Genome Duplication and Alternative Splicing: Gateways to functional diversity

Inaugural - Dissertation

zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von
Dale Newton Richardson III
aus Illinois

Köln, 2010

Berichterstatter: Prof. Dr. Thomas Wiehe
Prof. Dr. Peter Nürnberg

Tag der letzten mündlichen Prüfung: November 2010

Zusammenfassung

Eines der Ziele dieser Dissertation ist es die Rolle zweier fundamentaler biologischer Prozesse, die Genomduplikation und das alternative Spleißen, in der Regulation der Genexpression zu verstehen. Die Genomduplikation und das alternative Spleißen haben tiefgreifende Auswirkungen auf die Genregulation, wie zum Beispiel dass die kontrollierte Expression duplizierter Gene die Evolution von Genomen beeinflusst, während das alternative Spleißen regulatorischer Gene enorme Auswirkungen auf die Funktionalität nahezu aller expressionierten Gene hat. Die Gesamtgenom-Duplikation (WGD) hat die Entstehung neuer Spezies, die Formation von Genen mit neuen Funktionen, oder auch die Modifizierung von Expressionsmustern beschleunigt und Organismen eine Form genetischer Robustheit verliehen.

Wir haben die Langzeit-Evolution und das Zusammenspiel von 5' "upstream" regulatorischen Sequenzen (URSs), Protein-kodierenden Sequenzen (CDSs) und Expressionskorrelationen (EC) von duplizierten Gen-Paaren im Modellorganismus *Arabidopsis thaliana* untersucht. Drei verschiedene Methoden haben eine signifikante evolutionäre Konservierung zwischen paralogen URSs verdeutlicht und waren mit Microarray-basierten Expressions-korrelationen der betreffenden Gen-Paare hoch korreliert. Die positionale Information von genauen zwischensequenzlichen Übereinstimmungen hat den Beitrag mikro-chromosomaler Neuordnungen für die Expressionsdivergenz demonstriert. Eine Drei-Wege Ranganalyse der URS-Similarität, der CDS-Divergenz und der EC haben spezifische Genfunktionale Ver-zerrungen aufgezeigt. Transkriptionsfaktoraktivität wurde mit Gen-Paaren, die konservierte URSs und divergente CDSs aufweisen, assoziiert, während eine große Anzahl metabolischer Enzyme mit Gen-Paaren, die sich durch divergente URSs und konservierte CDSs auszeichnen, in Verbindung gebracht werden konnten. Bemerkenswerterweise wird die Mehrheit an duplizierten Genen in den verschiedenen Entwicklungsstadien von *Arabidopsis thaliana* unterschiedlich expressioniert, was darauf hindeutet, dass oft eine der beiden Genkopien bevorzugt wird, und dass der

Mechansimus der Subfunktionalisierung für die Genregulierung eine Rolle spielen könnte.

Zusammen mit der WGD ist das alternative Spleißen (AS) der pre-mRNA ein fundamentaler molekularer Prozess, der genetische Diversität im Transkriptom und Proteom verursacht. Zahlreiche Komponenten, wie Länge und Sequenz der Exons und Introns, Trans-Faktoren und Transkriptionsraten, beeinflussen die Spleißreaktion. SR-Proteine, eine Familie von Spleiß-Regulatoren mit einem oder zwei RNA-Erkennungsmotiven (RRMs) am N-Terminus und einem "arg/ser-rich" am C-Terminus, wirken sowohl beim konstitutiven als auch beim alternativen Spleißen.

Wir haben Datenbanksuchen für SR-Proteine 27 eukaryotischer Spezies durchgeführt, die die Taxone der Pflanzen, Tiere, Fungis und basalen Eukaryonten, die außerhalb dieser Abstammungslinien liegen, umfasst. Mithilfe von RRMs als phylogenetische Marker haben wir mindestens 12 SR-Protein-Subfamilien feststellen können, von denen vier in Pflanzen weit verbreitet sind. Zudem befinden sich RRMs innerhalb der Subfamilien von SR-Proteinen an hoch konservierten Positionen, jedoch sind ihre vorhergesagten RNA-Bindungsresiduen degeneriert. Damit einhergehend stellten wir fest, dass die Mehrheit pflanzlicher SR Gene unter purifizierender Selektion steht. Darüberhinaus ist die Mehrheit an paralogen SR-Genen in Arabidopsis und Reis in den diversen Entwicklungsstadien unterschiedlich expressioniert, was mit unserer Beobachtung bezüglich duplizierter Gene im Einklang steht. Wir haben das Ausmaß an SR-Gen betreffendes AS unter der Verwendung von Spleiß-Graphen, die auf multiple "alignments" von ESTs/cDNAs und SR-genomischen Sequenzen beruhen, abgeschätzt. Das AS von SR-Genen ist ein weit verbreitetes Phänomen über zahlreiche Abstammungslinien und ein häufiges Merkmal unter Eukaryonten. Zudem variiert die Art der Ausführung des AS unter Organismen und SR-Subfamilien. Abschließend suggerieren wir einen Zusammenhang zwischen der DNA-Methylation innerhalb kodierender Regionen von SR-Genen und deren Spleißmuster.

Abstract

One of the goals of this dissertation is to understand how two fundamental biological processes, genome duplication and alternative splicing, factor into the regulation of gene expression. Genome duplication and alternative splicing have profound implications on gene regulation, as the controlled expression of duplicated genes affects the evolution of genomes, whereas alternative splicing of regulatory genes has enormous ramifications on the functionality of nearly all expressed genes.

Whole genome duplication (WGD) has catalyzed the formation of new species, genes with novel functions, altered expression patterns, complexified signaling pathways and has provided organisms a level of genetic robustness. We studied the long-term evolution and interrelationships of 5' upstream regulatory sequences (URSs), protein coding sequences (CDSs) and expression correlations (EC) of duplicated gene pairs in the model organism, *Arabidopsis thaliana*. Three distinct methods revealed significant evolutionary conservation between paralogous URSs and were highly correlated with microarray-based expression correlation of the respective gene pairs. Positional information on exact matches between sequences unveiled the contribution of micro-chromosomal rearrangements on expression divergence. A three-way rank analysis of URS similarity, CDS divergence and EC uncovered specific gene functional biases. Transcription factor activity was associated with gene pairs exhibiting conserved URSs and divergent CDSs, whereas a broad array of metabolic enzymes was found to be associated with gene pairs showing diverged URSs but conserved CDSs. Strikingly, the majority of duplicate genes are differentially expressed in magnitude throughout various developmental stages in Arabidopsis, suggesting that often one of the two gene copies is preferred and may hint at a mechanism of sub-functionalization acting at the gene regulatory level.

Along with WGD, alternative splicing (AS) of pre-mRNA is a fundamental molecular process that generates diversity in the transcriptome and proteome of eukaryotic organisms. Multiple factors influence the splicing reaction, such as the length and sequence of exons, introns, the presence and

levels of trans-factors and the rate of transcription. SR proteins, a family of splicing regulators with one or two RNA recognition motifs (RRMs) at the N-terminus and an arg/ser-rich at the C-terminus, function in both constitutive and alternative splicing.

We performed database searches for SR proteins in 27 eukaryotic species, which included taxa from plants, animals, fungi and basal eukaryotes that lie outside of these lineages. Using RRMs as a phylogenetic marker, we observed at least 12 SR protein sub-families, four of which are vastly expanded in plants. Furthermore, RRMs are in highly conserved positions within SR proteins within sub-families, yet their predicted RNA binding residues are degenerate. In line with this finding is our observation that the majority of plant SR genes are under purifying selection. Moreover, the majority of paralogous SR genes in Arabidopsis and rice are divergently expressed in different developmental stages, suggesting that these gene pairs have subfunctionalized at the expression level, reminiscent of the patterns we observed in our duplicated genes study. We assessed the extent of SR gene AS by generating splice graphs based on multiple alignments of ESTs/cDNAs to SR genomic sequences. AS of SR genes is a widespread phenomenon throughout multiple lineages and is a common trait among eukaryotes. Furthermore, the types of AS vary by organism and by SR sub-family. Lastly, we suggest that there is a link between DNA methylation within coding regions of SR genes and their AS patterns.

This thesis is a culmination of work that has spanned many years and multiple life changing events. I dedicate it entirely to my family, who have spent their lives encouraging me and being nothing more than a bastion of love and patience. I could have never done any of this without you.

and most importantly, this is for you, Dad

Acknowledgements

I would like to convey my gratitude to those who have provided me with helpful scientific discussion, criticism and general commentary, without whose help the soundness of my studies would have been otherwise called into question. I would like to thank my supervisor, Dr. Thomas Wiehe, for his incredible understanding in my moments of desperation, exhaustion, petulance and insubordination. My thanks also go out to Dr. Heiko Schoof who has provided me with additional support and invaluable insight into the world of plant genomics. Dr. ASN Reddy of *Colorado State University* also deserves my sincere respect for always being open to collaboration and imparting his insurmountable expertise to me. Dr. Asa Ben-Hur, Adam Labadorf and Mark Rogers (also from *Colorado State University*) are acknowledged for their computational efforts to make feasible the study of alternative splicing across diverse phylogenetic taxa. Last but not least, I would also like to thank the members, colleagues and friends of the Wiehe lab, who have made my stay in Germany worthwhile and all together more interesting. Dr. Daniel ("JKL") Živković has always been there to tell the horrible truth, Dr. Ivana ("KillerBee") Vukusic who was the third member of our mighty triumvirate and who welcomed me to the hospitality of Hilden, to Robert Fuerst and Andreas ("Wolle") Wollstein for computational queries and software development, to Dr. Sabari ("S-Man") Sankar Thirupathy for his eastern wisdom and finally, to Anton Malina for his zeal for all things science-fiction and general linux-is-superior-to-all-that-exists attidude.

Contents

Contents

1

Preface and Aims

1.1 Preface

This dissertation is comprised of two major independent but thematically linked analyses as shown below in Fig 1.1:

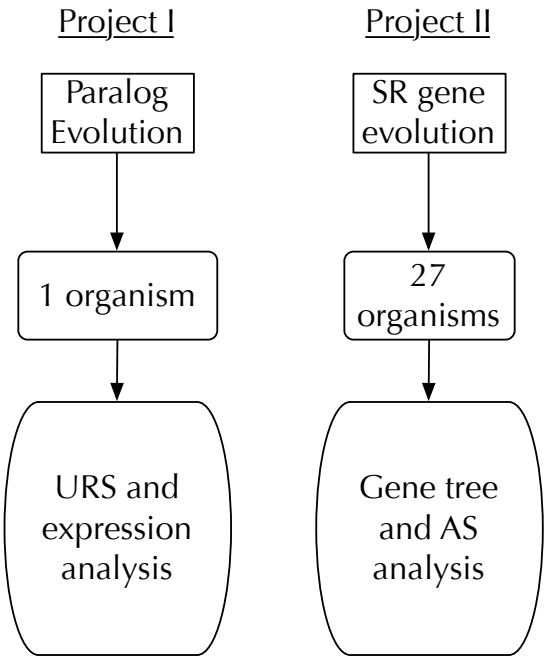

Figure 1.1: **The two projects** - Top-level organization of projects comprising this thesis.

The two projects are further described:

1. the analysis of whole genome duplication (WGD) derived genes in *Arabidopsis thaliana* with respect to upstream regulatory sequence (URS) and coding sequence (CDS) evolution within the context of paralogous expression correlation, and

2. a study of serine-arginine (SR) rich splicing factors, which focuses on their classification into sub-families, lineage-specific abundance, conservation and alternative splicing (AS) across 27 phylogenetically distinct eukaryotic genomes.

Consequently, the remainder of this thesis is partitioned accordingly and when possible, attempts will be made to unify the two topics into a cohesive whole. For the sake of clarity and ease, each of the projects will henceforth be referred to as "Project I" and "Project II". As each project was written in a modular format, the reader may choose either to begin with.

1.2 The criticality of gene regulation

How organisms develop into morphologically distinct species and respond to ever changing environmental conditions can be thought of in terms of gene regulation, that is, the timing at which particular genes are expressed, the quantity and quality of the resultant transcripts as well as the cellular localization of those transcripts. Certainly, there are many factors that influence gene regulation either before or after an mRNA transcript has been produced, such as accessibility of genes and their promoters based on chromatin states, competitive binding of transcription activators or repressors, differential expression of transcription factors, mRNA editing, or regulated degradation of nascent transcripts by RNA surveillance mechanisms (i.e., miRNA/siRNA induced degradation). While this list is far from exhaustive, it nevertheless serves to illustrate the importance of controlling the timing and location of gene expression in order for an organism to develop and dynamically adapt to a variable environment.

However, from where do genes arise? While the question of how the first gene came into being may be largely speculative, what is more certain is that new genes can arise from pre-existing genes by means of tandem, segmental or whole genome duplication events. These duplication events have consequences on gene regulation because an

organism must now deal with superfluous copies of genes. Too much or too little of a given gene product could severely affect the viability of the organism, but extra gene copies may also provide a pathway to adaptation through natural selection. The variability conferred by gene or genome duplication and the consequent regulation of these duplicated genes can lead to the formation of new species through reproductive isolation at the molecular level due to technicalities during meiosis.

In addition to the fundamental biological process of gene/genome duplication, many eukaryotic organisms have evolved another means to generate genetic variability: alternative splicing. Instead of a single gene always producing the same transcript or protein product, a single gene may give rise to multiple transcript isoforms or proteins by virtue of differential combinations of exons and introns. The resultant combinations may form a viable gene product, or result in subtle variations that confer different functions onto the protein. Often times, alternative splicing of nascent mRNA molecules results in truncated transcripts which are degraded by mRNA surveillance mechanisms in the cell. This may at first sound wasteful, but it is an elegant way to fine tune the quantity of available mRNA transcripts for translation into proteins, thus adding a layer of complexity to post-transcriptional gene regulation. Gene/genome duplication and alternative splicing are not mutually exclusive events, but instead are critical components of gene regulation, and thus, organismal complexity and development.

This thesis relies upon comparative genomics to understand how particular aspects of gene/genome duplication and alternative splicing have affected gene regulation and is primarily focused on photosynthetic eukaryotes.

1.2.1 Comparative genomics as a research tool

Comparative genomics has become one of the cornerstones of bioinformatic analyses in the post-genomic era (120). Comparative genomics is multifaceted and multi-layered. Not only is it a foundation for basic genome annotation, it is also the window in which we are able to observe the traces of evolution within and between species, which in turn, permits new insights into the biology and natural history of life.

Given its power and soundness, comparative genomics methodology is used throughout this dissertation from both an *intra-* and *inter*-species perspective. First, we examine 5' upstream regulatory regions of paralogous genes in *Arabidopsis thaliana* (*intra*-species comparative genomics in Project I) followed by an investigation of orthologous

and paralogous genes from 27 pyhlogenetically distinct species (*inter*-species comparative genomics in Project II). Both are described in detail in the following sections.

1.3 Aims

1.3.1 Project I – Intra-species comparative genomics in Arabidopsis

The primary goal of this project was to ascertain the inter-relationships of different gene components across a large group of paralogous gene pairs that arose from a single polyploidy event millions of years in the past. We will conduct analyses to address the evolution of 5' upstream regulatory sequences (URSs) versus that of protein coding sequences (CDSs). Furthermore, we will look at how each component (regulatory or coding) dictates expression correlation between paralogs, as well as biological function on a genome-wide scale.

1.3.2 Project II – Inter-species comparative genomics in 27 eukaryotes

In contrast to the above project, here the focus is on the recovery, classification and analysis of the "full" SR gene repertoire in 27 different eukaryotic species, with emphasis on 12 photosynthetic organisms and the extent of alternative splicing (AS) in the SR genes. Analyses will be performed to answer questions relating to SR gene sub-family expansions or losses in certain lineages, the extent of their AS, and characteristics of their sequence evolution within the context of gene duplication and speciation.

2

Project I Introduction

2.1 Project I – Intra-species comparative genomics in Arabidopsis

We begin by providing some background information on Arabidopsis and how it has become the model organism of choice to study basic plant biology.

2.1.1 General background information on *Arabidopsis thaliana*

Arabidopsis has become the prominent flowering plant, or angiosperm model to study plant morphogenesis, reproduction, evolution and molecular and cellular biology. Figure 2.1 depicts a photograph of an Arabidopsis adult sporophyte.

Arabidopsis was the first plant genome to be sequenced (54); consequently, it possesses one of the most curated and well-annotated genomes available within the plant kingdom. As with any model organism, Arabidopsis has many characteristics that make it an ideal organism for study:

- it has a short generation time (6 weeks)
- its genome is fully sequenced
- its genome is relatively small (5 chromosomes, 125 MB)
- there is a substantial amount of microarray expression data, and

Figure 2.1: **Arabidopsis sporophyte** - Wild-type, adult Arabidopsis photo taken from OpenWetWare.

- the fundamental aspects of its genetic evolution and molecular biology can be readily transferred to other, agronomically important crops.

The constantly refined sequence databases of Arabidopsis have permitted researchers to not only employ reverse genetic strategies in their studies, but also allows for the consideration of natural phenomena that have influenced the evolution of its genome. One important evolutionary force that has played a role in many plant lineages is that of whole genome duplication.

2.1.2 Whole genome duplication and Arabidopsis

Whole genome duplication (WGD) is a powerful force that has shaped the evolution of many, if not all, eukaryotic genomes. WGD is especially prevalent in the flowering plants, with duplications occurring multiple times throughout multiple lineages (27; 126) (Figure 2.2). WGD has had an important role in the origin and diversification of flowering plants (27) and it is estimated that at least 70% of flowering plants have polyploidy in their history (76). Arabidopsis has experienced at least three WGD events (106; 122), with the most recent WGD event having occurred between 20-60 million years ago (10; 14).

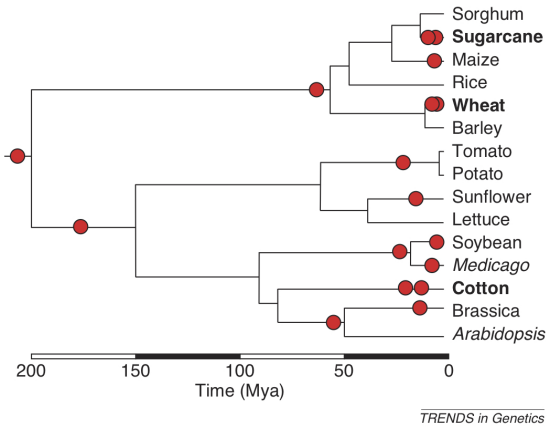

Figure 2.2: **Polyploidy in plants** - Figure showing the prevalence of polyploidization events (red circles) in selected flowering plants (70).

2.1.2.1 Prior research

Previous studies in Arabidopsis have focused primarily on how duplicated genes diverged in protein coding sequence (CDS) and expression divergence since the time of duplication. Blanc and Wolfe (2004) showed that gene pairs are not lost randomly over time, but lost according to gene functional biases, where transcription factors and signal transduction proteins were preferentially retained, whereas genes involved in DNA repair were preferentially lost. Furthermore, they reported an asymmetric rate of sequence evolution in the CDSs of more than 20% of the analyzed pairs (11). Haberer et al. (2004) reported that expression divergence occurs frequently between duplicated gene pairs and may be the primary mechanism behind preserving redundant genes (40). They also revealed a moderate but significant correlation between promoter sequence similarity and expression divergence for polyploidy derived pairs. Ganko et al. (2007) examined levels of expression divergence between duplicated genes in Arabidopsis and reported that the strength of purifying selection acting on CDSs is coupled to the corresponding pairs expression pattern (35).

2.1.2.2 Limitations and extensions of prior research

Most studies thus far have focused primarily on the properties of coding sequence evolution after duplication, measured in terms of non-synonymous (K_a), synonymous (K_s), or the ratio of non-synonymous to synonymous substitutions (K_a/K_s). However, the

evolution of upstream regulatory sequences (URSs) of duplicated genes has received less attention in Arabidopsis. This is partly due to the inherent difficulty of assessing sequence similarity in non-coding DNA, where the number, order and spacing of shared sequence elements may confound traditional, alignment-based approaches. Additionally, the well-defined models for coding sequence substitution rates are not applicable to URSs, which lack the discrete nature of codons. Moreover, the limited number of known in conjunction with the vast number of computationally predicted transcription factor binding sites (TFBSs) for Arabidopsis makes it difficult to recover a meaningful signal from background noise. One of the major public sources for obtaining TFBSs, PLACE (49), is no longer maintained and has not been updated since 2007 (only 409 motifs are readily available from PLACE).

Previously, Haberer et al. (2004) looked at similarities/dissimilarities in duplicated Arabidopsis promoters using a simple alignment-based approach (40). Here, by using the Shared Motif Method (21), DIALIGN-TX (an improved version of the algorithm used in (21)) (113; 114) and an alignment-free measure of word repetitiveness, (known as the IR (43; 45)) we were able to characterize aspects of Arabidopsis URS similarity in paralogous genes at a more detailed level. Furthermore, the incorporation of positional information on exact matches between paralogous URSs provided insights into URS sequence evolution and expression divergence that raw similarity values simply cannot provide. An evolutionary analysis of the protein coding sequences of the WGD-derived paralogs revealed distinct functional classifications of the duplicates, dependent on whether the URS or CDS is more conserved. Moreover, the joint consideration of URSs and CDSs revealed that different components of a gene experience different selective pressures following gene duplication.

3

Results – Project I

3.1 Similarity profiles of Arabidopsis upstream regulatory sequences (URSs)

After taking several steps to assemble a clean working data set of 815 paralogous URSs (Figure 3.1, and section 8), all of which are assumed to have arisen in a single WGD event between 20-60 million years ago (10), we applied three distinct means to assess similarity between each of the gene pairs: the Index of Repetitiveness (IR) (43; 45), Shared Motif Method (SMM) (21) and DIALIGN-TX (DTX) (113).

First, we confirmed that the IR and SMM values for the real data (Figure 3.2A and 3.2B, black boxes), as well as the measurements using DTX are significantly different from the randomized data (Figure 3.2A and 3.2B, red boxes; DTX data not shown).

A noticeable property of the IR was that its level of variance depended more heavily on the sequence lengths of analyzed URSs than the SMM. The inter-quartile range (IQR) for the IR values for sequences of length 600 bp was only 20% of the IQR for sequences of length 100 bp. For the SMM, the IQR was almost independent of sequence length. Furthermore, there are many more outliers and extreme values in the real data measured by IR (Figure 3.2A, black boxes) than by the SMM (Figure 3.2B, black boxes), whereas the converse is true for the random data (Figure 3.2A and 3.2B, red boxes). However, in general, median values of conservation for duplicated URSs decrease as the size of the TSS-anchored window increases.

Next, we tried to localize the regions within the URSs that harbor the highest

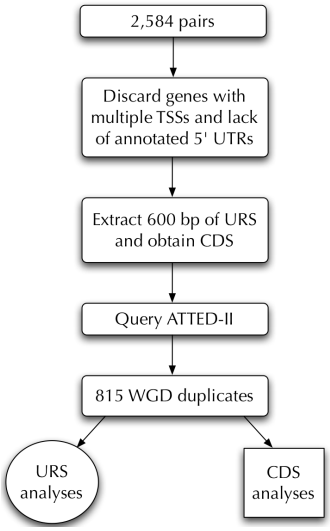

Figure 3.1: **Dataset construction** - Brief summary of the steps taken to generate the working data sets for Project I. For more details, the reader is referred to chapter 8.

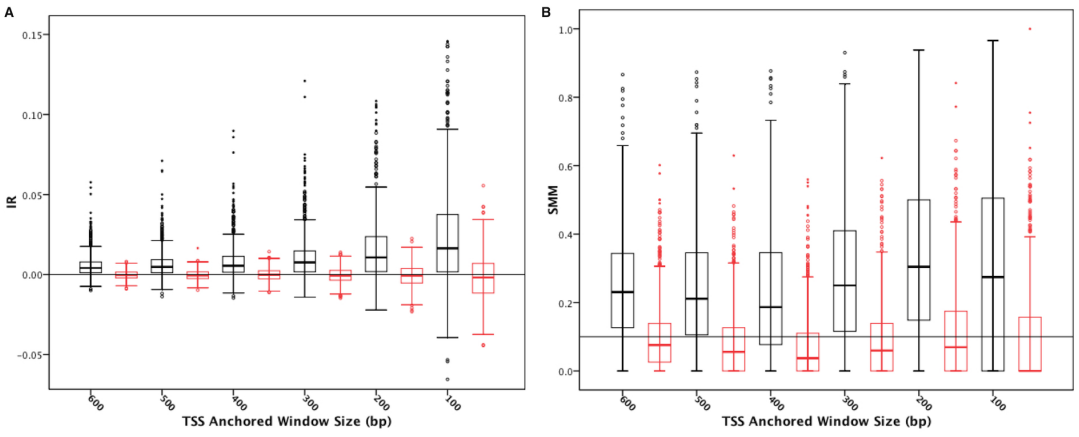

Figure 3.2: **Anchored window analysis** - Observed and randomized data (black and red box plots, respectively) for the IR (Panel A) and the SMM (Panel B).

conservation signals. We performed a sliding window analysis with the window size fixed at 200 bp and moving away from the 5' TSS in 50 bp steps (Figure 3.3).

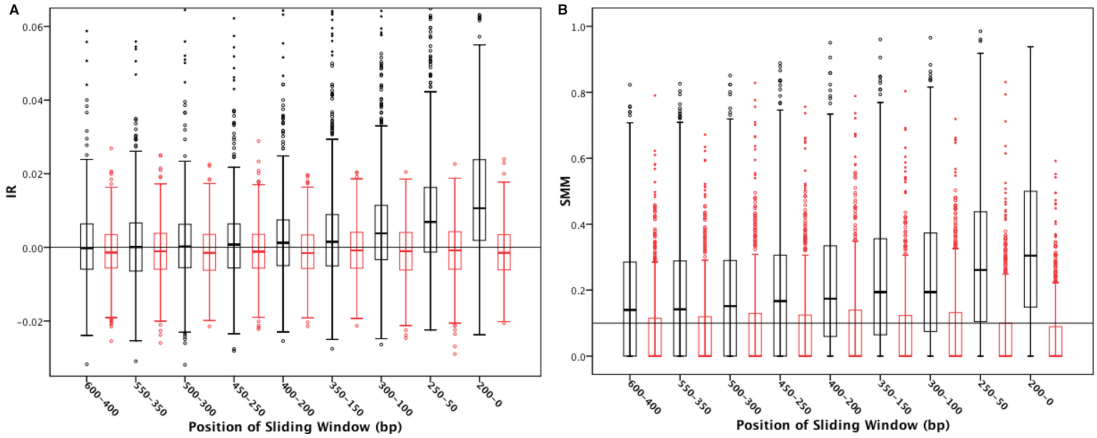

Figure 3.3: **Sliding window analysis** - Observed and randomized data (black and red box plots, respectively) for the IR (Panel A) and the SMM (Panel B).

Again, similar trends were observed with respect to the behavior of the two methods: all real data values were significantly different from random data (Wilcoxon Signed Ranks test, $p << 10^{-5}$) and the pattern of outliers was reflective of what was observed in Figure 3.2. Furthermore, note that the variability pattern in the IR and SMM for random data, exhibited fewer outliers for IR and a relatively constant level of variance (Figure 3.3A and 3.3B, red boxes). Nevertheless, despite the qualitative differences in both programs, a distinct pattern emerged with respect to the sliding window analysis. Both programs showed a nearly monotonic decrease as distance from the 5 TSS increased. Such a result suggests that most of the conservation signal (sequence similarity) is within the first 300 bp upstream of the TSS.

3.2 Inter-relationships between URSs, CDSs and expression correlation

Haberer et al. (2004) found only a marginal correlation ($r = 0.12$, $0.01 < p < 0.05$) between URS similarity and expression correlation in their study of WGD-derived duplicate pairs in Arabidopsis (40). Therefore, we correlated the IR, SMM and DTX data

compiled in the window analyses (Figures 3.2 and 3.3) with relative levels of gene expression between each gene pair (Table 3.1). We observed highly significant Spearman rank correlations (ρ) between URS similarity and expression correlation (range of correlation: 0.159 – 0.277, $p << 0.01$); in some cases, the correlation was more than twice that previously reported. The IR yielded higher correlations than the SMM, except for the first 100 bp TSS-anchored window. A peak in correlation was observed in the 300 bp window for IR ($\rho = 0.277$), whereas a bimodality was evident in the 100 bp and 200 bp windows of the SMM ($\rho = 0.226$). We also calculated the correlation between all three programs. In agreement with our observation that most of the conservation signal was found in the immediate upstream region of about 300 bp, we also observed that the three methods were most highly correlated at smaller window sizes (last three rows of Table 3.1).

Table 3.1: Spearman rank correlations in the anchored window

Anchored Window	IR-EC	SMM-EC	DTX-EC	IR-SMM	IR-DTX	SMM-DTX
600	0.222	0.159	0.158	0.451	0.251	0.391
500	0.228	0.169	0.153	0.471	0.273	0.401
400	0.247	0.192	0.191	0.510	0.269	0.437
300	0.277	0.207	0.175	0.525	0.295	0.388
200	0.252	0.226	0.178	0.582	0.313	0.412
100	0.214	0.226	0.097	0.619	0.292	0.380

Spearman rank correlations for the 815 pairs of URS regions, as well as inter-application correlations. All values for the anchored window URS regions are significant at the 1% level. A single asterisk denotes significance at the 0.05 level. IR, Index of Repetitiveness; SMM, Shared Motif Method; DTX, DIALIGN-TX; EC, expression correlation.

Table 3.2: Spearman rank correlations in the sliding window

Sliding Window	IR-EC	SMM-EC	DTX-EC	IR-SMM	IR-DTX	SMM-DTX
400-600	0.059	0.059	0.021	**0.287**	0.016	**0.188**
350-550	0.049	0.074*	0.022	**0.261**	-0.016	**0.225**
300-500	0.064	0.042	0.034	**0.308**	-0.044	**0.165**
250-450	**0.094**	0.036	0.040	**0.265**	0.019	**0.190**
200-400	**0.121**	0.063	0.041	**0.329**	0.065	**0.209**

Continued...

Sliding Window	IR-EC	SMM-EC	DTX-EC	IR-SMM	IR-DTX	SMM-DTX
150-350	**0.190**	0.085*	0.042	**0.356**	**0.115**	**0.217**
100-300	**0.217**	**0.150**	**0.079**	0.431	**0.139**	**0.299**
50-250	**0.257**	**0.198**	**0.108**	0.472	**0.232**	**0.316**
0-200	**0.253**	**0.226**	**0.170**	0.582	**0.314**	**0.404**

Spearman rank correlations for the 815 pairs of URS regions, as well as inter-application correlations. Values in bold for the sliding window URS regions are significant at the 1% level. A single asterisk denotes significance at the 0.05 level. IR, Index of Repetitiveness; SMM, Shared Motif Method; DTX, DIALIGN-TX; EC, expression correlation.

We also correlated the data from the sliding window analysis with expression correlation (Table 3.2). The pattern in correlation mirrors that of what was observed for URS conservation in Figure 3.3; that is, increasing distance from the 5 TSS translates not only into a reduced level of sequence conservation but also into a reduced correlation with expression. Considering only the windows that had correlations significant at the 1% level (bold values in Table 3.2), all programs reported high correlations with EC within the three sliding windows most proximal to the 5' TSS. The same pattern was evident in terms of the inter-application correlations reported for the anchored window analysis: the closer to the 5' TSS, the more congruent the methods were.

Note that In both the anchored window and sliding window analyses, the IR and the SMM yielded higher correlation values with EC than DTX. Additionally, the inter-program correlation values were higher between the IR and the SMM than either was with DTX.

3.2.1 Substitution rates of CDS and their relationship with EC

As our focus was not solely on the contribution of the URS to EC, we also measured properties of the coding sequences, such as sequence identity, synonymous (K_s) and non-synonymous (K_a) substitution rates and their ratio (K_a/K_s) (Table 3.3). Sequences were carefully aligned using the amino acid sequences of the respective gene pairs as a backbone to aid in the construction of the DNA coding sequence alignment (see chapter 8 for specifics).

Table 3.3: Spearman rank correlations for coding sequences

Coding Seq. Property	EC
Identity (protein)	**0.157**
K_s	0.032
K_a	**-0.185**
K_a/K_s	**-0.213**
Ti/Tv	-0.132
Δ Length	**-0.150**

Spearman rank correlations for coding sequence characteristics. Values in bold indicate significance at the 0.01 level.

A marginal but significant positive correlation was observed between protein sequence identity and expression correlation (Table 3.2). In complete agreement with this result, we also observed a significant negative correlation between expression correlation and the rate of non-synonymous substitutions, K_a. On the other hand, no correlation was found between expression correlation and the rate of synonymous substitutions, K_s. Also, the transition to transversion ratio (Ti/Tv) and the difference in length between duplicate pairs exhibited a significantly negative correlation. Taken together, there is concomitant functional constraint on the protein and its expression profile, as evidenced by the significantly negative correlation between expression correlation and K_a/K_s. Markedly, no significant correlation was observed between any of the URS windows and any evolutionary property of the coding sequences listed in Table 3.2 (data not shown).

3.3 Micro-chromosomal rearrangements of exact matches

Since similarity values alone do not encompass the entire range of sequence evolution, we analyzed the 300 bp URSs using the positional information and lengths of exact matches (7-19 bp) between these sequences. We analyzed the whole data set, gene pairs in the upper 25% and lower 25% quantiles based on EC. We considered four mutually exclusive arrangement classes for the type of exact matches that can occur: proximal exact matches (pem), distal exact matches (dem), inverse proximal exact matches (ipem) and inverse distal exact matches (idem) (Figure 3.4)

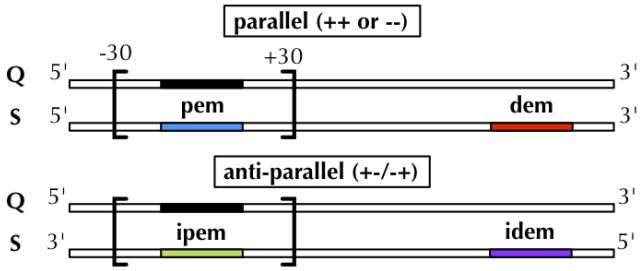

Figure 3.4: **Distribution of exact match arrangement classes in all gene pairs** - The four mutually exclusive arrangement classes are indicated by color. The solid black rectangle indicates the location of the exact match in the query. The large black brackets indicate the boundaries that define what we consider in this work to be "proximal". Pem, proximal exact match; dem, distal exact match; ipem, inverse proximal exact match; idem, inverse distal exact match.

Pem and ipem are defined as an exact match in the query sequence that is located within the subject sequence constrained by ±30 bp boundaries (blue and green boxes in Figure 3.4), whereas dem and idem are exact matches that are located outside of these boundaries (red and purple boxes in Figure 3.4).

We observed that as the length of the exact match increases, there is concordant increase and decrease in the fraction of pem and idem (blue and purple lines, Figure 3.5. By contrast, the fraction of ipem and dem (red and green lines) remained relatively constant as length increased.

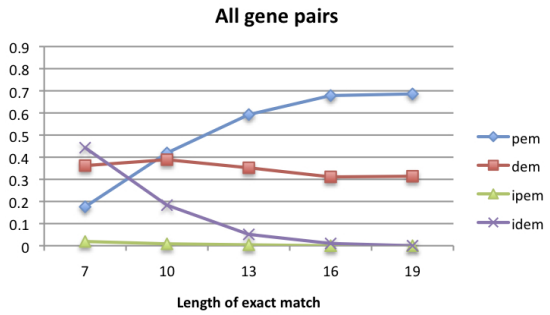

Figure 3.5: **Distribution of exact matches into the four arrangement classes for all gene pairs** - The four mutually exclusive arrangement classes are indicated by color. Pem, proximal exact match; dem, distal exact match; ipem, inverse proximal exact match; idem, inverse distal exact match.

A similar pattern was reflected in the upper 25% EC quantile gene pairs (Figure 3.6).

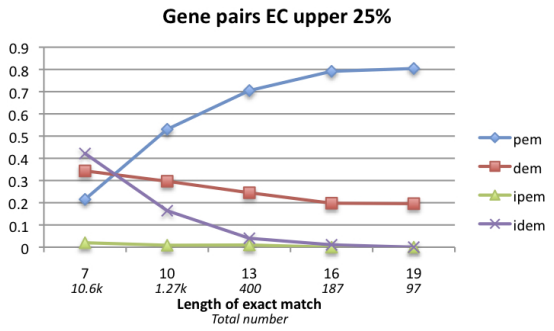

Figure 3.6: **Distribution of exact matches into the four arrangement classes for gene pairs in the upper 25% EC quantile** - The four mutually exclusive arrangement classes are indicated by color. Pem, proximal exact match; dem, distal exact match; ipem, inverse proximal exact match; idem, inverse distal exact match.

However, a distinctly different pattern emerged in the lower 25% EC quantile gene pairs Figure 3.7. Here, as length increased, the fraction of pem and dem decreased and increased, respectively. Also with increased length, the fraction of idem (purple line) was higher for the first three length categories.

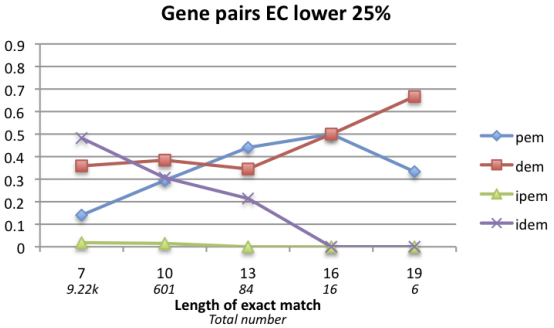

Figure 3.7: **Distribution of exact matches into the four arrangement classes for gene pairs in the lower 25% EC quantile** - The four mutually exclusive arrangement classes are indicated by color. Pem, proximal exact match; dem, distal exact match; ipem, inverse proximal exact match; idem, inverse distal exact match.

3.4 Three-way rank analysis

To understand what kind of gene functional biases result as a consequence of URS and CDS evolution within the context of expression correlation, we performed a joint rank analysis of three variables: the IR of the 300 bp anchored URS window, K_a of the CDS and the EC of the duplicate pairs. For each variable there were three equally populated classes ($n = 245$): high conservation, middle conservation and low conservation (Table 3.4).

Table 3.4: Rank classifications

	low	middle	high
IR	-0.014-0.004	0.004-0.012	0.012-0.123
K_a	0.159-2.221	0.103-0.159	0.000-0.103
EC	-0.578-0.273	0.273-0.499	0.500-0.958

A total of 735 gene pairs (after excluding pairs with $K_s > 1.5$) were ranked for the IR, K_a and expression correlation values. For each variable, the ranks were split into three equal classes ($n = 245$): low ranks, middle ranks and high ranks. The table presents the ranges of raw values for each of the variables for each rank class.

Of the 27 possible combinations for each rank class, we focused on two specific categories: those gene pairs that fell into high rank categories for IR and expression correlation but had low K_a ranks (26 pairs) and a second grouping that included high rank categories for K_a and expression correlation but low IR ranks (36 pairs). We checked for overrepresented GO slim terms using the web interface at AmiGO (20) for each of these data sets. The enriched molecular function terms for the gene pairs with high ranked K_a and expression correlation but low ranked IR values are reported in Table 3.5.

Interestingly, about half of the genes in this rank grouping are associated with transcription regulator activity (GO:0030528), transcription factor activity (GO:0003700) or DNA binding (GO:0003677) (first three rows in Table 3.5). In contrast, the rank grouping with high ranked K_a and expression correlation but low ranked IR values have completely different functional biases (Table 3.6).

Table 3.5: GO term enrichment in high IR but low K_a ranked gene pairs

GO Term	Description	P-value	Sample frequency	Background frequency
GO:0030528	transcription regulator activity	4.30E-27	37/73 (50.7%)	1813/32333 (5.6%)
GO:0003700	transcription factor activity	1.72E-23	33/73 (45.2%)	1659/32333 (5.1%)
GO:0003677	DNA binding	1.17E-22	35/73 (47.9%)	2083/32333 (6.4%)
GO:0003676	nucleic acid binding	1.23E-16	36/73 (49.3%)	3437/32333 (10.6%)
GO:0005488	binding	4.57E-08	37/73 (50.7%)	7006/32333 (21.7%)
GO:0016563	transcription activator activity	9.27E-05	4/73 (5.5%)	104/32333 (0.3%)
GO:0004568	chitinase activity	4.50E-04	2/73 (2.7%)	14/32333 (0.0%)
GO:0008308	voltage-gated anion channel activity	4.50E-04	2/73 (2.7%)	14/32333 (0.0%)
GO:0005253	anion channel activity	5.18E-04	2/73 (2.7%)	15/32333 (0.0%)
GO:0009982	pseudouridine synthase activity	6.69E-04	2/73 (2.7%)	17/32333 (0.1%)
GO:0016564	transcription repressor activity	2.10E-03	2/73 (2.7%)	30/32333 (0.1%)
GO:0004693	cyclin-dependent protein kinase activity	2.39E-03	2/73 (2.7%)	32/32333 (0.1%)
GO:0000156	two-component response regulator activity	2.85E-03	2/73 (2.7%)	35/32333 (0.1%)
GO:0022832	voltage-gated channel activity	3.18E-03	2/73 (2.7%)	37/32333 (0.1%)
GO:0005244	voltage-gated ion channel activity	3.18E-03	2/73 (2.7%)	37/32333 (0.1%)
GO:0016866	intramolecular transferase activity	4.88E-03	2/73 (2.7%)	46/32333 (0.1%)
GO:0022836	gated channel activity	7.92E-03	2/73 (2.7%)	59/32333 (0.2%)

Molecular Function GO terms for gene pairs with divergent coding sequences but conserved upstream sequences and expression correlation

Table 3.6: GO term enrichment in low IR but high K_a ranked gene pairs

GO Term	Description	P-value	Sample frequency	Background frequency
GO:0016835	carbon-oxygen lyase activity	1.46E-06	5/52 (9.6%)	120/32333 (0.4%)
GO:0008676	3-deoxy-8-phosphooctulonate synthase activity	2.54E-06	2/52 (3.8%)	2/32333 (0.0%)
GO:0051002	ligase activity, forming nitrogen-metal bonds	7.60E-06	2/52 (3.8%)	3/32333 (0.0%)
GO:0051003	ligase activity, forming nitrogen-metal bonds, forming coordination complexes	7.60E-06	2/52 (3.8%)	3/32333 (0.0%)
GO:0010280	UDP-L-rhamnose synthase activity	7.60E-06	2/52 (3.8%)	3/32333 (0.0%)
GO:0016851	magnesium chelatase activity	7.60E-06	2/52 (3.8%)	3/32333 (0.0%)
GO:0005198	structural molecule activity	8.68E-06	7/52 (13.5%)	460/32333 (1.4%)
GO:0016857	racemase and epimerase activity, acting on carbohydrates and derivatives	1.25E-05	3/52 (5.8%)	28/32333 (0.1%)
GO:0016854	racemase and epimerase activity	2.07E-05	3/52 (5.8%)	33/32333 (0.1%)
GO:0003978	UDP-glucose 4-epimerase activity	2.53E-05	2/52 (3.8%)	5/32333 (0.0%)
GO:0004739	pyruvate dehydrogenase (acetyl-transferring) activity	3.79E-05	2/52 (3.8%)	6/32333 (0.0%)
GO:0004738	pyruvate dehydrogenase activity	3.79E-05	2/52 (3.8%)	6/32333 (0.0%)
GO:0004386	helicase activity	6.93E-05	4/52 (7.7%)	136/32333 (0.4%)
GO:0016829	lyase activity	8.61E-05	5/52 (9.6%)	279/32333 (0.9%)
GO:0003824	catalytic activity	1.77E-04	22/52 (42.3%)	6410/32333 (19.8%)

Continued...

GO Term	Description	P-value	Sample frequency	Background frequency
GO:0016624	oxidoreductase activity, acting on the aldehyde or oxo group of donors, disulfide as acceptor	1.96E-04	2/52 (3.8%)	13/32333 (0.0%)
GO:0016836	hydro-lyase activity	1.99E-04	3/52 (5.8%)	70/32333 (0.2%)
GO:0003735	structural constituent of ribosome	3.00E-04	5/52 (9.6%)	365/32333 (1.1%)
GO:0004089	carbonate dehydratase activity	3.40E-04	2/52 (3.8%)	17/32333 (0.1%)
GO:0003724	RNA helicase activity	3.82E-04	2/52 (3.8%)	18/32333 (0.1%)
GO:0017111	nucleoside-triphosphatase activity	7.67E-04	5/52 (9.6%)	449/32333 (1.4%)
GO:0030570	pectate lyase activity	8.04E-04	2/52 (3.8%)	26/32333 (0.1%)
GO:0016837	carbon-oxygen lyase activity, acting on polysaccharides	8.04E-04	2/52 (3.8%)	26/32333 (0.1%)
GO:0016462	pyrophosphatase activity	9.41E-04	5/52 (9.6%)	470/32333 (1.5%)
GO:0016818	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	9.50E-04	5/52 (9.6%)	471/32333 (1.5%)
GO:0016817	hydrolase activity, acting on acid anhydrides	9.50E-04	5/52 (9.6%)	471/32333 (1.5%)
GO:0008374	O-acyltransferase activity	1.00E-03	2/52 (3.8%)	29/32333 (0.1%)
GO:0005200	structural constituent of cytoskeleton	1.07E-03	2/52 (3.8%)	30/32333 (0.1%)
GO:0016853	isomerase activity	3.31E-03	3/52 (5.8%)	185/32333 (0.6%)

Continued...

GO Term	Description	P-value	Sample frequency	Background frequency
GO:0016903	oxidoreductase activity, acting on the aldehyde or oxo group of donors	4.09E-03	2/52 (3.8%)	59/32333 (0.2%)
GO:0046983	protein dimerization activity	4.95E-03	2/52 (3.8%)	65/32333 (0.2%)
GO:0008026	ATP-dependent helicase activity	8.13E-03	2/52 (3.8%)	84/32333 (0.3%)
GO:0070035	purine NTP-dependent helicase activity	8.13E-03	2/52 (3.8%)	84/32333 (0.3%)

Molecular function GO terms for gene pairs with divergent upstream sequences but conserved coding sequences and expression correlation.

Almost all cases were associated with enzymatic activity, such as various lyases, ligases and hydrolases, etc., with 42% of the genes enriched for the term catalytic activity (GO:0003824). There was not a single transcription factor term annotated in this rank grouping. Furthermore, 13.5% of the genes were enriched for structural molecule activity (GO:0005198).

3.5 Cluster analysis of gene expression magnitude during Arabidopsis development

While we have established links between expression correlation, upstream regulatory and coding sequence divergence in the evolution of paralogous gene pairs, our analyses suffered from a lack of resolution regarding information on expression divergence. The use of pearson correlation coefficients (r) as an index for co- or divergent expression across multiple experiments between paralogs did not directly consider changes in expression across developmental stages or anatomical components. In order to address this shortcoming, new expression data were acquired from the Genevestigator database (137).

Genevestigator is a functional genomics database that permits meta-analyses of gene expression across ontologically distinct developmental stages and anatomical components by providing the user with a friendly interface to query genes of interest. The database is a conglomeration of thousands of microarray experiments that allows one to observe normalized average intensity values of gene expression across samples that share the same biological context (Figure 3.8).

We queried our 815 gene pairs against the Genevestigator repository and recovered unique probe sets for 807 gene pairs. Next, we calculated the log_2 signal intensity ratios for each of the gene pairs across nine separate developmental stages (germinated seed, seedling, young rosette, developed rosette, bolting, young flower, developed flower, flowers and siliques and mature siliques). After scaling these values to allow for comparisons, hierarchical clustering based on a euclidean distance matrix of these values was performed (Figure 3.9) and six clusters were selected based on the within group sum of squares as shown in Figure 3.10.

After clustering these data, the means of each cluster were calculated for each developmental stage and are shown with standard errors in Figure 3.11. Note that

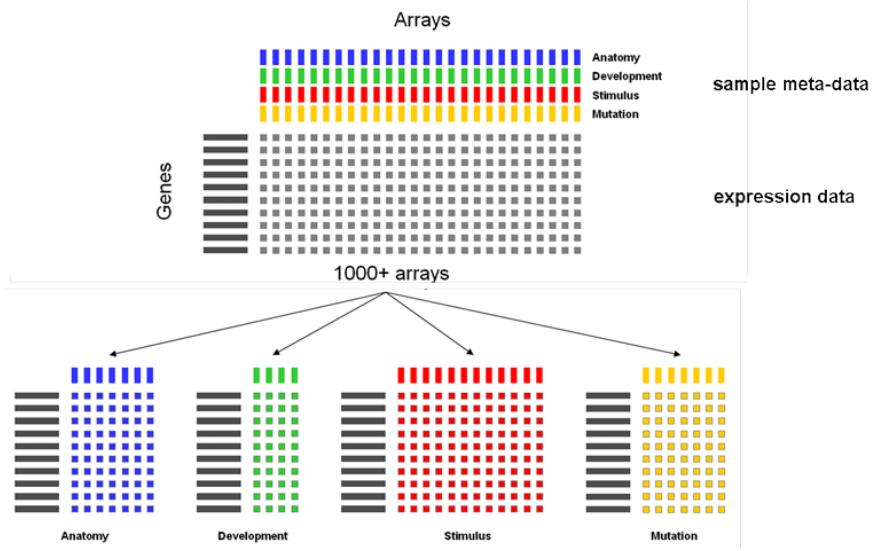

Figure 3.8: **Schematic representation of the meta-profile concept.** - Expression values from large number of arrays (top) are summarized into meta-profiles (bottom) according to their annotations. (Taken directly from the Genevestigator website).

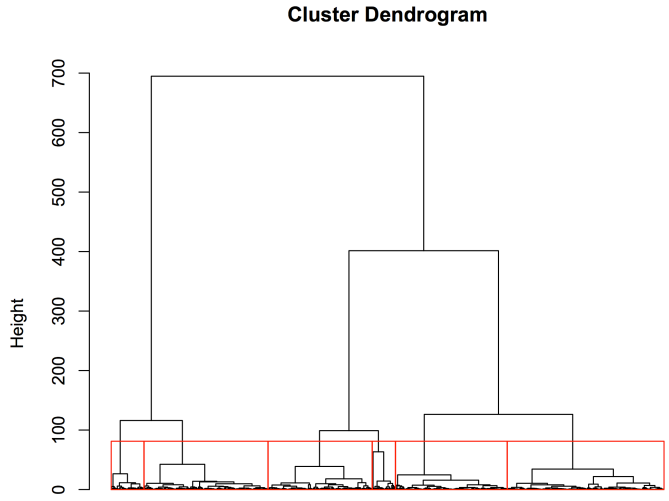

Figure 3.9: **Hierarchical clusters of the 807 paralogous genes.** - The 807 gene pairs grouped into six clusters that are indicated in red boxes.

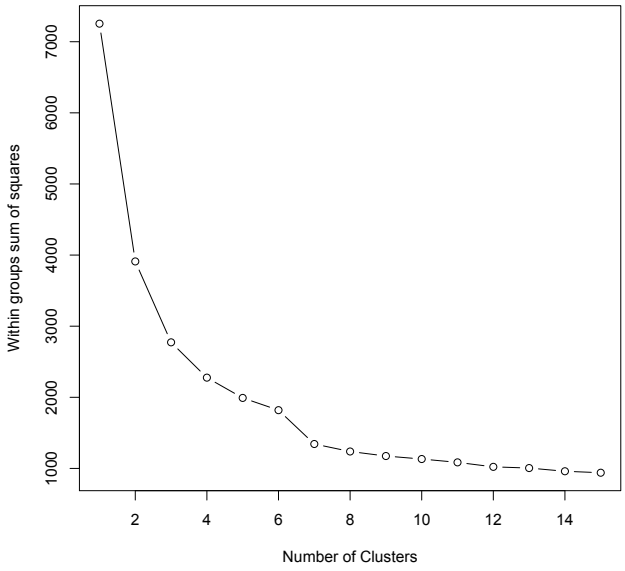

Figure 3.10: **Pseudo-scree plot.** - The number of clusters are plotted against the sum of squares for each group, which is similar to a scree test in principal component analysis.

some of the clusters appear to be mirror images of each other (e.g., Cluster 3 and Cluster 6). This simply reflects the nature of the ratio comparisons: in Cluster 3, the first member of a gene pair was expressed at a level of 1.5 to 2 times higher than the second member, whereas for Cluster 6, the converse was true. In all clusters and all developmental stages, there were only two instances where cluster means approached a log_2 ratio of zero, or in other words, where there was no fold-change in average expression intensity between gene pairs (germinated seed and mature siliques, red line in Figure 3.11. This is in stark contrast to the roughly 72% of the 807 paralogous gene pairs that showed consistent and differential expression across all developmental stages.

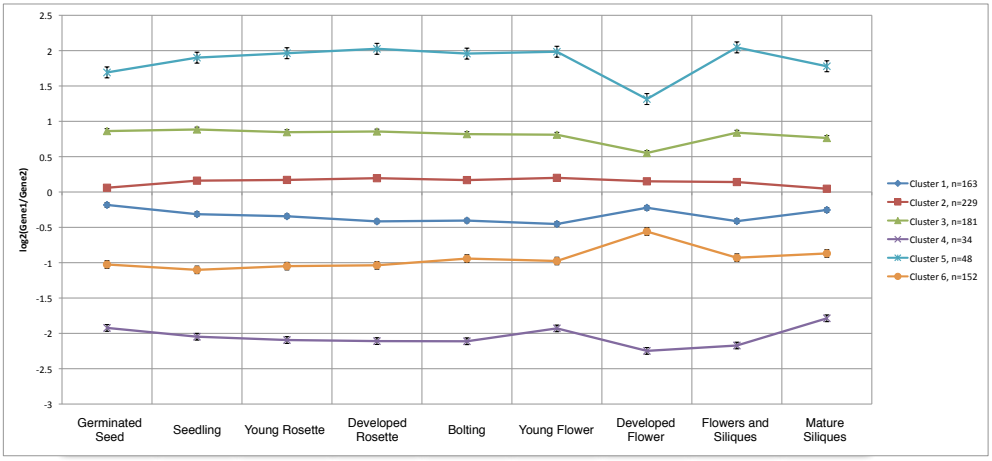

Figure 3.11: **Cluster means across 9 developmental stages.** - The six clusters with their mean fold-differences in expression intensities and standard errors. Sample sizes of each cluster are indicated in the legend. See text for details.

3.5.1 GO term enrichment by cluster

In addition to investigating the extent of expression divergence, we searched for significantly enriched GO terms for each of the six clusters. Interesting patterns emerged with respect to gene functional biases and involvement in biological processes based on the average ratios of differential gene expression. Cluster 1 (n = 163 pairs) was enriched for transcription factor activity and transcription factor regulator activity and contained gene pairs involved in the regulation of primary metabolic processes, such as RNA, carbohydrate and nitrogen compound metabolic processes (Table 3.7).

Table 3.7: GO term enrichment for molecular function and biological process in Cluster 1

GO Term	Aspect	Description	P-value	Sample Frequency	Background Frequency
GO:0003700	F	transcription factor activity	1.08E-13	17.8%	5.2%
GO:0030528	F	transcription regulator activity	7.75E-12	18.2%	5.9%
GO:0003677	F	DNA binding	4.05E-10	19.7%	7.4%
GO:0047262	F	polygalacturonate 4-alpha-galacturonosyltransferase activity	2.39E-04	1.8%	0.1%
GO:0005488	F	binding	4.01E-03	44.0%	32.0%
GO:0005215	F	transporter activity	9.44E-03	9.2%	3.8%
GO:0031323	P	regulation of cellular metabolic process	2.02E-10	16.9%	5.6%
GO:0019222	P	regulation of metabolic process	2.24E-10	17.8%	6.2%
GO:0051171	P	regulation of nitrogen compound metabolic process	4.97E-10	16.0%	5.2%
GO:0019219	P	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	1.15E-09	15.7%	5.1%
GO:0065007	P	biological regulation	1.46E-09	24.9%	11.1%
GO:0050789	P	regulation of biological process	1.71E-09	22.5%	9.4%
GO:0050794	P	regulation of cellular process	6.14E-09	20.3%	8.2%
GO:0080090	P	regulation of primary metabolic process	7.17E-09	15.7%	5.4%
GO:0010556	P	regulation of macromolecule biosynthetic process	7.73E-09	15.1%	5.0%

Continued...

GO Term	Aspect	Description	P-value	Sample Frequency	Background Frequency
GO:0045449	P	regulation of transcription	1.15E-08	14.8%	4.9%
GO:0009889	P	regulation of biosynthetic process	1.81E-08	15.1%	5.2%
GO:0031326	P	regulation of cellular biosynthetic process	1.81E-08	15.1%	5.2%
GO:0060255	P	regulation of macromolecule metabolic process	4.56E-08	15.7%	5.7%
GO:0010468	P	regulation of gene expression	5.58E-08	15.4%	5.5%
GO:0006350	P	transcription	1.00E-07	14.8%	5.2%
GO:0006355	P	regulation of transcription, DNA-dependent	4.17E-07	9.8%	2.7%
GO:0051252	P	regulation of RNA metabolic process	4.57E-07	9.8%	2.7%
GO:0006351	P	transcription, DNA-dependent	1.70E-06	9.8%	2.9%
GO:0032774	P	RNA biosynthetic process	1.80E-06	9.8%	2.9%
GO:0016051	P	carbohydrate biosynthetic process	6.59E-05	4.6%	0.8%
GO:0006807	P	nitrogen compound metabolic process	9.96E-05	20.0%	10.2%
GO:0050896	P	response to stimulus	1.97E-04	21.5%	11.6%
GO:0034641	P	cellular nitrogen compound metabolic process	2.22E-04	19.4%	10.0%
GO:0006139	P	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	2.53E-04	17.5%	8.7%
GO:0009058	P	biosynthetic process	4.07E-04	24.6%	14.2%
GO:0016485	P	protein processing	2.04E-03	1.5%	0.1%

Continued...

GO Term	Aspect	Description	P-value	Sample Frequency	Background Frequency
GO:0009719	P	response to endogenous stimulus	2.99E-03	8.3%	3.0%
GO:0042221	P	response to chemical stimulus	3.16E-03	12.9%	6.1%
GO:0051604	P	protein maturation	3.41E-03	1.5%	0.1%
GO:0090304	P	nucleic acid metabolic process	6.09E-03	15.1%	7.8%

Note that gene pairs in this first cluster were on average moderately differentially expressed, with one member of a gene pair expressed at about 1.2-1.4 fold difference. Gene pairs in Cluster 2 ($n = 229$ pairs) were even relatively less differentially expressed with fold differences hovering around 1.1. This perceived lack of differential expression between these duplicates in Cluster 2 is reflected in their GO terms, as the most enriched terms were related to enzymatic or catalytic activity in response to various stresses or stimuli (Table 3.8).

Table 3.8: GO term enrichment for molecular function and biological process in Cluster 2

GO Term	Aspect	Description	P-value	Sample Frequency	Background Frequency
GO:0016740	F	transferase activity	1.02E-09	18.5%	8.1%
GO:0003824	F	catalytic activity	1.40E-08	39.7%	25.4%
GO:0016301	F	kinase activity	6.72E-05	9.7%	4.0%
GO:0016772	F	transferase activity, transferring phosphorus-containing groups	9.82E-05	10.6%	4.6%
GO:0004674	F	protein serine/threonine kinase activity	7.02E-04	6.2%	2.1%
GO:0016773	F	phosphotransferase activity, alcohol group as acceptor	7.62E-04	8.2%	3.3%
GO:0016857	F	racemase and epimerase activity, acting on carbohydrates and derivatives	1.20E-03	1.3%	0.1%
GO:0004672	F	protein kinase activity	1.62E-03	7.3%	2.9%
GO:0016765	F	transferase activity, transferring alkyl or aryl (other than methyl) groups	2.02E-03	2.4%	0.4%
GO:0016854	F	racemase and epimerase activity	4.21E-03	1.3%	0.1%
GO:0042221	P	response to chemical stimulus	2.17E-10	15.7%	6.1%
GO:0050896	P	response to stimulus	1.01E-09	23.4%	11.6%
GO:0010033	P	response to organic substance	1.56E-09	11.5%	3.8%
GO:0009719	P	response to endogenous stimulus	2.10E-08	9.7%	3.0%
GO:0070887	P	cellular response to chemical stimulus	2.21E-06	5.7%	1.4%

Continued...

GO Term	Aspect	Description	P-value	Sample Frequency	Background Frequency
GO:0009725	P	response to hormone stimulus	3.04E-06	8.4%	2.8%
GO:0071310	P	cellular response to organic substance	7.43E-06	5.3%	1.3%
GO:0009651	P	response to salt stress	9.20E-06	5.1%	1.2%
GO:0006970	P	response to osmotic stress	4.11E-05	5.1%	1.3%
GO:0051716	P	cellular response to stimulus	4.71E-05	7.7%	2.7%
GO:0009755	P	hormone-mediated signaling pathway	4.99E-05	4.2%	0.9%
GO:0032870	P	cellular response to hormone stimulus	5.61E-05	4.2%	0.9%
GO:0071495	P	cellular response to endogenous stimulus	5.96E-05	4.4%	1.0%
GO:0006950	P	response to stress	7.86E-05	13.5%	6.5%
GO:0023033	P	signaling pathway	1.33E-04	6.8%	2.3%
GO:0009987	P	cellular process	1.65E-04	43.3%	31.7%
GO:0022622	P	root system development	5.90E-04	3.8%	0.9%
GO:0048364	P	root development	5.90E-04	3.8%	0.9%
GO:0009628	P	response to abiotic stimulus	6.58E-03	8.8%	4.1%
GO:0048528	P	post-embryonic root development	8.77E-03	1.8%	0.2%

Clusters 3 ($n = 181$ pairs) and 6 ($n = 152$ pairs) exhibited a 1.4-2.1 fold difference in expression for their gene pairs (green and orange lines in Figure 3.11. Cluster 3 was enriched for terms related to catalytic activity, binding and mainly processes linked to response to endogenous stimuli such as hormones, abiotic stimuli such as temperature and involved primarily in metabolic proceses (Table 3.9). Cluster 6, on the other hand, had molecular functions centered around transcription factor or regulator activity and contains differentially expressed gene pairs involved in organ and system development, gene regulation and regulation of primary metabolic processes (Table 3.10).

Table 3.9: GO term enrichment for molecular function and biological process in cluster 3

GO Term	Aspect	Description	P-value	Sample Frequency	Background Frequency
GO:0003824	F	catalytic activity	3.63E-03	36.2%	25.4%
GO:0005488	F	binding	7.21E-03	43.1%	32.0%
GO:0003677	F	DNA binding	7.45E-03	14.1%	7.4%
GO:0042221	P	response to chemical stimulus	4.02E-08	15.7%	6.1%
GO:0009725	P	response to hormone stimulus	1.10E-06	9.4%	2.8%
GO:0009987	P	cellular process	6.62E-06	46.1%	31.7%
GO:0009755	P	hormone-mediated signaling pathway	7.88E-06	5.0%	0.9%
GO:0032870	P	cellular response to hormone stimulus	8.84E-06	5.0%	0.9%
GO:0009719	P	response to endogenous stimulus	9.17E-06	9.4%	3.0%
GO:0023033	P	signaling pathway	4.01E-05	7.7%	2.3%
GO:0071495	P	cellular response to endogenous stimulus	4.41E-05	5.0%	1.0%
GO:0010033	P	response to organic substance	5.65E-05	10.2%	3.8%
GO:0050896	P	response to stimulus	9.71E-05	21.3%	11.6%
GO:0050789	P	regulation of biological process	1.60E-04	18.2%	9.4%
GO:0050794	P	regulation of cellular process	1.80E-04	16.6%	8.2%
GO:0070887	P	cellular response to chemical stimulus	2.88E-04	5.5%	1.4%
GO:0065007	P	biological regulation	3.29E-04	20.2%	11.1%
GO:0009415	P	response to water	1.08E-03	3.6%	0.7%
GO:0009628	P	response to abiotic stimulus	1.19E-03	9.9%	4.1%

Continued...

GO Term	Aspect	Description	P-value	Sample Frequency	Background Frequency
GO:0071310	P	cellular response to organic substance	1.39E-03	5.0%	1.3%
GO:0008152	P	metabolic process	2.40E-03	40.3%	28.9%
GO:0009737	P	response to abscisic acid stimulus	2.64E-03	4.1%	0.9%
GO:0009266	P	response to temperature stimulus	4.01E-03	4.7%	1.2%
GO:0009738	P	abscisic acid mediated signaling pathway	4.11E-03	2.2%	0.3%
GO:0071215	P	cellular response to abscisic acid stimulus	4.11E-03	2.2%	0.3%
GO:0009414	P	response to water deprivation	4.24E-03	3.3%	0.6%
GO:0009723	P	response to ethylene stimulus	9.21E-03	2.8%	0.5%

Table 3.10: GO term enrichment for molecular function and biological process in Cluster 6

GO Term	Aspect	Description	P-value	Sample Frequency	Background Frequency
GO:0030528	F	transcription regulator activity	3.17E-07	15.8%	5.9%
GO:0003700	F	transcription factor activity	5.58E-06	13.9%	5.2%
GO:0003677	F	DNA binding	6.24E-05	16.5%	7.4%
GO:0016563	F	transcription activator activity	1.45E-03	3.3%	0.5%
GO:0005488	F	binding	1.75E-03	44.9%	32.0%
GO:0050896	P	response to stimulus	2.25E-05	22.8%	11.6%
GO:0048438	P	floral whorl development	5.50E-05	3.6%	0.4%

Continued...

GO Term	Aspect	Description	P-value	Sample Frequency	Background Frequency
GO:0048513	P	organ development	1.07E-04	8.9%	2.8%
GO:0048731	P	system development	1.09E-04	8.9%	2.8%
GO:0042221	P	response to chemical stimulus	1.73E-04	14.2%	6.1%
GO:0048569	P	post-embryonic organ development	2.31E-04	4.3%	0.7%
GO:0009908	P	flower development	2.78E-04	5.3%	1.1%
GO:0048437	P	floral organ development	2.93E-04	3.6%	0.5%
GO:0051171	P	regulation of nitrogen compound metabolic process	4.79E-04	12.5%	5.2%
GO:0019219	P	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	9.81E-04	12.2%	5.1%
GO:0045449	P	regulation of transcription	1.01E-03	11.9%	4.9%
GO:0048443	P	stamen development	1.13E-03	2.3%	0.2%
GO:0048466	P	androecium development	1.13E-03	2.3%	0.2%
GO:0009889	P	regulation of biosynthetic process	1.16E-03	12.2%	5.2%
GO:0031326	P	regulation of cellular biosynthetic process	1.16E-03	12.2%	5.2%
GO:0048653	P	anther development	1.23E-03	2.0%	0.1%
GO:0010556	P	regulation of macromolecule biosynthetic process	1.76E-03	11.9%	5.0%
GO:0031323	P	regulation of cellular metabolic process	2.86E-03	12.5%	5.6%
GO:0080090	P	regulation of primary metabolic process	3.09E-03	12.2%	5.4%
GO:0006350	P	transcription	4.23E-03	11.9%	5.2%
GO:0002252	P	immune effector process	4.38E-03	1.7%	0.1%

Continued...

GO Term	Aspect	Description	P-value	Sample Frequency	Background Frequency
GO:0009901	P	anther dehiscence	4.53E-03	1.3%	0.0%
GO:0010468	P	regulation of gene expression	5.30E-03	12.2%	5.5%
GO:0060255	P	regulation of macromolecule metabolic process	9.76E-03	12.2%	5.7%

The two clusters that exhibited the starkest fold differences in duplicate gene pair expression were clusters 4 ($n = 34$ pairs) and 5 ($n = 48$ pairs). As just noted, these clusters also have the fewest number of gene pairs in them, and this is probably due to their 2.5-5 fold differences in expression (purple and blue lines in Figure 3.11). As a result of their sample sizes, enriched GO terms were relatively limited when compared to the other, more populated clusters. Nevertheless, Cluster 4 was enriched for terms related to transcription factor activity and transcription regulator activity (Table 3.11), whereas cluster 5 was enriched for terms dealing with kinase or transferase activity (Table 3.12).

Table 3.11: GO term enrichment for molecular function and biological process in Cluster 4

GO Term	Aspect	Description	P-value	Sample Frequency	Background Frequency
GO:0003700	F	transcription factor activity	7.40E-05	24.2%	5.2%
GO:0003677	F	DNA binding	3.67E-04	27.3%	7.4%
GO:0030528	F	transcription regulator activity	4.06E-04	24.2%	5.9%
GO:0016802	F	trialkylsulfonium hydrolase activity	2.06E-03	3.0%	0.0%
GO:0004013	F	adenosylhomocysteinase activity	2.06E-03	3.0%	0.0%
GO:0016801	F	hydrolase activity, acting on ether bonds	6.18E-03	3.0%	0.0%

Table 3.12: GO term enrichment for molecular function and biological process in Cluster 5

GO Term	Aspect	Description	P-value	Sample Frequency	Background Frequency
GO:0008792	F	arginine decarboxylase activity	4.94E-03	2.1%	0.0%
GO:0008415	F	acyltransferase activity	7.28E-03	6.4%	0.6%
GO:0004672	F	protein kinase activity	8.28E-03	12.8%	2.9%
GO:0016747	F	transferase activity, transferring acyl groups other than amino-acyl groups	9.82E-03	6.4%	0.6%

4

Discussion – Project I

4.1 Intra-species comparative genomics reveals insights into paralogous gene evolution

It has been stated that expression divergence is the first step in the functional divergence between duplicated genes and is a determinant of their evolutionary fates (17; 90). Here, we have profiled the characteristics of upstream regulatory sequence conservation between 815 duplicated gene pairs in Arabidopsis that originated in a single polyploidy event 20-60 million years ago using three distinct methodologies in order to evaluate the effects of URS and CDS evolution on expression correlation (see Figure 4.1).

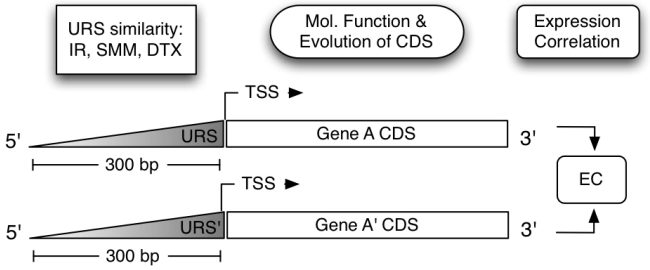

Figure 4.1: **Summary of our analyses.** - Depicted here are two paralogous genes, whereby each salient gene component is labeled according to the types of measurements performed upon them. Reflective of what was revealed in this study, the URS is shown as a diminishing gradient of 300 bp; darker at the transcription start site (TSS) and fading away in the 5' direction.

4.2 Traditional versus specialized methods for assessing URS similarity

Previously, it was reported in Arabidopsis that borderline significant similarity persists in URSs of WGD-derived duplicates and that this similarity was correlated to expression differences (40). However, the use of a simple alignment measure (DIALIGN2 (79)) likely underestimated the amount of conservation between duplicated URSs, and consequently, the correlations reported for URS similarity and expression correlation were probably also underestimated.

The outperformance of DIALIGN-TX (113; 114) by specialized programs (IR (43; 45) and SMM (21)) for measuring evolutionarily conserved regions in non-coding DNA sequences is not surprising (Table 3.1). This is partly explained by the nature of these specialized programs; both inherently check forward and reverse strands for similarity, whereas DIALIGN-TX considers only the forward strand in its alignments. Additionally, with the allowance for translocations and inversions of local regions of similarity, the specialized methods will be more sensitive at detecting conserved regions. These factors suggest that the failure to find stronger correlations in previous studies is due to the inferiority of traditional alignment-based approaches in measuring similarity in non-coding DNA.

4.2.1 Same sequences, different measures

Despite the high inter-program correlations observed between the SMM and the IR, their values for any individual sequence pair can be drastically different. For example, the sequence pair At3g49910 and At5g67510 (both code for 60S ribosomal protein) scored very differently with respect to the two programs. The IR value was 0.019 (median 300 bp anchored window = 0.007) and the SMM score was 0 (median 300 bp anchored window = 0.25). The number of exact matches shared between the two sequences reflects the difference in scores between the two programs. There are only 16 instances of exact matches of length greater than or equal to 8 bp, and no instances of length greater than or equal to 12 bp shared between At3g49910 and At5g67510. As mentioned in chapter 8, the SMM requires that an alignment score threshold be set ($L = 12$ for the 300 bp anchored window); consequently, our chosen L value excluded all alignments that did not have an exact match of at least 12 bp in these two sequences.

This scenario is likely repeated for other sequence pairs, giving rise to the observed differences in correlation values reported in Table 3.1.

4.3 Delimitation of regions of high similarity within paralogous URSs

One of the more subjective and arbitrary notions regarding the analysis of URSs is the definition what constitutes a biologically and statistically meaningful regulatory sequence space. More specifically, how long does a putative URS have to be in order to maximize the signal but minimize the amount of noise? What length is insufficient to acquire this optimal balance between sensitivity and specificity?

Our analysis extended previous work by revealing the most relevant location of the URS harboring the majority of shared sequence elements. Given our results for the sliding window analysis and considering only congruent levels of significance between the three programs and expression correlation ($p << 0.0$) (last three rows of Table 3.1), the most relevant region in the URS which is most strongly associated with duplicate gene expression falls within the first 300 bp of the TSS. It is possible that basal promoter elements will contribute to some of this signal; however, it is unlikely that basal promoter elements, which typically reside within the first 100 bp of the TSS (78) were solely responsible for the observed correlations, as significant correlations were observed well outside of this sequence range (Table 3.1).

Therefore, based on the patterns of similarity (Figures 3.2 and 3.3) and in lieu of the correlations reported in Table 3.1, duplicated URSs in Arabidopsis have diverged rapidly and have a very limited region of conservation restricted to the immediate vicinity of the TSS. More specifically, the sequence window which maximizes a biologically meaningful signal in Arabidopsis polyploidy derived duplicate URSs is about 300 bp.

4.4 Positional information on exact matches

We observed strong correlations between URSs and expression correlations in our data. Although these correlations are relatively strong compared to previous studies (40), they can only explain a minority of the shared variance between what occurs on the sequence level in the URS and what is actualized at the expression level. This is

to be expected since many factors can influence co- or divergent expression of duplicate genes, such as trans factors (136), possible stochastic epigenetic effects (125) and micro-chromosomal rearrangements in URSs. Therefore, by studying positional information on exact matches in paralogous URSs, we obtained a more comprehensive look at sequence evolution that was not previously seen on a genome-wide scale (3.3). Interestingly, longer exact matches tend to occur in homologous positions between query and subject sequences (Figures 3.5 and 3.6), whereas translocations and inversions of exact matches are more prominent in divergently expressed gene pairs (Figure 3.7). Regardless of length, the least prominent type of exact match arrangement is always "homologous inversion" or matches in homologous positions on opposite strands (green lines in Figures 3.5, 3.6 and 3.7). This suggests that this type of evolutionary event is unlikely to occur and may be unfavored as an adaptation for transcriptional regulation of duplicate genes. Moreover, the pattern in the upper and lower 25% quantiles based on EC (Figures 3.6 and 3.7) illustrates how micro-chromosomal rearrangements are an important consideration in studying what contributes to the expression divergence of duplicate genes.

4.5 Gene components appear to evolve independently

Intriguingly, while the evolution of CDSs is coupled to expression correlation, as is URS evolution (Table 3.1 and Table 3.3), neither of the two is themselves coupled. There is no evidence in the data to suggest that one is the consequence of the other, or that the order of events could be revealed. We hypothesize that this is because these regions face distinctly different selective constraints; that is, proteins have functions that are more constrained by form (implying function) and it is likely that they face heavier selective pressures against mutations, since most changes would probably be deleterious. On the other hand, URS regions are known to exhibit high plasticity, whereby due to their modular nature they are evolutionarily flexible (63). Furthermore, plant URSs possess the facility to tolerate a variety of evolutionary changes, such as insertions, rearrangements and other forms of mutation driven novelty that could therefore explain the evolutionary decoupling between CDSs and URSs (127) (see above). Nevertheless, these results suggest that different selective forces have acted on different components of the genes throughout their post- duplication history, depending on the genes functions.

Our GO analysis reveals that duplicates exhibiting conserved URSs and expression correlation tend to be associated with transcription factor activity/regulation, whereas those pairs with conserved CDS and EC tend to be involved in metabolic pathways or are structural proteins or enzymes. The incongruence between our results and those reported in (110) can be attributed to our data set being more diverse, e.g., we did not only consider gene pairs involved in oxidative stress response. With respect to the lack of correlation between K_s and EC (Table 3.3), it is evident that many of the duplicated genes have evolved at different rates, as evidenced by the large range of Ks and Ka values – a result in line with previous reports (35; 40).

Though the evolution of URSs and CDSs is uncoupled, both are intimately tied to the expression correlation of the paralogs. A three-way joint rank analysis on the IR, K_a and expression correlations revealed biases in molecular function for certain gene pair categories (Tables 3.5 and 3.6). The preponderance of terms associated with transcription factor activity for those gene pairs with relatively conserved URSs but diverged CDSs are in accord with previous reports on mammalian promoter sequences (56; 73). In contrast to this, the multitude of terms associated with enzymatic activity (Table 3.6) and primary metabolic processes (data not shown) for those gene pairs with non-conserved URSs but conserved CDSs, agrees with a report on mammalian housekeeping genes (32). This makes sense, because genes involved in primary metabolic processes are likely to be ubiquitously expressed in many tissues and consequently would not require the tight regulation of expression experienced by transcription factors. Taken together, the molecular functions of duplicate genes in Arabidopsis depend on the relative conservation profiles of either the URS or the CDS.

4.6 Arabidopsis paralogs are divergently expressed

While pearson correlation coefficients (r) can be used a measure of duplicate gene pair expression divergence, it is rather limited in scope and dimensionality. As a means to delve further into the relationships between URSs, CDSs and paralogous gene expression, we opted to construct a new data set based on expression information gathered from the Genevestigator database (137). These data allowed for the analysis of differential gene expression both in magnitude and developmental stages between Arabidopsis paralogs. The large number of differentially expressed paralogs (roughly 72%) is

striking and introduces new questions about the evolutionary forces acting to preserve duplicated genes in the dimension of gene expression. On relative terms, the least divergently expressed clusters (red and blue lines in Figure **??**) are enriched for primary biological processes, such as regulation of transcription, biosynthesis of macromolecules and response to stress. If these processes are considered as fundamental or "house keeping" processes, then it would make sense that drastic up- or down-regulation of one of the genes in a pair would be of minor consequence throughout the various developmental stages. However, one can not discount the possibility that averaged ratios of gene intensity between paralogs may nevertheless still be insufficient for capturing the subtleties involved in regulated gene expression. In contrast to these aforementioned clusters, the remaining four clusters contain gene pairs which exhibit an average fold difference 1.4-4.5 times in gene expression. This means that one copy of the gene pair is preferentially expressed over the other and that there is little overlap in their expression magnitudes. However, there are cases of decreased log_2 ratios, most clearly evident in the developed flower (Figure **??**). Four of the six clusters display a sharp reduction in their ratios during this developmental stage, which might suggest that many gene pairs become more coordinately expressed in the developed flower, but are relatively divergently expressed in the other flowering stages. Furthermore, these four clusters tend to be associated with more complex biological processes, such as organ and system development (Cluster 6) or involved in hormonal responses (Cluster 3). In contrast to the "house keeping" clusters, the greater differential expression of gene pairs in these clusters may have been favored by natural selection and could possibly point to subfunctionalization within the domain of gene expression as a means of duplicate gene retention.

Not surprisingly, there is little correlation between our one-dimensional measurements of various sequence properties (i.e., IR, SMM, EC, K_s, K_a and K_a/K_s) and the expression ratios in each cluster throughout Arabidopsis development. In fact, the averages of these diagnostics are nearly uniform, even with their varied sample sizes in each cluster (Table 4.1).

Table 4.1: Cluster means for various sequence diagnostics

Clusters	mean IR	mean SMM	EC	K_s	K_a	K_a/K_s
1	-0.016	0.274	0.374	0.954	0.160	0.171
2	0.009	0.255	0.357	1.042	0.170	0.169
3	0.012	0.291	0.399	0.969	0.150	0.171
4	0.008	0.224	0.385	1.009	0.141	0.148
5	0.009	0.282	0.362	0.974	0.141	0.154
6	0.012	0.298	0.402	1.011	0.165	0.166

Averages of the 300 bp window for the IR and SMM are shown, in addition to EC (pearson expression correlation) and measures of coding sequence substitution rates. Cluster sample sizes can be found in Figure **??**.

The lack of stark differences in the above diagnostics reflects the immense difficulty in ascribing clearly delineated sequence properties that govern differential gene expression between paralogs and instead suggests that a new perspective that includes sequence elements but also goes beyond them is warranted.

Nevertheless, taken all together, our results introduce new questions about the evolutionary forces acting to preserve duplicated genes in the dimension of gene expression.

What is the underlying reason for preferring the expression of one duplicate over the other in various developmental stages? What types of duplicates are co-expressed at similar or different levels throughout development? What factors contribute to the observed differential expression patterns seen in Figure **??**? How much of the differential expression pattern can be explained by changes in DNA sequence?

Such questions and their answers promise new insights into not only the regulation of gene expression between paralogs, but also the larger evolutionary forces acting to preserve them.

5

Introduction – Project II

Whereas the previous project concerned itself with an intra-species comparative genomics analysis, we now move onto an inter-species comparative analysis that spans 28 eukaryotic organisms.

5.1 Background

Pre-messenger RNA (pre-mRNA) splicing is a complex and critical molecular process that generates functional mRNA molecules via the precise removal of introns and ligation of exons and is an important gene regulatory step in eukaryotic gene expression (50; 99). Pre-mRNA splicing is carried out via a macromolecular protein complex known as the spliceosome, which contains five small nuclear ribonucleoprotein particles (snRNPs; U1, U2, U4/U6, and U5) and an immense number of auxiliary proteins [around 150 in animals (8; 28)] that act co-ordinately to catalyze the splicing reaction (23). Following the discovery that genes were comprised of coding exons and non-coding introns (24), it became evident that a single gene could give rise to multiple alternative mRNA transcript isoforms (37).

Alternative splicing (AS) of pre-mRNA is arguably one of the most important biological processes for expanding the eukaryotic proteome and might mitigate the apparent paradox between gene content and organismal complexity (39; 85). AS engenders more than one spliced mRNA isoform from a single gene by regulated selection of alternative splice sites (107), which typically give rise to four types of AS events: alternative 5' splice site choice, alternative 3' splice site choice, cassette-exon inclusion or skipping,

and intron retention (85). AS not only contributes to an increase in proteomic expansion (39), but also alters protein functionality (gain, loss or reduction in function), localization and may introduce premature termination codons leading to nonsense mediated decay (NMD) degradation of AS isoforms (107) (and references therein). Recent estimates based on high-throughput studies suggest that 95-100% of all human multi-exon genes undergo AS (94; 124), in contrast to the approximate 35% of multi-exon genes experiencing AS in plants (19; 123; 130).

Given the widespread prevalence of AS in eukaryotic lineages (61), what components contribute to its regulation? One pivotal family of splicing factors has stood out ever since their discovery in the 1990s: the serine/arginine-rich (SR) proteins (64; 132). The SR proteins were classified as a family based on their ability to restore splicing activity to splicing factor deficient cell extracts, their conservation across vertebrates and invertebrates (132) and their recognition by monoclonal antibody mAb104 (103). All SR proteins have a modular structure consisting of at least one N-terminal RNA recognition motif (RRM) and a variable length C-terminal domain rich in serine and arginine residues (the RS domain) (105). The RRM domains can recognize and bind to an array of mRNA cis-regulatory elements, albeit with specific yet degenerate RNA-binding specificities (105). The RS domain is required for essential SR protein function, but is highly intrinsically disordered, meaning that this domain exists in an ensemble of conformations in physiological conditions (46). However, by virtue of this disorder, RS domains are able to function as splicing activation domains by contacting the pre-mRNA directly to promote spliceosome assembly (47; 96; 104), foster protein-protein interactions (38), undergo heavy phosphorylation and dephosphorylation (thereby modulating interactions with other proteins or RNA) (112), and contain signals for nuclear localization and nucleocytoplasmic shuttling (18; 22).

Human SF2/ASF was the first SR protein identified (64), which was followed by the identification of the other classical SR proteins [SC35, SRp20, SRp75, SRp40, SRp55 and 9g8 (reviewed in (71))]. SF2/ASF (and the other SRs listed above) function in constitutive and alternative splicing (71). SF2/ASF facilitates 5' splice site recognition by promoting the recruitment of U1snRNP to the 5' splice site via interactions with U1-70K (38). SF2/ASF and both interact with U1-70K and U2AF35 to promote 3' splice site recognition via recruitment of U2AF65 to the 3' splice site (129). Engagement of the tri-snRNP complex U4/U6/U5 in addition to other proteins, including SRs, promotes

spliceosome assembly and permits the splicing reaction to occur (7) (and references therein). Besides their roles in constitutive and alternative splicing, SR proteins have also been implicated in mRNA export, RNA stability, nonsense mediated decay (NMD) and translation (7) (and references therein).

SR proteins have been found in all metazoans (132), in lower eukaryotes such as *Schizosaccharomyces pombe* (118) and *Trypanosoma cruzi* (98), and in plants such as Arabidopsis (74), rice (55) and maize (36). To date, plants possess the most SR proteins of any organism studied, with Arabidopsis encoding at least 19 SRs and rice encoding 24 (7). In addition to acting as regulators of AS, SR genes are also alternatively spliced. Recent studies in Arabidopsis indicated a six-fold increase in the SR gene transcriptome (15 SR genes giving rise to 95 distinct AS isoforms) in response to hormones and stresses (92), and extensive coupling of AS isoforms with NMD (93). Since SR genes are subjected to regulated AS in response to developmental or stress cues, they are most likely targets of multiple signalling pathways and may function as key components in the response to developmental and environmental signals (7).

As SR proteins are prominent players involved in spliceosome assembly, constitutive and alternative splicing of pre-mRNA transcripts, undergo AS themselves and are essential for proper gene expression, studying these master regulators within a comparative genomics context would allow for generalized inferences about SR gene evolution across multiple eukaryotic species. Much of the research focus has been on metazoan SR gene evolution and function, with ample studies conducted in human, drosophila and roundworm (c.f. (71)). However, in the plant kingdom the study of SR proteins and their AS events have either been restricted to a subset of plants e.g., Arabidopsis, rice and moss (53), and maize, pine and Chlamydomonas (60), or a subset of SR proteins, e.g., members of the plant specific RS subfamily or the RS2Z subfamily (60). Therefore, a comprehensive analysis which takes advantage of newly sequenced genomes of photosynthetic and non-photosynthetic eukaryotes to assess the inventory of SR proteins and updated expression data to measure the extent of their AS would contribute to our understanding of the evolution of SR proteins and their importance in generating transcriptome diversity.

By using existing and novel genome sequence data for phylogenetically diverse eukaryotes, we can address a series of questions about plant SR gene content and evolution. Specifically: i) do plants have a higher number of SR genes than other eukary-

otes? ii) how many SR gene families are truly plant specific? iii) is AS in plant SRs as widespread as in Arabidopsis? iv) what selective forces are acting upon SR genes? v) are SR genes alternatively spliced in all sampled organisms? vi) what are the most prevalent AS event types in SR genes? Vii) how do AS event types vary across SR sub-families? Viii) how is DNA methylation associated with AS?

To begin addressing these questions, we have mined SR genomic sequences, amino acid sequences and EST/cDNA sequences for 12 photosynthetic eukaryotes and 15 non-photosynthetic eukaryotes from publicly available databases. Tentative SR gene inventories for 10 of the 12 photosynthetic eukaryotes and 12 of the 15 non-photosynthetic eukaryotes were determined in this study. We show that the SR gene complement from these organisms falls into approximately 12 sub-families. Furthermore, it appears that it is a general characteristic of photosynthetic organisms to possess on average a larger inventory of SR genes than non-photosynthetic organisms. We go on to show that most SR genes in photosynthetic eukaryotes are under purifying selection, that paralogous SR genes in some photosynthetic organisms are divergently expressed throughout development and that alternative splicing of SR genes is a common phenomenon shared by the majority of eukaryotes analyzed here.

6

Results – Project II

6.1 SR genes comprise at least 12 sub-families

We undertook extensive database searches to acquire SR genomic, EST/cDNA and amino acid sequences for 27 different eukaryotic species that span a diverse array of lineages (Figure 6.1 and Table 6.1). We were able to retrieve the above sequences for 272 SR genes and using the amino acid sequences of the RRM regions, we carefully constructed a multiple alignment for use in gene-tree reconstruction (see section 8). The aim here was to consolidate the inventory of SR genes into robust sub-family classifications that have thus far never been shown at a multi-genomic scale.

Table 6.1: Organisms and databases used

Organism	#SRs	Reference	Database
Glycine max	26*+	AH	(89)
Populus trichocarpa	20	AH	(89)
Arabidopsis thaliana	19**+	(59)	(97)
Vitis vinifera	9	AH	(89)
Zea mays	22	AH	(67)
Sorghum bicolor	20*+	AH	(89)
Oryza sativa	24++	(53)	(91)
Selaginella moellendorffi	5*++	AH	(89)
Physcomitrella patens	13+++	AH	(89)
Chlamydomonas reinhardtii	5	AH	(89)
Chlorella vulgaris	3*	AH	(89)
Cyanidioschyzon merolae	2	AH	(86)

Continued...

Organism	#SRs	Reference	Database
Homo sapiens	11	(66)	(51)
Mus musculus	10	AH	(51)
Gallus gallus	10	AH	(51)
Xenopus tropicalis	11	AH	(51)
Danio rerio	14	AH	(51)
Branchiostoma floridae	11	AH	(58)
Ciona intestinalis	8	AH	(58)
Drosophila melanogaster	7	(81)	(29)
Anopeheles gambiae	6	AH	(51)
Aedes aegypti	6	AH	(51)
Caenorhabditis elegans	7	(72)	(42)
Schizosaccharomyces pombe	2	(118)	(48)
Dictyostelium discoidum	3+	AH	(34)
Plasmodium falciparum	3	AH	(6)
Phytophthora sojae	3	AH	(58)

Organisms are listed according to their groupings in Figure 6.1. **SR45a (116) was excluded; *these organisms may have more SRs than listed due to the exclusion of sequences that did not begin with methionine residues; +, number of SR45 genes included in these counts from preliminary gene-tree analyses; AH, analyzed here.

Using two maximum likelihood methods and one parsimony method, we inferred that there are at least 12 SR gene sub-families, with the 12 photosynthetic eukaryotes contributing to roughly 62% of the major groupings observed (green clades in Figures 6.2 and Additional File 1).

About 2% of the SR genes were unresolved in the gene tree analyses, which included taxa from the single celled eukaryotes *C. reinhardtii, S. pombe, P. sojae* and two multicellular eukaryotes, *C. elegans and B. floridae.* Sub-families were labelled according to pre-existing family nomenclature (SC35 (SFRS2), SCL, RS, RS2Z, 9g8/SRp20 (SFRS7/SFRS3), SF2 (SFRS1/SFRS9)), or by prominent SR genes populating a clade (SRp38 (SFRS13), SRp40 (SFRS5), SRp55/75 (SFRS6/SFRS4), RSZ, SF2(p), SRp54 (SFRS11). It should be noted that the clades RSZ and SF2(p) (consisting of only photosynthetic eukaryotes) have been considered as orthologous to the 9g8 and SF2/ASF sub-families, respectively (7). In the interest of highlighting expansions of SR genes in photosynthetic eukaryotes, we chose to designate these sister groupings separately. The parenthetical (p) is appended to the plant-enriched SF2(p) sub-family so as to

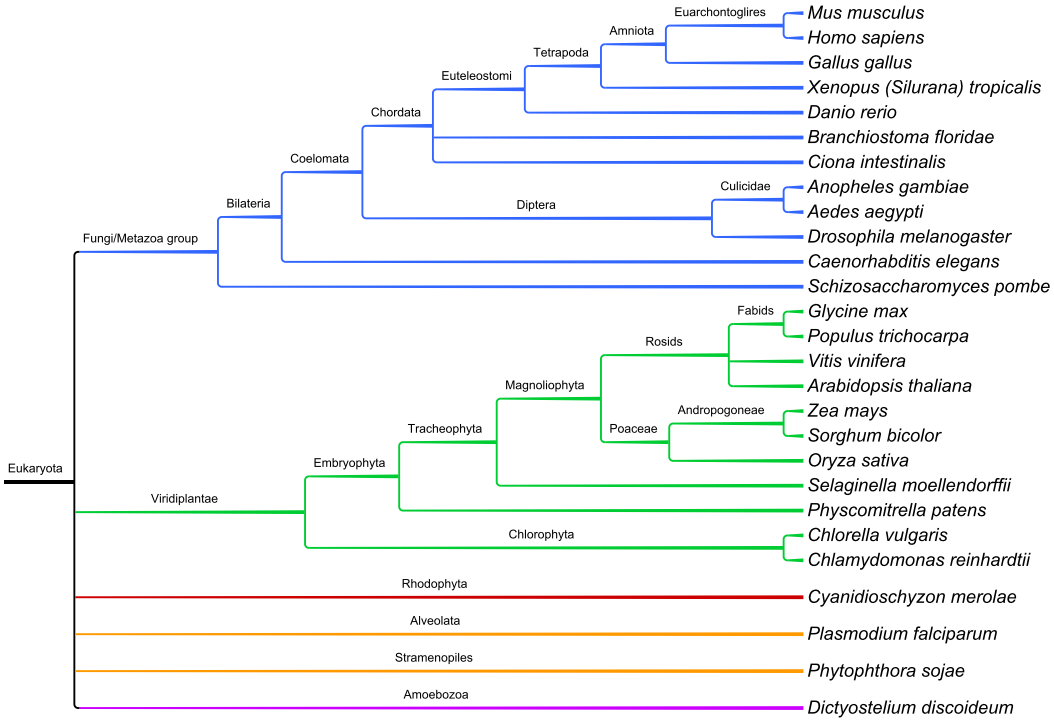

Figure 6.1: **Phylogeny of the 27 sampled organisms** - Phylogenetic positions were determined by using the NCBI taxonomy browser (9). Although the NCBI taxonomy browser is not an authoritative source for phylogenetics, for the purposes of illustrating the diversity inherent to the organisms sampled in this study, it readily describes the broad evolutionary relationships among them.

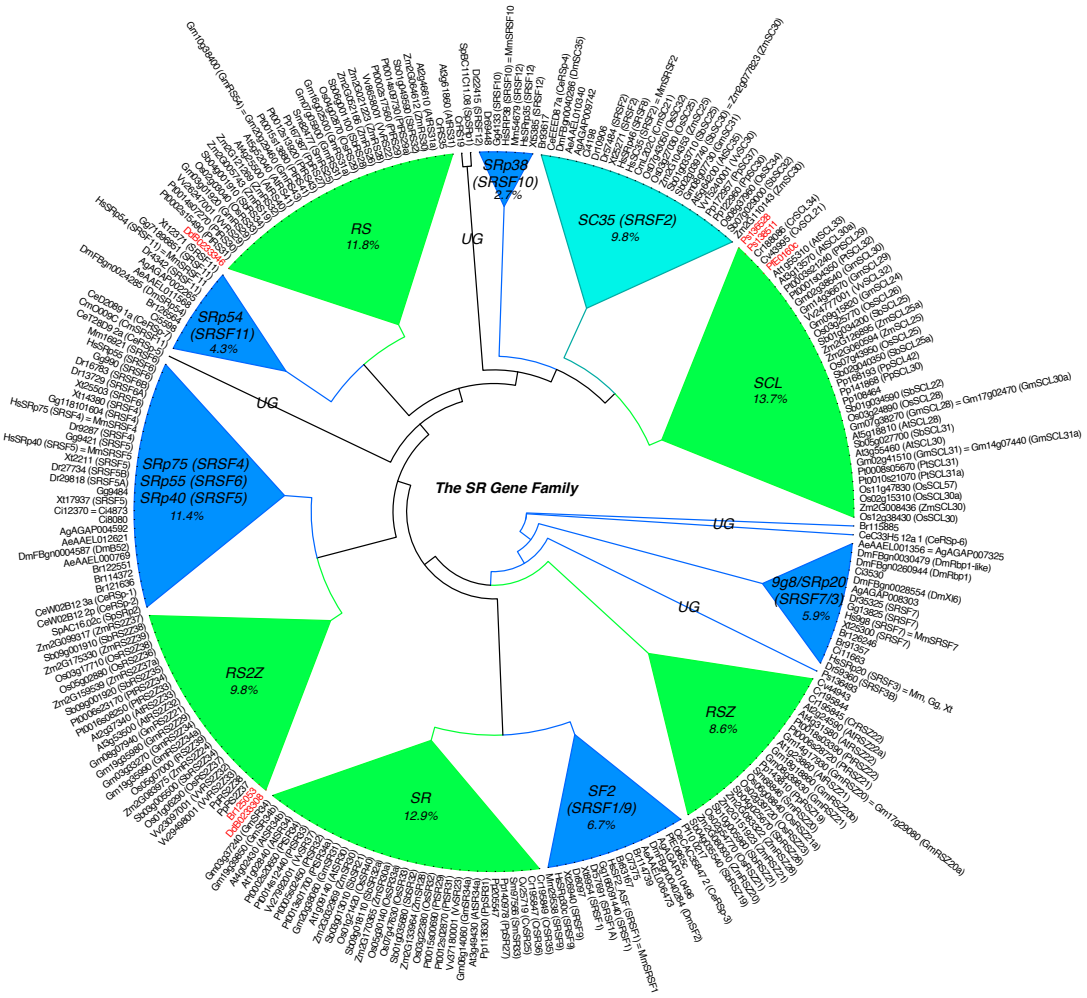

Figure 6.2: **Condensed SR gene family tree** - Schematic representation of the sub-family relationships among SR genes from the organisms sampled in this study. Green clades represent plant-enriched or plant-specific sub-families, whereas blue clades represent non-photosynthetic organisms. Taxa grouped into plant-enriched families that are non-photosynthetic are indicated in red. Species prefixes are as follows: *Gm, Glycine max; Pt, Populus trichocarpa; At, Arabidopsis thaliana; Vv, Vitis vinifera; Zm, Zea mays; Sb, Sorgum bicolour; Os, Oryza sativa; Sm, Selaginella moellendorfii; Pp, Physchomitrella patens; Cr, Chlamydomonas reinhardtii; Cv, Chlorella vulgaris; Cm, Cyanidioschyzon merolae; Hs, Homo sapiens; Mm, Mus musculus; Gg, Gallus gallus; Xt, Xenopus tropicalis; Dr, Danio rerio; Br, Branchiostoma floridae; Ci, Ciona intestinalis; Dm, Drosophila melanogaster; Ag, Anopheles gambiae; Aa, Aedes aegyptii; Ce, Caenorhabditis elegans; Sp, Schizosaccharomyces pombe; Dd, Dictyostelium discoideum; Pf, Plasmodium falciparum; Ps, Phytopthora sojae*; UG, ungrouped.

highlight the large number of photosynthetic eukaryotes populating this clade. In later sections, anytime a sub-family designation is followed by a parenthetical (p), only the photosynthetic members of the sub-family are under consideration.

6.2 No particular SR sub-family is broadly conserved across eukaryotes

If one follows our sub-family designations, there are ostensibly no SR sub-families shared across all of the sampled species (Figure 6.3). In photosynthetic organisms, with the exclusion of the red algae, *C. merolae* and the grapevine, *V. vinifera*, the only SR genes conserved are the one-zinc knuckle family, RSZ, and the SF2/ASF orthologous family, SF2(p) (Figure 6.3). However, if one excludes the algae from consideration (except for *C. reinhardtii*), all higher plants have sub-families RS, SF2(p), and RSZ in common. Further exclusion of the ancient lycophyte, *Selaginella*, results in the inclusion of SCL as one of the conserved sub-families among higher plant lineages.

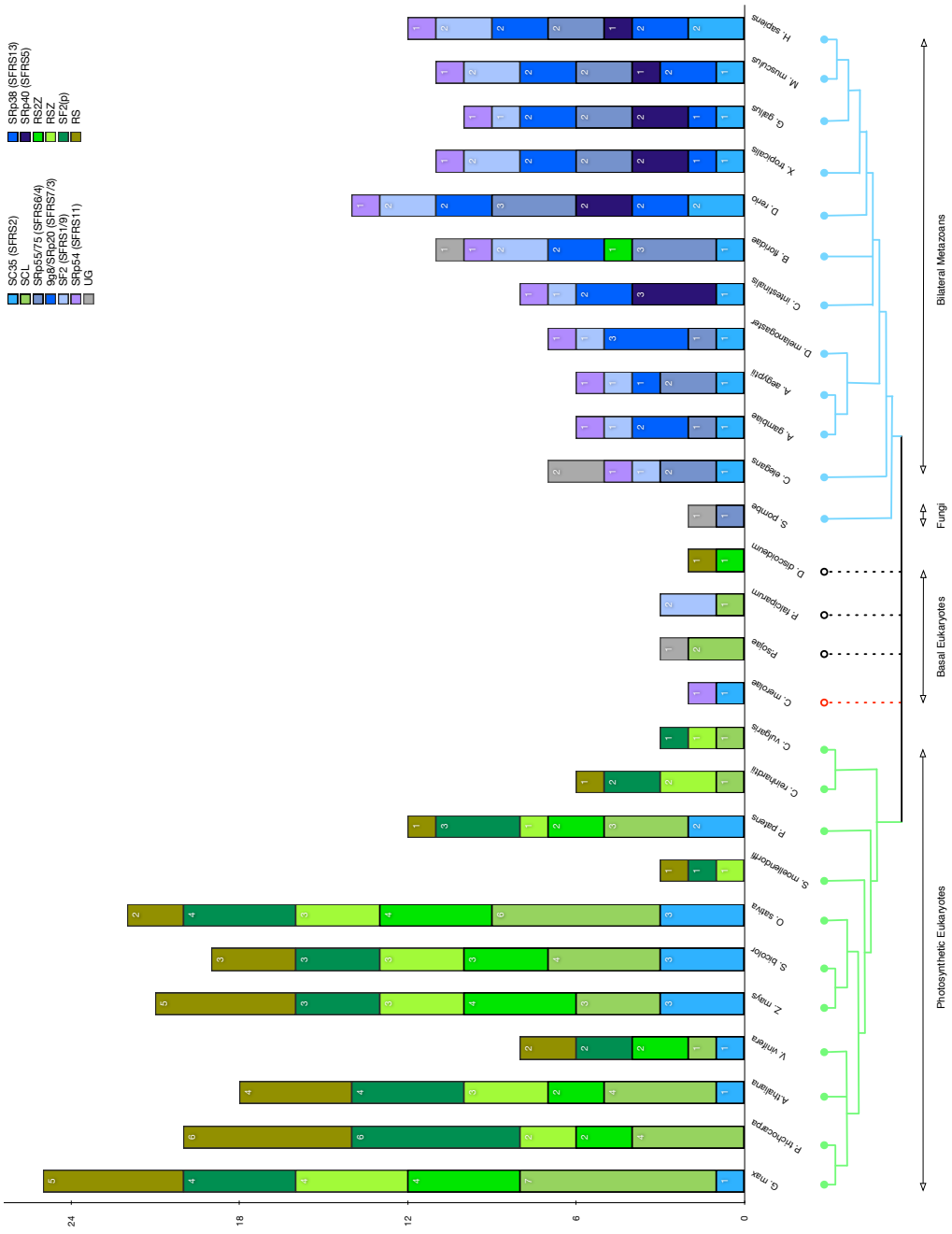

Figure 6.3: **Sub-family classification of SR genes** - According to the trees presented in Figures 6.2 and Additional File 1, we plotted the SR sub-families by organism. The inferred taxonomic grouping from Figure 6.1 is plotted below the bar chart and the number of SR genes per family is indicated by color codes as well as value labels.

6.3 SC35 (SFRS2) is likely an ancient SR gene

Though not shared across all eukaryotes, SC35 is present within eight of the photosynthetic organisms and all of the bilateral metazoans, but conspicuously absent from fungi and lower eukaryotes (Figures 6.3 and 6.4), except for *C. merolae*, the ancient red algae believed to have originated prior to the last common ancestor among plants, animals and fungi (111). The lack of SC35 in the fungi, lower eukaryotes and some of the multi-cellular plants is surprising, because SC35 is one of the core SR proteins that participates in 5 and 3 splice site recognition and interacts with U170-K and U2AF35 (129). However, in the photosynthetic eukaryotes it is likely that other SR proteins perform similar functions to SC35 thereby mitigating its loss in these genomes.

6.4 Five sub-families are vastly expanded in plants, with three of them plant-specific

Five particular sub-families contributed to the generally larger number of SR genes found in photosynthetic eukaryotes: RS, SF2(p), RSZ, RS2Z and SCL (with RS, RS2Z and SCL being plant specific; Figure 6.3). The RS sub-family (31 members) is unique to photosynthetic eukaryotes, except for a single SR protein from *D. discoideum* that also grouped into this family (Figure 6.5). Though this *D. discoideum* sequence possesses two RRMs, which is characteristic of RS family members, its relatively long branch (0.93, and indicated in red in Figure 6.5), long full-length sequence (737 aa) and modest bootstrap support values (36% RAxML, 23% Garli) call its grouping with the RS sub-family into question. Nevertheless, the hypothesis that this protein is indeed a distant member of the RS sub-family cannot be unequivocally disregarded. Bearing this in mind as a singular exception, the members of the RS sub-family are only present in the embryophyta and absent in the algal species, except for *C. reinhardtii*. Among the dicotyledenous plants, *P. trichocarpa* possesses the most RS sub-family members (six), whereas *V. vinifera* possesses the fewest (two) Figure 6.3). Interestingly, the low number of RS members in rice was not a characteristic feature among monocots (c.f. *Z. mays*, Figure 6.3). The moderately supported sister grouping of the SRp54 (SFRS11) clade (36% RAxML; Figure 6.5) suggests that the plant specific RS sub-family is orthologous to SRp54 (SFRS11) and therefore could be considered as plant-enriched

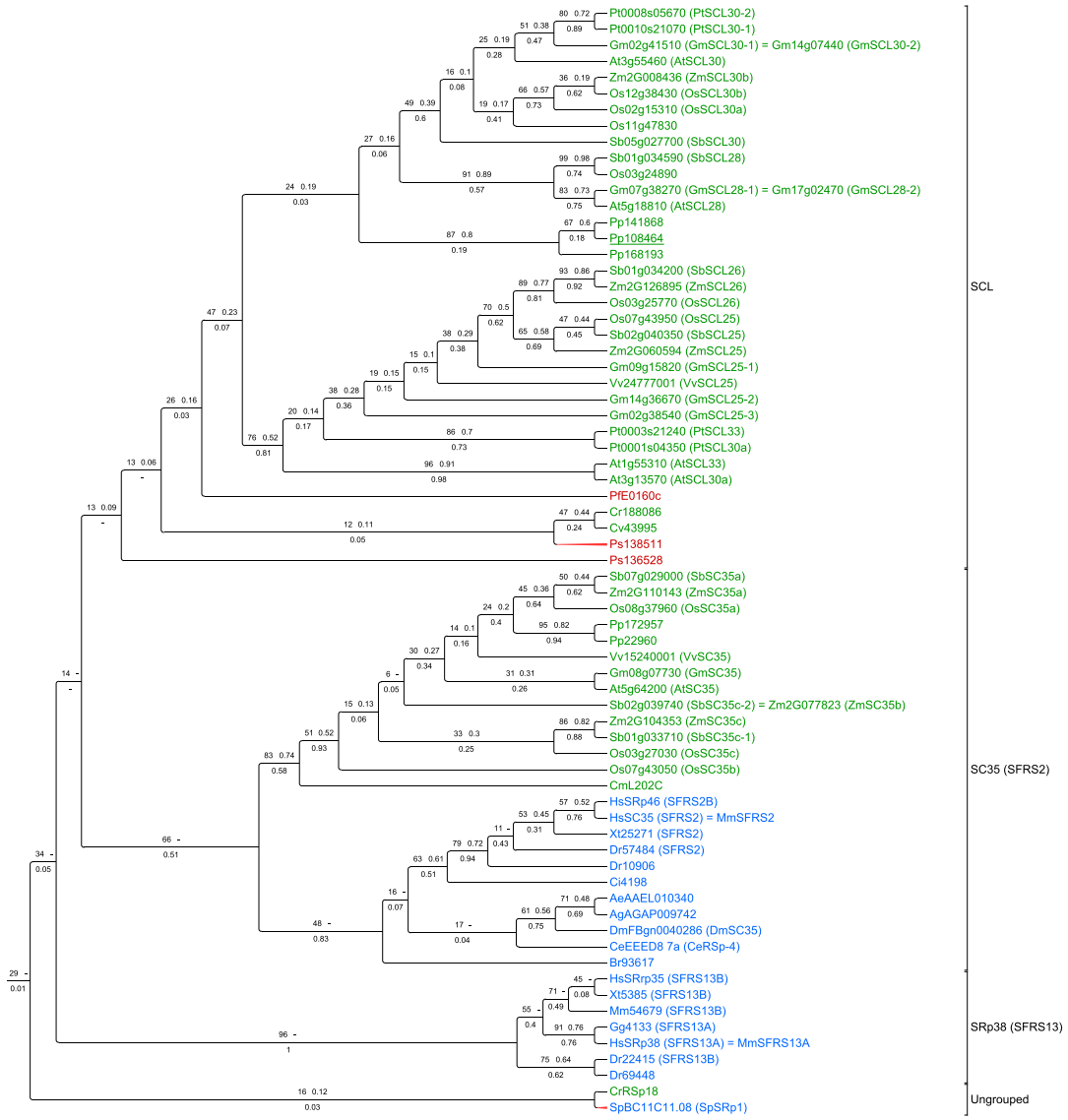

Figure 6.4: **Expansion of SCL, SC35 (SFRS2) and SRp38 (SFRS13) subfamilies** - The SCL and photosynthetic members of SC35 are shown in green, SRp38 (SFRS13) members are shown in blue. Plotted onto the branches are bootstrap support values from RAxML (top left), GARLI (top right) and maximum parisomony (bottom). The - symbols denote a lack of support for a particular grouping, which were typically from the parsimony analysis. If a sequence is followed by equality, it represents one or more other sequences that had exactly identical RRM(s) in the multiple alignment and were not included in the gene tree inference. Red branches indicate branch lengths greater than 0.75. The *P. patens* sequence is underlined because it contains a Zinc knuckle, whereas the remaining sequences do not (see text). Taxon labels use the same species prefixes as described in Figure 6.2.

rather than plant-specific, similar to RSZ and 9g8/SRp20 (SFRS7/3) or SF2(p) and SF2 (SFRS1/SFRS9) (Figure 6.2).

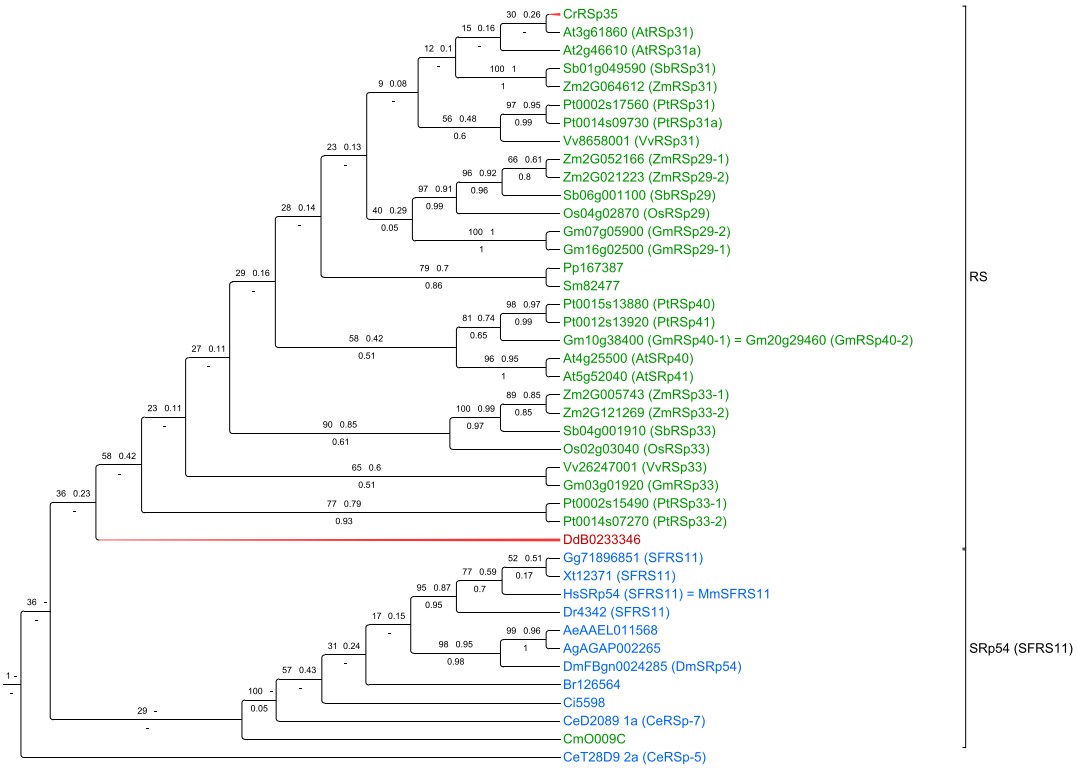

Figure 6.5: **Expansion of RS and SRp54 (SFRS11) sub-families** - RS (green) and SRp54 (blue) are shown in expanded form. Labeling conventions are as described in previous figures.

Another expanded plant-specific sub-family is the single RRM, two-zinc knuckle family, RS2Z (25 members) (Figures 6.2, 6.3 and 6.6). In contrast to the RS sub-family, RS2Z family members are restricted to the monocot and dicot lineages. In dicots, *G. max* has the most members (four) compared to Arabidopsis, *P. trichocarpa* and *V. vinifera*, which only have two members each. Each of the monocotyledonous organisms has four members (one member from *S. bicolor* was not officially counted because it did not pass our selection criteria; see section 8). Notably, one of the RS2Z members from *G. max*, GmRSZ33-4 (underlined in Figure 6), does not possess the dual zinc finger motifs characteristic of this sub-family and could therefore be grounds for exclusion from this sub-family. This could be an error in genome annotation, but it

should be noted that GmRSZ33-4 is relatively well supported by all three tree-searching methods (64% RAxML, 62% Garli, 60% parsimony).

Interestingly, two non-photosynthetic SR genes (one from *D. discoideum* and one from *B. flordiae*) grouped into the RS2Z sub-family with moderately weak support values and relatively long branches (DdB0233308 0.93 and Br125053 0.67; bootstrap support: 13% RAxML, 10% Garli, 27% parsimony). Notwithstanding the questionable support values, it should be noted that the *D. discoideum* sequence indeed possesses two zinc fingers and the *B. floridae* sequence possesses one zinc finger. When considering these additional features, it could very well be the case that the RS2Z sub-family is not a plant specific family, but rather, it is another ancient SR gene family that was lost in the Euteleostomi.

The largest plant-specific sub-family is the SCL family (containing a single RRM domain) with 37 members (Figures 6.2, 6.3 and 6.4). The family is present within the dicots, monocots and *P. patens* and the green algae, but absent from the remaining photosynthetic eukaryotes. *G. max* possesses the most SCL proteins (seven) among dicots, whereas rice possesses the most among the monocots (six). Interestingly, the bilateral metazoan conserved SRp39 (SFRS13) sub-family was a close sister group to the SCL sub-family (bottom clade in Figure 6.4. This similarity was previously acknowledged (7) as SRp38 members are splicing repressors. However, whether or not SCL proteins function as splicing repressors is an unanswered question. Strikingly, three sequences from *P. sojae*, a plant-attacking stramenopile also group into the SCL sub-family, albeit with long branches, poor bootstrap support or both (red taxa in Figure 6.4). Not only does this grouping of stramenopile sequences hint at the possibility of the SCL sub-family not being truly plant-specific, but also raises speculation into whether or not this evolutionary similarity is coupled to pathogenicity.

The remaining two sub-families, SF2(p) (33 members, Figure 6.7) and RSZ (23 members, Figure 6.8) are not plant-specific per se, but are orthologous to SF2/ASF (SFRS1/9) (Figure 6.7) and 9g8/SRp20 (SFRS7/3) (Figure 6.8), respectively. Orthology notwithstanding, these two families are greatly enriched in plants. *P. trichocarpa* contains six members of the SF2(p) sub-family, the most of any photosynthetic organism (Figure 6.3). As mentioned previously, SF2(p) is present in all photosynthetic lineages except for *C. merolae*, suggesting that this family was probably derived after the divergence of the red algae from plants and animals, but prior to the split of plants

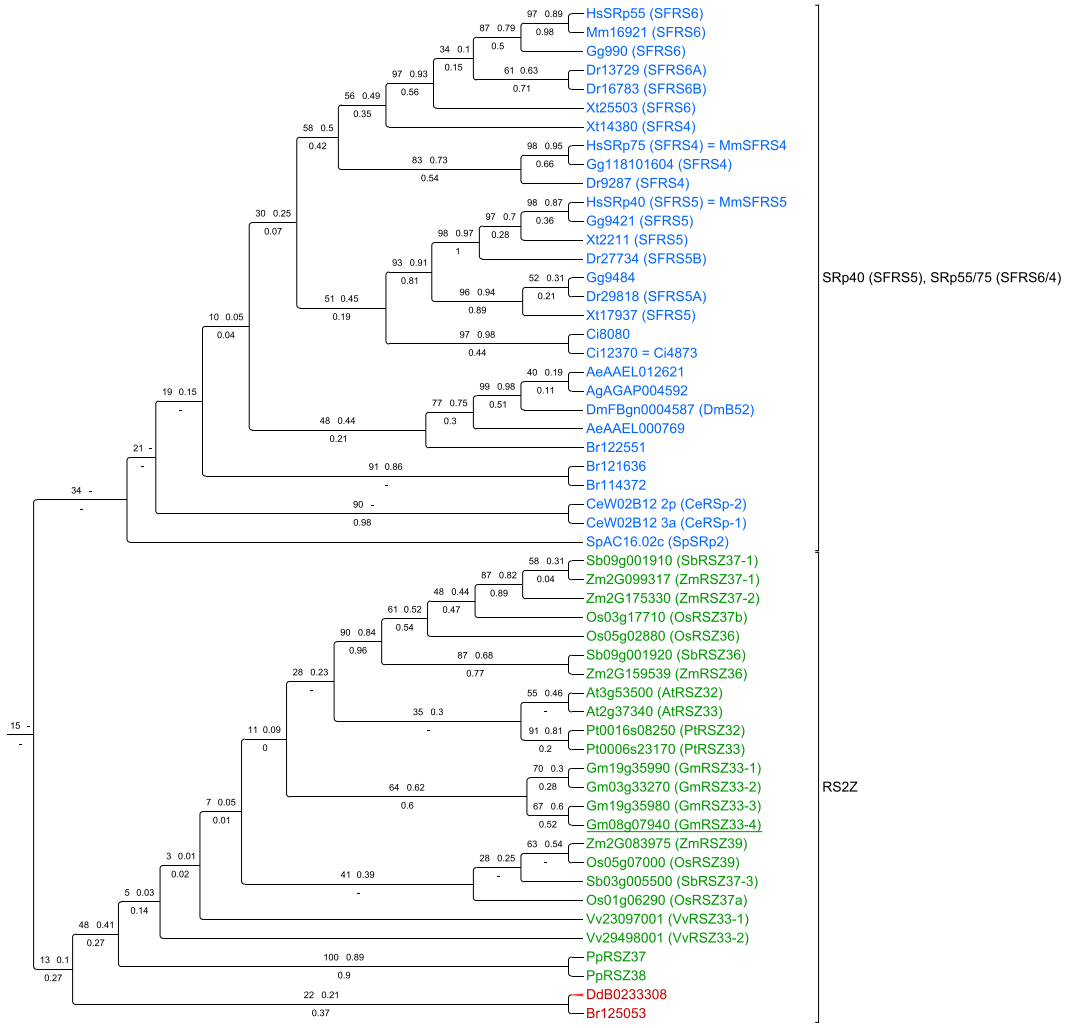

Figure 6.6: **Expansion of SRp40 (SFRS5), SRp55/75 (SFRS6/4) and RS2Z sub-families** - SRp55/75 (SFRS6/4) (top blue clade) and SRp40 (SFRS5) (middle and bottom blue clades) are shown in expanded form. The RS2Z plant-specific sub-family is shown in expanded form. A *G. max* sequence is underlined because it does not possess the canonical double Zinc knuckle domains characteristic of this sub-family (see text). Labeling conventions are as described in previous figures.

from animals. A similar situation is observed with respect to the RSZ sub-family: it is present in all photosynthetic eukaryotes (orthologous 9g8/SRp20 is present in all bilateral metazoans, as well), but absent in *C. merolae*, fungi and the other basal eukaryotes (dashed black lines in Figure 6.3).

6.5 Five SR sub-famiels are conserved across bilateral metazoans

Sub-families SRp54 (SFRS11), SF2/ASF (SFRS1/9), 9g8/SRp20 (SFRS7/3) SRp40 (SFRS5) and SC35 (SFRS2) are broadly conserved across the bilateral metazoans, whereas the remaining two families, SRp55/75 (SFRS6/4) (top blue clade in Figure 6.6 and SRp38 (SFRS13) (bottom blue clade in Figure 6.4) are only observed in the Euteleostomi (*D. rerio, X. tropicalis, G. gallus, M. musculus* and *H. sapiens*) [Figure 6.3]. Interestingly, *C. merolae* has a single member of the SRp54 sub-family that has moderate ML bootstrap support (29% RAxML) and branch length (0.73) (Figure 6.5). Therefore, a likely scenario is that SRp54 evolved prior to the divergence of plants from animals, but underwent several losses in multiple lineages. The 9g8/SRp20 sub-family also appears to have an early derivation given the sister grouping of a SR protein from *P. sojae* (Figure 6.8) as well as the zinc finger domain being shared between the plant-enriched RSZ sub-family.

6.6 Basal eukaryotes have the fewest SR sub-families

The lowest number of SRs was found in the basal eukaryotes (*P. sojae, P. falciparum, D. discoideum*), the algal species and the fission yeast, *S. pombe* (Figure 6.3 and Table 3.1). Each of these organisms, except for *C. merolae* and *D. discoideum* contained at least one SR protein that was not resolved in our gene tree analyses (Figures 6.2, 6.3 and Additional File 1). The low number of SR genes is likely a reflection of organismal complexity (single versus multi-cellularity) as well as the degree of multi-intron containing genes within a genome (e.g., only 43% of genes in *S. pombe* contain introns, of those only 25% have more than one intron, (128)).

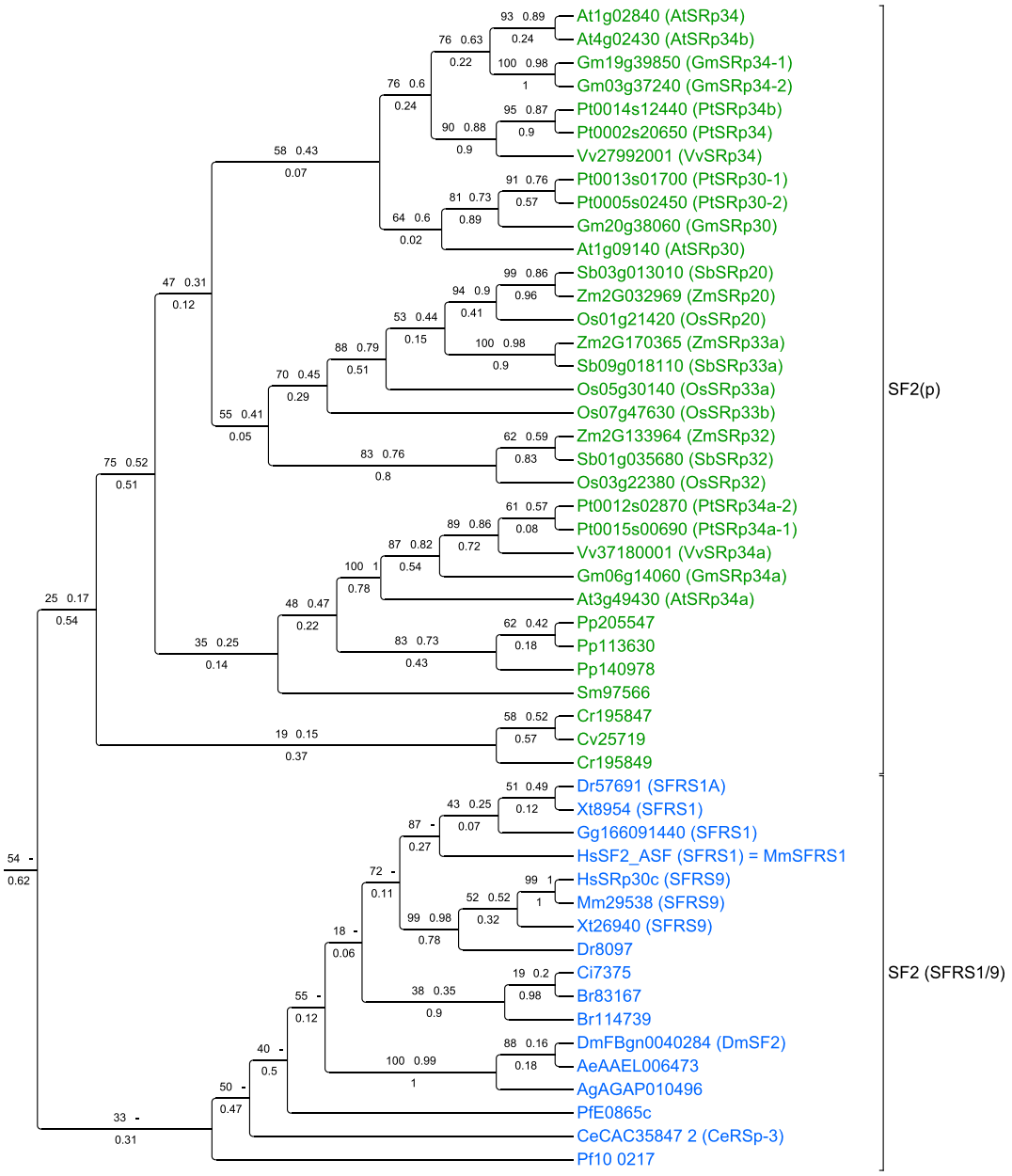

Figure 6.7: **Expansion of SF2(p) and SF2 (SFRS1/9) sub-families** - SF2(p) (green) and SF2 (blue) clades are shown in expanded form. Labelling conventions are as previously described.

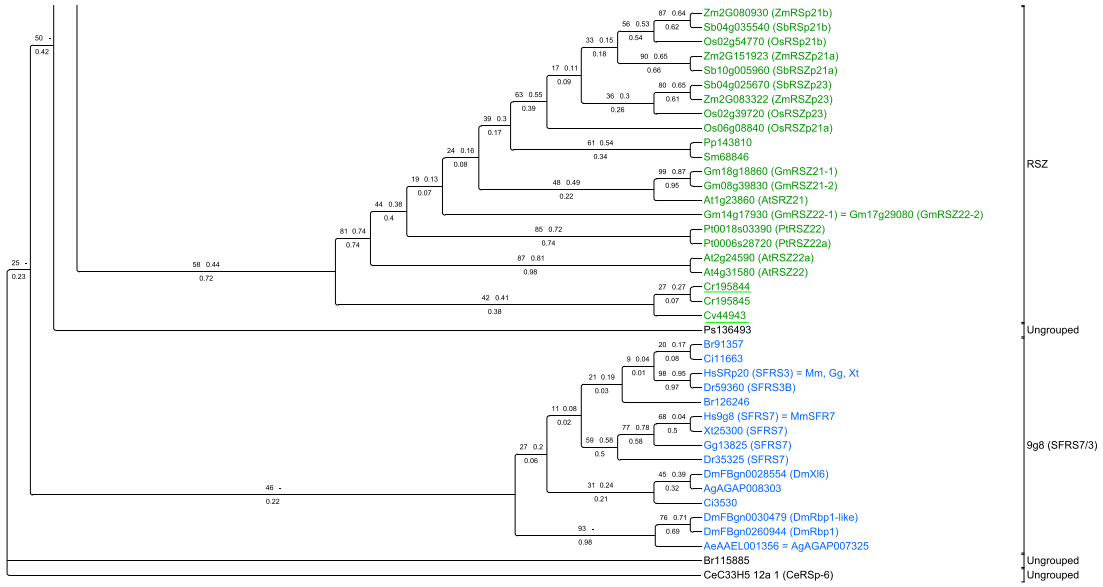

Figure 6.8: **Expansion of RSZ and 9g8/SRp20 (SFRS7/3) sub-families** - RSZ (green) and 9g8/SRp20 (SFRS7/3) (blue) are shown in expanded form. The two algal species are underlined because they do not possess the canonical Zinc knuckle domain that characterizes this sub-family. Labeling conventions are as described in previous figures.

6.7 RRM domains are highly collinear within sub-families and across species

Remarkably, the majority of SR proteins harbor RRM domains with identical start and end positions within the entire amino acid sequence (Additional File 2). Figure 6.9 distribution of N-terminal RRM start sites per SR sub-family. Only the conserved locations of the first RRM of the SR amino acid sequences are reported in Figure 6.9, even if there was more than one RRM in a particular sequence, whereas Additional File 2 shows the raw coordinates for all domains predicted by Interproscan (52; 82). Ten of the 13 groupings in Figure 6.9 are composed of SR proteins that have little to no inter quartile ranges in their RRM positions, with the most variance in the SCL, SF2 (SFRS1/9) and SRp54 (SFRS11) sub-families. However, the observed lack of variance in the remaining sub-families indicates that SR proteins are probably subjected to purifying selection to maintain domain organization, reflective of potential structural constraints necessary for splicing regulation (to be addressed in later sections). Further support for this idea was the occurrence of 17 RRM domains that were exactly identical in the multiple alignment within and across species (see the taxon annotations in Additional File 1). One particular example is the case of SR protein SRp20 (SFRS3) in *H. sapiens, M. musculus, G. gallus and X. tropicalis* (Figure 6.8). The identical nature of this RRM domain across four divergent taxa emphasizes the selective constraint on the molecular evolution of certain components of the SR protein.

6.8 Intron number is conserved within sub-families

To get an idea of how the number of introns varied within and across SR gene sub-families, we plotted the intron count of each gene using the longest transcript sequence as a reference from each organisms respective database (Figure 6.10). Intron number varied from a minimum of 0 (CeRSp-4, DdB0233749, AeEL000769 and Br122551) to a maximum of 16 (OsSRp20). The family with the highest median intron number (12) was that of SF2(p), whereas the two sub-families with the lowest median intron number (2) were the non-photosynthetic SC35 (SFRS2) members and SF2 (SFRS1/9). In general, sub-families enriched by photosynthetic eukaryotes had lower variability in

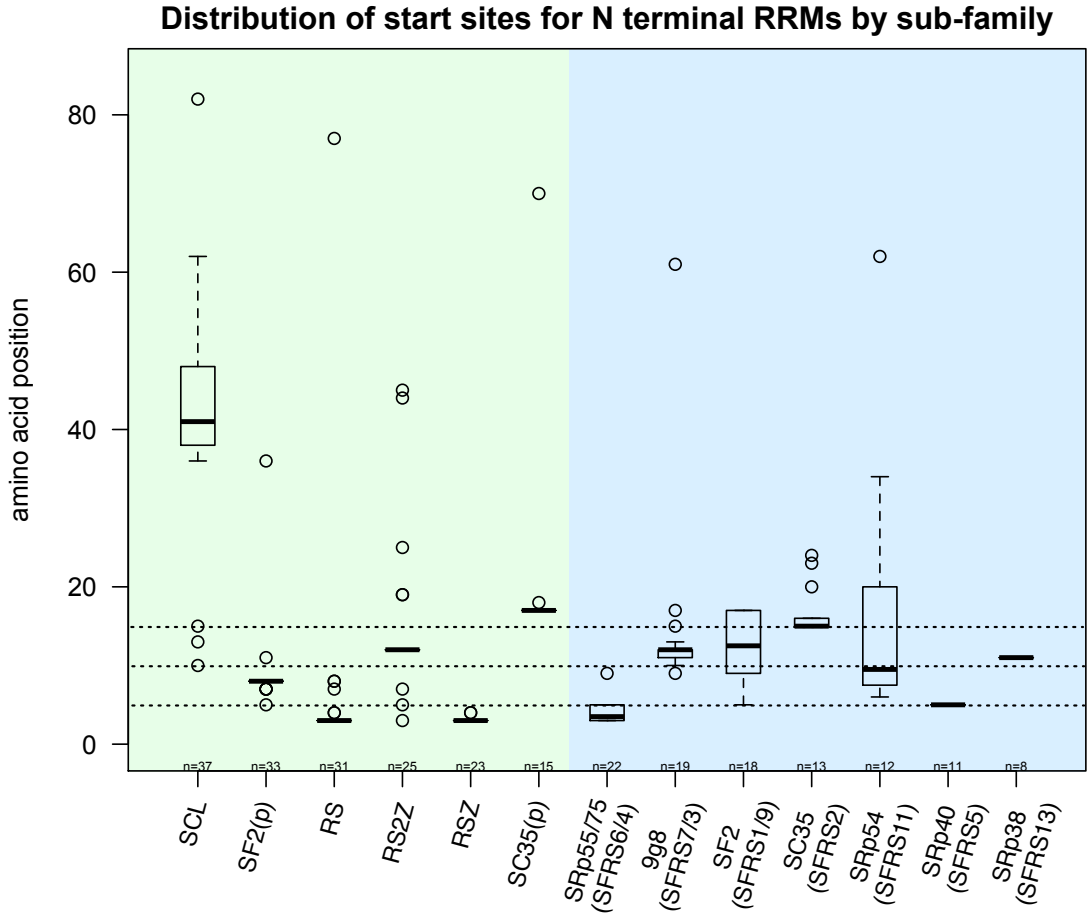

Figure 6.9: **Distribution of N-terminal RRM starting positions by SR gene sub-family** - N-terminal RRM start positions within the full amino acid sequence for each SR protein were organized by sub-family and plotted. The left half of the graph denotes the photosynthetic-enriched sub-families (green shaded area), whereas the right half denotes the non-photosynthetic organisms (blue shaded area). The three dashed horizontal lines from low to high demarcate positions 5, 10 and 15 respectively. Above the x-axis are the sub-family sizes comprising each box plot.

intron number. A striking example can be seen in the RS2Z sub-family, which has 24 members where 19 have an identical intron number of 6 (Figure 6.10).

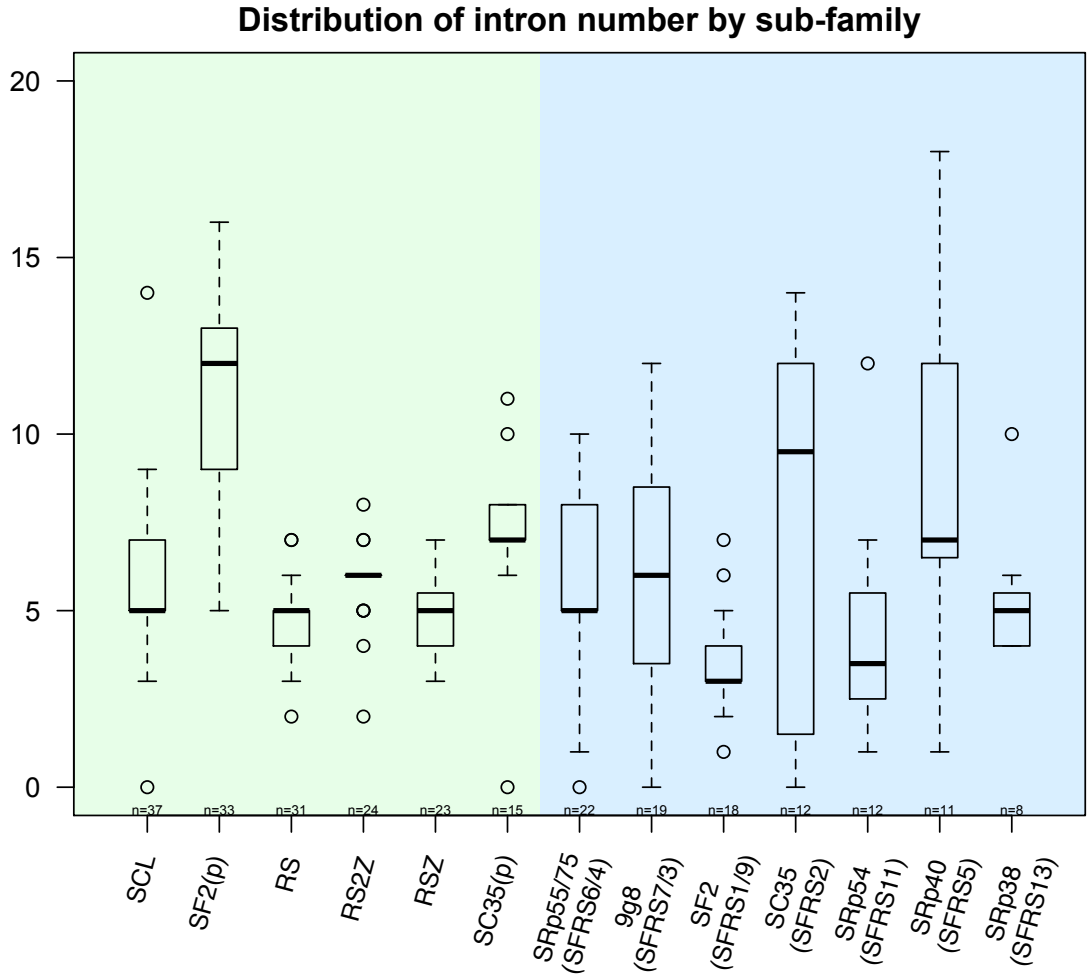

Figure 6.10: **Distribution of intron number by sub-family** - For each of the SR genes in each sub-family, the number of introns in the longest transcripts were assessed and plotted in a similar fashion to that of Figure 6.9.

6.9 RNA binding motifs are variable within RRM regions

In order to ascertain which residues within the highly conserved RRM regions of SR genes are involved in binding to mRNA molecules, we used the PiRanhA machine-learning web server to predict potential RNA binding residues (83; 108). Ten randomly

selected RRM sequences from each plant-enriched sub-family (RS, RSZ, RS2Z, SF2(p), SC35(p) and SCL) were submitted to the PiRanhA webserver for analysis. Boxes indicate potential amino acid residues implicated in RNA binding and motif regions are underlined in Figure 6.11. Interestingly, the majority of binding regions include highly variable positions within the RRM. Often, putative RNA binding residues are variable yet surrounded by a few highly conserved amino acid positions (Figure 6.11). In all sub-families, the first nine to 13 amino acids of the RRM are implicated in RNA binding and in the case of the RS and SF2(p) sub-families, the second RRM region is enriched for RNA binding regions.

6.10 SR genes in photosynthetic eukaryotes are mostly under purifying selection

As genome duplication has played a pivotal role in plant evolution (see Project I), we decided to investigate the impact of whole genome duplication on SR genes in the flowering plant lineages we sampled. Using the plant genome duplication database (`http://chibba.agtec.uga.edu/duplication/`) and following previously described methods (117), orthologous SR genes from Arabidopsis, *G. max*, rice, poplar, *S. bicolor* and *V. vinefera* were evaluated for their substitution rates, specifically, the ratio of the rate of non-synonymous to synonymous substitutions (K_a/K_s). Of the 132 orthologs analyzed from these species, only six genes (SbSRp33a, OsSRp33a; SbSRp33a, ZmSRp33a; SbSC35a, OsSC35a; SbSC35a, ZmSC35a) showed K_a/K_s ratios greater than 0.9 (red crosses in Figure 6.12). These results are in line with the high conservation of RRM start positions that were reported in section 6.7 and suggest that new substitutions in SR genes are most likely deleterious and selected against.

6.11 SR paralogs in photosynthetic eukaryotes are expressed at different magnitudes

To further investigate the influence of gene duplication in the SR gene family, we analyzed expression data for paralogous pairs in Arabidopsis, rice, maize and *S. bicolor*. For Arabidopsis, paralogous SR genes were determined by their groupings in Figure 6.2, and by referring to (11), whereas paralogy for the remaining plant species was based

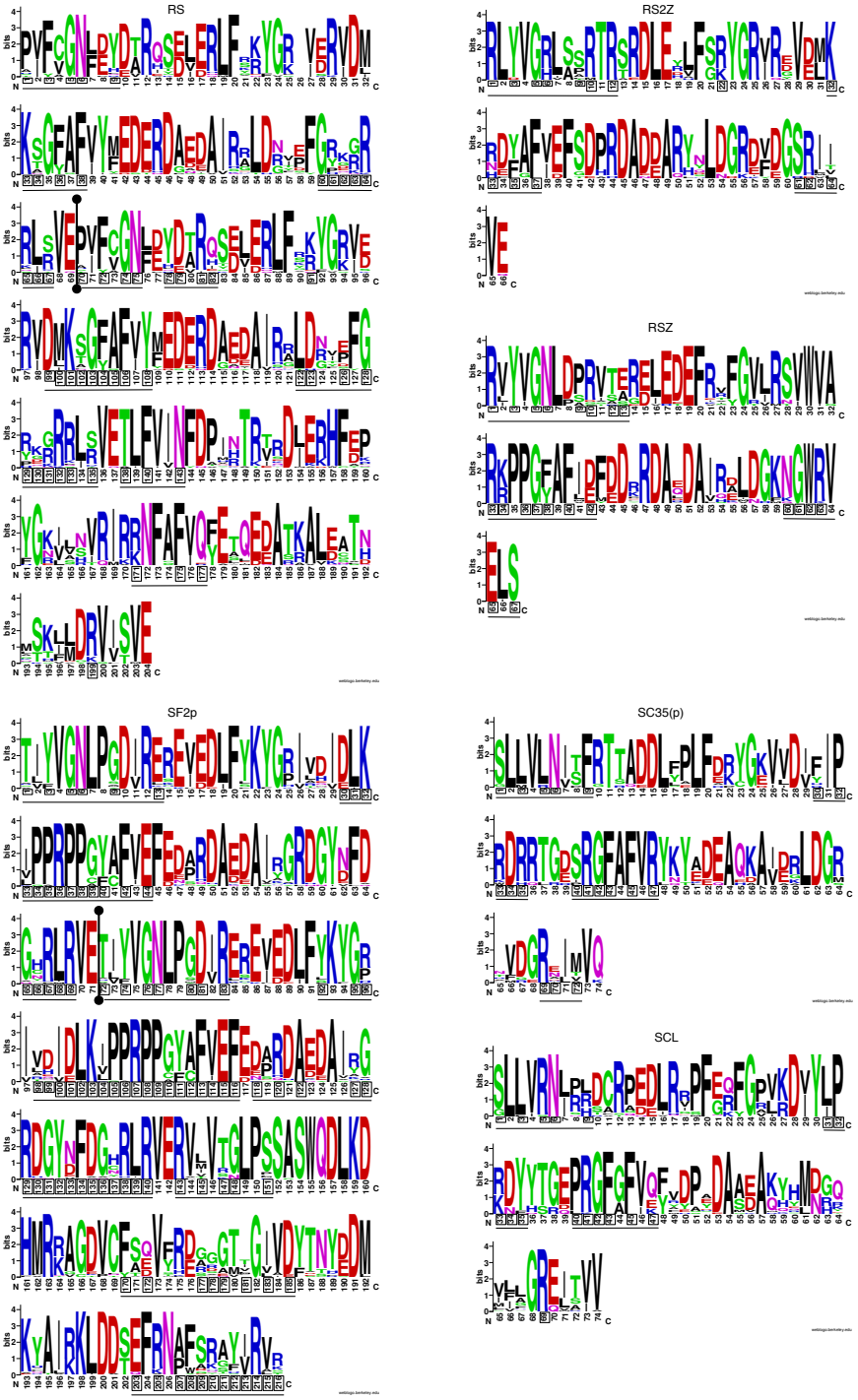

Figure 6.11: **RRM domain web logos for plant-enriched sub-families** - Web logos were created for each of the plant-enriched sub-families and putative RNA binding residues are indicated by boxes and underlined. Web logos were created by using the web logo server (26) and binding residues were predicted using the PiRanhA webserver (83; 108). In the cases of the RS and SF2(p) sub-families, the demarcation of RRMs is indicated by a vertical bar with circular endpoints.

Figure 6.12: **Orthologous pairwise K_a/K_s ratios for plant sub-families** - Pairwise comparisons of orthologous SR genes are shown. Ratios less than or equal to 0.1 are indicated by blue crosses, ratios in between 0.1 and 0.9 are shown as yellow crosses and ratios greater than or equal to 0.9 are depicted as red crosses.

solely on their groupings in Figure 6.2). Expression data for various developmental stages was extracted by using Genevestigator (137) and plots were generated for the paralogs.

In Arabidopsis, there are presumably six SR gene pairs and in every case in each developmental stage, none of the paralogs were expressed at the same levels (Figure 6.13). On average the fold difference in gene expression was around 1.5-2 times greater for one of the two genes in a pair, and sometimes as large as 7-12 times (see AtSRp34-AtSRp34b and AtRSp31-AtRSp31a in Figure 6.13). By contrast, the remaining Arabidopsis SR genes that do not exist as gene pairs and have overlapping expression patterns are depicted in Figure 6.14.

The pattern observed for the six Arabidopsis paralogs was also evident in rice, maize and soybean (Figures 6.15 and 6.16). There were only two cases of overlapping expression magnitudes, one during the stem elongation stage in maize for ZmSC35a and ZmSC35b and another case of overlap in the flowering stage in rice for Os01g72890 and Os05g01540 (members of the SR45 sub-family, which are only tentatively considered as bona-fide SR proteins).

6.12 Alternative splicing of SR genes is widespread

The next major component to our analysis of SR genes in eukaryotes was to assess the extent of alternative splicing (AS) among the organisms with sufficient EST/cDNA data. Of the 27 eukaryotes that were included in our phylogenetic analysis, 20 had enough expression information to be analyzed in our AS pipeline (Table 6.2; and see section 8 and online material: `http://combi.cs.colostate.edu/as/gmap/SRgenes/` for a description of the pipeline and resultant splice graphs). An example splice graph from our AS pipeline can be seen in Figure 6.17. While there were 20 organisms with sufficient expression information, the raw number of ESTs/cDNAs was highly variable between species (median EST/cDNAs per organism are shown in Figure 6.18). Therefore, we imposed a normalization procedure for measuring the extent of AS so that organisms would be comparable, similar to that of (61). We executed 100 resampling trials in triplicate of our AS pipeline requiring any given gene to have at least 15 ESTs/cDNAs. This procedure limited our dataset substantially, but conferred the

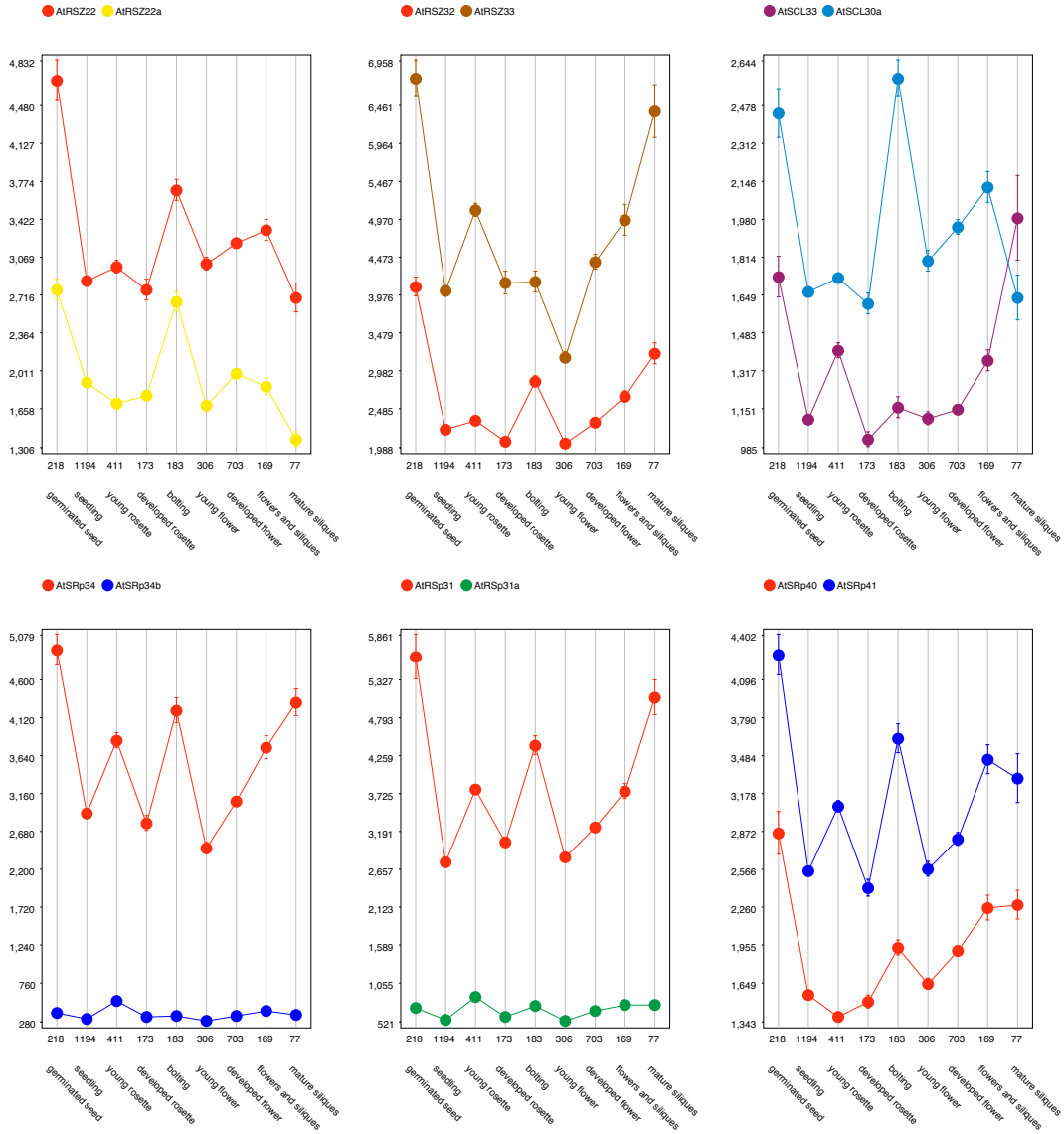

Figure 6.13: **Differential expression of Arabidopsis SR gene pairs** - Gene expression data for various developmental stages were taken from the Genevestigator database (137) and plotted for each of the six pairs of paralogous SR genes. The numbers below the x-axis indicate the number of microarray experiments that underlie the average intensity value plotted on the y-axis.

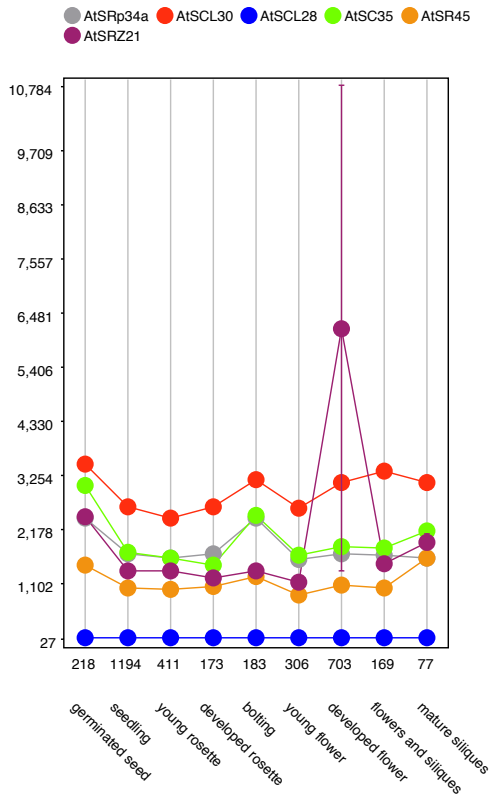

Figure 6.14: **Differential expression of non-paralogous Arabidopsis SR genes** - Gene expression data for various developmental stages were taken from the Genevestigator database (137) and plotted for each of the remaining SR genes. The numbers below the x-axis indicate the number of microarray experiments that underlie the average intensity value plotted on the y-axis.

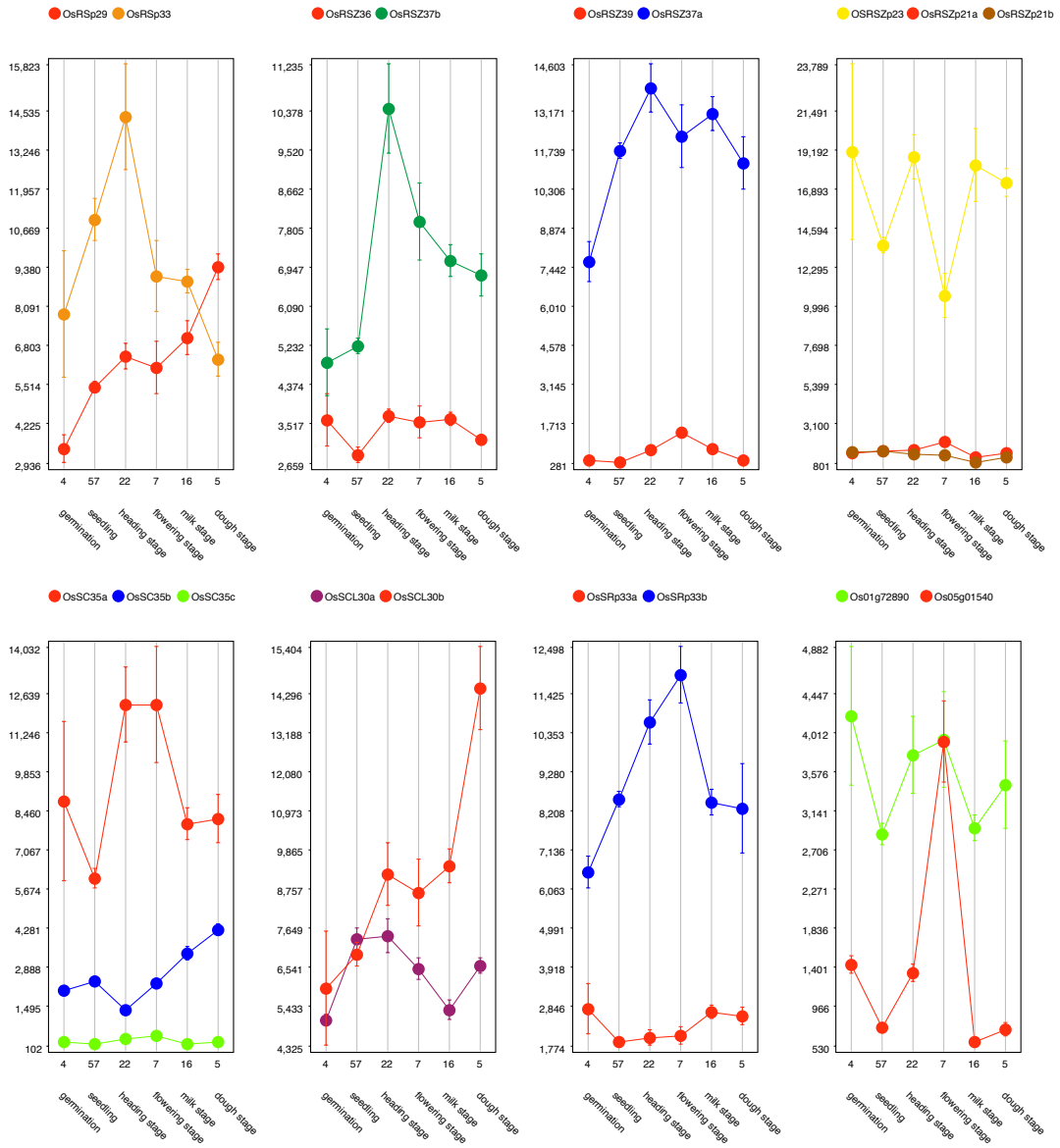

Figure 6.15: **Differential expression of paralogous rice SR genes** - Gene expression data for various developmental stages were taken from the Genevestigator database (137) and plotted for each of the pairs of SR genes. In some cases (fourth and fifth panels), there were three paralogs included. The numbers below the x-axis indicate the number of microarray experiments that underlie the average intensity value plotted on the y-axis.

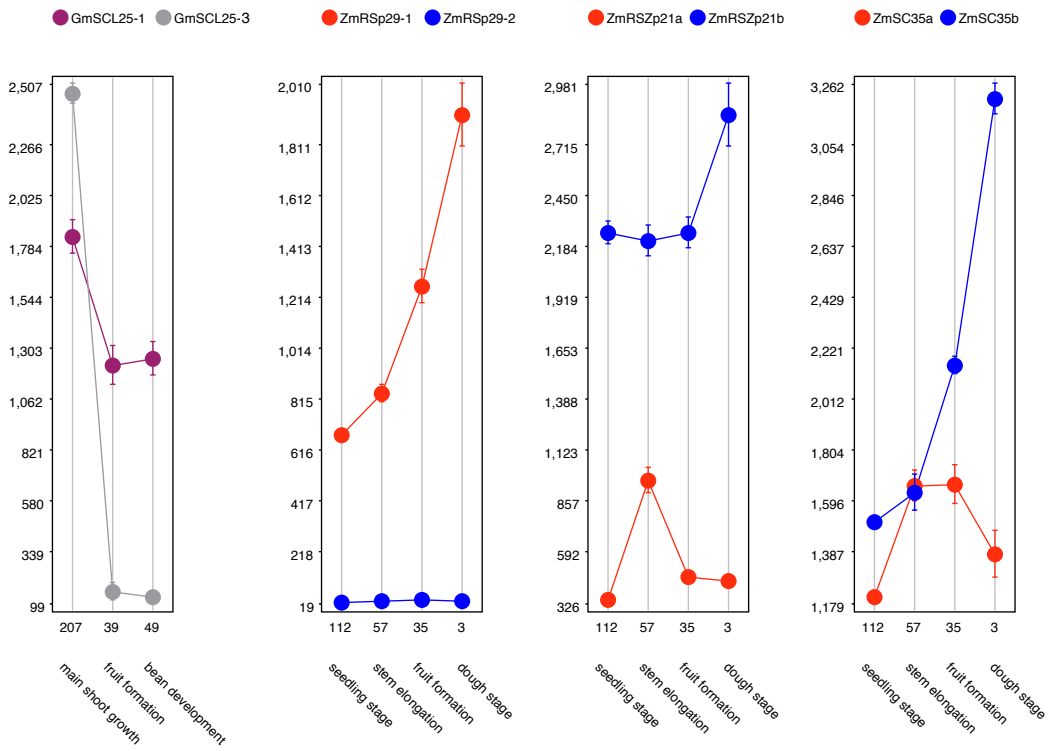

Figure 6.16: **Differential expression of paralogous soybean and maize SR genes** - Gene expression data for various developmental stages were taken from the Genevestigator database (137) and plotted for each of the pairs of SR genes. The numbers below the x-axis indicate the number of microarray experiments that underlie the average intensity value plotted on the y-axis.

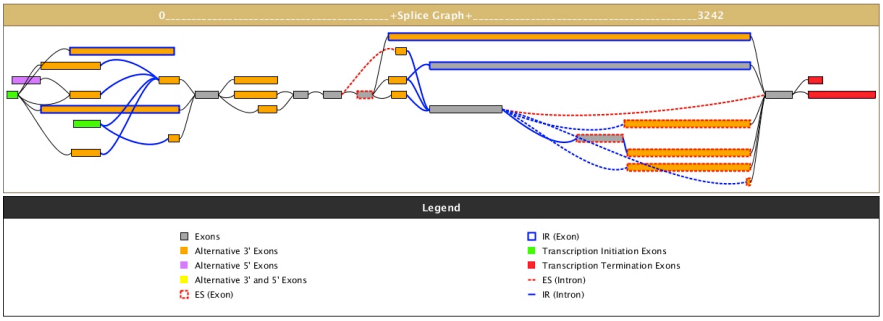

Figure 6.17: **Example splice graph for AtSRp34b** - Shown here is a typical splice graph from which AS event counts are taken.

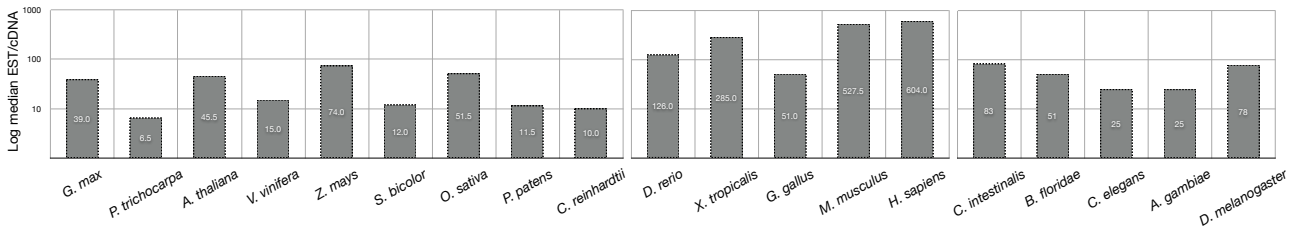

Figure 6.18: **Log Median ESTs/cDNAs per organism** - Log Median ESTs/cDNAs per organism.

ability to make comparisons across species. The non-normalized AS graphs are accessible from the website listed above and the non-normalized fraction of genes undergoing AS is presented in Table 6.2. Normalized fractions of AS for the three independent replicates are depicted in Figure 6.19.

Table 6.2: Alternatively spliced SR genes

Organism	Genes with AS	Total genes	Fraction
Glycine max	17	26	0.65
Populus trichocarpa	8	20	0.40
Arabidopsis thaliana	16	19	0.84
Vitis vinifera	6	9	0.66
Zea mays	21	22	0.95
Sorghum bicolor	11	20	0.55
Oryza sativa	20	24	0.83
Physcomitrella patens	8	13	0.61
Chlamydomonas reinhardtii	2	7	0.29
Danio rerio	12	14	0.85
Xenopus tropicalis	9	11	0.81
Gallus gallus	7	10	0.70
Mus musculus	10	11	0.91
Homo sapiens	11	12	0.91
Ciona intestinalis	7	8	0.87
Branchiostoma floridae	9	11	0.81
Caenorhabditis elegans	6	7	0.85
Anopeheles gambiae	5	6	0.83
Drosophila melanogaster	3	7	0.42
Aedes aegypti	1	6	0.16

Organisms are listed according to their groupings in Figure 6.1 and this table contains the non-normalized data from our AS pipeline. Though members of the SR45 sub-family were not included in our final gene-tree analyses, we nevertheless analyzed these genes for AS.

We observed negligible variance across each of the runs for most of the species, but it should be noted that some species have low sample sizes between 1-5 SR genes (due to the requirement that a gene have at least 15 ESTs/cDNAs for consideration). Bearing this in mind, the 100% AS of the single *P. trichocarpa* SR gene should not be considered reflective of the amount of AS in this organisms SR genes. Excluding

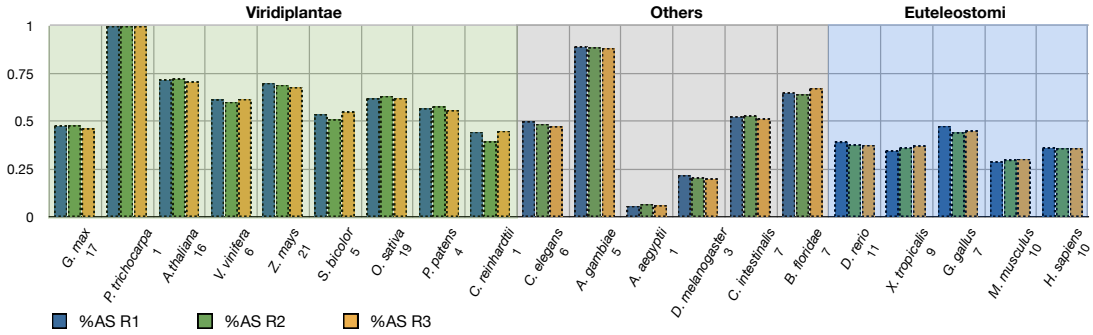

Figure 6.19: **EST/cDNA normalized %AS** - As detailed in the methods, we ran 100 trials in triplicate in order to compare alternative splicing evidence between SR genes from different organisms. The organisms are arranged from the Viridiplantae (green shaded area), to other eukaryotes (grey shaded area) and finally to the Euteleostomi (blue shaded area). Numbers below the taxon names indicate the number of SR genes that had at least 15 ESTs/cDNAs necessary for the normalization procedure.

those organisms that had only a single SR gene with at least 15 ESTs/cDNAs, all photosynthetic organisms (green shaded box in Figure 6.19) had greater than 50% of their SR genes undergoing AS, in contrast to the Euteleostomi (blue shaded box in Figure 6.19) that had AS percentages ranging from 30%-48%, while the other organisms had a much more variable range of %AS (grey shaded box in Figure 6.19).

We also measured the normalized average type of AS event, among five AS event types (IR, intron retention; SE, skipped exon; Alt 3, alternative 3 AS; Alt 5, alternative 5 AS; and Alt B, both 3 and 5 AS) per gene (Figure 6.20). Again, gene sample sizes should be taken into consideration when any comparisons are made and special attention given to those organisms that have extremely low sizes (i.e., *P. trichocarpa, C. reinhardtii, A. aegyptii*). Beginning with the Viridiplantae, Arabidopsis and maize had the highest incidence of intron retention events, with an average ranging from 0.84-1.94 events per SR gene (green shaded box in Figure 6.20). *V. vinifera*, Rice and *G. max* had the next highest incidence of IR, with *P. patens* having zero IR events but the highest average number of skipped exons (2.74 per SR gene) among all sampled organisms. Based on the available data, IR is not unilaterally the most prevalent AS type among the Viridiplantae. Instead, Alt 3, Alt 5 and SE events appeared to be just as prevalent or in some cases more prevalent (*G. max, V. vinifera, S. bicolor, O. sativa, P. patens*) than IR events.

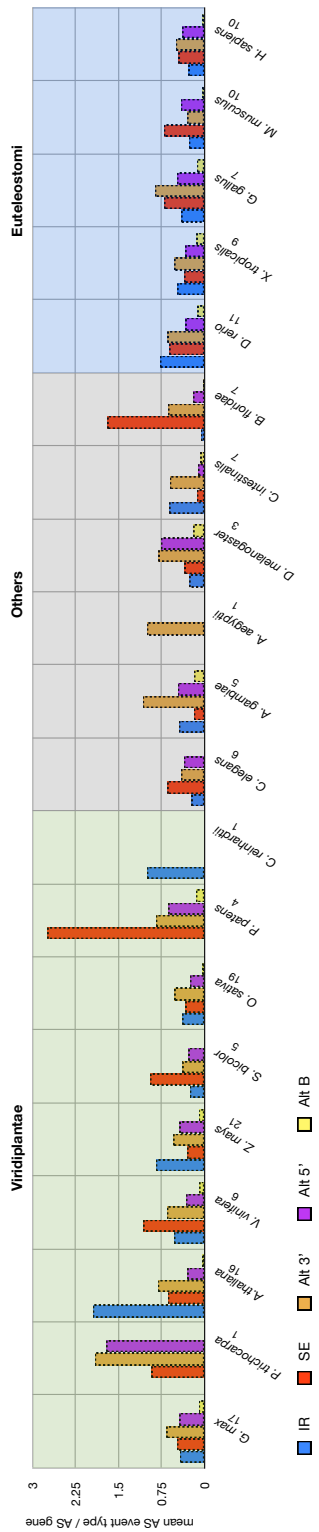

Figure 6.20: **AS event type prevalence by organism** - Based on the normalization procedure described in the section 8, five different AS event types were counted (IR, intron retention; SE, skipped exon; Alt 3, alternative 3; Alt 5, alternative 5 and Alt B, both Alt 3 and Alt 5 of the same intron). The y-axis shows the mean AS event type per gene experiencing AS in the normalization procedure. The arrangement of the shaded panels and numbers below the taxon names are similar to what is depicted in Figure 6.19.

Regarding the Eutelostomi (blue shaded box in Figure 6.20), Alt 3 AS events were generally the most prevalent followed by SE events and then Alt 5 or IR AS events. However, average IR events were the highest in *D. rerio*. This pattern is similar to what was observed in the Viridiplantae in the sense that there was no clearly preferable and broadly shared AS event type. Considering the final group of organisms (grey shaded box in Figure 6.20), the pattern of AS is reminiscent of the Viridiplantae and Euteleostomi, with considerable variance in average AS event types. Some organisms had a higher number of IR events (e.g., *C. intestinalis*), whereas others had a higher number of SE events (i.e., *C. elegans* [0.64] and especially *B. floridae* [1.69]).

Finally, we observed that in most organisms, Alt 3 AS (orange bars in Figure 6.20) was more prevalent than Alt 5 AS (purple bars in Figure 6.20) and that the simultaneous splicing of the 3 and 5 ends of introns was the least prominent AS event type (yellow bars in Figure 6.20). However, despite the differences in AS event type prevalence among these 20 organisms, AS of SR genes appeared to be a broadly shared characteristic among multiple eukaryotic lineages.

6.13 AS event types vary by sub-family

We next investigated how the percentage of AS and AS event type differed across the various SR sub-families. Using the classifications obtained from our gene tree analyses, the normalized measurements of family-wise %AS were calculated (Figure 6.21A). All photosynthetic sub-families (green shaded box in Figure 6.21A) were observed to have between 57%-88% of their SR genes experiencing some type of AS, in contrast to the non-photosynthetic sub-families (blue shaded box in Figure 6.21A) where the range was between 40%-54%.

For each sub-family, we also calculated the normalized AS event types (Figure **??**Figure20B). As was previously mentioned above, the occurrence of both Alt 3 and Alt 5 (Alt B) splicing of an intron was the least prevalent type of AS event and was also evident in the family-wise comparisons (yellow line in Figure 6.21B). The highest average number of Alt B events was observed in the RSZ sub-family (0.17 events per SR gene), followed by 9g8/SRp20 (SFRS7/3) and SRp54 (SFRS11) (0.09 events per SR gene, respectively). The sub-family with the highest amount of IR events was the plant-specific RS group (0.83 events per SR gene), whereas the family with the lowest

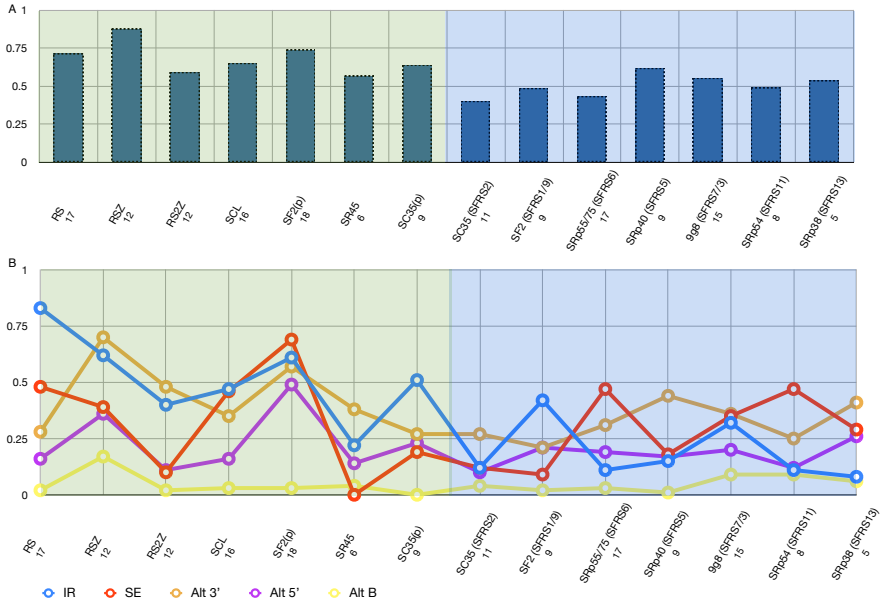

Figure 6.21: **Family-wise AS comparisons** - Panel A depicts the normalized proportion of genes undergoing AS per sub-family by averaging the values across the 100 trials in triplicate. Shading conventions are as previously described. Panel B shows the mean AS event type per gene experiencing AS in the normalization procedure but according to sub-family rather than organism (c.f. Figure 6.20). The Viridiplantae sub-families are shaded in green whereas the others are shaded in blue. The numbers below the sub-families designate the number of genes with AS in that particular sub-family.

amount of IR was SRp38 (SFRS13) (0.08 events per SR gene). Note that as the graph transitions into non-plant enriched sub-families (blue shaded area in Figure 6.21B), there was a tendency for the incidence of IR to decrease while SE events increased. The plant-specific SCL, RS and plant-enriched SF2(p) sub-families had SE events ranging from 0.46 to 0.69 events per SR gene, whereas the other plant-enriched sub-familes had much less SE events. Additionally, as previously stated, in nearly all sub-families, the incidence of Alt 3 AS was more frequent than Alt 5 AS.

6.14 DNA methylation is linked to alternatively spliced regions in Arabidopsis SR genes

Due to the observed differences in expression levels between paralogous SR genes in Arabidopsis, we checked to see whether Arabidopsis SR genes might be influenced by DNA methylation. We used the methylation data obtained from (134) (the first DNA methylation map of an entire genome) and cross-referenced the Arabidopsis SR genes against their database. Interestingly, 11 of the 19 SR genes are methylated in their coding sequences in wild type plants (Methylome track in Figure 6.22). Furthermore, paralogous SR genes tended to be methylated in different regions of their coding sequences (e.g., AtSRp34 and AtSRp34b, Figure 6.22) or one of the two paralogs was methylated and the other wasnt (e.g., AtSCL30a and AtSCL33, Figure 6.22). Even more striking was the localization of methylation as it relates to alternative splicing. It was readily apparent that methylated regions coincide with alternative spliced regions (Figure 6.23).

Figure 6.22: **DNA methylation in Arabidopsis SR genes** - SR genes were queried against the Arabidopsis methylome database (`http://signal.salk.edu/cgi-bin/methylome`) for methylation patterns. Strikingly, 11 of the 19 genes harbored methylation in their coding sequences. Paralogous SR genes are arranged adjacent to each other and labeled in bold.

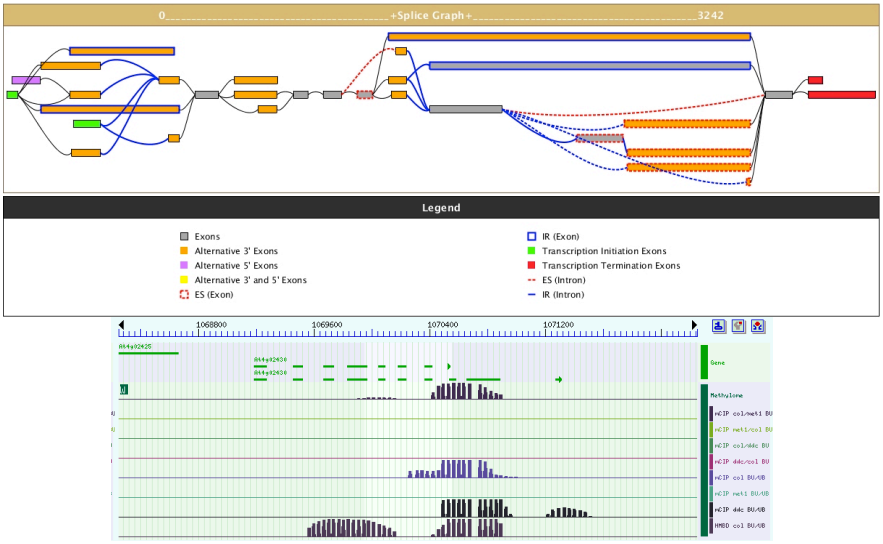

Figure 6.23: **DNA methylation in Arabidopsis SR34b** - The splice graph for AtSR34b is shown in the upper portion of the figure and below the graph, areas of DNA methylation are depicted. Note that the extensive alternative splicing in the 3 region of the splice graph coincides directly with areas of DNA methylation.

7

Discussion – Project II

The large-scale search and recovery of SR genes across 27 diverse eukaryotic species and the gene-tree analyses performed to classify them is the first of its kind. The work presented here provides a multi-genomic perspective on the similarities and differences in these key splicing regulators in terms of their gene number and their assortment into sub-families allowing for specific questions on SR gene evolution to be addressed.

7.1 The SR gene family is large and diverse

In order to understand the underlying evolutionary relationships between SR genes among a diverse set of eukaryotes, one must begin with phylogenetic or gene-tree analyses. Only by performing such a rigorous analysis is it possible to address the fundamental questions of how many sub-families comprise the greater SR gene family, and whether these sub-families are populated by diverse species from multiple domains of life or if they are instead restricted to lineage-specific groupings.

There are at least 12 SR gene sub-families, five of which are highly enriched by photosynthetic eukaryotes (RS, RSZ, RS2Z, SCL and SF2(p), six highly enriched by metazoan organisms (9g8/SRp20 (SFRS7/3), SRp38 (SFRS13), SRp40 (SFRS5), SRp55/75 (SFRS6/4), SF2 (SFRS1/9) and SRp54 (SFRS11), a single sub-family populated by both metazoans and plants (SC35/SFRS2), and a few ungrouped sequences (see Figure 6.2). Interestingly, the ungrouped organisms are primarily from the unicellular eukaryotes and their failure to fall into specific sub-families may be a reflection of their unique life histories. For example, putative SR proteins from the fission yeast, *S. pombe*

and the stramenopile, *P. sojae* fall into questionable sister groupings either adjacent to SRp38 (SFRS13) or sister to the 9g8/SRp20 (SFRS7/3) sub-family, respectively, with either long branches (in the case of SpSRp1) or lack of additional characteristic sub-family domains, such as the zinc finger domain (in the case of Ps136493). However, in a previous study, the two yeast proteins, SRp1 (UG) and SRp2 [SRp55/75 (SFRS6/4)] were shown to interact with each other and that their interactions were regulated by phosphorylation, hinting at a possible role in regulation of splicing in the mere 25% of multi-intronic genes of this organism (118). Unfortunately, in the previously mentioned study, there were no experiments conducted on alternative splicing. Furthermore, to date, there have not been any reports of alternative splicing in *S. pombe* (5). Therefore, it is plausible to consider that SR genes in basal unicellular eukaryotes perform rudimentary functions in regulated constitutive splicing. However, if we consider a recent report on the oomycete plant parasite, *P. sojae*, of which two of its three SR genes were resolved into the plant-specific SCL sub-family in the gene-tree analyses (Figure 6.2), there have been reported incidences of alternative intron processing in family 5 endoglucanase transcripts (25). It is tempting to speculate that perhaps these SCL-like SR genes could possibly be involved in pathogenic splicing regulation of endogenous plant mRNAs. This notwithstanding, it seems that alternative splicing in these organisms is a rare occurrence (neither of these organisms had EST/cDNA data to support AS in their SR genes), and instead, these SR genes might represent ancient prototypical SR genes that were either lost in higher lineages or adapted for new functionality.

Although we found no evidence for any broadly conserved sub-families, there was a single SR sub-family shared between members of the Viridiplantae, a red alga and the bilateral metazoans: SC35 (SFRS2). The sharing of this sub-family across so many diverse organisms might be due to its function not only in splicing (5 and 3 splice site recognition and interacting with U170-K and U2AF35) (129) but also because of its facilitation of transcription elongation of nascent transcripts (69). Presumably, this integration of transcription and splicing could very well be a fundamental biological process that has been conserved throughout multiple eukaryotic lineages.

Furthermore, our results support the idea that there are three plant-specific families: RS, RS2Z and SCL. Previously, it was thought that an additional sub-family (SR45) might be plant-specific, but our preliminary results (data not shown) suggest that SR45 did not appear later in evolution as previously thought (3), but is instead an ancient

SR-related gene found in monocots, dicots, mosses, fungi and slime molds (data not shown). Additionally, there are reports that human RNPS1 is homologous to plant SR45 (135), further lending credence to the idea that this is not a plant-specific SR sub-family.

Previous studies have often been limited in their phylogenetic scope, that is, often only a small subset of organisms and their SR gene repertoires are studied, such as human, drosophila, roundworm, fission yeast and Arabidopsis (105), or simply Arabidopsis, rice and moss (53). By including multiple species from divergent lineages, we were able to categorize SR genes into sub-families that will not only help in answering questions related to lineage-specific sub-family expansion (see below) and experimental design for gene knockout studies, but the classification performed here will aid in developing a standardized nomenclature for the SR genes (75).

It should be taken into account that the familial organization of the SR genes devised in this study are subject to change as genomes become more refined and annotations improve. Additionally, the use of the RNA recognition motif (RRM) domain as a phylogenetic marker may have certain pitfalls. By excluding other domains within the SR protein, such as the zinc knuckle domains, and simply by virtue of the shorter sequence length of RRMs, some phylogenetic signal may be lost (e.g., two SRs may be grouped into the same sub-family erroneously). However, the tradeoff between a slightly reduced signal versus a larger and more gap-filled alignment that increases the running time of our three tree searching methods was a factor we had to consider.

7.2 SR sub-family expansion in plants, structural constraints and selective pressures

Based on work done in Arabidopsis and rice, it was assumed that plants had the largest inventory of SR genes of any eukaryotes (7). The work presented here, with the inclusion of 27 different eukaryotic organisms, confirms this general trend (Figure 6.3). However, this trend is not universally seen in all plant species sampled. For example, *V. vinifera*, has the fewest SR genes of all the higher plants (Magnoliophyta). If one considers the influence of whole genome duplication events in the histories of flowering plants, this reduced number of SR genes in *V. vinifera* makes sense, since this genome has not undergone a recent duplication event, and instead experienced a paleo-hexaploidization

event after the divergence from the monocots but before the separation of the Eurosids (57).

The large number of SR genes in flowering plants can be attributed to whole genome duplication events, as previously mentioned. As whole genome duplication appears to be the rule rather than the exception within flowering plants, it is not surprising that these organisms would have a larger inventory of SR genes than organisms such as the metazoa. *Glycine max*, which has the most SRs of any organism we studied, is estimated to have undergone a recent genome duplication event, between 3-5 million years ago (2). Even the moss, *P. patens* is estimated to have a recent genome duplication in its past, occurring between 30-60 million years ago (100), around the same time that Arabidopsis experienced its most recent duplication event (10).

While there are bountiful SR genes in plants, what remains to be understood is why there is such a need for so many splicing regulators. Given our analysis using microarray expression data for Arabidopsis, rice, soybean and maize, it appears that expression levels between the members of a duplicate pair are tightly regulated, with very few instances of overlapping expression magnitudes within the same developmental stages (Figures 6.13-6.16). The high conservation of RRM location and intron number within SR sub-families combined with the overwhelming majority of SR homologs experiencing purifying selection all point to a post-duplication scenario of maintaining SR gene structure, form and function, albeit while reducing genetic redundancy via regulated gene expression. Such a situation might arise in evolution when there is a need for genetic robustness against potential null mutations and clear limits on gene dosage. However, an interesting case for novel function over redundancy is visible with respect to the SC35a gene in maize. ZmSC35a was one of the six genes with evidence to suggest that it is evolving under positive selection (see section 6.10 and Figure 6.12). Considering its expression profile against that of its paralog (last panel in Figure 6.16), it clearly overlaps in expression magnitude across 57 different arrays with ZmSC35b during the developmental stage of stem elongation. While most of the pairs may be experiencing purifying selection and may have redundant or sub-functions, ZmSC35a may be one of the salient examples of a neo-functionalized gene.

In addition to the highly conserved location of RRM regions within SR amino acid sequences, our analysis of RNA binding motifs within the plant-enriched sub-families is further indication that many residues within SR proteins are highly conserved and

under purifying selection. However, if many of the residues within RRM regions in a sub-family are conserved, how might binding specificity be achieved among sub-family members from a single species? Firstly, for each sub-family, there are multiple RNA binding motifs (underlined regions in Figure 6.11). Although many residues may be conserved within a sub-family in a particular binding region, the types of residues between binding regions are uniquely conserved. Moreover, in every predicted binding region there are at least three highly variable positions bordered by highly constant positions (except for the third binding motif in SC35, *sRGFAFVR*). Nevertheless, conserved and variable residues within binding regions are only partial players in RNA binding specificity. Other factors may influence specific binding or even be required to activate binding, such as phosphorylation of RS domains (115), even if RS domains may be interchangeable (121).

7.3 Alternative splicing of SR genes is a common characteristic among eukaryotes

There is a large interest in the SR gene family because they are important regulators in both constitutive and alternative splicing and are extensively alternative spliced themselves (71). Thus far, the investigation of AS of SR genes has been limited to a subset of model organisms, particularly mouse and human (66), drosophila (81), roundworm (72), Arabidopsis and rice (7). Though AS of SR genes has been shown to be a common occurrence in these organisms, what has not been addressed is the whether AS of SR genes is a common eukaryotic trait. While our data set is by no means comprehensive, consolidation of expression information for 27 organisms and their SR gene repertoires allows perspective into the extent of AS across organisms, the preferred types of AS events and how these events can vary by organism or specific SR sub-family.

We observed AS in SR genes across 20 organisms with sufficient EST/cDNA data (see Table 6.2). Mouse and human were the only two organisms to have AS events in each of their SR genes, which is probably reflective of their large EST/cDNA collections. Four of the original 27 organisms (*S. moellendorffi, C. vulgaris, C. merolae* and *S. pombe*) lacked sufficient ESTs, whereas no AS was found in the remaining three organisms with sufficient ESTs (*D. discoideum, P. falciparum* and *P. sojae*). It has

not escaped our attention that many of these organisms can be considered "basal" eukaryotes, with a highly reduced number of SR genes in their genomes relative to the remaining 20 organisms (see Figure 6.3). Their reduced number of SR genes is most likely indicative of their genomes having a relative low number of introns (31), and the lack of AS found in *D. discoideum, P. falciparum* and *P. sojae* SR genes further supports this idea. These organisms notwithstanding, the overwhelming occurrence of AS in SR genes is readily observable in Table 6.1 and Figures 6.19-6.20, and is highly suggestive of AS having a critical role in the regulation of SR genes in the Viridiplantae and Euteleostomi.

7.4 Not all AS event types are created equal

One of the advantages of performing a large comparative analysis of SR genes across species is the ability to discern which alternative splicing event types are predominant. Across the 20 organisms we sampled, alternative 3 splicing is the most common AS event type among SR genes (134 genes), followed by intron retention (111 genes), alternative 5 splicing (109 genes), skipped exons (106 genes) and finally alternative 3 and 5 events (29 genes). As we saw earlier, intron retention was not a universally abundant AS event type in the Viridiplantae and was only the most prevalent AS type in two of the nine photosynthetic eukaryotes (normalized averages in Figure 6.20). This suggests that different plant species might have specific preferences towards generating alternative splice forms of their SR genes or that the varying proportions of AS event types in Figure 6.20 remains biased by an uneven EST/cDNA distribution because we did not control for various tissue sources that ESTs/cDNAs may have been derived from. In contrast to the Viridiplantae, the Euteleostomi generally display a preference for exon skipping over intron retention, which agrees with previous genome-wide studies of alternative splicing in metazoans (61). The number of viable transcripts leading to functional SR proteins as a consequence of these five types of AS events was something beyond the scope of this study. However, it is interesting to consider the possibility that the majority of these alternative transcripts might lead to non-functional proteins and is instead a means for regulating levels of SR gene transcript abundance (62).

Interestingly, different SR sub-families show different levels of AS and preferences for AS event types. In general, there is a higher incidence of alternatively spliced SR

genes in plant-enriched sub-families as well as a higher number of IR and Alt 3 events per SR gene (green shaded boxes in Figure 6.21), whereas there is a lower number of alternatively spliced SR genes in non-photosynthetic sub-families and a lower incidence of IR events (blue shaded boxes in Figure 6.21) and higher number of skipped exons over the other types of events. These results suggest that specific sub-families rely on different types of AS to either generate novel protein forms with altered RRM binding domains (92), altered RS domains which may have implications on nuclear localization of the SR protein (105), or to affect the number of transcripts subjected to nonsense mediated decay (66).

After comparing methylated regions of the Arabidopsis genome to Arabidopsis SR genes, it became evident that there is a correlation between coding sequence methylation and alternative splicing. Given that 11 of the 19 SR genes in Arabidopsis are methylated in their coding regions and that these regions tend to overlap with areas of alternative splicing, it is highly likely that chromatin states can not be disregarded when considering factors and signals that influence patterns of alternative splicing and promises to be an exciting area for further research.

7.5 Summary and Outlook

In this work, we provided a large-scale comparative investigation into one of the most critical gene families involved in a fundamental biological process across multiple eukaryotic organisms. The SR gene family of splicing factors is pervasive throughout multiple lineages, is both conserved in sequence and domain organization yet differs in number and sub-family distribution across lineages and types of alternative splicing experienced. The work here has implications on the general evolution of homologous genes, for biological experimentation, differential regulation of SR gene expression by variable types of alternative splicing and hints at the importance of epigenetics to be incorporated into a code that can help to explain the regulation of alternative splicing not only in SR genes, but all the genes that these master regulators also have dominion over. However, there are many questions that remain to be addressed. One major issue that should be addressed is the level of divergence or conservation in alternative splicing events within paralogous and orthologous genes. A way to facilitate such an analysis on a large comparative genomics scale would be to implement a splice graph

alignment methodology, where a user could input their genes of interest and have the ability to automatically compare splicing patterns between their genes. Moreover, construction of a vast database of alternative splicing graphs in conjunction with DNA methylation data would go a long way to understanding the subtleties that underlie the complexities of alternative splicing. The further addition of a means for predicting the resultant amino acid sequence by following weighted paths through a splice graph (based on EST/cDNA preponderance) would aid in understanding the consequences of alternative splicing at a more functional level.

8

Materials & Methods

8.1 Project I – Intra-species comparative genomics in Arabidopsis

8.1.1 Arabidopsis duplicate gene pair sequences

A list of accession numbers for whole genome derived duplicate genes was obtained from (11). Genes were considered tandem duplicates and were excluded from this analysis if their protein alignments had a blast E-value $\leq 1\times10^{-10}$ and the corresponding sequences resided less than 15 genes apart on the same chromosome. Essentially, we employed the same criteria as that of (10). The TAIR accession numbers were used to obtain the 5' upstream regulatory sequences (URSs) and corresponding protein coding sequences (CDSs) for 2,584 gene pairs assumed to originate from the most recent whole genome duplication event in Arabidopsis (20-60 mya) (11).

Promoter annotation is dynamic and ever changing. The possibility for alternative 5' transcription start sites (TSS) increases as the amount of mRNA expression data increases. Therefore, to reduce the ambiguity regarding 5' TSS annotation, we maintained only those sequences that had annotated 5' untranslated regions (UTRs) supported by cDNA evidence from TAIR (101). We ensured that URSs were at least 600 bp long and did not interrupt other upstream, annotated genes. We discarded any pairs showing evidence for alternative 5' TSSs according to *blastn* searches against a database of Arabidopsis ESTs dowloaded from GenBank.

8.1.2 Arabidopsis expression information

Gene pairs that met the above criteria were then queried for *Pearson* (r) correlation coefficients of co-expression (EC). EC values were obtained from a database (87) that incorporates robust multi-array normalized (RMA) intensity values for 22,263 genes spanning 1,388 samples taken from the AtGenExpress project at TAIR. These samples are comprised of a variety of experimental conditions: i.e., different developmental stages, biotic, abiotic, nutrient, hormone and chemical treatments.

8.1.3 Working data set

The final list of usable duplicate gene pairs derived from the initial 2,584 was reduced to 815 through the above criteria. These 815 gene pairs (1,630 genes) are assumed to be polyploidy derived, unobtrusive to other genetic elements, clearly demarcated with a single 5' TSS, replete with expression information and have available protein coding sequences.

8.1.4 Upstream regulatory sequence anaylsis

Two data sets were constructed based on the URSs of the 815 Arabidopsis duplicate pairs. One URS data set consisted of six fixed-length sequence intervals, all anchored at the 5 TSS 8.1.

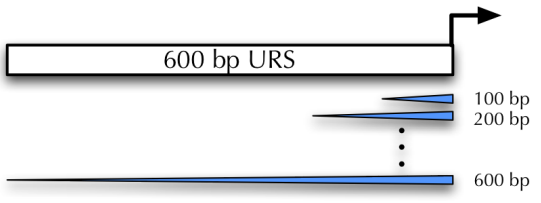

Figure 8.1: **Anchored window scheme** - Each URS fragment (blue arrow) increases by 100 bp until the maximum of 600 bp is reached.

The second URS data set consisted of nine sliding window intervals, each with a window size of 200 bp, and step size of 50 bp 8.2.

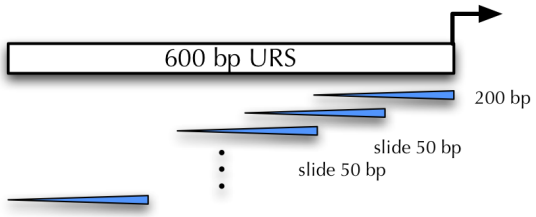

Figure 8.2: **Sliding window scheme** - Each URS fragment (blue arrow) slides by 50 bp until the maximum of 600 bp is reached.

For each data set and for each window interval, URS conservation was measured using three distinct applications:

1. the Shared Motif Method (SMM) (21)
2. the Index of Repetitiveness (IR) (43; 45)
3. DIALIGN-TX (DTX) (113; 114)

Each of the applications is explained in further detail below.

8.1.4.1 Shared Motif Method

The Shared Motif Method (SMM) was previously used in *ab initio* sequence divergence analysis of cis-regulatory DNA (21), or in other words, upstream regulatory/promoter sequences. The SMM was shown to be especially useful for measuring promoter divergence between homologous genes of organisms with poorly annotated transcription factor binding sites (TFBSs), such as *C. elegans*. It is essentially an alignment-based measure of similarity/divergence; however, what makes it unique is that it performs a series of recursive aligments after masking off high scoring segments during each iteration ((21), Supplementary information).

Prior to performing any alignments with the SMM, it is important to derive the alignment sensitivity (L) parameter. This is largely an empirical process of trial and error until the parameters are tuned in order to achieve an optimal signal to noise ratio. Following the recommendations of (21), L was determined by iterating through

progressively more stringent values until a level of "sequence divergence" of at least 90% was observable for each window size in each analysis.

As the SMM outputs a score (d_{SM}) reflective of the percentage of two sequences that do not share similar fragments, we report 1-d_{SM} so that the values would be reflective of similarity/conservation and would be comparable to programs that output values in terms of similarity.

See figure 8.3 for a pictorial representation of how the (d_{SM}) is calculated.

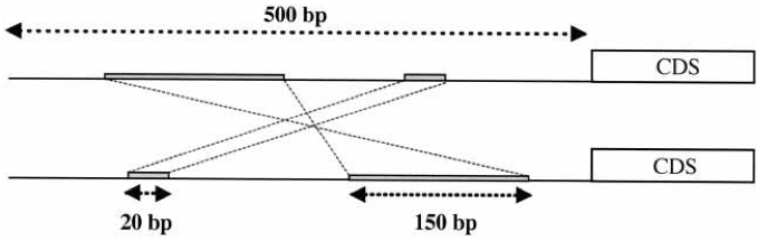

Figure 8.3: **The Shared Motif Method** - Illustration of the shared motif method (SMM). The SMM discovers regions of local similarity between DNA sequences without respect to their order, orientation, or spacing. In this example, two 500 bp noncoding sequences, upstream from homologous coding sequences (CDS), are compared. After iterative local alignment in both their native and inverted sequence orientations, two regions of significant local similarity between the sequences were discovered. One region is 150 bp long but has been inverted in one of the sequences. The other is 20 bp long but has been translocated. The fraction of shared motifs between these sequences is simply $(20 + 150)/500$, or 0.34. We define shared motif divergence (d_{SM}) as one minus this fraction, or $1 - 0.34 = 0.66$. Shared motif divergence is thus the fraction of the two sequences that does not contain a region of significant local alignment without respect to order, orientation, or spacing.Taken directly from (21).

8.1.4.2 The Index of Repetitiveness

The index of repetitiveness (IR) was first introduced and used to measure the repetitive nature of DNA on a genome- and organism-wide scale; more specifically, it was applied to a total of 336 genomes from all domains of life (45). Figure 8.4 describes how the IR is calculated for a single genome.

In this work, a modified version of the IR was utilized that makes a distinction between a query and a subject sequence, rather than computing the IR for a single

genome (43). The IR in this sense can then be thought of as an approximation for similarity or "repetitiveness" between two different DNA sequences. The algorithm for calculating this query/subject IR essentially determines for each position in the query the longest exact match in the subject. The resulting values are summed and the sum is divided by its expectation, assuming unrelatedness of query and subject. Finally, the logarithm of this ratio is taken to yield the IR (45). Therefore, the IR may take values between -∞ and +1, with an expectation of 0 for unrelated sequences.

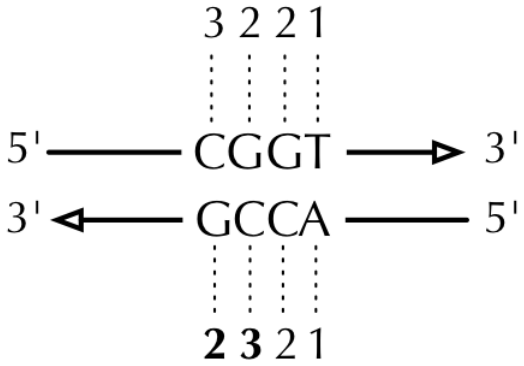

Figure 8.4: **The Index of repetitiveness** - Shortest unique substring lengths for the DNA sequence CGGT and its complement. Starting from, say, the first nucleotide, three steps in the 3' direction are necessary to generate a unique substring. The numbers in bold correspond to suffix length plus one. Taken directly from (45).

For each Arabidopsis duplicate pair, we calculated the IR as the arhmetic mean between the two IR values for the two posible query/subject configurations as shown in figure 8.5. Generally, sequence fragments that are identical between subject and query, and are longer than expected by chance (see (44) for the derivation of the aggregate shortest unique substring expectation), contribute to positive values of IR and have an upper bound of 1; however, random shuffles of the input sequences yield IR values close to 0.

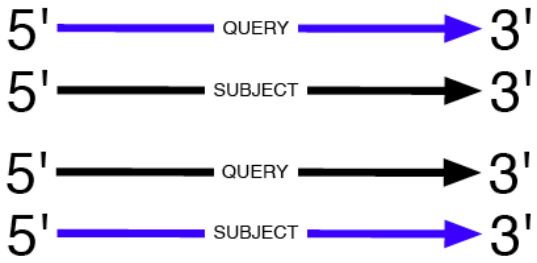

Figure 8.5: **Query/subject IR** - The IR values used in this work were calculated by taking the arithmetic mean for the two possible query/subject configurations (*minus strands not shown*). The black and blue lines represent the role reversal of query and subject.

8.1.4.3 DIALIGN-TX

DIALIGN-TX (DTX) (113; 114) is a greedy segment based alignment approach to measuring similarities between DNA and amino acid sequences. Three previous versions of DIALIGN have been in wide circulation throughout the scientific community (79; 80; 114). However, the latest version of DTX improves upon previous versions by including a progressive alignment approach as well as a vertex-cover approximation on a conflict graph in order to reduce susceptibility to spurious random similarities(113). For a detailed explanation of the algorithm used in DTX, the reader should consult the original publication (113).

In order to calculate an alignment similarity score for the 815 gene pairs used in this study, we used the raw counts of the number of alignable bases (see figure 8.6) in the URSs divided by the total number of bases of the URSs, as was done in (40).

```
AT3G05590  --------AC ACCATCCTAA TTATATAAAC AGACAGTGAT TAgta-----
AT5G27850  atggataaAC ATCATCCATA TCTTATGGAT AGAGATGAAT GAtggataat
```

alignment similarity score = 34/50 = 0.68

Figure 8.6: **DIALIGN-TX example alignment** - Alignment similarity was calculated by taking the number of alignable bases (boldface letters) and dividing by the length of the sequences considered. In this example, the number of alignable bases (mismatches considered as "alignable") is 34 bp and the total length is 50 bp.

8.1.5 Coding sequence analyses

For each of the 815 duplicate gene pairs, the synonymous (K_s), non-synonymous (K_a) and the ratio of non-synonymous to synonymous substitution rates (K_a/K_s) were calculated using a Python wrapper known as *MutationsHunter* (77). The MutationsHunter package integrates three different commonly used software packages:

- InParanoid (88),
- FASTA (95) and
- PAML (131).

MutationsHunter greatly facilitates the laborious process of conducting an evolutionary analysis (i.e., estimating substitution rates) on a large set of protein coding genes. There are essentially four steps to this process:

1. construct a list of homologous protein coding gene pairs (orthologs or paralogs) using the InParanoid launcher script, or by some other means.
2. perform a global alignment of the two corresponding protein coding sequences to detect gaps and mismatches (using FASTA).
3. convert the protein alignment into a DNA alignment
4. calculate the substitution rates between the two sequences using PAML.

Step 1 in this project, as described in 8.1.1 was performed by consulting literature, rather than using the InParanoid database. Steps 2–4 were fully conducted within the framework of *MutationsHunter*.

8.2 Project II - Inter-species comparative genomics in 28 eukaryotes

8.2.1 Species selection

We employed fundamental practical criteria to determine which organisms would be sampled in our study. These included:

1. Completeness and availability of genomic sequence
2. Availability and bulk of cDNA or EST data
3. Phylogeny

The major element influencing the selection of species was that of expression information, since this was the limiting factor. We used NCBI's dbEST (12) in order to glean information on the abundance of available transcripts per organism contained within the NCBI genome databases (see Table 8.1 for a sample).

Table 8.1: Subsample of NCBI's dbEST

Organism	ESTs
Homo sapiens (human)	8,296,280
Mus musculus + domesticus (mouse)	4,852,144
Zea mays (maize)	2,018,798
Bos taurus (cattle)	1,558,492
Sus scrofa (pig)	1,538,441
Arabidopsis thaliana (thale cress)	1,527,298
Danio rerio (zebrafish)	1,481,930
Glycine max (soybean)	1,422,604
Xenopus (Silurana) tropicalis (western clawed frog)	1,271,375
Oryza sativa (rice)	1,249,110
Ciona intestinalis	1,205,674
Triticum aestivum (wheat)	1,067,291

The top 12 organisms with more than 1 million ESTs in GenBank's dbEST division as of December 11, 2009. Five of the 12 organisms listed above are plants.

Based on the EST counts per organism and their phylogenetic nature, 27 species were selected and included in the alternative splicing analysis. Details of the data mining procedure are described below.

8.2.2 Organism sampling and SR sequence acquisition

To begin the assessment of the genomic inventory of SR genes in eukaryotes, we selected taxa based on completeness of genome sequencing efforts and their phylogenetic diversity as inferred by NCBIs taxonomy browser (`http://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi`; Figure 6.1). We sampled a total of 27 organisms with fully sequenced genomes that ranged from plants, animals and fungi (Opisthokonts) to Amoebozoa, Stramenopiles and the Alveolata (102). Once the organisms were chosen, SR amino acid sequences were obtained through either literature searches (*Homo sapiens* (66), *Caenorhabditis elegans* (72), *Drosophila melanogaster* (81), *Schizosaccharomyces pombe* (118), *Arabidopsis thaliana* (59)and *Oryza sativa* (53)) or via hidden markov model (HMM) searches using HMMER3 (30) (see Table 6.1) of downloaded protein databases.

We used a combination of HMM (30) and BLASTP (4) searches to find and then verify that putative sequences were SR gene homologs. We constructed three separate HMMs: one for the Viridiplantae (vHMM), one for the Fungi/Metazoa (fmHMM) and one for the Amoebozoa, Stramenopiles and Alveolata (asaHMM). The vHMM was composed of globally aligned (119) SR proteins of *Arabidopsis thaliana, Oryza sativa*, preliminary BLASTP candidate sequences from *Populus trichocarpa* and *Chlamydomonas reinhardtii*. Using this vHMM, we then searched downloaded protein databases of *Glycine max, Vitis vinifera, Zea mays, Sorghum bicolor, Selaginella moellendorfii, Physcomitrella patens, Chlorella vulgaris* and *Cyanidioschyzon merolae* (database references in Table 6.1). After re-searching downloaded databases of *Chalmydomonas reinhardtii* and *Populus trichocarpa* with this HMM, we then used the full sequence E-value from the HMMER3 output to exclude hits with an E-value greater than $1x10^{-03}$ to generate a set of candidate SR proteins. Next, we blasted each of the candidate SR proteins against the nr protein database at NCBI to cross-validate which of the candidate sequences could be further excluded based on sequence similarity to known non-SR proteins. All remaining candidates were then manually examined for the oc-

currence of at least three SR dipeptides and then submitted to Interproscan for domain searches to elucidate positions of their RNA recognition motifs (RRMs) (82; 133).

A similar process was performed with the fmHMM and the asaHMM. The only differences being the underlying sequences used in the construction of the respective HMMs. The fmHMM was composed of known SRs from *Homo sapiens, Caenorhabditis elegans, Drosophila melanogaster* and *Schizosaccharomyces pombe*, whereas the asaHMM was comprised of SRs from *Homo sapiens, Ciona intestinalis, Drosophila melanogaster, Neurospora crassa, Arabidopsis thaliana* and *Chlamydomonas reinhardtii*. Using the fmHMM, we searched downloaded protein sequence databases of *Mus musculus, Gallus gallus, Xenpus tropicalis, Danio rerio, Branchiostoma floridae, Ciona intestinalis, Anopheles gambiae, Aedes aegypti* and *Neurospora crassa* (references in Table 3.1). The asaHMM was used to search downloaded databases of *Plasmodium falciparum, Phytopthora sojae* and *Dictyostelium discoidum*. As with the vHMM search process, the same data filtering steps were taken to derive putative SR gene homologs within the Fungi/Metazoa and other eukaryotes.

Additionally, any sequences that did not begin with methionine residues were discarded from further consideration. Therefore, the numbers of SR genes per organism reported in Table 3.1 are likely to change as annotation efforts improve. The following sequences were removed: Chlv31017 (*Chlorella vulgaris*), Sb0514s002010 and Sb09g004685 (*Sorghum bicolor*), and Smo36388 (*Selaginella moellendorfii*). All accession numbers for all SR proteins used in these analyses are available online at `http://combi.cs.colostate.edu/as/gmap_SRgenes/`.

8.2.3 Alignment procedure

The resulting 272 SR proteins from the searches described above were initially aligned using DIALIGN-TX (113; 114) with default parameters. The RNA recognition motifs (RRMs) were extracted from the full-length amino acid sequences of the SR proteins based on their SMART (68) prediction coordinates from Interproscan searches (82; 133). First, all RRMs were aligned aggregately to ensure that sequences harbouring multiple RRMs were collinear. Imagine two SR proteins (Seq1, Seq2) with two RRMs, one located proximal to the N-terminus (RRM1) and the other located closer to the C-terminus (RRM2). In order to be certain that the N-terminal RRM of Seq1 aligned with the N-terminal RRM of Seq2, it was necessary to align all RRMs independently. A

preliminary UPGMA tree was constructed to evaluate the aligned RRMs. In all cases, there were no instances of a criss-crossed matching of RRM1 with RRM2. However, there was a single SR protein from *Branchiostoma floridae* (Br126246) that contained three RRM domains. Accordingly, we evaluated whether the three RRMs all had as their top BLASTP hit a SR protein with a single RRM. As this was the case, and because the three RRMs appeared to be tandem duplications of a single RRM, or a possible annotation error, we selected the most N-terminal RRM of Br126246 to be used in our gene tree inferences.

After the above determinations, all N-terminal RRMs were aligned separately from those sequences harbouring a C-terminal RRM, which were also aligned separately. Here, we used FSA (15) for the alignment of the RRMs because of its explicit consideration of insertions that should not align, which would otherwise confound our gene tree analyses by over-estimating the substitution rates. The disjoint alignments of sequences with two RRMs were then concatenated and any columns that would be considered gap-only if a single sequence did not cause an unalignable insertion to exist were removed. For those sequences with identical RRMs (17 sequences), only one representative was selected for use in gene tree construction, reducing the data matrix to 255 taxa. Twenty-eight columns of the 353 total characters in the alignment were constant, 267 were parsimony-informative and 58 were uninformative variable characters. This alignment file was then used for a series of gene tree searches as described below.

8.2.4 Gene tree inferences

The alignment described above was input into PROTTEST version 2.4 (1) and assessed for the best fitting model of amino acid substitution. The best scoring model with the fewest number of parameters was the LG model with a gamma shape distribution for rate heterogeneity (LG+G, lnl: -24515.47). Next, two maximum likelihood (ML) methods and a parsimony method were used to construct gene trees of the 255 SR proteins. We used the parallel threads implementation of RAxML version 7.2.6 (84; 109) to perform 2000 rapid bootstraps and search for the best known tree under the LG+G model (lnl: -23016.56). We used Garli version 1.0 as the second ML tree search method to conduct ML analyses on another 1000 bootstrap replicates (138). One thousand parsimony bootstrap replicate searches were conducted in Phylip version 3.69 using the protpars program and randomized input order of sequences (10 jumbles) (33).

Bootstrap support values from all three analyses were then mapped onto the best scoring ML tree from the RAxML analysis.

8.2.5 Genomic and cDNA/EST sequences for Alternative Splicing (AS) analysis

In addition to acquiring amino acid sequences of the SR genes, we also obtained full-length genomic sequences from the corresponding databases in Table 6.1. Next, we performed a series of MEGABLAST searches against NCBIs dbEST using each of the genomic sequences for each of the organisms in order to collect expression data to be used in the analysis of alternative splicing (AS) for the organisms under study. MEGABLAST searches were also conducted against the nr nucleotide database to acquire any full-length cDNAs that were available.

8.2.6 Alternative splicing analysis

Of the 27 eukaryotic organisms sampled in this study, 25 had expression data obtained from the MEGABLAST searches described above, except for (*Selaginella moellendorfii, Chlorella vulgaris* and *Cyanidioschyzon merolae*). The genomic sequences and transcript sequences were then fed into an in-house generated pipeline to assess the extent of AS among the SR genes in these 25 organisms.

We used Sircah (41) to detect possible AS events from a set of aligned transcripts. Sircah is an application written in Python that detects AS events and provides visualizations in the form of splice graphs. On the program's website (13), the authors have outlined the heuristics Sircah uses to detect AS events. The algorithm correctly detects most events, but we found cases in a previous study in which it identified spurious events. Thus, for our analysis we used a version of Sircah that we revised to correct these errors (see (65)).

To provide meaningful counts for alternative splicing events, we established rules for each event type. As the number of alternative splicing events increases in a gene, the number of combinations representing potential splice forms increases exponentially, but all splice forms are not equally probable. For our analysis we counted the number of events supported by EST transcripts. As a simple example, consider the transcripts

given in Figure 8.7A. Although there are two retained introns, the transcripts support just one intron retention event in which both introns are retained simultaneously. Consequently, for this graph we count a single intron retention event.

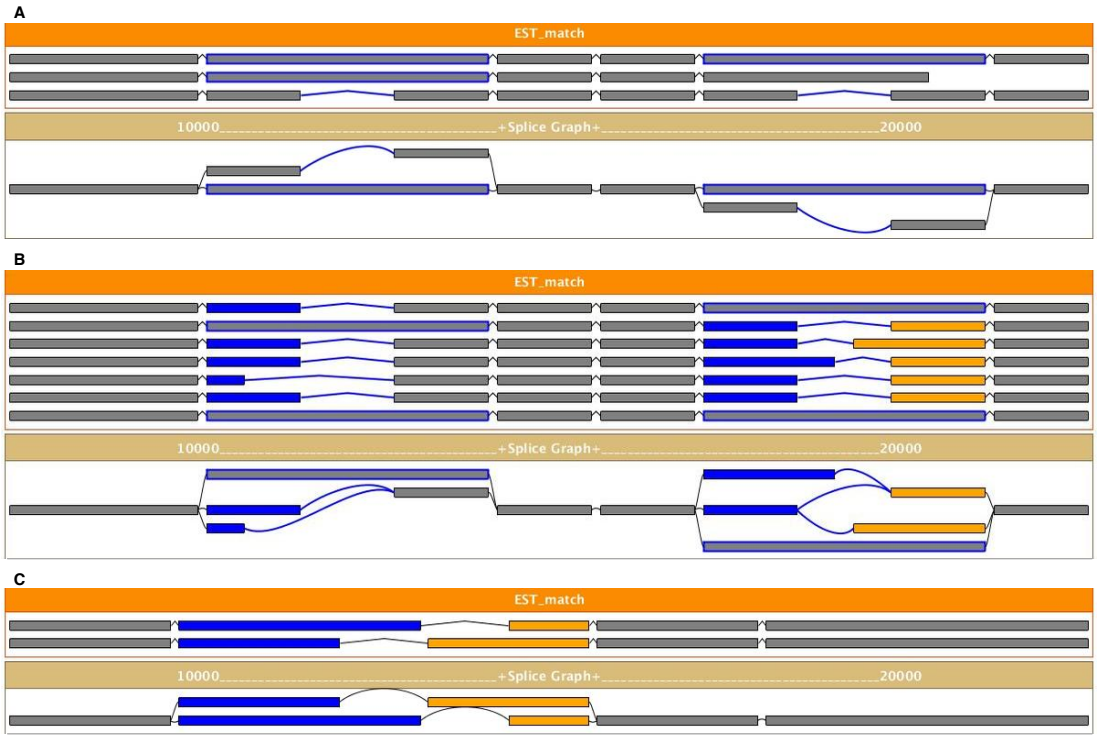

Figure 8.7: **Splice graph examples** - Generic splice graphs generated for illustration purposes. A. counting of intron retention events. B. alternate donor, acceptor and intron retention events. C. counting of Alt B events (simultaneous alternative splicing at 3' and 5' sites).

A more complicated example is shown in Figure 8.7B. The graph has two retained introns for which three combinations are supported by EST transcripts. In addition, there are two alternate 5' events supported by transcripts as well as an alternate 3' event. In this case, we count three intron retention events, two alternate 5' events and a single alternate 3' event.

The rules for cassette exons are analogous to those for intron retention: when there is evidence of multiple skipped exons in a gene, we count number of distinct EST transcripts that support each combination. For alternative 3' and 5' splice sites, we use the most prevalent splice site (the one supported by a plurality of EST transcripts)

and simply count the number of alternatives. When we cannot determine a prevalent form, we use the splice site that yields the longest intron.

We distinguish between alternate 3' sites (Alt 3'), alternate 5' sites (Alt 5') and simultaneous 3'/5' events (Alt B). We count Alt B events whenever an alternative 5' site is paired with the same alternate 3' site in all transcripts. For example, in Figure 8.7C the alternate 3' and 5' splice sites are paired, so this will be counted as a single Alt B event.

We incorporated our counting rules into our modified version of Sircah and generated statistics for each kind of AS event. For each type of event we counted the total number of events detected; the overall proportion of genes having each event, and the A-T composition of introns and exons involved in each event. In addition, for skipped exons we tracked the exon length, and for retained introns, we tracked the intron length as well as the lengths of its flanking exons.

Much of the evidence for AS in the SR genes is complex and complicates analysis even with the compact form of a splice graph. However, the splice graph format provides a way to assess this complexity. For each node in a gene's splice graph we counted the number of incoming and outgoing intron edges to establish the node's branching factor. We could thus measure a graph's complexity using the maximum branching factor for any node. For example, the graph in Figure 8.7A is relatively simple, with a maximum branching factor of 2, in contrast to Figure 8.7B, which is more complex, with a maximum branching factor of 4.

8.2.7 Normalization of Alternative splicing measurements

To compare alternative splicing evidence between SR genes from different organisms, we applied an approach similar to that used in (16; 61). We ran 100 trials in which we randomly selected a fixed number of 15 ESTs for each SR gene in each organism. Genes that had fewer than the required 15 EST alignments were omitted from our analysis. We ran a modified version of Sircah (41; 65) on the randomly selected ESTs to generate statistics on the number of alternative splicing events. In each trial and for each organism we counted the number of genes used in the trial, the number of genes that exhibited alternative splicing and the number of alternative splicing events: intron retention, skipped exon, alternative 5' site, alternative 3' site and simultaneous 3'/5' (altB).

To establish our threshold of 15 ESTs per gene, we examined the distribution of aligned ESTs across our SR gene data (see [Additional File 5] online). Raising the number of ESTs required in each trial improved the method's sensitivity, but reduced the number of genes available for comparison. *S. pombe*, provided only a single gene each and did not generate useful statistics. A third organism, *C. reinhardtii*, had just one gene with more than 10 aligned ESTs. The remaining species had enough genes and EST alignments to make meaningful comparisons. We selected a threshold of 15 ESTs to provide enough sensitivity to illuminate differences between species while permitting analysis on all but the three poorly represented species.

References

[1] F. Abascal, R. Zardoya, and D. Posada. Prottest: selection of best-fit models of protein evolution. *Bioinformatics*, 21(9):2104–5, May 1 2005. 103

[2] Keith L Adams and Jonathan F Wendel. Polyploidy and genome evolution in plants. *Curr Opin Plant Biol*, 8(2):135–41, Apr 2005. 88

[3] G. S. Ali, S. G. Palusa, M. Golovkin, J. Prasad, J. L. Manley, and A. S. Reddy. Regulation of plant developmental processes by a novel splicing factor. *PLoS One*, 2(5):e471, 2007. 86

[4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, Oct 5 1990. 101

[5] Gil Ast. How did alternative splicing evolve? *Nat Rev Genet*, 5(10):773–82, Oct 2004. 86

[6] C. Aurrecoechea, J. Brestelli, B. P. Brunk, J. Dommer, S. Fischer, B. Gajria, X. Gao, A. Gingle, G. Grant, O. S. Harb, M. Heiges, F. Innamorato, J. Iodice, J. C. Kissinger, E. Kraemer, W. Li, J. A. Miller, V. Nayak, C. Pennington, D. F. Pinney, D. S. Roos, C. Ross, Jr. Stoeckert C. J., C. Treatman, and H. Wang. Plasmodb: a functional genomic database for malaria parasites. *Nucleic Acids Res*, 37(Database issue):D539–43, Jan 2009. 52

[7] A. Barta, M. Kalyna, and Z. J. Lorkovic. Plant sr proteins and their functions. *Curr Top Microbiol Immunol*, 326:83–102, 2008. 49, 52, 60, 87, 89

[8] N. Behzadnia, M. M. Golas, K. Hartmuth, B. Sander, B. Kastner, J. Deckert, P. Dube, C. L. Will, H. Urlaub, H. Stark, and R. Luhrmann. Composition and three-dimensional em structure of double affinity-purified, human prespliceosomal a complexes. *EMBO J*, 26(6):1737–48, Mar 21 2007. 47

[9] Dennis A Benson, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. Genbank. *Nucleic Acids Res*, 38(Database issue):D46–51, Jan 2010. 53

[10] G. Blanc, K. Hokamp, and K. H. Wolfe. A recent polyploidy superimposed on older large-scale duplications in the arabidopsis genome. *Genome Res*, 13(2):137–44, 2003. Journal Article Research Support, Non-U.S. Gov't United States. 6, 9, 88, 93

[11] G. Blanc and K. H. Wolfe. Functional divergence of duplicated genes formed by polyploidy during arabidopsis evolution. *Plant Cell*, 16(7):1679–91, 2004. Journal Article Research Support, Non-U.S. Gov't United States. 7, 68, 93

[12] M S Boguski, T M Lowe, and C M Tolstoshev. dbest–database for "expressed sequence tags". *Nat Genet*, 4(4):332–3, Aug 1993. 100

[13] Bork. http://www.bork.embl.de/sircah/description.html, NaN. 104

[14] J. E. Bowers, B. A. Chapman, J. Rong, and A. H. Paterson. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, 422(6930):433–8, 2003. Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. England. 6

[15] Robert K Bradley, Adam Roberts, Michael Smoot, Sudeep Juvekar, Jaeyoung Do, Colin Dewey, Ian Holmes, and Lior Pachter. Fast statistical alignment. *PLoS Comput Biol*, 5(5):e1000392, May 2009. 103

[16] David Brett, Heike Pospisil, Juan Valcárcel, Jens Reich, and Peer Bork. Alternative splicing and genome complexity. *Nat Genet*, 30(1):29–30, Jan 2002. 106

[17] R. J. Britten and E. H. Davidson. Gene regulation for higher cells: a theory. *Science*, 165(891):349–57, 1969. Journal Article United states. 39

[18] J. F. Caceres, G. R. Screaton, and A. R. Krainer. A specific subset of sr proteins shuttles continuously between the nucleus and the cytoplasm. *Genes Dev*, 12(1):55–66, Jan 1 1998. 48

[19] M. A. Campbell, B. J. Haas, J. P. Hamilton, S. M. Mount, and C. R. Buell. Comprehensive analysis of alternative splicing in rice and comparative analyses with arabidopsis. *BMC Genomics*, 7:327, 2006. 48

[20] S. Carbon, A. Ireland, C. J. Mungall, S. Shu, B. Marshall, and S. Lewis. Amigo: online access to ontology and annotation data. *Bioinformatics*, 2008. the AmiGO Hub the Web Presence Working Group Journal article Bioinformatics (Oxford, England) Bioinformatics. 2008 Nov 25. 17

[21] C. I. Castillo-Davis, D. L. Hartl, and G. Achaz. cis-regulatory and protein evolution in orthologous and duplicate genes. *Genome Res*, 14(8):1530–6, 2004. Comparative Study Journal Article Research Support, Non-U.S. Gov't United States. 8, 9, 40, 95, 96

[22] D. Cazalla, J. Zhu, L. Manche, E. Huber, A. R. Krainer, and J. F. Caceres. Nuclear export and retention signals in the rs domain of sr proteins. *Mol Cell Biol*, 22(19):6871–82, Oct 2002. 48

[23] Mo Chen and James L Manley. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol*, 10(11):741–54, Nov 2009. 47

References

[24] L. T. Chow, R. E. Gelinas, T. R. Broker, and R. J. Roberts. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger rna. *Cell*, 12(1):1–8, Sep 1977. 47

[25] Stefano Costanzo, Manuel D Ospina-Giraldo, Kenneth L Deahl, C Jacyn Baker, and Richard W Jones. Alternate intron processing of family 5 endoglucanase transcripts from the genus phytophthora. *Curr Genet*, 52(3-4):115–23, Sep 2007. 86

[26] Gavin E Crooks, Gary Hon, John-Marc Chandonia, and Steven E Brenner. Weblogo: a sequence logo generator. *Genome Res*, 14(6):1188–90, Jun 2004. 69

[27] S. De Bodt, S. Maere, and Y. Van de Peer. Genome duplication and the origin of angiosperms. *Trends Ecol Evol*, 20(11):591–7, 2005. Journal Article England. 6

[28] J. Deckert, K. Hartmuth, D. Boehringer, N. Behzadnia, C. L. Will, B. Kastner, H. Stark, H. Urlaub, and R. Luhrmann. Protein composition and electron microscopy structure of affinity-purified human spliceosomal b complexes isolated under physiological conditions. *Mol Cell Biol*, 26(14):5528–43, Jul 2006. 47

[29] Rachel Drysdale and FlyBase Consortium. Flybase : a database for the drosophila research community. *Methods Mol Biol*, 420:45–59, 2008. 52

[30] Sean Eddy. Hmmer3, 2010. 101

[31] L. Eichinger, JA Pachebat, G. Gl "ockner, M.A. Rajandream, R. Sucgang, M. Berriman, J. Song, R. Olsen, K. Szafranski, Q. Xu, et al. The genome of the social amoeba Dictyostelium discoideum. *Nature*, 435(7038):43–57, 2005. 90

[32] D. Farre, N. Bellora, L. Mularoni, X. Messeguer, and M. M. Alba. Housekeeping genes tend to show reduced upstream sequence conservation. *Genome Biol*, 8(7):R140, 2007. Farre, Domenec Bellora, Nicolas Mularoni, Loris Messeguer, Xavier Alba, M Mar Research Support, Non-U.S. Gov't England Genome biology Genome Biol. 2007;8(7):R140. 43

[33] J Felsenstein. Phylip (phylogeny inference package) version 3.6, 2005. 103

[34] P. Fey, P. Gaudet, T. Curk, B. Zupan, E. M. Just, S. Basu, S. N. Merchant, Y. A. Bushmanova, G. Shaulsky, W. A. Kibbe, and R. L. Chisholm. dictybase–a dictyostelium bioinformatics resource update. *Nucleic Acids Res*, 37(Database issue):D515–9, Jan 2009. 52

[35] E. W. Ganko, B. C. Meyers, and T. J. Vision. Divergence in expression between duplicated genes in arabidopsis. *Mol Biol Evol*, 2007. Journal article. 7, 43

[36] H. Gao, W. J. Gordon-Kamm, and L. A. Lyznik. Asf/sf2-like maize pre-mrna splicing factors affect splice site utilization and their transcripts are alternatively spliced. *Gene*, 339:25–37, Sep 15 2004. 49

[37] W. Gilbert. Why genes in pieces? *Nature*, 271(5645):501, Feb 9 1978. 47

[38] B. R. Graveley. Sorting out the complexity of sr protein functions. *RNA*, 6(9):1197–211, 2000. Graveley, B R Review United states RNA (New York, N.Y.) RNA. 2000 Sep;6(9):1197-211. 48

[39] B. R. Graveley. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet*, 17(2):100–7, Feb 2001. 47, 48

[40] G. Haberer, T. Hindemitt, B. C. Meyers, and K. F. Mayer. Transcriptional similarities, dissimilarities, and conservation of cis-elements in duplicated genes of arabidopsis. *Plant Physiol*, 136(2):3009–22, 2004. Journal Article Research Support, Non-U.S. Gov't United States. 7, 8, 11, 40, 41, 43, 98

[41] E. D. Harrington and P. Bork. Sircah: a tool for the detection and visualization of alternative transcripts. *Bioinformatics*, 24(17):1959–60, Sep 1 2008. 104, 106

[42] T. W. Harris, I. Antoshechkin, T. Bieri, D. Blasiar, J. Chan, W. J. Chen, N. De La Cruz, P. Davis, M. Duesbury, R. Fang, J. Fernandes, M. Han, R. Kishore, R. Lee, H. M. Muller, C. Nakamura, P. Ozersky, A. Petcherski, A. Rangarajan, A. Rogers, G. Schindelman, E. M. Schwarz, M. A. Tuli, K. Van Auken, D. Wang, X. Wang, G. Williams, K. Yook, R. Durbin, L. D. Stein, J. Spieth, and P. W. Sternberg. Wormbase: a comprehensive resource for nematode research. *Nucleic Acids Res*, 38(Database issue):D463–7, Jan 2009. 52

[43] B. Haubold, M. Domazet-Loso, and T. Wiehe. An alignment-free distance measure for closely related genomes. *RECOMB-CG Proceedings*, 5267:87–99, 2008. 8, 9, 40, 95, 97

[44] B. Haubold, N. Pierstorff, F. Moller, and T. Wiehe. Genome comparison without alignment using shortest unique substrings. *BMC Bioinformatics*, 6:123, 2005. 1471-2105 (Electronic) Journal Article Research Support, Non-U.S. Gov't. 97

[45] B. Haubold and T. Wiehe. How repetitive are genomes? *BMC Bioinformatics*, 7:541, 2006. Journal Article Research Support, Non-U.S. Gov't England. 8, 9, 40, 95, 96, 97

[46] C. Haynes and L. M. Iakoucheva. Serine/arginine-rich splicing factors belong to a class of intrinsically disordered proteins. *Nucleic Acids Res*, 34(1):305–12, 2006. 48

[47] K. J. Hertel and B. R. Graveley. Rs domains contact the pre-mrna throughout spliceosome assembly. *Trends Biochem Sci*, 30(3):115–8, Mar 2005. 48

[48] Christiane Hertz-Fowler, Chris S Peacock, Valerie Wood, Martin Aslett, Arnaud Kerhornou, Paul Mooney, Adrian Tivey, Matthew Berriman, Neil Hall, Kim Rutherford, Julian Parkhill, Alasdair C Ivens, Marie-Adele Rajandream, and Bart Barrell. Genedb: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res*, 32(Database issue):D339–43, Jan 2004. 52

[49] K Higo, Y Ugawa, M Iwamoto, and T Korenaga. Plant cis-acting regulatory dna elements (place) database: 1999. *Nucleic Acids Res*, 27(1):297–300, Jan 1999. 8

References

[50] A. E. House and K. W. Lynch. Regulation of alternative splicing: more than just the abcs. *J Biol Chem*, 283(3):1217–21, Jan 18 2008. 47

[51] T. J. Hubbard, B. L. Aken, S. Ayling, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, L. Clarke, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Graf, S. Haider, M. Hammond, R. Holland, K. Howe, A. Jenkinson, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, D. Rios, M. Schuster, G. Slater, D. Smedley, W. Spooner, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, S. Wilder, A. Zadissa, E. Birney, F. Cunningham, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, A. Kasprzyk, G. Proctor, J. Smith, S. Searle, and P. Flicek. Ensembl 2009. *Nucleic Acids Res*, 37(Database issue):D690–7, 2009. Hubbard, T J P Aken, B L Ayling, S Ballester, B Beal, K Bragin, E Brent, S Chen, Y Clapham, P Clarke, L Coates, G Fairley, S Fitzgerald, S Fernandez-Banet, J Gordon, L Graf, S Haider, S Hammond, M Holland, R Howe, K Jenkinson, A Johnson, N Kahari, A Keefe, D Keenan, S Kinsella, R Kokocinski, F Kulesha, E Lawson, D Longden, I Megy, K Meidl, P Overduin, B Parker, A Pritchard, B Rios, D Schuster, M Slater, G Smedley, D Spooner, W Spudich, G Trevanion, S Vilella, A Vogel, J White, S Wilder, S Zadissa, A Birney, E Cunningham, F Curwen, V Durbin, R Fernandez-Suarez, X M Herrero, J Kasprzyk, A Proctor, G Smith, J Searle, S Flicek, P 062023/Wellcome Trust/United Kingdom WT062023/Wellcome Trust/United Kingdom Biotechnology and Biological Sciences Research Council/United Kingdom Medical Research Council/United Kingdom Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't England Nucleic acids research Nucleic Acids Res. 2009 Jan;37(Database issue):D690-7. Epub 2008 Nov 25. 52

[52] S. Hunter, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A. F. Quinn, J. D. Selengut, C. J. Sigrist, M. Thimma, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, and C. Yeats. Interpro: the integrative protein signature database. *Nucleic Acids Res*, 37(Database issue):D211–5, 2009. Hunter, Sarah Apweiler, Rolf Attwood, Teresa K Bairoch, Amos Bateman, Alex Binns, David Bork, Peer Das, Ujjwal Daugherty, Louise Duquenne, Lauranne Finn, Robert D Gough, Julian Haft, Daniel Hulo, Nicolas Kahn, Daniel Kelly, Elizabeth Laugraud, Aurelie Letunic, Ivica Lonsdale, David Lopez, Rodrigo Madera, Martin Maslen, John McAnulla, Craig McDowall, Jennifer Mistry, Jaina Mitchell, Alex Mulder, Nicola Natale, Darren Orengo, Christine Quinn, Antony F Selengut, Jeremy D Sigrist, Christian J A Thimma, Manjula Thomas, Paul D Valentin, Franck Wilson, Derek Wu, Cathy H Yeats, Corin BB/F010508/1/Biotechnology and Biological Sciences Research Council/United Kingdom GM081084/GM/NIGMS NIH HHS/United States Wellcome Trust/United Kingdom Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't England Nucleic acids research Nucleic Acids Res. 2009 Jan;37(Database issue):D211-5. Epub 2008 Oct 21. 65

[53] K. Iida and M. Go. Survey of conserved alternative splicing events of mrnas encoding sr proteins in land plants. *Mol Biol Evol*, 23(5):1085–94, 2006. Iida, Kei Go, Mitiko Research Support, Non-U.S. Gov't United States Molecular biology and evolution Mol Biol Evol. 2006 May;23(5):1085-94. Epub 2006 Mar 6. 49, 51, 87, 101

[54] The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, 408(6814):796–815, 2000. 5

[55] M. Isshiki, A. Tsumoto, and K. Shimamoto. The serine/arginine-rich protein family in rice plays important roles in constitutive and alternative splicing of pre-mrna. *Plant Cell*, 18(1):146–58, Jan 2006. 49

[56] H. Iwama and T. Gojobori. Highly conserved upstream sequences for transcription factor genes and implications for the regulatory network. *Proc Natl Acad Sci U S A*, 101(49):17156–61, 2004. Iwama, Hisakazu Gojobori, Takashi United States Proceedings of the National Academy of Sciences of the United States of America Proc Natl Acad Sci U S A. 2004 Dec 7;101(49):17156-61. Epub 2004 Nov 30. 43

[57] O. Jaillon, J. M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, A. Vezzi, F. Legeai, P. Hugueney, C. Dasilva, D. Horner, E. Mica, D. Jublot, J. Poulain, C. Bruyere, A. Billault, B. Segurens, M. Gouyvenoux, E. Ugarte, F. Cattonaro, V. Anthouard, V. Vico, C. Del Fabbro, M. Alaux, G. Di Gaspero, V. Dumas, N. Felice, S. Paillard, I. Juman, M. Moroldo, S. Scalabrin, A. Canaguier, I. Le Clainche, G. Malacrida, E. Durand, G. Pesole, V. Laucou, P. Chatelet, D. Merdinoglu, M. Delledonne, M. Pezzotti, A. Lecharny, C. Scarpelli, F. Artiguenave, M. E. Pe, G. Valle, M. Morgante, M. Caboche, A. F. Adam-Blondon, J. Weissenbach, F. Quetier, and P. Wincker. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449(7161):463–7, 2007. Jaillon, Olivier Aury, Jean-Marc Noel, Benjamin Policriti, Alberto Clepet, Christian Casagrande, Alberto Choisne, Nathalie Aubourg, Sebastien Vitulo, Nicola Jubin, Claire Vezzi, Alessandro Legeai, Fabrice Hugueney, Philippe Dasilva, Corinne Horner, David Mica, Erica Jublot, Delphine Poulain, Julie Bruyere, Clemence Billault, Alain Segurens, Beatrice Gouyvenoux, Michel Ugarte, Edgardo Cattonaro, Federica Anthouard, Veronique Vico, Virginie Del Fabbro, Cristian Alaux, Michael Di Gaspero, Gabriele Dumas, Vincent Felice, Nicoletta Paillard, Sophie Juman, Irena Moroldo, Marco Scalabrin, Simone Canaguier, Aurelie Le Clainche, Isabelle Malacrida, Giorgio Durand, Eleonore Pesole, Graziano Laucou, Valerie Chatelet, Philippe Merdinoglu, Didier Delledonne, Massimo Pezzotti, Mario Lecharny, Alain Scarpelli, Claude Artiguenave, Francois Pe, M Enrico Valle, Giorgio Morgante, Michele Caboche, Michel Adam-Blondon, Anne-Francoise Weissenbach, Jean Quetier, Francis Wincker, Patrick French-Italian Public Consortium for Grapevine Genome Characterization England Nature Nature. 2007 Sep 27;449(7161):463-7. Epub 2007 Aug 26. 88

[58] JGI. The joint genome institute. 52

[59] M. Kalyna and A. Barta. A plethora of plant serine/arginine-rich proteins: redundancy or evolution of novel gene functions? *Biochem Soc Trans*, 32(Pt

References

4):561–4, 2004. Kalyna, M Barta, A Research Support, Non-U.S. Gov't England Biochemical Society transactions Biochem Soc Trans. 2004 Aug;32(Pt 4):561-4. 51, 101

[60] M. Kalyna, S. Lopato, V. Voronin, and A. Barta. Evolutionary conservation and regulation of particular alternative splicing events in plant sr proteins. *Nucleic Acids Res*, 34(16):4395–405, 2006. Kalyna, Maria Lopato, Sergiy Voronin, Viktor Barta, Andrea Research Support, Non-U.S. Gov't England Nucleic acids research Nucleic Acids Res. 2006;34(16):4395-405. Epub 2006 Aug 26. 49

[61] E. Kim, A. Magen, and G. Ast. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res*, 35(1):125–31, 2007. Kim, Eddo Magen, Alon Ast, Gil Comparative Study Research Support, Non-U.S. Gov't England Nucleic acids research Nucleic Acids Res. 2007;35(1):125-31. Epub 2006 Dec 7. 48, 71, 90, 106

[62] Eddo Kim, Amir Goren, and Gil Ast. Alternative splicing: current perspectives. *Bioessays*, 30(1):38–47, Jan 2008. 90

[63] C. V. Kirchhamer, C. H. Yuh, and E. H. Davidson. Modular cis-regulatory organization of developmentally expressed genes: two genes transcribed territorially in the sea urchin embryo, and additional examples. *Proc Natl Acad Sci U S A*, 93(18):9322–8, 1996. Kirchhamer, C V Yuh, C H Davidson, E H HD-05753/HD/NICHD NIH HHS/United States Research Support, U.S. Gov't, P.H.S. Review United states Proceedings of the National Academy of Sciences of the United States of America Proc Natl Acad Sci U S A. 1996 Sep 3;93(18):9322-8. 42

[64] A R Krainer, G C Conway, and D Kozak. Purification and characterization of pre-mrna splicing factor sf2 from hela cells. *Genes Dev*, 4(7):1158–71, Jul 1990. 48

[65] A. Labadorf, A. Link, M. F. Rogers, J. Thomas, A. S. Reddy, and A. Ben-Hur. Genome-wide analysis of alternative splicing in chlamydomonas reinhardtii. *BMC Genomics*, 11:114, 2010. 104, 106

[66] L. F. Lareau, M. Inada, R. E. Green, J. C. Wengrod, and S. E. Brenner. Unproductive splicing of sr genes associated with highly conserved and ultraconserved dna elements. *Nature*, 446(7138):926–9, 2007. Lareau, Liana F Inada, Maki Green, Richard E Wengrod, Jordan C Brenner, Steven E Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. England Nature Nature. 2007 Apr 19;446(7138):926-9. Epub 2007 Mar 14. 52, 89, 91, 101

[67] C. J. Lawrence, Q. Dong, M. L. Polacco, T. E. Seigfried, and V. Brendel. Maizegdb, the community database for maize genetics and genomics. *Nucleic Acids Res*, 32(Database issue):D393–7, Jan 1 2004. 51

[68] I. Letunic, T. Doerks, and P. Bork. Smart 6: recent updates and new developments. *Nucleic Acids Res*, 37(Database issue):D229–32, Jan 2009. 102

[69] Shengrong Lin, Gabriela Coutinho-Mansfield, Dong Wang, Shatakshi Pandit, and Xiang-Dong Fu. The splicing factor sc35 has an active role in transcriptional elongation. *Nat Struct Mol Biol*, 15(8):819–26, Aug 2008. 86

[70] S. Lockton and B. S. Gaut. Plant conserved non-coding sequences and paralogue evolution. *Trends Genet*, 21(1):60–5, 2005. Journal Article Research Support, U.S. Gov't, Non-P.H.S. England Tig. 7

[71] Jennifer C Long and Javier F Caceres. The sr protein family of splicing factors: master regulators of gene expression. *Biochem J*, 417(1):15–27, Jan 2009. 48, 49, 89

[72] D. Longman, I. L. Johnstone, and J. F. Caceres. Functional characterization of sr and sr-related genes in caenorhabditis elegans. *EMBO J*, 19(7):1625–37, 2000. Longman, D Johnstone, I L Caceres, J F Research Support, Non-U.S. Gov't England The EMBO journal EMBO J. 2000 Apr 3;19(7):1625-37. 52, 89, 101

[73] N. Lopez-Bigas, S. De, and S. A. Teichmann. Functional protein divergence in the evolution of homo sapiens. *Genome Biol*, 9(2):R33, 2008. Lopez-Bigas, Nuria De, Subhajyoti Teichmann, Sarah A Research Support, Non-U.S. Gov't England Genome biology Genome Biol. 2008;9(2):R33. Epub 2008 Feb 15. 43

[74] Z. J. Lorkovic and A. Barta. Genome analysis: Rna recognition motif (rrm) and k homology (kh) domain rna-binding proteins from the flowering plant arabidopsis thaliana. *Nucleic Acids Res*, 30(3):623–35, Feb 1 2002. 49

[75] James L Manley and Adrian R Krainer. A rational nomenclature for serine/arginine-rich protein splicing factors (sr proteins). *Genes Dev*, 24(11):1073–4, Jun 2010. 87

[76] J. Masterson. Stomatal size in fossil plants: Evidence for polyploidy in majority of angiosperms. *Science*, 264(5157):421–424, 1994. Journal article. 6

[77] Aurélien Mazurie. Mutationshunter, 2005. 99

[78] C. Molina and E. Grotewold. Genome wide analysis of arabidopsis core promoters. *BMC Genomics*, 6(1):25, 2005. Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. England. 41

[79] B. Morgenstern. Dialign 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15(3):211–8, 1999. Comparative Study Journal Article England. 40, 98

[80] Burkhard Morgenstern. Dialign: multiple dna and protein sequence alignment at bibiserv. *Nucleic Acids Res*, 32(Web Server issue):W33–6, Jul 2004. 98

[81] S. M. Mount and H. K. Salz. Pre-messenger rna processing factors in the drosophila genome. *J Cell Biol*, 150(2):F37–44, Jul 24 2000. 52, 89, 101

[82] N. Mulder and R. Apweiler. Interpro and interproscan: tools for protein sequence classification and comparison. *Methods Mol Biol*, 396:59–70, 2007. Mulder, Nicola Apweiler, Rolf United States Methods in molecular biology (Clifton, N.J.) Methods Mol Biol. 2007;396:59-70. 65, 102

References

[83] Yoichi Murakami, Ruth V Spriggs, Haruki Nakamura, and Susan Jones. Piranha: a server for the computational prediction of rna-binding residues in protein sequences. *Nucleic Acids Res*, 38 Suppl:W412–6, Jul 2010. 67, 69

[84] M.Zola, J. Aluru, S. Stamatakis, and A. Ott. Large-scale maximum likelihood-based phylogenetic analysis on the ibm bluegene/l. In *SC '07: Proceedings of the 2007 ACM/IEEE conference on Supercomputing*, 2007. 103

[85] T. W. Nilsen and B. R. Graveley. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280):457–63, Jan 28 2010. 47, 48

[86] H. Nozaki, H. Takano, O. Misumi, K. Terasawa, M. Matsuzaki, S. Maruyama, K. Nishida, F. Yagisawa, Y. Yoshida, T. Fujiwara, S. Takio, K. Tamura, S. J. Chung, S. Nakamura, H. Kuroiwa, K. Tanaka, N. Sato, and T. Kuroiwa. A 100the hot-spring red alga cyanidioschyzon merolae. *BMC Biol*, 5:28, 2007. 51

[87] T. Obayashi, K. Kinoshita, K. Nakai, M. Shibaoka, S. Hayashi, M. Saeki, D. Shibata, K. Saito, and H. Ohta. Atted-ii: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in arabidopsis. *Nucleic Acids Res*, 35(Database issue):D863–9, 2007. Journal Article Research Support, Non-U.S. Gov't England. 94

[88] Kevin P O'Brien, Maido Remm, and Erik L L Sonnhammer. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res*, 33(Database issue):D476–80, Jan 2005. 99

[89] University of California Reagents. Phytozome, 2009. 51

[90] S Ohno. *Evolution by Gene Duplication*. Springer-Verlag, Heidelberg, 1970. 39

[91] S. Ouyang, W. Zhu, J. Hamilton, H. Lin, M. Campbell, K. Childs, F. Thibaud-Nissen, R. L. Malek, Y. Lee, L. Zheng, J. Orvis, B. Haas, J. Wortman, and C. R. Buell. The tigr rice genome annotation resource: improvements and new features. *Nucleic Acids Res*, 35(Database issue):D883–7, 2007. Ouyang, Shu Zhu, Wei Hamilton, John Lin, Haining Campbell, Matthew Childs, Kevin Thibaud-Nissen, Francoise Malek, Renae L Lee, Yuandan Zheng, Li Orvis, Joshua Haas, Brian Wortman, Jennifer Buell, C Robin Research Support, U.S. Gov't, Non-P.H.S. England Nucleic acids research Nucleic Acids Res. 2007 Jan;35(Database issue):D883-7. Epub 2006 Dec 1. 51

[92] S. G. Palusa, G. S. Ali, and A. S. Reddy. Alternative splicing of pre-mrnas of arabidopsis serine/arginine-rich proteins: regulation by hormones and stresses. *Plant J*, 49(6):1091–107, 2007. Palusa, Saiprasad Goud Ali, Gul Shad Reddy, Anireddy S N Research Support, U.S. Gov't, Non-P.H.S. England The Plant journal : for cell and molecular biology Plant J. 2007 Mar;49(6):1091-107. Epub 2007 Feb 22. 49, 91

[93] S. G. Palusa and A. S. Reddy. Extensive coupling of alternative splicing of pre-mrnas of serine/arginine (sr) genes with nonsense-mediated decay. *New Phytol*, 185(1):83–9, Jan 2009. 49

[94] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, 40(12):1413–5, 2008. Pan, Qun Shai, Ofer Lee, Leo J Frey, Brendan J Blencowe, Benjamin J Research Support, Non-U.S. Gov't United States Nature genetics Nat Genet. 2008 Dec;40(12):1413-5. Epub 2008 Nov 2. 48

[95] W R Pearson and D J Lipman. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85(8):2444–8, Apr 1988. 99

[96] D. Philipps, A. M. Celotto, Q. Q. Wang, R. S. Tarng, and B. R. Graveley. Arginine/serine repeats are sufficient to constitute a splicing activation domain. *Nucleic Acids Res*, 31(22):6502–8, Nov 15 2003. 48

[97] R. L. Poole. The tair database. *Methods Mol Biol*, 406:179–212, 2007. 51

[98] D. Portal, J. M. Espinosa, G. S. Lobo, S. Kadener, C. A. Pereira, M. De La Mata, Z. Tang, R. J. Lin, A. R. Kornblihtt, F. E. Baralle, M. M. Flawia, and H. N. Torres. An early ancestor in the evolution of splicing: a trypanosoma cruzi serine-arginine-rich protein (tcsr) is functional in cis-splicing. *Mol Biochem Parasitol*, 127(1):37–46, Mar 2003. 49

[99] A. S. Reddy. Plant serine/arginine-rich proteins and their role in pre-mrna splicing. *Trends Plant Sci*, 9(11):541–7, Nov 2004. 47

[100] Stefan A Rensing, Julia Ick, Jeffrey A Fawcett, Daniel Lang, Andreas Zimmer, Yves Van de Peer, and Ralf Reski. An ancient genome duplication contributed to the abundance of metabolic genes in the moss physcomitrella patens. *BMC Evol Biol*, 7:130, 2007. 88

[101] S. Y. Rhee, W. Beavis, T. Z. Berardini, G. Chen, D. Dixon, A. Doyle, M. Garcia-Hernandez, E. Huala, G. Lander, M. Montoya, N. Miller, L. A. Mueller, S. Mundodi, L. Reiser, J. Tacklind, D. C. Weems, Y. Wu, I. Xu, D. Yoo, J. Yoon, and P. Zhang. The arabidopsis information resource (tair): a model organism database providing a centralized, curated gateway to arabidopsis biology, research materials and community. *Nucleic Acids Res*, 31(1):224–8, 2003. Rhee, Seung Yon Beavis, William Berardini, Tanya Z Chen, Guanghong Dixon, David Doyle, Aisling Garcia-Hernandez, Margarita Huala, Eva Lander, Gabriel Montoya, Mary Miller, Neil Mueller, Lukas A Mundodi, Suparna Reiser, Leonore Tacklind, Julie Weems, Dan C Wu, Yihe Xu, Iris Yoo, Daniel Yoon, Jungwon Zhang, Peifen Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S. England Nucleic acids research Nucleic Acids Res. 2003 Jan 1;31(1):224-8. 93

[102] A. J. Roger and A. G. Simpson. Evolution: revisiting the root of the eukaryote tree. *Curr Biol*, 19(4):R165–7, Feb 24 2009. 101

[103] M. B. Roth, A. M. Zahler, and J. A. Stolk. A conserved family of nuclear phosphoproteins localized to sites of polymerase ii transcription. *J Cell Biol*, 115(3):587–96, Nov 1991. 48

References

[104] H. Shen, J. L. Kan, and M. R. Green. Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly. *Mol Cell*, 13(3):367–76, Feb 13 2004. 48

[105] P. J. Shepard and K. J. Hertel. The sr protein family. *Genome Biol*, 10(10):242, 2009. 48, 87, 91

[106] C. Simillion, K. Vandepoele, M. C. Van Montagu, M. Zabeau, and Y. Van de Peer. The hidden duplication past of arabidopsis thaliana. *Proc Natl Acad Sci U S A*, 99(21):13627–32, 2002. Journal Article Research Support, Non-U.S. Gov't United States. 6

[107] C. G. Simpson, S. Manthri, K. D. Raczynska, M. Kalyna, D. Lewandowska, B. Kusenda, M. Maronova, Z. Szweykowska-Kulinska, A. Jarmolowski, A. Barta, and J. W. Brown. Regulation of plant gene expression by alternative splicing. *Biochem Soc Trans*, 38(2):667–71, Apr 2010. 47, 48

[108] R V Spriggs, Y Murakami, H Nakamura, and S Jones. Protein function annotation from sequence: prediction of residues interacting with rna. *Bioinformatics*, 25(12):1492–7, Jun 2009. 67, 69

[109] A. Stamatakis. Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–90, Nov 1 2006. 103

[110] H. Stanley Kim, Y. Yu, E. C. Snesrud, L. P. Moy, L. D. Linford, B. J. Haas, W. C. Nierman, and J. Quackenbush. Transcriptional divergence of the duplicated oxidative stress-responsive genes in the arabidopsis genome. *Plant J*, 41(2):212–20, 2005. Stanley Kim, H Yu, Yan Snesrud, Erik C Moy, Linda P Linford, Lara D Haas, Brian J Nierman, William C Quackenbush, John Research Support, U.S. Gov't, Non-P.H.S. England The Plant journal : for cell and molecular biology Plant J. 2005 Jan;41(2):212-20. 43

[111] J. W. Stiller and B. D. Hall. The origin of red algae: implications for plastid evolution. *Proc Natl Acad Sci U S A*, 94(9):4520–5, Apr 29 1997. 57

[112] D. F. Stojdl and J. C. Bell. Sr protein kinases: the splice of life. *Biochem Cell Biol*, 77(4):293–8, 1999. 48

[113] A. R. Subramanian, M. Kaufmann, and B. Morgenstern. Dialign-tx: greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms Mol Biol*, 3:6, 2008. Subramanian, Amarendran R Kaufmann, Michael Morgenstern, Burkhard England Algorithms for molecular biology : AMB Algorithms Mol Biol. 2008 May 27;3:6. 8, 9, 40, 95, 98, 102

[114] A. R. Subramanian, J. Weyer-Menkhoff, M. Kaufmann, and B. Morgenstern. Dialign-t: an improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics*, 6:66, 2005. Subramanian, Amarendran R Weyer-Menkhoff, Jan Kaufmann, Michael Morgenstern, Burkhard Research Support, Non-U.S. Gov't England BMC bioinformatics BMC Bioinformatics. 2005 Mar 22;6:66. 8, 40, 95, 98, 102

[115] R. Tacke, Y. Chen, and J.L. Manley. Sequence-specific rna binding by an sr protein requires rs domain phosphorylation: creation of an srp40-specific splicing enhancer. *Proceedings of the National Academy of Sciences of the United States of America*, 94(4):1148, 1997. 89

[116] N. Tanabe, K. Yoshimura, A. Kimura, Y. Yabuta, and S. Shigeoka. Differential expression of alternatively spliced mrnas of arabidopsis sr protein homologs, atsr30 and atsr45a, in response to environmental stress. *Plant Cell Physiol*, 48(7):1036–49, 2007. Tanabe, Noriaki Yoshimura, Kazuya Kimura, Ayako Yabuta, Yukinori Shigeoka, Shigeru Research Support, Non-U.S. Gov't Japan Plant cell physiology Plant Cell Physiol. 2007 Jul;48(7):1036-49. Epub 2007 Jun 6. 52

[117] Haibao Tang, Xiyin Wang, John E Bowers, Ray Ming, Maqsudul Alam, and Andrew H Paterson. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res*, 18(12):1944–54, Dec 2008. 68

[118] Z. Tang, N. F. Kaufer, and R. J. Lin. Interactions between two fission yeast serine/arginine-rich proteins and their modulation by phosphorylation. *Biochem J*, 368(Pt 2):527–34, Dec 1 2002. 49, 52, 86, 101

[119] J. D. Thompson, T. J. Gibson, and D. G. Higgins. Multiple sequence alignment using clustalw and clustalx. *Curr Protoc Bioinformatics*, Chapter 2:Unit 2 3, Aug 2002. 101

[120] Abel Ureta-Vidal, Laurence Ettwiller, and Ewan Birney. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat Rev Genet*, 4(4):251–62, Apr 2003. 3

[121] W van Der Houven Van Oordt, K Newton, G R Screaton, and J F Cáceres. Role of sr protein modular domains in alternative splicing specificity in vivo. *Nucleic Acids Res*, 28(24):4822–31, Dec 2000. 89

[122] T. J. Vision, D. G. Brown, and S. D. Tanksley. The origins of genomic duplications in arabidopsis. *Science*, 290(5499):2114–7, 2000. Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. United states. 6

[123] Bing-Bing Wang and Volker Brendel. Genomewide comparative analysis of alternative splicing in plants. *Proc Natl Acad Sci U S A*, 103(18):7175–80, May 2006. 48

[124] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–6, Nov 27 2008. 48

[125] J. Wang, L. Tian, A. Madlung, H. S. Lee, M. Chen, J. J. Lee, B. Watson, T. Kagochi, L. Comai, and Z. J. Chen. Stochastic and epigenetic changes of gene expression in arabidopsis polyploids. *Genetics*, 167(4):1961–73, 2004. Gm67015/gm/nigms Journal Article Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S. United States. 42

[126] J. F. Wendel. Genome evolution in polyploids. *Plant Mol Biol*, 42(1):225–49, 2000. Journal Article Research Support, U.S. Gov't, Non-P.H.S. Review Netherlands. 6

References

[127] S. R. Wessler, T. E. Bureau, and S. E. White. Ltr-retrotransposons and mites: important players in the evolution of plant genomes. *Curr Opin Genet Dev*, 5(6):814–21, 1995. Wessler, S R Bureau, T E White, S E Comparative Study Research Support, U.S. Gov't, P.H.S. Review England Current opinion in genetics development Curr Opin Genet Dev. 1995 Dec;5(6):814-21. 42

[128] V. Wood, R. Gwilliam, M. A. Rajandream, M. Lyne, R. Lyne, A. Stewart, J. Sgouros, N. Peat, J. Hayles, S. Baker, D. Basham, S. Bowman, K. Brooks, D. Brown, S. Brown, T. Chillingworth, C. Churcher, M. Collins, R. Connor, A. Cronin, P. Davis, T. Feltwell, A. Fraser, S. Gentles, A. Goble, N. Hamlin, D. Harris, J. Hidalgo, G. Hodgson, S. Holroyd, T. Hornsby, S. Howarth, E. J. Huckle, S. Hunt, K. Jagels, K. James, L. Jones, M. Jones, S. Leather, S. McDonald, J. McLean, P. Mooney, S. Moule, K. Mungall, L. Murphy, D. Niblett, C. Odell, K. Oliver, S. O'Neil, D. Pearson, M. A. Quail, E. Rabbinowitsch, K. Rutherford, S. Rutter, D. Saunders, K. Seeger, S. Sharp, J. Skelton, M. Simmonds, R. Squares, S. Squares, K. Stevens, K. Taylor, R. G. Taylor, A. Tivey, S. Walsh, T. Warren, S. Whitehead, J. Woodward, G. Volckaert, R. Aert, J. Robben, B. Grymonprez, I. Weltjens, E. Vanstreels, M. Rieger, M. Schafer, S. Muller-Auer, C. Gabel, M. Fuchs, A. Dusterhoft, C. Fritzc, E. Holzer, D. Moestl, H. Hilbert, K. Borzym, I. Langer, A. Beck, H. Lehrach, R. Reinhardt, T. M. Pohl, P. Eger, W. Zimmermann, H. Wedler, R. Wambutt, B. Purnelle, A. Goffeau, E. Cadieu, S. Dreano, S. Gloux, V. Lelaure, S. Mottier, F. Galibert, S. J. Aves, Z. Xiang, C. Hunt, K. Moore, S. M. Hurst, M. Lucas, M. Rochet, C. Gaillardin, V. A. Tallada, A. Garzon, G. Thode, R. R. Daga, L. Cruzado, J. Jimenez, M. Sanchez, F. del Rey, J. Benito, A. Dominguez, J. L. Revuelta, S. Moreno, J. Armstrong, S. L. Forsburg, L. Cerutti, T. Lowe, W. R. McCombie, I. Paulsen, J. Potashkin, G. V. Shpakovski, D. Ussery, B. G. Barrell, and P. Nurse. The genome sequence of schizosaccharomyces pombe. *Nature*, 415(6874):871–80, Feb 21 2002. 62

[129] J. Y. Wu and T. Maniatis. Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell*, 75(6):1061–70, Dec 17 1993. 48, 57, 86

[130] Y. L. Xiao, S. R. Smith, N. Ishmael, J. C. Redman, N. Kumar, E. L. Monaghan, M. Ayele, B. J. Haas, H. C. Wu, and C. D. Town. Analysis of the cdnas of hypothetical genes on arabidopsis chromosome 2 reveals numerous transcript variants. *Plant Physiol*, 139(3):1323–37, Nov 2005. 48

[131] Z. Yang. Paml: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, 13(5):555–6, 1997. Yang, Z Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S. England Computer applications in the biosciences : CABIOS Comput Appl Biosci. 1997 Oct;13(5):555-6. 99

[132] A. M. Zahler, W. S. Lane, J. A. Stolk, and M. B. Roth. Sr proteins: a conserved family of pre-mrna splicing factors. *Genes Dev*, 6(5):837–47, may 1992. 48, 49

[133] E. M. Zdobnov and R. Apweiler. Interproscan–an integration platform for the signature-recognition methods in interpro. *Bioinformatics*, 17(9):847–8, 2001. Zdobnov, E M Apweiler, R England Bioinformatics (Oxford, England) Bioinformatics. 2001 Sep;17(9):847-8. 102

[134] X. Zhang, J. Yazaki, A. Sundaresan, S. Cokus, S. W. Chan, H. Chen, I. R. Henderson, P. Shinn, M. Pellegrini, S. E. Jacobsen, and J. R. Ecker. Genome-wide high-resolution mapping and functional analysis of dna methylation in arabidopsis. *Cell*, 126(6):1189–201, 2006. Zhang, Xiaoyu Yazaki, Junshi Sundaresan, Ambika Cokus, Shawn Chan, Simon W-L Chen, Huaming Henderson, Ian R Shinn, Paul Pellegrini, Matteo Jacobsen, Steve E Ecker, Joseph R GM60398/GM/United States NIGMS HG003523/HG/United States NHGRI T32 GM08666/GM/United States NIGMS Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. United States Cell Cell. 2006 Sep 22;126(6):1189-201. Epub 2006 Aug 31. 82

[135] Xiao-Ning Zhang and Stephen M Mount. Two alternatively spliced isoforms of the arabidopsis sr45 protein have distinct roles during normal plant development. *Plant Physiol*, 150(3):1450–8, Jul 2009. 87

[136] Z. Zhang, J. Gu, and X. Gu. How much expression divergence after yeast gene duplication could be explained by regulatory motif evolution? *Trends Genet*, 20(9):403–7, 2004. Journal Article Research Support, U.S. Gov't, P.H.S. England Tig. 42

[137] Philip Zimmermann, Oliver Laule, Josy Schmitz, Tomas Hruz, Stefan Bleuler, and Wilhelm Gruissem. Genevestigator transcriptome meta-analysis and biomarker search using rice and barley gene expression databases. *Mol Plant*, 1(5):851–7, Sep 2008. 22, 43, 71, 72, 73, 74, 75

[138] D. J. Zwickl. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Master's thesis, University of Texas at Austin, 2006. 103

Eidesstattliche Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit - einschlielich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen - noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Thomas Wiehe betreut worden.

Teilpublikationen

Richardson, D., Rogers, M., Reddy, ASN. (2010) Comparative analysis of serine/arginine-rich proteins across 27 eukaryotes: insights into subfamily classification and extent of alternative splicing, *Manuscript submitted*

Richardson, D. and Wiehe, T. (2009) Properties of sequence conservation in Upstream Regulatory and Protein Coding Sequences among Paralogs in *Arabidopsis thaliana* RECOMB-CG 5817: 217-228.

Curriculum vitae

Persönliche Angaben

Name	Dale N. Richardson III
Geburtsdatum	27.05.1979
Geburtsort	Great Lakes, Illinois
Staatsbürgerschaft	United States of America

Schulausbildung

08/1985 - 05/1993	Volksschule, El Paso, Texas
08/1993 - 05/1997	Andress High School, El Paso, Texas
05/1997	Abschluss der High School

Hochschulausbildung

05/2000 - 06/2006	Studium der Biologie an der Colorado State University
06/2006	Diplom (B.S., M.S.)
09/2006 - 09/2010	Promotion am Institut für Genetik, Arbeitsgruppe für Populationsgenetik und Bioinformatik der Universität zu Köln
11/2010	Voraussichtlicher Abschluss der Promotion