國立臺灣大學生命科學院基因體與系統生物學學位學程

碩士論文

Genome and Systems Biology Degree Program

College of Life Science

National Taiwan University

Master's Thesis

Paramecium bursaria 內含子：內共生期間的剪接調控與纖毛蟲物種中短內含子的演化

Paramecium bursaria introns: splicing regulation during endosymbiosis and short intron evolution in ciliate species

阮氏銀江

Nguyen Thi Ngan Giang

指導教授：林倩伶 博士 & 呂俊毅 博士

Advisor: Chien-Ling Lin, PhD & Jun-Yi Leu, PhD

中華民國 114 年 7 月

July 2025

doi:10.6342/NTU202502289

Acknowledgements

Throughout my master study, I have been so lucky to receive tremendous love and support. First and foremost, I am deeply grateful to my "big brother", Professor Nguyen Cong Tung, for encouraging me to apply to NTU from the very beginning and for his unwavering support throughout my academic journey. I also sincerely thank the Taiwan Ministry of Education for making my study in Taiwan possible through their generous support.

I want to thank my supervisors who have been my greatest influencers through my master study: Professor Chien-Ling Lin ("mẹ tui" *) and Professor Jun-Yi Leu ("ông ngoại tui" *). I want to thank Chien-Ling for her constant care, thoughtful guidance, and for inspiring me, not just to grow as a researcher, but also to become a more compassionate person. And to Jun-Yi, I am deeply grateful for all the advice he has given me, for teaching me how to think like a scientist, and above all, for his patience and encouragement throughout my learning and growth.

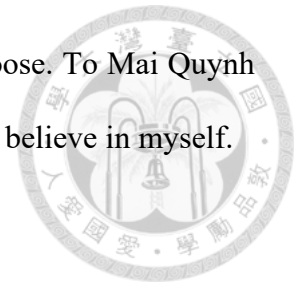
I want to express my heartfelt thanks to all the members of the N420 lab, Joy, Han-Han, Ang-Chu, Lun-Go, Wen-Chien, A-Wei, Willy, Jack, Xiao-Jie, and Yuan-Yuan, for welcoming me into the group and helping throughout my research at Academia Sinica. Your insights, thoughtful suggestions, and willingness to patiently answer my many (and most of the time naive) questions have meant so much to me. I am also deeply thankful to my mentors in the N411 lab, Jeff, Kamal, and Chi-Yen for their dedication and guidance during my time here.

Last but not least, I want to thank my family for always being there for me. To Bong (Cotton), for telling your endless elementary school stories and bringing laughter even on the hardest days. To Linh Bio, for being my best friend and always listening without judgments. And to

doi:10.6342/NTU202502289

my parents, for giving me their support to chase whatever dreams I choose. To Mai Quynh ("ck tui" *), for her to believe in me even before I had the confidence to believe in myself.

(*: my family game but I actually feel this way.)



doi:10.6342/NTU202502289

摘要

在內共生過程中，共生體整合進宿主細胞會顯著改變基因表現與細胞身分。雖然選擇性剪接已知可透過快速調節基因表現與蛋白質同功異構體的多樣性來促進細胞適應，但其在內共生期間扮演的調控角色仍所知甚少。*Paramecium bursaria* 是研究此現象的理想模式生物，因其細胞質中攜帶數百個 *Chlorella variabilis* 藻類共生體。此外，*P. bursaria* 及其相關的 *Paramecium* 物種具有極短的內含子（中位數約為 24 個核苷酸），因此為探討短內含子動態剪接、內共生期間的調控模式，以及纖毛蟲內含子演化提供了寶貴的系統。在本研究中，我們對含共生體（綠色）與無共生體（白色）的 *P. bursaria* 細胞進行時間進程 RNA 定序，以分析內共生期間的內含子剪接模式。一般而言，5’端近端內含子的剪接，有助於提高基因表現量。我們發現綠色與白色的 *P. bursaria* 細胞之間存在差異性剪接的內含子，且在跨膜轉運蛋白基因中有顯著富集，這些基因對於內共生期間宿主與藻類共生體間的養分交換至關重要。此外，我們鑑定出一系列剪接增強子與抑制子，位於保守的剪接體基因中，其表現量與內共生期間的差異性剪接內含子密切相關。透過跨纖毛蟲的內含子直系同源性分析，我們發現保留的內含子具有較高的剪接效率、較低的 GC 含量以及一致的內含

doi:10.6342/NTU202502289

子長度，顯示新演化出的內含子會經過剪接與表現上的優化。我們的研究結果提供了選擇性剪接在宿主於內共生過程中適應角色的新見解，並解釋了短內含子在真核基因體中的演化動態。

doi:10.6342/NTU202502289

Abstract

The integration of symbionts into host cells during endosymbiosis significantly alters gene expression and cell identity. While alternative splicing facilitates cellular adaptation through rapid modulation of gene expression and protein isoform diversity, its regulatory role during endosymbiosis remains poorly understood. *Paramecium bursaria* represents an ideal model for investigating this, as it harbors hundreds of *Chlorella variabilis* algae within its cytoplasm. Furthermore, *P. bursaria* and related *Paramecium* species possess exceptionally short introns (median ~24 nucleotides), providing a valuable system for examining short intron splicing dynamics, regulatory patterns during endosymbiosis, and ciliate intron evolution. In this study, we conducted time-course RNA sequencing on symbiont-bearing (green) and symbiont-free (white) *P. bursaria* cells to analyze intron splicing patterns during endosymbiosis. While in general, splicing, especially 5' proximal introns, enhances gene expression, we identified differentially spliced introns between green and white, with an enrichment in transmembrane transporter genes, which are crucial for establishing nutrient transport between the host cell and its algal symbionts during endosymbiosis. Additionally, we identified splicing enhancers and repressors among conserved spliceosome genes, whose expression was closely linked to differentially spliced introns during endosymbiosis. Through intron orthology analysis across ciliates, we found that conserved introns are spliced more efficiently, with lower GC content and uniform length, suggesting that newly evolved introns undergo refinement to optimize splicing and expression. Our findings provide insight into the role of alternative splicing in host adaptation during endosymbiosis while shedding light on the evolutionary dynamics of short introns in eukaryotic genomes.

doi:10.6342/NTU202502289

Table of Contents

Acknowledgements ..... **i**

摘要 ..... **iii**

Abstract ..... **v**

Table of Contents ..... **vi**

Introduction **1**

Materials and Methods **6**

Paramecium bursaria strain and culture conditions 6

Genome annotation file and intron annotation 7

RNA extraction and RNA sequencing 7

Gene expression analysis 8

Alternative splicing event analysis, intron retention rate (PSI) calculation & differential spliced intron (DSI) clustering 8

GO terms enrichment analysis 9

Spliceosome comparison using reciprocal best BLAST hit and snRNA finding in *P. bursaria* 10

Linear regression analysis of the relationship between PSI and splicing-related gene expression in *P. bursaria* 11

Intron orthologs between ciliate species and intron ages assignment 11

Data availability 12

Results **12**

Extremely short introns in *P. bursaria* exhibit enhancement effect on gene expression level 12

Conservation of the U2-dependent spliceosome in *P. bursaria* 14

Intron retention and alternative 3' splice site are the most abundant alternative splicing events in *P. bursaria* 16

Intron retention rates are associated with GC content, intron length and gene expression levels 17

Two distinct intron splicing patterns between symbiotic and aposymbiotic cells and their

doi:10.6342/NTU202502289

association with the expression of splicing factors in symbiotic cells.....19
Intron evolution in ciliate species reveals higher splicing efficiency and lower intron GC content in aged introns.....22
Alternative splicing as a mode for gene regulation in endosymbiosis.....25
Splicing factors specifically regulate DSIs between symbiotic and aposymbiotic cells 26
Intron evolution is associated with gene expression regulation.....28
References.....31
Figures 44
Tables.....66

doi:10.6342/NTU202502289

Introduction

Molecular basis of RNA splicing

RNA splicing is a critical post-transcriptional process in eukaryotes, involving exon ligation and intron removal. Pre-mRNA splicing proceeds through the stepwise assembly of the spliceosome on splice junctions. The U2-dependent splicing pathway is described in detail by Yang et al. (2022) (Yang, Beutler et al. 2022). The major components of the U2-dependent spliceosome include five essential small nuclear ribonucleoproteins - U1, U2, U4/U6, and U5 snRNPs, composed of their associated U-rich small nuclear RNAs (UsnRNAs) and other associated proteins. In human cells, the 5' splice site (5'SS), branch point site (BPS), and 3'SS are recognized by U1 snRNP, SF1, and U2AF, respectively, forming the E complex. SF1 is then displaced by U2 snRNP to form the A complex. The U4/U6.U5 tri-snRNP subsequently joins to create the B complex. Rearrangements of the B complex together with displacement of U1 and U4 snRNPs lead to the formation of the active B complex and then the B* complex, which catalyzes the first step of splicing, resulting in the C complex. Further rearrangements form the C* complex, which carries out the second catalytic step, ligating the exons, and allowing spliceosome components to be recycled.

Regulated alternative splicing enriches proteome diversity

Alternative splicing (AS), which allows for various combinations of exon inclusion or exclusion, intron retention, and the use of alternative splice sites, enables a single gene to produce multiple protein isoforms. In humans, for example, over 90% of multi-exon genes undergo alternative splicing, generating diverse proteins that play different roles in a wide

doi:10.6342/NTU202502289

range of cellular processes (Pan, Shai et al. 2008). There are five primary types of alternative splicing events: exon skipping, alternative 3' splice site (A3SS), alternative 5' splice site (A5SS), mutually exclusive exons (MXE), and intron retention (IR) (Verta and Jacobs 2022). Alternative splicing isoforms serve main functions: regulation at the transcriptional level and increasing proteome diversity. When alternative splicing introduces premature termination codons (PTCs), it activates the nonsense-mediated decay (NMD) pathway, leading to mRNA degradation (Kervestin and Jacobson 2012). The NMD pathway has been shown to regulate transcript levels and ensure the fidelity of the splicing process (Alonso 2005, Palacios 2013, Miller and Pearce 2014). In contrast, in-frame alternative splicing events that do not generate PTCs result in changes to protein sequences and structures, thereby altering protein activity or cellular localization (Birzele, Csaba et al. 2008, Kjer-Hansen and Weatheritt 2023). For example, alternative splicing of exons 6 and 7 in the cell division cycle protein 42 (CDC42) affects its lipid modification, thereby influencing its distribution in neuronal cell compartments (Lee, Zdradzinski et al. 2021). Similarly, exon skipping in the apoptotic protein Caspase-9 generates a short isoform, Caspase-9S, which competes with the long isoform to inhibit apoptosis (Seol and Billiar 1999).

Alternative splicing has been shown to be crucial for regulating various biological processes, including cell identity and circadian rhythms (Nilsen and Graveley 2010, Kalsotra and Cooper 2011, McGlincy, Valomon et al. 2012, Baralle and Giudice 2017). It plays a significant role in driving cell differentiation, with splicing patterns varying widely between tissues and cell types (Olivieri, Dehghannasiri et al. 2021, Zhang, Guo et al. 2025). For example, in humans, the Ribosomal Protein S24 (RPS24) +a-b+c isoform is the most dominant in epithelial cells, while it is barely detectable in non-epithelial cells (Olivieri,

doi:10.6342/NTU202502289

Dehghannasiri et al. 2021); Similarly, a previous study demonstrated that Mitogen-activated Protein Kinase 7 (MAP3K7) isoform usage shifts from a short isoform being predominant in undifferentiated progenitor cells to a long isoform being dominant in mature epidermal keratinocytes (Takashima, Sun et al. 2024).

Host adaptation during endosymbiosis through alternative splicing

The process of endosymbiosis involves the integration of symbionts into host cells, forming a mutually beneficial relationship. In this interaction, endosymbionts confer novel phenotypic traits to the host, enabling the exploitation of more diverse environmental resources, while in return, the symbionts rely on the host for essential nutrients (Archibald 2015). One of the most well-known and ancient examples is the "Endosymbiosis Theory", which explains the origin of mitochondria and chloroplasts through the integration of ancestral prokaryotes into early eukaryotic cells (Archibald 2015, Martin, Garg et al. 2015). This integration can fundamentally alter host cell biology: for example, by transforming photosynthetic cyanobacteria into chloroplasts within plant cells or driving the development of specialized organelles to accommodate symbionts (Archibald 2015, von der Dunk, Hogeweg et al. 2023). To maintain such symbiotic relationships, both host cells and endosymbionts must undergo extensive morphological and genomic adaptations, including horizontal gene transfer (HGT) and the evolution of mechanisms for nutrient and resource exchange (Keeling 2013, Martin, Garg et al. 2015, Wilson and Duncan 2015, Kelly 2021, Bennett, Kwak et al. 2024). Although previous research has demonstrated gene family expansion and shifts in gene expression related to host stress responses and metabolism during endosymbiosis (Kelly, Carlson et al. 2025), the contribution of alternative splicing to these adaptive processes is not well understood. Given its role in rapidly modulating gene

doi:10.6342/NTU202502289

expression and expanding proteomic complexity, alternative splicing may be a key, yet underexplored, mechanism in host adaptation to symbiotic integration.

RNA splicing regulation in Paramecium

Paramecium bursaria reflects exactly as a model to study endosymbiosis. *P. bursaria* stably hosts hundreds of algae cells in its cytoplasm, the algae are shielded from lysosome digestion by perialgal vacuole (PV) beneath host cell membrane (Fujishima and Kodama 2012, He, Wang et al. 2019, Jenkins 2024). *P. bursaria* green cells are typically increase in size, growth rate, decrease in mitochondria number beneath cell membrane, where the PV located (He, Wang et al. 2019, Kodama and Fujishima 2022). Comparative genomics studies with relative species showed that *P. bursaria* increase genes encoded nitrogen and mineral metabolism, suggesting changes due to nutrients exchange between algae and the host cells (He, Wang et al. 2019). Another analysis of differential gene expression between green and white *P. bursaria* cells showed downregulation of redox, aminotransferase and ribosomal proteins in host cells and upregulation in Hsp70, Myb transcriptional factor and histidine kinase pathway (Yuuki Kodama 2014). Even though there were studies in transcriptional regulation in endosymbiosis (Yuuki Kodama 2014, Suzuki, Ishida et al. 2016, Ferrarini, Vallier et al. 2023, Abresch, Bell et al. 2024), the functional role alternative splicing regulation in endosymbiosis host cell changes remains unexplored. Studying alternative splicing in *P. bursaria* has the potential to further explore the complexity of mechanism that drives host cell adaptation in endosymbiosis process.

Given its importance, RNA splicing is tightly regulated by cis-acting regulatory sequences and trans-acting splicing factors (Hang, Wan et al. 2015, Plaschka, Lin et al. 2018, Chao, Jiang et al. 2021). Cis-acting elements include exonic splicing enhancers (ESEs),

doi:10.6342/NTU202502289

intronic splicing enhancers (ISEs), exonic splicing silencers (ESSs), and intronic splicing silencers (ISSs). Enhancer elements are typically bound by splicing activators such as serine/arginine-rich (SR) proteins, which promote splicing; in contrast, silencer elements are bound by repressor proteins such as heterogeneous nuclear ribonucleoproteins (hnRNPs), leading to splicing inhibition (Wang, Xiao et al. 2006, Wang and Burge 2008, Wang, Liu et al. 2015). Since alternative splicing is regulated in a cell type-specific manner (Boutz, Stoilov et al. 2007, Yang, Hung et al. 2014, Olivieri, Dehghannasiri et al. 2021, Zhang, Guo et al. 2025), investigating the relationship between splicing factors and splicing efficiency in *Paramecium bursaria* may help uncover key regulators involved in the endosymbiosis process.

Evolution of extremely short introns in ciliates

Intron has a crucial role in evolutionary aspects due to its links with the evolution gene expression regulation. The characteristics of introns vary across species, with lengths ranging from 15 nucleotides in *Stentor coeruleus* to over 1 million nucleotides in humans (Yu, Yang et al. 2002, Nuadthaisong, Phetruen et al. 2022). Although most modern vertebrate species exhibit a bimodal distribution of intron lengths (Yu, Yang et al. 2002), ciliate species such as *Paramecium* and *Tetrahymena* are notable for harboring predominantly extremely short introns, with the majority being under 100 bp. This distinct intron architecture makes ciliates excellent models for studying the evolution of short introns splicing (Bondarenko and Gelfand 2016). Despite their extremely short sequences, ciliate intron splicing plays a significant role in gene regulation (Saudemont, Popa et al. 2017, Gnan, Matelot et al. 2022, Ryll, Rothering et al. 2022). Comparative studies of intron evolution across ciliates provides valuable insights into the evolutionary history and functional adaptation of short introns and

doi:10.6342/NTU202502289

their relationship with gene regulation.

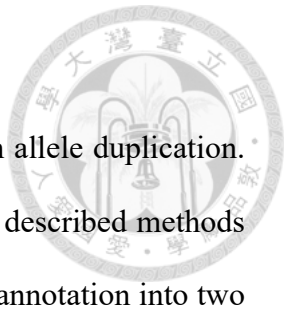
In this study, we examined time-course RNA sequencing data from both symbiont-bearing (green) and symbiont-free (white) *P. bursaria* cells to quantify alternative splicing across all stages. This allowed us to uncover the relationship between splicing efficiency and gene expression in *P. bursaria*. Differential analysis of intron retention rates between green and white cells over time revealed specific splicing regulation, particularly in genes associated with transmembrane transporter activity, functions that are closely linked to the nutrient exchanges during endosymbiosis. Furthermore, through comparative intron orthology across Paramecium species, we discovered that intron length and GC content have evolved into optimal splicing efficiency. Collectively, our findings highlight an evolutionary refinement of intron architecture to enhance splicing precision and metabolic fitness in the context of endosymbiotic adaptation.

Materials and Methods

Paramecium bursaria strain and culture conditions

This study used *P. bursaria* DK2 strain, which harbors the endogenous endosymbiont is *Chlorella variabilis*. DK2 is an offspring of the Dd1 and KM2 strains, and the detailed method for creating DK2 has previously been described (Cheng, Liu et al. 2020). *P. bursaria* cells were cultured on 2.5 % Boston lettuce in Dryl's solution medium and fed with *Klebsiella pneumoniae* (NBRC 100048 strain). Cell cultures were maintained at 23 °C with a 12-hour light/dark cycle. Every two days, bacteria-containing lettuce media was refreshed. To produce aposymbiotic (white) *Paramecium* strains, we used cycloheximide (10 μg/ml) to treat green cells as previously described (Kodama, Inouye et al. 2011).

doi:10.6342/NTU202502289



Genome annotation file and intron annotation

Like other ciliate species, *P. bursaria* has a diploid genome with allele duplication. The genome annotation file (GTF) was assembled following previously described methods (Cheng, Liu et al. 2020). We successfully separated the diploid genome annotation into two haplotypes and identified 14920 functional genes as representatives. Introns positions were determined using in-house R script, by extracting the sequences located between two consecutive exons.

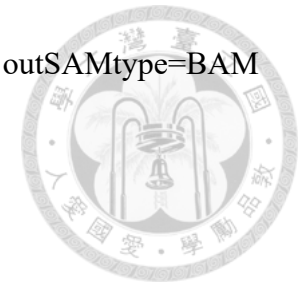
RNA extraction and RNA sequencing

For the RNA sequencing experimental design, *P. bursaria* green (algae-bearing) and white (algae-free) cells are harvested every three hours (AM2, AM5, AM8, AM11, PM2, PM5, PM8, PM11) for RNA extraction, with total 16 samples, each sample has 3 replicates. For RNA extraction, approximately 10^5 *P. bursaria* cells was collected in early stationary phases, washed twice with 1x Dryl's buffer, and concentrated using an 11-μm-pore-size nylon membrane. The RNeasy Mini Kit (Cat No. 74106, QIAGEN) and TRI Reagent (T9424, Sigma-Aldrich) were used to extract total RNA. An Illumina NextSeq platform was used to sequence the RNA sequencing libraries after they were generated using the Illumina TruSeq Stranded mRNA LT Sample Prep Kit (Illumina, San Diego, CA, USA).

The adapters and low quality reads were trimmed using fastp for paired-end reads with options cut_front_window_size=3, cut_tail_window_size=3, cut_right_window_size=4 , cut_right_mean_quality=30, length_required=36 (Chen, Zhou et al. 2018). Then, the trimmed reads are aligned to *P. bursaria* genome using STAR 2.7.11a to create BAM files with options twopassMode =Basic, overSJfilterOverhangMin=7 5 5 5, outSAMstrandField=intronMotif, alignIntronMin=10, alignIntronMax=10000,

doi:10.6342/NTU202502289

alignEndsType=EndToEnd, outSAMmapqUnique=3, outSAMtype=BAM SortedByCoordinate (Dobin, Davis et al. 2013).



Gene expression analysis

Salmon 0.13.1 was used to calculate raw gene read counts for each sample using quality-controlled FASTQ files, option gcBias, and numBoostraps 200 (Patro, Duggal et al. 2017). The tximport software was used to import raw gene counts from Salmon into R, which were then utilized in DESeq2 to normalize gene expression (Love, Huber et al. 2014, Soneson, Love et al. 2015). The differentially expressed genes (DEGs) were then analyzed using DESeq2 by computing the Log2FC of genes between conditions and the FDR of the difference in expression level significance (Love, Huber et al. 2014). For DESeq2 to continue processing, at least in three replicates, the number of normalized reads for each gene must be more than 10. If Log2 fold change ≥ 1 and FDR ≤ 0.05, the gene is defined as DEG between two groups considered.

Alternative splicing event analysis, intron retention rate (PSI) calculation & differential spliced intron (DSI) clustering

rMATs (replicate multivariate analysis of transcript splicing) was used to quantify AS events including SE (exon skipping), IR (intron retention), MXE (mutually exclusive exons), A3SS (alternative 3' splice site) and A5SS (alternative 5' splice site) (Shen, Park et al. 2014). Aligned BAM files from STAR were used as input for rMATs with options novelSS, allow-clipping, and mil=10 (minium intron size) to identify the AS events between green and white cells. We filtered the splicing efficiency difference on $|\Delta PSI| \geq 0.1$, Junction

doi:10.6342/NTU202502289

Read Count (inclusion read + exclusion read) >10 and FDR < 0.05 to be the differentially spliced introns between groups.

The individual intron retention rate was determined using SQUID (https://github.com/Xinglab/SQUID). PSI was calculated using the formula: total inclusion reads divided by total number of junction reads (Li, Wang et al. 2020).

$$\text{PSI} = \frac{\text{Inclusion read.}\ 0.5}{\text{Inclusion read.}\ 0.5\ + \text{Spliced read}}$$

The PSI of introns will only be included in the analysis if the number of Junction Read Counts is 10 or above, and the standard deviation for PSI for three replicates is less than 0.1 in all the replicates of the sample.

The PSI of each alternatively spliced intron between symbiotic and aposymbiotic cells is compared pairwisely at each time point using SQUID. Significant introns with |ΔPSI| > 0.1 and FDR < 0.05 between green and white cells were selected for further investigation.

GO terms enrichment analysis

GO IDs linked with each gene ID of *P. bursaria* were identified using InterProscan 5.72-103.0, which assigned each gene to GO terms based on its protein domains (Jones, Binns et al. 2014). The clusterprofiler R tool was used for GO enrichment analysis and visualization (Yu, Wang et al. 2012). Genes with False Discovery Rate (FDR) < 0.05 were considered significantly enriched in the gene set compared with the background gene set. Enrich score was calculated based on the ratio of genes in gene set divided by the ratio of genes in the background gene set.

doi:10.6342/NTU202502289

Spliceosome comparison using reciprocal best BLAST hit and snRNA finding in *P. bursaria*

To compare spliceosomal components between species, data on spliceosomal proteins in Homo sapiens was downloaded from the Spliceosome Database as the reference (Cvitkovic and Jurica 2013). The protein fasta files of other model species (*Mus musculus, Danio rerio, Drosophila melanogaster, Caenorhabditis elegans, Saccharomyces cerevisiae, Arabidopsis thaliana*) were downloaded from ENSEMBL database (Howe, Achuthan et al. 2021). P. caudatum and P. tetraurelia protein fasta files were downloaded from ParameciumDB (Arnaiz, Van Dijk et al. 2017); and protein fasta file of T. thermophila was obtained from TGD Wiki (Stover, Punia et al. 2012). The homologs of spliceosomal proteins with humans were then identified by utilizing reciprocal BLAST against the protein fasta files of each species (Ward and Moreno-Hagelsieb 2014). To improve specificity and sensitivity, RBBHs (reciprocal best BLAST hit) were allocated using both blastp and mmseq easy-rbh.

The UsnRNA multiple sequence alignment in Stockholm format was retrieved from the Rfam database 14.10 (Griffiths-Jones, Bateman et al. 2003). The INFERNAL program v1.1.5 was used to create a covariance model, calibrate the model, and search the UsnRNA sequence against the *P. bursaria* genome with cmbuild, cmcalibrate, and cmsearch, respectively (Nawrocki and Eddy 2013). The significant hit of UsnRNA sequence discovered by Infernal was extracted with BEDTools intersect (Quinlan and Hall 2010). The sequences of each UsnRNA candidate were then searched in R2DT with their corresponding templates to obtain secondary structures in DBN format (Sweeney, Hoksza et al. 2021). The structure was then annotated in the RNAcanvas (Johnson and Simon 2023).

doi:10.6342/NTU202502289

Linear regression analysis of the relationship between PSI and splicing-related gene expression in *P. bursaria*

To elucidate the potential regulation effect of 147 splicing-related DEGs on 992 DSIs, we performed a linear model:

$$\mathrm{PSI} = \beta_1 \,.\ln(\text{SF normalized counts} + 1) + \beta_2 \cdot (\text{cell type}) + \beta_0 + \varepsilon$$

PSI values and ln(SF normalized counts + 1) were standardized using z-transformation across 48 replicates from 16 samples. The significance of the regression coefficients β_1 and β_2 were assessed with FDR correction at a threshold of < 0.1. Since DSI values and splicing-related DEGs were identified from comparisons between different endosymbiosis states (green and white cells), we accounted for potential cell type confounding by setting β_1 to 0 when β_2 was found to be significant in the linear model. We interpreted a significant nonzero of β_1 as indicative of a potential regulatory effect of splicing-related gene expression on DSI splicing.

Intron orthologs between ciliate species and intron ages assignment

The process of identifying intron orthologs was adapted from a previously described method (Olthof, Schwoerer et al. 2024). Initially, protein homologs were retrieved using mmseqs2 easy-rbh to obtain reciprocal best-hit (RBH) proteins between species (Steinegger and Soding 2017). Next, the exon-exon junction amino acid sequences, consisting of 10 amino acids on each side of the junction, were extracted to define intron positions. These intron position sequences from protein homologs were then subjected to pairwise alignment using the pairwiseAlignment function in the Biostrings R package, using the BLOSUM-62 scoring system. Intron alignment were considered as RBHs if their aligned regions exhibited at least 40% sequence identity and achieved the highest alignment score in both forward and

doi:10.6342/NTU202502289

reverse directions.

For introns found in orthologous genes shared at least between three out of four species examined, their evolutionary age is determined using phylostratigraphy and the maximum parsimony method (Domazet-Loso, Brajkovic et al. 2007, Kannan and Wheeler 2012). The youngest intron group, assigned an age of 1, is specific to the reference species, while the oldest intron group consists of introns originated from first node in the phylogenetic tree examined.

Data availability

The transcriptomic data generated and analyzed during the current study are available in NCBI under the accession number BioProject PRJNA1279681 (https://dataview.ncbi.nlm.nih.gov/object/PRJNA1279681?reviewer=i5u6euocb0lp8fe6nc9e69ht6g).

Results

Extremely short introns in *P. bursaria* exhibit enhancement effect on gene expression level

To characterize the features of *P. bursaria* introns, we annotated 39,715 introns in the functional gene set from the DK2 genome annotation file (see Materials and Methods for details). More than 80% of genes contain at least one intron, with an average of 2.7 introns per gene (Figure 1A). Moreover, we discovered a unique distribution of introns in *P. bursaria* transcripts, which are enriched at both 5' and 3' end of the gene (Figure 1B). This bias of introns towards 5' end of transcript also observed in other model species, such as the budding yeast, fruit flies and mice (Sakurai, Fujimori et al. 2002, Lin and Zhang 2005), suggesting a

doi:10.6342/NTU202502289

general feature of eukaryotic introns. The majority of *P. bursaria* introns are less than 40 nucleotides (nt), with a median length of 24 nt, and minimum and maximum lengths of 15 and 100 nt, respectively (Figure 1C). The intron length distribution is very similar to that of *Paramecium tetraurelia*, which has a median length of 25 nt (Arnaiz, Van Dijk et al. 2017). The sequence logo plot shows that most of the annotated introns in *P. bursaria* are canonical introns with the conserved 5' and 3' splice boundaries GT-AG, suggesting an exclusive U2-dependent splicing mechanism (Figure 1D). While *P. bursaria* exon GC content is about 30% (average previous exon GC content = 30.7%, average next exon GC content = 30.5%), the GC content of introns (average = 17.6%) is substantially lower (Figure 1E).

Because the endosymbiosis of *P. bursaria* and algae (green cells) is influenced by the circadian clock and light conditions due to photosynthesis, we explore the regulation of splicing efficiency during the endosymbiosis process in relation to the circadian cycle. We compared the splicing efficiency of all introns across green and white cells, and across 8 timepoints through a light-dark cycle: 8AM, 11AM, 2PM, 5PM, 8PM, 11PM, 2AM, and 5AM. In eukaryotes, intron splicing plays a crucial role in gene expression regulation through the co-transcriptional splicing mechanism (Shaul 2017). We investigated whether small introns in *P. bursaria* have similar functions. Transcriptomic data from all 16 samples were pooled together and analyzed. We observed that genes containing introns exhibited significantly higher expression levels than those without introns (Wilcoxon rank-sum test, $p < 0.001$). Moreover, the expression level was positively correlated with the intron number (Figure 2A).

The positive correlation between intron number and gene expression levels in *P. bursaria* suggests the occurrence of intron-mediated enhancement (IME). Previous studies

doi:10.6342/NTU202502289

demonstrated that U1 snRNP can directly recruit transcription factors, such as TFIIH, TFIIB, and TFIID, to the promoter when the intron 5' splice site (5'SS) of introns is located near the promoter. (Das, Yu et al. 2007, Damgaard, Kahns et al. 2008). Since we observed a bias of intron position toward the 5' end of transcript in *P. bursaria* (Figure 1B), the presence of introns in this position is likely associated with enhanced gene expression in *P. bursaria*. To investigate that, we grouped and compared *P. bursaria*'s genes based on the relative position of the first intron: the 5' end group (i.e., the first intron is located in 0-25% of the total length), the middle group (25-75%) or the 3' end group (75-100%). Indeed, the group of genes with the first intron in the 5' end has significantly higher expression than the other two groups (Figure 2B).

Conservation of the U2-dependent spliceosome in *P. bursaria*

Compared to other eukaryotes with well-characterized splicing mechanisms (Zhu, He et al. 2010, Yang, Beutler et al. 2022), the intron size of *P. bursaria* is relatively small (Figure 1C). Spliceosome are huge protein complexes containing many snRNAs and associated proteins. According to Cryo-EM structure, human spliceosome A complex is ~205Å×195Å×150Å in size and occupies 79-125 nt of a stretched RNA (Behzadnia, Golas et al. 2007). The *P. bursaria* spliceosome likely requires extensive reorganization to function on small introns. To investigate this, we downloaded 1005 human splicing-related proteins from SpliceosomeDB (Cvitkovic and Jurica 2013). Using mmseqs2, we identified orthologs across three representative ciliate species (*P. bursaria*, *P. tetraurelia*, and *T. thermophila*), along with other model organisms, including *M. musculus*, *D. rerio*, *D. melanogaster*, *C. elegans*, *S. cerevisiae* and *A. thaliana* (Supplementary Table 1), using *H. sapiens* as the reference species (Steinegger and Soding 2017). Following SpliceosomeDB, we categorized

doi:10.6342/NTU202502289

human splicing-related proteins into Splicing Protein Classes. Among all species analyzed, 275 out of 1005 spliceosomal proteins were shared, representing the core components of spliceosome (Figure 3A).

Fourty splicing-related proteins are specifically lost in ciliate species, including a U1 snRNP protein (LUC7L), a U5 snRNP protein (CD2BP2), and an LSm protein (LSM1) (Figure 3A, Supplementary Table 2). In human cells, LUC7L enhances splicing by recognizing a strong consensus sequence downstream of the 5' splice site (within the intron), whereas its paralog LUC7L3 promotes splicing through interaction with an upstream consensus sequence located within the exon (Kenny, McGurk et al. 2025)(Daniels et al., 2021). The loss of LUC7L and the retention of LUC7L3 in *P. bursaria* corresponds with a conserved upstream consensus (positions -2 and -1 in the exon) and the absence of downstream consensus (positions +3 to +6 in the intron) to the 5' splice site (Figure 1D), implying a streamlined mechanism of 5' splice site recognition by U1 snRNP, likely a consequence of reduced splicing machinery in ciliates.

Interestingly, we observed a complete loss of the U11/U12 snRNP complex in the ciliate species group (Figure 3B). Additionally, there was a reduction in the number of proteins associated with regulatory classes, including SR proteins, hnRNPs, CSTF, and the EJC complex (Figure 3B). Notably, although the numbers of splicing-related proteins in ciliates and *S. cerevisiae* are pretty similar, different sets of proteins are uniquely lost in yeast and the ciliate species (Figure 3A, Supplementary Table 2, Supplementary Table 3). This likely reflects different directions of spliceosome reorganization, one for the reduction in intron number in *S. cerevisiae* and another for reduction in intron size in ciliates.

Overall, the numbers of U2-dependent core spliceosome proteins in *P. bursaria*,

doi:10.6342/NTU202502289

including U1, U2, U4, U5, U6, Sm, and Lsm proteins, are highly conserved. To further investigate the snRNA repertoire in *P. bursaria*, we analyzed conserved secondary structures of UsnRNA from the Rfam database (Griffiths-Jones, Bateman et al. 2003). Using Infernal, we searched for significant matches between *P. bursaria* RNA sequences and Rfam UsnRNA data, detecting U1, U2, U4, U5, and U6 snRNA with conserved secondary structure and key functional motifs, but not U11 and U12 snRNA (Figure 4). Together, our data suggest that a modified U2-dependent splicing mechanism is utilized to splice small introns in *P. bursaria*.

Intron retention and alternative 3' splice site are the most abundant alternative splicing events in *P. bursaria*

Since alternative splicing influences both protein isoform diversity and gene expression levels, it may play a key role in the adaptive regulation of the endosymbiosis process. To explore this, we used rMATs to analyze the regulation of five major types of alternative splicing events between symbiotic and aposymbiotic *P. bursaria* cells: Alternative 3' Splice Site (A3SS), Alternative 5' Splice Site (A5SS), Mutually Exclusive Exon (MXE), Intron Retention (RI), and Exon Skipping (SE), across eight timepoints (Figure 5). The results showed that among these five types, A3SS and RI events were the most frequently regulated between symbiotic and aposymbiotic cells (Figure 6A). We then filtered significant A3SS and RI events (FDR < 0.05) using junction read counts (total junction reads ≥ 10) and |PSI| ≥ 0.1, identifying 512 significant A3SS events (Figure 6B) and 992 significant RI events (Figure 6C).

Interestingly, the number of significant A3SS and RI events are greater at night (PM11, AM2, AM5, AM8) than during the day (AM11, PM2, PM5, PM8) (Figure 6B, 6C).

doi:10.6342/NTU202502289

Notably, during night timepoints, a higher ratio of A3SS events showed increased PSI in symbiotic cells, while the direction of change was evenly distributed between symbiotic and aposymbiotic cells (Figure 6B). In contrast, RI events exhibited a higher proportion of increased PSI in aposymbiotic cells at night, and also showed the overall balance in directional changes between the two cell types in daytime (Figure 6C). Next, to assess the functional impact of the alternative splicing events, we investigated the impact of A3SS events on translational reading frame by calculating the distance between canonical 3'SS and alternated 3'SS. Unexpectedly, we found that the majority of the distance is non 3n (n= 479 events out of 512 events), which cause the frameshift in open reading frame (Figure 6D). Unlike the 3' wobble splicing observed in mammalian cells (e.g., NAGNAG motifs), which often preserves functional reading frames from both 3'SS choices (Hiller, Huse et al. 2004, Akerman and Mandel-Gutfreund 2006), A3SS selection in *P. bursaria* predominantly serves as a mechanism to abrogate protein production. In contrast, the proportion of non-3n intron sequences among significantly retained introns (74.40%) closely mirrors that of the genome-wide intron pool (74.38%), indicating no apparent selection against frameshift-inducing introns in retained intron (RI) events during endosymbiosis. In conjunction with the observed changes in gene expression associated with retained introns (Figure 2), conventional intron splicing appears to promote gene expression by facilitating productive transcript maturation and contributing to alternative transcriptome diversity.

Intron retention rates are associated with GC content, intron length and gene expression levels

While A3SS regulation during endosymbiosis mainly causes frameshifts that may

doi:10.6342/NTU202502289

trigger mRNA degradation via the NMD pathway, intron retention events are more likely to play roles in regulating gene expression and protein isoforms ratio in endosymbiosis. Therefore, we began to examine intron retention rates more closely, focusing on both global introns and Differentially Spliced Introns (DSIs) between symbiotic and aposymbiotic cells.

After filtering out introns with low junction read counts (i.e., read count ≥ 10), 66% of total introns (26123 out of 39715) have sufficient counts for calculating the splicing efficiency in at least one sample. More than 87% of these introns have intron retention rates (PSI) lower than 0.1 for all samples, indicating that most introns in *P. bursaria* are spliced effciently (Figure 7A and Table 1).

Next, we investigated the intrinsic factors that can influence intron splicing efficiency. We observed a strong positive correlation between the GC content of introns and intron retention rates (Figure 7B). Although a similar trend has been observed in *P. tetraurelia* (Gnan, Matelot et al. 2022), the effect of the GC content on intron retention is much stronger in *P. bursaria.* To investigate whether intron length influences splicing efficiency in *P. bursaria*, we categorized introns into three length groups: 15–22 nucleotides (representing the shortest 25%), 23–26 nucleotides (within the interquartile range), and those longer than 27 nucleotides (representing the longest 25%). Notably, we found that introns within the interquartile range exhibited the highest splicing efficiency compared to both shorter and longer introns (Figure 7C).

As shown earlier in Figure 2, genes with a higher intron count and the presence of a first intron near the 5' end are strongly linked to elevated expression levels. In addition, we also observed a negative correlation between the expression level and intron retention rate (Pearson correlation r = -0.31, p value <2.2e-16) (Figure 7D), suggesting that efficient

doi:10.6342/NTU202502289

splicing contributes to transcript abundance. Moreover, the 5' end introns are spliced more efficiently compared with the middle introns (Figure 7E). These results showed that the intron splicing, especially of the introns located in the 5' end of transcripts, are associated with enhanced gene expression in *P. bursaria*, as observed in mammalian cells (Nott, Meislin et al. 2003, Park, Hannenhalli et al. 2014).

Two distinct intron splicing patterns between symbiotic and aposymbiotic cells and their association with the expression of splicing factors in symbiotic cells

To examine the pattern of these DSIs across all samples, we compiled DSIs from all timepoints, resulting in a total of 992 DSIs in 883 genes (Supplementary Table 4). PCA analysis was performed to cluster the samples (16 samples, each with three replicates) based on the PSI values of these DSIs. The PSI profiles of DSIs effectively distinguished symbiotic from aposymbiotic cells (Figure 8A), indicating that endosymbiosis influences the splicing pattern. Additionally, symbiotic cells formed two sub-clusters along PC2, one corresponding to nighttime, while the other includes cells from the daytime and the first nighttime time point (PM11) (Figure 8A). This may be caused by photosynthesis of endosymbiotic algae.

k-means clustering separated the identified DSIs into two clusters with distinct intron PSI patterns (Figure 8B, Supplementary Table 4). Cluster 1 consists of DSIs with higher PSI in symbiotic cells (n = 386 introns), whereas cluster 2 (n = 606 introns) includes DSIs with higher PSI in aposymbiotic cells (Figure 8B). Gene ontology (GO) enrichment analysis of genes containing DSIs revealed an overrepresentation of transport related terms including organic anion transmembrane transporter activity, organic anion transport, carboxylic acid

doi:10.6342/NTU202502289

transmembrane transport and mRNA splicing, via spliceosome (Figure 8C, Supplementary Table 5). Previous studies have shown that endosymbiotic algae are maintained in a membranous compartment (i.e., perialgal vacuole) inside *P. bursaria* host cells, and the host and endosymbionts continuously exchange several organic compounds to establish a stable mutualistic relationship (Kodama and Fujishima 2010). Our data suggest that intron splicing may play a pivotal role in regulating these transmembrane exchanges. Moreover, the regulation of intron retention rates in spliceosome genes suggests potential autoregulation of splicing-related genes during endosymbiosis, which may in turn influence the retention levels of other introns.

In our transcriptome-wide analysis, we observed that intron retention rate is negatively correlated with gene expression (Figure 7D). Consistently, we observed a similar negative correlation trend between gene expression and the PSI values of DSIs, with approximately 72% of significant Pearson correlations being negative (Figure 8D). To demonstrate this relationship between DSIs and their gene expression, we performed Pearson correlation test between DSIs and their gene expression and illustrate the top three correlation coefficient in Figure 8E. As expected, in cluster 1 DSIs (Figure 8E top), the PSI of symbiotic cell is lower and the gene have higher expression level compared with aposymbiotic cell. In contrast, in cluster 2 DSI (Figure 8E bottom), the PSI of white cell is lower and the gene have higher expression level compared with green cell. In summary, our data reveals white and green cells-specific splicing response in relation to circadian rhythms.

Certain splicing factors can specifically regulate introns of genes involved in cellular pathways. For instance, HNRNPK and SRSF1 had been shown to control intron retention during B cell development (Ullrich and Guigo 2020). Since the intron splicing efficiency was

doi:10.6342/NTU202502289

found to be regulated in endosymbiosis, potential splicing factors may influence the splicing of DSIs during this process. To explore this, we analyzed 6,949 differentially expressed genes (DEGs) between symbiotic and aposymbiotic cells across all time points (see Materials and Methods for details). Notably, k-means clustering revealed three distinct expression patterns among these DEGs (Figure 9A). Cluster 1 showed high expression levels in white cells, cluster 2 consisted of genes highly expressed in green cells, and cluster 3 contained genes with elevated expression specifically in AM11 green cell samples. GO term enrichment analysis of cluster 1 genes indicated an association with microtubule movement activity, suggesting an increase in cell transport-related gene expression in white cells compared to green cells. Meanwhile, cluster 2 genes were enriched in terms related to mitochondrial membrane, oxidoreductase activity, and nitrogen metabolism, indicating a change of respiratory status and nitrogen metabolic processes in green cells. Furthermore, cluster 3 genes were enriched in cell cycle-related terms, suggesting that the symbiont algae play a role in shaping the light-driven cell cycle response of *P. bursaria* (Figure 9B, Supplementary Table 6).

We identified 147 splicing-related genes that were DEGs between green and white cells. Notably, we observed a significant overlap between spliceosomal DEGs with cluster 2 DEGs (p-value = 3.73e-08, Chi-square test). This significant overlap indicates that the differential expression of spliceosomal genes may play a role in regulating splicing patterns during endosymbiosis process.

To investigate this, we examined the association between expression levels of 147 splicing factors and intron retention rates of 992 DSIs using a linear regression model. As both splicing factor-DEGs and DSIs were identified through comparisons between symbiotic

doi:10.6342/NTU202502289

and aposymbiotic cells, endosymbiosis status may confound their associations. To account for this, we incorporated endosymbiosis status as a covariate in the model (DSI ~ splicing factor-DEG + endosymbiosis status) and retained only DSIs whose coefficients were significant with splicing factor-DEGs but not for endosymbiosis status (see Materials and Methods for details). We further excluded DSIs that did not show significant associations with any of the 147 candidate splicing factors. This filtering yielded 477 introns, which likely represent direct regulatory targets of these splicing factors (Figure 9C). These 147 splicing-related genes thus are classified into two groups based on their mean coefficient with DSIs' PSI: "splicing enhancers", whose mean coefficients with the PSI of DSIs is negative, and "splicing repressors", whose mean coefficients is positive with the PSI (Figure 9C). Notably, splicing-enhancing factors are significantly upregulated in symbiotic cells (Figure 9D). It suggests that symbiotic cells regulate these splicing-related factors to fine-tune the splicing and expression of intron-containing genes.

Intron evolution in ciliate species reveals higher splicing efficiency and lower intron GC content in aged introns

We observed that some intron characteristics, such as the GC content and the intron distribution within genes, are closely associated with the splicing efficiency and gene expression regulation (Figures 2 and 7). Moreover, the symbiotic cells may adjust the expression of specific splicing factors to regulate the splicing efficiency and the mRNA levels of a group of introns (Figure 9). Given the functional significance of introns, we next examined the evolutionary patterns of introns within the Paramecium genus.

For this analysis, we selected three *Paramecium* species with well-annotated genomes, including *P. bursaria*, *P. tetraurelia*, and *P. caudatum*, and *T. thermophila* as an

doi:10.6342/NTU202502289

outgroup. We identified 8,886 orthologs shared across these species. We then identified the best-matching introns from paired orthologous genes between *P. bursaria* and the other species (see Materials and Methods for details).

To classify intron origins, we applied phylostratigraphy and maximum parsimony, assigning introns to different nodes on the phylogenetic tree. Our findings revealed that 10.1% of introns in orthologous genes belong to the oldest group (n = 2,680 introns, age group = 3), while 43.3% are Paramecium-specific introns (n = 11,475 introns, age group = 2). The remaining 46.6% are unique to *P. bursaria* (n = 12,336 introns, age group = 1), highlighting the dynamic evolution of introns in this lineage (Figure 10A).

Interestingly, older introns (age groups 2 and 3) exhibit better splicing efficiency compared to newer introns (age group 1) (Figure 10B). In *P. bursaria*, high-GC introns generally have higher PSI values compared to low-GC introns. This aligns with our observation that new introns in *P. bursaria* have higher GC contents than older ones (Figure 10C). We also observed a trend where younger introns have a wider range of length distributions compared to those in the older groups. (Figure 10D). Consistently, our earlier analysis showed that introns longer than 27 nt or shorter than 23 nt are spliced less efficiently compared to those within the 23–26 nt range, suggesting that the intron length distribution in *P. bursaria* has evolved toward an optimal range over time (Figure 7C). Given the positive association between intron presence, intron number, and gene expression, we further analyzed gene expression in relation to intron age, reflecting when introns were acquired in the evolutionary tree. We categorized the 8,886 orthologous genes into three groups based on the age of their oldest intron: genes with a maximum intron age of 3 were assigned to group 3 (n = 1,656 genes), group 2 included genes with a maximum intron age of 2 (n = 4,863

doi:10.6342/NTU202502289

genes), and group 1 contained genes with only young introns (n = 1,466 genes). Notably, genes with older introns exhibited significantly higher gene expression levels compared to those with only young introns (Figure 10E). This finding further supports the role of introns in enhancing gene expression.

To further examine whether the evolutionary patterns observed in *Paramecium* introns are also present in other ciliates, we extended our analysis to *Tetrahymena*. We selected four representative species from the Tetrahymena genus, including *T. thermophila*, *T. malaccensis*, *T. eliotti* and *T. borealis*. Using the same approach as in *P. bursaria*, we classified introns in *T. thermophila* that had gene orthologs in at least three *Tetrahymena* species into four intron age groups (Figure 11A).

Through intron orthology analysis, we detected 60,971 intron orthologs distributed among 12,039 ortholog genes shared among *Tetrahymena* species. Our results revealed that the majority of introns in *T. thermophila* belonged to the oldest group, with 42,030 introns accounting for 68.9% of all introns. In contrast, species-specific introns (age group 1) in *T. thermophila* made up only 9.3% of the total introns in shared orthologous genes (n = 5,678 introns) (Figure 11B). We found that younger introns also exhibited a higher retention rate, a pattern consistent with what we observed in *P. bursaria* (Figure 11B). Similarly, young introns in *T. thermophila* had higher GC content and a wider length range compared to older introns (Figure 11C, D). When we categorized genes based on their oldest intron age in *T. thermophila*, we observed a significant trend where genes containing older introns exhibited higher expression levels compared to those with only younger introns (Figure 11E). Taken together, our analysis of intron features across different evolutionary age groups reveals a consistent pattern of intron optimization, characterized by improved splicing efficiency and

doi:10.6342/NTU202502289

enhanced gene expression, during evolution in both *Paramecium* and *Tetrahymena* lineages.

Discussion

Alternative splicing as a mode for gene regulation in endosymbiosis

Adaptation of host cells when integrating symbionts was found to involve drastic gene expression changes. A previous study in *P. bursaria* showed nearly 7000 DEGs between symbiotic and aposymbiotic cells, which is consistent with our result (Yuuki Kodama 2014). Other analysis indicated the upregulation of glutamine biosynthesis correlated with symbionts abundance (He, Wang et al. 2019). Our analysis has consistently shown the enrichment of nitrogen metabolism genes and mitochondrial transmembrane activity genes. This phenomenon is universal in algal secondary endosymbiosis, for example, in *Hydra viridissima* algal endosymbiosis, genes involved in glutamine synthesis and phosphate transporters are also upregulated (Hamada, Schroder et al. 2018). Alternative splicing regulation could give the host tools to quickly adapt in the endosymbiotic relationship, especially when this relationship is varied between timepoints, possibly due to change of photosynthesis in algal cells.

Recent study in *P. bursaria* showed that 6mA DNA methylation is involved in endosymbiosis, and knock-down 6mA methyltransferase genes leads to reduced IR events (Pan, Ye et al. 2023). That prompted together with DNA methylation, alternative splicing is important in endosymbiosis establishment. Our study showed there are 992 introns differentially spliced between symbiotic and aposymbiotic cells in 8 timepoints, providing the evidence of alternative splicing regulation in endosymbiosis. Furthermore, those introns' PSI showed a clear pattern with one cluster being higher in apposymbiotic cells, and another being timepoint specific in green cells. We also observed the fluctuation of some introns' retention rate is changed across timepoints, especially in symbiotic cells, implying the function of alternative splicing regulation in the circadian adaptation of host cell-symbiont

doi:10.6342/NTU202502289

relationship. Regarding the function of genes regulated by alternative splicing in endosymbiosis, we found the enrichment of transporter related terms, including “organic anion transmembrane transporter activity”, “organic anion transport”, “carboxylic acid transmembrane transport”. Other studies on *P. bursaria* endosymbiosis revealed that silencing the transmembrane transporter protein in green cells resulted in cell death, while knocking down glutamine synthetase in symbiotic cells led to a decline in algal cell numbers (He, Wang et al. 2019). This implied that alternative splicing regulation plays an important role in nutrients exchange between algae cells and host cells, potentially ATP or carboxylic acid transportation.

Splicing factors specifically regulate DSIs between symbiotic and aposymbiotic cells

Alternative splicing is strictly regulated by many factors, such as co-transcriptional, chromatin structure and epigenetic modification, DNA sequence modification and splicing factors (Chen and Manley 2009). In *P. tetraurelia*, nucleosome positioning affects intron splicing, introns at the edges of nucleosomes have higher splicing efficiency (Gnan, Matelot et al. 2022). 6mA DNA methylation in *P. bursaria* was shown to be enriched in retained introns (Pan, Ye et al. 2023). Alternative splicing factors are proved to be regulated in a tissue-specific manner in humans, for example, polypyrimidine tract binding protein 1 (PTB1) expression is high in neuron progenitor cells, but significantly downregulated in differentiated neurons (Boutz, Stoilov et al. 2007, Keppetipola, Sharma et al. 2012). Since endosymbiosis induces morphological changes in host cells, including the formation of new organelles for symbionts, alternative splicing regulation during this process may be influenced by specific splicing factors. Our analysis identified differentially expressed

doi:10.6342/NTU202502289

splicing-related genes between green and white cells, suggesting a potential role for these splicing factors in intron splicing during endosymbiosis. Furthermore, linear regression model between splicing factor DEGs and differentially spliced introns across 16 samples revealed that a subset of DSIs is highly associated with those splicing factor DEGs. Two groups of splicing factors exhibit two different coefficient direction with PSI: one group is negative, while the other is positive, possibly representing "splicing enhancers" and "splicing repressors" of those DSIs in endosymbiosis.

To pinpoint key splicing factors highly associated with DSIs' PSI, we ranked 147 splicing factor DEGs based on their number of significant coefficients with DSIs' PSI. Top 30 splicing-related genes includes core splicing factors, such as U1 snRNP, Sm proteins (SNRNP70, SNRPD1, SNRPC), RNA helicase (DDX23, DDX46), U5 snRNP (TXNL4A), second-step factors (PRPF18) and chromatin-related (SMARCA5, SMARCA1) (Figure 12).

The PSI of potential regulated introns is negatively associated with ribosomal protein expression, including RPL22, RPS20, RPS11. Besides the canonical function in ribosomal structure, RPLs are proved to have alternative splicing regulation activity, for example, rpl-10 has autoregulation splicing activity (Takei, Togo-Ohno et al. 2016). This is possible that regulation of DSIs is closely linked to translational process; and there are possibilities that ribosomal proteins actually have function in alternative splicing regulation by RNA bindings or by bindings with other splicing factors.

In the top 30 SFs associated with DSIs, the coefficient directions of some splicing factors align with their known roles in splicing regulation. For instance, PRPF18 plays a crucial role in maintaining high splicing fidelity during the second splicing step; G3BP2 interacts with PSF (polypyrimidine tract-binding protein-associated splicing factor) to

doi:10.6342/NTU202502289

enhance mRNA stability (Han, Liu et al. 2022, Roy, Gabunilas et al. 2023, Takayama, Suzuki et al. 2024). However, we observed that some gene coefficients suggest regulatory effects that differ from previous studies. For example, PRPF4B was found to phosphorylate SRSF1 protein in fission yeast (Chen, Moore et al. 2007), which is a splicing enhancer, yet we found the major effect of negative association with DSIs in *P. bursaria*. These findings indicate that splicing regulation in *P. bursaria*, particularly in the context of endosymbiosis, operates differently. Further research is necessary to explore the regulation of DSIs involved in endosymbiosis.

Intron evolution is associated with gene expression regulation

Our study showed a strong relationship between intron splicing efficiency and gene expression level. This result is consistent with other studies showing that short intron sequences in *H. sapiens* and *C. elegans* are associated with highly expressed genes (Castillo-Davis, Mekhedov et al. 2002). This association intrigued us to question about how these short introns were shaped under evolutionary pressure.

We categorized *P. bursaria* introns into three age groups based on their presence in the phylogenetic tree we examined. Our results revealed that older introns are spliced more efficiently than newly acquired introns in *P. bursaria*. This suggests that introns gained over time can adapt to the splicing system, potentially through sequence modifications or changes in splicing factor binding sites (Yeo, Van Nostrand et al. 2007, Schirman, Yakhini et al. 2021). Another possible explanation for the higher efficiency of older introns is that new intron sequences are more diverse in both sequence composition and secondary structure, necessitating adaptation by the splicing machinery. A study on *S. cerevisiae* intron splicing

doi:10.6342/NTU202502289

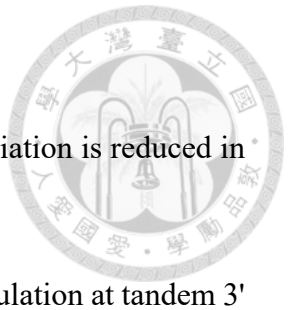
demonstrated that intron orthologs from species lacking U2AF1 and having shorter Branch Point Site (BPS) to 3'SS distance are spliced more efficiently in *S. cerevisiae* (which also lacks U2AF1 and has a short BPS to 3'SS distance) compared to intron orthologs from species that possess U2AF1 orthologs and longer BPS to 3'SS distance, highlighting the specialization of the splicing machinery for its own introns structure (Schirman, Yakhini et al. 2021). Our analysis consistently emphasized differences between new and old intron sequences, showing that new introns in *P. bursaria* have a higher GC content and are longer than older introns. Intron GC content was known to be negatively correlated with splicing efficiency in *P. tetraurelia* and *S. cerevisiae* (Schirman, Yakhini et al. 2021, Gnan, Matelot et al. 2022). Consistently, our findings indicate that introns with higher GC content have a higher retention rate, thus it is reasonable that new introns are spliced less efficiently. In terms of intron length, new introns are slightly longer than older ones in both *P. bursaria* and *T. thermophila*. These changes in new introns may contribute to intron sequence diversification, driving evolutionary shifts in intron features and prompting the necessary adaptations in splicing factors to accommodate variations in intron length and GC content.

We observed the higher expression of genes containing old introns compared with genes with only new introns. It was clear from our data that intron presence associates with higher gene expression levels. Through our intron orthology analysis in ciliate species, we found that some introns are highly conserved across species, raising questions about the potential role of these introns in fine-tuning gene expression during evolution. A previous study showed that genes retaining conserved introns in intron-poor species are enriched for functions related to ribosomal proteins and endosome organization, whereas genes that are depleted of introns are enriched in processes such as protein folding and small molecule

doi:10.6342/NTU202502289

synthesis (Lim, Weinstein et al. 2021). We found that genes with intron age 1 (introns only in *P. bursaria*) in *P. bursaria* are enriched with "phosphorelay signal transduction system" term (Figure 13A, Supplementary Table 7), which linked to the responses to wide ranges of environments (Koretke, Lupas et al. 2000). Moreover, we observed the enrichment of transmembrane transport and metal ion transportation-related terms in genes with intron age 2 (introns conserved in *Paramecium* species) (Figure 13B, Supplementary Table 7), which implied the specific change in expression of transportation-related genes in the *Paramecium* species group. Consistently, other research showed the increase of potassium ion channel gene numbers in the *Paramecium* group compared with other species (Haynes, Ling et al. 2003). Gaining introns in transportation genes emphasized another mechanism to upregulate the important genes in *Paramecium*. Genes gain with the oldest node (age = 3) are enriched in cilium terms (Figure 13C, Supplementary Table 7). This implies that the order of genes gaining introns affects gene expression regulation, and this regulation mechanism may appear when it is necessary for the species clade.

doi:10.6342/NTU202502289



References

Abresch, H., T. Bell and S. R. Miller (2024). "Diurnal transcriptional variation is reduced in a nitrogen-fixing diatom endosymbiont." ISME J **18**(1).

Akerman, M. and Y. Mandel-Gutfreund (2006). "Alternative splicing regulation at tandem 3' splice sites." Nucleic Acids Res **34**(1): 23-31.

Alonso, C. R. (2005). "Nonsense-mediated RNA decay: a molecular system micromanaging individual gene activities and suppressing genomic noise." Bioessays **27**(5): 463-466.

Archibald, J. M. (2015). "Endosymbiosis and Eukaryotic Cell Evolution." Curr Biol **25**(19): R911-921.

Arnaiz, O., E. Van Dijk, M. Betermier, M. Lhuillier-Akakpo, A. de Vanssay, S. Duharcourt, E. Sallet, J. Gouzy and L. Sperling (2017). "Improved methods and resources for paramecium genomics: transcription units, gene annotation and gene expression." BMC Genomics **18**(1): 483.

Baralle, F. E. and J. Giudice (2017). "Alternative splicing as a regulator of development and tissue identity." Nat Rev Mol Cell Biol **18**(7): 437-451.

Behzadnia, N., M. M. Golas, K. Hartmuth, B. Sander, B. Kastner, J. Deckert, P. Dube, C. L. Will, H. Urlaub, H. Stark and R. Luhrmann (2007). "Composition and three-dimensional EM structure of double affinity-purified, human prespliceosomal A complexes." EMBO J **26**(6): 1737-1748.

Bennett, G. M., Y. Kwak and R. Maynard (2024). "Endosymbioses Have Shaped the Evolution of Biological Diversity and Complexity Time and Time Again." Genome Biol Evol **16**(6).

doi:10.6342/NTU202502289

Birzele, F., G. Csaba and R. Zimmer (2008). "Alternative splicing and protein structure evolution." Nucleic Acids Res **36**(2): 550-558.

Bondarenko, V. S. and M. S. Gelfand (2016). "Evolution of the Exon-Intron Structure in Ciliate Genomes." PLoS One **11**(9): e0161476.

Boutz, P. L., P. Stoilov, Q. Li, C. H. Lin, G. Chawla, K. Ostrow, L. Shiue, M. Ares, Jr. and D. L. Black (2007). "A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons." Genes Dev **21**(13): 1636-1652.

Castillo-Davis, C. I., S. L. Mekhedov, D. L. Hartl, E. V. Koonin and F. A. Kondrashov (2002). "Selection for short introns in highly expressed genes." Nat Genet **31**(4): 415-418.

Chao, Y., Y. Jiang, M. Zhong, K. Wei, C. Hu, Y. Qin, Y. Zuo, L. Yang, Z. Shen and C. Zou (2021). "Regulatory roles and mechanisms of alternative RNA splicing in adipogenesis and human metabolic health." Cell Biosci **11**(1): 66.

Chen, M. and J. L. Manley (2009). "Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches." Nat Rev Mol Cell Biol **10**(11): 741-754.

Chen, S., Y. Zhou, Y. Chen and J. Gu (2018). "fastp: an ultra-fast all-in-one FASTQ preprocessor." Bioinformatics **34**(17): i884-i890.

Chen, Y. I., R. E. Moore, H. Y. Ge, M. K. Young, T. D. Lee and S. W. Stevens (2007). "Proteomic analysis of in vivo-assembled pre-mRNA splicing complexes expands the catalog of participating factors." Nucleic Acids Res **35**(12): 3928-3944.

Cheng, Y. H., C. J. Liu, Y. H. Yu, Y. T. Jhou, M. Fujishima, I. J. Tsai and J. Y. Leu (2020). "Genome plasticity in Paramecium bursaria revealed by population genomics." BMC Biol **18**(1): 180.

doi:10.6342/NTU202502289

Cvitkovic, I. and M. S. Jurica (2013). "Spliceosome database: a tool for tracking components of the spliceosome." Nucleic Acids Res **41**(Database issue): D132-141.

Damgaard, C. K., S. Kahns, S. Lykke-Andersen, A. L. Nielsen, T. H. Jensen and J. Kjems (2008). "A 5' splice site enhances the recruitment of basal transcription initiation factors in vivo." Mol Cell **29**(2): 271-278.

Das, R., J. Yu, Z. Zhang, M. P. Gygi, A. R. Krainer, S. P. Gygi and R. Reed (2007). "SR proteins function in coupling RNAP II transcription to pre-mRNA splicing." Mol Cell **26**(6): 867-881.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson and T. R. Gingeras (2013). "STAR: ultrafast universal RNA-seq aligner." Bioinformatics **29**(1): 15-21.

Domazet-Loso, T., J. Brajkovic and D. Tautz (2007). "A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages." Trends Genet **23**(11): 533-539.

Ferrarini, M. G., A. Vallier, C. Vincent-Monegat, E. Dell'Aglio, B. Gillet, S. Hughes, O. Hurtado, G. Condemine, A. Zaidman-Remy, R. Rebollo, N. Parisot and A. Heddi (2023). "Coordination of host and endosymbiont gene expression governs endosymbiont growth and elimination in the cereal weevil Sitophilus spp." Microbiome **11**(1): 274.

Fujishima, M. and Y. Kodama (2012). "Endosymbionts in paramecium." Eur J Protistol **48**(2): 124-137.

Gnan, S., M. Matelot, M. Weiman, O. Arnaiz, F. Guerin, L. Sperling, M. Betermier, C. Thermes, C. L. Chen and S. Duharcourt (2022). "GC content, but not nucleosome positioning,

doi:10.6342/NTU202502289

directly contributes to intron splicing efficiency in Paramecium." Genome Res **32**(4): 699-709.

Griffiths-Jones, S., A. Bateman, M. Marshall, A. Khanna and S. R. Eddy (2003). "Rfam: an RNA family database." Nucleic Acids Res **31**(1): 439-441.

Hamada, M., K. Schroder, J. Bathia, U. Kurn, S. Fraune, M. Khalturina, K. Khalturin, C. Shinzato, N. Satoh and T. C. Bosch (2018). "Metabolic co-dependence drives the evolutionarily ancient Hydra-Chlorella symbiosis." Elife **7**.

Han, B. Y., Z. Liu, X. Hu and H. Ling (2022). "HNRNPU promotes the progression of triple-negative breast cancer via RNA transcription and alternative splicing mechanisms." Cell Death Dis **13**(11): 940.

Hang, J., R. Wan, C. Yan and Y. Shi (2015). "Structural basis of pre-mRNA splicing." Science **349**(6253): 1191-1198.

Haynes, W. J., K. Y. Ling, Y. Saimi and C. Kung (2003). "PAK paradox: Paramecium appears to have more K(+)-channel genes than humans." Eukaryot Cell **2**(4): 737-745.

He, M., J. Wang, X. Fan, X. Liu, W. Shi, N. Huang, F. Zhao and M. Miao (2019). "Genetic basis for the establishment of endosymbiosis in Paramecium." ISME J **13**(5): 1360-1369.

Hiller, M., K. Huse, K. Szafranski, N. Jahn, J. Hampe, S. Schreiber, R. Backofen and M. Platzer (2004). "Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity." Nat Genet **36**(12): 1255-1257.

Howe, K. L., P. Achuthan, J. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, A. G. Azov, R. Bennett, J. Bhai, K. Billis, S. Boddu, M. Charkhchi, C. Cummins, L. Da Rin Fioretto, C. Davidson, K. Dodiya, B. El Houdaigui, R. Fatima, A. Gall, C. Garcia Giron, T. Grego, C. Guijarro-Clarke, L. Haggerty, A. Hemrom, T. Hourlier, O. G. Izuogu, T.

doi:10.6342/NTU202502289

Juettemann, V. Kaikala, M. Kay, I. Lavidas, T. Le, D. Lemos, J. Gonzalez Martinez, J. C. Marugan, T. Maurel, A. C. McMahon, S. Mohanan, B. Moore, M. Muffato, D. N. Oheh, D. Paraschas, A. Parker, A. Parton, I. Prosovetskaia, M. P. Sakthivel, A. I. A. Salam, B. M. Schmitt, H. Schuilenburg, D. Sheppard, E. Steed, M. Szpak, M. Szuba, K. Taylor, A. Thormann, G. Threadgold, B. Walts, A. Winterbottom, M. Chakiachvili, A. Chaubal, N. De Silva, B. Flint, A. Frankish, S. E. Hunt, I. I. GR, N. Langridge, J. E. Loveland, F. J. Martin, J. M. Mudge, J. Morales, E. Perry, M. Ruffier, J. Tate, D. Thybert, S. J. Trevanion, F. Cunningham, A. D. Yates, D. R. Zerbino and P. Flicek (2021). "Ensembl 2021." Nucleic Acids Res **49**(D1): D884-D891.

Jenkins, B. H. (2024). "Mutualism on the edge: Understanding the Paramecium-Chlorella symbiosis." PLoS Biol **22**(4): e3002563.

Johnson, P. Z. and A. E. Simon (2023). "RNAcanvas: interactive drawing and exploration of nucleic acid structures." Nucleic Acids Res **51**(W1): W501-W508.

Jones, P., D. Binns, H. Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S. Y. Yong, R. Lopez and S. Hunter (2014). "InterProScan 5: genome-scale protein function classification." Bioinformatics **30**(9): 1236-1240.

Kalsotra, A. and T. A. Cooper (2011). "Functional consequences of developmentally regulated alternative splicing." Nat Rev Genet **12**(10): 715-729.

Kannan, L. and W. C. Wheeler (2012). "Maximum Parsimony on Phylogenetic networks." Algorithms Mol Biol **7**(1): 9.

Keeling, P. J. (2013). "The number, speed, and impact of plastid endosymbioses in eukaryotic evolution." Annu Rev Plant Biol **64**: 583-607.

doi:10.6342/NTU202502289

Kelly, J. B., D. E. Carlson, M. Reuter, A. Sommershof, L. Adamec and L. Becks (2025). "Genomic Signatures of Adaptation to Stress Reveal Shared Evolutionary Trends Between Tetrahymena utriculariae and Its Algal Endosymbiont, Micractinium tetrahymenae." Mol Biol Evol **42**(2).

Kelly, S. (2021). "The economics of organellar gene loss and endosymbiotic gene transfer." Genome Biol **22**(1): 345.

Kenny, C. J., M. P. McGurk, S. Schuler, A. Cordero, S. Laubinger and C. B. Burge (2025). "LUC7 proteins define two major classes of 5' splice sites in animals and plants." Nat Commun **16**(1): 1574.

Keppetipola, N., S. Sharma, Q. Li and D. L. Black (2012). "Neuronal regulation of pre-mRNA splicing by polypyrimidine tract binding proteins, PTBP1 and PTBP2." Crit Rev Biochem Mol Biol **47**(4): 360-378.

Kervestin, S. and A. Jacobson (2012). "NMD: a multifaceted response to premature translational termination." Nat Rev Mol Cell Biol **13**(11): 700-712.

Kjer-Hansen, P. and R. J. Weatheritt (2023). "The function of alternative splicing in the proteome: rewiring protein interactomes to put old functions into new contexts." Nat Struct Mol Biol **30**(12): 1844-1856.

Kodama, Y. and M. Fujishima (2010). "Secondary symbiosis between Paramecium and Chlorella cells." International review of cell and molecular biology **279**: 33-77.

Kodama, Y. and M. Fujishima (2022). "Endosymbiotic Chlorella variabilis reduces mitochondrial number in the ciliate Paramecium bursaria." Sci Rep **12**(1): 8216.

doi:10.6342/NTU202502289

Kodama, Y., I. Inouye and M. Fujishima (2011). "Symbiotic Chlorella vulgaris of the ciliate Paramecium bursaria plays an important role in maintaining perialgal vacuole membrane functions." Protist **162**(2): 288-303.

Koretke, K. K., A. N. Lupas, P. V. Warren, M. Rosenberg and J. R. Brown (2000). "Evolution of two-component signal transduction." Mol Biol Evol **17**(12): 1956-1970.

Lee, S. J., M. D. Zdradzinski, P. K. Sahoo, A. N. Kar, P. Patel, R. Kawaguchi, B. J. Aguilar, K. D. Lantz, C. R. McCain, G. Coppola, Q. Lu and J. L. Twiss (2021). "Selective axonal translation of the mRNA isoform encoding prenylated Cdc42 supports axon growth." J Cell Sci **134**(7).

Li, S., Y. Wang, Y. Zhao, X. Zhao, X. Chen and Z. Gong (2020). "Global Co-transcriptional Splicing in Arabidopsis and the Correlation with Splicing Regulation in Mature RNAs." Mol Plant **13**(2): 266-277.

Lim, C. S., B. N. Weinstein, S. W. Roy and C. M. Brown (2021). "Analysis of Fungal Genomes Reveals Commonalities of Intron Gain or Loss and Functions in Intron-Poor Species." Mol Biol Evol **38**(10): 4166-4186.

Lin, K. and D. Y. Zhang (2005). "The excess of 5' introns in eukaryotic genomes." Nucleic Acids Res **33**(20): 6522-6527.

Love, M. I., W. Huber and S. Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biol **15**(12): 550.

Martin, W. F., S. Garg and V. Zimorski (2015). "Endosymbiotic theories for eukaryote origin." Philos Trans R Soc Lond B Biol Sci **370**(1678): 20140330.

doi:10.6342/NTU202502289

McGlincy, N. J., A. Valomon, J. E. Chesham, E. S. Maywood, M. H. Hastings and J. Ule (2012). "Regulation of alternative splicing by the circadian clock and food related cues." Genome Biol **13**(6): R54.

Miller, J. N. and D. A. Pearce (2014). "Nonsense-mediated decay in genetic disease: friend or foe?" Mutat Res Rev Mutat Res **762**: 52-64.

Nawrocki, E. P. and S. R. Eddy (2013). "Infernal 1.1: 100-fold faster RNA homology searches." Bioinformatics **29**(22): 2933-2935.

Nilsen, T. W. and B. R. Graveley (2010). "Expansion of the eukaryotic proteome by alternative splicing." Nature **463**(7280): 457-463.

Nott, A., S. H. Meislin and M. J. Moore (2003). "A quantitative analysis of intron effects on mammalian gene expression." Rna **9**(5): 607-617.

Nuadthaisong, J., T. Phetruen, C. Techawisutthinan and S. Chanarat (2022). "Insights into the Mechanism of Pre-mRNA Splicing of Tiny Introns from the Genome of a Giant Ciliate Stentor coeruleus." Int J Mol Sci **23**(18).

Olivieri, J. E., R. Dehghannasiri, P. L. Wang, S. Jang, A. de Morree, S. Y. Tan, J. Ming, A. Ruohao Wu, C. Tabula Sapiens, S. R. Quake, M. A. Krasnow and J. Salzman (2021). "RNA splicing programs define tissue compartments and cell types at single-cell resolution." Elife **10**.

Olthof, A. M., C. F. Schwoerer, K. N. Girardini, A. L. Weber, K. Doggett, S. Mieruszynski, J. K. Heath, T. E. Moore, J. Biran and R. N. Kanadia (2024). "Taxonomy of introns and the evolution of minor introns." Nucleic Acids Res **52**(15): 9247-9266.

Palacios, I. M. (2013). "Nonsense-mediated mRNA decay: from mechanistic insights to impacts on human health." Brief Funct Genomics **12**(1): 25-36.

doi:10.6342/NTU202502289

Pan, B., F. Ye, T. Li, F. Wei, A. Warren, Y. Wang and S. Gao (2023). "Potential role of N(6)-adenine DNA methylation in alternative splicing and endosymbiosis in Paramecium bursaria." iScience **26**(5): 106676.

Pan, Q., O. Shai, L. J. Lee, B. J. Frey and B. J. Blencowe (2008). "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing." Nat Genet **40**(12): 1413-1415.

Park, S. G., S. Hannenhalli and S. S. Choi (2014). "Conservation in first introns is positively associated with the number of exons within genes and the presence of regulatory epigenetic signals." BMC Genomics **15**(1): 526.

Patro, R., G. Duggal, M. I. Love, R. A. Irizarry and C. Kingsford (2017). "Salmon provides fast and bias-aware quantification of transcript expression." Nat Methods **14**(4): 417-419.

Plaschka, C., P. C. Lin, C. Charenton and K. Nagai (2018). "Prespliceosome structure provides insights into spliceosome assembly and regulation." Nature **559**(7714): 419-422.

Quinlan, A. R. and I. M. Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features." Bioinformatics **26**(6): 841-842.

Roy, K. R., J. Gabunilas, D. Neutel, M. Ai, Z. Yeh, J. Samson, G. Lyu and G. F. Chanfreau (2023). "Splicing factor Prp18p promotes genome-wide fidelity of consensus 3'-splice sites." Nucleic Acids Res **51**(22): 12428-12442.

Ryll, J., R. Rothering and F. Catania (2022). "Intronization Signatures in Coding Exons Reveal the Evolutionary Fluidity of Eukaryotic Gene Architecture." Microorganisms **10**(10).

Sakurai, A., S. Fujimori, H. Kochiwa, S. Kitamura-Abe, T. Washio, R. Saito, P. Carninci, Y. Hayashizaki and M. Tomita (2002). "On biased distribution of introns in various eukaryotes." Gene **300**(1-2): 89-95.

doi:10.6342/NTU202502289

Saudemont, B., A. Popa, J. L. Parmley, V. Rocher, C. Blugeon, A. Necsulea, E. Meyer and L. Duret (2017). "The fitness cost of mis-splicing is the main determinant of alternative splicing patterns." Genome Biol **18**(1): 208.

Schirman, D., Z. Yakhini, Y. Pilpel and O. Dahan (2021). "A broad analysis of splicing regulation in yeast using a large library of synthetic introns." PLoS Genet **17**(9): e1009805.

Seol, D. W. and T. R. Billiar (1999). "A caspase-9 variant missing the catalytic site is an endogenous inhibitor of apoptosis." J Biol Chem **274**(4): 2072-2076.

Shaul, O. (2017). "How introns enhance gene expression." Int J Biochem Cell Biol **91**(Pt B): 145-155.

Shen, S., J. W. Park, Z. X. Lu, L. Lin, M. D. Henry, Y. N. Wu, Q. Zhou and Y. Xing (2014). "rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data." Proc Natl Acad Sci U S A **111**(51): E5593-5601.

Soneson, C., M. I. Love and M. D. Robinson (2015). "Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences." F1000Res **4**: 1521.

Steinegger, M. and J. Soding (2017). "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets." Nat Biotechnol **35**(11): 1026-1028.

Stover, N. A., R. S. Punia, M. S. Bowen, S. B. Dolins and T. G. Clark (2012). "Tetrahymena Genome Database Wiki: a community-maintained model organism database." Database (Oxford) **2012**: bas007.

Suzuki, S., K. Ishida and Y. Hirakawa (2016). "Diurnal Transcriptional Regulation of Endosymbiotically Derived Genes in the Chlorarachniophyte Bigelowiella natans." Genome Biol Evol **8**(9): 2672-2682.

doi:10.6342/NTU202502289

Sweeney, B. A., D. Hoksza, E. P. Nawrocki, C. E. Ribas, F. Madeira, J. J. Cannone, R. Gutell, A. Maddala, C. D. Meade and L. D. Williams (2021). "R2DT is a framework for predicting and visualising RNA secondary structure using templates." Nature Communications **12**(1): 3494.

Takashima, S., W. Sun, A. B. C. Otten, P. Cai, S. I. Peng, E. Tong, J. Bui, M. Mai, O. Amarbayar, B. Cheng, R. J. Odango, Z. Li, K. Qu and B. K. Sun (2024). "Alternative mRNA splicing events and regulators in epidermal differentiation." Cell Rep **43**(3): 113814.

Takayama, K. I., T. Suzuki, K. Sato, Y. Saito and S. Inoue (2024). "Cooperative nuclear action of RNA-binding proteins PSF and G3BP2 to sustain neuronal cell viability is decreased in aging and dementia." Aging Cell **23**(12): e14316.

Takei, S., M. Togo-Ohno, Y. Suzuki and H. Kuroyanagi (2016). "Evolutionarily conserved autoregulation of alternative pre-mRNA splicing by ribosomal protein L10a." Nucleic Acids Res **44**(12): 5585-5596.

Ullrich, S. and R. Guigo (2020). "Dynamic changes in intron retention are tightly associated with regulation of splicing factors and proliferative activity during B-cell development." Nucleic Acids Res **48**(3): 1327-1340.

Verta, J. P. and A. Jacobs (2022). "The role of alternative splicing in adaptation and evolution." Trends Ecol Evol **37**(4): 299-308.

von der Dunk, S. H. A., P. Hogeweg and B. Snel (2023). "Obligate endosymbiosis enables genome expansion during eukaryogenesis." Commun Biol **6**(1): 777.

Wang, Y., J. Liu, B. O. Huang, Y. M. Xu, J. Li, L. F. Huang, J. Lin, J. Zhang, Q. H. Min, W. M. Yang and X. Z. Wang (2015). "Mechanism of alternative splicing and its regulation." Biomed Rep **3**(2): 152-158.

doi:10.6342/NTU202502289

Wang, Z. and C. B. Burge (2008). "Splicing regulation: from a parts list of regulatory elements to an integrated splicing code." RNA **14**(5): 802-813.

Wang, Z., X. Xiao, E. Van Nostrand and C. B. Burge (2006). "General and specific functions of exonic splicing silencers in splicing control." Mol Cell **23**(1): 61-70.

Ward, N. and G. Moreno-Hagelsieb (2014). "Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: how much do we miss?" PloS one **9**(7): e101850.

Wilson, A. C. and R. P. Duncan (2015). "Signatures of host/symbiont genome coevolution in insect nutritional endosymbioses." Proc Natl Acad Sci U S A **112**(33): 10255-10261.

Yang, H., B. Beutler and D. Zhang (2022). "Emerging roles of spliceosome in cancer and immunity." Protein Cell **13**(8): 559-579.

Yang, J., L. H. Hung, T. Licht, S. Kostin, M. Looso, E. Khrameeva, A. Bindereif, A. Schneider and T. Braun (2014). "RBM24 is a major regulator of muscle-specific alternative splicing." Dev Cell **31**(1): 87-99.

Yeo, G. W., E. L. Van Nostrand and T. Y. Liang (2007). "Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements." PLoS Genet **3**(5): e85.

Yu, G., L. G. Wang, Y. Han and Q. Y. He (2012). "clusterProfiler: an R package for comparing biological themes among gene clusters." OMICS **16**(5): 284-287.

Yu, J., Z. Yang, M. Kibukawa, M. Paddock, D. A. Passey and G. K. Wong (2002). "Minimal introns are not "junk"." Genome Res **12**(8): 1185-1189.

Yuuki Kodama, H. S., Hideo Dohra, Manabu Sugii, Tatsuya Kitazume, Katsushi Yamaguchi, Shuji Shigenobu, Masahiro Fujishima (2014). "Comparison of gene expression of Paramecium bursaria with and without Chlorella variabilissymbionts." BMC Genomics(1): 183.

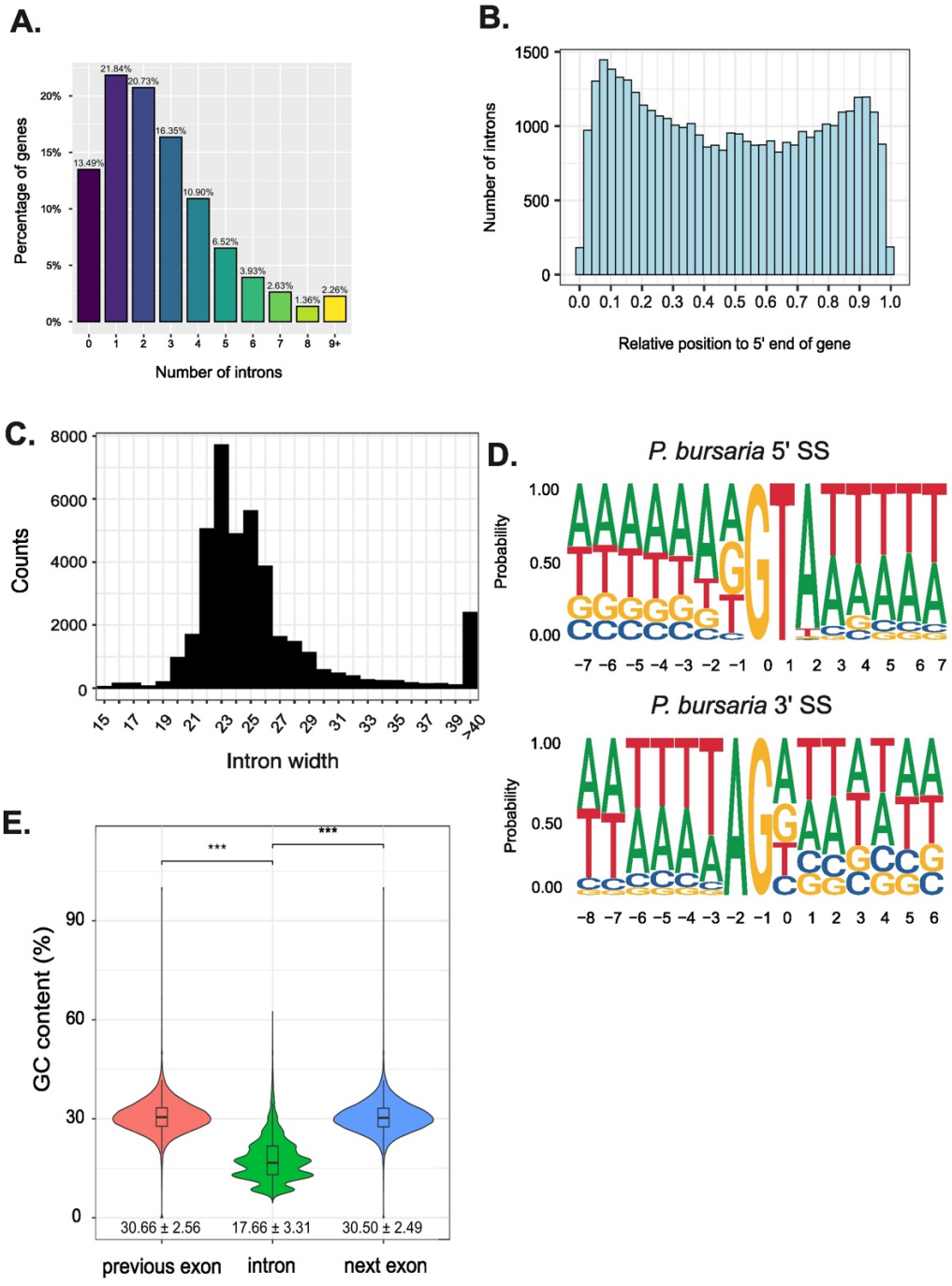
doi:10.6342/NTU202502289

Zhang, X., Z. Guo, Y. Li and Y. Xu (2025). "Splicing to orchestrate cell fate." Mol Ther Nucleic Acids **36**(1): 102416.

Zhu, J., F. He, D. Wang, K. Liu, D. Huang, J. Xiao, J. Wu, S. Hu and J. Yu (2010). "A novel role for minimal introns: routing mRNAs to the cytosol." PLoS One **5**(4): e10144.

doi:10.6342/NTU202502289

Figures



doi:10.6342/NTU202502289

Figure 1. Intron characteristics of *P. bursaria*

A. Bar plot shows the distribution of intron number in *P. bursaria*'s genes

B. Distribution of introns relative to the 5' end of transcripts. Intron positions are determined by dividing the distance from the translation start site by the total mRNA length.

C. Distribution of intron lengths (nucleotides) in *P. bursaria* (n = 39,715 introns).

D. Sequence logo of 5'SS and 3'SS in P. bursaria, generated using ggseqlogo R. The height of each letter represents its relative frequency at that position, with letters arranged in descending order of probability.

E. Boxplot of GC content distribution of intron and flanking exons in *P. bursaria.*

doi:10.6342/NTU202502289

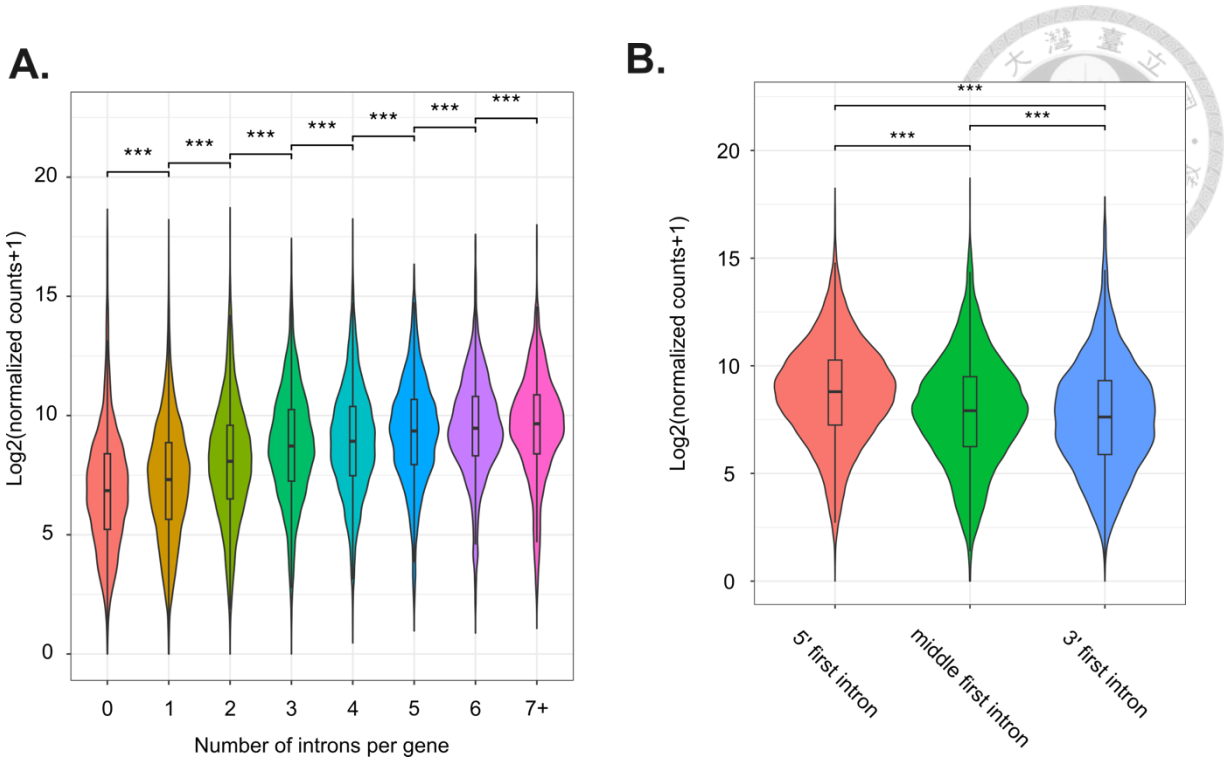


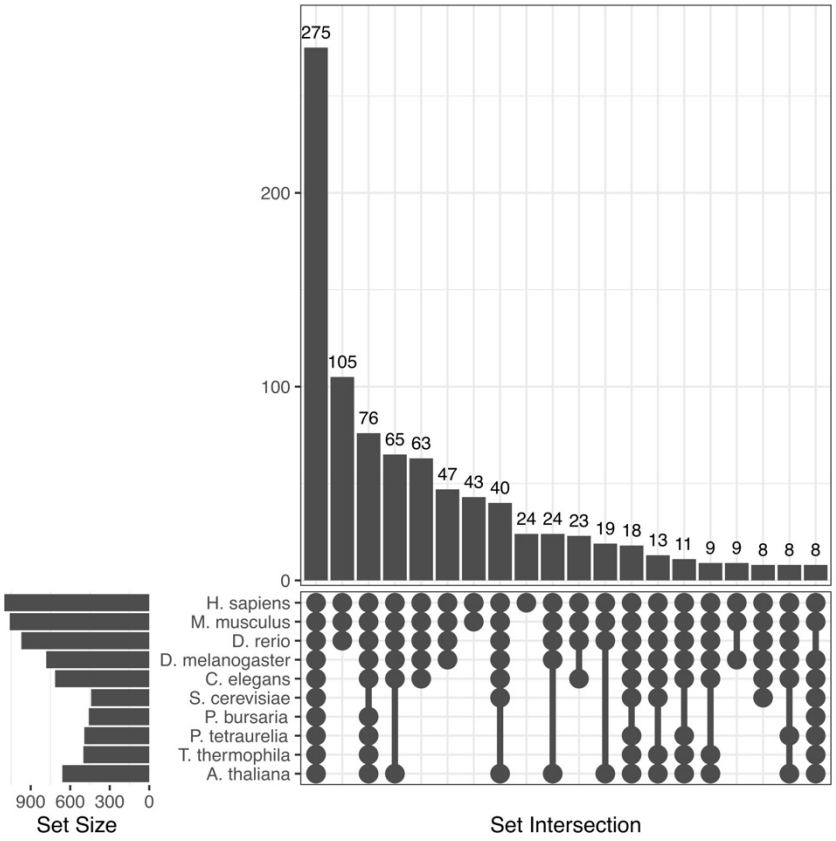
Figure 2. Intron position and Intron Mediated Enhancement (IME) in *P. bursaria*

A. Gene expression levels categorized by the number of introns per gene. *** indicates $p < 0.001$ (Mann-Whitney U test).

B. Gene expression levels grouped by the position of the first intron within the gene. A Mann-Whitney U test was performed (NS: $p > 0.05$, ***: $p < 0.001$).

doi:10.6342/NTU202502289

A.



B.

Class/ Family	*H. sapiens*	*M. musculus*	*D. rerio*	*D. melanogaster*	*C. elegans*	*S. cerevisiae*	*P. bursaria*	*P. tetraurelia*	*T. thermophila*	*A. thaliana*
U1 snRNP	4	4	4	3	3	4	3	2	2	3
Sm	7	7	7	7	7	6	4	5	5	7
U11/U12 snRNP	5	5	5	3	0	0	0	0	0	5
17S U2 snRNP	12	12	12	12	11	10	12	12	12	12
17S U2 snRNP associated	11	11	11	11	11	3	8	9	9	10
U5 snRNP	8	8	8	8	8	6	7	7	7	8
U4/U6 snRNP	5	5	5	5	5	4	5	5	5	5
U4/U6 recycling	1	1	1	1	1	0	0	1	1	1
LSm	8	8	8	8	8	7	6	7	7	8
tri–snRNP	2	2	2	2	2	2	2	2	2	2
Recruited at A complex	8	8	8	6	5	2	4	4	4	7
PRP19 complex	9	9	9	9	8	4	5	7	7	9
PRP19 related	9	8	8	8	8	6	6	6	5	7
Recruited at B complex	8	8	8	8	8	2	6	6	6	8
Recruited at Bact complex	12	12	11	12	12	5	10	10	10	11
Recruited at C complex	18	18	18	18	17	0	11	12	11	16
Second step factors	5	5	5	5	5	5	5	5	5	5
RES complex	3	3	3	3	3	3	2	3	3	3
EJC/mRNP	12	12	11	12	11	3	3	7	5	8
hnRNP	29	28	17	11	6	2	2	3	5	7
SR protein	16	15	15	10	9	1	2	5	2	7
SR related	2	2	2	2	2	0	2	2	2	2
TREX	6	6	6	6	5	2	3	3	3	6
CSTF	2	2	2	2	2	1	0	0	0	2
CPSF	7	7	6	6	6	3	2	2	1	6
GEM	3	3	3	3	2	0	0	0	0	1

max
0

doi:10.6342/NTU202502289

Figure 3. Conserved splicing-related proteins in model species and ciliate species

A. Upset plot displaying the number of shared splicing-related proteins among 10 species. Set size represents the number of spliceosomal proteins in each species, while numbers above the bars indicate the count of proteins in each shared group.

B. Number of splicing-related proteins in each Class/Family across 10 species. Colors represent the relative protein counts within each class, scaled from 0 to the maximum number of proteins in that class.

doi:10.6342/NTU202502289

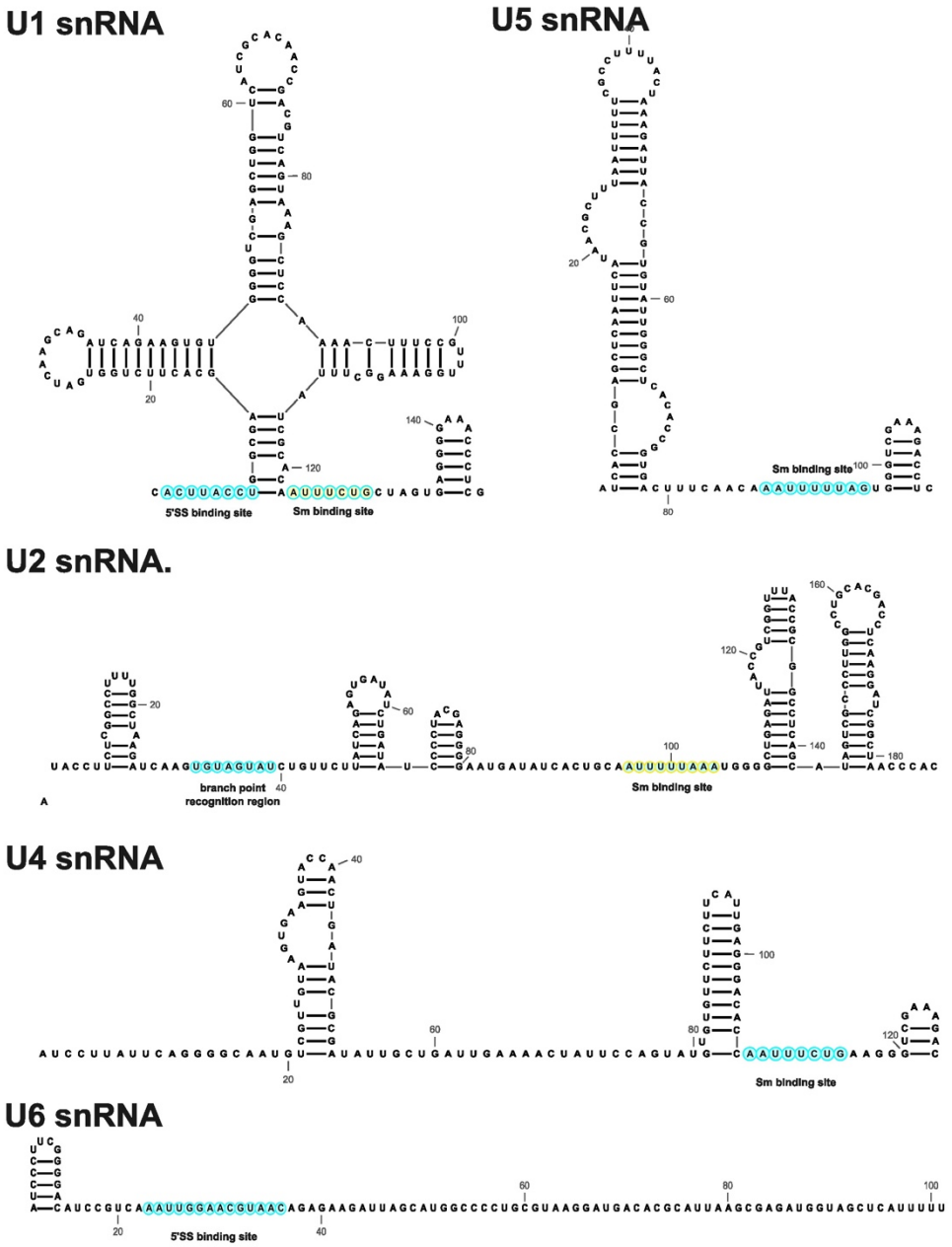


Figure 4. Sequence and secondary structure of UsnRNAs in *P. bursaria*

doi:10.6342/NTU202502289

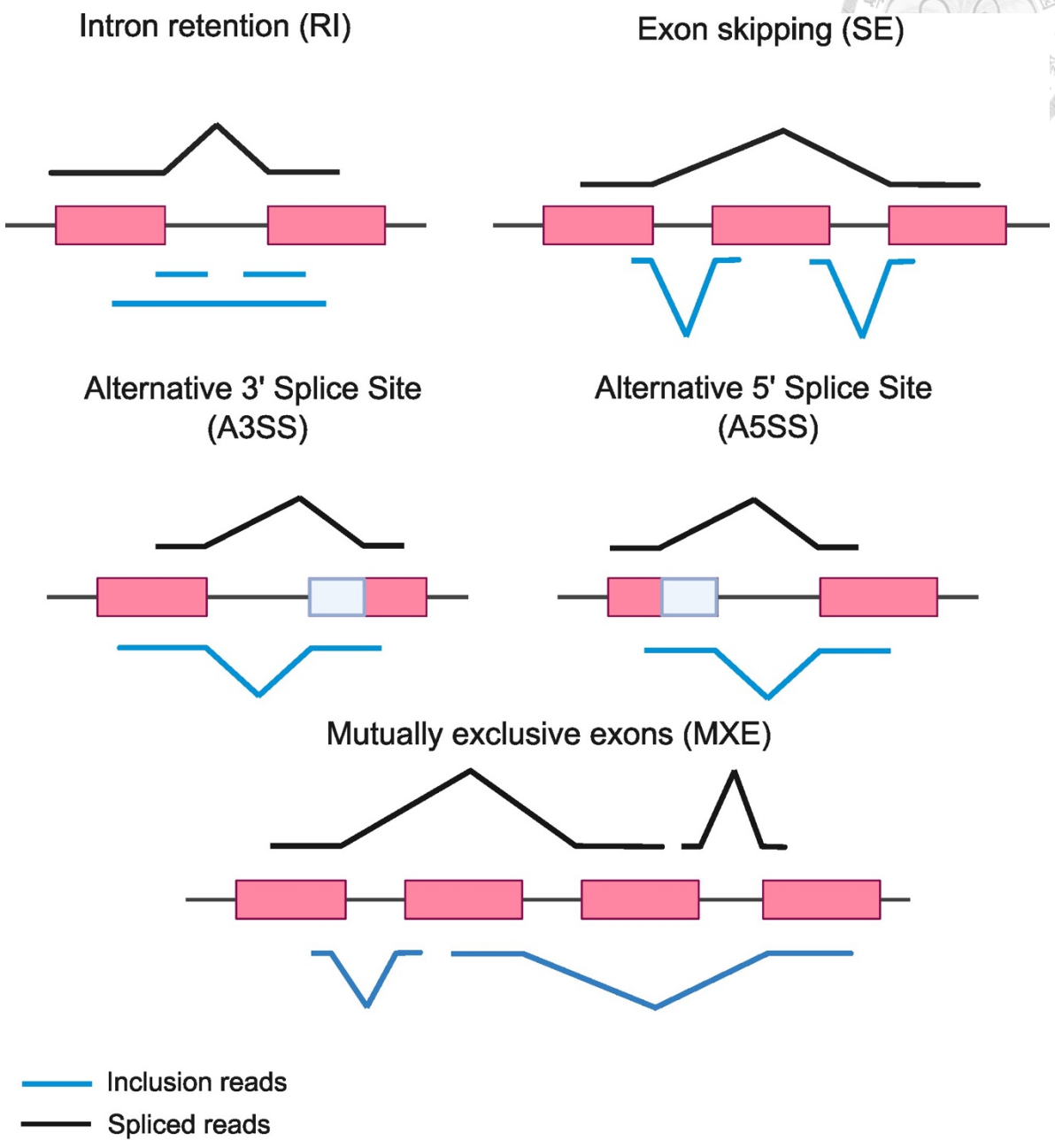
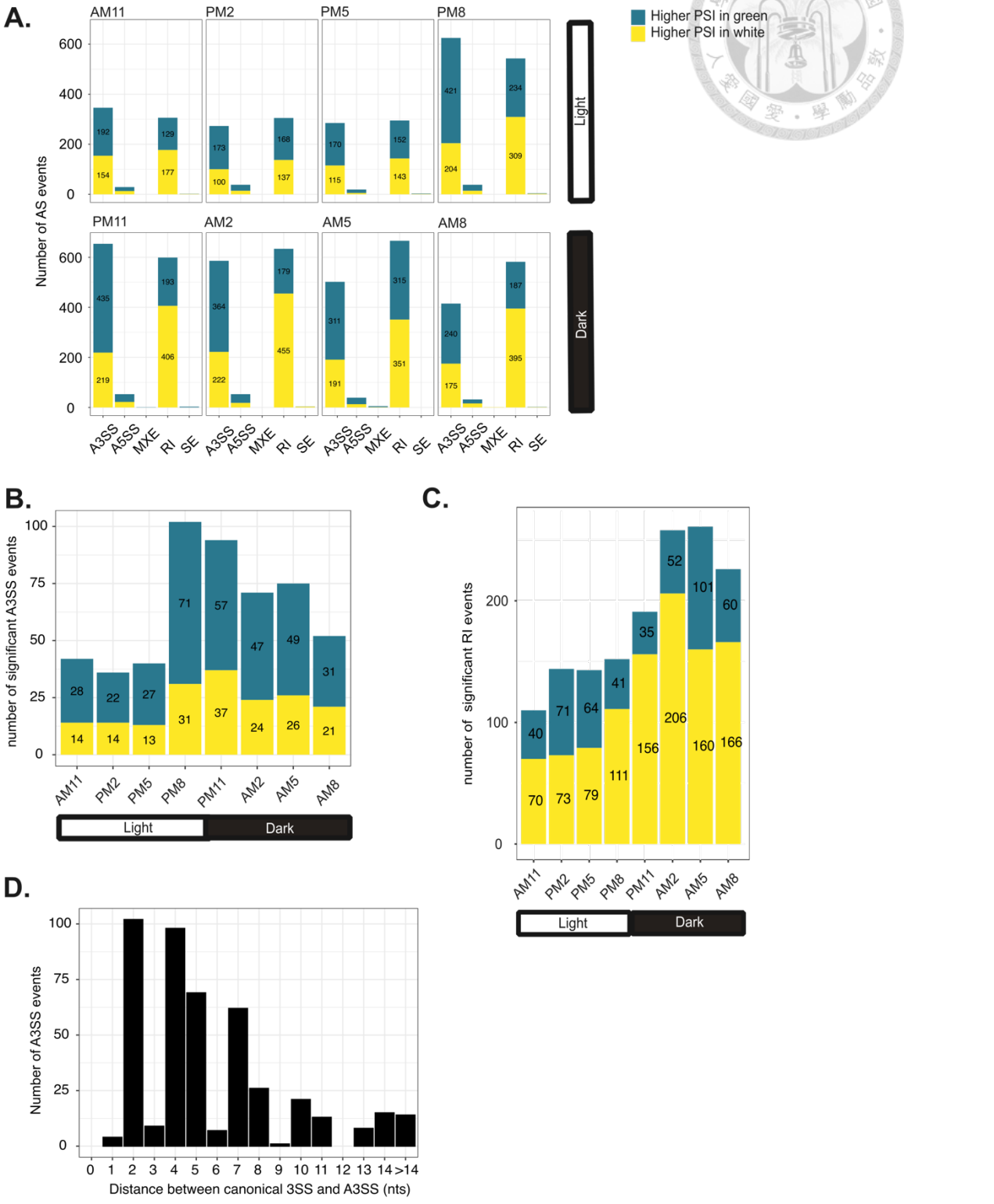


Figure 5. Illustration of five Alternative Splicing events

doi:10.6342/NTU202502289



doi:10.6342/NTU202502289

Figure 6. Alternative splicing events between symbiosis and asymbiosis *P. bursaria*

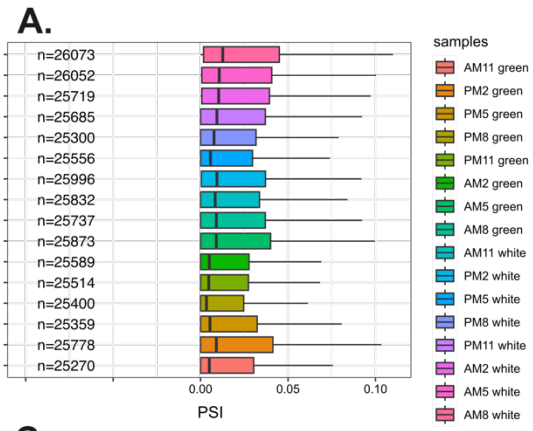
A. Number of significant Alternative Splicing events (FDR < 0.05) between green and white cells in 8 timepoints. Green: higher PSI in green cells, yellow: higher PSI in white cells

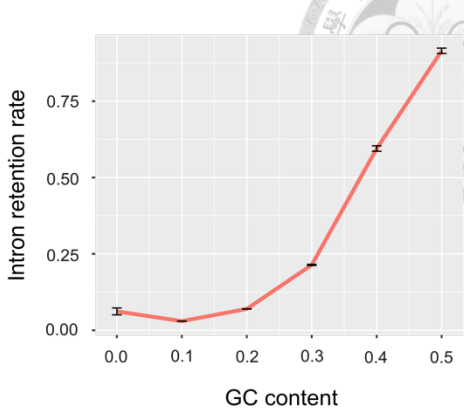
B. Number of significant A3SS events between green and white cells in 8 timepoints after filtering using Junction Counts. Green: higher PSI in green cells, yellow: higher PSI in white cells

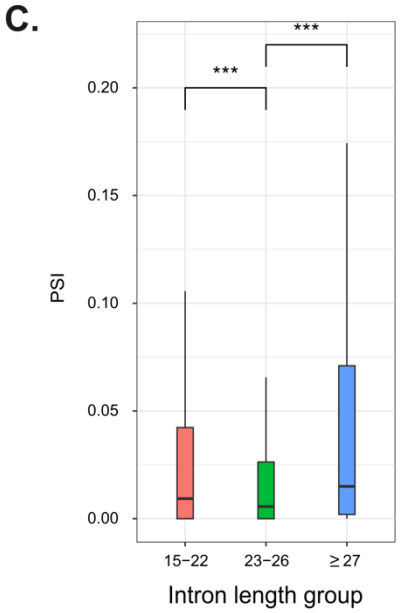
C. Number of DSIs between green and white cells in 8 timepoints after filtering using Junction Counts. Green: DSIs with higher PSI in green cells, yellow: DSIs with higher PSI in white cells

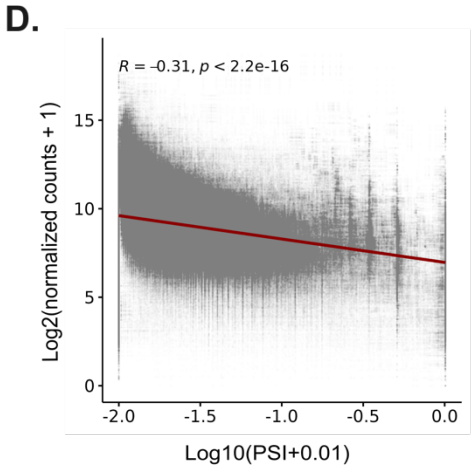
D. Distance (in nucleotides) from the Alternative 3' Splice Site to the Canonical 3' Splice Site in significant A3SS events between green and white cells.

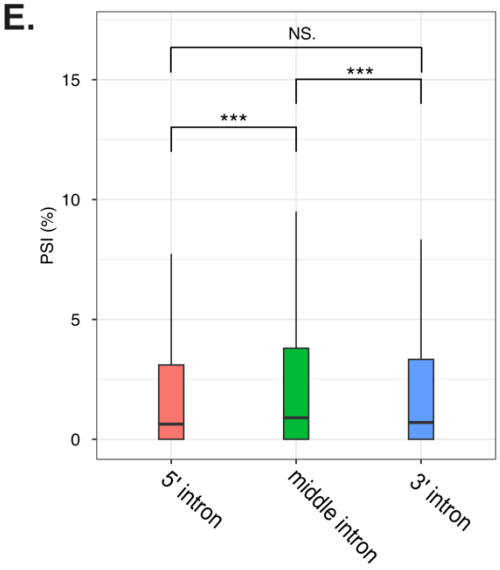
doi:10.6342/NTU202502289











doi:10.6342/NTU202502289

Figure 7. Intron retention rate correlates with intron sequence features and gene expression level

A. Violin plot shows intron retention rate (PSI) distribution across 16 samples, red dots indicate median of PSI, n = number of introns calculated in each sample.

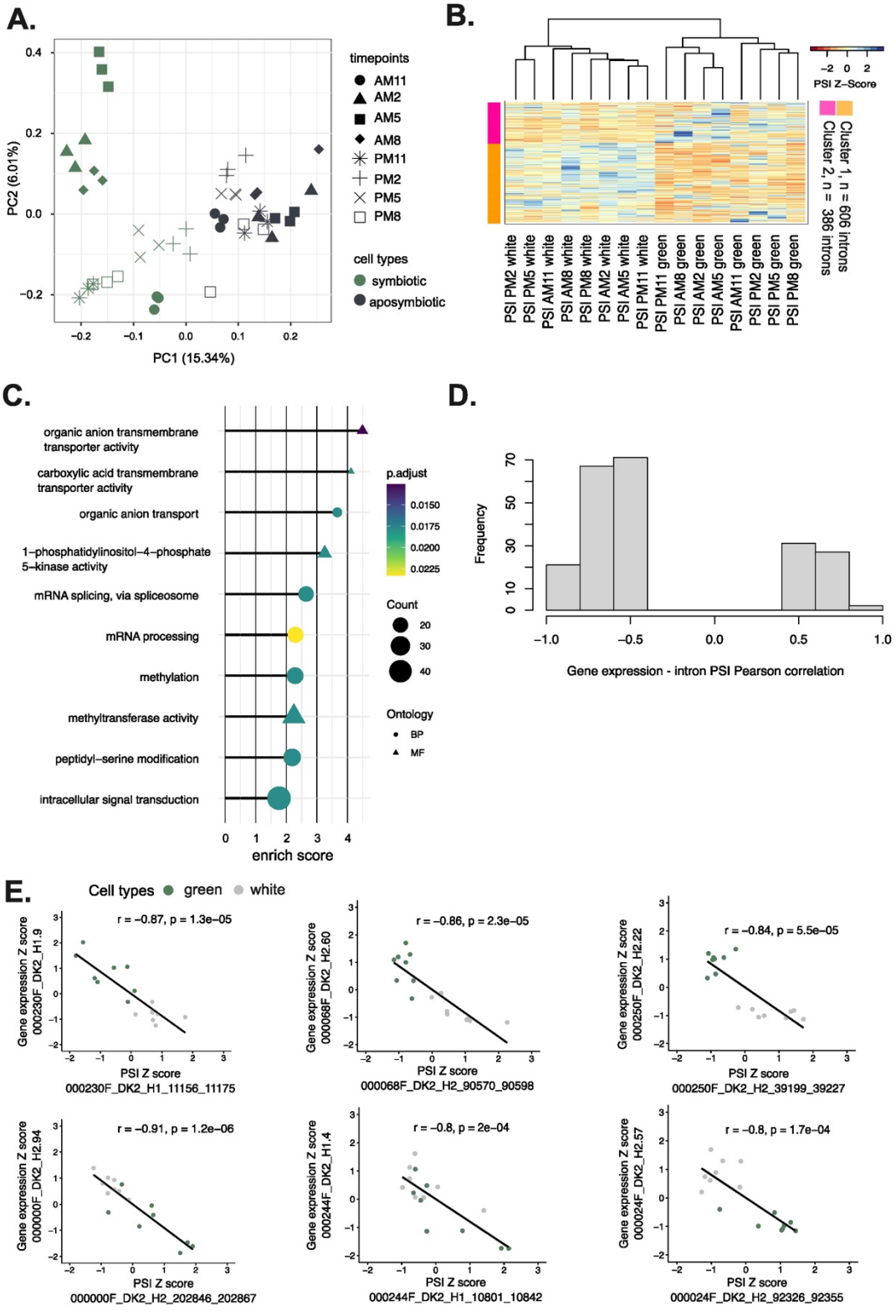
B. Intron retention rate across different intron GC content groups. GC content values are rounded to the nearest group. Error bars indicate SEM of intron retention rate in each GC content group.

C. Intron retention rate across different intron length group. The introns were divided into three groups: >=27 nts, 15-22 nts, 23-26 nts. Wilcoxon rank-sum test was performed, ***: p value < 0.001, NS: non-significant.

D. Scatter plot of gene expression level and corresponding intron's PSI shows negative correlation between PSI and gene expression level (Pearson R = -0.31, $p < 2.2e-16$).

E. Intron retention rates (%) based on intron position within genes. Introns are classified as 3' introns (relative position ≥ 0.75), 5' introns (relative position ≤ 0.25), and middle introns (0.25 < relative position < 0.75). Statistical significance was determined using the Mann-Whitney U test (NS: $p > 0.05$, ***: $p < 0.001$).

doi:10.6342/NTU202502289

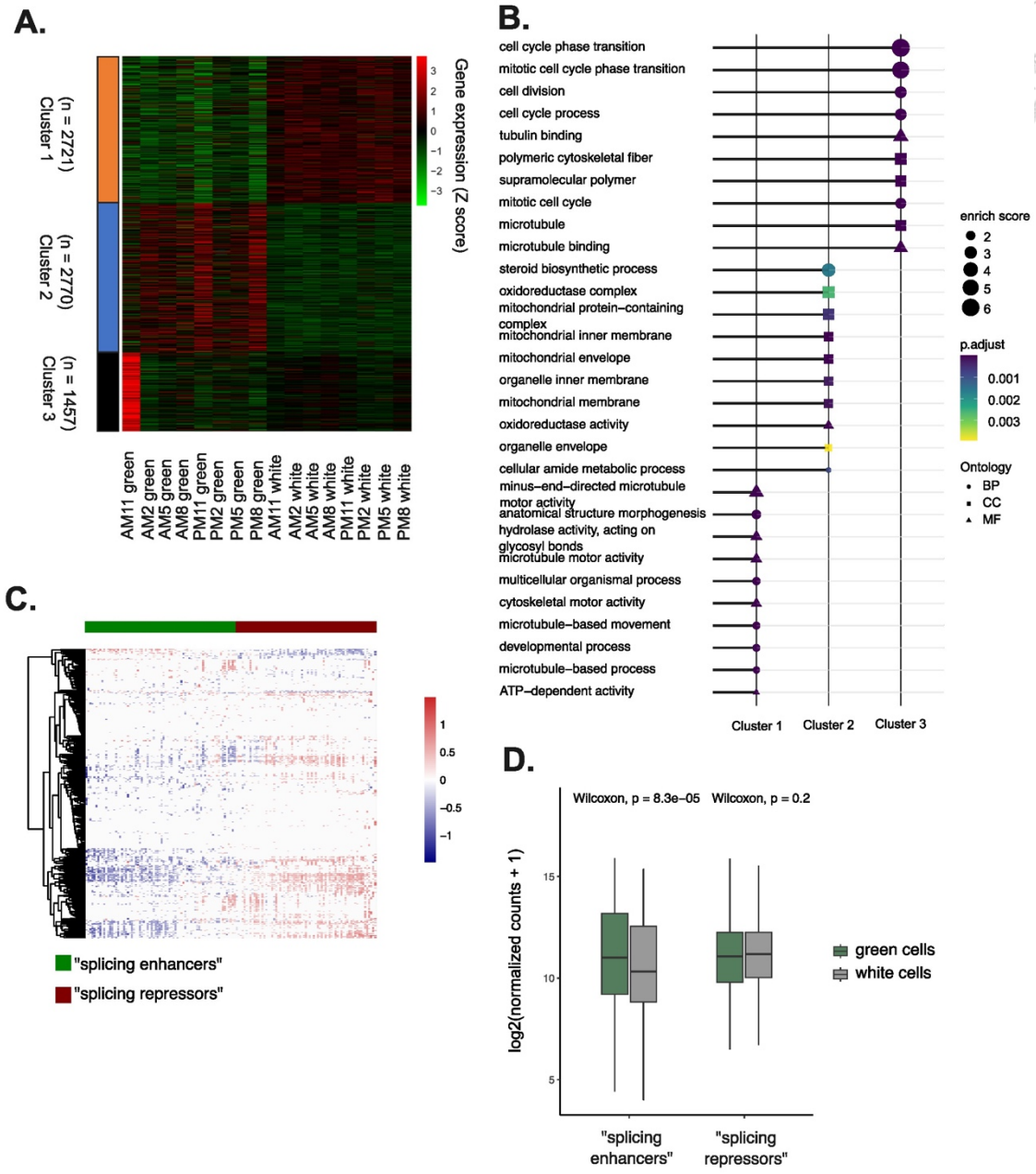


doi:10.6342/NTU202502289

Figure 8. Patterns of intron splicing between green and white cells

A. PCA cluster 16 samples (n=3 replicates for each sample) using PSI of 992 DSIs
B. k-means clustering results, the heatmap showing the PSI in 16 samples in two cluster of DSIs. Colour is scaled based on z-score of PSI. Cluster 1 (orange), n = 606 introns, Cluster 2 (pink), n = 386 introns,
C. Top 10 GO terms enrichment of genes containing 992 DSIs
D. Distribution of significant Pearson correlation coefficient between DSIs' PSI and their gene expression level.
E. Top 3 Pearson correlation between PSI Z and its Gene expression level Z score between DSIs in cluster 1 and cluster 2. Top: introns in cluster 2, bottom: introns in cluster 1. Green dots: white cells, blue dots: green cells.

doi:10.6342/NTU202502289



doi:10.6342/NTU202502289

Figure 9. Splicing-related DEGs associates with DSIs during endosymbiosis

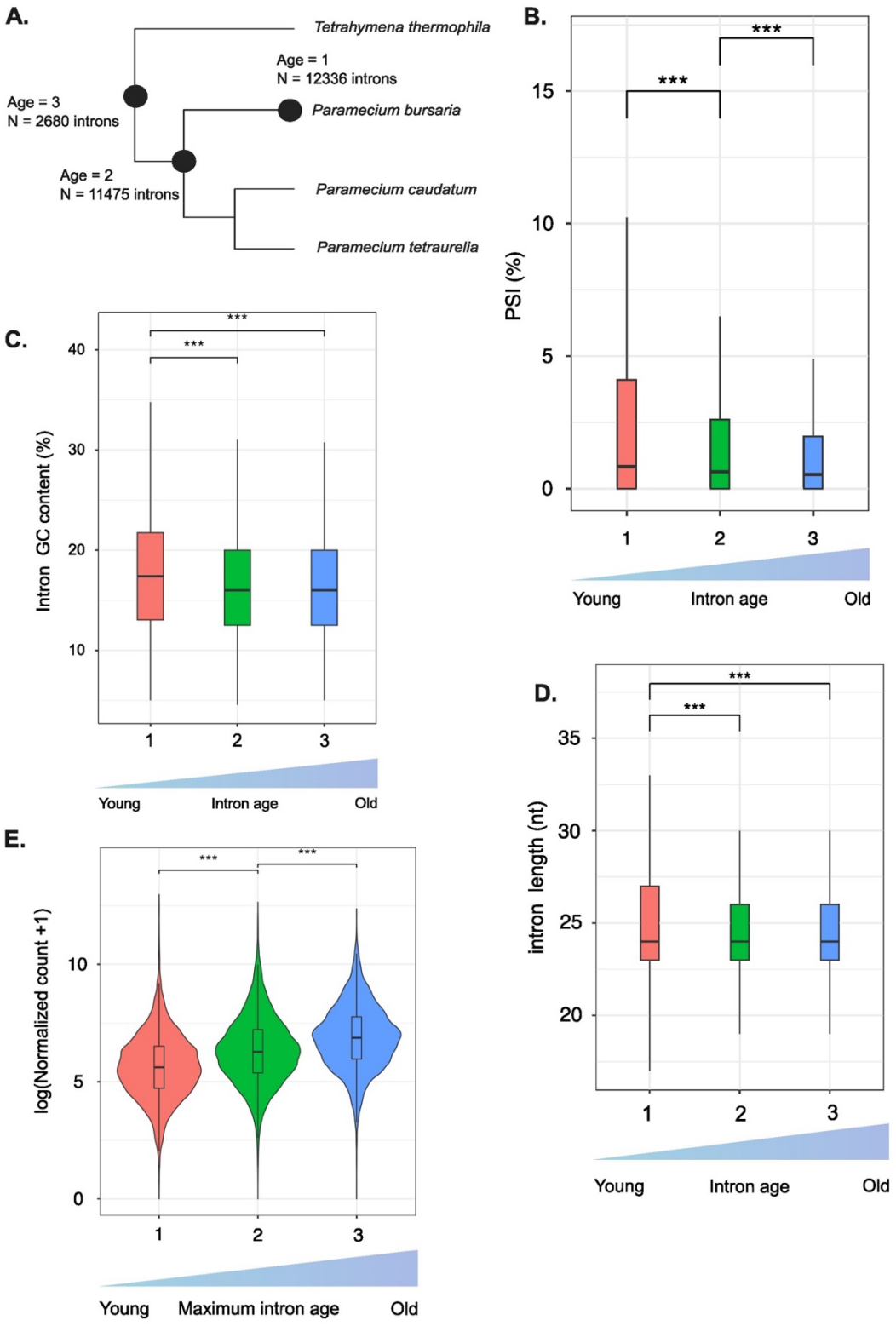
A. k-means clustering results of DEGs between green and white cells in 8 timepoints, heatmap showing the gene expression level (z score) in 16 samples. Cluster 1 = 2721 genes, Cluster 2 = 2770 genes, Cluster 3 = 1457 genes.

B. Top 10 GO terms enrichment of three different DEG clusters between *P. bursaria* green and white cells

C. Heatmap shows the coefficients (β_1) of splicing-related genes from the linear model of DSI PSI values, based on 992 DSIs (rows) and 147 differentially expressed splicing-related genes (columns).

D. Gene expression level between greeen and white cells of "splicing enhancers" group and "splicing repressors" group. Wilcoxon rank-sum test was performed.

doi:10.6342/NTU202502289



doi:10.6342/NTU202502289

Figure 10. Short intron evolution in *P. bursaria*

A. Introns in *P. bursaria* were assigned in age groups based on phylogenetic tree.

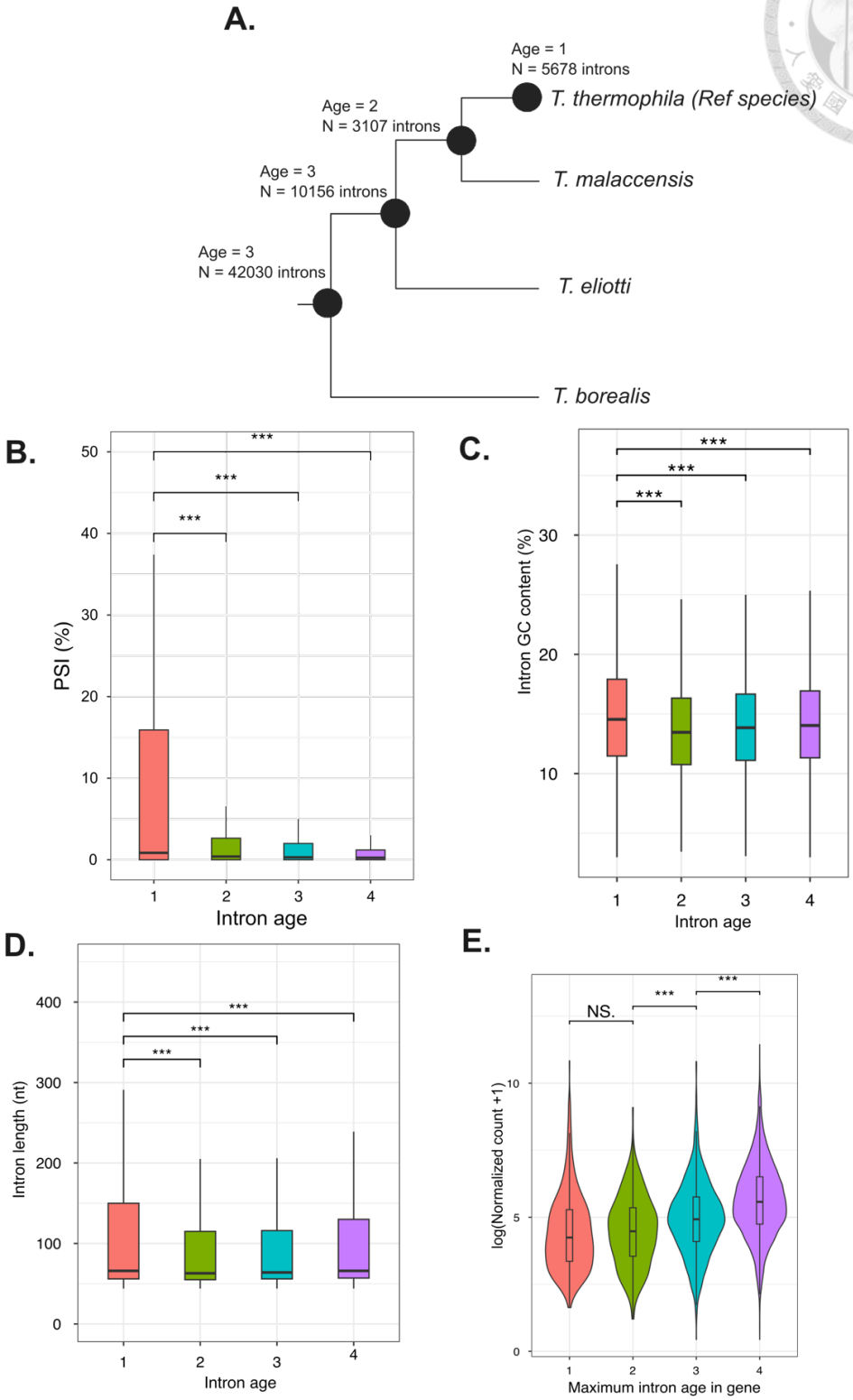
B. Boxplot showed intron retention rate (%) in each intron age group. Man-Whitney U test was performed, ***: p-value < 0.001, NS: p value > 0.05

C. Boxplot showed intron GC content (%) in each intron age group. Man-Whitney U test was performed, ***: p-value < 0.001, NS: p value > 0.05

D. Boxplot showed intron length (bp) in each intron age group. KS test was performed, ***: p-value < 0.001, NS: p value > 0.05

E. Boxplot showed expression in each gene group based on maximum intron age in gene. Man-Whitney U test was performed, ***: p-value < 0.001, NS: p value > 0.05

doi:10.6342/NTU202502289



doi:10.6342/NTU202502289

Figure 11. Intron evolution in *Tetrahymena*

A. Introns in *T. thermophila* were assigned in age groups based on phylogenetic tree.

B. Boxplot showed intron retention rate (%) in each intron age group. Man-Whitney U test was performed, ***: p-value < 0.001, NS: p value > 0.05

C. Boxplot showed intron GC content (%) in each intron age group. Man-Whitney U test was performed, ***: p-value < 0.001, NS: p value > 0.05

D. Boxplot showed intron length (bp) in each intron age group. KS test was performed, ***: p-value < 0.001, NS: p value > 0.05

E. Boxplot showed expression in each gene group based on maximum intron age in gene. Man-Whitney U test was performed, ***: p-value < 0.001, NS: p value > 0.05

doi:10.6342/NTU202502289

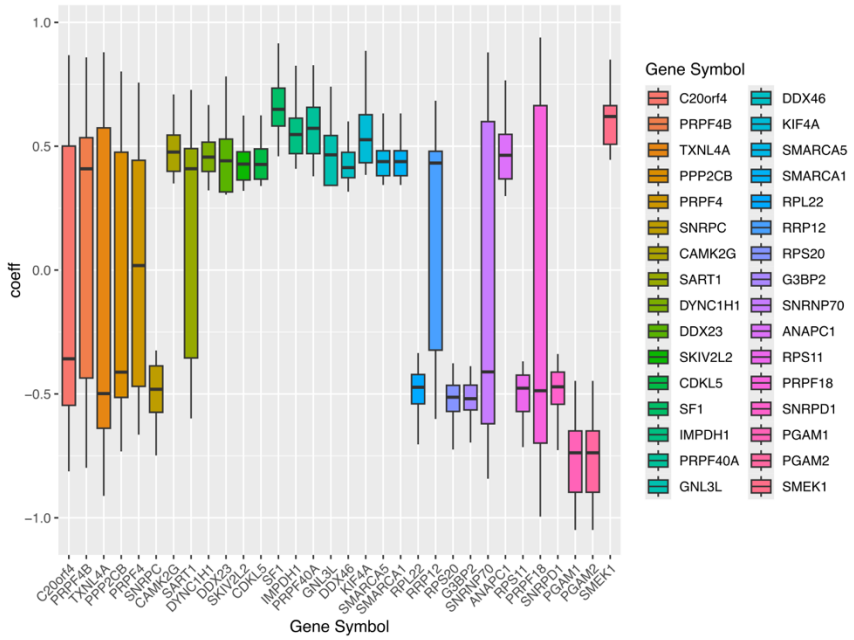
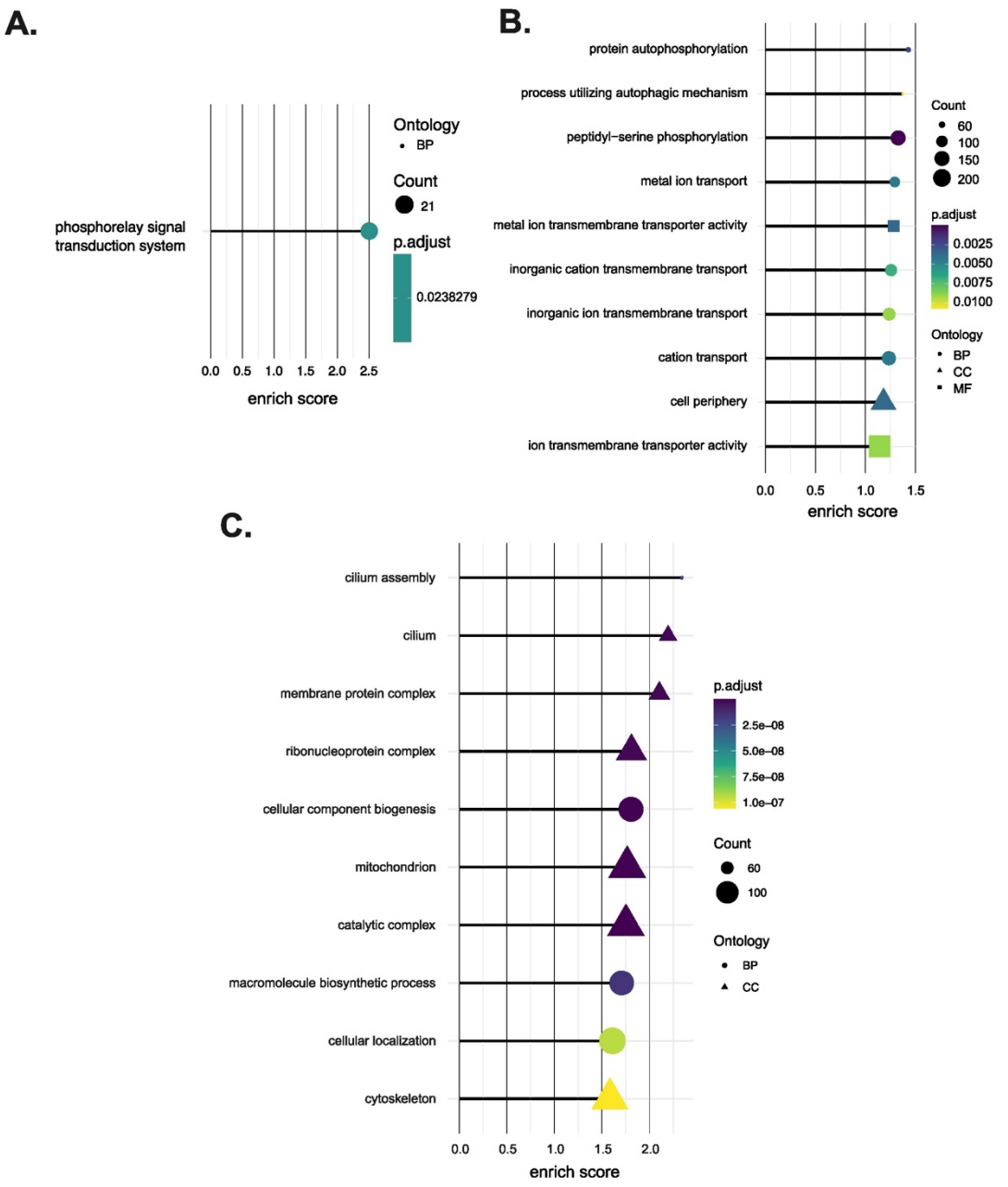


Figure 12. Coefficient distribution of top 30 Splicing-related DEGs have highest number of significant coefficient with DSIs

doi:10.6342/NTU202502289



doi:10.6342/NTU202502289

Figure 13. GO terms enrichment of genes based on maximum intron age groups

A. GO terms enrichment in genes with maximum intron age = 1 in *P. bursaria*

B. GO terms enrichment in genes with maximum intron age = 2 in *P. bursaria*

C. GO terms enrichment in genes with maximum intron age = 3 in *P. bursaria*

doi:10.6342/NTU202502289

Tables

Table 1. Summary of PSI distribution in 16 samples

Sample name	mean	median	Q1	Q3	Number of introns with PSI >= 0.1	% of introns with PSI <0.1
AM2 green	0.062	0.005	0.000	0.028	2195	89.45
AM2 white	0.070	0.010	0.001	0.039	2675	87.62
AM5 green	0.071	0.009	0.000	0.040	2726	87.33
AM5 white	0.073	0.011	0.001	0.041	2773	87.17
AM8 green	0.068	0.009	0.000	0.037	2613	87.93
AM8 white	0.076	0.013	0.002	0.045	2967	86.42
AM11 green	0.062	0.005	0.000	0.030	2305	89.05
Am11 white	0.068	0.008	0.000	0.034	2407	88.57
PM2 green	0.072	0.009	0.000	0.041	2837	86.91
PM2 white	0.071	0.009	0.000	0.037	2675	87.59
PM5 green	0.064	0.005	0.000	0.032	2408	88.53
PM5 white	0.065	0.006	0.000	0.030	2273	89.07
PM8 green	0.060	0.003	0.000	0.025	2054	89.94
PM8 white	0.065	0.008	0.000	0.032	2359	88.87
PM11 green	0.060	0.005	0.000	0.027	2072	89.93
PM11 white	0.069	0.009	0.000	0.037	2553	88.08

doi:10.6342/NTU202502289