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: RESEARCH ARTICLE

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

ONLINE ENHANCEMENTS: Online Supplementary Material

KEYWORDS: Gene duplication; Noise-plasticity coupling; Inter-individual variability; Expression noise; Evolvability; Polyploidy

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, remains 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 bias remain elusive. Here, to better understand the mechanisms and factors promoting gene retention of certain functional categories, we study a frequently overlooked aspect —expression variability—as measured by within-line expression variation. We find that, on average, genes with duplicates exhibit higher expression variability than singletons. We further find that patterns of duplicate retention are likely driven by the immediate increase in expression variability following small-scale duplications (SSDs) and prolonged evolutionary processes after whole-genome duplications (WGDs). These findings suggest a potential overarching mechanism that facilitates gene expression divergence and, consequently, promotes gene retention.

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 main 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 including 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

bias are not 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.

As posited by population genetic theory, 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, a key aspect of understanding the role of gene and genome duplication in evoluton is the earlies stages following duplication (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), escape from adaptive conflict (a special case of subfunctionalization)(Des Marais and Rausher 2008, Kuzmin et al. 2022), and retention to ensure genetic robustness (Gu et al. 2003).

Changes in gene expression are believed to 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 features) (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 subsequent functional biases in retained genes. 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. Gene expression is an inherently noisy process (Elowitz et al. 2002, Raser and O'shea 2005). Expression variability, an analogous concept of expression noise in single-cell organisms, is here 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). Such expression variability may be determined by the structure of transcriptional regulatory networks (Puzović et al. 2023), by *cis*- regulatory elements (Landry et al. 2007), or by differences in selection pressure (Puzović et al. 2023). Critically, phenotypic variability may be advantagous when an unpredicatable environment selects for a bet-hedging strategy (Simons 2011, Venable 2007, Abley et al. 2021). Furthermore, such variability may be indicative of the inherent ability of a system to vary (Cai et al. 2024). 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. Intuitively, if the transcriptional abundance of duplicates exhibit is more variable, the duplicate may have a higher chance of regulatory neofunctionalization. Here, we first show that duplicates arising from both SSDs and WGDs tend to have elevated expression variability. However, unlike WGDs, elevated expression variability following duplication may promote regulatory neofunctionalization for SSDs, thus increasing the chances of gene retention.

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 responses in expression 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. Responsiveness was quantified as the sum of log2 ratios between expression in a treatment and 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. 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 j tissue, 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.

Gene age and paralog pairs

Designations of genes as duplicates or singletons were acquired from Yang and Gaut 2011. Duplicated genes are classified as to whether they were derived from WGD events 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. Paralog pairs and their corresponding duplication modes were obtained from Coate et al. 2020, following the methods in Wang et al. 2013. In total, there are 11,630 pairs of paralogs. 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). For our analyses, we first used the maximum clade number within each paralog pair as a proxy for duplication time and then binned the paralog pair into three groups based on this number: early (strata 1 - 2), intermediate (strata 3 - 4), recent (strata 5 - 15). Paralog pairs are not divided with equal intervals in the mean clade number mainly due to unbalanced sample sizes in each clade.

Results and Discussion

Following Cortijo et al. 2019, we first obtained normalized expression variability for each gene at each time point during a day. Gene expression 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. 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 currently in the *A. thaliana* genome; N = 1241), duplicates arising from WGD (N = 3175), and duplicates arising from SSD (N = 3366) (Fig. 1A). The distribution of duplicates 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. 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, 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. 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. S1).

Prior work on the effects of naturally occurring mutations on gene expression found that genes with greater sensitivity to mutations exhibit higher stochastic noise in gene expression and

are also more sensitive to environmental perturbation (Landry et al. 2007). In a 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 speculate that elevated expression variability may increase the probability of gene retention following duplication, particularly through functional gains of stress and environmental responses.

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

We next test the hypothesis that an increase of gene expression variability helps gene copies to evolve environmental responses. We conducted GO enrichment analysis on pairs of duplicated genes that show the greatest divergence 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” paralog from each pair exhibit, as a group, 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 response to cold, zinc ion, fungal, and herbivore. 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 SSD are more likely to be involved in response to environmental stimuli (Hanada et al. 2008). Here, we provide evidence that in the case of SSDs, increasing expression variability may contribute to the regulatory neofunctionalization involved in abiotic or biotic responses.

In line with the above contrasting results between WGDs and SSDs, Casneuf et al. found that duplicates created from WGDs are often expressed in a larger number of tissues as compared to duplicates generated by SSDs (Casneuf et al. 2006). Moreover, functional redundancy — in which both duplicates are expressed in the same tissue — is common for WGDs. Conversely, Casneuf et al. observed that the tissue-specific expression of gene pairs resulting from SSDs has

diverged asymmetrically, where one gene is expressed in a high number of tissues, as opposed to its duplicate that is expressed in a small number of tissues.

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. List of TFs are downloaded from PlantTFDB (Jin et al. 2016). (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 WGDs (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 crucial for retentions among transcription factors.

Distinct mechanisms for elevated expression variability between WGDs and SSDs

Why do WGDs and SSDs exhibit contrasting functional enrichments of retained genes despite showing similarly elevated variability as compared to singletons? We hypothesize that this can be attributed to the timing and rate of gaining high expression variability. In principle, both the *cis*- and the *trans*- environments immediately upon duplication by WGD events should be conserved. Therefore, the observed increase in NEV for WGD duplicates — as compared to singletons — may be the result of accumuated differences over longer evolutionary timescales. By contrast, the observed increase in NEV for SSDs may arise from changes in the regulatory machinery during the process of duplication itself. 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 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, to survive the early fate of gene loss, an increase in gene expression variability could facilitate the gain of new functions in SSDs. Accordingly, elevated expression variability in the case of SSDs is more likely caused by biased gene retention: paralogs with higher expression variability tend to be preserved. Our results support the idea that elevated expression variability may potentiate the

retention of new gene because we observe contrasting behaviors of regulatory neofunctionalization (in environmental responses) for SSDs even after controlling for expression variability level as compared to WGDs.

Based on our observation of a link between the mode of duplication and post-duplication regulatory evolution, we predict that genes with higher NEVs are preferentially retained after SSD events. We predict that the retention probability for paralogs arising from WGDs is dictated by mechanisms such as reduced noise in expression (Rodrigo and Fares 2018), selection for genetic robustness (Gu et al. 2003), or dosage compensation (Wang et al. 2012, Freeling 2009). In other words, the elevated NEV we observed above in WGD genes may be due to neutral or adaptive processes over longer evolutionary time scale. Accordingly, we next test two specific hypotheses regarding the mechanisms of paralog retention. First, we hypothesize that younger SSD paralog pairs exhibit more divergence in NEV as compared to older SSD paralog pairs. Second, we hypothesize that, within a given pair of retained SSD paralogs, the younger paralog will exhibit higher NEV than the older paralog, while WGD paralogs will not show this age bias in NEV as two copies are symmetric following duplications.

We do, indeed, find support for the first hypothesis that the divergence of NEV within a paralog pair over time exhibits constrasting patterns between WGDs and SSDs (Fig. 4 and Fig. S3). Here, we use existing estimates of the timing of genome duplication to group genes arising from three major events of WGDs. To estimate the age of SSD events, we use estimates of gene clade age (Moutinho et al. 2022), categorizing paralog pairs into three groups, from the most recent to the earliest (see Materials and Methods). We find a trend of greater NEV divergence for younger SSD paralog pairs (more recent duplications), while NEV divergence exhibit no significant trend for WGD pairs (Fig. 4). Further, the number of synonymous substitutions between the two aligned protein coding sequences (Ks) can be viewed as a proxy of of duplication age (Ganko et al. 2007). In Fig. S4, we show a weak positive correlation between Ks and NEV divergence for WGDs, that is, we observe greater NEVs for older paralog pairs. On the other hand, no significant trend is observed in SSD paralog pairs.

For the second hypothesis, to test the age bias, we calculate the gene age difference and the NEV difference for each paralog pair. The positive correlation for SSDs indicates that the younger gene within a pair of paralogs tends to exhibit a higher NEV (Fig. 5). As expected, in contrast, no correlation is observed for WGDs (Fig. 5). Taken together, these suggest that unlike WGDs, new genes arising from SSDs with higher NEVs are more likely to be retained.

A mechanism for biased gene retention in SSDs and WGDs

Collectively, our analyses suggest two distinct models to explain the elevated expression variability observed for SSDs and WGDs. For SSDs, a translocation of a paralog to a novel genomic setting is likely to lead to regulatory changes, resulting in expression variability differences. Abrupt increases in expression variability potentially promote retention for genes, during a positive selection, especially for stress response genes. In line with this model, we observe a bias towards higher expression variability for younger duplicates in SSDs (Fig. 5). Indeed, Qiao et al. 2019 synthesized duplication data across plant lineages and reported that proximal and tandem duplicates exhibited relatively high Ka/Ks ratios but small Ks values, apparently 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 stage for SSDs. In contrast to SSDs, our data suggest that regulatory and expression divergence are less likely to contribute to the gene retention processes for WGDs. Thus, 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 be due to neutral drift or positive selection (Zhang et al. 2009) over prolonged evolutionary time. As such, stress-responses genes are more likely to be retained via SSD mode 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 may subjected 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 new model to reciprocal retention between tandem duplication and WGDs, alternative to dosage balance theory and other existing models (Freeling 2009, Tasdighian et al. 2017).

Low genetic variation is often interpreted as an evolutionary constraint. However, it is challenging to trace back and infer the genetic variation in an ancestral population. One approach is to use phenotypic variability. The variability in gene expression is an outcome of stochastic molecular fluctuations during transcription and the tendency of regulatory networks responding to micro-environmental perturbations. In many cases, such within-line phenotypic variabiliity 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 the congruence is that different perturbations may aligned because the particular responsive dynamics upon perturbation are more likely to occur as it reflects the intrinsic tendencies of developmental systems to produce certain forms rather than others

(Alberch 1989, 1982, Nuño de la Rosa and Müller 2024). 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

Abley, K., Formosa-Jordan, P., Tavares, H., Chan, E. Y., Afsharinafar, M., Leyser, O., and Locke, J. C. 2021. An aba-ga bistable switch can account for natural variation in the variability of arabidopsis seed germination time. Elife 10:e59485.

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.

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.

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.

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.

Gu, Z., Steinmetz, L. M., Gu, X., Scharfe, C., Davis, R. W., and Li, W.-H. 2003. Role of duplicate genes in genetic robustness against null mutations. Nature 421:63–66.

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.

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.

Jin, J., Tian, F., Yang, D.-C., Meng, Y.-Q., Kong, L., Luo, J., and Gao, G. 2016. Planttfdb 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic acids research page Gkw982.

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.

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

Moutinho, A. F., Eyre-Walker, A., and Dutheil, J. Y. 2022. Strong evidence for the adaptive walk model of gene evolution in drosophila and arabidopsis. PLoS biology 20:e3001775.

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.

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

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.

Rodrigo, G. and Fares, M. A. 2018. Intrinsic adaptive value and early fate of gene duplication revealed by a bottom-up approach. Elife 7:e29739.

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.

Simons, A. M. 2011. Modes of response to environmental change and the elusive empirical evidence for bet hedging. Proceedings of the Royal Society B: Biological Sciences 278:1601–1609.

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.

Venable, D., 2007. Bet hedging in a guild of desert annuals. ecology88: 1086-1090.

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.

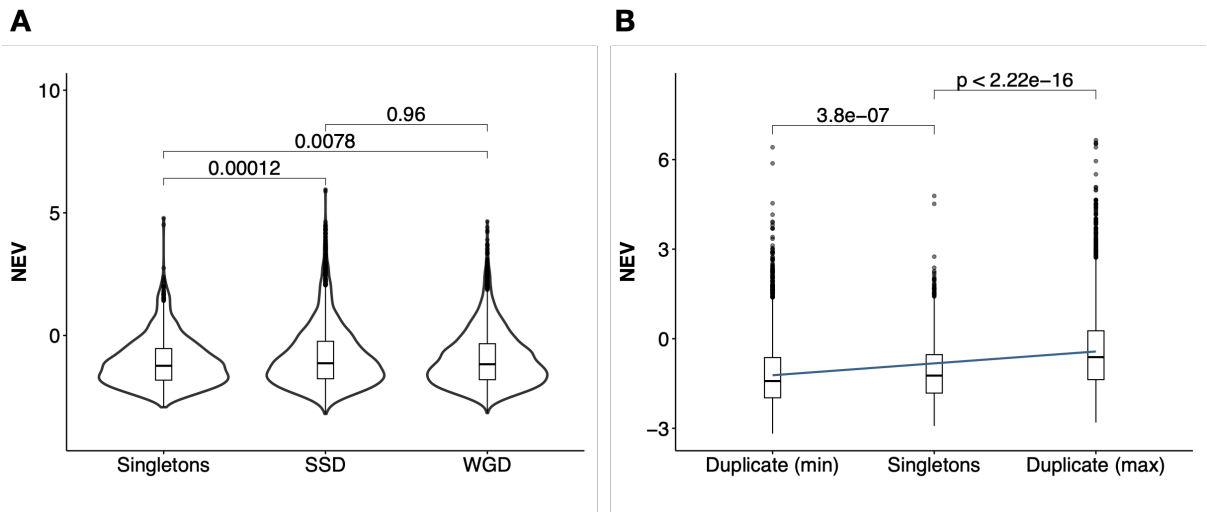

Figure 1: **Genes born from whole genome duplications (WGDs) and small-scale duplications (SSDs) exhibit higher NEV than singletons.** A). Distribution of NEV grouped by singletons, genes arising from SSD, and from WGD. B). Distributions of NEV grouped by genes with lower NEVs within a paralog pair,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

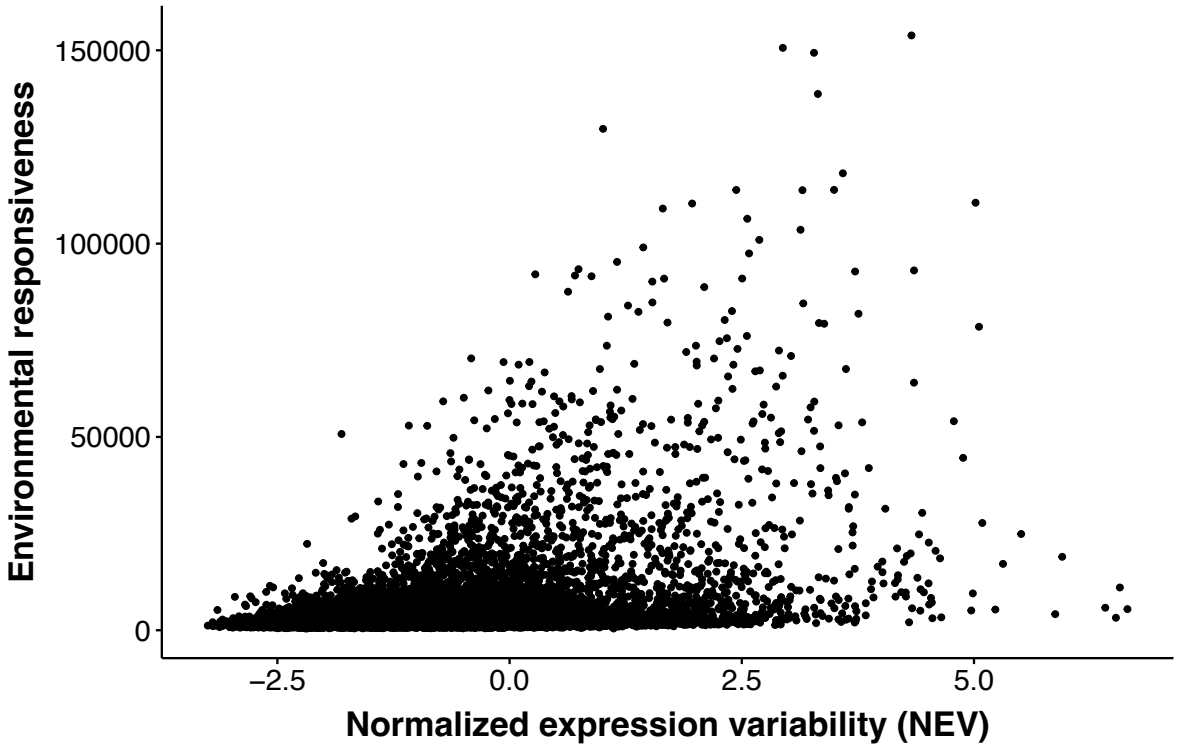

Figure 2: **Normalized expression variability (NEV) correlates with plasticity of transcriptional responses** The plasticity of transcriptional responses (responsiveness) summarizes the ability of a gene's transcript to vary in response to wide range of environmental perturbations. In our case, we obtained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

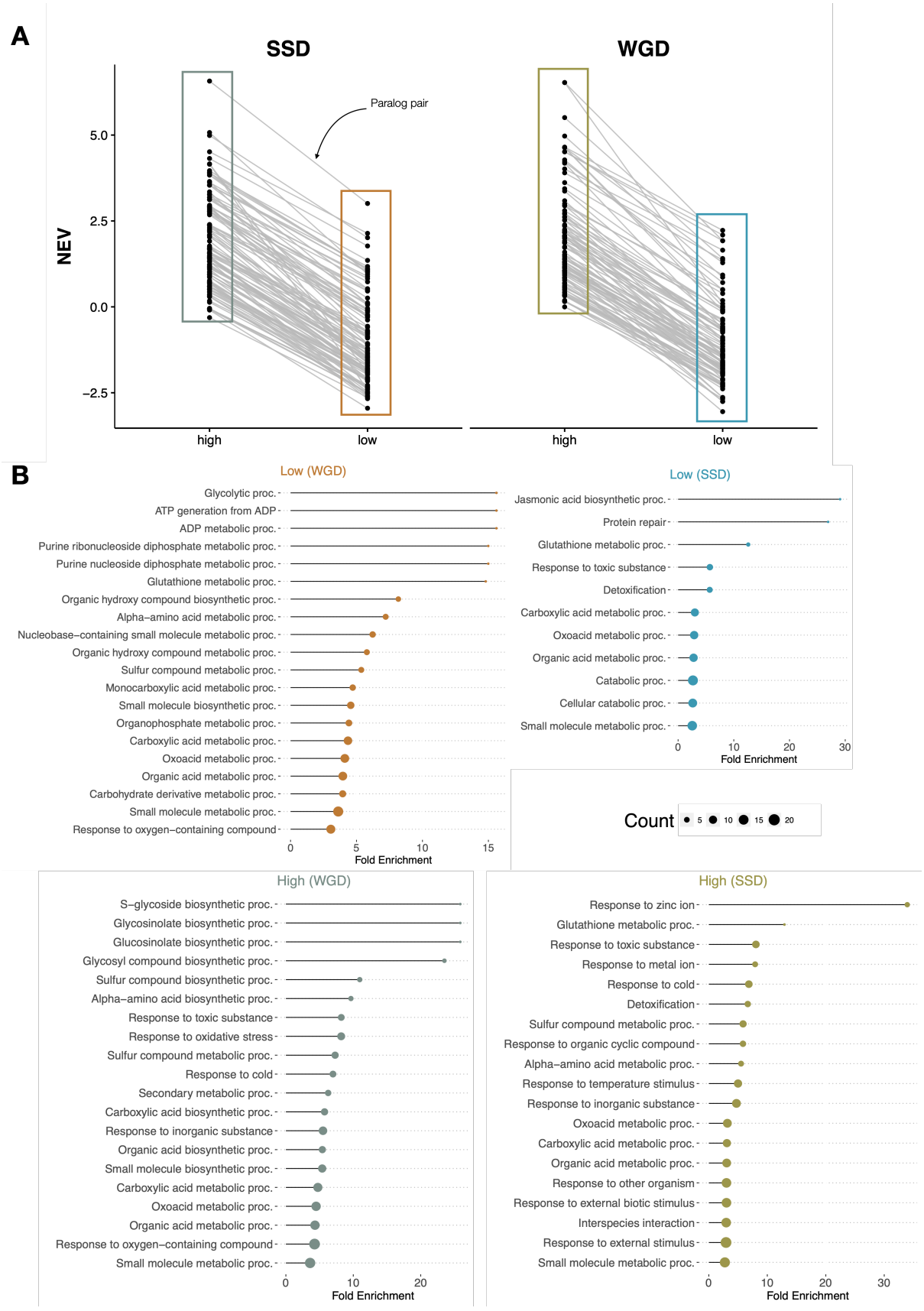
A
SSD
WGD
Paralog pair
NEV
5.0
2.5
0.0
−2.5
high
low
high
low
B
Low (WGD)
Glycolytic proc.
ATP generation from ADP
ADP metabolic proc.
Purine ribonucleoside diphosphate metabolic proc.
Purine nucleoside diphosphate metabolic proc.
Glutathione metabolic proc.
Organic hydroxy compound biosynthetic proc.
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
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
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
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

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. S2 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.

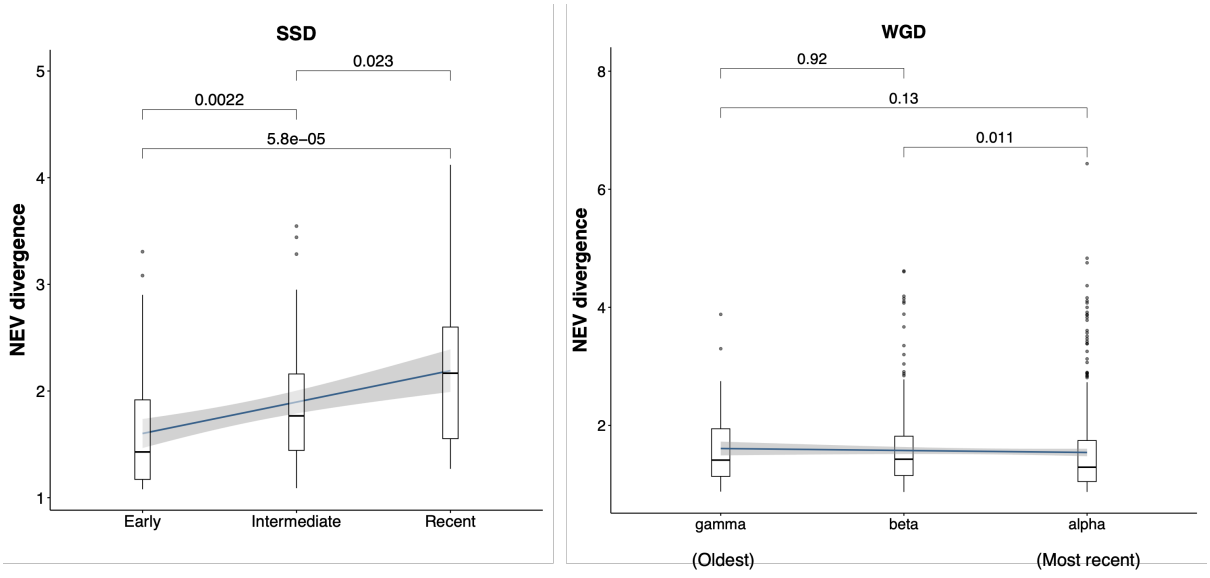

Figure 4: **Paralog pairs arising from SSDs and WGDs exhibit contrasting patterns in the NEV divergence over time.** A). Distribution of NEV divergence grouped by the time of duplications. For SSDs, the time is approximated by the averaged clade between two paralogs (See Materials and Methods); For WGD, we use the inferred age of WGD events from oldest (gamma) to the most recent (alpha). Highest 30% among all paralog pairs are used.

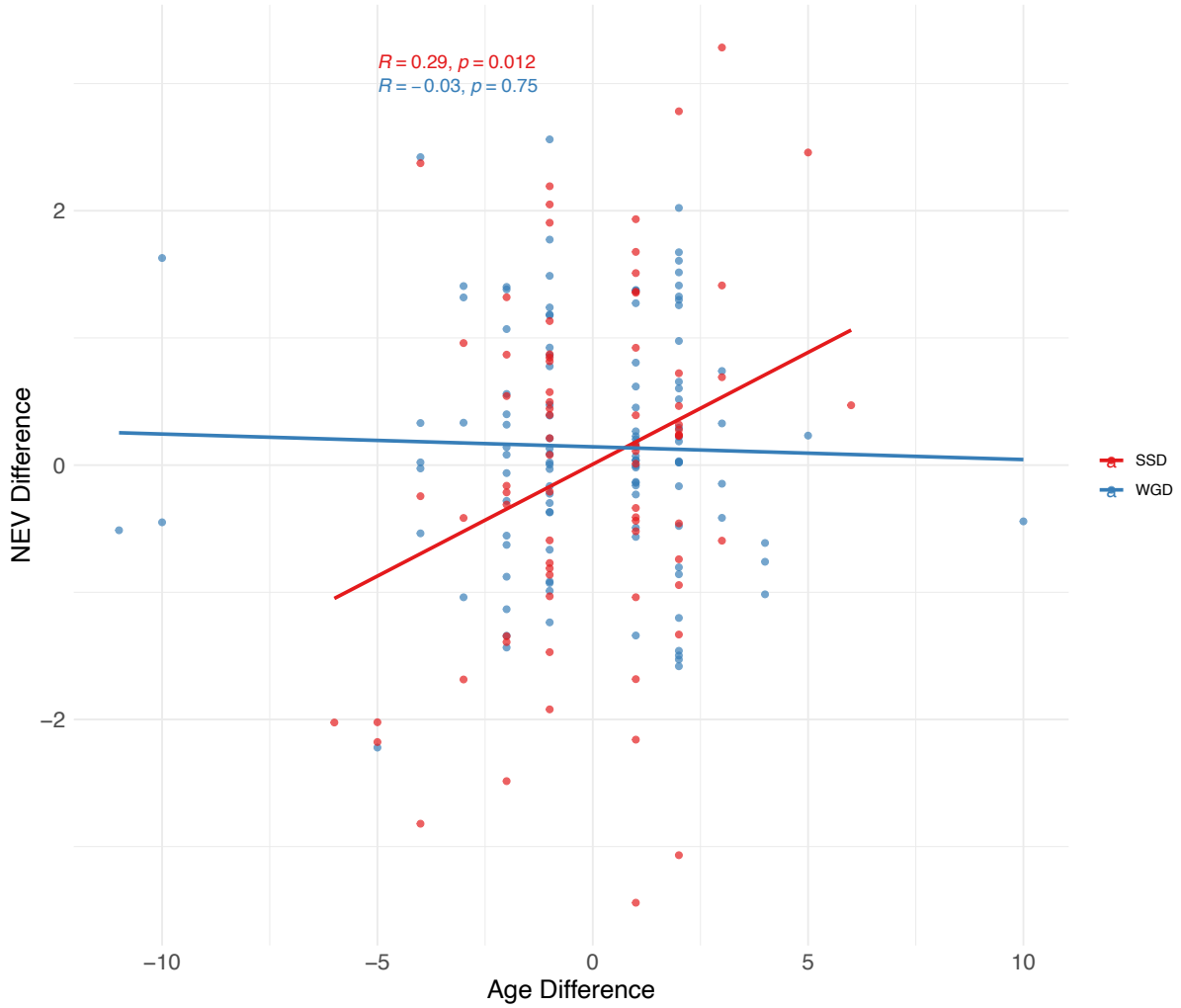

Figure 5: **Younger gene wthin a paralog exhibiting higher NEV for SSDs.** Values in x-axis are the difference of gene clades between paralogs and the y-axis is the difference of normalized expression variability (NEV) between paralogs. Paralog pairs with the same inferred clade are excluded (resulting in y = 0).