Title
Transcriptional Regulation in Algae, Fungi and Plants: Mating Loci, Splicing, and miRNAs

Author
Douglass, Stephen Michael

Publication Date
2014

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Transcriptional Regulation in Algae, Fungi and Plants: Mating Loci, Splicing, and miRNAs

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Bioinformatics

by

Stephen Michael Douglass

2014

ABSTRACT OF THE DISSERTATION

Transcriptional Regulation in Algae, Fungi and Plants: Mating Loci, Splicing, and miRNAs

by

Stephen Michael Douglass

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2014

Professor Matteo Pellegrini, Chair

Genome-wide analysis of gene expression and regulation is important for elucidating basic principles of how cells function. In addition to gene expression, this dissertation will also discuss two methods of RNA-mediated RNA regulation: mRNA splicing and microRNAs. Three projects investigating gene expression and regulation using the Illumina platform are discussed here. The first project describes mRNA expression analysis of dozens of genes on the mating locus of the multicellular green alga, *Volvox carteri* in both the male and female mating types. The analysis describes sex-specific genes for both male and female mating types, and reveals the evolutionary history of the locus. The second project describes very low frequency splice products in a unicellular fungus with a fairly simple splicing landscape, *Saccharomyces cerevisiae*. These extremely rare splice events shed light on the mechanism of selecting splice sites in eukaryotic organisms. The third project describes an algorithm for predicting novel candidate microRNAs from small RNA sequence data. We describe a Naïve Bayes Classifier to differentiate microRNAs from contaminants and provide experimental validation of top candidates.

The dissertation of Stephen Michael Douglass is approved.

Huiying Li

Sabeeha Merchant

Xinshu Xiao

Matteo Pellegrini, Committee Chair

University of California, Los Angeles

2014

Table of Contents

Acknowledgements

I would like to thank my collaborators for their contributions to the projects described here, and for giving their permission to reprint figures and text from our manuscripts. Specifically, I would like to thank James Umen, Patrick Ferris, and Bradley Olson for their work drafting the mating loci manuscript and experimental work on the mating loci project, Guillaume Chanfreau and Tad Kawashima for drafting the yeast splicing manuscript, Tad Kawashima and Jason Gabunilas for experiential work on the yeast splicing project, Ssu-Wei Hsu for experimental validation of predicted miRNAs for the miRNA project, and all co-authors (listed after the title of each chapter) for critical reading and editing of manuscripts that contributed to this dissertation. I would also like to thank my committee and especially my thesis advisor, Matteo Pellegrini.

Vita

Education

2004-2008 BS, University of California at Los Angeles

Molecular Cell and Developmental Biology Major

Computer Programming Specialization

Conferences

2010-2014 UCLA Bioinformatics Retreat

2012-2013 Seed Institute

Publications

Douglass S, Hsu SH, Cokus S, Goldberg W, Harada J, Pellegrini M. A Naive Bayesian Classifier for Detecting miRNAs in Plants. In preparation.

Simpson L, **Douglass S**, Pellegrini M, Li F. Comparison of the complete mitochondrial genomes and transcriptomes of two strains of the lizard parasite, Leishmania tarentolae, by Next Generation Sequencing. In preparation.

Patterson M, Gaeta X, Loo K, Edwards M, Smale S, Azghadi S, **Douglass S**, Pellegrini M, Lowry W. The LIN28/let-7 circuit acts through Notch to control developmental maturity in the human nervous system. Submitted.

Kawashima T, **Douglass S**, Gabunilas J, Pellegrini M, Chanfreau GF. Widespread Use of Non-Productive Alternative Splice Sites in Saccharomyces cerevisiae. *PLoS Genetics*. 2014 April; 10.1371.

Fang W, Si Y, **Douglass S**, Casero D, Merchant SS, Pellegrini M, Ladunga I, Liu P, Spalding MH. Transcriptome-wide changes in Chlamydomonas reinhardtii gene expression regulated by carbon dioxide and the CO2-concentrating mechanism regulator CIA5/CCM1. *Plant Cell*. 2012 May; 24(5):1876-93.

Ferris P, Olson BJ, De Hoff PL, **Douglass S**, Casero D, Prochnik S, Geng S, Rai R, Grimwood J, Schmutz J, Nishii I, Hamaji T, Nozaki H, Pellegrini M, and Umen JG. Evolution of an expanded sex-determining locus in Volvox. *Science*. 2010 Apr 16;328(5976):351-4.

Chapter 1: Introduction

The ability to characterize and quantify transcripts provides insights that cannot be gained from genomic data alone. Analyzing transcripts is vital for discovering the functional units of the genome and for gaining insights into development, disease, and cellular response to environmental stimuli. Transcriptomics refers to the study of all RNA in the cell, including rRNAs, tRNAs, and small RNAs, although recent research has been primarily focused on the study of mRNA (1). Nucleic acid sequencing technologies have advanced remarkably over the past few decades, but the availability of next generation sequencing (NGS) technologies has dramatically increased our ability to sequence RNA in just a few short years (2).

RNA-Seq through NGS technologies has largely replaced older technologies in transcriptome analysis (3). Microarrays, a powerful tool for estimating relative transcript abundance, have become outdated due to background and cross-hybridization problems, inability to detect small differences in gene expression, and relatively narrow applicability compared to modern RNA-Seq (4). Other transcriptomic methodologies, such as tag-based sequencing, Serial Analysis of Gene Expression (SAGE), Cap Analysis of Gene Expression (CAGE), and Polony Multiplex Analysis of Gene Expression (PMAGE) also only focus on gene expression estimation, and while they can determine novel sequences, they are too laborious to be efficiently applied to entire transcriptomes (4). NGS RNA-Seq is efficient on a whole-transcriptome scale and allows for accurate estimation of gene expression and detection of novel genes and isoforms (4).

An important method of increasing transcript diversity in eukaryotic cells is alternative mRNA splicing. Alternative splicing allows a single gene to produce multiple transcripts by selecting

different coding regions, called exons, to incorporate into the mature mRNA sequence. Alternative splicing impacts ~95% of human genes with at least two exons (5), and it has been shown probabilistically that more than 60% of human disease-causing mutations affect splicing, as opposed to modifying the gene's coding sequence (6). The most dramatic currently known example of alternative splicing is Dscam gene in *Drosophila melanogaster*, which can create 38,016 splice variants from a single gene (7). Here, we will discuss alternative splicing in *Saccharomyces cerevisiae*, a unicellular fungi with a relatively simple splicing landscape. Only ~300 genes in *Saccharomyces* have been shown to be spliced with no examples of functional alternative splicing.

Gene expression is regulated by many factors including chromatin state, DNA methylation, availability of activators and repressors, and many other factors. An important method of gene expression regulation that was discovered recently is mRNA knockdown and silencing by microRNAs (miRNAs). miRNAs are small, ~22 nucleotide non-coding RNAs that dramatically reduce expression in target mRNA genes by binding through the RNA-induced silencing complex (RISC). Since being discovered in the nematode worm, *Caenorhabditis elegans*, in 1993, miRNAs have been found to impact thousands of genes across eukaryotic organisms (8).

In this dissertation, I describe my work on three applications of RNA-Seq using the Illumina platform: 1) Quantification of differential gene expression, 2) characterization of non-canonical splicing events, and 3) identification of novel microRNAs. Although RNA-Seq research is advancing rapidly, there is still progress that can be made in these three fields to contribute to our understanding of gene expression and regulation. To date there has been extensive use of RNA-Seq to determine differential expression between many conditions in many organisms (9-19), but despite considerable research there is still no single, clear RNA-Seq-based differential expression

workflow (4). Software exists to identify splice sites from RNA-Seq datasets given information about known splice junctions, but little effort has been made to identify non-canonical splice sites (10, 20). While the importance of small RNAs in gene regulation is well documented (21), a recent review of RNA-Seq and its applications does not mention small RNAs as a "main goal" of RNA-Seq experiments (1).

References

1. Wang Z, Gerstein M, and Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 2009 Jan;10(1):57-63.

2. McPherson JD. Next-generation gap. *Nature Methods*. vol. 6, no. 11S, pp. S2–S5, 2009.

3. Ozsolak F and Milo PM. RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, 12, 87-98.

4. Costa V, Angelini C, De Feis I, and Ciccodicola A. Uncovering the Complexity of Transcriptomes with RNA-Seq. *J Biomed Biotechnol*. 2010;2010:853916. Epub 2010 Jun 27.

5. Pan Q, Shai O, Lee LJ, Frey BJ, and Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*. 2008;40(12):1413-5.

6. Lopez-Bigas N, Audit B, Ouzounis C, Parra G, and Guigo R. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Letters*. 2005;579(9):1900-3.

7. Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE, and Zipursky SL. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*. 2000;101(6):671-84.

8. Lewis BP, Burge CB, and Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*. 2005;120(1):15-20.

9. Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, Go A, Delaney A, Griffith M , Hickenbotham M, Magrin V, Mardis ER, Sadar MD, Siddiqui AS, Marra MA, and Jones SJM. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics*, vol. 7, article 246, 2006.

10. Mortazavi A, Williams BA, McCue K, Schaeffer L, and Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.

11. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, and Soldatov A. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Research*, vol. 37, no. 18, article e123, 2009.

12. Blekhman R, Marioni JC, Zumbo P, Stephens M, and Gilad Y. Sex-specific and lineage-specific alternative splicing in primates. *Genome Research*, vol. 20, no. 2, pp. 180–189, 2010.

13. Rösel TD, Hung LH, Medenbach J, Donde K, Starke S, Benes V, Rätsch G, and Bindereif A. RNA-Seq analysis in mutant zebrafish reveals role of U1C protein in alternative splicing regulation. *EMBO J*, 2011 Apr 5.

14. Petkov SG, Marks H, Klein T, Garcia RS, Gao Y, Stunnenberg H, and Hyttel P. In vitro culture and characterization of putative porcine embryonic germ cells derived from domestic breeds and Yucatan mini pig embryos at Days 20-24 of gestation. *Stem Cell Research*, 2011 Jan 31.

15. Shi CY, Yang H, Wei CL, Yu O, Zhang ZZ, Jiang CJ, Sun J, Li YY, Chen Q, Xia T, and Wan XC. Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. *BMC Genomics*, 2011 Feb 28;12:131.

16. Lott SE, Villalta JE, Schroth GP, Luo S, Tonkin LA, and Eisen MB. Noncanonical compensation of zygotic X transcription in early Drosophila melanogaster development revealed through single-embryo RNA-seq. *PLoS Biol*, 2011 Feb 8;9(2):e1000590.

17. Rosenthal AZ, Matson EG, Eldar A, Leadbetter JR. RNA-seq reveals cooperative metabolic interactions between two termite-gut spirochete species in co-culture. *ISME J*, 2011 Feb 17.

18. Yang H, Lu P, Wang Y, and Ma H. The transcriptome landscape of Arabidopsis male meiocytes from high-throughput sequencing: the complexity and evolution of the meiotic process. *Plant J*, 2011 Feb;65(4):503-16.

19. Bruno VM, Wang Z, Marjani SL, Euskirchen GM, Martin J, Sherlock G, and Snyder M. Comprehensive annotation of the transcriptome of the human fungal pathogen *Candida albicans* using RNA-seq. *Genome Research*, 2010 Oct;20(10):1451-8.

20. Cloonan N, Xu Q, Faulkner GJ, Taylor DF, Tang DTP, Kolle G, and Grimmond SM. RNA-MATE: a recursive mapping strategy for high-throughput RNAsequencing data. *Bioinformatics*, vol. 25, no. 19, pp. 2615–2616, 2009.

21. Jacquier A. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nature Reviews Genetics*, vol. 10, no. 12, pp. 833–844, 2009.

Chapter 2: Evolution of an expanded sex-determining locus in *Volvox*

Patrick Ferris[1,*], Bradley J.S.C. Olson[1,*], Peter L. De Hoff[1], Stephen Douglass[2], David Casero Diaz-Cano[2], Simon Prochnik[3], Sa Geng[1], Rhitu Rai[1,4], Jane Grimwood[5], Jeremy Schmutz[5], Ichiro Nishii[6], Takashi Hamaji[7], Hisayoshi Nozaki[7], Matteo Pellegrini[2], and James G. Umen[1]

[1]The Salk Institute for Biological Studies, La Jolla, California 92037, USA
[2]Institute for Genomics and Proteomics and Department of Molecular, Cell and Developmental Biology, University of California, Los Angeles, California, 90095, USA
[3]U.S. Department of Energy (DOE) Joint Genome Institute (JGI), Walnut Creek, California, 95498, USA
[4]Laboratory of Plant Microbe Interaction, National Research Center on Plant Biotechnology, Pusa Campus, Indian Agricultural Research Institute, New Delhi-110012, India
[5]Genome Sequencing Center, Hudson Alpha Institute for Biotechnology, Huntsville, Alabama, 35906, USA
[6]Department of Biological Science, Nara Women's University, Nara 630-8506, Japan
[7]Department of Biological Sciences, University of Tokyo, Tokyo 113-0033, Japan
[*]These authors contributed equally to this work

Abstract

Although dimorphic sexes have evolved repeatedly in multicellular eukaryotes, their origins are unknown. The mating locus (*MT*) of the sexually dimorphic multicellular green alga, *Volvox carteri,* specifies the production of eggs and sperm and has undergone a remarkable expansion and divergence relative to *MT* from *Chlamydomonas reinhardtii*, a closely related unicellular species that has equal-sized gametes. Transcriptome analysis revealed a rewired gametic expression program for *Volvox MT*genes relative to *Chlamydomonas,* and identified multiple gender-specific and sex-regulated transcripts. The retinoblastoma tumor suppressor homolog *MAT3* is a *Volvox MT* gene that displays sexually regulated alternative splicing and evidence of gender-specific selection, both

indicative of cooption into the sexual cycle. Thus, sex-determining loci impact the evolution of both sex-related and non-sex-related genes.

Sexually dimorphic gametes have evolved in every major group of eukaryotes, and are thought to be selected when parents can differentially allocate resources to progeny (1). However, the origins of oogamy (large eggs and small sperm) and the contribution of sex determining loci to such evolution are largely unknown (2, 3).

The Volvocine algae are a group of chlorophytes comprising unicellular species such as *Chlamydomonas reinhardtii* (hereafter *Chlamydomonas*) and a range of multicellular species of varying complexity such as *Volvox carteri* (hereafter *Volvox*). *Volvox* has a vegetative reproductive form containing 16 large germ cells (gonidia) and ~2000 terminally differentiated somatic cells (4, 5) (Figure 2-4).

Chlamydomonas and other Volvocine algae also undergo a sexual cycle where a large, haploid mating locus (*MT*) controls sexual differentiation, mating compatibility, and zygote development (6). *MT* in *Chlamydomonas* is a 200–300 kb multigenic chromosomal region (Figure 2-4A) within which gene order is rearranged between the two sexes (*MT*+ and *MT*−) and meiotic recombination is suppressed, thus leading to its inheritance as a single Mendelian trait. Within each *MT* allele are gender-limited genes (allele present in only one of the two sexes) required for the sexual cycle as well as shared genes (alleles present in both sexes), most of which have no known function in sex or mating (7). The rearrangements that suppress recombination serve to maintain linkage of gender-limited genes, but they also reduce genetic exchange between shared genes leading to their meiotic isolation. Thus, *Chlamydomonas MT*bears similarity to sex chromosomes and to expanded mating type regions of some fungi and bryophytes (8–10).

While *Chlamydomonas* is isogamous (producing equal-sized gametes), *Volvox* and several other Volvocine genera have evolved oogamy that is under the control of female and male *MT* loci (11) (Figure 2-4). Moreover, the *Volvox* sexual cycle is characterized by a suite of other traits not found in *Chlamydomonas*—such as a diffusible sex-inducer protein rather than nitrogen deprivation (−N) as a trigger for gametogenesis (Table 2-1). A detailed characterization of *MT* in *Volvox* would be expected to shed light on the transition from isogamy to oogamy and on other properties of the sexual cycle that evolved in this multicellular species (Table 2-1).

The *MT*+ allele of *Chlamydomonas* was previously sequenced and resides on chromosome 6 (Figures 2-1A, 2-5) (12). To enable a comparison of mating loci evolution between two related species with markedly different sexual cycles, we sequenced *Chlamydomonas MT*− and both alleles of *Volvox MT* (Figure 2-1) (4). *Volvox MT* was previously assigned to Linkage Group I (LG I) (5) but the locus had not been further characterized. We mapped *Volvox MT* to the genome sequence and assembled most of LG I (Table 2-2) (4). Extensive synteny with *Chlamydomonas* chromosome 6 indicates that *MT* has remained on the same chromosome in both lineages for ~200 million years since their estimated divergence, despite numerous intra-chromosomal rearrangements between the two (Figure 2-5) (13).

While the haploid *Volvox* genome is ~17% larger than that of *Chlamydomonas* (138 Mb versus 118 Mb) and the two have very similar predicted proteomes (12, 14), *Volvox MT* is ~500% larger than *Chlamydomonas MT* and contains over 70 protein coding genes in each allele (Figure 2-1B). Compared to autosomes *Volvox MT* is unusually repeat-rich (>3X the genomic average), has lower gene density, and has genes with more intronic sequence (Table 2-3)—all properties that suggest an unusual evolutionary history and distinguish it from *Chlamydomonas MT*.

Only two gender-limited genes from *Chlamydomonas MT−* have recognizable homologs in *Volvox*—*MID* and *MTD1*—that are both in male *MT* (Figure 2-1). *MID* is a conserved RWP-RK family transcription factor whose expression in other Volvocine algae is induced by −N (15–17) as is also the case for *MTD1* (18, 19). Surprisingly, both *MTD1* and *MID* are expressed constitutively in *Volvox*, indicating that their transcription is uncoupled from sexual differentiation. This result suggests that additional *MT* genes might play a role in gametogenesis.

We used differential deep transcriptome sequencing (4) to identify *MT* genes in *Volvox*, a method that helped to mitigate problems associated with automated gene prediction in atypical genomic regions such as *MT*. We identified transcripts for five new female-limited and eight new male-limited genes that do not have detectable homologs in *Chlamydomonas*, and found that most of these gender-limited genes are sex-regulated (expression induced or repressed during sexual differentiation) (Figure 2-1C) (4). *HMG1* encodes a female-limited HMG domain protein that belongs to a family of DNA binding proteins whose members regulate mammalian and fungal sex determination (20, 21). However, HMG proteins had not been previously implicated in the sexual cycles of green algae or plants. A second novel female-limited gene, *FSI1*, is strongly induced during gametogenesis and encodes a small predicted transmembrane protein with no identifiable homologs (Figure 2-1C).

Besides identification of new gender-limited genes, our transcriptome data provided empirical support for 51 of 52 single-copy shared genes in *Volvox MT* that previously had limited EST support for the female allele (33 of 52), and no EST support for the male allele. Moreover, some of these shared genes showed patterns of expression suggesting cooption into the *Volvox* sexual cycle. These patterns include gender-biased expression (male:female expression ratio $\neq 1$) and sex-regulated expression (Figure 2-1C) (4). This set of genes encodes putative signaling,

extracellular matrix, and chromatin-associated proteins with known or potential roles in gametogenesis and fertilization, and are candidates for further investigation (Figure 2-6).

In diploid species heterogametic sex chromosomes evolve rapidly (22) and lose genes that are not related to sex (23). Genes within large haploid mating loci are predicted to accumulate mutations more rapidly than genes in autosomal regions due to suppressed recombination, but they are continuously exposed to selection (24). Suppressed recombination also appears to have played a role in diversification of mating-locus linked genes in haploid fungi and bryophytes (8–10). Our data allowed us to compare the evolutionary history of *Volvox MT* genes from this oogamous species to each other, and to genes from*MT* of its isogamous relative *Chlamydomonas*.

Divergence was measured from synonymous (dS) and non-synonymous (dN) substitutions (Figures 2-2A, 2-2B), and from total nucleotide distances for shared genes (4). Unexpectedly, divergence for*Volvox MT* allelic pairs is up to two orders of magnitude larger than for allelic pairs in *Chlamydomonas MT*, suggesting that *Volvox MT* alleles may have been subject to more intense and/or more prolonged recombinational suppression than *Chlamydomonas MT* alleles have been. In contrast, two internal syntenic blocks within *Volvox MT* are relatively similar (Figures 2-1B, 2-2A) suggesting that they were acquired more recently in an ongoing stratification process as first described for the human X chromosome (25). *Volvox MT* genes also showed reduced codon usage bias relative to autosomal genes, most likely due to suppressed recombination (26).

Three *MT* genes and a flanking gene, *PRP4*, were sequenced from a set of related *Volvox* species to determine the extent of *MT* gene isolation (4). Phylogenies revealed the expected pattern for *PRP4* that grouped by species and geographical location (Figure 2-2C). In contrast,

the *MT* genes grouped by gender (Figure 2-2D). These data demonstrate that the shared genes in *Volvox MT* have essentially become gender specific and have remained genetically isolated during speciation. Thus, the *MT* locus in *Volvox*has become a repository of genetic diversity that is linked to the sexual cycle.

In *Chlamydomonas* the retinoblastoma (RB) tumor suppressor pathway controls cell division in response to cell size (27), and the RB homolog encoded by *MAT3* is adjacent to *MT* (28). *Volvox MAT3*, on the other hand, is within *MT* (Figure 2-1B) and we investigated its evolution and expression as a candidate regulator of sexually dimorphic cell divisions (Figure 2-4).
The *Volvox* male and female MAT3 proteins are exceptionally diverged from each other (Figure 2-7). Moreover male and female*Volvox MAT3* have different structures: the female allele contains an intron that is absent from males while the male allele contains an unusually large fourth intron compared to females (Figure 2-3). Although*MAT3* shows signs of having undergone purifying selection (dN/dS=0.23), several short sequences in the male and female proteins are asymmetric in their conservation pattern, suggesting that the two alleles are under different selective constraints. We also found dozens of alternatively processed *MAT3* mRNAs from both *Volvox* sexes, representing most types of alternative splicing (29) (Figure 2-3). In addition, sex-regulated pre-mRNA splicing of *MAT3* was found for both genders, and might be controlled by the *MT* encoded splicing factor, *SPL2*, whose expression level is sex-regulated in males (Figure 2-1C). Intriguingly, the predominant *MAT3* isoform in sexual males retains the first two introns, leading to inclusion of an early termination codon (Figure 2-3). *mat3* mutants in *Chlamydomonas* produce tiny gametes (28), and down-regulation of *MAT3* in *Volvox* males through alternative splicing may be linked to the production of small-celled sperm.

The accelerated divergence of sex chromosomes is usually associated with gene loss and degeneration (23), though adaptive evolution of sex chromosomes is an emerging theme (30). Our data suggest that expansion, loss of recombination, and rapid divergence can be mutually reinforcing properties of sex determining regions that facilitate cooption into the sexual cycle and provide novel sources of developmental innovation.

Figures

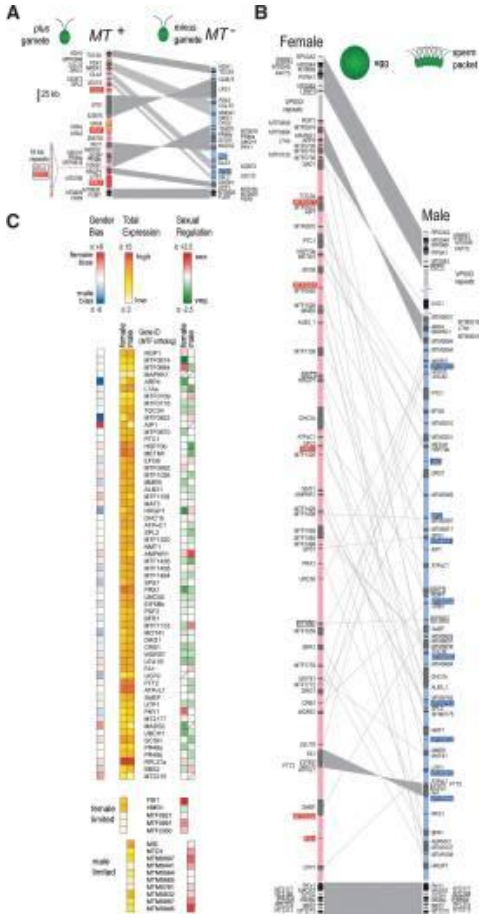

Figure 2-1. Expansion of *Volvox MT* and sex-regulated gene expression. (A) Schematic of *Chlamydomonas* mating locus with rearranged domains in light blue or pink. *MT*+ limited genes are shaded red if unique or orange if they have an autosomal copy. *MT*− limited genes are shaded blue. Flanking and shared genes are shaded black and gray respectively. Synteny is indicated by gray shading. (B) Schematic of*Volvox MT* scaled as in (A). Boxed genes were used for mapping. The broken segment represents a transposon repeat region containing copies of *VPS53*. (C) Expression heat maps of *Volvox MT* genes. Left panel, female/male expression ratio; middle panel, total expression; right panel, sexual induction (Sex) or repression (Veg). Diagonal hatch, insufficient data.

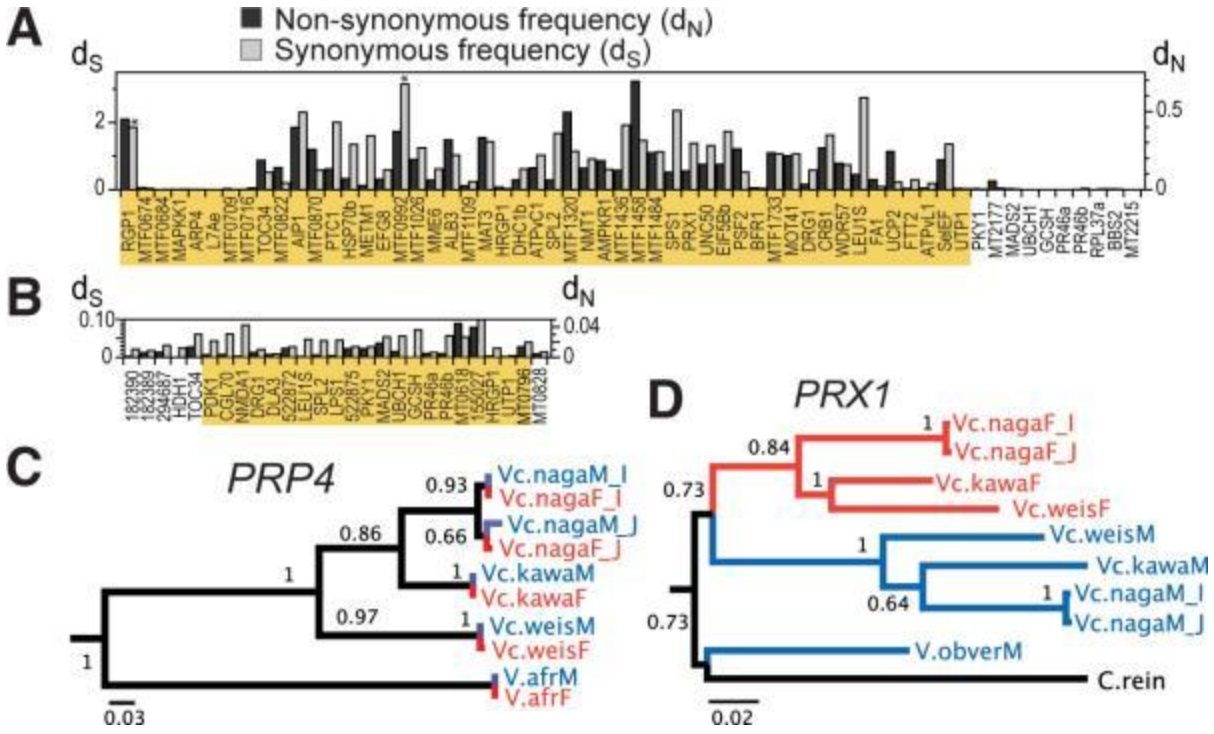

Figure 2-2. Divergence of *MT* genes. (A,B) d_N and d_S for shared *Volvox* (A) or *Chlamydomonas* (B) genes within *MT* (orange shading) or flanking*MT*. Asterisks mark saturated d_S values. (C,D) Maximum likelihood phylogenies for *PRP4* (C) and *MT* gene *PRX1* (D). Red/blue signify female/male strains and clades.

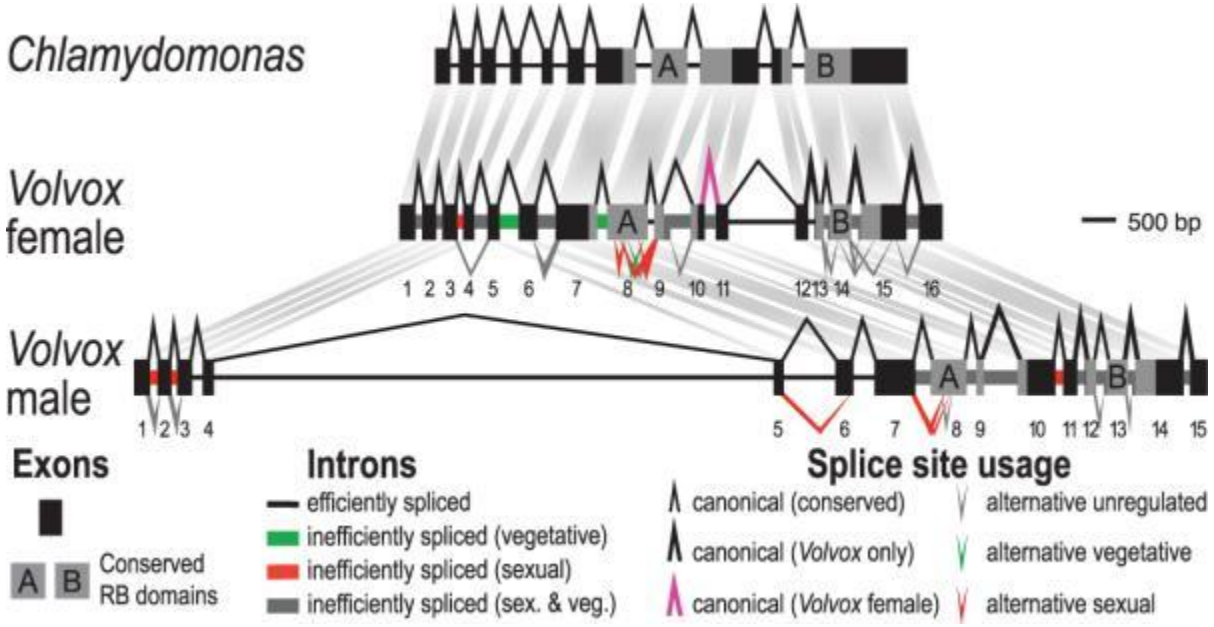

Figure 2-3. Gender-specific divergence and splicing of *MAT3*. Schematic of *MAT3* from *Chlamydomonas* (top), *Volvox* female (middle) and *Volvox*male (bottom). *Volvox* exons are numbered.

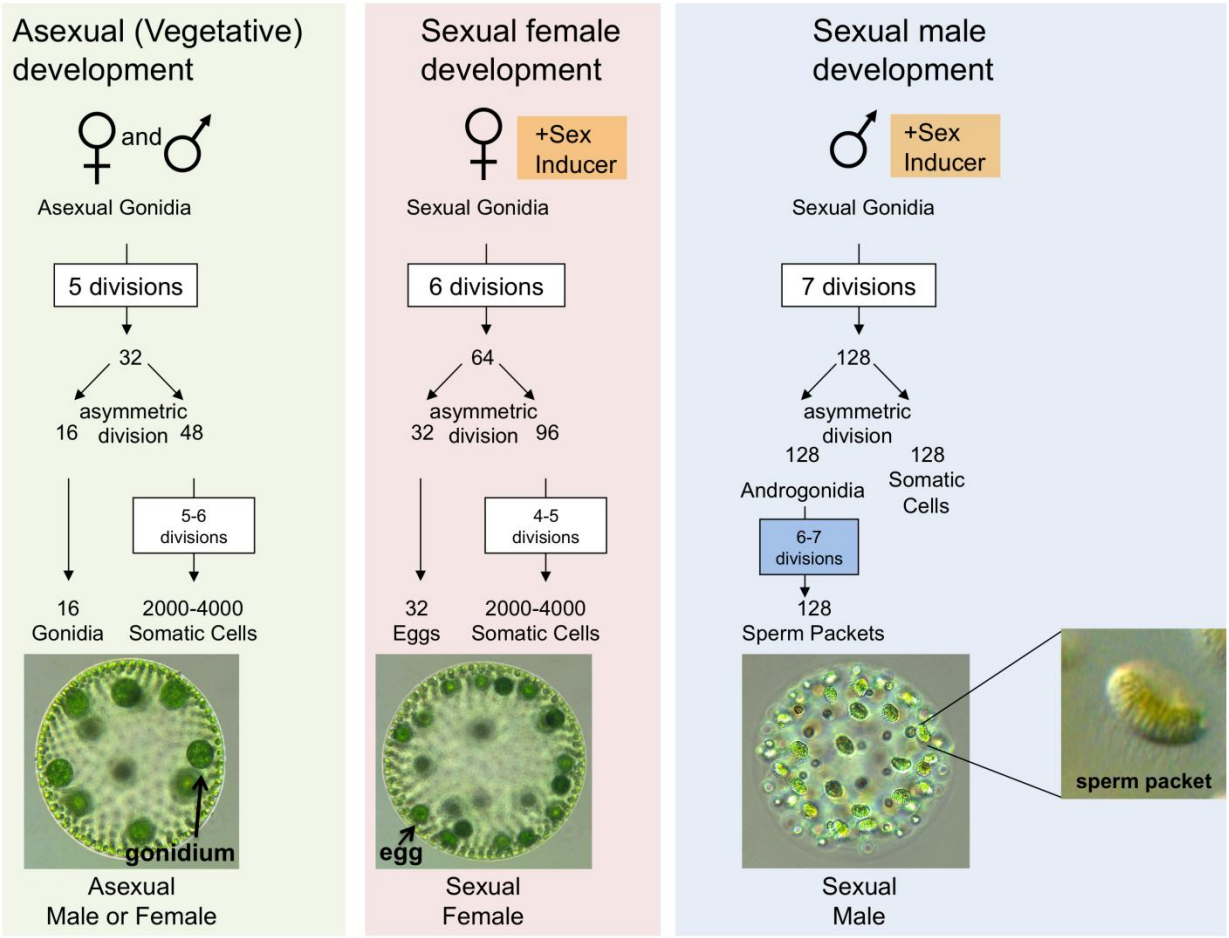

Figure 2-4. Sexual development in *Volvox*. Schematic of vegetative (asexual) and sexual development in *Volvox carteri* with photographs of each type of spheroid. Adapted from (*28*). The left panel with green shading shows male and female asexual development that have the same embryonic cleavage pattern for both sexes. Each mature gonidium cleaves a total of 11 or 12 times to make a new spheroid. Asymmetric division of 16 anterior cells at cycle 6 (32!64 cells) generates 16 gonidial precursors that form the asexual germ cells for the next generation. The middle panel (pink shading) shows the modified developmental pattern of females whose gonidia were exposed to sex inducer prior to cleavage. The pattern is similar to asexuals except that the asymmetric division occurs one cycle later (64→128 cells) in up to 48 anterior cells (typically in 32 cells as shown here) resulting in the production of eggs that are smaller than asexual gonidia and that can be fertilized by sperm. The right panel (blue shading) shows male sexual development where asymmetric cell division occurs in all cells two cycles later than in asexual (128→256 cells) to produce 128 small sexual somatic cells and 128 large androgonidia. The large androgonidia undergo a second set of cleavage divisions the following day (blue shaded box) to produce sperm packets containing 64 or 128 sperm cells. After their release individual sperm packets swim as a unit until they reach a female spheroid where they dissociate prior to entry and fertilization of eggs (*29*).

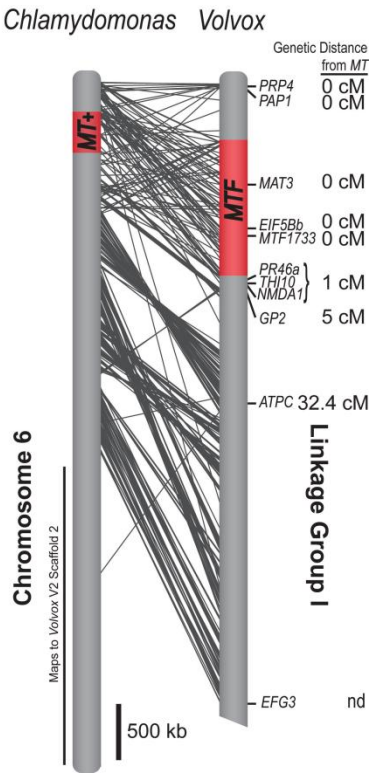

Figure 2-5. **Synteny between *C. reinhardtii* Chromosome 6 and *V. carteri* Linkage Group I.** Positions of orthologous genes from *Chlamydomonas* Chromosome 6 (*MT*+) and *Volvox* Linkage Group I (female) are depicted by connecting gray lines. The mating locus is shaded red. Molecular markers and their genetic distances from *MT* are indicated for *Volvox*. The distal portion of Chromosome 6 maps to *Volvox carteri* Version 2 Scaffold 2.

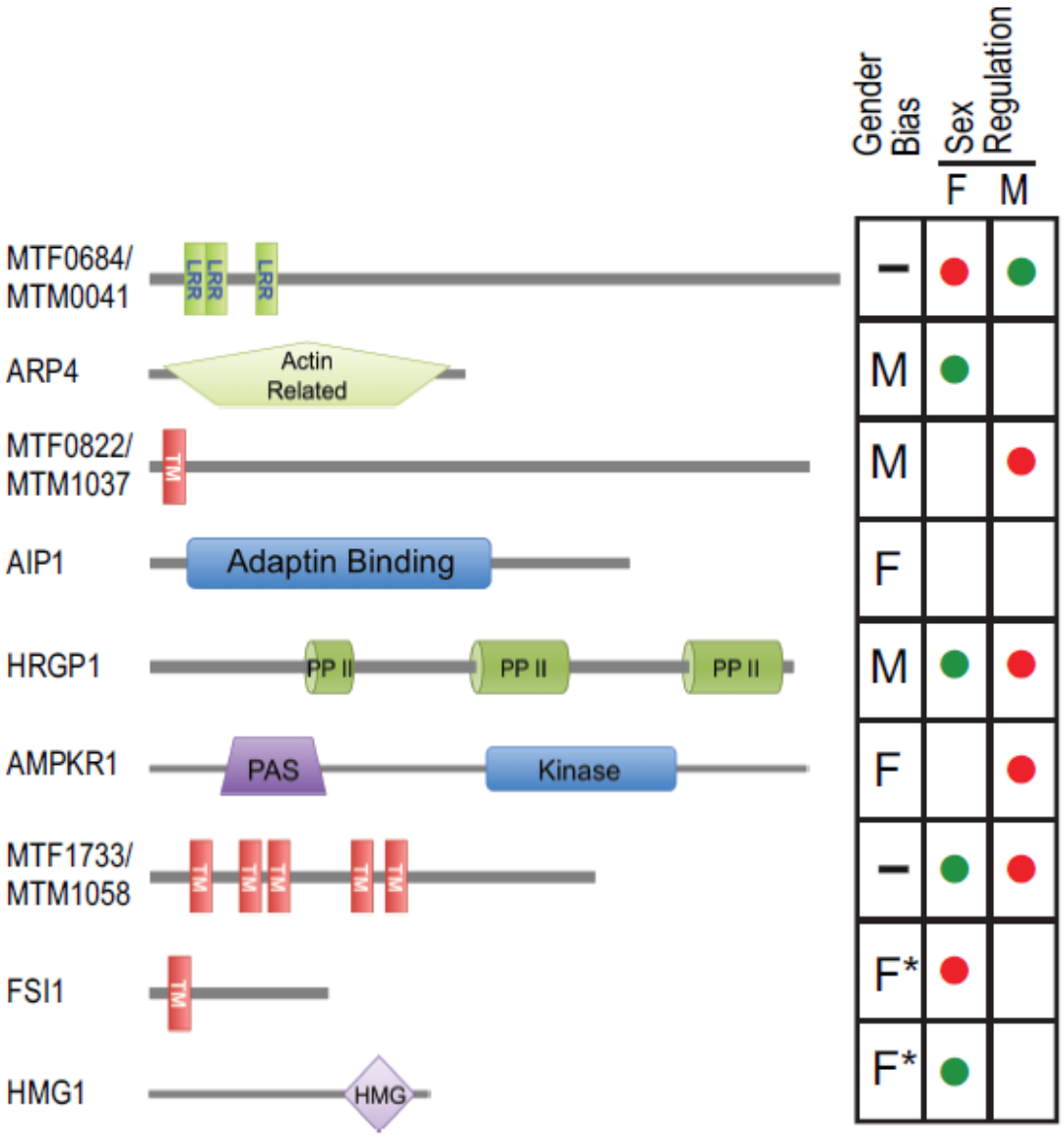

Figure 2-6. **Domain structures and expression patterns of sex regulated *Volvox* MT genes**. The left side shows domain structures of predicted proteins (*31*). LRR—leucine rich repeat (*32*). ARP4—nuclear actin related protein involved in chromatin formation(*33*).TM—transmembrane domain. Adaptin binding protein domain (*34*) involved in membrane trafficking (*35*). PP II—polyproline II helix (*36*) involved in sexual signaling (*37*). PAS-kinase domain involved in signal transduction (*38*). HMG—high mobility group protein some of which are involved in sex determination (*39*). The table to the right summarizes expression patterns from selected genes in Fig. 1C. Gender bias indicates greater total expression in females (F) or males (M). (-) indicates no overall expression bias. F*—female limited gene. Sex regulation is indicated by colored dots that represent higher expression in sexual (red) or vegetative (green) samples from within each sex as indicated in the column header, with F for female and M for male.

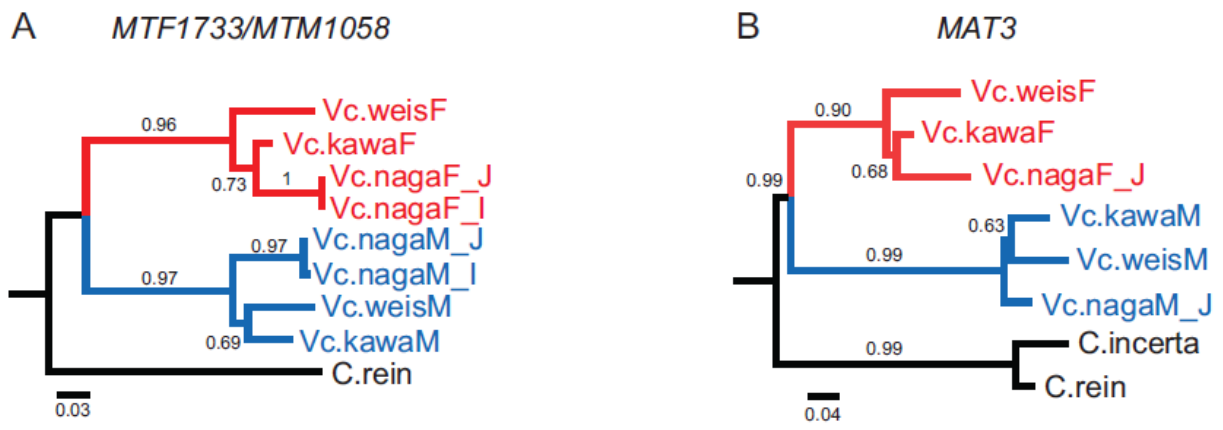

Figure 2-7. Gender-based phylogenies of genes residing in *Volvox* MT. (A) Maximum Likelihood tree of the coding sequence of *MTF1733/MTM1058* (unknown conserved gene). Species names and abbreviations are in Table S13. Female isolates and clades are colored blue. Male isolates and clades are colored red. Scale bars denote substitution rate. Numbers next to branch points indicate Likelihood Ratio Test values that are analogous to bootstrap values. (B) Maximum Likelihood tree of MAT3 deduced partial protein sequence. Labels and color coding are the same as in (A). Genbank IDs: *C. reinhardtii MTF1733/MTM1058*, GI:159477131; *C. reinhardtii* MAT3, GI:14573436; *C. incerta* MAT3, GI:55140503.

Tables

Chlamydomonas reinhardtii	***Volvox carteri f. nagariensis***
Nitrogen starvation is signal for gametogenesis.	Diffusible sex inducer protein is signal for gametogenesis
Isogamous (equal sized gametes)	Oogamous (eggs and sperm)
All vegetative cells can differentiate into gametes directly.	Eggs and sperm packets formed by only a subset of cells during sexual spheroid embryogenesis.
Gametes can de-differentiate back to vegetative cells.	Eggs can de-differentiate. Sperm are terminally differentiated.
Gametes are free-swimming single cells that find each other by chance.	Motile sperm packets must find a sexual female spheroid, gain entry, dissociate, and fertilize eggs.
Zygotes undergo meiosis to form tetrads (all four products viable)	One large meiotic product survives; 3 small polar bodies are non-viable.
Uniparental inheritance of chloroplast genome from MT+ parent and mitochondrial genome from MT– parent.	Uniparental inheritance of chloroplast and mitochondrial genomes both from female parent.

Table 2-1. Comparison of Chlamydomonas reinhardtii and Volvox carteri sexual cycles.

Gene Name	*V.cart.* JGI (V1) protein ID	*V.cart.* V2 scaffold 2 position	Position based on *V.cart.* female *MT* merged with genome V2 scaffold 2	Physical distance from *MT* border (kb)	Recombinants with *V.cart. MT*	Genetic distance from *V.cart. MT* in centimorgans (cM)	*C.rein.* JGI (V4) location
PRP4	100336		65424	nd	0/107	0.0 cM	Chromosome 6:122356-130070
PAP1	100339		93513	nd	0/107	0.0 cM	Chromosome_6:1484008-1489887
MAT3	127376		1162641	na (*MT* gene)	0/93	0.0 cM	Chromosome 6:939371-945298
EIF5Bb	121771		1650659	na (*MT* gene)	0/93	0.0 cM	Chromosome 6:358313-362377
MTF1733	84244		1733026	na (*MT* gene)	0/93	0.0 cM	Chromosome 6:109448-112101
PR46a	127492		2194703	58	1/106	1 cM	Chromosome 6:613946-615559
THI10	107906		2236670	100	1/106	1 cM	Chromosome 6:931565-934443
NMDA1	72329	4760911	2267879	130	1/106	1 cM	Chromosome 6:441396-443425
GP2	127271	4685286	2343504	206	4/78	5.1 cM	Chromosome 6:1374807-1379615
ATPC	102786	3436900	3591890	1455	12/37	32.4 cM	Chromosome 6:1514578-1517615
EFG3	78065	81578	6947212	4810	nd	nd	Chromosome 6:4156041-4160199

Table 2-2. Genetic mapping data for *Volvox carteri* MT. The *Volvox* protein IDs are from JGI http://genome.jgi-psf.org/Volca1/Volca1.home.html. nd, not determined. na, not applicable for genes within MT. Physical distances between PRP4 or PAP1 and MT border could not be determined because the size of intervening VPS53 repeat region is unknown.

	Female	*Male*	Genome
Total Sequence (Mbp)	1.51	1.13	138
% Repeats	72	70	21
% GC	52	53	56
Average gene density (per Mbp)	39	54	113
Average gene length (bp)	7198	6062	5300
Average coding (CDS) length (bp)	1883	1824	1800
Average total intron length per gene (bp)	5315	4542	3500
Average number of exons per gene	10	9	8
Average exon length (bp)	196	208	236
Average intron Length (bp)	618	584	500

Table 2-3. Gene content, repeats and statistics for *Volvox carteri* MT. The *Volvox* genome data is based on http://genome.jgi-psf.org/Volca1/Volca1.home.html.

References

1. Parker GA, Baker RR, Smith VG. J. Theor. Biol. 1972;36:529.

2. Charlesworth B. J. Theor. Biol. 1978;73:347.

3. Williams TM, Carroll SB. Nat Rev Genet. 2009;10:797.

4. Information on Materials and Methods is available on Science Online.

5. Kirk DL. Volvox. In: Bard JBL, Barlow PW, Green PB, Kirk DL, editors. Developmental and Cell Biology Series. Cambridge University Press; Cambridge, U.K.: 1998.

6. Goodenough U, Lin H, Lee J-H. Semin. Cell Dev. Biol. 2007;18:350.

7. Ferris PJ, Armbrust EV, Goodenough UW. Genetics. 2002;160:181.

8. Fraser JA, et al. Plos Biol. 2004;2:e384.

9. Menkis A, Jacobson DJ, Gustafsson T, Johannesson H. PLoS Genet. 2008;4:e1000030.

10. Yamato KT, et al. Proc Natl Acad Sci USA. 2007;104:6472.

11. Nozaki H. J Plant Res. 1996;109:353.

12. Merchant SS, et al. Science. 2007;318:245.

13. Herron MD, Hackett JD, Aylward FO, Michod RE. Proc. Natl. Acad. Sci. USA. 2009;106:3254

14. http://genomeportal.jgi-psf.org/Volca1/Volca1.home.html.

15. Ferris PJ, Goodenough UW. Genetics. 1997;146:859.

16. Hamaji T, et al. Genetics. 2008;178:283.

17. Nozaki H, Mori T, Misumi O, Matsunaga S, Kuroiwa T. Curr Biol. 2006;16:R1018.

18. Hamaji T, Ferris PJ, Nishii I, Nozaki H. J Phycol. 2009:1.

19. Lin H, Goodenough UW. Genetics. 2007;176:913.

20. Idnurm A, Walton FJ, Floyd A, Heitman J. Nature. 2008;451:193.

21. Waters PD, Wallis MC, Marshall Graves JA. Semin. Cell Dev. Biol. 2007;18:389.

22. Hughes JF, et al. Nature. 2010;463:536.

23. Charlesworth B, Charlesworth D. Philos Trans R Soc Lond, B, Biol Sci. 2000;355:1563.

24. Bull J. The American Naturalist. 1978;112:245.

25. Lahn BT, Page DC. Science. 1999;286:964.

26. Kliman RM, Hey J. Mol. Biol. Evol. 1993;10:1239.

27. Fang S-C, de los Reyes C, Umen JG. PLoS Genet. 2006;2:e167.

28. Umen JG, Goodenough UW. Genes Dev. 2001;15:1652.

29. Matlin A, Clark F, Smith C. Nat Rev Mol Cell Biol. 2005;6:386.

30. Bachtrog D. Curr. Opin. Genet. Dev. 2006;16:578.

Chapter 3: Widespread Use of Non-productive Alternative Splice Sites in Saccharomyces cerevisiae

Tadashi Kawashima[1], Stephen Douglass[2], Jason Gabunilas[1], Matteo Pellegrini[3], and Guillaume F. Chanfreau[1]

[1]Department of Chemistry and Biochemistry and the Molecular Biology Institute, UCLA, Los Angeles, California, United States of America
[2]Bioinformatics Interdepartmental Program, UCLA, Los Angeles, California, United States of America
[3]Department of Molecular, Cellular and Developmental Biology, UCLA, Los Angeles, California, United States of America

Abstract

Saccharomyces cerevisiae has been used as a model system to investigate the mechanisms of pre-mRNA splicing but only a few examples of alternative splice site usage have been described in this organism. Using RNA-Seq analysis of nonsense-mediated mRNA decay (NMD) mutant strains, we show that many *S. cerevisiae* intron-containing genes exhibit usage of alternative splice sites, but many transcripts generated by splicing at these sites are non-functional because they introduce premature termination codons, leading to degradation by NMD. Analysis of splicing mutants combined with NMD inactivation revealed the role of specific splicing factors in governing the use of these alternative splice sites and identified novel functions for Prp17p in enhancing the use of branchpoint-proximal upstream 3′ splice sites and for Prp18p in suppressing the usage of a non-canonical AUG 3′-splice site in *GCR1*. The use of non-productive alternative

splice sites can be increased in stress conditions in a promoter-dependent manner, contributing to the down-regulation of genes during stress. These results show that alternative splicing is frequent in *S. cerevisiae* but masked by RNA degradation and that the use of alternative splice sites in this organism is mostly aimed at controlling transcript levels rather than increasing proteome diversity.

Introduction

Nonsense-mediated mRNA decay (NMD) is an RNA degradation system that degrades RNAs containing premature termination codons (1, 2). In mammalian cells and higher eukaryotes, NMD can be used to regulate gene expression, for instance by reducing the level of alternatively spliced isoforms containing premature termination codons (3-8). This interplay between alternative splicing and NMD is involved in the autoregulation of SR proteins (3-5). In addition to its function in regulating non-productively spliced isoforms, NMD is also used in a variety of eukaryotes to degrade unspliced pre-mRNAs that have escaped the splicing machinery (9-13). Thus, NMD is widely involved in the proofreading of splicing efficiency and accuracy.

The yeast *Saccharomyces cerevisiae* has long been used as a model system to investigate the mechanisms of pre-mRNA splicing, as many components of the splicing machinery were identified through genetic screens in *S. cerevisiae* (14), and most splicing factors are highly conserved from yeast to mammalian cells (15). Despite the presence of *c.a.* 330 intron-containing genes in *S. cerevisiae*, the prevalence of alternative splicing in this organism remains largely unexplored, as only a few examples of alternative splice site selection have been documented. The *SRC1* gene encodes an integral transmembrane protein, for which the use of an alternative 5′-splice site changes the number of passes through the membrane and ultimately the location of the C-terminal end of Src1p (16, 17). Alternative 3′-splice site selection has been

shown to regulate expression of the *APE2* gene according to a temperature-dependent secondary structure in the transcript (18). A few other alternative 3′-splice sites have been described, and the use of some of these sites produces transcripts that are degraded by NMD (19). Recent work analyzing alternative splicing across fungal species has shown that *S. cerevisiae* has lost some of the alternative splicing events through gene duplication and sub-functionalization of the duplicated genes, which are otherwise produced by alternative splicing in other species (20).

Using RNA-Seq analysis of strains mutated for NMD factors, we identify here a large number of alternative splice sites in *S. cerevisiae*. However, we show that splicing at these sites is generally non-productive because it introduces premature termination codons (PTC), leading to degradation of the transcripts by NMD. Non-productive splicing can be increased during environmental stress to contribute to a global regulatory mechanism that down-regulates transcripts levels in response to environmental cues. These results show that non-productive splice sites are widely used in *S.cerevisiae*, but that transcripts spliced at these sites are eliminated by RNA quality control mechanisms. Thus, while alternative splicing is frequently utilized in higher eukaryotes to generate proteome diversity, it is mainly used in *S.cerevisiae* as a means to regulate transcript levels.

Results

RNA-Seq reveals the accumulation of a large number of non-productive splice variants in NMD mutants

We previously showed that NMD degrades unspliced transcripts arising from a large fraction of intron-containing genes in *S. cerevisiae*, due to suboptimal splice sites (12, 13), or upon splicing factor inactivation (21). In addition, recent data showed that transcripts generated by the use of

alternative 3′-splice sites can be degraded by NMD (19). To gain further insights into the function of NMD in the proofreading of spliced isoforms, we performed RNA sequencing of mRNAs from wild-type and isogenic*upf1Δ*, *upf2Δ* and *upf3Δ* strains defective for NMD. To identify transcripts spliced at alternative splice sites, we performed gapped alignment analysis of the RNA sequences (Table 3-1) using BLAT (22). This analysis revealed numerous occurrences of spliced transcripts arising from previously unknown splice sites, in both WT and the NMD mutants. We will refer to these new splicing events as alternative splicing events, even if these are found in wild-type cells, and to the annotated splicing events as the normal or canonical splicing events. Alternative splicing events were detected more frequently in RNA samples obtained from the NMD mutants (Figures 3-1A, 3-1B; Table 3-2), consistent with the fact that most of these alternative splicing events result in the introduction of a PTC, either by inducing a translational frameshift or by inserting an intronic PTC-containing sequence (Table 3-2). After adjusting for sequencing depth, *upf1Δ*, *upf2Δ* and *upf3Δ* showed a 1.67, 1.72, and 1.90-fold enrichment in alternative splicing events and 1.59, 1.70, and 1.79-fold enrichment in PTC-generating alternative splicing events, respectively, versus wild-type (Table 3-2). NMD mutants showed an approximately 1.7-fold increase in unspliced mRNAs compared to the wild-type (Table 3-3) when considering reads that map to intronic and exon-intron regions, confirming our previous results from tiling arrays showing the involvement of NMD in eliminating unspliced transcripts genome-wide (12). This enrichment for unspliced RNAs in NMD mutants is probably underestimated. Although there were 4-fold more reads that mapped only to intronic regions in NMD mutants compared to wild-type (Table 3-3), we observed an unanticipated high number of reads that mapped to exon-intron junctions in the wild-type strain (Table 3-3), which lowered the overall enrichment for unspliced RNAs in NMD mutants.

There was limited overlap in the alternative splicing events identified in the three *UPF* mutants (Figure 3-1C), suggesting that the depth of our sequencing analysis was not sufficient to saturate identification of all alternative splicing events, particularly those occurring at low frequencies. The list of intron containing genes (ICG) for which we did not find the use of alternative splice sites is provided in Table 3-4. 97 out of 304 intron containing genes analyzed did not exhibit alternative splicing (Table 3-4). Whether this reflects the absence of competing alternative sites or the lack of depth of our sequencing analysis remains to be determined.

To investigate if alternative splicing events are due to rare events or to splicing errors that occur randomly during transcript expression, we examined the abundance of ICG mRNAs that exhibited alternative splicing events and that of ICG mRNAs for which no alternative splicing events were detected (Figure 3-1D). This analysis showed that some low abundance transcripts exhibited alternative splicing, while some high abundance transcripts did not (Figure 3-1D). In addition, the median abundance of genes that showed alternative splicing was 117 RPKM, while the median abundance for genes with no alternative splicing events detected was 136 RPKM. Thus, even if the most highly expressed ICG (>2200 RPKM) all exhibited alternative splicing (Figure 3-1D), genes with no alternative splicing were in general expressed at higher level than genes for which alternative splicing events were detected, showing no clear correlation between transcript abundance and the detection of alternative splicing events. We conclude that the detection of alternative splicing events in our RNA-Seq analysis is not an indirect consequence of the higher number of reads for highly-expressed transcripts.

The consensus sequences derived from the alternative splicing events identified in wild-type and all three mutants exhibited differences from the consensus sequences derived from the canonical (normal) splicing events (Figure 3-1E). Alternative 5′-splice sites showed a relaxation of the

conserved sequences, especially at positions 4 and 6 compared to the consensus obtained from the canonical splicing events. The 3′-splice sites used in alternative splicing events also showed a decrease in conservation of the polypyrimidine sequence preceding the conserved YAG, as well as a weaker conservation of the pyrimidine preceding the conserved AG dinucleotide (Figure 3-1E). Thus, alternative splice sites identified by RNA sequencing showed a relaxed conservation, suggesting that these might correspond to lower efficiency splice sites, and possibly to regulated splicing events. Finally, we identified a number of alternative splicing events in either wild-type or NMD mutants that do not introduce a PTC and would potentially result in the production of proteins that differ from the SGD annotations. However, we did not investigate these alternative protein forms further because most of the RNAs that would result in the production of these proteins were found in low abundance compared to those resulting in the production of the annotated proteins.

Strategy for validation of alternative splicing events

The previous RNA-Seq analysis revealed the potential widespread usage of alternative splice sites (SS). Figure 3-2 depicts specific mRNAs that were chosen for validation and further characterization. These transcripts were classified into three classes: those with 1) alternative 5′-SS; 2) alternative 3′-SS; and 3) a combination of both. Transcripts from class 1 included *RPL22B* as well as the previously reported*SRC1* (16). Class 2 transcripts included genes encoding the RNA Polymerase III transcription factor*TFC3* with a downstream alternative 3′-SS, and the adenosine deaminase *TAN1* with two alternative 3′-SS flanking the normal 3′-SS. For the third class, we examined genes encoding the glycosylphosphatidylinositol biosynthetic enzyme *GPI15* and the transcriptional regulator *GCR1*. *GPI15*exhibited the use of an alternative 5′-SS with the normal 3′-SS, as well as the normal 5′-SS with an alternative 3′-SS (Figure 3-

2). *GCR1* showed a more complex splicing pattern with multiple combinations of 5′ and 3′-SS (Figure 3-2).

We analyzed alternative splicing events by RT-PCR using Cy3-end labeled primers, which allowed for relative comparison of the abundance of spliced and unspliced species, regardless of their size. Because we lacked an adequate size marker for Cy3 detection, the same RT-PCR analyses were initially performed with ^{32}P-end labeling with an appropriate ^{32}P-labelled ladder (data not shown) to confirm the sizes of all RT-PCR products and correlate the data back to gels obtained with Cy3-labeled primers. In addition to the wild-type and NMD-deficient *upf1Δ* strains, we analyzed the phenotypes of a number of *S. cerevisiae* splicing mutants. Knockout mutants of genes encoding Mud1p and Nam8p were chosen for their association with the U1 snRNP and role in 5′-SS selection (23-26). The *HUB1*knockout was also included, as Hub1p was recently implicated in 5′-SS selection for *SRC1* (17). Prp17p and Prp18p were selected for their involvement in the second step of splicing and potential effects on 3′-SS selection (27, 28). Finally, Isy1p was also included as a potential splicing fidelity factor (29). The splicing profiles were analyzed for each of the genes mentioned above in each of these mutant strains by fractionation of the RT-PCR products on polyacrylamide gels (Figure 3-3). For the splicing mutants for which the splicing pattern differed from the wild-type, additional RT-PCR experiments were performed in triplicate from three independent cultures and quantitated, as shown in Figures 3-4, 3-5, 3-6, 3-7 and 3-8.

RT-PCR analysis confirms the involvement of Prp17p and Hub1p in *SRC1* alternative splicing

As a first step in validating our RT-PCR strategy, we focused on *SRC1*, which exhibits two possible 5′-SS (Figure 3-2) and for which previous studies have demonstrated the roles of

various splicing factors (16, 17, 30). RT-PCR analysis of *SRC1* splice variants confirmed the use of these two alternative 5′-SS (Figure 3-3). Wild-type samples showed a 60/40 ratio of *SRC1-S*/*SRC1-L* (Figure 3-4), consistent with previous reports (16, 17, 30). Samples from the *upf1Δ* mutant showed a pattern similar to wild-type (Figures 3-3 and 3-4), indicating that both variants are stable and not targeted by NMD. This result is consistent with our RNA-Seq analysis, which showed high sequence counts for both forms in all strains. Samples from the*nam8Δ* strain showed a slight increase in the level of unspliced transcripts (Figure 3-3) due to reduced splicing efficiency (31). The *prp17Δ* and *prp18Δ* mutants both showed a slight increase in the usage of*SRC1*-L 5′-splice site relative to the *SRC-S* 5′-splice site (1.4 and 1.3 fold, respectively), as suggested previously for the *prp17Δ* mutant at the protein level (17), and the *prp18Δ* mutant also exhibited an increase in unspliced precursors accumulation, consistent with previous results for other transcripts (21). The *isy1Δ* mutant strain exhibited a clear accumulation of unspliced pre-mRNAs (Figure 3-3), in agreement with the documented role of Isy1p in maintaining the proper conformation needed for the 1st step of splicing (29). Hub1p inactivation resulted in a 3-fold reduction in the amount of *SRC1-S*, coinciding with an increase in *SRC1-L* (Figures 3-3 and 3-4), consistent with previous reports (17, 30). This reduction was also observed in the context of the *upf1Δ* mutant (Figures 3-3 and 3-4). Thus, the results described above confirmed the previously described effects of various splicing mutants on *SRC1* splicing patterns and showed that our RT-PCR strategy is effective in analyzing the impact of specific splicing factors on splice site usage.

Efficient use of the non-productive 5′-splice site of *RPL22B* is strongly dependent on the U1 snRNP components Nam8p and Mud1p

RPL22B showed the presence of an alternative 5′-SS in the intronic sequence, which unlike *SRC1*, yields a PTC-containing transcript potentially targeted to NMD (Figure 3-2). This alternatively spliced transcript was almost 10-fold more abundant in the *upf1Δ* mutant (Figures 3-3 and 3-5), further suggesting that it is targeted by NMD. We also detected a large accumulation of unspliced species in the *upf1Δ* mutant, indicating inefficient recognition of this splicing substrate. This may be the result of both the normal (GUACGU) and alternative (GUUUGU) 5′-SS having non-consensus sequences (see below). Interestingly, the abundance of the alternatively spliced product was found to decrease by two to three folds when Nam8p or Mud1p were inactivated in the context of the *upf1Δ* deletion (Figures 3-3 and 3-5). The deletion of either one of these two factors might hinder the ability of the U1 snRNP to bind to the alternative suboptimal 5′ GUUUGU splice site of *RPL22B*, resulting in decreased usage. This is consistent with the known roles of Mud1p and Nam8p in the first step of splicing (25), and suggest their direct involvement in modulating 5′-SS selection of *RPL22B*. By contrast, no major changes were observed in the *prp17Δ*, *prp18Δ*, *isy1Δ*, *hub1Δ* mutants, either alone or in combination with the *upf1Δ*deletion (Figure 3-3), showing the specificity of the effects detected with Nam8p and Mud1p. Thus, *RPL22B*exhibits two competing suboptimal 5′-SS, one of which is highly sensitive to perturbations in the U1 snRNP. The functional significance of the alternative 5′-SS of *RPL22B* in regulating transcript levels is investigated further below.

A novel role for Prp17p in promoting the use of branchpoint proximal alternative 3′-splice sites

Gapped sequence alignment showed that *TFC3* exhibits an alternative CAG 3′-SS 17 nt downstream of the annotated AAG (Figure 3-2). This product can be detected in samples from the wild-type and splicing mutants, but is 4.5-fold more abundant in the context of

the *upf1Δ* deletion, showing that a large fraction of this product is degraded by NMD (Figures 3-3 and 3-6). This non-productive isoform amounts to 27% of all spliced products (Figure 3-6), showing that a significant fraction of splicing generates NMD-targeted, non-productive transcripts. We observed a slight accumulation (1.7 fold) of the downstream alternative 3′-splice product in the *prp17Δ* mutant (Figures 3-3 and 3-6), showing that this second step splicing factor contributes to reducing the use of this alternative 3′-SS. As expected, inactivation of the first step splicing factors Mud1p or Nam8p had no effect on the pattern of 3′-SS selected (Figure 3-3).

TAN1 exhibits a more complex alternative 3′-SS pattern, where a canonical UAG 3′-SS is flanked by two alternative 3′ AAG sequences (Figure 3-2). The use of either of these sites would generate PTC-containing transcripts. The upstream AAG (AS 3′ #1) is only 6 nt away from the canonical 3′-SS. The retention of 6 nt of intronic sequence would maintain the proper reading frame but would result in a PTC because the UAG sequence of the normal 3′-splice site corresponds to an in-frame stop codon (32). The downstream AAG (AS 3′ #2) is 7 nt downstream of the normal 3′-SS, resulting in a frameshift-induced PTC. RT-PCR analysis of the wild-type and *upf1Δ* strains confirmed the RNA-Seq data by showing that these two alternative splice products are detected at extremely low levels, unless NMD is inhibited (Figures 3-3 and 3-7). In samples from the *upf1Δ* strain, the two alternatively spliced products accumulate to similar amounts, and both species are detected at lower levels than the normal spliced product (20% of all spliced products; Figure 3-7), possibly because these two suboptimal AAG sites do not compete well with the consensus canonical UAG site. Strikingly, the usage of these alternative 3′ splice sites was dramatically altered when Prp17p or Prp18p were inactivated. Inactivating Prp17p resulted in an increase in the use of the downstream alternative 3′-SS (AS 3′#2), while the upstream alternative 3′-SS (AS 3′#1) was no longer used (Figures 3-3 and 3-7), showing a

role of Prp17p in enhancing the use of upstream, branchpoint proximal 3′-SS. By contrast, Prp18p inactivation resulted in increased usage of the alternative 3′-SS most proximal to the branch point sequence (AS 3′ #1; Figures 3-3 and 3-7). This product is barely detectable in the wild-type strain but can be observed in the *prp18Δ* strain (Figure 3-3), and inactivation of Prp18p in the context of the *upf1Δ* deletion resulted in a 3-fold increase in the abundance of this species (Figure 3-7). The effect of Prp18p on this 3′-SS might be due to the identity of the sequences immediately following the 3′-SS, which have been shown to influence 3′-SS selection in the absence of a functional Prp18p (33). Isy1p inactivation resulted in an increase of unspliced species in a similar fashion to *SRC1*discussed above; however there was no effect of Isy1p, Hub1p, Mud1p and Nam8p on alternative 3′-SS selection of *TAN1* (Figure 3-3), showing the specificity of the effects observed with Prp17p and Prp18p. Finally, unspliced *TAN1* transcripts were generally not affected by NMD, except in the context of a*mud1Δ* mutant strain (Figure 3-3). This observation is consistent with a recent report showing that *TAN1*unspliced transcripts are retained in the nucleus by the RES complex, and are subject to NMD only when the RES complex is inactivated (34). Overall, analysis of *TFC3* and *TAN1* alternative 3′-SS patterns show that Prp17p and Prp18p have antagonistic roles in the selection of upstream and downstream 3′-SS *of TAN1*, and highlight the importance of Prp17p in enhancing the use of 3′-SS located closer to the branchpoint.

Alternative splicing patterns of *GPI15* and *GCR1* reveal the production of alternative non-functional protein products and the use of a non-canonical AUG 3′-splice site repressed by Prp18p

GPI15 in an interesting case where the two alternatively spliced products identified by our RNA-Seq analysis are not targeted by NMD. The use of an alternative GUACGU 5′-splice site results

in the deletion of 30 nucleotides from the 3′ end of exon 1 (Figure 3-2), which maintains the open-reading frame but generates a truncated protein. However, the protein product resulting from translation of this alternatively spliced product is likely to be non-functional, as this truncation removes a stretch of 10 amino acids at positions 187–197 in the most highly conserved region of this protein (35). This transcript can be detected in samples from the wild-type and the splicing factor mutants, and does not vary in intensity in the context of ***upf1Δ***, indicating that it is not targeted by NMD (AS 5′, Figure 3-3). In contrast, the alternatively spliced transcript generated by use of a downstream CAG 3′-SS results in a PTC. However, this PTC-containing transcript would exhibit a short 85 nt 3′-UTR, which might render it insensitive to NMD as suggested by the *faux 3′ UTR* model (36, 37). Indeed, the abundance of this transcript was not increased in the ***upf1Δ*** mutant (Figure 3-3). In addition, this transcript is expected to yield a non-functional protein due to C-terminal truncation and deletion of amino acids within the most conserved region of the protein (35). Analysis of the pattern of selection of these two alternatively spliced transcripts in the various splicing mutants did not reveal any major effect of these mutants (Figure 3-3) in contrast to the effects described above for *RPL22B, TAN1* or *TFC3*. However, there was a slight increase in the use of the downstream alternative 3′-SS in the ***prp17Δupf1Δ*** strain, consistent with the role of Prp17p in favoring the upstream 3′-SS, as described above for *TAN1* and *TFC3*.

GCR1 showed the most complex splicing pattern of all transcripts analyzed. Gapped alignments identified an intronic GUAUGG alternative 5′-SS as well as an upstream CAG alternative 3′-SS (Figure 3-2). In addition to these alternative splice sites identified by RNA-Seq, RT-PCR revealed the use of an additional GUAUGG alternative 5′-SS staggered 5-nt upstream of the normal 5′-SS and of a non-canonical AUG alternative 3′-SS 23 nt upstream of the other

alternative 3′-SS (Figure 3-2). The use of all of these sites was confirmed by RT-PCR, cloning and Sanger sequencing (Figure 3-9). The fact that some alternative splice sites escaped identification by mRNA sequencing indicates that a greater depth of coverage has the potential to identify even more alternative splice sites.

Based on *GCR1* annotation, the canonical spliced mRNA would use the GUAUGA 5′-SS along with the most downstream UAG 3′-SS (Figure 3-2). This product (labeled as $S^{(annot.)}$ in Figures 3-2 and and 3-3), however was detected at very low levels (Figure 3-3). The major spliced product observed resulted from the use of the most upstream GUAUGG 5′-SS and of an upstream CAG 3′-SS (labeled "S" in Figures 3-2 and 3-3). This splicing event does not introduce a PTC and results in a protein that is very similar to the translation product of the annotated spliced transcript $S^{(annot.)}$. The annotated amino acid sequence of *GCR1* from position 2 to 4 is VCT. In the major spliced product S, this sequence is replaced by QTSVDST. Thus, most of the protein is identical, except for a few N-terminal amino acids which are not expected to affect Gcr1p function, as all *GCR1* mutations with phenotypic effects have been mapped to a region downstream of this short sequence stretch (38-40). Based on the relative abundances of S and $S^{(annot.)}$, it is clear that S, and not $S^{(annot.)}$ is the main spliced product for the *GCR1* gene.

In addition to this major spliced product, we also detected a series of alternatively spliced products degraded by NMD (as denoted by asterisks in Figures 3-2 and 3-3). Splicing from the annotated GUAUGA 5′ splice site combined with the upstream CAG 3′ splice site resulted in a PTC-containing transcript labeled as *A in Figures 3-2 and 3-3. This transcript is degraded by NMD, as higher amounts are observed in all the strains containing a *upf1Δ* deletion, and it is the most abundant of all *GCR1* alternatively spliced products subject to NMD (Figures 3-3 and 3-8). Another product is generated from combining the upstream GUAUGG 5′-SS with the most

downstream UAG 3′-SS (*C in Figure 3-2). This splicing event results in a PTC, as it introduces a translational frameshift, which is not detected until the 43rd amino acid is translated. The corresponding transcript accumulates at low abundance in all samples and appears to be targeted by NMD, as its abundance increases slightly in all *upf1Δ* strains. In addition, the use of this most downstream 3′-SS increases almost 4-fold in the *prp17Δupf1Δ* strain when compared to the *upf1Δ*control (Figures 3-3 and 3-8). Because the 3′-SS used to generate this transcript corresponds to the most downstream one, this observation provides another example of the importance of Prp17p in favoring the selection of upstream 3′-SS, as shown above for *TFC3*, *TAN1* and to a lesser extent *GPI15*.

Another PTC-containing transcript that is degraded by NMD results from splicing of the downstream intronic GUAUGG 5′-SS with the CAG 3′-SS, (labeled *B in Figures 3-2 and 3-3). This product is faint, but detectable in all cases of NMD deactivation, except in combination with *nam8Δ* or *mud1Δ*, most likely because this 5′-SS has a higher sensitivity to U1 snRNP perturbations, as described above for *RPL22B*. Analysis of other mutants did not reveal any major influence on the pattern of 5′- or 3′-SS selection. Like *SRC1*, *GCR1* exhibits two staggered 5′ splice sites. However, unlike for *SRC1*, Hub1p has no influence on their selection (Figure 3-3).

A final set of NMD targets are produced by the use of the two most upstream 5′-SS with a highly unusual alternative AUG 3′-SS in the intronic sequence (labeled *D and *E in Figures 3-2 and 3-3). Interestingly, these products were only detected in the absence of Prp18p, suggesting that this factor is essential in preventing the use of this non-canonical 3′-SS. The use of this highly unusual AUG 3′ splice site was unambiguously confirmed through sequencing and RT-PCR analysis of RNAs derived from*prp18Δupf1Δ* samples (Figure 3-9). The ATPase Prp22p has been implicated in the fidelity of 3′-SS selection (41). Because Prp18p functions upstream from

Prp22p during the late stages of splicing (42), it is possible that the absence of Prp18p might indirectly hinder the function of Prp22p in proofreading 3′-SS selection, and that the use of this unusual 3′-SS might be the consequence of a reduced Prp22p function in the absence of Prp18p. To test this hypothesis, we analyzed *GCR1* splicing in a *prp22-1* mutant. RT-PCR analysis showed that the spliced product generated from the use of the AUG 3′-SS did not accumulate in a *prp22-1* splicing mutant (Figure 3-10). Thus, the accumulation of species resulting from the use of this unusual 3′-SS in the ***prp18Δupf1Δ*** samples is not an indirect consequence of hindered Prp22p function. The discovery of the splicing at this unusual 3′-SS sequence reveals the importance of Prp18p in ensuring proper 3′-SS selection for *GCR1* and in repressing the use of non-canonical 3′-SS sequences.

Alternatively spliced species of *RPL22B* and *GCR1* increase during stress conditions

The previous results validated our prediction that transcripts generated from the use of alternative non-productive splice sites are degraded by NMD and revealed the role of specific splicing factors in governing the choice between alternative sites. Strikingly, the sequence of some of these non-productive splice sites was found to be conserved across closely related yeast species. Because the level of sequence conservation in intronic sequences is usually very low, these peaks in sequence conservations for intronic alternative splice sites might reflect their functional importance. We hypothesized that the use of some of these alternative splice sites which lead to degradation by NMD might be favored under certain conditions to down-regulate gene expression. To test this hypothesis, we monitored changes in the splicing patterns of *RPL22B*, *TAN1*, and *TFC3* under stress conditions such as amino acid starvation, heat shock, LiCl-mediated hyperosmotic stress, and rapamycin treatment, as these have been reported to elicit diverse responses in the expression of intron containing genes (43, 44). In addition, various

stresses cause down-regulation in ribosomal protein gene expression (many of which contain introns), presumably to relieve the cell of massive energy requirements of ribosome biogenesis and focus those resources into regulations that are the most appropriate in response to the current stress condition (45-47). After 10 minutes of amino acid depletion, *RPL22B* showed an increase in unspliced species as well as well as a 4.5-fold increase in the level of the alternatively spliced product when compared to the SDC or YPD media controls (Figure 3-4A). In the ***upf1Δ*** strain shifted to amino acid starvation conditions, the levels of the alternatively spliced product increased compared to the wild-type strain grown in the same conditions, as would be expected when NMD transcripts are no longer degraded (Figure 3-11A lanes 2 and 4). The fact that the level of the alternatively spliced transcript is 2.5-fold higher in the ***upf1Δ*** sample than in the wild-type sample under amino acid starvation conditions argues that the increase in the abundance of these species in the wild-type strain in these conditions is not due to NMD inhibition in these conditions, but that a change in splice site selection occurs that favors the use of the alternative splice site. Significantly, amino acid starvation did not change the levels of the alternatively spliced species of *TAN1* and *TFC3* that are normally subject to NMD. This observation provides further evidence that the increase in the amount of alternatively spliced *RPL22B* transcript observed during amino acid starvation is due to a switch in splice site selection and not to an inhibition in NMD, since the level of alternatively spliced species of *TAN1* and *TFC3* that are normally degraded by NMD is unaffected in the same conditions.

We next investigated the effect of a 20 minute heat shock at 42°C on splicing patterns. Under these conditions and in the wild-type strain, *RPL22B* showed an increase in unspliced as well as a decrease in the relative amount of the normal spliced product (Figure 3-11A lane 5 vs. 7). More importantly, the NMD defective strain ***upf1Δ*** showed an even larger increase in unspliced pre-

mRNAs, as well a large accumulation of the alternatively spliced product that coincides with a decreased amount of canonical spliced product (Figure 3-11A lane 6 vs. 8). In these conditions, the alternatively spliced product now corresponds to more than half of all spliced species. Under heat shock, this alternatively spliced product is 4-fold more abundant in the *upf1Δ* strain than in the wild-type strain. These higher levels upon NMD inactivation show that the increased accumulation of these species under heat shock is not due to a decrease in NMD efficiency. Rather, this result shows that the use of the alternative splice site is being favored in heat shock conditions. By contrast, *TFC3* and *TAN1* exhibited an accumulation of unspliced species, but decreased levels of both the canonical and alternatively spliced species consistent with a general inhibition of pre-mRNA splicing under heat shock (48, 49). Thus the pattern of alternatively spliced species of *TFC3* and *TAN1* that are subject to NMD is very different from that of*RPL22B*, further proving that the accumulation of the alternatively spliced *RPL22B* transcript under heat shock conditions described above is not due to a general stabilization of spliced forms degraded by NMD.

Like heat shock, rapamycin treatment was shown to result in an inhibition of ribosomal proteins mRNA splicing based on microarray experiments (43). Within 20 minutes of rapamycin treatment, *RPL22B*indeed showed trends similar to those observed in heat shock, but to a lesser degree, with an increase of unspliced species and of alternatively spliced *RPL22B* species (Figure 3-11A), but no effect on the alternatively spliced *TAN1* and *TFC3* transcripts. Hyperosmotic shock (300 mM LiCl exposure for 10 min) only resulted in minimal effects; there were no changes observed for *TFC3* and *TAN1* targets under these stress conditions, and *RPL22B* showed only a slight increase in unspliced but the levels of spliced transcripts remained similar. Thus, *RPL22B* exhibits regulated use of its alternative 5′-splice site, mostly

under amino acid starvation and heat shock conditions, while other transcripts such as*TFC3* and *TAN1* did not exhibit any change in their alternative splicing profiles.

Because *GCR1* exhibited a very complex splicing pattern, especially in the absence of Prp18p, and because heat shock conditions resulted in the most dramatic changes in splicing for *RPL22B*, we next investigated the effect of heat shock on *GCR1* splicing in the wild-type, ***upf1Δ***, ***prp18Δ*** and ***prp18Δupf1Δ***mutants (Figure 3-11B). Under heat-shock, we detected a general inhibition of splicing, consistent with the data described above. However, we also observed an increase of the abundance of the A* form relative to the normal spliced product S, indicative of a switch from the normal GUAUGG site to the GUAUGA site. The absence of Prp18p resulted in an increase of the use of the non-canonical AUG site (*D species) in heat shock conditions, and this product now constituted one third of all spliced species. Thus we conclude that *GCR1*, like *RPL22B*, exhibits a switch in splice site selection during heat shock, and that Prp18p limits splicing at this non-canonical AUG site under stress conditions.

The alternative, suboptimal 5′-splice site of *RPL22B* contributes to the global down-regulation of*RPL22B* in stress conditions

To further analyze the importance of the alternative 5′-splice site of *RPL22B* on its splicing patterns and expression during normal and stress conditions, we investigated the effect of mutations of this alternative 5′-SS. The suboptimal GUUUGU alternative 5′-splice site was either deleted or mutated to the consensus GUAUGU sequence at the endogenous chromosomal locus (CS, consensus mutation and**Δ**, deletion, Figure 3-12A). Changing the alternative 5′-SS to the consensus GUAUGU sequence resulted in detectable amounts of alternatively spliced products at 25°C, even in a functional NMD background (Figure 3-12B), suggesting that the suboptimal GUUUGU sequence contributes to the low usage of this alternative site in normal

conditions. Inactivation of Upf1p in this context showed that 70% of all spliced species were now being produced by splicing from the alternative consensus site (Figure 3-12B, lane 5), and that splicing efficiency was improved, as shown by a decrease in unspliced species. By contrast, deleting the alternative splice site resulted in higher amount of unspliced transcripts, especially in the *upf1Δ* background. Thus, deleting the alternative 5′-splice site of *RPL22B* is not sufficient to improve splicing at the normal splice site, possibly because of the suboptimal sequence of the normal *RPL22B* 5′-splice site. In addition to RT-PCR, the same strains were analyzed by northern blot (Figure 3-12B, bottom panel), which yielded results similar to those obtained by RT-PCR. These results show that increasing the strength of the alternative 5′-SS of *RPL22B* is sufficient to enhance the overall splicing efficiency of this transcript, while deleting this site results in an overall increase of unspliced RNAs. Under heat shock and NMD inactivation, this effect was even more prominent, as mutation of the alternative splice site to the consensus resulted in the alternatively spliced product being the major spliced species (Figure 3-12B, lane 11). Thus, under heat shock conditions, *RPL22B* transcripts bearing the consensus alternative splice site mutation are now spliced almost exclusively at this site. Analysis of the mutant with a deletion of the alternative 5′-SS under heat shock conditions showed that the use of the normal 5′-SS is not increased at elevated temperatures when the competing alternative 5′-SS has been eliminated (Figures 3-12A and 3-12B lane 12). This mutant shows a larger accumulation of unspliced *RPL22B*transcript, hinting that the normal process of spliceosome assembly is perturbed on this transcript during heat shock, possibly due to the suboptimal 5′-SS. To obtain a more quantitative assessment of transcript levels, rather than just assessing the ratio between the different spliced forms, we analyzed the same samples by northern blot. This analysis showed that cells treated in heat shock conditions resulted in much weaker signal than in the samples obtained from cells grown at 25°C, consistent with a general down-regulation of ribosomal

protein genes under stress (43-47). Upon NMD inactivation, we observed a rescue of transcript levels, which mostly corresponded to unspliced RNAs and to some alternatively spliced transcripts (Figure 3-12B). However, changing the alternative 5′-SS to a consensus sequence in the context of NMD inactivation was sufficient to recover a large amount of spliced transcripts (Figure 3-12B, lane 11, lower panel). To investigate if this effect was specific to heat shock or is also observed during other stresses, we analyzed the expression of wild-type and mutated forms of*RPL22B* during amino acid starvation (Figure 3-12C). The results observed during amino acid starvation were similar to those described during heat shock, with a large increase in the level of spliced transcripts upon changing the alternative 5′-splice site to the consensus sequence. We also observed an increase in the use of the alternative 5′-SS under amino acid starvation (Figure 3-13). Interestingly, shifting the Upf1p-inactivated strain with the alternative 5′-splice site consensus sequence from SDC to amino acid starvation conditions resulted in only a minor increase of the use of the alternative 5′-SS, possibly because the level of transcripts spliced at that site is already very high in the *upf1***Δ** strain in normal conditions (75%; Figure 3-13). In conclusion, these results show that low splicing efficiency due to the suboptimal normal and alternative 5′ splice sites of *RPL22B*, combined with NMD degradation of the unspliced and alternatively spliced forms contribute to the general decrease in *RPL22B* levels as a means to rapidly halt production of this ribosomal protein under various stress conditions.

Usage of the alternative 5′-splice site of *RPL22B* is influenced by promoter identity

Ribosomal protein genes are known to be transcriptionally regulated in stress conditions. To investigate the use of *RPL22B* 5′-SS selection independently from transcriptional inhibition under heat shock, we replaced the natural *RPL22B* promoter with a galactose-inducible promoter. The wild-type and *upf1Δ*strains containing the natural *RPL22B* promoter showed no

detectable difference in *RPL22B* splicing patterns or expression when grown in galactose containing medium (YPGal) compared to glucose-containing medium (YPD) at 25°C (Figure 3-14 lanes 1–4), either by RT-PCR (top panel) or northern blot (bottom panel). Strikingly, replacement of the normal *RPL22B* promoter by the *GAL* promoter resulted in an increase in overall *RPL22B* transcript levels, but also in a decrease in the use of the alternative 5′-SS (Figure 3-14). The fact that the usage of the alternative splice site of *RPL22B* is reduced in this strain while transcript levels are higher overall argues against the hypothesis that alternative splice site usage is the result of splicing errors occurring at low frequencies, as if this were the case, one would expect higher levels of alternatively spliced *RPL22B* transcripts upon its overexpression in the strain in which the natural *RPL22B* promoter was swapped for the *GAL* promoter. Under heat shock conditions, the use of the alternative splice site was reduced 8.1 fold in the *upf1Δ* strain expressing *RPL22B* under the control of the *GAL* promoter compared to the *upf1Δ* strain expressing *RPL22B* from its natural promoter and grown in galactose medium (Figure 3-14, lanes 10 and 12). Thus, alternative splicing regulation of *RPL22B*upon heat shock is tightly linked to the identity of the *RPL22B* promoter, as switching the identity of the promoter is sufficient to favor the use of the normal 5′-splice site. The mechanism by which the identity of the promoter influences alternative splice site selection is unclear, but could be linked to the influence of the promoter on the speed of transcription. Nevertheless, we can conclude from these results that transcriptional down-regulation and the increased use of the alternative 5′-SS provide synergistic mechanisms to limit the expression of *RPL22B* during stress, consistent with the global down-regulation of ribosome biogenesis during stress conditions.

Discussion

A significant fraction of splicing events in *S. cerevisiae* generates non-functional RNA or protein products

In this study we show that the ensemble of transcripts generated by splicing from the *S. cerevisiae*genome is highly complex. Most of the splicing events that we have characterized in this study are non-productive, either because they result in transcripts that are targeted by NMD, or because the protein products generated from these transcripts are predicted to be non-functional (e.g. *GPI15*). The large number of additional splice sites identified, and their relaxed conservation (Figure 3-1D) imply that the rules governing splice selection are intrinsically more flexible than previously thought. This is further illustrated by the finding that a non-canonical AUG sequence in *GCR1* can be used as a 3′-SS in the absence of Prp18p (Figure 3-3). In some cases, non-productive alternatively spliced transcripts accumulate only at low levels (e.g. *GCR1*, *GPI15*, Figure 3-3). However, for other genes such as *TFC3*, the alternatively spliced non-productive transcripts represent a significant fraction (close to 30%) of all RNAs generated from this locus. Thus, non-productive splicing can significantly limit the expression of these genes. This was further demonstrated by mutagenesis of the non-productive splice site of *RPL22B*, as changing this site to a consensus sequence was sufficient to increase the splicing efficiency and the expression of this gene (Figure 3-12B). Thus, the presence of alternative and sometimes sub-optimal splice sites that compete with the normal splice site contributes to an overall decrease in the amount of productively-spliced transcripts. Because the overlap in the alternative splicing events detected in all three NMD-deficient strains was limited (Figure 3-1C), and because we detected by RT-PCR some alternative splicing events that escaped detection by RNA-Seq (e.g. *GCR1*), it is likely that we have not exhaustively identified the ensemble of splice

sites that can be used by *S. cerevisiae*, and that additional splice sites will be identified by deeper sequencing or systematic RT-PCR analysis.

Contribution of splicing factors to alternative splice site selection and splice site fidelity

The analysis of double mutants in which splicing factor mutations were combined with NMD inactivation revealed some important and unexpected functions for these factors on alternative splice site selection. We found that the Nam8p and Mud1p components are important for the selection of some, but not all of the alternative 5′-splice sites described here. In the case of *RPL22B*, this requirement was likely due to the fact that the alternative 5′-SS possesses a suboptimal splicing sequence, and therefore exhibits a weaker affinity for U1 binding, and a stronger requirement for Mud1p and Nam8p that impact the efficiency of U1 snRNP assembly on the alternative splice site. Strikingly, we identified a new role for Prp17p in favoring the use of upstream, branchpoint-proximal 3′-SS. In all cases that we have analyzed, Prp17p inactivation resulted in an increase in the use of the downstream 3′-SS. The mechanistic basis for this novel function that we describe here for Prp17p in promoting branchpoint proximal 3′-SS is not fully understood. Because 3′-SS close to the branchpoint are often the first ones that are being used, this novel function for Prp17p could be linked to promoting the ability of the spliceosome to scan and recognize 3′-SS close to the branchpoint, or to unwind secondary structures that mask branchpoint-proximal 3′-SS. The absence of Prp17p would result in a higher rate of misrecognition of 3′-SS and in the use of more distal 3′-SS. In addition, we found that the absence of Prp18p resulted in the selection of a non-canonical AUG 3′-SS in *GCR1*, and that this atypical 3′-SS was utilized to a greater extent during heat-shock, revealing a unique function for Prp18p in suppressing usage of a non-canonical 3′-SS. This function for Prp18p is independent from Prp22p's function in proofreading 3′-SS (41), but might complement its role to ensure the

overall proper fidelity of 3′-SS selection. While we have demonstrated this function for *GCR1* only, a full genomic analysis of 3′-SS usage in the absence of Prp18p might reveal further examples of non-canonical 3′-SS being used.

Spliceosome errors or bona-fide regulations?

The widespread occurrence of non-productive splice site usage described in this study begs the question of whether the use of these splice sites is the result of mistakes by the spliceosome, which occur at low frequency (as one might suggest based on their weaker consensus sequences) or whether they correspond to sites that have been selected throughout evolution for regulatory purposes. The sequence of some of these intronic, non-productive splice sites is conserved across closely related yeast species, which, given the low conservation of intronic sequences in general, argues that this might reflect some degree of functional relevance. In addition, there is no obvious correlation between transcript levels and the occurrence of alternative splicing events (Figure 3-1D), which argues against the suggestion that most of the alternative splicing events that we have mapped arise from low fidelity splicing events or errors that occur randomly, and which would be expected to be more frequently detected in highly abundant transcripts. Also, replacement of the *RPL22B* gene promoter results in higher transcript levels but reduces the usage of the alternative 5′-splice site of *RPL22B* (Figure 3-14), providing another independent argument to suggest that the level of usage of alternative splice sites is not solely a reflection of overall transcript abundance. Finally, we show that the use of some of these alternative splice sites can be up-regulated during stress conditions (*RPL22B*, *GCR1*), and that this increased use participates in the down-regulation of *RPL22B* in stress conditions. Thus, this phylogenetically conserved, alternative, non-productive 5′-SS of *RPL22B* is functionally important because it contributes to the down-regulation of *RPL22B* during stress. This is shown by the fact that

changing this sequence to a consensus sequence results in a significant increase in transcript levels upon NMD inactivation during stress (Figure 3-12). The transcriptional down-regulation of ribosomal proteins during stress has been documented previously (45). We show here that the promoter of the *RPL22B*gene is essential not only because it drives transcriptional repression during stress, but also because it controls the switch in 5′-SS selection that contributes to the overall repression of *RPL22B* during heat-shock. Thus, a combination of transcriptional and post-transcriptional regulations, through splicing inhibition (43, 44), degradation of unspliced RNAs by NMD (12, 47) and use of non-productive splice sites (this study) contributes to the repression of ribosomal protein production during stress. While several non-RPG transcripts analyzed in these stress conditions did not shown any changes, *GCR1* did exhibit a change in the use of alternative splice sites during stress (Figure 3-11B). This result raises the possibility that other intron-containing genes may be regulated similarly by alternative splicing as a function of different environmental growth conditions. Overall our study has revealed that the pattern of splicing events in the model eukaryote *S. cerevisiae* is highly complex, but masked by NMD-mediated degradation. Given the recent report that another single cell eukaryote, *S.pombe* shows alternative splicing patterns conserved in higher eukaryotes (50), these observations suggest that alternative splicing provides an important contribution to genetic regulations and adaptations to environmental changes in unicellular eukaryotes. However, in the case of *S.cerevisiae*, the use of alternative splice sites have evolved towards fine tuning transcript levels, rather than generating proteome diversity as shown in higher eukaryotes.

Methods

Yeast culture and RNA analysis

Yeast strains were grown at 25°C in YPD medium, unless indicated otherwise in the figures. For heat shock treatment, strains were pre-grown in YPD at 25°C, spun down in 50 mL Falcon tubes, resuspended in pre-warmed YPD medium and heat shocked for 20 min before harvesting. For LiCl treatment, yeast strains were grown to mid-log phase in YPD rich media at 30°C, harvested by centrifugation in 50 mL Falcon tubes, washed once with pre-warmed 50 mL of YPD+300 mM LiCl before being resuspended in pre-warmed YPD with 300 mM LiCl for 10 minutes. For Rapamycin treatment, cells were grown to mid-log phase in rich media (at 30°C), and rapamycin from a stock solution of 1 mg/mL in 90% ethanol, 10% Tween-20 was added to a final concentration of 200 ng/mL and cells were incubated for 20 minutes. The same volume of 90% ethanol, 10% Tween-20 solution used for the rapamycin treatment was added to the negative control. Sample preparation and RNA sequencing was performed by Illumina. RT-PCR analysis and northern blot was performed as described (21).

Mapping reads

High throughput sequencing data have been deposited in the GEO database (accession GSE55213). All sequence files were aligned against the 2008 SGD assembly of the *Saccharomyces cerevisiae* genome. The novoalign software package (www.novocraft.com) and the BLAT alignment tool (22) were used to align 75 base pair reads in two steps. In the first step, sequences were aligned with novoalign allowing for up to four mismatches and no gaps. In the second step, sequences that failed to align in the first step were aligned with BLAT allowing three mismatches and gaps up to 20000 nucleotides in length. A sequence was kept for further

analysis if it mapped with equal score to at most two genomic locations and did not contain a gap smaller than ten nucleotides.

Intronic sequences counts

Intronic sequence expression representative of unspliced RNAs was quantified for each ICG by summing reads that aligned to introns and exon-intron boundaries. Values between samples were normalized by total mapped reads to account for lane effects. p-values were computed by modeling each ICG wild-type count as a poisson random variable and calculating the probability of observing each mutant count if it were drawn from the same distribution.

Quantification of alternative splicing events

Alternative splicing events were defined as splicing events that are within ICGs and are supported by sequencing but that are not annotated in the *Saccharomyces* Genome Database (SGD). Counts of total alternative splicing events and PTC-generating alternative splicing events were quantified by summing all unique alternative splicing events in each sample. To determine if an alternative splicing event is PTC-generating we constructed the splice product's sequence using the novel splicing event in the otherwise canonical transcript sequence. Counts between samples were normalized by sequencing depth. p-values were calculated by modeling the wild-type count as a poisson random variable and calculating the probability of observing each mutant's count for both total alternative splicing events and PTC-generating alternative splicing events. Venn diagrams of agreement between samples were generated using BioVenn (51).

Splice site consensus sequence

Consensus sequences for 5′ and 3′ ends of both canonical splice sites and alternative splice sites were represented as sequence logos. Sequence logos were constructed using the MATLAB (MathWorks) seqlogo function.

Figures

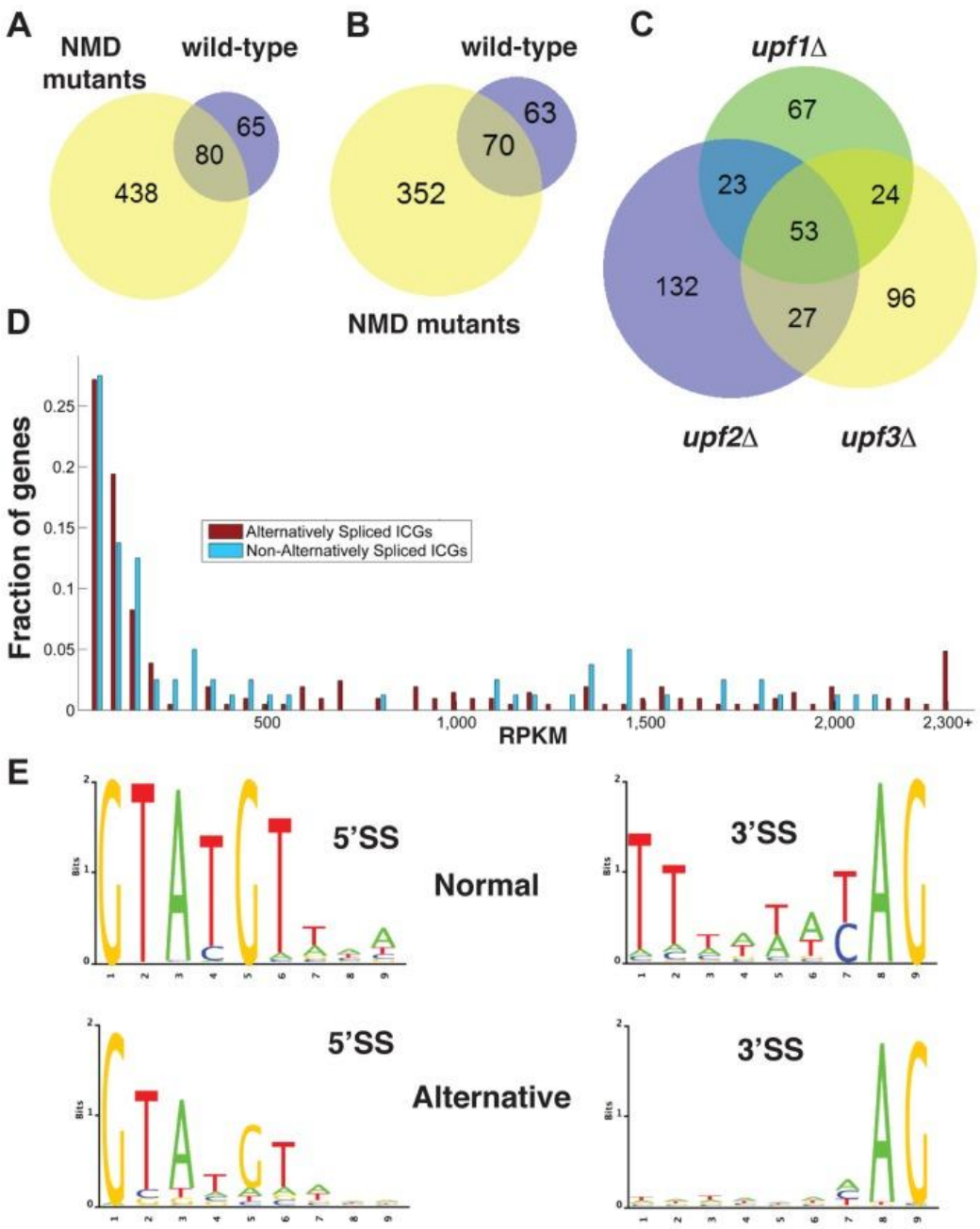

Figure 3-1. **Bioinformatics analysis of alternative splice site usage in wild-type and NMD mutants.** A. Venn diagram showing the overlap of alternative splice site usage between the wild-type and three NMD mutants pooled for all unique non-canonical splicing events (both PTC-generating and non-PTC-generating). B. Venn diagram showing the overlap of alternative splicing events between the wild-type and three NMD mutants pooled for all unique non-canonical splicing events resulting in a potential PTC. C. Venn diagram showing the overlap of alternative splicing events between the *upf1Δ*, *upf2Δ*, and *upf3Δ* strains for PTC-generating splicing events. D. Distributions of intron-containing gene transcripts showing alternative splicing events (red) or no alternative splicing events (blue) according to their overall abundance in RPKM. Transcripts for which the abundance was higher than 2,300 RPKM were grouped in the final bin. E. Sequence logo analysis of 5′- and 3′- splice sites for all normal and alternative splicing events detected by RNA-Seq in wild-type and NMD mutant strains.

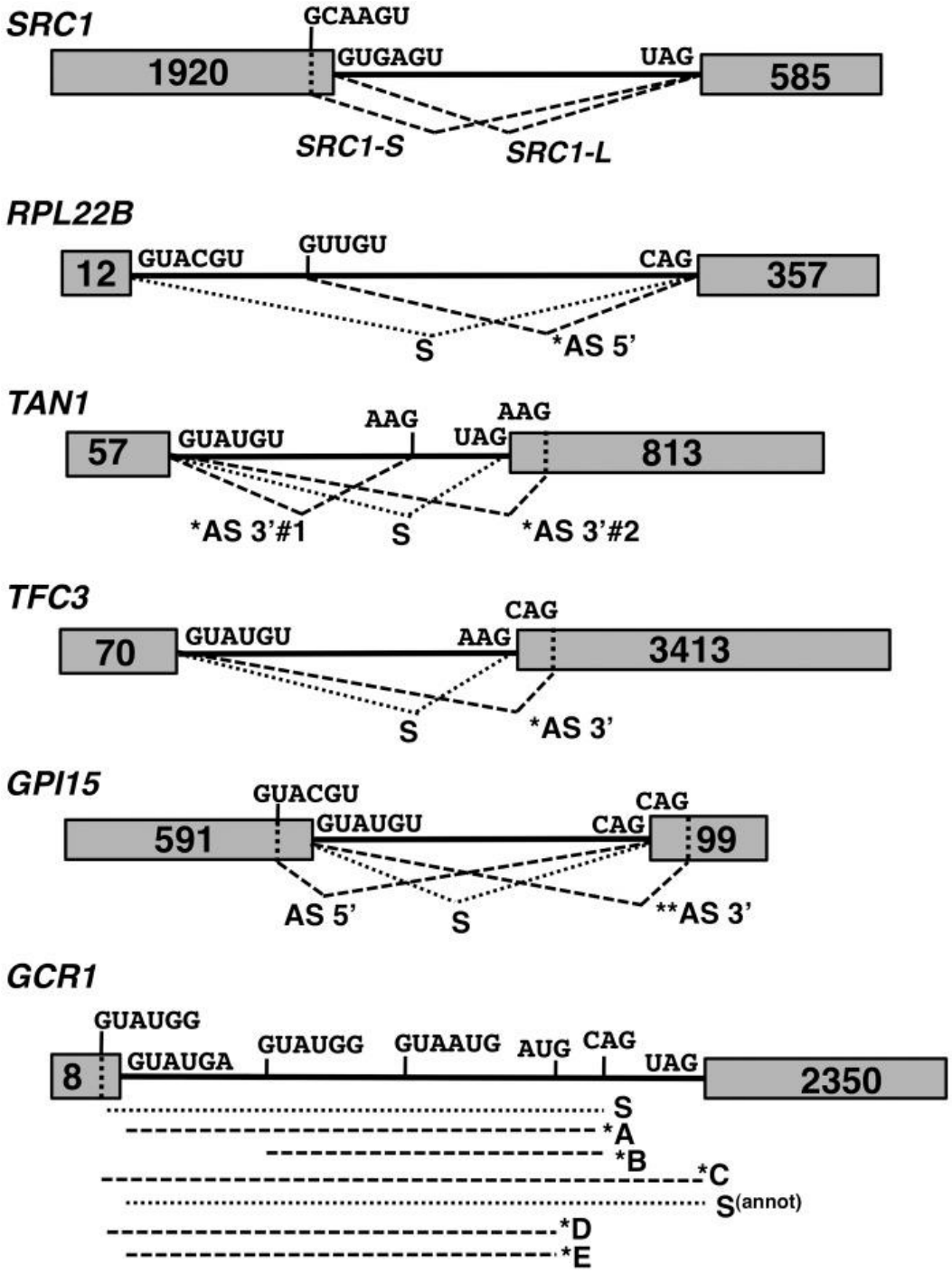

Figure 3-2. **Spliced species produced from the *SRC1, RPL22B, TAN1, TFC3, GPI15* and *GCR1* genes.** Species labeled with an asterisk are subject to NMD. Species labeled with two asterisks are predicted to be subject to NMD but were not observed to do so in subsequent experiments. The alternative 3′-SS of *SRC1* is located 4 nt upstream from the annotated 3′-SS. The alternative 3′-SS of *RPL22B* is located 64 nt downstream from the annotated 3′-SS. The alternative 3′-SS of *TAN1* are located 6 nt upstream and 7 nt downstream from the annotated 3′-SS. The alternative 3′-SS of *TFC3* is located 17 nt downstream from the annotated 3′-SS. The alternative 5′ and 3′-SS of *GPI15* are located 36 nt downstream and 14 nt upstream, respectively, from the annotated 5′ and 3′-SS. The alternative 3′-SS of *GCR1* are located 5 nt upstream (GUAUGG); 51 nt downstream (GUAUGG) and 627 nt downstream from the annotated 5′SS. The alternative 3′-SS of *GCR1* are located 40 nt upstream (AUG) and 17 nt downstream (CAG) from the annotated 3′-SS.

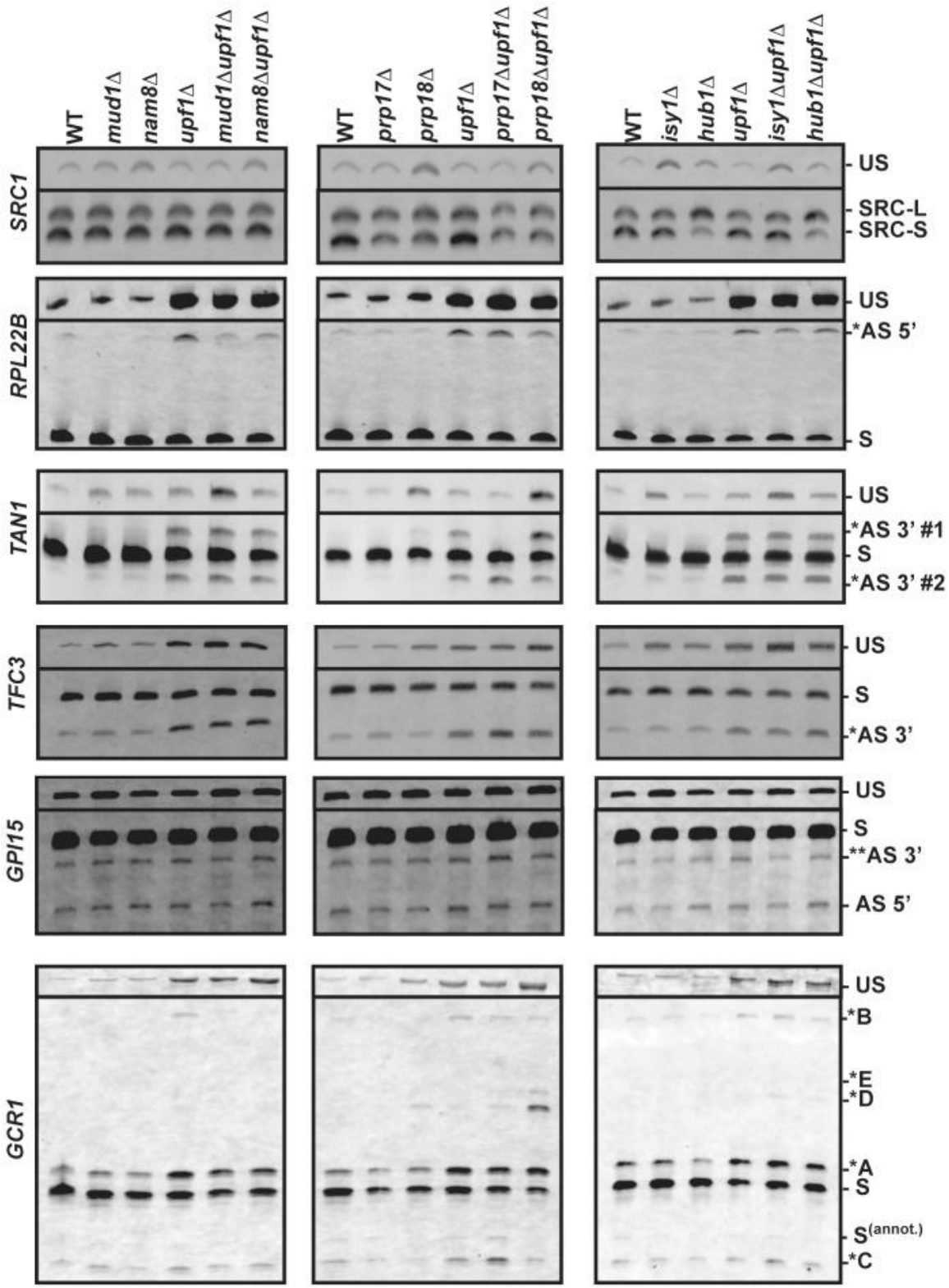

Figure 3-3. **RT-PCR analysis of alternatively spliced products for *SRC1*, *RPL22B*, *TAN1*, *TFC3*, *GPI15* and *GCR1* in wild-type, NMD and various splicing mutants.** The unspliced (US) species is also shown on top. The middle portions of the gel where no species were visible have been removed. In all cases, RT-PCR was performed with a Cy3-labeled primer. The labeling of the different alternatively spliced forms is according to the nomenclature shown in Figure 2.

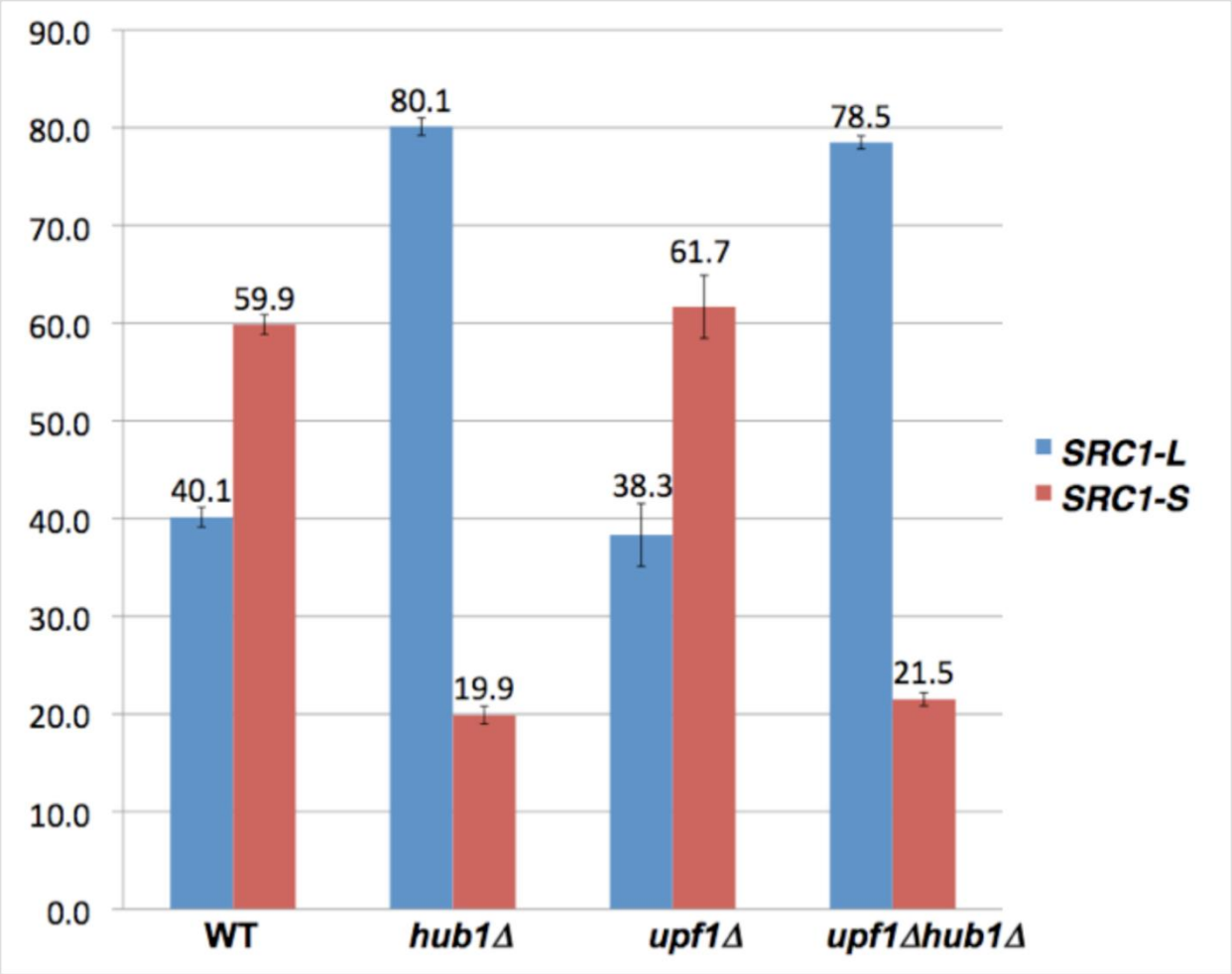

Figure 3-4. Quantification of the SRC1-L and SRC1-S isoforms in wild-type, *upf1Δ* and splicing mutants. Shown is the percentage of the *SRC1-L* and *SRC1-S* transcripts in various strains. Values shown are the average and standard deviations obtained from RT-PCR experiments of three independent cultures for each strain.

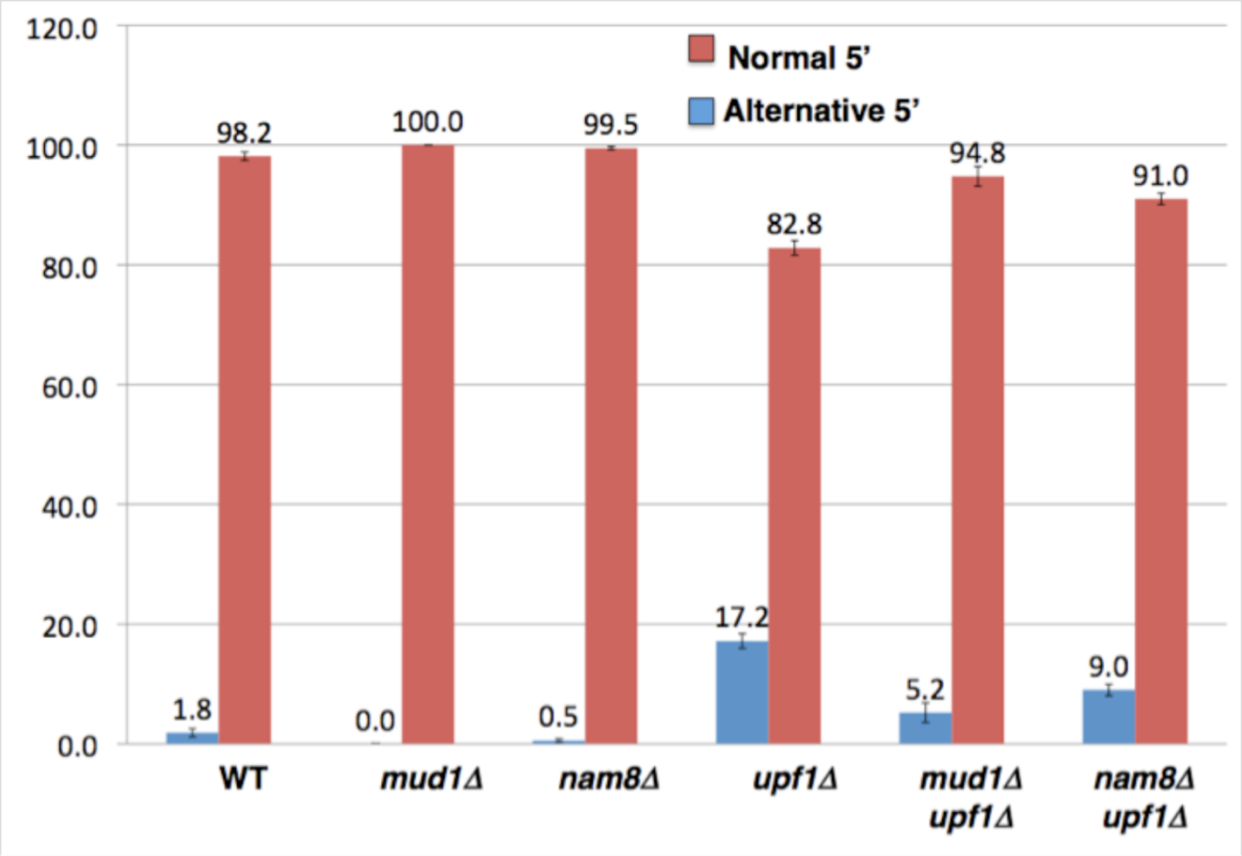

Figure 3-5. Quantification of the usage of the normal and alternative 5′-splice sites of *RPL22B* in wild-type,*upf1Δ* and splicing mutants. Shown is the percentage of transcripts spliced at the normal 5′-splice site (red) and at the alternative 5′-splice site (blue). Values shown are the average and standard deviations obtained from RT-PCR experiments of three independent cultures for each strain.

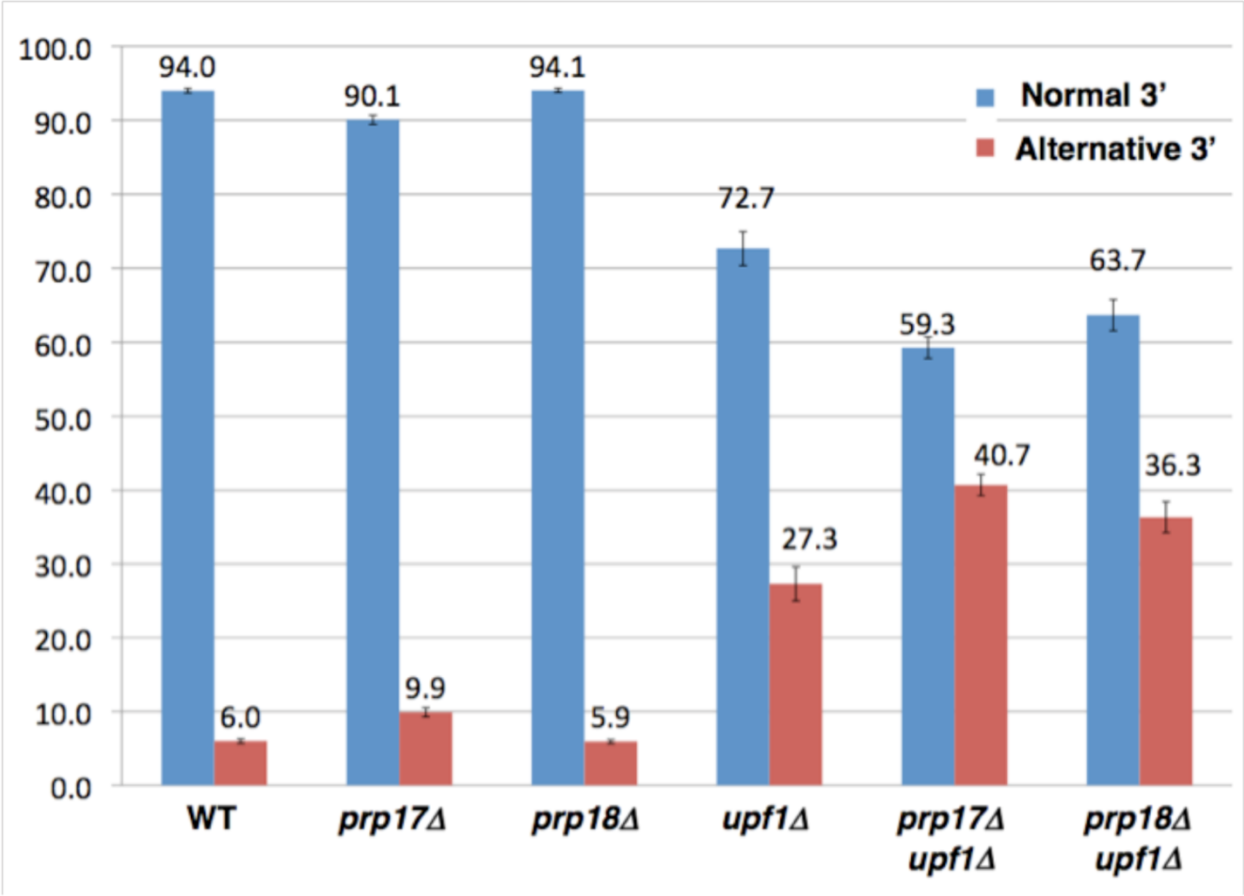

Figure 3-6. Quantification of the usage of the normal and alternative 3′-splice sites of *TFC3* in wild-type, *upf1Δ*and splicing mutants. Shown is the percentage of transcripts spliced at the normal 3′-splice site (blue) and at the alternative 5′-splice site (red). Values shown are the average and standard deviations obtained from RT-PCR experiments of three independent cultures for each strain.

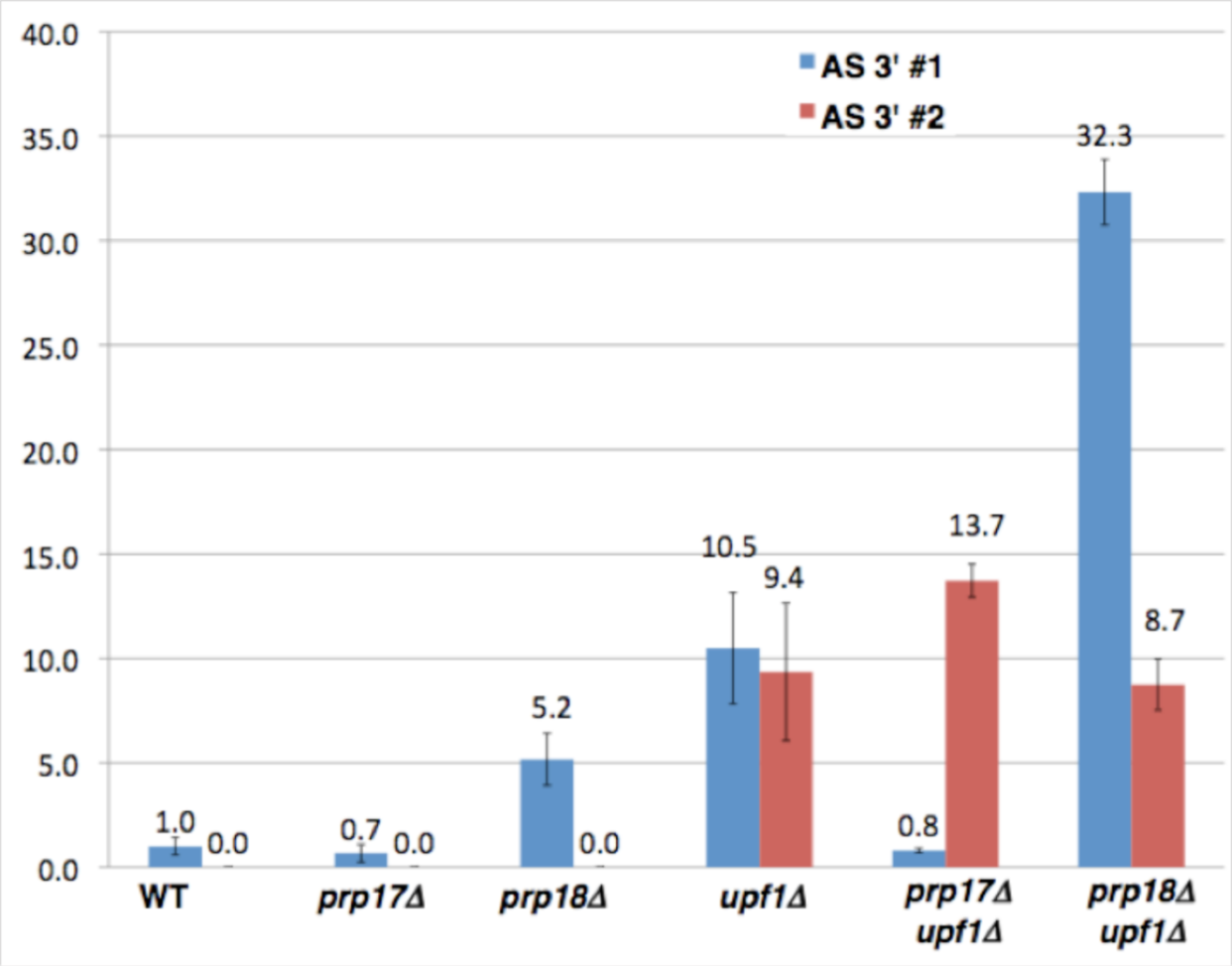

Figure 3-7. Quantification of the usage of the two alternative 3′-splice sites of *TAN1* in wild-type, *upf1Δ* and splicing mutants. Shown is the percentage of transcripts spliced at the alternative 3′-splice site #1 (blue) or #2 (red) compared to all the spliced transcripts. Values shown are the average and standard deviations obtained from RT-PCR experiments of three independent cultures for each strain.

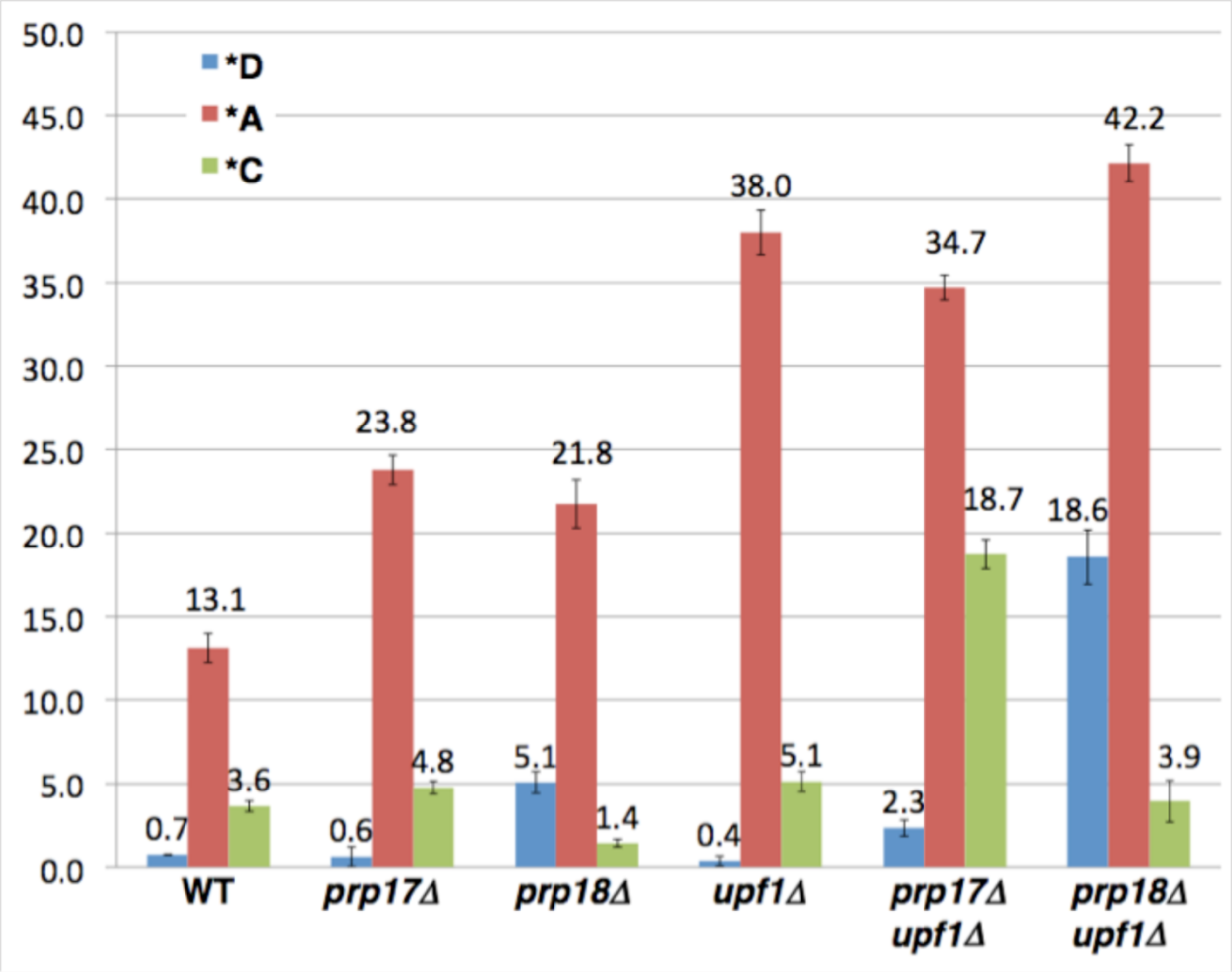

Figure 3-8. Quantification of the abundance of the major alternatively spliced forms of *GCR1* in wild-type,*upf1Δ* and splicing mutants. Shown is the percentage of the *D (blue), *A (red) or *C (green) spliced forms. Values shown are the average and standard deviations obtained from RT-PCR experiments of three independent cultures for each strain.

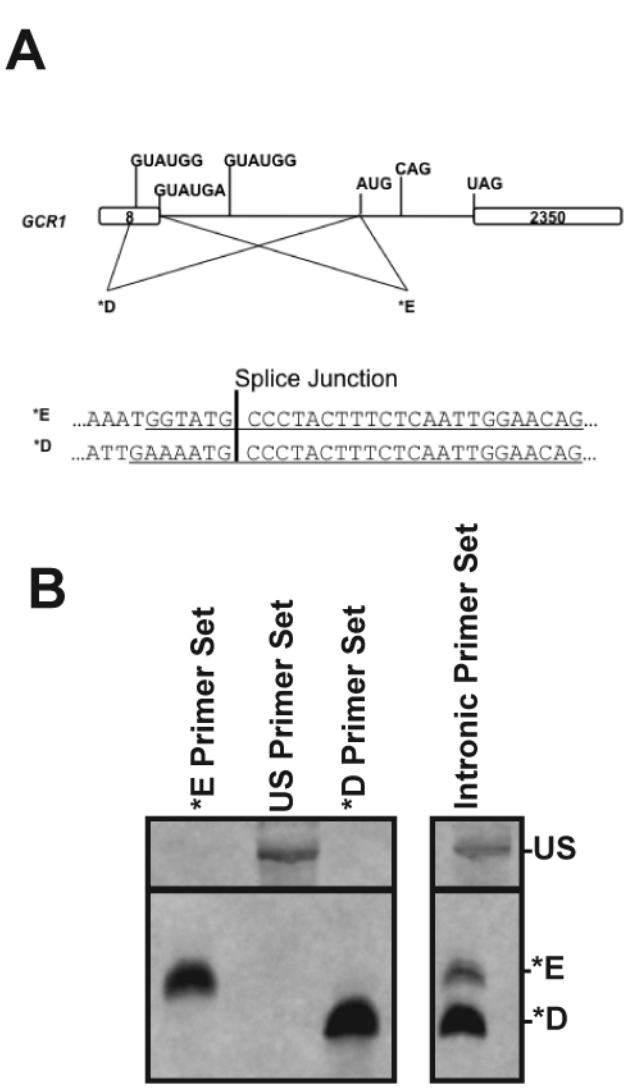

Figure 3-9. Validation of the use of the AUG alternative 3′ splice site of *GCR1* by RT-PCR. Sequencing of the cloned *D and *E cDNAs determined the location of the splice junction, while sequencing of unspliced cDNAs was used to confirm that this unusual alternative 3′-SS was indeed AUG, and not a SNP or other mutation of the *GCR1* gene that would have converted it into an AAG. RT-PCR confirmation of the use of this AUG 3′-SS was performed using reverse primers spanning the splice junction to specifically amplify distinct splicing events; either associated with *D, *E, or unspliced. The use of the AUG 3′ SS was also confirmed using an intronic reverse primer just downstream of the AUG sequence and detected *D, *E, and unspliced products, as predicted (Figure 4). A. RT-PCR strategy. All PCR include the same forward primer For, and various reverse primers that hybridize to the indicated regions of *GCR1*. B. RT-PCR data. Shown are the PCR products obtained from the different reverse primers shown in A.

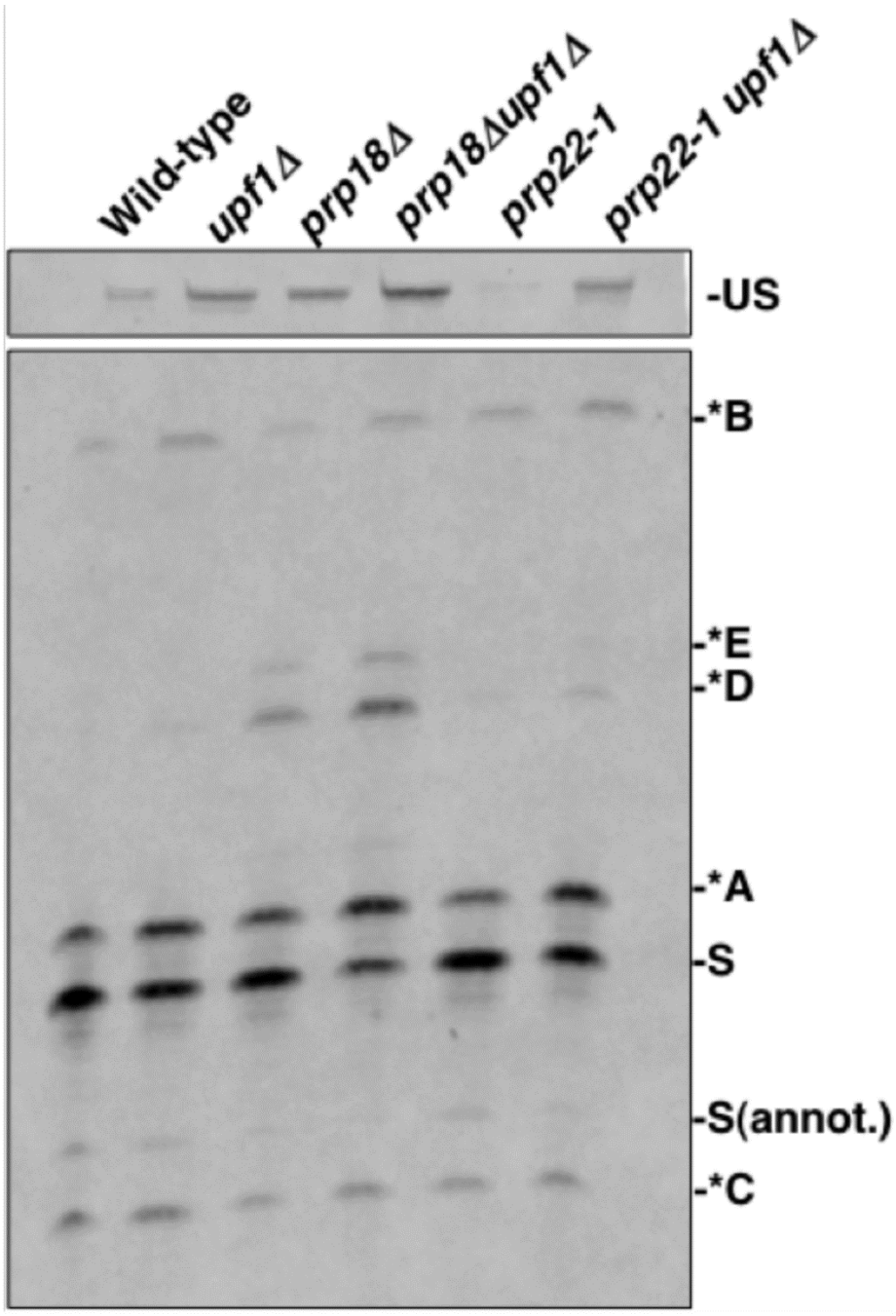

Figure 3-10. RT-PCR analysis of GCR1 splicing in the *prp18* and *prp22-1* mutant strains. The identity of the different spliced products is labeled according to Figure 2.

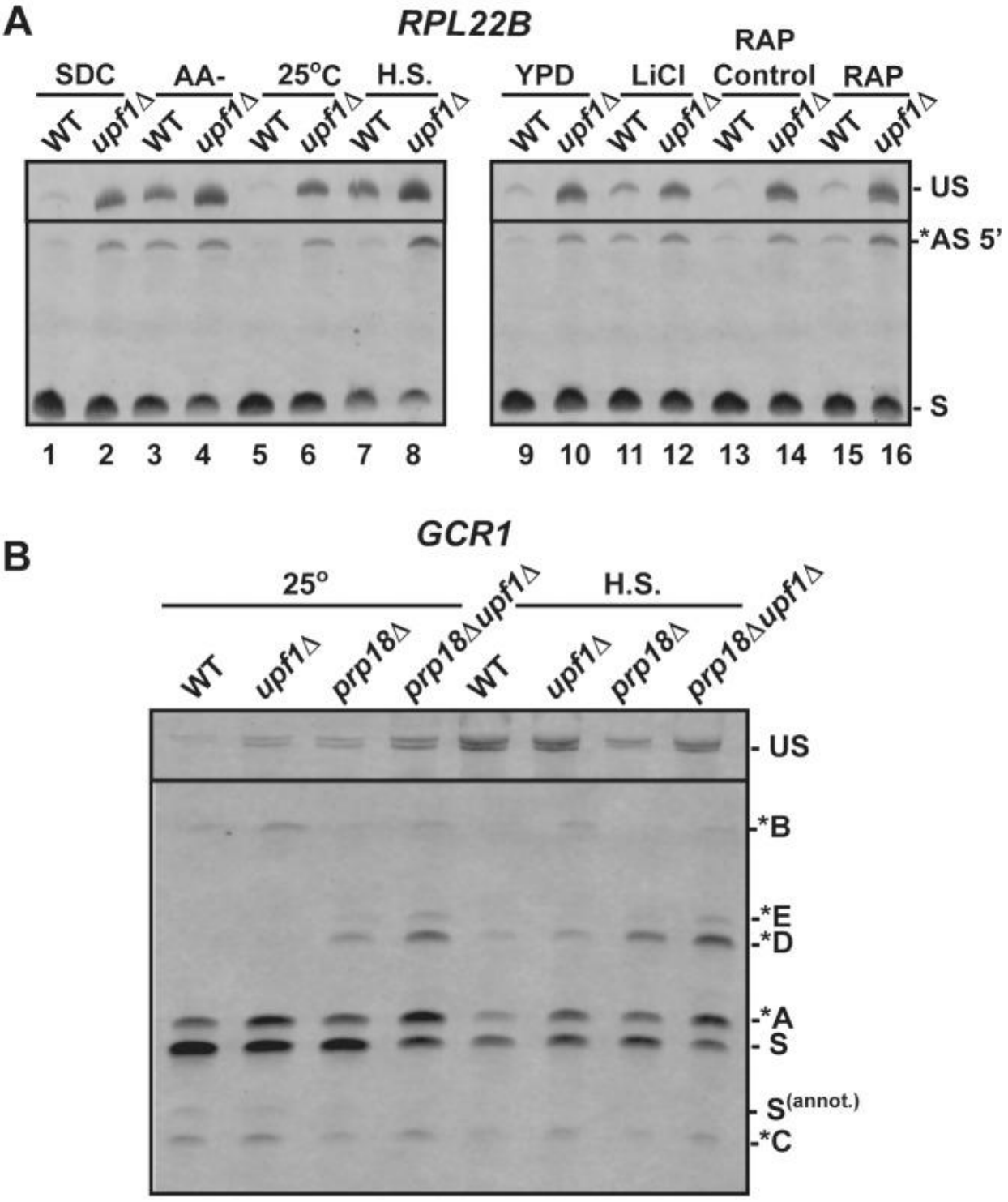

Figure 3-11. **RT-PCR analysis of alternatively spliced products under stress conditions.** A. Analysis of *RPL22B* in various stress conditions. Shown are the RT-PCR products obtained from the wild-type or *upf1Δ* mutant strain after growth in the following conditions: SDC, synthetic define complete medium at 30°C; -AA, 10 minutes in SDC medium at 30°C lacking amino acid (-AA); 25°C, log phase at 25°C in YPD; H.S, 20 minutes at 42°C in YPD; YPD: log phase at 30°C in YPD; LiCl, incubation with 300 mM LiCl in YPD at 30°C for 10 minutes; RAP control, see Methods; RAP, treatment with Rapamycin for 20 minutes. B. RT-PCR analysis of *GCR1* alternative splicing in heat-shock conditions. Labeling of the different species is similar to that of Figures 2 and 3.

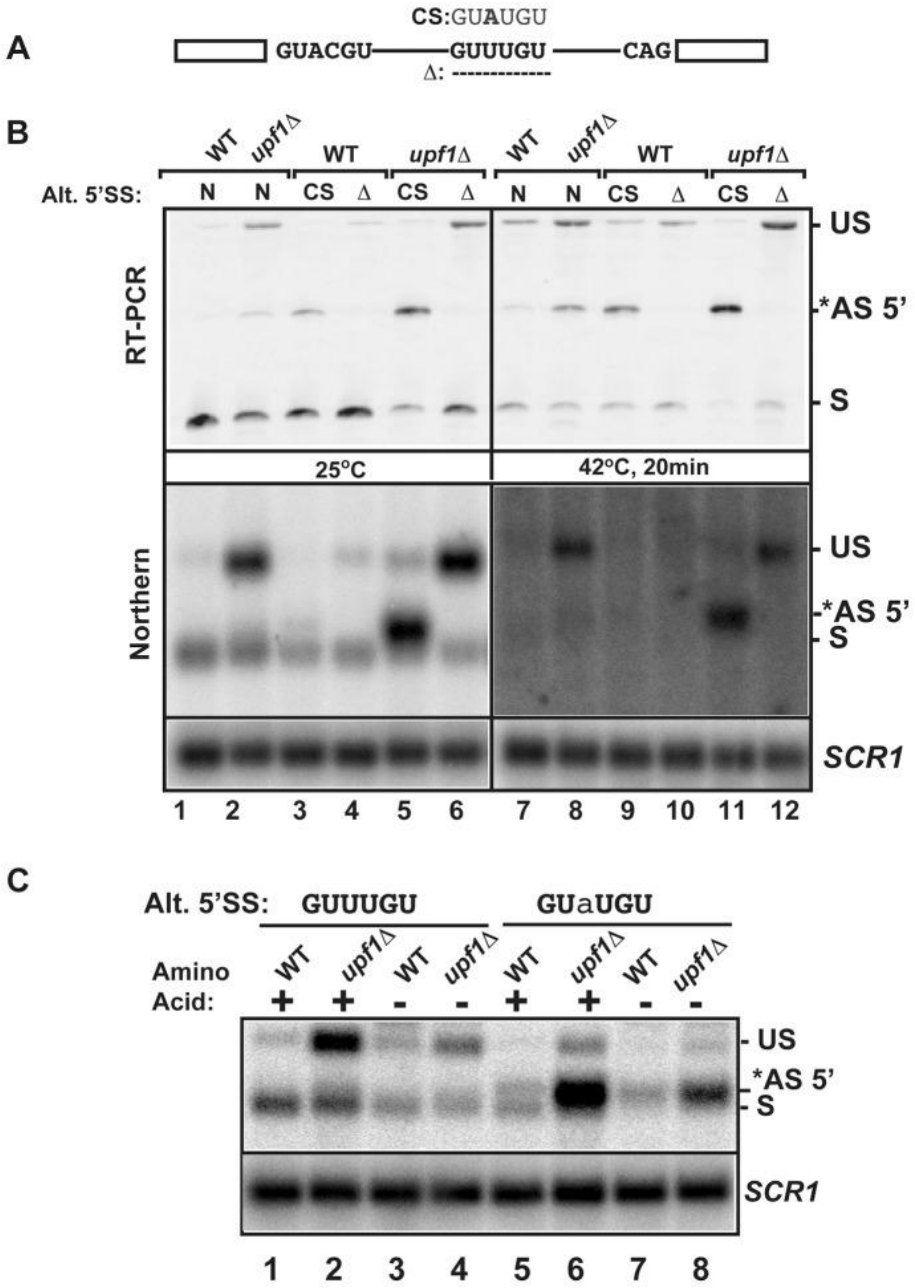

Figure 3-12. **Effects of mutations of the *RPL22B* alternative 5′ splice site on *RPL22B* splicing and expression in normal and stress conditions.** A. Organization of the *RPL22B* precursor, with the normal and alternative 5′-splice sites. Shown are the mutations to the consensus sequence (CS) GUAUGU or the deletion that entirely removes the GUUUGU sequence. B. Analysis of the effect of these mutations on*RPL22B* splicing and expression at normal temperatures (25°C) or after a 20 min heat shock at 42°C. N, natural 5′-splice site (GUUUGU); CS, consensus sequence (GUAUGU); Δ = deletion of the alternative 5′-splice site. Top panel: RT-PCR analysis. Bottom panel: northern blot analysis. US, *AS-5′, and S indicate the location of the products corresponding to the unspliced, alternatively spliced and normal spliced products, respectively. For the northern blot, *SCR1* was used as a loading control. C. Analysis of the effect of the*RPL22B* alternative splice site consensus mutation on *RPL22B* expression during amino acid starvation. Shown is a northern blot of RNA samples extracted from the indicated strains grown at 30°C in normal synthetic define complete (SDC) medium with amino acid (+) or in SDC medium lacking amino acid (−) for 10 minutes. Strains contained either the natural GUUUGU sequence at the alternative 5′-splice site of RPL22B, or the consensus GUaUGU sequence. The nucleotide mutated is highlighted in lower case. Labeling of the different species is similar to that of panel B. *SCR1* was used as a loading control.

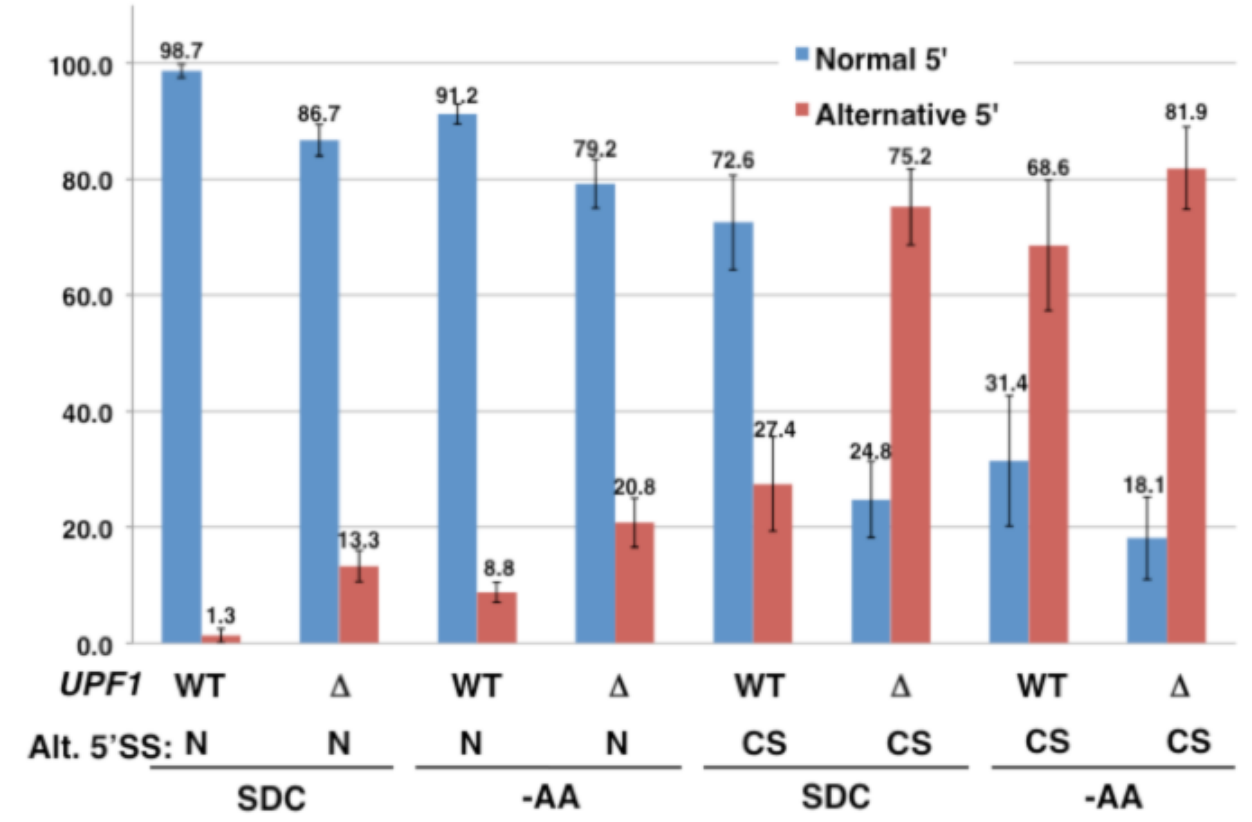

Figure 3-13. Quantitation of the use of the normal and alternative 5′-splice site of *RPL22B* under normal growth conditions in minimal medium (SDC) and after amino acid starvation (-AA) for the strains expressing the natural (N) GUUUGU sequence at the alternative 5′ splice site of *RPL22B*, or the consensus (CS) GUAUGU sequence in the context of wild-type *UPF1* (WT) or when *UPF1* has been deleted (**Δ**). Plotted are the amount of transcript spliced at the normal and alternative splice sites divided by the values obtained for all spliced species. Shown are the average of 3 independent experiments with the standard deviations.

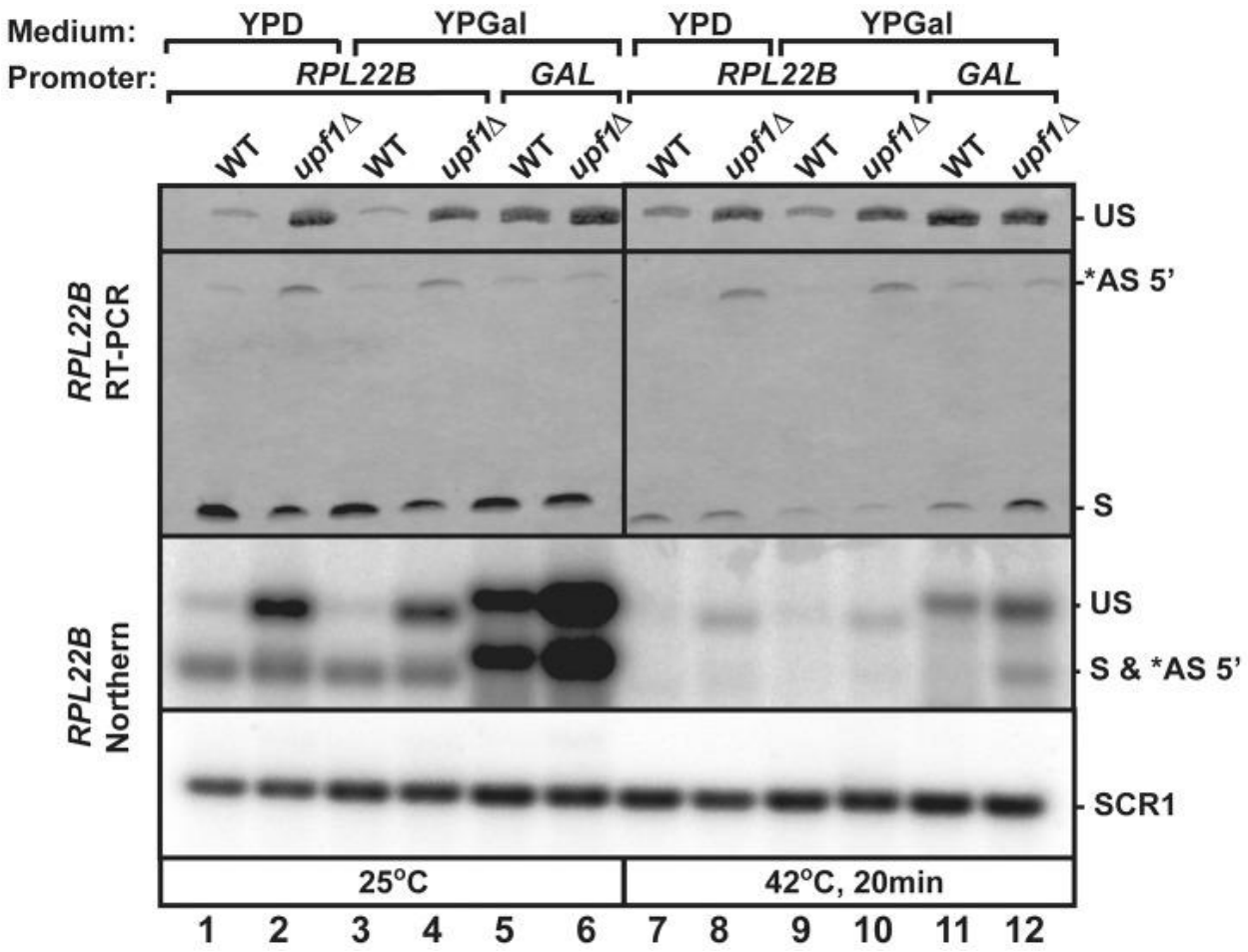

Figure 3-14. **Replacement of *RPL22B* promoter by the GAL promoter results in a decrease in alternative 5′-splice site usage.** Shown are the products generated when growing the indicated strains (wild-type or *upf1Δ* that contained the natural *RPL22B* promoter or the *GAL* promoter upstream *RPL22B*) in glucose (YPD) or galactose (YPGal)-containing media. Top panel, RT-PCR analysis. US, *AS 5′, and S indicate the location of the products corresponding to the unspliced, alternatively spliced and normal spliced products, respectively. Bottom Panel, Northern blot analysis. The labeling of the different species is similar to that of the top panel. SCR1 was used as a loading control.

Tables

	WT	***upf1Δ***	***upf2Δ***	***upf3Δ***
Total Sequences	17169945	14803384	15256510	14447952
Mapped	10530129	11782000	12245672	11476145
Low Quality	414	8226	8550	4941
Mapped to 2 positions	40566	41883	43421	41154
Repetitive	1759898	1633780	1782780	1636319
Did not match	4838938	1337495	1176087	1289393

	WT	***upf1Δ***	***upf2Δ***	***upf3Δ***
Total Sequences	100%	100%	100%	100%

Mapped	61.33%	79.59%	80.27%	79.43%
Low Quality	0.00%	0.06%	0.06%	0.03%
Mapped to 2 positions	0.24%	0.28%	0.28%	0.28%
Repetitive	10.25%	11.04%	11.69%	11.33%
Did not match	28.18%	9.04%	7.71%	8.92%

Table 3-1. Statistics of RNA-Seq analysis sequence alignments.

	Normalized Counts		
	Alternative Splicing	**PTC-generating**	**non-PTC-generating**
Wildtype	253	206	47
upf1 D	422	328	94
upf2 D	436	350	86
upf3 D	480	368	112
	p-values (vs wildtype)		
	Alternative Splicing	**PTC-generating**	**non-PTC-generating**
upf1 D	<1.0E-308	1.89E-15	5.15E-10
upf2 D	<1.0E-308	<1.0E-308	1.16E-07
upf3 D	<1.0E-308	<1.0E-308	2.22E-16

Table 3-2. Number of alternative splicing events detected in wild-type and NMD-deficient strains.

	WT	***upf1Δ***	***upf2Δ***	***upf3Δ***
Total Mapped	10530129	11782000	12245672	11476145
Mapped to intronless genes	8934033	9813692	10225911	9579339
Mapped outside of genes	516683	767386	759116	739233
Mapped only to exon of ICGs	1042627	1131395	1189492	1090340
Mapped to exon-intron junction of ICG	28972	36055	37142	34861
Mapped to only intron of ICG	7814	33472	34011	32372

	WT	***upf1Δ***	***upf2Δ***	***upf3Δ***
Total Mapped	100%	100%	100%	100%
Mapped to intronless genes	85.32%	83.29%	83.51%	83.47%
Mapped outside of genes	4.91%	6.51%	6.20%	6.44%
Mapped only to exon of ICGs	9.90%	9.60%	9.71%	9.50%
Mapped to exon-intron junction of ICG	0.28%	0.31%	0.30%	0.30%
Mapped to only intron of ICG	0.07%	0.28%	0.28%	0.28%

Table 3-3. Mapping of RNA-Seq reads in wild-type and NMD-deficient strains in various genomic elements and in intron-containing genes.

Gene	wild-type RPKM	*upf1D* RPKM	*upf2D* RPKM	*upf3D* RPKM	5' SS	3'SS
YBL091C-A	43.1662	42.7591	44.2333	42.5784	GTATGT	CAG
YBL059W	15.0117	18.5209	20.7662	17.9664	GTATGC	TAG
YBR090C	56.6191	66.9341	69.2684	61.6337	GTATGT	TAG
YBR186W	2.3531	12.5685	11.5627	11.5155	GTATGT	TAG
YBR219C	44.2678	39.7853	43.3827	43.7956	GTACGT	TAG
YBR230C	157.1036	150.8893	119.3669	138.3439	GTATGT	CAG
YCL005W-A_1	139.4621	112.6332	127.885	117.0051	GTATGT	TAG
YCL005W-A_2	238.208	190.4725	195.2775	193.2174	GTATGT	CAG
YCR028C-A	192.026	174.5351	183.9385	175.3424	GTATGT	CAG
YCR097W_2	32.5284	28.8771	22.5273	24.6388	GTATGT	TAG
YDL219W	37.525	59.7687	54.8015	54.6291	GTATGT	CAG
YDL189W	48.8606	48.2341	44.674	49.5196	GTATGT	AAG
YDL137W	325.318	311.9346	286.3038	284.9463	GTATGT	TAG
YDL125C	345.2209	438.611	440.1504	437.5132	GTATGT	CAG
YDL082W	784.0993	750.4385	772.5178	725.9987	GTATGT	CAG
YDL079C	7.3778	31.8987	33.0225	32.0545	GTATGT	TAG
YDL064W	78.537	79.3243	78.0436	74.0851	GTAAGT	CAG
YDR059C	27.4062	30.1905	22.2879	25.9268	GTATGT	AAG
YDR099W	306.0935	301.987	281.078	287.3123	GTATGT	CAG
YDR305C	22.5071	29.9789	28.4691	27.9799	GTATGC	CAG
YDR318W	12.0959	33.5054	32.6056	30.0691	GTATGT	CAG
YDR367W	55.1827	56.5835	55.7898	55.9981	GTATGT	TAG
YDR381W	250.3131	205.6448	227.7169	212.0212	GTATGT	TAG
YDR381C-A	28.0768	33.4581	27.2205	28.5406	GTATGT	CAG
YDR535C	2.6537	8.3012	6.1939	6.7831	GTACGT	CAG
YER003C	165.4168	136.6557	134.8364	134.9615	GTATGT	TAG
YER007C-A	104.3578	108.3481	117.108	115.2255	GTATGT	TAG
YER014C-A	32.0663	26.6383	27.0437	25.8394	GTCAGT	CAG
YER044C-A	1.8575	4.8421	5.1912	4.332	GTACGT	CAG
YER131W	1425.2754	1438.6352	1436.562	1392.5021	GTATGT	TAG
YER179W	4.6302	9.3743	10.3194	8.5837	GTATGT	TAG
YFL039C	1542.3491	1273.0533	1311.2173	1227.6468	GTATGT	TAG
YFL034C-B	40.1186	37.6241	38.4679	39.7362	GTATGT	CAG
YFL031W	309.0024	263.2671	317.4207	300.9092	CCGTGA	CCG
YFR045W	21.0354	19.9867	17.5616	19.5825	GTAAGT	CAG
YGL251C	4.0235	16.8369	15.787	17.7991	GTAGTA	TAG
YGL187C	126.423	96.0151	96.1224	85.1404	GTATGT	TAG
YGL183C	0.71944	1.5432	1.2373	1.0562	GTATGT	TAG
YGL033W	0.86727	0.51674	0.74577	0.39789	GTTAAG	CAG
YGR029W	59.3118	75.4943	81.6615	80.5638	GTATGT	TAG
YGR183C	119.5338	149.0595	128.3833	127.4545	GTATGT	TAG
YGR225W	1.3856	2.0481	2.4288	2.1026	GTACGT	CAG
YHR012W	170.9157	150.3561	150.6265	154.8765	GTATGT	TAG
YHR039C-A	497.1243	464.7227	399.3129	423.0578	GTATGT	TAG
YHR041C	73.6621	68.3829	71.857	78.0519	GTATGT	TAG

YHR079C-A	3.4408	4.6128	0.59175	5.6829	GTATGT	AAG
YHR123W	60.3225	68.2755	71.801	78.468	GTATGT	TAG
YHR141C	1449.3348	1309.6169	1292.3378	1288.0572	GTATGT	CAG
YHR218W	42.1895	43.2339	41.687	49.0508	GCATGT	CAG
YIL148W	2068.8759	1794.6616	1853.3152	1870.4117	GTATGC	CAG
YIL111W	108.0858	103.3021	96.7044	84.8442	GCATGT	CAG
YIL073C	5.7732	17.1606	16.8176	16.5168	GTATAT	AAG
YIL004C	75.0427	70.2348	74.8088	64.5913	GTATGA	TAG
YJL189W	3289.097	3190.9823	3107.3249	3230.782	GTATGT	TAG
YJL041W	64.8471	63.3474	60.4533	60.9467	GTATGT	TAG
YJL031C	29.725	38.6424	34.6066	34.4476	GTATGT	CAG
YJL024C	34.0902	57.3089	61.6998	54.0698	GTATGT	TAG
YJR079W	23.3097	20.833	20.2916	24.0288	GCATGT	TAG
YJR094W-A	1057.2156	1008.1596	1000.4266	1063.7619	GTATGC	CAG
YJR112W-A	71.0803	107.7658	117.7905	116.9752	GTATGT	AGG
YKL006C-A	172.8115	201.5065	188.0437	208.9516	GTATGT	CAG
YKR005C	4.333	6.6158	7.2968	8.0621	GTATGT	AAG
YLL050C	539.2375	548.742	507.738	523.0254	GTATGC	TAG
YLR054C	44.9286	50.0276	38.334	42.6272	GTATGG	CAG
YLR078C	80.6239	87.0693	99.6604	99.1112	GTATGT	TAG
YLR128W	21.4552	29.9683	27.6238	27.7548	GTATGT	TAG
YLR199C	68.5672	61.7924	62.106	57.1478	GTATGT	AAG
YLR202C	20.6194	22.5815	18.7297	21.8509	GTATGA	CAG
YLR211C	7.9487	16.327	17.5075	18.6814	GTAAGT	TAG
YLR275W	15.3998	21.9197	26.2396	22.7656	GTATGT	CAG
YLR333C	712.8006	647.0488	703.2497	661.2181	GTATGT	CAG
YLR445W	2.1773	1.4969	1.7283	1.6905	GTAAGT	CAG
YML085C	192.7576	143.9216	152.3254	147.887	GTATGT	CAG
YML067C	104.6505	110.442	113.7401	115.8539	GTATGT	CAG
YML036W	53.0485	74.6156	64.4617	66.3903	GTATGT	TAG
YML025C	194.1225	143.036	202.2094	161.9276	GTACGT	TAG
YML024W	1570.0517	1247.5214	1282.94	1275.0452	GTATGT	TAG
YML017W	45.6709	43.3902	41.0141	46.6492	GTATGT	CAG
YMR194C-B	53.0438	55.0542	72.8332	60.8391	GTTTGT	TAG
YMR242C	1521.6453	1459.2326	1486.1135	1382.108	GTGAGT	CAG
YMR292W	51.9236	44.7783	44.062	42.4193	GTATGT	TAG
YNL312W	67.0074	56.377	62.3886	61.5897	GTATGT	CAG
YNL138W-A	57.4211	88.165	84.8267	80.7202	GTATGT	CAG
YNL130C	98.9828	103.9039	100.5224	103.7243	GTATGT	TAG
YNL066W	222.4057	204.2997	221.0305	208.446	GTATGT	TAG
YNL050C	130.0083	116.0901	108.5807	107.2871	GTATGC	CAG
YNL044W	131.6284	127.5526	135.7949	136.8597	GTAAGT	TAG
YNR053C	61.1654	76.5097	75.5131	62.684	GTATGT	TAG
YOL047C	257.7638	240.2025	278.8386	246.6855	GTATGT	AAG
snR17A	7.9851	6.1171	9.0735	7.5885	GTATGT	CAG
YOR318C	609.8281	488.726	437.9298	479.5398	GTATGA	CAG
YPL241C	13.2975	31.1314	27.3217	29.1537	GTATGT	TAG
YPL230W	39.6883	33.2677	29.0671	36.9275	GTATGT	AAG
snR17B	13.4439	19.9406	21.1533	19.6846	GTATGT	CAG

YPR010C-A	355.1453	194.5542	387.426	222.021	GTATGT	TAG
YPR153W	48.0441	82.4674	81.4685	70.4514	GTATGT	CAG

Table 3-4. List of intron-containing genes for which no alternative splicing events were detected. Shown is the list of intron-containing genes for which no alternative splicing junctions were detected in any of the strains sequenced. The adjusted number of reads obtained from each strain (RPKM) and the sequence of the 5′- and 3′-splice sites is shown for each of these genes.

References

1. Isken O, Maquat LE (2007) Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. Genes Dev 21: 1833–1856.

2. Kervestin S, Jacobson A (2012) NMD: a multifaceted response to premature translational termination. Nat Rev Mol Cell Biol 13: 700–712.

3. Green RE, Lewis BP, Hillman RT, Blanchette M, Lareau LF, et al. (2003) Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes. Bioinformatics 19 (Suppl 1) i118–121.

4. Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE (2007) Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature 446: 926–929.

5. Ni JZ, Grate L, Donohue JP, Preston C, Nobida N, et al. (2007) Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev 21: 708–718.

6. Mendell JT, Sharifi NA, Meyers JL, Martinez-Murillo F, Dietz HC (2004) Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. Nat Genet36: 1073–1078.

7. Chan WK, Huang L, Gudikote JP, Chang YF, Imam JS, et al. (2007) An alternative branch of the nonsense-mediated decay pathway. EMBO J 26: 1820–1830.

8. Weischenfeldt J, Damgaard I, Bryder D, Theilgaard-Monch K, Thoren LA, et al. (2008) NMD is essential for hematopoietic stem and progenitor cells and for eliminating by-products of programmed DNA rearrangements. Genes Dev 22: 1381–1396.

9. He F, Peltz SW, Donahue JL, Rosbash M, Jacobson A (1993) Stabilization and ribosome association of unspliced pre-mRNAs in a yeast upf1- mutant. Proc Natl Acad Sci U S A 90: 7034–7038.

10. Mitrovich QM, Anderson P (2000) Unproductively spliced ribosomal protein mRNAs are natural targets of mRNA surveillance in C. elegans. Genes Dev 14: 2173–2184.

11. Jaillon O, Bouhouche K, Gout JF, Aury JM, Noel B, et al. (2008) Translational control of intron splicing in eukaryotes. Nature 451: 359–362.

12. Sayani S, Janis M, Lee CY, Toesca I, Chanfreau GF (2008) Widespread impact of nonsense-mediated mRNA decay on the yeast intronome. Mol Cell 31: 360–370.

13. Sayani S, Chanfreau GF (2012) Sequential RNA degradation pathways provide a fail-safe mechanism to limit the accumulation of unspliced transcripts in Saccharomyces cerevisiae. RNA 18: 1563–1572.

14. Vijayraghavan U, Company M, Abelson J (1989) Isolation and characterization of pre-mRNA splicing mutants of Saccharomyces cerevisiae. Genes Dev 3: 1206–1216.

15. Wahl MC, Will CL, Luhrmann R (2009) The spliceosome: design principles of a dynamic RNP machine. Cell 136: 701–718.

16. Grund SE, Fischer T, Cabal GG, Antunez O, Perez-Ortin JE, et al. (2008) The inner nuclear membrane protein Src1 associates with subtelomeric genes and alters their regulated gene expression. J Cell Biol 182: 897–910.

17. Mishra SK, Ammon T, Popowicz GM, Krajewski M, Nagel RJ, et al. (2011) Role of the ubiquitin-like protein Hub1 in splice-site usage and alternative splicing. Nature 474: 173–178.

18. Meyer M, Plass M, Perez-Valle J, Eyras E, Vilardell J (2011) Deciphering 3′ss selection in the yeast genome reveals an RNA thermosensor that mediates alternative splicing. Mol Cell 43: 1033–1039.

19. Plass M, Codony-Servat C, Ferreira PG, Vilardell J, Eyras E (2012) RNA secondary structure mediates alternative 3′ss selection in Saccharomyces cerevisiae. RNA 18: 1103–1115.

20. Marshall AN, Montealegre MC, Jimenez-Lopez C, Lorenz MC, van Hoof A (2013) Alternative splicing and subfunctionalization generates functional diversity in fungal proteomes. PLoS Genet 9: e1003376.

21. Kawashima T, Pellegrini M, Chanfreau GF (2009) Nonsense-mediated mRNA decay mutes the splicing defects of spliceosome component mutations. RNA 15: 2236–2247.

22. Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Res 12: 656–664.

23. Liao XC, Tang J, Rosbash M (1993) An enhancer screen identifies a gene that encodes the yeast U1 snRNP A protein: implications for snRNP protein function in pre-mRNA splicing. Genes Dev 7: 419–428.

24. Neubauer G, Gottschalk A, Fabrizio P, Seraphin B, Luhrmann R, et al. (1997) Identification of the proteins of the yeast U1 small nuclear ribonucleoprotein complex by mass spectrometry. Proc Natl Acad Sci U S A 94: 385–390.

25. Puig O, Gottschalk A, Fabrizio P, Seraphin B (1999) Interaction of the U1 snRNP with nonconserved intronic sequences affects 5′ splice site selection. Genes Dev 13: 569–580.

26. Gottschalk A, Tang J, Puig O, Salgado J, Neubauer G, et al. (1998) A comprehensive biochemical and genetic analysis of the yeast U1 snRNP reveals five novel proteins. RNA 4: 374–393.

27. Umen JG, Guthrie C (1995) Prp16p, Slu7p, and Prp8p interact with the 3′ splice site in two distinct stages during the second catalytic step of pre-mRNA splicing. RNA 1: 584–597.

28. Aronova A, Bacíková D, Crotti LB, Horowitz DS, Schwer B (2007) Functional interactions between Prp8, Prp18, Slu7, and U5 snRNA during the second step of pre-mRNA splicing. RNA 13: 1437–1444.

29. Villa T, Guthrie C (2005) The Isy1p component of the NineTeen complex interacts with the ATPase Prp16p to regulate the fidelity of pre-mRNA splicing. Genes Dev 19: 1894–1904.

30. Saha D, Banerjee S, Bashir S, Vijayraghavan U (2012) Context dependent splicing functions of Bud31/Ycr063w define its role in budding and cell cycle progression. Biochem Biophys Res Commun424: 579–585.

31. Rodriguez-Navarro S, Igual JC, Perez-Ortin JE (2002) SRC1: an intron-containing yeast gene involved in sister chromatid segregation. Yeast 19: 43–54.

32. Chanfreau GF (2010) A dual role for RNA splicing signals. EMBO Rep 11: 720–721

33. Crotti LB, Horowitz DS (2009) Exon sequences at the splice junctions affect splicing fidelity and alternative splicing. Proc Natl Acad Sci U S A 106: 18954–18959.

34. Zhou Y, Chen C, Johansson MJ (2013) The pre-mRNA retention and splicing complex controls tRNA maturation by promoting TAN1 expression. Nucleic Acids Res 41: 5669–5678.

35. Yan BC, Westfall BA, Orlean P (2001) Ynl038wp (Gpi15p) is the Saccharomyces cerevisiae homologue of human Pig-Hp and participates in the first step in glycosylphosphatidylinositol assembly.Yeast 18: 1383–1389.

36. Muhlrad D, Parker R (1999) Aberrant mRNAs with extended 3′ UTRs are substrates for rapid degradation by mRNA surveillance. RNA 5: 1299–1307.

37. Amrani N, Ganesan R, Kervestin S, Mangus DA, Ghosh S, et al. (2004) A faux 3′-UTR promotes aberrant termination and triggers nonsense-mediated mRNA decay. Nature 432: 112–118.

38. Uemura H, Jigami Y (1995) Mutations in GCR1, a transcriptional activator of Saccharomyces cerevisiae glycolytic genes, function as suppressors of gcr2 mutations. Genetics 139: 511–521.

39. Holland MJ, Yokoi T, Holland JP, Myambo K, Innis MA (1987) The GCR1 gene encodes a positive transcriptional regulator of the enolase and glyceraldehyde-3-phosphate dehydrogenase gene families in Saccharomyces cerevisiae. Mol Cell Biol 7: 813–820.

40. Clifton D, Fraenkel DG (1981) The gcr (glycolysis regulation) mutation of Saccharomyces cerevisiae.J Biol Chem 256: 13074–13078.

41. Mayas RM, Maita H, Staley JP (2006) Exon ligation is proofread by the DExD/H-box ATPase Prp22p. Nat Struct Mol Biol 13: 482–490.

42. James SA, Turner W, Schwer B (2002) How Slu7 and Prp18 cooperate in the second step of yeast pre-mRNA splicing. RNA 8: 1068–1077.

43. Bergkessel M, Whitworth GB, Guthrie C (2011) Diverse environmental stresses elicit distinct responses at the level of pre-mRNA processing in yeast. RNA 17: 1461–1478.

44. Pleiss JA, Whitworth GB, Bergkessel M, Guthrie C (2007) Rapid, transcript-specific changes in splicing in response to environmental stress. Mol Cell 27: 928–937.

45. Li B, Nierras CR, Warner JR (1999) Transcriptional elements involved in the repression of ribosomal protein synthesis. Mol Cell Biol 19: 5393–5404.

46. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, et al. (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11: 4241–4257.

47. Garre E, Romero-Santacreu L, Barneo-Munoz M, Miguel A, Perez-Ortin JE, et al. (2013) Nonsense-mediated mRNA decay controls the changes in yeast ribosomal protein pre-mRNAs levels upon osmotic stress. PLoS One 8: e61240.

48. Yost HJ, Lindquist S (1991) Heat shock proteins affect RNA processing during the heat shock response of Saccharomyces cerevisiae. Mol Cell Biol 11: 1062–1068.

49. Vogel JL, Parsell DA, Lindquist S (1995) Heat-shock proteins Hsp104 and Hsp70 reactivate mRNA splicing after heat inactivation. Curr Biol 5: 306–317.

50. Awan AR, Manfredo A, Pleiss JA (2013) Lariat sequencing in a unicellular yeast identifies regulated alternative splicing of exons that are evolutionarily conserved with humans. Proc Natl Acad Sci U S A110: 12762–12767.

51. Hulsen T, de Vlieg J, Alkema W (2008) BioVenn - a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. BMC Genomics 9: 488.

Chapter 4: A Naive Bayesian Classifier for Detecting miRNAs in Plants

Stephen Douglass[1], Ssu-Wei Hsu[2,3], Shawn Cokus[4],Robert Goldberg[4], John Harada[2], and Matteo Pellegrini[4]

[1]Bioinformatics Interdepartmental Program, UCLA, Box 951606, Los Angeles CA 90095-1606

[2]Department of Plant Biology, University of California, Davis, CA 95616

[3]NCHU-UCD Plant and Food Biotechnology Center, NCHU and Agricultural Biotechnology Center, NCHU, Taichung 40227,Taiwan

[4]Department of Molecular, Cell and Developmental Biology, University of California, Los Angeles, CA 90095, USA

Abstract

microRNAs (miRNAs) are important regulatory molecules in eukaryotic organisms. Existing methods for identification of mature miRNA sequences in plants rely extensively on the search for stem loop structures, leading to high false negative rates. Here, we describe a probabilistic method for ranking putative novel plant miRNAs using a naïve Bayes classifier. We use a number of properties to construct the classifier, including sequence length, number of observations, existence of detectable predicted miRNA* sequences, the distribution of nearby reads, and mapping multiplicity. We apply the methodology to small RNA data from soybean, peach, *Arabidopsis*, and rice and provide experimental validation of several predictions in soybean. The approach performs well overall and strongly enriches for known miRNAs over other types of sequences. By utilizing a Bayesian approach to rank putative miRNAs, our method is able to score miRNAs that would be eliminated by other methods, such as those that are lowly expressed or lack detectable miRNA* sequences. As a result, we are able to detect

several novel miRNA candidates in soybean, including some that are 24 nucleotides long, a class that is almost universally eliminated by other methods.

Introduction

MicroRNAs (miRNAs) are small, approximately 18-24 nucleotide long single-stranded, non-coding RNAs that function in eukaryotic gene regulation. miRNAs induce post-transcriptional gene silencing by base pairing with target mRNAs. miRNA-mediated gene silencing results from either mRNA cleavage or translational repression (1, 2). Longer, partially double-stranded RNAs called primary miRNAs (pri-miRNAs) are typically transcribed by RNA polymerase II and cleaved by the RNAse III Dicer-like1 (DCL1) into approximately 70 nucleotide long precursor miRNAs (pre-miRNAs) (3-5). Pre-miRNAs are then cleaved by DCL1 into approximately 22 nucleotide miRNA:miRNA* duplexes, in which miRNA* is partially complementary to the miRNA (6). The miRNA:miRNA* duplex is unwound, and the miRNA* is degraded while the miRNA is loaded into the RNA-induced silencing complex (RISC) to induce gene silencing (7).

Several computational methods have been developed to identify novel miRNAs. These include both *ab initio* methods that require only the genome sequence as input as well as methods that attempt to identify miRNAs within reads from small RNA libraries. To discover miRNAs, most computational approaches rely on finding stereotypical hairpin structures (8, 9, 10, 11), finding conserved sequences in related organisms (12, 13), or through a combination of these properties (14-18). The limitation of conservation-based approaches is that they require genome sequences from related species and cannot identify non-conserved species-specific miRNAs. Structure-based approaches tend to have high false negative rates, as they require that predicted miRNAs have hairpins, which are often difficult to detect computationally. As a result, methods have

been developed to separate miRNAs from non-miRNAs assuming both classes form hairpin structures using nearby sequence information such as promoters and splice sites (8).

While the core machinery of the miRNA biogenesis pathway is conserved between plant and animal lineages, other aspects, such as targeting mechanisms, diverge between the kingdoms (19). For example, animal miRNAs tend to target 3' UTRs while plant miRNA target sequences are more uniformly distributed along transcripts. As a result of these and other differences, the criteria for annotating miRNAs are usually distinct for plants and animals (20, 21). To date, most of the software available for miRNA prediction is animal-specific (13), mammal-specific (15), or specific to certain animal species (12, 16, 17, 22). The miRNA prediction tools that do not describe themselves as animal-specific or plant-specific are generally validated using exclusively animal genomes and datasets (8, 9, 14, 18, 23, 24). Thus, to date there have been only limited resources for identifying miRNAs in plants.

To overcome these limitations, here we describe a machine learning approach using a Naïve Bayes Classifier to identify plant miRNAs. The approach combines small RNA deep sequencing data and genomic features to determine the probability that a specific read is a miRNA. We have applied this methodology to datasets from soybean, peach, rice and *Arabidopsis*. We find that the overall accuracy of the method is high as it strongly enriches for known miRNAs among its top predictions. We also experimentally validate newly predicted miRNAs in soybean.

Results

We identify putative novel miRNAs from small RNA sequence data in four model plants, *Arabidopsis Thaliana*, *Glycine max*, *Oryza sativa*, and *Prunus persica* (hereafter *Arabidopsis*, soybean, rice, and peach, respectively) using a naïve Bayes classifier approach. We present

results of our miRNA prediction and bioinformatic validation in each of the four organisms and experimental validation in soybean.

Read Processing

Known miRNA sequences and small RNA sequence reads were processed in parallel in the four model plants. Known miRNAs were downloaded from miRBase (http://www.mirbase.org/). We kept only the known miRNAs with at least two independent lines of experimental evidence in *Arabidopsis* due to its relatively large number of available sequences. In contrast, for the other plant species we only required one line of evidence. It is likely that some of the annotated sequences with only a single line of evidence are not true miRNAs, but to be more stringent would dramatically reduce our statistical power. Many miRNAs exist as members of gene families in which multiple primary miRNAs give rise to identical mature miRNA sequences. For all downstream analyses we consider only unique mature miRNAs in each organism.

Small RNA-Seq Samples

Arabidopsis small RNA-seq data was gathered from two independent experiments, one from whole flowers (26) and the other from unopened flower buds (27). Reads from both samples were merged to form a single library. Rice small RNA-seq data was derived from leaf tissue (28). Peach small RNAs were generated from flower buds chilled to induce dormancy (29). Soybean small RNAs were isolated from soybean seeds during the early maturation stage of development. All samples were from untreated, wild-type organisms. The number of unique sequence reads in the soybean, *Arabidopsis*, rice, and peach datasets are 852982, 977497, 771765, and 17407051 respectively. It is unclear why peach has roughly 20X as many unique reads as the other organisms. The peach genome is an average size among the organisms

examined, so it appears that a much larger portion of the peach genome is expressed as small RNAs.

Construction of the Naïve Bayes Classifier

To differentiate mature miRNA sequences from contaminants, such as siRNAs and degraded larger RNAs in small RNA sequence data, we constructed a naïve Bayes classifier (NBC) using five parameters (Figure 4-1). Each unique read in a small RNA sequence library and each known miRNA were evaluated based on length, the number of times the read is found in the library, the distribution of mapped reads around the genomic location of the read (using entropy), the number of locations the read maps to in the genome, and the presence of a detectable miRNA* sequence.

The NBC was trained on each variable using known miRNAs as true positives and all reads in the small RNA-Seq library as true negatives. While there are undoubtedly some true miRNAs in the small RNA-Seq data, they are likely significantly outnumbered by non-miRNA sequences, so it is appropriate to use this set as our true negatives. Training and prediction of the NBC were done separately for each organism, and the distributions for these variables are shown in Figure 4-2 for soybean, Figure 4-3 for *Arabidopsis*, Figure 4-4 for rice, and Figure 4-5 for peach.

The number of times a read is observed in a library is a common criterion for miRNA annotation in smRNA-Seq based approaches. Sequences observed few times are more likely to be degraded larger RNAs, such as rRNAs or mRNAs. In soybean, for example, different values of read counts (Figure 4-2A) have differing power to discriminate miRNAs from other small RNAs. Sequences with a single read count are much less likely to be known miRNAs, and those observed twice are slightly depleted for known miRNAs. In contrast, sequences observed six or

more times are far more likely to be known miRNAs than other small RNAs, and this is especially true for those observed more than twenty times. Sequences with read counts between three and five, however, are relatively uninformative in determining if a sequence is likely a miRNA.

The entropy metric measures the distribution of reads within a window. If the reads are uniformly distributed across the window, then the entropy is high. In contrast, if all the reads within the window map to a single base, then the entropy is zero. The rational for using entropy to discriminate miRNAs from other types of small RNAs is that we expect miRNAs to originate from a single location within a window, and thus have low entropy. In contrast, siRNAs usually occur in clusters [28], and thus have high entropy. Therefore the entropy metric should discriminate miRNAs from siRNAs. In fact, we find that entropy values in soybean (Figure 4-2B) are the most effective variable for distinguishing true miRNAs. An entropy of zero corresponds to a sequence having a roughly 2-fold increased chance to be a known miRNA, while the next bin, with an entropy between zero and 0.005, is ~60X more likely to be associated with known miRNAs versus other small RNAs. In contrast, an entropy value between 0.005 and 0.01 is associated with a depletion of known miRNAs, with a stronger depletion for entropy values greater than 0.01.

Sequence length is a common criterion for miRNA annotation. miRNAs tend to be 18-24 nucleotides long, with the majority of annotated plant miRNAs having a length of 21 nucleotides (Figure 4-2C). A common contaminant in miRNA analysis are siRNAs, and their most common length is 24 nucleotides. As a result, many methods filter out potential miRNAs if they are 24 nucleotides long, even though there are many examples of 24 nucleotide miRNAs. To overcome this limitation, the naïve Bayes classifier calculates the relative frequency of any read length in

the known miRNA population, versus the whole population, to compute weights. For example, the soybean sequence length distributions reflect these trends (Figure 4-2C). The most dramatic effects of sequence length are a strong positive weight for being classified a miRNA for reads of length 21, and a strong penalty for those of length 24. Thus while 24 nucleotide sequences are penalized, if other properties of the read indicate that it is likely a miRNA, the read may still be given a high miRNA score. Sequences of length 18 and 23 are roughly even (receive no significant bonus or penalty), and sequences of length 19, 20, and 22 receive more modest positive weights for being classified as miRNAs.

We chose the multiplicity of reads as another criterion. Multiplicity involves the computation of the number of places a mature miRNA sequence or putative mature miRNA map to the genome. A significant fraction of miRNAs map to a number of locations in the genome because many miRNAs exist as members of families. These miRNA families have different primary miRNA transcript sequences but identical mature miRNA sequences (30). In soybean, sequences that map to a single location, that is those with a multiplicity of 1, receive a strong positive weight for being a miRNA (Figure 4-2D). Those that map to two locations receive a strong penalty, and those that map to more than two locations do not receive significant bonuses or penalties.

Finally, the presence of a detectable miRNA* sequence is a criterion used in many methods, and it is often required to annotate miRNAs. Here, we search for a miRNA* sequence without requiring that the predicted pri-miRNA sequences form a perfect hairpin between the putative miRNA and miRNA*. In contrast, the use of RNA folding software to predict stereotypical miRNA hairpin structures can lead other approaches to generate high false negative rates (31). In soybean, roughly half of the annotated miRNAs contain miRNA* sequences that meet these

criteria, compared to ~20% of all sequence reads (figure 4-2E). Details on the calculation of the variables can be found in the methods section.

While the examples we presented are from soybean, most of the corresponding distributions show similar trends in *Arabidopsis*, rice, and peach. In general, read counts have positive weights for reads observed more than 6 times and a penalty for those observed once or twice, but in the three other organisms the weights for those observed more than 100 times is far greater than in soybean (figures 4-3A, 4-4A, 4-5A). The entropy distributions in these organisms are also similar, but in peach (figure 4-5B), and particularly Arabidopsis (figure 4-3B), the separation of miRNAs and non-miRNAs is even more dramatic than in soybean (figure 4-2B) and rice (figure 4-4B). The read length distributions in all three organisms (figures 4-3C, 4-4C, 4-5C) show an even stronger bias for 21 nucleotide miRNAs and a weaker preference for 24 nucleotide non-miRNAs than in soybean (figure 4-2C). Here, rice and *Arabidopsis* seem more similar to one another than to soybean and peach. The multiplicity distribution in *Arabidopsis* (figure 4-3D) is similar to that of soybean (figure 4-2D) but is less informative in rice (figure 4-4D) and peach (figure 4-5D), with almost no distinction between miRNA and non-miRNA for any multiplicity value. This is an interesting finding because we suspected that multiplicity would be most informative for distinguishing true miRNAs in organisms with large, repetitive genomes, however *Arabidopsis* has the smallest genome of the four species, while soybean has the largest. Unsurprisingly, there are more detectable miRNA* sequences among known miRNAs than small RNA-seq reads in the other three organisms (figures 4-3E, 4-4E, 4-5E). What is surprising is that the differentiating power is strongest in soybean, with both the highest fraction of known miRNAs and lowest fraction of small RNA-seq reads with a detectable miRNA*. *Arabidopsis* showed the weakest signal with the highest fraction of small RNA-seq

reads with a detectable miRNA* of all four organisms. This is unexpected, as we suspected that *Arabidopsis* would have the strongest signal since it has the most high-confidence annotated miRNAs, and most of the work investigating the parameters for plant miRNA* detection was done in *Araibdopsis*. The lack of miRNA* sequences in many Arabidopsis miRNAs further argues against the use of stringent filters for identifying true miRNAs.

Bioinformatic Validation of Classifier

Each unique small RNA-seq read and known miRNA was scored by our classifier based on the criteria described above. Table 4-1 shows the enrichment of previously known miRNAs among our top predictions. In all four organisms, the top 100 candidates capture a significantly larger fraction of known miRNAs than the top 250, 500, and 1000 candidates. Performance of the classifier for each organism is shown as a ROC curve in Figure 4-6. In this figure we rank the small RNAs by their likelihood of being known miRNAs, based on our classifier, and then compute the relative fraction of known (y axis) and total (x-axis) unique reads. We score the classifier performance by computing the area under the curve (AUC) for each organism. A prefect classifier would predict all the knowns before capturing any of the other reads (upper left corner of the plot, AUC=1), while a random classifier would appear as a diagonal line from the lower left to the upper right (AUC=0.5). We find that *Arabidopsis* shows the strongest performance of the four species (AUC=0.9998), coming very close to the upper left corner, likely due to the availability of a relatively large number of high quality known miRNAs for training. Soybean had the weakest AUC value of 0.9948 with intermediate values of 0.9996 and 0.9988 for peach and rice respectively.

It is difficult to compare our method to other methods for identifying novel miRNAs because our method is a statistical approach that assigns likelihoods to individual reads, while most other

methods are based on filters. In addition, many annotated miRNAs that we use were discovered by these or similar methods, so test sets tend to be biased for miRNAs that meet these criteria.

Conservation of predicted miRNAs

To further characterize high scoring miRNAs, we tested the homology of our top candidates among the four plants. We performed all pairwise alignments of the annotated miRNAs and top 100 novel miRNA candidates from each organism. We then visualized the alignments in a network using Cytoscape, showing interactions of perfect homology and single base pair mismatches or indels (Figure 4-7, examples shown in Figure 4-8A and 4-8B). Many top candidates and known miRNAs are homologous to each other. Of the top 100 novel soybean candidates, 21 were homologous to novel and known miRNA sequences found in at least one of the four plants (Figure 4-8C). In contrast, the bottom 100 soybean candidates, that is, those predicted to not be miRNAs, have no homology to any known miRNAs or top scoring candidates.

Experimental Validation of Predicted Soybean miRNAs

To validate our classifier experimentally, we selected the top 13 predicted candidate miRNAs, and as a control, 14 with average prediction strength. We chose 14 average strength candidates that had read counts similar to those of the putative miRNAs to avoid expression biases leading to faulty confirmation of non-miRNAs. Results of experimental validation are shown in Table 4-2. Candidates are assigned a score from zero to one, with zero being least likely to be a true miRNA and one being most likely. This score corresponds to the classifier's predicted probability of being a miRNA using a uniform prior (Table 4-2, column 3). The choice of prior is unimportant since the scores are used to rank candidates, not to assign strict probabilities.

Using a different prior would preserve candidate rankings, and simply shift all scores up or down.

Candidate miRNAs were validated based on a number of criteria, including predicted stem-loop structure based on mfold secondary structure prediction, detection of putative miRNA and star strand sequences by stem-loop RT-PCR, detection of star strand sequence in small RNA sequence libraries, prediction of target mRNAs by SeqTar, anti-correlation between the abundance of a miRNA and the abundance of its target mRNA, and whether there is an annotated ortholog to the putative miRNA in other plant species.

None of these criteria alone is sufficient to conclusively determine if a sequence is or is not a miRNA. However, many of the results in our high scoring candidate miRNAs are distinct from those of our lower scoring miRNAs, suggesting that our method is successfully identifying true miRNAs.

We first sought to confirm whether the reads were detectable by an independent method that only tends to amplify short reads, even though this does not tell us whether the read is a miRNA or not. To accomplish this we used stem-loop RT-PCR (Table 4-2, column 7). Stem-loop RT-PCR allows for specific detection of short RNA sequences by using an oligo that forms a stem loop on one end of the amplification reaction. We found that 8/13 of our candidate high probability miRNAs and 10/14 of our lower probability miRNAs were detectable by stem-loop RT-PCR. This suggests, that as one might expect from the small RNA-seq data, both classes can be detected with nearly equal probability. The fact that we do not detect some of the sequences suggests that RNA-seq may have higher resolution the stem-loop RT-PCR.

We next asked whether sequences in the two groups contained stem loops, as predicted by mfold. We found that 7/13 of our putative miRNAs were predicted to form a stem-loop structure by mfold, compared to just 3/14 for the putative non-miRNAs (Table 4-2, column 6). It is not surprising that many of our putative miRNAs fail or that some of the putative non-miRNAs pass this test given the difficulty in accurately predicting RNA secondary structure, but the 2.5-fold enrichment in the putative miRNAs in encouraging.

Subsequently, we performed stem-loop RT-PCR on candidate miRNA* that had stem-loop structures predicted by mfold (Table 4-2, column 8). Of the candidates with predicted stem-loop structures, 4/7 of the high probability candidates and 2/3 of our lower probability candidates contained detectable miRNA*. The lack of detection of some start strands by RT-PCR could result from the fact that they are typically degraded far more rapidly than the mRNA itself. We also investigated whether the star strand sequence could be detected in small RNA sequencing libraries and found that all sequences detected by stem-loop RT-PCR were also in the sequencing library. In addition, one sequence from each group that was not detected by stem-loop RT-PCR was detected in the sequencing library (Table 4-2, column 9).

We next performed target prediction using the SeqTar software and a publicly available degradome library. Degradome libraries are generated by deep sequencing RNA fragments from the 5' end. This marks the location of miRNA binding due to the 5'-phosphate generated by miRNA target cleavage. This criterion performs the best at differentiating high scoring and low scoring miRNA candidates, with 7/13 of our high scoring miRNAs and 0/14 of our low scoring miRNAs having predicted mRNA targets (Table 4-2, column 10). For those with predicted targets, we performed RLM 5'-RACE to confirm degradation of the target, and found that 5/9 of the predicted targets showed degradation (Table 4-2, column 11).

Of the candidate miRNAs with low scores, miRC14, miRC15, and miRC16 had predicted stem-loop structures, detectable mature miRNA sequences, and detectable star strands by at least one method, with two of the three having star strands detected both by stem-loop RT-PCR and small RNA-seq. These represent three of the five 24 nucleotide putative non-miRNAs tested, which received low prediction scores primarily because of their length. We believe that these may be true miRNAs but were penalized strongly due to the bias against 24 nucleotide annotated miRNAs in the literature. As more 24 nucleotide miRNAs are discovered, the classifier will be better able to find this class of miRNAs.

Discussion

In this work we predicted putative novel miRNAs in several plant species using a Naïve Bayes classifier and characterized several candidates in soybean. Our experimental validation results show that our method is capable of detecting small RNAs that are likely to be true miRNAs. The results also suggest that our approach detects three 24 nucleotide small RNAs, a class that would be omitted by nearly all other methods for detecting miRNAs.

Traditional approaches to miRNA discovery are biased toward miRNAs with stereotypical 21 nucleotide length and predicted stem-loop structures. Computational approaches are explicit in disregarding potential miRNAs that lack these features, but many experimentalists use a similar process when deciding which potential miRNAs to confirm experimentally. Our approach instead uses a number of criteria that show distinctive distributions between known miRNAs and all other reads. These properties include the length of the read, the presence of a miRNA* sequence nearby, the distribution of additional reads around the putative miRNA, and the number of observations of each read and its mapping multiplicity. Using the known miRNAs from miRBase we have been able to train a naïve Bayes classifier to distinguish true miRNAs

from other types of reads that may represent degraded mRNAs or siRNAs, among other types. As a result, our approach is susceptible to the biases that have been used to identify known miRNAs in mirBase, but as each feature is used to assign weights, our approach does not eliminate any small RNAs from a small RNA sequence library that does not have the stereotypical length or stem loop structure, but rather assigns reads that deviate from this lower scores. This weighting allows these potential miRNAs to be retained for further analysis, such as mRNA target prediction. Thus, a miRNA that would be eliminated by most computational methods can be selected for validation based on a combination of prediction score and other metrics. For example, an experimentalist can choose to only validate putative miRNAs that target genes in a particular biological pathway. Very conservative approaches are unlikely to find any putative miRNAs that target mRNAs in such an instance, but our method will retain all small RNA sequences, so one can decide based on the candidate's score whether it is worth investigating further. Thus our classifier can be used as a hypothesis-generating method that allows users to determine whether or not to pursue potential candidates based on the candidate's score and the individual questions that the user wants to address. For example, a user might choose to only study candidates that are not found in other plant species or those not predicted to form stem-loop structures in order to better determine how these sequences hybridize with their miRNA*'s.

The criteria for annotation of plant miRNAs are not well defined, but predicted stem-loop structure and presence in a small RNA-Seq dataset have in the past been sufficient to annotate a read as a miRNA. Among our candidates, we find that those with predicted stem-loop structures are not more likely to have predicted targets (using SeqTar), but they are more likely to have detectable star strands by both stem-loop RT-PCR and small RNA sequencing. While this

indicates that predicted stem-loop structure is a powerful criterion for detecting miRNAs, there are several exceptions to this trend. For example, we find two candidates that have predicted stem-loop structures but cannot be found by stem-loop RT-PCR and lack detectable star strands. In our validation dataset, we find only a single sequence, miRC11, which passes all of our experimental tests. Of the sequences with lower prediction values, we find only three (miRC14, miRC15, and miRC16) that pass more than a single test and each of these has a predicted stem-loop structure, is found by stem-loop RT-PCR, and has a detectable star strand by at least one method. It is interesting to note that all three of the lower scoring candidates that pass multiple validation tests are 24 nucleotides. This suggests that 24 nucleotide miRNAs may be penalized too strongly by our method, suggesting that annotated miRNAs are not a representative sampling of the true miRNA population of this sequence length. We believe that this finding indicates that more effort should be placed into the discovery of 24 nucleotide miRNAs.

The homology among plant miRNAs and highly predicted candidates further indicates that there is strong enrichment for true miRNAs among our high probability candidates. It is worth noting, however, that while many miRNAs exist as members of large families conserved across species, most are found only in a single organism. This could be either because of a relatively high number of species-specific miRNAs or that many conserved miRNAs have not yet been annotated across plants. Similarly, most top soybean candidates had no homologous sequences in the plant species studied; however, most sequences with homology to at least one plant species had homology to all of them. It is also interesting that peach miRNAs seem to be members of very large families, which is true for both known miRNAs and the top predicted candidates. This is particularly surprising as peach has a relatively small genome compared to rice and

soybean, which lead us to speculate that it would have fewer identical mature miRNA sequences encoded in its genome.

In conclusion, we describe a method for ranking putative plant miRNAs from small RNA sequence datasets and provide experimental tests for validation of high and medium ranked soybean candidates. This study reveals some of the shortcomings of using stringent filter-based approaches to eliminate potentially interesting candidates and provides some examples of likely miRNAs that would typically be ignored by filter-based approaches. We also show potential pitfalls of using predicted stem-loop structures as the sole criteria for miRNA annotation, as candidates with stem-loop structures sometimes fail all or most other validation tests.

Methods

Datasets

Arabidopsis, rice, and peach small RNA sequencing datasets were downloaded from the Gene Expression Omnibus (GEO). The *Arabidopsis* sequence dataset was a combination of two experiments, one of whole flower tissue sequenced by 454 (GSM118372), and one from immature floral tissue sequenced by Illumina (GSM284747). The rice dataset is small RNA-seq from untreated Nipponbare leaves (GSE38480), and the peach data are time course small RNA-seq of flower buds after chilling to induce dormancy (GSE38535).

The soybean small RNA sequencing dataset was generated from early maturation stage soybean seeds (cv. Williams 82). Seeds were fixed, embedded in paraffin, sectioned, and placed on polyethylene napthalate (PEN)-membrane slides (Leica Microsystems). Several sections of whole seeds were scraped from slides, and RNA was isolated using the RNAqueous®-Micro Total RNA Isolation Kit (Ambion). A small RNA library was constructed according to the

procedures of the TruSeq Small RNA Preparation Kit (Illumina). RNA sequencing was performed with Illumina sequencing technology.

Known miRNA datasets for all four organisms were downloaded from miRBase (http://www.mirbase.org/). All known miRNAs from soybean, rice, and peach were used. Due to the greater available number of *Arabidopsis* miRNAs, these miRNAs were filtered to only keep miRNAs with two independent lines of experimental evidence (cloned, Northern blot, 454, Solexa, etc).

Arabidopsis thaliana complete genome sequence was downloaded from The *Arabidopsis* Information Resource (http://www.arabidopsis.org/). *Glycine max* complete genome sequence was downloaded from SoyBase (http://soybase.org/). *Oryza sativa* and *Prunus persica* complete genome sequences were both downloaded from the Plant Genome Database (http://www.plantgdb.org/).

Read Mapping

Small RNA sequence reads and known miRNAs were aligned to their respective genomes using the Bowtie software package (32). All known miRNAs and small sequence reads in *Arabidopsis*, rice, and peach were allowed a maximum of two mismatches to their respective genomes. Only the alignments with the fewest mismatches for each sequence were kept, allowing ties. We required zero mismatches in soybean sequence alignments to facilitate easier validation.

Entropy Calculation

Shannon entropy was computed for each mapped position of each sequence read in all datasets using the following formula:

$$-\sum_{i=1}^{k} p_i \ln p_i$$

Here, k is the number of positions in the genome interrogated in each test. In all of our datasets, k=101, and encompasses the position the read aligns to, 50 nucleotides upstream, and 50 nucleotides downstream. p_i is the fraction of observed counts at position *i* over the total counts in the 101 base pair region. We assigned a single read in a region an undetermined entropy value because entropy measures the spread of aligned reads across a genomic region, and a single read in a window cannot have variable entropy. Therefore, we excluded entropy values from the posterior probability calculation of these candidates.

miRNA* Detection

Presence or absence of miRNA* was determined based on parameters from the literature (20) using in-house scripts. Briefly, a sequence at least 10 nucleotides and no more than 100 nucleotides away from the candidate miRNA must be able to base pair with at most 4 mismatches and at most 1 bulge of length at most 2, with no penalty for G-U base pairs. Ability to base pair was computed using MATLAB's (Mathworks®) "nwalign" function.

Naïve Bayes Classifier

The Naïve Bayes Classifier is a probabilistic model that classifies based on Bayes Theorem, assuming independence between all variables. The Naïve Bayes probability is computed using the formula:

$$p(C \mid F_1, \ldots, F_n) = \frac{1}{Z} p(C) \prod_{i=1}^{n} p(F_i \mid C)$$

Here, $p(C|F_1,\ldots,F_n)$ is the probability that a given miRNA candidate belongs to class C (either "miRNA" or "not miRNA") given its values for each classification variable. p(C) is the prior probability of a candidate belonging to a given class in the absence of any evidence. We used a uniform prior (p(miRNA)=p(not miRNA)=0.5) because we used the classifier to rank candidates, not to compute absolute probabilities.

The classifier was trained independently for each organism using previously described miRNAs downloaded from miRBase as true positives and small RNA-Seq reads as true negatives. Variables were computed for each putative and known miRNA and distributions for true positive and negatives were generated. Counts and entropy values were discretized into bins as shown in Figures 4-2, 4-3, 4-4, and 4-5 to more accurately reflect underlying distributions.

Homology

Homology between sequence datasets was computed by aligning the top candidate miRNAs and known miRNAs from each organism against one another using MATLAB's (Mathworks®) "nwalign" function. Sequences were counted as being perfectly homologous if they were identical and imperfectly homologous if they had a single mismatch or single indel of length one. Networks of homology were visualized using Cytoscape (33).

miRNA and Star Strand RNA Detection

miRNAs and their star strands were detected by stem-loop reverse transcriptase- polymerase chain reaction (RT-PCR) (34). Total RNA was isolated from soybean seeds using Plant RNA Purification Reagent (Invitrogen). Stem-loop RT primers were designed with six-nucleotide extensions at their 3'-ends that were specific to the six nucleotides at the 3'-end of the miRNA or star strand. Reverse transcription was performed for 30 minutes at 16◦C followed by 60 cycles of

30◦C for 30 seconds, 42◦C for 30 seconds, and 50◦C for 1 seconds, and the reaction was inactivated at 85◦C for 5 minutes. PCR was then performed using miRNA or star strand-specific primers, excluding the last six nucleotides at the 3'-end of the miRNA or star strand, and universal reverse primers. Reactions were analyzed by gel electrophoresis on 3% agarose.

Stem-loop structure prediction

The stem-loop structure of each primary miRNA was predicted using the mfold web server (http://mfold.rna.albany.edu/?q=mfold) using default parameters. A candidate was determined to contain a stem-loop structure if any of the structures predicted by mfold contained a stem-loop that contained base pairing between the putative mature miRNA and its star strand.

mRNA target prediction

Potential mRNA targets of candidate miRNAs were predicted using the SeqTar program (35) and a public degradome library (GEO accession number GSM848963) for cotyledons of soybean seed at the early maturation stage. Target mRNA detected in the degradome analysis were tested experimentally by RNA ligase-modified 5' RNA amplification of cDNA ends (RLM 5'-RACE) (36). Total RNA was ligated to a 5'-RACE adaptor, and a poly (dT) oligonucleotide was used for cDNA synthesis. The first round of PCR was carried out using a primer corresponding to the 5'-RACE adaptor and a gene-specific primer. Nested PCR was performed using 1/25 of the first PCR reaction, a nested 5'-RACE primer, and a nested gene-specific primer. The PCR product was gel-eluted and sequenced.

Figures

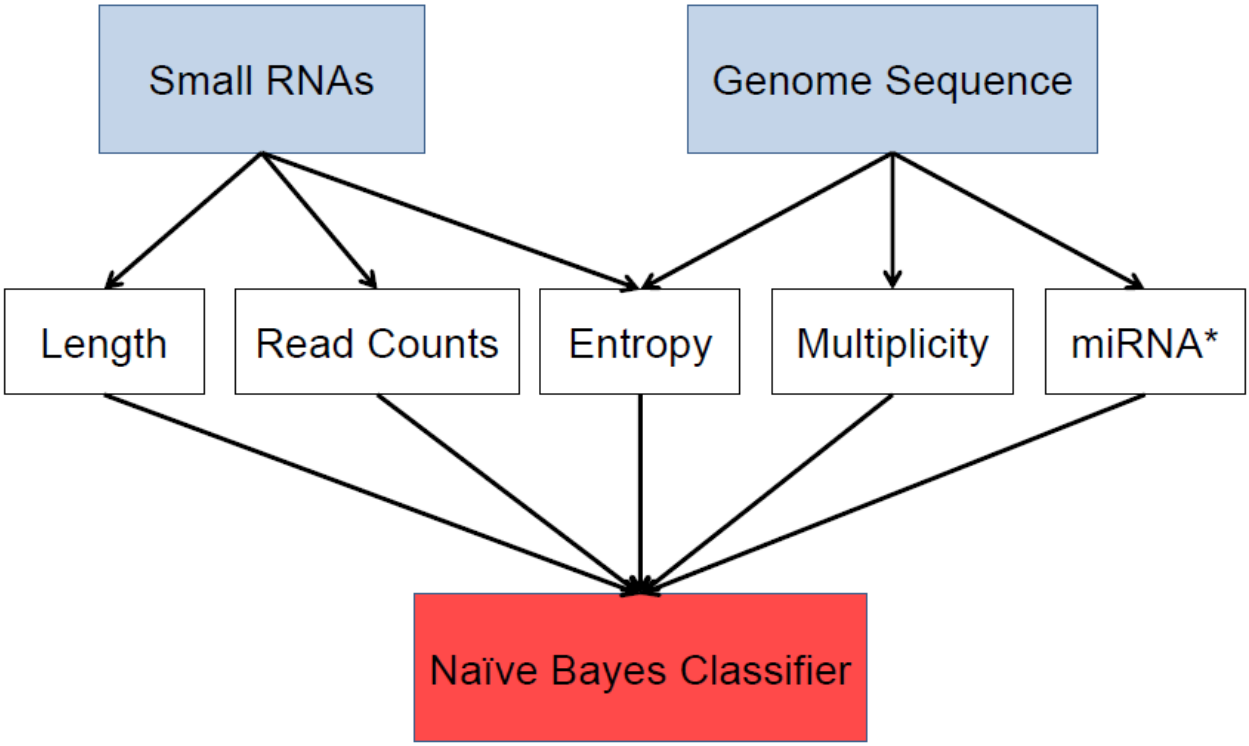

Figure 4-1. Workflow of the Naïve Bayes Classifier. The Naïve Bayes Classifier draws its information for each putative mature miRNA sequence from five criteria: length, read counts in sequencing library, entropy, the number of locations the sequence maps to in its respective genome (multiplicity), and presence of a detectable miRNA* sequence. A small RNA sequence library, but not a genome sequence, is needed In order to determine a putative miRNA's length and read counts, a genome sequence, but no small RNA sequence data, is needed to determine multiplicity and miRNA* presence, while both a small RNA sequence library and genome sequence are required to compute entropy.

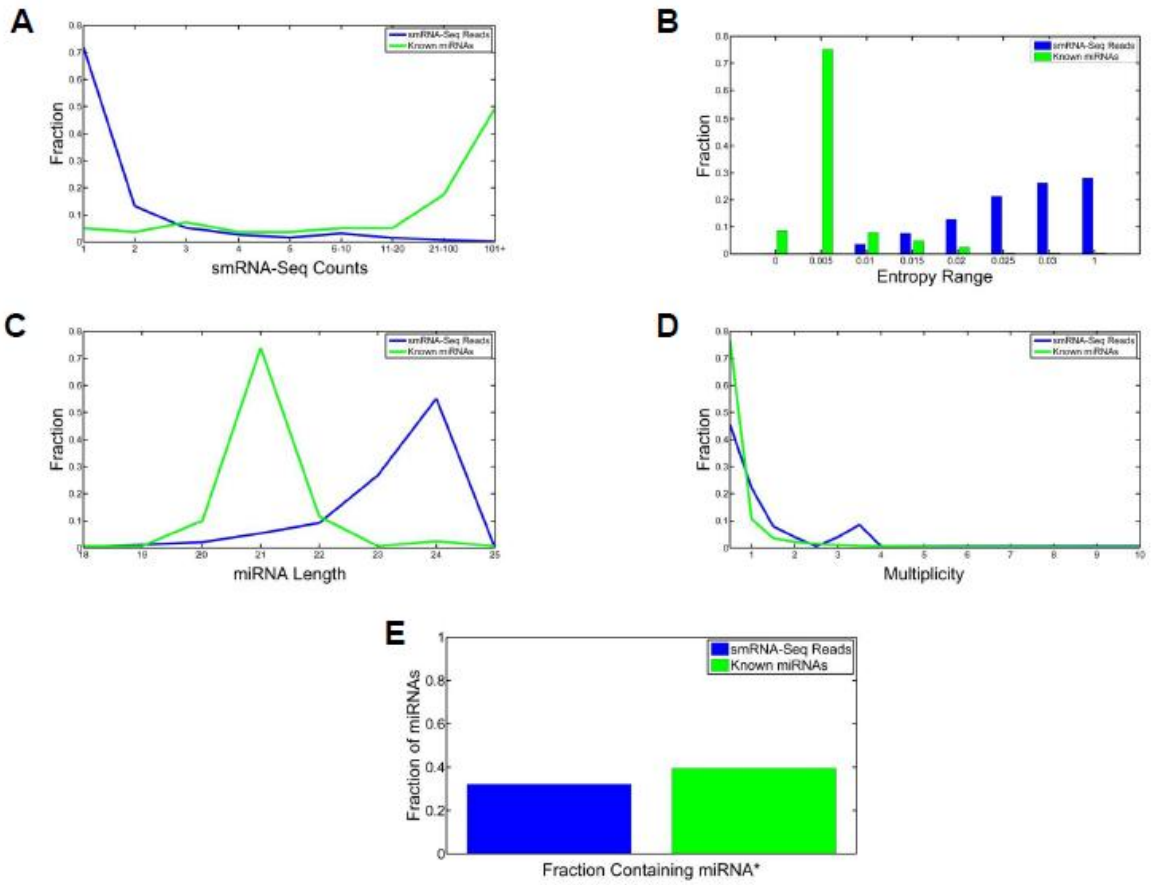

Figure 4-2. Distributions of NBC variables in soybean. All small RNA-Seq reads (blue) are taken to be true negatives, known miRNAs (green) as true positives. (A) Sequence counts for both groups, (B) Entropy, (C) miRNA length, (D) multiplicity, and (E) presence of miRNA*.

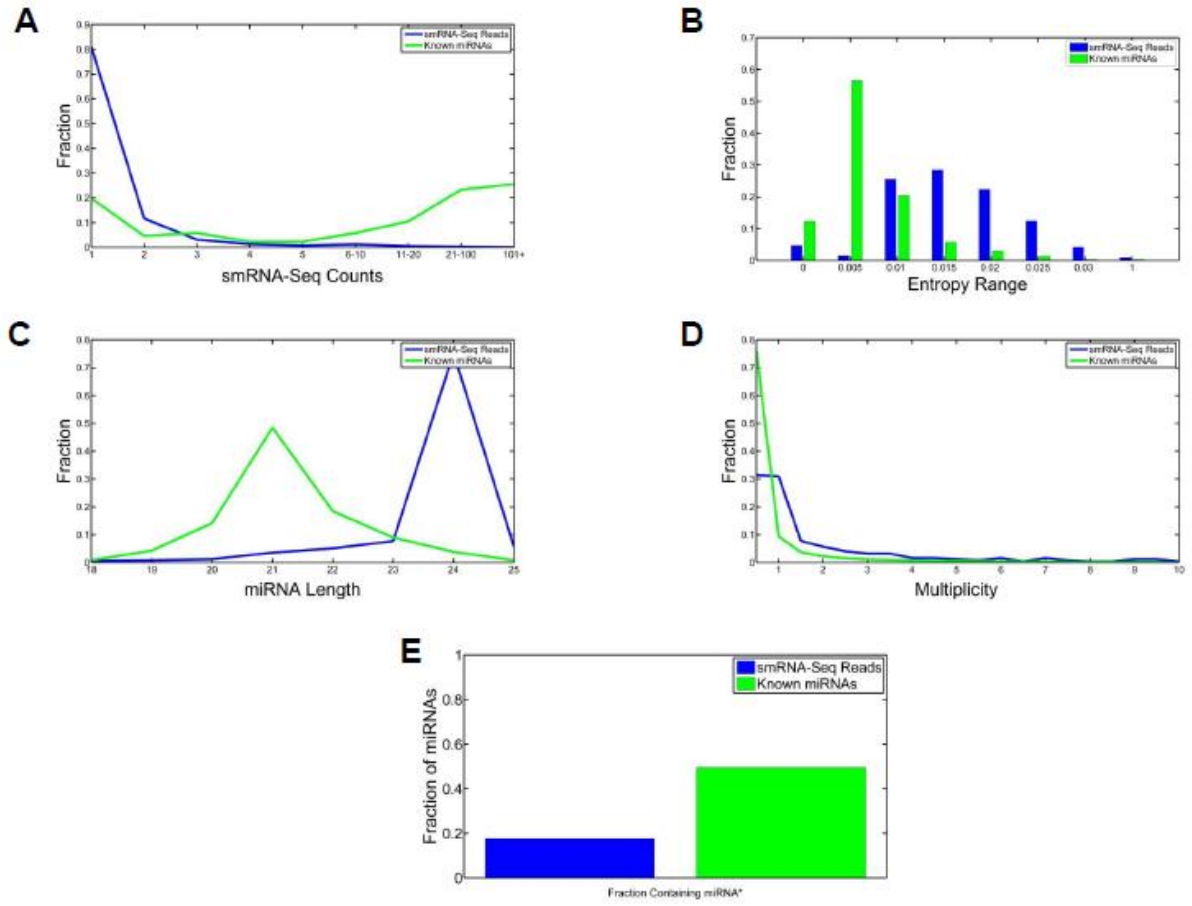

Figure 4-3. Distributions of NBC variables in *Arabidopsis*. All small RNA-Seq reads (blue) are taken to be true negatives, known miRNAs (green) as true positives. (A) Sequence counts for both groups, (B) Entropy, (C) miRNA length, (D) multiplicity, and (E) presence of miRNA*.

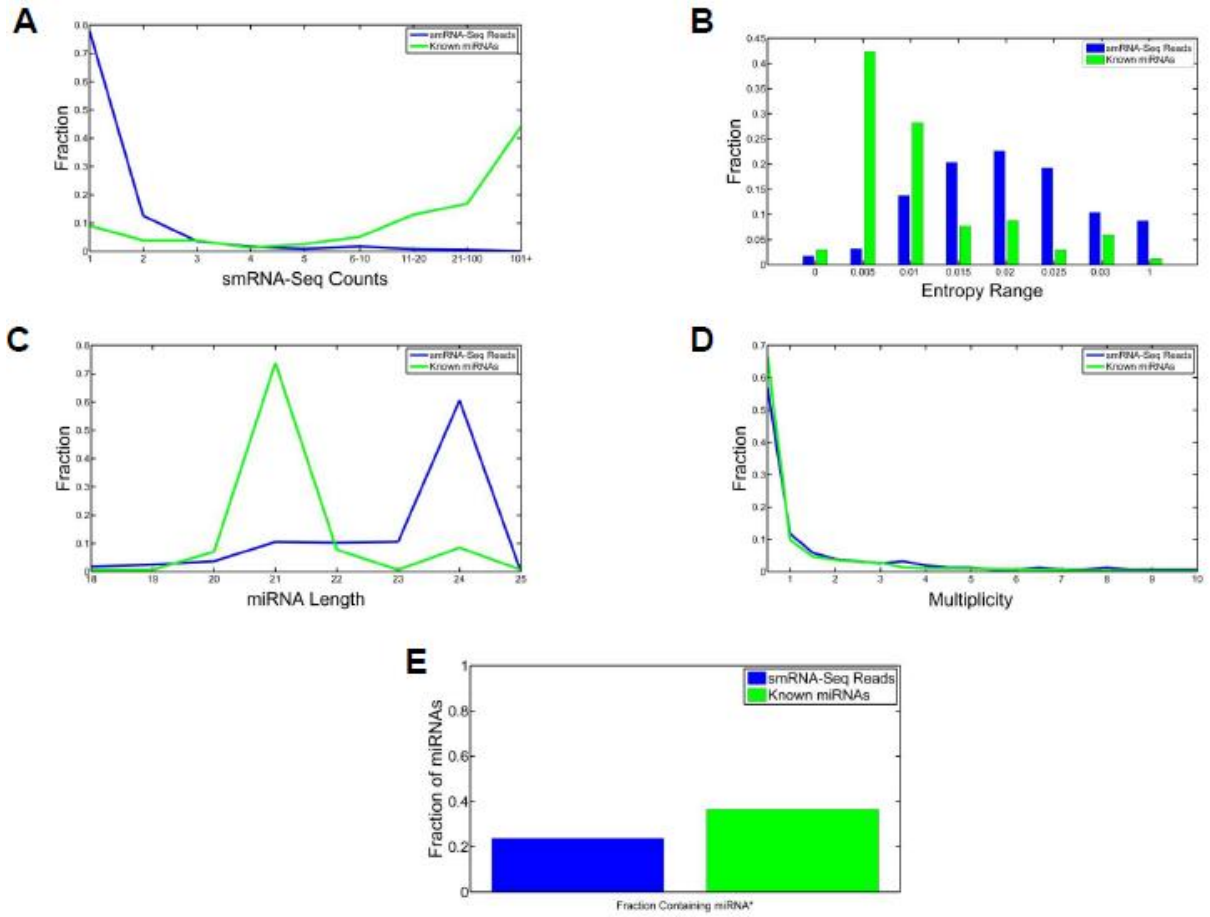

Figure 4-4. Distributions of NBC variables in rice. All small RNA-Seq reads (blue) are taken to be true negatives, known miRNAs (green) as true positives. (A) Sequence counts for both groups, (B) Entropy, (C) miRNA length, (D) multiplicity, and (E) presence of miRNA*.

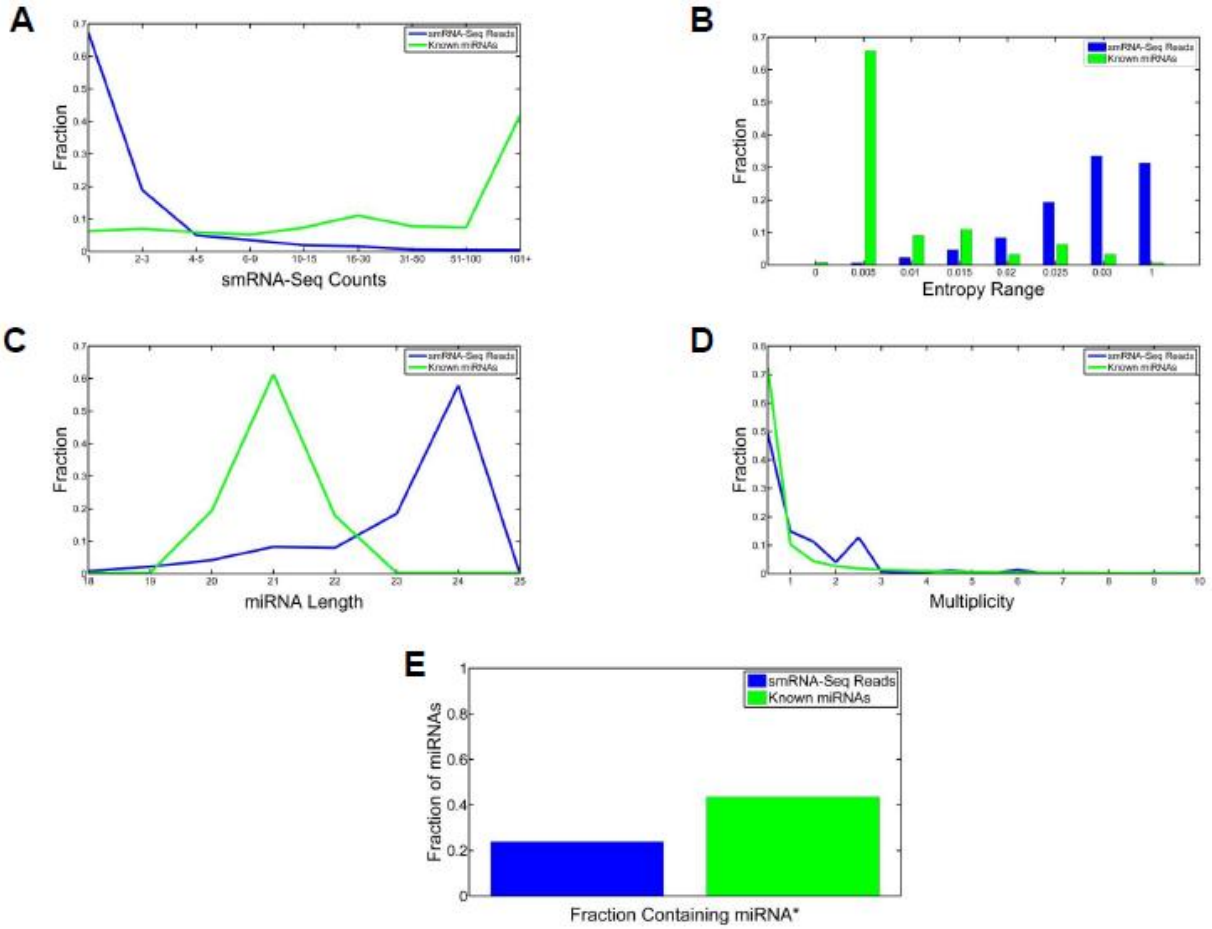

Figure 4-5. Distributions of NBC variables in peach. All small RNA-Seq reads (blue) are taken to be true negatives, known miRNAs (green) as true positives. (A) Sequence counts for both groups, (B) Entropy, (C) miRNA length, (D) multiplicity, and (E) presence of miRNA*.

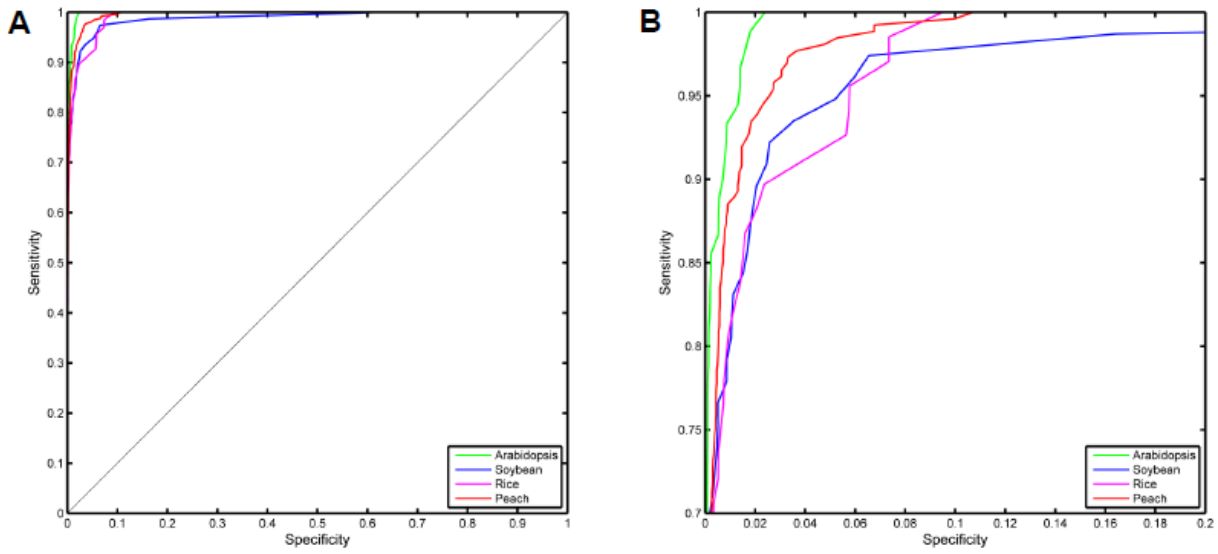

Figure 4-6. ROC curves of NBC. (A) Sensitivity vs specificity in four tested organisms. Sensitivity refers to the fraction of true positives (previously described known miRNAs) that are called as miRNAs by our classifier. Specificity refers to the fraction of true negatives (approximated by all small RNA-Seq reads) that are called as "not miRNAs" by our classifier. (B) Curves from Part A zoom in to show detail.

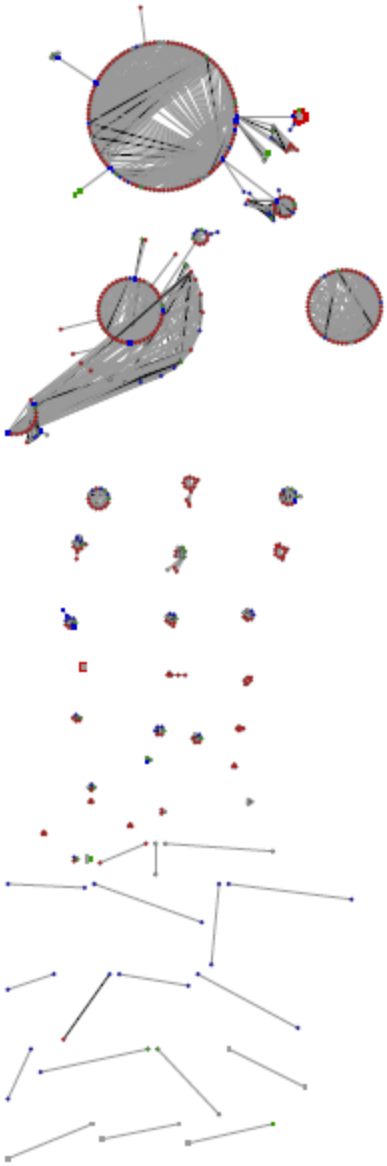

Figure 4-7. Network of homologous known and putative novel miRNAs in plants. Black lines indicate identical known or candidate miRNAs, grey lines indicate homology with at most 1 mismatch or indel.

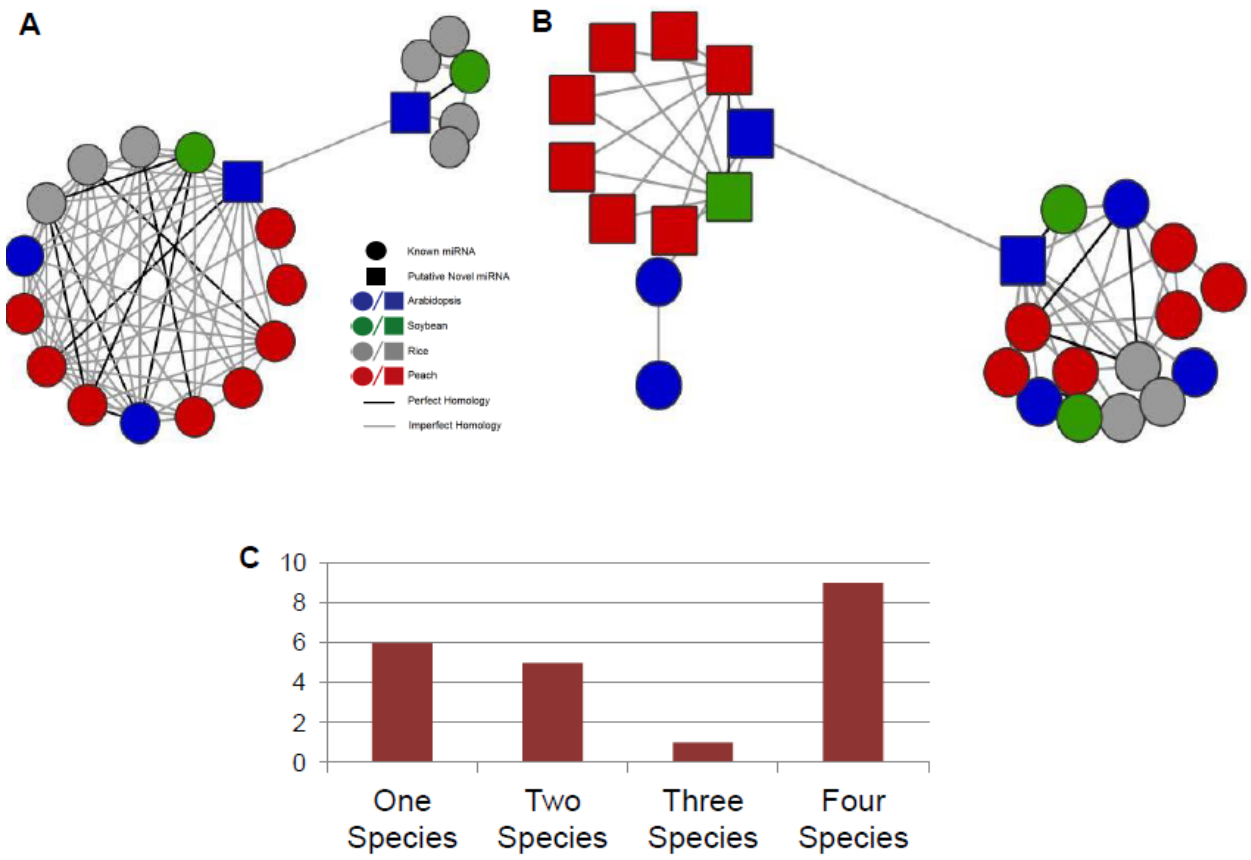

Figure 4-8. Homology of known and predicted miRNAs. (A & B) Example sub-networks of homologous known and putative novel miRNAs. (C) Distribution of homology of top 100 novel soybean candidates. 21 of 100 showed homology to at least one tested organism, the remainder are not shown.

Tables

miRNA	miRNA candidate sequence	miRNA score	Length	miRNA Reads (CPM)	Stem-loop Prediction	miRNA detection (Stem-loop RT-PCR)	Star strand detection (Stem-loop RT-PCR)	Star strand in libraries	Potential target (Reads in degradome library)	Target detection (RLM 5'-RACE)
miRC1	GGAATGGGCTGATTGGGAAG	1	20	1989	+	+	+	+	–	ND[a]
miRC2	GATGGGGAAGGGGGGCACATG	1	21	297	+	–	–	–	–	ND
miRC3	GGCATTCAGAATGAGTAGGA	1	20	738	–	–	---	---	Glyma10g39150 (9333)	–

miRC4	GGTGGC TGTAGTT TAGTGGT	1	20	580	−	+	---	---	Glyma04g0 8700 (17)	−
									Glyma08g4 1630 (44)	+
miRC5	GGGGGT GTAGCTC ATATGGT A	1	21	99	−	+	---	---	Glyma20g2 9820 (8)	−
									Glyma10g3 7980 (8)	
miRC6	GGGAAT GAAGCC TGGTCCG AA	1	21	2340	−	−	---	---	−	ND
miRC7	GGCATTC AGAATG AGTAGG AG	1	21	524	−	−	---	---	−	ND
miRC8	GGGGAT GTAGCTC AAATGG T	1	21	725	−	+	---	---	−	ND
miRC9	CCCGCCT TGCATCA ACTGAA T	1	21	135	+	+	+	+	Glyma01g2 4670 (47)	+
miRC10	GAGGAA TGAAGC CTGGTCC GA	1	21	1861	+	−	−	−	Glyma18g0 1860 (41)	+
miRC11	TGGGAA TGGGCT GATTGG GA	1	20	488	+	+	+	+	Glyma08g2 3050 (49)	+
miRC12	TTTCGGT GTCGGT GAATTG CC	1	21	28	+	+	−	+	Glyma20g3 5990 (47)	−
									Glyma12g1 4000 (21)	+
miRC13	CCTAGCT CCTGAA CCATCAC TTT	0.9	23	50	+	+	+	+	−	ND
miRC14	AGATGG TTGAGG AGCGTG AGAAGG	0.5	24	1123	+	+	−	+	−	ND
miRC15	TGGTTGA GGAGCG TGAGAA GGATT	0.5	24	241	+	+	+	+	−	ND

miRC16	CGAAGA TGAGGT CGACCA TGTGAC	0.3	24	117	+	+	+	+	−	ND
miRC17	TGCTGTT GGCCTC AATGAT CAGT	0.7	23	52	−	+	---	---	−	ND
miRC18	ATGAAT GAACAT GTTTCTG AGCTCT	0.7	25	2521	−	+	---	---	−	ND
miRC19	CCTAGCT CCTGAA CCATCAC TTTTT	0.7	25	463	−	+	---	---	−	ND
miRC20	TATTCTG GTGTCCT AGGCGT AGAGG	0.7	25	233	−	−	---	---	−	ND
miRC21	GCGAAT TTGTTGT TGGGCT ACAATT	0.7	25	135	−	+	---	---	−	ND
miRC22	TTTGTAT TAGCTCT ATCTGAT CATT	0.7	25	107	−	+	---	---	−	ND
miRC23	AAGGAG GGACTA GTGCTAT GGCT	0.6	23	30	−	−	---	---	−	ND
miRC24	TATCAA GCTCCTG AACCAT CATTTT	0.6	25	161	−	+	---	---	−	ND
miRC25	TGGTCGC ACGGTT GTCTGAC AGACC	0.6	25	106	−	+	---	---	−	ND
miRC26	TAGTACT AGGATG GGTGAT CTCCT	0.4	24	237	−	−	---	---	−	ND
miRC27	AATATA ACGCGT CGCCACT GGTGA	0.4	24	105	−	−	---	---	−	ND

Table 4-1. Summary of validation results. Potential target was predicted from a public degradome library (GSM848963) using SeqTar. Anti-correlation (Pearson's) between miRNA and its target was carried out

using the CPM of miRNA and target in different compartments of soybean seed at early maturation stage and a home made script. [a]Not determined.

	Top 100	Top 250	Top 500	Top 1000
Arabidopsis	30	42	55	67
Soybean	20	32	39	54
Rice	12	24	29	40
Peach	36	47	51	63

Table 4-2. Known miRNAs found in top predictions of the NBC. The top candidates from each organism show strong enrichment for known miRNAs. This enrichment is stronger as the number of top candidates decreases.

References

1. He L, Hannon GJ. MicroRNAs: small RNAs with a big role in gene regulation. *Nature* 5 (7): 3. 522–531.

2. Brodersen P, Sakvarelidze-Achard L, Bruun-Rasmussen M, Dunoyer P, Yamamoto YY, Sieburth L, Voinnet O. Widespread translational inhibition by plant miRNAs and siRNAs. *Science* 320 (5880): 1185–90.

3. Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, Kim VN. MicroRNA genes are transcribed by RNA polymerase II. *Embo J* 2004, 23(20):4051-4060.

4. Cai X, Hagedorn CH, Cullen BR. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA* 2004, 10(12):1957-1966.

5. Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S, Kim VN. The nuclear RNase III Drosha initiates micro-RNA processing. *Nature* 2003, 425(6956):415-419.

6. Chendrimada TP, Gregory RI, Kumaraswamy E, Norman J, Cooch N, Nishikura K, Shiekhattar R. TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature* 2005, 436(7051):740-744.

7. Schwarz DS, Hutvagner G, Du T, Xu Z, Aronin N, Zamore PD. Asymmetry in the assembly of the RNAi enzyme complex. *Cell* 2003, 115(2):199-208.

8. Xue C, Li F, He L, Liu G, Li Y, Zhang X. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics* 2005, 6: 310.

9. Yousef M, Nebozhyn M, Shatkay H, Kanterakis S, Showe LC, Showe MK. Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. *Bioinformatics* 2006, 22 (11): 1325-1334.

10. Pfeffer S, Sewer A, Lagos-Quintana M, Sheridan R, Sander C, Grässer FA, van Dyk LF, Ho CK, Shuman S, Chien M, Russo JJ, Ju J, Randall G, Lindenbach BD, Rice CM, Simon V, Ho DD, Zavolan M, Tuschl T. Identification of microRNAs of the herpesvirus family. *Nature Methods* 2005, 2(4): 269-76.

11. Kadri S, Hinman V, Benos PV. HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models. *BMC Bioinformatics* 2009, 10:S35.

12. Lai EC, Tomancak P, Williams RW, Rubin GM. Computational identification of Drosophila microRNA genes. *Genome Biology* 2003, 4(7): R42.

13. Legendre M, Lambert A, Gautheret D. Profile-based detection of microRNA precursors in animal genomes. *Bioinformatics* 2005, 1;21(7): 841-5.

14. Tao M. Thermodynamic and structural consensus principle predicts mature miRNA location and structure, categorizes conserved interspecies miRNA subgroups, and hints new possible mechanisms of miRNA maturization. *ARXIV* 2007, eprint arXiv:0710.4181.

15. Sheng Y, Engström PG, Lenhard B. Mammalian MicroRNA Prediction through a Support Vector Machine Model of Sequence and Structure. *PLoS ONE* 2007, 9: e946.

16. Grad Y, Aach J, Hayes GD, Reinhart BJ, Church GM, Ruvkun G, Kim J. Computational and experimental identification of C. elegans microRNAs. *Molecular Cell* 2003, 11(5): 1253-63.

17. Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP. The microRNAs of Caenorhabditis elegans. *Genes & Development* 2003, 15;17(8): 991-1008.

18. Terai G, Komori T, Asai K, Kin T. miRRim: A novel system to find conserved miRNAs with high sensitivity and specificity. *RNA* 2007, 13: 2081-2090.

19. Shabalina SA, Koonin EV. Origins and evolution of eukaryotic RNA interference. *Trends in Ecology and Evolution* 2008, 10 (10): 578–87.

20. Meyers BC, Axtell MJ, Bartel B, Bartel DP, Baulcombe D, Bowman JL, Cao X, Carrington JC, Chen X, Green PJ, Griffiths-Jones S, Jacobsen SE, Mallory AC, Martienssen RA, Poethig RS, Qi Y, Vaucheret H, Voinnet O, Watanabe Y, Weigel D, Zhui, JK. Criteria for Annotation of Plant MicroRNAs.

21. Berezikov E, Cuppen E, Plasterk RHA. Approaches to microRNA discovery. *Nature Genetics* 2006, 38: S2-S7.

22. Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, Barzilai A, Einat P, Einav U, Meiri E, Sharon E, Spector Y, Bentwich Z. Identification of hundreds of conserved and nonconserved human microRNAs. *Nature Genetics* 2005, 37(7): 766-70.

23. Sewer A, Paul N, Landgraf P, Aravin A, Pfeffer S, Brownstein MJ, Tuschl T, van Nimwegen E, Zavolan M. Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics* 2005, 6:267.

24. Leung,Y.Y., Ryvkin,P., Ungar,L., Gregory,B.D., and Wang, L.-S. CoRAL: Predicting non-coding RNAs from small RNA-sequencing. *Nucleic acids research* 2013, 41(14): e137.

25. Axtell MJ, Jan C, Rajagopalan R, Bartel DP. A two-hit trigger for siRNA biogenesis in plants. Cell 2006 Nov 3;127(3):565-77

26. Gregory BD, O'Malley RC, Lister R, Urich MA et al. A link between RNA metabolism and silencing affecting Arabidopsis development. Dev Cell 2008 Jun;14(6):854-66.

27. Chodavarapu RK, Feng S, Ding B, Simon SA et al. Transcriptome and methylome interactions in rice hybrids. Proc Natl Acad Sci U S A 2012 Jul 24;109(30):12040-5.

28. Matzke M, Kanno T, Daxinger L, Huettel B, Matzke AJ. RNA-mediated chromatin-based silencing in plants. *Curr Opin Cell Biol* 2009, 21(3):367-76.

29. Hong Zhu, Rui Xia, Christopher Dardick, Callahan, Yong-qiang An, Zongrang Liu (2012). Unique expression, processing regulation, and regulatory network of peach (Prunus persica) miRNAs. BMC Plant Biology 2012, 12:149. doi:10.1186/1471-2229-12-149.
30Subramanian S, Fu Y, Sunkar R, Barbazuk W, Zhu J, Yu O. Novel and nodulation-regulated microRNAs in soybean roots. *BMC Genomics* 2008, 9:160

31. Rivas E, Eddy SR. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*. 2000; 16(7):583–605.

32. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 2009, 10(3):R25.

33. Smoot M, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 2011, 27(3).

34. Varkonyi-Gasic E, Hellens RP. (2011) Quantitative stem-loop RT-PCR for detection of microRNAs. Methods Mol Biol. 744:145-157.

35. Zheng Y, Li YF, Sunkar R, Zhang W. (2012) SeqTar: an effective method for identifying microRNA guided cleavage sites from degradome of polyadenylated transcripts in plants. *Nucleic Acids Res*. 40(4):e28.

36. German MA, Pillay M, Jeong DH, Hetawal A, Luo S, Janardhanan P, Kannan V, Rymarquis LA, Nobuta K, German R, De Paoli E, Lu C, Schroth G, Meyers BC,Green PJ. (2008) Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nat Biotechnol*. 26(8):941-946.

Chapter 5: Concluding Remarks

This dissertation explored gene expression and some of the many mechanisms of gene regulation. It showed the power of next-generation sequencing for determining differential gene expression between cell types and conditions, for uncovering very low frequency alternative splice products, and for discovering novel microRNAs. Together, these results show the complexity of gene expression and regulation, and how modern technologies can help elucidate fundamental processes required for eukaryotic life.

The first part of this dissertation focused on gene expression in the mating loci of *Volvox carteri*. It described the evolutionary event that differentiated it from the closely related unicellular green alga *Chlamydomonas reinhardtii* with two mating types to the multicelluar alga with sexually dimorphic gametes. It also showed several sex-regulated and gender-specific transcripts, although like the human X chromosome, most were neither sex related nor gender-specific. We also demonstrate that while most of the alga's genome diverges based on geographic distance between individuals over gender, within the mating loci, male algae are genetically more similar to male algae and female algae more similar to female algae. We also show gender-specific splicing of MAT3, an important gene conserved tumor suppressor homolog.

Since the completion of this project, MAT3 has been shown to have quickly diverged from closely related colonial volvocine algae as well as the unicellular *Chlamydomonas reinhardtii*. The work indicates that the gender-specific divergence of MAT3 occurred after the ancestors of Volvox split off from the other volvocine algae.

The second part of this dissertation showed alternative splicing in the yeast *Saccharomyces cerevisiae*. It was previously thought that *Saccharomyces* did not alternatively splice its intron-

containing genes, but we showed that rare events occur in which the splicing machinery picks a sequence similar to the canonical splice site and creates a different splice product. Many of these products contain premature termination codons, and are degraded by the nonsense-mediated decay pathway, but our findings give insight into how alternative splicing may have evolved. We also found that stress conditions such as high temperature or nutrient starvation can induce higher rates of these alternative splice products and disabling the decay machinery can lead to a higher proportion of these events compared to the canonical splicing event. Together, our results show that alternative splicing is frequency in *Saccharomyces cerevisise*, but many of these transcripts are rapidly degraded and that alternative splicing is used for control transcript levels, not increasing the number of peptides produced by a single gene.

Since this project was completed, there has been increased interest in alternative splicing in *Saccharomyces cerevisiae*. Presently, there is a study looking at if defects in histone modifying enzymes cause an increase in alternative splicing. *Saccharomyces cerevisiae* is also a model organism for studying basic splicing mechanisms since the rate of alternative splicing is so low, it allows splicing to be studied in a relatively simple splicing landscape. If research into *Saccharomyces* alternative splicing continues, it would allow *Saccharomyces* to be used for basic alternative splicing research as well.

The third part of this dissertation focused on identifying novel miRNAs in plants. We showed the shortcomings of existing techniques for combining computational and experimental approaches for discovering new miRNAs, and applied our method to four plants. We showed experimental validation of soybean candidates. We found that our Bayesian approach successfully discriminated true miRNAs from contaminants, and we found several putative novel miRNAs, including three 24 nucleotide candidates, which is a class that is almost universally

eliminated by other methods. We also show how high quality annotated miRNAs increase the performance of the classifier, showing the importance of not mis-annotating miRNAs in the literature.

Together, these projects show the power and diverse applications of next-generation sequencing technology when applied to transcriptomics. We showed transcriptional regulation from organisms ranging from algae to fungi and plants. We showed how gene expression can change in the relatively small mating loci of *Volvox*, how alternative splicing may have evolved using *Saccharomyces cerevisiae* as a model, and how additional miRNAs can be discovered in plants. It is exciting to see how the field of transcriptional regulation will continue to evolve in the future.