Transcriptome analysis indicates considerable divergence in alternative splicing between duplicated genes in *Arabidopsis thaliana*.

David C. Tack*, William R. Pitchers §, Keith L. Adams*

*Department of Botany, 6270 University Boulevard, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada

§Department of Zoology, Program in Ecology Evolutionary Biology and Behavior, Michigan State University, East Lansing, MI, 48824, USA
GEO Series accession number GSE57579

Running title: Alternative splicing in duplicated genes

Key Words: genome duplication, gene duplication, alternative splicing, genome evolution, nonsense-mediated decay

Author for Correspondence:
Keith Adams, Department of Botany, University of British Columbia, Vancouver, BC, Canada
604-822-2355, keitha@mail.ubc.ca
Abstract

Gene and genome duplication events have created a large number of new genes in plants that can diverge by evolving new expression profiles and functions (neofunctionalization) or dividing extant ones (subfunctionalization). Alternative splicing (AS) generates multiple types of mRNA from a single type of pre-mRNA by differential intron splicing. It can result in new protein isoforms or down-regulation of gene expression by transcript decay. Using RNA-seq we investigated the degree to which alternative splicing patterns are conserved between duplicated genes in Arabidopsis thaliana. Our results revealed that 30% of AS events in alpha whole genome duplicates, and 33% of AS events in tandem duplicates, are qualitatively conserved within leaf tissue. Loss of ancestral splice forms, as well as asymmetric gain of new splice forms, may account for this divergence. Conserved events had different frequencies, as only 31% of shared AS events in alpha whole genome duplicates and 41% of shared AS events in tandem duplicates had similar frequencies in both paralogs, indicating considerable quantitative divergence. Analysis of published RNA-seq data from nonsense mediated decay (NMD) mutants indicated that 85% of alpha whole genome duplicates and 89% of tandem duplicates have diverged in their AS-induced NMD. Our results indicate that alternative splicing shows a high degree of divergence between paralogs such that qualitatively conserved alternative splicing events tend to have quantitative divergence. Divergence in AS patterns between duplicates may be a mechanism of regulating expression level divergence.
INTRODUCTION

Gene and genome duplications have created large numbers of duplicated genes, some of which have diverged in expression patterns and functions. All vertebrates have experienced at least two rounds of whole genome duplication, whereas plants persist with a characteristic propensity for iterative rounds of polyploidy (reviewed in Kasahara 2007, Blanc and Wolfe 2004a, Cui et al. 2006, Jiao et al. 2011). New duplicates must remain functionally relevant or be deleted or pseudogenized. Several models for duplicate gene retention and subsequent fates have been proposed, including genetic redundancy, gene dosage balance, genetic robustness, and divergence of protein sequence and expression patterns (reviewed in Sémon and Wolfe 2007; Han 2009). Complementary degenerate mutations may knock out one or more functions or expression patterns in each duplicate, referred to as subfunctionalization, further specializing each duplicate to their now partitioned function (Force et al. 1999, reviewed in Conant and Wolfe 2008). Neofunctionalization is the generation of a new function or expression pattern in one duplicate whereas the other copy retains the ancestral expression pattern or function.

In plants there has been particular interest in characterizing the fates of genes duplicated by whole genome duplication events. In Arabidopsis thaliana previous work has defined and examined the alpha whole genome (WG) duplicates, which originated from a polyploidy event at the base of the Brassicaceae family (Blanc et al. 2003; Bowers et al. 2003). These duplicates are a discrete set of paralogs that originated at the same time and some of them exhibit expression profile divergence (Blanc and Wolfe 2004b, Casneuf et al. 2006, Liu et al. 2011). Expression patterns of similar sets of genes derived by whole genome duplication have been characterized in other plants (Schnable et al. 2011, Renny-Byfield et al. 2014). Tandem duplicates contrast to
alpha WG duplicates in that they arise from small-scale duplications, often by unequal crossing over, and formed at various times during the evolution of most plant lineages.

Alternative splicing is integral to gene expression, as differential splicing of the primary mRNA transcript can alter protein functionality or transcript level. Much of the diversity of the eukaryotic proteome can be attributed to alternative splicing (Nilsen and Gravely 2010) where more transcripts and protein products exist than genes. The most prevalent class of alternative splicing in plants is intron retention (IR) where an intron is not spliced out of the transcript, whereas the most common type in animals is exon skipping (SKIP) where one or more modular exons are skipped to produce different transcript isoforms (reviewed in Reddy et al. 2013). Alternative donor (ALTD) and acceptor events (ALTA) cause a change in the 5' and 3' exon boundaries respectively, whereas an alternative position (ALTP) is a combination of both an alternative donor and acceptor. AS affects over 50% of genes in Arabidopsis thaliana and Oryza sativa; exact percentages depend on whether all genes or only intron-containing genes are counted (Lu et al. 2010; Marquez et al. 2012). Environmental and biological stresses are often met with changes in splicing in plants (e.g. Filichkin et al. 2010, reviewed in Staiger and Brown 2013). Numerous examples of alternative splicing being tied to the circadian clock exist (Sanchez et al. 2010, Seo et al. 2012, reviewed in Filichkin and Mockler 2012) as well as studies highlighting its key role in expression level regulation via nonsense-mediated decay (NMD) (Drechsel et al. 2013, Kalyna et al. 2011). For example, AtGRP8 auto-regulates itself via AS induced NMD during elevated protein levels (Schöning et al. 2008). While some alternatively spliced transcripts may simply be byproducts of random interaction between splicing factors, more functionally characterized forms in plants are continuing to be found mirroring the wealth of characterized AS isoforms found in animals.
A previous study used RT-PCR based methods to attempt to ascertain how divergent or conserved the AS patterns are among paralogs in *Arabidopsis thaliana* (Zhang et al. 2010). However this approach was limited by being confined to a small set of known events for which there was prior annotation. In the advent of RNA-seq, which allows for both qualitative discovery (Marquez et al. 2012, Filichkin et al. 2010) and quantitative comparisons, it is possible to re-evaluate the degree to which AS events have been conserved between duplicates. Thus studying conservation of alternative splicing in a single tissue type is a reflection of the regulatory divergence that has occurred between duplicates in the same trans environment. Alternative splicing is regulated by cis-elements known as exon splicing enhancers and silencers (ESEs and ESSs) and intron splicing enhancers and silencers (ISEs and ISSs) (Chen and Manley 2009, Huelga et al. 2012, reviewed in Reddy et al. 2013). Both pair members containing the same AS events and expressing them at similar frequencies within a tissue implies a lack of such regulatory divergence in AS, whereas qualitative or quantitative variation in AS between paralogs in the same tissue would show they have experienced asymmetrical change in their regulation.

We analyzed AS patterns in two types of duplicated genes in *Arabidopsis thaliana*, alpha WG duplicates and tandem duplicates, in leaf tissue using RNA-seq analysis to assess both qualitative and quantitative conservation. Our findings indicate considerable qualitative divergence in AS patterns between duplicates in leaves. Of those AS events that are qualitatively conserved, a large majority occur at different frequencies and show quantitative divergence. To assess the relationship between gene duplication and nonsense-mediated decay associated alternative splicing events, we analyzed published RNA-seq data from NMD mutants that indicated that most duplicates show AS induced NMD asymmetry.
MATERIALS AND METHODS

Plant Growth, RNA Extraction, and Library Preparation

*Arabidopsis thaliana* (Col-0) was grown at 20ºC with a 16h/8h light/dark cycle and 70% humidity in a growth chamber. The first true leaves were harvested at 20 days after germination and flash frozen in liquid nitrogen before being stored at -80ºC. Three biological replicates were used which consisted of pools of different plants from the same growth chamber. Total RNA was isolated from the first true leaves with the Ambion RNAqueous kit (AM1912) in conjunction with the Ambion Plant RNA Isolation Aid (AM9690). Extracted total RNA was DNase treated with the Ambion Turbo DNA-free kit (AM1907) and visualized on a 2% agarose gel to assess quality. Libraries were prepared with the Illumina TruSeq RNA Sample Prep Kit. The low-throughput (<48 samples) protocol was followed, using 4 µg total RNA input for each sample, and allowing fragmentation to proceed for 2 minutes.

Library Sequencing and Mapping/Processing

Libraries were sequenced to obtain 100 bp paired-end reads on an Illumina HiSeq 2000, multiplexing three libraries over 1.5 lanes. This yielded a total of 91315717, 90873817, and 168071194 million paired-end reads per sample. The RNA-seq reads were mapped against the *Arabidopsis thaliana* genome version Ensemble TAIR 10 from downloaded as an igenome from [http://tophat.ccbcb.umd.edu/igenomes.shtml](http://tophat.ccbcb.umd.edu/igenomes.shtml) using GSNAP version 2014-01-21. The command `gsnap -d thaliana --pairmax-rna=10000 --localsplicedist=10000 --clip-overlap -N 1 --nthreads=6 - -quality-protocol=sanger --ordered --format=sam --nofails R1.fastq R2.fastq > mapped.sam` was used to map. Mapped reads were processed with a set of custom Python scripts to both ascertain which genes they originated from, as well any splicing patterns, calling all splicing events based
on the representative gene models. A total of 84864583, 79167068, and 150297611 paired-end reads were mapped. Scripts are available on request. The RNA-seq data have been deposited in NCBI's Gene Expression Omnibus (Edgar et al. 2002) and are accessible through GEO Series accession number GSE57579 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=cdkjkawipzmdmp&acc=GSE57579).

**Equivalent Junction and Event Calling**

Exons from the representative models for alpha WGD and tandem duplicate pair genes (Liu et al. 2011) were compared against each other using BLASTn (-task dc-megablast). For each set of reciprocally matched and contiguous exons, an equivalent junction pairing was logged. This allowed investigation of events even if there has been a rearrangement or re-numeration of exons between duplicates' models, such as exon loss or fission/fusion between paralogs at homologous junctions. Only events that occurred within these junction pairs were compared. For AS event calling the minimum number of AS reads required was 3, with at least 1 read in support in each replicate as well as constitutive coverage of the junction. Overall 95% of all events had 5 or more total AS reads and 85% of events had 8 or more total AS reads. Analyses were repeated requiring 5 and 8 reads respectively (supplementary table 1) with almost no effect on the results.

**Analyses of Leaf RNA-Seq Data**

Expression data were used to exclude from the dataset any gene pairs with a member that had zero reads in any replicate. Gene pairs were also excluded if either member was part of the 5% of genes with the highest variance in number of reads among the three replicates. This filter thus excludes from our analysis gene pairs for which the repeatability of the read counts was low. The AS data were then further filtered to remove AS events wherein the expression of one
partner was represented by fewer than 1 read per replicate.

Qualitative conservation was assessed using a binary rule; the pattern of splicing was defined as conserved at a given junction if both paralogs had both constitutively and alternatively spliced reads mapped to the same equivalent junction, and as divergent if only one paralog had alternatively spliced reads. After counting qualitatively conserved junctions, the conserved junctions were tested for quantitative conservation. These tests were run as a logistic regression at each junction, using the model:

\[
pr(alt) \sim bi(p = \logit^{-1}(b_0 + b_{\text{paralog}}))
\]

where \(pr(alt)\) is the proportion of alternative splicing events and \(b_{\text{paralog}}\) is the paralog identity. Logistic regression was used because the proportion of alternative splicing events is by definition bounded between 0 and 1, and our outcome (difference or no difference between paralogs) is binary. Junctions were defined as quantitatively divergent if the model provided statistical support (i.e. a \(p\)-value <0.05) for the quantitative difference between the paralogs in the pattern of constitutively and alternatively spliced reads, and as quantitatively conserved otherwise (i.e. in the absence of statistical evidence for the existence of a difference between paralogs, a \(p\)-value \(\geq0.05\), we assume conservation). The number and percentage of conserved junctions are reported in Table 1. To visualize the difference in levels of a particular AS event between paralogs (i.e. effect size), the difference in number of AS reads was expressed as fold difference after standardizing for the mapped expression level of each gene (Figure 1).

Hypotheses about gene ontology were addressed using the web-based Gene Ontology enrichment analysis and visualization tool (GOriilla) at http://cbl-gorilla.cs.technion.ac.il/ (Eden et al. 2007; 2009). Lists of genes that our analysis classified as conserved or divergent were input into the GOriilla web tool to search for enrichment in GO terms. To determine if there was a
difference in alternative splicing conservation rate between events located in genes' UTR regions versus coding regions, the location of each event was used as a binary factor. Logistic regression was used to model the conservation status as a function of event location, and then a linear regression was used to model the size of difference in expression pattern between paralogs (fold change) as a function of event location.

**Analyses of NMD data**

The occurrence of nonsense-mediated decay (NMD) was investigated using data from Drechsel *et al.* (2013). Genes were classified as NMD positive if they were reported as showing NMD via alternative splicing in one or more mutant conditions. Genes where no significant increase in frequency in a splicing event in mutant conditions was reported were classified as NMD negative. NMD presence/absence data was used to test for differences between singleton and duplicate genes in the percentage of NMD positive genes using a Chi-Square test. These data were also used to quantify the percentages of tandem and alpha WG duplicates where NMD status was asymmetric, i.e. where one paralog was NMD positive and the other was NMD negative.

All statistical analyses were conducted in R (version 3.0.2) and analysis scripts are archived along with our data at Dryad DOIXXXX and at [https://github.com/willpitchers/petulant-dubstep](https://github.com/willpitchers/petulant-dubstep).

**Generation of Singletons and Duplicate sets**

Only terminal duplicate pairs without subsequent gene duplication were used to study alternative splicing. These were found by generating gene families following the analytical
procedure described in Liu et al. (2011). First, Arabidopsis thaliana gene families were obtained from PLAZA 1.0 (Proost et al. 2009) and a 50% consensus tree topology was obtained for each gene family based on 100 replicates of bootstrapping maximum-likelihood analyses by RAxML v.7.0.0 with an amino acid substitution matrix WAG and gamma-distributed rate variation (Stamatakis 2006). Then 2584 alpha-WG duplicate pairs identified by Blanc et al. (2003) and 1826 clusters of tandem duplicates identified by Liu et al. (2011) were used as the backbone for us to pull out the terminal duplicate pairs without subsequent gene duplications from each gene family consensus tree topology. A total of 1771 WG duplicate pairs and 1179 tandem duplicate pairs were used for subsequent analyses.

**RT-PCR Confirmations**

Primer sets were designed around 20 junctions chosen at random with one or more alternative events. The junction was encompassed to amplify the constitutive and alternative splicing(s) of the junction. Three cDNA pools were made via the Invitrogen SuperScript First Strand Synthesis for RT-PCR kit using the same RNA samples used to prepare Illumina libraries. PCR was performed using FroggaBio Ultra Pure Taq PCR Master Mix with dye. The PCR cycling conditions were 94° for 3 min; 30-35 cycles of 94° for 30 sec, 54° for 30 sec, 72° for 30 sec; and 72° for 7 min. PCR products were visualized on 1.2% agarose gels. Primer sets are listed in Table S1.

**RESULTS**

**Conservation of alternative splicing within paralog pairs**

Three biological replicates of RNA-seq from leaves of Arabidopsis thaliana were used to
investigate the splicing patterns of two classes of paralogs, alpha WG duplicates and tandem duplicates. Reads were mapped against the TAIR 10 assembly with GSNAP (Wu and Watanabe 2005) and interpreted with a suite of custom Python and R scripts (see methods). Representative gene models were BLAST searched against each other to create a set of homologous junctions between paralog pairs where events could be directly compared. Only splicing events that were present in every replicate were accepted, resulting in 62,909 unique AS events between all genes. A total of 43236 of these events were IR, 8817 were ALTA, 4986 were ALTD, and 1508 were ALTP class events. While we detected 995 SKIP events, as well as 3367 events in other complex categories, the difficulty in assigning the event to one homologous junction shared between paralogs excluded using them from further analysis.

First, a qualitative analysis comparing simple presence/absence of splice events at homologous junctions (Table 1) revealed 30% percent of events are conserved between alpha WG duplicates and 33% of events are conserved between tandem duplicates. Intron retention is both the most common type of event and the most conserved between paralogs, with alternative acceptor being second in both regards. The rate of conservation for IR events was higher than those for all other classes of AS in both alpha WG duplicates and tandem duplicates, and this difference had statistical support (alpha WG duplicates $p < 0.001$, tandem duplicates $p < 0.001$, chi-squared). Conserved and divergent AS events are listed in Supporting Table S2.

Next, we tested for changes in the frequency of qualitatively conserved events between paralogs by modeling the alternative and constitutive read counts at homologous junctions using binary logistic regression. This used data from all three replicates to provide robust statistical support for these quantitative changes. Because we used relative frequency (i.e. number of alternative reads compared to constitutive reads at each junction) differences in absolute
expression of the gene will have minimal effects on this analysis. Any splice event-junction model which produced a \( p \)-value <0.05, and thus had statistical support for a difference between frequencies of alternative versus constitutive reads, was tallied as divergent. Quantitative conservation (or more precisely the absence of statistical evidence for quantitative divergence) was found to be 31\% in alpha WG duplicates and 41\% in tandems (Table 1), with tandems having a significantly higher rate of quantitative conservation (\( p=0.002 \), chi-squared). However the median effect size difference in the quantitative comparison is 4.1 fold for alpha WG duplicates and 10.2 for tandems (\( p=0.001 \), chi-squared), again with a significant difference. More of the events in tandems duplicates are quantitatively conserved than events in alpha WG duplicates, but the events in tandems that are quantitatively divergent show a wider range of disparity on average than alpha WG events that are quantitatively divergent. When the qualitative and quantitative results are combined, only a small fraction of the events are fully conserved (Table 1). The alpha WG duplicates have 9\% conservation on both levels whereas the tandem duplicates again have a significantly higher rate of conservation (\( p=0.0003 \), chi-squared) at 14\%. Conserved and divergent AS events are listed in Supporting Table S2.

No significant differences in the rate of qualitative or quantitative conservation were found between events located in untranslated regions of transcripts (5'UTR and 3'UTR) versus the events found in coding sequences of transcripts (CDS). Additionally there was no differential enrichment of GO categories or terms between genes with conserved AS events versus genes with divergent AS, either qualitatively or quantitatively.

To compare RNA-seq data to RT-PCR data, as well as to validate our computational pipeline, twenty sets of primers were designed around a random set of junctions that contained one or more alternative splicing events. The vast majority (17/20, 85\%) amplified the two or
more expected fragments (Figure 3) for the constitutive and alternative form(s) while the remaining 3 either failed to amplify or had unspecific amplification. Events from these three junctions were validated by EST support in plant GDB (Duvick et al. 2008, http://www.plantgdb.org/ASIP/), resulting in support for all 22 events over the 20 random junctions.

**CCA1 and LHY have functionally characterized, divergent alternative splicing**

The qualitative and quantitative results were surveyed for genes with alternative splicing events that had been functionally characterized. The alpha WG pair of *LATE ELONGATED HYPOCOTYL* (*LHY*; AT1G01060) and *CIRCADIAN CLOCK-ASSOCIATED1* (*CCA1*; AT2G46830) was found to display two types of asymmetric AS events: an alternative donor unique to the exon 4-5 junction in *CCA1*, and an intron retention event of intron 4 which is shared qualitatively, albeit at exceptionally low levels, in *LHY* (Figure 4). The exon 5-6 junction of *LHY* corresponds to the exon 4-5 junction of *CCA1*. *LHY* was found to have 9 IR reads and 429 constitutive reads among all 3 replicates (~2% of reads contain the IR), while the corresponding junction in *CCA1* was found to have 2186 intron retention reads, 1811 constitutive reads, plus 114 alternative donor reads (~53% of reads contain the IR). Thus there has been a large quantitative change in IR levels as well as a qualitative change in the ALTD event. The functionality of the intron retention event has been detailed in Seto et al. (2012) which demonstrated that retention of the intron in *CCA1* makes CCA1β, which binds to CCA1α or LHY, creating protein dimers with reduced DNA binding affinity and preventing formation of normal CCA1α homodimers, LHY homodimers, and CCA1α-LHY heterodimers. The alternative splicing of *CCA1* into CCA1β is suppressed in lower temperatures allowing CCA1α and LHY to function in coordinating the cold response. The intron retention event is clearly the ancestral
state, as the same event was found in CCA1 orthologs in Oryza sativa, Brachypodium distachyon, and Populus trichocarpa (Filichkin et al. 2010), indicating that LHY clearly has an almost complete loss of the event in leaf tissue and that it was not a novel gain in CCA1 post-duplication.

Conservation of alternative splicing induced NMD within paralog pairs

To explore the relationship between gene duplication and nonsense-mediated decay (NMD) associated alternative splicing events, we analyzed RNA-seq data from NMD mutants reported in Drechsel et al. (2013). All genes reported as showing NMD via alternative splicing in one or more mutant conditions were said to be NMD positive, whereas genes without a significant increase in frequency in a splicing event in mutant conditions were considered to not have AS induced NMD. Genes were sorted by duplication status (singleton or duplicate, see methods) to ask if duplication status influenced NMD status. Singleton genes had an overall chance of having an NMD event of 6.37%, whereas duplicates had a 7.08% chance, with a non-significant difference (p=0.1496, chi-squared). However this does not preclude that it may still be a common mechanism of divergent evolution between duplicated genes. Comparing the NMD status of alpha WG duplicate pairs, 207 pairs had one member with an AS event(s) associated with NMD whereas only 35 pairs had an event in both members (Table 2), indicating that 85% of pairs exhibited AS induced NMD asymmetry. Tandem duplicate pairs show a similar pattern with 89% of cases of AS induced NMD being present in only one member of a pair (58 pairs asymmetric, 7 pairs symmetric). This compliments the previous finding of low conservation of AS, indicating that much of the divergence may have implications in further diverging expression levels and profiles between paralogs.
DISCUSSION

In this study we analyzed 3951 AS events in tandem and alpha WG duplicate pairs using RNA-seq data. We found that alternative splicing patterns are infrequently conserved between paralogs in *Arabidopsis thaliana* leaves. The majority of alternative splicing events are neither conserved qualitatively nor quantitatively, pointing to alternative splicing as being a rapidly evolving aspect of gene expression after gene duplication. The amount of divergence and asymmetry of AS events between paralogs suggests post-duplication specialization of AS is generally favored over conservation. The differences between paralogs could be accounted for by one gene losing an ancestral AS event or one gene gaining a new AS event. Intron retention events show the most conservation, with 37-40% being qualitatively conserved, whereas alternative donors and acceptors are much less conserved at 5-15% (Table 1A). One hypothesis to explain this difference would be that alternative donors and acceptors may be more prone to diverge due to requiring more specification; alternate exon boundaries have to be selected by specific splicing factors and specific binding motifs whereas intron retention may be more robust to sequence changes. The tandem duplicates have a higher rate of quantitative conservation than alpha WG duplicates (41% vs. 31%) which could be due to them being younger duplicates in general. Interestingly, the average fold difference between quantitative AS event levels is much higher in tandem duplicates (10.2 fold) than alpha WG duplicates (4.1 fold). Events in tandem duplicates tend to be either quantitatively conserved or present at very disparate frequencies, whereas quantitatively divergent AS events in alpha WG duplicates have less difference on average. Non-uniformity in age may help explain the more stochastic patterns of the tandem duplicates.

The rate of event conservation in leaves was found to be ~37% for intron retention events
in alpha WG duplicates, which are the most common type of AS event, while a further refinement of this to include a quantitative dimension reduced this to 11%. The more nuanced use of both qualitative and quantitative conservation status hints at incomplete or partial regulatory divergence between the duplicates' alternative splicing patterns. Among the qualitatively conserved AS events, both paralogs keeping the event itself but expressing them at different frequencies is common. These data indicate that the event itself may be qualitatively conserved though the regulation of it may not be. From an evolutionary standpoint, while many forms are certainly lost, this still implies that as many as 37% of splice forms, in the case of intron retentions, are conserved between duplicates but may be present at different levels within; thus, they have experienced quantitative divergence in regulation. Likewise, conserved AS events from Zhang et al. (2010) that varied in AS by organ type have also experienced divergence in AS regulation, perhaps in some cases analogous to neo- and subfunctionalization. Thus rewiring, or both paralogs keeping the event itself but either expressing them at different frequencies, or in different tissues, seems common. Rather than an AS event being lost from one paralog, instead there is differential regulation of the event to allow paralog specific specialization in terms of when, where, and how much the event is expressed; the AS event may still be useful in both paralogs depending on how and when it is deployed. This implies changes in \textit{cis} elements are the primary mechanisms for divergence of AS patterns between paralogs, and that most of the molecular evolution responsible is likely centered around ESEs, ESSs, ISEs, and ISSs to produce different effects in paralogs under the same \textit{trans} environment. Changes in \textit{cis}-elements were implicated to account for most species-specific splicing patterns (Barbosa-Morias \textit{et al.} 2012) with \textasciitilde80% of non exon-skipping events attributable to divergent \textit{cis} architecture (McManus \textit{et al.} 2014). Collectively most changes in alternative splicing are thus directed by changes in \textit{cis}. 
We have detected extensive qualitative and quantitative divergence in AS events between duplicates. In a concurrent study, Shen et al. (2014) identified AS events in soybean and compared the number of AS events present in genes duplicated by a paleopolyploidy event. They found that the number of AS events was different in the vast majority of duplicate pairs. Of those pairs where both genes had the same number of AS events, about 2/3 had divergence in the types of AS events. However that study did not examine qualitative or quantitative conservation of individual AS events between duplicates which was the focus of this study. Thus the two studies give complementary perspectives on AS events in duplicated genes.

Divergence in expression patterns, commonly seen in duplicates (e.g., Blanc and Wolfe 2004; Casneuf et al. 2006; Liu et al. 2011), may have an underpinning in alternative splicing in some cases. Our finding that 85% of alpha WG pairs are NMD asymmetric suggests that some cases of expression divergence are due to NMD lowering the level of expression of one paralog. Complex changes in regulation, especially those that invoke alternative splicing, may play a role in coordinating differential expression of paralogs. While diverged, non-splicing related, *cis*-elements differentially activate or repress paralogs to achieve differential expression, NMD via AS may also play a role. The evolution of an AS induced NMD event to selectively shut down or reduce the level of one paralog in a given *trans*-environment may be an alternative to mutations in *cis*-regulatory elements that change the transcription level. Alternatively, asymmetric NMD between duplicates in some cases might be due to one gene acquiring mutations that lead to AS due to a relaxation of selection, but the NMD is not involved in regulating protein levels (because transcript and protein levels do not always correlate well.)

The case of *CCA1* and *LHY* is a clear example of divergent AS between duplicated genes. While the IR event is ancestral, *LHY* only maintains the event at a very low frequency in leaf
tissue, leaving two distinct possibilities. One possibility is that selection has not yet completely eliminated the event from the gene, i.e. incomplete loss. The second possibility is that event is maintained in LHY due to selection for AS in another tissue or stress, and the small amount of AS detected in leaf is a consequence of non-specific trans-environment interactions, i.e. incomplete specialization. Specifically, in another tissue, intron retention in LHY may serve a functional purpose and be at a higher rate, but the same cis-elements that allow intron retention in another tissue produce an incidental amount in leaf tissue. In either case, the pattern CCA1/LHY displays is less drastic than full partitioning/subfunctionalization of ancestral splicing patterns (Lister et al. 2001, Cusack and Wolfe 2007, Marshall et al. 2013) where an ancestral gene with two splice forms becomes two specialized genes, each with only one of the two ancestral forms becoming the new constitutive form. Duplicated genes acting in concert and having diverged AS patterns to control the circadian clock highlights both the potential reservoir of functionality that alternative splicing offers, as well as the specialization power that duplication allows. Although alternative splicing is often thought of as a response to stress, in the case of CCA1, it is the absence of AS that invokes a cold response pathway and the AS form which invokes normal homeostasis. LHY and CCA1 have several other known roles, including regulating SVP (SHORT VEGETATIVE PHASE) (Fujiwara 2008) and the C-REPEAT BINDING FACTOR (CBF) cold-response pathway (Dong et al. 2011).

Although the functions of some AS events in plants have been characterized, the functions of the vast majority are unknown (reviewed in Reddy et al. 2013). Analyzing evolutionary conserved AS events between species that are somewhat distantly related is a way of finding AS-creating isoforms with potentially important functions, as compared with AS events with no function or ones created by splicing noise (e.g., Boue et al. 2003; Ast, 2004; Sorek et al. 2004;
Wang and Brendel, 2006; Darracq and Adams 2013). Likewise, cases where AS is conserved after gene duplication, especially those in anciently duplicated genes, may suggest that the AS event is functional in both duplicates. Thus the approach of comparing conservation of AS events in gene families may be a way of identifying AS events that are more likely to be functional.

Future studies may incorporate more tissues or even multiple cell types within a tissue and assay duplicate pairs to pry out more qualitative and quantitative changes in AS events between duplicates. These patterns could highlight fine-scale regulatory and cis-element divergence between duplicates. Additionally, investigating causal changes in splicing-related cis-elements may provide a mechanistic example of diverged AS patterns between paralogs. Further functional characterization of alternative splicing in plants would undoubtedly lead to more interesting example cases like *CCA1/LHY*. Perhaps the most engaging prospect is the continued escalation of the known role of alternative splicing in all aspects of plant biology, and that duplication and alternative splicing are indeed processes with a complex relationship.

**Acknowledgments**

We thank C. Grisdale for help with RT-PCR and providing feedback on analysis and methods, G. Baute and A. Hammel for providing feedback on analysis and methods, A. Gorski and H. Jhand for assistance with scripting and computation, A. Darracq for assistance with plant growth and computation and S-L. Liu for generating the duplicate gene sets. This research was supported by a grant from the Natural Science and Engineering Research Council of Canada.
Literature cited


Stamatakis, A., 2006 RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with


Table 1. Conservation of alternative splicing events in duplicated gene pairs.

### Qualitative Conservation of Alternative Splicing Events

<table>
<thead>
<tr>
<th>Event Class</th>
<th>Conserved</th>
<th>Divergent</th>
<th>% Conservation</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha WG duplicates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IR</td>
<td>881</td>
<td>1500</td>
<td>37.0</td>
</tr>
<tr>
<td>ALTA</td>
<td>37</td>
<td>346</td>
<td>9.7</td>
</tr>
<tr>
<td>ALTD</td>
<td>12</td>
<td>220</td>
<td>5.2</td>
</tr>
<tr>
<td>ALTP</td>
<td>0</td>
<td>35</td>
<td>0.0</td>
</tr>
<tr>
<td>total</td>
<td>930</td>
<td>2101</td>
<td>30.7</td>
</tr>
<tr>
<td>tandem duplicates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IR</td>
<td>279</td>
<td>411</td>
<td>40.4</td>
</tr>
<tr>
<td>ALTA</td>
<td>22</td>
<td>119</td>
<td>15.6</td>
</tr>
<tr>
<td>ALTD</td>
<td>8</td>
<td>62</td>
<td>11.4</td>
</tr>
<tr>
<td>ALTP</td>
<td>2</td>
<td>17</td>
<td>10.5</td>
</tr>
<tr>
<td>total</td>
<td>311</td>
<td>609</td>
<td>33.8</td>
</tr>
</tbody>
</table>

### Quantitative Conservation of Alternative Splicing Events

<table>
<thead>
<tr>
<th>Event Class</th>
<th>Conserved</th>
<th>Divergent</th>
<th>% Conservation</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha WG duplicates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IR</td>
<td>279</td>
<td>602</td>
<td>31.7</td>
</tr>
<tr>
<td>ALTA</td>
<td>14</td>
<td>23</td>
<td>37.8</td>
</tr>
<tr>
<td>ALTD</td>
<td>2</td>
<td>10</td>
<td>16.7</td>
</tr>
<tr>
<td>ALTP</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>total</td>
<td>295</td>
<td>635</td>
<td>31.7</td>
</tr>
<tr>
<td>tandem duplicates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IR</td>
<td>116</td>
<td>163</td>
<td>41.6</td>
</tr>
<tr>
<td>ALTA</td>
<td>8</td>
<td>14</td>
<td>36.4</td>
</tr>
<tr>
<td>ALTD</td>
<td>4</td>
<td>4</td>
<td>50.0</td>
</tr>
<tr>
<td>ALTP</td>
<td>2</td>
<td>0</td>
<td>100.0</td>
</tr>
<tr>
<td>total</td>
<td>130</td>
<td>181</td>
<td>41.8</td>
</tr>
</tbody>
</table>

### Overall Conservation of Alternative Splicing Events

<table>
<thead>
<tr>
<th>Event Class</th>
<th>Conserved</th>
<th>Divergent</th>
<th>% Conservation</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha WG duplicates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IR</td>
<td>279</td>
<td>2102</td>
<td>11.7</td>
</tr>
<tr>
<td>ALTA</td>
<td>14</td>
<td>369</td>
<td>3.7</td>
</tr>
<tr>
<td>ALTD</td>
<td>2</td>
<td>230</td>
<td>0.9</td>
</tr>
<tr>
<td>ALTP</td>
<td>0</td>
<td>35</td>
<td>0.0</td>
</tr>
<tr>
<td>total</td>
<td>295</td>
<td>2736</td>
<td>9.7</td>
</tr>
<tr>
<td>tandem duplicates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IR</td>
<td>116</td>
<td>610</td>
<td>16.8</td>
</tr>
<tr>
<td>ALTA</td>
<td>8</td>
<td>133</td>
<td>5.7</td>
</tr>
</tbody>
</table>
Table 2. NMD status of Paralog Pairs

<table>
<thead>
<tr>
<th></th>
<th>symmetric NMD</th>
<th>asymmetric NMD</th>
<th>neither NMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha WG</td>
<td>35</td>
<td>207</td>
<td>1529</td>
</tr>
<tr>
<td>tandem</td>
<td>7</td>
<td>58</td>
<td>1114</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>ALTD</th>
<th></th>
<th></th>
<th></th>
<th>ALTP</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
<td>66</td>
<td>5.7</td>
<td></td>
<td>2</td>
<td>17</td>
<td>10.5</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>130</td>
<td>790</td>
<td>14.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Figure Legends**

**Figure 1.** Distribution of fold change differences for alpha WG duplicates and tandem duplicates. Each point represents the difference in AS expression between a pair of paralogs. Boxplot overlays indicate means (heavy lines), 25-75% quantile regions (boxes) and 95% quantile regions (whiskers). Tandem duplicate events that are not quantitatively conserved have a very wide distribution of fold differences, whereas alpha WG duplicate events that are not quantitatively conserved have a more narrow distribution.

**Figure 2.** Graphical representation of conservation status of alternative splicing events between paralogs.

**Figure 3.** RT-PCR results. Example genes include #3 (AT1G06640 junction 2-3), #11 (AT4G37540 junction 1-2), and #9 (AT4G18593 junction 1-2). Genes 3 and 11 show amplification of the constitutive form as well as a predicted intron retention event, while gene 9 shows amplification the constitutive form, an intron retention, and an alternative acceptor event.

**Figure 4.** Alternative splicing divergence between *CCA1* and *LHY* paralogs. *CCA1*β is produced at high frequency from the retention of intron 4 in *CCA1* under normal conditions, blocking *CCA1*α and *LHY* from triggering a cold response. *LHY* has an extremely low, perhaps vestigial amount of intron retention at the equivalent junction. Cold conditions shift the splicing of *CCA1* to cease producing the beta form, allowing *CCA1*α and *LHY* to trigger a cold response. Only the junctions relevant to this alternative splicing event are shown for clarity.
Figure 1. Distribution of fold change differences for alpha WG duplicates and tandem duplicates.

Each point represents the difference in AS expression between a pair of paralogs. Boxplot overlays indicate means (heavy lines), 25-75% quantile regions (boxes) and 95% quantile regions (whiskers). Tandem duplicate events that are not quantitatively conserved have a very wide distribution of fold differences, whereas alpha WG duplicate events that are not quantitatively conserved have a more narrow distribution.
Figure 2. Graphical representation of conservation status of alternative splicing events between paralogs. Red indicates divergent, blue indicates qualitative conservation, and green indicates quantitative conservation.
Figure 3. RT-PCR results. Example genes include #3 (AT1G06640 junction 2-3), #11 (AT4G37540 junction 1-2), and #9 (AT4G18593 junction 1-2). Genes 3 and 11 show amplification of the constitutive form as well as a predicted intron retention event, while gene 9 shows amplification the constitutive form, an intron retention, and an alternative acceptor event.
Figure 4. Alternative splicing divergence between CCA1 and LHY paralogs. CCA1β is produced at high frequency from the retention of intron 4 in CCA1 under normal conditions, blocking CCA1α and LHY from triggering a cold response. LHY has an extremely low, perhaps vestigial amount of intron retention at the equivalent junction. Cold conditions shift the splicing of CCA1 to cease producing the beta form, allowing CCA1α and LHY to trigger a cold response. Only the junctions relevant to this alternative splicing event are shown for clarity.