GENES ENCODING FLOWER- AND ROOT-SPECIFIC FUNCTIONS ARE MORE RESISTANT TO FRACTIONATION THAN GLOBALLY EXPRESSED GENES IN *BRASSICA RAPA*

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Biology

by

Naiyerah Kolkailah

June 2016

COMMITTEE MEMBERSHIP

TITLE: Genes Encoding Flower- and Root-Specific Functions are More Resistant to Fractionation than Globally Expressed Genes in *Brassica rapa*

AUTHOR: Naiyerah Kolkailah

DATE SUBMITTED: June 2016

COMMITTEE CHAIR: Ed Himelblau, Ph.D.
Associate Professor of Biology

COMMITTEE MEMBER: Jenn Yost, Ph.D.
Assistant Professor of Biology

COMMITTEE MEMBER: Sandra Clement, Ph.D.
Assistant Professor of Biology

ABSTRACT

Genes Encoding Flower- and Root-Specific Functions are More Resistant to Fractionation than Globally Expressed Genes in *Brassica rapa*

Naiyerah Kolkailah

Like many angiosperms, *Brassica rapa* underwent several rounds of whole genome duplication during its evolutionary history. *Brassica rapa* is particularly valuable for studying genome evolution because it also experienced whole genome triplication shortly after it diverged from the common ancestor it shares with *Arabidopsis thaliana* about 17-20 million years ago. While many *B. rapa* genes appear resistant to paralog retention, close to 50% of *B. rapa* genes have retained multiple, paralogous loci for millions of years and appear to be multi-copy tolerant. Based on previous studies, gene function may contribute to the selective pressure driving certain genes back to singleton status. It is suspected that other factors, such as gene expression patterns, also play a role in determining the fate of genes following whole genome triplication. Published RNA-seq data was used to determine if gene expression patterns influence the retention of extra gene copies. It is hypothesized that retention of genes in duplicate and triplicate is more likely if those genes are expressed in a tissue-specific manner, as opposed to being expressed globally across all tissues. This study shows that genes expressed specifically in flowers and roots in *B. rapa* are more resistant to fractionation than globally expressed genes following whole genome triplication. In particular, there appears to have been selection on genes expressed specifically in flower tissues to retain higher copy numbers and for all three copies to exhibit the same flower-specific expression pattern. Future research to determine if these observations in *Brassica rapa* are consistent with other angiosperms that have undergone recent whole genome duplication would confirm that retention of flower-specific-expressed genes is a general feature in plant genome evolution and not specific to *B. rapa*.

Keywords: polyploidy, autopolyploidy, whole genome duplication, whole genome triplication, hexaploidy, gene expression, pseudogene, gene dosage

ACKNOWLEDGMENTS

I would like to thank all the undergraduate students in Dr. Ed Himelblau's lab who helped with identifying potential pseudogenes in the first year of my research. I thank Dr. John Walker for his insights on sample sizes and statistical analyses. I am very grateful to my committee members, Dr. Jenn Yost and Dr. Sandra Clement, for their generous advice and feedback on my research and thesis. Dr. Yost and Dr. Matt Ritter also inspired me to set daily goals to complete my thesis, for which I am very grateful. Special thanks goes to the Cal Poly Biological Sciences Department faculty, staff, and fellow graduate students for their motivation and constant encouragement. I would also like to express my gratitude to my parents, siblings and friends for their patience, and for believing in me. Finally, to my outstanding professor, mentor, and committee chair Dr. Himelblau: thank you for your deep wisdom, relentless guidance, and unwavering support—especially during challenging times.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

1. INTRODUCTION

An autopolyploid is an organism with more than two sets of chromosomes resulting from genome duplication within the same species (Wolfe, 2001; Ha et al., 2009). Autopolyploidy is a common occurrence in the evolutionary history of many plant species (Cui et al., 2006; Havananda et al., 2011; Parisod et al., 2016). Like many angiosperms, *Brassica rapa* underwent several rounds of whole genome duplication during its evolutionary history (Tang & Lyons, 2012). *Brassica rapa* is particularly valuable for studying genome evolution because it also experienced a hexaploidy event shortly after it diverged from the common ancestor it shares with *Arabidopsis thaliana* roughly 17-20 million years ago (Mun et al., 2009; Lin et al., 2014). This round of triplication is the most recent hexaploidy event known to have occurred in the angiosperm clade (Wang et al., 2011). Genome duplication in eukaryotes produces extensive genetic redundancy, which gives rise to novel gene functions over time (Ohno, 1970; Conant & Wolfe, 2008; Flagel & Wendel, 2009). This functional diversification may have contributed to the great morphological diversity observed in *B. rapa* today (Tang & Lyons, 2012).

Following its recent whole genome triplication, the three sub-genomes of *B. rapa* underwent differential gene loss, or biased fractionation, due to varying rates of mutation (mostly short deletions) occurring between the three sub-genomes (Cheng et al., 2012; Tang et al., 2012). The result of fractionation is that many genes present in three copies, or paralogs, immediately after triplication are today found in one or two copies. Many *B. rapa* genes appear resistant to paralog retention and rapidly return to single copy following duplication or triplication. Functional enrichment analysis was conducted in a

previous study to identify such multi-copy-resistant genes (i.e. genes found mostly in single copy status) across 20 different angiosperms, including *B. rapa* (De Smet et al., 2013). Genes involved in conserved cellular functions (i.e. DNA damage repair and replication) were found overrepresented among the orthologous groups (OGs) reverting back to single copy. Gene evolution simulation ruled out the possibility of random chance causing the observed number of single copy OGs, supporting the conclusion that selective pressure restores a set of common genes involved in core cellular processes back to single copy (De Smet et al., 2013).

Several hypotheses have been proposed to explain why some genes are under selective pressure to revert back to single copy. One hypothesis is that these particular genes are dosage sensitive; they may encode subunits of multi-protein complexes that require stoichiometric balance between the products (Birchler & Veitia, 2007; Veitia et al., 2008; Edger & Pires, 2009). For example, photosynthesis-related complexes require a balanced interaction between proteins produced from nuclear genes and chloroplast genes (Leister, 2003; De Smet et al., 2013). Since whole genome duplication affects the nuclear genome but not the chloroplast genome, extra nuclear protein production relative to chloroplast production can potentially disrupt the protein ratio required for normal photosynthetic activity. A second hypothesis is that the chance of developing dominant-negative alleles is reduced when genes revert back to single copy (De Smet et al., 2013). Dominant-negative alleles encode proteins that disrupt the function of the wild-type protein complexes (Herskowitz, 1987; Veitia, 2007). Restoring genes back to single-copy eliminates extra copies, which could potentially develop mutations and cause dominant-negative phenotypes.

Not all *B. rapa* genes are under such selective pressure to revert back to singleton status. While the *B. rapa* genome contains many multi-copy-resistant genes, other *B. rapa* genes are multi-copy tolerant. About 50% of *B. rapa* genes are thought to persist in multiple copies (Wang et al., 2011). These genes may be under reduced pressure to revert back to singleton status, or not enough time has lapsed before fractionation could take place. Some of these genes may play a role in environmental adaptation, in which case additive effects and finely regulated gene dosage may provide some selective advantage (Tang et al., 2012). Alternatively, functional divergence of duplicated genes (neofunctionalization) or divergence in expression patterns (subfunctionalization) may be mechanisms by which duplicated gene copies are retained in the genome (Lynch & Conery, 2000; Lynch & Force, 2000; Wolfe, 2001).

Investigating the expression pattern of multi-copy tolerant genes may help explain why some genes persist as duplicates and triplicates. If housekeeping genes perform conserved cellular functions in plant tissues, and most have reverted back to single-copy status, it may be that genes encoding highly tissue-specific functions are more tolerant to higher copy number and are therefore retained in two or three copies. The main goal of this study is to determine if there is a correlation between expression patterns of *B. rapa* genes and retention of these genes in duplicate or triplicate. Using the transcriptome of the *B. rapa* subspecies *pekinesis* (or Chiifu—a Chinese cabbage), this study aims to establish first if copy number distribution is the same for globally expressed genes (i.e. genes expressed in all tissues) and genes expressed in some or only one tissue. This study also aims to identify which tissue-specific-expressed genes show the same expression pattern across all paralogs. It is hypothesized that retention of genes in duplicate and

triplicate is more likely if those genes are expressed in a tissue-specific manner, as opposed to being expressed globally across all tissues.

2. METHODS

2.1 Existing *B. rapa* Genome and Transcriptome Data

In a previous study, RNA-seq data was generated from multiple tissues of the *B. rapa* accession Chiifu-401-42, the same Chinese cabbage variety used for whole genome sequencing (Tong et al., 2013). The raw RNA-seq data from this study was obtained from the NCBI Gene Expression Omnibus (accession number GSE43245). The retrieved file contains RNA expression data (in Fragments Per Kilobase of Transcript Per Million Fragments Mapped [FPKM]) for 41,020 *B. rapa* genes across six different plant tissues: root, stem, leaf, flower, silique and callus. Expression data is available for one sample each of stem, flower, silique and callus tissue, and for two root and two leaf samples. Three additional files were obtained from another study, containing *B. rapa* singleton, duplicate, and triplicate gene IDs, along with their corresponding *A. thaliana* orthologs (Wang et al., 2011).

2.2 Pseudogene Identification

R Studio software was used to subset the RNA expression data file by gene copy number, then by expression pattern (Fig. 1). First, three separate files were created with expression data for singleton, duplicate, and triplicate genes (average FPKM values for the two root and two leaf samples were calculated for each file and used in lieu of the two individual root and leaf tissue expression values for all subsequent data analyses). Once expression data was separated according to gene copy number (Fig. 1A), potential pseudogenes were removed from all three data files (Fig. 1B). Potential pseudogenes were defined as having zero FPKM values across all tissues. Genes showing zero

expression for all tissues were removed—along with any paralogs—before conducting any further analysis.

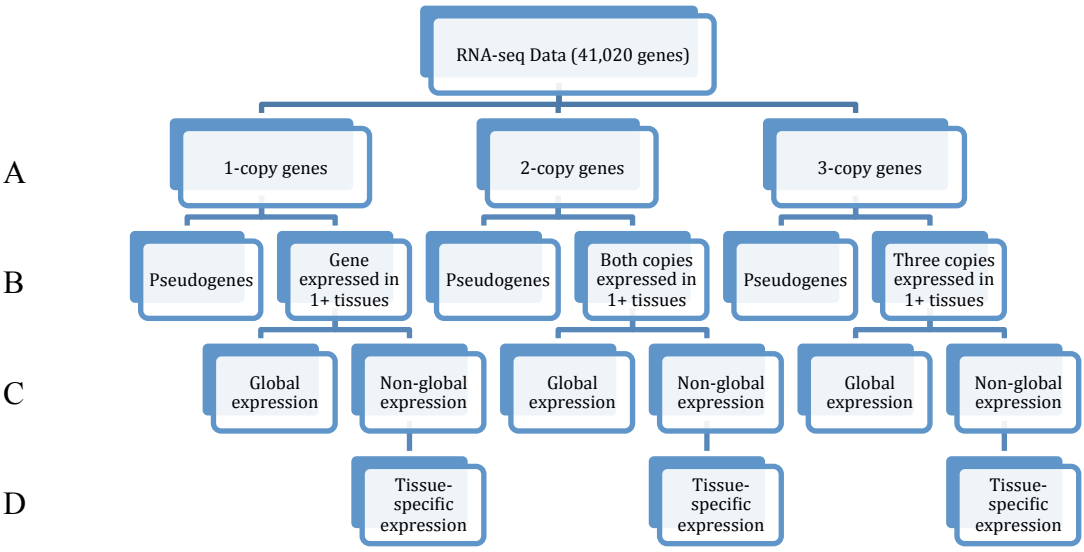

Figure 1. Flowchart for Subsetting RNA-seq Data. (A) Expression data was divided by copy number, (B) potential pseudogenes were removed, (C) globally expressed genes were isolated, and (D) non-globally expressed genes were divided into multiple sets of tissue-specific-expressed genes.

2.3 Subsetting Genes by Expression Pattern

2.3.1 Globally Expressed Genes

The remaining genes in all three files fall under one of three expression pattern categories: globally expressed genes (genes showing non-zero expression across all tissues), non-globally expressed genes (genes showing non-zero expression in one or more, but not all, tissues), or tissue-specific-expressed genes (genes showing >0.1 FPKM values in only one tissue and <0.1 FPKM values in all other tissues). The first category of genes to be removed and grouped separately from each of the three larger data sets was

globally expressed genes (Fig. 1C). Three additional gene sets were created for all the globally expressed singletons, and all duplicates and triplicate sets for which at least one of the paralogs exhibited global expression.

2.3.2 Tissue-Specific-Expressed Genes

From the non-globally expressed genes, genes with tissue-specific expression were grouped into separate files (Fig. 1D), but not removed from the original file with non-globally expressed genes. Since genes with multiple copies may exhibit overlap in gene expression categories (e.g. one paralog may show leaf-specific expression while another may show stem-specific expression), all tissue-specific-expressed genes remained in the file so they could be counted accurately. Flower-specific-expressed genes (and their paralogs) were grouped first, followed by leaf-specific, stem-specific and root-specific-expressed genes. Callus-specific and silique-specific-expressed genes were not considered in this study due to the minimal number of genes showing expression patterns specific to those tissues.

2.4 Chi-Square Analyses

To compare copy number distribution between globally expressed genes and non-globally expressed genes, as well as globally expressed genes and each group of tissue-specific expressed genes, total gene sets showing each expression pattern were first counted among singleton, duplicate, and triplicate genes. Then, five independent Chi-square analyses were conducted in JMP® Pro 11.2.0 to determine whether or not there was a significant difference in copy number distribution between 1) globally expressed

genes and non-globally expressed genes, 2) globally expressed genes and flower-specific-expressed genes, 3) globally expressed genes and leaf-specific-expressed genes, 4) globally expressed and stem-specific-expressed genes, and 5) globally expressed and root-specific-expressed genes. For all statistical analyses, expression pattern (global or non-global/tissue-specific) was the explanatory variable (X), copy number (singleton, duplicate, or triplicate) was the response variable (Y), and the observed count of singleton, duplicate, or triplicate sets exhibiting each expression pattern was inputted as the frequency. Each individual test was conducted at a 1% significance level.

2.5 Gene Expression Patterns Across Paralogs

In our scheme for identifying tissue-specific expression, it is possible that not all paralogs have the same pattern of expression. To identify which tissue-specific-expressed genes show the *same* expression pattern for all paralogs, expression data for each set of duplicate and triplicate genes with tissue-specific expression was observed. For duplicate genes, a count was made of all gene sets with only one of the two paralogs showing the same expression pattern. Another count was made of all sets in which both copies showed the same expression pattern. Percentages were generated using the total number of gene sets exhibiting that form of tissue-specific expression. The same calculations were conducted for triplicate genes, with an additional count for gene sets in which two of the three paralogs showed the same expression pattern.

2.6 Single-Copy Genes & Globally Expressed Genes

The list of single-copy genes identified in a previous study (De Smet et al., 2013) was compared to the globally expressed genes identified in this study. R Studio software was used to identify *A. thaliana* gene IDs that are common to both gene lists. The percent of single-copy *A. thaliana* orthologs found as globally expressed genes in *B. rapa* was calculated for singletons, duplicates, and triplicates.

3. RESULTS

3.1 Potential Pseudogenes

Out of 7,812 singleton *B. rapa* genes with corresponding *A. thaliana* orthologs, 260 genes (3.33%) were identified as potential pseudogenes. These genes showed no expression (i.e. FPKM is 0.00) across all six tissues. Out of 5,438 duplicate gene sets with *A. thaliana* orthologs, 502 duplicate sets (9.23%) had at least one potential pseudogene. Out of 1,674 triplicate gene sets with *A. thaliana* orthologs, 208 triplicate sets (12.43%) had at least one potential pseudogene (Figure 2).

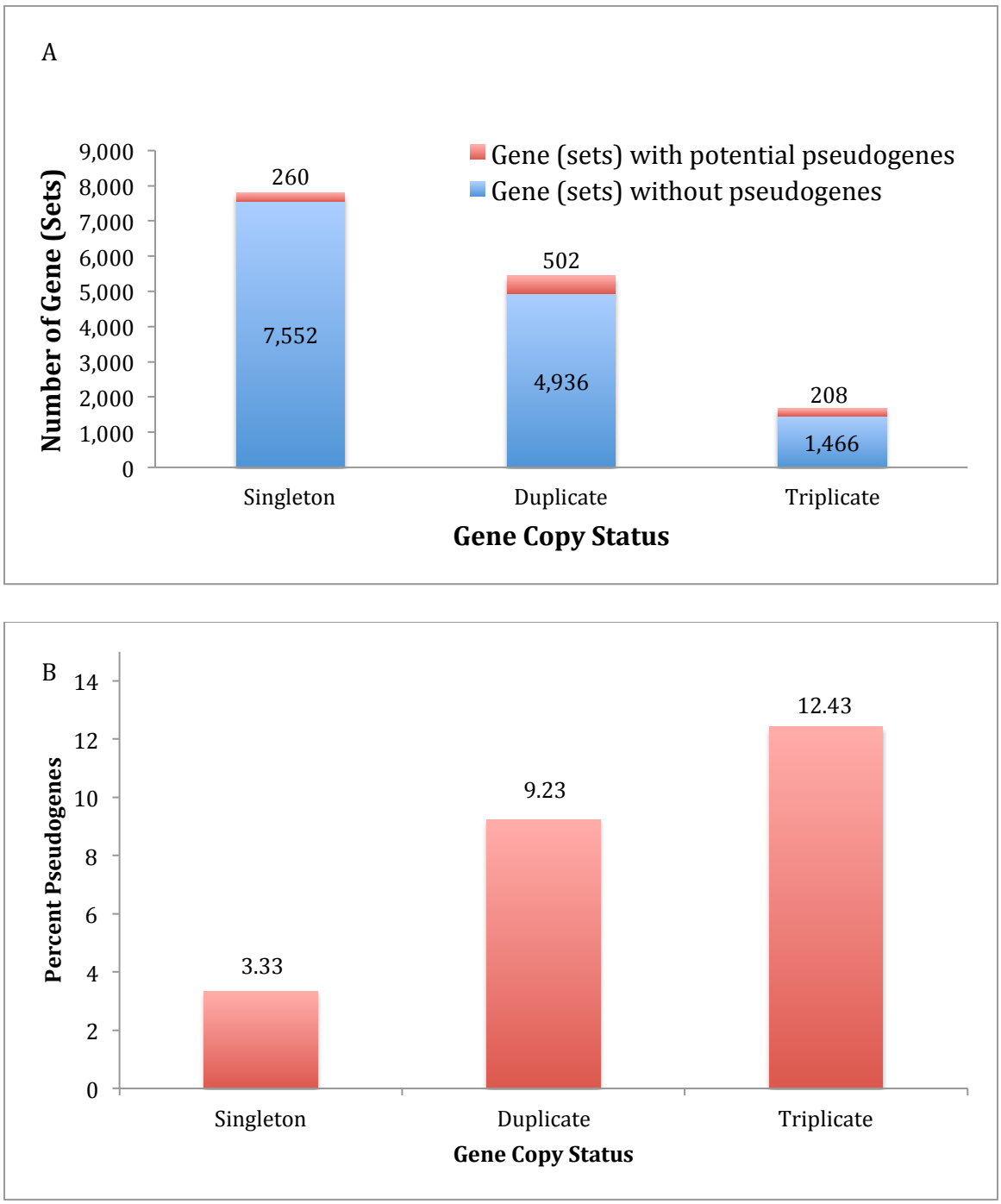

Figure 2. Pseudogenes and Non-Pseudogenes Among Singletons, Duplicates, and Triplicates. (A) Proportion (red) of total singleton, duplicate, and triplicate gene sets with at least one potential pseudogene. The majority of gene (sets), shown in blue, had non-zero expression for at least one of the tissues. (B) Percentage of total singleton,

duplicate, and triplicate gene sets with at least one potential pseudogene. Triplicates show the highest percentage of potential pseudogenes.

3.2 Comparing Globally Expressed and Non-Globally Expressed Genes

Genes were considered globally expressed if they had non-zero expression across all six tissues. Globally expressed genes were isolated from the singleton, duplicate and triplicate expression files, grouped with their paralogs and then counted. A total of 11,614 genes or gene sets included at least one globally expressed gene. Of this total, 6,053 (52.1%) were globally expressed singletons, and 4,261 duplicate sets (36.7%) and 1,300 triplicate sets (11.2%) had at least one globally expressed gene (Figure 2).

Non-globally expressed genes were genes showing non-zero expression in one or more, but not all, tissues. This set of genes includes all tissue-specific-expressed genes. A total of 1,462 genes or gene sets included at least one non-globally expressed gene. Of this total, 790 (54.0%) were non-globally expressed singletons; 506 duplicate sets (34.6%) and 166 triplicate sets (11.4%) had at least one non-globally expressed gene (Figure 3). There was no significant difference in copy number distribution between globally expressed and non-globally expressed genes (Chi-square=2.481, P>.2893).

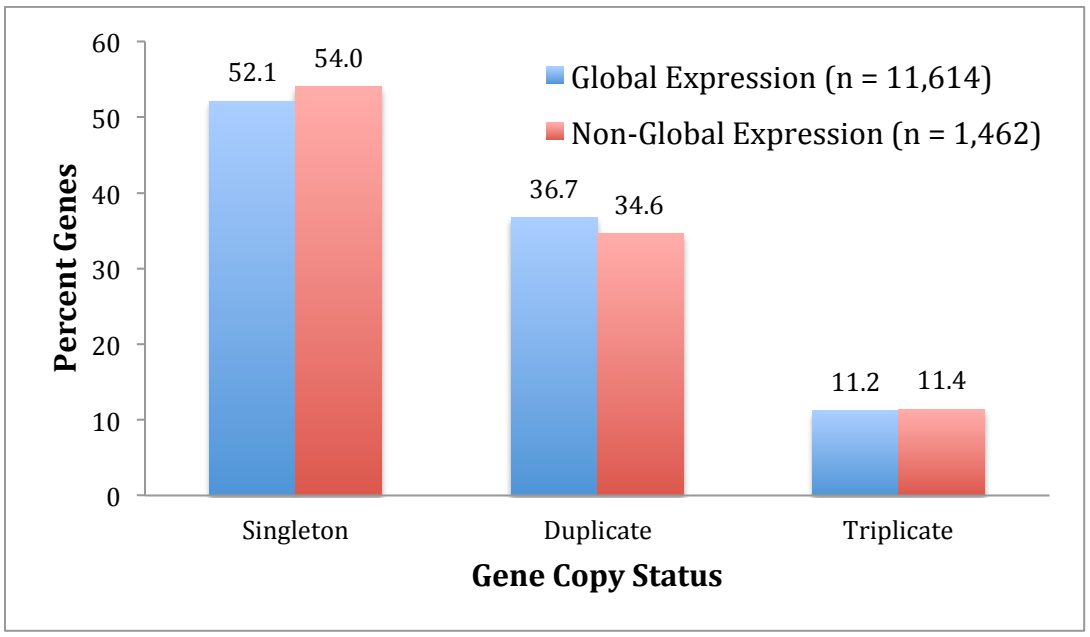

Figure 3. Global and Non-Global Expression Among Singletons, Duplicates, and Triplicates. Percentage of genes showing global expression (blue) and non-global expression (red).

3.3 Comparing Globally Expressed and Tissue-Specific-Expressed Genes

3.3.1 Flower-Specific-Expressed Genes

From the expression data containing non-globally-expressed genes, flower-specific-expressed genes were the first to be grouped with their paralogs and counted. Flower-specific-expressed genes show >0.1 FPKM values in the flower tissue and <0.1 FPKM values in all other tissues. In total, there were 201 genes or gene sets that included at least one flower-specific-expressed gene. Out of the 201 genes, 66 (32.8%) were singletons. There were 91 duplicate sets (45.3%) that had at least one flower-specific-expressed gene, and 44 triplicate sets (21.9%) with at least one flower-specific-expressed

gene (Figure 4). There was a significant difference in copy number distribution between globally expressed and flower-specific-expressed genes (Chi-square=38.013, P<.0001).

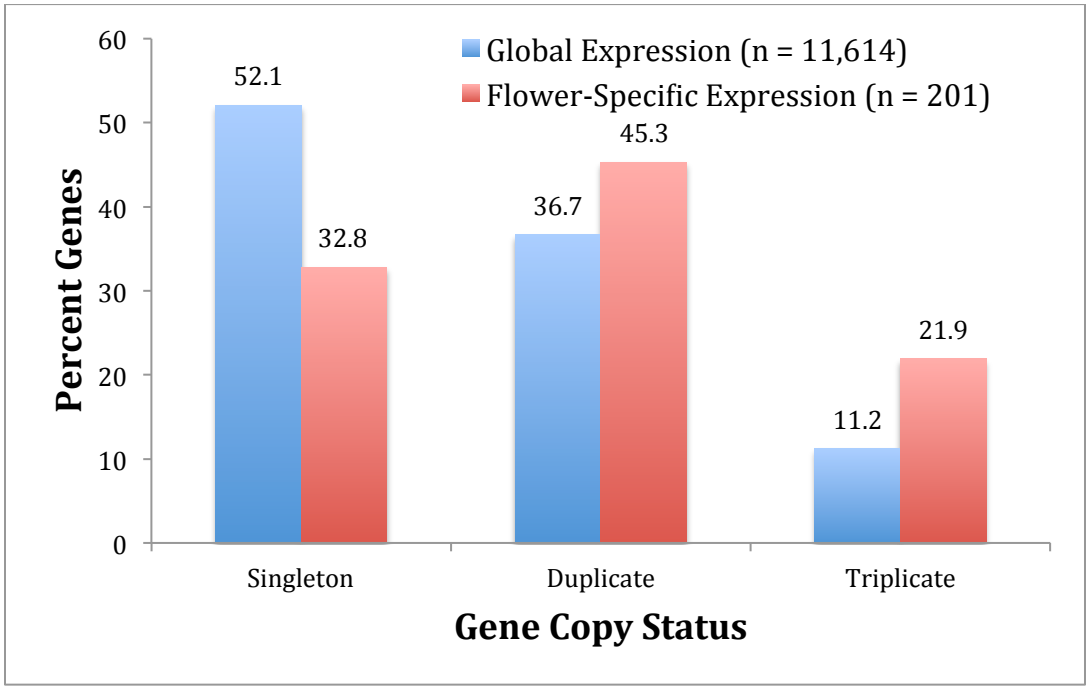

Figure 4. Global and Flower-Specific Expression Among Singletons, Duplicates, and Triplicates. Percentage of genes showing global expression (blue) and flower-specific expression (red).

3.3.2 Leaf-Specific-Expressed Genes

Leaf-specific-expressed genes were grouped with their paralogs and counted. These genes showed >0.1 FPKM values for the leaf tissue (averaged), and <0.1 FPKM values for all other tissues. There were 32 genes or gene sets that included at least one leaf-specific-expressed gene; 11 (34.4%) were singletons, 16 duplicate sets (50%) had at least one leaf-specific-expressed gene, and 5 triplicate sets (15.6%) had at least one leaf-specific-expressed gene. There was no significant difference in copy number distribution

between globally expressed and leaf-specific-expressed genes (Chi-square=4.028, P>.1334).

3.3.3 Stem-Specific-Expressed Genes

As with the last two sets of tissue-specific-expressed genes, genes showing stem-specific expression were grouped with their paralogs and counted without being removed from the expression data file containing non-globally expressed genes. A total of 20 genes or gene sets included at least one stem-specific-expressed gene; 9 (45%) were singletons, 6 duplicate sets (30%) had at least one leaf-specific-expressed gene, and 5 triplicate sets (25%) had at least one stem-specific-expressed gene. There was no significant difference in copy number distribution between globally expressed and stem-specific-expressed genes (Chi-square=3.831, P>.2265).

3.3.4 Root-Specific-Expressed Genes

Genes showing root-specific expression were the last of the tissue-specific-expressed genes to be grouped with their paralogs and counted. A total of 190 genes or gene sets included at least one root-specific-expressed gene; 63 (33.2%) were singletons, 89 duplicate sets (46.8%) had at least one root-specific-expressed gene, and 38 triplicate sets (20%) had at least one root-specific expressed gene. There was a significant difference in copy number distribution between globally expressed and root-specific-expressed genes (Chi-square=30.991, P<.0001).

3.4 Gene Expression Patterns Across Paralogs

In this analysis, duplicates and triplicates were designated as tissue-specific if at least one paralog showed tissue-specific expression. In these cases, it is possible that the other paralog(s) show the same expression pattern or a distinct pattern. The expression data for gene sets with at least one tissue-specific-expressed gene was examined to identify how many of the paralogs exhibited the same expression pattern.

3.4.1 Flower-Specific-Expressed Paralogs

A total of 201 genes and gene sets had at least one flower-specific-expressed gene. There were 66 singletons, 91 duplicates, and 44 triplicates that showed this expression pattern. For duplicate genes with at least one flower-specific-expressed gene, 38 sets (42%) showed flower-specific expression in only one of the two paralogs; 53 sets (58%) showed this same expression pattern in both copies (Figure 5). For triplicate sets, 8 sets (18%) showed flower-specific expression in one of the three paralogs; 7 (16%) showed it in two of the three paralogs; and 29 sets (22%) showed this expression pattern in all three paralogs (Figure 5).

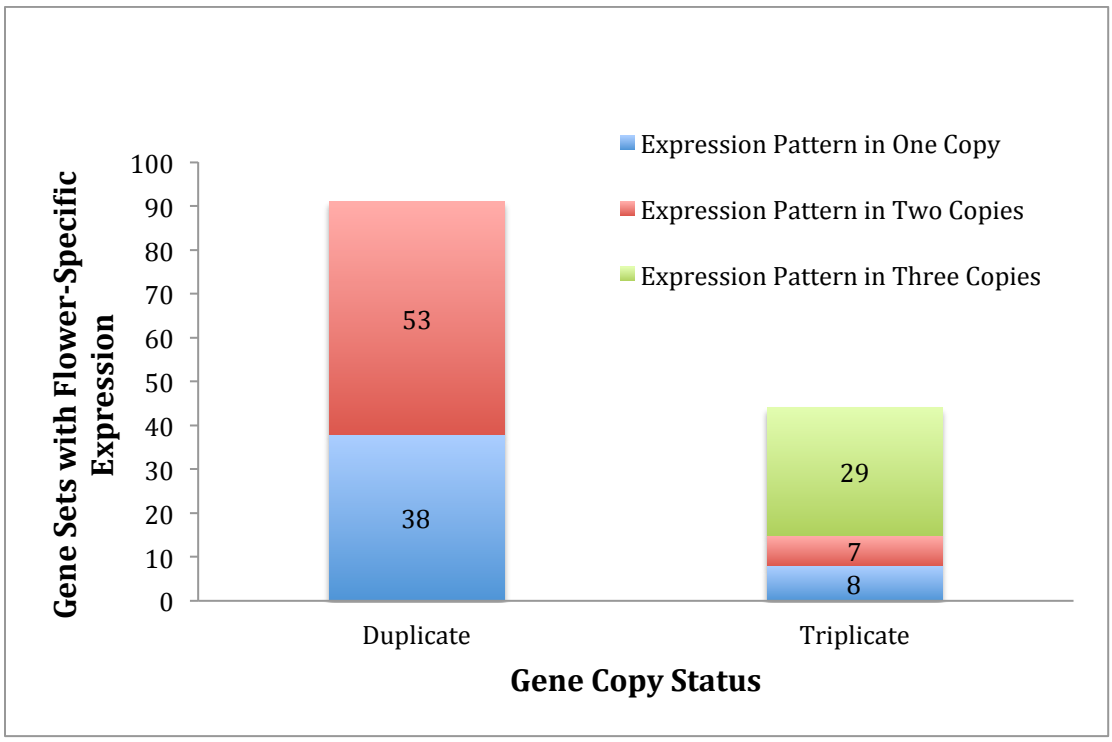

Figure 5. Flower-Specific Expression Among Duplicate and Triplicate Gene Sets. Bars are color-coded to show the proportion of sets with one copy (blue), two copies (red), or three copies (green) showing flower-specific expression. The majority of duplicate and triplicate sets showed flower-specific expression in all their respective gene copies.

3.4.2 Leaf-Specific-Expressed Paralogs

A total of 32 genes or gene sets had at least one leaf-specific-expressed gene; 11 singletons showed this expression pattern, and most duplicates and triplicate sets only showed leaf-specific expression in one of the paralogs. There were 15, out of 16 duplicate sets total, showing leaf-specific expression in only one of the two paralogs. For triplicate sets, 4 out of the 5 sets showed this same expression pattern in only one copy, and no sets showed leaf-specific expression in all three copies (Figure 6).

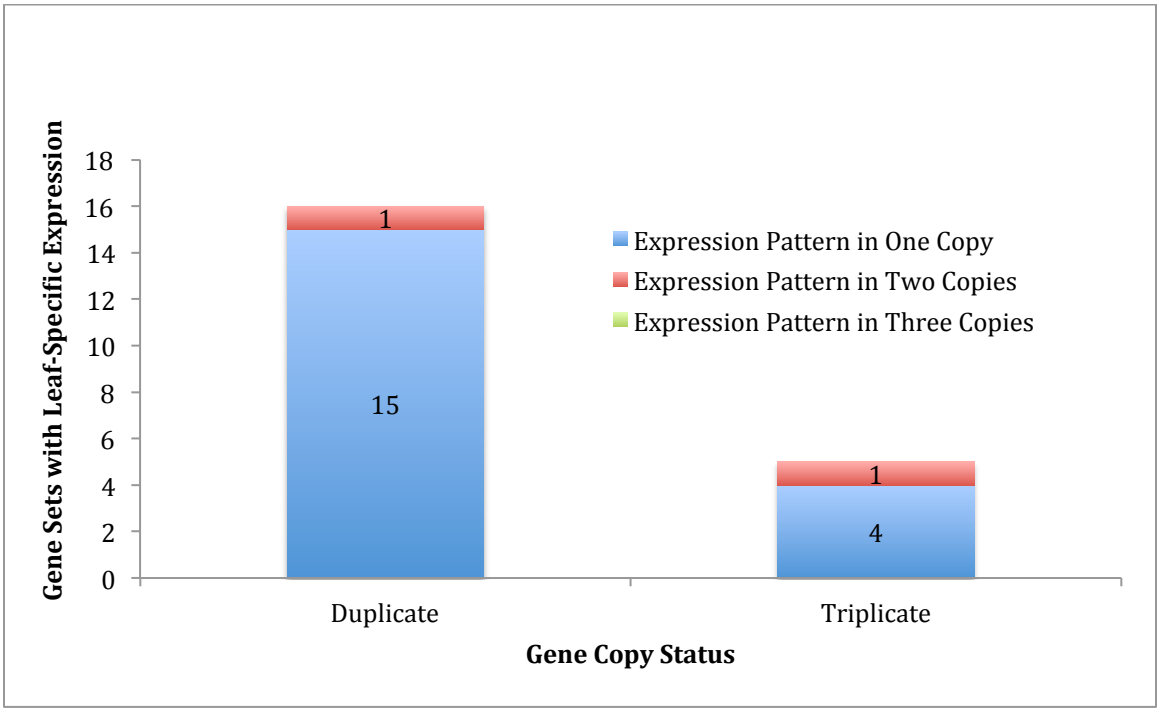

Figure 6. **Leaf-Specific Expression Among Duplicate and Triplicate Gene Sets.** Bars are color-coded to show the proportion of sets with one copy (blue), two copies (red), or three copies (green) showing leaf-specific expression. The majority of duplicate and triplicate sets showed leaf-specific expression in only one gene copy.

3.4.3 Stem-Specific-Expressed Paralogs

A total of 20 genes and gene sets had at least one stem-specific-expressed gene. There were 9 singletons showing this type of expression. All duplicates (6/6) showed leaf-specific expression in only one of the two paralogs. All triplicates (5/5) showed this same expression pattern in only one of the three paralogs (Figure 7).

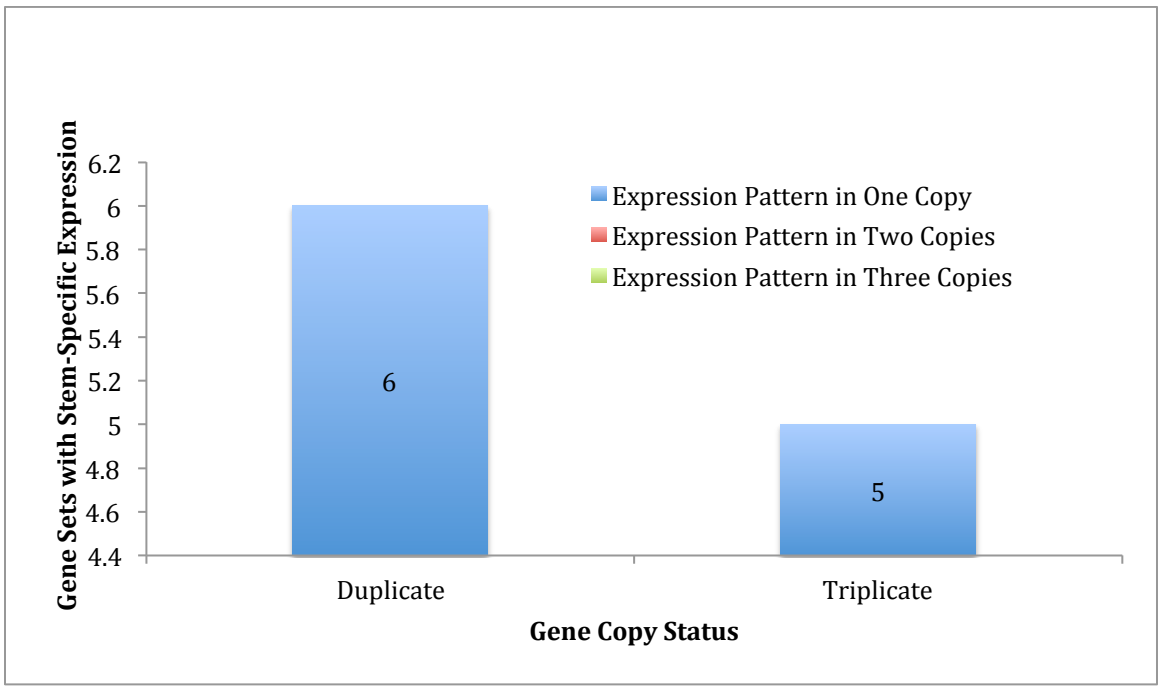

Figure 7. Stem-Specific Expression Among Duplicate and Triplicate Gene Sets. Bars are color-coded to show the proportion of sets with one copy (blue), two copies (red), or three copies (green) showing stem-specific expression. All duplicate and triplicate sets showed stem-specific expression in only one paralog; no gene sets showed stem-specific expression in multiple gene copies.

3.4.4 Root-Specific Expressed Paralogs

A total of 62 singleton genes showed root-specific expression. For duplicate genes with at least one root-specific-expressed gene, 58 sets (65.2%) showed root-specific expression in only one of the two paralogs; 31 sets (34.8%) showed this same expression pattern in both copies (Figure 6). For triplicate sets, 15 sets (39.5%) showed root-specific expression in one of the three paralogs; 16 (42.1%) showed it in two of the three paralogs; and 7 sets (18.4%) showed this expression pattern in all three copies (Figure 8).

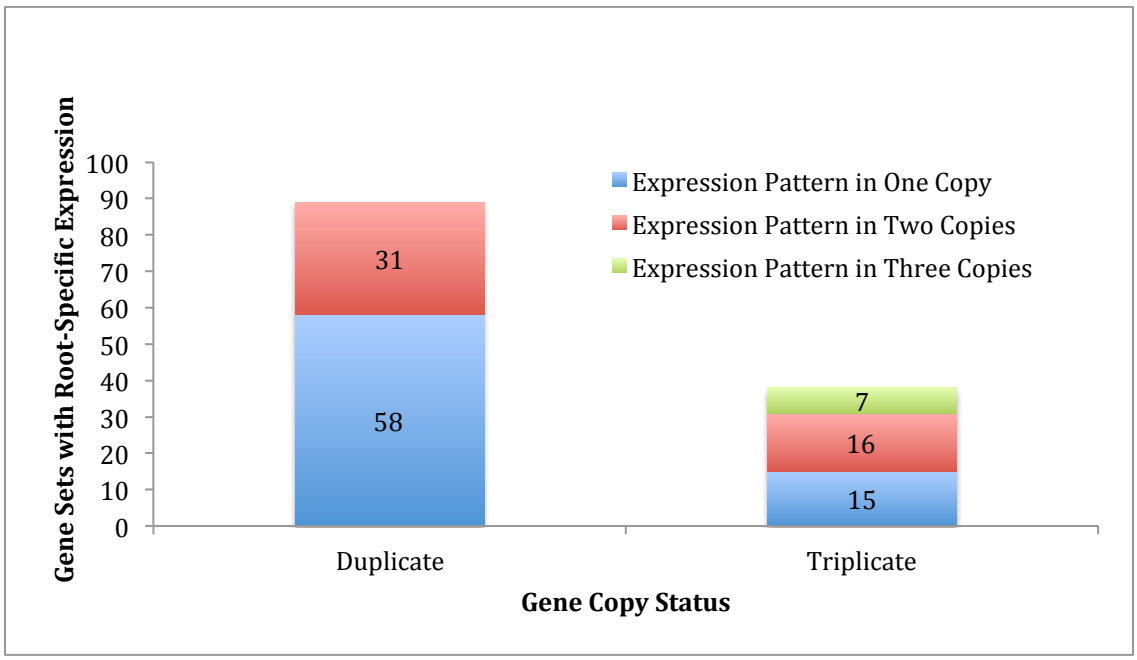

Figure 8. Root-Specific Expression Among Duplicate and Triplicate Gene Sets. Bars are color-coded to show the proportion of sets with one copy (blue), two copies (red), or three copies (green) showing root-specific expression. The majority of duplicates show root-specific expression in one paralog. Roughly the same number of triplicate sets showed root-specific expression in 1/3 and 2/3 copies.

3.5 Single-Copy Genes & Globally Expressed Genes

The final analysis compared percent overlap between single-copy genes identified in a previous study (De Smet et al., 2013), and globally and non-globally expressed genes identified in this study. The greatest overlap was observed between single-copy genes and globally expressed singleton genes. Of all the genes reverting back to single copy, 49.3% were globally expressed singleton genes; 19.1% were globally expressed duplicate genes; and 5.02% were globally expressed triplicates. Overlap was also examined between the single copy genes and non-globally expressed genes identified here; only

about 2.5% of single copy genes were non-globally expressed singletons, duplicates or triplicates.

4. DISCUSSION

Following whole genome triplication, the three sub-genomes of *B. rapa* underwent differential gene loss, which restored many triplicated genes back to singleton or duplicate status. This study utilized the transcriptome of *B. rapa* (subspecies *pekinesis)* to investigate the relationship between gene expression patterns and the retention of extra gene copies after whole genome triplication. Functional analysis in a previous study revealed that single copy genes in many angiosperm genomes tend to encode housekeeping functions (De Smet et al., 2013). While gene function may explain why some genes revert back to single copy, the present study examined if gene expression patterns across multiple *B. rapa* tissues influence the retention of genes in duplicate or triplicate—as opposed to the reduction to single copy status, which is the most common state in *B. rapa* (Cheng et al., 2012).

To compare expression patterns (i.e. globally expressed to non-globally expressed and tissue-specific-expressed), potential pseudogenes were first identified and removed—along with their paralogs—from the original gene expression file. One hallmark of pseudogenes is that they tend to have low or no expression and could, therefore, be miscounted in this analysis. Out of all *B. rapa* gene sets with corresponding *A. thaliana* orthologs, the greatest proportion of potential pseudogenes were found in triplicate gene sets, followed by those in duplicate gene sets and then singletons. This finding suggests that triplicate and duplicate genes may be undergoing pseudogenization to restore their status back to single copy. This also suggests that the published number of duplicates and triplicates is an overestimate and that diploidization is more advanced in *B. rapa* than previous studies have indicated (Cheng et al., 2012). Even in this study, the number of

duplicates and triplicates may be overestimated since genes with non-zero FPKM are considered viable when generally 0.1>FPKM represents no expression (Pat Edgar, personal communication).

To confirm pseudogene status, a small-scale study of 70 *B. rapa* triplicate genes revealed that paralogs with low sequence alignment scores had at least one gene copy with a large terminal deletion. That gene copy was also often missing exons relative to the paralogous loci (data not shown). Such deletions are strong indicators of gene inactivation and pseudogenization (Woodhouse et al., 2010; Tang et al., 2012). Applying a similar analysis to the 260 potential pseudogenes in this study can improve the annotation of the *B. rapa* genome and confirm the process of pseudogenization through DNA sequence examination.

Analysis of the remaining genes expressed in *B. rapa* revealed that retention of extra gene copies can be explained, in part, by gene expression patterns. The majority (52.1%) of globally expressed genes were singletons, 36.7% were a part of duplicate sets, and only 11.2% were part of triplicate sets. There was no significant difference in copy number distribution between globally expressed and non-globally expressed genes, showing that most genes expressed in all tissues or multiple tissues tend to revert back to singleton status. Only a small proportion of *B. rapa* genes performing functions across all or multiple tissues remain in triplicate.

A different result was observed for genes showing flower-specific expression. Copy number distribution for genes showing flower-specific expression differed significantly from the pattern observed for globally expressed genes. Singletons were more than 33%, duplicates were roughly 45% and triplicates were about 22%. Flower-

specific-expressed genes appear to be retained in duplicate and triplicate copies in significantly higher proportions than are globally expressed genes.

It is noteworthy that flowers express unique developmental pathways and reproductive processes (i.e. fertilization, meiosis and gamete development), which may be controlled by large regulatory networks (Franks, 2015) and multi-protein complexes that require stoichiometric balance between subunits to be maintained. Unlike small-scale duplications, whole genome duplication and triplication maintains the relative ratios between gene products and retains the stoichiometric balance between the different subunits of multi-protein complexes (Birchler & Veitia, 2010). Functional analysis of triplicate genes showing flower-specific expression could be conducted to identify the role these genes play in developmental pathways and regulatory networks unique to flowers, as well as the degree of networking between their gene products. Since this study showed that flower-specific-expressed genes were the most likely of all tissue-specific-expressed genes to show the same expression pattern across all three paralogs, there is an even greater possibility of paralogs contributing additively to the same conserved functions in flowers (Tang et al., 2012).

Although flower-specific-expressed genes showed the greatest difference in copy number distribution when compared to globally expressed genes, root-specific-expressed genes also exhibited a similar pattern to that of flower-specific-expressed genes (i.e. lower singleton count, and higher duplicate and triplicate counts than was observed for globally expressed genes). Roots express genes involved in environmental stress responses such as drought and salt stress (Tao et al., 2014), which are likely controlled by complex regulatory networks and would therefore be under pressure to retain copy

numbers similar to other root-specific, environmental response genes. A study of multi-copy genes involved in trace metal element responsive processes revealed that these genes are over-retained in the *B. rapa* genome, indicating a possible functional advantage for maintaining these genes in duplicate or triplicate (Li et al., 2014). Although this previous study analyzed differential gene expression in *B. rapa* leaves, similar processes may be at work in other plant tissues.

A relatively small number of genes showed stem-specific- and leaf-specific expression. The copy number distribution of both stem-specific and leaf-specific-expressed genes did not differ significantly from the distribution of globally expressed genes. Stems and leaves are both photosynthetic tissues, especially leaves. Functional enrichment analyses have revealed a class of single copy genes involved in organelle-related functions and photosynthetic processes (De Smet et al., 2013; Li et al., 2016). Since whole genome duplication only duplicates the nuclear genome and not the chloroplast genome, the stoichiometric balance between the nuclear and chloroplast-encoded subunits of photosynthetic complexes may be disrupted if more gene copies in the nuclear genome are expressed relative to chloroplast genes (De Smet et al., 2013). If genes encoding photosynthetic proteins are affected deleteriously by dosage imbalance, it is expected that genes expressed only in leaves and stems may be more resistant to retaining extra gene copies.

Based on the global and tissue-specific expression patterns observed in this study, it can be concluded that genes encoding flower- and root-specific functions are more resistant to fractionation than globally expressed genes in *B. rapa*. It is important to consider, however, that this study used RNA-seq data generated from *B. rapa* plants

grown only in greenhouse conditions, and tissue samples were from particular ages and developmental stages (Tong et al., 2013). Gene expression patterns may need to be re-examined under different growth conditions and at multiple developmental stages to determine if the observed expression patterns in this study are consistent throughout all plant stages of development, and whether or not they vary under different growth conditions.

The final analysis in this study showed overlap between globally expressed genes identified here and a previously published list of genes shown to rapidly return to single-copy status following whole genome duplication and whole genome triplication. The latter are considered multi-copy resistant genes. Approximately 50% of multi-copy resistant genes were present as single copy, globally expressed genes. However, multi-copy resistant genes were also found as two- and three-copy, globally expressed genes—but in lower abundance. Since the greatest overlap was found between multi-copy resistant genes involved mainly in core cellular processes and globally expressed singletons, it appears that many genes encoding housekeeping functions are expressed globally across all plant tissues. These results suggest that, along with function, gene expression pattern may also contribute to the selective pressure driving certain genes back to singleton status.

Future studies can further investigate the relationship between gene function and gene expression pattern as they relate to retention or loss of extra gene copies. These studies can employ functional analyses, gene knockout techniques and proteomics to investigate why globally expressed genes involved in housekeeping functions resist duplicate status. Retaining extra copies of globally expressed genes may have deleterious

effects, but these effects have yet to be examined in the light of both gene function and gene expression patterns. Studies can also aim to explain why certain tissue-specific-expressed genes retain their extra gene copies more readily than their globally expressed counterparts. Retaining extra copies of tissue-specific-expressed genes may enhance fitness or provide adaptive benefits—particularly flower- or root-specific-expressed genes showing the same expression pattern across all paralogs. These benefits have not been investigated sufficiently or considered in relation to both gene function and gene expression pattern.

This study revealed that in *B. rapa*, there appears to have been selection on flower genes to remain in three copies and for all three copies to be expressed in a narrow range of tissues. Future research to determine if these observations in *B. rapa* are consistent with other angiosperms that have undergone recent whole genome duplication would confirm that retention of flower-specific-expressed is a general feature in plant genome evolution, and not specific to *B. rapa*.

REFERENCES

Birchler, J. A., & Veitia, R. A. (2007). The gene balance hypothesis: from classical genetics to modern genomics. *The Plant Cell Online*, *19*(2), 395–402. doi:10.1105/tpc.106.049338

Birchler, J. A., & Veitia, R. A. (2010). The gene balance hypothesis: Implications for gene regulation, quantitative traits and evolution. *New Phytologist*, *186*(1), 54–62. doi:10.1111/j.1469-8137.2009.03087.x

Cheng, F., Wu, J., Fang, L., Sun, S., Liu, B., Lin, K., … Wang, X. (2012). Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. *PLoS ONE*, *7*(5), e36442. doi:10.1371/journal.pone.0036442

Conant, G. C., & Wolfe, K. H. (2008). Turning a hobby into a job: How duplicated genes find new functions. *Nature Reviews. Genetics*, *9*(12), 938–50. doi:10.1038/nrg2482

Cui, L., Wall, P. K., Leebens-Mack, J. H., Lindsay, B. G., Soltis, D. E., Doyle, J. J., … DePamphilis, C. W. (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Research*, *16*(6), 738–749. doi:10.1101/gr.4825606

De Smet, R., Adams, K. L., Vandepoele, K., Montagu, M. C. E. Van, & Maere, S. (2013). Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences*, 110(8):2898-903. doi:10.1073/pnas.1300127110

Edger, P. P., & Pires, J. C. (2009). Gene and genome duplications: The impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Research*, *17*(5), 699–717. doi:10.1007/s10577-009-9055-9

Flagel, L. E., & Wendel, J. F. (2009). Gene duplication and evolutionary novelty in plants. *New Phytologist*, *183*(3), 557–564. doi:10.1111/j.1469-8137.2009.02923.x

Franks, S. J. (2015). The unique and multifaceted importance of the timing of flowering. *American Journal of Botany*, *102*(9), 1401–1402. doi:10.3732/ajb.1500234

Ha, M., Kim, E.-D., & Chen, Z. J. (2009). Duplicate genes increase expression diversity in closely related species and allopolyploids. *Proceedings of the National Academy of Sciences*, *106*(7), 2295–2300. doi:10.1073/pnas.0807350106

Havananda, T., Charles Brummer, E., & Doyle, J. J. (2011). Complex patterns of autopolyploid evolution in alfalfa and allies(Medicago Sativa; Leguminosae). *American Journal of Botany*, *98*(10), 1633–1646. doi:10.3732/ajb.1000318

Herskowitz, I. (1987). Functional inactivation of genes by dominant negative mutations. *Nature*, *329*(6136), 219–222. doi:10.1038/329219a0

Leister, D. (2003). Chloroplast research in the genomic age. *Trends in Genetics*, *19*(1), 47–56. doi:10.1016/S0168-9525(02)00003-3

Li, J., Liu, B., Cheng, F., Wang, X., Aarts, M. G. M., & Wu, J. (2014). Expression profiling reveals functionally redundant multiple-copy genes related to zinc, iron and cadmium responses in *Brassica rapa*. *New Phytologist*, *203*(1), 182–194. doi:10.1111/nph.12803

Li, Z., Defoort, J., Tasdighian, S., Maere, S., Van de Peer, Y., & De Smet, R. (2016). Gene duplicability of core genes is highly consistent across all angiosperms. *The Plant Cell*, *32*(0), TPC2015–00877–LSB. doi:10.1105/tpc.15.00877

Lin, K., Zhang, N., Severing, E. I., Nijveen, H., Cheng, F., Visser, R. G., … Bonnema, G. (2014). Beyond genomic variation - comparison and functional annotation of three Brassica rapa genomes: a turnip, a rapid cycling and a Chinese cabbage. *BMC Genomics*, *15*(1), 250. doi:10.1186/1471-2164-15-250

Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science (New York, N.Y.)*, *290*(5494), 1151–1155. doi:10.1126/science.290.5494.1151

Lynch, M., & Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics*, *154*(1), 459–473. doi:10.1371/journal.pgen.0040029

Mun, J.-H., Kwon, S.-J., Yang, T.-J., Seol, Y.-J., Jin, M., Kim, J.-A., … Park, B.-S. (2009). Genome-wide comparative analysis of the Brassica rapa gene space reveals genome shrinkage and differential loss of duplicated genes after whole genome triplication. *Genome Biology*, *10*(10), R111. doi:10.1186/gb-2009-10-10-r111

Ohno, S. (1970). *Evolution by gene duplication.* New York, NY, USA: Springer-Verlag.

Parisod, C., Holderegger, R., Brochmann, C. (2016). Evolutionary consequences of autopolyploidy. *New Phytologist*, *186*(1), 5–17. doi:10.1111/j.1469-8137.2009.03142.

Tang, H., & Lyons, E. (2012). Unleashing the genome of *Brassica rapa*. *Frontiers in Plant Science*, *3*(July), 172. doi:10.3389/fpls.2012.00172

Tang, H., Woodhouse, M. R., Cheng, F., Schnable, J. C., Pedersen, B. S., Conant, G., … Pires, J. C. (2012). Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy. *Genetics*, *190*(4), 1563–1574. doi:10.1534/genetics.111.137349

Tao, P., Zhong, X., Li, B., Wang, W., Yue, Z., Lei, J., … Huang, X. (2014). Genome-wide identification and characterization of aquaporin genes (AQPs) in Chinese cabbage (*Brassica rapa* ssp. *pekinensis*). *Molecular Genetics and Genomics*, 1131–1145. doi:10.1007/s00438-014-0874-9

Tong, C., Wang, X., Yu, J., Wu, J., Li, W., Huang, J., … Liu, S. (2013). Comprehensive analysis of RNA-seq data reveals the complexity of the transcriptome in *Brassica rapa*. *BMC Genomics*, *14*(1), 689. doi:10.1186/1471-2164-14-689

Veitia, R. A. (2007). Exploring the molecular etiology of dominant-negative mutations. *The Plant Cell Online*, *19*(12), 3843–3851. doi:10.1105/tpc.107.055053

Veitia, R. A., Bottani, S., & Birchler, J. A. (2008). Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. *Trends in Genetics*, *24*(8), 390–397. doi:10.1016/j.tig.2008.05.005

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., … Zhang, Z. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics*. doi:10.1038/ng.919

Wolfe, K. H. (2001). Yesterday's polyploids and the mystery of diploidization. *Nature Reviews*, *2*(May), 333–341.

Woodhouse, M. R., Schnable, J. C., Pedersen, B. S., Lyons, E., Lisch, D., Subramaniam, S., & Freeling, M. (2010). Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homologs. *PLoS Biology*, *8*(6). doi:10.1371/journal.pbio.1000409