Analyses of *cis*-regulatory elements in duplicated genes of *Arabidopsis thaliana*

By Julian Booth

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF BACHELOR OF SCIENCE

in

THE FACULTY OF SCIENCE

(Combined Honours Chemical Biology)

THE UNIVERSITY OF BRITISH COLUMBIA

August 2017

We accept this thesis as conforming to the required standard

_____

_____

_____

_____

_____

Abstract

Much of the genetic diversity found in plants is due to gene duplication. Duplicated genes often diverge in their function, expression patterns and regulation. Regulatory changes in duplicate genes can be studied by comparing the various transcription factors that target them and their cis-regulatory elements. In this study, cis-regulatory elements in duplicated genes of *Arabidopsis thaliana* were examined using a reliable and detailed transcription factor binding site (TFBS) data set that was constructed using a new high-throughput technique known as DNA affinity purification sequencing (DAP-seq). The fates of duplicates were found to differ depending on the type of duplication. Tandem duplicates were found to possess fewer TFBSs than their whole genome duplicate counterparts. There is also less conservation of TFBSs between tandem duplicate paralogs. Duplicate transcription factors are regulated by more TFBSs than their non-transcription factor counterparts. Compared to whole genome duplicate transcription factors, tandem duplicate transcription factors are highly diverged in their regulation but not in their function. This study provides evidence that tandem duplicates are more divergent in their regulation than whole genome duplicates.

Table of Contents

Acknowledgements

I would like to thank Professor Keith Adams for supervising my thesis project and for giving me a chance when I really needed it.

I would also like to thank the rest of the lab, especially Yichun Qiu, for helping me out along the way.

Finally, a huge thank you to the countless professors, students and administration who helped me during my undergraduate career.

Introduction

Gene duplication is an important evolutionary process – it provides the raw genetic material for new gene formation and modification - with major duplication events shaping the genomes of species over time (Kellis et al., 2004; Jiao et al., 2011). Whole Genome Duplication (WGD) has occurred many times during the past 200 million years in plants (Renny-Byfield and Wendel, 2014; Lee et al., 2012). These duplication events provide a massive influx of redundant genetic material for evolution to act on. It is thought that the success of seed plants and angiosperms (flowering plants) can be attributed to innovations following local duplication and WGDs (De Bodt et al., 2005; Soltis et al., 2008; Fawcett et al., 2009). The diversification of seed plants was prolific shortly after a WGD 319 million years ago, and a similar diversification of angiosperms occurred after a WGD 192 million years ago (Jiao et al, 2011).

Tandem duplications (TDs) are local duplications resulting from unequal crossing-over that lead to two or more sequential or closely spaced paralogs (Zhang, 2003). These duplicates also make up a significant portion of flowering plant genomes. For example, tandem duplicates account for approximately 9% of the Arabidopsis genome (Panchy et al, 2016). These tandem duplicates may also be imperfectly duplicated and result in partially missing coding and/or regulatory regions (Panchy et al, 2016).

There are several possible fates for duplicated genes. Often one paralog undergoes relaxed purifying selection and is lost in a process known as pseudogenization (Panchy et al, 2016). This occurs most frequently through premature stop codons and frame shifts (Vanin, 1985). A few pseudogenes in Arabidopsis are expressed, with multiple theories explaining their presence (Zou et al., 2009b). One hypothesis is that pseudogenes may be in the process of complete decay, having only recently become pseudogenized. This is

supported by evidence showing that expressed pseudogenes tend to be derived from more recent duplication events (Zou et al. 2009a).

Both genes in a duplicate pair may be retained if one or both paralogs gain a modified or novel function (Ohno et al., 1968), or if the two genes partition the functions of the original gene in processes known respectively as neofunctionalization and subfunctionalization (Lynch et al., 2000b; Lynch et al., 2000a; Force et al., 1999). The expression pattern of duplicated genes can diverge through changes in gene regulation, with the addition of a novel expression pattern or a partitioning of the ancestral expression pattern being referred to as regulatory neofunctionalization and subfunctionalization respectively (Force et al., 1999). These are thought to be key mechanisms in the retention of duplicated genes.

Duplicates can also be conserved as mostly unchanged pairs (Force et al., 1999; Gu et al., 2003). Conservation may be dictated by gene product dosage according to the gene balance hypothesis (Birchler and Veitia, 2007). Gene dosage imbalances can sometimes have lethal consequences for organisms (Veitia and Potier, 2015). A locally duplicated gene pair may lead to a dosage imbalance since the components they interact with remain single copy. This leads to selective pressure against retention of both duplicated genes. However, in a WGD event all gene dosages end up being doubled and retention becomes favorable, as loss of one copy would lead to imbalance (Panchy et al, 2016). There is a greater correlation between the expression patterns of WGDs than between tandem duplicates (Casneuf et al., 2006). This may be a consequence of regulatory evolution in response to dosage based selective pressure. Clearly, expression changes are a key factor in determining the fates of duplicate genes.

Studying changes in cis-regulatory elements helps to understand changes in expression. For example, differences in putative transcription factor binding sites (TFBS) between

duplicate genes partly explain the differences in stress response seen by the genes in Arabidopsis (Zou et al., 2009b). In the past we have been held back from studying cis-regulatory evolution in flowering plants at the genome-wide scale by the lack of reliable, high-throughput methods. This has recently changed.

Arabidopsis (*Arabidopsis thaliana)* is an angiosperm with a plethora of duplicated genes of different ages, including two relatively recent WGD events (Wang et al., 2013). These WGD events are known respectively as the alpha and beta WGDs. The alpha duplication is more recent – predating the divergence of *Arabidopsis* from *Brassica* 15-20 million years and spanning all of the Brassicaceae family, but no closely related families (Edger et al., 2015). The beta duplication is older and it includes the Brassicaceae and a few closely related families in the Brassicales order. A third gamma duplication is even older but was not examined in this study. Along with these duplications of different ages, Arabidopsis is also a model organism with a well-annotated genome - making it a great species for understanding the role of duplication in genome evolution in plants.

In a previous study of cis-regulatory elements in *Arabidopsis thaliana* done in our lab (Arsovski et al., 2015), the researchers analyzed the fates of cis-regulatory elements in duplicated genes of Arabidopsis using a dataset of cis-element footprints found through the DNase hypersensitivity (DHS) method (Sullivan et al., 2014). The method works by treating genomic DNA with the enzyme DNase I that finds and cleaves exposed euchromatin (hypersensitive) sites, which usually correspond to actively expressed genes. The DNA that is bound by regulatory proteins is protected from DNase I cleavage. These are sequenced and mapped to the genome and are known as footprints. The dataset used was from 7 day old seedlings and therefore only captures the hypersensitive sites and footprints for that specific context, thus missing genes and transcription factors (TFs)

expressed in different tissues and at different times.

Other techniques have been designed for the high-throughput identification of transcription factor binding sites (TFBSs) with varying levels of convenience, comprehensiveness and biological relevance. ChIP-seq is a reliable and comprehensive method but difficult to perform in a high-throughput assay, whereas DHS provides an easier method, but offers little information on the identities of the targeting TFs (Kheradpour and Kellis, 2014; Stamatoyannopoulos et al., 2012). In vitro TFBS mapping approaches like SELEX (Jolma et al., 2010) and PBM (Berger and Bulyk, 2009) are convenient methods but use synthetic DNA, which lacks properties known to affect TF binding such as primary sequence context and epigenetic modifications like 5-methylcytosine.

In my study, I performed an analysis of a dataset produced through a novel procedure known as DAP-seq. DNA affinity purification sequencing (DAP-seq) is a high-throughput method that uses in-vitro-expressed tagged TFs to bind naked gDNA fragments. The method provides genomic-context preserved TFBSs and sequence motifs for hundreds of Arabidopsis TFs (O'Malley et al., 2016). DAP-seq peaks also appeared to be reliable in a comparison with ChIP-seq, overlapping with 69% to 97% of ChIP-seq peaks that ranked in the top 25% by motif score and containing a larger percentage of high-scoring motifs overall (O'Malley et al., 2016).

I was able to expand on the work of Arsovski et al. (2015) by analyzing the DAP-seq dataset. I found that TDs have fewer regulatory elements that are less conserved and are under more evolutionary pressure to diverge than their WGD counterparts. I also looked at duplicate TFs and found that, compared to WGD TFs, TD TFs are highly diverged in their regulation but not in their function.

Materials and Methods

DAP-Seq Data

DAP-seq transcription factor and transcription factor binding site data comes from O'Malley et al. (2016). Motif, peaks (FRiP ≥ 5%) and target genes (FRiP ≥ 5%) were downloaded from http://neomorph.salk.edu/dap_web/pages/browse_table_aj.php.

Duplicate Genes

TD pairs, and alpha and beta WGD pairs were obtained from Wang et al. (2013). TD pairs occasionally included overlapping sequences of pairs when clusters of three or more paralogs or more appeared i.e., a sequence of three duplicates A B C would be counted as two pairs A-B and B-C. Background genes, pairs and transcription factors were taken randomly from the set of protein coding genes from *Arabidopsis thaliana* minus the three duplicate sets. Protein coding genes were obtained from the TAIR10 genome release from the Arabidopsis Information Resource (TAIR) (https://www.arabidopsis.org/download/index.jsp).

Analysis of Transcription Factor Binding Sites Between Duplicates

I developed Python scripts to compare, count and filter TFBSs in duplicate pairs of interest. Many TFs analyzed in the DAP-seq dataset came from the same families and had very similar target sequences. Several members of the same TF family can bind the same high affinity TFBSs and differ only at the lower affinity sites (Siggers and Gordân,

2013). Therefore, to avoid counting one site many times I only counted the TFBSs of 49 TFs identified as spanning the space of motif diversity listed in table S1C of the O'Malley et al. (2016) supplemental information.

To calculate conservation of TFBSs per duplicate pair I calculated the number of shared TFBSs divided by the total number of TFBSs. A duplicate pair shares a TFBS when both genes are targeted by the same TF. I divided the number of shared TFBSs by the total number of TFBSs to control for the large variance in total TFBS number seen in Fig. 1. For example, if Gene 1 is targeted by TF A and TF B and Gene 2 is targeted by TF A and TF C and TF D then the number of TFBSs shared is 2 and the total number of TFBSs is 5 – therefore, there is 40% conservation.

Calculating Ks Values

Pairwise synonymous substitution rates (Ks) for all duplicate pairs were estimated using the KaKs Calculator (Zhang et al., 2006), with the MYN (Modified version of the Yang-Nielsen method) approximation method. Protein sequences were obtained from TAIR10 and then aligned using Clustal (Sievers et al., 2011) and were translated into codon alignments using CDS, also obtained from TAIR10 using PAL2NAL (Suyama et al., 2006).

Regulatory Analysis of Duplicate Pair Transcription Factors

I developed Python scripts to compare, count and filter the target genes of the TFs examined in the DAP-seq dataset. Target conservation was calculated in a similar manner to TFBS conservation.

Analysis of Position Weight Matrices

Position weight matrices (PWM) of duplicate pair TFs, in which both paralogs were examined in the DAP-Seq dataset, were compared using TomTom from the MEME Suite (version 4.12.0) (http://meme-suite.org/doc/overview.html) (Gupta et al., 2007). P-values were calculated by uploading motifs of interest against the DAP motifs target database. P-values represent the likelihood of finding a PWM of equal or greater similarity by chance in a similarly sized database filled with random PWMs.

Figures and Statistical Analysis

All figures were made using the python library plotly. Plotly was also used for statistical analysis. All statistical tests are two-tailed, independent t-tests unless otherwise specified. For all boxplots the dotted line indicates mean and the solid line indicates median. All outlying data points are either below the first quartile or above the third quartile by 1.5 times the interquartile range.

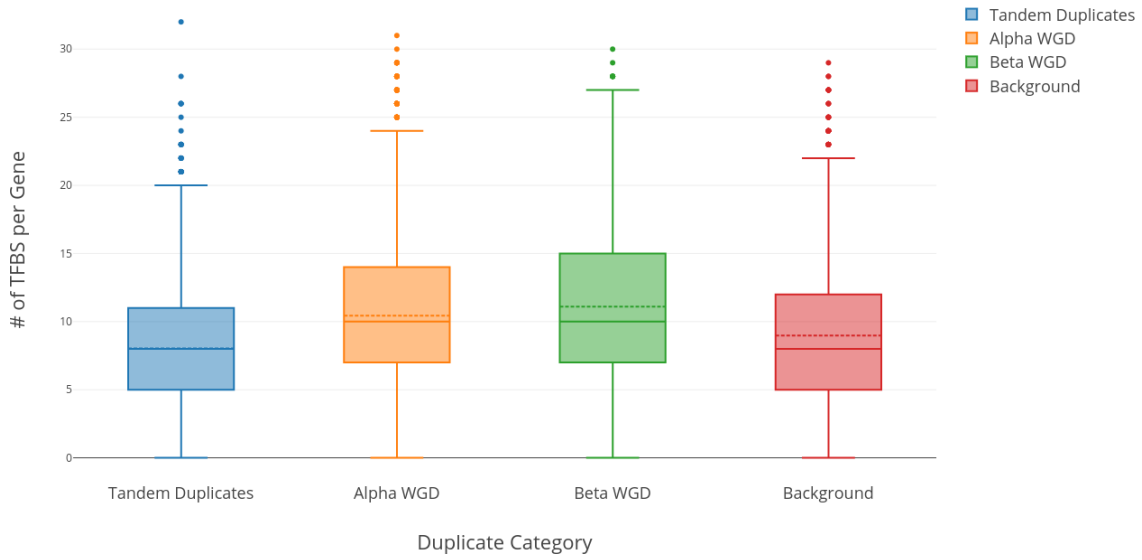
Results

Total Regulatory Elements Differ Between Types of Duplicates

I started by finding the average number of TFBSs per gene in four categories: tandem duplicates, alpha whole genome duplicates, beta whole genome duplicates and background genes (singletons). As in Arsovski et al. (2015), TDs were found, on average, to possess fewer regulatory elements than WGDs. Using the DAP-seq dataset, I found TDs had 8.02 TFBSs versus 10.43 TFBSs in alpha WGDs ($P < 10^{-6}$) and 11.10 TFBSs in beta WGDs ($P < 10^{-6}$). TDs had slightly fewer TFBSs than background (singleton) genes (8.97, $P < 10^{-6}$) (Fig. 1A). Beta WGDs also had significantly more TFBS than the younger WGDs (P = 1E-6).

To check if the lower totals for tandem duplications were due to incomplete duplications, I once again found the average number of TFBSs per gene in each category. However, this time I only looked at the gene in each pair that had more TFBSs. This was done to assess whether the average number of TFBSs per TD gene was lower simply because some TD pairs had one gene with an incompletely duplicated regulatory region. TDs were still found to have an average number of TFBSs (9.69) significantly lower than that of alpha WGDs (12.54, $P < 10^{-6}$) and beta WGDs (13.49, $P < 10^{-6}$) (Fig. 1B).

(A)


Tandem Duplicates
Alpha WGD
Beta WGD
Background
# of TFBS per Gene
0
5
10
15
20
25
30
Tandem Duplicates
Alpha WGD
Beta WGD
Background
Duplicate Category

(B)

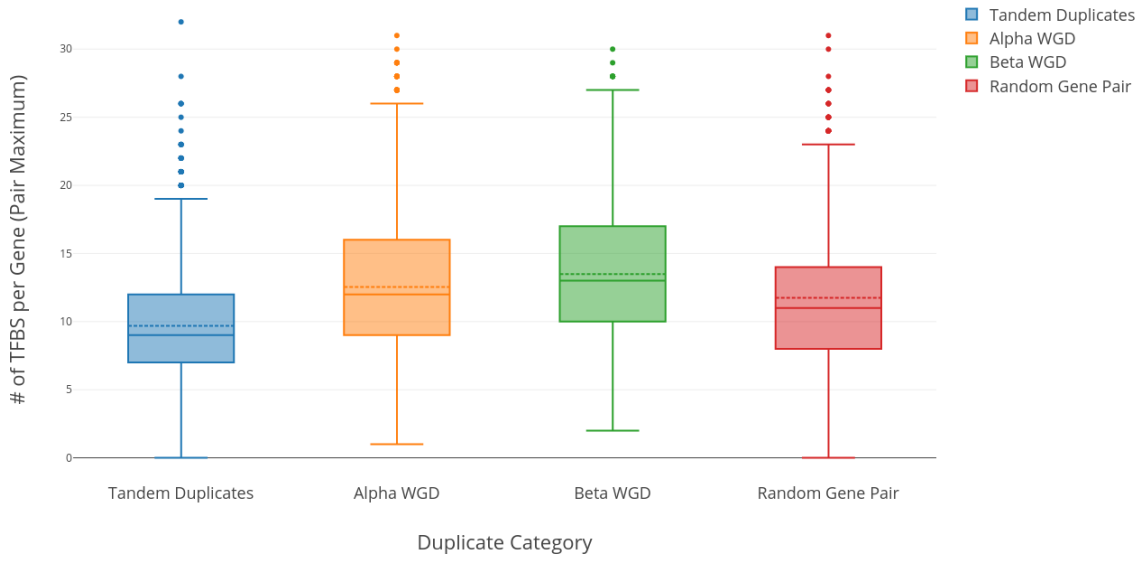
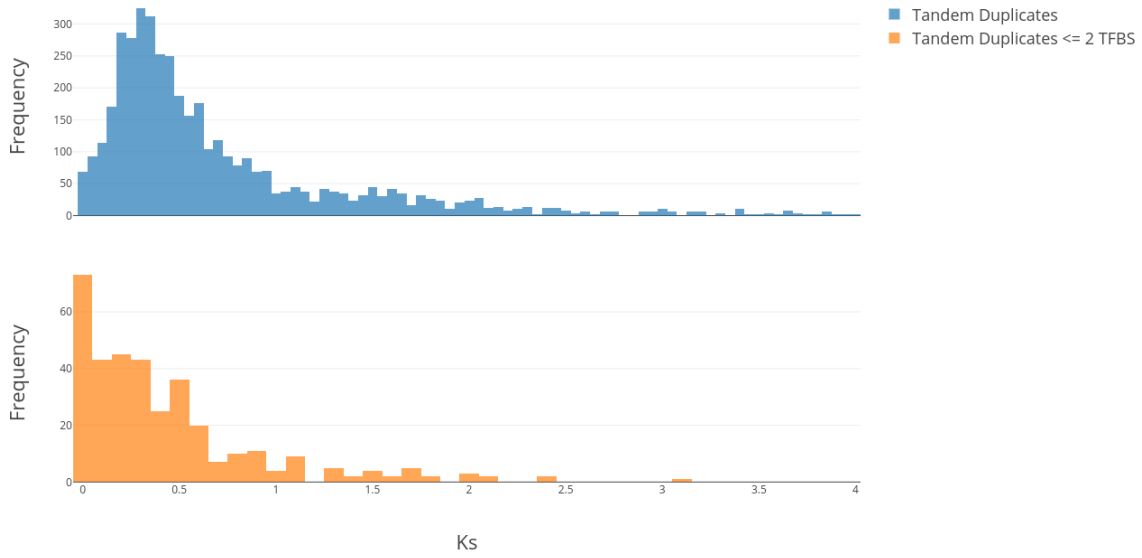

Tandem Duplicates
Alpha WGD
Beta WGD
Random Gene Pair
# of TFBS per Gene (Pair Maximum)
0
5
10
15
20
25
30
Tandem Duplicates
Alpha WGD
Beta WGD
Random Gene Pair
Duplicate Category

Figure 1. Transcription Factor Binding Sites (TFBS) per gene in each category. (A) Box plots show the number of TFBS assigned to genes within duplicate and random gene categories. (B) Box plots show the number of TFBS assigned to the gene with the greater number of TFBS in each pair within duplicate categories and randomly selected pairs.

Sparsely Regulated Duplicates Are Recent Duplicates

Duplicate gene sets may include pseudogenes (Arsovski et al., 2015). To detect likely pseudogenes, I analyzed the duplicate data sets for genes with no regulatory elements. I found 3.4% of my TDs to possess no TFBS versus 0.27 % and 0.53 % in alpha and beta WGDs respectively. Zou et al. (2009a) showed that expressed pseudogenes tend to be derived from more recent duplication events. I wanted to test whether TDs with few regulatory elements are younger than the average TD. I compared the age distribution - using Ks as a proxy for age of duplication - of the total population of TDs with the population of TDs with two or fewer TFBSs - two was chosen as a cutoff for sample size reasons. I found that the average TD with few or no TFBSs was significantly younger (Ks = 0.43) than the unfiltered average for TDs (Ks = 0.70) ($P < 10^{-6}$) (Fig. 2A). I ran the same experiment using the alpha WGDs (Fig. 2B). I found no difference in Ks (P=0.61) between the two populations.

(A)



(B)

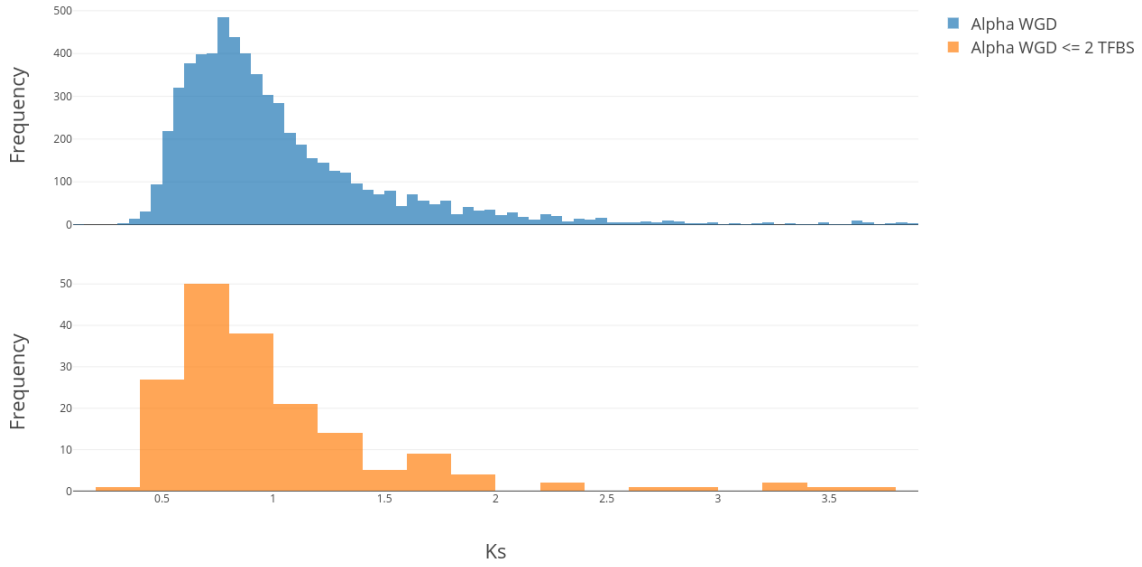


Figure 2. Tandem duplicates with fewer TFBSs are younger. (A) Histogram showing the difference in Ks frequency between TDs (blue) and TDs with 2 or fewer TFBSs (orange).

(B) Histogram showing the difference in Ks frequency between alpha WGDs (blue) and alpha WGDs with 2 or fewer TFBSs (orange).

Regulatory Elements Are Conserved Between Duplicate Pairs

Arsovski et al. (2015) compared how conserved the regulatory elements are between pairs of TDs and pairs of WGDs. Previous research showed that there is a greater correlation between the expression patterns of WGDs than between TDs (Casneuf et al., 2006). One would expect that to be mirrored in the regulatory elements governing those expression patterns. Arsovski et al. (2015) found somewhat conflicting results. They found TDs had a smaller net difference in regulatory element number between copies (4.37) than WGDs (6.32). However, they also found that WGDs had a higher percentage of pairs (36.5%) with some regulatory elements in common than TDs (16%). With knowledge of the actual TFs targeting the TFBSs in the DAP-Seq data set, I was able to make a more detailed analysis of regulatory conservation. TDs shared on average 32.9% of their total TFBS, significantly less than alpha WGDs (38.1%, $P < 10^{-6}$) and beta WGDs (37.0%, $P < 10^{-6}$) (Fig. 3). Thus WGDs show a higher degree of conservation of TFBSs than do TDs. Random gene pairs shared fewer TFBSs (25.9%) than all three duplicate types (Fig. 3). The difference in conservation between alpha and beta WGDs (Fig. 3) was not significant ($P = 0.073$).

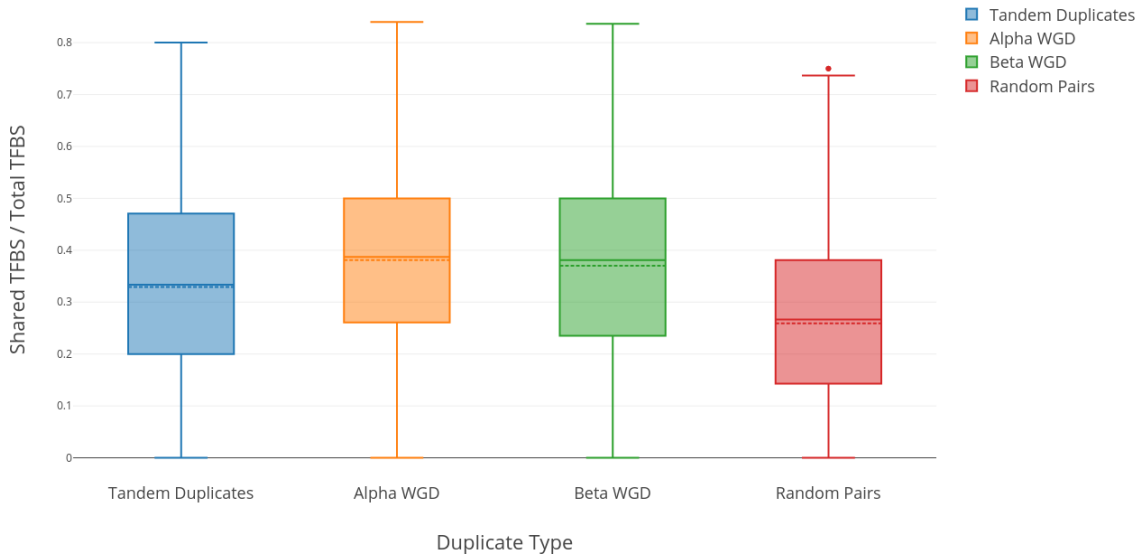


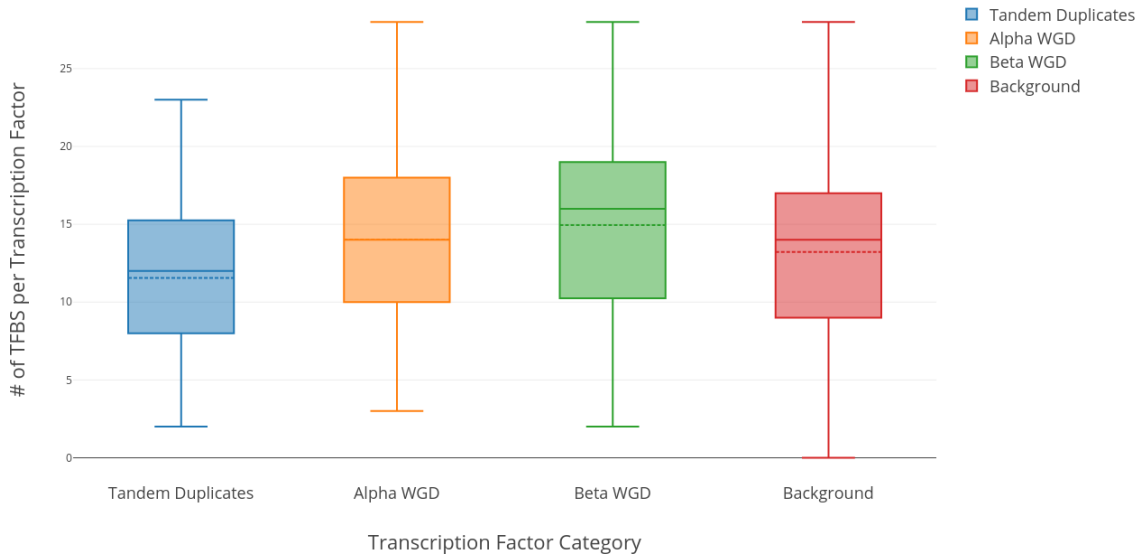
Figure 3. Tandem duplicates have less conserved TFBSs than WGDs. Shared TFBS are the TFBS found to be regulating both genes in a duplicate or random pair.

Regulatory Element and Target Gene Totals for Duplicate Transcription Factors

A duplicated gene with a large number of interactions should be more sensitive to dosage imbalances according to the gene balance hypothesis and transcription factors are a prime example of genes with many interactions (Birchler and Veitia, 2007). Therefore, I analyzed TFs that are themselves duplicates. There are only 25 transcription factors in the dataset that are also TDs, perhaps because TDs with many interactions are selected against, and thus low sample sizes made it harder to find statistically significant results. Still, TD TFs had significantly fewer TFBSs (11.56) than alpha (14.01, P = 0.035) and beta (14.95, P = 0.0072) WGDs (Fig. 4A). In general, TFs appear to have more TFBSs regulating them than non-TF genes (Fig. 1A). I then looked at the number of genes

targeted by duplicate TFs (Fig. 4B). No TFs in any category targeted a significantly different number of genes (ANOVA $P = 0.71$).

(A)


Tandem Duplicates
Alpha WGD
Beta WGD
Background
# of TFBS per Transcription Factor
25
20
15
10
5
0
Tandem Duplicates
Alpha WGD
Beta WGD
Background
Transcription Factor Category

(B)

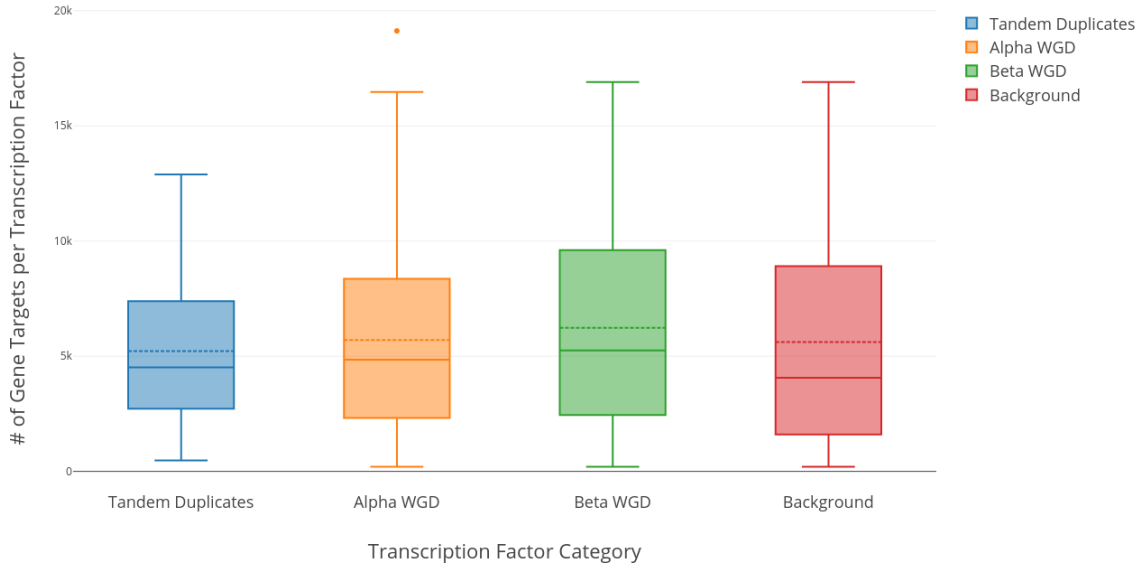
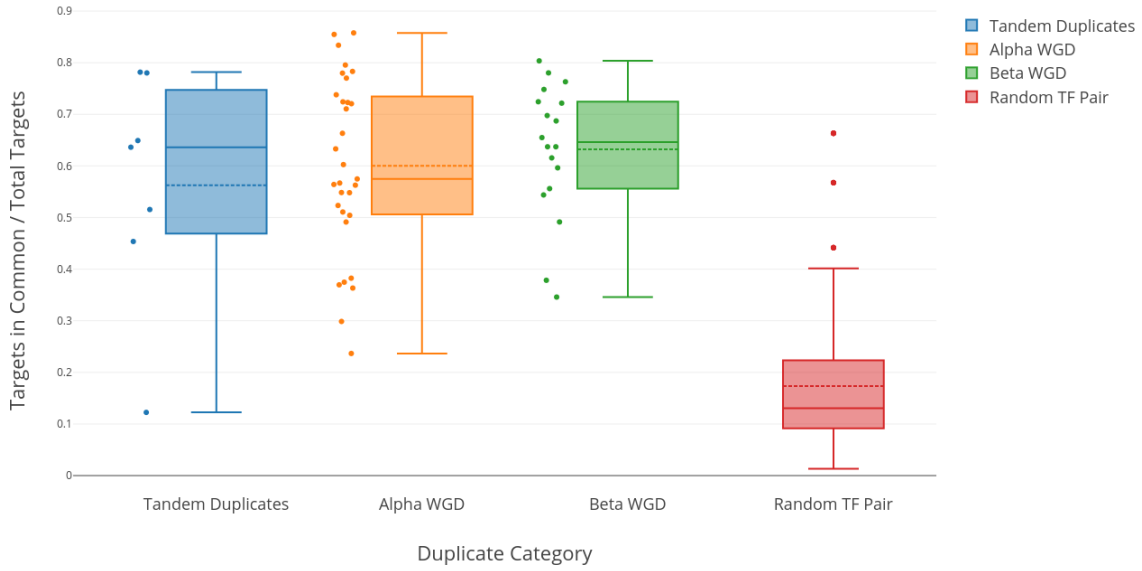

Tandem Duplicates
Alpha WGD
Beta WGD
Background
# of Gene Targets per Transcription Factor
20k
15k
10k
5k
0
Tandem Duplicates
Alpha WGD
Beta WGD
Background
Transcription Factor Category

Figure 4. Transcription factors are highly connected genes. (A) Box plots show the number of TFBS assigned to duplicate and random transcription factors. Mean and median are equal for alpha WGD box plot. (B) Box plots show the number of TFBS targeted by duplicate and random transcription factors.

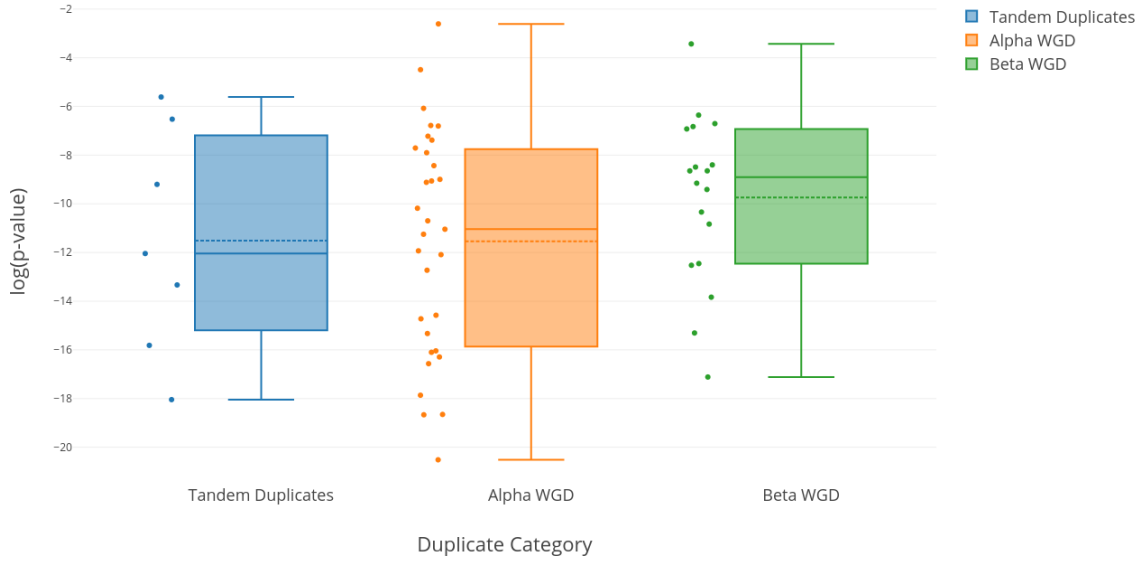
Conservation of Target Genes and Regulatory Elements for Duplicate Transcription Factors

I expected TD TFs to target fewer genes because I hypothesized that TD TFs might partition their original gene targets between themselves to avoid dosage imbalances. Another possibility is that TD TFs undergo a form of neofunctionalization where one paralog's target sequence changes in some way and it begins to target a whole new array of genes. To test this, I calculated the conservation of gene targets between TFs in duplicate pairs (Fig. 5A). I was not able to find significant evidence (ANOVA P = 0.62) that TD TFs share fewer targets in common than either type of WGD TFs. TFs in a TD pair share 56.3% of the same targets, a much higher percentage than for two transcription factors taken at random (17.3%, P = 0.0041) (Fig. 5A). I also compared the divergence of the target sequences themselves by using the position weight matrix comparison software TomTom (Fig. 5B). There was no significant difference (ANOVA P=0.35) in the mean log(p-value) for any duplicate category. Finally, I examined how the regulation of TFs themselves changes upon duplication (Fig. 5C). I examined the conservation of TFBSs between duplicate pair TFs. As before, pairs of TD TFs shared significantly fewer TFBSs (37.7%) on average than both alpha WGDs (51.2%, P = 0.019) and beta WGDs (56.6%, P = 0.0044). They shared a similar amount of TFBSs as randomly paired TFs (35.3%, P = 0.60). There was no significant difference between beta and alpha WGDs (P = 0.39) (Fig. 5C).

(A)


Tandem Duplicates
Alpha WGD
Beta WGD
Random TF Pair
Targets in Common / Total Targets
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Tandem Duplicates
Alpha WGD
Beta WGD
Random TF Pair
Duplicate Category

(B)


Tandem Duplicates
Alpha WGD
Beta WGD
log(p-value)
−2
−4
−6
−8
−10
−12
−14
−16
−18
−20
Tandem Duplicates
Alpha WGD
Beta WGD
Duplicate Category

(C)

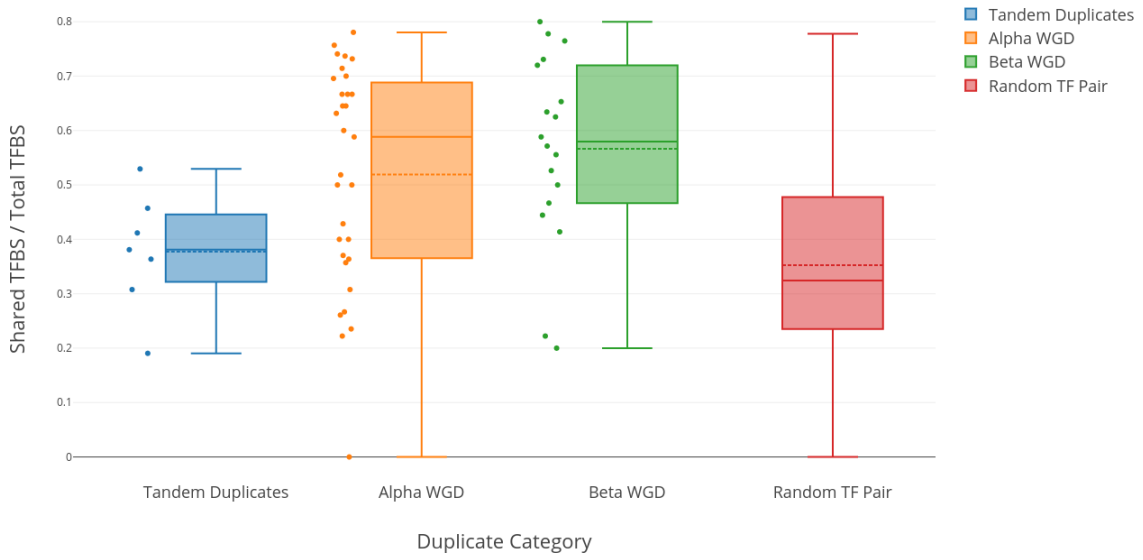


Figure 5. Divergence between duplicated transcription factors. (A) Box plots show the number of genes (counted once for each gene) that are targeted by both transcription factors over the total genes targeted. (B) Box plots show the similarity in position weight matrix (log(P-value)) for each transcription factor pair within the duplicate categories. P-values calculated using TomTom software from MEME Suite (version 4.12.0). (C) Box plots show the conservation of TFBSs between TFs in duplicate and random pairs. Data points are also plotted for the duplicate categories because box plots may not adequately capture the shape of sparse data.

Discussion

As in Arsovski et al. (2015), I found that TDs are regulated by fewer TFBSs than WGDs, although the difference does not seem to be as pronounced (Fig. 1a). This might be because the DHS dataset, used in Arsovski et al. (2015), included regulatory element data from only one tissue type (7 day old seedling), whereas DAP-seq is not tissue specific. WGDs are more likely to be highly connected genes such as TFs or involved in signal transduction, development etc. as predicted by the gene balance hypothesis (Blanc and Wolfe, 2004; Maere et al., 2005). These types of genes would be highly expressed in developing seedlings, potentially explaining why there is a larger discrepancy in the other study.

The difference in total regulatory elements between TDs and WGDs does not change after accounting for possibly incomplete TD regulatory regions (Fig. 1b). Background genes seem to be regulated by fewer TFBSs than WGDs as well. Since WGDs are enriched in highly connected genes the high number of TFBSs regulating them may just reflect their regulatory complexity.

There were a lot more TDs than alpha WGDs with two or fewer TFBSs (Fig. 2). The average TD with two or fewer TFBSs was a lot younger than the average for all TDs (Fig. 2a). There was seemingly no difference for alpha WGDs (Fig. 2b). However, since Ks is an approximation for the time since the duplication event, we should expect all alpha WGDs to have the same Ks value because all the pairs were duplicated at one time. The large range of Ks values seen make Ks an imperfect proxy for age. This result, at the very least, implies that the number of TFBSs does not somehow affect Ks values, thus we can be surer that the low Ks values seen in the filtered TDs represent young duplicates.

The pattern of expression or functional partitioning between genes in duplicate pairs tends to be highly asymmetric, with one copy hoarding significantly more subfunctions than the other (Zou et al., 2009b). This may happen in a regulatory fashion as well. TD genes with very few TFBSs may have survived by becoming highly specialized, expressed only in the presence of a few TFs and possibly in a very limited number of tissue types. However, this hypothesis does not explain why these genes tend to be younger. These sparsely regulated, young TDs could also be in the process of losing TFBSs. There is evidence that expressed pseudogenes are derived from recent duplications (Zou et al., 2009b). These could represent unannotated pseudogenes in the tandem duplicate set or even genes becoming pseudogenes. Normally genes become pseudogenes through point mutations resulting in premature stop codons or in-dels that result in frame shifts (Vanin, 1985). Arabidopsis pseudogenes usually have functional relatives involved in disease resistance, specialized (secondary) metabolism, cell wall modification, and protein degradation (Zou et al., 2009a). Before pseudogenization, one copy in a duplicate could become highly specialized, e.g. it aides in the defence against a specific virus or degrades a specific protein. A change in environment or biology might eliminate the need for that highly specialized function, removing any evolutionary pressure for keeping it in the genome. This would explain why these modestly regulated genes tend to be younger, as the older ones have long since been fully pseudogenized and are no longer recognizable due to purifying selection. A future study could look at the identities of the TFs that target these TD genes with few TFBSs and check if they are more likely to be involved in specialized responses.

In general, TDs are divergent in their regulation compared to WGDs. TDs have fewer conserved TFBSs between pairs than both alpha and beta WGDs (Fig. 3). The conservation of TFBSs between duplicates has been shown to decrease over time (Papp et al., 2003). The difference between alpha and beta WGDs is not statistically significant, but it seems that beta WGDs are slightly less conserved, as expected of older genes. By

the same logic, we should expect TDs to be more conserved than WGDs, as TDs are younger on average than WGDs (Haberer et al 2004). This conservation paradox is consistent with previous research and may be explained again by the gene balance hypothesis (Ganko et al., 2007). TDs are more likely to diverge in either their function or their regulation so as not to remain at 2x the dosage of background genes. Regulatory subfunctionalization and neofunctionalization would result in TFBSs unique to each gene in the duplicate - lowering the conserved ratio.

Highly connected genes –like TFs- would be most susceptible to dosage imbalances and therefore should be the most extreme examples of low TFBS conservation in TDs. This is exactly the case we see in Figure 5C. The conservation of TFBSs in TD TFs is at background level. TD TFs also appear to be regulated by fewer TFBSs than WGD TFs (Fig. 4.A). This could be evidence again of the gene balance hypothesis, well connected WGD TFs would have a higher chance of both being conserved than well connected TD TFs. TD TFs have even fewer TFBSs than non-duplicate TFs, although the difference is not significant. There is a strong correlation between the divergence of protein sequences and the similarity of expression between duplicates. More similar protein sequences have less similar expression patterns and vice versa (Ganko et al. 2007). We might expect TD TFs to target fewer genes than their WGD counterparts - partitioning of targets would be a form of subfunctionalization. However, we observe no significant difference in the number of targets per TF between any duplicate category (Fig. 4.B). TD TFs don't appear to undergo neofunctionalization of target genes either (Fig. 5.A), with TD and WGD TFs targeting roughly the same percentage of similar genes per pair. There is also remarkable conservation of similar targets between duplicate pairs of TFs – being much higher than randomly paired TFs. This is due to duplicate TFs having highly similar target sequences, regardless of duplicate family (Fig. 5.B). Overall, it seems that if there is a difference in how TD TFs evolve compared to WGD TFs it is in how the TFs are regulated and not their DNA binding domains. Duplicate TFs are similar in function but not in their

expression patterns.

I have provided evidence showing the divergence of regulatory elements is more extreme between TDs compared to WGDs. While not direct evidence of divergence in expression, I believe that examining regulatory elements is a useful way to assess potential expression profiles. Future studies can check my results using techniques like RNA Seq to see if they hold for direct expression. Future studies could also look at how epigenetics affects gene regulation and evolution. The DAP-Seq dataset includes the TFBSs for hundreds of transcription factors exposed to unmethylated DNA. I've only scratched the surface of what this dataset can tell us about gene evolution.

References

Arsovski, A. A., Pradinuk, J., Guo, X. Q., Wang, S., & Adams, K. L. (2015). Evolution of cis-regulatory elements and regulatory networks in duplicated genes of Arabidopsis. *Plant Physiology*, *169*, 2982-2991.

Berger, M.F., & and Bulyk, M.L. (2009). Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nature Protocols 4*, 393–411.

Birchler, J. A., & Veitia, R. A. (2007). The gene balance hypothesis: from classical genetics to modern genomics. *The Plant Cell*, *19*, 395-402.

Blanc, G., & Wolfe, K. H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant Cell*, *16*, 1667-1678.

Casneuf, T., De Bodt, S., Raes, J., Maere, S., & Van de Peer, Y. (2006). Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant Arabidopsis thaliana. *Genome Biology*, *7*, R13.

De Bodt, S., Maere, S., & Van de Peer, Y. (2005). Genome duplication and the origin of angiosperms. *Trends in Ecology & Evolution*, *20*, 591-597.

Edger, P.P., Heidel-Fischer, H.M., Bekaert, M., Rota, J.,…& Wheat, C.W. (2015). The butterfly plant arms-race escalated by gene and genome duplications. *Proceedings of the National Academy of Sciences USA, 112,* 8362-8366.

Fawcett, J. A., Maere, S., & Van de Peer, Y. (2009). Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proceedings of the National Academy of Sciences USA*, *106*, 5737-5742.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., & Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, *151*, 1531-1545.

Ganko, E. W., Meyers, B. C., & Vision, T. J. (2007). Divergence in expression between duplicated genes in Arabidopsis. *Molecular Biology and Evolution*, *24*, 2298-2309.

Gu, Z., Steinmetz, L. M., Gu, X., & Scharfe, C. (2003). Role of duplicate genes in genetic robustness against null mutations. *Nature*, *421*, 63.

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., & Noble, W. S. (2007). Quantifying similarity between motifs. *Genome Biology*, *8*, R24.

Haberer G, Hindemitt T, Meyers BC, & Mayer KFX (2004) Transcriptional similarities, dissimilarities, and conservation of cis-elements in duplicated genes of Arabidopsis. *Plant Physiology,* 136, 3009–3022.

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., ... & Soltis, D. E. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature*, *473*, 97.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., ... & Bonke, M. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research*, *20*, 861-873.

Kellis, M., Birren, B. W., & Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature*, *428*, 617.

Kheradpour, P., & Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Research*, *42*, 2976–2987.

Lee, T. H., Tang, H., Wang, X., & Paterson, A. H. (2012). PGDD: a database of gene and genome duplication in plants. *Nucleic Acids Research*, *41*, D1152-D1158.

Lynch, M., & Conery, J. S. (2000a). The evolutionary fate and consequences of duplicate genes. *Science*, *290*, 1151-1155.

Lynch, M., & Force, A. (2000b). The probability of duplicate gene preservation by subfunctionalization. *Genetics*, *154*, 459-473.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., & Van de Peer, Y. (2005). Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 5454-5459.

Ohno, S., Wolf, U., & Atkin, N. B. (1968). Evolution from fish to mammals by gene duplication. *Hereditas*, *59*, 169-187.

O'Malley, R. C., Huang, S. S. C., Song, L., Lewsey, M. G., Bartlett, A., Nery, J. R., ... & Ecker, J. R. (2016). Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell*, *165*, 1280-1292.

Panchy, N., Lehti-Shiu, M., & Shiu, S. H. (2016). Evolution of gene duplication in plants. *Plant Physiology*, *171*, 2294-2316.

Papp, B., Pál, C., & Hurst, L. D. (2003). Evolution of cis-regulatory elements in duplicated genes of yeast. *TRENDS in Genetics*, *19*, 417-422.

Renny-Byfield, S., & Wendel, J. F. (2014). Doubling down on genomes: polyploidy and crop plants. *American Journal of Botany*, *101*, 1711-1725.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., ... & Thompson, J. D. (2011). Fast, scalable generation of high- quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*, 539.

Sievert, C., Parmer, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M., & Despouy, P. (2016). plotly: Create interactive web graphics via Plotly's JavaScript graphing library [Software].

Siggers, T., & Gordân, R. (2013). Protein–DNA binding: complexities and multi-protein codes. *Nucleic Acids Research*, *42*, 2099-2111.

Soltis, D. E., Bell, C. D., Kim, S., & Soltis, P. S. (2008). Origin and early evolution of angiosperms. *Annals of the New York Academy of Sciences*, *1133*, 3-25.

Soltis, D. E., Visger, C. J., & Soltis, P. S. (2014). The polyploidy revolution then… and now: Stebbins revisited. *American Journal of Botany*, *101*, 1057-1078.

Stamatoyannopoulos, J. A., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D. M., ... & Giste, E. (2012). An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biology*, *13*, 418.

Sullivan, A. M., Arsovski, A. A., Lempe, J., Bubb, K. L., Weirauch, M. T., Sabo, P. J., ... & Stergachis, A. B. (2014). Mapping and dynamics of regulatory DNA and transcription factor networks in A. thaliana. *Cell Reports*, *8*, 2015-2030.

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, *34*, W609-W612.

Vanin, E. F. (1985). Processed pseudogenes: characteristics and evolution. *Annual Review of Genetics*, *19*, 253-272.

Veitia, R. A., & Potier, M. C. (2015). Gene dosage imbalances: action, reaction, and models. *Trends in Biochemical Sciences*, *40*, 309-317.

Wang, Y., Tan, X., & Paterson, A. H. (2013). Different patterns of gene structure divergence following gene duplication in Arabidopsis. *BMC Genomics*, *14*, 652.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, *18*, 292-298.

Zhang, Z., Li, J., Zhao, X. Q., Wang, J., Wong, G. K. S., & Yu, J. (2006). KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics, Proteomics & Bioinformatics*, *4*, 259-263.

Zou, C., Lehti-Shiu, M. D., Thibaud-Nissen, F., Prakash, T., Buell, C. R., & Shiu, S. H. (2009a). Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. *Plant Physiology*, *151*, 3-15.

Zou, C., Lehti-Shiu, M. D., Thomashow, M., & Shiu, S. H. (2009b). Evolution of stress-regulated gene expression in duplicate genes of Arabidopsis thaliana. *PLoS Genetics*, *5*, e1000581.