CHARACTERIZATION OF THE WHOLE-GENOME DUPLICATION IN *POTAMOPYRGUS ANTIPODARUM*

By
Angie Kalwies

A thesis submitted in partial fulfillment of the requirements for graduation

with Honors in the

Department of Biology

Dr. John Logsdon

Thesis Mentor

Spring 2020

All requirements for graduation with Honors in

the

Department of Biology

have been completed.

Lori Adams, PhD.

Biology Honors Advisor.

Abstract

Whole-genome duplications (WGDs) in animals and the mechanisms which drive them are not well understood. Previous studies have focused primarily on characterizing ancient WGDs, due to the relative rarity of recent WGDs, however, there is a unique opportunity to study a likely young whole-genome duplication in the snail *Potamopyrgus antipodarum*. Here, I outline the collection of data supporting the existence of this recent WGD and the development of a promising method for detecting and characterizing duplicated genes. I annotated 20 predicted single-copy genes and classified them by the presence of nucleotide insertions or deletions (indels) as a means to detect divergence between copies. I describe a series of methods which utilize bioinformatic and evolutionary expertise. I found that the pairwise distances between duplicate genes do not statistically differ. However, due to high variance among duplicate averages, I found that in order to detect divergence between copies, they must be normalized to each other instead of compared in two groups as averages. Upon normalizing the copies to each other, I found that a majority of copies that contain indels were accumulating more new mutations than their intact counterparts, which potentially indicates the early steps of process of pseudogenization. Confirming the existence of recent WGD in in *P. antipodarum* and initiating these analyses of its consequences on the genome opens the door to future studies to investigate the pattern of divergence among all types of duplicate genes and to study the effects of large-scale genomic changes in various functional gene groups.

Acknowledgements

I would like to express my deepest appreciation to my advisor Dr. John Logsdon and my graduate student mentor Joseph Jalinsky. I am very grateful for the invaluable insights they provided during my time researching in the lab. When I became discouraged due to new data interfering with my initial project, they encouraged me to formulate new ideas and expand my knowledge of bioinformatics and evolutionary biology. The completion of this thesis would not have been possible without their strong support over the last three years.

Table of Contents

Introduction……....1

Materials and Methods……..3

Obtaining Candidate Gens for Duplication Analysis……...3

Gene Set Curation……3

Identification and Annotation of Duplicate Genes……..4

Divergence Data Collection……5

Classification of Duplicate Genes……....5

Results and Discussion……...6

Predicted Single-Copy Genes……..6

Divergence Between Duplicate Genes……8

Overall Pairwise Distances Between Gene Copies……10

Pairwise Differences By Codon Position……...11

Gene Copy Normalization……15

Additional Exploratory Analysis: Introns……..17

Conclusions……...17

Supplementary Tables……19

Supplementary Table 1……..19

Supplementary Table 2……..19

Supplementary Table 3……..20

Supplementary Table 4……..20

Supplementary Table 5……..21

Supplementary Figures……...22

Supplementary Figure……22

Literature Cited……...23

Introduction

Whole-genome duplication (WGD) is an evolutionary phenomenon whereby the entire genetic make-up of an organism is doubled. WGDs have played a very important role in shaping the evolution of today's living organisms and are a vital source of genetic diversity because the parts of the genome which were previously under strong selection can now evolve freely as there is another copy to fulfill the necessary function (Crow, 2006). Specifically, vertebrate evolution is characterized by two rounds of ancient WGDs that gave rise to the complexity we see today (Dehal & Boore, 2005). There is a plethora of evidence as to how ancient and recent WGDs contribute to growing diversity in many plant species (del Pozo, 2015). However, there is a strikingly small amount of evidence for recent WGD in animals and therefore the mechanisms by which their genomes respond to a WGD are largely unknown. In characterizing an ancient WGD, we typically only see the end of the process. This is mostly due to rapid gene loss following duplication, as most of the new genetic material accumulates deleterious mutations before beneficial ones due to the lack of selective constraint on half of the existing material. (Dehal & Boore, 2005).

Following a WGD, at least one copy of every gene is expected to remain intact as it is vital to proper functioning, whereas the duplicate copy may be free to evolve without constraint. Furthermore, since it is important for one copy to retain the original function, the other can accumulate mutations which could potentially lead to a variety of outcomes such as pseudogenization, sub-functionalization or neofunctionalization (Force *et al.*, 1999). Deleterious mutations can accumulate even in duplicated genes that are functionally important because there remains a functional copy elsewhere in the genome. In the event of pseudogenization, mutations such as frameshifts caused by insertions or deletions interrupt the coding sequence of a gene

causing it to no longer be functional. Under this scenario, over very long periods of time, pseudogenes eventually become unrecognizable compared to their original twin as they accumulate mutations. Subfunctionalization, for example, can occur when a gene performs two functions, but after a duplication, each copy of the gene may perform one of those functions. Finally, neofunctionalization can occur when a gene develops a completely new function through accumulation of beneficial mutations.

Potamopyrgus antipodarum, a fresh water lake- and river-dwelling snail native to New Zealand, may present a unique opportunity to study a recent WGD. Previous analysis shows that *P. antipodarum*'s genome is twice the size of closely related species belonging to the same genus, *Potamopyrgus estuarinus* (Logsdon & Nieman Labs, personal communication). To investigate the possibility of a whole-genome duplication in *P. antipodarum*, I used the high-quality genome assembly of *P. antipodarum* to test for the presence of duplicated genes then analyze those genes and potentially detect early pseudogenization by extracting and comparing genes to their orthologs in *P. estuarinus* which represents a pre-duplication state. As a result, we may be able to gain insight into how many and which types of genes are subject to different types of selection following WGD and the evolutionary fate of duplicates of those genes.

In this study, I analyzed a set of 20 genes that are predicted to exist in single-copy but are instead found in duplicate in *P. antipodarum*. I found that a majority of the genes that are predicted to exist exclusively in single-copy exist commonly as duplicates or higher multiples in *P. antipodarum*. Of the gene duplicates used in this study, genes with insertions or deletions (indels), that results in a reading frameshift, showed trends indicating an excess accumulation of new mutations, which could point to an early process of pseudogenization.

Materials and Methods

Obtaining Candidate Genes for Duplication Analysis

The genome of *P. antipodarum* (v 1.0) was sequenced and assembled in such a way that both copies of the duplicated genome are maintained in sequence (un-collapsed) as opposed to collapsing the genome where each loci is only represented once (collapsed), so that duplicate genes are able to be individually extracted rather than being collapsed into a single contiguous sequence (Logsdon & Nieman Labs, personal communication). To search for the presence of duplicated genes, I obtained a list of Benchmarking Universal Single-Copy Orthologs (BUSCOs) from the BUSCOv3 application (Waterhouse *et al*., 2017). I used these genes as a queries to the *P. antipodarum* genome assembly since, by definition, these genes are predicted to be retained in single-copy across almost all Metazoa. BUSCO is a program that can produce an estimation of the single-copy genes in a genome using data from many other organisms (Waterhouse *et al.*, 1990). In BUSCO, a metazoan dataset was used to find versions of the highly conserved genes and indicate the sequences for those genes in *P. antipodarum* (Waterhouse *et al.*, 2017). Next, I used the predicted single-copy genes as a query against the entire genome using the basic local alignment search tool (BLAST) (Altschul *et al.*, 1990). I also counted the number of hits for each BUSCO gene, which indicates the number of copies of each gene in the *P. antipodarum* genome.

Gene Set Curation

The BUSCO gene queries of the *P. antipodaru*m genome returned scaffold numbers and coordinates for the genes within their corresponding scaffold (Altschul *et al*., 1990). Next, I searched for the predicted single-copy genes in *P. antipodarum*'s previously assembled transcriptome to obtain a sequence which was consistent with functional conservation using

BLASTn, which uses a nucleotide query to search a nucleotide database (Altschul *et al.*, 1990). The transcriptome is a genomic library made up of messenger RNA; all non-coding regions are spliced out, resulting in expressed gene sequences exclusively. For the final gene set, *P. antipodarum* genes were selected that matched above 85% identity and 90% coverage to the queried predicted single-copy genes. Selecting the genes in this way allowed for higher annotation confidence.

Identification and Annotation of Duplicate Genes

In order to be counted as duplicates for analysis, the chosen genes must only return two results with highly similar BLAST E-values to each other. For expediency and consistency, queries which returned more than two hits were classified by number of returned hits regardless of the quality of subsequent hits. I mapped the sequences which passed the quality control assessment *via* transcriptome comparison to the scaffolds and annotated the exon and intron regions using the annotation software Sequencher v5.1 ®. This is possible because the transcriptome only includes coding regions since it is made of mRNA sequence; thus, it will only map to exonic regions. I searched for and annotated the same gene in the genome of the closely related species, *P. estuarinus* v0.9 using Sequencher 5.1® to use as an ancestral state comparison with the gene duplicates (Logsdon & Nieman Labs, PC). The genome of *P. estuarinus* is not well assembled and annotated at this time. Therefore, I manually annotated the target genes in *P. estuarinus* by mapping Hi-seq reads to one of the duplicate copies from *P. antipodarum*. I employed the same method to annotate three representative introns from *P. estuarinus* to compare to the same introns in *P. antipodarum* for the purpose of evaluating a neutral rate of evolution.

Divergence Data Collection

To begin sequence analysis, I entered and aligned the duplicate *P. antipodarum* sequences and orthologous *P. estuarinus* sequence in MEGA using Clustal (which appeared to perform better than MUSCLE) (Kumar *et al*., 2018). Given the frequent occurrence of indels of nucleotides present in the *P. antipodarum* sequences, a stricter alignment more accurately aligned the indels I knew to be present from my manual annotation of the genes. Next, I used MEGA to calculate the total difference and pairwise distance for the aligned duplicate genes in *P. antipodarum* and the ortholog in *P. estuarinus* (Kumar *et al*., 2018). The total difference is the amount of unique nucleotide differences in each sequence. The pairwise distance is the total differences divided by the length of the gene. I obtained total difference measurements and pairwise distance calculations which include 1^{st} site, 2^{nd} site, 3^{rd} sites (positions within amino-acid encoding triplet codons), 1^{st}+2^{nd} sites, and 1^{st}+2^{nd}+3^{rd} sites. Because 3^{rd} site mutations infrequently result in an amino acid change (*i.e.* synonymous mutation), it was used an estimation of neutral mutation rate. In addition to calculating 3^{rd} site pairwise distances, I used the annotated introns as an additional measure of neutral evolution. Because introns are not protein-coding, changes in their sequence are typically tolerated, which makes them a potentially good model comparison for pseudogenizing genes. I used these intron data in comparison with the duplicate copies in *P. antipodarum* to indicate whether a gene copy is experiencing mutation levels consistent with relaxed selection.

Classification of Duplicate Genes

After establishing the presence of duplicate BUSCO genes, I manually scanned the duplicate *P. antipodarum* gene pair while they were aligned to *P. estuarinus* and recorded the number of

insertions and deletions. If a particular nucleotide was missing in one *P. antipodarum* copy, but present in the other copy as well as in *P. estuarinus*, I labeled it a deletion. If there was an additional base in one copy which is not present in the other copy or *P. estuarinus*, I labeled it an insertion. In the event that both *P. antipodarum* copies exhibited an insertion or deletion in the same position compared to *P. estuarinus*, I assumed *P. estuarinus* to be an estimate of the ancestral (pre-duplication) state of the gene; the presence of an indel in both *P. antipodarum* copies was treated the same as if only one contained an indel by comparing to *P. estuarinus*. Next, I entered all sequences into ExPasy to convert them to protein coding form (Gasteiger *et al*., 2003). If the protein-coding sequence of the copy with an indel was interrupted, it was grouped as "indel-containing" with the assumption of it being non-functional and if the copy was uninterrupted, it was grouped as "intact". The non-indel-containing copies which contained no other interruptions were also labeled as "intact". Then, I performed multiple statistical analyses based on data obtained by 3-way alignments and aforementioned classification system using RStudio (RStudio Team, 2015). For each duplicate gene set, I compared aligned nucleotides. I also normalized each gene copy to each other by subtracting the intact copy pairwise distance from the indel-containing copy pairwise distance for each gene.

Results and Discussion

Predicted Single-Copy Genes

BUSCO returned 664 predicted single-copy genes for *P. antipodarum*. When BLASTing the predicted single-copy genes against the *P. antipodarum* genome assembly, I found that 19 of these genes returned no hits or could not be located in the genome, 171 genes were found only

once, 209 genes were found twice, 127 genes were found three times, 61 genes were found four times, and 77 genes were found more than four times (Figure 1). The distribution of predicted

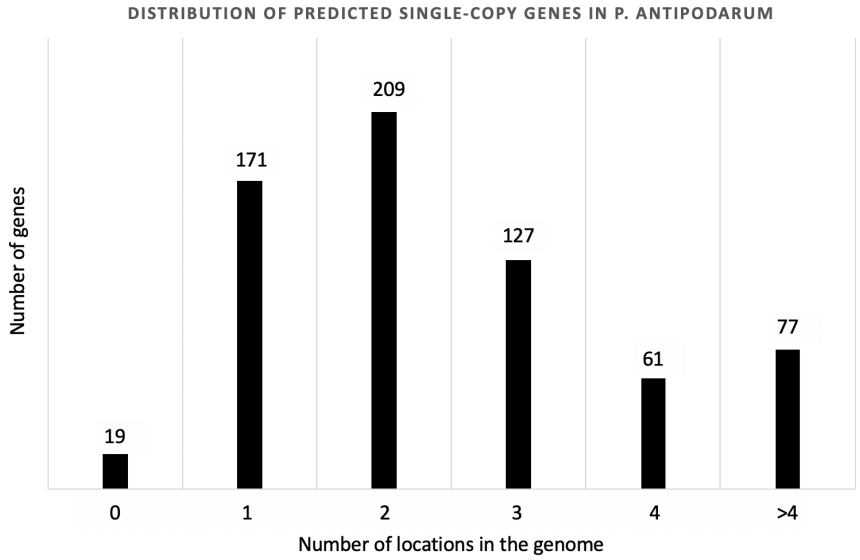

Figure 1. The distribution of 664 predicted single-copy genes showing the number of genes which exist in 0 to more than 4 locations in the genome. The x-axis is the number of "hits" returned upon querying predicted single-copy genes against the P. antipodarum *genome using BLAST. The y-axis is the number of genes which exist X times in the genome.*

single-copy genes existing heavily of duplicates and higher multiples supports the existence of an evolutionary recent WGD because, by nature, these genes are predicted to exist in single-copy throughout eukaryotic life (Simão *et al*, 2015). However, the duplication is not so recent that all genes are present as duplicates. Interestingly, I expected all predicted single-copy genes, with little exception, to fall in the category of zero, one, or two locations within the genome if there was a WGD. The predicted single-copy genes found in more than two locations is not expected and may be due to errors in the genome assembly. For expediency and consistency, I did not consider the quality of the sequence match as much as the overall number of matching sequences. Were I to re-evaluate the sorting criteria used to distinguish the quality of the sequence matches, I would likely find that a many of the genes found in more than two locations could actually be assigned to the duplicate category. At the time of my analysis, I was concerned

about potentially not identifying the true duplicates; thus, I used less strict criteria due to the uncertainty to detect gene duplicates with this method.

Divergence Between Duplicate Genes

I annotated 20 randomly selected sets of predicted single-copy genes; each set consisted of two versions of a *P. antipodarum* gene and the corresponding ortholog in *P. estuarinus* for a total of 60 genes. Among these gene sets, 15 *P. antipodarum* genes contained a frameshift while their paralogs were intact and uninterrupted (Table 1). While two of the intact versions did contain indels, genes 6 and 11, those indels are multiples of three nucleotides, therefore they should not interrupt the reading frame and can be considered intact (*i.e.* has an open reading frame) for the purpose of my analysis (Table 1). The number of indels within the frameshifted copies varied from a single base pair up to eight separate base pair deletions (Table 1). An example 3-way alignment of *P. estuarinus* and the corresponding *P. antipodarum* duplicates for Gene 5 is provided in Supplementary Figure 1.

Table 1. Summary of indels and gene length for P. antipodarum *duplicates. The BUSCO reference ID is the randomly assigned name of the predicted single-copy gene. The numbering of genes is arbitrary for analytical purposes.*

		indel copy		intact copy	
BUSCO reference	Gene	insertion	deletion	insertion	deletion
6Y3	1	-	1 bp	-	-
7Z5	2	-	1 bp	-	-
5OW	3	-	1 bp x 2	-	-
59G	4	-	1 bp x 2	-	-
BAG	5	-	1 bp	-	-
5D7	6	-	1 bp	-	-
6CB	7	-	1 bp	-	9 bp
7DR	8	-	1 bp	-	-
78D	9	-	1 bp x 2	-	-
GWF	10	-	1 bp	-	-
HZK	11	-	1 bp, 3 bp	-	-
2NO	12	1 bp	1 bp x 2	-	3 bp
C8H	13	2 bp		-	-
CNT	14	1 bp	1 bp x 4	-	-
CTU	15	1 bp x 3	1 bp x 8, 2 bp	-	-
3ZS	16	-	-	-	-
AYM	17	-	-	-	-
2P2	18	-	-	-	-
60M	19	-	-	-	-
7N5	20	-	-	-	-

Table 2. Summary of differences by site in P. antipodarum *duplicates and length of each sequence. Table values were obtained using MEGA (Kumar et al, 2018). The BUSCO reference ID is the randomly assigned name of the predicted single-copy gene. The numbering of genes is arbitrary for analytical purposes*

		indel copy				intact copy			
BUSCO reference	Gene	1st + 2nd	3rd	Total Distance	Gene length (nt)	1st + 2nd	3rd	Total Distance	Gene length (nt
6Y3	1	1	21	22	2198	2	20	22	2199
7Z5	2	2	6	8	1340	2	11	13	1341
5OW	3	2	7	9	1500	2	5	7	1502
59G	4	0	18	18	1522	0	24	24	1524
BAG	5	2	3	5	917	1	2	3	918
5D7	6	7	16	23	2255	5	16	21	2247
6CB	7	0	2	2	1382	1	2	3	1383
7DR	8	2	17	19	2453	2	18	20	2454
78D	9	4	6	10	1330	4	6	10	1332
GWF	10	0	4	4	1085	0	4	4	1086
HZK	11	10	15	25	1877	10	15	25	1878
2NO	12	0	9	9	2504	12	3	15	2505
C8H	13	11	15	26	1199	10	15	25	1197
CNT	14	2	4	6	984	2	3	5	987
CTU	15	10	9	19	1346	10	9	19	1353
3ZS	16	1	23	24	1650	1	23	24	1650
AYM	17	1	3	4	1140	1	3	4	1140
2P2	18	1	33	34	2292	4	21	25	2292
60M	19	4	24	28	1389	4	24	28	1389
7N5	20	2	5	7	1221	2	5	7	1221

Of the 20 gene sets, five were identical with respect to indels (Table 1). This gene set, while admittedly small, exemplifies what a recent WGD might look like: I consistently observed a majority of the *P. antipodarum* genes having one copy which was intact and the other accumulating frameshifts. A smaller proportion of gene pairs were identical in this regard. When considering the evolutionary fate of genes, I predicted that the 15 genes with frameshifts are likely heading toward pseudogenization while the five identical copies are predicted to be maintained and have the potential for either subfunctionalization or neofunctionalization (but also pseudogenization). The total analysis consisted of 38 indels found in 63,182 nucleotides across 40 sequences (Table 1 & 2). I analyzed an additional 31,609 nucleotides for the 20 corresponding *P. estuarinus* genes. Every gene set, with the exception of four completely identical *P. antipodarum* duplicates, contained unique differences (Table 2). Gene 18 is a notable outlier in the indel or no-indel classification system as they are not identical but contain no

frameshifts. The fact that four out of the five genes which contain no indels are completely identical emphasizes the utility of indels as a divergence marker. Overall, a majority of the duplicates are interrupted by indels rather than nonsense or missense changes resulting from nucleotide substitutions

Overall Pairwise Distances Between Gene Copies

I compared the pairwise distances for each *P. antipodarum* gene *versus* its P. estuarinus ortholog as a means of estimating divergence. When considering the pairwise distance including all sites, they do not differ on average between copies; the average pairwise distances are 0.0089 and 0.0091 for the intact and indel-containing copies, respectively, and p=0.9242 (Figure 2), indicating no statistical difference between groups. One explanation of this is that the WGD in *P. antipodarum* happened so recently in evolutionary history that extensive divergence is not detectable between gene copies. However, the calculated pairwise distance for comparing the copies to themselves displays high variability between copies (Figure 2, right). While indels are used as the identifying factor for mutation accumulation and therefore potential pseudogenization, frameshifts are not the only nucleotide change that may result in disruption of a coding sequence. However, frameshifts are easily identifiable and almost certain to disrupt gene function, making them good candidate identifiers for duplication analysis.

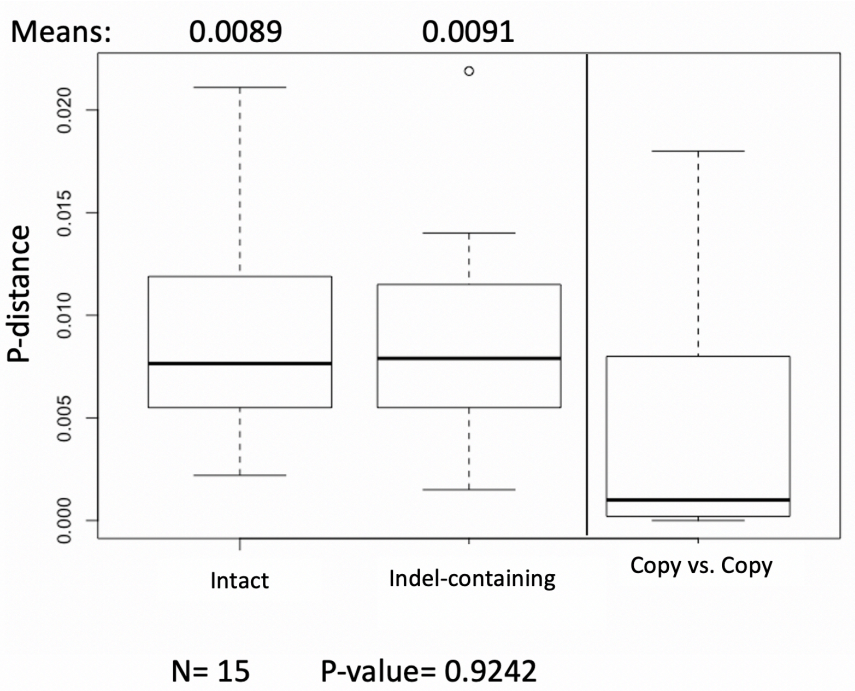

Figure 2. The y-axis is the pairwise distance calculated using MEGA (Kumar et al, 2018). All differences are included in the calculation of pairwise analysis for this plot. The black bars represent the median for each box. There are 15 genes in each box. Statistical analysis done in RStudio (RStudio Team, 2015). Refer to columns "indel", "intact", and "indel – intact" in Supplementary Table 5 for raw values.

Pairwise Differences By Codon Position

Each nucleotide position in a coding sequence is potentially under a different selective constraint. First and second site changes are seen less frequently than third site changes because first and second site changes typically result in amino acid changes. I utilized this concept in the analysis of my 20 gene-sets and took variations of 1st, 2nd, and 3rd site pairwise distances separately. It common for 1st and 2nd site changes to be combined together as likely causing nonsynonymous mutations. However, there are a limited number of changes in 1st site which could still be synonymous while that is not true of 2nd site, so a full breakdown of pairwise distances was necessary. I expected the indel-containing copy to contain more mutations as that copy may be under less selective constraint due to the frameshift mutation and the presence of an

intact copy. All subsequent site-based pairwise distance calculations exclude the five sets which do not contain frameshifts.

On average, the indel-containing copies have a larger 1st site pairwise distance than the intact copies (Figure 3). However, the intact copies have a larger 2nd site pairwise distance, on average, than the indel-containing copies (Figure 3). However, given the p-values of 0.845 and 0.698 for the 1st and 2nd site pairwise-distances, respectively, there is no statistical difference between the two groups (Figure 3). When comparing the 1st site calculated pairwise distance between the copies, there is a high level of variability, similar to the complete pairwise distance (Figure 3). Interestingly, there is very low level of variability, on average, when comparing the 2nd site pairwise distances between copies (Figure 3). This may be consistent with a scenario where a WGD happened in recent evolutionary history combined with the 2nd sites' low tolerance for substitution.

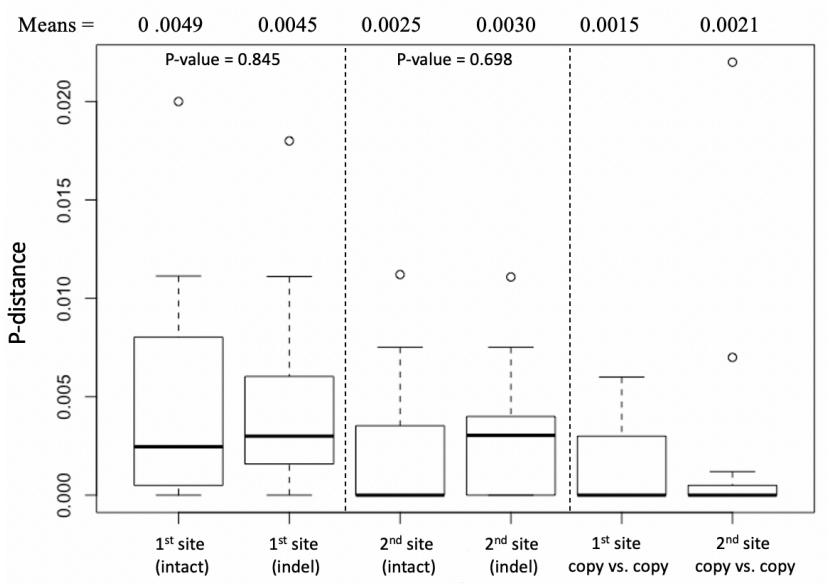

Figure 3. The y-axis is the pairwise distance calculated using MEGA (Kumar et al, 2018). Only 1st or 2nd site differences are included in calculation of pairwise analysis for this plot. The black bars represent the median for each box. There are 15 genes in each box. Statistical analysis done in RStudio (RStudio Team, 2015). Refer to Supplementary Table 1 for individual values.

On average, the 3rd site pairwise distances are nearly identical between the indel-containing copies and the intact copies (p= 0.9325) (Figure 4). When comparing pairwise differences of the 1st and 2nd site differences together, on average, they do not differ between copies (Figure 5). The variability between indel-containing and intact copies is still detectable upon combining the sites for analysis (Figure 5). These results are unexpected because, in principle, a pseudogenizing gene (defined by the presence of a frameshift) should be more tolerant of 1st and 2nd site mutations and would be accumulating more of these than its intact counterpart. Likely explanations for this observation include short divergence time or small sample size. In the case of 3rd site differences, a p-value of 0.9325 is actually expected (Figure 4). The case of 3rd differences is unique when considering WGDs because they are under little evolutionary constraint even in an unduplicated genome and therefore should not differ between copies. Any extreme differences between them could be explained as sporadic due to small sample size. The exception would be that in a large data set, there is potential to detect the miniscule occurrence of the scenario in which 3rd site mutations do change the amino acid.

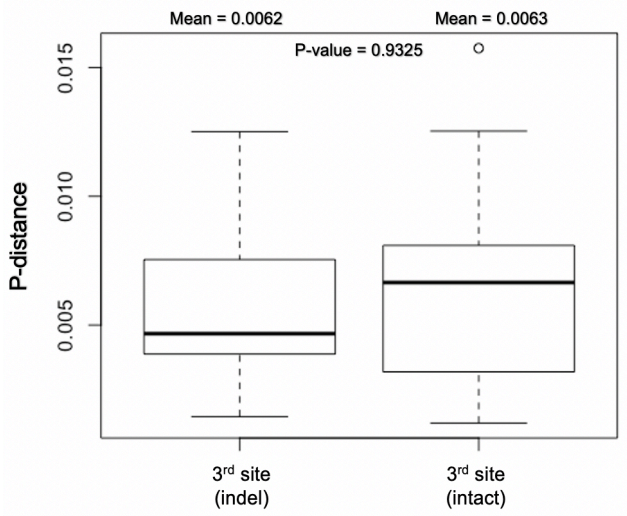

Figure 4. The y-axis is the pairwise distance calculated using MEGA (Kumar et al, 2018). Only 3rd site differences are included in the calculation of pairwise analysis for this plot. The black bars represent the median for each box. There are 15 genes in each box. Statistical analysis done in RStudio (RStudio Team, 2015). Refer to Supplementary Table 2 for individual values.

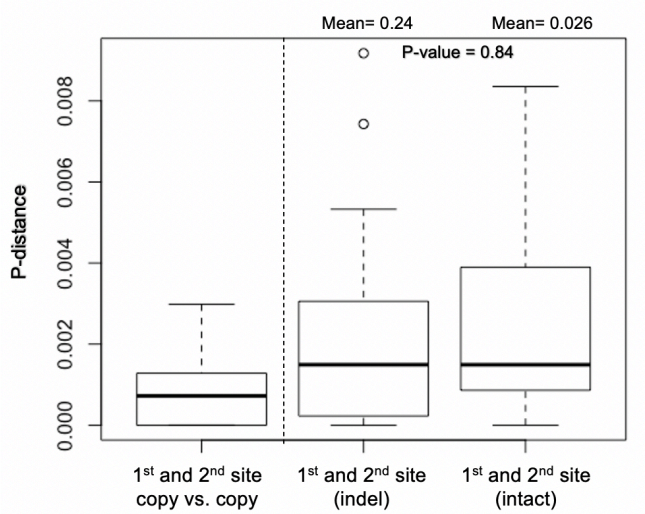

Figure 5. The y-axis is the pairwise distance calculated using MEGA (Kumar et al, 2018). The sum of 1st and 2nd site differences are included in the calculation of pairwise analysis for this plot. The black bars represent the median for each box. There are 15 genes in each box. Statistical analysis done in RStudio (RStudio Team, 2015). Refer to Supplementary Table 2 for individual values.

Calculating the sum of 1st and 2nd site pairwise distances and dividing it by the 3rd site pairwise distance is a typical estimation of positive or negative selection. If this value is larger than two, it means that nonsynonymous mutations are accumulating over synonymous mutations. If this value is below two, it means that synonymous(S) mutations are accumulating over nonsynonymous (NS) mutations. I expect the (1st + 2nd)/3rd or NS/S value to be higher for the indel-containing copy than for the intact copy because the indel-containing copy will be more tolerable to nonsynonymous mutations due to the mutational shielding created by the presence of the intact copy. On average, the NS/S value is higher in the intact copies than the indel-containing copies (Figure 6); however, the p-value for this comparison is 0.296 and thus not significant (Figure 6). A single outlier appears to be skewing the average for the intact copies; the average NS/S for the intact copies is calculated to be 0.42 excluding the outlier (Figure 6). While NS/S is still unexpectedly higher for the intact copies than the indel-containing copies, given the time of divergence and low samples size, the differences between the averages excluding the outlier align with the data from other site-based calculations.

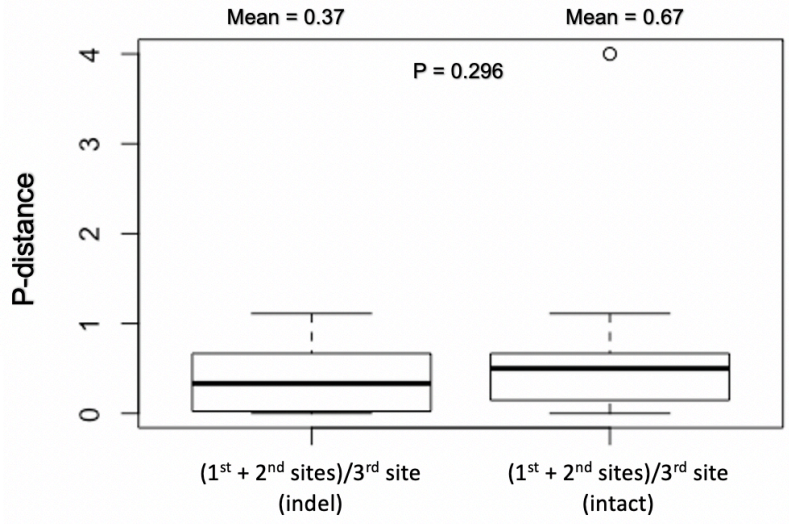

Figure 6. The y-axis is the pairwise distance calculated using MEGA (Kumar et al, 2018). The values for this plot were calculated by dividing the previously calculated pairwise distance for 1st +2nd site divided by the 3rd site. The black bars represent the median for each box. There are 15 genes in each box. Statistical analysis done in RStudio (RStudio Team, 2015). Refer to Supplementary Table 2 for individual values.

The fact that the aggregate analyses of all site-based comparisons of indel-containing copies and intact copies produced no statistical significance concerning the difference between copies further supports the idea that *P. antipodarum* underwent an evolutionarily recent WGD. That is, there has not been enough elapsed time since the WGD, given the rate of nucleotide substation, to generate a consistent and robust general signal. If the WGD had happened today, then the copies would be completely identical without exception. If the WGD happened too long ago, in some cases, one copy of a gene will have accumulated so many mutations that it will no longer be identifiable as a duplicate of the intact gene. However, lack of statistical significance regarding the difference between copies in *P. antipodarum* does not mean they are not diverging. These analyses involved sorting the intact copies and the indel-containing copies into two separate groups and comparing averages. In addition, if the duplication happened very recently, any observable site-based differences could have arisen prior to the indel. That being said, more time since divergence could also provide statistical significance given that the amount of time is not so long that genes are no longer classifiable duplicates. Including more genes in duplicate statistical analysis serves to increase the ability to detect true differences. In this small sample, the average differences are more easily skewed by outliers. It is possible that there are detectable site-based differences between duplicate genes, but that my sample size of 15 indel-containing genes is too small to detect it as an overall effect.

Gene Copy Normalization

Next, I compared the duplicate genes to each other rather than comparing averages across gene pairs. In contrast to the findings averaged across genes, I found that, of the 15 indel-containing gene sets, the indel-containing copy is accumulating more mutations than the intact copy a

majority of the time (Figure 7). These analyses provide strong evidence for the initial stages of pseudogenization in *P. antipodarum*. This is a clear demonstration of early processes following a WGD, and follows clear predictions from evolutionary models (Force *et al*., 1999). This result is based on a conservative analysis because it does not include the indels themselves as differences in the calculation of pairwise distance. By subtracting the pairwise distances between copies, I eliminated the confounding factor of highly variable averages among gene sets. In previous analyses based on grouping the genes based on their indel-containing status, true divergence based on my hypothesis is masked by the differences between pairwise distances per gene; this would allow a gene set which has the indel-containing copy to be accumulating more mutations than the intact copy to look as if this was not true because the individual values must be normalized to their counterpart in the genome.

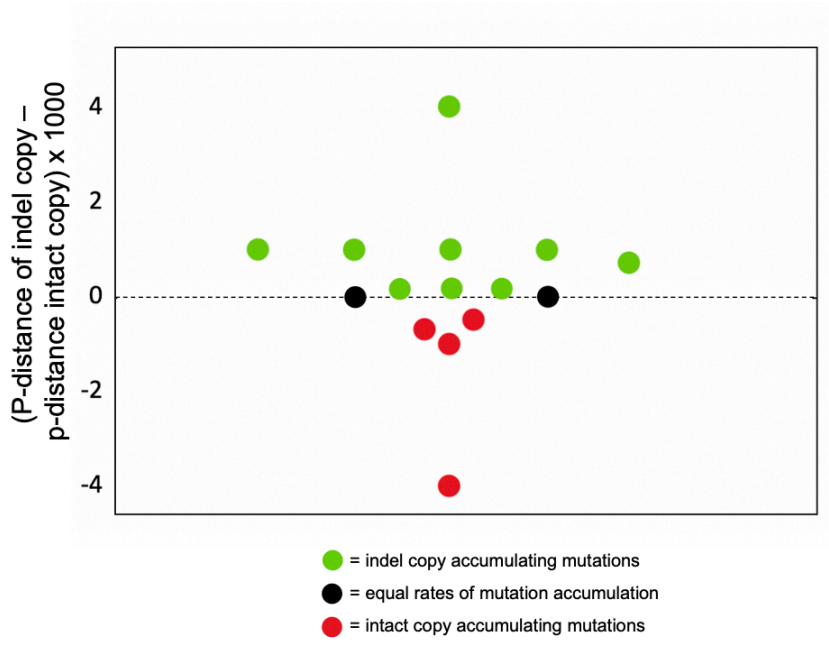

Figure 7. All differences are included in the calculation of pairwise analysis for this plot. The pairwise difference for the intact copy was subtracted from the pairwise distance of its indel-containing paralog and multiplied by 1000 for graph scaling purposes. The black bars represent the median for each box. There are 15 genes in each box. Statistical analysis done in RStudio (RStudio Team, 2015). Refer to Supplementary Table 5 for original values.

Additional Exploratory Analysis: Introns

Introns are often used in pseudogene analysis as an exemplar of neutral evolution. I attempted to implement intron analysis in the characterization of duplicate genes concerning the evolutionary fate of genes in *P. antipodarum*. All introns were annotated from raw reads and approximately three homologous introns from each gene were annotated for analysis. I was unable to annotate all introns for the gene set due to technical obstacles such as high variability among introns. My initial results (data not shown) indicated that the indel-containing copy may have a pairwise distance larger than that of the intact copy, but smaller than that its introns, thus indicating its journey toward pseudogenization. However, given the recency of the duplication, the quality of annotation, and the discovery that the averages among genes were highly variable, the utility of introns in this dataset was deemed inconclusive. Refer to Supplementary Table 4 for a summary of data.

Conclusions

Utilizing predicted single-copy genes is useful for analyzing the WGD in *P. antipodarum* and is a promising method for the development of a conservative model for estimating duplicate gene divergence in a young duplication. Successful analysis of WGDs are highly beneficial in characterizing genome evolution and has application in simple gene duplication as well. Single gene duplication is a more common and tolerable event than WGD, therefore a WGD provides an extensive catalogue of duplicate genes consisting of many functional types. The high level of similarity without being identical between a majority of duplicate genes indicates sufficient time has passed to be able to detect differences between duplicates. A study projecting how long it has been since divergence would supplement this finding and indicate approximately now long it

takes for even the most strictly conserved genes to begin diverging. Additionally, indels are common and present in nearly identical duplicate *P. antipodarum* genes. The presence of indels in one copy and effectively no indels in another copy of the same gene after a WGD may be an indicator of the initial steps leading to pseudogenization.

More research is needed as this project only covered 20 of the predicted 664 single-copy genes in *P. antipodarum*. Including more genes in these analyses would also serve to increase statistical power and potentially provide a framework for automation. Annotating and, in the case of *P. estuarinus*, compiling sequence from raw reads is extremely time-intensive, but necessary at this time due to the low number of detectable differences over thousands of nucleotides. Computer automated sequence annotation produced errors in an exploratory study of its use. In this instance, the number of introduced errors was unacceptable as it equaled or exceeded the number of differences estimated to be correct from manual annotation. Automation is still promising and should be pursued to reduce the amount of time and resources involved and increase statistical power so that trends can be more strongly supported and analyses can expand outside the predicted single-copy dataset. Because predicted-single copy genes are also predicted to be highly conserved, they are relatively easy to analyze for these types of analyses. Including other parts of the genome such as additional non-single-copy genes, intergenic regions, and introns will be beneficial but accounting for the increased acceptable level of baseline variability will be more difficult. Finally, functional analysis of any duplicate data set will be needed to provide information on which genes tolerate duplication, at least more so than others. In turn, such analyses may indicate which genes are under the most or least selective constraint from the time of duplication. The longer duplicates are retained, recognizable, and functional, the more potential there is for subfunctionalization or neofunctionalization.

Supplemental Tables

Supplemental Table 1.. 1st and 2nd site pairwise distances obtained for 15 duplicate genes as well as a direct comparison between the duplicate genes. Values obtained via analysis in MEGA (Kumar et al, 2018).

indel copy		intact copy		copy vs. copy	
1st	2nd	1st	2nd	1st	2nd
0.001	0	0.001	0.00120192	0.006	0.00119904
0.007	0.003	0.003	0.004	0.004	0.001
0	0	0.00217391	0	0.0021692	0
0.001	0	0.003	0	0.004	0
0.00246002	0	0.00246002	0	0	0
0.00223714	0.002	0.00223714	0.004	0	0.007
0	0.004	0	0.004	0	0.022
0	0	0	0	0	0
0.00904977	0	0.00902935	0	0	0
0.007	0	0.003	0.00328947	0.003	0
0.02	0.0075188	0.018	0.0075188	0.003	0
0.00304878	0.0030581	0.00303951	0.00303951	0	0
0.01113586	0.01121076	0.01111111	0.01108647	0	0
0	0	0	0	0	0
0.00967742	0.00642055	0.00967742	0.00641026	0	0

Supplemental Table 2. 1st + 2nd site pairwise distance measurements, 3rd site pairwise distance measurements, and (1st+2nd)/3rd calculations are included for both copies of 15 P. antipodarum *genes. Site-based values were obtained using MEGA (Kumar et al, 2018). (1st+2nd)/3rd calculations were performed in excel using previously obtained values.*

indel copy			intact copy		
1st + 2nd	3rd	(1st + 2nd)/3rd	1st + 2nd	3rd	(1st + 2nd)/3rd
0.000454959	0.00955414	0.047619048	0.0009095	0.009095043	0.1
0.001492537	0.004477612	0.333333333	0.00149142	0.008202834	0.181818182
0.001333333	0.004666667	0.285714286	0.00133156	0.003328895	0.4
0	0.011826544	0	0	0.015748031	0
0.002181025	0.003271538	0.666666667	0.00108932	0.002178649	0.5
0.003104213	0.007095344	0.4375	0.00222519	0.007120605	0.3125
0	0.001447178	0	0.00072307	0.001446132	0.5
0.000815328	0.006930289	0.117647059	0.000815	0.007334963	0.111111111
0.003007519	0.004511278	0.666666667	0.003003	0.004504505	0.666666667
0	0.003686636	0	0	0.003683241	0
0.005327651	0.007991476	0.666666667	0.00532481	0.00798722	0.666666667
0	0.003594249	0	0.00479042	0.001197605	4

Supplemental Table 3. Data collected for the direct comparison between 15 P. antipodarum *duplicates. The two right most columns labeled "1st + 2nd" and "3rd" are pairwise distances calculated in MEGA (Kumar et al, 2018). The left columns labeled "1st + 2nd" and "3rd" are the sheer number of differences between the two copies.*

Gene	1st + 2nd	3rd	avg. gene length	1st + 2nd	3rd
1	3	19	2198.5	0.001364567	0.008642256
2	4	13	1340.5	0.002983961	0.009697874
3	4	2	1501	0.00266489	0.001332445
4	0	10	1523	0	0.006565988
5	1	1	917.5	0.001089918	0.001089918
6	4	2	2256	0.00177305	0.000886525
7	1	0	1382.5	0.000723327	0
8	0	1	2454.5	0	0.000407415
9	0	0	1381	0	0
10	0	0	1085.5	0	0
11	0	0	1877.5	0	0
12	3	18	2504.5	0.001197844	0.007187063
13	1	0	1198	0.000834725	0
14	0	1	985.5	0	0.001014713
15	0	0	1349.5	0	0
16	0	0	1650	0	0
17	0	0	1140	0	0
18	3	28	2292	0.001308901	0.012216405

Supplementary Table 4. Summary of data collected for intron analysis. Left two columns are pairwise distances for all sites calculated in MEGA (Kumar et al, 2018). The third column from the left is the pairwise distance between copies. The left most column is the subtraction of the intact copy pairwise distance from the indel copy pairwise distance. Only 14 genes are included in this summary due to annotation errors in one of the indel-containing genes.

indel copy	intact copy	copy vs. copy	indel - intact
0.014	0.016	0.006	-0.002
0.4	0.3	0.3	0.1
0.02	0.01	0.01	0.01
0.054	0.056	0.002	-0.002
0.03	0.03	0.01	0
0.01438849	0.01436266	0	2.58321E-05
0.008	0.004	0.01	0.004
0.0339	0.0339	0	0
0.0217	0.0217	0	0
0.023	0.024	0.007	-0.001
0.0286	0.0286	0	0
0.001	0.003	0.001	-0.002
0.03	0.01	0.01	0.02
0.05	0.06	0.03	-0.01

Supplementary Table 5. Summary of pairwise distance measurements including differences from all sites obtained from analysis in MEGA (Kumar et al, 2018). The "indel – intact" refers to the pairwise distance of the intact copy subtracted from the pairwise distance of the indel copy. "copy vs. copy" refers to the pairwise distance calculated between P.antipodarum *duplicates, rather than compared to* P. estuarinus *as in the "indel" and "intact" columns.*

indel copy	intact copy	indel - intact	copy vs. copy
0.01	0.01	0	0.01
0.01	0.006	0.004	0.018
0.005	0.006	-0.001	0.009
0.012	0.016	-0.004	0.007
0.005	0.004	0.001	0.002
0.011	0.01	0.001	0.004
0.0015	0.0022	-0.0007	0.007
0.0079	0.0083	-0.0004	0.0004
0.00765697	0.00764526	1.17079E-05	0
0.00373832	0.003734827	3.4905E-06	0
0.0137893	0.013781698	7.6016E-06	0
0.007	0.006	0.001	0.014
0.0219	0.0211	0.0008	0.008
0.006	0.005	0.001	0.001
0.014	0.014	0	0

Supplemental Figure

Supplemental Figure 1. Example alignment of gene 5 or BUSCO reference BAG. Assumed heterozygosity in P. estuarinus *(P. est) is indicated by a yellow dot. Divergence between the* P. estuarinus *gene and the versions in* P. antipodarum *(tig00026969 and tig00035169) is indicated by a green dot. Presence of an indel is indicated by a red dot. Alignment performed using Clustal (Kumar et al, 2018). Refer to tables 1 and 2 for a summary of substitutions and indels in this gene*

Literature Cited

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool." *J. Mol. Biol.* 215:403-410.

Crow, K.D., Wagner, G.P. (2006). What Is the Role of Genome Duplication in the Evolution of Complexity and Diversity? *Molecular Biology and Evolution*, 23: 887–892.

Dehal P., Boore J.L. (2005) Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. PLoS Biol 3(10): e314. https://doi.org/10.1371/journal.pbio.0030314

del Pozo, Carlos, J., Ramirez-Parra, E., (2015). Whole genome duplications in plants: an overview from *Arabidopsis*. *Journal of Experimental Botany*, 66: 6991–7003

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., Postlethwait, J., (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151.4: 1531-1545.

Gasteiger E., Gattiker A., Hoogland C., Ivanyi I., Appel R.D., Bairoch A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis *Nucleic Acids Res*. 31:3784-3788

Kumar S., Stecher G., Li M., Knyaz C., and Tamura K. (2018). MEGA-X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution* 35:1547-1549

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Sequencher® version 5.1 DNA sequence analysis software, Gene Codes Corporation, Ann Arbor, MI USA.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, *Bioinformatics*, Volume 31, Issue 19, 1,, Pages 3210–3212, https://doi.org/10.1093/bioinformatics/btv351

Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., Zdobnov, E.M., (2017). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*. 35(3), 543-548 doi: 10.1093/molbev/msx319.