Transcriptome Analysis Indicates Considerable Divergence in Alternative Splicing Between Duplicated Genes in *Arabidopsis thaliana*

David C. Tack,* William R. Pitchers,[†] and Keith L. Adams*¹

*Department of Botany, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada, and [†]Department of Zoology, Program in Ecology, Evolutionary Biology, and Behavior, Michigan State University, East Lansing, Michigan 48824

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 downregulation 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 α -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 α -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 α -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.

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 Blanc and Wolfe 2004a; Cui *et al.* 2006; Kasahara 2007; 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 *et al.* 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 its 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 α -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.*

Copyright © 2014 by the Genetics Society of America

doi: 10.1534/genetics.114.169466

Manuscript received August 7, 2014; accepted for publication October 13, 2014; published Early Online October 16, 2014.

Supporting information is available online at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.169466/-/DC1>.

RNA-seq data have been deposited in the GEO database at NCBI as series GSE57579.

¹Corresponding author: Department of Botany, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada. E-mail: keitha@mail.ubc.ca

2014). Tandem duplicates contrast to α -WG duplicates in that they arise from small-scale duplications, often by unequal crossing over, and are 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 Graveley 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 alternative acceptor (ALTA) events 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 an alternative acceptor. Alternative splicing (AS) affects >50% of genes in *A. 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) (Kalyna *et al.* 2011; Drechsel *et al.* 2013). For example, AtGRP8 autoregulates itself via AS-induced NMD during elevated protein levels (Schöning *et al.* 2008). While some alternatively spliced transcripts may simply be by-products 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 *A. 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 (Filichkin *et al.* 2010; Marquez *et al.* 2012) and quantitative comparisons, it is possible to reevaluate 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 (ESEs) and exon splicing silencers (ESSs) and intron splicing enhancers (ISEs) and intron splicing silencers (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 *A. thaliana*, α -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

A. thaliana (Col-0) was grown at 20° with a 16-hr/8-hr 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 -80°. 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 μ g total RNA input for each sample and allowing fragmentation to proceed for 2 min.

Library sequencing and mapping/processing

Libraries were sequenced to obtain 100-bp paired-end reads on an Illumina HiSeq2000, multiplexing three libraries over 1.5 lanes. This yielded a total of 91,315,717, 90,873,817, and 168,071,194 paired-end reads per sample. The RNA-seq reads were mapped against the *A. thaliana* genome version Ensemble TAIR 10 downloaded as an igenome from <http://ccb.jhu.edu/software/tophat/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 ascertain which genes they originated from, as well as any splicing patterns, calling all splicing events based on the representative gene models. A total of 84,864,583, 79,167,068, and 150,297,611 paired-end reads were mapped. Scripts are available on request. The RNA-seq data have been deposited in NCBI's Gene Expression Omnibus (GEO) (Edgar *et al.* 2002) and are accessible through GEO Series accession

no. GSE57579 (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=cdkjkawipzmvdmpp&acc=GSE57579>).

Equivalent junction and event calling

Exons from the representative models for α -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 renumeration 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 three, with at least one read in support in each replicate as well as constitutive coverage of the junction. Overall 95% of all events had five or more total AS reads and 85% of events had eight or more total AS reads. Analyses were repeated requiring five and eight reads, respectively (supporting information, Table S2) with almost no effect on the results.

Analyses of leaf RNA-seq data

Expression data were used to exclude from the data set 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 one 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

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

where pr(alt) is the proportion of alternative splicing events and b_{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 ≥ 0.05 ,

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 (GO) enrichment analysis and visualization tool (GORilla) 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 GORilla web tool to search for enrichment in GO terms. To determine whether there was a difference in alternative splicing conservation rate between events located in genes' UTR regions vs. 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 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 were 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 α -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 g7b25 and at <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, *A. 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 α -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.

Table 1 Conservation of alternative splicing events in duplicated gene pairs

	Event class	Conserved	Divergent	% conservation
Qualitative conservation of alternative splicing events				
α -WG duplicates	IR	881	1500	37.0
	ALTA	37	346	9.7
	ALTD	12	220	5.2
	ALTP	0	35	0.0
	Total	930	2101	30.7
Tandems	IR	279	411	40.4
	ALTA	22	119	15.6
	ALTD	8	62	11.4
	ALTP	2	17	10.5
	total	311	609	33.8
Quantitative conservation of alternative splicing events				
α -WG duplicates	IR	279	602	31.7
	ALTA	14	23	37.8
	ALTD	2	10	16.7
	ALTP	0	0	0.0
	Total	295	635	31.7
Tandems	IR	116	163	41.6
	ALTA	8	14	36.4
	ALTD	4	4	50.0
	ALTP	2	0	100.0
	Total	130	181	41.8
Overall conservation of alternative splicing events				
α -WG duplicates	IR	279	2102	11.7
	ALTA	14	369	3.7
	ALTD	2	230	0.9
	ALTP	0	35	0.0
	Total	295	2736	9.7
Tandems	IR	116	610	16.8
	ALTA	8	133	5.7
	ALTD	4	66	5.7
	ALTP	2	17	10.5
	Total	130	790	14.1

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 complementary DNA (cDNA) pools were made via the Invitrogen (Carlsbad, CA) 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, and 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 *A. thaliana* were used to investigate the splicing patterns

of two classes of paralogs, α -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 *Materials and 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 43,236 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% of events are conserved between α -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 α -WG duplicates and tandem duplicates, and this difference had statistical support (α -WG duplicates $P < 0.001$, tandem duplicates $P < 0.001$, chi square). Conserved and divergent AS events are listed in Table S3.

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 that produced a P -value < 0.05 and thus had statistical support for a difference between frequencies of alternative vs. 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 α -WG duplicates and 41% in tandems (Table 1), with tandems having a significantly higher rate of quantitative conservation ($P = 0.002$, chi square). However, the median effect size difference in the quantitative comparison is 4.1-fold for α -WG duplicates and 10.2 for tandems ($P = 0.001$, chi square), again with a significant difference. More of the events in tandem duplicates are quantitatively conserved than events in α -WG duplicates, but the events in tandems that are quantitatively divergent show a wider range of disparity on average than α -WG events that are quantitatively divergent. When the qualitative and quantitative results are combined, only a small fraction of the events are fully

Figure 1 Distribution of fold change differences for α -WG duplicates and tandem duplicates. Each point represents the difference in AS expression between a pair of paralogs. Boxplot overlays indicate means (thick 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 α -WG duplicate events that are not quantitatively conserved have a more narrow distribution.

conserved (Table 1; Figure 2). The α -WG duplicates have 9% conservation on both levels whereas the tandem duplicates again have a significantly higher rate of conservation ($P = 0.0003$, chi square) at 14%. Conserved and divergent AS events are listed in Table S3.

No significant differences in the rate of qualitative or quantitative conservation were found in events located in untranslated regions of transcripts (5'-UTR and 3'-UTR) vs. the events found in coding sequences of transcripts (CDS). Additionally there was no differential enrichment of GO categories or terms in genes with conserved AS events vs. 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, 20 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; Table S1) for the constitutive and alternative form(s) while the remaining 3 either failed to amplify or had unspecific amplification. Events from these 3 junctions were validated by EST support in plantGDB (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 α -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 that 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 three replicates (~2% of reads contain the IR), while the corresponding junction in *CCA1* was found to have 2186 intron retention reads,

1811 constitutive reads, and 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 Seo *et al.* (2012), who demonstrated that retention of the intron in *CCA1* makes *CCA1* β , which binds to *CCA1* α or *LHY*, creating protein dimers with reduced DNA binding affinity and preventing formation of normal *CCA1* α homodimers, *LHY* homodimers, and *CCA1* α -*LHY* heterodimers. The alternative splicing of *CCA1* into *CCA1* β is suppressed in lower temperatures, allowing *CCA1* α 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 *O. 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 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 *Materials and Methods*) to ask whether 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 nonsignificant difference ($P = 0.1496$, chi square). However, this does not preclude that it may still be a common mechanism of divergent evolution between duplicated genes. Comparing the NMD status of α -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

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.

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 complements 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 α -WG duplicate pairs, using RNA-seq data. We found that alternative splicing patterns are infrequently conserved between paralogs in *A. thaliana* leaves. The majority of alternative splicing events are conserved neither 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 1). 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

Figure 3 RT-PCR results. Example genes include no. 3 (AT1G06640 junction 2–3), no. 11 (AT4G37540 junction 1–2), and no. 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 of the constitutive form, an intron retention, and an alternative acceptor event.

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 α -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 in α -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 α -WG duplicates have less difference on average. Nonuniformity 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 α -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 although 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; 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 expressing it either 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 *cis* elements are the primary mechanisms for

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 β form, allowing *CCA1*α and *LHY* to trigger a cold response. Only the junctions relevant to this alternative splicing event are shown for clarity.

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 *trans* environment. Changes in *cis* elements were implicated to account for most species-specific alternative splicing patterns in vertebrates (Barbosa-Morais *et al.* 2012), and in another study ~80% of non-exon-skipping events were attributable to divergent *cis* architecture in *Drosophila* (McManus *et al.* 2014). Collectively most changes in alternative splicing are thus directed by changes in *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 two-thirds 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 2004b; Casneuf *et al.* 2006; Liu *et al.* 2011), may have an underpinning in alternative splicing in some cases. Our finding that 85% of α-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 differen-

Table 2 NMD status of paralog pairs

	Symmetric NMD	Asymmetric NMD	Neither NMD
α-WG	35	207	1529
Tandem	7	58	1114

tially 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* maintains the event only 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 the 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 nonspecific *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 highlight 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 that invokes normal homeostasis. *LHY* and *CCA1* have several other known roles, including regulating SVP (short vegetative phase) (Fujiwara *et al.* 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, 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 ancient 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 gene 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

- Ast, G., 2004 How did alternative splicing evolve? *Nat. Rev. Genet.* 5: 773–782.
- Barbosa-Morais, N. L., M. Irimia, Q. Pan, H. Y. Xiong, S. Gueroussov *et al.*, 2012 The evolutionary landscape of alternative splicing in vertebrate species. *Science* 338: 1587–1593.
- Blanc, G., and K. H. Wolfe, 2004a Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16: 1667–1678.
- Blanc, G., and K. H. Wolfe, 2004b Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* 16: 1679–1691.
- Blanc, G., K. Hokamp, and K. H. Wolfe, 2003 A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res.* 13: 137–144.
- Boue, S., I. Letunic, and P. Bork, 2003 Alternative splicing and evolution. *BioEssays* 25: 1031–1034.
- Bowers, J. E., B. A. Chapman, J. Rong, and A. H. Paterson, 2003 Unraveling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422: 433–438.
- Casneuf, T., S. De Bodt, J. Raes, S. Maere, and Y. Van de Peer, 2006 Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*. *Genome Biol.* 7: R13.
- Chen, M., and J. L. Manley, 2009 Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat. Rev. Mol. Cell Biol.* 10: 741–754.
- Cui, L., P. K. Wall, J. H. Leebens-Mack, B. G. Lindsay, D. E. Soltis *et al.*, 2006 Widespread genome duplications throughout the history of flowering plants. *Genome Res.* 16: 738–749.
- Conant, G. C., and K. H. Wolfe, 2008 Turning a hobby into a job; how duplicated genes find new functions. *Nat. Rev. Genet.* 9: 938–950.
- Cusack, B. P., and K. H. Wolfe, 2007 When gene marriages don't work out: divorce by subfunctionalization. *Trends Genet.* 23: 270–272.
- Darracq, A., and K. L. Adams, 2013 Features of evolutionarily conserved alternative splicing events between Brassica and *Arabidopsis*. *New Phytol.* 199: 252–263.
- Dong, M. A., E. M. Farré, and M. F. Thomashow, 2011 Circadian clock-associated 1 and late elongated hypocotyl regulate expression of the C-repeat binding factor (CBF) pathway in *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* 108: 7241–7246.
- Drechsel, G., A. Kahles, A. K. Kesariwani, E. Stauffer, J. Behr *et al.*, 2013 Nonsense-mediated decay of alternative precursor mRNA splicing variants is a major determinant of the *Arabidopsis* steady state transcriptome. *Plant Cell* 10: 3726–3742.
- Duvick, J., A. Fu, U. Muppirala, M. Sabharwal, W. D. Wikerson *et al.*, 2008 PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Res.* 36: 959–965.
- Eden, E., D. Lipson, S. Yogeve, and Z. Yakhini, 2007 Discovering motifs in ranked lists of DNA sequences. *PLoS Comput. Biol.* 3: e39.
- Eden, E., R. Navon, I. Stienfeld, D. Lipson, and Z. Yakhini, 2009 GORilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10: 48.
- Edgar, R., M. Domrachev, and A. E. Lach, 2002 Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30: 207–210.
- Filichkin, S. A., H. D. Priest, S. A. Givan, R. Shen, D. W. Bryant *et al.*, 2010 Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res.* 20: 45–58.
- Filichkin, S. A., and T. C. Mockler, 2012 Unproductive alternative splicing and nonsense mRNAs: a widespread phenomenon among plant circadian clock genes. *Biol. Direct* 7: 20.
- Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan *et al.*, 1999 Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151: 1531–1545.
- Fujiwara, S., A. Oda, R. Yoshida, K. Niinuma, K. Miyata *et al.*, 2008 Circadian clock proteins LHY and CCA1 regulate SVP protein accumulation to control flowering in *Arabidopsis*. *Plant Cell* 20: 2960–2971.
- Han, M. V., J. P. Demuth, C. L. Mc, C. Grath, C. Casola *et al.*, 2009 Adaptive evolution of young gene duplicates in mammals. *Genome Res.* 5: 859–867.
- Huelga, S. C., A. Q. Vu, J. D. Arnold, T. Y. Liang, P. P. Liu *et al.*, 2012 Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell Rep.* 1: 167–178.
- Jiao, Y., N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr *et al.*, 2011 Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97–100.
- Kalyna, M., C. G. Simpson, N. H. Syed, D. Lewandowska, Y. Marquez *et al.*, 2011 Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in *Arabidopsis*. *Nucleic Acids Res.* 40: 2454–2469.
- Kasahara, M., 2007 The 2R hypothesis: an update. *Curr. Opin. Immunol.* 19: 547–552.

- Lister, J. A., J. Close, and D. W. Rabile, 2001 Duplicate *mitf* genes in zebrafish: complementary expression and conservation of melanogenic potential. *Dev. Biol.* 237: 333–344.
- Liu, S., G. J. Baute, and K. L. Adams, 2011 Organ and cell type-specific complementary expression patterns and regulatory neofunctionalization between duplicated genes in *Arabidopsis thaliana*. *Genome Biol. Evol.* 3: 1419–1436.
- Lu, T., G. Lu, D. Fan, C. Zhu, W. Li *et al.*, 2010 Function annotation of rice transcriptome at single nucleotide resolution by RNA-seq. *Genome Res.* 20: 1238–1249.
- Marquez, Y., J. W. Brown, C. Simpson, A. Barta, and M. Kalyna, 2012 Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*. *Genome Res.* 22: 1184–1195.
- Marshall, A. N., M. C. Montealegre, C. Jimenez-Lopez, M. C. Lorenz, and A. van Hoof, 2013 Alternative splicing and subfunctionalization generates functional diversity in fungal proteomes. *PLoS Genet.* 9: e1003376.
- McManus, J. C., J. D. Coolon, J. Eipper-Mains, P. J. Wittkopp, and B. R. Graveley, 2014 Evolution of splicing regulatory networks in *Drosophila*. *Genome Res.* 24: 786–796.
- Nilsen, T. W., and B. R. Graveley, 2010 Expansion of the eukaryotic proteome by alternative splicing. *Nature* 463: 457–463.
- Proost, S., M. Van Bel, L. Sterck, K. Billiau, T. Van Parys *et al.*, 2009 PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell* 21: 3718–3731.
- Reddy, A. S. N., Y. Marquez, M. Kalyna, and A. Barta, 2013 Complexity of the alternative splicing landscape in plants. *Plant Cell* 25: 3657–3683.
- Renny-Byfield, S., J. P. Gallagher, C. E. Grover, E. Szadkowski, J. T. Page *et al.*, 2014 Ancient gene duplicates in *Gossypium* (Cotton) exhibit near-complete expression divergence. *Genome Biol. Evol.* 6: 559–571.
- Sanchez, S. E., E. Petrillo, E. J. Beckwith, X. Zhang, M. L. Rugnone *et al.*, 2010 A methyl transferase links the circadian clock to the regulation of alternative splicing. *Nature* 468: 112–116.
- Schnable, J. C., N. M. Springer, and M. Freeling, 2011 Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. USA* 108: 4069–4074.
- Schöning, J. C., C. Streitner, I. M. Meyer, Y. Gao, and D. Staiger, 2008 Reciprocal regulation of glycine-rich RNA-binding proteins via an interlocked feedback loop coupling alternative splicing to nonsense-mediated decay in *Arabidopsis*. *Nucleic Acids Res.* 36: 6977–6987.
- Sémon, M., and K. H. Wolfe, 2007 Consequences of genome duplication. *Curr. Opin. Genet. Dev.* 17: 505–512.
- Seo, P. J., M. Park, M. Lim, S. Kin, M. Lee *et al.*, 2012 A self-regulatory circuit of CIRCADIAN CLOCK-ASSOCIATED1 underlies the circadian clock regulation of temperature responses in *Arabidopsis*. *Plant Cell* 24: 2427–2442.
- Shen, Y., Z. Zhou, Z. Wang, W. Li, C. Fang *et al.*, 2014 Global dissection of alternative splicing in paleopolyploid soybean. *Plant Cell* 26: 996–1008.
- Sorek, R., R. Shamir, and G. Ast, 2004 How prevalent is functional alternative splicing in the human genome? *Trends Genet.* 20: 68–71.
- Staiger, D., and J. W. S. Brown, 2013 Alternative splicing at the intersection of biological timing, development, and stress responses. *Plant Cell* 25: 3640–3656.
- Stamatakis, A., 2006 RAXML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22: 2688–2690.
- Wang, B. B., and V. Brendel, 2006 Genomewide comparative analysis of alternative splicing in plants. *Proc. Natl. Acad. Sci. USA* 103: 7175–7180.
- Wu, T. D., and C. K. Watanabe, 2005 Gmap: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21: 1859–1875.
- Zhang, P. G., S. Z. Huang, A. Pin, and K. L. Adams, 2010 Extensive divergence in alternative splicing patterns after gene and genome duplication during the evolutionary history of *Arabidopsis*. *Mol. Biol. Evol.* 27: 1686–1697.

Communicating editor: J. A. Birchler

GENETICS

Supporting Information

<http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.169466/-/DC1>

Transcriptome Analysis Indicates Considerable Divergence in Alternative Splicing Between Duplicated Genes in *Arabidopsis thaliana*

David C. Tack, William R. Pitchers and Keith L. Adams

Table S1. Primer sets for RT-PCR.

Gene	Junction	Event Class	Forward Primer	Reverse Primer	Confirmed?	product 1	product 2
AT1G01620	1,2	IR	GACAGCTTACGAGCCAAGAA	ATTGACAGTGATGGGAGTGAAG	1	174	334
AT3G20300	1,2	IR	CTCCTGAAGCATACCACCATATC	GTTTCGCTGCTCTTTCGTTTC	X	330	780
AT1G06640	2,3	IR	CTTACCGTCCTTCTACCAGATAAC	GGTCCGTACACTCTAGGATTTG	1	250	374
AT3G21360	2,3	IR	TGGACGTGGTTGGAAATCTAC	CGGTTTCTCGACTCGTCATATT	1	170	244
AT5G42850	2,3	IR	ACATTGAGCACGAAATCGTAAAG	CGCATCCTTGGAGAGTTGAT	X	300	374
AT1G26945	1,2	IR	CACCTTCTTCTCCACTCTCATTC	GGTCATCAACCTCTCTGTGTAAG	1	300	933
AT2G31360	4,5	IR	CCGAACAATGTACCACGAAATG	GTAGGAGCAGCATTGGAAGT	1	227	402
AT2G42810	5,7	SKIP	GGTAACCTCCTCTCCCTCAATTC	AGCAATCTCTGTGCCAGTATC	X	350	230
AT4G18593	1,2	AltA/IR	AATCAGTGTAGCAGACACCATC	CTTCACCCTTTCCTGGTTCAT	1	200	704
AT4G24800	1,2	AltA/IR	CTTCAGACTCGGGTTCTTCTTT	CATGAGACCTTCTGTGCTTCA	1	329	629
AT4G37540	1,2	IR	AGTTCCTGGTCCACAACATAC	CACCGTCTTCGTCGCTAAA	1	205	305
AT1G72170	1,2	IR	CCACCGGAATACGATGTGAA	CAATACCAATTCCAGCACCTAAAG	1	182	382
AT1G63880	1,2	IR	TCAATTCCCACCATACCATCAA	CAATGAACTTGTGCACGTAAGA	1	207	457
AT3G54810	2,3	IR	CATAGGTGACTCCGACTCTTTC	GGAGGTCTTTCACCTTAACA	1	167	467
AT5G04550	1,2	IR	CGCAACCTCAAACGCTAATAC	CCTCTCTCTCTCTCTCTCATC	1	320	620
AT1G66100	1,2	IR	CACCGCAAACAGAGGATACA	AGTGAGTCGACATGTCCTAGA	1	100	202
AT5G24165	1,2	IR	TCGTATCAAATTTCTCAGAGACA	GAGATGCTTTCCCTCCATCAG	1	150	271
AT1G52347	1,2	IR	TTGACGCCGGTGATTGTT	GCTTTCCTCGGTCTTGATAC	1	334	834
AT1G23030	1,2	IR	GGCCTCTCTACTCGACCTAAT	CTTTGTTTCTCCTCCTCCTCAC	1	553	743
AT5G52530	1,2	IR	CGAGCGGTGGATAGTGAAAT	CGCTTTGATTCTCCTGCTACTA	1	340	990

Tables S2-S3

Available for download as Excel files at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.169466/-/DC1>

Table S2 Conservation of alternative splicing events in duplicated gene pairs, using different numbers of minimum required reads.

Table S3