Published in final edited form as:

Trends Genet. 2007 April ; 23(4): 162–166.

External factors accelerate expression divergence between duplicate genes

Misook Ha¹, Wen-Hsiung Li², and Z. Jeffrey Chen¹

¹ Section of Molecular Cell and Developmental Biology and Institute for Cellular and Molecular Biology, One University Station, A-4800, University of Texas, Austin, TX 78712-0159, USA

²Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, IL 60637, USA

Abstract

We examined the evolution of expression of duplicate genes in *Arabidopsis thaliana*, by analyzing 512 data sets of gene expression microarrays and 2022 recent duplicate gene pairs. Expression divergence between gene duplicates is significantly greater in response to environmental stress than to developmental processes. A slow rate of expression divergence during development might offer dosage-dependent selective advantage, whereas rapid expression divergence in response to external changes might accelerate adaptation.

Introduction

The genomes of all eukaryotes have undergone at least one round of whole-genome duplication (WGD) (see Glossary) during their evolutionary history [1,2]. Duplicate genomes can undergo massive gene loss and genomic rearrangements, leading to a diploidized state, as shown in yeast [3], *Arabidopsis* [4-7] and rice [8]. During evolution one copy of the gene duplicate can be lost by the accumulation of deleterious mutations [9]. Evidently, many duplicate genes are retained because the redundancy conferred by duplicate genes might facilitate species adaptation [1] and genetic robustness against null mutations [10]. Both copies can be retained if a higher dosage is advantageous [11], or the function of the duplicate can diverge from that of the ancestral gene by subfunctionalization [12] (such as tissue specificity). Alternatively, one gene duplicate can evolve to possess a novel function by neofunctionalization [1,13].

How duplicate genes diverge in expression is a longstanding issue [1,14]. Of particular interest is what factors affect the rate of expression divergence between duplicate genes? This question has not been well explored, although some factors such as developmental constraint have been investigated [10,15,16]. Because environmental factors such as abiotic and biotic stresses tend to change faster than internal factors such as developmental programs, we hypothesize that environmental factors accelerate expression divergence between duplicate genes. Similarly, acceleration can also occur in the extracellular transport processes that are affected by environmental conditions. The duplicate genes in *Arabidopsis thaliana* that were derived from a WGD 20–40 million years ago [4-6] are ideal for testing our hypothesis because these duplicate genes are old enough to have accumulated a substantial degree of expression divergence but not too old to make statistical inferences difficult.

Corresponding author: Chen, Z.J. (zjchen@mail.utexas.edu).

Supplementary material associated with this article can be found at doi:10.1016/j.tig.2007.02.005.

Preferential induction of duplicate genes by abiotic and biotic stresses

To investigate the expression evolution of duplicate genes in response to exogenous processes, including external factors, we first studied how often these duplicate genes are induced by environmental stresses using microarray data analysis (see Methods, Figure S1 and Table S1 in Online Supplementary Material). The proportion of duplicate genes upregulated under abiotic stress in roots or shoots is significantly greater than that of other genes in the genome (2022 pairs of duplicate genes were compared with other genes in the genome; one-tailed t test, $P \leq 0.01$) (see Table S2a in Online Supplementary Material). We obtained the same conclusion for duplicate genes in response to biotic stress induced by pathogen infections or pathogenic molecules. Moreover, similar results were obtained from the gene duplicates that were downregulated in abiotic stress but not in biotic stress or cold stress in roots (see Table S2b in Online Supplementary Material), probably because systematic repression is a protective mechanism for susceptible genes [17-19]. The data suggest that duplicate genes are preferentially involved in stress responses, probably through preferential retention [7,11] or expression divergence of duplicate genes.

Expression diversity in response to developmental changes

We then studied how duplicate genes respond to endogenous processes, including developmental programs. The differentially regulated genes were detected across 79 different tissues using one-way analysis of variance (ANOVA) (see Table S3 in Online Supplementary Material). We found that the frequency of genes displaying differential expression in various developmental stages was greater in gene duplicates than in other genes. Among five representative tissues (leaf, flower, root, seed and pollen), the proportion of duplicate genes that were differentially expressed was significantly greater than that of the other genes in the genome (see Table S4 in Online Supplementary Material). The data suggest that duplicate genes increase expression diversity during development, similar to the findings in *Drosophila* and yeast [15].

Box 1. Expression correlation coefficients and gene ontology (GO)

In our analysis of expression correlations between gene duplicates we obtained similar results using rank correlation coefficients of both Pearson and Spearman, and the former results are presented in this study. The Pearson correlation coefficient (R_{ik}) between gene i and gene k was calculated as:

$$R_{ik} = \frac{S_{ik}}{\sqrt{S_{ii}}\sqrt{S_{kk}}} = \frac{\sum_{j=1}^n (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\sum_{j=1}^n (x_{ji} - \bar{x}_i)^2} \sqrt{\sum_{j=1}^n (x_{jk} - \bar{x}_k)^2}}$$

where x_{ji} is the expression value of gene i under condition j ; $\bar{x}_i$ is the mean expression value of gene i ; s_{ii} is a standard deviation of gene i expression across the conditions ($1 \dots n$) used in the analysis. R measures the strength of the linear association between two expression profiles.

The GO for *A. thaliana* was downloaded from The Arabidopsis Information Resource (TAIR) (ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/), released on 10 December 2005. GoSlim was used to classify 14 biological process categories. To ensure the accuracy of GO classification, only the GoSlim terms with experimental evidence were used for analysis. The evidence codes are IDA (inferred from direct assay), IGI (inferred from genetic interaction), IMP (inferred from mutant phenotype), IPI (inferred from

physical interaction), IEA (inferred from electronic annotation) and ISS (inferred from sequence or structural similarity). A duplicate gene pair was assigned to a GoSlim classification if at least one copy of the gene duplicate was annotated. The responses to abiotic, biotic and other stresses were combined into one category, namely, 'response to external stresses' because the GO terms included in the 'response to other stress' and 'response to abiotic and biotic stimuli' categories overlap considerably. The 'transport' category was divided into two groups: 'extracellular transport' (into or out of a cell) and 'intracellular transport', corresponding to the external and internal processes, respectively. Among the 2022 gene duplicate pairs, 1232 were assigned using GoSlim biological process classifications (see Methods and Tables S5 and S6 in Online Supplementary Material).

Faster expression divergence in response to environmental factors than to developmental processes

To evaluate the relative contributions of environmental and developmental factors to expression divergence between duplicate genes, we analyzed the Pearson correlation coefficient of expression (Box 1) between gene duplicates in developmental (R_{dev}) or environmental (R_{env}) processes using the same number of expression data sets: 63 in different developmental stages and 63 treatment and time-course combinations in roots and shoots, respectively. The distributions of expression correlation coefficients were compared using the Wilcoxon rank-sum test [20]. As expected, correlation coefficients of expression profiles between randomly chosen genes showed a normal distribution with mean zero, and there was no significant difference in expression variation among random gene pairs in all three conditions (data not shown). Interestingly, the expression divergence of duplicate genes under environmental stress is significantly greater than that under developmental process (Figure 1a, $P \leq 2.2 \times 10^{-16}$). Furthermore, we analyzed the correlation coefficient difference (D) of each gene duplicate in developmental and environmental processes ($D_i = R_{\text{dev},i} - R_{\text{env},i}$ for the i th duplicate gene pair). The cumulative probability difference for the gene duplicates and random gene pairs between environmental and developmental processes was significantly different by either the Kolmogorov–Smirnov (KS) test [21] or the Wilcoxon rank sum test (see Figure S2 in Online Supplementary Material). Taken together, the data indicate that expression divergence between gene duplicates occurs faster in response to the environmental stresses than to the developmental changes.

To test if external factors are more effective in promoting expression divergence than other biological processes, we classified recent WGD duplicate genes into Gene Ontology Slim (GOSlim) biological processes (ftp://ftp.Arabidopsis.org/home/tair/Ontologies/Gene_Ontology) [22] (Box 1) and analyzed expression correlation coefficients of gene duplicates in each category (Figure 1b). The levels of expression divergence between gene duplicates were greatest in extracellular transport, signal transduction, stress response and transcription, and lowest in the cellular and developmental processes such as energy pathway, protein metabolism, intracellular transport, DNA and RNA metabolism, and cell organization and biogenesis.

To infer the biological processes responsive to external conditions, duplicate genes in the 'transport' category were divided into extracellular and intracellular subgroups. Duplicate genes in the 'extracellular transport' subgroup showed the greatest level of expression divergence, whereas those in the 'intracellular transport' subgroup displayed a low level of expression divergence (Figure 1b). This supports the notion that gene expression divergence occurs at a faster rate in response to external than to internal factors. Note that biological processes in 'response to external stresses' and 'extracellular transport' would be directly affected by external conditions. So, we compared gene expression divergence in two groups:

‘other processes’ and ‘response to external factors’, which include external stresses (abiotic, biotic and other) and extracellular transport. The expression divergence is significantly faster in response to the external factors than to the other processes (Wilcoxon rank-sum test, $P \leq 1.29 \times 10^{-7}$) (Figure 1c). Note that this analysis might underestimate the difference because some potential external factors related to signal transduction and transcription were included in the ‘other processes’.

Biological implications

Duplicate genes display greater levels of expression diversity than do random gene pairs in response to external and internal processes. However, the duplicate genes involved in developmental processes tend to be coregulated, whereas the duplicate genes involved in abiotic and biotic stresses tend to diverge in expression (Figure 2). There is experimental support for the above conclusion. For example, *SEP1* (formerly *AGL2*) and *SEP2* (formerly *AGL4*) are gene duplicates expressed at the flower developmental stage [23], and their expression patterns are correlated throughout plant development (Figure 2, right lower panel). The two genes have a redundant function in the floral organ identity, and single-gene knockout shows no developmental defect [24]. By contrast, cyclophilin gene duplicates, *At2g21130* (*CYP1*) and *At4g38740* (*ROC1*), are induced by abiotic and biotic stresses (Figure 2, left lower panel) [25,26]. Their expression levels differ greatly in response to various external stimuli, suggesting that the gene duplicate is involved in different regulatory networks. Although some extreme examples of duplicate genes that show expression similarity or divergence exist in large data sets, the experimental data collectively support the above notion.

Environmental stresses are often associated with a short-term cascade or simple signal amplification (or both), leading to rapid changes in gene expression [27]. Therefore, external conditions could promote the acquisition by organisms of an adaptive mechanism, as predicted by McClintock [28], through diversification of duplicate genes [1,27] after WGD [4-7]. Many plants respond to environmental stresses (e.g. drought and salt) by inducing the expression of stress-related genes or gene products (or both) [17,18]. By contrast, developmental programs affect gene expression through long-term, multistage, complex molecular interactions, corresponding to a relatively slow rate of expression divergence between duplicate genes.

Concluding remarks

We propose a model (Figure 2) for different evolutionary fates of duplicate genes in response to external and internal processes. Duplicate genes diverge in expression relatively rapidly in response to abiotic and biotic stresses, which might facilitate subfunctionalization [12], neofunctionalization [1,13] and the evolution of an adaptive mechanism to cope with environmental changes [28]. In development, duplicate genes diverge in expression relatively slowly and tend to be coregulated. A relatively slow rate of expression divergence between the duplicates might provide genetic robustness against null mutations [10] and selective advantage by dosage-dependent gene regulation [11,29] that enables organisms to fine-tune complex regulatory networks through genetic and epigenetic mechanisms [30]. Therefore, duplicate genes could promote an adaptive mechanism for environmental changes or provide genetic robustness and dosage-dependent regulation during organismal development, which might facilitate polyploid evolution.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Justin Borovitz and reviewers for critical comments on the manuscript, and members in the Chen and Li laboratories for valuable suggestions. We especially thank Detlef Weigel and AtGenExpress Consortium for sharing the expression array data. The work was supported by grants from the NIH (W-H.L. and Z.J.C.) and NSF (Z.J.C.).

References

1. Ohno, S. Evolution by Gene Duplication. Springer-Verlag; 1970.
2. Wolfe KH, Li W-H. Molecular evolution meets the genomics revolution. *Nat Genet* 2003;33:255–265. [PubMed: 12610535]
3. Wolfe K, Shields DS. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 1997;387:708–713. [PubMed: 9192896]
4. Blanc G, et al. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res* 2003;13:137–144. [PubMed: 12566392]
5. Vision TJ, et al. The origins of genomic duplications in *Arabidopsis*. *Science* 2000;290:2114–2117. [PubMed: 11118139]
6. Bowers JE, et al. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 2003;422:433–438. [PubMed: 12660784]
7. Blanc G, Wolfe KH. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* 2004;16:1679–1691. [PubMed: 15208398]
8. Yu J, et al. The genomes of *Oryza sativa*: A history of duplications. *PLoS Biol* 2005;3:e38. [PubMed: 15685292]
9. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. *Science* 2000;290:1151–1155. [PubMed: 11073452]
10. Gu Z, et al. Role of duplicate genes in genetic robustness against null mutations. *Nature* 2003;421:63–66. [PubMed: 12511954]
11. Thomas BC, et al. Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res* 2006;16:934–946. [PubMed: 16760422]
12. Lynch M, Force A. The probability of duplicate gene preservation by subfunctionalization. *Genetics* 2000;154:459–473. [PubMed: 10629003]
13. Lynch M, et al. The probability of preservation of a newly arisen gene duplicate. *Genetics* 2001;159:1789–1804. [PubMed: 11779815]
14. Li WH, et al. Expression divergence between duplicate genes. *Trends Genet* 2005;21:602–607. [PubMed: 16140417]
15. Gu Z, et al. Duplicate genes increase gene expression diversity within and between species. *Nat Genet* 2004;36:577–579. [PubMed: 15122255]
16. Stolc V, et al. A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* 2004;306:655–660. [PubMed: 15499012]
17. Xiong L, et al. Cell signaling during cold, drought, and salt stress. *Plant Cell* 2002;14(Suppl):S165–S183. [PubMed: 12045276]
18. Seki M, et al. Monitoring the expression pattern of 1300 *Arabidopsis* genes under drought and cold stresses by using a full-length cDNA microarray. *Plant Cell* 2001;13:61–72. [PubMed: 11158529]
19. Chisholm ST, et al. Host-microbe interactions: shaping the evolution of the plant immune response. *Cell* 2006;124:803–814. [PubMed: 16497589]
20. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics Bulletin* 1945;1:80–83.
21. Massey FJJ. The Kolmogorov-Smirnov test of goodness of fit. *J Am Stat Assoc* 1951;46:68–78.
22. Berardini TZ, et al. Functional annotation of the *Arabidopsis* genome using controlled vocabularies. *Plant Physiol* 2004;135:745–755. [PubMed: 15173566]
23. Jack T. Relearning our ABCs: new twists on an old model. *Trends Plant Sci* 2001;6:310–316. [PubMed: 11435170]
24. Pelaz S, et al. B and C floral organ identity functions require SEPALLATA MADS-box genes. *Nature* 2000;405:200–203. [PubMed: 10821278]

25. Luan S, et al. pCyP B: a chloroplast-localized, heat shock-responsive cyclophilin from fava bean. *Plant Cell* 1994;6:885–892. [PubMed: 8061522]
26. Marivet J, et al. Effects of abiotic stresses on cyclophilin gene-expression in maize and bean and sequence-analysis of bean cyclophilin cDNA. *Plant Sci* 1992;84:171–178.
27. Seoighe C, Gehring C. Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. *Trends Genet* 2004;20:461–464. [PubMed: 15363896]
28. McClintock B. The significance of responses of the genome to challenge. *Science* 1984;226:792–801. [PubMed: 15739260]
29. Bomblies K, Doebley JF. Pleiotropic effects of the duplicate maize FLORICAULA/LEAFY genes *zfl1* and *zfl2* on traits under selection during maize domestication. *Genetics* 2006;172:519–531. [PubMed: 16204211]
30. Chen ZJ, Ni Z. Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. *Bioessays* 2006;28:240–252. [PubMed: 16479580]

Glossary

Endogenous process

a biological process, such as organ differentiation, that involves internal signals and developmental switches from vegetative to reproductive growth

Exogenous process

a biological process that involves external stimuli such as abiotic and biotic stresses

Neofunctionalization

gain of a novel function (or expression pattern) from a duplicate gene

Polyploid

an individual or cell that has two or more basic sets of chromosomes

Subfunctionalization

divergence of function (or expression) of a duplicate gene from that of the ancestral gene (e.g. tissue-specific expression of a duplicate gene)

Whole-genome duplication (WGD)

doubling of the entire basic set of chromosomes. WGD can lead to autopolyploidy (doubling of a single genome) or allopolyploidy (combination of divergent genomes)

Figure 1.

(a) Distributions of expression correlations between gene duplicates with regard to environmental factors and developmental processes. The probability density of expression correlation coefficient is plotted against correlation coefficients. The distribution of expression correlations between duplicate genes in various developmental stages (red) is significantly shifted to the right of those subjected to abiotic treatments in roots (green) and in shoots (blue) (Wilcoxon rank sum test, $P \leq 2.2 \times 10^{-16}$). The right shift indicates less expression divergence. The approximately normal distributions observed in the roots and shoots suggest a high level of expression divergence. (b) External factors accelerate expression divergence between duplicate genes. The boxplots show distribution of expression correlation between duplicate genes in biological process classifications (except for biological functions unknown). Boxplots of expression correlations indicate distributions of 75th percentiles, median (black dots) and 25th percentiles. The levels of gene expression divergence were arranged in order based on the median values, which increased from the external factors (left) to the internal factors (right). The horizontal line shows the median correlation coefficient using a total of 2022 gene duplicates. (c) Gene duplicates that respond to external factors are significantly more diverged

than gene duplicates in other processes. Gene duplicates in the extracellular transport and in response to external stresses were grouped into one category ('Response to external factors'), and all other gene duplicates were considered to be in the 'Other processes' category. Dotted vertical lines indicate the range of gene expression divergence.

Figure 2.

Different strategies for the evolution of expression of duplicate genes in the external and internal processes. Upper panels: simple and short-term external signals (e.g. abiotic stress) could induce rapid expression divergence of duplicate genes, whereas complex and long-term internal factors (e.g. developmental program) could decelerate expression divergence of duplicate genes. Lower panels: expression intensities were standardized into the z score (z_{ij}) of gene i in condition j as:

$$z_{ij} = \frac{(x_{ij} - x_i)}{\sqrt{s_{ii}}}$$

where x_{ij} is an expression value before standardization; x_i is a mean expression value of gene i ; and s_{ii} is a standard deviation of gene i expression across the conditions ($1 \dots j$). In response to external factors such as abiotic and biotic stresses, gene duplicates (*CYP1* and *ROC1*) underwent fast expression divergence under various stress responses, leading to the evolution of an adaptive mechanism for adverse growth environments. By contrast, duplicate genes (*SEP1* and *SEP2*) encoding proteins involved in floral organ identity are both differentially regulated during developmental processes, but their expression patterns are greatly correlated. Abbreviations: C, cold; D, drought; G, genotoxic; H, heat; O, osmosis; OX, oxidative stress; S, salt; UV, UV-B; W, wounding; L, leaf; F, flower; R, root; SS, seed and seedling; and O, others.