1-1-2015

Evolution Of New Duplicate Genes In Arabidopsis Thaliana

Nicholas Curtis Marowsky
Wayne State University,

EVOLUTION OF NEW DUPLICATE GENES IN ARABIDOPSIS THALIANA

by

NICHOLAS MAROWSKY

THESIS

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

2015

MAJOR: BIOLOGICAL SCIENCES

Approved By:

_____

Advisor Date

DEDICATION

This thesis is dedicated to the memory of my late mother Denise Marowsky. Without her love and guidance none of this would have been possible.

ACKNOWLEDGMENTS

I would like to thank everyone who helped me in this endeavor. Dr. Chuanzhu Fan for guiding me in the project and teaching me everything I needed to know to complete this task. I would also like to thank my committee, Dr. Edward Golenberg and Dr. Weilong Hao, for their input and guidance. Thank you Dr. Jun Wang for aiding me in the computational analysis, and Feng Tao for helping me with day to day lab tasks and the statistical analysis. I want to extend my appreciation to the Fan Lab undergrads for assisting me in my work, thank you Muhieldean Ibrahim, Muhieldean Ibrahim, Muhadia Rafi, Eun Young Kim, Richa Bhatia, Reekan Honest, Abdullah Islam, Benish Alam, Jessica Cobey, Diebh Faraj, and Barah Alden for all of your hard work. I would like to thank Wayne State University for facilitating my research and education, and the Biological Sciences staff for their support. And finally I would like to thank my family and friends, without their love and support I would not be the person I am today and would not be where I am today.

TABLE OF CONTENTS

Dedication ii

Acknowledgments iii

Introduction 1

Gene Duplication 4

GINS Complex Genes 7

Phosphoribosylanthranilate Isomerase Genes 7

Chloroplast Proteins of Unknown Function 9

Materials and Methods 9

Results 14

GINS Complex Genes Results 14

Phosphoribosylanthranilate Isomerase Genes Results 19

Chloroplast Proteins of Unknown Function Results 22

Discussion 30

Appendix 33

Works Cited 48

Abstract 51

Autobiographical Statement 53

LIST OF TABLES

Table 1 List of Gene Pairs used in Analysis 6

Table 2 Tajima's D and McDonald-Kreitman Tests 33

Table 3 Primers used for PCR 35

Table 4 ANOVA Table for Phenotype Assay 1 37

Table 5 ANOVA Table for Phenotype Assay 2 38

Table 6 ANOVA Table for Phenotype Assay 3 39

Table 7 ANOVA Table for Phenotype Assay 4 40

Table 8 List of Accession Lines used for Sequencing Analysis 41

LIST OF FIGURES

Figure 1 Phylogeny of Brassicaceae Species 5

Figure 2 Structural Comparison of Gins Genes 7

Figure 3 Structural Comparison of Pai1, Pai2, and Pai3 Genes 8

Figure 4 Structural Comparison of *AT2G05310* and *AT4G13500* Genes 9

Figure 5 Developmental Stages of *Arabidopsis thaliana* 14

Figure 6 PCR Example of Deletion in *AT1G19080* 15

Figure 7 Neighbor Joining Tree for *AT3G55490* 17

Figure 8 Neighbor Joining Tree for *AT1G19080* 18

Figure 9 Neighbor Joining Tree for *AT1G29410* 21

Figure 10 Expression Analysis for *AT1G29410* 22

Figure 11 Neighbor Joining Tree for *AT2G05310* 23

Figure 12 Neighbor Joining Tree for *AT4G13500* 24

Figure 13 Expression Analysis for *AT4G13500 and AT2G05310* 25

Figure 14 PCR Confirmation for ML68 tDNA Insertion 26

Figure 15 Development Times for First Set of Phenotype Analyses 27

Figure 16 Development Times for Second Set of Phenotype Analyses 29

Figure 17 Positive Control for cDNA used in Expression Analyses 36

Figure 18 Additional N-J Tree for *AT3G55490* without Homologs 43

Figure 19 Additional N-J Tree for *AT1G19080* without Homologs 44

Figure 20 Additional N-J Tree for *AT1G29410* without Homologs 45

Figure 21 Additional N-J Tree for *AT2G05310* without Homologs 46

Figure 22 Additional N-J Tree for *AT4G13500* without homologs 47

INTRODUCTION

Understanding the mechanisms by which new genetic information arises is important in our understanding of how genes and genomes change throughout evolutionary time. By deciphering the mechanism by which new genetic sequence and new genes come into being, we will have more of an understanding of how organisms evolve novel functions. Gene duplications are believed to be a driving force in the process of speciation (Bikard et al., 2009), possibly offering the organism an advantage over their relatives (Roux et al., 2011). By understanding how new genes change over time, we can begin to understand how new genes contribute to the evolutionary process on the larger scale.

New genes can arise from a variety of mechanisms, the most common fall into two major categories, duplication and *de novo* gene formation. *De novo* genes are those which arise as expressed genes from a region of DNA previously not expressed and showing no traceable paralogues in closely related species (Tautz and Domazet-Loso, 2011). These genes, and the mechanisms by which they come into being, are still poorly understood. The second class, duplicate genes, can arise through a variety of mechanisms which include RNA based retroposition events and DNA based duplication events including but not limited to, whole genome duplication, segmental duplication, unequal crossing over, and transposition (Long et al., 2003). While all new gene development is interesting, this analysis will focus on new duplicate genes.

Duplication of a DNA sequence is a relatively common occurrence, which offers an organism new material for the possible development of new genes and functions (Zhang, 2003). Often, during these events an active gene is copied. The new sequence leaves the organism with an extra copy of the gene, as the system was previously functioning without this new copy. This situation allows for one of several possible outcomes. Firstly, if an excess of the original gene product is beneficial for survival, then the new and old copies of the gene could continue on producing the original product unmodified, simply making more of it. This is known as gene redundancy. Secondly, the genes could become altered in either expression pattern or amino acid sequence in such a way that each copy to perform a subset of the original functions; this is termed subfunctionalization. Another case involves one gene copy diverging in sequence enough that the changes in its product would allow it to perform a function completely unrelated to its parental function, termed neofunctionalization. The last case, when the new copy of the gene does not benefit the organism, one of the genes simply mutates until it is no longer functional and the sequence becomes non expressed DNA, a pseudogene (Prince and Pickett, 2002). These actual outcomes are dependent on the selective pressures present in the environment, and investigating their occurrence will offer us insight in the process of new gene evolution. Gene duplication is a powerful process, allowing an organism to change more quickly than by mutating one nucleotide at a time. We have chosen to study a handful of new duplicate genes within the plant *Arabidopsis thaliana,* in order to more accurately describe these occurrences within the plant genome.

With the advent of high throughput sequencing we have seen a veritable explosion in of new gene discovery. The availability of complete genome sequences for many closely related species has allowed for very powerful computation based comparisons between species. Due to the availability of whole genome sequencing for *Arabidopsis thaliana* and its close relatives *A. lyrata*, *Capsella rubella* and *Brassica rapa*, we have chosen to use *A. thaliana* as our target organism for this study.

Arabidopsis thaliana is one of the most commonly used model organisms for biological study. It is a small selfing plant found throughout the world in a variety of climates. It originates from Eurasia, but can now be found on all continents excluding Antarctica. *A. thaliana* is short lived, easy to maintain, and its high fecundity make it an ideal model for laboratory studies. *Arabidopsis thaliana* was the first plant to have its genome fully sequences. Its genetic makeup is relatively simple, with one of the smallest genomes of any angiosperm at 121Mbp and 5 chromosomes (Arabidopsis Genome, 2000). For comparisons sake, the other species used in this project have significantly larger genomes, *A. lyrata* 210Mb with 8 chromosomes, *C. rubella* 136 Mb with 8 chromosomes, and *B. rapa* has 290Mb and 10 chromosomes. Despite the relatively simple appearance of the *A. thaliana* genome it contains a disproportionate amount of duplication; it is estimated that roughly 65% of its genes arose from duplication compared to *Drosophila melanogaster* at 41% or *Caenorhabditis elegans* at 49% (Zhang, 2003). This makes it a perfect model for the discovery and investigation of duplicate genes.

Gene duplication events appear to be very common in all domains of life. However, older duplicates will have had more time to fixate within the population and will likely display less sequence dynamics. We are interested in how these duplicate genes change after their creation, so we will need to find duplicate genes which have been recently copied. To identify these new duplicate genes, we first had to analyze and compare genes within the species *A. thaliana*, and between *A. thaliana* and its close relatives. Based on genetic analyses, it is estimated that *A. thaliana* and *A. lyrata* separated 5-10 MYA (million years ago), while *C. rubella* split around 15MYA and *B. rapa* roughly 20 MYA (Figure 1). Therefore if we can locate *A. thaliana* genes which show significant similarity to other genes within the *A. thaliana* genome, one of these genes is likely a copy of the other. To determine whether the new copy is lineage specific, thus establishing an age limit on the new gene, we need to search for this sequence in the other related species. If the sequence shared by the gene pair is similar to genes which exist as single copy genes in the other 3 species, it is likely that the new duplicate gene in *A. thaliana* was created in the last 5-10 Ma (Million years). These are the genes we wish to identify and study, termed lineage specific new duplicate genes. Identification of these lineage specific new duplicate genes relied on two criteria: the new gene must not be located in a similar syntenic region compared to the other species, and the gene could not have any reciprocal ortholog in the related species (Wang et al., 2013).

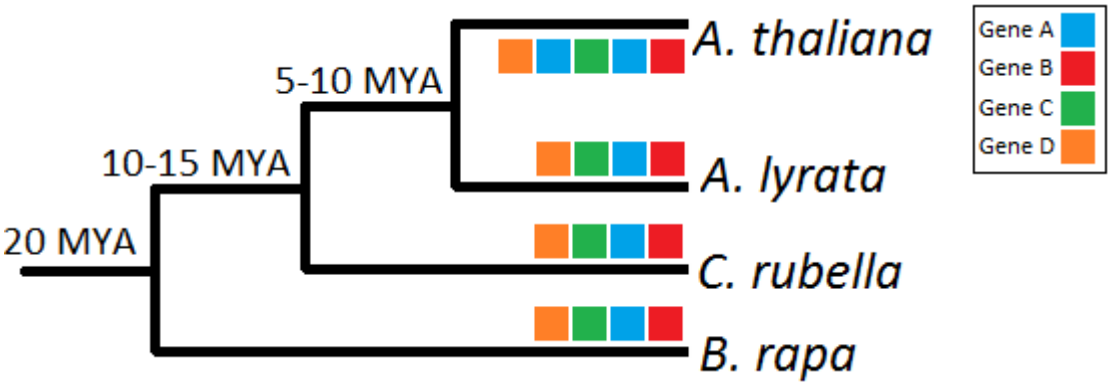


Figure 1: Phylogeny of Brassicaceae species included in this study with divergence times noted at forks (Kleffmann et al., 2004; Koch et al., 2008). Colored squares used to represent genes and display lineage specific gene new duplication in A. thaliana genome.

By this method we found 137 lineage specific NDGs, however 37 of these genes appeared to be generated by duplication of another lineage specific gene. As we could not establish a parent/duplicate relationship in these pairs, they were ignored for this study. This left us with 100 new duplicate genes (Wang et al., 2013). Of these 100 genes we wanted to take a more precise look at their evolution in the *A. thaliana* population at large. We focused on 6 genes, 3 pairs of parental and new genes, which we analyzed in order to investigate their evolution more thoroughly. We studied parental genes *AT3G55490*, *AT1G07780* and *AT2G05310* as well as their respective duplicates *AT1G19080*, *AT1G29410*, and *AT4G13500*. Based on the analysis of the published genomes and expression data, these 3 sets all exhibited differential expression, contain substitutions and display evidence of non-neutral selection. The latter of the 3 sets is a gene family of unknown function, while the former 2 gene families had published functions.

Table 1: Gene pairs, their Ka/Ks value and divergent expression profile

Duplicate Gene	Parental Gene	NDG Branch Specific Ka/Ks	Parental Enriched Gene Expression	NDG Enriched Gene Expression
AT1G19080	*AT3G55490*	0.0001	Non-Specific	Leaf
AT1G29410	*AT1G07780*	0.62481	Inflorescence	Silique
AT4G13500	*AT2G05310*	0.07545	Flower	Non-specific

In order to gain a more thorough understanding of these new duplicate genes, we will study their sequences, expression patterns and attempt to determine the function of the pair not currently classified. To understand the evolutionary dynamics of these gene pairs, we will sequence these genes as they exist in *A. thaliana* populations in a variety of regions. These sequences will be the basis of a series of statistical analyses in order to determine what selective pressures may be affecting the genes. To look for evidence of subfunctionalization we will also be conducting expression assays in order determine the actual expression patterns of these genes. Lastly, in an attempt to identify the function of the gene set with no annotated function, we will use tDNA knockout mutants to determine any visible phenotypic differences when compared to wild-type plants. By these assays we will attempt to more accurately depict the nature and the dynamics of these duplicate gene pairs.

For use in sequencing, we obtained *A. thaliana* 80 population lines from the Arabidopsis Biological Resource Center (ABRC) stock center (Table 3). These lines are derived from populations in a variety of locations in Europe,

northern Africa, the Middle East and western Asia. By sequencing the genes in populations from varied environments and locals, we will develop a greater understanding of the evolutionary dynamics of these duplicate genes and their parents.

GINS COMPLEX GENES

The gene pair *AT3G55490* and its duplicate *AT1G19080* are highly conserved and believed to function as a part of the GINS complex. The identification is solely based on protein BLAST analysis reported to www.arabidopsis.org and www.phytozome.net, blastp comparison with the *Homo sapiens* and *Xenopus laevis* proteomes both offer matches to GINS complex subunit 3 scoring 89% query cover with E-values of 2e -14 and 2e -15 respectively. The GINS complex is involved in chromosomal DNA replication (Takayama et al., 2003). These two genes are highly similar in their coding sequences, having over 99% sequence similarity.

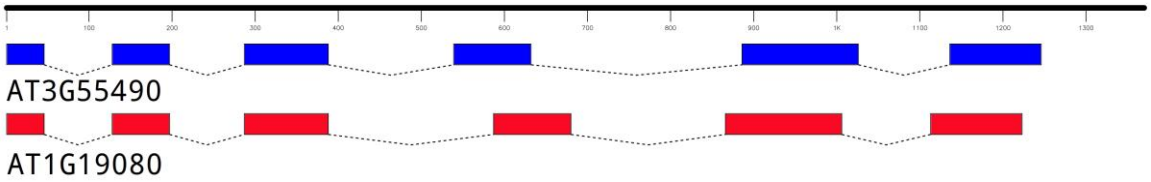


Figure 2: Comparison of gene structures between *AT3G55490* and its duplicate *AT1G19080*.

Phosphoribosylanthranilate Isomerase (PAI) Genes

The genes *AT1G07780* (Pai1) and *AT1G29410* (Pai3) are encode proteins noted for their involvement in the tryptophan biosynthesis pathway. There is third gene within this family, *AT5G05590* (Pai2). However the Pai2 gene

appears to be another copy of Pai1, as it shares more sequence similarity, and the new gene Pai3 is more similar to Pai1 than to the Pai2 gene. The outgroup species each contain only one copy of this sequence. So this system still fits within our parameters, as Pai3 is a copy of the parental gene Pai1. Pai1 appears to perform the majority of the enzymatic activity, while Pai2 is activated when the plant is under stress (He and Li, 2001). Based on deletion assays they posit that Pai3 is a non-functioning copy. This is supported by the fact that the sequence of Pai3 displays a single base pair deletion, which causes a -1 frame shift in the 3rd exon. This frame shift offsets a splicing site and causes an early termination. Pai1 and 2 have very similar sequences, the same cDNA length, and the same exon and intron makeup.

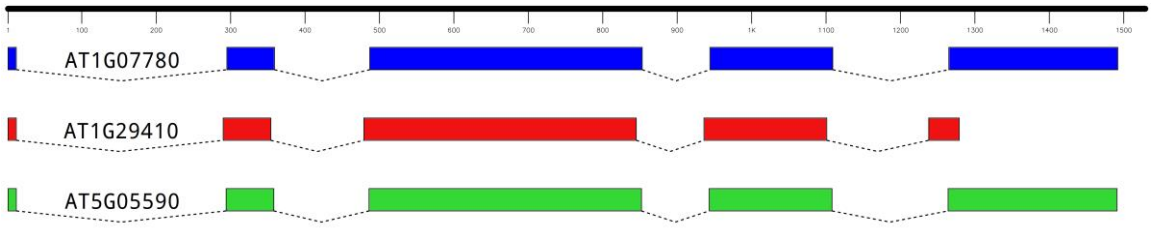


Figure 3: Comparison of gene structures between *AT1G07780* (Pai1) and *AT1G29410* (Pai3), also shown *AT5G05590* (Pai2).

CHLOROPLAST PROTEINS OF UNKNOWN FUNCTION

Our last gene pair is *AT2G05310* and *AT4G13500*, of which little is known. These genes reportedly lead to the production of proteins, which can be found in the chloroplasts (Kleffmann et al., 2004). However nothing about the proteins' function has been reported to our knowledge.

The sequences of these two genes are divergent enough to allow us to fully analyze the two with our methods. The genetic makeup of the DNA is very similar between the two copies; they have the same cDNA length and the same splice sites. The introns however are dissimilar, but in coding sequence they are >90% similar.

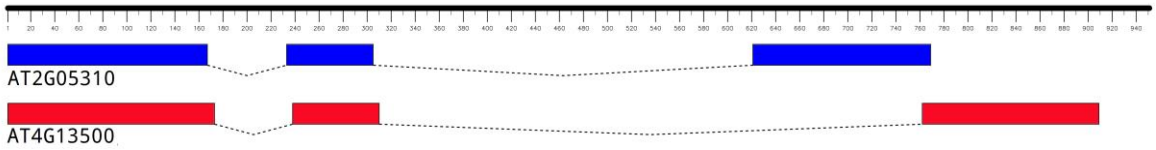


Figure 4: Comparison of gene structures between *AT2G05310* and *AT4G13500*.

MATERIALS AND METHODS

Due to the published *A. thaliana* genomes being based on high throughput computationally compiled sequences, we required a more thorough analysis in order to determine the polymorphisms present in the populations more accurately. We used Sanger termination sequencing in order to recover SNPs which are not properly annotated or missing from the Illumina SNP data due to ambiguous sequence reads, low coverage, or mapping errors in the publically available reference genomes for the 80 population lines. Primers were designed to target the sequences in question (see Material and Methods) and the

sequences were amplified using PCR. The products were then sent to Eton Biosciences for sequencing. The sequences were then compiled into FASTA files and aligned using Muscle. The compiled sequences were then transferred to DnaSP and MEGA statistical analysis and phylogenetic tree construction.

To investigate expression patterns, RNA was extracted from five tissues, leaf, root, stem, flower and silique using Qiagen RNeasy Plant Mini Kits. Our cDNA libraries were created by RT-PCR. The cDNA was run through PCR using gene specific primers, the products were visualized by gel electrophoresis.

For the genes that do not have a known function, tDNA mutant lines were acquired from the ABRC stock center for phenotypic analysis. We planted the mutant lines alongside a control group of Colo-0 plants and noted their development times. These developmental rates were then analyzed to determine whether the mutation has any effect on development. After flowering, a random sample of plants was tested to confirm the tDNA insertion using gene specific primers and a primer specific to the left border of the tDNA insertion.

DNA EXTRACTION

Plant tissues were frozen with liquid nitrogen and ground before DNA extraction with CTAB Buffer (see Solutions). Buffer was added to finely ground plant tissue and incubated in a 55° C water bath at least one hour. Afterward 500µl of 24:1 chloroform: isoamyl alcohol was added and the mixture was centrifuged at 15,000 rpm for 10 minutes. The aqueous phase was extracted, to which was added 0.08 volumes of 7.5 M cold ammonium acetate and 0.54

volumes of cold isopropanol. This mixture was placed in the freezer for at least 15 minutes, followed by centrifugation for 3 minutes at 15,000 rpm. The pellet was retained and was washed with 700µl cold 70% ethanol vortexed and centrifuged at 15,000 rpm for 1 minute. The supernatant was discarded and the pellet was washed again with 700µl cold 95% ethanol, vortexed and centrifuged at 15,000 rpm for 1 minute. The supernatant was discarded and the samples were allowed to air dry before resuspension with 30µl TE Buffer. These DNA samples were then stored in a freezer for future use.

PCR PRIMERS

Due to the nature of the DNA sequence similarities very specific regions had to be selected and rigorously tested to assure that both gene copies were not amplified. Due to the length limitations on Eton Bioscience capillary sequencing (700bp per reaction), several of the genes had to be sequenced in segments using primers that would produce overlapping products. Once suitable regions were mapped, Primer-BLAST from ncbi.nlm.nih.gov was used to determine the best primers for the region and to check for false hits throughout the genome. Primers were designed for multiple uses. External primers lie outside of the coding region and can be used together to amplify the entire gene region. Internal primers are based in exons so that they can be used for expression analysis if possible. And using an external primer with the opposite internal primer allows for amplification of a product that would be smaller than the entire gene but contain similar sequence to the opposite pairing of primers, this

allows for sequencing of the larger genes (>1400bp) in smaller segments for higher quality. LBb1.3 was a primer recommended by the Salk Genotyping Project in order to detect the tDNA insertion for the mutant lines.

SEQUENCING AND ANALYSIS

PCR products were sent to Eton Biosciences for capillary sequencing using the primers mentioned earlier. Sequence histogram (.ab1) files were analyzed for sequence integrity and the DNA sequences were mapped out based on overlapping sequence and concatenated into complete sequences formatted in FASTA for analysis. FASTA files of all sequences were aligned using Muscle, then edited for length for further analysis. Sequences were further analyzed using Mega6 for phylogenetic tree formation and DnaSP and statistical analysis (Librado and Rozas, 2009). The Neighbor-Joining trees were constructed using distances computed by the Jukes-Cantor method assessed by bootstrap (1,000 replicates) (Felsenstein, 1985; Jukes and Cantor, 1969; Saitou and Nei, 1987; Tamura et al., 2013). All codon positions were included and positions containing gaps or missing data were removed.

TDNA MUTANTS

We obtained our mutant lines from the ABRC stock center. SALK_06954.55.00.x (ML68) contains a tDNA insertion within the coding sequence, reportedly in an exon. However when we analyzed the placement by sequencing, we determined that it is rather inserted 40bp upstream of the transcription start site of the gene. This however is well placed for disruption of expression.

PHENOTYPE ANALYSIS

Phenotyping analysis was performed by growing mutant plants alongside control Col-0 plants in a growth chamber. 2 sets of conditions were used; 12h (25° C) Day 12h (20° C) night, and 16h (22° C) day/8h (18° C) night. We noted the timing of their various developmental stages, comparing the developmental timings between the two groups. These stages included cotyledon formation, inflorescence emerging, bolting, flowering, silique development, silique shattering and senescence. Development times were analyzed by Two Factor ANOVA using SPSS.

Figure 5: Developmental stages noted during phenotype analysis.

Top: Cotyledon (Left), Inflorescence (Center), Bolting (Right)

Bottom: Flowering (Left), Silique Development (Center), Senescence (Right)

RESULTS

GINS COMPLEX GENES

The *AT3G55490* and *AT1G19080* sequences are remarkably similar (>99% sequence similarity) based on the published reference genomes. This similarity prevents us from studying the expression of these genes with the techniques available to us. However the sequences flanking the genes seemed sufficiently polymorphic to allow us to amplify each specifically. This worked well for *AT3G55490*; however the new gene *AT1G19080* seems to have been the

target of a rather large deletion event which is found throughout a multitude of populations (Figure 6). These deletions were found in populations 16, 17, 20, 22, 40-44, 49, 52, 62, 64, and 68. These deletion carrying populations are also not geographically isolated; they exist in almost all of the locals from which our stocks originate. The deletion is exactly the same in all of the populations with this feature, so it is likely derived from a common ancestor. Also note that we were unable to sequence 19 of the 80 lines we worked with, so this deletion is present in a significant number of the populations assayed (21%). Due to the constraints of the analytical tools, we had to ignore the length variant sequences when conducting our statistical analyses. The duplicate gene also appears to contain a multitude of mutations in the first half of the gene preventing us from sequencing the entire gene for any population outside of the Colombia group. The Colombia group, which is the basis for the reference, appears to have a unique DNA sequence upstream of the new gene. We were therefore able to successfully obtain population line sequences after the 2nd exon; it was with these sequences that we ran our statistical analyses.

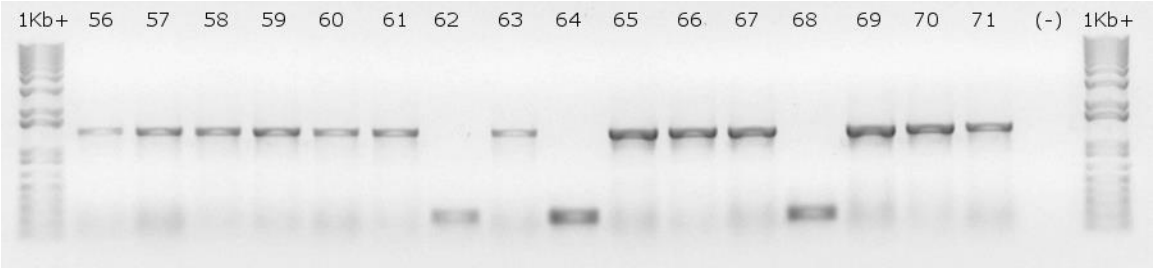


Figure 6: Example population lines amplified with *AT1G19080* cF and R1 primers, photo provided as an example of the size variations seen in 62, 64 and 68.

The negative Tajima's D values and the high neutrality index values (Table 2) suggest that these genes are both under negative selection, or there has been a recent population bottleneck. The parental copy appears to be strongly influenced by this, while the new gene's statistics are not significant. BLAST searching the sequence for *AT3G55490* and *AT1G19080* identify the orthologs in our outgroup species *A. lyrata* (*485976*), *C. rubella* (*Carubv10018083m.g*), and *B. rapa* (*Brara.D02428*). All three outgroup genes score over 93% similarity when compared with the *A. thaliana* genes. As noted earlier the sequences given by the reference genomes are very similar when comparing the coding sequence, we saw a similar invariability in the sequences from the population lines. Other than the deletion events the sequences all seem to be very similar, this is displayed in both the Neighbor-Joining trees generated by the data, as well as the statistical test run which all signify an excess of low frequency alleles. The trees do not show any real clustering by region, displaying the lack of variability in these genes. The *B. rapa* ortholog had to be ignored when aligning the genes for these trees as it appears to have a series of insertions and deletions including multiple frame shifts. This made it impossible to align with the other copies while preserving the integrity of the analysis.

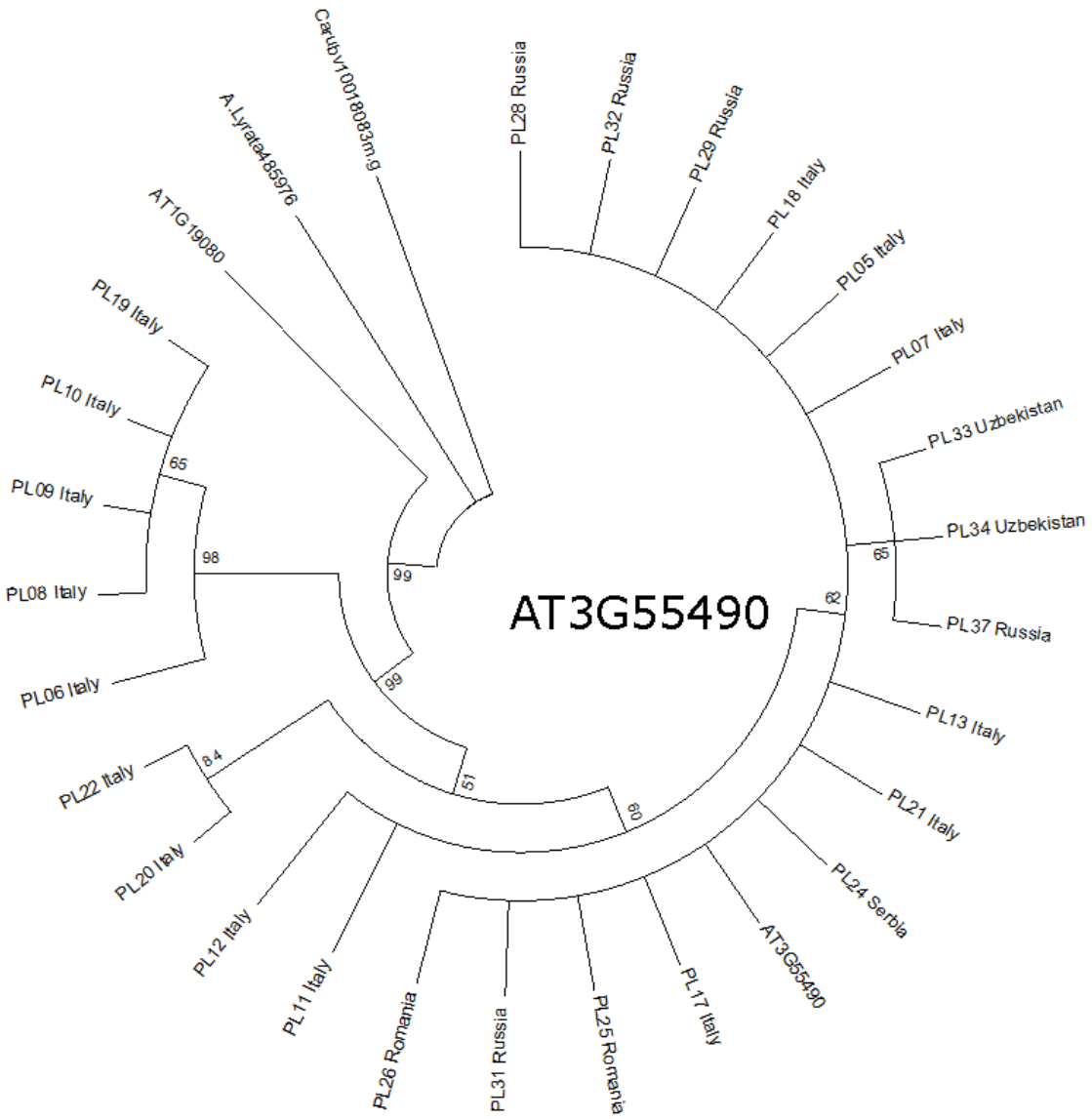


Figure 7: Neighbor-Joining tree created using the *AT3G55490* DNA sequences of the *A. thaliana* populations, with *C. rubella* and *A. lyrata* used as out groups. Note that *B. rapa* is absent due to large deletions in the gene.

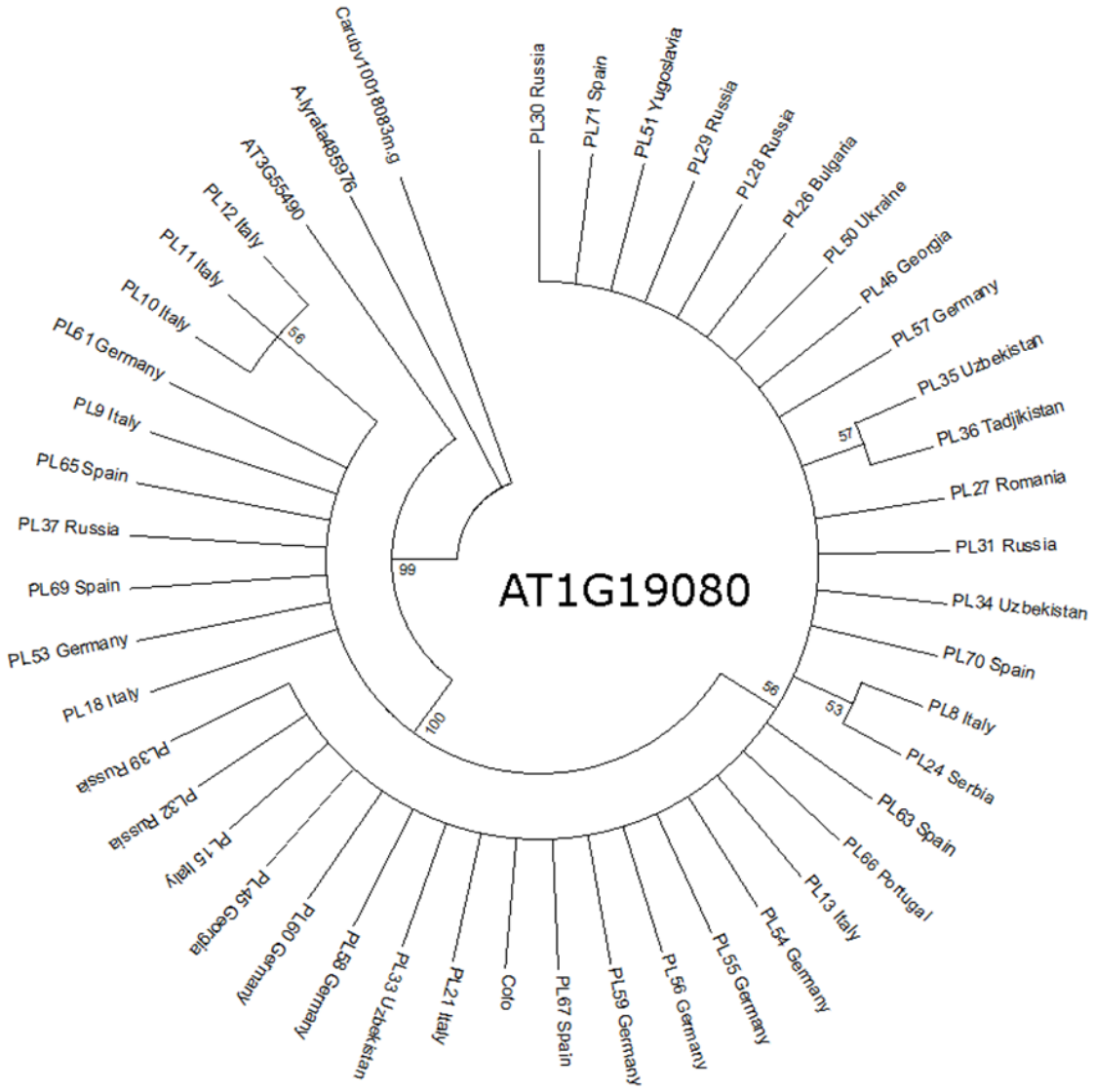


Figure 8: Neighbor-Joining tree created using the *AT1G19080* DNA sequences of the *A. thaliana* populations, with *C. rubella* and *A. lyrata* used as out groups. Note that *B. rapa* is absent due to large deletions in the gene.

PHOSPHORIBOSYLANTHRANILATE ISOMERASE (PAI) GENES

I designed primers to sequence Pai1 and 3, but to specifically exclude Pai2; Pai3 appears to be a copy of Pai1 as its sequence is less similar with Pai2. BLAST searching the sequence for Pai1 and Pai3 identifies the orthologs in our outgroup species *A. lyrata* (*487359)*, *C. rubella* (*Carubv10001692m.g)*, and *B. rapa* (*Brara.B00183*). Interestingly the other species do not have a second copy of the Pai gene, while *A. thaliana* has 3. The outgroup genes also share much more in common with the parental copy (>90% similarity) than they do with the new duplicate (<70%), most likely due to the deletion mentioned earlier. Several sets of external primers were designed for Pai1, however no sequencing for the gene ever came back without significant noise. This could be due to a variety of circumstances. The presence of a near identical copy of the gene could be a large factor in this error. Also as previously noted the fact that the reference sequence is taken from the Colombia genome and not from wild populations may offer a small sample of the variance in the flanking sequence. The flanking sequence before and after Pai1 is highly varied from that of Pai2, which leads me to believe that the sequence is not fixed in the populations. Large sequence variations in the population lines not present in the reference makes it difficult to create primers universal to the populations and unique to Pai1. In order to align the coding region of the new gene to the others for statistical analysis, all of the coding sequence beyond the frameshift had to be ignored. Inclusion of this sequence would force sequence in the outgroup introns to be treated as exons and would skew results. Due to this restriction only the coding regions shared by

Pai3 and the other copies were used for the phylogenetic tree, and only the matching loci in *A. lyrata* (*487359)* were included in the calculations for Tajima's D and the MK tests (Table 2). Once again a strongly negative Tajima's D value of -1.90267 suggests that this new gene is also under negative selection, or adds support to the theory of a recent population bottleneck.

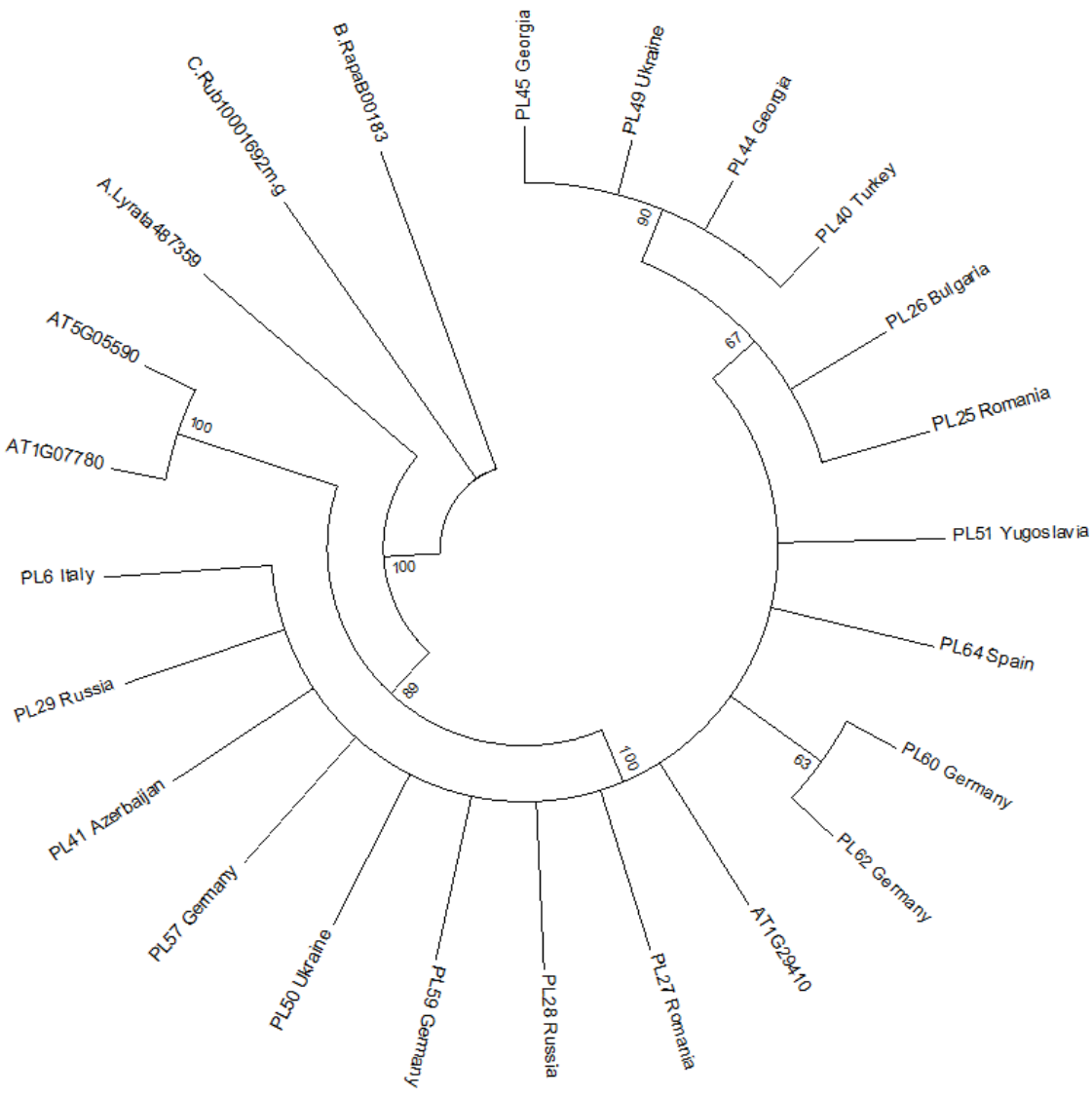


Figure 9. Neighbor-Joining tree created using the *AT1G29410* (Pai3) DNA sequences of the *A. thaliana* populations, *AT1G07780* (Pai1) reference, *AT5G05590* (Pai2) reference, with *A. lyrata, C. rubella,* and *B. rapa* used as outgroups.

Expression analysis based on my assay displays differential expression. However the differential expression in this study differs from the expression pattern hinted at in the mass study using the published data. The published data claims that Pai1 is enriched in inflorescence while Pai3 is enriched in silique

(Wang et al., 2013). Neither seems to express in the silique. Also the parental gene seems to be inactive in the leaf and expressed weakly in the root, while the new gene is active in the leaf and inactive in the root. This expression profile divergence is somewhat perplexing if the new gene is non-functional.

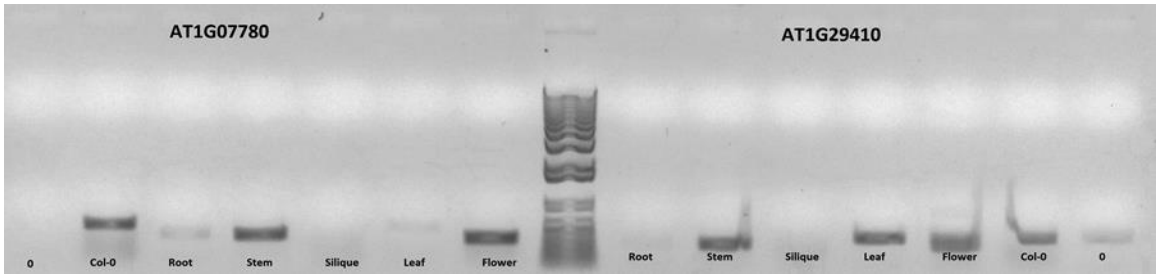


Figure 10: Expression analysis: PCR amplification of cDNA using internal primers for the *AT1G07780* (Left) and *AT1G29410* (Right). Positive control located in appendix.

CHLOROPLAST PROTEINS OF UNKNOWN FUNCTION

These genes are similar to the Gins pair in the fact that they are very similar in coding sequence and exon structure to one another. The sequences are however dissimilar enough for us to do a complete analysis on each independently. While these genes maintain their structure for the most part, it must be noted that one population (PL40) contains a highly mutated *AT2G5310* which contains a 447bp insertion in intron 2, and a deletion in exon 3. This is counter to the cases in the Gins set where the large insertion and deletions appear to be limited to the new copy. However due to the similarity between the two gene copies, it is likely that the intact new duplicate gene would be able to function on its own. The only significant statistic for these two genes is the 6.071 neutrality index for the new gene (Table 2). This high value is once again

suggestive of negative selection or a population bottleneck. Neither of the statistics for the parental copy carried any significance. The comparisons were conducted using the outgroup orthologs *A. lyrata (480141), C. rubella (Carubv10014914m.g),* and *B. rapa (Brara.G00684.1).*

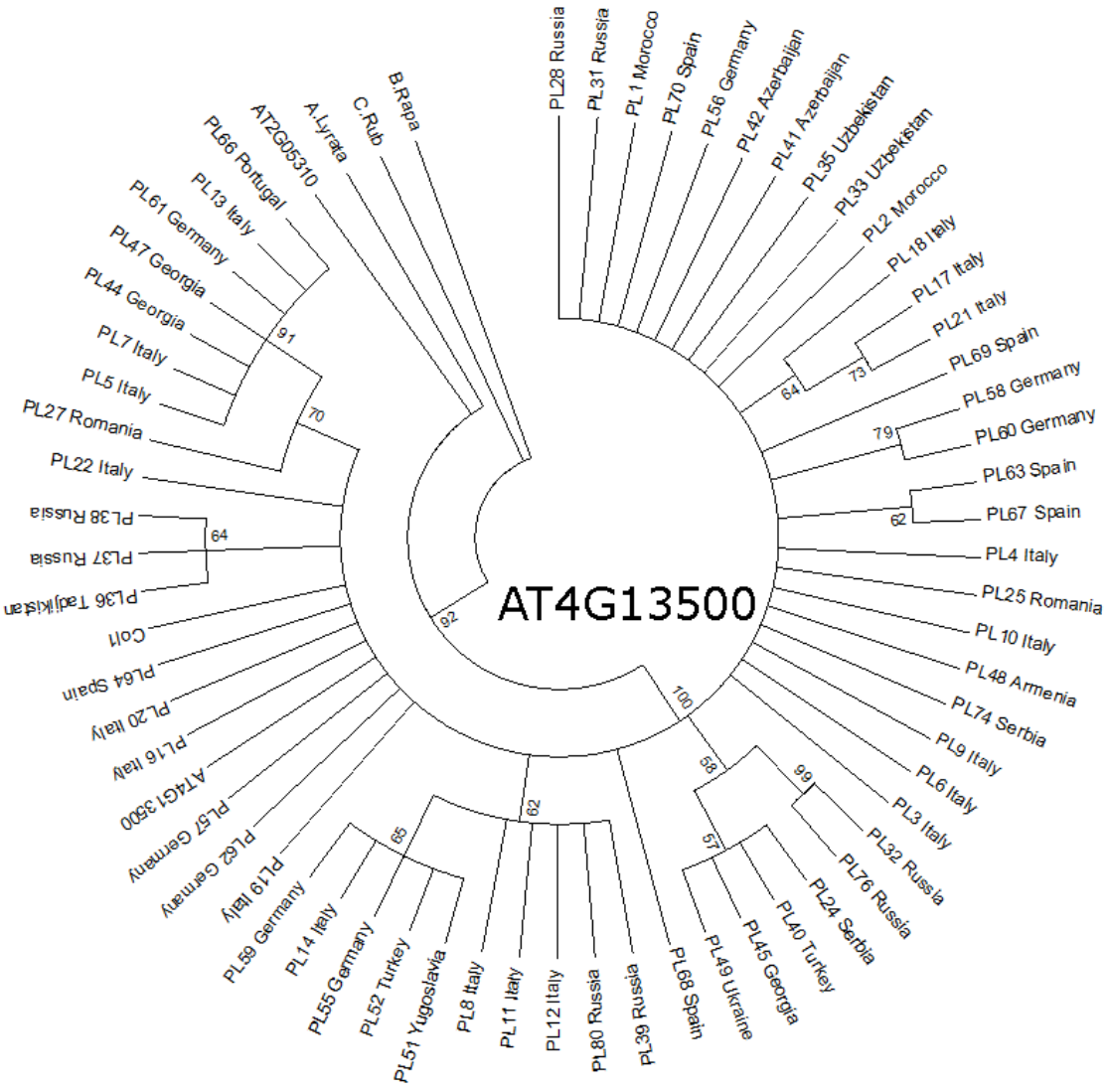


Figure 11: Neighbor-Joining trees created using the *AT2G05310* DNA sequences of the *A. thaliana* populations, with *A. lyrata, C. rubella,* and *B. rapa* used as outgroups.

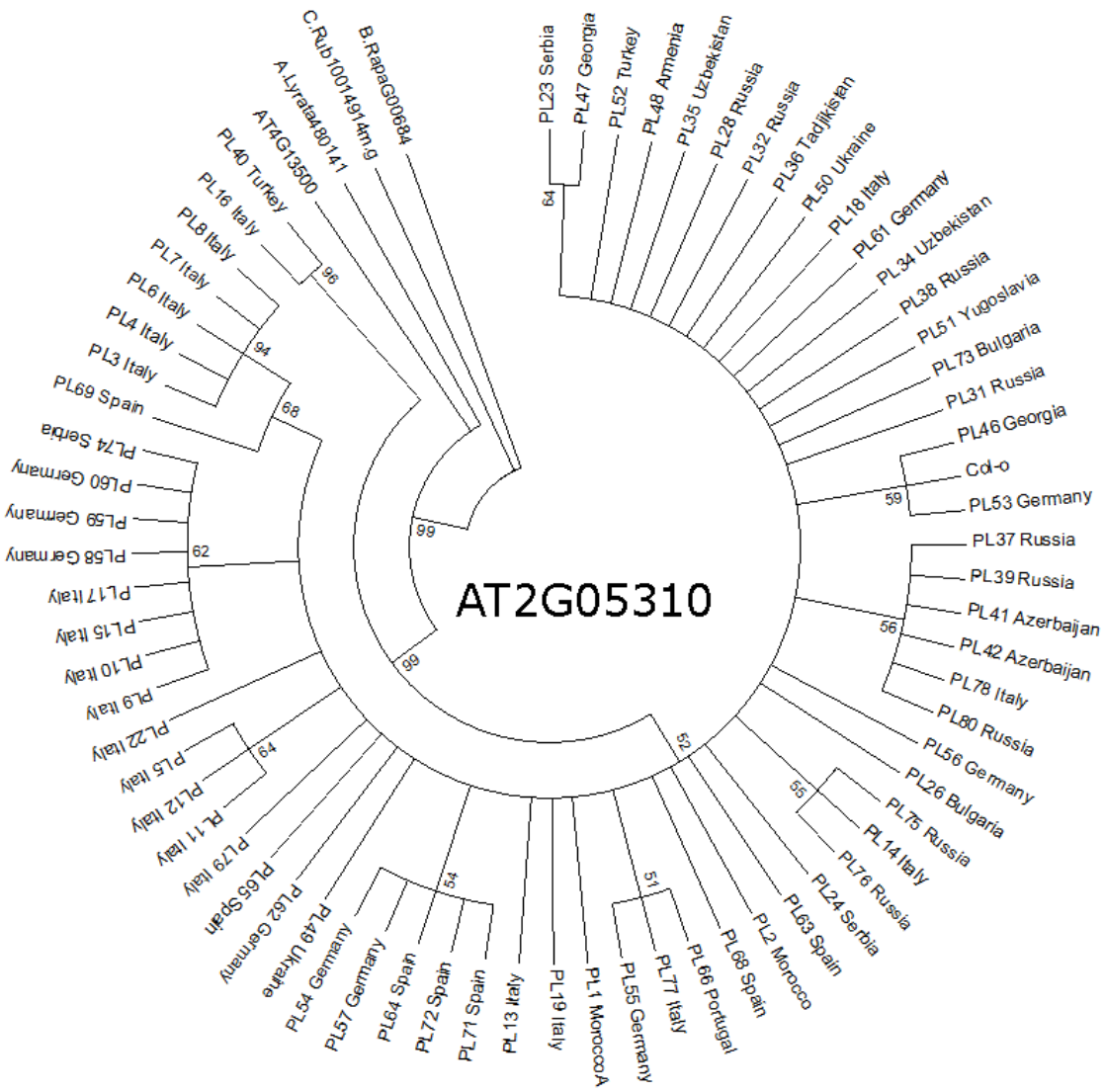


Figure 12: Neighbor-Joining trees created using the *AT4G13500* DNA sequences of the *A. thaliana* populations, with *A. lyrata, C. rubella,* and *B. rapa* used as outgroups

EXPRESSION

Based on our assay the expression profiles of these two genes do not appear to differ at all. This goes against the previously noted expression data showing that *AT2G05310* should have enriched expression in the flowers (Wang et al., 2013).

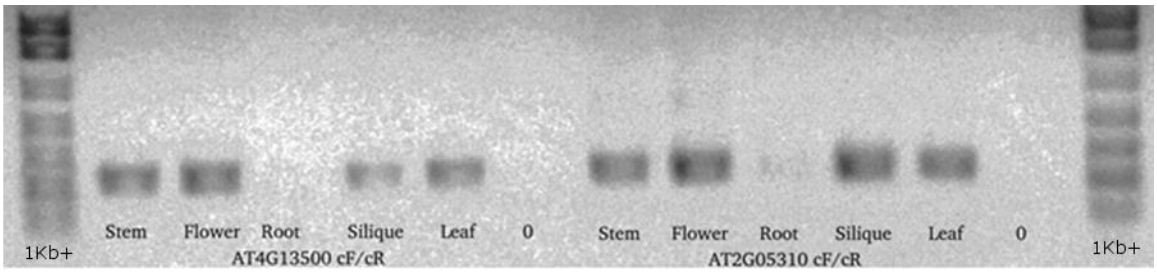


Figure 13: Expression of the new gene *AT4G13500* (Left) and its parental paralog *AT2G05310* (Right) produced by PCR of A. thaliana cDNA using internal primers for the respective genes Positive control located in appendix.

MUTANT LINE

We obtained our mutant lines from the ARBC stock center. The mutant line that seemed best fitted to our purposes was SALK_06954.55.00.x. This line reportedly contains a tDNA insertion in an intron of the new gene AT4G13500. However our sequencing data shows that the insertion appears to be 40bp upstream from the gene, which should be sufficient to knock out function of the gene. The tDNA insertion was confirmed by PCR and sequencing using a gene specific primer set and primer LBb1.3 supplied by SALK. Samples from every generation of mutants were randomly selected alongside control Col-0 individuals and screened by PCR. Mutants displayed a small ~500bp fragment, which was sequenced to confirm presence of the tDNA insertion. Homozygotes displayed 2 products, one mutant band and one band with similar size to the Col-0 control ~2000bp in size. All of the individuals tested in these studies were homozygous for the insertion.

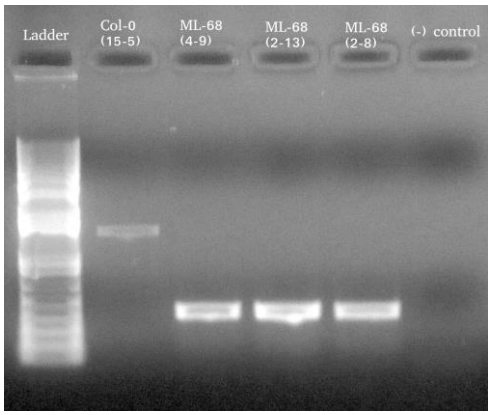
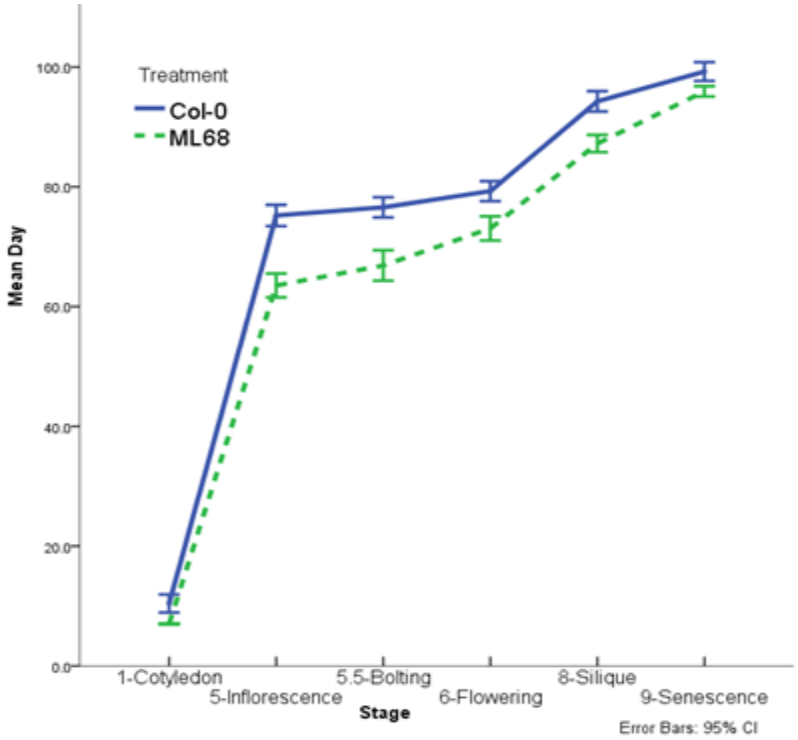


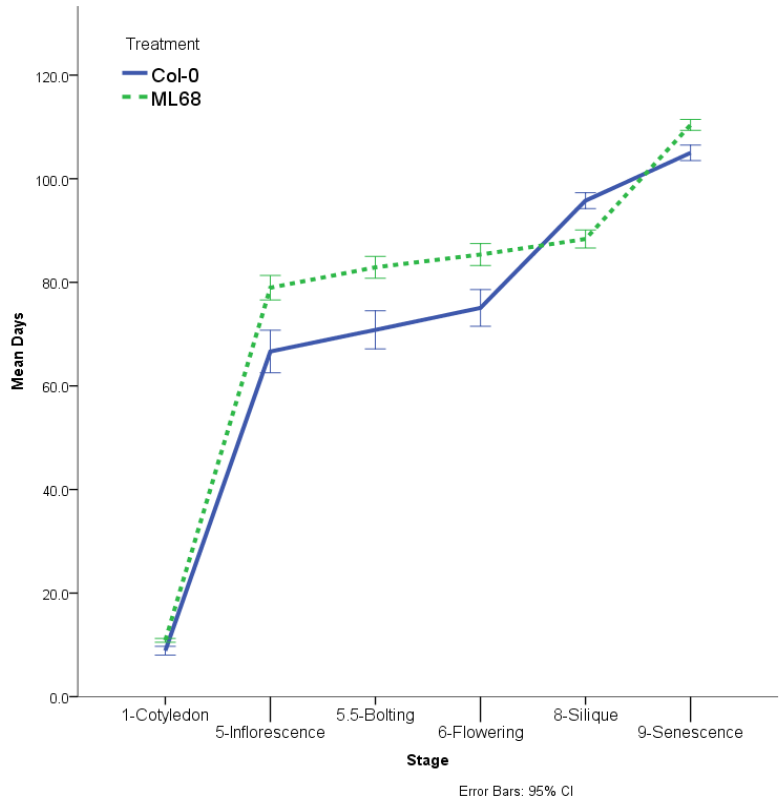
Figure 14: PCR product screening ML68 samples using *AT4G13500* external primers and LBb1.3 to target the tDNA insertion. All tested individuals appear homozygous.

PHENOTYPING

Mutant plants containing the tDNA insertion were grown alongside Col-0 control plants under the same conditions. We ran 4 assays under 2 different condition sets. The first set of conditions was 12h (25° C) days/ 12h (20° C) nights. Under these conditions we noticed a substantial delay in development times for the mutant strain. In the first run it took roughly 10% longer for the mutant plants to come to each developmental stage, although their average development times seem to converge after flowering and actually cross over during silique formation. Our second assay 16h (22° C) day/8h (18° C), while still presenting us with a differential development time for the mutant; displayed the exact opposite trend. We noted a delay in the Col-0 development, while the mutant line stayed in line with the development times for the control group from the previous assays.

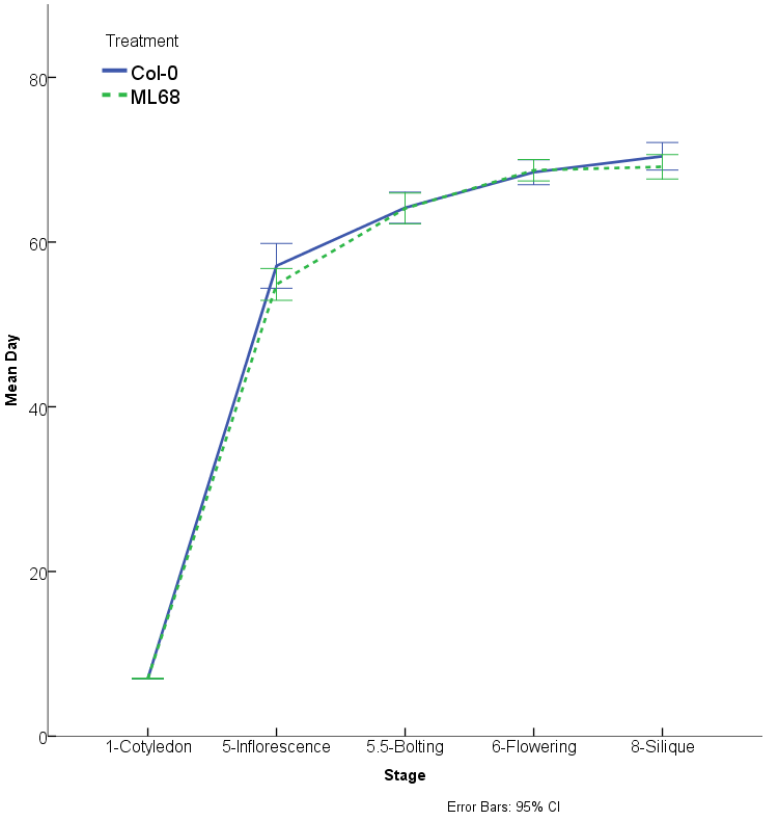
Figure 15: Average time to develop, first 12/12 Day Night cycle trial (Top) and first 16/8 cycle (bottom). ANOVA tables are located in the appendix (Tables 4-5).

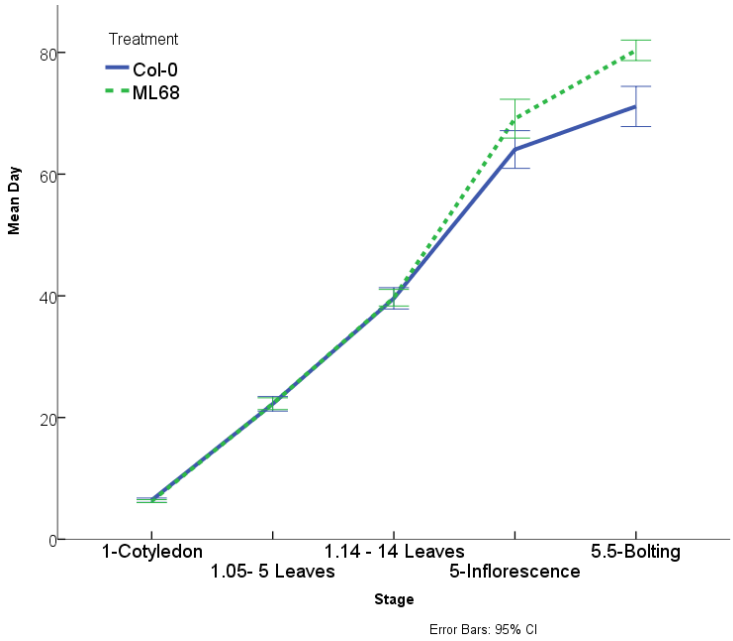




However upon replication these effects seem to be non-repeatable. We ran the same set of assays for a second time and in these seem to offer no discernable difference in growth rate in the 12/12 day/night conditions, and the 16/8 cycle is ongoing as of this writing but seems to contradict the previous assay as the control plants seem to be developing more quickly. So unfortunately these data do not offer us any insight into the workings of this new gene, or its parent as these growth differences seem to follow no discernable pattern. No noticeable visible phenotype differences were observed when comparing ML68 individuals to the control plants. This outcome could be expected, as the genes are so similar to one another that they likely are interchangeable in their operations in the cell. It was possible that reducing the availability of one or the other would have an effect, but it appears that they do not.

Figure 16: Average time to develop, second trial set. 12/12 day/night cycle (Top), 16/8 day/night cycle (Bottom). ANOVA tables are located in the appendix (Tables 6-7).





DISCUSSION

Gene duplication is a very powerful tool for increasing the size of the genome, however as it works by relatively random processes these events are most likely detrimental or neutral to the organism's survival. When a gene is duplicated the organism could now experience a new evolutionary pressure on it, while the new gene is tested as a possible aid or detriment to the organism. By sequencing and comparing new duplicate genes throughout different populations we can begin to understand these pressures and how they affect the genomes of the organism.

Looking at the gene sets, we can see some patterns in the data that can lead us to some interesting conclusions and possible explanations. Firstly, the parental genes were defined by their existence in a syntenic region similar to the outgroup gene. In the two sequenced sets we see two different outcomes. The *AT3G55490* parental gene displays a higher rate of mutation compared with its heavily deleted duplicate gene. Where in *AT2G05310* and *AT4G13500* we observe more divergence in the duplicate gene, compared to the parental gene. All of the genes exhibited negative Tajima's D values and the MK tests all report negative α values and neutrality indexes greater than 1, however not all of the statistics are significant (Table 2). Those that are suggest a couple possible conclusions; negative selection, recent bottlenecks, or selective sweeps (Eyre-Walker, 2002; Tajima, 1989). Whole genome population analyses have established that *A. thaliana* genes en masse tend to skew toward high rates of low frequency alleles (Cao et al., 2011; Nordborg et al., 2005). These studies

similarly posit that a likely explanation is a population bottleneck, although they concede that it is not the only possible interpretation of the data. There have been studies showing evidence of a possible *A. thaliana* population bottleneck within the last roughly 17,000 years due to glacial activity in Europe (Sharbel et al., 2000), however this is not the only theory.

Somewhat interesting are the presence of the deletions in the new genes for Pai and Gins families. Why do single nucleotide polymorphisms appear to be limited while large indels including frameshifts are common? I can only speculate, but it may be a lack of tolerance for sub-optimal functioning for these genes which are involved in tryptophan synthesis and DNA replication. Perhaps mutating the individual nucleotides could be problematic, the malfunctioning mutant copy could impede the functional protein and disrupt the system. The only way to get rid of the new gene without reducing the fitness of the organism may be to delete it. As I have no evidence of the mutant genes actually impeding the wild type, I do admit that this is simply a theory. However it is interesting how these two cases are so similar. Although the deletion in Pai3 seems to be fixed, while the deletions in the new Gins gene exist in around a quarter of the populations assayed.

The chloroplast pair appears to differ from the other sets. In this pair both paralogs seem to be relatively intact in all populations studied; the only notable example was PL40 which contains a mutated and partially deleted parental gene. This gene pair also seems to have been affected by selective sweeps as the nucleotide diversity is relatively low. So while the other new genes seem to be

targeted by deletions, this gene seems to be relatively stable. This gene pair could be evidence of redundant gene duplication. It would be interesting however, if the new gene were to persist while the parental copy becomes a pseudogene, although this case only exists on one population thus far so it is far from a trend.

APPENDIX

Table 2: Statistical Analyses conducted using DnaSP (* 0.01<P<0.05; ** 0.001<P<0.01; *** P<0.001)

Tajima's D	*AT3G55490* (Parent)	*AT1G19080* (NDG)	*AT1G29410* (NDG)	*AT2G05310* (Parent)	*AT4G13500* (NDG)
Sequences Used	37	44	18	67	62
Total # of Sites	1633	1098	2367	890	1319
# of Polymorphic Sites	72	14	38	34	81
# of Mutations, η	75	15	38	34	82
Avg. # of differences, k	8.38288	2.32981	5.88235	5.53053	12.33210
Nucleotide diversity, π	0.00513	0.00212	0.00273	0.00621	0.00935
Θ (per sequence) from η	17.96597	3.44828	11.04795	7.12127	17.46069
Θ (per site) from η	0.011	0.00314	0.00513	0.00800	0.01324
Tajima's D	-1.9566*	-1.01660 (NS)	-1.90267*	-0.72199 (NS)	-1.01020 (NS)
Syn polymorphisms	1	4	3	5	4
Non-Syn polymorphisms	7	1	6	8	10
McDonald–Kreitman					
Neutrality Index, NI	20.000	2.875	1.565	1.371	6.071
α value	-19.000	-1.875	-0.565	-0.371	-5.071
G-value	9.888	0.562 (NS)	0.345	0.158	6.531
P-value	0.00166**	0.45326 (NS)	0.55687(NS)	0.69124 (NS)	0.01060*

PCR

For the polymerase chain reaction we used Choice Taq supplied by Denville Scientific. Our typical reaction consisted of: 0.5µl DNA, 0.1µl ChoiceTaq DNA polymerase, 0.2µl per Primer, 0.4µl dNTP mix 100mmol, 2.0µl 10x PCR Reaction Buffer (w/ Mg2+), and ddH20 was then added to bring the final volume to 20ul

Biorad T100 Thermo cycler program:

94° C for 3:00

94° C for 0:30

55° C for 0:30(annealing temp was varied based on primer annealing temp)

72° C for 1:00(time varied due to length of product ~1 minute per Kb of target)

------Repeat steps 2-4 34x-------

72° C for 10:00

4° C Hold

Table 3: Primers used for PCR amplification and sequencing.

Gene Name	External Primers			Internal Primers
AT3G55490	F1	TTCGGGCTTCAATAGAGCTG	cF	AAAACAGCAAACGGAGTGAC
	R1	AGCGGATCCTCACAACTTC	cR	CCTGAACTTCAAGCCTCGTC
AT1G19080	F1	TCCCACGTGTACCCTCTCTT	cF	Coding sequence too similar used same primers as AT3G55490
	R1	GAAGCAGAAGAGAATGATTCCA	cR	
AT1G07780	N/A		cF	ATCCACTGATCTCCATGTCCA
			cR	CAGCTGCTCTCAGTATCGTGT
AT1G29410	F2	CCATATCACACTCTGTCCTTTTG	cF	TATTTCCCAAGTGGCCAGGG
	R2	GCCCATAATTCGATCGTCTC	cR	ACTCGTTCCTCGAAGGAATACA
AT2G05310	F1	GCCAAATGTCACTACAAATGC	cF	GGCAGCAAAACTCTGCATTC
	R1	TTGGAGAGATGACCAATGTTTG	cR	GGCGTTGACATACCGAAGAT
AT4G13500	F1	GGTTGACTTTTCAATCCTGA	cF	CGTTTAGCCACAGGGCTAGT
	R1	TTCACATGTACTTACAAACAAAAA	cR	GGCAAAAGCAATGGCTAAGA
Actin1	N/A		cF	TACAATGAGCTCCGTGTTGC
			R	CACGACCAGCAAGATCAAGA
LBb1.3	N/A			ATTTTGCCGATTTCGGAAC

GEL ELECTROPHORESIS

PCR Electrophoresis was performed using 1% agarose gels made with peqGold Universal Agarose supplied by PeqLab and 0.005% by volume Ethidium Bromide 1% solution. Buffer was made by 10mM Tris-HCl and 10mM EDTA.

10X Loading Dye: 0.5% bromophenol blue, 0.5% xylene cyanol FF, 50% glycerol, 1Kb+ used as ladder.

SOLUTIONS

CTAB Buffer: 100 ml 1 M Tris HCl (pH 8.0), 280 ml 5 M NaCl, 40 ml of 0.5 M EDTA, 20 g of CTAB (cetyltrimethyl ammonium bromide). Total volume brought to 1 L with ddH_2O. Then to each .5ml aliquot .02g polyvinylpyrrolidone and 2.5 µl β-mercaptoethanol was added before use.

Figure 17: RT-PCR Control using Actin1 Primers.

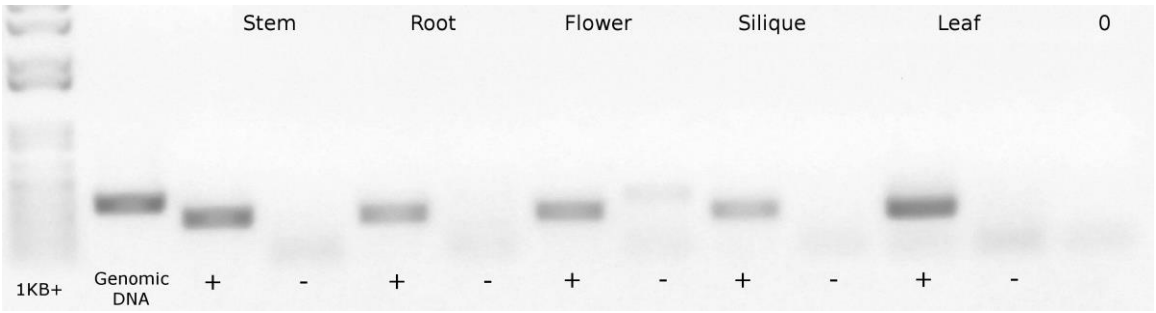


Table 4: ANOVA Table for Phenotype Assay 1 (12/12 Night/Day)

Tests of Between-Subjects Effects

Dependent Variable: Days

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	270091.818[a]	11	24553.802	968.042	.000
Intercept	1397466.620	1	1397466.620	55095.621	.000
Stage	243888.783	5	48777.757	1923.080	.000
Treat	2176.379	1	2176.379	85.805	.000
Stage * Treat	3159.024	5	631.805	24.909	.000
Error	6848.384	270	25.364		
Total	1823393.000	282			
Corrected Total	276940.202	281			

a. R Squared = .975 (Adjusted R Squared = .974)

Table 5: ANOVA Table for Phenotype Assay 2 (16/8 Night/Day)
Tests of Between-Subjects Effects

Dependent Variable: Day

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	349317.097[a]	11	31756.100	1348.748	.000
Intercept	1970365.171	1	1970365.171	83685.539	.000
Stage	343130.602	5	68626.120	2914.695	.000
Treatment	4909.084	1	4909.084	208.499	.000
Stage * Treatment	973.646	5	194.729	8.271	.000
Error	9465.038	402	23.545		
Total	2324320.000	414			
Corrected Total	358782.135	413			

a. R Squared = .974 (Adjusted R Squared = .973)

Table 6: ANOVA Table for Phenotype Assay 3 (12/12 Night/Day)

Tests of Between-Subjects Effects

Dependent Variable: Day

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	191817.725[a]	9	21313.081	948.563	.000
Intercept	941196.778	1	941196.778	41889.023	.000
Treatment	38.008	1	38.008	1.692	.194
Stage	191541.625	4	47885.406	2131.194	.000
Treatment * Stage	77.203	4	19.301	.859	.489
Error	7324.834	326	22.469		
Total	1114146.000	336			
Corrected Total	199142.560	335			

a. R Squared = .963 (Adjusted R Squared = .962)

Table 7: ANOVA Table for Phenotype Assay 4 (16/8 Night/Day)

Tests of Between-Subjects Effects

Dependent Variable: Day

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	172702.886[a]	9	19189.210	811.947	.000
Intercept	442622.590	1	442622.590	18728.551	.000
Treatment	508.670	1	508.670	21.523	.000
Stage	170168.273	4	42542.068	1800.069	.000
Treatment * Stage	874.193	4	218.548	9.247	.000
Error	5672.058	240	23.634		
Total	622910.000	250			
Corrected Total	178374.944	249			

a. R Squared = .968 (Adjusted R Squared = .967)

Table 8 continued: List of A. thaliana accession lines used in population analysis.

code	Acc #	Name	Abb Name	Other Names	Country
PL1	CS76347	Ait Barka	Aitba-2	ice49, N76347	Morocco
PL2	CS76348	Toufliht	Toufl-1	ice50, N76348	Morocco
PL3	CS76349	Vezzano	Vezzano-2	ice226, N76349, Vezzano-2.1	Italy
PL4	CS76350	Vezzano	Vezzano-2	ice228, N76350, Vezzano-2.2	Italy
PL5	CS76351	Rovero	Rovero-1	ice216, N76351	Italy
PL6	CS76352	Voeran	Voeran-1	ice79, N76352	Italy
PL7	CS76353	Altenburg	Altenb-2	ice163, N76353	Italy
PL8	CS76354	Mitterberg	Mitterberg-1	ice181, N76354	Italy
PL9	CS76355	Castel Feder	Castelfed-4	ice212, N76355, Castelfed-4.1	Italy
PL10	CS76356	Castel Feder	Castelfed-4	ice213, N76356, Castelfed-4.2	Italy
PL11	CS76357	Bozen-Guntschnaberg	Bozen-1	ice169, N76357, Bozen-1.1	Italy
PL12	CS76358	Bozen-Guntschnaberg	Bozen-1	ice173, N76358, Bozen-1.2	Italy
PL13	CS76359	Cisterna de Latina	Ciste-1	ice97, N76359	Italy
PL14	CS76360	Cisterna de Latina	Ciste-2	ice98, N76360	Italy
PL15	CS76361	Montesano Scalo	Monte-1	ice111, N76361	Italy
PL16	CS76362	Sant Angelo	Angel-1	ice91, N76362	Italy
PL17	CS76363	Morane	Moran-1	ice112, N76363	Italy
PL18	CS76364	Mammola	Mammo-2	ice107, N76364	Italy
PL19	CS76365	Mammola	Mammo-1	ice106, N76365	Italy
PL20	CS76366	Ponte Angitola	Angit-1	ice92, N76366	Italy
PL21	CS76367	Rocigliano-Lago	Lago-1	ice104, N76367	Italy
PL22	CS76368	Sant Piedro Apostolo	Apost-1	ice93, N76368	Italy
PL23	CS76369	Dobranovci	Dobra-1	ice36, N76369	Serbia
PL24	CS76370	Petrovac	Petro-1	ice21, N76370	Serbia
PL25	CS76371	Lechovo	Lecho-1	ice7, N76371	Romania
PL26	CS76372	Jablokovec	Jablo-1	ice33, N76372	Bulgaria
PL27	CS76373	Bolintin Vale	Bolin-1	ice1, N76373	Romania
PL28	CS76374	Shiguljovsk	Shigu-2	ice72, N76374	Russia
PL29	CS76375	Shiguljovsk	Shigu-1	ice71, N76375	Russia
PL30	CS76376	Kidrjasovo	Kidr-1	ice73, N76376	Russia
PL31	CS76377	Stepnoje	Stepn-2	ice60, N76377	Russia
PL32	CS76378	Stepnoje	Stepn-1	ice61, N76378	Russia
PL33	CS76379	Sijak	Sij-1	ice150, N76379	Uzbekistan
PL34	CS76380	Sijak	Sij-2	ice152, N76380	Uzbekistan
PL35	CS76381	Sijak	Sij-4	ice153, N76381	Uzbekistan
PL36	CS76382	Shakdara	Sha	Shahdara, N76382	Tadjikistan
PL37	CS76383	Kolyvanskoe ozero bei Sawwuschka	Koz-2	ice134, N76383	Russia
PL38	CS76384	Kolyvan	Kly-4	ice130, N76384	Russia
PL39	CS76385	Kolyvan	Kly-1	ice127, N76385	Russia
PL40	CS76386	Dogruyol	Dog-4	N76386	Turkey

Table 8: List of A. thaliana accession lines used in population analysis.

code	Acc #	Name	Abb Name	Other Names	Country
PL41	CS76387	Xanbulan	Xan-1	N76387	Azerbaijan
PL42	CS76388	Lerik	Lerik1-3	N76388	Azerbaijan
PL43	CS76389	Istisu	Istisu-1	N76389	Azerbaijan
PL44	CS76390	Lagodechi	Lag2-2	N76390	Georgia
PL45	CS76391	Vashlovani	Vash-1	N76391	Georgia
PL46	CS76392	Bakuriani	Bak-2	N76392	Georgia
PL47	CS76393	Bakuriani	Bak-7	N76393	Georgia
PL48	CS76394	Yeghegis	Yeg-1	N76394	Armenia
PL49	CS76395	Kastel Mountain	Kastel-1	N76395	Ukraine
PL50	CS76396	Kocherov	Koch-1	N76396	Ukraine
PL51	CS76397	Deliblato	Del-10	N76397	Yugoslavia
PL52	CS76398	Nemrut Mountain	Nemrut-1	N76398	Turkey
PL53	CS76399	Eyach	Ey1.5-2	MPI_A24_Acc_941_H02', N76399	Germany
PL54	CS76400	Starzach	Star-8	MPI_A16_Acc_0177_G11', N76400	Germany
PL55	CS76401	Tübingen - Schaal	Tu-Scha-9	N76401	Germany
PL56	CS76402	Niederreutin	Nie1-2	MPI_A26_Acc_1070_C03', N76402	Germany
PL57	CS76403	Tubingen - Scho”nblick 30	Tu-SB30-3	N76403	Germany
PL58	CS76404	Heiligkreuztal 2	HKT2-4	MPI_A17_Acc_0200_A10', N76404	Germany
PL59	CS76405	Tubingen - Wanne	Tu-Wa1-2	N76405	Germany
PL60	CS76406	Rubgarten - 3	Ru3.1-31	N76406	Germany
PL61	CS76407	Tubingen - Volksbank	Tu-V-13	N76407	Germany
PL62	CS76408	Walddorf-Haslach	Wal-HasB-4	N76408	Germany
PL63	CS76409	Aguaron	Agu-1	N76409	Spain
PL64	CS76410	Caldas de Miravete	Cdm-0	N76410	Spain
PL65	CS76411	Donana	Don-0	N76411	Spain
PL66	CS76412	St. Maria d. Feiria	Fei-0	N76412	Portugal
PL67	CS76413	San Leonardo de Yague	Leo-1	N76413	Spain
PL68	CS76414	Merida	Mer-6	N76414	Spain
PL69	CS76415	Pedriza	Ped-0	N76415	Spain
PL70	CS76416	Pradena del Rincon	Pra-6	N76416	Spain
PL71	CS76417	Quintela	Qui-0	N76417	Spain
PL72	CS76418	Viella	Vie-0	N76418	Spain
PL73	CS76419	Slavianka	Slavi-1	ice29, N76419	Bulgaria
PL74	CS76420	Copac	Copac-1	ice63, N76420	Serbia
PL75	CS76421	Borskoje	Borsk-2	ice70, N76421	Russia
PL76	CS76422	Krasnaja Zorka	Krazo-2	ice75, N76422	Russia
PL77	CS76423	Galdo	Galdo-1	ice102, N76423	Italy
PL78	CS76424	Timpo Ulivi	Timpo-1	ice119, N76424	Italy
PL79	CS76425	Valsinnica	Valsi-1	ice120, N76425	Italy
PL80	CS76426	Lebjashje	Leb-3	ice138, N76426	Russia

Figure 18: Neighbor-Joining tree of only *AT3G55490*, no homologs, to accentuate sequence interrelatedness. Bootstrap and branch lengths displayed. Branch Lengths below 0.001 are not displayed. Computed using the same parameters as previous trees.

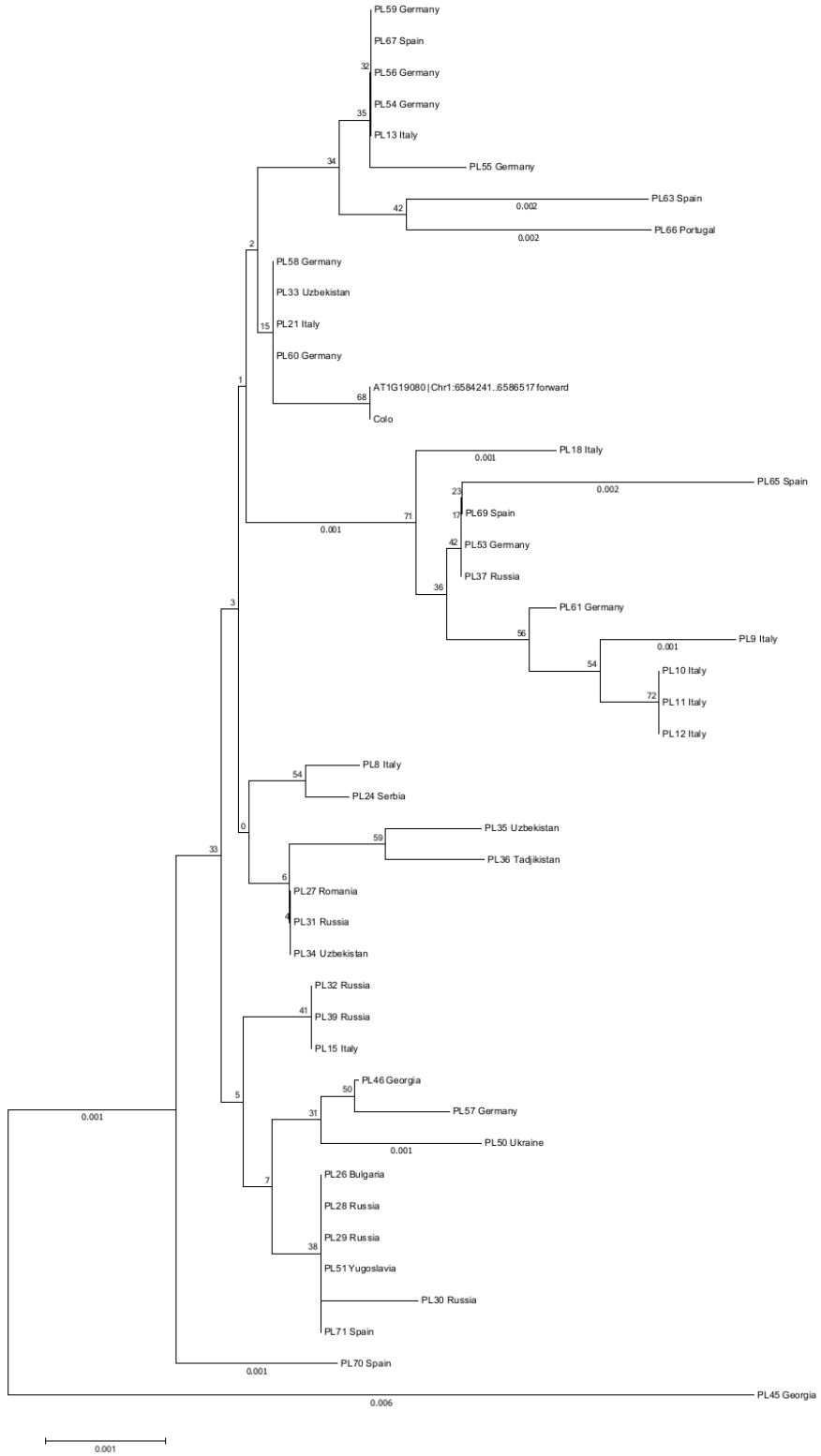


Figure 19: Neighbor-Joining tree of only *AT1G19080*, no homologs, to accentuate sequence interrelatedness. Bootstrap and branch lengths displayed. Branch Lengths below 0.001 are not displayed. Computed using the same parameters as previous trees.

Figure 20: Neighbor-Joining tree of only *AT1G29410*, no homologs, to accentuate sequence interrelatedness. Bootstrap and branch lengths displayed. Branch Lengths below 0.001 are not displayed. Computed using the same parameters as previous trees.

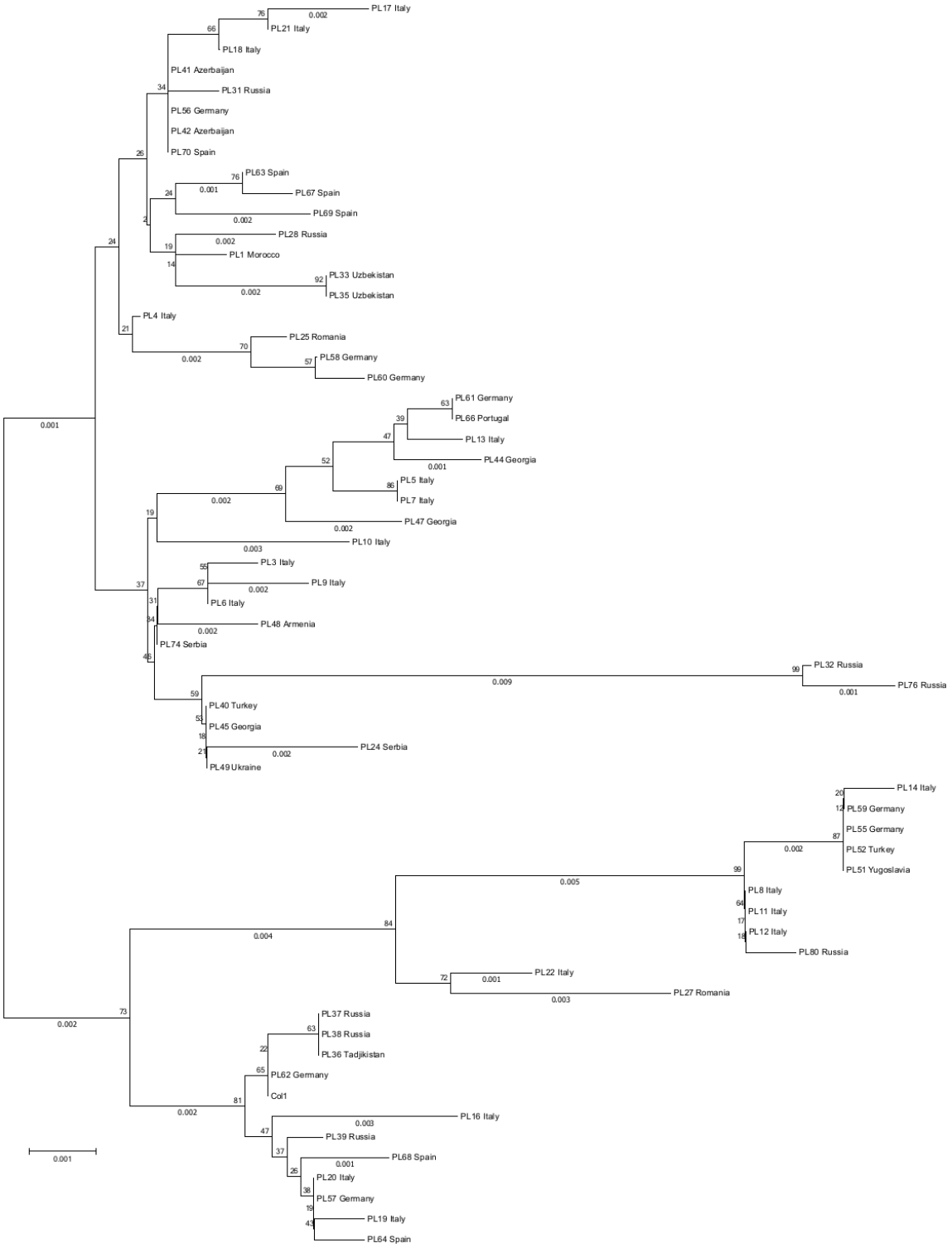


Figure 21: Neighbor-Joining tree of only *AT4G13500*, no homologs, to accentuate sequence interrelatedness. Bootstrap and branch lengths displayed. Branch Lengths below 0.001 are not displayed. Computed using the same parameters as previous trees.

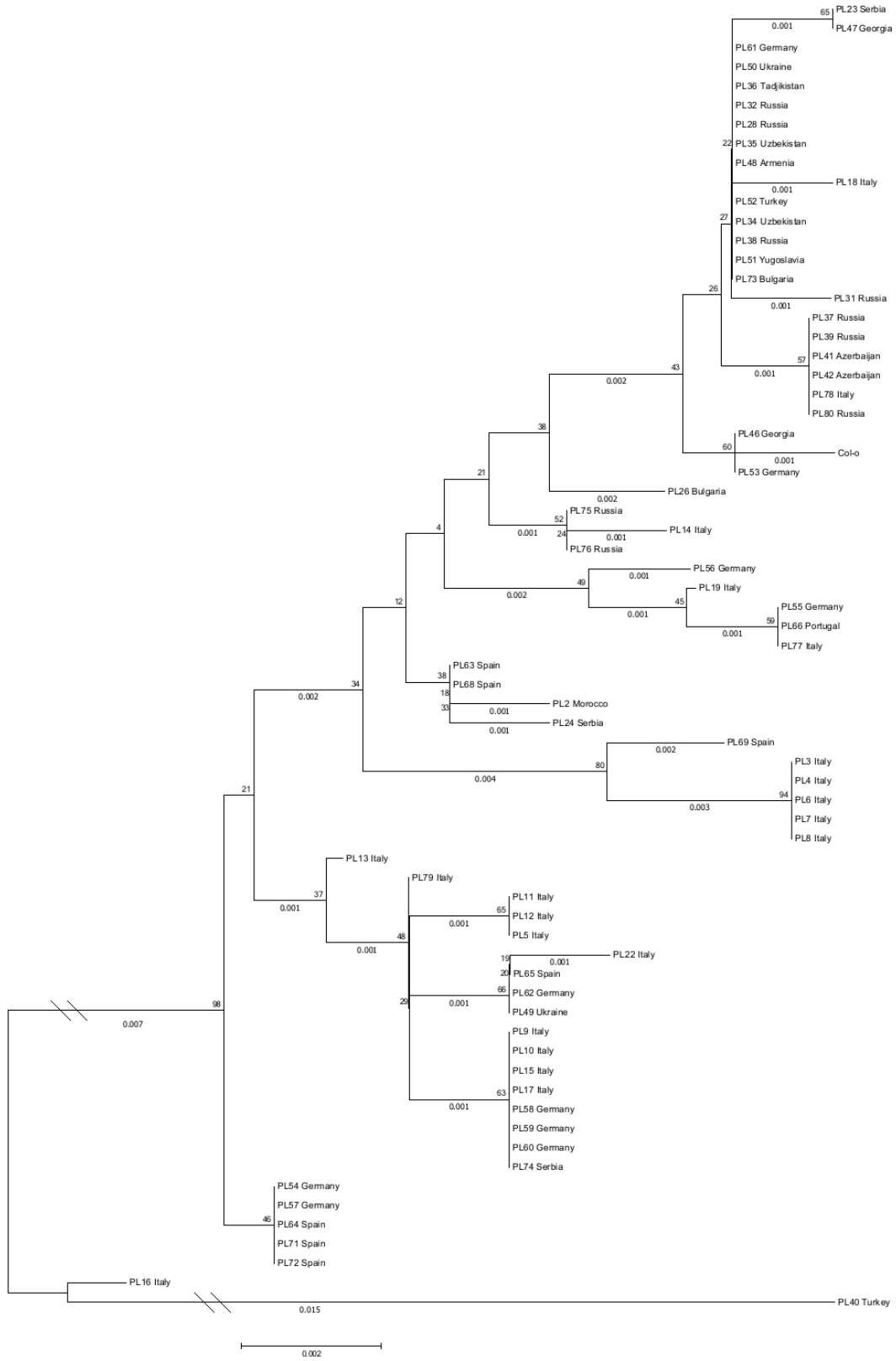


Figure 22: Neighbor-Joining tree of only *AT2G05310*, no homologs, to accentuate sequence interrelatedness. Bootstrap and branch lengths displayed. Branch Lengths below 0.001 are not displayed. Computed using the same parameters as previous trees.

Arabidopsis Genome, I. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature *408*, 796-815.

Bikard, D., Patel, D., Le Mette, C., Giorgi, V., Camilleri, C., Bennett, M.J., and Loudet, O. (2009). Divergent evolution of duplicate genes leads to genetic incompatibilities within A. thaliana. Science *323*, 623-626.

Cao, J., Schneeberger, K., Ossowski, S., Gunther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C.*, et al.* (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet *43*, 956-963.

Eyre-Walker, A. (2002). Changing effective population size and the McDonald-Kreitman test. Genetics *162*, 2017-2024.

Felsenstein, J. (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution *39*, 783-791.

He, Y., and Li, J. (2001). Differential expression of triplicate phosphoribosylanthranilate isomerase isogenes in the tryptophan biosynthetic pathway of Arabidopsis thaliana (L.) Heynh. Planta *212*, 641-647.

Jukes, T.H., and Cantor, C.R. (1969). Evolution of Protein Molecules (Academy Press).

Kleffmann, T., Russenberger, D., von Zychlinski, A., Christopher, W., Sjolander, K., Gruissem, W., and Baginsky, S. (2004). The Arabidopsis thaliana chloroplast proteome reveals pathway abundance and novel protein functions. Curr Biol *14*, 354-362.

Koch, M.A., Wernisch, M., and Schmickl, R. (2008). Arabidopsis thaliana's Wild Relatives: An Updated Overview on Systematics, Taxonomy and Evolution. Taxon *57*, 933-943.

Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics *25*, 1451-1452.

Long, M., Betran, E., Thornton, K., and Wang, W. (2003). The origin of new genes: glimpses from the young and old. Nat Rev Genet *4*, 865-875.

Nordborg, M., Hu, T.T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R.*, et al.* (2005). The pattern of polymorphism in Arabidopsis thaliana. PLoS Biol *3*, e196.

Prince, V.E., and Pickett, F.B. (2002). Splitting pairs: the diverging fates of duplicated genes. Nat Rev Genet *3*, 827-837.

Roux, C., Castric, V., Pauwels, M., Wright, S.I., Saumitou-Laprade, P., and Vekemans, X. (2011). Does speciation between Arabidopsis halleri and Arabidopsis lyrata coincide with major changes in a molecular target of adaptation? PLoS One *6*, e26872.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol *4*, 406-425.

Sharbel, T.F., Haubold, B., and Mitchell-Olds, T. (2000). Genetic isolation by distance in Arabidopsis thaliana: biogeography and postglacial colonization of Europe. Mol Ecol *9*, 2109-2118.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585-595.

Takayama, Y., Kamimura, Y., Okawa, M., Muramatsu, S., Sugino, A., and Araki, H. (2003). GINS, a novel multiprotein complex required for chromosomal DNA replication in budding yeast. Genes Dev *17*, 1153-1165.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol *30*, 2725-2729.

Tautz, D., and Domazet-Loso, T. (2011). The evolutionary origin of orphan genes. Nat Rev Genet *12*, 692-702.

Wang, J., Marowsky, N.C., and Fan, C. (2013). Divergent evolutionary and expression patterns between lineage specific new duplicate genes and their parental paralogs in Arabidopsis thaliana. PLoS One *8*, e72362.

Zhang, J.Z. (2003). Evolution by gene duplication: an update. Trends Ecol Evol *18*, 292-298.

ABSTRACT

EVOLUTION OF NEW DUPLICATE GENES IN ARABIDOPSIS THALIANA

by

NICHOLAS MAROWSKY

August 2015

Advisor: Dr. Chuanzhu Fan

Major: Biological Sciences

Degree: Master of Science

Gene duplication is one of the major mechanisms by which organisms expand their genomes. The material added to the genome can then be acted upon by mutation and natural selection to increase the fitness of the species. By studying these duplicate sequences we can understand the process by which species evolve new functional genes. In a previous paper we identified 100 new duplicate genes through a genome wide comparison between A. thaliana and related species. We selected three of these new duplicate genes and investigated more closely their sequence and expression divergence from their parental gene. The three new duplicate genes selected were *AT1G19080, AT1G29410 and AT4G13500* and their parents *AT3G55490 AT1G07780 and AT2G05310* respectively. These genes were sequenced using *A. thaliana* accession lines from a multitude of locations, and the sequences were used in population analyses. The genes were also tested for differential expression patterns. The genes all show evidence of negative selection or a recent population bottleneck. Notably we detected a large number of populations

carrying deletions for the new genes. The second set (*AT1G07780/ AT1G19080)* displayed differential expression, while the third set shows no divergence. The *AT4G13500*/*AT2G05310* gene family has no known function. In an attempt to discern their function we obtained mutant plants and grew them alongside control plants in an attempt to detect a phenotype for the knockout. We noticed divergent growth patterns between the groups under different light cycles, however they require further testing.

AUTOBIOGRAPHICAL STATEMENT

I, Nicholas Marowsky, have been interested in the sciences since I was a young child, from astronomy and physics to computation and biology. It was however while attending Churchill High School that I became focused on genetics and decided that I must pursue a career in the field. As I delved deeper into biology I became interested in aging, and thus finished my Bachelors of Science at Wayne State working in Dr. Robert Arking's lab. There I worked on delaying senescence using diet and drugs, and studied their effects on lifespan and healthspan. As I spent my work days on these biological problems I found more of my free time was spend learning about computers, and programming. It was only when I applied to the Wayne State graduate program that the department happened to hire some bioinformatics professors. Dr. Chuanzhu Fan was the first to arrive and it seemed like fate, so I joined his lab as soon as possible. Being the first student in a new lab was an interesting experience, I spent much of my time trouble shooting and setting up machines and computer systems. But I learned a lot in the process and still found time to finish my research project, which is the subject of this writing. This study of new genes and their evolution has been very interesting to me, and hopefully will be to others.

I now plan to pursue a PhD, hopefully returning to the field of aging, and I believe that the skills I have developed over these past few years will be indispensable in my future work.