Expression variability following gene duplication facilitates gene retention

Haoran Cai*, David L. Des Marais

Department of Civil and Environmental Engineering, MIT.
15 Vassar St., Cambridge, MA, 02139 USA

ARTICLE TYPE: Brief Communication

LIST OF ELEMENTS: Manuscript, color version of figures 1-5

ONLINE ENHANCEMENTS: Online Supplementary Material

KEYWORDS: Gene duplication; Inter-individual variability; Evolvability

*To whom correspondence should be addressed. hrcai@ucla.edu

Abstract

Genome analyses reveal that gene duplication in eukaryotes is pervasive, providing a primary source for the emergence of new genes. The mechanisms governing the retention of duplicated genes, particularly in the early stages post-duplication, remain ambiguous. Patterns of divergence between duplicated genes vary, leading to biases in the functions of genes retained following duplication. For example, genes that arise from tandem duplication tend to be involved in environmental responses. However, the mechanisms that cause such functional biases remain elusive. Here, to better understand the mechanisms and factors determining gene retention, we study a frequently overlooked molecular feature —expression variability—as measured by within-line expression variation, termed variability. We find that, on average, genes with duplicates exhibit higher expression variability than singletons. We further find that such elevated expression variability for small-scale duplications (SSDs) is likely driven by the immediate increase in expression variability following duplication events. These findings suggest a potential overarching mechanism that facilitates gene expression divergence, functional gains of environmental responses, and duplicate retention following SSDs.

Introduction

Gene duplication is recognized as the primary source of new genes and gene functions and has consequently been credited with great evolutionary importance (Ohno 1970, Des Marais and Rausher 2008). Typically, gene duplication generates two gene copies; this theoretically allows one or both copies to evolve under reduced selective constraint and, on some occasions, to acquire novel gene functions that contribute to adaptation.

Duplicate gene copies can be generated through one of two broad mechanisms, namely small-scale or large-scale duplication events. Among the most dramatic forms of large-scale duplication is whole-genome duplication (WGD), which leads to a sudden expansion in genome size and, significantly, preserves the chromosmal context of most genes (Kuzmin et al. 2022, Wang et al. 2012). WGD have likely occurred multiple times throughout 200 million years of angiosperm evolution (Vision et al. 2000). Conversely, small-scale gene duplications (SSD), including proximal duplication, tandem duplication, and transposed duplication, result in duplication of a specific genomic region, often comprising a single gene (Wang et al. 2012, Kuzmin et al. 2022, Zhang 2003). Plant genomes, on average, harbor duplicate copies for approximately 65% of annotated genes. Most duplicates originate from WGD events and, accordingly, paleopolyploidization has been cited as an important source of evolutionary novelty for land plants (Panchy et al. 2016, Tiley et al. 2016). While WGD events lead to a sudden and massive expansion of the genome, they are often followed by extensive gene loss over evolutionary time (Lynch and Conery 2000). Nevertheless, WGD-derived duplicates tend to be retained at a higher rate compared to SSD-derived duplicates (Panchy et al. 2016).

Considerable evidence suggests that duplicates originating from WGD and SSD events differ in their evolutionary rate, essentiality, and function (Qiao et al. 2019, Freeling 2009). For example, SSD duplicates are more prone to neo-functionalization whereas WGD duplicates tend to partition ancestral functions (Fares et al. 2013). This difference is apparent in the functional similarity observed among WGD paralog pairs compared to SSD paralogs (Hakes et al. 2007). In addition, models of gene and genome duplication suggest that the rate of gene retention following duplication is notably different for WGD and SSD events, with biases towards specific functional classes (Maere et al. 2005, Freeling 2009). Particularly, genes originating from SSDs often play roles in environmental and defense responses, while genes retained from WGDs are associated with intracellular regulation and include transcription factors, ribosomal proteins, and other core celluar processes. (Hanada et al. 2008, Rizzon et al. 2006, Wang et al. 2012, Qiao et al. 2019, Seoighe and Gehring 2004). However, the underlying mechanisms driving such

functional biases are not well understood. As such, we remain uncertain whether genes retained following the different duplication sources are subjected to different evolutionary constraints, or whether the two classes of genes possess equivalent potential to generate novel function. The gene balance theory offers one plausible explanation for why WGD events predominantly retain transcription factors (Freeling 2009). It has been hypothesized that the more connected the gene product, the more likely the phenotype will change if dosage imbalance occurs. Therefore, the gene balance theory posits that if a tandem duplicate is highly connected, it is more prone to cause dosage imbalance and thus is very likely to be removed. Conversely, after WGD, loss of a highly connected duplicate would be more likely to cause dosage imbalance and deleterious effects and thus is more likely to be retained (Wang et al. 2012, Freeling 2009).

Theoretical and empirical results show that most duplicated genes return to single copies after duplication because a functionally redundant duplicate will accumulate deleterious mutations and evolve into a pseudogene (Wagner 1998, Tautz 1992, Lynch and Conery 2000). Therefore, the earliest stages following duplication is key to understanding the role of gene and genome duplication in evoluton (Innan and Kondrashov 2010). Classic models of duplicate retention include neofunctionalization (Kuzmin et al. 2022, Ohno 1970), subfunctionalization (e.g., the duplication–degeneration–complementation model, DDC) (Van Hoof 2005, Marshall et al. 2013), and escape from adaptive conflict (a special case of subfunctionalization)(Des Marais and Rausher 2008, Kuzmin et al. 2022).

Changes in gene expression can play an important role in the preservation of duplicated genes (Francino 2005, Ganko et al. 2007, Duarte et al. 2006, Doebley and Lukens 1998, Casneuf et al. 2006, Duarte et al. 2006). One observation is that gene expression exhibits greater divergence for pairs originating from SSD as compared to WGD (Casneuf et al. 2006), which may arise from a higher likelihood of altered regulatory features during the process of SSD (e.g., if the duplication event does not faithfully replicate *cis*-regulatory elements) (Rogers et al. 2017). Although seemingly contradictory to such observations, WGD pairs in fact have a higher average number of *cis*-element differences between paralogs than do SSDs (Arsovski et al. 2015). Thus, a question remains as to whether changes in transcription contribute to the probability of gene retention and the subsequent functional biases observed in retained paralog. Another relevant and intriguing question is whether variants affecting gene expression dosage experience the same degree of relaxed selection as do variants that affect protein coding sequences; it is generally accepted that functional redundancy of a protein-coding gene leads to relaxed selection constraints (Ohno 1970, Kondrashov et al. 2002). If this is also the case with transcriptional

expression, it would naturally increase the 'search space' of a new copy, and potentially facilitate gene retention via neofunctionalization.

Here, we study expression variability among gene duplicate pairs to test the hypothesis that biases in the functions of gene copies retained can be explained by changes in regulatory architecture immediately after duplication. Expression variability is defined as the variation of gene expression among genetically identical individuals in a controlled environment (Cortijo et al. 2019, Cortijo and Locke 2020). In multicellular organisms, such as plants, genetically identical or clonal individuals are used to assess variability, also referred to as inter-individual variability (Cortijo et al. 2019) or intra-genotypic variability (Metcalf and Ayroles 2020). In the context of gene duplication, a duplicate may change its genomic context, via transposition or incomplete duplication in which regulatory elements are reshuffled (Arsovski et al. 2015), thereby changing expression variability of the new copy as it becomes fixed through neofunctionalized. Here, we first show that duplicates arising from both SSDs and WGDs tend to exhibit elevated expression variability. Our subsequent analyses point to distinct mechanisms that could explain the increasing expression variability observed for SSDs and WGDs.

Materials and Methods

Normalized expression variability

Gene expression data was obtained from Cortijo et al. 2019. Briefly, RNA-sequencing of 16 individual seedlings of *A. thaliana* (Col-0 accession) were used to calculate by-gene inter-individual expression variability. The square coefficient of variation ($CV^2 = variance/(average^2)$) is first calculated. To account for the coupling between the mean gene expression level and CV^2 and to allow for comparisons among genes, we calculated the normalized expression variability (NEV) as log2(CV^2/trend for the same expression level)(Cortijo et al. 2019).

Calculation of environmental responsiveness from expression atlas

We used a large gene expression atlas (Roberts and Josephs 2023), encompassing a range of experimental conditions and tissues, to characterize the plastic expression responses of individual genes. To calculate environmental responsiveness, we focused on the 'Col-0' ecotype and 'seedling' tissue across environmental treatments (e.g., cold exposure, day length, hormonal treatment)

and developmental stages in the atlas. Responsiveness was quantified as the sum of log2 ratios between expression in a treatment and a baseline condition:

$$\Sigma \log_2 \frac{expr}{expr_{baseline}}$$

Tissue specificity

Tissue specificity of individual genes were obtained from Yang and Gaut 2011, which were calculated from Arabidopsis Development Atlas (ADA), and from the expression atlas (Roberts and Josephs 2023). The expression level of a gene was estimated by the average value of all 79 samples in Roberts and Josephs 2023. The tissue specificity was measured with the index τ (Yanai et al. 2005):

$$\tau = \frac{\Sigma_{j=1}^{n}(1 - \frac{log_2 S_{i,j}}{log_2 S_{i,max}})}{n-1}$$

where n is the number of tissues, $S_{i,j}$ is the expression of gene i in tissue j , and $S_{i,max}$ is the highest expression of gene i across n tissues. The index τ ranges from 0 to 1, with a higher value indicating a higher specificity. If a gene is expressed in only one tissue, τ approaches 1.

Paralog pairs

The designations of genes as duplicates or singletons were acquired from Yang and Gaut 2011 according to the assignments of Blanc et al. 2003. k_s were similarly obtained from Yang and Gaut 2011 and estimated in PAML (Yang et al. 1997), using default parameters. Paralogs and their corresponding duplication modes were obtained from Coate et al. 2020 (Fig. S3), following the methods in Wang et al. 2013.

Results

Following Cortijo et al. 2019, we first obtained normalized expression variability for each gene at each time point during a day. Gene expression data were acquired from 16 seedlings of the Col-0 wild type of *Arabidopsis thaliana*. The goal of normalization is to account for the effect of coupling between mean expression and variation around the mean, with the aim to compare expression variability across genes. Hereafter, we use median normalized expression variability

among all time points to quantify a focal gene's expression variability, termed normalized expression variability (NEV).

Elevated gene expression evolvability for duplicates

We first categorized genes into three groups using their duplication status acquired from Yang and Gaut 2011: singletons (i.e., no known duplicates in the *A. thaliana* genome; N = 1241), duplicates arising from WGD (N = 3175), and duplicates arising from SSD (N = 3366) (Fig. 1A). The distributions of duplicates of either type are significantly shifted towards higher NEV than genes without duplicates ($p = 0.00012$ with SSDs, $p = 0.0078$ with WGDs). We detect no significant difference in NEV between SSDs and WGDs, which suggests a higher rate of changes for SSDs over the timeframe considered because the evolutionary time since duplication is generally longer for WGDs than SSDs in this dataset. We also conducted similar analyses by stratifying the data based on specific features of genes; significant differences between singletons and duplicates were consistently observed, except in groups of genes with low GC content or low intron number (Fig. S1). Further analysis reveals that this elevated expression variation is primarily driven by divergence in NEV between paralogs, rather than uniformly high NEV across all duplicates (Fig. 1B).

Expression variability as an evolvability index for gene expression

We next ask whether expression variability is coupled with plasticity that caused by macro-environmental variation, finding that genes with higher NEV tend to show higher sensitivity to environmental perturbations (Fig. 2). Conversely, genes with lower NEV are more often insensitive to environmental perturbation. While one of the drivers of NEV is micro-environmental variation among individuals, such differences are small, random, and distinguishable from macro-environmental variation that characterizing systematic differences of environmental parameters in experimental settings (Masel and Siegal 2009). This distinction is apparent as indicated in Fig. 2: a higher NEV does not necessarily lead to a higher sensitivity to environmental perturbations in an experiment, while a lower NEV constrain the environmental responsiveness of a gene (inaccessible top left area in the scatter plot of Fig. 2).

Notably, such noise–plasticity coupling has been observed in single cell organisms, and has been interpreted as a constraint on the evolution of gene expression more broadly (Lehner 2010, Singh 2013, Chapal et al. 2019). We also observe a positive correlation between tissue specificity and normalized expression variability in two independent datasets, which were calculated from

Arabidopsis Development Atlas (ADA) (Yang and Gaut 2011), and from the expression atlas (Roberts and Josephs 2023) (Fig. S2).

Prior work on the effects of naturally occurring mutations on gene expression found that genes with greater sensitivity to mutation exhibit higher stochastic noise in gene expression and are also more sensitive to environmental perturbation (Landry et al. 2007). In single-cell organisms, stochastic noise describes the variation within a population of genetically identical cells under a single environment. Together with our study, these earlier findings in yeast mutation accumulation lines imply that expression variability may play a role in the evolution of environmental responses more broadly. We thus hypothesize that elevated expression variability increases the probability of gene retention following duplication, particularly through functional gains of environmental responses.

Constrasting functions among SSD and WGD paralogs despite similar levels of NEV

We conducted GO enrichment analysis on pairs of duplicated genes that show the greatest divergence in NEV between copies (the 5% of genes exhibiting the greatest divergence, comprising 116 SSD pairs and 101 WGD pairs). We specifically tested whether the “high NEV” paralogs from each pair, as a group, exhibit different functional enrichments than the group of “low NEV” paralogs (Fig. 3A, Supplemental tables 1-4). For WGDs, both the low and high NEV groups are enriched with central metabolism and biosythetic processes. However, for SSDs, genes with higher NEV in a SSD paralog pair show novel functions as compared to the lower group. In particular, paralogs in the SSD high group are enriched for functions in biotic and abiotic environmental responses such as responses to cold, zinc ion, fungi, and herbivores. Overall, the number of plasticity-related genes (those with GO terms including "Response to") in enriched cagegories increases from 6.8% to 41.4% for SSDs and from 17.8% to 31.7% for WGDs when we compare low and high NEV classes, respectively. Previous studies found that genes derived from small scale duplications (SSD) are more likely to be involved in response to environmental stimuli (Hanada et al. 2008). Here, we provide evidence that, for SSDs, abiotic or biotic responses are specifically associated with increased expression variability.

We also observed bias in transcription factor retention between two duplication types. We assessed the number of transcription factor (TF) and non-TF paralog pairs in the SSD and WGD pools. (Cases in which a pair of paralogs comprise both a TF and a non-TF gene are rare and thus excluded.) SSDs including proximal, tandem, and transposed duplications have far fewer TF

paralog pairs as compared to genes arising from WGD (Fig. S5, p-value = 5.61e-96, chi-squared test), in line with previous results (Maere et al. 2005). Importantly, for tandem and proximal duplications, TF paralog pairs exhibit lower NEVs as compared to non-TF paralogs, and the distribution of NEV between TF and non-TF pairs is more divergent than other duplication types (Fig. S6). This suggests, for small-scale duplications, lower expression variability may be essential for retentions among transcription factors.

Distinct mechanisms leading to elevated expression variability of WGDs and SSDs

Despite comparable elevation in normalized expression variability (NEV) relative to singletons, WGDs and SSDs display distinct functional enrichment patterns among retained genes. We hypothesize that these differences stem from the timing and rate at which high expression variability is acquired. In principle, the *cis*- and *trans*-regulatory environments are likely conserved immediately after WGD events. Consequently, the increased NEV observed in WGD duplicates, relative to singletons, may result from the gradual accumulation of differences over extended evolutionary timescales. In contrast, the elevated NEV in SSDs may emerge during the duplication process itself. Mechanisms such as translocation via transposable elements or reverse transcription can disrupt a paralog's regulatory machinery, independently of selection or other long-term evolutionary pressures. As a result, the increased NEV observed in SSDs may reflect both the immediate disruptive effects of novel genomic contexts post-duplication and the preferential retention of duplicates with inherently higher NEVs, as those with lower variability may be more prone to early gene loss.

We tested whether SSDs and WGDs exhibit different patterns of NEV divergence between paralogs over time. Our findings support the hypothesis that NEV divergence within paralog pairs follows contrasting evolutionary trajectories for WGDs and SSDs (Fig. 4 and Fig. S4). The number of synonymous substitutions between aligned protein-coding sequences (K_s) serves as a proxy for duplication age (Ganko et al. 2007), though saturation can occur over extended periods. The pattern observed for WGDs provides a null expectation for how NEV differences accumulate over time (Fig. 4, right). We find that NEV divergence increases with higher K_s for WGDs, indicating a greater NEV divergence in older paralog pairs. In contrast, no significant trend between K_s and NEV divergence is detected in SSD paralog pairs (Fig. 4, left).

Next, we further test the age bias for SSDs and WGDs. We examined the differences in gene age and NEV within each paralog pair. We utilized gene age estimates for 17,732 Arabidopsis genes,

inferred based on homologs (a phylostratigraphy approach (Arendsee et al. 2014, Domazet-Lošo et al. 2007)) without paralog information. Many inferred paralogs exhibit differences in their estimated clade of origin (obtained from Arendsee et al. 2014). The positive trend for SSDs in Fig. 5A and B indicates an age bias that younger genes arising from SSD tend to have a higher NEV. By contrast, no such bias is observed for WGDs in both Fig. 5A and B. Notably, we also observe an overall postive trend that younger genes among all singletons and duplicates exhibit higher NEV (Fig. S7).

In sum, our findings support distinct models underlying elevated NEV in SSDs and WGDs. The NEV divergence over time displays distinct patterns between SSDs and WGDs (Fig. 4 and Fig. S4). Furthermore, WGDs show no age bias, whereas SSDs exhibit a bias toward higher NEV for younger genes (Fig. 5), driving the overall positive relationship between gene age and NEV (Fig. S7).

Discussion

Although direction of causality between expression variability and gene retention cannot be determined conclusively, our analyses suggest a possible model to explain the elevated expression variability observed for SSDs and WGDs. Translocation of a paralog to a novel genomic setting is likely to lead to regulatory changes, resulting in expression variability differences between the new paralogs (Fig. 4). SSDs that arise from unequal crossing over during meiosis may not faithfully duplicate the complete *cis*-regulatory sequences, leading to mis-regulation as compared to the ancestral single copy gene (Rogers et al. 2017). SSDs that arise from translocation, e.g. via transposable elements or reverse transcription, may likewise find themselves in novel genomic contexts where local features such as chromatin structure, methylation, or *cis*-elements introduce novel expression phenotypes (Wang et al. 2012). Therefore, the elevated NEV observed in SSDs likely results from immediate regulatory changes post-duplication and a higher retention probability for paralogs with higher NEVs. In line with this model, we observe a bias towards higher NEV for younger genes in SSDs (Fig. 5) and greater NEV divergence for younger SSD paralog pairs (Fig. 4 and Fig. S4), compared to WGDs. Indeed, Qiao et al. 2019 synthesized duplication data across plant lineages and reported that proximal and tandem duplicates exhibited relatively high K_a/K_s ratios but small K_s values, possibly a signature of more rapid functional divergence as compared to other gene classes and supporting the hypothesis that positive selection plays an important role in the early stages following small scale duplications. Future work could use synteny-based methods to determine the parental copy in SSD paralog

pairs, testing whether in general parental copies are less variable for SSDs compared to WGDs. In contrast to SSDs, we observe no significant or, at most, weakly opposite patterns for WGDs in Fig. 4 and Fig. 5. We speculate that the elevated expression variability for WGD could potentially be due to relaxed selection due to extra copies or positive selection (Zhang et al. 2009) over prolonged evolutionary time.

As such, stress-responsive genes are more likely to be retained via SSD than via WGD. Conversely, the retention of transcription factors follow a different pattern, for which dramatic regulatory and expression variability changes are likely deleterious as these can be central and highly 'connected' in gene regulatory networks and therefore subject to selective pressure to minimize transcriptional noise (Puzović et al. 2023). Therefore, transcription factors are less likely to be retained following SSDs than WGDs. Our results propose a plausible explanation for reciprocal retention between tandem duplication and WGDs, potentially complementary to dosage balance theory and other existing models (Freeling 2009, Tasdighian et al. 2017).

The concept of inter-individual variability we study here is distinct from the expression noise in single-cell organisms, which is cell-to-cell heterogeneity in gene expression. The cell-to-cell heterogeneity in gene expression is due to the stochastic nature of the molecular interactions involved in (post)transcription regulation (Elowitz et al. 2002, Raser and O'shea 2005), including variation in the internal states such as the abundance of key cellular components (Pedraza and van Oudenaarden 2005, Das Neves et al. 2010), and external random variation such as micro-environment variation among genetically-identical cells (Battich et al. 2015). Besides these sources of variation, the inter-individual expression variability can arise from micro-environmental variation among individuals (during development) and stochastic variation in cell-type compositions among individuals (bulk RNA-seq experiments measure total gene expression from heterogeneous tissues). Micro-environmental variation, which leads to cell-to-cell heterogeneity and inter-individual variability, is distinct from macro-environmental variation, which regulates stress responses and phenotypic plasticity (Hall et al. 2007, Siegal and Leu 2014, Paaby and Testa 2021, Cortijo et al. 2019). In experimental settings, macro-environmental variation refers to systematic environmental differences deliberately introduced by the experimenter, whereas micro-environmental variation encompasses mild, non-directional random fluctuations (Masel and Siegal 2009).

Low genetic variation is often interpreted as an evolutionary constraint. In many cases, the within-line phenotypic variability and noise is associated with genetic variation within population as well as environmental plasticity - the so-called congruence hypothesis (Wagner et al. 1997,

Gibson and Wagner 2000, Hansen 2006). One explanation for congruence is that different causes of perturbations may be aligned because the responsive dynamics reflect the intrinsic tendencies of developmental systems to produce certain forms rather than others (Alberch 1989, 1982, Nuño de la Rosa and Müller 2024, Landry et al. 2007). Our study demonstrates a case where such variation, apparently arising from stochastic factors, may be used as an index of evolvability (Draghi and Ogbunugafor 2023). Our study emphasizes the role of expression variability in shaping environmental plasticity and expression divergence, pointing to the need to assess phenotypic noise in evolutionary studies (Cai et al. 2024).

References

Alberch, P., 1982. Developmental constraints in evolutionary processes. *in* Evolution and Development: Report of the Dahlem Workshop on Evolution and Development Berlin 1981, May 10–15, Pages 313–332. Springer.

Alberch, P. 1989. The logic of monsters: evidence for internal constraint in development and evolution. Geobios 22:21–57.

Arendsee, Z. W., Li, L., and Wurtele, E. S. 2014. Coming of age: orphan genes in plants. Trends in plant science 19:698–708.

Arsovski, A. A., Pradinuk, J., Guo, X. Q., Wang, S., and Adams, K. L. 2015. Evolution of cis-regulatory elements and regulatory networks in duplicated genes of arabidopsis. Plant physiology 169:2982–2991.

Battich, N., Stoeger, T., and Pelkmans, L. 2015. Control of transcript variability in single mammalian cells. Cell 163:1596–1610.

Blanc, G., Hokamp, K., and Wolfe, K. H. 2003. A recent polyploidy superimposed on older large-scale duplications in the arabidopsis genome. Genome research 13:137–144.

Cai, H., Melo, D., and Des Marais, D. 2024. Disentangling variational bias: the roles of development, mutation and selection. Trends in Genetics .

Casneuf, T., De Bodt, S., Raes, J., Maere, S., and Van de Peer, Y. 2006. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant arabidopsis thaliana. Genome biology 7:1–11.

Chapal, M., Mintzer, S., Brodsky, S., Carmi, M., and Barkai, N. 2019. Resolving noise–control conflict by gene duplication. PLoS biology 17:e3000289.

Coate, J. E., Farmer, A. D., Schiefelbein, J. W., and Doyle, J. J. 2020. Expression partitioning of duplicate genes at single cell resolution in arabidopsis roots. Frontiers in Genetics 11:596150.

Cortijo, S., Aydin, Z., Ahnert, S., and Locke, J. C. 2019. Widespread inter-individual gene expression variability in arabidopsis thaliana. Molecular systems biology 15:e8591.

Cortijo, S. and Locke, J. C. 2020. Does gene expression noise play a functional role in plants? Trends in plant science 25:1041–1051.

Das Neves, R. P., Jones, N. S., Andreu, L., Gupta, R., Enver, T., and Iborra, F. J. 2010. Connecting variability in global transcription rate to mitochondrial variability. PLoS biology 8:e1000560.

Des Marais, D. L. and Rausher, M. D. 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. Nature 454:762–765.

Doebley, J. and Lukens, L. 1998. Transcriptional regulators and the evolution of plant form. The Plant Cell 10:1075–1082.

Domazet-Lošo, T., Brajković, J., and Tautz, D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends in Genetics 23:533–539.

Draghi, J. A. and Ogbunugafor, C. B. 2023. Exploring the expanse between theoretical questions and experimental approaches in the modern study of evolvability. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution 340:8–17.

Duarte, J. M., Cui, L., Wall, P. K., Zhang, Q., Zhang, X., Leebens-Mack, J., Ma, H., Altman, N., and DePamphilis, C. W. 2006. Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of arabidopsis. Molecular biology and evolution 23:469–478.

Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. 2002. Stochastic gene expression in a single cell. Science 297:1183–1186.

Fares, M. A., Keane, O. M., Toft, C., Carretero-Paulet, L., and Jones, G. W. 2013. The roles of whole-genome and small-scale duplications in the functional specialization of saccharomyces cerevisiae genes. PLoS genetics 9:e1003176.

Francino, M. P. 2005. An adaptive radiation model for the origin of new gene functions. Nature genetics 37:573–578.

Freeling, M. 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annual review of plant biology 60:433–453.

Ganko, E. W., Meyers, B. C., and Vision, T. J. 2007. Divergence in expression between duplicated genes in arabidopsis. Molecular biology and evolution 24:2298–2309.

Ge, S. X., Jung, D., and Yao, R. 2020. Shinygo: a graphical gene-set enrichment tool for animals and plants. Bioinformatics 36:2628–2629.

Gibson, G. and Wagner, G. 2000. Canalization in evolutionary genetics: a stabilizing theory? BioEssays 22:372–380.

Hakes, L., Pinney, J. W., Lovell, S. C., Oliver, S. G., and Robertson, D. L. 2007. All duplicates are not equal: the difference between small-scale and genome duplication. Genome biology 8:1–13.

Hall, M. C., Dworkin, I., Ungerer, M. C., and Purugganan, M. 2007. Genetics of microenvironmental canalization in arabidopsis thaliana. Proceedings of the National Academy of Sciences 104:13717–13722.

Hanada, K., Zou, C., Lehti-Shiu, M. D., Shinozaki, K., and Shiu, S.-H. 2008. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant physiology 148:993–1003.

Hansen, T. F. 2006. The evolution of genetic architecture. Annu. Rev. Ecol. Evol. Syst. 37:123–157.

Innan, H. and Kondrashov, F. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nature Reviews Genetics 11:97–108.

Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V. 2002. Selection in the evolution of gene duplications. Genome biology 3:1–9.

Kuzmin, E., Taylor, J. S., and Boone, C. 2022. Retention of duplicated genes in evolution. Trends in Genetics 38:59–72.

Landry, C. R., Lemos, B., Rifkin, S. A., Dickinson, W., and Hartl, D. L. 2007. Genetic properties influencing the evolvability of gene expression. Science 317:118–121.

Lehner, B. 2010. Conflict between noise and plasticity in yeast. PLoS genetics 6:e1001185.

Lynch, M. and Conery, J. S. 2000. The evolutionary fate and consequences of duplicate genes. science 290:1151–1155.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van de Peer, Y. 2005. Modeling gene and genome duplications in eukaryotes. Proceedings of the National Academy of Sciences 102:5454–5459.

Marshall, A. N., Montealegre, M. C., Jiménez-López, C., Lorenz, M. C., and van Hoof, A. 2013. Alternative splicing and subfunctionalization generates functional diversity in fungal proteomes. PLoS genetics 9:e1003376.

Masel, J. and Siegal, M. L. 2009. Robustness: mechanisms and consequences. Trends in genetics 25:395–403.

Metcalf, C. J. E. and Ayroles, J. F. 2020. Why does intragenotypic variance persist? Unsolved Problems in Ecology page 43.

Nuño de la Rosa, L. and Müller, G. B. 2024. The legacy and evolvability of pere alberch's ideas. Interface Focus 14:20240011.

Ohno, S., 1970. Evolution by gene duplication. Berlin-Heidelberg-New York: Springer-Verlag.

Paaby, A. B. and Testa, N. D. 2021. Developmental plasticity and evolution. Evolutionary Developmental Biology: A Reference Guide Pages 1073–1086.

Panchy, N., Lehti-Shiu, M., and Shiu, S.-H. 2016. Evolution of gene duplication in plants. Plant physiology 171:2294–2316.

Pedraza, J. M. and van Oudenaarden, A. 2005. Noise propagation in gene networks. Science 307:1965–1969.

Puzović, N., Madaan, T., and Dutheil, J. Y. 2023. Being noisy in a crowd: Differential selective pressure on gene expression noise in model gene regulatory networks. PLoS Computational Biology 19:e1010982.

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., Zhang, S., and Paterson, A. H. 2019. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. Genome biology 20:1–23.

Raser, J. M. and O'shea, E. K. 2005. Noise in gene expression: origins, consequences, and control. Science 309:2010–2013.

Rizzon, C., Ponger, L., and Gaut, B. S. 2006. Striking similarities in the genomic distribution of tandemly arrayed genes in arabidopsis and rice. PLoS computational biology 2:e115.

Roberts, M. and Josephs, E. B. 2023. Weaker selection on genes with treatment-specific expression consistent with a limit on plasticity evolution in arabidopsis thaliana. Genetics 224:iyad074.

Rogers, R. L., Shao, L., and Thornton, K. R. 2017. Tandem duplications lead to novel expression patterns through exon shuffling in drosophila yakuba. PLoS genetics 13:e1006795.

Seoighe, C. and Gehring, C. 2004. Genome duplication led to highly selective expansion of the arabidopsis thaliana proteome. Trends in Genetics 20:461–464.

Siegal, M. L. and Leu, J.-Y. 2014. On the nature and evolutionary impact of phenotypic robustness mechanisms. Annual review of ecology, evolution, and systematics 45:495–517.

Singh, G. P. 2013. Coupling between noise and plasticity in e. coli. G3: Genes, Genomes, Genetics 3:2115–2120.

Tasdighian, S., Van Bel, M., Li, Z., Van de Peer, Y., Carretero-Paulet, L., and Maere, S. 2017. Reciprocally retained genes in the angiosperm lineage show the hallmarks of dosage balance sensitivity. The Plant Cell 29:2766–2785.

Tautz, D. 1992. Problems and paradigms: Redundancies, development and the flow of information. BioEssays 14:263–266.

Tiley, G. P., Ane, C., and Burleigh, J. G. 2016. Evaluating and characterizing ancient whole-genome duplications in plants with gene count data. Genome biology and evolution 8:1023–1037.

Van Hoof, A. 2005. Conserved functions of yeast genes support the duplication, degeneration and complementation model for gene duplication. Genetics 171:1455–1461.

Vision, T. J., Brown, D. G., and Tanksley, S. D. 2000. The origins of genomic duplications in arabidopsis. Science 290:2114–2117.

Wagner, A. 1998. The fate of duplicated genes: loss or new function? BioEssays 20:785–788.

Wagner, G. P., Booth, G., and Bagheri-Chaichian, H. 1997. A population genetic theory of canalization. Evolution 51:329–347.

Wang, Y., Tan, X., and Paterson, A. H. 2013. Different patterns of gene structure divergence following gene duplication in arabidopsis. BMC genomics 14:1–9.

Wang, Y., Wang, X., and Paterson, A. H. 2012. Genome and gene duplications and gene expression divergence: a view from plants. Annals of the New York Academy of Sciences 1256:1–14.

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21:650–659.

Yang, L. and Gaut, B. S. 2011. Factors that contribute to variation in evolutionary rate among arabidopsis genes. Molecular Biology and Evolution 28:2359–2369.

Yang, Z. et al. 1997. Paml: a program package for phylogenetic analysis by maximum likelihood. Computer applications in the biosciences 13:555–556.

Zhang, J. 2003. Evolution by gene duplication: an update. Trends in ecology & evolution 18:292–298.

Zhang, Z., Qian, W., and Zhang, J. 2009. Positive selection for elevated gene expression noise in yeast. Molecular systems biology 5:299.

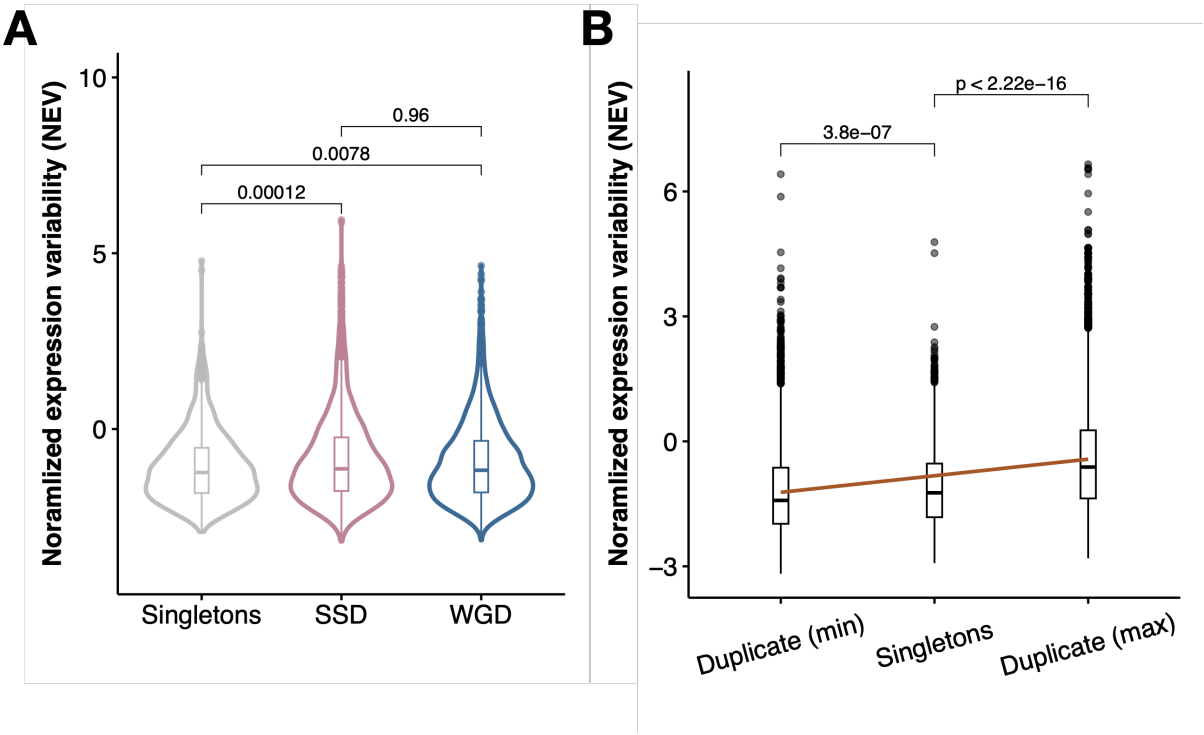

Figure 1: **Genes born from whole genome duplications (WGDs) and small-scale duplications (SSDs) exhibit higher NEV than singletons.** A). Distribution of NEV grouped by singletons, genes arising from SSD, and from WGD. B). Distributions of NEV grouped by genes with lower NEVs within a paralog pair, singletons, and genes with higher NEVs within a paralog pair. Numbers indicate single-sided P-values from Wilcox test.

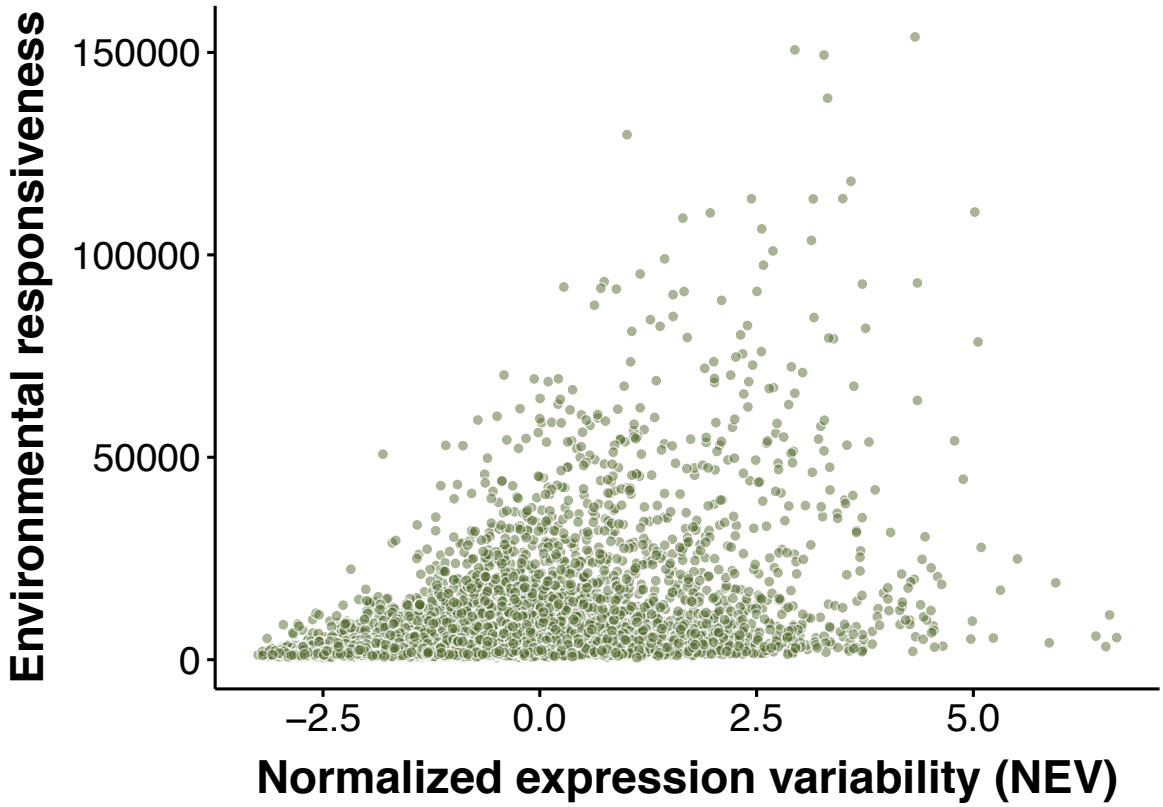

Figure 2: **Normalized expression variability (NEV) correlates with plasticity of transcriptional responses** The plasticity of transcriptional responses (responsiveness) summarizes the ability of a gene's transcript to vary in response to wide range of environmental perturbations. In our case, we obtained an expression atlas and calculated responsiveness using expression data in seedlings and col-1 ecotype of *Arabidopsis* (Roberts and Josephs 2023) .

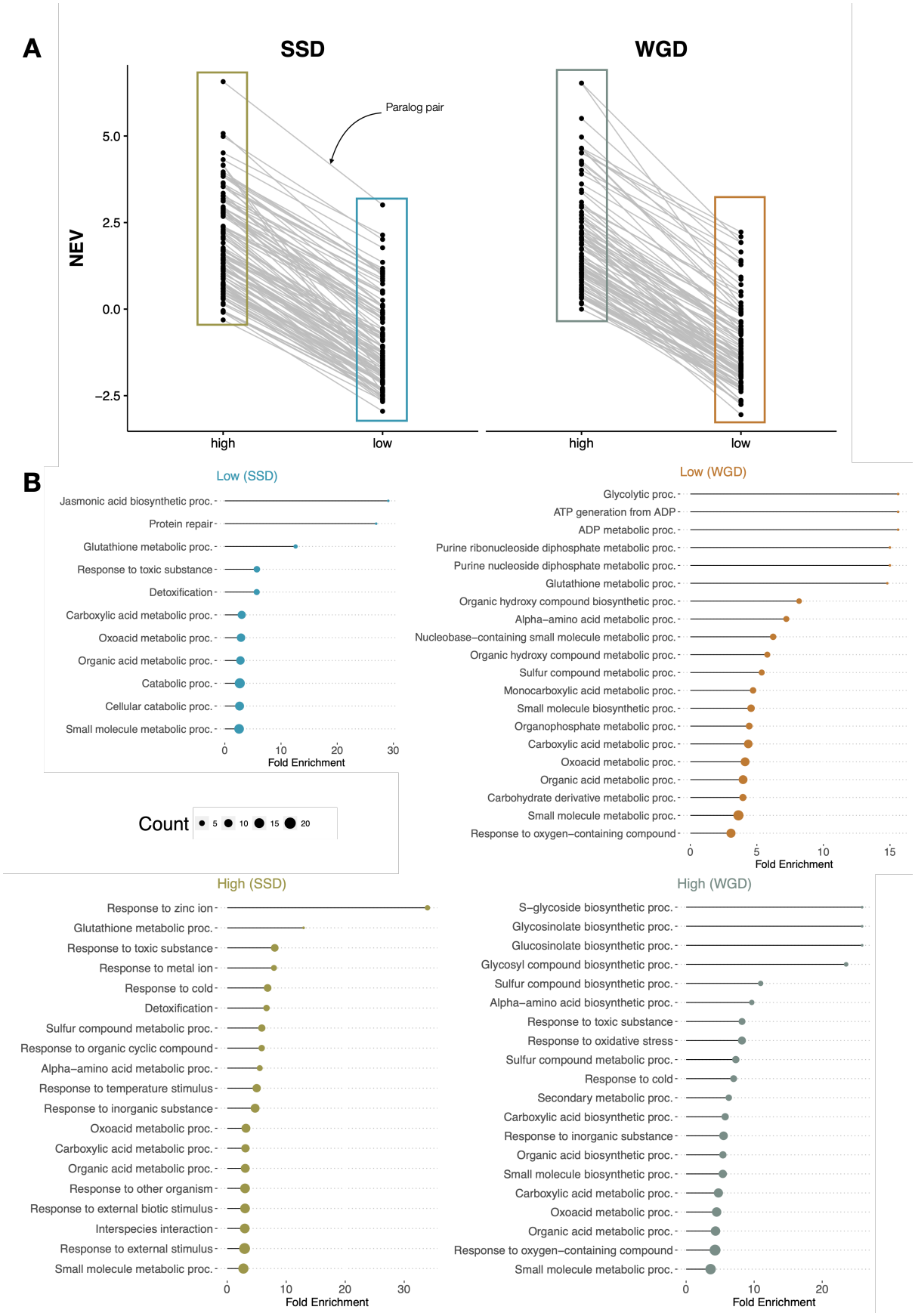
A
SSD
WGD
Paralog pair
NEV
5.0
2.5
0.0
−2.5
high
low
high
low
B
Low (SSD)
Jasmonic acid biosynthetic proc.
Protein repair
Glutathione metabolic proc.
Response to toxic substance
Detoxification
Carboxylic acid metabolic proc.
Oxoacid metabolic proc.
Organic acid metabolic proc.
Catabolic proc.
Cellular catabolic proc.
Small molecule metabolic proc.
0
10
20
30
Fold Enrichment
Count
5
10
15
20
Low (WGD)
Glycolytic proc.
ATP generation from ADP
ADP metabolic proc.
Purine ribonucleoside diphosphate metabolic proc.
Purine nucleoside diphosphate metabolic proc.
Glutathione metabolic proc.
Organic hydroxy compound biosynthetic proc.
Alpha−amino acid metabolic proc.
Nucleobase−containing small molecule metabolic proc.
Organic hydroxy compound metabolic proc.
Sulfur compound metabolic proc.
Monocarboxylic acid metabolic proc.
Small molecule biosynthetic proc.
Organophosphate metabolic proc.
Carboxylic acid metabolic proc.
Oxoacid metabolic proc.
Organic acid metabolic proc.
Carbohydrate derivative metabolic proc.
Small molecule metabolic proc.
Response to oxygen−containing compound
0
5
10
15
Fold Enrichment
High (SSD)
Response to zinc ion
Glutathione metabolic proc.
Response to toxic substance
Response to metal ion
Response to cold
Detoxification
Sulfur compound metabolic proc.
Response to organic cyclic compound
Alpha−amino acid metabolic proc.
Response to temperature stimulus
Response to inorganic substance
Oxoacid metabolic proc.
Carboxylic acid metabolic proc.
Organic acid metabolic proc.
Response to other organism
Response to external biotic stimulus
Interspecies interaction
Response to external stimulus
Small molecule metabolic proc.
0
10
20
30
Fold Enrichment
High (WGD)
S−glycoside biosynthetic proc.
Glycosinolate biosynthetic proc.
Glucosinolate biosynthetic proc.
Glycosyl compound biosynthetic proc.
Sulfur compound biosynthetic proc.
Alpha−amino acid biosynthetic proc.
Response to toxic substance
Response to oxidative stress
Sulfur compound metabolic proc.
Response to cold
Secondary metabolic proc.
Carboxylic acid biosynthetic proc.
Response to inorganic substance
Organic acid biosynthetic proc.
Small molecule biosynthetic proc.
Carboxylic acid metabolic proc.
Oxoacid metabolic proc.
Organic acid metabolic proc.
Response to oxygen−containing compound
Small molecule metabolic proc.
0
10
20
Fold Enrichment

Figure 3

Figure 3: **Contrasting GO enrichment results for SSDs and WGDs with NEV diverged gene pairs.** Higher NEV genes within SSD paralog pairs are enriched with functions environmental responses including both biotic and abiotic factors. A). Top 5% paralog pairs with strongest NEV divergence between two paralogs are shown as single lines for two duplication types. See Fig. S3 for number of genes in each sub-category of duplication. Repeated genes in each group are removed. B). GO enrichedment analyses are conducted using ShinyGo `http://bioinformatics.sdstate.edu/go/` (Ge et al. 2020) , for each group among the four in A separately (with corresponding colors). Lists of GO biological functions with FDR of 0.05 are shown. GO terms with less than 10 genes are excluded. Fold enrichment and gene count are shown with lollipop plot with the dot.

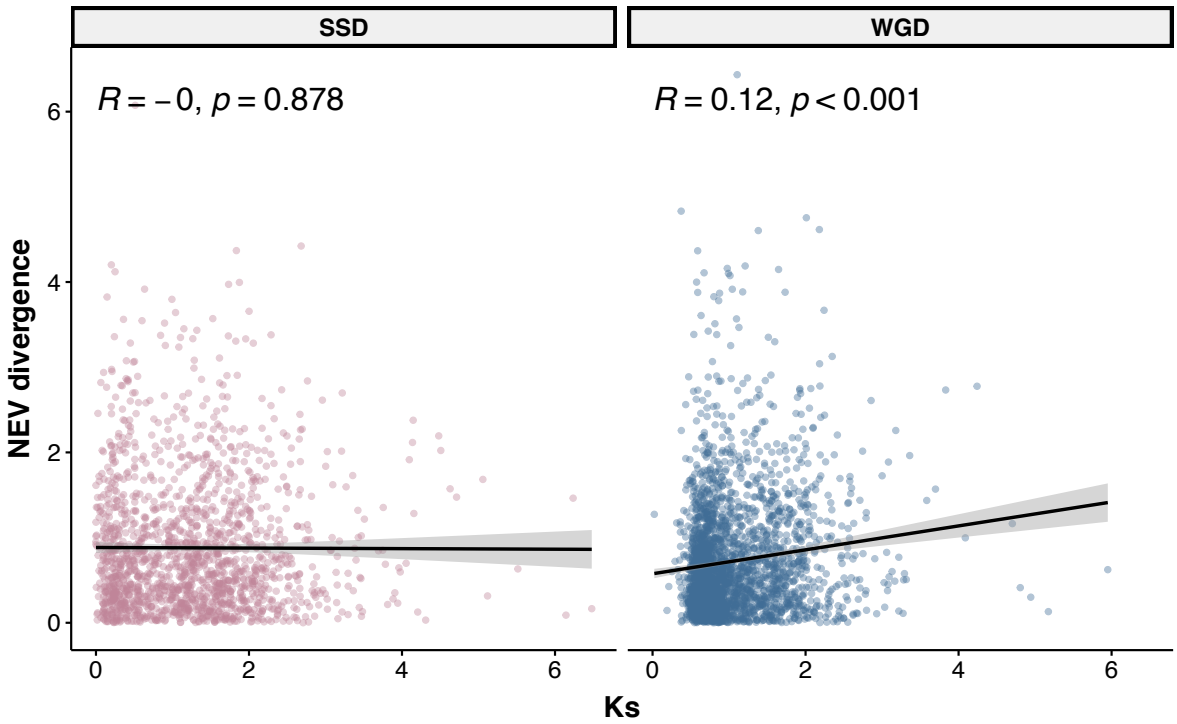

Figure 4: **Paralog pairs arising from SSDs and WGDs exhibit different patterns in the NEV divergence over time.** Scatter plot between NEV divergence and K_s for paralog pairs arising due to SSD and WGD, respectively. Each dot represents a paralog pair.

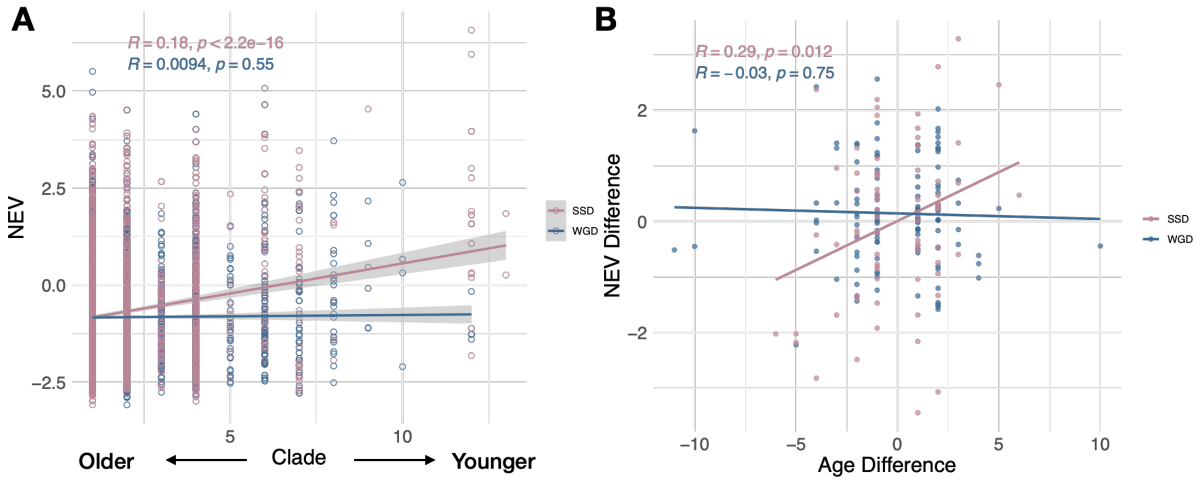

Figure 5: **Younger gene exhibiting higher NEV for SSDs.** A). For duplicates arising from SSD, younger genes tend to exhibit higher NEV while no significant patterns observed for WGD duplicates. B). Constrasting patterns between SSD and WGD for gene age and NEV within a paralog pair. In a SSD paralog pair, the younger gene tends to exhibit higher NEV while no significant patterns observed for WGD paralog pairs. Values in x-axis are the difference of gene clades between paralogs in each pair and the y-axis is the difference of normalized expression variability (NEV) between paralogs in each pair. Paralog pairs with the same inferred clade are excluded (resulting in y = 0). Gene age data were obtained from published data sets (Arendsee et al. 2014), comprising 17,732 Arabidopsis genes stratified into 15 clades, numbered from the oldest (stratum 1) to the most recent (stratum 15) (Arendsee et al. 2014).