Knocking out a gene from a genome often causes no phenotypic effect. This phenomenon has been explained in part by the existence of duplicate genes. However, it was found that in mouse knockout data duplicate genes are as essential as singleton genes. Here, we study whether it is also true for the knockout data in Arabidopsis. From the knockout data in Arabidopsis thaliana obtained in our study and in the literature, we find that duplicate genes show a significantly lower proportion of knockout effects than singleton genes. Because the persistence of duplicate genes in evolution tends to be dependent on their phenotypic effect, we compared the ages of duplicate genes whose knockout mutants showed less severe phenotypic effects with those with more severe effects. Interestingly, the latter group of genes tends to be more anciently duplicated than the former group of genes. Moreover, using multiple-gene knockout data, we find that functional compensation by duplicate genes for a more severe phenotypic effect tends to be preserved by natural selection for a longer time than that for a less severe effect. Taken together, we conclude that duplicate genes contribute to genetic robustness mainly by preserving compensation for severe phenotypic effects in A. thaliana.

Materials and Methods

Phenotypic Effects of Single-Gene Knockout

We obtained 5041 insertional mutant lines from our mutant sources (Kuromori et al. 2006) and from the Arabidopsis Biological Resource Center (Columbus, OH). In each insertional mutant line, the tag was inserted into the coding region of a TAIR7 gene. In many cases, a gene had multiple insertional mutant lines, but the phenotypic effect was the same for all lines. Thus, from the 5041 insertional mutant lines, we found 3871 genes that had an insertion(s) of Ds transposon or T-DNA. Among these 3871 genes with insertion, only 253 showed visible phenotypic changes. In addition, we found 1489 genes that had been described in the literature as inducing phenotypic change. In summary, among the 5360 genes included in this study, 1742 did not show induced phenotypic changes, whereas 3618 did not (Table S1, Supplementary Material online).

Identification of Singletons, Duplicates, and Closest Paralog Genes

The amino acid sequences of A. thaliana (TAIR7) and mouse genes (NCBIM37.55) were obtained from TAIR (www.arabidopsis.org) and Ensembl (www.ensembl.org).
The longest amino acid sequence in a gene locus was used as the representative protein sequence. Similarity searches of all against all genes were conducted using BlastP (Altschul et al. 1997). A singleton protein was defined as a protein that did not match any other protein in the Blast search with the E value ≤1. To find the closest paralog of a protein, we aligned the protein and all proteins obtained in the Blast search by ClustalW (Thompson et al. 1994) and estimated the amino acid similarity between sequences. The protein with the highest similarity to the one under study was defined as the closest paralog. Following the definition of Gu et al. (2003), a gene is defined as a duplicate if it showed ≥30% similarity and ≥50% alignable region at the protein level with the closest paralog. The synonymous divergence (Ks) and the nonsynonymous divergence (Ka) between the target gene and the closest paralog were estimated by the modified Nei–Gojobori method (Zhang et al. 1998).

### Results and Discussion

#### Functional Compensation by Duplicate Genes

We examined the phenotypic changes in each of the 3871 knockout mutants and found phenotypic changes in only 253 genes. Furthermore, from the literature we found 1489 genes whose knockout mutants showed phenotypic changes. We then computed the ratio of duplicate and singleton genes with and without phenotypic changes (Table 1). Genes with phenotypic changes have a significantly lower proportion of duplicate genes and a higher proportion of singleton genes than genes without phenotypic changes ($P = 2.0 \times 10^{-11}$, $\chi^2$ test), indicating that the functional compensation is significantly more common in duplicate genes than in singleton genes in the Arabidopsis genome.

Because the age (sequence divergence) of duplicate genes may influence the chance of functional compensation, we computed the p-distance (proportion of amino acid differences) and the synonymous distance (Ks) between a knockout gene and its closest paralog. Because tandem duplicates may have a higher chance of gene conversion than nontandem duplicates (Gao and Innan 2004), we also did the same analyses using the data set without tandem duplicates. In both data sets, genes without phenotypic effects have, on average, significantly lower p-distance and Ks to the closest paralog than genes with phenotypic effects ($P < 0.01$ in p-distance, $P < 0.01$ in Ks). See Table S2 (Supplementary Material online) for $P$ values.

#### Defining Gene Families and Tandem Duplicate Clusters

To define gene families in *A. thaliana*, all-against-all similarity searches of protein sequences were conducted using BlastP with an E value cutoff of $10^{-5}$. Based on the E values, we generated similarity clusters representing gene families with the Markov clustering program (van Dongen 2000). As reported earlier (Hanada et al. 2008), tandem duplicates were defined as genes in any gene pair, T1 and T2, that 1) belong to the same gene family, 2) are located within 100 kb of each other, and 3) are separated by at most ten nonhomologous (not in the same gene family as T1 and T2) genes.

#### Knockout Effects on Phenotypes

Numbers of Duplicates and Singletons with or without Knockout Effects on Phenotypes

<table>
<thead>
<tr>
<th>Phenotypic changes</th>
<th>Number of duplicates (D)</th>
<th>Number of singletons (S)</th>
<th>D/S ratio</th>
<th>$P$ value (Yes vs. No)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes (either our study or literature)</td>
<td>1,168</td>
<td>98</td>
<td>12.0</td>
<td>$2.0 \times 10^{-11}$</td>
</tr>
<tr>
<td>Yes (our study)</td>
<td>166</td>
<td>19</td>
<td>8.7</td>
<td>$1.6 \times 10^{-5}$</td>
</tr>
<tr>
<td>Yes (literature)</td>
<td>1,002</td>
<td>79</td>
<td>12.7</td>
<td>$4.4 \times 10^{-8}$</td>
</tr>
<tr>
<td>No (our study)</td>
<td>2,903</td>
<td>122</td>
<td>23.8</td>
<td></td>
</tr>
</tbody>
</table>

![Fig. 1.—Sequence divergence in genes with and without phenotypic effects in single-gene knockout. Sequence divergences are in terms of the p-distance (proportion of amino acid differences) and the synonymous distance (Ks) between a knockout gene and the closest paralog. The three data sets (O, L, and N) used are 1) with phenotypic changes observed in our study (O), 2) with phenotypic changes reported in the literatures (L), and 3) no phenotypic changes observed in our study (N). A. The p-distance in the entire data. B. The p-distance in the data without tandem duplicates. C. The Ks in the entire data. D. The Ks in the data without tandem duplicates. The distributions of p-distance and Ks values are shown as box plots with the thick solid horizontal line indicating the median value, the box representing the interquartile range (25–75%), and the dotted lines indicating the first to the 99th percentile. The p-distance and Ks are compared between genes with and without phenotypic effects. Genes with phenotypic effects have a significantly higher p-distance and Ks to the closest paralog than genes without any phenotypic effects ($P < 0.01$ in p-distance, $P < 0.01$ in Ks). See Table S2 (Supplementary Material online) for $P$ values.](image-url)
or an abiotic treatment. Among the 1742 genes whose knockout mutants showed phenotypic changes, 279, 219, 882, and 362 were classified as genes whose knockout mutants showed changes as the seed, reproductive, vegetative, and conditional phenotypes, respectively. From the 1742 genes, we randomly chose 10,000 pairs of genes either within a gene family or between gene families, and examined the ratio between the number of gene pairs with the same kind of phenotypes (seed, reproductive, vegetative, or conditional) and the number of gene pairs with different kind of phenotypes. We compared the ratios within and between gene families (fig. 2) and found that genes in the same gene family tend to have higher ratios than genes from different families (\(P < 1 \times 10^{-15}\), the \(t\)-test), indicating that a gene family tends to be persistently maintaining its phenotype.

Differential Persistence in Functional Compensation among Phenotypes

To examine the evolutionary persistence of functional compensation for different phenotypes, we compared the ages of duplicate genes among the four phenotypes in either the entire data set or the data set excluding tandem duplicates (fig. 3). In both data sets, genes whose knockout mutants showed changes in seed, reproductive, vegetative, conditional, and no-phenotypes are ranked in that order of oldest to youngest duplication dates. That is, it is likely that seed, reproductive, vegetative, conditional, and no-phenotypes are ranked in the order of longest to the shortest persistence of functional compensation.

If natural selection affects the persistence of functional compensation, the turnover rate is likely to be slower in genes with functional compensation to more severe phenotypes than genes with functional compensation to less severe phenotypes. We assume that the different phenotypes can be ranked in the order of highest to lowest significance for plant survival. The relationship between phenotype and significance can be explained as follows. Among the four phenotypes studied, an effect on seed phenotype might be most severe because a mutant with an altered seed phenotype may have embryonic defects and may not be able to germinate. Thus, functional compensation by duplicate genes for this phenotype may persist longer than the three other phenotypes. In the case of mutants with changes in reproductive phenotype, these changes largely affect the next generation and seem to reduce survivability. Therefore, the reproductive phenotype is generally severe but less severe for survival than the seed phenotype. For the vegetative phenotype, some mutations in this phenotype do not affect survival, so that this phenotype is likely to be less severe than that of either seed or reproductive phenotype. Changes in conditional phenotype appear only under certain treatments, so the significance of this phenotype is likely to be minimal. No-change phenotype is undoubtedly the least severe for plant survival. We therefore propose the order from the highest to the lowest significance in phenotypes as seed, reproductive, vegetative, and conditional and no-change.

Under this ranking of significance, we tested differences in either p-distance or Ks distance between different pairs of phenotypes by the Wilcoxon test (fig. 3 and Table S2,
Supplementary Material online). We found that genes whose knockout mutants showed more severe phenotypes tend to be more anciently duplicated, whereas those whose knockout mutants showed less severe phenotypes are more recently duplicated. Thus, this result supports our expectation that functional compensation to more severe phenotypes tends to persist for a longer time, whereas functional compensation to less severe phenotypes tends to disappear faster. Thus, this relationship between duplication age and phenotypic significance supports the hypothesis that natural selection affects the persistence of functional compensation.

Different Selection Pressures on Functional Compensation among Phenotypes

We then tried to evaluate selection pressures using duplicate genes with direct evidence of functional
compensation. “Direct evidence” means the observation of phenotypic change only upon knocking out multiple duplicate genes; knocking out a single gene did not induce phenotypic change (Table S1, Supplementary Material online). A total of 163 nontandem duplicate genes that were related to 75 kinds of phenotypic change were classified into 26, 89, 32, and 16 genes whose multiple-knockout mutants showed phenotypic changes in the seed, vegetative (V), reproductive (R), or conditional (C) phenotypes, respectively. Using the above genes, we examined whether the pressure of natural selection on functional compensation by duplicate genes depends on the phenotype involved. Because selection pressure on genes is commonly estimated by the ratio of the nonsynonymous substitution rate (Ka) to the synonymous substitution rate (Ks), we compared the Ka/Ks ratios of the genes whose knockout mutants showed phenotypic changes in deletion of multiple duplicate genes. Differences in Ka/Ks were tested between phenotypes by the Wilcoxon test (fig. 4 and Table S2, Supplementary Material online). Although the difference in Ka/Ks was not significant between some phenotype pairs, the Ka/Ks ratio increased with decreasing phenotypic significance as a whole, indicating different selection pressures. Indeed, duplicate genes with functional compensation for more severe phenotypes tend to be under strong purifying selection (a low Ka/Ks ratio), and those with functional compensation for less severe phenotypes are subject to relaxed purifying selection (a high Ka/Ks ratio). Because recently (low Ks and p-distance values) and anciently (high Ks and p-distance values) duplicated genes tend to have high and low Ka/Ks ratios, respectively (Lynch and Conery 2000), the differences in Ks and p-distance may mislead the relationship between selection pressures and phenotypes in the case where genes with functional compensation for more severe phenotypes have high Ks and p-distance values. However, genes with functional compensation for more severe phenotypes tend to have lower Ks and P-distance values (fig. 4 and Fig. S1, Supplementary Material online). That is, genes with functional compensation for more severe phenotypes have lower Ks values and lower Ka/Ks ratios in comparison with genes with functional compensation for less severe phenotypes (fig. 4 and Fig. S1, Supplementary Material online). Therefore, duplication age unlikely affects our conclusion. These results indicate that there is a stronger tendency for selection pressure to preserve functional compensation to severe phenotype by duplicate genes. Thus, duplicate genes contribute to genetic robustness mainly by preserving compensation of severe phenotypic effects in the A. thaliana genome.

Conclusion

From single-gene knockout data in A. thaliana obtained in our study and from the literature, we found that duplicate genes play a significant role in functional compensation. This conclusion is similar to that in nematode and yeast but different from that in mouse (Gu et al. 2003; Conant and Wagner 2004; Liang and Li 2007; Liao and Zhang 2007). Because mouse is a more complex organism than nematode and yeast, it was proposed that mouse duplicate genes may tend to undergo functional divergence instead of preserving functional compensation (Zhang 2003; Liao and Zhang 2007). However, this view is not supported by our observation in A. thaliana. A possible reason for the difference between mouse and A. thaliana duplicate genes is differences in duplication ages because genes with functional compensation tend to be more recently duplicated (Fig. 1). To examine the duplication age between mouse and A. thaliana, we calculated the p-distance and the Ks between each duplicate gene and its closest paralog in mouse and A. thaliana (Fig. S2, Supplementary Material online). Both p-distance and Ks are, on average, significantly lower in A. thaliana than in mouse (\(P < 1 \times 10^{-167}\) in p-distance, \(P < 1 \times 10^{-120}\) in Ks) by the Wilcoxon test, indicating that A. thaliana genome has a higher proportion of recently duplicated genes than the mouse genome.

Another interesting finding in this study is that the evolutionary persistence of functional compensation by duplicate genes depends on the phenotype involved. That is, functional compensation by duplicate genes tends to persist for a longer time for a more severe phenotype than a less severe phenotype. We therefore propose that natural
selection retains functional compensation to increase genetic robustness, especially by preserving the functional compensation to more significant phenotype. Evidence of functional compensation by duplicate genes was provided by the observation of phenotypic change only upon knocking out multiple duplicate genes. Moreover, we found that the duplicate genes with functional compensation of more severe phenotypes tend to be under strong purifying selection (a low Ka/Ks ratio), and those with functional compensation of less severe phenotypes are subject to relaxed selection (a high Ka/Ks ratio). In conclusion, one reason why many duplicated genes are retained in the Arabidopsis genome is because some or many of them contribute to genetic robustness and are retained by natural selection.

Supplementary Material

Supplementary materials Table S1, Table S2, Figure S1 and Figure S2 are available at Genome Biology and Evolution online (http://www.oxfordjournals.org/our_journals/gbe/).

Acknowledgments

We thank Kenneth H Wolfe, Takashi Makino, and Nishiyama Tominaga for discussions. We also thank TAIR (http://www.arabidopsis.org/) for giving us gene sequence data. This work was supported by the RIKEN Plant Science Center and by National Institutes of Health grant (GM30998 to W.H.L.).

Literature Cited


Tischler J, Lehner B, Chen N, Fraser AG. 2006. Combinatorial RNA interference in Caenorhabditis elegans reveals that redundancy between gene duplicates can be maintained for more than 80 million years of evolution. Genome Biol. 7:R69.


Marta Wayne, Associate Editor

Accepted October 21, 2009