THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF BIOLOGY

Development of a R package for Classifying Duplicate Gene Retention Mechanisms (CDROM) and Application to Duplicate Genes in Grasses

BRENT R. PERRY
SPRING 2017

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Biology
with honors in Biology

Reviewed and approved* by the following:

Raquel Assis
Assistant Professor of Biology
Thesis Supervisor

Timothy Jegla
Associate Professor of Biology
Honors Adviser

* Signatures are on file in the Schreyer Honors College.

ABSTRACT

Gene duplication is a major source of new genes and is thought to have an important role in genomic evolution. Though there are several proposed mechanisms of long-term retention of duplicate genes, their genome-wide prevalence remains unclear in a majority of species. Assis and Bachtrog (2013) developed a phylogenetic approach for classifying these duplicate gene retention mechanisms on a genome-wide scale. In Chapter 1, we implement their phylogenetic approach as the R package, *CDROM*, short for Classification of Duplicate gene RetentiOn Mechanisms (Perry and Assis 2016). *CDROM* is the first tool capable of classifying duplicate gene retention mechanisms on a genome-wide scale, can be applied to a number of species and datasets, runs quickly, and is user-friendly. In Chapter 2, we apply *CDROM* to duplicate genes in three grass species: *Brachypodium distachyon*, *Oryza sativa*, and *Sorghum bicolor*. Our findings reveal that a variety of mechanisms may retain duplicate genes in grasses, though interestingly, also indicate that subfunctionalization may not play as significant a role as hypothesized. Thus, we have developed a useful tool for studying duplicate gene retention mechanisms in a variety of species, as well as applied it to gain novel insight into the mechanisms retaining duplicate genes in grasses over long evolutionary timescales.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Dr. Raquel Assis who has provided me with opportunities that I had never dreamed of as an undergraduate researcher, such as the opportunity to publish research in a peer-reviewed journal. I will always be grateful for your support and advice during my college career. Second, I would like to thank the current and past members of Assis Group: Jacob Heldenbrand, Xueyuan Jiang, and Balaji Kumar, for becoming my friends over the past few years. I wish you all the best of luck with your future research and careers. Third, I would like to thank my honors advisor Dr. Timothy Jegla for his help in reviewing my thesis. Finally, I would like to thank my parents who constantly supported and encouraged me to challenge myself. My college career might not have been as successful without all of those extra math problems I complained about when I was younger.

Introduction

Since the first genome sequence of a multicellular organism, *Caenorhabditis elegans,* was released (*C. elegans* Sequencing Consortium 1998), the genomes of countless other species have been sequenced. As these genome sequences became available, computational and statistical methods were and continue to be developed to study a variety of genomic features and evolutionary phenomena. For example, microarray and RNAseq experiments provide information about gene expression levels across the genome. Large databases were created to store these datasets, making most of the information widely available to researchers. This has led to the expansion of the field of computational biology, as knowledge of computer programming provides a platform for interpretation of these large datasets.

Often, the methods developed to perform such analyses require quantitative skills. While these methods are extremely useful and yield impactful results, they are limited to those researchers with knowledge of computer programming. As a result, many of these methods may not be used as widely as they could be if they were accessible to researchers without programming knowledge. Therefore, bioinformatics researchers may choose to turn their methods into software that is accessible to researchers with minimal programming knowledge. In addition to making their methods more accessible, software often enables faster and easier data analysis.

In Chapter 1 of this thesis, we develop a user-friendly R package that implements Assis and Bachtrog's (2013) phylogenetic approach for classifying duplicate gene retention mechanisms on a genome-wide scale. We call this tool *CDROM*, as it is used for Classification of Duplicate gene RetentiOn Mechanisms. *CDROM* is the first tool developed for performing this type of analysis, can be applied to a variety of datasets from any species, runs quickly, and can be used by researchers without programming knowledge. In Chapter 2, we apply *CDROM* to duplicate genes from three species of grasses to perform the first genome-wide analysis of duplicate gene retention mechanisms in this lineage. Though we uncover support for several mechanisms, we find little evidence of subfunctionalization in grasses, contrary to what has been hypothesized for many plants. Thus, in this thesis, we design the first tool for studying genome-wide roles of duplicate gene retention mechanisms, and use this tool to shed light on these mechanisms in the evolution of duplicate genes in grasses.

Chapter 1 Development of *CDROM*

1.1 Introduction

Gene duplication is thought to be a primary driving force in genome evolution, as known duplicates account for 17-65% of genes in organisms from all three domains of life (Moore 2003; Zhang 2003). In the simplest scenario, a gene duplication event produces two identical copies of an ancestral single-copy gene. The gene orthologous to the ancestral gene is referred to as the parent, and the product of the duplication event is referred to as the child. After duplication, it is thought that parent and child copies are functionally redundant, resulting in a period of relaxed purifying selection during which either the parent or child may accumulate mutations that may have been deleterious to the single-copy ancestral gene (Ohno 1970). Though this process typically leads to loss of function in the copy under relaxed selection (Lynch and Conery 2000), it also provides opportunities for functional divergence and evolutionary retention of the duplicate genes.

There are four major mechanisms that can result in long-term retention of duplicate genes (Figure 1). First, under conservation, the ancestral function is retained in both the parent and the child copy (Ohno 1970). Second, under neofunctionalization, the ancestral function is retained in one copy, and a new function emerges in the other copy (Ohno 1970). Third, under subfunctionalization, the ancestral function is divided between the parent and child copies (Hughes 1994, Force *et. al.* 1999, Stoltzfus 1999). Fourth, under specialization, both copies acquire new functions via the combined actions of subfunctionalization and neofunctionalization (He and Zhang 2005).

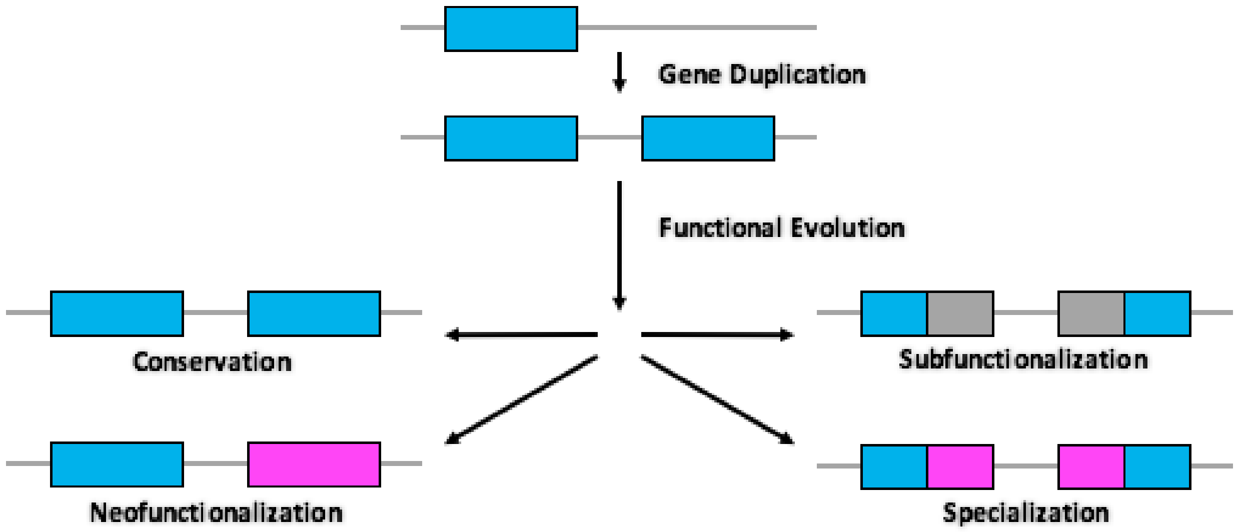

Figure 1. Duplicate Gene Retention Mechanisms

Figure 1. Gene duplication results in two identical gene copies, which undergo functional evolution and can be retained by preservation of the ancestral function in both copies (conservation), preservation of the ancestral function in one copy and acquisition of a new function in the other copy (neofunctionalization), division of the ancestral function between copies (subfunctionalization), or acquisition of a new function in each copy (specialization). Rectangles represent genes, and colors their functions (gray – no function, blue – ancestral function, pink – new function).

In 2013, Assis and Bachtrog developed a phylogenetic approach that is capable of classifying these duplicate gene retention mechanisms on a genome-wide scale. Their method uses the Euclidian distance between expression profiles of genes as a measure of their expression divergence, and expression divergence between genes as a proxy for their functional divergence. Application of their method requires gene expression data from parent and child genes that arose by duplication in one species, from ancestral genes that gave rise to these duplicates in a second species, and from single-copy genes that did not undergo duplication in both species. Euclidian distances are calculated between expression profiles of parent and ancestral genes ($E_{\mathrm{P,A}}$), between expression profiles of child and ancestral genes ($E_{\mathrm{C,A}}$), between combined parent-child gene expression profiles and ancestral gene expression profiles ($E_{\mathrm{P+C,A}}$), and between expression profiles of single-copy genes in the two species ($E_{\mathrm{S1,S2}}$). The distribution of $E_{\mathrm{S1,S2}}$ is used to establish a cutoff for expression divergence (E_{div}). For example, if $E_{\mathrm{P,A}} < E_{\mathrm{div}}$, then it is assumed that the ancestral function is retained in the parent. Using this reasoning, they applied the phylogenetic rules listed in Table 1 to classify the mechanisms retaining duplicate genes.

Table 1. Rules for Classifying Duplicate Gene Retention Mechanisms

Duplicate Gene Retention Mechanism	Phylogenetic Classification Rule
Conservation	$E_{\mathrm{P,A}} \leq E_{\mathrm{div}}$ and $E_{\mathrm{C,A}} \leq E_{\mathrm{div}}$
Neofunctionalization	$E_{\mathrm{P,A}} > E_{\mathrm{div}}$ and $E_{\mathrm{C,A}} \leq E_{\mathrm{div}}$ or $E_{\mathrm{P,A}} \leq E_{\mathrm{div}}$ and $E_{\mathrm{C,A}} > E_{\mathrm{div}}$
Subfunctionalization	$E_{\mathrm{P,A}} > E_{\mathrm{div}}$, $E_{\mathrm{C,A}} > E_{\mathrm{div}}$, and $E_{\mathrm{P+C,A}} \leq E_{\mathrm{div}}$
Specialization	$E_{\mathrm{P,A}} > E_{\mathrm{div}}$, $E_{\mathrm{C,A}} > E_{\mathrm{div}}$, and $E_{\mathrm{P+C,A}} > E_{\mathrm{div}}$

Table 1. Rules used by Assis and Bachtrog (2013) to classify the retention mechanism of each pair of duplicate genes.

Assis and Bachtrog's (2013) rules are based on expectations of the retention mechanisms. Under conservation, the ancestral function is retained in both the parent and child copy. Therefore, we expect that $E_{\mathrm{P,A}} \leq E_{\mathrm{div}}$ and $E_{\mathrm{C,A}} \leq E_{\mathrm{div}}$. With neofunctionalization, the ancestral function is preserved in one copy and a new function emerges in the other copy. Thus, we expect $E_{\mathrm{P,A}} > E_{\mathrm{div}}$ and $E_{\mathrm{C,A}} \leq E_{\mathrm{div}}$, if the new function emerges in the parent copy and $E_{\mathrm{P,A}} \leq E_{\mathrm{div}}$ and $E_{\mathrm{C,A}} > E_{\mathrm{div}}$

if the new function emerges in the child copy. In subfunctionalization, the ancestral function is divided between parent and child copies. As a result, we expect $E_{\mathrm{P,A}} > E_{\mathrm{div}}$ and $E_{\mathrm{C,A}} > E_{\mathrm{div}}$ because neither copy has the ancestral function, as well as $E_{\mathrm{P+C,A}} \leq E_{\mathrm{div}}$ because the functions of these copies together compose the ancestral function. Under specialization, a new function emerges in both the parent and child copies. Therefore, we expect $E_{\mathrm{P,A}} > E_{\mathrm{div}}$ and $E_{\mathrm{C,A}} > E_{\mathrm{div}}$, because neither copy has retained the ancestral function, as well as $E_{\mathrm{P+C,A}} > E_{\mathrm{div}}$ because the functions of the copies together would not compose the ancestral function.

Prior to the development of the R package described in this chapter, no software tools existed for classifying duplicate gene retention mechanisms. Additionally, the phylogenetic approach developed by Assis and Bachtrog (2013) has only been applied in *Drosophila* (Assis and Bachtrog 2013) and mammals (Assis and Bachtrog 2015). Yet, duplicate genes are abundant in many other species (Moore 2003; Zhang 2003), and availability of genomic data is rapidly increasing. Thus, the goal of this chapter was to provide users with a fast and simple tool for studying the mechanisms retaining duplicate genes in any species for which there is appropriate data available. To accomplish this goal, we implemented the phylogenetic classification method developed by Assis and Bachtrog (2013) as an R package *CDROM,* short for Classification of Duplicate gene RetentiOn Mechanisms (Perry and Assis 2016). R is a programming language that is freely available for download at www.r-project.org and widely used by researchers. *CDROM* can be downloaded from the CRAN package repository, which is where the majority of R packages are located, at https://CRAN.R-project.org/package=CDROM. One of the major advantages of *CDROM* is that it is easy to use, even if the user has minimal programming knowledge. Once downloaded, the user simply provides necessary input files, and all calculations and classifications are made by *CDROM* and output to the user.

1.2 Methods

CDROM requires three input files from the user. First, it requires a table in which each row contains names of a pair of parent and child genes in one species, as well as the name of the ancestral gene in a second species. Second, the user must provide a table in which each row contains the names of orthologous single-copy genes in the two species. Third, *CDROM* requires a table in which each row contains the name of a gene and measurements of any type of data that can be used as a proxy for function of that gene. This table should contain data from as many genes as possible, and ideally from most or all of the genes in the genomes of the two species. The most commonly available data that can be provided are gene expression levels obtained from microarray or RNA-seq studies. Measurements can be from a single sample or from multiple samples, such as different tissues, developmental time points, or experimental conditions. A greater number of samples will result in more reliable classifications by *CDROM*.

CDROM first checks to ensure that all the necessary input files have been provided by the user. If any files are missing, *CDROM* returns an error indicating which files are missing. Second, *CDROM* obtains the "functional" measurements for each gene mentioned in the input files. If any parent, child, or ancestral genes are missing functional data, the classification will be "unavailable". The user can then check to see which of the three gene names is missing expression values and attempt to correct the problem. Third, if the functional data are raw expression levels (indicated by the user as a parameter), *CDROM* converts the raw values to relative expression levels, as this has been shown to be more robust when computing Euclidian distances (Pereira 2009). Fourth, *CDROM* calculates all $E_{\mathrm{P,A}}$, $E_{\mathrm{C,A}}$, $E_{\mathrm{P+C,A}}$, and $E_{\mathrm{S1,S2}}$. Fifth, $E_{\mathrm{S1,S2}}$ values are used to set E_{div}. *CDROM* automatically uses the semi-interquartile range from the median as E_{div} because it is robust to distribution shape and outliers and was the cutoff used in previous analyses

(Assis and Bachtrog 2013, 2015). However, *CDROM* also allows the user to choose E_{div}. Last, *CDROM* uses the phylogenetic rules outlined in Table 1 to classify the retention mechanism of each pair of duplicate genes.

A limitation of Assis and Bachtrog's (2013) method is that it requires knowledge of the identities of parent and child copies. However, often directionality of a duplication event cannot be determined, which limits the potential usage of the method. Therefore, *CDROM* defaults to a setting that does not require this knowledge and instead simply refers to the first copy listed as "Duplicate 1" and the second copy listed as "Duplicate 2". This allows classifications to be made as outlined above, with the major difference being in the interpretation of neofunctionalization. In particular, when using the more general default method, the user cannot determine whether neofunctionalization is more prevalent in parent or child copies. As a result, the interpretation of classifications will be less specific.

1.3 Results

After all necessary calculations and classifications are made, *CDROM* outputs two tables and one figure. The first table provides the counts of retention mechanism classifications for five different E_{div} values (e.g., Table 2). Each row of this table contains the method used to obtain E_{div}, the value of E_{div}, and the number of each retention mechanism based on the E_{div} value. This table enables the user to evaluate the robustness of the classifications to E_{div}, and also aids the user in selecting an appropriate E_{div} value. The second table provides the retention mechanism classification for each duplicate gene pair (e.g., Table 3). Each row of this table contains names of parent, child, and ancestral genes, values of $E_{\mathrm{P,A}}$, $E_{\mathrm{C,A}}$, $E_{\mathrm{P+C,A}}$, and the resulting classification. The figure provides the distributions of Euclidean distances calculated and the position of the chosen E_{div} value (e.g., Figure 2). This figure provides a general picture of the functional divergence of the duplicates, and also enables the user to evaluate whether the selected E_{div} value is appropriate for the data. If the user chooses the default method with no knowledge of parent or child copies, the figure will combine the Euclidean distances for Duplicate 1 and Duplicate 2 ($E_{\mathrm{D1,A}}$ and $E_{\mathrm{D2,A}}$) into a single distribution (e.g., Figure 2A). Otherwise, the figure will display separate distributions for $E_{\mathrm{P,A}}$ and $E_{\mathrm{C,A}}$ (e.g., Figure 2B).

Table 2. Sample *CDROM* Output Table 1

E_div type	E_div Value	Cons	Neo	Sub	Spec
meanSD	0.615	45	4	1	4
mean2SD	0.833	52	2	0	0
medSIQR	0.460	32	13	1	8
medIQR	0.573	42	7	0	5
quant75	0.478	36	11	0	7

Table 2. In the first output table, *CDROM* provides counts of classifications (cons – conservation, neo – neofunctionalization, sub – subfunctionalization, spec – specialization) obtained with five E_{div} values (meanSD - one standard deviation from the mean, mean2SD - two standard deviations from the mean, medSIQR - semi-interquartile range from the median, medIQR – interquartile range from the median, quant75 - 75th percentile).

Table 3. Sample CDROM Output Table 2

Duplicate 1	Duplicate 2	Ancestral gene	$E_{D1,A}$	$E_{D2,A}$	$E_{D1+D2,A}$	Classification
BD36	BD37	OS82	0.35	0.57	0.40	Neo of D2
BD46	BD24	OS21	0.30	0.28	0.20	Cons
BD08	BD92	OS49	0.37	0.28	0.30	Cons

Table 3. In the second output table, *CDROM* provides classifications of retention mechanisms (neo of D2 – neofunctionalization of Duplicate 2, cons – conservation) for all duplicate gene pairs.

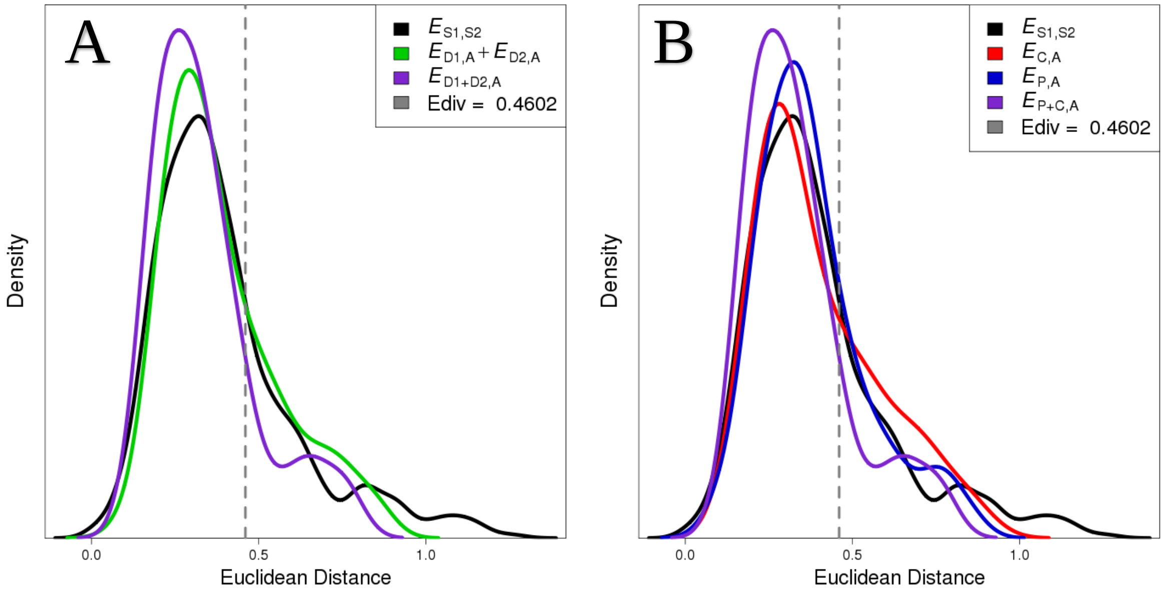

Figure 2. Sample *CDROM* Output Figure

Figure 2. *CDROM* outputs a figure that shows the distributions of all Euclidean distances that were calculated. Figure 2A shows the more generalized default method of *CDROM*, which assumes no knowledge of the parent child copies. Figure 2B shows the more specific method, which assumes knowledge of the parent and child copy.

Figure 2 can be used to illustrate the differences between interpretations made when parent and child copies are unknown (Figure 2A) or known (Figure 2B). In Figure 2A, the $E_{\mathrm{D1,A}} + E_{\mathrm{D2,A}}$ distribution is slightly to the left of the $E_{\mathrm{S1,S2}}$ distribution, indicating that a majority of the duplicate genes are retained by conservation. However, this result can also be reached from the table of duplicate gene classifications (Table 2). In contrast, Figure 2B shows that the $E_{\mathrm{P,A}}$distribution is shifted slightly to the right of the $E_{\mathrm{C,A}}$ distribution. This indicates that parent copies generally have greater functional divergence than child copies. Thus, the more specific method of *CDROM* allows enhanced interpretation of the outcomes.

1.4 Discussion

In the majority of species, the mechanisms driving the functional evolution of duplicate genes remains unclear. Assis and Bachtrog (2013) developed the first approach for classifying these mechanisms on a genome-wide scale and used this approach to classify retention mechanisms of *Drosophila* (Assis and Bachtrog 2013) and mammalian (Assis and Bachtrog 2015) duplicate genes. These studies revealed that duplicate genes in *Drosophila* diverge faster and appear to be primarily retained by neofunctionalization of the child copy, while mammalian duplicate genes diverge slower and are primarily retained by conservation. In this chapter, we developed *CDROM*, which implements this phylogenetic approach in a simple and flexible R package (Perry and Assis 2016). *CDROM* is freely available on the CRAN repository. It is also accessible to all researchers, even those with minimal programming knowledge. *CDROM* can be used in any species, even prokaryotes, provided that functional measurements of genes are available. Thus, researchers can easily apply *CDROM* to many different datasets leading to increased knowledge about the duplicate gene retention mechanisms in many diverse species.

Chapter 2 : Application of *CDROM* to Duplicate Genes in Grasses

2.1 Introduction

Recent evidence has shown that a major characteristic of plants is that gene families are highly conserved over long evolutionary timescales (Rensing *et. al.* 2008). Yet, among different species, there is a lot of variation in gene family size, diversity, and phenotype. An explanation may be that much of the observed diversity is a result of functional evolution of duplicates (Rensing *et. al.* 2008, Ming *et. al.* 2008). Therefore, gene duplication may play a central role in plant diversity because it generates new genes that can undergo adaptive evolution. It is speculated that no other group of organisms has a greater occurrence of duplication events than plants (Flagel 2009). One of the defining characteristics of most plants is polyploidy, consisting of more than two paired sets of chromosomes. In fact, almost all but the most recently formed plant gene families have experienced expansion through polyploidy (Flagel 2009). The polyploidy events are generally a result of a whole genome duplication, hence providing ample opportunity for the functional evolution of duplicate genes.

The Poaceae family of grasses consists of more than 10,000 plant species and is the fifth largest plant family. These grasses are monocots and angiosperms, or flowering plants. Grasses are very important for agriculture around the world, as they include important cereal crops, such as wheat, oat, maize, sugarcane, and rice (Davidson 2012). Prior to the divergence of the Poaceae family 56-70 million years ago, a whole genome duplication event occurred, resulting in many homologs in various species of grasses (Wang 2011).

Brachypodium distachyon is a member of the Poaceae family of grasses. It is native to southern Europe, northern Africa, and Southwest Asia. It has a genome size of 270 Mb and 25,532 genes. Although it has little agricultural significance, it is commonly used as a model organism when experimenting with temperate grasses (Brachypodium Initiative 2010). *Oryza sativa*, commonly known as rice, feeds over half the population in the world. It has a genome size of 380 Mb and contains 40,331 genes (International Rice Genome Project 2005). *Sorghum bicolor* is an African grass that is grown for food, feed, and fuel. It has potential use as a crop for biofuel. It has a genome size of 730 Mb and 34,497 genes (Paterson *et. al.* 2009). These species represent three of the major subfamilies of the grasses: the Pooideae, the Panicoideae, and the Ehrhartoideae. Their ancestry is depicted in Figure 3.

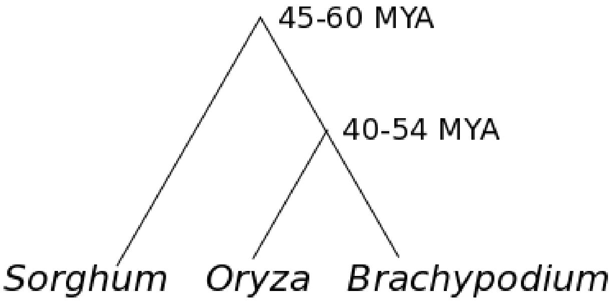

Figure 3. Phylogeny of Grasses

Figure 3. Presented is the phylogenetic history of the 3 grasses: *S. bicolor, O. sativa,* and *B. distachyon*. These three groups of grasses represent three subfamilies of the Poaceae. *O. sativa* and *B. distachyon's* last common ancestor was between 40-54 million years ago, and all three groups share a common ancestor approximately 45-60 million years ago (Davidson 2012).

Though plants have experienced many gene duplication events in their history, and these grasses are known to have undergone at least one whole genome duplication, little is known about how the duplicate genes are retained. In grasses, most research on duplicate genes has focused on the number that have been retained or lost since the whole genome duplication event. In *S. bicolor*, after the whole genome duplication, most duplicate genes lost one of their copies (Paterson *et. al.* 2009). However, of great interest is what happens to those duplicate genes that are retained. In this chapter, we perform the first genome-wide classification of the mechanisms that led to the retention of duplicate genes in grasses. In particular, we apply *CDROM* to duplicate genes from *B. distachyon, O. sativa,* and *S. bicolor*. These genomes have all recently been sequenced, and their duplicates genes and are well annotated. Moreover, Davidson *et. al.* (2012) used RNA-seq to measure genome-wide expression levels for nine tissues in all three species. Thus, these grasses contain an ideal dataset for application of *CDROM*.

2.2 Methods

Input data for *CDROM* was provided by Dr. Raquel Assis. The table of single-copy orthologs and the table of pairs of duplicate genes and their ancestral genes were created by parsing annotations from PLAZA version 2.5, a database for plant genomic information (Vandepoele 2013). The eight sets of triplets (the pair of duplicate genes in one species and the ancestral gene in a second species) analyzed are summarized in Table 4, and the pairings used for single-copy genes are shown in Table 5. The table of functional measurements of genes was constructed from the RNA-seq dataset of Davison *et al.* (2012), which consists of the FPKM (fragments per kilobase of transcript per million mapped reads) of each gene in the same nine tissues of *B. distachyon, O. sativa*, and *S. bicolor.* The nine tissues are anther, endosperm, leaf, pistil, embryo, seed five days after pollination, seed ten days after pollination, early inflorescence and the emerging inflorescence. A major advantage of this dataset is that all experiments were performed under identical conditions, which is ideal when comparing data from different species.

Table 4. Classes of Triplets used for Analysis

Class #	Species Duplication Event Occurred In	Ancestral Gene Sister Species
Class 1	*S. bicolor*	*B. distachyon*
Class 2	*S. bicolor*	*O. sativa*
Class 3	*O. sativa*	*B. distachyon*
Class 4	*O. sativa*	*S. bicolor*
Class 5	*B. distachyon*	*S. bicolor*
Class 6	*B. distachyon*	*O. sativa*
Class 7	Before *O. sativa* and *B. distachyon* Divergence (using *O. sativa* genes)	*S. bicolor*
Class 8	Before *O. sativa* and *B. distachyon* Divergence (using *B. distachyon* genes)	*S. bicolor*

Table 4. Eight classes of triplets were used for analysis of the three species *B. distachyon*, *O. sativa*, and *S. bicolor*. Each class requires a pair of duplicate genes in one species and an ancestral gene in a second species. Classes 1 and 2 consist of duplicate genes that arose in *S. bicolor* after its divergence from the other two species. Classes 3 and 4 consist of duplicate genes that arose in *O. sativa* after its divergence from *B. distachyon*. Classes 5 and 6 consist of duplicate genes that arose in *B. distachyon* after its divergence from *O. sativa*. Classes 7 and 8 consist of duplicate genes that arose in the common ancestor of *O. sativa* and *B. distachyon* after its divergence from *S. bicolor*, but before its divergence of *B. distachyon* and *O. sativa*. Thus, these duplicate genes appear in both *O. sativa* and *B. distachyon*. Class 7 uses the duplicate gene names from *O. sativa*, whereas class 8 uses the duplicate gene names from *B. distachyon*.

Table 5. Pairs of Single-Copy Orthologous Genes

Species 1	Species 2
S. bicolor	*O. sativa*
S. bicolor	*B. distachyon*
B. distachyon	*O. sativa*

Table 5. Three pairings of orthologous single-copy genes were used for different runs of *CDROM*. Group 1 contains the orthologous single-copy genes between *S. bicolor* and *O. sativa*, group 2 contains the orthologous single-copy genes between *S. bicolor* and *B. distachyon*, and group 3 contains the orthologous single-copy genes between *B. distachyon* and *O. sativa*.

After all necessary input data were obtained, *CDROM* was used to classify the retention mechanisms of duplicate genes that arose at each of the time points indicated in Table 4. For example, for Class 1, the input files *CDROM* used were the table containing the duplicate genes that arose in *S. bicolor* with their ancestral genes in *B. distachyon*, the table containing the orthologous single-copy genes between *S. bicolor* and *B. distachyon*, and the table of FPKM values for all nine tissues in both *S. bicolor* and *B. distachyon*. Because the identities of parent and child copies were ambiguous, the default method of *CDROM* was used for all runs. Also, the medSIQR, the semi-interquartile range from the median, was used as E_{div} for each class.

2.4 Results

Table 6. Counts of Duplicate Gene Retention Mechanisms in Grasses

Class	Cons	Neo	Sub	Spec
1	43 (75.44%)	11 (19.30%)	0 (0%)	3 (5.26%)
2	32 (58.18%)	16 (29.09%)	0 (0%)	7 (12.73%)
3	29 (53.70%)	17 (31.48%)	0 (0%)	8 (14.81%)
4	21 (58.33%)	10 (27.78%)	0 (0%)	5 (13.89%)
5	34 (60.71%)	9 (16.07%)	1 (1.79%)	12 (21.43%)
6	32 (59.26%)	13 (24.07%)	1 (1.85%)	8 (14.81%)
7	12 (57.14%)	7 (33.33%)	0 (0%)	2 (9.52%)
8	15 (62.50%)	5 (20.83%)	0 (0%)	4 (16.67%)
All	218 (61.06%)	88 (24.65%)	2 (0.56%)	49 (13.72%)

Table 6. Here, the counts and proportions of classifications (cons – conservation, neo – neofunctionalization, sub – subfunctionalization, spec – specialization) for each of the eight classes outlined in Table 4 are provided.

There are three main results from this study. First, there are consistent levels of the duplicate gene retention mechanism conservation across all of the duplicate genes studied in these three species of grasses, with the level of conservation varying between 53% - 75%. Second, on average, neofunctionalization appeared to be the next most common retention mechanism, with levels varying between 16% - 33%. The one exception was class 5, which consists of duplicate genes that emerged in *B. distachyon* using *S. bicolor* as the ancestral copy, which showed a greater proportion of specialization (21.43%) than neofunctionalization (16.07%). Third, the most interesting result may be the lack of evidence supporting subfunctionalization. Only two duplicate genes that emerged in *B. distachyon* showed evidence for subfunctionalization in this analysis.

For each class, *CDROM* outputs the classifications based on five standard E_{div} values (e.g. Table 2). This table provides insight into the robustness of classifications to E_{div}. For the majority of the eight classes of duplicates, varying E_{div} resulted in either zero or one classification of subfunctionalization. The only exception was class 8, where setting E_{div} to medIQR resulted in two subfunctionalization classifications. This analysis indicates that our classifications are robust to E_{div} and also suggests that there may be a lack of subfunctionalization in duplicate genes in grasses.

Unfortunately, there were some duplicate gene pairs that were unable to be classified as a result of missing expression data. There were three pairs of duplicates that were unable to be classified from class 1, five from class 2, 27 from class 3, 23 from class 4, two from class 6, and three from class 7. In total, there were 63 pairs that were unable to be classified, representing approximately 15% of the initial dataset.

2.5 Discussion

A defining characteristic of *S. bicolor*, *O. sativa*, and *B. distachyon* are the whole genome duplication events that have taken place in their history. Polyploidy may have a significant influence on plant diversity, as it enables genes to evolve under relaxed selective constraint, potentially resulting in their functional evolution (Wendel 2000). Following polyploid events, each duplicate gene has the potential to be retained in the genome by conservation, neofunctionalization, subfunctionalization, or specialization. In *Arabidopsis*, which is hypothesized to have undergone two or three whole genome duplication events, it was shown that significant levels of subfunctionalization and neofunctionalization contribute to the diversity of duplicate genes (Duarte 2006). Roulin *et. al.* (2012) found that the approximately 50% of gene duplicates in soybean, which underwent two whole genome duplication events, show evidence of conservation. The other 50% of duplicate genes show differential expression between the two copies of duplicate genes. This indicates that there may be significant levels of neofunctionalization, as well as the other duplicate gene retention mechanisms which result in functional evolution. It has also been demonstrated that there are significant levels of subfunctionalization in both soybean and cotton (Chaudhary 2009, Roulin 2012).

Thus, our results support previous research in plants in that the mechanism of conservation appears to be the primary mechanism retaining duplicate genes in grasses. However, one of the most interesting results from this study was that only two pairs of duplicates showed evidence of subfunctionalization. Previous studies have shown that other angiosperms show significant levels of subfunctionalization (Chaudhary 2009, Roulin 2012). One observation is that while grasses are examples of angiosperms, grasses are monocots. In contrast, Chaudhary (2009) and Roulin (2012) studied dicots. This indicates that retention mechanisms of duplicate genes may vary across different plant clades.

One of the major limitations of this study was the absence of expression data for 15% of duplicate gene pairs, as these pairs were unable to be classified. A major problem in bioinformatics research is the conversion of gene IDs. Gene IDs are constantly updated and reorganized. Different databases will use different gene IDs and, unfortunately, universal conversion tools are not widely available. As the expression data and duplicate gene IDs were downloaded from separate databases, the issue of different gene IDs occurred as many of the gene IDs found in the duplicate gene files were missing in the corresponding expression files. Future studies may be needed to specifically address this issue, such as using the location of the sequence in the genome to determine the corresponding gene ID in a different database.

Another limitation in this study is that only expression data was used as a proxy for functional divergence of the duplicate genes. Expression data only represents one of the many facets of gene function; other examples include protein-protein interactions and gene knockout effect data. Currently, gene expression data is the only type available on a genome-wide scale. However, *CDROM* can easily be applied to any quantitative trait that provides a measure of gene function. The results of this study could be further supported by applying *CDROM* to other quantitative traits which represent other facets of gene function. As future measurements of gene function become available on genome-wide scales, *CDROM* can be used to study this data and provide more holistic classifications of functional divergence for duplicate genes.

There are many other future directions for this study. First, knowledge of the age of the duplicates (Table 4) can allow for further refinement of the conclusions, in particular whether retention mechanisms differ over time. Also, another interesting direction may be to separate duplicate genes into different categories based on which tissue they are expressed most highly in. An interesting hypothesis to test is whether duplicate genes that are primarily expressed in reproductive organs tend to evolve new functions, as has been found in animals (Assis and Bachtrog 2013, 2015).

Last, this chapter shows how *CDROM* can be applied in species with sequenced genomes and an available quantitative measure of gene function. With the growing number of species with sequenced genomes, expression data, and techniques to identify homologous genes, *CDROM* has potential future usage in many species in which duplicate gene retention mechanisms have not been studied before. Here, *CDROM* was used to show that subfunctionalization may not play a significant role in the retention of duplicate genes in grasses, which is a result of great interest that can be further tested and refined in future studies.

Conclusions

Assis and Bachtrog (2013) developed the first approach for classifying duplicate gene retention mechanisms on a genome-wide scale and applied this method to data from *Drosophila* (Assis and Bachtrog 2013) and mammals (Assis and Bachtrog 2015). However, in a majority of species, the mechanisms driving the functional evolution of duplicate genes still remain unclear. To enhance knowledge of the prevalence of these mechanisms, we developed software that can quickly and easily use the approach of Assis and Bachtrog (2013) to classify duplicate gene retention mechanisms in a variety of species and datasets, and we demonstrate its utility by applying it to a novel dataset.

Chapter one introduces *CDROM*, which implements the Assis and Bachtrog (2013) phylogenetic approach as a R package (Perry and Assis 2016). *CDROM* can be freely downloaded from the CRAN repository, making it accessible to all researchers. It can also be used in any species provided that functional measurements of genes are available. Thus, researchers can easily apply *CDROM* to many different datasets, leading to increased knowledge about the duplicate gene retention mechanisms in many diverse species.

Chapter two applies *CDROM* to duplicate genes in grasses, for which levels of retention mechanisms had not previously been studied. Though we uncover several retention mechanisms, our results suggest that subfunctionalization may not play as significant of a role as hypothesized. Thus, this application yields novel insights about the evolution of duplicate genes in plants, as well as demonstrating the utility of *CDROM* in many species in which duplicate gene retention mechanisms have not been studied before.

BIBLIOGRAPHY

Assis, R., Bachtrog, D. (2013). Neofunctionalization of young duplicate genes in *Drosophila*. *Proc. Natl. Acad. Sci. USA*, 110, 17409–17414.

Assis, R., Bachtrog, D. (2015). Rapid divergence and diversification of mammalian duplicate gene functions. *BMC Evol. Biol,* 15, 138.

C. elegans sequencing consortium. (1998). Genome Sequence of the nematode C. elegans: A platform for investigating biology. *Science.* 282(5396) 2012-8.

Chaudhary, B., Flagel, L., Stupar, R.M., Udall, J.A., Verma, N. Springer, N.M, Wendel, J.F. (2009). Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (grossypium). *Genetics*. 182: 503-517.

Davidson *et. al*. (2012). Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. *The Plant Journal.* 71(3): 492-502.

Dissecting Plant Genomes with the PLAZA Comparative Genomics Platform. *Plant Physiology.* 158: 590-600.

Duarte *et. al.* Expression Pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of *Arabidopsis*. *Mol. Biol. Evol.* 23(2): 469-478.

Flagel, L. E., Wendel, J. F. (2009). Gene duplication and evolutionary novelty in plants. *New Phytologist*. 183(3): 557-564.

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

Fuller, J. C., Khoueiry, P., Dinkel, H., Forslund, K., Stamatakis, A., Barry, J., Budd, A. Soldatos, T. G., Linssen, K., Rajput, A. M. (2013). Biggest Challenges in Bioinformatics. *EMBO Reports*, 14(4), 302-304.

He, X., Zhang, J. (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, 169, 1157–1164.

Hughes, A.L. (1994). The evolution of functionally novel proteins after gene duplication. *Proc. Royal Soc. B*, 256, 119–124.

International Rice Genome Sequencing Project. (2005). The map-based sequence of the rice genome. *Nature*. 436: 793-800.

Lynch, M., Conery, J.S. (2000). The Evolutionary Fate and consequences of duplicate genes. *Science*. 290(5494): 1151 – 1155.

Ming *et. al*. (2008). The draft genome of transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature*. 452: 991-996.

Moore, R. C., Purugganan, M. D. (2003). The Early Stages of Duplicate Gene Evolution. *PNAS*, 100, 15682-15687.

Paterson, A.H., Bowers, J.E., Bruggmann, R. *et. al.* (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature*. 457: 551-556.

Pereira, V., Waxman, D., Eyre-Walker, A. (2009). A problem with the correlation coefficient as a measure of gene expression divergence. *Genetics*. 183(4): 1597-1600.

Perry B. R., Assis R. (2016). CDROM: Classification of Duplicate Gene Retention Mechanisms. *BMC Evolutionary Biology*. 16:82.

Ohno, S. (1970). Evolution by gene duplication. *Springer-Verlag*, Berlin.

Rensing *et. al.* (2008). The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science*. 319: 64-69.

Roulin *et. al.* (2012). The fate of duplicated genes in a polyploid plant genome. *The Plant Journal*. 73(1): 143-153.

Stoltzfus, A. (1999). On the possibility of constructive neutral evolution. *J Mol. Evol*. 49, 169 181.

The International Brachypodium Initiative. (2010). Genome Sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*. 463: 763-768.

Vandepoele, K., Van Bel, M. *et. al.* (2013). Pico-PLAZA, a Genome database of microbial photosynthetic eukaryotes. *Environmental Microbiology*. 15(8): 2147 – 2153.

Wang, X., Tang, H., Paterson, A. H. (2011). Seventy million years of concerted evolution of a homologous chromosome pair, in parallel, in major Poaceae lineages. *Plant Cell*. 23: 27-37.

Wendel, J.F. (2000). Genome Evolution in polyploids. *Plant Mol. Biol.* 42: 225 – 249.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology and Evolution*. 18(6): 292 – 298.

Academic Vita

BRENT R. PERRY

bperry2328@psu.edu
(724)-777-2530

EDUCATION

The Pennsylvania State University, University Park, PA May 2017

Schreyer Honors College
Bachelor of Science in Biology – Vertebrate Physiology Option
Minor in Global Health

RESEARCH EXPERIENCE

Undergraduate Research Assistant Fall 2014 - Present

Assis Research Group, University Park, PA

- Studied the functional evolution of duplicate genes in animals
- Investigated non-allelic gene conversion mechanisms
- Analyzed genomic data with the programming languages Perl and R
- Published software that classifies duplicate gene retention mechanisms
- Attended and presented at Penn State Bioinformatics research meetings
- Presented in the Eberly College of Science Fall 2016 Undergraduate Poster Exhibition

PUBLICATIONS

Perry B. R. (2017). The Evolution of Duplicate Gene Retention Mechanisms. Undergraduate Honors Thesis, Penn State. *In preparation*.

Perry B. R., Assis R. (2016). CDROM: Classification of Duplicate Gene Retention Mechanisms. *BMC Evolutionary Biology*. 16:82.

CLINICAL EXPERIENCE

Global Health Minor Field Work Summer 2016

Hospital of M'Bour, M'Bour, Senegal

- Observed Senegalese medical students for six-weeks
- Shadowed in the surgical, emergency, pediatric, and HIV departments
- Prepared and presented a hygiene assessment of the emergency department
- in coordination with the hospital staff

Heart Center Volunteer Summer 2014 - 2015

Trinity Health System, Steubenville, OH

- Shadowed physicians and nurses during day to day activities
- Transported patients and patients' families around the hospital

LEADERSHIP EXPERIENCE

Chemistry 112: Inorganic Chemistry Tutor Fall 2014 - Present

LionTutors, State College, PA

- Taught exam reviews for groups of 30-40 students
- Worked approximately 10 hours a week

Biology 230: Molecular Biology Lecture Assistant Fall 2015

Penn State, University Park, PA

- Encouraged active engagement in the classroom
- Guided weekly office hours and graded assignments

Sociology 119: Race and Ethnic Relations Discussion Leader Spring 2014

Penn State, University Park, PA

- Trained to facilitate dialogues on topics such as racism and inequality
- Directed multiple 15 student discussions each week

Biology 110: Introduction to Biology Peer Leader Fall 2014

Penn State, University Park, PA

- Explained and discussed relevant course material with students
- Tutored one class of 10 students a week

EXTRACURRICLAR ACTIVITIES

Weer Africa Volunteer Summer 2016

M'Bour, Senegal

- Organized supplies for rural medical clinics and diabetes screening days
- Directed individuals to appropriate medical evaluation station

Phi Gamma Nu Professional Business Fraternity Fall 2014 - Present

Penn State, University Park, PA

- Attended regular meetings, philanthropy events, and professional development sessions

AWARDS AND GRANTS

Eberly College of Science Travel Grant Summer 2016

- Obtained to support field work experience in Senegal

Schreyer Travel Grant Summer 2016

- Obtained to support field work experience in Senegal

Undergraduate Research Grant Summer 2015

- Obtained to support summer research with Assis Research Group

Edward C. Hammond Jr. Memorial Scholarship Spring 2015

- Outstanding accomplishment in the Eberly College of Science

Presidents Freshman Award Spring 2014

- 4.00 cumulative GPA first semester