me3, H3K4me2, H3K4me1, H3K27ac, H3K9ac, H3K36me3, H3K27me3, and H3K9me3. Seven types of AS events and the PSI value for each spliced exon were computed: A3’, A5’, AF, AL, MX, RI, and SE (Figure 1B). Sequence features, ATAC signals, and ChIPseq signals of histone modifications were extracted accordingly. Our analyses cover all seven AS classes and span five developmental time points.

The sequence context around the splice sites from both constitutive and alternative exons is derived to examine potential patterns associated with AS. Upstream and downstream sequences from each exon junction are extracted: +/- 150 nt flanking region surrounding 5' splice site, +/- 150 nt flanking region surrounding 3' splice site. After removing the conserved 5' and 3' splice site sequences, the GC content in each flanking region is calculated. Lengths of related exons and introns for AS events are computed.

Sequence features are part of genomic architecture, contributing to splice site and splicing unit recognition and affecting the splicing process [41, 42]. The GC content is an important aspect of genomic organization and function. It is known that higher GC content has higher stability for DNA and secondary structures of mRNAs. Our results show that flanking regions around alternative splice sites in AF and SE classes have a higher GC content in the high PSI group compared to its corresponding low PSI group, suggesting the alternative splice site with a higher GC content likely has a higher chance of being selected (Figure 4A-B).

Besides that, we examined how the lengths of introns and exons affect splicing. In AF, AL, and SE events, larger alternative exons tend to be spliced in (Figure 4C, D, F). On the contrary, for RI, the lengths of retained introns vary dramatically in both PSI high and low groups. However, the upstream and downstream exons of a retained intron appear to be significantly larger in the PSI high group (Figure 4E).

Selection of alternative splice sites is inherently competitive and depends on a number of factors. Splice site strength is one of them. We observed that the splice site strengths are significantly different between the PSI high versus its corresponding low groups for all AS classes except MX. In general, we found that the splice site strength is positively correlated with exon inclusion levels of 5 types of AS events except MX

and RI (Figure 4G-J, M). For instance, for SE events, 5' and 3' splice sites flanking the skipped exons in the high PSI group appear to have significantly higher splice site strength than its corresponding low PSI group. Since PSI refers to the percent spliced-in value, our results suggest that given two 5' or 3' splice sites, the higher scoring ones have a higher probability of being selected (Figure 4M). Interestingly, for RI, the PSI high group has lower strength in both 5' and 3' splice sites flanking the retained intron than its corresponding PSI low group (Figure 4L), indicating that these retained introns probably have weak splice sites flanked so that RNA polymerase II can read through and include the intron in the mRNA.

Lastly, we assessed the association of epigenetic features with AS events. For ATAC signal and each histone modification, we summarized the ChIPseq signal in the ± 150 nt regions of each alternative splice site and its paired constitutive splice site. ATACseq is a popular method used to measure genome-wide chromatin accessibility. ATAC signal abundance is highly indicative of active transcription. We found that increased ATAC signal is significantly associated with the alternative first exon selection in AF events (Figure 5A). Moreover, it has been well known that histone modifications change chromatin architecture and regulate gene expression [43]. Recent findings suggest histone modifications directly act on the regulated exon and its vicinity to affect splicing outcomes [18], likely through a combinatorial manner [44]. Here, we examined the ChIPseq signal distributions of 8 histone marks in the adjacent regions of splice sites across E12.5 to E16.5. Based on distribution patterns, the enrichment of histone marks varies significantly in different AS classes across time points. Using E16.5 as an example, 7 out of the 8 histone marks exhibit a significant difference between the PSI high versus low groups in AF (Figure 5B-I). Furthermore, in AF events, H3K4me2, H3K4me3, H3K9ac and H3K27ac tend to favor the inclusion of an alternative first exon. In contrast, H3K4me1, H3K27me3 and H3K36me3 appear to have the opposite effect, i.e., promoting exon exclusion. Meanwhile, no clear association was observed for H3K9me3 (Figure 5F). In SE

events, H3K36me3 positively correlates with the PSI high groups (Figure 5R). However, H3K4me1, H3K4me2, and H3K27me3 appear to have negative correlations (Figure 5K-L, Q).

Model predictions and feature importance

To elucidate how sequence and epigenetic features contribute to different types of AS events, we constructed a random forest-based binary classifier to identify features important for AS regulation. Both sequence and epigenetic features are used as the input variables in our random forest model to classify each AS type into PSI high versus PSI low groups. The model performance was evaluated using accuracy and AUC (area under the ROC curve) based on 5-fold cross validation.

As illustrated in Table 1, when predicting PSI high versus low using combined sequence features, the model accuracy varied significantly for seven AS classes across different developmental stages: A3' (0.75-0.77), AF (0.74-0.76), A5' (0.71-0.72), RI (0.69-0.70), SE (0.68-0.69), AL (0.66-0.70), MX (0.44-0.56). At the same time, we also measured the AUC score, a solution to any potential imbalanced data. AUC scores are in line with accuracy values, A3' (0.84-0.85), AF (0.82-0.83), A5' (0.78-0.79), RI (0.76-0.77), SE (0.74-0.76), AL (0.70-0.74), MX (0.47-0.59). Overall, sequence features alone, only including the GC content around splice sites, the lengths of spliced exons and their adjacent introns, and the splice site strength, offer a decent predictive ability for different AS classes except for MX.

Subsequently, we investigated the effects of eight histone modifications and chromatin accessibility on AS. We found that epigenetic factors significantly impact how the first alternative exon is selected and how the alternative exon is skipped (Table 2). The prediction accuracy on PSI high versus low groups using these epigenetic features is approximately 0.77 and 0.59 for AF and SE, respectively. Taking one step further, we ran the model with pooled sequence and epigenetic

features. It turns out that adding all the features together only modestly increased the prediction accuracy for AF (Table 3).

To reveal the regulatory mechanisms of AS, we were in search of features with major contributions to the classification. In our random forest model, variable importance is quantified by Gini importance. Consequently, using E16.5 as an example, we obtained a ranked feature importance list for every AS class. One of the major findings is that seven AS classes exhibit distinctive regulatory features. Since the accuracy of MX is fluctuating around 50%, a random guessing level, we focused on the other six types of AS events. The results indicate that splice site strength is the single most important factor to determine PSI high versus low for A3 and A5 classes. Contrariwise, multiple sequence and epigenetic features significantly contribute to the classification in AF, including GC contents, the lengths of introns and exons, splice site strength, ATAC signal, H3K4me3, H3K9ac, H3K27ac, and H3K4me2. Like AF, sequence features and specific epigenetic features are strongly associated with SE classification. In AL, only two top features were identified: the lengths of introns and exons, and splice site strength. Moreover, all three sequence features significantly influence the classification in RI (Table 4).

Comparison with other software

Many software packages have been developed to support AS analysis. Several programs are widely used for the detection and quantification of RNA splicing: rMATS, SUPPA2, vast-tools [45], and EventPointer. Software packages that can perform AS functional enrichment analysis include ExonOntology [46] and NEASE. rMAPS2 is designed to identify the spatial distribution of RBPs binding motifs [47]. Matt [48] and SpliceTools can extract multiple sequence features as well. In addition, SpliceTools provides summary statistics and functional significance of AS changes. We compared our ASTK with these popular tools (Figure 8). Regarding the analysis modules, our software essentially covers almost all the function modules implemented

in other software. Importantly, ASTK also includes extra modules to extend its functionality. The major differences between ASTK and other toolkits are clustering AS genes and spliced exons into three categories according to the exon size, providing functional enrichment analysis at both the gene- and exon-level, extracting epigenetic features, and decoding splicing codes with machine learning algorithms. Collectively, ASTK is more comprehensive in AS analysis and more user-friendly.

Discussion

Rapid development in sequencing technologies enables large-scale transcriptomic analysis. In the past few years, one of the most prominent applications of RNAseq is gnome-wide profiling of alternative splicing. For example, comprehensive databases for developmentally related (MeDAS) and cancer associated AS (ASCancer) have been created recently [49, 50]. As a fundamental regulatory process of gene expression, AS maintains normal cell physiology and cellular behavior. On the other hand, defects in AS have been reported to be associated with human diseases and cancer [3]. Approximately 20% of human disease-associated mutations affect splicing [51].

Thus, studies of AS will fundamentally advance our understanding of gene regulation and disease etiology and assist the development of AS-targeting clinical therapies. In this study, we aim to develop a comprehensive pipeline to fully characterize the functional and regulatory aspects of AS. Even though various computational tools have been developed to quantify transcript isoforms and identify splicing events [21-25], functional characterization of AS events has proven to be challenging. Gene-level enrichment analyses such as GO, KEGG, and GSEA generally focus on differential AS genes. The underlying assumption of such analysis is that all AS events will affect the gene equally. However, a single gene may have multiple AS events and splicing patterns. Not all AS events have a functional impact. In our software package, we integrated the NEASE method to specifically address the functional consequences at

the exon level. Additionally, previous work has found that spliced exons less than 250 nt are significantly conserved in terms of length and sequence across vertebrate species [4, 11, 52, 53]. We clustered differential AS genes and exons into three groups based on the length of the corresponding spliced exons: ≤ 51 nt, 52-250 nt, and ≥ 251 nt. It appears that each group is enriched in different functions in our case studies.

AS is a hallmark of the complexity and diversity of gene regulation. During the process, genomic architecture such as splice site strength, GC contents, and exon-intron structure defines the splicing unit. In addition, epigenetic features including histone modifications and chromatin structures regulate exon splicing by recruiting splice machinery and modulating RNA polymerase II elongation rate [17, 18, 39, 40, 44, 54, 55]. Eventually, splicing factors occupy their specific binding motifs along splice sites to assist intron folding and splicing [1, 15]. A recent study shows that AS genes shared with specific histone marks tend to have similar RBP binding motifs and enrich in common functional pathways [44], suggesting a strong functional coupling between epigenomic features and coordinately regulated cell-specific alternative splicing. Given complex genomic structures, numerous histone modifications and splicing regulators, it will be computationally intensive to find the specific splicing-associated chromatin signatures and regulatory features. Therefore, we created modules in our ASTK to efficiently identify sequence features, RBP binding motifs, and extract enriched histone marks and chromatin states for splice sites and their flanking regions.

More importantly, we applied cutting-edge machine learning algorithms to decipher the splicing regulatory codes for different AS classes during embryonic mouse forebrain development. First, based on the PSI values, AS events were classified as PSI high versus low groups, which were treated as output variables in our predictive model. Sequence and epigenetic features were used as input. We found sequence features to be most predictive for 6 out of 7 types of AS except for MX. Splice site

strength stood out as a strong individual predictor for A3 and A5. Interestingly, besides sequence features, ATAC signal, H3K4me3, and H3K9ac are strong individual predictors for AF. Taken together, different AS classes may have distinct splicing regulatory mechanisms, in which sequence and epigenetic features likely have various degrees of contributions.

To our knowledge, ASTK is the first platform integrated with machine learning for comprehensive upstream and downstream analysis of AS events. Through testing of experimental multi-omics datasets, we showed that ASTK can systematically and effectively address the needs. Overall, our ASTK software will facilitate the functional studies of AS, as well as provide mechanistic insights into how AS is regulated and evolved.

Supplementary data

Supplementary data are available at *Nucleic Acids Research* online.

Acknowledgements

We thank Dr. Jianming Zeng (University of Macau), and all the members of his bioinformatics team, biotrainee, for generously sharing their experience.

Funding

This work was supported by grants from the National Natural Science Foundation of China (No. 31771165, No.81971064 to Y. Z.), and funding from Wenzhou Medical University.

Conflict of interest

The authors declare no conflict of interest.

References:

1. Lee, Y. and D.C. Rio, *Mechanisms and Regulation of Alternative Pre-mRNA Splicing.* Annu Rev Biochem, 2015. **84**: p. 291-323.
2. Fredericks, A.M., et al., *RNA-Binding Proteins: Splicing Factors and Disease.* Biomolecules, 2015. **5**(2): p. 893-909.
3. Kim, H.K., et al., *Alternative splicing isoforms in health and disease.* Pflugers Arch, 2018. **470**(7): p. 995-1016.
4. Movassat, M., et al., *Exon size and sequence conservation improves identification of splice-altering nucleotides.* RNA, 2019. **25**(12): p. 1793-1805.
5. Song, S., et al., *Comparative component analysis of exons with different splicing frequencies.* PLoS One, 2009. **4**(4): p. e5387.
6. Gudlaugsdottir, S., et al., *Exon size distribution and the origin of introns.* Genetica, 2007. **131**(3): p. 299-306.
7. Li, Y.I., et al., *RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts.* Genome Res, 2015. **25**(1): p. 1-13.
8. Lee, J., et al., *QUAKING Regulates Microexon Alternative Splicing of the Rho GTPase Pathway and Controls Microglia Homeostasis.* Cell Rep, 2020. **33**(13): p. 108560.
9. Johnson, V., H.J. Junge, and Z. Chen, *Temporal regulation of axonal repulsion by alternative splicing of a conserved microexon in mammalian Robo1 and Robo2.* Elife, 2019. **8**.
10. Ratcliffe, C.D.H., et al., *HGF-induced migration depends on the PI(3,4,5)P3-binding microexon-spliced variant of the Arf6 exchange factor cytohesin-1.* J Cell Biol, 2019. **218**(1): p. 285-298.
11. Lee, J.S., N. Lamarche-Vane, and S. Richard, *Microexon alternative splicing of small GTPase regulators: Implication in central nervous system diseases.* Wiley Interdiscip Rev RNA, 2021: p. e1678.
12. Jurica, M.S., et al., *Purification and characterization of native spliceosomes suitable for three-dimensional structural analysis.* RNA, 2002. **8**(4): p. 426-39.
13. Yeo, G. and C.B. Burge, *Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals.* J Comput Biol, 2004. **11**(2-3): p. 377-94.
14. Tammer, L., et al., *Gene architecture directs splicing outcome in separate nuclear spatial regions.* Mol Cell, 2022. **82**(5): p. 1021-1034 e8.
15. Shi, Y., *Mechanistic insights into precursor messenger RNA splicing by the spliceosome.* Nat Rev Mol Cell Biol, 2017. **18**(11): p. 655-670.
16. Corley, M., M.C. Burns, and G.W. Yeo, *How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms.* Mol Cell, 2020. **78**(1): p. 9-29.
17. Zhou, H.L., et al., *Regulation of alternative splicing by local histone modifications: potential roles for RNA-guided mechanisms.* Nucleic Acids Res, 2014. **42**(2): p. 701-13.
18. Luco, R.F., et al., *Regulation of alternative splicing by histone modifications.* Science, 2010. **327**(5968): p. 996-1000.
19. Ullah, F., et al., *Exploring the relationship between intron retention and chromatin accessibility in plants.* Bmc Genomics, 2018. **19**.

20. Petrova, V., et al., *Increased chromatin accessibility facilitates intron retention in specific cell differentiation states.* Nucleic Acids Res, 2022. **50**(20): p. 11563-11579.
21. Katz, Y., et al., *Analysis and design of RNA sequencing experiments for identifying isoform regulation.* Nat Methods, 2010. **7**(12): p. 1009-15.
22. Louadi, Z., et al., *Functional enrichment of alternative splicing events with NEASE reveals insights into tissue identity and diseases.* Genome Biol, 2021. **22**(1): p. 327.
23. Shen, S., et al., *rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data.* Proc Natl Acad Sci U S A, 2014. **111**(51): p. E5593-601.
24. Trincado, J.L., et al., *SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions.* Genome Biol, 2018. **19**(1): p. 40.
25. Li, Y.I., et al., *Annotation-free quantification of RNA splicing using LeafCutter.* Nat Genet, 2018. **50**(1): p. 151-158.
26. Flemington, E.K., et al., *SpliceTools, a suite of downstream RNA splicing analysis tools to investigate mechanisms and impact of alternative splicing.* Nucleic Acids Res, 2023.
27. Romero, J.P., et al., *EventPointer: an effective identification of alternative splicing events using junction arrays.* BMC Genomics, 2016. **17**: p. 467.
28. Ferrer-Bonsoms, J.A., et al., *EventPointer 3.0: flexible and accurate splicing analysis that includes studying the differential usage of protein-domains.* NAR Genom Bioinform, 2022. **4**(3): p. lqac067.
29. Yu, G., et al., *clusterProfiler: an R package for comparing biological themes among gene clusters.* OMICS, 2012. **16**(5): p. 284-7.
30. Gu, Z. and D. Hubschmann, *Simplify enrichment: A bioconductor package for clustering and visualizing functional enrichment results.* Genomics Proteomics Bioinformatics, 2022.
31. Bailey, T.L. and P. Machanick, *Inferring direct DNA binding from ChIP-seq.* Nucleic Acids Res, 2012. **40**(17): p. e128.
32. Giudice, G., et al., *ATtRACT-a database of RNA-binding proteins and associated motifs.* Database (Oxford), 2016. **2016**.
33. Ray, D., et al., *A compendium of RNA-binding motifs for decoding gene regulation.* Nature, 2013. **499**(7457): p. 172-7.
34. Buitinck, L., et al., *API design for machine learning software: experiences from the scikit-learn project.* In ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 2013.
35. Inoue, F., et al., *Identification and Massively Parallel Characterization of Regulatory Elements Driving Neural Induction.* Cell Stem Cell, 2019. **25**(5): p. 713-727 e10.
36. Hentze, M.W., et al., *A brave new world of RNA-binding proteins.* Nat Rev Mol Cell Biol, 2018. **19**(5): p. 327-341.
37. Carazo, F., J.P. Romero, and A. Rubio, *Upstream analysis of alternative splicing: a review of computational approaches to predict context-dependent splicing factors.* Brief Bioinform, 2019. **20**(4): p. 1358-1375.
38. Barash, Y., et al., *Deciphering the splicing code.* Nature, 2010. **465**(7294): p. 53-9.
39. Hu, Q., et al., *Histone posttranslational modifications predict specific alternative exon subtypes in mammalian brain.* PLoS Comput Biol, 2017. **13**(6): p. e1005602.

40. Hu, Q., C.S. Greene, and E.A. Heller, *Specific histone modifications associate with alternative exon selection during mammalian development.* Nucleic Acids Res, 2020. **48**(9): p. 4709-4724.
41. Sterner, D.A., T. Carlo, and S.M. Berget, *Architectural limits on split genes.* Proc Natl Acad Sci U S A, 1996. **93**(26): p. 15081-5.
42. Fox-Walsh, K.L., et al., *The architecture of pre-mRNAs affects mechanisms of splice-site pairing.* Proc Natl Acad Sci U S A, 2005. **102**(45): p. 16176-81.
43. Stillman, B., *Histone Modifications: Insights into Their Influence on Gene Expression.* Cell, 2018. **175**(1): p. 6-9.
44. Agirre, E., et al., *Splicing-associated chromatin signatures: a combinatorial and position-dependent role for histone marks in splicing definition.* Nat Commun, 2021. **12**(1): p. 682.
45. 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.* Genome Res, 2017. **27**(10): p. 1759-1768.
46. Tranchevent, L.C., et al., *Identification of protein features encoded by alternative exons using Exon Ontology.* Genome Res, 2017. **27**(6): p. 1087-1097.
47. Hwang, J.Y., et al., *rMAPS2: an update of the RNA map analysis and plotting server for alternative splicing regulation.* Nucleic Acids Res, 2020. **48**(W1): p. W300-W306.
48. Gohr, A. and M. Irimia, *Matt: Unix tools for alternative splicing analysis.* Bioinformatics, 2019. **35**(1): p. 130-132.
49. Li, Z., et al., *MeDAS: a Metazoan Developmental Alternative Splicing database.* Nucleic Acids Res, 2021. **49**(D1): p. D144-D150.
50. Wu, S., et al., *ASCancer Atlas: a comprehensive knowledgebase of alternative splicing in human cancers.* Nucleic Acids Res, 2022.
51. Lim, K.H., et al., *Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes.* Proc Natl Acad Sci U S A, 2011. **108**(27): p. 11093-8.
52. Ustianenko, D., S.M. Weyn-Vanhentenryck, and C. Zhang, *Microexons: discovery, regulation, and function.* Wiley Interdiscip Rev RNA, 2017. **8**(4).
53. Irimia, M., et al., *A highly conserved program of neuronal microexons is misregulated in autistic brains.* Cell, 2014. **159**(7): p. 1511-23.
54. Li, T., et al., *Integrative analysis reveals functional and regulatory roles of H3K79me2 in mediating alternative splicing.* Genome Med, 2018. **10**(1): p. 30.
55. Spies, N., et al., *Biased chromatin signatures around polyadenylation sites and exons.* Mol Cell, 2009. **36**(2): p. 245-54.

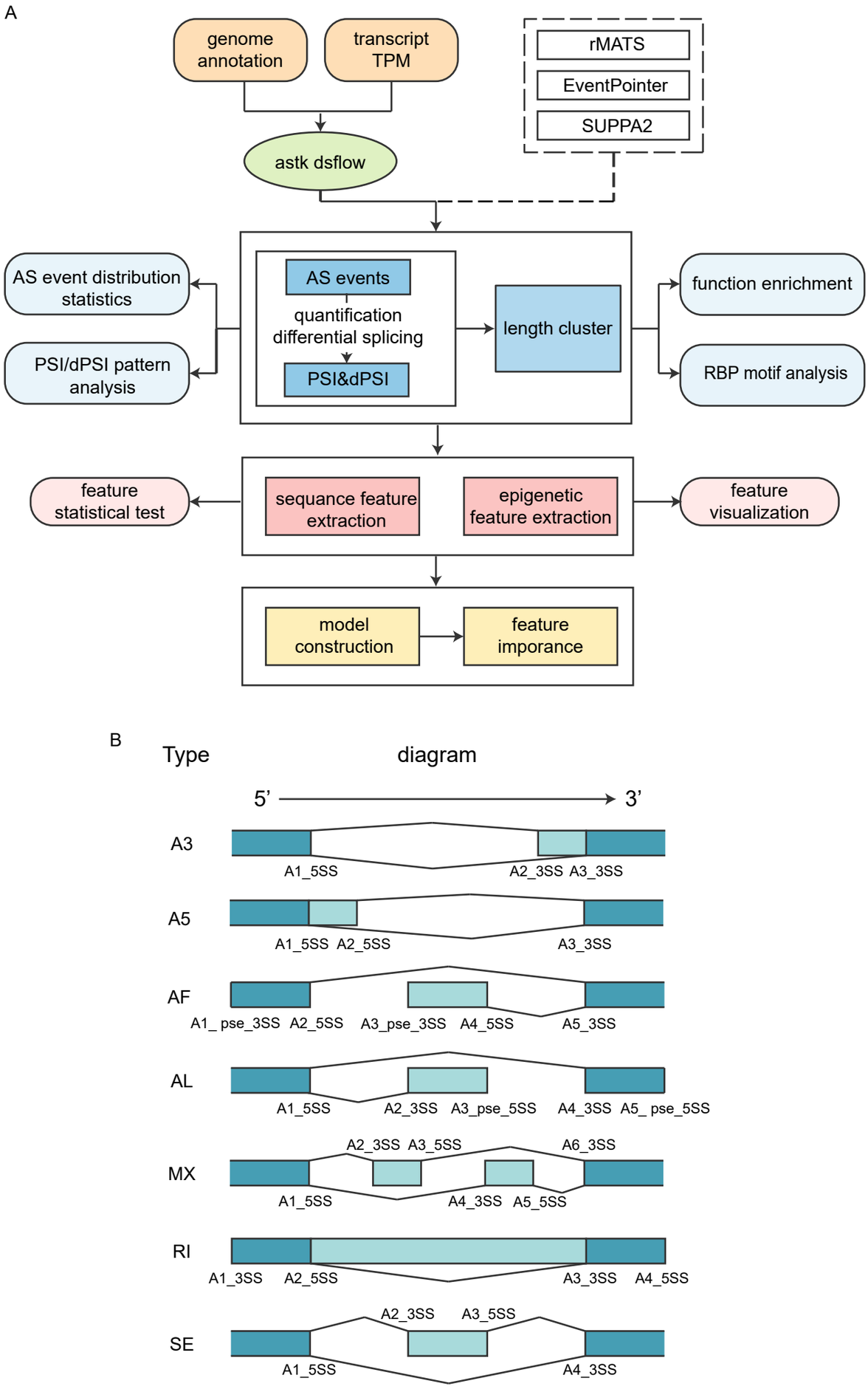


Figure 1 ASTK workflow

A. Schematic illustration of the workflow

B. Diagram of seven AS types: alternative 3' splice site (A3), alternative 5' splice site (A5), alternative first exons (AF), alternative last exons (AL), mutually exclusive exons (MX), retained intron (RI), and skipping exon (SE). "A1" "A2", "A3", "A4", and "A5" denote the order of splice sites in AS events. "5SS" and "3SS" denote the 5' and 3' splice site, respectively.

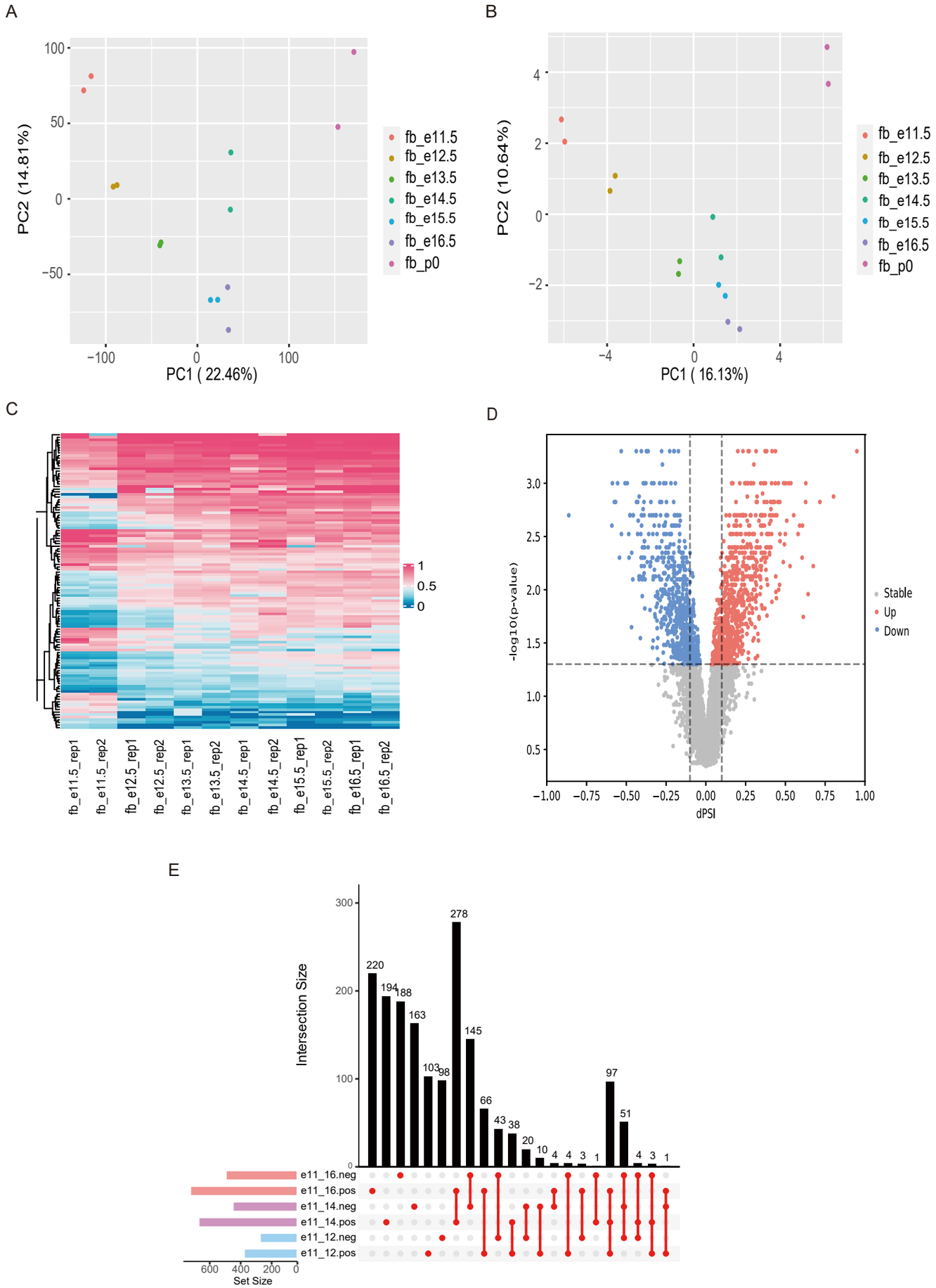


Figure 2 AS events distribution during embryonic mouse forebrain development: E11.5 - P0

A. PCA analysis based on gene expression data

B. PCA analysis based on PSI values

C. Heatmap on PSI values

D. Volcano plot on ΔPSI values

E. Upset plot for visualizing the intersection of different AS events sets

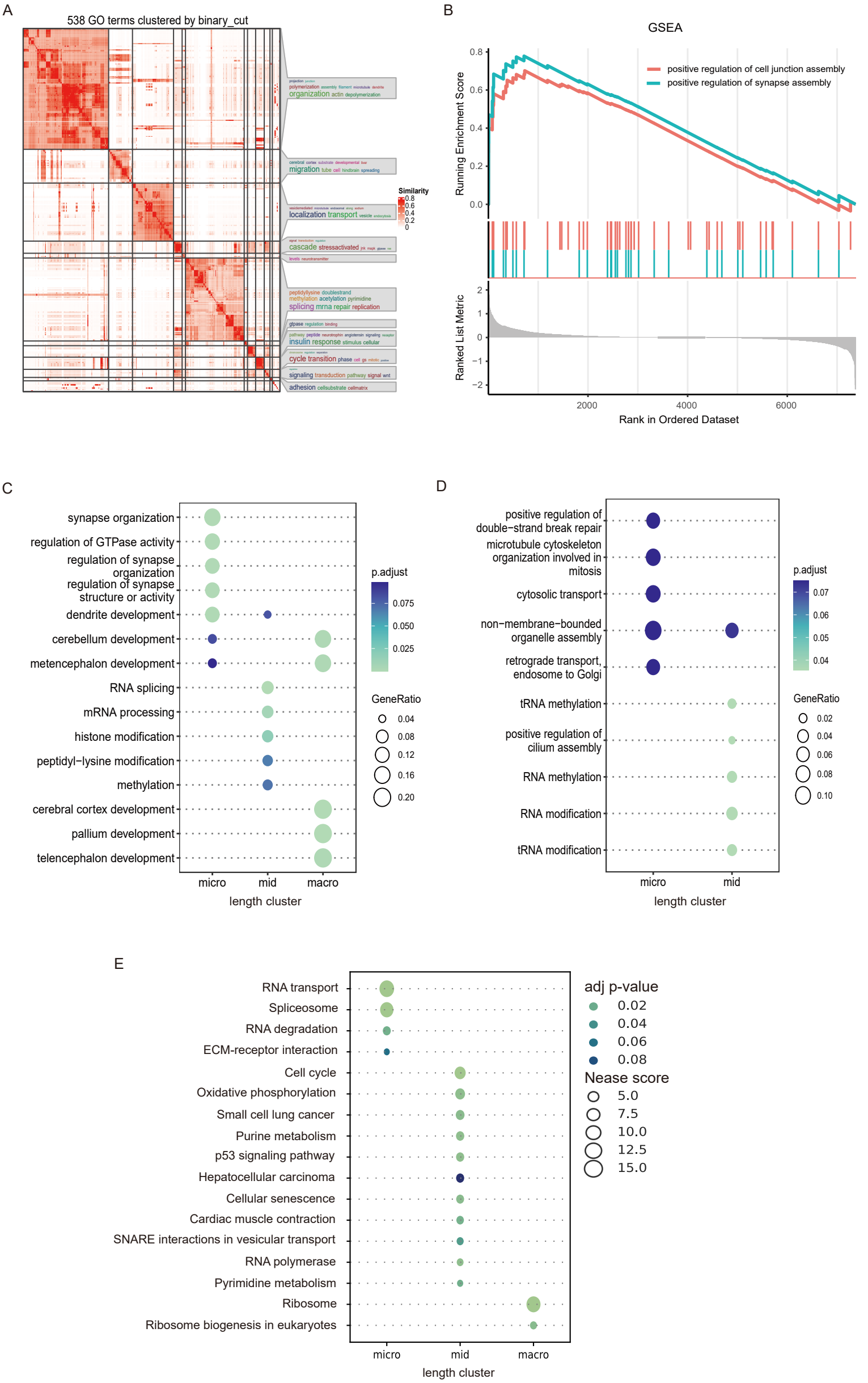


Figure 3 Functional enrichment analysis of differential AS at gene- and exon-level

A. Clustering similar GO terms using simplifyEnrichment R package for differential AS genes identified between embryonic mouse forebrain E11.5 and E16.5

B. Gene-level GSEA analysis shows enriched pathways for AS genes with large ΔPSI values between embryonic mouse forebrain E11.5 and E16.5

C. Enriched GO terms for three groups of differential AS genes identified between embryonic mouse forebrain E11.5 and E12.5, which are clustered on the size of spliced exons

D. Enriched GO terms for three groups of differential AS genes identified between human stem cells 0h and 6h, which are clustered on the size of spliced exons

E. Exon-level KEGG enrichment for three groups of spliced exons identified between human stem cells 0h and 6h

Figure 4

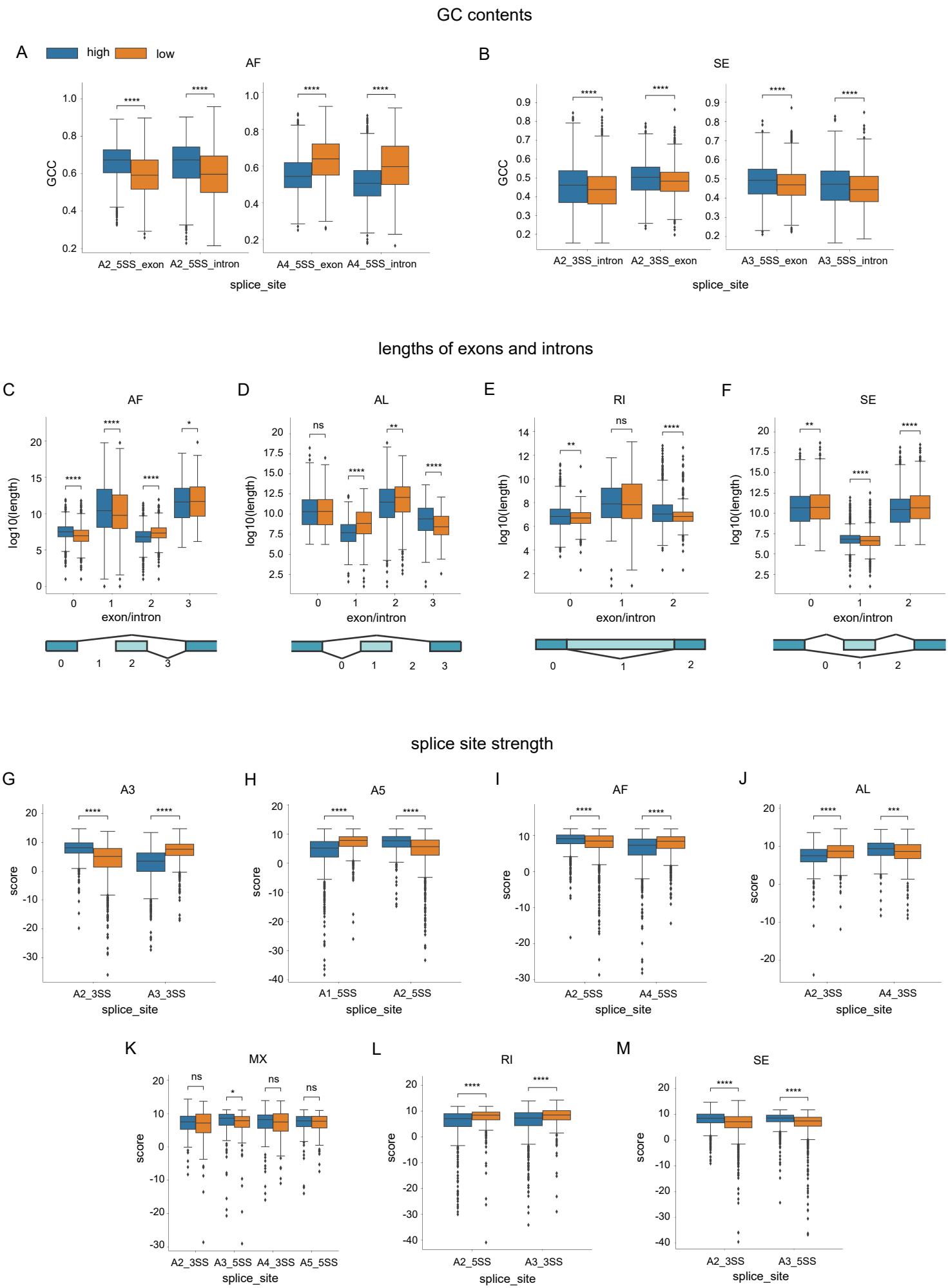

GC contents
A
high
low
AF
GCC
A2_5SS_exon
A2_5SS_intron
A4_5SS_exon
A4_5SS_intron
splice_site
B
SE
A2_3SS_intron
A2_3SS_exon
A3_5SS_exon
A3_5SS_intron
lengths of exons and introns
C
AF
log10(length)
exon/intron
D
AL
E
RI
F
SE
splice site strength
G
A3
score
A2_3SS
A3_3SS
H
A5
A1_5SS
A2_5SS
I
AF
A4_5SS
J
AL
A4_3SS
K
MX
A3_5SS
A5_5SS
L
RI
M
SE

Figure 4 Sequence features associated with AS events at E16.5

A-B. +/- 150 nt flanking regions of each 5' or 3' splice site were extracted, and their GC contents were computed after removing the conserved 5' or 3' splice site sequences. In AF (A) and SE (B) classes, the high PSI group have a higher GC content in these flanking regions than the low PSI group. In addition, "_exon" and "_intron" suffix denote the exonic and intronic flanking regions of the splice site, respectively.

C-F. Lengths of up- or downstream exons/ introns for each alternative splice site, converted to log base 2 values. In AF, AL, and ES classes (C, D, F), larger alternative exons are more likely to be spliced in than smaller ones. In RI (E), the PSI high group has significantly larger upstream and downstream exons of a retained intron than the PSI low group.

G-M. Splice site strength is computed for each alternative splice site based on MaxEntScan. Higher splice site strength favors the inclusion of alternative exons in A3, A5, AF, AL, and ES (G-J, M). In contrast, lower spice site strength promotes the intron retention in RI (L).

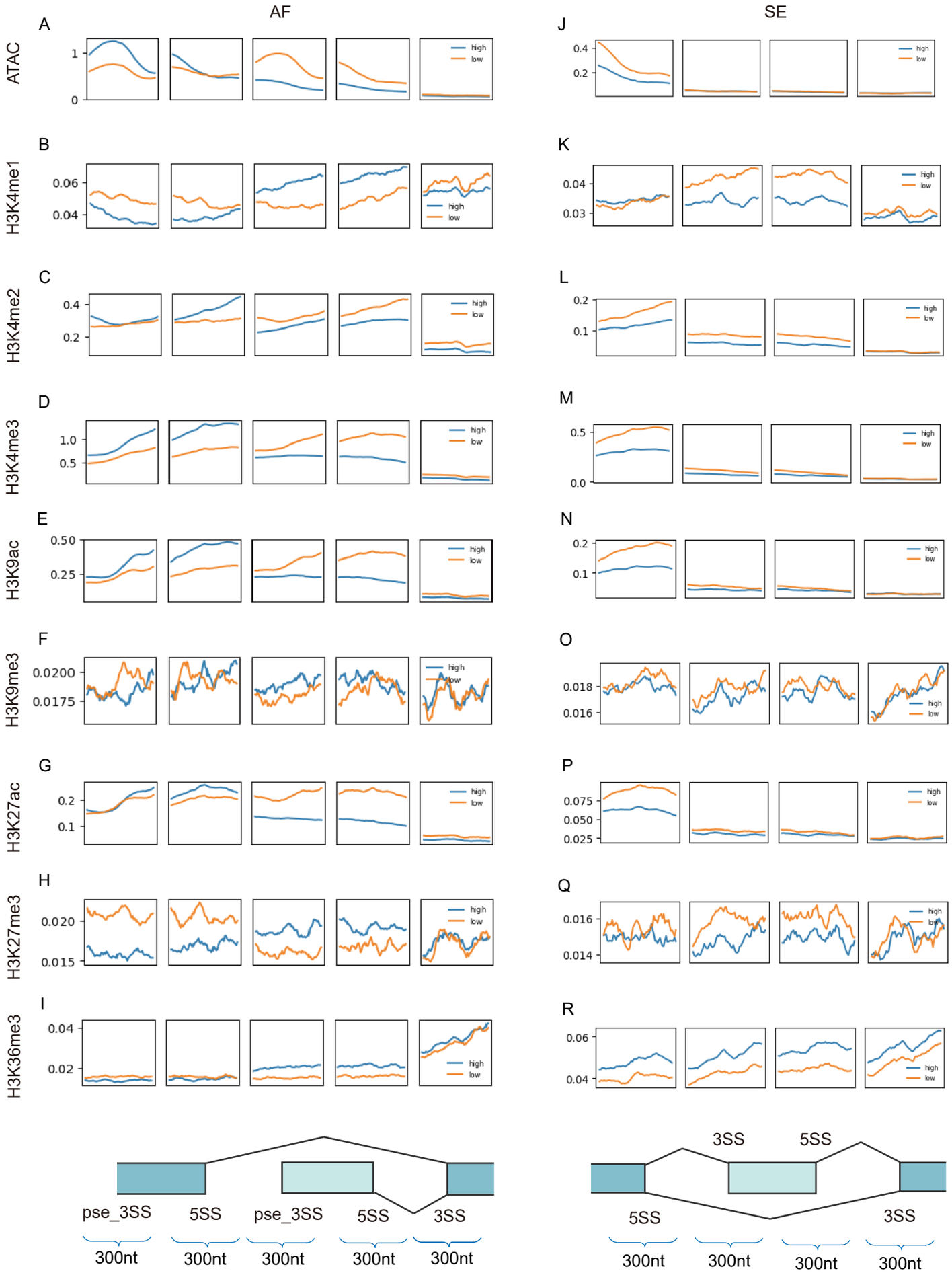


Figure 5 Epigenetic features associated with AS events at E16.5

A-I. In AF, several epigenetic features are positively correlated with the PSI high group: ATAC signal, H3K4me2, H3K4me3, H3K9ac, and H3K27ac. Splice sites are illustrated at the bottom.

J-R. In SE, H3K36me3 has a positive correlation with the PSI high groups, while H3K4me1, H3K4me2, and H3K27me3 have negative correlations. Splice sites are illustrated at the bottom.

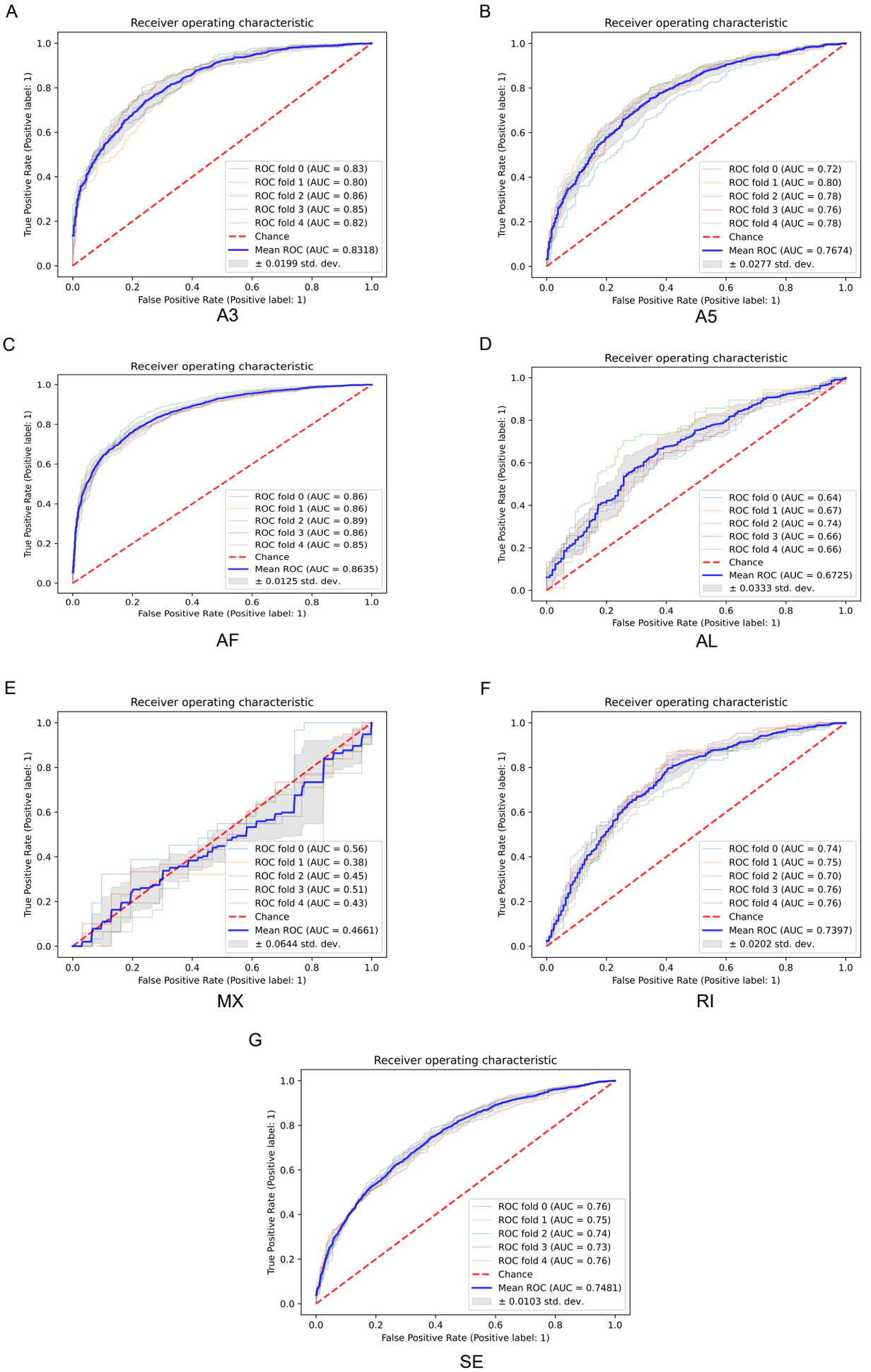


Figure 6 ROC curves of PSI high versus low classification for seven AS types at E16.5 using pooled sequence and epigenetic features

A-G. ROC was measured with 5-fold cross validation for each AS type. The mean ROC AUC scores are reported: 0.83 (A3), 0.77 (A5), 0.86 (AF), 0.67 (AL), 0.47 (MX), 0.74 (RI), and 0.75 (SE).

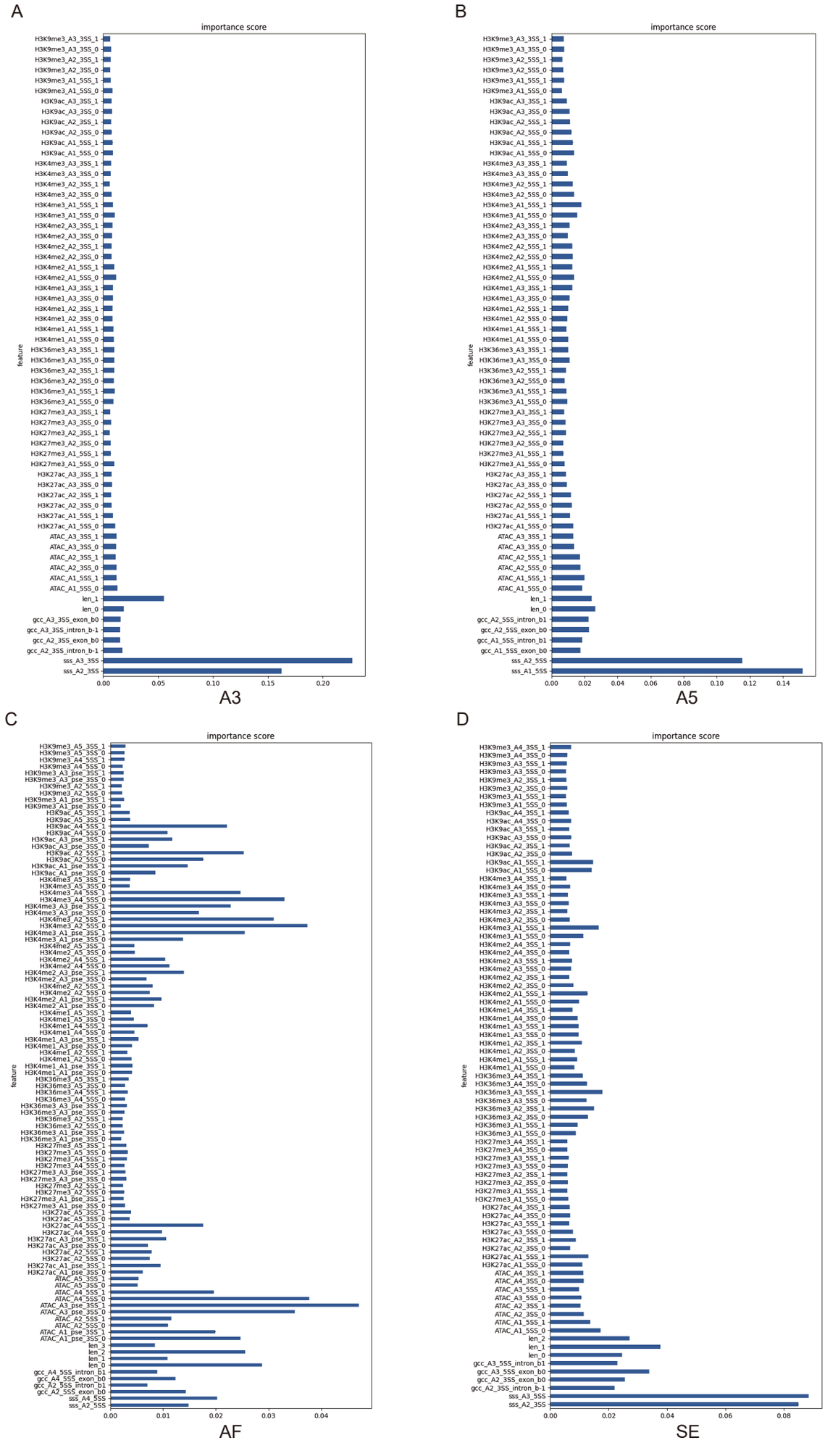


Figure 7 Feature importance in A3, A5, AF, and SE

A-B. In A3 and A5, the splice site strength is the most important feature deciding PSI high versus low classification.

C-D. In AF and SE, multiple sequence features and epigenetic factors strongly influence PSI high versus low classification.

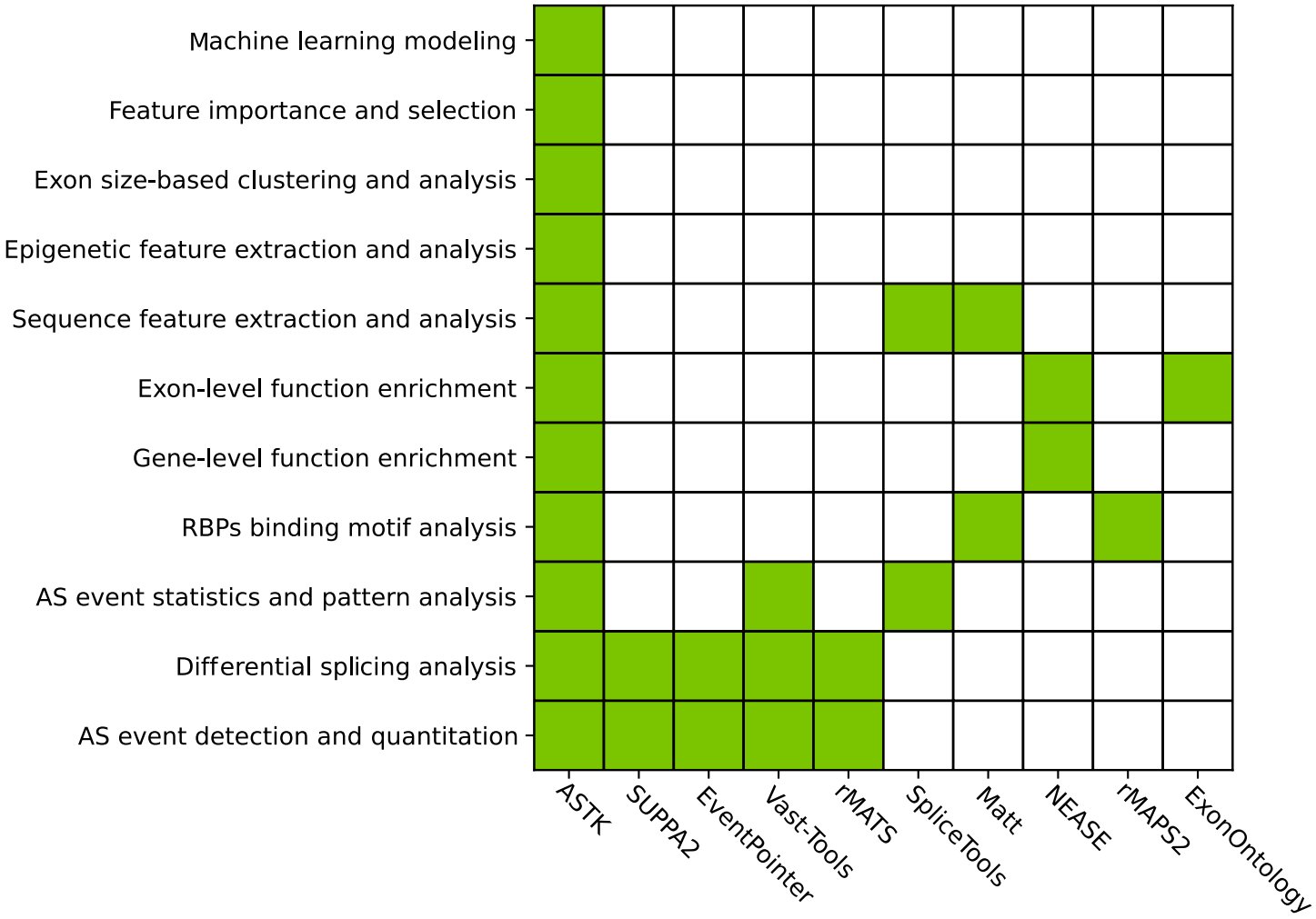
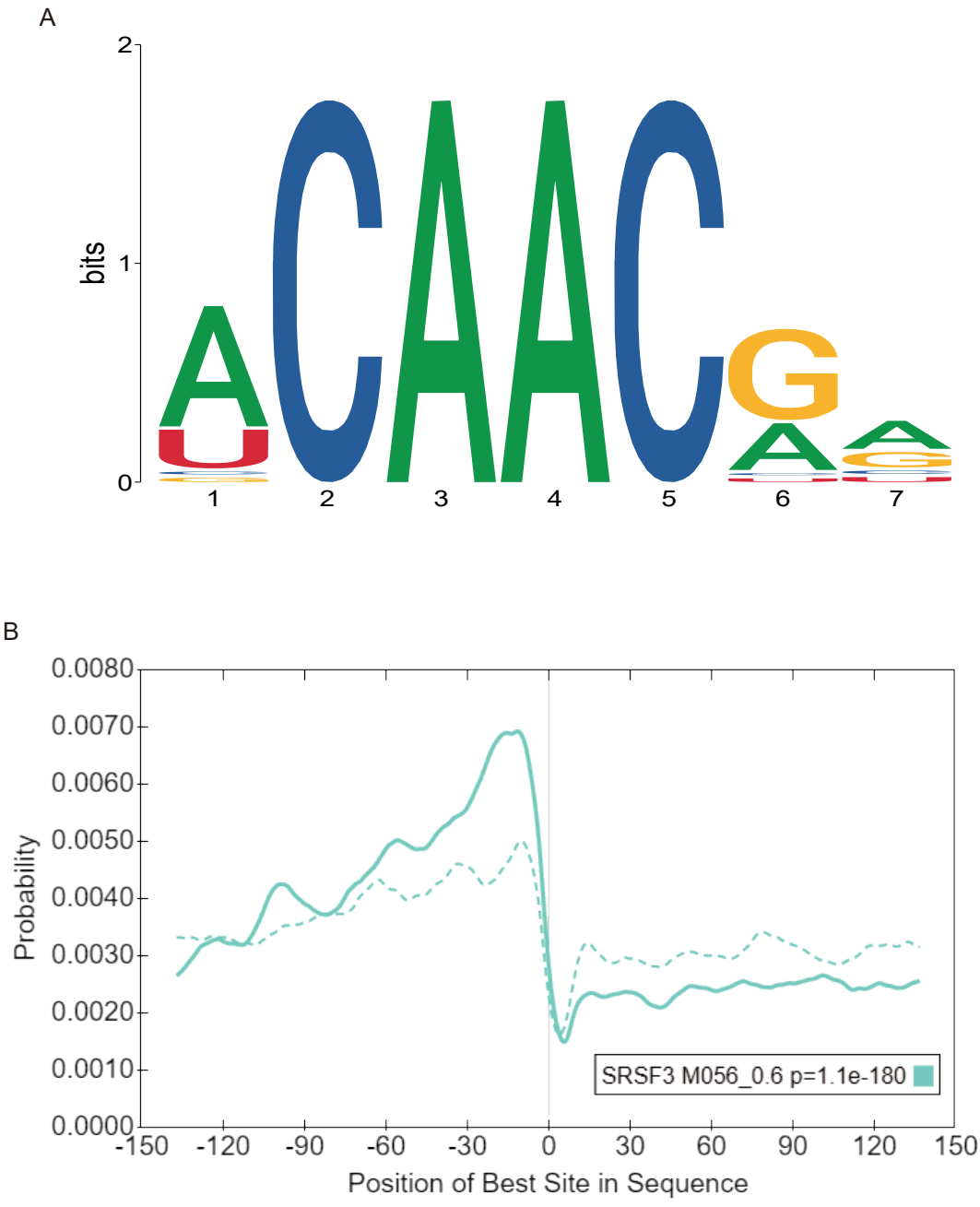


Figure 8 Comparison of ASTK with other software



Supplemental Figure 1 RBPs binding motif analysis

A. The sequence logo of the enriched SRSF3 binding motif

B. The density curve of the SRSF3 binding motif in the flanking region of alternative 5' splice site in SE for human embryonic stem cells at 72 hours, solid line – high PSI group, dashed line – low PSI group

Table 1. Accuracies (ACC) and AUC values of random forest models to predict PSI high versus low groups for seven AS types at each developmental time point using combined sequence features as input

	e12.5		e13.5		e14.5		e15.5		e16.5	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
A3	0.7597	0.8430	0.7597	0.8453	0.7695	0.8532	0.7546	0.8421	0.7532	0.8410
A5	0.7077	0.7794	0.7165	0.7869	0.7232	0.7911	0.7159	0.7909	0.7168	0.7825
AF	0.7585	0.8333	0.7424	0.8170	0.7464	0.8254	0.7466	0.8249	0.7469	0.8246
AL	0.6990	0.7402	0.6606	0.6987	0.6800	0.7284	0.6692	0.7183	0.6606	0.7110
MX	0.5610	0.5928	0.5286	0.5573	0.5068	0.5014	0.4443	0.4731	0.5324	0.5372
RI	0.6908	0.7569	0.6951	0.7629	0.6962	0.7554	0.6963	0.7642	0.7001	0.7651
SE	0.6912	0.7586	0.6834	0.7467	0.6753	0.7431	0.6798	0.7466	0.6768	0.7414

Table 2. Accuracies (ACC) and AUC values of random forest models to predict PSI high versus low groups for seven AS types at each developmental time point using combined epigenetic features as input

	e12.5		e13.5		e14.5		e15.5		e16.5	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
A3	0.5008	0.5112	0.5089	0.5126	0.5025	0.4991	0.5131	0.5196	0.4973	0.5053
A5	0.5204	0.5282	0.5295	0.5560	0.5410	0.5513	0.5292	0.5514	0.5349	0.5486
AF	0.7732	0.8514	0.7648	0.8418	0.7696	0.8456	0.7698	0.8444	0.7678	0.8483
AL	0.5788	0.6308	0.5661	0.6126	0.5521	0.5907	0.5359	0.5543	0.5124	0.5374
MX	0.4557	0.4382	0.5077	0.4833	0.5035	0.5129	0.5295	0.4920	0.4678	0.4458
RI	0.5500	0.5607	0.5402	0.5840	0.5526	0.5780	0.5220	0.5393	0.5063	0.5348
SE	0.5852	0.6245	0.5843	0.6194	0.5742	0.6125	0.5849	0.6229	0.5901	0.6262

Table 3. Accuracies (ACC) and AUC values of random forest models to predict PSI high versus low groups for seven AS types at each developmental time point using pooled sequence and epigenetic features as input

	e12.5		e13.5		e14.5		e15.5		e16.5	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
A3	0.7441	0.8286	0.7537	0.8336	0.7629	0.8420	0.7473	0.8266	0.7437	0.8318
A5	0.7027	0.7693	0.6949	0.7714	0.7052	0.7794	0.7113	0.7713	0.7013	0.7674
AF	0.7918	0.8716	0.7892	0.8599	0.7828	0.8628	0.7823	0.8622	0.7771	0.8635
AL	0.6625	0.7353	0.6342	0.6791	0.6411	0.7017	0.6257	0.6765	0.6398	0.6725
MX	0.5134	0.5052	0.5071	0.4956	0.5137	0.5065	0.4507	0.4272	0.4805	0.4661
RI	0.6900	0.7333	0.6692	0.7288	0.6822	0.7371	0.6691	0.7288	0.6756	0.7397
SE	0.6923	0.7602	0.6828	0.7475	0.6773	0.7393	0.6858	0.7490	0.6776	0.7481

Table 4. Accuracies (ACC) and AUC values of random forest models to predict PSI high versus low groups for seven AS types at E16.5 using each or combined sequence and epigenetic features as input

	A3		A5		AF		AL		MX		RI		SE	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
sss	0.7559	0.8324	0.7172	0.7703	0.5856	0.6171	0.5295	0.5499	0.4934	0.5165	0.6187	0.6480	0.6274	0.6809
gcc	0.5072	0.5213	0.5424	0.5537	0.6858	0.7515	0.4962	0.5190	0.5003	0.5010	0.5514	0.5739	0.5743	0.5959
len	0.5683	0.6027	0.5143	0.5216	0.6277	0.6777	0.6292	0.6851	0.5617	0.5741	0.6590	0.6983	0.5775	0.6115
ATAC	0.4920	0.5011	0.5134	0.5212	0.7543	0.8282	0.4906	0.5099	0.4576	0.4205	0.5158	0.5103	0.5594	0.5697
H3K4me1	0.4958	0.5094	0.4975	0.4992	0.6176	0.6640	0.5780	0.5892	0.4869	0.4997	0.5143	0.5241	0.5596	0.5806
H3K4me2	0.5125	0.5135	0.5193	0.5273	0.7077	0.7878	0.4971	0.5236	0.5002	0.4917	0.5182	0.5426	0.5621	0.5716
H3K4me3	0.5129	0.5019	0.5176	0.5320	0.7597	0.8400	0.5371	0.5802	0.4647	0.4465	0.5435	0.5367	0.5523	0.5743
H3K9ac	0.4901	0.4869	0.5172	0.5339	0.7507	0.8288	0.5190	0.5194	0.4222	0.4304	0.5364	0.5463	0.5628	0.5810
H3K9me3	0.5076	0.5050	0.4853	0.4934	0.5179	0.5274	0.5171	0.5280	0.5002	0.4885	0.5000	0.4929	0.5132	0.5181
H3K27ac	0.4837	0.4798	0.5202	0.5245	0.7082	0.7823	0.5209	0.5177	0.5097	0.5294	0.5308	0.5312	0.5600	0.5796
H3K27me3	0.5144	0.5268	0.5340	0.5353	0.5969	0.6397	0.5399	0.5414	0.4967	0.4828	0.5253	0.5313	0.5153	0.5139
H3K36me3	0.4920	0.4958	0.5063	0.4994	0.5862	0.6281	0.5257	0.5212	0.5451	0.5293	0.5554	0.5723	0.5642	0.5915
seq	0.7532	0.8410	0.7168	0.7825	0.7469	0.8246	0.6606	0.7110	0.5324	0.5372	0.7001	0.7651	0.6768	0.7414
8hm+ATAC	0.4973	0.5053	0.5349	0.5486	0.7678	0.8482	0.5124	0.5374	0.4678	0.4458	0.5063	0.5348	0.5901	0.6262
8hm+ATAC+seq	0.7437	0.8318	0.7013	0.7674	0.7771	0.8635	0.6398	0.6725	0.4805	0.4661	0.6756	0.7397	0.6776	0.7481

sss: splice site strength; gcc: GC content; len: lengths of exons and introns; seq: combined sequence features; hm: histone modifications