Article Title:

ASTK: a software for comprehensive analysis of alternative splicing

Author names and affiliation:

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

1 Key Laboratory of Laboratory Medicine, Ministry of Education, 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 Department of Mathematics, School of Science & Engineering, Tulane University, New Orleans, LA, USA

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

6 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) is pervasive and a key regulatory mechanism leading to diverse protein isoforms with distinct functions. Next-generation sequencing has enabled genome-wide studies of AS. However, AS analysis of massive amount of transcriptomic data generated by RNAseq is challenging due to the limited functions of current software packages, calling for a comprehensive and flexible pipeline. Here, we present ASTK, a software package combing diverse tools to systematically analyze AS events: including clustering AS genes based on spliced exon lengths, sequence feature extraction, functional enrichment analysis for spliced genes and spliced exons respectively, identifying binding motifs of RNA-binding proteins (RBPs) surrounding splice sites, and characterizing splicing-associated histone marks and chromatin states. ASTK was tested using experimental datasets to show its effectiveness and reliability. In our test cases, we grouped differential AS genes/spliced exons into three classes based on the length of spliced exons, which tend to be enriched into discrete functions. Furthermore, we identified several sequence features and epigenetic marks associated with the inclusion level of spliced exons. Therefore, ASTK will facilitate the functional characterization of AS events and improve our understanding of splicing regulation.

Key words: Alternative splicing, systematical analysis, functional enrichment, sequence features, epigenetic marks, RNA binding proteins

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]. As a highly regulated process, AS occurs by differential selection of splice sites. Several sequence features, such as the splice site strength, the exon/intron architecture, and GC content, impact the exon recognition and selection [3, 4]. Moreover, pre-mRNA splicing is catalyzed by the spliceosome, a complex macromolecular machine consisting of five core snRNPs and more than 300 different RNA binding proteins (RBPs) [5]. Through RNA binding motifs, RBPs interact with their mRNA targets and regulate alternative splicing processes [6]. In addition to the interactions between trans-acting factors and their corresponding cis-regulatory sequences, it has been shown that histone modifications, chromatin conformation and accessibility also influence the splicing outcomes [1, 7-11].

Under evolutionary pressure, splicing is tightly regulated to ensure correct splicing of the pre-mRNAs. Exon conservation in length and sequence has been implicated in splicing regulation [12]. Statistically, majority of spliced exons are less than 300 nt in the human and mouse genomes [13, 14]. Comparative genomics of 76 vertebrate species, including human, demonstrated that exons of < 250 nucleotides (nt) are highly conserved in length. And more importantly, a strong correlation between exon length and exon sequence conservation is observed for these exons [12], suggesting that exon size is an evolutionary conserved feature. Meanwhile, exons longer than 250 nt are low in length-conservation [12]. In another study, over 13，000 small sized exons (≤ 51 nt), named micro-exons, are identified in the human genome and are highly conserved across species [15]. Recent studies have shown that micro-exons exhibit tissue-specific expression patterns and play important roles in cell

differentiation, cell migration, neuronal and glial functions [16-18]. Mis-regulated micro-exons appear to contribute to the development of neurological disorders and cancer [19]. In our ASTK pipeline, we provide the option to cluster AS genes according to spliced exon lengths before running downstream analysis. The exon length cutoffs are as follows: micro-exons (≤ 51 nt), mid-sized exons (52-250 nt), and macro-exons (≥ 251 nt). In our test cases, we grouped differential AS genes into three classes based on the spliced exon length. We found that the three groups of AS genes seem to be functionally enriched in distinct biological processes. Exon-level enrichment analysis indicates that the three groups of exons are associated with different pathways as well.

With the development of high-throughput sequencing technologies and the reduction of cost, huge amounts of transcriptomic data accumulated so far have provided unprecedented resources and challenges to fully characterize AS events. Despite multiple software available for AS analysis, most of them only focus on one or few aspects [20-24]. In this study, we propose a new command-line software package, ASTK, for comprehensive analysis of AS, including clustering AS genes based on the size of spliced exons, differential alternative splicing analysis, functional enrichment analysis for AS genes at gene-level and exon-level separately, sequence feature extraction, identifying RBP binding motifs, and detecting histone marks and chromatin states in flanking regions of splice sites. We demonstrated the use of ASTK with published RNAseq datasets.

Implementation

The pipeline process and components of ASTK are illustrated in Figure 1A. ASTK consists of five main 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 identification

SUPPA2 [23] is the default algorithm to detect AS events in ASTK. Splice events found by other software packages such as rMATs [22], can be imported into ASTK for downstream analysis too. Differential AS events can be determined using cutoffs of ΔPSI (Percent Spliced-In) and p-values. After the AS event detection, users have an option to cluster spliced exons. Based on previous work, we recommend clustering spliced exons at different lengths: ≤ 51 nt, 52-250 nt, and ≥ 251 nt. Principal component analysis (PCA) using PSI values can be used to visualize variation between samples, useful for quality control. Heatmap 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 spliced genes and exons

Gene-level gene ontology (GO), KEGG, and gene set enrichment analysis (GSEA) for differential AS genes are conducted by running clusterProfiler R package [25]. However, gene functional enrichment analysis often produces redundant GO terms. Hence, we use simplifyEnrichment R package [26] to simplify enrichment results generated by clusterProfiler.

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 in human [21]. Furthermore, alternative exons often encode intrinsically disordered regions, which are enriched with functional peptide modules, named short linear motifs (SLiMs) [27, 28]. We developed a function to search SLiMs from the most comprehensive eukaryotic linear motif (ELM) database [29] to identify potential functional changes due to the addition or removal of SLiM-containing exons.

Sequence feature extraction

ASTK provides a function for statistic comparison of individual sequence features between conditions. The sequence context around the splice sites from both constitutive and alternative exons is derived to examine potential patterns associated with AS. Using exon skipping as an example, the +/- 150 nt RNA sequences from each exon junction are extracted: +/- 150 nt flanking region surrounding 5' splice site, +/- 150 nt flanking region surrounding 3' splice site, namely, A1, A2, A3, and A4, respectively (Figure 1B). After removing the conserved 5' and 3' splice site sequences, GC content in the flanking region of each 5' or 3' splice site is calculated. The length distributions of spliced exons, their upstream and downstream introns are also analyzed.

Splice site strength is estimated using the maximum entropy score [3]. The maximum entropy model simultaneously considers adjacent and non-adjacent dependencies between positions. The score is calculated using a log-odd ratio for a 9-mer (5' splice site) or a 23-mer (3' splice site) sequence.

Histone modifications and chromatin accessibility profiling analysis

To associate a specific chromatin signature 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 files format. The reads are calculated for the promoter region and gene body region separately to distinguish the strong promoter effect. Heatmaps of ChIP/ATAC signals are generated with deepTools [30]. The enrichment levels of histone marks and chromatin accessibility profiling are visualized with metagene plots.

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.

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 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/.

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). The description and accession numbers of datasets used in this study are listed in Supplemental Table 1.

In general, AS events are classified into seven groups: 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 exon skipping (ES). As the most common AS type, ES events were chosen for further demo analysis. The usage of each module is illustrated with examples as follows.

AS events identification

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 gene sets, for example, the overlapping of differential AS genes 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 ES class during C57BL/6 mouse forebrain development (E11.5 to P0) were chosen as the 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. ASTK employs the simplifyEnrichment R package to calculate semantic similarities among enriched GO terms and cluster highly similar GO terms. Figure 3A shows 569 significant GO terms are clustered into a few major clusters for 970 differential AS genes identified between E11.5 and E16.5. However, GO terms sometimes overstate the significance of individual differentially expressed genes that

are simply selected using arbitrary cutoffs. In contrast, GSEA (Gene set enrichment analysis) takes all gene expression values into consideration and associate a biological pathway to a set of co-regulated genes. Both GO 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, prior to functional enrichment analysis, we attempted to cluster differential AS genes into three groups based on the size of spliced exons. In total, 509 differential AS genes were identified between E11.5 and E12.5, 78 genes with micro-exons spliced, 33 genes with macro-exons spliced, and 398 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. We took advantage of NEASE [21], a bioinformatics tool integrating pathways with structural annotations of protein-protein interactions for human genes, to functionally characterize AS events in the human transcriptomes. Using RNAseq data of human embryonic stem cells at seven early neural differentiation time points (0-72 h) [34], we demonstrated that NEASE can reveal 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 at the exon-level, which displays clear differences for micro-, mid-sized, and macro-exons.

Extraction of sequence features

Sequence features are part of genomic architecture, contributing to splice site and splicing unit recognition and affecting the splicing process [35, 36]. In this paper, several distinctive features were examined, including GC contents around splice sites, the lengths of spliced exons and their upstream/downstream introns, and the splice site strength.

The GC content is an important aspect of genomic organization and function. It is known that higher GC content is positively correlated with higher stability of DNA and secondary structures of mRNAs. In both human embryonic stem cells and mouse forebrain development datasets, our analysis results show that almost all A1 to A4 regions have higher GC content in high PSI groups compared to low PSI groups, suggesting that higher GC content in the flanking regions around both constitutive and alternative splice sites may favor inclusion of alternative exons (Figure 4A-H). Moreover, we also found that the high PSI groups tend to have relatively larger spliced exons compared to low PSI groups (Figure 4I-J).

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 indicated by maximum entropy scores are quite similar at 5' splice sites adjacent to the upstream constitutive exons and 3' splice sites adjacent to the downstream constitutive exons (constitutive splice sites) for both high PSI and low PSI groups. However, 5' and 3' splice sites flanking the skipped exons in high PSI groups appear to have significantly higher maximum entropy scores than in low PSI groups. 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 4K-L).

Characterization of histone marks and chromatin accessibility

Accumulating evidence indicates that local histone modifications and chromatin structure are critical to ensure proper regulation of alternative splicing. We systematically surveyed a variety of histone modifications and chromatin accessibility in A1 to A4 sites using ATACseq and ChIPseq datasets of mouse forebrain development. ChIPseq data contains 8 major histone modifications: H3K4me3, H3K4me2, H3K4me1, H3K27ac, H3K9ac, H3K36me3, H3K27me3, and H3K9me3. As shown in Figure 5A, the overall distribution of H3K27me3 signals in A1-A4 regions at P0 is not correlated with low or high PSI exons. However, at P0, H3K36me3 is specifically enriched in A1, A2, and A3 regions for high PSI exons (Figure 5D). Since H3K36me3 is known to play a role in alternative splicing, our finding is consistent with previous experimental results [10, 37]. Additionally, chromatin accessibility increases in the A4 region for high PSI exons at E16.5 (Figure 5E).

RBP motif enrichment analysis

RBPs broadly regulate RNA transcripts at both transcriptional and post-transcriptional levels, such as splicing, modification, and translation [6, 38]. Regulation of AS by RBPs depends on their binding site positions and preferences along the splice sites. Figure 5F shows the sequence logo of a RBP, SRSF3. The binding affinity difference of SRSF3 in the A3 region between high PSI and low PSI events for human embryonic stem cells at 72 hours is observed (Figure 5G).

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. Comprehensive databases for developmentally related and cancer associated AS have been created recently [39, 40]. 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 [41]. Approximately 15% of human disease-causing mutations are splicing mutations [42]. 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 toolkit for i) systematically analyzing functional effects of AS events, ii) dissecting out splicing-associated regulatory features and epigenetic marks.

Even though various computational tools have been developed to quantify transcript isoforms and identify splicing events [20-24], 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 functional impact. More often, spliced exons encode short linear motifs and affect protein-protein interactions. In our software, we integrated the eukaryotic linear motif (ELM) database and 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 [12, 19, 43, 44]. 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 study.

AS is a hallmark of the complexity and diversity of gene regulation. During the process, genomic architecture such as splice site strength, GC content, 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 [9, 10, 37, 45-48].

Eventually, splicing factors are recruited to their specific binding motifs along splice sites to assist intron folding and splicing [1, 7]. 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 [45], 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, RNA binding motifs, and extract enriched histone marks and chromatin states for flanking regions of splice sites. In our testing cases, we found that higher splice site strength, higher GC content, and larger exon size likely promote the inclusion of alternatively spliced exons. On top of that, our preliminary results suggested that histone marks, for example, H3K36me3, and open chromatin state might be determinants of exon inclusion.

To our knowledge, ASTK is the first integrated platform to thoroughly analyze and annotate AS events. ASTK streamlines different bioinformatics tools into a single, easy-to-use software. Through testing of experimental RNAseq datasets, we showed that ASTK can systematically and effectively address the needs. Overall, our ASTK pipeline will facilitate the functional studies of AS, as well as provide mechanistic insights into how AS is regulated and evolved.

Key Points

- We developed a flexible and comprehensive software, ASTK, for alternative splicing analysis.
- ASTK performs clustering of AS genes/alternative exons based on the length of spliced exons, thus providing additional functional insights.
- ASTK can conduct various functional enrichment analysis of AS events.

- ASTK can efficiently extract sequence features, histone marks, chromatin states, and RBP binding motifs associated with the inclusion level of spliced exons.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

Acknowledgements

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

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. 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.
4. Tammer, L., et al., *Gene architecture directs splicing outcome in separate nuclear spatial regions.* Mol Cell, 2022. **82**(5): p. 1021-1034 e8.
5. Jurica, M.S., et al., *Purification and characterization of native spliceosomes suitable for three-dimensional structural analysis.* RNA, 2002. **8**(4): p. 426-39.

6. 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.
7. Shi, Y., *Mechanistic insights into precursor messenger RNA splicing by the spliceosome.* Nat Rev Mol Cell Biol, 2017. **18**(11): p. 655-670.
8. 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.
9. 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.
10. Luco, R.F., et al., *Regulation of alternative splicing by histone modifications.* Science, 2010. **327**(5968): p. 996-1000.
11. Ullah, F., et al., *Exploring the relationship between intron retention and chromatin accessibility in plants.* Bmc Genomics, 2018. **19**.
12. Movassat, M., et al., *Exon size and sequence conservation improves identification of splice-altering nucleotides.* RNA, 2019. **25**(12): p. 1793-1805.
13. Song, S., et al., *Comparative component analysis of exons with different splicing frequencies.* PLoS One, 2009. **4**(4): p. e5387.
14. Gudlaugsdottir, S., et al., *Exon size distribution and the origin of introns.* Genetica, 2007. **131**(3): p. 299-306.
15. 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.
16. 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.
17. 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**.
18. 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.
19. 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.
20. Katz, Y., et al., *Analysis and design of RNA sequencing experiments for identifying isoform regulation.* Nat Methods, 2010. **7**(12): p. 1009-15.
21. 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.
22. 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.
23. Trincado, J.L., et al., *SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions.* Genome Biol, 2018. **19**(1): p. 40.
24. Li, Y.I., et al., *Annotation-free quantification of RNA splicing using LeafCutter.* Nat Genet, 2018. **50**(1): p. 151-158.
25. Yu, G., et al., *clusterProfiler: an R package for comparing biological themes among gene clusters.* OMICS, 2012. **16**(5): p. 284-7.

26. Gu, Z. and D. Hubschmann, *Simplify enrichment: A bioconductor package for clustering and visualizing functional enrichment results.* Genomics Proteomics Bioinformatics, 2022.
27. Romero, P.R., et al., *Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms.* Proc Natl Acad Sci U S A, 2006. **103**(22): p. 8390-5.
28. Fuxreiter, M., P. Tompa, and I. Simon, *Local structural disorder imparts plasticity on linear motifs.* Bioinformatics, 2007. **23**(8): p. 950-6.
29. Kumar, M., et al., *ELM-the eukaryotic linear motif resource in 2020.* Nucleic Acids Res, 2020. **48**(D1): p. D296-D306.
30. Ramirez, F., et al., *deepTools2: a next generation web server for deep-sequencing data analysis.* Nucleic Acids Res, 2016. **44**(W1): p. W160-5.
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. 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.
35. 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.
36. 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.
37. Hu, Q., et al., *Histone posttranslational modifications predict specific alternative exon subtypes in mammalian brain.* PLoS Comput Biol, 2017. **13**(6): p. e1005602.
38. Hentze, M.W., et al., *A brave new world of RNA-binding proteins.* Nat Rev Mol Cell Biol, 2018. **19**(5): p. 327-341.
39. Li, Z., et al., *MeDAS: a Metazoan Developmental Alternative Splicing database.* Nucleic Acids Res, 2021. **49**(D1): p. D144-D150.
40. Wu, S., et al., *ASCancer Atlas: a comprehensive knowledgebase of alternative splicing in human cancers.* Nucleic Acids Res, 2022.
41. Kim, H.K., et al., *Alternative splicing isoforms in health and disease.* Pflugers Arch, 2018. **470**(7): p. 995-1016.
42. Blencowe, B.J., *Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases.* Trends Biochem Sci, 2000. **25**(3): p. 106-10.
43. Ustianenko, D., S.M. Weyn-Vanhentenryck, and C. Zhang, *Microexons: discovery, regulation, and function.* Wiley Interdiscip Rev RNA, 2017. **8**(4).
44. Irimia, M., et al., *A highly conserved program of neuronal microexons is misregulated in autistic brains.* Cell, 2014. **159**(7): p. 1511-23.
45. 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.
46. Li, T., et al., *Integrative analysis reveals functional and regulatory roles of H3K79me2 in mediating alternative splicing.* Genome Med, 2018. **10**(1): p. 30.

47. Spies, N., et al., *Biased chromatin signatures around polyadenylation sites and exons.* Mol Cell, 2009. **36**(2): p. 245-54.
48. 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.

Figure 1 workflow of ASTK pipeline

A. Schematic illustration of the workflow

B. Diagram of the flanking regions of splice sites in exon skipping events

Figure 2 AS events identification for mouse forebrain development between E11.5 and 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 differential AS gene sets

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

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

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

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

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

Figure 4 Sequence features associated with AS events

A-D. For exon skipping in human embryonic stem cells, A1 to A4 flanking regions all have higher GC content in high PSI groups than in low PSI groups.

E-H. For exon skipping in mouse forebrain development, A1 to A4 flanking regions except A4 3'SS in exons have higher GC content in high PSI groups than in low PSI groups.

I. For exon skipping in human embryonic stem cells, spliced exons in high PSI groups are statistically larger than in low PSI groups. Meanwhile, the downstream introns are statistically smaller than in low PSI groups.

J. For exon skipping in mouse forebrain development, spliced exons in high PSI groups are statistically larger than in low PSI groups.

K. For exon skipping in human embryonic stem cells, 5' and 3' splice sites flanking the skipped exons in high PSI groups have higher splice site strength than in low PSI groups.

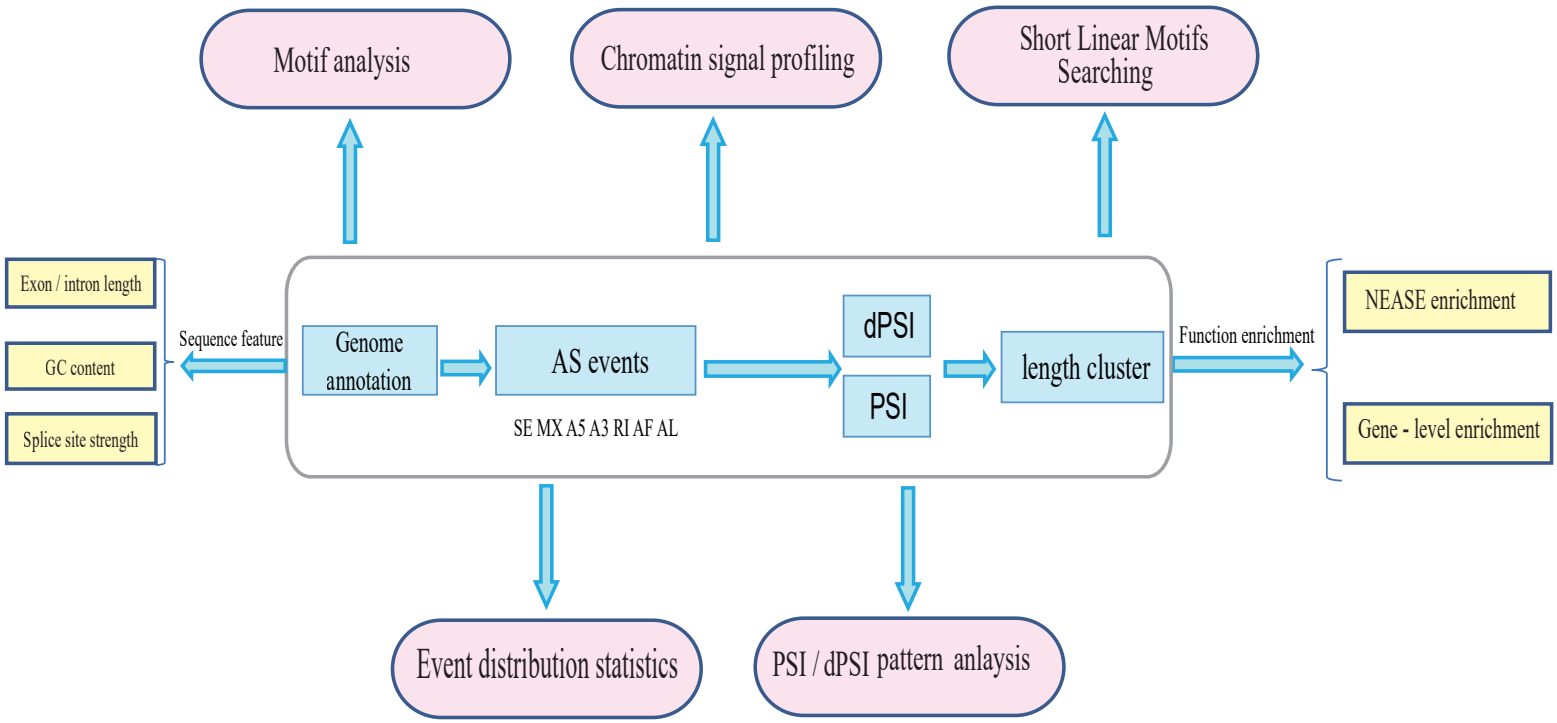
L. For exon skipping in mouse forebrain development, 5' and 3' splice sites flanking the skipped exons in high PSI groups have higher splice site strength than in low PSI groups.

Figure 5 AS associated histone marks, chromatin states, and RBPs

A. Distribution of H3K27me3 signals in A1-A4 regions for significant exon-skipping (ES) events at P0
B. Distribution of H3K27me3 signals in A1-A4 regions with high PSI (PSI > 0.8) for significant ES events at P0
C. Distribution of H3K27me3 signals in A1-A4 regions with low PSI (PSI < 0.2) for significant ES events at P0
D. Distribution of H3K36me3 signals in A1-A4 regions for significant ES events at P0
E. Distribution of ATAC signals in A1-A4 regions for significant ES events at E16.5
F. The sequence logo of the enriched SRSF3 binding motif
G. The density curve of the SRSF3 binding motif in the A3 region between high PSI and low PSI events for human embryonic stem cells at 72 hours, solid line – high PSI events, dashed line – low PSI events

Figure 1

A


Motif analysis
Chromatin signal profiling
Short Linear Motifs Searching
Exon / intron length
GC content
Splice site strength
Sequence feature
Genome annotation
AS events
SE MX A5 A3 RI AF AL
dPSI
PSI
length cluster
Function enrichment
NEASE enrichment
Gene - level enrichment
Event distribution statistics
PSI / dPSI pattern anlaysis

B

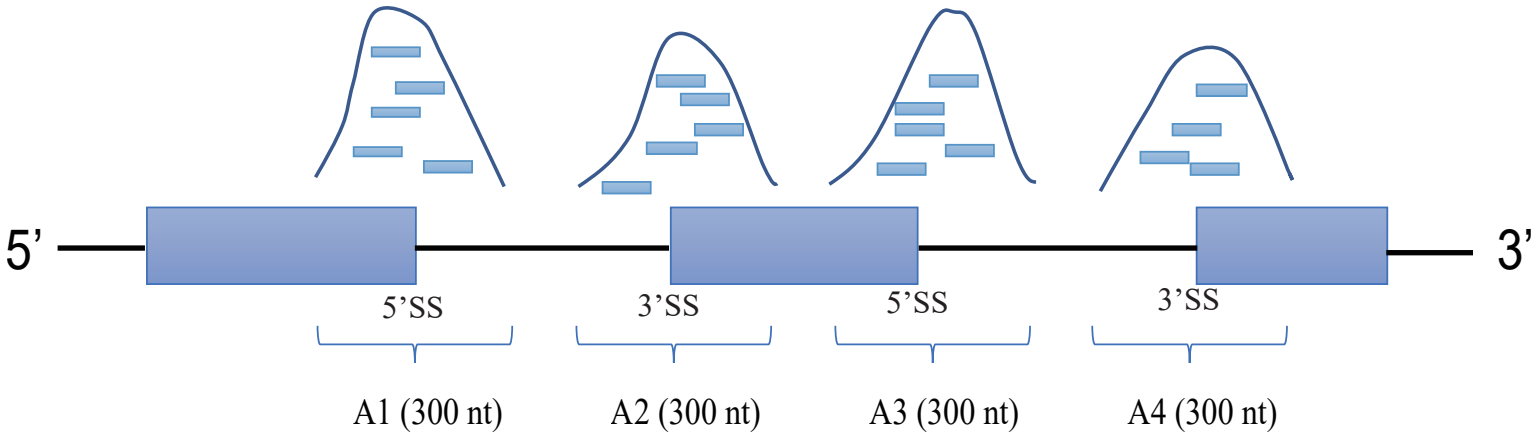

5'
3'
5'SS
3'SS
5'SS
3'SS
A1 (300 nt)
A2 (300 nt)
A3 (300 nt)
A4 (300 nt)

Figure 2

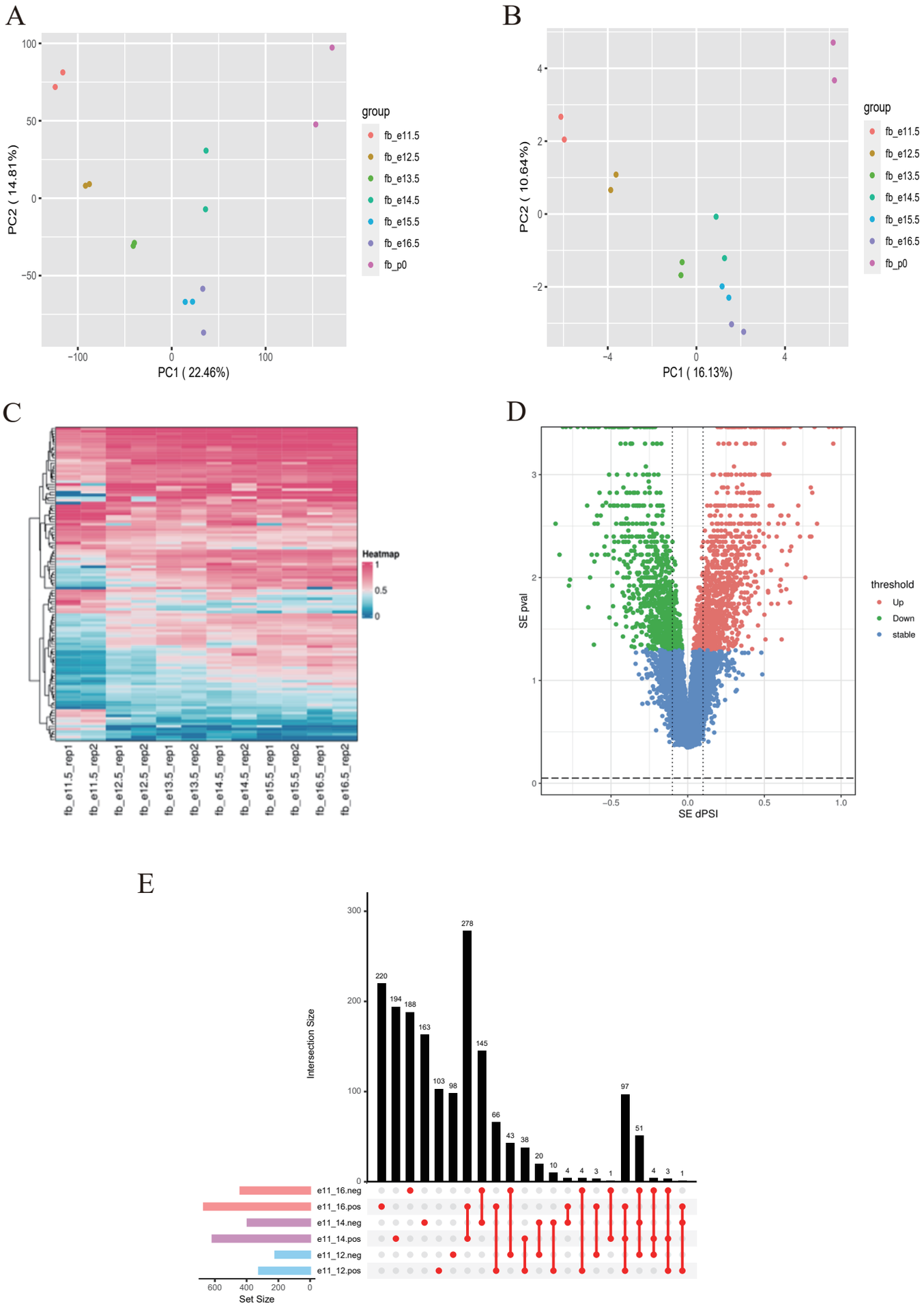

A
PC2 (14.81%)
PC1 (22.46%)
group
fb_e11.5
fb_e12.5
fb_e13.5
fb_e14.5
fb_e15.5
fb_e16.5
fb_p0
B
PC2 (10.64%)
PC1 (16.13%)
C
Heatmap
fb_e11.5_rep1
fb_e11.5_rep2
fb_e12.5_rep1
fb_e12.5_rep2
fb_e13.5_rep1
fb_e13.5_rep2
fb_e14.5_rep1
fb_e14.5_rep2
fb_e15.5_rep1
fb_e15.5_rep2
fb_e16.5_rep1
fb_e16.5_rep2
D
SE pval
SE dPSI
threshold
Up
Down
stable
E
Intersection Size
Set Size
e11_16.neg
e11_16.pos
e11_14.neg
e11_14.pos
e11_12.neg
e11_12.pos

Figure 3

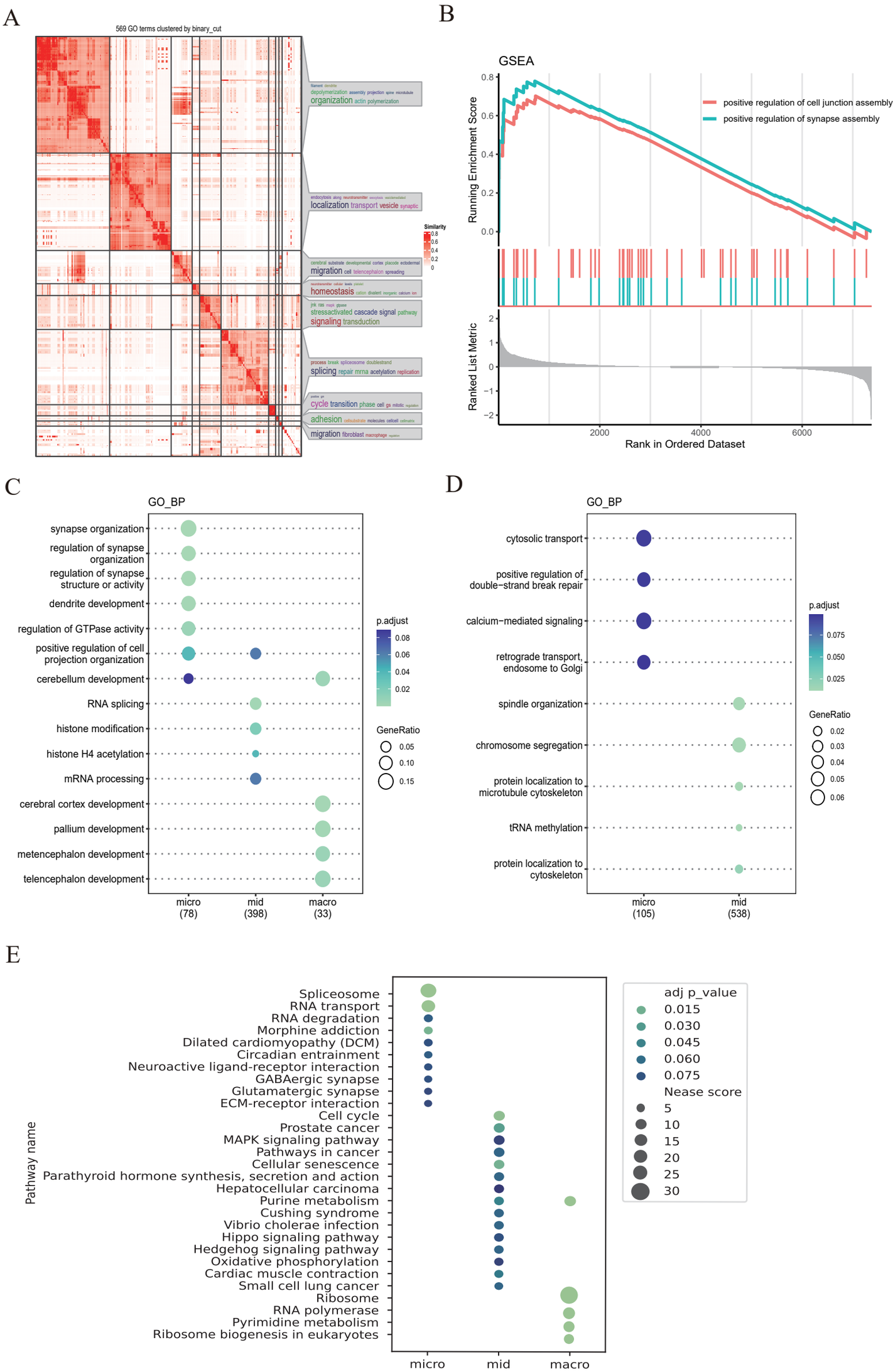

A
569 GO terms clustered by binary_cut
organization
localization transport vesicle
migration
homeostasis
signaling
splicing
cycle transition phase
adhesion
migration
Similarity
B
GSEA
positive regulation of cell junction assembly
positive regulation of synapse assembly
Running Enrichment Score
Ranked List Metric
Rank in Ordered Dataset
C
GO_BP
synapse organization
regulation of synapse organization
regulation of synapse structure or activity
dendrite development
regulation of GTPase activity
positive regulation of cell projection organization
cerebellum development
RNA splicing
histone modification
histone H4 acetylation
mRNA processing
cerebral cortex development
pallium development
metencephalon development
telencephalon development
micro (78)
mid (398)
macro (33)
p.adjust
GeneRatio
D
GO_BP
cytosolic transport
positive regulation of double−strand break repair
calcium−mediated signaling
retrograde transport, endosome to Golgi
spindle organization
chromosome segregation
protein localization to microtubule cytoskeleton
tRNA methylation
protein localization to cytoskeleton
micro (105)
mid (538)
p.adjust
GeneRatio
E
Pathway name
Spliceosome
RNA transport
RNA degradation
Morphine addiction
Dilated cardiomyopathy (DCM)
Circadian entrainment
Neuroactive ligand-receptor interaction
GABAergic synapse
Glutamatergic synapse
ECM-receptor interaction
Cell cycle
Prostate cancer
MAPK signaling pathway
Pathways in cancer
Cellular senescence
Parathyroid hormone synthesis, secretion and action
Hepatocellular carcinoma
Purine metabolism
Cushing syndrome
Vibrio cholerae infection
Hippo signaling pathway
Hedgehog signaling pathway
Oxidative phosphorylation
Cardiac muscle contraction
Small cell lung cancer
Ribosome
RNA polymerase
Pyrimidine metabolism
Ribosome biogenesis in eukaryotes
micro
mid
macro
adj p_value
Nease score

Figure 4

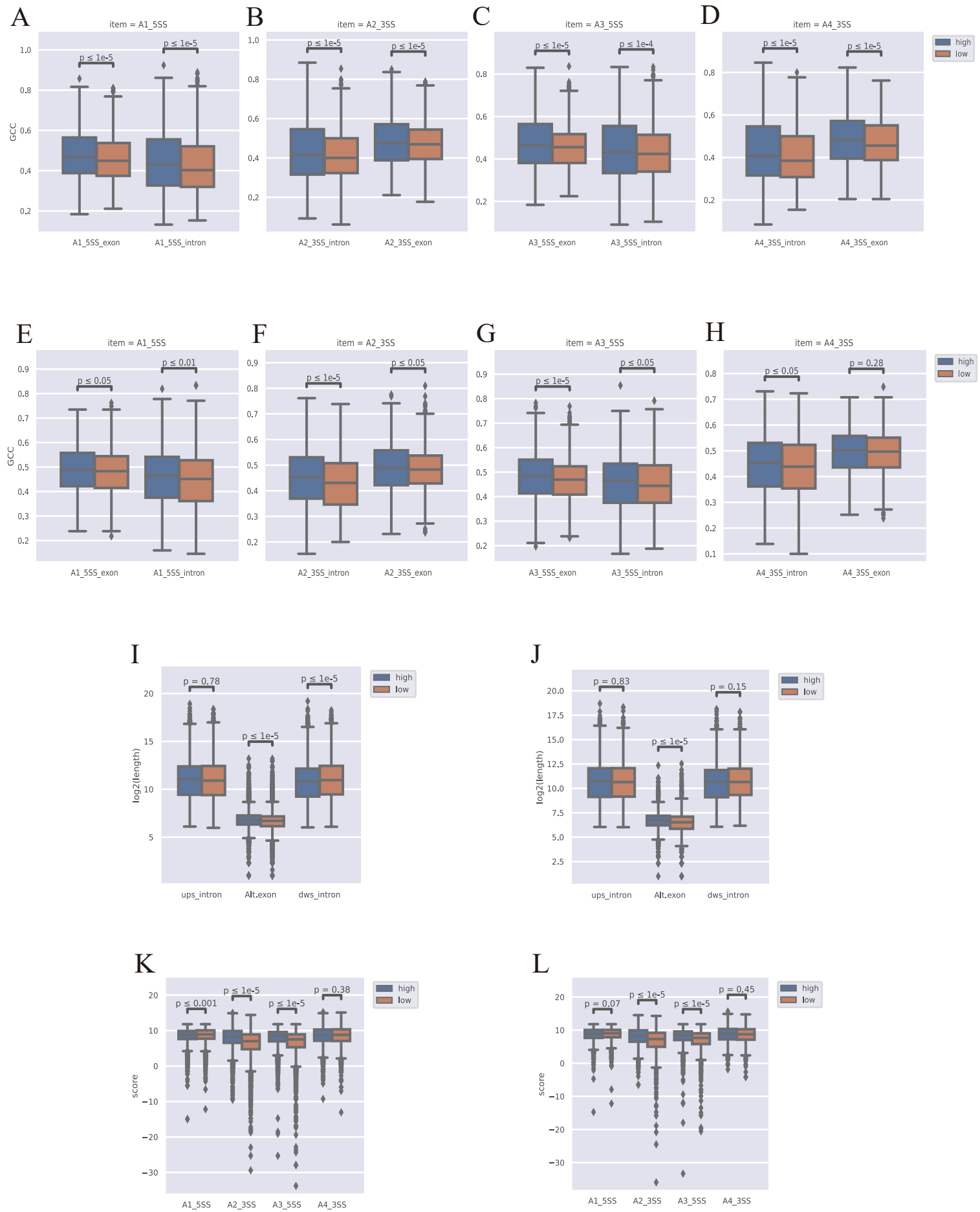
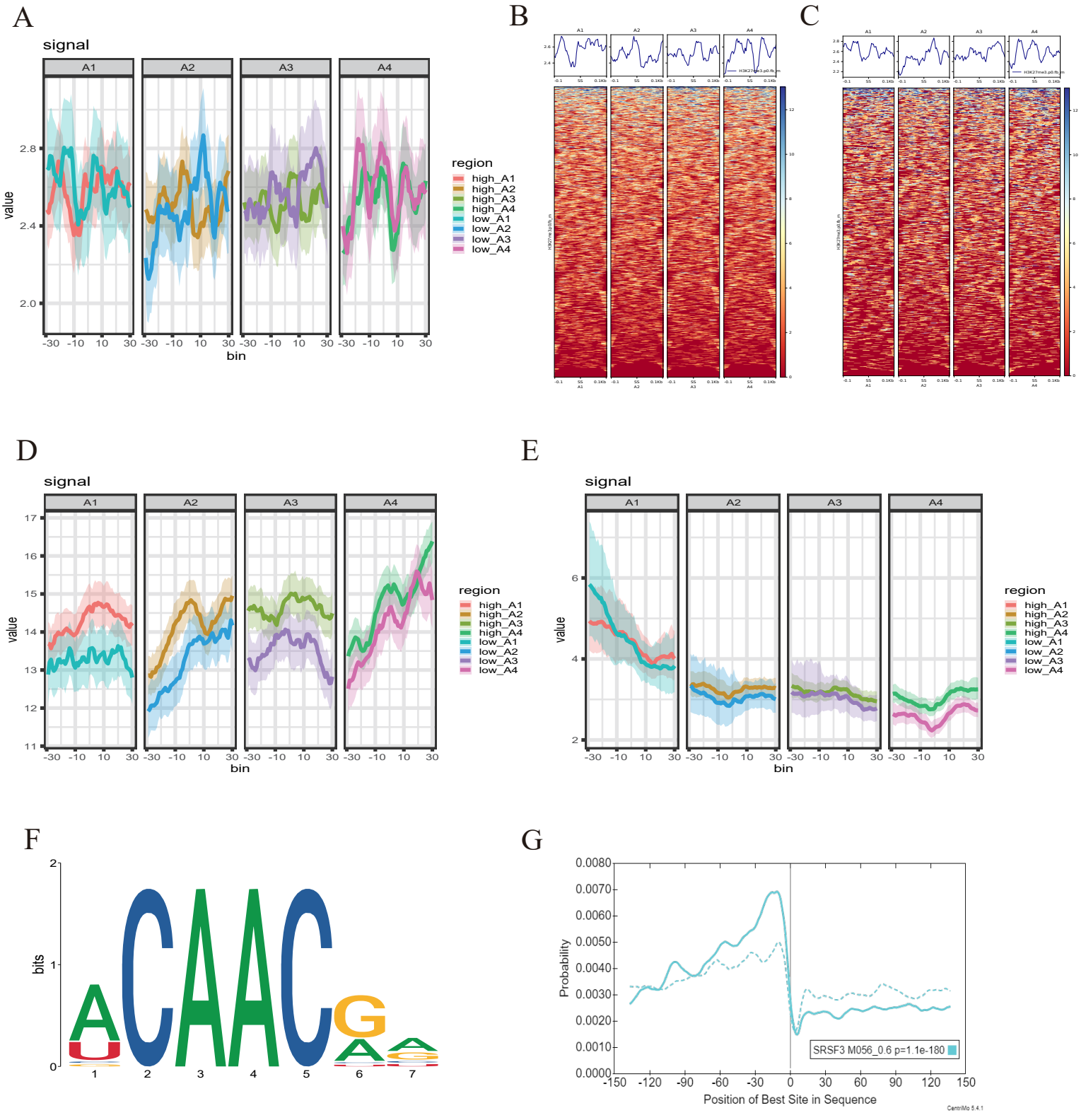

A
item = A1_5SS
p ≤ 1e-5
p ≤ 1e-5
GCC
A1_5SS_exon
A1_5SS_intron
B
item = A2_3SS
p ≤ 1e-5
p ≤ 1e-5
A2_3SS_intron
A2_3SS_exon
C
item = A3_5SS
p ≤ 1e-5
p ≤ 1e-4
A3_5SS_exon
A3_5SS_intron
D
item = A4_3SS
p ≤ 1e-5
p ≤ 1e-5
A4_3SS_intron
A4_3SS_exon
high
low
E
item = A1_5SS
p ≤ 0.05
p ≤ 0.01
GCC
A1_5SS_exon
A1_5SS_intron
F
item = A2_3SS
p ≤ 1e-5
p ≤ 0.05
A2_3SS_intron
A2_3SS_exon
G
item = A3_5SS
p ≤ 1e-5
p ≤ 0.05
A3_5SS_exon
A3_5SS_intron
H
item = A4_3SS
p ≤ 0.05
p = 0.28
A4_3SS_intron
A4_3SS_exon
high
low
I
p = 0.78
p ≤ 1e-5
p ≤ 1e-5
log2(length)
ups_intron
Alt.exon
dws_intron
high
low
J
p = 0.83
p ≤ 1e-5
p = 0.15
log2(length)
ups_intron
Alt.exon
dws_intron
high
low
K
p ≤ 0.001
p ≤ 1e-5
p ≤ 1e-5
p = 0.38
score
A1_5SS
A2_3SS
A3_5SS
A4_3SS
high
low
L
p = 0.07
p ≤ 1e-5
p ≤ 1e-5
p = 0.45
score
A1_5SS
A2_3SS
A3_5SS
A4_3SS
high
low

Figure 5


A
signal
A1
A2
A3
A4
value
bin
region
high_A1
high_A2
high_A3
high_A4
low_A1
low_A2
low_A3
low_A4
B
C
D
E
F
bits
G
Probability
Position of Best Site in Sequence
SRSF3 M056_0.6 p=1.1e-180
CentriMo 5.4.1